JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

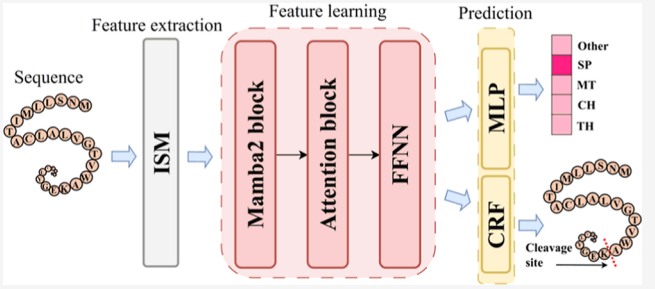
DeepMaT: Prediction of Target Peptide Classification and Cleavage Site by Combining Mamba2 and Multiple Attention Mechanisms

Qianmao Wen, Aoyun Geng, Junlin Xu, Yajie Meng, Leyi Wei, Zilong Zhang, Quan Zou, and Feifei Cui*

Cite This: J. Chem. Inf. Model. 2025, 65, 10011−10024 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Signal peptides and transit peptides are essential for directing mature proteins to their proper cellular locations, particularly through cleavage following transport. Although various prediction tools achieve strong performance in identifying and classifying targeting peptides, their accuracy in determining cleavage sites remains limited. We introduce DeepMaT, a deep learning model that integrates Mamba2 and a multihead self-attention mechanism, leveraging the global modeling capabilities of Mamba2 and the localized focus of self-attention. Experimental results show that DeepMaT significantly outperforms state-of-the-art models in cleavage site prediction, achieving an accuracy of 0.867 for thylakoid transit peptides and also performing well on other peptides. Moreover, DeepMaT can accurately learn the amino acid distribution of real samples. Ablation experiments show that the combination of Mamba and attention mechanisms can improve model efficiency, further proving the effectiveness of the combination. It also enables prediction of targeting peptides with unspecified cleavage sites, offering a valuable tool for protein database annotation. DeepMaT is freely available on GitHub at https://github.com/qianmao2001/DeepMaT.



1. INTRODUCTION

Proteins are synthesized within cells and guided to specific regions by sorting signals, a process known as subcellular localization, which is critical to protein function.[1−4] These sorting signals vary in both length and destination. Among them, N-terminal targeting peptides[5] are common and direct proteins to locations such as the secretory pathway, plasma membrane, mitochondria, chloroplasts, and endosomes.[6] In the secretory pathway and plasma membrane translocation, the main targeting signal is the signal peptide (SP), which guides proteins to the secretory system in eukaryotes and to the plasma membrane in prokaryotes.[7] Mitochondrial transit peptides (MT) possess an amphipathic α-helix structure and transport proteins through the TOM and TIM complexes of the outer and inner mitochondrial membranes, respectively.[8] For chloroplast targeting, the chloroplast transit peptide (CH) delivers proteins through the TOC and TIC complexes into the chloroplast stroma,[9] where it is cleaved by the stromal processing peptidase (SPP). If the protein contains a thylakoid transit peptide (TH), it is subsequently directed to the thylakoid membrane or lumen. These targeting peptides are typically excised by specific enzymes after successful delivery, enabling the mature protein to function in its destined compartment. Notably, Markus Kunze and Berger reported that mitochondrial, chloroplast, and thylakoid targeting peptides share structural similarities, potentially resulting in dual localization of some proteins.[10]

Various molecular and cellular biology techniques are commonly employed to identify whether a protein contains a targeting peptide—such as signal peptide, mitochondrial transit peptide, or chloroplast transit peptide—and corresponding cleavage site. A widely used approach is the fluorescence fusion localization assay,[11−13] in which a green fluorescent protein is fused to the protein of interest. Subcellular localization and function are then inferred by observing fluorescent signal aggregation in live cells. For cleavage site determination, Edman degradation,[14] developed by Pehr Edman, is typically used. This method sequentially removes N-terminal amino acids to determine the mature protein's starting position but requires highly purified samples and involves labor-intensive procedures. Overall, experimental identification of targeting peptides is complex and may fail after multiple attempts.

To reduce experimental costs and improve efficiency, predictive tools based on artificial intelligence techniques are increasingly being adopted. With advances in artificial intelligence, numerous prediction tools for protein have been

Received: June 28, 2025
Revised: September 21, 2025
Accepted: September 23, 2025
Published: September 26, 2025

https://doi.org/10.1021/acs.jcim.5c01489
J. Chem. Inf. Model. 2025, 65, 10011−10024

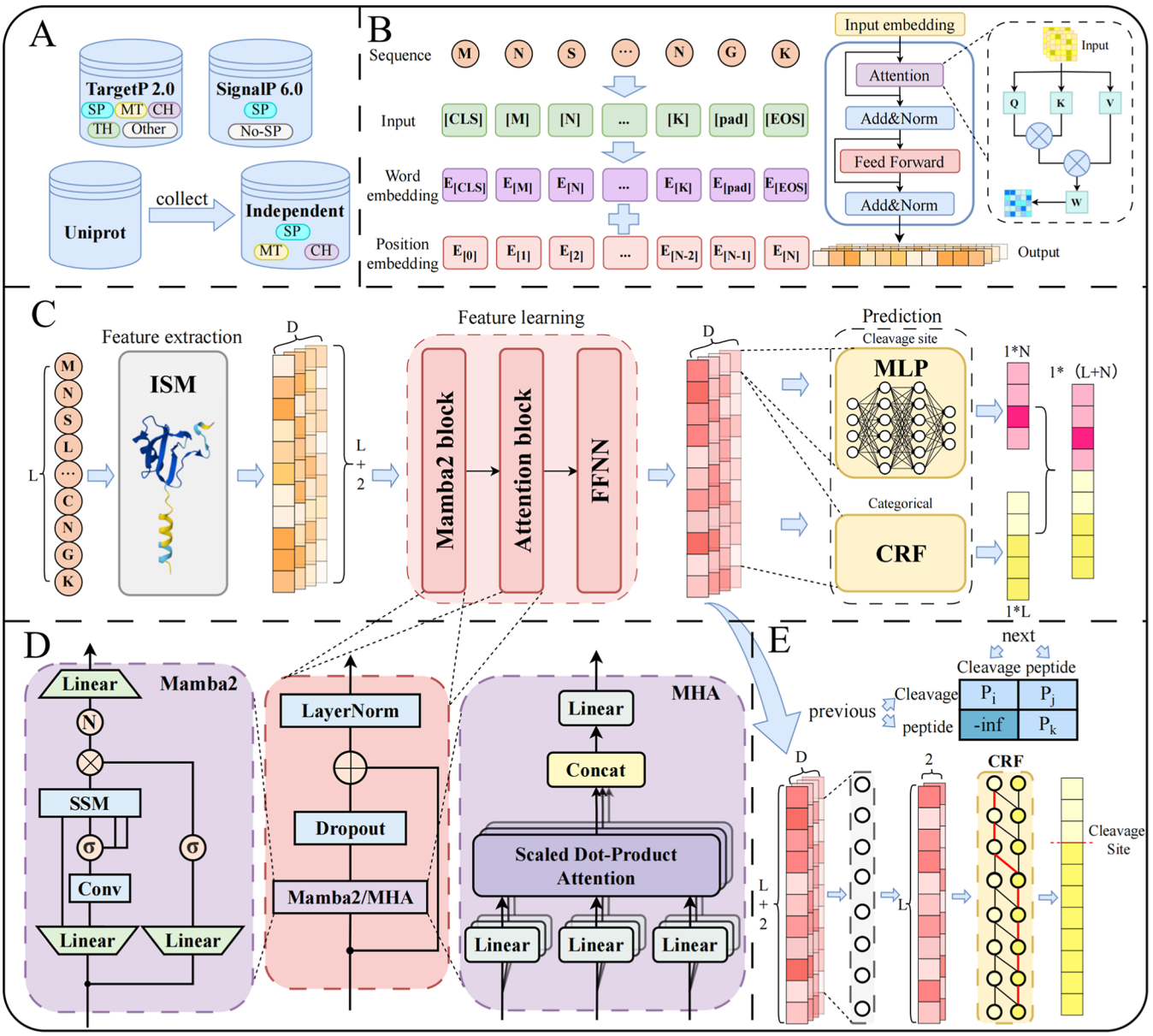


Figure 1. Overview of the DeepMaT model framework. (A) Data collection: data sets were obtained from TargetP 2.0, comprising SP, MT, CH, TH, and other peptides, as well as from SignalP 6.0, which includes various signal peptide isoforms. Independent test sets from UniProt were also collected to evaluate model generalization. (B) ISM feature extraction: Raw sequences are tokenized into Input IDs, Word Embeddings, and Position Embeddings, as well as <CLS>, < pad>, and <EOS> tokens, then processed by the ISM model to generate sequence embeddings. (C) DeepMaT architecture: consists of feature extraction, feature learning, and prediction modules. Feature extraction uses ISM to compute L × 1280-dimensional peptide representations, where L is sequence length. Feature learning integrates Mamba2, MHA, and a feed-forward network to capture global and local patterns. Prediction employs a Conditional Random Field (CRF) for cleavage site identification and a Multi-Layer Perceptron (MLP) for peptide classification. (D) Structure of the Mamba2 and attention block. (E) Prediction module: The CRF layer initializes a transition probability matrix and is applied for cleavage site prediction.

developed,[15] such as SBSM-Pro.[16] MitoFates,[17] by Fukasawa et al., uses support vector machines (SVM) to classify mitochondrial transit peptides and predict cleavage sites. DeepMito,[18] developed by Savojardo et al., employs convolutional neural networks to predict mitochondrial protein localization. ChloroP,[19] by Emanuelsson et al., utilizes neural networks to predict chloroplast transit peptides and cleavage sites. Westerlund et al. created LumenP,[20] also based on neural networks, to predict vesicle-like protein localization and cleavage sites, facilitating taxonomic identification. Savojardo et al. further developed TPpred3,[21] which combines neural networks and other architectures for classification and cleavage site prediction of mitochondrial and chloroplast transit peptides. DeepLoc 1.0[22] uses a three-stage deep learning approach to predict subcellular localization. Subsequent DeepLoc 2.0[23] uses a protein language model for 10-type subcellular localization. For signal peptides, the well-known SignalP family includes the latest SignalP 6.0[24] by Teufel et al., which predicts five types of signal peptides and their subregions (N-terminal n-region, hydrophobic h-region, and C-terminal c-region) underlying SP function. Subsequent tools such as PEFT-SP[25] and USPNet[26] predict classification and cleavage sites but not subregions. DeepSig[27] is one of the first models to replace HMM/SVM methods with deep learning.

https://doi.org/10.1021/acs.jcim.5c01489
J. Chem. Inf. Model. 2025, 65, 10011−10024

TSignal[28] is a multitask model that uses the ProtT5 model to extract amino acid sequence features. Most models predict only one or two transit peptides, lacking simultaneous prediction of all four targeting peptides. TargetP 1.1[29] by Emanuelsson et al. was an early model able to predict multiple targeting peptides and cleavage sites simultaneously. Its successor, TargetP 2.0[5] by Armenteros et al., leverages deep learning methods including LSTM and multihead attention to improve classification and cleavage site prediction. However, for certain transit peptides, TargetP 2.0s cleavage site accuracy remains low (∼0.5), possibly because its attention mechanism focuses on short-range residue features, limiting its ability to capture long-range dependencies.

Previous methods often converted targeting peptide sequences directly into digital representations for prediction without extracting informative features,[30] leading to suboptimal cleavage site identification, as observed in TargetP 2.0. To address this, we propose DeepMaT (Figure 1), which integrates Mamba2,[31] a multihead self-attention mechanism[32] (MHA), and Implicit Structural Model (ISM) -extracted sequence features[33] to enhance target peptide prediction accuracy. On the TargetP 2.0 data set, DeepMaT outperforms TargetP 2.0 across most metrics, notably improving classification and cleavage site prediction for the underrepresented thylakoid peptides. Analysis of classification heatmaps, Sankey diagrams, and *t*-SNE visualizations reveals that chloroplast and thylakoid transit peptides are often confused, an issue TargetP 2.0 struggles to resolve. For signal peptides, DeepMaT achieves performance comparable to TargetP 2.0. To further evaluate generalizability, we tested DeepMaT on the SignalP 6.0 data set, where it surpasses state-of-the-art methods on multiple metrics. The main contributions of this study are summarized as follows:

1 developed a multitasking model based on ISM embeddings capable of performing both targeted peptide multiclassification and cleavage site prediction;
2 collected and curated data sets of peptides with unspecified cleavage sites from the UniProt database, annotated using DeepMaT to assist further research;
3 applied Mamba2 to sequential tasks, demonstrating its effectiveness in this context;
4 experimentally validated that DeepMaT achieves 0.8 accuracy in classifying thylakoid transit peptides despite limited samples, and 0.867 accuracy in cleavage site prediction.

2. MATERIALS AND METHODS

DeepMaT (Figure 1) addresses two deep learning tasks: sequence classification and cleavage site prediction. Both tasks are performed simultaneously using distinct approaches. The model inputs ISM-extracted features, which are first processed by the Mamba layer to learn and integrate global features while capturing local dependencies. These global features are then passed to MHA layer, enhancing the model's capacity to learn diverse local patterns. A feedforward layer follows to increase model complexity and mitigate overfitting. Finally, CRF predicts cleavage sites, while MLP performs classification.

2.1. Data Set Selection and Collection. In this study, we utilized the TargetP 2.0 data set (Table 1) and the SignalP 6.0 data set (Table 2) for model training and evaluation. The TargetP 2.0 data set,[5] compiled by J.J. Almagro Armenteros et al., was derived from the UniProt database[34] (version 2018_04) and contains 2697 signal peptides, 499 mitochondrial transit peptides, 227 chloroplast transit peptides, and 45 thylakoid transit peptides, each annotated with cleavage sites. Additionally, it includes 9537 nontargeting peptides. The SignalP 6.0 data set[24] is an extension of SignalP 5.0,[35] which originally contained three types of signal peptides—Sec/SPI, Sec/SPII, and Tat/SPI. The expanded SignalP 6.0 data set incorporates annotations for multiple signal peptide isoforms, including Sec/SPI, Sec/SPII, Sec/SPIII, Tat/SPI, and Tat/SPII, along with their corresponding cleavage site annotations. Both data sets exhibit significant class imbalance, with some peptide classes represented by substantially more samples than others. To facilitate efficient batch training, sequence lengths were standardized to 200 prior to tokenization: sequences exceeding the fixed length were truncated, while shorter sequences were padded to ensure uniformity across all input data.

Table 1. Number of Samples for Each Peptide Category in the TargetP 2.0 Dataset

type	SP	MT	CH	TH	other
number	2697	499	227	45	9537

Given that the TargetP 2.0 data set is relatively outdated and that recent years have seen significant advancements in cleavage site research, we curated two independent test sets from the UniProt database[36] to further validate our model. A query was conducted in UniProt using the following search string: (name: (date_created: [2018-05-01 TO 2025-03-31]) reviewed: true), where the name field includes terms such as "Signal peptide", "thylakoid transit peptide", "Mitochondrial transit peptides" and "Chloroplast transit peptide". The search was restricted to manually reviewed entries created after 2018-05-01, corresponding to the final update of the TargetP 2.0 data set. Based on the presence of cleavage site annotations, we grouped sequences into two sets: those with clearly defined cleavage sites were assigned to Independent Data set 1, while those lacking such annotations were placed in Independent Data set 2. Sample counts are detailed in Table 3. We then compared the performance of the proposed DeepMaT model against TargetP 2.0 on these two independent test sets to evaluate DeepMaT's generalization ability and robustness on newly collected data. To further demonstrate the performance of DeepMaT, we also collected and organized a balanced positive and negative data set, Independent Data set 3. We searched the UniProt database for "cytosolic proteins", selecting sequences with reviewed and Annotation score of 5 and sequence lengths of 1−200, totaling 536 sequences. To balance the data set, we randomly selected 536 target peptides from Independent Data set 1, including 373 signal peptides, 120 mitochondrial transit peptides, and 43 chloroplast transit peptides.

2.2. DeepMaT Model Building. *2.2.1. Feature Extraction of Peptide Sequences.* DeepMaT employs ISM to extract features from amino acid sequences. ISM is a self-supervised learning framework that encodes local structural information into a sequence model, producing embeddings that are both evolutionarily conserved and structurally informed. ISM includes a specialized tokenizer that processes amino acid sequences by converting them into a specific encoded format. This tokenizer adds start and end identifiers to the sequence, resulting in a final length of L. ISM then outputs a feature array of size $(L + 2) \times 1{,}280$, denoted as H_{ISM}, representing the

https://doi.org/10.1021/acs.jcim.5c01489
J. Chem. Inf. Model. 2025, 65, 10011−10024

Table 2. Number of Signal Peptide Isoforms Across Different Species in the SignalP 6.0 Dataset and Their Totals

organism	Sec/SPI	Sec/SPII	Sec/SPIII	Tat/SPI	Tat/SPII	TM/globular (NO-SP)	total
Eukaryotes	2040					14,356	16,396
Gram-positive	142	516	4	39	8	226	935
Gram-negative	356	1087	56	313	19	933	2764
Archaea	44	12	10	13	6	110	195

Table 3. Number of Samples for Each Peptide Category in the Three Independent Test Sets

Type	independent data set 1	independent data set 2	independent data set 3
SP	721	2	373
MT	43	3	43
CH	232	23	120
other			536

structurally enriched sequence embeddings, including the start symbol "[CLS]", the end symbol "[EOS]", and the fill symbol "[pad]".

$$H_{\mathrm{ISM}} = \mathrm{ISM}(X_{\mathrm{seq}}) \tag{1}$$

where X_{seq} denotes the original amino acid sequence.

2.2.2. Feature Learning Module. The DeepMaT model utilizes Mamba2 and MHA to learn ISM features. It then performs cleavage site prediction and classification using CRF and MLP, respectively.

Mamba2, an advanced version of the original Mamba, leverages the Structured State Space Duality algorithm to efficiently process sequential data. By internally employing numerous torch.einsum operations, Mamba2 enables highly efficient matrix computations. It also incorporates algorithmic improvements to support data-parallel computation, significantly accelerating data processing, as illustrated in Figure 1D. Mamba has been widely adopted in sequence modeling tasks, including ChiMamba,[37] PTM-Mamba,[38] and Caduceus.[39]

The ISM-derived features are directly input into the Mamba layer, which models global dependencies and outputs a feature array consistent with the dimensions of H_{ISM}, denoted as M

$$M = \mathrm{LayerNorm}(H_{\mathrm{ISM}} + \mathrm{Dropout}(\mathrm{Mamba2}(H_{\mathrm{ISM}}))) \tag{2}$$

MHA is an enhanced attention mechanism that builds upon the traditional attention framework. It employs multiple attention "heads", each of which independently learns and processes input features. This parallel processing enables the model to capture a richer set of representations and improves its learning capacity and expressiveness.

In MHA, the input features are projected into three distinct vectors: query (Q), key (K), and value (V), using corresponding weight matrices W_q, W_k, and W_v. These vectors are used to compute the attention scores, which are then normalized via a Softmax function to produce attention weights.

$$Q = W_q X;\ K = W_k X;\ V = W_v X \tag{3}$$

$$\text{Self-Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{4}$$

The output of MHA layer is a refined representation of the input features, obtained after parallel attention heads have independently learned and weighted different aspects of the data. To enhance model stability and prevent overfitting, MHA layer is followed by Layer Normalization (LayerNorm) and Dropout. The output array from this layer maintains the same dimensionality as the output of the Mamba layer and is denoted as A.

$$A = \mathrm{LayerNorm}(M + \mathrm{Dropout}(\text{Multi head Attention}(M))) \tag{5}$$

2.2.3. Processing Features for Prediction after Learning. CRF[40] are widely used in sequence labeling tasks, particularly those involving multilabel annotation, due to their ability to capture complex dependencies between output labels and input features. In the DeepMaT model, the CRF module operates on the sequence of states $y = y_1 \ldots y_t$ produced after MHA layer, where y is the 1−200 tokens in $A(A_1 - A_{200})$. It assigns a corresponding hidden state sequence $h = h_1 \ldots h_t$ to each position in the input, effectively modeling the structured relationships within the output sequence

$$P(y|h) = \frac{1}{Z(h)} \prod_{t=1}^{T} \exp(\psi(h_t)) \prod_{t=1}^{T-1} \exp(\varphi_{y_t, y_{t+1}}) \tag{6}$$

where Z(h) is the modeled normalization constant; φ is the learning transfer matrix of the CRF with shape C*C and C is the number of modeled labels; T is the length L of the sequence; and ψ is the learnable linear transformation from the hidden state h to the labels

$$\psi(h_t) = W_\psi h_t + b_\psi \tag{7}$$

MLP operates on the first token of the sequence output by MHA layer. Specifically, it takes the corresponding 1280-dimensional feature vector and maps it to an n-dimensional output, where n represents the number of peptide classes. A softmax function is then applied to the resulting vector to compute the classification probabilities.

$$P_{\mathrm{class}} = \mathrm{softmax}(\mathrm{MLP}(A_0)) \tag{8}$$

2.3. Model Training. To evaluate the robustness of the model and ensure that every sample was included in the evaluation process, TargetP 2.0 data set was evenly divided into five nonoverlapping subsets. Specifically, TargetP 2.0 data set was divided sequentially into five subsets: the first 0−20% as the first fold, 20−40% as the second fold, and so on, resulting in five equal folds. In each iteration, four subsets were used for training and the remaining one for testing. However, the SignalP 6.0 data set is divided according to the sample labels provided by the authors. Specifically, the authors of SignalP 6.0 assigned each sample a subset label, which has three categories, indicating that the data set is divided into three subsets. In each round of training, two subsets are used for training and one for testing. The final performance report is an average. DeepMaT simultaneously performs peptide classification and cleavage site prediction, each requiring a separate objective during optimization. For the classification

https://doi.org/10.1021/acs.jcim.5c01489
J. Chem. Inf. Model. 2025, 65, 10011−10024

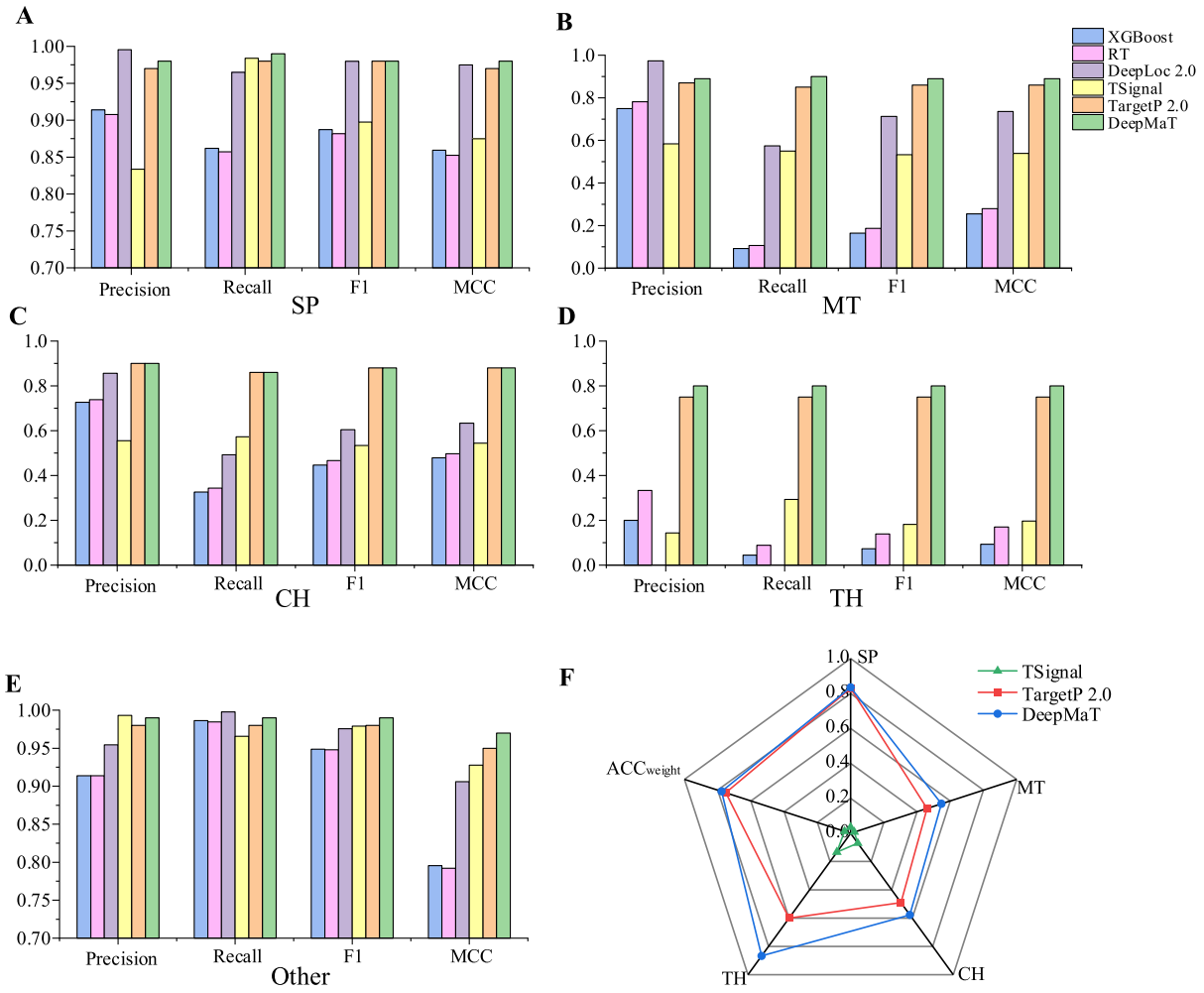


Figure 2. Performance of DeepMaT on classification and cleavage site prediction tasks for targeting peptides. (A–E) These five bar graphs present classification metrics for both targeting and nontargeting peptides, including SP, MT, CH, TH, and other peptide types. (F) The radar map illustrate the accuracy of cleavage site prediction across different targeting peptides.

task, we used the cross-entropy loss function,[41] which quantifies the divergence between predicted class probabilities and ground truth labels, thereby enhancing the model's discrimination across peptide types. For the cleavage site prediction task, we applied the negative log-likelihood loss derived from the CRF, which effectively models label dependencies across sequences and improves cleavage site precision. The final training objective is a weighted sum of both loss functions, where tunable hyperparameters are introduced to balance performance between classification and sequence labeling tasks.

$$-\log(P(y|h)) = \log(Z(h)) - \log\left(\exp\left(\sum_{t=1}^{T}\sum_{y_i\in M_t}\psi(\mathbf{h_t}) + \varphi(y_t, y_{t-1})\right)\right) \quad (9)$$

$$\mathrm{Loss_{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} p_i^k \log(\hat{p}_i^k) \quad (10)$$

$$\mathrm{Loss} = w_1 \times \mathrm{Loss_{CRF}} + w_2 \times \mathrm{Loss_{CE}} \quad (11)$$

where T is the sequence length of the samples, N is the number of samples in a batch, K is the number of categories, p_i is the true label of the samples, and $\hat{p}_i$ is the predicted probability value.

2.4. Model Performance Evaluation. In this study, we evaluated model performance using four standard metrics:[42] Precision, Recall, F1 Score, Matthews Correlation Coefficient (MCC),[43] Accuracy (ACC), and Weighted ACC ($\mathrm{ACC_{weight}}$). Among these, recall reflects the model's sensitivity—its ability to correctly identify true positive samples.

For SignalP 6.0, MCC is further subdivided into MCC1 and MCC2. MCC1 treats one specific signal peptide class as the positive class and all nonsignal peptides as negatives. In contrast, MCC2 considers the target class as positive while treating all other signal peptide classes and nonsignal peptides as negatives. While MCC1 evaluates the model's ability to distinguish one class from nonsignal peptides, MCC2 is more appropriate for assessing performance in single-class detection under multiclass classification settings. We follow this evaluation scheme to ensure consistency and comparability across models.

For the categorize prediction task, we primarily use Precision, Recall, F1 Score, and MCC.

https://doi.org/10.1021/acs.jcim.5c01489
J. Chem. Inf. Model. 2025, 65, 10011−10024

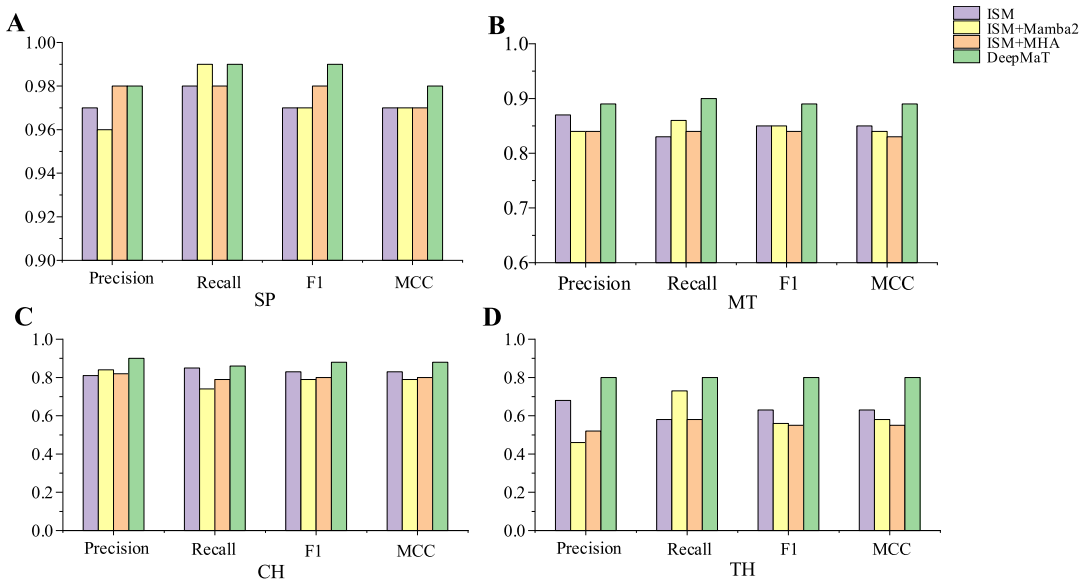


Figure 3. Comparison of ablation experiment results. (A–D) Show the performance metrics of the ablation experiment on signal peptides (SP), mitochondrial transit peptides (MT), chloroplast transit peptides (CH), thylakoid transit peptides (TH).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{12}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{13}$$

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{14}$$

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

where TP denotes True Positive, TN denotes True Negative, FP denotes False Positive, and FN denotes False Negative.

For the cleavage site prediction task, we primarily use ACC, and $\text{ACC}_{\text{weight}}$. We determine that the cleavage site prediction is accurate only when the string of sequence labels predicted by CRF is completely consistent with the actual labels we defined.

$$\text{ACC} = \frac{\sum_{t=1}^{N} \sum_{y_i \in Y_t} F(\hat{y}_i = y_i)}{N} \tag{16}$$

$$\text{ACC}_{\text{weight}} = \sum_{c=1}^{C} \frac{N_c}{N} \text{ACC}_c \tag{17}$$

where N is the sample quantity, Y is the label of a sequence, F is used to calculate whether the labels are consistent, outputting 1 if consistent and 0 if not, and C is the number of categories.

3. RESULTS AND DISCUSSION

3.1. Evaluation of DeepMaT for Targeted Peptide Recognition and Cleavage Site Prediction on the TargetP 2.0 Data Set. DeepMaT demonstrated strong performance on the TargetP 2.0 data set (Figure 2). In the classification tasks for signal peptides, mitochondrial transit peptides, chloroplast transit peptides, and thylakoid transit peptides, DeepMaT consistently outperformed TargetP 2.0 across key evaluation metrics, including Accuracy, F1 Score, and Recall. While TargetP 2.0 utilizes an attention mechanism that has limited capacity for modeling long-range dependencies in sequences, DeepMaT leverages the Mamba2 architecture—a state space model (SSM)[31]—to enhance global sequence modeling and extract more informative features from longer input sequences. As shown in Figure 2A–E, DeepMaT surpasses other models such as TargetP 2.0 in most metrics. Although chloroplast and thylakoid transit peptides share certain sequence characteristics,[9] DeepMaT improves classification performance for thylakoid peptides without compromising the accuracy of chloroplast peptide predictions. Notably, the recall for mitochondrial transit peptides increased from 0.85 to 0.90—a 5-point improvement. Figure 2E further illustrates DeepMaT's enhanced capability in distinguishing both targeted and nontargeted peptides. This improvement is also evident in the clustering results shown in Figure 9. DeepMaT achieves clear separation between targeted and nontargeted peptides on the t-SNE plot, with well-formed clusters and significant intergroup distance.

In addition to its superior classification performance, the model also significantly improves cleavage site prediction accuracy for MT and TH (Figure 2F). We performed a chi-squared test on the cleavage site prediction results from DeepMaT and TargetP 2.0. Although the p-values did not fall below 0.05 for SP (p = 0.6354) and CH (p = 0.0599), indicating no significant improvement, the p-values for MT (p = 0.0071) and TH (p = 0.0042) were significantly less than 0.05, demonstrating a significant improvement. Notably, the accuracy of cleavage site prediction for thylakoid transit peptides increased from 0.60 in TargetP 2.0 to 0.867—a substantial improvement. Importantly, this gain in prediction accuracy is achieved without sacrificing classification performance, highlighting the robustness and effectiveness of DeepMaT in multitask learning.

On the TargetP 2.0 data set, we implemented XGBoost and RT two machine learning methods. From their classification results (Figure 2), it appears that this task cannot be effectively predicted by machine learning, and more complex deep

https://doi.org/10.1021/acs.jcim.5c01489
J. Chem. Inf. Model. 2025, 65, 10011–10024

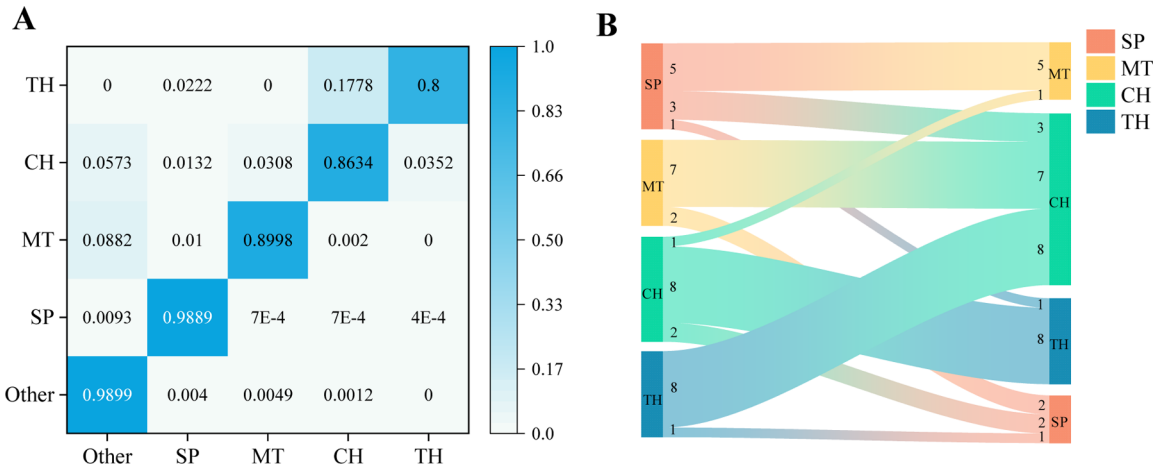


Figure 4. Analysis of misclassification. (A) Heatmap display prediction results, each row is the real label, and each column is the predicted label; (B) Putting the four categories of targeted peptide misclassification through the Sankey diagram display.

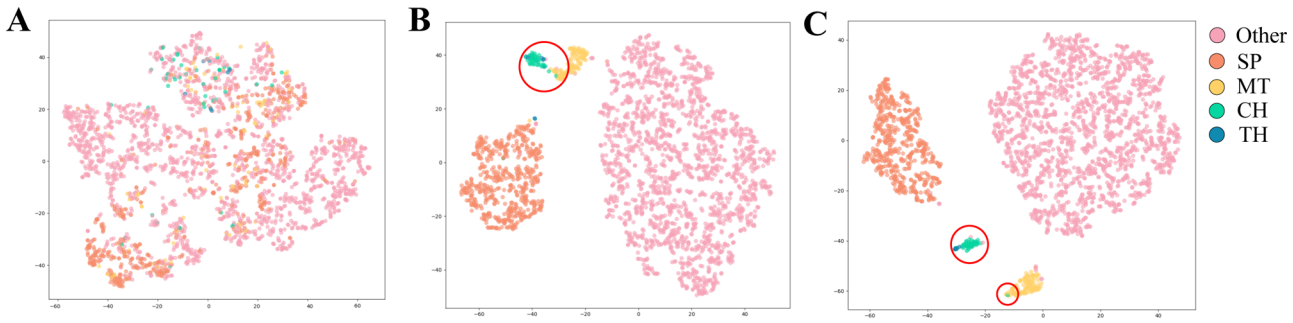


Figure 5. tSNE visualization. (A) Sequence features extracted by ISM; (B) Feature data after performing learning; (C) Data after MLP prediction.

learning is still needed. We also replicated and compared Deeploc 2.0 and TSignal. Deeploc 2.0, as a tool specifically designed for subcellular prediction, performs well for multiclass tasks. On the TargetP 2.0 data set, Deeploc 2.0 shows poor performance for CH, which is attributed to the imbalance in the data set. TSignal, as a multitask model, can identify multiple signal peptides and recognize cleavage sites simultaneously. However, due to the lack of more detailed peptide labels in the data set, the model performs poorly in this task.

3.2. Ablation Experiments. We conducted ablation experiments on the feature learning module of DeepMaT by retraining ablated versions of the model on the TargetP 2.0 data set. Specifically, we examined the effects of removing Mamba2,[31] MHA,[32] and both components simultaneously. As shown in Figure 3 and Supporting Information Table S1, the removal of either Mamba2 or MHA individually resulted in significant performance degradation, particularly for thylakoid peptides, which are underrepresented in the data set. Performance also declined across other peptide categories to varying degrees. Interestingly, when both components were removed simultaneously, performance showed partial recovery but remained inferior to the full DeepMaT model. The phenomenon is due to the fact that Mamba2 alone is only able to learn long-range dependent features, whereas MHA alone focuses attention on localized features, which results in the use of the two alone leading to an inability to learn comprehensive features to facilitate the model's prediction of classifications and prediction of cleavage sites. These results suggest that Mamba2 and MHA provide complementary benefits in capturing global and local dependencies, allowing DeepMaT to learn richer and more informative features.

3.3. DeepMaT Utilizes Different Coding to Represent Peptides. Among the encoding approaches, we experimented with four approaches with different degrees, which include One-hot, Blosum62, ESM-2, and ISM.[44,45] One-hot encoding is able to represent peptide sequences in a simple and direct way. Blosum62 takes into account amino acid conservatism and substitution possibilities. ESM-2 is a protein language model utilizing deep learning, which can further represent the peptides, including peptide evolutionary information and sequence information. Experimental results (Figure S1) show that ISM encoding enables DeepMaT to achieve better performance metrics. In classification (Figure S1B–D), One-hot and Blosum62 can hardly predict accurately on MT, CH, and TH with fewer samples, probably due to the fact that these two encodings are only a single representation of amino acids, which cannot characterize sequence information. For the CH prediction task (Figure S1D), ESM-2 cannot make effective recognition, probably because ESM-2 has too much sequence similarity for TH and CH to distinguish the two from the structure. In the cleavage site task (Figure S1F), ISM is more effective than ESM-2, ISM introduces local structural information encoded into the sequence model and is able to characterize the structural information on the peptide, which allows the model to learn the structural information on the peptide and thus make an effective recognition of the cleavage site.

3.4. Analysis of DeepMaT's Misclassification on the TargetP 2.0 Data Set. We used a heatmap (Figure 4A) to

https://doi.org/10.1021/acs.jcim.5c01489
J. Chem. Inf. Model. 2025, 65, 10011−10024

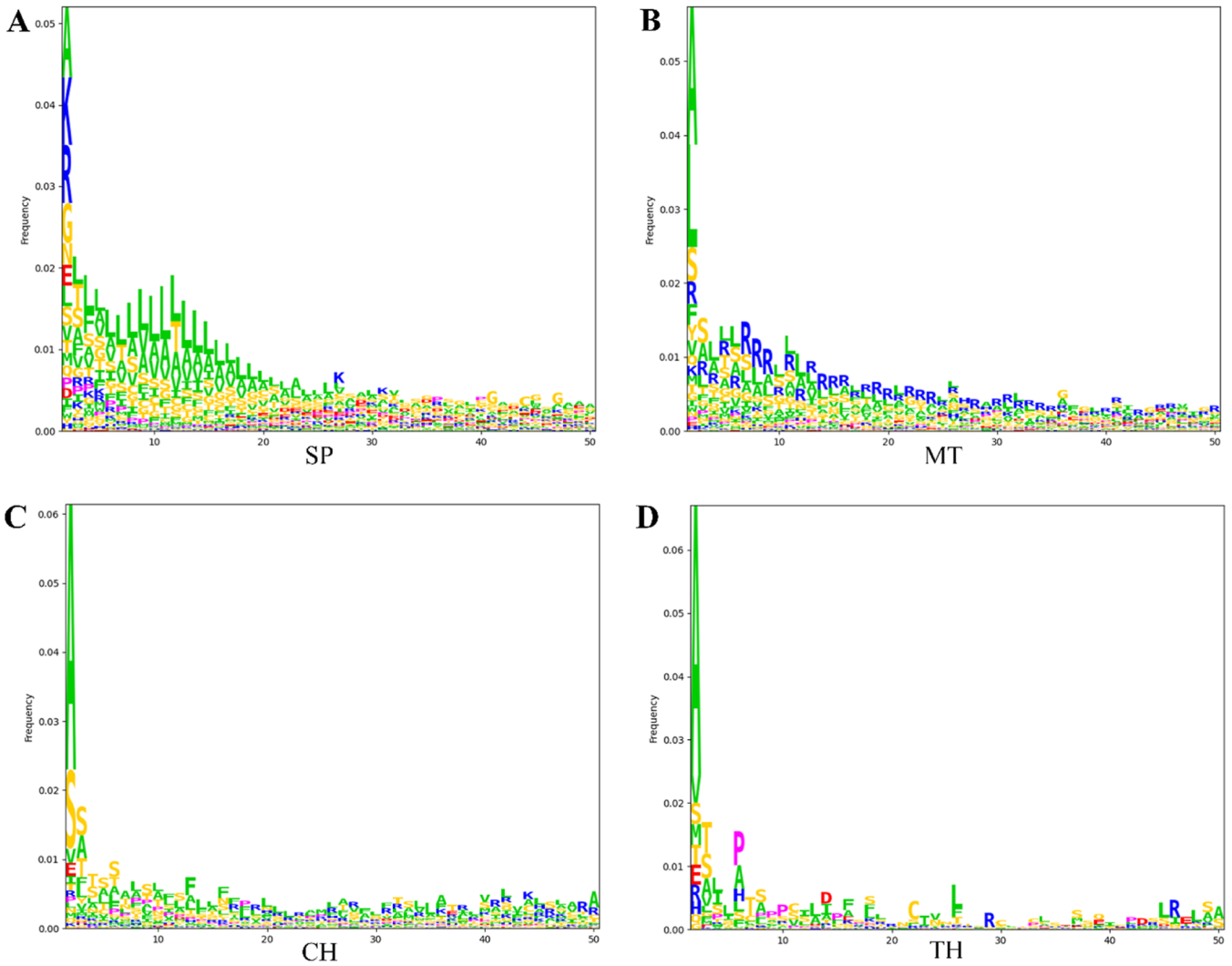


Figure 6. LOGOs for attentional layers showing the attentional weights of amino acids. (A) Visualization of SP's attention weights. (B) Visualization of MT's attention weights. (C) Visualization of CH's attention weights. (D) Visualization of TH's attention weights.

clearly illustrate the confusion patterns in the classification of the four transit peptide classes: signal peptides, mitochondrial, chloroplast, and thylakoid transit peptides. The bright diagonal blocks indicate that the model achieves high recall and specificity for most categories—for example, 98.9% recall for signal peptides. Most misclassified signal and transit peptides were incorrectly predicted as nontargeting peptides. This can be attributed to the class imbalance in the data set, where nontargeting peptides constitute the majority of samples. As a result, the model tends to favor predictions toward this dominant class, leading to a classification bias.[46]

To further examine the misclassification behavior, we visualized the misclassified sequences using a Sankey diagram (Figure 4B). Notably, chloroplast and thylakoid transit peptides are often misclassified as each other. This is likely due to the shared structural motif of N-terminal amphipathic α-helices present in both classes,[47,48] suggesting that the similarity in amphiphilic domains contributes to the classification ambiguity between them.

The *t*-SNE visualizations (Figure 5) demonstrate how the feature distributions of the four transit peptide types and nontargeting peptides become progressively more distinguishable as the model learns. In Figure 5A, the features extracted by ISM appear disorganized, with substantial overlap between classes, indicating poor separability. In Figure 5B, which shows the output from the feature learning module, the clustering has significantly improved compared to the ISM features. However, there remains some overlap among three peptide types within the red-circled region. Finally, Figure 5C, representing the output of the MLP layer, reveals a clearly defined clustering structure. Except for a degree of overlap between thylakoid and chloroplast peptides, most categories are well-separated, indicating strong feature disentanglement. The observed overlap between thylakoid and chloroplast transit peptides is likely due to shared sequence features,[9] which contribute to their similarity in the learned representation space.

3.5. Model Interpretability Evaluation Based on Attention to Amino Acid Sequences. To further investigate how our model attends to specific regions in the sequence data, we utilized LogoMaker[49] to visualize the self-attention weights. As shown in the LOGO plots (Figure 6), the second most attended amino acid—alanine—closely matches its frequency in the actual sequences (Figure 7), indicating that the model effectively learns the true amino acid distribution. For signal peptides, the model consistently focuses on stretches

https://doi.org/10.1021/acs.jcim.5c01489
J. Chem. Inf. Model. 2025, 65, 10011−10024

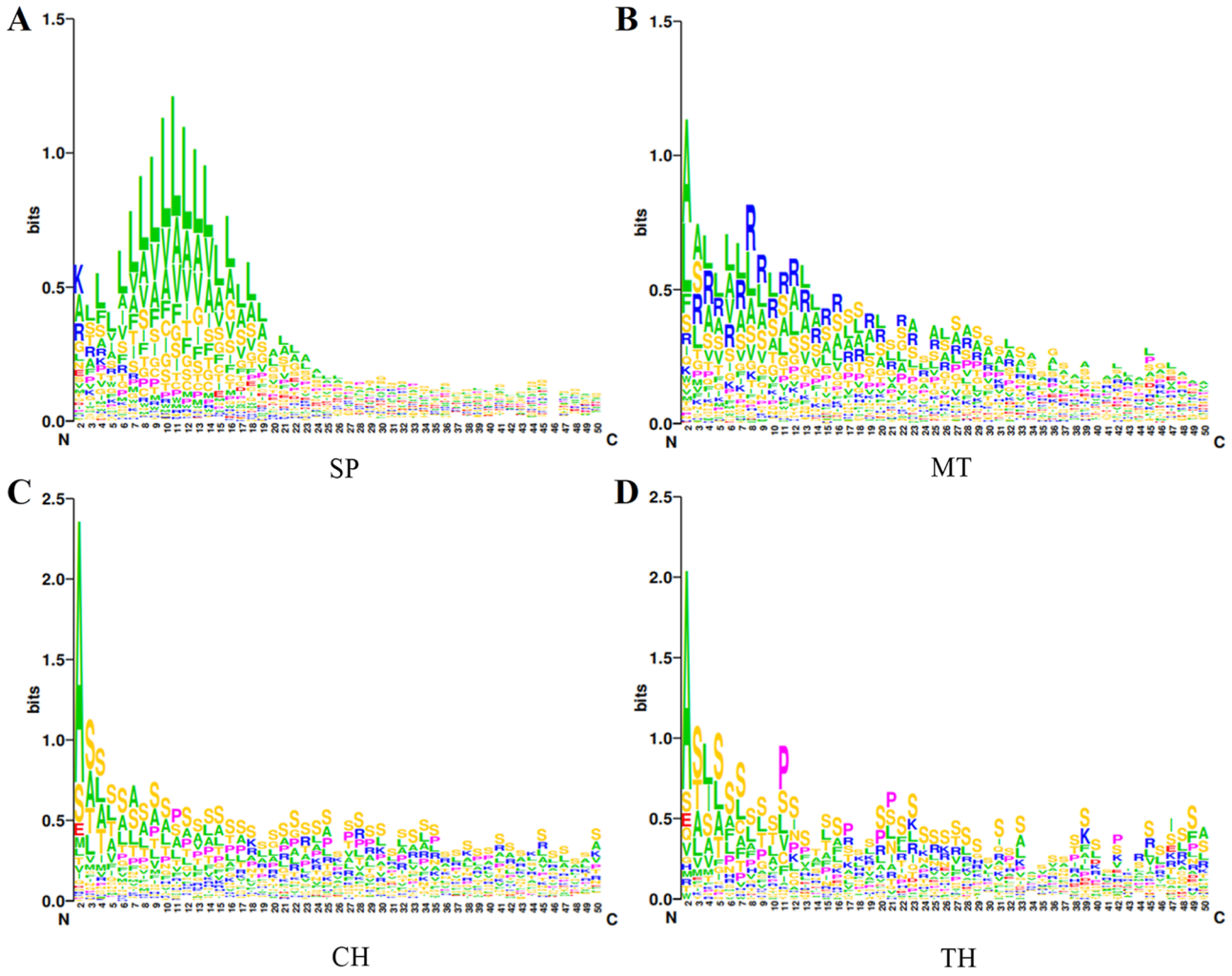


Figure 7. LOGOs for sequences showing true amino acid weights. (A) Visualization of SP's true amino acid weight. (B) Visualization of MT's true amino acid weight. (C) Visualization of CH's true amino acid weight. (D) Visualization of TH's true amino acid weight.

of consecutive leucine residues. These leucine-rich sequences are known to form stable α-helical structures, which enhance hydrophobic interactions with the lipid bilayer, thereby improving the efficiency of transmembrane translocation.[50]

In the case of mitochondrial transit peptides, after excluding the second alanine, our model shifts its attention to subsequent arginine residues. Notably, the model captures a recurring pattern of alternating arginine and hydrophobic residues—an arrangement capable of forming amphipathic α-helices.[51] Voisine et al.[52] reported that such amphipathic structures not only facilitate binding to the mitochondrial membrane but also prevent precursor protein misfolding by interacting with the molecular chaperone Hsp70 via their hydrophobic surfaces.

3.6. Comparison with SOTA Models on the SignalP 6.0 Data Set to Validate Signal Peptide Recognition Performance. In addition to evaluating performance on targeted peptides, we also assessed the generalization capability of our model in the task of signal peptide prediction. To this end, we compared DeepMaT with three established models—SignalP 6.0,[24] PEFT-SP,[25] and USPNet[26]—all of which can predict both signal peptides and their cleavage sites. The results for these baseline models were taken from the Supporting Information of their respective publications. For a fair comparison, we retrained DeepMaT using the same data set split employed by SignalP 6.0. This data set contains 20,290 samples spanning four species and five signal peptide types, and is characterized by a significant class imbalance (as shown in Table 2). PEFT-SP is a fine-tuned model based on ESM-2[53] that incorporates the LoRA[54,55] mechanism, while USPNet combines a BiLSTM module with a protein language model for signal peptide prediction.

As shown in Figure 8A,B, DeepMaT achieves improved classification performance on certain signal peptide types—for example, the Archaea Sec/SPI class—where it outperforms others by approximately 3% to 12%. Although it may not lead on all SP types, DeepMaT maintains competitive performance, likely due to the data set's extreme class imbalance. In terms of cleavage site prediction, DeepMaT also outperforms other models for the Sec/SPI-labeled SPs. Figure 8C−G shows the model's overall performance across the entire SignalP 6.0 data set. DeepMaT demonstrates strong performance, particularly in predicting cleavage sites for Sec/SPIII- and TAT/SPII-labeled SPs, and shows enhanced ability to identify uncommon SP types. This may be attributed to the Mamba2 module, which effectively captures long-range dependencies and learns global sequence features critical for detecting less common signal peptide patterns.

3.7. Model Performance Evaluation on Additional Independent Test Set. To further demonstrate the generalization capability of our model, we constructed three

https://doi.org/10.1021/acs.jcim.5c01489
J. Chem. Inf. Model. 2025, 65, 10011−10024

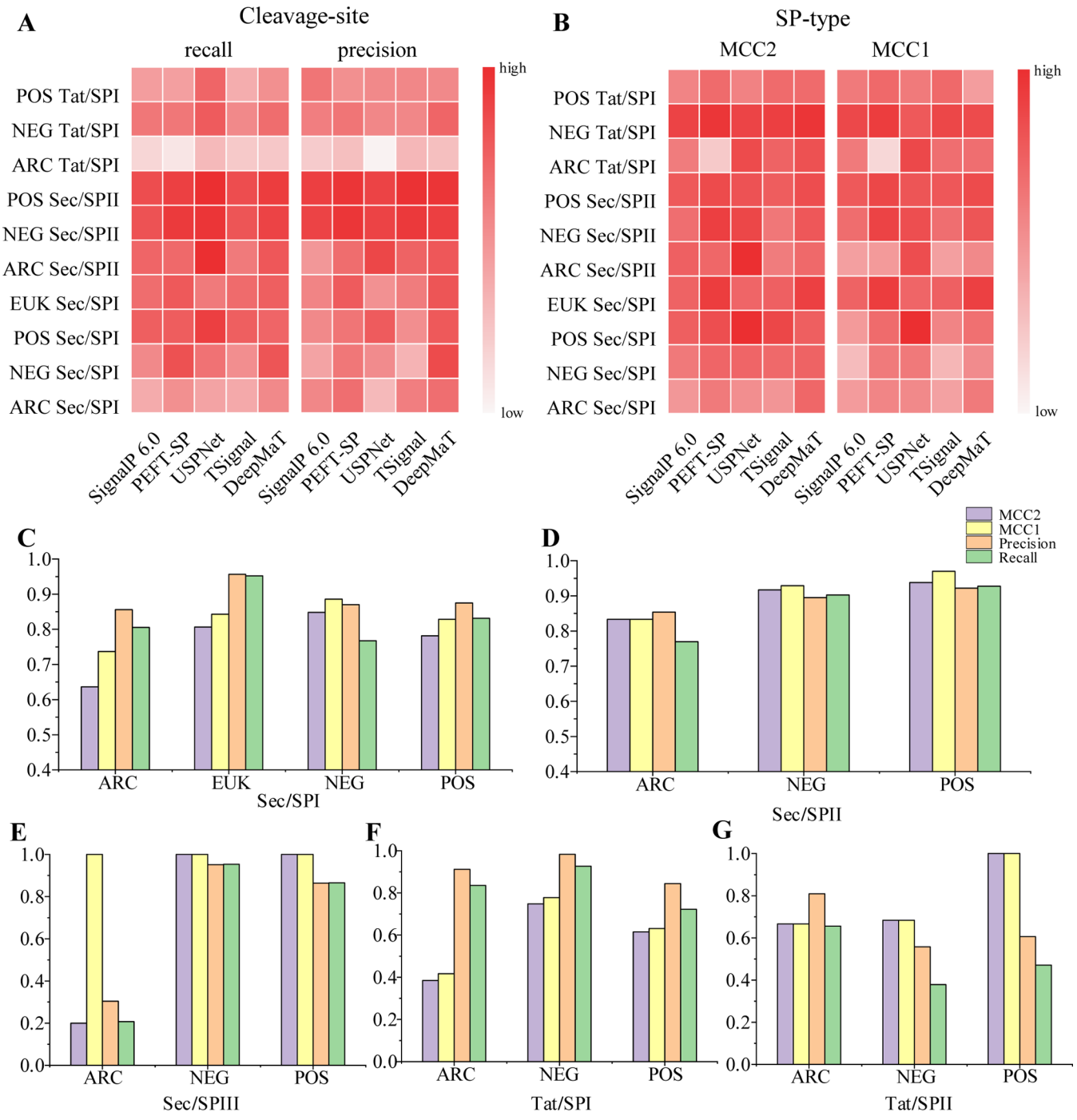


Figure 8. Comparison of signal peptide prediction results. (A,B) Show performance comparisons for specific signal peptide categories, where DeepMaT is evaluated alongside three baseline models: SignalP 6.0, PEFT-SP, and USPNet. (C−G) These five bar graphs show the performance of DeepMaT on five signal peptides, including Sec/SPI, Sec/SPII, Sec/SPIII, Tat/SPI, and Tat/SPII.

independent data sets—Independent Data set 1, Independent Data set 2 and Independent Data set 3—each comprising signal peptides, mitochondrial transit peptides, and chloroplast transit peptides (Table 3). In Figure 9, "Data 1" and "Data 2" represent the original data types, and the predictions from our model are compared with those of TargetP 2.0. The results from these independent test sets provide preliminary evidence that DeepMaT exhibits superior generalization performance compared to TargetP 2.0. We present the classification results on these data sets in Figure 9 and Supporting Information Table S2 to visualize the model's effectiveness. DeepMaT achieves higher classification accuracy than TargetP 2.0 for both mitochondrial and chloroplast transit peptides, as well as for cleavage site prediction, indicating a notable improvement. However, we also observed that both DeepMaT and TargetP 2.0 tend to misclassify a significant portion of nontarget peptides, contributing to classification confusion.

For the cleavage site prediction results on Independent Data set 2 (Supporting Information Table S3), it is evident that our model can generate up to five candidate predictions, providing insight into the probability distribution of potential cleavage sites. Among these, only two sequences—A0FKE6 and

https://doi.org/10.1021/acs.jcim.5c01489
J. Chem. Inf. Model. 2025, 65, 10011−10024

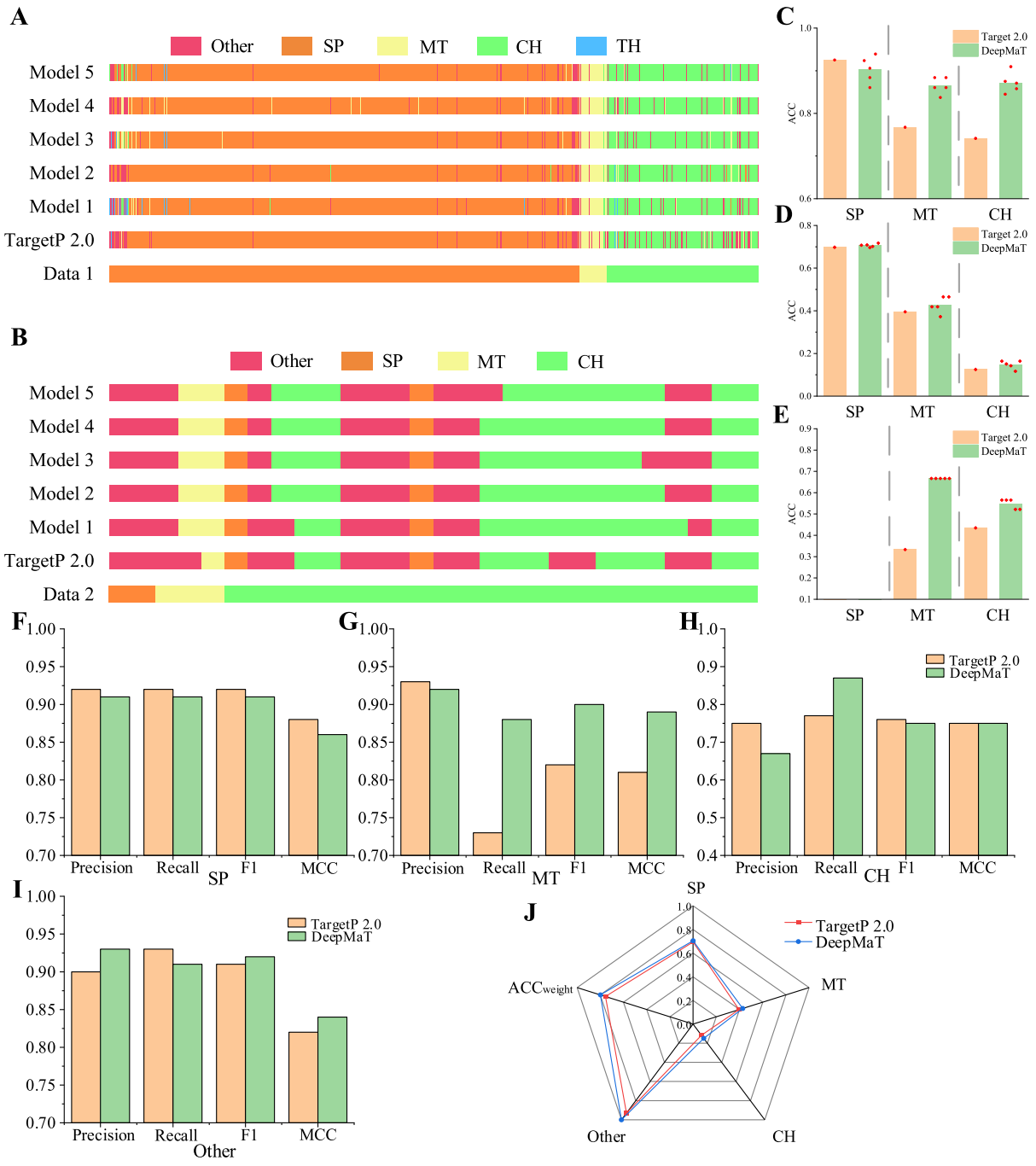


Figure 9. Independent test results. Models 1−5 correspond to the five DeepMaT models obtained through 5-fold training on the TargetP 2.0 data set. In panels (C−E), each red dot represents the evaluation metric of an individual model. (A) Classification performance comparison on Independent Data set 1. (B) Classification performance comparison on Independent Data set 2. (C) Classification accuracy on Independent Data set 1. (D) Cleavage site prediction accuracy on Independent Data set 1. (E) Classification accuracy on Independent Data set 2. (F−I) Classification performance metrics of DeepMaT on Independent Data set 3. (J) Cleavage site prediction performance metrics of DeepMaT on Independent Data set 3 include ACC for four peptides and ACC_{weight}.

P0DO76—were predicted identically by both DeepMaT and TargetP 2.0, suggesting that these sequences hold the highest reference value for consistent cleavage site prediction.

Independent Data set 3 we built is a balanced positive−negative data set, but it is not perfectly balanced in the positive class. The results (Figure 9F−J) from Independent Data set 3 show that our model performs better than TargetP 2.0 in some classifications and all cleavage sites. In the prediction task for negative cleavage sites, DeepMaT can predict all of them as no cleavage sites, which well demonstrates DeepMaT's excellent performance in cleavage site prediction tasks.

4. CONCLUSIONS

In this study, we present DeepMaT, a novel model designed to enhance the accuracy of targeted peptide classification and cleavage site prediction. DeepMaT integrates ISM for extracting evolutionary and structural features from sequences, followed by a feature learning module that combines Mamba2

https://doi.org/10.1021/acs.jcim.5c01489
J. Chem. Inf. Model. 2025, 65, 10011−10024

and MHA. This combination allows the model to capture both long-range and local dependencies effectively. The output is then processed by MLP for classification and CRF for cleavage site prediction. By leveraging the complementary strengths of Mamba2 and MHA, DeepMaT demonstrates improved predictive performance on the TargetP 2.0 data set, outperforming the original model. These results highlight DeepMaT as a promising tool for the accurate annotation of cleavage sites.

Despite the strong performance of DeepMaT, the model has certain limitations. DeepMaT integrates Mamba2 with MHA; however, it remains unclear whether these components capture distinct aspects of the input features. To enhance model interpretability, we recommend incorporating a multilayer attention mechanism to visualize the attention weights at each layer, which would help elucidate the positional focus and contribution of different layers. Furthermore, although ISM is effective at extracting evolutionary and structural information, it lacks the capacity to represent sequence-specific details and residue-level physicochemical properties. This limitation may introduce feature bias, potentially making the model overly sensitive to specific types of evolutionary or structural signals and increasing the risk of errors in cleavage site prediction. To address this, we propose enriching the feature set with additional biological descriptors to improve the model's predictive accuracy across diverse peptide types.

■ ASSOCIATED CONTENT

Data Availability Statement

Code and data sets of this study are available at https://github.com/qianmao2001/DeepMaT.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c01489.

Figure S1 DeepMaT experimental results for different encodings. Table S1 Comparison of ablation experiment results. Table S2 Performance metrics of DeepMaT and TargetP 2.0 on independent datasets. Table S3 Cleavage site prediction results for independent dataset 2 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Feifei Cui − *School of Computer Science and Technology, Hainan University, Haikou 570228, China;* orcid.org/0000-0001-7055-3813; Email: feifeicui@hainanu.edu.cn

Authors

Qianmao Wen − *School of Computer Science and Technology, Hainan University, Haikou 570228, China*

Aoyun Geng − *School of Computer Science and Technology, Hainan University, Haikou 570228, China*

Junlin Xu − *School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430081, China*

Yajie Meng − *School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan 430200, China*

Leyi Wei − *Centre for Artificial Intelligence Driven Drug Discovery, Faculty of Applied Science, Macao Polytechnic University, Macao SAR 999078, China;* orcid.org/0000-0003-1444-190X

Zilong Zhang − *School of Computer Science and Technology, Hainan University, Haikou 570228, China;* orcid.org/0000-0002-4934-1258

Quan Zou − *Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, China;* orcid.org/0000-0001-6406-1142

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c01489

Author Contributions

Q.W.: Writing—original draft, Designed, Experiments and Investigation. A.G.: writing-review and editing. J.X.: Writing—review and editing, Supervision. Y.M.: Writing—review and editing, Supervision. L.W.: Writing—review and editing, Supervision, Funding acquisition. Z.Z.: Writing—review and editing, Supervision. Q.Z.: Writing−review and editing, Supervision, Funding acquisition. F.C.: Writing—review and editing, Supervision.

Notes

Declaration of Generative AI and AI-assisted Technologies in the Writing Process: During the preparation of this work, the authors used Generative AI tools ChatGPT in order to improve the readability and language of the manuscript. The authors reviewed and verified all AI-assisted translations to ensure accuracy and consistency with the original content and take full responsibility for the content of the published article. The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The work is supported by the National Natural Science Foundation of China (No. 62450002) and the Science and Technology Development Fund of Macau (No. 0177/2023/RIA3).

■ REFERENCES

(1) Donnes, P.; Hoglund, A. Predicting protein subcellular localization: past, present, and future. *Genomics, Proteomics Bioinf.* **2004**, *2* (4), 209−215.

(2) Nakai, K. Protein sorting signals and prediction of subcellular localization. *Adv. Protein Chem.* **2000**, *54*, 277−344.

(3) Ispano, E.; Bianca, F.; Lavezzo, E.; Toppo, S. An overview of protein function prediction methods: a deep learning perspective. *Curr. Bioinf.* **2023**, *18* (8), 621−630.

(4) Rezende, S. B.; Lima, L. R.; Macedo, M. L.; Franco, O. L.; Cardoso, M. H. Advances in peptide/protein structure prediction tools and their relevance for structural biology in the last decade. *Curr. Bioinf.* **2023**, *18* (7), 559−575.

(5) Almagro Armenteros, J. J.; Salvatore, M.; Emanuelsson, O.; Winther, O.; von Heijne, G.; Elofsson, A.; Nielsen, H. Detecting sequence signals in targeting peptides using deep learning. *Life Sci. Alliance* **2019**, *2* (5), No. e201900429.

(6) Le, N. Q. K. Explainable artificial intelligence for protein function prediction: a perspective view. *Curr. Bioinf.* **2023**, *18*, 205−207.

(7) Owji, H.; Nezafat, N.; Negahdaripour, M.; Hajiebrahimi, A.; Ghasemi, Y. A comprehensive review of signal peptides: Structure, roles, and applications. *Eur. J. Cell Biol.* **2018**, *97* (6), 422−441.

(8) Jain, N.; Chacinska, A.; Rehling, P. Understanding mitochondrial protein import: a revised model of the presequence translocase. *Trends Biochem. Sci.* **2025**, *50*, 585−595.

(9) Ballabani, G.; Forough, M.; Kessler, F.; Shanmugabalaji, V. The journey of preproteins across the chloroplast membrane systems. *Front. Physiol.* **2023**, *14*, 1213866.

(10) Kunze, M.; Berger, J. The similarity between N-terminal targeting signals for protein import into different organelles and its evolutionary relevance. *Front. Physiol.* **2015**, *6*, 259.
(11) Snapp, E. Design and Use of Fluorescent Fusion Proteins in Cell Biology. *Curr. Protoc. Cell Biol.* **2005**, *27* (1), 21.4.1.
(12) Simpson, J. C.; Wellenreuther, R.; Poustka, A.; Pepperkok, R.; Wiemann, S. Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Rep.* **2000**, *1* (3), 287−292.
(13) Tanz, S. K.; Castleden, I.; Small, I. D.; Millar, A. H. Fluorescent protein tagging as a tool to define the subcellular distribution of proteins in plants. *Front. Plant Sci.* **2013**, *4*, 214.
(14) Edman, P. H.; Högfeldt, E.; Sillén, L. G.; Kinell, P.-O. Method for determination of the amino acid sequence in peptides. *Acta Chem. Scand.* **1950**, *4*, 283−293.
(15) Soylu, N. N.; Sefer, E. DeepPTM: protein post-translational modification prediction from protein sequences by combining deep protein language model with vision transformers. *Curr. Bioinf.* **2024**, *19* (9), 810−824.
(16) Wang, Y.; Zhai, Y.; Ding, Y.; Zou, Q. SBSM-Pro: support bio-sequence machine for proteins. *Sci. China Inf. Sci.* **2024**, *67* (11), 212106.
(17) Fukasawa, Y.; Tsuji, J.; Fu, S.-C.; Tomii, K.; Horton, P.; Imai, K. MitoFates: Improved Prediction of Mitochondrial Targeting Sequences and Their Cleavage Sites. *Mol. Cell. Proteomics* **2015**, *14* (4), 1113−1126.
(18) Savojardo, C.; Bruciaferri, N.; Tartari, G.; Martelli, P. L.; Casadio, R. DeepMito: accurate prediction of protein sub-mitochondrial localization using convolutional neural networks. *Bioinformatics* **2020**, *36* (1), 56−64.
(19) Emanuelsson, O.; Nielsen, H.; Heijne, G. V. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci.* **1999**, *8* (5), 978−984.
(20) Westerlund, I.; Gv, H.; Emanuelsson, O. LumenP–a neural network predictor for protein localization in the thylakoid lumen. *Protein Sci.* **2003**, *12* (10), 2360−2366.
(21) Savojardo, C.; Martelli, P. L.; Fariselli, P.; Casadio, R. TPpred3 detects and discriminates mitochondrial and chloroplastic targeting peptides in eukaryotic proteins. *Bioinformatics* **2015**, *31* (20), 3269−3275.
(22) Almagro Armenteros, J. J.; Sønderby, C. K.; Sønderby, S. K.; Nielsen, H.; Winther, O. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* **2017**, *33* (21), 3387−3395.
(23) Thumuluri, V.; Almagro Armenteros, J. J.; Johansen, A. R.; Nielsen, H.; Winther, O. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Res.* **2022**, *50* (W1), W228−W234.
(24) Teufel, F.; Almagro Armenteros, J. J.; Johansen, A. R.; Gíslason, M. H.; Pihl, S. I.; Tsirigos, K. D.; Winther, O.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* **2022**, *40* (7), 1023−1025.
(25) Zeng, S.; Wang, D.; Jiang, L.; Xu, D. Prompt-Based Learning on Large Protein Language Models Improves Signal Peptide Prediction. In *Research in Computational Molecular Biology*; Springer Nature Switzerland: Cham, 2024; pp 400−405.
(26) Shen, J.; Yu, Q.; Chen, S.; Tan, Q.; Li, J.; Li, Y. Unbiased organism-agnostic and highly sensitive signal peptide predictor with deep protein language model. *Nat. Comput. Sci.* **2024**, *4* (1), 29−42.
(27) Savojardo, C.; Martelli, P. L.; Fariselli, P.; Casadio, R. DeepSig: deep learning improves signal peptide detection in proteins. *Bioinformatics* **2018**, *34* (10), 1690−1696.
(28) Dumitrescu, A.; Jokinen, E.; Paatero, A.; Kellosalo, J.; Paavilainen, V. O.; Lähdesmäki, H. TSignal: a transformer model for signal peptide prediction. *Bioinformatics* **2023**, *39* (Supplement_1), i347−i356.
(29) Emanuelsson, O.; Henrik, N.; Søren, B.; Heijne, G. v. Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence. *J. Mol. Biol.* **2000**, *300* (4), 1005−1016.
(30) Yang, Y.; Zuo, X.; Das, A.; Xu, H.; Zheng, W. Representation learning of biological concepts: a systematic review. *Curr. Bioinf.* **2024**, *19* (1), 61−72.
(31) Dao, T.; Gu, A. Transformers are SSMs: Generalized Models and Efficient Algorithms Through Structured State Space Duality. In *Proceedings of the 41st International Conference on Machine Learning; Proceedings of Machine Learning Research*; PMLR, 2024, pp 10041−10071.
(32) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Lu.; Polosukhin, I. Attention is All you Need. In *Advances in Neural Information Pro-cessing Systems*; Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; NIPS, 2017; pp 5998−6008.
(33) Ouyang-Zhang, J.; Gong, C.; Zhao, Y.; Kraehenbuehl, P.; Klivans, A.; Diaz, D. J. Distilling Structural Representations into Protein Sequence Models. In *The Thirteenth International Conference on Learning Representations*, 2025.
(34) UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **2015**, *43* (D1), D204−D212.
(35) Almagro Armenteros, J. J.; Tsirigos, K. D.; Sønderby, C. K.; Petersen, T. N.; Winther, O.; Brunak, S.; von Heijne, G.; Nielsen, H. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **2019**, *37* (4), 420−423.
(36) Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Adesina, A.; Ahmad, S.; Bowler-Barnett, E. H.; Bye-A-Jee, H.; Carpentier, D.; Denny, P.; et al. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* **2025**, *53* (D1), D609−D617.
(37) Zhang, J.; Song, C.; Cui, T.; Li, C.; Ma, J. ChiMamba: Predicting Chromatin Interactions Based on Mamba. In *Advanced Intelligent Computing in Bioinformatics*; Springer Nature: Singapore, 2024; pp 50−61.
(38) Peng, F. Z.; Wang, C.; Chen, T.; Schussheim, B.; Vincoff, S.; Chatterjee, P. PTM-Mamba: a PTM-aware protein language model with bidirectional gated Mamba blocks. *Nat. Methods* **2025**, *22* (5), 945−949.
(39) Schiff, Y.; Kao, C. H.; Gokaslan, A.; Dao, T.; Gu, A.; Kuleshov, V. Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling. In *Proceedings of the 41st International Conference on Machine Learning; Proceedings of Machine Learning Research*; PMLR, 2024, pp 43632−43648.
(40) Lafferty, J. D.; McCallum, A.; Pereira, F. C. N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*; Morgan Kaufmann Publishers Inc., 2001, pp 282−289.
(41) de Boer, P.-T.; Kroese, D. P.; Mannor, S.; Rubinstein, R. Y. A Tutorial on the Cross-Entropy Method. *Ann. Oper. Res.* **2005**, *134* (1), 19−67.
(42) Li, Y.; Geng, A.; Zhou, Z.; Cui, F.; Xu, J.; Meng, Y.; Wei, L.; Zou, Q.; Zhang, Q.; Zhang, Z. AVP-HNCL: Innovative Contrastive Learning with a Queue-Based Negative Sampling Strategy for Dual-Phase Antiviral Peptide Prediction. *J. Chem. Inf. Model.* **2025**, *65*, 5868−5886.
(43) Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A. F.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**, *16* (5), 412−424.
(44) Geng, A.; Luo, Z.; Li, A.; Zhang, Z.; Zou, Q.; Wei, L.; Cui, F. ACP-CLB: An Anticancer Peptide Prediction Model Based on Multichannel Discriminative Processing and Integration of Large Pretrained Protein Language Models. *J. Chem. Inf. Model.* **2025**, *65*, 2336−2349.
(45) Lv, J.; Li, K.; Wang, Y.; Xu, J.; Meng, Y.; Cui, F.; Wei, L.; Zhang, Q.; Zhang, Z. ACP-EPC: an interpretable deep learning framework for anticancer peptide prediction utilizing pre-trained protein language model and multi-view feature extracting strategy. *Mol. Diversity* **2025**, 1.

https://doi.org/10.1021/acs.jcim.5c01489
J. Chem. Inf. Model. 2025, 65, 10011−10024

(46) Johnson, J. M.; Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *J. Big Data* **2019**, *6* (1), 27.
(47) Jeong, J.; Hwang, I.; Lee, D. W. Functional Organization of Sequence Motifs in Diverse Transit Peptides of Chloroplast Proteins. *Front. Physiol.* **2021**, *12*, 795156.
(48) von Heijne, G.; Steppuhn, J.; Herrmann, R. G. Domain structure of mitochondrial and chloroplast targeting peptides. *Eur. J. Biochem.* **1989**, *180* (3), 535−545.
(49) Tareen, A.; Kinney, J. B. Logomaker: Beautiful sequence logos in python. *bioRxiv* **2019**, 635029.
(50) Rapoport, T. A. Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature* **2007**, *450* (7170), 663−669.
(51) Nishimura, T.; Lazzeri, G.; Mizushima, N.; Covino, R.; Tooze, S. A. Unique amphipathic α helix drives membrane insertion and enzymatic activity of ATG3. *Sci. Adv.* **2023**, *9* (25), No. eadh1281.
(52) Voisine, C.; Craig, E. A.; Zufall, N.; von Ahsen, O.; Pfanner, N.; Voos, W. The Protein Import Motor of Mitochondria. *Cell* **1999**, *97* (5), 565−574.
(53) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **2023**, *379* (6637), 1123−1130.
(54) Edward, J. H.; yelong, s.; Phillip, W.; Zeyuan, A.-Z.; Yuanzhi, L.; Shean, W.; Lu, W.; Weizhu, C. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022.
(55) Xiao, C.; Zhou, Z.; She, J.; Yin, J.; Cui, F.; Zhang, Z. PEL-PVP: Application of plant vacuolar protein discriminator based on PEFT ESM-2 and bilayer LSTM in an unbalanced dataset. *Int. J. Biol. Macromol.* **2024**, *277*, 134317.

https://doi.org/10.1021/acs.jcim.5c01489
J. Chem. Inf. Model. 2025, 65, 10011−10024