




GOGCN: Graph Convolutional Network on Gene Ontology for Functional Similarity Analysis of Genes

Zhen Tian , Haichuan Fang , Zhixia Teng , and Yangdong Ye 

Abstract—The measurement of gene functional similarity plays a critical role in numerous biological applications, such as gene clustering, the construction of gene similarity networks. However, most existing approaches still rely heavily on traditional computational strategies, which are not guaranteed to achieve satisfactory performance. In this study, we propose a novel computational approach called **GOGCN** to measure gene functional similarity by modeling the Gene Ontology (**GO**) through Graph Convolutional Network (**GCN**). GOGCN is a graph-based approach that performs sufficient representation learning for terms and relations in the GO graph. First, GOGCN employs the GCN-based knowledge graph embedding (KGE) model to learn vector representations (i.e., embeddings) for all entities (i.e., terms). Second, GOGCN calculates the semantic similarity between two terms based on their corresponding vector representations. Finally, GOGCN estimates gene functional similarity by making use of the pair-wise strategy. During the representation learning period, GOGCN promotes semantic interaction between terms through GCN, thereby capturing the rich structural information of the GO graph. Further experimental results on various datasets suggest that GOGCN is superior to the other state-of-the-art approaches, which shows its reliability and effectiveness.

Index Terms—Graph convolutional network, knowledge graph embedding, Gene Ontology, gene functional similarity

1 INTRODUCTION

IN recent years, many gene functional similarity approaches have been proposed and widely used in computational molecular biology, such as gene function analysis and prediction [1], [2], [3], gene clustering [4], protein interaction prediction [5], disease gene prioritization [6]. Compared with sequence and structure similarity, the functional similarity is more informative for understanding the biological roles and functions of genes [7].

GO is a directed acyclic graph comprising three orthogonal ontologies: biological process (BP), cellular component (CC), and molecular function (MF). Genes and their products, collectively called genes in this article, are usually annotated with diverse GO terms, which is useful for describing the behavior of genes. Accordingly, it is quite effective that utilizing GO annotation to measure gene functional similarity. In this premise, many approaches based

on the GO graph and GO annotation have been proposed for measuring gene functional similarity. On the whole, these approaches could be generally divided into two categories: pair-wise and group-wise approaches.

Generally speaking, for pair-wise approaches, the functional similarity between two genes can be inferred from the semantic similarity between their corresponding annotated terms. In consequence, there are two main steps for pair-wise approaches when estimating gene functional similarity. The first step is computing semantic similarity values between annotated terms by taking advantage of term comparison measures. The second step is to calculate gene functional similarity values using the obtained semantic similarity values. Three main rules—average rule (AVG), maximum rule (MAX), and best match average rule (BMA)—are applied into the second step. For example, Resnik [8] proposed a pair-wise approach where the semantic similarity between two terms is equal to the information content (IC) of their lowest common ancestor (LCA) term. Jiang and Conrath [9], and Lin [10] took the specificity of terms themselves into consideration. By contrast, Wang [11] designed an improved algorithm to measure semantic similarity between terms by exploiting the inherited relationship of terms. Each edge in the GO graph is assigned a weight parameter called the semantic contribution factor. Bien [12] improved method Wang by taking the entire GO structure into consideration. Method simDEF [13] measured semantic similarity between terms by exploiting their definition in a specific corpus. Besides, Yu [14] put forward HPHash model to compute the taxonomic similarity between GO terms.

Although traditional pair-wise approaches are used widely in past decades, they may have some drawbacks.

- Zhen Tian, Haichuan Fang, and Yangdong Ye are with the Department of School of Information Engineering, Zhengzhou University, Zhengzhou, Henan 450001, China. E-mail: {ieztian, iezydye}@zzu.edu.cn, hcfang@gs.zzu.edu.cn.
- Zhixia Teng is with the Department of College of Information and Computer Engineering, Northeast Forestry University, Harbin, Heilongjiang 150040, China. E-mail: tengzhixia@nefu.edu.cn.

Manuscript received 25 July 2021; revised 26 Jan. 2022; accepted 3 June 2022. Date of publication 10 June 2022; date of current version 3 Apr. 2023.

This work was supported in part by the National Natural Science Foundation of China under Grants 62176239, 61901103, and 61801432, in part by the Natural Science Foundation of Heilongjiang Province under Grant LH2019F002, and in part by the Postdoctoral Science Foundation of Heilongjiang Province of China under Grant LBH-Z19106.

(Corresponding author: Yangdong Ye.)

Digital Object Identifier no. 10.1109/TCBB.2022.3181300

For example, method Resnik and Lin may suffer from 'shallow annotation' problem [11]. Concerning method Wang, it may cause semantic loss or overload when we assign 0.8 and 0.6 weight values to the relation 'is a' and 'part of' respectively. Moreover, method Wang measured the semantic value of terms considering their ancestor terms only. For method simDEF, its performance depends on the context definition of terms to a large extent.

According to the true path rule, a gene annotated with some terms also is annotated with the ancestors of these terms [7]. Accordingly, group-wise approaches measure gene functional similarity by integrating the annotated terms and their ancestors into a set firstly. Then, the functional similarity between genes is translated into the similarity between their corresponding annotated term sets. Gentleman [15] proposed an approach called simUI that takes the ratio of the number of terms between term sets as the final functional similarity. Subsequently, many approaches improved simUI by considering the IC of terms. For example, Yu [16] computed functional similarity of proteins with their overlap of GO annotations term sets. Inspired by Sanchez, all descendants of the term contributed to the calculation of IC in SORA [7]. After that, WIS [17] suggested that the IC of a term has a great connection with its depth in the GO graph and made some improvements. Furthermore, SORA and WIS considered that the IC of a term contains two parts: the first one called extended semantics is originated from itself and the second one called inherited semantics is inherited from its direct ancestor, which demonstrated that SORA and WIS made the best use of the specificity of annotation terms. Apart from the aforementioned, some effective models [18], [19] based on vector space were proposed. In this study, we refer to the basic vector space model mentioned by Zhang [19] as VSM. Specifically, VSM model adopted the one-hot encoding style to translate the term set into a vector whose dimension is equal to the term number and each dimension corresponds to one annotation term, indicating whether it annotates the gene or not. Afterward, VSM estimated the gene functional similarity by computing the cosine similarity of two vectors. Yu [20] proposed a novel method called HashGO and measured the semantic similarity between proteins based on their low-dimensional representations, which was constructed by protein-term association matrix with a series of hash functions.

To the best of our knowledge, group-wise approaches also have some shortcomings. Taking the aforementioned methods as examples, Method simUI ignored the semantics of terms and the relationship between terms. Method Sanchez only took the leaf terms and directed descendant terms into consideration when calculating the IC of terms. SORA and WIS relied heavily on the IC of terms, which may lead to higher time consumption. VSM ignored the structure information of the GO graph and the specificity of terms.

In summary, how to measure gene functional similarity reliably is still a challenging task. In this paper, we propose a novel GCN framework with KGE techniques to measure gene functional similarity. The main contributions of GOGCN are summarized as follows:

- We put forward a novel gene functional similarity measurement framework utilizing GCN to model

the GO graph. The framework can effectively capture the structure information of the GO graph.

- In system biology, GOGCN is the first work leveraging GCN to model the GO graph for learning the vector representations of terms so as to measure the gene functional similarity.
- GOGCN makes a great improvement in the measurement of semantic similarity between terms.
- Extensive experimental results on various datasets show the effectiveness and reliability of GOGCN by comparing it with the state-of-the-art baselines.

2 RELATED WORK

In this section, we briefly introduce GCN and its applications in KGE and bioinformatics. To the best of our knowledge, so far no work has applied GCN into the measurement of gene functional similarity.

Kipf [21] first proposed GCN to model graph-structure data, whose core idea is to implement convolution operation on a graph based on neighborhood structure for learning the vector representations of nodes, which successfully generalizes Convolutional Neural networks (CNNs) to non-euclidean data. GCN and some of its variants have been achieved significant performances in node classification [21], [22] and link prediction [23]. However, most GCN methods focus on representation learning of nodes in undirected graphs such as Cora, Citeseer and Pubmed [24]. For directed acyclic graphs with specific relationships between two nodes, such as GO graph, applying GCN directly into representation learning of nodes will result in bad effectiveness of learned representations since the relationships and their direction are ignored. To address this issue, several researches [25], [26], [27] on GCN for modeling multi-relational graphs attempting representation learning of both nodes and relationships are shown remarkable performance.

KGE is also called knowledge representation learning, whose basic idea is to learn the representations of entities and relations for predicting the missing links of knowledge graphs. R-GCNs [23] is the first to apply the GCN framework into KGE, followed by other approaches, such as VR-GCN [28], COMPGCN [29]. KGE models based on GCN consist of an encoder: GCN layers producing latent feature representations of entities and relations, and a decoder: a link prediction model utilizing these representations to predict labeled edge and update these representations. Link prediction models can be classified into three categories: translation models like TransE [30], tensor decomposition models like DistMult [31], and neural network models like ConvE [32].

Very recently, researchers have developed numerous GCN-based approaches to tackle various bioinformatics tasks [33]. For example, Huang [34] used GCNs to predict associations between miRNA and drug resistance. Li [35] proposed a GCN-based method with neural inductive matrix completion called NIMCGCN to cope with the problem of miRNA-disease association prediction. Long [33] developed a framework named GCNMDA for the human microbe-drug association prediction. Yu employed the traditional model to predict GO annotations for maize proteins [36] and isoform[37]. In this paper, owing to these successful applications of GCN on bioinformatics, we make use of

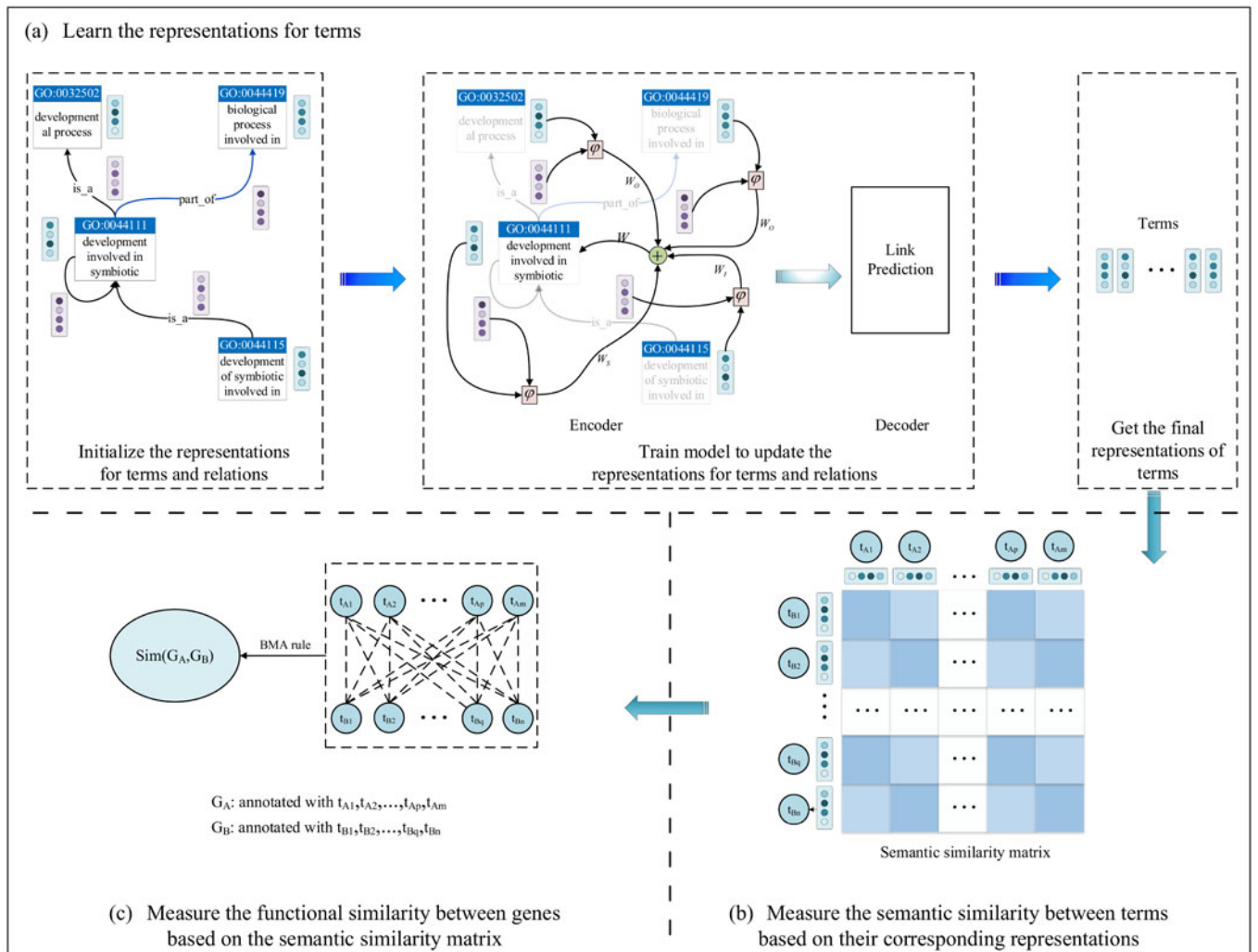


Fig. 1. The framework of GOGCN. (a) GOGCN devises a GCN-based encoder and a decoder driven by link prediction to learn the representations for all terms and relations. (b) The representations of terms are then used to measure the semantic similarity between terms. (c) the functional similarity between genes can be calculated by the BMA rule.

GCN with simultaneously learning term and relationship representations to measure gene functional similarity based on the GO graph, which can fully capture the structural information of GO graph.

3 METHODS

The framework of GOGCN is displayed in Fig. 1. First, to learn the representations for terms, GOGCN designs a KGE model which consists of an encoder and a decoder. The encoder is made up of GCN layers that update the initialized representations of terms and relations according to the semantic interaction between terms. The decoder is driven by the link prediction task which takes the output of the encoder as input so as to perform secondary learning for the representations of terms and relations. Then, GOGCN calculates the semantic similarity between terms by exploiting their corresponding representations. Finally, the functional similarity between genes is measured by the BMA rule.

3.1 Learn the Representations for Terms

In this paper, we express the GO graph as $\mathcal{G} = (\mathcal{V}, \mathcal{R}, \mathcal{E}, \mathcal{X}, \mathcal{Z})$. \mathcal{V} denotes the set of entities (*i.e.* terms). \mathcal{R}

represents the set of relations. \mathcal{E} represents the edge set. $\mathcal{X} \in \mathbb{R}^{|\mathcal{V}| \times d_0}$ denotes the initialized entity representations. $\mathcal{Z} \in \mathbb{R}^{|\mathcal{R}| \times d_0}$ represents the initialized relation representations. Inspired by COMPGCN, GOGCN takes the original direction and inverse direction of edges into account, and extends \mathcal{R} , *i.e.*, $\mathcal{R}' = \mathcal{R} \cup \mathcal{R}_{inv} \cup \{\mathcal{T}\}$, where \mathcal{T} denotes the self loop. For a triple (h, r, t) , its reverse triple can be inferred as (t, r^{-1}, h) . Then the set of the inverse relations \mathcal{R}_{inv} can be represented as $\mathcal{R}_{inv} = \{r^{-1} | r \in \mathcal{R}\}$.

In the encoder, for fully capturing the structural information of the GO graph, GOGCN adds a propagation weight between GCN layers based on COMPGCN[29]. Its propagation formula is defined as:

$$h_i^{k+1} = f \left(W^k \sum_{(j,r) \in \mathcal{N}(i)} W_{\lambda(r)}^k Aggr(h_j^k, h_r^k) \right) \quad (1)$$

where $\mathcal{N}(i)$ is the first-order neighbors set of i , j and r represent the entity and relation connected to i , h_j^k and h_r^k indicate the representations of entity j and relation r after the $(k - 1)$ -th GCN layer respectively, h_j^0 and h_r^0 are j -th and r -th entry of the initialized \mathcal{X} and \mathcal{Z} respectively, $W^k \in$

$\mathbb{R}^{d_{k+1} \times d_{k+1}}$ represents the propagation weight of the k -th GCN layers, f is the activation function like $ReLU$, $W_{\lambda(r)}^k \in \mathbb{R}^{d_{k+1} \times d_k}$ is the convolutional filter related to the direction of relations. $Aggr$ denotes the aggregation operation of neighbor terms and relations. We choose the circular correlation operation in HoLE[38] as the aggregation operation, defined as:

$$Aggr(h_j, h_r) = h_j \star h_r \quad (2)$$

Since the direction of relationships between two terms may indicate a specific semantic meaning, we adopt the direction-specific weights in COMPGCN to define $W_{\lambda(r)}$, expressed as:

$$W_{\lambda(r)} = \begin{cases} W_O, & r \in \mathcal{R} \\ W_I, & r \in \mathcal{R}_{inv} \\ W_S, & r = T(\text{self} - \text{loop}) \end{cases} \quad (3)$$

Further, the representations of relations are updated as follows:

$$h_r^{k+1} = W_{rel}^k h_r^k \quad (4)$$

where $W_{rel}^k \in \mathbb{R}^{d_{k+1} \times d_k}$ indicates the relation update matrix of k -th GCN layer and h_r^{k+1} represents the updated representations of relation r after the k -th GCN layer.

In the decoder, we first take the GO graph as a knowledge graph and store its information in the form of triples. The triple, also called fact, consists of three parts: head entity, relation, and tail entity. Then, the corresponding representations of these triples are made use of to train the link prediction model. The model gives higher scores for true triples and lower scores for corrupt triples and in turn updates these representations. Finally, the model is iteratively trained to reach the convergence state where the scores of true and corrupt triples can be measured as accurately as possible. The 1-n scoring strategy is applied in the training phase, that is, for each true triple, the corresponding corrupt triples dataset is constructed by replacing its head entity and tail entity with other entities. In this study, we choose the model developed by ConvE[32] as the scoring function defined as:

$$\varphi(e_s, e_o) = f(\text{vec}(f([\bar{e}_s; \bar{r}_r] * w))W)e_o \quad (5)$$

where $e_s \in \mathbb{R}^k$ and $e_o \in \mathbb{R}^k$ represent the vector representation of head entity s and tail entity o updated by (1) respectively, $r_r \in \mathbb{R}^k$ denotes the vector representation of relation r updated by (4), $\bar{e}_s \in \mathbb{R}^{k_w k_h}$ and $\bar{r}_r \in \mathbb{R}^{k_w k_h}$ indicate a 2D reshaping of e_s and e_o in which $k = k_w k_h$, $*$ denotes 2D convolutional operation and w is a convolutional filter, $\text{vec}(\cdot)$ represents vectorization that transforms a tensor into a vector. W is a linear transformation matrix which transforms the vector obtained by convolutional neural networks into a k -dimension space.

GOGCN chooses the binary cross entropy (BCE) loss as loss function, given as:

$$\mathcal{L} = -\frac{1}{N} \sum_i ((t_i \cdot \log(p_i)) + (1 - t_i) \cdot \log(1 - p_i)) \quad (6)$$

where p_i represents the score of i -th triple. t_i denotes the label of i -th triple with the value is 1 for true triples and 0 for corrupt triples.

After training, the representations of entities and relations are updated again and the semantic similarity between entities can be estimated by exploiting the representations of entities. The above processes demonstrate GOGCN sufficiently promotes the semantic interaction between terms, thereby capturing the structural information of the GO graph.

3.2 Measure Semantic Similarity Between Terms

The representations of entities (*i.e.* terms), can be obtained through Section 3.1. Since the most common strategy for measuring the similarity between vectors is cosine similarity, GOGCN also leverages it to measure the semantic similarity between two terms. The formula is expressed as:

$$S_T(t_1, t_2) = \frac{e_{t_1} \cdot e_{t_2}}{|e_{t_1}| |e_{t_2}|} \quad (7)$$

where e_{t_1} and e_{t_2} represent the vector representations of the term t_1 and t_2 .

3.3 Measure Functional Similarity Between Genes

In this section, we measure the functional similarity between genes based on the BMA rule. Suppose there are two genes G_A and G_B annotated with two term sets $T_A = \{t_{A1}, t_{A2}, \dots, t_{Am}\}$ and $T_B = \{t_{B1}, t_{B2}, \dots, t_{Bn}\}$ respectively. GOGCN first estimates the similarity between terms and genes. Taking term t_{A1} and gene G_B as examples, the calculation formula is:

$$S_{TG}(t_{A1}, T_B) = \max_{1 \leq i \leq n} (S_T(t_{A1}, t_{Bi})) \quad (8)$$

Subsequently, the functional similarity between gene G_A and gene G_B is defined as:

$$S_G(G_A, G_B) = \frac{\sum_{1 \leq i \leq m} S_{TG}(t_{Ai}, T_B) + \sum_{1 \leq j \leq n} S_{TG}(t_{Bj}, T_A)}{m + n} \quad (9)$$

4 MATERIALS

4.1 GO and GO Annotations

In this research, the GO graph is reconstituted as triples for training the KGE model to obtain the final vector representations of terms and relations. GO files can be downloaded from the Gene Ontology database (<http://geneontology.org/docs/download-ontology/>). The version exploited by GOGCN is dated September 2020 which contains 28,922 BP terms, 4,193 CC terms, and 11,157 MF terms. The number of terms, relations, and triples in the GO graph are displayed in Table 1. Given that the measurement of gene functional similarity only needs to consider 'is a' and 'part of' relations, GOGCN also follows this rule.

GO annotations are composed of the annotation information of genes, which is essential for estimating the functional similarity of genes. Each GO annotation is assigned together with an evidence code (EC) that refers to the process used to assign the specific GO term to a given gene [39]. In this study, we refer to the annotation set that contains annotations created by the Inferred from Electronic Annotation (IEA) as IEA+, and the one that does not contain IEA annotations is called IEA-. We download the GO annotation files

TABLE 1
The Number of Terms, Relations, and Triples in the GO Graph

	Entities (Terms)	Total Entities	Relations	Triples
BP	28,922	44,272	2	81,952
CC	4,193			
MF	11,157			

for *Homo sapiens* (dated October 2020) and *Saccharomyces cerevisiae* (dated October 2020) from the Gene Ontology database (<http://geneontology.org/docs/download-go-annotations/>).

4.2 Protein-Protein Interaction Dataset

We rebuild the *Homo sapiens* datasets and *Saccharomyces cerevisiae* datasets mainly collected from WIS [17] and Zhang [19] respectively. The new dataset contains 1,251 positive Protein-Protein Interactions (PPIs) for *H.sapiens* and 1,113 positive PPIs for *S.cerevisiae* respectively. In the meantime, the corresponding negative PPIs of *H.sapiens* and *S.cerevisiae* with the same number of positive ones are created randomly which are absent from the dataset of all possible positive PPIs.

4.3 Gene Expression Dataset for *Saccharomyces Cerevisiae*

Since the genes involved in the same biological process or function have higher expression values, the correlation between gene expression data and functional similarity of genes can be used as an effective evaluation criterion. In this paper, GOGCN makes use of the gene expression dataset downloaded from Jain [40] which contains 6,000 different gene pairs and their corresponding expression value, including 2,000 pairs for BP, CC, and MF ontology respectively.

4.4 CESSM Dataset for Correlation With Pfam Similarity

The Collaborative Evaluation of GO-based Semantic Similarity Measures (CESSM) is an online tool for the evaluation of GO-based SSMs against sequence, and protein family (Pfam) similarities[41]. However, its dataset is an old version that was dated August 2008. The GO graph and GO annotation file have changed simultaneously. To this end, we update the dataset by removing the proteins not existing in the GO annotation file, then obtain 10,774 pairs of proteins from various species. We then employ these protein pairs to find correlation against the Pfam (download from the UniPort database) similarity because terms in the GO graph have the manifest ability to distinguish the functional aspect of gene [42].

4.5 Biological Pathways Dataset

In this study, we collect two kinds of biological pathways for two experiments: Functional classification of genes in a biological pathway and Set-discriminating power of different KEGG pathways. First, genes in the same reaction stage (i.e., have the same EC number) in a biological pathway tend to exhibit similar functions[11]. Hence, to explore the

TABLE 2
The Main Hyperparameter Setting of GOGCN

Hyperparameters	Value
optimizer	Adam
learning rate	0.001
GCN layers	1
initial embedding size	100
GCN embedding size	200
GCN dropout	0.1
epoch	50
batch size	128

functional classification of genes, we extract a yeast pathway that contains 10 genes from the SGD database to estimate the classification performance based on MF ontology. Second, because a biological pathway shows the accomplishment processes of a specific biological process in a cell, proteins involved in a pathway are more likely to interact among themselves than the proteins belonging to different pathways [42]. To this end, we collect five yeast KEGG pathways which all contain the number of genes between 11 to 14 for measuring the discriminating power based on CC ontology.

5 RESULTS

In this section, we introduce the experimental setup for learning the representations of terms and discuss the experimental results in detail. Before discussing the experimental results, we briefly introduce baselines and evaluation metrics. Subsequently, we evaluate GOGCN in several convincing and pervasive experiments: Protein-Protein Interaction of *S. cerevisiae* and *H. sapiens*, Pearson's correlation coefficient analysis based on gene expression data, Correlation with Pfam, Functional classification of genes in a biological pathway, and Set-discriminating power of KEGG Pathways. The details are described as follows.

5.1 Experimental Setup

The main hyperparameter setting of GOGCN when learning the representations of terms are listed in Table 2. We adopt a single layer GCN to train the model for learning the representations of terms and relations. In the encoder, the input embedding size (i.e., initial embedding size) and output embedding size (i.e., GCN embedding size) are set as 100 and 200 respectively. And the dropout through GCN layers is equal to 0.1. In the decoder, a dropout with a rate of 0.3 is assigned to feature maps and the hidden layer. During training the model, we use Adam optimizer with a learning rate of 0.001 to update model parameters. All model parameters are initialized by Xavier initialization [43]. The experimental code is implemented by Pytorch.

5.2 Baselines

In this section, we briefly introduce the baseline methods as follows.

- Resnik [8]: the most representative pair-wise method that measures the IC of terms based on specific corpora.

- simUI [15]: a classic group-wise method that uses the ratio of the number of the intersection and union set annotating two genes to measure gene functional similarity.
- simGIC [44]: an extension of simUI where the ratio of the number of the intersection and union set is converted to the overlapping IC value of the intersection and union set.
- Wang [11]: an innovative pair-wise method that assigns a weight value to the relationship connecting two terms.
- SORA [7]: an innovative group-wise method that fully incorporates the structure of GO into measuring gene functional similarity.
- VSM: a basic vector-based method that uses a vector to represent the annotating term set for measuring gene functional similarity based on vectors.

5.3 Evaluation Metrics

In this study, there are mainly four group experiments which are Protein-Protein Interaction of *S. cerevisiae* and *H. sapiens*, Pearson's correlation coefficient analysis based on gene expression data, Correlation with Pfam similarity. Four metrics are employed to evaluate the performance of the gene functional calculation methods.

The first one is area under curve (AUC) values used in the Protein-Protein Interaction experiment. In this experiment, we construct the dataset containing positive Protein-Protein Interactions (PPIs) for *H.sapiens* and *S.cerevisiae* respectively. In the meantime, the corresponding negative PPIs of *H.sapiens* and *S.cerevisiae* are created randomly (see Section 4.2). Lastly, we calculate all the functional similarity for both positive and negative PPIs. Functional similarity values of all gene pairs are varied from 0 to 1. To draw the receiver operating characteristic curves (ROC) plots, gene pairs of which functional similarity values are greater than the specific threshold are treated as positive samples, while those gene pairs which are smaller than the specific threshold are regarded as negative samples. Thereafter, four indicators which are the true positive, true negative, false positive, and false negative values are obtained. Further, true positive rate (TPR) and false positive rate (FPR) can be computed. We varied the threshold from 0 to 1 and ROC curves can also be plotted based on TPR and FPR values [45]. According to ROC curves, we can get the AUC values. In this study, we compare the performance of the GOGCN and other state-of-the-art methods based on the AUC values. Generally, the higher the AUC value is, the better the performance of the functional calculation method is.

The second is Pearson's correlation coefficient analysis based on gene expression data. For this metric, we first measure the expression similarity for gene pairs. Then the functional similarity for these gene pairs are also computed with functional similarity calculation methods. Lastly, we measure the Pearson correlation coefficient based on the expression similarity and functional similarity for gene pairs. The higher the Pearson correlation coefficient value is, the better the performance of the functional calculation method is.

The third one is the Correlation with Pfam similarity. Similar to Pearson's correlation coefficient analysis, we first

need to compute the Pfam family similarity between gene pairs [41]. Then, the functional similarity for these gene pairs are also measured by functional similarity calculation methods. Lastly, the correlation of functional similarity with Pfam similarity can be measured by Pearson's correlation coefficient. The higher the Pearson correlation coefficient value is, the better the performance of the functional calculation method is.

The last one is the DP (discriminating power), which we take as the evaluation metric [42]. Specifically, let $P = \{P_1, P_2, \dots, P_n\}$ and S_G as the KEGG pathways set and an approach for measuring gene functional similarity respectively. For a KEGG pathway P_k , $\{g_{k_1}, g_{k_2}, \dots, g_{k_p}\}$ denotes the gene set involved in this pathway. For calculating the DP value of P_k , the *intra-set average similarity* is computed firstly, given as:

$$Intra_sim(P_k) = \frac{\sum_{i=1}^{k_p} \sum_{j=1}^{k_p} S_G(g_{k_i}, g_{k_j})}{k_p^2} \quad (10)$$

Next, let $\{g_{l_1}, g_{l_2}, \dots, g_{l_q}\}$ represent the gene set contained by KEGG pathway P_l , the *inter-set average similarity* between P_k and P_l is then defined as:

$$Inter_sim(P_k, P_l) = \frac{\sum_{i=1}^{k_p} \sum_{j=1}^{l_q} S_G(g_{k_i}, g_{l_j})}{k_p \times l_q} \quad (11)$$

Finally, For the KEGG pathways set P , the DP value of pathway P_k can be calculated as:

$$DP(P_k) = \frac{(n-1) \times Intra_sim(P_k)}{\sum_{i=1, i \neq k}^n Inter_sim(P_k, P_i)} \quad (12)$$

From the above calculation process, the DP value of a KEGG pathway quantifies the ability of an approach to discriminate the genes involved in the current pathway from the genes contained by other pathways. Thus, the higher the DP value, the better the performance of the corresponding approach.

5.4 Protein-Protein Interaction of *S. Cerevisiae* and *H. Sapiens*

In this section, we conduct the PPI experiment to evaluate the effectiveness of GOGCN. Here, we choose four mainstream group-wise approaches (simUI[15], simGIC[44], and SORA[7]) and two classical pair-wise approaches (Resnik [8], and Wang[11]) as the baselines to evaluate the experimental results. The AUC values of these approaches on *S. cerevisiae* datasets are displayed in Table 3. GOGCN gets the best result on BP_IEA+, BP_IEA-, CC_IEA+, MF_IEA+, and MF_IEA- except for CC_IEA-. For example, the AUC value of GOGCN is 0.8457 on the CC_IEA+ experiment. From Table 3, GOGCN shows great advantages on BP and MF ontologies. For example, the AUC value of GOGCN on the BP_IEA+ experiment is 0.8968 followed by 0.8784 calculated by simGIC. Further, even if the AUC value of GOGCN on the CC_IEA- experiment is inferior to simGIC and Wang, the gap between them is not significant.

Table 4 depicts the AUC values of functional similarity approaches on *H. sapiens* datasets. GOGCN gets the best results on each experiment compared with the other six baselines. From Table 4, although Wang and simGIC show

TABLE 3

The AUC Values of Functional Similarity Approaches for *S. cerevisiae* Datasets

Approaches	BP_IEA+	BP_IEA-	CC_IEA+	CC_IEA-	MF_IEA+	MF_IEA-
simUI[15]	0.8515	0.8376	0.8002	0.7809	0.7600	0.7711
simGIC[44]	0.8784	0.8680	0.8262	0.8145	0.7843	0.7940
VSM	0.8545	0.8394	0.8010	0.7824	0.7615	0.7713
SORA[7]	0.8762	0.8653	0.8140	0.8031	0.7899	0.7985
Resnik[8]	0.7926	0.7977	0.7852	0.7762	0.7506	0.7611
Wang[11]	0.8718	0.8676	0.8416	0.8138	0.7699	0.8000
GOGCN	0.8968	0.8837	0.8457	0.8090	0.8132	0.8312

The best results are in bold.

extraordinary performance, GOGCN is still in the lead. For instance, in comparison with Wang and simGIC, GOGCN increases by 1.30% and 6.07% respectively on CC_IEA-experiment. Besides, GOGCN is 1.89% and 5.88% higher than Wang and simGIC on the MF_IEA+ experiment respectively.

In brief, GOGCN achieves the best performance on PPI experiments for both *S. cerevisiae* and *H. sapiens* datasets. In addition, we also summed up two points from the PPI experiment. On one hand, regardless of *S. cerevisiae* datasets or *H. sapiens* datasets, the experimental results show that the performances on BP, CC, and MF experiments decrease in turn. On the other hand, although GOGCN has achieved excellent results, the performance of group-wise approaches is better than pair-wise approaches in some circumstances, which is consistent with SORA[7].

5.5 Pearson's Correlation Coefficient Analysis Based on Gene Expression Data

We calculate the Pearson's correlation coefficient between gene functional similarity and gene expression data. The results are depicted in Table 5. In general, the higher the Pearson's correlation coefficient, the better the approach [46].

As Table 5 shows, most approaches demonstrate a higher correlation on CC ontology. We can find that GOGCN achieves the highest correlation on BP_IEA+, CC_IEA+, CC_IEA-, and MF_IEA- experiments. Additionally, the achievements of GOGCN far surpass other baseline approaches on the four experiments. A convincing case in point is that GOGCN scores 0.4500 on CC_IEA- experiment which is 0.1524 higher than Resnik. Furthermore, the results show the performances of GOGCN and simGIC are neck and neck on the BP_IEA- experiment. Apart from the aforementioned, for the MF_IEA+ experiment, the achievement of GOGCN is only inferior to SORA and simUI. In summary, GOGCN successfully demonstrates the highest correlation with the gene expression dataset overall, which reflects its effectiveness.

5.6 Correlation With Pfam Similarity

A Pfam generally denotes the evolutionary process of proteins, that is, proteins sharing with same Pfams show similar functions. Following CESSM[41], the Pfam similarity of two proteins is calculated by the Jaccard index which is defined as the ratio of the number of protein families they share to the total number of protein families they belong to.

Authorized licensed use limited to: Zhengzhou University. Downloaded on September 14, 2023 at 13:26:12 UTC from IEEE Xplore. Restrictions apply.

TABLE 4

The AUC Values of Functional Similarity Approaches for *H. sapiens* Datasets

Approaches	BP_IEA+	BP_IEA-	CC_IEA+	CC_IEA-	MF_IEA+	MF_IEA-
simUI[15]	0.8993	0.8922	0.7614	0.7605	0.6520	0.6551
simGIC[44]	0.9227	0.9154	0.7955	0.7937	0.7399	0.7579
VSM	0.9081	0.8986	0.7648	0.7641	0.6428	0.6519
SORA[7]	0.9147	0.9096	0.7676	0.7722	0.7014	0.7040
Resnik[8]	0.8550	0.8604	0.7488	0.7536	0.7435	0.7615
Wang[11]	0.9247	0.9195	0.8344	0.8311	0.7689	0.7545
GOGCN	0.9323	0.9228	0.8378	0.8419	0.7834	0.7653

The best results are in bold *sapiens* datasets.

TABLE 5

The Pearson's Correlation Coefficient Between the Results of Functional Similarity Approaches and Gene Expression Data

Approaches	BP_IEA+	BP_IEA-	CC_IEA+	CC_IEA-	MF_IEA+	MF_IEA-
simUI[15]	0.3449	0.3542	0.3872	0.3465	0.3575	0.3069
simGIC[44]	0.3892	0.3917	0.3583	0.3450	0.3259	0.3208
VSM	0.2898	0.3017	0.4077	0.3611	0.3280	0.2861
SORA[7]	0.3128	0.3535	0.4017	0.3808	0.3622	0.3377
Resnik[8]	0.3111	0.2690	0.2922	0.2976	0.3023	0.3127
Wang[11]	0.3086	0.2939	0.4221	0.4322	0.3048	0.3084
GOGCN	0.4051	0.3864	0.4329	0.4500	0.3303	0.3529

The best results are in bold.

Subsequently, the correlation of functional similarity with Pfam can be measured by Pearson's correlation coefficient.

As Table 6 shows, we divide the protein pairs into 1,000, 2,000, and 5,000 groups and then calculate the Pearson's correlation coefficient with their corresponding Pfam similarity. Overall performance of GOGCN, simGIC, and SORA are better than the rest approaches. Particularly, The correlation of them on BP (1,000 groups), CC (1,000 groups), and MF (1,000 groups) are all more than 0.8. Meanwhile, it may be noted that the correlations decrease in turn when dividing protein pairs into 1,000, 2,000, and 5000 groups, which shows that Pearson's correlation coefficient is negatively correlated with the number of groups.

Further, GOGCN performs best in eight metrics. By way of illustration, GOGCN has an overwhelming advantage in CC (2,000 groups) where GOGCN scores 0.842 and is 0.300 more than VSM which gets the second best result. In brief, GOGCN shows the highest correlation with Pfam and therefore is more superior compared with the state-of-the-art baselines.

5.7 Experimental Analysis of Biological Pathways

5.7.1 Functional Classification of Genes in a Biological Pathway

As is shown in Table 7, we take the 'L-tyrosine degradation III' pathway as an example. There are 10 genes involved in three biological processes, corresponding to three EC numbers. The functional similarity between these genes are calculated by GOGCN and three baseline methods which are Resnik [8], simGIC[44], and simUI[15].

From Fig. 2, GOGCN, Resnik, simUI, and simGIC all divide the 10 genes into their corresponding category. From the perspective of GOGCN, the functional similarity

TABLE 6
The Pearson's Correlation Coefficient Between the Results of Functional Similarity Approaches and Pfam

Approaches	BP			CC			MF		
	1000	2000	5000	1000	2000	5000	1000	2000	5000
simUI[15]	0.851	0.753	0.596	0.877	0.807	0.669	0.881	0.808	0.665
simGIC[44]	0.849	0.758	0.606	0.854	0.780	0.646	0.893	0.824	0.683
VSM	0.860	0.762	0.599	0.884	0.812	0.672	0.846	0.765	0.618
SORA[7]	0.876	0.788	0.631	0.875	0.800	0.658	0.863	0.786	0.641
Resnik[8]	0.695	0.576	0.414	0.725	0.631	0.474	0.727	0.623	0.467
Wang[11]	0.836	0.733	0.564	0.861	0.774	0.610	0.859	0.782	0.626
GOGCN	0.887	0.794	0.641	0.898	0.842	0.706	0.896	0.829	0.682

The best results are in bold.

between genes involved in the same biological process is much higher than the functional similarity between genes belonging to different EC numbers. Besides, the lowest similarity value between genes belonging to the same EC numbers is 0.63. What's more, it is reasonable to a large extent that the functional similarity between genes involved in different biological processes calculated by GOGCN is in the range of 0.2 to 0.4. For method Resnik, although it can also classify these genes into their corresponding categories, there are some problems with the functional similarity in some cases. For example, the functional similarity between gene 'PDC1' and 'PDC5' is 0.79 instead of 1.00 computed by Resnik, though, gene 'PDC1' and 'PDC5' are jointly annotated with GO term 'GO:0004737', 'GO:0047433', and 'GO:0016831'. Meanwhile, the functional similarity between gene 'ADH1' and gene 'ADH2', 'ADH3', 'ADH4', and 'ADH5' are all only 0.43, but they jointly participate in the same biological process. For simUI, there are some unsatisfactory gene functional similarities. As a case in point, the functional similarity between gene 'ADH1' and gene 'ADH2', 'ADH3', 'ADH4', and 'ADH5' is 0.54 which is lower than 0.63 calculated by GOGCN. From the results of simGIC, the functional similarity between genes involved in different biological processes are all 0, which is irrational to some extent. In a nutshell, GOGCN achieves the best results in the experiment of Functional classification of genes in a biological pathway.

5.7.2 Set-Discriminating Power of KEGG Pathways

The information of the selected KEGG pathways is listed in Table 8, and Table 9 depicts the discriminating power of

TABLE 7
Function of Genes in 'L-Tyrosine Degradation III' Pathway

Class ID	EC number	Gene Name
1	1.1.1.1	ADH1
	1.1.1.1	ADH2
	1.1.1.1	ADH3
	1.1.1.1	ADH4
	1.1.1.1	ADH5
2	2.6.1-	ARO8
	2.6.1-	ARO9
3	4.1.1.80	PDC1
	4.1.1.80	PDC5
	4.1.1.80	PDC6

GOGCN and three comparison approaches (Wang[11], VSM, and Resnik[8]) based on CC ontology.

Overall, the DP value of GOGCN is higher than other approaches for 4 of 5 KEGG pathways, thereby showing the better discriminating power of GOGCN. Further, although the discriminating power of GOGCN for the 'sce00514' pathway is inferior to Resnik, GOGCN yields an excellent performance than the rest two approaches.

6 ABLATION STUDY

We conduct several PPI experiments on Homo sapiens datasets to investigate the effectiveness of the components of our model. The impact of different types of aggregation operation, the use of relationships, and the encoder and decoder on model performance are explored in the following sections.

6.1 Ablation Study on Different Types of Aggregation Operation

To testify the effectiveness of the circular correlation operation in aggregating neighbor terms and relations, we introduce two other types of aggregation operation to compare with the circular correlation operation, which are 'add' and 'mult' denoting add and multiply respectively, given as:

$$Aggr(h_j, h_r)_{add} = h_j + h_r \quad (13)$$

$$Aggr(h_j, h_r)_{mult} = h_j * h_r \quad (14)$$

The circular correlation operation (corr) in HoLE [38] are given in (2).

As is shown in Fig. 3, the overall results indicate that adopting the circular correlation operation can achieve the best performance on 5 sub-tasks except for MF_IEA-, while using the other simple aggregation operations leads to bad performance. Specifically, although adopting the 'add' operation gets the best result on MF_IEA-, employing the circular correlation operation outperform it on the other 5 sub-tasks, which shows adopting more complex aggregation operations can benefit our model.

6.2 Ablation Study on the Use of Relationships

We train the encoder-decoder model to learn meaningful representations of both terms and relationships, then use the learned term representations to measure gene functional similarity. Note that although we only use the learned term representations to measure gene functional similarity after

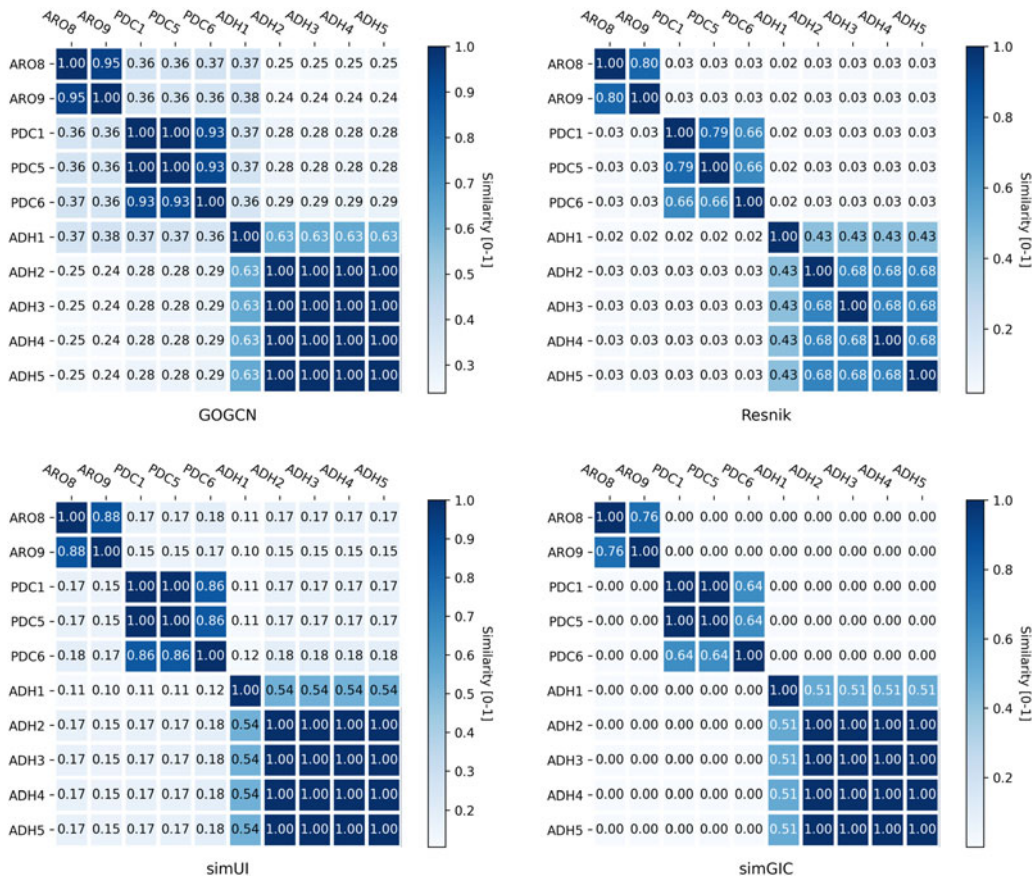


Fig. 2. Functional classification of genes in 'L-tyrosine degradation III' pathway based on MF ontology using method GOGCN, Resnik, simUI, and simGIC.

training our model, the relation representations also contribute to the learning process of term representations. To explore the validity of incorporating relationship 'is a' and 'part of' into our model, we introduce two variants of GOGCN as follows: 1) noRelation: remove representation learning of relationships and only use the information of neighbor terms to update the representation of the central term. 2) oneRelation: treat relationship 'is a' and 'part of' as equal and obtain only one relationship in our model.

As Fig. 4 shows, oneRelation achieves better results than noRelation on 4 out of 6 sub-tasks, indicating incorporating representation learning of relationships into our model can benefit to representation learning of terms. Moreover, GOGCN outperforms noRelation on all sub-tasks and performs better than oneRelation on 5 sub-tasks, which demonstrates treating 'is a' and 'part of' differently can learn more

representative terms representations and further improve the performance of the model.

6.3 Ablation Study on the Encoder and Decoder

Considering different neighbor terms and relationships may contribute to the central term differently, we introduce an attention-based model to compare with GOGCN. In the attention model, the propagation formula of representation learning of terms is defined as:

$$h_i^{k+1} = f \left(W^k \sum_{(j,r) \in N(i)} W_{\lambda(r)}^k \alpha_{ijr}^k \text{Aggr}(h_j^k, h_r^k) \right) \quad (15)$$

where α_{ijr}^k is calculated as:

$$\alpha_{ijr}^k = \frac{\exp(\text{Aggr}(h_j^k, h_r^k))}{\sum_{(m,n) \in N(i)} \exp(\text{Aggr}(h_m^k, h_n^k))} \quad (16)$$

TABLE 8
The Number of Genes in Five Yeast KEGG Pathways

Pathway ID	Pathway Name	Number of Genes
sce00053	Ascorbate and aldarate metabolism	11
sce00290	Valine, leucine and isoleucine biosynthesis	12
sce00350	Tyrosine metabolism	13
sce00514	Other types of O-glycan biosynthesis	14
sce00790	Folate biosynthesis	11

TABLE 9
The Discriminating Power of Different Approaches for the Selected Five KEGG Pathways Based on CC Ontology

Approaches	sce00053	sce00290	sce00350	sce00514	sce00790
VSM	1.0439	1.2285	1.1703	1.5872	1.1457
Resnik[8]	0.9087	1.2750	0.9901	2.3778	1.1812
Wang[11]	1.0003	1.2292	1.1840	1.6300	1.1705
GOGCN	1.0589	1.2886	1.2192	1.7684	1.1877

The best results are in bold.

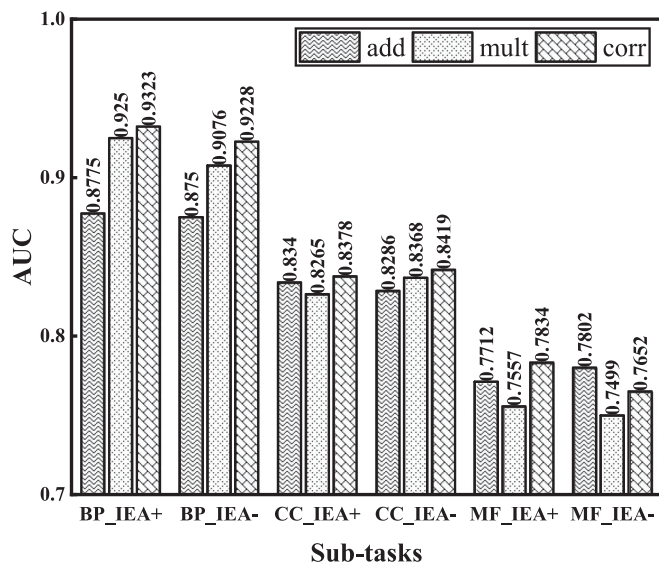


Fig. 3. Ablation study on different types of aggregation operation.

Additionally, we first employ the encoder to fuse the information of neighbor terms and relationships, then use the fused term and relation representations as the input of the decoder model to perform a secondary update. To validate the reasonability of our encoder model, we only use the decoder model ConvE to train the model for learning term and relation representations.

As illustrated in Fig. 5, GOGCN outperforms the attention-based model on all sub-tasks, manifesting that treating the neighbor terms differently results in worse performance. The reason for this is that the representations of relationships may already provide underlying importance for each neighbor term when fusing the aggregated information of neighbor terms and relationships. In addition, GOGCN achieves better than ConvE on 5 sub-tasks. Especially, GOGCN significantly outperforms ConvE on BP_IEA+, BP_IEA-, CC_IEA+, and CC_IEA-, which strongly demonstrates the reasonability of our encoder model, and further indicates that the information of neighbor terms and relationships fused by our encoder model is valuable.

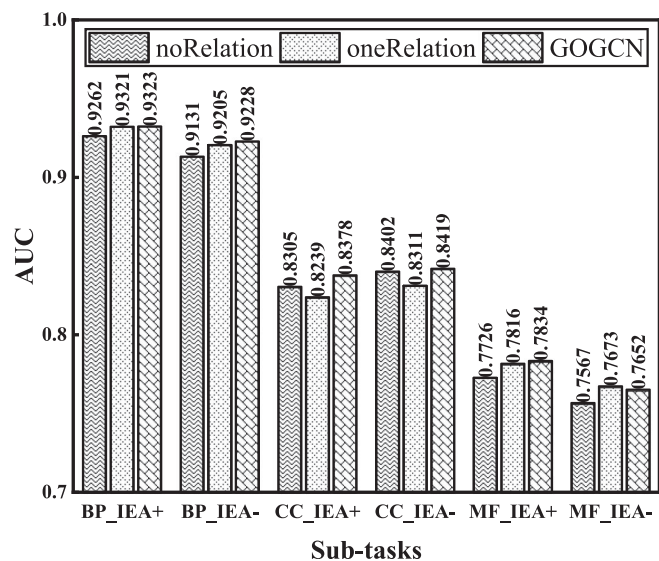


Fig. 4. Ablation study on the use of relationships.

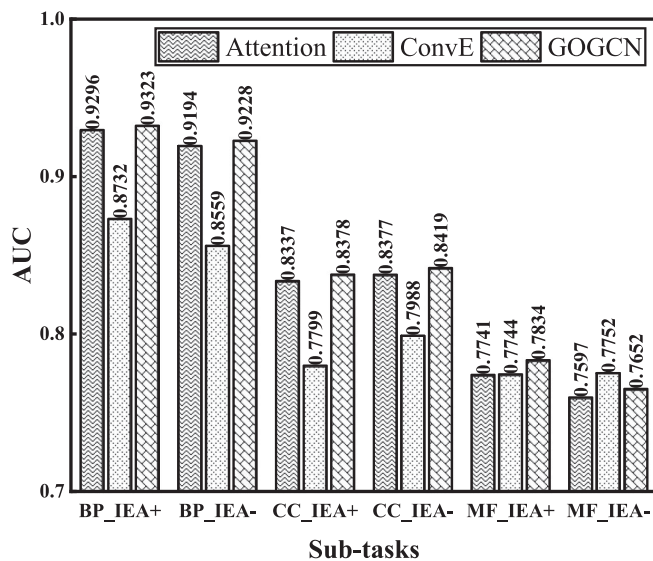


Fig. 5. Ablation study on the encoder and decoder.

7 DISCUSSION

In this manuscript, a novel approach called GOGCN for measuring gene functional similarity based on GCN is proposed, which is the first attempt to model the GO graph by GCN. The experimental results strongly validate the innovation and effectiveness of GOGCN. In the meantime, several questions deserved discussion in this section.

- 1) Why the performance of GOGCN is better than other approaches?

For one thing, in the encoder, GOGCN exploits GCN to aggregate the vector representations of neighbor terms and connected relations for updating the representation of central term where the weight parameter based on the direction of the relation is considered as the convolution kernel to implement the convolution operation on the GO graph. In this process, semantic interactions between terms are realized through relations, which largely capture the structural information of the GO graph. For another, in the decoder, GOGCN takes all the real triples existing in the GO graph as the training set and constructs negative samples by destroying the head terms or tail terms during the training process so as to perform secondary learning on the representations of terms and relations. Therefore, after the joint training of the encoder and decoder, GOGCN has fully modeled all the terms and the relation between them, which is pivotal for measuring the semantic similarity between terms.

- 2) Why GOGCN applies the technique of KGE based on GCN into the measurement of gene functional similarity?

GCNs have recently been shown to be quite successful in modeling graph-structured data [29]. Meanwhile, numerous researches indicated that employing GCN to model biological knowledge graphs is effective. What's more, the GO graph is a special knowledge graph where terms and relations can be seen as the entities and relations of knowledge

graphs respectively. Thereafter, the representations of terms can be learned for purpose of the measurement of semantic similarity between terms.

- 3) Why GOGCN learns the representations of relations in encoder but does not use the learned representations of relations to measure gene functional similarity?

For one thing, in order to converge as quickly as possible when training the model, it is indispensable for the decoder to learn the representations of relations in the encoder. Simultaneously, learning the representations of relations can facilitate semantic interaction between terms, which makes full use of the structural information of the GO graph to a certain extent. For another, given the framework of GCN, the measurement of semantic similarity between terms is merely based on the representations of terms.

8 CONCLUSION

In this study, we put forward a novel pair-wise approach utilizing GCN to model the GO graph for measuring gene functional similarity. Subsequently, we conduct four experiments so as to estimate the performance of GOGCN. In addition, the ablation study is conducted to explore the validity of the components of GOGCN. In comparison with mainstream approaches, GOGCN has the following innovations.

On the whole, GOGCN designs a GCN-based framework to learn the vector representations of the GO terms for measuring gene functional similarity. What's more, because GO terms interact with each other through relations, GOGCN can fully collect the structural information of the GO graph, which effectively captures the specificity of terms and relations.

In terms of details, GOGCN transforms the measurement of semantic similarity between terms into the similarity between their corresponding vector representations. On one hand, compared with approaches that need to measure the IC of terms, GOGCN can skip this sophisticated stage. On the other hand, it is convenient to measure the similarity between vectors, thereby there is no need to design a complicated semantic similarity strategy. As a consequence, GOGCN is advantageous in measuring semantic similarity between terms.

In conclusion, GOGCN is the first attempt that makes use of GCN to model the GO graph for measuring gene functional similarity. Aside from this, the corresponding experimental results demonstrate the effectiveness and innovation of GOGCN. In the future, we will further improve GOGCN in detail and apply GOGCN to some other biological applications.

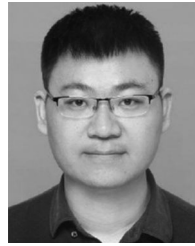
ACKNOWLEDGMENTS

Zhen Tian and Haichuan Fang contribute to this paper equally.

REFERENCES

- [1] J. Peng, H. Xue, Z. Wei, I. Tuncali, J. Hao, and X. Shang, "Integrating multi-network topology for gene function prediction using deep neural networks," *Brief. Bioinf.*, vol. 22, no. 2, pp. 2096–2105, 2021.
- [2] G. Yu, G. Fu, J. Wang, and Y. Zhao, "NewGOA: Predicting new GO annotations of proteins by bi-random walks on a hybrid graph," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 4, pp. 1390–1402, Jul./Aug. 2018.
- [3] G. Yu, K. Wang, G. Fu, M. Guo, and J. Wang, "NMFGO: Gene function prediction via nonnegative matrix factorization with gene ontology," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 17, no. 1, pp. 238–249, Jan./Feb. 2020.
- [4] Q. Zou, G. Lin, X. Jiang, X. Liu, and X. Zeng, "Sequence clustering in bioinformatics: An empirical study," *Brief. Bioinf.*, vol. 21, no. 1, pp. 1–10, 2020.
- [5] M. Zeng, F. Zhang, F.-X. Wu, Y. Li, J. Wang, and M. Li, "Protein-protein interaction site prediction through combining local and global features with deep neural networks," *Bioinformatics*, vol. 36, no. 4, pp. 1114–1120, 2020.
- [6] Y. Liu, X. Zeng, Z. He, and Q. Zou, "Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 4, pp. 905–915, Jul./Aug. 2017.
- [7] Z. Teng, M. Guo, X. Liu, Q. Dai, C. Wang, and P. Xuan, "Measuring gene functional similarity based on group-wise comparison of go terms," *Bioinformatics*, vol. 29, no. 11, pp. 1424–1432, 2013.
- [8] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *J. Artif. Intell. Res.*, vol. 11, pp. 95–130, 1999.
- [9] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proc. 10th Res. Comput. Linguistics Int. Conf.*, 1997, pp. 19–33.
- [10] D. Lin et al., "An information-theoretic definition of similarity," in *Proc. Int. Conf. Mach. Learn.*, 1998, pp. 296–304.
- [11] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, "A new method to measure the semantic similarity of go terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, 2007.
- [12] S. J. Bien, C. H. Park, H. J. Shim, W. Yang, J. Kim, and J. H. Kim, "Bi-directional semantic similarity for gene ontology to optimize biological and clinical analyses," *J. Amer. Med. Inform. Assoc.*, vol. 19, no. 5, pp. 765–774, 2012.
- [13] A. Pesaranghader, S. Matwin, M. Sokolova, and R. G. Beiko, "SimDEF: Definition-based semantic similarity measure of gene ontology terms for functional similarity analysis of genes," *Bioinformatics*, vol. 32, no. 9, pp. 1380–1387, 2016.
- [14] Y. Zhao, G. Fu, J. Wang, M. Guo, and G. Yu, "Gene function prediction based on gene ontology hierarchy preserving hashing," *Genomics*, vol. 111, no. 3, pp. 334–342, 2019.
- [15] S. Falcon and R. Gentleman, "Using gostats to test gene lists for go term association," *Bioinformatics*, vol. 23, no. 2, pp. 257–258, 2007.
- [16] G. Yu, G. Fu, J. Wang, and H. Zhu, "Predicting protein function via semantic integration of multiple networks," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 2, pp. 220–232, Mar./Apr. 2016.
- [17] Z. Tian, C. Wang, M. Guo, X. Liu, and Z. Teng, "An improved method for functional similarity analysis of genes based on gene ontology," *BMC Syst. Biol.*, vol. 10, no. 4, pp. 465–484, 2016.
- [18] S. Benabderrahmane, M. Smail-Tabbone, O. Poch, A. Napoli, and M.-D. Devignes, "Intelligo: A new vector-based semantic similarity measure including annotation origin," *BMC Bioinf.*, vol. 11, no. 1, pp. 1–16, 2010.
- [19] J. Zhang, K. Jia, J. Jia, and Y. Qian, "An improved approach to infer protein-protein interaction based on a hierarchical vector space model," *BMC Bioinf.*, vol. 19, no. 1, pp. 1–14, 2018.
- [20] G. Yu, Y. Zhao, C. Lu, and J. Wang, "HashGO: Hashing gene ontology for protein function prediction," *Comput. Biol. Chem.*, vol. 71, pp. 264–273, 2017.
- [21] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learn. Representations*, 2017.
- [22] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proc. 6th Int. Conf. Learn. Representations*, 2018.
- [23] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *Proc. Eur. Semantic Web Conf.*, 2018, pp. 593–607.
- [24] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, "Collective classification in network data," *AI Mag.*, vol. 29, no. 3, pp. 93–93, 2008.
- [25] D. Yu, Y. Yang, R. Zhang, and Y. Wu, "Knowledge embedding based graph convolutional network," in *Proc. Web Conf.*, 2021, pp. 1619–1628.
- [26] D. Nathani, J. Chauhan, C. Sharma, and M. Kaul, "Learning attention-based embeddings for relation prediction in knowledge graphs," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4710–4723.

- [27] Z. Zhang, F. Zhuang, H. Zhu, Z. Shi, H. Xiong, and Q. He, "Relational graph neural network with hierarchical attention for knowledge graph completion," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 9612–9619.
- [28] R. Ye, X. Li, Y. Fang, H. Zang, and M. Wang, "A vectorized relational graph convolutional network for multi-relational network alignment," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 4135–4141.
- [29] S. Vashishth, S. Sanyal, V. Nitin, and P. Talukdar, "Composition-based multi-relational graph convolutional networks," in *Proc. 8th Int. Conf. Learn. Representations*, 2020.
- [30] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. Neural Informat. Process. Syst.*, 2013, pp. 1–9.
- [31] B. Yang, S. W.-T. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015.
- [32] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, "Convolutional 2D knowledge graph embeddings," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1811–1818.
- [33] Y. Long, M. Wu, C. K. Kwoh, J. Luo, and X. Li, "Predicting human microbe–drug associations via graph convolutional network with conditional random field," *Bioinformatics*, vol. 36, no. 19, pp. 4918–4927, 2020.
- [34] Y.-A. Huang, P. Hu, K. C. Chan, and Z.-H. You, "Graph convolution for predicting associations between miRNA and drug resistance," *Bioinformatics*, vol. 36, no. 3, pp. 851–858, 2020.
- [35] J. Li, S. Zhang, T. Liu, C. Ning, Z. Zhang, and W. Zhou, "Neural inductive matrix completion with graph convolutional networks for mirna-disease association prediction," *Bioinformatics*, vol. 36, no. 8, pp. 2538–2546, 2020.
- [36] G. Zhou, J. Wang, X. Zhang, M. Guo, and G. Yu, "Predicting functions of maize proteins using graph convolutional network," *BMC Bioinf.*, vol. 21, no. 16, pp. 1–16, 2020.
- [37] G. Yu, G. Zhou, X. Zhang, C. Domeniconi, and M. Guo, "DMIL-IsoFun: Predicting isoform function using deep multi-instance learning," *Bioinformatics*, vol. 37, no. 24, pp. 4818–4825, 2021.
- [38] M. Nickel, L. Rosasco, and T. Poggio, "Holographic embeddings of knowledge graphs," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 1955–1961.
- [39] M. F. Rogers and A. Ben-Hur, "The use of gene ontology evidence codes in preventing classifier assessment bias," *Bioinformatics*, vol. 25, no. 9, pp. 1173–1177, 2009.
- [40] S. Jain and G. D. Bader, "An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology," *BMC Bioinf.*, vol. 11, no. 1, pp. 1–14, 2010.
- [41] C. Pesquita, D. Pessoa, D. Faria, and F. Couto, "CESSM: Collaborative evaluation of semantic similarity measures," *JB2009 Challenges Bioinf.*, vol. 157, 2009, Art. no. 190.
- [42] M. Paul and A. Anand, "A new family of similarity measures for scoring confidence of protein interactions using gene ontology," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 19, no. 1, pp. 19–30, Jan./Feb. 2022.
- [43] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.
- [44] C. Pesquita, D. Faria, H. Bastos, A. E. Ferreira, A. O. Falcão, and F. M. Couto, "Metrics for GO based protein semantic similarity: A systematic evaluation," *BMC Bioinf.*, vol. 9, pp. 1–16, 2008.
- [45] S. Bandyopadhyay and K. Mallick, "A new path based hybrid measure for gene ontology similarity," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 11, no. 1, pp. 116–127, Jan./Feb. 2014.
- [46] J. L. Sevilla *et al.*, "Correlation between gene expression and GO semantic similarity," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 2, no. 4, pp. 330–338, Oct./Dec. 2005.



Zhen Tian received the BSc degree in computer science and technology from Harbin Engineering University, China, in 2011, the MSc and PhD degrees in computer science and technology from the Harbin Institute of Technology, in 2013 and 2017, respectively. He is now working with the school of information and engineering, Zhengzhou University, China. His current research interests include bioinformatics and machine learning.



Haichuan Fang is currently working toward the master's degree of engineering with Zhengzhou University, Zhengzhou, China. His research interests include knowledge graph embedding, bioinformatics, and deep learning.



Zhixia Teng received the PhD degree in computer science and technology from the Harbin Institute of Technology, in 2016. She is currently a lecturer with the College of Information and Computer Engineering, Northeast Forestry University, Harbin, China. Her current research interests include bioinformatics, computational system biology, and machine learning.



Yangdong Ye (Member, IEEE) received the PhD degree from the China Academy of Railway Sciences, Beijing, China, in 2002. He is a professor with the School of Information Engineering, Zhengzhou University, Zhengzhou, China. He worked one year as a senior visiting scholar with Deakin University, Australia. He has published some papers in peer-reviewed prestigious journals and conference proceedings, such as *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, *IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS*, *IEEE TRANSACTIONS ON MULTIMEDIA*, *Information Fusion*, *Neural Networks*, *Pattern Recognition*, *Information Sciences*, *IEEE CVPR*, *IJCAI*, and *ACM Multimedia*. He has wide research interests, mainly including machine learning, pattern recognition, knowledge engineering, and intelligent systems.

▷ For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/csdl.