JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

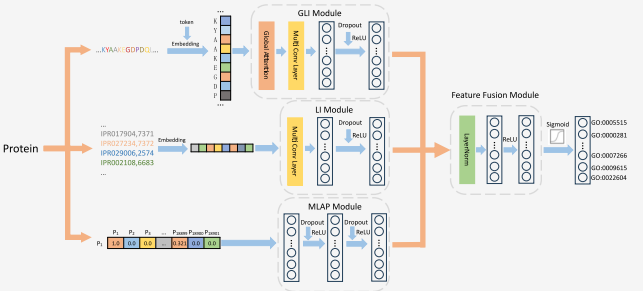
MSF-PFP: A Novel Multisource Feature Fusion Model for Protein Function Prediction

Xinhui Li, Yurong Qian,* Yue Hu, Jiaying Chen, Haitao Yue, and Lei Deng

Cite This: J. Chem. Inf. Model. 2024, 64, 1502−1511 Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Protein function prediction is essential for disease treatment and drug development; yet, traditional biological experimental methods are less efficient in annotating protein function, and existing automated methods fail to fully leverage protein multisource data. Here, we present MSF-PFP, a computational framework that fuses multisource data features to predict protein function with high accuracy. Our framework designs specific models for feature extraction based on the characteristics of various data sources, including a global-local-individual strategy for local location features. MSF-PFP then integrates extracted features through a multisource feature fusion model, ultimately categorizing protein functions. Experimental results demonstrate that MSF-PFP outperforms eight state-of-the-art models, achieving F_{Max} scores of 0.542, 0.675, and 0.624 for the biological process (BP), molecular function (MF), and cellular component (CC), respectively. The source code and data set for MSF-PFP are available at https://swanhub.co/TianGua/MSF-PFP, facilitating further exploration and validation of the proposed framework. This study highlights the potential of multisource data fusion in enhancing protein function prediction, contributing to improved disease therapy and medication discovery strategies.



INTRODUCTION

Proteins, as the cornerstone of life, are indispensable for the maintenance and repair of living tissues. Their complex functions are critical for the execution of essential biological processes.[1] In recent years, the exploration of proteins' role in life activities and the accurate identification of their functions have become a major focus in the field of biomedical research. As such, research on protein functions is essential for the development of new treatments for a wide range of biological processes.[2]

Protein function prediction is a vital tool in modern biological research, offering a deeper understanding of proteins and their roles in various biological processes. By employing computational and mathematical techniques, scientists can predict a protein's structure, stability, and activity. This information can lead to the identification of novel drug targets for diseases such as cancer,[3] HIV,[4] and Alzheimer's[5,6] and improve the development of new drug therapies.[7,8] Accurate and efficient protein function prediction methods are urgently needed, and further research is required to develop and apply these methods to a broader range of biological processes.

In traditional protein function prediction, it is assumed that proteins with similar sequences will exhibit similar functions. For example, tools such as BLAST,[9] PSI-BLAST,[10] and Diamond[11] use dynamic programming algorithms to perform approximate comparisons of two protein sequences. They utilized sequence homology theory to generate a comparison score, which is then used to predict the function of the queried protein. Functional annotation of proteins using sequence homology theory is inefficient and consumes computer resources. Machine learning models for protein function show good performance and use features extracted from large amounts of data to enhance the learning process and get the best prediction results through continuous iterative updates. For example, MultiLoc[12] predicts novel subcellular localization based on support vector machines (SVMs); SherLoc[13] predicts eukaryotic protein localization using SVMs; ECPred[14] builds multiple binary classifiers using a machine learning approach; and Srivastava et al.[15] compare the performance of two data mining methods, SVM and Random Forest, to predict protein function. Machine learning algorithms can extract hidden patterns from data more quickly and with guaranteed accuracy than traditional manual labeling. Thus, machine learning methods are used for protein prediction and consistently achieve good results.

Received: November 10, 2023
Revised: January 31, 2024
Accepted: February 21, 2024
Published: February 27, 2024

https://doi.org/10.1021/acs.jcim.3c01794
J. Chem. Inf. Model. 2024, 64, 1502−1511

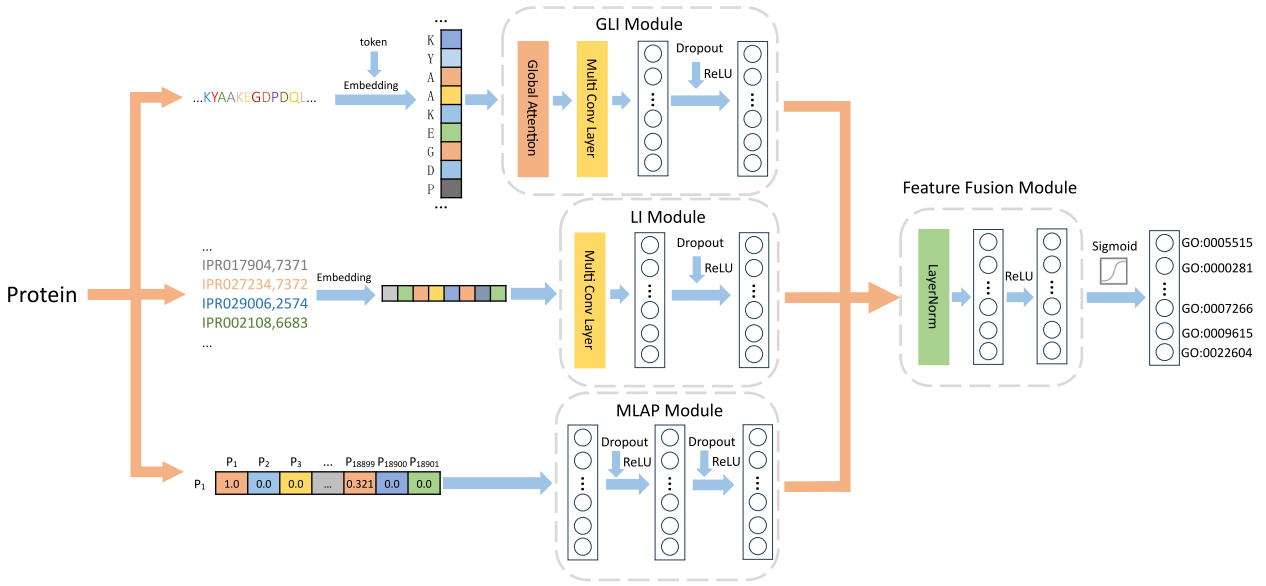


Figure 1. MSF-PFP overview: the model takes the sequence, domain, and PPI of proteins as inputs, which are processed by the GLI, LI, MLAP and finally fed into the multilayer fully connected network for fusion and classification, where GLI is the Global-local-individual feature extraction module, LI is the Local-individual feature extraction module, and MLAP is the Multi-Layer Adjacency Perception module.

However, machine learning requires extensive prior feature engineering, and the quality of the processed features has a significant impact on the results. Moreover, model training becomes more difficult when dealing with high-dimensional features and performs poorly when dealing with long sequences. Feature engineering for deep learning is not complicated and requires only preprocessing of the data. Protein sequences are a kind of text that is difficult to feature engineer, so deep learning is more friendly when dealing with textual data. The most typical sequence prediction method is DeepGOPlus,[16] which removes PPIN information from DeepGO and uses deep neural networks and sequence similarity for GO classification. DeepGOPlus uses the whole sequence for prediction, making it difficult to focus on local information. ProtVecGen-Ensemble[17] normalizes seqences by length and log-normalization and then divides the preprocessed sequence into base sequences of different lengths. The information carried by the sequences is highly accurate and reliable, but it is difficult to extract; and the influence of other factors, such as sequence folding and mutation, is not considered at the same time, which increases the difficulty of prediction and leads to low prediction accuracy. The STRING database[18] contains many protein interaction networks (PPINs), which are involved in or share a certain physiological function when two proteins have interaction information. Protein function can be inferred by querying the neighboring nodes of the protein. In recent years, PPIs have also become important data sources for protein function prediction. DeepNF[19] uses the STRING network processing as a low-dimensional representation of protein features. It adds different autoencoders at different network layers, but it does not take into account the deeper features in PPIN. FunPred 3.0[20] utilizes the domain-specific and physicochemical properties of amino acids in PPIN to incorporate four basic classifiers (XGBoost, Random Forest, Extra Tree, and RFE), all of which play a crucial role in the performance of predictions.

Furthermore, the 3D structure of proteins, which encapsulates fundamental physicochemical properties, in addition to positional and folding information, is more advantageous for protein function prediction. DeepFRI[21] employs sequence features extracted from protein language models and structures, which are subsequently fed into graph neural networks for predicting function and link structure and sequence. However, obtaining protein structures is exceedingly challenging and expensive. AlphaFold2,[22] with an accuracy of over 92% and an average error of 1 Å, predicts protein structures. Ma et al.[23] utilize PyTorch to replicate the DeepFRI model and AlphaFold2 to generate a "virtual structure" data set, demonstrating the usability of AlphaFold2's predicted structures. Currently, most models leverage cross-fertilization techniques for protein function prediction and predominantly rely on sequences, structures, structural domains, biological networks, medical literature, cross-references, and so forth.

Although there are many methods to predict protein function, there are still some limitations. Due to the large amount of protein data and the complex structure of protein data, the existing models do not properly extract the deep-level features in the data, and models with a large number of parameters rely too much on high-performance computing resources with high time complexity. Protein data sources are abundant, the existing models do not make full use of protein multisource data, and the fusion of multiple features is also a problem that urgently needs to be solved. The existing models have a weak generalization ability and cannot predict multiple protein functions with high accuracy.

In this work, we develop a model called MSF-PFP that is designed to predict protein function by fusing feature extraction and feature fusion. To extract features, MSF-PFP employs submodules focused on protein sequence, domain, and interaction information. By borrowing concepts from natural language processing and computer vision processing, the model learns deep features embedded in the data. This enables MSF-PFP to generate precise and inclusive features from protein sequences, enhancing its ability to predict protein function accurately. For feature fusion, MSF-PFP develops a module that combines features from the three submodules.

https://doi.org/10.1021/acs.jcim.3c01794
J. Chem. Inf. Model. 2024, 64, 1502−1511

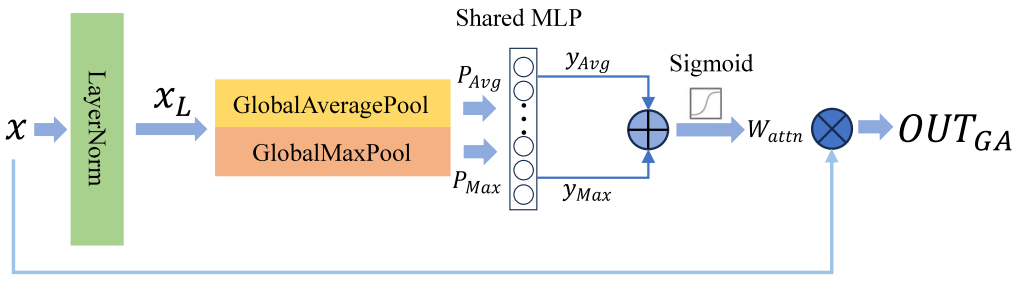


Figure 2. Global Attention module overview.

This approach makes full use of multisource data and solves the problem of inadequate expression of protein function by a single feature. Multisource data complement each other's discrepancies between data, and this fusion allows for a more comprehensive use of multisource data and more accurate prediction of protein function. Our method ensures that the model effectively consolidates information from multiple sources, resulting in improved accuracy in protein function predictions.

Overall, the design of MSF-PFP highlights the importance of using advanced techniques and methods to predict protein function. By incorporation of ideas from natural language processing and computer vision processing, MSF-PFP learns accurate features from protein sequence data and makes more accurate predictions about protein function. The experimental results showed that MSF-PFP excelled when compared to the existing superior models. Ablation experiments demonstrated the effectiveness of the designed feature extraction module in protein function prediction.

METHOD

MSF-PFP. MSF-PFP consisted of two stages for fusing multiple source features of proteins (Figure 1). In the first stage, feature learning employs natural language processing techniques and computer vision ideas to extract deep features of proteins. This stage also introduces a global-local-individual feature extraction strategy. The second stage, fusion, combines the feature representations from the previous stage with multilayer fully connected networks to predict protein functions.

Global-Local-Individual Feature Extraction Module. Protein sequences are arrangements of amino acids closely related to the functions of proteins. To fully explore the individual amino acids, the local arrangement of amino acids, and the overall arrangement of the sequence, MSF-PFP proposed the global local individual feature extraction module.

To ensure that each amino acid has a relatively independent and complete representation, MSF-PFP employs Token[24] for the disambiguation of words. Subsequently, the word embeddings generated by the tokenizer are applied to the segmentation results, ensuring that each amino acid letter is uniformly distributed across the high-dimensional feature space. The calculation of this process is publicized below

$$Token = D[input] \tag{1}$$

$$x_{Em} = \text{Embedding}(Token) \tag{2}$$

where D is the dictionary of Token.

This approach leverages natural language processing techniques to convert amino acids into features that are evenly distributed in the high-dimensional space, preserving the individual information on each amino acid and making it easier for the feature extraction module to learn the deeper, more complex features embedded in the protein sequence.

The traditional self-attention mechanism module is unable to simultaneously consider both the individual information on amino acids and the overall information on amino acid sequences. In contrast, computer vision processing methods are better able to capture the location information in the data. MSF-PFP introduces the Global Attention module (GA), compressing the token sequence into the embedding dimension to maximize the retention of the overall sequence information, resulting in a more robust and accurate feature representation. The overall structure is shown in Figure 2.

As shown in Figure 2, the result of word embedding is represented as x, and the Layernorm can be expressed as

$$x_L = \text{Layernorm}(x) = \frac{x - \mu}{\sigma} \tag{3}$$

Here, the shape of x is represented as *(batch_size,embed_dim,seq_len)*, where *embed_dim* stands for the word embedding dimension, and *seq_len* represents the sequence length. The mean of the input data for the current layer is denoted as μ, and the standard deviation is represented as σ.

Subsequently, the results of Layernorm are fed into the global average pooling layer and global max pooling layer to extract the maximum and average information from the data, respectively. At this stage, MSF-PFP treats *embed_dim* as the compressed dimension and performs pooling operations at each embedding dimension to compress the feature and obtain valid information in the compressed dimension. The pooling can be expressed as

$$P_{Avg} = \text{Avgpool}(x_L) \tag{4}$$

$$P_{Max} = \text{Maxpool}(x_L) \tag{5}$$

where the shape of P after pooling is represented as *(batch_size,embed_dim)*.

P_{Max} and P_{Avg} are fed into the MLP with shared weights for feature extraction, which can be expressed as

$$h_j = \sum_{i=0}^{N} W_{ij}P_j + b_j \tag{6}$$

$$a_j = g(h_j) = g\left(\sum_{i=0}^{N} W_{ij}P_j + b_j\right) \tag{7}$$

$$y = a_k = g(h_k) = g\left(\sum_{i=0}^{N} W_{ik}P_k + b_k\right) \tag{8}$$

https://doi.org/10.1021/acs.jcim.3c01794
J. Chem. Inf. Model. 2024, 64, 1502−1511

$$g(h) = ReLU(h) \tag{9}$$

where h_j denotes the weighted sum of all the inputs of the current node, g denotes the activation function, and $a_j = P_k$ represents the current output and the next layer of inputs. y is the result of the value of the output layer.

After MLP processing, y_{Max} and y_{Avg} are summed, the attention weight W_{attn} is calculated using the sigmoid activation function, and then W_{attn} is multiplied by input x to obtain final result OUT_{GA}. The calculation process can be expressed as

$$y = y_{Avg} + y_{Max} \tag{10}$$

$$W_{attn} = Sigmoid(y) \tag{11}$$

$$OUT_{GA} = W_{attn} \times x \tag{12}$$

where the shape of OUT_{GA} is $(batch\_size,embed\_dim,seq\_len)$.

In the context of the above mandate, the incorporation of word embedding and attention modules aims to maximize the utilization of both the individual information and the overall arrangement information on amino acids. Previous studies have established a correlation between protein functions and the local arrangement order of amino acids. Given the ability of convolutional neural networks (CNNs)[25] to effectively learn local position information on features, MSF-PFP employed a multilayer CNN to comprehensively exploit the local positional features embedded within protein sequences. The calculation of this process is publicized below

$$OUT(i, Dim_{OUT_j}) = bias(Dim_{OUT_j}) + \sum_{k=0}^{Dim_{IN}-1} weight(Dim_{OUT_j}, k) \bigstar OUT_{GA}(i, k) \tag{13}$$

where Dim_{OUT} is the output dimension of the 1D convolution, and Dim_{IN} is the embedding dimension of the $OUT_{GA}(embed\_dim)$. ★ is the one-dimensional convolutional cross-correlation operation, calculated as follows

$$(x \bigstar kernel)_i = \sum_j x_j \times kernel_{i-j} \tag{14}$$

where x denotes the sequence being convolved, $kernel$ is the convolution kernel, i is the index of the convolution result, and j is the index of the convolution kernel.

This approach facilitated the extraction of protein sequence features from multiple scales, encompassing the individual, local, and overall perspectives, thereby yielding a more accurate and stable feature representation. The adoption of this multiscale feature extraction strategy, encompassing the individual, local, and overall perspectives, is anticipated to enhance the performance of a diverse range of NLP tasks and provide valuable insights into the process of sequence feature extraction.

Local-Individual Feature Extraction Module. In the field of protein function prediction, each protein domain is considered an independent amino acid fragment that possesses a distinct functional role. According to the idea of the Global-local-individual feature extraction module, MSF-PFP proposes a Local-individual (LI) feature extraction module for protein domains.

To fully capture the individual features of structural domains, MSF-PFP employed a word embedding method[26] to encode each protein domain ID. This approach facilitates the generation of features that are uniformly distributed within a high-dimensional space, thereby providing a comprehensive representation of the individual features of each protein domain. The calculation of this process is publicized below

$$ID = D|input| \tag{15}$$

$$x_{Em} = \text{Embedding}(ID) \tag{16}$$

where D is the dictionary of domain IDs.

To obtain localized information more accurately about the domains, MSF-PFP constructed a corresponding CNN module. The calculation of this process is publicized below

$$OUT(i, Dim_{OUT_j}) = bias(Dim_{OUT_j}) + \sum_{k=0}^{Dim_{IN}-1} weight(Dim_{OUT_j}, k) \bigstar x_{Em}(i, k) \tag{17}$$

where Dim_{OUT} is the output dimension of the 1D convolution, and Dim_{IN} is the embedding dimension of x_{Em}.

This module places significant emphasis on local positional information within protein domains, ensuring that the global context of each domain is considered. The integration of the CNN module with the word embedding method enables LI to explore deeper features within the protein domain data at both the individual and local scales. Consequently, this combination leads to a more comprehensive and accurate representation of the features within the protein domain data.

Multi-Layer Adjacency Perception Module. In the protein−protein interaction (PPI), the scoring matrix is a crucial component for assessing the strength of interactions between proteins. To effectively learn the intricate features of this matrix, the Multi-Layer Adjacency Perception module (MLAP), as part of the MSF-PFP framework, is designed. The detailed computational process of the multilayer perceptron is demonstrated in eqs 6-9 above, and the computational process of this step is simply calculated here

$$z_i = W_i a_{i-1} + b_i \tag{18}$$

$$a_i = \sigma_i(z_i) \tag{19}$$

where z_i is the weighted output of layer i, W_i is the weight matrix linking layer $i-1$ to layer i, a_i is the activation output of layer i, b_i is the bias vector of layer i, and σ_i is the activation function for layer i.

The employment of MLAP is justified by its robust fault tolerance and exceptional generalization capabilities.

The MLAP module combines multiple layers of architecture, enabling the capture of complex linear relationships present within the scoring matrix. The interlayer information transfer further refines this characterization, resulting in a more comprehensive understanding of the protein interactions. Additionally, the strong classification ability of MLAP allows for the provision of more accurate and robust feature expressions for the subsequent feature fusion module.

Feature Fusion Module. In the field of multifeature fusion tasks, the issues of exorbitant parameter growth, sluggish computation velocity, and elevated computational expenses due to excessive model complexity are pervasive. To tackle these challenges, MSF-PFP adopted MLP as the fundamental structure of the feature fusion module.

The adoption of MLP confers upon the fusion process both speedy and precise feature integration thanks to its superior

https://doi.org/10.1021/acs.jcim.3c01794
J. Chem. Inf. Model. 2024, 64, 1502−1511

classification as well as generalization capabilities. This designation serves as an accurate foundation for the subsequent classification task, effectively ameliorating the computational efficiency and accuracy of traditional protein function prediction methodologies.

The MSF-PFP approach not only mitigates the issues arising from excessive model complexity but also enhances the model's ability to generalize, leading to more accurate predictions. Moreover, MSF-PFP demonstrates a higher computational efficiency compared to its counterparts, making it a viable choice for protein function prediction tasks.

In summary, MSF-PFP addresses the key concerns associated with multifeature fusion tasks by employing MLP as the basic structure of the feature fusion module. This design provides a solid foundation for the final classification task and significantly improves the computational efficiency and accuracy of traditional protein function prediction approaches.

EXPERIMENTS AND RESULTS

The data sets from all the comparison experiments were divided into 5-fold cross-validation data sets, with BP, MF, and CC serving as the corresponding labels for model training. F_{Max} score, *AUC*, *Recall*, and *Precision* were used as the evaluators to assess the model's performance during the 5-fold cross-validation. Among them, F_{Max} is one of the key metrics in the CAFA Challenge. The evaluators are defined as follows:

$$F_{Max} = Max_t \frac{2\cdot \mathrm{AvgPr}(t) - \mathrm{AvgRc}(t)}{\mathrm{AvgPr}(t) + \mathrm{AvgRc}(t)} \tag{20}$$

$$\mathrm{AvgPr}(t) = \frac{1}{k(t)}\cdot \sum_{i=1}^{k(t)} pr_i(t) \tag{21}$$

$$\mathrm{AvgRc}(t) = \frac{1}{n}\cdot \sum_{i=1}^{n} rc_i(t) \tag{22}$$

$$pr_i(t) = \frac{\sum_j T(G_j, p_i)\cdot J(S(p_i, G_j) \geq t)}{\sum_j J(S(p_i, G_j) \geq t)} \tag{23}$$

$$rc_i(t) = \frac{\sum_j T(G_j, p_i)\cdot J(S(p_i, G_j) \geq t)}{\sum_j T(G_j, p_i)} \tag{24}$$

Here, t denotes the threshold for prediction, where $t \in [0, 1]$, and the step size is 0.1. J denotes the value to determine whether the protein prediction is true or not, if true, then 1, otherwise 0. Note that the total number of proteins is n.

F_{Max}, *Precision*, and *Recall* were used to measure the accuracy of the model's predictions, while the *AUC* was used to evaluate the model's ability to identify relevant proteins. The *AUC* calculation is derived from the confusion matrix with the following formula:

$$AUC = \int_{-\infty}^{\infty} TPR(t)(-FPR(t))\mathrm{d}t \tag{25}$$

Among others,

$$TPR(t) = \frac{TP(t)}{TP(t) + FN(t)} \tag{26}$$

$$FPR(t) = \frac{FP(t)}{FP(t) + TN(t)} \tag{27}$$

In the confusion matrix, *TP*, *FP*, *TN*, and *FN* are indicated as the number of true positives, false positives, true negatives, and false negatives, respectively.

Additionally, the cross-validation method was chosen to ensure that the data sets used for training and testing were representative of the entire data set, thereby providing a reliable evaluation of the model's performance.

Data Set. In this study, a multisource feature fusion strategy was adopted for the task of protein functional multilabel classification, so the data sets used include protein sequences, protein domains, and protein interaction information scoring matrices. Existing publicly available data sets cannot simultaneously satisfy the need for multisource information, so this study collected and organized protein data sets with all three of these data sources.

The first data set (2019 Human)[27] of this study was collated in 2019. Protein sequence data were retrieved from the UniProt[28] database for human data and downloaded as FASTA files. Protein structural domains were obtained from the public database InterPro[29] and matched to protein sequences. PPI data were downloaded from the STRING[18] database version V10 human data and processed into scoring matrices. The contents of the data set are shown in Table 1.

Table 1. 2019 Human Data Set Details

Data Set	BP	MF	CC
Train	9521	9392	10038
Test	2381	2349	2510

All GO terms were screened for BP terms with a frequency ≥ 40 occurrences, MF terms with a frequency ≥ 20 occurrences, and CC terms with a frequency ≥ 20 occurrences. The GO term labeling data contained 491 BP terms, 321 MF terms, and 240 CC terms. This data set retains data that are missing some kind of GO term labeling, so the amount of data for BP, MF, and CC will be unequal.

The second data set (2024 Human) of this study is 2024 retrieved and matched data from the above databases, where PPI data were downloaded from STRING version V11 human data and processed as the scoring matrix. The content of the data set is shown in Table 2.

Table 2. 2024 Human Data Set Details

Data Set	Proteins
Train	4496
Test	1124

All GO terms were screened for BP terms with a frequency ≥ 20 occurrences, MF terms with a frequency ≥ 15 occurrences, and CC terms with a frequency ≥ 15 occurrences. The GO term labeling data contained 494 BP terms, 397 MF terms, and 351 CC terms. This data set retains only data that are eligible and fully labeled, so the amount of data for BP, MF, and CC will be equal.

Comparative Experiments. *SDN2GO.* SDN2GO[27] leverages protein multisource data fusion to predict protein functional labels. In the sequence model, SDN2GO employs N-gram coding to process protein sequence data and subsequently utilizes a convolution module to extract features from the sequence. Notably, SDN2GO has made significant

https://doi.org/10.1021/acs.jcim.3c01794
J. Chem. Inf. Model. 2024, 64, 1502−1511

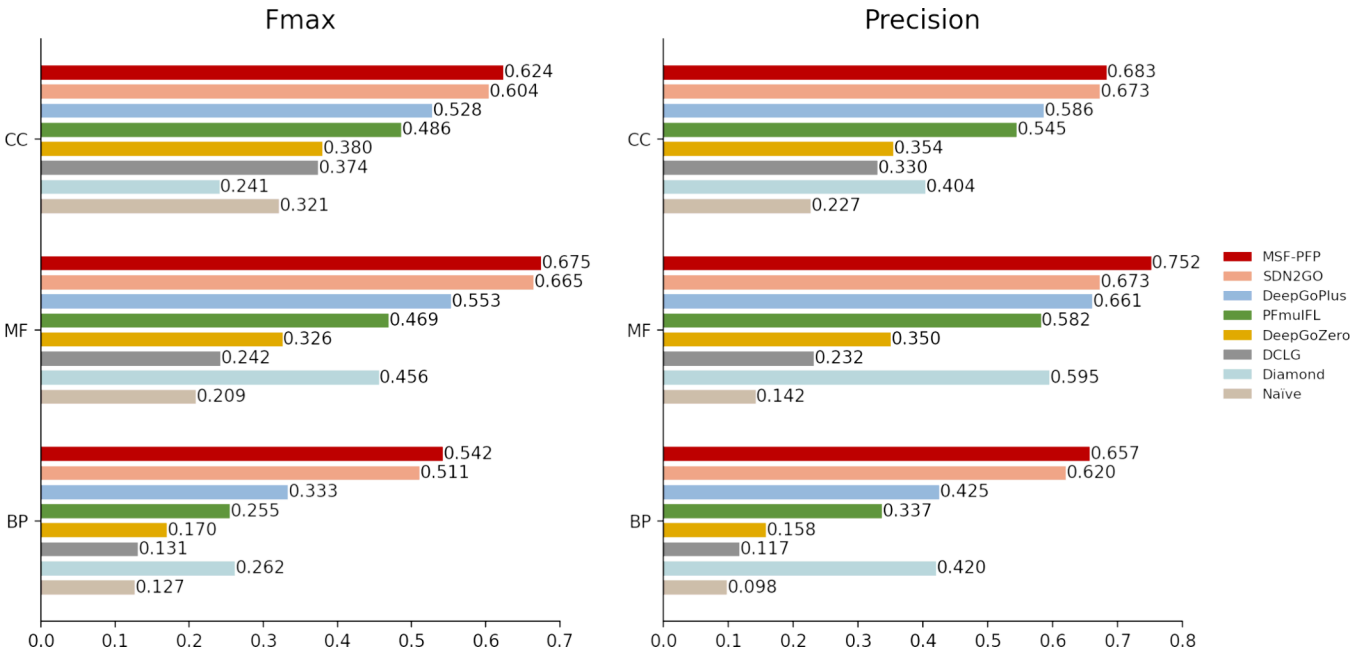


Figure 3. Comparison of MSF-PFP with seven correlation models of 2019 Human.

Table 3. Performance Comparison of 2019 Human between naïve, Diamond, DCLG, DeepGOZero, PFmulDL, DeepGOPlus, SDN2GO, and MSF-PFP, where the naïve and Diamond Methods Are Baseline Models in CAFA Competitions[a]

	Performance evaluation metric											
	F_{max}			*AUC*			*Recall*			*Precision*		
Method	BP	MF	CC	BP	MF	CC	BP	MF	CC	BP	MF	CC
naïve	0.127	0.209	0.321	0.581	0.684	0.760	0.181	0.395	0.552	0.098	0.142	0.227
Diamond	0.262	0.456	0.241	0.672	0.770	0.704	0.212	0.400	0.193	0.420	0.595	0.404
DCLG	0.131	0.242	0.374	0.742	0.840	0.882	0.147	0.254	0.432	0.117	0.232	0.330
DeepGoZero	0.170	0.326	0.380	0.769	0.867	0.889	0.184	0.305	0.412	0.158	0.350	0.354
PFmulFL	0.255	0.469	0.486	0.803	0.900	0.912	0.207	0.394	0.439	0.337	0.582	0.545
DeepGoPlus	0.333	0.553	0.528	0.864	0.940	0.933	0.274	0.476	0.481	0.425	0.661	0.586
SDN2GO	0.511	0.665	0.604	0.923	0.959	**0.953**	0.434	0.547	0.547	0.620	0.673	0.673
MSF-PFP(ours)	**0.542**	**0.675**	**0.624**	**0.935**	**0.964**	0.951	**0.461**	**0.613**	**0.575**	**0.657**	**0.752**	**0.683**

[a]Bold indicates optimal values, and underlining indicates suboptimal values.

advancements in feature extraction compared with other models.

DeepGOPlus. DeepGOPlus[16] overcomes the limitations of sequence length while maintaining excellent prediction accuracy. The model utilizes a multiscale CNN to extract high-level features from the motif. In addition, it incorporates the Diamond tool for calculating sequence similarity and utilizes a combination of Diamondscore's K-nearest neighbor algorithm and DeepGOCNN deep CNNS to efficiently perform the prediction task. It is worth noting that DeepGOPlus performed most prominently in the CAFA3 challenge, demonstrating its potential for real-world applications.

DeepGOZero. DeepGOZero[30] is the first machine learning method to use GO axioms for zero-shot predic-tions. The background knowledge of GO terminology is incorporated, and only sequence data are used for proteins with few or no annotations in the sequence database, improving the classification performance of specific GO terms.

Deep_CNN_LSTM_GO. Deep_CNN_LSTM_GO (DCLG)[31] is also a sequence-based prediction method. It uses the CNN module and Long Short-Term Memory (LSTM) network. The model combines the advantages of both architectures to improve the prediction.

PFmulDL. PFmulDL[32] combines the architectural strengths of multicore CNNs and RNNs to perform functional annotation of proteins using amino acid sequences. Notably, the model shows an exceptional ability to predict both "major class" proteins, which are usually well predicted by existing mainstream models, and "rare class" proteins, which are less common.

Performance Evaluation. To validate the effectiveness of MSF-PFP, a comprehensive comparative experiment was conducted with seven other excellent models. The experimental results of 2019 Human are presented in Figure 3. *Precision* of MSF-PFP was determined to be 0.657, 0.752, and 0.683 in BP, MF, and CC terms, respectively. These results indicated that MSF-PFP significantly outperformed the other models in terms of protein function prediction accuracy. Furthermore, F_{Max} of MSF-PFP was found to be 0.542, 0.675, and 0.624 in the BP, MF, and CC data sets, which were higher than the other models. These results clearly demonstrate the reliability and efficiency of MSF-PFP in predicting protein functions.

https://doi.org/10.1021/acs.jcim.3c01794
J. Chem. Inf. Model. 2024, 64, 1502−1511

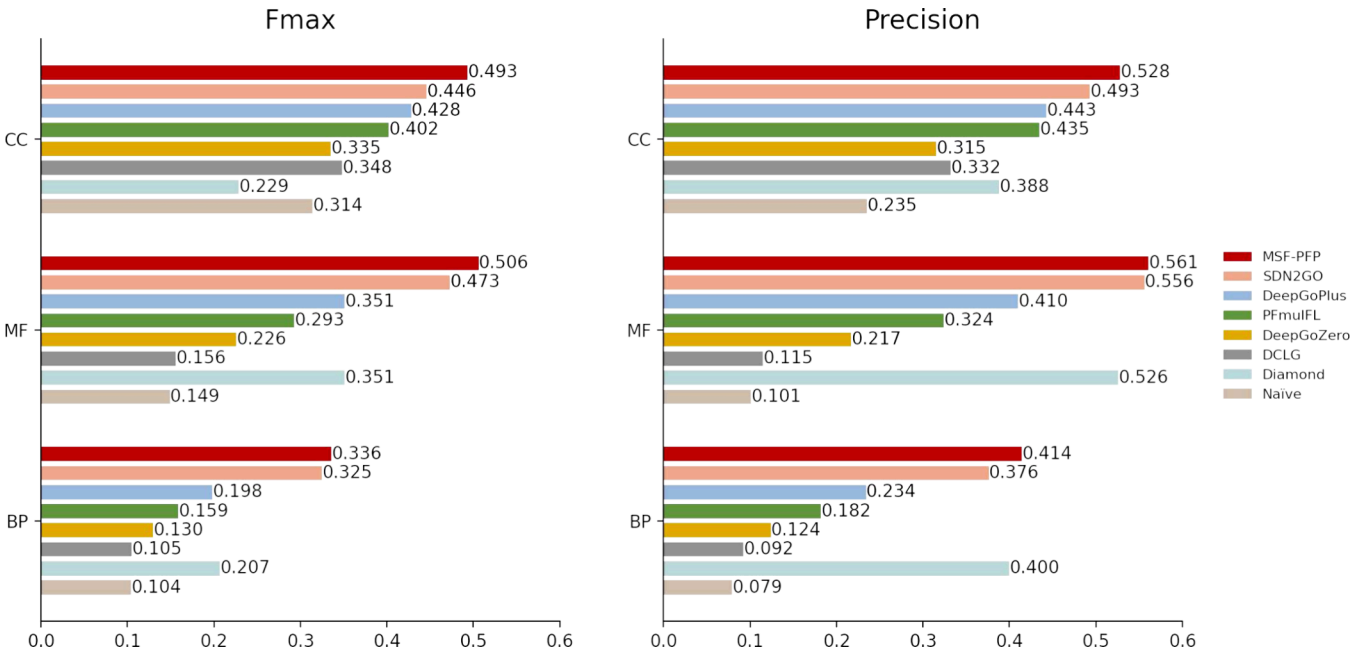


Figure 4. Comparison of MSF-PFP with seven correlation models of 2024 Human.

Table 4. Performance Comparison of 2024 Human between naïve, Diamond, DCLG, DeepGOZero, PFmulDL, DeepGOPlus, SDN2GO, and MSF-PFP[a]

	Performance evaluation metric											
	F_{max}			*AUC*			*Recall*			*Precision*		
Method	BP	MF	CC	BP	MF	CC	BP	MF	CC	BP	MF	CC
naïve	0.104	0.149	0.314	0.566	0.631	0.725	0.151	0.285	**0.417**	0.079	0.101	0.235
Diamond	0.207	0.351	0.229	0.768	0.873	0.832	0.148	0.285	0.174	0.400	0.526	0.388
DCLG	0.105	0.156	0.348	0.680	0.795	0.849	0.125	0.244	0.371	0.092	0.115	0.332
DeepGoZero	0.130	0.226	0.335	0.713	0.849	0.852	0.140	0.241	0.359	0.124	0.217	0.315
PFmulFL	0.159	0.293	0.402	0.748	0.866	0.893	0.145	0.271	0.373	0.182	0.324	0.435
DeepGoPlus	0.198	0.351	0.428	0.778	0.893	0.901	0.171	0.307	0.414	0.234	0.410	0.443
SDN2GO	0.325	0.473	0.446	0.846	**0.921**	0.917	**0.288**	0.413	0.408	0.376	0.556	0.493
MSF-PFP(ours)	**0.336**	**0.506**	**0.493**	**0.865**	0.917	**0.926**	0.283	**0.461**	0.462	**0.414**	**0.561**	**0.528**

[a]Bold indicates optimal values, and underlining indicates suboptimal values.

The complete evaluation results are listed in Table 3. The evaluation metrics of MSF-PFP across BP, MF, and CC showed optimal or suboptimal values, demonstrating that MSF-PFP outperforms the other models in this task. These results clearly confirm the efficacy of MSF-PFP in predicting protein functions.

The results of the 2024 Human experiments are shown in Figure 4. *Precision* of MSF-PFP in BP, MF and CC data sets is 0.414, 0.561, and 0.528, respectively, and F_{Max} of MSF-PFP is 0.336, 0.506, and 0.493, respectively. The experimental results demonstrate the effectiveness of MSF-PFP in the protein function prediction task.

The results of the complete experiment are shown in Table 4.

In conclusion, the superiority of the MSF-PFP method over the other eight methods may be attributed to its exceptional ability of multisource feature fusion. This capability enables MSF-PFP to learn information from multiple sources and capture more complex features. When compared to other encoding methods, tokenized protein sequence data better preserve the individual order of each amino acid and its meaning, resulting in a more densely packed feature matrix and reduced loss of critical information in deep learning networks. The introduction of GLI allows the model to efficiently extract rich features in sequences, surpassing the capabilities of the conventional CNN modular feature extraction methods and enabling the model to fully utilize sequence data to enhance its performance. In conclusion, MSF-PFP proved to be a reliable method for predicting protein function.

Ablation Experiment. In this study, we present a deep-learning-based MSF-PFP model for predicting protein functions based on protein sequences, structural domains, and PPI information. To verify the efficacy of the token embedding module, as well as the proposed GA module in the GLI, and the necessity of the newly proposed GLI, we conducted ablation experiments.

The baseline model for this study employed the 3-Gram approach to slice and code proteins; however, MSF-PFP took a different approach by utilizing the token method. Each amino acid was regarded as a word in the context of the protein sequence, and word embeddings were subsequently processed. This approach allowed MSF-PFP to fully leverage individual information on amino acids and effectively disambiguated protein sequences. In this study, the proposed method was

https://doi.org/10.1021/acs.jcim.3c01794
J. Chem. Inf. Model. 2024, 64, 1502−1511

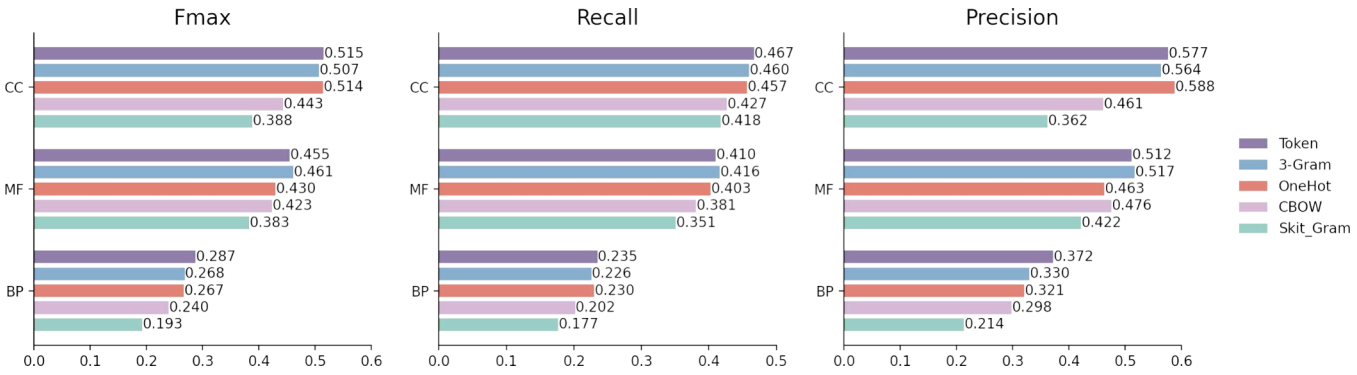


Figure 5. Ablation experiments with the Encoding module.

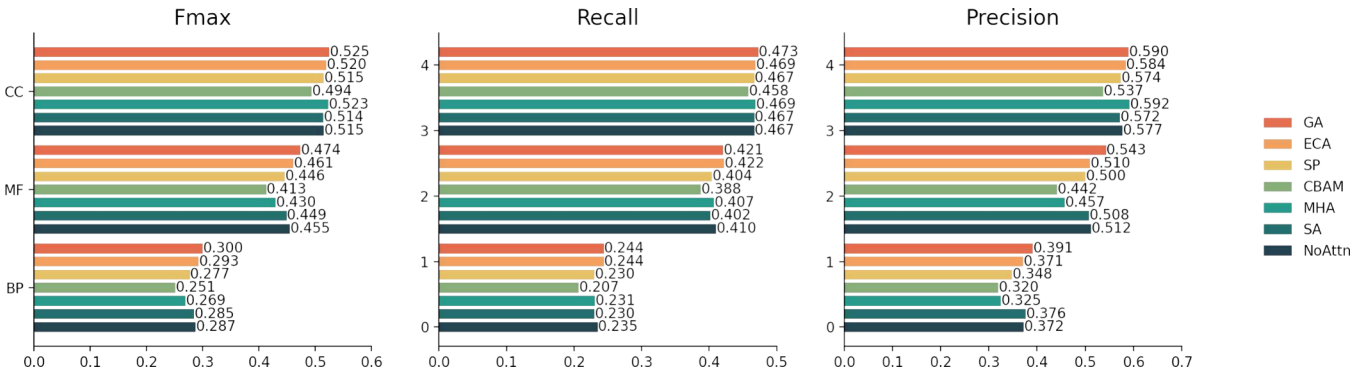


Figure 6. Ablation experiments with Attention modules. NoAttn denotes a model that does not introduce an attention mechanism.

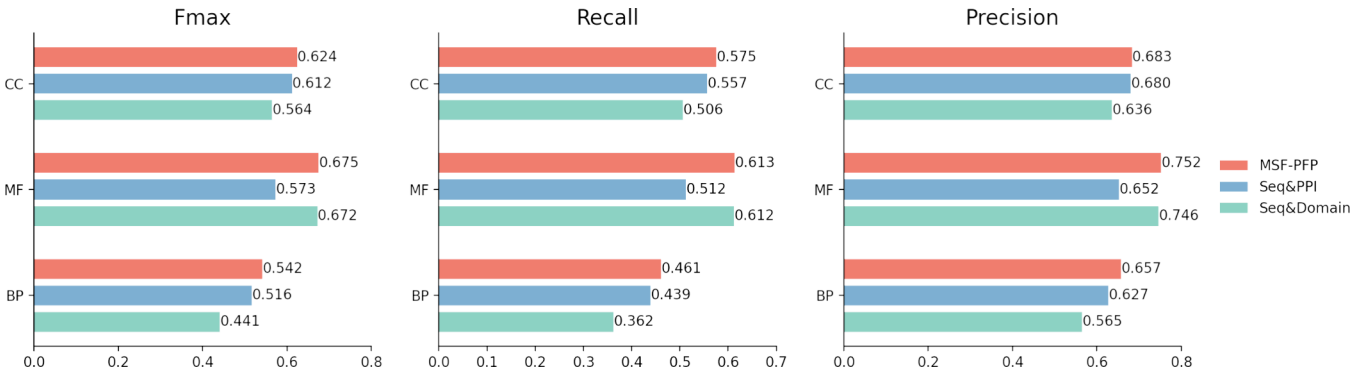


Figure 7. Ablation experiments with Multi-Source Feature, where Seq&Domain indicates the use of sequence and domain data, Seq&PPI indicates the use of sequence and interaction data, and MSF-PFP indicates the use of sequence, domain, and PPI data.

compared with the commonly used One-Hot, CBOW, Skit-Gram, and 3-Gram methods, and the experimental results are presented in Figure 5.

As presented in Figure 5, the Token module achieves improved performance in all evaluation metrics for BP, MF, and CC.

In terms of feature extraction, the baseline model employed a CNN structure to process the features. However, traditional CNNs may struggle to effectively hand over the correlation features between the embedded dimensions in encoded protein sequences. By treating these embedded dimensions as compressed dimensions and designing the GA module, MSF-PFP can fully use the correlation features among the embedded dimensions to improve feature extraction. In the second part of the ablation experiments, MSF-PFP compared its performance across various attention modules, demonstrating its effectiveness in this task.

The experimental results of this study are listed in Figure 6. MSF-PFP was compared to a variety of attention mechanism modules, including the Self Attention (SA), Multi-Head Attention (MHA), Convolutional Block Attention Module (CBAM), spatial attention (SP), Efficient Channel Attention (ECA), and GA modules. It turned out that the GA module performs the best in terms of protein sequence feature extraction. This demonstrates the efficacy of the GA module in protein sequence feature extraction.

The primary objective of MSF-PFP is to integrate multisource protein data for accurate protein function prediction, which provides the rationale for the last set of ablation experiments. In this study, multiple data sources are incrementally added to verify the effectiveness of multisource feature fusion. The experimental results are shown in Figure 7: multiple data sources can effectively compensate for the feature richness of single data sources. Multisource feature fusion can

https://doi.org/10.1021/acs.jcim.3c01794
J. Chem. Inf. Model. 2024, 64, 1502−1511

fully utilize the existing protein data to achieve accurate prediction of protein function.

DISCUSSION AND CONCLUSION

Numerous studies have demonstrated that understanding protein function is essential for exploring life processes and biological principles. Predicting protein function based on deep learning methods can enhance the efficiency of current protein function annotation. However, existing approaches fail to fully leverage the diversity of protein data as well as effectively learn location information.

To address these limitations, we propose a novel method called MSF-PFP that combines natural language processing and computer vision techniques to integrate multisource protein data. The method consists of three modules: GLI, LI, and MLAP, respectively, designed for learning sequence, domain, and PPI feature information.

After fusing multisource information, MSF-PFP predicts protein function. Experimental results indicate that MSF-PFP outperforms other state-of-the-art models. The F_{Max} scores for BP, MF, and CC reached 0.542, 0.675, and 0.624, respectively.

Ablation experiments demonstrate the effectiveness of the GLI module, which extracts protein sequence features. Furthermore, the feature fusion strategy of MSF-PFP significantly enhances the protein function prediction. Overall, MSF-PFP represents a promising approach for predicting protein function by using deep learning techniques.

In future investigations, we intend to expand our data set by incorporating additional species and introducing novel features, such as protein structural information and physicochemical properties, to enhance the predictive accuracy of protein function. Moreover, we also plan to integrate large language models for more precise protein sequence representations.

ASSOCIATED CONTENT

Data Availability Statement

The code and data sets of MSF-PFP are available at https://swanhub.co/TianGua/MSF-PFP.

AUTHOR INFORMATION

Corresponding Author

Yurong Qian − *School of Software, Xinjiang University, Urumqi 830091, China; Key Laboratory of Signal Detection and Processing in Xinjiang Uygur Autonomous Region, Xinjiang University, Urumqi 830046, China; Key Laboratory of Software Engineering, Xinjiang University, Urumqi 830091, China*; Email: qyr@xju.edu.cn

Authors

Xinhui Li − *School of Software, Xinjiang University, Urumqi 830091, China; Key Laboratory of Signal Detection and Processing in Xinjiang Uygur Autonomous Region, Xinjiang University, Urumqi 830046, China; Key Laboratory of Software Engineering, Xinjiang University, Urumqi 830091, China*; orcid.org/0009-0009-7990-6828

Yue Hu − *School of Software, Xinjiang University, Urumqi 830091, China; Key Laboratory of Signal Detection and Processing in Xinjiang Uygur Autonomous Region, Xinjiang University, Urumqi 830046, China; Key Laboratory of Software Engineering, Xinjiang University, Urumqi 830091, China*

Jiaying Chen − *School of Software, Xinjiang University, Urumqi 830091, China; Key Laboratory of Signal Detection and Processing in Xinjiang Uygur Autonomous Region, Xinjiang University, Urumqi 830046, China; Key Laboratory of Software Engineering, Xinjiang University, Urumqi 830091, China*

Haitao Yue − *School of Future Technology and Laboratory of Synthetic Biology, School of Life Science and Technology, Xinjiang University, Urumqi 830017, China*

Lei Deng − *School of Software, Xinjiang University, Urumqi 830091, China; School of Computer Science and Engineering, Central South University, Changsha 410083, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c01794

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work is supported by the Third Xinjiang Scientific Expedition Program (No. 2022XJKK-020603), the Tianshan Innovation Team Program of Xinjiang Uygur Autonomous Region of China (No. 2023D14012), the Excellent Youth Foundation of Xinjiang Uygur Autonomous Region of China (No. 2023D01E01), the Natural Science Foundation of Xinjiang Uygur Autonomous Region of China (No. 2022D01C692), the Basic Research Foundation of Universities in the Xinjiang Uygur Autonomous Region of China (No. XJEDU2023P012), and the National Natural Science Foundation of China (Nos. 62266043, 61966035, 62272490, and 61972422).

REFERENCES

(1) Agaton, C.; Uhlén, M.; Hober, S. Genome-based proteomics. *Electrophoresis* **2004**, *25*, 1280−1288.

(2) Avery, C.; Patterson, J.; Grear, T.; Frater, T.; Jacobs, D. J. Protein Function Analysis through Machine Learning. *Biomolecules* **2022**, *12*, 1246.

(3) Wu, Y.; Zhao, J.; Tian, Y.; Jin, H. Cellular functions of heat shock protein 20 (HSPB6) in cancer: A review. *Cell. Signalling* **2023**, *112*, 110928.

(4) Namba, M. D.; Xie, Q.; Barker, J. M. Advancing the preclinical study of comorbid neuroHIV and substance use disorders: Current perspectives and future directions. *Brain, Behav., Immun.* **2023**, *113*, 453.

(5) Wilkins, H. M. Interactions between amyloid, amyloid precursor protein, and mitochondria. *Biochem. Soc. Trans.* **2023**, *51*, 173−182.

(6) Pradeepkiran, J. A.; Baig, J.; Selman, A.; Reddy, P. H. Mitochondria in Aging and Alzheimer's Disease: Focus on Mitophagy. *Neuroscientist* **2023**, DOI: 10.1177/10738584221139761.

(7) Barabási, A.-L.; Gulbahce, N.; Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **2011**, *12*, 56−68.

(8) Xuan, P.; Sun, C.; Zhang, T.; Ye, Y.; Shen, T.; Dong, Y. Gradient boosting decision tree-based method for predicting interactions between target genes and drugs. *Front. Genet.* **2019**, *10*, 459.

(9) Ye, J.; McGinnis, S.; Madden, T. L. BLAST: improvements for better sequence analysis. *Nucleic Acids Res.* **2006**, *34*, W6−W9.

(10) Madeira, F.; Pearce, M.; Tivey, A. R.; Basutkar, P.; Lee, J.; Edbali, O.; Madhusoodanan, N.; Kolesnikov, A.; Lopez, R. Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.* **2022**, *50*, W276−W279.

(11) Buchfink, B.; Reuter, K.; Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **2021**, *18*, 366−368.

(12) Höglund, A.; Dönnes, P.; Blum, T.; Adolph, H.-W.; Kohlbacher, O. MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics* **2006**, *22*, 1158−1165.
(13) Shatkay, H.; Höglund, A.; Brady, S.; Blum, T.; Dönnes, P.; Kohlbacher, O. SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. *Bioinformatics* **2007**, *23*, 1410−1417.
(14) Dalkiran, A.; Rifaioglu, A. S.; Martin, M. J.; Cetin-Atalay, R.; Atalay, V.; Doğan, T. ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinf* **2018**, *19*, 334.
(15) Srivastava, A.; Mahmood, A.; Srivastava, R. A comparative analysis of SVM random forest methods for protein function prediction. *CTCEEC 2017*; 2017; pp 1008−1010.
(16) Kulmanov, M.; Hoehndorf, R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* **2020**, *36*, 422−429.
(17) Ranjan, A.; Fernández-Baca, D.; Tripathi, S.; Deepak, A. An ensemble tf-idf based approach to protein function prediction via sequence segmentation. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2022**, *19*, 2685−2696.
(18) Szklarczyk, D.; Kirsch, R.; Koutrouli, M.; Nastou, K.; Mehryary, F.; Hachilif, R.; Gable, A. L.; Fang, T.; Doncheva, N. T.; Pyysalo, S.; et al. The STRING database in 2023: protein−protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* **2023**, *51*, D638−D646.
(19) Gligorijević, V.; Barot, M.; Bonneau, R. deepNF: deep network fusion for protein function prediction. *Bioinformatics* **2018**, *34*, 3873−3881.
(20) Saha, S.; Chatterjee, P.; Basu, S.; Nasipuri, M.; Plewczynski, D. FunPred 3.0: improved protein function prediction using protein interaction network. *PeerJ.* **2019**, *7*, No. e6830.
(21) Gligorijević, V.; Renfrew, P. D.; Kosciolek, T.; Leman, J. K.; Berenberg, D.; Vatanen, T.; Chandler, C.; Taylor, B. C.; Fisk, I. M.; Vlamakis, H.; et al. Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* **2021**, *12*, 3168.
(22) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.
(23) Ma, W.; Zhang, S.; Li, Z.; Jiang, M.; Wang, S.; Lu, W.; Bi, X.; Jiang, H.; Zhang, H.; Wei, Z. Enhancing protein function prediction performance by utilizing AlphaFold-predicted protein structures. *J. Chem. Inf. Model.* **2022**, *62*, 4008−4017.
(24) Ganesan, D.; Tendulkar, A. V.; Chakraborti, S. Protein Word Detection using Text Segmentation Techniques. *BioNLP 2017*; 2017; pp 238−246.
(25) LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278−2324.
(26) Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. arXiv:1301.3781. *arXiv preprint*. 2013. https://arxiv.org/abs/1301.3781 (accessed 2024-02-26).
(27) Cai, Y.; Wang, J.; Deng, L. SDN2GO: an integrated deep learning model for protein function prediction. *Front. Bioeng. Biotechnol.* **2020**, *8*, 391.
(28) The UniProt Consortium.; et al. UniProt :the universal protein knowledgebase in 2023. *Nucleic Acids Res.* **2023**, *51*, D523−D531.
(29) Paysan-Lafosse, T.; Blum, M.; Chuguransky, S.; Grego, T.; Pinto, B. L.; Salazar, G. A.; Bileschi, M. L.; Bork, P.; Bridge, A.; Colwell, L.; et al. InterPro in 2022. *Nucleic Acids Res.* **2023**, *51*, D418−D427.
(30) Kulmanov, M.; Hoehndorf, R. DeepGOZero: improving protein function prediction from sequence and zero-shot learning based on ontology axioms. *Bioinformatics* **2022**, *38*, i238−i245.
(31) Elhaj-Abdou, M. E.; El-Dib, H.; El-Helw, A.; El-Habrouk, M. Deep_CNN_LSTM_GO: protein function prediction from amino-acid sequences. *Comput. Biol. Chem.* **2021**, *95*, No. 107584.
(32) Xia, W.; Zheng, L.; Fang, J.; Li, F.; Zhou, Y.; Zeng, Z.; Zhang, B.; Li, Z.; Li, H.; Zhu, F. PFmulDL: a novel strategy enabling multi-class and multi-label protein function annotation by integrating diverse deep learning methods. *Comput. Biol. Med.* **2022**, *145*, No. 105465.

https://doi.org/10.1021/acs.jcim.3c01794
J. Chem. Inf. Model. 2024, 64, 1502−1511