IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON FEATURE REPRESENTATION AND LEARNING METHODS WITH APPLICATIONS IN LARGE-SCALE BIOLOGICAL SEQUENCE ANALYSIS

Received April 21, 2020, accepted April 27, 2020, date of publication May 6, 2020, date of current version May 22, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2992468

The Classification of Enzymes by Deep Learning

ZHIYU TAO, BENZHI DONG, ZHIXIA TENG, AND YUMING ZHAO
Information and Computer Engineering College, Northeast Forestry University, Harbin 150040, China

Corresponding author: Yuming Zhao (zym@nefu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61971117 and Grant 61901103, and in part by the Natural Science Foundation of Heilongjiang Province under Grant LH2019F002.

ABSTRACT Enzymes, as a group of crucial biocatalysts produced by living cells, enable the chemical reactions in organisms to be more efficient. According to the properties of the reactions catalyzed by enzymes, the Enzyme Commission (EC) number system divided enzymes into 6 primary main classes in 1961: oxidoreductases (EC1), transferases (EC2), hydrolases (EC3), lyases (EC4), isomerases (EC5), and ligases (EC6). These six categories did not change for many years until a new class, the translocases (EC7), was added in August 2018. Different enzymes have different properties of catalytic reaction, and the prediction of enzyme classes is a very important research topic, allowing us to further study the structure and function of enzyme molecules when we know the category of enzyme. Because the number of enzymes whose function remains unknown is enormous, it is time-consuming to use biological experiments to determine enzyme characteristics. Thus, devising various computational models to predict enzyme classes has become a feasible scheme. In hope of giving researchers more inspiration and ideas for predicting the EC number of enzymes by machine learning, we summarize a variety of research methods used in the prediction of enzyme families in this research.

INDEX TERMS Commission, enzyme classification, machine learning, bioinformatics.

I. INTRODUCTION

As a type of very important biocatalyst, enzymes have a vital role in maintaining the life activities of organisms. They dominate the metabolism, nutrition, energy conversion and many other chemical reactions closely related to the life process. Most enzymes are proteins, while others are ribonucleic acid. To date, research on the prediction of enzyme classes and subclasses has always focused on the enzyme whose chemical essence is protein; that is, when we use computational models to classify enzymes, the feature extraction method we adopt is always for proteins. To facilitate the further study of enzymes, the International Union of Biochemistry (IUB) established an International Commission called the Enzyme Commission in charge of developing a nomenclature for enzymes. The Commission has classified enzymes into 7 main classes. The nomenclature of the EC is composed of four parts of figures that identify the main class, subclass, sub-subclass and substrate class of the enzyme. Most classifiers designed by scholars can classify enzymes to the level of subclass [1]–[3]. To make it easier to understand the classes of enzymes, they are visualized in Table 1. Since the category of translocases had not been proposed for many years, most prediction methods divide the enzymes into the other six categories [4], [5]. It is worth mentioning that in recent years, classifiers have appeared that can classify enzymes to the level of substrate class.

In the field of biology, wet laboratory-based functional identification procedures were adopted to determine the function and category of enzymes, but this kind of experiment is costly and time consuming [6]–[12]. Therefore, classifying enzymes using bioinformatic tools is suitable. With the development of bioinformatics and deep learning [13]–[28], scholars have designed many models for the prediction of enzyme classes [29]. In 2009, Nasibov *et al.* [30] adopted the method of K-nearest neighbor (KNN) classification. In 2010, Qiu *et al.* [31] used support vector machine (SVM), obtaining good results. In 2010, Concu *et al.* [32] used linear discriminant analysis (LDA) and artificial neural networks (ANN) and compared the final classification results. In addition, in order to achieve better prediction results, scholars usually combine various feature extraction Methods and classification methods in their prediction process. For instance, Shen *et al.* [33] combined functional domain (FunD) and pseudo position-specific scoring matrix (PsePSSM) to extract features in 2009. Wang *et al.* [34] combined composition,

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

TABLE 1. The main classes and subclasses of enzymes

Main Classes	Subclasses
Oxidoreductases(EC1)	Acting on the CH-OH group, Acting on the CH-NH group, Acting on the CH-CH group ...
Transferases(EC2)	Transferring one-carbon, Transferring nitrogenous, Transferring sulfur-containing, Glycosyltransferases ...
Hydrolases(EC3)	Acting on piptide bonds, Acting on acid anhydrides, Acting on ester bonds, Acting on nitrogen bonds ...
Lyases(EC4)	Phosphorus-oxygen lyases, Carbon-nitrogen lyases, Carbon-carbon lyases, Carbon-sulfur lyases ...
Isomerases(EC5)	Intramolecular lyases, Intramolecular transferases, Intramolecular oxidoreductases, Cis-trans-isomerases ...
Ligases(EC6)	Forming carbon-carbon bonds, Forming carbon-oxygen bonds, Forming nitrogen-metal bonds ...
Translocases(EC7)	Translocation of proton, Translocation of amino acids, Translocation of cation, Translocation of anion ...

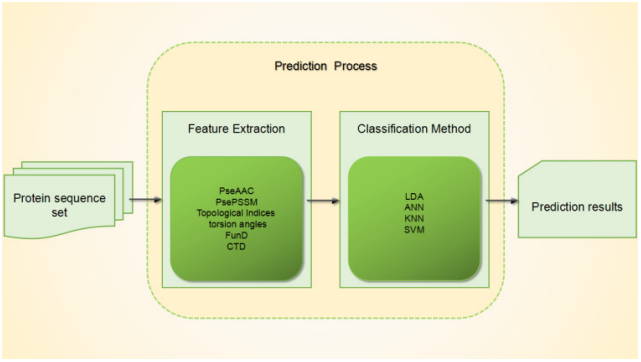

FIGURE 1. The prediction process of enzymes.

transition and distribution (CTD) and pseudo-amino acid composition (PseAAC) to extract features and classify sequences with the combination of the methods of random-k-label-random forest (RAkEL-RF) and multi-label KNN (MLKNN) in 2014. In 2019, Ryu *et al.* [35] used DeepEC, consisting of three different convolutional neural network (CNN) structures in enzyme classification. Generally, the prediction process can be roughly described as two steps: extracting features from sequences first, and then classifying the feature set by a classification model. The specific prediction process is shown in Fig 1. The methods of feature extraction and classification shown in Fig 1 are often adopted by scholars. To facilitate research for scholars in this field, we summarize some recent machine learning methods utilized in predicting enzyme classes that are novel and classic. Our summary consists of four parts. The first part briefly introduces the protein sequence set. In the second and third parts, we introduce some methods adopted in the latest papers or used frequently for feature extraction and classification. Finally, we present a statistical analysis and comment on recent published results.

II. PROTEIN SEQUENCE SET

There are some databases that contain large numbers of protein sequences, such as ENZYME (http://enzyme.expasy.org/), UniProt (http://www.uniprot.org/) and PDB (http://www.rcsb.org/). Scholars obtain sequences from these databases and use them to train and test their prediction models. To make the prediction process more rational, processing of the data sets is required [3], [12], such as deleting some sequences with high similarity [36]. Some specific tools or algorithms can be used to implement this process, such as PISCES or CD-HIT. Scholars use these processed data sets to help build prediction models [37], [38]. We summarize the source and composition of the data in Table 2 adopted by scholars in the past few years.

III. FEATURE EXTRACTION

Before classification, we need to extract the features of the protein sequences in the data set [39], [40]. Formulating the sequences using rational mathematical expressions obtains the quantification of various kinds of protein characteristics and is conducive to the classification in the next step.

A. PSEUDO AMINO ACID COMPOSITION (PseAAC)

This method evolved from the method of AAC, which counts the proportion and the rate of each amino acid, and generates a 20-dimension vector. Since PseAAC was proposed by Chou in 2001, it has been widely adopted in bioinformatics [41]–[46].

By PseAAC, a sequence P can be formulated as a vector as follows:

$$P = [x_1, x_2, \ldots, x_{20}, x_{20+1}, \ldots, x_{20+\lambda}]^T, \quad \lambda < L \tag{1}$$

In (1), a sequence is represented by a vector P. Among them, the first 20 elements represent the composition of 20 amino acids in a sequence and the latter λ elements represent the sequence-order information. Each element in vector P is formulated as follows:

$$x_\theta \begin{cases} \dfrac{f_\theta}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{\lambda} \delta_j}, & (1 \le \theta \le 20) \\ \dfrac{\omega \delta_{\theta-20}}{\sum_{i=1}^{20} f_i + \omega \sum_{j=1}^{\lambda} \delta_j}, & (20+1 \le \theta \le 20+\lambda) \end{cases} \tag{2}$$

TABLE 2. The composition of data sets

database	thesis	Number of each parts of the dataset collected								
		EC1	EC2	EC3	EC4	EC5	EC6	EC7	multi-label	Non-enzyme
ENZYME	[33]	1618	3450	2791	679	518	776	0	0	9850
	[34]	1411	1935	2703	1626	602	177	0	0	0
	[47]	800	1931	1351	655	166	139	0	1793	0
PDB	[32]	3295	5278	5041	1322	934	746	10327	0	55413
	[48]	11194	18733	23861	4476	2762	2532	0	0	0
	[49]	7256	10665	15451	2694	1642	1543	0	783	0
	[50]	16669	1893	1757	3102	7968	7968	0	0	0
	[51]	13187	31033	24794	5983	3833	4072	0	0	0
UniProt	[52]	36577	86163	59551	22368	13615	29233	0	0	42382
	[53]	9167	24219	12683	5804	4174	5956	0	0	0
ALL	[54]	3343	8517	5917	1532	1193	1666	0	1085	0

*The 'All' in the column of 'database' means that the data comes from all the databases mentioned in the table. The 'multi-label' means that the enzymes belongs to more than one class.

where f_i is the content of amino acid i in protein P and δ_j is the i-tier correlation factor calculated by (3). The ω and λ are user-defined parameters. The λ reflects the maximal distance between one contiguous residue and the other.

$$\delta_j = \frac{1}{L-j}\sum_{i=1}^{L-j}\Theta(R_i, R_{i+j}) \tag{3}$$

where correlation function $\Theta(R_i, R_j)$ can be defined as follows:

$$\Theta(R_i, R_{i+j}) = \frac{1}{3}\Big\{\big[F_1(R_j) - F_1(R_i)\big]^2 + \big[F_2(R_j) - F_2(R_i)\big]^2 + \big[F_3(R_j) - F_3(R_i)\big]^2\Big\} \tag{4}$$

where $F_1(R_i)$, $F_2(R_i)$, and $F_3(R_i)$ are the values of quantified physicochemical properties such as hydrophobicity, hydrophilicity and side chain mass. These three values after standardization are described as follows:

$$F_k(i) = \frac{F_k^0(i) - \sum_{i=1}^{20}\frac{F_k^0(i)}{20}}{\sqrt{\frac{\sum_{i=1}^{20}\left[F_k^0(i) - \frac{F_k^0(i)}{20}\right]^2}{20}}} \quad (k = 1, 2, 3) \tag{5}$$

where $F_1^0(R_i)$, $F_2^0(R_i)$, and $F_3^0(R_i)$ are the original values of quantified physicochemical properties of the residue that can be obtained from other academic theses.

For model PseAAC, the setting of parameters ω and λ is critical in that different values of these two parameters often lead to different accuracy of final classification in the case of using the same classification algorithm. In addition, we can add more quantifiable indexes about protein character to correlation function $\Theta(R_i, R_j)$, such as $F_4^0(R_i)$, $F_5^0(R_i)$ and so on, to obtain a vector helpful for classification.

B. PSEUDO POSITION SPECIFIC SCORING MATRIX (PsePSSM)

This method depends on the Position Specific Scoring Matrix (PSSM), which represents the changes in amino acids at specific positions in a protein during the long process of evolution. A protein P can be described by PSSM as follows:

$$P_{PSSM} = \begin{bmatrix} a_{1A} & a_{1C} & \cdots & a_{1x} & \cdots & a_{1Y} \\ a_{2A} & a_{2C} & \cdots & a_{2x} & \cdots & a_{2Y} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ a_{nA} & a_{nC} & \cdots & a_{nx} & \cdots & a_{nY} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ a_{NA} & a_{NC} & \cdots & a_{Nx} & \cdots & a_{NY} \end{bmatrix} \tag{6}$$

For example, the element a_{nx} represents the standardized and quantified results by estimating the degree that the nth amino acid in the peptide segments converted into amino acid x and n is the number of amino acids and the rows of the matrix P_{pssm} represent 20 amino acids in nature. This matrix is obtained by searching a comprehensive sequence database by PSI-BLAST [55], [56].

To let the matrix of PSSM become a vector, we represent a column by the average value of this column of the matrix, as described in the following formula:

$$[\overline{a_A}, \overline{a_C}, \ldots, \overline{a_Y}]^T \tag{7}$$

where

$$\overline{a_x} = \frac{1}{N}\sum_{n=1}^{N} a_{nx} \quad (x = A, C, \ldots, Y) \tag{8}$$

However, Eq. (7) cannot represent the sequence-order information. To express the amino acid order information in a sequence, the vector PsePSSM is proposed and given by:

$$\left[a_A^0, a_C^0, \ldots, a_Y^0, a_A^1, a_C^1, \ldots, a_Y^1 \ldots, a_A^k, a_C^k, \ldots, a_Y^k\right] \tag{9}$$

where

$$a_n^k = \frac{1}{N-k}\sum_{n=1}^{N-k}\left[a_{nx} - a_{(n+k)x}\right]^2 \quad (x = A, C, \ldots, Y; k < N) \tag{10}$$

where k is the parameter that needs to be set. It is noteworthy that the value of parameter N is the minimum number of amino acids in the sequences and the value of k must be smaller than N. When the condition $k = 0$ appears, (9) degenerates into (7).

C. TOPOLOGICAL INDICES

This method has been recently used to predict enzyme function. The protein sequence is represented in a star graph (SG) designed by Milan *et al.* and it is widely used in the field of bioinformation. The primary structure of the protein is expressed by a star graph. We use the distance matrix and the degree matrix to describe a star graph.

We calculate particular topological indices from a star graph to characterize a protein sequence. The software S2SNet can accomplish this process. The results file from S2SNet contains a series of indices as shown below:

Trace of the n connectivity matrices (*Trn*):

$$Tr_n = \sum\nolimits_i (Mn)ii \tag{11}$$

where M is the graph connectivity matrix and n is the power of matrix M and ii represents the ith diagonal element.

Schultz topological index(s):

$$S = \sum\nolimits_{i<j} (\deg_i + \deg_j) \times d_{ij} \tag{12}$$

where d_{ij} is the element of the distance matrix and deg_i is the elements of the degree matrix.

Besides the two simple indexes mentioned above, there is an extra set of indicators: the features of a sequence and the different feature vectors composed of different indices selected and given different weights resulting in different prediction results.

D. TORSION ANGLES

Information for the amide plane can be used as structural features. Only the single bond formed by carbon atom α can rotate in the peptide chain, so it is the root cause of peptide chain curling and folding. We use two angles, φ and ψ, and call them torsion angles to describe the rotation angle of the peptide plane produced by carbon atom α.

Owing to φ and $\psi \in [-180°, 180°]$, we use the probability density of the torsion angles with equally sized bins based on the 2D sample histogram and smoothed with a 2D Gaussian kernel to represent this information. A matrix of 19×19 bins is used to describe the range of angles φ and ψ. The value of each bin represents the frequency information of the torsion angles. Finally, a feature vector used to describe the distribution of different angles is obtained.

We can also select a specific number of amino acid sequence fragments in a protein and use 19×19 bins to represent the information of an amino acid. Finally, the peptide bond information of each amino acid is combined as a feature of a protein.

E. FUNCTION DOMAIN (FunD)

A protein can be divided into several fixed modules or regions, called functional domains, which can often be found in other proteins. The information as to whether these fixed modules or regions appear in a protein or not can be regarded as the characteristics of a protein. There are many databases, such as Pfam, HMMER, SMART, COG, KOG and CDD, to help us search a query protein.

First, use a specific program such as RPS-BLAST to retrieve functional domain information of enzymes from the database. Second, the protein P can be formulated as:

$$P_{FunD} = [D_1, D_2, \ldots, D_i, \ldots, D_n]^T \tag{13}$$

where n is the number of protein domains determined by the selected database and T is the transpose operator. Element D_i can be obtained by (14).

$$D_i = \begin{cases} 1 & \text{a hit of } P \text{ is found in the } i\text{th profile} \\ 0 & \text{otherwrise} \end{cases} \tag{14}$$

F. COMPOSITION, TRANSITION AND DISTRIBUTION (CTD)

With this method, a sequence is described from the following three different perspectives [57]. First is the specific amino acid content, defined as:

$$composition = \frac{A}{N} \tag{15}$$

where N is the total number of amino acids and A is the selected one with specific physicochemical properties. The second perspective is the combination content with two amino acids selected and it is calculated as follows:

$$transition = \frac{AB + BA}{N-1} \tag{16}$$

In (16), AB and BA represent a combination mode, where the amino acids A and B meeting specific conditions are next to each other. The third one calculates the distribution of amino acids with specific residues in the sequence, as shown below:

$$distribution = \frac{A}{N_1} \quad (i = 1, 2, \ldots, N) \tag{17}$$

which is the proportion of type A amino acids in the first number of N_i amino acids of the protein sequence.

IV. CLASSIFICATION METHODS

With the development of enzyme function prediction, the classification process becomes more and more complex. However, no matter how abstruse the classification algorithm is, it is basically composed of the classic classification algorithms in different ways. Here, we briefly introduce some common classification methods.

A. LINEAR DISCRIMINANT ANALYSIS (LDA)

In a sense, a linear classifier is a dimension reduction operation. The specific operation uses the following transformation to map points in a multidimensional space to a line:

$$y = \omega^T x \tag{18}$$

where x and y separately express the coordinates of high dimensional space and the result of the projection of vector x in one-dimensional space. Then, we compare the result y with the threshold value we set, and determine the category of vector x by comparing the result as shown below:

$$\begin{cases} C_1 y \geq y_0 \\ C_2 y < y_0 \end{cases} \tag{19}$$

Here, we take two categories as an example, where C_1 and C_2 are categories and y_0 is the threshold value. Different values of ω in (18) lead to different results. We need to maximize the distance between classes and minimize the distance within classes to get the best effect. According to the Fisher linear discriminant, ω can be obtained by the following formula:

$$\omega = S_\omega^{-1}(m_1 - m_2) \tag{20}$$

where m_1-m_2 represents the distance between classes and S_ω represents total variance within classes.

B. ARTIFICIAL NEURAL NETWORK (ANN)

ANN, as a mathematical model to simulate the processing mechanism of complex information in the human brain, is very suitable as a classifier. It is often used in the classification of enzymes and other bioinformatics problems [16], [58], [59].

The neural network is composed of many computing units called neurons. By adjusting the connection between a large number of internal neurons and synthesizing the operation results of all neurons, the purpose of information processing can be achieved. Each neuron carries a threshold (θ), an activation function (F), and a set of weights (ω) for the input data. The output of each neuron can be obtained by the following formula:

$$y = f(\sum_{i=1}^{n} \omega_i x_i - \theta) \tag{21}$$

where n is the number of inputs.

By using back propagation arithmetic based on a strategy of gradient descent, the thresholds (θ) and weights (ω) of each neuron are constantly modified until it has a satisfactory classification ability. With improvements in the technical applications of ANN, many ANNs with specific structures are designed according to different situations, such as CNN and LNN.

C. K-Nearest-Neighbor (KNN)

This is a relatively simple algorithm compared with other algorithms, and its logic is uncomplicated and direct: classifying a sample into the category belonging to the closest sample to it. The algorithm consists of the following 4 steps:

(1) Calculate all distances from each point in the known class to the current point.
(2) Sort the values of distances from the previous step by increasing the distance.
(3) Define a range around the point to be classified, which contains k other samples.
(4) Determine the category to which most of the k samples belong, and then classify this sample into this category.

In this algorithm, the k, representing the number of the samples around point to be classified, is very important because it directly affects the effect of classification and if its value is small, the noise in the sample will have a great impact on the classification. On the other hand, if the value of k is larger, there will be a larger classification error.

There are also many improved algorithms based on KNN, such as ML-KNN and OET-KNN.

D. SUPPORT VECTOR MACHINE (SVM)

As a classic two-classification model, SVM is widely used in bioinformatics [60]–[78]. The basic idea of SVM is finding a hyperplane to segment samples meeting the condition of the sum of the distance from two kinds of samples from which it is the farthest.

The process of classification can be simply described as:

$$\begin{cases} \omega^T x_i + b \geq +1 & y_i = +1 \\ \omega^T x_i + b \leq -1 & y_i = -1 \end{cases} \tag{22}$$

where xi is the vector of sample i, for which the class is determined by the result of the formula $\omega^T x_i + b$. For finding the sum of the distance, the following conditions are derived:

$$\max \frac{2}{\|\omega\|} \quad s.t.\ y_i(\omega^T x_i + b) \geq +1 \tag{23}$$

However, it is difficult to solve (22) directly. SVM is divided into hard margins and soft margins, because the samples cannot be separated completely in the actual classification process. As a result, usually one adopts the strategy of a soft margin and solves the formula whose solution is the same as (22), as follows:

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \\ &s.t. \sum_{i=1}^{m} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C,\ i = 1, 2, \ldots, m \end{aligned} \tag{24}$$

where α is trained by maximizing the Lagrangian expression and k represents the kernel function. Finally, the function for the solution is obtained:

$$f(x) = \sum_{i=1}^{m} \alpha_i y_i k(x_i, x_j) + b \tag{25}$$

Two parameters, C and g, need to be set before the classification process, which can be calculated by software such as LibSVM.

V. EVALUATION AND COMMENT

In addition to the above aspects, the measures of evaluating the prediction result and the performance of a predictor is worth mentioning. A statistical analysis and comments on recent published results follows.

A. EVALUATION MEASURE

It is often not objective to measure the result of a prediction only by its accuracy. The following metrics are usually adopted to evaluate prediction quality [4], [11], [79]–[91]:

$$\begin{aligned} recall &= \frac{TP}{TP + FN} \\ precision &= \frac{TP}{TP + FP} \\ accuracy &= \frac{TP + TN}{TP + FN + FP + FN} \\ F_1 &= \frac{2 \times recall \times precision}{recall + precision} \end{aligned} \tag{26}$$

where TP, FP, FN represent positive class determined as positive class, negative class determined as positive class, and positive class determined as negative class, respectively. In addition, other indicators such as ROC and AUC are often used to evaluate the prediction results.

For data set processing, there are three strategies: independent test, n-fold cross validation test, and jackknife test [92]–[103]. Of these strategies, jackknife is the most reasonable and widely used. Its process is to (1) divide the original data set into many parts, (2) select one part as the testing set and the rest as the training set, and (3) integrate the results of each testing set. As shown in (26) below:

$$\begin{aligned} X &= (x_1, x_2, \ldots, x_n) \\ x_i &= (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \end{aligned} \tag{27}$$

where data set X is divided into n parts and x_i is the ith training data set.

B. COMMENT ON PUBLISHED RESULTS

Here, we select the theses listed in Table 2 and evaluate them in detail.

In 2007, Shen *et al.* [33] adopted OET-KNN to predict enzyme function. In the process of prediction, first they used FunD and PseAAC to extract features and obtained testing sets S1 and S2. Second, they used OET-KNN to classify S1 and S2. Finally, they fused the classification results from S1 and S2 to obtain the final prediction results. The results showed that the ability to predict the enzyme in the subclass of oxidoreductases is somewhat poor, with a success rate by the jackknife test of 86.7%. In 2014, Wang *et al.* [34] adopted several different methods in feature extraction and classification of feature extraction methods to match one classification method and compared the prediction results of four prediction models. They found that the best prediction model is the combination of RAkEL-RF and CTD, with which the highest accuracy with 10-fold cross validation of the training data reached 97.99% and the test data reached 97.57%. In 2014, Zou *et al.* [47] used three methods to extract features and classify them and then compared the classification results. The best results with the jackknife test have an accuracy of 91.64%. It is worth noting that the classifier designed by Zou *et al.* can classify multi-label enzymes. In 2019, Zou *et al.* [54] developed the prediction model MIDEEPre, which combines three methods of feature extraction. It can also classify multi-label enzymes.

There are also some simple prediction models with good classification results. In 2019, Concu *et al.* [32] separately used LDA and ANN methods for classification with only four kinds of characteristics and the topological indices calculated from software S2SNet. Both achieved high accuracy (98.73% in LDA and 100% in ANN), and the models were validated using the cross-validation tool. Their prediction model is also one of the few designed to divide enzymes into seven main classes. In another thesis [51], the authors used more topological indices in a predictor model and obtained the best overall accuracy of 91.2%, which can classify enzymes into the level of subclasses. Not only are the order information of amino acids or the information about amino acid residues in a protein sequence featured, but the 3D structure of the proteins can also be extracted. Amidi *et al.* [48] designed a coordinate system to represent the 3D structure of proteins. He used CNN to classify a test set and achieved an overall accuracy of 77.6%. The accuracy was relatively low, perhaps because only the 3D structure of the protein was adopted as the feature. The two torsion angles of the polypeptide chain φ and ψ are also an embodiment of the protein 3D structure. In 2019, Amidi *et al.* [49] extracted two kinds of features, the two torsion angles and similarity quantification. They fused the features in two strategies, the feature and the decision, classified by SVM and NN. The best result was achieved by the strategy of decision-level fusion and combining the classification method of SVM and NN. The best overall accuracy was 85.4%, and their predictor model can also classify multi-label enzymes. In the same year, Gao *et al.* [50] extracted features using PSSM and 3D Structure and then used three CNN to complete three feature maps. Finally, KNN was used for classification. Representing the protein feature information by a feature map in this prediction model is a novel approach. The best overall accuracy was 92.34%. Dalkiran *et al.* [52] combined three prediction models: SPMap, Blast-KNN, and Pepstats. He gave different weights to the prediction results of each prediction model, and then combined the results as the prediction results. The precision of prediction was 99%. The strength of this prediction model lies in that it can classify enzymes to the level of substrate class. There are also predictors involved in the classification of multifunctional enzymes. In 2016, Che *et al.* [53] adopted the ACC method developed from

TABLE 3. The information of prediction models

thesis	Method		accuracy (%)	precision(%)	Predicting level
	Feature extraction	Classification			
[33]	FunD, PsePSSM	OET-KNN	94.2	none	2
[34]	CTD, PseAAC	RAkEL-RF, MLKNN	97.6	none	1
[47]	PseAAC, SAAC, GM	ML-KNN	90.6	91.6	1
[32]	topological indices	LDA, ANN	100.0	none	1
[48]	3D structure	CNN	77.6	none	1
[49]	torsion angles, Similarity assessment	SVM, NN	90.6	94.2	1
[50]	PSSM, 3D structure	CNN, KNN	92.3	none	1
[51]	topological indices	LDA, ANN	91.2	none	2
[52]	SPMap, BLAST, Pepstats	SPMap, KNN, SVM	none	99.0	4
[53]	PSSM, ACC	KNN	none	94.1and95.5	1
[54]	PSSM, Sequence One-Hot, FunD	CNN	97.6	none	2

*Predicting level(i.e. 0: enzyme or non-enzyme, 1: main class, 2: subclass, 3: sub-subclass and 4: substrate classes)

PSSM and used KNN in classification. With 5-fold cross-validation, the precision of distinguishing monofunctional enzymes approached 94.1% and 95.54% of multifunctional enzymes.

Table 3 shows the information of these prediction models mentioned in this theses more intuitively. In this table, if the accuracy or precision are not clearly given in the theses, we use ''none'' to represent it.

VI. CONCLUSION

For the classification of enzymes, the functions of different prediction models are not identical. For example, some models support the classification of multi-labeled enzymes, and some models can classify an enzyme to the level of subclass. Therefore, it is meaningless to compare the accuracy of one model to another.

We find that in the prediction models described in these recent papers, the algorithm of ANN is used frequently. This implies that with the increase in research, the function of the prediction model becomes more and more powerful, and many traditional classification algorithms such as SVM and RF cannot meet the increasingly complex classification requirements. Since the classification algorithm is designed based on the fixed mathematical principle, its performance is always stable. Therefore, the key to improve the classification effect is to adopt better feature extraction methods. This view has been confirmed in many prediction models. For instance, the model DEEPre uses five methods of feature extraction.

The types of feature extraction can be roughly divided into three categories. The first category extracts information from the sequence, such as AAC and PseAAC. The second category is about the structure information of proteins; for example, the representation of 3D structure and the torsion angles. The third category is the result of comparing the protein with the information in the corresponding database, such as FunD and PSSM. It should be noted that although structure is more thorough in describing protein characteristics, the features from structural information do not necessarily contribute to classification.

For many years, the development of feature extraction focused on quantifying the primary structure of proteins, such as PSSM, PseAAC, FunD, and CTD. Recently, some scholars have tried to adopt some specifically designed algorithms such as torsion angles to extract some secondary structure information for the protein. However, the key to the function of a protein is the physicochemical properties caused by its different spatial structure. That is, the tertiary structure of a protein is the crux to accurately judge its different functions. The complexity and unpredictability of advanced protein structure will be the bottleneck in the development of enzyme function prediction.

In brief, we believe that a satisfactory classification result can be obtained by using the features synthetically. Moreover, with the development of bioinformatics, more feature extraction methods will be found [104]–[107], and the classification effect will be further improved.

AUTHOR CONTRIBUTIONS

Yuming Zhao conceived and designed the project. Zhiyu Tao conducted experiments and analyzed the data. Zhiyu Tao and Benzhi Dong wrote the paper. Zhixia Teng and Yuming Zhao revised the manuscript. All authors read and approved the final manuscript.

COMPETING INTERESTS

The authors have declared no competing interests.

REFERENCES

[1] X.-Y. Cheng, W.-J. Huang, S.-C. Hu, H.-L. Zhang, H. Wang, J.-X. Zhang, H.-H. Lin, Y.-Z. Chen, Q. Zou, and Z.-L. Ji, ''A global characterization and identification of multifunctional enzymes,'' *PLoS ONE*, vol. 7, no. 6, 2012, Art. no. e38979.

[2] J. Yin, W. Sun, F. Li, J. Hong, X. Li, Y. Zhou, Y. Lu, M. Liu, X. Zhang, N. Chen, X. Jin, J. Xue, S. Zeng, L. Yu, and F. Zhu, ''VARIDT 1.0: Variability of drug transporter database,'' *Nucleic Acids Res.*, vol. 48, no. D1, pp. D1042–D1050, 2020.

[3] B. Malysiak-Mrozek and D. Mrozek, "An improved method for protein similarity searching by alignment of fuzzy energy signatures," *Int. J. Comput. Intell. Syst.*, vol. 4, no. 1, pp. 75–88, Feb. 2011.

[4] J.-X. Tan, H. Lv, F. Wang, F.-Y. Dao, W. Chen, and H. Ding, "A survey for predicting enzyme family classes using machine learning methods," *Current Drug Targets*, vol. 20, no. 5, pp. 540–550, Mar. 2019.

[5] D. Mrozek, P. Gosk, and B. Małysiak-Mrozek, "Scaling ab *initio* predictions of 3D protein structures in microsoft azure cloud," *J. Grid Comput.*, vol. 13, no. 4, pp. 561–585, Dec. 2015.

[6] L. Cheng, C. Qi, H. Zhuang, T. Fu, and X. Zhang, "GutMDisorder: A comprehensive database for dysbiosis of the gut microbiota in disorders and interventions," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D554–D560, Jan. 2020.

[7] L. Cheng, H. Yang, H. Zhao, X. Pei, H. Shi, J. Sun, Y. Zhang, Z. Wang, and M. Zhou, "MetSigDis: A manually curated resource for the metabolic signatures of diseases," *Briefings Bioinf.*, vol. 20, no. 1, pp. 203–209, Jan. 2019.

[8] B. Li, J. Tang, Q. Yang, S. Li, X. Cui, Y. Li, Y. Chen, W. Xue, X. Li, and F. Zhu, "NOREVA: Normalization and evaluation of MS-based metabolomics data," *Nucleic Acids Res.*, vol. 45, no. W1, pp. W162–W170, Jul. 2017.

[9] . W. Xue, F. Yang, P. Wang, G. Zheng, Y. Chen, X. Yao, and F. Zhu, "What contributes to serotonin–norepinephrine reuptake inhibitors' dual-targeting mechanism? The key role of transmembrane domain 6 in human serotonin and norepinephrine transporters revealed by molecular dynamics simulation," *ACS Chem Neurosci*, vol. 9, no. 5, pp. 1128–1140, 2018.

[10] B. Li, J. Tang, Q. Yang, X. Cui, S. Li, S. Chen, Q. Cao, W. Xue, N. Chen, and F. Zhu, "Performance evaluation and online realization of data-driven normalization methods used in LC/MS based untargeted metabolomics analysis," *Sci. Rep.*, vol. 6, no. 1, Dec. 2016, Art. no. 38881.

[11] B. Liu, X. Gao, and H. Zhang, "BioSeq-Analysis2.0: An updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches," *Nucleic Acids Res.*, vol. 47, no. 20, pp. e127–e127, Nov. 2019.

[12] B. Malysiak-Mrozek, M. Stabla, and D. Mrozek, "Soft and declarative fishing of information in big data lake," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 5, pp. 2732–2747, Oct. 2018.

[13] L. Peng, M. Peng, B. Liao, G. Huang, W. Li, and D. Xie, "The advances and challenges of deep learning application in biological big data processing," *Current Bioinf.*, vol. 13, no. 4, pp. 352–359, Jul. 2018.

[14] Z. Lv, C. Ao, and Q. Zou, "Protein function prediction: From traditional classifier to deep learning," *Proteomics*, vol. 19, no. 14, Jul. 2019, Art. no. 1900119.

[15] L. Yu, X. Sun, S. Tian, X. Shi, and Y. Yan, "Drug and nondrug classification based on deep learning with various feature selection strategies," *Current Bioinf.*, vol. 13, no. 3, pp. 253–259, May 2018.

[16] L. Nie, L. Deng, C. Fan, W. Zhan, and Y. Tang, "Prediction of protein S-Sulfenylation sites using a deep belief network," *Current Bioinf.*, vol. 13, no. 5, pp. 461–467, Sep. 2018.

[17] Q. Zou, P. Xing, L. Wei, and B. Liu, "Gene2vec: Gene subsequence embedding for prediction of mammalian N6-methyladenosine sites from mRNA," *RNA*, vol. 25, no. 2, pp. 205–218, Feb. 2019.

[18] S. Wen, H. Wei, Z. Yan, Z. Guo, Y. Yang, T. Huang, and Y. Chen, "Memristor-based design of sparse compact convolutional neural network," *IEEE Trans. Netw. Sci. Eng.*, early access, Aug. 14, 2019, doi: 10.1109/TNSE.2019.2934357.

[19] S. Wen, M. Dong, Y. Yang, P. Zhou, T. Huang, and Y. Chen, "End-to-end detection-segmentation system for face labeling," *IEEE Trans. Emerg. Topics Comput. Intell.*, early access, Nov. 6, 2019, doi: 10.1109/TETCI.2019.2947319.

[20] J. Han, X. Han, Q. Kong, and L. Cheng, "PsSubpathway: A software package for flexible identification of phenotype-specific subpathways in cancer progression," *Bioinformatics*, vol. 36, no. 7, pp. 2303–2305, Apr. 2020.

[21] L. Cheng, Y. Hu, J. Sun, M. Zhou, and Q. Jiang, "DincRNA: A comprehensive Web-based bioinformatics toolkit for exploring disease associations and ncRNA function," *Bioinformatics*, vol. 34, no. 11, pp. 1953–1956, Jun. 2018.

[22] L. Yu, S. Yao, L. Gao, and Y. Zha, "Conserved disease modules extracted from multilayer heterogeneous disease and gene networks for understanding disease mechanisms and predicting disease treatments," *Frontiers Genet.*, vol. 9, p. 745, Jan. 2019.

[23] L. Yu and L. Gao, "Human pathway-based disease network," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 4, pp. 1240–1249, Jul. 2019.

[24] J. Hong, Y. Luo, Y. Zhang, J. Ying, W. Xue, T. Xie, L. Tao, and F. Zhu, "Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning," *Briefings Bioinf.*, 2019, doi: 10.1093/bib/bbz081.

[25] B. Liu, C. Li, and K. Yan, "DeepSVM-fold: Protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks," *Briefings Bioinf.*, to be published, doi: 10.1093/bib/bbz098.

[26] C.-C. Li and B. Liu, "MotifCNN-fold: Protein fold recognition based on fold-specific features extracted by motif-based convolutional neural networks," *Briefings Bioinf.*, Nov. 2019, doi: 10.1093/bib/bbz133.

[27] X. Zhao, Q. Jiao, H. Li, Y. Wu, H. Wang, S. Huang, and G. Wang, "ECFS-DEA: An ensemble classifier-based feature selection for differential expression analysis on expression profiles," *BMC Bioinf.*, vol. 21, no. 1, p. 43, Dec. 2020.

[28] D. Mrozek, B. Socha, S. Kozielski, and B. Małysiak-Mrozek, "An efficient and flexible scanning of databases of protein secondary structures," *J. Intell. Inf. Syst.*, vol. 46, no. 1, pp. 213–233, Feb. 2016.

[29] Q. Zou, W. Chen, Y. Huang, X. Liu, and Y. Jiang, "Identifying multi-functional enzyme by hierarchical multi-label classifier," *J. Comput. Theor. Nanosci.*, vol. 10, no. 4, pp. 1038–1043, Apr. 2013.

[30] E. Nasibov and C. Kandemir-Cavas, "Efficiency analysis of KNN and minimum distance-based classifiers in enzyme family prediction," *Comput. Biol. Chem.*, vol. 33, no. 6, pp. 461–464, Dec. 2009.

[31] J.-D. Qiu, J.-H. Huang, S.-P. Shi, and R.-P. Liang, "Using the concept of chous pseudo amino acid composition to predict enzyme family classes: An approach with support vector machine based on discrete wavelet transform," *Protein Peptide Lett.*, vol. 17, no. 6, pp. 715–722, Jun. 2010.

[32] R. Concu and M. N. D. S. Cordeiro, "Alignment-free method to predict enzyme classes and subclasses," *Int. J. Mol. Sci.*, vol. 20, no. 21, p. 5389, 2019.

[33] H.-B. Shen and K.-C. Chou, "EzyPred: A top–down approach for predicting enzyme functional classes and subclasses," *Biochem. Biophy. Res. Commun.*, vol. 364, no. 1, pp. 53–59, Dec. 2007.

[34] Y. Wang, R. Jing, Y. Hua, Y. Fu, X. Dai, L. Huang, and M. Li, "Classification of multi-family enzymes by multi-label machine learning and sequence-based descriptors," *Anal. Methods*, vol. 6, no. 17, pp. 6832–6840, 2014.

[35] J. Y. Ryu, H. U. Kim, and S. Y. Lee, "Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers," *Proc. Nat. Acad. Sci. USA*, vol. 116, no. 28, pp. 13996–14001, Jul. 2019.

[36] Q. Zou, G. Lin, X. Jiang, X. Liu, and X. Zeng, "Sequence clustering in bioinformatics: An empirical study," *Briefings Bioinf.*, vol. 21, no. 1, pp. 1–10, Sep. 2018.

[37] D. Mrozek, B. Malysiak, and S. Kozielski, "An optimal alignment of proteins energy characteristics with crisp and fuzzy similarity awards," in *Proc. IEEE Int. Fuzzy Syst. Conf.*, Jun. 2007, pp. 1–6.

[38] B. Małysiak-Mrozek, T. Baron, and D. Mrozek, "Spark-IDPP: High-throughput and scalable prediction of intrinsically disordered protein regions with spark clusters on the cloud," *Cluster Comput.*, vol. 22, no. 2, pp. 487–508, Jun. 2019.

[39] M. L. Liu, W. Su, Z. X. Guan, D. Zhang, W. Chen, L. Liu, and H. Ding, "An overview on predicting protein subchloroplast localization by using machine learning methods," *Current Protein Peptide Sci.*, 2020, doi: 10.2174/1389203721666200117153412.

[40] H. Ding and D. Li, "Identification of mitochondrial proteins of malaria parasite using analysis of variance," *Amino Acids*, vol. 47, no. 2, pp. 329–333, Feb. 2015.

[41] H. Tang, W. Chen, and H. Lin, "Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique," *Mol. BioSyst.*, vol. 12, no. 4, pp. 1269–1275, 2016.

[42] X.-X. Chen, H. Tang, W.-C. Li, H. Wu, W. Chen, H. Ding, and H. Lin, "Identification of bacterial cell wall lyases via pseudo amino acid composition," *BioMed Res. Int.*, vol. 2016, Jun. 2016, Art. no. 1654623.

[43] . H. Yang, H. Tang, X.-X. Chen, C.-J. Zhang, P.-P. Zhu, H. Ding, W. Chen, and H. Lin, "Identification of secretory proteins in mycobacterium tuberculosis using pseudo amino acid composition," *BioMed Res. Int.*, vol. 2016, Aug. 2016, Art. no. 5413903.

[44] W. Chen, P. Feng, T. Liu, and D. Jin, "Recent advances in machine learning methods for predicting heat shock proteins," *Current Drug Metabolism*, vol. 20, no. 3, pp. 224–228, May 2019.

[45] L. Wei, Y. Ding, R. Su, J. Tang, and Q. Zou, "Prediction of human protein subcellular localization using deep learning," *J. Parallel Distrib. Comput.*, vol. 117, pp. 212–217, Jul. 2018.

[46] B. Liu, "BioSeq-analysis: A platform for DNA, RNA and protein sequence analysis based on machine learning approaches," *Briefings Bioinf.*, vol. 20, no. 4, pp. 1280–1294, Jul. 2019.
[47] H.-L. Zou and X. Xiao, "Classifying multifunctional enzymes by incorporating three different models into Chou's general pseudo amino acid composition," *J. Membrane Biol.*, vol. 249, no. 4, pp. 551–557, Aug. 2016.
[48] A. Amidi, S. Amidi, D. Vlachakis, V. Megalooikonomou, N. Paragios, and E. I. Zacharaki, "EnzyNet: Enzyme classification using 3D convolutional neural networks on spatial representation," *PeerJ*, vol. 6, p. e4750, May 2018.
[49] S. Amidi, A. Amidi, D. Vlachakis, N. Paragios, and E. I. Zacharaki, "Automatic single- and multi-label enzymatic function prediction by machine learning," *PeerJ*, vol. 5, p. e3095, Mar. 2017.
[50] R. Gao, M. Wang, J. Zhou, Y. Fu, M. Liang, D. Guo, and J. Nie, "Prediction of enzyme function based on three parallel deep CNN and amino acid mutation," *Int. J. Mol. Sci.*, vol. 20, no. 11, p. 2845, 2019.
[51] R. Concu, M. N. D. S. Cordeiro, C. R. Munteanu, and H. González-Díaz, "PTML model of enzyme subclasses for mining the proteome of biofuel producing microorganisms," *J. Proteome Res.*, vol. 18, no. 7, pp. 2735–2746, Jul. 2019.
[52] A. Dalkiran, A. S. Rifaioglu, M. J. Martin, R. Cetin-Atalay, V. Atalay, and T. Doğan, "ECPred: A tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature," *BMC Bioinf.*, vol. 19, no. 1, p. 334, Dec. 2018.
[53] Y. Che, Y. Ju, P. Xuan, R. Long, and F. Xing, "Identification of multifunctional enzyme with multi-label classifier," *PLoS ONE*, vol. 11, no. 4, 2016, Art. no. e0153503.
[54] Z. Zou, S. Tian, X. Gao, and Y. Li, "MlDEEPre: Multi-functional enzyme function prediction with hierarchical multi-label deep learning," *Frontiers Genet.*, vol. 9, p. 714, Jan. 2019.
[55] Y. Xiong, J. Liu, W. Zhang, and T. Zeng, "Prediction of heme binding residues from protein sequences with integrative sequence profiles," *Proteome Sci.*, vol. 10, no. 1, p. S20, 2012.
[56] Y. H. Li, C. Y. Yu, X. X. Li, P. Zhang, J. Tang, Q. Yang, T. Fu, X. Zhang, X. Cui, G. Tu, Y. Zhang, S. Li, F. Yang, Q. Sun, C. Qin, X. Zeng, Z. Chen, Y. Z. Chen, and F. Zhu, "Therapeutic target database update 2018: Enriched resource for facilitating bench-to-clinic research of targeted therapeutics," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1121–D1127, Jan. 2018.
[57] J.-H. Cheng, H. Yang, M.-L. Liu, W. Su, P.-M. Feng, H. Ding, W. Chen, and H. Lin, "Prediction of bacteriophage proteins located in the host cell using hybrid features," *Chemometric Intell. Lab. Syst.*, vol. 180, pp. 64–69, Sep. 2018.
[58] X. Zeng, W. Wang, G. Deng, J. Bing, and Q. Zou, "Prediction of potential disease-associated MicroRNAs by using neural networks," *Mol. Therapy-Nucleic Acids*, vol. 16, pp. 566–575, Jun. 2019.
[59] J. Tang, J. Fu, Y. Wang, Y. Luo, Q. Yang, B. Li, G. Tu, J. Hong, X. Cui, Y. Chen, L. Yao, W. Xue, and F. Zhu, "Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome quantification by optimizing data manipulation chains," *Mol. Cellular Proteomics*, vol. 18, no. 8, pp. 1683–1699, Aug. 2019.
[60] Y. Wang, F. Shi, L. Cao, N. Dey, Q. Wu, A. S. Ashour, R. S. Sherratt, V. Rajinikanth, and L. Wu, "Morphological segmentation analysis and texture-based support vector machines classification on mice liver fibrosis microscopic images," *Current Bioinf.*, vol. 14, no. 4, pp. 282–294, Apr. 2019.
[61] C. Meng, S. Jin, L. Wang, F. Guo, and Q. Zou, "AOPs-SVM: A sequence-based classifier of antioxidant proteins using a support vector machine," *Frontiers Bioeng. Biotechnol.*, vol. 7, p. 224, Sep. 2019.
[62] N. Zhang, Y. Sa, Y. Guo, W. Lin, P. Wang, and Y. Feng, "Discriminating Ramos and Jurkat cells with image textures from diffraction imaging flow cytometry based on a support vector machine," *Current Bioinf.*, vol. 13, no. 1, pp. 50–56, Feb. 2018.
[63] . L. Chao, L. Wei, and Q. Zou, "SecProMTB: A SVM-based classifier for secretory proteins of Mycobacterium tuberculosis with imbalanced data set," *Proteomics*, vol. 19, Aug. 2019, Art. no. e1900007.
[64] H. Ding, W. Yang, H. Tang, P.-M. Feng, J. Huang, W. Chen, and H. Lin, "PHYPred: A tool for identifying bacteriophage enzymes and hydrolases," *Virologica Sinica*, vol. 31, no. 4, pp. 350–352, Aug. 2016.
[65] H. Yang, W. Yang, F. Y. Dao, H. Lv, H. Ding, W. Chen, and H. Lin, "A comparison and assessment of computational method for identifying recombination hotspots in Saccharomyces cerevisiae," *Briefings Bioinf.*, 2019, doi: 10.1093/bib/bbz123.
[66] L. Wei, C. Zhou, H. Chen, J. Song, and R. Su, "ACPred-FL: A sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides," *Bioinformatics*, vol. 34, no. 23, pp. 4007–4016, Jun. 2018.
[67] Y. Ding, J. Tang, and F. Guo, "Identification of drug-target interactions via multiple information integration," *Inf. Sci.*, vols. 418–419, pp. 546–560, Dec. 2017.
[68] Y. Shen, J. Tang, and F. Guo, "Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC," *J. Theor. Biol.*, vol. 462, pp. 230–239, Feb. 2019.
[69] X. Zhu, J. He, S. Zhao, W. Tao, Y. Xiong, and S. Bi, "A comprehensive comparison and analysis of computational predictors for RNA N6-methyladenosine sites of saccharomyces cerevisiae," *Briefings Funct. Genomics*, vol. 18, no. 6, pp. 367–376, Oct. 2019.
[70] Y. Xiong, Y. Qiao, D. Kihara, H.-Y. Zhang, X. Zhu, and D.-Q. Wei, "Survey of machine learning techniques for prediction of the isoform specificity of cytochrome P450 substrates," *Current Drug Metabolism*, vol. 20, no. 3, pp. 229–235, May 2019.
[71] J. He, T. Fang, Z. Zhang, B. Huang, X. Zhu, and Y. Xiong, "PseUI: Pseudouridine sites identification based on RNA sequence information," *BMC Bioinf.*, vol. 19, no. 1, p. 306, Dec. 2018.
[72] L. Deng, J. Wang, and J. Zhang, "Predicting gene ontology function of human MicroRNAs by integrating multiple networks," *Frontiers Genet.*, vol. 10, p. 3, Jan. 2019.
[73] Y. H. Li, X. X. Li, J. J. Hong, Y. X. Wang, J. B. Fu, H. Yang, C. Y. Yu, F. C. Li, J. Hu, W. W. Xue, Y. Y. Jiang, Y. Z. Chen, and F. Zhu, "Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs," *Briefings Bioinf.*, vol. 21, no. 2, pp. 649–662, Mar. 2020.
[74] J. Tang, J. Fu, Y. Wang, B. Li, Y. Li, Q. Yang, X. Cui, J. Hong, X. Li, Y. Chen, W. Xue, and F. Zhu, "ANPELA: Analysis and performance assessment of the label-free quantification workflow for metaproteomic studies," *Briefings Bioinf.*, vol. 21, no. 2, pp. 621–636, Mar. 2020.
[75] B. Liu and K. Li, "IPromoter-2L2.0: Identifying promoters and their types by combining smoothing cutting window algorithm and sequence-based features," *Mol. Therapy-Nucleic Acids*, vol. 18, pp. 80–87, Dec. 2019.
[76] Y. Zhao, F. Wang, S. Chen, J. Wan, and G. Wang, "Methods of MicroRNA promoter prediction and transcription factor mediated regulatory network," *BioMed Res. Int.*, vol. 2017, Jun. 2017, Art. no. 7049406.
[77] Y. Zhao, F. Wang, and L. Juan, "MicroRNA promoter identification inArabidopsisUsing multiple histone markers," *BioMed Res. Int.*, vol. 2015, 2015, Art. no. 861402.
[78] Q. Jiang, G. Wang, S. Jin, Y. Li, and Y. Wang, "Predicting human microRNA-disease associations based on support vector machine," *Int. J. Data Mining Bioinf.*, vol. 8, no. 3, pp. 282–293, 2013.
[79] L. Cheng, H. Zhao, P. Wang, W. Zhou, M. Luo, T. Li, J. Han, S. Liu, and Q. Jiang, "Computational methods for identifying similar diseases," *Mol. Therapy-Nucleic Acids*, vol. 18, pp. 590–604, Dec. 2019.
[80] J.-X. Tan, S.-H. Li, Z.-M. Zhang, C.-X. Chen, W. Chen, H. Tang, and H. Lin, "Identification of hormone binding proteins based on machine learning methods," *Math. Biosci. Eng.*, vol. 16, no. 4, pp. 2466–2480, 2019.
[81] W. Yang, X.-J. Zhu, J. Huang, H. Ding, and H. Lin, "A brief survey of machine learning methods in protein sub-golgi localization," *Current Bioinf.*, vol. 14, no. 3, pp. 234–240, Mar. 2019.
[82] W. Zhang, Z. Li, W. Guo, W. Yang, and F. Huang, "A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Jul. 29, 2019, doi: 10.1109/TCBB.2019.2931546.
[83] W. Zhang, X. Yue, G. Tang, W. Wu, F. Huang, and X. Zhang, "SFPEL-LPI: Sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions," *PLOS Comput. Biol.*, vol. 14, no. 12, 2018, Art. no. e1006616.
[84] L. Cheng, "Computational and biological methods for gene therapy," *Current Gene Therapy*, vol. 19, no. 4, pp. 210–210, Nov. 2019.
[85] Y. Chu, A. C. Kaushik, X. Wang, W. Wang, Y. Zhang, X. Shan, D. R. Salahub, Y. Xiong, and D. Q. Wei, "DTI-CDF: A cascade deep forest model towards the prediction of drug-target interactions based on hybrid features," *Briefings Bioinf.*, 2019, doi: 10.1093/bib/bbz152.

[86] L. Deng, W. Li, and J. Zhang, "LDAH2V: Exploring meta-paths across multiple networks for lncRNA-disease association prediction," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Nov. 12, 2019, doi: 10.1109/TCBB.2019.2946257.

[87] . X. Shan, X. Wang, C. D. Li, Y. Chu, Y. Zhang, Y. Xiong, and D.-Q. Wei, "Prediction of CYP450 enzyme–substrate selectivity based on the network-based label space division method," *J. Chem. Inf. Model.*, vol. 59, no. 11, pp. 4577–4586, 2019.

[88] G. Wang, Y. Wang, W. Feng, X. Wang, J. Y. Yang, Y. Zhao, Y. Wang, and Y. Liu, "Transcription factor and microRNA regulation in androgen-dependent and-independent prostate cancer cells," *BMC Genomics*, vol. 9, no. 2, p. S22, 2008.

[89] G. Wang, Y. Wang, M. Teng, D. Zhang, L. Li, and Y. Liu, "Signal transducers and activators of transcription-1 (STAT1) regulates microRNA transcription in interferon γ-stimulated HeLa cells," *PLoS ONE*, vol. 5, no. 7, 2010, Art. no. e11794.

[90] Y. Xiong, Q. Wang, J. Yang, X. Zhu, and D.-Q. Wei, "PredT4SE-stack: Prediction of bacterial type IV secreted effectors from protein sequences using a stacked ensemble method," *Frontiers Microbiol.*, vol. 9, p. 2571, Oct. 2018.

[91] L. Yu, J. Zhao, and L. Gao, "Predicting potential drugs for breast cancer based on miRNA and tissue specificity," *Int. J. Biol. Sci.*, vol. 14, no. 8, pp. 971–982, 2018.

[92] L. Wei, H. Chen, and R. Su, "M6APred-EL: A sequence-based predictor for identifying N6-methyladenosine sites using ensemble learning," *Mol. Therapy-Nucleic Acids*, vol. 12, pp. 635–644, Sep. 2018.

[93] L. Wei, P. Xing, J. Zeng, J. Chen, R. Su, and F. Guo, "Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier," *Artif. Intell. Med.*, vol. 83, pp. 67–74, Nov. 2017.

[94] L. Wei, S. Wan, J. Guo, and K. K. Wong, "A novel hierarchical selective ensemble classifier with bioinformatics application," *Artif. Intell. Med.*, vol. 83, pp. 82–90, Nov. 2017.

[95] Y. Ding, J. Tang, and F. Guo, "Predicting protein-protein interactions via multivariate mutual information of protein sequences," *BMC Bioinf.*, vol. 17, no. 1, p. 398, Dec. 2016.

[96] Y. Ding, J. Tang, and F. Guo, "Identification of protein–protein interactions via a novel matrix-based sequence representation model with amino acid contact information," *Int. J. Mol. Sci.*, vol. 17, no. 10, p. 1623, 1623.

[97] Y. Pan, D. Liu, and L. Deng, "Accurate prediction of functional effects for variants by combining gradient tree boosting with optimal neighborhood properties," *PLoS ONE*, vol. 12, no. 6, 2017, Art. no. e0179314.

[98] Y. Xiao, J. Zhang, and L. Deng, "Prediction of lncRNA-protein interactions using HeteSim scores based on heterogeneous networks," *Sci. Rep.*, vol. 7, no. 1, p. 3664, Dec. 2017.

[99] Y. Wang, S. Zhang, F. Li, Y. Zhou, Y. Zhang, Z. Wang, R. Zhang, J. Zhu, Y. Ren, Y. Tan, C. Qin, Y. Li, X. Li, Y. Chen, and F. Zhu, "Therapeutic target database 2020: Enriched resource for facilitating research and early development of targeted therapeutics," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D1031–D1041, Nov. 2019.

[100] F. Li, Y. Zhou, X. Zhang, J. Tang, Q. Yang, Y. Zhang, Y. Luo, J. Hu, W. Xue, Y. Qiu, Q. He, B. Yang, and F. Zhu, "SSizer: Determining the sample sufficiency for comparative biological study," *J. Mol. Biol.*, 2020, doi: 10.1016/j.jmb.2020.01.027.

[101] K. Yan, X. Fang, Y. Xu, and B. Liu, "Protein fold recognition based on multi-view modeling," *Bioinformatics*, vol. 35, pp. 2982–2990, Sep. 2019.

[102] L. Cheng, P. Wang, R. Tian, S. Wang, Q. Guo, M. Luo, W. Zhou, G. Liu, H. Jiang, and Q. Jiang, "LncRNA2Target v2.0: A comprehensive database for target genes of lncRNAs in human and mouse," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D140–D144, Jan. 2019.

[103] G. Wang, X. Luo, J. Wang, J. Wan, S. Xia, H. Zhu, J. Qian, and Y. Wang, "MeDReaders: A database for transcription factors that bind to methylated DNA," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D146–D151, Jan. 2018.

[104] K. Yan, J. Wen, J.-X. Liu, Y. Xu, and B. Liu, "Protein fold recognition by combining support vector machines and pairwise sequence similarity scores," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Jan. 13, 2020, doi: 10.1109/TCBB.2020.2966450.

[105] Q. Yang, B. Li, J. Tang, X. Cui, Y. Wang, X. Li, J. Hu, Y. Chen, W. Xue, Y. Lou, Y. Qiu, and F. Zhu, "Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data," *Briefings Bioinf.*, 2019, doi: 10.1093/bib/bbz049.

[106] N. Zheng, K. Wang, W. Zhan, and L. Deng, "Targeting virus-host protein interactions: Feature extraction and machine learning approaches," *Current Drug Metabolism*, vol. 20, no. 3, pp. 177–184, May 2019.

[107] X.-J. Zhu, C.-Q. Feng, H.-Y. Lai, W. Chen, and L. Hao, "Predicting protein structural classes for low-similarity sequences by evaluating different features," *Knowl.-Based Syst.*, vol. 163, pp. 787–793, Jan. 2019.

ZHIYU TAO was born in Liaoning, China, in 1993. He received the B.S. degree in computer science and technology from Northeast Forestry University, in 2016. He is currently pursuing the master's degree. During the undergraduate period, he served as the Project Leader of the Science and Technology Innovation Project of National College Students and published an article and a patent. In 2019, he came to the University of Electronic Science and technology as a Visiting Student for further study. His research interests include machine learning, classification of enzymes, and the prediction of protein structure.

BENZHI DONG was born in Heilongjiang, China, in 1975. He received the B.S. degree in computer engineering from Shenyang Ligong University, in 1997, and the M.S. degree in computer science and technology form the Harbin Institute of Technology, in 2004, and the Ph.D. degree in mechanical design and theory form Northeast Forestry University, in 2010. From 1997 to 2008, he was a Lecturer with the College of Computer Science and Engineering, Northeast Forestry University, Harbin, China, where he has been an Assistant Professor with the College of Computer Science and Engineering, since 2008. He has authored more than 40 articles. His research interests include CAD/CAM, computer vision, and computer recognition of plants and insects. Dr. Dong was a member of the China Computer Society.

ZHIXIA TENG was born in Lanzhou, Gansu, China, in 1982. She received the B.S. degree in information management and information system and the M.S. degree in computer application technology from Northeast Forestry University, in 2005 and 2008, respectively, and the Ph.D. degree in computer science and technology from the Harbin Institute of Technology, in 2016. Since 2011, she has been a Lecture with the College of Information and Computer Engineering, Northeast Forestry University, Harbin, China. Her current research interests focus on bioinformatics, computational system biology, and machine learning.

YUMING ZHAO was born in Harbin, Heilongjiang, China, in 1978. She received the B.S. degree in computer science and technology from Yanshan University, in 2001, and the M.S. and Ph.D. degrees in computer science and technology form the Harbin Institute University, in 2003 and 2009, respectively. From 2010 to 2017, she was a Lecturer with the College of Computer Science and Engineering, Northeast Forestry University, Harbin, China, where she has been an Assistant Professor with the College of Computer Science and Engineering, since 2018. She has authored more than 20 articles. Her research interests are bioinformatics, machine learning, and algorithms.