Abstract

Disease-gene association through Genome-wide association study (GWAS) is an arduous task for researchers. Investigating single nucleotide polymorphisms (SNPs) that correlate with specific diseases needs statistical analysis of associations, considering the huge number of possible disease mutations. The most important drawback of GWAS analysis in addition to its high cost is the large number of false-positives. Thus, researchers search for more evidence to cross-check their results through different sources. To provide the researchers with alternative low-cost disease-gene association evidence, computational approaches come into play. Since molecular networks are able to capture complex interplay among molecules in diseases, they become one of the most extensively used data for disease-gene association prediction. In this survey, we aim to provide a comprehensive and an up-to-date review of network-based methods for disease gene prediction. We also conduct an empirical analysis on 14 state-of-the-art methods. To summarize, we first elucidate the task definition for disease gene prediction. Secondly, we categorize existing network-based efforts into network diffusion methods, traditional machine learning methods with handcrafted graph features and graph representation learning methods. Thirdly, an empirical analysis is conducted to evaluate the performance of the selected methods across seven diseases. We also provide distinguishing findings about the discussed methods based on our empirical analysis. Finally, we highlight potential research directions for future studies on disease gene prediction.

Key words: disease gene prediction; network-based methods; graph representation learning
Potential disruptions on the biological pathways due to the disease-causing factors. As such, they have been extensively used by the computational approaches to complement and enrich existing linkage analysis and GWAS studies.

Currently, there are several molecular networks available to describe relationships between genes, such as protein-protein interaction (PPI) networks, gene regulatory networks, gene co-expression and metabolic interaction (MI) networks. Among these networks, PPI networks are the most extensively leveraged for disease-gene association prediction. One key reason is that proteins perform a vast array of critical functions to sustain an organism’s well-being. Some of these functions include biochemical reactions, metabolic reaction catalysis, DNA replication, transmission of signals between cells, and maintaining structure for cells and tissues. More specifically, proteins collaborate and interact with each other to perform biological functions, leading to many protein interactions, which can be integrated and modelled as a graph/network data structure. Given a protein interaction, a corresponding gene mutation from one of the two proteins could make the existing interaction impossible, and thus loses certain important biological functions and causes diseases. Another reason is, several studies using a PPI network embody the assumption that the position of a protein is not random. Proteins associated with a common set of biological properties tend to have common topological properties in the network such as node degree and centrality [3]. Therefore, PPI networks can be employed in revealing protein-disease associations through the useful network-based features for proteins. However, existing PPI networks are incomplete, i.e., only a fraction of real protein interactions are detected through high-throughput experiments. Likewise they are noisy due to biased experimental evidence towards much studied disease genes [3]. Thus, an integrative approach covering various aspects of proteins such as GWAS, gene-expression, gene ontology, and other domain knowledge is both important and necessary.

In this survey, we focus on reviewing the computational methods leveraging network/graph data for disease gene prediction. First, we introduce the problem definition (i.e., node classification and link prediction) for disease gene prediction with different types of graph inputs. Second, we classify the network-based methods into three categories and provide a brief introduction for these methods. Third, we select representative methods from each category and conduct a comprehensive empirical study on them. The three categories are listed in the following:

- **Network diffusion methods.** The diffusion methods employ random walk techniques for influence propagation in different networks (e.g., PPI networks or phenotype-gene networks) for disease gene prediction.
- **Machine learning methods with handcrafted graph features.** Various features for diseases and genes are first extracted from input graph data, and then fed into traditional machine learning models (e.g., Random Forest) for predicting disease-gene associations.
- **Graph representation learning methods.** Instead of using handcrafted features for disease gene prediction, graph representation learning methods automatically learn the latent features or embeddings for diseases and genes by matrix factorization, graph embedding and graph neural network techniques.

Former surveys on network-based methods [1][2][3], which were published about ten years ago, present pioneering endeavours on network analysis and trying to highlight the importance of network data for disease gene prediction. Nowadays, we witness the usefulness of incorporating network data in several areas ranging from drug discovery to disease gene identification through novel network-based algorithms. In this survey, we aim to bring together up-to-date methods for disease gene prediction and provide a wider perspective to this important problem. In addition, there are two recent surveys which review the recent network embedding efforts on biomedical networks [4][5]. In [5], the authors present an overview of the existing network embedding methods and their applications in biomedical data science, e.g., pharmacetical data analysis, multi-omics data analysis and clinical data analysis. However, they do not conduct any empirical evaluation for the introduced methods. In [4], the authors introduce the recent graph embedding methods and their applications in biomedical networks. In particular, they further select 11 graph embedding methods and perform a systematic evaluation on 5 different tasks, i.e., drug-disease association prediction, drug-drug interaction prediction, PPI prediction, protein function prediction and medical term semantic type classification.

Yet, they do not touch the task on disease gene prediction and its various state-of-the-art approaches. In this survey, we aim to conduct an empirical analysis on disease gene prediction task using different methods from the above three categories. Besides, we further apply and evaluate recent graph embedding methods including heterogeneous network embedding and multi-view network embedding for disease gene prediction.

The rest of this survey is organized as follows. In Section ‘Problem Definition for Disease Gene Prediction’, we cast the task of disease gene prediction as an instance of node classification or link prediction based on different graph inputs. Then, we introduce various network-based disease-gene prediction methods in details in Section ‘Network-based Methods for Disease Gene Prediction’. Next, in Section ‘Empirical Comparison’, we perform a comprehensive empirical evaluation for 14 representative methods. Finally, in Section ‘Future Perspectives’, we discuss potential future research directions in disease-gene association prediction problem and we conclude this paper in Section ‘Conclusion’.

**Problem Definition for Disease Gene Prediction**

Network-based disease gene prediction leverages graph/network data as inputs to predict disease-causing genes. In particular, different types of graphs have been exploited for this purpose, including homogeneous graphs, heterogeneous graphs and multi-view/multiplex graphs. A homogeneous graph refers to a graph with a single type of nodes and a single type of edges, while a heterogeneous graph contains different types of nodes and edges. In addition, a multi-view or multiplex graph is a collection of graphs with the same set of nodes and different types of edges (e.g., edges from different views). Figure 1(a) shows the examples of different types of graph inputs. PPI network in Figure 1(a) is a homogeneous graph, while phenotype-gene network in Figure 1(b) is a heterogeneous graph. Figure 1(c) shows a multi-view graph for proteins, containing three views from the perspectives of PPI, GO similarity and gene expression.

Based on the above graph inputs, network-based disease gene prediction can be treated as a node classification or link prediction task. Node classification aims to infer the disease label of the unlabelled genes by utilizing the known disease genes, whereas link prediction aims to predict disease causing genes by utilizing gene-disease associations. Next, we give a formal definition for these two tasks from the perspective of disease gene prediction.

**Node Classification**

Figure 2(a) shows the node classification tasks, which is to predict the label of the genes of which disease associations are unknown, given known labels on some genes/nodes.

More formally, assume that we have a homogeneous graph \( G = (V, E) \), where \( V \) is the set of nodes/genes and \( E \) shows relationships between nodes. A subset of genes \( V_{\text{known}} \subseteq V \) represents known disease-causing genes, while \( V_{\text{unknown}} = V \setminus V_{\text{known}} \) represents the set of genes of which disease associations are unknown. The formulation of node classification on \( G \) is to predict the labels for the nodes in \( V_{\text{unknown}} \).

Figure 2(b) shows the link prediction tasks, which aims to infer the interaction of unknown gene pairs by using known interactions between genes. More formally, given a heterogeneous graph \( G = (V, E) \), where \( V \) is the set of nodes/genes and \( E \) shows relationships between nodes, the task is to predict the existence of edges in \( E \) for a subset of nodes \( U = \{ u, v \} \), where \( u, v \in V \). The formulation of link prediction on \( G \) is to predict the existence of edges in \( E \) for the nodes in \( U \).

**Link Prediction**

Figure 2(c) shows the link prediction tasks, which aims to infer the interaction of unknown gene pairs by using known interactions between genes. More formally, given a heterogeneous graph \( G = (V, E) \), where \( V \) is the set of nodes/genes and \( E \) shows relationships between nodes, the task is to predict the existence of edges in \( E \) for a subset of nodes \( U = \{ u, v \} \), where \( u, v \in V \). The formulation of link prediction on \( G \) is to predict the existence of edges in \( E \) for the nodes in \( U \).

**Problem Definition for Disease Gene Prediction**

Network-based disease gene prediction leverages graph/network data as inputs to predict disease-causing genes. In particular, different types of graphs have been exploited for this purpose, including homogeneous graphs, heterogeneous graphs and multi-view/multiplex graphs. A homogeneous graph refers to a graph with a single type of nodes and a single type of edges, while a heterogeneous graph contains different types of nodes and edges. In addition, a multi-view or multiplex graph is a collection of graphs with the same set of nodes and different types of edges (e.g., edges from different views). Figure 1(a) shows the examples of different types of graph inputs. PPI network in Figure 1(a) is a homogeneous graph, while phenotype-gene network in Figure 1(b) is a heterogeneous graph. Figure 1(c) shows a multi-view graph for proteins, containing three views from the perspectives of PPI, GO similarity and gene expression.

Based on the above graph inputs, network-based disease gene prediction can be treated as a node classification or link prediction task. Node classification aims to infer the disease label of the unlabelled genes by utilizing the known disease genes, whereas link prediction aims to predict disease causing genes by utilizing gene-disease associations. Next, we give a formal definition for these two tasks from the perspective of disease gene prediction.

**Node Classification**

Figure 2(a) shows the node classification tasks, which is to predict the label of the genes of which disease associations are unknown, given known labels on some genes/nodes.

More formally, assume that we have a homogeneous graph \( G = (V, E) \), where \( V \) is the set of nodes/genes and \( E \) shows relationships between nodes. A subset of genes \( V_{\text{known}} \subseteq V \) represents known disease-causing genes, while \( V_{\text{unknown}} = V \setminus V_{\text{known}} \) represents the set of genes of which disease associations are unknown. The formulation of node classification on \( G \) is to predict the labels for the nodes in \( V_{\text{unknown}} \).

**Link Prediction**

Figure 2(b) shows the link prediction tasks, which aims to infer the interaction of unknown gene pairs by using known interactions between genes. More formally, given a heterogeneous graph \( G = (V, E) \), where \( V \) is the set of nodes/genes and \( E \) shows relationships between nodes, the task is to predict the existence of edges in \( E \) for a subset of nodes \( U = \{ u, v \} \), where \( u, v \in V \). The formulation of link prediction on \( G \) is to predict the existence of edges in \( E \) for the nodes in \( U \).
Network-based Methods for Disease Gene Prediction

In this section, we present a comprehensive review of network-based methods for disease gene prediction, namely diffusion-based methods, traditional feature-based methods and graph representation learning methods as illustrated in Figure 3.

Network Diffusion Methods

Diffusion-based methods working on biological networks are mostly adopted from pioneer graph-based semi-supervised algorithms such as 

In disease gene prediction problem, they are widely exploited in the analysis of the biological pathways especially the pathways that are closest to the known disease genes. Starting from known disease genes, they diffuse along the biological network through random walks. Next, we introduce diffusion methods by walking in different networks.

Random Walk in PPI Network

Random walk with restart (RWR) is the most extensively used algorithm especially in prioritization of disease genes. Differently than a basic random walk, it is able to jump back to any seed node with probability \( r \) at each iteration. Formally, it is defined as follows:

\[
p^{t+1} = (1 - r)Wp^t + rp^0,
\]

where \( W \) is the column-wise normalized adjacency matrix of the network. In particular, \( p_i \) is the probability vector of being at node \( i \) at time step \( t \) in its \( i \)-th entry and \( p^0 \) is the initial probability vector holding the probabilities of being at known disease nodes (seeds) \( \{ s_i \} \). Initially, each known disease protein has a uniformly distributed probability as the sum of the probabilities equal to 1. Total number of iterations is determined by the condition which satisfies the L1 norm difference between \( p^t \) and \( p^{t+1} \) smaller than a pre-defined threshold (e.g., \( 10^{-6} \)).

PRINCE [14] adopts RWR to a weighted PPI network through a weight function and also utilizes the disease similarities as prior probabilities. On an unweighted PPI network, PRINCE basically performs random walk with restart (Eq. 4) with adjusted prior probabilities based on the disease similarity scores.

VAVIEN [13] is proposed to measure the topological similarity of proteins by formulating their interactions with Pearson correlation coefficient as a topological profile. To compute this profile, they employ the random walk proximity as a feature between seed and candidate proteins, so that the proteins with similar interactions will have a similar topological profile. Then, they prioritize the candidate disease genes based on the topological profile scores. In [15], the authors introduce a method called ORIENT, which prioritizes the candidate disease genes with RWR. They perform RWR on the adjacency matrix where the weights of interactions close to the known disease genes are properly reinforced by computing the shortest path distance to the disease genes. In [16], the authors introduce a method called DPLCC to construct diffusion profiles for the disease genes and the candidate genes separately based on their RWR on the PPI and the phenotype similarity network. Candidate genes are prioritized according to their diffusion profile similarities with the query disease. NPDE [17] utilizes non-disease essential proteins by formulating a dual flow network propagation method to prioritize candidate disease genes.
genes. It uses the assumption that if non-disease essential proteins exist as neighbors of a candidate protein, the protein is unlikely to cause a disease. Eventually it formulates this assumption as the negative flow in the prior information of the network propagation and the positive flow is allocated for the disease proteins.

Random Walk in Heterogeneous Network

For diffusion methods in heterogeneous networks, we further divide them into two categories with respect to their task as node classification or link prediction. We start with node classification methods. BioGraph, [19], is a data integration and data mining platform. It utilizes stochastic model of random walks with restarts for a given candidate gene prioritization query as incorporating multiple data sources such as disease, pathway and GO annotation. In [20], two methods, CrossRank and CrossRankStar, formulate disease gene prioritization problem as optimization problems based on network propagation. They model two types of networks, network of networks (NoN) for CrossRank and networks of star networks (NoSN) model for CrossRankStar and both models incorporate tissue specific molecular networks. In SLN-SRW [21], the authors first aim to integrate biomedical data from heterogeneous sources including multiple ontologies and databases. For this propose, a simplified Laplacian normalization based supervised random walk algorithm is employed to learn the edge weights of the integrated network. RWR is then performed on this integrated network for disease gene prediction.

Link prediction methods aim to predict disease-gene associations in heterogeneous phenotype-gene network, where PPI network and phenotype network are connected through known phenotype-gene relationships. RWRH [22] aims to prioritize proteins based on their relevance to disease proteins. Basically, RWRH extends RWR algorithm to the heterogeneous phenotype-gene network through inter and intra transitions between PPI network and phenotype network. It is performed for a given query disease and corresponding disease genes are considered as seed nodes. In [23], a RWRH-based method called RWPCN is proposed to predict and prioritize disease genes on an integrated network comprising human protein complexes, protein interaction network, and phenotype similarity network. The basic difference from RWRH is that RWPCN operate random walks in an additional protein-complex networks. Likewise, a RWRH-based approach to predict malaria-associated genes on an integrated network by integrating human-human, parasite-parasite and human-parasite protein interactions in [24]. BiRW [25], aims to capture circular bigraph patterns based on the assumption that the relation among phenotype-gene associations can be characterized by these patterns. For this purpose, they employ a bi-random walk algorithm and aim to

Table 1. Various network diffusion methods for disease gene prediction.

| Categories                  | Methods          | Method descriptions                                      |
|-----------------------------|------------------|---------------------------------------------------------|
| RW in PPI network (Node classification) | PRINCE [14]     | Network propagation on a PPI network.                   |
|                             | VAVIEN [15]     | Random walk and correlation on a PPI network.           |
|                             | ORIENT [16]     | Neighbor-favoring RWR on a PPI network.                 |
|                             | DP-LCC [17]     | RWR-based diffusion profiles based on PPI and phenotype similarity network. |
|                             | NPDE [18]       | Network propagation with dual flow on a PPI network.    |
| RW in heterogeneous network (Node classification) | BioGraph [19]     | Stochastic RW on an integrated network containing 21 publicly available curated databases. |
|                             | CrossRank [20]  | Optimization based on network propagation on a heterogeneous network with various aspects. |
|                             | SLN-SRW [21]    | Supervised RW to learn edge weights of the integrated network and RWR for prediction. |
| RW in heterogeneous network (Link prediction) | RWRH [22]       | RWR on a heterogeneous phenotype-gene network.          |
|                             | RWPCN [23]      | RWR on a heterogeneous network with phenotypes, genes and protein complexes. |
|                             | RWRH-Malaria [24]| RWR on cross-species PPI networks for human and parasite. |
|                             | BiRW [25]       | RWR on a heterogeneous network with a phenotype, gene and the phenotype-gene associations. |
|                             | RWRMH [26]      | RWR on a multiplex heterogeneous network of protein interactions and disease networks. |
Machine Learning Methods with Handcrafted Features

**Supervised machine learning methods**

Supervised machine learning methods employ features and/or kernels to integrate various biological concepts, which have been extensively studied in disease gene prediction. In [27], DERanking is proposed to benchmark four different strategies and prioritize candidate genes according to network analysis of their differential expression. The reason of incorporating differential expression as prior knowledge is based on the assumption that the strong disease candidates tend to be surrounded by differentially expressed neighbors. Three of the four benchmarking strategies are distinct random walk-based strategies from exponential diffusion kernel approach and one of them is direct neighborhood analysis. These strategies are performed on four networks comprising functional or psychical interactions of proteins for disease gene prediction. In [23], a method called BRIDGE is introduced for prioritization of disease genes by integrating various gene aspects including PPI, protein sequence data, gene expression data, KEGG database and GO, through a weighting scheme. This scheme is attained through a multiple linear regression model with lasso penalty, which determines the phenotypic similarity between two diseases based on the functional similarities between their associating genes. Accordingly, the model is capable of identifying genes associated with the diseases whose genetic bases are completely unknown. In [29], multiple aspects of proteins including known gene-disease associations, protein complexes, PPIs, pathways and gene expression profiles are integrated by Markov random field (MRF) and Bayesian analysis for disease gene prioritization. Initial prior probability of the MRF is set by Gibbs sampling. Metagraph representations [30] are proposed for disease gene prediction, which utilize both PPI network and biological annotations called keywords through heterogeneous subgraphs. By counting the occurrences of proteins within a specific subgraph type in terms of connectivity patterns, feature vectors of proteins are constructed, so that the proteins, which are functionally similar but located far away in a PPI network still have a chance to have similar representations in case they co-occur within the same subgraph type. In [31], the authors employ a multimodal deep belief net named dgMDL to predict disease-gene associations for all known diseases instead of predicting associated genes for a specific disease. This could alleviate the risk of overfitting due to the small number of positives and a large number of features in disease gene prediction problem. They first train two multi-modal DBN, one on PPI network and the other on GO-based similarity network, and then they train a final joint DBN for prediction based on the outputs of the initial DBNs.

**PU learning methods**

Supervised models in disease gene prediction may suffer due to all provided evidence serves as positive samples and the negative samples remain unlabeled. Thus, positive-unlabeled (PU) methods arise to tackle this problem. They are usually based on weighting the positive samples and unlabeled samples to promote the prediction performance during the learning. ProDBGe [32] is a PU learning method for prioritization of candidate genes by enabling multiple data source integration such as PPI network and phenotype similarities through multitask learning strategy. In this way, scoring of the genes for prioritization is performed by considering disease-gene pairs instead of the individual genes. PUDS [33] partitions unlabeled set (i.e., negatives) into four sets namely, reliable negative set RN, likely positive set LP, likely negative set LN and weak negative set WN based on features extracted from PPI, GO terms and protein domain. The weighted support vector machines are then used to build a multi-level classifier based on these sets and positive training set. CATAPULT [35] performs a biased SVM to classify gene-phenotype pairs on a combined heterogeneous gene-phenotype interactions across multiple species. Features for SVM are derived from walks in the network. EPU [27] is an ensemble-based PU learning method, which integrates data from multiple biological sources for training PU learning classifiers consists of weighted KNN, weighted naive Bayes and multi-level support vector machine. It starts with individual label propagation on gene expression network, PPI network and GO similarity network. Then, the obtained gene weights are combined and fed into the ensemble learner. PEGPUL [37] is an ensemble PU learning model to combine a multi-level SVM, a weighted KNN and a weighted decision tree for disease gene prediction. First, it extracts reliable negative genes by utilizing a co-training algorithm. Then, it constructs a similarity graph through metric learning by using PPI, protein domain and GO terms and performs a multi-rank-walk on the constructed graph to propagate labels prior to feeding the genes into the ensemble. A summary of supervised methods and PU learning methods is presented in Table 2.

| Categories         | Methods       | Method descriptions |
|--------------------|---------------|---------------------|
| Supervised learning| DERanking [23]| A ranking strategy based on whether a gene is surrounded by highly differentially expressed genes. |
|                    | BRIDGE [23]   | A regression model with lasso penalty to weight 5 genomic data sources. |
|                    | IMRF [25]     | An improved MRF method on multiple networks, e.g., PPI, pathways, protein complexes, etc. |
|                    | Metagraphs [30]| Using metagraph features extracted from a heterogeneous network with PPI and gene keywords. |
|                    | dgMDL [31]    | Multi-modal deep belief nets on PPI network and gene network based on GO similarities. |
| PU learning        | ProDBGe [32]  | Biased SVM using features derived from multiple data sources [33] including PPI. |
|                    | PUDS [33]     | Multi-level classification with biased SVM on features extracted from PPI, protein domain and GO. |
|                    | CATAPULT [35] | Biased SVM with features derived from walks in a heterogeneous phenotype-gene network. |
|                    | EPU [27]      | Ensemble PU learning using various features derived from PPI and GO similarity networks. |
|                    | PEGPUL [37]   | A perceptron ensemble of graph-based PU learning from 3 base classifiers SVM, KNN and CART. |

Graph representation learning methods

Recall that we need to manually extract various features for the methods introduced in Table 2 which is tedious and requires domain knowledge. Recently, graph representation learning methods [38, 39], which can automatically learn the latent features/representations for the nodes, have acclaimed wide attentions in bioinformatics and biomedical applications.
In this section, we review graph representation learning methods designed for disease gene prediction. In particular, we divide them into four categories, namely, matrix factorization, graph embedding, graph neural network and hybrid methods as shown in Figure 6.

Matrix factorization
Matrix factorization (MF) techniques have been widely used for link prediction in bioinformatics, e.g., PPI prediction, drug-target prediction, miRNA-disease association prediction, drug-pathway association prediction, etc. Here, we introduce various MF methods developed for disease-gene association prediction. Basically, MF methods in gene-disease association prediction, such as matrix completion (MC) [24], probability-based collaborative filtering (PCFM) [11] and manifold learning [13], aim to learn the latent factors for diseases and genes from the gene-disease association matrix. In addition, they can also specify the factorization process by including additional constraints into the objective function and thus MF techniques are useful for revealing important associations between diseases and genes.

In [25], IMC constructs gene and disease features using different sources including human gene, disease associations, gene-expression from different tissue samples, functional interactions between genes, gene-phenotype associations of other species, disease similarities and disease-related textual data from OMIM database. Then, they form the general disease association matrix using these features and formulate an optimization problem to recover unknown low-rank matrix using observations from the constructed associations matrix. Their inductive approach is capable of making predictions for a query disease with no previously known gene associations. In [12], PCFM exploits the gene-gene, gene-disease relationships, gene-disease linkages between orthologous genes and eight non-human species diseases and disease-gene similarity associations. They propose two probability-based collaborative filtering models, one with an average heterogeneous regularization and another with personal heterogeneous regularization using vector space similarity to predict gene-disease associations. The advantage of PCFM than the collaborative filtering is that PCFM enables probabilistic consideration in gene-disease associations instead of binary evaluation. Manifold learning [23] uses gene-disease association data for prediction and assumes that the geodetic distance between any associated gene-disease pairs is shorter than non-associated gene-disease pairs in a lower dimensional manifold. An optimization function is defined based on this assumption and singular value decomposition is employed to solve the optimization problem. Collage [23] applies collective matrix factorization (CMF) to combine a wide range of 14 data sets including RPKM-normalized RNA-seq transcriptional profiles and phenotype ontology annotations. Then, it performs chaining on the learned latent matrices to obtain gene profiles for prioritizing bacterial response genes in Dictyostelium. Similarly, Medusa [59] also builds a collective matrix factorization model for data fusion of a large-scale collections of heterogeneous data and performs chaining on the learned latent matrices to establish connections between non-neighboring nodes in the fusion graph. It formulates the growing of the modules as a sub-modular optimization program. The proposed method is capable of both associating genes with diseases and detecting disease modules. In [31], an unsupervised learning model based on matrix tri-factorization (tri-NMF) framework is proposed to detect disease causing genes from pan-cancer data. In particular, it exploits both the similarities of mutation profiles of different cancer types and gene interaction network. A method called GeneHound [52] adopts Bayesian probabilistic matrix factorization (BPMF) for disease gene prioritization problem. GeneHound uses gene-disease associations as partially observed data and a raw fusion is employed to integrate multiple genomic data sources including literature-based phenotypic and literature-based genomic information. Then, the Bayesian data fusion model jointly learns gene and disease latent factors and corresponding gene and disease-association matrices for disease-gene association prediction.

Graph embedding methods
Graph embedding methods learn low-dimensional and continuous vector representations of nodes through a neural network. For example, SkipGram [53] architecture is an extensively used architecture to construct associations between the node and its neighborhood. The neighborhood of the node is extracted through the random walks. Endeavours [54], [55], such as DeepWalk [55] and node2vec [54], generate node representations such that the nodes lying within the short random walk distance have similar embeddings. There are also several random-walk based embedding methods proposed for disease gene prediction as follows.

SmuDEG [57] combines disease-gene, phenotype and gene-gene, phenotype associations with interactions between genes to generate a corpus for SkipGram-based representation learning. Then, it builds an artificial neural network (ANN) to predict gene-gene disease associations. In [56], HereWalk builds a weighted heterogeneous network by joining six public data sources including PPI, miRNA similarity network, and disease phenotype similarity network and then performs SkipGram based network embedding. It further applies meta-path selection to eliminate potential redundant and misleading information caused by the heterogeneous walks with multiple entities. HerGePred [59] also performs SkipGram-based graph embedding on a heterogeneous network consisting of a PPI, disease-protein associations, protein-GO associations, and gene-disease associations. Eventually, HerGePred predicts novel disease-gene associations in two different manners. It can directly calculate the cosine similarity between the embedding vectors of the query disease and the proteins. It can also perform random walk on disease-gene network, which is reconstructed based on the calculated cosine similarities between embeddings.

Graph neural networks
Graph neural network (GNN) is an advanced deep learning model for graph data [9] and has been applied for various bioinformatics tasks [9,23]. Graph convolutional network (GCN), graph attention network (GAT) and graph auto-encoder (GAE) are representative GNN models. For example, GCN aims to learn node embeddings by implementing the convolution operation on a graph based on the properties of neighborhood nodes. Here, we will have a quick review on GNN models for disease gene prediction as follows.

In [24], the authors introduce GCAS, which is an adaptation of graph convolutional network for disease gene prioritization. First, they construct a heterogeneous network consisting of ontological associations as well as curated associations including genes, diseases and pathways from multiple sources. They then perform direct spectral convolution to successively propagate the influence to the neighborhood and infer novel disease-gene associations. In [41], a graph convolutional network-based disease gene prioritization method called PGCN exploits a heterogeneous phenotype-gene network and the additional information for the nodes (e.g., disease ontology similarity, gene expression data, etc) for disease gene prioritization. In [42], the authors introduce the variational graph auto-encoder (VGAE) for gene-disease association prediction in a heterogeneous disease-gene network.

Hybrid methods
Here, we denote the methods, which combine the representation learning methods with other techniques to derive features, as hybrid methods. A number of methods combine graph embedding (e.g., node2vec) with other methods for feature engineering. In [43], a method called N2V
build a graph using the similarity between proteins based on G-SESAME on GO terms [74]. In particular, we first calculate the pairwise functional IntAct database [73] and a protein functional similarity network based We collect two networks for proteins, namely, a PPI network from the Datasets for various network-based methods for disease gene prediction. In this section, we conduct experiments to evaluate the performance of prediction. Table 3 summarizes various graph representation learning models for disease-gene associations.

Table 3. Graph representation learning methods for predicting disease-gene associations.

| Categories            | Methods               | Method descriptions                                                                 |
|-----------------------|-----------------------|-------------------------------------------------------------------------------------|
| Matrix factorization   | IMC [60]              | Inductive matrix completion incorporating gene and disease features from multiple data sources. |
|                       | PCFM [61]             | Probability-based collaborative filtering models with different regularization terms. |
|                       | Collage [29]          | Collective matrix factorization using 14 datasets including genes, GO terms, KEGG pathways, etc. |
|                       | Medusa [59]           | Collective matrix factorization on a data fusion graph with 16 data matrices. |
|                       | Tri-NMF [31]          | Matrix tri-factorization on mutation profile similarities for different cancer types and PPI network. |
|                       | GeneHound [57]        | Bayesian matrix factorization on OMIM associations with genomic and phenotypic data sources. |
| Graph embedding        | SmoGDL [62]           | RW-based embedding and an ANN on pair of disease and gene feature vectors. |
|                       | HetGePred [59]        | RW-based embedding on a heterogeneous network with diseases, symptoms, genes and GO terms. |
| Graph neural networks  | GCAS [29]             | Graph convolution on heterogeneous association network for rare diseases (HANRD). |
|                       | PGCN [61]             | GCN to learn the embeddings for phenotypes and genes in a disease-gene heterogeneous network. |
|                       | VGAE [62]             | VGAE to learn embeddings for diseases and genes in disease-gene networks for gene prioritization. |
| Hybrid methods         | N2VGKo [63]           | Combines node2vec embeddings and handcrafted features. |
|                       | N2A-SVM [64]          | Combines node2vec embeddings with an autoencoder on a PPI network. |
|                       | N2Vmotif [65]         | High-order PPI structures combining node2vec embeddings and network motifs. |
|                       | GCN-MF [66]           | Combines GCN with matrix factorization using both gene similarities and disease similarities. |
|                       | HNEEM [67]            | An ensemble of 6 graph embedding methods in a disease-gene-chemical heterogeneous network. |
|                       | DW-GCN [68]           | Integrates graph embedding (DeepWalk) and graph convolutional network. |

integrates the node2vec embeddings extracted from PPI network with the biological annotations for disease-gene association prediction. In [63], the proposed method N2A-SVM employs node2vec embedding of the genes from a PPI network and then performs dimension reduction with auto-encoder to predict Parkinsonism [79]’s disease genes. In [65], the authors propose to combine graphlet representations with node2vec embeddings for disease gene prediction. [66] builds a network-based framework for gene-disease association prediction, which exploits disease similarity and gene similarity graphs to construct two GCNs separately for diseases and genes. Then, GCNs are trained by using corresponding disease and gene features and optimized with label information with matrix factorization. In [67], the authors propose a heterogeneous network embedding method, HNEEM, for gene-disease association prediction by ensemble learning. The constructed heterogeneous network consists of gene-disease associations, gene-chemical associations ans disease-chemical associations. HNEEM first extracts graph embeddings with six embedding methods and then feeds these embeddings into a random forest classifier for disease gene prediction. DW-GCN [68] is proposed to combine DeepWalk and GCN for gene-disease association prediction on a heterogeneous disease-gene network. Final predictions are derived using the output of a GCN decoder and the probability distribution derived from DeepWalk. Table 4 summarizes various graph representation learning models for disease gene prediction.

Emirical Comparison

In this section, we conduct experiments to evaluate the performance of various network-based methods for disease gene prediction.

Experimental Setup

Datasets

We collect two networks for proteins, namely, a PPI network from the IntAct database [29] and a protein functional similarity network based on GO terms [29]. In particular, we first calculate the pairwise functional similarity between proteins based on G-SESAME [47], and subsequently build a graph using the K nearest neighbor (KNN) algorithm and we set $K = 10$ in all of our experiments. Overall, there are a total of 12,901 nodes (i.e., proteins) in both networks, with 96,845 edges in the PPI network and 107,308 edges in protein functional similarity network. There are several publicly available databases for gene-disease associations as shown in Table 4. In our experiments, we acquired the associations from the OMIM database. Given a specific disease/phenotype (e.g., Alzheimer’s disease), we extract MIM IDs from OMIM Morbid Map and retrieve their corresponding protein IDs from the UniProt [76] conversion tool. Each protein node is subsequently assigned a binary label to represent whether it is a causative protein for this disease. In our experiments, we focus on seven diseases, namely, Alzheimer’s disease (11), breast cancer (24), colorectal cancer (34), diabetes mellitus (37), obesity (19), lung cancer (15) and prostate cancer (17). Note that the number of positive proteins for each disease is included in the parentheses. The data and supplementary materials are available at https://github.com/seiminah/SurveyDGP.

Table 4. Publicly available databases for gene-disease associations.

| Name       | URL                                      | Latest Update |
|------------|------------------------------------------|---------------|
| OMIM       | https://omim.org                          | July, 2020    |
| DoGeNet    | https://www.disgenet.org                  | June, 2020    |
| COSMIC     | https://cancer.sanger.ac.uk               | April, 2020   |
| GyGeNET    | http://www.gygenet.org                    | January, 2018 |
| CTDBase    | http://ctdbase.org                        | July, 2020    |

Selected methods for evaluation

We select a subset of methods from different categories for evaluation. For example, we select RWRH [22] from network diffusion methods and IMC [46] from matrix factorization methods. For machine learning methods with hand-crafted features, we select Metagraphs [60] from supervised methods, and Catapult [15] and ProDiGe [32] from PU learning methods. Furthermore, we select a graph embedding method node2vec [54] and a hybrid method N2VGKo [63]. Table 4 shows the availability of the source codes of these methods.

As aforementioned, disease gene prediction is a typical node classification or link prediction task. Therefore, the state-of-the-art social
network analysis methods for node classification or link prediction can be exploited for disease gene prediction. In our experiments, we also include HIN2Vec [83], HeGAN [84], MVE [85], mvn2vec [86], DMNE [87] and MANE [88] in our evaluation study. These methods are introduced as follows and Table 5 shows the availability of their source codes.

- **HIN2Vec** [83]: A heterogeneous network embedding approach, which samples heterogeneous paths called meta-paths and feeds them into a neural network. We combined PPI network (PPI view) and functional similarity network (GO view) to form a single heterogeneous graph, where edges from different views are assigned to a different type of relation.

- **HeGAN** [84]: A heterogeneous network embedding approach, which utilizes the adversarial principle. We perform it on the constructed single network as described in HIN2Vec.

- **MVE** [85]: A state-of-the-art multi-view network embedding approach which maintains the collaboration between views by regularizing the Euclidean norm between view-specific embeddings and the final embeddings. The parameter $\gamma$ is used to control the weight of regularization. Since we conduct our experiments on unsupervised models, we adopt the unsupervised version.

- **mvn2vec** [86]: A state-of-the-art multi-view embedding approach. There are two proposed versions. Since they both have similar results we report only mvn2vec-v1 version which regularizes the Euclidean norm between view-specific embeddings, controlled by a hyperparameter $\gamma$. When $\gamma$ is set to zero, it becomes equivalent to node2vec performed on a single view.

- **DMNE** [87]: A multi-view network embedding algorithm which is also capable of generating embeddings for many-to-many node mappings across views. Note that in our dataset only one-to-one mappings exist. We adopt their proximity disagreement formulation, due to its flexible assumption and better empirical performance.

- **MANE** [88]: A random-walk sampling-based multi-view embedding algorithm, which unifies diversity, first-order collaboration and novel second-order collaboration principles in a framework. There are two hyperparameters $\alpha$ and $\beta$ to regularize the contribution of the principles.

The datasets that we utilize for the studied methods in this survey are as follows. node2vec and RWR are simply performed on the PPI network. In IMC, Catapult, ProDiGe and RWRH we use the PPI network, phenotype similarity network [89] and the protein-phenotype associations. Metagraph and N2VKO leverage the PPI network and protein-keyword associations retrieved from Uniprot [77], HIN2Vec, HeGAN, MVE, mvn2vec-r, DMNE and MANE utilize GO and PPI networks described in Datasets section. For methods using protein-phenotype associations, we eliminate the test data associations to prevent data leakage. All methods are performed using the implementations provided by their respective authors unless stated otherwise, and we apply their suggested parameter settings for the hyperparameters. For multi-view graph embedding models (i.e., MVE, mvn2vec-r, DMNE and MANE) we set walk length to 10, number of walks per node to 5, negative sampling size to 10, windows size to 3 and random walk parameters of node2vec $(p,q)$ to 1. All methods have $D = 128$ as the dimension of the final embedding. Differently, for heterogeneous graph embedding model HIN2Vec we set walk length to 30 and number of walks per node to 10. For graph embedding methods (i.e., HIN2Vec, HeGAN, MVE, mvn2vec-r, DMNE and MANE), the learned final embeddings are fed into the logistic regression model and for all methods five-fold cross-validation is performed for training.

Table 5. Selected methods for disease gene prediction and their source codes.

| Methods   | Source Code Availability                                      |
|-----------|--------------------------------------------------------------|
| IMC       | https://bio.data.oden.utexas.edu/project/gene-disease       |
| Catapult  | http://www.marcotte.lab.org/index.php/Catapult              |
| ProDiGe   | http://chao.enst.fr/prodigie                                |
| RWRH      | https://github.com/alberto-valdeolivas/RWR-RH               |
| Metagraph | https://github.com/sezinata/Metagraph                       |
| node2vec  | https://github.com/aditya-grover/node2vec                   |
| N2VKO     | https://github.com/sezinata/N2VKO                           |

Table 6. Graph embedding methods proposed for social network analysis and their availability.

| Methods   | Source Code Availability                                      |
|-----------|--------------------------------------------------------------|
| HIN2Vec   | https://github.com/ciesheep/hin2vec                          |
| HeGAN     | https://github.com/librahau/HeGAN                            |
| MVE       | https://github.com/impse/MVE                                 |
| mvn2vec-r | https://github.com/sezinata/mvn2vec-code                     |
| DMNE      | https://github.com/njingchaol/dmne                           |
| MANE      | https://github.com/sezinata/MANE                             |

Results and Discussion

We demonstrate the performance comparison of 14 state-of-the-art methods in Table 1 in terms of AUC. In particular, we group these methods into three groups based on their input data for easy comparison, namely homogeneous graph, heterogeneous graph and multi-view graph. In addition, we also performed over-sampling with SMOTE [92] for the network embedding methods. The over-sampling results as well as the comparison results in terms of ranking-aware metrics are provided in our supplementary materials. Based on the results in Table 1 we can have the following observations.

Firstly, we can observe that RWR and node2vec working on a homogeneous PPI network achieve low performance as expected. Therefore, we are motivated to integrate different data sources into a heterogeneous network or multi-view network to improve the prediction performance.

Secondly, various methods working on a heterogeneous network perform well for disease gene prediction. ProDiGe and Catapult, as PU learning methods, are very promising to address the imbalanced data issue for disease gene prediction and both incorporate the phenotype similarity network and phenotype-gene associations. In particular, ProDiGe achieves an AUC of 0.7763 and is the second best performer among these 14 methods. N2VKO and Metagraph first work on feature extraction/selection in a PPI and keyword network, and then undergo an oversampling procedure. Both of the methods achieve a decent performance. Note that the random walk parameters of node2vec $(p,q)$ in N2VKO are not tuned to maintain consistency with random walk sampling based embedding methods. Heterogeneous graph embedding methods including HIN2Vec and HeGAN hinge on different principles and perform very well. HIN2Vec utilizes random walks and negative sampling with a neural network model, while HeGAN utilizes adversarial learning principle in which a discriminator and a generator compete with each other as in a mini-max game. In particular, HeGAN’s performance highly depends on the used pre-trained embeddings as inputs and thus careful initialization is critical in this model. RWRH also performs very well and it is a good alternative to the network embeddings methods. However, its performance depends on the tuning process, e.g., the setting of the restart probability.

Lastly, multi-view methods are also robust and useful tools for disease gene prediction. mvn2vec-r and MANE are two random walk-based embedding methods employing Skip-gram for learning the final
The model to focus on more important neighbours. Due to the self-attention layers, in particular, this self-attention operation enables networks to assign different weights to different neighbours with masked Attention Mechanisms and Data Integration learning or oversampling for disease gene prediction in the future. It would be very interesting to see the efforts that combine GAN with PU learning and oversampling techniques. For example, PU learning methods are also common strategies to address this challenge. However, both oversampling of minority class samples (e.g., SMOTE [32, 35, 34]) select likely-positive samples from the unlabeled data to tackle the problem. Oversampling of minority class samples (e.g., SMOTE) is also a common strategy to address this challenge. However, both strategies might need a tremendous tuning effort for difficult scenarios to attain a satisfactory performance. Developing more efficient and accurate models leveraging these strategies could be a promising direction.

Recently, generative adversarial networks (GAN) have been successfully applied to augment the data for various tasks, e.g., image classification [23, 24], speech recognition [25], etc. It is thus worth investigating GAN-based techniques for disease gene prediction with limited labels in the graph data. In addition, researchers also employ GAN to boost the PU learning [26] and oversampling processes [27]. Therefore, it would be very interesting to see the efforts that combine GAN with PU learning or oversampling for disease gene prediction in the future.

Attention Mechanisms and Data Integration

Graph Attention Networks (GAT) [29], as an extension of graph convolutional networks, assign different weights to different neighbours with masked self-attention layers. In particular, the self-attention operation enables the model to focus on more important neighbours. Due to the attention mechanism, GAT have been applied to generate accurate graph embeddings for various tasks in recommendation systems [29, 30] and bioinformatics [29]. Therefore, it is also promising to develop node-level attention mechanisms for diseases and genes in different types of graph inputs for network-based disease gene prediction. Moreover, it is common to integrate different graph data sources through multi-view techniques [90]. In these multi-view techniques, it is important to reveal the contributions of each network to the final prediction performance, so that it would be possible to prioritize the networks based on their significance. Graph-level attention mechanisms [31] can be useful in multi-view graphs for this purpose. In addition, hierarchical attention mechanisms [30] can also be applied for disease gene prediction by combining both node-level and graph-level attentions.

FUTURE PERSPECTIVES

In this section, we present possible future directions that may address current challenging issues for more accurately predicting disease genes.

Learning with Limited Labeled Data

Learning with limited labelled data has been a challenging task in disease gene prediction. Existing efforts to overcome this difficulty include PU learning and oversampling techniques. For example, PU learning methods [32, 33] select likely-positive samples from the unlabelled data to tackle the problem. Oversampling of minority class samples (e.g., SMOTE [32]) is also a common strategy to address this challenge. However, both strategies might need a tremendous tuning effort for difficult scenarios to attain a satisfactory performance. Developing more efficient and accurate models leveraging these strategies could be a promising direction.

Sampling Strategies for Multi-View Inference

The study N2VKO [63] demonstrates that node2vec embeddings achieve an inferior prediction performance on some diseases such as the obesity and prostate cancer. The reason is that the disease-associated proteins are scattered in the network with a greater hop-distance between them, which limits the prediction power of a RW-based sampling method (e.g., node2vec). In these cases, it is good to develop models that can utilize other structures in the network. Specifically, this problem can be tackled by sampling strategies which may not consider only the local neighborhood of a node in a network. For example, we can adopt a collaborative sampling strategy, which can consider all the views of multi-view networks or utilize the attributed nodes during the sampling procedure. It would thus minimize the effect of local neighborhoods and help to increase the prediction performance.

Single-cell Data

Recently, single-cell RNA-seq (scRNA-seq) techniques become popular with the advancements in next-generation sequencing. Compared to traditional bulk RNA-seq analysis, they enable cell-level sequencing to capture cell-to-cell heterogeneity. In particular, several methods have been developed to infer gene relationships [40] and gene regulatory networks from single-cell gene expression data. Disease gene prediction can thus benefit from such inferred gene relationships and gene regulatory networks.

In addition, precise identification of cell states and types is crucial to understand the disease related mechanisms so the scientists can detect correlated expression levels of genes across a homogeneous population of cells. Community detection algorithms such as louvain [41] and its similar version leiden [42] are very popular tools to identify cell clusters.
Furthermore, it is possible to model single-cell gene expression data as graphs and we can thus employ graph embedding methods for cell type identification through graph clustering.\cite{10}

Explainable Machine Learning Models

Decision trees, linear regressions and logistic regressions are commonly used explainable machine learning models. The rise of the neural networks and deep learning in many areas such as video, speech and text processing comes with the need for explanation for the “black box” nature of these models. As we know, it is very important for medical disease-related predictions to provide information on why the model performs a certain prediction. However, the graph representation learning methods covered in this review are also lack of interpretability. We are thus highly motivated to develop explainable machine learning models for disease gene prediction.

Recently, knowledge graphs (KG) have been integrated with user-item graphs for accurate and explainable recommendation.\cite{9,10} A recent work\cite{10} adopts KG and graph neural networks for explainable drug-drug interaction (DDI) prediction. Similarly, we can also construct knowledge graphs for diseases and genes using various side information. For example, we can obtain the medical KG with diseases, drugs and symptoms, and gene ontology KG with genes and their functional annotations. By integrating advanced graph neural network techniques (e.g., GAT) and knowledge graphs, we expect to achieve accurate and explainable predictions for novel disease-gene associations.

Conclusion

Discovering disease causing genes and analyzing their roles in the disease are not only critical for understanding disease formation mechanism, but also extremely important for designing appropriate drugs for corresponding clinical therapies. Linkage analysis and genome-wide association studies (GWAS) form the basis of disease gene prediction. However, they generate a large number of false positives in their statistical analysis of biomarkers. Computational approaches are efficient and complementary tools to help biologists filter out noisy false positives and provide a list of genes which are worth for further clinical study. In this survey, we focus on network-based research, leveraging various networks in their problem formulation for disease gene prediction. We provide an organized, up-to-date overview of state-of-the-art network-based approaches. We also perform an empirical comparison study on different computational methods based on different graph inputs.

Generally, the methods in both heterogeneous network and multi-view network perform very well for disease gene prediction. In particular, multi-view methods with Skip-gram architecture can serve as robust and useful tools, not only for disease gene prediction but also for visualization and clustering purposes. Its low computational cost with a minimal tuning necessity and high prediction performance, make it a good alternative to the other learning methods. These techniques could be further investigated by considering attention mechanisms or different network sampling strategies instead of random walks. Moreover, for network-based disease gene prediction analysis, constructing reliable networks is critical, so future research need to focus on predicting novel interactions and removing noisy interactions. The ultimate success of the disease gene prediction will depend on the parallel improvements both in the experimental techniques by biologists and clinicians to provide rich and reliable biological data sets, and in the advanced computational techniques by computer scientists to provide efficient and robust ways to discover novel knowledge and insights from the biological data.

Key Points

- Uncovering disease-causing genes is a fundamental objective of human genetics, while computational prediction of disease-genes provides a low-cost alternative.
- We classified and reviewed state-of-the-art network-based approaches for disease gene prediction with different types of graph inputs.
- We empirically evaluated various selected methods, including some advanced methods for social network analysis, for disease gene prediction.
- We also discussed possible future directions that may address current challenging issues for more accurately predicting disease genes.

References

1. Fides Zenk, Eva Loeser, Rosaria Schiavo, Fabian Kilpert, Ozren Bogdanović, and Nicola Iovino. Germ line-inherited hlk2?m3 restricts enhancer function during maternal-to-zygotic transition. Science, 357(6347):212–216, 2017.
2. Sora Yoon et al. Efficient pathway enrichment and network analysis of GWAS summary data using GSA-SNP2. Nucleic Acids Research, 46(10):e60–e60, 03 2018.
3. Xiujuan Wang, Natali Gulbahce, and Haiyuan Yu. Network-based methods for human disease gene prediction. Briefings in Functional Genomics, 10(5):280, 2011.
4. Han-Yu Chuang, Eunjung Lee, Yu-Tsoung Lia, Doheon Lee, and Trey Ideker. Network-based classification of breast cancer metastasis. Molecular Systems Biology, 3(1):140-n/a, 2007.
5. Trey Ideker and Roded Sharan. Protein networks in disease. Genome Research, 18(4):644-652, 2008.
6. Jörg Menche, Amitabh Sharma, Maksim Kitsak, Susan Dina Ghiaassian, Marc Vidal, Joseph Loscalzo, and Albert-Lázló Barabási. Uncovering disease-disease relationships through the incomplete interactome. Science, 347(6224), 2015.
7. Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. Nature reviews genetics, 12(1):56, 2011.
8. Chang Su, Jie Tong, Yongjun Zhu, Peng Cui, and Fei Wang. Network embedding in biomedical data science. Briefings in Bioinformatics, 21(1):182–197, 12 2018.
9. Xiang Yue, Zhen Wang, Jingong Huang, Srinivasan Parthasarathy, Soheil Moosavinasab, Yugui Huang, Simon M Lin, Wen Zhang, Ping Zhang, and Huan Sun. Graph embedding on biomedical networks: methods, applications and evaluations. Bioinformatics, 36(4):1241–1251, 2020.
10. Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In Proceedings of the Twentieth International Conference on International Conference on Machine Learning, ICML’03, page 912â€–919, AAAI Press, 2003.
11. Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. Random walk with restart: Fast solutions and applications. Knowl. Inf. Syst., 14(3):327â€–346, March 2008.
12. Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In Proceedings of the 15th International Conference on Neural Information Processing Systems, NIPS’02, page 321â€–328, Cambridge, MA, USA, 2003. MIT Press.
Survey for disease gene prediction

[13] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N. Robinson. Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics*, 82(4):949 – 958, 2008.

[14] Oron Vanunu, Oded Magger, Eytan Ruppin, Tomer Shimoni, and Roded Sharan. Associating genes and protein complexes with disease via network propagation. *PLOS Computational Biology*, 6(1):1–9, 01 2010.

[15] Sinan Erten, Gurkan Bebek, and Mehmet Koyutürk. Vavien: An algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks. *J Comput Biol*, 18(11):1561–1574, Nov 2011.

[16] Jie Zhu, Yulang Qin, Taigang Liu, Jun Wang, and Xiaqi Zheng. Prioritization of candidate disease genes by topological similarity between disease and protein diffusion profiles. *BMC Bioinformatics*, 14(5):S5, 2013.

[17] Shuyao Wu, Fengjiong Shao, Jun Ji, Rencheng Sun, Rizhuang Dong, Yueke Zhou, Shaojie Xu, Yi Sui, and Junlong Hu. Network propagation with dual flow for gene prioritization. *PLOS ONE*, 10(2):1–15, 02 2015.

[18] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N. Robinson. Walking the interactome for prioritization of candidate disease genes based on topological similarity of proteins in interaction networks. *J Comput Biol*, 18(11):1561–1574, Nov 2011.

[19] Anthony ML Liekens, Jeroen De Knijf, Walter Delemaes, Bart Goethals, Peter De Rijck, and Jurgen Del-Favero. Biograph: unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biology*, 12(6):R57, 2011.

[20] Jie Zhu, Yulang Qin, Taigang Liu, Jun Wang, and Xiaqi Zheng. Prioritization of candidate disease genes by topological similarity between disease and protein diffusion profiles. *BMC Bioinformatics*, 14(5):S5, 2013.

[21] Shuyao Wu, Fengjiong Shao, Jun Ji, Rencheng Sun, Rizhuang Dong, Yueke Zhou, Shaojie Xu, Yi Sui, and Junlong Hu. Network propagation with dual flow for gene prioritization. *PLOS ONE*, 10(2):1–15, 02 2015.

[22] Yong Jin Li and Jagdish C. Patra. Genome-wide inferring drug-target interaction prediction with dual flow for gene prioritization. *PLOS ONE*, 10(2):1–15, 02 2015.

[23] Peng Yang, Xiao-Li Li, Jian-Xing Wei, Chee-Keong Kwoh, and Xing-Chang Wu. Enhancing the prediction of disease-gene associations with multimodal deep learning. *Bioinformatics*, 35(19):3735–3742, 03 2019.

[24] Fantine Mordelet and Jean-Philippe Vert. Progde: Prioritization of disease genes with multitask learning from positive and unlabeled examples. *BMC Bioinformatics*, 12(1):389, 2011.

[25] Tijl De Bie, Leon-Charles Tranchevent, Liesbeth MM Van Olftelen, and Yves Moreau. Kernel-based data fusion for gene prioritization. *Bioinformatics*, 23(13):i125–i132, 2007.

[26] Peng Yang, Xiao-Li Li, Jian-Ping Mei, Chee-Keong Kwoh, and See-Kiong Ng. Positive-unlabeled learning for disease gene identification. *Bioinformatics*, 28(20):2640–2647, 08 2012.

[27] U. Martin Singh-Blom, Nagarajan Natarajan, Ambuj Tewari, John O. Woods, Inderjit S. Dhillon, and Edward M. Marcotte. Prediction and validation of gene-disease associations using methods inspired by social network analyses. *PLoS ONE*, 8(5):1–17, 05 2013.

[28] Peng Yang, Xiaoli Li, Hon-Nian Chua, Chee-Keong Kwoh, and Xing-Chang Wu. Ensemble positive unlabeled learning for disease gene identification. *Plos one*, 9(5), 2014.

[29] Ghomam-Hossein Jowkar and Egbhal G Mansoori. Perceptron ensemble of graph-based positive-unlabeled learning for disease gene identification. *Computational biology and chemistry*, 64:263–270, 2016.

[30] Peng Cui, Xiao Wang, Jian Pei, and Wenwu Zhu. A survey on network embedding. *IEEE Transactions on Knowledge and Data Engineering*, 31(5):833–852, 2018.

[31] Hongyun Cai, Vincent W Zheng, and Kevin Chang. A comprehensive survey of graph embedding: problems, techniques and applications. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[32] Hua Wang, Heng Huang, Chris Ding, and Feiping Nie. Predicting protein–protein interactions from multimodal biological data sources via nonnegative matrix tri-factorization. *Journal of Computational Biology*, 20(6):344–358, 2013.

[33] Yong Liu, Min Wu, Chunyan Miao, Peilin Zhao, and Xiao-Li Li. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS Computational Biology*, 12(2):e1004760, 2016.

[34] Ali Ezzat, Peilin Zhao, Min Wu, Xiao-Li Li, and Chee-Keong Kwoh. Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM transactions on computational biology and informatics*, 14(3):646–656, 2016.

[35] Xing Chen, Lei Wang, Jia Qu, Na-Nan Guan, and Jian-Qiang Li. Predicting mirna–disease association based on inductive matrix completion. *Bioinformatics*, 34(24):4256–4265, 2018.

[36] Zhi-Chao Zhang, Xiao-Fang Zhao, Min Wu, Le Qu-Yang, Xing-Ming Zhao, and Xiao-Li Li. A graph regularized generalized matrix factorization model for predicting links in biomedical bipartite networks. *Bioinformatics*, 36(11):3474–3481, 2020.

[37] Chun-Chun Wang, Yan Zhao, and Xing Chen. Drug-pathway association prediction: from experimental results to computational models. *Briefings in Bioinformatics*, 2020.

[38] Nagarajan Natarajan and Inderjit S. Dhillon. Inductive matrix completion for predicting gene–disease associations. *Bioinformatics*, 30(12):i60–i68, Jun 2014.
[47] Xiangxiang Zeng, Ningxiang Ding, Alfonso Rodríguez-Patón, and Quan Zou. Probability-based collaborative filtering model for predicting gene-disease associations. BMC Medical Genomics, 10(5):76, Dec 2017.

[48] Ping Luo, Li-Ping Tian, Bolin Chen, Quanxia Xiao, and Fang-Xiang Wu. Predicting gene-disease associations with manifold learning. In Bioinformatics Research and Applications, pages 265–271, Cham, 2018.

[49] Marinka Žitnik, Edward A Nam, Christopher Dinh, Adam Kuspu, Gad Shaulesky, and Blaz Zupan. Gene prioritization by compressive data fusion and chaining. PLoS Computational Biology, 11(10), 2015.

[50] Marinka Žitnik and Blaz Zupan. Jumping across biomedical contexts using compressive data fusion. Bioinformatics (Oxford, England), 32(12):i90–i100, 2016.

[51] Jianing Xi, Ao Li, and Minghui Wang. A novel unsupervised learning model for detecting driver genes from pan-cancer data through matrix tri-factorization framework with pairwise similarities constraints. Neurocomputing, 296:64–73, 2018.

[52] Pouya Zakeri, Jaak Simm, Adam Arany, Sarah ElShal, and Yves Moreau. Gene prioritization using bayesian matrix factorization with genomic and phenotypic side information. Bioinformatics, 34(13):4475–4486, 2018.

[53] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In ICLR, 2013.

[54] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. KDD, 2016:855–864, Aug 2016.

[55] Bryan Perozzi, Ram Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In KDD, pages 701–710, 2014.

[56] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In Proceedings of the 24th International Conference on World Wide Web, pages 1067–1077, 2015.

[57] Mona Alshahrani and Robert Hoendorf. Semantic disease gene embeddings (smudge): phenotype-based disease gene prioritization without phenotypes. Bioinformatics, 34(17):i901–i907, 2018.

[58] Yun Xiong, Mengjie Guo, Lu Ruan, Xiangnan Kong, Chunlei Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Xueying Zhu. Integrating node embeddings prediction. models harnessing graph neural networks for disease-gene prediction. IEEE Bioinformatics Research and Applications, 2018.

[59] Peng Han, Peng Yang, Peilin Zhao, Shao Shang, Yong Liu, Jiayu Zhou, Xin Gao, and Panos Kalnis. Gcn-mf: Disease-gene association identification by graph convolutional networks and matrix factorization. In KDD, pages 705–713, 2019.

[60] Xiaochuan Wang, Yuchong Gong, Jing Yi, and Wen Zhang. Predicting gene-disease associations from the heterogeneous network using graph embedding. In BIBM, pages 504–511, 2019.

[61] Livxing Zhu, Zhaolin Hong, and Haoran Zheng. Predicting gene-disease associations via graph embedding and graph convolutional networks. In BBM, pages 382–389. IEEE, 2019.

[62] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Y Su Philip. A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, 2020.

[63] Mengying Sun, Sendong Zhao, Cormandar Gilvary, Olivier Elemento, Jiayu Zhou, and Frei Wang. Graph convolutional networks for computational drug development and discovery. Briefings in Bioinformatics, 06 2019. bb042.

[64] Ruizhe Cai, Xuexin Chen, Yuan Fang, Min Wu, and Yuexing Hao. Dual-dropout graph convolutional network for predicting synthetic lethality in human cancers. Bioinformatics, 2020.

[65] Yahui Long, Min Wu, Chee Keong Kwok, Jiawei Luo, and Xiaoli Li. Predicting human microbe-drug associations via graph convolutional network with conditional random field. Bioinformatics, 2020.

[66] Daniel Smilkov, David Ziegler, and Ian Goodfellow. A simple solution to a common problem. In ICML, 2017.

[67] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. KDD, 2016:855–864, Aug 2016.

[68] Bryan Perozzi, Ram Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In KDD, pages 701–710, 2014.

[69] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In Proceedings of the 24th International Conference on World Wide Web, pages 1067–1077, 2015.

[70] Mon Alishahram and Robert Hoendorf.Semantic disease gene embeddings (smudge): phenotype-based disease gene prioritization without phenotypes. Bioinformatics, 34(17):i901–i907, 2018.

[71] Yun Xiong, Mengjie Guo, Lu Ruan, Xiangnan Kong, Chunlei Tang, Yangzong Zhu, and Wei Wang. Heterogeneous network embedding enabling accurate disease association predictions. BMC Medical Genomics, 12(10):186, 2019.

[72] K. Yang, R. Wang, G. Liu, Z. Shu, N. Wang, R. Zhang, J. Yu, J. Chen, X. Li, and X. Zhou. Hergepred: Heterogeneous network embedding representation for disease gene prediction. IEEE Journal of Biomedical and Health Informatics, 23(4):1805–1815, 2019.

[73] Aditya Rao, Saipradeep VG, Thomas Joseph, Sujatha Kotte, Naveen Sivadasan, and Rajgopal Srinivasan. Phenotype-driven gene prioritization for rare diseases using graph convolution on heterogeneous networks. BMC Medical Genomics, 11(1):57, Jul 2018.

[74] Yu Li, Hiroyuki Kusuhara, Peng Yang. Le Song, and Xin Gao. Pgcn: Disease gene prioritization by disease and gene embedding through graph convolutional neural networks. bioRxiv, page 532226, 2019.

[75] Vikash Singh and Pietro Lio. Towards probabilistic generative models harnessing graph neural networks for disease-gene prediction. ICML Workshop on Computational Biology, 2019.

[76] Sezin Kircali Ata, Le Ou-Yang, Yuan Fang, Chee-Keong Kwoh, Min Wu, and Xiao-Li Li. Integrating node embeddings and biological annotations for genes to predict disease-gene associations. BMC Systems Biology, 12(9):138, 2018.
Marta Torrens, Alex Bravo, Janet Pinerio, Ferran Sató, and Laura I. Furlong. PsyGeNET: a knowledge platform on psychiatric disorders and their genes. *Nucleic Acids Research*, 31(18):3075–3077, 05 2015.

[92] Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Roy McCarron, Jolene Wiegars, Thomas C Wiegars, and Carolyn I Mattingly. The Comparative Toxicogenomics Database: update 2019. *Nucleic Acids Research*, 47(D1):D948–D954, 09 2019.

[93] Tao-yang Fu, Wang-Chien Lee, and Zhen Lei. HIN2Vec: multi-network embedding. In conference proceedings of the 23rd International Conference on Discovery Science (DS) 2020.

[94] Ye Yuan and Ziv Bar-Joseph. Deep learning for inferring gene regulatory network plasticity. *Genome biology*, 20(1):110, 2019.

[95] Sankha Subhra Mullick, Shounak Datta, and Swagatam Das. Generative adversarial minority oversampling. In *ICCV*, pages 2255–2261, 2018.