Computers in Biology and Medicine 145 (2022) 105465

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

PFmulDL: a novel strategy enabling multi-class and multi-label protein function annotation by integrating diverse deep learning methods

Weiqi Xia [a], Lingyan Zheng [a,b], Jiebin Fang [a], Fengcheng Li [a], Ying Zhou [a], Zhenyu Zeng [b], Bing Zhang [b], Zhaorong Li [b], Honglin Li [c,*], Feng Zhu [a,b,**]

[a] College of Pharmaceutical Sciences, The Second Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang University, Hangzhou, 310058, China
[b] Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare, Hangzhou, 330110, China
[c] School of Pharmacy, East China University of Science and Technology, Shanghai, 200237, China

ARTICLE INFO

Keywords:
Protein function prediction
Deep learning
Gene ontology
Convolutional neural network
Recurrent neural network

ABSTRACT

Bioinformatic annotation of protein function is essential but extremely sophisticated, which asks for extensive efforts to develop effective prediction method. However, the existing methods tend to amplify the representativeness of the families with large number of proteins by misclassifying the proteins in the families with small number of proteins. That is to say, the ability of the existing methods to annotate proteins in the '*rare classes*' remains limited. Herein, a new protein function annotation strategy, PFmulDL, integrating multiple deep learning methods, was thus constructed. *First*, the recurrent neural network was integrated, for the first time, with the convolutional neural network to facilitate the function annotation. *Second*, a transfer learning method was introduced to the model construction for further improving the prediction performances. *Third*, based on the latest data of Gene Ontology, the newly constructed model could annotate the largest number of protein families comparing with the existing methods. *Finally*, this newly constructed model was found capable of significantly elevating the prediction performance for the '*rare classes*' without sacrificing that for the '*major classes*'. All in all, due to the emerging requirements on improving the prediction performance for the proteins in '*rare classes*', this new strategy would become an essential complement to the existing methods for protein function prediction. All the models and source codes are freely available and open to all users at: https://github.com/idrblab/PFmulDL.

1. Introduction

Proteins participate in most of the physiological functions/biological processes in a living system by two main ways [1–4]. One is to join the formation of tissues or organs as the structural proteins, and the other is to involve in the signal transduction, immune response, and biochemical reaction as the functional proteins [5–8]. The functional study of protein can help to understand its various mechanisms of cellular response and is of great significance for the discovery of drug targets and the study of the physiological/pathological process [9–11]. With the maturity of next-generation sequencing, a large number of protein sequences have been produced [12]. The UniProt database provides ~200 million sequences, but <1% of these proteins have been experimentally annotated [13–18]. Due to the time-consuming and labor-intensive nature of the experiment annotation [19–21], it is urgently needed to develop the methods enabling bioinformatic annotation [22].

However, bioinformatic annotation of protein function, as well-known, is very sophisticated [23], and the extensive effort has been devoted to the development of related methods [24]. There are two distinct types of bioinformatic annotation approaches which are based on *sequence similarity* and *machine learning*, respectively [25]. The fundamental principle of *sequence similarity*-based methods is that the proteins with similar sequences are likely to have similar functions, which is also known as the sequence homologous transfer [26–32]. So far, a variety of tools based on this method have been publicly available for all users, which include BLAST [33], GoFDR [34], *etc.* Although these tools have attracted extensive interests from relevant research communities, their innate limitations still need to be carefully considered [34,35]. *First*, when the sequence identity is lower than a certain point (~60%), the annotation accuracy of this method will be significantly

* Corresponding author. School of Pharmacy, East China University of Science and Technology, Shanghai, China.
** Corresponding author. College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, China.
E-mail addresses: hlli@ecust.edu.cn (H. Li), zhufeng@zju.edu.cn (F. Zhu).

https://doi.org/10.1016/j.compbiomed.2022.105465
Received 4 March 2022; Received in revised form 22 March 2022; Accepted 25 March 2022
Available online 28 March 2022

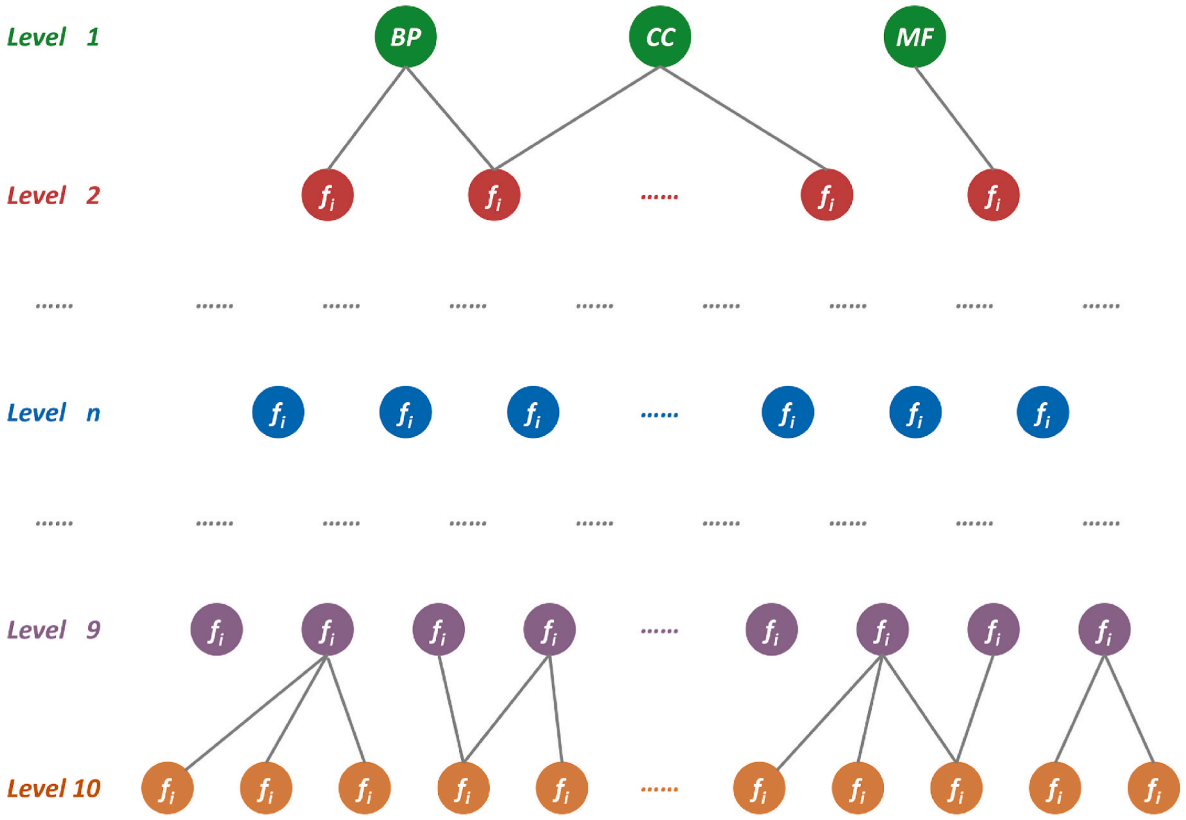

Fig. 1. Schematic illustration of the hierarchical multi-label structure of GO families (labeled by f_i). BP, MF and CC were three root nodes at the top of the structure, and the remaining families were hierarchically connected to them. In this study, the level of root nodes was defined as 'Level 1' (green). The child families directly connected to the root nodes were labeled as 'Level 2' (red). Then, the families of 'Level 3' were defined by those child families directly connected to 'Level 2'. The following levels can be thus deduced in the similar manner. Based on our comprehensive evaluation on all GO data, the bottom level of GO's hierarchical multi-label structure was 'Level 10', which had no child family and composed of the smallest number of proteins comparing with the families in other levels (Level 1 to Level 9).

affected [36,37]. *Second*, the time cost in similarity search needs to be reduced [38].

To cope with these limitations, the *machine learning* (ML) based methods have been constructed [39–50]. Particularly, the ML-based methods are unique in their ability to identify the functional homology irrespective of sequence similarity, which makes them powerful in annotating proteins of low sequence identity [51]. Moreover, the model constructed using ML-based methods usually demonstrates the fast process of functional annotations [52]. Till now, diverse tools based on this method have been developed, which are publicly available to all users for the function annotation [53,54]. As one of the most powerful and popular annotation tools, DeepGO has been developed, which predicts the protein functions from sequence and interactions using a deep ontology-aware classifier [53]. Based on the first version of DeepGO, two upgraded versions (DeepGOCNN and DeepGOPlus) have also been developed using multi-kernel convolutional neural network (CNN) and the combination of CNN and sequence similarity, respectively [54]. However, both methods (based on *sequence similarity & machine learning*) tend to amplify the representativeness of the families with large number of proteins by misclassifying the protein in families with less proteins [55]. In other words, the ability of the existing methods/tools to annotate the proteins in the '*rare classes*' remains limited [56]. Due to the key role played by the proteins in such '*rare classes*' as calcium ion homeostasis family important to cell signaling, hormone regulation, and bone health [57], adenine nucleotide transporters key to attenuate the myocardial ischemia-reperfusion injury [58], microtubule motor activity family essential for the transportation of substances within cells [59], *etc.*, it is essential to construct new strategy to significantly elevate annotation performance for *rare classes* without sacrificing that for the *major ones*.

In this study, a novel protein functional annotation strategy that integrated multiple deep learning methods (PFmulDL) was therefore constructed. *First*, a recurrent neural network (RNN) method was integrated, for the first time, with a multi-kernel CNN method to facilitate protein functional annotation. *Second*, the transfer learning (TL) method was introduced to model construction for further improving annotation performance. *Third*, based on the latest dataset collected from the *Gene Ontology* (GO) [60], PFmulDL became a tool capable of annotating the largest number of GO families. *Finally*, our newly developed PFmulDL was compared with various available tools (both *sequence similarity*-based & *machine learning*-based), and found to be able to significantly elevate the annotation performance for '*rare classes*' without sacrificing that for the '*major ones*'. All in all, due to the emerging requirement on improving the annotation performance for proteins in *rare classes*, the PFmulDL would become an essential complement to those available tools in the field of protein function annotation.

2. Materials and Methods

2.1. Data Collection for Model Construction and Assessment

Gene Ontology (GO) [60] provided comprehensive descriptions on genes and their products, and divided the biological functions of protein into three groups: *Biological Process* (BP), *Molecular Function* (MF), and *Cellular Component* (CC). The GO encapsulated protein function annotation into a directed acyclic graph. Each node in the graph is called a GO term, and the edges represent specified parent to child relations between the terms [61]. Protein annotations are thus considered as the hierarchical multi-label classification (HMC) task. In this case, if a protein is predicted to a certain GO term, then it will be labeled to all the

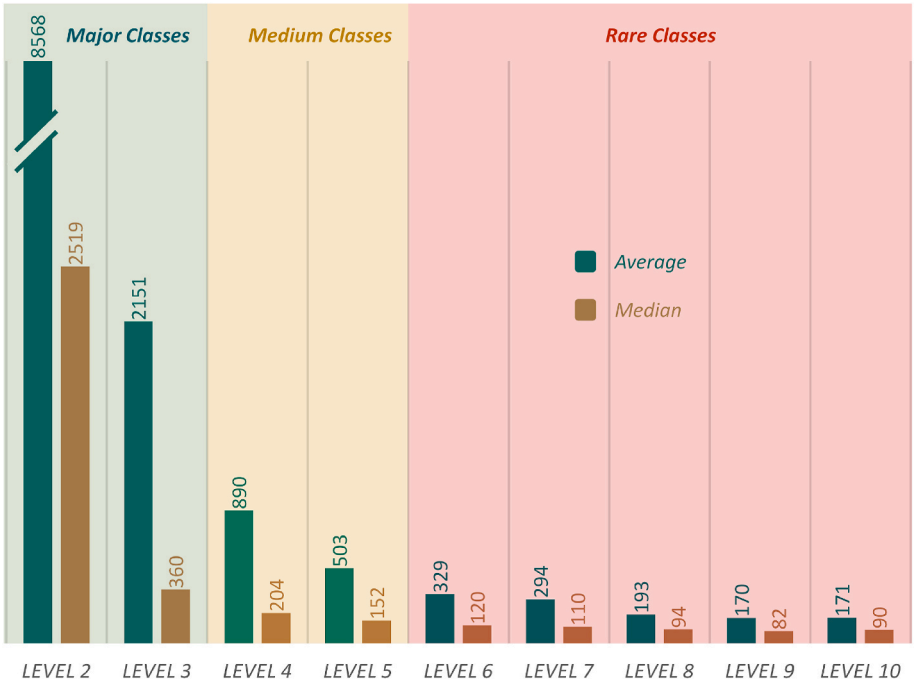

Fig. 2. Average (green) and median (brown) numbers of proteins in those GO families of nine different levels (Level 2 to Level 10). There was a clear descending trend of average and median numbers from the top level (Level 2) to the bottom one (Level 10). Since the numbers of proteins in particular families indicated the representativeness of the corresponding families in the model construction, it was reasonable to say that the representativeness of a family gradually decreased with the penetration into a deeper level. Therefore, the representativeness of each GO family was classified into three groups using both average and median numbers of proteins in nine levels of GO hierarchical structure. *Rare Classes* (average < 500 & median <150), *Medium Classes* (500 < average < 1000 & 150< median <300), and *Major Classes* (average > 1000 & median >300).

parent terms of that specific GO term [62,63]. Particularly, the hierarchical structure information about GO data was from the go.obo file which was downloaded from GO website, and the protein sequences were from a uniprot_sprot.dat file downloaded from UniPort. As a result, a total of 67,888 protein sequences with explicit GO term were collected. Like existing tools [53,54], only those GO families with relatively large number of proteins (>50) were included into model construction, which contained a total of 5825 protein sequences (77% BP terms, 12% MF terms, and 11% CC terms).

Within the downloaded go.obo file, GO families were represented in the hierarchical multi-label structure. Particularly, BP, MF, and CC were three root nodes at the top of the structure, and the remaining GO families were hierarchically connected to those root nodes. In this study, the level of root nodes was defined as 'Level 1' (also known as the *Top Node*, illustrated in Fig. 1). The child families directly connected to the root nodes were grouped to 'Level 2'. Then, the families of 'Level 3' were defined by those child families directly connected to the families in 'Level 2'. The following *levels* can be therefore deduced in the same manner. Based on our comprehensive evaluation on all GO data, the bottom level of GO's hierarchical multi-label structure was 'Level 10', which had no child family and composed of the smallest number of proteins comparing with the families in other levels (Level 1 to Level 9). As demonstrated in Fig. 2, the average (green) and median (brown) numbers of proteins in the families of nine different levels (Level 2 to Level 10) were described. As illustrated, there was a clear descending trend of both average and median numbers from the top level (Level 2) to the bottom one (Level 10). Since the numbers of proteins in particular families indicated the representativeness of the corresponding families in the model construction, it was reasonable to say that the representativeness of a family gradually decreased with the penetration into a deeper level. Therefore, the representativeness of each GO family was classified into three groups (as shown in Fig. 2) based on the average and median numbers of proteins in nine different levels of GO structure, which included *Rare Classes* (average < 500 & median <150), *Medium Classes* (500 < average < 1000 & 150< median <300), and *Major Classes* (average > 1000 & median >300). In other words, based on such classification, it is feasible now to evaluate the performance of any protein function annotation tool/method.

To realize the model assessment, the CAFA data [64] were further collected. CAFA was designed to enable the large-scale assessment of bioinformatic methods dedicated to predicting the protein function in a time challenge manner. Here, the criteria applied for data processing and evaluation were based on CAFA3 [65]. Some popular metrics (F*max*, AUC and AUPRC) were then used to assess the constructed model. F*max* considered the precision and recall of the constructed models and was commonly used in protein function annotation [66]. A large AUC value represented the overall capacity of correctly predicting positive & negative samples [67]. AUPRC was a standard metric to evaluate classification performance by punishing false positives more than AUC, which resulting in being more frequently applied when high costs are required for obtaining labels [68]. The values of all the metrics (F*max*, AUC and AUPRC) ranged from 0 to 1. The larger the values were, the better the constructed model performed, and a value of 1 indicated the best performance as measured by any of these metrics.

2.2. Encoding of protein sequences using the one-hot strategy

As one of the most popular encoding methods to represent protein sequence, the one-hot strategy [69–77] was applied in this study to represent each protein. The dictionary of amino acids in this strategy equaled to 21 (20 for the common amino acids, and 1 additional for the remaining amino acids). Particularly, during the encoding, the corresponding position of a specific amino acid will be represented using number '1', and the remaining 20 positions will be set to number '0'. Thus, each amino acid in a sequence was represented by a 21-dimensional vector. Moreover, all protein sequences were encoded by their first 2000 amino acids, since over 99.5% of protein sequences from the well-established *Swiss-Prot* database [15] were with their sequence length less than 2000. The ZERO codes were added to proteins with insufficient length (<2000). As a result, all protein sequences could be finally represented by the matrix of 2000*21 dimensions.

2.3. Recurrent neural network (RNN) integrated in this work

In this study, a RNN method titled *gated recurrent unit* (GRU) was integrated, for the first time, with multi-kernel CNN method to facilitate the protein functional annotation, and the framework of this newly proposed deep learning strategy was explicitly illustrated in Fig. 3.

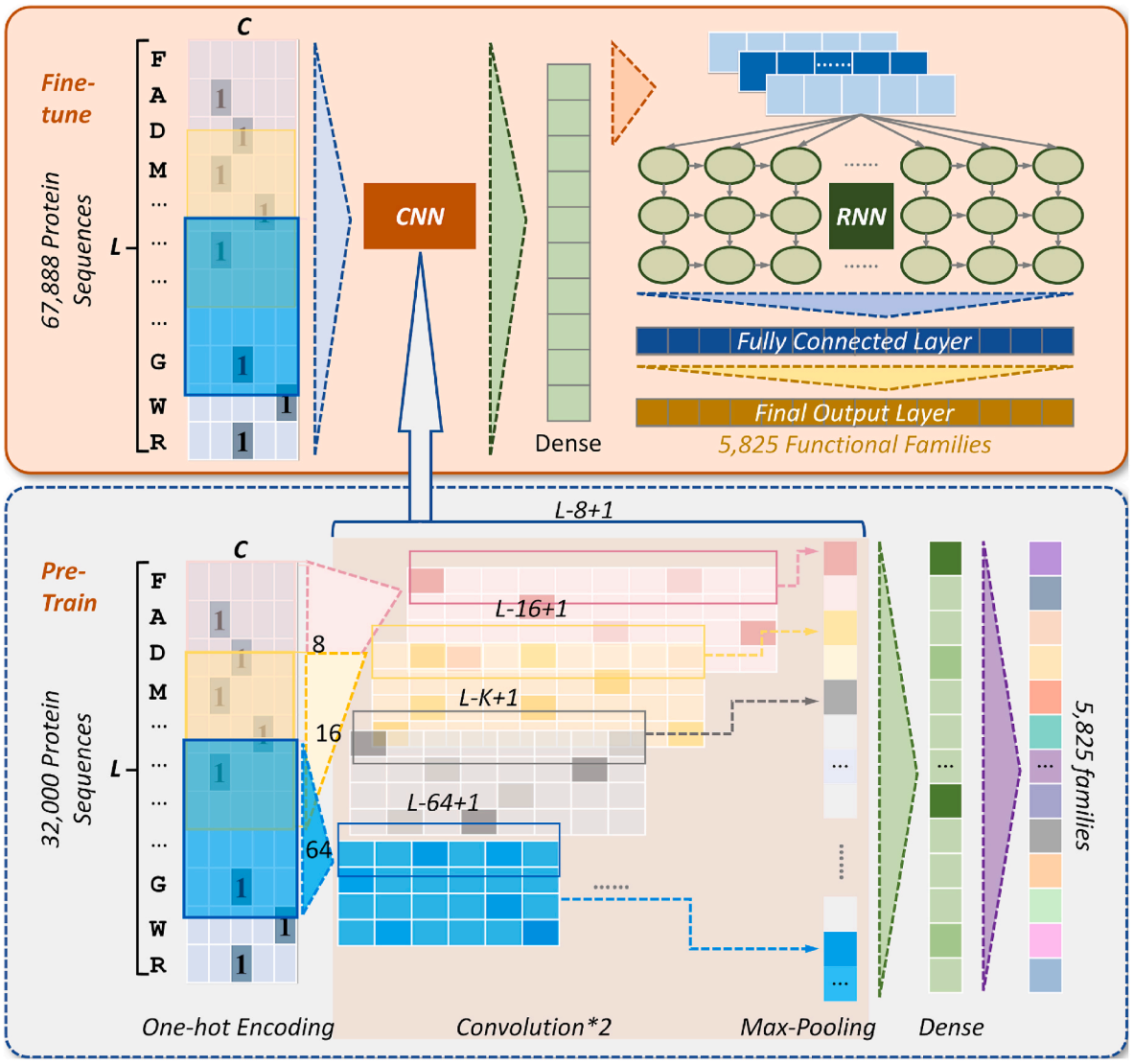

Fig. 3. Deep learning strategy proposed in this study for model construction. A RNN method was integrated with multi-kernel CNN method to facilitate the protein functional annotation. As shown, all (67,888) sequences were *first* represented by the one-hot strategy. *Second*, all encoded proteins were used as input to construct model based on multi-kernel CNN technique, which was fine-tuned by the pre-train process as illustrated at the bottom. *Third*, the output layer (dense) of the CNN model was further input to a RNN method, which led to a fully connected layer. *Finally*, to facilitate the comprehensive annotations of all (5,825) GO families, the dimension of the fully connected layer was reduced through setting the number of neural units to 5825. An output layer was finally developed to enable the annotation for all 5825 GO families.

For the CNN method used in both pre-train and fine-tune processes of Fig. 3, L indicated the length of the calculated protein sequence (L equaled to 2000 in this study), and m_k referred to the size of each kernel. There were 8 kernels ($k = 1, 2, \ldots, 8$), which led to 8 kernel sizes of $8 \times k$. Then, the length of output matrix after convolution can be indicated as $L - m_k + 1$ (Fig. 3). In the convolution process of CNN, the value of the i^{th} neuron ($h_{k,i}$) in those output matrix of the k^{th} kernel could therefore be shown by the following equation (Equation (1)):

$$h_{k,i} = ReLU\left(b_k + \sum_{l=1}^{m_k}\sum_{j=1}^{C} x_{i+l-1,j} w_{k,l,j}\right) \quad (1)$$

The nonlinear activation function applied in this study is ReLU (*rectified linear unit*). The bias parameter is indicated by b_k for the corresponding kernel k. The j indicated the j^{th} location of the amino acid dictionary and the C referred to the dictionary dimension (C equaled to 21 in this work). The l indicated the l^{th} location of the vertical dimension of the sliding window for different kernel, and the m_k gave the size of the kernel k. The weight parameter for kernel k in the $[j, l]$ location of a studied sliding window was shown by $w_{k,l,j}$. The value of the one-hot encoding matrix in the $[j, l]$ location of the studied sliding window was indicated using $x_{i+l-1,j}$. After convolution, an additional max-pooling layer was applied, which took the maximum value (h_k^{max}) among $L - m_k + 1$ neurons for kernel k using the following Equation (2).

$$h_k^{max} = \max\left(h_{k,i}\right) \ (i = 1, 2, \cdots, L - m_k + 1) \quad (2)$$

The output vectors $[h_1^{max}, h_2^{max}, \ldots\ldots, h_k^{max}]$ after max-pooling generated by different kernels ($k = 1, 2, \ldots, 8$) were then concatenated (Equation (3)) and processed using a batch regularization. Moreover, a random dropout strategy was added to prevent over-fitting of the model.

$$s = concat\left(\left[h_1^{max}, h_2^{max}, \ldots\ldots, h_k^{max}\right]\right) \quad (3)$$

Finally, the vector s was input into the fully connected layer for dimensionality reduction. The output (a_i) of i^{th} neuron in the fully connected layer could be represented by Equation (4).

$$a_i = Sigmoid\left(b_i + \sum_{j=1}^{n} s*w_{ij}\right) \quad (4)$$

The nonlinear activation function applied here is Sigmoid [78]. The bias parameter is indicated by b_i for the corresponding neuron i (i is from 1 to 5,825, denoting all 5825 families in GO database). The j indicated the j^{th} location of the dense layer (Fig. 3), and the n provided the entire length of the dense layer. The weight parameter between the location j of dense layer and the location i of fully connected layer was described using w_{ij}.

2.4. Convolutional neural network (CNN) applied in this work

The *recurrent neural network* (RNN) has been successfully applied to process sequence data and had a wide range of applications in various fields including the processing of medical image [79]. natural language [80], and so on [81–85]. Its implementation process is similar to a *Markov* chain, and it predicts *current* state based on *previous* state [86]. Particularly, the neurons in current state is used as the reference input of the neurons in the next state, thus it can be applied to the problem of serial correlation [86]. RNN is well known for its capacity of encoding contextual information in the sequential data, and it only requires the limited number of network parameters. Particularly, the RNN method applied in this study was the *gated recurrent unit* (GRU), which added a gating mechanism (with two gates of "*reset*" and "*update*") to make decision on how much information of the current step will be transferred to the next one [87].

3. Results and discussion

3.1. Deep learning strategy proposed for model construction

As reported, both convolutional (CNN) and deep (DNN) neural networks were applied to predict protein function, such as DeepGO [53], DeepGOCNN [54] and DeepGOPlus [54]. The recurrent neural network (RNN) was reported as very powerful in dealing with the time-series problem of sequential process, and it was therefore expected to elevate the annotation performances for '*rare classes*' [88]. In this study, a RNN method titled *gated recurrent unit* (GRU) was integrated, for the first time, with multi-kernel CNN method to facilitate the protein functional annotation, and the framework of this newly proposed deep learning strategy was explicitly illustrated in Fig. 3. As shown, all (67,888) protein sequences were *first* represented based on the one-hot strategy. *Second*, all encoded proteins were used as input to construct a model based on multi-kernel CNN technique, which was fine-tuned using the pre-train process as illustrated at the bottom of Fig. 3. *Third*, the output layer (dense) of the CNN model was further input to the *gated recurrent unit* (GRU), which led to a fully connected layer. *Finally*, to facilitate the comprehensive annotations of all (5,825) GO families, the dimension of the fully connected layer was reduced by setting the number of neural units to 5825. The output layer was finally developed to enable annotation for all 5825 GO families. The detailed information of this framework was elaborated below.

(a) *Application of the Convolutional Neural Network*

For the pre-train process (as shown in Figs. 3) and 2 convolutional layers and 8 kernels of different sizes were used for feature extraction. In the one-dimensional convolution processing of the input matrix, a total of 256 filters were set for each convolution kernel. Then, after the one-dimensional maximum pooling layer, each kernel will result in a vector of 256 dimensions. By concatenating the vectors resulted from eight kernels, a dense layer (illustrated at the bottom of Fig. 3) with a length of 2048 were generated. Based on the resulting dense layer, a new output layer of 5825 dimensions were finally produced by setting the number of output neural unit to 5825.

(b) *Integration of the Recurrent Neural Network*

For the specific research direction discussed in this work, a protein's function is described by the hierarchical multi-label structure of Gene Ontology [60]. Starting from the ontologies of the root nodes, thousands of leaf nodes were hierarchically arranged into a directed acyclic graph (DAG). In this DAG, the annotations of the child nodes can be referred back to their parent nodes, which is very similar to the state transition process of RNN method. In other words, since the RNN was reported to be powerful in dealing with the time-series problem of sequential process [88], it was highly anticipated to elevate the annotation performance for the proteins within the '*rare classes*'. Therefore, the RNN method (particularly, the GRU) was integrated, for the first time, with multi-kernel CNN method to facilitate the protein functional annotation in this study.

(c)Model Construction Based on Transfer Learning

To further elevate the performance of protein functional annotation, the state-of-the-art technique of transfer learning (TL) was further introduced to this study for model construction. Particularly, the adoption of TL was expected to prevent the disappearance of gradients and the overfitting of the model on the multi-label problem of protein functional annotation. In this study, based on all 67,888 protein sequences with explicit GO terms collected in this study, the half of the sequences (32,000) were *first* selected to construct a CNN-based pre-train model based on a similar random selection strategy of the previous study [89]. *Second*, the optimized parameters (such as *bias* and *weight*) of the pre-train process were fed into the CNN in fine-tune process (orange box in Fig. 3) as the starting point for further tuning the parameter. *Third*, all the collected protein sequences were used to construct CNN model based on those optimized parameters, which helped to extract the common characteristics of the studied proteins [90]. *Finally*, these characteristics (the dense layer in the upper section of Fig. 3) were used to construct a RNN model, which were followed by a fully connected layer and a final output layer. As a result, the resulting model could annotate the protein function of 5825 families, which, to the best of our knowledge, was one of the models covering the largest number of GO families.

(d)Environment Setup for Realizing the Proposed Strategy

To realize the novel strategy proposed above, the computational environment was systematically configured. Particularly, the program of this strategy was written using *Python* language, and the *Keras* framework implemented in *Tensorflow* 2.1 was used. Batch normalization technology [91] was added after the convolutional layer to prevent over-fitting and accelerate the convergence of the model. Dropout technique [92] was added to the fully connected layer for randomly removing certain number of neurons at each training step to prevent the overfitting of neural networks. The *Adam* optimizer [93] was adopted for the optimization during back-propagation, and the learning rate was set to 0.0005. The batch size was set to 64, and the early stopping technology was used to avoid the overfitting problem. The loss function used here was the binary cross-entropy (BCE) [94], which was used to calculate the difference between the true value and the predicted one.

3.2. Comparing the overall performance with existing tools

In this study, a novel protein functional annotation strategy that integrated multiple deep learning methods (PFmulDL) was constructed by **1**) integrating RNN with multi-kernel CNN method, **2**) introducing transfer learning, and **3**) enabling functional prediction for the largest number of GO families. To assess the prediction performance of this

Table 1
Performance comparison among existing tools popular in protein functional annotations and the novel strategy PFmulDL proposed in this study based on the CAFA3 benchmark dataset. These existing tools included two *sequence similarity*-based (*SS*-based, *BLAST* [64] and *GoFDR* [34]) and another three *machine learning*-based (*ML*-based, *DeepGO* [53], *DeepGOCNN* [54] & *DeepGOPlus* [54]) tools. The prediction performances of the studied tools were compared using the benchmark dataset CAFA3 of three types: *molecular function* (MF), *biological process* (BP), and *cellular component* (CC), and the F*max* and AUPRC were used as the assessing metrics. All the best performing values among existing tools were highlighted by the double underlines, and all the worst performing values among existing tools were highlighted using the wavy lines. The best performing values among all studied tools were highlighted using bold font.

Method		BP		CC		MF	
		F*max*	AUPRC	F*max*	AUPRC	F*max*	AUPRC
PFmulDL		0.459	0.452	0.677	0.729	0.508	0.509
SS-based	*BLAST* [64]	0.262	0.071	0.513	0.311	0.435	0.263
	GoFDR [34]	0.193	0.183	0.413	0.357	0.475	0.424
ML-based	*DeepGO* [53]	0.362	0.213	0.502	0.446	0.392	0.312
	DeepGOCNN [54]	0.388	0.213	0.582	0.523	0.411	0.402
	DeepGOPlus [54]	0.393	0.346	0.640	0.670	0.432	0.396

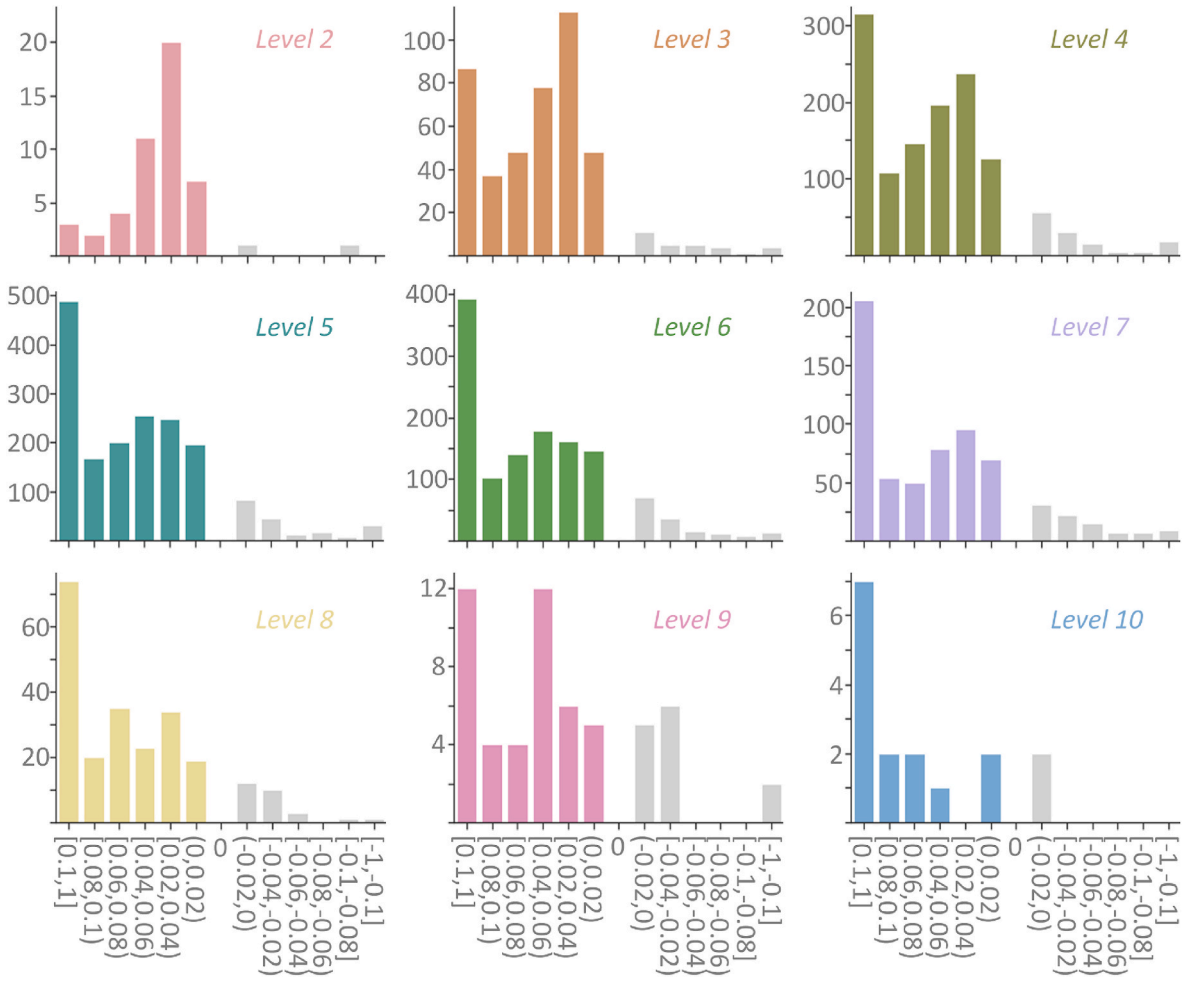

Fig. 4. Performance comparison between *DeepGOPlus* and PFmulDL using benchmark data of CAFA3 and AUC. The degree of PFmulDL's performance elevations from *DeepGOPlus* were provided, and all levels (Level 2 to 10) of GO's hierarchical structure were assessed. The *x*-axis indicated the degree of PFmulDL's performance elevations from *DeepGOPlus* (with the positive value denoting the performance elevation and the negative value demonstrating the performance reduction). The *y*-axis indicated the number of GO families. The PFmulDL showed the improved performances for the vast majority (~88.7%) of GO families when comparing with *DeepGOPlus*.

newly developed PFmulDL, the prediction performance of several existing tools popular in protein functional annotations was assessed and compared with that of PFmulDL based on the CAFA3 benchmark dataset (described in the "Data Collection for Model Construction and Assessment" of the **Materials and Methods**). As shown in Table 1, two typical *sequence similarity*-based (*SS*-based, BLAST [64] and GoFDR [34]) and another three well-established *machine learning*-based (*ML*-based, DeepGO [53], DeepGOCNN [54] and DeepGOPlus [54]) were considered. Their prediction performance and the performance of PFmulDL were compared using the same benchmark dataset CAFA3. Particularly, there were three types of CAFA3: molecular function (MF), biological process (BP) and cellular component (CC), and the F*max* and AUPRC were adopted as the assessing metrics. As provided in Table 1, among the five existing tools, the *DeepGOPlus* gave the highest value of

Fig. 5. Comparing the overall performances between PFmulDL and *DeepGOPlus*. The overall performances were represented using AUC values in predicting CAFA3 data, and the prediction performances of PFmulDL and *DeepGOPlus* were colored in red and green, respectively. In the '*major classes*' (Level 2 and 3 as defined in Fig. 2), the performance of PFmulDL was slightly better than that of *DeepGOPlus* (~1.2% increase). For '*medium classes*' (Level 4 and 5 in Fig. 2), the performance of PFmulDL was moderately better than that of *DeepGOPlus* (~+2.2%). For '*rare classes*' (Level 6 to 10 shown in Fig. 2), the performance of PFmulDL was significantly elevated by ~5.0% from that of *DeepGOPlus*. The PFmulDL was discovered able to significantly elevate the annotation performances for the proteins in '*rare classes*' without sacrificing that for the '*major ones*'.

both F*max* and AUPRC under two CAFA3 data types (BP and CC), and the *GoFDR* resulted in the highest values of both F*max* and AUPRC under the data type of MF (all these best performing values were highlighted by the double underlines in Table 1). Moreover, the *GoFDR* and *BLAST* performed the worst in two data types (BP and CC) as assessed using F*max* and AUPRC, respectively, and the *DeepGO* and *BLAST* performed the worst in MF as assessed by F*max* and AUPRC, respectively (all these worst performing values were highlighted using the wavy line in Table 1). On the one hand, this result indicated that it was difficult to find a tool with the consistently best or worst performances. On the other hand, the *DeepGOPlus* demonstrated a better performance in most data types when comparing with other existing tools.

The same benchmark dataset CAFA3 as that used by existing tools was also adopted to evaluate the performance of PFmulDL. As shown in Table 1, the newly constructed PFmulDL performed the consistently best performances for all data types (BP, CC and MF) as assessed by all metrics (F*max* and AUPRC). Its performances were highlighted by bold fonts in Table 1. Moreover, the percentages of performance enhancements comparing to the best-performing existing tools were also provided in the bracket of Table 1. Particularly, PFmulDL was compared with *DeepGOPlus* and *GoFDR*, and the percentages of performance enhancements from these two well-performing tools varied from +5.8% to 30.6%, which indicated a dramatical elevation in the performance of protein function prediction by the deep learning strategy proposed in this study.

3.3. Family-based performance comparison with DeepGOPlus

Based on the above assessment of the overall performance on CAFA3 benchmark, *DeepGOPlus* was found to perform the best in most data types among all existing tools, which inspired us to conduct in-depth evaluation on the performance of each GO family. In other words, all those GO families in CAFA3 were predicted by *DeepGOPlus* & PFmulDL, and the degree of performance enhancements between these tools at the family level were explicitly analyzed. As demonstrated in Fig. 4, the degree of PFmulDL's performance elevations from *DeepGOPlus* were provided, and all levels (from Level 2 to Level 10) of GO's hierarchical multi-label structure were assessed. The *x*-axis indicated the degree of PFmulDL's performance elevations from *DeepGOPlus* (with the positive value denoting the performance elevation and the negative value demonstrating the performance reduction). The *y*-axis indicated the number of GO families. As described in Figs. 4 and 95.9%, 93.2%, 89.6%, 89.0%, 87.9%, 85.9%, 88.4%, 76.8%, and 87.5% of GO families were found to be predicted with better accuracy (as assessed using AUC values) by the PFmulDL than *DeepGOPlus* in Level 2, Level 3, Level 4, Level 5, Level 6, Level 7, Level 8, Level 9, and Level 10, respectively. In other words, the PFmulDL demonstrated improved performances for the vast majority (88.7%) of GO families when comparing with *DeepGOPlus*.

Moreover, the percentages of families with >10% elevation in annotation performances (Fig. 4) were 6.1%, 20.0%, 25.0%, 28.0%, 30.8%, 31.9%, 31.9%, 21.4%, and 43.8% in Level 2, Level 3, Level 4, Level 5, Level 6, Level 7, Level 8, Level 9, and Level 10, respectively. It is clear that these data illustrate an obvious ascending trend from Level 2 to 10, which reminds us to evaluate the overall performances of the proposed strategy on predicting the functions of proteins in those '*rare classes*' that were defined as from level 6 to level 10 in Fig. 2.

Therefore, the comparison of the overall performances between PFmulDL and *DeepGOPlus* was also conducted. As shown in Fig. 5, the overall performances were represented using the AUC value in predicting CAFA3 data, and the predictive performances of PFmulDL and *DeepGOPlus* were colored in red and green, respectively. For the 'major classes' (Level 2 and 3 in Fig. 2), the performance of PFmulDL was slightly better than that of *DeepGOPlus* (about 1.2% increase). For 'medium classes' (Level 4 and 5 in Fig. 2), the performance of PFmulDL was also slightly better than that of *DeepGOPlus* (about 2.2% increase). For 'rare classes' (Level 6 to 10 in Fig. 2), the performance of PFmulDL was significantly elevated by ~5.0% from that of *DeepGOPlus*. Taking the Level 10 as examples, it showed the largest performance increase from *DeepGOPlus* to PFmulDL (about 7.0%). In conclusion, based on the result of Fig. 5, the PFmulDL proposed in this study was discovered to be capable of significantly elevating the annotation performances for the proteins in '*rare classes*' without sacrificing that for the '*major ones*'.

To test whether the PFmulDL can accurately predict the function of specific protein, an exemplar protein (*Escherichia coli* L-asparaginase 1, which has not been included into model construction) was annotated by both PFmulDL and *DeepGOPlus*. As a result, PFmulDL annotated this protein to 29 GO terms, 24 (~82.8%) out of which were consistent with the annotation label of the Gene Ontology. However, *DeepGOPlus* predicted that protein to 44 GO terms, only 19 (~43.2%) out of which were consistent with the annotation label of the Gene Ontology. Therefore, this further demonstrated the good annotation performance of PFmulDL proposed in this study.

4. Conclusions

In this study, a novel protein functional annotation strategy that integrated multiple deep learning methods (PFmulDL) was constructed by **1**) integrating RNN with multi-kernel CNN method, **2**) introducing transfer learning, and **3**) enabling functional prediction for the largest number of GO families. Based on a systematical comparison with some existing tools popular in current protein functional annotation, this strategy was found as capable of significantly elevating the annotation performance for '*rare classes*' without sacrificing that for the '*major ones*'. All in all, due to the emerging requirement on improving the annotation performance for proteins in *rare classes*, the PFmulDL proposed in this study would become an essential complement to those available tools in the fields of drug target discovery [95,96], drug transportation [97,98], drug metabolism [99], OMICS [100–102], protein-relevant interactions [103–106], and so on [107–109]. The model and source codes are freely available and open to all users at: https://github.com/idrblab/PFmulDL.

Funding

Funded by Natural Science Foundation of Zhejiang Province (LR21H300001); National Natural Science Foundation of China (81872798 & U1909208); Leading Talent of the 'Ten Thousand Plan' - National High-Level Talents Special Support Plan of China; Fundamental Research Fund for Central Universities (2018QNA7023); "Double Top-Class" University Project (181201*194232101); Key R&D Program of Zhejiang Province (2020C03010). This work was supported by Westlake Laboratory (Westlake Laboratory of Life Sciences and Biomedicine); Alibaba-Zhejiang University Joint Research Center of Future Digital Healthcare; Alibaba Cloud; Information Technology Center of Zhejiang University.

Declaration of competing interest

The Authors Declare That There is No Conflict of Interest.

References

[1] A.C. Goldstrohm, T.M.T. Hall, K.M. McKenney, Post-transcriptional regulatory functions of mammalian pumilio proteins, Trends Genet. 34 (2018) 972–990.
[2] J. Hong, Y. Luo, Y. Zhang, et al., Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning, Brief. Bioinform. 21 (2020) 1437–1447.
[3] Y.H. Li, J.Y. Xu, L. Tao, et al., SVM-Prot 2016: a web-server for machine learning prediction of protein functional families from sequence irrespective of similarity, PLoS One 11 (2016), e0155290.
[4] X. Wang, F. Li, W. Qiu, et al., SYNBIP: synthetic binding proteins for research, diagnosis and therapy, Nucleic Acids Res. 50 (2022) D560–D570.
[5] Y. Ma, Y. Yokota, Y. Kozutsumi, et al., Structural and functional roles of the amino-terminal region and collagen-like domain of human serum mannan-binding protein, Biochem. Mol. Biol. Int. 40 (1996) 965–974.
[6] W. Xue, F. Yang, P. Wang, et al., 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. 9 (2018) 1128–1140.
[7] J. Tang, J. Fu, Y. Wang, et al., Simultaneous improvement in the precision, accuracy, and robustness of label-free proteome quantification by optimizing data manipulation chains, Mol. Cell. Proteomics 18 (2019) 1683–1699.
[8] J. Tang, J. Fu, Y. Wang, et al., ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies, Brief. Bioinform. 21 (2020) 621–636.
[9] A. Sureyya Rifaioglu, T. Dogan, M. Jesus Martin, et al., DEEPred: automated protein function prediction with multi-task feed-forward deep neural networks, Sci. Rep. 9 (2019) 7344.
[10] A.K. Sharma, R. Srivastava, Protein secondary structure prediction using character bi-gram embedding and Bi-LSTM, Curr. Bioinform. 16 (2021) 333–338.
[11] C. Ao, L. Yu, Q. Zou, Prediction of bio-sequence modifications and the associations with diseases, Brief. Funct. Genomics 20 (2021) 1–18.
[12] S. Das, C.A. Orengo, Protein function annotation using protein domain family resources, Methods 93 (2016) 24–34.
[13] Y. Cai, J. Wang, L. Deng, SDN2GO: an integrated deep learning model for protein function prediction, Front. Bioeng. Biotechnol. 8 (2020) 391.
[14] B. Zhao, S. Hu, X. Li, et al., An efficient method for protein function annotation based on multilayer protein networks, Hum. Genom. 10 (2016) 33.
[15] C. UniProt, UniProt: a worldwide hub of protein knowledge, Nucleic Acids Res. 47 (2019) D506–D515.
[16] C. Sun, Y. Feng, Identification of disordered regions of intrinsically disordered proteins by multi-features fusion, Curr. Bioinform. 16 (2021) 1126–1132.
[17] J.R. Hamre, D.K. Klimov, M.D. McCoy, et al., Machine learning-based prediction of drug and ligand binding in BCL-2 variants through molecular dynamics, Comput. Biol. Med. 140 (2022) 105060.
[18] W. Xue, P. Wang, G. Tu, et al., Computational identification of the binding mechanism of a triple reuptake inhibitor amitifadine for the treatment of major depressive disorder, Phys. Chem. Chem. Phys. 20 (2018) 6606–6616.
[19] R. You, X. Huang, S. Zhu, DeepText2GO: improving large-scale protein function prediction with deep semantic text representation, Methods 145 (2018) 82–90.
[20] Q. Yang, B. Li, S. Chen, et al., MMEASE: online meta-analysis of metabolomic data by enhanced metabolite annotation, marker selection and enrichment analysis, J. Proteomics 232 (2021) 104023.
[21] J. Fu, J. Tang, Y. Wang, et al., Discovery of the consistently well-performed analysis chain for SWATH-MS based pharmacoproteomic quantification, Front. Pharmacol. 9 (2018) 681.
[22] C. Zhang, P.L. Freddolino, Y. Zhang, COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information, Nucleic Acids Res. 45 (2017) W291. W9.
[23] V. Gligorijevic, P.D. Renfrew, T. Kosciolek, et al., Structure-based protein function prediction using graph convolutional networks, Nat. Commun. 12 (2021) 3168.
[24] A. Ranjan, M.S. Fahad, D. Fernandez-Baca, et al., Deep robust framework for protein function prediction using variable-length protein sequences, IEEE/ACM Trans. Comput. Biol. Bioinform. 17 (2020) 1648–1659.
[25] Z. Du, Y. He, J. Li, et al., DeepAdd: protein function prediction from k-mer embedding and additional features, Comput. Biol. Chem. 89 (2020) 107379.
[26] J. Hong, Y. Luo, M. Mou, et al., Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery, Brief. Bioinform. 21 (2020) 1825–1836.
[27] S.J. Giri, P. Dutta, P. Halani, et al., MultiPredGO: deep multi-modal protein function prediction by amalgamating protein structure, sequence, and interaction information, IEEE J. Biomed. Health Inform. 25 (2021) 1832–1838.
[28] W.R. Pearson, Protein function prediction: problems and pitfalls, Curr. Protoc. Bioinformatics 51 (2015) 4121–4128.
[29] Z. Basharat, U. Akhtar, K. Khan, et al., Differential analysis of orientia tsutsugamushi genomes for therapeutic target identification and possible intervention through natural product inhibitor screening, Comput. Biol. Med. 141 (2022) 105165.
[30] Q. Yang, B. Li, J. Tang, et al., Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data, Brief. Bioinform. 21 (2020) 1058–1068.
[31] F. Zhu, X.X. Li, S.Y. Yang, et al., Clinical success of drug targets prospectively predicted by in silico study, Trends Pharmacol. Sci. 39 (2018) 229–231.
[32] Y.H. Li, X.X. Li, J.J. Hong, et al., Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs, Brief. Bioinform. 21 (2020) 649–662.
[33] S.F. Altschul, T.L. Madden, A.A. Schaffer, et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[34] Q. Gong, W. Ning, W. Tian, GoFDR: a sequence alignment based method for predicting protein functions, Methods 93 (2016) 3–14.
[35] C. Zhang, W. Zheng, P.L. Freddolino, et al., MetaGO: predicting gene ontology of non-homologous proteins through low-resolution protein structure prediction and protein-protein network mapping, J. Mol. Biol. 430 (2018) 2256–2265.
[36] A. Zielezinski, S. Vinga, J. Almeida, et al., Alignment-free sequence comparison: benefits, applications, and tools, Genome Biol. 18 (2017) 186.
[37] H. Seligmann, Alignment-based and alignment-free methods converge with experimental data on amino acids coded by stop codons at split between nuclear and mitochondrial genetic codes, Biosystems 167 (2018) 33–46.
[38] S. Seo, M. Oh, Y. Park, et al., DeepFam: deep learning based alignment-free method for protein family modeling and prediction, Bioinformatics 34 (2018) 254–262.
[39] S. Uddin, A. Khan, M.E. Hossain, et al., Comparing different supervised machine learning algorithms for disease prediction, BMC Med. Inform. Decis. Mak. 19 (2019) 281.
[40] E. Ebrahimie, F. Zamansani, I.O. Alanazi, et al., Advances in understanding the specificity function of transporters by machine learning, Comput. Biol. Med. 138 (2021) 104893.
[41] J.M. Cunningham, G. Koytiger, P.K. Sorger, et al., Biophysical prediction of protein-peptide interactions and signaling networks using machine learning, Nat. Methods 17 (2020) 175–183.
[42] N. Ikram, M.A. Qadir, M.T. Afzal, SimExact - an efficient method to compute function similarity between proteins using gene ontology, Curr. Bioinform. 15 (2020) 318–327.
[43] J. Fu, Y. Zhang, Y. Wang, et al., Optimization of metabolomic data processing using NOREVA, Nat. Protoc. 17 (2022) 129–151.
[44] B. Li, J. Tang, Q. Yang, et al., NOREVA: normalization and evaluation of MS-based metabolomics data, Nucleic Acids Res. 45 (2017) W162–W170.
[45] Q. Yang, Y. Wang, Y. Zhang, et al., NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data, Nucleic Acids Res. 48 (2020) W436–W448.
[46] F. Li, Y. Zhou, X. Zhang, et al., SSizer: determining the sample sufficiency for comparative biological study, J. Mol. Biol. 432 (2020) 3411–3421.
[47] Q. Yang, J. Hong, Y. Li, et al., A novel bioinformatics approach to identify the consistently well-performing normalization strategy for current metabolomic studies, Brief. Bioinform. 21 (2020) 2142–2152.
[48] J. Hu, H.L. Chen, A.A. Heidari, et al., Orthogonal learning covariance matrix for defects of grey wolf optimizer: insights, balance, diversity, and feature selection, Knowl-Based Syst. 213 (2021) 106684.
[49] Y.N. Zhang, R.J. Liu, X. Wang, et al., Boosted binary harris hawks optimizer and feature selection, Eng. Comput. 37 (2021) 3741–3770.
[50] Y.N. Zhang, R.J. Liu, A.A. Heidari, et al., Towards augmented kernel extreme learning models for bankruptcy prediction: algorithmic behavior and comprehensive analysis, Neurocomputing 430 (2021) 185–212.
[51] M.R. Bakhtiarizadeh, M. Moradi-Shahrbabak, M. Ebrahimi, et al., Neural network and SVM classifiers accurately predict lipid binding proteins, irrespective of sequence homology, J. Theor. Biol. 356 (2014) 213–222.
[52] L. Patel, T. Shukla, X. Huang, et al., Machine learning methods in drug discovery, Molecules 25 (2020) 5277.
[53] M. Kulmanov, M.A. Khan, R. Hoehndorf, et al., DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier, Bioinformatics 34 (2018) 660–668.
[54] M. Kulmanov, R. Hoehndorf, DeepGOPlus: improved protein function prediction from sequence, Bioinformatics 36 (2020) 422–429.
[55] Q. Shi, W. Chen, S. Huang, et al., Deep learning for mining protein data, Brief. Bioinform. 22 (2021) 194–218.

[56] K. De Angeli, S. Gao, I. Danciu, et al., Class imbalance in out-of-distribution datasets: improving the robustness of the TextCNN for the classification of rare cancer types, J. Biomed. Inform. 125 (2022) 103957.
[57] Z. Ma, J.E. Tanis, A. Taruno, et al., Calcium homeostasis modulator (CALHM) ion channels, Pflugers Arch 468 (2016) 395–403.
[58] J. Traba, J. Satrustegui, A. del Arco, Adenine nucleotide transporters in organelles: novel genes and functions, Cell. Mol. Life Sci. 68 (2011) 1183–1206.
[59] S. Niekamp, N. Coudray, N. Zhang, et al., Coupling of ATPase activity, microtubule binding, and mechanics in the dynein motor domain, EMBO J. 38 (2019), e101414.
[60] C. Gene Ontology, The gene ontology resource: 20 years and still going strong, Nucleic Acids Res. 47 (2019) D330–D338.
[61] S.B. Zhang, Q.R. Tang, Protein-protein interaction inference based on semantic similarity of gene ontology terms, J. Theor. Biol. 401 (2016) 30–37.
[62] R. Cerri, R.C. Barros, Y. Jin, Reduction strategies for hierarchical multi-label classification in protein function prediction, BMC Bioinformatics 17 (2016) 373.
[63] F.K. Nakano, M. Lietaert, C. Vens, Machine learning for discovering missing or wrong protein function annotations: a comparison using updated benchmark datasets, BMC Bioinformatics 20 (2019) 485.
[64] N. Zhou, Y. Jiang, T.R. Bergquist, et al., The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens, Genome Biol. 20 (2019) 244.
[65] Y. Jiang, T.R. Oron, W.T. Clark, et al., An expanded evaluation of protein function prediction methods shows an improvement in accuracy, Genome Biol. 17 (2016) 184.
[66] F. Zhang, H. Song, M. Zeng, et al., DeepFunc: a deep learning framework for accurate prediction of protein functions from protein sequences and interactions, Proteomics 19 (2019), e1900019.
[67] Y. Wang, Y. Yang, S. Chen, et al., DeepDRK: a deep learning framework for drug repurposing through kernel-based multi-omics integration, Brief. Bioinform. 22 (2021) bbab048.
[68] R. You, Z. Zhang, Y. Xiong, et al., GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank, Bioinformatics 34 (2018) 2465–2473.
[69] X. Tang, Y. Sun, Fast and accurate microRNA search using CNN, BMC Bioinformatics 20 (2019) 646.
[70] L. Guo, S. Wang, M. Li, et al., Accurate classification of membrane protein types based on sequence and evolutionary information using deep learning, BMC Bioinformatics 20 (2019) 700.
[71] M. Niu, Y. Lin, Q. Zou, sgRNACNN: identifying sgRNA on-target activity in four crops using ensembles of convolutional neural networks, Plant Mol. Biol. 105 (2021) 483–495.
[72] S. Akbar, A. Ahmad, M. Hayat, et al., iAtbP-Hyb-EnC: prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model, Comput. Biol. Med. 137 (2021) 104778.
[73] Y. Fan, H. Xu, Prediction of off-target effects in CRISPR/Cas9 system by ensemble learning, Curr. Bioinform. 16 (2021) 1169–1178.
[74] Y. Zhang, J.B. Ying, J.J. Hong, et al., How does chirality determine the selective inhibition of histone deacetylase 6? A lesson from trichostatin A enantiomers based on molecular dynamics, ACS Chem. Neurosci. 10 (2019) 2467–2480.
[75] J. Li, Z. Li, J. Luo, et al., ACNNT3: attention-CNN framework for prediction of sequence-based bacterial type III secreted effectors, Comput. Math. Meth. Med. 2020 (2020) 3974598.
[76] T. Fu, G. Zheng, G. Tu, et al., Exploring the binding mechanism of metabotropic glutamate receptor 5 negative allosteric modulators in clinical trials by molecular dynamics simulations, ACS Chem. Neurosci. 9 (2018) 1492–1502.
[77] S. Zhang, K. Amahong, C. Zhang, et al., RNA-RNA interactions between SARS-CoV-2 and host benefit viral development and evolution during COVID-19 infection, Brief. Bioinform. 23 (2022) bbab397.
[78] V.B. Sinha, S. Kudugunta, A.R. Sankar, et al., DANTE: deep alternations for training neural networks, Neural Network. 131 (2020) 127–143.
[79] Y. Xu, A. Hosny, R. Zeleznik, et al., Deep learning predicts lung cancer treatment response from serial medical imaging, Clin. Cancer Res. 25 (2019) 3266–3275.
[80] C. Lyu, B. Chen, Y. Ren, et al., Long short-term memory RNN for biomedical named entity recognition, BMC Bioinformatics 18 (2017) 462.
[81] H.R. Yun, G. Lee, M.J. Jeon, et al., Erythropoiesis stimulating agent recommendation model using recurrent neural networks for patient with kidney failure with replacement therapy, Comput. Biol. Med. 137 (2021) 104718.
[82] S. Naseer, W. Hussain, Y.D. Khan, et al., NPalmitoylDeep-pseaac: a predictor of N-palmitoylation sites in proteins using deep representations of proteins and PseAAC via modified 5-steps rule, Curr. Bioinform. 16 (2021) 294–305.
[83] W. Liang, K. Zhang, P. Cao, et al., Rethinking modeling Alzheimer's disease progression from a multi-task learning perspective with deep recurrent neural network, Comput. Biol. Med. 138 (2021) 104935.
[84] W. Xue, T. Fu, S. Deng, et al., Molecular mechanism for the allosteric inhibition of the human serotonin transporter by antidepressant escitalopram, ACS Chem. Neurosci. 13 (2022) 340–351.
[85] F. Zhu, C. Qin, L. Tao, et al., Clustered patterns of species origins of nature-derived drugs and clues for future bioprospecting, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 12943–12948.
[86] M. Geravanchizadeh, H. Roushan, Dynamic selective auditory attention detection using RNN and reinforcement learning, Sci. Rep. 11 (2021) 15497.
[87] B. Zhang, D. Xiong, J. Xie, et al., Neural machine translation with GRU-gated attention model, IEEE Trans. Neural Netw. Learn. Syst. 31 (2020) 4688–4698.
[88] J. Lv, C. Wang, W. Gao, et al., An economic forecasting method based on the LightGBM-optimized LSTM and time-series model, Comput. Intell. Neurosci. 2021 (2021) 8128879.
[89] Z. Chai, H. Jin, S. Shi, et al., Hierarchical shared transfer learning for biomedical named entity recognition, BMC Bioinformatics 23 (2022) 8.
[90] T. Han, C. Liu, W. Yang, et al., Learning transferable features in deep convolutional neural networks for diagnosing unseen machine conditions, ISA Trans. 93 (2019) 341–353.
[91] S.H. Wang, C. Tang, J. Sun, et al., Multiple sclerosis identification by 14-layer convolutional neural network with batch normalization, dropout, and stochastic pooling, Front. Neurosci. 12 (2018) 818.
[92] X. Shen, X. Tian, T. Liu, et al., Continuous dropout, IEEE Trans. Neural Netw. Learn. Syst. 29 (2018) 3926–3937.
[93] S.R. Dubey, S. Chakraborty, S.K. Roy, et al., DiffGrad: an optimization method for convolutional neural networks, IEEE Trans. Neural Netw. Learn. Syst. 31 (2020) 4500–4511.
[94] S.G. Zadeh, M. Schmid, Bias in cross-entropy-based training of deep survival networks, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2021) 3126–3137.
[95] F. Zhu, Z. Shi, C. Qin, et al., Therapeutic target database update 2012: a resource for facilitating target-oriented drug discovery, Nucleic Acids Res. 40 (2012) D1128–D1136.
[96] H. Yang, C. Qin, Y.H. Li, et al., Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information, Nucleic Acids Res. 44 (2016) D1069–D1074.
[97] J. Yin, W. Sun, F. Li, et al., VARIDT 1.0: variability of drug transporter database, Nucleic Acids Res. 48 (2020) D1042–D1050.
[98] T. Fu, F. Li, Y. Zhang, et al., VARIDT 2.0: structural variability of drug transporter, Nucleic Acids Res. 50 (2022) D1417–D1431.
[99] J. Yin, F. Li, Y. Zhou, et al., INTEDE: interactome of drug-metabolizing enzymes, Nucleic Acids Res. 49 (2021) D1233–D1243.
[100] J. Tang, M. Mou, Y. Wang, et al., MetaFS: performance assessment of biomarker discovery in metaproteomics, Brief. Bioinform. 22 (2021) bbaa105.
[101] J. Fu, Y. Zhang, J. Liu, et al., Pharmacometabonomics: data processing and statistical analysis, Brief. Bioinform. 22 (2021) bbab138.
[102] S. Zhang, K. Amahong, X. Sun, et al., The miRNA: a small but powerful RNA for COVID-19, Brief. Bioinform. 22 (2021) 1137–1149.
[103] Y. Zhang, Q.C. Zheng, In silico analysis revealed a unique binding but ineffective mode of amantadine to influenza virus B M2 channel, J. Phys. Chem. Lett. 12 (2021) 1169–1174.
[104] Y. Zhang, H.X. Zhang, Q.C. Zheng, In silico study of membrane lipid composition regulating conformation and hydration of influenza virus B M2 channel, J. Chem. Inf. Model. 60 (2020) 3603–3615.
[105] B. Lin, H. Zhang, Q. Zheng, How do mutations affect the structural characteristics and substrate binding of CYP21A2? An investigation by molecular dynamics simulations, Phys. Chem. Chem. Phys. 22 (2020) 8870–8877.
[106] Y. Zhang, Q.C. Zheng, What are the effects of the serine triad on proton conduction of an influenza B M2 channel? An investigation by molecular dynamics simulations, Phys. Chem. Chem. Phys. 21 (2019) 8820–8826.
[107] H.L. Chen, G. Wang, C. Ma, et al., An efficient hybrid kernel extreme learning machine approach for early diagnosis of Parkinson's disease, Neurocomputing 184 (2016) 131–144.
[108] L. Liu, D. Zhao, F.H. Yu, et al., Ant colony optimization with cauchy and greedy Levy mutations for multilevel COVID 19 X-ray image segmentation, Comput. Biol. Med. 136 (2021) 104609.
[109] F. Li, Y. Zhou, Y. Zhang, et al., POSREG: proteomic signature discovered by simultaneously optimizing its reproducibility and generalizability, Brief. Bioinform. 23 (2022) bbac040.