PDGNet: Predicting Disease Genes Using a Deep Neural Network With Multi-View Features

Kuo Yang, Yi Zheng, Kezhi Lu, Kai Chang, Ning Wang, Zixin Shu, Jian Yu, Baoyan Liu, Zhuye Gao, and Xuezhong Zhou

Abstract—The knowledge of phenotype-genotype associations is crucial for the understanding of disease mechanisms. Numerous studies have focused on developing efficient and accurate computing approaches to predict disease genes. However, owing to the sparseness and complexity of medical data, developing an efficient deep neural network model to identify disease genes remains a huge challenge. Therefore, we develop a novel deep neural network model that fuses the multi-view features of phenotypes and genotypes to identify disease genes (termed PDGNet). Our model integrated the multi-view features of diseases and genes and leveraged the feedback information of training samples to optimize the parameters of deep neural network and obtain the deep vector features of diseases and genes. The evaluation experiments on a large data set indicated that PDGNet obtained higher performance than the state-of-the-art method (precision and recall improved by 9.55 and 9.63 percent). The analysis results for the candidate genes indicated that the predicted genes have strong functional homogeneity and dense interactions with known genes. We validated the top predicted genes of Parkinson's disease based on external curated data and published medical literatures, which indicated that the candidate genes have a huge potential to guide the selection of causal genes in the 'wet experiment'. The source codes and the data of PDGNet are available at https://github.com/yangkuoone/PDGNet.

Index Terms—Disease gene prediction, deep neural network, multi-view features

1 Introduction

With the developments of modern medicine, identification of disease genes has been a critical step to understand disease mechanisms [1], [2], [3], improve clinical diagnosis and therapy and achieve precision medicine [4] in final. Owing to a lot of labours and money that need to be invested for the causing gene selection in wet lab [1], a large number of computing-based approaches to predict disease genes [5], [6], [7], [8], [9] have been proposed. There are several types of typical prediction methods, e.g., network propagation-based methods (e.g., PRINCE [10], DADA [11] and pgWalk [12]), network feature-based methods [9], [13], [14], traditional classification-based methods [7], [15], [16], and network embedding-based methods [17], [18], [19]. For example, in previous work, we proposed a heterogeneous network embedding method HerGePred, which integrated heterogeneous biological associations to obtain embedding features of diseases and genes and obtained high performance on gene prediction. Li *et al*. [20] proposed a graph convolutional network-based disease gene prediction method with high performance.

In recent years, as an important research direction in machine learning area, deep learning have made breakthroughs in image classification [21], speech recognition [22] and natural language processing [23]. Many classic deep neural networks, e.g., convolutional neural networks (CNN) [24], long short term memory network (LSTM) [25] and generative adversarial networks (GAN) [26] have been proposed. Meanwhile, many related researchers have tried to solve medical prediction problems using deep neural networks, e.g., Decagon [27] for predicting polypharmacy side effects, DeepTACT [28] to predict chromatin contacts between regulatory elements, CNNLDA [29] for predicting the non-coding RNA genes of diseases and N2A-SVM [30] for identifying the candidate genes of Parkinson's disease.

Multi-view learning is a significant research direction for exploiting the complementarity among multiple modalities or multiple features all the time. As the focus of multi-view learning, multi-view neural network models are mainly applied to image and vision area, e.g., MVCNN [31] for 3D shape recognition, DDFD [32] for face detection and FCN-VGG [33] for 6D pose estimation in the beginning. Recently these methods are increasingly applied to biomedical filed [34] and obtain high performance. For example, Zhang *et al*. [35] proposed a multi-layer multi-view

- *K. Yang is with the Institute of Medical Intelligence, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China, and also with the BNRIST/Department of Automation, Tsinghua University, Beijing 10084, China. E-mail: yangkuo@bjtu.edu.cn.*
- *Y. Zheng, K. Lu, K. Chang, N. Wang, Z. Shu, and X. Zhou are with the Institute of Medical Intelligence, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China. E-mail: {18120459, lukezhi, changkai, 15120442, 18112032, xzzhou}@bjtu.edu.cn.*
- *J. Yu is with the Beijing Key Lab of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China. E-mail: jianyu@bjtu.edu.cn.*
- *B. Liu is with the Data Center of Traditional Chinese Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China. E-mail: liuby@mail.cintcm.ac.cn.*
- *Z. Gao is with Xiyuan Hospital, China Academy of China Academy of Chinese Medical Sciences, Beijing 100091, China. E-mail: zhuyegao@126.com.*

Manuscript received 21 Oct. 2019; revised 27 May 2020; accepted 10 June 2020. Date of publication 16 June 2020; date of current version 3 Feb. 2022. (Corresponding author: Xuezhong Zhou.)
Recommended for acceptance by Y. Wu.
Digital Object Identifier no. 10.1109/TCBB.2020.3002771

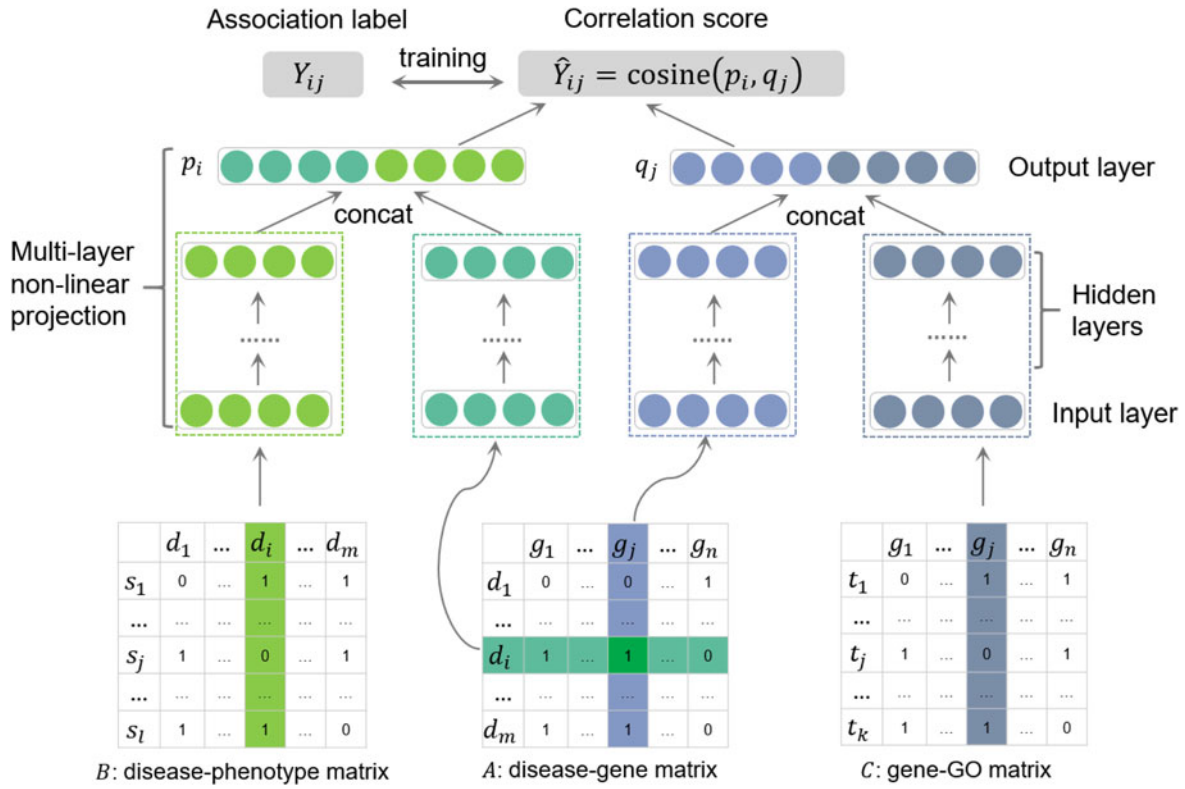

Fig. 1. *PDGNet overview.* PDGNet is a multi-view non-linear neural network to predict disease genes. First, the model extracted input features of disease and genes from the matrices of disease-gene, disease-phenotype, and gene-GO associations. Then four sub neural networks are constructed to learn vector representations of diseases and genes. We obtained two syncretic vectors for the disease p_i and gene q_j using 'vector concat'. The relevant scores between diseases and genes are measured by the cosine similarity of their vectors. Finally, the cross entropy between the true labels and relevant scores are measured as the feedback information to continuously optimize the deep model.

classification approach for Alzheimer's disease diagnosis. Ma *et al*. [36] have conducted drug similarity integration through attentive multi-view graph auto-encoders. Therefore, multi-view neural network models have huge advantages on automatically fusing multi-view features and learning deep features. However, owing to the sparseness and complexity of medical association data, developing an efficient deep neural network model based on multi-view features of phenotypes and genotypes to identify disease genes is still a huge challenge.

To address the above challenges, we propose a deep neural network model PDGNet to identify disease genes (Fig. 1). The key features of PDGNet are the following:

- PDGNet is a non-linear neural network model that integrated the multi-view information (i.e., known disease genes, disease phenotypes and gene annotations) to predict disease genes.
- Compared with network embedding-based prediction algorithms [19], [37] with two-step strategy (i.e., separate learning and prediction), PDGNet leveraged the priori association data as the feedback information to learn the deep features of diseases and genes and optimize the neural network at the same time in each step of training.
- The systematic experiments on a large data set of disease-gene associations indicated that the multi-view features of PDGNet is able to effectively improve the performance and obtained better performance than the state-of-the-art methods.

2 Materials and Methods

2.1 Dataset

Disease-Gene Associations. We collected 130,820 disease-gene associations (Figs.2 A and 2B) between 13,074 diseases and 8,947 genes from DisGeNet database [38], which integrated disease-gene associations from multiple databases, e.g., PsyGeNET [39], ClinVar [40], OrphaNet [41], and the GWAS Catalog [42]. The average number of causing genes per disease is ∼10, and the average number of related diseases per gene is ∼15. We conducted the cross-validation experiments based on this dataset of disease-gene associations. To conduct the external validation in case study section, we also used the disease-gene associations of MalaCards [43] that contains a large number of curated disease-gene associations and DISEASES [44] that offers a lot of literature-based evidences for disease-gene associations.

Disease-Phenotype Associations. We integrated the human disease-related phenotypes from HPO [45] and OrphaNet [41] databases. Finally, we obtained 99,087 disease-phenotype associations that include 5,424 diseases and 2,691 phenotypes (Fig. 2 C).

Gene Functional Annotations. We collected the Gene Ontology (GO) annotations from STRING 10 [46] database and obtained 218,337 annotation records with 18,584 genes and 14,204 GO terms (Fig. 2 D).

Protein-Protein Interactions. The data set of protein-protein interactions (PPI) are from Menche *et al*. [47] that integrated several high-quality databases (e.g., HPRD [48], IntAct [49] and PINA [50]). The PPI data includes 213,888 protein-protein interactions with 15,964 proteins.

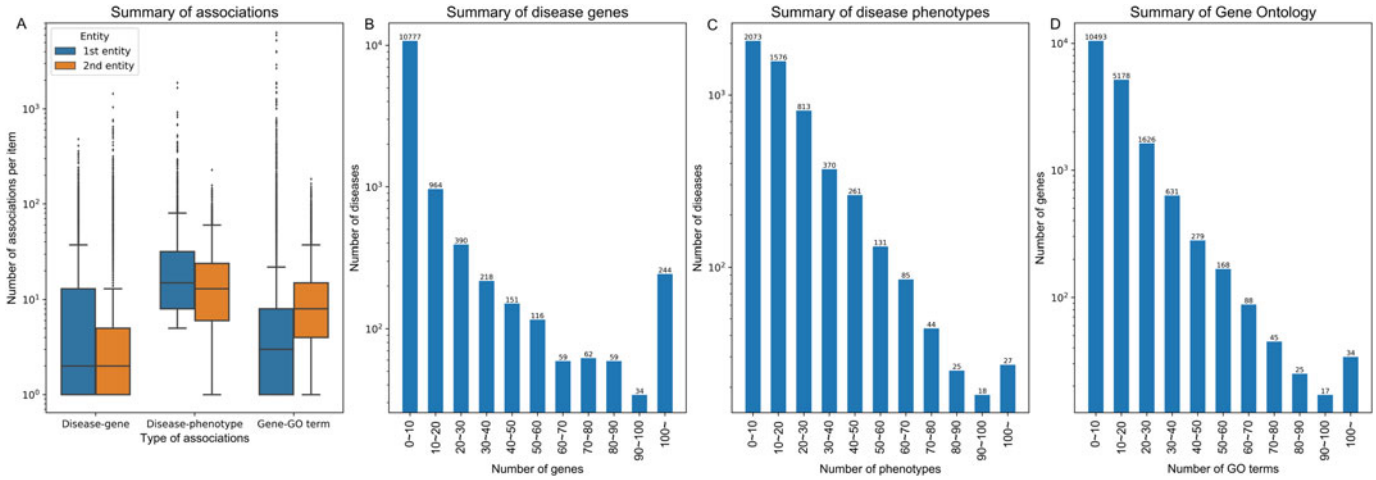

Fig. 2. *Summary of disease and gene related associations.* (**A**) The number distribution of the two types of entities for disease-gene, disease-phenotype and gene-GO term associations, e.g., the first (1st) and second (2nd) entities for disease-gene associations denote diseases and genes, respectively. (**B**) The number distribution of disease-related genes. (**C**) The number distribution of disease-related phenotypes. (**D**) The number distribution of gene-GO terms.

2.2 PDGNet Model

We developed a multi-view deep neural network model PDGNet (Fig. 1) to predict disease genes. In this model, to learn multi-view deep features of diseases and genes, we constructed a non-linear neural network that fed multi-view vectors as input layers. The relevant scores between diseases and genes are measured by the cosine similarity of their deep vectors. Finally, we utilized the cross entropy between the true labels and relevant scores of disease-gene associations as the feedback information to continuously optimize the deep model.

2.2.1 Notations

Mathematically, first, suppose there are m diseases $D = \{d_1, \ldots, d_m\}$ and n genes $G = \{g_1, \ldots, g_n\}$ that are associated with these diseases. We use $A \in \mathbb{R}^{m \times n}$ to denote the association matrix between these diseases and genes, where $A_{ij} = 1$ denotes the gene j is associated with the disease i, and $A_{ij} = 0$ denotes the unknown association between the gene j and the disease i. And the ith row of A is denoted as $A_{i,:}$ and the jth column is denoted as $A_{:,j}$. Similarly, we constructed the matrices of disease-phenotype associations $B \in \mathbb{R}^{m \times l}$ and gene-GO associations $C \in \mathbb{R}^{n \times k}$ based on the known disease-phenotype and gene-GO associations, respectively.

It's necessary to note that the ones in the matrices A, B and C represent the existing associations and the zeros in these matrices represent unknown associations, which contain two kinds of associations, i.e., (1) the undiscovered associations that have not been discovered or confirmed by literatures, (2) the truly negative associations that do not exist between given two entities (e.g., diseases and genes).

2.2.2 Multi-View Neural Network for Predicting Disease Genes

To obtain multi-view deep features in PDGNet, the disease-related gene vectors and phenotype vectors are fed to initial disease inputs of the neural network, and gene-related disease vectors and GO term vectors are fed to initial gene inputs. Specifically, for a disease-gene pair i and j, the vectors $A_{i,:}$ and $B_{i,:}$ were extracted from the matrices A and B as two input views of disease i, and the vectors $A_{:,j}$ and $C_{j,:}$ were extracted from the matrices A and C as two input views of gene j.

Four multi-layer non-linear sub-neural networks for the four input views are constructed to separately extract deep features of diseases and genes. In each sub-neural network, we denote the input vector by x, the output layer by r and the intermediate hidden layers by $h_i, i = 1, \ldots, N$. The symbols W_i and b_i denote the weight matrix and bias matrix for ith hidden layer. We define the first hidden layer $h_1 = W_1 x$ and the ith hidden layer, as follows:

$$h_i = f(W_i h_{i-1} + b_i), i = 2, \ldots, N-1. \tag{1}$$

The output layer is denoted by $r = f(W_N h_{N-1} + b_N)$. The two deep output vectors of diseases are denoted by r_{d1} and r_{d2}, and the two deep output vectors of genes are denoted by r_{g1} and r_{g2}. For every hidden layer, the activation function f is ReLU [51], i.e., $f(x) = max(0, x)$.

To integrate the features from different views, we constructed a symphysic layer $p_i = [r_{d1}; r_{d2}]$ for the disease i by combining the two vectors of the disease using 'vector concat' operation. Similarly, a symphysic vector $q_j = [r_{g1}; r_{g2}]$ for the gene j is also defined. We use the cosine similarity of feature vectors to measure the relevant score $\hat{Y}_{ij}$ between the disease i and the gene j, as follows:

$$\hat{Y}_{ij} = \text{cosine}(p_i, q_j) = \frac{p_i^T q_j}{\|p_i\| \cdot \|q_j\|}. \tag{2}$$

To make the predicted score $\hat{Y}_{ij}$ be as possible as close to the true label Y_{ij}, we define a cross-entropy [52] loss function, as follows:

$$\begin{aligned} L = &- \sum_{(i,j) \in Y} Y_{ij} \log \hat{Y}_{ij} + (1 - Y_{ij}) \log (1 - \hat{Y}_{ij}) \\ &+ \lambda \left(\|p_i\|_L^2 + \|q_j\|_L^2 \right), \end{aligned} \tag{3}$$

where $\| \bullet \|_L^2$ denotes the L2-norm [53] and λ denotes the weight of regularization. The deep feature vectors p_i and q_j

are regularized by L2-norm in order to prevent the overfitting [54] phenomenon of the neural network.

2.2.3 Model Training

To train a lot of parameters of the neural network, we use the back propagation to update these parameters with batch strategy. We defined a feedback set Y that contains positive (marked by Y^+) and negative (marked by Y^-) samples as the implicit feedback of this neural network. Based on 10-fold cross-validation strategy, 117,738 (=130820*0.9) known disease-gene associations from the matrix A (i.e., $A_{ij} = 1$) are randomly selected as the set Y^+ of positive samples in the training of model.

There is still no benchmark set that contains enough truly negative samples of disease-gene associations. We obtained the negative samples of disease-gene associations using a conventional random generation strategy, which has been used to multiple high-quality studies of disease gene prediction [18], [30], [55]. The strategy would randomly generate disease-gene association combinations (i.e., Y^-) from 13,074 diseases and 8,947 genes that don't exist in Y^+. Actually, the current random selected negative samples would incorporate potential novel disease-gene associations, which is also the feasible condition for the disease-gene prediction algorithm with the capability of detecting novel associations. However, it would be an empirical assumption for successful model training that the probability of true association for a random selected disease-gene association from the current negative samples would be much lower than that of the associations from positive samples. In addition, we set the number of negative samples as a parameter that need to be tuned.

2.3 Experimental Settings

To validate the performance of different prediction algorithms, we applied 10-fold cross-validation on a benchmark dataset of disease-gene associations.

2.3.1 Baselines for Comparison

Several classic disease-gene prediction algorithms were selected as baseline algorithms, as follows:

CIPHER [13] is a global network-based inference algorithm for disease gene prediction.

PRINCE [10] is a network propagation-based prediction algorithm based on a PPI network.

DADA [11] is a statistical adjustment-based prediction method to account for the degree distribution of known and candidate genes in a PPI network.

GUILD [56] is a genome-wide network-based algorithm for disease gene prediction.

PgWalk [12] is a random walk-based prediction algorithm in a heterogeneous network that integrated multiple phenomic and genomic datasets.

HerGePerd [19] is a disease-gene prediction framework that fused heterogeneous network representation and network propagation.

2.3.2 Evaluation Metrics

We select the association precision (AP) [19], precision (PR), recall (RE) and F1-score (F1) [57] as evaluation metrics. In detail, given a test set D with M diseases, $T(d)$ denotes the test gene set of the disease $d \in D$. Based on the gene ranking list of disease d, we selected the top i genes $R_i(d)$ of the list (TOP@i) as the final candidate genes of the disease. We define the precision, recall and F1-score for TOP@i, as follows:

$$\text{Precision } = \frac{\sum_{d \in D} |T(d) \cap R_i(d)|}{\sum_{d \in D} |R_i(d)|} \tag{4}$$

$$\text{Recall } = \frac{\sum_{d \in D} |T(d) \cap R_i(d)|}{\sum_{d \in D} |T(d)|} \tag{5}$$

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{6}$$

Meanwhile, we selected the top k genes $R_k(d)$ of the ranking list (k equals to the number of test genes) to calculate the association precision, as follows:

$$\text{AP} = \frac{\sum_{d \in D} |T(d) \cap R_k(d)|}{\sum_{d \in D} |R_k(d)|}. \tag{7}$$

In addition, all metric results are mean values with standard deviation in the 10-fold cross experiments.

2.4 Related Analysis Methods for Prediction Results

Functional Homogeneity Analysis. To further evaluate the candidate genes predicted by our model, we measured the functional homogeneity (FH) of the candidate genes according to the well-defined measures [58]. The FHs are calculated based on the annotations of Gene Ontology that includes biological process (BP), molecular function (MF) and cellular component (CC).

Complex Network Analysis. Based on the PPI network, we calculated the shortest path lengths of protein pairs [59] and the average of shortest path lengths (ASPL) between every candidate gene and the set of known genes for the given disease. The betweenness centrality (BC) and node degree (ND) of every protein in PPI network are calculated for the case study of Parkinson's disease.

Statistical Analysis. We conducted the related random experiments using random permutation [60]. In comparison experiments, we applied the Student's t-test [61] to calculate the statistical significance for the performance of different methods. The Pearson Correlation Coefficient (PCC) and P-value are calculated to measure the correlation for the given two variables (e.g., precision and the number of disease genes).

3 Results

3.1 The Overall Performance of PDGNet

To evaluate the performance of PDGNet, we applied 10-fold cross-validation on a large-scale data with 130,820 disease-gene associations. First, The results showed that PDGNet obtained the best performance (AP=0.373±0.004, PR=0.136±0.001 and RE=0.379±0.004 for TOP@10) among all the algorithms (Table 1 and Figs. 3 A, 3 B, 3 C, and 3

TABLE 1
Performance Comparison of Prediction Algorithms for Disease Candidate Genes

Prediction algorithm	Association precision	TOP@3			TOP@10		
		Precision	Recall	F1-score	Precision	Recall	F1-score
pgWalk	0.258 ± 0.003	0.222 ± 0.003	0.186 ± 0.002	0.202 ± 0.002	0.105 ± 0.001	0.294 ± 0.002	0.155 ± 0.002
PRINCE	0.019 ± 0.003	0.006 ± 0.003	0.005 ± 0.005	0.005 ± 0.004	0.005 ± 0.002	0.014 ± 0.010	0.007 ± 0.003
CIPHER	0.003 ± 0.002	0.002 ± 0.001	0.002 ± 0.001	0.002 ± 0.001	0.001 ± 0.000	0.004 ± 0.001	0.002 ± 0.001
DADA	0.087 ± 0.007	0.079 ± 0.006	0.097 ± 0.006	0.087 ± 0.006	0.035 ± 0.001	0.141 ± 0.006	0.056 ± 0.002
GUILD	0.091±0.007	0.082 ± 0.003	0.101 ± 0.004	0.091 ± 0.003	0.036 ± 0.002	0.147 ± 0.008	0.058 ± 0.003
HerGePred	0.294 ± 0.005*	0.243± 0.003*	0.203 ± 0.002*	0.221 ± 0.003*	0.124 ± 0.002*	0.346 ± 0.005*	0.183 ± 0.003*
PDGNet	**0.373 ± 0.004**	**0.249 ± 0.003**	**0.208 ± 0.002**	**0.227 ± 0.003**	**0.136 ± 0.001**	**0.379 ± 0.004**	**0.200 ± 0.002**
Improvement	26.71%	2.30%	2.58%	2.55%	9.55%	9.63%	9.31%
(P-value)	(6.18E-19)	(1.52E-03)	(1.93E-04)	(3.16E-04)	(2.07E-10)	(1.45E-11)	(6.49E-11)

*PDGNet obtain best performance (marked by bold text) in all the algorithms and HerGePred obtain best performance (marked by *) in baselines. The proportions and P-values of the improvement of PDGNet compared with HerGePred are shown.*

D). DeepGN achieves higher AP (increased 26.71 percent, P=6.18E-19), PR (increased 9.55 percent, P=2.07E-10) and RE (increased 9.63 percent, P=1.45E-11) for TOP@10 than HerGePred with the best performance in all of baselines (Fig. 3 A). The precision-recall curve (Fig. 3 B) also showed similar results. Second, we showed the tendency of the precision and recall under different TOP@K with the range from 1 to 100. The results (Figs. 3 C and 3 D) indicated that with the constant increase of TOP@K, the precision rate decreases and the recall rate increases for all the algorithms. From the term of recall rate, DeepGN showed the bigger improvement than the baseline algorithms as TOP@K increases (Fig. 3 D).

To evaluate the contributions of different views (A: disease-gene matrix; B: disease-symptom matrix; C: gene-GO matrix) involved in PDGNet, we designed five variants of PDGNet with different views (i.e., A, AB, AC, BC and ABC) and compared the prediction performance of them (Fig. 3 E). First, the performance rankings (i.e., ABC>AB>AC>A>BC) of these algorithm variants indicated that PDGNet with

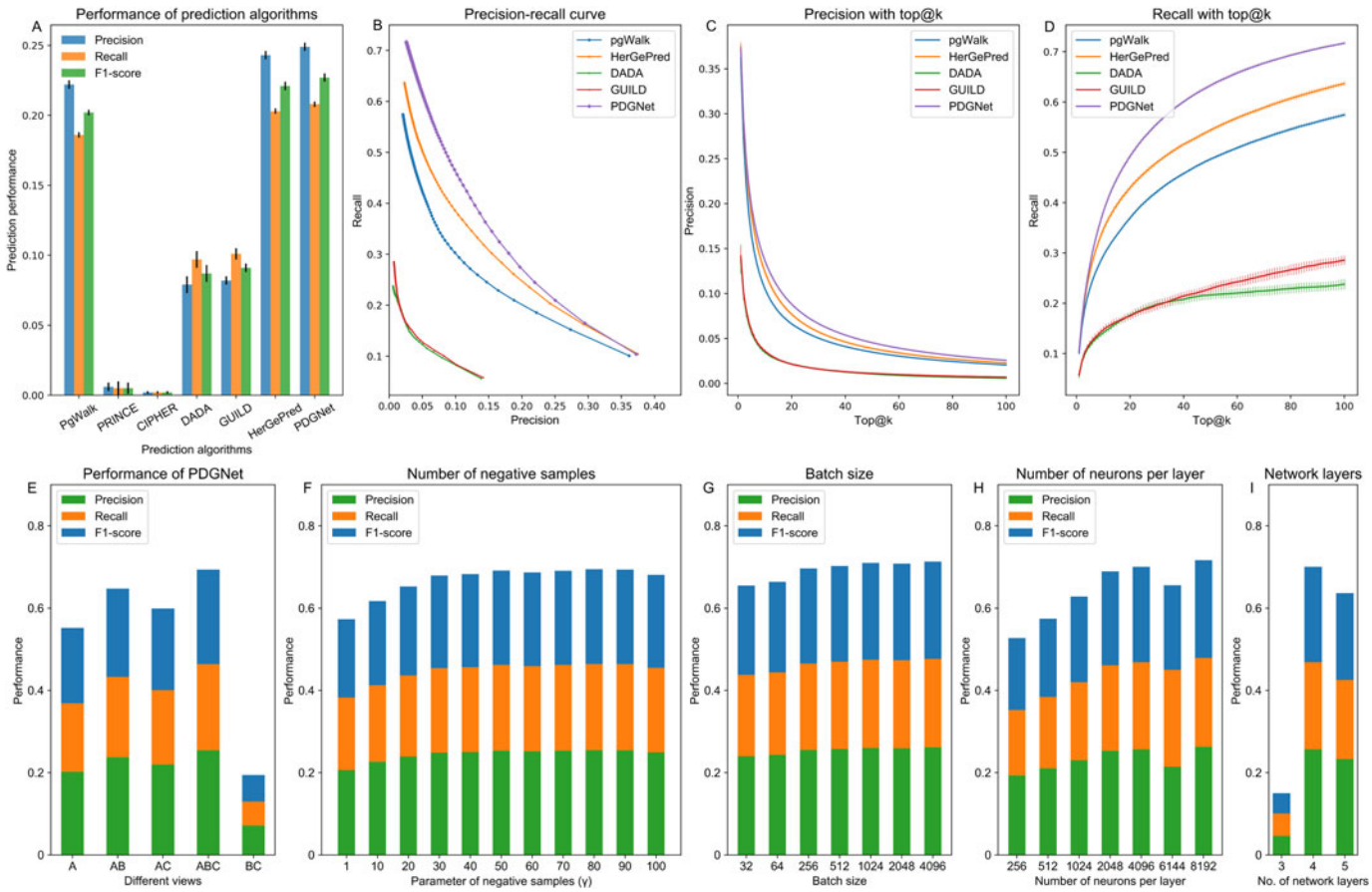

Fig. 3. *Performance of prediction algorithms.* (**A-D**) A performance comparison of prediction algorithms. The precision, recall and F1-score under Top@3 (**A**), the precision-recall curve (**B**), precision (**C**), recall (**D**) under different Top@k indicates that the performance of PDGNet is better than that of all baseline algorithms. (**E**) Contribution of different views of PDGNet. (**F-I**) The prediction performance of the four tuning parameters are shown: parameter γ of negative samples (**F**), batch size (**G**), number of neurons per layer (**H**) and number of network layers (**I**).

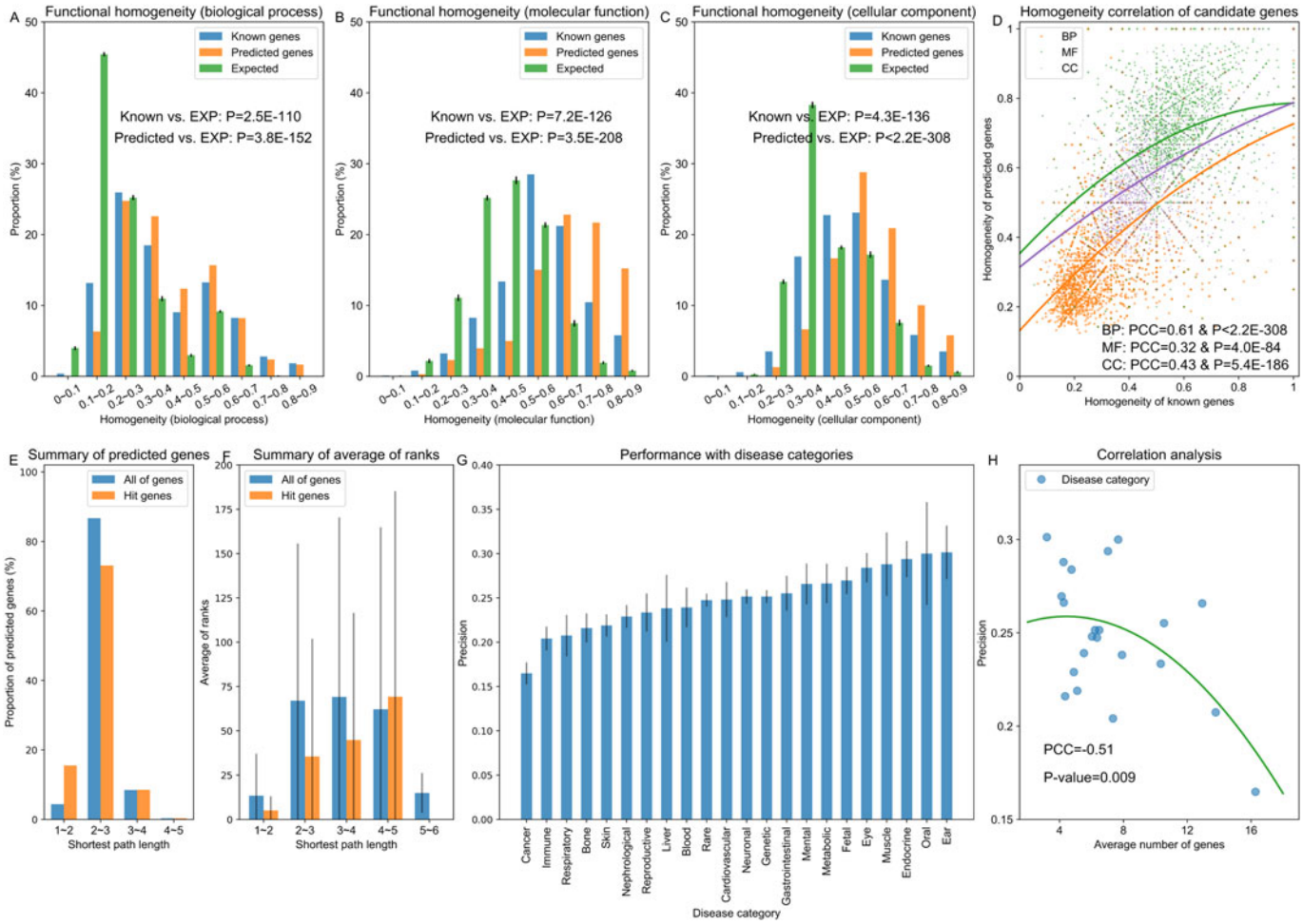

Fig. 4. *Evaluation and analysis of candidate genes.* (**A-C**) The functional homogeneity of the candidate genes and the known genes of diseases. The analysis of functional homogeneity included biological process (BP) (**A**), molecular function (MF) (**B**) and cellular component (CC) (**C**). We compared candidate genes and known genes with expected results (termed EXP) and showed the p-values (termed P) of significance testing. (**D**) Correlation analysis of functional homogeneity for known and candidate genes. (**E**) Summary of proportion of predicted genes corresponding to shortest path lengths. (**F**) Summary of ranks of predicted gene corresponding to shortest path lengths. (**G**) The performance with different disease categories. (**H**) Correlation analysis between the average of the number of disease genes and the prediction precision.

multi-views can improve the performance on predicting disease genes. For example, PDGNet with three views (i.e., ABC) obtained higher performance than that with two views (i.e., AB and AC). However, the performance of PDGNet with A view is much higher than that with BC views, which indicated that the disease-gene view is more effective information than disease-symptom and gene-GO views.

3.2 Further Analysis of the Candidate Genes Predicted by PDGNet

3.2.1 Functional Homogeneity Analysis

To have a full evaluation for the candidate genes predicted by the PDGNet, we have conducted an in-depth functional homogeneity analysis for the candidate genes. First, we calculated the functional homogeneity (see Methods) of the candidate genes and the known genes of test diseases and compared them with random results. The results (Figs. 4 A, 4 B, and 4 C) showed both candidate genes and known genes have stronger FHs than random experiment. The FHs of candidate genes are stronger than that of known genes. Taking BP as an example, there are 1,340 diseases (∼42 percent, P=3.8E-152) for candidate genes and 1,181 diseases (∼38 percent, P=2.5E-110) for known genes with FH bigger than 0.4, respectively. Other annotations of GO (i.e., MF and CC) also showed similar results. Meanwhile, the correlation analysis (Fig. 4 D) showed the FHs of the candidate genes and the known genes have strong correlation (PCC=0.61 and P<2.2E-308 for BP; PCC=0.32 and P=4.0E-84 for MF; PCC=0.43 and P=5.4E-186 for CC), which indicated that the results predicted by PDGNet have good consistency with existing knowledge of disease genes.

3.2.2 Network Interaction Analysis

To investigate the interaction relationships between the candidate genes and the known genes of diseases, we calculated the information of the shortest path lengths among these genes in PPI network (see Methods). First, the results of ASPL (Fig. 4 E) showed that the ASPLs of ∼16 percent hit genes (i.e., the genes exist in the test set) are the range from 2 to 3 compared with ∼4 percent predicted genes with this range (P=7.1E-320), which validated that the hit genes (i.e., the known genes of diseases) have closer interactions with the known genes of diseases. Second, we analysed the ranks of the candidate genes corresponding to the ASPL of these

TABLE 2
Top Ten Candidate Genes for Parkinson's Disease

Rank	Candidate gene	Degree of genes	Betweenness of genes	Recorded in MalaCards	Co-occurrence in related reference	Top related references (PubMed ids)
1	DNAJC13	18	4.73E-06	✓	✓	30788857, 30537300
2	SNCB	15	6.59E-06	✓	✓	30711526, 30040713
3	PTGS2	28	3.92E-05		✓	30413118, 30368226
4	VEGFA	30	2.31E-04			
5	SNCAIP	28	1.18E-05	✓	✓	30316984, 28653979
6	**HMOX1**	15	2.24E-05	✓	✓	30810907, 30739428
7	SIRT1	297	0.00137		✓	30929586, 30826419
8	VPS13C	3	3.10E-07	✓	✓	30245141, 30093493
9	NR4A2	36	2.73E-05	✓	✓	30859219, 30949504
10	IGF1	19	2.80E-05		✓	28221705, 29149058

The bold gene HMOX1 exists in the test set.

genes. The distribution of their ranks (Fig. 4 F) indicated that the smaller ASPL of hit genes are, the higher their ranks, while the ranks of all predicted genes don't show similar tendency. This implied that PDGNet has the power of giving high scores to some candidate genes that have close interactions with the known genes of given diseases in PPI network.

3.2.3 Performance Analysis of Different Disease Categories

According to the 22 disease categories derived from MalaCards database, we showed that the prediction performance of PDGNet model on the diseases of every category. The results (Fig. 4 G) indicated that the diseases of different categories have an obvious difference on the prediction performance. To explore the factors that affect the performance for different diseases, we conducted the correlation analysis for the average of the number of disease's known genes (ANDG) and the performance metrics. The results (Fig. 4 H) showed that the performance metrics have negative correlations (PCC=-0.51; P=0.009) with the ANDG. For example, the ear (ANDG=3.2; PR=0.3) and muscle (ANDG=4.24; PR=0.29) diseases obtain better performance than the cancer (ANDG=16.27; PR=0.16) and respiratory (ANDG=13.78; PR=0.21) diseases. This indicated that most of diseases with a large number of known genes are more difficult to identify novel genes than the diseases with a small number of known genes.

3.3 Parameter Sensitivity Analysis

In PDGNet model, there are four hyper-parameters that need to be tuned. The first two parameters are the number of neural network layers (α=3, 4 and 5) and the number of neurons per layer (β=from 256 to 8,192) that are tuned in order to optimize the structure of the proposed neural network. In each epoch of the training, there are two parameters, i.e., the number of negative samples (γ=from 1 to 100; γ=10 represents there are 10 times as many negative samples as positive samples) and the batch size (θ=from 32 to 4,096) that need to be tuned. Each parameter is tuned when the remaining three parameters are fixed in turn. First, the results (Figs. 3 F, 3 G, 3 H, and 3 I) indicated that PDGNet obtained high performance when α is 4, but both smaller α and bigger α degrade the prediction performance.

Second, the number of neurons per layer had a strong influence on the performance of PDGNet. The precision and recall have an increase in fluctuations as β increases. However, too big β maybe bring about lots of memory overhead and lead to the substantial increase of the training time for PDGNet. Third, the left two parameters (γ and θ) showed slight influence for PDGNet, and the model obtained high performance when γ is bigger than 30 and θ exceeds 512. Finally, according to the results of parameter tuning, the optimized PDGNet is a four-layer neural network (includes an input layer, an output layer and two hidden layers) and every layer contains 8,129 neurons. The number of negative samples is 50 and the batch size equals to 2,048.

3.4 Case Study: Parkinson's Disease

To show the capability that PDGNet seeks out novel and credible candidate genes, we obtained the top ten candidate genes (Table 2) for Parkinson's disease (PD) [62].

First, we used an independent database of disease-gene association (i.e., MalaCards) and two literature databases (i.e., PubMed and DISEASES) to validate the credibility of these candidate genes. We found that there are six candidate genes (precision=60 percent), namely, DNAJC13 (rank=1), SNCB (rank=2), SNCAIP (rank=5), HMOX1 (rank=6), VPS13C (rank=8) and NR4A2 (rank=9), that are recorded in MalaCards database. Especially the gene HMOX1 also exists in the test set. Second, except for the gene VEGFA, the remaining nine genes have co-occurrence with PD in current published literatures, which declared these genes are likely to be associated with PD. For example, the recent studies [63], [64] (PMID: 30678325 and 30810907) manifest that PTGS2 (rank=3) is likely to associate with PD. Wang *et al.* [65] (PMID: 30826419) found miR-9-5p modulates the progression of PD by targeting SIRT1 (rank=7).

Third, we showed the node degrees and betweenness centrality (see Methods) of these candidate genes. It showed that most of genes (9/10=90 percent; with $ND < 50$ and $BC < 0.0001$) are marginal genes (i.e., the genes with low ND and BC) of the PPI network, while the gene SIRT1 (ND=297 and BC=0.00137) is a core gene (i.e., the gene with high ND and BC). This indicated compared with network propagation algorithms that have bias toward high degree proteins [66], PDGNet has the ability of identifying both marginal and core candidate genes.

4 Discussion

In this study, we proposed a deep neural network model that fused the multi-view features of phenotypes and genotypes to predict disease genes. The evaluation experiments indicated that our model obtained higher prediction performance than the existing methods. Systematic analysis for the predicted candidate genes also manifested the reliability and high quality of the prediction results of PDGNet. We summarized the several advantages of PDGNet compared with other prediction algorithms. First, the current network-based methods for disease genes often are focused on the methods of network propagation (DADA [11] and pgWalk [12]) and correlation analysis (CIPHER [13]). However, the deep neural network model we designed have the virtues of fitting complex nonlinear functions and automatically learning the deep features of input data [67]. Second, compared with the algorithms of network embedding [19], [37] that is two-step methods (i.e., first learn the embedding features of nodes, then conduct candidate gene prediction based on the features), PDGNet would utilize the feedback information of training samples to learn the deep features of diseases and genes and optimize the neural network-based prediction model at the same time in each step of model training. Finally, PDGNet applied the multi-view information of diseases and genes to enrich the initial features of them and improve the robustness of PDGNet.

There are several works that are worth to explore in the future. First, the original intention of the model is to obtain initial features of diseases and genes from multi-view information. If neither the gene vector and phenotype vector of a given disease are present, our model would not be able to predict the new candidate genes for this disease. Therefore, in the future, more views of disease and genes (e.g., disease modules [2], [47] and protein interactions) should be considered to enrich the input features. Second, owing to the poor performance on complex diseases (e.g., cancer diseases) for PDGNet, we would try to design a specific PDGNet model with high performance on these complex diseases. Third, in the field of disease genes prediction, the random generation strategy of negative samples is still a limitation. Furthermore, the similar strategy as the screening reliable negative samples proposed by Liu *et al*. [68] could be incorporated in our future work for possible improvement.

Fourth, inspired by the graph convolutional network-based method for disease gene prediction [20], we will try to combine graph neural network [69] with multi-task learning [70] to construct prediction model for disease genes. For example, we can construct a large-scale heterogeneous network that contains disease-gene, disease-phenotype and gene-GO associations, and try to design a specific multi-task-based graph neural network that contains a main task (i.e., predicting disease genes) and two deputy tasks (i.e., predicting disease phenotypes and predicting GO terms), which would be trained synchronously in order to obtain better embeddings of diseases and genes. Finally, although PDGNet obtain high performance on predicting disease genes, there is a poor interpretability [71] with PDGNet that most of current neural networks have. The technology of deep learning with knowledge graph [72] may be a good answer to improve the interpretability.

Acknowledgments

This work was partially supported by the National Key Research and Development Program [2017YFC1703506 and 2017YFC1703502], the National Science and Technology Major Project [2019ZX09201005-002-006], the Fundamental Research Funds for the Central Universities [2018JBZ006], the Special Programs of Traditional Chinese Medicine [JDZX2015170 and JDZX2015171]. Kuo Yang and Yi Zheng are contributing equally to this work.

References

[1] D. Botstein and N. Risch, "Discovering genotypes underlying human phenotypes: Past successes for mendelian disease, future approaches for complex disease," *Nat. Genet.*, vol. 33, no. 33 Suppl, pp. 228–237, 2003.
[2] B. Albert-László, G. Natali, and L. Joseph, "Network medicine: A network-based approach to human disease," *Nat. Rev. Genet.*, vol. 12, no. 1, pp. 56–68, 2011.
[3] L. Kasper *et al.*, "A human phenome-interactome network of protein complexes implicated in genetic disorders," *Nat. Biotechnol.*, vol. 25, no. 3, pp. 309–316, 2007.
[4] D. C. Crawford, A. A. Morgan, J. C. Denny, B. J. Aronow, and S. E. Brenner, "Precision medicine: From diplotypes to disparities towards improved health and therapies," in *Proc. Pacific Symp. Biocomputing*, 2018, vol. 23, pp. 389–399.
[5] R. A. George, J. Y. Liu, L. L. Feng, R. J. Bryson-Richardson, D. Fatkin, and M. A. Wouters, "Analysis of protein sequence and interaction data for candidate disease gene prediction," *Nucleic Acids Res.*, vol. 34, no. 19, 2006, Art. no. 130.
[6] S. Karni, H. Soreq, and R. Sharan, "A network-based method for predicting disease-causing genes," *J. Comput. Biol.*, vol. 16, no. 2, pp. 181–189, 2009.
[7] F. Mordelet and J. P. Vert, "ProDiGe: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples," *BMC Bioinf.*, vol. 12, no. 1, pp. 1–15, 2011.
[8] D. S. Himmelstein and S. E. Baranzini, "Heterogeneous network edge prediction: A data integration approach to prioritize disease-associated genes," *PLOS Comput. Biol.*, vol. 11, no. 7, 2015, Art. no. e1004259.
[9] X. Zeng, Y. Liao, Y. Liu, and Q. Zou, "Prediction and validation of disease genes using HeteSim scores," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 3, pp. 687–695, May/Jun. 2017.
[10] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation," *PLOS Comput. Biol.*, vol. 6, no. 1, 2010, Art. no. e1000641.
[11] S. Erten, G. Bebek, R. M. Ewing, and M. Koyutürk, "DADA: Degree-aware algorithms for network-based disease gene prioritization," *Biodata Mining*, vol. 4, no. 1, 2011, Art. no. 19.
[12] R. Jiang, "Walking on multiple disease-gene networks to prioritize candidate genes," *J. Mol. Cell Biol.*, vol. 7, no. 3, 2015, Art. no. 214.
[13] X. Wu, R. Jiang, M. Q. Zhang, and S. Li, "Network-based global inference of human disease genes," *Mol. Syst. Biol.*, vol. 4, no. 1, 2008, Art. no. 189.
[14] Y. Xin, H. Han, Y. Li, and L. Shao, "Modularity-based credible prediction of disease genes and detection of disease subtypes on the phenotype-gene heterogeneous network," *BMC Syst. Biol.*, vol. 5, no. 1, pp. 1–11, 2011.
[15] P. Radivojac *et al.*, "An integrated approach to inferring gene-disease associations in humans," *Proteins-Struct. Function Bioinf.*, vol. 72, no. 3, pp. 1030–1037, 2008.
[16] H. Zhou and J. Skolnick, "A knowledge-based approach for predicting gene-disease associations," *Bioinformatics*, vol. 32, no. 18, pp. 2831–2838, 2016.
[17] M. Wu, W. Zeng, W. Liu, Y. Zhang, T. Chen, and R. Jiang, "Integrating embeddings of multiple gene networks to prioritize complex disease-associated genes," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2017, pp. 208–215.
[18] M. Alshahrani and R. Hoehndorf, "Semantic disease gene embeddings (SmuDGE): Phenotype-based disease gene prioritization without phenotypes," *Bioinformatics*, vol. 34, no. 17, pp. i901–i907, 2018.
[19] K. Yang *et al.*, "HerGePred: Heterogeneous network embedding representation for disease gene prediction," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 4, pp. 1805–1815, Jul. 2019.

[20] Y. Li, H. Kuwahara, P. Yang, L. Song, and X. Gao, "PGCN: Disease gene prioritization by disease and gene embedding through graph convolutional neural networks," *bioRxiv*, 2019, Art. no. 532226.
[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[22] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.
[23] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018.
[24] Y. L. Cun *et al.*, "Handwritten digit recognition with a back-propagation network," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 1990, vol. 2, pp. 396–404.
[25] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Comput.*, vol. 12, no. 10, pp. 2451–2471, 2000.
[26] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[27] M. Zitnik, M. Agrawal, and J. Leskovec, "Modeling polypharmacy side effects with graph convolutional networks," *Bioinformatics*, vol. 34, no. 13, pp. i457–i466, 2018.
[28] W. Li, W. H. Wong, and R. Jiang, "DeepTACT: Predicting 3D chromatin contacts via bootstrapping deep learning," *Nucleic Acids Res.*, vol. 47, no. 10, 2019, Art. no. e60.
[29] P. Xuan, Y. Cao, T. Zhang, R. Kong, and Z. Zhang, "Dual convolutional neural networks with attention mechanisms based method for predicting disease-related lncRNA genes," *Front. Genet.*, vol. 10, no. 1, 2019, Art. no. 416.
[30] J. Peng, J. Guan, and X. Shang, "Predicting parkinson's disease genes based on node2vec and autoencoder," *Front. Genet.*, vol. 10, no. 1, 2019, Art. no. 226.
[31] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3D shape recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 945–953.
[32] S. S. Farfade, M. J. Saberian, and L.-J. Li, "Multi-view face detection using deep convolutional neural networks," in *Proc. 5th ACM Int. Conf. Multimedia Retrieval*, 2015, pp. 643–650.
[33] A. Zeng *et al.*, "Multi-view self-supervised deep learning for 6D pose estimation in the Amazon picking challenge," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2017, pp. 1386–1383.
[34] K. J. Geras *et al.*, "High-resolution breast cancer screening with multi-view deep convolutional neural networks," 2017, *arXiv: 1703.07047*.
[35] C. Zhang, E. Adeli, T. Zhou, X. Chen, and D. Shen, "Multi-layer multi-view classification for alzheimer's disease diagnosis," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 4406–4413.
[36] T. Ma, C. Xiao, J. Zhou, and F. Wang, "Drug similarity integration through attentive multi-view graph auto-encoders," in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2018, pp. 3477–3483.
[37] K. Yang *et al.*, "Heterogeneous network embedding for identifying symptom candidate genes," *J. Amer. Med. Informat. Assoc.*, vol. 25, no. 11, pp. 1452–1459, 2018.
[38] J. Piñero *et al.*, "DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes," *Database J. Biol. Databases Curation*, vol. 2015, no. 3, 2015, Art. no. bav028.
[39] A. Gutiérrezsacristàn *et al.*, "PsyGeNET: A knowledge platform on psychiatric disorders and their genes," *Bioinformatics*, vol. 31, no. 18, pp. 3075–3077, 2015.
[40] M. J. Landrum *et al.*, "ClinVar: Public archive of relationships among sequence variation and human phenotype," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. 980–985, 2014.
[41] R. Mangon, J. J. Sikkens, M. Teeuw, and M. C. Cornel, "Orphanet: A european database for rare diseases," *Nederlands Tijdschrift Voor Geneeskunde*, vol. 152, no. 9, pp. 518–519, 2008.
[42] D. Welter *et al.*, "The NHGRI GWAS catalog, a curated resource of SNP-trait associations," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. 1001–1006, 2014.
[43] N. Rappaport *et al.*, "MalaCards: An amalgamated human disease compendium with diverse clinical and genetic annotation and structured search," *Nucleic Acids Res.*, vol. 45, no. Database issue, pp. D877–D887, 2017.
[44] S. Pletscher-Frankild, A. Pallejà, K. Tsafou, J. X. Binder, and L. J. Jensen, "DISEASES: Text mining and data integration of disease–gene associations," *Methods*, vol. 74, pp. 83–89, 2015.
[45] S. Köhler *et al.*, "The human phenotype ontology in 2017," *Nucleic Acids Res.*, vol. 45, no. Database issue, pp. D865–D876, 2017.
[46] D. Szklarczyk *et al.*, "The string database in 2017: Quality-controlled protein–protein association networks, made broadly accessible," *Nucleic Acids Res.*, vol. 45, no. Database issue, pp. D362–D368, 2017.
[47] M. Jörg *et al.*, "Disease networks. Uncovering disease-disease relationships through the incomplete interactome," *Science*, vol. 347, no. 6224, 2015, Art. no. 1257601.
[48] G. R. Mishra *et al.*, "Human protein reference database–2006 update," *Nucleic Acids Res.*, vol. 34, no. Database issue, pp. 411–414, 2006.
[49] S. Orchard *et al.*, "The mintact project—intact as a common curation platform for 11 molecular interaction databases," *Nucleic Acids Res.*, vol. 42, pp. 358–363, 2014.
[50] M. J. Cowley *et al.*, "PINA v2.0: Mining interactome modules," *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. 862–865, 2012.
[51] K. He, X. Zhang, S. Ren, and S. Jian, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.
[52] R. Rubinstein, "The cross-entropy method for combinatorial and continuous optimization," *Methodol. Comput. Appl. Probability*, vol. 1, no. 2, pp. 127–190, 1999.
[53] A. Neumaier, "Solving ill-conditioned and singular linear systems: A tutorial on regularization," *SIAM Rev.*, vol. 40, no. 3, pp. 636–666, 1998.
[54] I. V. Tetko, D. J. Livingstone, and A. I. Luik, "Neural network studies. 1. Comparison of overfitting and overtraining," *J. Chem. Inf. Comput. Sci.*, vol. 35, no. 5, pp. 826–833, 1995.
[55] J. Xu and Y. Li, "Discovering disease-genes by topological features in human protein–protein interaction network," *Bioinformatics*, vol. 22, no. 22, pp. 2800–2805, 2006.
[56] G. Emre and O. Baldo, "Exploiting protein-protein interaction networks for genome-wide disease-gene prioritization," *PLOS One*, vol. 7, no. 9, 2012, Art. no. e43557.
[57] G. Adomavicius and Y. O. Kwon, "Improving aggregate recommendation diversity using ranking-based techniques," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 5, pp. 896–911, May 2012.
[58] G. Liu, H. Wang, H. Chu, J. Yu, and X. Zhou, "Functional diversity of topological modules in human protein-protein interaction networks," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 16199.
[59] K. Yang *et al.*, "Heterogeneous network propagation for herb target identification," *BMC Med. Informat. Decision Making*, vol. 18, no. 1, 2018, Art. no. 17.
[60] R. A. Fisher and F. Yates, "Statistical tables for biological, agricultural and medical research," *Can. J. Comparative Med. Veterinary Sci.*, vol. 22, no. 1, 1958, Art. no. 8.
[61] J. F. Box, "Guinness, gosset, fisher, and small samples," *Statist. Sci.*, vol. 2, no. 1, pp. 45–52, 1987.
[62] W. Poewe *et al.*, "Parkinson disease," *Nat. Rev. Disease Primers*, vol. 3, 2017, Art. no. 17013.
[63] G. P. Selvakumar *et al.*, "CRISPR/cas9 editing of glia maturation factor regulates mitochondrial dynamics by attenuation of the NRF2/HO-1 dependent ferritin activation in glial cells," *J. Neuroimmune Pharmacol.*, vol. 14, pp. 537–550, 2019.
[64] Y.-L. Yang, X. Cheng, W.-H. Li, M. Liu, Y.-H. Wang, and G.-H. Du, "Kaempferol attenuates LPS-induced striatum injury in mice involving anti-neuroinflammation, maintaining BBB integrity, and down-regulating the HMGB1/TLR4 pathway," *Int. J. Mol. Sci.*, vol. 20, no. 3, 2019, Art. no. e491.
[65] Z. Wang, L. Sun, K. Jia, H. Wang, and X. Wang, "miR-9–5p modulates the progression of parkinson's disease by targeting sirt1," *Neurosci. Lett.*, vol. 701, pp. 226–233, 2019.
[66] H. Biran, M. Kupiec, and R. Sharan, "Comparative analysis of normalization methods for network propagation," *Front. Genet.*, vol. 10, 2019, Art. no. 4.
[67] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[68] H. Liu, J. Sun, J. Guan, J. Zheng, and S. Zhou, "Improving compound–protein interaction prediction by building up highly credible negative samples," *Bioinformatics*, vol. 31, no. 12, pp. 221–229, 2015.
[69] S. Franco, G. Marco, T. Ah Chung, H. Markus, and M. Gabriele, "The graph neural network model," *IEEE Trans. Neural Netw.*, vol. 20, no. 1, pp. 61–80, Jan. 2009.

[70] X. Chu, Y. Lin, Y. Wang, L. Wang, J. Wang, and J. Gao, "MLRDA: A multi-task semi-supervised learning framework for drug-drug interaction prediction," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 4518–4524.
[71] A. A. Cruz-Roa, J. E. A. Ovalle, A. Madabhushi, and F. A. G. Osorio, "A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2013, vol. 16, pp. 403–410.
[72] H. Wang, F. Zhang, X. Xie, and M. Guo, "DKN: Deep knowledge-aware network for news recommendation," in *Proc. Int. World Wide Web Conf.*, 2018, pp. 1835–1844.

Kuo Yang received the PhD degree in computer science from Beijing Jiaotong University, China, in 2020. He is currently a postdoctoral associate with the BNRIST/ Department of Automation, Tsinghua University, China. His research interests include machine learning, deep learning, bioinformatics, and artificial intelligence in medicine.

Yi Zheng is currently working toward the graduate degree at Beijing Jiaotong University, China. His research interests include machine learning , big data analysis, and statistical learning in bioinformatics.

Kezhi Lu is currently working toward the graduate degree in the Institute of Machine Learning and Cognitive Computing, Beijing Jiaotong University, China. His research interests include natural language processing, intelligent medicine, bioinformatics and machine learning.

Kai Chang is currently working toward the PhD degree at the School of Beijing Jiaotong University, China. His research interests focus on artificial intelligence in medicine, medical data analysis, and mining.

Ning Wang is currently working toward the PhD degree at the School of Beijing Jiaotong University, China. His research interests focus on artificial intelligence in medicine, medical data analysis and mining.

Zixin Shu is currently working toward the PhD degree in computer science at Beijing Jiaotong University, China. Her research interests include data analysis of medicine, ontology of symptom and knowledge graph construction.

Jian Yu received the BS and MS degrees in mathematics and the PhD degree in applied mathematics from Peking University, Beijing, China, in 1991, 1994, and 2000, respectively. He is currently a professor with the School of Computer and Information Technology and the director of the Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, China. His research interests include machine learning, image processing, and pattern recognition.

Baoyan Liu is a chair professor of the China Academy of Chinese Medical Sciences, China. His research interests include clinical acupuncture, clinical evaluation methodology and information technology framework for traditional Chinese medicine. He has published more than 200 peer reviewed papers, including JAMA and Annals of Internal Medicine, etc.

Zhuye Gao received the MD degree from the Beijing University of Chinese Medicine, China, in 2009. He is currently a professor with National Clinical Research Center for Chinese Medicine Cardiology. His research interests cover the diagnosis and the treatment of cardiovascular disease, and evidence-based medicine. He has published more than 30 peer reviewed papers, including the *Frontiers in Pharmacology*, the *Journal of the American Heart Association*, the *E-CAM*, etc.

Xuezhong Zhou received the PhD degree in computer science from Zhejiang University, China, in 2005. He is currently a professor with the School of Computer and Information Technology, Beijing Jiaotong University. His research interests cover medical data mining, network medicine, clinical decision support, and medical knowledge engineering. He has published more than 100 peer reviewed papers, including the *Nature Communications*, the *Ebiomedicine*, the *Nucleic Acid Research*, etc.