Improving PTM Site Prediction by Coupling of Multi-Granularity Structure and Multi-Scale Sequence Representation

Zhengyi Li[1], Menglu Li[1], Lida Zhu[1*], Wen Zhang[1,2,3*]

[1]College of Informatics, Huazhong Agricultural University, Wuhan 430070, China
[2]Hubei Key Laboratory of Agricultural Bioinformatics, Huazhong Agricultural University, Wuhan 430070, China
[3]Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education, Wuhan 430070, China
{lzy_gnn, mengluli}@webmail.hzau.edu.cn, {ldzhu, zhangwen}@mail.hzau.edu.cn

Abstract

Protein post-translational modification (PTM) site prediction is a fundamental task in bioinformatics. Several computational methods have been developed to predict PTM sites. However, existing methods ignore the structure information and merely utilize protein sequences. Furthermore, designing a more fine-grained structure representation learning method is urgently needed as PTM is a biological event that occurs at the atom granularity. In this paper, we propose a **PTM** site prediction method by **C**oupling of **M**ulti-**G**ranularity structure and **M**ulti-**S**cale sequence representation, **PTM-CMGMS** for brevity. Specifically, multi-granularity structure-aware representation learning is designed to learn neighborhood structure representations at the amino acid, atom, and whole protein granularity from AlphaFold predicted structures, followed by utilizing contrastive learning to optimize the structure representations. Additionally, multi-scale sequence representation learning is used to extract context sequence information, and motif generated by aligning all context sequences of PTM sites assists the prediction. Extensive experiments on three datasets show that PTM-CMGMS outperforms the state-of-the-art methods. Source code can be found at https://github.com/LZY-HZAU/PTM-CMGMS.

Introduction

Post-translational modification (PTM) refers to the biological events of adding small molecular groups to the side chains of amino acids, which greatly enhances the functional diversity of the proteome. Accurate determination of PTM sites contributes to a deepening understanding of protein functions(Yang and Gibson 2019) and their specific roles in various complex cellular processes (Wei et al. 2017; Fernando et al. 2019; Fang et al. 2021), and facilitates precision therapy by illuminating the regulatory systems underlying disease states (Cai et al. 2022). Several wet experiments have been used for identifying PTM sites, such as high-sensitivity mass spectrometry, antibodies, isotope labeling, and probes (Bos and Muir 2018), but they are often time-consuming and labor-intensive, hindering large-scale investigations into PTM.
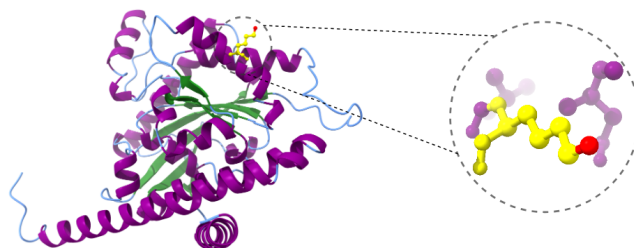
Figure 1: The example of lysine crotonylation, which occurs on the side chain nitrogen atom marked in red. The other atoms of lysine are marked in yellow.

In recent years, plenty of computational methods have been developed for predicting PTM sites, and these methods are roughly classified into three categories: machine learning-based methods, deep learning-based methods, and pre-trained language model-based methods. The machine learning-based methods (Xu et al. 2013; Qiu et al. 2017; Ning et al. 2018; Hasan et al. 2019; Liu et al. 2020) extracted the protein sequence-derived features manually and then fed them into the classifiers, such as support vector machine (SVM), random forest, and conditional random field algorithm, for PTM site prediction. These methods heavily rely on feature engineering and cannot learn abstract feature representations. Deep learning-based methods (Xie et al. 2018; Thapa et al. 2020; Lv et al. 2021; Li et al. 2022) employed a deep learning model, such as bidirectional long short-term memory network(BiLSTM), convolutional neural network(CNN), deep neural network(DNN) and transformer, to learn representations automatically, which have been demonstrated to be powerful techniques for PTM site prediction. However, these methods roughly address PTM site prediction as a sequence modeling problem, ignoring the structure information. The pre-trained language models (PLMs) on large-scale protein sequence corpora have achieved impressive performance for various downstream protein understanding tasks. A series of pre-trained language model-based methods (Qiao, Zhu, and Gong 2022; Pokharel et al. 2022) have been developed, which utilized the sequence representation extracted from PLMs, such as BERT (Devlin et al. 2018) and ProtTrans (Elnaggar et al. 2021), for PTM site prediction. PLMs can learn implicit structure in-

formation, but cannot be directly aware of the explicit neighborhood structure of PTM sites. More importantly, PLMs only generate sequence representations at the amino acid or whole protein granularity, which may not be suitable for the specific biological problem as PTM occurs at the atom granularity, shown in Figure 1.

Recently, researchers have paid more attention to the structure information of PTM sites (Gao et al. 2017; Bludau et al. 2022). For example, Li et al. (2020) investigated the structure background of PTM sites and provided detailed annotations on structure features such as secondary structure, tertiary structure context, and solvent accessibility of PTM sites. These studies indicate that PTM sites exhibit structure preferences, which may play a vital role in improving the performance of PTM site prediction. On the other hand, the appearance of AlphaFold (Jumper et al. 2021; Varadi et al. 2022) enables us to obtain protein structures of a wide range of organisms without resorting to traditional experimental techniques, inspiring us to design a structure-aware PTM site prediction model.

In this paper, we propose a novel PTM site prediction method by Coupling of Multi-Granularity Structure and Multi-Scale Sequence Representation, namely PTM-CMGMS. Specifically, we design a multi-granularity structure-aware representation learning, including the atom, amino acid, and whole protein granularity, to learn neighborhood structure representations of PTM sites from AlphaFold predicted structures. Further, we adopt contrastive learning to optimize the structure representations. Additionally, we use a multi-scale sequence representation learning to extract context sequence information of PTM sites, and obtain the motif-based information by detecting the differences between context sequences and motif which is generated by aligning all context sequences of PTM sites. Finally, the combination of structure representations, context sequence information, and motif-based information is fed into an MLP to produce the prediction probability.

In summary, the main contributions of this paper are described as follows:

- We design a multi-granularity structure-aware representation learning that integrates structure information of amino acid, atom, and whole protein granularity to learn more comprehensive structure representations. To the best of our knowledge, this is the first method to model PTM sites at the atom granularity, which is more consistent with biochemical facts.

- We utilize contrastive learning to optimize structure representations of PTM sites for enhancing the generalization capacity of PTM-CMGMS.

- We design a multi-scale sequence representation learning to simulate local amino acid interactions at different scales for obtaining sequence representations, which is used to assist the prediction of our model.

Related Work

Existing methods mainly focus on the context sequence information of PTM sites.

A line of works manually extract the sequence information as feature vectors and train a classic classifier to predict PTM sites (Xue et al. 2010; Lee et al. 2011; Xu et al. 2015; Jia et al. 2016a,b; Hasan et al. 2016; Ju and He 2017; Hasan and Kurata 2018). For example, Qiu et al. (2017) design an encoding scheme called Position Weighted Amino Acid Composition (PWAA) and use an SVM as the classifier for PTM site prediction. PSuccE (Ning et al. 2018) adopts an ensemble learning algorithm with a feature selection scheme to enhance the prediction performance. LightGBM-CroSite (Liu et al. 2020) integrates multiple encoding schemes to extract more comprehensive sequence features and uses the elastic net to remove redundant information. Although machine learning-based methods have achieved great success in PTM site prediction, they heavily rely on hand-crafted features and domain expertise, making it difficult to capture the implicit information of protein sequences. The other line of works usually train a deep learning model that learn representations automatically. For instance, DeepNitro (Xie et al. 2018) constructs four types of encoding features (positional amino acid distributions, sequence contextual dependencies, physicochemical properties, and position-specific scoring features), and uses a DNN to learn the high-dive information for representing the PTM sites. Deep-Kcr (Lv et al. 2021) combines sequence-based features, physicochemical property-based features, and numerical space-derived information to a CNN for further representation learning to predict PTM sites. Similar to Deep-Kcr, DeepSuccinylSite (Thapa et al. 2020) adopts one-hot and word embeddings as features and inputs them into a CNN for PTM site prediction. Adapt-Kcr (Li et al. 2022) utilizes an adaptive word embedding algorithm to encode the sequence, employs CNN, a BLSTM, and an attention mechanism to better capture the latent information of adaptive embedding for predicting PTM sites. These deep learning-based methods merely utilize the protein sequence while ignoring structure information.

Additionally, there are also a series of works based on pre-trained language models. BERT-Kcr (Qiao, Zhu, and Gong 2022) transfers each amino acid into a word as the input of BERT (Devlin et al. 2018), and adopts BiLSTM as the classifier for PTM site prediction. They also attempted to pre-train BERT using protein sequences and fine-tune the model for this prediction task. LMSuccSite (Pokharel et al. 2022) utilizes the embeddings derived from ProtTrans in conjunction with supervised word embedding to improve the prediction performance. The pre-trained language models typically involve tens of millions of parameters, and require too much time and computational resources, especially for large-scale sequences input. Additionally, the pre-trained language models can not generate representations of atom granularity for PTM site prediction.

Different from the above methods, we integrate multi-granularity structure and multi-scale sequence information to construct a PTM site prediction method.

Method

In this section, we first formulate the PTM site prediction problem. After that, we elaborately enumerate all compo-
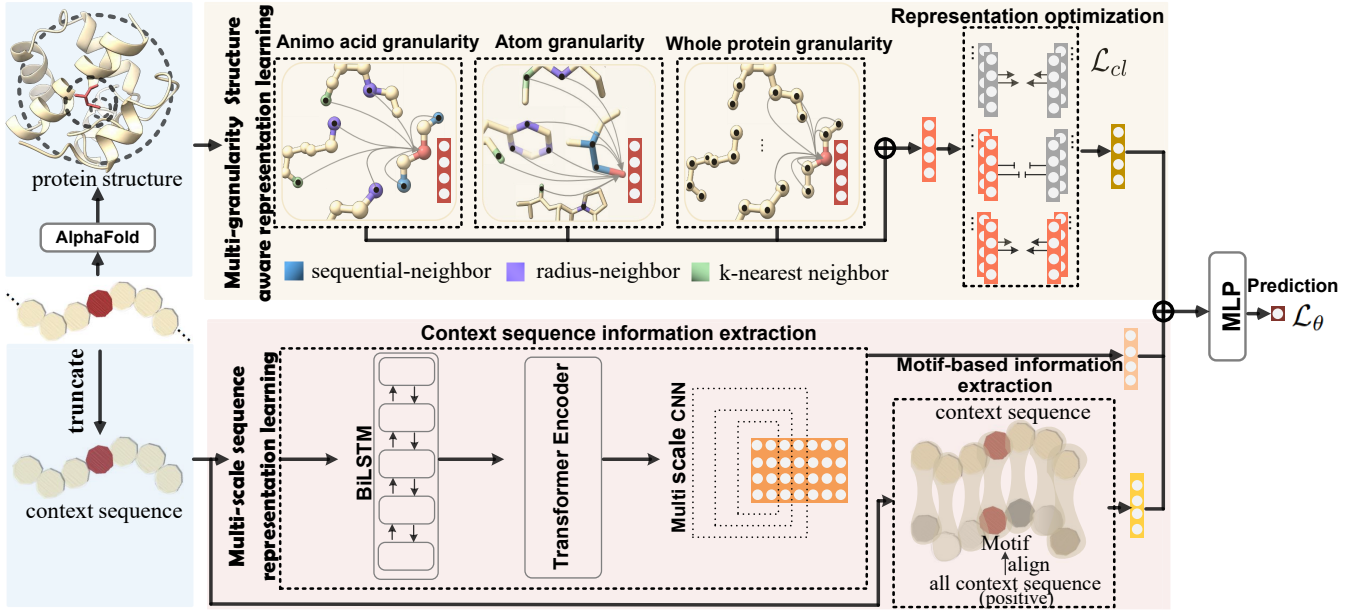
Figure 2: The overview of our proposed PTM-CMGMS.

nents of the proposed method PTM-CMGMS, shown in Figure 2. At last, we provide an exposition of model training.

Problem Formulation

Given a protein sequence $S = \{s_i | i = 1, 2, ..., L\}$ with L residues. $\mathcal{X} = \{x_i \in \mathbb{R}^{Z \times 3} | i = 1, 2, ..., L\}$ denotes the coordinates in 3D space for the residues with Z atoms (e.g., N, C_α, C, O) obtained by AlphaFold. The context sequence P is obtained by truncating 15aa short peptide fragments upstream and downstream of s_i in S, where the residue s_i is to be predicted, and the length of P is 31aa. In this paper, our goal is to construct a model that predicts whether s_i is a PTM site (positive) or not (negative) based on the protein structure and context sequence, which is a binary classification problem.

Multi-Granularity Structure-Aware Representation Learning

In this section, we learn multi-granularity structure representations of s_i at the amino acid, atom, and whole protein granularity. Then contrastive learning is utilized to optimize the structure representations.

Amino Acid Granularity. We capture three types of neighbors within residue s_i to learn the neighborhood structure representation at amino acid granularity, including sequential-neighbors, radius-neighbors, and k-nearest neighbors (Zhang et al. 2022; Zeng et al. 2023a). They are generated as follows: sequential-neighbors $\{s_j | |s_j - s_i| < d_{seq}\}$, where $|s_j - s_i|$ represents the sequential distance between residues s_j and s_i at S; radius-neighbors $\{s_j | D(x_j(C_\alpha) - x_i(C_\alpha)) < d_{spa}\}$, where $D(x_j(C_\alpha) - x_i(C_\alpha))$ represents the Euclidean distance between C_α of residues s_j and s_i; k-nearest neighbors $\{s_j\}_{j=1}^k$, where s_j

is the top j nearest neighbors of s_i based on the Euclidean distance. In this study, we set $d_{seq} = 3$, $d_{rad} = 10\mathring{A}$ and $k = 10$ (Zeng et al. 2023a).

Based on these neighbors, we can obtain the amino acid granularity representation of the residue s_i:

$$r_{s_i}^{aa} = \sigma(W \sum_{j \in \mathcal{N}(s_i)} (H_j || E_j)) \tag{1}$$

where W is the learnable parameter, $\sigma(\cdot)$ denotes an activation function LeakyReLU, $||$ represents concatenation operation, and $\mathcal{N}(s_i)$ is a set of three type neighbors of residue s_i. H_j represents the embedding of residue type, secondary structure, backbone torsion angle ϕ and ψ, and relative solvent accessibility of residue s_j (Kabsch and Sander 1983). E_j represents the embedding of specific neighbor type of s_j related to s_i, the sequential and Euclidean distance between s_j and s_i.

Atom Granularity. We capture three types of neighbors within the particular side chain atom of residue s_i to learn the neighborhood structure representation at atom granularity, which is due to PTM occurring at the particular side chain atom of residue s_i (e.g. crotonylation occurs on the side chain nitrogen atom). Here, we take the prediction of whether s_i is a crotonylation site as an example to illustrate. For the residue s_i, the generation process of three types of neighbors of side chain nitrogen atom is similar to the amino acid granularity, the atom granularity representation is as follows:

$$r_{s_i}^{atom} = \sigma(W \sum_{j \in \mathcal{N}(s_i^{(N)})} (h_j || e_j)) \tag{2}$$

where h_j represents the embedding of atom type of j and its residue type, e_j represents the embedding of specific neigh-

bor type of j related to the nitrogen atom of s_i, the sequential and Euclidean distance between them. $s_i^{(N)}$ represents the nitrogen atom of s_i.

Whole Protein Granularity. We capture all residue information to learn the neighborhood structure representation at whole protein granularity, and the whole protein granularity representation of s_i is as follows:

$$r_{s_i}^{pro} = \sigma(W \sum_{j \in S}(H_j)) \tag{3}$$

After that, we integrate these three representations to obtain the multi-granularity structure representation of s_i, denoted as r^{MG}:

$$r^{MG} = \text{MLP}(r_{s_i}^{atom}||r_{s_i}^{aa}||r_{s_i}^{pro}) \tag{4}$$

where MLP is a multi-layer perceptron.

Representation Optimization. In order to learn discriminative multi-granularity structure representations, we introduce contrastive learning in our model. For the structure representations of residue s_i and s_j, we calculate the contrastive learning loss as follows (Zhang et al. 2023):

$$\begin{aligned}\mathcal{L}_{cl} = \frac{1}{2}((1 - Y)D(r_{s_i}^{MG}, r_{s_j}^{MG})^2 + \\ Y\{\max(0, M - D(r_{s_i}^{MG}, r_{s_j}^{MG}))^2\})\end{aligned} \tag{5}$$

where Y is equal to 0 if and only if s_i and s_j belong to the same label (both positive or both negative samples), and 1 in other cases. M represents the cut-off distance. If s_i and s_j belong to different labels and the distance of their structure representations is less than M, we optimize them so that they are far away from each other.

Multi-Scale Sequence Representation Learning

In this section, we learn multi-scale sequence representation by integrating context sequence information and motif-based information.

Context Sequence Information Extraction. For the context sequence P of s_i, we initialize embedding vectors for each residue by fusing position and token embeddings (He et al. 2021). The embedding of P is computed as follows:

$$r^P = embedding_{(position)} + embedding_{(token)} \tag{6}$$

where $r^P \in \mathbb{R}^{(31 \times d)}$, d represents the embedding dimension.

We then utilize BiLSTM (Hochreiter and Schmidhuber 1997) to integrate forward and reverse context sequence information, and feed the output of BiLSTM into the transformer encoder (Vaswani et al. 2017) for further feature extraction. The core of the transformer encoder is the multi-head self-attention mechanism, which enables s_i to aggregate other residue embeddings with different attention weights.

$$Q_i = r^P W_i^Q; K_i = r^P W_i^K; V_i = r^P W_i^V \tag{7}$$

$$\text{head}_i = \text{Softmax}(\frac{Q_i K_i^T}{\sqrt{d}})V_i \tag{8}$$

$$r^P = \|_{i=1}^h \text{head}_i \cdot W \tag{9}$$

where W_i^Q, W_i^K, W_i^V are query, key, value matrix of i-th head ($i = 1, 2, ..., h$), respectively, h is the number of heads, W is a linear transformation matrix, and d is the scaling factor.

In order to capture local patterns, we further design a multi-scale CNN (MCNN) with multiple convolutional kernels $\{c_i\}_{i=1}^C$ to simulate the local interaction of residues at different scales.

$$r^P = \sigma(W(\text{mean}(\|_{i=1}^C \text{Pool}(\text{Conv}(r^P, c_i))))) \tag{10}$$

Based on the combination of BiLSTM, Transformer encoder, and MCNN, we extract the context sequence information to obtain the representations r^P of context sequence P.

Motif-Based Information Extraction. We align the context sequences of all known positive sites in the training set to obtain the frequency of residues at each position. The residues with the highest frequency (denoted as w) at each position make up the motif M (the length is 31aa) which reflects the contextual pattern of the PTM sites. In general, s_i is more likely to belong to a PTM site if context sequence P is similar to M.

Here, we adopt physicochemical properties AAIndex (AAI) and evolutionary relationships BLOSUM62 (Ning and Li 2022) to characterize the difference between P and M:

$$\begin{aligned}r^M = \text{MLP}(w \cdot D(\text{AAI}_M||\text{BLOSUM62}_M, \\ \text{AAI}_P||\text{BLOSUM62}_P))\end{aligned} \tag{11}$$

where AAI_M and AAI_P denote the physicochemical property embedding matrices of M and P, BLOSUM62_M and BLOSUM62_P denote the evolutionary relationship embedding matrices of M and P, respectively.

After that, we combine these two representations to obtain a multi-scale sequence representation of P:

$$r^{MS} = r^P||r^M \tag{12}$$

Model Training

For the residue s_i, the multi-granularity structure representations r^{MG} and multi-scale sequence representation r^{MS} are concatenated and fed into an MLP to get the probability of PTM site:

$$\hat{y} = \text{MLP}\left(r^{MG}||r^{MS}\right) \tag{13}$$

The training objective is to minimize the loss function:

$$\mathcal{L}_\theta = -\sum_{i=1}^N \left(y_i \log\left(\hat{y}_i\right) + (1 - y_i) log\left(1 - \hat{y}_i\right)\right) \tag{14}$$

where N represents the total number of samples in the training set, y_i represents the true label of the i-th sample, and $\hat{y}_i$ represents the probability predicted by our model.

Experiments

In this section, we first introduce the experimental settings, and then compare our proposed PTM-CMGMS with baselines. We then employ the ablation study to investigate the effectiveness of each component. Further, we also explore the importance of structure information for PTM site prediction.

Experimental Settings

Datasets. To test the performance of our proposed PTM-CMGMS, we evaluate it on three publicly available datasets with different PTM types, including Crotonylation (Qiao, Zhu, and Gong 2022; Yu et al. 2020), Succinylation (Pokharel et al. 2022), and Nitrosylation (Pratyush et al. 2023). The division scheme of these datasets is consistent with the previous studies, and the redundant sites in the datasets are removed. The statistics of three datasets are in Table 1.

Datasets	Training set		Test set	
	Positive	Negative	Positive	Negative
Crotonylation	6,975	6,975	2,889	1,939
Succinylation	4,749	4,750	253	2,973
Nitrosylation	3,276	3,276	351	3,168

Table 1: The statistics of three datasets.

Baselines. We compare our proposed PTM-CMGMS with several baselines, which can be categorized as follows.

- PSuccE (Ning et al. 2018) adopts an ensemble learning algorithm to predict succinylation sites combining multiple features with a feature selection scheme (information gain).

- LightGBM-CroSite (Liu et al. 2020) uses the elastic net to select the optimal feature subset from multiple sequence features, and inputs them into a LightGBM to predict crotonylation sites.

- DeepNitro (Xie et al. 2018) utilizes four types of encoding features, and trains a deep neural network to predict nitrosylation sites.

- Deep-Kcr (Lv et al. 2021) combines sequence-based, physicochemical property-based, and numerical space-derived information to predict crotonylation sites based on a CNN.

- DeepSuccinylSite (Thapa et al. 2020) combines one-hot and word embeddings as features, and uses a CNN to predict succinylation sites.

- Adapt-Kcr (Li et al. 2022) utilizes adaptive embedding and is based on a CNN together with a BiLSTM and attention architecture to predict crotonylation sites.

- BERT-Kcr (Qiao, Zhu, and Gong 2022) employs pre-trained BERT to transfer each amino acid into a word, and adopts a BiLSTM to predict crotonylation sites.

- LMSuccSite (Pokharel et al. 2022) combines supervised word embedding and embedding learned from a pre-trained protein language model, and uses a neural network to predict succinylation sites.

Implementation Details. To measure the performance of our method, we implement the 10-fold cross-validation on the training set, conduct independent testing on the test set, in which the model is trained on the training set. And we adopt three evaluation metrics, including the Matthews correlation coefficient (MCC), the area under the receiver-operating characteristic curve (AUC), and the area under the precision-recall curve (AUPR). The experiments are performed on the machine with Intel(R)Core(TM)i9-10980XE CPU@3.00GHz and 2 GPUs(NVIDIA GeForce RTX 3090).

Comparison with Baselines

Table 2 shows the evaluation results of PTM-CMGMS and the baselines on three datasets. According to the results shown in Table 2, PTM-CMGMS achieves the best performance on three datasets. We also have the following observation: (1) Compared with PSuccE and LightGBM-CrotSite which rely on hand-crafted features, PTM-CMGMS significantly exceeds all these baselines on three datasets, deep learning-based methods (DeepNitro, Deep-Kcr, DeepSuccinylSite, and Adapt-Kcr) and pre-trained language model-based methods (BERT-Kcr and LMSuccSite) also achieve better performance than PSuccE and LightGBM-CrotSite on most datasets, which indicates that using deep learning techniques to automatically extract sequence information can enhance the prediction for PTM sites. (2) Compared with DeepNitro, Deep-Kcr, DeepSuccinylSite, and Adapt-Kcr which only consider protein sequence information, PTM-CMGMS achieves the best performance on all datasets, which implys that considering structure information benefits PTM site prediction. (3) Compared with BERT-Kcr and LMSuccSite that utilize representations at the amino acid granularity generated by pre-trained language models, PTM-CMGMS makes improvements of 13.56% and 7.83% on Crotonylation dataset, 9.49% and 0.82% on Succinylation dataset, and 12.49% and 3.38% on Nitrosylation dataset in terms of MCC, indicating the information of atom granularity is also beneficial for PTM sites prediction.

Ablation Study

To illustrate the importance and effectiveness of each component designed in our model, we consider the following variants of PTM-CMGMS.

- PTM-CMGMS without amino acid granularity (w/o AA) removes the structure representations at the amino acid granularity in the multi-granularity structure-aware representation learning.

- PTM-CMGMS without atom granularity (w/o Atom) removes the structure representations at the atom granularity in the multi-granularity structure-aware representation learning.

- PTM-CMGMS without whole protein granularity (w/o WP) removes the structure representations at the whole

Method	Crotonylation			Succinylation			Nitrosylation		
	MCC	AUC	AUPR	MCC	AUC	AUPR	MCC	AUC	AUPR
PSuccE	0.5142	0.8427	0.8744	0.1665	0.7187	0.1839	0.2699	0.7807	0.2509
LightGBM-CrotSite	0.5396	0.8587	0.8858	0.1803	0.7372	0.1887	0.2562	0.7687	0.2221
DeepNitro	0.5982	0.8873	0.9082	0.2831	0.8073	0.2837	0.2344	0.7544	0.2193
Deep-Kcr	0.5004	0.8445	0.8734	0.2009	0.7307	0.1944	0.2575	0.7625	0.2257
DeepSuccinylSite	0.6422	0.9011	0.9181	0.2272	0.7753	0.2136	0.1629	0.6845	0.1790
Adapt-Kcr	<u>0.6426</u>	<u>0.9020</u>	<u>0.9188</u>	0.2360	0.7782	0.2258	0.2064	0.7311	0.1999
BERT-Kcr	0.5649	0.8660	0.8942	0.2123	0.7575	0.2190	0.2149	0.7489	0.2284
LMSuccSite	0.6222	0.8914	0.9078	<u>0.2990</u>	<u>0.8169</u>	<u>0.2966</u>	<u>0.3060</u>	<u>0.8112</u>	<u>0.2850</u>
PTM-CMGMS	**0.7005**	**0.9245**	**0.9386**	**0.3072**	**0.8306**	**0.3024**	**0.3398**	**0.8369**	**0.3136**

Table 2: Comparison results of PTM-CMGMS and baselines on three datasets (Crotonylation, Succinylation, and Nitrosylation). Note that the highest score in each column is in bold and the second-best score is underlined.
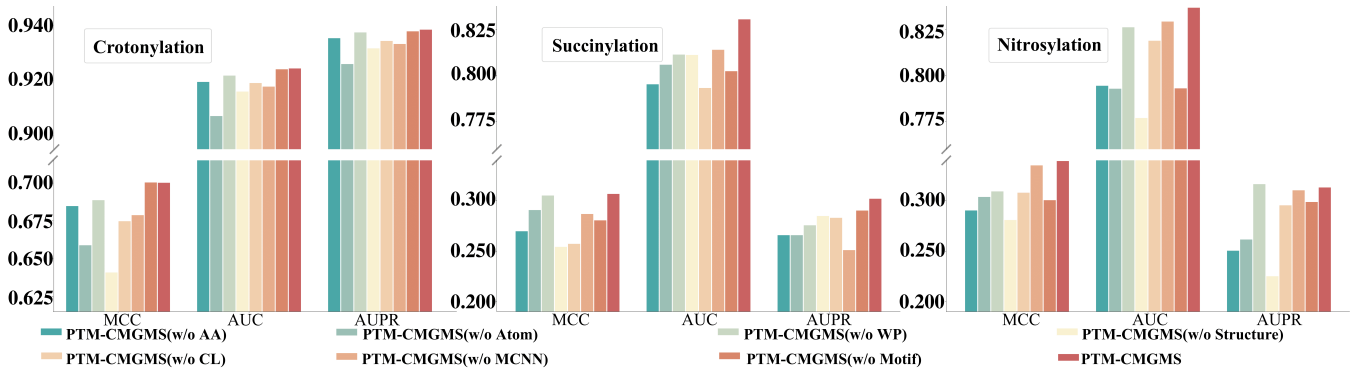

Figure 3: Results of PTM-CMGMS and its variants on three datasets.

protein granularity in the multi-granularity structure-aware representation learning.

- PTM-CMGMS without contrastive learning (w/o CL) removes representation optimization based on contrastive learning in the multi-granularity structure-aware representation learning.

- PTM-CMGMS without structure information (w/o Structure) removes multi-granularity structure-aware representation learning.

- PTM-CMGMS without multi-scale CNN (w/o MCNN) removes the multi-scale CNN in multi-scale sequence representation learning.

- PTM-CMGMS without Motif-based information extraction (w/o Motif) removes the motif-based information extraction in multi-scale sequence representation learning.

As shown in Figure 3, PTM-CMGMS with the integration of multi-granularity structure and multi-scale sequence representation learning achieves superior performances, and the removal of any will undermine the predictive capacity of PTM-CMGMS on most datasets. Besides, we have the following observations: (1) PTM-CMGMS outperforms the PTM-CMGMS(w/o Structure) and PTM-CMGMS(w/o MCNN), which demonstrates that the combination of multi-granularity structure-aware information and multi-scale sequence information helps improve the performance of PTM site prediction. In addition, PTM-CMGMS(w/o Structure) gets an obvious performance drop, indicating the prediction performance is boosted mostly by the structure information and multi-granularity structure-aware representation learning is the core component of the model architecture. (2) PTM-CMGMS outperforms the PTM-CMGMS(w/o AA), and PTM-CMGMS(w/o WP), which demonstrates that structure representations at these granularities are all helpful for PTM site prediction. Note that PTM-CMGMS outperforms the PTM-CMGMS without the information of atom granularity (PTM-CMGMS(w/o Atom)), which verifies the rationality to model the PTM site into finer-grained atom level. These granularity information provide a more comprehensive characterization of the structure feature of PTM sites. (3) PTM-CMGMS(w/o CL) also gets a performance drop on all metrics, which shows that the discriminative structure representations obtained by contrastive learning can provide useful information for PTM site prediction. (4) PTM-CMGMS performs better than PTM-CMGMS(w/o Motif), which implies that the motif generated by aligning all context sequences of PTM sites can assist the prediction.

To explore the effectiveness of the combination of BiLSTM, Transformer, and MCNN (BiL-

STM+Transformer+MCNN), we also investigate the performance of various combination of context sequence information extraction architectures on the Crotonylation dataset, including BiLSTM+Transformer, Transformer+MCNN, BiLSTM+MCNN, and BiLSTM+Transformer+2DCNN that replaces the MCNN with a 2DCNN. The PLMs pre-trained on large-scale protein sequence corpora have achieved impressive performance on various downstream protein understanding tasks, thus ESM(Lin et al. 2022) and ProtTrans are also within our consideration. Results in Figure 4 show that BiLSTM+Transformer+MCNN achieves the best performance, which indicates that the combination of BiLSTM, Transformer, and MCNN can comprehensively capture sequence information to promote PTM site prediction. Besides, we have the following observation: (1) BiLSTM+Transformer+MCNN is superior to PLMs (ESM and ProtTrans), which demonstrates the effectiveness of our designed multi-scale sequence information extraction strategy on PTM site prediction. (2) BiLSTM+Transformer+MCNN outperforms any combination of pairwise architectures, this may be due to the complementary strengths of the three architectures. (3) BiLSTM+Transformer+2DCNN that replaces the MCNN with a 2DCNN gets a performance drop on all metrics, suggesting that MCNN may be more suitable for extracting local information of context sequences for PTM site prediction.
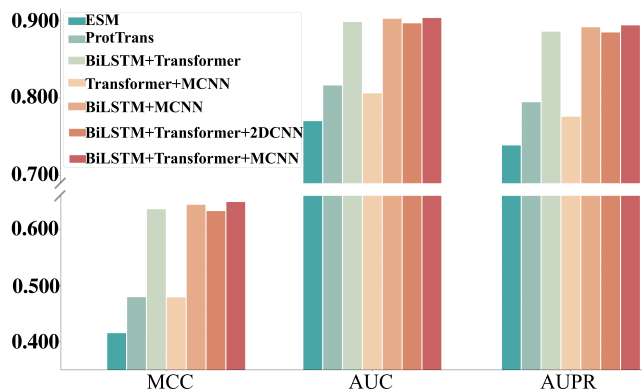

Figure 4: Results of various context sequence information extraction architectures on the Crotonylation dataset.

Effectiveness of Structure Information for PTM Site Prediction

To verify the effectiveness of structure information for predicting PTM sites, we compare the performance of PTM-CMGMS and Adapt-Kcr (a baseline that achieves the best performance on the Crotonylation dataset) in dealing with PTM sites with different numbers of non-local contact residues. Residues are considered as non-local contact with the PTM sites if they are separated with greater than 20 residues in the primary structure and their spatial distance is less than 12Å (Yuan et al. 2022; Zeng et al. 2023b). Structure information of PTM sites mainly refers to these non-local contact residues. We have the following conclusions:

(1) PTM-CMGMS consistently surpasses Adapt-Kcr when the number of non-local contact residues is greater than 0, which suggests that using of structure information can effectively enhance the prediction for PTM sites with multiple non-local contact residues. (2) PTM-CMGMS is also better than Adapt-Kcr when the number of non-local contact residues is equal to 0 (i.e. the structure information of PTM sites is relatively insufficient), implying the advantage of integrating both aspects of structure and sequence information. These results demonstrate the effectiveness of structure information for PTM site prediction.
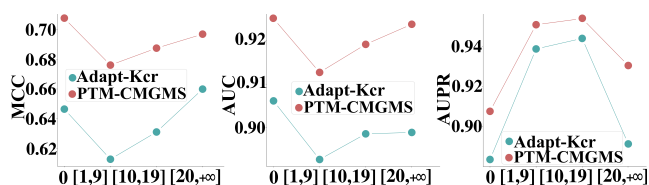

Figure 5: Results of PTM-CMGMS and Adapt-Kcr in dealing with PTM sites with different numbers of non-local contact residues on the Crotonylation dataset.

Conclusion

In this paper, we propose a PTM site prediction method (PTM-CMGMS), which couples the multi-granularity structure and multi-scale sequence representation. PTM-CMGMS designs a multi-granularity structure-aware representation learning(including the atom, amino acid, and whole protein granularity) to capture neighborhood structure representations of PTM sites based on AlphaFold predicted structures, and adopts contrastive learning to optimize structure representations. Further, PTM-CMGMS leverages a multi-scale sequence representation learning to extract context sequence information of PTM sites, and utilizes the motif generated by aligning all context sequences of positive sites to assist the prediction. Experimental results demonstrate the effectiveness of PTM-CMGMS and the importance of multi-granularity structure and multi-scale sequence information for PTM site prediction.

In future work, we have several directions to optimize the PTM site prediction method, such as considering crosstalk between different types of PTMs and providing interpretable PTM site prediction method.

Acknowledgments

References

Bludau, I.; Willems, S.; Zeng, W.-F.; Strauss, M. T.; Hansen, F. M.; Tanzer, M. C.; Karayel, O.; Schulman, B. A.; and Mann, M. 2022. The structural context of posttranslational modifications at a proteome-wide scale. *PLoS biology*, 20(5): e3001636.

Bos, J.; and Muir, T. W. 2018. A chemical probe for protein crotonylation. *Journal of the American Chemical Society*, 140(14): 4757–4760.

Cai, W.; Xu, D.; Zeng, C.; Liao, F.; Li, R.; Lin, Y.; Zhao, Y.; Dong, W.; Wang, Q.; Yang, H.; et al. 2022. Modulating lysine crotonylation in cardiomyocytes improves myocardial outcomes. *Circulation Research*, 131(5): 456–472.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. 2021. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10): 7112–7127.

Fang, Y.; Xu, X.; Ding, J.; Yang, L.; Doan, M. T.; Karmaus, P. W.; Snyder, N. W.; Zhao, Y.; Li, J.-L.; and Li, X. 2021. Histone crotonylation promotes mesoendodermal commitment of human embryonic stem cells. *Cell Stem Cell*, 28(4): 748–763.

Fernando, V.; Zheng, X.; Walia, Y.; Sharma, V.; Letson, J.; and Furuta, S. 2019. S-nitrosylation: an emerging paradigm of redox signaling. *Antioxidants*, 8(9): 404.

Gao, J.; Prlić, A.; Bi, C.; Bluhm, W. F.; Dimitropoulos, D.; Xu, D.; Bourne, P. E.; and Rose, P. W. 2017. BioJava-ModFinder: identification of protein modifications in 3D structures from the Protein Data Bank. *Bioinformatics*, 33(13): 2047–2049.

Hasan, M. M.; and Kurata, H. 2018. GPSuc: Global Prediction of Generic and Species-specific Succinylation Sites by aggregating multiple sequence features. *PLoS one*, 13(10): e0200283.

Hasan, M. M.; Manavalan, B.; Khatun, M. S.; and Kurata, H. 2019. Prediction of S-nitrosylation sites by integrating support vector machines and random forest. *Molecular omics*, 15(6): 451–458.

Hasan, M. M.; Yang, S.; Zhou, Y.; and Mollah, M. N. H. 2016. SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties. *Molecular BioSystems*, 12(3): 786–795.

He, W.; Wang, Y.; Cui, L.; Su, R.; and Wei, L. 2021. Learning embedding features based on multisense-scaled attention architecture to improve the predictive performance of anticancer peptides. *Bioinformatics*, 37(24): 4684–4693.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Jia, J.; Liu, Z.; Xiao, X.; Liu, B.; and Chou, K.-C. 2016a. iSuc-PseOpt: identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. *Analytical biochemistry*, 497: 48–56.

Jia, J.; Liu, Z.; Xiao, X.; Liu, B.; and Chou, K.-C. 2016b. pSuc-Lys: predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. *Journal of theoretical biology*, 394: 223–230.

Ju, Z.; and He, J.-J. 2017. Prediction of lysine crotonylation sites by incorporating the composition of k-spaced amino acid pairs into Chou's general PseAAC. *Journal of Molecular Graphics and Modelling*, 77: 200–204.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589.

Kabsch, W.; and Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12): 2577–2637.

Lee, T.-Y.; Chen, Y.-J.; Lu, T.-C.; Huang, H.-D.; and Chen, Y.-J. 2011. SNOSite: exploiting maximal dependence decomposition to identify cysteine S-nitrosylation with substrate site specificity. *PLoS one*, 6(7): e21849.

Li, F.; Fan, C.; Marquez-Lago, T. T.; Leier, A.; Revote, J.; Jia, C.; Zhu, Y.; Smith, A. I.; Webb, G. I.; Liu, Q.; et al. 2020. PRISMOID: a comprehensive 3D structure database for post-translational modifications and mutations with functional impact. *Briefings in bioinformatics*, 21(3): 1069–1079.

Li, Z.; Fang, J.; Wang, S.; Zhang, L.; Chen, Y.; and Pian, C. 2022. Adapt-Kcr: a novel deep learning framework for accurate prediction of lysine crotonylation sites based on learning embedding features and attention architecture. *Briefings in Bioinformatics*, 23(2): bbac037.

Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; dos Santos Costa, A.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; et al. 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv*, 2022: 500902.

Liu, Y.; Yu, Z.; Chen, C.; Han, Y.; and Yu, B. 2020. Prediction of protein crotonylation sites through LightGBM classifier based on SMOTE and elastic net. *Analytical biochemistry*, 609: 113903.

Lv, H.; Dao, F.-Y.; Guan, Z.-X.; Yang, H.; Li, Y.-W.; and Lin, H. 2021. Deep-Kcr: accurate detection of lysine crotonylation sites using deep learning method. *Briefings in bioinformatics*, 22(4): bbaa255.

Ning, Q.; and Li, J. 2022. DLF-Sul: a multi-module deep learning framework for prediction of S-sulfinylation sites in proteins. *Briefings in Bioinformatics*, 23(5): bbac323.

Ning, Q.; Zhao, X.; Bao, L.; Ma, Z.; and Zhao, X. 2018. Detecting succinylation sites from protein sequences using ensemble support vector machine. *BMC bioinformatics*, 19(1): 1–9.

Pokharel, S.; Pratyush, P.; Heinzinger, M.; Newman, R. H.; and Kc, D. B. 2022. Improving protein succinylation sites prediction using embeddings from protein language model. *Scientific Reports*, 12(1): 16933.

Pratyush, P.; Pokharel, S.; Saigo, H.; and Kc, D. B. 2023. pLMSNOSite: an ensemble-based approach for predicting protein S-nitrosylation sites by integrating supervised word embedding and embedding from pre-trained protein language model. *BMC bioinformatics*, 24(1): 41.

Qiao, Y.; Zhu, X.; and Gong, H. 2022. BERT-Kcr: prediction of lysine crotonylation sites by a transfer learning method with pre-trained BERT models. *Bioinformatics*, 38(3): 648–654.

Qiu, W.-R.; Sun, B.-Q.; Tang, H.; Huang, J.; and Lin, H. 2017. Identify and analysis crotonylation sites in histone by using support vector machines. *Artificial intelligence in medicine*, 83: 75–81.

Thapa, N.; Chaudhari, M.; McManus, S.; Roy, K.; Newman, R. H.; Saigo, H.; and Kc, D. B. 2020. DeepSuccinylSite: a deep learning based approach for protein succinylation site prediction. *BMC bioinformatics*, 21(3): 1–10.

Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; et al. 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1): D439–D444.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wei, W.; Liu, X.; Chen, J.; Gao, S.; Lu, L.; Zhang, H.; Ding, G.; Wang, Z.; Chen, Z.; Shi, T.; et al. 2017. Class I histone deacetylases are major histone decrotonylases: evidence for critical and broad function of histone crotonylation in transcription. *Cell research*, 27(7): 898–915.

Xie, Y.; Luo, X.; Li, Y.; Chen, L.; Ma, W.; Huang, J.; Cui, J.; Zhao, Y.; Xue, Y.; Zuo, Z.; et al. 2018. DeepNitro: prediction of protein nitration and nitrosylation sites by deep learning. *Genomics, proteomics & bioinformatics*, 16(4): 294–306.

Xu, Y.; Ding, J.; Wu, L.-Y.; and Chou, K.-C. 2013. iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PloS one*, 8(2): e55844.

Xu, Y.; Ding, Y.-X.; Ding, J.; Lei, Y.-H.; Wu, L.-Y.; and Deng, N.-Y. 2015. iSuc-PseAAC: predicting lysine succinylation in proteins by incorporating peptide position-specific propensity. *Scientific reports*, 5(1): 10184.

Xue, Y.; Liu, Z.; Gao, X.; Jin, C.; Wen, L.; Yao, X.; and Ren, J. 2010. GPS-SNO: computational prediction of protein S-nitrosylation sites with a modified GPS algorithm. *PloS one*, 5(6): e11290.

Yang, Y.; and Gibson, G. E. 2019. Succinylation links metabolism to protein functions. *Neurochemical research*, 44: 2346–2359.

Yu, H.; Bu, C.; Liu, Y.; Gong, T.; Liu, X.; Liu, S.; Peng, X.; Zhang, W.; Peng, Y.; Yang, J.; et al. 2020. Global crotonylome reveals CDYL-regulated RPA1 crotonylation in homologous recombination–mediated DNA repair. *Science advances*, 6(11): eaay4697.

Yuan, Q.; Chen, S.; Rao, J.; Zheng, S.; Zhao, H.; and Yang, Y. 2022. AlphaFold2-aware protein–DNA binding site prediction using graph transformer. *Briefings in Bioinformatics*, 23(2): bbab564.

Zeng, Y.; Wei, Z.; Yuan, Q.; Chen, S.; Yu, W.; Lu, Y.; Gao, J.; and Yang, Y. 2023a. Identifying B-cell epitopes using AlphaFold2 predicted structures and pretrained language model. *Bioinformatics*, 39(4): btad187.

Zeng, Y.; Wei, Z.; Yuan, Q.; Chen, S.; Yu, W.; Lu, Y.; Gao, J.; and Yang, Y. 2023b. Identifying B-cell epitopes using AlphaFold2 predicted structures and pretrained language model. *Bioinformatics*, 39(4): btad187.

Zhang, X.; Wei, L.; Ye, X.; Zhang, K.; Teng, S.; Li, Z.; Jin, J.; Kim, M. J.; Sakurai, T.; Cui, L.; et al. 2023. SiameseCPP: a sequence-based Siamese network to predict cell-penetrating peptides by contrastive learning. *Briefings in Bioinformatics*, 24(1): bbac545.

Zhang, Z.; Xu, M.; Jamasb, A.; Chenthamarakshan, V.; Lozano, A.; Das, P.; and Tang, J. 2022. Protein representation learning by geometric structure pretraining. *arXiv preprint arXiv:2203.06125*.