IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS

IEEE Xplore® Digital Library
http://ieeexplore.ieee.org/Xplore

# HerGePred: Heterogeneous Network Embedding Representation for Disease Gene Prediction

Kuo Yang, Ruyu Wang, Guangming Liu, Zixin Shu, Ning Wang, Runshun Zhang, Jian Yu, Jianxin Chen, Xiaodong Li, Xuezhong Zhou*

***Abstract*—The discovery of disease-causing genes is a critical step towards understanding the nature of a disease and determining a possible cure for it. In recent years, many computational methods to identify disease genes have been proposed. However, making full use of disease-related (e.g., symptoms) and gene-related (e.g., Gene Ontology and protein-protein interactions) information to improve the performance of disease-gene prediction is still an issue. Here, we developed a heterogeneous disease-gene-related network (HDGN) embedding representation framework for disease gene prediction (called HerGePred). Based on the framework, low-dimensional vector representation (LVR) of the nodes in the HDGN can be obtained. Then, we proposed two specific algorithms, LVRSim and RW-RDGN, to predict disease genes with high performance. First, to validate the rationality of the framework, we analyzed the similarity-based overlap distribution of disease pairs and designed an experiment for disease-gene association recovery, the results of which revealed that the LVR of nodes performs well at preserving the local and global network structure of the HDGN. Then, we applied 10-fold cross validation and external validation to compare our methods to other well-known disease-gene prediction algorithms. The experimental results showed that RW-RDGN performed better than the state-of-the-art algorithm. The prediction results of disease candidate genes is essential for molecular mechanism investigation and experimental validation. The source code of HerGePred and experimental data are available at https://github.com/yangkuoone/HerGePred.**

***Index Terms*—Disease gene prediction, network embedding representation, heterogeneous network, network propagation.**

## I. Introduction

The identification of genes involved in genetic and rare diseases is a primary step towards revealing the underlying molecular mechanisms of diseases and can potentially improve clinical therapies for diseases. The investigation of susceptible loci by linkage analysis would involve hundreds of genes and require labor intensive efforts to experimental identification of the disease-causing genes [1, 2]. In the last several decades, computational methods for the identification of disease genes have been developed. The basic criterion of in silico gene prediction methods is the "guilt by association" principle with respect to all known genes related to the given query disease [3]. We divided current prediction algorithms into four types according to the data types that they utilize: (1) methods using protein-protein interaction (PPI) data [4-11]; (2) methods integrating PPI and disease phenotypic data [12-17]; (3) methods using PPI and other single types of gene-related data [18-20], such as gene expression, pathway, transcriptional regulation, or Gene Ontology (GO); and (4) methods using multiple types of data [1, 21-25], such as literature data, disease phenotypic data, PPI data, GO annotations, gene expression, protein domain-dependent sequences, protein pathway data, protein sequences, signaling networks and transcriptional regulation. Current prediction methods were mainly divided into four types based on their algorithm principles: (1) network propagation on PPI networks or heterogeneous networks [1, 6, 8, 14, 16, 25]; (2) shortest path analysis on PPI networks or integrated networks [22, 26]; (3) correlation analysis methods [5, 12, 13, 17-19, 21, 23, 27]; and (4) cluster-based or classification-based methods [4, 7, 9, 20, 28-30]. In addition to the aforementioned algorithms, there are other kinds of prediction algorithms, such as AlignPI [15] with network alignment, ProDiGe [24] with kernel data fusion, CATAPULT [31] with Katz measure and positive-unlabeled learning techniques, Gentrepid [32] with Gentrepid system and multiple-locus-based approach, DADA [10] and GUILD [11] with node degree-bias of related network, and Beegle [33] with literature mining and genomic data fusion, to identify disease genes. In fact, heterogeneous and multi-source data offer multi-

This work was partially supported by the Fundamental Research Funds for Central Universities (2017YJS057 and 2017JBM020), National Key Research and Development Program (2017YFC1703506), the National Science Foundation of China (61105055 and 81230086), Special Programs of Traditional Chinese Medicine (201407001, JDZX2015170 and JDZX2015171), and the National Key Technology R&D Program (2013BAI02B01 and 2013BAI13B04).

K. Yang, R. Y. Wang, G. M. Liu, N. Wang, J. Yu and X. Z. Zhou, are at the School of Computer and Information Technology and Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, 100044, China (e-mail: yangkuo@bjtu.edu.cn; 17125236@bjtu.edu.cn; 13112078@bjtu.edu.cn; 15120442@bjtu.edu.cn; jianyu@bjtu.edu.cn; xzzhou@bjtu.edu.cn).

Z. X. Shu and X. D. Li are at Hubei Provincial Hospital of Traditional Chinese Medicine, Wuhan, 430061, and China Hubei Provincial Hospital of Traditional Chinese Medicine, Wuhan, 430061, China (542067880@qq.com; lixiaodong555@126.com).

R. S. Zhang is at Guanganmen Hospital, China Academy of Chinese Medical Sciences, Beijing 100053, China (runshunzhang@139.com).

J. X. Chen is at Beijing University of Chinese Medicine, Beijing 100029, China (cjx@bucm.edu.cn).

*Correspondence should be addressed to: xzzhou@bjtu.edu.cn

Digital Object Identifier 10.1109/JBHI.2018.2870728

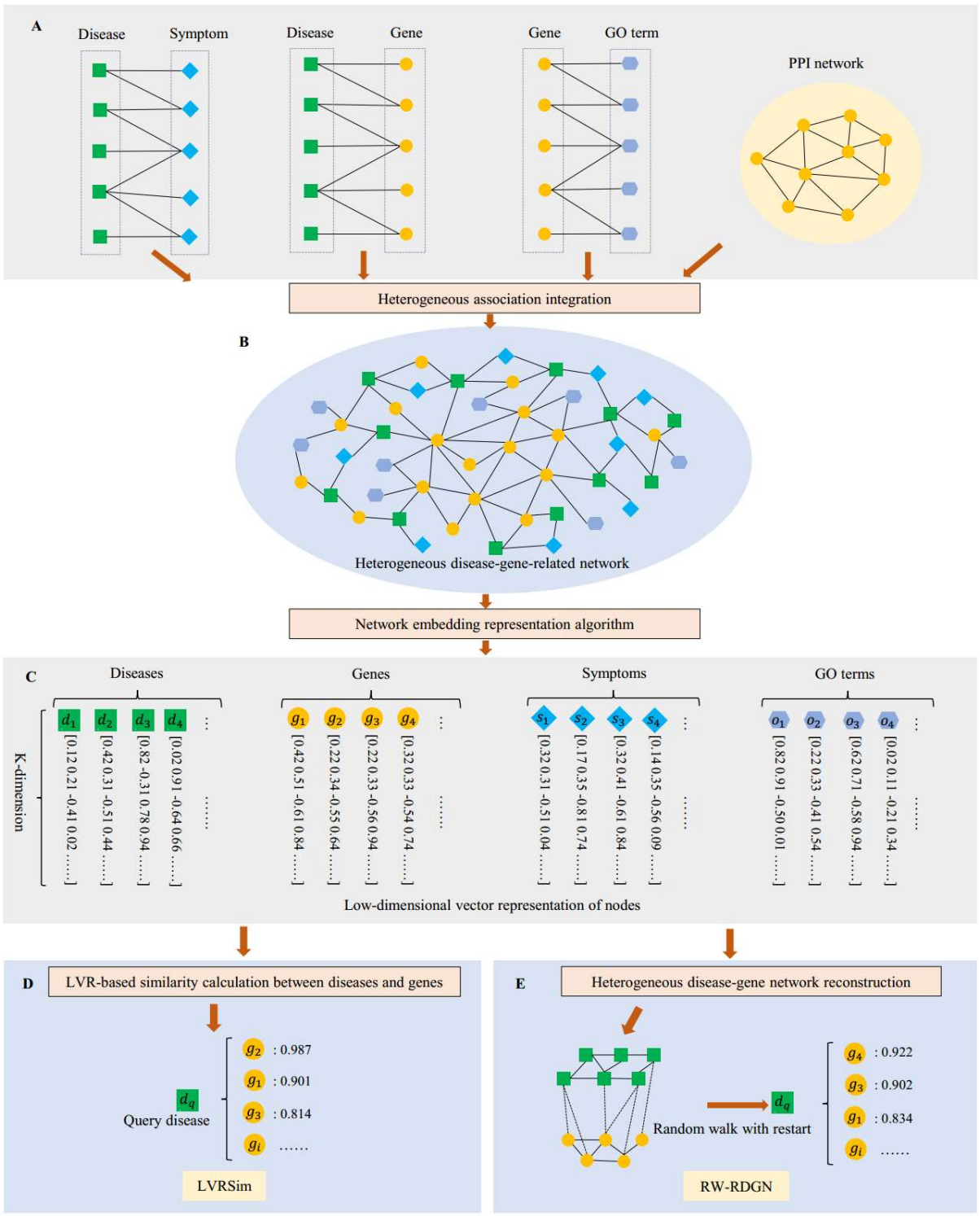


Fig. 1. An overview of HerGePred. Firstly, by integrating disease-symptom and disease-gene associations, GO and protein-protein interactions (PPI) (Fig 1A), a heterogeneous disease-gene-related network (Fig 1B) was built. Then, a network embedding representation algorithm was applied to learn low-dimensional vector representations of the nodes (disease, gene, symptom, and GO) (Fig 1C). Based on the framework, we proposed two specific algorithms, LVR-based similarity prediction (LVRSim, Fig 1D) and random walk with restart on reconstructed heterogeneous disease-gene network (RW-RDGN, Fig 1E), to predict disease genes.

dimensional and complementary information representation, which has a huge advantage in disease gene prediction over homogeneous data. On the one hand, based on these heterogeneous data types, random walk on a heterogeneous network [25] or heterogeneous data fusion [33] were proposed and proven to be feasible schemes. On the other hand, applying embedding method to the discovery of biomedical relationships (e.g., disease associations [34] and drug-target interaction prediction [35]) has been proved with high performance prediction results. However, effectively extracting the

TABLE I
HETEROGENEOUS DISEASE-GENE-RELATED NETWORK

| Networks | Edge types of network | | | | Node types of network | | | |
|---|---|---|---|---|---|---|---|---|
| | Disease-gene associations | Disease-symptom associations | Protein-protein interactions | Gene-GO associations | Disease | Gene | Symptom | GO |
| embDG | √ | × | × | × | √ | √ | × | × |
| embDGS | √ | √ | × | × | √ | √ | √ | × |
| embDGG | √ | × | √ | × | √ | √ | × | × |
| embDGGG | √ | × | √ | √ | √ | √ | × | √ |
| embDGSG | √ | √ | √ | × | √ | √ | √ | × |

The symbols √ and × represent the given networks contain and do not contain the corresponding edges (disease-gene associations, disease-symptom associations, protein-protein interactions, or gene-GO associations) or nodes (diseases, genes, symptoms or GO).

underlying feature representations of disease phenotypes and genes from these heterogeneous data sources to improve prediction performance still needs be investigated.

In this study, we developed a heterogeneous disease-gene-related network embedding representation framework for disease gene prediction (Fig. 1). First, we built a heterogeneous disease-gene-related network (HDGN) that includes disease-gene associations and disease-related and gene-related information. Then, we applied network embedding representation algorithms on the HDGN to obtain low-dimensional vector representations (LVR) of the nodes under a unified dimension. To validate the reliability of the framework, we analyzed the similarity-based overlap distribution of disease pairs and designed an experiment for disease-gene association recovery, the results of which revealed that the LVR of nodes performed well at preserving local and global network structure information in the HDGN. Second, we proposed two specific algorithms, LVR-based similarity prediction (LVRSim) and random walk with restart on a reconstructed heterogeneous disease-gene network (RW-RDGN), to predict disease genes. Finally, we applied 10-fold cross validation and external validation to evaluate our algorithms, which showed that RW-RDGN performed better than the other algorithms.

## II. MATERIALS AND METHODS

### A. Datasets

*Disease-gene associations*. We collected disease-gene associations from two databases: DisGeNet [36] and MalaCards [37]. First, we obtained 130,820 disease-gene associations between 13,074 diseases and 8,947 genes from DisGeNet (i.e. curated disease-gene set), which were used for cross validation. Second, to evaluate the capability in predicting new candidate genes of the proposed algorithms, we collected two independent datasets from DisGeNet that integrated animal model and literature data and MalaCards databases. Meanwhile, we removed the overlapped associations between cross validation and each independent dataset to guarantee true independence. Finally, 430,286 disease-gene associations that exclude curated set among 11,925 diseases and 15,993 genes and 65,905 disease-gene association among 5,783 diseases and 8,045 genes were collected from the DisGeNet and MalaCards, respectively.

*Protein-protein interactions*. We collected 213,888 protein-protein interactions containing 15,964 proteins from Menche et al. [38]. Since the PPI network included most of the protein-protein interactions of pathways, we did not consider the pathway information of proteins when constructing the heterogeneous disease-gene-related network.

*Disease-symptom associations*. Disease-symptom associations were collected from HPO [39] and Orphanet [40] databases. In particular, besides of the disease-symptom associations derived from HPO, we integrated the disease-symptom associations from the Orphanet database as well. To unify disease codes, we manually mapped OMIM and Orphanet disease identifiers to UMLS codes, which resulted in the collection of 99,087 disease-symptom associations between 5,423 distinct disease UMLS codes and 6,540 genes.

*Gene functional annotations.* GO annotations were extracted from STRING 10 [41]. Finally, we collected 218,337 annotation records containing 18,584 genes and 14,204 GO terms.

### B. Heterogeneous Disease-gene-related Network Embedding Representation

Network embedding representation learning is an algorithm for learning low-dimensional feature vectors, and it can effectively preserve the local and global structure information of given network. Network embedding representation methods are useful in many tasks, such as in visualization, label classification and link prediction. In this study, we applied two well-known network embedding algorithms, node2vec [42] and LINE [43], to get a low-dimensional vector representation of nodes in the HDGN.

By integrating disease-gene-related information, we built heterogeneous disease-gene-related networks. According to edge (association) types in the given network, we defined five networks (Table I): (1) embedding disease-gene associations (DGA) (termed embDG); (2) embedding DGA and disease-symptom associations (DSA) (termed embDGS); (3) embedding DGA and protein-protein interaction associations (PPI) (termed embDGG); (4) embedding DGA, PPI and gene-GO term associations (GGA) (termed embDGGG); and (5) embedding DGA, PPI and DSA (termed embDGSG). Given a heterogeneous network $G = (V, E)$, $V$ and $E$ represent nodes and edges of the network, respectively. Then, we applied network embedding representation algorithms (node2vec and LINE) to learn the LVR of nodes in the network. The node $v$ can be mapped to a low-dimensional vector $N(v)$. For any given heterogeneous network, the algorithm can learn the LVR of nodes in the corresponding network.

### C. LVR-based Similarity Calculation of Disease Pairs and Gene Pairs

Based on the heterogeneous network embedding representation framework, we can get the LVR of nodes in the HDGN. The LVR-based cosine similarity of node pairs (disease pairs and gene pairs) can be measured. Taking similarity calculation of disease pairs as examples, given the disease pair $v_{dx}$ and $v_{dy}$, $N(v_{dx})$ and $N(v_{dy})$ are their vector representations. Then, the LVR-based cosine similarity of the disease pair can be measured by (1) as follows:

$$Cos\big(N(v_{dx}), N(v_{dy})\big) = Cos(x, y) = \frac{x \cdot y}{|x| \cdot |y|} \quad (1)$$

Similarly, based on the LVR of genes, the LVR-based cosine similarity of gene pairs can also be measured. For the five defined networks, we can obtain the LVR of diseases and genes and measure the LVR-based cosine similarity of disease pairs and gene pairs. Otherwise, we can also calculate the gene-based and symptom-based cosine similarity of disease pairs.

### D. Similarity-based Overlap Analysis of Disease Pairs

Gene associated with phenotypically similar diseases exhibited functional similarities across different genomic data [25]. To validate the correlations between the LVR-based and gene-based or symptom-based similarities of disease pairs, we quantified the LVR-based average similarity of disease pairs under the corresponding gene-based similarities of disease pairs. The Pearson Correlation Coefficient (PCC) between gene-based and LVR-based similarities of disease pairs was also calculated. Similarly, symptom-based similarity of disease pairs was also calculated and compared to the LVR-based similarity. To further illustrate the accordance results of disease pairs, we compared the overlap results to the results of random control using Fisher- Yates method [44]. We randomly shuffled LVR of diseases for 100 times, and calculated the average ratios of overlap results.

### E. LVR-based Similarity Method to Predict Disease Genes

The low-dimensional vector representations of nodes not only fused local structure information but also considered global structure information of the network, which was verified by the similarity-based overlap results. Therefore, we proposed a LVR-based similarity method to predict disease genes. Mathematically, given the disease node $v_d$ and gene node $v_g$, we can measure the correlation between them by calculating the LVR-based cosine similarity $Cos\big(N(v_d), N(v_g)\big)$ of their vectors $N(v_d)$ and $N(v_g)$. By calculating and sorting the correlation between the query disease and all candidate genes, we can get a ranking list of candidate genes for the query disease.

### F. Random Walk on Reconstructed Disease-gene Network to Predict Disease Genes

Since the LVR of nodes (diseases and genes) in the HDGN can fuse more related information (local and global information of HDGN), LVR-based disease similarity may be a more appropriate similarity measure than the gene-based or symptom-based similarity of disease pairs, which was applicable to the LVR-based similarity of gene pairs. Hence, we proposed a random walk with restart method on the reconstructed heterogeneous disease-gene network (RW-RDGN) to predict disease genes.

First, the LVR-based similarity of disease pairs and gene pairs were calculated. We reconstructed the LVR-based disease-disease network and gene-gene network, in which nodes represent diseases or genes and edges represent disease pairs or gene pairs with a similarity larger than the given threshold. The network may contain many low confident edges with small similarities. We select only $\alpha$ neighbor disease nodes and $\beta$ neighbor gene nodes with the highest similarity for each disease and gene to build more confident disease-disease and gene-gene networks. Given a disease-disease network, a gene-gene network and known disease-gene associations, we built a reconstructed heterogeneous disease-gene network (RDGN), which included a disease layer, a gene layer, and the interconnections between the two layers.

Afterwards, given a query disease, we simulated a random walk with an initial probability $\boldsymbol{p}^{(0)}$. Then, for each next step, the walker would start a new journey with a probability $\theta$ or move to neighbors of the current node with a probability $1-\theta$. When moving to neighbors, the walker would jump from the disease layer to the gene layer or vice versa with a probability $\varphi$ or wander in either disease layer or gene layer with a probability $1-\varphi$. The RDGN is denoted by $\boldsymbol{I} = (\boldsymbol{D}, \boldsymbol{G}, \boldsymbol{R})$, where $\boldsymbol{D} = (d_{ij})_{m\times m}$ is the weight matrix of the disease-disease network, $\boldsymbol{G} = (g_{ij})_{n\times n}$ is the weight matrix of the gene-gene network, and $\boldsymbol{R} = (a_{ij})_{m\times n}$ is the adjacency matrix of the disease-gene network in which $m$ and $n$ are the numbers of disease and genes, respectively. For disease matrix $\boldsymbol{D}$ and gene matrix $\boldsymbol{G}$, the weights of disease-disease or gene-gene edges equal to the similarities between disease nodes or gene nodes. In the matrix $\boldsymbol{R}$, the weights of disease-gene edges equal to 1 (the gene is associated with the disease) or 0 (if not). By row-normalizing $\boldsymbol{D}, \boldsymbol{G}, \boldsymbol{R}$ and $\boldsymbol{R^T}$, we can obtain $\boldsymbol{U} = (u_{ij})_{m\times m}$, $\boldsymbol{V} = (v_{ij})_{n\times n}$, $\boldsymbol{A} = (r_{ij})_{m\times n}$, and $\boldsymbol{B} = (s_{ij})_{n\times m}$, where $u_{ij} = d_{ij}/\sum_{j=1}^{m} d_{ij}$, $v_{ij} = g_{ij}/\sum_{j=1}^{n} g_{ij}$, $r_{ij} = a_{ij}/\sum_{j=1}^{n} a_{ij}$ and $s_{ij} = a_{ij}/\sum_{j=1}^{m} a_{ij}$. Then, the transition matrix $\boldsymbol{T}$ is defined as follows:

$$\boldsymbol{T} = (t_{ij})_{(m+n)\times(m+n)} = \begin{pmatrix} (1-\varphi)\boldsymbol{U} & \varphi\boldsymbol{A} \\ \varphi\boldsymbol{B} & (1-\varphi)\boldsymbol{V} \end{pmatrix} \quad (2)$$

We row-normalized $\boldsymbol{T}$ to get $\boldsymbol{W} = (w_{ij})_{(m+n)\times(m+n)}$, where $w_{ij} = t_{ij}/\sum_{j=1}^{m+n} t_{ij}$. Random walk with restart was simulated by the following iteration equation:

$$\boldsymbol{p}^{(t)} = (1-\theta)\boldsymbol{W}^T\boldsymbol{p}^{(t-1)} + \theta\boldsymbol{p}^{(0)} \quad (3)$$

where the initial probability $\boldsymbol{p}^{(0)}$ is set to $\left(\left(\boldsymbol{u}^{(0)}\right)^T, \left(\boldsymbol{v}^{(0)}\right)^T\right)^T$, and $\boldsymbol{u}^{(0)} = (u^{(0)})_{m\times 1}$ and $\boldsymbol{v}^{(0)} = (v^{(0)})_{n\times 1}$ are the initial probabilities of the disease and gene layers, respectively, which were extracted from disease matrix $\boldsymbol{D}$ and gene matrix $\boldsymbol{G}$. After a number of steps, when $L_1$ norm of probability $\Delta\boldsymbol{p}$ $(= \boldsymbol{p}^{(t+1)} - \boldsymbol{p}^{(t)})$ is smaller than the given threshold, the iteration reached a steady state $\boldsymbol{p}^{(t)}$. The steady-state probability $\boldsymbol{p}^{(t)}$ included two

parts: disease scores $\boldsymbol{u}^{(t)}$ and gene scores $\boldsymbol{v}^{(t)}$. For the above parameters ( $\alpha, \beta, \varphi$ and $\theta$ ), we adopted the optimized parameters that were tuned in the study [45] of herb-target prediction. Sorting all predicted genes by the scores, the top $n$ genes of the ranking list were selected as the candidate genes of the query disease.

### G. Experimental Setting and Evaluation Metrics

We selected curated disease-gene associations from the DisGeNet database as a benchmark dataset and applied the conventional 10-fold cross validation to evaluate the disease-gene prediction algorithms. We adopted several classic disease-gene prediction algorithms, CIPHER [13], PRINCE [14], pgWalk [25], DADA [10] and GUILD [11] as baseline methods. We adopted the default parameters suggested in the original studies in our experimental settings. In addition, if there were multiple versions for the proposed algorithms, we selected the version with best performance as baselines. For example, for CIPHER and GUILD, the prediction results of CIPHER- SP and NetCombo were showed, respectively.

We used precision (PR), recall (RE), F1-score (F1) [46] and association precision (AP) as evaluation metrics in our experiments. Given a test disease set $D$ with $M$ diseases, for every test disease $d \in D$, $T(d)$ represents the test gene set of disease $d$. Given the gene ranking list of disease $d$, we selected the top $i$ genes $R_i(d)$ of the ranking list ( $i = 3$ or $10$ ) as candidate genes. Then, the precision, recall and F1-score in TOP@i can be defined as follows:

$$Precision = \frac{1}{M}\sum_{d \in D} \frac{|T(d) \cap R_i(d)|}{|R_i(d)|} \tag{4}$$

$$Recall = \frac{1}{M}\sum_{d \in D} \frac{|T(d) \cap R_i(d)|}{|T(d)|} \tag{5}$$

$$\mathrm{F1-score} = 2 \cdot \frac{precision \cdot \mathrm{recall}}{precision + recall} \tag{6}$$

Otherwise, for every test disease $d$, the top $k$ genes $R_k(d)$ of the ranking list were also selected ($k$ equals the number of test genes of disease $d$). Then, the association precision can be defined as follows:

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

In general, the more known genes of disease $d$ in $R_k(d)$, the better the performance of the prediction algorithm, that is, a bigger $AP$ indicates better performance ($0 \le AP \le 1$) of the prediction algorithm.

## III. Results

### A. Rationality Validation of Heterogeneous Network Embedding Representation Method

To validate the rationality of the heterogeneous disease-gene-related network embedding representation framework, we analyzed the gene-based (and symptom-based) and LVR-based overlap distribution of disease pairs and designed an experiment of disease-gene association recovery, the results of

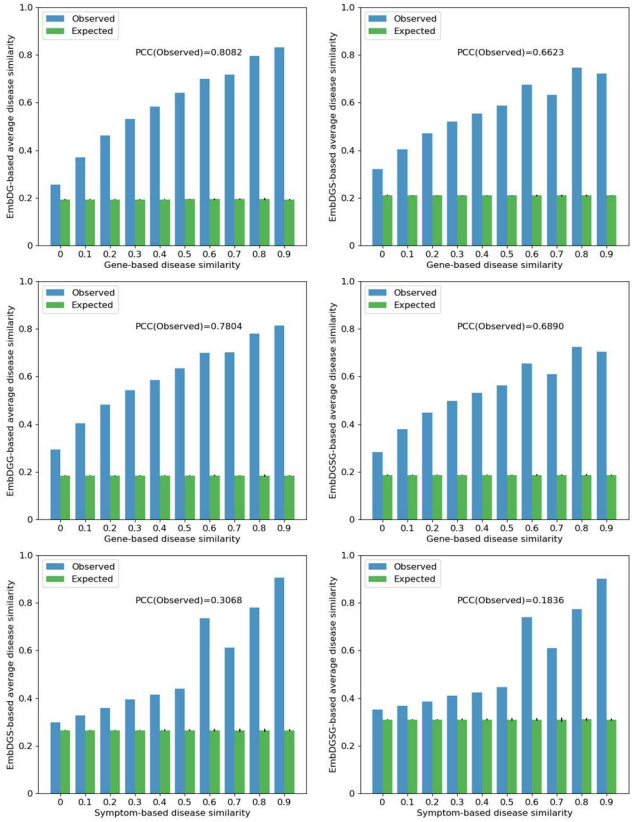


Fig. 2. The similarity-based overlap analysis of disease pairs. We showed the gene-based and LVR-based overlap distribution of disease pairs (Fig 1A, B, C, and D). The horizontal axis represents the gene-based similarity bins of disease pairs and the vertical axis represents the LVR-based average similarity of disease pairs under the corresponding similarity bins. The mazarine and green bars represent observed and expected overlap results, respectively. Fig 1E and 1F showed the symptom-based and LVR-based overlap distribution of disease pairs. The horizontal axis represents the symptom-based similarity bins of disease pairs.

which indicated that the low-dimensional vector representation of nodes exhibits good performance in preserving the local and global network structure of the HDGN.

Based on the obtained networks, we compared actual overlap distribution with that of random permutation. The results (Fig. 2) showed there exists the consistency of disease pairs with gene-based similarity (shared genes) and LVR-based similarity. For example, for the embDG network, the disease pairs with weak gene-based similarity (0.1-0.2) have weak LVR-based average similarity (0.37; Expected: 0.19±0.0015). Nevertheless, for the disease pairs with high gene similarity (0.9-0.1), their LVR-based average similarity are also high (0.83; Expected: 0.19±0.0021), which indicated that the disease pairs with more shared genes are more likely to have higher LVR-based similarity (PCC=0.81). Meanwhile, we measured the accordance between the symptom-based and LVR-based similarity of disease pairs and found that positive correlations exist as well (embDGS: PCC=0.31; embDGSG: PCC=0.18).

Since the LVR-based similarity method can be used to measure the correlation of given disease-gene pairs, we try to recover the known genes of diseases using LVRSim when retaining all known disease-gene associations. We listed the

TABLE II
THE EXPERIMENTAL RESULTS OF DISEASE-GENE ASSOCIATION RECOVERY

| Networks | Algorithms | AP | TOP@3 | | TOP@10 | |
|---|---|---|---|---|---|---|
| | | | PR | RE | PR | RE |
| EmbDG | LVRSim_N | 0.594 | 0.626 | 0.778 | 0.341 | 0.897 |
| | LVRSim_L | 0.162 | 0.159 | 0.250 | 0.079 | 0.272 |
| EmbDGS | LVRSim_N | 0.690 | 0.623 | 0.767 | **0.346** | 0.892 |
| | LVRSim_L | 0.214 | 0.212 | 0.212 | 0.113 | 0.244 |
| EmbDGG | LVRSim_N | **0.762** | **0.630** | 0.780 | **0.346** | **0.899** |
| | LVRSim_L | 0.259 | 0.304 | 0.283 | 0.152 | 0.309 |
| EmbDGSG | LVRSim_N | 0.597 | 0.616 | 0.757 | 0.341 | 0.885 |
| | LVRSim_L | 0.244 | 0.281 | 0.213 | 0.147 | 0.251 |
| EmbDGGG | LVRSim_N | 0.743 | **0.630** | **0.781** | 0.345 | **0.899** |
| | LVRSim_L | 0.261 | 0.331 | 0.308 | 0.156 | 0.334 |

The symbols AP, PR, and RE represent the association precision, precision and recall metrics, respectively. TOP@3 and TOP@10 represent the top 3 and 10 candidate genes of a ranked gene list, respectively. The meaning of these symbols are applied to the following tables as well.

recovery performance of each algorithm (LVRSim_N based on node2vec and LVRSim_L based on LINE) under each network (Table II). Among all the results, the LVRSim_N algorithm with the embDGG network yielded the best performance (AP=0.762; PR=0.630 and RE=0.780 for TOP@3; PR=0.495 and RE=0.840 for TOP@10). For the algorithm LVRSim_N, the embDGG network exhibited the best performance, whereas for the algorithm LVRSim_L, the embDGGG network exhibited the best performance (AP=0.261; PR=0.331 and RE=0.308 for TOP@3; PR=0.346 and RE=0.899 for TOP@10). For all the networks, the LVRSim_L algorithm showed low performance (AP<0.3; PR<0.4 and RE <0.4 for TOP@3; PR<0.2 and RE<0.4 for TOP@10) in disease-gene recovery. However, the LVRSim_N algorithm exhibited high performance (AP>0.5; PR>0.6 and RE>0.7 for TOP@3; PR>0.33 and RE>0.87 for TOP@10), which indicated that LVRSim_N had better performance in the embedding representation of the HDGNs. For different networks, the results showed that the algorithm with embDGS and embDGG networks had better performance than the algorithm with embDG network, which indicates that considering more information (e.g., disease-symptom associations or PPI network) can improve the recovery performance in disease-gene associations. Compared with the embDGS network (AP=0.690; PR=0.623 and RE=0.767 for TOP@3; PR=0.346 and RE=0.892 for TOP@10), the embDGG network had better performance, which indicated that the PPI network would be more effective for improving the recovery performance in disease-gene associations. However, when fusing more information, such as in the embDGGG network (AP=0.743; PR=0.630 and RE=0.781 for TOP@3; PR=0.345 and RE=0.899 for TOP@10), there was no obvious improvement in association recovery. Meanwhile, LVRSim_N with embDGSG network (AP=0.597) had lower performance than the embDGG and embDGS networks, which indicated that mutual interference may exist in disease-symptom associations and PPI network information.

### *B. LVR-based Similarity Model (LVRSim and RW-RDGN) to Predict Disease Genes*

Based on the heterogeneous disease-gene-related network embedding representation framework, we proposed the LVRSim and RW-RDGN algorithms to predict disease genes. We collected a benchmark dataset and two external datasets to evaluate the performance of the prediction algorithms. For LVRSim and RW-RDGN algorithms, we adopted the four networks, embDG, embDGS, embDGG and embDGSG, as experimental networks. The experimental results of disease gene recovery indicated LVRSim with node2vec performs much better than LINE, so the proposed LVRSim and RW-RDGN are based on node2vec method in the following experiments.

We showed the performance of disease gene prediction algorithms (Table III) and marked the best performance among baseline algorithms and all algorithms by italic and bold text, respectively. First, among these baseline algorithms, pgWalk obtained the best performance (AP: 0.258; PR=0.222 and RE=0.305 for TOP@3). We conducted Student's t test [47] on best prediction results of baseline and our algorithms to indicate the statistical significance with p-value (values in the bracket of Table III). The RW-RDGN algorithm with the embDGG network yielded the highest performance: AP improved by 13.80% (P=1.42E-13), PR and RE improved by 9.40% (P=1.80E-10) and 6.36% (P=8.19E-08), respectively, for TOP@3, PR and RE improved by 18.10% (P=6.39E-14) and 14.66% (P=2.00E-14), respectively, for TOP@10 than pgWalk, which indicated that compared with the current classic algorithms, RW-RDGN had better performance in disease gene prediction. In detail, as for different networks, in accordance with the experimental results of disease-gene association recovery, RW-RDGN with the embDGG network obtained the best performance. In the term of AP, the LVRSim algorithm with the embDGG network showed high performance (AP=0.239). However, in the terms of precision and recall for TOP@3, the LVRSim with embDGS network showed high performance (PR=0.213 and RE=0.285 for TOP@3). The performance of the LVRSim algorithm was better than the performance of PRINCE and CIPHER but was worse than that of pgWalk, which indicated that random walk on the heterogeneous disease-gene network had better prediction performance than LVRSim. Therefore, the RW-RDGN algorithm integrating LVR-based similarity and random walk with restart would be a potential method for disease gene prediction that had better performance than pgWalk. In addition, to fully show the performance of prediction algorithms, we showed the prediction results with two additional test disease sets, one filtered from the terminology type of diseases belonging to “T047” and another for having at least ten candidate genes. The prediction results (Table S1 and S2 at https://github.com/yangkuoone/HerGePred/tree/master/prediction_results) indicated the proposed algorithms perform better than the baseline algorithms. For example, in two additional experiments, RW-RDGN yielded the highest performance: PR and RE improved by 14.31% and 18.21%, 27.09% and 19.26% for TOP@3 than pgWalk (best in baselines), respectively.

TABLE III
PERFORMANCE COMPARISON OF DISEASE GENE PREDICTION ALGORITHMS

| Networks | Algorithms | AP | TOP@3 | | | TOP@10 | | |
|---|---|---|---|---|---|---|---|---|
| | | | PR | RE | F1 | PR | RE | F1 |
| - | pgWalk | *0.258±0.003 (1.42E-13)* | *0.222±0.003 (1.80E-10)* | *0.305±0.005 (8.19E-08)* | *0.219±0.003 (9.06E-09)* | *0.105±0.001 (6.39E-14)* | *0.416±0.003 (2.00E-14)* | *0.145±0.002 (1.22E-14)* |
| - | PRINCE | 0.019±0.003 | 0.006±0.003 | 0.008±0.007 | 0.005±0.004 | 0.005±0.002 | 0.020±0.011 | 0.006±0.003 |
| - | CIPHER | 0.003±0.002 | 0.002±0.001 | 0.003±0.003 | 0.002±0.001 | 0.001±0.0003 | 0.006±0.003 | 0.001±0.001 |
| - | DADA | 0.087±0.001 | 0.010±0.006 | 0.021±0.011 | 0.012±0.007 | 0.004±0.001 | 0.029±0.009 | 0.007±0.002 |
| - | GUILD | 0.091±0.007 | 0.010±0.003 | 0.021±0.010 | 0.013±0.004 | 0.004±0.002 | 0.030±0.011 | 0.007±0.003 |
| EmbDG | LVRSim | 0.138±0.003 | 0.152±0.004 | 0.242±0.006 | 0.164±0.003 | 0.070±0.001 | 0.338±0.004 | 0.104±0.001 |
| EmbDGS | LVRSim | 0.216±0.004 | 0.213±0.003 | 0.285±0.004 | 0.207±0.002 | 0.098±0.002 | 0.394±0.003 | 0.136±0.001 |
| EmbDGG | LVRSim | 0.239±0.004 | 0.200±0.003 | 0.264±0.006 | 0.191±0.003 | 0.101±0.002 | 0.386±0.007 | 0.136±0.002 |
| EmbDGSG | LVRSim | 0.166±0.003 | 0.184±0.002 | 0.256±0.007 | 0.184±0.003 | 0.078±0.002 | 0.347±0.007 | 0.113±0.002 |
| EmbDG | RW-RDGN | 0.239±0.005 | 0.211±0.002 | 0.316±0.003 | 0.220±0.002 | 0.112±0.002 | 0.470±0.006 | 0.158±0.002 |
| EmbDGS | RW-RDGN | 0.268±0.005 | 0.232±0.002 | 0.315±0.004 | 0.227±0.002 | 0.118±0.002 | 0.465±0.005 | 0.163±0.002 |
| EmbDGG | RW-RDGN | **0.294±0.005** | **0.243±0.003** | **0.325±0.004** | **0.233±0.003** | **0.124±0.002** | **0.477±0.008** | **0.167±0.003** |
| EmbDGSG | RW-RDGN | 0.184±0.003 | 0.204±0.003 | 0.299±0.004 | 0.211±0.003 | 0.092±0.001 | 0.428±0.006 | 0.136±0.002 |

The italic and bold values represent the best performance in baseline algorithms and all prediction algorithms, respectively. The values in brackets are P-values with t-test of the baseline and our algorithms with best performance.

TABLE IV
PERFORMANCE COMPARISON OF PREDICTION ALGORITHMS ON EXTERNAL DATASET

| Networks | Algorithms | TOP@3 (MC) | | TOP@10 (MC) | | TOP@3 (DGN) | | TOP@10 (DGN) | |
|---|---|---|---|---|---|---|---|---|---|
| | | PR | RE | PR | RE | PR | RE | PR | RE |
| - | pgWalk | 0.105±0.006 | 0.027±0.002 | 0.076±0.004 | 0.061±0.003 | 0.135±0.003 | 0.018±0.001 | 0.116±0.002 | 0.039±0.002 |
| - | PRINCE | 0.024±0.005 | 0.004±0.002 | 0.026±0.005 | 0.015±0.005 | 0.067±0.007 | 0.002±0.001 | 0.076±0.003 | 0.011±0.001 |
| - | CIPHER | 0.013±0.002 | 0.002±0.000 | 0.007±0.001 | 0.004±0.001 | 0.022±0.003 | 0.001±0.000 | 0.015±0.002 | 0.002±0.001 |
| - | DADA | 0.113±0.009 | 0.030±0.004 | *0.080±0.004 (*3.31E-08*)* | *0.071±0.005 (3.13E-04)* | ***0.147±0.005*** *(1.66E-04)* | 0.021±0.002 | 0.121±0.004 | *0.044±0.003 (0.319)* |
| - | GUILD | *0.113±0.007 (0.053)* | *0.033±0.004 (0.349)* | *0.080±0.004 (*3.69E-07*)* | 0.070±0.006 (5.49E-04) | 0.144±0.006 | ***0.022±0.002*** *(1.42E-02)* | ***0.124±0.004*** ***(0.835)*** | *0.044±0.004 (0.477)* |
| EmbDG | LVRSim | 0.099±0.005 | 0.028±0.002 | 0.073±0.005 | 0.065±0.005 | 0.083±0.003 | 0.013±0.001 | 0.069±0.002 | 0.030±0.002 |
| EmbDGS | LVRSim | 0.092±0.007 | 0.025±0.002 | 0.070±0.004 | 0.062±0.004 | 0.081±0.003 | 0.012±0.001 | 0.071±0.002 | 0.030±0.002 |
| EmbDGG | LVRSim | 0.111±0.007 | 0.031±0.002 | 0.088±0.004 | 0.078±0.004 | 0.100±0.003 | 0.014±0.001 | 0.089±0.002 | 0.035±0.003 |
| EmbDGSG | LVRSim | 0.099±0.005 | 0.026±0.002 | 0.073±0.004 | 0.062±0.003 | 0.092±0.003 | 0.012±0.001 | 0.080±0.002 | 0.030±0.002 |
| EmbDG | RW-RDGN | 0.117±0.005 | 0.033±0.002 | 0.094±0.006 | 0.080±0.005 | 0.123±0.004 | 0.018±0.002 | 0.116±0.003 | 0.044±0.003 |
| EmbDGS | RW-RDGN | 0.114±0.007 | 0.032±0.002 | 0.094±0.004 | 0.080±0.002 | 0.129±0.005 | 0.018±0.002 | 0.119±0.002 | **0.045±0.003** |
| EmbDGG | RW-RDGN | **0.121±0.006** | **0.034±0.002** | **0.097±0.004** | **0.083±0.005** | 0.132±0.005 | 0.018±0.003 | **0.124±0.003** | **0.045±0.003** |
| EmbDGSG | RW-RDGN | 0.121±0.008 | 0.033±0.001 | 0.096±0.004 | 0.080±0.003 | 0.136±0.003 | 0.018±0.002 | 0.120±0.002 | 0.043±0.003 |

The symbols MC and DGN represent external validation of MalaCards and DisGeNet, respectively.

However, there are candidate genes that are not in the test dataset but are still likely to be potential disease genes that have been recorded in other databases. Hence, we showed the performance of prediction algorithms (Table IV) on an external dataset (MalaCards and DisGeNet). From the MalaCards dataset, RW-RDGN with embDGG obtained higher performance than GUILD (best in baselines): PR improved by 7.07% (P=0.053) for TOP@3; PR and RE improved by 21.34% (P=3.69E-07) and 17.59% (P=5.49E-04), respectively, for TOP@10. From the DisGeNet dataset, RW-RDGN with embDGG and GUILD (best in baselines) obtained similar performance. As a whole, our method obtained higher performance than the baseline algorithms for both cross-validations on benchmark data and external validations.

### *C. Case Study: the Top 10 Predicted Genes of Specific Diseases*

To further illustrate the biological insights of our algorithms, we evaluated the prediction results of RW-RDGN with the embDGG network of six diseases: amyotrophic lateral sclerosis

TABLE V
PREDICTION PERFORMANCE OF SPECIFIC DISEASES

| Disease (CUI) | TOP@10 | | |
|---|---|---|---|
| | PR | PR (MC) | PR (DGN) |
| Amyotrophic lateral sclerosis (C0002736) | 0.4 | 0.1 | 0.2 |
| Anemia (C0002871) | 0.4 | 0.3 | 0.2 |
| Renal cell carcinoma (C0007134) | 0.4 | 0.3 | 0.4 |
| Idiopathic pulmonary arterial hypertension (C3203102) | 0.4 | 0.1 | 0.3 |
| Hypothyroidism (C0020676) | 0.5 | 0.1 | 0.1 |
| Pulmonary hypertension (C0020542) | 0.5 | 0.2 | 0.2 |

TABLE VI
TOP 10 CANDIDATE GENES OF SPECIFIC DISEASES

| Rank | Amyotrophic lateral sclerosis (C0002736) | Anemia (C0002871) | Renal cell carcinoma (C0007134) | Idiopathic pulmonary arterial hypertension (C3203102) | Hypothyroidism (C0020676) | Pulmonary hypertension (C0020542) |
|---|---|---|---|---|---|---|
| 1 | GRN (MC, DGN) | HFE2 (MC) | EP300 (TS) | ENG (MC, DGN) | TPO (TS) | CSF2RB |
| 2 | TBK1 (TS) | PRF1 (TS) | FHIT (MC, DGN) | DLL4 | LHX3 (RF3) | ENG (MC, DGN) |
| 3 | DAO (TS) | SLC7A7 (TS) | DIRC3 (MC) | TNNT2 | NKX2-1 (MC, DGN) | NOS2 (MC, DGN) |
| 4 | PRPH (TS) | HFE (MC, DGN) | HSPBAP1 (MC, DGN) | MEGF10 | HESX1 (TS) | COL1A1 (TS) |
| 5 | PLA2G6 | BCS1L | CREBBP | KCNK3 (TS) | THRB (TS) | SMAD9 (TS) |
| 6 | VAPB (TS) | TFR2 (MC, DGN) | TP53 (TS) | EIF2AK4 (DGN) | SLC5A5 (TS) | ACTA2 (TS) |
| 7 | CHD2 | TERT (TS) | PTEN (TS) | LIFR (TS) | KCNJ10 (TS) | CACNA1D (TS) |
| 8 | HNRNPA2B1 (DGN) | ERCC4 (RF1, RF2) | DCC (TS) | LIPT1 (TS) | GLI2 (RF4) | TGFB2 |
| 9 | POLG | WRAP53 (TS) | TMEM127 (DGN) | GJA1 (TS) | OTX2 (RF3) | MFAP5 |
| 10 | FTL | IVD | KRAS (DGN) | MIR204 (DGN) | TBX3 | PAM16 (TS) |

The symbols TS, MC, DGN, and RF (1-4) represent that the associations between candidate genes and given diseases are validated by the benchmark test dataset, MalaCards dataset, DisGeNet dataset, or biomedical literatures, respectively.

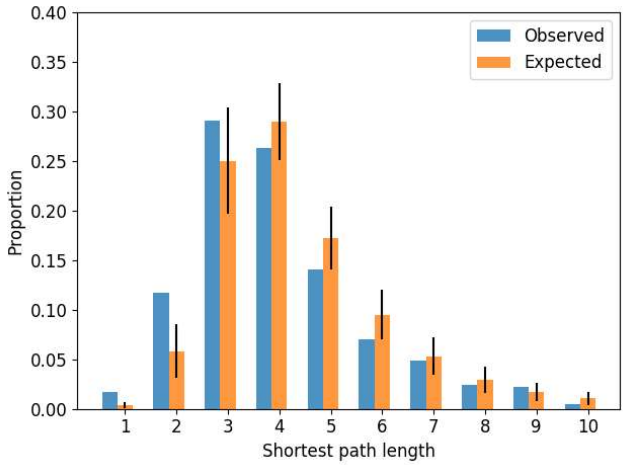


Fig. 3. The distribution of shortest path length between top 10 predicted genes and known genes of anemia in PPI network. The horizontal axis represents shortest path length, and the vertical axis represents the proportion with respect to the length. The comparison of observed and expected results indicated the predicted genes of anemia are closer to the known genes than random results.

(CUI: C0002736), anemia (CUI: C0002871), renal cell carcinoma (CUI: C0007134), idiopathic pulmonary arterial hypertension (CUI: C3203102), hypothyroidism (CUI: C0020676) and pulmonary hypertension (CUI: C0020542) as examples. The prediction performance (Table V) and the top 10 predicted genes (Table VI) of these diseases were listed. For example, for renal cell carcinoma, the four candidate genes (marked by TS): EP300 (rank=1), TP53 (rank=6), PTEN (rank=7) and DCC (rank=8) in the top 10 candidate genes are the known genes in the benchmark (precision=0.4). In addition, FHIT (rank=2), DIRC3 (rank=3) and HSPBAP1 (rank=4) are known genes of renal cell carcinoma in the MalaCards database (marked by MC). Besides of CREBBP (rank=5), which is not verified, the two remaining genes, TMEM127 (rank=9) and KRAS (rank=10), were included in the DisGeNet literature dataset (marked by DGN). For hypothyroidism, the genes: TPO (rank=1), HESX1 (rank=4), THRB (rank=5), SLC5A5 (rank=6) and KCNJ10 (rank=7) were well-known associated genes (precision=0.5). Furthermore, the novel gene NKX2-1 (rank=2) appeared in both MalaCards database and DisGeNet literature database. To fully evaluate the candidate genes, we manually searched the published biomedical literature to perform final confirmations. Recent study [48] (RF3) showed that LHX3 and OTX2 were associated with hypothyroidism. Similarly, another research [49] (RF4) indicated that GLI2 is associated with hypothyroidism. In addition, the association between anemia and ERCC4 also has been verified by Manandhar et al. [50] (RF1) and Bogliolo et al. [51] (RF2). Meanwhile, we conducted the shortest path analysis [45] between the top 10 predicted genes and known genes of anemia. The result (Fig 3) indicated the candidate genes are closer to the known genes compared to random expectations. The reason is that network embedding method could make the closer genes located in the network obtain more similar vector representations. These results indicated that a high degree of the novel genes predicted by our methods might be true associated genes for a given disease.

## IV. DISCUSSION

In genetic research, network propagation methods, e.g., random walks [52], information diffusion [53] and electrical resistance [54], which act as a universal amplifier of genetic associations [55], have been applied successfully to identify gene functions [56], disease characterization [57], and drug targets [58]. Meanwhile, network embedding representation methods, such as DeepWalk [59], LINE [43] and node2vec [42], have been widely applied in network classification [60] and link prediction [42]. Here, we developed a heterogeneous disease-gene-related network embedding representation framework for disease gene prediction. The experimental results indicated that RW-RDGN performed better than current existing methods of disease-gene prediction.

The advantages of RW-RDGN are attributed to several aspects. First, we integrated multiple types of disease-related and gene-related data into a heterogeneous disease-gene-related network and applied the network embedding representation method to obtain low-dimensional vectors of the nodes, which fused and preserved local and global structure information of the given network. Second, compared with the original disease-gene network in pgWalk, the reconstructed disease-gene network based on LVR-based similarity of nodes can take full

advantage of heterogeneous network information. Compared with the algorithm LINE that capture the first-order or second-order proximity between the nodes of network, the node2vec with a biased random walk that efficiently explores diverse neighborhoods can obtain richer vector representations. Finally, we made full use of the advantages of the network embedding representation and network propagation method (i.e., random walk with restart in the heterogeneous network) to improve prediction performance.

Two potential applications of our disease gene prediction framework could be considered. On the one hand, the prediction results can be used to guide the selection of candidate genes in diseases that have not been studied yet or find new disease-causing genes in common diseases, which would benefit the treatment of complex diseases, such as cancer [61]. On the other hand, the HerGePred method provided a universal heterogeneous network embedding prediction framework that can be extended easily to other prediction tasks, such as disease-single nucleotide polymorphism (SNP) identification and disease-microRNA identification.

Our approach can be improved with respect to the following aspects. First, in our framework, the embedding representation methods (node2vec and LINE) that we used were adept at processing homogeneous node types. Designing a specific network embedding representation method for heterogeneous disease-gene-related networks is still a challenge. In general, for the specific prediction task (e.g., disease genes), different node or edge types should be distinguished with different weights. Therefore, our further experiments would be focusing on adapting the strengths of typical nodes (e.g., diseases and genes) and their edges (e.g., disease-gene associations that be used to train) for learning the network embedding representations. Second, the appropriate data fusion with more useful biomedical features, such as drug-target network, tissue specific network and gene expressions, can improve the prediction performance of HerGePred, which would be investigated in our future work as well. Meanwhile, investigating the contribution of hierarchical structures of different ontologies (e.g., symptom ontology, Gene Ontology and disease ontology) for disease-gene prediction will be a meaningful work in future. Finally, disease-gene prediction could be modelled as a classification task if the different links between diseases and genes have been annotated with positive and negative labels. In this direction, we could adopt the low-dimensional features of node pairs (e.g., disease-gene associations) [42] to represent links and use various classification methods to train the prediction models. However, it should be noted that the disjoint cross validation [62-64] would be a necessary option to assure more reliable prediction results.

## REFERENCES


[1] D. H. Le, and V. T. Dang, "Ontology-based disease similarity network for disease gene prediction," *Vietnam Journal of Computer Science,* vol. 3, no. 3, pp. 1-9, 2016.

[2] B. Calvo, N. López-Bigas, S. J. Furney, P. Larrañaga, and J. A. Lozano, "A partially supervised classification approach to dominant and recessive human disease gene prediction," *Computer Methods & Programs in Biomedicine,* vol. 85, no. 3, pp. 229-37, 2007.

[3] C. J. Wolfe, I. S. Kohane, and A. J. Butte, "Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks," *Bmc Bioinformatics,* vol. 6, no. 1, pp. 1-10, 2005.

[4] 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.

[5] M. Oti, B. Snel, M. A. Huynen, and H. G. Brunner, "Predicting disease genes using protein–protein interactions," *Journal of Medical Genetics,* vol. 43, no. 8, pp. 691, 2006.

[6] D. Horn, "Walking the interactome for prioritization of candidate disease genes," *American Journal of Human Genetics,* vol. 82, no. 4, pp. 949-58, 2008.

[7] J. Sun, J. C. Patra, and Y. Li, "Functional link artificial neural network-based disease gene prediction," in International Joint Conference on Neural Networks, Atlanta, Georgia, USA, 2009, pp. 3003-3010.

[8] S. Navlakha, and C. Kingsford, "The power of protein interaction networks for associating genes with diseases," *Bioinformatics,* vol. 26, no. 8, pp. 1057, 2010.

[9] H. Zhou, and J. Skolnick, "A knowledge-based approach for predicting gene-disease associations," *Bioinformatics*, no. 18, pp. btw358, 2016.

[10] S. Erten, G. Bebek, R. M. Ewing, and M. Koyutürk, "D A D A : Degree-Aware Algorithms for Network-Based Disease Gene Prioritization," *Biodata Mining,* vol. 4, no. 1, pp. 19-19, 2011.

[11] E. Guney, and B. Oliva, "Exploiting Protein-Protein Interaction Networks for Genome-Wide Disease-Gene Prioritization," *Plos One,* vol. 7, no. 9, pp. e43557, 2012.

[12] K. Lage, E. O. Karlberg, Z. M. Størling, P. Í. Ólason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tümer, F. Pociot, and N. Tommerup, "A human phenome-interactome network of protein complexes implicated in genetic disorders," *Nature Biotechnology,* vol. 25, no. 3, pp. 309, 2007.

[13] X. Wu, R. Jiang, M. Q. Zhang, and S. Li, "Network-based global inference of human disease genes," *Molecular Systems Biology,* vol. 4, no. 1, pp. 189, 2008.

[14] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation," *Plos Computational Biology,* vol. 6, no. 1, pp. e1000641, 2010.

[15] X. Wu, Q. Liu, and R. Jiang, "Align human interactome with phenome to identify causative genes and networks underlying disease families," *Bioinformatics,* vol. 25, no. 1, pp. 98-104, 2009.

[16] Y. Li, and J. C. Patra, "Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network," *Bioinformatics,* vol. 26, no. 9, pp. 1219, 2010.

[17] 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 Systems Biology,* vol. 5, no. 1, pp. 1-11, 2011.

[18] 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 Research,* vol. 34, no. 19, pp. e130, 2006.

[19] S. Karni, H. Soreq, and R. Sharan, "A network-based method for predicting disease-causing genes," *Journal of Computational Biology A Journal of Computational Molecular Cell Biology,* vol. 16, no. 2, pp. 181-189, 2009.

[20] P. G. Sun, L. Gao, and S. Han, "Prediction of human disease-related gene clusters by clustering analysis," *International Journal of Biological Sciences,* vol. 7, no. 1, pp. 61-73, 2011.

[21] S. Aerts, D. Lambrechts, S. Maity, L. P. Van, B. Coessens, S. F. De, L. C. Tranchevent, M. B. De, P. Marynen, and B. Hassan, "Gene prioritization through genomic data fusion," *Nature Biotechnology,* vol. 24, no. 5, pp. 537, 2006.

[22] P. Radivojac, K. Peng, W. T. Clark, B. J. Peters, A. Mohan, S. M. Boyle, and S. D. Mooney, "An integrated approach to inferring gene-disease associations in humans," *Proteins-structure Function & Bioinformatics,* vol. 72, no. 3, pp. 1030-1037, 2008.

[23] B. Linghu, E. S. Snitkin, Z. Hu, X. Yu, and C. Delisi, "Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network," *Genome Biology,* vol. 10, no. 9, pp. 1-17, 2009.

[24] F. Mordelet, and J. P. Vert, "ProDiGe: prioritization of disease genes with multitask machine learning from positive and unlabeled examples," *Bmc Bioinformatics,* vol. 12, no. 1, pp. 1-15, 2011.

[25] R. Jiang, "Walking on multiple disease-gene networks to prioritize candidate genes," *Journal of Molecular Cell Biology,* vol. 7, no. 3, pp. 214, 2015.

[26] F. L, v. B. H, F. L, d. J. ED, E.-P. M, and W. C, "Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes," *American Journal of Human Genetics,* vol. 78, no. 6, pp. 1011, 2006.

[27] S. Zickenrott, V. E. Angarica, B. B. Upadhyaya, and S. A. Del, "Prediction of disease-gene-drug relationships following a differential network analysis," *Cell Death & Disease,* vol. 7, no. 1, pp. e2040, 2016.

[28] J. Freudenberg, and P. Propping, "A similarity-based method for genome-wide prediction of disease-relevant human genes," *Bioinformatics,* vol. 18 Suppl 2, no. suppl_2, pp. S110-5, 2002.

[29] M. A. Care, J. R. Bradford, C. J. Needham, A. J. Bulpitt, and D. R. Westhead, "Combining the interactome and deleterious SNP predictions to improve disease gene identification," *Human Mutation,* vol. 30, no. 3, pp. 485-92, 2009.

[30] X. Zeng, Y. Liao, Y. Liu, and Q. Zou, "Prediction and validation of disease genes using HeteSim Scores," *IEEE/ACM Transactions on Computational Biology & Bioinformatics,* vol. 14, no. 3, pp. 687, 2017.

[31] S. B. U. Martin, N. Nagarajan, T. Ambuj, J. O. Woods, I. S. Dhillon, and E. M. Marcotte, "Correction: prediction and validation of gene-disease associations using methods inspired by social network analyses," *Plos One,* vol. 8, no. 5, pp. e58977, 2013.

[32] S. Ballouz, J. Y. Liu, M. Oti, B. Gaeta, D. Fatkin, M. Bahlo, and M. A. Wouters, "Candidate disease gene prediction using Gentrepid: application to a genome-wide association study on coronary artery disease," *Molecular Genetics & Genomic Medicine,* vol. 2, no. 1, pp. 44-57, 2014.

[33] S. Elshal, L. C. Tranchevent, A. Sifrim, A. Ardeshirdavani, J. Davis, and Y. Moreau, "Beegle:from literature mining to disease-gene discovery," *Nucleic Acids Research,* vol. 44, no. 2, pp. e18-e18, 2016.

[34] D. Gligorijevic, J. Stojanovic, N. Djuric, V. Radosavljevic, M. Grbovic, R. J. Kulathinal, and Z. Obradovic, "Large-Scale Discovery of Disease-Disease and Disease-Gene Associations," *Scientific Reports,* vol. 6, pp. 32404, 2016.

[35] Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, J. Peng, L. Chen, and J. Zeng, "A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information," *Nature Communications,* vol. 8, no. 1, 2017.

[36] J. Piñero, N. Queraltrosinach, À. Bravo, J. Deupons, A. Bauermehren, M. Baron, F. Sanz, and L. I. Furlong, "DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes," *Database the Journal of Biological Databases & Curation,* vol. 2015, no. 3, pp. bav028, 2015.

[37] N. Rappaport, M. Twik, I. Plaschkes, R. Nudel, T. I. Stein, J. Levitt, M. Gershoni, C. P. Morrey, M. Safran, and D. Lancet, "MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search," *Nucleic Acids Research,* vol. 45, no. Database issue, pp. D877-D887, 2017.

[38] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A. L. Barabási, "Disease networks. Uncovering disease-disease relationships through the incomplete interactome," *Science,* vol. 347, no. 6224, pp. 1257601, 2015.

[39] S. Köhler, N. A. Vasilevsky, M. Engelstad, E. Foster, J. Mcmurry, S. Aymé, G. Baynam, S. M. Bello, C. F. Boerkoel, and K. M. Boycott, "The Human Phenotype Ontology in 2017," *Nucleic Acids Research,* vol. 45, no. Database issue, pp. D865-D876, 2017.

[40] R. Mangon, J. J. Sikkens, M. Teeuw, and P. D. M. C. Cornel, "Orphanet," no. 1, pp. 139-152, 2006.

[41] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, and P. Bork, "The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible," *Nucleic Acids Research,* vol. 45, no. Database issue, pp. D362-D368, 2017.

[42] A. Grover, and J. Leskovec, "node2vec: Scalable Feature Learning for Networks," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, 2016, pp. 855-864.

[43] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, and Q. Mei, "LINE: Large-scale Information Network Embedding," in Proceedings of the 24th International Conference on World Wide Web, Florence, Italy, pp. 1067-1077.

[44] R. A. Fisher, and F. Yates, "Statistical tables for biological, agricultural and medical research," *Canadian Journal of Comparative Medicine & Veterinary Science,* vol. 22, no. 1, pp. 8, 1958.

[45] K. Yang, G. Liu, N. Wang, R. Zhang, J. Yu, J. Chen, and X. Zhou, "Heterogeneous network propagation for herb target identification," *Bmc Medical Informatics & Decision Making,* vol. 18, no. 1, pp. 17, 2018.

[46] D. Billsus, and M. J. Pazzani, "Learning collaborative information filters," in Proceedings of the Fifteenth International Conference on Machine Learning, San Francisco, CA, USA, 1998, pp. 46-54.

[47] J. F. Box, "Guinness, Gosset, Fisher, and Small Samples," *Statistical Science,* vol. 2, no. 1, pp. 45-52, 1987.

[48] N. Schoenmakers, K. S. Alatzoglou, V. K. Chatterjee, and M. T. Dattani, "Recent advances in central congenital hypothyroidism," *Journal of Endocrinology,* vol. 227, no. 3, pp. 51-71, 2015.

[49] D. Ma, R. Marion, N. P. Punjabi, E. Pereira, J. Samanich, C. Agarwal, J. Li, C. K. Huang, K. H. Ramesh, and L. A. Cannizzaro, "A de novo 10.79 Mb interstitial deletion at 2q13q14.2 involving PAX8 causing hypothyroidism and mullerian agenesis: a novel case report and literature review," *Molecular Cytogenetics,* vol. 7, no. 1, pp. 1-6, 2014.

[50] M. Manandhar, K. S. Boulware, and R. D. Wood, "The ERCC1 and ERCC4 (XPF) genes and gene products," *Gene,* vol. 569, no. 2, pp. 153, 2015.

[51] M. Bogliolo, B. Schuster, C. Stoepker, B. Derkunt, Y. Su, A. Raams, J. P. Trujillo, J. Minguillón, M. J. Ramírez, and R. Pujol, "Mutations in ERCC4, encoding the DNA-repair endonuclease XPF, cause Fanconi anemia," *American Journal of Human Genetics,* vol. 92, no. 5, pp. 800, 2013.

[52] H. Tong, C. Faloutsos, and J. Y. Pan, *Random walk with restart: fast solutions and applications*: Springer-Verlag New York, Inc., 2008.

[53] D. Ben-Avraham, and S. Havlin, *Diffusion and reactions in fractals and disordered systems*: Cambridge University Press, 2000.

[54] P. G. Doyle, and J. L. Snell, *Random Walks and Electric Networks*: Mathematical Association of America, 1984.

[55] L. Cowen, T. Ideker, B. J. Raphael, and R. Sharan, "Network propagation: a universal amplifier of genetic associations," *Nature Reviews Genetics,* vol. 18, no. 9, pp. 551–562, 2017.

[56] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular Systems Biology,* vol. 3, no. 1, pp. 88-88, 2007.

[57] D. Y. Cho, Y. A. Kim, and T. M. Przytycka, "Chapter 5: Network biology approach to complex diseases," *Plos Computational Biology,* vol. 8, no. 12, pp. e1002820, 2012.

[58] P. Csermely, T. Korcsmáros, H. J. Kiss, G. London, and R. Nussinov, "Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review," *Pharmacology & Therapeutics,* vol. 138, no. 3, pp. 333-408, 2012.

[59] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: online learning of social representations," in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, 2014, pp. 701-710.

[60] J. Li, J. Zhu, and B. Zhang, "Discriminative deep random walk for network classification," in Meeting of the Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1004-1013.

[61] S. Gao, C. Tibiche, J. Zou, N. Zaman, M. Trifiro, M. O'Connor-Mccourt, and E. Wang, "Identification and construction of combinatory cancer hallmark-based gene signature sets to predict recurrence and chemotherapy benefit in stage II colorectal cancer," *Jama Oncology,* vol. 2, no. 1, pp. 1-9, 2015.

[62] Y. Park, and E. M. Marcotte, "A flaw in the typical evaluation scheme for pair-input computational predictions," *Nature Methods,* vol. 9, no. 12, pp. 1134-1136, 2012.

[63] T. Pahikkala, A. Airola, S. Pietilä, S. Shakyawar, A. Szwajda, J. Tang, and T. Aittokallio, "Toward more realistic drug–target interaction predictions," *Briefings in Bioinformatics,* vol. 16, no. 2, pp. 325-337, 2015.

[64] E. Guney, "Reproducible drug repurposing: When similarity does not suffice," *Pac Symp Biocomput,* vol. 22, pp. 132-143, 2016.