Liu, Lizhi; Mamitsuka, Hiroshi; Zhu, Shanfeng

**HPOD Nets: deep graph convolutional networks for predicting human protein–phenotype associations**

*Published in:*
Bioinformatics

**DOI:**
10.1093/bioinformatics/btab729

Published: 01/02/2022

*Document Version*
Peer reviewed version

*Please cite the original version:*
HPODNets: deep graph convolutional networks for predicting human protein-phenotype associations

Lizhi Liu 1, Hiroshi Mamitsuka 2,3 and Shanfeng Zhu 4,5,6,7,8,9,∗

1 School of Computer Science, Fudan University, Shanghai 200433, China. 2 Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto Prefecture, Japan. 3 Department of Computer Science, Aalto University, Espoo, Finland. 4 Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China. 5 Ministry of Education, Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Fudan University), Shanghai 200433, China. 6 MOE Frontiers Center for Brain Science, Fudan University, Shanghai 200433, China. 7 Zhangjiang Fudan International Innovation Center, Shanghai 200433, China. 8 Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China. 9 Institute of Artificial Intelligence Biomedicine, Nanjing University, Nanjing, China.

∗ To whom correspondence should be addressed.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Deciphering the relationship between human genes/proteins and abnormal phenotypes is of great importance in the prevention, diagnosis and treatment against diseases. The Human Phenotype Ontology (HPO) is a standardized vocabulary that describes the phenotype abnormalities encountered in human disorders. However, the current HPO annotations are still incomplete. Thus, it is necessary to computationally predict human protein-phenotype associations. In terms of current, cutting-edge computational methods for annotating proteins (such as functional annotation), three important features are 1) multiple network input, 2) semi-supervised learning, and 3) deep graph convolutional network (GCN), whereas there are no methods with all these features for predicting HPO annotations of human protein.

Results: We develop HPODNets with all above three features for predicting human protein-phenotype associations. HPODNets adopts a deep GCN with eight layers which allows to capture high-order topological information from multiple interaction networks. Empirical results with both cross-validation and temporal validation demonstrate that HPODNets outperforms seven competing state-of-the-art methods for protein function prediction. HPODNets with the architecture of deep GCNs is confirmed to be effective for predicting HPO annotations of human protein and, more generally, node label ranking problem with multiple biomolecular networks input in bioinformatics.

Availability: https://github.com/liulizhi1996/HPODNets
Contact: zhusf@fudan.edu.cn
Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Uncovering genetic causes of human disorders is a critical step in designing diagnosis, therapy and prevention strategies against diseases (Opap and Mulder, 2017). To facilitate the description of phenotype abnormalities encountered in human diseases, Human Phenotype Ontology (HPO) (Köhler et al., 2021) was built upon a standardized vocabulary to provide unified representation of phenotypic anomalies and of their semantic relationships. Akin to Gene Ontology (GO) (Ashburner et al., 2000), the terms in HPO are organized as a directed acyclic graph (DAG), in which a more specialized term (child) can be related to more than one less specialized term (parent). Through gene-disease links derived from OMIM (Hamosh et al., 2002) and Orphanet (Pavan et al., 2017), hereditary diseases can be annotated with HPO terms, resulting in gene-HPO term associations. Determining the HPO annotations of human genes can promote disease gene identification and prioritization and hence assist clinical diagnostics.
As of October 2020, over 4,000 proteins have been annotated with HPO terms, while the current HPO annotations would be still incomplete (Gao et al., 2019; Liu et al., 2021). Fig. 1 shows the average number of annotated proteins over HPO terms used to curated in March 2018 release. In the past two years, an average of 22.09 novel proteins were annotated per HPO term, which implies that more proteins possibly have yet to be annotated. Nevertheless, it is laborious to experimentally confirm an association between a protein and an abnormal phenotype. Therefore, it is imperative to develop a computational approach to predict HPO annotations of human proteins (which hereafter we call simply ‘HPO prediction’).

In recent years, several HPO predicting methods has been proposed (Van Ljundt et al., 2013; Fang and Gough, 2013). In CAFA challenge, the organizers evaluate the participating methods by both protein-centric and term-centric metrics (Radojnov, 2013). The former measures the accuracy of predicted annotations for a given novel protein, while the latter measures the accuracy of predicted proteins associated with a given particular term. For the same task, however, different performance measures can lead to different rankings of methods. A method which shows impressive term-centric performance may fail in the protein-centric evaluation, and vice versa (Kahanda et al., 2015a). Typically, molecular biologists and physicians are interested in knowing not only the phenotypic abnormalities associated with a particular human protein but also the set of proteins associated with a certain HPO term (Notaro et al., 2017). While the optimization of most of existing HPO predicting methods is targeted towards the higher protein-centric performance (Kahanda et al., 2015b; Notaro et al., 2017; Dogan, 2018; Liu et al., 2020; Komorar and Hoehndorf, 2020), for facilitating disease gene prioritization and clinical diagnosis, it is imperative to develop a method which particularly has a superior term-centric performance.

HPO prediction can be a problem analogous to protein function prediction, which predicts protein-GO term associations (which hereafter we call ‘GO prediction’). This is because HPO as well as GO are both ontologies with the form of DAG for relations among terms. In spite of lacking term-centric methods for HPO prediction, there already exist a certain number of methods for GO prediction in term-centric manner. Thus a reasonable way to develop a powerful method for solving HPO prediction would be to check the latest methods for GO prediction and to make the most of the cutting-edge techniques of these methods.

We raise the following three features from the literature of term-centric GO prediction methods: 1) Multiple protein networks: first of all, the current successful approaches have used multiple protein-protein interaction (PPI) networks (Mostafavi et al., 2008; Mostafavi and Morris, 2010; Cho et al., 2016; Gligorijevic et al., 2018; Xue et al., 2019; Forster et al., 2021), being based on an assumption that strongly interacted proteins are more likely to be associated with similar function (Altshuler et al., 2000). 2) Semi-supervised learning: most of the above methods with multiple PPI networks have two steps: they first derive low-dimensional embeddings for each protein from PPI networks, and then plug them into off-the-shelf machine learning models to obtain the prediction results. Obviously, generating protein representations in the first step is totally unsupervised learning. That is, only protein features are used and known GO annotations are not incorporated into the embeddings, which does not fully use all available information, resulting in limited prediction performance. On the other hand, graph-based semi-supervised learning methods (Mostafavi et al., 2008; Mostafavi and Morris, 2010; Gligorijevic et al., 2018) derive functional annotations of a protein from its annotated neighboring proteins in the networks, which only consider the local topology and ignore the high-order connectivity. 3) Deep graph neural networks: in deep learning, currently graph convolutional network (GCN) has been developed to capture the high-order connectivity in a given graph (Kipf and Welling, 2017), achieving successful results in various applications, such as prioritizing polypharmacy side effects (Zitnik et al., 2018), drug repurposing (Wang et al., 2020), and disease gene prioritization (Han et al., 2019). Additionally, GCN has been applied to protein function prediction to better exploiting protein structure (Gligorijevic et al., 2021; Swenson et al., 2020). Recently, BIONIC (Forster et al., 2021) applied GCN on PPI networks to GO prediction, which only uses a shallow architecture in an unsupervised manner.

Based on the above features of the latest term-centric GO prediction methods, we develop HPOND Nets, a systematic deep GCN-based approach with multiple input networks for HPO prediction. Methodologically HPOND Nets has the following two advantages: 1) Deep GCN: GCN learns the ‘representation’ of each node through the edge connectivity, to capture not only local (low-order) but also long-range (high-order) interactions among nodes, by repeating updating (learning) the representations through a large number of neural layers. This likes to cause a problem so-called ‘over-smoothing’, in which representations of entire nodes become too similar (Li et al., 2020). Because of this problem, shallow GCNs, which are trained by using only a few neural layers, have been often used and are eventually unable to extract high-order information in a given graph good enough. Recently two simple strategies of changing the manner of updating the representation of each node are proposed, both being based on the idea of constraining the representation update and demonstrated to be useful (Chen et al., 2020). HPOND Nets incorporates these two ideas, which allows to design a deep GCN: i) Initial residual connection: not only the information from neighboring nodes but also the initial representation itself are used to update. In other words, the initial representation is (partly) kept and the representation is updated partially by neighboring information every time. ii) Identity mapping: an identity matrix is added to the weight matrix so in each aggregation step only a part of information will be transformed and the rest is retained, thus low-order topological information can be passed to the higher layers. 2) Side-by-side GCN architecture: in HPOND Nets, each PPI network is fed into a deep GCN, which means that multiple networks are trained by multiple deep GCNs independently, and then finally the outputs of the multiple GCNs are combined for prediction.

In this work, we address the term-centric prediction of HPO annotations. Since current HPO prediction methods are mostly in the protein-centric manner, in our experiments, we compared HPOND Nets with latest term-centric GO prediction methods under both cross-validation and temporal validation. The experimental results show the performance advantage of HPOND Nets over all compared methods in all metrics used. In particular, HPOND Nets outperforms all other methods in area under the precision-recall curve (AUROC) for all cases, except only one. We further perform various ablation studies to examine the significance
of each component or parameter settings of HPOND Nets, revealing the importance of initial residual connection (out of the two solutions for over-smoothing), STRING (out of the three input networks) and adjustment weight between positive and negative samples. Finally, we successfully find latest supporting evidence in the biomedical literature for several false positives, suggesting great potentiality of HPOND Nets. Overall, the deep graph neural network architecture of HPOND Nets has been demonstrated to be effective for predicting protein-HPO term associations from multiple protein interaction networks.

2 Related work

Since GO prediction is an analogous setting of HPO prediction, we review existing methods of term-centric protein function prediction, which can be summarized as two categories: i) unsupervised learning methods and ii) semi-supervised learning methods.

Unsupervised learning. Most of proposed approaches were in an unsupervised manner. They first derived low-dimensional embeddings for each protein, and then plugged them into off-the-shelf machine learning models to obtain the prediction results. Mashup (Cho et al., 2016) characterized the topological structure by analyzing the diffusion states in each network, but suffered from high computational complexity when it came to large-scale networks. As a deep learning model, deepNF (Gligorijevic et al., 2018) used a multimodal deep auto-encoders to extract high-level features from heterogeneous networks. Differing from deepNF which did not take into account the correlation between multiple networks, DeepMNE (Xue et al., 2019) designed a communication mechanism that allowed information to flow across branches, but which made the model more sensitive to the noise. Both deepNF and DeepMNE did not incorporate network structures in the training stage. In contrast, BIONIC (Forster et al., 2021) adopted prevalent graph convolutional networks to integrate multiple biological networks. However, the over-smoothing problem would occur when chaining multiple GCN layers in BIONIC. The significant deficiency of unsupervised models was inability of considering known annotations in learning, by which the resultant embeddings are too general to be used for GO prediction.

Semi-supervised learning. There were two major methods, RANKS (Valenti et al., 2016) and GeneMANIA (Mostafavi et al., 2008; Mostafavi and Morris, 2010). RANKS was a general semi-supervised algorithmic scheme for solving node label ranking problems in biological networks. It adopted both global and local learning, corresponding to kernelization and guilt-by-association score function, respectively. However, RANKS took as input only a single network, while GeneMANIA (Mostafavi et al., 2008; Mostafavi and Morris, 2010) integrated multiple networks in a linear regression fashion, which allowed to learn the weights over multiple networks, by which the impact of potential noise in a single network could be diminished. Nonetheless, GeneMANIA may suffer from information loss incurred when combining multi-networks to a single network (Karasuyama and Mamitsuka, 2013).

3 Methods

3.1 Problem formulation

Let $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$ be the set of HPO terms, where $n$ represents the total number of terms. We assume a collection of $m$ proteins $\mathcal{P} = \{p_1, \ldots, p_i, \ldots, p_m\}$, and $r$ protein-protein interaction networks of them which are expressed as adjacency matrices $\{A_i\}_{i=1}^r$ where $A_i \in \mathbb{R}^{m \times m}$. The first $l$ proteins have been annotated, where the HPO annotations of the $i$-th protein $p_i$ are denoted by $y_i = (y_{i,1}, y_{i,2}, \ldots, y_{i,n})$ and each $y_{i,j} \in \{0, 1\}$ indicates whether $p_i$ is annotated by term $t_j$ or not. Our task is by using the known annotations $\{(y_i, t_j)^i\}_{j=1}^l$ and multiple protein-protein association networks $\{A_i\}_{i=1}^r$ to predict the annotations $\{y_{i}^{-1, m}\}$ of the remaining $m-l$ proteins. We formulate this task as the node label ranking problem to prioritize candidate proteins for each HPO term, or node classification problem. It can be viewed as multiple semi-supervised flat classification problems on biological network, because both labeled and unlabeled proteins are seen in the PPI networks in the training phase, but only the annotations of labeled training proteins are available.

3.2 Model architecture

In this section, we introduce our framework for predicting associated proteins for each HPO term from multiple interaction networks. Fig. 2 shows the architecture of HPOND Nets which consists of three steps: i) pre-processing step that generates initial feature vector, ii) encoding step with deep GCNs for refining protein representation; iii) output step that produces predictive scores for each HPO term.

3.2.1 Pre-processing: generates initial node representation

Protein-Protein Interactions (PPIs) refer to specific contacts between two or more protein molecules (Nooren and Thornton, 2003). By assembling the PPIs data stored in PPI databases, researchers can construct the protein-protein interaction networks, where the node is protein and the edge represents PPI strength. Current popular interaction databases not only contain experimentally verified “physical interactions” from the source publications, but also integrate “logical interactions” from multiple omics data. The score between two proteins reflects the strength of their association. In this work, we utilize three PPI networks including STRING (Szklarczyk et al., 2019), GeneMANIA-Net (Feaata et al., 2018) and HumanNet (Hwang et al., 2019). Specifically, the associations in the STRING are derived from high-throughput experimental data, from the mining of biological databases and literature, and from predictions based on genomic context analysis. GeneMANIA-Net integrates various association data sources like physical and genetic interactions, co-expression, co-localization, pathways, and shared protein domains. HumanNet is constructed by an integration of multiple types of data including not only physical and genetic interactions, literature, co-expression, and domains, but also the transferred knowledge from model organisms.

Formally, STRING, GeneMANIA-Net, and HumanNet are represented by adjacency matrices $\{A_{GM}\}$ where $G = \text{STR, GM, HN}$. These matrices are then symmetrized and guilt-by-association score function, respectively. However, RANKS took as input only a single network, while GeneMANIA (Mostafavi et al., 2008; Mostafavi and Morris, 2010) integrated multiple networks in a linear regression fashion, which allowed to learn the weights over multiple networks, by which the impact of potential noise in a single network could be diminished. Nonetheless, GeneMANIA may suffer from information loss incurred when combining multi-networks to a single network (Karasuyama and Mamitsuka, 2013).

3.2.2 Encoding: refines representations by deep GCNs

The dimension of initial node representation is equal to the number of nodes in the network, which is too high to be readily used. In consequence, we first apply a single dense layer with activation to reduce the dimensionality. Note that the Batch Normalization (Ioffe and Szegedy, 2015) is conducted for standardization. We denote the generated initial representation for the subsequent GCN layers, we adopt the strategy proposed by Cao et al. (2016). Specifically, for the normalized adjacency matrix $\hat{A}_{GM}$ of PPI network $G$, we compute the $(i, j)$-th entry in positive pointwise mutual information (PPMI) matrix as $X_{GM,j} = \max \left(0, \log_2 \left( \frac{\hat{A}_{GM,j} \sum_{i} \sum_{k} \hat{A}_{GM,k} \hat{A}_{GM,i}}{\sum_{i} \hat{A}_{GM,j} \sum_{k} \hat{A}_{GM,k}} \right) \right)$, and use the i-th row of it as the initial feature vector for protein $p_i$. Based on a widely accepted assumption that the strongly interacted proteins are more likely to be associated with similar phenotypes (Oh et al., 2008).
Fig. 2. Overview of HPODNets. The model consists of three steps: (i) pre-processing: the model takes as input three kinds of PPI networks including STRING, GeneMANIA-Net, and HumanNet. The positive pointwise mutual information (PPMI) matrices are calculated separately and served as proteins’ initial features; (ii) encoding: high-dimensional PPMI matrices are then fed into a dense layer to reduce the dimensionality. Straight afterwards, eight consecutive GCN blocks are built to extract low-order and high-order topological information from PPI networks and update the node representation; (iii) output: concatenating generated latent representations of each branch and applying a dense layer to fuse heterogeneous information, we finally obtain the compact embeddings. The decoder takes these embeddings to produce prediction scores for each HPO term.

Fig. 3. Illustration of GCN block. The components inside are arranged by Batch Normalization → Activation → Dropout. The i-th GCN block receives the output of the previous block $H^{(i-1)}_G$ and the initial representation $H^{(0)}_G$ as input, and outputs the updated representation $H^{(i)}_G$.

2006, Goh et al., 2007), we can build upon message-passing operation to form embedding propagation between the interacted proteins. Kipf and Welling (2017) proposed the classical graph convolutional layer as follows:

$$H^{(1)}_G = \sigma \left( \tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} W (H^{(0)}_G) \right)$$

where $\tilde{P}_G = \tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2}$ and $\tilde{A} = \tilde{A}_G + \mathbf{I}$, $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$, $W(\cdot) \in \mathbb{R}^{d(1) \times d(0)}$ is the trainable weight matrix to map the $d(0)$-dimensional input vector to $d(1)$-dimensional output vector. $\sigma(\cdot)$ is a non-linear activation function. The idea of above GCN layer is to learn a transformation function to generate the new representation of a node by aggregating its own features and its neighbors’ features considering the edge weights.

With the representations augmented by first-order message propagation, we can stack more such GCN layers to explore the high-order connectivity information. However, chaining multiple vanilla GCN layers tends to degrade the performance due to the intractable over-smoothing phenomenon. That is, as the number of layers increases, the representations of the nodes are inclined to converge to a certain similar value and thus become indistinguishable (Li et al., 2018). In addition, deeper neural network often suffers gradient vanishing or exploding that brings the difficulty to the model training (Hannin, 2018). ResNet (He et al., 2016) solved similar problem and has achieved great success in computer vision by adding skip connection between two consecutive layers in the deep convolutional network. But unfortunately, skip connections in the GCN models can merely slow down the over-smoothing (Kipf and Welling, 2017). As a result, most current GCN models only stay in the shallow architecture, which are unable to capture high-order connectivity. HPODNets addresses this issue by adopting two simple strategies, which both are recently proposed, i.e. (i) initial residual connection and (ii) identity mapping.

Initial residual connection. The idea is to ensure that the final representation can retain a fraction of initial feature even if we stack many layers. The message passing operation can be formulated as follows (Klicpera et al., 2019).

$$H^{(i)}_G = \sigma \left( (1 - \alpha) \tilde{P}_G H^{(i-1)}_G + \alpha H^{(0)}_G \right),$$

where $\alpha$ is a balance factor between the current representation and the initial representation. Note that due to the introduction of initial feature, the dimension of the output in the i-th layer $H^{(i)}_G$ should be the same as $H^{(0)}_G$, i.e. $d^{(i)} = d^{(0)}$.

Identity mapping. Although initial residual connection can partially relieve the over-smoothing, the performance still degrades as the model goes deeper (Klicpera et al., 2019). Thus, an identity matrix is added to the weight matrix, and the message passing operation can be formulated as the following equation which is called the GCNII layer (Chen et al., 2020).

$$H^{(i)}_G = \sigma \left( (1 - \alpha) \tilde{P}_G H^{(i-1)}_G + \alpha H^{(0)}_G \right) \left( 1 - \beta_i \mathbf{I} + \beta_i W(\cdot) \right),$$

where $\beta_i = \log(\frac{d_i}{d_i + 1}) \approx \frac{d_i}{d_i + 1}$ and $\theta$ is a hyperparameter. The advantage of adding an identity matrix is that in each aggregation step only a part of information will be transformed and the rest is retained, thus the low-order connectivity information can be passed more to the higher layers. It is noteworthy that the factor $\beta_i$ is inversely proportional to the depth $i$, which means that the influence of identity mapping is higher in the deeper layers, or in other words, less high-order information will be incorporated. It is intuitively reasonable because in most cases multi-hop neighbors are less likely to share the same phenotypic abnormality as direct neighbors.

HPODNets stacks eight GCNII layers to capture both low-order and high-order information from the interaction networks (see Fig. 3-4B for the reason of the number of layers). In order to prevent over-fitting and make the model training more stable, HPODNets adopts Dropout (Srivastava et al., 2014) and Batch Normalization (Ioffe and Szegedy, 2015) after every GCNII layer. Li et al. (2020) found that the order of these components had quite an impact on the final performance. On their suggestion, HPODNets builds a GCN block by arranging these components in this order: GCNII → Batch Normalization → Activation → Dropout, as illustrated in Fig. 3. Here, LeakyReLU (Maas et al., 2013) is selected as the activation function.

3.2.3 Output: gives prediction scores for each HPO term

After going through L GCN layers, we obtain the embeddings of proteins from each PPI network, namely $H^{(L)}_{HG}$, $H^{(L)}_{HGC}$ and $H^{(L)}_{HH}$. These embeddings are then concatenated together to constitute the combined
embedding:

\[ \mathbf{H} = \mathbf{H}^{(L)}_{E \rightarrow STR} \| \mathbf{H}^{(L)}_{G \rightarrow GM} \| \mathbf{H}^{(L)}_{H \rightarrow HN}, \]

where \( \| \) is the concatenation operation. Then we use a dense layer with LeakyReLU activation to reduce the dimension as well as perform information fusion, and finally obtain the final embedding \( \mathbf{E} \).

After passing the above embedding through the following Dropout module, we finally distribute them to each output neuron and produce the predictive score for the corresponding HPO term:

\[ \hat{y}_{i,t} = \sigma \left( \mathbf{e}_i \cdot \mathbf{\theta}_t \right) = \frac{1}{1 + \exp \left( -\mathbf{e}_i \cdot \mathbf{\theta}_t \right)}, \]

where \( \mathbf{e}_i \) refers to the embedding of protein \( p_i \) being the \( i \)-th row of \( \mathbf{E} \), \( \mathbf{\theta}_t \) is the weight for the \( j \)-th HPO term in logistic regression, and \( \cdot \) denotes dot product.

### 3.3 Model training

As we obtain outputs from per-term predictors, we calculate the binary cross-entropy loss for each term and sum them up. Formally, the objective function to be optimized is as follows,

\[ \mathcal{L} = - \sum_{t=1}^{n} \sum_{i=1}^{l} \gamma_t \left( y_{i,t} \log (\hat{y}_{i,t}) + (1 - y_{i,t}) \log (1 - \hat{y}_{i,t}) \right), \]

where \( \gamma_t \) is the adjustment weight to the positive samples of term \( t \) to mitigate the impact of drastic class imbalance, which is assigned with

\[ \gamma_t = \frac{m^+_t}{m^-_t}, \]

where \( m^+_t \) and \( m^-_t \) refer to the number of positive and negative samples in the training set in terms of \( t \), respectively.

### 4 Experiments

#### 4.1 Data

To measure the performance of HPODNets, we adopt two evaluation strategies: a) cross-validation and b) temporal validation.

##### 4.1.1 Data preparation for cross-validation

We downloaded human gene-HPO term associations released by 2020-03-27 from HPO project website (https://hpo.jax.org/). Then the genes in raw HPO annotations were mapped into proteins using the UniProt ID mapping tool (https://www.uniprot.org/mapping/). For the sake of data quality, we filtered out proteins that were not stored in Swiss-Prot. The true-path rule was applied to propagate annotations. In this work, only HPO terms belonging to the biggest sub-ontology, Phenotypic Abnormality (PA), remained. Besides, terms with no more than 10 annotated proteins were removed (Jiang et al., 2016). Due to the incompleteness of HPO annotations of newly added proteins (Liu et al., 2020), we deleted the proteins added after 2018-03-19 to ensure the accuracy of annotations. As Table 1 shown, after processing, the dataset consisted of 3,652 proteins and 3,704 HPO terms. Akin to previous methods (Mostafavi and Morris, 2010; Cho et al., 2016; Gligorijevic et al., 2018; Xue et al., 2019), we further organized the terms into four groups according to the number of annotated proteins: 11-30, 31-100, 101-300 and \( \geq \)301. For the PPI networks, we downloaded STRING v11 released by 2019-01-19, GeneMANIA-Net released by 2017-03-14, and HumanNet v2 released by 2018-11. Specifically, STRING was obtained from the file 9606.protein.links.v11.0.txt.gz in https://string-db.org/cgi/download. GeneMANIA-Net was built from the file COMBINED.DEFAULT_NETWORKS.BP_COMBINING.txt in http://genemania.org/data/current/Homo_sapiens. COMBINED/, and HumanNet was got from the fully extended functional gene network HumanNet-XN in https://www.inetbio.org/humannet/download.php. The networks used in this work were constructed from the downloaded data directly without any processing. The details of these PPI networks are provided in Supplementary Materials. The performance was measured by 5-fold cross-validation, and the hyperparameters were tuned on a nested 5-fold cross-validation.

##### 4.1.2 Data preparation for temporal validation

In the temporal validation, we adopted a similar strategy as proposed in the CAFA challenge (Jiang et al., 2016). The training set comprised HPO annotations released by 2019-02-12, and the test set comprised the new annotations added from 2019-02-12 to 2020-10-12. HPO annotations in the test set were aligned to 2019-02-12 version and thus the newly created HPO terms were discarded. We also divided terms into four groups based on its frequency. The statistics of the dataset are shown in Table 2, and the distributions are given in Fig. 5.3. Note that the average number of annotations per protein in the training and test set vary greatly (Liu et al., 2020). The versions of PPI networks were the same as those in cross-validation. There was no potential information leakage because the released dates were all earlier than 2019-02-12.

#### 4.2 Evaluation metrics

Two evaluation metrics were chosen: (i) area under the precision-recall curve (AUPR); (ii) F1 score (F1). Considering the imbalance of HPO terms, we reported performance of each group by two average modes: (i) Macro-averaged (M): calculate metrics for each term, and find their unweighted mean; (ii) Micro-averaged (m): vectorize the predicted score matrix for each protein-HPO term pair and calculate metrics based on the resulting vector. The difference is that the macro-averaged metric is term-centric, while the micro-averaged metric is pairwise. These two evaluation strategies have been widely adopted in the previous literature (Cho et al., 2016; Gligorijevic et al., 2018; Xue et al., 2019) as their only options.

#### 4.3 Competing methods and implementation details

We compared HPODNets with seven state-of-the-art GO prediction methods including deepNF (Gligorijevic et al., 2018), DeepMNE (Xue et al., 2019), BIONIC (Forster et al., 2021), LP (Zhou et al., 2003), RANKS (Valentini et al., 2016), Mashup (Cho et al., 2016), and GeneMANIA (Mostafavi et al., 2008; Mostafavi and Morris, 2010). For
LP and RANKS which takes a single network as input, we adopted the pre-processing strategy in (Cho et al., 2016) to integrate the networks by \( \bar{A}_{ij} = 1 - \prod_{k=1}^5 (1 - A_{ij}^{(k)}) \), and for the others, we used the same PPI networks as ours. Since deepNF, DeepMNE, Mashup, and BIONIC are unsupervised methods, we trained multiple SVM classifiers for each HPO term using the generated embeddings except for BIONIC which opted for Logistic Regression according to the original papers. We tuned parameters of LP and RANKS by nested 5-fold cross-validation, and set hyperparameters of others except GeneMANIA according to the guideline in their original papers, where GeneMANIA has no hyperparameters. The details of parameter settings are provided in Supplementary Materials.

For our method, we set the dimension of embeddings \( d^{(1)} = 500 \), the factors in GCNII \( \alpha = 0.5 \) and \( \theta = 0.5 \), and the dropout rate to 0.5. We used the Adam optimizer (Kingma and Ba, 2015) with epoch of 300 and learning rate of 0.001 to train the model. The algorithm was implemented using PyTorch and PyTorch Geometric. To align the nodes of multiple PPI networks, we took the union of protein sets of input networks. If a protein did not appear in one PPI network but appeared in other networks, we added a node in this network and treated it as isolated vertex.

4.4 Results and discussion of cross-validation

4.4.1 Performance comparison

Tables 3 and 4 show macro- and micro-averaged results, respectively. HPODNets achieves the best in all 16 cases except only one case, being followed by GeneMANIA in M-AUPR and M-F1. Although GeneMANIA is a semi-supervised learning approach with multiple networks, HPODNets achieves gains ranging from 6.3% to 7.7% in terms of M-AUPR, which demonstrates the power of deep GCN in exploiting the network structure and our capability of integrating multiple PPI networks. Furthermore, HPODNets surpasses three deep learning-based unsupervised methods (deepNF, DeepMNE and BIONIC) by a large margin, indicating the effectiveness of end-to-end fashion of our model.

Interestingly, we notice that traditional machine learning methods basically outperform the deep learning methods, which highlights the difficulty in designing effective neural network architecture. Additionally, we calculate the average AUPR for each of the 25 categories under the PA sub-ontology (Fig. S4). HPODNets defeats the competing methods on almost all groups. We further evaluate HPODNets under protein-centric metric, and the results are listed in Table S3. These results indicate that protein-centric prediction is different from term-centric prediction, whereas the latter one is our focus in this work. Although \( F_{\text{max}} \) of HPODNets is not the best, it still outperforms GeneMANIA in protein-centric prediction.

4.4.2 Effect of two strategies for alleviating over-smoothing

Fig. 4(A) shows M-AUPR of HPODNets, obtained by removing initial residual connection or/and identity mapping from the GCNII layer. The full model consistently outperforms other variants significantly. For all four groups, initial residual connection always contributes more to the performance than identity mapping. For example, for group 11-30, M-AUPR is reduced from 0.3073 to 0.1935 by removing initial residual connection but 0.2801 by identity mapping. Besides, removing any part of GCNII will result in degrading performance, suggesting that all connections but 0.2801 by identity mapping. Besides, removing any part of GCNII will result in degrading performance, suggesting that all connections are important. We further evaluate HPODNets using multiple numbers of layers. For all groups, HPODNets achieves the best performance when the depth is 8. Moreover, as the model becomes deeper, the performance keeps stable. For example, for group 31-100, M-AUPR is 0.3205 (2), 0.3259 (4), 0.3302 (8), 0.3297 (16), 0.3266 (32) and 0.3273 (64). It shows that the sophisticated design of GCNII layer can help alleviate the over-smoothing. Moreover, we analyze the change of M-AUPR with respect to the number of GCN layers without using the initial residual and/or identity mapping. As shown in Fig. 5, the initial residual contributes the most between two strategies: the performance drastically degrades as the model deepens.

4.4.3 Effect of the depth of GCN

Fig. 4(B) summarizes the results with various numbers of layers. For all groups, HPODNets achieves the best performance when the depth is 8. Moreover, as the model becomes deeper, the performance keeps stable. For example, for group 31-100, M-AUPR is 0.3205 (2), 0.3259 (4), 0.3302 (8), 0.3297 (16), 0.3266 (32) and 0.3273 (64). It shows that the sophisticated design of GCNII layer can help alleviate the over-smoothing. Moreover, we analyze the change of M-AUPR with respect to the number of GCN layers without using the initial residual and/or identity mapping. As shown in Fig. 5, the initial residual contributes the most between two strategies: the performance drastically degrades as the model deepens.
Fig. 5. Analysis of GCNII layer. M-AUPRs of each group are reported. (A) Ablation study that removes initial residual connection and/or identity mapping from GCNII layer. (B) Effect of the number of layers on model performance. The gray dotted line represents the number of layers selected in this paper. Results are summarized over five trials (standard deviation shown as error bars), and asterisks represent where full mode’s improvement over other variants is significant (one-tailed t-test, *: P < 0.05, **: P < 0.01).

Fig. 6. Performance degradation without adjustment weights in binary cross-entropy loss function. The percentage of change is relative to the original model. Results are summarized over five trials (standard deviation shown as error bars).

Besides, although the performance remains stable as the number of layers increases after removing the identity mapping, the overall performance is still inferior to that of GCNII in the full mode. Such results again demonstrate the necessity of initial residual and identity mapping.

4.4.4 Effect of multiple interaction networks

We also apply HPODNets to individual networks without integration. From Fig. 5(A), we note that the integrated mode is better than the model with individual network, which demonstrates the necessity of combining multiple networks. Besides, we observe that the model taking STRING only achieves the M-AUPR closest to integrated one, highlighting that STRING is the most informative network. For example, for group 101-300, M-AUPR with all three networks is 0.3778, while 0.3374, 0.1708 and 0.2351 by STRING, GeneMANIA and HumanNet, respectively. Fig. 5(B) shows the performance decrease after removing one of PPI networks. The reduction by removing STRING is the largest among three networks for all groups, which again shows the significance of STRING. For example, for 101-300, the reductions are 25.01%, 1.03% and 4.01% by STRING, GeneMANIA and HumanNet, respectively. From Fig. S2, we notice that three networks have less direct overlap in PPIs than might be expected given that they use similar data types (see Supplementary Materials and (Huang et al., 2018)). The impressive performance of STRING might be contributed to its effective data integration methodology. Such results demonstrate that multiple networks can be complements to each other and network fusion can help boost the final performance.

4.4.5 Effect of adjustment weights in loss function

Fig. 6 shows the performance decrease when removing adjustment weights $\gamma_t$ from the loss function. The performance gets worse especially in the low-frequency groups, where the samples suffer from heavier class-imbalance. For example, for group 11-30, the performance reduction (percentage) was 16.7957 (M-AUPR) and 13.7894 (M-F1), while for group $\geq 301$, 2.4684 (M-AUPR) and 2.3446 (M-F1). It indicates the effectiveness of adjustment weights for improving the accuracy on imbalanced data.

4.4.6 Performance analysis regarding splitting with sequence identity thresholds

Two proteins with very different sequences may have similar structure and maintain similar functions as a consequence of the evolution (Creighton, 1993). Quantifying these evolutionary biases is quite important for preventing undesired information leakage between data splits (Jones, 2019; Rao et al., 2019). Typically, the sequence identity is adopted which measures the percentage of exact amino acid matches between two different proteins (Rost, 1999). To further examine the generalizability of HPODNets, we filtered the training set at the 30% sequence identity threshold. Specifically, we still conducted 5-fold cross-validation by splitting the protein list into the training set (80%) and the test set (20%). Then, for each protein in the test set, we removed all the proteins from the training set that have greater than 30% sequence identity with the corresponding test protein based on BLAST (Altschul et al., 1990) alignments. As a result, there are no two proteins in the training and test set have large similarity, which avoids potential information leakage. After processing, the test split contains 20% of the total proteins, while the training split only remains 53.6% of the total proteins on average. The results are listed in Tables S4 and S5. Despite the stringent experimental setting, HPODNets still achieves the best performance in all 16 cases. The results demonstrate the generalizability of our model to the unfamiliar proteins.

4.5 Results and discussion of temporal validation

4.5.1 Performance comparison

Tables 5 and 6 report the performance under macro- and micro-averaged metrics, respectively. HPODNets achieves clearly the best performance in all 16 cases, except only two cases. These two cases are all obtained by the group $\geq 301$, implying that HPODNets is more powerful for low-frequency terms. The results in temporal validation are entirely worse than those in cross-validation, which may be caused by the incompleteness of HPO annotations. In fact, as shown in Table 2, the average number of annotations per protein of the test set is fairly lower than that of training set by over 30%. Thus, the performance is probably under-estimated, as pointed out by Liu et al. (2020).
Table 5. Temporal validation performance under macro-averaged metrics

<table>
<thead>
<tr>
<th>Method</th>
<th>11-30</th>
<th>31-100</th>
<th>101-300</th>
<th>≥301</th>
</tr>
</thead>
<tbody>
<tr>
<td>deepNF</td>
<td>0.0769</td>
<td>0.1172</td>
<td>0.0992</td>
<td>0.2364</td>
</tr>
<tr>
<td>DeepMNE</td>
<td>0.0727</td>
<td>0.1229</td>
<td>0.1217</td>
<td>0.3108</td>
</tr>
<tr>
<td>BIONIC</td>
<td>0.0685</td>
<td>0.1040</td>
<td>0.0501</td>
<td>0.3099</td>
</tr>
<tr>
<td>LP</td>
<td>0.0611</td>
<td>0.1044</td>
<td>0.1105</td>
<td>0.2215</td>
</tr>
<tr>
<td>RANKS</td>
<td>0.0567</td>
<td>0.0958</td>
<td>0.0819</td>
<td>0.2152</td>
</tr>
<tr>
<td>Mashup</td>
<td>0.0679</td>
<td>0.1016</td>
<td>0.0984</td>
<td>0.3249</td>
</tr>
<tr>
<td>GeneMANIA</td>
<td>0.0789</td>
<td>0.1143</td>
<td>0.0813</td>
<td>0.2671</td>
</tr>
<tr>
<td>HPODNets</td>
<td>0.0864</td>
<td>0.1273</td>
<td>0.1196</td>
<td>0.3442</td>
</tr>
</tbody>
</table>

Notes: The boldface items in the table represent the best performance, and the runner-ups are underlined.

Table 6. Temporal validation performance under micro-averaged metrics

<table>
<thead>
<tr>
<th>Method</th>
<th>11-30</th>
<th>31-100</th>
<th>101-300</th>
<th>≥301</th>
</tr>
</thead>
<tbody>
<tr>
<td>deepNF</td>
<td>0.0274</td>
<td>0.0574</td>
<td>0.0334</td>
<td>0.3485</td>
</tr>
<tr>
<td>DeepMNE</td>
<td>0.0207</td>
<td>0.0456</td>
<td>0.0219</td>
<td>0.3936</td>
</tr>
<tr>
<td>BIONIC</td>
<td>0.0242</td>
<td>0.0456</td>
<td>0.0300</td>
<td>0.4062</td>
</tr>
<tr>
<td>LP</td>
<td>0.0124</td>
<td>0.0478</td>
<td>0.0196</td>
<td>0.3874</td>
</tr>
<tr>
<td>RANKS</td>
<td>0.0134</td>
<td>0.0542</td>
<td>0.0202</td>
<td>0.3952</td>
</tr>
<tr>
<td>Mashup</td>
<td>0.0184</td>
<td>0.0419</td>
<td>0.0136</td>
<td>0.3560</td>
</tr>
<tr>
<td>GeneMANIA</td>
<td>0.0110</td>
<td>0.0475</td>
<td>0.0221</td>
<td>0.3920</td>
</tr>
<tr>
<td>HPODNets</td>
<td>0.0359</td>
<td>0.0672</td>
<td>0.0471</td>
<td>0.3927</td>
</tr>
</tbody>
</table>

Notes: The boldface items in the table represent the best performance, and the runner-ups are underlined.

4.5.2 Evaluation using the CAFA2 challenge dataset
To validate the performance advantages of HPODNets against HPO prediction methods, we conduct the evaluation using the CAFA2 challenge dataset (Jiang et al., 2016) and compare our results with the top performing participating methods in CAFA2, and famous HPO predicting methods proposed after CAFA2, such as HPO2GO (Do˘gan, 2018), DeepPheno (Kulmanov and Hoehndorf, 2020), and HPOLabeler (Liu et al., 2020). The results are reported in Fig. S6. HPODNets wins the first place under the term-centric metric AUC (area under the receiver-operating characteristic curve), and defeats six out of eight best CAFA2 participators under the protein-centric metric Fmax. In particular, compared with the runner-up DeepPheno, HPODNets increases AUC by 8.97%.

4.5.3 Case study on genes predicted related to an HPO term with supporting literature
As mentioned before, the current HPO annotations are still incomplete. To further assess the quality of predictions and check the false positives, we retrieve the latest literature for the top 10 new hits associated with Pneumonia (HP:0002090) and compare our results with the top performing methods. We find supporting evidence for 7 out of them. Specifically, we perform online searches using the pair of protein/gene name and phenotype name as the query for the search engine, which results in a list of publications. Then we manually check if those papers contain supporting evidence that suggests the positive association between the queried protein and the phenotype. Table 7 lists the details. For the first prediction, Monnetter et al. (2020) reported that interferon-7 (IL7) can help restore lymphocyte count and improve immune functions in critically ill COVID-19 patient. Sa Ribero et al. (2020) illustrated that a weak production of IFN-I to IFNAR leads to the activation of Janus kinase 1 (JAK1), which in turn phosphorylate the signal transducer and activator of transcription (STATs). Phosphorylated STATs heterodimerize and associate with the DNA binding protein IRF9 to form a complex and finally influences the establishment of antiviral state. Schaefer et al. (2020) reported that ciliated cells of two patients showing SARS-CoV-2 maculopapulosis positivity consistently showed the absence of FOXJ1 expression. Not only COVID-19, our method also finds genes/proteins related to other pneumonia. Zhang et al. (2020) provided evidence that Pneumocystis infection promotes PD-1 deficiency enhances the phagocytic function of macrophages and the pulmonary T-helper cell type 1 (Th1)/Th17 response, which might contribute to Pneumocystis clearance. Nepal et al. (2019) demonstrated that STAT6 induces expression of Gas6 in macrophages to clear apoptotic neutrophils and thereby promotes resolution of inflammatory lung injury. The above remarkable results suggest that our approach has great potential to uncover new candidate genes/proteins associated with abnormal phenotypes.

4.5.4 Case study on HPO terms predicted related to a novel protein with supporting literature
We further use the HPO annotations of 4,424 proteins released in October 2020 as the training set and apply HPODNets to predict the phenotype annotations of 15,033 proteins stored in STRING, GeneMANIA-Net and HumanNet databases that have not been annotated. According to...
Table 9. Novel disease-gene associations found by HPODNets by bridging between the protein-HPO term predictions and known disease-HPO term annotations. Top 5 confirmed predictions that are newly added to the latest database are shown below.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Disease ID</th>
<th>Disease name</th>
<th>HPO term ID</th>
<th>HPO term name</th>
<th>Protein</th>
<th>Gene</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HP:0011109</td>
<td>Chronic sinusitis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.999952</td>
</tr>
<tr>
<td>12</td>
<td>HP:0011539</td>
<td>Atrial situs ambiguous</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.999499</td>
</tr>
<tr>
<td>13</td>
<td>HP:0011535</td>
<td>Abnormal atrial arrangement</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.999822</td>
</tr>
<tr>
<td>17</td>
<td>HP:0000433</td>
<td>Abnormality of the nasal mucosa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.999499</td>
</tr>
<tr>
<td>22</td>
<td>HP:0001748</td>
<td>Polysplenia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.999426</td>
</tr>
</tbody>
</table>

Note: ‘HPO term’ refers to the predicted HPO term annotation of corresponding protein by HPODNets.

the predictive score, all predicted protein-HPO term pairs are sorted in descending order, and the top 5 literature-supported results are listed in Table 8. Axonemal dynein light intermediate polypeptide 1 (UniProt ID: O14645) is the protein encoded by gene DNALI1 whose precise function is not known now. Höben et al. (2018) demonstrated that C11orf70 deficiency resulted in severe reduction or complete absence of DNALI1 from the ciliary axonemes, and all individuals with loss-of-function mutations in C11orf70 in their research showed classical primary ciliary dyskinesia (PCD) symptoms such as chronic sinusitis. Tarkar et al. (2013) found that DNALI1 was absent from ciliary axonemes in DYX1C1-mutant cells, and 16% individuals with biallelic DYX1C1 mutations had atrial situs ambiguous and 16% patients suffered from polysplenia. We note that Abnormal atrial arrangement (HP:0011535) is the parent of Atrial situs ambiguous (HP:0011539), thus DNALI1 also related to the abnormalities of atrial arrangement. In addition, HPODNets found the association between DNALI1 and nasal mucosa, which has been revealed by Peng et al. (2018) that the mRNA levels of DNALI1 were significantly reduced in patients with allergic nasal mucosa. These concrete cases affirm practical usefulness of our approach to identify relevant phenotypic abnormalities of novel human proteins.

4.5.5 Application to find disease-gene associations

Predicted protein-HPO term associations together with known HPO annotations of diseases can help to identify disease caused genes. Specifically, we downloaded disease-HPO term relationships released by 2019-02-12 and built up the prediction of disease-gene associations. Our prediction was then compared with the disease-gene associations added in the period from 2019-02-12 to 2020-10-12. Table 9 lists top 5 predictions that has been confirmed to be correct. Surprisingly, the highest ranked prediction is hit in the latest database. It is noteworthy that their predictive scores are extremely high and close, which demonstrates the ability of HPODNets to discover the genetic causes of human disorders.

5 Conclusion

In this paper, we presented HPODNets for predicting human protein-HPO term associations from three types of PPI networks. The key features of HPODNets could be summarized into the following two aspects: 1) Deep graph convolutional network (GCN): differing from conventional shallow GCNs which usually caused over-smoothing, HPODNets adopted eight GCN layers with initial residual connection and identity mapping, which allowed to capture not only local (low-order) information but also long-range (high-order) information from input networks. This feature provided HPODNets with higher predictive performance than existing approaches. 2) Side-by-side GCN architecture: HPODNets learned representation from multiple PPI networks separately, and then performs information fusion to project them into an unified embedding space. Extensive experiments including cross-validation and temporal validation revealed the clear advantage of HPODNets over existing, cutting-edge GO prediction methods. Moreover, the usefulness of HPODNets was further demonstrated by a lot of case studies to predict unknown protein-HPO term associations. HPODNets has been developed for term-centric HPO prediction, which can be viewed as a set of large-scale flat classification problems with the input of multi-networks. Due to the generalizability of inputs and outputs, it will be interesting to transfer the methodology cultivated in this work to other domains in bioinformatics such as protein function prediction and drug repositioning.

Funding

S.Z. has been supported by National Natural Science Foundation of China (No. 61872094), Shanghai Municipal Science and Technology Major Project (No.2018SHZDZX01), ZJ Lab, and Shanghai Center for Brain Science and Brain-Inspired Technology. L.L. has been supported by the 111 Project (No. B18015), Shanghai Municipal Science and Technology Major Project (No. 2017SHZDZX01) and Information Technology Facility, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences. H.M. has been supported partially by Academy of Finland (315896), JST ACCEL (JPMJAC1503), NEXT KAKENHI (19H04169).


