PANDA: Prioritization of Autism-genes using Network-based Deep-learning Approach

by

© *Yu Zhang*

A thesis submitted to the

School of Graduate Studies

in partial fulfilment of the

requirements for the degree of

Master of *Science*

Department of *Computer Science*

Memorial University of Newfoundland

August 2019

St. John's Newfoundland

Abstract

Autism is a neuropsychiatric disorder characterized by impairments in reciprocal social interaction and communication, and the presence of restricted and repetitive behaviours. Autism is predominantly heritable, but the underlying genetic associations are still largely unknown. Understanding the genetic background of complex diseases, such as autism, plays an essential role in the promising precision medicine. The evaluation of candidate genes, however, requires time-consuming and expensive experiments given the large number of possibilities. Thus, computational methods have seen increasing applications in predicting gene-disease associations. In this thesis, we proposed a bioinformatics framework, *P*rioritization of *A*utism-genes using *N*etwork-based *D*eep-learning *A*pproach (PANDA). Our approach aims to identify autism-genes across the human genome based on patterns of gene-gene interactions and topological similarity of genes in the interaction network. PANDA trains a graph deep learning classifier using the input of the human molecular interaction network (HMIN) and predicts and ranks the probability of autism association of every node (gene) in the network. PANDA was able to achieve a high classification accuracy of 89%, outperforming three other commonly used machine learning algorithms. Moreover, the gene prioritization ranking list produced by PANDA was evaluated and validated using a large-scale independent exome-sequencing study. The top decile (top 10%) of PANDA ranked genes were found significantly enriched for autism association.

Acknowledgements

My first thanks go to Dr. Ting Hu, my supervisor. Without her support, motivation, guidance and overabundance of patience, I could not have even begun a career in Bioinformatics, let alone accomplish this long-held goal. Furthermore, I am deeply grateful to the School of Graduate Studies for supporting me financially throughout my degree.

My friends and family have been by my side through the best and the worst, have endured much, and I have gained much because of them. To my parents and my sister, you are the best. Everything else I can say is insufficient. Throughout my academic life, I have been in touch with many great teachers and colleagues; I express my special gratitude to them.

Finally, my deepest gratitude and admiration must go to Zachary. He always supports me and encourages me to step out of my comfort zone. Without his continual grace, persistence, and love, I could never have achieved this goal.

Contents

Abstract **ii**

Acknowledgements **iii**

List of Tables **vii**

List of Figures **viii**

1 Introduction **1**

2 Related Work **6**

2.1 Disease-gene association studies 6

2.2 Network science-based disease-gene association studies 8

2.3 Machine learning-based disease-gene association studies 10

2.4 Deep learning on biological data 11

2.5 Graph neural network 13

2.6 Summary 15

3 Disease-Gene Association Studies for Autism **16**

3.1 Background 16

3.1.1 Autism spectrum disorder 16
3.1.2 Autism gene databases 17
3.1.3 Network biology for molecular interaction network 18
3.2 Method 20
3.2.1 Training set of genes compilation 20
3.2.2 Human molecular interaction network 22
3.3 Results 26
3.3.1 Investigation of the HMIN 26
3.3.2 Investigation of the known autism-associated genes 27
3.4 Summary 29

4 Prioritization of Autism-Genes Using Network-Based Deep-Learning Approach 30

4.1 Background 30
4.1.1 Learning on networks 30
4.1.1.1 Network embedding 30
4.1.1.2 Challenges of network embedding 32
4.1.2 Graph neural network 33
4.2 Method 35
4.2.1 PANDA overview 35
4.2.2 Our GNN in PANDA 37
4.2.3 Loss function and optimization 40
4.2.4 Training and cross-validation 42
4.3 Results 42

4.3.1 Classification performance of PANDA 43
4.3.2 Model sensitivity and specificity 44
4.3.3 Methods comparison 47
4.4 Biological interpretation 48
4.4.1 Characterization of prioritized autism-associated genes 49
4.4.2 Enrichment analysis 52
4.5 Summary 53

5 Discussion **58**
5.1 Contribution summary 58
5.2 Future work 60

Bibliography **62**

List of Tables

3.1 Gene labels based on their confidence levels of autism association . . 22

3.2 Network properties of the HMIN 26

3.3 Measurements of subgraphs extracted using autism-associated genes in HMIN 29

4.1 Terms and notations used in PANDA algorithm 40

4.2 Enriched functional annotation terms of the ten communities 55

4.3 Enriched disease category annotations of the ten communities 56

4.4 Enriched pathway category annotations of the ten communities . . . 57

List of Figures

2.1 Schematic flow of deep learning approaches 13

3.1 Score categories of SFARI genes 21

3.2 The orbits of 4- and 5-node graphlets 25

3.3 Node degree distribution of the HMIN 27

3.4 The degree distribution of autism-associated genes 28

4.1 The illustration of network embedding 31

4.2 Overview of a general graph neural network 35

4.3 PANDA overview 36

4.4 Schematic demonstration of PANDA for node classification 38

4.5 Learning the node embeddings 39

4.6 *Precision@k* chart 45

4.7 The enrichment distribution of autism-associated genes of the independent sequencing study 46

4.8 Specificity enrichment tests of neuron-related disease 47

4.9 Comparison of classification accuracies of PANDA, support vector machine, random forest, and linear genetic programming 48

4.10 The histogram of the community sizes in the autism module 50

4.11 The pairwise correlations of six network metrics 51

Chapter 1

Introduction

Biomedicine studies acknowledge that susceptibility to common diseases is multifactorial, and both genetic and environmental factors play a crucial role [1]. Abnormalities in certain genes can either predispose individuals to a disease or directly account for the manifestation of a disease phenotype [2]. Thereby, deciphering the association of genes with a specific disease helps better understand the etiology of the disease, leading to better diagnosing the disease, designing therapeutic strategies, and even preventing the disease [3, 4, 5].

Understanding the genetic etiology of complex diseases is one of the greatest challenges in modern biomedicine research [6, 7, 8]. Many common diseases are speculated to have complex genetic architecture, and a substantial number of genes may contribute collectively to the manifestation of a disease [9, 10]. However, the identification of genes associated with a disease, such as linkage studies [11], genome-wide association studies [12] and RNA interference screens [13], requires time-consuming and expensive biological experiments to evaluate a considerable number of possible

candidates [14, 15, 16, 17].

Computational methods *in silico* can successfully facilitate more targeted downstream biological evaluation experiments [18]. Cooperative endeavours are requested from various research fields, ranging from computer science and statistics to biochemistry. Due to the interdependencies of molecular components, identifying genetic variants contributory to a disease needs not only to systematically study molecular functionality independently but also to look into the interconnectivity of molecular components [19]. In order to identify disease-associated genes, systems biology has seen increasing applications of computational approaches that model the interactions among multiple constitutes in human cellular systems [6, 20, 21, 22, 23].

The inheritable disease we are particularly interested in is autism spectrum disorder. Autism is a neuropsychiatric disorder characterized by impaired social interaction and communication, repetitive behaviour, and restricted interests [24]. Sequencing-based studies suggest that complex neurodevelopmental phenotypes of autism are driven by a multitude of genomic variants across the genome [25, 26, 27]. Large-scale family-based exome sequencing studies have unraveled autism-associated genetic variants [28, 29, 30, 31]. Although over 1,000 genes have been identified to influence autism susceptibility, only about 7% of them have shown significant associations with the disease [32, 33, 34].

Technological advances in genomics have led to an explosion of molecular and cellular data from large number of samples. The rapid increase in biological data dimension and acquisition rate is challenging conventional analysis of disease-gene associations. With impressive recent advances made in applications ranging from computer vision to natural-language processing, deep learning methods, a class of

machine learning techniques, have demonstrated the promising capability of identifying highly complex patterns in large datasets. The crux of deep learning problems is searching for appropriate representations for input data, which makes the learned representations amenable to the task at hand.

The combination of genetics and molecular biology has greatly facilitated the identification of candidate genes for human diseases. Genes associated with similar disorders show a higher likelihood of physical interactions between their products. The molecular interconnectivity in human suggests the complex genetic architecture of autism, meaning that the genetic abnormality of autism is not restricted to the activity of single gene aberration, but can involve multiple genes from different molecular pathways [35]. Although the data are complex, the network information is very important for bioinformatics analysis of autism-associated genes, since the topological and interaction information often have a clear biological meaning.

To address these challenges, network science [36] recently has seen ever increasing applications to disease-gene association studies. Network science studies entities (as nodes) and their pairwise relationships (as edges), and can be a powerful tool to discover interaction patterns among biological components. To unravel the links between genes and diseases, network algorithms rely on the premise that phenotypically similar diseases are caused by genes that are functionally related. Previous network-based studies have demonstrated that cellular interaction networks, i.e., protein-protein interaction networks, gene co-expression networks, and metabolic networks, can be used to understand the molecular basis of gene-disease associations [37, 38, 39, 40, 41, 42].

On the other hand, machine learning techniques have been extensively explored in disease genomics research [43]. Particularly, with the impressive recent advances

in computer vision and natural-language processing, deep learning has been rapidly gaining popularity in bioinformatics as well [44]. The flexibility of multi-layered neural networks extends their usage of diverse datasets, ranging from DNA and RNA microarrays to gene expression profiles [45, 46, 47, 48, 49]. However, standard deep learning approaches intuitively take data in Euclidean domain as input, such as images (2-dimensional space), text(1-dimensional space), and gene expression profiles (n-dimensional space) [50], and do not explicitly process data from the non-Euclidean domain, such as graphs [51]. This limitation holds back its utilities for data naturally represented as graphs, such as biological networks, social and knowledge networks, and physical systems.

Graph neural network (GNN) has recently become a promising technique of learning with graph-structured data [52]. GNN is a family of deep learning methods that directly analyze data structured as graphs [53]. GNN can extract local spatial features on both node- and graph- levels directly from a graph and compose them to build highly expressive representations, not only capturing the high nonlinearity of the graph but also preserving the spatial patterns of the nodes [51]. Due to its outstanding performance on various applications, such as social networks [54], protein interface inference [55], physical systems [56], and knowledge graphs [57], GNN has been receiving a surge of attention on different graph inference tasks, such as node classification, link prediction, and graph classification. Although many machine learning approaches have been proposed in order to predict disease-associated genes, to the best of our knowledge, few have explored the idea of designing GNN approaches for the task.

In this thesis, we proposed a bioinformatic framework, *P*rioritization of *A*utism-

genes using *N*etwork-based *D*eep-learning *A*pproach (PANDA), and aimed to identify potential genes associated with autism across the human genome by designing a GNN classifier that used the human molecular interaction network (HMIN) as input for training. Our research starts by constructing the HMIN, which provides a scaffold of the connectivity patterns and structural properties of autism-associated genes. We then compile the set of autism-associated genes, which was used to train the GNN and to discover novel genes that may associate with autism.

Chapter 2

Related Work

2.1 Disease-gene association studies

Many common human diseases can be observed passing from one generation to the next. The information about the genetic basis of human diseases lies at the heart of precision medicine and drug discovery [58]. Understanding what genes cause a specific disease helps better diagnose and treat the disease, and may even prevent the disease if predicted accurately at early stages and effective preventive actions are taken. However, the associations of genes and diseases are still largely unknown, and research is required to study and predict such relationships.

In recent years, with the completion of the *Human Genome Project* [59], genetic markers spanning the entire human genome have empowered widespread mapping efforts based on linkage analysis using families with a number of affected individuals, leading to the discovery of multiple genes for Mendelian diseases. Yet, linkage studies have had only limited success in identifying genes for complex diseases, such as

autism, heart disease, cancer, and psychiatric disorders. With the improvement in genotyping technology, focuses have been shifted from gene mapping in humans to genetic association studies [60]. With greater power and resolution of disease-associated gene locations than linkage analysis, genetic association studies provide renewed hope for mapping genes to complex diseases.

Disease-gene association studies are a group of approaches to associate candidate genes with common diseases [61]. Research on the relationships between genes and diseases has accelerated as a result of both the completion of the human genome [59] and the advance of the Next Generation Sequencing technologies [62]. In contrast to genome-wide association studies (GWAS), which scan the entire human genome for common genetic variations, disease-gene association studies often focus on exploiting genetic alterations with pre-specified genes of interest with a priori knowledge about their functional impact on the disease in question.

A disease-gene association study is considered as a useful initial step in exploring potential causal pathways between genetic markers and complex diseases. Once a statistically significant association of a gene is ascertained, the same gene and its variants can be further examined in independent populations [63]. In addition, the identified genes and variants allow molecular biologists designing experiments to find their functional roles in biological processes and disease pathology, providing strong support for causality.

2.2 Network science-based disease-gene association studies

Networks are ubiquitous [36]. A network is a collection of *nodes* joined together in pairs by *edges*. They are natural representations for encoding relational structures encountered in many complex systems, such as biological systems, social networks, and physical systems [36]. Networks capture the patterns of interactions between the components of a complex system. For instance, connections in a social network affect how people learn, shape opinions, and diffuse ideas, as well as other less obvious phenomena, such as the spread of disease. In biology fields, biological networks agglomerate the interactions of molecules in living cells, which provide the perspectives in understanding the dynamics and mechanisms of complex biological systems.

A major challenge in systems biology and medical genetics is to understand how interactions among genetic variants contribute to complex diseases. Networks have become an emerging trend for representing the structure of a cellular system that creates a bridge between biological data and a large toolkit of powerful analysis techniques. Indeed, the enormously complex interactions among molecules within a cell and even among cells present researchers difficulties to acquire knowledge from them. Over the past decades, systems biology and medical genetics have seen rapid advances in network biology [64]. Increasing number of studies have utilized network-based analysis to unravel the relationships between a group of genes and a disease of interest [23, 65, 66].

Yonan *et al.* [67] constructed gene pathway networks based on prior genetic evidence from the allelic association literature to query the known transcripts within the

1-LOD (logarithm of the odds) support interval for each region. They used biological databases and the pathway networks to identify a subset of biologically meaningful, high priority candidates, which contained 383 positional autism candidate genes. Corominas *et al.* [68] introduced an interactome mapping approach by experimentally identifying interactions between brain-expressed alternatively spliced variants of autism risk factors. Suthram *et al.* [69] presented the first approach integrating high-throughput datasets such as mRNA expression and large-scale protein-protein interaction networks to discover human disease relationships in a systematic and quantitative way.

Other network-based studies on disease-gene association have also focused on the protein-protein interaction (PPI) network. Sun *et al.* [6] analyzed and compared four publicly available disease-gene association datasets and applied three disease similarity measures, namely annotation-based measure, function-based measure, and topology-based measure, to estimate the similarity scores between diseases. The result demonstrated that the predicted disease associations correlated with disease associations generated from genome-wide association studies significantly higher than expected at random. A recent study [41] focused on studying the protein-protein interaction network structure of disease pathways. They defined different sets of proteins associated with diseases and found that disease pathways are fragmented and sparsely represented in the PPI network, and that spatial clustering of disease pathways within the PPI network is statistically insignificant.

2.3 Machine learning-based disease-gene association studies

Machine learning techniques have been extensively explored in computational biology under various scenarios, such as classification, regression, clustering, and feature selection [43]. A machine learning approach optimizes a performance criterion by using example data or past experience. Machine learning-based computational approaches provide a convenient framework for taking advantages of the exponential growth of the amount of available biological data [70].

The use of machine learning methods in the context of genome-wide data on genetic variants has yielded an ever-growing number of studies in recent years, compared to a large number of machine learning studies on other types of genomic datasets, especially genome-wide gene expression profiles [71, 72]. Further, the combination of predictive modeling and disease-gene association studies have yielded quite positive results [71]. Indeed, many studies have suggested that the use of machine learning approaches are capable of identifying genes contributing to certain diseases [72, 73].

In recent years, advanced computational and engineering methodologies have been employed to meet the needs of cross-disciplinary applications in biomedicine. Brown et al. [74] introduced a method of functionally classifying genes by using gene expression data from DNA microarray hybridization experiments. The method was based on support vector machines (SVMs). The method demonstrated that SVMs could accurately classify genes into functional categories based on expression data. SVM as a popular supervised learning algorithm has been used by other studies aiming to prioritize autism genes. Kou et.al [75] applied several SVM kernels to prioritize

autism gene candidates based on curated lists of known autism genes. The proposed approach aimed to learn an SVM classifier for prioritizing pluripotency stem cell regulators from RNAi screens using microarray and ChIP-seq data. Duda et al. [76] constructed a machine learning model by leveraging a brain-specific functional relationship network (FRN) of genes to produce a genome-wide ranking of autism risk genes. Through functional enrichment analysis on their highly prioritized candidate gene set, they identified a small number of pathways that are key in early neural development, providing further support for their potential role in autism. Recently, Krishnan et al. [77] developed a machine-learning approach based on a human brain-specific gene network to present a genome-wide prediction of autism risk genes. By using the linkage of nodes, they identified a large set of autism genes converges on a smaller number of key pathways and developmental stages of the brain.

2.4 Deep learning on biological data

While conventional machine learning approaches have seen great potential for making use of genetic data, they were limited in their ability to process natural data in their raw form [78]. Moreover, the rapid increase in biological data dimension and acquisition rate is challenging conventional analysis strategies. Decades-long efforts have been put to constructing machine-learning systems that required carefully manually-engineered features. Considerable domain experts need to design an effective feature extractor that transformed the raw data into a suitable internal representation from which a model is able to learn.

Impressive advances in computer vision and natural-language processing have been

revolutionary over the past decade [79]. Deep learning, a family of neural network-based machine learning methods, has been rapidly gaining in popularity among bioinformatic scientists [44]. The flexibility of multi-layer neural networks extends their applications on diverse genetic datasets from DNA and RNA microarrays to gene expression profiles [45, 46, 47, 48, 49]. However, standard deep learning approaches intuitively take data in Euclidean domain as input, such as images (2-dimensional space), text(1-dimensional space), and gene expression profiles (n-dimensional space) [50], and do not explicitly process data from the non-Euclidean domain, such as graphs [51]. This limitation holds back its utilities for data naturally represented as graphs, such as biological networks, social and knowledge networks, and physical systems.

To better utilize the graph-structured datasets, representation learning on graphs has shown its effectiveness in many learning tasks, such as prediction or classification [80]. Deep-learning methods can serve as representation-learning methods with multi-level representation, allowing a model to deal with raw data and to automatically discover suitable representations for downstream learning tasks. Deep neural networks (DNNs) are efficient algorithms based on the use of compositional layers of neurons. Figure 2.1 displays the general workflow of a deep neural network. In computational biology, their appeal is the ability to derive predictive models without a need for strong assumptions about underlying mechanisms, which are frequently unknown or insufficiently defined.

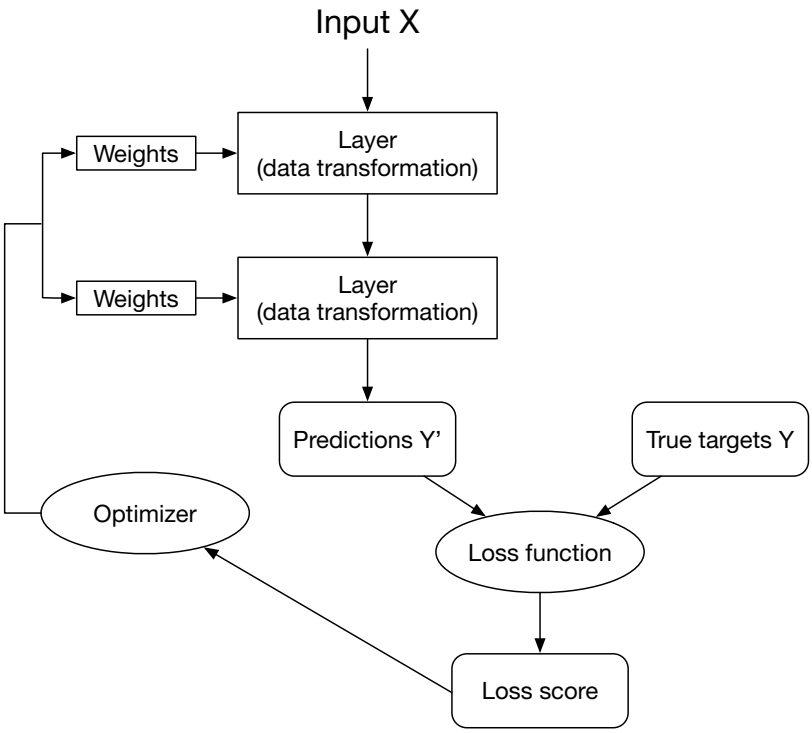

Figure 2.1: Schematic flow of deep learning approaches. A deep neural network is parameterized by its weights on each layer. A loss function measures the quality of the network output. The loss score is used as a feedback signal to adjust the weights.

2.5 Graph neural network

As stated in the previous section, data in many scientific fields can be represented as graphs, including biological, financial, social, and knowledge networks [36]. Graphs can not only serve as useful structured knowledge repositories but also play a vital role in modern machine learning [81]. Thus researches of learning on graphs have been receiving ever-growing attention due to the great expressive power of graphs. For example, modeling physics systems, learning molecular fingerprints, predicting

protein interfaces and classifying diseases require that a computational model learns from graph inputs.

Graph-structured data contains rich geometric relational information among elements of a complex system. As a unique non-Euclidean data structure for machine learning, statistical analysis of graphs focuses on node classification, link prediction and clustering [82]. Traditionally, machine learning algorithms rely on domain expert-defined heuristics to extract features in order to encoding structural information about a graph [83]. In recent years, exploiting graph-structured data effectively inspired the advances of graph neural networks [84].

Graph neural network (GNN) is a family of deep learning algorithms that operate on a graph and organize their computation over the graph structure [53]. GNNs are connectionist models that capture the dependence of graphs via message passing between nodes of graphs. The primary goal of GNNs is finding an effective way to encode graph structure, so that the downstream learning tasks can exploit the graph.

For various graph learnings tasks, e.g., node classification, clustering, and edge prediction, GNNs often require different designs. GNNs broadly follow a recursive neighborhood aggregation scheme [53]. Similar to message passing, each node aggregates feature vectors of its neighbors to compute its new feature vector. After k iterations of aggregation, a node is represented by its transformed feature vector, which captures the structural information within the node's k-hop neighborhood. In chapter 4, we describe in detail the framework and computation procedures of a general GNN.

2.6 Summary

In this section, we introduced the concept of disease-gene association and its position in exploring potential causal pathways between genetic markers and complex diseases. We then described two main trends of computational approaches to disease-gene association studies. The first trend is to apply the concepts and tools of network science to model and analyze the underlying interaction relationships among genes in human cells. The second trend is to design machine learning-based methods to find the models aiming at predicting the disease-associated genes and interpreting their roles in the etiology of the diseases. Finally, we introduced the concepts and applications of deep learning, especially graph neural network, on the biological data, which demonstrated their power and flexibility to explore the genomic data.

Chapter 3

Disease-Gene Association Studies for Autism

3.1 Background

3.1.1 Autism spectrum disorder

Autism spectrum disorder is a neurodevelopmental disorder typified by striking deficits in social communication and genetically by a mixture of *de novo* and inherited variation contributing to liability [85]. Numerous epidemiology studies have reported that 1 in 68 children is diagnosed with autism, with a 3 to 4-fold increased risk for boys [86]. Family and twin studies have found that autism is highly heritable, most caused by a combination of genetic and environmental influencers [87]. The genetic risk factors behind autism, however, are highly heterogeneous and over a thousand genes across the genome are estimated to be involved with no single gene accounting for more than 1-2% of the cases [23, 88]. Almost all genetic risk factors for autism can

be found in the general population, but the effects of these risk factors are unclear in people not ascertained for neuropsychiatric symptoms [89].

Sequencing-based discovery efforts have produced valuable catalogs of genetic variants that point toward potential causal autism genes [90]. Decades-long efforts to explain the causes of autism have produced an impressive list of disease-gene associations. Yet only a small fraction of potentially casual genes are known with strong genetic evidence from sequencing studies [77]. Unraveling the genetic background of autism serves a number of goals. One aim is to identify genes that modify the susceptibility to autism. When a set of autism risk genes can be identified with significantly high frequencies, we have the potential information to learn about the pathogenesis of the disease, and we can identify possible targets for therapeutic interventions. Another goal is to classify autism patients on the spectrum according to their risk for autism or to make risk predictions on autism.

Recently, copy number variations (CNVs) were strongly associated with autism [91]. Additionally, autism is consistently associated with a number of specific genetic disorders such as Fragile X syndrome amongst others [92]. Single-gene variants are also linked to rare cases of autism [93]. The high genetic heterogeneity of autism poses an enormous challenge for understanding disease etiology.

3.1.2 Autism gene databases

To establish our understanding of autism-associated genes, we explored three major databases, including SFARI Gene 2.0 [94], AutDB [95] and Online Mendelian Inheritance in Man (OMIM) [96]. Each resource gathered data from sources with

different levels of evidence, ranging from recurrent mutations in patients with autism to nebulous links gleaned from text-mining thousands of PubMed abstracts.

SFARI Gene 2.0 database is a web-platform developed for the ongoing collection, curation, and visualization of genes linked to autism in order to enable systematic community driven assessment of genetic evidence for individual genes with regard to autism. Several types of genetic variations, such as common variants of small effect or single-gene variants of large effect, can contribute to autism [89]. Structural variations in the genome, such as microdeletions or duplications, are also associated with the disorder [97]. Therefore, SFARI database keeps track of numerous susceptibility genes uncovered by advanced high-throughput approaches.

AutDB is a publicly available web-portal repository for on-going assembling, manual annotation, and visualization of genes associated with autism [95]. The content of AutDB is gleaned entirely from published scientific literature and is manually annotated by expert biologists. Online Mendelian Inheritance in Man (OMIM) is a comprehensive, authoritative compendium of human genes and genetic phenotypes that is freely available [96].

3.1.3 Network biology for molecular interaction network

A key aim of research on postgenomic biomedicine is to systematically study all molecules and their interactions within a living cell. Network biology [39] has brought about a shift in the paradigm of elucidating disease pathologies from analyzing the impacts of single genes to understanding the structures and dynamics of molecular interaction networks [98]. Indeed, most phenotypes reflect the interplay of multiple

molecular components that interact with each other. Network-based analysis strategies offer a quantifiable description of various biological systems [69].

Molecular interaction networks have been exploited to nominate novel candidate disease-associated genes, based on the assumption that the neighbors of disease-related genes in a network are more likely to be involved in similar disease traits. Such interconnectivity implies that the abnormal function of a molecule is not only confined to that molecule but can spread along the links in the network, which results in a dense cluster or community, called disease module. Yet, due to the limited knowledge of disease-associated genes, only a fraction of the known disease-associated genes are known to physically interact with each other. This suggests that analyzing dense communities of known disease-associated genes restricts the discovery of all disease-associated genes.

The elements of molecular interaction networks range from metabolites to proteins. Understanding the function and dysfunction of over 20,000 protein-coding genes reveals how proteins assemble into functional modules and networks engaged in specific biological activities. In this thesis, we construct the human molecular interaction network (HMIN) in order to provide a scaffold of gene-gene relationships that helps identify candidate genes according to their structural similarities to known autism genes. In the following section, we describe in detail our HMIN.

3.2 Method

3.2.1 Training set of genes compilation

To collect prior known autism-associated genes as training samples, we compiled the supervised training set of genes using SFARI Gene 2.0 [94] and Online Mendelian Inheritance in Man (OMIM) [96]. The SFARI Gene 2.0 database is exclusively for the autism research community and has a collection of manually annotated autism-associated genes. It assigned each gene a score ranging from 1 (highest association confidence) to 6 (no role in autism) to quantify its association with autism, see Figure 3.1. Score 1 and 2 represent the strongest evidence of autism association, score 3 and 4 show relaxed criteria of autism association, score 5 marks genes hypothesized but without tested associations, and score 6 genes have no supporting evidence to be related to autism. In addition, the OMIM database is a comprehensive, authoritative and updated knowledge base of human genes and genetic disorders compiled to support human genetics research and education and the practice of clinical genetics.

As described above, score 5 includes genes for which the only evidence comes from studies of model organisms, without statistical or genetic support in human studies and score 6 marks genes whose evidence argues against a role in autism. Therefore, we retrieved 732 genes of categories 1 to 4 from the SFARI database. In OMIM, we extracted 28 genes from OMIM using the terms of "autism", "autism spectrum disorder", "ASD". All data were retrieved in November 2018. Overall, a total number of 760 autism-associated genes were used as positive instances for the supervised training in the proposed method (a.k.a. PANDA).

On the other hand, we retrieved 1,146 genes that have shown no association as

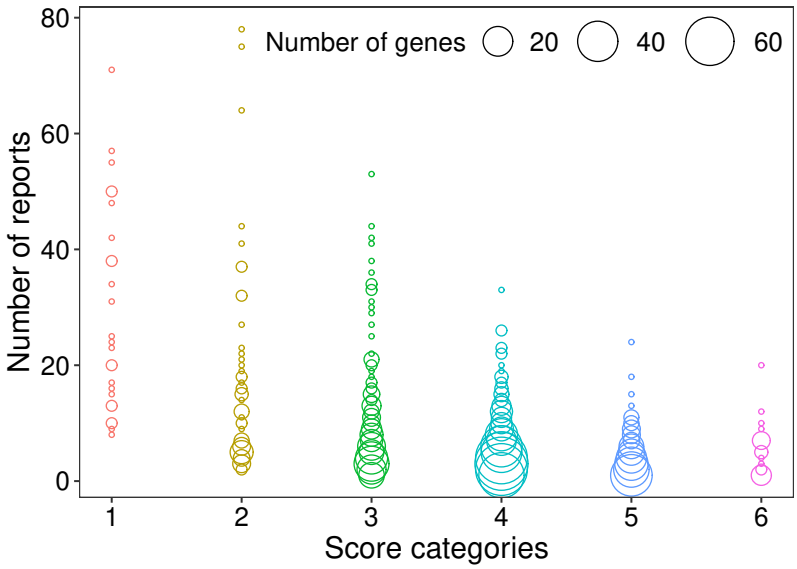

Figure 3.1: Score categories of SFARI genes. Score 1 and 2 represent the strongest evidence of autism association, score 3 and 4 show relaxed criteria of autism association, score 5 marks genes hypothesized but without tested associations, and score 6 genes have no supporting evidence to be related to autism. In the figure, each point represents the mapping of a score category and its associated number of reports, and the size of a point denotes the number of genes that have the given score and the corresponding number of reports.

the negative instances curated by brain-disease experts [76]. Our integrated HMIN covered 1,102 out of the 1,146 genes. Given the fact that genes in the positive set have different strengths of association with autism, we placed them into three confidence levels, where each level was assigned with a confidence value from {0.5, 0.75, 1.0} (see Table 3.1). For the set of negative genes, their confidence values were set to -1.0. The association with autism of the positive and negative genes served as the labels of our training set genes for the subsequent supervised training of the classifier in PANDA.

Table 3.1: Gene labels based on their confidence levels of autism association

Confidence	**Data Source**
1.0	Genes in SFARI category 1
0.75	Genes in SFARI category 2 and OMIM autism entry
0.5	Genes in SFARI category 3 and 4
-1.0	Genes without autism association

3.2.2 Human molecular interaction network

We integrated the HMIN based on two sources. The first data source was a previously well-established human protein-protein interaction network [99]. This network included data curated up to 2015. We updated this network by integrating newly discovered protein-protein interactions using BioGRID version 3.5.167, released on November 25th, 2018 [100]. BioGRID is a public repository with data of genetic and protein interactions.

Our HMIN has 23,472 nodes (genes) and 405,618 edges (interactions), representing their pairwise relationships. The HMIN covers 732 positive autism-associated genes and 1,102 negative genes in our training set (Section 3.2.1). It included physical interactions experimentally annotated in human cells, such as transcription factor regulatory interactions, metabolic enzyme-coupled interactions, and protein-protein interactions. Note that we treat the HMIN as an unweighted and undirected network. The hypothesis is that the manifestation of autism is unlikely a consequence of the dysfunction of a single gene product, but implies various pathological processes that interact as captured in the HMIN [19]. The interaction patterns of known autism-associated genes may imply these pathological processes, and can be utilized to predict

novel autism genes. Therefore, we use such an interaction network and aim to discover candidate genes that are structurally similar to known autism-genes.

In the first-stage experiments, we investigated both global and local network properties of the HMIN and autism-associated genes in the HMIN. Here, we briefly describe the definitions of the global network properties.

- **Giant connected component:** A component is subgraph $\mathcal{G}_0$ of network $\mathcal{G}$ that the sets of its nodes and edges are subsets of those of $\mathcal{G}$. A graph is connected if there is a path, which is a sequence of edges, between any pair of its nodes. A giant connected component refers to a connected subgraph of the network larger than any other connected components that includes the majority of nodes.

- **Average shortest path length:** A path is a sequence of nodes $P = (v_1, v_2, ..., v_m)$ where v_i and v_{i+1} are connected by an edge for $i \in [1, m)$. In an unweighted network, the path length equals the number of edges traveled in the path. Due to the possible existence of multiple paths between any pair of nodes, the shortest path between a pair of nodes is defined as the one with the shortest length. Thus the average path length in the network is the average length of the shortest paths between every pair of nodes.

- **Network diameter:** The diameter of a network refers to the longest shortest path in the network. The definition is as $D = \max(d_{i,j}), \forall v_i, v_j \in \mathcal{V}(\mathcal{G})$, where $d_{i,j}$ is path length for nodes within the same connected component.

- **Network density:** The density of a network refers to the ratio of the number of

edges and the number of possible edges. Network density measures the sparsity of the network.

- **Clustering coefficient:** It is defined as the fraction of length-two paths in the network that are closed. The path $P = (v_1, v_2, v_3)$ is closed if there exists a third edge from v_3 to v_1. Clustering coefficient quantifies the degree of transitivity in the network, indicating the likelihood of an interaction (i.e. connected by an edge) between gene products A and C given that there are edges between gene products A and B as well as B and C.

As local network properties, centrality quantifies how important nodes are in a network. We used four network centrality metrics, whose definitions are as follows:

- **Degree centrality** is the number of edges connected to a node.
- **Betweenness centrality** measures the extent to which a node lies on paths between other nodes.
- **Closeness centrality** is a centrality score that measures the mean distance from a node to other nodes.
- **Eigenvector centrality** is an extension of degree centrality. Eigenvector centrality awards a number of points proportional to the centrality scores of the neighbors.

Apart from the network properties stated above, we also utilized three other network metrics, including k-coreness, personalized PageRank and graphlets. Their definitions are as follows:

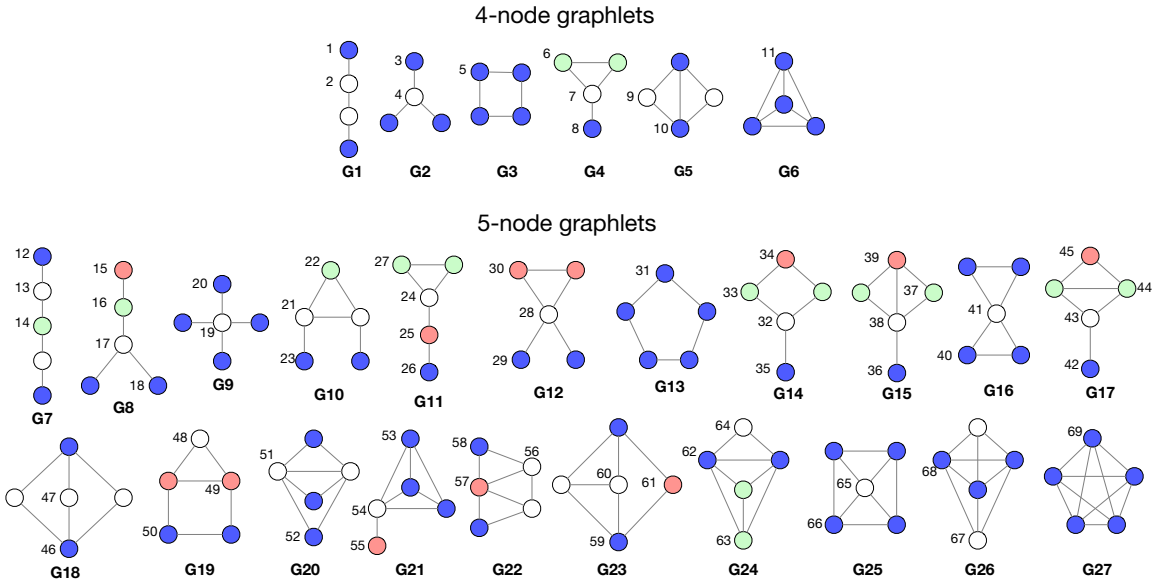

Figure 3.2: The orbits of 4- and 5-node graphlets. Each isomorphic orbit is rendered using four different colors – white, blue, green and red. The colors do not distinguish the importance of each orbit.

- **k-coreness** The k-core of graph is a maximal subgraph in which each node has at least degree k. The coreness of a node is k if it belongs to the k-core but not to the $(k+1)$-core.

- **Personalized PageRank** is a weighted version of conventional PageRank [101]. The weights in personalized PageRank are defined by users.

- **Graphlets** A graphlet is a small connected non-isomorphic induced subgraph of a large network [102]. Orbits refer to distinct positions of vertices in a graphlet. There are 69 different orbits in 4- and 5-node graphlets, see Figure 3.2.

Table 3.2: Network properties of the HMIN

Property	**Value**
Number of nodes	23,472
Number of edges	405,618
Number of connected component	4
Network Diameter	8
Network Density	0.001473
Clustering coefficient	0.107435
Average node degree	34.561861
Average shortest path length	3.203211

3.3 Results

3.3.1 Investigation of the HMIN

First, we investigated the global network properties of HMIN. Table 3.2 shows some fundamental network properties of the HMIN. In the HMIN, each node has a degree ranging from 1 to 2,393 with an average of 34.562. The degree distribution of the HMIN can be seen in Figure 3.3, which is approximately a power-law distribution, suggesting a scale-free structure. Most of the genes only interact with a handful of other genes, while some can interact with one or two thousands of others.

The HMIN is highly connected, with only four components, and the giant connected component (GCC) includes 23,465 nodes. The other seven nodes form islands of two to three nodes. We used the GCC of the HMIN as the input network to train the GNN classifier in PANDA, described in Chapter 4. Moreover, the average shortest

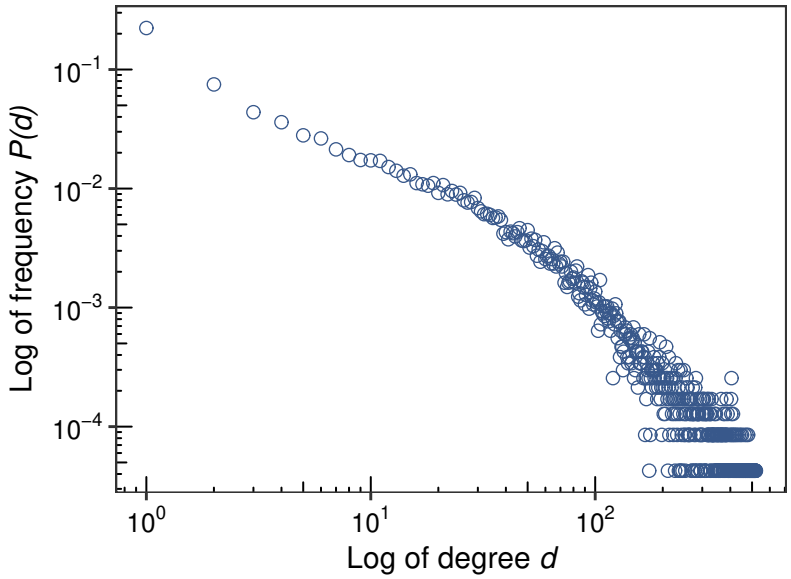

Figure 3.3: The node degree distribution of the HMIN. The distribution is approximately power-law, suggested by the straight line correlation in a log-log scale.

path length is roughly 3, suggesting that the HMIN has a small-world property [103]. The small-world effect of HMIN implies that the neighbors of two given nodes are likely to be neighbors of each other and most genes can be reached from every other gene by a small number of steps.

3.3.2 Investigation of the known autism-associated genes

We inspected the degree distribution of autism-genes in the GCC of HMIN. Figure 3.4 shows the distribution of autism-genes. The degrees of autism-associated genes range from 1 to 665 with the average degrees of 76.32368. Nodes of unusually high degree are called hubs. The hubs of the GCC are at least 1267 degrees. Compared with the hubs of the HMIN, autism-genes usually are distributed peripherally in the HMIN.

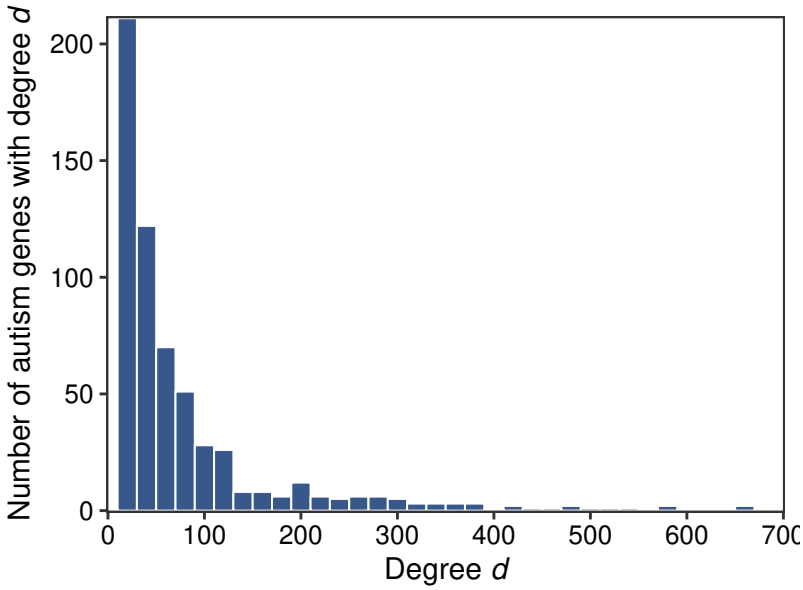

Figure 3.4: The degree distribution of autism-associated genes in the GCC. In the figure, the scales of x-axis and y-axis are natural numbers. The majority of autism-associated genes are distributed peripherally in the GCC of the HMIN.

To further investigate the distribution structures of autism-associated genes, we extracted the subgraphs from the HMIN using the autism-associated genes as seeds. Table 3.3 summarizes the size and number of detected subgraphs. The majority of known autism-associated genes are connected within a large component with 643 nodes. On the other hand, the rest of the autism nodes are scattered in the HMIN, suggesting that there may exist other uncovered autism-associated genes that can joint the large component with the remaining nodes.

Table 3.3: Measurements of subgraphs extracted using autism-associated genes in HMIN.

Subgraph sizes	**Number of occurrence**
634	1
2	2
1	121

3.4 Summary

In this section, we first introduced the definition of autism spectrum disorders and described the databases of autism genes. Then, we described in detail the human molecular interaction network and the network properties we used as node properties in later study. We performed initial network analysis of our HMIN and the known autism-associated gene in the GCC. The results demonstrated an approximate power-law distribution of the degrees in the HMIN. Further analysis of the degree distribution of the autism-associated genes indicated that most of the known autism genes are peripherally distribution in the GCC of the HMIN. In the coming section, we will describe in detail our design of PANDA algorithm.

Chapter 4

Prioritization of Autism-Genes Using Network-Based Deep-Learning Approach

4.1 Background

4.1.1 Learning on networks

4.1.1.1 Network embedding

Network embedding is an important approach to learn low-dimensional representations of nodes in networks, aiming to capture and preserve the network structure [104]. Modeling the interactions between entities as graphs has enabled scientists to understand the various network systems in a systematic manner [36]. Network embedding is capable of supporting subsequent network analytic tasks, such as node classifica-

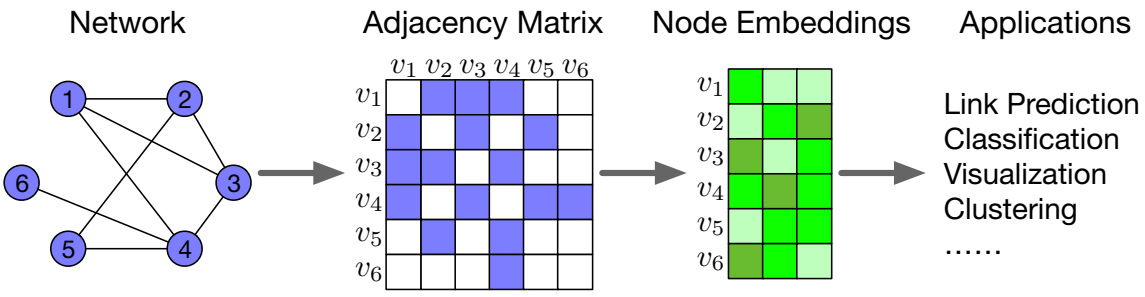

Figure 4.1: The illustration of network embedding [84]. In the third panel, the different colors denote that after embedding, different nodes are transformed to distinct representation of the original graph space. Each row of the embedding matrix is the embedding of the corresponding node.

tion, link prediction, clustering, and network visualization, yielding insight into the structure of society, communication patterns, and different mechanisms of biological components, see Figure 4.1 [84].

Typically, a model defined to solve graph-based problems either operates on the original graph adjacency matrix or on a derived vector space. Essentially, network embedding exploits the latter way to represent network structures. Many approaches have been proposed to represent network structure [84, 83]. The traditional techniques of network representation raise several issues on performing network processing and analysis. On the one hand, network data represented in the traditional way, e.g. adjacency matrices, causes severe difficulties to design and implementation of parallel algorithms. On the other hand, traditional network representation is inapplicable for machine learning methods, especially deep learning.

In the past few decades, neural network-based methods have become widely popular for the node classification. The primary goal of such methods is that the learned

embedding space can effectively support the inference of node labels. These methods can be broadly abstracted into two categories of approaches – methods which use random walks to propagate the labels, and methods which extract features from nodes and apply classifiers on them. The embeddings are input as features to a model and the parameters are learned based on the training data. This obviates the need for complex classification models which are applied directly on the graph.

In summary, the structure and property preserving network embedding plays a vital role in subsequent learning tasks. If one cannot preserve well the network structure and retain the important network properties in the embedding space, serious information is loss, which consequently hurts the downstream analytic tasks [84].

4.1.1.2 Challenges of network embedding

Obtaining a vector representation of each node in a network is inherently difficult and presents several challenges which have been motivating research in this field. The first challenge is choosing the property of the graph which the embedding should preserve. A "good" vector representation of nodes should preserve the structure of the graph and capture the connection patterns of individual nodes. Due to the plethora of distance metrics and properties defined for graphs, this choice can be difficult and the performance may depend on the applications.

Second, most real-world networks are large and contain millions of nodes and edges. Embedding methods should be scalable and able to process large-scale graphs. Defining a scalable model can be challenging especially when the model is aimed to preserve global properties of the network.

Last but not least, find the optimal dimensions of the representation can be hard.

The choice can rely on application-specific demands on the approach. For example, lower number of dimensions may result in better link prediction accuracy if the chosen model only captures local connections between nodes.

4.1.2 Graph neural network

Graph neural network (GNN) is a family of deep learning algorithms [53]. The concept of GNNs extends existing neural networks for processing the data represented in graph domains. GNNs operate on a graph and organize their computation over the graph structure. In many research problems, data can be represented as graphs, including biological, financial, social, and knowledge networks [36]. Exploiting graph structured data effectively inspired the advances of GNN [84].

For various graph learnings tasks, e.g., node classification, clustering, and edge prediction, GNNs often require different designs. In particular, node classification aims at determining the labels of nodes based on other labeled nodes and the topology of the network. In this thesis, we focused on predicting the gene association with autism, i.e., node classification.

Based on convolutional neural network [105] and graph embedding, GNNs are proposed to collectively aggregate information from graph structure. Thus they can model input and output consisting of elements and their dependency. Further, graph neural network can simultaneously model the diffusion process on the graph with recurrent neural network kernel [106].

Though experimental results showed that GNN is a powerful architecture for modeling structural data, the original GNN is inefficient to update the hidden states of

nodes iteratively for the fixed points [51]. In contrast, a multi-layer version of GNN relaxes the assumption of the fixed point, resulting in the capability to obtain a stable representation of node and its neighborhood.

A network can be mathematically represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of edges. In the context of classification, each node has a class label. A GNN is able to learn a d-dimensional real-valued representation $\mathbf{h}_v \in \mathbb{R}^d$ for every node v in $\mathcal{V}$, which is called its *node embedding*. A GNN constructs the node embedding $\mathbf{h}$ by aggregating the neighborhood information of a node. The node embedding can be seen as predictors, which will be used in turn for the prediction of the class label of a node. Figure 4.2 shows the general framework of GNNs in the proposed research [53].

The embedding of node v captures its spatial structure in its neighborhood and in the graph, and can be used to produce an output $\mathbf{o}_v$ such as the class label. A GNN has two computation modules, i.e., embedding generation and output generation. First, a *local transition function* f, shared among all nodes, aggregates the information from the neighborhood of each node to update the node embedding. Then, a *global output function* g takes a node's embedding as input and returns the prediction of its label as output. The node embedding $\mathbf{h}_v$ and the output $\mathbf{o}_v$ are defined as

$$\begin{aligned} \mathbf{h}_v &= f(\mathbf{x}_v, \mathbf{x}_{nb[v]}, \mathbf{x}_{co[v]}, \mathbf{h}_{nb[v]}), \\ \mathbf{o}_v &= g(\mathbf{h}_v, \mathbf{x}_v), \end{aligned} \tag{4.1}$$

where $\mathbf{x}_v, \mathbf{x}_{nb[v]}, \mathbf{x}_{co[v]}, \mathbf{h}_{nb[v]}, \mathbf{h}_v$ denote the network properties of v, those of v's neighborhood nodes, the properties of v's edges, the embeddings of v's neighborhood nodes, and the embedding of v, respectively. The learning of functions f and g can be implemented using feed-forward neural networks [53].

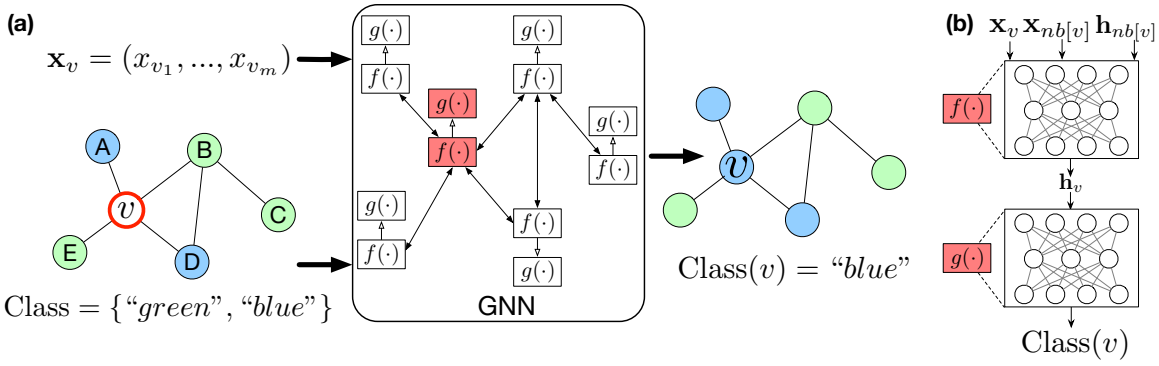

Figure 4.2: Overview of a general graph neural network (GNN). **(a)** This example graph has six nodes {v, A, B, C, D, E}, where the class label of v needs to be assessed. A GNN has two essential functions, the local transition function $f(\cdot)$, which learns the embedding representation of a node, and the global output function $g(\cdot)$. The vector $\mathbf{x}_v$ includes m network properties of node v, such as node degree, centralities, etc. First, f generates an embedding $\mathbf{h}_i$ of each node i by iteratively aggregating its $\mathbf{x}_i$ and its neighboring nodes' network properties ($\mathbf{x}_{nb[i]}$) and embeddings ($\mathbf{h}_{nb[i]}$). Then, g uses the embeddings to predict the class label for node v. **(b)** Both f and g can be learned using feed-forward neural networks.

4.2 Method

4.2.1 PANDA overview

Our proposed framework, *P*rioritization of *A*utism-genes using *N*etwork-based *D*eep-learning *A*pproach (PANDA), include the following steps. We started by integrating the human molecular interaction network (HMIN) using the literature of physical protein interactions experimentally documented, where nodes are proteins mapped to their corresponding genes, and an edge indicates the existence of interactions between

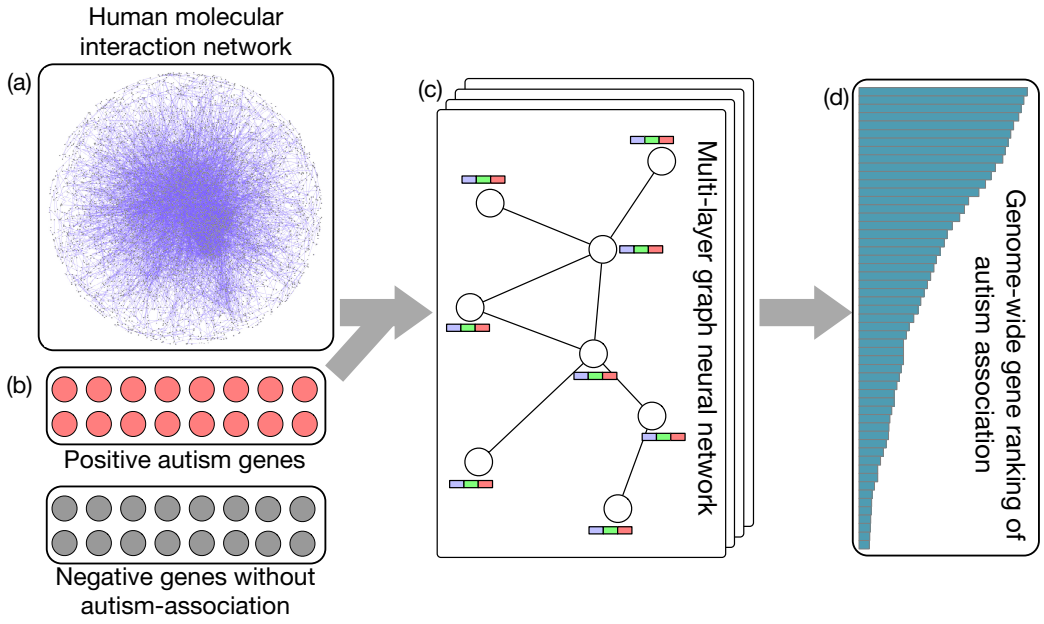

Figure 4.3: PANDA overview. Using the graph input of **(a)** the human molecular interaction network (HMIN), PANDA employs **(b)** a ground truth set of known autism-associated genes as positive instances and a set of genes that have been shown without autism association as negative instances. The classifier learns the structural patterns of autism-associated genes through **(c)** a multi-layer graph neural network (GNN) and **(d)** predicts and ranks autism associations for all genes in the HMIN.

two genes (Figure 4.3a). We then compiled a set of known autism-associated genes and a set of confirmed genes without autism association from databases including SFARI [94] and Online Mendelian Inheritance in Man (OMIM) [96] (Figure 4.3b). These two sets of genes were used as positive and negative instances for training the classifier. Next, we designed a multi-layer graph neural network (GNN) that used the HMIN as input (Figure 4.3c). In the final step, we used the trained GNN classifier to predict and rank the probabilities of all genes in the HMIN that influence autism (Figure 4.3d).

4.2.2 Our GNN in PANDA

In PANDA, we utilized the gene-gene interaction patterns captured by the HMIN, and designed a deep graph neural network (GNN) that directly takes the graph of the HMIN as input and outputs a predicted ranking of genes in the HMIN based on their autism associations.

Following the ideas of the general GNN outlined in the previous section, we proposed a GNN tailored for our PANDA framework. Figure 4.4 presents the node classification procedure of PANDA. Specifically, here we discuss the design of the node embedding $\mathbf{h}$, the local transition function f, and the global output function g and its loss function.

We used six network metrics and graphlet orbit frequencies to describe the local structural properties $\mathbf{x}_v$ of node v. These six network metrics include *betweenness*, *closeness*, *eigenvector centrality*, *personalized PageRank centrality* [107], *degree*, and *coreness* [108]. A graphlet is a small connected non-isomorphic induced subgraph of a large network [102]. Orbits refer to distinct positions of vertices in a graphlet. There are 69 different orbits in 4- and 5-node graphlets we extracted from the HMIN.

The local transition function f aggregates the network properties of node v and its direct neighbors (Figure 4.5). Recall that only the training set of genes in the HMIN have labels, so v's neighbors can contain both labeled and unlabeled nodes. For the unlabeled nodes u', we performed the conventional element-wise averaging over them. For the labeled neighbors u, we computed a weighted average using their labels (confidence values, see Table 3.1). The aggregation was computed as follows

$$\mathbf{h}_{nb[v]}^{k-1} = \frac{1}{\sum_u \xi_u} \sum_u \xi_u \mathbf{h}_u^{k-1} + \frac{1}{B} \sum_{u'} \mathbf{h}_{u'}^{k-1}, \tag{4.2}$$

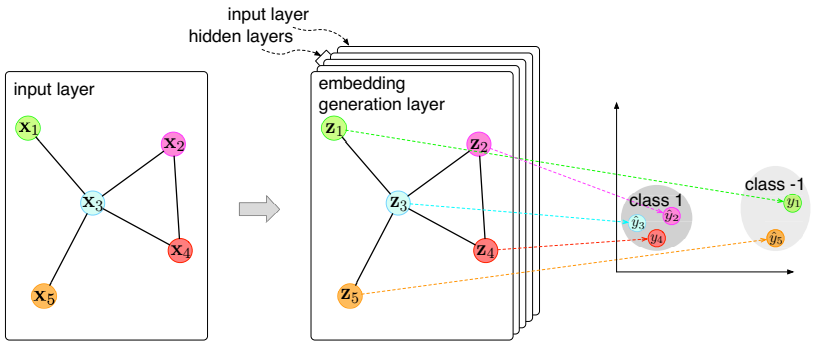

Figure 4.4: Schematic demonstration of PANDA for node classification. Here, the graph with five nodes and five edges is a miniature of HMIN. PANDA learns the node embeddings over the structure of HMIN via a multi-layer graph neural network, aiming to classify the set of nodes into autism-associated gene ($y = 1$) and no-autism-association gene ($y = -1$). Start from the input layer, each node has a m-dimensional network property vector $\mathbf{x}_{i=\{1,2,3,4,5\}} \in \mathbb{R}^m$. Transformed through multiple hidden layers, every node obtains an d-dimensional real-valued embedding $\mathbf{h}_{i=\{1,2,3,4,5\}} \in \mathbb{R}^d$. The output of PANDA is the predicted probabilities of all nodes. In our study, only a subset of nodes have labels. As shown in the figure, only node $\mathbf{x}_1$ and $\mathbf{x}_4$ have labels y_1 and y_4, respectively. By aggregating information from neighbors, the five nodes embeddings transform the representations of the nodes from the original m-dimensional space to the 2-dimensional space.

where $\mathbf{h}_{nb[v]}^{k-1}$ is the $(k-1)th$-layer aggregated embedding of node v's neighborhood, $\mathbf{h}_u^{k-1}$ and ξ_u are the $(k-1)th$-layer embedding and the label of the neighbor node u of v, respectively, and $\mathbf{h}_{u'}^{k-1}$ is the $(k-1)th$-layer embedding of the total number of B unlabeled nodes u'.

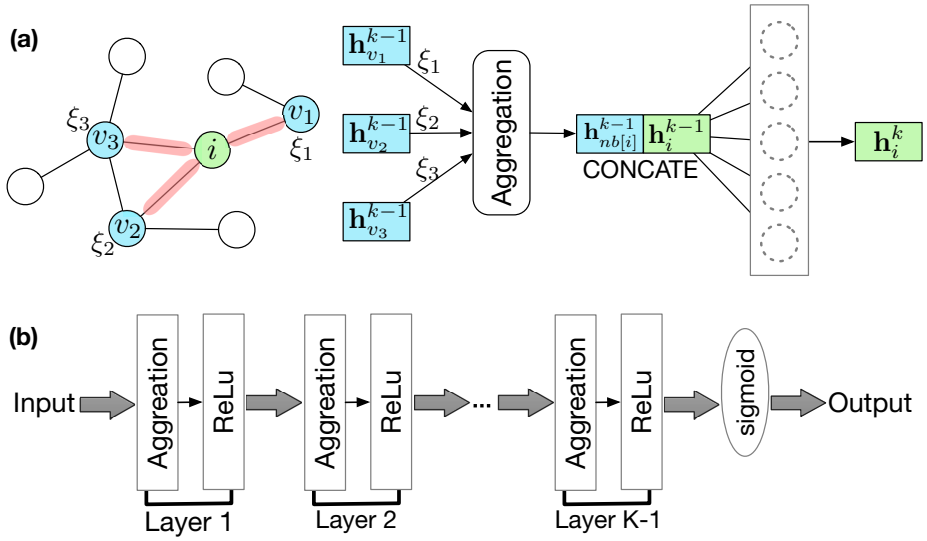

Figure 4.5: Learning the node embeddings. **(a)** Node i has three neighbors v_1, v_2, v_3, each assigned with a confidence value ξ_1, ξ_2, ξ_3. To update the embedding of i at layer k, the local transition function f first aggregates the (k-1)th-layer's embeddings of i's neighbors, i.e., $\mathbf{h}_{nb[i]}^{k-1} = (1/\sum_{j=1}^{3}\xi_j)\sum_{j=1}^{3}\xi_j\mathbf{h}_{v_j}^{k-1}$. Next, it concatenates the embedding of i at layer $k-1$ with $\mathbf{h}_{nb[i]}^{k-1}$, i.e., $\mathbf{h}_i^{k-1} \oplus \mathbf{h}_{nb[i]}^{k-1}$. Finally, we used a ReLU function as the activation function that takes as input the concatenated vector and generates the kth-layer's embedding of i, i.e., $\mathbf{h}_i^k = \text{ReLU}(\mathbf{W}^k \cdot (\mathbf{h}_i^{k-1} \oplus \mathbf{h}_{nb[i]}^{k-1}) + \mathbf{b}^k)$. **(b)** For a K-layer GNN, the embedding generation process is repeated $K-1$ times.

Then, we combined the embedding $\mathbf{h}_v^{k-1}$ of v and the aggregated embedding $\mathbf{h}_{nb[v]}^{k-1}$, using the concatenation operation $\oplus$ proposed in GraphSAGE [109], which is to combine two embedding vectors side by side,

$$\text{CONCATE}(\mathbf{h}_v^{k-1}, \mathbf{h}_{nb[v]}^{k-1}) = \mathbf{h}_v^{k-1} \oplus \mathbf{h}_{nb[v]}^{k-1}. \tag{4.3}$$

After the training of the local transition function f and receiving the embeddings of all nodes in the HMIN, the global output function g was used to predict the

Table 4.1: Terms and notations used in PANDA algorithm

Notation	**Definition**
$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	Human Molecular Interaction Network
$\lvert\mathcal{V}\rvert$	Number of nodes
$\lvert\mathcal{E}\rvert$	Number of edges
$nb[v]$	The set of v's neighbors
$\mathbf{x}_v \in \mathbb{R}^m$	Properties of v in m-dimensional space
$y_v \in \{1, -1\}$	Desired label for v
$\hat{y}_v \in \{1, -1\}$	Predicted label for v
$\hat{p}_v \in [0, 1]$	Predicted autism association probability for v
K	Number of layers in PANDA algorithm
$\theta = (\mathbf{W}, \mathbf{b})$	Hyperparameters

class of each node, i.e., the association with autism. We used the *sigmoid* function that transformed the prediction result to a probability. We let the nominal variable $y_v \in \{1, -1\}$ be the desired label for the node v, where $y_v = 1$ indicates that the node v is associated with autism, otherwise it is not related to autism (i.e. $y_v = -1$), and let $\hat{p}_v$ be the predicted probability of v being associated with autism. We then projected these probabilities to one of the two labels.

4.2.3 Loss function and optimization

Before we introduce the loss function, Table 4.1 gives the notation used in the PANDA algorithm.

Since only a subset of genes have labels while the others do not, we designed a

two-term loss function in PANDA. The first term of the loss function is a binary classification cross-entropy loss function for the supervised learning using labeled nodes [50]. Cross-entropy loss increases as the predicted probability diverges from the actual label, defined as

$$\mathcal{L}_{\text{su}} = \sum_{i=1}^{n_{\text{su}}} -y_i \log \hat{p}_i - (1 - y_i) \log(1 - \hat{p}_i), \tag{4.4}$$

where n_{su} is the number of labeled nodes, y_i is the desired label, and $\hat{p}_i$ is the predicted probability.

For the second term of unsupervised learning using unlabeled nodes, a similarity measurement was employed. Recall that the node embeddings were generated under the principle that neighboring nodes should have similar embedding vectors. The similarity between two embedding vectors can be naturally quantified using Euclidean distance. The second term of the loss function is to penalize the embeddings that encode neighboring nodes very differently, defined as

$$\mathcal{L}_{\text{un}} = \sum_{j=1}^{n_{\text{un}}} \sum_{i \in nb[j]} ||\mathbf{h}_j - \mathbf{h}_i||_2^2, \tag{4.5}$$

where n_{un} is the number of the unlabeled nodes, $\mathbf{h}_j$ and $\mathbf{h}_i$ are the embeddings of node j and its direct neighbors.

The final loss function is thus computed as

$$\mathcal{L} = \mathcal{L}_{\text{su}} + \alpha \mathcal{L}_{\text{un}} + \lambda \mathcal{L}_{\text{reg}}, \tag{4.6}$$

where α balances the influence of unlabeled nodes in the learning, λ is the regularization weight, and $\mathcal{L}_{\text{reg}}$ is an L2-norm regularization term to prevent overfitting, defined as

$$\mathcal{L}_{\text{reg}} = \frac{1}{2} \sum_{k=1}^{K} ||\mathbf{W}^k||_2^2, \tag{4.7}$$

where K is the number of layers of the GNN.

4.2.4 Training and cross-validation

We adopted a five-fold cross-validation scheme to train the classifier in PANDA. First, we randomly split the labeled genes into five partitions. In each of the five iterations, we trained the classifier using four partitions and evaluated the classifier on the remaining testing partition. For an unlabeled node, its final predicted label is computed by averaging the predictions of the five iterations. After obtaining the predictions for all nodes in the HMIN, we ranked the nodes in the descending order of their predicted probabilities. This rank list represents PANDA's prioritization of candidate autism-associated genes. The pseudocode of this learning algorithm in PANDA is shown in Algorithm 1.

4.3 Results

We implemented the GNN training in PANDA using TensorFlow [110] with Adam optimizer [111]. We trained the GNN classifier using the giant connected component of the HMIN as the input network. The number of layers of our GNN was set to 4. We set the dimension of the embeddings to 10. The hyperparameters of α, λ were tuned by grid search on the cross-validation partitions. Considering the model variance, we repeated the five-fold cross-validation 20 epochs. We reported the accuracy from the epoch with the lowest validation error. We evaluated the autism gene ranks predicted by PANDA by validating using data from an independent sequencing study and comparing the prediction performance with three other machine learning

methods.

Algorithm 1: The learning algorithm in PANDA.

Input: HMIN $\mathcal{G} = (\mathcal{V}, \mathcal{E})$; node properties and labels$(\mathbf{X}, \mathbf{Y})$; confidence vector ξ; number of PANDA layers K; model parameters $\theta = (\mathbf{W}, \mathbf{b})$

Output: Predicted probabilities $\hat{\mathbf{P}}$, node label predictions $\hat{\mathbf{Y}}$, and trained parameters θ_{updated}

$\mathbf{h}_v^0 \leftarrow \mathbf{x}_v, \forall v \in \mathcal{V}$

repeat

  for $k = 1, .., K-1$ **do**

    for $v \in \mathcal{V}$ **do**

      $\mathbf{h}_{nb[v]}^{k-1} \leftarrow \frac{1}{\sum_u \xi_u} \sum_u \xi_u \mathbf{h}_u^{k-1} + \frac{1}{B} \sum_{u'} \mathbf{h}_{u'}^{k-1}$

      $\mathbf{h}_v^k \leftarrow \text{ReLU}(\mathbf{W}^k \cdot (\mathbf{h}_v^{k-1} \oplus \mathbf{h}_{nb[v]}^{k-1}) + \mathbf{b}^k)$

    end

    $\mathbf{h}_v^k \leftarrow \mathbf{h}_v^k / ||\mathbf{h}_v^k||, \forall v \in \mathcal{V}$

  end

  $\hat{y}_v, \hat{p}_v = \sigma(\mathbf{W}^K \cdot \mathbf{h}_v^{K-1} + \mathbf{b}^K)$

  $\mathcal{L}_{\text{hybrid}} = \mathcal{L}_{\text{su}} + \alpha \mathcal{L}_{\text{un}} + \lambda \mathcal{L}_{\text{reg}}$

  use stochastic gradient descent to update the parameters θ

until *converge*;

return $\hat{\mathbf{P}} = \{\hat{p}_v, \forall v \in \mathcal{V}\}, \hat{\mathbf{Y}} = \{\hat{y}_v, \forall v \in \mathcal{V}\}, \theta_{updated}$

4.3.1 Classification performance of PANDA

We computed several performance measurements to evaluate the classifier in PANDA. The GNN classifier has a sensitivity of 0.95, a specificity of 0.86, and a classification

accuracy of 0.89. We further looked at *Precision@k*, defined as the proportion of genes in the top-k ranking list by PANDA that are known associated with autism (positive genes), and computed as

$$precision@k = \frac{|\Delta_{\text{top-k}} \cap \Delta_{\text{asd}}|}{k}. \tag{4.8}$$

Here, $\Delta_{\text{top-k}}$ and Δ_{asd} are the set of top k genes from the predicted ranking list and the set of known autism-associated genes, respectively. $\cap$ is the operation of set intersection. Figure 4.6 shows that 24.7% of the top-2000 ranked genes are known autism-associated genes, i.e., 494 ($= 2000 \times 24.7\%$) out of a total 760 autism-associated genes have been successfully identified by PANDA in the first decile of its ranking list.

4.3.2 Model sensitivity and specificity

Apart from testing the computational results *in silico*, we used an independent sequencing study [29] to validate the autism genes prioritized by our method. This whole exome-sequencing study examined 2,508 autism probands (autism affected children), 1,911 unaffected siblings, and their parents in the Simons Simplex Collection (SSC) [112]. This study focused on identified *de novo* likely gene-disrupting (DN-LGD) mutations. It reported 353 target genes identified in autism probands with 27 recurrent genes, and 176 DN-LGD genes in unaffected siblings.

We looked at the genes in each decile of the rank list R produced by PANDA and compared them with the three gene sets, i.e., DN-LGD in probands, recurrent DN-LGD in probands, and DN-LGD in unaffected siblings, reported in the independent validation study. For each comparison, we applied a one-tail binomial test [113] in order to assess the significance of the overlap. Figure 4.7 shows the overlap of these

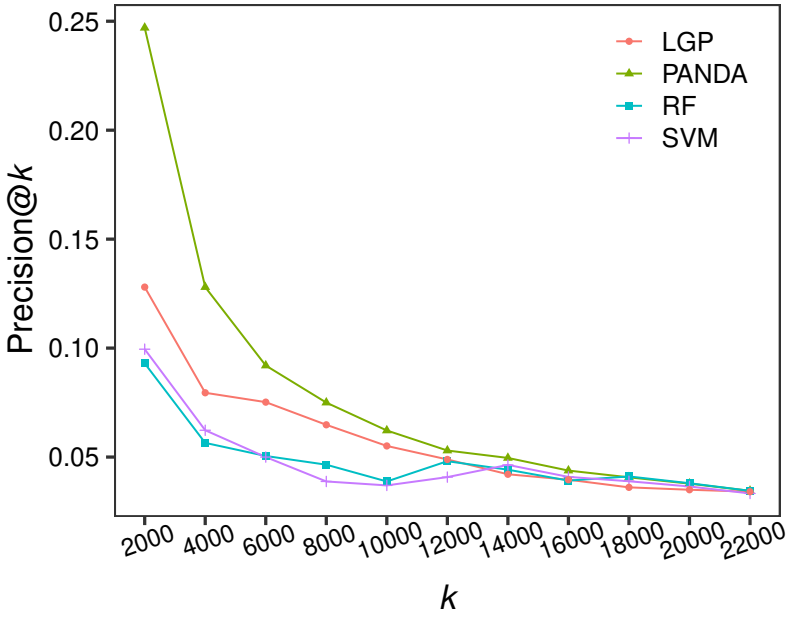

Figure 4.6: *Precision@k* chart. *Precision@k* is the proportion of genes in the top-k ranking list that are known associated with autism. We compared our PANDA with support vector machine (SVM), random forest (RF), and linear genetic programming (LGP) using the measure of *Precision@k*.

independently discovered autism-genes in our PANDA rank R. DN-LGD genes were found enriched in our top 10% of the ranking list (103/353, $p = 2.467 \times 10^{-5}$). Moreover, the recurrent DN-LDG genes validated in probands were enriched in the first decile (15/27, $p = 4.786 \times 10^{-5}$). No significant enrichment of the DN-LGD genes in unaffected siblings was found by PANDA (26/176, $p = 0.3744$).

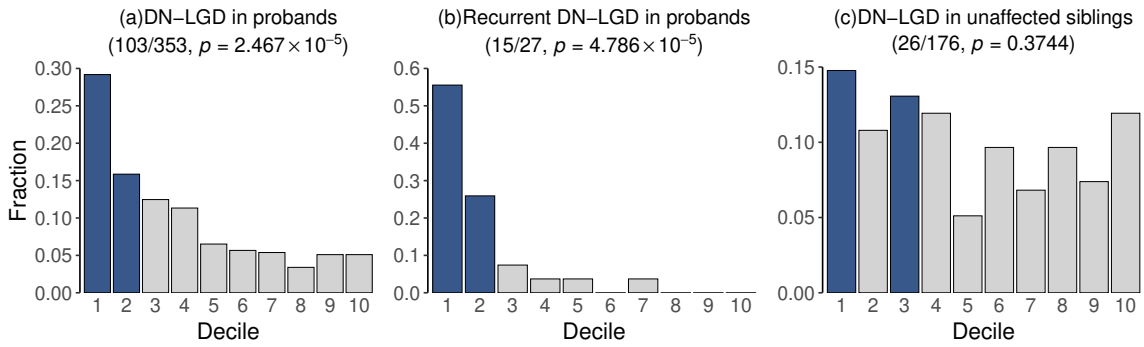

Figure 4.7: Distribution of the PANDA ranks for autism genes identified in the independent exome-sequencing study. (a) Genes with *de novo* likely gene-disrupting (DN-LGD) mutations in probands (autism affected children). (b) Genes with recurrent DN-LGD in probands. (c) DN-LGD genes identified in unaffected siblings. The first and second highest deciles are highlighted in blue.

Furthermore, we evaluated the specificity of the gene rankings by PANDA, which is the ability to downweight non-autism genes or to ensure we were not simply observing an enrichment for genes involved in brain function. It specifically can be used to test if our enriched genes were associated with general neurological disorders, rather than specifically with autism. Figure 4.8 shows the distributions of the PANDA ranks for genes associated with Alzheimer's disease, Parkinson's disease, and epilepsy. None of the three neurological disease genes were significantly enriched in the top decile of the PANDA ranking list. The results suggested that our method was able to identify autism genes rather than genes associated with a broader range of neurological disorders.

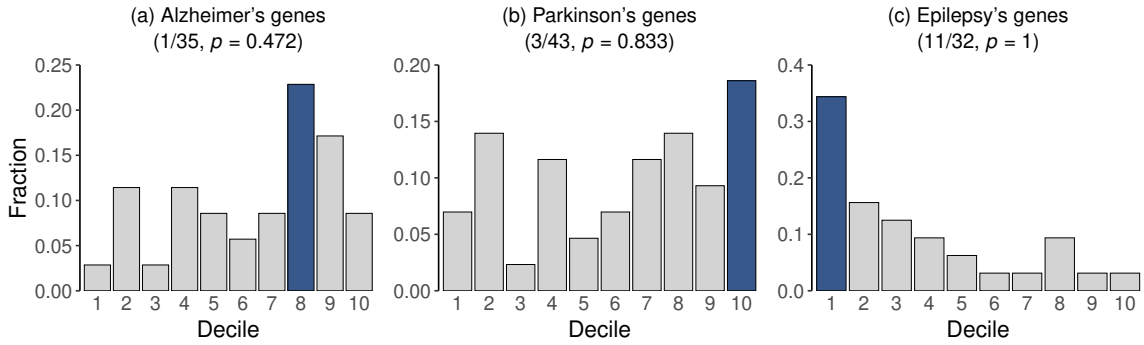

Figure 4.8: Specificity results of PANDA gene ranking list. (a) The distribution of PANDA ranks of genes associated with Alzheimer's disease (AD). 1 out of the total 35 AD genes was in the top decile of PANDA ranking and its binomial test p-value is 0.472. (b)The distribution of PANDA ranks of Parkinson's disease (PD) genes. 3 out of 43 PD genes were ranked in the top decile ($p = 0.833$). (c) The distribution of PANDA ranks of epilepsy genes. 11 out of 32 genes were in the top decile ($p = 1$).

4.3.3 Methods comparison

To evaluate the classification performance of PANDA, we compared the results of PANDA with using three other machine learning methods, including support vector machine (SVM) [114], random forest (RF) [115], and linear genetic programming (LGP) [116]. The parameters of the compared methods are selected by grid search with cross-validation [117]. We used the Radial Basis Function kernel for SVM. The number of estimators in RF was set to 100. We set the population size, number of generations and number of runs in LGP to 1000, 1000 and 100, respectively, and the crossover and mutation rates to 0.9 and 0.05. To ensure a fair comparison, all methods used the same five-fold cross-validation partitions. Figure 4.6 shows the

comparison result using precision@k. PANDA achieved the best ranking quality of the top-2000 genes among all four methods. Figure 4.9 shows the comparison result using the measure of prediction accuracy. PANDA achieved the best classification accuracy among all four methods.

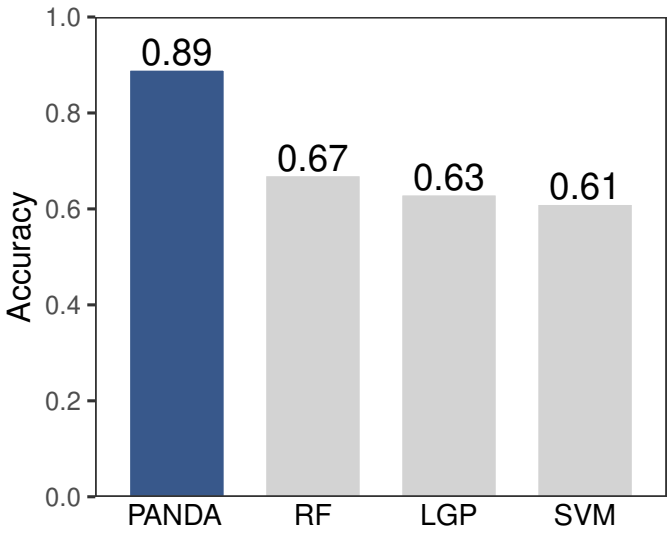

Figure 4.9: Comparison of the classification accuracy using PANDA, support vector machine (SVM), random forests (RF), and linear genetic programming (LGP).

4.4 Biological interpretation

Prioritizing the candidate autism-associated genes from the HMIN is just one step toward deciphering the etiology of the disease. As our goal is to facilitate downstream biological validation tests on potential autism-associated genes, the next step after ranking the genetic variants via PANDA classifier is to explore the top-ranked genes. To better understand the genetic attributes of the top 10% ranked autism-associated

candidate genes(˜2300 genes), we conducted experiments from two aspects. First, we extracted the disease module from the HMIN using the top 10% ranked genes. Then, we utilized DAVID database [118] to analyze the biological characteristics of the top ranked genes.

4.4.1 Characterization of prioritized autism-associated genes

We took a closer look at the top 10% genes in the PANDA ranking list and investigated their properties as nodes in the HMIN. Inspired by the proposed concept of disease module in network biology [19], we defined the *autism module* as an induced subgraph of the HMIN that includes the top decile genes predicted by PANDA with regard to their autism association. The autism module has 2,346 nodes and 11,668 edges.

Following the intuition that functionally related genes tend to share many connections and develop local "neighborhoods" within a larger network, we applied a fast-greedy community detection algorithm [119] to the autism module in order to identify the functional neighborhoods. Figure 4.10 shows the communities identified in the autism module. The modularity rate of the autism module is 0.385, indicating a significant occurrence of community structures, i.e., nodes have significantly more connections with others within a community than with nodes from different communities. Figure 4.11 shows the pairwise correlations among the six network metrics in the autism module.

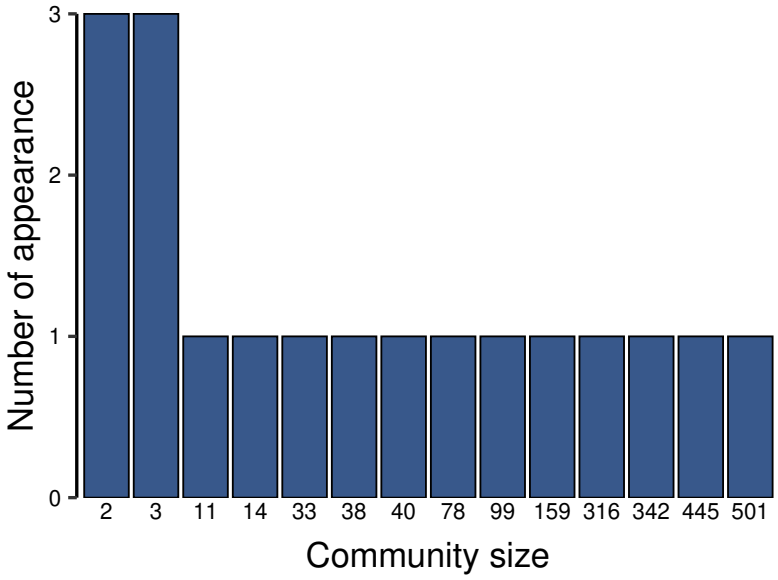

Figure 4.10: The histogram of the community sizes in the autism module. For the consideration of clarity, the figure only displays communities with at least 2 nodes. There are 255 communities that only contain a single node, which is not shown in the figure.

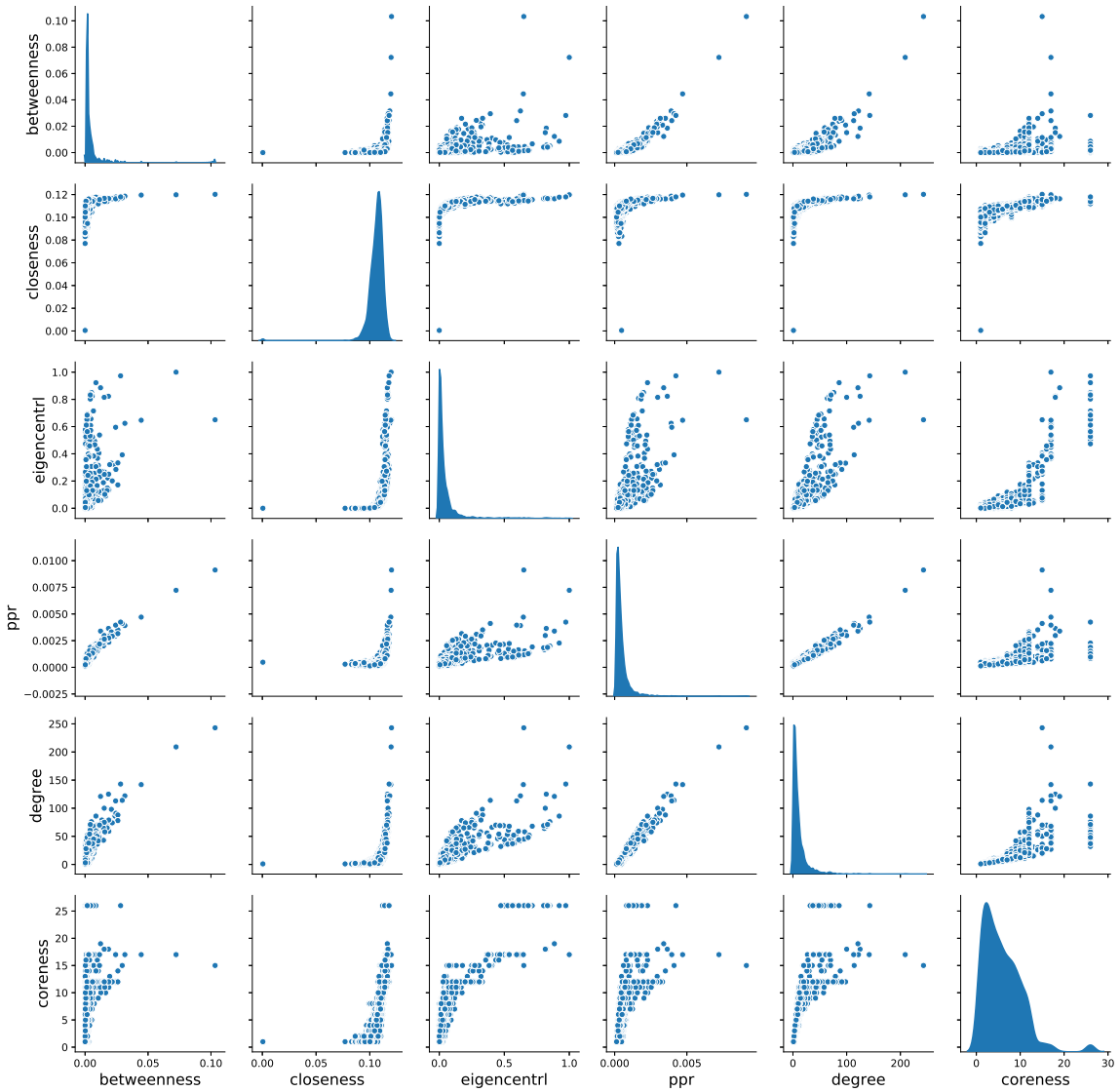

Figure 4.11: The pairwise correlations of six network metrics in the autism module, including betweenness centrality, closeness centrality, eigenvector centrality(eigencontrl), personalized pagerank(ppr), degree centrality, k-coreness.

4.4.2 Enrichment analysis

The **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (DAVID) provides a comprehensive set of functional annotation tools to understand biological meaning behind a given list of genes [118]. We first submitted the first 10% genes to DAVID database and extracted the results under the categories of Gene Ontology (GO), Disease, Pathways. Table 4.2 presents the most significantly enriched functional categories.

We analyzed the functional enrichment tests of the first 10 communities detected in the autism module and applied multiple testing correction to adjust the resulting p-values. For each community, the threshold of gene count is set to 15, indicating that only functional categories containing more than 15 of the total submitted genes were included. The significance cutoff as 0.05, i.e. Fisher's exact test $p < 0.05$. Table 4.3 shows most significantly annotated diseases enriched for the top 10 communities. We found that tobacco use disorder had the highest appearance in 8 of the 10 communities under the *GAD DISEASE* category, with Schizophrenia for community 4 and Neuroblastoma for community 7.

We then looked into the enriched KEGG pathways in the 10 communities. For the pathway enrichment tests, the threshold of gene count is set to 5, indicating that only functional categories containing more than 5 of the total submitted genes were included. The significance cutoff as 0.05. Table 4.4 shows the most significantly enriched pathways for each community. The results suggest several pathways, which have been examined in order to determine the involvement in autism.

For example, MAPK signaling pathway is enriched in community 1. The MAPK

pathway is a prominent intracellular signaling pathway regulating various intracellular functions [120]. Studies have revealed that components of this pathway are mutated in a related collection of disorders that are associated with autism [121]. Another example is cAMP signaling pathway in community 5. cAMP (a.k.a. Cyclic AMP) is a second messenger involved in many processes including mnemonic processing and anxiety [122]. It has previously been investigated with regard to its role in autism [123].

DAVID enrichment tests have yielded insights into the top-ranked genes. First, the autism module derived using these genes has organized into several functionally distinct communities. These communities have relevant associations with autism. Community 1 was enriched for several signaling pathways with previous evidence of association with autism, such as PI3K signaling pathway, HTLV-1 infection, and MAPK signaling pathway. Communities 2 and 4 reflected multiple biological processes, including chromatin remodeling, histone modification, and transcriptional regulation that have prior implication in autism. Furthermore, the detected communities captured a number of developmental processes. Community 5 showed embryonic development. Neuron fate commitment and nervous system development were enriched in community 6. Enrichment of axonal and dendritic development and morphogenesis was contained in community 3.

4.5 Summary

In this section, we first introduced the concepts and challenges of network embedding, from which we introduced graph neural network. Section 4.1.2 described in detail the

fundamental ideas of graph neural network. Based on the original proposal of GNN, Section 4.2.1 presented the overview of PANDA framework. Section 4.2.2 described the node classification procedure of PANDA. Finally, we assessed our method from three aspects, including the model performance, the validation on an independent study, and biological enrichment test. The results showed that PANDA was able to achieve high classification accuracy, and that its output gene ranking was successfully validated using an independent sequencing dataset.

Table 4.2: Enriched functional annotation terms of the ten communities. The options of gene count and significant cutoff are set to 15 and 0.05, respectively.

Category	**Term**	**Count**	p**-value**
GOTERM_CC_DIRECT	Nucleus	233	1.6×10^{-27}
GOTERM_CC_DIRECT	Nucleoplasm	155	1.9×10^{-27}
GOTERM_BP_DIRECT	Negative regulation of transcription from RNA polymerase II promoter	71	7.4×10^{-24}
GOTERM_MF_DIRECT	Zinc ion binding	92	1.2×10^{-23}
GOTERM_BP_DIRECT	Transcription, DNA-templated	123	1.9×10^{-23}
GOTERM_BP_DIRECT	Positive regulation of transcription from RNA polymerase II promoter	81	3.4×10^{-22}
GOTERM_BP_DIRECT	Negative regulation of transcription, DNA-templated	52	1.2×10^{-18}
GOTERM_MF_DIRECT	Protein binding	304	1.7×10^{-17}
GOTERM_CC_DIRECT	Cytoplasm	198	3.2×10^{-15}
GOTERM_MF_DIRECT	Sequence-specific DNA binding	65	1.9×10^{-13}
GAD_DISEASE_CLASS	Chemdependency	150	8.1×10^{-10}
GOTERM_MF_DIRECT	DNA binding	82	1.9×10^{-9}
GOTERM_BP_DIRECT	Regulation of DNA-templated transcription	73	2.2×10^{-8}
GOTERM_MF_DIRECT	ATP binding	65	7.7×10^{-6}
GAD_DISEASE_CLASS	Psych	75	8.9×10^{-4}
GAD_DISEASE_CLASS	Developmental	56	6.5×10^{-3}

Table 4.3: Enriched disease category annotations of the ten communities. The options of gene count and significant cutoff are set to 15 and 0.05, respectively. The category is selected as GAD_DISEASE.

Community	**Disease Term**	**Gene Count**	p**-value**
Community 1	Tobacco Use Disorder	134	2.0×10^{-10}
Community 2	Tobacco Use Disorder	91	6.5×10^{-7}
Community 3	Tobacco Use Disorder	214	5.4×10^{-34}
Community 4	Schizophrenia	18	1.6×10^{-4}
Community 5	Tobacco Use Disorder	52	8.0×10^{-6}
Community 6	Tobacco Use Disorder	33	5.1×10^{-4}
Community 7	Neuroblastoma	16	6×10^{-3}
Community 8	Tobacco Use Disorder	76	2.1×10^{-3}
Community 9	Tobacco Use Disorder	17	2.2×10^{-2}
Community 10	Tobacco Use Disorder	19	2.5×10^{-2}

Table 4.4: Enriched pathways category annotations of the ten communities. The options of gene count and significant cutoff are set to 5 and 0.05, respectively. The category is selected as KEGG_PATHWAYS.

Community	**Pathway Term**	**Gene Count**	p**-value**
Community 1	MAPK signaling pathway	16	1.5×10^{-4}
Community 2	Axon guidance	10	4.5×10^{-5}
Community 3	Glutamatergic synapse	30	4.9×10^{-18}
Community 4	Purine metabolism	15	4.5×10^{-14}
Community 5	cAMP signaling pathway	8	1.7×10^{-6}
Community 6	Serotonergic synapse	6	8.7×10^{-5}
Community 7	Lysine degradation	7	6.7×10^{-3}
Community 8	mRNA surveillance pathway	18	6.5×10^{-15}
Community 9	RNA degradation	14	1.7×10^{-3}
Community 10	N-Glycan biosynthesis	6	5.6×10^{-3}

Chapter 5

Discussion

5.1 Contribution summary

Elucidating the genetic etiology of complex diseases is one of the greatest challenges in modern biomedical research. Disease-gene association study is a crucial step in understanding disease etiology. Deciphering the link between genes and diseases is an open problem in biomedical sciences, but it presents an opportunity to better understand disease etiology, thereby allowing for the design and development of better mitigation strategies. As a result, various *in silico* methods for predicting associations from large-volume biological data have been developed using different approaches.

Computational methods have seen increasing applications in biomedicine, thanks to their powerful abilities to analyze large-volume, high-dimensional data. Prioritizing disease association genes is a research problem that can benefit from using carefully designed and tailored computational methods. Many common and complex diseases have observable high inheritance, however, our understanding of their genetic

architecture is still limited. Complex network analysis and deep learning provide a powerful toolset that can be used to characterize the inter-connective relationships of genes in the human genome and to identify potential genes associated with diseases based on such interaction patterns.

In this thesis, we proposed PANDA, a graph deep learning approach to prioritizing autism genes across the entire human genome. In PANDA, we first constructed the human molecular interaction network (HMIN) to characterize the patterns and structure of gene-gene interactions in the human genome. We then designed a graph neural network that was able to input graph structured data, to learn the similarity of nodes and their interaction patterns, and to predict potential genes that were similar and related to known autism genes. The results showed that PANDA was able to achieve high classification accuracy, and that its output gene ranking was successfully validated using an independent sequencing dataset.

We contribute to the understanding of autism etiology in the following aspects. Firstly, we developed a deep graph neural network method that can operate directly on a biological network. PANDA utilized the interaction relationships embedded in the HMIN to learn hidden features for every node in the network. In line with our hypothesis that these hidden features encode the connectivity patterns of the autism-associated genes and distinguish from those of non-disease genes, the PANDA classifier achieved high classification performance with 89% accuracy. Secondly, PANDA not only trained a predictive classifier but also generated a prioritization ranking of genes across the genome. We tested the prioritization quality using an independent sequencing study, and the results suggest that our top 10% (2346) genes were significantly enriched for the *de novo* likely gene-disrupting mutations that are identified

in autism-affected children, while no significant enrichment was shown in the autism unaffected siblings. The result elucidates the capability of the PANDA to nominate candidate autism-associated genes, which can facilitate downstream biological researches.

In summary, our study showcased the potential of designing advancing network analysis and machine learning methods for the prioritization of disease-associated genes. Apart from many existing studies where new genes were often predicted with disease association based on their direct relationship with known disease genes, we hypothesized that genes involved in the development of a disease may not directly interact with each other, but may exhibit similar topological properties in the HMIN. Using the graph deep learning approach enabled searching for these potential genes that were multi-hops away from known autism-associated genes. We hope this study opens future research avenues of employing more advanced modeling and learning algorithms in order to better characterize the genetic architecture of complex diseases.

5.2 Future work

However, there are limitations in our study. First, we consider that the HMIN are invariant for autism, meaning that the mutations on certain genes associated with autism do not introduce a new topology in the HMIN. Second, although Alzheimer and Parkinson diseases often occur late than autism, we ignored the onset and development timepoints of autism and the three neurological disorders.

In future studies, we expect to explore more generic models on prioritizing genes associated with other diseases or disease subtypes. For the top-ranked candidate

autism-associated genes, further explorations, such as genome-wide association studies, can be conducted to gain a deeper understanding of the associations with autism. In addition, a comparative analysis can be conducted by including more machine learning methods, particularly other graph learning-targeted algorithms in order to obtain robust and reasonable conclusions. Moreover, we can develop an application with a graphical user interface for other biologists to adopt the methods exploited in this research.

On the other hand, genes carrying mutations associated with genetic diseases are present in all human cells [124]. Clinical manifestations of complex diseases, however, are usually highly tissue-specific [125]. Genes with tissue-specific expressions have shown to be involved in important physiological processes for complex organisms [126]. Although some disease-causing genes are expressed only in selected tissues, the expression patterns of disease-causing genes alone cannot explain the observed tissue specificity of human diseases. In the future, we expect to explore the application of PANDA over tissue-specific biological networks, which are built using tissue specificity expression data.

Bibliography

[1] Ebony B Bookman, Kimberly McAllister, Elizabeth Gillanders, Kay Wanke, David Balshaw, Joni Rutter, et al. Gene-environment interplay in common complex diseases: forging an integrative model–recommendations from an nih workshop. *Genetic epidemiology*, 35(4):217–225, 2011.

[2] Neil Risch and Kathleen Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, 1996.

[3] Tracy Murray-Stewart, Yanlin Wang, Andrew Goodwin, Amy Hacker, Alan Meeker, and Robert A Casero Jr. Nuclear localization of human spermine oxidase isoforms–possible implications in drug response and disease etiology. *The FEBS journal*, 275(11):2795–2806, 2008.

[4] David JA Goldsmith and Adrian Covic. Coronary artery disease in uremia: Etiology, diagnosis, and therapy. *Kidney international*, 60(6):2059–2078, 2001.

[5] Roy C Page. The etiology and pathogenesis of periodontitis. *Compendium of continuing education in dentistry*, 23(5 Suppl):11–14, 2002.

[6] Kai Sun, Joana P Gonçalves, Chris Larminie, and Nataša Pržulj. Predicting disease associations via biological network analysis. *BMC bioinformatics*, 15(1):304, 2014.

[7] Joseph Loscalzo, Isaac Kohane, and Albert-László Barabási. Human disease classification in the postgenomic era: A complex systems approach to human pathobiology. *Molecular Systems Biology*, 3(1):124, 2007.

[8] Alan Wright and Nicholas Hastie. *Genes and common diseases: Genetics in modern medicine*. Cambridge University Press, 2007.

[9] Anne M Glazier, Joseph H Nadeau, and Timothy J Aitman. Finding genes that underlie complex traits. *Science*, 298(5602):2345–2349, 2002.

[10] Anthony JF Griffiths, Jeffrey H Miller, David T Suzuki, Richard C Lewontin, et al. *An introduction to genetic analysis*. WH Freeman and Company, 2000.

[11] M Dawn Teare and Jennifer H Barrett. Genetic linkage studies. *The Lancet*, 366(9490):1036–1044, 2005.

[12] Joel N Hirschhorn and Mark J Daly. Genome-wide association studies for common diseases and complex traits. *Nature reviews genetics*, 6(2):95, 2005.

[13] Andrew G Fraser, Ravi S Kamath, Peder Zipperlen, Maruxa Martinez-Campos, Marc Sohrmann, and Julie Ahringer. Functional genomic analysis of C. elegans chromosome I by systematic RNA interference. *Nature*, 408(6810):325, 2000.

[14] David B Goldstein, Kourosh R Ahmadi, Mike E Weale, and Nicholas W Wood. Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. *Trends in genetics*, 19(11):615–622, 2003.

[15] Jennifer M Kwon and Alison M Goate. The candidate gene approach. *Alcohol research and health*, 24(3):164–168, 2000.

[16] Mengjin Zhu and Shuhong Zhao. Candidate gene identification approach: progress and challenges. *International journal of biological sciences*, 3(7):420–427, 2007.

[17] Holly K Tabor, Neil J Risch, and Richard M Myers. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature reviews genetics*, 3(5):391–397, 2002.

[18] Barbara Di Ventura, Caroline Lemerle, Konstantinos Michalodimitrakis, and Luis Serrano. From *in vivo* to *in silico* biology and back. *Nature*, 443(7111):527–533, 2006.

[19] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature reviews genetics*, 12(1):56–68, 2011.

[20] Ting Hu, Nicholas A. Sinnott-Armstrong, Jeff W. Kiralis, Angeline S. Andrew, Margaret R. Karagas, and Jason H. Moore. Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinformatics*, 12(1):364, 2011.

[21] Ting Hu, Yuanzhu Chen, Jeff W Kiralis, Ryan L Collins, Christian Wejse, Giorgio Sirugo, Scott M Williams, and Jason H Moore. An information-gain approach to detecting three-way epistatic interactions in genetic association studies. *Journal of the American Medical Informatics Association*, 20(4):630–636, 2013.

[22] Ting Hu, Marco Tomassini, and Wolfgang Banzhaf. Complex network analysis of a genetic programming phenotype network. In *European Conference on Genetic Programming*, pages 49–63. Springer, 2019.

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

[24] Catherine Lord, Edwin H Cook, Bennett L Leventhal, and David G Amaral. Autism spectrum disorders. *Neuron*, 28(2):355–363, 2000.

[25] Joseph T Glessner, Kai Wang, Guiqing Cai, Olena Korvatska, Cecilia E Kim, Shawn Wood, et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature*, 459(7246):569–573, 2009.

[26] Eric Fombonne. The epidemiology of autism: a review. *Psychological medicine*, 29(4):769–786, 1999.

[27] Ryan KC Yuen, Bhooma Thiruvahindrapuram, Daniele Merico, Susan Walker, Kristiina Tammimies, Ny Hoang, et al. Whole-genome sequencing of quartet families with autism spectrum disorder. *Nature medicine*, 21(2):185–191, 2015.

[28] Stephan J Sanders, Michael T Murtha, Abha R Gupta, John D Murdoch, Melanie J Raubeson, A Jeremy Willsey, et al. *De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*, 485(7397):237–241, 2012.

[29] Ivan Iossifov, Brian J O'roak, Stephan J Sanders, Michael Ronemus, Niklas Krumm, Dan Levy, et al. The contribution of *de novo* coding mutations to autism spectrum disorder. *Nature*, 515(7526):216, 2014.

[30] Silvia De Rubeis, Xin He, Arthur P Goldberg, Christopher S Poultney, Kaitlin Samocha, A Ercument Cicek, et al. Synaptic, transcriptional and chromatin genes disrupted in autism. *Nature*, 515(7526):209–215, 2014.

[31] Brian J O'Roak, Pelagia Deriziotis, Choli Lee, Laura Vives, Jerrod J Schwartz, Santhosh Girirajan, et al. Exome sequencing in sporadic autism spectrum disorders identifies severe *de novo* mutations. *Nature genetics*, 43(6):585–589, 2011.

[32] Trent Gaugler, Lambertus Klei, Stephan J Sanders, Corneliu A Bodea, Arthur P Goldberg, Ann B Lee, et al. Most genetic risk for autism resides with common variation. *Nature genetics*, 46(8):881–885, 2014.

[33] Michael Ronemus, Ivan Iossifov, Dan Levy, and Michael Wigler. The role of *de novo* mutations in the genetics of autism spectrum disorders. *Nature reviews genetics*, 15(2):133–141, 2014.

[34] Stephan J Sanders, Xin He, A Jeremy Willsey, A Gulhan Ercan-Sencicek, Kaitlin E Samocha, A Ercument Cicek, et al. Insights into autism spectrum

disorder genomic architecture and biology from 71 risk loci. *Neuron*, 87(6):1215–1233, 2015.

[35] Brian J O'Roak, Laura Vives, Santhosh Girirajan, Emre Karakoc, Niklas Krumm, Bradley P Coe, et al. Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. *Nature*, 485(7397):246—250, 2012.

[36] Mark Newman. *Networks: an introduction*. Oxford university press, 2010.

[37] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[38] Tijana Milenković and Nataša Pržulj. Uncovering biological network function via graphlet degree signatures. *Cancer informatics*, 6:257–273, 2008.

[39] Albert-László Barabási and Zoltan N Oltvai. Network biology: understanding the cell's functional organization. *Nature reviews genetics*, 5(2):101–113, 2004.

[40] Ting Hu, Karoliina Oksanen, Weidong Zhang, Ed Randell, Andrew Furey, Guang Sun, and Guangju Zhai. An evolutionary learning and network approach to identifying key metabolites for osteoarthritis. *PLoS computational biology*, 14(3):e1005986, 2018.

[41] Monica Agrawal, Marinka Zitnik, and Jure Leskovec. Large-scale analysis of disease pathways in the human interactome. In *Pacific symposium on biocomputing*, volume 23, pages 111–122. World Scientific, 2018.

[42] Seyed M. Almasi and Ting Hu. Measuring the importance of vertices in the weighted human disease network. *PloS ONE*, 14(3):e0205936, 2019.

[43] Maxwell W Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nature reviews genetics*, 16(6):321–332, 2015.

[44] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. *Briefings in bioinformatics*, 18(5):851–869, 2017.

[45] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

[46] Christof Angermueller, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. Deep learning for computational biology. *Molecular systems biology*, 12(7):878, 2016.

[47] Yongjin Park and Manolis Kellis. Deep learning for regulatory genomics. *Nature biotechnology*, 33(8):825–826, 2015.

[48] Yifei Chen, Yi Li, Rajiv Narayan, Aravind Subramanian, and Xiaohui Xie. Gene expression inference with deep learning. *Bioinformatics*, 32(12):1832–1839, 2016.

[49] Michael KK Leung, Hui Yuan Xiong, Leo J Lee, and Brendan J Frey. Deep learning of the tissue-regulated splicing code. *Bioinformatics*, 30(12):i121–i129, 2014.

[50] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[51] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, and Maosong Sun. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*, 2018.

[52] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[53] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2009.

[54] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[55] Alex Fout, Jonathon Byrd, Basir Shariat, and Asa Ben-Hur. Protein interface prediction using graph convolutional networks. In *Advances in Neural Information Processing Systems 30*, pages 6530–6539. Curran Associates, Inc., 2017.

[56] Alvaro Sanchez-Gonzalez, Nicolas Heess, Jost Tobias Springenberg, Josh Merel, Martin Riedmiller, Raia Hadsell, and Peter Battaglia. Graph networks as learnable physics engines for inference and control. *arXiv preprint arXiv:1806.01242*, 2018.

[57] Takuo Hamaguchi, Hidekazu Oiwa, Masashi Shimbo, and Yuji Matsumoto. Knowledge transfer for out-of-knowledge-base entities: A graph neural network approach. *arXiv preprint arXiv:1706.05674*, 2017.

[58] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research*, 45(D1):D833–D839, 2016.

[59] Francis S Collins, Michael Morgan, and Aristides Patrinos. The human genome project: lessons from large-scale biology. *Science*, 300(5617):286–290, 2003.

[60] Joel N Hirschhorn, Kirk Lohmueller, Edward Byrne, and Kurt Hirschhorn. A comprehensive review of genetic association studies. *Genetics in medicine*, 4(2):45, 2002.

[61] Daniel J Schaid and SS Sommer. Genotype relative risks: methods for design and analysis of candidate-gene association studies. *American journal of human genetics*, 53(5):1114, 1993.

[62] Michael L Metzker. Sequencing technologie–the next generation. *Nature reviews genetics*, 11(1):31, 2010.

[63] Jonathan K Pritchard, Matthew Stephens, Noah A Rosenberg, and Peter Donnelly. Association mapping in structured populations. *The American Journal of Human Genetics*, 67(1):170–181, 2000.

[64] Sebastian Okser, Tapio Pahikkala, and Tero Aittokallio. Genetic variants and their interactions in disease risk prediction–machine learning and network perspectives. *BioData mining*, 6(1):5, 2013.

[65] Xiujuan Wang, Natali Gulbahce, and Haiyuan Yu. Network-based methods for human disease gene prediction. *Briefings in functional genomics*, 10(5):280–293, 2011.

[66] Rui Jiang, Mingxin Gan, and Peng He. Constructing a gene semantic similarity network for the inference of disease genes. *BMC systems biology*, 5(2):S2, 2011.

[67] AL Yonan, AA Palmer, KC Smith, I Feldman, HK Lee, JM Yonan, SG Fischer, P Pavlidis, and TC Gilliam. Bioinformatic analysis of autism positional candidate genes using biological databases and computational gene network prediction. *Genes, brain and behavior*, 2(5):303–320, 2003.

[68] Roser Corominas, Xinping Yang, Guan Ning Lin, Shuli Kang, Yun Shen, Lila Ghamsari, et al. Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. *Nature communications*, 5:3650, 2014.

[69] 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):e1000662, 2010.

[70] Pedro Larranaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Inaki Inza, et al. Machine learning in bioinformatics. *Briefings in bioinformatics*, 7(1):86–112, 2006.

[71] Enrico Glaab, Jaume Bacardit, Jonathan M Garibaldi, and Natalio Krasnogor. Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. *PloS one*, 7(7):e39932, 2012.

[72] Hong-Woo Chun, Yoshimasa Tsuruoka, Jin-Dong Kim, Rie Shiba, Naoki Nagata, Teruyoshi Hishiki, and Junichi Tsujii. Extraction of gene-disease relations from medline using domain dictionaries and machine learning. In *Biocomputing 2006*, pages 4–15. World Scientific, 2006.

[73] 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, 2012.

[74] Michael PS Brown, William Noble Grundy, David Lin, Nello Cristianini, Charles Walsh Sugnet, Terrence S Furey, Manuel Ares, and David Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1):262–267, 2000.

[75] Yan Kou, Catalina Betancur, Huilei Xu, Joseph D Buxbaum, and Avi Ma'Ayan. Network- and attribute-based classifiers can prioritize genes and pathways for

autism spectrum disorders and intellectual disability. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, 160(2):130–142, 2012.

[76] Marlena Duda, Hongjiu Zhang, Hong-Dong Li, Dennis P Wall, Margit Burmeister, and Yuanfang Guan. Brain-specific functional relationship networks inform autism spectrum disorder gene prediction. *Translational psychiatry*, 8(1):56, 2018.

[77] Arjun Krishnan, Ran Zhang, Victoria Yao, Chandra L Theesfeld, Aaron K Wong, Alicja Tadych, et al. Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. *Nature neuroscience*, 19(11):1454, 2016.

[78] Jaume Bacardit and Xavier Llorà. Large-scale data mining using genetics-based machine learning. *WIREs Data Mining and Knowledge Discovery*, 3(1):37–61, 2013.

[79] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[80] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Chang. Network representation learning with rich text information. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[81] William L Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. *arXiv preprint arXiv:1709.05584*, 2017.

[82] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE signal processing magazine*, 34(4):18–42, 2017.

[83] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE transactions on pattern analysis and machine intelligence*, 29(1):40–51, 2007.

[84] Peng Cui, Xiao Wang, Jian Pei, and Wenwu Zhu. A survey on network embedding. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[85] Judith H. Miles. Autism spectrum disorders–a genetics review. *Genetics in medicine*, 13:278–294, 2011.

[86] Rachel Loomes, Laura Hull, and William Polmear Locke Mandy. What is the male-to-female ratio in autism spectrum disorder? a systematic review and meta-analysis. *Journal of the American Academy of Child & Adolescent Psychiatry*, 56(6):466–474, 2017.

[87] Dan E Arking, David J Cutler, Camille W Brune, Tanya M Teslovich, Kristen West, Morna Ikeda, et al. A common genetic variant in the neurexin superfamily member CNTNAP2 increases familial risk of autism. *The American Journal of Human Genetics*, 82(1):160–164, 2008.

[88] Eitan E Winter, Leo Goodstadt, and Chris P Ponting. Elevated rates of protein secretion, evolution, and disease among tissue-specific genes. *Genome research*, 14(1):54–61, 2004.

[89] Elise B Robinson, Beate St Pourcain, Verneri Anttila, Jack A Kosmicki, Brendan Bulik-Sullivan, Jakob Grove, et al. Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. *Nature genetics*, 48(5):552, 2016.

[90] Silvia De Rubeis and Joseph D Buxbaum. Genetics and genomics of autism spectrum disorder: embracing complexity. *Human molecular genetics*, 24(R1):R24–R31, 2015.

[91] Jonathan Sebat, B Lakshmi, Dheeraj Malhotra, Jennifer Troge, Christa Lese-Martin, Tom Walsh, et al. Strong association of *de novo* copy number mutations with autism. *Science*, 316(5823):445–449, 2007.

[92] Arlene Mannion and Geraldine Leader. Comorbidity in autism spectrum disorder: A literature review. *Research in Autism Spectrum Disorders*, 7(12):1595–1616, 2013.

[93] EM Kenny, P Cormican, S Furlong, E Heron, G Kenny, C Fahey, E Kelleher, S Ennis, D Tropea, R Anney, et al. Excess of rare novel loss-of-function variants in synaptic genes in schizophrenia and autism spectrum disorders. *Molecular psychiatry*, 19(8):872, 2014.

[94] Brett S Abrahams, Dan E Arking, Daniel B Campbell, Heather C Mefford, Eric M Morrow, Lauren A Weiss, et al. Sfari gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). *Molecular autism*, 4(1):36, 2013.

[95] Saumyendra N Basu, Ravi Kollu, and Sharmila Banerjee-Basu. Autdb: a gene reference resource for autism research. *Nucleic acids research*, 37:D832–D836, 2008.

[96] Ada Hamosh, Alan F Scott, Joanna S Amberger, Carol A Bocchini, and Victor A McKusick. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(suppl_1):514–517, 2005.

[97] Vishal Saxena, Shweta Ramdas, Courtney Rothrock Ochoa, David Wallace, Pradeep Bhide, and Isaac Kohane. Structural, genetic, and functional signatures of disordered neuro-immunological development in autism spectrum disorder. *PloS one*, 7(12):e48835, 2012.

[98] Fredrik Barrenas, Sreenivas Chavali, Petter Holme, Reza Mobini, and Mikael Benson. Network properties of complex human disease genes identified through genome-wide association studies. *PloS one*, 4(11):e8090, 2009.

[99] 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):1257601, 2015.

[100] Rose Oughtred, Chris Stark, Bobby-Joe Breitkreutz, Jennifer Rust, Lorrie Boucher, Christie Chang, et al. The biogrid interaction database: 2019 update. *Nucleic acids research*, 47(D1):D529–D541, 2018.

[101] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[102] Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.

[103] Luis A Nunes Amaral, Antonio Scala, Marc Barthelemy, and H Eugene Stanley. Classes of small-world networks. *Proceedings of the national academy of sciences*, 97(21):11149–11152, 2000.

[104] Daixin Wang, Peng Cui, and Wenwu Zhu. Structural deep network embedding. In *ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1225–1234. ACM, 2016.

[105] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[106] Alex Graves, Marcus Liwicki, Santiago Fernández, Roman Bertolami, Horst Bunke, and Jürgen Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):855–868, 2008.

[107] David F Gleich. Pagerank beyond the web. *SIAM Review*, 57(3):321–363, 2015.

[108] Vladimir Batagelj and Matjaz Zaversnik. An o(m) algorithm for cores decomposition of networks. *arXiv preprint cs/0310049*, 2003.

[109] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems 30*, pages 1024–1034. Curran Associates, Inc., 2017.

[110] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[111] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[112] Gerald D Fischbach and Catherine Lord. The simons simplex collection: a resource for identification of autism genetic risk factors. *Neuron*, 68(2):192–195, 2010.

[113] Ulrich Kaempf. The binomial test: A simple tool to identify process problems. *IEEE transactions on semiconductor manufacturing*, 8(2):160–166, 1995.

[114] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[115] Tin Kam Ho. Random decision forests. In *The international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[116] Markus F Brameier and Wolfgang Banzhaf. *Linear genetic programming*. Springer Science & Business Media, 2007.

[117] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(2):281–305, 2012.

[118] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols*, 4(1):44—57, 2009.

[119] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):66–111, 2004.

[120] Joseph Vithayathil, Joanna Pucilowska, and Gary E Landreth. ERK/MAPK signaling and autism spectrum disorders. *Progress in brain research*, 241:63–112, 2018.

[121] Joanna Pucilowska, Joseph Vithayathil, Emmanuel J Tavares, Caitlin Kelly, J Colleen Karlo, and Gary E Landreth. The 16p11. 2 deletion mouse model of autism exhibits altered cortical progenitor proliferation and brain cytoarchitecture linked to the ERK MAPK pathway. *Journal of Neuroscience*, 35(7):3190–3200, 2015.

[122] Eric R Kandel. The molecular biology of memory: cAMP, PKA, CRE, CREB-1, CREB-2, and CPEB. *Molecular brain*, 5(1):14, 2012.

[123] Daniel J Kelley, Anita Bhattacharyya, Garet P Lahvis, Jerry CP Yin, Jim Malter, and Richard J Davidson. The cyclic AMP phenotype of fragile X and autism. *Neuroscience & Biobehavioral Reviews*, 32(8):1533–1543, 2008.

[124] Maksim Kitsak, Amitabh Sharma, Jörg Menche, Emre Guney, Susan Dina Ghiassian, Joseph Loscalzo, and Albert-László Barabási. Tissue specificity of human disease module. *Scientific reports*, 6:35241, 2016.

[125] Casey S Greene, Arjun Krishnan, Aaron K Wong, Emanuela Ricciotti, Rene A Zelaya, Daniel S Himmelstein, et al. Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics*, 47(6):569, 2015.

[126] Kasper Lage, Niclas Tue Hansen, E Olof Karlberg, Aron C Eklund, Francisco S Roque, Patricia K Donahoe, Zoltan Szallasi, Thomas Skøt Jensen, and Søren Brunak. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proceedings of the National Academy of Sciences*, 105(52):20870–20875, 2008.