Bioinformatics, 2022, 1–7
https://doi.org/10.1093/bioinformatics/btac520
Advance Access Publication Date: 28 July 2022
Original Paper

Data and text mining

CoGO: a contrastive learning framework to predict disease similarity based on gene network and ontology structure

Yuhao Chen[1,†], Yanshi Hu [1,†], Xiaotian Hu[1], Cong Feng[1] and Ming Chen [1,2,3,*]

[1]Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou, 310058, China, [2]Biomedical Big Data Center, The First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou, 310058, China and [3]Institute of Hematology, Zhejiang University, Hangzhou, 310058, China

*To whom correspondence should be addressed.
†The authors wish it to be known that first two authors contributed equally.

Associate Editor: Zhiyong Lu

Received on January 27, 2022; revised on June 16, 2022; editorial decision on July 16, 2022

Abstract

Motivation: Quantifying the similarity of human diseases provides guiding insights to the discovery of micro-scope mechanisms from a macro scale. Previous work demonstrated that better performance can be gained by integrating multiview data sources or applying machine learning techniques. However, designing an efficient framework to extract and incorporate information from different biological data using deep learning models remains unexplored.

Results: We present CoGO, a Contrastive learning framework to predict disease similarity based on Gene network and Ontology structure, which incorporates the gene interaction network and gene ontology (GO) domain knowledge using graph deep learning models. First, graph deep learning models are applied to encode the features of genes and GO terms from separate graph structure data. Next, gene and GO features are projected to a common embedding space via a nonlinear projection. Then cross-view contrastive loss is applied to maximize the agreement of corresponding gene-GO associations and lead to meaningful gene representation. Finally, CoGO infers the similarity between diseases by the cosine similarity of disease representation vectors derived from related gene embedding. In our experiments, CoGO outperforms the most competitive baseline method on both AUROC and AUPRC, especially improves 19.57% in AUPRC (0.7733). The prediction results are significantly comparable with other disease similarity studies and thus highly credible. Furthermore, we conduct a detailed case study of top similar disease pairs which is demonstrated by other studies. Empirical results show that CoGO achieves powerful performance in disease similarity problem.

Availability and implementation: https://github.com/yhchen1123/CoGO.

Contact: mchen@zju.edu.cn

1 Introduction

Uncovering the associations among human diseases draws great attention from researchers. The similarity between different diseases can be measured within different biological scopes. It can be extended to address numerous questions at the forefront of network medicine (Menche *et al.*, 2015), from interpreting genome-wide association study data to drug target identification and repurposing. Current studies have confirmed that similar diseases tend to be caused by similar molecules (Cáceres and Paccanaro, 2019; Hu *et al.*, 2017), share similar biomarkers (Franke *et al.*, 2006; Tang *et al.*, 2018), and can be diagnosed by similar symptoms or be cured by similar drugs (Csermely *et al.*, 2013; Luo *et al.*, 2016). Therefore, it has attracted increasing attention to design accurate and efficient algorithms to make full use of the big data and prior biological knowledge in existing researches and databases.

Disease similarity can be calculated by shared molecular features (e.g. disease-associated genes, proteins, noncoding RNAs), phenotypes or semantic descriptions. Molecule-based methods utilize the qualitative associations between molecules and diseases from related data sources (Mathur and Dinakarpandian, 2012; Suthram *et al.*, 2010). Phenotype-based methods are analogous to that of the previously stated molecular-based methods. Instead, they use phenotype data to measure diseases (Freudenberg and Propping, 2002). Semantic-based approaches are based on the graph structure of disease-related ontologies (Wang *et al.*, 2007).

Apart from treating disease-related molecules separately, graph structure data, such as gene regulatory network, protein–protein

interaction network (PPIN) and pathways, is also used for disease similarity methods. Cheng *et al.* (2014) designed SemFunSim to calculate similarity by integrating the functional interactions of genes from HumanNet Kim (2022) and the relationship among diseases from disease ontology graph. Hamaneh and Yu (2015) used random walk to explore nontrivial similarity between diseases with known gene associations. Ni *et al.* (2020) followed disease module theory and measured associations between diseases by using disease-gene association data and PPIN. Oerton *et al.* (2019) integrated six types of biological data to construct disease networks and predicted disease relationships through similarity fusion. NETSIM2 (Peng *et al.*, 2018) considered both gene network structure and gene ontology (GO) graph structure to decrease the noise information.

However, the predictive power of the current feature extractors from raw data or graphs can be impeded because of human bias. In order to circumvent this issue, the graph deep learning methods allow for leveraging graph structure data about diseases to refine genetic and phenotypic disease relationships. Han *et al.* (2019) used graph convolutional networks (GCN) and matrix factorization to capture disease-gene association. Wang *et al.* (2020) applied GCN to predict circRNA-disease associations. Deep-DRM (Zhao *et al.*, 2021) used graph deep learning approaches to encode structure features of disease network and metabolite network and predicted disease-related metabolite. Li *et al.* (2021) integrated gene association network and GO graph to reconstruct gene network and utilized graph representation learning model to calculate disease similarity.

In parallel, contrastive learning is a self-supervised framework that learns by comparison which can be performed between positive pairs of 'similar' inputs and negative pairs of 'dissimilar' inputs. MoCo (He *et al.*, 2019) and SimCLR (Chen *et al.*, 2020) show that it has largely closed the gap between unsupervised and supervised representation learning in vision tasks. This framework has been extended to solve problems in bioinformatics. The scNAME (Wan *et al.*, 2022) combines contrastive paradigm into their method to learn underlying feature representation of scRNA-seq data for cell clustering. SMILE Xu (2022) adopts contrastive learning to perform single-cell omics data integration.

In summary, the current trends in disease researches are to incorporate the ontology-based background knowledge and disease-related molecular data, or construct disease molecular network and use graph deep learning model to extract the latent structure of raw data. However, both avenues of research have encountered their limitations: first, it can be inefficient and biased to use nonparametric models to process disease-related data. Second, graph deep learning model cannot be extended to multiview networks directly. To overcome these obstacles, we aim to develop an effective method for computation of gene and disease embedding in biological networks by integrating two key insights: (i) Graph deep learning models are potentially effective in computing powerful node embedding for biological networks. (ii) Using a contrastive learning framework to combine biological data and prior knowledge gains better performance in downstream tasks without introducing human bias.

In this study, we present CoGO (Contrastive learning framework to predict disease similarity based on Gene network and Ontology structure), a contrastive learning model which uses an intra-view GCN model to learn node embedding and applies cross-view model to jointly encode both the gene network and GO graph. The cross-view contrastive loss can make the genes and their corresponding GO annotations in similar positions in the embedding space.

Furthermore, we compare and evaluate the performance of our method against other disease similarity methods on the benchmark set. Experimental results demonstrate our method outperforms others to discover potential similar diseases. In addition, our method can identify corresponding similar diseases that do not appear in the benchmark but have been confirmed by publications.

2 Materials and methods

In this section, we first describe the data used in our model. Then we will introduce the overview of CoGO and its implementation and application as shown in Figure 1. During the training stage, it joint learns structure features from gene network and GO knowledge graph using two model components: *intra-view model* and *cross-view model*. In the inference stage, we can only use part of the intra-view model to calculate the gene representations and get the discriminative disease representations.

2.1 Datasets

We integrate the gene interaction network, the GO graph, the gene-GO associations and the gene-disease associations into a unified computational framework.

The gene interaction network is derived from HumanNet (Kim, 2022). This database covers 99.8% protein-encoding genes of *Homo sapiens* and is constructed by means of the expanded data with network inference algorithms. The weight of links in this network is calculated as log-likelihood score (LLS) based on their Bayesian statistics algorithms. It supports a three-tier model: a protein–protein physical interaction network HumanNet-PI, a functional gene network HumanNet-FN and a functional network extended by co-citation HumanNet-XC. We use HumanNet-FN as the training data.

The GO knowledgebase is the world's largest source of information on the functions of genes (Ashburner *et al.*, 2000; Carbon *et al.*, 2021). GO annotations are divided into three top-level branch terms: *biological_process* (BP), *molecular_function* (MF), *cellular_component* (CC). All GO terms and the set of inclusion relationships form a hierarchy and directed acyclic graph of GO. This GO graph can be modeled as a multirelational graph in our framework. Unlike other disease similarity methods, we use all three branches of GO instead of only considering the BP branch. Besides, the gene-GO associations are downloaded from NCBI Gene database to bridge gene instances and ontology concepts.

The gene-disease associations are obtained from DisGeNET database (Piñero *et al.*, 2020). DisGeNET is a knowledge management platform integrating and standardizing data about disease-associated genes and variations, which covers more than 24 000 diseases, 17 000 genes and 117 000 genomics variations.

2.2 Networks construction

To formulize all the graph structure data above, we use $\mathcal{G}_g$, $\mathcal{G}_o$ and $\mathcal{D}$ to denote gene interaction network, GO graph and the disease set separately. The gene network is defined as an undirected weighted graph $\mathcal{G}_g \in (\mathcal{V}, \mathcal{E}, W)$, where $\mathcal{V}$ represents the $|\mathcal{V}| = n$ vertices in gene network, $\mathcal{E}$ is a set of edges and $W \in \mathbb{R}^{n\times n}$ is a weighted adjacency matrix encoding the edge weight between two genes. The original weight between g_i and g_j is their associated LLS provided by HumanNet. We normalize the edge weight as follow to rescale it between 0 and 1:

$$w'_{i,j} = \frac{w_{i,j} - w_{min}}{w_{max} - w_{min}}$$

The GO graph is defined as a directed and labeled multirelational graph as $\mathcal{G}_o \in (\mathcal{V}, \mathcal{E}, \mathcal{R})$ with vertices and labeled edges. $(v_i, r, v_j) \in \mathcal{E}$, where $r \in \mathcal{R}$ is a relation type. For example, glucose transport, *is_a*, monosaccharide transport denotes that the glucose transport is a subtype of monosaccharide transport. To make the GO graph more suitable for our intra-view model processing, we add inverse relation and self-connection to $\mathcal{R}$. Though most GO-based methods only use relation 'is_a' and BP branch to construct models, in our case, we use all GO terms and relationships in GO.

2.3 Intra-view model

The intra-view model consists of two graph encoders $f(\cdot)$ to extract the original structure information in gene interaction network $\mathcal{G}_g$ and GO graph $\mathcal{G}_o$ separately to corresponding embedding spaces. We opt for simplicity and adopt the commonly used two-layer GCN to obtain gene embedding $\boldsymbol{u} = f_g(\mathcal{G}_g)$, where $\boldsymbol{u} \in \mathbb{R}^d$ is the output of GCN. Similarly, we use two-layer relational graph convolutional network (RGCN) to obtain the GO embedding $\boldsymbol{v} = f_o(\mathcal{G}_o)$, where

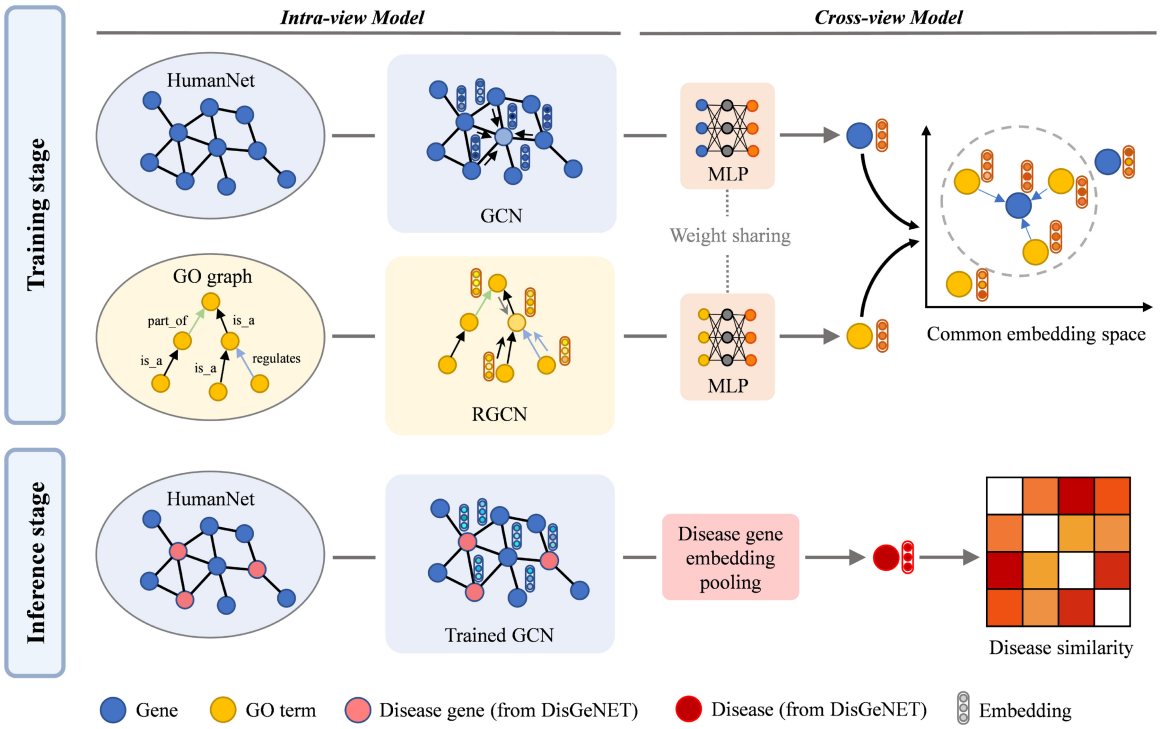

Fig. 1. Overview of CoGO. In the training stage, GCN and RGCN are implemented to encode features of gene interaction network and GO graph. MLP is applied to map the output of GCN and RGCN to the common embedding space. Contrastive loss is used to maximize the agreement of corresponding genes and GO terms. In the inference stage, only trained GCN is preserved to calculate the gene embedding. And disease representation is derived from related gene embedding by average pooling

$\boldsymbol{v} \in \mathbb{R}^d$ is the output of RGCN. The output dimension of GCN and RGCN should be the same for the ease of projection and loss calculation.

The GCN model uses both the graph structure and node features to learn the embedding of nodes. In original GCN (Kipf and Welling, 2017), multilayer perceptron (MLP) and summation operation over connected nodes are combined together to propagate features among neighbors:

$$\mathbf{X}^{(l+1)} = \sigma(\hat{A}\mathbf{X}^{(l)}W^{(l)})$$

where $\mathbf{X}^{(l)}$ means the l^{th} layer features and $\mathbf{X}^0$ is the original node features. $W^{(l)}$ is the weight of l^{th} linear layer and σ is an activation function. We initialize $\mathbf{X}^0$ as one-hot coding of each gene and use leaky-ReLU as the activation function. The normalized adjacency matrix $\hat{A}$ is utilized to encode structure information:

$$\hat{A} = \mathbf{D}^{-\frac{1}{2}}(A + I)\mathbf{D}^{-\frac{1}{2}}$$

where I is an identity matrix and denotes self-connection, $\mathbf{D}$ is a diagonal degree matrix where $\mathbf{D}_{(i,i)} = \sum_j A_{(i,j)}$. The adjacency matrix can include other values than one representing edge weights.

Different from GCN, RGCN model introduces relation-specific transformations depending on the type and direction of an edge (Schlichtkrull *et al.*, 2017). It is motivated as an extension of GCN to the large-scale relational graph. In GCN, weight $W^{(l)}$ is shared by all edges in layer l. In contrast, in RGCN, different edge types use different weights and only edges of the same relation type r are associated with the same transformation weight $W_r^{(l)}$. It defines a similar propagation model for calculating the forward-pass update of the node feature denoted by v_i in a directed and labeled multirelational graph:

$$v_i^{l+1} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r}\frac{1}{c_{i,r}}W_r^{(l)}v_j^{(l)} + W_0^{(l)}v_i^{(l)}\right)$$

where $\mathcal{N}_i^r$ denotes the set of neighbor indices of node i under relation $r \in \mathcal{R}$. $c_{i,r} = |\mathcal{N}_i^r|$ is a normalization constant. $W_r^{(l)}$ is the relation-specific weight matrix in layer l and $W_0^{(l)}$ is the weight of self-connection in layer l. In our case, we use one-hot coding as the initial feature of GO terms and leaky-ReLU as the activation function.

2.4 Cross-view model

To enable the feature extraction with gene-GO associations, it is intuitive that relevant genes and GO annotations should be similar to each other and unrelated pairs should be far away in common embedding space. This is consistent with the assumption of contrastive learning framework. With such motivation, we propose a cross-view contrastive loss that enables model to learn gene-GO connections to improve the embedding of the previous intra-view model.

The cross-view model is a neural network to perform nonlinear projection that maps intra-view embedding vectors to the space where contrastive loss is applied. We use MLP with one hidden layer to obtain $\boldsymbol{z} = g(\boldsymbol{h}) = W^{(2)}\sigma(W^{(1)}\boldsymbol{h})$ where σ is a ReLU function and $\boldsymbol{h}$ can be either gene embedding $\boldsymbol{u}$ or GO embedding $\boldsymbol{v}$. To maximize the agreement between the gene instances and corresponding ontology concepts, we propose a cross-view contrastive loss to learn the embedding $\boldsymbol{z}$ and model parameters. We treat the genes and their corresponding GO terms as positive pairs and others as negatives. Let $sim(z_i, z_j) = z_i^\top z_j / ||z_i|| ||z_j||$ denotes the dot product between ℓ_2 normalized z_i and z_j. Then the cross-view contrastive loss function for gene i is defined as

$$\ell_i = -\log\frac{\sum_{j=1}^{N_o}\mathbb{I}_{i,j}\exp\left(sim(\boldsymbol{z}_i^g, \boldsymbol{z}_j^o)/\tau\right)}{\sum_{k=1}^{N_o}\exp\left(sim(\boldsymbol{z}_i^g, \boldsymbol{z}_j^o)/\tau\right)}$$

where $\mathbb{I}_{i,j} \in \{0, 1\}$ is an indicator function equal to 1 if gene i is annotated by GO term j. And τ denotes a temperature parameter controlling the scale of distribution and we set it equal to 1.

The cross-view contrastive learning paradigm provides each gene with some intuitive GO annotation information and promotes the automatic cohesion of genes engaged in biological process, performing the similar molecular function or occurring in the same cell component. Whereas mutual repulsion forces exist between genes with

different GO annotations. This process refines gene representations by incorporating the information from GO annotations and results in meaningful disease representations.

2.5 Model training

CoGO uses the GCN and RGCN mentioned in the intra-view model to obtain the structure information in gene network $\mathcal{G}_g$ and GO graph $\mathcal{G}_o$ parallelly. And we adopt a weight-sharing nonlinear projection before calculating the contrastive loss. We use the adaptive learning rate algorithm RMSProp as an optimizer. After several rounds of parameter optimization, the model output tends to converge. Then we only preserve the GCN encoder $f_g(\cdot)$ and throw away other components for downstream computation.

2.6 Disease similarity inference

In the inference stage of CoGO, We calculate the similarity between diseases through the embedding vector of disease-related genes. First, we extract gene set G_i relevant to disease $d_i \in \mathcal{D}$. Second, we perform average pooling on gene embedding to obtain the representation of each disease:

$$d_i = \frac{1}{|G_i|} \sum_{g_j \in G_i} \boldsymbol{u}_j$$

where $\boldsymbol{u}_i$ is the representation of a gene g_i in gene set G_i. And we can get the similarity between two diseases by measuring the cosine similarity of their representation vectors.

We use the gene embedding $\boldsymbol{u}$ after GCN rather than $\boldsymbol{z}$ after nonlinear projection for the reason: previous work (Chen *et al.*, 2020) about contrastive learning demonstrated that a nonlinear projection head improves the representation quality of the layer before it because the contrastive loss can induce loss of information. In our cases, the nonlinear projection in the cross-view model can weaken the graph structure information that is crucial for the disease similarity calculation. And our ablation experiments also confirmed this point.

3 Results

3.1 Prediction of disease similarity

3.1.1 Benchmark

It is important to choose high-quality benchmark as ground truth to examine predictive models. We follow other disease similarity methods and use the same benchmark as theirs (Li *et al.*, 2021). It integrates highly similar disease pairs derived from two different data sources: one is predicted by multiple human molecular networks (Suthram *et al.*, 2010), and it has been further verified by Mathur and Dinakarpandian (2012) according to literature. The other is discovered by Pakhomov *et al.* (2010) according to the electronic health records (EHR) of the US population. Thus, this benchmark can balance the impact of phenotypic and molecular level on disease similarity assessment. Besides, this benchmark data is highly imbalanced because of few positive pairs compared to the number of all potential disease pairs.

3.1.2 Experimental setup

A previous study (Li *et al.*, 2021) demonstrated Li's method achieved state-of-the-art (SOTA) performance and improved the Area Under Receiver Operating Characteristic (AUROC) score by 10.1%. For comparison, we use this method as the baseline:

- Li method (Li *et al.*, 2021) integrates gene interaction network and GO hierarchy to reconstruct a novel gene network, and utilizes graph representation learning model LINE (Tang *et al.*, 2015) to learn gene representations. And disease similarity is calculated by the cosine similarity of disease embedding derived from related gene representations.

We run CoGO for 200 epochs with a learning rate of 0.003 until convergence. In addition, we set hyper-parameter τ to 1.0 the dimension of all hidden layers in our model to 32. For Li's method, the implementation is consistent with the settings in their paper.

3.1.3 Evaluation metrics

Evaluation metric plays a critical role in achieving the optimal discriminator in the inference stage. Although most disease similarity researches adopt Receiver Operating Characteristic (ROC) curve and AUROC to evaluate their model, we introduce Precision–Recall Curve (PRC) and Area Under PRC (AUPRC) to evaluate our model for two main reasons: first, ROC curve can present an overly optimistic view of a model's performance when applied to imbalanced data sets. Second, PRC gives a more informative picture of an algorithm when dealing with highly skewed datasets. And AUPRC is useful when we care about our model handling the positive examples correctly.

Both AUROC and AUPRC are restricted to lie in $[0, 1]$. A larger numerical metric represents better model performance.

3.1.4 Experimental results

Figures 2 and 3 show the ROC, PRC and corresponding AUROC, AUPRC obtained by CoGO and Li. The difference between AUROC reported here against the original paper is caused by the negative sampling strategy and stochastic parameter initialization and optimization methods. The AUROC score of our method is higher than the previous SOTA by 2.43%. This demonstrates that our method achieves the best performance in contrast to traditional feature design methods and even the graph representation learning method. Although in AUROC, the gap between CoGO and Li is not as much as Li and the third-best method (which is 10.1%), CoGO shows significantly high AUPRC than Li by 19.57%. In other words, whereas Li can achieve competitive performance in identifying true similar disease pairs, the false positive rate is very high. In comparison, CoGO can not only effectively find true positive disease similar pairs, but also discriminate the false positive samples.

In order to examine the evaluation effect of different ground truth datasets on disease similarity prediction task, we compare CoGO and Li's model in the two latest datasets (Dong *et al.*, 2021; Westergaard *et al.*, 2019; Xu, 2022). Westergaard *et al.* constructed a comprehensive map of disease co-occurrences in the complete Danish population. Dong *et al.* investigated the multimorbid relations among 439 common diseases using hospital inpatient data in the UK Biobank. Both of them assess disease relations in phenotypic level based on patients' diagnoses in EHR. Meanwhile, considering that disease pairs of original benchmark dataset are derived from US EHR and molecular network, each of which was treated as baselines

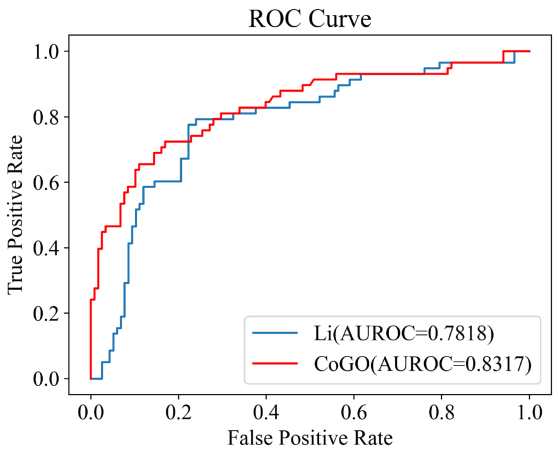

Fig. 2. ROC curve and AUROC scores from CoGO and previous SOTA method

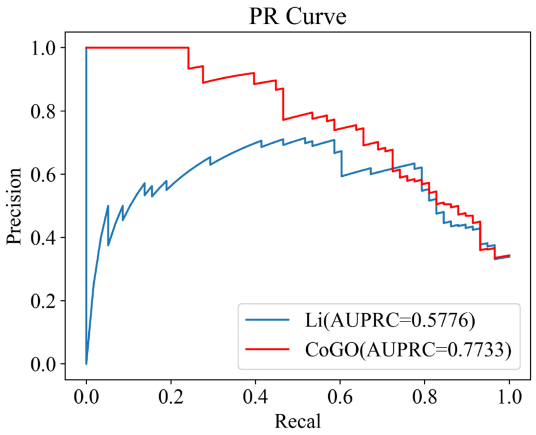

Fig. 3. PR curve and AUPRC scores from CoGO and previous SOTA method

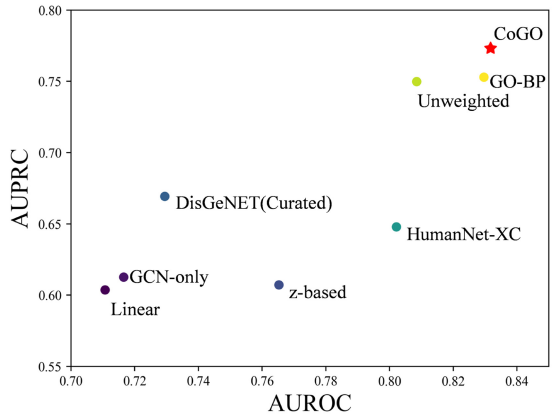

Fig. 4. Performance of CoGO and its variants in ablation experiment

Table 1. Model comparisons in different baseline datasets

Dataset	AUROC		AUPRC	
	Li	CoGO	Li	CoGO
Westergaard *et al.* (Danish EHR)	0.4453	0.5006	0.4746	0.5232
Dong *et al.* (UK EHR)	0.4427	0.4773	0.4823	0.4918
Benchmark (US EHR)	0.4280	0.5857	0.4495	0.7003
Merged (Danish+UK+US EHR)	0.4499	0.5339	0.4514	0.5723
Benchmark (molecular network)	0.6354	0.7438	0.5739	0.8063
Benchmark (US EHR+molecular network)	0.7818	0.8317	0.5776	0.7733

as well. Table 1 shows that CoGO outperforms Li consistently in various baseline datasets. Prediction performance on EHR-based datasets is relatively unsatisfactory and merging all three EHR-based baselines from Danish, the UK and the US population cannot improve prediction performance. The above results might indicate limited contributions of EHR-based phenotypic features to model assessment. In contrast, performance of CoGO and Li on molecular network-based dataset improves significantly compared to counterparts on EHR-based datasets, demonstrating prominent advantage of molecular over phenotypic features in model evaluation. Notably, both CoGO and Li achieve optimal performance on the original benchmark, indicating its superiority in orchestrating phenotypic and molecular level data.

In addition, one of the most advantages of CoGO is the way it processes the input raw data. Traditional nonparametric models take hours to extract features in raw data which are time-consuming and hard to extend to high throughput computation. Li's model, though takes an advanced graph learning model, spends more than 10 h in constructing gene network and 192 min in model training in Intel Core i7-10700 CPU @ 2.90 GHz. In contrast, CoGO follows the end-to-end learning paradigm and can finish training and disease similarity inference within 32 min in the same equipment.

Overall, the results above demonstrate CoGO achieves SOTA performance in disease similarity problems by integrating multiview data into a contrastive learning framework.

3.2 Ablation experiment

To demonstrate the relative importance of each dataset and our model structure, we conduct ablation analysis by selecting different network data or removing some characteristics of our model components.

Specifically, we denote our original model as 'CoGO' and use detailed descriptions to the variants. From the view of the dataset, we select BP branch in GO graph, HumanNet-XC and curated DisGeNET respectively. To further investigate the power of components in the intra-view model, we replace RGCN model with GCN to evaluate the effect of relation-specific transformations in GO graph and denote it as 'GCN-only'. In addition, we treat the interaction between genes uniformly and use unweighted HumanNet-FN as the input of GCN and denote it as 'Unweighted'. We also modify the cross-view model by replacing the nonlinear projection with linear projection and denote it as 'Linear'. We also utilize the embedding z after projection to infer the representation of disease and denote it as 'z-based'.

The results are shown in Figure 4. Using BP branch for model training has a trivial influence effect on model performance. This may be because BP branch contains most of the annotation information and close genes in the network are more likely to engage in the same biological process rather than have similar molecular functions or appear in the same cellular component. HumanNet-XC is the functional gene network extended network by co-citation, which has more noise in gene interactions and leads to model performance degradation. The curated gene-disease associations in DisGeNET are high confidence but may lose some disease-related gene information.

To assess the contribution of individual factors to CoGO, we evaluate both intra-view model and cross-view model. As shown in Figure 4, the RGCN model and the nonlinear projection are the two most important components. The relation-specific modeling in RGCN model shows a promising effect in knowledge graph information extraction. Nonlinear projection before calculating contrastive loss plays crucial importance in our full model. And as illustrated in Section 2.5, using embedding z will result in a decrease in model performance. In addition, we can gain better performance by incorporating edge weight in our GCN model. The result shows that a variety of different mechanisms contribute to CoGO's performance.

3.3 Comparison with other disease similarity studies

To validate the credibility of the disease relationships found by CoGO, we compare them to the disease-associations related studies conducted by Dong *et al.* (2021), Zhou *et al.* (2014) and Sánchez-Valle *et al.* (2020). Dong *et al.* investigated the multimorbid relations among 439 common diseases using hospital inpatient data in the UK Biobank. Zhou *et al.* used biomedical literature database to investigate the symptom-based similarity of two diseases. Sánchez-Valle *et al.* inferred disease interactions from similarities between patients' gene expression profiles. To infer credible disease similarity pairs, we select the similarity score which maximizes the F1 score in our benchmark datasets as the threshold. We calculate the odds ratio (OR) and use Fisher exact test to evaluate the overlap between different studies. Results are shown in Table 2.

There are 435 diseases commonly used by Dong *et al.* and us. The comparison results show that the multimorbid relations identified by Dong *et al.* and our model have significant overlap (OR = 2.3, P = 1.1e-288). Zhou's study shared 1217 diseases with

Table 2. Disease similarity comparisons between Dong *et al*., Zhou *et al*., Sánchez-Valle *et al*. and ours

Study 1	Study 2	Shared diseases	Disease pairs in Study 1	Disease pairs in Study 2	Overlapping disease pairs	OR	*P*-values
Ours	Dong *et al.*	435	15 463	12 005	3435	2.3	1.1e-288
	Zhou *et al.*	1217	39 561	74 741	9992	3.3	0
	Sánchez-Valle *et al.*	103	2077	2241	846	0.9	0.9897
Sánchez-Valle *et al.*	Dong *et al.*	60	798	345	147	0.9	0.8623
	Zhou *et al.*	57	611	579	233	1.1	0.1229
Dong *et al.*	Zhou *et al.*	343	3327	4146	1093	8.4	0

ours and the overlap of disease relations is also highly significant (OR = 3.3, $P = 0$). While there are a total of 103 diseases commonly used by Sánchez-Valle *et al.* and us, the comparison shows that our result is inconsistent with theirs (OR = 0.9, $P = 0.9897$). To further investigate the reliability of Sánchez-Valle's result, we compare these three disease similarity studies with each other. Result shows that disease relationships found by Dong *et al.* and Zhou *et al.* share significantly (OR = 8.4, $P = 0$). However, there is no obvious overlap between Sánchez-Valle's and Dong's (OR = 0.9, $P = 0.8623$) or Zhou's (OR = 1.1, $P = 0.1229$). Thus, from the perspective of the above comparative studies, disease similarity result from Sánchez-Valle *et al.* should be treated with caution.

Taken together, disease relationships inferred by CoGO could be confirmed by previous parallel studies and highly credible.

3.4 Case study

To further verify the generalization power of our model, we address two intuitive questions: first, does our model perform equally in both widely studied diseases and relatively rare diseases? Second, does our model discriminate similarity simply based on the intersection of genes?

Thus, we select several diseases with different sizes of associated genes and find their top five similar diseases to do a case study. We study the asthma with 3443 related genes, bipolar disorder with 1183 related genes and Chronic Progressive External Ophthalmoplegia (CPEO) with only 52 related genes. Furthermore, we use Jaccard Index (JI_g) based on common genes to evaluate the similarity between two disease-related gene sets G_i and G_j as follow:

$$JI_g(G_i, G_j) = \frac{|G_i \cap G_j|}{|G_i \cup G_j|}$$

As shown in Table 3, 14 of 15 predicted similar diseases have been reported by other researchers and two of them are further included in the benchmark. For example, recent research conducted a cross-sectional study of electronic health record information for 56.6 million Americans. Results showed that asthma is significantly more common in those with multiple sclerosis than in the general population—particularly in the young and elderly—irrespective of gender and race. Besides, although bipolar disorder and drug dependence share few common genes according to their JI_g, Leventhal Adam and Zimmerman (2010) used individual logistic regression models to indicate that presence of lifetime Bipolar Disorder was associated with significant increases in rates of lifetime drug dependence. The result shows that CoGO maintains good performance in both widely studied diseases and relatively rare diseases. This is attributed to CoGO aiding gene representation learning with corresponding ontology concepts.

Although most of the similar diseases possess relatively low JI_g, they have been confirmed to have comorbidity by literature. This illustrates that although many similar diseases cannot be discriminated based on their related genes, CoGO is able to learn meaningful disease representation containing gene interaction patterns and gene annotation information.

Table 3. Top five similar diseases predicted by CoGO

Aim	Similar diseases	JI_g	Evidence
Asthma	Multiple sclerosis	0.259	PMID: 30557818
	Rheumatoid arthritis	0.291	PMID: 32906033
	Autoimmune diseases	0.261	PMID: 31219041
	Inflammatory bowel diseases	0.253	PMID: 30250122
	Psoriasis	0.243	PMID: 29490768
Bipolar disorder	Schizophrenia	0.250	PMID: 29906448
	Anxiety	0.173	PMID: 25617037
	Drug dependence	0.065	PMID: 20565163
	Obesity	0.132	PMID: 24194362
	Mood disorders	0.219	PMID: 12071513
CPEO	MERRF syndrome	0.180	PMID: 9436447
	Pigmentary retinopathy	0.045	PMID: 22993469
	Nocturia	0.086	None
	Respiratory insufficiency	0.019	PMID: 21533826
	MELAS syndrome	0.183	PMID: 8363452

Note: Red: the similar disease in benchmark.

4 Conclusion

In this work, we present a novel method 'CoGO' to incorporate the disease-related molecular data and ontology-based domain knowledge using competitive graph deep learning models, demonstrating its implementation and performance on benchmark evaluation and detailed case study. Our approach fully follows the modern deep learning paradigm and circumvents the manual feature extraction step which is highly biased and time-consuming.

By combining our findings, we show that contrastive learning, which dominants the self-supervised learning domain in vision tasks, can learn densely informative representations from multiview data sources and be extended to solve other problems like integrated bioinformatics analysis. In addition, although using GCNs in the molecular graph has been widely adopted, relation-specific modeling of knowledge graph shows promising direction to extract human-curated domain knowledge.

We also find current obstacles in disease similarity problems. The benchmark is out-of-date and only covers a small portion of disease space, which hinders the development of machine learning models. In addition, the similarity of diseases can be measured from different aspects such as phenotypes, molecular features or medicine treatments. Thus, fine-grained measurement is needed to transform disease similarity problem from binary classification tasks to multi-label classification problems.

Acknowledgements

The authors thank the anonymous reviewers for their constructive comments and the members of Ming Chen's laboratory for helpful discussions and valuable comments.

Funding

This work was supported by the National Natural Sciences Foundation of China [32070677]; the 151 Talent Project of Zhejiang Province (first level); Jiangsu Collaborative Innovation Center for Modern Crop Production and Collaborative Innovation Center for Modern Crop Production cosponsored by province and ministry.

Conflict of Interest: none declared.

Data availability

The dataset for the experiment can be obtained at https://github.com/yhchen1123/CoGO.

References

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Cáceres,J.J. and Paccanaro,A. (2019) Disease gene prediction for molecularly uncharacterized diseases. *PLoS Comput. Biol.*, **15**, e1007078.

Carbon,S. *et al.* (2021) The gene ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.

Chen,T. *et al.* (2020). A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*.

Cheng,L. *et al.* (2014) SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association. *PLoS One*, **9**, e99415.

Csermely,P. *et al.* (2013) Structure and dynamics of molecular networks: a novel paradigm of drug discovery. *Pharmacol. Ther.*, **138**, 333–408.

Dong,G. *et al.* (2021) A global overview of genetically interpretable multimorbidities among common diseases in the UK biobank. *Genome Med.*, **13**, 1–20.

Franke,L. *et al.* (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.*, **78**, 1011–1025.

Freudenberg,J. and Propping,P. (2002) Prediction of Disease-Relevant human genes. *Bioinformatics*, **18**, S110–S115.

Hamaneh,M.B. and Yu,Y.K. (2015) DeCoaD: determining correlations among diseases using protein interaction networks. *BMC Res. Notes*, **8**, 1–7.

Han,P. *et al.* (2019). Gcn-mf: Disease-gene association identification by graph convolutional networks and matrix factorization. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 705–713, New York, NY, USA. Association for Computing Machinery.

He,K. *et al.* (2019). Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*.

Hu,Y. *et al.* (2017) Measuring disease similarity and predicting disease-related ncRNAs by a novel method. *BMC Med. Genomics.*, **10**, 71.

Kim,C.Y. *et al.* (2022) HumanNet v3: An improved database of human gene networks for disease research. *Nucleic Acids Res.*, **50**, D632–D639. https://doi.org/10.1093/nar/gkab1048.

Kipf,T.N. and Welling,M. (2017). Semi-supervised classification with graph convolutional networks. In: *International Conference on Learning Representations (ICLR)*. Palais des Congrès Neptune, Toulon, France.

Leventhal,A.M. and Zimmerman, M. (2010) The relative roles of bipolar disorder and psychomotor agitation in substance dependence. *Psychol. Addict. Behav.*, **24**, 360–365.

Li,Y. *et al.* (2021) Evaluating disease similarity based on gene network reconstruction and representation. *Bioinformatics*, **37**, 3579–3587.

Luo,H. *et al.* (2016) Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. *Bioinformatics*, **32**, 2664–2671.

Mathur,S. and Dinakarpandian,D. (2012) Finding disease similarity based on implicit semantic similarity. *J. Biomed. Inform.*, **45**, 363–371.

Menche,J. *et al.* (2015) Uncovering disease-disease relationships through the incomplete interactome. *Science*, **347**, 1257601.

Ni,P. *et al.* (2020) Constructing disease similarity networks based on disease module theory. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **17**, 906–915.

Oerton,E. *et al.* (2019) Understanding and predicting disease relationships through similarity fusion. *Bioinformatics*, **35**, 1213–1220.

Pakhomov,S. *et al.* (2010) Semantic similarity and relatedness between 520 clinical terms: An experimental study. *AMIA Annual Symposium 521 proceedings/AMIA Symposium. AMIA Symposium*, 2010, pp. 572–576.

Peng,J. *et al.* (2018) Improving the measurement of semantic similarity by combining gene ontology and co-functional network: a random walk based approach. *BMC Syst. Biol.*, **12**, 18.

Piñero,J. *et al.* (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.*, **48**, D845–D855.

Sánchez-Valle,J. *et al.* (2020) Interpreting molecular similarity between patients as a determinant of disease comorbidity relationships. *Nat. Commun.*, **11**, 1–13.

Schlichtkrull,M. *et al.* (2017). Modeling relational data with graph convolutional networks. *arXiv preprint arXiv:1703.06103*.

Suthram,S. *et al.* (2010) Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Comput. Biol.*, **6**, e1000662.

Tang,J. *et al.* (2015). Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pp. 1067–1077, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Tang,W. *et al.* (2018) Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinformatics*, **34**, 398–406.

Wan,H. *et al.* (2022) scNAME: neighborhood contrastive clustering with ancillary mask estimation for scRNA-seq data. *Bioinformatics*, **38**, 1575–1583.

Wang,J.Z. *et al.* (2007) A new method to measure the semantic similarity of GO terms. *Bioinformatics*, **23**, 1274–1281.

Wang,L. *et al.* (2020) GCNCDA: a new method for predicting circRNA-disease associations based on graph convolutional network algorithm. *PLoS Comput. Biol.*, **16**, e1007568.

Westergaard,D. *et al.* (2019) Population-wide analysis of differences in disease progression patterns in men and women. *Nat. Commun.*, **10**, 1–14.

Xu,Y. *et al.* (2022) SMILE: mutual information learning for integration of single-cell omics data. *Bioinformatics*, **38**, 476–486. https://doi.org/10.1093/bioinformatics/btab706.

Zhao,T. *et al.* (2021) Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep learning approaches. *Briefings in Bioinformatics*, **22**, 1–9.

Zhou,X. *et al.* (2014) Human symptoms–disease network. *Nat. Commun.*, **5**, 1–10.