Learning graph representations for disease gene prediction

Ata, Kircali Sezin

2019

Ata, K. S. (2019). Learning graph representations for disease gene prediction. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/137913>

<https://doi.org/10.32657/10356/137913>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 20:52:57 SGT

Learning Graph Representations for Disease Gene Prediction

ATA KIRCALI SEZIN

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
2019

Learning Graph Representations for Disease Gene Prediction

ATA KIRCALI SEZIN

School of Computer Science & Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2019

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

06/08/2019

.....

Date

.....

ATA KIRCALI SEZIN

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

06/08/2019

.....

Date

.....

Assoc. Prof. Li XIAOLI

Authorship Attribution Statement

This thesis contains material from 2 papers published in the following peer-reviewed journals and 1 book chapter in which I am listed as an author.

Chapter 3 contains material from a study published as [S.K. Ata, Y. Fang, M. Wu, X. Li, X. Xiao. Disease gene classification with metagraph representations. *Methods* 131, 83-92 \(2017\). DOI: 10.1016/j.ymeth.2017.06.036.](#) and

The contributions of the co-authors are as follows:

- All the co-authors guided the project direction and edited the manuscript drafts.
- I prepared the manuscript drafts. The manuscript was revised by all the co-authors.
- I performed all coding regarding data preprocessing, feature engineering, training and validation.
- Assist. Prof. Y. Fang. provided guidance about tools Grami and SubMatch for metagraph mining.

Chapter 3 contains material from a study published as [S.K. Ata, Y. Fang, M. Wu, X. Li, X. Xiao. Disease gene classification with metagraph representations. *Data mining for systems biology: Methods and protocols*, 211-224 \(2018\). DOI: 10.1007/978-1-4939-8561-6_16](#) and

The contributions of the co-authors are as follows:

- All the co-authors guided the project direction and edited the manuscript drafts.
- I prepared the manuscript drafts. The manuscript was revised by all the co-authors.
- I performed all coding regarding data preprocessing, feature engineering, training and validation.

Chapter 4 contains material from a study published as [S.K. Ata, L. Ou-Yang, Y. Fang, C. Kwok, M. Wu, X. Li. Integrating node embeddings and biological annotations for genes to predict disease-gene associations. *BMC Systems Biology* 12\(Suppl 9\), 138 \(2018\). DOI: 10.1186/s12918-018-0662-y.](#)

The contributions of the co-authors are as follows:

- All the co-authors guided the project direction and edited the manuscript drafts.
- I prepared the manuscript drafts. The manuscript was revised by all the co-authors.
- I performed all coding regarding data preprocessing, feature engineering, training and validation.

- Dr. W. Min. provided guidance in the interpretation and analysis of the results.

.....06/08/2019.....

Date

.....

ATA KIRCALI SEZIN

Acknowledgements

I would like to express my sincere thanks and appreciation to my supervisor Assoc. Prof. Li Xiaoli (I²R), co-supervisor Assoc. Prof. Kwoh Chee Keong (NTU), former supervisor Assoc. Prof. Xiao Xiaokui (NUS), and advisors Dr. Wu Min (I²R) and Assist. Prof. Fang Yuan (SMU) for their meticulous guidance, vital suggestions and supportive attitudes. Their knowledge and our common research interests help me a lot to overcome the difficulties encountered during my entire research. Additionally, I would like to thank dear TAC members: Prof. Cong Gao and Assoc. Prof. Pan Sinno Jialin for their valuable feedbacks during our meetings. I am also grateful to NTU laboratory managers, Loo Kian Hock (DMAL) and Irene Goh (PDCL) for their understanding and support which made extensive experiments easier.

Last but not least, I want to thank my family for their constant encouragement.

Abstract

The analysis of disease-causing conditions based on genes and their protein products plays a crucial role in the diagnosis and treatment of several serious diseases such as cancer and diabetes. Since experimental techniques are time-consuming and expensive, computational methods preserve their significance in revealing the functional roles of genes/proteins in the context of many diseases. Molecular networks such as protein-protein interaction and co-expression networks are useful computational tools to figure out the underlying mechanism of diseases by studying the complex interplay between genes. However, molecular networks are often noisy and incomplete. Therefore, computational approaches are designed in such a way that they are able to complement and enhance network data. The vast amount of cumulated data from various scientific domains are exploited to extract useful information for disease gene prediction efficiently. The goal of this research is to develop techniques which extract robust network-based feature representations for prediction of disease-causing genes with a wider perspective by combining both topological characteristics of molecular networks and biological knowledge such as gene ontology and protein domain. However, exploiting the topological arrangements of proteins in the context of not only the interactions between them but also their relevance based on other biological properties is another challenge. Furthermore, hand-engineering to extract complex network features needs tedious efforts with domain expertise. Automating the extraction of these topological features and combining them with the biological aspects of proteins is a challenge to be addressed. For this purpose, node embedding models are applied to automate the extraction of useful feature representations for disease gene prediction. Besides, there are various molecular networks which provide deeper insight into genes and diseases from multiple perspectives. Thus, a unified computational framework can leverage these various perspectives to generate more robust representations.

The design of models, which are able to automate the extraction of feature representations through multiple networks is essential to achieve higher prediction performance. In this thesis, we propose three computational frameworks to predict candidate disease genes:

1. The first proposed computational method to predict candidate disease genes is called *Metagraph*. To complement and enrich the protein-protein interaction (PPI) networks, *Metagraph* leverages the biological properties of the individual proteins by integrating the ontological properties, named as *keywords*, of proteins into the PPI network, and constructs a novel PPI-Keywords (PPIK) network composed of both proteins and keywords as two different types of nodes. As disease proteins tend to exhibit similar topological properties on the PPIK network, we further propose to represent proteins with metagraphs. Apart from a traditional network motif or subgraph, a metagraph is able to capture both topological arrangements involving the interactions between proteins and the associations between the proteins and keywords. Thus, proteins which are not neighbors in a noisy PPI network have a better chance to be topologically similar through keywords. Extended metagraph representations considering the disease occurrences, called *Metagraph+*, are fed into various classifiers for disease protein prediction in a supervised manner. Conducted experiments show that *Metagraph+* consistently improves disease protein prediction on three different PPI databases and outperforms the state-of-the-art baselines including both diffusion-based methods and module-based methods. In addition, predictions of *Metagraph+* attain better correlations with the literature findings from PubMed database.

2. Numerous studies for the discovery of disease-associated genes and their part in the development of a disease have been proposed over the last decades. Yet, automatically extracted features from a molecular network, have not been exploited for disease gene prediction. The second proposed technique in this thesis is an integrative framework called N2VKO which adopts a well-known representation learning

method, *node2vec*. We combine network-based feature representations of genes obtained by node2vec with biological aspects of proteins. Then, we apply various feature selection methods to analyze their performance on disease gene prediction task. As the data for disease gene prediction is imbalanced, we further address this data imbalance issue by applying oversampling techniques on our novel representations to improve the prediction performance. Extensive experiments show that N2VKO significantly outperforms four state-of-the-art methods for disease gene prediction across seven diseases. Moreover, the categories of the biological aspects within N2VKO representations are listed to analyze their role in disease formation. We also provided literature evidence for N2VKO biological features over lung cancer. Finally, the literature evidence from PubMed database reveals the effectiveness of our proposed N2VKO framework for disease gene prediction.

3. The third proposed framework in this thesis addresses the candidate disease gene prediction problem. Since there are various biological networks providing different insights of genes, combining them as complementing each other in a unified framework is necessary to improve disease gene prediction performance. We propose a novel unsupervised algorithm for “**Multi-view network embedding with Intra-Cross**” consistencies (MICROS). This approach learns low-dimensional representations to be fed into various downstream tasks through a multi-view network embedding framework. MICROS is based on two well-known principles: diversity and collaboration. Former enables views to maintain their topological characteristics, latter enables views to work together and reinforce each other. Unlike existing methods, we also examine a novel form of higher-order collaboration that has not been explored previously on multi-view networks, and further integrate it into a unifying framework of consistencies to provide more robust, superior node representations. Finally, we conduct extensive experiments on three real-world multi-view networks. Our results demonstrate that our learned representations consistently outperform state-of-the-art approaches on various downstream tasks

namely node-level tasks (i.e., classification and clustering), relationship mining and link prediction.

Contents

Acknowledgements	ii
Abstract	iii
List of Figures	x
List of Tables	xiii
1 Introduction	1
1.1 Background	1
1.1.1 Motivation and Objectives	1
1.1.2 Challenges	4
1.2 Related Prior Work	5
1.3 Major Contributions and Organization	9
1.4 Outline	12
2 Literature Review	14
2.1 Disease and Network	14
2.2 Biological Data and Network Representation	16
2.3 Integrative Approaches	20
2.3.1 Single-view Approaches	21
2.3.1.1 Homogeneous Network Integration	21
2.3.1.2 Heterogeneous network integration	22
2.3.2 Multi-view Approaches	24
2.4 Network-Based Techniques in Disease Gene Prediction	25
2.4.1 Module-based Methods	25
2.4.2 Diffusion-based Methods	32
2.4.3 Representation-based Methods	37
2.5 Summary	42
3 Disease Gene Classification with Metagraph Representations	46
3.1 Introduction	47
3.2 Materials and Methods	49
3.2.1 Preliminaries	49

3.2.2	PPI-Keyword network	50
3.2.3	Metagraph representations	52
3.2.4	General framework	55
3.3	Results and Discussion	57
3.3.1	Data and setup	57
3.3.2	Benefits of metagraph-based representations	59
3.3.3	Comparison to baselines	62
3.3.4	Further analysis of the predicted disease proteins	65
3.3.5	Biological analysis for <i>Metagraph</i> representations ($\mathbf{m}_p$)	68
3.4	Summary	70
4	Integrating Node Embeddings and Biological Annotations for Genes to Predict Disease-Gene Associations	71
4.1	Introduction	72
4.2	Method	74
4.2.1	Data Preparation	75
4.2.2	Learning the Feature Representations for Genes	76
4.2.2.1	node2vec	76
4.2.2.2	N2VK Representations	78
4.2.3	Feature Selection	79
4.2.4	Imbalance Correction	80
4.3	Results and Discussion	81
4.3.1	Experimental Setup	81
4.3.2	Classification Model Selection	82
4.3.3	Impacts of Keywords and Oversampling	85
4.3.4	Feature Selection and Importance Analysis	88
4.3.5	Comparison with State-of-the-arts	91
4.3.6	Further Analysis on the Effect of Biological Annotations	97
4.3.7	Comparison of the Integration Methods	97
4.3.8	Case Studies	99
4.4	Summary	100
5	Unsupervised Multi-View Network Embedding with Higher-Order Collaboration	101
5.1	Introduction	102
5.2	Method	107
5.2.1	Problem formulation	107
5.2.2	Background on skip-gram model	108
5.2.3	Proposed Approach: MICROS	109
5.2.3.1	Intra-cross consistencies	110
5.2.4	The MICROS algorithm	114
5.2.4.1	Overall framework	114
5.2.4.2	Loss function	115
5.2.4.3	Algorithm	116

5.2.4.4	Complexity analysis	116
5.3	Results and Discussion	117
5.3.1	Experimental setup	118
5.3.1.1	Datasets	118
5.3.1.2	Baselines	119
5.3.1.3	Parameter configuration	120
5.3.2	Baseline comparison on evaluation tasks	121
5.3.2.1	Algorithms and metrics for tasks	121
5.3.2.2	Node-level tasks	121
5.3.2.3	Edge-level tasks	123
5.3.3	Model analysis	124
5.3.3.1	Ablation study	125
5.3.3.2	Scalability study	126
5.3.3.3	Convergence analysis	126
5.3.3.4	Weights α and β	127
5.3.3.5	Number of embedding dimensions D	127
5.3.3.6	Methods for Combining the Final Embeddings	129
5.3.4	Further Analysis on the Effect of Second-order Collaboration in Disease Gene Prediction	130
5.3.5	Comparison among Proposed Studies: Metagraph+, N2VKO, MICROS	131
5.4	Summary	132
6	Conclusions and Future Directions	134
6.1	Conclusions and Discussion	134
6.2	Future Directions	137
6.2.1	Network reconstruction - link prediction	137
6.2.2	Attention-based end-to-end multi-view network embedding	139
6.2.3	Convolutional neural networks (CNNs)	139
6.2.4	PU learning techniques	140
6.3	Final Remarks	141
	List of Author's Awards, Patents, and Publications	143
	Bibliography	144

List of Figures

1.1	Categories of the existing network-based studies for disease gene prediction.	7
1.2	Overall schema for the major contributions.	10
2.1	Perturbations in molecular networks. The figure is adapted from the study [1].	16
2.2	Homogeneous network integration.	22
2.3	Projected heterogeneous network integration (PHNI). There are three different networks: gene-gene interaction, disease-disease similarity network and gene-disease association network (dashed orange lines). In the resultant network, thicker lines represent the stronger association.	23
2.4	Pathway identification. Two disease module members, protein <i>A</i> , protein <i>B</i> and predicted disease protein <i>C</i> , which falls within the shortest path between <i>A</i> and <i>B</i>	26
2.5	Unweighted PPI network to describe ICN [2] method. The shortest path between proteins <i>A</i> and <i>D</i> includes nodes <i>B</i> and <i>C</i>	27
2.6	2-, 3-, and 4-node graphlets.	31
2.7	Diffusion-based Method.	34
2.8	Heterogeneous network of study RWRH [3].	36
3.1	Part of the PPIK network based on UniProt and IntAct databases.	52
3.2	Example metagraphs: common structures of subgraphs on the PPIK network.	53
3.3	Example subgraph instances of a metagraph, where some contain disease proteins and some do not (Q9BRI3 is a known disease protein).	54
3.4	Metagraphs up to 4-nodes in this study. Green nodes represent the <i>keywords</i> , blue nodes represent the proteins.	55
3.5	General framework of the proposed method.	56
3.6	Performance of Metagraph and Metagraph+ compared to Keywords for <i>all disease</i>	59
3.7	Performance of combinations among $\mathbf{m}_p$, $\mathbf{d}_p$, $\mathbf{k}_p$ for <i>all disease</i>	61
3.8	Performance comparison between metagraphs with up to 4-nodes and up to 5-nodes for <i>all disease</i>	61
3.9	Performance of Metagraph and Metagraph+ compared to Keywords for <i>breast cancer</i>	62

3.10	Average number of PubMed publications per prediction based on DisGeNET.	66
4.1	Framework of our N2VKO for Predicting Disease-Gene Associations	74
4.2	node2vec transition probabilities on a small PPI network. A random walk traversed from node t to node v and now resides at node v . The transition probabilities from node v to other proteins x_1, x_2, x_3 and x_4 are shown in this Figure.	78
4.3	AUC comparison for various classifiers on node2vec embeddings through 7 diseases. “Average” in the last column shows the average performance of classifiers over all the seven diseases on a) IntAct dataset and b) NCBI dataset.	83
4.4	AUC performances over different p and q pairs for disease Alzheimer’s and breast cancer.	84
4.5	The effect of embedding dimension D on the average disease performance on IntAct dataset.	85
4.6	Comparison among node2vec, N2VO, N2VK and N2VKO on a) IntAct dataset and b) NCBI dataset.	86
4.7	FS and oversampling applied solely on <i>node2vec</i> embeddings.	87
4.8	FS and oversampling applied on <i>Metagraph+</i> representations.	88
4.9	Comparison between ROSE and SMOTE on a) IntAct dataset and b) NCBI dataset. ROSE refers to applying ROSE on node2vec embeddings, while ROSE-FS refers to applying ROSE after the feature selection on node2vec embeddings.	89
4.10	AUC comparison among 4 feature selection methods on N2VK representations.	90
4.11	AUC comparison for various combinations of feature selection and oversampling techniques in N2VKO framework on cancer-related diseases. Four cancer-related diseases include <i>breast cancer</i> (BC), <i>colorectal cancer</i> (CC), <i>lung cancer</i> (LC) and <i>prostate cancer</i> (PC).	91
4.12	The percentages of <i>Keywords</i> categories used in N2VKO representations through 7 diseases. The percentage contribution of <i>Keywords</i> categories in N2VKO is investigated a) IntAct dataset. b) NCBI dataset.	92
4.13	Comparison between our N2VKO and various state-of-the-arts for disease gene prediction.	94
4.14	One-hop sub-networks for various diseases.	95
4.15	Connected component and clustering coefficient analysis for one-hop disease sub-networks.	96
4.16	To show the effect of biological annotations (i.e., <i>keywords</i>), two models consists of the same feature selection and oversampling procedure applied on solely node2vec embeddings (i.e., N2V-FSO) and its keywords concatenated version (i.e., N2VKO) are compared across seven diseases on datasets a) IntAct b) NCBI.	98

5.1	A real-world multi-view network consists of two views: PPI view (a part of PPI network from IntAct database) and GO view (a part of Gene Ontology network).	103
5.2	A toy protein multi-view network with three views.	104
5.3	Major characteristics of a multi-view network.	104
5.4	Overall framework of MICROS , capturing diversity and two-orders of collaboration using intra-cross consistencies on a multi-view network.	111
5.5	Intra-cross consistencies in four quadrants.	114
5.6	Visualization of protein nodes: alzheimer (■), breast cancer (▲), colorectal cancer (●), and lung cancer (★).	122
5.7	Comparison of MICROS with two simpler variants.	125
5.8	Scalability analysis of MICROS	126
5.9	Convergence analysis of MICROS.	127
5.10	Impact of hyperparameter α	128
5.11	Impact of hyperparameter β	128
5.12	Impact of dimension D	128
5.13	Ablation study to demonstrate the contribution of second-order collaboration through seven diseases.	131
5.14	AUC comparison among proposed methods. Unlike the proposed N2VKO, random walk parameters p and q are not tuned for a reasonable comparison between N2VKO and MICROS. Besides, the dataset of MICROS consists of GO network and IntAct PPI interactions. For a better comparison we perform N2VGO, i.e., the same feature selection and oversampling procedure of N2VKO is applied on the concatenated node2vec embeddings from these two networks.	132
6.1	Unlike the social networks, in a PPI network, proteins with many number of shared protein neighbors do not necessarily interact with each other (X and Y). Instead, proteins occurred together in many $l = 3$ paths, are more likely to interact and form a link (red link). The figure is adopted from the study [4].	138

List of Tables

2.1	Module-based models surveyed in this thesis. Columns present the name, main technique(s), view types (i.e., single or multi-view), integration type (i.e., none (-), homogeneous or heterogeneous), network(s), complexity (i.e., linear (L) or polynomial (P)), formulation (i.e., hand-crafted (H) or automated (A)), solution (i.e., problem-specific (P) or generic (G)) and release years of the studies.	33
2.2	Diffusion-based models surveyed in this thesis. Columns present the name, main technique(s), view types (i.e., single or multi-view), integration type (i.e., none (-), homogeneous or heterogeneous), network(s), complexity (i.e., linear (L) or polynomial (P)), formulation (i.e., hand-crafted (H) or automated (A)), solution (i.e., problem-specific (P) or generic (G)) and release years of the studies.	38
2.3	Representation-based models surveyed in this thesis. Columns present the name, main technique(s), view types (i.e., single or multi-view), integration type (i.e., none (-), homogeneous or heterogeneous), network(s), complexity (i.e., linear (L) or polynomial (P)), formulation (i.e., hand-crafted (H) or automated (A)), solution (i.e., problem-specific (P) or generic (G)) and release years of the studies.	43
3.1	A summary of keywords from the UniProt database.	48
3.2	Summary of the three PPIK networks.	58
3.3	Performance of Metagraph+ compared to random walk and subgraph baselines.	64
3.4	Running time (minutes) comparisons. For Subgraph+ , Metagraph , and Metagraph+ the classifier with the best AUC performance, GLM, is chosen.	65
3.5	Top 10 genes predicted by Metagraph+ (GLM) with the most number of publications.	67
3.6	Top 5 associated diseases reported by DisGeNET for the 10 genes reported in Table 3.5.	67
3.7	Novel disease genes (mapped from predicted proteins) discovered by Metagraph+ based on recent PubMed publications from years 2014–2016.	68
3.8	Confusion matrix for clustering performance on metagraph structures $\mathbf{m}_p$	69

3.9	Notable biological properties of false negative (FN), true positive (TP), false positive (FP) and true negative (TN) predictions with the number of predictions and their percentages (Pct.%) over the total number of samples presented in the parenthesis (i.e., 2188, 759, 1562, 8554, respectively).	70
4.1	The number of disease-genes in IntAct and NCBI data, respectively.	76
4.2	<i>Keywords</i> in selected features for Lung Cancer on NCBI dataset	93
4.3	Comparison among the integration methods for IntAct dataset.	99
4.4	Predicted disease genes on IntAct with top 10 prediction scores and at least 10 publications in PubMed records.	100
4.5	Predicted disease genes on NCBI with top 10 prediction scores and at least 10 publications in PubMed records.	100
5.1	Summary of main notations.	108
5.2	Summary of datasets.	118
5.3	Performance of node classification tasks on IntAct (bold: best; underline: runner-up).	122
5.4	Performance of node clustering task on IntAct (bold: best; underline: runner-up).	122
5.5	Performance of relationship mining on LinkedIn and link prediction on YouTube (bold: best; underline: runner-up).	123
5.6	Performance of varying integration methods in disease gene prediction.	130
5.7	Biological processes of the proteins, which <i>first-order collaboration</i> is insufficient to identify in prostate cancer analysis.	131

Abbreviations

Al:	Alzheimer’s disease
BC:	Breast Cancer
BFS:	Breadth-First Sampling
CC:	Colorectal Cancer
CNN:	Convolutional Neural Networks
DFS:	Depth-First Sampling
DL:	Deep Learning
DM:	Diabetes Mellitus
DTI:	Drug-Target Interactions
FS:	Feature Selection
GLM:	Generalized Linear Model
GO:	Gene Ontology
HNI:	Heterogeneous Network Integration
IMC:	Inductive Matrix Completion
kNN:	k-Nearest Neighbor
LC:	Lung Cancer
MRMR:	Minimum Redundancy - Maximum Relevance
N2VK:	FS applied on <i>node2vec</i> + <i>Keywords</i> , proposed representations
N2VKO:	Oversampling is applied on the train set obtained by N2VK
N2VO:	Oversampling is applied on <i>node2vec</i> protein representations
NMF:	Nonnegative Matrix Factorization
Ob:	Obesity
OMIM:	Online Mendelian Inheritance in Man
PC:	Prostate Cancer
PHNI:	Projected Heterogeneous Network Integration
PPI:	Protein-Protein Interaction
PTM:	Post-Translational Modification
PU:	Positive-Unlabeled
RF:	Random Forest
RWR:	Random Walk with Restart [5]
RWRH:	Random Walk with Restart on Heterogeneous Network [3]
SVM:	Support Vector Machine
TT:	Technical Term

Chapter 1

Introduction

1.1 Background

1.1.1 Motivation and Objectives

A disease can be defined as an anomalous state affecting an organism [6]. We can assort human diseases into three different categories: pathogenic/infectious, deficiency, and genetic diseases. Pathogenic diseases are caused by external factors such as bacteria, viruses, fungi, and parasites which are called pathogens. Deficiency diseases, on the other hand, are caused by insufficient essential dietary elements such as vitamins and minerals over a long period of time. The scope of our research does not include the first two categories. Our study is based on genetic diseases. Contrary to popular belief, genetic disorders may or may not be hereditary. Particularly, hereditary diseases have the potential of being passed down a line of generations such as hemophilia, and sickle cell anemia. However, in both cases, they are mostly caused by a mutation. Besides, recent studies reveal that epigenetic factors can also play a role in both types of genetic diseases [7].

Unsurprisingly, disease-causing genes or disease-associated genes and their protein products play a crucial role in the diagnosis and treatment of diseases such as Alzheimer’s, diabetes mellitus, and obesity. For instance, mutations in gene APP affect its encoded protein product the amyloid precursor protein and could cause Alzheimer’s disease. Besides, Alzheimer’s is a genetically heterogeneous disorder, i.e., there are other genes such as the presenilin-1 gene (PSEN1), a disintegrin and metalloprotease family gene (ADAM10), and a subfamily of ATP-binding cassette gene (ABCA7) which their mutations could play role in Alzheimer’s disease as well. Despite the recent advances in this area [8–11], it is still a challenging research issue to produce new computational methods to reveal interactions of disease causing proteins and functions in the context of many diseases.

In order to figure out the underlying mechanism of genetic diseases, several approaches have been dedicated, such as linkage analysis and genome-wide association studies (GWAS). These studies are based on genetic variations among humans known as single nucleotide polymorphisms (SNPs). Correlated SNPs with specific diseases provide predictive biomarkers. However, statistical analysis of the biomarkers demands vast amounts of money and time due to a high number of false positives that need to be filtered out [12]. Additionally, these techniques are based on a simple genotype-phenotype relationship. However, the analysis of the collaborations and the participations of the molecules in biological functions or disease formations can be performed more systematically by projecting their relationships onto a wired network/graph structure. Furthermore, it is known that the diseases are consequences of perturbations in the molecular networks caused by various complex factors such as genetic mutations, epigenetic factors, and pathogens. Potential perturbations can be easily captured by tracking possible edge or node removals in the occurrence of a disease. Therefore, molecular networks are much more efficient data representations in the analysis of diseases by tracking the potential disruptions on the biological pathways due to the perturbations in the molecular

networks [1]. Thus, for the prediction of disease genes, molecular networks are important instruments.

There are several molecular networks such as protein-protein interaction (PPI), regulatory, genetic and co-expression networks. In this thesis, we mainly work on PPI networks which have been largely exploited over the years [13–15] for disease gene prediction problem. Proteins are responsible for several vital duties within an organism such as biochemical reactions, transmission of signals between cells, and providing structure for cells and tissues. To complete these tasks they work in collaboration and interact with each other. Modeling these interactions is simple and effective as a graph/network data structure. Numerous studies have fortified the notion that the position of a protein in a PPI network is not random. In fact, the proteins having the same phenotype or function tend to display common topological characteristics in the network such as the degree, coreness, closeness, etc [16, 17]. Therefore, PPI networks can be employed in the extraction of useful features for proteins towards predicting the associations between proteins and diseases. However, they are known to be incomplete and noisy [18]. Thus, leveraging multiple aspects of proteins such as GWAS, gene-expression, gene ontology, and domain knowledge is necessary. Since both GWAS and gene-expression data are known to be biased based on mostly studied diseases and genes, in this thesis we use ontology of proteins obtained from Uniprot database [19], called keywords, providing several insights of proteins, including gene ontology and domain aspects of proteins.

In this thesis, our focus is to develop techniques, which extract robust network-based (i.e., using network data) feature representations for prediction of disease-causing genes, with a wider perspective by combining both topological characteristics of molecular networks and biological knowledge, such as gene ontology and protein domain. Provided candidate disease genes can be analyzed further by the scientists within a shorter time and at a lower cost compared to aforementioned

techniques. Note that, in this research we mainly work on the protein-protein interaction networks, and thus disease genes and disease proteins are used interchangeably.

1.1.2 Challenges

Molecular networks such as protein-protein interaction and co-expression networks are useful computational tools to understand the underlying mechanism of diseases by studying the complex interplay between genes. However, molecular networks are often noisy and incomplete [20]. For instance, a small percentage of all potential pairwise protein interactions are revealed through the high-throughput experiments and protein-interaction networks are almost 80% incomplete. Additionally, both molecular networks and disease association lists are biased towards much-studied disease genes [18]. One cause of this bias is that some diseases are not well studied, and thus the amount of available clinical data limits the performance of the algorithms for disease gene prediction. Another cause is that the clinical data mostly comprises evidence for frequently mutated genes, consequently, genes with a less number of mutation records tend to be missed by the disease gene prediction algorithms [21, 22]. Therefore, computational approaches should be designed in such a way that they are able to complement and enhance the network data.

To complement and enhance PPI networks, designing models which are able to exploit the topological arrangements of the proteins in the context of not only the physical interactions between them but also their relevance based on other biological properties is another challenge to be addressed. By this way, incomplete and noisy interactions have a chance to be reconstructed through the strengthened, heterogeneous network.

Since the disease gene prediction problem usually suffers from the data imbalanced issue, this may cause the classifier to favor the majority class samples (negatives),

resulting in a lower rate of true positives. Thus, either integrating an oversampling technique into the feature extraction framework or providing competent representations overcoming the imbalanced issue is inevitable.

Furthermore, hand-engineering to extract complex network features, such as degree and closeness, needs arduous efforts with domain expertise. Automating the extraction of these topological features and combining these with biological aspects of proteins is a challenge to be addressed.

Besides, there are various molecular networks which provide deeper insight into genes and diseases from multiple perspectives. A unified computational framework can leverage these various perspectives to generate more robust representations. However, existing studies in heterogeneous network integration mostly fail to take into account the topological structure of multiple networks by preserving their individual connectivity structure. They usually project networks onto a single node type network, which result in connectivity information loss for the other node type networks [23]. Thus, the design of models which are able to automate the extraction of feature representations through multiple networks by preserving their topological structure is essential to achieve higher prediction performance.

1.2 Related Prior Work

Many computational approaches for the discovery of disease-associated/disease-causing genes and their function in the development of a disease have been proposed over the last decades. In the light of these findings, the assumption that the genes residing within the same neighborhood in a molecular network tend to be associated with the same or similar diseases has been exploited.

Numerous studies widely examine the molecular networks such as protein-protein interaction (PPI), metabolic, regulatory, genetic and co-expression networks [24–26]. The reason is in the current modern molecular biology, biological activities can be modeled as a wired network of molecules. One of the most widely exploited molecular networks is the PPI networks which represent the physical interactions between the proteins [27–30]. And they still form a basis for several approaches in the discovery of genes associated with a disease [31, 32]. However, PPI networks are incomplete and only a small percentage of all potential pairwise protein interactions are provided through the high-throughput experiments [18]. Thus, experiments combining multiple types of data with PPI networks have been conducted in recent years.

Several studies approach the incomplete and noisy PPI problem by combining clinical data with PPI networks [21, 22]. Some of these studies [33–35] combine GWAS-based data with PPI network, and some of them integrate gene-expression data into PPI network [36, 37]. Basically, the former maps single-nucleotide polymorphisms and their corresponding P-values to the gene-level, and then combines them with PPI network. The latter analyzes the expression level of each gene in patients compared to normal people and then detects differentially expressed genes or rewired coexpressions (i.e., differential coexpressions caused by a change in co-expression between genes). However, both of these approaches struggle with biased clinical data problem, as clinical data is biased towards more studied diseases and more frequently observed mutated genes [21].

To address the clinical-data problems, a common strategy is combining PPI network with several non-clinical data, such as gene ontology (GO) terms, protein domains, disease similarities, gene similarities, etc. [22, 38, 39]. Over the past decade, representing these different aspects of data as networks and studying them through multiple networks [3, 40] or an integrated single network [41–44] has been widely exploited for disease-gene association prediction.

FIGURE 1.1: Categories of the existing network-based studies for disease gene prediction.

In this thesis we specify the methods utilizing network/graph data in their problem formulation as network-based methods. Network-based methods working on PPI network in protein-disease association prediction can be grouped into three main categories: module-based, diffusion-based, and representation-based methods. These categories are concisely introduced in this section as in Figure 1.1.

Module-based methods are based on the hypothesis that proteins within the same topological or functional module are more likely to be associated with the same disease [45–47]. Clustering [48, 49], k-cores [17], and frequent subgraph mining [50] are variants of this category. Mostly, they are designed to detect subnetworks that contain most of the disease-associated genes in the interactome. Disease module identification techniques performed solely on neighborhood consideration, such as direct neighbors, shared neighborhood and local shortest paths have the risk of sticking to the local evaluation. However, these methods have the advantage of being parameter-free, i.e, they do not have parameters to specify the boundary of clusters or subnetworks.

Diffusion-based methods are designed to traverse through the pathways which are closest to the known disease genes. They initialize known disease proteins as seeds and diffuse along the interactome through random walks [5] or propagation

flow [38]. In particular, these studies adopt pioneer graph-based semi-supervised algorithms such as [51–54] to a PPI network. Upon convergence, the frequency of being visited for a protein is used to rank the proteins with respect to their association with a disease. The proteins, which are placed farther from the seeds (i.e., labeled nodes), are scored with a lower likelihood of being associated with the disease. As a matter of course, diffusion-based methods are widely used for prioritizing candidate disease proteins. Moreover, since they consider the full network topology besides the placement of the known disease genes, they are superior to the module-based methods [27, 29, 55–57].

Representation-based methods construct representations for proteins covering their various topological (degree, the average distance to disease-genes, common neighbors with disease-associated protein, etc.) or biological aspects (molecular function, biological process, cellular component, etc.), and feed them into a machine learning predictor. They transform the original data to a higher-dimensional space, i.e. the feature space, to be fed into a pattern analysis method. Nonnegative matrix factorization (NMF) methods such as inductive matrix completion (IMC) [58], probability-based collaborative filtering [59] and manifold learning [60] as well as positive-unlabeled (PU) learning [42, 61, 62] methods are some endeavors in this field. NMF methods are based on learning latent factors of gene-disease association matrix. They are able to predict disease genes even for a disease that has no known associated genes. However, they generally do not provide global optimal solutions and have difficulty in convergence even for local optimal solutions [63]. PU learning methods, instead of treating the unknown genes as negative, they treat them as samples in unlabeled set U . They basically train biased SVM classifiers to assign specific penalties on different sets of samples.

Use of graph kernels such as subgraphs and graphlets is another technique to obtain descriptive representations for proteins. For instance, graphlet degree signatures [64, 65] are adopted in some studies [24] for feature construction of proteins in

disease-gene association prediction. In this thesis, we present our study Metagraph representations which utilizes both topological and biological information through heterogeneous subgraphs in which proteins participate (Chapter 3).

Recently, network embedding methods [66–69] emerged as a promising direction that aims to learn low-dimensional and continuous vector representations for nodes through a neural network by considering both node itself and its neighborhood. For this purpose, Skip-Gram architecture [70] is extensively used to build models which are capable of constructing associations between the node and its neighborhood. Random walk-based neural embedding methods DeepWalk [68] and its generalized version in terms of random walks, node2vec [66] generates node representations as nodes within the short random walk distance have similar embeddings [71]. Another work [72] exploits neural representations through a multi-layer protein interaction network for multicellular function prediction. A recent study [73] combines graphlet representations with neural embeddings for disease pathway analysis. In this thesis, we present our study N2VKO which is to the best of our knowledge, the first attempt of integrating node2vec embeddings and biological annotations in disease-gene association prediction (Chapter 4). Furthermore, multi-view embedding or multiplex network embedding methods [74–77] exhibit increased popularity especially in social data analysis. While learning the embeddings collaboratively through the views, they also achieve to preserve the individual topological structure of each view. In this thesis, we present our recent multi-view embedding model which is able to predict disease-gene associations (Chapter 5).

1.3 Major Contributions and Organization

The goal of this research is providing descriptive representations of proteins for identifying disease genes. For this purpose we present three contributions as follows:

(A) Metagraph representations

(B) N2VKO representations

(C) MICROS representations

FIGURE 1.2: Overall schema for the major contributions.

- *Metagraph+*: Our first contribution *Metagraph+* provides representations based on novel extracted heterogeneous subgraphs called metagraphs from PPI-Keywords (PPIK) network which consists of biological annotations (Keywords) from Uniprot database [19] and PPI network (Figure 1.2a). Thus, proteins which are not neighbors in a noisy PPI network have a better chance to be topologically similar through keywords. Conducted experiments show that *Metagraph+* consistently improves disease protein prediction on three different PPI databases and outperforms the state-of-the-art baselines including both diffusion-based methods and module-based methods. In addition, predictions of *Metagraph+* exhibit better correlations with literature findings from PubMed database.
- *N2VKO*: Our second proposed framework, *N2VKO*, aims to provide representative feature vectors for proteins using automatically extracted features

from PPI network (Figure 1.2b). To the best of our knowledge, it is the first attempt of using node embeddings for disease gene prediction. We combine network-based feature representations of genes obtained by known node embedding method, *node2vec*, with biological properties of proteins. Then, we apply various feature selection methods to analyze their performance on disease gene prediction task. As the data for disease gene prediction is imbalanced, we further address the data imbalance issue by applying oversampling techniques on our novel representations to improve the prediction performance. Extensive experiments show that N2VKO significantly outperforms four state-of-the-art methods for disease gene prediction across seven diseases: Alzheimer’s disease, breast cancer, colorectal cancer, diabetes mellitus, lung cancer, obesity, and prostate cancer. Moreover, the categories of the biological properties of N2VKO representations are listed to analyze their role in disease formation. We also provided literature evidence for N2VKO biological features over lung cancer. Finally, the literature evidence from PubMed database reveals the effectiveness of our proposed N2VKO framework for disease gene prediction.

- *MICROS*: The third proposed framework is a novel unsupervised algorithm, “Multi-view network embedding with Intra-Cross” consistencies (MICROS), which can serve in a wide range of application area from node clustering and node classification to network construction and recommender system design. Since various biological networks provide different insights of genes, combining them as complementing each other in a unified framework is necessary to improve disease gene prediction performance. This approach learns low dimensional representations while preserving their topological characteristics and enabling views to work together and reinforce each other (Figure 1.2c). For disease gene identification problem we worked on two biological networks: PPI and Gene Ontology (GO) networks. To the best of our knowledge, this study is the first attempt of multi-view network embedding for disease gene

prediction. We also examine a novel form of higher-order collaboration that has not been explored previously on multi-view networks and further integrate it into a unifying framework of consistencies to provide more robust, superior node representations. We performed extensive experiments for not only disease gene prediction task but also link prediction task on real-world social network analysis. Our results demonstrate that our learned representations consistently outperform the state-of-the-art approaches on various downstream tasks namely node-level tasks, relationship mining and link prediction. Moreover, experiments show that MICROS is also capable of generating superior descriptive representations for disease gene clustering.

1.4 Outline

The PhD thesis is organized as follows:

Chapter 1 introduces the scope of this research in terms of disease type, data, and purpose. We explain our motivation of working on disease gene-association prediction problem using network data, followed by the challenges. Then, we summarize related prior work. Finally, we explain our major contributions and provide an outline of our work.

Chapter 2 provides a survey on the literature in terms of biological data and network-based methods including module-based, diffusion-based and representation-based algorithms.

Chapter 3 introduces representation-based model *Metagraph+* for disease gene prediction which initially constructs a novel PPIK network composed of biological properties of individual proteins and protein-protein interaction network. Then, the novel extracted heterogeneous subgraphs called metagraphs are used as representations for proteins to be fed into classifiers.

Chapter 4 introduces a framework on an integrated biological data with a PPI network, called *N2VKO*, which uses a network embedding model to learn low-dimensional and continuous vector representations for proteins through a neural network.

Chapter 5 proposes a novel multi-view embedding framework, *MICROS*, which presents comprehensive experimental results for not only disease gene prediction but also link prediction in social network analysis.

Finally, Chapter 6 concludes the thesis and presents possible future directions to follow.

Chapter 2

Literature Review

In the following sections first, we point out the importance and the advantage of using molecular networks in disease gene prediction problem. Second, we introduce widely used biological data types in this area such as GO annotations and existing strategies for the integration of them into the network-based models. Third, we provide a comprehensive survey on network-based computational models for disease-gene association prediction.

2.1 Disease and Network

Human diseases can be categorized into three different categories: pathogenic/infectious, deficiency, and genetic diseases. The scope of our research does not include the first two categories. Our study is based on genetic diseases, contrary to common belief, they might be hereditary or non-hereditary. Genetic diseases are mostly caused by a mutation. However, recent genome-wide human population studies have revealed that environmental factors, epigenetic changes, even pathogens (e.g., viruses, parasitic worms) play a role in disease-causing mutations [7, 78].

Disease-gene association is an arduous task for researchers. They must have detailed genetic and clinical information of several people with and without particular disease phenotypes. Genome-wide association study (GWAS) uses genetic information, which reveals the variation among humans, as known as single nucleotide polymorphisms (SNPs). Researchers aim to find out SNPs that correlate with specific diseases. Statistical analysis of associations, considering the huge number of possible disease mutations, takes time and cost. Early efforts for this task were based on recurred conditions in families, generation after generation. In recent years, researchers search for the evidence of a specific disease by observing common mutations in people with the disease. However, by doing so they may misevaluate the cases that the same mutation of interest occurs also in people without the disease [79], i.e., false-positives. The most important drawback of GWAS study is the large number of false-positives. Thus, researchers should search for more evidence such as the functional role of the gene in the disease or cross-check analysis through different sources is needed.

To provide the researchers with alternative low-cost disease-gene association evidence, computational approaches come into play. Note that computational approaches cannot be sufficient alone without the support of GWAS studies. The most important reason is the computational methods need to use known disease-gene associations derived by genome-wide studies as input to predict disease-causing candidates. Thus, these methods are considered as complementary tools in revealing potential disease-associated genes.

The most extensively used computational tools are molecular networks. The reason is unlike the GWAS studies which are based on the simple genotype-phenotype relationship, molecular networks are able to capture complex interplay among the proteins in a disease. The diseases are consequences of perturbations in the molecular networks caused by various complex factors such as genetic mutations, epigenetic factors, and pathogens. For instance, Figure 2.1 demonstrates two possible

FIGURE 2.1: Perturbations in molecular networks. The figure is adapted from the study [1].

perturbations: edge removal or node removal caused by a single mutated protein. These two perturbations may result in two different diseases. This illuminates the ability of molecular networks to model the effects of disease factors by taking into account the factors itself, as well as the affected neighbors. Therefore, molecular networks are efficient tools in the analysis of diseases by tracking the potential disruptions on the biological pathways due to the perturbations in the molecular networks [1].

2.2 Biological Data and Network Representation

Biological networks have illuminated our understanding of biological systems and diseases. A network (or a graph), can be represented as $G = (V, E)$, consists of a set of nodes, V , and set of edges. Network edges can be directed or undirected, weighted or unweighted based on the relationship between the nodes. For instance, edges in a PPI network are always undirected because they represent physical interactions between proteins, whereas the edges in a metabolic interaction network can

be either directed or undirected based on the chemical reaction between metabolites. Networks can also be designed to have different types of nodes such as genes and diseases to model the relations between the nodes. Disease-gene association is usually represented by bipartite networks [23]. Protein-protein interaction (PPI), phenotype, gene co-expression (Co-Ex), gene regulatory and metabolic interaction (MI) networks are widely used biological networks. Also, the integration of complementary biological concepts such as gene ontology and domain data into the biological networks has been widely examined [26, 80–82] through the years. We briefly introduce these biological concepts in this section.

Gene Ontology Data

Ontology data is an efficient, structured data to represent complex biological knowledge in a formal way [83]. It describes relationships between essential terms. The most popular ontology database is Gene Ontology (GO) [84]. It is a dynamic database which updates itself with new biological knowledge. The objective of the Gene Ontology database is to produce evidence-based annotations to reveal biological activities of genomic products (e.g., genes, proteins). There are three categories in GO: the molecular function, biological process, and cellular component. The molecular functions represent activities such as catalytic and binding without time and location information such as kinase activity. They may have subtypes, e.g., carbohydrate kinase activity and cyclin-dependent protein kinase activity. Biological processes such as growth, and homeostasis (i.e., regulation of the internal environment to maintain a constant state) are accomplished by one or more consecutive molecular functions. A cellular component represents the locations in terms of subcellular sections (e.g., mitochondrion) and macromolecular complexes (e.g., ribosome). Each of these three categories is represented by a directed acyclic graph (DAG) structure starting from a separate root ontology term and terms within each category may have more than one parent.

Protein Domain

Protein domains are subunits that define functional or structural aspects of proteins. A protein may have multiple domains and the same domain may occur in several proteins. The domains of proteins may change over time due to the evolutionary factors and they can keep their functional roles in different proteins. Pfam [85] and InterPro [86] are some of the most popular protein domain databases.

Protein-protein Interaction (PPI) Network

Protein molecules are the products of genes. They work in collaboration to perform various vital tasks. They have crucial duties within organisms, for example, enzymes are responsible of catalytic reactions or proteins such as insulin, carry out transmission of a signal across cells. To complete their tasks, they interact with each other in a transient way or stable way. Transient interactions carry out short-term actions such as binding of transcription factors. Stable interactions include the physically docking and forming a complex structure of proteins. The success of the interactions in a cell depends on several circumstances such as cell type, cell cycle phase, and state, developmental stage, environmental conditions, protein modifications (e.g., phosphorylation), etc. A protein-protein interaction (PPI) network represents these interactions as nodes are the proteins and edges (i.e., they are undirected) are the interactions. Since these interactions form the basis of biological processes, studying PPI networks is crucial for better understanding of human metabolism [87]. STRING [88], NCBI Entrez Gene[89], IntAct [90] are example PPI databases. However, PPI networks are incomplete; that is only a small percentage of all potential pairwise protein interactions are provided through the high-throughput experiments [18]. Thus, they can be modeled as weighted networks so that the trustworthiness of the interactions can be represented by the weights.

Phenotype Network

Since it is known that the human diseases occur due to the perturbations in the highly interlinked cellular networks [1, 91] (Figure 2.1), their relations can be

examined through a phenotype (i.e., disease is a phenotype) network. Phenotype networks consist of nodes representing the phenotypes and edges representing the molecular or phenotypic relationships between the phenotypes. Known gene-disease associations are obtained from databases such as OMIM database [92], then two diseases are linked if they share at least one common gene association [91].

Gene Co-expression Network

Gene co-expression network is an undirected graph which represents the correlations between genes (nodes) based on their expression levels across various samples, or experimental conditions. Since the genes with similar expression level in various conditions have the same transcriptional regularity scheme, by means of co-expression network revealing functionally similar genes is possible [93]. To compute the co-expression patterns, usually, the Pearson correlation coefficient (PCC) of the expression levels of all pairs of genes is computed. Two genes are coexpressed if their PCC is higher than a determined threshold value. There are several statistical approaches to specify the threshold [23].

Gene Regulatory Networks

In gene regulatory networks, usually, nodes represent a transcription factor or a DNA regulatory element. The edges are directed and represent the physical binding of transcription factors to the regulatory elements. The direction of the edges points the regularity element affected by the transcription factor [91]. Basically transcription factors bind to the promoter region and activate or repress the gene transcription which also affects the production of other proteins and mRNAs [94].

Protein Domain Co-occurrence Network

A protein domain network, consists of individual domains which are represented with nodes and their co-occurrences in a protein are represented with edges/links. They are used to study the function and interaction of proteins at the proteome level. It has been exploited to explore domain combinations which predominately

play a part in protein functions such as cell-cell contacts and signal transduction [95].

Metabolic interaction (MI) networks

Within a cell, a series of chemical reactions take place known as the biological pathway. Produced intermediates in the pathway are referred to as metabolites. Lactic acid, amino acids, alcohols, vitamins (B2 and B12) are some examples. Metabolic networks try to map potential biochemical reactions for a particular cell or organism. Most of the time, nodes represent biochemical metabolites and edges are either the reactions that transform one metabolite into another or the enzymes that catalyze these reactions [91]. The type of the edges whether directed or undirected depends on the reversibility of the reaction. In some cases, nodes may represent enzymes and edges may represent the product or substrate relationship between them. Studies [96–98] provide examples of human metabolic network construction methods.

2.3 Integrative Approaches

In this section, we introduce several network-based integration approaches in candidate disease-gene identification. Based on the outcome of the integration, we divide integrative approaches into two groups: single-view approaches and multi-view approaches. Single-view approaches take multiple networks as input $G_1, G_2, \dots, G_n$, where n is the number of networks and construct one single unified network (or a graph) as output to perform inference models on it. The constructed network can be represented as $G = (V, E)$, consists of a set of nodes, V , and set of edges, E . The edges of the constructed network can be directed or undirected, weighted or unweighted based on the relationship between the nodes and the nodes of the network can be homogeneous or heterogeneous. Whereas, the multi-view approaches

preserve the topological structure of each network and perform the inference models on the whole multi-view infrastructure.

2.3.1 Single-view Approaches

Based on the type of data they combine, single-view approaches can be grouped into two categories: homogeneous and heterogeneous integration [23].

2.3.1.1 Homogeneous Network Integration

In this integration different networks containing the same set of nodes (i.e., $G_i = (V, E_i), i \in 1, \dots, n$) are merged together to construct a single unified network which consists of all the edges from each network. Thus, adjacency matrix of the constructed network is $A = \sum_i^n A_i$, where A_i is the adjacency matrix of each network i (Figure 2.2). However, in weighted networks this type of integration may affect the clusters within each network. To prevent this kind of distortion in the network structure, edge weights are also considered. Thus, adjacency matrix of the constructed network becomes $A = \sum_i^n w_i A_i$, where A_i is the adjacency matrix and w_i is the weight of each network i as $w_i \geq 0$. For instance, MAXDRIVER [99] to identify potential cancer driver genes constructs a fused gene functional similarity network consists of four network data sources, including PPI, gene co-expression patterns, gene sequence similarities, and pathway co-occurrence relationships. It utilizes a multiple regression model to calculate a functional similarity score between a gene pair in the fused network by a heuristic filtration strategy to filter out noises that indicate low confidence relationships for gene pairs. Another study [100], for predicting a set of complementary prostate cancer genes, exploits PPI, genetic interaction (GI), gene regulatory interaction (GRI) and inferred protein-protein interaction from protein complexes (IPPI) to construct a cancer-specific gene network. It utilizes the Pearson Correlation Coefficient (PCC) of mRNA expressions

FIGURE 2.2: Homogeneous network integration.

to identify cancer-specific edges where two genes show different expression patterns between cancer and normal phenotypes. A recent study [101], to predict disease genes, integrates four different weighted networks: a PPI, two gene-function association networks and a gene association network, based on information entropy measure. In particular, the study defines the average uncertain degree of an edge between a gene pair by information entropy and use it to compute integration parameters.

2.3.1.2 Heterogeneous network integration

This category can be divided into two sub-categories based on the node types in the constructed network: *Projected heterogeneous network integration* (PHNI) and *heterogeneous network integration* (HNI). PHNI methods construct a single network by projecting used networks onto the one they are interested in [23], so that combined network consists of the same type of nodes. In Figure 2.3, the links of the resultant projected network is constructed based on the genes sharing the same disease association. For instance, the study IDN [102] to infer disease-disease

FIGURE 2.3: Projected heterogeneous network integration (PHNI). There are three different networks: gene-gene interaction, disease-disease similarity network and gene-disease association network (dashed orange lines). In the resultant network, thicker lines represent the stronger association.

associations, integrates different types of biological data, including disease-gene associations, disease-chemical associations, biological pathways and GO annotations into an Integrated Disease Network (IDN), a heterogeneous network where nodes are bio-entities such as genes, diseases, etc. and edges between the nodes represent their associations. They propose a similarity measure to construct disease-disease associations as a projection using IDN. Davis et. al. [103] constructs two different disease networks: phenotypic, which is based on the clinical history of 700,00 patients and genetic, which is based on a significant number of shared gene associations by pair of diseases. These two networks are integrated as an example of heterogeneous projections of multiple networks, i.e., disease-gene associations and disease-patient associations onto disease-disease associations. *Heterogeneous network integration* (HNI) methods, unlike PHNI methods, develop algorithms for heterogeneous networks consist of different type of nodes. For instance, CIPHER [43], to prioritize candidate genes for a given phenotype, constructs an integrated network composed of phenotype, gene-phenotype, and protein network (e.g., three networks demonstrated in Figure 2.3 excluding the projection network). Then, they score phenotype-gene pairs based on two profiles: phenotype similarity profile of the query phenotype and topological closeness profile of the gene. A recent study SLN-SRW (Simplified Laplacian Normalization-Supervised Random Walk) [104], to predict disease-related genes, constructs a heterogeneous network containing

both ontology terms and genes, based on eleven heterogeneous data sources including ontologies, networks, unified vocabularies, and relational databases. Another recent study [105] constructs a heterogeneous network, HANRD (Heterogeneous Association Network for Rare Diseases), and develops a graph convolution-based technique to infer new phenotype-gene associations from it. HerGePred [106] using four types of associations: disease-symptom, disease-gene, gene-GO term, and PPI network, constructs a heterogeneous disease-gene related network, then applies a well-known node embedding method [66] to obtain similarity scores of disease pairs and gene pairs for disease gene prediction.

2.3.2 Multi-view Approaches

Multi-view approaches, instead of projecting one network onto another or working on an integrated single network, they simultaneously explore the structure of each network in a unified framework. By this way, they preserve the topological structure of each network. There is an increased interest in multi-view network approaches in the literature [74–76, 107, 108]. However, most of the algorithms for disease gene predictions are based on single networks.

Most of the studies in this category are based on random walks and work on a multi-view system consists of two networks [23, 109, 110]. RWRH [3] is a network propagation method which is able to perform random-walk-and-restart (RWR) algorithm over a heterogeneous network (i.e., composed of phenotype, gene-phenotype, and protein network) through intra-subnetwork and inter-subnetwork transitions. A recent study RWRMH [40] generalizes RWRH to multiplex-heterogeneous networks and constructs three layers of physical and functional interactions of genes and proteins, a disease-disease network based on phenotype similarities, and gene-disease bipartite associations. Ni et. al. [111] utilizes tissue-specific networks in two ways: a network of networks (NoN) and networks of star networks (NoSN). Then, they

model a disease gene prioritization algorithm as an optimization problem based on the generalized version of network propagation methods to multiple networks. Another work [72] exploits neural representations through a multi-layer protein interaction network for multicellular function prediction.

A recent study, Medusa [112], applies the collective matrix factorization method to capture semantic relations among different views. Then, it designs a submodular optimization model to derive significant associations.

2.4 Network-Based Techniques in Disease Gene Prediction

Network-based techniques in disease gene prediction utilize a network (or a graph) as input of the model. The network can be represented as $G = (V, E)$, consists of a set of nodes, V , and set of edges E . The edges can be directed or undirected, weighted or unweighted based on the relationship between the nodes. The output of these techniques is a list of candidate disease-causing/disease-associated genes.

2.4.1 Module-based Methods

Module-based methods are based on the notion of proteins within the same topological or functional module in a network tend to be associated with the same disease [16, 27, 45, 46]. Their aim is to capture a subnetwork that contains most of the disease-associated molecules (i.e., proteins/genes) in the interactome (i.e., interaction network) [29]. For this aim, mostly the protein interaction networks are exploited based on the functional and/or topological similarities of the proteins.

In cases where the detected modules contain a large number of proteins, to attain more accurate predictions, pathway identification can be an option. The basic

FIGURE 2.4: Pathway identification. Two disease module members, protein *A*, protein *B* and predicted disease protein *C*, which falls within the shortest path between *A* and *B*.

approach addressing this problem is the assumption of the shortest path between known disease proteins forms a specific molecular pathway, therefore a disruption on these pathways may induce a disease phenotype [29]. In Figure 2.4, a pathway identification example is shown. Validating disease associations between the protein *C* and disease proteins could be tedious. Instead, by means of pathway assumption, predicting *C* as a disease protein is possible because *C* falls within the shortest path between two disease proteins: *A* and *B* in the module.

Disease module identification algorithms are mostly based on solely neighborhood considerations such as direct neighbors, shared neighborhood, and local shortest paths. This causes them to be more vulnerable in sticking to the local evaluation. Thus, the support of additional biological knowledge such as phenotype similarities and functional information may increase their prediction performance. However, one advantage of direct neighbor-based methods is they rarely require parameter tuning stage. For example, interconnectedness (ICN) [2] is a method for prioritizing candidate disease genes, which computes the proximity of each candidate protein to known disease genes by means of the direct link and shared neighbors without any parameter tuning. However, its performance is not better than the traditional diffusion-based methods, RWR [5] and PRINCE [38]. Whereas, combining ICN with RWR and PRINCE, results in higher performance. Another parameter-free approach [113], VS, differently than ICN, follows “extended guilt

FIGURE 2.5: Unweighted PPI network to describe ICN [2] method. The shortest path between proteins A and D includes nodes B and C.

by association” so that it is possible to assign a similarity score between proteins based on the shortest path between them. In Figure 2.5, based on ICN, the similarity score between node A and D is zero because they do not have any direct link or shared neighbors between them. Whereas, the study [113] computes the similarity between these nodes based on the shortest path between them considering the similarity scores between pairs: $A-B$, $B-C$, and $C-D$. Consequently, this approach outperforms ICN.

Lage et al. [114] propose one of the first endeavors in the computational integration of phenotype similarity scores to a PPI network in identifying the disease-causing genes. Their framework first identifies protein complexes for each candidate disease genes, then using a phenotype-scoring and a clinical text-mining scoring employs a Bayesian predictor. Wu et al. [43] propose a method called CIPHER. Differently, than Lage et al., they construct an integrated network consisting of phenotype, gene-phenotype, and protein network. The similarity profile between query phenotype and all phenotypes is derived from the phenotype network. And the closeness profile of the candidate gene to all known disease genes is computed by means of topological distances, then grouped according to the phenotypes. Finally, with the use of linear regression, they compute the correlation between these two profiles and ranked the candidate protein based on the correlation score. According to study [3], CIPHER performs better than Lage et al. [114] on overlapped data. Another study [115], merges integrative data approach with a module-based method,

which determines protein modules directly from topological modules using a network topological algorithm.

Protein function generally describes the biochemical, cellular, and intercellular activities of a protein [116]. Module-based methods are also extensively used in identifying functional modules based on the assumption that interacting proteins have a higher chance of sharing the same function than any two proteins that are not interacting with each other. Snel et al. [117] integrate genomic associations into a protein interaction network and use properties of the network to detect functional modules.

The dense subnetworks in an interactome can be considered as functional modules. Due to the conceptual similarity between identifying functional modules and detecting clusters in a network, clustering algorithms are largely used in the detection of functional modules. Since functional modules mostly do overlap, most of the endeavors applied in this area are designed to detect overlapping clusters [49]. Wang et al. [118] propose a fast hierarchical clustering algorithm, HC-PIN, on yeast protein interaction network to identify functional modules of proteins.

A graph clustering algorithm, MCL [119], a widely used method for detection of the modular structures in a network [49, 120], which is based on increasing the contrast between the regions of high and low flow. For this aim, it simulates random walks by performing successive expansion (matrix squaring) and inflation (Hadamard power followed by a scaling) operations on the adjacency matrix. The expansion operator is responsible for assigning new probability values for all pairs of nodes, whereas the inflation operator alters the probabilities by boosting the intra-cluster interactions and demoting inter-cluster walks [25, 121].

Adamcsek et al. [122] design a tool called CFinder to identify functional modules in PPI networks. It exploits Clique Percolation Method [123] to detect the k -clique percolation clusters as functional modules. These clusters are formed as a result

of linking adjacent k -cliques (i.e., two k -cliques are adjacent if they share $(k - 1)$ common nodes) [25]. Detection of k -cores is also used in bioinformatics domain. The k -core of a network is defined as the largest subnetwork which every node has at least k degree. Wuchty et al., propose a pioneering work which exploits k -cores [95] to explore protein domain role in the functional activities of proteins. For this aim, they construct a domain co-occurrence network using Pfam database [85].

Another module-based instrument, which plays a key role in biological network analysis is *network motifs* [124, 125]. The concept has been discovered by Milo et al. [124] based on the fact that the real-world networks from different domains share common small topological structures more frequently than expected by chance. P-Finder tool [126] exploits network motifs to rebuild signaling networks for proteins. A signaling process might rely on distinctive flow patterns. The motivation of the proposed network motif-based (NM) algorithm is that the signaling patterns of proteins can be unveiled by mapping the network motifs extracted from a PPI network. In particular, the source and target nodes (proteins) need to be fed into the algorithm. As a two-step approach, first, all network motifs in the genome-wide PPI network are extracted by using a publicly available network motif detection tool. Second, starting from the source node, method iteratively chooses following edges in a greedy manner until arrival at the target. The edges which are members of many distinct network motifs are chosen at iteration.

Module-based approaches are used for protein function prediction problem as well. Cho and Zhang [50] for function prediction of proteins, explore frequent patterns of functional associations in a protein interaction network. They assign labels to proteins based on the set of functions of that protein. Labeled subnetwork represents a functional association pattern. Then, based on the assumption of downward closure of frequency, they performed the a priori algorithm [127] which detects frequent functional association patterns.

Another similar concept in protein function prediction problem is identifying protein complexes in a PPI network. The difference between these two concepts are to construct a functional module of the proteins without time and place constraint for the interaction of the proteins, whereas to form a protein complex, proteins should interact with each other at the same time and place.

Let us introduce some of the algorithms in protein complex detection. Bader et al. [128] describe graph-based clustering algorithm, “Molecular Complex Detection” (MCODE) to identify molecular complexes. The algorithm detects densely connected regions in large PPI networks. It basically applies vertex weighting using the local neighborhood density and traversal from a locally dense seed protein.

The study [65] proposes a novel graph-theoretic method for uncovering the protein functions based on *graphlets* on a PPI network. By exploiting local structures of node neighborhoods, it concludes that the biological function of a protein and its local network structure are closely related. The proposed method, expresses the local topology of a protein with the vector of *graphlet degrees*, namely “the signature of the protein”, and computes the similarities between them for all protein pairs. In particular, the proposed new measure expresses the topology of a node with a vector of *graphlet degrees* by counting the number of graphlets that the node contacts. Graphlets are non-isomorphic induced subgraphs of a large network. In other words, they are the subgraphs containing all edges of its nodes from the original network. Figure 2.6 displays 2, 3, and 4-node graphlets. The study considers up to five nodes and 29 graphlets in total.

Amin et al. [129] propose a clustering algorithm for the identification of protein complexes called DPClus. It is based on keeping track of the periphery of a detected cluster. Initially, it assigns weights to each edge computing the common neighbors between its two proteins and further, it assigns weights to each node by summing the weights of each edge connected to the node. In the cluster formation stage, it forms a protein complex starting from selecting the node with the highest weight

FIGURE 2.6: 2-, 3-, and 4-node graphlets.

(seed node) as the initial cluster. Then, iteratively enlarges the cluster by adding nodes/proteins according to their relevance priority.

Wu et al. [47] present a core-attachment based method (COACH) for detection of protein complexes. The first stage of their method detects protein-complex cores. Then it includes attachments into the extracted cores to form protein complexes. Zhao et al. [130] detect protein complexes based on the uncertain graph model DCU. They use both the density degree of a subgraph with expected density concept and the relationship between a protein and a subgraph in a PPI network with relative degree concept. Yang et al. [131] integrate PPI networks with the PPI data from the biomedical literature for protein complex detection. In order to obtain PPI data from the biomedical literature, a natural language processing system, PPIExtractor, is applied. Peng et al. [26] propose two methods on enriched protein-protein interaction (PPI) network with protein domain for the function prediction of the proteins. The first method is based on domain combination similarity (DCS), which consists of the domain similarity of both proteins and their neighbors. The other method adds the context of protein complexes (DSCP) to include functional similarities as well.

The study [48], RNSC, proposes a two-step approach to detect protein complexes

on the PPI network. First, they cluster the network using the proposed restricted neighborhood search algorithm which partitions the network into clusters using a cost function. Second, they filter the results according to cluster size, density, and functional homogeneity. Finally, the clusters exhibiting common properties with known biological complexes are preserved.

Qi et al. [132] propose a supervised graph clustering approach for the identification of protein complexes. It learns critical topological and biological properties of known protein complexes by developing a supervised strategy without any prior knowledge on the topological properties of protein complexes. Consequently, by using the learned representations it aims to detect new complexes in the PPI networks.

In order to discriminate functional modules and protein complexes, subcellular localization data and expression profile data can be integrated to the PPI network [133, 134]. The summary table for the module-based methods is presented in Table 2.1. In this summary, in addition to their *technique*, *view*, *integration*, *network* types and release *years* we examine their different properties such as *complexity*: 1) scalability to large graphs (linear time complexity) and/or 2) polynomial time complexity, *formulation*: 1) score-based with many hand-crafted features and lots of intermediary calculations or 2) algorithm-based (automated), and *solution*: 1) problem specific or 2) applicable to other problems from different domains (generic).

2.4.2 Diffusion-based Methods

Diffusion-based methods are mostly used in the analysis of the biological pathways especially the pathways that are closest to the known disease genes. They consider known disease proteins as seeds and diffuse along the interactome through random walks starting from the seeds [29]. Since they consider the full network topology

TABLE 2.1: Module-based models surveyed in this thesis. Columns present the name, main technique(s), view types (i.e., single or multi-view), integration type (i.e., none (-), homogeneous or heterogeneous), network(s), complexity (i.e., linear (L) or polynomial (P)), formulation (i.e., hand-crafted (H) or automated (A)), solution (i.e., problem-specific (P) or generic (G)) and release years of the studies.

Algorithm	Tech.	View	Int.	Network(s)	(Complex., Formula., Solution)	Year
Snel et al. [117]	Clustering	Single	-	PPI	(P, H, P)	2002
MCODE [128]	Clustering	Single	-	PPI	(P, A, G)	2003
RNSC [48]	Cost-based clustering	Single	-	PPI	(P, H, P)	2004
Wuchty et al. [95]	k-core decomposition	Single	-	Domain co-occurrence, domain interaction	(P, H, P)	2005
CFinder [122]	Clique percolation	Single	-	PPI	(P, A, G)	2006
DPCLUS [129]	Clustering	Single	-	PPI	(P, A, G)	2006
Lage et al. [114]	Bayesian predictor	Single	Homo.	PPIs	(P, H, P)	2007
CIPHER [43]	Regression model	Single	Hete. (HNI)	Phenotype, gene-phenotype, PPI	(P, H, P)	2008
Milenković and Pržulj [65]	Modular features	Single	-	PPI	(P, H, G)	2008
Qi et al. [132]	Supervised Bayesian network (BN) model	Single	-	PPI	(P, H, G)	2008
COACH [47]	Core-attachment	Single	-	PPI	(P, A, G)	2009
Cho and Zhang [50]	A priori algorithm-based subgraph mining	Single	-	PPI	(L, A, G)	2010
ICN [2]	Parameter-free scoring	Single	Homo.	PPIs or functional association network	(P, H, G)	2011
VS [113]	Parameter-free scoring	Single	-	PPI	(P, H, P)	2012
Shih et al. [49]	Markov clustering	Single	-	PPI	(P, A, G)	2012
DCU [130]	Core-attachment	Single	-	PPI	(P, A, G)	2014
Yang et al. [131]	NLP, clustering	Single	Homo.	PPIs	(P, H, P)	2014
Peng et al. [26]	Similarity scoring	Single	-	PPI	(P, H, P)	2014
Liu et al. [115]	Clustering	Single	Hete. (HNI)	Function, disease, PPI	(L, A, G)	2015
P-Finder [126]	Network motif	Single	-	PPI	(L, A, G)	2015

together with the placement of the known disease genes, they are dominant over module-based methods [27, 29, 56, 57]. Figure 2.7 describes the overall functioning of the diffusion-based methods. The proteins, which are placed farther from the known disease proteins, are scored with a lower likelihood of being associated with the disease. Thus, diffusion-based methods are very suitable for prioritizing

FIGURE 2.7: Diffusion-based Method.

candidate disease proteins and widely used for this aim.

Random Walk with Restart [5] is the most extensively used algorithm as a basis of many studies. It is based on the iterative transitions of a walker starting from the seed nodes (i.e., known disease-associated nodes) to a randomly selected neighbor in the interactome. Differently than a basic random walk, it is able to jump back to any seed node with probability r at each iteration. Formally, it is defined as,

$$p^{t+1} = (1 - r)Wp^t + rp^0 \quad (2.1)$$

where W is the column-wise normalized adjacency matrix of the interactome and p_t is a vector keeping the probability of being at node i at time step t in its i -th entry. p^0 is the initial probability vector holding the probabilities of being at known disease nodes (seeds). Every member of the disease has uniformly distributed probabilities as the sum of the probabilities equal to 1. In other words, random walker commences from each of the known disease proteins/genes with equal probability. Thus, candidate proteins/genes are prioritized according to the values in the steady-state probability vector p^∞ . The number of iterations, N , is determined by the condition which satisfies the L1 norm difference between p^t and p^{t+1} smaller than 10^{-6} .

Vanunu et al. propose a method called PRINCE (PRIoritization and Complex Elucidation) [38] which basically generalizes the RWR by adapting a weight function for the PPI network and also taking into account the disease similarities. On an unweighted PPI network, this method basically performs random walk with restart with (Eq. 2.1) one major difference than RWR which is prior probabilities. In RWR the initial probability vector consists of equal probabilities uniformly distributed between known disease-associated proteins. Whereas, in PRINCE prior probabilities are assigned to each disease-associated-protein based on a logistic function: $L(x) = \frac{1}{1 + e^{(cx+d)}}$ and $x = S(q, p)$ where $S(q, p)$ is the similarity score between query disease q and the associated disease p with the protein x . If the protein is associated with more than one disease then, p is selected to be the most similar one to q . The study uses similarity scores of phenotypes obtained from MimMiner [135].

Li and Patra [3] propose the aforementioned multi-view approach in Section 2.3.2, called RWRH (random walk with restart on the heterogeneous network) to prioritize proteins based on their relevance to disease proteins. They combine the protein network and phenotype network based on the protein-phenotype associations. Figure 2.8 illustrates the constructed a heterogeneous network. The dashed lines in the Figure represent the associations with the disease phenotypes (labeled as ph). The other lines represent either the interactions between proteins or similarities among phenotypes. Li and Patra adapt the random walk with restart (RWR) algorithm to the constructed heterogeneous network over a certain disease phenotype, using their associated disease proteins as seed nodes. A recent study RWRMH [40] generalizes RWRH to a multiplex-heterogeneous network containing three layers of physical and functional interactions of genes and proteins, a disease-disease network based on phenotype similarities, and gene-disease bipartite associations.

Random walks have been widely used in the literature to identify or prioritize potential disease genes. Liekens et al. [136] provide a data integration and data

FIGURE 2.8: Heterogeneous network of study RWRH [3].

mining platform, called BioGraph, based on stochastic random walks. Erten et al. [137] propose VAVIEN which uses random walk proximity as a feature (not core algorithm just feature) to estimate the topological similarity between seed and candidate disease proteins and merges these features with Pearson correlation coefficient measure to capture the topological similarities. ORIENT (neighbor-favoring weight reinforcement) [138] is a method to improve the performance of RWR by means of a proper increase of the weights of interactions close to the known disease genes. In the study of Zhu et al. [56] diffusion profiles of a gene and phenotype are obtained separately based on RWR. Then, candidate disease genes are prioritized by comparing these profiles by the linear correlation coefficient (DP-LCC) and cosine of the angle (DP-COS) between them. Yang et al. [139] propose a random walk-based method called RWPCN (Random Walker on Protein Complex Network) to predict and prioritize disease genes. In the study, initially, an integrated network including human protein complexes, protein interaction network, and phenotypes is constructed. Then, RWR is performed on the constructed scheme.

Chen and Xu [140] propose a network-based approach to predict malaria-associated genes. They construct a network by integrating human-human, parasite-parasite and human-parasite protein interactions. Then, they perform the random walk

algorithm on the network by setting malaria genes as the seeds. Wu et al. [141] based on the assumption that the disease proteins are not well connected with the essential proteins; they introduce a novel approach to gene prioritization on PPI networks assigning positive flow to disease genes and negative flow to essential genes. Then, perform network propagation for gene prioritization.

Xie et al. [142] propose bi-random walk (BiRW) algorithm to study the relationship between the network information and phenotype-gene associations by analyzing the circular bigraphs (CBGs) in OMIM human disease phenotype-gene association network and MGI mouse phenotype-gene association network. The proposed algorithm detects the CBG patterns in the networks for revealing human and mouse phenome-genome association. BiRW is applied on gene interaction network and phenotype similarity network at the same time to obtain gene paths and phenotype paths in CBGs for the association predictions.

In a recent study SLN-SRW (Simplified Laplacian Normalization-Supervised Random Walk), Peng et al. [143] aim to integrate huge amounts of omics data into a network for disease gene prediction. They propose a disease gene prediction method, to produce and model the edge weights of the constructed biomedical network that integrates biomedical data from heterogeneous sources. The summary table for the diffusion-based methods is presented in Table 2.2. The summary table contains the same aspects with Table 2.1 described at the end of the previous section.

2.4.3 Representation-based Methods

Representation-based methods for disease gene prediction can be grouped into four categories: feature-based methods (e.g., the representations are derived from

TABLE 2.2: Diffusion-based models surveyed in this thesis. Columns present the name, main technique(s), view types (i.e., single or multi-view), integration type (i.e., none (-), homogeneous or heterogeneous), network(s), complexity (i.e., linear (L) or polynomial (P)), formulation (i.e., hand-crafted (H) or automated (A)), solution (i.e., problem-specific (P) or generic (G)) and release years of the studies.

Algorithm	Tech.	View	Int.	Network(s)	(Complex., Formula., Solution)	Year
PRINCE [38]	Network propagation	Single	-	PPI	(L*, A, G)	2010
RWRH [3]	RWR	Multi-view	Hete.	Phenotype, PPI, gene-phenotype	(P, A, G)	2010
BioGraph [136]	Stochastic RWs	Single	Hete. (HNI)	21 publicly available curated databases	(P, A, G)	2011
VAVIEN [137]	Random walk and correlation	Single	-	PPI	(P, A, G)	2011
RWPCN [139]	RWRH	Multi-view	Hete.	PPI, phenotype, protein complex network	(P, A, G)	2011
ORIENT [138]	RWR	Single	-	PPI	(P, A, G)	2013
Zhu et al. [56]	RWR-based diffusion profiles	Single	-	PPI, phenotype, gene-phenotype	(P, A, G)	2013
Chen and Xu [140]	RWRH	Multi-view	Hete.	Cross-species genetic networks	(P, A, G)	2015
Wu et al. [141]	Network propagation	Single	-	PPI	(P, A, G)	2015
BiRW [142]	RWR	Multi-view	Hete.	PPI, phenotype, gene-phenotype	(P, A, G)	2015
Ni et. al. [111]	Network propagation-based optimization	Multi-view	Hete. (PHNI)	Disease similarity, disease-tissue mapping, tissue-specific PPIs, tissue-specific gene co-expression	(P, A, G)	2016
SLN-SRW [143]	Supervised RW	Multi-view	Hete.	Gene, disease, phenotype ontologies, PPI	(P, A, G)	2017
RWRMH [40]	RWRH	Multi-view	Hete.	Physical and functional interactions, phenotype, gene-phenotype	(P, A, G)	2018

* RWR is performed on the fly, i.e., solving random walk with restart as an iterative method till the convergence [53]. In other cases, its time complexity is polynomial [144].

module-based, diffusion-based or various integrative/fusion approaches), nonnegative matrix factorization (NMF), positive-unlabeled learning, and network embedding. The common characteristic of representation-based methods is that the extracted representations are fed into a downstream classifier for the prediction task. Most of the representation-based endeavors in disease gene prediction are supervised or semi-supervised models. Note that, in this category, we also cover

the methods which do not require an automated learning of the representations/features.

For the last decade, feature-based methods, as they integrate various biological concepts, have been extensively studied. For instance, Xiong et al. [145] propose a new method with two phases for protein function prediction problem. First, the algorithm combines PPI information and protein sequence information. Then, a collective classification algorithm, Gibbs sampling with bootstrapping and iterative classification, is employed on the constructed network to predict protein-function associations. Chen et al. [146] combine PPI datasets with biological features of the protein complex, human gene expression profiles, and pathway data based on the theory of Markov random field (MRF), then perform Bayesian analysis to predict human disease genes. Chen et al. [147] propose a method called BRIDGE which integrates disease phenotypic similarities with protein-protein interactions, gene sequence similarities, gene expression patterns, gene ontology annotations, and gene pathway memberships. For prioritization of the candidate disease genes, they perform multiple regression models with lasso penalty to automatically weight these different data sources. Nitsch et al. [148] compute the differential expression level of each gene using the gene expression data and combine this measure with the PPI network. Then, they apply three diffusion-based kernels to rank candidate disease genes, which are kernel ridge regression, heat kernel, and Arnoldi kernel approximation.

NMF methods such as inductive matrix completion [58], probability-based collaborative filtering [59] and manifold learning [60] are based on learning latent factors of gene-disease association matrix. Their motive is to predict disease genes even for a disease that has no known associated genes. Natarajan and Dhillon propose IMC [58], which combines a wide range of features including microarray expression data, disease-related textual data and gene network. They assume the constructed

gene-disease association matrix is low rank. Then, they solve an optimization problem to learn the latent factors of the association matrix, that satisfy the observed gene-disease associations, so that the model can be generalized to diseases that do not have any known association. Zeng et al. [59] propose a probability-based collaborative filtering model (PCFM), which uses alternating least squares and stochastic gradient descent method to learn the latent factors. In addition, they include two regularizations in the objective function based on the similarities between human genes and their neighbor genes. Luo et al. [60], for predicting disease-gene associations, propose a manifold learning method which is based on the assumption that the geodesic distance between a disease and the associated genes should be shorter than the other genes. The learned geodesic distances via NMF technique are used to prioritize disease-gene pairs. Another study Medusa [112], builds a collective matrix factorization model for disease-gene association prediction, which aims to work on large-scale collections of heterogeneous data but at the same time to preserve their individual diverse semantics. However, NMF-based methods generally do not provide global optimal solutions and have difficulty in convergence even for local optimal solutions [63].

Supervised techniques may struggle in the identification of candidate disease genes, as known disease-gene associations provide evidence only for positive samples and the negative samples remain unlabeled. Therefore, several studies work on positive-unlabeled (PU) strategies. For example, Mordelet and Vert [61] propose a method called ProDiGe. Apart from traditional disease gene identification techniques, it computes the relative similarity of both known and candidate disease genes to score and prioritize all candidates. It works on an integrated platform consists of sequence features, expression levels of the genes in different conditions and PPI network. Similar to the other PU learning models, ProDiGe also adopts the biased support vector machine (SVM) classifier, which assigns higher weights to the

positive examples, as they stand for high-confidence samples. However, it has similar performance with PRINCE [38] (i.e, a random walk-based method) if they are compared under fair conditions such as solely on a PPI network without any integration. Singh-Blom et al. [42] propose a method called CATAPULT, which is based on PU learning using the biased SVM for disease-gene association prediction. They use walks on the heterogeneous network including PPI network as the feature set and differently from diffusion-based methods such as PRINCE [38] and RWRH [3] the length of the walks is limited. The method is basically a generalization of Katz measure to a heterogeneous network to learn appropriate feature weights. Yang et al. [62] propose PUDI, apart from other PU learning techniques, they divide unlabeled training genes into multiple sets, then perform a multi-level classification over the partitioned sets using biased SVMs.

Recently, representation learning through deep learning (DL) [149–151] has exhibited an increased demand not only in social network analysis but also in bioinformatics. We can mainly categorize DL-based node embedding techniques into two classes: *network embedding methods with random walks* and *network embedding methods without random walks* [149, 151]. The first category aims to learn low-dimensional and continuous vector representations for nodes through a neural network by considering both node itself and its neighborhood. For this purpose, Skip-Gram architecture [70] is extensively used to build models which are capable of constructing associations between the node and its neighborhood. The neighborhood of the nodes is sampled through the random walks. The second category is based on multi-layered learning architecture. Auto-encoders [152–155] and deep neural networks e.g., convolutional neural networks (CNN) [156, 157] are in this category. However, since these techniques should transform graph structure into a grid-like domain, they have higher computation cost than random walk-used methods.

Network embedding methods with random walks [66, 68, 69] such as DeepWalk [68]

and generalized version of DeepWalk in terms of random walks, node2vec [66], generate node representations such that the nodes within the short random walk distance have similar embeddings [71]. A recent work OhmNet [72] exploits neural representations through a multi-layer protein interaction network for multicellular function prediction. Another study [73] combines graphlet representations with neural embeddings for disease pathway analysis. In this thesis, we present our study which is to the best of our knowledge, the first attempt of integrating node embeddings and biological annotations in disease-gene association prediction (Chapter 4).

In addition, recently researchers have taken a keen interest in multi-view embedding or multiplex network embedding methods [74–76] especially for social data analysis. They are capable of learning the embeddings collaboratively through the views as well as preserving the individual topological structure of each view. In this thesis, we present our recent multi-view embedding model which is able to predict candidate disease genes. The summary table for the representation-based models is presented in Table 2.3. It contains the same aspects with Table 2.1 and Table 2.2 described at the end of the previous two sections.

2.5 Summary

This chapter presents a literature review of the network-based computational approaches on disease gene prediction problem in terms of their data representation design (i.e., single-view or multi-view), data type (i.e., heterogeneous, homogeneous) and problem formulation (i.e., module-based, diffusion-based and representation-based).

Diffusion-based methods demonstrate superiority over module-based approaches from two aspects: i) most of the module-based techniques need parameter tuning, ii) more capable of traversing the whole network topology rather than sticking

TABLE 2.3: Representation-based models surveyed in this thesis. Columns present the name, main technique(s), view types (i.e., single or multi-view), integration type (i.e., none (-), homogeneous or heterogeneous), network(s), complexity (i.e., linear (L) or polynomial (P)), formulation (i.e., hand-crafted (H) or automated (A)), solution (i.e., problem-specific (P) or generic (G)) and release years of the studies.

Algorithm	Tech.	View	Int.	Network(s)	(Complex., Formula., Solution)	Year
Nitsch et al. [148]	Diffusion-based kernels	Single	-	PPI or functional associations	(P , H , G)	2010
ProDiGe [61]	PU learning, biased SVM	Single	-	PPI	(P , A , G)	2011
PUDI [62]	PU learning, multi-level classification with biased SVM	Single	-	PPI	(P , H , P)	2012
Xiong et al. [145]	Gibbs sampling based collective classification	Single	Homo.	PPI, implicit edges from protein sequence similarity	(P , A , G)	2013
BRIDGE [147]	Multiple regression model with lasso penalty	Single	-	PPI, gene sequence similarities, gene expression patterns, gene ontology annotations, and pathway	(P , H , P)	2013
CATAPULT [42]	PU learning, biased SVM	Single	Hete. (HNI)	Gene-phenotype (human and non-human), PPI	(P , A , G)	2013
Chen et al. [146]	Markov random field (MRF), Bayesian analysis	Single	-	Gene-phenotype, protein complexes, PPIs, pathways and gene expression profiles	(P , A , G)	2014
IMC [58]	Inductive matrix completion	Single	-	Orthologous phenotype, PPI, phenotype	(P , A , G)	2014
Medusa [112]	Collective matrix factorization, submodular optimization	Multi-view	Hete.	Genes, chemicals, pathways, GO terms, etc.	(P , A , G)	2016
PCFM [59]	Probability-based collaborative filtering model	Single	Hete. (HNI)	PPI and functional associations	(P , A , G)	2017
OhmNet [72]	Neural representation	Multi-view	Homo.	Tissue-specific interactome	(L , A , G)	2017
Luo et al. [60]	Manifold learning	Single	-	Gene-phenotype	(P , A , G)	2018
Agrawal et al. [73]	Neural representation, graphlet	Single	-	PPI	(P , A , G)	2018
HANRD [105]	Graph convolution-based	Single	Hete. (HNI)	Curated, ontological associations involving genes, diseases, phenotypes and pathways	(L , A , G)	2018
HerGePred [106]	RW-based embedding	Multi-view	Hete.	Disease-symptom, disease-gene, gene-GO term, PPI	(L* , A , G)	2019

* RWR is performed on the fly, i.e., solving random walk with restart as an iterative method till the convergence [53]. In other cases, its time complexity is polynomial [144].

to the local neighborhood. As a result, they have been widely used in disease gene prediction and especially in disease-gene prioritization. However, they are prioritizing disease genes based on the assumption that disease genes should be located near each other. This assumption may not hold for all diseases.

Among the module-based approaches network motifs and graphlet kernels can be the most competent ones to make a structurally global inference. However, they are still weak if the network is noisy and incomplete as a PPI network, therefore they need to exploit additional biological data. To address this issue, in Chapter 3, we introduce our Metagraph representations which are capable of working on a more robust PPIK network (i.e., consists of both PPI network and biological annotations called keywords). Our proposed PPIK heterogeneous network also enables proteins which are functionally similar but locationally distant in the original PPI network to be closer to each other, so that it can perform on the network more globally compared to other module-based methods.

NMF methods generally do not provide global optimal solutions and have difficulty in convergence even for local optimal solutions [63], PU learning methods are usually based on weighting the positive samples and unlabeled samples differently during the prediction, as a result, tuning is inevitable, which is computationally expensive [158]. Multi-layered learning architectures including auto-encoders and convolutional neural networks (CNN) are powerful tools, however, since these techniques should transform graph structure into a grid-like domain, they have higher computation cost than Skip-gram-based methods. In addition, they could be too complex models resulting in over-fitting for diseases with a small number of known disease genes. To eliminate these problems, in Chapter 4, we propose a novel N2VKO framework, which combines Skip-gram embeddings with biological annotations for disease gene prediction.

In addition, multi-view embedding or multiplex network embedding methods [74–76] exhibit increased popularity especially in social data analysis. They are able to

provide more robust node embeddings, as learning the embeddings collaboratively through the views as well as preserving the individual topological structure and semantics of each view. In this thesis, we present our recent, to the best of our knowledge first attempt, multi-view embedding model *MICROS*, for prediction of candidate disease genes in Chapter 5. The experimental results demonstrate that *MICROS* significantly outperforms the state-of-the-art methods on a wide range of supervised and unsupervised evaluation tasks, including node classification, visualization, and clustering, as well as relationship mining and link prediction.

Chapter 3

Disease Gene Classification with Metagraph Representations

Protein-protein interaction (PPI) networks are important tools in studying the functional roles of proteins, including their association with diseases. However, they are often noisy, and thus need complementary biological knowledge of proteins such as molecular functions and biological processes. To complement and enrich PPI networks, in this work we integrate *keywords* describing protein properties into the PPI network, and construct a novel *PPI-Keywords* (PPIK) network consisting of both proteins and keywords as two different types of nodes. Since disease proteins tend to have similar topological properties on the PPIK network, we further propose to represent proteins with *metagraphs*. Different from a traditional network motif or subgraph, a metagraph can capture a particular topological arrangement involving the interactions/associations between both proteins and keywords. Based on the novel metagraph representations for proteins, we further build classifiers for disease protein classification through supervised learning. Our experiments on three different PPI databases demonstrate that the proposed method consistently improves disease protein prediction across various classifiers compared to baselines including the diffusion-based methods (e.g., RWR) and the

module-based methods. Lastly, our predictions also turn out to exhibit better correlations with literature findings from PubMed.

3.1 Introduction

To understand the underlying mechanism of diseases, *protein-protein interaction* (PPI) networks have been widely exploited as lighting upon the collaboration of proteins. This is because several studies [16, 17] have demonstrated that the locality of proteins in a PPI network is not random. Rather, proteins with the same phenotype or function tend to exhibit common topological characteristics in a PPI network, including the degree, coreness, and closeness. Given such characteristics, PPI networks could be instrumental towards predicting the associations between proteins and diseases [30].

Most of the PPI-based methods do not consider the properties of proteins themselves, such as their Gene Ontology (GO) annotations like biological processes, molecular functions, cellular components, etc. For example, it is known that only the proteins which are localized at the same subcellular compartments can interact with each other [159, 160]. Thus, we propose to use keywords from the Universal Protein Resource (UniProt) database [19] to enrich the PPI network. The keywords cover various biological aspects of the proteins, as summarized in Table 3.1. A previous study [161] reveals the relationship between these keywords and intrinsic disorders: some keywords for cellular components, domains, technical terms, developmental processes, and coding sequence diversities indicate strong positive or negative correlations with long intrinsically disordered regions. Furthermore, it is known that intrinsically disordered proteins are associated with many diseases [162]. Additionally, several investigations [163–165] consider the role of post-translational modifications (PTM) in disease and functional complexes. Based on these findings, we integrate the *keywords* of Uniprot database directly into the PPI

TABLE 3.1: A summary of keywords from the UniProt database.

Keyword Category	Examples
Biological Process	<i>Apoptosis, Cell cycle, cAMP biosynthesis</i>
Cellular component	<i>Golgi apparatus, Vacuole, Cytoplasm</i>
Coding sequence diversity	<i>Polymorphisms, RNA-editing, alternative splicing</i>
Domain	<i>SH2 domain, Kelch repeat, Transmembrane</i>
Ligand	<i>cAMP, S-adenosyl-L-methionine, cGMP</i>
Molecular function	<i>RNA-binding, Protein kinase inhibitor, Chromatin regulator</i>
Post-translational modification	<i>Phosphorylation, Ubiquitination, Acetylation</i>
Technical term	<i>Allosteric enzyme, Transposable element</i>

networks. With this integration we are able to capture the network characteristics between proteins and keywords as well.

Although the concept of integrating additional biological knowledge into a PPI network has been studied in recent years [26, 42, 80, 115, 131], to the best of our knowledge, our work is the first attempt to integrate the PPI network with the *keywords* in Uniprot database and form a heterogeneous network for disease protein prediction. We call this heterogeneous network a *PPI-Keyword* (PPIK) network, which contains network structures accounts for not only protein interactions with one another but also their functional and structural similarities.

In this chapter, we propose to address the problem of disease protein classification on a PPIK network, hinging on the notion of *metagraphs* defined in the study [166]. Different from a traditional network motif or subgraph, a metagraph in a PPIK network, is a graph structure capturing a particular topology of both proteins and keywords on the PPIK network. In other words, each metagraph describes a particular heteronomous biological arrangement between one or more proteins and keywords. Each protein can be subsequently represented as a series of metagraphs that describe its interactions with other proteins and associations with keywords. The key intuition is that proteins with similar functional roles, such as their disease-causing property, tend to have similar metagraph representations, *i.e.*, they tend to interact with other proteins and and associate with certain keywords in a similar arrangement on the PPIK network. Thus, we further build a

classifier for disease proteins based on their metagraph representations. Finally, we conduct comprehensive experiments on three PPI databases, namely IntAct [90], STRING [88] and NCBI [89], and demonstrate the superior predictive power of our proposed metagraph-based prediction model.

3.2 Materials and Methods

In this section, we describe the proposed method. We start with some preliminaries in Section 3.2.1, including the problem statement as well as the motivations of our approach. Next, we introduce the proposed PPIK network and metagraph representations, the basis of our method in Section 3.2.2 and Section 3.2.3, respectively. Lastly, we present a general framework of our method in Section 3.2.4.

3.2.1 Preliminaries

Problem Statement. The problem of disease protein classification aims to identify human disease proteins in a given protein database. Let P be the protein space, and $C = \{\text{disease}, \text{non-disease}\}$ be the set of classes. Assume we have a training set P_{train} and test set P_{test} such that $P = P_{\text{train}} \cup P_{\text{test}}$ and $P_{\text{train}} \cap P_{\text{test}} = \emptyset$. The goal is to learn a classifier $\beta : P \rightarrow C$ based on P_{train} . Ultimately, for any protein in the test set $p \in P_{\text{test}}$, we can predict its class to be $\beta(p)$ with minimal prediction errors.

There have been numerous efforts in designing and learning different classifiers, with considerable success. All these classifiers generally assume a vector representation $\phi(p)$ for each protein $p \in P$. In this work, the main focus is proposing a novel representation $\phi(\cdot)$ for proteins to consistently improve classifiers across the spectrum of classifiers, rather than developing a new classification model.

Motivations. First, while protein-protein interaction (PPI) networks have offered some insight into how proteins work with one another, most existing PPI networks are noisy and incomplete [20, 167]. Interestingly, an individual protein also exhibits a number of biological properties that may reveal how it works, ranging from molecular function and biological process to cellular component and protein domain. Previous work shows that disease proteins are likely to share common properties in Gene Ontology annotations [80, 81]. Furthermore, proteins with the same protein domain as a disease protein could be associated with the disease as well [81, 82]. Therefore, these properties can also be associated with different proteins together, to complement and enhance the existing PPI networks which only encode protein-protein interactions.

Second, disease proteins tend to have similar topological arrangements in a PPI network. This hypothesis has been validated to some extent by previous module-based methods such as clustering [48, 49], k-cores [37], graphlets [65], network motifs [125, 126] and frequent subgraph mining [50]. However, the challenge here is how we can leverage the topological arrangements of proteins in the context of not only the interactions between proteins but also their relevance based on other biological properties.

The two motivations inspire the proposed PPI-Keyword network in Section 3.2.2 and metagraph representation in Section 3.2.3, respectively.

3.2.2 PPI-Keyword network

To exploit the biological properties of individual proteins, we leverage the keywords associated with each protein from the Universal Protein Resource (UniProt) database [19]. These keywords describe the various biological mechanisms of the proteins, as summarized in Table 3.1. To take the biological properties of individual genes into consideration, we extract the rich *Keywords* associated with each

gene from the Universal Protein Resource (UniProt) database [168]. These keywords describe the various biological aspects of the genes, including their *biological processes, cellular components, molecular functions, coding sequence diversity, ligand, protein domains, post-translational modifications* (PTMs), etc. We enrich the PPI network with such biological keywords, to construct a PPI-Keyword (PPIK) network. Since *Keywords* involves GO annotations and these terms have a hierarchical structure in the form of a directed acyclic graph, it needs further discussion to use them directly in our studies. To investigate it, the semantic similarity between pairs of *Keywords* is computed. There are 768 keywords associated with a GO term and 36 of them are associated with more than one GO terms. We performed GO semantic analysis with GOSemSim package in R to compare the similarity between these 768 GO terms. The results show us only 326 pairs out of 294528 (0.1%) pairs have semantic similarity score greater than 0.5. Based on the statistics, using exact match for the keywords is adequate and reasonable.

Formally, a PPIK network is modeled by an undirected graph and unweighted graph $G = (V, E, \ell)$, such that V is the set of nodes, E is the set of edges, and ℓ is the label function on V . In particular, each node can be either a protein or a keyword. They can be differentiated by the label function $\ell : V \rightarrow \{\text{protein}, \text{keyword}\}$. Furthermore, an edge can connect either two proteins, or a protein and a keyword. The former represents the mapped interactions between the two proteins, whereas the latter represents the association between the protein and keyword. Figure 3.1 shows a part of the constructed PPIK network based on the UniProt and IntAct [90] databases. Note that the color scheme of the nodes essentially serves as the label function.

The PPIK network integrates keywords that describe biological mechanisms of proteins into the traditional PPI network. This integration complements the original noisy and incomplete network. On the one hand, protein-keyword associations

FIGURE 3.1: Part of the PPIK network based on UniProt and IntAct databases.

could reinforce useful protein-protein interactions. On the other hand, proteins with no direct interactions can now become related through keywords.

3.2.3 Metagraph representations

On the PPIK network, proteins with similar roles (*e.g.*, their disease-causing functions) tend to have similar topological characteristics. To model topological similarities, previous module-based methods resort to structures such as network motifs [125, 126] and k-cores [37]. However, on a PPIK network, both proteins and keywords exist, and traditional structures do not differentiate between different labels of nodes. Fortunately, the recent emergence of metagraphs [166] has enabled the representation of common structures on a *heterogeneous* graph, where nodes with different labels connect with each other.

In the PPIK network as shown in Figure 3.1, we observe multiple subgraphs with a common structure, which are illustrated in Figure 3.2. More specifically, Figure 3.2(a) showcases two 3-node subgraphs of the PPIK network, both with a common structure “protein–keyword–protein”. Likewise, Figure 3.2(b) illustrates two 4-node subgraphs with a common structure consisting of a triangle of three proteins and one keyword. We call such common structures metagraphs, and the corresponding subgraphs are their instances, i.e, metagraph instances based on the study [166].

FIGURE 3.2: Example metagraphs: common structures of subgraphs on the PPIK network.

Formally, a graph $S = (V_S, E_S, \ell)$ is a *subgraph* of graph $G = (V, E, \ell)$ iff $V_S \subseteq V$ and $E_S \subseteq E$. A graph $M = (V_M, E_M, \ell_M)$ is a *metagraph* for some label function ℓ_M , where each node is defined by its label and its value is immaterial. We say that S is an *instance* of M iff there exists a bijection ω between the nodes of S and M such that

- $\forall v \in V_s, \ell(v) = \ell_M(\omega(v))$, and
- $\forall v, u \in V_s, (v, u) \in E_S$ holds iff $(\omega(v), \omega(u)) \in E_M$ holds.

As a metagraph defines a specific topological arrangement of proteins and keywords, two proteins associated with the same metagraph tend to have similar functional roles. Therefore, we can use metagraphs to construct the vector representation of proteins. Let $\mathcal{M} \triangleq \{M_1, M_2, \dots, M_{|\mathcal{M}|}\}$ denote the set of metagraphs on the PPIK network. Let $\mathcal{I}(M_i)$ be the set of instances of $M_i \in \mathcal{M}$. A protein p can be represented by a vector $\mathbf{m}_p$ of length $|\mathcal{M}|$, where the i -th element is the number of instances of M_i containing the protein p . That is,

$$\mathbf{m}_p[i] \triangleq |\{S \in \mathcal{I}(M_i) : p \in V_S\}|. \quad (3.1)$$

FIGURE 3.3: Example subgraph instances of a metagraph, where some contain disease proteins and some do not (Q9BRI3 is a known disease protein).

Furthermore, the same metagraph can have multiple subgraph instances of different “utilities”. That is, a protein appearing in a subgraph together with a disease protein is more likely to be a disease protein, which implies that such subgraphs have a higher utility towards identifying disease proteins. As shown in Figure 3.3, some subgraphs contain disease proteins and some do not, based on biological knowledge from a disease database. To quantify such utilities, for each metagraph we compute the fraction of its subgraph instances containing any of the known disease proteins. The label function $\varphi : P \rightarrow \{\text{disease}, \text{non-disease}\}$ differentiates known disease proteins from other proteins. Formally, let $\mathbf{d}_p$ be a vector of length $|\mathcal{M}|$, where the i -th element is defined as follows:

$$\mathbf{d}_p[i] \triangleq \frac{|\{S \in \mathcal{I}(M_i) : p \in V_S \wedge (\exists v \in V_S : v \neq p \wedge \varphi(v) = \text{disease})\}|}{\mathbf{m}_p[i]}. \quad (3.2)$$

$\mathbf{m}_p$ and $\mathbf{d}_p$ are $|\mathcal{M}|$ -dimensional representations of protein p based on *metagraphs*. In addition, keywords describing each protein naturally become part of its vector representation. Given a set of keywords $\mathcal{K}$, let $\mathbf{k}_p$ be a vector of length $|\mathcal{K}|$, where the i -th element is 1 iff the i -th keyword is associated with protein p . Thus, $\phi(p)$, the *overall* vector representation of protein p is a vector with $(2|\mathcal{M}| + |\mathcal{K}|)$ dimensions, which is the concatenation of the above representations as shown in Equation (3).

FIGURE 3.4: Metagraphs up to 4-nodes in this study. Green nodes represent the *keywords*, blue nodes represent the *proteins*.

$$\phi(p) \triangleq [\mathbf{m}_p, \mathbf{d}_p, \mathbf{k}_p]. \quad (3.3)$$

We demonstrate the metagraphs up to 4-nodes that we use throughout this study in Figure 3.4.

3.2.4 General framework

Based on the proposed PPIK network and metagraph representations, we describe a overall framework to learn a classifier for disease protein prediction. In particular, the framework consists of three main steps, as summarized in Figure 5.4.

First, from the PPIK network, we mine the collection of metagraphs $\mathcal{M}$. This is an active research area and many off-the-shelf solutions exist. Therefore, we apply an existing state-of-the-art approach GRAMI [169] for this step. Note that the

FIGURE 3.5: General framework of the proposed method.

number of metagraphs becomes 362 if we use metagraphs up to 6 nodes. Thus, we only consider metagraphs up to 5 nodes (164 metagraphs), which is a good balance between efficiency and accuracy as we demonstrate a performance comparison among 5-nodes and 4-nodes in Section 3.3.

Second, we derive the *metagraph* representations for proteins based on the mined metagraphs $\mathcal{M}$. We employ the SymISO algorithm [166] to compute the set of instances $\mathcal{I}(M_i)$ for each metagraph $M_i \in \mathcal{M}$ and then construct the metagraph representations $\mathbf{m}_p$ and $\mathbf{d}_p$ as defined in Eq. 3.1–3.2.

Third, based on the protein representations $\phi(\cdot)$ and training data P_{train} , we build a classifier β through supervised learning. Note that the main focus of this work is to propose the metagraph representations based on the PPIK network, which aims to improve disease prediction across various supervised learning techniques including

random forest, SVM and Generalized Linear Models, as we will demonstrate in the experiments.

3.3 Results and Discussion

In this section, we empirically evaluate the effect of metagraph representations in the context of disease protein prediction. Results show that the proposed representations can significantly improve prediction across various classifiers and substantially outperforms random walk baselines RWR [5] and PRINCE [38]. The reason of choosing these diffusion/propagation based methods is their dominant power over the clustering and neighborhood methods [27, 56, 57].

3.3.1 Data and setup

In this work, to demonstrate the effects of our proposed novel protein representation, we work on three different human PPI databases, namely IntAct [90], NCBI [89] and STRING [88]. We also exploit protein keywords from the UniProt database [19] as illustrated in Table 3.1, to construct a PPIK network for each PPI database. Note that in order to maintain a homogeneous integration of the protein interactions with the *keyword*-associations, PPINs that we use are unweighted and undirected graphs. However, they both can be adopted to this study. Weighted PPINs can be used to eliminate metagraphs which fails to satisfy a certain threshold about the reliability of the protein interactions. Then, they can be treated as the unweighted graphs. Besides, if directed PPINs are used, then each metagraph becomes a directed subgraph. In that case, the complexity of the subgraph matching problem dramatically increases. Thus, limiting the size of the metagraphs to 3 or 4 nodes can be a good option.

Disease labels for proteins are obtained from the UniProt and OMIM databases [92]. In particular, we first obtain disease genes from OMIM, and further map these genes to their product proteins based on UniProt. We conduct disease protein prediction under two different scenarios as follows. First, whether a protein is associated with *all disease*, i.e, all phenotypes in OMIM; second, whether a protein is specifically associated with *breast cancer*, i.e, phenotype breast cancer. Table 3.2 summarizes the three PPIK networks. As we can see from Table 3.2, the three PPIK networks are very different in terms of number of proteins, number of PPI edges, as well as number of PPIK edges.

TABLE 3.2: Summary of the three PPIK networks.

	Proteins	Disease proteins		Keywords	PPI edges	PPIK edges
		All	Breast cancer			
IntAct	13 063	2 947	29	554	97 652	246 092
NCBI	15 951	3 476	31	567	227 004	405 632
STRING	17 668	3 539	29	567	3 912 853	4 107 335

We split each dataset into training and testing sets, containing 80% and 20% proteins respectively. The split is repeated 5 different times. All results reported are averaged over the 5 splits.

To evaluate the effectiveness of the proposed method, we employ the standard metric of Area Under the ROC Curve (AUC), which is a robust measure of the classifiers’ predictive power even with unbalanced classes (*e.g.*, breast cancer).

To combine different type of vectors we apply standardization on the concatenated training data as centering it around zero and scaling with respect to the standard deviation.

Note that we prefer to use metagraphs up to 5-nodes which provide a good balance between efficiency and accuracy which is discussed further in the following sections.

FIGURE 3.6: Performance of Metagraph and Metagraph+ compared to Keywords for *all disease*.

3.3.2 Benefits of metagraph-based representations

As discussed in Section 3.2, given a protein representation $\phi(\cdot)$, different supervised learning models can be applied to derive a disease protein classifier. In particular, we consider three progressively richer representations as follow.

- **Keyword.** We only consider protein keywords from UniProt database (see Table 3.1), *i.e.*, $\phi(p) = \mathbf{k}_p$. These keywords describe various biological properties of proteins, and thus encompass reasonable predictive power for disease proteins.
- **Metagraph.** We enhance the keyword-based representation with metagraph statistics (see Eq. 3.1), *i.e.*, $\phi(p) = [\mathbf{k}_p, \mathbf{m}_p]$. To extract these features we used GRAMI [169]¹ and SubMatch [166]² tools as described in Section 3.2.4. The vector $\mathbf{m}_p$ captures the topological arrangement on the PPIK network for interactions between both proteins and keywords.
- **Metagraph+.** We further incorporate metagraph representations based on the utilities of their subgraph instances (see Eq. 3.2), *i.e.*, $\phi(p) = [\mathbf{k}_p, \mathbf{m}_p, \mathbf{d}_p]$. In particular, $\mathbf{d}_p$ differentiates metagraphs based on the disease class of the proteins in their subgraph instances.

¹GraMi: <https://github.com/ehab-abdelhamid/GraMi>

²SubMatch: <http://www.yfang.site/data-and-tools/submatch>

The above representations are meant to work across different models of supervised learning. In our experiments, we adopted 3 well-known classification models, namely, Random Forest (**RF**), Support Vector Machine (**SVM**) and Generalized Linear Model (**GLM**). For RF, we used `randomforest` package in R and tuned the `mtry` parameter between 1 and the cardinality of protein representation with `tuneRF` function based on OOB error. For SVM, we used `e1071` package in R, and tuned the `gamma` parameter over $\{10^{-5}, 10^{-4}, 0.001, 0.01, 0.1\}$ and the `cost` parameter over $\{0.1, 1, 10\}$ with grid-based `tune.svm` function (based on classification error). For GLM, we used `stats` package in R and adopted default Gaussian distribution. Furthermore, in the case of *breast cancer*, the classes are highly unbalanced. Therefore, we used the `ROSE` package in R to oversample the *breast cancer* class with probability 0.2.

Figure 3.6 shows the AUC performance of the three representations in each of the classification method, on each of the three datasets, for *all diseases*. Two main observations can be made from the results. First, the metagraph representation (**Metagraph**) can significantly and consistently outperform the keyword-only representation (**Keyword**) across all circumstances. Averaging over all classifiers and datasets, **Metagraph** improves AUC over **Keyword** by 12.6%. The results imply that interactions/associations on the PPIK network are powerful towards disease prediction; in particular, proteins with similar functional roles tend to appear in similar topological arrangements. On the contrary, it is inadequate to only consider keywords for individual proteins. Second, the utility-based metagraph representation **Metagraph+** can further enhance the performance. Overall, **Metagraph+** can achieve an average AUC of 90.9, as compared to 88.2 for **Metagraph** and 75.6 for **Keyword**. Thus, metagraphs can effectively incorporate different utilities based on disease proteins in the subgraph instances.

Further comparisons among $\mathbf{m}_p$, $\mathbf{d}_p$, and $\mathbf{k}_p$ vector representations are investigated

FIGURE 3.7: Performance of combinations among m_p, d_p, k_p for *all disease*.

FIGURE 3.8: Performance comparison between metagraphs with up to 4-nodes and up to 5-nodes for *all disease*.

in Figure 3.7 with SVM classifier across all datasets for *all disease*. **Metagraph+** representations outperform the second best combination by between 0.3% and 2.2%. This shows the effectiveness of all three vector representations combined in disease gene prediction problem.

We also demonstrate AUC performance comparison between using metagraphs up to 4-nodes and up to 5-nodes in Figure 3.8 with a tuned SVM classifier. It shows that using metagraphs up to 5-nodes outperforms up to 4-nodes by 31 – 38% across datasets IntAct, NCBI and STRING for *all disease*. Thus, we prefer to use metagraphs up to 5-nodes which provide a good balance between efficiency and accuracy.

Next, we zoom into the results for *breast cancer* only, as shown in Figure 3.9. The performance differences of **Keyword**, **Metagraph** and **Metagraph+** are similar to those for all diseases. Averaging across all classifiers and datasets, **Metagraph+** attains an AUC of 75.8, beating **Metagraph** and **Keyword** by 2.3% and 14.4%, respectively.

FIGURE 3.9: Performance of Metagraph and Metagraph+ compared to Keywords for *breast cancer*.

The results reaffirm that disease proteins tend to appear in the same neighborhood of the PPIK network, and our metagraph representations carry strong predictive power.

3.3.3 Comparison to baselines

Having demonstrated the benefits of metagraph representations, we now compare our proposed work to other baselines for protein disease prediction, as follows.

- RWR or Random walk with Restart [5]. On the PPI network, consider a particle initially at one of the known disease proteins, *i.e.*, the initial position of the particle has a uniform distribution over the disease proteins in the training data. Next, in each step, the particle makes a move on the network: either moving to a randomly selected neighbor with $1 - \alpha$ probability, or jumping to one of the disease proteins in the training data with α probability. Note that the “jumping” effectively returns the particle to the initial condition, and thus “restarts” the random walk. The process is repeated until it converges to a stationary distribution over all the proteins. In the end, candidate proteins in the test data are ranked according to the stationary distribution. We chose RANKS package³ in R as the implementation. The α parameter is tuned over $\{0.1, 0.2, \dots, 0.9\}$ based on AUC performances.

- **RWRK**. The same method as RWR, except that the random walk with restart is performed on the PPIK network described in Section 3.2.2. The same tuning procedure is applied and in the final evaluation the same set of proteins with *RWR* are considered.
- **PRINCE** [38]. In our case, i.e, on an unweighted PPI network, this method basically performs random walk with restart with one major difference than RWR that is prior probabilities. In RWR the initial probability vector consist of equal probabilities apportioned between known disease associated proteins. On the other hand, in **PRINCE** prior probabilities are assigned to each disease associated protein based on a logistic function: $L(x) = \frac{1}{1 + e^{(cx+d)}}$ and $x = S(q, p)$ where $S(q, p)$ is the similarity score between query disease q and the associated disease p with the protein. If the protein is associated with more than one disease then, p is chosen to be the most similar one to q . We use the recommended values as in paper [38] and set c to -15 and d to $\log(9999)$. The similarity scores between phenotypes are derived from the study [135]. The α restart probability parameter is tuned over $\{0.1, 0.2, \dots, 0.9\}$ based on AUC performances as in RWR. We adopt **RANKS** package³ in R as the implementation.
- **PRINCEK**. The same method as **PRINCE**, except that the algorithm is performed on the PPIK network described in Section 3.2.2. The same tuning procedure is applied and in the final evaluation the same set of proteins as *PRINCE* are considered.
- **Subgraph+**. We also compare proposed work to subgraph-based approaches. Unlike metagraphs, traditional subgraphs or network motifs do not differentiate heterogeneous types of nodes. Thus, even we use the same number of nodes (up to 5), the number of subgraphs are 30 and 164 in Subgraph+ and Metagraph+, respectively. In this baseline, we consider subgraphs with

³**RANKS** package: <https://rdrr.io/cran/RANKS/man/do.RWR.html>

TABLE 3.3: Performance of Metagraph+ compared to random walk and sub-graph baselines.

	All Disease			Breast Cancer		
	IntAct	NCBI	STRING	IntAct	NCBI	STRING
RWR	0.551	0.567	0.622	0.578	0.665	0.605
RWRK	0.590	0.587	0.629	0.587	0.664	0.612
PRINCE	-	-	-	0.506	0.716	0.632
PRINCEK	-	-	-	0.634	0.717	0.596
Classifier: RF						
Subgraph+	0.745	0.756	0.826	0.611	0.696	0.713
Metagraph+	0.886	0.862	0.916	0.820	0.748	0.796
Classifier: SVM						
Subgraph+	0.739	0.741	0.808	0.687	0.682	0.597
Metagraph+	0.913	0.902	0.930	0.698	0.715	0.743
Classifier: GLM						
Subgraph+	0.751	0.758	0.818	0.733	0.715	0.796
Metagraph+	0.918	0.921	0.937	0.748	0.734	0.819

only protein nodes, and their statistics and utilities are formulated as protein representations similar to the case of metagraphs (see Eq. 3.1–3.2), which are then concatenated with keyword representations for individual proteins. To extract subgraph features we use GRAMI [169] and SubMatch [166] as in the case of metagraphs. We call this method **Subgraph+**, analogous to **Metagraph+**. We leveraged **Subgraph+** by different classifiers with the same tuning routine as **Metagraph** and **Metagraph+**.

Table 3.3 compares the AUC of the baseline methods with **Metagraph+**, for *all diseases* and *breast cancer*. The foremost observation is that with classifiers GLM and RF, **Metagraph+** is significantly better than all baselines. On average considering all the classifiers, for *all diseases*, it outperforms RWR, RWRK, and **Subgraph+** by 32.9%, 30.7% and 13.8% respectively; for *breast cancer* it outperforms RWR, RWRK, PRINCE, PRINCEK and **Subgraph+** by 14.2%, 13.7%, 14%, 11% and 6.6% respectively. We attribute the better performance to the more predictive representations enabled by metagraphs. Second, the results also show that PPIK network is more effective than PPI network, as evident from the comparison between RWR and RWRK methods. More specifically, random walks on the PPIK network (RWRK) produce more accurate predictions than those on the PPI network (RWR) under

TABLE 3.4: Running time (minutes) comparisons. For **Subgraph+**, **Metagraph**, and **Metagraph+** the classifier with the best AUC performance, GLM, is chosen.

	All Disease			Breast Cancer		
	IntAct	NCBI	STRING	IntAct	NCBI	STRING
RWR	1.0	1.3	2.0	1.4	1.5	1.0
RWRK	1.4	2.3	3.9	1.4	2.2	2.9
PRINCE	-	-	-	0.7	1.0	1.1
PRINCEK	-	-	-	0.9	1.4	1.6
Subgraph+	0.4	1.0	0.8	0.6	0.6	0.8
Metagraph	0.2	0.2	0.2	0.4	0.4	0.2
Metagraph+	0.8	1.0	1.6	1.4	2.0	2.6

most circumstances. In other words, protein keywords can indeed complement and enrich the PPI network.

Table 3.4 examines the running times of the baselines RWR, RWRK, PRINCE, PRINCEK, Subgraph+ and the proposed approach(es). It can be concluded that for RWR and PRINCE, *keywords* integrated into the PPI network induced an increase in the running times. Moreover, for Metagraph+ we observe unfavorable effect of additional vector $\mathbf{d}_p$ of length $|\mathcal{M}|$ (metagraph set size) compare to Metagraph.

3.3.4 Further analysis of the predicted disease proteins

We further study the disease proteins predicted by our proposed methods. Since GLM generally has the best performance among the three classifiers, we only focus on the results from this classifier. A protein is classified as a disease protein if its prediction score is higher than 0.5. Moreover, for each proposed method, we ran disease classification on all three datasets (IntAct, NCBI and STRING). Then, we combine the predictions for all three datasets to obtain a predicted disease gene set for each method.

To enable the analysis, DisGeNET database [170] is used to search the PubMed Ids of the up-to-date publications⁴ reporting the gene-disease associations. In particular, we transform our predicted disease proteins to their producer gene Id's

⁴Up to the year 2016, when we conducted this analysis.

FIGURE 3.10: Average number of PubMed publications per prediction based on DisGeNET.

based on UniProt. Figure 3.10 illustrates the average number of publications per prediction that support the disease gene predictions of our proposed methods. The results are consistent with the AUC performance reported earlier, where methods with higher average publications attain higher AUC scores. Table 3.5 further lists the 10 genes with the most number of PubMed publications for our best proposed method, **Metagraph+**. Table 3.6 zooms into each of the 10 genes, and illustrates the top diseases associated with each gene based on DisGeNET scoring (which combines both curated content and literature).

We also evaluate novel proteins predicted by **Metagraph+**, based on recent publications from years 2014–2016, where each prediction has fewer than 20 publications. There are 27 such novel genes predicted (mapped from the predicted proteins), as listed in Table 3.7. For example, PDE9A (GeneId: 5152) has been reported to be associated with heart failure [171], and ZNF366 (GeneId: 167465) has been reported to be associated with estrogen metabolism and progression of breast cancer, and a new candidate for endometriosis as well [172]. Furthermore, a study [173] presents supportive evidence for KCND2 (GeneId: 3751) being a causal gene for epilepsy, whereas another study [174] proposes KCND2 as a novel cause of J-wave syndrome associated with sudden cardiac arrest.

TABLE 3.5: Top 10 genes predicted by Metagraph+ (GLM) with the most number of publications.

GeneId	Symbol	Full Name
3565	IL4	interleukin 4
3552	IL1A	interleukin 1, alpha
4513	COX2	cytochrome c oxidase subunit II
3133	HLA-E	major histocompatibility complex, class I, E
6696	SPP1	secreted phosphoprotein 1
2272	FHIT	fragile histidine triad
7298	TYMS	thymidylate synthetase
1813	DRD2	dopamine receptor D2
2100	ESR2	estrogen receptor 2 (ER beta)
2078	ERG	v-ets avian erythroblastosis virus E26 oncogene homolog

TABLE 3.6: Top 5 associated diseases reported by DisGeNET for the 10 genes reported in Table 3.5.

Disease UMLS and Name	#PM	Disease UMLS and Name	#PM
GeneId: 3565		GeneId: 3552	
C0004096 Asthma	199	C0018843 Heat Stroke	2
C0011615 Dermatitis, Atopic	65	C0035126 Reperfusion Injury	2
C0034069 Pulmonary Fibrosis	6	C0021368 Inflammation	17
C0035455 Rhinitis	8	C0011633 Dermatomyositis	3
C0993582 Arthritis, Experimental	4	C0037274 Dermatologic disorders	8
GeneId: 4513		GeneId: 3133	
C0268237 Cytochrome-c Oxidase Deficiency	5	C0030491 Parapsoriasis	1
C0027819 Neuroblastoma	5	C0151744 Myocardial Ischemia	2
C0013080 Down Syndrome	1	C0011854 Diabetes Mellitus, Insulin-Dependent	107
C0011853 Diabetes Mellitus, Experimental	1	C0004364 Autoimmune Diseases	105
C0151786 Muscle Weakness	1	C0019693 HIV Infections	86
GeneId: 6696		GeneId: 2272	
C0022650 Kidney Calculi	9	C0024121 Lung Neoplasms	25
C0017638 Glioma	15	C0038356 Stomach Neoplasms	5
C0006663 Calcinosi	3	C0033578 Prostatic Neoplasms	4
C0027627 Neoplasm Metastasis	147	C0007131 Non-Small Cell Lung Carcinoma	29
C1458155 Mammary Neoplasms	18	C0025500 Mesothelioma	2
GeneId: 7298		GeneId: 1813	
C0009404 Colorectal Neoplasms	47	C0036341 Schizophrenia	162
C0038356 Stomach Neoplasms	17	C1834570 Myoclonic dystonia	6
C0009375 Colonic Neoplasms	19	C0030567 Parkinson Disease	22
C1458155 Mammary Neoplasms	10	C0236736 Cocaine-Related Disorders	13
C0034885 Rectal Neoplasms	9	C0001973 Alcoholic Intoxication, Chronic	163
GeneId: 2100		GeneId: 2078	
C1458155 Mammary Neoplasms	48	C0033578 Prostatic Neoplasms	37
C0282612 Prostatic Intraepithelial Neoplasias	2	C0023467 Leukemia, Myelocytic, Acute	34
C0024668 Mammary Neoplasms, Experimental	2	C0023418 leukemia	18
C0014175 Endometriosis	23	C0553580 Ewings sarcoma	22
C0033578 Prostatic Neoplasms	18	C0013080 Down Syndrome	10

TABLE 3.7: Novel disease genes (mapped from predicted proteins) discovered by Metagraph+ based on recent PubMed publications from years 2014–2016.

GeneId	Symbol	PubMedId	Diseases
319	APOF	25726912	Liver neoplasms; Liver carcinoma
506	ATP5B	25666834	Non-alcoholic Fatty Liver Disease; Acute kidney injury
988	CDC5L	26089329	Ischemic Cerebrovascular Accident; Ischemic stroke
1635	DCTD	25735499	Neutropenia; Leukopenia
2686	GGT7	25884624	Glioblastoma; Glioma; Carcinogenesis
3751	KCND2	24501278	Epilepsy; Autistic Disorder; Seizures
		25214526	Cardiac Arrest
		25878292	Alzheimer’s Disease
5152	PDE9A	25799991	Heart failure; Heart Diseases; Congestive heart failure
5634	PRPS2	25149475	Lupus Erythematosus, Systemic
		26004865	Congenital absence of germinal epithelium of testes
6723	SRM	25889691	Prostate carcinoma; Malignant neoplasm of prostate
6942	TCF20	25228304	Autism Spectrum Disorders; Atrial Septal Defects; Moderate mental retardation
8974	P4HA2	25741866	Severe myopia
		26001784	Disorder of skeletal system
9620	CELSR1	25117632	Ischemic stroke
10584	COLEC10	25495265	Chronic Lymphocytic Leukemia
		25786252	Chronic Lymphocytic Leukemia
10940	POP1	26275995	Inflammatory disorder
11097	NUPL2	25584925	Chronic Obstructive Airway Disease; Chronic Obstructive Airway Disease
22938	SNW1	26103569	Skin carcinoma
23513	SCRIB	24802235	Neoplasm Metastasis
23710	GABARAPL1	24879149	Malignant neoplasm of breast; Breast Carcinoma
26270	FBXO6	25811541	Stevens-Johnson Syndrome
55576	STAB2	25989359	Ankylosing spondylitis
84870	RSPO3	24430505	Osteoporosis
115426	UHRF2	25664994	Hepatitis B
144568	A2ML1	26121085	Otitis Media
162515	SLC16A11	25839936	Diabetes Mellitus, Non-Insulin-Dependent; Diabetes; Diabetes Mellitus
		25973943	Gestational Diabetes; Diabetes Mellitus, Non-Insulin-Dependent
167465	ZNF366	25722978	Breast Carcinoma; Malignant neoplasm of breast; Endometriosis; Endometrioma
285671	RNF180	24833402	Stomach Carcinoma; Malignant neoplasm of stomach
340419	RSPO2	25769727	Pancreatic carcinoma; Malignant neoplasm of pancreas

3.3.5 Biological analysis for *Metagraph* representations ($\mathbf{m}_p$)

We further investigate the biological effect of using proposed *Metagraph* representations ($\mathbf{m}_p$) in identifying disease proteins. As mentioned earlier, these vector representations are constructed with metagraph structures up to 5 nodes forming 164 features in total. We apply K-means clustering (K=10) with 44 repetitions on IntAct dataset having 13063 proteins in total. Figure 3.8 shows the confusion

matrix of the clustering performance. Note that the used features do not leak any label information, thus it is performed in an unsupervised fashion. This figure shows that proposed representations successfully identify 9313 proteins. The number of incorrectly clustered instances is 3750 which is 28.707% of all proteins.

TABLE 3.8: Confusion matrix for clustering performance on metagraph structures $\mathbf{m}_p$

Confusion Matrix		Predicted	
		0	1
Actual	0	TN=8554	FP=1562
	1	FN=2188	TP=759

We further investigate the successful (true positive (TP) and true negative (TN)) and unsuccessful predictions (false positive (FP) and false negative (FN)) through several biological aspects such as domain information, molecular function, biological process and pathway. We summarize the distinctive properties among the predictions in Table 3.9. This table focuses on revealing the notable differences among the predictions. Bold values represent the distinctive properties of the predictions. For instance, *lipid metabolism* displays the highest percentage (0.049) for FN proteins. We observe that, overall, the properties of FN and TN proteins exhibit a positive correlation. Similarly, to a lesser extent, we observe a positive correlation among the properties of TP and FP proteins. This could explain the failure of the representations in some cases due to their ability to capture also the biological similarity among the proteins. Thus, more descriptive *keywords* in disease gene identification can lead to a better prediction performance.

TABLE 3.9: Notable biological properties of false negative (FN), true positive (TP), false positive (FP) and true negative (TN) predictions with the number of predictions and their percentages (Pct.%) over the total number of samples presented in the parenthesis (i.e., 2188, 759, 1562, 8554, respectively).

Biological Aspect	Property	FN (2188)	Pct. (%)	TP (759)	Pct. (%)	FP (1562)	Pct. (%)	TN (8554)	Pct. (%)
Biological Process	Transcription	32	0.015	346	0.456	1164	0.745	337	0.039
	Lipid Metabolism	107	0.049	17	0.022	9	0.006	193	0.023
	Immunity	96	0.044	46	0.061	65	0.042	221	0.026
	Protein Biosynthesis	34	0.016	5	0.007	4	0.003	76	0.009
	Neurogenesis	27	0.012	41	0.054	37	0.024	107	0.013
Molecular Function	Hydrolase	221	0.101	86	0.113	88	0.056	721	0.084
	Oxidoreductase	155	0.071	12	0.016	24	0.015	200	0.023
	Chromatin Regulator	9	0.004	47	0.062	155	0.099	62	0.007
Domain	Transit peptide	178	0.081	13	0.017	9	0.006	249	0.029
	Homeobox	11	0.005	37	0.049	34	0.022	48	0.006
	Transmembrane	577	0.264	147	0.194	93	0.060	1720	0.201

3.4 Summary

Disease protein prediction is crucial to the diagnosis and treatment of many diseases. In this study, we integrated protein-protein interaction and biological keywords of proteins, to construct a novel PPIK network. Based on the PPIK network, we further proposed metagraph representations for proteins. Such novel representations can improve the classification of disease proteins consistently across different classifiers, outperforming them by 15.3% in AUC on average. Our method also beats random walk and subgraph baselines by 13.8–32.9%. Finally, our literature search based on PubMed revealed that the proposed method can indeed better predict disease proteins that mapped with newly discovered biological knowledge.

Chapter 4

Integrating Node Embeddings and Biological Annotations for Genes to Predict Disease-Gene Associations

Predicting disease genes has been critical for understanding the genetic basis of human diseases and further developing a treatment scheme. Recent increasing availability of biological information for genes, motivated researchers to leverage these valuable data sources and extract useful information for accurately predicting disease genes. While various computational methods have been proposed for disease gene prediction, the use of fully automated feature extraction on a network data (i.e, embeddings) combined with biological annotations of the genes has not been studied.

We present an *integrative framework* called N2VKO for disease gene prediction problem. Firstly, we learn the node embeddings from protein-protein interaction (PPI) network for genes by adapting Skip-gram based *node2vec* representation

learning method. To the best of our knowledge, it is the first attempt of exploiting node embeddings in disease gene prediction. Secondly, we combine the learned node embeddings with various biological annotations as rich feature representation for genes, and subsequently build binary classification models for disease gene prediction. Finally, as the data for disease gene prediction is usually imbalanced (i.e., the number of the causative genes for a specific disease is much less than that of its non-causative genes), we further address this serious data imbalance issue by applying oversampling techniques for imbalance data correction to improve the prediction performance. Comprehensive experiments demonstrate that our proposed N2VKO significantly outperforms four state-of-the-art methods for disease gene prediction across seven diseases.

4.1 Introduction

Studying disease-causing genes plays critical role in both diagnosis and treatment of various diseases such as cancer and diabetes. Since traditional linkage analysis provides a huge number of candidates, it is expensive and time-consuming to identify real disease genes among these massive candidates via laboratory experiments. Therefore, various computational methods have been developed to address disease gene identification problem.

Protein-protein interaction (PPI) networks where nodes are proteins (gene products) and edges are the physical interactions between proteins [27, 28, 30, 175] are one of the most extensively used tools in prediction of disease genes. We can categorize computational methods on PPI networks into two main groups: Module-based methods and diffusion-based methods. Former are based on the *guilty-by-association* concept that genes within the same topological or functional modules are more likely to be associated with the same disease [45, 46]. Whereas, the latter are designed to traverse through the pathways to the known disease

genes [3, 5, 38]. As diffusion-based methods consider both full network topology as well as the placement of the known disease genes, they are superior than the module-based methods which only consider the local neighborhood. Nevertheless, diffusion-based methods focus on network propagation but usually ignore other valuable data sources (e.g., functional annotations for genes, gene expression profiles, etc), thus still less satisfactory for disease gene prediction.

As an alternative to the above methods, feature-based machine learning methods, are widely used for disease gene prediction. They use constructed representations or feature vectors for genes derived from the PPI networks [30] based on some network properties, such as degree, average distance to disease-genes, common neighbors with disease associated genes, metagraphs (Chapter 3), etc. Then, they train supervised learning models (e.g., SVM [42, 61]) for disease gene classification. While such features *manually* derived from PPI networks are easy to understand, we need to have good knowledge about PPI networks to manually define them and extraction of those hand-crafted features needs tedious efforts as well. Moreover, we may still lose some other useful features.

To address the limitations of existing studies, in this study, we propose a framework called N2VKO for disease gene prediction. Firstly, we *automatically* learn node embeddings as novel features for genes from the PPI networks by adapting a well-known graph embedding method called *node2vec* [66]. Secondly, we further integrate the node embeddings with other valuable biological information to construct comprehensive descriptors for genes and then build up machine learning models for disease gene prediction. Finally, for a specific disease, as the number of its causative genes is much less than the number of non-causative genes and thus the data is highly imbalanced, we apply oversampling techniques for imbalance correction to improve the prediction performance.

We conduct comprehensive experiments on various PPI networks. Experimental results show that integration of node embeddings and biological annotations as

gene features, as well as the oversampling techniques, can improve the prediction performance. The proposed N2VKO framework is demonstrated to significantly and consistently outperform four state-of-the-arts for disease gene prediction across seven diseases. In addition, literature search also shows that N2VKO can effectively predict novel disease genes for various diseases.

4.2 Method

To better illustrate our proposed framework N2VKO, in Figure 4.1, the left side shows the generic pipeline for building classification models, while the right side describes our proposed framework for predicting disease genes or discovering disease-gene associations. More specifically, after data preparation, we performed node2vec to obtain graph embeddings. We combined biological annotations with these embeddings and applied oversampling techniques before we constructed classification models. Next, we will introduce each step in N2VKO in more details.

FIGURE 4.1: Framework of our N2VKO for Predicting Disease-Gene Associations

4.2.1 Data Preparation

To predict novel disease-gene associations, we exploit three types of data in our study, namely, protein-protein interaction (PPI) networks, biological annotations for proteins (i.e., *Keywords*) downloaded from UniProt [168] database (Table 3.1 in Chapter 3), and the existing disease-gene associations obtained from OMIM [92] database.

Proteins interact with each other to perform specific biological functions or tasks¹. In a PPI network, proteins and interactions are represented by nodes and edges, respectively. Formally, a PPI network is modeled as an undirected graph $G = (V, E)$, where V stands for the set of proteins and E represents the set of interactions. In particular, we work on two PPI networks, namely IntAct [90] and NCBI [89]. There are 13,063 unique proteins and 97,652 interactions in IntAct and 15,951 unique proteins and 227,004 interactions in NCBI, respectively.

Disease-gene associations involve known human diseases and the human genes whose mutations causing these diseases. We extracted disease-gene associations from OMIM [92] database, which is the best-curated resource for known phenotype-genotype relationships. In this study, we aim to predict causative genes for 7 specific diseases, namely, Alzheimer's Disease (Al), Breast Cancer (BC), Colorectal Cancer (CC), Prostate Cancer (PC), Diabetes Mellitus (DM), Lung Cancer (LC) and Obesity (Ob). Table 4.1 shows the number of confirmed causative genes from OMIM for specific diseases. For example, there are 13 genes for Alzheimer in IntAct data and 47 genes for Diabetes Mellitus in NCBI data, respectively.

¹We work on PPI networks for disease gene prediction and thus we will use genes and proteins interchangeably.

TABLE 4.1: The number of disease-genes in IntAct and NCBI data, respectively.

Diseases	IntAct	NCBI
Alzheimer' s Disease (Al)	13	14
Breast Cancer (BC)	30	32
Colorectal Cancer (CC)	34	35
Prostate Cancer (PC)	18	20
Diabetes Mellitus (DM)	37	47
Lung Cancer (LC)	18	20
Obesity (Ob)	21	33

4.2.2 Learning the Feature Representations for Genes

We first briefly introduce node2vec [66], which we adapt it to learn the feature representations for genes from the PPI networks. Then, we introduce our N2VK representations (node2vec+Keywords) which integrate both node2vec embeddings and keywords for genes.

4.2.2.1 node2vec

node2vec [66] is an algorithmic framework for learning feature representations for nodes in networks. Given a network, it can learn continuous feature representations for the nodes and then use these learned features for various downstream machine learning tasks, e.g., clustering, node classification and link prediction.

Inspired by the models (e.g., word2vec [176]) for natural language processing, node2vec treats a network as a “document”, where nodes are “words” and sampled walks (or paths) are “sentences”. node2vec has a flexible strategy to sample the neighborhood, which enables to alternate between breadth-first sampling strategy (BFS) and depth-first sampling (DFS) strategy. In particular, in our study, BFS consecutively extends its sampling space with immediate neighbors of the source protein in PPI network, whereas DFS extends the sampling space with sequential neighbors at increased distances from the source protein.

Search bias α consists of two parameters p and q and leads the walks across two strategies. Note that, $d_{tx} \in \{0, 1, 2\}$ stands for the shortest path distance between proteins from t to x in a walk. Parameter p determines the likelihood of immediately revisiting a node in the walk. If it is set to a high value ($> \max(q, 1)$) it is less likely to sample an already visited node. Whereas, parameter q controls the DFS-BFS behavior of the walk. If $q < 1$, the walk tends to have a higher depth in DFS behavior by visiting the nodes which are far from the node t . Consider a random walk that traversed from node t to node v and now resides at node v . The transition probability from node v to node x , denoted as π_{vx} , is defined as $\pi_{vx} = \alpha_{pq}(t, x) \cdot w_{vx}$, where w_{vx} is static original edge weight between proteins v and x . In particular, the search bias function $\alpha_{pq}(t, x)$ is defined as follows:

$$\alpha_{pq}(t, x) = \begin{cases} \frac{1}{p} & \text{if } d_{tx} = 0 \\ 1 & \text{if } d_{tx} = 1 \\ \frac{1}{q} & \text{if } d_{tx} = 2 \end{cases}$$

In Figure 4.2, the random walk that traversed from node t (i.e. Q495A1) to node v (i.e. P15151) and now resides at node v . The transition probabilities from v to x_1 , x_2 , x_3 and x_4 are then determined by the shortest path distances from t to x_1 , x_2 , x_3 and x_4 . For example, the shortest path distance between t and x_2 (i.e. P60409) is 2 and $\alpha_{pq}(t, x_2)$ is thus $\frac{1}{q}$. Assuming the edge weight w_{vx_2} is 1, the transition probability π_{vx_2} is then $\frac{1}{q}$ as shown in Figure 4.2.

Sampled walks or paths from the PPI networks will then be fed into a single-layer neural network, which will learn vector representation for each protein. In the learning process, skip-gram architecture [176] on network data is applied to learn the feature representation by optimizing the neighborhood preserving likelihood objective function. The goal is to maximize the log-probability of observing a network neighborhood $N_S(u)$ for a node u conditioned on its feature representation:

FIGURE 4.2: node2vec transition probabilities on a small PPI network. A random walk traversed from node t to node v and now resides at node v . The transition probabilities from node v to other proteins x_1 , x_2 , x_3 and x_4 are shown in this Figure.

$$\max_f \sum_{u \in V} \log P(N_S(u) | f(u)) \quad (4.1)$$

where $f(u)$ is the feature representation for a given source node u and $N_S(u) \subset V$ represents random walk outcome for the source node u with sampling strategy S .

4.2.2.2 N2VK Representations

node2vec embeddings are learned from the PPI networks. Meanwhile, many databases, e.g., Gene Ontology and UniProt, provide additional information (e.g, GO annotations and keywords) for genes. To better represent genes in feature vectors for disease-gene prediction, we combine both the node2vec embeddings with the additional information for genes.

N2VKO framework 4.1 has an extensive feature selection procedure, which allows us to combine multiple data sources in a flexible way through the concatenation operation. In this study, we focus on the same *keywords* for genes obtained from UniProt, as summarized in Chapter 3, Table 3.1. In total, there are 554 keywords for 13,063 genes in IntAct data. As such, given a gene in IntAct, we have a

554-dimensional feature vector with binary values to describe all the keywords it has (e.g. if a gene has a certain function). Similarly, we have 567-dimensional feature vectors for genes in NCBI data. Eventually, we concatenate the node2vec embeddings with the feature vectors for keywords as our N2VK representations, i.e. node2vec embedding + keywords.

Note that as described in Chapter 3 (Section 3.2.2), the *keywords* involves many aspects of proteins such as their domain or GO annotations. GO terms are known to have a hierarchical structure in the form of a directed acyclic graph. After a semantic similarity analysis among GO terms in *keywords*, we conclude that only 326 pairs out of 294528 (0.1%) GO term pairs have semantic similarity scores greater than 0.5. This analysis shows us that there is no severe harm to use exact keywords in N2VK representations through the concatenation operation.

4.2.3 Feature Selection

After we have the N2VK representations for genes, some features may be irrelevant for predicting disease-gene associations. In addition, different feature subsets may be used to predict the disease-genes for specific diseases. Therefore, feature selection is an important step before we build the classification model for predicting disease-gene associations.

In this study, we investigate four different feature selection (FS) techniques, namely, *Earth* [177], *varImp*, *Stepwise Linear Model*, and *MRMR* (minimum redundancy – maximum relevance) [178]. We briefly introduce these FS techniques as follows.

Earth is a package in R [177]. It implements the non-parametric regression technique “Multivariate Adaptive Regression Splines”, commonly known as MARS. MARS is extension to the linear regression which captures nonlinearities and interactions between variables/features. In particular, the function *evimp* in *Earth* package [177] is used to estimate variable importance.

varImp is a function implemented in Caret package [179], which calculates variable/feature importance for regression and classification models. In particular, for linear models, it extracts variable importance based on the absolute value of the t-statistics for each model parameter.

Stepwise Linear Model has also been employed for feature selection. It starts with a full model with all the features, and iteratively drops a less important feature, which gives the minimum AIC (Akaike Information Criterion) value when dropped. It stops when there is no significant drop in AIC achieved and returns the remaining features as selected features.

MRMR [178] is a feature selection approach which aims to minimize the redundancy and maximize the relevance for the selected features. Firstly, it selects features maximally dissimilar to each other to ensure the minimum redundancy among them. Secondly, it will utilize the mutual information between the features and class labels to maximize the relevance for the features. We used the *mRMRe* package [180] in *R* in this study.

4.2.4 Imbalance Correction

As we can observe from the Table 4.1, our data for disease gene prediction is highly imbalanced. For example, we have 30 and 32 disease genes (or positive samples) for breast cancer in IntAct and NCBI, respectively. Meanwhile, IntAct and NCBI have 13,063 and 15,951 genes in total. This data imbalance issue often results in degraded performance as standard machine learning tools tend to bias towards the majority class, i.e. negative or normal genes. To address this issue, we apply two overampling techniques, namely SMOTE [181] and ROSE [182], for imbalance correction in our N2VKO.

SMOTE (Synthetic Minority Oversampling Technique)² creates synthetic samples based on two parameters, namely k and N . In the feature space of minority class, SMOTE selects a sample and its k nearest neighbors. SMOTE further generates a data point between the selected sample and one of its k nearest neighbors as the synthetic data point. In addition, the parameter N will determine the number of synthetic data points to be generated.

ROSE (Random Over-Sampling Examples) is a R package³ to deal with binary classification problems for imbalanced data. In ROSE, synthetic samples are generated according to a smoothed bootstrap approach. In particular, it creates new samples from a conditional kernel density estimate of the two classes. The parameter p in ROSE determines the number of minority class examples in the resulting data created by ROSE.

After applying SMOTE or ROSE, we build various classification models on the balanced data, e.g., Random Forest and Support Vector Machines, for disease gene prediction. In summary, our proposed N2VKO consists of three main steps in Figure 4.1. Firstly, we learn N2VK representations by integrating node2vec embeddings and keywords. Secondly, we conduct feature selection to extract subset of important features for classification. Finally, we apply oversampling techniques including SMOTE and ROSE for imbalance correction.

4.3 Results and Discussion

4.3.1 Experimental Setup

In our experiments, we worked on two different human PPI databases, namely IntAct [90] and NCBI [89]. We also exploited *Keywords*, i.e, biological annotations

²SMOTE: <https://www.rdocumentation.org/packages/DMwR/versions/0.4.1/topics/SMOTE>

³ROSE: <https://cran.r-project.org/web/packages/ROSE/ROSE.pdf>

of proteins from UniProt [168] database as illustrated in Table 3.1. Disease labels for genes were extracted from OMIM [92] database based on phenotype entries. The number of positive examples (i.e., causative genes) for each disease is listed in Table 4.1.

We employed the standard metric of Area Under the ROC Curve (AUC), which is a popular and robust measure for models even upon imbalanced data and it can effectively capture the ranking effect of potential disease genes. In addition, we conduct five-fold cross-validation for each specific disease where train datasets contain 80% of genes while test data sets contain 20% genes respectively. Finally, the 5 AUC results from five-fold cross valuation were averaged to obtain the final prediction results. Note that, all our feature selection experiments were performed on the training data.

4.3.2 Classification Model Selection

We investigated and compared 4 well-known classification models for disease gene prediction, namely, k-nearest neighbors (kNN), random forest (RF), support vector machine (SVM) and generalized linear model (GLM). kNN algorithm computes k nearest training set vectors for each testing sample based on Euclidean distance and the classification is performed upon majority vote. RF is an ensemble learning technique. It samples training set by sampling with replacement and uses out-of-bag data to compute an unbiased estimate of the classification error. SVM aims to find the maximum-margin hyperplane in a transformed feature space using various kernels. GLM will fit a linear equation to model the association between features and the class label.

We used different R packages for these 4 models - `class` package for kNN, `randomforest` package for RF, `e1071` package for SVM and `stats` package for GLM. The parameters for these models are tuned to achieve their optimal performance. For example,

we tuned the parameter *mtry* in RF with `tuneRF` function and we performed grid-based tuning for the parameter *gamma* over $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1\}$ and the parameter *c* over $\{0.1, 1, 10\}$ in SVM.

We first ran these 4 models on node2vec embeddings to predict the causative genes for the seven diseases. As shown in Figure 4.3, GLM achieved the highest average AUC scores on both IntAct and NCBI. Thereafter, all the results obtained under different scenarios were based on GLM.

FIGURE 4.3: AUC comparison for various classifiers on node2vec embeddings through 7 diseases. “Average” in the last column shows the average performance of classifiers over all the seven diseases on a) IntAct dataset and b) NCBI dataset.

We performed grid-based tuning over the two parameters of node2vec, namely *p* and *q* (both *p* and *q* are selected over $\{0.25, 0.5, 1, 2, 4\}$). We chose the embedding which gives the highest AUC score and use this embedding throughout the experiments for each specific disease. For example, we set *p* as 0.25 and *q* as 2

FIGURE 4.4: AUC performances over different p and q pairs for disease Alzheimer's and breast cancer.

for Alzheimer's Disease (Al) on IntAct data. Note that different p and q values are used for different diseases. Figure 4.4 shows $p - q$ effect in performances over breast cancer and Alzheimer's Disease for both datasets. Overall different $p - q$ values yield similar performances. To see the effect of embedding dimension D in our experiments, we examine $D = \{32, 64, 128, 256, 512\}$ in Figure 4.5. We set *node2vec* embedding dimension as default $D = 128$, which attains the highest AUC performance outperforming the runner-up $D = 64$ by 16%.

Note that, as introduced in Section 4.2.4, we set k as 5, tune parameter N for SMOTE over $\{200, 300, \dots, 900\}$ and; for ROSE, we tune parameter p over $\{0.1, 0.2, \dots, 0.9\}$ on each fold of five-fold cross validation independently.

FIGURE 4.5: The effect of embedding dimension D on the average disease performance on IntAct dataset.

4.3.3 Impacts of Keywords and Oversampling

In this section, we evaluated the impacts of keywords and oversampling in our N2VKO. We compared four scenarios as follows.

- *node2vec*: node2vec embeddings were used as features without feature selection or oversampling.
- *N2VO*: Oversampling was applied node2vec embeddings without feature selection. On a specific dataset, we used the oversampling method which achieved the higher AUC performance.
- *N2VK*: Feature selection was applied on N2VK representations. On a specific dataset, we used the feature selection method (e.g., Earth, varImp, StepwiseLM and MRMR), which achieved the higher AUC performance.
- *N2VKO*: Feature selection was applied on N2VK representations. Oversampling was then applied on the selected feature subset. On a specific dataset, we used the best combination of feature selection and oversampling.

In Figure 4.6, we compared node2vec, N2VO, N2VK and N2VKO in terms of AUC on both IntAct and NCBI data sets. Similarly, the last column “Average” in Figure 4.6 is the average AUC over all the seven diseases. Firstly, we observe that N2VK performs better than node2vec on both IntAct and NCBI, indicating that

keywords are good complements to node embeddings for disease gene prediction. Secondly, N2VKO outperforms N2VK by 2.71% on IntAct and 2.69% on NCBI, respectively. The comparison between N2VKO and N2VK clearly demonstrates that imbalance correction by oversampling techniques indeed enhances the prediction performance.

FIGURE 4.6: Comparison among node2vec, N2VO, N2VK and N2VKO on a) IntAct dataset and b) NCBI dataset.

To further emphasize the contribution of biological annotations, we applied FS and oversampling solely on *node2vec* embeddings. Figure 4.7 shows that node embeddings with feature selection and oversampling procedure is not sufficient alone without additional support of biological knowledge.

We also present Figure 4.8 to show that same biological annotations, as well as same feature selection and oversampling procedures result in lower performance on Metagraph+ (our previous work), compared to our integrative approach. The

final Metagraph+FSO has already been displayed in our benchmark comparison as *Metagraph+*.

FIGURE 4.7: FS and oversampling applied solely on *node2vec* embeddings.

In addition, N2VO and node2vec achieve comparable performance on IntAct and NCBI as shown in above Figure 4.6. Note that we did not perform feature selection for node2vec embeddings in N2VO. It is usually more practical to perform feature selection before applying the oversampling techniques, and this helps to explain N2VO and node2vec achieve comparable performance. Moreover, we can clearly observe that both ROSE and SMOTE achieve better performance on FS applied features in Figure 4.9. These results are consistent with the above comparison between N2VKO and N2VK in Figure 4.6. In addition, ROSE achieves higher average AUC scores than SMOTE on both IntAct and NCBI, as shown in Figure 4.9.

FIGURE 4.8: FS and oversampling applied on *Metagraph+* representations.

4.3.4 Feature Selection and Importance Analysis

In Figure 4.10, we compared 4 feature selection techniques on N2VK representations. Overall MRMR and VarImp perform better than Earth and Stepwise Linear Model.

We also investigated the combinations of feature selection and oversampling to obtain the final N2VKO results. We showed the AUC scores for all the eight combinations on four cancer-related diseases in Figure 4.11. For each specific disease, we selected the combination resulting in the best performance. Taking lung cancer (LC) on NCBI as example, we used the combination of MRMR and ROSE in our N2VKO framework. In addition, we can draw similar conclusions that ROSE

FIGURE 4.9: Comparison between ROSE and SMOTE on a) IntAct dataset and b) NCBI dataset. **ROSE** refers to applying ROSE on node2vec embeddings, while **ROSE-FS** refers to applying ROSE after the feature selection on node2vec embeddings.

outperforms SMOTE, and VarImp and MRMR perform better than the other two methods as shown in previous figures.

Moreover, we analyzed the selected keywords for each disease in our N2VKO and showed the percentage of their categories in Figure 4.12. “Biological process” has the highest percentage and is thus the most representative category for disease gene prediction.

We further investigated the keywords selected in N2VKO to predict lung cancer genes on NCBI dataset as shown in Table 4.2. There are 13 keywords selected for lung cancer in Table 4.2, and some of these keywords are indeed important for lung cancer based on our literature search. For example, the 2nd keyword in Table 4.2,

FIGURE 4.10: AUC comparison among 4 feature selection methods on N2VK representations.

SH3 domain, is important for lung cancer. There are approximately 300 SH3 domains encoded in the human genome, and a total of 56 human SH3 domains have been reported to be involved in the growth, proliferation, apoptosis, invasiveness, and metastasis of lung cancer [183]. In addition, the levels of “Branched-chain amino acids” such as leucine, isoleucine and valine significantly change in blood and cells of lung cancer patients as shown in [184]. Therefore, a disruption on the proteins which is responsible in synthesis of branched-chain amino acids (12th keyword in the Table) might cause the development of lung cancer.

FIGURE 4.11: AUC comparison for various combinations of feature selection and oversampling techniques in N2VKO framework on cancer-related diseases. Four cancer-related diseases include *breast cancer* (BC), *colorectal cancer* (CC), *lung cancer* (LC) and *prostate cancer* (PC).

4.3.5 Comparison with State-of-the-arts

We compared our proposed N2VKO with four state-of-the-art methods for disease gene prediction, including RWR [5], RWRH [3], Catapult [42], Prodige [61] and Metagraph+ (Chapter 3).

- *RWR* (Random Walk with Restart). RWR simulates iterative traversals of a walker starting from the seed genes (i.e., known disease genes) to a randomly selected neighbor in the PPI network. Different from traditional random

FIGURE 4.12: The percentages of *Keywords* categories used in N2VKO representations through 7 diseases. The percentage contribution of *Keywords* categories in N2VKO is investigated a) IntAct dataset. b) NCBI dataset.

walk, RWR is able to jump back to any seed gene with a pre-defined probability at each iteration.

- *RWRH* (Random Walk with Restart on Heterogeneous Network). RWRH prioritizes genes according to their relevance with disease genes. It combines the PPI network and phenotype network relying on the protein-phenotype associations. RWRH performs random walk with restart across inter and intra transitions in both networks.
- *Catapult*. Catapult is a Positive-Unlabeled learning model for disease gene prediction [42]. It is a supervised machine learning method that employs a

TABLE 4.2: *Keywords* in selected features for Lung Cancer on NCBI dataset

	Keyword	Category
1	Peroxisome biogenesis	Biological process
2	SH3 domain	Domain
3	Triplet repeat expansion	Coding sequence diversity
4	Serine biosynthesis	Biological process
5	Repeat	Domain
6	Antiport	Biological process
7	Dipeptidase	Molecular function
8	Cobalt	Ligand
9	Heparin-binding	Molecular function
10	Vacuole	Cellular component
11	Acetylation	PTM
12	Branched-chain amino acid biosynthesis	Biological process
13	Urea cycle	Biological process
14	Topoisomerase	Molecular function

biased support vector machine (SVM) upon walk-based features in a heterogeneous gene-disease network. In biased SVM, the bias assigns more penalty on false negatives than false positives to address the class imbalance issue for disease gene prediction.

- *Prodige*. Prodige is another Positive-Unlabeled learning model for disease gene prediction [61]. It builds a support vector machine model that calculates similarity scores for gene pairs. It defines two kernels for pairs of genes and pairs of phenotypes. By exploiting these two kernels, it provides the resulting kernel for gene-phenotype pairs.
- *Metagraph+*. Metagraph+ (Chapter 3) provides representations for genes which include both topological features (i.e. metagraphs) and biological annotations (i.e. keywords extracted from UniProt). In our experiments, similar to N2VKO, we also applied feature selection and oversampling to predict disease genes with GLM as classification model in Metagraph+.

In Figure 4.13, we compared AUC performances of our N2VKO with state-of-the-art methods in this field. N2VKO on IntAct data outperforms RWR, RWRH, Catapult, Prodiges and Metagraph+ by 31%, 15%, 14%, 11%, and 6% respectively. Similarly, on NCBI data, N2VKO outperforms by 27%, 11%, 14%, 12%, and 7%. Clearly, N2VKO performs better than these existing algorithms for disease gene prediction on both IntAct and NCBI data.

FIGURE 4.13: Comparison between our N2VKO and various state-of-the-arts for disease gene prediction.

We observed that N2VKO achieved low performance on obesity (Ob.) and prostate cancer (PC) data. To investigate the reasons behind, we first visualized the interactions for all the disease genes in PPI networks. For a specific disease, we constructed a sub-network where all the nodes are either the genes for this disease or their direct neighbours (one-hop neighbours). Figure 4.14 shows the sub-networks

FIGURE 4.14: One-hop sub-networks for various diseases.

for all the 7 diseases and we can roughly observe that Ob and PC sub-networks are very sparse.

Moreover, we calculated the clustering coefficient scores and counted the number of connected components for these sub-networks as shown in Figure 4.15. We normalized both the clustering coefficient scores and the number of connected components

FIGURE 4.15: Connected component and clustering coefficient analysis for one-hop disease sub-networks.

into the range $[0.5, 1]$. For DM, PC and Ob on IntAct, they have more connected components and smaller clustering coefficient scores than others, and they achieve low AUC. On NCBI, PC, LC and Ob have much smaller clustering coefficient scores than others and they achieve lower performance. Therefore, it is clear that high sparsity (low clustering coefficient scores) will result in low prediction performance. High sparsity prohibits the influence propagation in the network, and RWR (Random Walk with Restart) thus achieves extremely low performance for PC and Ob on IntAct and Ob on NCBI as shown in 4.13. Hence, it is also reasonable that node2vec, a random walk based graph embedding algorithm, achieves poor results in those sparse networks.

4.3.6 Further Analysis on the Effect of Biological Annotations

To evaluate the effect of *keywords* in a disease specific analysis, in Figure 4.16 we demonstrate a comparison between proposed method N2VKO and the same feature selection and oversampling procedure applied on solely node2vec embeddings (i.e., N2V-FSO) without the *keywords*. It can be clearly observed that for the two cases, obesity and prostate cancer, which have a scattered known-disease distribution as discussed earlier in Section 4.3.5, the contribution of *keywords* is higher than the other diseases. This outcome expresses the importance of using an additional biological evidence to compensate the inefficacy of a random walk-based embedding algorithm on a scattered known-disease distribution. Exceptionally, on NCBI dataset, *keywords* exhibit a small negative effect on lung cancer prediction by 2% in AUC performance. It is hard to derive a conclusion about lung cancer proteins due to the dataset’s highly imbalanced nature, i.e., the number of known-disease proteins is 20 as shown in Table 4.1. Moreover, keywords yield an improvement over AUC performance of IntAct PPI embeddings for lung cancer. This is an indication about the prediction superiority of the embeddings derived from NCBI network over selected *keywords*. As an additional evidence on the effectiveness of NCBI embeddings, we can point to the difference on the average AUC performances between N2V-FSO and N2VKO for two datasets. To summarize, the derived embeddings from NCBI network exhibit a slight supremacy over Intact embeddings, however biological annotations are capable of closing the gap.

4.3.7 Comparison of the Integration Methods

Note that, apart from the concatenation of the *keywords* and the embeddings, we investigate several integration techniques such as the Hadamard product, the auto-encoder (AE) and the summation of the node2vec embeddings on the PPI

FIGURE 4.16: To show the effect of biological annotations (i.e., *keywords*), two models consists of the same feature selection and oversampling procedure applied on solely node2vec embeddings (i.e., N2V-FSO) and its keywords concatenated version (i.e., N2VKO) are compared across seven diseases on datasets a) IntAct b) NCBI.

and “keywords network” (i.e., a constructed protein network based on the common/shared *keywords* between the proteins). The auto-encoder we use consists of two linear layers and a sigmoid layer right after the linear layers. We feed the concatenation of the embeddings into the auto-encoder and extract the features for prediction. The Hadamard product simply gives us the element-wise multiplication of the two embeddings and the summation is obtained by the element-wise addition of the embeddings. Besides, since our N2VKO framework conducts extensive feature selection experiments, the concatenation of PPI embeddings with *keywords* terms is a reasonable and a flexible method to combine multiple sources. Table 4.3 shows the effectiveness of the concatenation in disease gene prediction compared to the other methods on the node2vec embeddings with the same parameter setting.

TABLE 4.3: Comparison among the integration methods for IntAct dataset.

Disease	AE	Concat	Sum	Hadamard
Al	0.5899	0.6988	0.4593	0.5481
BC	0.6770	0.7964	0.8110	0.6703
CC	0.7165	0.8568	0.7528	0.5839
DB	0.5752	0.5588	0.5821	0.4472
LC	0.5497	0.5900	0.4613	0.4068
Ob	0.6599	0.6395	0.4952	0.7097
PC	0.6213	0.6328	0.5322	0.4643
Average	0.6271	0.6819	0.5848	0.5472

4.3.8 Case Studies

We conducted further analysis for the disease genes predicted by our N2VKO. In particular, we prioritized the candidate genes based on their prediction scores. For each top-ranked gene, we searched in DisGeNET database [170] for the PubMed records reporting the diseases associated with this gene.

In Table 4.4, we showed the genes predicted by N2VKO on IntAct, which are ranked as top 10 predictions and also have at least 10 PubMed records supporting our predictions. The column “#PMs” shows the number of PubMed records which support the predicted disease-gene associations. As shown in Table 4.4, STK11 is a candidate gene for breast cancer (BC) with the highest prediction score. There are also 23 PubMed publications reporting the association between STK11 with breast cancer and “Breast Carnicoma” is the most frequent label in these publications. In addition, RAD51C’s rank is 8 and it has 14 supporting PubMed records. Similarly, genes on NCBI with top 10 prediction scores and at least 10 PubMed supports are listed in Table 4.5. Both Tables 4.4 and 4.5 demonstrate that our N2VKO is capable of predicting novel genes for different diseases effectively.

TABLE 4.4: Predicted disease genes on IntAct with top 10 prediction scores and at least 10 publications in PubMed records.

Disease	Rank	Gene Symbol	# PMs	Related diseases in PubMed Records
Breast Cancer (BC)	1	STK11	23	Breast Carcinoma
	6	AKT2	11	Malignant neoplasm of breast
	7	PGR	465	Malignant neoplasm of breast, Breast Carcinoma
	8	RAD51C	14	Malignant neoplasm of breast, Hereditary Breast and Ovarian Cancer Syndrome
Diabetes Mellitus (DM)	1	AKT1	15	Diabetes Mellitus Non-Insulin-Dependent, Diabetes Mellitus
Lung Cancer (LC)	1	TNFSF10	39	Carcinoma of lung, Non-Small Cell Lung Carcinoma
	3	AURKA	13	Malignant neoplasm of lung
	4	IL24	20	Lung Neoplasms, Carcinoma of lung

TABLE 4.5: Predicted disease genes on NCBI with top 10 prediction scores and at least 10 publications in PubMed records.

Disease	Rank	Gene Symbol	# PMs	Related diseases in PubMed Records
Breast Cancer (BC)	2	IL24	16	Breast Carcinoma
	3	NAT2	36	
	4	NBN	37	
Colorectal Cancer (CC)	6	MRE11	11	Colorectal Cancer, Malignant tumor of colon
Diabetes Mellitus (DM)	2	EIF2AK3	14	Diabetes Mellitus Insulin-Dependent, Diabetes
Lung Cancer (LC)	3	HRAS	32	Lung Neoplasms, Malignant neoplasm of lung
Obesity (Ob)	2	NPY	65	Obesity
	4	SREBF1	21	
Prostate Cancer (PC)	1	ERBB2	87	Prostate carcinoma, Malignant neoplasm of prostate

4.4 Summary

Identification of disease-related genes by means of computational approach provides cost-effective techniques for the diagnosis and treatment of many diseases. In this study, we proposed a framework N2VKO to exploit human PPI networks and biological aspects of genes and predict disease-gene associations. In particular, we learned node embeddings from the PPI networks and extracted the keywords from UniProt as features for genes to build various classification models. Considering the data is highly imbalanced, we also applied oversampling techniques for imbalance correction. Comprehensive experimental results show that integration of node embeddings and keywords as gene features, as well as the oversampling techniques, can effectively improve the prediction performance. The proposed N2VKO significantly outperforms the state-of-the-arts including RWR, RWRH, Catapult, Prodigy and Metagraph+ for disease gene prediction across 7 diseases. Literature search also demonstrates that N2VKO can effectively predict novel genes for various diseases.

Chapter 5

Unsupervised Multi-View Network Embedding with Higher-Order Collaboration

For the last decade, integrating multiple biological network data has been popular in disease gene prediction and gene functional analysis [40, 72, 110, 111]. Unlike traditional single-view networks, multiple-view networks maintain distinct semantics to complement each other. While node embedding applications on single-view networks emerged as a promising direction, due to sparse and noisy circumstances, exploiting multiple views jointly towards extracting more robust representations is imperative. To the best of our knowledge, our recent study introduced in this chapter, is the first attempt of multi-view network embedding in disease gene prediction.

In this chapter, we propose MICROS, an unsupervised approach for multi-view network embedding, to learn low-dimensional and continuous vector representation to be fed into various downstream tasks. Similar to existing multi-view network embedding studies, MICROS is based on diversity and collaboration principles. While

diversity principle enables views to maintain their individual topological characteristics, collaboration principle enables views to work together and reinforce each other. However, we also investigate a novel form of collaboration (i.e., second-order collaboration) that has not been explored previously on multi-view networks, and further integrate it into a unifying framework of principles/consistencies to extract superior node representations. Finally, we conduct comprehensive experiments and our results demonstrate that our learned representations consistently outperform state-of-the-art approaches on various downstream tasks, across three real-world multi-view networks including a PPI network and a gene ontology (GO) network.

5.1 Introduction

Throughout the years researchers in disease-gene discovery provided various molecular networks ranging from PPI networks and gene co-expression networks to gene regulatory networks and metabolic interaction networks. They provide deeper insight into genes and diseases from multiple perspectives through classification and clustering problems on the network. Naturally, deriving robust and effective representations through leveraging multiple networks becomes inevitable. In particular, network embedding approach has arisen as a desirable strategy, which aims to learn a low-dimensional and continuous vector representation for each node on the network. While many network embedding studies [66, 68, 69] focus on single-view networks, there are also several multi-view network embedding methods [74, 76, 108, 185]. The main advantage of multi-view network embedding methods is, they are capable of leveraging various perspectives. For instance in a biological concept, the resultant embeddings can keep both the protein’s physical interactions and gene ontology aspects all at once. Furthermore, biological networks tend to be noisy, thus it is inevitable to design efficient frameworks which enable multiple networks to complement each other.

FIGURE 5.1: A real-world multi-view network consists of two views: PPI view (a part of PPI network from IntAct database) and GO view (a part of Gene Ontology network).

A toy multi-view network is shown in Figure 5.1, consists of two views: the same set of proteins can be connected in two distinct ways, through having physical interaction or having Gene Ontology (GO) [186] similarity to a certain degree. In particular, PPI view is a subnetwork from IntAct database [90] and displayed node names are actual protein IDs from UniProt database [168]. Protein Q9BRI3 is a disease protein related to zinc deficiency and it is isolated in GO view. This causes random-walk based methods, edgewise most of the embedding methods, to be vulnerable in disease gene prediction. However, multi-view platforms can make use of neighborhood of nodes even if they are isolated in one of the views. In addition, a single view is often sparse and noisy, it becomes imperative to work on multiple views *jointly* towards learning more robust representations.

In this chapter, we investigate the problem of unsupervised multi-view network embedding. That is, we learn node representations from multiple views without any labels, in order to perform a wide-range of downstream tasks, including unsupervised node clustering and network visualization, as well as supervised node classification and link prediction. Although numerous multi-view learning algorithms exist [185, 187–192], very limited research has studied multi-view network embedding [74–76, 108]. Existing approaches investigate two common principles of multi-view tasks, namely, *diversity* and *collaboration*. Diversity is based on retaining the uniqueness of each view. As each view is constructed from distinct

topological properties, different representations should be learned for them. Let us give a multi-view network example which comprises a PPI view, a GO view and a gene co-expression view as shown in Figure 5.2. Proteins a and c are connected in PPI view as they interact with each other, but they do not have GO similarity as can be observed from GO view. Thus, a and c 's representations should be close to one another in PPI view but not in GO view. Forcing the two views to adopt the same representations would disregard their diversity demonstrated in Figure 5.3(a). In terms of collaboration, as the same node in different views ultimately describes the same instance (e.g., a particular protein on a PPI network), its representations across views should be aligned (i.e., made closer/similar to each other), as illustrated in Figure 5.3(b). We call this form of collaboration the first-order, since it only concerns the representations of the same node across views.

FIGURE 5.2: A toy protein multi-view network with three views.

FIGURE 5.3: Major characteristics of a multi-view network.

Motivation for second-order collaboration. While first-order collaboration has been widely explored in state-of-the-art methods [74, 76, 108], they do not

study collaboration of a higher-order in multi-view networks. We observe that mutual reinforcement often exists between node associations in two views, which means that two nodes linked in one view tend to have a better than a random chance to also be linked in other views. Our analysis of three benchmark multi-view networks reveals that, two nodes linked in one view are 2.5–27.8 times more likely to also form links in another view than two random nodes do. In particular, this score is 27.8 for our dataset consisting of a PPI view and a GO view.

Taking Figure 5.3(c) as an example, protein d and e have GO similarity in GO view, implying that they are more likely than a random pair of proteins to also interact with each other but due to noisy PPI network they might be currently not connected in PPI view. Thus, it is expected to obtain some positive reinforcement between the two views, to enhance d and e ’s representations in both views. We call this form of collaboration the second-order, since it deals with the representations of a pair of nodes across views.

As a remark, not all multi-view networks would express second-order collaboration to the same extent. Even in some extreme cases, node associations in two views might be contradictory. Such networks are not the focus of this research, yet our proposed model can be flexibly adjusted to restrain the second-order collaboration.

It is non-trivial to explore diversity and the two orders of collaboration simultaneously, accordingly two major challenges arise.

First, second-order collaboration is directly at odds with the diversity of views. In our example above, since d and e are not currently connected in PPI view, their representations in that view should not be aligned explicitly. Otherwise, both views become uniform and diversity principle is violated. Instead, to ensure reinforcement between the two views while preserving a certain extent of the diversity, we propose a novel indirect approach. To give an illustration, in Figure 5.3(c), we “push” the representation of d in PPI view closer to that of e in GO view, and likewise

for e in PPI view and d in GO view, as indicated by the dotted lines. This way the co-existence of both diversity and second-order collaboration becomes attainable. Furthermore, second-order collaboration is also distinct from first-order collaboration. In the first order, for every node, its representations in all views are pushed closer, regardless of their associations with other nodes in each view. On the contrary, for second order, only pairs of nodes in a view that are associated in other views are involved.

Second, it is crucial to model the three characteristics in a unifying framework to enable their integration. Current multi-view network embedding approaches typically learn view-specific embeddings to attain diversity through skip-gram [76, 108] or graph factorization [75] models, whereas first-order collaboration is performed through partial parameter sharing or direct regularization. In contrast, to capture diversity and the two-orders of collaboration, we unify them with three types of consistency: intra-view-cross-node consistency, cross-view-intra-node consistency, and cross-view-cross-node consistency, respectively. We collectively call them intra-cross consistencies, each of which can be modeled by a different form of node pairs.

We introduce a novel unsupervised algorithm for **Multi-view network embedding with Intra-CROSSs consistencies (MICROS)**, which aims to learn node representations that simultaneously attain the diversity of views, achieve the first-order collaboration among nodes, and supports the second-order collaboration between views. We summarize our major contributions as follows.

1. We discover a new form of collaboration (i.e., second-order) on multi-view networks, in addition to the traditionally exploited diversity and compatibility. To the best of our knowledge, we are the first to model second-order collaboration property.

2. We develop a novel multi-view network embedding algorithm MICROS , integrating different characteristics as intra-cross consistencies in a unifying framework.
3. We conduct extensive experiments on three public datasets, and empirically demonstrate that MICROS significantly outperforms state-of-the-art competitors.

5.2 Method

We formalize the problem of unsupervised multi-view network embedding in 5.2.1, then provide a brief background on existing skip-gram-based network embedding models in 5.2.2. Later, we propose our model MICROS for unsupervised multi-view network embedding in 5.2.3.

5.2.1 Problem formulation

The main notations used throughout this study are summarized in Table 5.1. We begin with a formal definition of multi-view network and describe problem formulation right after that.

Definition 5.2.1 (Multi-view network). A multi-view network consists of a set of nodes V and a set of views U , where each view $u \in U$ contains a set of edges $E^{(u)}$ between the nodes V . Specifically, an edge $e_{ij}^{(u)} \in E^{(u)}$ joins nodes $i, j \in V$ in view $u \in U$, which may be weighted or directed depending on the application. We denote this multi-view network as $G = (V, U, \{E^{(u)} : u \in U\})$. $\square$

Given such a network G , the goal of unsupervised multi-view network embedding is to learn a low-dimensional vector representation $\mathbf{f}_i \in \mathbb{R}^D$ for each node $i \in V$, where D is the dimension of the embedding space, without any task-specific label.

TABLE 5.1: Summary of main notations.

Notation	Definition
G	a multi-view network
$V, U, E^{(u)}$	the set of nodes V , views U , edges $E^{(u)}$ in view u
$\mathbf{f}_i^{(u)}$	the center embedding of node i specific to view u
$\tilde{\mathbf{f}}_i^{(u)}$	the context embedding of node i specific to view u
$\mathbf{f}_i$	the final (output) vector representation of node i
$\Omega^{(u)}$	the set of node pairs in view u
$P_u(j i)$	probability of intra-view cross-node consistency
$P_{u'u}(i)$	probability of cross-view intra-node consistency
$P_{u'u}(j i)$	probability of cross-view cross-node consistency
Θ	the set of parameters in optimization
α, β	hyperparameters to balance different consistencies

More specifically, to maintain the diversity of views, for each node $i \in V$ and each view $u \in U$, we learn an intermediate view-specific representation $\mathbf{f}_i^{(u)} \in \mathbb{R}^{[D/|U|]}$. The final representation $\mathbf{f}_i$ is aggregated from these intermediate representations:

$$\mathbf{f}_i = \oplus_{u \in U} \mathbf{f}_i^{(u)}, \quad (5.1)$$

where $\oplus$ denotes vector concatenation. Note that our formulation is flexible to incorporate other forms of aggregation, such as taking the summation or applying an autoencoder. Investigating these alternatives is not the primary interest of this study.

Since the representations are learned in an unsupervised manner, they can be used as features in a variety of downstream tasks ranging from node classification, clustering and visualization, to relationship mining and link prediction.

5.2.2 Background on skip-gram model

We briefly recap the popular skip-gram model in single-view network embedding [66, 193, 194]. Assume a single-view network with a set of nodes V . Each node $i \in V$ is associated with a set of *context* nodes. A node, say $j \in V$, is said

to be a context node of i , if j is topologically related to i in the given network. For example, j can be a neighbor of i , or can be sampled by random walks starting from i . While j is known as the context of i , i is known as the *center* of j . Subsequently, we can predict the context j of a given center i by estimating the conditional probability $P(j|i)$, which is often defined using a softmax:

$$P(j|i) = \frac{\exp(\tilde{\mathbf{f}}_j \cdot \mathbf{f}_i)}{\sum_{v \in V} \exp(\tilde{\mathbf{f}}_v \cdot \mathbf{f}_i)}, \quad (5.2)$$

where $\tilde{\mathbf{f}}_i, \mathbf{f}_i \in \mathbb{R}^D$ is known as the context and center embedding of node i , when i serves as the context or center, respectively.

To learn the embeddings $\Theta = \{\tilde{\mathbf{f}}_i, \mathbf{f}_i : i \in V\}$, we minimize the negative log-likelihood given the set of all the observed center-context pairs Ω , as shown in Equation (5.3). The optimization can be solved with stochastic gradient descend [193]. To cope with the expensive computation of the softmax function, hierarchical softmax [193] or negative sampling [66] is often employed.

$$\Theta^* = \arg \min_{\Theta} - \sum_{(i,j) \in \Omega} \log P(j|i; \Theta). \quad (5.3)$$

Finally, the learned center embeddings, $\{\mathbf{f}_i : i \in V\}$, can be used as final node features to perform downstream tasks.

5.2.3 Proposed Approach: MICROS

We propose our model MICROS for unsupervised multi-view network embedding. First, we introduce the notion of intra-cross consistencies that unify the characteristics of diversity and both orders of collaboration. Then, we will further discuss the algorithm including the overall framework, loss function, and complexity analysis.

5.2.3.1 Intra-cross consistencies

As motivated in Section 5.1, three characteristics can be explored on multi-view networks, namely, diversity, first-order, and second-order collaboration. As shown in Figure 5.4, these major characteristics on multi-view networks can be exploited in a uniform manner, respectively through *intra-view-cross-node*, *cross-view-intra-node*, and *cross-view-cross-node* consistencies. The three types of consistency are collectively called *intra-cross consistencies*. In the following, we describe each characteristic and its corresponding consistency.

- **Diversity: Intra-view-cross-node consistency**

Views have different topological characteristics. To attain the diversity of each view, we seek the consistency within each view. More specifically, each node has a view-specific representation, and topologically related nodes in the same view should have similar representations in that view. For each view, we generate center-context pairs based on random walks just as the skip-gram model does [66, 193]. As shown in Figure 5.4, example pairs include $(b^{(1)}, c^{(1)})$ in View 1, $(a^{(2)}, d^{(2)})$ in View 2 and $(b^{(3)}, d^{(3)})$ in View 3, where the superscripts indicate the specific views from which the nodes are observed. Given that every pair is observed within a single view and across two different nodes, we name this intra-view-cross-node (ivcn) consistency.

Formally, for a given view $u \in U$, we assume a set of center-context pairs $\Omega^{(u)} \subset V \times V$. We could subsequently maximize the log-likelihood of all views individually in Equation (5.4). In this formulation, all views are optimized in a decoupled manner without interactions with each other, to retain their diversity.

$$L_{\text{ivcn}} = \sum_{u \in U} \sum_{(i,j) \in \Omega^{(u)}} \log P_u(j|i), \quad (5.4)$$

FIGURE 5.4: Overall framework of MICROS , capturing diversity and two-orders of collaboration using intra-cross consistencies on a multi-view network.

where $P_u(j|i)$ symbolizes the conditional probability of observing the context node j in view u given that the center node i has been observed in view u too. Since it involves the same view u and different nodes i, j are paired, intra-view-cross-node consistency is captured. Specifically, we define

$$P_u(j|i) = \frac{\exp\left(\tilde{\mathbf{f}}_j^{(u)} \cdot \mathbf{f}_i^{(u)}\right)}{\sum_{v \in V} \exp\left(\tilde{\mathbf{f}}_v^{(u)} \cdot \mathbf{f}_i^{(u)}\right)}, \quad (5.5)$$

where $\mathbf{f}_i^{(u)}$ or $\tilde{\mathbf{f}}_i^{(u)}$ respectively denotes the center or context embedding of node i specific to view u .

- **First-order collaboration: Cross-view-intra-node consistency**

While different views exhibit diversity, they ultimately converge on a common set of nodes. Since the same node across different views fundamentally describes the same instance, such as a particular user on a social network, its view-specific representations should be collaborating with one another. To attain such first-order collaboration, we seek the consistency on each node

across views. As shown in Figure 5.4, pairs such as $(a^{(1)}, a^{(2)})$ and $(a^{(2)}, a^{(3)})$ can be generated across different views for the same node. In each pair, the two view-specific embeddings of the same node should become similar. We term this as cross-view-intra-node (cvin) consistency, formulated as follows.

$$L_{\text{cvin}} = \sum_{u \in U} \sum_{u' \in U \setminus \{u\}} \sum_{i \in V} \log P_{u'u}(i), \quad (5.6)$$

where we further define

$$P_{u'u}(i) = \frac{\exp(\mathbf{f}_i^{(u')} \cdot \mathbf{f}_i^{(u)})}{\sum_{v \in V} \exp(\mathbf{f}_v^{(u')} \cdot \mathbf{f}_i^{(u)})}. \quad (5.7)$$

- **Second-order collaboration: Cross-view-cross-node consistency**

While a few studies on multi-view network embedding [74–76, 108] has addressed diversity and the first-order collaboration (albeit in a non-unified manner), no previous work has integrated the second-order collaboration. As motivated in Section 5.1, our key intuition is that two nodes linked by an edge in one view are often more likely than a random pair of nodes to also associate with each other in another view. Given two views $u \neq u'$, this intuition can be quantified by the following ratio:

$$\frac{P(e_{ij}^{(v')} \in E^{(v')} | e_{ij}^{(v)} \in E^{(v)})}{P(e_{ij}^{(v')} \in E^{(v')} | e_{ij}^{(v)} \notin E^{(v)})}. \quad (5.8)$$

On the three benchmark datasets used in our experiments, the ratios for all pairwise views range from 2.5 to 27.8 with a median of 7.0. That means on these networks, node pairs linked in one view are indeed more likely to form links with each other in another view. Thus, the second-order collaboration is a fair assumption on many multi-view networks, although it is not universal.

As we shall see later, our model is flexible to adjust the contribution from the second-order collaboration.

Using Figure 5.4 as an example, nodes d and e are topologically related in view 2 but not in view 1. Under the second-order collaboration, the two nodes may entail a latent, unknown association in view 1, which is not currently manifested, possibly due to incomplete and noisy data, or a time delay—they may develop an explicit association in future. It is important to note that this assumption merely suggests a tendency. Thus, it is not appropriate to directly or explicitly “push” d and e ’s representations closer in view 1. In fact, doing so makes all views uniform, which is a complete contradiction of the diversity requirement.

To encourage the second-order collaboration between views whilst retaining the diversity, we propose an implicit mechanism to enable such collaboration across views. As shown in Figure 5.4, we consider pairs like $(d^{(1)}, e^{(2)})$ and $(d^{(2)}, e^{(1)})$. That is, we bias d ’s representation in view 1 towards e ’s in view 2, and symmetrically, bias e ’s representation in view 1 towards d ’s in view 2. This implicit approach enables the node representations from two views to mutually reinforce each other, and simultaneously allows for views’ diversity since we are not directly forcing d and e ’s representations in view 1 to become closer. Since each pair involves two different nodes observed across two views, we call it cross-view-cross-node (cvcn) consistency. Formally, we maximize the following log-likelihood.

$$L_{\text{cvcn}} = \sum_{u \in U} \sum_{u' \in U \setminus \{u\}} \sum_{(i,j) \in \Omega(u)} \log P_{u'u}(j|i), \quad (5.9)$$

where it is further defined that

$$P_{u'u}(j|i) = \frac{\exp\left(\tilde{\mathbf{f}}_j^{(u')} \cdot \mathbf{f}_i^{(u)}\right)}{\sum_{v \in V} \exp\left(\tilde{\mathbf{f}}_v^{(u')} \cdot \mathbf{f}_i^{(u)}\right)}. \quad (5.10)$$

FIGURE 5.5: Intra-cross consistencies in four quadrants.

In other words, for each pair of nodes within a single view u , say $(i, j) \in \Omega^{(u)}$ (which means nodes i and j are topologically connected in view u), we exploit their second-order collaboration through biasing j 's representations in another view u' towards i 's representation in view u . In this formulation, we are effectively generating a cross-view-cross-node pair $(i^{(u)}, j^{(u')})$.

We summarize the three types of intra-cross consistencies in Figure 5.5 using a quadrant diagram. On each axis (view or node), as we progress from “intra” to “cross”, data become more complex and more characteristics can be leveraged. Interestingly, the intra-view-intra-node quadrant corresponds to traditional machine learning on single-view non-network data.

5.2.4 The MICROS algorithm

Based on the unification with intra-cross consistencies, we further present the proposed approach.

5.2.4.1 Overall framework

We outline the overall framework of MICROS in Figure 5.4. Given a multi-view network as input, different types of pairs are generated to materialize the three

types of intra-cross consistencies. These pairs are then used to optimize the intermediate view-specific node embeddings, which are further aggregated to produce the final node embeddings as output. The process is unsupervised, without any application-specific label given. In particular, our framework unifies the materialization of diversity and the two orders of collaboration, since all of them boils down to maximizing the probabilities of the generated pairs.

5.2.4.2 Loss function

Our overall loss function integrates the three types of intra-cross consistencies, to optimize the likelihood of our parameters $\Theta = \{\tilde{\mathbf{f}}_i^{(u)}, \mathbf{f}_i^{(u)} : (i, u) \in V \times U\}$ given all the generated pairs. Specifically, we minimize the following objective function, which builds upon three different kinds of generated pair.

$$\Theta^* = \arg \min_{\Theta} -L_{\text{ivcn}} - \alpha \cdot L_{\text{cvin}} - \beta \cdot L_{\text{cvcn}}, \quad (5.11)$$

where $\alpha, \beta \geq 0$ are hyperparameters to control the relative importance among the three components. Note that while we can adjust α and β , the three types of consistencies are not at a comparable scale. In particular, L_{cvin} in Equation (5.6) sums over all nodes V once, whereas Equations (5.4) and (5.9) sum over the pairs $\Omega^{(u)}$ in each view. Given that $|\Omega^{(u)}| \gg |V|$ in general, their scales vastly differ. We can rewrite Equation (5.6) as follows:

$$L_{\text{cvin}} = \sum_{u \in U} \sum_{u' \in U \setminus \{u\}} \sum_{(i, *) \in \Omega^{(u)}} \log P_{u'u}(i). \quad (5.12)$$

That is, we sum over each of the center node observed in every pair. Thus, a frequently observed center node, which participates more in L_{ivcn} and L_{cvcn} , will also participate more in L_{cvin} , making the three consistencies comparable in scale.

As a further remark, when $\alpha = \beta = 0$, MICROS reduces to skip-gram models such as DeepWalk [193] and node2vec [66] on individual views in a decoupled manner, without leveraging any collaboration between views. When $\alpha > 0$ and $\beta = 0$, MICROS only captures diversity and first-order collaboration, which becomes similar in spirit, but not identical, to MVE [76], MVN [74] and mvn2vec [108]. In particular, they model diversity and collaboration in non-unifying manners—skip-gram model for diversity, and partial parameter sharing or regularization for collaboration.

5.2.4.3 Algorithm

Following previous work [66, 193], we employ stochastic gradient descent and negative sampling to optimize the loss function. Our approach is summarized in Algorithm 1. We first initialize the parameters randomly (line 1). Subsequently, for each view u , we sample random walks starting from every center node, and further generate center-context node pairs $\Omega^{(u)}$ within view u (line 2-4). Note that $\Omega^{(u)}$ consists of the intra-view-cross-node pairs only. In the next block, we perform gradient updates on the parameters (line 5–12). In particular, we exploit different formulations of node pairs and view pairs, to optimize the three types of intra-cross consistencies (line 9, 11 and 12, respectively). Finally, we concatenate the view-specific center embeddings of each node to obtain the final embedding (line 13-14).

5.2.4.4 Complexity analysis

Based on existing studies [66, 76], learning view-specific embeddings over all views takes $O(|E| \cdot D/|U| \cdot K)$ time, where $|E| = \sum_{u \in U} |E^{(u)}|$ denotes the total number of edges in all views, $\lfloor D/|U| \rfloor$ is the dimension for view-specific embeddings, and K is the number of negative samples. This is equivalent to the time complexity of

Algorithm 1: MICROS

Input : $G = (V, U, \{E^{(u)} : u \in U\})$, α , β **Output**: final embeddings $\{\mathbf{f}_i : i \in V\}$

```
1 Randomly initialize embeddings;
2 for each view  $u \in U$  do
3    $R^{(u)} = \text{SAMPLERANDOMWALKS}(V, E^{(u)})$ ;
4    $\Omega^{(u)} = \text{GENERATEPAIRS}(R^{(u)})$ ;
5 while not converged do
6   for each view  $u \in U$  do
7     Shuffle pairs in  $\Omega^{(u)}$ ;
8     for each pair  $(i, j) \in \Omega^{(u)}$  do
9       Update  $\tilde{\mathbf{f}}_j^{(u)}, \mathbf{f}_i^{(u)}, \tilde{\mathbf{f}}_u^{(u)}$  based on  $L_{\text{ivcn}}$ ;
10      for each view  $u' \in U \setminus \{u\}$  do
11        Update  $\tilde{\mathbf{f}}_i^{(u')}, \mathbf{f}_i^{(u)}, \tilde{\mathbf{f}}_u^{(u')}$  based on  $L_{\text{cvin}}$ ;
12        Update  $\tilde{\mathbf{f}}_j^{(u')}, \mathbf{f}_i^{(u)}, \tilde{\mathbf{f}}_u^{(u')}$  based on  $L_{\text{cvcn}}$ ;
13 for  $i \in V$  do
14    $\mathbf{f}_i = \oplus_{u \in U} \mathbf{f}_i^{(u)}$ ;
```

intra-view-cross-node consistency (line 9 in Algorithm 1). On the other hand, the time complexities for cross-view-intra-node and cross-view-cross-node consistencies (line 11–12) are both $O(|E| \cdot D / |U| \cdot K \cdot |U|)$. Thus, the overall time complexity is $O(|E| \cdot D \cdot K)$. Thus, MICROS scales linearly in the total number of edges in all views.

5.3 Results and Discussion

To evaluate the performance of our proposed MICROS method, we conducted extensive experiments on not only biological networks but also social networks, three multi-view networks in total, through various downstream tasks. Results demonstrate the superior performance of MICROS against a comprehensive suite of state-of-the-art baselines. We perform additional model analysis to understand the various aspects of our model, including an ablation and a scalability study, a convergence analysis, as well as an investigation on the impact of parameters.

TABLE 5.2: Summary of datasets.

Dataset	# Core nodes	# Edges in each view		
IntAct	12 901	PPI 96 845	GO 107 508	
LinkedIn	10 197	College 1 527 681	Employer 1 765 927	Friend 29 434
YouTube	7 558	Friend 1 003 923	Subscriber 1 547 091	Video 1 918 208

Outline of this section is as follows: first we introduce our experimental setting, second we describe evaluation tasks employed for performance comparison, lastly we present model analysis.

5.3.1 Experimental setup

5.3.1.1 Datasets

Table 5.2 summarizes three multi-view networks used in our experiments. On each dataset, we worked with the set of nodes that participate in at least one view, known as the core nodes. We describe each dataset below.

- **IntAct**, a protein network includes two views, namely, PPI and GO. The PPI view captures the physical protein-protein interactions based on the IntAct database [90], whereas the GO view captures the functional associations between proteins based on their Gene Ontology (GO) annotations [186]. In construction of GO view, firstly, we mapped the protein names to their producer gene names by using the converter tool of UniProt database [168]. Then, we calculated the pairwise functional similarity between proteins based on the study [195] (G-SESAME¹), and subsequently, constructed GOs with respect to $\mathcal{K}$ -th nearest neighbors for different $\mathcal{K}$ over $\mathcal{K} = \{5, 10, 20, 50\}$. We evaluated the performance of GOs based on their single view node representations.

¹G-SESAME: <http://bioinformatics.clemson.edu/G-SESAME/>

Throughout this study, we conducted our IntAct experiments with $\mathcal{K} = 10$ which has the best performance.

- **LinkedIn** [196], a user professional network with three views, namely, college, employer and friend, extracted from LinkedIn. The college view is based on users' education history and the edges in this view indicate common colleges. Similarly, the edges in the employer view indicate common workplace and those in the friend view represent the friendship in LinkedIn.
- **YouTube**[197] [198], a user video-sharing network with three views, namely, friend, subscriber and video. In these three views, two users have an edge if they have shared friends, shared subscribers and shared favorite videos, respectively.

5.3.1.2 Baselines

We compare MICROS against several common strategies and state-of-the-art algorithms for multi-view network embedding.

- **Single:** A Skip-gram model equivalent to DeepWalk [193] applied to a single view which corresponds to the lines up to 9 in the afore-mentioned Algorithm 1. Our version uses negative sampling. For each task, we run the model on every single view, and report the average performance of all views.
- **Decoupled:** Each view is independently trained by the single-view Skip-gram model, with the number of dimensions set to $\lfloor D/|U| \rfloor$. All views' embeddings are then concatenated to form the final embedding.
- **Merged:** All views are merged into one network by taking the union of the edges in all views, as if there is only a single view. The single-view Skip-gram model is then trained on the merged network.

- **MVE** [76]²: A state-of-the-art multi-view network embedding approach. We adopt their unsupervised version to match the setting of MICROS. It tries to promote the (first-order) collaboration between views by regularizing the Euclidean norm between view-specific embeddings and the final embeddings. A parameter η is used to control the weight of regularization.
- **mvn2vec** [108]: A state-of-the-art multi-view embedding approach with two variants, namely, **mvn2vec-con** and **mvn2vec-reg**. To enforce the first-order collaboration, mvn2vec-con variant employs partial parameter sharing across views, with a hyperparameter θ to adjust the trade-off between the sharing and views' diversity. On the other hand, mvn2vec-reg regularizes the Euclidean norm between view-specific embeddings directly, controlled by a hyperparameter γ . When γ is set to zero, it becomes the same model as MICROS with only diversity version. We implement mvn2vec in the same way with MICROS .
- **MNE** [74]³: A state-of-the-art multiplex network embedding algorithm. For each view, the model learns a final embedding which is composed of three parts: common embedding, relation-based embedding, and a transformation matrix. The first-order collaboration is achieved through partially sharing the common embedding. There is a hyper-parameter r to control the Frobenius norm of the transformation matrix.

5.3.1.3 Parameter configuration

To sample random walks for Skip-gram-based methods, we apply the same settings: walk length of 10, 5 walks per node, windows size of 3, and 10 samples per negative sampling. All methods adopt $D = 128$ as the dimensions of the final embeddings. During training, we run 10 epochs with a batch size of 256. For hyper-parameters

²MVE: <https://github.com/mnqu/MVE>

³MNE: <https://github.com/HKUST-KnowComp/MNE>

of the baselines, we set $\eta = 0.05$ for MVE, $\theta = 0.8$ for mvn2vec-con and $\gamma = 0.01$ for mvn2vec-reg, and $r = 1000$ for MNE based on their empirical performance. For our method, we set $\alpha = \beta = 1$ as the default to have an equal weight between the three types of intra-cross consistencies. Nonetheless, we also conducted a parameter sensitivity study to show the impact of α and β on the performance in Section 5.3.3.

5.3.2 Baseline comparison on evaluation tasks

All algorithms are evaluated with a series of tasks including (i) node-level tasks on node classification, visualization and clustering, and (ii) edge-level tasks on relationship mining and link prediction. We defer the description of each task until we encounter them.

5.3.2.1 Algorithms and metrics for tasks

For all classification-based tasks (*i.e.*, all tasks except node visualization and clustering), we trained a logistic regression model under five-fold cross-validation with five repetitions. For binary classification, we used the areas under the ROC and Precision-Recall curve (ROC-AUC and PR-AUC) as evaluation metrics; for multi-class settings, we used micro- and macro-averaged F-scores (micro-F and macro-F) as well as accuracy. For node visualization, we adopted the popular *t*-SNE algorithm [199]. For node clustering, we employed the K-means algorithms with 100 random initializations, and evaluated the results across purity, normalized mutual information (NMI) and rand index.

5.3.2.2 Node-level tasks

We performed three node-level tasks on IntAct, *i.e.*, node classification, visualization and clustering, given the availability of node-level labels.

TABLE 5.3: Performance of node classification tasks on IntAct (bold: best; underline: runner-up).

Disease	Metric	Single	Decoupled	Merged	MVE	mvn2vec-con	mvn2vec-reg	MNE	MICROS
Alzheimer's	ROC-AUC	0.6933	0.8507	0.7908	0.6581	0.8759	0.8944	<u>0.9357</u>	0.9704
	PR-AUC	0.0191	0.0822	0.0103	0.0116	0.1297	0.0312	0.2078	<u>0.2063</u>
breast cancer	ROC-AUC	0.8523	0.8836	0.8435	0.8487	0.8802	0.9064	<u>0.8951</u>	0.9154
	PR-AUC	0.0486	0.1035	0.0421	0.0772	0.0773	0.1010	<u>0.1063</u>	0.1629
colorectal cancer	ROC-AUC	0.8258	0.8988	0.8327	0.8552	<u>0.9269</u>	0.9076	0.8649	0.9328
	PR-AUC	0.0432	0.0923	0.0806	0.0756	<u>0.1407</u>	0.1795	0.1096	0.1255
lung cancer	ROC-AUC	0.7090	<u>0.7341</u>	0.6732	0.5626	0.7019	0.6174	0.6995	0.7552
	PR-AUC	0.0081	0.0233	0.0059	0.0173	0.0180	0.0061	<u>0.0305</u>	0.0665
average	ROC-AUC	0.7701	0.8418	0.7850	0.7311	0.8462	0.8314	<u>0.8488</u>	0.8935
	PR-AUC	0.0297	0.0753	0.0347	0.0455	0.0914	0.0795	<u>0.1136</u>	0.1403

FIGURE 5.6: Visualization of protein nodes: alzheimer (■), breast cancer (▲), colorectal cancer (●), and lung cancer (★).

TABLE 5.4: Performance of node clustering task on IntAct (bold: best; underline: runner-up).

Metric	Single	Decoupled	Merged	MVE	mvn2vec-con	mvn2vec-reg	MNE	MICROS
Purity	0.4324	0.4279	0.4214	0.4117	<u>0.4381</u>	0.4275	0.4340	0.4711
NMI	0.0683	0.1020	0.0586	0.0428	0.1020	0.0992	<u>0.1269</u>	0.1523
Rand	0.0207	0.0242	0.0138	0.0194	0.0310	0.0253	<u>0.0415</u>	0.0679

Our node classification deals with the identification of disease-causing proteins. Given a specific disease (e.g., breast cancer), each protein node has a binary label to show whether it is the causative protein for this disease. In our experiments, we worked on four diseases, namely, Alzheimer's disease (the number of positive proteins for this disease: 11), breast cancer (24), colorectal cancer (34), lung cancer (15). Note that the classes are highly skewed with very few positive nodes for each disease. Thus, we have also tried to over sample with SMOTE [181], although it yielded similar results.

Table 5.3 reports the performance on the binary classification of each disease. In terms of ROC-AUC, MICROS achieves the best performance on all diseases. In

TABLE 5.5: Performance of relationship mining on LinkedIn and link prediction on YouTube (bold: best; underline: runner-up).

Dataset	Metric	Single	Decoupled	Merged	MVE	mvn2vec-con	mvn2vec-reg	MNE	MICROS
LinkedIn	Micro-F	0.3797	0.4110	0.3448	0.4030	0.4179	0.4233	<u>0.4282</u>	0.4522
	Macro-F	0.3295	0.3611	0.3052	0.2870	0.3636	<u>0.3655</u>	0.3460	0.3900
	Accuracy	0.4837	0.5200	0.4578	0.5225	0.5139	0.5283	<u>0.5367</u>	0.5552
YouTube	ROC-AUC	0.6331	0.6700	0.6564	0.6279	0.6635	0.6704	<u>0.6749</u>	0.6918
	PR-AUC	0.1564	0.1650	<u>0.1764</u>	0.1650	0.1621	0.1670	0.1604	0.2038

terms of PR-AUC, MICROS is the best or a close runner-up on all but one disease. Among the baselines, no consistent winner emerges, with decoupled, mvn2vec and MNE being generally competitive. Overall, MICROS attains the highest average ROC-AUC and PR-AUC over all the diseases, outperforming the runner-up by 5% and 24%, respectively.

We further visualize the embeddings generated by MICROS and four competitive baselines in Figure 5.6. We showcase the separation of the protein nodes based on their disease labels. Observe that the plots generated by Decoupled, MVE and mvn2vec-con are visibly inferior, where the four diseases are mixed with each other without forming clusters. The plot by MNE is somewhat improved with a dense cluster for alzheimer (left), but the other three diseases are still quite inseparable. In contrast, MICROS forms relatively dense clusters especially for alzheimer (top right), lung cancer (top left) and breast cancer (bottom left).

To obtain a more quantitative evaluation, we also conducted node clustering over all the disease proteins, hoping that each disease would form one cluster. We present the results in Table 5.4 on the five algorithms used in the visualization. Results show that MICROS attains the best result in all clustering metrics, significantly outperforming the runner-up by 8% or more.

5.3.2.3 Edge-level tasks

Edge-level tasks involve a pair of nodes, and we concatenate their embeddings as the final features for the task. We consider two tasks, as follows.

- Relationship mining [200], to uncover latent relationship types between two nodes on the network. Each pair of nodes has a label to indicate their relationship, such as advisor/advisee, professional contact, or ex-colleague. Note that two nodes might be connected on a network, but the type of their relationship is unknown. We evaluated relationship mining on LinkedIn, where there are eight classes of relationship and is thus a multi-class problem. The relationship classes are latent, hidden from the three views on LinkedIn.
- Link prediction, to predict links that are most likely to form, and can be formulated as a binary classification problem. On YouTube, we performed link prediction using a contact network as the ground truth for the links. Note that the friend view excludes all the contacts from the contact network. To provide negative instances, we randomly selected five times as many non-contact core users.

We report the results in Table 5.5. MICROS outperforms all the baselines on both datasets. Note that we apply different metrics based on their binary or multi-class nature. In particular, MICROS significantly outperforms the runner-up by 3–7% on LinkedIn and 3–16% on YouTube across various metrics. It is also worth noting that there is no consistent winner among the baselines, with Decoupled, mvn2vec and MNE emerge as relatively competitive methods.

5.3.3 Model analysis

We conduct further experiments to investigate various aspects of MICROS through a series of studies on ablation, scalability, convergence and parameter sensitivity.

FIGURE 5.7: Comparison of MICROS with two simpler variants.

5.3.3.1 Ablation study

Since MICROS integrates diversity and two orders of collaboration, we further examine the contribution of each characteristic. Specifically, we compare MICROS with the following two variants:

- MICROS $\setminus C^1 C^2$, *i.e.*, without both orders of collaboration (*i.e.*, $\alpha = \beta = 0$), which leverages diversity only;
- MICROS $\setminus C^2$, *i.e.*, without the second-order collaboration (*i.e.*, $\alpha = 1$ and $\beta = 0$), which leverages diversity and the first-order collaboration only.

We evaluate the above variants on the three datasets, for node classification, relationship mining and link prediction, respectively. As shown in Figure 5.7, MICROS outperforms MICROS $\setminus C^2$ consistently, validating that the novel second-order collaboration is indeed beneficial for multi-view networks. Furthermore, MICROS $\setminus C^2$ also consistently outperforms MICROS $\setminus C^1 C^2$, as expected. Finally, the gradual improvements from the simplest to the most sophisticated model also imply that our unification of diversity and collaboration with intra-cross consistencies works well.

FIGURE 5.8: Scalability analysis of MICROS .

5.3.3.2 Scalability study

To investigate the scalability of MICROS, we sampled smaller multi-view networks from the original datasets. In Figure 5.8, we plot the time taken per epoch of training *w.r.t.* the total number of edges in all views. Note that the scalability experiments were conducted on the same standalone machine using a PyTorch-based implementation. We observe that MICROS scales linearly in the total number of edges, which is consistent with our complexity analysis in Section 5.2.4.4.

5.3.3.3 Convergence analysis

We further analyze the convergence property of MICROS on the three datasets. Figure 5.9(a) shows that our proposed objective function can converge quickly, typically between 5 and 8 epochs on all datasets. Likewise, Figure 5.9(b) demonstrates that the evaluation scores (ROC-AUC for IntAct and YouTube, Micro-F for LinkedIn) begin to converge from the 5th epoch as well. The fast convergence implies that the three types of consistencies work well together, and MICROS can be efficiently trained.

FIGURE 5.9: Convergence analysis of MICROS.

5.3.3.4 Weights α and β

We investigate the choice of α and β in Equation (5.11). The parameter α controls the contribution of the cross-view-intra-node consistency (*i.e.*, first-order collaboration), whereas β controls the contribution of the cross-view-cross-node consistency (*i.e.*, second-order collaboration). By adjusting their values, we trade off the importance between different characteristics. We first vary $\alpha \in \{0.1, 0.5, 1, 2, 5\}$ while setting $\beta = 0$. We observe that the performance is generally robust in the range $[0.5, 2]$ (Figure 5.10 and Figure 5.11). Very large values such as $\alpha = 5$ may result in performance deterioration due to over-emphasis on the first-order collaboration and less importance on the diversity. Next, we fix α at its optimal value, and vary β . A similar conclusion can be drawn, where the range $[0.5, 2]$ generally gives stable outcomes, and large values like $\beta = 5$ would often hurt the performance.

5.3.3.5 Number of embedding dimensions D

Lastly, we vary the dimension of the final embeddings, D , as shown in Figure 5.12. It can be concluded that MICROS performs well when $D \in \{128, 256\}$ on all datasets. However, when the number of dimensions is very small (*e.g.*, $D = 32$),

FIGURE 5.10: Impact of hyperparameter α .

FIGURE 5.11: Impact of hyperparameter β .

FIGURE 5.12: Impact of dimension D .

the performance suffers significantly, since the view-specific embeddings would become too small (which is set to $\lfloor D/|U| \rfloor$) to capture meaningful representations. In particular, relationship mining on LinkedIn appears to be the most sensitive task to the small embedding dimension. This could be explained by the large number of relationship categories which might be difficult to model without sufficient dimensions. On the other hand, when the number of dimensions becomes large (*e.g.*, $D = 512$), the performance may also deteriorate, due to the increase in the number of learnable parameters.

5.3.3.6 Methods for Combining the Final Embeddings

In this study, apart from the concatenation of the final embeddings from multiple views, we investigate various integration methods including the auto-encoder (AE), the Hadamard product, and the summation. The auto-encoder we use consists of two linear layers and a sigmoid layer right after the linear layers. We feed the concatenation of the embeddings into the auto-encoder and extract the features for the classifier. The Hadamard product is simply element-wise multiplication of two embeddings and the summation is obtained by the element-wise addition of the embeddings. Table 5.6 shows their performance comparison analysis for IntAct dataset. All results are performed by the same parameter setting as described in “Parameter configuration” (Section 5.3.1.3).

Since the concatenation achieves the highest prediction performance, it is the preferred method for combining the final embeddings of multiple views before the prediction performance analysis in this study.

TABLE 5.6: Performance of varying integration methods in disease gene prediction.

Disease	Metric	Sum	AE	Concat	Hadamard
Alzheimer's	ROC-AUC	0.6100	0.6499	0.8507	0.7745
	PR-AUC	0.0029	0.0042	0.0822	0.0052
Breast Cancer	ROC-AUC	0.9123	0.6127	0.8836	0.7538
	PR-AUC	0.0669	0.0036	0.1035	0.0066
Colorectal Cancer	ROC-AUC	0.8611	0.7367	0.8988	0.6629
	PR-AUC	0.0588	0.0099	0.0923	0.0055
Lung Cancer	ROC-AUC	0.7386	0.5905	0.7341	0.6156
	PR-AUC	0.0111	0.00286	0.0233	0.0027
Average	ROC-AUC	0.7805	0.6475	0.8418	0.7017
	PR-AUC	0.0349	0.0052	0.0753	0.0050

5.3.4 Further Analysis on the Effect of Second-order Collaboration in Disease Gene Prediction

To understand the contribution of the novel *second-order collaboration* in disease gene prediction problem, we perform an additional ablation study (as in Section 5.3.3.1) and demonstrate AUC performances across seven diseases in Figure 5.13. We observe that MICROS outperforms both $\text{MICROS} \setminus C^2$, $\text{MICROS} \setminus C^1C^2$ consistently, validating that the novel second-order collaboration is effective in disease gene prediction. However, *first-order collaboration*, $\text{MICROS} \setminus C^2$, is prominently harmful for lung cancer and prostate cancer by 4% and 20%, respectively. It is hard to derive an overall conclusion due to the data imbalance issue, however we compare K-means ($K = 10$) performance of $\text{MICROS} \setminus C^2$ and MICROS embeddings to understand the drastic effect on prostate cancer. The number of true positives of $\text{MICROS} \setminus C^2$ is four proteins less than the MICROS. When we further investigate these four mis-clustered proteins of *first-order collaboration*, we realize that this method could be less effective to identify the proteins with the biological processes displayed in Table 5.7. Additionally, one protein with *immunoglobulin domain* and one protein with *transmembrane helix* domain are mis-identified in prostate cancer analysis.

FIGURE 5.13: Ablation study to demonstrate the contribution of second-order collaboration through seven diseases.

TABLE 5.7: Biological processes of the proteins, which *first-order collaboration* is insufficient to identify in prostate cancer analysis.

Biological Processes	# of proteins
Immune system process	3
Homophilic cell adhesion	1
Reproductive process	1
Leukocyte migration	1

5.3.5 Comparison among Proposed Studies: Metagraph+, N2VKO, MICROS

In this thesis we present our Metagraph+ representations and N2VKO representations in Chapter 3 and Chapter 4, respectively. Previously we compare these two studies in Section 4.3.5 “Comparison with State-of-the-arts” under the same experimental settings, i.e., the same feature selection (FS) and oversampling (O) procedure (FSO) is applied and the five-fold cross validation on GLM classifier is employed. Note that, originally N2VKO representations are tuned across $p, q \in \{0.25, 0.5, 1, 2, 4\}$ values. However, for MICROS representations the sampled walks have $p, q = 1$ parameter setting. Thus, for a reasonable comparison we repeated N2VKO with $p, q = 1$ setting. In Figure 5.14, we represent it with N2VKO. Besides, N2VKO feature representations consist of node2vec embeddings on a PPI network and *keywords* feature vectors, however MICROS is performed on the PPI network and the GO protein network described in Section 5.3.1.1. To maintain consistency, we perform N2VGO, i.e., we employ node2vec on the GO

FIGURE 5.14: AUC comparison among proposed methods. Unlike the proposed N2VKO, random walk parameters p and q are not tuned for a reasonable comparison between N2VKO and MICROS. Besides, the dataset of MICROS consists of GO network and IntAct PPI interactions. For a better comparison we perform N2VGO, i.e., the same feature selection and oversampling procedure of N2VKO is applied on the concatenated node2vec embeddings from these two networks.

protein network and the PPI network individually and perform FSO on the concatenated embeddings of these two networks. Eventually, we applied the same FSO on MICROS embeddings and demonstrate it in Figure 5.14 as MICROS. MICROS embeddings outperforms the runner-up, N2VKO, by 2% in disease gene prediction. This shows the power of multi-view algorithm, which achieves the highest AUC performance without the support of *keywords* feature vector concatenation. Note that, in this analysis a more recent OMIM (2018) database is used for the known disease proteins.

5.4 Summary

In this chapter, we investigated a critical problem of *unsupervised multi-view network embedding*. Our approach MICROS not only utilizes the traditional characteristics of diversity and (first-order) collaboration, but also leverages on our newly discovered higher-order collaboration. In particular, the second-order collaboration is materialized through an implicit mechanism so as to retain diversity. Furthermore, to integrate diversity and both orders of collaboration in a principled manner, we proposed a novel framework of intra-cross consistencies, which systematically

unifies each characteristic as one type of consistencies. Finally, we evaluated the performance of our proposed approach on three real-world multi-view networks on a wide range of supervised and unsupervised evaluation tasks, including node classification, visualization and clustering, as well as relationship mining and link prediction. Our extensive experiments demonstrate that MICROS outperforms several state-of-the-art competitors consistently and significantly on all tasks.

Chapter 6

Conclusions and Future Directions

In this chapter, we briefly explain our research contributions included in the Ph.D. thesis and subsequently present our possible future directions.

6.1 Conclusions and Discussion

In this research, we work on providing robust and reliable representations for proteins on network data to improve disease-gene prediction performance. We initially propose our novel metagraph representations, which is, differently from the traditional network motifs and graphlet kernels, able to capture not only topological characteristics of the proteins but also ontological associations between protein and term pairs. Afterwards, we provide representations including both node embeddings extracted from network data and biological annotations consist of several functional and cellular features. Lastly, we design a multi-view embedding framework to enable preserving the semantics and topological structures of each view/network, while learning the embeddings.

In Chapter 3, we describe our first work *Metagraph* which is a representation-based approach on a novel heterogeneous single network. In this work, we provide two types of representations *Metagraph* and *Metagraph+*. *Metagraph* representations are extracted from heterogeneous subgraphs including both physical interaction of proteins and *protein-biological term* associations. Whereas, *Metagraph+* provides extended representations, which also include disease gene co-occurrences within these subgraphs. We observe that extended representations achieve higher prediction performance. Of course, while computing the disease gene co-occurrences we consider the disease genes in the train set. This result shows us, disease genes tend to occur within the same metagraphs (i.e., heterogeneous subgraphs). As most of the module-based approaches, which are based on network motifs or graphlet kernels, we work with subgraphs containing up to five nodes due to the computational cost.

In Chapter 4, we present our novel integrative framework *N2VKO* for disease gene prediction problem. In this study, we utilize both human PPI networks and biological aspects of genes and predict disease-gene associations. In particular, we learn node embeddings from the PPI network using Skip-gram architecture and use the *keywords* from UniProt database as properties of genes. To the best of our knowledge, it is the first attempt of using node embeddings for candidate disease gene prediction. Then, we apply various feature selection techniques (i.e., *N2VK* representations are obtained). Among them, overall, MRMR and VarImp demonstrate the highest prediction performances. Afterwards, we examine two different oversampling techniques: ROSE and SMOTE (i.e., final *N2VKO* representations). Overall, ROSE attains slightly higher performance than SMOTE. Finally, we evaluate the performance of the *N2VKO* representations through various binary classification models.

We observe that our *N2VKO* representations perform better than feature selection and oversampling applied node2vec embeddings (i.e., *N2V-FSO*). Furthermore, we

compare our *Metagraph+* representations (i.e., same FS and oversampling procedure is applied) with *N2VKO* embeddings. *N2VKO* representations achieve better prediction performance. It is worth noting that, our same feature selection and oversampling procedure increases the performance of *Metagraph+* representations as well. In addition, the success of *N2VKO* compared to *Metagraph+* could show the superiority of the diffusion-based methods over module-based methods as pointed out in several studies [27, 29, 56, 57]. We also examine the contribution of the keywords in the selected feature sets and realize that *biological process* has the highest percentage in *N2VKO* representations.

However, we observe that *N2VKO* attain lower prediction performance on diseases such as obesity, prostate cancer and diabetes mellitus due to lower interconnectedness among known disease genes, that is one-hop neighbors of disease genes exhibit high sparsity (i.e., low clustering coefficient scores). This shows the weakness of diffusion-based models concerning the distribution of disease genes in the interactome. We discuss this issue in Section 6.2.

In Chapter 5, we describe our recent unsupervised multi-view network embedding algorithm, *MICROS*, which learns robust and reliable representations to be used on a wide range of supervised and unsupervised evaluation tasks, including node classification, visualization, and clustering, as well as relationship mining and link prediction. To the best of our knowledge, this is the first attempt of building a multi-view network embedding algorithm for disease gene prediction. Existing studies focus on social network analysis such as MVE [76], MNE [74], and MVNE [75]. However, our approach *MICROS* not only utilizes the traditional characteristics of diversity and (first-order) collaboration, but also leverages on our newly discovered higher-order collaboration. In particular, the second-order collaboration is materialized through an implicit mechanism so as to retain diversity. Furthermore, to integrate diversity and both orders of collaboration in a principled manner, we proposed a novel framework of intra-cross consistencies, which

systematically unifies each characteristic as one type of consistencies.

For disease-gene prediction problem, *MICROS* learns representations, which are not only suitable for supervised models but also qualified for unsupervised evaluation by capturing the label information of the nodes/proteins naturally. For this specific problem, we work on two views: protein-protein network based on gene ontology (GO) semantic similarity scores and PPI network. *MICROS* embeddings outperform the state-of-the-art multi-view network embedding approaches on clustering tasks including quantitative, visual as well as supervised classification. The most important contributions of this model are: i) learns embeddings while preserving each view’s semantics and topological structure, ii) provide automated features without arduous hand engineering and domain expertise.

6.2 Future Directions

In this section, we present our possible solutions and future directions to address the current issues for prediction of candidate disease genes.

6.2.1 Network reconstruction - link prediction

Link prediction in biomedical networks has various roles such as PPI prediction and drug-target interactions (DTI) prediction [103, 201, 202]. The goal of the link prediction task is predicting missing interactions or possible interactions that may occur in the future in non-static networks. Presented *network embedding algorithm MICROS*, in Chapter 5, can serve as a useful tool for link prediction on biomedical graphs. A recent study [203] explores the effectiveness of network embedding models including LINE [69], node2vec [66], DeepWalk [68] and SDNE [155] on DTI and PPI prediction via a neural link predictor. It shows that these methods outperform the baselines including Jaccard Index and Adamic-Adar. Since

FIGURE 6.1: Unlike the social networks, in a PPI network, proteins with many number of shared protein neighbors do not necessarily interact with each other (X and Y). Instead, proteins occurred together in many $l = 3$ paths, are more likely to interact and form a link (red link). The figure is adopted from the study [4].

MICROS embeddings are known to provide more robust and reliable embeddings than these methods, MICROS could be a preferred framework for the link prediction task. Another survey [204] compares several network embedding methods such as random walk-based DL method node2vec [66], autoencoder-based DL method SDNE [155] and matrix factorization-based method HOPE [205] (i.e., designed for directed graph). HOPE and SDNE seem to be the most promising ones among these methods. However, they do not investigate the contribution of additional biological data such as GO in PPI prediction. Since MICROS is able to collaboratively integrate multiple networks during embedding learning, it may attain higher prediction performance.

It is also worth noting that, a very recent study [4] reveals that proteins interact not if they are similar to each other, but if one of them is similar to the other's partners. That is, unlike the social networks, in a PPI network, proteins with many numbers of shared protein neighbors do not necessarily interact with each other (Figure 6.1). Instead, proteins occurred together in several 3-link distance ($l = 3$) paths, are more likely to interact and form a link. This information could be very useful to increase the prediction performance of random walk-based approaches.

6.2.2 Attention-based end-to-end multi-view network embedding

Attention mechanism, in general terms, allows a model to weigh the contribution of its components during the learning phase. For instance, a recent study [206] on compound protein interaction (CPI) prediction, adapts an attention mechanism to analyze which subsequences in a protein are decisive for a drug compound to interact with the protein.

Attention mechanism has gained popularity [76, 206, 207] with increased interest in multi-view approaches as well. We are currently working on a system to further improve the prediction performance of MICROS with an adopted attention mechanism to weigh the contributions of the views in disease-gene prediction. For instance, on a multi-view system, consists of PPI and GO networks, the relative importance of the networks can be learned simultaneously with the embeddings so that we can learn more robust representations. This way, we can also make an inference on which network is more useful for candidate disease gene prediction.

Moreover, most of the network embedding approaches follow an inefficient strategy, which they perform an external evaluation stage for the downstream tasks after the model is trained. However, it could be tiresome to demonstrate the power of a model compared to other models through various predictors. Whereas, an end-to-end model, which provides the evaluation scores as well as the embeddings is more efficient.

6.2.3 Convolutional neural networks (CNNs)

We observe that our random-walks based approach, N2VKO, attain lower prediction performance on diseases such as obesity, prostate cancer and diabetes mellitus

due to lower interconnectedness among known disease genes, that is one-hop neighbors of disease genes exhibit high sparsity (i.e., low clustering coefficient scores). This shows the weakness of diffusion-based models with respect to the distribution of disease genes in the interactome. This problem can be tackled by graph convolutions [105, 151, 156, 157, 208], which aim to capture the latent node feature information through the aggregation of several local neighborhoods so that characterize global neighborhood in the whole network. We can broadly categorize the approaches based on their construction of convolutions filters as spatial and spectral. Spatial filters are performed over the original graph via the adjacency matrix, whereas spectral filters are performed in graph-Laplacian domain [204].

However, CNNs may face the risk of over-fitting due to small number of positives and a large number of features. That is, the constructed model could be much more complex than the required, which is contrary to Ockham’s razor principle. Regularization techniques could be adopted to penalize the complexity of the model.

6.2.4 PU learning techniques

Candidate disease gene prediction problem can be tackled with a PU learning algorithm. The reason is, for a specific disease there are only a few known disease-associated genes from the human genome and the rest of the genes is waiting for further analysis and remain unlabeled. Therefore, unlike traditional machine learning methods, treating the rest of the genes as an unlabeled set rather than a negative training set is a suitable strategy for this problem.

Biased SVMs or weighted logistic regression algorithms have been studied for disease gene prediction problem [42, 61, 62], however, the prediction performance of these approaches highly depends on tuning the weights of the classifier. In addition, these methods are more advantageous in cases where the class prior in the

unlabeled dataset is known [209, 210]. Thus, *unbiased PU learning* algorithms [158] can be adopted for disease gene prediction algorithm.

It is also worth noting that a recent study [211] develops a generative adversarial networks (GAN)-based PU learning algorithm, which can open our minds to alternative ideas. Basically, GAN is a generative algorithm, which is designed as a two-player game. One of the players is the generator, which synthesizes fake data from random noise. The other player is the discriminator, which examines both fake and real data to determine whether it is real or not. Their design provides a more powerful approach to adapt PU learning into deep neural networks without over-fitting issue.

6.3 Final Remarks

Disease-causing genes and their protein products play a crucial role in the diagnosis and treatment of many diseases such as Alzheimer's, type 2 diabetes, and coronary heart disease. It is still a challenging research issue to produce new computational methods to reveal their interactions and functions in the context of many diseases.

In order to figure out the underlying mechanism of diseases, several approaches have been dedicated such as linkage analysis and genome-wide association studies (GWAS). These studies are based on genetic variations among humans known as single nucleotide polymorphisms (SNPs), which are used as predictive biomarkers. However, statistical analysis of the biomarkers demands vast amounts of money and time due to a high number of false positives. Additionally, these techniques are based on a simple genotype-phenotype relationship. However, it is known that diseases are consequences of perturbations in the molecular networks caused by various complex factors such as genetic mutations, epigenetic factors, and pathogens. Therefore, molecular networks are much more efficient tools in the analysis of diseases by tracking the potential disruptions on the biological pathways due to the

perturbations in the molecular networks. Thus, in the prediction of disease genes, molecular networks are important instruments.

In this thesis, we present our novel approaches, which extracts robust network-based feature representations for prediction of disease-causing genes with a wider perspective, by combining both topological characteristics of molecular networks and biological knowledge. The extensive experiments show that our representations perform better than the state-of-the-the-arts for candidate disease gene prediction.

List of Author's Awards, Patents, and Publications

Journal Articles

- **Ata Kircali Sezin**, Fang Yuan, Wu Min, Li Xiaoli and Xiao, Xiaoukui. Disease Gene Classification with Metagraph Representations. *Methods*, 131:83-92, 2017.¹
- **Ata Kircali Sezin**, Ou-Yang Le, Fang Yuan, Kwoh Chee-Keong, Wu Min and Li Xiaoli. Integrating node embeddings and biological annotations for genes to predict disease-gene associations. *BMC Systems Biology*, 12(9):138, 2018.²

Book Chapter

- **Ata Kircali Sezin**, Fang Yuan, Wu Min, Li Xiaoli and Xiao, Xiaoukui. *Data mining for systems biology: Methods and protocols*, chapter Disease gene classification with metagraph representations, pages 211-224. Humana Press, 2018.³

¹Presented in Chapter 3

²Presented in Chapter 4

³Presented in Chapter 3

Submitted Work

- **Ata Kircali Sezin**, Fang Yuan, Shi Jiaqi, Wu Min, Li Xiaoli, Kwoh Chee-Keong. “Unsupervised Multi-View Network Embedding with Higher-Order Collaboration,” *CIKM*, 2019.⁴

⁴Presented in Chapter 5

Bibliography

- [1] Xiujuan Wang, Natali Gulbahce, and Haiyuan Yu. Network-based methods for human disease gene prediction. *Briefings in Functional Genomics*, 10(5): 280, 2011. [x](#), [3](#), [16](#), [18](#)
- [2] Chia-Lang Hsu, Yen-Hua Huang, Chien-Ting Hsu, and Ueng-Cheng Yang. Prioritizing disease candidate genes by a gene interconnectedness-based approach. *BMC Genomics*, 12(3):S25, 2011. [x](#), [26](#), [27](#), [33](#)
- [3] Yongjin Li and Jagdish C. Patra. Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, 26(9):1219–1224, 2010. [x](#), [xv](#), [6](#), [24](#), [27](#), [35](#), [36](#), [38](#), [41](#), [73](#), [91](#)
- [4] István A. Kovács, Katja Luck, Kerstin Spirohn, Yang Wang, Carl Pollis, Sadie Schlabach, Wenting Bian, Dae-Kyum Kim, Nishka Kishore, Tong Hao, Michael A. Calderwood, Marc Vidal, and Albert-László Barabási. Network-based prediction of protein interactions. *Nature Communications*, 10(1):1240, 2019. ISSN 2041-1723. [xii](#), [138](#)
- [5] Sebastian Köhler, Sebastian Bauer, Denise Horn, and Peter N. Robinson. Walking the interactome for prioritization of candidate disease genes. *The American Journal of Human Genetics*, 82(4):949 – 958, 2008. [xv](#), [7](#), [26](#), [34](#), [57](#), [62](#), [73](#), [91](#)

- [6] David King and Christina M. King. Health management: Prevention awareness and wellness exceptionalism. *International Journal of Social Health Information Management*, 4(8):43–72, 2011. [1](#)
- [7] Fides Zenk, Eva Loeser, Rosaria Schiavo, Fabian Kilpert, Ozren Bogdanović, and Nicola Iovino. Germ line-inherited h3k27me3 restricts enhancer function during maternal-to-zygotic transition. *Science*, 357(6347):212–216, 2017. ISSN 0036-8075. [1](#), [14](#)
- [8] Matthew R. Nelson, Hannah Tipney, Jeffery L. Painter, Judong Shen, Paola Nicoletti, Yufeng Shen, Aris Floratos, Pak Chung Sham, Mulin Jun Li, Junwen Wang, Lon R. Cardon, John C. Whittaker, and Philippe Sanseau. The support of human genetic evidence for approved drug indications. *Nat Genet*, 47(8):856–860, Aug 2015. [2](#)
- [9] William A. Flavahan, Yotam Drier, Brian B. Liao, Shawn M. Gillespie, Andrew S. Venteicher, Anat O. Stemmer-Rachamimov, Mario L. Suvá, and Bradley E. Bernstein. Insulator dysfunction and oncogene activation in idh mutant gliomas. *Nature*, 529(7584):110–114, Jan 2016.
- [10] Aswin Sekar, Allison R. Bialas, Heather de Rivera, Avery Davis, Timothy R. Hammond, Nolan Kamitaki, Katherine Tooley, Jessy Presumey, Matthew Baum, Vanessa Van Doren, Giulio Genovese, Samuel A. Rose, Robert E. Handsaker, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Mark J. Daly, Michael C. Carroll, Beth Stevens, and Steven A. McCarroll. Schizophrenia risk from complex variation of complement component 4. *Nature*, 530(7589):177–183, Feb 2016.
- [11] Priti Singh and John C. Schimenti. The genetics of human infertility by functional interrogation of snps in mice. *Proc Natl Acad Sci U S A*, 112(33):10431–10436, Aug 2015. [2](#)

- [12] Sora Yoon et al. Efficient pathway enrichment and network analysis of GWAS summary data using GSA-SNP2. *Nucleic Acids Research*, 46(10):e60–e60, 03 2018. ISSN 0305-1048. [2](#)
- [13] M. Li, Y. Lu, J. Wang, F.-X. Wu, and Y. Pan. A topology potential-based method for identifying essential proteins from ppi networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(2):372–383, 2015. [3](#)
- [14] T. Gui, X. Dong, R. Li, Y. Li, and Z. Wang. Identification of hepatocellular carcinoma-related genes with a machine learning and network analysis. *Journal of Computational Biology*, 22(1):63–71, 2015.
- [15] L. Fu, S. Zhang, L. Zhang, X. Tong, J. Zhang, Y. Zhang, L. Ouyang, B. Liu, and J. Huang. Systems biology network-based discovery of a small molecule activator bl-ad008 targeting ampk/zipk and inducing apoptosis in cervical cancer. *Oncotarget*, 6(10):8071–8088, 2015. [3](#)
- [16] Han-Yu Chuang, Eunjung Lee, Yu-Tsueng Liu, Doheon Lee, and Trey Ideker. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3(1):140–n/a, 2007. [3](#), [25](#), [47](#)
- [17] Trey Ideker and Roded Sharan. Protein networks in disease. *Genome Research*, 18(4):644–652, 2008. [3](#), [7](#), [47](#)
- [18] Jörg Menche, Amitabh Sharma, Maksim Kitsak, Susan Dina Ghiassian, Marc Vidal, Joseph Loscalzo, and Albert-László Barabási. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224), 2015. ISSN 0036-8075. [3](#), [4](#), [6](#), [18](#)
- [19] UniProt Consortium et al. Uniprot: a hub for protein information. *Nucleic acids research*, page gku989, 2014. [3](#), [10](#), [47](#), [50](#), [57](#)

- [20] Paolo Marcatili and Anna Tramontano. *Network cleansing: Reliable interaction networks*, pages 80–97. IGI Global, 2009. ISBN 9781605663982. [4](#), [50](#)
- [21] Teresa Davoli et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell*, 155(4):948–962, Nov 2013. ISSN 0092-8674. [4](#), [6](#)
- [22] Ping Luo, Yuanyuan Li, Li-Ping Tian, and Fang-Xiang Wu. Enhancing the prediction of disease–gene associations with multimodal deep learning. *Bioinformatics*, 03 2019. ISSN 1367-4803. [4](#), [6](#)
- [23] Vladimir Gligorijević and Nataša Pržulj. Methods for biological data integration: perspectives and challenges. *Journal of The Royal Society Interface*, 12(112), 2015. [5](#), [17](#), [19](#), [21](#), [22](#), [24](#)
- [24] Kai Sun, Joana P. Gonçalves, Chris Larminie, and Nataša Pržulj. Predicting disease associations via biological network analysis. *BMC Bioinformatics*, 15(1):304, 2014. [6](#), [8](#)
- [25] Xiaoli Li, Min Wu, Chee-Keong Kwoh, and See-Kiong Ng. Computational approaches for detecting protein complexes from protein interaction networks: a survey. *BMC Genomics*, 11(1):S3, Feb 2010. [28](#), [29](#)
- [26] Wei Peng, Jianxin Wang, Juan Cai, Lu Chen, Min Li, and Fang-Xiang Wu. Improving protein function prediction using domain and protein complexes in ppi networks. *BMC Syst Biol*, 8:35–35, Mar 2014. [6](#), [17](#), [31](#), [33](#), [48](#)
- [27] Saket Navlakha and Carl Kingsford. The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, 26(8):1057, 2010. [6](#), [8](#), [25](#), [33](#), [57](#), [72](#), [136](#)

- [28] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews genetics*, 12(1):56, 2011. [72](#)
- [29] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nat Rev Genet*, 12, 2011. [8](#), [25](#), [26](#), [32](#), [33](#), [136](#)
- [30] Jianzhen Xu and Yongjin Li. Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics*, 22(22): 2800–2805, 2006. [6](#), [47](#), [72](#), [73](#)
- [31] Ravindra Kumar, Sabindra K. Samal, Samapika Routray, Rupesh Dash, and Anshuman Dixit. Identification of oral cancer related candidate genes by integrating protein-protein interactions, gene ontology, pathway analysis and immunohistochemistry. *Scientific Reports*, 7(1):2472, 2017. [6](#)
- [32] Edward L. Huttlin et al. Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545:505, May 2017. [6](#)
- [33] Peilin Jia, Siyuan Zheng, Jirong Long, Wei Zheng, and Zhongming Zhao. dmngwas: dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics*, 27(1):95–102, 11 2010. ISSN 1367-4803. [6](#)
- [34] Insuk Lee, U. Martin Blom, Peggy I. Wang, Jung Eun Shim, and Edward M. Marcotte. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Research*, 21(7):1109–1121, 2011.
- [35] Y. Liu et al. Network-assisted analysis of gwas data identifies a functionally-relevant gene module for childhood-onset asthma. *Scientific Reports*, 7(1): 938, 2017. ISSN 2045-2322. [6](#)

- [36] P. Luo, L. Tian, J. Ruan, and F. Wu. Disease gene prediction by integrating ppi networks, clinical rna-seq data and omim data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(1):222–232, Jan 2019. ISSN 1545-5963. [6](#)
- [37] Y. Zhang, Z. Li, M. Yang, D. Wang, L. Yu, C. Guo, X. Guo, and N. Lin. Identification of grb2 and gab1 coexpression as an unfavorable prognostic factor for hepatocellular carcinoma by a combination of expression profile and network analysis. *PLoS ONE*, 8(12), 2013. [6](#), [50](#), [52](#)
- [38] Oron Vanunu, Oded Magger, Eytan Ruppin, Tomer Shlomi, and Roded Sharan. Associating genes and protein complexes with disease via network propagation. *PLOS Computational Biology*, 6(1):1–9, 01 2010. [6](#), [8](#), [26](#), [35](#), [38](#), [41](#), [57](#), [63](#), [73](#)
- [39] Gert R. G. Lanckriet, Tijl De Bie, Nello Cristianini, Michael I. Jordan, and William Stafford Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626, 2004. [6](#)
- [40] Alberto Valdeolivas, Laurent Tichit, Claire Navarro, Sophie Perrin, Gaelle Odelin, Nicolas Levy, Pierre Cau, Elisabeth Remy, and Anais Baudot. Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics*, 35(3):497–505, 07 2018. ISSN 1367-4803. [6](#), [24](#), [35](#), [38](#), [101](#)
- [41] Xingli Guo, Lin Gao, Chunshui Wei, Xiaofei Yang, Yi Zhao, and Anguo Dong. A computational method based on the integration of heterogeneous networks for predicting disease-gene associations. *PLOS ONE*, 6(9):1–10, 09 2011. [6](#)
- [42] U. Martin Singh-Blom, Nagarajan Natarajan, Ambuj Tewari, John O. Woods, Inderjit S. Dhillon, and Edward M. Marcotte. Prediction and validation of gene-disease associations using methods inspired by social network analyses. *PLOS ONE*, 8(5):1–17, 05 2013. [8](#), [41](#), [43](#), [48](#), [73](#), [91](#), [92](#), [140](#)

- [43] Xuebing Wu, Rui Jiang, Michael Q. Zhang, and Shao Li. Network-based global inference of human disease genes. *Mol Syst Biol*, 4:189–189, May 2008. [23](#), [27](#), [33](#)
- [44] Yongjin Li and Jinyan Li. Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data. *BMC genomics*, 13 Suppl 7(Suppl 7):S27–S27, Dec 2012. ISSN 1471-2164. [6](#)
- [45] Silpa Suthram, Joel T. Dudley, Annie P. Chiang, Rong Chen, Trevor J. Hastie, and Atul J. Butte. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLOS Computational Biology*, 6(2):1–10, 02 2010. [7](#), [25](#), [72](#)
- [46] Guanming Wu and Lincoln Stein. A network module-based method for identifying cancer prognostic signatures. *Genome Biology*, 13(12):R112, 2012. [25](#), [72](#)
- [47] Min Wu, Xiaoli Li, Chee-Keong Kwoh, and See-Kiong Ng. A core-attachment based method to detect protein complexes in ppi networks. *BMC Bioinformatics*, 10(1):169, 2009. [7](#), [31](#), [33](#)
- [48] A.D. King, Nataša Pržulj, and I. Jurisica. Protein complex prediction via cost-based clustering. *Bioinformatics*, 20(17):3013–3020, 2004. [7](#), [31](#), [33](#), [50](#)
- [49] Y.-K. Shih and S. Parthasarathy. Identifying functional modules in interaction networks through overlapping markov clustering. *Bioinformatics*, 28(18):i473–i479, 2012. [7](#), [28](#), [33](#), [50](#)
- [50] Young-Rae Cho and Aidong Zhang. Predicting protein function by frequent functional association pattern mining in protein interaction networks. *Trans. Info. Tech. Biomed.*, 14(1):30–36, January 2010. [7](#), [29](#), [33](#), [50](#)
- [51] Xiaojin Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the*

Twentieth International Conference on International Conference on Machine Learning, ICML'03, page 912–919. AAAI Press, 2003. ISBN 1577351894. [8](#)

- [52] László Lovász. Random walks on graphs: A survey, combinatorics, paul erdos is eighty. *Bolyai Soc. Math. Stud.*, 2, 01 1993.
- [53] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. Random walk with restart: Fast solutions and applications. 14(3), 2008. ISSN 0219-1377. [38](#), [43](#)
- [54] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'03, page 321–328, Cambridge, MA, USA, 2003. MIT Press. [8](#)
- [55] Lenore Cowen, Trey Ideker, Benjamin J. Raphael, and Roded Sharan. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, 18:551, Jun 2017. [8](#)
- [56] Jie Zhu, Yufang Qin, Taigang Liu, Jun Wang, and Xiaoqi Zheng. Prioritization of candidate disease genes by topological similarity between disease and protein diffusion profiles. *BMC Bioinformatics*, 14(5):S5, 2013. [33](#), [36](#), [38](#), [57](#), [136](#)
- [57] Jung Eun Shim, Sohyun Hwang, and Insuk Lee. Pathway-dependent effectiveness of network algorithms for gene prioritization. *PLOS ONE*, 10(6): 1–10, 06 2015. [8](#), [33](#), [57](#), [136](#)
- [58] Nagarajan Natarajan and Inderjit S. Dhillon. Inductive matrix completion for predicting gene–disease associations. *Bioinformatics*, 30(12):i60–i68, Jun 2014. [8](#), [39](#), [43](#)
- [59] Xiangxiang Zeng, Ningxiang Ding, Alfonso Rodríguez-Patón, and Quan Zou. Probability-based collaborative filtering model for predicting gene–disease

- associations. *BMC Medical Genomics*, 10(5):76, Dec 2017. ISSN 1755-8794. [8](#), [39](#), [40](#), [43](#)
- [60] Ping Luo, Li-Ping Tian, Bolin Chen, Qianghua Xiao, and Fang-Xiang Wu. Predicting gene-disease associations with manifold learning. In Fa Zhang, Zhipeng Cai, Pavel Skums, and Shihua Zhang, editors, *Bioinformatics Research and Applications*, pages 265–271, Cham, 2018. Springer International Publishing. ISBN 978-3-319-94968-0. [8](#), [39](#), [40](#), [43](#)
- [61] Fantine Mordelet and Jean-Philippe Vert. Prodige: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples. *BMC Bioinformatics*, 12(1):389, 2011. [8](#), [40](#), [43](#), [73](#), [91](#), [93](#), [140](#)
- [62] Peng Yang, Xiao-Li Li, Jian-Ping Mei, Chee-Keong Kwoh, and See-Kiong Ng. Positive-unlabeled learning for disease gene identification. *Bioinformatics*, 28(20):2640–2647, 08 2012. ISSN 1367-4803. [8](#), [41](#), [43](#), [140](#)
- [63] L. Li, L. Wu, H. Zhang, and F. Wu. A fast algorithm for nonnegative matrix factorization and its convergence. *IEEE Transactions on Neural Networks and Learning Systems*, 25(10):1855–1863, Oct 2014. ISSN 2162-237X. [8](#), [40](#), [44](#)
- [64] Nataša Pržulj, D. G. Corneil, and I. Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508, 2004. [8](#)
- [65] T. Milenković and Nataša Pržulj. Uncovering biological network function via graphlet degree signatures. *Cancer Informatics*, 6:257–273, 2008. [8](#), [30](#), [33](#), [50](#)
- [66] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. *KDD*, 2016:855–864, Aug 2016. ISSN 2154-817X. [9](#), [24](#), [41](#), [42](#), [73](#), [76](#), [102](#), [108](#), [109](#), [110](#), [114](#), [116](#), [137](#), [138](#)

- [67] Shaosheng Cao, Wei Lu, and Qiongkai Xu. Grarep: Learning graph representations with global structural information. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 891–900, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3794-6.
- [68] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 701–710, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2956-9. [9](#), [41](#), [102](#), [137](#)
- [69] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 1067–1077, Republic and Canton of Geneva, Switzerland, 2015. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-3469-3. [9](#), [41](#), [102](#), [137](#)
- [70] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations (ICLR)*, 2013. [9](#), [41](#)
- [71] William L. Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *CoRR*, abs/1709.05584, 2017. [9](#), [42](#)
- [72] Marinka Zitnik and Jure Leskovec. Predicting multicellular function through multi-layer tissue networks. *Bioinformatics*, 33(14):i190–i198, 2017. [9](#), [25](#), [42](#), [43](#), [101](#)
- [73] Monica Agrawal, Marinka Zitnik, and Jure Leskovec. Large-scale analysis of disease pathways in the human interactome. *Pac Symp Biocomput*, 23: 111–122, 2018. [9](#), [42](#), [43](#)

- [74] Hongming Zhang, Liwei Qiu, Lingling Yi, and Yangqiu Song. Scalable multiplex network embedding. In *IJCAI*, pages 3082–3088, 2018. [9](#), [24](#), [42](#), [44](#), [102](#), [103](#), [104](#), [112](#), [114](#), [116](#), [120](#), [136](#)
- [75] Yiwei Sun, Ngoc Bui, Tsung-Yu Hsieh, and Vasant G. Honavar. Multi-view network embedding via graph factorization clustering and co-regularized multi-view agreement. *CoRR*, abs/1811.02616, 2018. [106](#), [136](#)
- [76] Meng Qu, Jian Tang, Jingbo Shang, Xiang Ren, Ming Zhang, and Jiawei Han. An attention-based collaboration framework for multi-view network representation learning. In *CIKM*, pages 1767–1776. ACM, 2017. [24](#), [42](#), [44](#), [102](#), [103](#), [104](#), [106](#), [112](#), [114](#), [116](#), [120](#), [136](#), [139](#)
- [77] Yifeng Li, Fang-Xiang Wu, and Alioune Ngom. A review on machine learning principles for multi-view biological data integration. *Briefings in Bioinformatics*, 19(2):325–340, 12 2016. ISSN 1477-4054. [9](#)
- [78] Matteo Fumagalli, Manuela Sironi, Uberto Pozzoli, Anna Ferrer-Admettla, Linda Pattini, and Rasmus Nielsen. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLOS Genetics*, 7(11):1–14, 11 2011. [14](#)
- [79] Monkol Lek et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536:285, Aug 2016. [15](#)
- [80] Kai Sun, Joana P. Gonçalves, Chris Larminie, and Nataša Pržulj. Predicting disease associations via biological network analysis. *BMC Bioinformatics*, 15(1):304, 2014. [17](#), [48](#), [50](#)
- [81] Rui Jiang, Mengmeng Wu, and Lianshuo Li. Pinpointing disease genes through phenomic and genomic data fusion. *BMC Genomics*, 16(2):S3, 2015. [50](#)

- [82] Jia-Shing Chen, Wei-Shiang Hung, Hsiang-Han Chan, Shaw-Jenq Tsai, and H. Sunny Sun. In silico identification of oncogenic potential of fyn-related kinase in hepatocellular carcinoma. *Bioinformatics*, 29(4):420, 2013. [17](#), [50](#)
- [83] Jonathan B. L. Bard and Seung Y. Rhee. Ontologies in biology: design, applications and future challenges. *Nat Rev Genet*, 5(3):213–222, Mar 2004. [17](#)
- [84] Gene Ontology Consortium. The gene ontology (go) database and informatics resource. *Nucleic Acids Research*, 32(suppl-1):D258, 2004. [17](#)
- [85] Robert D. Finn et al. The pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279, 2016. [18](#), [29](#)
- [86] Philip Jones et al. Interproscan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236, 2014. [18](#)
- [87] Javier De Las Rivas and Celia Fontanillo. Protein-protein interactions essentials: Key concepts to building and analyzing interactome networks. *PLoS Comput Biol*, 6(6):e1000807, Jun 2010. [18](#)
- [88] Damian Szklarczyk et al. String v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1):D447, 2015. [18](#), [49](#), [57](#)
- [89] Donna Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic Acids Research*, 35(suppl-1):D26, 2007. [18](#), [49](#), [57](#), [75](#), [81](#)
- [90] Sandra Orchard et al. The mintact project intact as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, 42 (Database issue):D358–63, January 2014. [18](#), [49](#), [51](#), [57](#), [75](#), [81](#), [103](#), [118](#)
- [91] Marc Vidal, Michael E. Cusick, and Albert-László Barabási. Interactome networks and human disease. *Cell*, 144(6):986–998, Mar 2011. [18](#), [19](#), [20](#)

- [92] Ada Hamosh, Alan F. Scott, Joanna Amberger, Carol Bocchini, David Valle, and Victor A. McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 30(1):52, 2002. [19](#), [58](#), [75](#), [82](#)
- [93] Joshua M. Stuart, Eran Segal, Daphne Koller, and Stuart K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302(5643):249–255, 2003. [19](#)
- [94] Eric Davidson and Michael Levin. Gene regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(14):4935, 2005. [19](#)
- [95] Stefan Wuchty and Eivind Almaas. Evolutionary cores of domain co-occurrence networks. *BMC Evolutionary Biology*, 5(1):24, Mar 2005. [20](#), [29](#), [33](#)
- [96] Hongwu Ma, Anatoly Sorokin, Alexander Mazein, Alex Selkov, Evgeni Selkov, Oleg Demin, and Igor Goryanin. The edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol*, 3:135–135, Sep 2007. [20](#)
- [97] Ming Wu and Christina Chan. Human metabolic network: Reconstruction, simulation, and applications in systems biology. *Metabolites*, 2(1):242–253, 2012.
- [98] Natapol Pornputtpong, Intawat Nookaew, and Jens Nielsen. Human metabolic atlas: an online resource for human metabolism. *Database (Oxford)*, 2015:bav068, Jul 2015. [20](#)
- [99] Yong Chen, Jingjing Hao, Wei Jiang, Tong He, Xuegong Zhang, Tao Jiang, and Rui Jiang. Identifying potential cancer driver genes by genomic data integration. *Scientific Reports*, 3:3538, Dec 2013. [21](#)

- [100] Jaegyeon Ahn, Youngmi Yoon, Chihyun Park, Eunji Shin, and Sanghyun Park. Integrative gene network construction for predicting a set of complementary prostate cancer genes. *Bioinformatics*, 27(13):1846–1853, 05 2011. ISSN 1367-4803. [21](#)
- [101] Fan Yang, Duzhi Wu, Limei Lin, Jian Yang, Tinghong Yang, and Jing Zhao. The integration of weighted gene association networks based on information entropy. *PLOS ONE*, 12(12):1–19, 12 2017. [22](#)
- [102] Kai Sun, Natalie Buchan, Chris Larminie, and Natasa Przulj. The integrated disease network. *Integrative Biology*, 6(11):1069–1079, 08 2014. ISSN 1757-9708. [22](#)
- [103] Darcy A. Davis and Nitesh V. Chawla. Exploring and exploiting disease interactions from multi-relational gene and phenotype networks. *PLOS ONE*, 6(7):1–9, 07 2011. [23](#), [137](#)
- [104] Jiajie Peng, Kun Bai, Xuequn Shang, Guohua Wang, Hansheng Xue, Shuilin Jin, Liang Cheng, Yadong Wang, and Jin Chen. Predicting disease-related genes using integrated biomedical networks. *BMC Genomics*, 18(1):1043, 2017. ISSN 1471-2164. [23](#)
- [105] Aditya Rao, Saipradeep VG, Thomas Joseph, Sujatha Kotte, Naveen Sivadasan, and Rajgopal Srinivasan. Phenotype-driven gene prioritization for rare diseases using graph convolution on heterogeneous networks. *BMC Medical Genomics*, 11(1):57, Jul 2018. ISSN 1755-8794. [24](#), [43](#), [140](#)
- [106] K. Yang, R. Wang, G. Liu, Z. Shu, N. Wang, R. Zhang, J. Yu, J. Chen, X. Li, and X. Zhou. Hergpred: Heterogeneous network embedding representation for disease gene prediction. *IEEE Journal of Biomedical and Health Informatics*, 23(4):1805–1815, 2019. ISSN 2168-2194. [24](#), [43](#)

- [107] Guixiang Ma, Chun-Ta Lu, Lifang He, S Yu Philip, and Ann B Ragin. Multi-view graph embedding with hub detection for brain network analysis. In *ICDM*, pages 967–972. IEEE, 2017. [24](#)
- [108] Yu Shi, Fangqiu Han, Xinwei He, Xinran He, Carl Yang, Jie Luo, and Jiawei Han. mvn2vec: Preservation and collaboration in multi-view network embedding. *arXiv preprint arXiv:1801.06597*, 2018. [24](#), [102](#), [103](#), [104](#), [106](#), [112](#), [114](#), [116](#), [120](#)
- [109] Xingli Guo, Lin Gao, Chunshui Wei, Xiaofei Yang, Yi Zhao, and Anguo Dong. A computational method based on the integration of heterogeneous networks for predicting disease-gene associations. *PLOS ONE*, 6(9):1–10, 09 2011. [24](#)
- [110] Arda Halu, Manlio De Domenico, Alex Arenas, and Amitabh Sharma. The multiplex network of human diseases. *npj Systems Biology and Applications*, 5(1):15, 2019. ISSN 2056-7189. [24](#), [101](#)
- [111] Jingchao Ni, Mehmet Koyuturk, Hanghang Tong, Jonathan Haines, Rong Xu, and Xiang Zhang. Disease gene prioritization by integrating tissue-specific molecular networks using a robust multi-network model. *BMC Bioinformatics*, 17(1):453, Nov 2016. ISSN 1471-2105. [24](#), [38](#), [101](#)
- [112] Marinka Zitnik and Blaz Zupan. Jumping across biomedical contexts using compressive data fusion. *Bioinformatics (Oxford, England)*, 32(12):i90–i100, 2016. ISSN 1367-4811. [25](#), [40](#), [43](#)
- [113] Cheng Zhu, Akash Kushwaha, Kenneth Berman, and Anil G. Jegga. A vertex similarity-based framework to discover and rank orphan disease-related genes. *BMC Syst Biol*, 6(Suppl 3):S8–S8, Dec 2012. [26](#), [27](#), [33](#)

- [114] Kasper Lage et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotech*, 25(3):309–316, Mar 2007. [27](#), [33](#)
- [115] Wei Liu, Aiping Wu, Matteo Pellegrini, and Xiaofan Wang. Integrative analysis of human protein, function and disease networks. *Scientific Reports*, 5: 14344, Sep 2015. [27](#), [33](#), [48](#)
- [116] Predrag Radivojac et al. A large-scale evaluation of computational protein function prediction. *Nat Meth*, 10(3):221–227, Mar 2013. [28](#)
- [117] Berend Snel, Peer Bork, and Martijn A. Huynen. The identification of functional modules from the genomic association of genes. *Proceedings of the National Academy of Sciences*, 99(9):5890–5895, 2002. [28](#), [33](#)
- [118] J. Wang, M. Li, J. Chen, and Y. Pan. A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(3):607–620, May 2011. [28](#)
- [119] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, 30(7):1575–1584, Apr 2002. [28](#)
- [120] Jose B. Pereira-Leal, Anton J. Enright, and Christos A. Ouzounis. Detection of functional modules from protein interaction networks. *Proteins: Structure, Function, and Bioinformatics*, 54(1):49–57, 2004. [28](#)
- [121] Jaya Thomas, Dongmin Seo, and Lee Sael. Review on graph clustering and subgraph similarity based analysis of neurological disorders. *Int J Mol Sci*, 17(6):862, Jun 2016. [28](#)

- [122] Balázs Adamcsek, Gergely Palla, Illés J. Farkas, Imre Derényi, and Tamás Vicsek. Cfinder: Locating cliques and overlapping modules in biological networks. *Bioinformatics*, 22(8):1021–1023, April 2006. [28](#), [33](#)
- [123] Gergely Palla, Imre Derenyi, Illes Farkas, and Tamas Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, Jun 2005. [28](#)
- [124] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002. [29](#)
- [125] Ngoc Tam L. Tran, Sominder Mohan, Zhuoqing Xu, and Chun-Hsi Huang. Current innovations and future challenges of network motif detection. *Briefings in Bioinformatics*, 16(3):497, 2015. [29](#), [50](#), [52](#)
- [126] Y. R. Cho, Y. Xin, and G. Speegle. P-finder: Reconstruction of signaling networks from protein-protein interactions and go annotations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(2):309–321, March 2015. [29](#), [33](#), [50](#), [52](#)
- [127] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proc. of 20th Intl. Conf. on VLDB*, pages 487–499, 1994. [29](#)
- [128] Gary D. Bader and Christopher WV Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4:2–2, Jan 2003. [30](#), [33](#)
- [129] Md Altaf-Ul-Amin, Yoko Shinbo, Kenji Mihara, Ken Kurokawa, and Shigehiko Kanaya. Development and implementation of an algorithm for detection of protein complexes in large interaction networks. *BMC Bioinformatics*, 7(1):207, Apr 2006. [30](#), [33](#)

- [130] B. Zhao, J. Wang, M. Li, F. X. Wu, and Y. Pan. Detecting protein complexes based on uncertain graph model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(3):486–497, May 2014. [31](#), [33](#)
- [131] Zhi Hao Yang, Feng Ying Yu, Hong Fei Lin, and Jian Wang. Integrating ppi datasets with the ppi data from biomedical literature for protein complex detection. *BMC Med Genomics*, 7(Suppl 2):S3–S3, Oct 2014. [31](#), [33](#), [48](#)
- [132] Yanjun Qi, Fernanda Balem, Christos Faloutsos, Judith Klein-Seetharaman, and Ziv Bar-Joseph. Protein complex identification by supervised graph local clustering. *Bioinformatics*, 24(13):i250–i268, Jul 2008. [32](#), [33](#)
- [133] Hongchao Lu, Baochen Shi, Gaowei Wu, Yong Zhang, Xiaopeng Zhu, Zhihua Zhang, Changning Liu, Yi Zhao, Tao Wu, Jie Wang, and Runsheng Chen. Integrated analysis of multiple data sources reveals modular structure of biological networks. *Biochemical and Biophysical Research Communications*, 345(1):302 – 309, 2006. [32](#)
- [134] Bolin Chen, Weiwei Fan, Juan Liu, and Fang-Xiang Wu. Identifying protein complexes and functional modules from static ppi networks to dynamic ppi networks. *Briefings in Bioinformatics*, 15(2):177, 2014. [32](#)
- [135] Marc A. van Driel, Jorn Bruggeman, Gert Vriend, Han G. Brunner, and Jack A. M. Leunissen. A text-mining analysis of the human phenome. *Eur J Hum Genet*, 14(5):535–542, Feb 2006. [35](#), [63](#)
- [136] Anthony ML Liekens, Jeroen De Knijf, Walter Daelemans, Bart Goethals, Peter De Rijk, and Jurgen Del-Favero. Biograph: unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biology*, 12(6):R57, 2011. ISSN 1474-760X. [35](#), [38](#)
- [137] Sinan Erten, Gurkan Bebek, and Mehmet Koyutürk. Vavien: An algorithm for prioritizing candidate disease genes based on topological similarity of

- proteins in interaction networks. *J Comput Biol*, 18(11):1561–1574, Nov 2011. [36](#), [38](#)
- [138] Duc-Hau Le and Yung-Keun Kwon. Neighbor-favoring weight reinforcement to improve random walk-based disease gene prioritization. *Computational Biology and Chemistry*, 44:1 – 8, 2013. [36](#), [38](#)
- [139] Peng Yang, Xiaoli Li, Min Wu, Chee-Keong Kwoh, and See-Kiong Ng. Inferring gene-phenotype associations via global protein complex network propagation. *PLOS ONE*, 6(7):1–11, 07 2011. [36](#), [38](#)
- [140] Yang Chen and Rong Xu. Network-based gene prediction for plasmodium falciparum malaria towards genetics-based drug discovery. *BMC Genomics*, 16(7):S9, 2015. [36](#), [38](#)
- [141] Shunyao Wu, Fengjing Shao, Jun Ji, Rencheng Sun, Rizhuang Dong, Yuanke Zhou, Shaojie Xu, Yi Sui, and Jianlong Hu. Network propagation with dual flow for gene prioritization. *PLOS ONE*, 10(2):1–15, 02 2015. [37](#), [38](#)
- [142] MaoQiang Xie, YingJie Xu, YaoGong Zhang, TaeHyun Hwang, and Rui Kuang. Network-based phenome-genome association prediction by bi-random walk. *PLOS ONE*, 10(5):1–18, 05 2015. [37](#), [38](#)
- [143] Jiajie Peng, Kun Bai, Xuequn Shang, Guohua Wang, Hansheng Xue, Shuiling Jin, Liang Cheng, Yadong Wang, and Jin Chen. Predicting disease-related genes using integrated biomedical networks. *BMC Genomics*, 18(1):1043, 2017. [37](#), [38](#)
- [144] Kathy Macropol, Tolga Can, and Ambuj K. Singh. Rrw: repeated random walks on genome-scale protein networks for local cluster discovery. *BMC Bioinformatics*, page 283, 09 2009. [38](#), [43](#)

- [145] Wei Xiong, Hui Liu, Jihong Guan, and Shuigeng Zhou. Protein function prediction by collective classification with explicit and implicit edges in protein-protein interaction networks. *BMC Bioinformatics*, 14(12):S4, 2013. [39](#), [43](#)
- [146] Bolin Chen, Jianxin Wang, Min Li, and Fang-Xiang Wu. Identifying disease genes by integrating multiple data sources. *BMC Medical Genomics*, 7(2):S2, 2014. [39](#), [43](#)
- [147] Yong Chen, Xuebing Wu, and Rui Jiang. Integrating human omics data to prioritize candidate genes. *BMC Medical Genomics*, 6(1):57, 2013. [39](#), [43](#)
- [148] Daniela Nitsch, Joana P. Gonçalves, Fabian Ojeda, Bart de Moor, and Yves Moreau. Candidate gene prioritization by network analysis of differential expression using machine learning approaches. *BMC Bioinformatics*, 11(1):460, 2010. [39](#), [43](#)
- [149] Hongyun Cai, Vincent W. Zheng, and Kevin Chen-Chuan Chang. A comprehensive survey of graph embedding: Problems, techniques and applications. *IEEE Transactions on Knowledge and Data Engineering*, 09 2017. [41](#)
- [150] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–444, May 2015.
- [151] Peng Han, Peng Yang, Peilin Zhao, Shuo Shang, Yong Liu, Jiayu Zhou, Xin Gao, and Panos Kalnis. Gcn-mf: Disease-gene association identification by graph convolutional networks and matrix factorization. pages 705–713, 07 2019. [41](#), [140](#)
- [152] Ji-Wei Chang, Yuduan Ding, Muhammad Tahir ul Qamar, Yin Shen, Junxiang Gao, and Ling-Ling Chen. A deep learning model based on sparse auto-encoder for prioritizing cancer-related genes and drug target combinations. *Carcinogenesis*, 40(5):624–632, 04 2019. ISSN 0143-3334. [41](#)

- [153] Tianle Ma and Aidong Zhang. Multi-view factorization autoencoder with network constraints for multi-omic integrative analysis. *CoRR*, abs/1809.01772, 2018.
- [154] Davide Chicco, Peter Sadowski, and Pierre Baldi. Deep autoencoder neural networks for gene ontology annotation predictions. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '14, pages 533–540, New York, NY, USA, 2014. ACM.
- [155] Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1225–1234, New York, NY, USA, 2016. ACM. [41](#), [137](#), [138](#)
- [156] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 3844–3852, USA, 2016. Curran Associates Inc. ISBN 978-1-5108-3881-9. [41](#), [140](#)
- [157] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13): i457–i466, 06 2018. ISSN 1367-4803. [41](#), [140](#)
- [158] Ryuichi Kiryo, Gang Niu, Marthinus Christoffel du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. *CoRR*, abs/1703.00593, 2017. [44](#), [141](#)
- [159] Xiwei Tang, Xiaohua Hu, Xuejun Yang, Yetian Fan, Yongfan Li, Wei Hu, Yongzhong Liao, Ming cai Zheng, Wei Peng, and Li Gao. Predicting diabetes mellitus genes via protein-protein interaction and protein subcellular localization information. *BMC Genomics*, 17(Suppl 4):433, Aug 2016. [47](#)

- [160] Janos X. Binder, Sune Pletscher-Frankild, Kalliopi Tsafoi, Christian Stolte, Seán I. O'Donoghue, Reinhard Schneider, and Lars Juhl Jensen. Compartments: unification and visualization of protein subcellular localization evidence. *Database (Oxford)*, 2014:bau012, Feb 2014. [47](#)
- [161] Slobodan Vucetic, Hongbo Xie, Lilia M. Iakoucheva, Christopher J. Oldfield, A. Keith Dunker, Zoran Obradovic, and Vladimir N. Uversky. Functional anthology of intrinsic disorder. 2. cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. *Journal of Proteome Research*, 6(5):1899–1916, 2007. [47](#)
- [162] Vladimir N. Uversky, Christopher J. Oldfield, and A. Keith Dunker. Intrinsically disordered proteins in human diseases: Introducing the d2 concept. *Annual Review of Biophysics*, 37(1):215–246, 2008. [47](#)
- [163] Tejaswita M. Karve and Amrita K. Cheema. Small changes huge impact: The role of protein posttranslational modifications in cellular homeostasis and disease. *J Amino Acids*, 2011:207691, Jul 2011. [47](#)
- [164] Matthias Mann and Ole N. Jensen. Proteomic analysis of post-translational modifications. *Nat Biotech*, 21(3):255–261, Mar 2003.
- [165] Jonathan Woodsmith, Ulrich Stelzl, and Arunachalam Vinayagam. *Bioinformatics Analysis of PTM-Modified Protein Interaction Networks and Complexes*, pages 321–332. Springer New York, New York, NY, 2017. ISBN 978-1-4939-6783-4. [47](#)
- [166] Yuan Fang, Wenqing Lin, Vincent Wenchen Zheng, Min Wu, Kevin Chen-Chuan Chang, and Xiaoli Li. Semantic proximity search on graphs with metagraph-based learning. In *32nd IEEE International Conference on Data Engineering, ICDE 2016, Helsinki, Finland, May 16-20, 2016*, pages 277–288, 2016. [48](#), [52](#), [56](#), [59](#), [64](#)

- [167] L. Zhu, S.-P. Deng, and D.-S. Huang. A two-stage geometric method for pruning unreliable links in protein-protein networks. *IEEE Transactions on Nanobioscience*, 14(5):528–534, 2015. [50](#)
- [168] UniProt Consortium et al. Uniprot: a hub for protein information. *Nucleic acids research*, page gku989, 2014. [51](#), [75](#), [82](#), [103](#), [118](#)
- [169] Mohammed Elseidy, Ehab Abdelhamid, Spiros Skiadopoulos, and Panos Kalnis. Grami: Frequent subgraph and pattern mining in a single large graph. *Proc. VLDB Endow.*, 7(7):517–528, March 2014. [55](#), [59](#), [64](#)
- [170] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I Furlong. Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research*, 2016. [65](#), [99](#)
- [171] Dong I. Lee, Guangshuo Zhu, Takashi Sasaki, Gun-Sik Cho, Nazha Hamdani, Ronald Holewinski, Su-Hyun Jo, Thomas Danner, Manling Zhang, Peter P. Rainer, Djahida Bedja, Jonathan A. Kirk, Mark J. Ranek, Wolfgang R. Dostmann, Chulan Kwon, Kenneth B. Margulies, Jennifer E. Van Eyk, Walter J. Paulus, Eiki Takimoto, and David A. Kass. Phosphodiesterase 9a controls nitric-oxide-independent cgmp and hypertrophic heart disease. *Nature*, 519 (7544):472–476, Mar 2015. [66](#)
- [172] Bruno Borghese et al. Identification of susceptibility genes for peritoneal, ovarian, and deep infiltrating endometriosis using a pooled sample-based genome-wide association study. *BioMed Research International*, 2015. [66](#)
- [173] Hane Lee, Meng-chin A. Lin, Harley I. Kornblum, Diane M. Papazian, and Stanley F. Nelson. Exome sequencing identifies de novo gain of function missense mutation in *kcnk2* in identical twins with autism and seizures that

- slows potassium channel inactivation. *Human Molecular Genetics*, 23(13): 3481, 2014. [66](#)
- [174] Mark J. Perrin, Arnon Adler, Sharon Green, Foad Al-Zoughool, Petro Doroshenko, Nathan Orr, Shaheen Uppal, Jeff S. Healey, David Birnie, Shubhayan Sanatani, Martin Gardner, Jean Champagne, Chris Simpson, Kamran Ahmad, Maarten P. van den Berg, Vijay Chauhan, Peter H. Backx, J. Peter van Tintelen, Andrew D. Krahn, and Michael H. Gollob. Evaluation of genes encoding for the transient outward current (ito) identifies the *kcnd2* gene as a cause of j-wave syndrome associated with sudden cardiac deathclinical perspective. *Circulation: Cardiovascular Genetics*, 7(6):782–789, 2014. [66](#)
- [175] XueZhong Zhou, Jörg Menche, Albert-László Barabási, and Amitabh Sharma. Human symptoms–disease network. *Nature communications*, 5: 4212, 2014. [72](#)
- [176] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. [76](#), [77](#)
- [177] S. Milborrow. Derived from mda:mars by T. Hastie and R. Tibshirani. *earth: Multivariate Adaptive Regression Splines*, 2018. R package. [79](#)
- [178] C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. In *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*, pages 523–528, Aug 2003. [79](#), [80](#)
- [179] Max Kuhn. Caret: classification and regression training. *Astrophysics Source Code Library*, 2015. [80](#)
- [180] Nicolas De Jay, Simon Papillon-Cavanagh, Catharina Olsen, Nehme El-Hachem, Gianluca Bontempi, and Benjamin Haibe-Kains. mrmre: an r

- package for parallelized mrmr ensemble feature selection. *Bioinformatics*, 29(18):2365–2368, 2013. [80](#)
- [181] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. [80](#), [122](#)
- [182] Nicola Lunardon, Giovanna Menardi, and Nicola Torelli. ROSE: a Package for Binary Imbalanced Learning. *R Journal*, 6(1):82–92, 2014. [80](#)
- [183] Keping Wang, Tiefeng Qiu, and Xianwen Li. Structure-based identification, characterization, and disruption of human securin-binding sh3 domains in lung cancer. *Cancer Investigation*, 34(5):231–236, 2016. [90](#)
- [184] Ana M. Proenza, Jordi Oliver, Andreu Palou, and Pilar Roca. Breast and lung cancer are associated with a decrease in blood cell amino acid content. *The Journal of Nutritional Biochemistry*, 14(3):133 – 138, 2003. [90](#)
- [185] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013. [102](#), [103](#)
- [186] Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049, 2015. [103](#), [118](#)
- [187] Bokai Cao, Lifang He, Xiangnan Kong, S Yu Philip, Zhifeng Hao, and Ann B Ragin. Tensor-based multi-view feature selection with applications to brain diseases. In *ICDM*, pages 40–49. IEEE, 2014. [103](#)
- [188] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *ICML*, pages 129–136. ACM, 2009.
- [189] Derek Greene and Pádraig Cunningham. A matrix factorization approach for integrating multiple data views. In *ECML-PKDD*, pages 423–438. Springer, 2009.

- [190] Paramveer Dhillon, Dean P Foster, and Lyle H Ungar. Multi-view learning of word embeddings via cca. In *NIPS*, pages 199–207, 2011.
- [191] Guangxia Li, Steven CH Hoi, and Kuiyu Chang. Two-view transductive support vector machines. In *SDM*, pages 235–244. SIAM, 2010.
- [192] Chun-Ta Lu, Lifang He, Hao Ding, Bokai Cao, and Philip S Yu. Learning from multi-view multi-way data via structural factorization machines. In *WWW*, pages 1593–1602, 2018. [103](#)
- [193] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. *CoRR*, abs/1403.6652, 2014. [108](#), [109](#), [110](#), [114](#), [116](#), [119](#)
- [194] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. metapath2vec: Scalable representation learning for heterogeneous networks. In *KDD*, pages 135–144. ACM, 2017. [108](#)
- [195] James Z Wang, Zhidian Du, Rapeeporn Payattakool, Philip S Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281, 2007. [118](#)
- [196] Rui Li, Chi Wang, and Kevin Chen-Chuan Chang. User profiling in an ego network: co-profiling attributes and relationships. In *WWW*, pages 819–830, 2014. [119](#)
- [197] Lei Tang, Xufei Wang, and Huan Liu. Uncovering groups via heterogeneous interaction analysis. In *ICDM*, pages 503–512, Washington, DC, USA, 2009. IEEE Computer Society. ISBN 978-0-7695-3895-2. [119](#)
- [198] R. Zafarani and H. Liu. Social computing data repository at ASU, 2009. [119](#)
- [199] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(Nov):2579–2605, 2008. [121](#)

- [200] Chi Wang, Jiawei Han, Yuntao Jia, Jie Tang, Duo Zhang, Yintao Yu, and Jingyi Guo. Mining advisor-advisee relationships from research publication networks. In *KDD*, pages 203–212, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0055-1. [124](#)
- [201] Daniel S. Himmelstein and Sergio E. Baranzini. Heterogeneous network edge prediction: A data integration approach to prioritize disease-associated genes. *PLOS Computational Biology*, 11(7):1–27, 07 2015. [137](#)
- [202] Zengrui Wu, Weihua Li, Guixia Liu, and Yun Tang. Network-based methods for prediction of drug-target interactions. *Frontiers in pharmacology*, 9:1134–1134, Oct 2018. ISSN 1663-9812. [137](#)
- [203] Gamal Crichton, Yufan Guo, Sampo Pyysalo, and Anna Korhonen. Neural networks for link prediction in realistic biomedical graphs: a multi-dimensional evaluation of graph embedding-based approaches. *BMC Bioinformatics*, 19(1), May 2018. ISSN 1471-2105. [137](#)
- [204] Palash Goyal and Emilio Ferrara. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, 151:78 – 94, 2018. ISSN 0950-7051. [138](#), [140](#)
- [205] Mingdong Ou, Peng Cui, Jian Pei, Ziwei Zhang, and Wenwu Zhu. Asymmetric transitivity preserving graph embedding. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pages 1105–1114, New York, NY, USA, 2016. ACM. [138](#)
- [206] Masashi Tsubaki, Kentaro Tomii, and Jun Sese. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, 35(2):309–318, 07 2018. ISSN 1367-4803. [139](#)

- [207] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018. [139](#)
- [208] Yu Li, Hiroyuki Kuwahara, Peng Yang, Le Song, and Xin Gao. PgcN: Disease gene prioritization by disease and gene embedding through graph convolutional neural networks. *bioRxiv*, 2019. [140](#)
- [209] Marthinus C. du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS’14, pages 703–711, Cambridge, MA, USA, 2014. MIT Press. [141](#)
- [210] Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI’01, pages 973–978, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-812-5, 978-1-558-60812-2. [141](#)
- [211] Ming Hou, Qibin Zhao, Chao Li, and Brahim Chaib-draa. A generative adversarial framework for positive-unlabeled classification. *CoRR*, abs/1711.08054, 2017. [141](#)