De-Shuang Huang, Bo Li, Haiming Chen, Chuanlei Zhang

Poster Proceedings
Volume 0002

The 2025 Third
International Conference on Applied Intelligence
(ICAI 2025)

November 6-9, 2025

Nanning, Guangxi, China

Editors

De-Shuang Huang
Ningbo Institute of Digital Twin, Eastern Institute of Technology, Ningbo, China
Tongxin Road No.568, Kaiyuan Xinqingnian Plaza, Building 1, Floor 8, Ningbo 315201, Zhejiang, China
dshuang@eitech.edu.cn

Bo Li
School of Computer Science & Technology, Wuhan University of Science and Technology
Heping road 947, 430081, Wuhan, China
libo@wust.edu.cn

Haiming Chen
Faculty of Electrical Engineering and Computer Science, Ningbo University
Fenghua Road No.818, Ningbo University, Ningbo 315211, Zhejiang, China
chenhaiming@nbu.edu.cn

Chuanlei Zhang
College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin, China
Thirteenth Street No.9, Western Campus at Taida, Building 8, Floor 4, Tianjin 300457, Tianjin, China
97313114@tust.edu.cn

ISSN 3134-5670
Poster Proceedings of the International Conference on Applied Intelligence
Open Access: http://poster-openaccess.com/
Publication Date: 2025/12/20

© The Editor(s), The Author(s), and the International Conference on Applied Intelligence (ICAI) / The Society of International Computing (SIC), 2025.

This work is subject to copyright. All rights are reserved by the International Conference on Applied Intelligence (ICAI) and The Society of International Computing (SIC), specifically the rights of reproduction, translation, reprinting, reuse of illustrations, broadcasting, adaptation, and storage in electronic form.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publishers, editors, and authors are safe to assume that the advice and information in this proceeding are believed to be true and accurate at the date of publication. Neither the publishers nor the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

This publication is issued by the International Conference on Applied Intelligence (ICAI) and The Society of International Computing (SIC).

Preface

The third International Conference on Applied Intelligence (ICAI 2025) was started to provide an annual forum dedicated to the emerging and challenging topics in artificial intelligence, machine learning, pattern recognition, bioinformatics, and computational biology. It aims to bring together researchers and practitioners from both academia and industry to share ideas, problems, and solutions related to the multifaceted aspects of intelligent computing.

ICAI 2025, held in Nanning, Guangxi, China, November 6-9, 2025, constituted the third International Conference on Applied Intelligence. It built upon the success of ICAI 2024 (Zhengzhou, China) and ICAI 2023 (Nanning, China).

This year, the conference concentrated mainly on the theories and methodologies as well as the emerging applications of Applied Intelligence. Its aim was to unify the picture of contemporary Applied Intelligence techniques as an integral concept that highlights the trends in advanced computational intelligence and bridges theoretical research with applications. Therefore, the theme for this conference was "Advanced Applied Intelligence Technology and Applications". Papers that focused on this theme were solicited, addressing theories, methodologies, and applications in science and technology.

ICAI 2025 received 289 submissions from 5 countries and regions. All papers went through a rigorous peer-review procedure and each paper received at least two review reports. Based on the review reports, the Program Committee finally selected 70 high-quality papers for presentation at ICAI 2025, included in volumes of proceedings published by Springer: two volumes of Communications in Computer and Information Science (CCIS).

In addition, this year we selected 3 Poster papers from the left papers, which will be arranged on the open access website http://poster-openaccess.com/.

The organizers of ICAI 2025, including the Society of International Computing, Nanning University, Ningbo Key Laboratory of Multi-Omics Multimodal Biomedical Data Mining and Computing, Guangxi Key Laboratory of Human-Machine Interaction and Intelligent Decision, Guangxi Association for Artificial Intelligence, made an enormous effort to ensure the success of the conference. We hereby would like to thank the members of the Program Committee and the referees for their collective effort in reviewing and soliciting the papers. In particular, we would like to thank all the authors for contributing their papers. Without the high-quality submissions from the authors, the success of the conference would not have been possible. Finally, we are especially grateful to the International Neural Network Society, and the National Science Foundation of China for their sponsorship.

De-Shuang Huang
Bo Li
Haiming Chen
Chuanlei Zhang

Contents

Protein Structure and Function Prediction

ProAttUnet: Protein Secondary-Structure Prediction Re-imagined via ESM2-Enhanced U-Net Dual-Fusion 1

Long Cheng, Zhiqiang Hui, and Anchi Sun

Biomedical Data Modeling and Mining

Optimization and Validation of Transformer-Based PTM Site Prediction Model for Paeonia lactiflora 5

Kai Xiao and Wenzheng Bao

Prediction of Post-Translational Modification Sites of Paeonia lactiflora Proteins Based on Attention Mechanism 19

Yingyue Tang and Wenzheng Bao

2025 International Conference on Applied Intelligence
November 6-9, Nanning, Guangxi, China
http://www.icai.org.cn/2025/index.php

ProAttUnet: Protein Secondary-Structure Prediction Re-imagined via ESM2-Enhanced U-Net Dual-Fusion

Long Cheng, Zhiqiang Hui, Anchi Sun

Suzhou University of Science and Technology

Abstract. Protein secondary structure prediction remains a pivotal concern within the domain of bioinformatics. In this innovative research, we introduce a novel methodology to further enhance a protein prediction model grounded in single sequences. Our key contribution lies in integrating the state-of-the-art (SOTA) model ESM2, which hails from the field of universal protein language models. By leveraging ESM2, we are able to acquire residual embeddings and contact maps for the protein sequences under study. Regarding the model architecture, we employ a unique dual-way U-Net framework for effective feature fusion. This framework is complemented by the integration of a cross-attention mechanism, enabling the model to capture more comprehensive context information. To better capture both local and global characteristics of protein sequences, we introduce the GCU_SE module into both the encoder and decoder. This module cascades a Gated Convolutional Unit (GCU) with a Squeeze-and-Excitation (SE) block: the GCU employs a dual-branch convolutional structure—one branch with ReLU activation and the other with Sigmoid gating—to selectively extract local features; the subsequent SE block performs global channel-wise recalibration, emphasizing informative channels. Unlike the standard SE block that only reweights channels globally, GCU_SE synergizes local feature selection with global channel refinement, enabling the model to more effectively perceive complex structural motifs. These innovative enhancements enable the ProAttUnet model to outperform the benchmark model SPOT-1D-Single by 1.6%, 3.5%, 1.0%, 4.6%, and 7.2% for ss3, and by 5.5%, 7.8%, 4.1%, 8.1%, and 10.1% for ss8 across five test sets (SPOT-2016, SPOT-2016-HQ, SPOT-2018, SPOT-2018-HQ and TEST2018, respectively). This significant improvement vividly demonstrates the effectiveness and novelty of our proposed model.

1 Methods

The study leverages the 650M-parameter ESM2 pretrained protein language model (based on BERT with RoPE) to extract residual embeddings (1280-dimensional per amino acid) and contact maps of protein sequences. Its architecture is a dual-path U-Net, including an encoder (3 downsampling modules + TCN layer, with max pooling and residual connections) and a decoder (3 upsampling modules via 1D convolution, with feature concatenation). It integrates cross-attention (query from embeddings, key/value from contact maps; 128/256/512 kernels) and GCU_SE modules in both encoder and decoder. Each GCU_SE block first applies a Gated Convolutional Unit: the

input is processed by two parallel 3×1 convolutions—one followed by ReLU to generate main features, the other by Sigmoid to produce a soft gate; their element-wise product suppresses noisy positions while preserving local motifs. The gated output is then fed into a Squeeze-and-Excitation layer that globally pools, excites channel-wise weights through a bottleneck MLP (reduction 16×) and recalibrates the feature maps, achieving synergistic local selection and global channel refinement. A dynamic sliding window (window size 10–100, step size calculated via formula) handles variable sequence lengths. Evaluation uses accuracy, precision, recall, F1-score, and SOV. Experiments run on an Intel Xeon Platinum 8362 CPU, NVIDIA RTX 3090 GPU (CUDA 11.2), with TensorFlow 2.5 and Python 3.8.

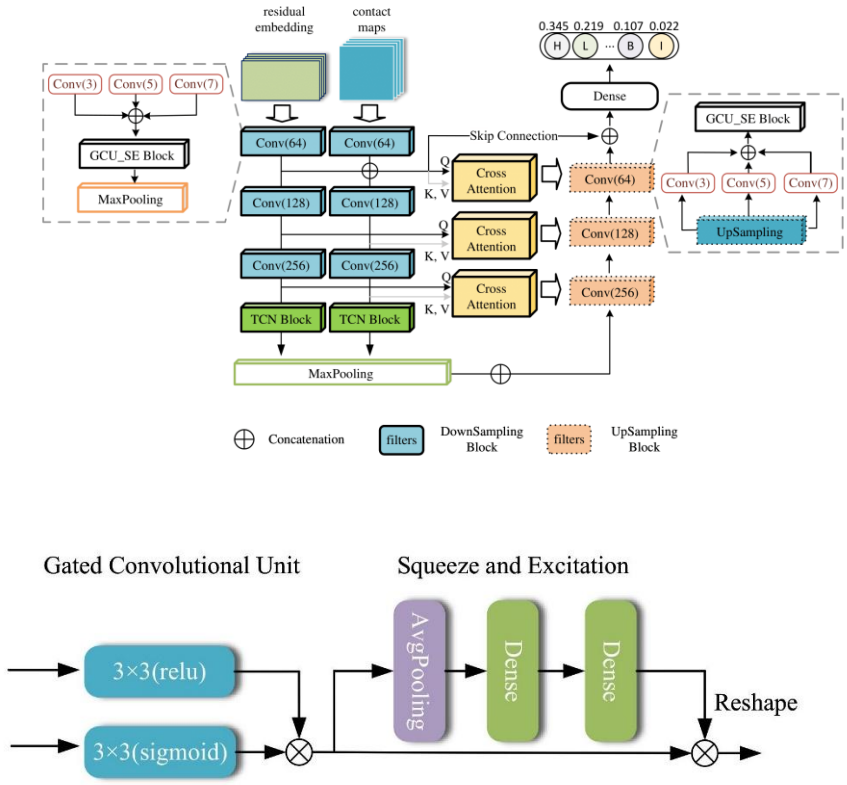


Figure. 1. The model's schematic diagram and the architecture of GCU_SE Block.

2 Datasets

Training relies on ProteinNet: initially 50,914 PDB entries (2016), 39,120 left after 95% truncation, 38,915 post-cleaning, and 25,166 selected (length 100–400, avoiding redundancy/noise). Evaluation uses 5 test sets: SPOT-2016 (2016–2020 PDB proteins, E-cutoff <0.1 excluded), SPOT-2016-HQ (SPOT-2016 with resolution <2.5Å, R-free

<0.25), SPOT-2018 (SPOT-2016-based, post-2018 proteins), SPOT-2018-HQ (SPOT-2018 with same HQ constraints), and TEST2018 (2018 proteins, 25% pre-2018 filter). Ablation uses CB513, TS115, CASP12 (sequence similarity >25%), and NEW364 (post-2019 proteins, length ⩾20, resolution <2.5Å). Dataset size experiments test 1000–15,000 sequences, with 100 sequences as validation.

Table 1. The 8 category labels in 4 testsets

ss types\testsets	SPOT-2016	SPOT-2016-HQ	SPOT-2018	SPOT-2018-HQ
H	99298	17486	34747	7183
L	93610	15251	36779	6476
B	2014	421	805	199
E	31884	10165	13503	4488
G	6516	1766	2466	772
I	299	10	19	5
T	23153	4908	8315	2105
S	23018	3585	8237	1596
Overall Samples	1473	295	682	125

3 Results

ProAttUnet outperforms SPOT-1D-Single (benchmark) across 5 test sets: for ss3, it is 1.6% (SPOT-2016), 3.5% (SPOT-2016-HQ), 1.0% (SPOT-2018), 4.6% (SPOT-2018-HQ), 7.2% (TEST2018) higher; for ss8, 5.5%, 7.8%, 4.1%, 8.1%, 10.1% higher (TEST2018 accuracy 72.2%). It has better precision/recall/F1 (e.g., superior AUC-PR for L/H/E/T on TEST2018) and SOV (10%/13%/8% higher for H/L/E vs SPOT-1D-Single; 17% higher for S, 2% for B). Removing GCU_SE reduces accuracy by 1.0–1.8% (e.g., TEST2018 -1.7%) and SOV. Dataset size ≥13,000 slows accuracy gain, 14,000 is optimal. Sliding window (25–150) boosts rare structures (B/G/S); 3 downsampling modules balance accuracy/error.

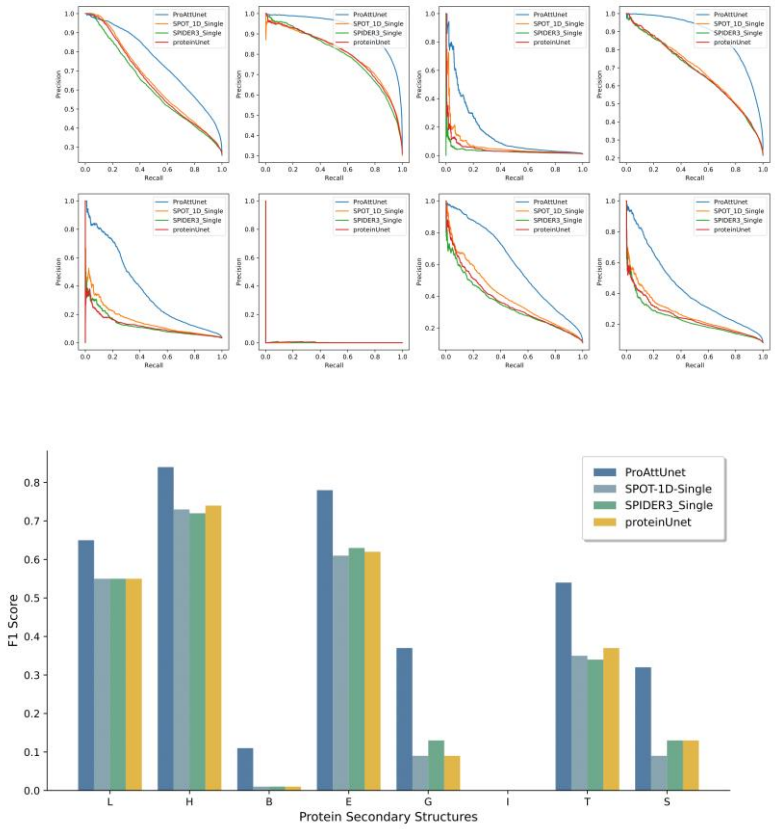


Figure. 2. The precision-recall curve(L,H,B,E,G,I,T,S in order) and F1 score on 8 categories and 4 models.

4 Conclusions

This study proposes ProAttUnet, a single-sequence-based model advancing protein secondary structure prediction. By integrating the ESM2 pretrained protein language model (650M parameters) for embeddings/contact maps, a dual-path U-Net with cross-attention for feature fusion, and GCU_SE modules for feature refinement, it outperforms the benchmark SPOT-1D-Single across five test sets (ss3: 1.0–7.2% higher; ss8: 4.1–10.1% higher). Ablation experiments confirm GCU_SE, dynamic sliding window (aiding rare structures like B/G/S), 3 downsampling modules, and 14,000-sequence dataset size optimize performance. Future work will explore knowledge graph multimodal learning and novel attention mechanisms to further enhance prediction accuracy.

2025 International Conference on Applied Intelligence
November 6-9, Nanning, Guangxi, China
http://www.icai.org.cn/2025/index.php

Optimization and Validation of Transformer-Based PTM Site Prediction Model for Paeonia lactiflora

Kai Xiao[1] and Wenzheng Bao[2,3]

[1] School of Information Science, University of Jinan, Jinan, China, 250022
[2] Institute for Regenerative Medicine, Medical Innovation Center and State Key Laboratory of Cardiology, Shanghai East Hospital, School of Medicine, Tongji University, Shanghai 200123, P. R. China
[3] Xuzhou University of Technology, Xuzhou, China, 221018

Abstract. Post-translational modifications (PTMs) significantly regulate peony's growth, stress resistance, and biosynthesis of active pharmaceutical ingredients. However, traditional experimental methods for peony PTM site identification are cumbersome and inefficient; existing computational models are further limited by reliance on manual features, single modification type support, and poor interpretability—failing to meet precise identification needs. To address this, we first built a peony PTM site dataset: retrieving peony proteins from TCMSP, truncating sequences via sliding window to generate 1080 positive samples and 1976 negative sample, with sequence lengths of 3–41 amino acid residues. We then used the Transformer model for prediction: it fuses word vectors and position vectors for initial sequence representation, while its multi-head self-attention captures long-range residue interactions to explore PTM site patterns. 10-fold cross-validation showed optimal performance at a sliding window length of 31; key metrics (accuracy, MCC, F1) significantly outperformed existing models, validating the approach's effectiveness for peony PTM site identification.

Keywords: Transformer, PTM,Paeonia lactiflora, Classification

1 Introduction

Paeonia lactiflora is a core herbaceous species in traditional Chinese medicine. It contains active components such as paeoniflorin and exhibits diverse pharmacological effects, such as enriching the blood to normalize the menstrual cycle, astringing yin to reduce sweating, hepatic pacification for pain relief, coupled with the suppression of liver-yang hyperactivity. The proteins of *Paeonia lactiflora* regulate biological processes through post-translational modification (PTM) mechanisms [1-3]: specifically, phosphorylation mediates signal transduction and cell cycle regulation, glycosylation promotes protein folding and immune recognition, acetylation governs reproductive development and metabolic enzyme activity, and methylation regulates growth stage transition and biosynthesis of medicinal components [4-6]. These PTM mechanisms act synergistically, exerting a profound impact on the growth, disease

resistance, and medicinal quality of *Paeonia lactiflora*, while also laying a molecular foundation for elucidating its pharmacological effects [7]. Consequently, pinpointing the exact locations of PTM sites is essential for elucidating the molecular mechanisms that drive the therapeutic efficacy of *Paeonia lactiflora.*

Traditional experimental techniques have established a comprehensive system for PTM site identification. Anchored by liquid chromatography-tandem mass spectrometry [8,9], this system integrates high-resolution mass spectrometry and fragmentation techniques [10], enabling precise localization of modification sites. Meanwhile, complementary approaches such as antibody affinity enrichment [11], chemical probe labeling[12], and selective enrichment strategies [13] significantly improve the detection efficiency of low-abundance modified peptides. Two-dimensional gel electrophoresis [14] can reveal differences between modified proteins; mutant construction and in vitro enzymatic reactions validate functional modification sites; and hydrogen/deuterium exchange mass spectrometry [15] assists in analyzing the impact of modifications on protein structure. These methods cover the entire workflow from enrichment and separation to verification and structural analysis, ensuring the accuracy and comprehensiveness of PTM site identification. However, such methods are generally time-consuming and labor-intensive.

To circumvent these bottlenecks, in silico approaches have been increasingly adopted as crucial substitutes. Previous studies have made progress in predicting protein modification sites using computational approaches, such as the prediction of protein acetylation and lysine 2-hydroxyisobutyrylation sites [16,17]. Alongside the continuous evolution of machine learning and deep neural networks, more PTM site prediction models have been developed: for example, tools for predicting S-nitrosylation (SNO) sites include GPS-SNO, SNOSite, and iSNOPSeAAC [18-20]. Among these, SNOSID developed by Hao et al. [21] is likely the first computational tool of this type; GPS-SNO is built based on the GPS 3.0 algorithm; and iSNO-PseAAC developed by Xu et al. achieves prediction by representing protein sequences through pseudo-amino acid composition. In terms of lysine crotonylation (Kla) site prediction, FSL-Kla developed by Jiang et al. [22] uses 343 Kla sites from 3 species as training data, encodes sequences by combining amino acid composition features and structural features, and then integrates deep learning models via an ensemble method (note: lysine crotonylation is associated with diseases such as colon cancer and acute kidney injury). DeepKla proposed by Lv et al. [23] adopts a convolutional neural network-bi-directional gated recurrent unit-attention (CNN-BiGRU-attention) mechanism, specifically designed to predict Kla sites in rice. In addition, TransPTM developed by Meng et al. [24] predicts non-histone acetylation sites based on a Transformer network; Pokharel et al. improved the protein language model (PLM [25]) to enhance the performance of succinylation site prediction; PTM-CGGMS developed by Li et al. [26] optimizes prediction results through multi-granularity structure and multi-scale sequence representation; and Liu et al. improved the accuracy of lactylation site prediction by combining structural features predicted by AlphaFold 2 [27] with sequence information [28].

In 2016, Qiu et al. proposed iPTM-mLys [29], the first computational method capable of identifying four types of lysine PTM sites (acetylation, crotonylation,

methylation, and succinylation). It adopts a four-step workflow and uses simple undersampling to address data imbalance. Subsequent methods such as predML-Site, mLysPTMpred, and iMul-kite [30-32] have improved upon iPTM-mLys by optimizing sampling schemes and single-label classification algorithms. CNN+SGT and MLysPRED [33,34] extract additional sequence features and directly employ multi-label classification algorithms (e.g., CNN, MLKNN) for prediction, also using sampling to resolve data imbalance. RMTLysPTM developed by Chen Lei et al. [35] is another multi-label classification model that can identify the aforementioned four types of lysine PTM sites.

While existing approaches have achieved notable success, inherent drawbacks persist: primarily, the widespread reliance on handcrafted feature extraction often fails to model intricate sequential dependencies, thereby yielding a fragmented representation of protein characteristics. second, most models target only a single PTM type or specific lysine PTM sites, lacking a universal model applicable to diverse protein sequences.

To resolve these issues, we introduce a versatile framework for predicting protein modification sites. As depicted in Figure 1, the workflow begins with data collection and the generation of positive and negative samples. Figure 3 details the model's architecture. Initially, sequences are mapped into numerical vectors by fusing word and position embeddings, thereby effectively capturing both amino acid composition and sequential dependencies. A Transformer architecture then processes the generated embeddings to simultaneously extract local and global contextual features. This approach more richly represents the underlying biological context and enables the model to capture the complex interactions between different feature representations. The model achieved the following performance in 10-fold cross-validation: accuracy (Acc) of 92%, sensitivity (Sn) of 88%, specificity (Sp) of 92%, Matthews correlation coefficient (MCC) of 91, and F1 score of 86.45. These results demonstrate that the proposed model is a stable and efficient model for predicting post-translational modification sites in peony.

2 Methods and Materials

2.1 DAtaset Construction

Post-translational modification (PTM) sites in protein sequences are scattered, which poses considerable challenges to directly retrieving PTM site information of *Paeonia lactiflora* sequences from existing databases. To construct a high-quality research dataset (see Figure 1), we used the protein sequences from the *Paeonia lactiflora* sequence database within the Traditional Chinese Medicine Systems Pharmacology Database and Analysis Platform (TCMSP) as the original data source. The specific construction process is as follows:

First, the target protein sequences were batch-downloaded from this database; subsequently, the complete protein sequences were cleaved into short peptide sequences, and the length of the cleaved short peptides was set to $2\theta+1$ (where θ is an integer, used as a sequence index identifier). This length setting method is derived from

Chou's formula [36], which enables the standardized unification of different short peptide sequence lengths. In this study, the number "0" represents the target amino acid in the short peptide, and the relevant short peptide sequences can be described by the following equation:

$$A_{-\theta}A_{-(\theta-1)}\cdots A_{-2}A_{-1}0A_1A_2\cdots A_{\theta-1}A_{\theta} \quad (1)$$

In this equation, "0" corresponds to the amino acid residue at the central position of the peptide sequence; "As" represents the amino acids adjacent to the lysine (K)-site; "A-θ" denotes the θ-th amino acid residue upstream of the central amino acid, and "A+θ" denotes the θ-th amino acid residue downstream of the central amino acid.

In terms of sample classification, we defined short peptide sequences containing PTM sites as positive samples, and short peptide sequences centered on residues adjacent to PTM sites as negative samples. To investigate the impact of variations in window size on subsequent analysis performance, we performed 20 equal-step gradient adjustments of the θ value from 1 to 20 [37]. After the above series of processes, this process yielded 1080 positive and 1976 negative instances, thereby finalizing the dataset.

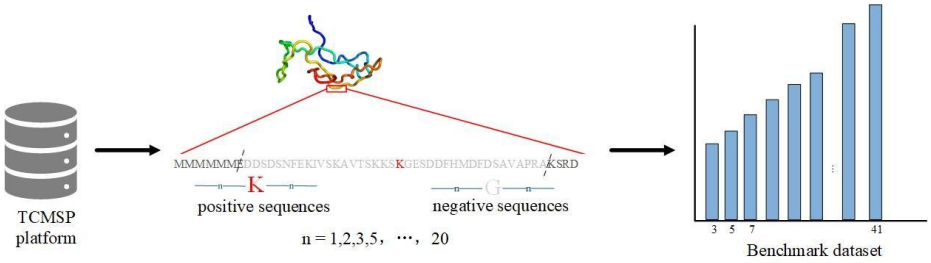


Fig. 1. Workflow of the dataset construction

2.2 Amino Acid Encoding

When processing non-histone sequence data, One-hot Encoding method generates features by mapping every amino acid within the protein chain to a 20-dimensional vector, where this dimension setting corresponds to the 20 natural amino acids present in organisms. The choice of this encoding method is mainly based on two considerations: first, as verified in the feature representation of biological sequences such as proteins [38] and RNA [39], One-hot Encoding is direct and effective, and can intuitively reflect the category differences of amino acids; second, in early studies on protein modification site prediction, One-hot Encoding was often used as the baseline encoding scheme for amino acid chains. For example, MusiteDeep, a tool with high citation frequency in this field [40], uses it as the core encoding method. Therefore, this study also adopts One-hot Encoding to represent peptide chain sequences. The specific encoding rule is: each of the 20 different amino acids is mapped to a 20-dimensional binary vector containing only 0s and 1s, where only the dimension corresponding to the amino acid takes the value of 1, and the remaining dimensions take the value of 0.

Taking alanine (A) and lysine (K) as examples, the former is encoded as (0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0), and the latter is encoded as (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0). For peptide chain sequences with a window size set to 2θ+1 (where θ takes values of 5, 7, 10, 12, 15, 17, 20, 22, 25, 27, 30), they can be converted into feature vectors with a dimension of 20×(2θ+1) after One-hot Encoding processing, and the vector dimension is adjusted accordingly with the change of window size.

As the fundamental form of input embedding [41], Word Embedding encodes each amino acid into a dense, low-dimensional vector, so as to capture the potential correlation information between residues. In the Transformer model architecture, the word embedding layer usually exists in the form of a trainable matrix: for each "token" (i.e., amino acid residue) in the input sequence, the corresponding word vector is retrieved from this matrix through a table-lookup operation, and the parameters of this matrix are continuously optimized according to task requirements during the model training process. However, Word Embedding itself cannot convey the positional information of each residue in the sequence [42]—since the Transformer model only relies on the attention mechanism for feature interaction and lacks the inherent sequence order perception ability of Recurrent Neural Networks (RNNs), using only Word Embedding will make the model unable to distinguish the relative positional differences of residues in the sequence. Therefore, positional embedding must be added on the basis of Word Embedding: by appending a position vector related to the sequence position to each input token, This enables the model to precisely capture the spatial arrangement of individual residues, facilitating a deep comprehension of the peptide's sequence topology.

In recent years, Language Models (LMs) have attracted wide attention because they extract context-aware representations from massive unannotated corpora, thereby overcoming the limitations of rigid, context-blind word vectors. This technology has now been extended to the field of protein research, forming protein Language Models (pLMs) [43]. Benefiting from the massive resources of protein sequence databases, researchers have developed a variety of pLMs. These models can mine deep feature information of protein sequences from the databases [44] and and apply these learned representations to subsequent applications, including the forecasting of protein attributes and the localization of modification sites, and have shown better ability in understanding sequence relationships compared with traditional encoding methods. In this study, the encoder output of the pre-trained model ProtT5-XL-U50 was selected as the source of embedding features [45]. Built upon the Transformer framework, this tool functions as a dedicated protein language model, containing 3 billion parameters, and its training process is divided into two stages: the initial training stage uses the "Big Fantastic Database (BFD)" as the training set, which covers 65 million protein families, and these families are classified and annotated through Multiple Sequence Alignments (MSA) [46] and Hidden Markov Models [47]; The subsequent fine-tuning phase leverages the UniRef50 dataset, which aggregates clustered protein sequences extracted from UniProtKB alongside specific UniParc entries, and can further optimize the accuracy of the model in capturing protein sequence features. After inputting the peptide chain sequences into ProtT5-XL-U50, the encoder output of the model is the

embedding feature of the peptide chain, where each amino acid residue corresponds to a 1024-dimensional embedding vector. This embedding feature is not only position-dependent and can reflect the positional differences of residues in the sequence, but also can effectively capture the contextual correlation features of each residue, providing more abundant sequence information support for subsequent modification site prediction tasks.

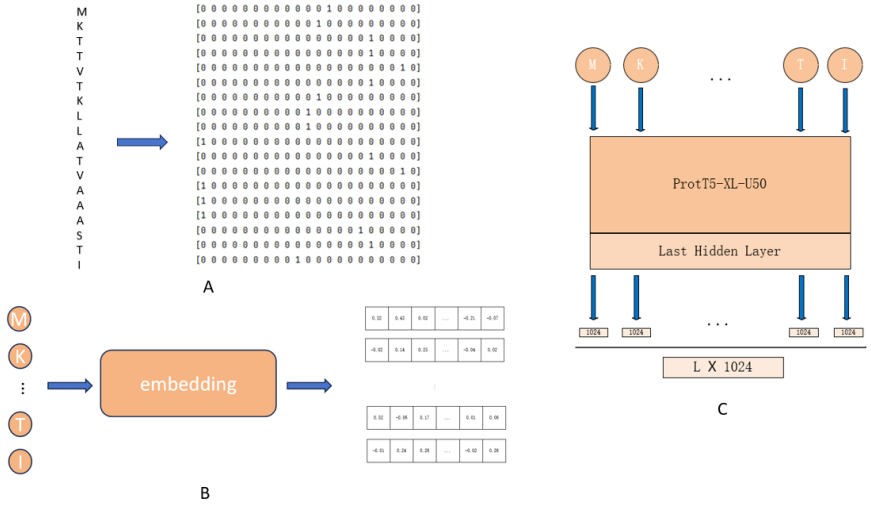


Fig. 2. Amino acid encoding. (A)One-hot encoding,(B) Word embedding ,(C) Embeddings of ProtT5.

2.3 Attention

The self-attention mechanism captures the internal dependencies of a sequence by calculating the correlations between elements within the sequence[49]. Its core idea is to map the input sequence into three matrices: Query, Key, and Value. The weights of the value are determined by the similarity between the Query and the Key. The specific calculation process is as follows:

$$Attention(Q,K,V) = soft\,max(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

Here, d is the dimension of the Key, $\sqrt{d_k}$ which is used to scale the dot-product value to avoid gradient vanishing. This calculation method enables the model to adaptively focus on the important parts of the sequence.

The Transformer introduces the multi-head attention mechanism to further enhance the model's expressive ability. Multi-head attention projects the input into multiple sub-spaces, calculates the attention independently in each sub-space, and finally concatenates the outputs of each head:

$$MultiHead(Q,K,V) = Concat(head_1,\ldots,head_h)W^O \quad (3)$$

$$head_i = Attention(QW_i^Q, QW_i^K, QW_i^V) \quad (4)$$

Where h is the number of attention heads, W_i^Q, W_i^K, W_i^V and W^O are learnable parameter matrices.

Beyond the attention mechanism, every layer within the Transformer block incorporates a position-wise feed-forward module that processes each sequence element independently yet uniformly. Structurally, this network employs a pair of linear projections separated by a ReLU non-linear activation function.

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (5)$$

2.4 Framework of the Proposed Model

To map the protein sequences into a continuous vector space, our framework integrates both positional and token embeddings. According to a pre-established reference dictionary, every discrete residue within the chain is assigned a distinct integer identifier. As an illustration, the residues alanine (A), cysteine (C), and aspartic acid (D) are designated by the scalar values 3, 4, and 5, respectively. Formally, for any given amino acid x_i located in the sequence, its structural representation begins by extracting a corresponding integer $E(x_i)$ from this global vocabulary:

$$E(x_i) = Index(x_i) \quad (6)$$

To capture the sequential order within the protein chain, learnable positional embeddings are introduced. Specifically, the spatial location of the i-th amino acid is encoded by a trainable matrix P_i

$$P_i = \text{Position Embedding}(i) \quad (7)$$

The final vector representation H_i is obtained by adding its word embedding and positional embedding:

$$H_i = E_i + P_i \quad (8)$$

The resulting embedding matrix, $H \in \mathbb{R}^{L\times d}$, is subsequently fed into a deep network optimized to extract multi-scale sequence dependencies.

Initially, the embedding matrix H is fed into a stacked BiLSTM network to extract hierarchical, bidirectional dependencies. By integrating context from surrounding amino acids, this module outputs an enriched sequence representation H_{LSTM}:

$$H_{LSTM} = StackedBiLSTM(H) \quad (9)$$

Here, $H_{LSTM} \in \mathbb{R}^{L\times d}$ denotes the bidirectionally enriched representation. To capture global dependencies, this matrix is subsequently routed through N Transformer blocks equipped with multi-head self-attention mechanisms, computed as follows:

$$H_{tran} = transformerblock(H_{LSTM}) \tag{10}$$

The resulting refined embedding, $H_{tran} \in \mathbb{R}^{L\times d}$, is finally processed by an adaptive max-pooling layer. This step effectively collapses the sequence dimension, compressing the variable-length data into a unified, fixed-size feature vector.

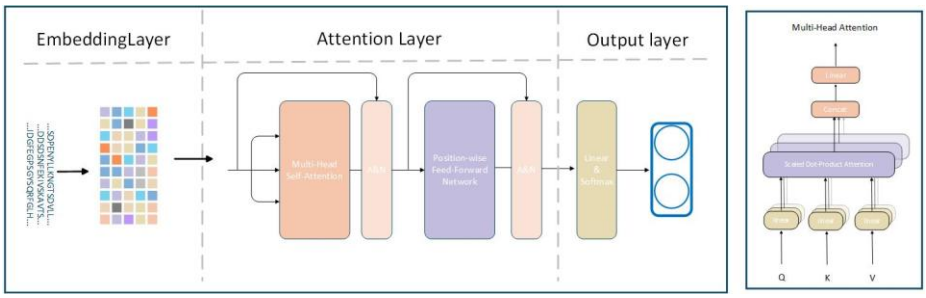


Fig. 3. The overall architecture of the model

2.5 Model Evaluation

In this study, we regarded the protein sequences containing post-translational modification sites as positive samples, and those without such sites as negative samples. During the prediction process, we refer to the correct identification of positive samples as true positives (TP), and the correct identification of negative samples as true negatives (TN). We label the misclassification of a negative sample as a positive one as a false positive (FP), and the misclassification of a positive sample as a negative one as a false negative (FN). We averaged all performance indicators and reported the results. We used four indicators, namely accuracy, sensitivity, precision, and the Matthew Correlation Coefficient (MCC), to evaluate the performance, and we set the decision probability threshold at 0.5.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{11}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \tag{12}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \tag{13}$$

$$\text{F1score} = 2 \times \frac{\text{TP}}{\text{2TP+FP+FN}} \tag{14}$$

$$\text{MCC} = \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\sqrt{(\text{TP+FP})(\text{TP+FN})(\text{TN+FP})(\text{TN+FN})}} \tag{15}$$

3 Results and Discussion

3.1 Window Size Selection

During modeling, many studies employ a fixed local sliding window. However, it's important to emphasize that different sliding windows can produce different prediction results. Optimizing the window size can significantly assist feature selection and improve prediction performance. To determine the optimal window size, we experimented with a range of window sizes: 3, 5, 7, ..., 37, 39, and 41. As shown in Figure 4, the average MCC across different window sizes indicates that model performance peaks at a window size of 31, as the length of the protein sequence changes. Generally, we expect longer sequences to contain more semantic and contextual information. We hypothesize that this may be because when the sequence is too short, the model is unable to learn its intrinsic information, while when the sequence is too long, this information is diluted. Therefore, in subsequent analyses, we selected 31 as the optimized window size for acetylated residues.

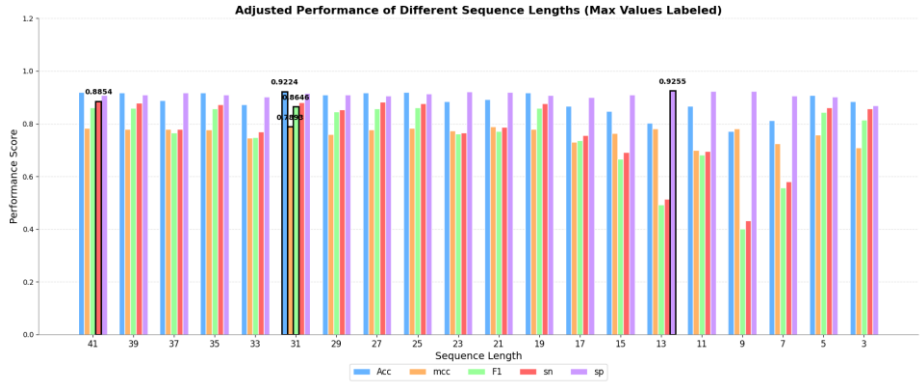


Fig. 4. Indicators of stackedbilstm_transformer on the training dataset for sliding window size range

3.2 Performance Evaluation of Model

Table 1. Performance comparison of transformer with baseline

	Acc(%)	Sn(%)	sp(%)	Mcc(%)	F1(%)
tranformer	92.24%	88.01%	91.51%	78.93%	86.46%
lstm	89.62%	88.44%	90.31%	77.89%	87.34%
CNN	91.26%	81.68%	96.51%	78.59%	82.54%
Svm	49.67%	40.52%	55.06%	-4.29%	50.31%
rf	51.47%	3.96%	79.48%	-2.27%	44.47%

To rigorously evaluate our framework, we conducted a ten-fold cross-validation against four baseline classifiers: two traditional models (RF and SVM utilizing PseAAC features) and two deep learning architectures (CNN and LSTM). As detailed in Table

1, our proposed approach significantly outperformed these alternatives, particularly dominating in ACC, MCC, and F1 metrics. Specifically, our model achieved scores of 92%, 88%, 92%, 91%, and 86% across ACC, SN, SP, MCC, and F1, respectively. These robust metrics validate the framework as a highly reliable tool for identifying post-translational modification sites in peony.

To further demonstrate the advantages of our model, we tested it on a validation set along with several other encoding methods that have performed well in PTM site prediction. The first model was a one-hot plus Transformer model, and the second was a Prott5 plus Transformer model. We then used the hidden layer vectors from the Transformer as features for machine learning classification and observed the results.

Table 2. Performance comparison of tranformer with other encoding methods

	Acc	sn	sp	MCC	F1
transformer	85.57%	83.37%	88.83%	73.51%	83.38%
one-hot+transformer	64.78%	0.62%	99.66%	2.05%	1.22%
prott5+tranformer	61.42%	28.21%	73.51%	23.45%	45.65%
RF	84.35%	82.88%	89.34%	72.87%	82.89%
SVM	83.13%	81.98%	89.51%	72.32%	82.49%
Xgboost	82.91%	80.07%	87.97%	72.11%	82.60%
Ensemble+Learning	81.13%	80.28%	89.34%	72.36%	82.54%

In the PTM site prediction task, Transformer and Random Forest (RF) performed best, with accuracy (Acc) exceeding 87% and balanced sensitivity (sn ≈ 85%) and specificity (sp ≈ 89%). The Prott5 series of models performed poorly due to issues such as feature encoding and data imbalance. Traditional machine learning models (RF, SVM, and Xgboost) outperformed most deep learning models due to their robustness.

4 COnclusion

Post-translational modifications (PTMs) are critical regulators of biological processes in *Paeonia lactiflora*—a core herbaceous species in traditional Chinese medicine—governing its growth, disease resistance, and biosynthesis of medicinal components such as paeoniflorin. However, traditional experimental methods for PTM site identification (e.g., LC-MS/MS-based workflows) are time-consuming and labor-intensive, while existing computational models suffer from three key limitations: over-reliance on manually designed features, support for only single PTM types, and poor interpretability. To address these challenges, this study focused on constructing a species-specific PTM dataset and developing an optimized prediction model, with key findings and contributions summarized as follows.

First, this study established a high-quality, standardized *Paeonia lactiflora* PTM site dataset, filling the gap of scarce species-specific PTM research resources. Using protein sequences retrieved from the TCMSP as the original data source, we employed a sliding window method (derived from Chou's formula) to cleave full-length proteins into short peptide fragments. The window size was adjusted in 20 equal steps (corresponding to sequence lengths of 3–41 amino acid residues), yielding 1080 positive samples

(peptides containing PTM sites) and 1976 negative samples (peptides centered on non-PTM adjacent residues). This dataset not only provides a reliable benchmark for subsequent *Paeonia lactiflora* PTM site prediction studies but also offers a replicable framework for dataset construction in other medicinal plants.

Second, we proposed a Transformer-based PTM site prediction model that addresses the inherent limitations of existing computational methods. The model integrates two core embedding strategies: word embeddings (to capture the semantic and chemical properties of amino acids) and positional embeddings (to encode sequence order information, compensating for the Transformer's lack of intrinsic sequence perception). Its multi-head self-attention mechanism enables efficient modeling of long-range interactions between amino acid residues, while the fusion of stacked BiLSTM and Transformer encoder modules enhances the capture of both local and global sequence dependencies. An adaptive max-pooling layer further converts variable-length sequence embeddings into fixed-size feature vectors, ensuring the model's stability in practical applications.

Third, systematic experiments validated the superiority and robustness of the proposed model. Through 10-fold cross-validation across window sizes (3–41 amino acid residues), we identified **31 amino acids** as the optimal window size: at this length, the model achieved peak performance with an accuracy (Acc) of 92.24%, sensitivity (Sn) of 88.01%, specificity (Sp) of 91.51%, Matthews correlation coefficient (MCC) of 78.93%, and F1 score of 86.46%. Comparative analysis with baseline models confirmed its advantages: it outperformed traditional machine learning models (SVM, RF) by a large margin (e.g., Acc of 92.24% vs. 49.67% for SVM and 51.47% for RF) and surpassed deep learning counterparts (LSTM, CNN) in key metrics such as MCC and Acc. Additionally, when compared with alternative encoding strategies (One-hot+Transformer, ProtT5+Transformer), the proposed model maintained superior performance, verifying the rationality of its feature representation and architectural design.

This study makes two distinct contributions to the field of PTM research in medicinal plants. On the data front, the constructed *Paeonia lactiflora* PTM dataset provides a standardized foundation for exploring the molecular mechanisms by which PTMs regulate the plant's medicinal properties. On the methodological front, the Transformer-based model overcomes the limitations of existing computational tools, offering an efficient and stable solution for PTM site prediction in *Paeonia lactiflora*.

Despite these achievements, this study has room for improvement. Future work will focus on three directions: (1) expanding the dataset to include more PTM types (e.g., phosphorylation, glycosylation) and larger sample sizes to enhance the model's generalization ability; (2) integrating structural features (e.g., protein 3D structures predicted by AlphaFold 2) with sequence information to further improve predictive accuracy; and (3) validating the model's performance on PTM sites of other medicinal plants to confirm its cross-species applicability, thereby promoting broader progress in PTM research for traditional Chinese medicine.

References

1. Seo, J.-W., Lee, K.-J.: Post-translational modifications and their biological functions: proteomic analysis and systematic approaches. BMB Rep 37(1), 35–44 (2004)
2. Krassowski, M., Paczkowska, M., Cullion, K., et al.: Activedriverdb:human disease mutations and genome variation in post-translational modification sites of proteins. Nucleic Acids Res 46(D1), D901–10 (2018)
3. Keenan, E.K., Zachman, D.K., Hirschey, M.D.: Discovering the landscape of protein modifications. Mol Cell 81(9), 1868–78 (2021)
4. Walsh, C.T., Garneau-Tsodikova, S., Gatto Jr, G.J.: Protein posttrans-lational modifications: the chemistry of proteome diversifications. Angew Chem Int Ed 44(45), 7342–72 (2005)
5. Jensen, O.N.: Interpreting the protein language using proteomics. Nat Rev Mol Cell Biol 7(6), 391–403 (2006)
6. Yang, X.-J., Seto, E.: Lysine acetylation: codified crosstalk with other posttranslational modifications. Mol Cell 31(4), 449–61 (2008)
7. Millán-Zambrano, G., Burton, A., Bannister, A.J., et al.: Histone post-translational modifications—cause and consequence of genome function. Nature Reviews Genetics 23(9), 563-580 (2022)
8. Zhao, W., et al.: Systematic identification of the lysine lactylation in the protozoan parasite toxoplasma gondii. Parasit. Vectors 15(1), 180 (2022)
9. Wang, J.H., et al.: Beyond metabolic waste: lysine lactylation and its potential roles in cancer progression and cell fate determination. Cell. Oncol. (Dordr.) 46(3), 465–480 (2023)
10. Guthals, A., Bandeira, N.: Peptide identification by tandem mass spectrometry with alternate fragmentation modes. Molecular & Cellular Proteomics 11(9), 550-557 (2012)
11. Roque, A.C.A., Silva, C.S.O., Taipa, M.Â.: Affinity-based methodologies and ligands for antibody purification: advances and perspectives. Journal of Chromatography A 1160(1-2), 44-55 (2007)
12. Horisawa, K.: Specific and quantitative labeling of biomolecules using click chemistry. Frontiers in physiology 5, 457 (2014)
13. Li, X.S., Yuan, B.F., Feng, Y.Q.: Recent advances in phosphopeptide enrichment: strategies and techniques. TrAC Trends in Analytical Chemistry 78, 70-83 (2016)
14. Meleady, P.: Two-dimensional gel electrophoresis and 2D-DIGE. Difference gel electrophoresis: methods and protocols, 3-14 (2018)
15. Engen, J.R., Botzanowski, T., Peterle, D., et al.: Developments in hydrogen/deuterium exchange mass spectrometry. Analytical chemistry 93(1), 567-582 (2020)
16. Bao, W., Yang, B.: Protein acetylation sites with complex-valued polynomial model. Frontiers of Computer Science 18(3), 183904-null (2024)
17. Bao, W., Yang, B., Chen, B.: 2-hydr_ensemble: lysine 2-hydroxyisobutyrylation identification with ensemble method. Chemom. Intel. Lab. Syst. 4, 104351 (2021)
18. Xue, Y., Liu, Z., Gao, X., et al.: GPS-SNO: computational prediction of protein S-nitrosylation sites with a modified GPS algorithm. PloS one 5(6), e11290 (2010)
19. Lee, T.Y., Chen, Y.J., Lu, T.C., et al.: SNOSite: exploiting maximal dependence decomposition to identify cysteine S-nitrosylation with substrate site specificity. PloS one 6(7), e21849 (2011)
20. Xu, Y., Ding, J., Wu, L.Y., et al.: iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PloS one 8(2), e55844 (2013)

21. Hao, G., Derakhshan, B., Shi, L., et al.: SNOSID, a proteomic method for identification of cysteine S-nitrosylation sites in complex protein mixtures. Proceedings of the National Academy of Sciences 103(4), 1012-1017 (2006)
22. Jiang, P., Ning, W., Shi, Y., et al.: FSL-Kla: A few-shot learning-based multi-feature hybrid system for lactylation site prediction. Computational and structural biotechnology journal 19, 4497-4509 (2021)
23. Lv, H., Dao, F.Y., Lin, H.: DeepKla: an attention mechanism-based deep neural network for protein lysine lactylation site prediction. Imeta 1(1), e11 (2022)
24. Meng, L., Chen, X., Cheng, K., et al.: TransPTM: a transformer-based model for non-histone acetylation site prediction. Briefings in Bioinformatics 25(3), bbae219 (2024)
25. Pokharel, S., Pratyush, P., Heinzinger, M., et al.: Improving protein succinylation sites prediction using embeddings from protein language model. Scientific reports 12(1), 16933 (2022)
26. Li, Z., Li, M., Zhu, L., Zhang, W.: Improving PTM site prediction by coupling of multi-granularity structure and multi-scale sequence representation. Proceedings of the AAAI Conference on Artificial Intelligence 38(1), 188–196 (2024)
27. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D.: Highly accurate protein structure prediction with AlphaFold. Nature 596 (7873), 583–589 (2021)
28. Yang, Y.H., Yang, J.T., Liu, J.F.: Lactylation prediction models based on protein sequence and structural feature fusion. Brief. Bioinform. 25(2) (2024).
29. Qiu, W.-R., Sun, B.Q., Xiao, X., et al.: iPTM-mLys: identifying multiple lysine PTM sites and their different types. Bioinformatics 32(20), 3116–23 (2016)
30. Ahmed, S., Rahman, A., Hasan, M.A.M., et al.: predML-site: predicting multiple lysine PTM sites with optimal feature representation and data imbalance minimization. IEEE/ACM Trans Comput BiolBioinform 19(6), 3624–34 (2022)
31. Hasan, M.A.M., Ahmad, S.: mLysPTMpred: multiple lysine PTM siteprediction using combination ofSVMwith resolving data imbal-ance issue. Natural Science 10(9), 370–84 (2018)
32. Ahmed, S., Rahman, A., Hasan, M.A.M., et al.: Computational identification of multiple lysine PTM sites by analyzing theinstance hardness and feature importance. Sci Rep 11(1), 18882 (2021)
33. Sua, J.N., Lim, S.Y., Yulius, M.H., et al.: Incorporating convolutional neural networks and sequence graph transform for identifying multilabel protein lysine PTM sites. Chemom Intel Lab Syst 206, 104171 (2020)
34. Zuo, Y., Hong, Y., Zeng, X., et al.: MLysPRED: graph-based multi-view clustering and multi-dimensional normal distribution resampling techniques to predict multiple lysine sites. Brief Bioinform 23(5), bbac277 (2022)
35. Chen, L., Chen, Y.: RMTLysPTM: Recognizing multiple types of lysine PTM sites by deep analysis on sequences. Briefings in Bioinformatics 25(1), bbad450 (2024)
36. Chou, K.C.: Prediction of signal peptides using scaled window. peptides 22(12), 1973-1979 (2001)
37. Li, W., Godzik, A.: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13), 1658-1659 (2006)

38. ElAbd, H., Bromberg, Y., Hoarfrost, A., et al.: Amino acid encoding for deep learning applications. BMC Bioinformatics 21, 1–14 (2020)
39. El Allali, A., Elhamraoui, Z., Daoud, R.: Machine learning applications in rna modification sites prediction. Comput Struct Biotechnol J 19, 5510–24 (2021)
40. Wang, D., Liu, D., Yuchi, J., et al.: MusiteDeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization. Nucleic Acids Research 48(W1), W140-W146 (2020)
41. Li, Y., Yang, T.: Word embedding for understanding natural language: a survey. Guide to big data applications, 83-104 (2018)
42. Tng, S.S., Le, N.Q.K., Yeh, H.Y., et al.: Improved prediction model of protein lysine crotonylation sites using bidirectional recurrent neural networks. Journal of proteome research 21(1), 265-273 (2021)
43. Heinzinger, M., Littmann, M., Sillitoe, I., et al.: Contrastive learning on protein embeddings enlightens midnight zone. NAR GenomicsBioinf 4(2), lqac043 (2022)
44. Vu, M.H., Akbar, R., Robert, P.A., et al.: Linguistically inspired roadmap for building biologically reliable protein languagemodels. Nature. Mach Intell 5(5), 485–96 (2023)
45. Pratyush, P., Bahmani, S., Pokharel, S., et al.: LMCrot: an enhanced protein crotonylation site predictor by leveraging an interpretable window-level embedding from a transformer-based protein language model. Bioinformatics 40(5), btae290 (2024)
46. Edgar, R.C., Batzoglou, S.: Multiple sequence alignment. Current opinion in structural biology 16(3), 368-373 (2006)
47. Eddy, S.R.: Hidden markov models. Current opinion in structural biology 6(3), 361-365 (1996)
48. Zhou, J., Lu, Y., Dai, H.N., et al.: Sentiment analysis of Chinese microblog based on stacked bidirectional LSTM. IEEE Access 7, 38856-38866 (2019)
49. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. arXiv preprint arXiv:1803.02155 (2018)
50. Huang, L., Lin, J., Liu, R., et al.: CoaDTI: multi-modal co-attention based framework for drug–target interaction annotation. Briefings in bioinformatics 23(6), bbac446 (2022)
51. Vacic, V., Iakoucheva, L.M., Radivojac, P.: Two sample logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics 22(12), 1536–7 (2006)
52. Huang, L., Lin, J., Liu, R., et al.: Coadti: multi-modal co-attention based framework for drug target interaction annotation. BriefBioinform 23(6), bbac446 (2022)
53. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. JMachLearn Res 9(11) (2008)

2025 International Conference on Applied Intelligence
November 6-9, Nanning, Guangxi, China
http://www.icai.org.cn/2025/index.php

Prediction of Post-Translational Modification Sites of Paeonia lactiflora Proteins Based on Attention Mechanism

Yingyue Tang[1] and Wenzheng Bao[2,3]

[1] School of Information Engineering, Yancheng Institute of Technology, Yancheng, China 224051
[2] Institute for Regenerative Medicine, Medical Innovation Center and State Key Laboratory of Cardiology, Shanghai East Hospital, School of Medicine, Tongji University, Shanghai 200123, P. R. China

[3] Xuzhou University of Technology, Xuzhou, China, 221018

Abstract. The precise prediction of post-translational modification sites of proteins was an important aspect of elucidating the mechanisms of traditional Chinese medicine. Existing computational models struggled to meet the needs of herbal pharmacological research due to their neglect of the specificity of herbal materials and the complexity of modification sites. This study focused on the traditional Chinese medicinal material Paeonia lactiflora and constructed a deep learning model based on convolutional neural networks and attention mechanisms by integrating multi-dimensional features of amino acid sequences. A dataset for Paeonia lactiflora was built using the TCMSP database, which included 1080 positive samples and 1976 negative samples, and the input space was optimized through feature normalization and dimensionality reduction. Experimental results indicated that the model effectively captured the modification patterns of proteins from different herbal materials, and SHAP feature selection significantly improved the prediction accuracy of post-translational modification sites. Compared to traditional single algorithm models, the proposed integrated architecture demonstrated significant advantages in balancing sequence conservation and functional specificity, providing a new computational tool for elucidating the mechanisms of action of active components in traditional Chinese medicinal materials and guiding rational clinical medication.

Keywords: Post-translational Modification Sites, Attention Mechanism, Transfer Learning, Paeonia lactiflora

1 Introduction

Paeonia, as a traditional medicinal herb widely used in Chinese medicine, received significant attention due to its various pharmacological activities [1][2]. Recent studies

indicated that Paeonia exhibited remarkable efficacy in treating oral diseases, particularly positively influencing conditions such as xerostomia [3][4]. The active components contained not only possessed anti-inflammatory and analgesic properties but also demonstrated hepatoprotective and antioxidant functions, which made Paeonia play an important role in alleviating oral discomfort and improving quality of life [5].

Especially among patients with Sjögren's syndrome, Paeonia effectively promoted saliva secretion, thereby alleviating dry mouth symptoms and providing tangible assistance to patients [6-8]. This not only improved the patients' oral health but also further enhanced their overall quality of life. Additionally, research found that formulations related to white Paeonia could significantly enhance therapeutic effects when used in conjunction with chemical drugs while reducing drug toxicity [9]. This finding not only provided new ideas for the rational combination of clinical medications but also opened new avenues for the treatment of autoimmune diseases [10].

Therefore, in-depth exploration of the pharmacological mechanisms of Paeonia and its potential in clinical applications not only aided in better understanding its status in traditional Chinese medicine but also provided important references for the advancement of modern medicine [11][12]. As research on Paeonia deepened, it was anticipated that more innovative therapies based on its active components would emerge, bringing new hope for the treatment of various diseases [13][14].

Post-Translational Modifications (PTMs) of proteins referred to a series of chemical modification processes that occurred after protein synthesis [15][16]. These modifications involved various biochemical reactions, forming a complex and precise network, which not only influenced protein function and stability but also played key roles in biological processes such as cell signaling and metabolic regulation [17-18]. Thus, in-depth understanding of PTMs was crucial for revealing protein diversity and its biological functions. This indicated the unique complexity of PTMs in maintaining biological activity and mediating various biological interactions [19]. The biosynthetic mechanisms of tanshinones and phenolic acids in Salvia miltiorrhiza suggested that these modifications not only affected protein structure and function but also played significant regulatory roles in the mechanisms of action between white Paeonia-derived compounds and various diseases [20]. For instance, certain PTMs might enhance their antitumor or anti-inflammatory efficacy, providing new insights for exploring the modern applications of traditional Chinese medicine [21].

Although traditional experimental methods for studying PTMs, such as mass spectrometry, Western blotting, immunoprecipitation, and flow cytometry, had developed relatively maturely, these methods often proved time-consuming, costly in terms of equipment, and complex in process [22-23]. This made traditional experimental methods increasingly inadequate in the face of the growing amount of plant protein data. Therefore, to more efficiently process and analyze this data, AI-based predictive methods emerged [24].

In recent years, deep learning algorithms made significant contributions to bioinformatics. In the field of drug discovery, Hakime Öztürk et al. proposed a deep learning-based model, DeepDTA, which effectively predicted drug-target binding affinity using target and drug sequence information, overcoming limitations in previous methods [25]. Rohan Gupta et al. reviewed the extensive applications of AI and

machine learning in drug design, highlighting the modernizing impact of new technologies on the drug discovery process [26]. Ehsaneddin Asgari et al. introduced the BioVec biological sequence representation method, showcasing the potential of deep learning in proteomics and genomics research [27]. The P2Rank tool developed by Krivák and Hockx provided an efficient solution for rapid and accurate prediction of ligand binding sites using machine learning [28]. In the prediction of phosphorylation sites, the MusiteDeep framework proposed by Duolin Wang et al. significantly improved prediction accuracy through convolutional neural networks and attention mechanisms [29]. The DEEPre model by Li, Yu et al. enhanced enzyme function prediction performance through automated feature selection and classification training methods [30]. Schubach et al. developed the CADD v1.7 version, which integrated new annotation features to further improve the accuracy of genome-wide variant predictions, providing important tools for related research [31]. These studies collectively propelled advancements in the fields of drug discovery and bioinformatics.

Currently, attention mechanisms are widely used in deep learning due to their ability to reduce information redundancy, handle long sequence data, and possess high interpretability. In the field of drug-protein interaction prediction, Dayu Tan et al. constructed a dual-channel neural network model, DCA-DPI, which utilized drug molecular graphs and protein sequences as inputs, learning the feature representations of drugs and proteins through residual graph neural networks and residual convolutional networks, combined with efficient path attention mechanisms, thereby improving the accuracy of DPI predictions [32]. Ying Xu and Jinyong Cheng proposed a protein secondary structure prediction model based on multi-scale convolutional attention neural networks, aiming to effectively extract both local and long-range information from amino acid sequences, enhancing the reconstruction of feature maps through a multi-channel multi-scale parallel architecture, ultimately improving prediction outcomes [33]. The deep learning tool DLBWE-Cys developed by Zhengtao Luo et al. aimed to identify cysteine S-carboxyethylation sites, combining CNN, BiLSTM, Bahdanau attention mechanism, and fully connected neural networks, with experimental results showing its superior performance in cross-validation and independent testing compared to other models, validating the effectiveness of its encoding method [34]. Yuhao He et al. established a novel sound-based COVID-19 diagnostic framework, TFA-CLSTMNN, which significantly improved detection accuracy by extracting time-frequency domain features from cough recordings and integrating attention convolutional long short-term memory neural networks, achieving an accuracy of over 0.95 on public real-world datasets [35].

Although these emerging methods improved prediction efficiency to some extent, they still struggled to comprehensively meet all demands for PTM site prediction, necessitating the urgent development of new models and algorithms to enhance prediction accuracy and broad applicability [36].

To address the above issues, our experimental contributions included the following points:

(1) A deep learning model based on convolutional neural networks and attention mechanisms was constructed, where the convolutional neural network was responsible for extracting local information, and global attention could establish long-distance

interactions for each feature information; randomly generated orthogonal filters could adaptively adjust channel importance, and we preliminarily explored the impact of different attention mechanisms on encoding results.

(2) The prediction accuracy of PTM sites was significantly improved through SHAP feature selection. Compared to traditional single-algorithm models, our proposed ensemble architecture exhibited significant advantages in balancing sequence conservation and functional specificity.

(3) A dataset integrating various encoding features was created to verify the impact of encoding before and after feature selection on the model.

2 Feature input

2.1 Dataset Construction

The data used in this study was sourced from the Traditional Chinese Medicine Systems Pharmacology Database and Analysis Platform (TCMSP), which provided extensive information regarding the relationships between various herbal targets and diseases. We segmented the complete protein sequence of Paeonia lactiflora (Ps) according to Zhou's formula to obtain protein sequence fragments (Pss), which can be described by formula (1) [37]。

$$P_{ss} = [A_{-20}, A_{-19}, \ldots A_{-2}, A_{-1}, A_0, A_1, A_2, \ldots A_{19}, A_{20}], \quad P_{ss} \subseteq P_s \tag{1}$$

The amino acid at the center position A0 of Pss determined the positive or negative nature of the fragment. We used the amino acids at post-translational modification sites as A0, with the resulting protein sequence fragments cut from this position serving as positive examples. Specifically, we took 20 consecutive amino acids both forward and backward from A0, resulting in Pss having a length of 41 amino acids. Next, we identified the amino acids adjacent to the center position A0, namely A-1 and A1. If A-1and A1 were non-post-translational modification site amino acids, we redefined A-1 and A1 as the new center position A0 for the negative example protein fragment segmentation. During this process, if the center amino acid A0 of Pss was too close to the beginning or end of the protein sequence, it would lead to the problem of the protein sequence fragment being too short. To address this issue, we supplemented the sequence by adding the most homologous amino acid residues.

In the end, we obtained a dataset for Paeonia lactiflora consisting of 1,080 positive samples and 1,976 negative samples.

2.2 Amino acid encoding

2.2.1 Block Substitution Matrix

The Block Substitution Matrix (BLOSUM62) utilized a predefined amino acid substitution probability matrix to map individual amino acid residues to numerical

vectors[38]. Specifically, each residue was represented by an m×n element matrix, where n denoted the length of the peptide chain and m represented the number of amino acid types. Each residue was represented by a row in the matrix, allowing it to reflect the evolutionary information of proteins; however, this method also fixed it to encode equally long peptide segments.

2.2.2 Composition of k-spaced amino acid group pairs

The k-spaced amino acid group pair (CKSAAGP) was an improved version of the CKSAAP descriptor[39]. This metric quantified features by statistically measuring the frequency of occurrence of group pairs separated by any k amino acid residues, aiding in the capture of distant correlation information. The cksaagp feature vector was defined by equation (2).

$$\left(\frac{N_{g1g1}}{N_{total}}, \frac{N_{g1g2}}{N_{total}}, \frac{N_{g1g3}}{N_{total}}, \ldots, \frac{N_{g5g5}}{N_{total}}\right)_{25} \tag{2}$$

The value of each descriptor represented the composition of the corresponding residue pairs in the protein or peptide sequence, where N_{total} was the total number of spaced residue pairs in the protein. When k=0, there were 25 types of 0-spaced group pairs (i.e., g1g1, g1g2, g1g3, ..., g5g5). When the length of the protein was P and k=0, 1, 2, 3, 4, and 5, the corresponding N_{total} values were P-1, P-2, P-3, P-4, P-5, and P-6, respectively.

2.2.3 Adaptive skip dinucleotide composition

Adaptive skip dinucleotide composition (ASDC) fully considered the correlation information between adjacent residues, including those between intervening residues, balancing the associations of both adjacent and non-adjacent residues[40]. Its feature vector was represented by equation (3).

$$\begin{cases} ASDC = (f_{v1}, f_{v1}, \ldots, f_{v400}), \\ f_{vi} = \dfrac{\sum_{g=1}^{L-1} O_i^g}{\sum_{i=1}^{400}\sum_{g=1}^{L-1} O_i^g} \end{cases} \tag{3}$$

where fvi denoted the frequency of occurrence of all possible dipeptides (with intervening nucleotide numbers ≤ L-1).

2.2.4 Enhanced amino acid composition

Enhanced amino acid composition (EAAC) employed a fixed-length sliding window that moved along the peptide chain, dynamically calculating the proportion of each amino acid within the window[41]. Specifically, EAAC slid continuously from the N-terminus to the C-terminus of each peptide chain to compute the amino acid composition encoding (AAC), where the ACC formula was as shown in equation (4), and the EAAC formula was as indicated in equation (5).

$$f(t) = \frac{N(t)}{N}, t \in \{A, C, D, \ldots, Y\} \tag{4}$$

AAC encoding calculated the occurrence frequency of the 20 natural amino acids ("ACDEFGHIKLMNPQRSTVWY") in the protein or peptide sequence. In this context, N(t) represented the quantity of amino acid type t, while N denoted the length of the protein or peptide sequence.

$$f(t,win) = \frac{N(t,win)}{N(win)}, t \in \{A, C, D, \ldots Y\}, \quad win \in \{window1, window2, window3, \ldots window17\} \tag{5}$$

In this context, $N(t,win)$ represented the quantity of amino acid type t within the sliding window win, while $N(win)$ denoted the size of the sliding window.

2.2.5 DPC (Di-Peptide Composition)

The Di-peptide Composition, as shown in equation (6), comprised 400 descriptors, where Nrs represented the quantity of dipeptides corresponding to amino acid types r and s. This method characterized the overall dipeptide abundance in protein sequences; however, its spatial order was disrupted and was often applied in scenarios requiring rapid global feature extraction.

$$D(r,s) = \frac{N_{rs}}{N-1}, \quad r,s \in \{A, C, D, \ldots Y\}, \tag{6}$$

3 Model construction

This study integrated various amino acid sequence-based feature extraction methods to create a classification model using Convolutional Neural Networks (CNN) and attention mechanisms, classifying the fused features. The flowchart is shown in Figure 1.

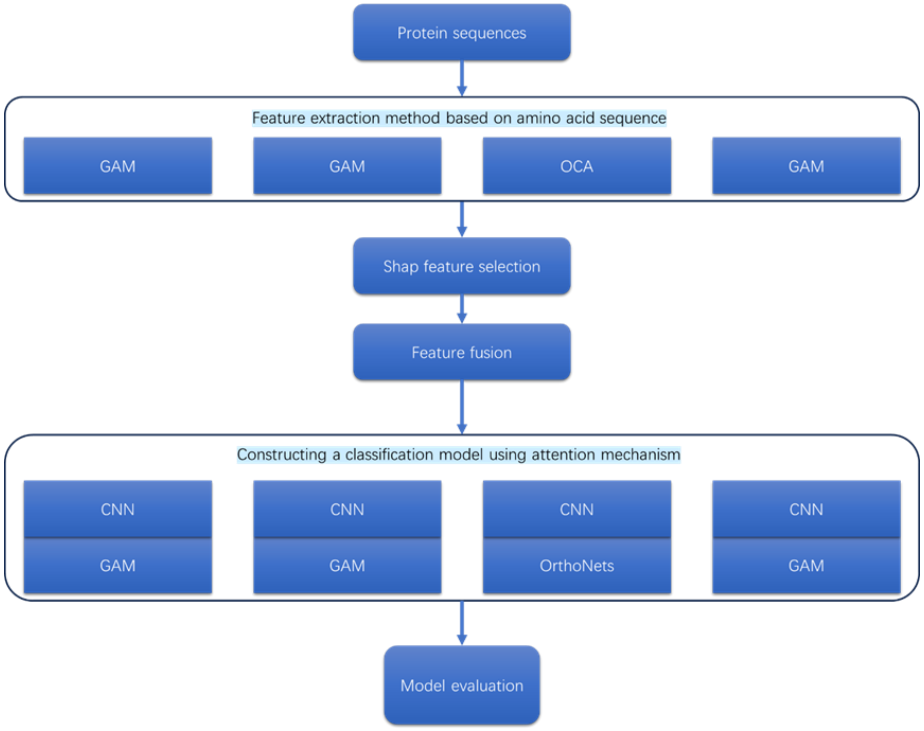


Fig. 1. Experimental flow chart

3.1 Attention Mechanisms

3.1.1 Global Attention Mechanism

Global Attention Mechanism (GAM) was a mechanism capable of reducing information loss and enhancing global interaction information[42], represented by equations (7) and (8). Originally designed for computer vision tasks, in the context of PTM site prediction, protein sequences could be encoded as multi-channel features. The channel attention sub-module of GAM dynamically allocated the importance of different feature channels, with the potential to automatically filter key biochemical signals. In contrast, traditional CNNs faced limitations in modeling long-range residue associations due to the constraints of convolution kernel size (such as interactions between enzyme active sites and substrate binding regions). The spatial attention sub-module of GAM, through global context awareness, could directly establish associations between any two residues to address this limitation.

$$F_2 = M_c(F_1) \otimes F_1 \tag{7}$$

$$F_3 = M_s(F_2) \otimes F_2 \tag{8}$$

In this context, $F_1 \in \mathbb{R}^{C\times H\times W}$ represented the given input feature map, F_2 denoted the intermediate state, and F_3 indicated the output result. M_c and M_s represented the channel and spatial attention maps, respectively, while $\otimes$ indicated element-wise multiplication.

GAM included two important sub-modules: the channel attention sub-module and the spatial attention sub-module. The channel attention sub-module utilized 3D permutation to retain information across three dimensions, and then amplified the inter-dimensional channel-space dependencies through a two-layer multi-layer perceptron (MLP). The spatial attention sub-module employed two convolutional layers to aggregate spatial information, with the reduction rate r being the same as that of the channel attention sub-module. Group convolution and channel shuffling were implemented to prevent a significant increase in parameters after de-pooling[43].

3.1.2 Polarized Self-Attention

The Polarized Self-Attention (PSA) block comprised two components: a pure channel module and a pure spatial module[44]. It accomplished filtering by fully collapsing features in one direction while maintaining high resolution in its orthogonal direction. The dynamic range of attention was increased by performing Softmax normalization at the bottleneck tensor (the minimal feature tensor within the attention block), using the Sigmoid function for tone mapping.

The pure channel module maintained high internal resolution during the calculations of channel and spatial attention while fully collapsing the corresponding dimensions of the input tensor $A^{ch}(X) \in \mathbb{R}^{C\times 1\times 1}$, which could be represented by the equation (9).

$$A^{ch}(X) = F_{SG}\left[W_{z|\theta_1}\left(\sigma_1\left(W_v(X)\right) \times Soft\,max\left(\sigma_2\left(W_q(X)\right)\right)\right)\right] \quad (9)$$

In this context, W_q, W_v, and W_z were each a 1 × 1 convolution layer, while σ_1 and σ_2 were two tensor reshaping operators. The symbol "×" represented the matrix dot product operation $Soft\,max(X) = \sum_{j=1}^{N_p} \frac{e^{xj}}{\sum_{m=1}^{N_p} e^{xm}} x_j$. The number of internal channels between W_v and $W_q|W_z$ was C/2. The output of the channel branch alone was $Z^{ch} = A^{ch}(X) \odot^{ch} X \in \mathbb{R}^{C\times H\times W}$, where $\odot^{ch}$ was the channel multiplication operator.

The pure spatial module was capable of combining non-linearity to directly adapt to the output distribution of typical fine regression $A^{sp}(X) \in \mathbb{R}^{1\times H\times W}$, as represented by the following equation (10).

$$A^{sp}(X) = F_{SG}\left[\sigma_3\left(Soft\,max\left(\sigma_1\left(F_{GP}\left(W_q(X)\right)\right)\right) \times \sigma_2\left(W_v(X)\right)\right)\right] \quad (10)$$

In this context, W_q and W_v were standard 1 x 1 convolution layers, while σ_1, σ_2, and σ_3 were three tensor reshaping operators. $F_{GP}(\cdot)$ was a global pooling operator $F_{GP}(X) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} X(:,i,j)$, and "×" represented the matrix dot product operation. The output of the pure spatial branch was $Z^{sp} = A^{sp}(X) \odot^{sp} X \in \mathbb{R}^{C\times H\times W}$, where $\odot^{sp}$ was the spatial multiplication operator.

The outputs of the pure channel module and the pure spatial module could be arranged in either a parallel layout or a serial layout. In this case, a parallel layout method was adopted, as represented by the following equation (11).

$$PSA_p(X) = Z^{ch} + Z^{sp} = A^{ch}(X) \odot^{ch} X + A^{sp}(X) \odot^{sp} X \tag{11}$$

3.1.3 Multi-Attention

The MAB module was originally designed to address the single-image super-resolution (SR) problem, which is an important task in the field of computer vision aimed at reconstructing missing high-frequency information from low-quality inputs[45]. We could simply understand this attention module as a form of information completion. Since amino acid sequence information does not contain structural information about the amino acids, and the process of feature extraction may lead to the loss of inter-relationships between various amino acids, this completion of information is theoretically reasonable for the prediction of amino acid sequences.

The MAB module internally included Multi-Scale Large Kernel Attention (MLKA) and Gated Spatial Attention Units (GSAU), which could be represented by the following equation (12).

$$\begin{cases} N = LN(X) \\ X = X + \lambda_1 f_3\left(MLKA\left(f_1(N)\right) \otimes f_2(N)\right) \\ N = LN(X) \\ X = X + \lambda_2 f_6\left(GSAU\left(f_4(N), f_5(N)\right)\right) \end{cases} \tag{12}$$

Where $LN(\cdot)$ and λ represented layer normalization and learnable scaling factors, respectively. $MLKA(\cdot)$ and $GSAU(\cdot)$ denoted the MLKA and GSAU modules, respectively. $\otimes$ indicated element-wise multiplication, and $f_i(\cdot)$ represented the i-th pointwise convolution that maintained dimensions. To preserve instance details and accelerate convergence, MAB utilized layer normalization instead of batch normalization or no normalization. MLKA combined large kernel attention (LKA) and a multi-scale mechanism, dynamically adjusting the attention maps through a gating mechanism to avoid potential block artifacts. This could be expressed using the following formula (13-15).

$$LKA(X) = f_{PW}\left(f_{DWD}\left(f_{DW}(X)\right)\right) \tag{13}$$

$$MLKA_i(X_i) = G_i(X_i) \otimes LKA_i(X_i) \tag{14}$$

$$GSAU(X, Y) = f_{DW}(X) \otimes Y \tag{15}$$

Given the input feature map $X \in \mathbb{R}^{C \times H \times W}$, LKA decomposed the convolution into three components to establish long-range relationships, including depthwise convolution, depthwise dilated convolution, and pointwise convolution, using K×K. Specifically, LKA achieved this by decomposing into one (2d-1)×(2d-1) depthwise

convolution, one $\left\lceil \frac{K}{d} \times \frac{K}{d} \right\rceil$ depthwise dilated convolution $f_{DWD}(\cdot)$, and one pointwise convolution $f_{DW}(\cdot)$. GSAU simplified the feedforward network through spatial attention and a gating mechanism, reducing computational costs while incorporating spatial information.

3.1.4 Orthogonal Channel Attention

Orthogonal Channel Attention (OrthoNets) randomly generated orthogonal filters and utilized an attention mechanism to weight the channels while incorporating residual connections, aiming to enhance the neural network's ability to process and represent input features[46]. This design integrated various techniques, including randomization, orthogonalization, attention mechanisms, and residual connections. The random generation of orthogonal filters introduced a certain degree of randomness and diversity, while the attention mechanism adaptively adjusted the importance of the channels. The residual connections helped alleviate issues such as gradient vanishing in deep networks, collectively enhancing the model's performance and generalization capability.

The filters were represented by $A(X) = E\left(F_{ortho}(X)\right)$, and the squeeze process was expressed using formula $F_{\mathrm{ortho}}(X)_c = \sum_{h=1}^{H} \sum_{w=1}^{W} K_{c,h,w} X_{c,h,w}$. The channel attention formula was denoted by $A(X) = E\left(F_{ortho}(X)\right)$.

In the first stage, random filters were initialized, with sizes matching those of the feature layer. The Gram-Schmidt process was used to orthogonalize these filters. Random numbers entered the Gram-Schmidt module, and through the Gram-Schmidt orthogonalization process, orthogonal filters were generated, with dimensions of H×W×C, where H×W×C represented height, width, and the number of channels, respectively.

In the second stage, the filters were utilized to extract compressed vectors, and the excitation proposed by SENet was employed to obtain attention vectors. By multiplying the attention vectors with the input features, the weighted output features were computed, and the residuals were added.

In the third stage, the input data was processed through convolution blocks, resulting in input feature maps with dimensions of H×W×C. The orthogonal filters from Stage 0 were operated on the input feature maps, and after a squeeze operation, the feature maps were compressed to dimensions of 1×1×C. The squeezed features underwent an excitation operation to generate attention weights. The attention weights were then multiplied by the input feature maps processed through the convolution blocks, resulting in weighted feature maps.

The residual connection from the original input X was added to the weighted feature maps, ultimately yielding the output with dimensions of H×W×C.

3.2 SHAP

Shapley Additive exPlanations (SHAP) was a model interpretation method based on game theory, aimed at providing clear explanations for the outputs of machine learning models[47-48]. It fairly distributed the contributions of features to the prediction results

using the classical Shapley values, helping us understand the decision-making process of the model.

When using SHAP for interpretation, it was first necessary to create an explainer that supported various model types, such as tree models and deep learning models. We utilized the default XGBoost model. The calculation of SHAP values helped us understand the impact of each feature on the model's predictions. SHAP provided several visualization tools, such as Summary Plot, Bar Plot, and Beeswarm Plot, which intuitively displayed the importance of features and their effects on the prediction results, as illustrated in Figure 2.

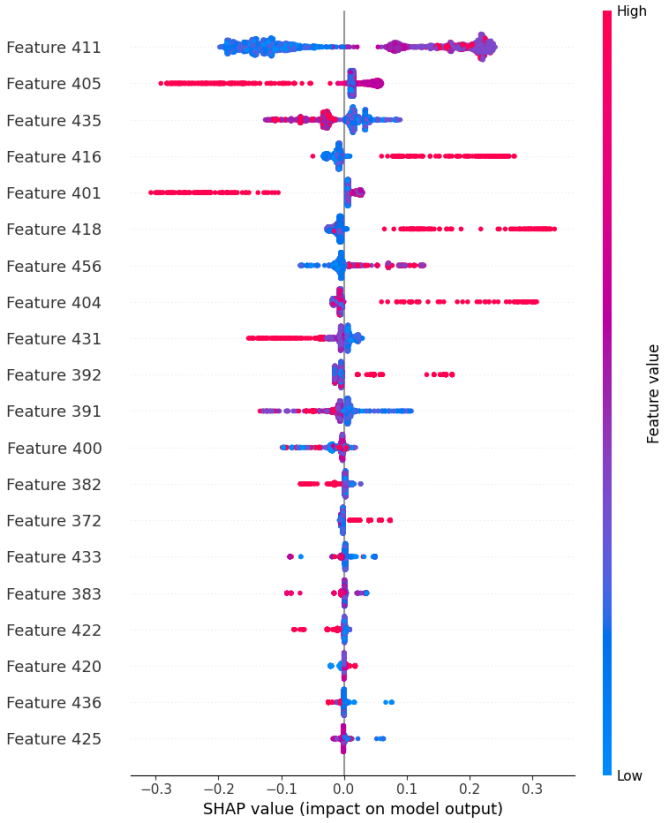


Fig. 2. SHAP Performance on BLOSUM62 Features

These visualizations not only helped identify important features but also revealed the interactions between features.

3.3 Model evaluation

We used accuracy (ACC), sensitivity (SN), specificity (SP), F1 score (F1), and Matthews correlation coefficient (MCC) to evaluate model performance[49] (see

formulas (16-20)). We took the average from ten-fold cross-validation to present the final results. In the prediction process, true positives (TP) referred to the positive samples correctly identified by the model, while true negatives (TN) referred to the negative samples accurately recognized by the model. False positives (FP) indicated negative samples incorrectly classified as positive, and false negatives (FN) indicated positive samples incorrectly classified as negative[50]. Accuracy (ACC) reflected the overall predictive accuracy of the samples.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

SN and SP represented the predictive accuracy of negative and positive samples, respectively.

$$SN = \frac{TP}{TP + TN} \tag{17}$$

$$SP = \frac{TN}{TN + FP} \tag{18}$$

The Matthews correlation coefficient (MCC) comprehensively considered true positives, true negatives, false positives, and false negatives, making it suitable for datasets with class imbalance. It effectively reduced the impact of class distribution on the results and summarized classification quality with a single value.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{19}$$

The F1 score achieved a balance between precision and recall by harmonizing the two. In cases of severe class imbalance, MCC was regarded as a more comprehensive performance evaluation metric, while F1 score could be prioritized as the evaluation criterion when focusing on the balance between precision and recall.

$$F1 = 2 \times \frac{TP}{2TP + FP + FN} \tag{20}$$

4 Model results and analysis

In this experiment, we concatenated convolutional neural networks with different attention mechanisms. The convolutional neural networks extracted rich local relational information in upstream tasks, while the attention mechanisms provided redundancy reduction and key information enhancement in downstream tasks, with important parameters required for each constructed model shown in Table 1.

Table1. Model Parameters

Model	Key parameters	values
CNN+PSA	channels	64
CNN+GAM	channels, rate	64, 4
CNN+MAB	channels	243
CNN+OrthoNets	channels, height	256, 64

After constructing the classification models, we input five different encoded features and their combined all_str feature into the model for evaluation, with results presented in Table 2. The BLOSUM62 feature performed the best in the model, followed by the combined all_str feature, while the ASDC, CKSAAGP, and DPC features were almost completely misclassified. In this round of experiments, the CNN+OrthoNets model achieved the highest accuracy, with an ACC value of 85.06%, followed by CNN+MAB, which reached an ACC value of 83.59%. The performances of the other two models, CNN+PSA and CNN+GAM, were relatively similar.

Table2. The performance of the CNN+attention model on various features

Attention mechanism: PSA					
Feature	ACC	SP	SN	MCC	F1
ASDC	0.4989	0.9949	0.0028	-0.0184	0.6650
EAAC	0.7281	0.8488	0.6074	0.4701	0.7574
CKSAAGP	0.4992	0.9985	0.0000	-0.0276	0.6660
BLOSUM62	0.8134	0.8777	0.7491	0.6320	0.8246
DPC	0.4922	0.9789	0.0056	-0.0677	0.6585
All_str	0.8003	0.8701	0.7306	0.6066	0.8134
Attention mechanism: GAM					
Feature	ACC	SP	SN	MCC	F1
ASDC	0.4989	0.9985	0.0000	-0.0184	0.6650
EAAC	0.7281	0.7823	0.6065	0.4701	0.7574
CKSAAGP	0.4992	0.9574	0.0102	-0.0276	0.6660
BLOSUM62	0.8134	0.8684	0.8028	0.6320	0.8246
DPC	0.4922	0.9970	0.0000	-0.0677	0.6585
All_str	0.8003	0.8902	0.7074	0.6066	0.8134
Attention mechanism: MAB					
Feature	ACC	SP	SN	MCC	F1
ASDC	0.4986	0.9944	0.0028	-0.0215	0.6648
EAAC	0.7333	0.8518	0.6148	0.4803	0.7616
CKSAAGP	0.4990	0.9980	0.0000	-0.0319	0.6658

BLOSUM62	0.8359	0.8792	0.7926	0.6743	0.8427
DPC	0.4845	0.9578	0.0111	-0.0964	0.6501
All_str	0.8086	0.8736	0.7435	0.6224	0.8203

Attention mechanism: OrthoNets

Feature	ACC	SP	SN	MCC	F1
ASDC	0.4798	0.9254	0.0343	-0.0889	0.6402
EAAC	0.7750	0.8853	0.6648	0.5640	0.7974
CKSAAGP	0.4784	0.9169	0.0398	-0.0900	0.6374
BLOSUM62	0.8506	0.8938	0.8074	0.7038	0.8568
DPC	0.4654	0.8901	0.0407	-0.1310	0.6248
All_str	0.8117	0.8631	0.7602	0.6266	0.8209

Building on the above, in order to further investigate whether the ASDC, CKSAAGP, and DPC features truly lacked experimental value, we utilized the interpretable model SHAP to visualize the selected features. We then filtered out a maximum of 20 useful features from each reduced feature file and added the suffix "select" to the filtered feature files. To understand the impact of the ASDC, CKSAAGP, and DPC features on the overall feature set All_str, we created a new set that included the results from the filtered feature files, termed All_select_str, with a clear explanation provided in Figure 3.

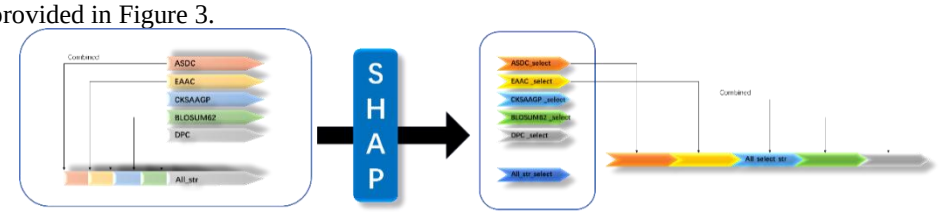


Fig. 3. feature selection

The benefit of this process was that it achieved dimensionality reduction for each feature, which could be interpreted as the pixel contribution of feature channels to the categories.

Subsequently, we input the selected features into the classification model again, and the performance results were presented in Table 3. The BLOSUM62 feature continued to perform the best across all models, but this time the model with the highest accuracy became CNN+MAB, with an ACC value increasing by 5.32% before and after feature selection, and an MCC value increasing by 0.1050. Additionally, an interesting phenomenon was observed in the regrouping results. Theoretically, the All_str_select features, as the selection results of All_str, should have drawn from the optimal parts of each feature; however, their performance on each model was generally lower than that of the All_select_str features. This indicated that the ASDC, CKSAAGP, and DPC feature sets also held certain research value.

Table3. The performance of the CNN + attention model on the selected features

Attention mechanism: PSA					
Feature	ACC	SP	SN	MCC	F1
ASDC_select	0.5010	0.9919	0.0102	0.0108	0.6653
EAAC_select	0.7861	0.9185	0.6537	0.5934	0.8111
CKSAAGP_select	0.5000	1.0000	0.0000	0.0000	0.6667
BLOSUM62_select	0.8805	0.9074	0.8537	0.7622	0.8837
DPC_select	0.5000	1.0000	0.0000	0.0000	0.6667
All_str_select	0.8758	0.8988	0.8528	0.7524	0.8786
All_select_str	0.8829	0.9074	0.8583	0.7667	0.8857
Attention mechanism: GAM					
Feature	ACC	SP	SN	MCC	F1
ASDC_select	0.5010	0.9959	0.0037	0.0108	0.6653
EAAC_select	0.7861	0.8685	0.6713	0.5934	0.8111
CKSAAGP_select	0.5000	0.9980	0.0009	0.0000	0.6667
BLOSUM62_select	0.8805	0.9090	0.8259	0.7622	0.8837
DPC_select	0.5000	1.0000	0.0000	0.0000	0.6667
All_str_select	0.8758	0.9054	0.8463	0.7524	0.8786
All_select_str	0.8829	0.9090	0.8389	0.7667	0.8857
Attention mechanism: MAB					
Feature	ACC	SP	SN	MCC	F1
ASDC_select	0.5020	0.9939	0.0102	0.0228	0.6662
EAAC_select	0.7797	0.8928	0.6667	0.5743	0.8021
CKSAAGP_select	0.4990	0.9980	0.0000	-0.0319	0.6658
BLOSUM62_select	0.8891	0.9170	0.8611	0.7793	0.8921
DPC_select	0.4923	0.9801	0.0046	-0.0695	0.6588
All_str_select	0.8810	0.9130	0.8491	0.7636	0.8847
All_select_str	0.8865	0.9099	0.8630	0.7738	0.8891
Attention mechanism: OrthoNets					
Feature	ACC	SP	SN	MCC	F1
ASDC_select	0.4998	0.9838	0.0157	-0.0018	0.6629
EAAC_select	0.7688	0.8811	0.6565	0.5517	0.7922
CKSAAGP_select	0.4876	0.9668	0.0083	-0.0873	0.6536
BLOSUM62_select	0.8768	0.9165	0.8370	0.7559	0.8815
DPC_select	0.5000	1.0000	0.0000	0.0000	0.6667
All_str_select	0.8750	0.9120	0.8380	0.7520	0.8794
All_select_str	0.8789	0.9079	0.8500	0.7591	0.8823

5 Conclusions

In recent years, with the deepening research into the functions of plant proteins, the precise prediction of post-translational modification sites of proteins became a critical bottleneck in elucidating the mechanisms of traditional Chinese medicine. However, traditional experimental methods faced limitations such as low efficiency, high costs, and difficulty in handling large-scale data. Therefore, developing efficient and reliable computational models emerged as an important research direction in this field.

For Paeonia lactiflora, a traditional Chinese medicinal material with significant pharmacological activity, this study innovatively integrated multi-dimensional features of amino acid sequences and successfully constructed a protein post-translational modification site prediction model based on a deep learning framework combining convolutional neural networks and attention mechanisms, suitable for the characteristics of different herbal materials. By utilizing a dedicated dataset for Paeonia lactiflora (1080 positive samples and 1976 negative samples) built from the TCMSP database, the feasibility of this method in predicting complex herbal protein modifications was systematically validated.

Although the current model demonstrated effective recognition capabilities for key modification sites, its classification accuracy still required further improvement, particularly in optimizing the distinction of boundary cases between adjacent non-modified sites. In the future, we can improve in the following aspects: (1) explore more encoding methods and perform pairwise combinations; (2) investigate the integration of three-dimensional structural information and multimodal data fusion strategies; (3) leverage advanced techniques such as transfer learning and federated learning for feature fusion, which are expected to significantly enhance the model's generalization ability and predictive accuracy. This provides important theoretical support for revealing the mechanisms of action of active components in traditional Chinese medicine and reducing clinical medication risks.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (Grant No. 62333018), Xuzhou Science and Technology Plan Project (KC21047), Jiangsu Provincial Natural Science Foundation (No. SBK2019040953), Natural Science Fund for Colleges and Universities in Jiangsu Province (No. 19KJB520016) and Young Talents of Science and Technology in Jiangsu and ghfund202302026465.

References

1. Yan, L., Wu, Y., Guan, R., Jin, C., Han, R., Ou, J., Tong, X.: Nontargeted Metabolomic Profiling of a Single Paeonia Lactiflora Plant and Its Quality Marker Identification. ChemistryOpen. 2400520 (2025)
2. Sun, S., Jimu, R. B., Lema, A. K., Elmamoune, H., Fan, Z., Jin, C., Han, R.: A Systematic Review on the Origin, Anti-Inflammatory Effect, Mechanism, Pharmacokinetics, and Toxicity of Albiflorin. Arabian Journal of Chemistry. 17(7), 105836 (2024)
3. Lim, J. W., Kang, M. K., Kim, H. E.: Antibacterial Effects of Paeonia Lactiflora Extract on Oral Microcosm Biofilms. Applied Sciences. 14(23), 11290 (2024)

4. Wu, F., Wu, G., Li, T., Lu, W., Fu, T., Zhang, Z.: Exploring the Target and Mechanism of Radix Paeoniae Alba on Sjogren's Syndrome. Combinatorial Chemistry & High Throughput Screening. 26(6), 1224--1232 (2023)
5. Huang, Y., He, M., Zhang, J., Cheng, S., Cheng, X., Chen, H., Zeng, S.: White Tea Aqueous Extract: A Potential Anti-Aging Agent Against High-Fat Diet-Induced Senescence in Drosophila melanogaster. Foods. 13(24), 4034 (2024)
6. Chen, L., Kan, J., Zheng, N., Li, B., Hong, Y., Yan, J., Li, H.: A Botanical Dietary Supplement from White Peony and Licorice Attenuates Nonalcoholic Fatty Liver Disease by Modulating Gut Microbiota and Reducing Inflammation. Phytomedicine. 91, 153693 (2021)
7. Mu, X., Luan, R., Gao, Y., Zhao, B., Wang, J., Ni, X., Gao, D.: The Traditional Applications, Phytochemistry, Pharmacology, Pharmacokinetics, Quality Control and Safety of Paeoniae Radix Alba: A Review. The American Journal of Chinese Medicine. 52(08), 2337--2376 (2024)
8. Fang, C., Xu, X., Lu, F., Liu, S.: Study on the Collaborative Protective Mechanism of Scutellariae Radix and Paeoniae Radix Alba Against Diabetic Cardiomyopathy Through the Gut-Heart Axis. Frontiers in Microbiology. 16, 1500935 (2025)
9. Jiang, H., Li, J., Wang, L., Wang, S., Nie, X., Chen, Y., He, Y.: Total Glucosides of Paeony: A Review of Its Phytochemistry, Role in Autoimmune Diseases, and Mechanisms of Action. Journal of Ethnopharmacology. 258, 112913 (2020)
10. Luan, X., Zhang, X., Nie, M., Zhao, Y.: Traditional Chinese Medicine Integrated Responsive Microneedles for Systemic Sclerosis Treatment. Research. 6, 0141 (2023)
11. Xu, M. Y., Zhang, J. B., Peng, Y. Z., Liu, M. C., Ma, S. Y., Zhou, Y., Ma, S.: Network Pharmacology and Experimental Validation Identify Paeoniflorin as a Novel SRC-Targeted Therapy for Castration-Resistant Prostate Cancer. Pharmaceuticals. 18(8), 1241 (2025)
12. Wang, Y., He, Q. Q., Zhu, Y. T., Zhang, Y., Yan, J., Liang, L. F., Ke, P. F.: Total Glucosides of Paeony Ameliorates Lupus Nephritis by Suppressing ZBP1-Mediated PANoptosis in Podocytes. Phytomedicine. 156996 (2025)
13. Liao, T., Kang, J., Ma, Z., Jie, L., Feng, M., Liu, D., Xing, R.: Total Glucosides of White Paeony Capsule Alleviate Articular Cartilage Degeneration and Aberrant Subchondral Bone Remodeling in Knee Osteoarthritis. Phytotherapy Research. 39(4), 1758--1775 (2025)
14. Yan, M., Wang, Q., Yang, H., Liu, D., Liang, W., Chen, H.: The Paeonol of Total Glucosides of White Paony Regulates the Differentiation of CD4+ Treg Cells Through the EP300/Foxp3 Axis to Relieve Pulmonary Fibrosis in Mice. Cell Biochemistry and Biophysics. 1--12 (2025)
15. Lee, J. M., Hammarén, H. M., Savitski, M. M., Baek, S. H.: Control of Protein Stability by Post-Translational Modifications. Nature Communications. 14(1), 201 (2023)
16. Agrata, R., Komander, D.: Ubiquitin—A Structural Perspective. Molecular Cell. 85(2), 323--346 (2025)
17. Wu, X., Xu, M., Geng, M., Chen, S., Little, P. J., Xu, S., Weng, J.: Targeting Protein Modifications in Metabolic Diseases: Molecular Mechanisms and Targeted Therapies. Signal Transduction and Targeted Therapy. 8(1), 220 (2023)
18. Wang, R., Li, Y., Ji, J., Kong, L., Huang, Y., Liu, Z., Lu, L.: The Emerging Role of Herbal Medicines in Cancer by Interfering with Posttranslational Modifications. Antioxidants & Redox Signaling. 42(1-3), 150--164 (2025)
19. Zhang, N., Wang, X., Li, Y., Lu, Y., Sheng, C., Sun, Y., Jiao, Y.: Mechanisms and Therapeutic Implications of Gene Expression Regulation by circRNA-Protein Interactions in Cancer. Communications Biology. 8(1), 77 (2025)

20. Shao, S., Lv, B., Wang, M., Zeng, S., Wang, S., Yang, Z., Ma, P.: Biosynthesis and Regulatory Mechanism of Tanshinones and Phenolic Acids in Salvia miltiorrhiza. The Plant Journal. 123(2), e70358 (2025)
21. Huang, Z., Zhu, J., Zhou, Y. L., Shi, J.: The cGAS-STING Pathway: A Dual Regulator of Immune Response in Cancer and Therapeutic Implications. Journal of Translational Medicine. 23(1), 766 (2025)
22. Geffen, Y., Anand, S., Akiyama, Y., Yaron, T. M., Song, Y., Johnson, J. L., Zhou, D. C.: Pan-Cancer Analysis of Post-Translational Modifications Reveals Shared Patterns of Protein Regulation. Cell. 186(18), 3945--3967 (2023)
23. van der Gaag, B. L., Deshayes, N. A., Breve, J. J., Bol, J. G., Jonker, A. J., Hoozemans, J. J., van de Berg, W. D.: Distinct tau and alpha-synuclein Molecular Signatures in Alzheimer's Disease with and without Lewy Bodies and Parkinson's Disease with Dementia. Acta Neuropathologica. 147(1), 14 (2024)
24. Ertelt, M., Mulligan, V. K., Maguire, J. B., Lyskov, S., Moretti, R., Schiffner, T., Schoeder, C. T.: Combining Machine Learning with Structure-Based Protein Design to Predict and Engineer Post-Translational Modifications of Proteins. PLOS Computational Biology. 20(3), e1011939 (2024)
25. Öztürk, H., Özgür, A., Ozkirimli, E.: DeepDTA: Deep Drug–Target Binding Affinity Prediction. Bioinformatics. 34(17), i821--i829 (2018)
26. Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K., Kumar, P.: Artificial Intelligence to Deep Learning: Machine Intelligence Approach for Drug Discovery. Molecular Diversity. 25(3), 1315--1360 (2021)
27. Asgari, E., Mofrad, M. R.: Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. PloS One. 10(11), e0141287 (2015)
28. Krivák, R., Hoksza, D.: P2Rank: Machine Learning Based Tool for Rapid and Accurate Prediction of Ligand Binding Sites from Protein Structure. Journal of Cheminformatics. 10(1), 39 (2018)
29. Wang, D., Zeng, S., Xu, C., Qiu, W., Liang, Y., Joshi, T., Xu, D.: MusiteDeep: A Deep-Learning Framework for General and Kinase-Specific Phosphorylation Site Prediction. Bioinformatics. 33(24), 3909--3916 (2017)
30. Li, Y., Wang, S., Umarov, R., Xie, B., Fan, M., Li, L., Gao, X.: DEEPre: Sequence-Based Enzyme EC Number Prediction by Deep Learning. Bioinformatics. 34(5), 760--769 (2018)
31. Schubach, M., Maass, T., Nazaretyan, L., Röner, S., Kircher, M.: CADD v1.7: Using Protein Language Models, Regulatory CNNs and Other Nucleotide-Level Scores to Improve Genome-Wide Variant Predictions. Nucleic Acids Research. 52(D1), D1143--D1154 (2024)
32. Tan, D., Jiang, H., Li, H., Xie, Y., Su, Y.: Prediction of Drug–Protein Interaction Based on Dual Channel Neural Networks with Attention Mechanism. Briefings in Functional Genomics. 23(3), 286--294 (2024)
33. Xu, Y., Cheng, J.: Secondary Structure Prediction of Protein Based on Multi Scale Convolutional Attention Neural Networks. Mathematical Biosciences and Engineering: MBE. 184, 3404--3422 (2021)
34. Luo, Z., Wang, Q., Xia, Y., Zhu, X., Yang, S., Xu, Z., Gu, L.: DLBWE-Cys: A Deep-Learning-Based Tool for Identifying Cysteine S-Carboxyethylation Sites Using Binary-Weight Encoding. Frontiers in Genetics. 15, 1464976 (2025)
35. He, Y., Zheng, X., Miao, Q.: TFA-CLSTMNN: Novel Convolutional Network for Sound-Based Diagnosis of COVID-19. International Journal of Wavelets, Multiresolution and Information Processing. 21(03), 2250058 (2023)

36. Liang, B., Zhu, Y., Shi, W.: SARS-CoV-2 Spike Protein Post-Translational Modification Landscape and Its Impact on Protein Structure and Function via Computational Prediction. Res (Washington DC). 6, 0078 (2023)
37. Chou, K. C.: Prediction of Signal Peptides Using Scaled Window. Peptides. 22(12), 1973--1979 (2001)
38. Lee, T. Y., Chen, S. A., Hung, H. Y., Ou, Y. Y.: Incorporating Distant Sequence Features and Radial Basis Function Networks to Identify Ubiquitin Conjugation Sites. PloS One. 6(3), e17331 (2011)
39. Chen, Z., Zhao, P., Li, C., Li, F., Xiang, D., Chen, Y. Z., Song, J.: iLearnPlus: A Comprehensive and Automated Machine-Learning Platform for Nucleic Acid and Protein Sequence Analysis, Prediction and Visualization. Nucleic Acids Research. 49(10), e60--e60 (2021)
40. Wei, L., Zhou, C., Chen, H., Song, J.: Acpred-fl: A Sequence-Based Predictor Using Effective Feature Representation to Improve the Prediction of Anti-Cancer Peptides. Bioinformatics. 34, 4007--4016 (2018). DOI: https://doi.org/10.1093/bioinformatics/bty451
41. Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., Song, J.: iFeature: A Python Package and Web Server for Features Extraction and Selection from Protein and Peptide Sequences. Bioinformatics. 34(14), 2499--2502 (2018)
42. Liu, Y., Shao, Z., Hoffmann, N.: Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions. arXiv Preprint arXiv:2112.05561 (2021)
43. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6848--6856) (2018)
44. Liu, H., Liu, F., Fan, X., Huang, D.: Polarized Self-Attention: Towards High-Quality Pixel-Wise Regression. arXiv Preprint arXiv:2107.00782 (2021)
45. Wang, Y., Li, Y., Wang, G., Liu, X.: Multi-Scale Attention Network for Single Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5950--5960) (2024)
46. Salman, H., Parks, C., Swan, M., Gauch, J.: Orthonets: Orthogonal Channel Attention Networks. In 2023 IEEE International Conference on Big Data (BigData) (pp. 829--837). IEEE (2023)
47. Jabeur, S. B., Mefteh-Wali, S., Viviani, J. L.: Forecasting Gold Price with the XGBoost Algorithm and SHAP Interaction Values. Annals of Operations Research. 334(1), 679--699 (2024)
48. Lundberg, S. M., Lee, S. I.: A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems. 30 (2017)
49. Araf, I., Idri, A., Chairi, I.: Cost-Sensitive Learning for Imbalanced Medical Data: A Review. Artificial Intelligence Review. 57(4), 80 (2024)
50. Hassanzadeh, R., Farhadian, M., Rafieemehr, H.: Hospital Mortality Prediction in Traumatic Injuries Patients: Comparing Different SMOTE-Based Machine Learning Algorithms. BMC Medical Research Methodology. 23(1), 101 (2023)

Author Index

Anchi Sun 1
Kai Xiao 5
Long Cheng 1
Wenzheng Bao 5, 19
Yingyue Tang 19
Zhiqiang Hui 1

Table of Email Addresses from the Corresponding Authors

0001 allen_brown@163.com
0019 baowz55555@126.com
0005 baowz55555@126.com