Chemometrics and Intelligent Laboratory Systems 200 (2020) 103999

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

DNNAce: Prediction of prokaryote lysine acetylation sites through deep neural networks with multi-information fusion

Bin Yu [a,b,c,*,1], Zhaomin Yu [a,b,1], Cheng Chen [a,b], Anjun Ma [d], Bingqiang Liu [e], Baoguang Tian [a,b], Qin Ma [d]

[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 Life Sciences, University of Science and Technology of China, Hefei, 230027, China
[d] Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH, 43210, USA
[e] School of Mathematics, Shandong University, Jinan, 250100, China

ARTICLE INFO

Keywords:
Prokaryote lysine acetylation sites
Multi-information fusion
Group Lasso
Deep neural networks

ABSTRACT

As a reversible and widely existing post-translational modification of proteins, acetylation plays a crucial role in transcriptional regulation, apoptosis, and cytokine signaling. To better understand the molecular mechanism of acetylation, identification of acetylation sites is vital. The traditional experimental methods are time-consuming and cost-prohibitive, and the majority of acetylation sites remain unknown. It is necessary to develpoe an effective and accurate computational approach to predict the acetylation sites, especially utilizing the advanced deep learning technique. In this study, we propose a prokaryote acetylation sites prediction method based on the deep neural networks, named DNNAce. We extract the protein features from sequence information, physicochemical information, and evolution information of amino acid residues and obtain the initial feature set All. Moreover, we use Group Lasso to remove the irrelevant features that are not effective for classification in the sites prediction area. Compared to other machine learning methods, we use deep neural networks to predict acetylation sites. The 10-fold cross-validation on independent test datasets indicates that DNNAce has the highest values of accuracy for predicting acetylation sites. That DNNAce accurately identifies acetylation sites assists us in understanding its molecular mechanism, and provides related theoretical foundation for the development of drug targets for various diseases. The source code and all datasets are available at https://github.com/QUST-AIBBDRC/DNNAce/.

1. Introduction

Protein post-translational modification (PTM) is an essential regulatory mechanism and plays a vital role in normal and pathological cell physiology [1]. Researches have shown that most proteins in a cell (e.g., those make up the cytoskeleton, signaling network, and metabolic pathways) are regulated by reversible post-translational modifications so that the cell can quickly respond to changes in the external environment and signal stimulation [2]. Hundreds of protein post-translational modification types have been discovered [3] including methylation [4,5], nitrotyrosine [6], phosphorylation [7,8], phosphoglycerylation [9], prenylation [10], ubiquitination [11] crotonylation [12], acetylation [13], and so on. The regulation of lysine acetylation on metabolism is one of the remarkable progress in the post-translational modification research field [14].

The acetylation modification regulation of metabolism occurs from prokaryotic cells to mammals including humans, which is widely present in the nucleus and cytoplasm and essential for many cell processes [15]. Acetylation regulates protein expression, stability, localization and synthesis [16], and affects gene expression and metabolism [17]. Most importantly, it is associated with specific human diseases, as abnormal KATs/KDACs function can affect cell division [18]. Protein acetylation is also common in prokaryotes [19], and many metabolic enzymes involved in central metabolism and intermediate metabolism have acetylation modification. Study of the regulation of acetylation will enhance the understanding of cell metabolism and epigenetic inheritance, promote revealing of the underlying biological processes of acetylation and its consequences, promote the development of tumor cell growth inhibitors

* Corresponding author. College of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao, 266061, China.
E-mail address: yubin@qust.edu.cn (B. Yu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.chemolab.2020.103999
Received 26 August 2019; Received in revised form 31 January 2020; Accepted 16 March 2020
Available online 19 March 2020

[20].

Identification of acetylation sites is the first step in understanding the mechanism of acetylation. Traditional experimental methods for acetylation sites detection include the radioactive chemical method [21], chromatin immunoprecipitation (ChIP) [22], and mass spectrometric detection [23]. Since these methods are time-consuming and do not recognize a large number of acetylation sites, it is urgent to develop computational methods to identify acetylation sites. Logistic regression (LR), random forest (RF), support vector machine (SVM), and other computational methods are used to predict acetylation sites.

Logistic regression is a regression analysis algorithm based on logic function. Hou et al. [24] proposed lysine acetylation sites prediction system LAceP using LR classifier in 2014. Random forest builds a forest of many decision trees, and each decision tree is unrelated. Li et al. [25] developed a method termed as SSPKA for species-specific lysine acetylation prediction, using RF classifiers that combine sequence-derived and functional features with two-step feature selection. SVM is a supervised learning model in which input samples are mapped into high-dimensional space by kernel function, and the optimal hyperplane is searched for classification. Gnad et al. [26] used the SVM to predict acetylated residues, and the precision of their acetylation sites predictor is 78%. Shi et al. [27] developed a novel algorithm called PLMLA for lysine acetylation prediction from the protein primary sequence. The prediction accuracy based on 10-fold cross-validation is 83.08%. Wuyun et al. [28] developed a novel tool, KA-predictor, to predict species-specific lysine acetylation sites based on SVM, and firstly introduced predicted half sphere exposure (HSE) features. Lee et al. [29] proposed the computational model called N-Ace based on a two-stage SVM. The predictive accuracy achieved is 5%–14% higher than trained model using only amino acid sequences. Bao et al. [30] proposed LAIPT, which identify lysine acetylation sites with polynomial tree method, and polynomial style was enriched by the physico-chemical properties of amino-acid residues. Xu et al. [31] constructed EnsemblePail to predict lysine acetylation sites based on an ensemble of support vector machine classifiers, which outperformed single support vector machine classifier. Wang et al. [32] proposed a novel computational method by using the combination of multiple kernel SVM for predicting acetylation sites, and used 10-fold cross-validation to evaluate performance.

It is worth noting that there are still many deficiencies in the acetylation sites prediction methods. First, many researchers pay attention to acetylation sites analysis in eukaryotes, ignoring the importance of acetylation in prokaryotes. Second, researchers underestimate the importance of multi-information fusion for identifying acetylation sites. Finally, the current classifiers for acetylation sites prediction are limited to the models of SVM, RF, and LR. In this study, we construct a novel deep learning model for acetylation sites prediction, DNNAce (Deep Neural Networks for Acetylation sites prediction). First, to efficiently convert protein sequence character signals into numerical signals, we employ binary encoding (BE), pseudo-amino acid composition (PseAAC), AAindex, normalized Moreau-Broto autocorrelation (NMBroto), encoding based on grouped weight (EBGW), multivariate mutual information (MMI), BLOSUM62, k nearest neighbor (KNN) to extract multiple amino acid residue information and fuse the encoded feature vectors to obtain the feature vector set. Second, Group Lasso is used to screen the optimal feature subset, for the first time, to preserve the effective and important features for improving the calculation speed of the model. Finally, the nine acetylation sites datasets of prokaryotes are predicted based on DNN and compared with AdaBoost, Naïve Bayes (NB), XGBoost, K-nearest neighbor (KNN), RF, SVM, Convolutional Neural Network (CNN) and Long Short-Term Memory Network (LSTM). The experimental results show that the method DNNAce can effectively predict the acetylation sites of different prokaryote datasets. The area under the ROC curve (AUC) of the independent test datasets are higher than 0.9, and have achieved satisfactory results.

Table 1
The numbers of positive and negative of training dataset and independent test dataset among nine species.

Species	Training dataset		Independent test dataset	
	positive	negative	positive	negative
Archaea	193	193	21	21
B.*subtilis*	1040	1040	115	115
C.*glutamicum*	1021	1021	113	113
E.*amylovora*	95	95	10	10
E.*coli*	1919	1919	213	213
G.*kaustophilus*	189	189	21	21
M.*tuberculosis*	866	866	96	96
S.*typhimuricum*	174	174	19	19
V.*parahemolvticus*	1065	1065	118	118

2. Materials and methods

2.1. Datasets

We selected the nine categories lysine acetylation sites datasets of prokaryote constructed by Chen et al. [33], which are E. *coli*, S. t*yphimurium,* Bacillus subtilis (B. *subtilis*), Vibrio parahemolvticus (V. *parahemolvticus*), *Mycobacterium tuberculosis* (M. *tuberculosis*), Corynebacterium glutamicum (C. *glutamicum*), Erwinia amylovora (E. *amylovora*), Geobacillus kaustophilus (G. *kaustophilus*), and *Archaea* (including T. *themophilus* and other species). The 9 datasets are originally from UniProtKB/Swiss-Prot (UniProt, 2016), NCBI (NCBI, 2016), CPLM database [34] and other relevant literature [35–41]. By removing the mistaken sequences with modification sites among the nine datasets, Chen et al. collected 5316 experimentally validated acetylated proteins, containing 8787 lysine acetylation sites and 87,585 lysine non-acetylation sites. Then the protein sequence was clustered using CD-HIT [42] with a 30% homology threshold to obtain 7288 lysine acetylated fragments and 41,638 lysine non-acetylated fragments. Afterward, they randomly selected 10% non-homologous lysine acetylated fragments and non-acetylated fragments as independent test datasets for nine species.

In this paper, a sequence containing an acetylation site is referred to as a positive sample, and the sequence without a known acetylation site is referred to as a negative sample. Because the number of negative samples is more than positive samples for training dataset and independent test dataset, the negative samples are randomly selected in the negative datasets, for achieving the balance of positive and negative samples. The window size of the *Archaea* dataset is 13 (−6–6), the window size of the V. *parahemolvticus* dataset is 17 (−8–8), and the sample window size of the remaining seven datasets is 21 (−10–10). When the length of the positive and negative samples is insufficient, the virtual amino acid O is defined to achieve the desired window size. The numbers of positive and negative of training datasets and independent test datasets among nine species are shown in Table 1. For the convenience of follow-up work, nine datasets *Archaea,* B. *subtilis,* C. *glutamicum,* E. *amylovora,* E. *coli,* G. *kaustophilus,* M. *tuberculosis,* S. t*yphimurium*, and V. *parahemolvticus* are denoted by A., B., C., E., E. *coli*, G., M., S., and V., respectively.

2.2. Feature extraction

Feature extraction is a critical step in the classification task, meaning that protein sequence information is converted into numerical information. In this paper, sequence-based features, physicochemical property-based features, and evolutionary-derived features are selected to extract the information of protein sequences. These can be further divided into eight subtypes: BE, PseACC, AAindex, NMBroto, EBGW, MMI, BLOUSM62, KNN.

2.2.1. Sequence-based features

BE: Binary encoding converts each amino acid residue in the protein

sample sequence into a 21-dimension numerical vector, which is a feature extraction method based on sequence information. The 20 common amino acids are encoded in the order of 'ACDEFGHIKLMNPQRSTVWY' [43]. For example, aspartic acid D is represented by (00100000000000000000), tryptophan W is encoded as the vector (00000000000000000100), and for a virtual amino acid O, which is represented by (00000000000000000001). Therefore, for a sample with a sequence window length of L, the binary encoding dimension is $21 \times L$.

PseAAC: Proteins are composed of a series of amino acid residues at different positions. PseACC avoids losing the position and sequence relationship in the amino acid sequence, which is widely used in the field of bioinformatics [44–46]. Chou et al. [47] proposed pseudo-amino acid composition to extract the feature vectors in the protein sequence, sequence p is defined the $20+\lambda$ dimensional pseudo amino acid composition as follows:

$$p = [p_1, p_2, \cdots, p_{20}, p_{20+1}, \cdots, p_{20+\lambda}]^T \tag{1}$$

Each component in the vector p is defined as follows:

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{u=1}^{20} f_u + w\sum_{k=1}^{\lambda}\tau_k} & 1 \le u \le 20 \\ \dfrac{w\tau_{u-20}}{\sum_{u=1}^{20} f_u + w\sum_{k=1}^{\lambda}\tau_k} & 20+1 \le u \le 20+\lambda \end{cases} \tag{2}$$

where w is the weighting factor, set to 0.05, τ_k is $k-tier$ sequence correlation factor, and f_u is the u -th amino acid occurrence frequency in the protein sequence ($u = 1, 2, \cdots, 20$). According to formula (2), the first 20 dimensions of the feature vector p represent the amino acid composition, and the latter λ dimensions reflect the sequence correlation factors of different levels in the amino acid sequence information. The sequence correlation factor is obtained by the physicochemical properties of the amino acids. By setting different parameter values λ, the optimal value $\lambda = 1$ can be determined from the prediction accuracy.

2.2.2. Physicochemical property-based features

AAindex: The physical and chemical properties of amino acids represent the most intuitive features of biochemical reactions and have been widely used in bioinformatics research. The AAindex database [48] collected 544 amino acid indices with a set of values for each amino acid index. If we select the full amino acid index that not only does not adequately reflect the physicochemical properties around the acetylation sites but also produce redundancy and noise information. So 12 types of physicochemical properties are selected for reference [49]. By utilizing these 12 physicochemical properties, acetylated fragments and non-acetylated fragments are converted into numerical signals, and $12\times L$ dimensional vectors are generated for samples of L window size.

NMBroto: Autocorrelation descriptors (AD) are defined by amino acids with different physicochemical properties. In this paper, eight amino acid indexes are selected from the AAindex database [48], and normalized Moreau-Broto autocorrelation (NMBroto) [50] is utilized to transform protein residues sequence into the numerical signals.

For a given protein residue sequence $P = R_1R_2R_3\cdots R_L$ of length L, the values of the eight physicochemical properties corresponding to the 20 common amino acids are normalized by equation (3):

$$\varphi_i^{(\xi)} = \frac{\varphi_i^{(\xi)} - \langle\varphi_i^{(\xi)}\rangle}{SD(\varphi_i^{\xi})} \ (\xi = 1, 2, \cdots, 8;\ i = 1, 2, \cdots, L) \tag{3}$$

where $\langle\ \rangle$ represents the average of the ξ -th physicochemical properties and SD represents the standard deviation of the ξ -th physicochemical properties.

Normalized Moreau-Broto autocorrelation is defined as follows:

$$NMBA(d) = \frac{MBA(d)}{N-d},\ d = 1, 2, \cdots, lag \tag{4}$$

where $MBA(d) = \sum_{i=1}^{N-d} P_iP_{i+d}$, P_i and P_{i+d} represent the normalization physicochemical values of the amino acid at positions i and $i+d$, respectively. lag represents the lag interval of the autocorrelation, and the NMBroto can be used to extract the $8 \times lag$ -dimensional vector of the protein sequence.

EBGW: Zhang et al. [51] proposed encoding based on grouped weight (EBGW) to extract the protein sequence based on the physicochemical properties of amino acid residues. Considering the hydrophobicity and charged character, 20 amino acid residues are divided into four different classes as follows:

neutral and non-polarity residue $C1 = \{A, F, G, I, L, M, P, V, W\}$
neutral and polarity residue $C2 = \{C, N, Q, S, T, Y\}$
acidic residue $C3 = \{D, E\}$
basic residue $C4 = \{K, H, R\}$

Combining the above four divisions then three combinations are obtained, each of which divides 20 amino acid residues into disjoint parts, $C1+C2$ vs. $C3+C4$, $C1+C3$ vs. $C2+C4$, $C1+C4$ vs. $C2+C3$. For a protein sequence $P = p_1p_2\cdots p_L$, it will be converted into three binary sequences as follows:

$$\begin{aligned} H_1(p_j) &= \begin{cases} 1\ if\ p_j \in C1+C2 \\ 0\ if\ p_j \in C3+C4 \end{cases} \\ H_2(p_j) &= \begin{cases} 1\ if\ p_j \in C1+C3 \\ 0\ if\ p_j \in C2+C4 \end{cases} \\ H_3(p_j) &= \begin{cases} 1\ if\ p_j \in C1+C4 \\ 0\ if\ p_j \in C2+C3 \end{cases} \end{aligned} \tag{5}$$

Each binary sequence is divided into J sub-sequences whose length is increased in order. For example, for H_1, the j -th sub-sequence is expressed as $X_1(j) = Sum(j)/D(j)$, where $Sum(j)$ represents the number of the number 1 in the sub-sequence of j, $D(j) = \mathrm{int}(j*L/J)$ represents the length of the j -th subsequence, and L represents the length of the protein sequence. In summary, for a protein sequence P of length L, $3*J$ -dimensional vector $X = (X_1, X_2, X_3)$ can be obtained.

MMI: Ding et al. [52] proposed the multivariate mutual information (MMI) algorithm to adequately represent the amino acids in the sequence. Based on the dipoles and volume of the side chain of the amino acid residues, the 20 amino acids are divided into seven functional groups (Table S1). Any three contiguous amino acids are regarded as a unit, Ding et al. only think about the basic ingredient of the unit and don't consider the order of three amino acids. The type of 3-gram can be represented by 'C_0, C_0, C_0' , 'C_0, C_0, C_1' , $\cdots$, 'C_6, C_6, C_6', the type of 2-gram is represented by, 'C_0, C_0', $\cdots$, 'C_6, C_6'. The number of 3-gram and 2-gram are counted by the sliding window as shown in Fig. S1.

The entropy and mutual information (MI) refer to the mutual dependence between two amino acids on the sequence. The 3-tuple MI for 3-gram is defined as follows:

$$I(a, b, c) = I(a, b) - I(a, b|c) \tag{6}$$

where a, b and c are three conjoint amino acids in a unit, the MI of $I(a, b)$ and the conditional MI of $I(a, b|c)$ are defined as:

$$I(a, b) = f(a, b)\log\left(\frac{f(a, b)}{f(a)f(b)}\right) \tag{7}$$

$$I(a, b|c) = H(a|c) - H(a|b, c) \tag{8}$$

where $f(a, b)$ is the frequency of a and b arising in 2-gram on one sequence, and $f(a)$ is the frequency of a in the sequence. $H(a|c)$ and $H(a|b, c)$ are calculated as follows:

$$H(a|c) = -f(a|c)\log f(a|c) = -\frac{f(a,c)}{f(c)}\log\left(\frac{f(a,c)}{f(c)}\right) \tag{9}$$

$$H(a|b,c) = -f(a|b,c)\log f(a|b,c) = -\frac{f(a,b,c)}{f(b,c)}\log\left(\frac{f(a,b,c)}{f(b,c)}\right) \tag{10}$$

where $f(a,b,c)$ denotes the frequency of a, b and c appearing in 3-tuples on one amino acid sequence.

To avoid the infinity of the 3-tuple and 2-tuple mutual information values, the frequency is defined as:

$$f(a) = \frac{n_a + 1}{L + 1} \tag{11}$$

where L represents the length of the sequence and n_a denotes the number of occurrence of category a appearing on the amino acid sequence. $f(a,b)$ and $f(a,b,c)$ are also calculated by a similar method.

The MI of $I(a,b,c)$ and $I(a,b)$ of the 3-tuple (84-dimensional) and 2-tuple (28-dimensional) is extracted from the amino acid sequence, respectively. A 119-dimension vector is finally generated by calculating the frequency of each category appearing on the amino acid sequence.

2.2.3. Evolutionary-derived features

BLOSUM62: The BLOSUM62 matrix is used to represent the primary sequence information of the proteins. Each residue in the training dataset is represented by the matrix containing $m \times L$ elements, where L is the length and $m = 20$ is the 20 amino acids. Each row of the normalized BLOSUM62 matrix represents one of the 20 common amino acids. The BLOSUM62 descriptor [50] can be used to encode peptides of equal length.

KNN: The KNN algorithm [53] predicts PTM sites by clustering information of local sequences, extracting features from similar sequences of positive and negative datasets to capture local sequence similarity around post-translational modification sites. Specifically, for the two sequence fragments $s_1 = (s_1(1), s_1(2), \cdots, s_1(L))$ and $s_2 = (s_2(1), s_2(2), \cdots, s_2(L))$, the distance $Dist(s_1, s_2)$ between the sequences s_1 and s_2 is defined as follows:

$$Dist(s_1, s_2) = 1 - \frac{\sum_{i=1}^{L} Sim(s_1(i), s_2(i))}{L} \tag{12}$$

$$Sim(a,b) = \frac{M(a,b) - \min\{M\}}{\max\{M\} - \min\{M\}} \tag{13}$$

where L represents the protein sequence window size, Sim is the normalized amino acid substitution matrix. M is the substitution matrix, derived from the BLOSUM62 matrix, a and b represent two amino acids, and $max/min\{M\}$ represents the maximum/minimum in the substitution matrix M, respectively.

For the query sequence $p = (p_1, p_2, \cdots, p_L)$, the corresponding KNN score is calculated in the following three steps. First, calculate the distance between the query sequence p and all comparison datasets that contain the same number of positive and negative datasets. Second, sort by distance and select k nearest neighbors. Finally, the percentage of positive neighbors (samples containing acetylation sites) among the k nearest neighbors is calculated as KNN score.

The above steps are repeated for different k values to obtain the acetylation predictor multiple features. In this paper, considering that the dataset E. contains 190 samples, for nine different acetylation site datasets, k is set to 2, 4, 8, 16, 32, 64, 128 in order. So for each protein sequence segment, the KNN encoding produces a 7-dimensional vector.

2.3. Group Lasso

To reduce the risk of over-fitting, Tibshirani [54] proposed the least absolute shrinkage and selection operator (Lasso) in 1996. However, the Lasso method does not apply to the situations where there is a connection between features, because these connected features can be analyzed as a whole. Group Lasso overcomes the shortcoming of feature selection from the group level that Lasso cannot achieve by adding constraints to a set of coefficient vectors under the condition that the coefficient vectors are grouped in advance. Each set of coefficients are selected as a "single" variable, that is, if the set of coefficients are not zero, the features corresponding to the set of coefficients are all selected; conversely, if the set of coefficients are all zero, the all features corresponding to the set of coefficients are discarded, which significantly improve the sparsity between feature groups. Group Lasso is defined as follows:

$$\arg\min_{\beta} \frac{1}{2}\|Y - \sum_{\ell=1}^{L} X_\ell \beta_\ell\|_2^2 + \alpha \sum_{\ell=1}^{L} \|\beta_\ell\| \tag{14}$$

where Y is N observation vector, X is $N \times P$ a feature matrix, β is P -dimensional coefficient vector. The P features are divided into L groups, and the submatrix X_ℓ of the matrix X represents the sign matrix matching the ℓ group ($\ell = 1, 2, \cdots, L$), the corresponding coefficient vector is represented by β_ℓ, and the block coordinate descent algorithm [55] is used to solve the Group Lasso parameters.

2.4. Deep neural networks

The basic structure of the neural network consists of an input layer, multiple hidden layers, and an output layer. The neural network structure with two or more hidden layers is called the deep neural networks (DNN). Each layer in the DNN is fully connected, and the units in the hidden or output layer are connected to all units in the previous layer (Fig. 1A) The output values are calculated sequentially along the network layer (Fig. 1B) and are converted in a non-linear manner before the output layer calculates the final output (Fig. 1C). In this paper, we select ReLU [56] as the activation function. The Adam [57] algorithm is used to optimize the classification cross-entropy loss function. The dropout rate [58] between different layers is set to 0.5. The softmax function is employed as a particular class possibility.

$$ReLU(x)\begin{cases} x \; for \; x < 0 \\ 0 \; for \; x \geq 0 \end{cases} \tag{15}$$

$$u_k = \frac{e^{y_k}}{\sum_{i=1}^{P} e^{y_i}} \tag{16}$$

where u_k is the output of the k -th neuron, representing the observed probability of class k, y^k is the associated linear output from the previous hidden layers; and P is the total number of output neurons in the softmax layer. Our model is implemented by Keras and TensorFlow, and the parameter tuning of DNN is shown in Table 2.

2.5. Performance evaluation

In statistical prediction, independent dataset test [59,60], K-fold cross-validation [61,62], and jackknife test [63,64] are often used to examine predictor for its effectiveness. In this paper, we use the 10-fold cross-validation method to evaluate the performance of the model. To more intuitively measure the quality of the prediction model, four metrics are usually used in literature [65]: sensitivity (Sn), specificity (Sp), overall accuracy (ACC) and Matthew's correlation coefficient (MCC), which are defined as follows:

$$Sn = \frac{TP}{TP + FN} \tag{17}$$

$$Sp = \frac{TN}{FP + TN} \tag{18}$$

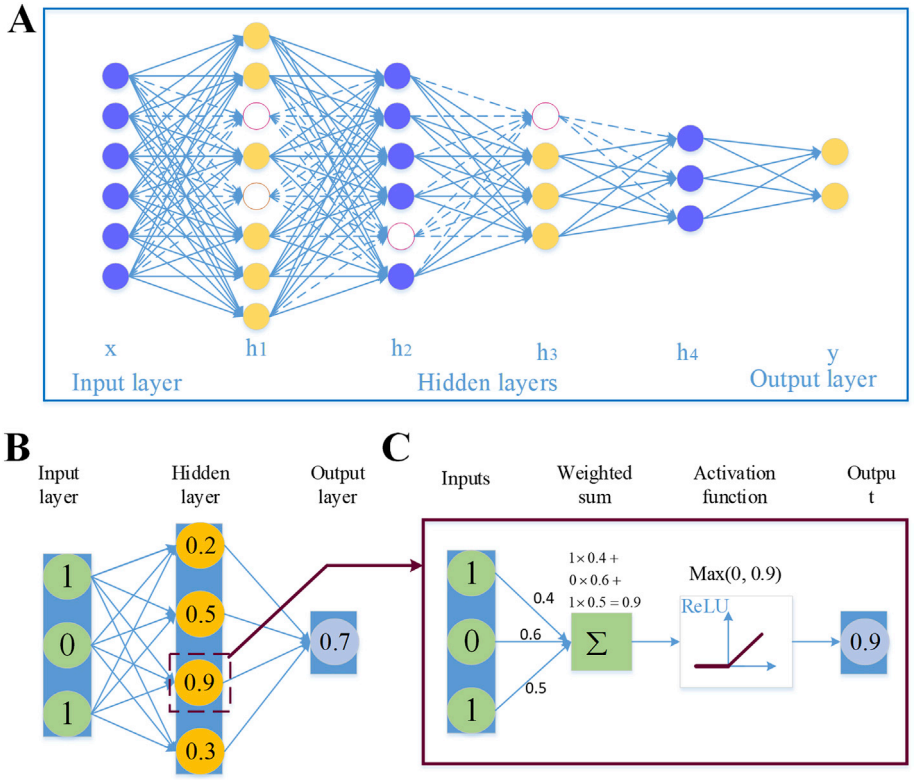

Fig. 1. DNN network structure and network training process. (A) shows the underlying network structure of the DNN, including the input layer unit, the four hidden layer units, and the output layer unit. (B) shows the calculation of the output of each layer of hidden layers. (C) shows the calculation of the output value by a nonlinear activation function.

Table 2
Parameters range and settings of the DNN.

Name	Range	Setting
Learning rate	1, 0.1, 0.01, 0.001, 0.0001	0.01
Weight initialization	uniform, normal, glorot_normal, glorot_uniform, lecun_uniform,he_normal, he_uniform	glorot_normal
Per-parameter adaptive learning rate methods	SGD, RMSprop, Adagrad, Adadelta, Adam, Adamax, Nadam	Adam
Activation function	relu, tanh, sigmoid, softmax, softplus, softsign, hard_sigmoid	relu
Dropout rate	0.1, 0.2, 0.5, 0.8	0.5

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{19}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{20}$$

where *TP*, *TN*, *FP* and *FN* respectively denote the number of true positives, true negatives, false positives, and false negatives. The receiver operating characteristic (ROC) curve is an important index to measure the robustness of the model, which is widely used in the bioinformatics field [66–68]. When the curve crosses, the AUC values can be more reasonable for the predictive model. Therefore, we also use ROC and AUC to measure the quality of the prediction model.

2.6. Illustration of the DNNAce workflow

The workflow of DNNAce is shown in Fig. 2, with the detailed steps listed below.

Step 1: The nine prokaryote acetylation sites datasets are obtained, and the positive and negative datasets of the protein sequence and the corresponding category labels are input.
Step 2: Feature extraction. By encoding the character signal into a numerical signal, the protein sequence features are extracted from the sequence information, physicochemical information and evolution information of the amino acid residue and the optimal parameters in the PseAAC, NMBroto and EBGW feature encoding methods are determined. Then the eight extracted features are combined to obtain an initial feature set.
Step 3: Feature selection. For the initial feature space All, Group Lasso is applied to remove redundant and irrelevant information, and to preserve the essential features related to classification.
Step 4: According to ***Step 2*** and ***Step 3***, the optimal feature subset and the corresponding class labels are input into the DNN classifier, and the acetylation sites of the nine datasets are predicted.
Step 5: Model performance evaluation. The AUC, ACC, Sn, Sp, and MCC values are calculated using 10-fold cross-validation, and the ROC curve is drawn to evaluate the prediction performance of the model.
Step 6: Using the model built in ***Step 1-Step 5***, the model is tested using independent test datasets.

3. Results and discussion

All the experiments are carried out on a Windows Server 2012R2 Intel (R) Xeon (TM) CPU E5-2650 @ 2.30 GHz 2.30 GHz with 32.0 GB of RAM, MATLAB2014a and Python 3.6 programming implementation.

3.1. Analysis of the difference between amino acids near acetylation sites and non-acetylation sites in 9 datasets

For the analysis of sequence bias information around prokaryote

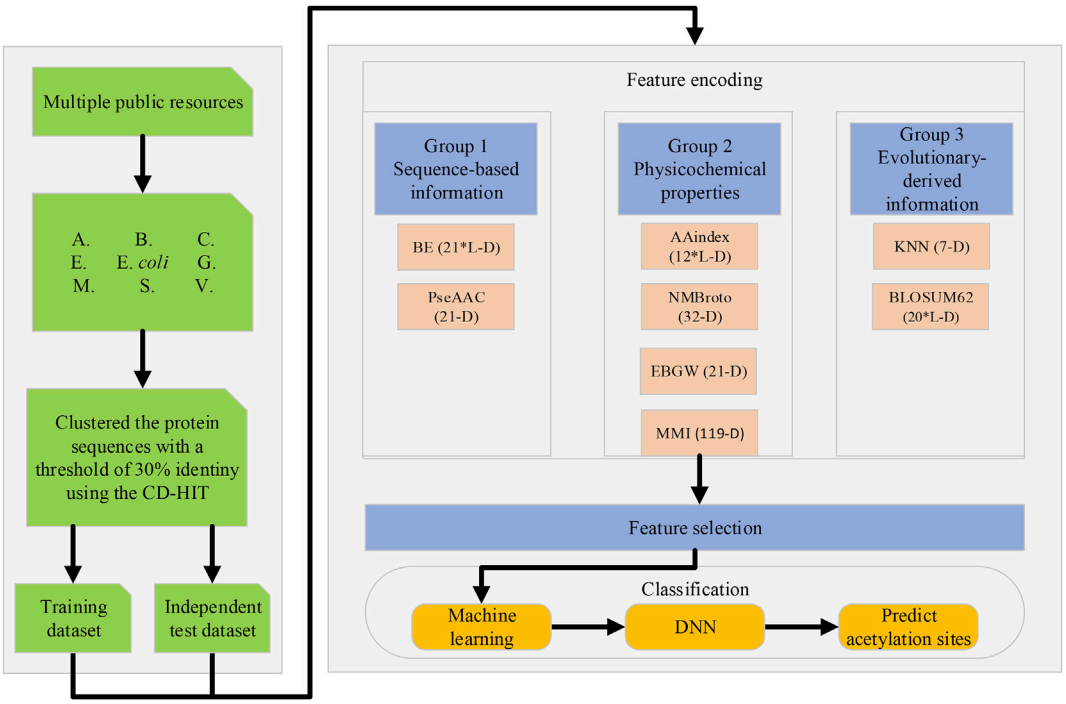

Fig. 2. The flowchart of DNNAce, which include datasets construction, feature extraction and feature selection, classifier modeling and acetylation sites prediction. Datasets construction: through the CD-HIT tool and other operations, the nine datasets are divided into training datasets and independent datasets. Feature extraction and feature selection: optimal parameters in PseAAC, NMBroto, and EBGW are determined and group lasso is applied to select features. Classifier modeling and acetylation sites prediction: we employ DNN to predict acetylation sites and use AUC, ACC, Sn, Sp, and MCC to evaluate model.

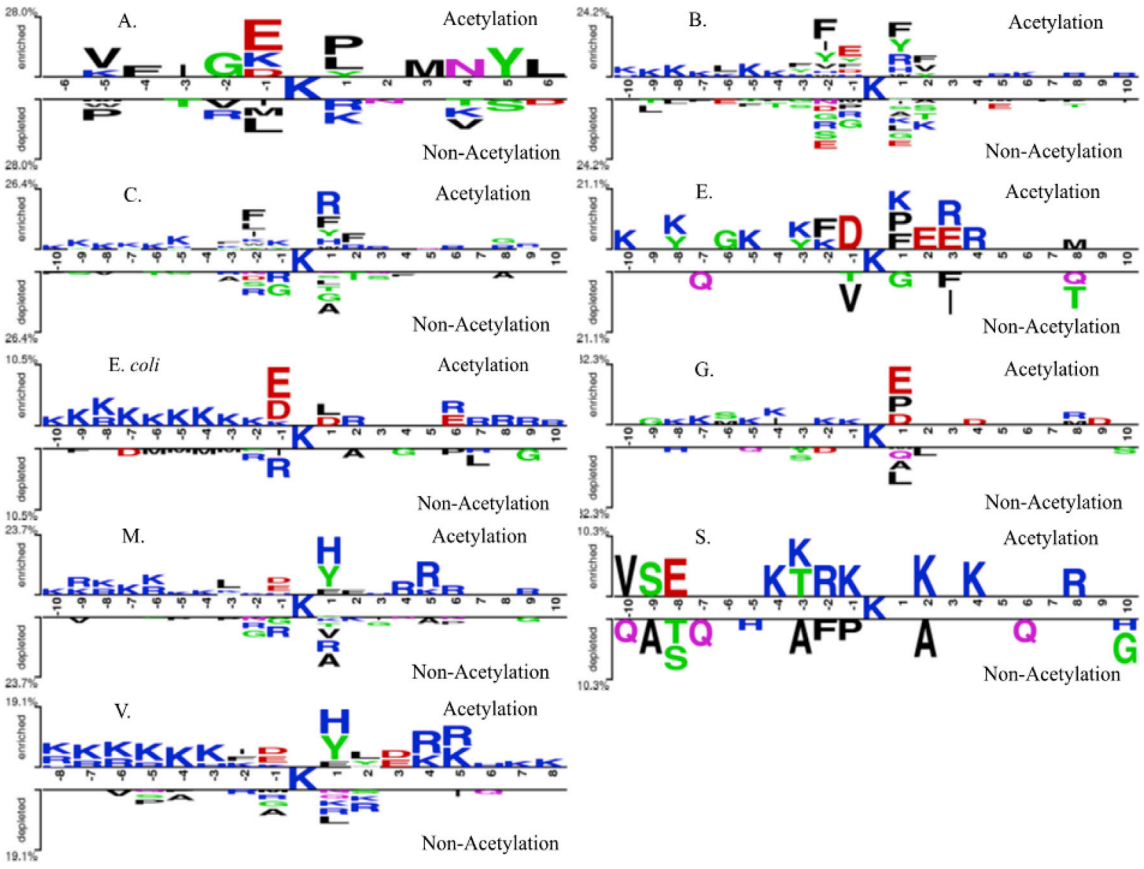

Fig. 3. Comparison of the amino acid composition of the nine prokaryote datasets near the acetylation sites and the non-acetylation sites.

Table 3
ACC values corresponding to different λ in PseAAC.

ACC (%)	A.	B.	C.	E.	E. *coli*	G.	M.	S.	V.
$\lambda = 1$	**58.57**	60.62	59.06	60.83	**58.10**	56.89	60.79	55.25	64.89
$\lambda = 2$	54.68	60.77	59.31	60.33	57.19	52.37	61.60	57.19	**66.07**
$\lambda = 3$	55.70	**60.96**	58.97	55.67	57.56	54.53	62.41	**57.48**	65.55
$\lambda = 4$	57.03	60.87	60.83	56.67	56.49	**57.44**	**63.28**	53.73	65.64
$\lambda = 5$	58.30	59.28	**60.92**	**62.11**	56.75	54.17	61.14	55.42	65.18
$\lambda = 6$	52.59	60.24	60.63	58.67	58.03	54.72	61.83	53.46	65.41

acetylation sites, we use Two Sample Logos [69] (http://www.twosamplelogo.org/cgi-bin/tsl/tsl.cgi), which is used to analyze the amino acid composition near the post-translational modification sites of proteins [70], to obtain a comparison of the double sequence identifiers of the nine datasets.

It can be seen from Fig. 3 that the statistical significance of the relative position frequencies of the nine datasets sequences is different, and there is also a significant distinction between the local comparisons of each dataset sequence. For example, the dataset A. near the acetylation sites, the negatively charged residue glutamic acid (E), the non-polar amino acid M, and the polar amino acids N and Y are significantly more at positions −1, 3, 4, and 5, respectively. The acetylation sites of dataset S. is near the lysine, the non-polar amino acid V and the polar amino acid S frequency of occurrence are higher at positions −10 and −9, and there are no significant amino acids in the depletion positions −6, −4, 1, 3, 4, 5, 7, 8, and 9. Based on these characteristics, it can be inferred that there is an apparent difference between the amino acids near the acetylation sites and non-acetylation sites in nine datasets.

According to the analysis in EBGW, we also find that acetylation sites and non-acetylation sites have distinct physicochemical characteristics. It can be intuitively found that the overall distribution of EBGW values between different datasets is approximately the same (Supplementary Fig. S2). The ratio of EBGW values of H_1 and H_2 between acetylation and non-acetylation is less than 0, and the ratio of EBGW values of H_3 is greater than 0, indicating more positive charge existed around the acetylation sites than that around the non-acetylation ones. It also shows that the subtle differences between acetylation and non-acetylation of each dataset (Supplementary Fig. S2). For dataset E., EBGW values indicate a more significant difference in H_1 and H_3, and there is a small difference in H_2. But for dataset V., the curves show a slight difference in H_1 and H_2, and much more difference can be observed in H_3.

3.2. Determination of optimal parameter λ in PseAAC

PseAAC method needs to select the optimal parameter, which has a crucial impact on the model construction. For different parameter values, the prediction accuracy of the acetylation sites is used as a measure to select the best feature parameter. Since the samples of the training dataset contain the virtual amino acid O, the value of λ is set 1, 2, 3, 4, 5, and 6, respectively, and the corresponding accuracy values of the nine prokaryote datasets under different values λ are shown in Table 3.

It can be seen from Table 3 that for different datasets, the corresponding λ value is inconsistent when the accuracy rate is the highest. For example, for datasets A. and *E. coli*, when λ is 1, the accuracy ACC reaches the highest value of 58.57% and 58.10%, respectively. For datasets C. and E., when the λ value is 5, the accuracy ACC reaches the maximum value. Considering comprehensively, we analyze the overall prediction accuracy of the nine datasets under different λ values. The overall prediction accuracy reaches peak value at interval value of $\lambda = 1$. Therefore, when we use the PseAAC algorithm for extracting protein sequences feature, the optimal parameter λ is determined to be 1, and each protein sequence generates a 21-dimensional feature vector.

3.3. Determination of optimal parameter lag in NMBroto

NMBroto is defined by the physicochemical properties of the amino acids in the sequence. Since the sequence samples with virtual amino acids in the dataset, the interval *lag* in NMBroto is set to 1, 2, 3, 4, 5, and 6, respectively. The ACC values of different *lag* on nine acetylation sites datasets are shown in Table 4.

As can be seen from Table 4, for nine acetylation sites datasets, the corresponding NMBroto interval *lag* is inconsistent when the accuracy reaches the highest. For dataset A., when the *lag* is 6, the accuracy reaches the highest value of 65.96%; and for the dataset *E. coli*, when the *lag* is 5, the highest prediction accuracy 51.62% is reached. We comprehensively consider the influence of different interval values on the prediction results. By analyzing the prediction accuracy of the nine datasets in NMBroto with different *lag*, when the *lag* is 4, the overall prediction accuracy of the model is the highest. Considering that the optimal parameter of the NMBroto is set to 4 when we use NMBroto to extract the feature of the protein sequence, each protein sequence generates a 32-dimensional feature vector.

3.4. Determination of optimal parameter J in EBGW

For the dataset A., the sample window size is 13, so the number of sub-sequences in the EBGW is set to 1, 2, 3, …, 12, respectively. For different parameter values, the DNN is selected to classify the acetylation sites. The ACC values corresponding to the different subsequence numbers for the nine datasets are shown in Table 5.

It is known from Table 5 that for the prediction of nine prokaryote acetylation sites datasets, the corresponding subsequence number of EBGW is inconsistent when the accuracy rate is the highest. For example, for dataset A., when the J is 13, the accuracy reaches the highest value. For dataset G., dataset M., and dataset V., when the J is 7, the ACC reaches the maximum, 56.64%, 64.09%, and 67.14%, respectively. Overall, by analyzing the overall prediction accuracy under the different subsequence number of 9 datasets, we find when the number of the subsequence is 7, the overall prediction accuracy of the nine datasets achieve the maximum ACC value. Therefore, when the ENGW algorithm is selected to extract the feature of the protein sequence, a $3 \times 7 = 21$ dimensions feature vector is obtained for each protein sequence.

3.5. The effect of feature extraction algorithm

Using a single feature encoding method, the obtained feature vector is too monotonous to characterize acetylation sites information. The different sequence information, physicochemical properties, and evolutionary information can complement each other. To acquire more feature information of amino acid residues, the eight feature methods of each dataset are fused to obtain the initial feature space called All. To analyze the influence of different feature extraction methods and feature fusion for acetylation sites recognition. The dimensions of the individual features and fusion feature All about nine datasets are detailed as shown in Supplementary Table S2. The nine datasets prediction accuracy of various features and feature fusion All is shown in Table 6.

It can be seen from Table 6 that the corresponding prediction accuracy of the nine acetylation sites datasets is different under different feature extraction methods. For nine different datasets, we do not find any feature extraction method that exceeds the others. In general, in the single feature encoding method, the BE feature based on sequence information, the AAindex feature based on physicochemical properties and

Table 4
ACC values corresponding to different *lag* in NMBroto.

ACC (%)	A.	B.	C.	E.	E. *coli*	G.	M.	S.	V.
lag = 1	55.22	54.33	51.52	45.61	51.09	46.84	51.44	52.63	51.22
lag = 2	57.63	54.23	52.60	**52.61**	51.43	**50.51**	52.71	55.03	53.16
lag = 3	58.66	**56.01**	53.33	49.67	50.42	46.55	**54.50**	**55.23**	**54.75**
lag = 4	61.74	54.66	55.93	51.39	51.38	46.02	53.23	54.12	54.51
lag = 5	60.46	54.62	**56.08**	51.39	**51.62**	46.61	52.53	51.75	53.95
lag = 6	**65.96**	53.75	54.36	49.22	50.08	43.99	54.16	52.94	54.65

Table 5
ACC values corresponding to different *J* in EBGW.

ACC (%)	A.	B.	C.	E.	E. *coli*	G.	M.	S.	V.
J = 1	51.05	53.08	50.78	52.72	50.65	51.32	53.07	52.06	52.62
J = 2	52.11	58.75	56.32	53.39	54.30	54.23	60.10	53.17	61.64
J = 3	53.66	61.01	58.42	59.39	58.23	52.11	62.07	52.27	63.77
J = 4	51.03	**62.26**	59.16	60.39	57.04	49.49	62.82	55.41	64.09
J = 5	57.05	62.12	58.57	60.22	56.57	53.70	62.82	56.91	65.83
J = 6	56.18	60.53	58.47	61.78	58.29	54.80	62.99	58.45	65.23
J = 7	58.36	62.07	58.87	62.56	57.58	**56.64**	**64.09**	58.97	**67.14**
J = 8	56.21	60.29	60.09	64.17	57.92	53.77	61.95	58.38	66.21
J = 9	58.07	61.25	60.43	**65.72**	**59.25**	55.32	62.94	56.18	64.38
J = 10	57.01	60.34	**60.73**	63.11	56.72	55.85	63.04	56.73	66.16
J = 11	57.50	61.35	59.01	63.50	58.10	54.01	63.34	**61.24**	66.21
J = 12	60.33	61.30	59.26	64.61	57.61	52.60	63.05	60.15	65.98
J = 13	**63.74**	60.72	59.55	64.50	57.77	53.23	63.51	58.68	58.68

Table 6
ACC values corresponding to different feature extraction methods.

ACC (%)	A.	B.	C.	E.	E. *coli*	G.	M.	S.	V.
BE	62.03	67.12	66.95	61.5	55.34	53.77	66.98	53.42	67.66
PseAAC	58.57	60.63	59.06	60.83	58.10	56.89	60.79	55.25	64.89
AAindex	66.09	66.20	63.28	59.94	55.50	52.89	67.50	58.06	65.78
NMBroto	61.74	54.66	55.93	51.39	51.38	46.02	53.23	54.12	54.51
EBGW	58.36	62.07	58.87	62.56	57.58	56.64	64.09	**58.97**	67.14
MMI	58.34	55.77	54.80	58.22	55.21	50.48	59.42	53.46	60.38
BLOSUM62	63.84	65.87	65.82	56.94	55.89	54.05	68.08	51.09	66.34
KNN	**66.14**	65.34	62.15	54.28	52.87	49.47	67.68	50.02	66.21
All	63.03	**67.50**	**68.86**	**67.78**	**60.63**	**57.19**	**68.47**	57.01	**69.96**

the BLOSUM62 feature based on evolutionary information are three relatively important feature extraction methods, which can better extract protein sequence information and have higher ACC values. For the AAindex feature extraction method, the prediction accuracy of the dataset M. is the highest, which is 67.50%, and the prediction accuracy of the dataset G. is the lowest, which is 52.89%. For the EBGW feature encoding method, the prediction accuracy of the dataset V. is the highest, which is 67.14%, and the dataset G. has the lowest prediction accuracy, which is 56.64%. By combining eight feature encoding methods, the initial feature dataset All is obtained. In addition to the dataset A. and dataset S., the prediction accuracy of the other seven acetylated datasets is improved to some extent. They reached 67.50%, 68.86%, 67.78%, 60.63%, 57.19%, 68.47%, and 69.96%, respectively. It shows that instead of using a single feature encoding method alone, it is necessary and important to use various feature types, because a single feature type has limited predictive power of the trained model.

3.6. Feature selection using Group Lasso

Group Lasso is used to select the optimal feature subset, and parameter α is set to 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, and 0.1, respectively. When the value of α is set too large, more "sparse" solutions are generated, and the variables with lower correlation are removed. When the value of α is small, all feature attributes will be selected. As α increases, the corresponding dimension of each dataset decreases, and the optimal parameters in Group Lasso are selected according to the prediction accuracy. The ACC values of the datasets A., E., E. *coli* and M. reach the maximum when the parameter α value is 0.03, which are 84.47%, 96.89%, 63.08%, and 76.62%, respectively. When the parameter is 0.04, the ACC value of the dataset B. reaches the maximum, and the ACC value of the dataset V. reaches the maximum at the same time when the parameters are 0.03 and 0.04. The ACC values of datasets C., G., and S. reach the maximum when parameters are 0.05, 0.1, and 0.06, respectively. Therefore, 0.03 is the optimal parameter of the Group Lasso algorithm (Supplementary Table S3). The composition and distribution of different features in each optimal feature set as shown in Fig. 4.

By analyzing the composition and distribution of different features in the optimal feature subset, we can visually find that the feature composition and distribution of the nine datasets optimal feature subset roughly similar, but there are still some subtle differences. We analyze the percentage of each type of feature in the optimal feature subsets, and the necessity and importance of effective combinations of feature types from various aspects are again illustrated. In addition to the datasets A., C., and E. *coli,* the evolutionary features KNN of the remaining six datasets have the largest percentage (Fig. 4A). Although the evolutionary feature KNN dimension of the datasets is only 7-dimensional in this paper, the important features of KNN are still selected and included in the optimal feature subsets, indicating that the evolutionary information affects the recognition of acetylation sites to some extent. Specifically, for the nine different species datasets, the proportion of the sequence feature BE to the optimal feature set reaches the maximum, both of which reach more than 40% (Fig. 4B). The evolutionary feature BLOSUM62 is the second

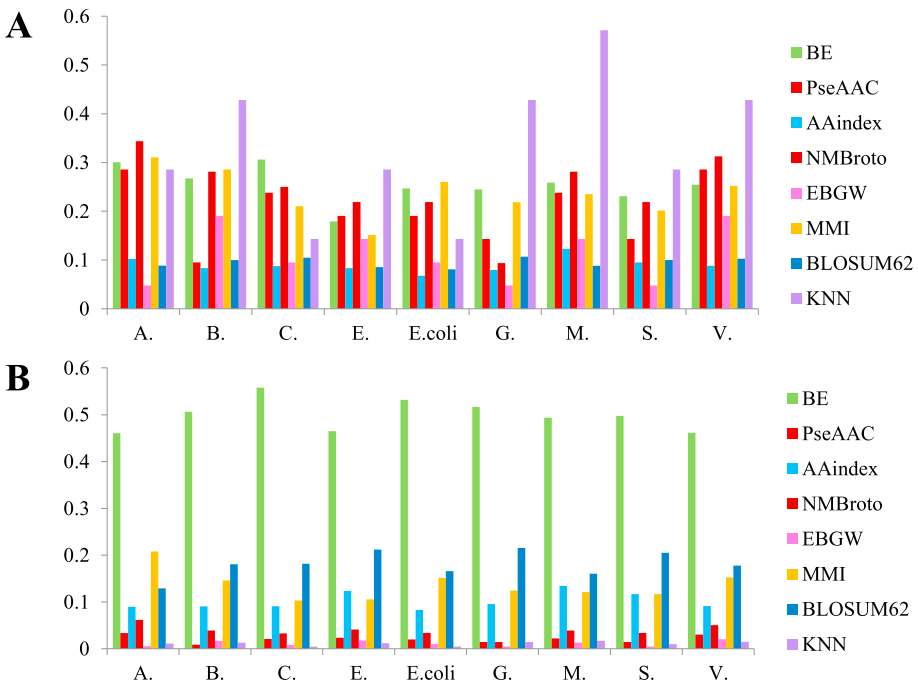

Fig. 4. Analysis of the optimal feature set of the nine datasets. (A) the percentage of each type of feature in the optimal feature sets, and (B) the proportion of each type of feature selected in the optimal feature sets.

most crucial information for acetylation sites prediction, and then the MMI based on physicochemical information accounts for a large proportion in the optimal feature subset. Finally, we also noticed that the physicochemical feature AAindex represents the most significant proportion in the optimal feature subsets (Fig. 4B). The result is also consistent with the result of section 3.5, that is, the single feature BE, AAindex and BLOSUM62 have high prediction accuracy for different datasets.

3.7. Comparison of Group Lasso and other dimensional reduction methods

Although multi-information fusion improves the prediction performance of the model to a certain extent, it also brings redundant feature information, which affects the classification accuracy of the model and reduces the calculation speed. We select singular value decomposition (SVD) [71], mutual information (MI) [72], information gain (IG) [50], Extra-Trees (ET) [73], Elastic net [74], logistic regression (LR) [75], and Group Lasso to reduce the initial feature space dimension and the difficulty of learning task. The comparison of the overall prediction accuracy and dimensions of sevendimensional reduction methods for the nine datasets are shown in Supplementary Table S4.

Different feature selection methods have different dimensional reduction effects on the initial feature space of the nine datasets. When MI and IG are used to select feature, the dimension is consistent with Group Lasso. The feature subsets of Elastic net and ET still have higher dimension, and LR feature selection method only significantly reduces the dataset E. feature set dimension. However, SVD and Group Lasso can significantly reduce the dimensions of the original feature set for nine datasets. We analyze the different feature selection algorithms ACC values of the nine datasets, and the results show that the determination of the best feature subsets is closely related to the dimensional reduction method. Most of the models trained on feature subsets have improved performance compared to models trained on a complete feature set. Group Lasso has a better dimensional reduction effect compared with the other six-dimensional reduction methods, which ACC values reach 84.47%, 73.89%, 75.38%, 96.89%, 63.08%, 89.15%, 76.62%, 90.51%, 75.46%, respectively. At the same time, the ROC curves corresponding to different dimensional reduction methods of nine datasets are shown in Fig. 5.

It can be seen intuitively from Fig. 5 that for the nine datasets, the ROC curve corresponding to the Group Lasso feature selection method contains the ROC curves corresponding to other methods and the AUC values are 0.9098, 0.8163, 0.8228, 0.9968, 0.6747, 0.9624, 0.8409, 0.9623, 0.8315, respectively. The AUC values corresponding to the datasets B., C., E. *coli*, M., and V. are 4.06%, 5.34%, 3.3%, 5.28%, and 4.79% higher than the values corresponding to LR, respectively. The AUC values corresponding to the datasets A., E., G., and S. are 25.82%, 38.68%, 30.74%, and 27.54% higher than the values corresponding to SVD, respectively. Group Lasso compresses the corresponding group of regression coefficients to select important group variables, which provides good feature information for the input DNN and improving the accuracy of prediction results. This paper chooses Group Lasso as the best feature selection algorithm, which considers multiple variables as one entire group.

3.8. Influence of different classifiers on prediction results and feature visualization analysis

To construct a highly efficient acetylation sites prediction model, we comprehensively utilize various types of information to generate strong discriminative ability features and select the DNN with strong generalization ability by comparing other classifiers. The best feature subsets selected by Group Lasso are input into AdaBoost [76], NB [77], XGBoost [78], KNN [79], RF [80], SVM [81], CNN [82], LSTM [83] and DNN to identify the acetylation sites, respectively. The following is the parameter setting of comparison classifiers: both the AdaBoost algorithm and the NB algorithm use default parameters. The learning rate of XGBoost is set to 0.01, and the number of iterations is 500. The Euclidean distance is used in the KNN algorithm, and the number of neighboring points is 10. In the RF, the Gini coefficient is selected to divide the node, and the number of decision trees is set to 500. The polynomial is selected as a kernel function in the SVM algorithm. CNN using two convolutional and pooling layers is followed by a fully connected output layer. LSTM includs two long short-term memory (LSTM) layers and one fully connected layer and each layer used the ReLU as the activation function. DNN has four hidden layers, and the detailed parameter settings are shown in Table 2. The nine

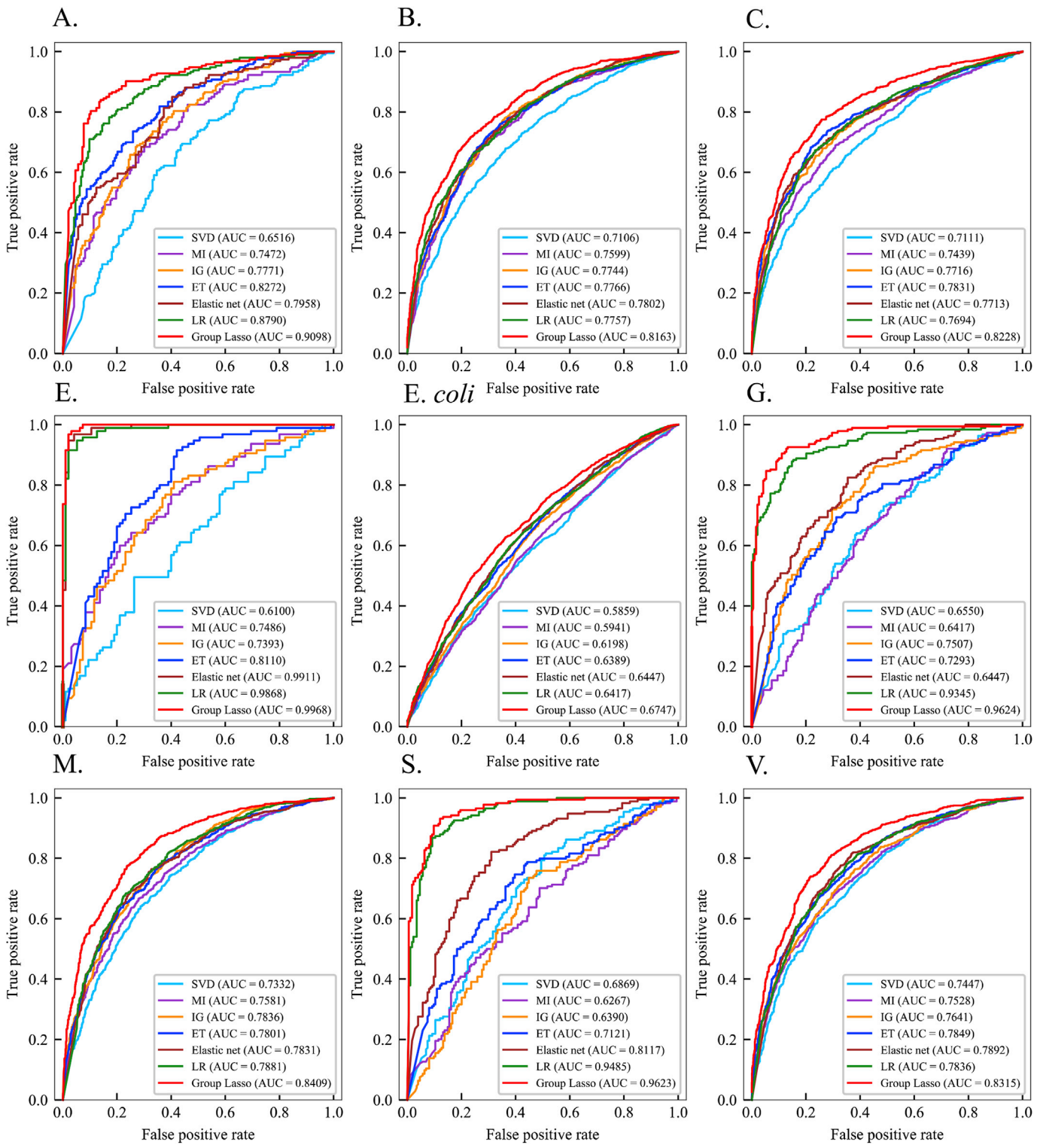

Fig. 5. ROC curves of different dimensional reduction methods on nine datasets.

datasets ACC values for different classification algorithms are shown in Supplementary Table S5. For the datasets A., C., E., G., M., S., V., the prediction accuracy of DNN reaches the highest, which are 84.47%, 75.38%, 96.89%, 89.15%, 76.62%, 90.51%, and 75.46%, respectively. The accuracy shows that for nine datasets, DNN can improve the robustness of the model and have higher prediction accuracy, compared with traditional machine learning methods and other deep learning approaches (CNN and LSTM). The comparison of ROC curves for different classification algorithms is shown in Fig. 6.

As can be seen from Fig. 6, the nine datasets AUC values corresponding to the DNN are 0.9098, 0.8163, 0.8228, 0.9968, 0.6747, 0.9624, 0.8409, 0.9623, and 0.8315, respectively. In particular, the ROC curves corresponding to the datasets A., E., G., and S. contain the ROC curves corresponding to the otherclassifiers, and the AUC values are 5.78%, 6.24%, 32.59%, and 16% higher than the values corresponding to the SVM, respectively. For the datasets B., C., M., V., the AUC values of

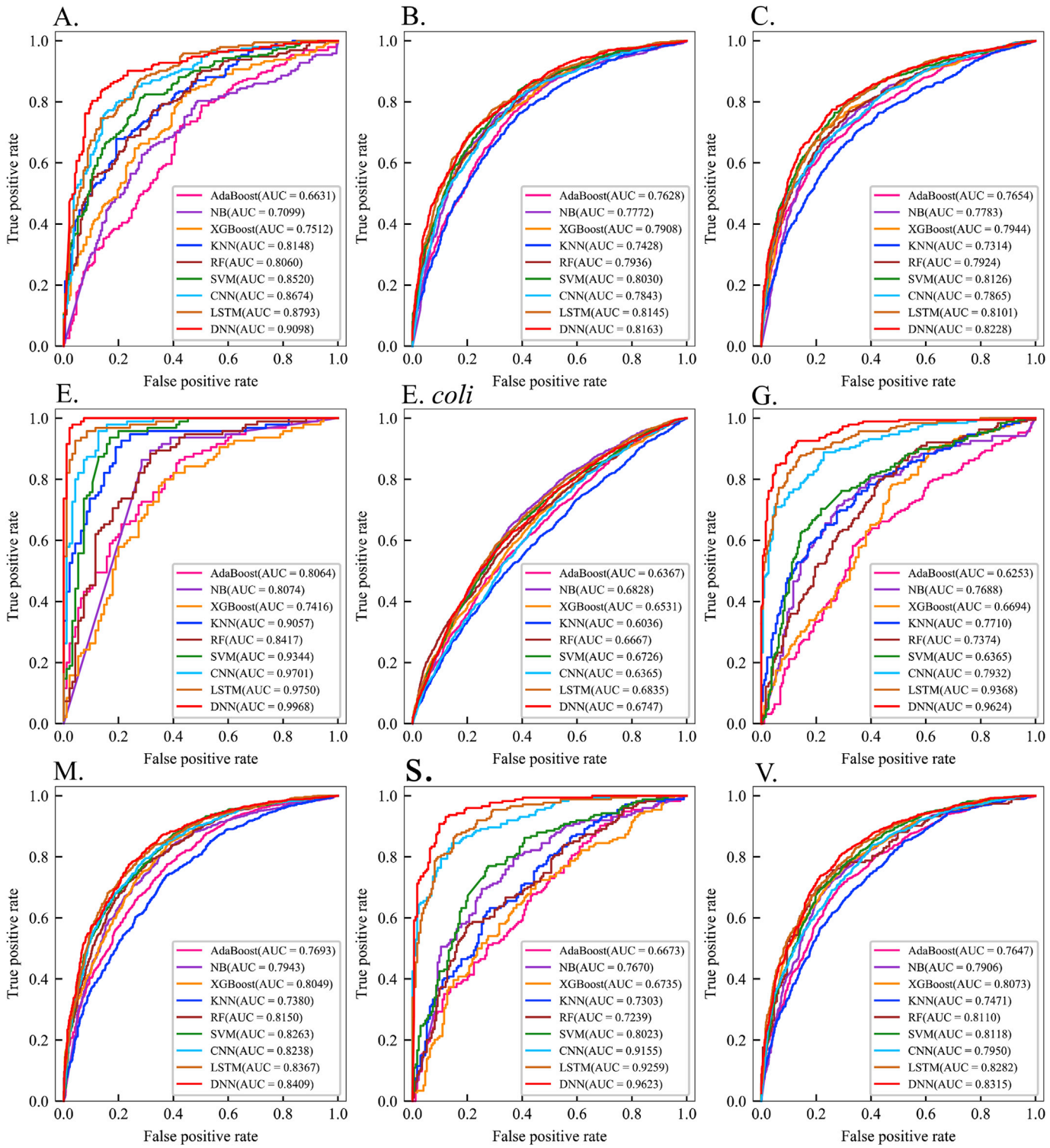

Fig. 6. ROC curves of different classifiers on nine datasets.

DNN are 3.2%, 3.63%, 1.71%, 3.65% higher than the values corresponding to the CNN, respectively. Although the AUC value of DNN is 0.88% lower than that of LSTM, DNN has more competitiveness with other datasets. Considering comprehensively, we choose DNN as the best classifier which can achieve better prediction results by providing deep layers of multi-layer networks and nonlinear mappings for revealing their internal correlations. DNNAce obtains more discriminant features from the original input data through the four hidden layers, demonstrating the validity and rationality of the deep structure. To visually observe the difference between acetylation sites and non-acetylation sites, we use t-SNE [84,85] to visualize the features learned from the data. The features of learning in the input layer and the last layer of the hidden layers of the DNN are visualized, and the distinction between the nine datasets test samples is plotted in 2D coordinates, as shown in Fig. 7.

From Fig. 7, we can visually see that the raw data of the input layer is very confusing, and the positive and negative datasets are mixed. DNNAce extracts abstract features from protein sequences through the hidden layers, learn the advanced features and sequence specificity of the

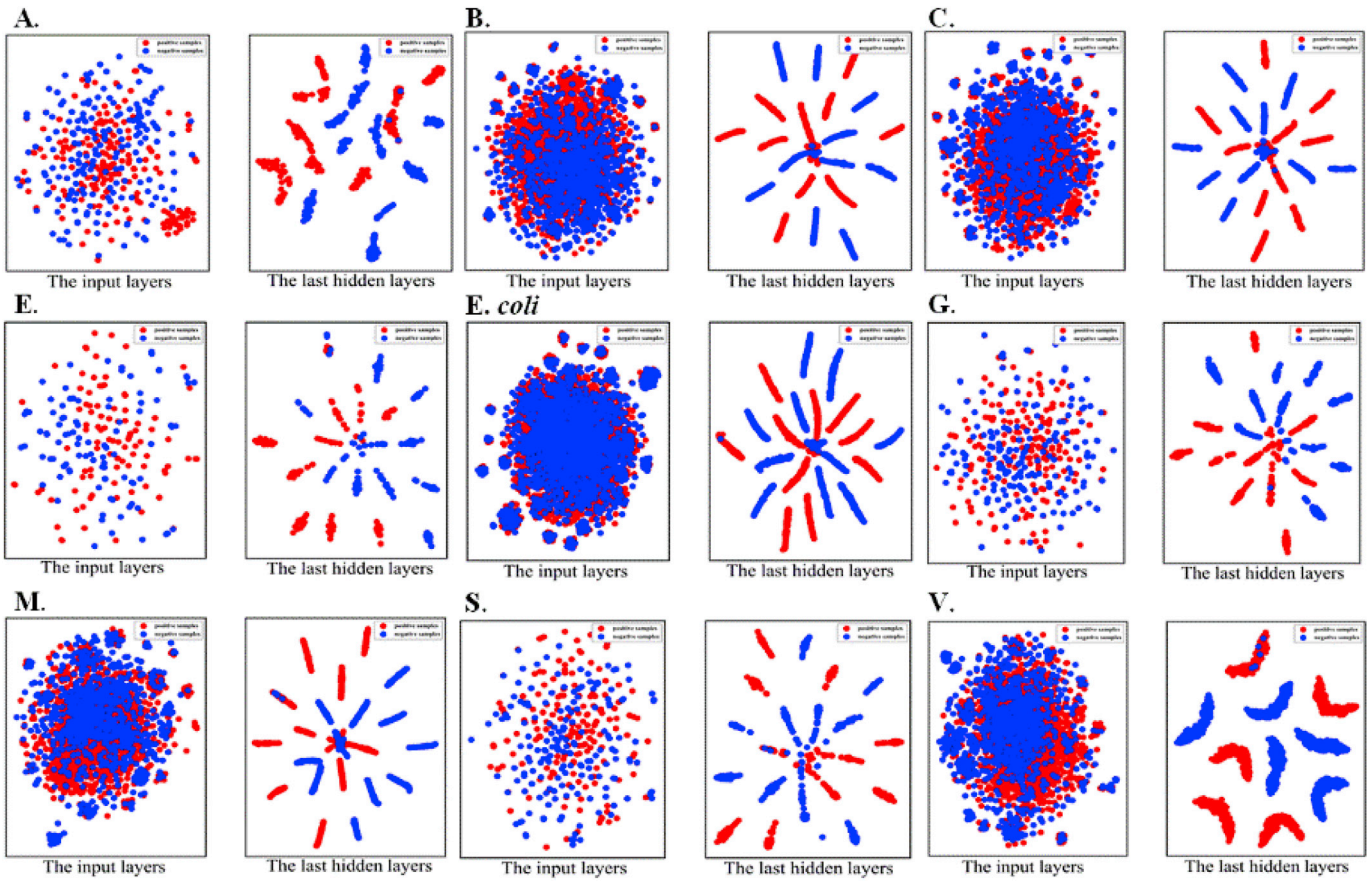

Fig. 7. t-SNE visualization of nine datasets features. Red represents the acetylation sites, and blue represents the non-acetylation sites. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

original protein input features. We also find that the high-level features of the last layer of the hidden layers become more and more clear, positive and negative classes samples can be separated. The above analysis shows that the DNN has the superiority of feature learning compared with the traditional machine learning methods.

3.9. Acetylation sites prediction performance of DNNAce

To evaluate the prediction performance of the DNNAce method, its results are compared with ProAcePred [33] on training datasets (Supplementary Table S6). To make a fair and objective comparison with other existing methods, we strictly implement the algorithm process on the independent test datasets, including the best parameters in feature extraction and optimal feature subset are consistent with the training datasets. In addition, we also compare the performance of the DNNAce with other acetylation sites prediction models. Considering the effectiveness of the acetylation sites prediction tool, LAcep [24], PLMLA [27] and ProAcePred [33] are selected to predict acetylation sites in the independent test datasets, the prediction results comparison of different methods on independent test datasets are shown in Fig. 8 and the detailed comparison results are shown in Supplementary Table S7.

From Fig. 8, we can intuitively see that, except for the datasets E and E. *coli*, for most independent test datasets, the prediction results of DNNAce are higher than LAcep, PLMLA and ProAcePred, and the ACC, Sn, and Sp reach above 80%, the MCC and AUC reach above 0.8. DNNAce's ACC values for nine independent test datasets are 90.00%, 98.26%, 92.88%, 90.00%, 86.18%, 97.50%, 96.44%, 95.00%, and 94.02% (Supplementary Table S7). The AUC value of DNNAce on the independent test datasets are greater than 0.9, which are 12%, 5.8%, 11.32%, 0.4%, 5.03%, 14.6%, 13.7%, 22.2% and 13.57% respectively than the values corresponding to the model ProAcePred (Supplementary Table S7). The independent test datasets results are better than other tools, which indicates that our model DNNAce can significantly improve the prediction performance of acetylation sites and has stronger competitiveness for acetylation sites prediction in different species.

4. Conclusion

Protein post-translational modification at specific sites affects almost every aspect of cell biology and pathogenesis, and various regulatory enzymes involved in PTM have become targets of many drugs. Hence, protein post-translational modification research has become a new focus in bioinformatics research. This paper proposes a new prediction method DNNAce based on the DNN to predict nine prokaryote datasets acetylation sites. The eight feature extraction methods are fused, considering the influence of sequence information, physicochemical information, and evolution information on the model performance. We use Group Lasso effectively retaining the important features of predicting acetylation sites. By analyzing different feature percentages and contribution rates in the optimal feature subset, it is found that the BE, AAindex, and BLOSUM62 have great influence on the performance of the prediction model. Then the optimal feature subset is input into DNN. Compared with other eight machine learning methods, the t-SNE visualization indicates that DNNAce can generate discriminative features through the network structure. At the same time, the DNNAce avoids other costs and associated limitations, pushing prokaryote acetylation research to a novel stage.

Although DNNAce improves the accuracy of acetylation sites prediction to a certain extent, there is still a room to improve in terms of prediction accuracy and algorithm efficiency. In the future, we will try to

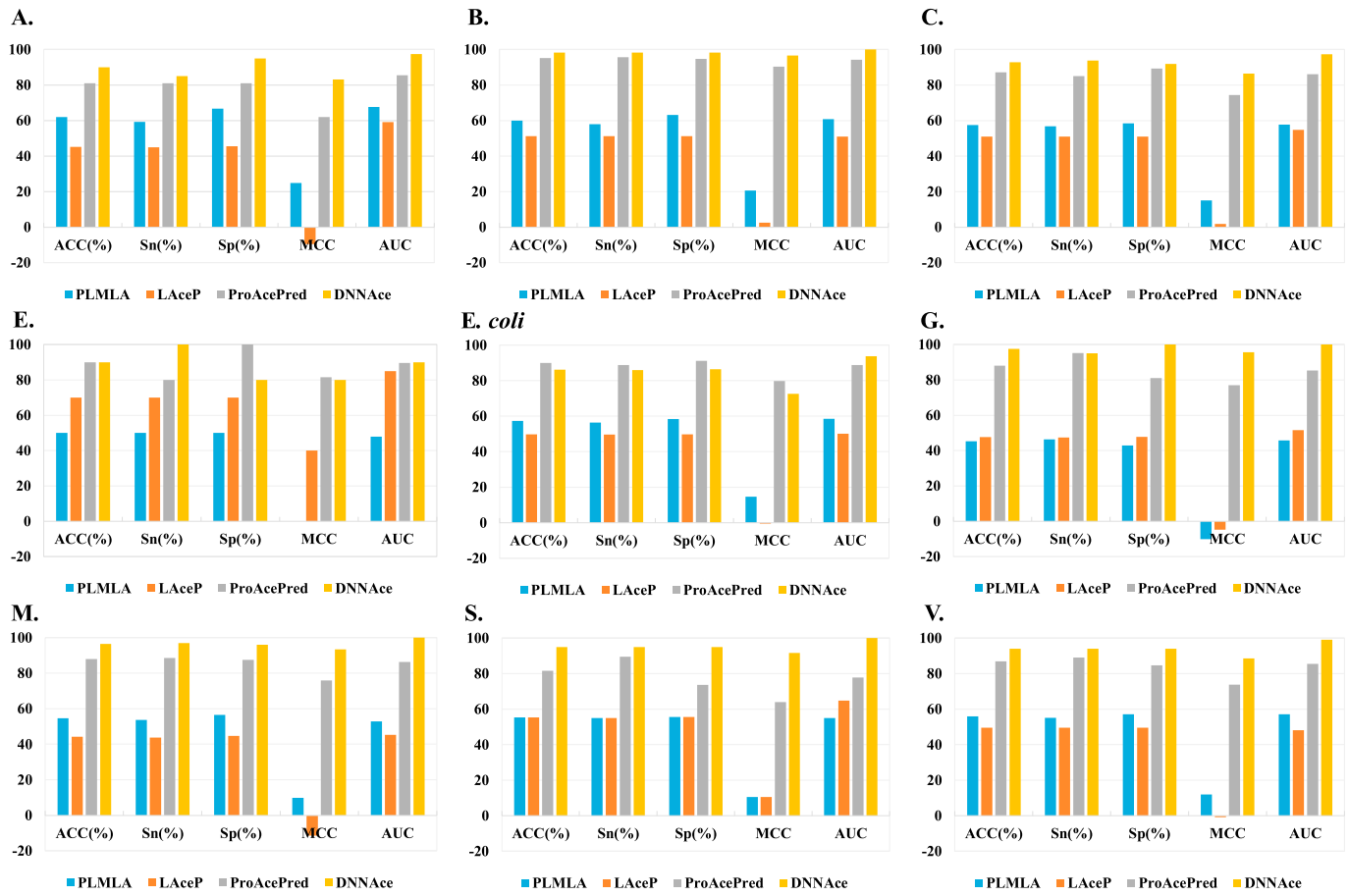

Fig. 8. Comparison of LAcep, PLMLA, ProAcePred and DNNAce on independent test datasets.

consider the structural information of proteins to generate more effective features, consider biological significance, and combine some effective algorithms, such as capsule networks and generative adversarial networks.

Declaration of competing interest

The authors declare no conflict of interest.

CRediT authorship contribution statement

Bin Yu: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft, Validation, Writing - review & editing. **Zhaomin Yu:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft, Validation, Visualization. **Cheng Chen:** Formal analysis, Investigation, Methodology, Validation, Visualization. **Anjun Ma:** Formal analysis, Investigation, Methodology, Validation, Visualization. **Bingqiang Liu:** Formal analysis, Investigation, Methodology, Writing - original draft. **Baoguang Tian:** Formal analysis, Investigation, Writing - original draft. **Qin Ma:** Formal analysis, Writing - original draft, Writing - review & editing.

Acknowledgments

The authors sincerely thank the anonymous reviewers for their valuable comments. This work was supported by the National Natural Science Foundation of China (No. 61863010), the Key Research and Development Program of Shandong Province of China (No. 2019GGX101001), and the Natural Science Foundation of Shandong Province of China (Nos. ZR2017MA014 and ZR2018MC007). This work used the Extreme Science and Engineering Discovery Environment, which is supported by the National Science Foundation (No. ACI-1548562).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2020.103999.

References

[1] G. Grotenbreg, H. Ploegh, Chemical biology: dressed-up proteins, Nature 446 (2007) 993–995.
[2] M. Mann, O.N. Jensen, Proteomic analysis of post-translational modifications, Nat. Biotechnol. 21 (2003) 255–261.
[3] 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.
[4] H. Qiu, Y. Guo, L. Yu, X. Pu, M. Li, Predicting protein lysine methylation sites by incorporating single-residue structural features into Chou's pseudo components, Chemometr. Intell. Lab. Syst. 179 (2018) 31–38.
[5] Y. Shi, Y. Guo, Y. Hu, M. Li, Position-specific prediction of methylation sites from sequence conservation based on information theory, Sci. Rep. 5 (2015) 12403.
[6] Y. Xu, X. Wen, L.S. Wen, L.Y. Wu, N.Y. Deng, K.C. Chou, iNitro-Tyr, Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition, PloS One 9 (2014), e105018.
[7] D. Wang, S. Zeng, C. Xu, W. Qiu, Y. Liang, T. Joshi, D. Xu, MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction, Bioinformatics 33 (2017) 3909–3916.
[8] F. Luo, M. Wang, Y. Liu, X.M. Zhao, A. Li, DeepPhos: prediction of protein phosphorylation sites with deep learning, Bioinformatics 35 (2019) 2766–2773.
[9] Q.Y. Chen, J. Tang, P.F. Du, Predicting protein lysine phosphoglycerylation sites by hybridizing many sequence based features, Mol. Biosyst. 13 (2017) 874–882.
[10] W. Hussain, Y.D. Khan, N. Rasool, S.A. Khan, K.C. Chou, SPrenylC-PseAAC: A sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-prenylation sites in proteins, J. Theor. Biol. 468 (2019) 1–11.

[11] J.R. Wang, W.L. Huang, M.J. Tsai, K.T. Hsu, H.L. Huang, S.Y. Ho, ESA-Ubisite: accurate prediction of human ubiquitination sites by identifying a set of effective negatives, Bioinformatics 33 (2017) 661–668.
[12] W.R. Qiu, B.Q. Sun, H. Tang, J. Huang, H. Lin, Identify and analysis crotonylation sites in histone by using support vector machines, Artif. Intell. Med. 83 (2017) 75–81.
[13] M. Wu, Y. Yang, H. Wang, Y. Xu, A deep learning method to more accurately recall known lysine acetylation sites, BMC Bioinf. 20 (2019) 49.
[14] S. Zhao, W. Xu, W. Jiang, W. Yu, Y. Lin, T. Zhang, J. Yao, L. Zhou, Y. Zeng, H. Li, Y. Li, J. Shi, W. An, S.M. Hancock, F. He, L. Qin, J. Chin, P. Yang, X. Chen, Q. Lei, Y. Xiong, K.L. Guan, Regulation of cellular metabolism by protein lysine acetylation, Science 327 (2010) 1000–1004.
[15] V.V. Ogryzko, R.L. Schiltz, V. Russanova, B.H. Howard, Y. Nakatani, The transcriptional coactivators p300 and cbp are histone acetyltransferases, Cell 87 (1996) 953–959.
[16] R. Behnia, B. Panic, J.R.C. Whyte, S. Munro, Targeting of the Arf-like GTPase Arl3p to the Golgi requires N-terminal acetylation and the membrane protein Sys1p, Nat. Cell Biol. 6 (2004) 405–413.
[17] P.V. Damme, T. Arnesen, K. Gevaert, Protein alpha-N-acetylationstudied by N-terminomics, FEBS J. 278 (2011) 3822–3834.
[18] J.E. Bradner, N. West, M.L. Grachan, E.F. Greenberg, S.J. Haggarty, T. Warnow, R. Mazitschek, Chemical phylogenetics of histone deacetylases, Nat. Chem. Biol. 6 (2010) 238–243.
[19] Q. Wang, Y. Zhang, C. Yang, H. Xiong, Y. Lin, J. Yao, H. Li, L. Xie, W. Zhao, Y. Yao, Z.B. Ning, R. Zeng, Y. Xiong, K.L. Guan, S. Zhao, G.P. Zhao, Acetylation of metabolic enzymes coordinates carbon source utilization and metabolic flux, Science 327 (2010) 1004–1007.
[20] V.M. Richon, S. Emiliani, E. Verdin, Y. Webb, R. Breslow, R.A. Rifkind, P.A. Marks, A class of hybrid polar inducers of transformed cell differentiation inhibits histone deacetylases, Proc. Natl. Acad. Sci. U.S.A. 95 (1998) 3003–3007.
[21] D.J. Welsch, G.L. Nelsestuen, Amino-terminal alanine functions in a calcium-specific process essential for membrane binding by prothrombin fragment 1, Biochemistry 27 (1988) 4939–4945.
[22] D. Umlauf, Y. Goto, R. Feil, Site-specific analysis of histone methylation and acetylation, Methods Mol. Biol. 287 (2004) 99–120.
[23] H. Zhou, J.A. Ranish, J.D. Watts, R. Aebersold, Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry, Nat. Biotechnol. 19 (2002) 512–515.
[24] T. Hou, G. Zheng, P. Zhang, J. Jia, J. Li, L. Xie, C. Wei, Y. Li, LAceP: lysine acetylation site prediction using logistic regression vlassifiers, PloS One 9 (2014), e89575.
[25] Y. Li, M. Wang, H. Wang, H. Tan, Z. Zhang, G.I. Webb, J. Song, Accurate in silico identification of species-specific acetylation sites by integrating protein sequence-derived and functional features, Sci. Rep. 4 (2014) 5765.
[26] F. Gnad, S. Ren, C. Choudhary, J. Cox, M. Mann, Predicting post-translational lysine acetylation using support vector machines, Bioinformatics 26 (2010) 1666–1668.
[27] S. Shi, J. Qiu, X. Sun, S. Suo, S. Huang, R. Liang, Plmla: prediction of lysine methylation and lysine acetylation by combining multiple features, Mol. Biosyst. 8 (2012) 1520–1527.
[28] Q. Wuyun, W. Zheng, Y. Zhang, J. Ruan, G. Hu, Improved species-specific lysine acetylation site prediction based on a large variety of features set, PloS One 11 (2016), e0155370.
[29] T.Y. Lee, J.B.K. Hsu, F.M. Lin, W.C. Chang, P.C. Hsu, H.D. Huang, N-Ace: using solvent accessibility and physicochemical properties to identify protein N-acetylation sites, J. Comput. Chem. 31 (2010) 2759–2771.
[30] W. Bao, B. Yang, Z. Li, Y. Zhou, LAIPT: lysine acetylation site identification with polynomial tree, Int. J. Mol. Sci. 20 (2019) 113.
[31] Y. Xu, X. Wang, J. Ding, L. Wu, N. Deng, Lysine acetylation sites prediction using an ensemble of support vector machine classifiers, J. Theor. Biol. 264 (2010) 130–135.
[32] B. Wang, M. Wang, A. Li, Prediction of post-translational modification sites using multiple kernel support vector machine, Peer J 5 (2017), e3261.
[33] G. Chen, M. Cao, K. Luo, L. Wang, P. Wen, S. Shi, ProAcePred: prokaryote lysine acetylation sites prediction based on elastic net feature optimization, Bioinformatics 34 (2018) 3999–4006.
[34] Z. Liu, Y. Wang, T. Gao, Z. Pan, H. Cheng, Q. Yang, Z. Cheng, A. Guo, J. Ren, Y. Xue, CPLM: a database of protein lysine modifications, Nucleic Acids Res. 42 (2014) D531–D536.
[35] S. Kosono, M. Tamura, S. Suzuki, Y. Kawamura, A. Yoshida, M. Nishiyama, M. Yoshida, Changes in the acetylome and succinylome of Bacillus subtilis in response to carbon source, PloS One 10 (2015), e0131169.
[36] D.W. Lee, D. Kim, Y.J. Lee, J.A. Kim, J.Y. Choi, S. Kang, J.G. Pan, Proteomic analysis of acetylation in thermophilic *Geobacillus kaustophilus*, Proteomics 13 (2013) 2278–2282.
[37] Y. Mizuno, M. Nagano-Shoji, S. Kubo, Y. Kawamura, A. Yoshida, H. Kawasaki, M. Nishiyama, M. Yoshida, S. Kosono, Altered acetylation and succinylation profiles in *Corynebacterium glutamicum* in response to conditions inducing glutamate overproduction, Microbiologyopen 5 (2016) 152–173.
[38] H. Okanishi, K. Kim, R. Masui, S. Kuramitsu, Acetylome with structural mapping reveals the significance of lysine acetylation in *Thermus thermophiles*, J. Proteome Res. 12 (2013) 3952–3968.
[39] J. Pan, Z. Ye, Z. Cheng, X. Peng, L. Wen, F. Zhao, Systematic analysis of the lysine acetylome in *Vibrio parahemolyticus*, J. Proteome Res. 13 (2014) 3294–3302.
[40] X. Wu, A. Vellaichamy, D. Wang, L. Zamdborg, N.L. Kelleher, S.C. Huber, Y. Zhao, Differential lysine acetylation profiles of Erwinia amylovora strains revealed by proteomics, J. Proteomics 79 (2013) 60–71.
[41] L. Xie, X. Wang, J. Zeng, M. Zhou, X. Duan, Q. Li, Z. Zhang, H. Luo, L. Pang, W. Li, G. Liao, X. Yu, Y. Li, H. Huang, J. Xie, Proteome-wide lysine acetylation profiling of the human pathogen Mycobacterium tuberculosis, Int. J. Biochem. Cell Biol. 59 (2015) 193–202.
[42] W. Li, A. Godzik, CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (2006) 1658–1659.
[43] Z. Ju, J.J. He, Prediction of lysine propionylation sites using biased SVM and incorporating four different sequence features into Chou's PseAAC, J. Mol. Graph. Model. 76 (2017) 356–363.
[44] B. Yu, S. Li, C. Chen, J. Xu, W. Qiu, X. Wu, R. Chen, Prediction subcellular localization of Gram-negative bacterial proteins by support vector machine using wavelet denoising and Chou's pseudo amino acid composition, Chemometr. Intell. Lab. Syst. 167 (2017) 102–112.
[45] B. Tian, X. Wu, C. Chen, W. Qiu, Q. Ma, B. Yu, Predicting protein-protein interactions by fusing various Chou's pseudo components and using wavelet denoising approach, J. Theor. Biol. 462 (2019) 329–346.
[46] C. Chen, Q. 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.
[47] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins 43 (2001) 246–255.
[48] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, AAindex: amino acid index database, progress report 2008, Nucleic Acids Res. 36 (2008) D202–D205.
[49] M.M. Hasan, D. Guo, H. Kurata, Computational identification of protein S-sulfenylation sites by incorporating the multiple sequence features information, Mol. Biosyst. 13 (2017) 2545–2550.
[50] Z. Chen, P. Zhao, F. Li, A. Leier, T.T. Marquez-Lago, Y. Wang, G.I. Webb, A.I. Smith, R.J. Daly, K.C. Chou, J. Song, iFeature: a python package and web server for features extraction and selection from protein and peptide sequences, Bioinformatics 34 (2018) 2499–2502.
[51] Z.H. Zhang, Z.H. Wang, Z.R. Zhang, Y.X. Wang, A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine, FEBS Lett. 580 (2006) 6169–6174.
[52] Y. Ding, J. Tang, F. Guo, Identification of drug-target interactions via multiple information integration, Inf. Sci. 418–419 (2017) 546–560.
[53] J. Gao, J.J. Thelen, A.K. Dunker, D. Xu, Musite, a tool for global prediction of general and kinase-specific phosphorylation sites, Mol. Cell. Proteomics 9 (2010) 2586–2600.
[54] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. Roy. Stat. Soc. B. 58 (1996) 267–288.
[55] M. Blondel, K. Seki, K. Uehara, Block coordinate descent algorithms for large-scale sparse multiclass classication, Mach. Learn. 93 (2013) 31–52.
[56] V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, in: Proceedings of the 27th International Conference on Machine Learning, 2010, pp. 807–814.
[57] D. Kingma, J. Ba, Adam: a Method for Stochastic Optimization, 2014 arXiv: 1412.6980.
[58] N. Srivastava, G. 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.
[59] X. Sun, T. Jin, C. Chen, X. Cui, Q. Ma, B. Yu, RBPro-RF: Use Chou's 5-steps rule to predict RNA-binding proteins via random forest with elastic net, Chemometr. Intell. Lab. Syst. 197 (2020) 103919.
[60] X. Wang, B. Yu, A. Ma, C. Chen, B. Liu, Q. Ma, Protein-protein interaction sites prediction by ensemble random forests with synthetic minority oversampling technique, Bioinformatics 35 (2019) 2395–2402.
[61] B. Yu, W. Qiu, C. Chen, A. Ma, J. Jiang, H. Zhou, Q. Ma, SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting, Bioinformatics 36 (2020) 1074–1081.
[62] 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.
[63] H. Zhou, C. Chen, M. Wang, Q. Ma, B. Yu, Predicting Golgi-resident protein types using conditional covariance minimization with XGBoost based on multiple features fusion, IEEE Access 7 (2019) 144154–144164.
[64] B. Yu, L. Lou, S. Li, Y. Zhang, W. Qiu, X. Wu, M. Wang, B. Tian, Prediction of protein structural class for low-similarity sequences using Chou's pseudo amino acid composition and wavelet denoising, J. Mol. Graph. Model. 76 (2017) 260–273.
[65] B. Yu, S. Li, W. Qiu, M. Wang, J. Du, Y. 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.
[66] W. Qiu, S. Li, X. Cui, Z. Yu, M. Wang, J. Du, Y. Peng, B. Yu, Predicting protein submitochondrial locations by incorporating the pseudo-position specific scoring matrix into the general Chou's pseudo-amino acid composition, J. Theor. Biol. 450 (2018) 86–103.
[67] B. Yu, S. Li, W. Qiu, C. Chen, R. Chen, L. Wang, M. Wang, Y. Zhang, Accurate prediction of subcellular location of apoptosis proteins combining Chou's PseAAC and PsePSSM based on wavelet denoising, Oncotarget 8 (2017) 107640–107665.
[68] J. Lin, H. Chen, S. Li, Y. Liu, X. Li, B. Yu, Accurate prediction of potential druggable proteins based on genetic algorithm and Bagging-SVM ensemble classifier, Artif. Intell. Med. 98 (2019) 35–47.
[69] 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.

[70] X. Cui, Z. Yu, B. Yu, M. Wang, B. 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.
[71] M.E. Wall, A. Rechtsteiner, L.M. Rocha, Singular value decomposition and principal component analysis, in: A Practical Approach to Microarray Data Analysis, 2002, pp. 91–109.
[72] A. Kraskov, H. Stögbauer, P. Grassberger, Estimating mutual information, Phys. Rev. E - Stat. Nonlinear Soft Matter Phys. 69 (2004), 066138.
[73] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (2006) 3–42.
[74] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. Roy. Stat. Soc. B 67 (2005) 301–320.
[75] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, C.J. Lin, LIBLINEAR: a library for large linear classification, J. Mach. Learn. Res. 9 (2008) 1871–1874.
[76] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to Boosting, J. Comput. Syst. Sci. 55 (1997) 119–139.
[77] N. Friedman, G. Dan, M. Goldszmidt, Bayesian network classifiers, Mach. Learn. 29 (1997) 131–163.
[78] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
[79] F. Nigsch, A. Bender, B.V. Buuren, J. Tissen, E. Nigsch, J.B.O. Mitchell, Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization, J. Chem. Inf. Model. 46 (2006) 2412–2422.
[80] L. Breiman, Random forest, Mach. Learn. 45 (2001) 5–32.
[81] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297.
[82] C. Angermueller, T. Pärnamaa, L. Parts, O. Stegle, Deep learning for computational biology, Mol. Syst. Biol. 12 (2016) 878.
[83] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997) 1735–1780.
[84] L. Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.
[85] E. Becht, L. McInnes, J. Healy, C.A. Dutertre, I.W.H. Kwok, L.G. Ng, F. Ginhoux, E.W. Newell, Dimensionality reduction for visualizing single-cell data using UMAP, Nat. Biotechnol. 37 (2019) 38–44.