Bioinformatics, 39(3), 2023, btad123
https://doi.org/10.1093/bioinformatics/btad123
Advance Access Publication Date: 8 March 2023
Original Paper

Systems biology

CFAGO: cross-fusion of network and attributes based on attention mechanism for protein function prediction

Zhourun Wu[1], Mingyue Guo[2], Xiaopeng Jin[3], Junjie Chen [1,*], Bin Liu [1,4,5,*]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, Guangdong 518055, China
[2]School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China
[3]College of Big Data and Internet, Shenzhen Technology University, Shenzhen, Guangdong 518118, China
[4]School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China
[5]Advanced Research Institute of Multidisciplinary Science, Beijing Institute of Technology, Beijing 100081, China

*Corresponding author. junjiechen@hit.edu.cn (J.C.); bliu@bliulab.net (B.L.)
Associate Editor: Pier Luigi Martelli

Received on 30 November 2022; revised on 28 February 2023; accepted on 5 March 2023

Abstract

Motivation: Protein function annotation is fundamental to understanding biological mechanisms. The abundant genome-scale protein–protein interaction (PPI) networks, together with other protein biological attributes, provide rich information for annotating protein functions. As PPI networks and biological attributes describe protein functions from different perspectives, it is highly challenging to cross-fuse them for protein function prediction. Recently, several methods combine the PPI networks and protein attributes via the graph neural networks (GNNs). However, GNNs may inherit or even magnify the bias caused by noisy edges in PPI networks. Besides, GNNs with stacking of many layers may cause the over-smoothing problem of node representations.

Results: We develop a novel protein function prediction method, CFAGO, to integrate single-species PPI networks and protein biological attributes via a multi-head attention mechanism. CFAGO is first pre-trained with an encoder–decoder architecture to capture the universal protein representation of the two sources. It is then fine-tuned to learn more effective protein representations for protein function prediction. Benchmark experiments on human and mouse datasets show CFAGO outperforms state-of-the-art single-species network-based methods by at least 7.59%, 6.90%, 11.68% in terms of m-AUPR, M-AUPR, and Fmax, respectively, demonstrating cross-fusion by multi-head attention mechanism can greatly improve the protein function prediction. We further evaluate the quality of captured protein representations in terms of Davies Bouldin Score, whose results show that cross-fused protein representations by multi-head attention mechanism are at least 2.7% better than that of original and concatenated representations. We believe CFAGO is an effective tool for protein function prediction.

Availability and implementation: The source code of CFAGO and experiments data are available at: http://bliulab.net/CFAGO/.

1 Introduction

Annotating protein functions is the key for unveiling the mechanism of disease, bringing great benefits for biomedical and pharmaceutical (Radivojac et al. 2013). Currently, protein functions are standardized by Gene Ontology (GO) (Ashburner et al. 2000; Carbon et al. 2021), which covers three aspects: biological process ontology (BPO), molecular function ontology (MFO), and cellular component ontology (CCO). Because biochemical experiments are expensive and time-consuming, it is impractical to experimentally annotate protein functions in large scale. In fact, only about 0.25% of known proteins have been experimentally annotated their functions (UniProt 2021). Therefore, to fill the huge vacancy of protein function annotation, developing effective and accurate computational protein function prediction methods is of great importance (Radivojac et al. 2013; Jiang et al. 2016; Zhou et al. 2019).

In the past decades, a lot of computational protein function prediction methods have been developed (Friedberg 2006; Lee et al. 2007; Rentzsch and Orengo 2009; Kihara 2016). Depending on their paradigms of feature extraction, these methods can be divided into four categories: sequence-based methods, structure-based methods, protein–protein interaction (PPI) network-based methods, and multi-source-based methods. As high sequence identity implies a similar function (Kimura and Ohta 1974; Lord et al. 2003), sequence-based methods infer protein functions by retrieving similar sequences (Cozzetto et al. 2013; Radivojac et al. 2013; Gong et al.

2016; You et al. 2018; Makrodimitris et al. 2019; Kulmanov and Hoehndorf 2020; Villegas-Morcillo et al. 2021; Kulmanov and Hoehndorf 2022). However, many proteins are similar in function, but not in sequence, so sequence-only-based methods are unable to predict functions for proteins with low sequence similarity. Protein structure determines its function, and proteins with similar structures usually share similar functions even when their sequence similarities are very low (Brenner et al. 1996; Holm and Sander 1996; Rost 1999). Therefore, structure-based methods detect the structure similarity between proteins to determine the functions of target proteins (Holm and Sander 1995; Gibrat et al. 1996; Laskowski et al. 2005). However, it is expensive to determine protein structures, and the amount of protein structure data is small. Although AlphaFold2 (Jumper et al. 2021) can predict protein structures from sequences, it has a limitation on the prediction of protein multi-chain structure that is the true structure for most proteins in living cells (Varadi et al. 2022). These facts limit the application of structure-based methods. On the other hand, as high-throughput techniques can screen PPIs in genome scale, predicting protein functions from PPI networks is desirable. PPI network-based methods assume similar functions usually shared by proteins with interaction (Sharan et al. 2007) or proteins with similar topological roles in PPI networks (Milenkovic and Przulj 2008). They predict protein functions either by label propagation among network nodes (Mostafavi et al. 2008; Mostafavi and Morris 2010) or by graph embedding of PPI network (Cho et al. 2016; Gligorijevic et al. 2018). However, high-throughput PPI data are incomplete and noisy due to the technical bias (De Las Rivas and Fontanillo 2010; Luck et al. 2020). Therefore, PPI networks alone cannot compactly describe protein functions. Protein information from multiple sources is complementary, such as PPI network and sequence (Kulmanov et al. 2018; Barot et al. 2021; You et al. 2021), in addition to subcellular location (Fan et al. 2020), text and sequence (You et al. 2018), structure and sequence (Gligorijevic et al. 2021; Lai and Xu 2022), etc. The last three Critical Assessment of Functional Annotation (CAFA) challenges have shown that the combination of different information indeed achieved the best performance on protein function prediction (Radivojac et al. 2013; Jiang et al. 2016; Zhou et al. 2019).

There are two main ways to combine protein information from multiple sources. The intuitive and simple way is concatenation, which directly concatenates the representations of multiple sources as the input of classifiers (Kulmanov et al. 2018). However, the concatenation fails to remove the effect of noise information from various sources. The widely used way is the graph neural networks (GNNs), which take the PPI network as graph and other information as node attribute features (Fan et al. 2020; You et al. 2021). But the message-passing mechanism of GNNs may inherit or even magnify the noise effect in networks (Dai and Wang 2021). Besides, GNNs with deep layers may cause the over-smoothing problem that all nodes tend to learn the same representation (Li et al. 2018; Cai and Wang 2020). Thus, it is urgent to propose a new method to integrate PPI network and other protein attributes into a more powerful representation.

In this study, we propose a new method called CFAGO to cross-fuse single-species PPI network and protein biological attributes via a multi-head attention mechanism. CFAGO contains a pre-training step and a fine-tuning step, and both of them use the multi-head attention mechanism to focus on important information. The pre-training step consists of an autoencoder, which can cross-fuse the effective information while ignoring the noise in the sources. The fine-tuning step learns more distinguishing protein representations for protein function annotation. The experimental results on human and mouse datasets show that CFAGO outperforms state-of-the-art single-species network-based protein function prediction methods, including pure PPI network-based methods and GNN-based fusion methods. Both the ablation study and protein representation visualization show the multi-head attention mechanism has an important contribution to fuse features of PPI network and other sources for single-species protein function prediction.

2 Materials and methods

2.1 Datasets

We conduct experiments on two species: *Homo sapiens* (human) and *Mus musculus* (mouse). The PPI data and protein sequence data are retrieved from the STRING (v11.5) database (Szklarczyk et al. 2021). In particular, we use the 'combined' type PPI data, which includes all of the 'experimental', 'coexpression', 'coocurrence', 'neighborhood', 'fusion', 'database', and 'textmining' types of PPI data. The protein function annotation data are retrieved from Gene Ontology Resource (http://geneontology.org) (version 2022-01-13 release) (Ashburner et al. 2000; Carbon et al. 2021). Protein subcellular location and protein domain data are retrieved from the UniProt database (v3.5.175) (UniProt 2021). Specifically, the protein domain from the pfam database (Mistry et al. 2021) is used.

Following the standard CAFA protocol (Radivojac et al. 2013; Jiang et al. 2016; Zhou et al. 2019), we extract experimental annotations of protein functions with evidence 'IDA', 'IPI', 'EXP', 'IGI', 'IMP', 'IEP', 'IC', or 'TA', and use two time points: t_0 (1 January 2018), t_1 (31 December 2020), to divide annotated proteins into training, validation, and testing sets. Concretely, the training set consists of proteins that have been annotated no later than t_0, validation set consists of proteins that only have been annotated in $(t_0, t_1]$, testing set consists of proteins that only have been annotated after t_1. We only use GO terms that have at least 10, 5, and 1 training, validation, and testing proteins, respectively. Furthermore, to reduce the effect of the dependence relationship between GO terms, we remove those GO terms annotating more than 5% of the species' PPI network proteins, following a previous study (Barot et al. 2021). The statistics of GO terms, training, validation, and testing sets used in this study is shown in Table 1.

2.2 Method

CFAGO introduces multi-head attention layers to cross-fuse protein information from different sources in two steps (Fig. 1). The first step is the pre-training, which is an encoder–decoder model that learns protein hidden embedding vectors by reconstructing original source features. The second step is fine-tuning, which combines the pre-trained encoder with a two-layer fully-connected neural network to predict protein functions.

2.2.1 PPI network structure and node attribute representations

For a protein, we use its first-order neighborhood of the PPI network to represent its network structure. Specifically, we first convert the PPI network into a weighted adjacency matrix, in which elements are weights of edges, then normalize elements to range $[0, 1]$ by min-max normalization. A column vector of the normalized adjacency matrix is a protein representation that contains normalized weights to its first-order neighborhoods.

For the protein attributes, we select the widely used protein domain and subcellular location information. Protein attributes are represented as binary vectors by bag-of-words encoding, which assigns 1 to an element in the binary vector if the protein is annotated with the corresponding domain or subcellular location. We filter out protein domain terms that appear less than 6 times in the

Table 1. Data statistics considered for each organism and Gene Ontology branch

Species	Statistics	BPO	MFO	CCO
Human	#GO terms	45	38	35
	#training proteins	3197	2747	5263
	#validation proteins	304	503	577
	#testing proteins	182	719	119
Mouse	#GO terms	42	17	37
	#training proteins	2714	1185	4014
	#validation proteins	336	232	694
	#testing proteins	155	126	147

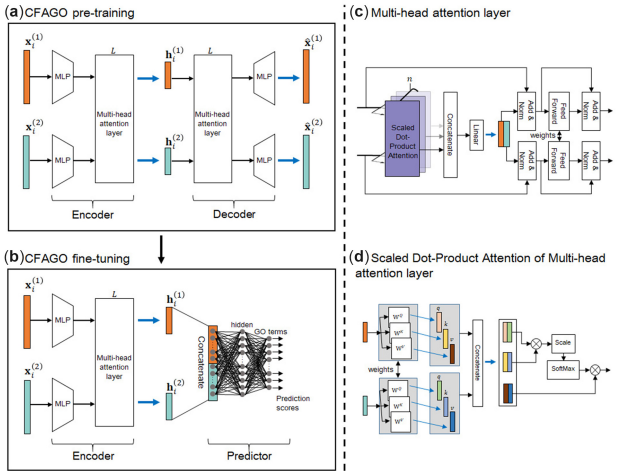

Figure 1 The flowchart of CFAGO. (a) Architecture of encoder–decoder for CFAGO pre-training step, where MLP stands for multilayer perceptron. (b) Architecture of CFAGO fine-tuning step. (c) Multi-head attention layer of Encoder and Decoder. (d) Scaled Dot-Product Attention of Multi-head attention layer.

dataset, following a previous study (Fan et al. 2020). Without prior knowledge of different attributes and GO aspects, we concatenate the two attribute vectors as the protein attribute vector representation for all GO aspects.

2.2.2 Multi-head attention layer

Here we define the multi-head attention layer following a previous study (Vaswani et al. 2017). The multi-head attention layer consists of multi-head attention, residual connection, normalization, and position-wise feed-forward networks. The core of multi-head attention is the Scaled Dot-Product Attention (Vaswani et al. 2017):

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{1}$$

where $\mathbf{Q},\mathbf{K},\mathbf{V}$ are the query matrix, key matrix, and value matrix, respectively, and d_k is the dimension size of key matrix. The multi-head attention is defined as (Vaswani et al. 2017):

$$\text{MultiHead}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{Concat}(head_1,\ldots head_n)\mathbf{W}^{\text{O}} \tag{2}$$

where n is the number of heads, $head_i$ is defined as:

$$head_i = \text{Attention}\left(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V\right) \tag{3}$$

where $\mathbf{W}_i^Q \in \mathrm{R}^{d\times d_k}$, $\mathbf{W}_i^K \in \mathrm{R}^{d\times d_k}$, $\mathbf{W}_i^V \in \mathrm{R}^{d\times d_k}$, $\mathbf{W}^O \in \mathrm{R}^{d\times d_k}$ are projection weight parameter matrices. Here d is the dimension size of hidden embedding vectors, and $d_k = d/n$.

The feed-forward network consists of two fully connected layers with a nonlinear activation function (Vaswani et al. 2017):

$$\text{FFN}(\boldsymbol{h}) = \mathbf{W}_2 f(\mathbf{W}_1\boldsymbol{h} + \boldsymbol{b}_1) + \boldsymbol{b}_2 \tag{4}$$

where f is the nonlinear activation function, $\mathbf{W}_1 \in \mathrm{R}^{d\times d_{ff}}$, $\mathbf{W}_2 \in \mathrm{R}^{d_{ff}\times d}$ are the weight parameter matrices of feed-forward network, d_{ff} is the output dimension size of the first linear layer, $\boldsymbol{b}_1 \in \mathrm{R}^{d_{ff}}$, $\boldsymbol{b}_2 \in \mathrm{R}^{d}$ are bias parameter vectors.

2.2.3 Pre-training with a self-supervised encoder–decoder

The pre-training step uses an encoder–decoder model to cross-fuse information from two sources. For protein i, its two original source features are represented as $\mathbf{x}_i^{(1)} \in \mathrm{R}^{d(1)}$ and $\mathbf{x}_i^{(2)} \in \mathrm{R}^{d(2)}$, where $d(m)$ is the feature dimension of source m.

Encoder: the encoder has two parallel multilayer perceptrons (MLPs), each for a source feature, and L multi-head attention layers. As original features of different sources may be sparse and differ in dimension, the original feature vector of protein i from source m, $\mathbf{x}_i^{(m)}$, is projected to a common vector with d dimensions by a two-layer MLP, which is defined as:

$$\text{MLP}(\mathbf{x}) = f\left(\text{LN}\left(\mathbf{W}_2 f(\text{LN}(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1))\right) + \mathbf{b}_2\right) \tag{5}$$

where f is the nonlinear activation function, LN is the layer normalization function (Ba et al. 2016), $\mathbf{W}_1 \in R^{d(m)\times d_e}$ and $\mathbf{W}_2 \in R^{d_e\times d}$ are the weight matrices, $\mathbf{b}_1 \in R^{d_e}$ and $\mathbf{b}_2 \in R^{d}$ are the bias vectors, d_e is the size of the MLP hidden layer. Then, the projected vectors of two sources are cross-fused by multi-head attention layers to generate protein-hidden embedding vectors.

Decoder: the structure of the decoder is symmetric to the encoder. The decoder first feeds the hidden embedding vectors into L multi-head attention layers. Then for protein i, the feature vector of source m is reconstructed by an MLP whose structure is symmetric to the corresponding MLP in encoder, denoting as $\hat{\mathbf{x}}_i^{(m)}$. In decoder, the sigmoid function is used as the activation function of the output layer of MLPs.

The aim of the encoder–decoder is to minimize the sample-wise binary cross-entropy loss between original and reconstructed source features:

$$loss(\Theta) = \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{2}\sum_{j=1}^{d(m)} -\left[x_{ij}^{(m)}\log\hat{x}_{ij}^{(m)} + \left(1 - x_{ij}^{(m)}\right)\log\left(1 - \hat{x}_{ij}^{(m)}\right)\right] \tag{6}$$

where N is the number of total proteins in PPI network, $x_{ij}^{(m)}$ and $\hat{x}_{ij}^{(m)}$ are the jth dimension value of $\mathbf{x}_i^{(m)}$ and $\hat{\mathbf{x}}_i^{(m)}$, respectively, Θ is the set of all parameters in the pre-training step.

2.2.4 Fine-tuning for protein function prediction

In this study, protein function prediction is modeled as a multi-label task. We extract the pre-trained encoder and attach it with a predictor, which is a two-layer perceptron, to predict protein labels. Let the number of target GO terms to be K. The predictor takes the concatenation of the embedding vectors, denoting as $\mathbf{h}_i^{(1)}$ and $\mathbf{h}_i^{(2)}$, of the two sources generated by the encoder as input, and output a K-dimension score vector for GO terms Formally, the prediction score vector $[p_{i1},\ldots p_{iK}]^T$ of protein i is defined as:

$$[p_{i1},\ldots p_{iK}]^T = \sigma\left(\mathbf{W}_o\sigma\left(\mathbf{W}_h\left(\left|\left|\boldsymbol{h}_i^{(m)}\right.\right) + \mathbf{b}_h\right) + \mathbf{b}_o\right) \tag{7}$$

where $||$ is concatenation operator, and σ is the sigmoid function, d_h is the size of the predictor's hidden layer, $\mathbf{W}_h \in R^{2d\times d_h}$ and $\mathbf{W}_o \in R^{d_h\times K}$ are the weight matrices of predictor's hidden and output layers, respectively. $\mathbf{b}_h \in R^{d_h}$ and $\mathbf{b}_o \in R^{K}$ are the bias vectors of predictor's hidden and output layers, respectively.

For GO terms, negative proteins are much more than positive proteins in training set. Therefore, we use the asymmetric loss (Ridnik et al. 2021) as the prediction loss:

$$\begin{aligned} ASL(\Phi) = \frac{1}{N_{train}K}\sum_{i=1}^{N_{train}}\sum_{k=1}^{K} & -y_{ik}(1-p_{ik})^{\gamma_+}\log(p_{ik}) \\ & -(1-y_{ik})(p_{ik})^{\gamma_-}\log(1-p_{ik}) \end{aligned} \tag{8}$$

where $y_{ik} \in \{0,1\}$ and $p_{ik} \in [0,1]$ are the true label and predicted score of protein i in terms of GO term k, γ_+, and γ_- are the positive and negative focusing parameters, respectively, N_{train} is the number of proteins in training set, Φ is the set of all parameters in the fine-tuning step. In this study we set $\gamma_+ = 0$ and $\gamma_- = 2$.

2.3 Evaluation metrics

In this study, we use five metrics to evaluate prediction performance, including two types of area under the precision–recall curve (AUPR), e.g. micro-averaged AUPR (m-AUPR) and Macro-averaged AUPR (M-AUPR), F1-score (F1), accuracy (ACC), and F-max score (Fmax). The first three metrics are function-centric measures that evaluate proteins annotated to each GO term, while the last two

metrics are protein-centric measures that evaluate GO terms annotated to each protein.The m-AUPR, M-AUPR, F1, and ACC are widely used to evaluate protein function prediction (Mostafavi et al. 2008; Cho et al. 2016; Gligorijevic et al. 2018; Barot et al. 2021). Specifically, m-AUPR is the AUPR across the vectorized results of true label and prediction matrices, M-AUPR is the average of AUPRs of all GO terms. F1 is computed by taking the top three prediction scores for each protein, then constructing a two-by-two confusion matrix for each GO term, and calculating the harmonic mean of precision and recall on the summed-up confusion matrix of all GO terms. Accuracy is the proportion of proteins that the predicted GO terms are exactly the same as the true GO terms, using 0.5 as the predicted threshold. Fmax is used for the CAFA challenging (Radivojac et al. 2013; Jiang et al. 2016; Zhou et al. 2019) and many protein function prediction studies (Kulmanov and Hoehndorf 2020; Barot et al. 2021; Gligorijevic et al. 2021; You et al. 2021; Lai and Xu 2022), which is defined as following:

$$\mathrm{Fmax} = \max_{\tau} \left\{ \frac{2 \times pr(\tau) \times rc(\tau)}{pr(\tau) + rc(\tau)} \right\} \tag{9}$$

where τ is the threshold value, $pr(\tau)$ and $rc(\tau)$ are the precision and recall in terms of τ, respectively, which are defined as:

$$\begin{cases} pr(\tau) = \dfrac{1}{q(\tau)} \displaystyle\sum_{i=1}^{q(\tau)} \frac{\sum_k \mathbb{I}(p_{ik} \geq \tau \wedge y_{ik} \equiv 1)}{\sum_k \mathbb{I}(p_{ik} \geq \tau)} \\ rc(\tau) = \dfrac{1}{g} \displaystyle\sum_{i=1}^{g} \frac{\sum_k \mathbb{I}(p_{ik} \geq \tau \wedge y_{ik} \equiv 1)}{\sum_k \mathbb{I}(y_{ik} \equiv 1)} \end{cases} \tag{10}$$

where $\mathbb{I}(\cdot)$ is the indicator function, $q(\tau)$ is the number of proteins whose max predicted score is not less than τ, g is the number of target proteins.

In addition, we use the Davies Bouldin Score (Davies and Bouldin 1979) to evaluate the goodness of feature representations. Lower Davies Bouldin Score means proteins with same functions are clustered together better.

3 Experiments

3.1 Experimental setup

We evaluate CFAGO on the three GO aspects: BPO, MFO, and CCO separately. We use the same empirical hyperparameter set for both species and all of the three GO aspects, and merge the validation and training sets for each aspect to train our model. Specifically, the batch size is 32 for both of pre-training and fine-tuning steps, the encoder MLP hidden dimension size $d_e = 1024$, hidden embedding vector dimension size $d = 512$, the number of multi-head attention layers $L = 6$, feed-forward network hidden dimension $d_{ff} = 2048$, number of attention heads $n = 8$, predictor hidden layer dimension $d_h = 256$, normalization function is set as the layer normalization (Ba et al. 2016). We use the Gaussian Error Linear Unit (Hendrycks and Gimpel 2016) as the nonlinear activation in all hidden layers in encoder and decoder, followed by a dropout layer. In predictor, the dropout rate is set as 0.3, while in encoder and decoder, the dropout rate is set as 0.1. The pre-training step is trained with 5000 epochs, learning rate of 1e-5 for the first 2500 epochs and 1e-6 for the remaining epochs. The fine-tuning step is trained with 100 epochs. In the first 50 epochs, we freeze the pre-trained encoder, and set the learning rate of 1e-4 for predictor. In the last 50 epochs, we set the learning rate of 1e-6 for the encoder, and set the learning rate of 1e-5 for predictor. The optimizer is AdamW (Loshchilov and Hutter 2019).

Here, we compare CFAGO with eight state-of-the-art PPI network-based methods, including two baseline methods, four network integrate methods, and two GNN-based methods. The baseline methods include the Naïve and BLAST methods of CAFA (Radivojac et al. 2013). The network integrate methods include deepNF (Gligorijevic et al. 2018), Mashup (Cho et al. 2016), GeneMANIA (Mostafavi et al. 2008), and NetQuilt (Barot et al. 2021). The deepNF, Mashup, and GeneMANIA integrate multiple types of single-species PPI networks into a single kernel or compact low-dimensional representations, while NetQuilt globally aligns different species' PPI networks into a meta-network profile. The GNN-based methods include Graph2GO (Fan et al. 2020) and DeepGraphGO (You et al. 2021). Since this study focuses on the information cross-fusion from multi-sources of single species, all of the competing methods are fitted on single species datasets by hyperparameters reported on their papers. Besides, their features are generated by using their feature-generating tools or procedures. Because GeneMANIA, Mashup, deepNF, and Graph2GO did not conduct experiments on mouse, we use their reported structure and hyperparameters for human to evaluate their performance on both species. For Graph2GO, the validation and training sets of each aspect are merged to train the model. All of the results are averaged by five random repeats.

3.2 CFAGO outperforms competing methods

Figure 2 shows the performance of different methods on testing datasets of human and mouse for the three GO aspects. It is clear that CFAGO outperforms all of the competing methods in terms of m-AUPR, M-AUPR, and Fmax measures. Specifically, in terms of m-AUPR, CFAGO achieves (7.59%, 37.58%), (78.65%, 43.69%), and (89.89%, 85.91%) higher than that of competing methods on (human, mouse) datasets for BPO, MFO, and CCO, respectively. CFAGO achieves (6.90%, 45.86%), (16.10%, 13.45%), and (38.01%, 47.65%) higher in terms of M-AUPR, and (11.68%, 31.88%), (26.47%, 22.01%), and (23.57%, 34.30%) higher in terms of Fmax than that of competing methods on (human, mouse) datasets for BPO, MFO, and CCO, respectively.

For the F1 measurement, CFAGO also outperforms the competing methods, only except on mouse MFO where BLAST and Graph2GO achieve better performance. This is not surprising,

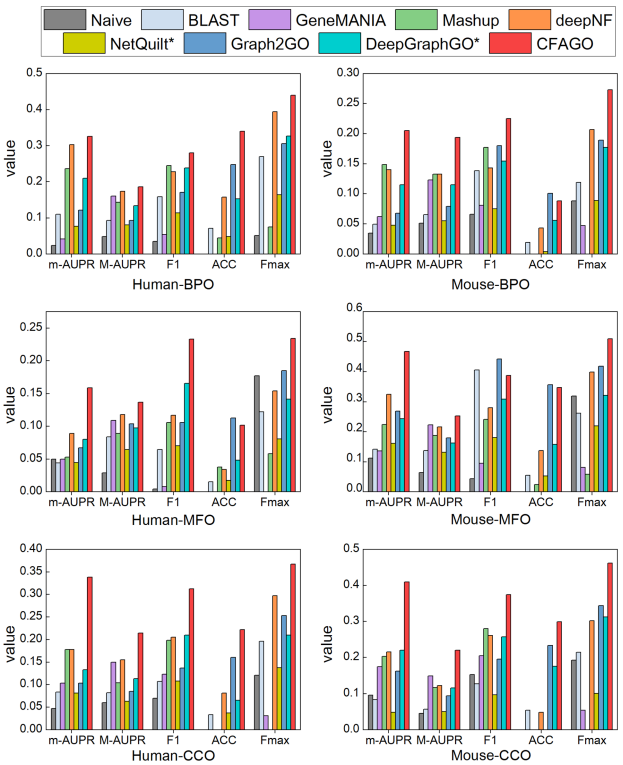

Figure 2 Performance comparison of CFAGO with competing methods. CFAGO achieves better or comparable performance compared to the competing methods in terms of all measurements. The blank gaps on the ACC and Fmax measurements mean the value of corresponding methods are 0. Methods labelled with an asterisk are multi-species methods but have been trained as single-species methods in the comparisons.

because several studies have pointed out that MFO is highly correlated with sequence information (Lord et al. 2003; Fan et al. 2020), and Graph2GO has the additional information from protein sequence similarity networks, in addition to PPI networks. For the ACC measurement, CFAGO outperforms competing methods on half of tasks. For the worst case of CFAGO, it has comparable performance with Graph2GO on human MFO, mouse BPO, and MFO, while outperforming other seven competing methods. We also noticed that the frequencies of individual labels on all datasets except on mouse BPO are much lower than 0.5 in the training dataset. Since Naïve uses the frequency of each label on the training dataset as the prediction score for every test protein, it achieves a value of 0 in terms of the ACC measurement on the corresponding datasets. Besides, as training proteins cover less than 30% of total proteins in the PPI network, GeneMANIA cannot effectively propagate labels from training proteins to test proteins, and Mashup is unable to learn protein compact low-dimensional via matrix factorization, leading them achieve a value of 0 in terms of ACC and Fmax on several tasks.

We further investigate the AUPR performance of CFAGO in individual GO terms. The results in Supplementary Figs S1–S6 show that the performance of linked GO terms is not correlated. These results are expected. The first reason is that proteins are annotated to the most granular term (Ashburner et al. 2000), the second reason is that we remove those annotating more than 5% of proteins in the PPI network to reduce the effect of the dependence relationship between GO terms.

Such outstanding results of CFAGO demonstrate the feature cross-fusion via multi-head attention mechanism has obvious advantages compared with pure PPI network-based methods and the multi-source combining method based on GNN.

3.3 Attention mechanism learns more distinguishing representation via cross-fusing information from multiple sources

To quantitatively evaluate the effectiveness of cross-fusion by attention mechanism and pre-training, we compare the distinguishing power of protein representations by Davies Bouldin Score (Davies and Bouldin 1979).

We compare four types of protein representations, including the original PPI network representation, original attribute representation, and hidden embedding representations learned by CFAGO with and without attention mechanism. The structure of CFAGO without multi-head attention mechanism is shown in Supplementary Fig. S7. The GO term sets are used as the protein cluster labels, that is, two proteins with exactly the same GO term set are considered to be in the same cluster. The results on the union of training and validation set are listed in Table 2, which shows that the hidden embedding representations learned by full CFAGO model achieve the best performance. Specifically, its Davies Bouldin Score is (6.15%, 8.39%), (8.02%, 5.51%), and (6.30%, 7.63%) lower than that of comparison representations on (human, mouse) datasets for BPO, MFO, and CCO, respectively. These results indicate that cross-fusion of CFAGO can effectively fuse protein features from multiple sources to generate better protein representation for function prediction.

Table 2 Davies Bouldin Score comparison of different protein feature represents

Representation	Human			Mouse		
	BPO	MFO	CCO	BPO	MFO	CCO
o_PPI[a]	1.855	2.250	2.243	1.991	3.133	2.204
o_attribute[b]	2.128	2.387	2.128	2.209	3.349	2.122
c_embedding[c]	1.884	2.183	2.201	1.943	2.924	2.179
cf_embedding[d]	**1.741**	**2.008**	**1.994**	**1.780**	**2.763**	**1.960**

Note: Smaller Davies Bouldin Score value means the cluster of protein representations are more clearly separated.

[a]The original PPI network structure feature.

[b]The original attribute feature.

[c]The concatenation of hidden embedding vectors output by CFAGO without attention mechanism.

[d]The concatenation of hidden embedding vectors output by CFAGO.

In addition, we visualize the distribution of above protein representations on human and mouse datasets for the three aspects of GO via t-Distributed Stochastic Neighbor Embedding (van der Maaten and Hinton 2008) by assigning a unique color for each cluster label (GO term set), as shown in Fig. 3. It is clear that the distribution of original PPI network structure and attribute features are different, indicating they are complementary for annotating protein function. The original PPI network structure feature shows a relatively clear distribution, while the original attribute feature isolates proteins into a ring part and a nucleus part that each of which shows more vague cluster segmentation.

The representation learned by CFAGO without cross-fusion mixes up the shape of distribution of original PPI network structure and original attribute, showing an edge that is similar to the ring part in the distribution of original attribute. The hidden embedding representation learned by full CFAGO model shows a much better cluster segmentation than compared representations. These results show that the features from two sources are indeed cross-fused by multi-head attention mechanism.

3.4 The contribution of self-supervised pre-training and multi-head attention mechanism

Here we conduct ablation experiments to study the contribution of self-supervised pre-training and attention mechanism to performance improvement (Fig. 4). Supplementary Fig. S7 shows the CFAGO model that removed multi-head attention mechanism. We only use the m-AUPR, M-AUPR, and Fmax measures, as F1 and ACC measures depend on special thresholds.

We observe that the performance drops significantly if not applying self-supervised pre-training. The reason is that the CFAGO model contains a huge amount of parameters. As there are only several thousands of training proteins, the CFAGO model is over-fitted without applying pre-training. For the multi-head attention mechanism, the performance of CFAGO drops clearly without it, except on Fmax measure of mouse MFO. The reason is that there are only 1185 mouse MFO training proteins, therefore adding the attention mechanism makes the CFAGO model overfitted, leading the performance drops down. Overall, these results also demonstrate the feature cross-fusion via multi-head attention mechanism has obvious advantages compared with concatenation of features.

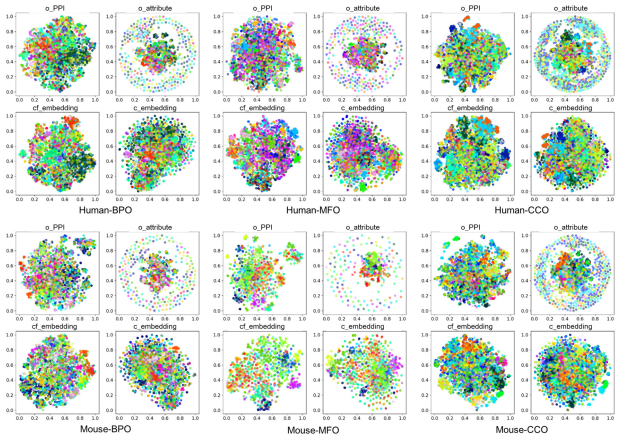

Figure 3 Visualization of different feature representations on human and mouse dataset in terms of the BPO, MFO, and CCO, respectively. o_PPI is the original PPI network structure feature, o_attribute is the original biological attribute feature, cf_embedding and c_embedding is the concatenation of hidden embedding vectors output by CFAGO, with and without attention mechanism, respectively.

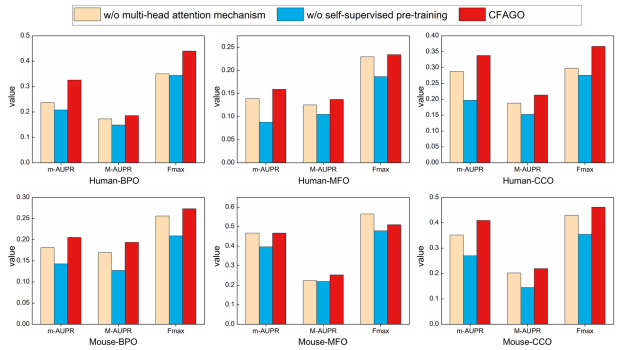

Figure 4 Ablation studies of self-supervised pre-training and multi-head attention mechanism, the meaning of different colors is shown on top of the graph. m-AUPR, M-AUPR, and Fmax are performance measurements.

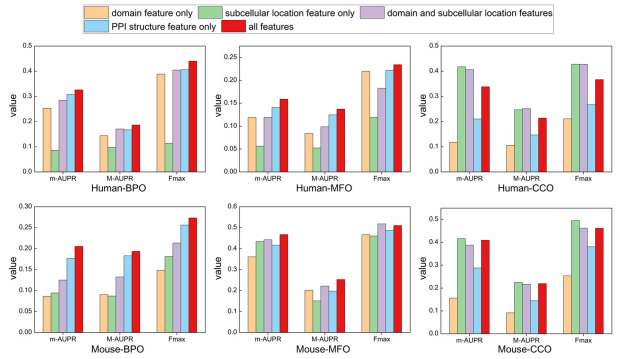

Figure 5 Performance comparison of different feature combinations. The meaning of different color is shown on top of the graph. m-AUPR, M-AUPR, and Fmax are performance measurements.

3.5 Contribution of different features

Here, we analyze the contribution of different features for the three GO aspects. We test five feature conditions: protein domain feature only, protein subcellular location feature only, concatenation of protein domain and subcellular location features, PPI network structure feature only, and combination of all features. Figure 5 shows the results. For BPO task, combining all of the features shows the best performance, and the PPI network structure feature contributes the most for improving prediction performance. For MFO task, combining all of the features shows the best performance, except on Fmax measure of mouse MFO. The PPI network structure feature contributes the most on human, while the combination of protein domain and subcellular location feature contributes the most on mouse. For CCO task, using the protein subcellular location feature only shows the best performance. By combining protein domain feature, the performance drops a bit, by combining PPI network structure feature, the performance drops significantly.

The reduced performance in terms of Fmax on mouse MFO is caused by the noise in protein attribute data, and insufficient number of training proteins. Protein domains came from the pfam (Mistry et al. 2021) database that contains unverified domains (Mistry et al. 2021), and the subcellular location data came from the UniProt database (UniProt 2021) that contains unreviewed records (MacDougall et al. 2020). Therefore the protein domain and subcellular location features contain noise. The reduced performance on CCO is caused by the noise in both protein attribute data and PPI data. The PPI data are produced by high-throughput techniques which contain inherent bias noise or predicted computationally (Szklarczyk et al. 2021). Besides, in the STRING database, the ratio of PPIs derived from 'textmining' in human dataset is 27.75%, and the ratio in mouse dataset is 19.28%. Since this kind of PPIs is predicted from scientific literature, they likely connect proteins from different subcellular locations (Szklarczyk et al. 2021). Therefore, the PPIs derived from 'textmining' became noise for CCO prediction.

4 Conclusion

In this study, we propose CFAGO, an attention mechanism-based neural network model, for protein function prediction. It cross-fuses the information from multiple sources of single species using an attention mechanism to learn more effective protein representations. Specifically, CFAGO is the first pre-trained via an encoder–decoder architecture to learn the universal protein representations and then fine-tuned to further improve protein function prediction. We show that CFAGO outperforms state-of-the-art single-species network-based protein function prediction methods in both human and mouse organisms. CFAGO would be an effective tool for understanding disease mechanisms or finding drug targets.

Several studies (Barot et al. 2021; You et al. 2021) have shown that integrating PPI networks of multiple species can further improve the accuracy of protein function prediction. In future work, we will try more types of protein attributes such as sequence features, and explore effective ways that can use full homology information to integrate PPI networks of multi-species for protein function prediction.

Acknowledgements

The authors are very much indebted to the anonymous reviewers, whose constructive comments are very helpful for strengthening the presentation of this article.

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 62102118, 62271049, and U22A2039), Project of Educational Commission of Guangdong Province of China (Grant No. 2021KQNCX274), the Shenzhen Colleges and Universities Stable Support Program (Grant No. GXWD20220811170504001 and 20220715183602001).

Conflict of interest: None declared.

References

Ashburner M, Ball CA, Blake JA *et al.* Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat Genet* 2000;**25**:25–9.

Ba JL, Kiros JR, Hinton GE. Layer Normalization. 2016. https://doi.org/10.48550/ARXIV.1607.06450.

Barot M, Gligorijević V, Cho K *et al.* NetQuilt: deep multispecies network-based protein function prediction using homology-informed network similarity. *Bioinformatics* 2021;**37**:2414–2422.

Brenner SE, Chothia C, Hubbard TJ *et al.* Understanding protein structure: using scop for fold interpretation. *Methods Enzymol* 1996;**266**:635–43.

Cai C, Wang Y. A note on over-smoothing for graph neural networks. *CoRR* 2020;abs/2006.13318

Carbon S, Douglass E, Good BM *et al.* The Gene Ontology Consortium et al. The gene ontology resource: enriching a GOld mine. *Nucleic Acids Res* 2021;**49**:D325–D334.

Cho H, Berger B, Peng J. Compact integration of multi-network topology for functional analysis of genes. *Cell Syst* 2016;**3**:540–8. e545.

Cozzetto D, Buchan DWA, Bryson K *et al.* Protein function prediction by massive integration of evolutionary analyses and multiple data sources. *BMC Bioinform* 2013;**14 Suppl 3**:S1.

Dai E, Wang S. Say no to the discrimination: learning fair graph neural networks with limited sensitive attribute information. In: *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2021. pp. 680–8.

Davies DL, Bouldin DW. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell* 1979;**1**:224–7.
De Las Rivas J, Fontanillo C. Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput Biol* 2010;**6**:e1000807.
Fan K, Guan Y, Zhang Y. Graph2GO: a multi-modal attributed network embedding method for inferring protein functions. *Gigascience* 2020;**9**: giaa081.
Friedberg I. Automated protein function prediction–the genomic challenge. *Brief Bioinform* 2006;**7**:225–42.
Carbon S, Douglass E, Good BM, The Gene Ontology Consortium *et al.* The gene ontology resource: enriching a GOld mine. *Nucleic Acids Res* 2021;**49**: D325–D334.
Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. *Curr Opin Struct Biol* 1996;**6**:377–85.
Gligorijevic V, Barot M, Bonneau R. deepNF: deep network fusion for protein function prediction. *Bioinformatics* 2018;**34**:3873–81.
Gligorijevic V *et al.* Structure-based protein function prediction using graph convolutional networks. *Nat Commun* 2021;**12**:3168.
Gong Q, Ning W, Tian W. GoFDR: a sequence alignment based method for predicting protein functions. *Methods* 2016;**93**:3–14.
Hendrycks D, Gimpel K. Gaussian error linear units (gelus). 2016.
Holm L, Sander C. Dali: a network tool for protein structure comparison. *Trends Biochem Sci* 1995;**20**:478–80.
Holm L, Sander C. Mapping the protein universe. *Science* 1996;**273**:595–603.
Jiang Y, Oron TR, Clark WT *et al.* An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol* 2016; **17**:1–19.
Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.
Kihara D. Computational protein function predictions. *Methods* 2016;**93**:1–2.
Kimura M, Ohta T. On some principles governing molecular evolution. *Proc Natl Acad Sci USA* 1974;**71**:2848–52.
Kulmanov M, Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. *Bioinformatics* 2020;**36**:422–9.
Kulmanov M, Hoehndorf R. DeepGOZero: improving protein function prediction from sequence and zero-shot learning based on ontology axioms. *Bioinformatics* 2022;**38**:i238–i245.
Kulmanov M, Khan MA, Hoehndorf R *et al.* DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics* 2018;**34**:660–8.
Lai B, Xu J. Accurate protein function prediction via graph attention networks with predicted structure information. Brief Bioinform 2022;23:bbab502.
Laskowski RA, Watson JD, Thornton, JM. Protein function prediction using local 3D templates. *J Mol Biol* 2005;**351**:614–26.
Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol* 2007;**8**:995–1005.
Li Q, Han Z, Wu X-M. Deeper insights into graph convolutional networks for semi-supervised learning. *CoRR* 2018;abs/1801.07606.
Lord PW, Stevens RD, Brass A *et al.* Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics* 2003;**19**:1275–83.
Loshchilov I, Hutter F. Decoupled weight decay regularization. In: *7th International Conference on Learning Representations*. New Orleans, LA, USA: OpenReview.net, 2019.
Luck K, Kim D-K, Lambourne L *et al.* A reference map of the human binary protein interactome. *Nature* 2020;**580**:402–8.
MacDougall A, Volynkin V, Saidi R, UniProt Consortium *et al.* UniRule: a unified rule resource for automatic annotation in the UniProt Knowledgebase. *Bioinformatics* 2020;**36**:4643–8.
Makrodimitris S, van Ham R, Reinders MJT. Improving protein function prediction using protein sequence and GO-term similarities. *Bioinformatics* 2019;**35**:1116–24.
Milenkovic T, Przulj N. Uncovering biological network function via graphlet degree signatures. *Cancer Inform* 2008;**6**:CIN.S680–273.
Mistry J, Chuguransky S, Williams L *et al.* Pfam: the protein families database in 2021. *Nucleic Acids Res* 2021;**49**:D412–19.
Mostafavi S, Morris Q. Fast integration of heterogeneous data sources for predicting gene function with limited annotation. *Bioinformatics* 2010;**26**: 1759–65.
Mostafavi S, Ray D, Warde-Farley D *et al.* GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol* 2008;**9 Suppl 1**:S4.
Radivojac P, Clark WT, Oron TR *et al.* A large-scale evaluation of computational protein function prediction. *Nat Methods* 2013;**10**:221–7.
Rentzsch R, Orengo CA. Protein function prediction–the power of multiplicity. *Trends Biotechnol* 2009;**27**:210–9.
Ridnik T *et al.* Asymmetric loss for multi-label classification. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 82–91.
Rost B. Twilight zone of protein sequence alignments. *Protein Eng* 1999;**12**: 85–94.
Sharan R, Ulitsky I, Shamir R. Network-based prediction of protein function. *Mol Syst Biol* 2007;**3**:88.
Szklarczyk D, Gable AL, Nastou KC *et al.* The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* 2021;**49**: D605–12.
UniProt C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2021;**49**:D480–9.
van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–605.
Varadi M, Anyango S, Deshpande M *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* 2022;**50**:D439–44.
Vaswani A *et al.* Attention is all you need. *Adv Neur In* 2017;**30**:6000–6010.
Villegas-Morcillo A, Makrodimitris S, van Ham RCHJ *et al.* Unsupervised protein embeddings outperform hand-crafted sequence and structure features at predicting molecular function. *Bioinformatics* 2021;**37**:162–70.
You R, Huang X, Zhu S. DeepText2GO: improving large-scale protein function prediction with deep semantic text representation. *Methods* 2018;**145**: 82–90.
You R, Yao S, Mamitsuka H *et al.* DeepGraphGO: graph neural network for large-scale, multispecies protein function prediction. *Bioinformatics* 2021; **37**:i262–71.
You R, Zhang Z, Xiong Y *et al.* GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics* 2018; **34**:2465–73.
Zhou N, Jiang Y, Bergquist TR *et al.* The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol* 2019;**20**:1–23.