Network-based methods for disease-gene prediction

Lorenzo Madeddu1,4, Giovanni Stilo2,4, and Paola Velardi3,4

1 Translational and Precision Medicine Department, Sapienza Università di Roma, Roma, Italy
2 Department of Engineering and Information Sciences and Mathematics, Università degli Studi dell’Aquila, L’Aquila, Italy
3 Computer Science Department, Sapienza Università di Roma, Roma, Italy
4 Intelligent Information Mining - Research Group
http://iim.di.uniroma1.it, http://iim.disim.univaq.it
madeddu@di.uniroma1.it, giovanni.stilo@univaq.it, velardi@di.uniroma1.it

Abstract. We predict disease-genes relations on the Human Interactome network using a methodology that jointly learns functional and connectivity patterns surrounding proteins. Contrary to other data structures, the Interactome is characterized by high incompleteness and absence of explicit negative knowledge, which makes predictive tasks particularly challenging. To exploit at best latent information in the network, we propose an extended version of random walks, named Random Watcher-Walker (RW²), which is able to learn rich representations of disease genes (or gene products) features. Our method successfully compares with the best known system for disease gene prediction, and other state-of-the-art graph-based methods. We perform sensitivity analysis and apply perturbations to ensure robustness. In contrast with previous studies, our results demonstrate that connectivity alone is not sufficient to classify disease-related genes.

1 Introduction

In the last decades, the evolution of medical knowledge has been supported by academic research and technological developments, on the one hand providing a continuously growing set of biomedical data and on the other hand revealing a complexity only perceived until now. In this context, biological networks have become a central hub of multidisciplinary research [1], to address important challenges on both diagnostic and therapeutic aspects, such as drug development and disease classification [2][3][4].

Network Medicine [5] (NM) is a relatively recent approach to analyze the complexity of biomolecular structures. NM proposes to exploit the network topology (e.g. the relations among biological entities) and the network dynamics (e.g. the information flow across the network) to better understand the pathogenic behavior of complex molecular interconnections, that standard reductionist5 approaches cannot detect. A central finding of NM is the following [5]: "If a gene or molecule is involved in a specific biochemical process or disease, its direct interactors might also be suspected to have some

5 according to reductionism, a single disease cause can be identified by decoupling the complex biological or medical phenomenons in multiple components.
role in the same biochemical process. In line with this ‘local’ hypothesis, proteins that are involved in the same disease show a high propensity to interact with each other”. Several studies have been published in support of this hypothesis, such as [6], [7], [8] and others.

It is important to stress the potential impact of network methods to progress in this field. In fact, traditional ways to assess the role of genes in diseases involve time-consuming and extremely expensive\(^6\) statistical studies based on sequencing the DNA of a large number of patients affected by a given disease, known as Genome-Wide association studies (GWAS). In this context, network science and machine learning methods can be effective in reducing the number of alternatives to be explored in clinical experiments.

The objective of this paper is to contribute to the problem of predicting disease-related genes. We present a graph-based approach, based on an extended notion of random walks, to extract topological information and functional properties of local substructures within the human interactome network. Detected patterns are then used to train a machine learning predictor. Our method advances the state of the art, by successfully comparing with the best known system for disease gene prediction, and other known graph-based methods. Furthermore, our study brings new insight to the subject matter, showing that connectivity alone is not sufficient to reliably identify disease-related genes.

2 Related Work

Recent research fields such as System Biology and Network Medicine (NM) [5] has led to new approaches integrating the so called -omics fields of study (genomics, proteomics and metabolomics) and network science. In these studies, complex physical and structural interactions between molecules are modeled as a graph structure, called Interactome. As mentioned in the introduction, the driving idea of NM is that the study of network topology and dynamics can accelerate discovery of new biological interactions and pathways [9], which in turn will drive progress on disease treatments and personalized medicine. However, the application of "standard" graph-based methods to the human Interactome is complicated by the fact that it is estimated that only 20% of existing interactions are actually modeled [10]. Another problem is absence of explicit encoding of negative knowledge: when a relation is not present between two biological entities, we don’t know if the relation actually does not exist or if it is still unknown\(^7\). Our intuition is that these specific characteristics of the Interactome network cause many commonly used graph-mining methods to be ineffective. For example, [11] and [7] found that community detection algorithms and centrality measures fail to identify relevant structures. On the other hand, the interest of scientists in NM remains high, since a growing interdisciplinary effort gives hope for an acceleration of results in this field.

Disease Gene Prediction (DGP) is a relevant, but still open, research topic, since the genetic bases of diseases are largely unknown. Currently, only 10% of genes have a

---

\(^{6}\) https://www.genome.gov/27541954/dna-sequencing-costs-data/

\(^{7}\) This is motivated by the inherent complexity of negative hypothesis testing in this field.
known association with some disease [5]. Genome-Wide associations studies (GWAS) have led to the collections of such associations in databases, like OMIM [12] and DisGenNet [13]. However, as mentioned in the introduction, GWAS studies are very expensive and labor intensive. Following the categorization in [5], there exist several interactome-based approaches for the DGP problem, divided into three categories:

1. **Linkage Methods**: Candidate disease genes (or gene products\(^8\)) are hypothesized to have a direct interaction with known disease genes in the interactome. For example, in [14] candidate disease genes are identified in the direct neighborhood of known disease genes and then filtered using their chromosomal locations (the authors check if the chromosomal location falls within one or more "disease loci").

2. **Module-based Methods**: Candidate genes belonging to the same neighborhoods are more likely to be involved in the same diseases. These approaches are also know as "disease module-based", like for example [7,8]. Note that the notions of "neighborhood" and "module" are vague here, and standard community detection algorithms do not work. Both approaches expand disease modules by exploiting the structure of the network. The main idea of DIAMOnD [7] is based on the use of a connectivity significance measure, designed to take advantage of the weak interconnection properties of the interactome. Using this metric, DIAMOnD first generates a connection ranking for each node, with respect to a chosen disease module. DIAMOnD works by iteratively expanding a single disease module with the first ranked node identified in each iteration. Unlike DIAMOnD, Gladiator [8] considers multiple disease modules simultaneously. Gladiator is based on the intuition that nodes with common phenotypes (common sets of symptoms) are also likely to share molecular mechanisms\(^9\). In order to predict gene-disease relationships, Gladiator uses a simulated annealing algorithm that considers both information on phenotypic similarity and protein interconnection.

3. **Diffusion Methods**: are ranking approach based on the random-walk technique [15,16,17,18,19]. For example, in order to find novel disease-gene candidates, [16] introduce random walk with restart, starting from genes known to be associated with a given disease.

In this paper, we introduce a fourth category, that we call Representation Learning Methods, consisting in the use of graph-based methods [20][21] to learn rich representations of disease genes (or gene products) features, followed by a machine learning predictor for detecting candidate genes. As an example in this category, we present the Random Watcher-Walker (\(RW^2\)), described in Section 3.

### 3 Description of the Method

We predict disease genes using a graph-based methodology which jointly learns functional and connectivity patterns surrounding proteins in the human interactome. The

---

\(^8\) in gene prediction literature, proteins are commonly used in place of genes, since each gene tells the cell how to put together the building blocks for one specific protein. Therefore, studies in this field consider protein-protein interaction networks in place of genes.

\(^9\) One of the problems of Gladiator is that the use of phenotypic information as a predictor of disease-gene relatedness for multi-factorial diseases is not assessed in literature.
network model $G(V,E)$ is shown in Figure 1: nodes $v \in V$ are proteins, and edges $e(i,j) \in E$, $i,j \in V$ are protein-protein interactions (as derived from three databases: IntAct [22], BioGrid [23], MINT [24]. This structure is also referred to in literature as PPI (Protein-Protein Interaction) network\(^\text{10}\).

In our approach, each node $v$ is further described by a multi-dimensional feature vector $f(v)$. In the present study, besides diseases (from OMIM [12]), we consider two additional relevant\(^\text{11}\) dimensions: tissues (from the The Human Protein Atlas [25]) and pathways\(^\text{12}\) (from KEGG [26]). For each dimension, a valued cell in $f(v)$ indicates that $v$ is related with a given disease, tissue or pathway category, hereafter denoted as "category label" or label. Category labels are provided in the same datasets (OMIM, etc.).

The methodology to predict disease-related genes is in three steps:

**Step 1: Random Watcher Walker**: we collect network connectivity patterns using a novel method, Random Watcher-Walker ($RW^2$). In $RW^2$, the walker, when landing on node $v$, "watches" the node features and selects one label at random with uniform probability in those cells of $f(v)$ that are equals to 1. Next, it "walks" with uniform probability to one of $v$’s neighbors. In this way, random walks embody both functional features of traversed nodes (pathobiological properties such as related diseases, tissues and pathways), and structural features (connected proteins in the interactome).

$RW^2$ can be seen as a label sequence generation where, $v^e$ denote the $e^{th}$ node in the walk, and $l^e$ denote the selected label of $v^e$. The generation process satisfies the following distribution:

$$P(v^e = x, l^e = a | v^{e-1} = y) = \ldots \ldots$$

\(^\text{10}\) Note that different studies have used different knowledge bases to build the PPI.

\(^\text{11}\) following the indications of domain experts

\(^\text{12}\) https://en.wikipedia.org/wiki/Biological_pathway
\[
\begin{cases}
\pi(y, x) \cdot \sigma(x, a) & \text{if } (y, x) \in E \text{ and } a \text{ is a label of } x \\
0 & \text{otherwise}
\end{cases}
\]

where \(\pi(y, x)\) is the normalized transition probability between nodes \(y\) and \(x\); \(\sigma(x, a)\) is the normalized probability of selecting the node-label \(a\) in \(f(x)\).

Note that our Random Watcher Walker approach is meant to exploit one relevant findings of Network Medicine, the "modular" structure of diseases in the interactome, as summarized in Section 2. Furthermore, our intuition is that random walks crossing nodes associated with disease modules that are either close, or intersect each other in the interactome, should have similar label subsequences, since they are extracted from a similar neighborhood. Given the "loose" notion of neighborhood implemented by random walks, similarity patterns might be captured even in presence of highly incomplete knowledge.

**Step 2: Label Embeddings:** collected network connectivity patterns are treated as "contexts" for individual labels, (as shown in Figure 2) much in same way as sentences are contexts for individual words. Context are used to train a Skip-Gram [27] model and learn label embeddings ("dense" vector representations of each label). Label embeddings are used to enrich the multidimensional feature vector \(f(v)\) of each node of \(G\): valued cells are replaced by the respective embedding vectors, producing the enriched feature matrix \(\mathcal{F}(v)\).

![Fig. 2: Example of context for the label "Stomach"](image)

**Step 3: Training:** feature matrices \(\mathcal{F}(v)\) are used to train a fully connected neural network (NN) with Softmax activation function, for predicting disease-gene associations (Figure 3). The system’s output is a \(|D| + 1\)-dimensional probability vector, where \(|D|\) is the number of considered disease labels and the additional class label is UNK, to state absence of known disease relations for a given node.

### 4 Evaluation Methodology

In this section we describe the dataset used for our experiments, the adopted data transformation methodology, and the experimental strategies and setup.

**Dataset:** As we already remarked, biomedical data suffer from two problems:

- **Absence of Negative Instances:** Biomedical data are not characterized by negative instances. Databases usually do not collect the information about the absence of a
relationship between a gene and a disease.

- **Incompleteness**: The high incompleteness of the human interactome (see Section 1) implies that the information to be exploited can be very sparse.

For this reason, evaluating on synthetic data may lead to inconclusive results, since the generated datasets can hardly satisfy the statistical properties of the real interactome. In our experiment, for the purpose of comparison, we use the same interactome (PPI) network used in DIAMOnD [7]. In this network, there are 13458 proteins and 70 disease categories. Only 1536 nodes are related to at least 1 disease, of which, 766 have co-morbidities (i.e., they are known to be related to more than one disease). In addition to information on disease categories, we enrich feature vectors $f(v)$ with information on 57 tissue types\(^\text{13}\) and 321 pathways.

**Data Transformation**: Figure 4 shows, for each node $v$ of the interactome, the enriched multidimensional feature matrix $\mathcal{F}(v)$ (left) and the corresponding ground-truth output vectors $\mathcal{D}$ (right) to be used for training. Dark cells in $\mathcal{F}(v)$ represent embedding vectors associated to valued feature labels in the original $f(v)$, while white cells are zero vectors. The $(|\mathcal{D}| + 1)$-dimensional ground-truth vector $\mathcal{D}$ has the $i^{th}$ cell equal to 1 if the node is known to be associated to the corresponding disease $d \in \mathcal{D}$. The last cell of this vector is 1 if no diseases are known to be associated to the node.

Clearly, the dataset in Figure 4 cannot be used for training, because the NN would trivially learn that if a disease vector is valued in $\mathcal{F}(v)$, then the corresponding cell of the output should be 1. To avoid trivial learning, we train the NN using modified feature matrices , as shown in Figure 5.

1. If a node $v$ is known to be related to a single disease $d$, $D^v : \{d\}$, $|D^v| = 1$ , then the corresponding embedding vector from the feature matrix $\mathcal{F}(v)$ is replaced

\(^{13}\)taken from https://www.proteinatlas.org/humanproteome/tissue
with a zero vector. For example, nodes 2), 3) and 5) of Figure 4 are modified as in Figure 5. Note that in this way a node with no valued cells in the disease dimension (instances 1 and 2 of Figure 5), can either be “unknown” - which corresponds to a 1 in the last cell of the ground-truth vector - or known to be related with one disease. The NN is therefore encouraged to learn distinguishing between these two cases based on the embedding vectors of the other dimensions, pathways and tissues.

2. If a node is known to be related to diseases \( D_v = \{d_1, \ldots, d_m\} \), then its feature matrix is duplicated into \( m \) matrices. Each duplicated matrix is associated to only one disease \( d_k \in D_v \). In each duplicated feature matrix \( F(v)_k \), we replace the corresponding embedding vector of the associated disease \( d_k \) with a zero vector, and we keep only the 1 associated with \( d_k \) in the related ground-truth vector. For example, node 4 in Figure 4 is duplicated in 4.a and 4.b in Figure 5. This encourages the NN to learn also from co-morbidities.

**Experimental Strategies:** We first adopt a computationally expensive, but unbiased, strategy (method 1). With method 1, the Random Watcher-Walker \( RW^2 \) collects connectivity patterns from a “degraded” network from which 20% of the information on node labels has been removed. Node labels are transformed as for in Figure 5. Next, we learn label embeddings and train the NN network. Note that in method 1, we repeat all the three steps of the algorithm (random walks, label embedding and NN training) for each of the \( K \) different train/test split of the dataset used during the \( K \)-fold evaluation of the system. We also introduce a less computationally expensive strategy (method 2) in which the previously described data transformation is performed after label embedding without degrading the interactome at each fold evaluation. The dataset shown in Figure 5 is used to train the NN with 80% 20% split and 10-fold evaluation (\( K = 10 \)). How-
ever, learning the embeddings on the complete Interactome network could potentially bias the results of the subsequent NN training phase.

**Experimental setup:** Tables 1 and 2 show the system parameters for our best experiment. Sensitivity to parameters is discussed in Section 5.2.

| Random Walk Parameter                     | Value |
|-------------------------------------------|-------|
| Label Embedding Length                    | 128   |
| Walk Length                               | 20    |
| Number of random walks per node           | 100   |
| $p$                                       | 1     |
| $q$                                       | 1     |
| Skip-Gram context window                  | 3     |
| Skip-Gram Epochs                          | 3     |

Table 1: Model parameters.

| NN Parameter                     | Value         |
|----------------------------------|---------------|
| Hidden Layer                    | 0             |
| Activation Function             | Softmax       |
| Loss                             | Binary Cross-entropy |
| Optimizer                       | Adam          |
| Batch Size                      | 100           |
| Epochs                          | 5             |

Table 2: Model parameters.

5 Experiments

5.1 Comparison with other methods

Given the previously outlined characteristics of biomedical data, evaluation measures such as precision, accuracy and f-score are ineffective. In line with other works [7][8] in this domain, we use Recall@k, the fraction of correctly predicted items at rank k. In all our experiments, we set the $k$ value of Recall@k to 1, since as explained in Section 1, the intended use of network methods in medicine is to exploit the results with the highest confidence, to narrow the scope of expensive and labor intensive clinical tests.

We compare our system (trained according to methods 1 and 2, described in Section 4) with:

1. A baseline method which uses only functional information, i.e. the feature vectors $f(v)$ without label embeddings. This corresponds to exploiting only functional (feature) similarity.
2. Node2Vec [20] combined with the NN Classifier. Node2Vec (N2V) is a state-of-the-art random walk algorithm for Node-Embedding generation. Contrary to our approach, Node2Vec uses a standard random-walk strategy which exploits only the connectivity patterns of a network. Consequently, Node2Vec generates embeddings containing only connectivity information.

3. Graph Convolutional Networks (GCN) [21], an inductive semi-supervised method to classify nodes in a network, based on an efficient variant of convolutional neural networks, which operates directly on graphs. GCN handle both the functional and connectivity information of the network but in a completely different way compared to our method. We test GCN with default parameters (as specified in [21]).

4. DIAMOnD, which is commonly considered the state of art and most cited study on disease gene prediction (see Section 2). DIAMOnD, like N2V, exploits only connectivity information.

For all these methods, during the training phase we remove 20% of the information concerning disease-node relationships, and use these data for testing. Each experiment is repeated 10 times with different splits of the learning and test set (10-fold cross evaluation). Next, we compute the Recall@1 and average over all folds. Note that computing Recall@1 for DIAMOnD is not straightforward. In DIAMOnD evaluation experiments, described in [7], diseases are considered one at the time. For each disease $d$, they remove node-disease associations from a given fraction of nodes $N'_d$ among those known to be related with $d$. Next, they apply an iterative method in which, at each iteration, they add a new node $n$ (the most likely node among those considered) to the current set of nodes believed to be related to $d$. In their paper, the authors perform 200 iterations and lastly they compute the recall, i.e. the fraction of disease nodes retrieved by their method, among those $(N'_d)$ that were initially removed. Although the authors do not explicitly set/report a k value for the Recall, we can assume that setting $k=1$ for their system is an upper bound of the real system performances. In our experiments, we use the software made available by the authors, and adopt exactly the same iterative methodology, removing 20% of disease-node associations, like for the other compared methods.

The results of all comparative experiments are shown in Table 3. The first two rows of Table 3 show the performance of our system. The difference between methods 1 and 2 (see Section 4) is less than 1% and is not statistically significant ($p = 0.14$). This was expected, given the very sparse nature of the interactome graph. A comparison between the two methods was however necessary to exclude any learning bias. Given the negligible difference between the two methods, the less computationally intensive method 2 can be adopted instead of method 1.

Table 3 also shows that N2V and GCN perform very poorly: they are unable to identify patterns that help to predict disease nodes. As a result, they always predict the majority class (the label "unknown") which explains why the STD is zero. For DIAMOnD, we report the recall value obtained as previously explained, averaging over all diseases, different splits, and 200 iterations for each experiment. We remark that in [7] the reported performance is higher, but limited to two diseases, lysosomal storage diseases and lipid metabolism disorders, that show higher density of the respective modules. For the purpose of completeness, Table 4 compares $RW^2$ and DIAMOnD on
these very same diseases. Furthermore, the authors of [8], in an experiment considering all diseases (on a slightly different dataset), reported that DIAMOnD was “able to recover 13.3% of the removed associations”, which in line with the performance value (14%) in Table 3.

| NN Parameter | R@1 | STD |
|--------------|-----|-----|
| RW\(^2\) (method 1) | 35% | 0.008 |
| RW\(^2\) (method 2) | 36% | 0.008 |
| Baseline | 2% | 0.002 |
| N2V | 1% | 0 |
| GCN | 1% | 0 |
| DIAMOnD | 14%* | 0.011 |

Table 3: Average Recall@1 and standard deviation over 10 folds. *Note that, as explained, for DIAMOnD the reported performances represent an upper-bound.

| Disease Module | RW\(^2\) | DIAMOnD |
|----------------|---------|---------|
| lysosomal storage diseases | 79% | 50% |
| lipid metabolism disorders | 32% | 31% |

Table 4: Comparison on two diseases

The results of Table 3 seem rather unexpected: none of the considered methods is able to predict disease nodes, except RW\(^2\) and, with some “caveat” given the adopted evaluation procedure, DIAMOnD. The purpose of next Sections is to interpret more clearly our results.

### 5.2 Robustness and sensitivity analysis

First off, we perform a robustness test of RW\(^2\), by randomly rewiring edges\(^14\), relabeling node features, and both. The result is shown in Table 5. The table shows that while relabeling severely affects system’s performance, rewiring only results in a 3% points decrease in performance. Although this difference is statistically significant \((p < 0.02)\), it clearly shows that connectivity is a weak feature. This is in line with Table 3, that shows extremely poor performances for methods based only on connectivity patterns (N2V and GCN).

\(^{14}\) using the method in http://bit.ly/2N0sKHf
Next, we analyze the network sensitivity to parameters. First, we found that increasing the number of layers of the neural network (step 3 of the pipeline) does not improve results\(^\text{15}\). Considering the entire pipeline, only two parameters were found to affect the performance: the dimension of embedding vectors, and the dimension of the window (context) used during the label embedding phase. Figure 6 shows that a sufficiently high number of dimensions are needed (>100), which is in line with other domains in which embeddings have been used. Figure 7 shows that best performances are obtained with smaller left-right contexts (a window size between 1 and 5). This confirms that the diameter of disease modules (remember that disease modules are vaguely defined as an "area" where nodes related to the same disease tend to reside) is relatively small, in line with other studies, for example [28], stating that the median distance between components in a module is almost 2.9, and [4] where the diameter of a disease module is estimated to be 1.8 in the average.

\[ \text{Table 5: Robustness test: rewiring and relabeling the network} \]

| Method       | STD  |
|--------------|------|
| R\(W^2\)     | 0.0086 |
| Rewiring     | 0.0006 |
| Relabel      | 0.0003 |
| Both         |       |

![Fig. 6: Sensitivity analysis: performance as a function of the dimension of embedding vectors.](image)

**5.3 Discussion**

To summarize, all experiments confirm that connectivity patterns only marginally contribute to identifying gene-disease relationships. This is demonstrated both by the very

\(^\text{15}\) Although more experiments with different and more complex learners might be needed, our intuition is that data quality -namely, incompleteness and absence of negative evidence - is too low for deep methods to learn regularities.
poor performances of methods, such as N2V and GCN, that only exploit connectivity, and by our robustness tests. Similarly, using only functional features does not produce relevant results, as shown by the baseline method performance. Instead, $RW^2$ is able to discover specific combinations of connectivity and functional features that have a higher probability of being found in the vicinity of a node related with a given disease. The scarce contribution of connectivity features to the discovery of disease modules contradicts previous results in this domain, but is in agreement with a very recent, although preliminary, study published by a prestigious institution [28] were the authors demonstrate that 90% of disease-related nodes do not correspond to single well-connected components in the human interactome network. Instead, nodes associated with a single disease tend to form many separate connected components/regions in the network. In particular, the authors observe that "current methods disregard loosely connected proteins when making predictions, causing many disease module components in the network to remain unnoticed". Our study confirms this finding, and demonstrates that $RW^2$ is a better method to capture common features of such sparse regions: first, the Random Watcher Walker jointly captures connectivity and functional patterns in the vicinity of nodes; second, label embeddings allow to optimize the combination of features types that are more predictive of each disease. Note that the notion of "vicinity" in embedding methods is more relaxed than "connectivity", since the relative distance between two labels is not fixed, but only constrained by the length of the context window. We also found that performance degrades when the window length exceeds ±5, which implies that "some" vicinity among nodes related with the same disease does exist.

6 Conclusions

We have shown that $RW^2$ is able to learn regularities in complex networks characterized by high incompleteness and absence of negative knowledge. Our study, together with [28], also poses new problems to the Network Medicine field of research. On the one hand, the data quality of the interactome network seems too low for many popular and high performing network methods to produce useful results; on the other hand,
the initial hypotheses of Network Medicine are not denied by our results, since $RW^2$ shows that some structure in the data indeed exists. Rather, the challenge is to design effective methods to capture at best such sparse regularities (as we have shown, $RW^2$ is a promising solution), and secondly, to enrich or combine in a more effective way different available sources of information. In this work, we use tissues and pathways, since they are a readily available and sufficiently reliable information, however we plan to investigate other sources in our future work.

References

1. D. Yu, M. Kim, G. Xiao, and T. H. Hwang, “Review of biological network data and its applications,” *Genomics & informatics*, vol. 11, pp. 200–210, 12 2013.
2. F. Cheng, R. J. Desai, D. Handy, and et al., “Network-based approach to prediction and population-based validation of in silico drug repurposing,” in *Nature Communications*, 2018.
3. Z. Wu, Y. Wang, and L. Chen, “Network-based drug repositioning,” *Molec. bioSystems*, vol. 9, 03 2013.
4. J. Menche, A. Sharma, M. Kitsak, and et al., “Uncovering disease-disease relationships through the incomplete interactome,” *Science*, vol. 347, no. 6224, 2015.
5. A.-L. Barabási, N. Gulbahce, and J. Loscalzo, “Network medicine: A network-based approach to human disease,” *Nature reviews. Genetics*, vol. 12, pp. 56–68, 01 2011.
6. S. Vlaic, T. Conrad, C. Tokarski-Schnelle, and et al., “Modulediscreever: Identification of regulatory modules in protein-protein interaction networks,” *Scientific Reports*, vol. 8, no. 1, p. 433, 2018.
7. S. D. Ghiassian, J. Menche, and A. Barabási, “A disease module detection (diamond) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome,” *PLoS Computational Biology*, vol. 11, no. 4, 2015.
8. Y. Silberberg, M. Kupiec, and R. Sharan, “Gladiator: a global approach for elucidating disease modules,” *Genome Medicine*, vol. 9, p. 48, May 2017.
9. S. Y. Chan and J. Loscalzo, “The emerging paradigm of network medicine in the study of human disease,” *Circulation Research*, vol. 6, p. 359–374, 3 2012.
10. K. Venkatesan, J.-F. Rual, A. Vazquez, and et al., “An empirical framework for binary interactome mapping,” *Nature methods*, vol. 6, pp. 83–90, 01 2009.
11. A. An, B. Andreopoulos, M. Schroeder, and X. Wang, “A roadmap of clustering algorithms: finding a match for a biomedical application,” *Briefings in Bioinformatics*, vol. 10, pp. 297–314, 02 2009.
12. A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, “Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders,” *Nucleic Acids Research*, vol. 33, no. suppl_1, pp. D514–D517, 2005.
13. J. Piñero, A. Bravo, N. Queralt-Rosinach, and et al., “Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants,” *Nucleic Acids Research*, vol. 45, no. Database-Issue, pp. D833–D839, 2017.
14. M. Oti, B. Snel, M. A. Huynen, and H. G. Brunner, “Predicting disease genes using protein–protein interactions,” *J. of Medical Genetics*, vol. 43, no. 8, pp. 691–698, 2006.
15. O. Vanunu, O. Magger, E. Ruppin, and et al., “Associating genes and protein complexes with disease via network propagation,” *PLoS Computational Biology*, vol. 6, pp. 1–9, 01 2010.
16. S. Köhler, S. Bauer, D. Horn, and P. Robinson, “Walking the interactome for prioritization of candidate disease genes,” *American journal of human genetics*, vol. 82, pp. 949–58, 05 2008.
17. X. Wu, R. Jiang, and M. Zhang, “Network-based global inference of human disease genes,” *Molecular Systems Biology*, vol. 4, pp. 189–1, 01 2008.

18. Y. Li and J. Chandra Patra, “Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network,” *Bioinformatics (Oxford, England)*, vol. 26, pp. 1219–24, 03 2010.

19. M. Leiserson, F. Vandin, H. Wu, and et al., “Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes,” *Nature Genetics*, 12 2014.

20. 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*, KDD ’16, (New York, NY, USA), pp. 855–864, ACM, 2016.

21. T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *CoRR*, vol. abs/1609.02907, 2016.

22. S. Orchard, M. Ammari, B. Aranda, and et al., “The mintact project-intact as a common curation platform for 11 molecular interaction databases,” *Nucleic Acids Research*, vol. 42, no. D1, pp. D358–D363, 2014.

23. C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, “Biogrid: a general repository for interaction datasets,” *Nucleic Acids Research*, vol. 34, no. suppl_1, pp. D535–D539, 2006.

24. A. Ceol, A. Chatr Aryamontri, L. Licata, and et al., “Mint, the molecular interaction database: 2009 update,” *Nucleic Acids Research*, vol. 38, pp. D532–D539, 2010.

25. M. Uhlén, L. Fagerberg, B. Hallström, and et al., “Tissue-based map of the human proteome,” *Science*, vol. 347, no. 6220, 27–30, 2000.

26. M. Kanehisa and S. Goto, “Kegg: Kyoto encyclopedia of genes and genomes,” *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

27. T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *CoRR*, vol. abs/1301.3781, 2013.

28. M. Agrawal, M. Zitnik, and J. Leskovec, “Large-scale analysis of disease pathways in the human interactome,” in *Pacific Symposium on Biocomputing*, vol. 23, pp. 111–122, 2018.