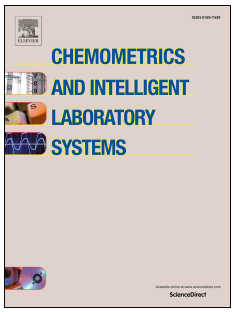
Journal Pre-proof



DeepMal: Accurate prediction of protein malonylation sites by deep neural networks

Minghui Wang, Xiaowen Cui, Shan Li, Xinhua Yang, Anjun Ma, Yusen Zhang, Bin Yu

PII: S0169-7439(20)30436-6

DOI: https://doi.org/10.1016/j.chemolab.2020.104175

Reference: CHEMOM 104175

To appear in: *Chemometrics and Intelligent Laboratory Systems*

Received Date: 16 July 2020

Revised Date: 8 October 2020

Accepted Date: 8 October 2020

Please cite this article as: M. Wang, X. Cui, S. Li, X. Yang, A. Ma, Y. Zhang, B. Yu, DeepMal: Accurate prediction of protein malonylation sites by deep neural networks, *Chemometrics and Intelligent Laboratory Systems* (2020), doi: https://doi.org/10.1016/j.chemolab.2020.104175.

Author Contributions - CRediT

Minghui Wang: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft, Validation. **Xiaowen Cui:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft, Validation, Visualization. **Shan Li:** Formal analysis, Investigation, Methodology, Validation, Visualization. **Xinhua Yang:** Methodology, Validation, Writing - review & editing. **Anjun Ma:** Formal analysis, Investigation, Methodology, Validation. **Yusen Zhang:** Formal analysis, Investigation, Methodology, Writing - original draft. **Bin Yu:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft, Writing- Reviewing and Editing.

DeepMal: accurate prediction of protein malonylation sites by deep neural networks

Minghui Wang [a,b,1], Xiaowen Cui [a,b,1], Shan Li [c], Xinhua Yang [a,b,1], Anjun Ma [d], Yusen Zhang [e], Bin Yu [a,b,f,*]

[a] *College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China*

[b] *Artificial Intelligence and Biomedical Big Data Research Center, Qingdao University of Science and Technology, Qingdao 266061, China*

[c] *School of Mathematics and Statistics, Central South University, Changsha 410083, China*

[d] *Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, Ohio 43210, USA*

[e] *School of Mathematics and Statistics, Shandong University, Weihai 264209, China*

[f] *School of Life Sciences, University of Science and Technology of China, Hefei 230027, China*

ABSTRACT

Lysine malonylation is one of the important post-translational modification (PTM) of proteins. Malonylated proteins can affect various cell functions of eukaryotes and prokaryotes, which play an important role in metabolic pathways, mitochondrial functions, fatty acid oxidation and other life activities. However, accurate identification of the malonylation sites is the key to further understand the molecular mechanism of malonylation. Currently, experimental identification is still a challenging task, which usually requires a large amount of laboratory work and considerable cost. Regarding this situation, there is an urgent need to establish useful calculation methods and develop effective predictors. We propose a new deep learning network model called DeepMal. Firstly, features are extracted by enhanced amino acid composition (EAAC), enhanced grouped amino acid composition (EGAAC), dipeptide deviation from the expected mean (DDE), k nearest neighbors (KNN) and BLOSUM62 matrix. Secondly, the linear convolutional neural network is used to extract the specific features of malonylation sites, select the relevant features and reduce the feature dimension through maximum pooling. Finally, malonylation sites and non-malonylation sites are classified by a multilayer neural network, an independent dataset is used to assess the predictive ability of the model DeepMal. On the independent datasets

* Corresponding author.

E-mail address: yubin@qust.edu.cn (B. Yu).

[1] These authors contributed equally to this work.

Escherichia coli (*E. coli*), *Homo sapiens* (*H. sapiens*), *Mus musculus* (*M. musculus*), the AUC values are 0.974, 0.956 and 0.944, and the accuracy are 96.5%, 95.5% and 94.5%, respectively. Compared with other prediction models, the prediction accuracy is increased by 9.5%-18.5%, further indicating the effectiveness of the prediction model DeepMal. Using deep learning network can enhance the robustness of DeepMal model for predicting malonylation sites. The source codes and all datasets are publicly available at https://github.com/QUST-AIBBDRC/DeepMal/.

Keywords: Lysine malonylation sites; Multi-information fusion; Convolutional neural network; Deep neural network.

1. Introduction

Post-translational modification (PTM) can change protein interactions, stability, activity, and affect protein various functions by adding modifying groups with one or more amino acid residues. Currently, more than 400 types of modification have been discriminated [1], such as phosphorylation, glycosylation, methylation, ubiquitination, malonylation, acetylation, S-sulfonation, formylation and S-nitrosylation. PTM of proteins, as an important mechanism for regulating protein functions, is involved in all kinds of cell life activities [2]. Many proteins interact and coordinate with each other to maintain the normal activities of cell life. However, serious diseases may be caused due to abnormal post-translational modification [3], such as cancer [4], diabetes [5], and autoimmune disease [6]. Therefore, making a profound study of PTM of proteins is important in revealing the mechanism of life activities, screening clinical drugs of diseases and identifying drug targets.

As an important protein post-translational modification sites, malonylation [7] was firstly discovered in 2011 as an evolutionarily conserved type of post-translational modification of lysine. Studies proved that malonylated proteins present in multiple metabolic pathways, for instance glucose and fatty acid metabolism, fatty acid synthesis and oxidation [8], impaired mitochondrial function [9]. Meanwhile, malonylation sites are related to muscle contraction, myocardial ischemia and hypothalamic regulation of appetite, diabetes, and cancer [10]. Given the importance of malonylation, it is very important to accurately identify the malonylation sites, which can supply useful news for biomedical research to better understand molecular functions. At present, the experimental methods have time and cost constraints, and the experiments are difficult. Hence, it is necessary to exploit computing method that can accurately identify malonylation sites.

Recently, due to their important role in life activities, there have been some published works that used machine learning and deep learning methods to predict malonization sites of proteins. These methods are different in various aspects, including datasets, feature extraction algorithm, feature selection algorithm, classifier, evaluation verification method and evaluation index. Xu et al. [11] put forward malonylation sites predictor Mal-Lys, based on sequence feature k-grams, position-specific amino acid propensity and physical and chemical information AAindex feature extraction method to convert protein character information into numerical vectors. The mRMR is applied to choose the optimal feature subset. Wang et al.

[12] developed a species-specific malonylation sites classifier, MaloPred, which used amino acid compositions, binary encoding, encoding based on grouped weight, KNN and position specific scoring matrix to convert character information into numerical vectors. Reduce the dimension of feature vector through information gain, and then input it to the support vector machine for classification. Zhang et al. [13] constructed a malonylation sites predictor kmal-sp, using 11 feature extraction methods to extract protein feature information, using GainRatio for feature analysis and selection, and integrating RF, SVM, KNN, LR and Light Gradient Boosting Machine learning methods for classification. Xiang et al. [14] applied pseudo amino acid composition to extract features, and selected nu-SVM as a classifier to construct a prediction model. Taherzadeh et al. [15] constructed a protein malonylation predictor SPRINT-Ma using binary coding, position-specific scoring matrix, AAindex, accessible surface area, secondary structure, half-sphere exposure, and intrinsically disordered region to extract protein features. The SVM of the radial basis kernel function was classified. Bao et al. [16] constructed a malonylation sites prediction model IMKPse, which used the general PseAAC to extract classification features and used a exible neural tree as a classifier. Chen et al. [17] used EAAC, AAindex, and one-hot coding methods to integrate long short-term memory with word embedding and RF to construct malonylation sites classifier LEMP. With the development of computer technology and big data in recent years, deep learning has become one of the popular machine learning algorithms [18]. Deep learning is a means in view of data representational learning. In essence, it is a neural network used to build and simulate human brain for analysis and learning, and learn high-dimensional data by constructing complex networks[19]. In addition, with the application of deep learning in various domain, researchers have begun to apply in bioinformatics research, including protein folding recognition [20], prediction of hydroxylation sites [21], prediction of protein submitochondrial localization [22], prediction of protein nitration and nitrosation sites [23], protein-protein interaction sites [24], identification of clathrin proteins [25], deep learning has been used to build different prediction models, and better prediction results have been achieved. In addition, with the application of deep learning in various fields, researchers have begun to use it for bioinformatics research. Specific applications can be seen in the Supplementary Si1.

Although the above methods have obtained exciting results and promoted predictive studies of malonylation sites, these methods have some shortcomings. i) Existing methods use only limited features, and other potential features have yet been determined; ii) Redundant features in model construction will reduce prediction performance.

We put forward a new deep learning framework, DeepMal, to accurately predict malonylation sites. Firstly, an enhanced amino acid composition, enhanced grouped amino acid composition, dipeptide deviation from the expected mean, K nearest neighbors and BLOSUM62 matrix are used for feature extraction. Secondly, the feature vectors are inputted into the convolutional neural network framework to extract significant features related to malonylation sites. Thirdly, the most relevant features are selected through maximum pooling and the feature dimension is reduced. Finally, deep neural network and softmax classification function are used to classify malonylation and non-malonylation sites. On the independent

datasets *E. coli*, *H. sapiens* and *M. musculus*, the ACC of DeepMal were 96.5%, 95.5% and 94.5%, respectively. DeepMal significantly improved the prediction of lysine malonylation sites compared with other existing tools. The accurate prediction of lysine malonylation sites will help people to understand the pathogenesis and treatment of related diseases.

2. Materials and methods

2.1. Datasets

In this study, an experimentally validated lysine malonylation dataset was taken from the study [13]. The dataset comes from UniProt database [26], and CD-HIT [27] is used to reduce the similarity and sequences homology, and retain protein sequences with sequence identities <30%, so as to avoid latent errors in model training. The dataset includes 1,746 malonylation sites from 595 *E. coli* proteins, 3,435 malonylation sites from 1,174 proteins in *M. musculus*, and 4,579 malonylation sites from 1,660 proteins in *H. sapiens* [13]. Each protein contains 25 residues, and lysine (K) is in the center. The nonredundant datasets were randomly separated into the training datasets and the independent datasets. Finally, we gained experimental datasets as shown in Table 1.

Table 1

The number of malonylation and non-malonylation samples in the training sets and independent test datasets.

Dataset	Species	Number of malonylation samples	Number of non-malonylation samples
Training set	*E. coli*	1,453	1,453
	H. sapiens	3,585	3,585
	M. musculus	2,606	2,606
Independent test	*E. coli*	100	100
	H. sapiens	300	300
	M. musculus	600	600

2.2. Feature extraction

It is an important step in the classification task converting character information of a sequence into a numeric vector through feature extraction algorithm, which directly affects the prediction results of the model. In this study, we use EAAC [28], EGAAC [28], DDE [29], KNN [12] and BLOUSM62 [30] algorithms to extract protein sequence features. Each sample is converted into 1431-D numerical vector.

2.2.1. Enhanced amino acid composition

The enhanced amino acid composition algorithm [28] extracts protein sequence information and calculates amino acid frequency information. The EAAC can be calculated as:

$$g(m,n)=\frac{H(m,n)}{H(n)},\ m\in\{A,C,D,\cdots,Y\},\ n\in\{\text{ wl, w2},\cdots,\text{ wL}\} \tag{1}$$

where $H(m,n)$ is the number of amino acid type m, $H(n)$ is the length of the window n [28].

2.2.2. Enhanced grouped amino acid composition

The Enhanced grouped amino acid composition algorithm converts character information of protein sequences into numerical vectors. It is an effective feature extraction algorithm that has been used in the field of bioinformatics research, such as virus post-translational modification sites prediction [31], lysine succinylation sites prediction [32]. Lee et al. [33] divided 20 amino acid types into five categories based on the five physicochemical properties of amino acids:

aliphatic group ($g1:GAVLMI$),

aromatic group ($g2:FYW$),

positive charge group ($g3:KRH$),

negative charged group ($g4:DE$),

uncharged group ($g5:STCPNQ$)

The calculation formula is as follows:

$$G(g,n)=\frac{H(g,n)}{H(n)},\ \ g\in\{g1,g2,g3,g4,g5\},\ \ n\in\{w1,w2,\cdots,wL\} \tag{2}$$

where $H(g,n)$ is the number of amino acids in group g within the window n, $H(n)$ is the length of the window n [28]. The fixed length sequence window size defaults to 5.

2.2.3. Dipeptide deviation from expected mean

Saravanan et al. [29] proposed a feature extraction algorithm, dipeptide deviation from the expected mean when studying B cell epitope prediction. First calculate the protein sequence dipeptide composition (DC), the calculation formula is as follows:

$$DC(m,n)=\frac{H_{mn}}{H-1},\qquad m,n\in\{A,C,D,\cdots,Y\} \tag{3}$$

where H_{mn} is the number of amino acid pairs mn and H is the size of the protein sequence.

Secondly, to calculate the theoretical mean (TM) and theoretical variance (TV) of the protein sequence, the calculation formula is as follows:

$$TM(m,n) = \frac{C_m}{C_H} * \frac{C_n}{C_H} \tag{4}$$

where C_m is the number of codons encoding the first amino acid, C_n is the number of codons encoding the second amino acid, and C_H is the total number of possible codons.

$$TV(m,n) = \frac{TM(m,n)(1-TM(m,n))}{H-1} \tag{5}$$

Finally, calculate DDE based on DC, TM and TV, DDE feature vector calculation is as follows:

$$DDE(m,n) = \frac{DC(m,n)(1-TM(m,n))}{\sqrt{TV(m,n)}} \tag{6}$$

2.2.4. K nearest neighbors

Using local sequence clustering information to predict malonylation sites, clustering information can be obtained by merging protein sequence fragments and comparing the positive and negative data with the sequences of the alignment group. The KNN algorithm [12] has been widely used in bioinformatics, such as prediction of lysine formylation sites [34], acetylation sites prediction [35], and prediction of phosphorylation sites [36]. The detailed description is as follows:

(1) For two local sequences s_1 and s_2, the distance $Dist(s_1, s_2)$ is defined as:

$$Dist(s_1, s_2) = 1 - \frac{\sum_{i=-p}^{p} Sim(s_1(i), s_2(i))}{2p+1} \tag{7}$$

$$sim(a,b) = \frac{M(a,b) - \min(a,b)}{\max(M) - \min(M)} \tag{8}$$

where a and b are two amino acids, M is the BLOSUM62 substitution matrix, and $\max(M)/\min(M)$ is the maximum / minimum number in the BLOSUM62 substitution matrix.

(2) Then extract the corresponding KNN features: *i)* according to formulas 7) and 8), calculate the sequence local similarity, *ii)* we sort the distances in ascending order, and select K number of nearest neighbors by the given K value, *iii)* the KNN score is calculated as the percentage of positive samples (ie malonylation sites) in K samples. Different K values are used to obtain multiple features. In this paper, K is set to 2, 4, 8, 16, 32, 64.

2.2.5. BLOSUM62

The amino acid substitution matrix is based on the alignment of amino acid sequences. The identity between the two peptide sequences does not exceed 62%. The character information of the protein sequence is converted into a numerical vector. It has been widely used in bioinformatics, such as prokaryote lysine acetylation sites prediction [37], and S-sulphenylation sites prediction [38]. Using the BLOSUM62 [28] feature extraction algorithm, the fragment sequence with a window length of 25 amino acids can be encoded as a 500-dimensional feature vector.

2.3. Deep learning

In order to accurately predict the malonylation sites of proteins, we construct a deep neural network framework called DL. DL has a hybrid architecture, including convolutional neural network (CNN) [39] and dense fully-connected layer [40]. This framework includes input layer, convolution layer, pooling layer , dropout layer, dense fully-connected layer and output layer [41]. The numerical vector transformed by EAAC, EGAAC, DDE, KNN and BLUSM62 is taken as the input features, and its size is $N*1431$, N is the number of protein sequences. In this paper, we use a 3×3 matrix with sliding windows for convolution operations. A rectified linear unit (ReLU) [20] is used to the normalized results as activation function, which is principal activation function in all deep neural networks, setting the negative value to zero. Given the input sample X, the convolution computation [39] in CNN layer is expressed as follows:

$$Convolution(X)_{ik} = \mathrm{Re}lu(\sum_{p=0}^{P-1}\sum_{q=0}^{Q-1} w_{pq}^{k} x_{i+p,q}) \tag{9}$$

$$\mathrm{Re}lu(x) = \max(0, x) \tag{10}$$

where w is the size of the convolution kernel, $W^k = (w_{pq}^k)_{P*Q}$ is the weight matrix of the k-th convolution kernel, and the size is $P*Q$. In order to reduce the feature vectors output by the convolutional layer and reduce redundant features of the operation in the network, the maximum pooling layer is embedded in the convolutional layer. Maxpooling only selects the maximum value in a filter range, the filter of size is 2×2. For a given feature vector of malonylation sites, it will output a smaller representation of convolution operation results. The pooling operation [39] is expressed as:

$$pooling(X)_{ik} = \max(x_{iM,\mathrm{k}}, x_{iM+1,\mathrm{k}}, x_{iM+2,\mathrm{k}}, \cdots, x_{iM+M-1,\mathrm{k}}) \tag{11}$$

Adding a dropout [42] layer which prevents overfitting, which improves predictive performance of the model. From the convolutional layer to the fully connected layer [22], the Flattened layer is used for "flattened" input. Since the output layer requires output of probability distribution of all classes, the flattening layer converts the input matrix into a vector.

We then use dense fully connected layers to learn non-linear local features, where the inputs are features learned from convolutional layers and pooling layers. Here, three fully

connected layers in our DL, denoted as fc3, fc4 and fc5, have 128, 64 and 32 neurons, respectively. As the activation function, ReLU, which is used for classification in the model construction process, once again play an important role. Softmax [43] activation function is used in output layer to calculate final output $o_k \in [0,1]$ of the target sites. In probability theory, softmax functions show the classification results in the form of probability.

In oredr to classify malonylation sites and non malonylation sites, we imported a softmax classifier in the final output layer of the model [23].

$$o_k = \frac{e^{y_k}}{\sum_{i=1}^{H} e^{y_k}} \tag{12}$$

where o_k is the output of the k-th neuron, which indicates that the k class probability is observed, y_k is the associated linear output of the previous hidden layer. The specific framework of deep learning can be found in Table 2. Our model is implemented by Keras and TensorFlow, and the parameter tuning of DeepMal is shown in Table 3.

Table 2

Deep learning architecture and hyperparameter settings. The size column describes the kernel size of the convolutional layer, the size of the largest pooling layer and the fully connected layer.

Layer no.	Layer type	Size	Activation
0	INPUT	-	-
1	CONV	64*3*1	Relu
2	MaxPooling	4*1	-
3	CONV	128*3*1	Relu
4	MaxPooling	4*1	-
5	Dropout	0.2	-
6	Flatten	-	-
7	Dense	128	Relu
8	Dense	64	Relu

9	Dense	32	Relu
10	Dense	2	Softmax
11	Optimizer	-	Adam

Table 3

Parameters range and settings of the DeepMal.

Name	Range	Setting
Activation function	relu, tanh, sigmoid, softmax, softplus, softsign, hard_sigmoid	relu
Dropout rate	0.1, 0.2, 0.5	0.2
Per-parameter adaptive learning rate methods	SGD, RMSprop, Adagrad, Adadelta, Adam, Adamax, Nadam	Adam
Learning rate	0.1, 0.01, 0.001, 0.0001	0.01

2.4. Model evaluation

The jackknife validation test, independent dataset test, and K-fold cross-validation are generally applied to evaluate the effectiveness of the prediction model [44-49]. We use 10-fold cross-validation method on the training dataset to train the model and use independent dataset to evaluate the predictive performance of the model. We used sensitivity (*Sn*), specificity (*Sp*), precision (*Pre*), accuracy (*ACC*) and Matthew's correlation coefficient (*MCC*) as the evaluation indicators [45]. The five evaluation indicators are defined as follows:

$$Sn = \frac{TP}{TP + FN} \tag{13}$$

$$Sp = \frac{TN}{TN + FP} \tag{14}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

$$Pre = \frac{TP}{TP + FP} \tag{16}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{17}$$

where TP represents the number of true positives, FP represents the number of false positives, TN represents the number of true negatives, and FN represents the number of false negatives. In addition, ROC [50-52] is a curve based on sensitivity and specificity.

2.5. Description of the DeepMal process

We propose a model for predicting malonation sites, DeepMal. The process of building the predictive model DeepMal is shown in Fig. 1. The specific procedures are as follows:

Step 1: *Get the datasets.* Obtain the training datasets *E. coli, M. musculus* and *H. sapiens* protein sequences and corresponding class labels, and the independent test set *E. coli, M. musculus* and *H. sapiens.*

Step 2: *Feature extraction.* Character information is converted into a feature vector through feature coding. (a) Extraction of protein sequence features using EAAC and EGAAC coding. (b) Physicochemical properties of protein sequences extracted from the BLOSUM62 matrix. (c) Use KNN coding to obtain the evolution information of the sequence. (d) The 400-D feature vector is obtained by DDE feature coding. Through fusion feature extraction, each protein sequence is transformed into 1431 dimensional feature vector.

Step 3: *Constructing a deep learning model.* The feature vectors and class labels are inputted into deep learning framework to predict the PTM of malonylation sites.

Step 4: *Model performance evaluation.* The 10-fold cross-validation is utilized to evaluate the prediction performance of the classification model, compute AUC, ACC, Pre, Sn, Sp and MCC values, draw the ROC curve, and evaluate the robustness of the model [53]. Use the model constructed in 1) -3) to test the model using an independent test sets and compare it with other prediction methods.

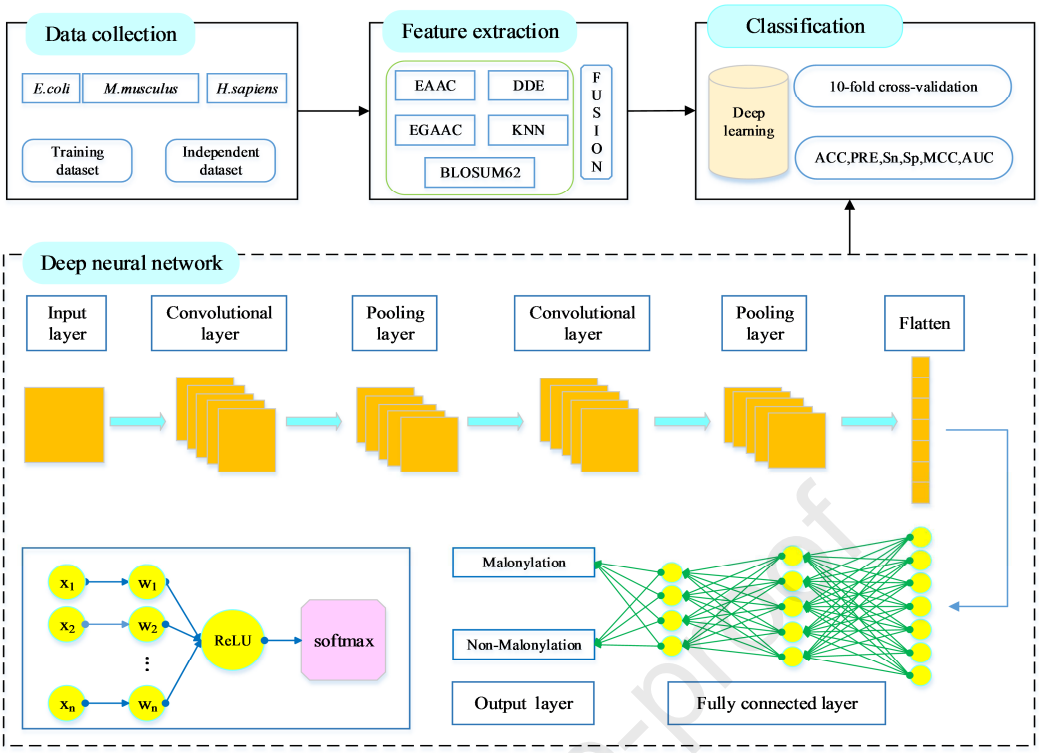


Fig. 1. The workflow of DeepMal for classifying malonylation and non-malonylation sites using convolutional neural networks and dense neural networks. It contains four steps: data collection, feature extraction, neural networks framework and model evaluation.

3. Results

3.1. Sequence analysis

For the sake of studying the distribution of amino acids around malonylation sites and non-malonylation sites, we submitted datasets *E. coli*, *H. sapiens* and *M. musculus* positive and negative training samples to the double sequence identification Two Sample Logo web server [54] (http://www.twosamplelogo.org/cgi-bin/tsl/tsl.cgi), to detect and display statistically significant differences between sequences. Two Sample Logos analysis displayed significant differences among the protein sequence of the malonylation sites and the protein sequence of non-malonylation sites on the datasets *E. coli*, *H. sapiens* and *M. musculus*, are shown in Fig. 2, respectively. Lysine is situated in the centre of the sequence, and positions of flanking residue amino acids that are significantly enriched or depleted are described in the range of -12 to +12.

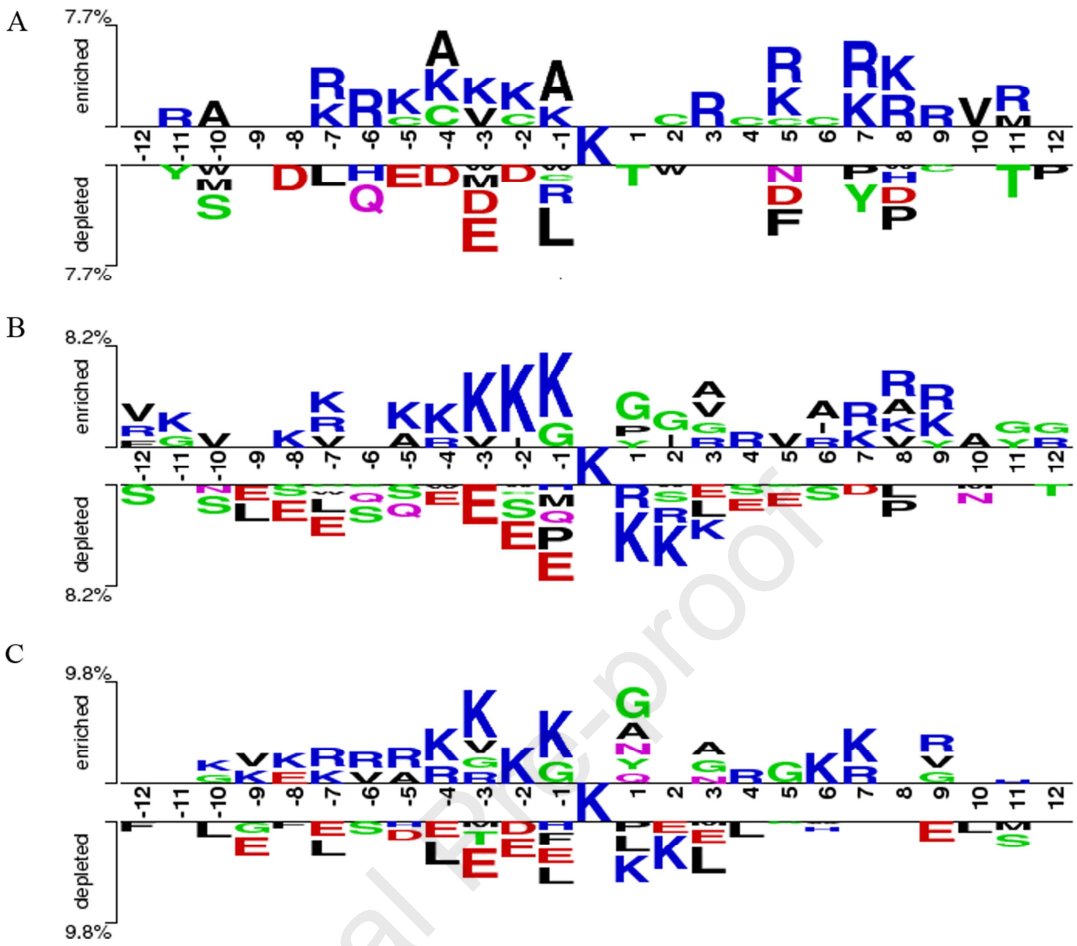


Fig. 2. Comparison of sequence identification of malonylation and non-malonylation sites in (A) *E. coli*, (B) *H. sapiens* and (C) *M. musculus* on the training datasets.

Fig. 2 shows significant differences in amino acid composition near the lysine malonylation sites and non-malonylation sites in the training datasets *E. coli*, *H. sapiens* and *M. musculus*. A detailed analysis can be seen in Supplementary Si3. The frequency of amino acid residues around lysine is significantly different, which is helpful to identify the malonylation sites of lysine.

3.2. Prediction performance of deep learning for protein malonylation sites

Recently, more and more attention is paid to the analysis and prediction of biological sequences using deep learning methods, with the accumulation of a large number of genomic data and the breakthrough of deep learning methods in the research field. In view of deep learning such as deep neural networks [23], recurrent neural networks [55], methods DL and DL-1, we predict the malonylation sites of proteins. The deep neural network includes five fully connected layers. The recurrent neural network includes two Gated Recurrent Units layers and two fully connected layers. Method DL-1 includes four layers, one convolutional layer, and three fully connected layers. Method DL includes five layers, two convolutional layers, and three fully-connected layers. On account of the binary cross-entropy loss function Adam, the parameters in the deep learning network are trained and optimized. To avoid

overfitting and reducing the complicacy of the model, a simple and powerful dropout mechanism of regularization technology is introduced and prevents overfitting, which randomly discards some neurons and their connections during the training procedure, in order to increase the generalization performance of the model. Through EAAC, EGAAC, DDE, KNN and BLOSUM62 feature extraction algorithm, after fusion, each sample gets 1431-D feature vector. Through 10-fold cross-validation, the fused feature vector is input to a SVM with a radial basis kernel function [47], XGBoost [56], DL-1, deep neural network, recurrent neural network and DL learning framework for feature learning and prediction, the prediction results are obtained on the datasets *E. coli*, *H. sapiens* and *M. musculus*, as shown in Table 4. Fig. 3 is the ROC curve obtained by the datasets *E. coli*, *H. sapiens* and *M. musculus* under different prediction models. We use the Keras package with the Tensorflow backend [57] to implement our deep learning.

Table 4

Prediction results of different classification algorithms in the datasets *E. coli*, *H. sapiens* and *M. musculus*.

Dataset	Classifier	ACC (%)	Sn (%)	Sp (%)	MCC	AUC
E. coil	SVM	61.91	59.60	64.21	0.2392	0.6716
	XGBoost	64.18	64.62	63.74	0.2840	0.7002
	DNN	8297	97.67	68.26	0.6898	0.9267
	RNN	8582	94.10	77.54	0.7273	0.8867
	DL-1	89.51	92.60	86.42	0.7930	0.9300
	DL	93.01	91.71	94.31	0.8607	0.9513
H. sapiens	SVM	64.33	61.53	67.12	0.2872	0.6980
	XGBoost	69.00	67.34	70.66	0.3803	0.7674
	DNN	85.47	95.57	75.36	0.7242	0.9226
	RNN	85.00	89.58	80.43	0.7032	0.9031
	DL-1	86.50	85.12	87.88	0.7312	0.9163
	DL	90.92	91.61	90.22	0.8186	0.9447
M. muculus	SVM	64.12	62.28	65.96	0.2828	0.7002
	XGBoost	69.49	69.92	69.07	0.3902	0.7722
	DNN	87.70	96.63	78.76	0.7670	0.9431
	RNN	86.14	80.29	91.99	0.7283	0.9199

SVM	61.91	59.60	64.21	0.2392	0.6716
DL-1	88.91	87.39	90.42	0.7800	0.9282
DL	91.93	92.30	91.57	0.8045	0.9534

As shown in Table 4, for the dataset *E. coli*, the performance of the DL prediction model is the best. The ACC of the DL model is 3.5%-31.1% higher than DL-1, XGBoost, SVM, DNN and RNN classifiers. The MCC value of the DL model is 6.77%, 57.67%, 62.15%, 17.10% and 13.35% higher than DL-1, XGBoost, SVM, DNN and RNN, respectively. On the dataset *H. sapiens*, when using the DL model to predict the malonylation sites, the highest prediction result is obtained. In terms of the evaluation index AUC value, the DL model is increased by 2.21%-24.67% compared with other classification algorithms. On the dataset *M. musculus*, our proposed deep learning framework DL has the best prediction performance for protein malonylation sites, far exceeding SVM, XGBoost, DL-1, DNN, and RNN classifiers. In terms of prediction accuracy, the model DL is 3.02%, 22.44%, 27.81%, 4.24%, and 5.79% higher than DL-1, XGBoost, SVM, DNN and RNN, respectively. The Sn and Sp values of the model DL were 4.91% and 1.15% higher than DL-1, respectively. For the AUC value, the model DL is 2.52%-25.32% higher than DL-1, XGBoost, SVM, DNN, and RNN, respectively. In summary, on different species *E. coli*, *H. sapiens* and *M. musculus* datasets used DL model to obtain the highest prediction result. This is why deep neural networks can better learn the characteristics of malonylation sites through convolution operations and fully-connected layers. The model DL can better distinguish malonylation sites and non-malonylation sites of proteins and obtain the best prediction results.

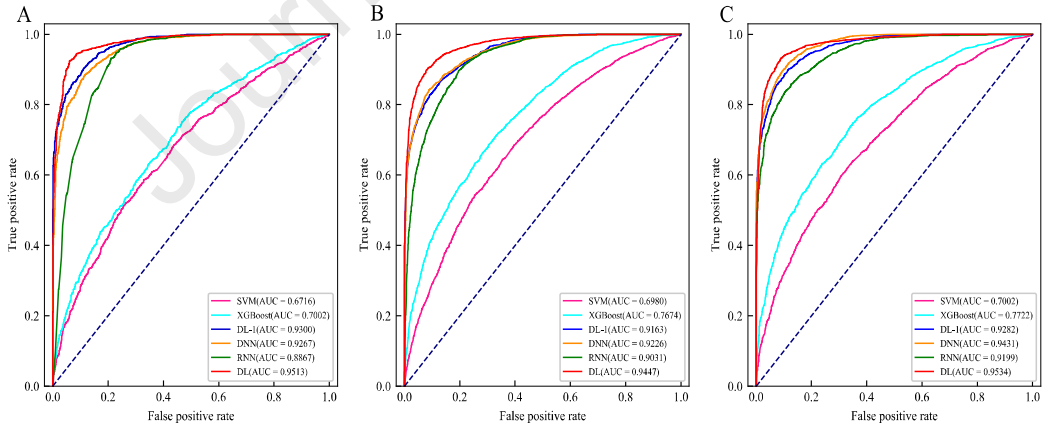


Fig. 3. Prediction performance of different classification algorithms. (A), (B) and (C) are the ROC curves of the training datasets *E. coli*, *H. sapiens* and *M. musculus*, respectively.

Fig. 3 displays that the ROC curve of DL includes the ROC curves corresponding to the classifiers SVM, XGBoost, DL-1, DNN and RNN. The area covered by the ROC curve of DL is the largest, significantly higher than other five classification algorithms, demonstrating that the classification algorithm has strong generalization ability and has good prediction performance for protein lysine malonylation and non-malonylation sites.

3.3. Visualization of learning features

For the sake of distinguishing the features produced by our deep learning framework, we visualize the features extracted by DeepMal and the original protein sequence. To visually observe the difference between malonylation and non-malonylation sites, we use visualization algorithm t-SNE [37, 58] to visualize feature vector. We compress the high-dimensional features into a two-dimensional space, and the values are normalized to $(-1,1)$. On the datasets *E. coli*, *H. sapiens* and *M. musculus*, the original sequence features and the abstract features extracted by DeepMal are visualized by t-SNE, as shown in Fig. 4.

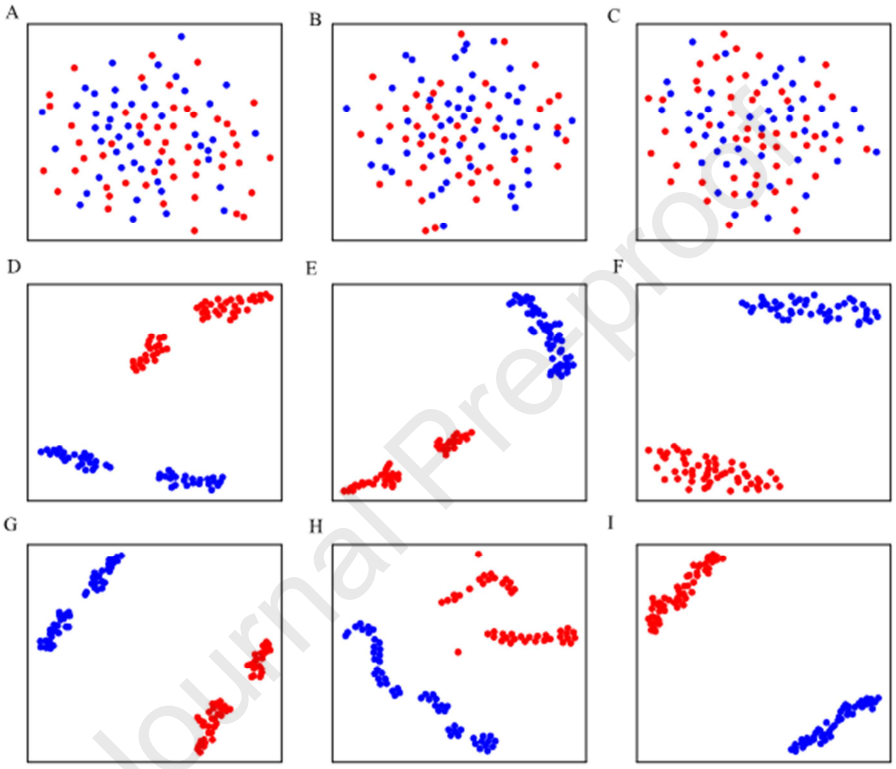


Fig. 4. On the training datasets *E. coli*, *H. sapiens* and *M. musculus*, the original sequence features and the abstract features of different layers extracted by DeepMal are visualized with t-SNE. Fig. (A), (B) and (C) are the original sequence features of the training datasets. Fig. (D), (E), (F), (G), (H) and (I) graphs are abstract features of different layers extracted by DeepMal. Blue dots represent malonylation sites, and red dots represent non-malonylation sites.

It can be seen from Fig. 4A, 4B and 4C that it is hard to distinguish the malonylation from non-malonylation sites of feature extraction of the original sequence. From the abstract representation of DeepMal, it can be seen from Fig. 4D, 4E, 4F, 4G, 4H and 4I that we can more easily classify malonylation and non-malonylation sites. By the visualization of t-SNE, we mean that the sequence of the original protein can be generated to a meaningful

representation through the non-linear change mapped by DeepMal, which contribute to further analysis of malonylation and non-malonylation sites.

3.4. Performance comparisons with other tools

For more justly evaluating the effectiveness of the prediction model for protein malonylation sites proposed in this paper, we compare the prediction performance of the tools in the same dataset. Table 5 details the comparison results of this method with other prediction methods on *E. coli*, *H. sapiens* and *M. musculus*. For the sake of more intuitively analyzing prediction performance of the independent test sets on different prediction models, we draw a bar chart of the evaluation indicators ACC, AUC, PRE and MCC value, as shown in Fig. 5.

Table 5

Comparison of prediction results of different methods on *E. coli*, *H. sapiens* and *M. musculus* independent test datasets.

Datasets	Methods	PRE	Sn	Sp	AUC	ACC	MCC
E. coli	kmal-sp [13]	0.856	0.830	0.860	0.930	0.845	0.690
	MaloPred[12]	0.798	0.750	0.810	0.870	0.780	0.561
	DeepMal	0.971	0.950	0.980	0.974	0.965	0.931
H. sapiens	kmal-sp[13]	0.867	0.849	0.870	0.944	0.860	0.720
	MaloPred[12]	0.824	0.829	0.824	0.932	0.827	0.653
	DeepMal	0.952	0.967	0.943	0.956	0.955	0.910
M. musculus	kmal-sp[13]	0.835	0.829	0.837	0.923	0.833	0.667
	MaloPred[12]	0.798	0.806	0.797	0.874	0.802	0.603

DeepMal	0.945	0.947	0.943	0.944	0.945	0.890

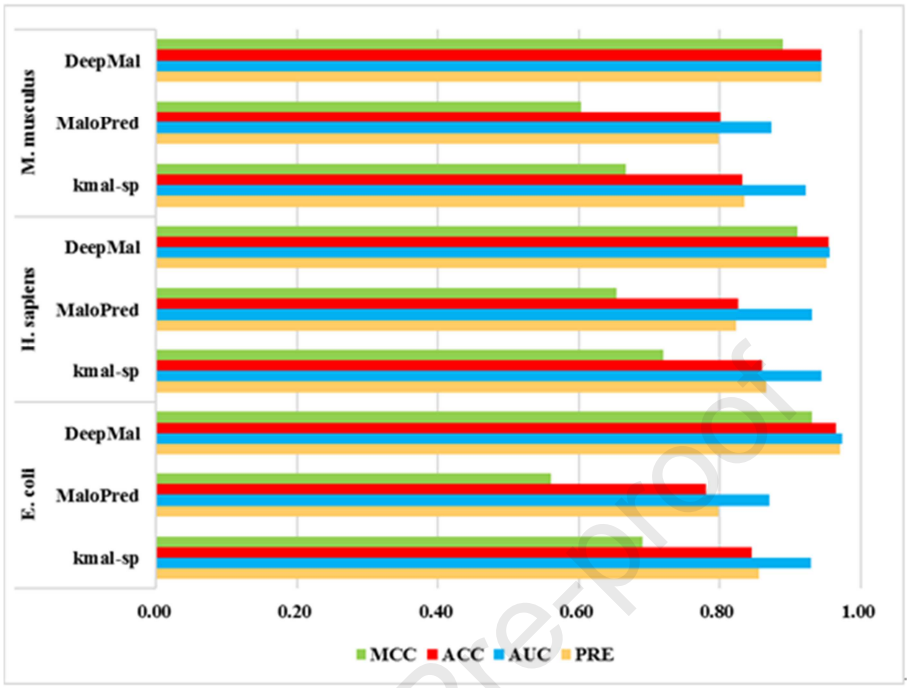


Fig. 5. Comparison of DeepMal and other methods on independent test sets *E. coli*, *H. sapiens* and *M. musculus*.

As shown in Table 5, PRE, ACC, AUC and MCC values of the DeepMal model far exceed the models MaloPred [12] and kmal-sp [13] in the independent test set *E. coli*, *H. sapiens and M. musculus*. For *E. coli* dataset, PRE of DeepMal is 0.971, which is 11.5% and 17.3% higher than kmal-sp and MaloPred, respectively. For the *H. sapiens* dataset, the AUC index of DeepMal achieved 0.956, which is 1.2%-2.4% higher than other tools. Compared with the two methods in ACC and PRE, DeepMal achieved 95.5% and 0.952, which signify the ACC and PRE have been remarkably improved 9.5%-12.8% and 8.5%-12.8%, respectively. For the *M. musculus* dataset, the prediction accuracy of the method DeepMal in this paper is 94.5%, which is higher than other prediction methods 11.2%-14.3%. The MCC values of these methods can vary from 0.5 to 0.7. However, the MCC values of the prediction tool DeepMal constructed in this paper is greater than 0.85, which is better than the other two tools, and significantly improves the prediction results of the malonylation sites of proteins. In conclusion, the prediction results show that DeepMal can significantly improve the prediction performance of protein malonylation sites. Furthermore, it shows that DeepMal tool has good generalization performance and robustness.

4. Discussion

In this study, we construct a novel model DeepMal to predict malonylation sites and non-malonylation sites. Through five kinds of feature extraction algorithms, including EAAC, EGAAC, DDE, KNN, and BLOSUM62 matrix, the character information of the protein

sequence is converted into a numerical vector. A convolutional neural network is used to extract important features related to malonylation sites. Using deep neural network and softmax classification function to predict malonylation sites and non-malonylation sites. Through 10-fold cross-validation, on the training data set *E. coil*, ACC, Sn, Sp, MCC and AUC are 93.01%, 91.71%, 94.31%, 0.8607 and 0.9513, respectively. On the *H. sapiens* dataset, ACC, Sn, Sp, MCC and AUC are 90.92%, 91.61%, 90.22%, 0.8186 and 0.9447, respectively. On the *M. muculus* dataset, ACC, Sn, Sp, MCC and AUC are 91.93%, 92.30%, 91.57%, 0.8045 and 0.9534 respectively. On the independent datasets *E. coli*, *H. sapiens*, *M. musculus*, the AUC values are 0.974, 0.956 and 0.944, and the accuracy are 96.5%, 95.5% and 94.5%, respectively. Compared with other prediction models, the prediction accuracy is increased by 9.5%-18.5%, and other evaluation indicators have also been significantly improved.

The model DeepMal can effectively predict protein malonylation sites on the independent test datasets of different species. The prediction performance of the model DeepMal is superior to the current latest malonylation sites prediction method for the following reasons:

1. Protein sequence feature information is extracted from sequence features, physical and chemical properties, and evolutionary information. The fusion of multiple features further improves the prediction performance of the model for malonylation sites.

2. The linear convolutional neural network is used to extract the specific features of malonylation sites, and then select the relevant features and reduce the feature dimension through maximum pooling.

3. The deep neural network as a classifier for predicting malonylation sites and non-malonylation sites effectively improves the prediction accuracy of the model and has good generalization performance.

Malonylation sites regulate a variety of life activities, such as glucose and fatty acid metabolism pathways. A calculation method is proposed to accurately identify the sites, to help people better understand its molecular mechanism and function, so as to reveal the mechanism of its life activity, and to screen clinical drugs for diseases.

5. Conclusion

The identification of PTM sites of proteins is of great significance to understand and reveal the structure and function of proteins, as well as the occurrence and treatment of diseases. Protein malonylation sites are participated in a multiple of physiological progress, and are related to muscle contraction, myocardial ischemia and hypothalamic regulation of appetite, and are closely related to type 2 diabetes. Constructing accurate classification models of malonylation sites is essential to help biologists understand their molecular functions and drug design based on their influence on mankind disease. This study proposes an effective prediction model DeepMal to identify the malonylation sites of proteins through

convolutional neural networks and deep neural networks. Compared with other state-of-the-art models, we have achieved significant improvements in evaluation indicators. From t-SNE visualization, we found that DeepMal can generate powerful features and distinguish between protein sequences malonylation sites and non-malonylation sites. In addition, this study helps to further promote the application of deep learning in bioinformatics research, especially protein function research. Although DeepMal can validly improve the prediction accuracy of malonylation sites, there is still room for promoting. As for our future research work, we will construct a new large-scale dataset for PTM sites prediction, and use deep learning to further improve the prediction results and operation efficiency of protein PTM sites.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Nature Science Foundation of China (Nos. 61863010, 61877064, U1806202), the Key Research and Development Program of Shandong Province of China (No. 2019GGX101001), and the Natural Science Foundation of Shandong Province of China (No. ZR2018MC007).

References

[1] G.A. Khoury, R.C. Baliban, C.A. Floudas, Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database, Sci. Rep. 1 (2011) 90.

[2] M. Matthias, O.N. Jensen, Proteomic analysis of posttranslational modifications, Nat. Biotechnol 21 (2003) 255-261.

[3] M.H. Wang, X.W. Cui, B. Yu, C. Chen, Q Ma, H.Y. Zhou, SulSite-GTB: identification of protein S-sulfenylation sites by fusing multiple feature information and gradient tree boosting, Neural Comput. Appl. 32 (2020) 13843-13862.

[4] L. Luna, V. Rolseth, G.A. Hildrestrand, M. Otterlei, F. Dantzer, M. Bjoras, E. Seeberg, Dynamic relocalization of hOGG1 during the cell cycle is disrupted in cells harbouring the hOGG1-Cys326 polymorphic variant, Nucleic Acids Res. 33 (2005) 1813-1824.

[5] M.R. Nicolls, The clinical and biological relationship between Type II diabetes mellitus and Alzheimer's disease, Curr. Alzheimer. Res. 1 (2004) 47-54.

[6] A. Visser, N. Hamza, F.G. Kroese, N.A Bos, Acquiring new N-glycosylation sites in variable regions of immunoglobulin genes by somatic hypermutation is a common feature of autoimmune diseases, Ann. Rheum. Dis. 77 (2017) 2212568.

[7] C. Peng, Z. Lu, Z. Xie, Z. Cheng, Y. Chen, M. Tan, Y. Zhao, The first identification of lysine malonylation substrates and its regulatory enzyme, Mol. Cell Proteomics 10 (2011).

[8] Y.P. Du, T.X. Cai, T.T. Li, P. Xue, B. Zhou, X.L. He, P. Wei, P.S. Liu, F.Q. Yang, T.T. Wei, Lysine malonylation is elevated in type 2 diabetic mouse models and enriched in metabolic associated proteins, Mol. Cell Proteomics 14 (2015) 227-236.

[9] Y.Y. Nishida, M.J. Rardin, C. Carrico, W.J. He, A.K. Sahu, P. Gut, R. Najjar, M. Fitch, M.K. Hellerstein, B.W. Gibson, E. Verdin, SIRT5 regulates both cytosolic and mitochondrial protein malonylation with Glycolysis as a major target, Mol. Cell 59 (2015) 321-332.

[10] Z.Y. Xie, J.B. Dai, L.Z. Dai, M.J. Tan, Z.Y. Cheng, Y.M. Wu, J.D Boeke, Y.M. Zhao, Lysine succinylation and lysine malonylation in histones, Mol. Cell Proteomics 11 (2012) 100-107.

[11] Y. Xu, Y.X. Ding, J Ding, L.Y. Wu, Y. Xue, Mal-Lys: prediction of lysine malonylation sites in proteins integrated sequence-based features with mRMR feature selection, Sci. Rep. 6 (2016) 38318.

[12] L.N. Wang, S.P. Shi, H.D. Xu, P.P. Wen, J.D. Qiu, Computational prediction of species-specific malonylation sites via enhanced characteristic strategy, Bioinformatics 33 (2017) 1457-1463.

[13] Y.J. Zhang, R.P. Xie, J.W. Wang, A. Leier, T.T. Marquezlago, T. Akutsu, G.I Webb, K.C. Chou, J.N. Song, Computational analysis and prediction of lysine malonylation sites by exploiting informative features in an integrative machine-learning framework, Brief. Bioinform. 20 (2019) 2185-2199.

[14] Q.L. Xiang, K.Y. Feng, B. Liao, Y.W. Liu, G.H. Huang, Prediction of lysine malonylation sites based on pseudo amino acid compositions, Comb. Chem. High T. Scr. 20 (2017) 622-628.

[15] G. Taherzadeh, Y.D. Yang, H.D. Xu, Y. Xue, A.W. Liew, Y.Q. Zhou, Predicting lysine-malonylation sites of proteins using sequence and predicted structural features, J. Comput. Chem. 31 (2018) 1757-1763.

[16] W. Bao, B. Yang, D. S. Huang, D. Wang, Q. Liu, Y. H. Chen, IMKPse: identification of protein malonylation sites by the key features into general PseAAC, IEEE Access 7 (2019) 54073-54083.

[17] Z. Chen, N.N. He, Y. Huang, W.T. Qin, X.H. Liu, L. Li, Integration of a deep learning classifier with a random forest approach for predicting malonylation sites, Genom. Proteom. Bioinf. 16 (2018) 451-459.

[18] Y. Li, C. Huang, L.Z. Ding, Z.X. Li, Y.J. Pan, X. Gao, Deep learning in bioinformatics: Introduction, application, and perspective in the big data era, Methods 166 (2019) 4-21.

[19] M. Wang, H. Dong, W. Zhang, W. Shu, H. Li,. An end-to-end auto-driving method based on 3D lidar. Journal of Physics Conference Series 1288 (2019), 012061.

[20] C.C. Li, B. Liu, MotifCNN-fold: protein fold recognition based on fold-specific features extracted by motif-based convolutional neural networks, Brief. Bioinform. (2019), https://doi.org/10.1093/bib/bbz133.

[21] H.X. Long, B. Liao, X.Y. Xu, J.L. Yang, A hybrid deep learning model for predicting protein hydroxylation sites, Int. J. Mol. Sci. 19 (2018) 2817.

[22] C. Savojardo, N. Bruciaferri, G. Tartari, P.L. Martelli, R. Casadio, DeepMito: accurate prediction of protein submitochondrial localization using convolutional neural networks, Bioinformatics 36 (2020) 56-64.

[23] Y.B. Xie, X.T. Luo, Y.P. Li, L. Chen, W.B. Ma, J.J. Huang, J. Cui, Y. Zhao, Y. Xue, Z.X. Zuo, J. Ren, DeepNitro: prediction of protein nitration and nitrosylation sites by deep learning, Genom. Proteom. Bioinf. 16 (2018) 294-306.

[24] B.Z. Zhang, J.Y. Li, L.J. Quan, Y. Chen, Q. Lüa, Sequence-based prediction of protein-protein interaction sites by simplified long short-term memory network, Neurocomputing 357 (2019) 86-100.

[25] N.Q. Le, T. Huynh, E.K. Yapp, H. Yeh,. Identification of clathrin proteins by incorporating hyperparameter optimization in deep learning and PSSM profiles. Comput. Meth. Prog. Bio. 177 (2019) 81-88.

[26] A. Morgat, R. Apweiler, M.J. Martin, C. Odonovan, M. Magrane, Y. Alamfaruque, R. Antunes, D. Barrell, B. Bely, M Bingley, D. Binns, L. Bower, P. Browne, C. Wm, E. Dimmer, R.Y. Eberhardt, F. Fazzini, A. Fedotov, R.E. Foulger, J.S Garavelli, C. Lg, Ongoing and future developments at the Universal Protein Resource, Nucleic Acids Res. 39 (2011) 214-219.

[27] Y. Huang, B. Niu, Y. Gao, L. Fu, W. Li, CD-HIT suite: a web server for clustering and comparing biological sequences, Bioinformatics 26 (2010) 680-682.

[28] Z. Chen, P. Zhao, F.Y. Li, A. Leier, T.T. Marquezlago, Y.N. Wang, G.I. Webb, A.I. Smith, R.J. Daly, K.C. Chou, J.N. Song, iFeature: a python package and web server for features extraction and selection from protein and peptide sequences, Bioinformatics 34 (2018) 2499-2502.

[29] Saravanan V, Gautham N, Harnessing computational biology for exact linear B-cell epitope prediction: a novel amino acid composition-based feature descriptor, Omics 19 (2015) 648-658.

[30] Y.J. Chen, C. Lu, K.Y. Huang, H. Wu, Y.J. Chen, T. Lee, GSHSite: exploiting an iteratively statistical method to identify S-glutathionylation sites with substrate specificity, PLoS One 10 (2015) e0118752.

[31] Y.J. Xiang, Q. Zou, L.L. Zhao, VPTMdb: a viral post-translational modification database. bioRxiv, (2020) https://doi.org/10.1101/2020.04.01.019562.

[32] Y. Zhu, C.Z. Jia, F.Y. Li, J.N. Song, Inspector: a lysine succinylation predictor based on edited nearest-neighbor undersampling and adaptive synthetic oversampling, Anal. Biochem. 593 (2020) 113592

[33] T.Y. Lee, Z.Q. Lin, S.J. Hsieh, N.A. Bretaña, C.T. Lu, Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences, Bioinformatics 27 (2011) 1780-1787.

[34] C.Z. Jia, M. Zhang, C.S. Fan, F.Y. Li, J.N. Song, Formator: predicting lysine formylation sites based on the most distant undersampling and safe-level synthetic minority oversampling, IEEE/ACM Trans Comput. Biol. Bioinform. (2019), https://doi.org/ 10.1109/TCBB.2019.2957758.

[35] G.D. Chen, M. Cao, K. Luo, L.N. Wang, P.P. Wen, S.P. Shi, ProAcePred: prokaryote lysine acetylation sites prediction based on elastic net feature optimization, Bioinformatics 34 (2019) 3999-4006.

[36] F.Y. Li, C. Li, T.T. Marquez-Lago, A. Leier, T Akutsu, A.W.Purcell, A.I. Smith, T. Lithgow, R.J. Daly, J.N. Song, K.C. Chou, Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome, Bioinformatics, 34 (2018) 4223-4231.

[37] B. Yu, Z.M. Yu, C. Chen, A.J Ma, B.Q. Liu, B.G. Tian, DNNAce: Prediction of prokaryote lysine acetylation sites through deep neural networks with multi-information fusion, Chemometr. Intell. Lab. Syst. 200 (2020) 103999.

[38] X.C. Wang, C. Li, F.Y. Li, V.S. Sharma, J.N. Song, G.I. Webb, SIMLIN: a bioinformatics tool for prediction of S-sulphenylation in the human proteome based on multi-stage ensemblelearning models, BMC Bioinformatics 20 (2019) 602.

[39] Q. Liu, F. Xia, Q.J. Yin, R. Jiang, Chromatin accessibility prediction via a hybrid deep convolutional neural network, Bioinformatics 34 (2018) 732-738.

[40] F.Y. Li, J.X. Chen, A. Leier, T.T. Marquezlago, Q.Z. Liu, Y.Z. Wang, J. Revote, A.I. Smith, T. Akutsu, G.I. Webb, L. Kurgan, J.N. Song, DeepCleave: a deep learning predictor for caspase and matrix metalloprotease substrates and cleavage sites, Bioinformatics 36 (2020) 1057-1065.

[41] J.Y. Yang, A.J. Ma, A.D. Hoppe, C.K. Wang, Y. Li, C. Zhang, Y. Wang, B.Q. Liu, Q. Ma, Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework, Nucleic Acids Res. 47 (2019) 7809-7824.

[42] N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (2014) 1929-1958.

[43] C. Savojardo, P.L. Martelli, P. Fariselli, R. Casadio, DeepSig: deep learning improves signal peptide detection in proteins, Bioinformatics 34 (2018) 1690-1696.

[44] Y.H. Huo, L.H. Xin, C.Z. Kang, M.H. Wang, Q. Ma, B. Yu, SGL-SVM: A novel method for tumor classification via support vector machine with sparse group LASSO, J. Theor. Biol. 486 (2020) 110098.

[45] C. Chen, Q.M. Zhang, Q. Ma, B. Yu, LightGBM-PPI: predicting protein-protein interactions through LightGBM with multi-information fusion, Chemometr. Intell. Lab. Syst. 191 (2019) 54-64.

[46] H. Shi, S. Liu, J. Chen, X. Li, Q. Ma, B. Yu, Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure, Genomics 111 (2019) 1839-1852.

[47] C. Chen, Q.M. Zhang, B. Yu, Z.M. Yu, P. J. Lawrence, Q. Ma, Y. Zhang, Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier, Comput. Biol. Med. 123 (2020) 103899.

[48] X.M. Sun, T.Y. Jin, C. Chen, X.W. Cui, Q. Ma, B. Yu, RBPro-RF: use Chou's 5-steps rule to predict RNA-binding proteins via random forest with elastic net, Chemomet. Intell. Lab. 197 (2020) 103919.

[49] Q. Zhang, S. Li, B. Yu, Q.M. Zhang, Y. Han, Y. Zhang, Q. Ma. DMLDA-LocLIFT: Identification of multi-label protein subcellular localization using DMLDA dimensionality reduction and LIFT classifier, Chemometr. Intell. Lab. Syst. 206 (2020) 104148.

[50] X.Y. Wang, B. Yu, A.J. Ma, C. Chen, B.Q. Liu, Q. Ma, Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique, Bioinformatics 35 (2019) 2395-2402.

[51] B. Yu, S. Li, W.Y. Qiu, M.H. Wang, J.W. Du, Y.S. Zhang, X. Chen, Prediction of subcellular location of apoptosis proteins by incorporating PsePSSM and DCCA coefficient based on LFDA dimensionality reduction, BMC Genom. 19 (2018) 478.

[52] W. Bao, B. Yang, D. Li, Z. Li, Y. Zhou, R. Bao, CMSENN: computational modification sites with ensemble neural network, Chemometr. Intell. Lab. Syst. 185 (2019) 65-72.

[53] X.W. Cui, Z.M. Yu, B. Yu, M.H. Wang, B.G. Tian, Q. Ma, UbiSitePred: A novel method for improving the accuracy of ubiquitination sites prediction by using LASSO to select the optimal Chou's pseudo components, Chemometr. Intell. Lab. Syst. 184 (2019) 28-43.

[54] V. Vacic, L.M. Iakoucheva, P. Radivojac, Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments, Bioinformatics 22 (2006) 1536-1537.

[55] M. Hamid, I. Friedberg, Identifying antimicrobial peptides using word embedding with deep recurrent neural networks, Bioinformatics 35 (2019) 2009-2016.

[56] B. Yu, W.Y. Qiu, C. Chen, A.J. Ma, J. Jiang, H.Y. Zhou, Q. Ma, SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting, Bioinformatics 36 (2020) 1074-1081.

[57] F. Chollet, Keras: The Python Deep Learning Library, (2018) Astrophysics Source Code Library. Available online at: https://keras.io.

[58] L.V. Der Maaten, G.E. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579-2605.

Highlights

- A novel method (DeepMal) to predict the lysine malonylation sites of proteins.
- Protein sequence feature information is extracted from sequence features, physical and chemical properties, and evolutionary information.
- Convolutional neural networks can learn to extract significant features associated with malonylation sites.
- Deep neural networks can improve the accurate prediction of protein lysine malonylation sites of proteins.
- The proposed method increases the prediction performance on independent test datasets compared with other state-of-the-art methods.

Declaration of Interest statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.