**Thesis Paper:** TM Susana Nunes (Pdf on group’s github)

**Topic:**

**"Predicting Gene-Disease Associations with knowledge graph embeddings over multiple ontologies"**

->scientific research or computational approach

->relationships between genes and diseases using a combination of knowledge graph embeddings and multiple ontologies.

**knowledge graph**

->entities (such as genes and diseases) are represented as nodes, and the relationships between them are represented as edges.

**Ontologies**

->define the concepts and relationships within a specific domain.

An ontology is a technique or technology used to represent the knowledge about a domain, by modeling concepts and the relationships between them, being that these relationships describe the properties of those concepts.

Ontologies have **two major components:**

**(i)** a set of classes (concepts) that define the entities in a domain; and

**(ii)** a set of semantic links (relationships) between the classes that describe interactions between classes or properties of classes.

facilitating text mining, intelligent searching, and unambiguous identification.

**"knowledge graph embeddings"**

->represent entities and relationships in a knowledge graph as numerical vectors in a high-dimensional space.

->enabling computational algorithms to perform various tasks, including prediction.

I**n the specific contex**t of predicting gene-disease associations, multiple ontologies are used to provide a comprehensive representation of the biological knowledge related to genes and diseases.

**The overall goal of this approach is to leverage the power of knowledge graph embeddings and multiple ontologies to predict potential associations between genes and diseases. This predictive capability can be valuable in various biomedical research areas, such as identifying potential drug targets, understanding disease mechanisms, and aiding in the development of personalized medicine.**

**Abstract**

Our approach integrates different methods for building the shared semantic space, as well as multiple knowledge graph embedding algorithms and machine learning methods.

The prediction performance was evaluated on curated gene-disease associations from **DisGeNET** and compared to classical semantic similarity measures.

\*\*\*DisGeNET is a discovery platform containing one of the largest publicly available collections of genes and variants associated to human diseases\*\*\*

**Keywords**: ontologies, semantic similarity, knowledge graph, knowledge graph embedding, machine learning

**Objective:**

The main goal of this work is to investigate approaches to predict gene-disease associations that explore the semantic richness of knowledge graphs.The work tackles semantic richness from different perspectives: semantic similarity vs. knowledge graph embeddings; one ontology vs. two ontologies; disconnected ontologies vs. linked ontologies.

**semantic similarity vs. knowledge graph embeddings**

->Semantic similarity and knowledge graph embeddings are both techniques used in natural language processing and computational linguistics.

**Semantic Similarity:** Semantic similarity measures quantify the degree of similarity or relatedness between two pieces of text based on their semantic content or meaning.

**Knowledge Graph Embeddings**: Knowledge graph embeddings, on the other hand, are techniques used to represent entities and relationships in a knowledge graph as numerical vectors in a high-dimensional space.

**one ontology vs. two ontologies**

**One Ontology:**

Simplicity: Utilizing a single ontology can simplify the overall knowledge representation and management. It reduces the complexity of maintaining and aligning multiple ontologies.

**Two Ontologies:**

Domain Separation: When dealing with distinct domains or subject areas, it may be beneficial to have separate ontologies. This allows for focused representation and organization of knowledge specific to each domain.

**disconnected ontologies vs. linked ontologies**

**Disconnected Ontologies:** Disconnected ontologies refer to separate ontologies that are developed and maintained independently, without explicit links or connections between them. Each ontology represents a specific domain or subject area, and the knowledge within each ontology is self-contained.

**Linked Ontologies:**

Linked ontologies, also known as ontology networks or ontology alignments, involve establishing explicit links or mappings between concepts and relationships across multiple ontologies.

Our guiding hypothesis is that richer representations, powered by knowledge graph embeddings computed over multiple linked ontologies achieve a better predictive performance than simpler approaches based on semantic similarity or knowledge graph embeddings using a single ontology.

**The work aims to answer three research questions:**

1. **RQ1**: What are the advantages of knowledge graph embedding over semantic similarity measures as a representation strategy?

2. **RQ2**: How can different knowledge graph embedding approaches be computed over multiple biomedical ontologies to represent both genes and diseases as vectors?

3. **RQ3**: What is the impact of employing multiple ontologies and having logical links between them?

**Machine Learning:**

It is defined as a field in computer science that studies the use of computers to simulate human learning by exploring patterns in the data and applying self-improvement to continually enhance the performance of learning tasks.

The main learning task is to learn a mapping from this feature vector to an output prediction of some form (Nickel et al., 2016a). The algorithms can be divided into supervised and unsupervised learning algorithms.

**Supervised Learning:**

Supervised learning can be divided further into two categories of problem:

1. **Classification**: Common classification algorithms are linear classifiers, support vector machines, decision trees, k-nearest neighbor, and random forest, etc.
2. **Regression**: It is used for the prediction of continuous variables. Linear regression, logistical regression, and polynomial regression are popular regression algorithms.

In this work, we focused on **four different supervised learning algorithms**:

* Random Forest (RF) (Breiman, 2001),
* Gradient Boosting (Chen and Guestrin, 2016),
* Na¨ıve Bayes (NB) (Friedman et al., 1997), and
* Multi-Layer Perceptron (MLP) (Rumelhart et al., 1986).

**Gene-Disease Prediction:**

we categorize them into two types of approaches: those that do not employ ontologies, and those that do.

**1. Non Ontology-Based approaches**:

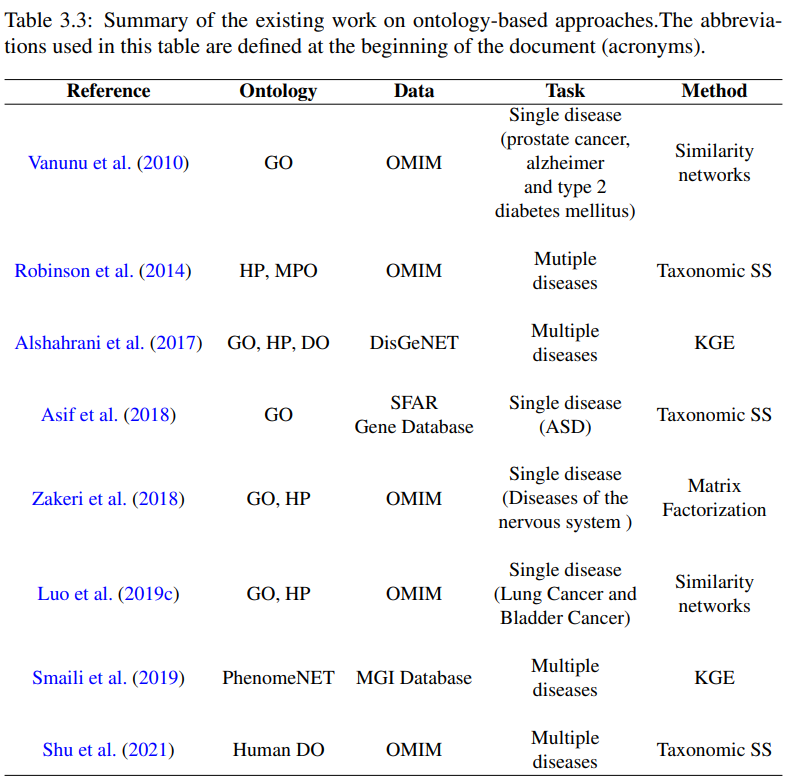
The majority of gene-disease prediction methods train a machine learning classifier with various types of features extracted from different kinds of data (Mordelet and Vert, 2011; Yang et al., 2012; Singh-Blom et al., 2013; Luo et al., 2019b,a).

Matrix completion methods address these issues by jointly predicting gene-disease associations and leveraging disease similarities during calculation (Zeng et al., 2017a)

. Zeng et al. (2017b) also used phenotype–phenotype similarity to prioritize novel gene–phenotype associations.

Centrality indices, random walk, and network energy are used in many methods to predict disease gene associations (Kohler et al. ¨ , 2008; Chen et al., 2014a,b).

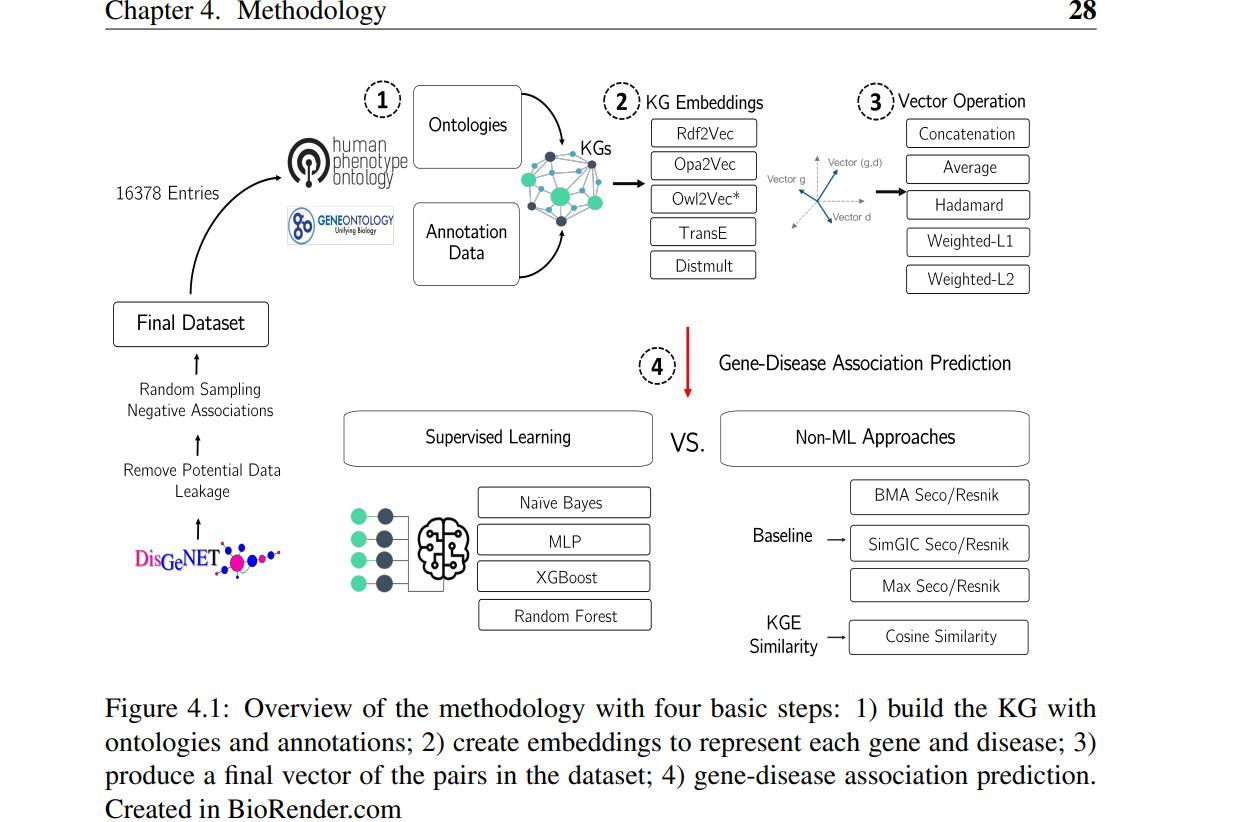
**2.Ontology-Based approaches:**

****

anunu et al. (2010) presented a novel network-based approach for predicting causal genes and protein complexes that are involved in a disease of interest (Prostate Cancer, Alzheimer or type 2 Diabetes Mellitus) and explored the Gene Ontology for manually annotated protein complexes

**Methodology:**

The methodology proposed in this work can be divided into four main steps as depicted in Figure 4.1. The first step in the approach is to integrate the different ontologies and annotation data to build the knowledge graph. In a second step, the embeddings that represent the gene and the disease according to their annotations in different knowledge graphs are created. In a third step, these embeddings are combined using different vector operators producing a representation of genes and diseases in what is effectively a shared semantic space. Finally, in a fourth step, supervised learning algorithms are trained over the combined embeddings to predict gene-disease associations. This approach is evaluated against non-machine learning approaches based on classical semantic similarity measures (baseline) and knowledge graph embedding similarity.



**To have accurate annotations for all the pairs, they were filtered and excluded if they did not correspond to one of these three criteria:**

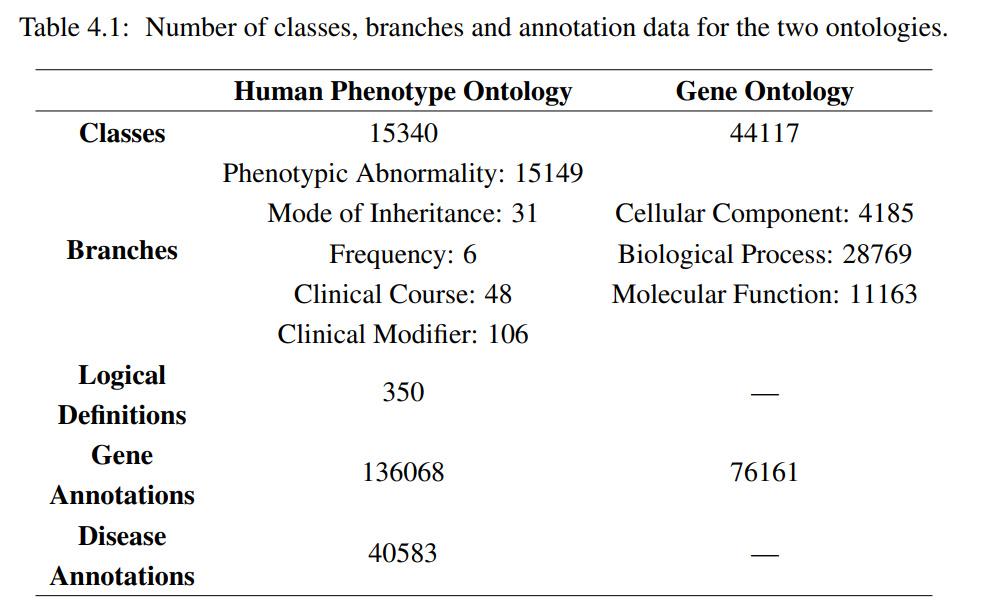
(i) the genes must have correspondence with a protein from **Uniprot** that is annotated with GO classes;

(ii) the genes must be annotated with HP classes;

(iii) the diseases must be annotated with HP classes.

**Ontologies and Knowledge Graphs:**

Two of the most popular biomedical ontologies were chosen: Human Phenotype Ontology and Gene Ontology.



1https://hpo.jax.org/app/

2http://geneontology.org/

**Knowledge Graph Integration;**

A first step was to merge the Human Phenotype Ontology and Gene Ontology through a common virtual root and save into an **OWL file** because both OWL2Vec and OPA2Vec methods only accepted a single ontology file with a single root. This was achieved by using the **RDFlib library**. RDFLib contains parsers for most of the known RDF serializations, including RDF/XML (OWL)

**Knowledge Graph Embeddings and Representation: (page50)**

three types of popular knowledge graph embedding approaches (AppendixA for default parameters):

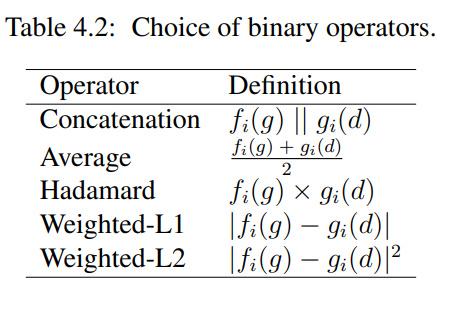
• Translational Distance: TransE4 (Bordes et al., 2013) with default parameters;

• Semantic Matching: DistMult4 (Yang et al., 2015) with default parameters;

• Path-based:

Random Walk: -RDF2Vec5 (Ristoski and Paulheim, 2016a) with sequences generated using the Weisfeiler-Lehman algorithm with walks depth 8 and a limited number of 500 by entity. The corpora of sequences were used to build a Skip-Gram model with the default parameters for Word2Vec; - OWL2Vec\*6 (Chen et al., 2021) with the same parameters used with RDF2Vec.

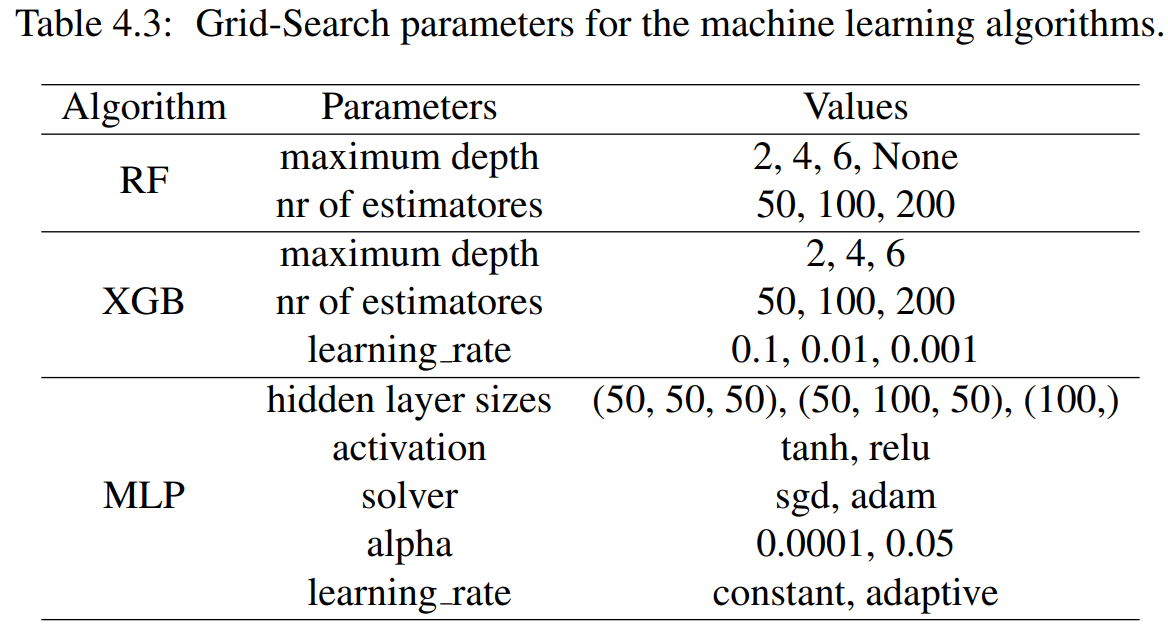
Non-Random Walk: - OPA2Vec7 (Smaili et al., 2019) with default parameters;



**Gene-Disease Prediction:** (pg-52)

To evaluate the approach, we tested the performance of supervised classifiers to predict gene-disease associations using the proposed feature vectors.

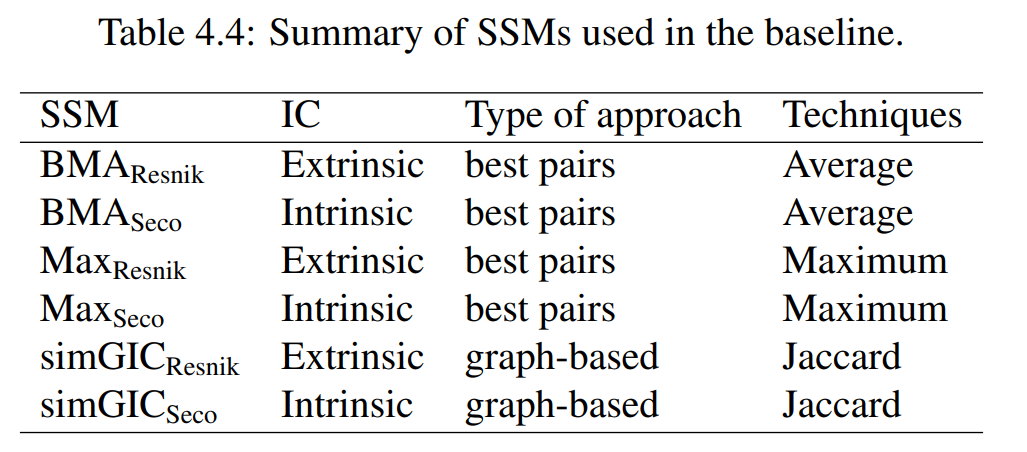
We used **four** different machine learning algorithms: **RF** (Breiman, 2001), **XGB** (Chen and Guestrin, 2016), **NB** (Friedman et al., 1997) and **MLP** (Rumelhart et al., 1986).



**Baseline and Experiments:**

The baseline aims to establish the performance of methods that use a single ontology and classical semantic similarity measures.

Semantic similarity computations were run using the tool SSMC8 which was designed to measure semantic similarity between a set of objects annotated by ontology classes.



the **six state-of-the-art** semantic similarity measures used: BMAResnik, BMASeco, MaxResnik, MaxSeco, simGICResnik, and simGICSeco. These six measures are representative of the most successful approaches for the baseline calculation using a single ontology.

**Results and Discussion:**

there are three important aspects that need to be considered when elucidating the performance impact:

1. How can we combine the gene and disease vector?

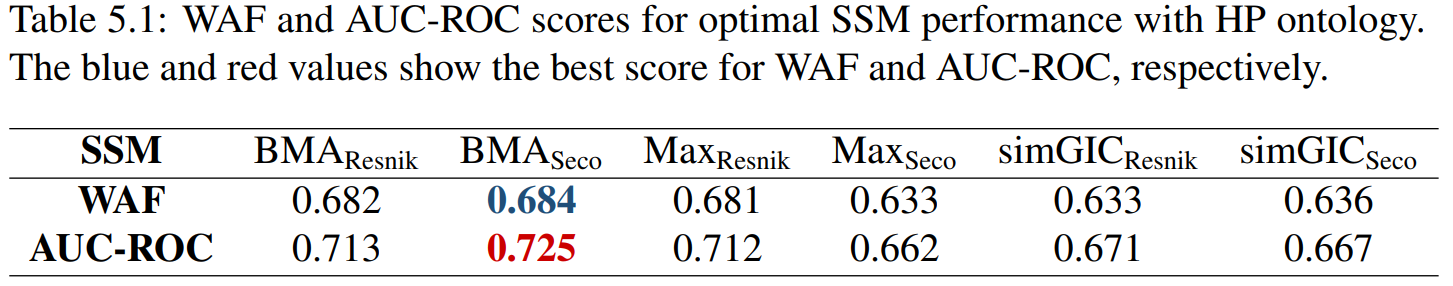
2. Which knowledge graph embedding methods are more suitable for this task?

3. What is the impact of considering more than one ontology?

First, the results for the gene-disease prediction baseline established using semantic similarity measures are described. Then, the results obtained using a rich semantic representation with different embeddings methods, knowledge graphs, and machine learning techniques are presented and compared to the best semantic similarity measure that the baseline achieved as well as the calculated cosine similarity for each knowledge graph embedding.

**Baseline Performance :**

Concerning the performance of the six semantic similarity measures in terms of information content for WAF score with a single ontology, we observe that the Max approachis more sensitive to the IC measure employed.



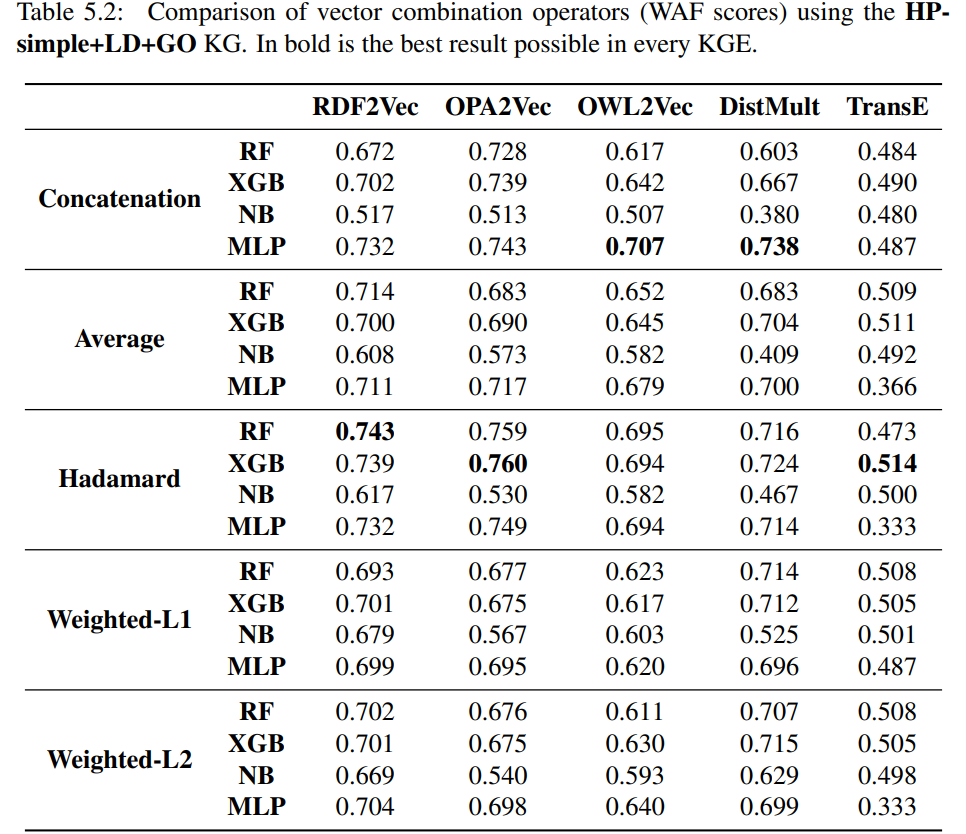
This may be explained by the fact that the Max approach only considers the best pair of classes when measuring similarity, with the IC measure playing a major role in the similarity score, whereas both BMA and simGIC consider all annotating classes.

Regarding the combination approach, the pairwise approach followed by BMA achieves the best results with a top WAF of 0.684 and AUC-ROC of 0.725 when combined with ICSeco. This highlights two interesting aspects.

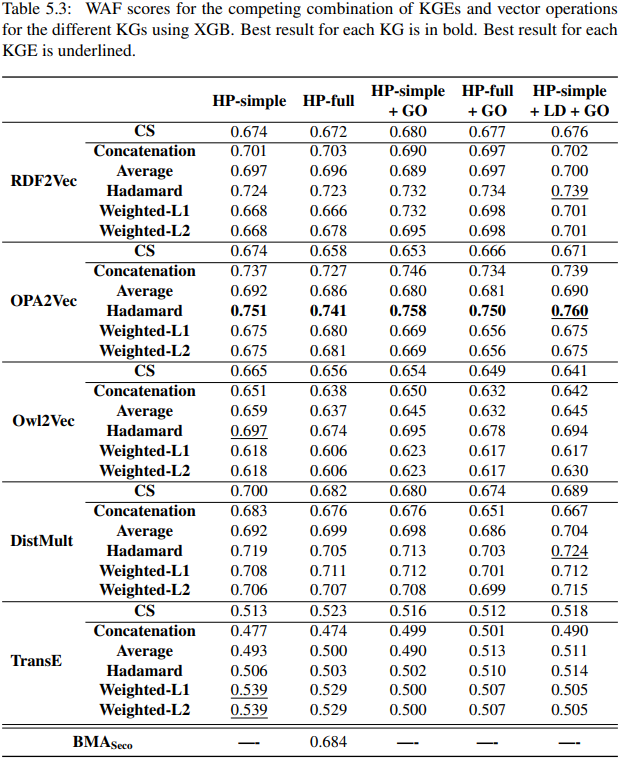
**5.2 Rich Semantic Representations Performance**

**5.2.1 Comparison of Vector Combination Approaches**

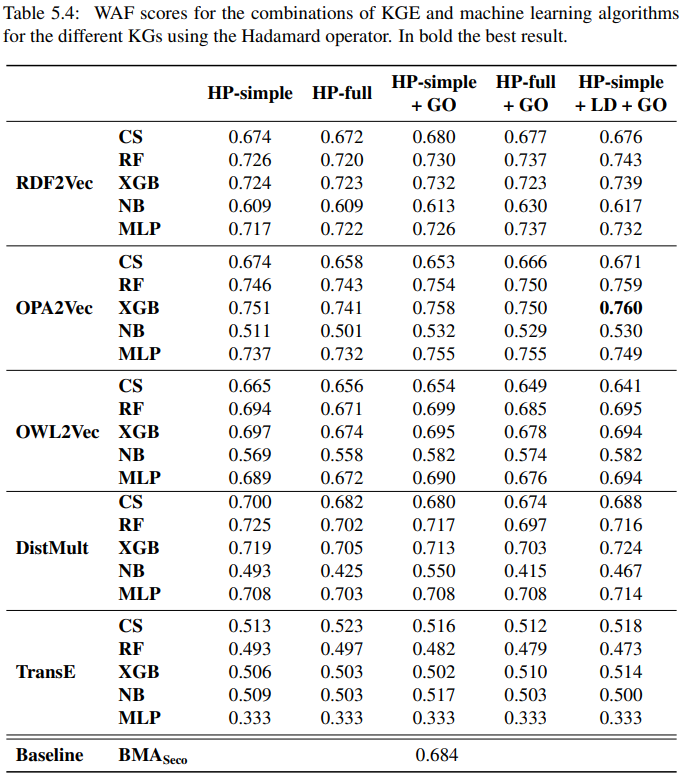
summarizes the comparison of the five chosen vector operations with AUC-ROC evaluated using the three best knowledge graph embedding methods (RDF2Vec, OPA2Vec, and DistMult) coupled with Random Forest classifier (one of the best-performing machine learning algorithms) using the richest knowledge graph (HP-simple + LD + GO).



**5.2.2 Comparison of Knowledge Graph Embedding Methods:**



**5.2.3 Comparison of different Knowledge Graphs**



**5.2.4 A case study on gene BACH2 and KPD disease:**

This example illustrates the limitations of semantic similarity-based approaches in handling complex diseases which are related to more than one gene, and genes related to more than one disease. On one hand, a single score view coupled with a simple prediction approach based on a similarity threshold can fail to identify the association between a gene and a disease with few closely related annotations when using a measure such as BMA that considers all annotations. On the other, the Max approach is less suitable when several annotations support a prediction. Moreover, machine learning models coupled with embeddings afford the formulation of a more complex solution.

**Conclusions:**

The impact of different approaches to build a shared rich semantic representation for genes and diseases was investigated, as well as multiple knowledge graph embedding methods such as RDF2Vec and OPA2Vec. An unbiased benchmark dataset was created to support evaluation, ensuring its appropriateness for gene-disease prediction.

The experiments provided answers to the research questions. Namely, they showed that knowledge graph embeddings when coupled with machine learning algorithms achieve a better performance than semantic similarity measures, answering RQ. They also illustrated that some vector combination approaches support gene disease association prediction better, but that a simple cosine similarity between vectors can support predictions as well as semantic similarity, answering RQ2. Finally, the experiments also revealed that differences between using a single ontology or combining two ontologies are comparatively small regardless of a richer integration using the logical definitions, answering RQ3.

However, there is a clear advantage for most knowledge graph embeddings methods to employ graph versions where logical definitions are encoded as direct links between classes. We hypothesize that the information provided by the Gene Ontology and links to it does not provide substantial additional information comparing with what is already present in the Human Phenotype Ontology.

**WEBSITES or IMPORTANT CLUES**

[**DisGeNET**](https://www.disgenet.org/)**: DisGeNET integrates data from expert curated repositories, GWAS catalogues, animal models and the scientific literature.**

[**BioPortal**](https://bioportal.bioontology.org/)**: the world's most comprehensive repository of biomedical ontologies.**

[**XGB**](https://xgboost.ai/)**:** is an open-source library that provides an efficient and effective implementation of the gradient boosting algorithm. Repository: [dmlc/xgboost: Scalable, Portable and Distributed Gradient Boosting (GBDT, GBRT or GBM) Library, for Python, R, Java, Scala, C++ and more. Runs on single machine, Hadoop, Spark, Dask, Flink and DataFlow (github.com)](https://github.com/dmlc/xgboost)

**Liu, W., Liu, J., and Rajapakse, J. (2018). Gene ontology enrichment improves performances of functional similarity of genes. Scientific Reports, 8. 13**

**OMIM knowledgebase**

**Uniprot**

**RDFlib library**. RDFLib contains parsers for most of the known RDF serializations, including RDF/XML (OWL)

**https://github.com/RDFLib/rdflib;** [**https://rdflib.readthedocs.io/en/stable**](https://rdflib.readthedocs.io/en/stable)

**https://github.com/thunlp/OpenKE 5https://github.com/IBCNServices/pyRDF2Vec https://github.com/KRR-Oxford/OWL2Vec-Star https://github.com/bio-ontology-research-group/opa2vec**

**https://github.com/liseda-lab/SSMC**

**Papers to read:**

1. Breiman, L. (2001). Machine learning, volume 45, number 1 - springerlink. Machine Learning, 45:5–32.
2. Chen, B., Shang, X., Li, M., Wang, J., and Wu, F.-X. (2016). Identifying individual cancer-related genes by rebalancing the training samples. IEEE Transactions on
3. NanoBioscience, 15:309–315. Chen, B., Wang, J., Li, M., and Wu, F.-x. (2014b). Identifying disease genes by integrating multiple data sources. BMC Medical Genomics, 7(Suppl 2):1–12.
4. Caceres, J. J. and Paccanaro, A. (2019). Disease gene prediction for molecularly uncharacterized diseases. PLOS Computational Biology, 15(7):1–14.
5. Kulmanov, M., Smaili, F. Z., Gao, X., and Hoehndorf, R. (2020a). Machine learning with biomedical ontologies. bioRxiv.
6. Lamy, J.-B. (2017). Owlready: Ontology-oriented programming in python with automatic classification and high level constructs for biomedical ontologies. Artificial Intelligence in Medicine, 80.
7. Luo, P., Xiao, Q., Wei, P.-J., Liao, B., and Wu, F.-X. (2019c). Identifying disease-gene associations with graph-regularized manifold learning. Frontiers in Genetics, 10.
8. Nickel, M., Murphy, K. P., Tresp, V., and Gabrilovich, E. (2016a). A review of relational machine learning for knowledge graphs. Proceedings of the IEEE, 104:11–33
9. Opap, K. and Mulder, N. (2017). Recent advances in predicting gene–disease associations. F1000Research, 6:578.
10. Sun, P. G., Gao, L., and Han, S. S. (2011). Prediction of human disease-related gene clusters by clustering analysis. International Journal of Biological Sciences, 7:61 – 73.

**Title: Disease gene prediction for molecularly uncharacterized diseases (Mahin)**

**Abstract:**

a novel network approach to prioritize gene-disease associations that is able to also predict genes for diseases with no known molecular basis. Our method, which we have called **Cardigan** (ChARting DIsease Gene AssociatioNs).

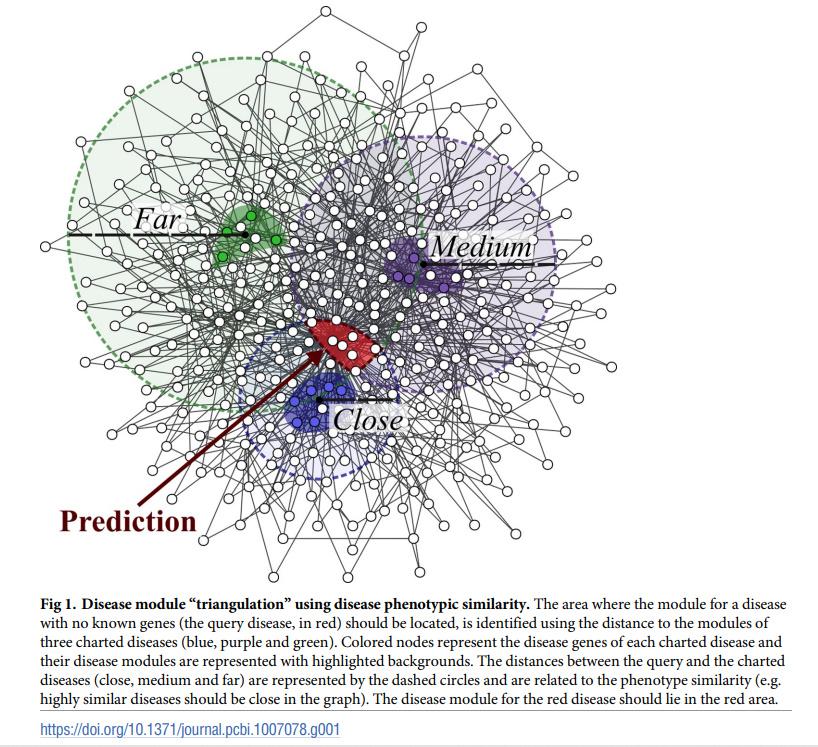
We evaluated its performance at predicting genes for both **molecularly characterized** and **uncharacterized diseases** in **OMIM**, using both weighted and binary interactomes, and compared it with state-of-the-art methods. Our tests, which use datasets collected at different points in time to replicate the dynamics of the disease gene discovery process, prove that **Cardigan** is able to **accurately predict** disease genes for molecularly uncharacterized diseases.

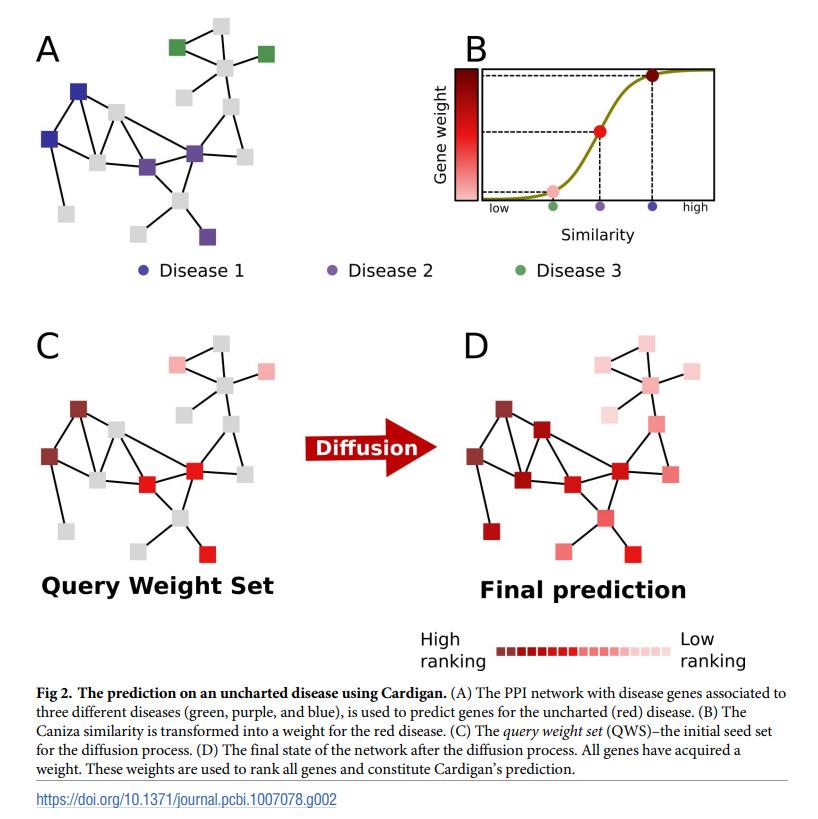
**Introduction:**

In this paper, we present a **disease gene prediction** method that predicts disease genes for both **charted and uncharted diseases** in **OMIM**, and can also predict disease modules. Our approach, which we have called **Cardigan** (ChARting DIsease Gene AssociatioNs), is based on a semi-supervised algorithm that propagates labels on the interactome. These labels integrate disease phenotypic information expressed as a similarity measure between diseases, which is obtained by **mining and comparing** **sets of MeSH terms** relevant for the diseases.

**The Cardigan algorithm:**

To predict disease genes for a given disease **(query disease)**, Cardigan begins by calculating its **phenotypic similarity** to **every other disease** in **OMIM.**

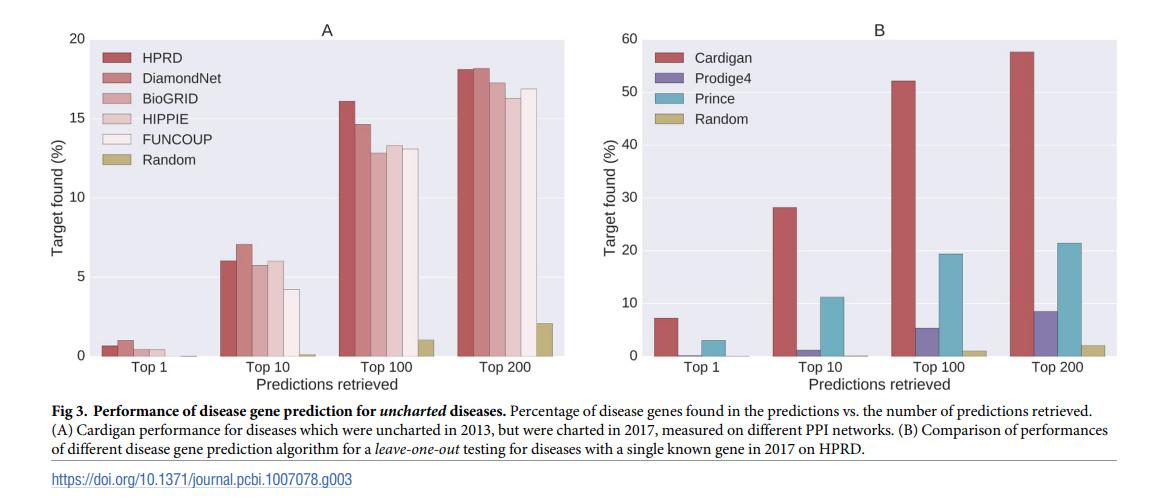




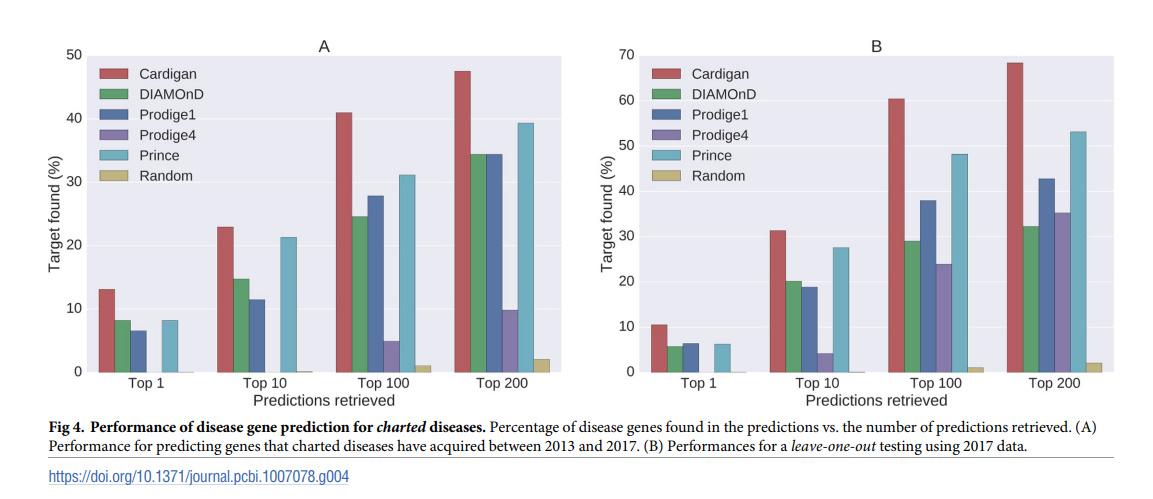
**Performance evaluation:**

We compared the performance of **Cardigan** against **PRINCE, ProDiGe1, ProDiGe4 and DIAMOnD** at predicting disease genes for OMIM diseases (these algorithms are described in the Methods section). PRINCE and Cardigan were run using both binary protein-protein interaction networks (HPRD [38], BioGRID [39], DiamondNet [34]) as well as weighted networks (HIPPIE [40] and FUNCOUP [41]), while ProDiGe1, ProDiGe4 and DIAMOnD can run only on binary networks (see Methods for details).

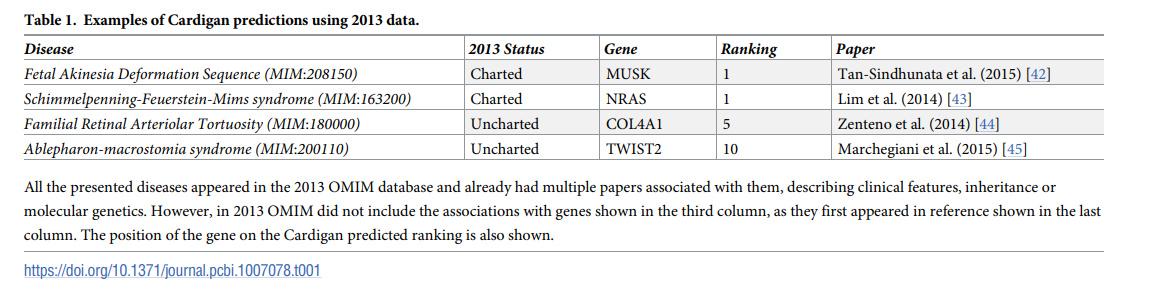
**Performance on uncharted diseases:**

****

**Performance on charted diseases:**



**Discussion:**



**Methods:**

**1.Disease data:**

Our experiments were carried out using disease data from the **OMIM database**.We also used **the Ghiassian et al. diseases** module dataset.

**2.Protein-protein interaction networks;**

we performed our tests using different types of protein interaction networks including **weighted and binary networks** with both experimental and predicted data: HPRD [48], DiamondNet [34] and BioGRID [39] are binary experimental networks; HIPPIE [49] is a weighted experimental network; **FUNCOUP** is a large weighted network including both experimental and predicted data.

**3.Other prediction methods;**

We compared Cardigan to four methods: ProDiGe1, ProDiGe4[20], PRINCE [19] and DIAMOnD [34]. These were chosen because they are state-of-the-art representatives of the disease gene prediction methods and of the disease module prediction methods described earlier.

**4.The Caniza similarity:**

Caniza et al. [36] recently proposed a measure to quantify the phenotypical similarity between hereditary diseases. Their method begins by collecting, for each disease, the set of MeSH terms assigned to the scientific publications relevant for that disease.

**Implementation;**

Our method is available as a fast, industrial strength library for **Python 2.7** which implements sparse matrices and lazy loading for disease similarities to reduce the memory footprint. The code is publicly available from the paper website at **http://www.paccanarolab.org/cardigan.**

**Supporting information:**

S1 Dataset. Ghiassian disease dataset to OMIM identifier mapping.

S2 Dataset. Cardigan prediction on the entire 2017 OMIM dataset.

**Title: Identifying disease genes by integrating multiple data sources (Shuvo)**

**Abstract:**

Results: In this paper, we propose a multiple data integration method based on the theory of Markov random field (MRF) and the method of Bayesian analysis for identifying human disease genes.

**Background:**

Many human genetic diseases or disorders are resulted from mutations of multiple genes [1]. The identification of those disease genes is not only important in understanding genetic disease mechanisms, but is also helpful in developing new methods in diagnostics and therapeutics.

Here the author mentioned some techniques that were proposed or developed by many scientists to predict gene-disease associations. E.g. Wu et al. [5] develop a tool called CIPHER to predict disease genes based on a global concordance between a PPI network and a phenotype network.

In this paper, we introduce a multiple data integration method for disease gene identifications, which considers comprehensive characters of a set of heterogeneous datasets to capture the complex relationship between genotypes and phenotypes. The method is based on the theory of MRF and the method of Bayesian analysis. Two previous algorithms of Deng et al. [30] and Ma et al. Their method cannot be directly employed to identify human disease genes. It is not clear how to integrate more kinds of biological data by using their method. In paper [32], we have developed a basic modified MRF model for human disease gene prioritization. In this study, we will further improve it by introducing a new parameter estimation strategy and a new Gibbs sampling strategy. The improved MRF algorithm is not only stable in terms of parameter estimation, but also reliable in terms of its prediction accuracy.

**Methods:**

In this paper, we first briefly describe how the problem is formulated as a Bayesian labelling problem. The labelling configuration assumes to follow a Gibbs distribution. After that, a MRF model is introduced to solve this problem by integrating multiple kinds of biological data, including known gene-disease associations, protein complexes, PPIs, pathways and gene expression profiles.

In this section the authors described The Bayesian labelling problem and Gibbs distribution in MRF.

**The Bayesian labelling problem:**

The best estimation of f is the one maxizing a posteriori probability (MAP), which is

P(f|r) = P(r|f )P(f )/P(r) (1)

where P(r) is the probability that we get the observation r. The Bayesian labelling problem [33] is that given a set of observation r, find the MAP configuration of labelling

f ∗ = arg maxf∈F P(f|r).

Here, as P (r) is not a function of f , it does not affect the MAP estimation of f.

**Gibbs distribution in MRF**:

It is usually hard to specify a prior probability of a MRF for a real problem. Fortunately, the Hammersley-Clifford theorem [34] provides a solution for this. According to the theorem, F is a MRF on S w.r.t. N if and only if the probability distribution of P (F = f ) of the configuration is a Gibbs distribution w.r.t. N.

The Gibbs distribution has a form of



Then the posterior probability of the Gibbs distribution has form



where the posterior energy is

U(f|r) = U(f ) + U(f|r).

Based on this, suppose the collection of whole human genes G = {g1, g2, ..., gN} is the site set, and {1, 0} is the label set, where 1 represents a gene is a disease gene and 0 otherwise. The problem of human disease gene identification is actually to find the best configuration of G according to what is currently known about human diseases.

**The MRF model for identifying human disease genes**

Suppose human genome consists of a set of N genes G = {g1, g2, ..., gN}. Some of them are already known to be associated with genetic diseases, while associations of most other genes are still not known.

**The Gibbs sampling**

The Gibbs distribution (14) gives a prior probability distribution of the configuration for all genes. In the study of identifying human disease genes, the objective is to find the posterior probability of X1, X2, · · ·, Xn conditional on known disease genes.

**Parameter estimation**

In practice, we do not know parameters of the model and they need to be estimated according to those known informations.The step-by-step description of this procedure is given in this section.

**Estimating a prior probability**

Now, the only problem left is to estimate the prior probability of πi. Similarly as the method used in Deng et al. [30], we also estimate them according to known protein complexes. Since genes that encode proteins in a same complex tend to associated with similar diseases.

**Data sources**:

The gene-disease association data are obtained from Goh et al. [3], which contain 1 284 disorders and 1 777 disease genes. These data are originally collected from the Morbid Map list of the Online Mendelian Inheritance in Man (OMIM) [38]. In this study, we consider only those disease classes that consist of at least 30 genes.

The protein complex data are collected from the database of CORUM [39] and PCDq [40]. There are 1677 and 1103 protein complexes in the dataset that consist of at least two proteins, respectively. There are in total 3881 proteins in those protein complexes.

The PPI datasets are derived from the database of HPRD (Release 9) [9], BioGrid (Release 3.2.108) [10] and IntAct (downloaded on Jan 26, 2014) [11], respectively. Duplicated edges between the same pair of vertices and edges connecting to itself are deleted.

Each dataset is processed independently, and three PPI networks are obtained finally. The HPRD PPI network consists of 9465 vertices and 37039 edges. The BioGrid PPI network consists of 15298 vertices and 127612 edges. The IntAct PPI network consists of 13449 vertices and 63825 edges.

The pathway datasets are obtained from the database of KEGG [12], Reactome [13], PharmGKB [14] and PIN [15], There are 280, 1469, 99 and 2679 pathways in datasets, respectively. There are in total 8614 proteins in those pathways.

The human gene expression profiles are obtained from BioGPS (GSE1133) [16,17], which contain 79 human tissues in duplicates, measured using the Affymetrix U133A array.

Hence, five biological networks are constructed by collecting data from various databases. All protein IDs are mapped onto the form of the gene symbol. In order to test the performance of multiple data integration of our methods, we select those genes that appears at least four times in the five networks. The final datasets consist of 7311 human genes, 815 out of which are known associated with 12 disease classes.

**Validation method and evaluation criteria:**

The accuracy of predictions is validated by the leaveoneout method. For each known disease gene with at least one annotated interaction partner in a biological network, we assume it is an unknown gene and predict its posterior probability by our proposed methods. We use the receiver operating characteristic (ROC) curve to show the relationship between the true positive rate and the false positive rate by varying the threshold for declaring positives. The area under the ROC curve (AUC) is also employed to show an overall measure of the performance. The negative control set consists of known disease genes that do not belong to current disease class, and they are also validated by using the leave-one-out method.

**Decision score and declaration of positives:**

One can directly use the posterior probabilities obtained by the Gibbs sampling to select candidate disease genes. The greater the probability is for a gene, the more likely it is to associated with specific disease. However, different disease classes consist of different numbers of known disease genes, and thus the prediction results may not be good if a global threshold is used for all classes. Hence, we propose to use a percentage as a decision score to generate the finial predictions. All the ROC curves and the AUC scores of our “IMRF1“ and “IMRF2“ method are calculated according to the decision score hereafter.

**Results and discussion:**

Firstly, since ideas of our improved methods (IMRF1 and IMRF2) are initially inspired by the MRF-Deng method, the direct comparison illustrates how much improvement can be made results from our methods.

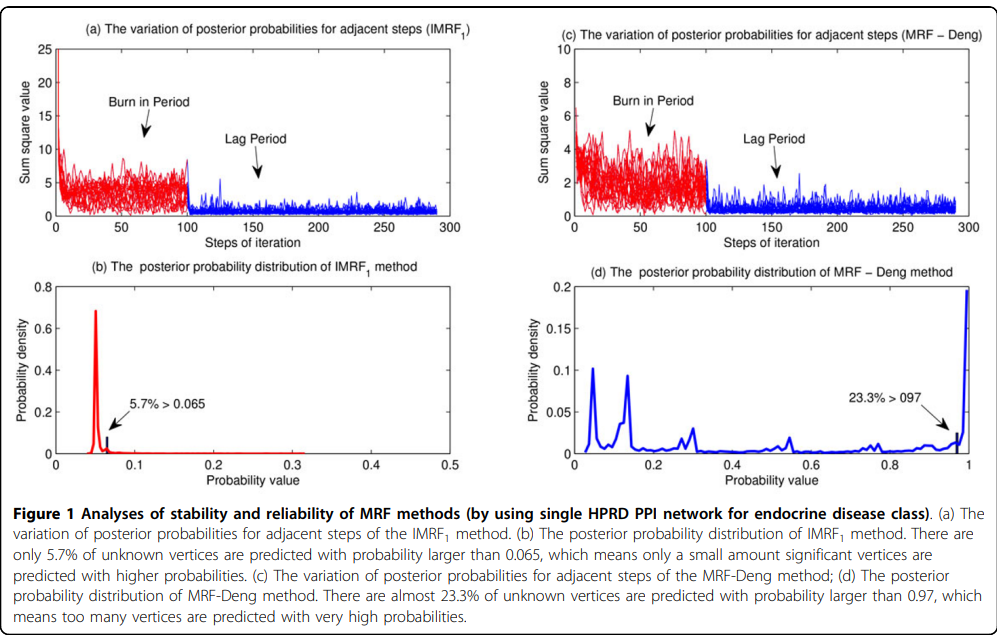
Secondly, we compare our methods with the RWR algorithm to show which manner of multiple data integration is better. The RWR algorithm is a typical data integration method that uses a mixed network, where vertices and edges of several biological networks are simply merged together, while our methods integrate different networks separately.

Finally, the DIR algorithm has a very good performance among multiple data integration methods, which also integrates different networks separately. It is the same with our methods in terms of the data integration method.

**Stability and reliability of MRF methods:**

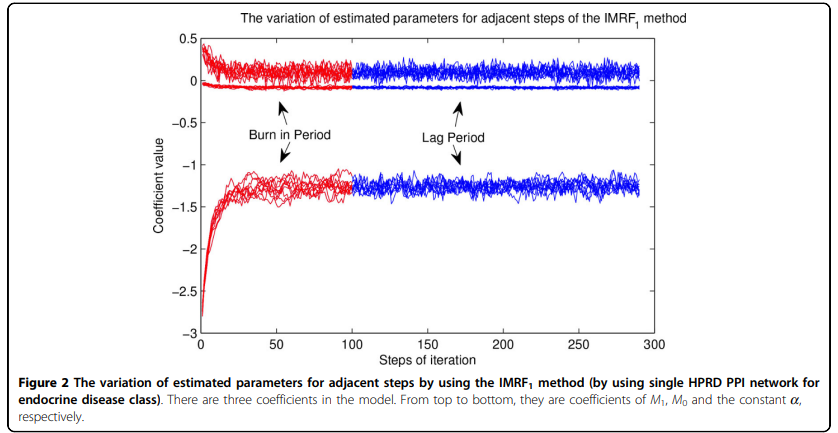
We first investigate the stability and reliability MRF methods, by analyzing Markov processes of the IMRF1 method and the MRF-Deng method.

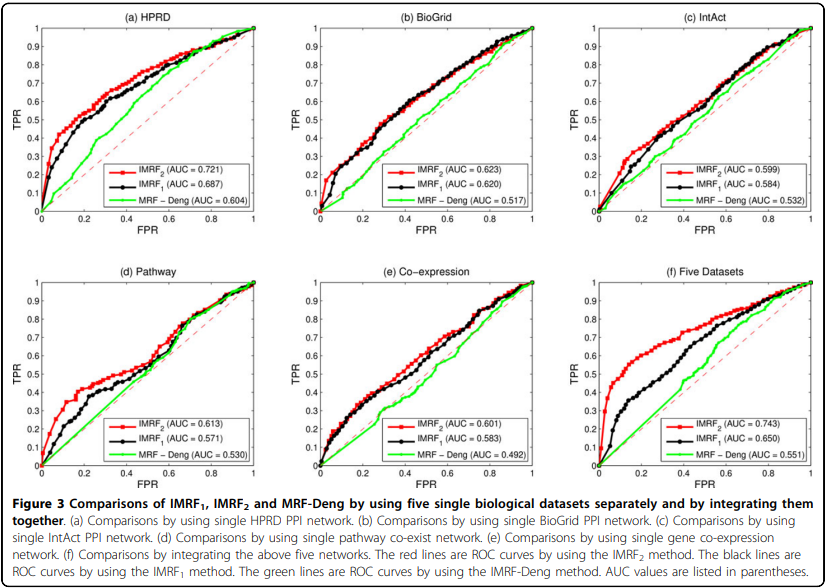
However, for disease gene identifications, only dozens of disease genes are available for individual disease classes. The estimated parameters of the MRF-Deng method becomes unreliable. stabilized Markov processes and parameters do not indicate they converge to expected results. It seems the performance of the MRF-Deng method is better.



**Comparisons with the MRF-Deng method:**

Our improved methods are significantly better than the MRF-Deng method in terms of identifying disease genes. Predictions of the IMRF1 method is significantly better than that by using the MRF-Deng method, but is a little worse than the IMRF2 method. In terms of informativeness of each biological network, the HPRD PPI network (shows in Figure 3 (a)) is the most informative data source, which obtains the highest AUC value in all three methods.





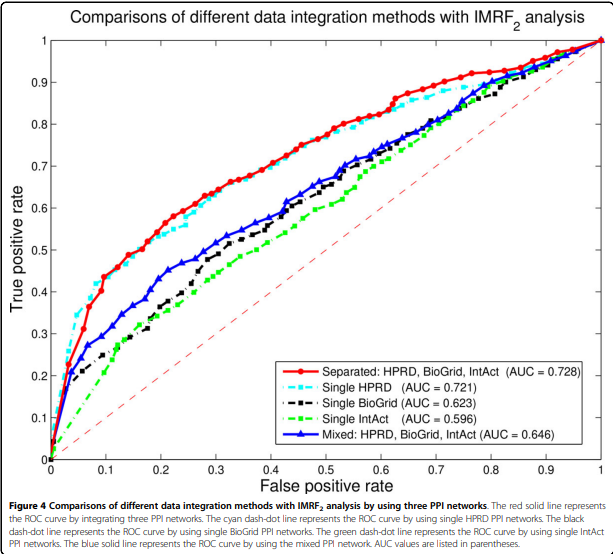
**Integration of heterogeneous data sources:**

Different biological datasets are commonly heterogeneous. When information in those data is integrated, noises are also integrated. Hence, an inappropriate method may result in a set of worse predictions than using only single dataset. Generally, various data integration methods can be divided into two categories: (1) by using a mixed network and (2) by using several separated networks. Generally, separated networks contain more information than the mixed network, since it is very easy to generate the mixed network from several separated networks but not vice versa. One advantage of the MRF model is that it takes the whole network into consideration, which potentially yields better performance than those using mixed network ones.

The separated network method achieves the best performance among all predictions, while the mixed network method achieves only modest performance.

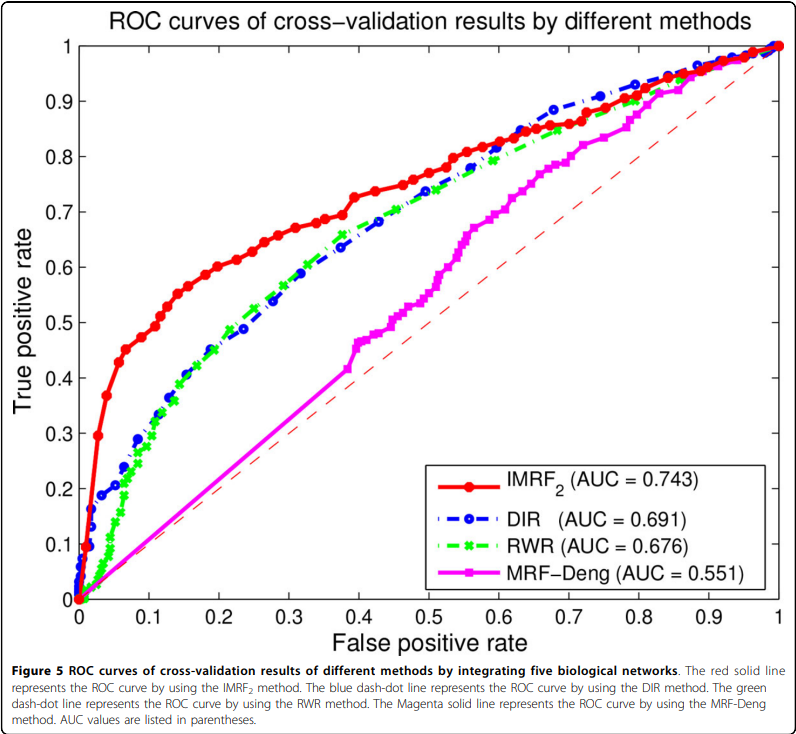
**Comparisons by using multiple data sources:**

The IMRF2 method is compared with the RWR algorithm, the DIR algorithm and the MRF-Deng algorithm respectively. Figure 5 illustrates ROC cross-validation results by integrating all five biological networks.



The IMRF2 method achieves the highest AUC score at 0.743, followed by the DIR algorithm (AUC = 0.691) and the RWR algorithm (AUC = 0.676). The MRF-Deng method achieves the AUC score only at 0.551. It also shows that the separated network interaction method performs better than the mixed

network RWR method.



**Conclusions:**

In this paper, we have presented an improved multiple data integration method for prioritizing human disease genes, which is based on the theory of MRF and the method of Bayesian analysis. Compared to the MRF-Deng method [30], two strategies have been developed to significantly improve the performance of the MRF method for disease gene identifications.

Firstly, parameters of our improved MRF methods are estimated according to all labelled vertices in integrated biological networks, instead of estimating them according to only known vertices. Moreover, parameters are updated together with sampling labels during iterations, instead of using fixed parameters. The improved parameter estimation method makes our MRF methods more stable and more reliable.

Secondly, a new “prediction period” is added to Gibbs sampling process. Parameters of this period is obtained by taking average parameters in the previous “lag period” and is fixed during iterations of this period. The input probability is also obtained by taking average of posterior probabilities in the “lag period”. This strategy significant improves the prediction accuracy of our method.

Predictions when integrating known gene-disease associations, protein complexes, PPIs, pathways and gene expression profiles achieve the AUC score of 0.743, which is better than the RWR method and the DIR method by using the same datasets.

**List of abbreviations**:

MRF, Markov random field; PPI, protein-protein interaction; RWR, random walk with restart; DIR, data integration rank; MLE, maximum likelihood estimation; MCMC, Markov chain Monte Carlo; OMIM, online Mendelian inheritance in man; PCC, Pearson correlation coefficient; ROC, receiver operating characteristic; AUC, area under the ROC curve.

**Title: Recent advances in predicting gene–disease associations [version 1; peer review: 2 approved] (আরণ্যক)**

Deciphering gene–disease association is a **crucial step in designing therapeutic strategies against diseases**.There are **experimental methods for identifying gene–disease associations**, such as **genome-wide association studies** and **linkage analysis**, but these can be **expensive and time consuming**.As a result, various in **silico methods for predicting associations from these and other data have been developed using different approaches.**

**Introduction:**

Aberrations in certain genes have been observed to either predispose individuals to disease or be directly responsible for the development of a disease phenotype, as in the case of **Huntington’s disease1** and **sickle cell disease2** . **Deciphering the link between genes and diseases is an open problem in biomedical sciences**, here we are describing only the links or associations between genes and disease rather than suggesting causality, as the issue of causality is still under debate

Experimental methods for gene–disease association, such as **linkage studies3** , **genome-wide association studies (GWAS)4** , and **RNA interference screens5** , are expensive and time consuming to run

**Genome variation**

GWAS and genetic linkage studies3 are the main methods used for identifying variations across the genomes of individuals and associating these with diseases or phenotypes. The idea behind GWAS is to establish whether there is a significant genetic variation between case and control populations for a given phenotype under investigation. The most common type of variation studied for diseases is the variation at a single nucleotide position, otherwise known as the single nucleotide polymorphism (SNP), although other types of variation such as copy number or chromosomal rearrangements have also been linked to many diseases. GWAS identify marker SNPs that are associated with the phenotype/trait under investigation. Once the marker SNPs have been identified, the next challenge is to determine how the variants are responsible for the phenotypes. This entails finding the location of the SNPs in relation to genes and, if associated with a gene, then identifying the pathways the gene is involved in. Genetic linkage studies, on the other hand, identify linked regions on the genomes of related

**Crowdsourcing**

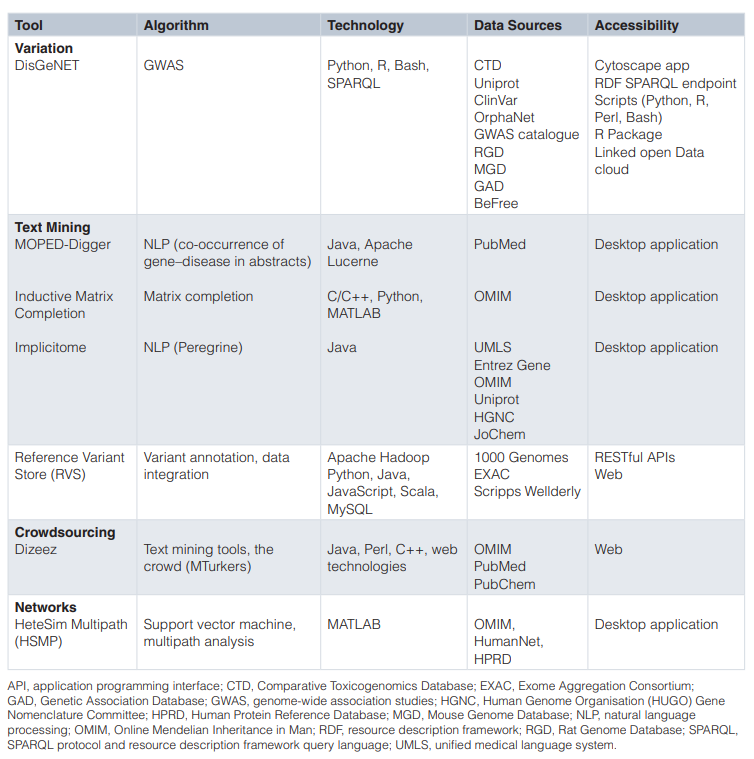
Crowdsourcing refers to the act of delegating a job traditionally assigned to a dedicated agent (usually an employee) to a large group of people in the form of an open call32. The immense quantity of data that biomedical scientists need to deal with today has prompted the search for innovative ways of solving scientific problems. The following qualities identify suitable candidates for crowdsourcing solutions: 1) Few individuals with rare abilities could solve the problem. It is sometimes difficult to harness all the necessary skills for a particular task in one organization or through traditional ways of collaboration. 2) The problems are simple tasks that require human intelligence, e.g. annotating images. 3) The problems can be broken into tasks with definite endpoints. The possibility of breaking jobs into smaller tasks translates to the possibility of sharing the incentives with a larger group of people and, in essence, simplifying the problem

**Text mining**

The bulk of scientific knowledge is still kept in textual format, although the availability of these data in scientific databases is also growing exponentially. For instance, Burger et al. 22 estimate that articles about gene–disease associations that are deposited in public databases grow at the rate of about 10,000 papers per year (approximately one paper every hour of every day). As a result, there is an increasing need to find better and faster ways of retrieving and processing knowledge from scientific databases. Databases that are manually curated by experts provide highquality data, albeit at a very slow pace, so text mining algorithms are now being used to automate some manual processes.

**Networks and semantic similarity-based algorithms**

Network algorithms rely on the premise that phenotypically similar diseases are caused by genes that are functionally related40. The idea is to find a set of genes that are already linked to the disease or phenotype in question and then find genes that are functionally related to that set. Many examples of network-based methods have been reviewed in Piro & Cunto6 and two are mentioned below. HeteSim41 integrates heterogeneous networks of protein–protein interaction (PPI), gene–phenotype association, and phenotype–phenotype similarity to prioritise novel gene– phenotype associations. Natarajan & Dhillon42 formulate the gene–disease association problem in a similar way to a recommendation problem in which the players are genes as the “recommenders”, and diseases are the “items” that they recommend or “prefer”. The goal is to identify which diseases a given set of genes would prefer given a set of observed preferences provided as biological entities

****

# **Title:** Identifying Disease-Gene Associations With Graph-Regularized Manifold Learning (Symum)

Complex diseases are known to be associated with disease genes.

A manifold learning-based method is proposed for predicting disease-gene associations by assuming that the geodesic distance between any disease and its associated genes should be shorter than that of other non-associated disease-gene pairs.

In the 3-fold cross-validation experiments, our method achieves scores of 0.882 and 0.854 in terms of the area under of the receiver operating characteristic (ROC) curve (AUC) for diseases with more than one known associated genes and diseases with only one known associated gene, respectively. Further de novo studies on Lung Cancer and Bladder Cancer also show that our model is capable of identifying new disease-gene associations.

**Introduction:**

Identifying disease-gene associations is of critical importance since it helps us unravel the mechanisms of diseases, which has many applications such as diagnosis, treatment and prevention of disease.

Computational methods that translate the experimental data into legible disease-gene associations are necessary for in-depth experimental validation.

Currently, many algorithms have been developed to predict disease-gene associations, and they can be briefly divided into two categories: **the machine learning-based methods and the network-based methods**

**The typical machine learning-based methods** extract gene-related features and train models that can discriminate disease genes and passenger genes

these algorithms are usually single-task algorithms which once can only predict disease genes for a specific disease.

**Matrix completion methods**, as a type of machine learning methods, can solve the above two issues by jointly predicting **diseasegene** associations and leveraging the similarities among diseases during the calculation (Natarajan and Dhillon, 2014; Zeng et al., 2017)

**Networkbased methods** are based on the assumption that genes close related in the network are associated with the same diseases. Centrality indices, random walk and network energy are used in many methods to predict disease-gene associationss (Köhler et al., 2008; Vanunu et al., 2010; Chen et al., 2014a,b)

y, we propose a manifold learning-based method (**dgManifold**) to predict disease-gene associations.

In our dgManifold, genes and diseases are regarded as points in the same high-dimensional Euclidean space

In this study, we propose a manifold learning-based method (**dgManifold**) to predict disease-gene associations. In our dgManifold, genes and diseases are regarded as points in the same high-dimensional Euclidean space. Our assumption is that diseases and their associated genes should be consistent in some lower dimensional manifold, and the geodesic distance between a disease and its associated genes should be shorter than that of other non-associated disease-gene pairs. Although the Euclidean distance between diseases and genes in the high dimensional space may not reflect their true geodesic distance, we can map the diseases and genes into a low-dimensional manifold based on the experimentally verified disease-gene associations (Tenenbaum et al., 2000; Ham et al., 2005). Then, the true geodesic distance between all the disease-gene pairs can be calculated. In the meantime, the mapping process is regularized by two affinity graphs, disease similarity network and gene similarity network, so that the learned representations with the similarity information can further increase the prediction accuracy. Additionally, since our dgManifold is a supervised method, and it is difficult (if possible) to learn valuable representations for diseases that only have a few or no known associated genes. A prior information vector calculated with the disease similarities and known disease-gene associations should be combined with the original association data to solve this issue. Similar strategies have been applied to calculate the initial probabilities used in the random walk, which have improved the accuracy of predicting miRNA-disease associations (Chen et al., 2016b, 2018a,b).

2. MATERIALS AND METHODS

2.1. General Model

2.2. Similarity Network

2.2.1. Gene Similarity

2.2.2. Disease Similarity

2.3. Prior Information

2.4. Data Sources

2.5. Evaluation Metrics

3. RESULTS

3.1. Model Parameters

3.2. Cross-Validation

3.3. De novo Study

4. CONCLUSION

DATA AVAILABILITY

AUTHOR CONTRIBUTIONS

F-XW conceived this study. F-XW, PL, QX, P-JW, and BL discussed the methods. PL implemented the algorithm, designed and performed the experiments. PL and F-XW wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

This work is supported in part by Natural Science and Engineering Research Council of Canada (NSERC) and China Scholarship Council (CSC).

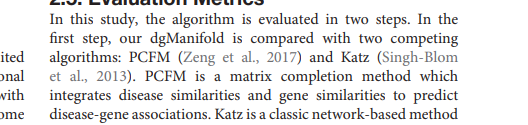
**There are two type of method:**

**machine learning method**

**extract gene-related features and train models**

**Matrix completion methods, as a type of machine learning methods**

**Network based method**

****