2025 International Conference on Applied Intelligence
November 6-9, Nanning, 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, including nourishing blood to regulate menstruation, astringing yin to reduce sweating, soothing the liver to alleviate pain, and suppressing hyperactivity of liver-yang. 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]. Therefore, accurate identification of PTM sites is crucial for exploring the functional mechanisms underlying the pharmacological activities of Paeonia lactiflora.

Traditional experimental techniques have established a comprehensive system for PTM site identification. Centered on liquid chromatography-tandem mass spectrometry (LC-MS/MS [8,9]), this system integrates high-resolution mass spectrometry and fragmentation techniques (e.g., collision-induced dissociation, CID [10]; electron transfer dissociation, ETD [10]), enabling precise localization of modification sites. Meanwhile, complementary approaches such as antibody affinity enrichment (e.g., phosphorylation antibodies [11]), chemical probe labeling (e.g., click chemistry [12]), and selective enrichment strategies (e.g., immobilized metal ion affinity chromatography, IMAC; titanium dioxide, TiO_2 [13]) significantly improve the detection efficiency of low-abundance modified peptides. Two-dimensional gel electrophoresis (2D-PAGE [14]) can reveal differences between modified proteins; mutant construction and in vitro enzymatic reactions validate functional modification sites; and hydrogen/deuterium exchange mass spectrometry (HDX-MS [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 overcome these limitations, computational methods have emerged as important alternatives. 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]. With the development of machine learning and deep learning, 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.

Despite the progress made by these methods, limitations remain: first, most models rely on manually designed feature extraction techniques, which struggle to capture complex sequence relationships, leading to an incomplete understanding of protein sequences; second, most models target only a single PTM type or specific lysine PTM sites, lacking a universal model applicable to diverse protein sequences.

To address these shortcomings, this paper proposes a novel universal site prediction framework for identifying sites in protein sequences. Figure 1 illustrates the data preparation process, including data collection, generating positive and negative samples. Figure 3 illustrates the construction of the site prediction model. First, we transform the amino acid sequence into a numerical feature representation by combining word embeddings and position embeddings, which more comprehensively encodes amino acid composition and sequence order. The sequence embeddings are then processed through a Transformer architecture, which captures both local and global dependencies within the sequence. 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 Method and material

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 2θ+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, 1080 positive samples and 1976 negative samples were finally obtained, completing the dataset construction.

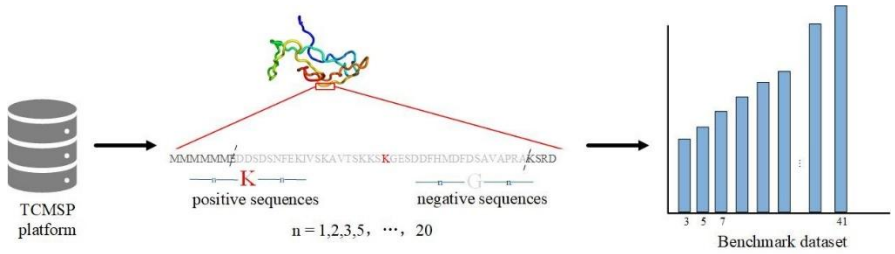

Figure 1. Workflow of the dataset construction

2.2 Amino acid encoding

When processing non-histone sequence data, One-hot Encoding extracts features by converting each amino acid in the protein sequence into 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 $2\theta+1$ (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 $20\times(2\theta+1)$ 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 mainly functions to convert each amino acid residue in the peptide chain into a low-dimensional dense vector representation, 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, the model can accurately perceive the positional information of each residue in the sequence, thereby understanding the sequential structural characteristics of the peptide chain.

In recent years, Language Models (LMs) have attracted wide attention because they can obtain contextualized embeddings from large-scale unlabeled language datasets, rather than static and context-insensitive word embeddings. 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 transfer it to downstream tasks such as protein property prediction and modification site identification, 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]. This model is a protein language model based on the Transformer architecture, 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 fine-tuning stage is carried out on the UniRef50 database, which provides clustered sequence data from UniProtKB and selected UniParc records, 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.

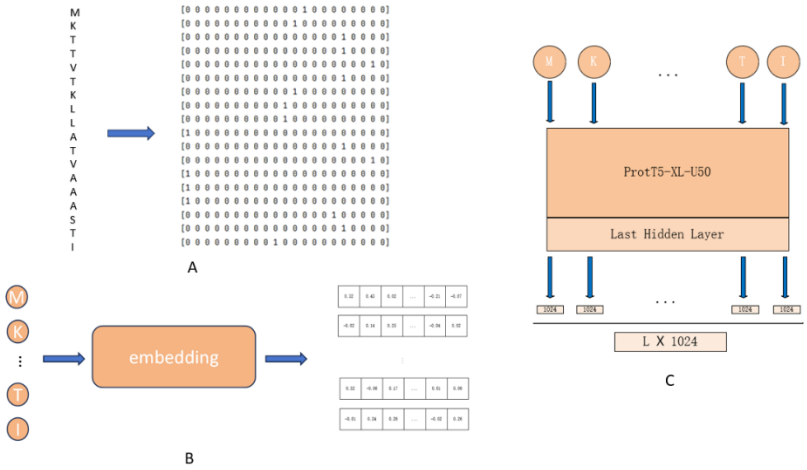

Figure 2. Amino acid encoding. (A)One-hot encoding,(B) Word embedding ,(C) Embeddings of ProtT5.

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.

In addition to the attention sub - layer, each layer of the Transformer block contains a fully - connected feed - forward network, which is applied individually and identically to each position. It consists of two linear transformations with a ReLU activation in between.

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

2.4 Framework of the proposed model

We employed a combination of word and positional embeddings to convert protein sequences into vector representations. Each amino acid in the protein sequence is mapped to a unique numerical index based on a predefined amino acid vocabulary. For example, amino acids such as alanine (A), cysteine (C), and aspartic acid (D) are mapped to indices 3, 4, and 5, respectively. Mathematically, each amino acid x_i in the sequence is represented by its corresponding index, where the index $E(x_i)$ is derived from the vocabulary:

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

Positional embeddings are used to encode the positional information of each amino acid, enabling the model to capture the sequence order within the protein chain. The positional embedding of the i-th amino acid is denoted as P_i, which is a learnable embedding matrix:

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

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

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

Once the protein sequence is converted into an embedding matrix H $\in \mathbb{R}^{L\times d}$, this matrix H serves as the input to a deep learning model that is specifically designed to capture both local and global dependencies within the sequence.

First, the embedding matrix H is processed through the stacked_BiLSTM layer. The LSTM layer is designed to hierarchically capture dependencies along the sequence in

both the forward and backward directions, enabling the model to learn contextual information from the preceding and following amino acids. The BiLSTM layer processes the input H to generate an enhanced sequence representation H_{LSTM}:

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

where $H_{LSTM} \in \mathbb{R}^{L \times d}$ is the output of the LSTM layer. This representation integrates the context of each amino acid from both directions. Then, H is fed into a series of Transformer encoder modules. Each Transformer block consists of multi-head self-attention and feed-forward operations, enabling the model to capture complex long-range dependencies between amino acids. Mathematically, the Transformer block processes the input as follows:

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

where, $H_{tran} \in \mathbb{R}^{L \times d}$ represents the refined sequence embedding after passing through N Transformer blocks. The multi-head self-attention mechanism in the Transformer block allows the model to focus on different parts of the sequence simultaneously, thereby enhancing its ability to model the global relationships between amino acids. Finally, to convert the variable-length sequence information into a fixed-size representation, an adaptive max-pooling layer is applied to the output of the Transformer block. This operation reduces the sequence dimension, thus producing a compressed feature vector.

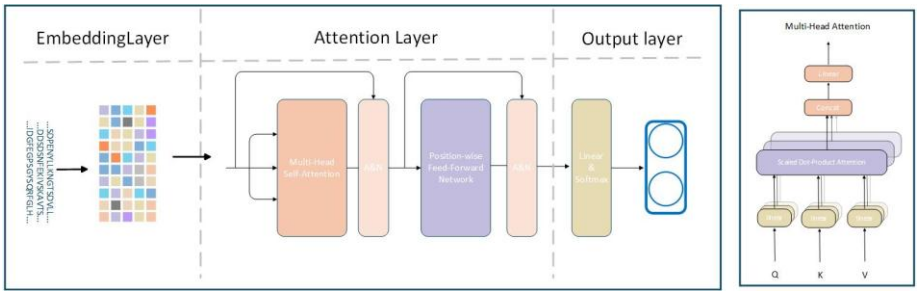

Figure 3. The overall architecture of the model

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.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (11)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \quad (12)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \quad (13)$$

$$\text{F1score} = 2 \times \frac{\text{TP}}{\text{2TP+FP+FN}} \quad (14)$$

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

.

3 Results and discussions

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.

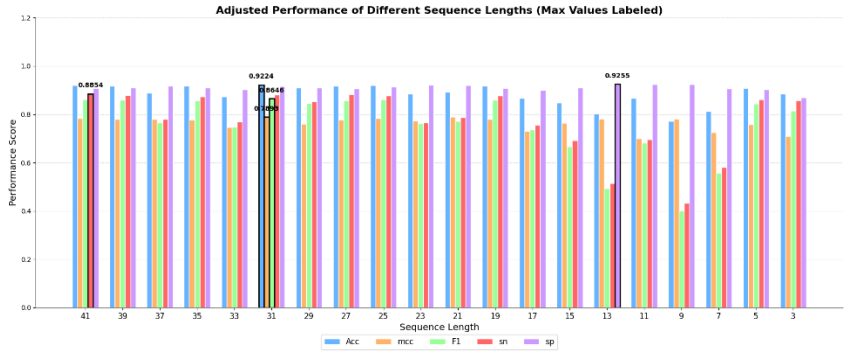

Figure 4. Indicators of stackedbilstm_transformer on the training dataset for sliding window size range

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%

Comparative analysis can provide insights into the strengths and weaknesses of different methods and guide future research. Therefore, we further employed five different classifiers for ten-fold cross-validation, including two baseline machine learning models (RF and SVM) and two baseline deep learning models (CNN and LSTM), to compare the predictive performance of our model. We used PseAAC as the data input for the machine learning models. Table 1 lists the outputs of all five methods on the dataset, including ACC, SN, SP, F1 score, and MCC value. Our method significantly outperformed the other models in terms of ACC, MCC, and F1 value. Transformer achieved ACC, SN, SP, MCC, and F1 values of 92%, 88%, 92%, 91%, and 86%, respectively. These results demonstrate that our model is stable and performs well, and can serve as an effective model for predicting 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 Conclusions

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 Traditional Chinese Medicine Systems Pharmacology Database (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.

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. Seo J-W, Lee K-J. Post-translational modifications and their biological functions: proteomic analysis and systematic approaches. BMB Rep 2004;37(1):35–44.
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 2018;46(D1):D901–10.
3. Keith Keenan E, Zachman DK, Hirschey MD. Discovering the landscape of protein modifications. Mol Cell 2021;81(9):1868–78.
4. Walsh CT, Garneau-Tsodikova S, Gatto Jr GJ. Protein posttrans-lational modifications: the chemistry of proteome diversifications. Angew Chem Int Ed 2005;44(45):7342–72.
5. Jensen ON. Interpreting the protein language using proteomics.Nat Rev Mol Cell Biol 2006;7(6):391–403.
6. Yang X-J, Seto E. Lysine acetylation: codified crosstalk with other posttranslational modifications. Mol Cell 2008;31(4):449–61.
7. Millán-Zambrano G, Burton A, Bannister A J, et al. Histone post-translational modifications—cause and consequence of genome function[J]. Nature Reviews Genetics, 2022, 23(9): 563-580.
8. W. Zhao, et al., Systematic identification of the lysine lactylation in the protozoan parasite toxoplasma gondii, Parasit. Vectors 15 (1) (2022) 180.
9. J.H. Wang, et al., Beyond metabolic waste: lysine lactylation and its potential roles in cancer progression and cell fate determination, Cell. Oncol. (Dordr.) 46 (3) (2023) 465–480.
10. Guthals A, Bandeira N. Peptide identification by tandem mass spectrometry with alternate fragmentation modes[J]. Molecular & Cellular Proteomics, 2012, 11(9): 550-557.

11. Roque A C A, Silva C S O, Taipa M Â. Affinity-based methodologies and ligands for antibody purification: advances and perspectives[J]. Journal of Chromatography A, 2007, 1160(1-2): 44-55.
12. Horisawa K. Specific and quantitative labeling of biomolecules using click chemistry[J]. Frontiers in physiology, 2014, 5: 457.
13. Li X S, Yuan B F, Feng Y Q. Recent advances in phosphopeptide enrichment: strategies and techniques[J]. TrAC Trends in Analytical Chemistry, 2016, 78: 70-83.
14. Meleady P. Two-dimensional gel electrophoresis and 2D-DIGE[J]. Difference gel electrophoresis: methods and protocols, 2018: 3-14.
15. Engen J R, Botzanowski T, Peterle D, et al. Developments in hydrogen/deuterium exchange mass spectrometry[J]. Analytical chemistry, 2020, 93(1): 567-582.
16. W. Bao, B. Yang, Protein acetylation sites with complex-valued polynomial model, Frontiers of Computer Science 18 (3) (2024), p. 183904-null.
17. W. Bao, B. Yang, B. Chen, 2-hydr_ensemble: lysine 2-hydroxyisobutyrylation identification with ensemble method, Chemom. Intel. Lab. Syst. 4 (2021) 104351.
18. Xue Y, Liu Z, Gao X, et al. GPS-SNO: computational prediction of protein S-nitrosylation sites with a modified GPS algorithm[J]. PloS one, 2010, 5(6): e11290.
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[J]. PloS one, 2011, 6(7): e21849.
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[J]. PloS one, 2013, 8(2): e55844.
21. Hao G, Derakhshan B, Shi L, et al. SNOSID, a proteomic method for identification of cysteine S-nitrosylation sites in complex protein mixtures[J]. Proceedings of the National Academy of Sciences, 2006, 103(4): 1012-1017.
22. Jiang P, Ning W, Shi Y, et al. FSL-Kla: A few-shot learning-based multi-feature hybrid system for lactylation site prediction[J]. Computational and structural biotechnology journal, 2021, 19: 4497-4509.
23. H. Lv, F.Y. Dao, H. Lin, DeepKla: an attention mechanism-based deep neural network for protein lysine lactylation site prediction, Imeta 1 (1) (2022) e11.
24. Meng L, Chen X, Cheng K, et al. TransPTM: a transformer-based model for non-histone acetylation site prediction[J]. Briefings in Bioinformatics, 2024, 25(3): bbae219.
25. Pokharel S, Pratyush P, Heinzinger M, et al. Improving protein succinylation sites prediction using embeddings from protein language model[J]. Scientific reports, 2022, 12(1): 16933.
26. Z. Li, M. Li, L. Zhu, W. Zhang, 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) (2024) 188–196.
27. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger,K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler,T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger,M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior,K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature 596 (7873) (2021) 583–589.
28. Y.H. Yang, J.T. Yang, J.F. Liu, Lactylation prediction models based on protein sequence and structural feature fusion, Brief. Bioinform. 25 (2) (2024).
29. Qiu W-R, Sun BQ, Xiao X, et al. iPTM-mLys: identifying multiple lysine PTM sites and their different types. Bioinformatics 2016;32(20):3116–23.

30. Ahmed S, Rahman A, Hasan MAM, et al. predML-site: predicting multiple lysine PTM sites with optimal feature representation and data imbalance minimization. IEEE/ACM Trans Comput BiolBioinform 2022;19(6):3624–34.
31. Hasan MAM, Ahmad S. mLysPTMpred: multiple lysine PTM siteprediction using combination ofSVMwith resolving data imbal-ance issue. Natural Science 2018;10(9):370–84.
32. Ahmed S, Rahman A, Hasan MAM, et al. Computational identification of multiple lysine PTM sites by analyzing theinstance hardness and feature importance. Sci Rep 2021; 11(1):18882.
33. Sua JN, Lim SY, Yulius MH, et al. Incorporating convolutional neural networks and sequence graph transform for identifying multilabel protein lysine PTM sites. Chemom Intel Lab Syst2020;206:104171.
34. Zuo Y, HongY, ZengX, et al. MLysPRED: graph-based multi-view clustering and multi-dimensional normal distribution resampling techniques to predict multiple lysine sites. Brief Bioinform2022;23(5):bbac277.
35. Chen L, Chen Y. RMTLysPTM: Recognizing multiple types of lysine PTM sites by deep analysis on sequences[J]. Briefings in Bioinformatics, 2024, 25(1): bbad450.
36. Chou K C. Prediction of signal peptides using scaled window[J]. peptides, 2001, 22(12): 1973-1979.
37. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences[J]. Bioinformatics, 2006, 22(13): 1658-1659.
38. ElAbd H, Bromberg Y, Hoarfrost A, et al. Amino acid encoding for deep learning applications. BMC Bioinformatics 2020;21:1–14.
39. El Allali A, Elhamraoui Z, Daoud R. Machine learning applications in rna modification sites prediction. Comput Struct Biotechnol J 2021;19:5510–24.
40. Wang D, Liu D, Yuchi J, et al. MusiteDeep: a deep-learning based webserver for protein post-translational modification site prediction and visualization[J]. Nucleic Acids Research, 2020, 48(W1): W140-W146.
41. Li Y, Yang T. Word embedding for understanding natural language: a survey[J]. Guide to big data applications, 2018: 83-104.
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[J]. Journal of proteome research, 2021, 21(1): 265-273.
43. Heinzinger M, Littmann M, Sillitoe I, et al. Contrastive learning on protein embeddings enlightens midnight zone. NAR GenomicsBioinf 2022;4(2):lqac043.
44. Mai Ha V, Akbar R, Robert PA, et al. Linguistically inspired roadmap for building biologically reliable protein languagemodels. Nature. Mach Intell 2023;5(5):485–96.
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[J]. Bioinformatics, 2024, 40(5): btae290.
46. Edgar R C, Batzoglou S. Multiple sequence alignment[J]. Current opinion in structural biology, 2006, 16(3): 368-373.
47. Eddy S R. Hidden markov models[J]. Current opinion in structural biology, 1996, 6(3): 361-365.
48. Zhou J, Lu Y, Dai H N, et al. Sentiment analysis of Chinese microblog based on stacked bidirectional LSTM[J]. IEEE Access, 2019, 7: 38856-38866.
49. Shaw P, Uszkoreit J, Vaswani A. Self-attention with relative position representations[J]. 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[J]. Briefings in bioinformatics, 2022, 23(6): bbac446.
51. Vacic V, Iakoucheva LM, Radivojac P. Two sample logo: a graphical representation of the differences between two sets of sequence alignments. Bioinformatics 2006;22(12):1536–7.
52. Huang L, Lin J, Liu R, et al. Coadti: multi-modal co-attention based framework for drug target interaction annotation. BriefBioinform 2022;23(6):bbac446.
53. Van der Maaten L, Hinton G. Visualizing data using t-sne. JMachLearn Res 2008;9(11).