IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 5, 2021, accepted March 21, 2021, date of publication March 24, 2021, date of current version April 6, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3068413

DeepKcrot: A Deep-Learning Architecture for General and Species-Specific Lysine Crotonylation Site Prediction

XILIN WEI[1], YUTONG SHA[1], YIMING ZHAO[1], NINGNING HE[2], AND LEI LI[1,2]
[1]School of Data Science and Software Engineering, Qingdao University, Qingdao 266021, China
[2]School of Basic Medicine, Qingdao University, Qingdao 266021, China

Corresponding author: Lei Li (leili@qdu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 31770821 and Grant 32071430.

ABSTRACT Lysine crotonylation (Kcrot), as a post-translational modification (PTM) originally identified at histone proteins, is involved in diverse biological processes. Several conventional machine-learning (ML) predictors were developed based on the Kcrot sites from histone proteins. Recently, thousands of Kcrot sites have been experimentally verified on non-histone proteins from multiple species. Accordingly, a few predictors have been developed for predicting the Krot sites for specific organisms (i.e. humans and papaya). Nevertheless, there is a lack of research on the comparison of the crotonylomes of different organisms. Here, we collected around 20,000 Kcrot sites experimentally identified from four different species as the benchmark data set. We present the deep-learning (DL) architecture dubbed DeepKcrot for predicting Kcrot sites on the proteomes across various species. DeepKcrot includes species-specific and general classifiers using a convolutional neural network with the word embedding (CNN_{WE}) encoding approach. CNN_{WE} performs better than both the traditional ML-based and other DL-based classifiers in terms of ten-fold cross-validation and independent test, independent of the size of the training set. Additionally, cross-species performance for each species-specific predictor is not as good as the self-species performance whereas the cross-species performance generally increases with the size of the training dataset. Moreover, the predictors developed based on the non-histone Kcrot sites are unsuccessful for the histone Kcrot prediction, suggesting that the Kcrot-containing peptides from non-histone and histone proteins have significantly different characteristics and data integration is required. Overall, DeepKcrot is an efficient prediction tool and freely available at http://www.bioinfogo.org/deepkcrot.

INDEX TERMS Deep learning, convolutional neural network, lysine crotonylation, non-histone protein, random forest.

I. INTRODUCTION

Lysine crotonylation (Kcrot) is a conserved type of PTMs and it was originally found on histone proteins [1]. Histone crotonylation affects chromatin structure and gene expression [1]–[4]. Recently, it has been discovered on non-histone proteins and involved in various cellular activities [5], [6]. The rapid progress in the development of the state-of-the-art techniques led to the identification of thousands of Kcrot sites from different species through affinity enrichment followed by high-throughput mass spectrometry. 10,163 Kcrot sites on 2445 non-histone proteins were determined from human A549 cells [7] and 2696 Kcrot sites on 1024 non-histone proteins were identified from the human H1299 cell line [8]. Besides, 5995/1265/2044 different Kcrot sites were experimentally verified from Carica papaya L. (papaya)/Oryza sativa L. japonica (rice)/Nicotiana tabacum (tabacum), respectively [9]–[11]. Recently, CDYL-regulated crotonylome was investigated in Hela cell lines [12].

To understand and elucidate modification kinetics and molecular mechanisms of lysine crotonylation, a fundamental but important step is to accurately predict the crotonylation sites. Currently, five *in-silico* approaches were developed based on the histone Kcrot sites [13]–[17]. Although these algorithms have made great contributions to the Krot prediction, they fail to identify non-histone Kcrot sites [18]. Additionally, we collected the papaya Kcrot sites and developed a panel of classifiers for the Kcrot prediction [18].

The associate editor coordinating the review of this manuscript and approving it for publication was Liangxiu Han.

Among the traditional ML approaches, the random forest (RF) model with the Enhanced Grouped Amino Acid Composition (EGAAC) encoding feature, dubbed RF_{EGAAC}, had the best performance [18]. Additionally, the one-dimensional Convolutional Neural Network (CNN) with the word-embedding (WE) encoding approach, named CNN_{WE}, showed superior performance in all the models [18]. Moreover, Wang and coworkers used our collected papaya Kcrot sites and a limited number (167) of mammalian Kcrot sites for the construction of Kcrot predictors using the RF and SVM (support vector machine) architectures with the combination of different features [19]. Lv *et al.* [20] developed a DL model called Deep-Kcr based on experimentally verified human crotonylome [12]. These developed models were mainly based on either the crotonylome of the specific organism (e.g. papaya or humans) or a limited number of proteins (e.g. histones). With the identification of thousands of Kcrot sites from various species, it is of interest to study the diversity of crotonylomes across the different organisms and compare the performance of the developed methods and investigate whether any other model with better performance than the previously developed models.

In this study, we constructed the Long Short-Term Memory (LSTM) model and compared it with our previously developed models including RF_{EGAAC} and CNN_{WE} models. We found that the CNN_{WE} model still showed the best performance. Additionally, the CNN_{WE} model compared favourably to the reported model Deep-Kcr. Moreover, we constructed DeepKcrot based on the CNN_{WE} architecture that included four orgasm-specific predictors and a general predictor. We find that cross-species performances for species-specific CNN_{WE} predictors are not as good as the self-species performance. The general CNN_{WE} predictor based on the integration of the training data from different species shows superior performance to the species-specific predictors except for one organism. Overall, the general CNN_{WE} models have excellent performance for predicting Kcrot sites on proteomes across different species.

II. MATERIALS AND METHODS

A. DATA COLLECTION AND PREPROCESSING

We collected 10,702/1265/2044/5995 Kcrot sites on non-histone proteins from human/rice/tabacum/papaya, respectively [9]–[11]. We took the human species as an example to describe the data preprocessing. To prepare the benchmark data sets with high confidence for training and testing, we referred to the procedure established by Chen *et al.* [21], [22].

The 10,702 Kcrot sites from the 2836 human proteins were considered as positive sites, and the remaining lysine residues (775,123) on the same proteins were deemed as negative sites. The 2836 proteins with sequence identities $>30\%$ were classified into 2064 clusters using CD-HIT [23]. In each cluster, the protein with the largest number of Kcrot sites remained as the representative, in which the Kcrot sites were considered as positive sites and the rest lysine sites were taken as negative sites. Note that the lysine sites in the representative were removed if the aligned counterparts from other members of the same cluster can be crotonylated. According to our previous study [18], the optimal sequence window for model construction was 29. Accordingly, the dataset contained 8,170 positive sites and 76,673 negative sites from 2064 representatives. The representatives were randomly divided into two groups: 4/5 (1651) for cross-validation and the rest 1/5 (413) for an independent test. Finally, the cross-validation data set contained 6687 positives and 67,106 negatives, and the independent test dataset contained 1483 positives and 16,497 negatives (Figure 1). The same data preprocessing was performed for papaya, rice and tabacum, respectively (Figure 1).

B. CONVENTIONAL MACHINE LEARNING ALGORITHMS

The RF algorithm was selected and trained with the EGAAC feature by randomly generating 1600 decision trees. In the EGAAC feature, the types of amino acids were categorized into five groups (g1: GAVLMI, g2:FYW, g3: KRH, g4: DE and g5: STCPNQ) according to their physicochemical properties and the frequencies of the groups were calculated in the window of fixed length (the default value is 5) continuously sliding from the N- to C-terminal of each peptide sequence.

C. DEEP LEARNING ALGORITHMS

We constructed a DL framework based on a one-dimensional CNN with the WE encoding approach [18]. Figure 2 showed that this framework included the five layers: the input layer, the embedding layer, the convolutional layer, the fully connected layer and the output layer. These layers were described in our previous study [18]. In the embedding layer, each type of amino acid was converted into a predefined certain dimension word vector. The parameters in the vectors were updated with subsequent layers during the learning process under the supervision of a class label. We investigated the effect of the dimension size on the prediction performance (Table 1). Within the range of the dimension from three to seven, the prediction performance increased from three to five and reach the plateau starting from five. Therefore, we chose the dimension of the word vector as five.

We also constructed the Long Short-Term Memory (LSTM) model with the WE encoding approach. This model contained five layers (Figure 3).

1) *The Input Layer:* Each peptide segment is converted into an integer vector with the NUM encoding approach, where each type of amino acid residues was mapped to a different integer [24].
2) *The Word Embedding Layer:* Each integer of the vector from the input layer is encoded into a predefined five-dimension word vector.
3) *The LSTM Layer:* Each of the word vectors is input sequentially into the LSTM cell that contained 32 hidden neuron units.

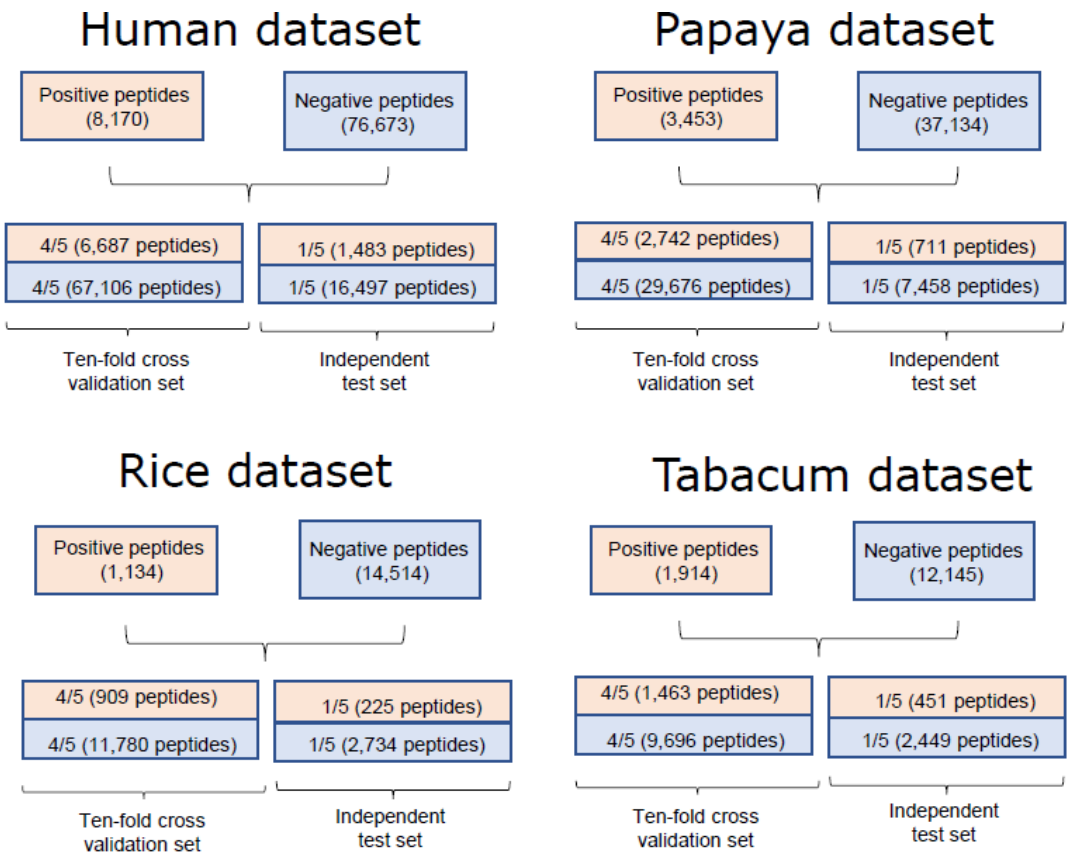

FIGURE 1. Separation of the datasets for the four organisms into the cross-validation set and the independent test set.

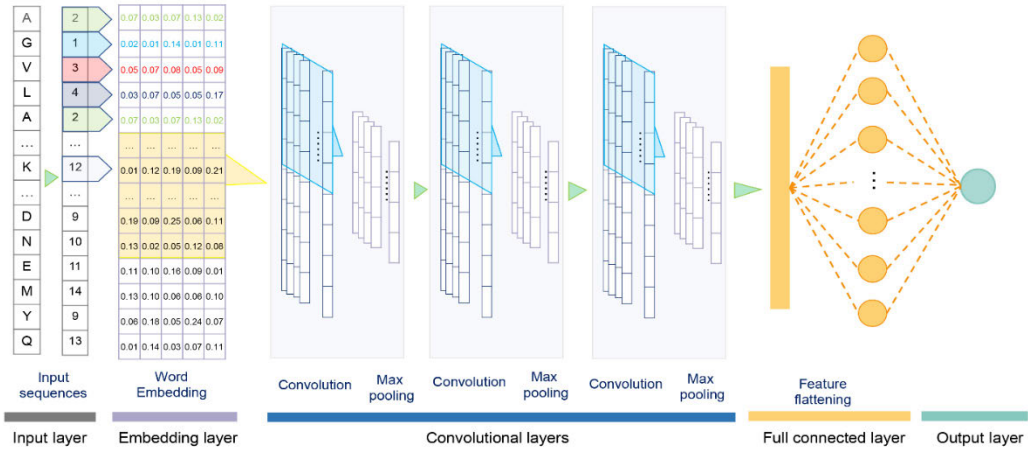

FIGURE 2. The architecture of CNN_{WE}. It contained five layers. The input layer received a peptide sequence of 29 residues with K in the center. In the embedding layer, each residue of the sequence was converted into a five-dimensional word vector. In the convolution layer, 29 five-dimension word vectors were input into the CNN cell that contained 128 hidden neuron units. In the fully connected layer, 128 neuron units were built in which the ReLU was chosen for its activation function. The last layer included a single unit outputting the prediction scores.

4) *The Dense Layer:* It contains a single dense sublayer that has 16 neurons with the ReLU activation function.
5) *The Output Layer:* This layer has only one neuron activated by sigmoid function, outputting the probability of the Kcrot modification.

The parameters in the DL models were trained and optimized using the Adam algorithm. The dropout rate was set as 0.5 to avoid overfitting. We set the learning rate as 0.001, determined using the maximum number of epochs as 500. The early-stopping strategy was applied, where the training process was stopped early if the performance did not improve within 50 epochs.

D. PERFORMANCE ASSESSMENT OF THE PREDICTORS

Four measurements of accuracy (Acc), sensitivity (Sn), specificity (Sp) and Mathew Correlation Coefficient (MCC)

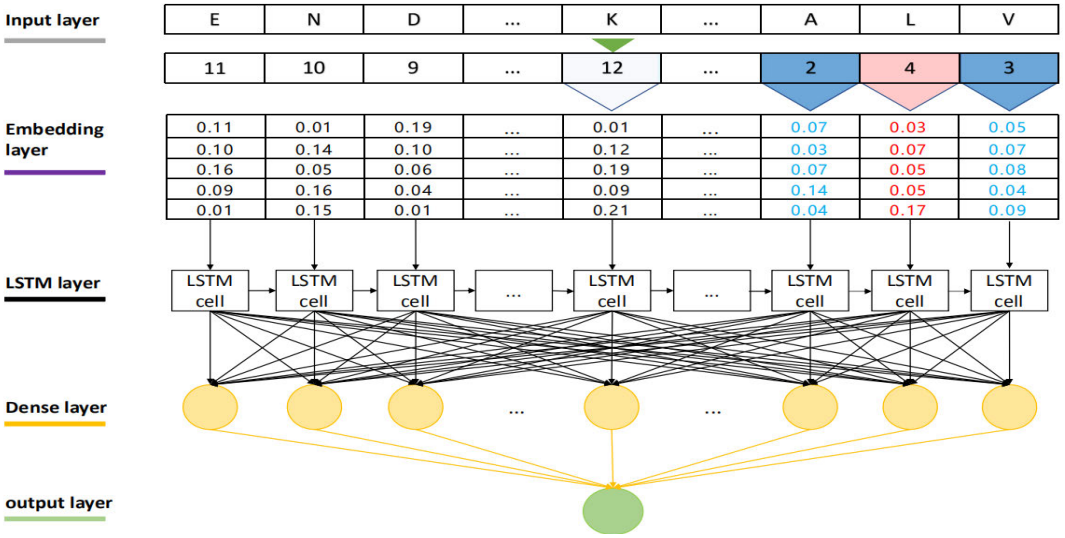

FIGURE 3. The architecture of $LSTM_{WE}$.

TABLE 1. Performances of the CNN_{WE} model with different dimensions of the word vector in terms of the human independent test.

Dimension	AUC*
3	0.852±0.002
4	0.858±0.002
5	0.861±0.001
6	0.860±0.002
7	0.861±0.001

*AUC means Area Under The Curve.

were calculated. They are defined as:

$$Sn = \frac{TP}{TP + FN} \tag{1}$$

$$Sp = \frac{TN}{TN + FP} \tag{2}$$

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \tag{3}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{4}$$

where TP/TN is the number of the correctly predicted Kcrot/K sites, separately, whereas FP/FN is the number of the Kcrot/K sites incorrectly predicted, respectively.

For each algorithm, a ten-fold cross-validation test was performed. The ROC curves were illustrated for Sn vs. 1-Sp scores and the AUC values were calculated. The area under the ROC curve with <10% false-positive rate (AUC01) was considered because it reflects the performance of the predictor in a low false-positive rate, which is significant in a practical application.

III. RUSULT AND DISCUSSTIONS

A. THE CNN APPROACH WITH WORD EMBEDDING SHOWED SUPERIOR PERFORMANCE

We collected from literature 10,702/1265/2044/5995 non-redundant Kcrot sites experimentally verified from human/rice/tabacum/papaya, respectively [7]–[11]. For each organism, we first eliminated homologous Kcrot sites and conducted the species-specific dataset. The species-specific dataset was further separated into two groups: 4/5 for ten-fold cross-validation and the rest 1/5 for an independent test (see Methods for details, Figure 1). For instance, the cross-validation dataset for the human species contained 6687 positives and 67,106 negatives while the independent dataset covered 1483 positives and 16,497 negatives. As the largest dataset is derived from the human species, our study focused on the human species followed by the expansion to other species.

Many computational approaches have been developed for the prediction of PTM sites. They are generally based on different ML algorithms combined with various pre-defined features encoded from peptide sequences [25]. The RF algorithm is widely applied to the PTM prediction as it is robust and insensitive to data imbalance [24]. We ever compared the effect of the imbalanced dataset on the potential overfitting of the classifiers and found that the RF model constructed using an imbalanced training dataset had a similar performance to that built using a balanced training dataset [24]. According to our previous study on the papaya proteome, we constructed RF-based predictors with the EGAAC encoding scheme, dubbed RF_{EGAAC} that showed the best performance in the RF-based models [18].

Deep learning algorithms have recently been applied to the field of modification prediction and have shown superior performance to traditional ML algorithms [21], [26]. We ever applied the CNN models for the prediction of the papaya

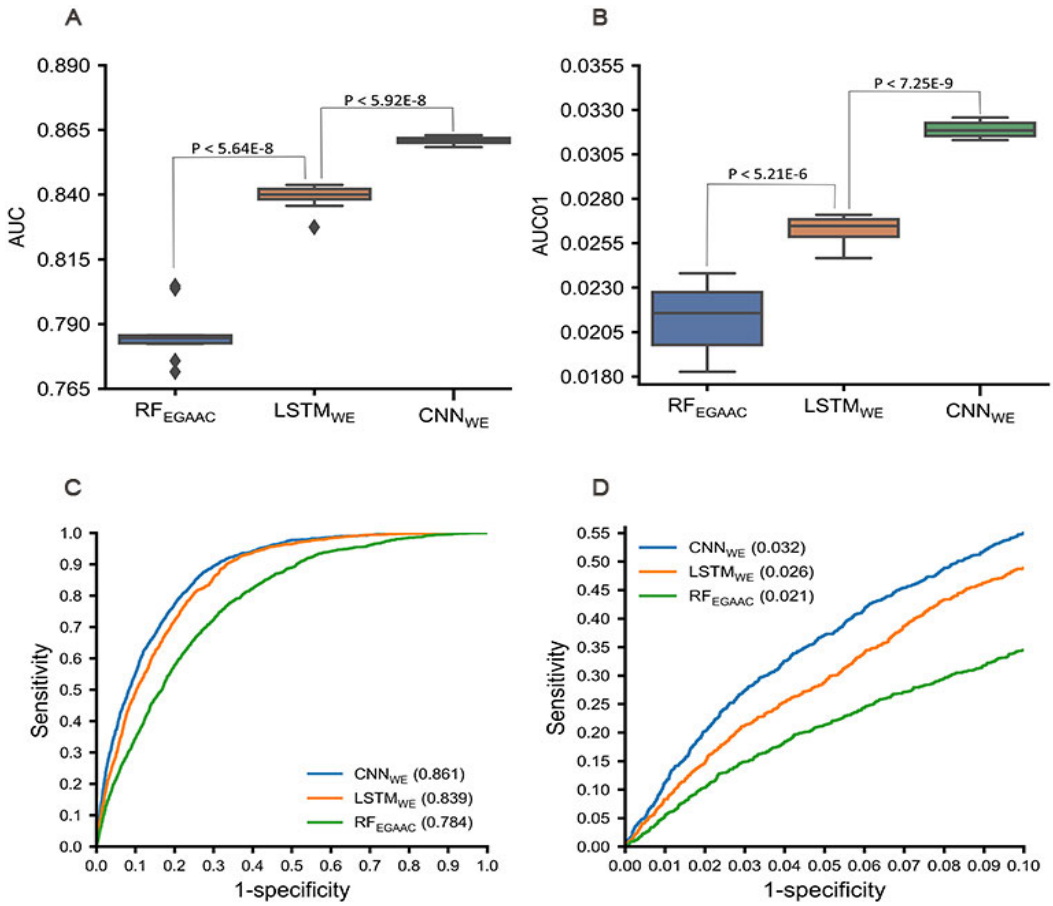

FIGURE 4. Performance comparison of the Kcrot predictors. The performances of different predictors constructed for human species were compared in terms of AUC (A) and AUC01 (B), respectively, for ten-fold cross-validation. AUC (C) and AUC01 (D) curves were also generated using the independent test. P values were calculated using a paired Student's t-test. A detailed performance comparison using different measurements is provided in Table 2.

Kcrot sites and the CNN model with the word embedding encoding approach compared favourably to other CNN-based approaches [18]. Additionally, we developed here the classifier based on long short-term memory (LSTM) with word embedding named LSTM_{WE}, which was previously constructed to predict Cysteine S-Sulphenylation Sites and had better performance than CNN_{WE} [24].

Among the three models, CNN_{WE} performed the best in the prediction of human Kcrot sites for both the ten-fold cross-validation and independent test, followed by LSTM_{WE} and RF_{EGAAC} ($P < 5.92 \times 10^{-8}$ for CNN_{WE} and LSTM_{WE}; $P < 5.64 \times 10^{-8}$ for LSTM_{WE} and RF_{EGAAC}; Figure 4 & Table 2). For instance, the MCC value and AUC value for CNN_{WE} are 0.342 and 0.864 in terms of cross-validation (Figure 4 & Table 2). As prediction performance at a low false-positive rate is highly useful in practice, we applied AUC01, in which the specificity was determined to be >90%, to the estimation of these predictors. CNN_{WE} again compared favourably to other models in terms of the cross-validation test as well as the independent test (Figure 4 & Table 2). Because CNN_{WE} showed its superior performance for the Kcrot prediction in two different proteomes (*ie*, human and papaya [18]), we concluded that CNN_{WE} was the best and robust model for the Kcrot prediction.

To understand the performance of the CNN_{WE} model, we visualized the sample distributions from the outputs of the embedding layer and the last convolutional layer of the human CNN_{WE} model using the t-SNE algorithm [27], based on the independent dataset (Figure 5A&5B). In the word embedding layer, all the samples were mixed (Figure 5A), whereas the positive and negative samples were separated after the convolutional operation (Figure 5B). This comparison indicates that the distinctive features of the positives and negatives were detected by the convolutional layer, and our CNN_{WE} model could produce a deep representation that is more discriminating than the original input sequences.

The word embedding approach is widely applied to the natural language process, in which each word is converted into a low-dimension vector. This approach avoids a sparse vector space and infers the semantic similarity of words. We applied this concept to peptide sequences. Each amino acid was converted into a five-dimension word vector in the embedding layer. Finally, a 20 × 5 matrix was generated after training where every row represented a five-dimensional

TABLE 2. Performances of the different classifiers for the human organism.

	Classifier[2]	Acc[3]	Sn[3]	Sp[3]	MCC[3]	AUC[3]	AUC01[3]
Ten-fold cross-validation[1]	RF_{EGAAC}	0.853±0.002	0.383±0.026	0.90	0.245±0.016	0.791±0.008	0.022±0.002
	$LSTM_{WE}$	0.863±0.001	0.458±0.010	0.90	0.294±0.009	0.839±0.015	0.027±0.002
	CNN_{WE}	**0.871±0.005**	**0.537±0.008**	**0.90**	**0.342±0.009**	**0.864±0.002**	**0.033±0.001**
Independent test[1]	RF_{EGAAC}	0.851±0.002	0.365±0.021	0.90	0.228±0.020	0.784±0.013	0.021±0.002
	$LSTM_{WE}$	0.860±0.004	0.462±0.008	0.90	0.306±0.010	0.839±0.005	0.026±0.001
	CNN_{WE}	**0.869±0.001**	**0.524±0.008**	**0.90**	**0.338±0.005**	**0.861±0.001**	**0.032±0.000**

Note: [1] The datasets for ten-fold cross-validation and an independent test were derived from experimentally verified Kcrot-containing peptides. The details of the two datasets are described in Materials and Methods.
[2] The RF classifier with the EGAAC approach was named RF_{EGAAC}.
The CNN classifier with the word embedding approach was named CNN_{WE}, and another classifier was called $LSTM_{WE}$.
[3]Acc=accuracy, Sn=sensitivity, Sp=specificity, MCC=Matthew's Correlation Coefficient, AUC=area under the receiver operating characteristic, AUC01 = AUC with a <10% false-positive rate (*i.e.*, specificity>90%). Sp was fixed as 0.9 across the models to fairly compare other measures (i.e. Acc, Sn and MCC) of the models.

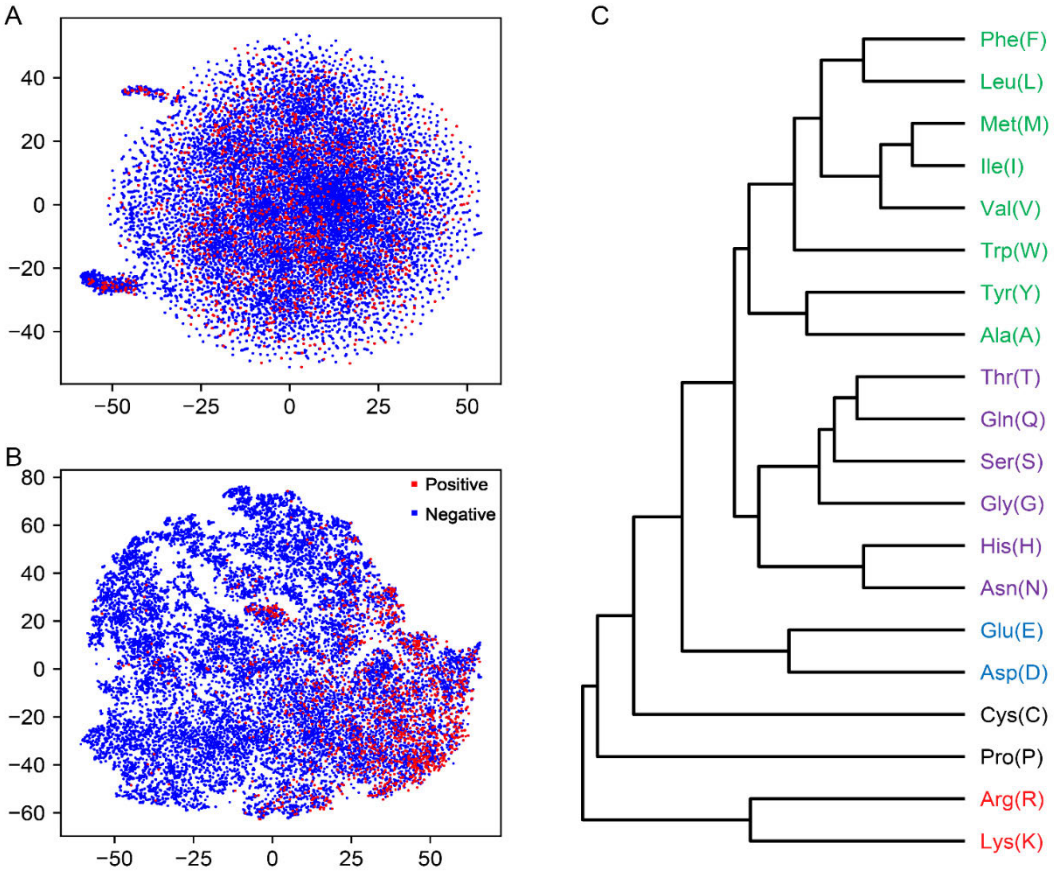

FIGURE 5. T-SNE visualization of the sample distributions and classification of the amino acids based on the information from the human CNN_{WE} model. The t-SNE visualization of the output of the embedding layer (A) and the last convolutional layer (B) of the human CNN_{WE} model. (C) Hierarchical clustering of the 20 residues based on their related five-dimensional word vectors in the embedding layer and the calculation of Euclidean distance in the average linkage. The residues were grouped into four major groups: (i) the alkaline residues K and R (red colour), (ii) the amino acids with negative charged side chains D and E (blue colour), (iii) the hydrophobic amino acids F, L, M, I, V, W, Y and A (green colour); (iv) the mainly polar uncharged residues T, Q, S, G, H and N (purple colour).

word vector of the amino acid. Based on the matrix, we investigated the similarity of amino acid residues around the Kcrot sites. The 20 amino acids were hierarchically clustered using Euclidean distance in average linkage (Figure 5C). The amino acids were distributed into four major clusters: (i) the alkaline amino acids K and R, (ii) the amino acids with negative charged side chains D and E, (iii) the hydrophobic amino acids F, L, M, I, V, W, Y and A, (iv) the mainly

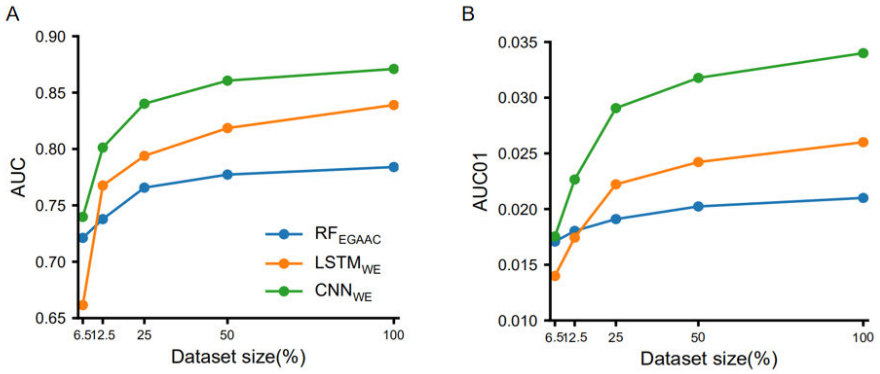

FIGURE 6. Impact of the training set data size on the prediction performance of independent test sets. The AUC (A) and AUC01 (B) curves were generated using five different data sizes: a sixteenth, an eighth, a quarter, a half, and the whole independent dataset from the human species. The whole dataset contained 6,687 positive peptides and 67,106 negative peptides.

polar uncharged residues T, Q, S, G, H and N. The special amino acid C and P were separated from these clusters. This clustering is similar to the classification of 20 amino acids according to their physicochemical properties, indicating that physicochemical properties are important as the features of classification and our model is capable of elucidating the significance of the correlation between amino acid properties.

B. ESTIMATION OF THE IMPACT OF DATA SIZE ON PREDICTION ACCURACY

The performance of an ML algorithm is generally sensitive to the size of the training data. We previously constructed the predictors for lysine malonylation and found that the DL algorithm has better performance than the traditional ML approach for the large-sized dataset but it might not be true for the small-sized dataset [21]. Here, we estimated whether the previous observation existed for lysine crotonylation. We selected the predictors and compared their performances constructed based on a sixteenth, eighth, a quarter, a half of, and the whole training dataset and evaluated them using the independent dataset (Figure 6A&6B). The overall performances of all the approaches increased with the size of the training dataset. Additionally, CNN_{WE} performed better than the traditional algorithms (RF_{EGAAC}) and $LSTM_{WE}$ in terms of AUC and AUC01 values in the range of the data size between 1/16 (including 418 positives) and the whole (including 6,687 positives) datasets. On the contrary, $LSTM_{WE}$ had an inferior performance compared to RF_{EGAAC} when the data size is the smallest whereas the former compared favourably to the latter when the data size increased. These observations indicate that the performances of the DL models are largely affected by the data size compared with the RF models and CNN_{WE} is a robust and reliable model with high performance.

C. EVALUATION OF SPECIES-SPECIFIC CNN_{WE} MODELS AND THEIR CROSS-SPECIES PERFORMANCES

The lysine crotonylation sites have been investigated from four different species, including human and three plant species (i.e. papaya, rice and tabacum). The number of identified Kcrot sites ranged from 1265 for the rice organism to over 10,000 for humans. We developed the classifier human-specific CNN_{WE} and found that CNN_{WE} outperformed other predictors for both large-sized and small-sized datasets (Figure 6). As the numbers of positives identified in the three plant species are within this range (Figure 1), we did not repeat our analyses performed above for the three plant species and constructed the CNN_{WE} predictors directly for these species, using the same data processing method as the human dataset (Figure 1). All the species-specific classifiers have the AUC/AUC01 values larger than 0.838/0.033 using species-specific independent test sets, respectively (Figure 7).

The crotonylation is catalyzed by crotonyltransferases. Some of them are evolutionarily conserved such as MOF that is found in both yeast and human [28] while others are not such as CBP and p300 that only exist in mammalian cells. Therefore, the enzymes from different species may have diverse characteristics and the produced Kcrot sites may have different features. To compare these modifications between different species, we interrogated the cross-species performance of CNN_{WE}. The test dataset was the independent test dataset from each species. Expectedly, the cross-species performances for each species-specific predictor are not as good as the self-species performance in terms of AUC and AUC01 values (Figure 7). For instance, the rice-specific model had the AUC value of 0.858 whereas other specific models only had the AUC value with the range of 0.76 to 0.80 (Figure 7). Additionally, we combined the training datasets from all four species and constructed a general CNN_{WE} classifier. The general predictor based on the large dataset had better performance than cross-species performance. Furthermore, the general predictor compared favourably to a few species-specific predictors (Figure 7). For example, the general classifier had the AUC/AUC01 value of 0.890/0.0377 for the papaya test set while the value reduced to 0.878/0.0355 for the papaya-specific classifier, respectively. However, the general classifier also showed inferior performance to the Tabacum-specific classifier in terms of the Tabacum independent test set. The former had the AUC/AUC01 value of 0.833/0.0319 whereas the latter had the values of 0.838/0.0337, respectively (Figure 7). These suggest that the large size of the Kcrot training dataset has comprehensive coverage of the Kcrot commonality across

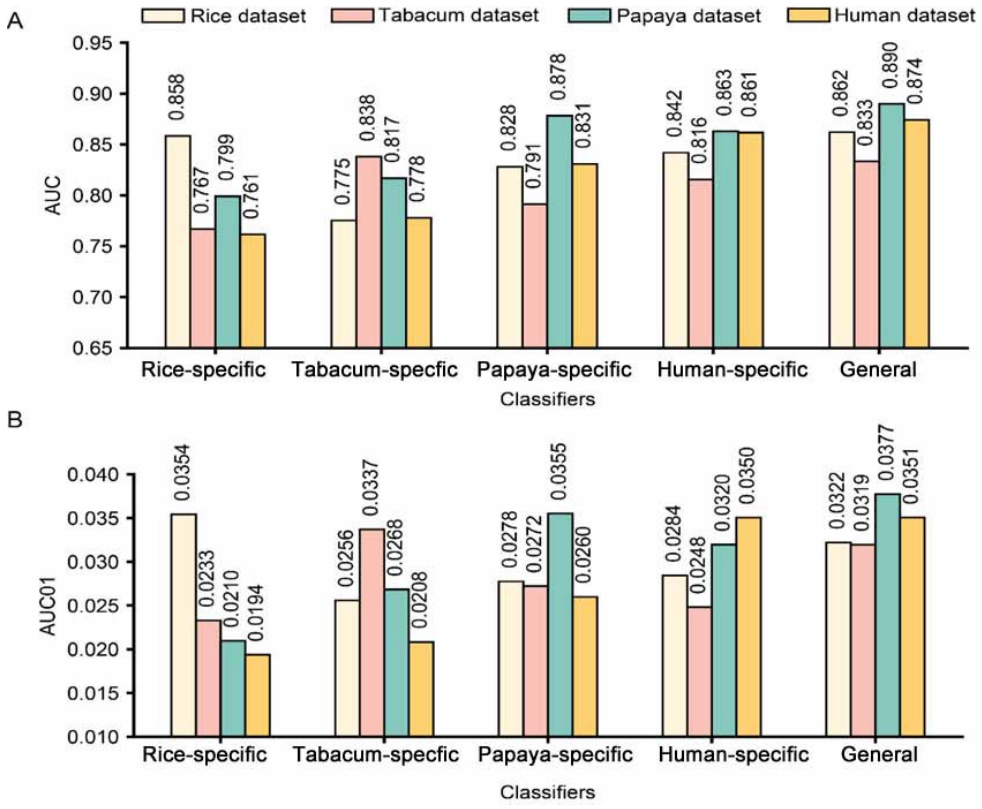

FIGURE 7. Comparison of prediction performances for species-specific CNN_{WE} and the general CNN_{WE} classifier. The AUC (A) and AUC01 (B) values were calculated for self-species and cross-species Kcrot prediction using the species-specific CNN_{WE} and the general CNN_{WE}. The former classifier was constructed using the species-specific training dataset while the latter was developed using the combination of species-specific training datasets. The test datasets were the independent test dataset from different species (Figure 1). The species were ordered according to the number of positives used for predictor construction.

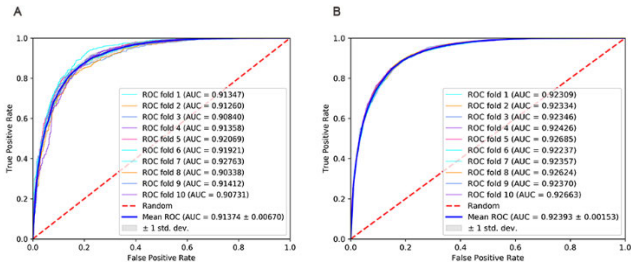

FIGURE 8. Prediction performances of the CNN_{WE} model were developed and evaluated using the dataset from Deep-Kcr. The performance of the CNN_{WE} model in terms of ten-fold cross-validation (A) and the independent test (B).

different species and thus increases cross-species prediction performance, although the exceptional species (i.e. tabacum) exist (Figure 7). These observations may be consistent with the fact that some crotonyltransferases are conserved and others are not.

D. COMPARISON OF CNN_{WE} WITH REPORTED KCROT PREDICTORS

The current classifier CNN_{WE} was constructed using Kcrot sites on non-histone proteins as the benchmark dataset. We estimated whether CNN_{WE} could efficiently predict Kcrot sites of histone proteins. The histone test set contained 169 positive sites and the rest 816 K-containing peptides as negative sites [15]. CNN_{WE} failed to distinguish the positives from the negatives (AUC = 0.623, AUC01 = 0.003). It indicates that the Kcrot sites from non-histone and histone proteins have distinct characteristics. We re-constructed CNN_{WE} by adding 4/5 (i.e. 134 positives) of known histone Kcrot peptides into the training set and considered the rest positives and all negatives (i.e. 35 positives and 816 negatives) as the test set. The new CNN_{WE} model showed improved accuracy for the histone Kcrot prediction (AUC = 0.966, AUC01 = 0.081). The larger AUC value for the histone Kcrot prediction than the AUC value for the non-histone Kcrot suggests that the histone Kcrot sites have common features whereas the features of non-histone Kcrot sites are relatively diverse. In summary, it is necessary to build CNN_{WE} by including Kcrot sites from histone proteins.

TABLE 3. Performances of papaya-specific CNN_{WE} with the models developed by Wang *et al.* [19].

Model	Sn	Sp	AUC
CNN_{WE}[1]	**0.894±0.007**	**0.73**	**0.879±0.004**
LGBM_RF[2]	0.85	0.72	0.84
MRMD_svm[2]	0.75	0.73	0.84

We compared our CNN_{WE} architecture with the Deep-Kcr model. As the developers of the Deep-Kcr model shared the training data set and the independent test set through https://github.com/linDing-group/Deep-Kcr, we developed the CNN_{WE} model using the same dataset and evaluated

it through ten-fold cross-validation and independent test. The average AUC values were 0.914 ± 0.007 and 0.924 ± 0.002 in terms of ten-fold cross-validation (Figure 7A) and independent test (Figure 7B), respectively. Both AUC values are significantly larger than the counterparts of Deep-Kcr (0.885 and 0.859). Therefore, the CNN_{WE} architecture compared favourably to Deep-Kcr.

We further compared our papaya-specific CNN_{WE} model with the models developed by Wang *et al.*. [19]. All the models were based on the same papaya dataset (3453 positive and 37,134 negative sequences). We separated the data into the cross-validation dataset (2742 positives and 29,676 negatives) and independent test dataset (711 positives and 7458 negatives), whereas Wang *et al.* generated the training dataset (2548 positives and 2548 negatives) and the testing dataset (669 positives and 6720 negatives) [19]. Please note that the sum of positives in Wang's datasets is 3217, which is smaller than 3453. It may be due to the filtering of the sequences of length less than 31 amino acid compositions or those containing uncertain composition, described by Wang *et al.* As our independent dataset was larger than Wang's testing dataset, we randomly selected from our independent test dataset 669 positives and 6720 negatives for ten times as the test dataset for the evaluation of our CNN_{WE} model. Wang *et al.* developed several models with different features and found that the models with the incorporated and selected features had the best performances. According to Table 4 [19], the RF model with the LGBM selection method (LGBM_RF) and the SVM model with the MRMD selection method (MRMD_svm) had the best performances. Therefore, we selected these two models for comparison. Table 3 showed that DeepKcrot had the largest AUC value and had the largest sensitivity when specificity was fixed at 0.72. In summary, the Papaya-specific CNN_{WE} model compared favourably to the models developed by Wang *et al.*.

E. CONSTRUCTION OF THE ONLINE KCROT PREDICTOR

We developed an easy-to-use online tool for the prediction of the Kcrot sites, dubbed DeepKcrot. DeepKcrot contained four species-specific CNN_{WE} predictors and a general CNN_{WE} classifier. The users could select the general model or species-specific model at the input interface and input the query protein sequences directly or upload the sequence file. The prediction results are output in tabular form with five columns: sequence header, position, sequence, prediction score, and prediction result that was colour-coded with at the specificity levels of 80, 90, and 95%, respectively.

IV. CONCLUSION

The common PTM prediction approaches are based on ML that requires experts to pre-define informative features. They have been widely applied to the prediction of lysine crotonylation based on the Kcrot sites on histone proteins. Recently, thousands of Kcrot sites have been identified on non-histone proteins from different species but it is unclear whether lysine crotonylation on these proteins could be effectively predicted. In this study, we compiled a benchmark data set of known Kcrot sites and evaluated the performance of different machine-learning approaches, including deep-learning algorithms. We found that the DL-based classifier CNN_{WE} had the best performance compared with the traditional ML model and the $LSTM_{WE}$ model that showed superior to CNN_{WE} for the prediction of cysteine sulphenylation sites, even for the limited training dataset [24]. It suggests that CNN and LSTM may have distinct characteristics that are feasible to extract different PTM features. Furthermore, these models were compared using different sizes of the training data set and CNN_{WE} again shows the best performance, suggesting its superior performance and robustness. We also compared CNN_{WE} with the recently reported Deep-Kcr model and CNN_{WE} showed better performance. Additionally, the CNN_{WE} model constructed based on non-histone Kcrot sites failed to predict the histone Kcrot sites, suggesting the histone and non-histone Kcrot sites may have different features. Accordingly, we reconstructed the CNN models by including the histone Kcrot sites. Moreover, we constructed four organism-specific CNN_{WE} models and found that cross-species performances for species-specific CNN_{WE} predictors were not as good as the self-species performances. It suggests that the crotonylome for each organism has its specific features. The general CNN_{WE} predictor based on the integration of the training data from different species showed superior performance to the species-specific predictors except for one organism. Taken together, we developed the first DL architecture DeepKcrot for predicting Kcrot sites on proteomes across various species. The outstanding performance of the DL algorithms in the prediction of Kcrot sites suggests that DL may be applied broadly to predicting other types of PTM sites.

AUTHORS' CONTRIBUTIONS

Lei Li conceived and designed the project. Xilin Wei, Yutong Sha, and Yiming Zhao constructed the algorithms under the supervision of Lei Li. Xilin Wei, Yutong Sha, Yiming Zhao, and Ningning He analyzed the data. Xilin Wei, Yutong Sha, Yiming Zhao, and Lei Li wrote the manuscript. All authors reviewed the manuscript.

ACKNOWLEDGMENT

(Xilin Wei, Yutong Sha, and Yiming Zhao contributed equally to this work.)

REFERENCES

[1] M. Tan, H. Luo, S. Lee, F. Jin, J. S. Yang, E. Montellier, T. Buchou, Z. Cheng, S. Rousseaux, N. Rajagopal, Z. Lu, Z. Ye, Q. Zhu, J. Wysocka, Y. Ye, S. Khochbin, B. Ren, and Y. Zhao, "Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification," *Cell*, vol. 146, no. 6, pp. 1016–1028, Sep. 2011.

[2] E. Montellier, S. Rousseaux, Y. Zhao, and S. Khochbin, "Histone crotonylation specifically marks the haploid male germ cell gene expression program: Post-meiotic male-specific gene expression," *Bioessays*, vol. 34, no. 3, pp. 93–187, Mar. 2012.

[3] S. Liu *et al.*, "Chromodomain protein CDYL acts as a crotonyl-CoA hydratase to regulate histone crotonylation and spermatogenesis," *Mol. Cell*, vol. 67, no. 5, pp. 853–866e5, Sep. 2017.

[4] B. R. Sabari, Z. Tang, H. Huang, V. Yong-Gonzalez, H. Molina, H. E. Kong, L. Dai, M. Shimada, J. R. Cross, Y. Zhao, R. G. Roeder, and C. D. Allis, "Intracellular crotonyl-CoA stimulates transcription through p300-catalyzed histone crotonylation," *Mol. Cell*, vol. 58, no. 2, pp. 203–215, Apr. 2015.
[5] W. Wei, A. Mao, B. Tang, Q. Zeng, S. Gao, X. Liu, L. Lu, W. Li, J. X. Du, J. Li, J. Wong, and L. Liao, "Large-scale identification of protein crotonylation reveals its role in multiple cellular functions," *J. Proteome Res.*, vol. 16, no. 4, pp. 1743–1752, Apr. 2017.
[6] H. Huang, D.-L. Wang, and Y. Zhao, "Quantitative crotonylome analysis expands the roles of p300 in the regulation of lysine crotonylation pathway," *Proteomics*, vol. 18, no. 15, Aug. 2018, Art. no. 1700230.
[7] Q. Wu, W. Li, C. Wang, P. Fan, L. Cao, Z. Wu, and F. Wang, "Ultradeep lysine crotonylome reveals the crotonylation enhancement on both histones and nonhistone proteins by SAHA treatment," *J. Proteome Res.*, vol. 16, no. 10, pp. 3664–3671, Oct. 2017.
[8] W. Xu, J. Wan, J. Zhan, X. Li, H. He, Z. Shi, and H. Zhang, "Global profiling of crotonylation on non-histone proteins," *Cell Res.*, vol. 27, no. 7, pp. 946–949, Jul. 2017.
[9] K. Liu, C. Yuan, H. Li, K. Chen, L. Lu, C. Shen, and X. Zheng, "A qualitative proteome-wide lysine crotonylation profiling of papaya (Carica papaya L.)," *Sci. Rep.*, vol. 8, no. 1, p. 8230, May 2018.
[10] H. Sun, X. Liu, F. Li, W. Li, J. Zhang, Z. Xiao, L. Shen, Y. Li, F. Wang, and J. Yang, "First comprehensive proteome analysis of lysine crotonylation in seedling leaves of Nicotiana tabacum," *Sci. Rep.*, vol. 7, no. 1, p. 3013, Jun. 2017.
[11] S. Liu, C. Xue, Y. Fang, G. Chen, X. Peng, Y. Zhou, C. Chen, G. Liu, M. Gu, K. Wang, W. Zhang, Y. Wu, and Z. Gong, "Global involvement of lysine crotonylation in protein modification and transcription regulation in rice," *Mol. Cellular Proteomics*, vol. 17, no. 10, pp. 1922–1936, Oct. 2018.
[12] H. Yu, C. Bu, Y. Liu, T. Gong, X. Liu, S. Liu, X. Peng, W. Zhang, Y. Peng, J. Yang, L. He, Y. Zhang, X. Yi, X. Yang, L. Sun, Y. Shang, Z. Cheng, and J. Liang, "Global crotonylome reveals CDYL-regulated RPA1 crotonylation in homologous recombination–mediated DNA repair," *Sci. Adv.*, vol. 6, no. 11, Mar. 2020, Art. no. eaay4697.
[13] G. H. Huang and W. F. Zeng, "A discrete hidden Markov model for detecting histone crotonyllysine sites," *Match-Commun. Math. Comput. Chem.*, vol. 75, no. 3, pp. 717–730, 2016.
[14] W.-R. Qiu, B.-Q. Sun, X. Xiao, Z.-C. Xu, and K.-C. Chou, "IPTM-mLys: Identifying multiple lysine PTM sites and their different types," *Bioinformatics*, vol. 32, no. 20, pp. 3116–3123, Oct. 2016.
[15] W.-R. Qiu, B.-Q. Sun, X. Xiao, Z.-C. Xu, J.-H. Jia, and K.-C. Chou, "IKcr-PseEns: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier," *Genomics*, vol. 110, no. 5, pp. 239–246, Sep. 2018.
[16] W.-R. Qiu, B.-Q. Sun, H. Tang, J. Huang, and H. Lin, "Identify and analysis crotonylation sites in histone by using support vector machines," *Artif. Intell. Med.*, vol. 83, pp. 75–81, Nov. 2017.
[17] Z. Ju and J.-J. He, "Prediction of lysine crotonylation sites by incorporating the composition of k -spaced amino acid pairs into Chou's general PseAAC," *J. Mol. Graph. Model.*, vol. 77, pp. 200–204, Oct. 2017.
[18] Y. Zhao, N. He, Z. Chen, and L. Li, "Identification of protein lysine crotonylation sites by a deep learning framework with convolutional neural networks," *IEEE Access*, vol. 8, pp. 14244–14252, 2020.
[19] R. Wang, Z. Wang, H. Wang, Y. Pang, and T.-Y. Lee, "Characterization and identification of lysine crotonylation sites based on machine learning method on both plant and mammalian," *Sci. Rep.*, vol. 10, no. 1, p. 20447, Nov. 2020.
[20] H. Lv, F.-Y. Dao, Z.-X. Guan, H. Yang, Y.-W. Li, and H. Lin, "Deep-Kcr: Accurate detection of lysine crotonylation sites using deep learning method," *Briefings Bioinf.*, pp. 1–2, Oct. 2020.
[21] Z. Chen, N. He, Y. Huang, W. T. Qin, X. Liu, and L. Li, "Integration of a deep learning classifier with a random forest approach for predicting malonylation sites," *Genomics, Proteomics Bioinf.*, vol. 16, no. 6, pp. 451–459, Dec. 2018.
[22] Z. Chen, Y. Zhou, Z. Zhang, and J. Song, "Towards more accurate prediction of ubiquitination sites: A comprehensive review of current methods, tools and features," *Briefings Bioinf.*, vol. 16, no. 4, pp. 640–657, Jul. 2015.
[23] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: Accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, Dec. 2012.
[24] X. Lyu, S. Li, C. Jiang, N. He, Z. Chen, Y. Zou, and L. Li, "DeepCSO: A deep-learning network approach to predicting cysteine S-sulphenylation sites," *Frontiers Cell Develop. Biol.*, vol. 8, Dec. 2020, Art. no. 594587.
[25] W. He, L. Wei, and Q. Zou, "Research progress in protein posttranslational modification site prediction," *Briefings Funct. Genomics*, vol. 18, no. 4, pp. 220–229, Jul. 2019.
[26] Y. Huang, N. He, Y. Chen, Z. Chen, and L. Li, "BERMP: A cross-species classifier for predicting m6A sites by integrating a deep learning algorithm and a random forest approach," *Int. J. Biol. Sci.*, vol. 14, no. 12, pp. 1669–1677, 2018.
[27] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[28] X. Liu, W. Wei, Y. Liu, X. Yang, J. Wu, Y. Zhang, Q. Zhang, T. Shi, J. X. Du, Y. Zhao, M. Lei, J.-Q. Zhou, J. Li, and J. Wong, "MOF as an evolutionarily conserved histone crotonyltransferase and transcriptional activation by histone acetyltransferase-deficient and crotonyltransferase-competent CBP/p300," *Cell Discovery*, vol. 3, no. 1, p. 17016, Dec. 2017.

XILIN WEI received the B.Eng. degree from the School of Medical Information Engineering, Jining Medical University, China, in 2018. He is currently pursuing the master's degree with the School of Data Science and Software Engineering, Qingdao University, China. His research interests include bioinformatics and deep learning.

YUTONG SHA received the B.Eng. degree from the School of Computer and Software, Weifang University of Science and Technology, China, in 2018. She is currently pursuing the master's degree with the School of Data Science and Software Engineering, Qingdao University, China. Her research interests include bioinformatics and deep learning.

YIMING ZHAO received the master's degree from the School of Data Science and Software Engineering, Qingdao University, China, in 2020. His research interests include bioinformatics and deep learning.

NINGNING HE received the Ph.D. degree from Sookmyung Women's University, South Korea, in 2015. She is currently an Associate Professor with the School of Basic Medicine, Qingdao University, China. Her research interests include bioinformatics, genomics, and proteomics.

LEI LI received the Ph.D. degree from Nanyang Technological University, Singapore, in 2006. He is currently a Professor with the School of Data Science and Software Engineering and the School of Basic Medicine, Qingdao University, China. His research interests include deep-learning, bioinformatics, systems biology, and proteomics.