A Feature-Learning based method for the disease-gene prediction problem

Lorenzo Madeddu

Translational and Precision Medicine Department,
Sapienza University of Rome,
Rome, Italy
E-mail: lorenzo.madeddu@uniroma1.it

Giovanni Stilo

Computer Science Department,
University of L'Aquila,
L'Aquila, Italy
E-mail: giovanni.stilo@univaq.it

Paola Velardi

Computer Science Department,
Sapienza University of Rome,
Rome, Italy
E-mail: 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 (RW^2), which is shown to perform better than other state-of-the-art algorithms. We also show that the performance of RW^2 and other compared state-of-the-art algorithms is extremely sensitive to the interactome used, and to the adopted disease categorizations, since this influences the ability to capture regularities in presence of sparsity and incompleteness.

Keywords: network medicine; disease gene prediction; disease gene prioritization; node embedding; random walks; graph-based methods; biological networks; complex diseases.

Biographical notes:
- *Lorenzo Madeddu* is a PhD student at the Department of Translational and Precision Medicine at Sapienza University of Rome with a Computer Science Master Degree. His research interests focus on machine learning, graph mining and Network Medicine. He is involved in interdisciplinary projects in the fields of Healthcare and Precision Medicine. He received his master degree in Computer Science from the Sapienza University of Rome in 2018.

- *Giovanni Stilo* is an Assistant Professor in the Department of Information Engineering, Computer Science and Mathematics at the University of L'Aquila.

He received his PhD in Computer Science in 2013, and in 2014 he was a visiting researcher at Yahoo! Labs in Barcelona. Between 2015 and 2018, he was a researcher at the Computer Science Department of La Sapienza University, in Rome. His research interests are in the areas of machine learning and data mining, and specifically temporal mining, social network analysis, network medicine, semantics-aware recommender systems, and anomaly detection. He has organized several international workshops, held in conjunction with top-tier conferences (ICDM, CIKM, and ECIR), and he is involved as editor and reviewer of top-tier journals, such as TITS, TKDE, DMKD, AI, KAIS, and AIIM.

- *Paola Velardi* is a Full Professor of Computer Science at Sapienza University in Rome, Italy. Her main interests are in the areas of natural language processing, social networks, machine learning, ontology learning and the semantic web. Main recent application areas are in the domain of recommender systems, e-health, e-learning and social networks. She published more than 160 papers on top-rated international journals, books and conferences. She is Associate Editor of KAIS and ACM-TKDD. She has been unit PI and PI of many national, regional and international projects. She received several awards and recognitions. Learn more on her Wikipedia page: `https://en.wikipedia.org/wiki/Paola_Velardi`.

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 (Yu et al. [2013]), to address important challenges in both diagnostic and therapeutic aspects, such as drug development and disease classification (Cheng et al. [2018], Wu et al. [2013], Menche et al. [2015]). Barabási et al. [2011]:"Given the functional interdependencies between the molecular components in a human cell, a disease is rarely a consequence of an abnormality in a single gene, but reflects the disruptions of the complex intracellular network". This complexity is hard to interpret using the traditional reductionist approach, according to which a single disease cause can be identified decoupling the complex biological or medical phenomenons in multiple components, consequently providing a cure. Instead, keeping in mind the complexity means to analyse the interaction between multiple components, which work dynamically in a system to pursue one or more purposes.

Network Medicine (Barabási et al. [2011]) (NM) is a relatively recent approach to analyse 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 reductionist (according to reductionism, a single disease cause can be identified by decoupling the complex biological or medical phenomenons in multiple components) approaches cannot detect. A central finding of NM (Barabási et al. [2011]) 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 studies have been published in support of this hypothesis, such as reported by Vlaic et al. [2018], Ghiassian et al. [2015], Silberberg et al. [2017] 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 (https://www.genome.gov/27541954/dna-sequencing-costs-data/) 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 sub-structures 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. In particular, the main contributions of this work are:

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 further show that, given high incompleteness of the interactome network, a careful aggregation of diseases into categories might considerably help predictive methods.

2 Related work and Background Knowledge

2.1 Survey of Protein-Protein Interaction (PPI) prediction methods

Recent research fields such as System Biology and Network Medicine (NM) (see Barabási et al. [2011]) 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*. The driving idea of NM is that the study of network topology and dynamics can accelerate the discovery of new biological interactions and pathways as noted by Chan and Loscalzo [2012], which in turn will drive progress on disease treatments and personalised medicine. The broader aim of the NM is to study and expands knowledge in the medical and biological domain, using computational methods to leverage the information embedded into the biological networks or databases. A typical challenge, in the NM domain, is to expand and organise biological information of the interactomes, such as the *Protein-Protein Interaction (PPI)* network (an undirected network, where nodes are proteins, and edges express physical interactions among them). This task is particularly crucial because this kind of networks are currently suffering from many issues:

- Incompleteness: it is estimated that only 20-30% of existing interactions have been discovered as noted by Venkatesan et al. [2009].

- Reliability: although many protein-protein interaction datasets are available in the literature (see Section 3), relationships are not fully reliable, unless experimentally tested on multiple assays. Only few recent experimental efforts are based on this idea, like HI-III-19 (see Katja and et al. [2019]).

- Negative Knowledge: when a relationship between two biological entities is not present, there is no assurance if it actually does not exist or if it is still unknown. Several empirical methods have been proposed in literature to simulate negative knowledge. Unfortunately, as discussed by L et al. [2018], training and testing with these negative datasets may significantly overestimate performances. Despite this, negative sampling is commonly used in many studies (e.g., PhenoRank Cornish et al. [2018]).

The characteristics of interactome networks highlighted above, cause, many commonly used, computational methods or graph-mining methods to be ineffective. For example, An et al. [2009] and Ghiassian et al. [2015] shows that community detection algorithms and centrality measures fail to identify relevant structures, because of incompleteness. For the reasons mentioned above, the interest of scientists to solve NM challenges remains high, and the growing interdisciplinary effort gives hope for an acceleration of results in this field.

From a higher perspective, a challenge of the NM domain consists of solving a predictive task where the goal is to discover the correlation of a biological entity to a specific class or its association with other biological entities. In this work, we are concerned with a specific predictive task - the *Disease Gene Prediction (DGP)* - where the involvement of a particular protein/gene with one or more diseases must be suggested.

DGP is a relevant, but still open, research topic, since, as pointed before, the genetic bases of diseases are largely unknown. Currently, only 10% of genes have a known association with some disease (Barabási et al. [2011]). Genome-Wide associations studies (GWAS) have led to the collections of such associations in databases, like OMIM (Hamosh et al. [2005]) and DisGenNet (Piñero et al. [2017]). However, as mentioned in the introduction, GWAS studies are very expensive and labor intensive. Several PPI-based computational approaches have been presented in the literature to solve the DGP problem. Barabási et al. [2011] have grouped them in the following three categories:

- **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" (`https://www.medicinenet.com/script/main/art.asp?articlekey=4166`). Linkage methods are based on the idea that genes associated with a given disease or disease category are often in a given linkage interval (the chromosomal location falls within one or more "disease loci"), and this information can be used to restrict the number of candidate genes for a given disease. Supposing that $\mathbb{D}$ is the set of all the diseases, and $d \in \mathbb{D}$ is a disease, then let be $V^d \subseteq V$ the set of nodes (genes) in the linkage interval of d. From the Linkage methods perspective, V^d is the set of *candidates* among which the genes related to disease d must be predicted. Further, let be $\hat{D} \subseteq \mathbb{D}$, a set of diseases which are functionally *similar* to d (i.e. $\hat{D}$ can be seen as a disease category), then $V_{\hat{D}}$ are the set of genes which are already known to be related with $\hat{D}$. One possible approach, as proposed by Oti et al. [2006], is to predict the disease-gene association among the candidates V^d and those in the direct neighbourhood of $V_{\hat{D}}$. The main problem with linkage methods is that identifying the causal genes at disease loci is often difficult, as noted in Cornish et al. [2018] and furthermore, the co-occurrence of genes in the same chromosomal location is a probable, but not a necessary, condition.

- **Diffusion Methods:** the majority of these methods, like those proposed by Vanunu et al. [2010], Köhler et al. [2008], Wu et al. [2008], Li and Chandra Patra [2010], are still based on linkage intervals to reduce the number of candidate genes for a given disease, but rely on more complex connectivity approaches to filter candidates. For example, in order to find novel disease-gene candidates, Köhler et al. [2008] introduces random walk with restart (RwR), starting from genes known to be associated with a given disease category. RwR and phenotypic information are also used in a recently published method, PhenoRank (Cornish et al. [2018]). PhenoRank does not rely on linkage intervals to reduce candidates, but rather it exploits the phenotypic similarity of an input disease (query disease) with other diseases. The similarity values are used to score the nodes of a PPI network, and these scores are then propagated across the network, so that genes that interact with many high scoring genes are also scored highly. To avoid bias induced by the fact that less studied genes are less connected within the PPI network, simulated sets of phenotype terms are used to calculate, for each gene, the probability of observing a gene score at least or great as the one actually observed.

- **Module-based Methods:** the approaches proposed by Ghiassian et al. [2015] and Silberberg et al. [2017] are solely based on network connectivity properties. The hypothesis is that candidate genes belonging to the same neighbourhood (or module) are more likely to be involved in the same diseases. Note that the notions of "neighbourhood" and "module" are vague here, and standard community detection

algorithms do not work. To have an idea of how a disease module looks like see figure 3. Both approaches start with a given disease d (or a disease category), consider the set of genes known to be associated with d - the initial "disease module" - and expand the module by exploiting the structure of the network. The main idea of Ghiassian et al. [2015] (DIAMOnD) 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 (Silberberg et al. [2017]) 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 (as noted by Cornish et al. [2018]), potentially influencing the performance of this method.

In this paper, we present *Random Watcher-Walker (RW^2)*, a graph-based method where the association between gene and disease is discovered using an artificial neural network predictor that exploit a new rich representation of the disease genes (or gene products). RW^2 does not fall into the categorisation shown above, but belongs to a fourth category, namely "Representation Learning", which exploits latent information and regular patterns to detect candidate genes. As best as we know there are no other methods crafted to solve the DGP problem, that fall in this category.

Moreover, 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 artificially boost performance;
- rather than using ad-hoc PPIs and categorisations, we analyse the influence of different PPIs and disease categorisations on systems' performance.

Table 1 summarises the methods surveyed in this Section, by providing an overview of the networks, features, and diseases databases used by the most relevant DGP methods, including the one presented in this paper.

2.2 Background Knowledge

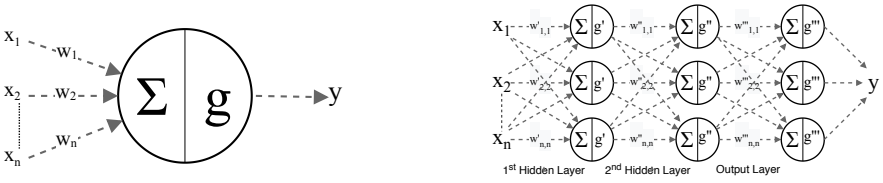
As stated before, NM is a cross-domains research field, where biology networks meet with computational methods. More specifically, our method falls in the Representation Learning category. In what follows, we present some background concepts for readers who are less familiar with computational methods based on Machine Learning.

- **Artificial Neural Networks (ANN)**, presented here in a simplified version, are Machine Learning methods vaguely inspired by the human brain. The model of an artificial

Work (Year, Name)	Type	Features	PPI Networks	Disease Databases
Oti et al. [2006]	Linkage	PPI	HPRD	OMIM
RwR Köhler et al. [2008]	Diffusion	PPI	BIND BioGRID DIP HPRD IntAct	OMIM
CIPHER Wu et al. [2008]	Diffusion	PPI	BIND HPRD MINT OPHID	OMIM
RWRH Li and Chandra Patra [2010]	Diffusion	PPI	HPRD	OMIM
PRINCE Vanunu et al. [2010]	Diffusion	PPI	HPRD	OMIM
DIAMOnD Ghiassian et al. [2015]	Module-based	PPI	Menche et al. [2015]	OMIM PheGenI
GLADIATOR Silberberg et al. [2017]	Module-based	PPI Phenotypes	Menche et al. [2015]	Menche et al. [2015]
PhenoRank Cornish et al. [2018]	Diffusion	PPI Phenotypes	BioGRID HI-II-14 HPRD IntAct	ClinVar OMIM UniProtKB
RW^2 (our)	Representation Learning	PPI	HI-III-19 Menche et al. [2015]	DIAMOnD DisGeNET PhenoRank
Menche et al. [2015]	Dataset	—	BIGG BIND BioGRID HPRD IntAct KEGG MINT OPHID PhosphositePlus TRANSFAC	Mottaz et al. [2008] OMIM PheGenI

Table 1 Summary of disease prediction approaches and adopted datasets. Since many works use, and refer to, the datasets described in Menche et al. [2015], to avoid repetitions we list in the lower part of the table Menche's et al. datasets.

neuron - introduced by Rosenblatt [1958] as the building block of an ANN - is a mathematical representation of a biological neuron, where the weighted sum of the inputs $\boldsymbol{x}$, is passed as an argument to non-linear function g, namely activation function (AF), to produce the output value y, as described by equation $y = g\Big(\sum_{i \in \boldsymbol{x}} (w_i \cdot x_i)\Big)$ and figure 1(a).



(a) (b)

Figure 1 Schematisation of artificial neuron fig. (a), and multi-layer ANN fig. (b)

Moreover, several neurons can be aggregated together (using a layered topology as it is shown in figure 1(b)) to compose a more complex ANN, called multi-layer or deep artificial neural network. In a deep artificial neural network (DANN), the outputs produced by the neurons of the previous layer are used as inputs of the next layer. In the supervised learning scenario, the input data $\boldsymbol{x}$ (i.e. an instance), the corresponding ground-truth value $\hat{y}$, and an ANN are given. The learning process consists of deciding the weights of the ANN that minimise, for every couple of input $\boldsymbol{x}$ and ground-truth $\hat{y}$, the difference $|\hat{y} - y|$ (namely error or loss function) between the produced output y and the ground-truth $\hat{y}$. One of the most employed techniques to learn the weights and minimise the loss function is the back-propagation algorithm presented by Rumelhart et al. [1986]. To conclude, even though there are several activation functions, all of them are ideally non-linear and differentiable. The nonlinearity is important to permit to learn the non-linear relationships hidden in the data. Differentiability is required to utilise the back-propagation techniques employed to learn the weights. The most popular activation function are ReLu, Tanh, Softmax and Sigmoid. Nwankpa et al. [2018] provide a more comprehensive overview of the activation functions used in deep learning applications.

- The **Skip-Gram model** is part of the Word2Vec models proposed by Mikolov et al. [2013] in the field of the Natural Language Processing. It is designed by leveraging the hypothesis that words in similar contexts (i.e. words close in the sentence) tend to have the same or close meaning. The goal of Skip-Gram is to predict the context words $\boldsymbol{y}$ of a given target word x. The core part of the Skip-Gram algorithm is made by an ANN trained using as input x a word and as the ground truth $\boldsymbol{y}$ its context words. More specifically the Skip-Gramm ANN is composed by one linear hidden layer and one softmax-based output layer. The network depicted in figure 2 is trained using the Cross-Entropy Loss function.

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. Figure 3 shows the network model G(V,E) where: nodes $v \in V$ are proteins or gene products, and edges

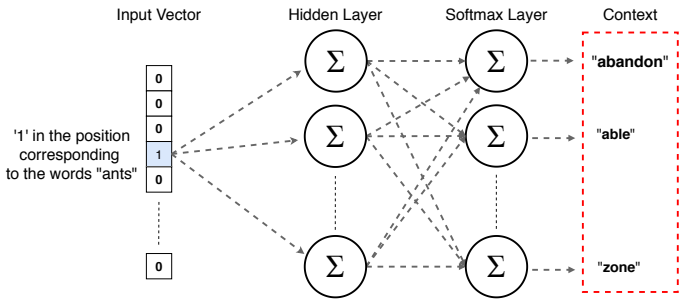


Figure 2 ANN architecture of the Skip-Gram model.

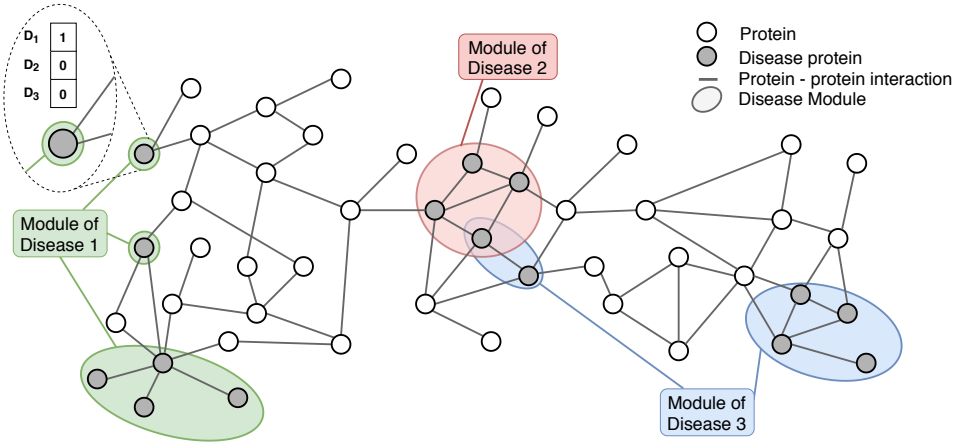


Figure 3 The network model considered in our work. Nodes are associated to one or more disease modules, and this information is reported in a feature vector, as shown in the upper left magnification.

$e(i,j) \in E, i,j \in V$ are interactions. Coloured clusters are disease modules, that is, set of disease-related genes (Barabási et al. [2011]). Each node (a protein or gene product), can be associated with one or more disease module, as reported in its feature vector. Note, as better shown in Section 4.2, that disease modules are not necessarily dense communities but well-localised neighbourhoods that can overlap.

In our approach, each node $v \in G$ is further described by a feature vector $f(v)$,

a one-hot vector where a "1", in position k, indicates that a specific disease d_k is associated to a node v. Note that we consider mono and multi-factorial diseases (those influenced by more than one gene). Furthermore, a gene might be associated to more than one disease. Since our method requires that every node has at least one value in the feature vector, we introduce an "empty" label, called wildcard W (see Figure 4), for all those nodes that are not associated with any disease.

The methodology to predict disease-related genes can be summarised in three steps:

- **Step 1: Random Watcher Walker**: we collect *network connectivity patterns* using a novel method, namely Random Watcher-Walker (RW^2) (exposed in Figure 4). The

RW^2 walker, when landing on node v, "watches" the node features and selects one disease label at random with uniform probability in those cells of the features vector $f(v)$ that are equals to 1. Next, it "walks" with uniform probability to one of v's neighbours. In this way, the visit made by the walker embodies both *functional* features of traversed nodes (disease labels), and *structural* features (connected proteins in the PPI). 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:

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

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

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

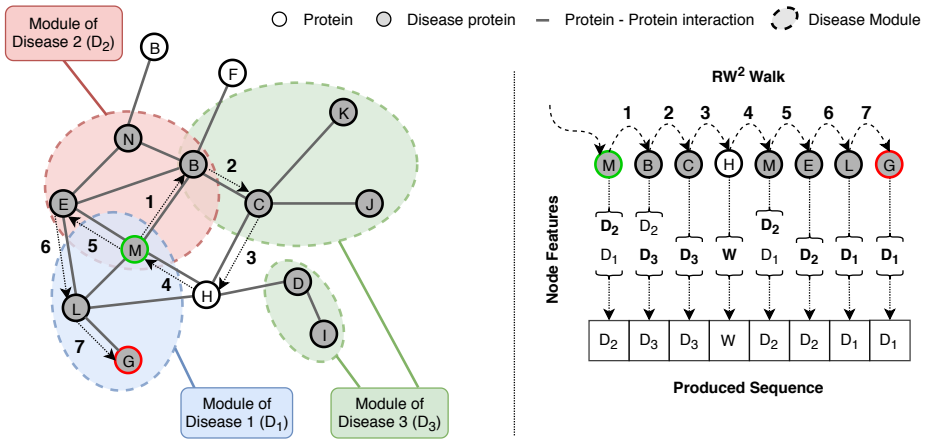


Figure 4 An example of the RW^2 step 1 applied on the network depicted in the left part of the figure. The random watcher-walker start its visit from the node M (highlighted in green) and after traversing node B,C,H,E,L, end the visit in the node G (highlighted in red).

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* (Barabási et al. [2011], Oti et al. [2006], Goh et al. [2007]). 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 5) much in the same way as sentences are contexts for individual words. Contexts are used to train a Skip-Gram (Mikolov et al. [2013]) model and learn *label embeddings* (embeddings are "dense" vector representations of feature labels, a very 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 $\mathscr{F}(v)$.

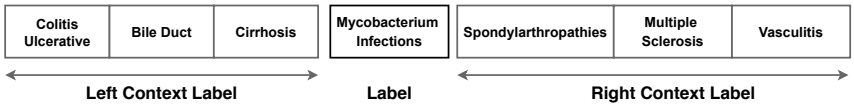


Figure 5 Example of "context" for the disease category *mycobacterium infections*. In each step t of the walk, a node v^t is randomly selected among those connected with the previous node v^{t-1}, and next, a label is randomly extracted from $f(v^t)$. The figure shows a fragment of the produced specific (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 $t+1, t+2\ldots$.

.

- **Step 3: Training**: feature matrices $\mathscr{F}(v)$ are used to train a fully connected artificial neural network (ANN) with Softmax activation function (depicted in Figure 6), for predicting disease-gene associations . 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 (unknown), to state absence of known disease relations for a given node.

Most existing approaches require that all nodes in the graph are present during the training of the system; these approaches are inherently transductive and do not naturally generalise to unseen nodes. Instead, our method is intrinsically inductive leverages node feature information to efficiently classify previously unseen data.

4 Evaluation Methodology

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 Data Sources

PPI networks: Protein-protein interactions are mostly derived from databases curated from the literature (hypothesis-driven), like those in *IntAct* (Orchard and et al. [2013]), *BioGrid* (Stark et al. [2006]), *MINT* (Ceol et al. [2009]). These datasets may be affected by inspection bias (also termed study bias or investigational bias as in Loscalzo et al. [2017]) 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

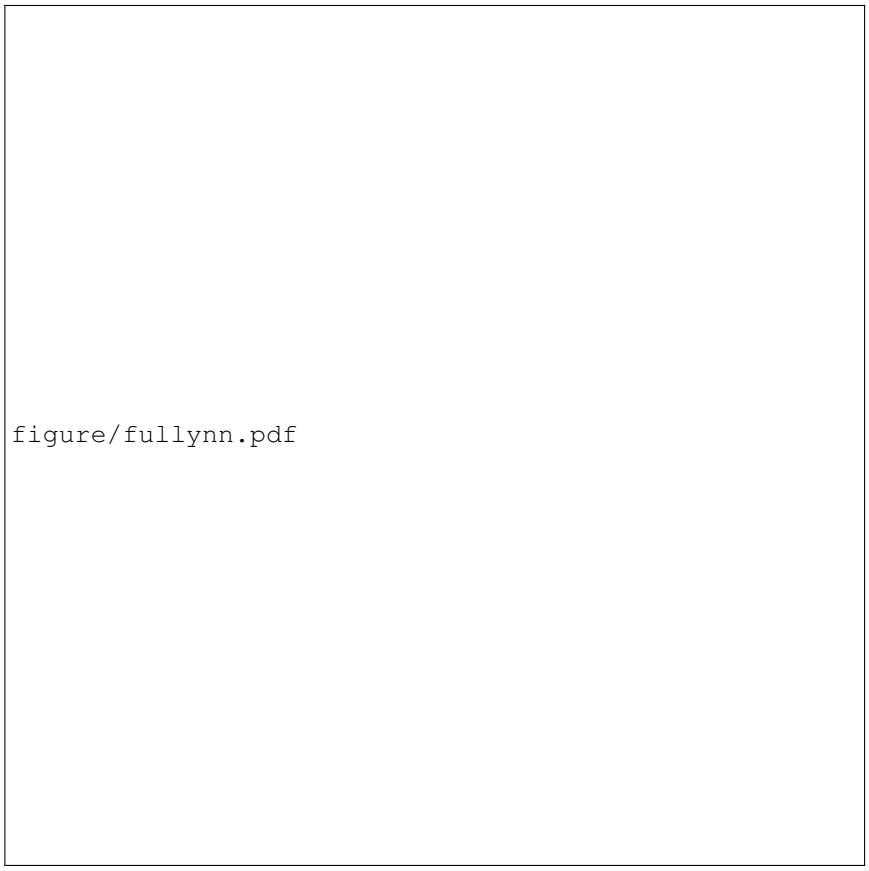


Figure 6 Training the ANN with feature matrices

a higher 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 the purpose of comparison, we use the same interactome (PPI) network used in DIAMOnD by Ghiassian et al. [2015], obtained by integrating several data sources as described by Menche et al. [2015]. This network is one of the most complete since includes most of PPI sources used by others DGP methods (see previous Table 1), as: HPRD (Keshava Prasad et al. [2009]), BioGRID (Stark et al. [2006]), IntAct (Orchard and et al. [2013]), MINT (Ceol et al. [2009]).

- **HI-III-19**: 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 inspect bias on this type of dataset is negligible (Loscalzo et al. [2017]). HI-III-19 is publicly available on the HuRI website (`http://interactome.baderlab.org/about/`).

Table 2 shows some network statistics. We note that DIAMOnD is slightly more connected, and larger, than HI-III-19. Another important difference is that nodes in HI-III-19 are isoform proteins, while in DIAMOnD they are genes. More importantly, relationships in HI-III-19 are considered highly reliable.

Disease categories: Disease categories with a genetic basis were obtained from DIAMOnD Ghiassian et al. [2015] (disease-gene associations from OMIM by Hamosh et al. [2005] and by PheGeni Ramos et al. [2014]), Phenorank (Cornish et al. [2018]) or Disgenet (Piñero et al. [2017])(selected disease-gene associations from OMIM (Hamosh et al. [2005]), Uniprot (Consortium [2017]), ClinVar (Landrum et al. [2016]), etc.). Table 3 shows the effect of applying these three categorisation types to the DIAMOnD and HI-III-19 PPIs. Clearly, since the DIAMOnD categorisation has been manually conceived for the DIAMOnD PPI, all categories (70) map to some of the nodes of 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-19, 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-19 and DIAMOnD. Finally, only 16 Phenorank categories could be associated to HI-III-19 and 12 to DIAMOnD.

PPI network	N. Nodes	N. Edges	Graph Density	Connected Components	Avg. Number of Neighbours
HI-III-19	8490	54495	0.0015	71	13
DIAMOnD	13458	141272	0.0016	89	20

Table 2 Network statistics

PPI network		DIAMOnD Category Labels (CL)	Disgenet CL	Phenorank CL
HI-III-19	N. of Different Diseases	10	31	16
	Diseases Nodes (%)	4%	14%	10%
	N. of Disease-Node associations	479	1828	1008
DIAMOnD	N. of Different Diseases	70	31	12
	Diseases Nodes (%)	11%	8%	5%
	N. of Disease-Node associations	2843	1717	850

Table 3 Effect of using different disease categorisations on different PPIs

4.2 Topological Analysis of the Disease Module

We conducted an in-depth topological analysis of the adopted networks, to obtain insight on the structure of disease modules. As modules, we used the induced subgraph of the DIAMOnD and the DisGeNET disease categories, in the DIAMOnD and the HI-III-19 network. For every disease module induced subgraph we measured, following Agrawal et al. [2018]: the number connected components, the number of proteins included in the largest connected component, density, the conductance against the remaining graph, the distance among disease nodes. According with Menche et al. [2015] we also measured the *modular separation* of the disease modules.

We found that disease modules are fragmented over the PPI network. The median of the connected components of each module is 21 (Avg. 27.2) in DIAMOnD and 34 (Avg.

55) in HI-III-19. The median of proteins included in the largest connected component is only 15% (Avg. 19%) in DIAMOnD and 2% (Avg. 2%) in HI-III-19. We also found that disease modules are not densely connected, with a median density of 0.04 in DIAMOnD (Avg. 0.05, the overall network density is 0.0016) and 0.005 in HI-III-19 (Avg. 0.006, the whole network density is 0.0015).

Note that the majority (90%) of the disease modules has a density below 0.09 in DIAMOnD and 0.01 in HI-III-19. Furthermore, the modules are somewhat well connected externally, having a median conductance of 0.96 (Avg. 94%) in DIAMOnD and 0.98 (Avg. 98%) in HI-III-19. This kind of conductance values highlights that there are more edges pointing outside the module than edges lying inside. Finally, the median distance between proteins in the same disease module is 3.36 (Avg. 3.31) in DIAMOnD and 3.92 (Avg. 3.94) in HI-III-19, and the modular separation of diseases is 56% in DIAMOnD and 49% in HI-III-19.

These analysis show that disease modules, following a definition based on the notion of the induced subgraph, do not express a topological structure (which is well connected internally and has few edges pointing outside the cluster) which is instead typical of communities, as defined in Network Science.

The absence of a topological structure in conjunction with the expressed modular separation of the PPI, suggests that methods able to exploit the *local* structure of the modules and to capture their overlaps, as is the case of RW^2, are more promising than those based solely on a broader topological structure of the modules.

4.3 Data Transformation

Figure 7(a) 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. Coloured cells in $\mathscr{F}(v)$ represent embedding vectors associated to 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 i^{th} cell equal to 1 if the node is known to be associated to the corresponding disease $d \in D$. The last cell of this vector is 1 if no diseases are known to be associated with the node.

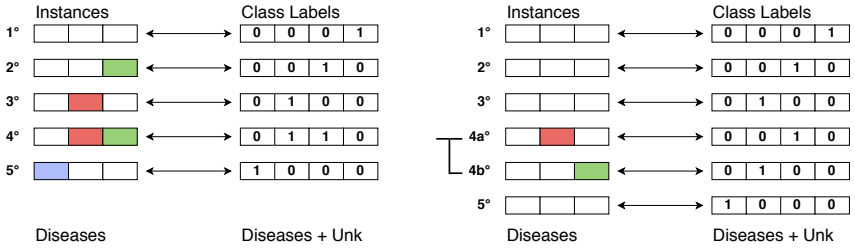


Figure 7 Feature matrix with ground-truth vectors fig. (a), and their modified version used for the learning phase fig. (b)

The dataset in Figure 7(a) cannot be used for training, because the ANN 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 ANN using modified feature matrices, as shown in Figure 7(b).

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 7(a) are modified as in Figure 7(b). Note that in this way a node with no valued cells in the *disease* dimension (instances 1 and 2 of Figure 7(b)), 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 to distinguish 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 to 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. For example, node 4 in Figure 7(a) is duplicated in 4.a and 4.b in Figure 7(b). In this case the ANN is encouraged to learn also from *co-morbidities*.

4.4 Experimental Setup

The dataset transformed as in Figure 7(b) is used to train the ANN with 80-20% train-test split and then averaging on 10 experiments. Tables 4 and 5 show the system parameters for our best experiment, when using DIAMOnD PPI and categories. Sensitivity to parameters is discussed 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 4 Best RW^2 parameters.

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

Table 5 Best ANN 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, since there is no assessed experimental method to create negative examples. In line with other works (Ghiassian et al. [2015], Silberberg et al. [2017]) in this domain, we use Recall@k, the fraction of correctly predicted items at rank k. Note 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, 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 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 art and most cited study on disease gene prediction (see Section 2). DIAMOnD exploits only connectivity information.

3. RWR (We used the following implementation `https://github.com/TuftsBCB/Walker`) (Random Walks with Restart) (Köhler et al. [2008]) that, like for DIAMOnD, uses only connectivity information. RwR is commonly used as a comparison in the recent literature on DGP.

Note 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.

Note that computing Recall@1 for DIAMOnD is not straightforward. In DIAMOnD evaluation experiments, described in Ghiassian et al. [2015] , 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 6.

Methods	Datasets used for PPI and disease categories			
	DIAMOnD (DIAMOnD)	HI-III-19 (DIAMOnD)	HI-III-19 (Disgenet)	HI-III-19 (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 6 Macro Recall@1 and standard deviation over 10 folds.

Table 6 shows variable performance depending mainly on the combination of PPI and disease categorisation 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-19, 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;

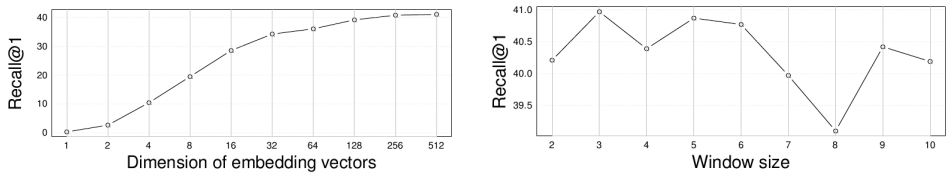
3. As also demonstrated in Agrawal et al. [2018] RwR does not perform worst than DIAMOnD, on the contrary, it seems to work better especially in the experiment of column 1;

4. In general, performances 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 artificially boost performances.

Concerning DIAMOnD, we remark that in Ghiassian et al. [2015] 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 the purpose of completeness, Table 7 compares RW^2 and DIAMOnD on these very same diseases. Furthermore, Silberberg et al. [2017], 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 6.

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

Table 7 Comparison between DIAMOnD and RW^2 for two diseases (R@1)



(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 8 Sensitivity analysis (DIAMOnD PPI and DIAMOnD categories)

5.2 Sensitivity analysis

Finally, we analyse 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 8(a) shows that a sufficiently high number of dimensions is needed (> 100) Figure 8(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 Agrawal et al. [2018], stating that the median distance between components in a module is almost 2.9, and Menche et al. [2015] where the diameter of a disease module is estimated to be 1.8 in the average.

6 Discussion and Concluding Remarks

The main advantage of RW^2 appears to be its 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 performances measured in our experiments appear 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, favors the characterisation of the disease-genes neighbourhood.

We also noted that, in the majority of cases, the performances of compared systems are 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 is also demonstrated by the results of Section 4.2, dedicated to the analysis of disease modules.

This result is in agreement with a very recent study Agrawal et al. [2018] 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, Agrawal et al.

[2018] 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 optimise 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. As shown in Figure 8(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 grant “Dipartimenti di eccellenza 2018-2022” of the Department of Computer Science of Sapienza University, by the “Sapienza information-based Technology InnovaTion Center for Health - STITCH” and by the “Territori Aperti project" funded by “Fondo Territori Lavoro eConoscenza CGIL, CSIL and UIL".

References

Monica Agrawal, Marinka Zitnik, and Jure Leskovec. Large-scale analysis of disease pathways in the human interactome. In *Pacific Symposium on Biocomputing*, volume 23, pages 111–122, 2018.

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

Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: A network-based approach to human disease. *Nature reviews. Genetics*, 12:56–68, 01 2011. doi: 10.1038/nrg2918.

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

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

Feixiong Cheng, Rishi J Desai, Diane E. Handy, Ruisheng Wang, Sebastian Schneeweiss, Albert-László Barabási, and Joseph Loscalzo. Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat Commun*, 9(1):2691, 2018 07 12 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-05116-5.

The UniProt Consortium. Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, 2017. doi: 10.1093/nar/gkw1099.

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

Susan Dina Ghiassian, Jörg Menche, and Albert-László 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*, 11(4), 2015.

Kwang-Il Goh, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690, 2007. ISSN 0027-8424.

Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(suppl_1):D514–D517, 2005.

Luck Katja Katja and et al. A reference map of the human protein interactome. *bioRxiv*, 2019. doi: 10.1101/605451.

TS Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, et al. Human protein reference database—2009 update. *Nucleic acids research*, 37(suppl_1):D767–D772, 2009.

Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter Robinson. Walking the interactome for prioritization of candidate disease genes. *American journal of human genetics*, 82:949–58, 05 2008. doi: 10.1016/j.ajhg.2008.02.013.

Tran L, Hamp T, and Rost B. Profppidb: Pairs of physical protein-protein interactions predicted for entire proteomes. *PLoS One*, 13, 07 2018.

Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, et al. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, 44(D1):D862–D868, 2016.

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

Joseph Loscalzo, Albert-László Barabási, and Edwin K. Silverman. *Network Medicine: Complex Systems in Human Disease and Therapeutics*, volume 1 of *1*. Harvard University Press, 1 edition, 2 2017.

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

Anaïs Mottaz, Yum L Yip, Patrick Ruch, and Anne-Lise Veuthey. Mapping proteins to disease terminologies: from uniprot to mesh. In *BMC bioinformatics*, volume 9, page S3. BioMed Central, 2008.

Chigozie Nwankpa, Winifred Ijomah, Anthony Gachagan, and Stephen Marshall. Activation functions: Comparison of trends in practice and research for deep learning. *CoRR*, abs/1811.03378, 2018.

Sandra Orchard and et al. The mintact project—intact as a common curation platform for 11 molecular interaction databases. *Nuc. acids res.*, 42, 11 2013. doi: 10.1093/nar/gkt1115.

M. Oti, B. Snel, M. A. Huynen, and H. G. Brunner. Predicting disease genes using protein–protein interactions. *J. of Medical Genetics*, 43(8):691–698, 2006. ISSN 0022-2593. doi: 10.1136/jmg.2006.041376.

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

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

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986. doi: 10.1038/323533a0.

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

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

Oron Vanunu, Oded Magger, Eytan Ruppin, Tomer Shlomi, and Roded Sharan. Associating genes and protein complexes with disease via network propagation. *PLoS comp. bio.*, 6, 01 2010. doi: 10.1371/journal.pcbi.1000641.

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. An empirical framework for binary interactome mapping. *Nature methods*, 6:83–90, 01 2009. doi: 10.1038/nmeth.1280.

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

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

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

Donghyeon Yu, Minsoo Kim, Guanghua Xiao, and Tae Hyun Hwang. Review of biological network data and its applications. *Genomics & informatics*, 11:200–210, 12 2013. doi: 10.5808/GI.2013.11.4.200.