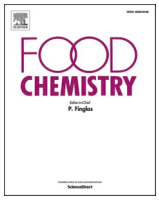
Food Chemistry 482 (2025) 143955

Contents lists available at ScienceDirect

Food Chemistry

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



Grain protein function prediction based on improved FCN and bidirectional LSTM☆

Jing Liu [a], Kun Li [a], Xinghua Tang [a], Yu Zhang [b,c], Xiao Guan [b,c,*]

[a] College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China
[b] School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China
[c] National Grain Industry (Urban Grain and Oil Security) Technology Innovation Center, Shanghai 200093, China

ARTICLE INFO

Dataset link: dataset (Original data)

Keywords:
Grain protein function prediction
Deep learning
Amino acid sequence order
Full convolutional networks
Bidirectional long short-term memory
Squeeze excitation

ABSTRACT

With the development of high-throughput sequencing technologies, predicting grain protein function from amino acid sequences based on intelligent model has become one of the significant tasks in bioinformatics. The soybean, maize, indica, and japonica are selected as grain dataset from the UniProtKB. Aiming at the problem of neglecting the sequence order of amino acids and the long-term dependence between amino acids, the PBiLSTM-FCN model is proposed for predicting grain protein function in this paper. The sequence of amino acid sequences is considered in the Fully Convolutional Networks (FCN), and the long-term dependence between amino acids is addressed by the bidirectional Long Short-Term Memory network (BiLSTM). The experimental results show that the PBiLSTM-FCN model is superior to existing models, and can predict more accurately by solving the problem of capturing long-range dependencies and the order of amino acid sequences. Finally, the interpretability analyses are performed by the actual protein function compared with the predicted protein function which proves the effectiveness of the PBiLSTM-FCN model in predicting grain protein function.

1. Introduction

Grain is one of the main food sources of human beings and provides most of the energy requirements (Raubenheimer & Simpson, 2016). Grain protein content is the main source of protein in human diet. The protein content in oats can be as high as about 16 % (Poutanen et al., 2022), and the protein content in wheat can reach approximately 18 % (Poutanen et al., 2022). Protein is a key factor determining nutritional quality. At the same time, proteins are widely distributed in organisms and are among the most important biomolecules. The basic units of proteins, which are large molecular compounds, are amino acids. The sequence of amino acids determines the structure of the protein, and the structure of the protein influences its function. Therefore, the study of grain protein function is of great significance to the development of human daily and proteomics.

With the improvement of the Human Genome Project and the continuous development and refinement of high-throughput sequencing technologies, a vast amount of protein sequence data has been generated, leading to a sharp increase in the number of unannotated proteins (Alex et al., 2022). Traditional experimental approaches to annotating protein functions are time-consuming and labor-intensive. The computational methods have become one of the mainstream methods for protein function prediction.

Traditional computational methods for protein function prediction, such as BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990), PSI-BLAST (Altschul et al., 1997), FASTA (Radivojac et al., 2013) primarily rely on the analysis of protein sequence similarity to predict functions. The greater the sequence similarity between two proteins, the more similar or even identical their protein functions (Gillis & Pavlidis, 2013). With the development of artificial intelligence, machine learning has been applied to the prediction of protein functions. The semantic similarity and the K-Nearest-Neighbors were integrated for predicting the protein function (Pandey et al., 2009). Support Vector Machine (SVM) was proposed to categorize proteins from different functional classes based on their amino acid sequences (Li et al., 2016). Decision tree and Random Forest were proposed for domain-based prediction of protein interactions (Chen & Liu, 2005). Co-learning (Nam et al., 2005) and Naive Bayes (Cao, Katheleen, Duong, & Ciso, 2008; Malik, Segun, Louise, & Showe, 2008) models were used for protein function prediction, achieving promising results. Compared to traditional machine

☆ This article is part of a Special issue entitled: 'AI for Food Chemistry' published in Food Chemistry.

* Corresponding author at: School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China.
E-mail address: gnxo@163.com (X. Guan).

https://doi.org/10.1016/j.foodchem.2025.143955
Received 14 May 2024; Received in revised form 10 March 2025; Accepted 17 March 2025
Available online 4 April 2025

learning models, deep learning models can learn directly from the original amino acid sequence data, without the need for human feature extraction. At the same time, the issues such as excessive dimensionality, redundancy, and noise in protein sequence data were successfully addressed by deep learning (Lv et al., 2019).

DeepGO (Kulmanov et al., 2018), DeepGOPlus (Kulmanov & Hoehndorf, 2020), ProtConv (Sara, Hasan, Ahmad, & Shatabda, 2021), and Deep_CNN_LSTM_GO (Elhaj-Abdou, El-Dib, El-Helw, & EII-Habrouk, 2021), DeepFRI (Gligorijević et al., 2021) and MMSNet (Liu, Zhang, Huang, Wei, & Guan, 2025) are the deep learning models designed to predict protein functions. DeepGO was proposed to predict protein functions using Convolutional Neural Network (CNN) model based on protein sequence and protein interaction networks. Based on the DeepGO, DeepGOPlus predicts functional annotations of proteins by combining DeepGOCNN, which predicts functions from the amino acid sequence of a protein using a 1D convolutional neural network, with the DiamondScore. ProtConv was proposed by Sara et al. to converted protein into vector representations, which are subsequently transformed into single-channel 2D images for processing through the CNN. Elhaj-Abdou et al. combined the CNN and Long Short-Term Memory (LSTM) networks to propose the Deep_CNN_LSTM_GO model for protein function prediction. This model does not require specialized GPU training and can be trained on any standard CPU. A graph convolutional network (GCN) was used in the DeepFRI to link sequence and structural data and achieves better performance in protein function prediction by using experi-mentally determined protein structure data. A multiscale one-dimensional convolutional neural network (1DCNN) was combined with a two-dimensional convolutional neural network (2DCNN) to enable the MMSNet to capture features.

Although existing models have achieved promising predictive performance in protein function prediction, there are still several issues with the present protein function prediction model. First, because the protein sequence has a certain length, existing models have difficulty capturing the long-range dependencies between amino acids when processing long protein sequences. At the same time, due to the specific environmental adaptability of grain proteins, their protein composition varies across different periods(Eugène et al., 2003), leading to more complex and stronger dependencies between amino acids. Mature grain proteins show great complexity and interaction (Anam et al., 2023). Secondly, the valid protein features information and the invalid protein information in the protein sequence cannot be identified, which impedes the identification of critical amino acid sequences that significantly influence protein function. Finally, the order of amino acid sequence is not considered by the existing models. When the amino acids in the sequence change, existing models may struggle to adapt effectively to these variations.

Given the current problems, PBiLSTM-FCN was proposed for predicting the functions of grain proteins in this study. In this study, a unique prediction method PBiLSTM-FCN is proposed to handle the problem of protein function prediction, taking into account the current problems mentioned above. First, the Fully Convolutional Networks (FCN) (Karim, Majumdar, Darabi, & Harford, 2019; Shelhamer et al., 2017) are used to automatically train effective features directly from raw protein sequence data, while accounting for the amino acid sequence order and adjusting for variations in amino acids. The combination of a deep coarse layer with a shallow fine layer is employed to achieve accurate and detailed information. To ensure precise and comprehensive feature extraction, the Squeeze-Excitation (SE) block is included concurrently to adaptably calibrate the input features. Furthermore, the bidirectional Long Short-Term Memory networks (BiLSTM) was proposed to extract the global and local features of proteins for addressing the challenge of the long-range dependencies. Importantly, the BiLSTM is better equipped to handle the inherent complexities of grain proteins.

In summary, the PBiLSTM-FCN model has been proposed to predict the functions of grain proteins, enabling the investigation of unannotated proteins within public protein databases. Moreover, new varieties that are more nutritious and higher yielding can be developed with the assistance of this model, thereby promoting advancements in agronomy. Protein data from four types of grains—soybean, maize, indica, and japonica—were obtained to predict grain protein functions in UniProtKB, ensuring diversity across the selected species. For evaluating the model's performance, a thorough analysis was conducted using five evaluation metrics, and comparisons with existing models were made, demonstrating the superior generalization and robustness of the PBiLSTM-FCN. Furthermore, through interpretability analysis methods, mis-predicted proteins were identified, and a comparison between the actual annotated protein functions and the predicted outcomes from the model was conducted. This process revealed potential annotations for grain proteins, establishing a foundation for subsequent functional validation and biological research.

The architecture of the study was shown in Fig. 1, which includes data acquisition, data processing, data division, PBiLSTM-FCN, and model evaluation. The PBiLSTM-FCN mode is proposed for grain protein function prediction, which combines BiLSTM branch that the long-term trend of protein data was learned and FCN branch that local feature information of protein data was extracted. The comparative experiments between PBiLSTM-FCN and existing models in five evaluation metrics were demonstrated that the performance of PBiLSTM-FCN outperforms existing models. The results demonstrate the generalization and effectiveness of the model. The interpretability analysis is used by the actual protein function compared with the predicted protein function. The interpretability analysis of the results also identified potential function annotations of protein. Finally, the study is summarized with emphasis placed on the application value of the PBiLSTM-FCN model in predicting grain protein functions.

2. Materials and methods

2.1. Datasets

Deep learning models typically require large-scale datasets to ensure their generalization and robustness. When selecting specific proteins, the data was obtained for oats and other grains type. However, the dataset sizes for oats were too small to effectively train deep learning models. Soybeans is considered as an important food source to meet protein demand for the human body, and has been regarded as the most popular plant protein with the highest industrial production (Pingxu et al., 2022). Maize and its products comprised 30 % of food supply for Americas, 38 % for Africans and 6.5 % for Asian (Maqbool et al., 2021). Rice is one of the most important food grains in the world (Ma et al., 2022). Indica and japonica rice are the two subspecies of cultivated rice. More than half of the population in the world relies on rice as a staple food (Min, Chengjing, Jiaxin, Jiana, & Fangbo, 2023). Rice is not only an energy source but also an important nutritional source for rice consumers. The four grains are important among grains, covering the diversity of grain species. Plant proteins typically lack one or two essential amino acids. In soybeans, methionine and cysteine (SAA) are lacking (Pingxu et al., 2022). Lysine and tryptophan are deficient in maize (Maqbool et al., 2021). Lysine and threonine are not sufficiently present in Indica and Japonica rice (Jiang, Ma, Xie, & Ramachandran, 2016). The amino acid profiles of these four grains can complement each other, thus providing a more comprehensive coverage of essential amino acids when used in combination, sufficient to cover biodiversity. Therefore, the four grain types (soybean, maize, indica rice, and japonica rice) with sufficiently large datasets were chose for the study.

Considering the diversity of grains and the requirement for large datasets, grain protein data for four types of grain proteins—soybean, maize, indica, and japonica—were selected from the SwissProt of UniProtKB (https://www.uniprot.org/uniprotkb). The data in the SwissProt database are manually annotated function data from the UniProtKB. Gene Ontology (GO) annotations are currently widely recognized as the

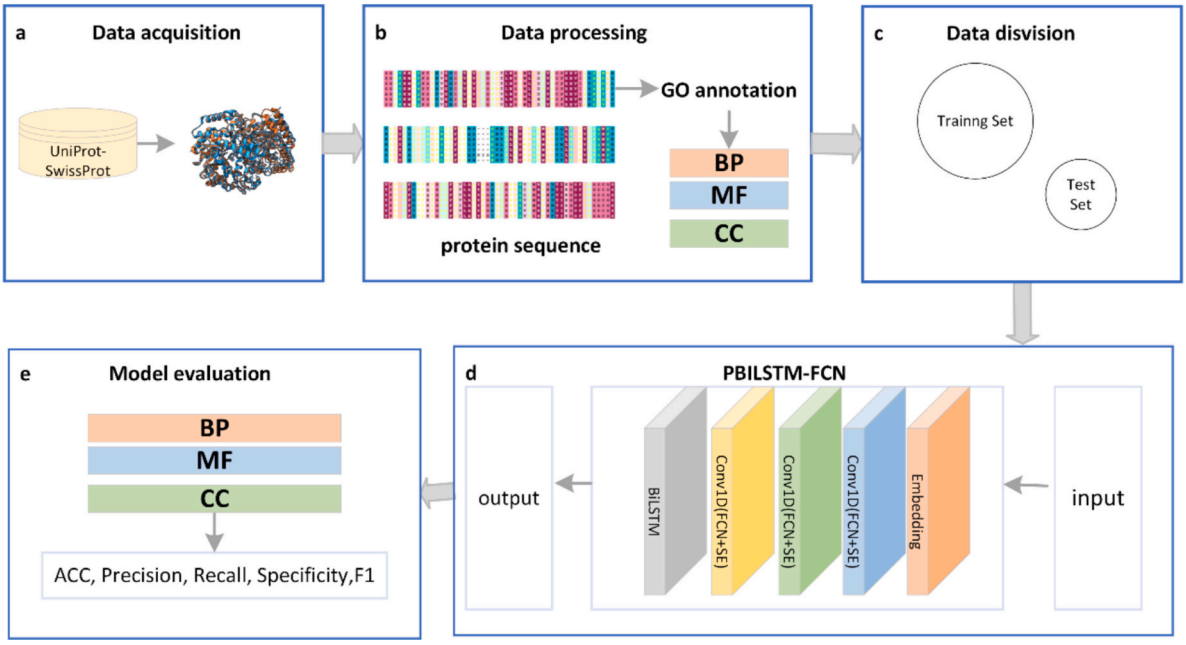


Fig. 1. The architecture of the study.

gold standard for protein function annotation (Huntley, Sawford, Martin, & Donovan, 2014). The GO is divided into three distinct sub-ontologies according to different functional categories: Molecular Function (MF), Biological Process (BP), and Cellular Component (CC) (Suzi et al., 2023). Separate datasets were constructed for each of the three GO sub-ontologies—MF, BP, and CC—based on the scope of GO functional annotations. When selecting protein sequences for each type of grain, the following criteria was adopted: the protein sequences with complete GO annotations were selected to ensure the accuracy of the labels. Based on the current Gene Ontology (GO), grain proteins with experimental evidence codes (EXP, IDA, IPI, IMP, IGI, IEP, TAS, and IC) were filtered and retained, removing any proteins without annotations. When annotating grain proteins, the hierarchical structure of GO terms was utilized. During the annotations of protein, the issue of proteins having GO terms from different sub-ontologies was considered. For example, if a protein P has both BP and MF terms, it would be included in both the BP and MF datasets. In the study, each GO class contains at least 50 annotated proteins. The data volume of four grain types (soybean, maize, indica and japonica) and their distribution under different sub-ontology (BP, MF, CC) were shown in the Table 1.

The original dataset was divided into an 80 % training set and a 20 % test set, ensuring that the test set was entirely independent of the training process, thereby allowing for an objective evaluation of the model's generalization capability. From the training set, a validation set was partitioned to tune model parameters and prevent overfitting, with the data samples between the training, validation, and test sets kept mutually independent. Furthermore, experiments were conducted 10 times, with results averaged, reducing the impact of random factors and enhancing the reliability of the conclusions.

Table 1
The distribution of protein sequences samples in the grain protein dataset.

Grain type	Sub-ontology	Training samples	Test samples	Total
Soybean	BP	264	67	331
	MF	283	71	354
	CC	256	65	321
Maize	BP	534	134	668
	MF	608	153	761
	CC	612	153	765
Indica	BP	652	164	816
	MF	756	189	945
	CC	754	189	943
Japonica	BP	2588	647	3235
	MF	2858	715	3573
	CC	2796	700	3496

2.2. Data representation

The amino acid composition method(n-gram) was used (Kabli et al., 2018) to statistically analyze amino acid sequences and digitize protein sequence data in this study. The amino acid was divided into 1-g and 2-g groups, as shown in Fig. 2 and Fig. 3, respectively. The protein sequence is consisted of 20 natural amino acids {A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y} , which are encoded as nature numbers{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20}. The trainable temporary embedding coding approach was used to initialize each item as a random vector with varying word vector dimension sizes. During the training process, the dimension of the word vector is trained together with other network parameters, and Tables 2 and 3 show the encoding strategy of the trainable embedding approach. The embedding vector size was designed using the research of Zuallaert, Pan, Saeys, Wang, & Neve (2019). Specifically, the word vector dimension size was found to be [5,10,15] for a 1-g amino acid composition and [5, 10, 15, 32, 64, 128] for a 2-g amino acid composition. For instance, the vector randomly assigned to amino acid A is [0.01, −0.02, 0.03,0.01, −0.01] when the size is 5.

2.3. FCN and BiLSTM

Convolutional layers and pooling layers were combined to form the FCN, a particular kind of neural network (Villa et al., 2018). This version of CNN is devoid of completely connected layers. Fixed-length feature vectors are obtained by applying numerous fully connected layers after convolutional processes. On the other hand, the convolutional layers with varying parameters were substituted for each fully connected layer. The input data that have the same length and have a fixed size was required in the CNNs. Zero-padding is necessary to make the input data

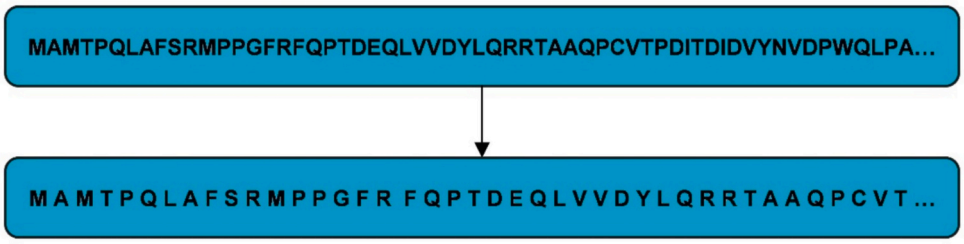


Fig. 2. Example of 1-g splitting.

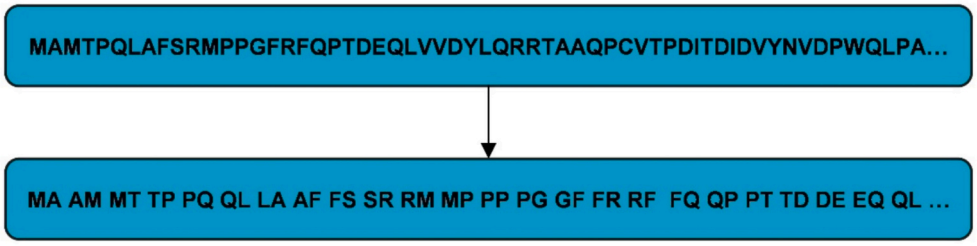


Fig. 3. Example of 2-g splitting.

Table 2
Random encoding strategy for trainable embedding methods.

A	C	D	...	Y
0.01	0.02	−0.01	…	0.01
−0.02	0.03	−0.03	…	0.01
0.03	−0.01	0.02	…	−0.04
…	…	…	…	…
−0.01	−0.03	0.01	…	−0.02

the same size if it is not long enough. However, FCN does not impose any limitations on the amount of the input data and is capable of accepting files of any size, including varying their dimensions. When processing lengthy sequence data, FCN is more adaptable than CNN and retains more characteristics from the protein data. This method is used in this work to predict the functional properties of grain proteins using an FCN-based model.

While knowledge from previous data can was extracted in the Recurrent Neural Networks (RNNs) (Zaremba et al., 2015), their context range for information storage is constrained. Learning from extended sequences can be severely hampered by this restriction, which can result in problems like vanishing gradients, bursting gradients, and lengthy training times. The-LSTM were presented by Hochreiter and Schmidhuber (1997) and Hochreiter (1998) as a solution to the RNN's noted issues. By including three control gates and a memory cell, the sequential input was handled by LSTM and improves the model's ability to represent sequences with long-term dependencies. It allows the model to learn longer sequences of dependencies and helps mitigate the issue of vanishing or ballooning gradients. Consequently, NLP and other sequence data processing tasks frequently employ LSTMs. Fig. 4 displays the architecture of the BiLSTM network.

2.4. Squeeze-excitation

To better extract distinctive features between amino acid segments with high similarity in protein sequences, the Squeeze-and-Excitation (SE) block proposed by Hu et al. in 2017 is introduced on the basis of the FCN algorithm (Hu et al., 2018). The four components of the SE block are Transform, Squeeze, Excite, and Scale. These four methods can be used to increase channel attentiveness, which will increase the model's capacity to extract characteristics from protein sequences. The four components of the Squeeze-Excitation block are explained in detail below.

(i) The formula for a convolutional transform is $F_{tr}: X \rightarrow U$, which entails that the input $X \in \mathbb{R}^{W \times H' \times C'}$ is mapped to the unweighted feature$U \in \mathbb{R}^{W \times H \times C}$ through the F_{tr} operation. H,W and C are the height, width, and number of feature channels of the feature map, respectively. The output of F_{tr} is expressed as $U = [u_1, u_2, \cdots u_C]$, where u_c is defined as shown in Eq. (1).

$$u_c = v_c * X = \sum_{S=1}^{C'} v_c^s * x^s \tag{1}$$

Where * represents convolution operation, and the 2D spatial kernel is represented by v_c^s. The single channel of v_c acts on the corresponding channel of the model, and the interdependence of the channels is divided into two steps of squeeze and excitation to adjust the filter response.

(ii) In order to fully extract the global feature information of protein sequence from the feature transformed by convolution, the feature U is squeezed. The squeeze operation mainly aggregates information within the channels along the spatial dimensions, so as to obtain the channel statistics that only cover the relationship

Table 3
Example of input sequences for trainable embedding.

C	M	A	A	A	D	Y	Y	Y	A	...
0.02	0.02	0.01	0.01	0.01	−0.01	0.01	0.01	0.01	0.01	…
0.03	−0.02	−0.02	−0.02	−0.02	−0.03	0.01	0.01	0.01	−0.02	…
−0.01	0.03	0.03	0.03	0.03	0.02	−0.04	−0.04	−0.04	0.03	…
…	…	…	…	…	…	…	…	…	…	…
−0.03	0.01	−0.01	−0.01	−0.01	0.01	−0.02	−0.02	−0.02	−0.01	…

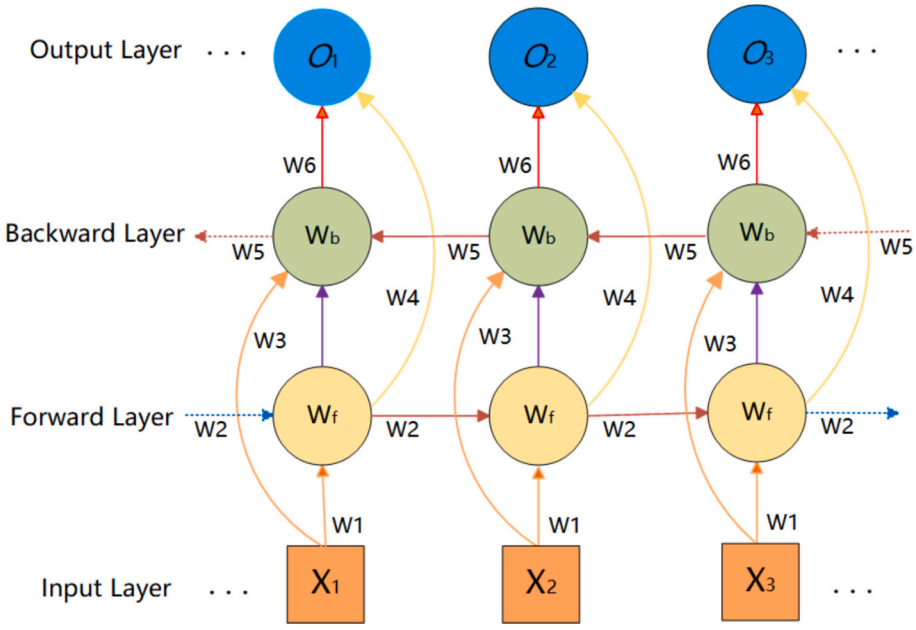


Fig. 4. BiLSTM network architecture.

between channels, and eliminate the interference of spatial distribution information. The squeeze operation is usually implemented by using the global average pooling operation to compress each channel of the transformed feature U along the spatial dimension $H \times W$ thereby obtaining the channel statistics $z \in \mathbb{R}^c$. Where C represents the number of channels, and the nth element of Z is computed using $F_{sq}(u_n)$ as defined in Eq. 2.

$$z_n = F_{sq}(u_n) = \frac{1}{W \times H}\sum_{j=1}^{W}\sum_{j=1}^{H} u_n(i,j) \tag{2}$$

Where $u_n(i,j)$ represents the element at the ith row and jth column of the nth channel in the feature U where n takes values from $1,2,\cdots,C$. $F_{sq}(u_n)$ is the function that compresses the input variable of size $W \times H \times C$ into a feature variable of size $1 \times 1 \times C$.

(iii) In order to obtain the dependence between channels comprehensively and improve the expression ability of protein sequence features of the neural network, the compressed channel statistic z is stimulated. The excitation operation is achieved through a fully connected layer + ReLU function + fully connected layer to form a bottleneck structure. The first fully connected layer compresses the original input C channels into C/rchannels, The second fully connected layer restores C/r channels to C channels. Finally, the normalized weight in the range of [0,1] can be obtained through the sigmoid function, and the output s is the weight that describes the importance of each channel in the feature U. The excitation process is represented by the excitation function F_{ex}, as shown in Eq. 3.

$$s = F_{ex}(z, W) = \sigma[g(z, W)] = \sigma[W_2\delta(W_1 z)] \tag{3}$$

Where σ and δ represent the sigmoid activation function and the ReLU activation function, respectively. $W_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ represents the dimension reduction parameter of the first connection layer. $W_2 \in \mathbb{R}^{\frac{C}{r}\times C}$ represents the dimension growth parameter of the second connection layer. r represents the reduction factor that balances the ratio between model performance and computational complexity.

(iv) The feature vector U is multiplied channel-wise with the corresponding weight coefficients s using F_{scale}, resulting in the final output $\widetilde{X}$, as shown in Eq. 4 and Eq. 5.

$$\widetilde{x_c} = F_{scale}(s_c, u_c) = u_c s_c \tag{4}$$

$$\widetilde{X} = \left[\widetilde{x_1}, \cdots, x_c, \widetilde{\cdots}, \widetilde{x_C}\right] \tag{5}$$

The complete squeeze-excitation operation is described above. In order to reflect the varying relevance of distinct feature information, each channel feature is given an unequal weight. This results in the suppression of less important or irrelevant information and the stimulation and amplification of information that is useful for the goal task. Consequently, the performance of the PBiLSTM-FCN is enhanced. Furthermore, there is no need to modify the neural network topology or change the hyperparameters while using the Squeeze-Excitation block. It fits well into the current network infrastructure and is quite easy to use. The model's complexity and computing load are not appreciably increased by the Squeeze-Excitation block. It is frequently used to improve model performance in a variety of lightweight models.

2.5. PBiLSTM-FCN

The PBiLSTM-FCN was composed of five main modules. As shown in Fig. 5, BiLSTM networks were combined Dropout layer to form BiLSTM branch. Conv1D, Batch Normalization (BN) and ReLU were composed to form FCN branch. The parallel combination of FCN branch and BiLSTM branch forms the PBILSTM-FCN model. The BiLSTM and FCN branch run in parallel, which comprehensively utilizes the feature extraction advantages of FCN and the information mining ability of BiLSTM for long sequences.

FCN was formed by combining three Conv1D layers to process input data. Ordered amino acid sequence data can be better processed for feature extraction, thereby enabling the local information of protein data to be captured more effectively. The local features were extracted by sliding filters over a one-dimensional sequence in the Conv1D layer. Local patterns in protein sequences, such as functional domains, can be effectively captured by Conv1D, aiding in the identification of important functional domains. To enhance the stability and robustness of the FCN, BN layer and ReLU were added after each Conv1D layer. Vanishing or exploding gradients was prevented by the BN and dispersed feature

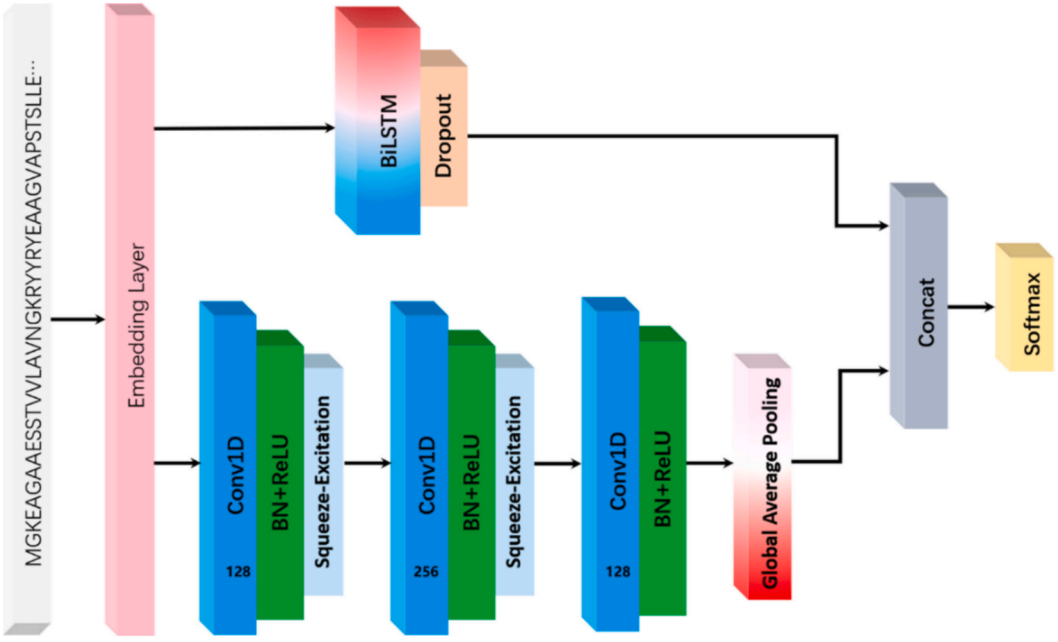


Fig. 5. PBiLSTM-FCN Model Structure.

distributions in samples was addressed by the BN. A stable distribution of input to each layer is ensured by BN, which accelerates the training process and improves model stability. Sensitivity to initialization parameters is also reduced, making training more robust. Nonlinear transformations are provided by ReLU, enabling the model to capture complex biological signals. A simple and efficient activation function that is easy to compute and optimize, ReLU is used to avoid the vanishing gradient problem and to enhance the training efficiency of neural networks. Additionally, SE modules were added between every two Conv1D layers. Channel features are reweighted by the SE, enhancing the FCN's focus on important local features and thereby improving prediction accuracy and reliability.

BiLSTM was composed of two LSTMs, one LSTM processing sequence and the other backward processing. This is particularly useful for understanding protein sequences, where the function of an amino acid can be influenced by its neighboring residues. To address the overfitting issue in the neural network model, the BiLSTM branch incorporates Dropout alongside the BiLSTM network. Finally, to classify the extracted feature information, the outputs from both branches are concatenated and passed through a softmax layer. LSTM selectively retains or forgets information through its internal memory units and gating mechanisms (input gate, forget gate, and output gate), effectively capturing long-term trends in time series data.

In summary, the long-term trend of protein data was learned by BiLSTM branch and local feature information of protein data was extracted by FCN branch. Long-range dependencies in amino acid sequences are captured by BiLSTM, aiding in the understanding of the overall structure and function of proteins. Conserved patterns and motifs within short peptide fragments are identified by FCN to identified the order of amino acid sequence. Therefore, PBiLSTM-FCN has been proposed to solve the problem of difficulty in capturing long-range dependencies and the order of amino acid sequences, improving the accuracy of model prediction.

The Conv1D in each convolutional block contain numerous filters, with filter sizes of 128, 256, and 128 in that order. There are matching kernels for the Conv1D in the sizes of 8, 5, and 3. The two parameters used in BN are ε and the momentum of the dynamic mean, which are set at 0.99 and 0.001, respectively. BN can assist prevent vanishing or bursting gradients and successfully address the issue of scattered feature distribution in samples. ReLU activation function offers neural network training that is effective and simple to compute and optimize. Furthermore, SE are added to the first two Conv1D of the FCN branch, and all SE have a dimensionality reduction factor r set to 16. In order to reduce model parameters and model calculation, the final layer of the FCN branch substitutes a global average pooling layer for the conventional fully connected layer. An addition to the FCN that can adaptively adjust input features is the SE. This enables the overall model size to increase by only 3 %–10 %.

2.6. Evaluation metrics

In protein function classification tasks, model evaluation criteria including Accuracy, Precision, Recall, F1-score are frequently employed (Xu & Wang, 2019). In order to evaluate the performance of PBiLSTM-FCN in comprehensive aspects, the specificity evaluation metric was added (Gireen et al., 2023). The five metrics were used in this study as evaluation criteria to assess the performance of the protein function prediction models. The percentage of the model was shown that is genuinely a positive example among the samples that are projected to be positive instances, while the percentage of samples was measured by the Accuracy that the model predicts is correct. The percentage of samples that the model can accurately identify as positive examples is shown by the Recall. The F1-score is a comprehensive metric that assesses the balance between prediction Accuracy and coverage by taking into account both Precision and Recall of the model. Among the five performance measurements, Accuracy and F1 are the two most significant metrics. These five metrics are specifically defined by Eqs. 6–10 respectively. By calculating these metrics, the performance of the proposed model was comprehensively evaluated in protein function prediction tasks.

$$\text{Accurary} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{8}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (10)$$

3. Results and discussion

3.1. The prediction results and analysis of the PBILSTM-FCN on different grain types

Each experiment was run 10 times, and the average of 10 times function prediction results was taken. The results of protein function for four different grain types across three sub-ontologies (BP, MF, CC) were shown in the Table 4. The word vector size is 128, and amino acid composition is 2. Among all sub-ontologies, soybean generally has high accuracy, specificity, and F1, especially in BP and MF. The results indicate that the soybean has high predictive ability within these two ontologies. Specificity is particularly prominent in all types of grains, which is crucial for reducing false positive predictions. Accuracy is represented by the proportion of samples that the model correctly classifies. The F1 is defined as the harmonic mean of precision and recall, aimed at balancing the trade-off between the two metrics. The high accuracy is achieved through correct classification of the majority of samples, and high recall is maintained with effective minimization of the false positive rate. As a result, excellent overall performance and superior handling of class imbalance issues are demonstrated, ensuring precise differentiation between positive and negative instances. From the Table 4, it can be seen that models of different grain types perform differently in each sub-ontology.

For a clear comparison of grain performance across BP, MF, and CC, the results are shown in Fig. 6. In soybeans, the specificity (red bar chart) is particularly prominent in all sub-ontologies, indicating that the model is effective in correctly identifying true negatives. Similar to soybeans, the specificity of maize performs well in all sub-ontologies, demonstrating the model's powerful ability to distinguish between positive and negative predictions. The specificity of indica and japonica are high, especially in the BP and MF sub-ontologies, while there is a slight decrease in the CC sub-ontology.

3.2. The comparison results and analysis of the PBILSTM-FCN with other models on different grain types

To demonstrate the effectiveness of the PBiLSTM-FCN model, multiple comparative experiments were conducted with other models. First, the PBiLSTM-FCN model was compared with basic models (CNN, LSTM, and FCN). As test subjects for predicting the function of proteins, the soybean, maize, indica, and japonica proteins are selected. The aim of these comparisons is to show the advantages of the PBiLSTM-FCN model over existing basic models. To further validate the superiority of the PBiLSTM-FCN model, an additional comparative experiment was conducted on japonica grain, which has a larger dataset, against the latest protein function prediction models (DeepGOCNN, DeepFRI, and MMSNet). At the same time, in order to observe the influence of amino acid composition and word vector composition on the model, a comparison was made between different amino acid compositions and word vector compositions on the basic model. The PBiLSTM-FCN is used, and word vector sizes are varied according to 1- and 2-g weights. The four grain proteins will be used in tests to predict protein function. The ten iterations of experimentation were used in every model, after which the function predictions are averaged. Additionally, the experimental results are shown as a percentage, with 1- and 2-g values for n, which stands for amino acid composition. The 2-g amino acid composition and 128-word vectors were selected for use in the comparisons with the latest models.

The Soybean protein function prediction results are shown in Table 5, over 81 % F1 and accuracy for varying parameter values were achieved in the PBiLSTM-FCN. In the BP dataset, an impressive 87.273766 % was reached in the F1 when n-gram is 2-g and size is 128. In the MF dataset, the well performance was shown across various parameter settings in the PBiLSTM-FCN, particularly showing outstanding performance when using higher dimensions for word embeddings. In the MF dataset for soybean, when using a 2-g and a size of 128, an impressive Accuracy of 89.566377 % was achieved in the PBiLSTM-FCN. Moreover, compared to the CNN, the Accuracy improvement of 6.686274 % and F1-score improvement of 9.533449 % were shown in the PBiLSTM-FCN. In the CC dataset, the LSTM performed better than other models under various parameter settings. The lower Accuracy and F1-score were achieved by the LSTM compared to the PBiLSTM-FCN, while PBiLSTM-FCN algorithm performs better when the dimension of word vector size is higher. In general, the performance of different data sets under different parameter settings and models are quite different. The excellent performance of the PBiLSTM-FCN model across different datasets was shown, particularly when using a higher dimensional word vector size, indicating its effectiveness in extracting protein features and accurately classifying protein functions.

Maiz protein function prediction results are shown in Table 6, the PBiLSTM-FCN model results are better than the other three models (CNN, LSTM, and FCN) in regard to Accuracy and F1-score in different combinations of word vector size, n-gram, and sub-ontology subsets. In the results of maize protein function prediction, the PBiLSTM-FCN model's accuracy is above 79 %, and the F1-score is over 70 %. Specifically, in the maize MF data set, the maximum Accuracy and F1-score (88.846801 % and 78.360687 %, respectively) were achieved by the

Table 4
The prediction results of the PBILSTM-FCN on difficult grain types.

Grain type	Sub-ontology	Accuracy(%)	Precision(%)	Recall(%)	F1(%)	Specificity(%)
Soybean	BP	88.333446	81.609949	93.830093	87.273766	83.783784
	MF	89.566377	86.721400	87.693513	87.181050	90.843524
	CC	86.384032	84.606908	86.991593	85.706566	85.840480
Maize	BP	83.22393	74.111709	74.163607	74.114245	87.570820
	MF	88.846801	77.602347	79.223470	78.360687	92.150421
	CC	83.118925	77.161496	69.400850	73.006987	89.878377
Indica	BP	88.055110	79.344762	74.746799	76.908342	92.906584
	MF	84.745584	66.047824	61.947255	63.735461	91.110128
	CC	81.665445	77.667135	78.478301	78.040732	83.934473
Japonica	BP	79.340080	69.295369	62.716729	65.780635	87.072885
	MF	85.867983	80.135982	87.610547	83.610528	84.635097
	CC	80.085478	78.749999	79.328485	79.010658	80.765669

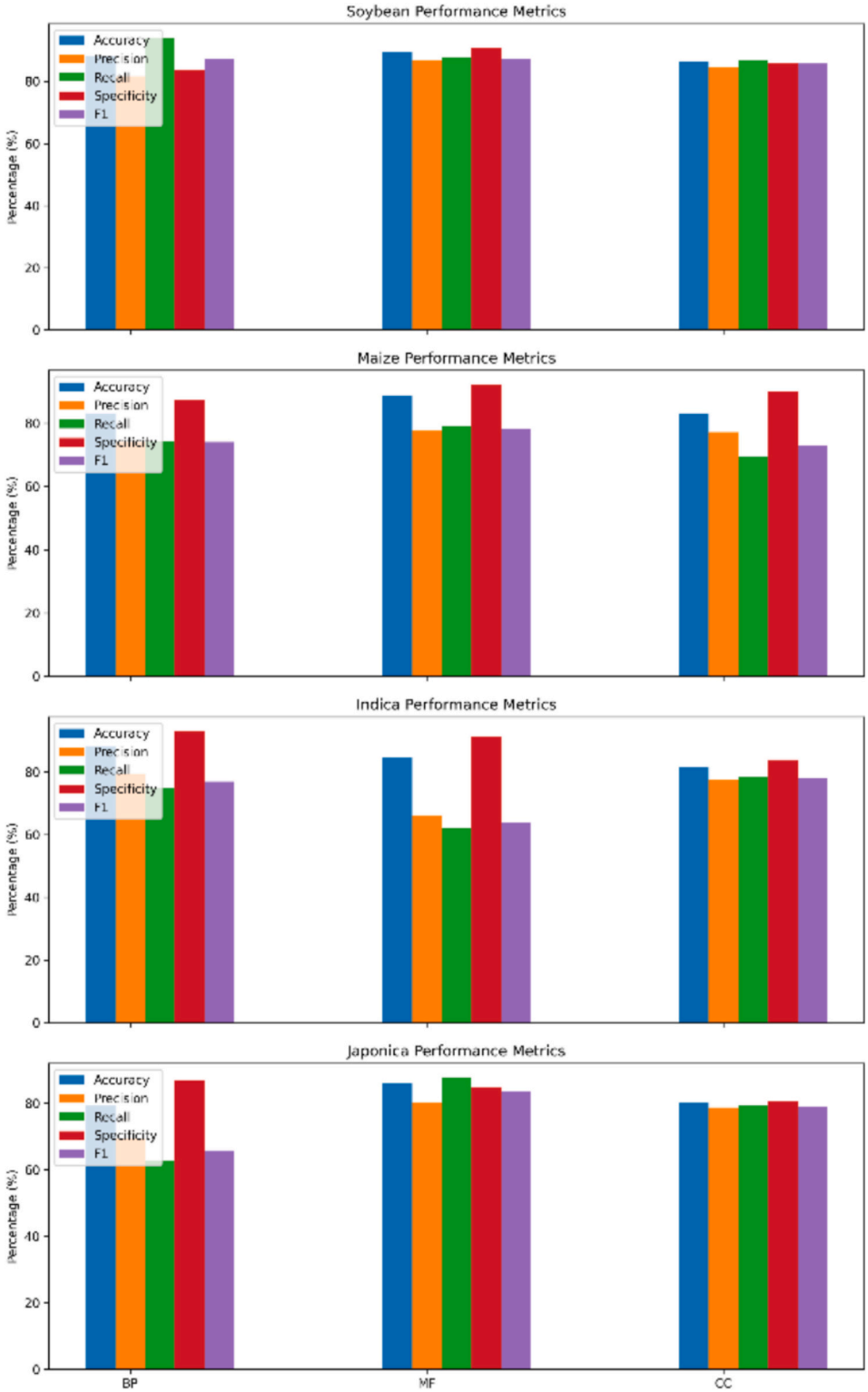


Fig. 6. The bar chart of results on different grain type.

PBiLSTM-FCN when the n-gram is 2 g and the size is 128. These results are significantly higher than the highest Accuracy and F1-score of other models. Comparing the results of PBiLSTM-FCN model with other models, it can be observed that in the MF dataset of maize, when n-gram is 2-g and size is 64, the Accuracy and F1-score of PBiLSTM-FCN are respectively increased by 8.751357 % and 11.073443 % compared with CNN. In general, provides the best performance in most cases was shown in the PBiLSTM-FCN, and it is particularly obvious when the n-gram is 2. This indicates that PBiLSTM-FCN can better identify the relevant information of protein sequence, and improve the Accuracy of maize protein function prediction models. Therefore, it can be concluded that the PBiLSTM-FCN exhibits superiority in predicting maize protein functions.

Indica protein function prediction results are demonstrated in Table 7 that the PBiLSTM-FCN of the protein function with an accuracy of more than 81 %. The PBiLSTM-FCN is best when the n-gram is 1 and

Table 5
Soybean protein function prediction results.

Sub-ontology	n	size	Algorithm	Accuracy(%)	Precision(%)	Recall(%)	F1(%)	Specificity(%)
BP	1	5	CNN	82.834667	75.868729	89.866942	82.175838	77.23356
			LSTM	82.931251	78.763315	85.122721	81.794114	81.12943
			FCN	83.684108	76.597773	89.106400	82.379638	79.62617
			PBiLSTM-FCN	**83.749574**	76.758085	89.123185	**82.471818**	79.70914
		10	CNN	83.774969	76.203455	89.657296	82.359733	79.45941
			LSTM	82.575701	79.006537	84.400276	81.588724	81.03256
			FCN	83.820465	76.399530	89.727363	82.489345	79.43862
			PBiLSTM-FCN	**84.119299**	77.783074	88.757872	**82.888021**	80.5629
		15	CNN	83.860919	76.657388	89.434810	82.469605	79.70767
			LSTM	83.035207	78.810990	85.254838	81.890340	81.21626
			FCN	83.678622	76.601723	89.082741	82.371741	79.63384
			PBiLSTM-FCN	**84.086603**	77.883666	88.631852	**82.876492**	80.57937
	2	5	CNN	83.635053	76.759020	88.728470	82.225708	79.80652
			LSTM	83.464150	78.040678	87.192978	82.327730	80.50058
			FCN	83.645182	76.669462	88.930376	82.345253	79.67581
			PBiLSTM-FCN	**85.773566**	79.449465	90.376932	**84.526490**	82.2852
		10	CNN	83.073111	76.860328	87.625583	81.577528	79.54285
			LSTM	83.452794	78.606597	86.655546	82.406627	80.85253
			FCN	83.660384	76.657933	88.999032	82.368302	79.65181
			PBiLSTM-FCN	**85.788450**	79.722255	90.536966	**84.697848**	82.09678
		15	CNN	82.761675	77.760748	85.589036	81.065510	80.51053
			LSTM	83.428532	78.531973	86.406079	82.256026	81.03826
			FCN	83.636660	76.718093	88.891664	82.356967	79.67789
			PBiLSTM-FCN	**86.722037**	81.017884	90.686623	**85.529498**	83.67615
		32	CNN	83.543773	77.168833	88.401511	82.271845	79.79014
			LSTM	83.469416	79.220993	85.965848	82.418250	81.41136
			FCN	83.658680	76.641582	88.991942	82.355760	79.65888
			PBiLSTM-FCN	**86.713482**	80.871179	91.017203	**85.574588**	83.3938
		64	CNN	82.777965	78.490359	85.695575	81.908253	80.33482
			LSTM	83.214532	79.306163	85.597690	82.287213	81.20964
			FCN	83.653720	76.664562	88.969181	82.359204	79.66168
			PBiLSTM-FCN	**87.228277**	81.477041	91.597277	**86.210966**	83.83717
		128	CNN	83.063722	77.717302	87.015175	81.713270	79.87647
			LSTM	83.490417	78.987836	85.950057	82.295827	81.5003
			FCN	83.621220	76.730092	88.821937	82.333902	79.70236
			PBiLSTM-FCN	**88.333446**	81.609949	93.830093	**87.273766**	84.23512
MF	1	5	CNN	84.918129	69.152386	94.389374	79.782133	80.54099
			LSTM	84.506414	71.615831	91.451696	80.206771	80.83389
			FCN	84.141597	67.576682	94.939979	78.953553	79.21423
			PBiLSTM-FCN	**85.169246**	80.234054	83.787658	**81.964120**	86.09964
		10	CNN	83.688996	71.464574	93.046063	80.753080	78.19811
			LSTM	84.433128	70.110571	92.783671	79.797987	80.26737
			FCN	84.014794	67.206609	95.145581	78.771686	78.97369
			PBiLSTM-FCN	**84.764751**	80.157732	82.856342	**81.460114**	86.06166
		15	CNN	84.815538	69.381968	93.952487	79.776206	80.52338
			LSTM	84.716341	71.281173	91.900680	80.193637	81.03686
			FCN	83.990119	67.116145	95.203567	78.729448	78.92334
			PBiLSTM-FCN	**85.647028**	80.952619	84.079389	**82.464855**	86.70087
	2	5	CNN	84.945325	71.472787	91.503378	79.839945	81.6504
			LSTM	85.971761	76.834809	88.911471	82.287914	84.24386
			FCN	84.088725	67.294106	95.150682	78.833472	79.08598
			PBiLSTM-FCN	**87.929360**	82.695014	87.403243	**84.965545**	88.26686
		10	CNN	85.202161	69.827946	94.078379	80.094308	81.06835
			LSTM	85.864156	77.296272	88.485550	82.327610	84.27847
			FCN	84.034368	67.288495	95.089718	78.808351	79.01592
			PBiLSTM-FCN	**87.791403**	82.957439	86.298626	**84.511449**	88.73953
		15	CNN	84.784501	69.327079	94.168210	79.768902	80.36138
			LSTM	85.830325	78.566775	87.244930	82.463206	84.93492
			FCN	84.016342	67.162576	95.190374	78.756851	78.96636
			PBiLSTM-FCN	**87.567754**	83.545077	85.462697	**84.464177**	88.94969
		32	CNN	85.160914	71.314384	92.765454	80.434876	81.36266
			LSTM	86.707825	80.000607	87.737968	83.559934	86.05278
			FCN	84.043224	67.165585	95.220296	78.769186	79.00111
			PBiLSTM-FCN	**88.786214**	84.574610	87.355546	**85.901044**	89.71019
		64	CNN	83.382452	66.779389	94.768227	78.285794	78.09128
			LSTM	86.664459	80.790975	86.817141	83.559957	86.56508
			FCN	84.009960	67.198518	95.147555	78.766309	78.96584
			PBiLSTM-FCN	**89.071571**	85.734568	87.285188	**86.475579**	90.26854
		128	CNN	82.880103	65.995073	94.752258	77.647601	77.37947
			LSTM	87.167469	81.310257	87.447227	84.170391	86.98634
			FCN	83.975518	67.135579	95.166787	78.729429	78.90887
			PBiLSTM-FCN	**89.566377**	86.721400	87.693513	**87.181050**	90.84352

(*continued on next page*)

Table 5 (*continued*)

Sub-ontology	n	size	Algorithm	Accuracy(%)	Precision(%)	Recall(%)	F1(%)	Specificity(%)
CC	1	5	CNN	80.786536	83.266121	76.445676	79.337679	85.01932
			LSTM	80.505676	79.517515	80.205336	79.824033	80.78501
			FCN	80.224802	79.350018	80.001822	79.674222	80.43435
			PBiLSTM-FCN	**81.930389**	80.623759	82.088870	**81.323850**	81.78406
		10	CNN	80.804231	83.395986	75.948788	79.110403	85.46986
			LSTM	80.974944	79.623989	81.041266	80.299060	80.91391
			FCN	79.958334	77.439382	81.440821	79.056568	78.62262
			PBiLSTM-FCN	**82.270045**	81.677594	81.717679	**81.673136**	82.78866
		15	CNN	80.258253	81.608930	76.363558	78.620882	83.9016
			LSTM	80.658589	79.148572	80.963359	80.041883	80.37822
			FCN	80.224802	79.350018	80.001822	79.674222	80.43435
			PBiLSTM-FCN	**82.335377**	81.057466	82.483492	**81.745839**	82.19859
	2	5	CNN	80.794637	82.887310	76.777237	79.614553	84.67793
			LSTM	80.795165	78.329756	81.775862	80.003439	79.92494
			FCN	79.951003	77.394816	81.499365	79.074054	78.55616
			PBiLSTM-FCN	**85.483372**	85.156716	84.672510	**84.811621**	86.23939
		10	CNN	80.617528	82.712718	76.228123	78.847011	84.80413
			LSTM	81.353864	79.703558	81.625628	80.618123	81.10685
			FCN	80.227746	79.364784	79.978626	79.670012	80.46181
			PBiLSTM-FCN	**86.171688**	85.707166	85.556483	**85.569227**	86.7433
		15	CNN	80.651799	82.548343	77.091763	79.508464	84.12042
			LSTM	80.431736	78.995721	80.657444	79.801143	80.2236
			FCN	80.235614	79.434473	79.941701	79.686527	80.51236
			PBiLSTM-FCN	**86.465337**	85.239140	86.676219	**85.914020**	86.27247
		32	CNN	81.017571	84.636587	74.721839	78.962529	87.03529
			LSTM	80.797646	78.904096	81.278187	80.027133	80.36341
			FCN	80.224802	79.350018	80.001822	79.674222	80.43435
			PBiLSTM-FCN	**86.191387**	82.909412	88.548912	**85.587360**	84.14338
		64	CNN	80.118485	78.620009	80.567277	79.570466	79.70272
			LSTM	81.870444	81.026722	81.289989	81.121238	82.40693
			FCN	80.218947	79.307176	80.033379	79.668344	80.39318
			PBiLSTM-FCN	**86.686592**	84.137618	88.141146	**86.006131**	85.40943
		128	CNN	80.797391	82.365842	77.674034	79.594850	83.83372
			LSTM	81.167120	79.449073	81.412888	80.371060	80.94474
			FCN	80.247933	79.524185	79.870856	79.696436	80.60358
			PBiLSTM-FCN	**86.384032**	84.606908	86.991593	**85.706566**	85.84048

the size is 15, but otherwise, all of the models perform fairly similarly in the BP data set of indica. At this point, the PBiLSTM-FCN algorithm's accuracy, which stands at 88.29014 %, reaches its maximum. With an F1-score of 78.040732 %, the top results in the CC data set were delivered by the PBiLSTM-FCN when the n-gram is 2 g and the size is 128. When n-gram is one gram and size is 5, an increase of 5.541422 % in the F1-score in the MF data set was observed compared to other models. Among the different algorithms, the best performance was exhibited by the PBiLSTM-FCN, followed by the FCN algorithm, LSTM algorithm, and CNN. In terms of precision, the best performance on BP and CC datasets was delivered by the PBiLSTM-FCN, with a slightly lower performance on MF dataset, whereas the worst performance across all sub-ontology subsets was shown by the CNN. The highest F1-score continues to be achieved by PBiLSTM-FCN, while the lowest performance in this regard is consistently demonstrated by CNN. The performance of the FCN and LSTM is extremely similar. A trend similar to the previously indicated Accuracy and F1-score performance is exhibited by recall and accuracy. Overall, the other three models were outperformed by the PBiLSTM-FCN in various parameter settings and sub-ontology subsets. The PBiLSTM-FCN has better adaptability in the selection of sub-ontology, n-gram and word vector dimension, and has better performance in Indica protein function classification.

Japonica protein function prediction results are shown in Table 8 that the PBiLSTM-FCN's Accuracy above 78 %. Specifically, in the BP dataset of japonica, the PBiLSTM-FCN is better than the other three models in terms of both Accuracy and F1-score. In addition, the high Precision was shown in the PBiLSTM-FCN, but the recall is slightly lower than the other three models. The model has fewer false positive predictions compared other algorithms, but the coverage of actual positive results is slightly lower. In the MF dataset for japonica, the PBiLSTM-FCN performance is better than the other three models in terms of Accuracy, Precision, and F1-score. It is only slightly inferior in terms of recall, especially when n-gram is 1-g and size is 5. In this case, the Accuracy and F1-score of the PBiLSTM-FCN are the highest, at 87.251059 % and 85.377289 %, respectively. On the japonica CC dataset, PBiLSTM-FCN still has excellent performance, achieving the highest scores in both Accuracy and F1-score, and the performance is relatively stable. Compared with the results of other models, it can be found that when n-gram is 1-g and size is 10, the Accuracy of PBiLSTM-FCN in MF dataset is improved by 3.648091 % compared with LSTM, and the F1-score in CC dataset is improved by 4.413416 % compared with CNN. It is worth noting that compared to the BP dataset and CC dataset, all models and different parameter combinations performance is better on the MF dataset in terms of various performance evaluation metrics. This may be due to the fact that the MF dataset of japonica is easier to classify protein functions. In general, the Accuracy and F1-score of PBiLSTM-FCN are significantly superior to CNN, LSTM and FCN under the combination of different sub-ontology subsets, n-gram and size, and could effectively improve the performance of japonica protein function prediction model.

A substantial amount of data was required in the deep learning models to train and learn effectively. Given that Japonica has a large and representative dataset, it was chosen for comparison experiments with the latest models. In previous comparisons with basic models (Table 8), the results showed that the configuration using 2-g and 128-word vectors yielded the best performance. Therefore, this configuration was adopted for the comparison with the latest models. The performance comparisons were made between CNN, LSTM, FCN, DeepGOCNN, DeepFRI, MMSNet, and PBiLSTM FCN on different sub-ontologies (BP, CC, and MF). In order to visually demonstrate the differences between these models on sub-ontologies, multiple types of charts were compared in presentation, and radar charts were ultimately chosen to present the results. In the radar charts, lines closer to the outer edge indicate better

Table 6
Maize protein function prediction results.

Sub-ontology	n	size	Algorithm	Accuracy(%)	Precisioin(%)	Recall(%)	F1(%)	Specificity(%)
BP	1	5	CNN	78.544200	69.484262	66.851200	68.095159	84.655484
			LSTM	78.197051	72.432247	66.265051	69.155263	85.199344
			FCN	77.778116	68.740618	68.172917	68.455446	83.034566
			PBiLSTM-FCN	**79.324042**	74.226171	66.990699	**70.400204**	86.488119
		10	CNN	77.789239	69.061987	67.349360	**68.116635**	83.499010
			LSTM	78.471467	74.208613	64.879823	**69.148248**	86.567809
			FCN	77.799325	68.740169	68.153683	68.445530	83.068672
			PBiLSTM-FCN	**79.550006**	75.080869	66.082800	**70.263879**	87.329809
		15	CNN	77.733140	69.486513	67.131978	68.200125	83.622622
			LSTM	78.873225	73.846241	66.044326	69.708007	86.356302
			FCN	77.788903	68.833927	68.082414	68.455586	83.107915
			PBiLSTM-FCN	**79.304355**	75.418300	65.770018	**70.240933**	87.313834
	2	5	CNN	77.770300	67.710944	66.719982	67.122164	83.500729
			LSTM	79.473625	73.794533	66.896786	70.101109	86.578962
			FCN	77.202792	66.375590	70.886632	67.749596	80.612815
			PBiLSTM-FCN	**80.781856**	70.439344	71.671246	**71.029952**	85.249752
		10	CNN	77.863730	69.338235	67.352899	68.254985	83.637958
			LSTM	79.483164	73.425955	67.346760	70.216837	86.303176
			FCN	77.772870	68.805500	68.080977	68.440491	83.084360
			PBiLSTM-FCN	**81.118251**	70.891927	71.938831	**71.381859**	85.594888
		15	CNN	77.565466	68.726226	66.913484	67.785587	83.379753
			LSTM	79.537081	73.452183	67.919880	70.564885	86.110251
			FCN	77.764336	68.787936	68.126872	68.455610	83.049112
			PBiLSTM-FCN	**81.728627**	71.292109	72.951529	**72.077375**	85.931934
		32	CNN	78.579041	70.866177	65.472565	67.933192	85.593852
			LSTM	80.554165	75.425513	67.861928	71.329337	87.645946
			FCN	77.798410	68.810606	68.098495	68.451887	83.107060
			PBiLSTM-FCN	**82.025695**	71.648687	73.076722	**72.319633**	86.273683
		64	CNN	77.910185	68.001107	68.665615	68.305155	82.824279
			LSTM	79.849080	74.956964	67.402873	70.943732	87.021951
			FCN	77.754791	68.692731	68.226789	68.458032	82.972349
			PBiLSTM-FCN	**82.899403**	73.666631	73.726967	**73.666960**	87.314425
		128	CNN	78.210751	68.569391	67.159357	67.807984	83.966642
			LSTM	80.389522	75.086803	67.706268	71.160333	87.466088
			FCN	77.775899	68.808447	68.063509	68.433102	83.096803
			PBiLSTM-FCN	**83.223933**	74.111709	74.163607	**74.114245**	87.570824
MF	1	5	CNN	82.665433	64.214089	72.643216	67.948973	86.105265
			LSTM	83.184186	72.052242	67.842182	69.846914	89.377378
			FCN	82.395773	63.145929	74.638662	68.410869	85.056750
			PBiLSTM-FCN	**83.298106**	71.313612	68.842869	**70.029768**	89.026284
		10	CNN	82.706426	64.190925	73.365976	68.297554	85.919807
			LSTM	83.415015	72.913298	68.327245	70.485641	89.598004
			FCN	82.385911	63.147251	74.646915	68.416263	85.042978
			PBiLSTM-FCN	**83.878053**	72.329188	69.817783	**71.016353**	89.437956
		15	CNN	82.555668	63.374824	73.548647	67.959453	85.605872
			LSTM	83.226002	73.486198	66.667557	69.817811	90.065132
			FCN	82.367902	63.056389	74.717097	68.390478	84.991073
			PBiLSTM-FCN	**84.837906**	73.323179	70.774505	**72.004601**	90.193716
	2	5	CNN	83.078022	65.921695	70.340334	67.785492	87.467960
			LSTM	83.722096	73.265214	69.484289	71.262167	89.575710
			FCN	82.408243	63.103175	74.650352	68.388727	85.063066
			PBiLSTM-FCN	**86.582262**	74.663636	75.459899	**75.031185**	90.645461
		10	CNN	82.065649	63.647246	73.394158	67.877775	85.139648
			LSTM	83.751892	74.131105	67.862654	70.760154	90.276168
			FCN	82.370532	63.059372	74.665528	68.372146	85.010948
			PBiLSTM-FCN	**86.759393**	74.412395	76.142436	**75.250777**	90.579421
		15	CNN	82.648063	64.830268	73.078981	68.478237	86.021966
			LSTM	84.613220	74.561053	70.987830	72.687351	90.152913
			FCN	82.380595	63.202110	74.609323	68.432525	85.054355
			PBiLSTM-FCN	**87.390558**	76.535178	75.942011	**76.217695**	91.547008
		32	CNN	81.567773	61.933490	74.850617	67.667930	83.916125
			LSTM	84.292080	74.622741	69.627308	71.894036	90.299742
			FCN	82.375057	63.225173	74.566566	68.428014	85.064079
			PBiLSTM-FCN	**87.650529**	76.789647	76.627631	**76.660382**	91.632680
		64	CNN	79.389690	56.942675	79.693187	65.975375	79.285362
			LSTM	84.238127	73.634201	70.201686	71.828354	89.885884
			FCN	82.372394	63.132767	74.677569	68.420704	85.016279
			PBiLSTM-FCN	**88.141047**	76.275187	77.897764	**77.048818**	91.664830
		128	CNN	82.787238	64.312458	71.939112	67.521059	86.465285
			LSTM	84.139045	73.961428	70.069517	71.875871	89.899708
			FCN	82.392933	63.264475	74.566911	68.451548	85.088183
			PBiLSTM-FCN	**88.846801**	77.602347	79.223470	**78.360687**	92.150422

(*continued on next page*)

Table 6 (*continued*)

Sub-ontology	n	size	Algorithm	Accuracy(%)	Precisioin(%)	Recall(%)	F1(%)	Specificity(%)
CC	1	5	CNN	78.185709	77.561683	63.369886	69.742597	87.935672
			LSTM	78.878478	76.218984	64.138203	69.570813	87.837308
			FCN	78.004268	77.529430	63.691499	69.931249	87.610443
			PBiLSTM-FCN	**81.044463**	77.790176	68.589455	**72.887108**	88.413581
		10	CNN	78.174182	78.689051	62.446960	69.403559	88.690901
			LSTM	79.622871	76.178392	65.775754	70.544802	87.821577
			FCN	77.976071	77.526540	63.731980	69.954789	87.564171
			PBiLSTM-FCN	**82.337635**	79.209481	70.559356	**74.610220**	89.203759
		15	CNN	77.625576	76.554856	64.301194	69.868000	86.655079
			LSTM	80.106069	76.866759	66.992312	71.569875	87.946254
			FCN	77.951542	77.461372	63.783798	69.959420	87.496570
			PBiLSTM-FCN	**81.668515**	78.930816	69.206691	**73.727005**	89.053074
	2	5	CNN	78.588359	78.196562	62.473162	69.249997	88.877846
			LSTM	80.484234	70.387438	67.230977	68.753333	86.711021
			FCN	77.993250	77.655068	63.622024	69.940892	87.671382
			PBiLSTM-FCN	**80.644190**	77.782247	67.672825	**72.314798**	88.416944
		10	CNN	78.046020	76.997603	62.451090	68.935483	88.041758
			LSTM	80.702221	70.274279	68.692593	69.455002	86.345757
			FCN	77.960813	77.504458	63.751924	69.957178	87.533596
			PBiLSTM-FCN	**81.098652**	76.896533	67.845640	**72.065323**	88.545769
		15	CNN	78.090689	78.373118	62.870601	69.681254	88.329283
			LSTM	80.926188	71.039886	69.429046	70.210438	86.436015
			FCN	77.978005	77.574789	63.686415	69.946910	87.602175
			PBiLSTM-FCN	**81.351010**	78.465520	67.667439	**72.618083**	89.262595
		32	CNN	77.975808	78.555973	61.903892	68.743670	88.711864
			LSTM	81.545267	71.233233	69.725234	70.424038	87.001683
			FCN	77.993053	77.600098	63.658603	69.941049	87.637114
			PBiLSTM-FCN	**81.609338**	77.060869	69.437193	**72.996525**	88.425443
		64	CNN	77.678972	75.513835	65.536873	70.100321	85.795137
			LSTM	81.717448	72.222206	70.020869	71.092075	87.253457
			FCN	78.007757	77.565465	63.668113	69.932365	87.635747
			PBiLSTM-FCN	**82.225707**	78.393772	69.001048	**73.345516**	89.516133
		128	CNN	77.480268	73.860585	66.164418	69.501986	84.817393
			LSTM	81.668575	72.980674	72.376252	72.647223	86.388762
			FCN	77.981243	77.570671	63.698060	69.953035	87.598538
			PBiLSTM-FCN	**83.118925**	77.161496	69.400850	**73.006987**	89.878377

performance, with the red line representing the PBiLSTM-FCN model. Five evaluation metrics (Accuracy, Precision, Recall, F1 score, and Specificity) were used comprehensively to assess the model performance. The performance comparisons of PBiLSTM-FCN with other latest models based on these five metrics are presented in Figs. 7 to11. The Accuracy evaluation metric is presented in Fig. 7. PBiLSTM-FCN is positioned at the outermost edge on all three sub-ontologies, demonstrating its high accuracy. In contrast, the performance of the DeepGOCNN and DeepFRI is poorer, being closer to the center. The result was indicated that PBiLSTM-FCN has a significant advantage in overall classification accuracy. For Precision (Fig. 8), the line of MMSNet is on the outermost layer, while PBiLSTM-FCN follows closely behind. The MMSNet performs better in reducing false positives, especially on the BP sub-ontology. In the Recall metric (Fig. 9), the CC sub ontology of PBiLSTM-FCN has the best results, while the results of FCN and DeepFRI in the BP sub-ontology are better than PBiLSTM-FCN. On the BP sub-ontology, all relevant instances can be better identified by the FCN and DeepFRI, although more false positives may be included. Regarding F1 Score (Fig. 10), for the CC and MF sub-ontologies, the highest F1 score was achieved by the PBiLSTM-FCN, indicating its better comprehensive performance in these two sub-ontologies. On the BP sub-ontology, the line of DeepFRI is on the outermost layer, the result is superior to the PBiLSTM-FCN. For Specificity (Fig. 11) , the line of PBiLSTM-FCN is located at the outermost edge on all three sub-ontologies, showing its excellent performance in excluding negative examples. Conversely, the yellow line of DeepGOCNN is at the innermost position, indicating its poor performance in specificity. The excellent performance of the PBiLSTM-FCN model was validated by the experiments conducted on the representative Japonica dataset. Overall, the best performance of the PBiLSTM-FCN was shown in accuracy and specificity, making it suitable for protein function prediction. It can correctly predict protein functions while minimizing false negatives.

3.3. The ablation experiment results of the PBILSTM-FCN on different grain types

The ablation experiments aim to analyze and compare the impact of different components of the PBiLSTM-FCN on overall performance, proving the effectiveness and robustness of the PBiLSTM-FCN model. To clearly observe the performance on each type of grain in different sub-ontologies, the ablation experiment results for BP, MF, and CC sub-ontologies are presented in Figs. 12 to 14. Based on the three sub-ontologies, the performances of four grains (Soybean, Maize, Indica, Japonica) are depicted in separate subplots. The performance of different models (LSTM module, FCN module, PBiLSTM-FCN) was illustrated in each subplot by five evaluation metrics (Accuracy, Precision, Recall, F1, Specificity). In Fig. 12 of the BP sub-ontology, the results of the PBiLSTM-FCN model (green) are superior to the LSTM module (blue) and FCN module (orange), particularly in terms of Recall and F1 Score. For Maize and Japonica, the results of the LSTM are notably outstanding in Specificity, while PBiLSTM-FCN shows better performance in Precision and Recall. The advantages of bidirectional LSTM and FCN were combined into the PBiLSTM-FCN, providing a more balanced performance in biological process prediction tasks. On the MF sub-ontology (Fig. 13), the results of the PBiLSTM-FCN model exhibit similarly high performance as observed on the BP sub-ontology. High Recall and F1 scores indicate that the model effectively balances precision and recall. The performance of models differs significantly across various grains and sub-ontologies, indicating that the type of grain has a notable impact on model performance. In some cases, the FCN may have higher Recall values, while in other scenarios, PBiLSTM-FCN shows greater advantages in Specificity. In Fig. 14 of the CC sub-ontology, the

Table 7
Indica protein function prediction results.

Sub-ontology	n	size	Algorithm	Accuracy(%)	Precision(%)	Recall(%)	F1(%)	Specificity(%)
BP	1	5	CNN	86.813007	76.195646	73.225268	74.153398	91.726918
			LSTM	86.808542	75.728316	75.185831	75.359077	91.100766
			FCN	87.085973	77.328716	74.517002	75.896742	91.802280
			PBiLSTM-FCN	**87.886512**	80.230723	72.574561	**76.191182**	93.474184
		10	CNN	87.759726	78.637383	74.986489	76.391424	92.476950
			LSTM	87.041642	77.152861	74.671089	75.891331	91.690400
			FCN	87.143132	77.392401	74.586605	75.963299	91.843669
			PBiLSTM-FCN	**88.100619**	80.288550	73.574784	**76.760845**	93.404507
		15	CNN	86.465999	75.610213	72.655662	73.600835	91.484089
			LSTM	87.070656	77.247928	74.589986	75.895678	91.754760
			FCN	87.098390	77.371218	74.494212	75.904170	91.826712
			PBiLSTM-FCN	**88.290140**	80.811683	73.578366	**76.978780**	93.643124
	2	5	CNN	87.178381	78.239075	72.980179	75.359134	92.455560
			LSTM	86.981494	77.083805	74.635365	75.838716	91.635637
			FCN	87.091368	77.319063	74.532158	75.899924	91.801125
			PBiLSTM-FCN	**87.549469**	77.787507	74.682741	**76.132581**	92.234617
		10	CNN	87.765650	79.519203	73.693228	76.323212	92.974513
			LSTM	87.335576	77.283909	75.130125	76.188071	91.843650
			FCN	87.148388	77.465588	74.476114	75.940099	91.891343
			PBiLSTM-FCN	**87.995726**	78.807674	74.120630	**76.369013**	92.922527
		15	CNN	87.071601	78.454445	72.868940	75.341547	92.438295
			LSTM	87.071556	77.256870	74.572213	75.890669	91.762033
			FCN	87.149715	77.625135	74.203959	75.859710	91.994891
			PBiLSTM-FCN	**87.151056**	77.400067	74.570977	**75.958646**	91.856213
		32	CNN	86.832136	77.035232	74.522094	75.699403	91.527111
			LSTM	87.025848	77.131051	74.678054	75.884145	91.670810
			FCN	87.109247	77.343232	74.568550	75.930389	91.810735
			PBiLSTM-FCN	**87.690141**	78.957840	74.097752	**76.406013**	92.708848
		64	CNN	85.329870	73.882428	76.013471	74.204382	89.106686
			LSTM	86.968765	76.905244	74.876839	75.864612	91.525847
			FCN	87.099311	77.351768	74.530749	75.914615	91.814033
			PBiLSTM-FCN	**87.506584**	78.235795	74.458943	**76.276943**	92.334900
		128	CNN	87.135499	76.991288	71.780925	74.026896	92.502232
			LSTM	86.998262	77.120983	74.643394	75.861576	91.654494
			FCN	87.093572	77.348392	74.495725	75.894164	91.818189
			PBiLSTM-FCN	**88.055110**	79.344762	74.746799	**76.908342**	92.906585
MF	1	5	CNN	84.400349	63.316945	55.832941	58.198222	91.726918
			LSTM	84.035105	63.148672	64.276698	63.510023	91.100766
			FCN	84.595705	64.214035	62.925509	63.562421	91.802280
			PBiLSTM-FCN	**84.678555**	64.380486	63.402942	**63.739644**	93.474184
		10	CNN	84.812000	64.155434	62.500741	63.256320	92.476950
			LSTM	84.526326	64.275036	62.974760	63.618073	91.690400
			FCN	84.583172	64.658708	62.720694	63.672179	91.843669
			PBiLSTM-FCN	**85.105873**	67.731112	61.218865	**64.256407**	93.404507
		15	CNN	84.815156	66.637185	60.455162	63.025971	91.484089
			LSTM	83.513691	60.111107	62.862815	61.022675	91.754760
			FCN	84.582509	64.348120	62.922482	63.626541	91.826712
			PBiLSTM-FCN	**85.124825**	66.682912	61.112580	**63.721936**	93.643124
	2	5	CNN	84.174061	62.867263	59.407548	60.849522	92.455560
			LSTM	84.421133	63.306432	63.764190	63.423118	91.635637
			FCN	84.565529	64.305721	62.897281	63.590117	91.801125
			PBiLSTM-FCN	**84.585584**	64.349641	62.907350	**63.620313**	92.234617
		10	CNN	84.342157	65.498072	61.105408	62.837149	92.974513
			LSTM	84.472833	64.183612	63.041051	63.605063	91.843650
			FCN	84.587389	63.784846	63.353977	63.459663	91.891343
			PBiLSTM-FCN	**84.791360**	64.506857	62.787906	**63.634142**	92.922527
		15	CNN	83.380128	61.023311	61.606537	61.101429	92.438295
			LSTM	84.517011	64.281354	62.937435	63.602224	91.762033
			FCN	84.591944	64.337975	62.934747	63.628201	91.994891
			PBiLSTM-FCN	**84.625506**	64.174885	63.862855	**63.949578**	91.856213
		32	CNN	84.580382	64.348804	62.908689	63.620585	91.527111
			LSTM	84.549639	64.323976	62.924489	63.616490	91.670810
			FCN	84.590876	64.217719	62.964521	63.411352	91.810735
			PBiLSTM-FCN	**84.650897**	64.409227	62.875101	**63.632406**	92.708848
		64	CNN	82.367710	58.520908	68.990773	63.031028	89.106686
			LSTM	84.245732	62.999323	60.171101	61.437791	91.525847
			FCN	84.546491	65.624621	61.507766	63.419576	91.814033
			PBiLSTM-FCN	**84.570113**	64.322195	62.922650	**63.614581**	92.334900
		128	CNN	81.981558	57.270568	66.233084	59.908327	92.502232
			LSTM	84.531986	64.282330	62.959250	63.613859	91.654494
			FCN	84.578221	64.283376	62.933801	63.600314	91.818189
			PBiLSTM-FCN	**84.745584**	66.047824	61.947255	**63.735461**	92.906585

(*continued on next page*)

Table 7 (*continued*)

Sub-ontology	n	size	Algorithm	Accuracy(%)	Precision(%)	Recall(%)	F1(%)	Specificity(%)
CC	1	5	CNN	81.289218	76.329920	79.026144	77.644600	91.726918
			LSTM	79.404355	71.412619	76.738163	73.874248	91.100766
			FCN	81.314468	78.272956	77.405237	77.809771	91.802280
			PBiLSTM-FCN	**81.368627**	76.647768	79.079438	**77.843791**	93.474184
		10	CNN	81.450292	76.066319	79.382545	77.659109	92.476950
			LSTM	81.271592	76.420526	79.247822	77.808393	91.690400
			FCN	81.317437	76.687459	79.031272	77.841100	91.843669
			PBiLSTM-FCN	**81.542032**	77.611176	78.196506	**77.885501**	93.404507
		15	CNN	81.175849	74.197874	80.014286	76.857104	91.484089
			LSTM	81.307731	76.507547	79.198285	77.829660	91.754760
			FCN	81.316317	76.701343	79.021093	77.843334	91.826712
			PBiLSTM-FCN	**81.568010**	78.221837	77.896436	**78.053146**	93.643124
	2	5	CNN	80.850484	74.642725	78.089993	76.269443	92.455560
			LSTM	81.311403	76.621451	79.073534	77.827650	91.635637
			FCN	81.327935	76.830659	78.906090	77.852372	91.801125
			PBiLSTM-FCN	**81.411149**	78.103882	77.845025	**77.961109**	92.234617
		10	CNN	80.781551	74.464879	78.636537	76.472733	92.974513
			LSTM	81.307856	76.503804	79.207096	77.831962	91.843650
			FCN	81.013089	75.278577	79.794867	77.342705	91.891343
			PBiLSTM-FCN	**81.381332**	77.654114	78.208283	**77.881364**	92.922527
		15	CNN	80.958337	75.392109	79.028912	77.151318	92.438295
			LSTM	81.299173	76.509282	79.189888	77.826411	91.762033
			FCN	81.324585	76.964362	78.567572	77.717818	91.994891
			PBiLSTM-FCN	**81.333566**	76.684527	79.039280	**77.842329**	91.856213
		32	CNN	81.448806	78.898521	76.374713	77.452698	91.527111
			LSTM	81.305654	76.538250	79.175352	77.834306	91.670810
			FCN	81.318743	76.759440	78.976529	77.852111	91.810735
			PBiLSTM-FCN	**81.684080**	77.128250	78.731746	**77.864345**	92.708848
		64	CNN	81.269358	76.848526	78.834403	77.813892	89.106686
			LSTM	81.240340	76.364856	79.249968	77.779405	91.525847
			FCN	81.318563	76.697456	79.030111	77.845122	91.814033
			PBiLSTM-FCN	**81.532665**	77.251850	78.654962	**77.935799**	92.334900
		128	CNN	81.278522	74.250585	79.499240	76.398860	92.502232
			LSTM	81.109007	76.047210	78.487507	77.246731	91.654494
			FCN	81.335143	76.835224	78.914511	77.859888	91.818189
			PBiLSTM-FCN	**81.665445**	77.667135	78.478301	**78.040732**	92.906585

results of the FCN module outperform the PBiLSTM-FCN in terms of Recall, indicating that FCN may have a stronger ability to identify positive examples. The stable performance was shown in the LSTM, but is generally slightly inferior to the results of PBiLSTM-FCN. The performance of Soybean and Maize grains remains relatively stable across different models, with PBiLSTM-FCN still demonstrating high performance by five evaluation metrics.

Overall, the best comprehensive performance across the four types of grains was demonstrated, indicating high robustness and generalization capability of the PBiLSTM-FCN. The effectiveness of combining LSTM and FCN was confirmed in the PBiLSTM-FCN model. Additionally, the influence of different sub-ontologies on model performance highlights their importance, providing valuable insights for future research.

3.4. The comparison results and analysis of PBILSTM-FCN prediction function and real protein function

To further enhance the predictive performance of the model and improve the interpretability of the results, The protein from each of four important cereal proteins was selected for analysis. By comparing the predicted protein functions from PBILSTM-FCN with the known functions in the SwissProt database, we aim not only to optimize the accuracy of our model but also to reveal the biological significance of these proteins through enhanced interpretability. This approach ensures that our research efforts are scientifically rigorous and provide valuable insights for practical applications. The examples of protein function prediction results for four types of grains were shown in Table 9.

First of all, by observing the predicted results of soybean protein function and the real functions in the database, it can be found that in the SwissProt database, the Q07185 (AOX1_SOYBN) protein has been confirmed to have five protein functional categories: GO: 0070469, GO: 0043229, GO: 005739, GO: 0016020, and GO: 005743. However, the GO: 0005743 function was not successfully predicted in the prediction results of this experiment. On the gene_ontology website, corresponding protein functions can be found based on GO annotations (https://www.informatics.jax.org/vocab/gene_ontology). This study used soybean to examine mitochondrial gene expression, demonstrating the importance of mitochondria in soybean (Manavski et al., 2025). The GO:0005743 function is the mitochondrial inner membrane, which mainly refers to the interior of the mitochondrial envelope, the lumen facing lipid bilayer. It is highly folded to form cristae. Further research on the definitions of the other four protein functions reveals that the GO:0005739 refers to the mitochondrion, and the GO:0016020 function is defined as the membrane, which are all related to the Inner mitochondrial membrane of the GO: 0005743 function. Therefore, the failure to correctly predict the GO:0005743 function may be attributed to the fact that the model in this experiment categorizes GO:0005743 as separate mitochondrial and membrane functions. But fails to more accurately regard the Inner mitochondrial membrane composed of mitochondria and membrane as the function of Q07185 (AOX1_SOYBN) protein.

Similarly, the functional annotations of the maize dataset's K7U9N8 (OP1_MAIZE) protein were shown for GO:0030050, GO:0030048, and GO:0007015 in the SwissProt database for its biological process classification. However, the protein functions obtained by using PBiLSTM-FCN only have GO:0030050 function and GO:0030048 function, and the GO:0007015 function is missing. The study has shown the importance of actin was shown in the study in plant growth and development (Lv et al., 2024). The GO:0007015 function represents actin filament organization, which is defined as a process occurring at the cellular level. It includes processes that control the spatial distribution of actin filaments, such as transforming actin filament structures into mesh works, bundles, or other structures through cross-linking. The

Table 8
Japonica protein function prediction results.

Sub-ontology	n	size	Algorithm	Accuracy(%)	Precision(%)	Recall(%)	F1(%)	Specificity(%)
BP	1	5	CNN	79.041284	66.492925	64.359168	65.384455	91.716196
			LSTM	78.734515	67.069210	64.387518	65.623113	89.543257
			FCN	79.003351	66.324266	65.203680	65.757526	90.479128
			PBiLSTM-FCN	**80.128164**	71.187949	61.762326	**66.121567**	90.462909
		10	CNN	79.204480	67.472462	64.106713	65.674252	90.732943
			LSTM	78.979456	66.225108	65.289364	65.752932	90.423044
			FCN	78.636537	65.213901	66.196822	65.626593	90.586434
			PBiLSTM-FCN	**80.309621**	71.041370	62.488964	**66.463998**	91.811888
		15	CNN	79.236169	67.418772	63.675182	65.294068	91.586496
			LSTM	79.003997	66.417790	65.113797	65.758563	88.972723
			FCN	78.921480	66.103142	65.450315	65.772392	90.490705
			PBiLSTM-FCN	**79.982223**	70.454589	62.421996	**66.173720**	91.673152
	2	5	CNN	78.286999	64.020337	63.269576	63.633649	90.722359
			LSTM	78.968323	67.447140	63.941884	65.544442	90.006607
			FCN	78.955983	66.365633	65.155167	65.753583	90.476338
			PBiLSTM-FCN	**79.143090**	66.226661	65.306230	**65.762252**	90.496781
		10	CNN	78.926187	65.384670	65.250893	65.237200	90.906681
			LSTM	78.906183	66.337537	65.183382	65.755240	90.351148
			FCN	78.933304	66.320755	65.339473	65.722163	90.308365
			PBiLSTM-FCN	**78.980223**	66.322446	65.256828	**65.783586**	90.708918
		15	CNN	78.382504	63.561181	62.274140	62.755919	89.300575
			LSTM	78.991765	66.255453	65.252237	65.749462	90.425194
			FCN	78.951346	68.160965	63.601238	65.743583	90.493729
			PBiLSTM-FCN	**79.334115**	66.101268	65.447147	**65.771380**	90.283989
MF	1	5	CNN	78.888679	65.394120	65.801467	65.210170	90.492263
			LSTM	78.983340	66.257117	65.285164	65.766974	90.460590
			FCN	78.850393	66.203692	65.281987	65.737387	90.467108
			PBiLSTM-FCN	**79.388060**	67.931472	64.042952	**65.869138**	90.564657
		10	CNN	79.009537	65.862326	64.742753	65.296597	86.155081
			LSTM	78.281128	63.522958	64.566487	63.823770	90.629356
			FCN	78.972508	66.221527	65.254596	65.733766	90.989569
			PBiLSTM-FCN	**79.426808**	67.580299	64.612258	**66.006134**	90.476814
		15	CNN	78.178490	61.420696	64.539006	62.647305	86.336100
			LSTM	78.956137	66.358924	65.170737	65.758835	90.432187
			FCN	79.002347	66.288924	65.276508	65.775559	90.474500
			PBiLSTM-FCN	**79.340080**	69.295369	62.716729	**65.780635**	91.110129
	2	5	CNN	83.682173	74.612952	88.077381	80.754571	82.871028
			LSTM	83.636044	74.013386	88.501420	80.610379	81.049141
			FCN	83.630570	73.998266	88.509853	80.593902	84.190817
			PBiLSTM-FCN	**87.251059**	83.890997	86.960091	**85.377289**	82.985270
		10	CNN	84.206087	75.152537	87.924893	80.999617	82.868532
			LSTM	83.423515	75.100525	86.971230	80.372261	82.703174
			FCN	83.603123	73.835720	88.616518	80.551093	82.940753
			PBiLSTM-FCN	**87.071606**	83.211200	87.178821	**85.117954**	83.925969
		15	CNN	84.240610	77.128004	87.174076	81.617532	81.930185
			LSTM	83.639002	74.054335	88.472532	80.623237	82.799715
			FCN	83.590890	73.849681	88.599921	80.544990	82.946950
			PBiLSTM-FCN	**87.174025**	82.822284	87.823214	**85.228270**	84.236650
	1	5	CNN	83.976007	75.479854	87.674757	81.047194	82.655337
			LSTM	82.793734	72.528755	87.037078	79.114802	82.897721
			FCN	83.567572	73.748433	88.663190	80.513569	83.052798
			PBiLSTM-FCN	**85.208944**	78.062078	88.007587	**82.675046**	84.022059
		10	CNN	83.303524	74.665169	86.959113	80.285634	82.197659
			LSTM	83.530979	73.411786	88.864610	80.371935	82.793738
			FCN	83.659786	74.349034	88.264378	80.709510	81.856558
			PBiLSTM-FCN	**85.091786**	77.484505	88.430773	**82.536508**	83.682018
		15	CNN	83.170238	76.148501	85.111219	80.155083	82.283506
			LSTM	83.260262	74.172490	86.897363	79.982972	82.792014
			FCN	83.687522	74.285664	88.325552	80.695611	83.284367
			PBiLSTM-FCN	**85.018665**	77.196811	88.684353	**82.519912**	82.960346
	2	5	CNN	83.714357	74.393148	88.290815	80.737223	85.140114
			LSTM	83.667627	74.237641	88.351059	80.681515	82.814138
			FCN	83.667709	74.224966	88.360509	80.672765	82.985362
			PBiLSTM-FCN	**85.217890**	79.145270	87.188424	**82.939250**	83.740185
		10	CNN	83.712645	74.059509	88.622413	80.687557	83.011149
			LSTM	83.471569	73.510705	88.772198	80.388306	82.648366
			FCN	83.641764	74.038620	88.485605	80.613870	82.944116
			PBiLSTM-FCN	**85.726523**	79.612037	87.862734	**83.503503**	83.573583
		15	CNN	82.955666	72.368094	87.730557	79.275100	82.413515
			LSTM	83.641605	74.245885	88.252456	80.639323	82.920179
			FCN	83.652742	74.115925	88.433819	80.639677	83.058791
			PBiLSTM-FCN	**85.867983**	80.135982	87.610547	**83.610528**	83.934473

(*continued on next page*)

Table 8 (*continued*)

Sub-ontology	n	size	Algorithm	Accuracy(%)	Precision(%)	Recall(%)	F1(%)	Specificity(%)
CC	1	5	CNN	78.991195	77.076006	78.614477	77.826950	85.572647
			LSTM	79.078131	78.192843	78.031012	78.111426	85.372649
			FCN	79.079218	78.013372	78.225033	78.117527	85.180475
			PBiLSTM-FCN	**80.556895**	79.961107	79.133017	**79.535170**	88.544616
		10	CNN	78.327782	75.600981	76.373478	75.967467	86.029422
			LSTM	79.083232	78.075645	78.159619	78.117128	85.103885
			FCN	79.087740	77.828782	78.418170	78.118863	84.202081
			PBiLSTM-FCN	**81.378459**	80.620710	80.162339	**80.380883**	88.414741
		15	CNN	78.629729	76.568621	76.931008	76.737658	86.209714
			LSTM	78.926530	77.745448	77.490127	77.155846	85.233851
			FCN	79.083962	77.816944	78.423652	78.118991	84.958744
			PBiLSTM-FCN	**80.423801**	79.288948	79.567192	**79.417898**	88.019253
	2	5	CNN	78.631436	76.550655	76.966676	76.727244	84.734301
			LSTM	79.079110	77.982577	78.246838	78.113973	85.855408
			FCN	79.086063	78.124584	78.093308	78.106321	85.158759
			PBiLSTM-FCN	**79.484033**	77.915216	78.988711	**78.440611**	85.265050
		10	CNN	78.874401	76.713993	78.580422	77.543757	84.902790
			LSTM	78.707866	78.497065	76.502023	77.371527	85.091413
			FCN	79.072954	78.154028	78.059121	78.103823	85.055775
			PBiLSTM-FCN	**79.949582**	78.329016	79.618328	**78.961567**	85.136084
		15	CNN	78.468380	75.539089	77.807758	76.576999	85.103734
			LSTM	79.080807	78.080171	78.141224	78.110057	85.136624
			FCN	79.082968	78.271922	77.925666	78.095449	86.121889
			PBiLSTM-FCN	**80.018575**	78.597174	79.368938	**78.960845**	85.382271
		32	CNN	79.074312	77.489448	78.191415	77.801606	84.655661
			LSTM	79.084452	78.016328	78.233094	78.124016	85.115694
			FCN	79.075018	78.359906	77.841421	78.092922	84.969838
			PBiLSTM-FCN	**79.951223**	78.747649	79.147258	**78.942324**	86.327650
		64	CNN	79.079313	78.074932	78.146960	78.110272	85.271359
			LSTM	79.079542	78.144583	78.087849	78.115701	84.145691
			FCN	79.075992	78.231161	77.989676	78.108946	85.109398
			PBiLSTM-FCN	**80.059753**	78.468588	79.734132	**79.086030**	86.079936
		128	CNN	79.076529	78.156808	78.058552	78.107113	83.672181
			LSTM	79.082850	77.980293	78.260634	78.119786	85.151621
			FCN	79.076745	78.218399	78.014943	78.115024	85.144670
			PBiLSTM-FCN	**80.085478**	78.749999	79.328485	**79.010658**	87.072886

GO:0030050 function represents the movement of vesicles along actin filaments mediated by motor proteins, and the GO:0030048 function refers to the process of movement of organelles or other particles along actin filaments or the dynamic processes between actin filaments mediated by motor proteins. The GO:0030050 function and GO:0030048 function can be summarized as the processes involving the movement of a certain structure along the actin filaments represented by the GO:0007015 function. The GO:0007015 function has a broader definition, while the GO:0030050 function and GO:0030048 function provide more specific details within the larger scope of the GO:0007015. Therefore, the failure to predict the function of the GO:0007015 function in this experiment may be due to the fact that the experimental model focused on more precise and specific protein functions, while overlooking the broader definition of protein functions.

Subsequently, examining the actual situation of Indica protein function prediction. The experimental results were shown that there are GO:0003676, GO:0003724, GO:0005524, GO:0000166, GO:0003723, GO:0016787, GO:0004386, GO:0016887 and GO:0140657 proteins in the molecular function classification of the A2XKG2 (RH10_ORYSI) protein function. The acid binding, RNA helicase activity, ATP binding, nucleotide binding, RNA binding, hydrolase activity, helicase activity and ATP hydrolysis activity were represented, respectively. However, in the SwissProt database, the A2XKG2 (RH10_ORYSI) protein does not have the function of GO:0140657. The ATP of indica has been shown to have biological efficacy against urolithiasis in mice (Sathya, Kokilavani, Teepa, & Balakrishnan, 2011). The definition of GO: 0140657 function represents the activity dependent on ATP, which is characterized by coupling ATP hydrolysis with other steps in the reaction mechanism to provide energy advantage for the reaction, such as catalyzing a reaction or facilitating transport against a concentration gradient. Obviously, the function of GO:0140657 is closely related to the hydrolase activity represented by the function of GO:0016787 and the ATP hydrolysis activity represented by the function of GO:0016887, which leads to the speculation of the A2XKG2 (RH10_ORYSI) protein that the function of GO:0140657 may exist in the PBiLSTM-FCN model.

Finally, in the japonica protein dataset, Q9AV71 (CESA7_ORYSJ) protein was found to have the functions GO:0016759, GO:0016760, GO:0016740, GO:0016757, and GO:0046872 through functional annotation verification in the SwissProt database. These functions are defined as cellulose synthase activity, cellulose synthase (UDP-forming) activity, transferase activity, glycosyltransferase activity, and metal ion binding, respectively. The experimental results of this study not only correctly predicted the five protein functions mentioned above, but also predicted the GO:0016758 function, which represents hexyltransferase activity, specifically referring to the catalysis of the transfer of a hexyl group from one compound (donor) to another compound (acceptor). Transferase is one of the key enzymes in Japan (Li et al., 2019). Similarly, the GO:0016757 function representing glycosyltransferase activity specifically refers to the catalysis of the transfer of a glycosyl group from one compound (donor) to another compound (acceptor). Obviously, the two functions of GO:0016757 and GO:0016758 belong to the transferase activity represented by the GO:0016740 annotation. Therefore, this paper speculates that the Q9AV71 (CESA7_ORYSJ) protein has the function of GO:0016758.

In the analysis of four types of grain proteins, mismatches between predicted and actual protein functions were categorized into two types for handling. If the model predicts a protein function that does not actually exist (false positives), these proteins will be provided to laboratory researchers to help them identify potential new functions. If an actual protein function known in the SwissProt database is not predicted by the model (false negatives), we will compare the original function from the SwissProt database with the predicted function, analyze the

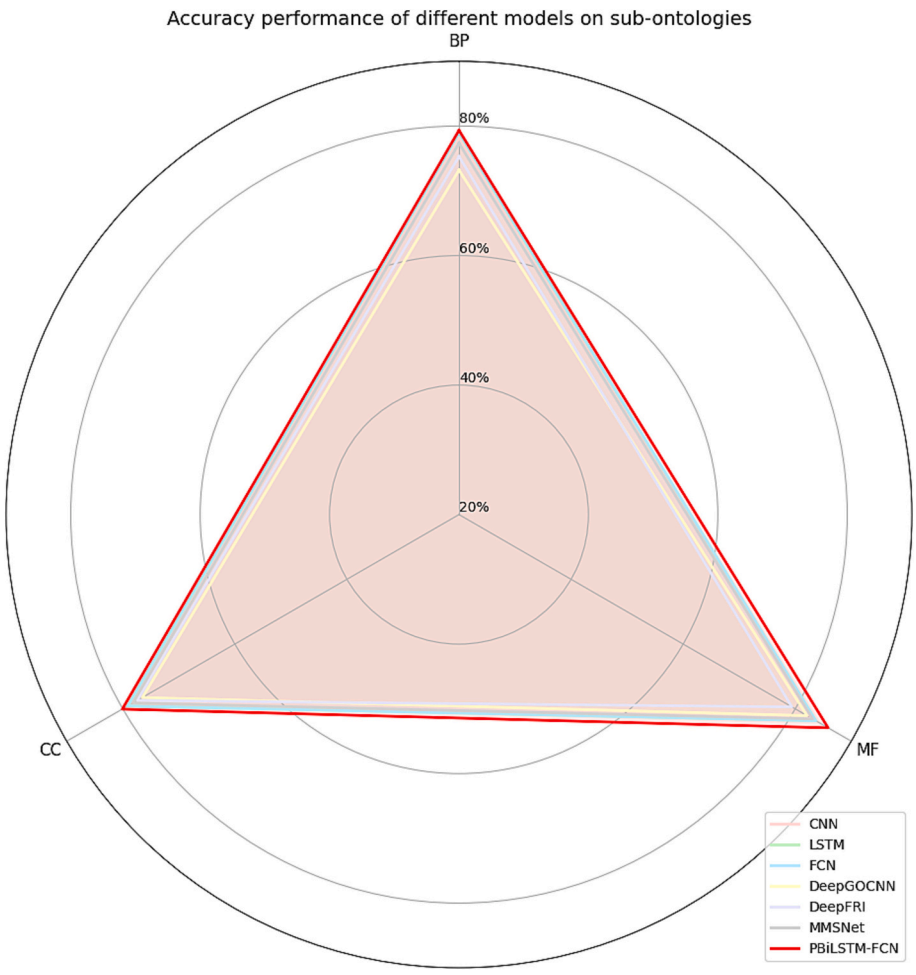


Fig. 7. The Accuracy performance of different models on sub-ontologies.

relationship between the two, and determine whether further refinement of the model is needed to improve its predictive performance, ensuring it can accurately predict these missing functions in the future. Through this approach, we not only enhance the accuracy of the model but also improve the interpretability and biological significance of the results.

3.5. Discussion

In the study, experiments were conducted on three sub-ontologies of four grains type. The results of PBiLSTM-FCN were obtained on accuracy, precision, recall, F1 and specificity metrics in Section 3.1. To demonstrate the superiority of PBiLSTM FCN over existing models, comparisons were made with basic models (CNN, LSTM, and FCN) in Section 3.2. At the same time, in order to observe the influence of amino acid composition and word vector composition on the model, a comparison was made between different amino acid compositions and word vector compositions. Based on the experimental results of the Section 3.2, it can be found that the values of n-gram and size are not necessarily related to the optimal results. Whether it is 1-g or 2-g, and whether the size is 5 or 128, both accuracy and F1 have the potential to achieve the highest score. However, it is possible that the accuracy and F1 can obtain higher scores when 3-g, 4-g, or size are other values, which requires future research to further explore the optimal n-gram and size for each model. On the Japanese dataset, comparisons were also made with the latest models (DeepGO CNN, DeepFRI, and MMSNet). The superior performance in grain protein function prediction experiments were demonstrated in the PBiLSTM-FCN model, effectively leveraging protein feature information to enhance the robustness and generalization capability of the protein function prediction model. This confirms the effectiveness of the PBiLSTM-FCN model. In Section 3.3, a comparison was made between the protein functions predicted incorrectly and the actual protein functions of four grain proteins for interpretability analysis. The interpretability analysis of results not only enhances the accuracy of the model but also discovers potential GO annotations based on existing protein functions. For example, by performing GO annotation predictions on the protein dataset of japonica, specifically for Q9AV71 (CESA7_ORYSJ), the predictive model confirmed the existing GO annotations of Q9AV71. Furthermore, Q9AV71 is infered to have annotation GO: 0016758, providing new insights into the role of this protein in cellular metabolism. In general, the most protein functions could be accurately predicted by the PBiLSTM-FCN model, but it may be difficult to accurately distinguish extremely similar functions. In addition, the unconfirmed GO functional annotations in the SwissProt database could be predicted in the PBiLSTM-FCN algorithm model, which provides some assistance for using biological experimental methods to determine the true function of proteins in the future.

In grain protein prediction, there are many factors that affect the prediction accuracy of the model. Firstly, the encoding method used to convert protein sequences into numerical representations is crucial for the prediction results. High-quality input data can significantly enhance prediction performance. In the experiments, various encoding methods (such as One-Hot encoding, ESM embeddings, and PSSM matrices) were employed and verified the effectiveness of the Embedding layer through multiple experiments. Secondly, the number of convolutional layers in the FCN and the number of BiLSTM layers are also important for the

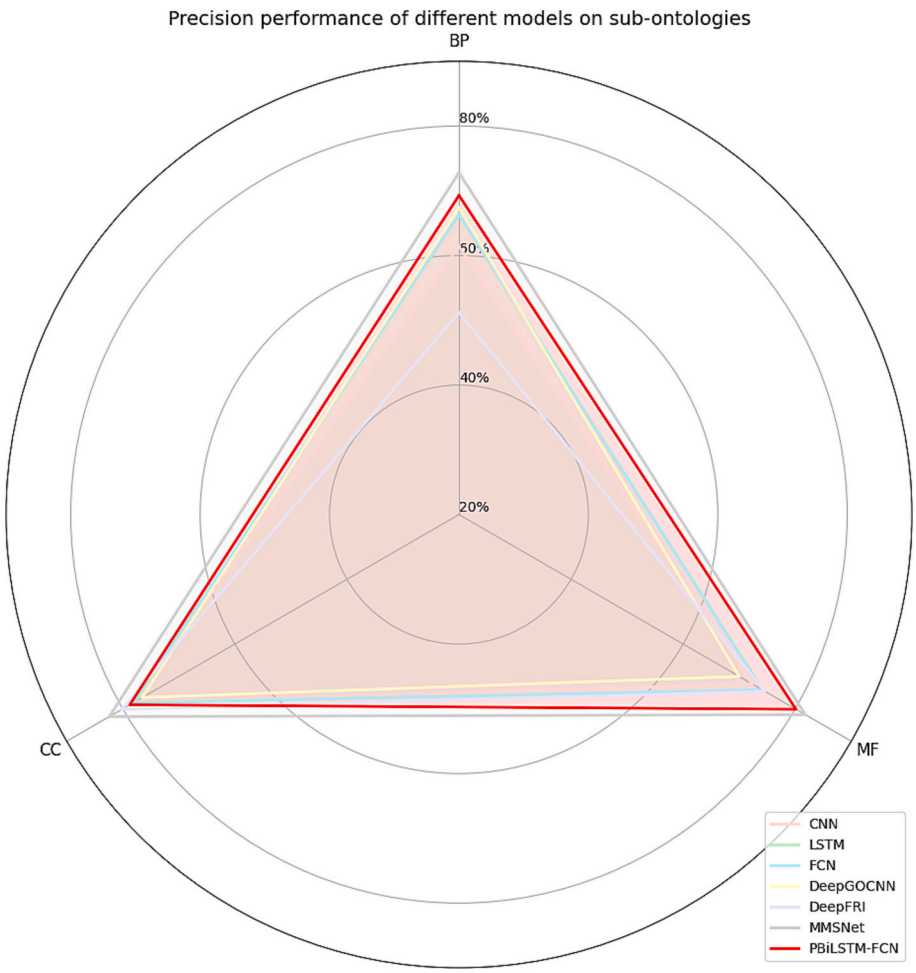


Fig. 8. The Precision performance of different models on sub-ontologies.

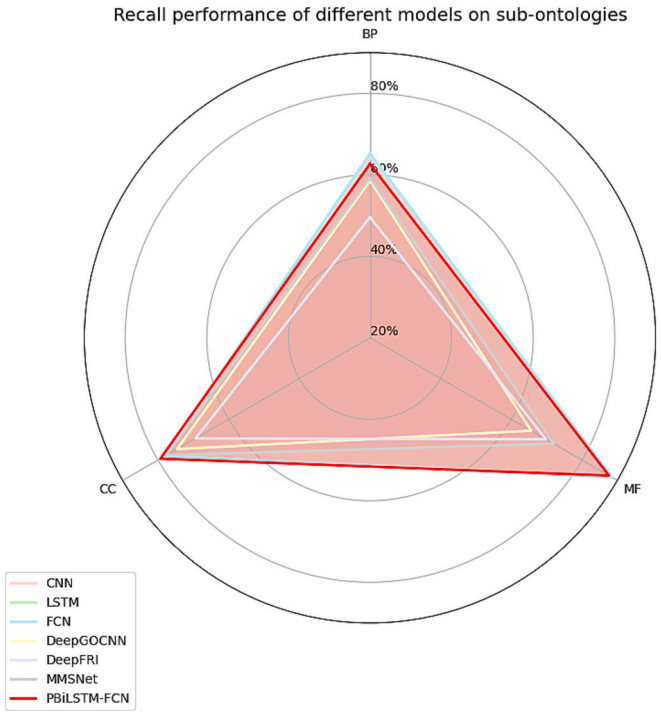


Fig. 9. The Recall performance of different models on sub-ontologies.

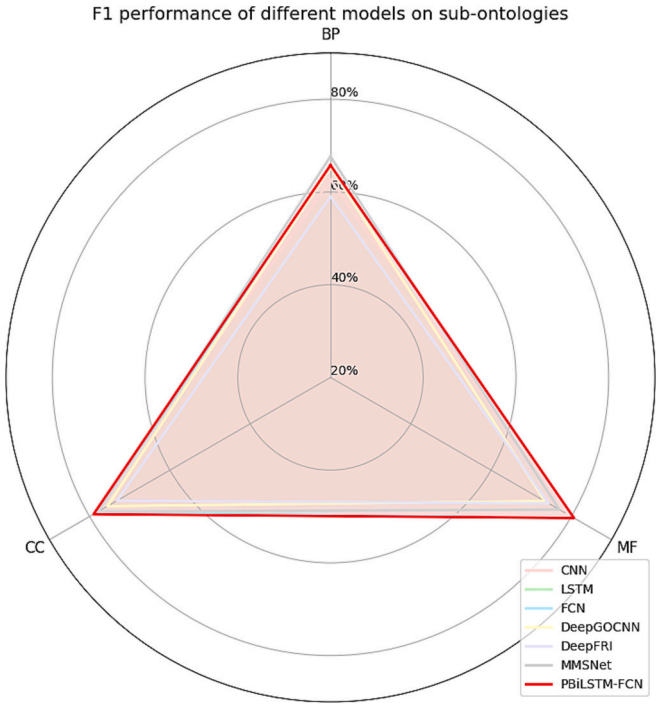


Fig. 10. The F1 performance of different models on sub-ontologies.

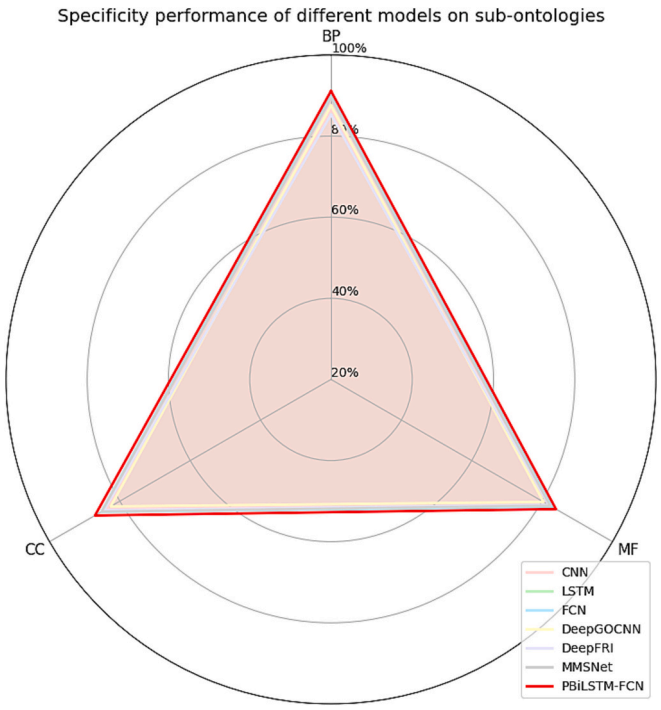


Fig. 11. The Specificity performance of different models on sub-ontologies.

model's composition. Overly deep or wide networks can lead to overfitting or vanishing gradient problems. Excessive BiLSTM layers can make the model difficult to train and cause it to memorize details in the training data rather than generalize to new data. Through multiple experiments, the optimal model—PBiLSTM-FCN was determined and its effectiveness was validated using ablation experiments. Finally, the impact of data quantity on model accuracy cannot be overlooked. The four types of grains we selected have ample data, but grains like oats have relatively smaller datasets, which can easily lead to overfitting in deep learning models. For grains with large amounts of data, the model can learn features well and generalize to new data.

In summary, the superior performance in protein function prediction was demonstrated in the PBiLSTM-FCN model, enhancing the robustness and generalization capability of the model. It aids in determining the true functions of proteins using biological experimental methods, particularly providing assistance for unconfirmed GO function annotations in the SwissProt database.

4. Conclusion

Grain protein function prediction is a significant subject because there is an enormous amount of unannotated grain protein. In previous protein predictions, the long-range dependencies between amino acids were difficulty captured when processing long protein sequences. And the order of amino acid sequence is not considered by the existing models. When the amino acids in the sequence change, existing models may struggle to adapt effectively to these variations. To address these issues, the PBiLSTM-FCN was proposed in order to overcome the problem of amino acid sequence order and efficiently capture long-range relationships. Four grain proteins—Soybean, Maize, Indica, and Japonica—were utilized as experimental datasets to predict the protein functions. The comparison experiments were conducted based on variations in the amino acid composition and the size of the word vector. At the same time, comparisons were made with the basic models (CNN, LSTM, FCN) and the latest models (DeepGOCNN, DeepFRI, and MMSNet). The comparisons results were showed that the PBiLSTM-FCN performed noticeably better than other existing models. Additionally, the interpretability analyses are conducted by comparing the actual protein functions with those predicted by the PBiLSTM-FCN. The auxiliary support is provided for identifying potential drug targets and

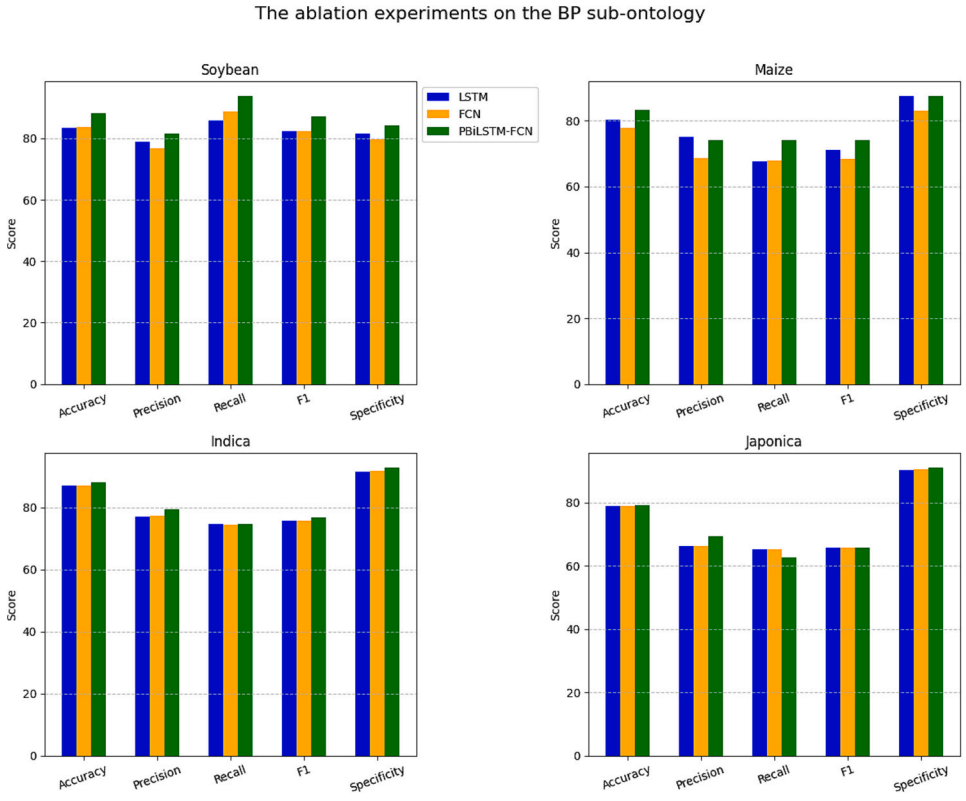


Fig. 12. The ablation experiments on the BP sub-ontology.

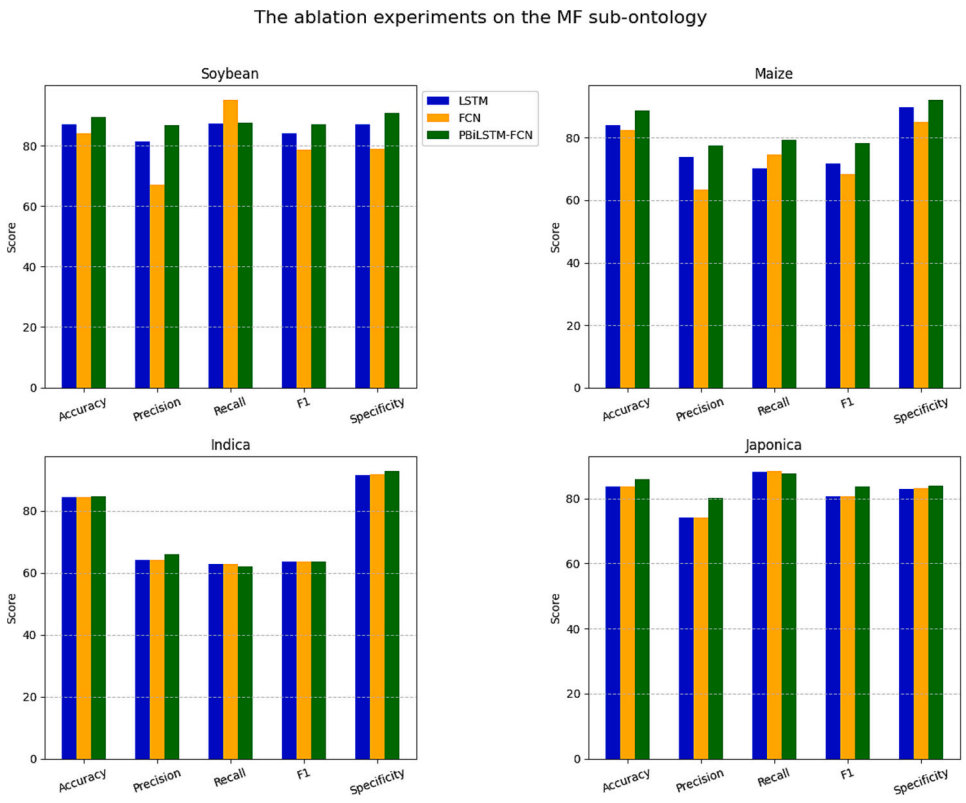


Fig. 13. The ablation experiments on the MF sub-ontology.

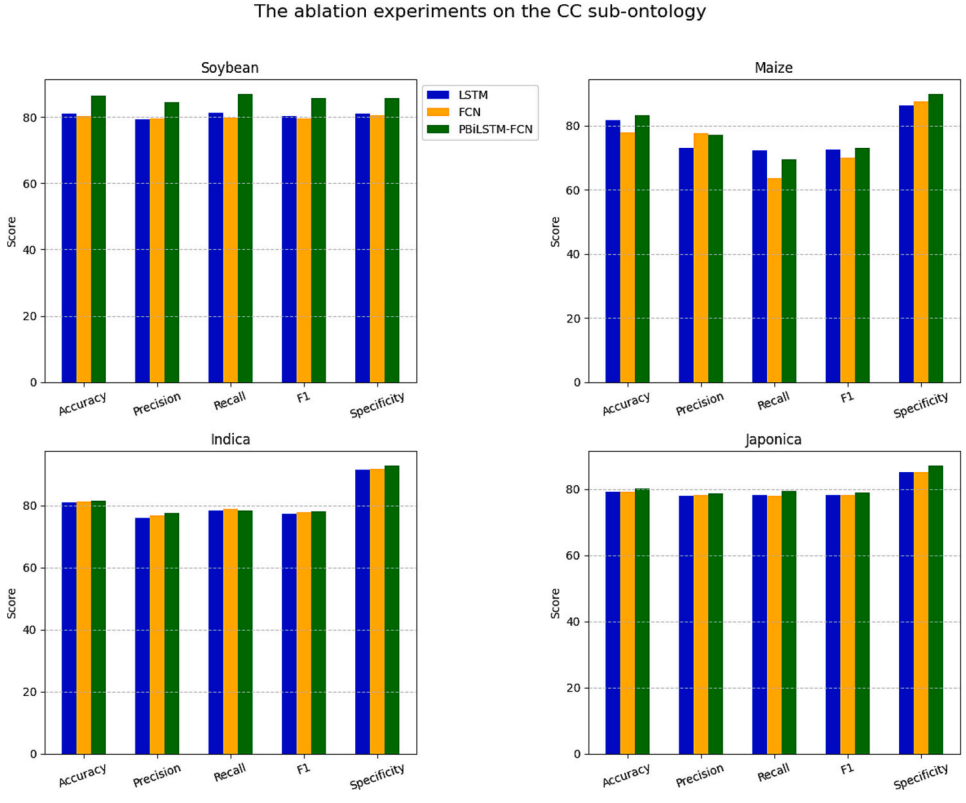


Fig. 14. The ablation experiments on the CC sub-ontology.

Table 9
Examples of protein function prediction results for four types of grains.

Datasets	Protein	Real Function	Predicted function
Soybeans	Q07185(AOX1_SOYBN)	GO:0070469	GO:0070469
		GO:0043229	GO:0043229
		GO:0005739	GO:0005739
		GO:0016020	GO:0016020
		GO:0005743	
Maize	K7U9N8(OP1_MAIZE)	GO:0030050	GO:0030050
		GO:0030048	GO:0030048
		GO:0007015	
Indica	A2XKG2(RH10_ORYSI)	GO:0003676	GO:0003676
		GO:0003724	GO:0003724
		GO:0005524	GO:0005524
		GO:0000166	GO:0000166
		GO:0003723	GO:0003723
		GO:0016787	GO:0016787
		GO:0004386	GO:0004386
		GO:0016887	GO:0016887
			GO:0140657
Japonica	Q9AV71(CESA7_ORYSJ)	GO:0016759	GO:0016759
		GO:0016760	GO:0016760
		GO:0016740	GO:0016740
		GO:0016757	GO:0016757
		GO:0046872	GO:0046872
			GO:0016758

improving grain food processing.

While the issue of long-term reliance between various amino acids in lengthy sequences was resolved by the PBiLSTM-FCN, the amino acid sequence was not precisely represented that is crucial to the function of proteins. Deep learning models are usually "black box" in nature, and it is difficult to understand how they make predictions. This lack of transparency makes it challenging to understand the model's behavior and trust its predictions. Future research will focus on developing new methods and techniques to uncover the decision-making mechanisms of these models, enabling users to better comprehend how the model operates.

CRediT authorship contribution statement

Jing Liu: Writing – review & editing, Writing – original draft, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kun Li:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation. **Xinghua Tang:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Investigation, Data curation. **Yu Zhang:** Writing – review & editing, Validation, Supervision, Resources, Investigation. **Xiao Guan:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

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

Acknowledgments

This work was supported by the Program of Shanghai Municipal Education Commission (Z-2024-312-099).

Research code availability

You can access our GitHub repository via the following link: https://github.com/Kun-Li-lab/PBiLSTM-FCN.git.

Appendix A. Supplementary data

The supplementary data has been submitted together with the manuscript.

Data availability

I have shared the link to my data and code at the Attach File step, and added in the manuscript.
dataset (Original data) (uniprot)

References

Alex, B., Maria-Jesus, M., Sandra, O., Michele, M., Shadab, A., Emanuele, A., et al. (2022). UniProt: The universal protein knowledgebase in 2023. *Nucleic Acids Research, 51*, 2699.
Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215*, 403–410.
Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research, 25*, 3389–3402.
Anam, K., Amjad, H., & Muhammad, F. T. (2023). Wheat quality: A review on chemical composition, nutritional attributes, grain anatomy, types, classification, and function of seed storage proteins in bread making quality. *Frontiers in Nutrition, 10*(1053196), 1–14.
Cao, D. N., Katheleen, J. G., Duong, N., & Ciso, K. J. (2008). Prediction of protein functions from protein interaction networks: A naïve Bayes approach. PRICAI 2008. *Trends in Artificial Intelligence, 5351*, 788–789.
Chen, X. W., & Liu, M. (2005). Prediction of protein–protein interactions using random decision forest framework. *Bioinformatics, 21*, 4394–4400.
Elhaj-Abdou, M. E. M., El-Dib, H., El-Helw, A., & EII-Habrouk, M. (2021). Deep_CNN_LSTM_GO: Protein function prediction from amino-acid sequences. *Computational Biology and Chemistry, 95*, 107584–107597.
Eugène, T., Pierre, M., & Anne-Marie, T.-B. (2003). Environmentally-induced changes in protein composition in developing grains of wheat are related to changes in total protein content. *Journal of Experimental Botany, 54*, 1731–1742.
Gillis, J., & Pavlidis, P. (2013). Characterizing the state of the art in the computational assignment of gene function: Lessons from the first critical assessment of functional annotation (CAFA). *BMC Bioinformatics, 14*, 1–12.
Gireen, N., Tranos, Z., & Elias, M. S. (2023). A review of evaluation metrics in machine learning algorithms. *Artificial Intelligence Application in Networks and Systems Lecture Notes in Networks and Systems*, 15–25.
Gligorijević, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Berenberg, D., Vatanen, T., et al. (2021). Structure-based protein function prediction using graph convolutional networks. *Nature Communications, 12*, 3168–3172.
Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 6*, 107–116.
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*, 1735–1780.
Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 8*, 2011–2023.
Huntley, R. P., Sawford, T., Martin, M. J., & Donovan, C. O. (2014). Understanding how and why the gene ontology and its annotations evolve: The GO within UniProt. *GigaScience, 3*, 24–32.
Jiang, S. Y., Ma, A., Xie, L., & Ramachandran, S. (2016). Improving protein content and quality by over-expressing artificially synthetic fusion proteins with high lysine and threonine constituent in rice plants. *Scientific Reports, 6*(34427), 1–14.
Kabli, F., Hamou, R. M., & Amine, A. (2018). Protein classification using n-gram technique and association rules. *International Journal of Software Innovation (IJSI), 6*, 77–89.
Karim, F., Majumdar, S., Darabi, H., & Harford, S. (2019). Multivariate LSTM-FCNs for time series classification. *Neural Networks, 116*, 237–245.
Kulmanov, M., & Hoehndorf, R. (2020). DeepGOPlus: Improved protein function prediction from sequence. *Bioinformatics, 36*, 422–429.
Kulmanov, M., Khan, M. A., & Hoehndorf, R. (2018). DeepGO: Predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics, 34*, 660–668.
Li, Y., Kong, D., Bai, M., He, H., Wang, H., & Wu, H. (2019). Correlation of the temporal and spatial expression patterns of HQT with the biosynthesis and accumulation of chlorogenic acid in Lonicera japonica flowers. *Horticulture Research, 6*(73), 1–14.
Li, Y., Li, W., He, K. Y., Li, P., Huang, Y., Nie, Z., et al. (2016). A biomimetic colorimetric logic gate system based on multi-functional peptide-mediated gold nanoparticle assembly. *Nanoscale, 8*, 8591–8599.

Liu, J., Zhang, X., Huang, K., Wei, Y., & Guan, X. (2025). Grain protein function prediction based on CNN and residual attention mechanism with AlphaFold2 structure data. *Applied Sciences, 15*(4), 1890–1909.
Lv, G., Li, Y., Wu, Z., Zhang, Y., Li, X., Wang, T., et al. (2024). Maize actin depolymerizing factor 1 (ZmADF1) negatively regulates pollen development. *Biochemical and Biophysical Research Communications, 703*, 149637–149646.
Lv, Z., Ao, C., & Zou, Q. (2019). Protein function prediction: From traditional classifier to deep learning. *Proteomics, 19*, Article 1900119.
Ma, Z., Ma, H., Chen, Z., Chen, X., Liu, G., Hu, Q., et al. (2022). Quality characteristics of Japonica rice in southern and Northern China and the effect of environments on its quality. *Agronomy, 12*.
Malik, Y., Segun, J., Louise, C. S., & Showe, M. K. (2008). Learning from positive examples when the negative class is undetermined- microRNA gene identification. *Algorithms for Molecular Biology, 3*, 2.
Manavski, N., Abdel-Salam, E., Schwenkert, S., Kunz, H. H., Brachmann, A., Leister, D., et al. (2025). Targeted introduction of premature stop codon in plant mitochondrial mRNA by a designer pentatricopeptide repeat protein with C-to-U editing function. *The Plant Journal, 121*(3), 1–12.
Maqbool, M. A., Beshir Issa, A. R., & Khokhar, E. S. (2021). Quality protein maize (QPM): Importance, genetics, timeline of different events, breeding strategies and varietal adoption. *Plant Breeding, 140*, 375–399.
Min, H, Chengjing, L, Jiaxin, X, Jiana, C, & Fangbo, C (2023). Lysine content and its relationship with protein content in indica rice landraces of China. *Food Chemistry, X,*, 17.
Nam, J. W., Shin, K. R., Han, J., Lee, Y., Kim, V., & Zhang, B. (2005). Human Microrna prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Research, 33*, 3570–3581.
Pandey, G., Myers, C. L., & Kumar, V. (2009). Incorporating functional inter-relationships into protein function prediction algorithms. *BMC Bioinformatics, 10*, 1–22.
Pingxu, Q., Taoran, W., & Yangchao, L. (2022). A review on plant-based proteins from soybean: Health benefits and soy product development. *Journal of Agriculture and Food Research, 7*, Article 100265.
Poutanen, K. S., Karlund, A. O., Carlos, G.-G., Daniel, P. J., Nathalie, M. S., Ingela, M. M., et al. (2022). Grains – A major source of sustainable protein for health. *Nutrition Reviews, 80*, 1648–1663.
Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., et al. (2013). A large-scale evaluation of computational protein function prediction. *Nature Methods, 10*, 221–227.
Raubenheimer, D., & Simpson, S. J. (2016). Nutritional ecology and human health. *Annual Review of Nutrition, 36*, 603–626.
Sara, S. T., Hasan, M. M., Ahmad, A., & Shatabda, S. (2021). Convolutional neural networks with image representation of amino acid sequences for protein function prediction. *Computational Biology and Chemistry, 92*, 107494–107498.
Sathya, M., Kokilavani, R., Teepa, K. S. A., & Balakrishnan, A. (2011). Biopotency of *Acalypha indica* Linn on membrane bound ATPases and marker enzymes urolithic rats. *Ancient Science of Life, 31*(1), 3–9.
Shelhamer, E., Long, J., & Darrell, T. (2017). Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*, 640–651.
Suzi, A. A., James, B., Seth, C., Cherry, J. M., Drabkin, H. J., Ebert, D., et al. (2023). The gene ontology knowledgebase in 2023. *Genetics, 224*(1).
Villa, M., Dardenne, G., Nasan, M., Letissier, H., & Hamitouche, C. (2018). FCN-based approach for the automatic segmentation of bone surfaces in ultrasound images. *International Journal of Computer assisted Radiology and Surgery, 13*, 1707–1716.
Xu, J., & Wang, S. (2019). Analysis of distance-based protein structure prediction by deep learning in CASP13. *Proteins: Structure, Function, and Bioinformatics, 87*, 1069–1081.
Zaremba, W., Sutskever, I., & Vinyals, O. (2015). Recurrent neural network regularization. *Proceedings of International Conference on Learning Representations, 1409*, 1–8.
Zuallaert, J., Pan, X., Saeys, Y., Wang, X., & Neve, W. D. (2019). Investigating the biological relevance in trained embedding representations of protein sequences. In *Biology at the 36th international conference on machine learning* (pp. 1–10).