Predicting Disease Genes for Complex Diseases using Random Watcher-Walker

Lorenzo Madeddu
Sapienza university of Rome
Rome, Italy
lorenzo.madeddu@uniroma1.it

Giovanni Stilo
University of L'Aquila
L'Aquila, Italy
giovanni.stilo@univaq.it

Paola Velardi
Sapienza university of Rome
Rome, Italy
velardi@di.uniroma1.it

ABSTRACT

In this paper we propose an extended version of random walks, named Random Watcher-Walker (RW^2), to predict disease-genes relations on the Human Interactome network. RW^2 is able to learn rich representations of disease genes (or gene products) features by jointly considering functional and connectivity patterns surrounding proteins. 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. Differently from previous studies, our results demonstrate that connectivity alone is not sufficient to classify disease-related genes.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Learning latent representations**; *Supervised learning*; *Neural networks*; • **Applied computing** → **Biological networks**; **Bioinformatics**.

KEYWORDS

Protein-Protein Interaction, Biological, Disease, Prediction, Neural Network

ACM Reference Format:
Lorenzo Madeddu, Giovanni Stilo, and Paola Velardi. 2020. Predicting Disease Genes for Complex Diseases using Random Watcher-Walker. In *The 35th ACM/SIGAPP Symposium on Applied Computing (SAC '20), March 30-April 3, 2020, Brno, Czech Republic.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3341105.3373979

1 INTRODUCTION

In the last decades, academic research and technological developments support the evolution of medical knowledge, on the one hand, providing a continuously growing set of biomedical data and on the other side revealing a complexity only perceived until now. In this context, biological networks have become a central hub of multidisciplinary research [29], to address essential challenges on both diagnostic and therapeutic aspects such as drug development and disease classification [6, 15, 28].

Network Medicine [3] (NM) is a relatively recent approach to analyze the complexity of biomolecular structures. The standard reductionist approach tries to identify a single disease by decoupling the complex biological or medical phenomenons into multiple components. NM, surpass the standard reductionist approach, exploiting the network topology (e.g. the relations among biological entities) and the network dynamics (e.g. the information flow across the network) to understand the pathogenic behaviour of complex molecular interconnections. A central finding of NM[3] is the following: "If a gene or molecule is involved in a specific biochemical process or disease, *its direct interactors might also be suspected to have some 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 published studies, such as [8, 22, 26] support this hypothesis.

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[1] 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. Recent studies in Network Medicine and genetic research, focusing on the analysis of disease-gene relationships, highlight that only the 10% of genes is related to a disease (disease gene). In addition, it has been noted that disease genes related to the same disease, tend to be limited in a structural context in protein-protein interaction (PPI) networks. 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 sub-structures within the human interactome network. Detected patterns are then used to train a machine learning predictor. The main contributions of this paper are summarised as follows:

(1) We present a new framework for disease gene prediction based on a variant of random walks, named Random Watcher-Walker (RW^2).
(2) We show that exploiting connectivity properties alone is not sufficient to reliably identify disease-related genes.
(3) We show that, given high incompleteness of the interactome network, a careful aggregation of diseases into categories might considerably help predictive methods.

2 RELATED WORK

Recent research fields such as System Biology and Network Medicine (NM) [3] have led to new approaches integrating the so-called *-omics* fields of study (i.e. genomics, proteomics and metabolomics) and

[1] https://www.genome.gov/27541954/dna-sequencing-costs-data/

SAC '20, March 30-April 3, 2020, Brno, Czech Republic

ACM ISBN 978-1-4503-6866-7/20/03…$15.00
https://doi.org/10.1145/3341105.3373979

network science. In these studies, complex physical and structural interactions between molecules are modelled as a graph structure, called *interactome*. As mentioned in the introduction 1, the driving idea of NM is that the study of network topology and dynamics can accelerate the discovery of new biological interactions and pathways [5], which in turn will drive progress on disease treatments and personalised medicine. Several issues complicate the application of computational methods to the human interactome:

(1) Incompleteness: it is estimated that only 20-30% of the existing interactions have been discovered [25]. Predictive tasks in such incomplete environments are particularly challenging.
(2) Reliability: although much protein-protein interaction (hereafter PPI) datasets are available in literature (see Section 3), relationships are not sufficiently reliable, unless experimentally tested on multiple assays (see Section 4 for details). Only few recent experimental efforts are based on this idea, like HI-III [10].
(3) Negative knowledge: when a relationship is not present between two biological entities, we are not sure if it actually does not exist or if it is still unknown. For some types of predictive task, like link prediction, empirical methods have been proposed to simulate negative knowledge (e.g. negative sampling) used to predict negative PPIs. In this approach, negative instances are chosen by randomly pairing proteins and then removing pairs already included in the positive examples. However, the limited reliability of negative sampling in link prediction has been recently demonstrated in more than one study (among the others, [19]), especially since negative links generated by this approach are highly influenced by the presence of hubs in the set of positive PPIs. As a consequence, training and testing with these negative datasets may significantly overestimate performances [12]. More recent efforts are based on the idea of generating reliable datasets of negative PPIs based on multiple experimental assays; however, experimentally tested negative PPIs have not yet been publicly released. Note that, It is practically impossible for biologists, given available techniques, to demonstrate the absence of a relationship. Despite this, the use of negative knowledge (e.g. generated by random sampling) is commonly employed in many studies (e.g., PhenoRank [7]).

These specific characteristics of the interactome network cause many commonly used graph-mining methods to be ineffective. For example, [2] and [8] found that community detection algorithms and centrality measures fail to identify relevant structures, because of incompleteness of the knowledge. 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.

In this paper, we are concerned with a specific predictive task, Disease Gene Prediction (DGP). 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 known association with some disease [3]. Genome-Wide associations studies (GWAS) have led to the collections of such associations in databases, like OMIM [9] and DisGenNet [20]. However, as mentioned in the introduction, GWAS studies are very expensive and labour-intensive. Several PPI-based computational approaches for the DGP problem have been presented in the literature. The most relevant approaches can be divided into three categories[3]:

(1) **Linkage Methods:** Linkage is the "tendency for genes and other genetic markers to be inherited together because of their location near one another on the same chromosome"[2]. Linkage methods are based on the idea that genes, associated with a given disease or disease category, are often in a given linkage interval (i.e. the chromosomal location falls within one or more "disease loci"). The information of the linkage interval can be used to restrict the number of candidate genes for a given disease. Given a disease d_j, let V_j^c be the set of nodes (genes) in the linkage interval of d_j. V_j^c is the set of *candidates* among which the genes related to d_j must be predicted. Let D_j be a set of diseases which are functionally similar to d_j (we can refer to D_j as a disease category). Moreover, let V_j denote the genes which are already known to be related with D_j. Disease genes among the candidates V_j^c are predicted among those in the direct neighbourhood of V_j as in [18]. The main problem with linkage methods is that identifying the causal genes at disease loci is often difficult, as noted in [7], and the co-occurrence of genes in the same chromosomal location is a probable, but not necessary, condition.
(2) **Diffusion Methods:** the majority of these methods, like [11, 13, 24, 27], are still based on linkage intervals to reduce the number of candidate genes for a given disease. However, they rely on more complex connectivity approaches to filter candidates. For example, in order to find novel disease-gene candidates, [11] introduce random walk with restart (*RWR*), starting from genes known to be associated with a given disease category. *RWR* and phenotypic information are used in a recently published method, PhenoRank [7]. PhenoRank exploits the phenotypic similarity of an input disease (query disease) with other human and mouse-mutant diseases. The similarity values between nodes of a PPI network are propagated across the network so that genes that interact with many high scoring genes are highly scored. Eventually,to avoid the bias induced by the fact that less studied genes are less connected within the PPI network, the p-value of each gene score is computed comparing it to the distribution of scores the gene receives for simulated sets of phenotype terms.
(3) **Module-based Methods:** these approaches [8, 22] are based on network connectivity properties. The base hypothesis suppose those candidate genes belonging to the same neighbourhood (or module) are more likely to be involved in the same diseases. Note that notions of "neighbourhood" and "module" are vague here, and standard community detection algorithms fail. Both approaches start with a given disease d_j (or disease category), consider the set of genes known to be associated with d_j - the initial "disease module" - and expand the module by exploiting the structure of the network. The

[2] https://www.medicinenet.com/script/main/art.asp?articlekey=4166

main idea of DIAMOnD [8] 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, concerning 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 [22] considers multiple disease modules simultaneously. Gladiator is based on the intuition that diseases with common phenotypes (common sets of symptoms) are also likely to share molecular mechanisms. In order to predict gene-disease relationships, Gladiator uses a simulated annealing algorithm that considers both information on phenotypic similarity and protein interconnections. One of the problems with this approach is that phenotypic data is not available for all genes [7], potentially influencing the performance of this method.

In this paper, we introduce a graph-based method, Random Watcher-Walker (RW^2), to learn rich representations of gene (or gene products) *features*, followed by a neural network predictor to detect candidate genes.

Differently from other methods surveyed in this Section:

- we do not rely on the linkage interval hypothesis;
- we do not consider diseases (or disease categories) one at the time, but jointly predict all disease-related genes;
- we do not rely on heuristic methods to simulate negative knowledge, which, as already noted, tend to boost performance artificially;
- rather than using ad-hoc PPIs and categorizations, we analyze the influence of different PPIs and disease categorizations on the systems' performance.

3 RW^2 METHODOLOGY

We predict disease genes using a graph-based methodology which jointly learns functional and connectivity patterns surrounding proteins in the human interactome. The network model G(V, E) is shown in Figure 1: nodes $v \in V$ are proteins or protein products, and edges $e(u, v) \in E$, $u, v \in V$ are interactions. In our approach, each node $v \in G$ is further described by a feature vector $f(v)$, which is a one-hot vector where a "1" indicates that a specific disease f_k^j is associated to a node v. Note that we consider mono and poly-genic diseases (those influenced by more than one gene). Furthermore, a gene might be associated with more than one disease. The methodology to predict disease-related genes can be summarized 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 disease 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 neighbours. In this way, random walks embody both *functional* features of traversed nodes (disease labels), and *structural* features (connected proteins in the PPI network). RW^2 can be seen as a label sequence generation where v^e denote the e^{th} node in the walk, and l^e denotes the selected label of v^e. The generation process satisfies the following distribution:

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

$$\begin{cases} \pi(y,x) \cdot \sigma(x,a) & \text{if } (y,x) \in E \text{ and } a \text{ is a label of x} \\ 0 & \text{otherwise} \end{cases}$$

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 finding of Network Medicine, the "modular" structure of diseases in the interactome: 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 neighbourhood.* Given the "loose" notion of neighbourhood implemented by random walks, similarity patterns might be captured even in the 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 the same way as sentences are contexts for individual words. Contexts are used to train a Skip-Gram [16] model and learn *label embeddings* (embeddings are "dense" vector representations of feature labels, a popular method used in Machine Learning to cope with feature sparsity). 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 $\mathscr{F}(v)$.
- **Step 3 - Training**: feature matrices $\mathscr{F}(v)$ 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 the absence of known disease relations for a given node.

4 EVALUATION ENVIRONMENT

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

4.1 Interactomes datasets

PPI networks: Protein-protein interactions are mostly derived from databases curated from the literature (hypothesis-driven), like *IntAct* [17], *BioGrid* [23], *MINT* [4]. These datasets may be affected by inspection bias (also termed study bias or investigational bias[14]) since they depend on the purposes of a study. In our experiments, however, we aim at using highly reliable PPI datasets obtained via clinical tests (discovery-driven), although this may lead to a higher

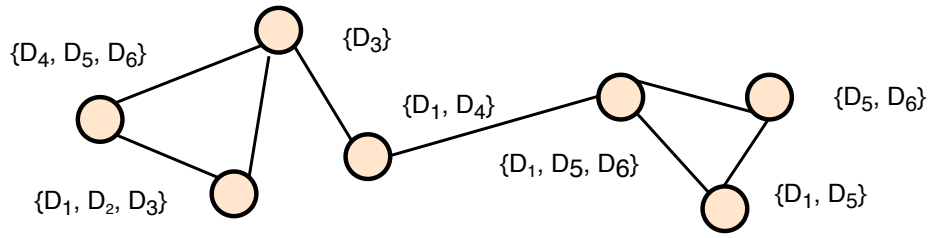


Figure 1: The network model: each node (a gene or gene product) is described by a feature vector. A "1" in a cell means that the considered node is associated with a specific disease category.

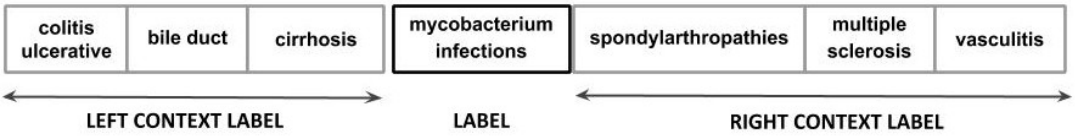


Figure 2: Example of "context" for the disease category *mycobacterium infections*. In each step t of the walk, a node $v(t)$ is randomly selected among those connected with the previous node $v(t-1)$, and next, a label is randomly extracted from $f(v(t))$. The figure shows a fragment of a specific (double) random walk, centred on the label *mycobacterium infections*, a disease label, extracted in step t of the random walk. Left context labels have been extracted in steps $t-1, t-2\ldots$ while right context labels have been extracted in steps $t+1, t+2\ldots$.

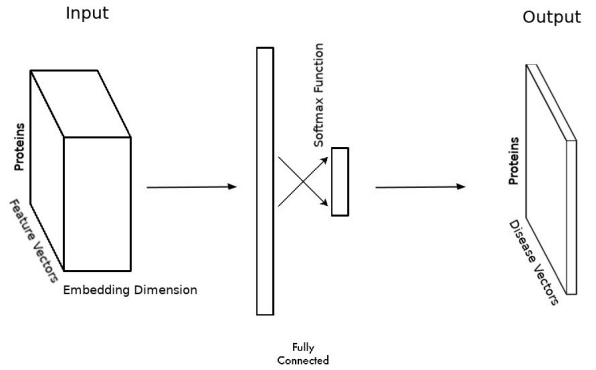


Figure 3: Training the NN with feature matrices

sparsity. Finally, we do not use synthetic datasets, since these generated datasets can hardly satisfy the statistical properties of the real interactome, and may lead to overestimated performances.

In our experiments, we used the following PPIs:

- **DIAMOnD**: For comparison, we use the same interactome (PPI) network used in DIAMOnD [8], obtained from curated literature.
- **HI-III**: this dataset contains protein-protein interactions identified by high throughput yeast two-hybrid screens applied systematically on pairwise combinations of human protein-coding genes using high throughput yeast two-hybrid screens (Discovery-driven or hypothesis-free). The quality of these interactions is further validated in multiple orthogonal assays. The effect of the inspect bias on this type of dataset is negligible[14]. HI-III is publicly available on the HuRI website[3].

Table 1 shows some network statistics. We note that DIAMOnD is slightly more connected, and larger, than HI-III. Another important difference is that nodes in HI-III are isoform proteins, while in DIAMOnD, they are genes. More importantly, relationships in HI-III are considered highly reliable.

Disease categories: Disease categories with a genetic basis were obtained from DIAMOnD[8] (disease-gene associations from OMIM[9] and GWAS[21]), Phenorank[7] or Disgenet [20]. Table 2 shows the effect of applying these three categorization types to the DIAMOnD and HI-III PPI network. Clearly, since the DIAMOnD categorization has been manually conceived for the DIAMOnD PPI network, all categories (70) map to some of the nodes in the network. Similarly to what the authors do, we consider only diseases modules with at least 20 genes associated with it. When we apply the same classification and dimensionality filter to HI-III, only 10 category labels out of 70 have at least one module associated with it. Disgenet categories with the same dimensionality filter are 31 both in HI-III and DIAMOnD. Finally, only 16 Phenorank categories could be associated to HI-III and 12 to DIAMOnD, and we had to reduce the dimensionality filter to the size of 10.

4.2 Data Preparation

Figure 4 shows, for each node v of the interactome, the enriched multidimensional feature matrix $\mathscr{F}(v)$ (left) and the corresponding ground-truth output vectors $\mathscr{D}$ (right) to be used for training. Dark cells in $\mathscr{F}(v)$ represent embedding vectors associated with valued feature labels in the original $f(v)$, while white cells are zero vectors. The $(|D|+1)$-dimensional ground-truth vector $\mathscr{D}$ has the

[3] http://interactome.baderlab.org/about/

PPI network	N. Nodes	N. Edges	Graph Density	Connected Components	Avg. Number of Neighbors
HI-III	8490	54495	0,0015	71	13
DIAMOnD	13458	141272	0,0016	89	20

Table 1: Network statistics

PPI network		DIAMOnD Category Labels (CL)	Disgenet CL	Phenorank CL
HI-III	N. of Different Diseases	10	31	16
	Diseases Nodes (%)	7%	14%	6%
	N. of Disease Nodes	479	1187	1008
DIAMOnD	N. of Different Diseases	70	31	12
	Diseases Nodes (%)	11%	8%	9%
	N. of Disease Nodes	2843	1084	850

Table 2: Effect of using different disease categorizations on different PPIs

i^{th} cell equal to 1 if the node is known to be associated with the corresponding disease $d \in D$. The last cell of this vector is 1 if no disease is known to be associated with the node.

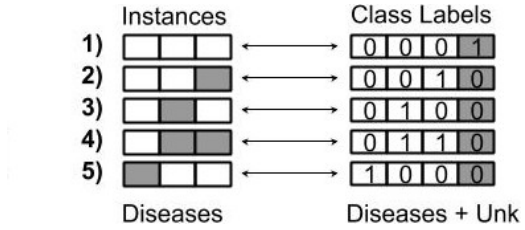


Figure 4: Feature Matrices and ground-truth vectors

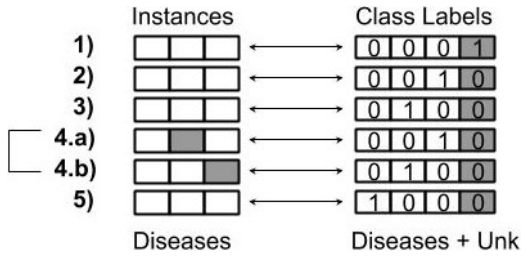


Figure 5: Modified Feature Matrices for the learning phase

The dataset in Figure 4 cannot be used for training, because the NN would trivially learn that if a disease vector is valued in $\mathscr{F}(v)$, then the corresponding cell of the output should be 1. To avoid *trivial learning*, we train the NN using a 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 $\mathscr{F}(v)$ is replaced with a zero vector. For example, nodes 2), 3) and 5) of Figure 4 are modified as in Figure 5. Notice 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. Only connectivity properties may allow a distinction between these cases;

(2) If a node v is known to be related to m diseases $D^v : \{d_1, \ldots, d_m\}$, $|D^v| = m$, then its feature matrix is duplicated into m matrices. Each duplicated matrix is associated with only one disease $d_k \in D^v$. In each duplicated feature matrix $\mathscr{F}(v)^k$ 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 $\mathscr{D}^k$. For example, node 4 in Figure 4 is duplicated in 4.a and 4.b in Figure 5. In this case, the NN is encouraged to learn also from *co-morbidities*.

4.3 Method Settings

The dataset shown in Figure 5 is used to train the NN with a 80-20 train-test split. Then, we average the performances on 10 experiments. Tables 3 and 4 show the system parameters for our best experiment when using DIAMOnD PPI and categories. We discuss the parameter sensitivity in Section 5.

Random Walk Parameter	Value
Label Embedding Length	300
Walk Length	20
Number of random walks per node	300
p	1
q	1
Skip-Gram context window	3
Skip-Gram Epochs	10

Table 3: Best RW^2 parameters.

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

Table 4: Best NN parameters.

5 EXPERIMENTS

5.1 Performances Evaluation

Given the previously outlined characteristics of biomedical data, evaluation measures such as *precision, accuracy* and *f-score* are ineffective, since there is no assessed experimental method to create negative examples. In line with other works [8, 22] in this domain, we use Recall@k, the fraction of correctly predicted items at rank k. Notice that, since reliable knowledge on negative interactions is not available, measures such as precision and AUC cannot be used. In all our experiments, we set the k value of Recall@k to 1 because the intended use of network methods in medicine is to exploit the results with the highest confidence, and to narrow the scope of expensive and labour-intensive clinical tests. We compare our system 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) DIAMOnD, which is commonly considered the state of the art and most cited study on disease gene prediction (see Section 2). DIAMOnD exploits only connectivity information.
(3) RWR[4] (Random Walks with Restart) [11] that, like for DIAMOnD, uses only connectivity information. RWR is commonly used as a comparison in recent literature on DGP.

Notice that we do not compare with Phenorank since it is a data-dependent algorithm. In order to rank the genes in the network, it needs to compute similarities between diseases and mouse mutants genes, exploiting their common phenotypes. In this context, Phenorank works only with specific datasets of mouse mutants and phenotypes, making it hard to re-use these data on a new network of proteins or genes.

For all the above-listed 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. Next, we compute the Recall@1 and average over all folds.

Notice that computing Recall@1 for DIAMOnD is not straightforward. In the evaluation experiments of DIAMOnD, diseases are considered one at the time. The node-disease associations are removed from a given fraction of the nodes N'_d known to be related with disease 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 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 5.

Table 5 shows variable performance depending mainly on the combination of PPI and disease categorization adopted: not surprisingly, all systems perform better on the DIAMOnD PPI when using DIAMOnD category labels, since this classification is more fine-grained and has been manually curated by medical experts specifically for this PPI (in fact, as shown in Table II when applied to HI-III, only 10 out of 70 defined disease categories could be mapped onto the PPI). We further observe that:

(1) Contrary to DIAMOnD and RWR, RW^2 exploits both node attributes and connectivity features, which systematically results in better performances; however, when fewer, or more coarse disease categories are used (as in columns 2-4), RW^2 reduces its ability to retrieve context-dependent differences in the neighbourhood of a disease-node, and its advantage over the other connectivity-based methods is reduced (or even lost, as in column 4);
(2) Using only similarity of feature vectors $f(v)$ (the Baseline method) does not allow to learn regularities, which is motivated by the high incompleteness and sparsity of available features. In other terms, co-morbidity alone is an extremely weak predictor of disease-genes;
(3) As also demonstrated in [1], RWR does not perform worse than DIAMOnD; on the contrary, it seems to work better especially in the experiment of column 1;
(4) In general, the performance of all systems are much lower than claimed in the respective papers: as already discussed, these methods use negative sampling (except for DIAMOnD) that appears to boost performances artificially.

Concerning DIAMOnD, we remark that in [8] the reported Recall is higher, but limited to two diseases, lysosomal storage diseases and lipid metabolism disorders, that show the higher density of the respective modules. For completeness, Table VI compares RW^2 and DIAMOnD on these very same diseases. Furthermore, the authors of [22], 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 5.

5.2 Robusteness and Sensitivity analysis

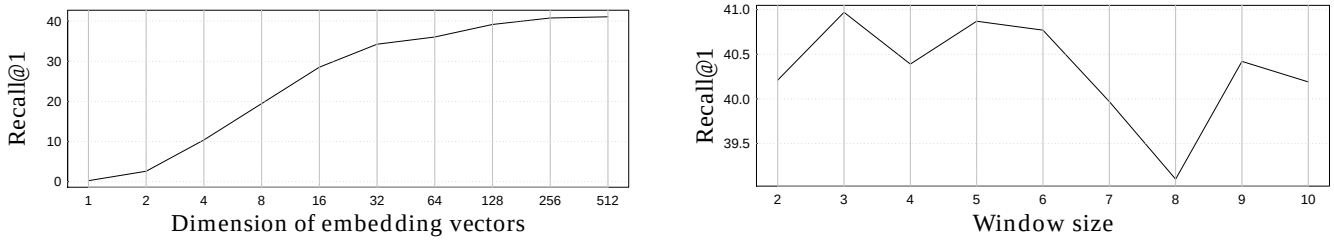
We perform a robustness test of RW^2 by randomly rewiring edges[5], by relabeling nodes features, and by exploiting both. The result is shown in Table 7. The table shows that while a severely relabeling affects the system's performance, rewiring only results in a ~3% points decrease in performance. Although this difference is statistically significant ($p < 0.02$), it clearly shows that connectivity is a weak feature. This is in line with Table 5, that shows extremely

[4]We used the following implementation https://github.com/TuftsBCB/Walker

[5]We use the method in http://bit.ly/2N0sKHf

Methods	Datasets used for PPI and disease categories			
	DIAMOnD (DIAMOnD)	HI-III (DIAMOnD)	HI-III (Disgenet)	HI-III (Phenorank)
RW^2	40.97% (0.90%)	33.29% (1.12%)	7.56% (0.71%)	4.87% (0.73%)
Baseline	0.26% (0.24%)	0.01% (0.36%)	0.47% (0.43%)	0.03% (0.57%)
DIAMOnD	14.05% (1.32%)	4.99% (1.46%)	3.38% (1.34%)	6.06% (2.06%)
RWR	22.29% (1.34%)	5.76% (1.71%)	3.80% (0.96%)	5.03% (2.79%)

Table 5: Macro Recall@1 and standard deviation (in parenthesis) over 10 folds.



(a) Performance as a function of the dimension of embedding vectors.

(b) Performance as a function of the dimension of Skip-gram context window.

Figure 6: Sensitivity analysis (DIAMOnD PPI and DIAMOnD categories)

Disease Module	RW^2	DIAMOnD
lysosomal storage diseases	85%	53%
lipid metabolism disorders	33%	31%

Table 6: Comparison between DIAMOnD and RW^2 for two diseases (R@1)

poor performances for methods based only on connectivity patterns (DIAMOnD and RWR).

Then, 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. Although more experiments with different and more complex learners might be needed, our intuition is that data quality - namely, incompleteness and sparsity of features - is too low for deep methods to learn regularities.

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(a) shows that a sufficiently high number of dimensions are needed (> 100) Figure 6(b) shows that the 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 [1], stating that the median distance between components in a module is almost 2.9, and [15] where the diameter of a disease module is estimated to be 1.8 in the average. To conclude the robustness and sensitivity analysis section, we tested several depths (1,2,3) of the classification network. We found out that, contrary to the typical expectations, increasing the number of hidden layers - up to 3 - decrease to 24%, the performance of RW^2 on Diamond dataset.

6 HIGHLIGHTS AND CONCLUSIONS

The main advantage of RW^2, is the ability to discover *specific combinations of connectivity and functional features that have a higher probability of being found in the vicinity of a node related to a given disease*. Although RW^2 surpasses other compared systems in most experimental settings, the performance measured in our experiments appears to be highly dependent on the adopted PPI and the specificity of considered disease categories. A larger number of fine-grained disease categories, as shown in Table V, favours the characterization of the disease-genes neighbourhood.

We also noticed that, in the majority of cases, the performance of compared systems is quite low in comparison with the values reported in the literature (either for specific diseases or specific networks), showing that connectivity features alone do not allow to discover disease modules in general.

This result is in agreement with a very recent study [1] where 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 in [1] 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

Type	DIAMOnD (DIAMOnD)	HI-III (DIAMOnD)	HI-III (Disgenet)	HI-III (Phenorank)
Rewiring	38.79% (1.87%)	30.26% (2.42%)	7.27% (1.39%)	4.31% (1.42%)
Relabel	0.24% (0.15%)	0.60% (0.66%)	1.24% (0.55%)	0.62% (0.55%)
Both	0.30% (0.11%)	0.22% (0.34%)	0.71% (0.28%)	0.85% (0.32%)

Table 7: Robustness test: rewiring and relabeling the network

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. As shown in Figure 6(b), we also found that performance degrades when the window length exceeds ±5, which implies that "some" vicinity among nodes related to the same disease does exist.

ACKNOWLEDGMENTS

Dr. Lorenzo Madeddu is attending the PhD program in Innovative Biomedical Technologies in Clinical Medicine of the Sapienza University.

This research has been supported by the MIUR under the grant "Dipartimenti di eccellenza 2018-2022" of the Department of Computer Science of Sapienza University and by the "Sapienza information-based Technology InnovaTion Center for Health - STITCH".

REFERENCES

[1] Monica Agrawal, Marinka Zitnik, and Jure Leskovec. 2018. Large-scale Analysis of Disease Pathways in the Human Interactome. In *Pacific Symposium on Biocomputing*, Vol. 23. 111–122.

[2] Aijun An, Bill Andreopoulos, Michael Schroeder, and Xiaogang Wang. 2009. A roadmap of clustering algorithms: finding a match for a biomedical application. *Briefings in Bioinformatics* 10, 3 (02 2009), 297–314.

[3] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. 2011. Network Medicine: A Network-based Approach to Human Disease. *Nature reviews. Genetics* 12 (01 2011), 56–68. https://doi.org/10.1038/nrg2918

[4] Arnaud Ceol, Andrew Chatr Aryamontri, Luana Licata, Daniele Peluso, Leonardo Briganti, Livia Perfetto, Luisa Castagnoli, and Gianni Cesareni. 2009. MINT: The molecular interaction database: 2009 update. *Nuc. acids res.* 38 (11 2009). https://doi.org/10.1093/nar/gkp983

[5] Stephen Y Chan and Joseph Loscalzo. 2012. The emerging paradigm of network medicine in the study of human disease. *Circulation research* 111 3 (2012), 359–74.

[6] Feixiong Cheng, Rishi J Desai, Diane E. Handy, Ruisheng Wang, Sebastian Schneeweiss, Albert-László Barabási, and Joseph Loscalzo. 2018. Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat Commun* 9, 1 (2018 07 12 2018), 2691. https://doi.org/10.1038/s41467-018-05116-5

[7] Alex J Cornish, Alessia David, and Michael J E Sternberg. 2018. PhenoRank: reducing study bias in gene prioritization through simulation. *Bioinformatics (Oxford, England)* 34-12 (12 2018), 2087–2095.

[8] Susan Dina Ghiassian, Jörg Menche, and Albert-László Barabási. 2015. A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome. *PLoS Computational Biology* 11, 4 (2015).

[9] Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research* 33, suppl_1 (2005), D514–D517.

[10] Luck Katja Katja and et al. 2019. A reference map of the human protein interactome. *bioRxiv* (2019). https://doi.org/10.1101/605451 arXiv:https://www.biorxiv.org/content/early/2019/04/10/605451.full.pdf

[11] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter Robinson. 2008. Walking the Interactome for Prioritization of Candidate Disease Genes. *American journal of human genetics* 82 (05 2008), 949–58. https://doi.org/10.1016/j.ajhg.2008.02.013

[12] Tran L, Hamp T, and Rost B. 2018. ProfPPIdb: Pairs of physical protein-protein interactions predicted for entire proteomes. *PLoS One* 13 (07 2018).

[13] Yongjin Li and Jagdish Chandra Patra. 2010. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics (Oxford, England)* 26 (03 2010), 1219–24. https://doi.org/10.1093/bioinformatics/btq108

[14] Joseph Loscalzo, Albert-László Barabási, and Edwin K. Silverman. 2017. *Network Medicine: Complex Systems in Human Disease and Therapeutics* (1 ed.). 1, Vol. 1. Harvard University Press.

[15] Jörg Menche, Amitabh Sharma, Maksim Kitsak, Susan Dina Ghiassian, Marc Vidal, Joseph Loscalzo, and Albert-Laszlo Barabasi. 2015. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science (New York, N.Y.)* 347 (02 2015). https://doi.org/10.1126/science.1257601

[16] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013). http://dblp.uni-trier.de/db/journals/corr/corr1301.html#abs-1301-3781

[17] Sandra Orchard and et al. 2013. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nuc. acids res.* 42 (11 2013). https://doi.org/10.1093/nar/gkt1115

[18] M. Oti, B. Snel, M. A. Huynen, and H. G. Brunner. 2006. Predicting disease genes using protein–protein interactions. *J. of Medical Genetics* 43, 8 (2006), 691–698. https://doi.org/10.1136/jmg.2006.041376

[19] Yungki Park and Edward Marcotte. 2011. Revisiting the negative example sampling problem for predicting protein-protein interactions. *Bioinformatics (Oxford, England)* 27 (09 2011), 3024–8. https://doi.org/10.1093/bioinformatics/btr514

[20] J. Piñero, À. Bravo, N. Queralt-Rosinach, and et al. 2017. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nuc. Acids Res.* 45, Database-Issue (2017).

[21] Erin M Ramos, Douglas Hoffman, Heather A Junkins, Donna Maglott, Lon Phan, Stephen T Sherry, Mike Feolo, and Lucia A Hindorff. 2014. Phenotype–Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eu. J. of Human Genetics* 22, 1 (2014), 144.

[22] Yael Silberberg, Martin Kupiec, and Roded Sharan. 2017. GLADIATOR: a global approach for elucidating disease modules. *Genome medicine* 9, 1 (2017), 48.

[23] Chris Stark, Bobby-Joe Breitkreutz, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. 2006. BioGRID: A general repository for interaction datasets. *Nucleic acids research* 34 (01 2006), D535–9. https://doi.org/10.1093/nar/gkj109

[24] Oron Vanunu, Oded Magger, Eytan Ruppin, Tomer Shlomi, and Roded Sharan. 2010. Associating Genes and Protein Complexes with Disease via Network Propagation. *PLoS comp. bio.* 6 (01 2010). https://doi.org/10.1371/journal.pcbi.1000641

[25] Kavitha Venkatesan, Jean-Francois Rual, Alexei Vazquez, Ulrich Stelzl, Irma Lemmens, Tomoko Hirozane-Kishikawa, Tong Hao, Martina Zenkner, Xiaofeng Xin, Kwang-Il Goh, Muhammed A Yildirim, Nicolas Simonis, Kathrin Heinzmann, Fana Gebreab, Julie M Sahalie, Sebiha Cevik, Christophe Simon, Anne-Sophie de Smet, Elizabeth Dann, and Marc Vidal. 2009. An empirical framework for binary interactome mapping. *Nature methods* 6 (01 2009), 83–90. https://doi.org/10.1038/nmeth.1280

[26] Sebastian Vlaic, Theresia Conrad, Christian Tokarski-Schnelle, Mika Gustafsson, Uta Dahmen, Reinhard Guthke, and Stefan Schuster. 2018. ModuleDiscoverer: Identification of regulatory modules in protein-protein interaction networks. *Scientific reports* 8, 1 (2018), 433.

[27] Xuebing Wu, Rui Jiang, and M.Q. Zhang. 2008. Network-based global inference of human disease genes. *Molecular Systems Biology* 4 (01 2008), 189–1. https://doi.org/10.1142/9789812790088_0018

[28] Zikai Wu, Yong Wang, and Luonan Chen. 2013. Network-based drug repositioning. *Mol. BioSystems* 9, 6 (2013).

[29] Donghyeon Yu, Minsoo Kim, Guanghua Xiao, and Tae Hyun Hwang. 2013. Review of Biological Network Data and Its Applications. *Genomics & informatics* 11 (12 2013), 200–210. https://doi.org/10.5808/GI.2013.11.4.200