MULTITASK LEARNING OF GENE RISK FOR AUTISM SPECTRUM DISORDER AND INTELLECTUAL DISABILITY

A THESIS SUBMITTED TO

THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR

THE DEGREE OF

MASTER OF SCIENCE

IN

COMPUTER ENGINEERING

By

İlayda Beyreli

October 2020

MULTITASK LEARNING OF GENE RISK FOR AUTISM SPECTRUM DISORDER AND INTELLECTUAL DISABILITY
By İlayda Beyreli
October 2020

We certify that we have read this thesis and that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Abdullah Ercüment Çiçek(Advisor)

Can Alkan

Tunca Doğan

Approved for the Graduate School of Engineering and Science:

Ezhan Karaşan
Director of the Graduate School

ABSTRACT

MULTITASK LEARNING OF GENE RISK FOR AUTISM SPECTRUM DISORDER AND INTELLECTUAL DISABILITY

İlayda Beyreli
M.S. in Computer Engineering
Advisor: Abdullah Ercüment Çiçek
October 2020

Autism Spectrum Disorder (ASD) and Intellectual Disability (ID) are comorbid neurodevelopmental disorders with complex genetic architectures. Despite large-scale sequencing studies only a fraction of the risk genes were identified for both. Here, we present a novel network-based gene risk prioritization algorithm named DeepND that performs cross-disorder analysis to improve prediction power by exploiting the comorbidity of ASD and ID via multitask learning. Our model leverages information from gene co-expression networks that model human brain development using graph convolutional neural networks and learns which spatio-temporal neurodevelopmental windows are important for disorder etiologies. We show that our approach substantially improves the state-of-the-art prediction power. We observe that both disorders are enriched in transcription regulators. Despite tight regulatory links in between ASD risk genes, such is lacking across ASD and ID risk genes or within ID risk genes. Finally, we investigate frequent ASD and ID associated copy number variation regions and confident false findings to suggest several novel susceptibility gene candidates. DeepND can be generalized to analyze any combinations of comorbid disorders.

Keywords: Autism Spectrum Disorder, Intellectual Disability, Comorbidity, Multitask Learning, Graph Convolutional Networks, Deep Learning.

ÖZET

OTİZM SPEKTRUM BOZUKLUĞU VE ZEKA GERİLİĞİ İÇİN ÇOK GÖREVLİ GEN RİSK ÖĞRENİMİ

İlayda Beyreli
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Danışmanı: Abdullah Ercüment Çiçek
Ekim 2020

Otizm Spectrum Bozukluğu (OSB) ve Zihinsel Engel (ZE) karmaşık genetik yapıları olup birbirlerine eşlik eden nörogelişimsel bozukluklardır. Büyük ölçekli gen sıralamaları araştırmalarına rağmen iki bozukluk için de sorumlu risk genleinin ancak küçük bir yüzdesi belirlenebilmiştir. Bu araştırmada OSB ve ZE için eşlik bilgisini de kullanarak çapraz hastalık analizi ile tahmin gücünü arttıran DeepND adında yeni bir çizge tabanlı gen risk önceliklendirme algoritması tasarladık. Model insan beyin gelişimini modelleyen gen eş ifade çizgelerini çizge evrişimsel sinir ağları ile işleyerek iki hastalık için de önemli mekansal-zamansal nörogelişimsel aralıkları öğrenebilir. Sonuçlarımız sunduğumuz mimarinin alandaki son teknolojinin sağladığı tahmin gücünü arttırdığını göstermektedir. İki bozukluk için de transkripsiyon regulatörlerinindeki zeginleştirilmeler gözlemlendi. OSB genleri arasındaki sıkı düzenleyici bağlantılara rağmen, bu bağlantıların OSB ve ZE genleri ile sadece ZE genleri arasında aynı derecede olmadığı belirlendi. Son olarak, OSB ve ZE bağlantılı kopya numara varyasyonlarını ve modelin emin olduğu yanlış bulguları inceleyerek birkaç yeni şüpheli risk geni belirlendi. DeepND mimarisi kolayca birbirlerine eşlik eden birçok hastalık için genellenebilir.

Anahtar sözcükler: Otizm Spektrum Bozukluğu, Zihinsel Engel, Eşlik Eden Hastalık, Çoklu Görev Öğrenme, Çizge Evrişimsel Ağlar, Derin Öğrenme.

Acknowledgement

First and foremost, I would like to express my sincere gratitude to my advisor Asst.Prof. A. Ercüment Çiçek for his guidance, support, and help throughout my master's studies. It was a great honor to work with him. With his guidance, I was able to conduct this study and discover a way to turn my interest into a career path.

I also would like to thank my jury members Can Alkan and Tunca Doğan for reading my thesis and accepting to be on my thesis committee. I also thank Simons Foundation Autism Research Initiative for supporting this research via SFARI 640935 pilot grant awarded to A. Ercüment Çiçek.

I would like to thank Oğuzhan Karakahya for his valuable friendship, support, and efforts throughout this study. His work created the baseline for this thesis and I am grateful for his contributions.

I also thank Mustafa, Emre, Gülşah, and other friends I am yet to mention for their support, encouragement, and wonderful memories. I would also like to thank Alpay for his immense support, understanding and always being by my side through highs and lows.

Last but not least, I would like to thank my family for their support and love. They have presented me opportunities for my education which I cannot repay all along.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Comorbidity refers two or more physiological or psychological conditions co-occuring independently or related to one another. In the context of neuropsychiatric disorders, comorbidity refers to coexisting disorders in the same person. Overwhelming evidence suggests that genetic architectures of neuropsychiatric disorders overlap [4, 5, 6]. For instance, out of the twenty five SFARI Cat I Autism Spectrum Disorder (ASD) risk genes (i.e., highest risk), only five are solely associated with ASD. Intellectual Disability (ID), characterized by impaired mental capabilities, is one of such comorbid disorders with ASD [7]. Both ID and ASD have complex genetic backgrounds with hundreds of risk genes involved and identified by rare *de novo* disruptive mutations observed in whole exome and genome sequencing studies [8, 9, 10, 11, 12, 13, 14]. In , CDC reported that 31% of children with ASD were also diagnosed with ID and 25% were borderline [15]. They also share a large number of risk genes [16]. Despite these similarities, Robinson *et al.* also point to differences in genetic architectures and report that intelligence quotient (IQ) positively correlates with family history of psychiatric disease and negatively correlates with *de novo* disruptive mutations [17]. Yet, the shared functional circuitry behind is mostly unknown.

The current lack of understanding on how comorbid neuropsychiatric disorders relate mostly stems from the incomplete profiling of individual genetic

architectures. Statistical methods have been used to assess gene risk using excess genetic burden from case-control and family studies [18] which are recently extended to work with multiple traits [19]. It is often of interest to use these as prior risk and obtain a posterior gene interaction network-adjusted risk which can also assess risk for genes with no prior signal. Network-based computational gene risk prediction methods come handy for (i) imputing the insufficient statistical signal and providing a genome-wide risk ranking, and (ii) finding out the affected cellular circuitries such as pathways and networks of genes [18, 20, 21, 22, 23, 24, 25, 26, 27, 28]. While these methods have helped unraveling the underlying mechanisms, they have several limitations. First, by design, they are limited to work with a single disorder. In order to compare and contrast comorbid disorders such as ASD and ID using these tools, one approach is to bag the mutational burden observed for each disorder assuming two are the same. However, disorder specific features are lost as a consequence [20]. The more common approach is to perform independent analyses per disorder and intersect the results. Unfortunately, this approach ignores valuable source of information coming from the shared genetic architecture and lose prediction power as per-disorder analyses have less input (i.e., samples, mutation counts) and less statistical power [29, 30, 22]. Second, current network-based gene discovery methods can work with one or two integrated gene interaction networks [24, 31, 22, 32, 27]. This means numerous functional interaction networks (e.g., co-expression, protein interaction etc.) are disregarded which limits and biases the predictions. Gene co-expression networks that model brain development are a promising source of diverse information regarding gene risk, but currently cannot be fully utilized, as the signal coming from different networks cannot be deconvoluted. Usually, investigating such risky neurodevelopmental windows is an independent downstream analysis [3, 32].

In this thesis, we address these challenges with a novel cross-disorder gene discovery algorithm (Deep Neurodevelopmental Disorders algorithm - *DeepND.*) DeepND analyzes comorbid neurodevelopmental disorders simultaneously over multiple gene co-expression networks and explicitly learns the shared and disorder-specific genetic features using multitask learning. The proposed DeepND

architecture uses graph convolution to extract associations between genes from gene co-expression networks. This information is processed by a mixture-of-experts model that can self-learn critical neurodevelopmental time windows and brain regions for each disorder etiology which makes the model interpretable.

We provide a genome-wide risk ranking for each disorder and show that the prediction power is improved. DeepND identifies mediodorsal thalamus and cerebellar cortex brain region and infancy to childhood period as the highest neurodevelopmental risk window for both disorders. We observe that top percentile risk genes for both disorders are enriched in transcription regulators. Despite tight regulatory links in between ASD risk genes, we observe loose connectivity across ASD and ID risk genes or within ID risk genes. Finally, we investigate frequent ASD and ID associated copy number variation regions and confident false findings to suggest several novel risk gene candidates. This neural network architecture can easily be generalized to other disorders with a shared genetic component and can be used to prioritize focused functional studies and possible drug targets.

Chapter 2

Background Information

2.1 Copy Number Variations, *De Novo* Loss-of-function and Missense Mutations

Genes are the basic units of heredity that contain the information needed to specify traits by coding for a specific protein or segment of protein leading to a particular characteristic or function within the organism.

Copy number variation (CNV) is a structural variation in genome where a stretch of DNA, where possibly multiple genes are located, are duplicated or deleted [33]. CNVs not only result in changes in the physical arrangement of genes on chromosomes but they also effect the dosage of the products of the genes which are located in the CNV regions. As a result, CNVs in several regions are associated with different types of diseases [34, 35]. However, most of the time these CNVs occur in relatively large regions and it is not clear which genes in particular are the actual driver genes of target diseases.

A genetic alteration that is present for the first time in one family member as a result of either a mutation in a germ cell of one of the parents or a mutation arises in the fertilized egg itself during early embryogenesis is called a *de novo* mutation

[33]. A missense mutation is single base pair change in DNA which results in substitution of a different amino acid in the resulting protein [33]. While this substitution may have no or very little effect, it also can increase or decrease the effectiveness of the protein [36]. If the missense *de novo* variant in question disturbs the function of the molecular product coded by the gene, it is classified as *de novo* gene disrupting or loss-of-function mutation (dnLoF). Such variants are important because they explain why reproductively lethal disorders, such as neurodevelopmental disorders, continue to occur in the population [37, 38, 39, 40].

In this study, we will use information obtained by identification of missense and dnLoF functions as features and several CNV locations previously associated with either ASD or mental retardation to conduct an extended analysis on our results.

2.2 The Transmission and De Novo Association Test (TADA)

De novo mutations have high importance in gene discovery due to their high signal-to-noise ratio. However, during sequencing studies the majority of the identified mutations are inherited variants and these analyses often fail to incorporate data from case-control studies. The Transmission And De novo Association test (TADA) is a Bayesian model that overcomes these challenges by combining data from *de novo* mutations, inherited variants in families, and standing variants in the population [41]. The model learns the parameters of disease risk gene and non-risk gene groups by likelihood maximization using a dataset including mutation counts and frequency of *de novo* variants in ASD case and control groups of different cohorts, the probability of loss-of-function intolerance (pLI) of genes and number of inherited and nonhereditary gene disruptive mutations in ASD trios.

TADA pipeline is extended and applied for three other neurodevelopmental

disorders: intellectual disability, schizophrenia and epilepsy [26]. The extended TADA (extTADA) provides genome-wide risk probabilities for mentioned disorders based on published mutation counts for these disorders. Results from TADA for ASD and extTADA for ID are used as features for the model that is proposed in this thesis.

2.3 Gene Co-expression Networks

Correlation among genes can be represented in a gene co-expression network, where each node correspond to a gene and genes are paired if a significant co-expression association between them exists. Gene co-expression networks are built by scoring similarity among all gene pairs, choosing a score threshold and connecting all gene pairs whose scores exceed this value [42]. The expression level for a gene is measured using the amount of mRNA transcribed from it within a particular cell. Considering the cell differentiation in multicellular organisms, different transcription levels can be measured using cells from different tissues for a particular gene. Also, gene activations change during the life cycle of the organism. Therefore, gene co-expression networks may change depending on the subject tissue and observation time.

There are different strategies to calculate co-expressions among genes, normalization of the data and similarity metrics to use to built gene co-expression networks. We use Pearson correlation coefficient or bivariate correlation to calculate gene similarity from measured expression values. Bivariate correlation measures linear correlation between two genes and ranges from -1 to $+1$. Then, we set a threshold to obtain the adjacency matrix for the gene co-expression network from calculated correlation values where correlations stronger than the threshold is represented with an undirected and unweighted edge in the network. Since the selected threshold effects the number of total edges (i.e. the size of the network), it can be adjusted according to the memory limitations of the computational unit. In this thesis, we used networks created by using the data from different brain regions of patients with varying ages for our analysis.

2.4 Biological Pathways and Transcription FactorsAssociated with ASD or ID

Biological pathways are series of interactions of molecules and signals in a cell that lead to certain changes. They control the cellular functions and its responses to the environment. The affects of biological pathways can spread on short or long distances such as triggering mitosis on nearby cells or produce hormones to target distance cells [33]. Mutations in the genes that take part in a particular pathway may disturb the functions carried out by that pathway which may cause various diseases such as cancer, diabetes and neurodevelopmental disorders [43, 44, 45]. Analyzing these pathways is important to identify risk genes for diseases, and to develop effective treatments and diagnostic and prognostic markers. Recent studies identified a number of biologic pathways related to ASD and ID as these pathways are shown to distracted commonly in ASD and ID case groups such as WNT signaling [46] and MAPK signaling [47].

There are also proteins such that can be related to neurodevelopmental disorders since they regulate expression of genes in the cell. Some of the transcription regulators that are previously linked with ASD or ID includes CHD8 [48], FMRP [49, 50, 51], RBFOX [52, 53, 51] and TOP1 [54]. Enrichment of targets of mentioned transcription factors and pathways can be used as a performance measure to validate risk predictions for disorders since it provides a comparison within the disrupted circuitries affected by ASD and ID. In this thesis, we also include genes that affect post-synaptic density complex [55, 56], and histone proteins[51, 57]) in our analyses since these two structural molecules highly impact gene expression.

Chapter 3

Related Work

In this chapter, we describe related methods from the literature.

3.1 DAWN

One of the network based gene discovery algorithms in the literature is Detecting Association With Networks (DAWN) [31]. As a Hidden Markov Random Field (HMRF) based model, DAWN consists of three stages: Partial Neighborhood Selection (PNS), HMRF parameter estimation and Bayesian FDR correction. To increase the power of HMRF, the adjacency matrix of the network should be as precise as possible. Due to the the high dimensionality and small sample size, computing the entire adjacency matrix becomes a challenge. PNS stage aims to decrease the size of network to overcome this challenge by screening key genes with relatively small TADA p-value and exclude ones that are not co-expressed with any other measured genes. To screen genes, PNS uses two parameters that are set depending on prior subject knowledge, problem size and computational resources. In the study, the performance of PNS is found to be stable as long as the parameters are not overly small. In the second step, HMRF parameters are estimated on the network where normalized Z-scores ofrisk and non-risk genes

are assumed to have a Gaussian mixture distribution. The mixture membership of a gene is determined by its hidden state which indicates whether it is a risk gene or not. After computing the posterior probability for a gene with HMRF, Bayesian FDR correction is applied to obtain the final risk gene set.

A major setback of DAWN is that it only uses a single network to predict risk scores which cannot utilize evolving gene-interactions over time simultaneously. Therefore, results from two separate spatio-temporal windows, PFC3-5 and PFC4-6, among 4 brain regions and 13 neurodevelopmental periods covered by BrainSpan trascriptome dataset [58] are reported as these networks are previously reported to be the most critical networks for ASD [3].

3.2 DAMAGES

Disease-associated mutation analysis using gene expression signatures (DAMAGES) analysis is proposed to investigate autism-susceptibility genes by using cell-type specific gene expression profiles [59]. The study uses gene expression data from central nervous system cells of 24 mice, coming from 6 brain regions where 76% of the genes within the database are orthologous to human genes. Instead of building a co-expression network, the gene expression levels in this set are directly used. For each gene in the set, DAMAGES score (D-score) is calculated by adding a constant to the log2-transformed intensity vector of a gene projected onto second principal component which has the highest contribution to prediction according to the lasso analysis on principal components. The constant to be added is determined by leave-one-out cross validation.

To obtain the final scoring, DAMAGES combines D-score with Exome Aggregation Consortium (ExAC) metrics by using them as features (X_i) to train a logistic regression model. After estimating the effect size (β_i) for each feature by a standard logistic regression, the final prediction score, i.e. ensemble score, for

each gene, i, is calculated as follows:

$$S = \frac{1}{[1 + e^{-(\beta_0 + \sum_l \beta_i X_i)}]} \tag{3.1}$$

3.3 SVM Based Genome-wide Risk Prediction for ASD

One of the state-of-art methods for gene risk assessment is an evidence weighted support vector machine (SVM) based algorithm proposed by Krishnan *et al.* [32]. Initially, they integrate thousands of large-scale genomic data using a regularized Bayesian approach to obtain the gene interaction network. This network is stored as a weighted edge matrix X, where $.X_{ij}$ is the naive Bayes functional-relationship posterior probability of genes i and j participating in the same pathway in the brain. The matrix is used to train the SVM classifier to predict autism-association genes. Due to the small number of labeled genes, they use 5-fold cross validation on labeled genes. Given network features X_i and label y_i or all the training genes, the linear SVM solves the following optimization problem:

$$min_w \frac{1}{2} w^T w + \sum_{i=1}^{m} l_i \cdot max(0, 1 - y_i w^T w)^2 \tag{3.2}$$

In this equation, m is the number of training samples which is 20% of all labeled genes and l_i is the penalty parameter that represents how costly it is to misclassify gene i. In the study, l_i is set to be equal to the evidence-weight of the labeled gene which is determined by combining information from various sources.

This method does not provide direct information on which development periods or time windows are more informative for ASD risk prediction. However, Krishnan *et al.* analyzes the enrichment of genes in spatio-temporal brain windows and resulting heatmap suggest that genes associated with ASD are more active during early fetal to midfetal periods in almost all brain regions.

3.4 ST-Steiner

ST-Steiner is an algorithm based on prize collecting Steiner trees that tries to utilize evolving gene-interactions over time [23]. ST-Steiner uses a cascade of networks representing spatio-temporal windows of neurodevelopment which are referred as plates. Plate 1 is the precursor network and represents the early fetal time window in development whereas Plate 2 represents the mid-fetal period in general.

ST-Steiner considers a spatio-temporal system $G = G_1, G_2, \ldots, G_T$, a list of T consecutive time windows. Given a spatio-temporal system G, the problem is finding a minimum spatio-temporal sub-system $F = F1, F2, \ldots, FT$ by optimizing the following equation.

$$o_F(F) = \sum_{e \in E_F} c(e) + \beta \sum_{v \notin V_F} p(v) + \omega \kappa_F, \beta \geq 0, \omega \geq 0 \tag{3.3}$$

In this equation, β is a sparsity parameter to adjust trade-off between node prizes and edge costs, $p(v)$ is the prize function for nodes and $c(e)$ is the cost function for edges. $\kappa_F \in N$ is the number of connected trees in the subgraph F and ω is a parameter corresponding penalty. In this setup, the subgraph found in Plate 2 is affected by its precursor as the artificial prize function of a node in Plate 2 is changed depending on whether the same node is included in the subgraph for Plate 1. Although ST-Steiner can model the development of brain with this approach, it does not give a genome-wide ranking since it only provides the optimized subgraph. Also, the method does not have a straight forward policy for the size of this subgraph and the experiments for pruning or extending the graph may be costly and impractical.

3.5 DeepASD

Deep Autism Spectrum Disorder algorithm (DeepASD), proposed by [60], is a deep learning model that prioritizes ASD risk genes and learns important neurodevelopmental periods by processing both mutational burden and gene co-expression networks. DeepASD uses graph convolutional (GCN) layers which estimates convolutional filters to be applied on the arbitrarily structured graph data by considering the node features as signals and the filters as approximations of the graph Laplacian. In this thesis, we use 52 spatio-temporal gene co-expression networks constructed from 4 brain region clusters and 15 time windows, as explained in Willsey *et al.* [3]. These are shown in TableA.2 and Table A.1. These networks are used as the input for GCN layers and the mutational burden is used as node features for these networks where each node is a gene and an edge exists between nodes if genes are coexpressed in that neurodevelopmental window. Then, the mixture of experts (MoE) method is used to combine information from GCN layers as weak learners and produce the final output as a weighted sum of the outputs of these weak learners.

DeepASD improves the performance in terms of area under ROC curve and area under precision-recall curve distributions. In addition, the gating module inside MoE structure enables DeepASD to identify significant spatio-temporal networks during its end-to-end training period. Despite its advantages, DeepASD gives the genome-wide risk ranking for a single disorder at a time and overlooks the comorbidity of neurodevelopmental disorders and the information coming from their shared genetic architectures.

Chapter 4

Methods

This section contains the description of our multitask risk assessment model for comorbid neurodevelopmental diseases.

4.1 Problem Description

In this study, we propose a multitask learning pipeline for comorbid neurodevelopmental disease to capture shared genetic architecture and increase prediction power for risk assessment. Let the graph representations of each 52 gene co-expression network j be $G_j = (V, E_j)$ where vertices $V = (v_1, ..., v_n)$ corresponds to genes in human genome and $E_j \in \{0, 1\}^{n \times n}$ is the binary adjacency matrix. Then, let $X_D \in \mathbb{R}^{n \times d}$ as the feature matrix for disorder D where each row $X_D[i]$ is a list of d features associated with gene and node $i, \forall i \in [1, n]$. Consider two vectors, y_{ASD} and y_{ID}, of dimension l and m respectively. $y_{ASD}[i] = 1$ if the node i is a risk gene for ASD and $y_{ASD}[i] = 0$ if the node corresponds to a non-mental health related gene $\forall i \in [1, l], l < n$. y_{ID} is defined similarly where $y_{ID}[i]$ is either 1 or 0 $\forall i \in [1, m], m < n$. In our semi-supervised learning task, our goal is to learn a function $f(X_{ASD}, X_{ID}, y_{ASD}, y_{ID}, G_1, ..., G_{52}) \rightarrow P \in \mathbb{R}^{n \times 2}$ where $P[i][ASD]$ denotes $p(y_{ASD}[i] = 1)$, and $P[i][ID]$ denotes $p(y_{ID}[i] = 1), \forall i \in [1, n]$.

4.2 DeepASD

Here, we describe the technical details of the DeepASD method which is an important building block in DeepND architecture [60].

4.2.1 Graph Convolutional Network Model

One of the revolutionary architectures in machine learning and computer vision fields are Convolutional Neural Networks (CNNs). They are widely used on grid-structured n-dimensional regular signals, such as images, to extract local patterns [61]. While applications of the same principle on arbitrarily structured graph data have also improved the state-of-the art [62, 63, 64], these spectral approaches are computationally expensive. DeepASD uses graph convolutional networks (GCNs) proposed by Kipf and Welling as a solution to the excessive computational requirement of previous methods [65]. GCNs approximates the filters in the regular convolution operation as a Chebyshev expansion of the graph Laplacian [66] and let them operate on the 1-hop neighborhood of each node by limiting the order of Chebyshev polynomials in the expansion. GCN modules in DeepASD inputs the normalized adjacency matrix $\hat{E}_j$ with self loops (i.e., $\hat{E}_j[i,i] = 1, \forall i \in [1,n]$) of gene co-expression network G_j and the feature vector $X_D[i]$, for gene i and for disorder D. Then, each subsequent layer k of a GCN module is defined as:

$$H_k[i] = ReLU(\hat{D}^{-0.5}\hat{E}\hat{D}^{-0.5}H_{k-1}[i]W_{k-1}) \tag{4.1}$$

where W_0 is the weight matrix at the input layer to be learnt, H_0 is the feature vector X_D, $ReLU$ is the rectified linear unit function and $\hat{D}$ is the normalized version of a diagonal matrix where $\hat{D}_{ii} = \sum_j \hat{E}_{ij}$. In DeepASD, 2 GCN layers are stacked with embedding dimensions $d_1 = 4$ and $d_2 = 1$. The final layer uses softmax as the activation function that outputs probabilities for positive class. The final output of a GCN module j for a gene i is given as:

$$GCN(G_j, X_D[i]) = softmax(H_2[i]) = v_i^j \in R \tag{4.2}$$

4.2.2 Mixture of Experts Model

In order to combine the information from 52 GNC modules, DeepASD uses the mixture of experts model (MoE) which is used to determine informative models among a collection [67]. MoE module for disease D inputs the feature vector X_D and consider 52 GCN modules which act on gene co-expression networks $(G_1, \cdots, G_{52})$ as individual experts. For every gene i, MoE produces a vector w of length 52 by which GCNs are weighted after it is passed through a softmax layer (i.e., $\sum_{h=1}^{52} w_i[h] = 1$). Then, the weighted sum of expert results are used to assign the final risk probability for every gene i using softmax. That is, the $MoE(X_D[i]) = w_i \in R^{52}$ and the final risk probability of gene i for disorder D is found by $w_i \cdot v_i = P[i][D]$.

4.3 Proposed Algorithm

4.3.1 Multitask Learning Model

Above-mentioned GCN and MoE cascade inputs $X_D[i]$ and co-expression networks $G_1, .., G_{52}$ along with labels for a single disorder to predict the risk probability for every gene. We hypothesize that analyzing comorbid neurodevelopmental disorders simultaneously can increases the prediction power since the prediction for the disorders depend on each other's genetic architecture. To analyze multiple disorders simultaneously, we propose a pipeline that utilizes multitask learning (MTL). In MTL paradigm, the algorithm first learns a shared component for all task at hand and then also focuses on specific tasks. Learning the shared component has shown to boost the performance in various tasks such as language and image processing [68, 69]. Our proposed algorithm, DeepND uses the MTL paradigm where multiple tasks are solved concurrently (i.e., genome wide risk assessments for ASD and ID.) The system figure of the proposed method is described in Figure 4.1

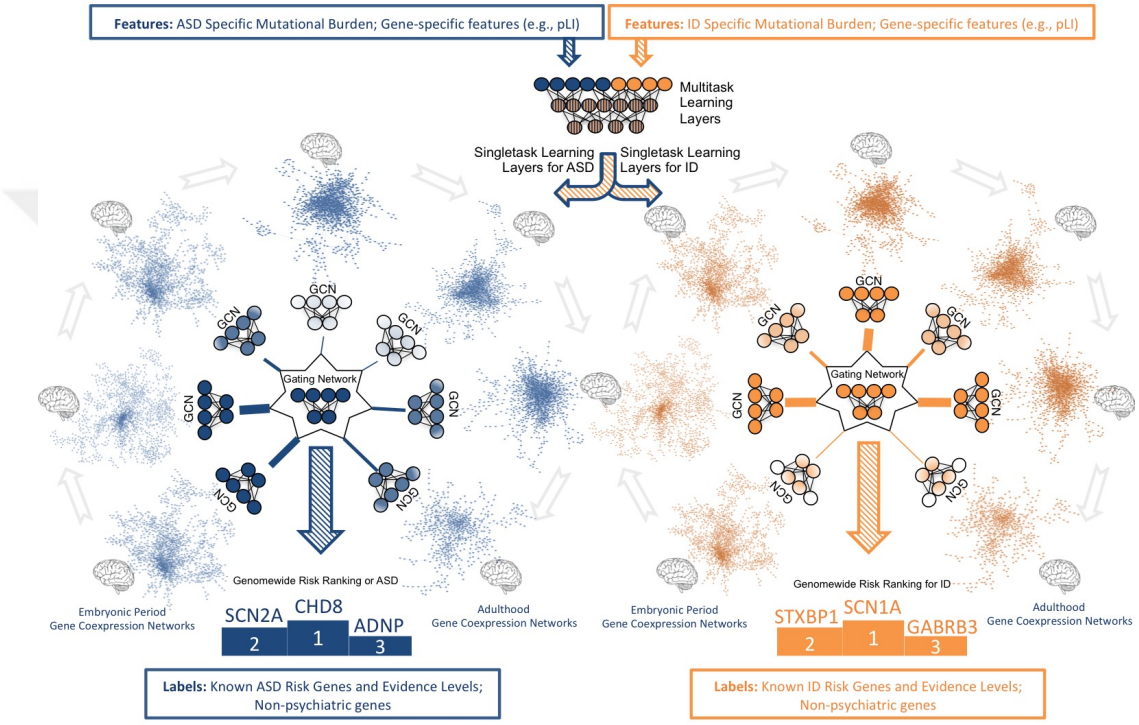

Figure 4.1: System model of the proposed deep learning architecture for genome-wide cross-disorder risk assessment (DeepND). The features are passed through fully-connected multitask layer that learn shared weights for both disorders and produces a new feature representation. This new representation is then input to singletask GCNs, each processing one of fifty-two gene co-expression networks that represent different brain regions and neurodevelopmental time windows. The output of GCNs are then weighted by the Gating Network to learn which networks are informative for the gene risk assessment for each disorder's etiology. The final output is a genome-wide risk probability ranking per disorder, which are then used for various downstream analyses to understand the underlying functional mechanisms and to compare/contrast both disorders. The singletask layers are exclusively trained with the ground truth genes of the disorder they belong. Thus, they learn only disorder specific parameters and disorder-specific networks that implicate risk.

DeepND employs one DeepASD cascade per disorder and puts a multi-layer perceptron (MLP) as a precursor to two DeepASD subnetworks. The weights learnt on these subnetworks are only affected by the back-propagated loss of the corresponding disorder, and hence, these are single task parts of the architecture. On the contrary, the weights learnt on the MLP part are affected by the loss back-propagated from both subnetworks. Thus, this part corresponds to the multitask component of the DeepND architecture. The MLP layer inputs the union of X_{ASD} and X_{ID} and passes it through a fully connected layer followed by Leaky ReLU activation (negative slope = -1.5) to learn a weight matrix W_{MLP} and output a d' dimensional embedding to be input DeepASD instead of $X_{ASD}[i]$ and $X_{ID}[i]$. The function learned by MLP module is defined as:

$$MLP(X_{ASD}[i] \cup X_{ID}[i]) = LRELU(W_{MLP}^{\intercal} \cdot (X_{ASD}[i] \cup X_{ID}[i]) + b_{MLP}) \quad (4.3)$$

4.4 Experimental Setup

4.4.1 Ground Truth Risk Gene Sets

DeepND is based on a semi-supervised learning framework in which only a subset of the samples are labeled as either positive or negative. The algorithm uses a different gold set for each disorder considered. For ASD, the ground truth set from Krishnan *et al.* [32] is used. The ASD related genes are collected from various sources and classified into four evidence levels (E1-E4, E1 indicating the highest risk). The final risk contains 46 E1, 67 E2 and 525 E3 and E4 genes as positively labeled ASD-risk genes. The negative set 1185 non-mental health related genes from Krishnan *et al.* [32]. In Table A.3 and A.4, the final positive and negative gold standards for ASD are presented.

For ID, we have curated a positive gold set from various review studies from the literature as the provide in depth analyses of ID risk genes. The base set contains *known gene* ID risk genes from 2 landmark reviews [10, 70] and these genes are divided into evidence level according considering information from 3

other studies [71, 72, 73] where the genes marked as ID risk gene by all 5 studies are assigned as E1 genes and the genes marked as ID risk gene by 4 studies are assigned as E2 genes. The negative gene set is kept the same as ASD as they are non-mental health related genes. However, the overlapping genes within E1 and E2 are discarded from the negative set. The final list contains 56 E1, 131 E2 and 1074 negative genes for ID. In Table A.5 and A.6, the final positive and negative gold standards for ASD are presented.

The weighted losses for both disorders are calculated according to the following weights: E1 genes (1.0), E2 genes (0.50), E3-E4 genes (0.25) and negative genes (1.0). The validation and test performance are evaluated and reported only on E1 genes.

4.4.2 Gene Features

Each branch (subnetwork) of DeepND uses the mutation burden information for related disease as the node features for 52 GCN modules. For ASD branch, the mutational burden from the largest whole exome study to date is used to generate features [29]. The feature matrix for ASD, $X_{ASD} \in \mathbb{R}^{n \times d}$ where $n = 25,825$ and $d = 17$ contains the following information from multiple case-control studies: (i) the number of *de novo* loss of function mutations, (ii) the number of damaging missense loss of function mutations, (iii) the number of protein protein truncating variants (PTVs) for ASD case and control groups, (iv) the number of transmitted loss of function mutations, (v) the number of *de novo* protein truncating variants (PTVs) and missense mutations in published ASD probands, (vi) frequency of *de novo* variants in ASD individuals, and (Vii) pLI of the gene.

For ID branch, a similar set is obtained from [26]. The feature matrix for ID, $X_{ID} \in \mathbb{R}^{n \times d}$ where $n = 25,825$ and $d = 8$ contains the following features for each gene $i, \forall i \in [1, n]$: (i) the number of *de novo* loss of function mutations, (ii) the number of damaging missense loss of function mutations, (iii) extTADA Bayes factor which is calculated using an extension of TADA, (iv) the number of *de novo* PTVs and missense mutations in published ID/DD (developmental delay)

probands, (v) frequency of *de novo* variants in ID/DD individuals, and (Vi) pLI of the gene.

The MLP architecture which facilitates information sharing between two disorders uses the union of these two feature sets for each gene $i, \forall i \in [1, n]$ as its input.

4.4.3 Spatio-Temporal Gene Co-expression Networks

BrainSpan dataset of Allen Brain Atlas [74, 58] contains gene expression levels in samples from 16 regions of 57 postmortem brains that span 15 consecutive neurodevelopmental period from 8 post-conception weeks to 40 years. We constructed 52 spatio-temporal networks by partitioning the the dataset into developmental periods and clusters of brain regions as explained in Willsey *et al.* [3]. Brain regions clusters and development windows are given in Table A.1 and Table A.2. Each spatio-temporal windows contains three consecutive time periods as shown in Figure 5.4. Consider the graph representations of a gene co-expression network, $G = (V, E)$ where vertices $V = (v_1, ..., v_n)$ corresponds to genes in human genome and $E \in \{0, 1\}^{n \times n}$ is the binary adjacency matrix. For two genes i and j, $\forall i, j \in [1, n]$ where $n = 25,825$, $E_{ij} = E_{ji} = 1$ if their absolute Pearson correlation coefficient $|r_{ij}|$ is greater or equal to 0.8 in the related partition of the data set.

4.4.4 Learning and the Cross Validation Setting of DeepND

We use a five-fold cross-validation scheme where all labeled genes are uniformly and randomly distributed to the folds. At each training session, one fold for validation and one fold for testing are left out and the remaining three folds are used for training. The model uses categorical cross entropy loss (evidence weighted) which is compatible with binary-classification nature of the problem. The model

weights are initialized using Xavier initialization [75] to avoid vanishing gradients problem which is a concern for multi layer neural network models such as DeepND. All weights are updated by ADAM optimizer [76] up to 1000 epochs with early stop which is determined with respect to the loss calculated on the validation fold using only E1 genes as the positives and all negative genes. The ADAM optimizer provides a faster learning as it uses first and second moments to calculate the step direction on the hyperplane defined by high dimensional loss function.

DeepND uses the learning rates of 7×10^{-4}, 7×10^{-4} and 7×10^{-3}, for the shared layer, ASD single task layer, and ID single task layer, respectively. To ensure proper and balanced convergence of DeepND, an overfit flag for each single task module is assigned and tracked. Once a flag for a single task module is set, meaning that the subnetwork is converged, the layers of that subnetwork and the shared layer are frozen. This lets the yet underfit subnetwork to keep learning till its own overfit flag is set or the epoch limit is reached. Once both overfit flags are set, i.e. the model converges, the test performance is reported on the test fold in the same manner as in the validation fold.

Following the procedure which is mentioned above, 4 results are reported as for every left-out test fold. This process is repeated 10 times with random initialization and a total of 200 results are used for performance comparison.

4.4.5 Computational Unit

All experiments for DeepND are run on a SuperMicro SuperServer 4029GP-TRT with 2 Intel Xeon Gold 6140 Processors (2.3GHz, 24.75M cache), 251GB RAM, 6 NVIDIA GeForce RTX 2080 Ti (11GB, 352Bit) and 2 NVIDIA TITAN RTX GPUs (24GB, 384Bit). For all training and testing, we used 5 RTX 2080 and 1 TITAN RTX cards and the cross validation setup took approximately 16 hours.

Chapter 5

Results

5.1 Genome-wide Risk Predictions for ASD and ID

Our results show that the genome-wide risk assessment of DeepND is robust, precise and sensitive and substantially improves the state of the art both in terms of performance and interpretability of the predictions.

We benchmark the performances of DeepND and other gene discovery algorithms for neurodevelopmental disorders. First, we compare DeepND and the state of the art algorithm by Krishnan *et al.* using their experimental settings. To quantify the contribution of the co-analyzing comorbid disorders (i.e., multitask) as opposed to individual analysis (i.e., singletask), we also compare our results with DeepASD. Evaluating performances of the algorithms using E1 genes through 5-fold cross shows that DeepND achieves a median AUC of 92% (Wilcoxon rank-sum test, $P = 1 \times 10^6$) and a median AUPR of 66% (Wilcoxon rank-sum test, $P = 1 \times 10^6$) for ASD, which correspond to 2% and 30% improvements over the evidence weighted SVM algorithm of Krishnan *et al.*, respectively (Figures 5.1). Similarly for ID, DeepND achieves a median AUC of 81% and a median AUPR of 35%, which improves the state of the art by 9% and 18%,

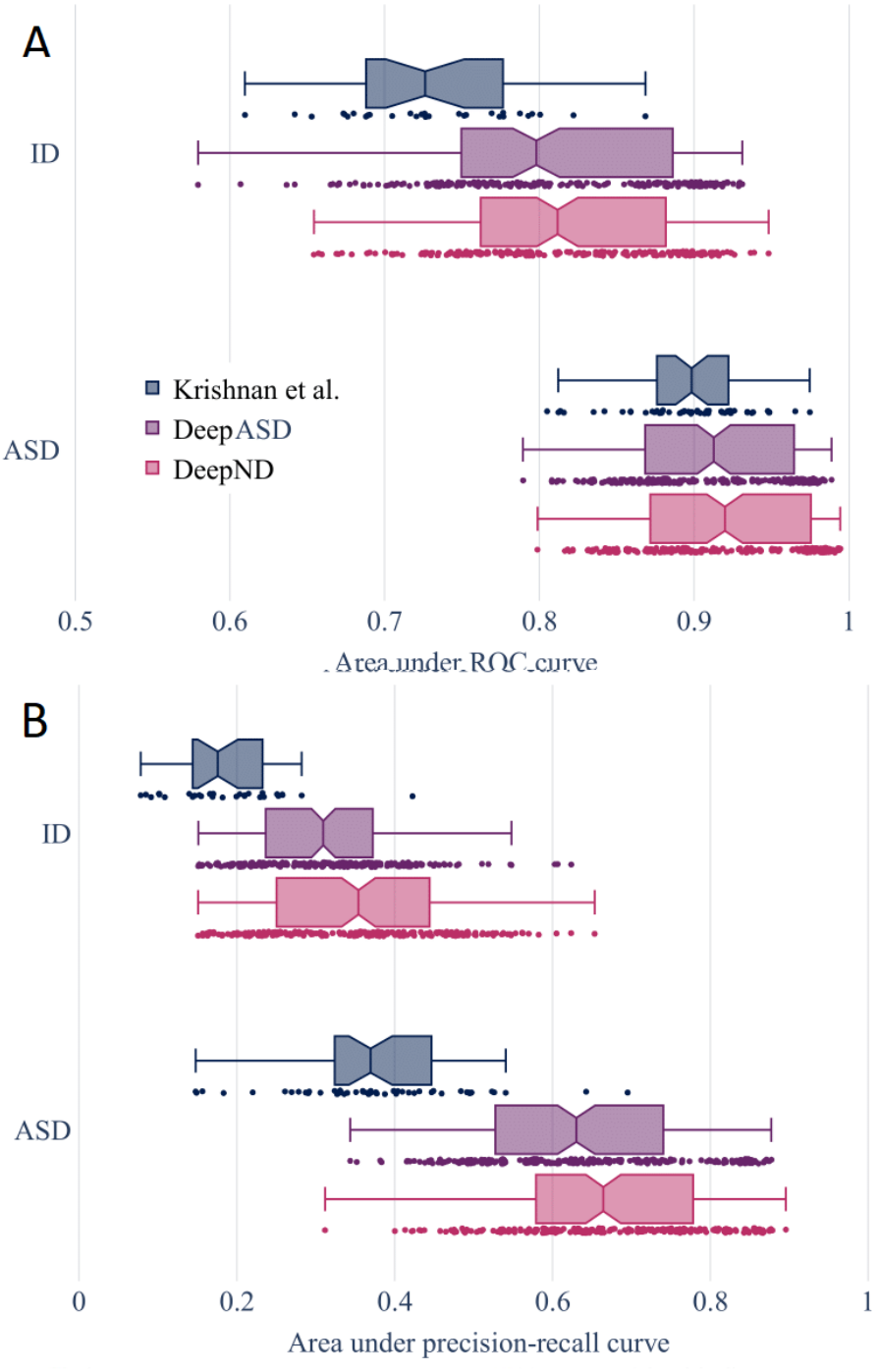

Figure 5.1: (**A**) The area under the receiver operating characteristic curve (ROC AUC) distribution of 200 results for DeepND, DeepASD and Krishnan *et al.* (**B**) The area under precision-recall curve (AUPRC) distributions for the same methods as in (**A**)

respectively. As expected, the multitask setting of DeepND performs better than DeepASD and leads to improved AUC (up to 1.4%) and AUPR (up to 5%) for both disorders.

We also compare DeepND and Krishnan *et al.*'s approach gene-by-gene for both disorders. Figure 5.2C shows the probabilities assigned to each gene (red: E1 genes; black: non-mental health genes; and grey: all other genes). The better performing algorithm would have the E1 genes on its own side of the diagonal (i.e., relatively higher risk probability assigned) and have the black genes on the other side (i.e., relatively lower risk probability assigned). We observe that DeepND consistently assigns higher probabilities to E1 genes and lower probabilities to non-mental health relate genes and thus, provides a better risk assessment (Wilcoxon Rank Sum test, DeepND ASD $P = 3.35 \times 10^{-22}$, DeepND ID $P = 5.07 \times 10^{-19}$, Krishnan *et al.* ASD $P = 5.07 \times 10^{-20}$ and Krishnan *et al.* ID $P = 7.12 \times 10^{-9}$).

Finally, we compare DeepND, DeepASD and Krishnan *et al.* with other neurodevelopmental disorder gene discovery algorithms from the literature that had output a genome-wide risk ranking: DAWN [31] and DAMAGES [59]. We compare them using precision-recall (PR) curves over the final genome-wide predictions of all methods. We consider 2 scenarios, where in the first one E1 genes are considered as true risk genes and in the second both E1 and E2 genes are considered as the ground truth. (Figure 5.3). For both disorders, DeepND consistently performs better than others.

5.2 Critical Neurodevelopmental Windows for ASD and ID

To further investigate the shared genetic component between ASD and ID, we focus on the spatio-temporal neurodevelopmental windows that are found to be important by DeepND for accurate ranking of risk genes for both disorders.

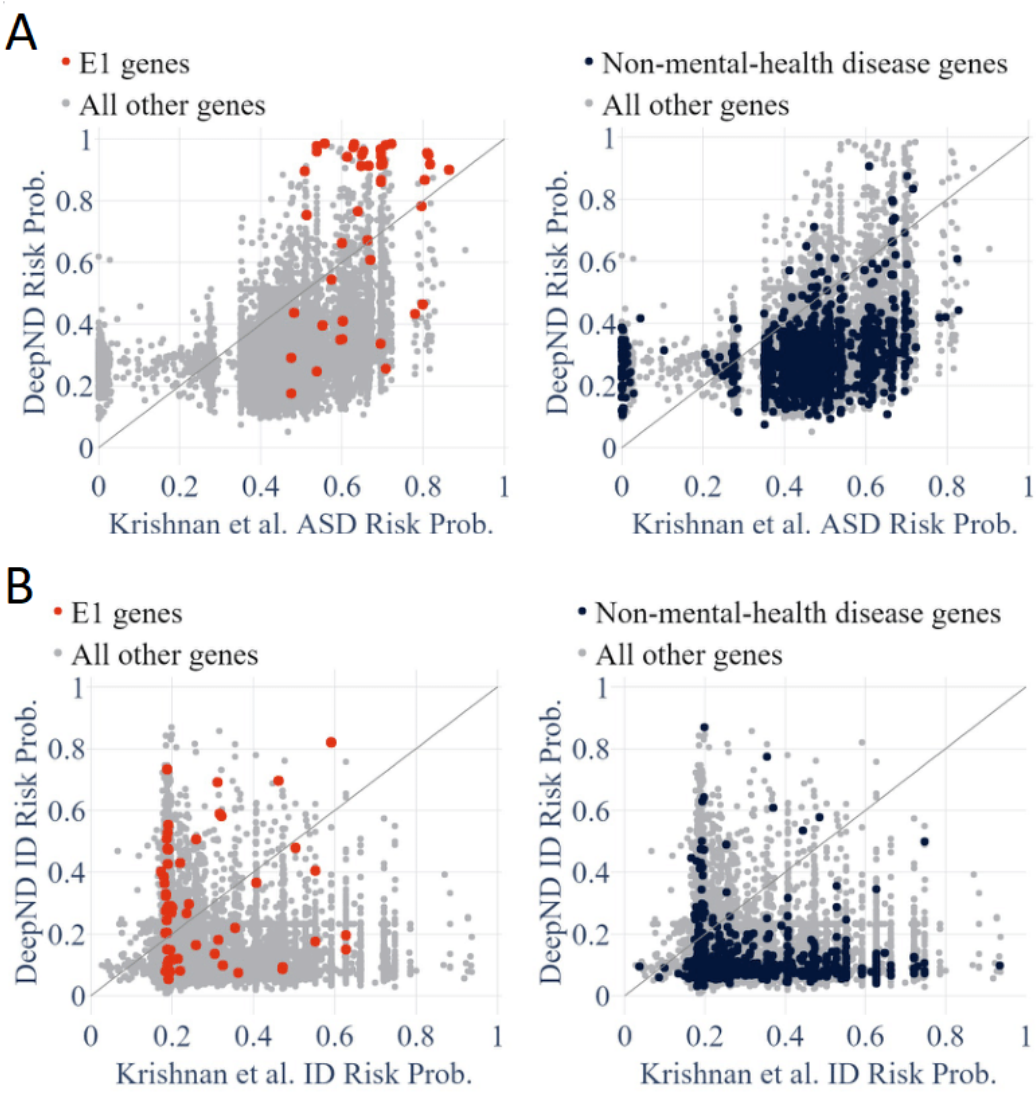

Figure 5.2: Scatter plots that presents assigned probabilities for each gene by DeepND and *et al.* for ASD (**A**) and ID (**B**). Red: E1 genes; Black: Non-mental-health disease genes; and Grey: all other genes. A better assessment method would have the E1 genes towards its side of the diagonal and have the black genes on the other side of the diagonal.

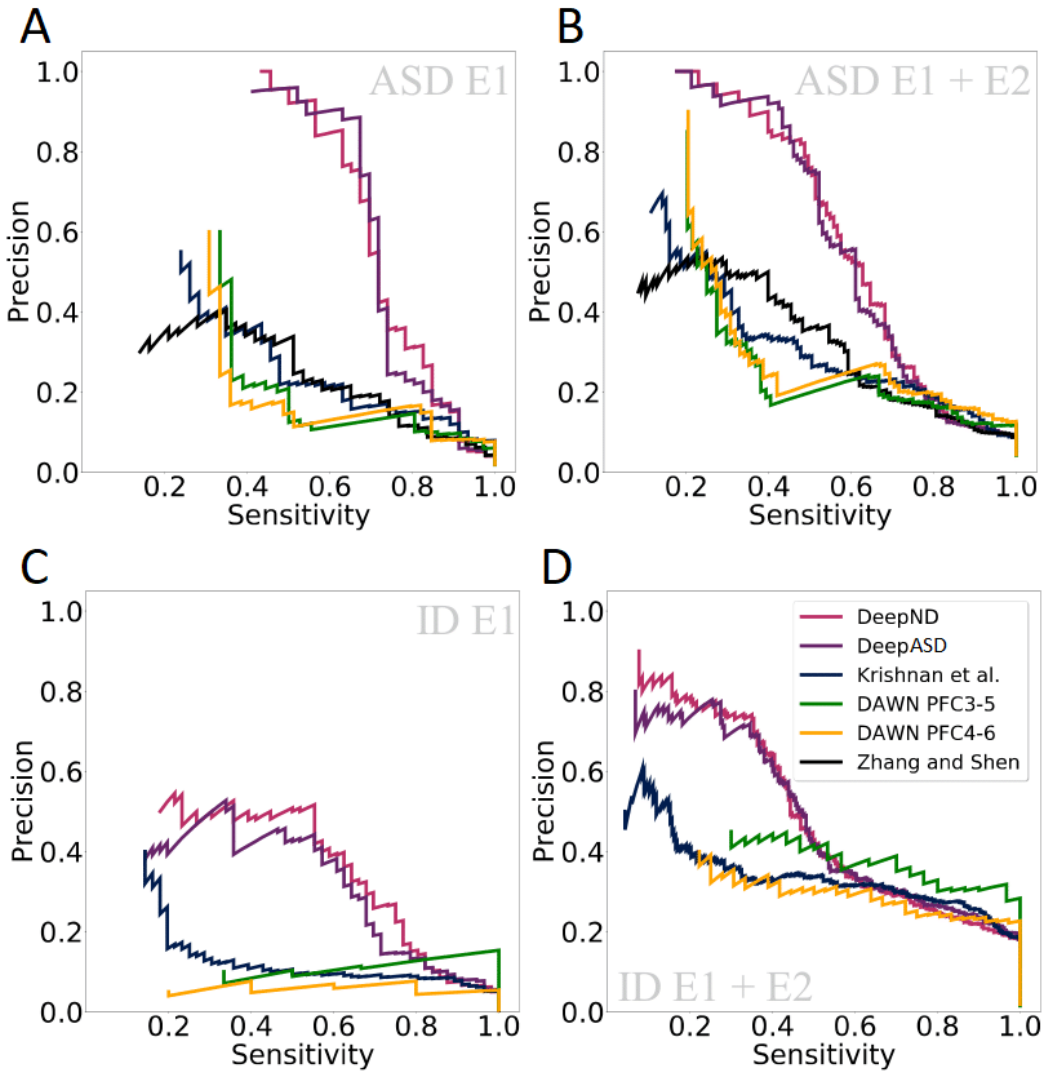

Figure 5.3: Precision - Recall curves to compare DeepND, DeepND - ST, Evidence weighted SVM of Krishnan *et al.*, DAWN, and DAMAGES score of Zhang and Shen. Results for ASD (**A** & **B**) and ID (**C** & **D**) shown when (i) E1 genes are used as the true risk genes and (ii) E1+E2 genes are considered as the true risk genes.

Our method analyzes co-expression networks that represent 13 neurodevelopmental time windows (from embryonic period to late adulthood; Figure 5.4) and 4 brain region clusters (PFC-MSC: Prefrontal and motor-somatosensory cortex; MDCBC: Mediodorsal nucleus of the thalamus and cerebellar cortex; V1C-STC: Primary visual and superior temporal cortex; SHA: Striatum, Hippocampus, Amygdala) generated using Brainspan RNA-Seq dataset [58] in accordance with Willsey *et al.* [3] (Figure 5.4).

We investigate which neurodevelopmental windows more confidently distinguish disorder risk genes. Figure 5.5 and Figure 5.6 show normalized average probabilities assigned to top percentile risk genes for ASD and ID, respectively. For each gene i, the average risk probability assigned by the corresponding GCN is calculated considering all iterations where gene i is in the test fold. Results of DeepASD shows that the networks of the MDCBC brain region, spanning several time windows from late/late-mid fetal to young adulthood (9-11) periods consistently are better predictors for ASD and ID risk. This is inline with the fact that ID or ASD linked genes are highly expressed during prenatal and early postnatal time windows in development [77, 78]. The period 8-10 (from neonatal and early infancy to early childhood) was also previously indicated as one of the highest risk regions and was subject to network analyses for ASD gene discovery [3, 31]. Prefrontal cortex mid-fetal (PFC-MSC 3-5) window was another region deemed important in the same study for ASD, but no such significant signal is picked up by DeepASD. Overall, we observe that earlier time windows are relatively more important for ID compared to ASD.

To compare the effect of co-analyzing comorbid disorders, we compare the results from DeepND and DeepASD for both disorders. In the multitask setting of DeepND, the change in the informativeness of networks for ASD is subtle compared to the singletask setting. For ID, we see a more scattered picture and a more diverse set of networks are more informative when cross-talk among ID and ASD genes are considered. This is also evident in the higher prediction performance gain in ID compared to ASD (Figure 5.1). Although the later periods become more informative, MDCBC infancy - childhood region is still one of the most important windows.

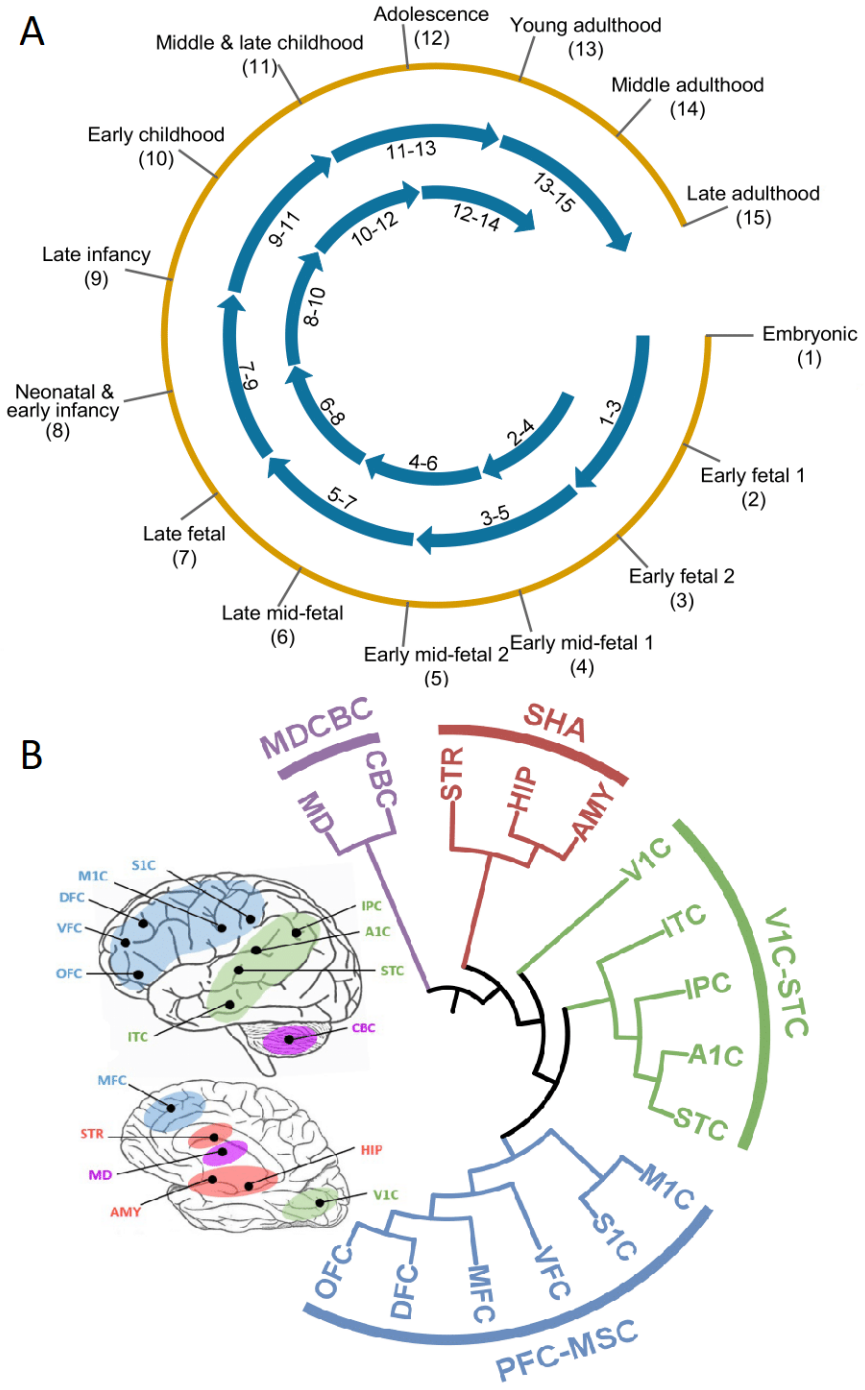

Figure 5.4: Neuroevelopmental windows (**A**) and brain regions clusters (**B**) used for generating co-expression networks from Brainspan RNA-Seq dataset

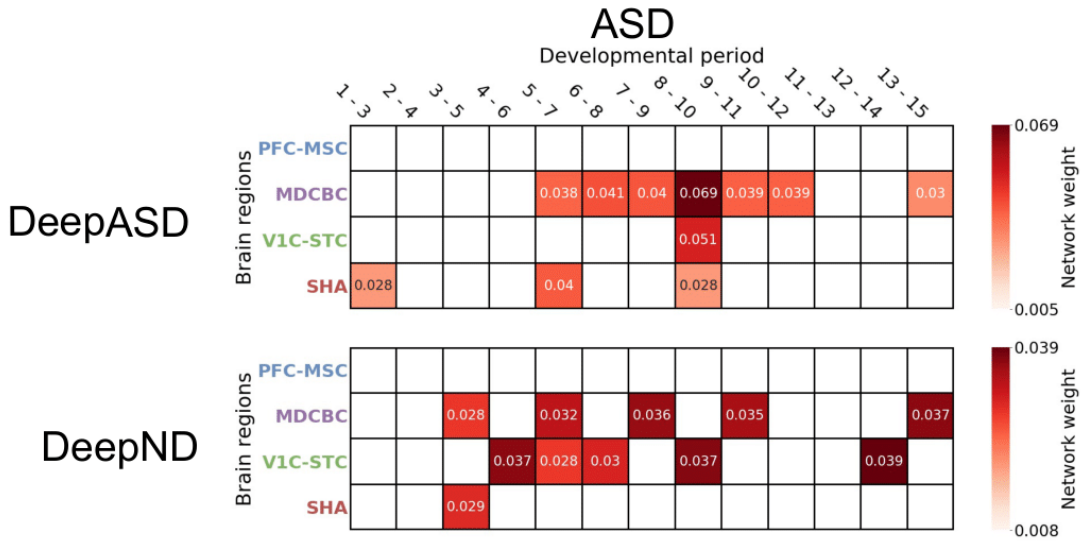

Figure 5.5: Normalized average probabilities assigned to top percentile risk genes by DeepASD (top) and DeepND (bottom) for ASD

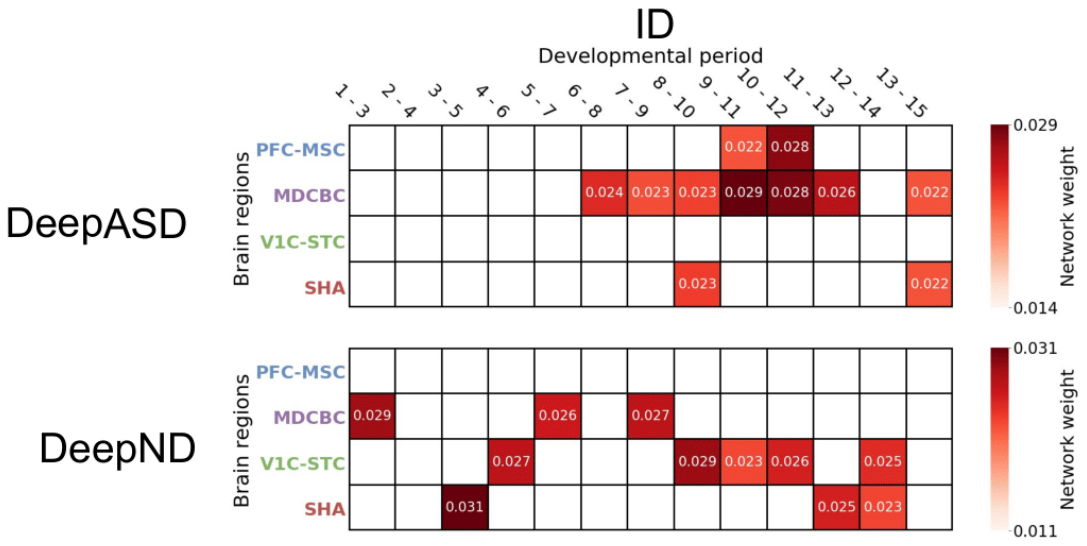

Figure 5.6: Normalized average probabilities assigned to top percentile risk genes by DeepASD (top) and DeepND (bottom) for ID

Using the MoE model, DeepND architecture associates a weight with each network which is proportional to its informativeness for the ranking task at hand. Assigned weights for each network for ASD and ID are shown in Figure 5.7 and Figure 5.8 respectively. It is observed that MDCBC late/late-mid fetal to late childhood/adolescence periods are consistently selected to be highly informative as these periods are associated with high weights. MDCBC 9-11 window is the top in 3 out of 4 analyses indicating that this spatio-temporal window is critical for the etiology of both disorders.

MDCBC 2-4 and MDCBC 9-11 networks which are marked as weakest and strongest sources of information by DeepND are visualized in Figure 5.9 and 5.10. There are roughly 12.5k links between top percentile ASD and ID genes in MDCBC 2-4 network, although the network contains close to 37.5m links in total. On the other hand, MDCBC 9-11 contains close to 13m edges. Yet, there exists close to 30k links between top percentile ASD and ID genes which is the reason behind DeepND focusing on this window as its top predictor.

5.3 Enrichment Analysis of the Predicted Risk Genes

We also analyze enrichment of the ASD and ID risk gene rankings in gene sets which are previously shown to be related to these disorders. Enrichment of member genes of these gene lists in higher deciles of the genome-wide ranking given by DeepND indicates the wellness of the predicted ranking. These gene sets consist of (i) targets of transcription regulators (CHD8 [48], FMRP [49, 50, 51], RBFOX [52, 53, 51] and TOP1 [54]); (ii) Susceptibility pathways (WNT signaling [46] and MAPK signaling [47]); and (iii) affected biological processes and molecular functions (post-synaptic density complex [55, 56], histone modification [51, 57]).

Figure 5.11 shows the enrichment of ASD and ID risk rankings in mentioned gene sets. We observe a similar trend for top decile ASD and ID genes such that

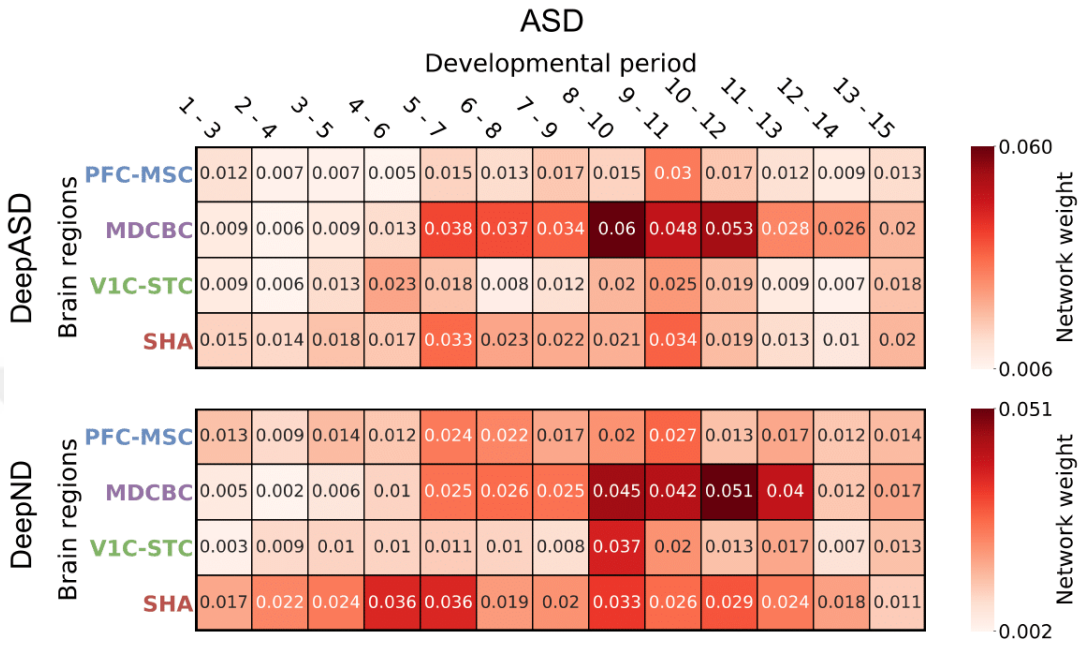

Figure 5.7: Heapmap shows average weights assigned by MoE module of DeepASD (top) and DeepND (bottom) to each spatio-temporal window for the top percentile genes for ASD.

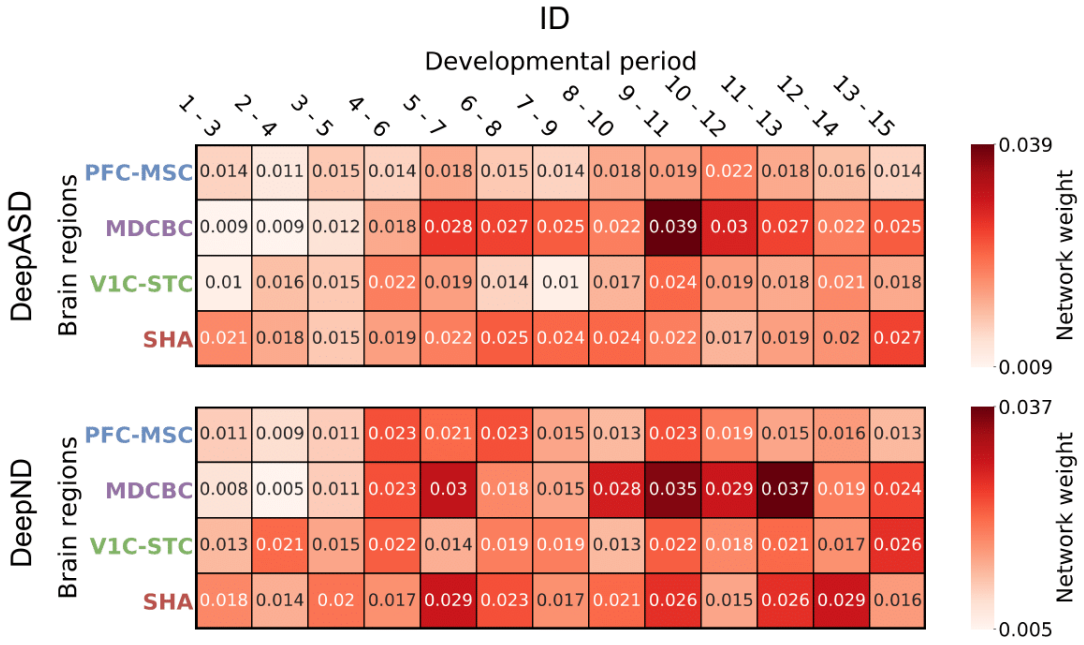

Figure 5.8: Heapmap shows average weights assigned by MoE module of DeepASD (top) and DeepND (bottom) to each spatio-temporal window for the top percentile genes for ID.

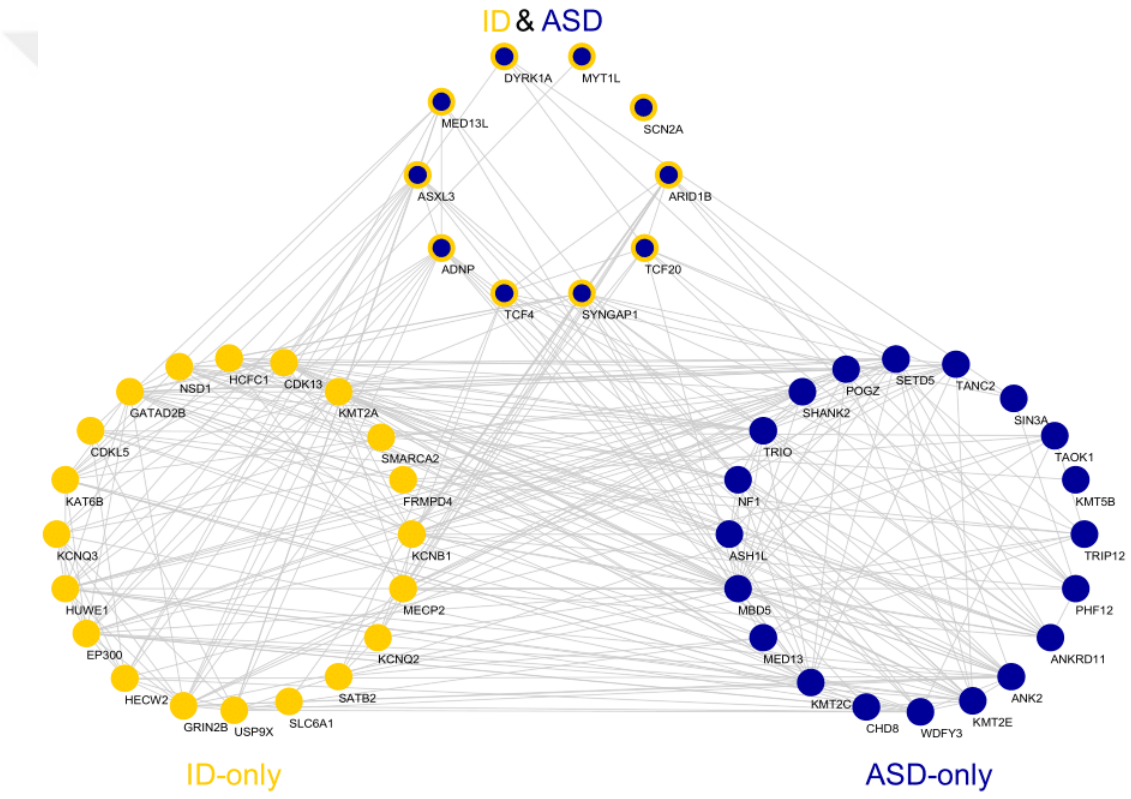

Figure 5.9: The summary of co-expression relationship between top percentile genes in each disorder on MDCBC2-4 network which is observed to be the weakest source of information. Although the full MDCBC2-4 network has 37.5m edges, only approximately 12.5k of them are between top percentile risk genes.

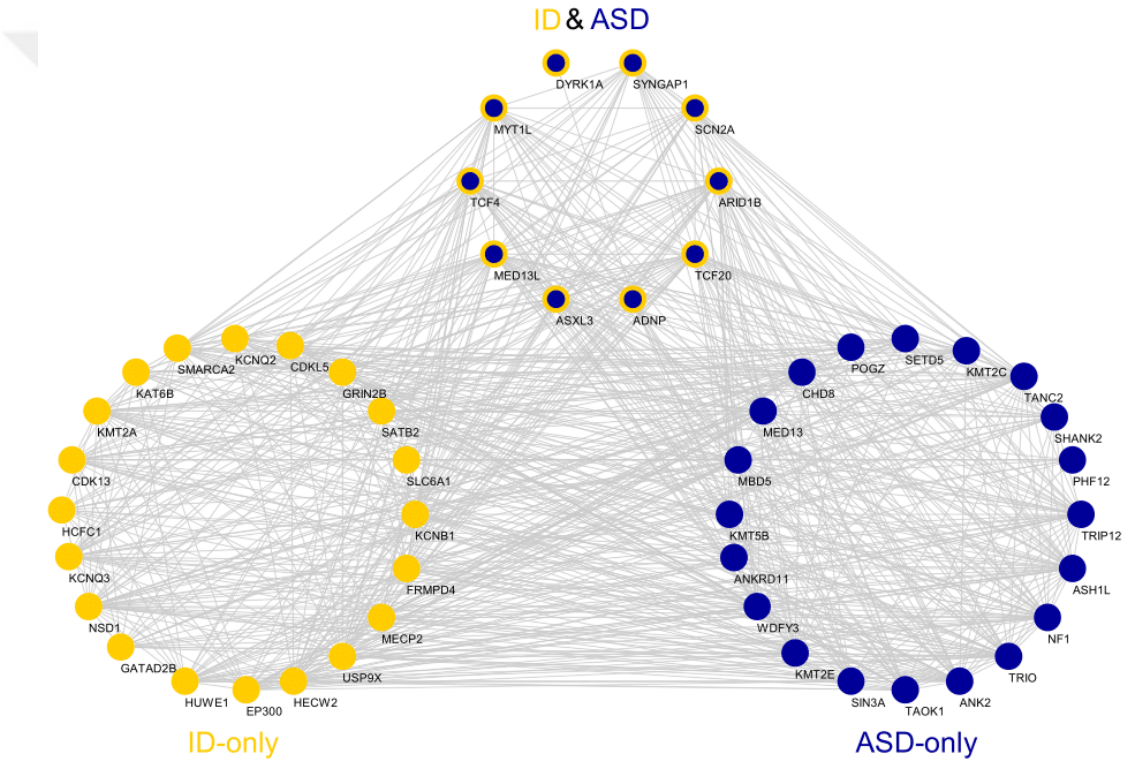

Figure 5.10: The summary of co-expression relationship between top percentile genes in each disorder on MDCBC9-11 network. The full MDCBC network contains 13m edges and 30k of them are between top percentile risk genes. The ratio of links between top percentile genes to total edges also indicate the importance of MDCBC9-11 for both disorders' ethiologies.

the top decile risk genes for both disorders have the highest enrichment in all categories. All enrichments for ASD are shown to be significant with respect to Binomial test, while the P values for ID are more subtle. For example, the fifth decile of the ID risk ranking is the most enriched set among the CHD8 targets and the distribution into deciles is not significant (Binomial test, $P = 5.9 \times 10^{-2}$). On the other hand, $CHD8$ (Chromodomain Helicase DNA Binding Protein 8) is the highest ASD-risk gene known to date with the highest mutation burden in large ASD cohorts [51, 29] and with downstream functional analysis [48]. Accordingly, DeepND ranks $CHD8$ as the third top risk gene whereas Krishnan *et al.* ranks it 1943^{th} genome-wide. While it has a solid association to ASD etiology, the ties to ID is not well-established. Bernier *et al.* reports that 9 out of 15 ASD probands with mutations in $CHD8$ also has ID. While it is a comorbid condition to ASD, $CHD8$ is not a well-established susceptibility gene for ID. Lower enrichment of ID genes in CHD8 targets is in line with this hypothesis. In accordance, DeepND places CHD8 as the 150^{th} highest risk gene for ID.

Another analysis regarding enrichment of genes is performed using Enrichr [79, 80] system for the top percentile risk genes. Enrichr is a platform that contains a large collection of diverse gene set libraries available and enables untargeted enrichment analysis in these gene sets. We find that *regulation of transcription from RNA polymerase II promoter* is the top enriched Biological Process GO term (Fisher's exact test, $P = 6.4 \times 10^{-19}$ for ASD and $P = 6.1 \times 10^{-13}$ for ID) which indicates that transcriptional regulation is a shared function that is disrupted in the etiology of both disorders. As for the Gene Ontlogy (GO) Molecular Function enrichment where genes are assigned to a set of predefined bins based on their functional characteristics, two disorders diverge in the mechanism: the top function affected for ASD is *RNA polymerase II transcription co-factor activity* whereas for ID the top affected function is *histone-lysine N-methyltransferase activity*. As the top predicted genes are responsible for regulating gene activity, we further investigate if there are any master transcription regulators upstream that regulate the high risk genes for ASD and ID in the ChEA Database [81]. We find that DMRT1 is the top transcription factor with 77 of its targets coincide with 258 the top percentile DeepND predicted ASD risk genes (Chi-Square test,

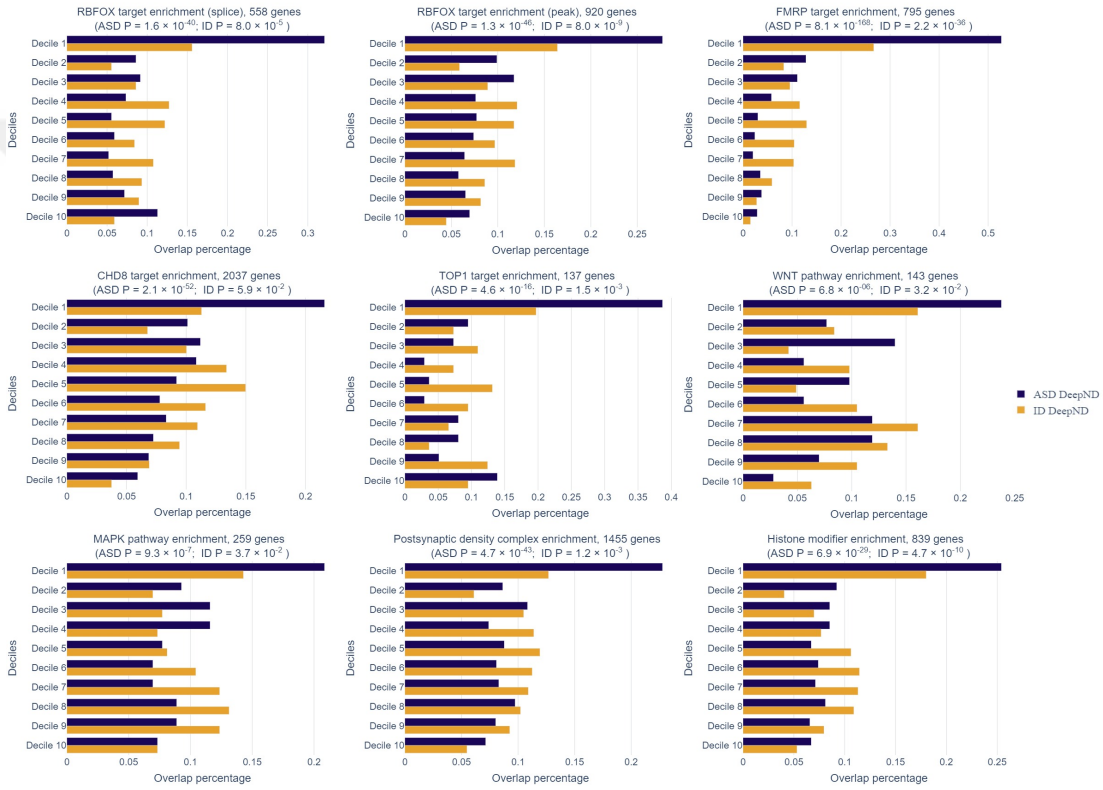

Figure 5.11: The enrichment of DeepND ASD and ID rankings in various disease related gene sets is given in each panel. The horizontal axis shows the percentage of genes in the corresponding gene set. The results for genome-wide risk ranking for ASD and ID are shown in blue and yellow respectively. The considered gene sets are as follows: (i) Targets of RBFOX (splice), (ii) Targets of RBFOX (splice target), (iii) Targets of FMRP (all peak), (iv) Targets of CHD8, (v) Targets of TOP1, (vi) WNT Pathway, (vii) MAPK Signaling Pathway, (viii) GTPase regulator activity, (x) Postsynaptic density complex genes, (xi) Synaptic genes (xii) Histone modifier genes.

$P = 2.84 \times 10^{-18}$). It is the top third regulator for ID by regulating 46 of 258 top percentile ID risk genes (Chi-Square test, $P = 1.74 \times 10^{-4}$), 22 targets are shared among ASD and ID. Note that only 69 genes overlap in top percentile ASD and ID risk genes. This means 30% of the shared set of genes among two disorders are targeted by DMRT1. When the union set of the top percentile genes for ASD and ID are considered (447 genes), we find that the top transcription factor targeting these is again DMRT1 (105 out of 2071 targets; Chi-Square test, $P = 2.33 \times 10^{-15}$).

While $DMRT1$ is not ranked as a high risk gene, Pinto *et al.* report a rare experimentally-validated *de novo* CNV (deletion) in an ASD proband which encompasses 9p24.34-p24.2 and $DMRT1$ [47]. $DMRT1$ is in the 9p region with the following genes: $DOCK8, DMRT2, DMRT3, VLDLR$, and $ANKRD15$ and several works indicate that mutations affecting these these genes might be related to neurodevelopmental disorders [82, 83, 84, 85]. DMRT1 is a transcriptional regulator and plays role in male sex determination. This finding suggests an interesting link between *DMRT1* and the strong male bias in ASD [**?**] and ID [**?**, **?**] prevalence.

5.4 Regulatory Interactions between ASD and ID Genes

The union of top percentile genes for ASD and ID contains 447 genes. We further investigate the existence of regulatory relation between these genes by using ChEA Database [81]. 25 of the common top percentile genes are found on this database as transcription factors (TF) with experimental target information (permutation test, $P < 1 \times 10^{-3}$). 9 of these 25 TFs are ASD-only risk genes whereas 9 of them ID-only and remaining 7 are both ASD and ID risk genes. Using a stringent matched randomization method based on mutation rate, pLI and gene length, we generate 1000 random gene sets for all 3 categories such that the set sizes are preserved and the original TFs are included in their relative risk gene

sets. The strongest connection among these gene sets is observed between ASD-only and ASD and ID gene sets both ways. TFs in both groups regulate each other significantly (permutation test - 1000 draws, $P < 1 \times 10^{-3}$; Figure 5.12). While TFs in ASD-only set regulate the genes in its group more frequently than expected by change, the significance is lower (permutation test - 1000 draws, $P = 3.1 \times 10^{-2}$). The same is true for TFs in ASD & ID set (permutation test, $P = 3.8 \times 10^{-2}$). This is in striking contrast with the relationships between ASD-only/ASD & ID TFs and ID-only risk genes. The strength of the connectivity is significantly less than expected (permutation test, $P < 1 \times 10^{-3}$). Similarly, TFs in the ID-only risk set do not significantly interact with other groups.

We further investigate the protein-protein interactions (PPI) between the top percentile risk genes in tissue specific PPI network for frontal cortex obtained from DifferentialNet database [1] using the NetworkAnalyst system [86]. This analysis reveals several hub proteins such as HECW2, EP300, CUL3, and CTTNB1 as shown in Figure 5.12. In this list, $HECW2$ gene stands out as it has the highest degree in this network and has very low prior risk for both ASD (TADA $P = 0.95$) and ID (extTADA $Q = 0.97$). Yet, it is the top 4^{th} ID risk gene identified by DeepND and in the top percentile for ASD risk. Note that DeepND did not use any PPI network information in its reasoning, and yet, was able to identify this hub protein which has been linked with ASD [87] and ID [88] via *de novo* disruptive mutations in simplex families.

5.5 Identification of Novel Candidates within ASD and ID Associated with CNV Regions

We investigate the DeepND risk ranking within (i) six regions which frequently harbor ASD-related CNVs (16p11.2, 15q11-13, 15q13.3, 1q21.1 and 22q11) [32], and (ii) six regions which were reported to harbor mental retardation related CNVs (16p11.2, 1q21.1, 22q11.21, 22q11.22, 16p13 and 17p11.2) [89]. In addition to supporting the linkage of previously reported confident genes, such as

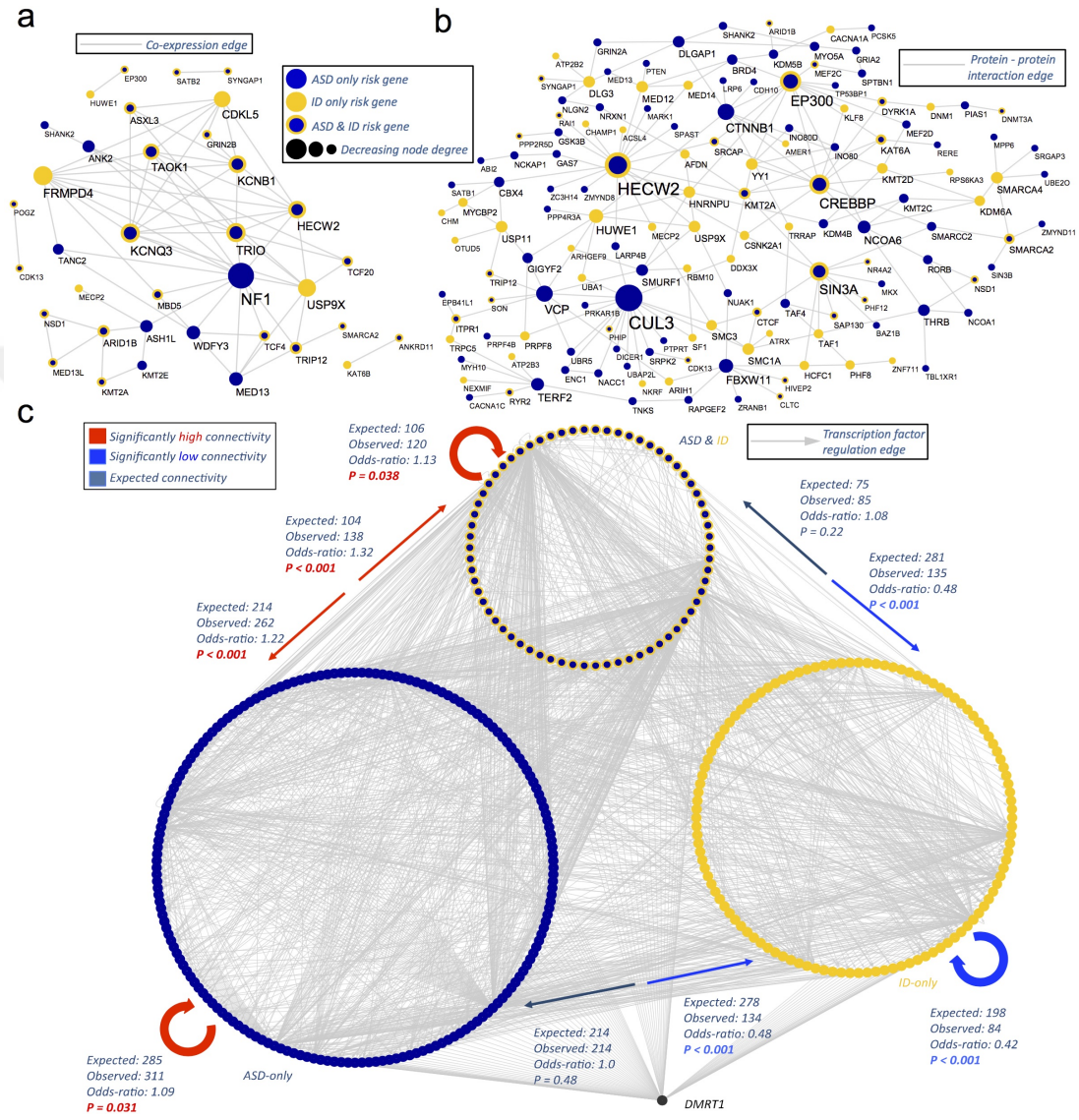

Figure 5.12: Network Analyses of the risk genes. (a) The co-expression relationships between top 30 ASD and ID risk genes in MDCBC 9-11 region. Only links for at least 0.95 absolute correlation and genes with at least one connection are shown. (b) The protein-protein interactions between the top percentile risk genes are shown. The PPIs are obtained from the tissue specific PPI network of frontal cortex in DifferentialNet database [1]. (c) The regulatory relationships between the top percentile ASD-only and ID-only risk genes in the ChEA transcription factor regulation database are shown. 69 genes are ranked in the top percentile for both disorders. Red arrows indicate that the TFs in the source has more targets than expected by chance in the target group. Cobalt blue arrows indicate that the TFs in the source has less targets than expected by chance in the target group and aegean blue arrows indicate expected connectivity.

CREBBP, ARID1B, CUX2 and FOXP1 to autism and intellectual disability, risk ranking of DeepND indicates several new candidate pathogenic genes for ASD and ID located in CNV regions under analysis. DeepND ranking of some selected genes located in these CNV regions are shown in Table A.7 to A.14.

One of the genes located in 16p11.2 as given in Table A.7 and ranked in 1st percentile by DeepND with TADA $Q = 0.55$, SRCAP, is known to be included in Wnt signaling pathway [90] which is linked to ASD by its functional role in the dysregulation of excitatory/inhibitory neuronal activity [91]. With its multitask nature, DeepND suggests that SRCAP is also an important candidate for ID by ranking it in top 1st percentile for ID. Another gene in importin beta family, XPO6, suggest a new linkage between 16p11.2 and autism as it is related to Arts syndrome according to evidence from [92].

DeepND also highly ranks several genes with low prior risk and low posterior risk assigned by other methods within these CNV regions. *GABRG3* is one of such genes with low prior risk (TADA $P = 0.75$) located in 15p11-13 as shown in Table A.8. Despite being an E3-E4 level risk gene for ASD, it does not participate any relevant gene set. Although none of the other algorithms rank it in the top thousand (Krishnan *et al.*, 1297^{th}, DAWN 3257^{th}, DAMAGES score 1472^{nd}), DeepND ranks *GABRG3* within top percentile (250^{th}) for ASD.

Similar to *GABRG3*, *MICAL3*, a gene related to actin and Rab GTPase binding, ranked in as the 352^{nd} for ASD risk despite having TADA $Q = 0.77$ and not being in relevant risk gene groups. Krishnan *et al.* rank it 3165^{th} and DAMAGES score ranks it 1916^{th} as DAWN provides no ranking since it is not co-expressed with any other gene in networks of interest for DAWN. While disruption in Rab GTPase cycle gene could lead intellectual disability [93], there are no established ties between *MICAL3* and ASD or ID. However, MICAL family genes are related to cytoskeletal organization which recently has been deemed an important molecular function in ASD etiology [29, 23]. Another gene from the same family, *MICAL2*, is an ASD-risk gene and also ranked in the top percentile by DeepND since it utilizes multiple co-expression networks concurrently and is able to capture the link between *MICAL2* and *MICAL3* in 7 networks spanning late fetal

to adolescence in SHA and V1C-STC regions. This enables DeepND to predict it as a risk gene in a relevant CNV region (22q11, shown in Table A.11).

FOXO3B is located in 17p11.2 (Table A.13) which is one of the ID related CNV regions. While this gene does not even have a prior TADA score in Nguyen *et al.* [26], DeepND marks it as the 57^{th} risk gene. Deletion of FOXO transcription factors is shown to cause axonal degeneration and elevated mTORC1 activity [94] which was previously associated with several neurological and neurodevelopmental disorders including autism and epilepsy. Krishnan *et al.* ranks this gene in the fifth decile and DAWN cannot capture it due to lack of correlation in prefrontal cortex during mid-fetal period. Interestingly, all discussed genes, *GABRG3, MICAL3* and *FOXO3B* have similar RNA expression patterns in brain. The RNA expression patterns of these genes are shown in Figure 5.13. All genes are either most expressed or second most expressed in cerebral cortex and have significant expression levels in several other brain regions and major endocrine glands [2].

5.6 Evaluation of Novel Predictions

While it is always important to classify ground truth samples correctly in machine learning, it is also interesting to investigate false positives and false negatives since it can highlight the specificity and explain a different aspect of the problem, such as overlapping relation between many neurodevelopmental disorders. As a result, we also focus on the most confident predictions of DeepND which are clashing with the current consensus in our analyses.

One of E1 risk genes for ASD, *SLC9A9*, is a sodium/hydrogen exchange protein. Although it has a record of rare variants in ASD cohort, it is ranked in the last decile by DeepASD [60] as other algorithms rank it in the fifth decile at the lowest. The protein and RNA expressions of *SLC9A9* are enriched most in the spinal cord, as shown in Figure 5.14, which suggest that it is unlikely for this gene to be an ASD risk candidate. DeepND also ranks this gene in the last decile which

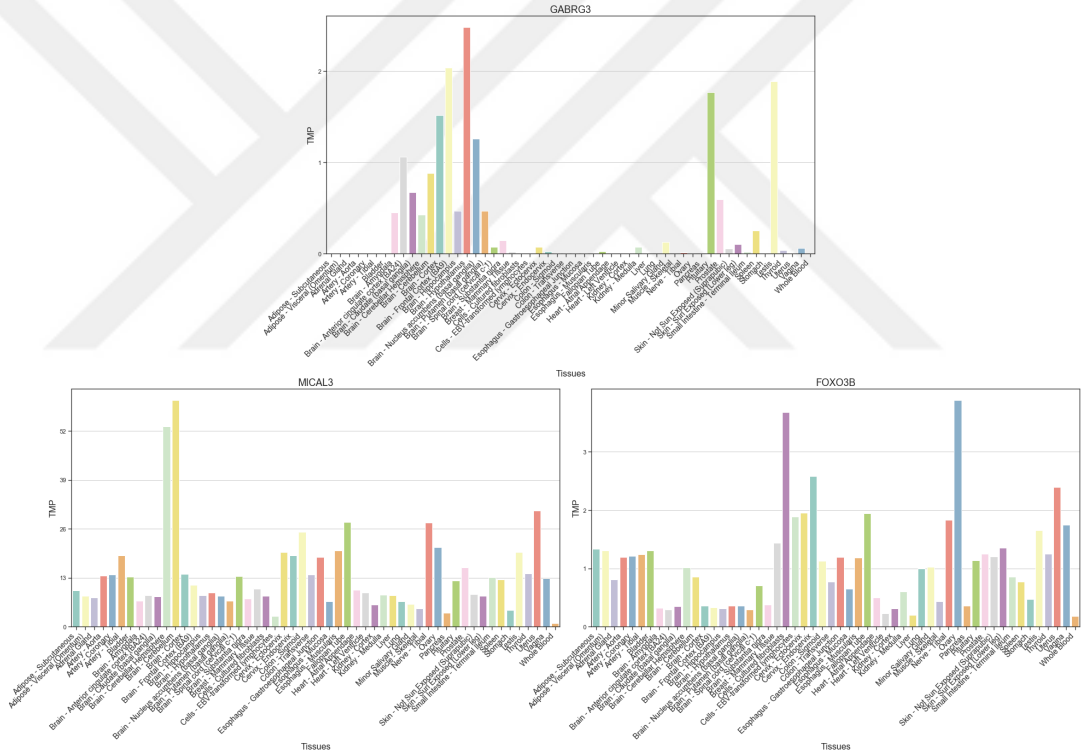

Figure 5.13: Median RNA expression patterns of *GABRG3*(top), *MICAL3* (bottom-left) and *FOXO3B*(bottom-right) across various tissues in the GTEx dataset [2] (transcripts per million).

shows that our method carries out the superior features of its precedent method which is to distinguish true risk genes from noise due to potential off-target labeling. Figure 5.14 also shows the RNA expression levels for *CACNA1H* which is another ASD E1 risk gene that DeepND assigns a low probability (0.246). It is mostly expressed in basal ganglia and pituitary gland in comparison to other brain regions which reduces the chances of it to be an ASD risk gene.

For ID, *TAT* can be shown as a similar example since it is ranked in the last decile with a very low probability of 0.074 despite being an ID E1 risk gene. It encodes a tyrosine aminotransferase and the high RNA expression levels of *TAT* in the liver, as shown in Figure 5.14, calls its involvement in ID in question.

DeepND also listed several genes with high probability of ID association that were not included in the positive gene set. For instance, *EP300* is one of the high ASD risk genes listed in [95] and ranked as highest risk gene for ID by DeepND. A similar scenario also holds for its closely related paralog, *CREBBP*, which is ranked in the first decile of risk prediction for ID and ASD by DeepND. Although they are included in the negative gold set, these genes are major lysine acetyltransferases in metazoans and their RNA expression levels are either most expressed or second most expressed in cerebellum which is previously linked to ASD as developmental damage to the cerebellum is a common occurrence in ASD [96, 97, 98].

Finally, DeepND ranks *GIGFY2* which is a negative ground truth gene, and its paralog *GIGFY1* in top quartile for ASD. *GIGYF2* is located in 2q37.1, which is linked to Parkinson disease type 11 [99] and it is argued to play a role in regulation of tyrosine. Likely gene disrupting mutations in *GIGYF2* has linked it to ASD in a Chinese cohort [100]. *GIGFY1* is also associated with Tourette syndrome (TS) which frequently co-occurs with other neurodevelopmental disorders, including ASD, as $3-20\%$ of children with TS have co-morbid ASD [101].

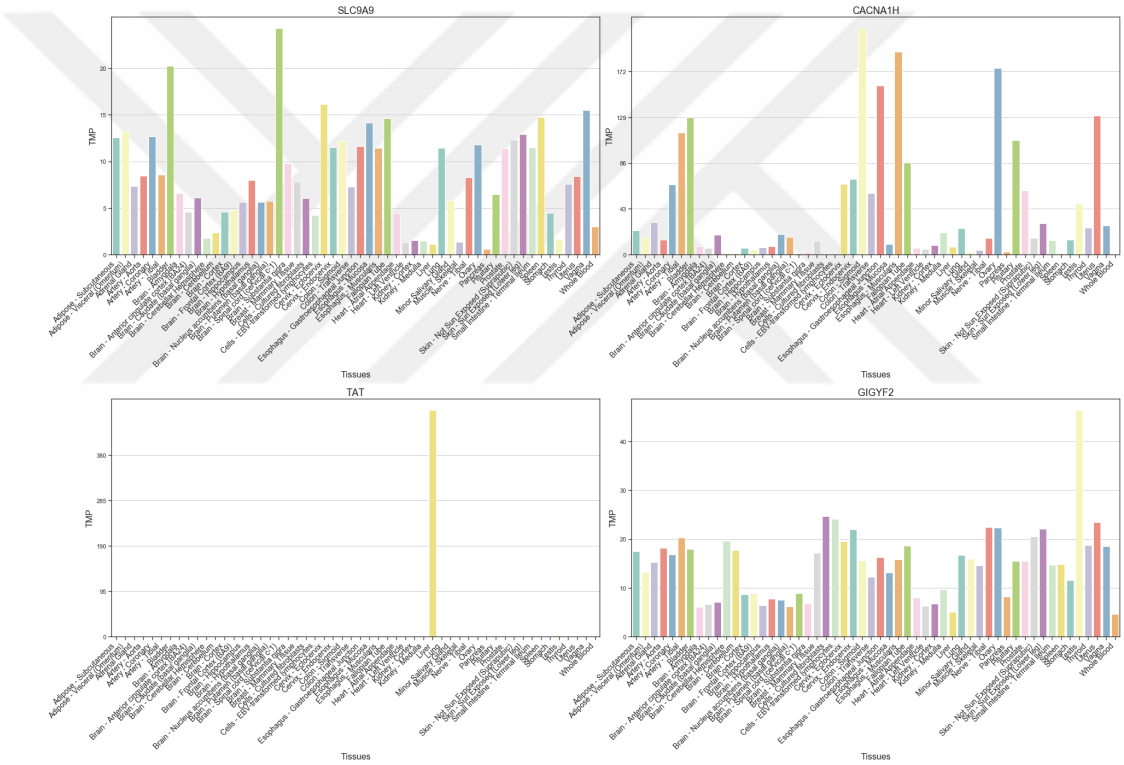

Figure 5.14: Median RNA expression patterns for confident novel predictions of DeepND across various tissues in the GTEx dataset [2] (transcripts per million): *SLC9A9* (top-left), *CACNA1H* (top-right), *TAT* (bottom-left) and *GIGYF2* (bottom-right).

Chapter 6

Conclusion

In this thesis, we propose *DeepND*, a cross-disorder gene discovery algorithm using deep learning principles to analyze comorbid neurodevelopmental disorders simultaneously. Our algorithm inputs multiple gene co-expression networks and explicitly learns the shared and disorder-specific genetic features using multitask learning. Associations between genes are processed by graph convolution layers using gene co-expression networks and the model self-learns critical neurodevelopmental time windows and brain regions for each disorder etiology through its disorder specific mixture-of-experts modules.

We apply our method to two highly comorbid neurodevelopmental disorders, ASD and ID. The model is tested using 5-fold cross validation with disorder specific trigger flags to achieve proper and balanced learning though out the layers. After each training session, the performance reported on the test fold and the final gene risk predictions for both disorders are given as the averages of results that DeepND produces when the gene in question is present in the test fold.

We compare the risk gene rankings for ASD and ID with other methods from the literature and show that our methods has a higher prediction power. DeepND performs better than all single task settings and improves the best performing

single task method, DeepASD, by a maximum of 1.4% in terms of AUC and by 5% AUPR for both disorders.DeepND assigns mediodorsal thalamus and cerebellar cortex brain region and infancy to childhood period (MDCBC 8-10 and MDCBC 9-11 networks) as the highest neurodevelopmental risk window for both disorders.

We show that top decile genes predicted by DeepND for both disorders are highly enriched in gene sets that are previously associated with ASD and ID which indicates the wellness of the ranking given by our model. We also show the connection among top percentile ASD and ID genes and observe that the strongest connection is between ASD-only and ASD & ID gene sets. We also investigate the highest ranking genes in mental retardation and ASD associated CNV regions. Additionally, we focus on and explain the most confident predictions of DeepND which are clashing with the current consensus. As the multi-task neural network architecture of DeepND can easily be generalized to other disorders, we show that the predicted gene-risk rankings by DeepND is applicable for different types of biological analyses which also makes it applicable for inter-disciplinary studies.

Bibliography

[1] O. Basha, R. Shpringer, C. M. Argov, and E. Yeger-Lotem, "The differentialnet database of differential protein–protein interactions in human tissues," *Nucleic acids research*, vol. 46, no. D1, pp. D522–D526, 2018.

[2] "The gtex portal," Agu 2019. Online; accessed 20 February 2020.

[3] A. J. Willsey, S. J. Sanders, M. Li, S. Dong, A. T. Tebbenkamp, R. Muhle, S. K. Reilly, L. Lin, S. Fertuzinhos, J. A. Miller, M. Murtha, C. Bichsel, W. Niu, J. Cotney, A. G. Ercan-Sencicek, J. Gockley, A. R. Gupta, W. Han, X. He, E. J. Hoffman, L. Klei, J. Lei, W. Liu, L. Liu, C. Lu, X. Xu, Y. Zhu, S. M. Mane, E. S. Lein, and L. Wei, "Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism," *Cell*, vol. 155, 2013.

[4] K. J. Mitchell, "The genetics of neurodevelopmental disease," *Current opinion in neurobiology*, vol. 21, no. 1, pp. 197–203, 2011.

[5] J. Roeper, "Closing gaps in brain disease—from overlapping genetic architecture to common motifs of synapse dysfunction," *Current opinion in neurobiology*, vol. 48, pp. 45–51, 2018.

[6] P. H. Lee, V. Anttila, H. Won, Y.-C. A. Feng, J. Rosenthal, Z. Zhu, E. M. Tucker-Drob, M. G. Nivard, A. D. Grotzinger, D. Posthuma, *et al.*, "Genome wide meta-analysis identifies genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders.," *bioRxiv*, p. 528117, 2019.

[7] A. Moreno-De-Luca, S. M. Myers, T. D. Challman, D. Moreno-De-Luca, D. W. Evans, and D. H. Ledbetter, "Developmental brain dysfunction: revival and expansion of old concepts based on new genetic evidence," *The Lancet Neurology*, vol. 12, no. 4, pp. 406–414, 2013.

[8] J. De Ligt, M. H. Willemsen, B. W. Van Bon, T. Kleefstra, H. G. Yntema, T. Kroes, A. T. Vulto-van Silfhout, D. A. Koolen, P. De Vries, C. Gilissen, *et al.*, "Diagnostic exome sequencing in persons with severe intellectual disability," *New England Journal of Medicine*, vol. 367, no. 20, pp. 1921–1929, 2012.

[9] H. V. Firth, C. F. Wright, and D. study, "The deciphering developmental disorders (ddd) study," *Developmental Medicine & Child Neurology*, vol. 53, no. 8, pp. 702–703, 2011.

[10] C. Gilissen, J. Y. Hehir-Kwa, D. T. Thung, M. van de Vorst, B. W. van Bon, M. H. Willemsen, M. Kwint, I. M. Janssen, A. Hoischen, A. Schenck, *et al.*, "Genome sequencing identifies major causes of severe intellectual disability," *Nature*, vol. 511, no. 7509, p. 344, 2014.

[11] A. Rauch, D. Wieczorek, E. Graf, T. Wieland, S. Endele, T. Schwarzmayr, B. Albrecht, D. Bartholdi, J. Beygo, N. Di Donato, *et al.*, "Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study," *The Lancet*, vol. 380, no. 9854, pp. 1674–1682, 2012.

[12] S. Srivastava, J. S. Cohen, H. Vernon, K. Barañano, R. McClellan, L. Jamal, S. Naidu, and A. Fatemi, "Clinical whole exome sequencing in child neurology practice," *Annals of neurology*, vol. 76, no. 4, pp. 473–483, 2014.

[13] L. E. Vissers, J. de Ligt, C. Gilissen, I. Janssen, M. Steehouwer, P. de Vries, B. van Lier, P. Arts, N. Wieskamp, M. del Rosario, *et al.*, "A de novo paradigm for mental retardation," *Nature genetics*, vol. 42, no. 12, p. 1109, 2010.

[14] X. Zhu, S. Petrovski, P. Xie, E. K. Ruzzo, Y.-F. Lu, K. M. McSweeney, B. Ben-Zeev, A. Nissenkorn, Y. Anikster, D. Oz-Levi, *et al.*, "Whole-exome

sequencing in undiagnosed genetic diseases: interpreting 119 trios," *Genetics in Medicine*, vol. 17, no. 10, p. 774, 2015.

[15] J. Baio, L. Wiggins, D. L. Christensen, M. J. Maenner, J. Daniels, Z. Warren, M. Kurzius-Spencer, W. Zahorodny, C. R. Rosenberg, T. White, *et al.*, "Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, united states, 2014," *MMWR Surveillance Summaries*, vol. 67, no. 6, p. 1, 2018.

[16] J. Li, T. Cai, Y. Jiang, H. Chen, X. He, C. Chen, X. Li, Q. Shao, X. Ran, Z. Li, *et al.*, "Genes with de novo mutations are shared by four neuropsychiatric disorders discovered from npdenovo database," *Molecular psychiatry*, vol. 21, no. 2, p. 290, 2016.

[17] E. B. Robinson, K. E. Samocha, J. A. Kosmicki, L. McGrath, B. M. Neale, R. H. Perlis, and M. J. Daly, "Autism spectrum disorder severity reflects the average contribution of de novo and familial influences," *Proceedings of the National Academy of Sciences*, vol. 111, no. 42, pp. 15161–15165, 2014.

[18] X. He, S. J. Sanders, L. Liu, S. D. Rubeis, E. T. Lim, J. S. Sutcliffe, G. D. Schellenberg, R. A. Gibbs, M. J. Daly, J. D. Buxbaum, and et al., "Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes," *PLoS Genetics*, vol. 9, no. 8, 2013.

[19] T.-H. Nguyen, A. Dobbyn, R. C. Brown, B. P. Riley, J. D. Buxbaum, D. Pinto, S. M. Purcell, P. F. Sullivan, X. He, and E. A. Stahl, "mtada is a framework for identifying risk genes from de novo mutations in multiple traits," *Nature Communications*, vol. 11, no. 1, pp. 1–12, 2020.

[20] A. J. Gonzalez-Mantilla, A. Moreno-De-Luca, D. H. Ledbetter, and C. L. Martin, "A cross-disorder method to identify novel candidate genes for developmental brain disorders," *JAMA psychiatry*, vol. 73, no. 3, pp. 275–283, 2016.

[21] M. J. Gandal, J. R. Haney, N. N. Parikshak, V. Leppa, G. Ramaswami, C. Hartl, A. J. Schork, V. Appadurai, A. Buil, T. M. Werge, *et al.*, "Shared

molecular neuropathology across major psychiatric disorders parallels polygenic overlap," *Science*, vol. 359, no. 6376, pp. 693–697, 2018.

[22] F. Hormozdiari, O. Penn, E. Borenstein, and E. E. Eichler, "The discovery of integrated gene networks for autism and related disorders," *Genome Research*, vol. 25, pp. 142–154, May 2014.

[23] U. Norman and A. E. Cicek, "St-steiner: a spatio-temporal gene discovery algorithm," *Bioinformatics*, vol. 35, no. 18, pp. 3433–3440, 2019.

[24] S. R. Gilman, I. Iossifov, D. Levy, M. Ronemus, M. Wigler, and D. Vitkup, "Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses," *Neuron*, vol. 70, no. 5, pp. 898–907, 2011.

[25] S. R. Gilman, J. Chang, B. Xu, T. S. Bawa, J. A. Gogos, M. Karayiorgou, and D. Vitkup, "Diverse types of genetic variation converge on functional gene networks involved in schizophrenia," *Nature neuroscience*, vol. 15, no. 12, pp. 1723–1728, 2012.

[26] H. T. Nguyen, J. Bryois, A. Kim, A. Dobbyn, L. M. Huckins, A. B. Munoz-Manchado, D. M. Ruderfer, G. Genovese, M. Fromer, X. Xu, *et al.*, "Integrated bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders," *Genome medicine*, vol. 9, no. 1, p. 114, 2017.

[27] L. Brueggeman, T. Koomar, and J. J. Michaelson, "Forecasting risk gene discovery in autism with machine learning and genome-scale data," *Scientific Reports*, vol. 10, no. 1, pp. 1–11, 2020.

[28] M. Duda, H. Zhang, H.-D. Li, D. P. Wall, M. Burmeister, and Y. Guan, "Brain-specific functional relationship networks inform autism spectrum disorder gene prediction," *Translational psychiatry*, vol. 8, no. 1, pp. 1–9, 2018.

[29] F. K. Satterstrom, J. A. Kosmicki, J. Wang, M. S. Breen, S. De Rubeis, J.-Y. An, M. Peng, R. L. Collins, J. Grove, L. Klei, *et al.*, "Large-scale exome

sequencing study implicates both developmental and functional changes in the neurobiology of autism," 2019.

[30] J. Buxbaum, E. Cicek, B. Devlin, L. Klei, K. Roeder, and S. De Rubeis, "Combining autism and intellectual disability exome data implicates disruption of neocortical development in both disorders," *European Neuropsychopharmacology*, vol. 27, p. S437, 2017.

[31] L. Liu, J. Lei, S. J. Sanders, A. J. Willsey, Y. Kou, A. E. Cicek, L. Klei, C. Lu, X. He, M. Li, R. A. Muhle, A. Ma'ayan, J. P. Noonan, N. Šestan, K. A. McFadden, M. W. State, J. D. Buxbaum, B. Devlin, and K. Roeder, "Dawn: a framework to identify autism genes and subnetworks using gene expression and genetics," *Molecular Autism*, vol. 5, p. 22, Mar 2014.

[32] A. Krishnan, R. Zhang, V. Yao, C. L. Theesfeld, A. K. Wong, A. Tadych, N. Volfovsky, A. Packer, A. Lash, and O. G. Troyanskaya, "Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder," *Nature neuroscience*, vol. 19, no. 11, pp. 1454–1462, 2016.

[33] "Talking glossary of genetic terms." https://www.genome.gov/genetics-glossary, January 10, 2020. Online; Accessed 08 August 2020.

[34] T. H. Shaikh, "Copy number variation disorders," *Current genetic medicine reports*, vol. 5, no. 4, pp. 183–190, 2017.

[35] F. Zhang, W. Gu, M. E. Hurles, and J. R. Lupski, "Copy number variation in human health, disease, and evolution," *Annual review of genomics and human genetics*, vol. 10, pp. 451–481, 2009.

[36] Z. Zhang, M. A. Miteva, L. Wang, and E. Alexov, "Analyzing effects of naturally occurring missense mutations," *Computational and mathematical methods in medicine*, vol. 2012, 2012.

[37] J. A. Veltman and H. G. Brunner, "De novo mutations in human genetic disease," *Nature Reviews Genetics*, vol. 13, no. 8, pp. 565–575, 2012.

[38] P. Lemay, M.-C. Guyot, É. Tremblay, A. Dionne-Laporte, D. Spiegelman, É. Henrion, O. Diallo, P. De Marco, E. Merello, C. Massicotte, *et al.*, "Loss-of-function de novo mutations play an important role in severe human neural tube defects," *Journal of Medical Genetics*, vol. 52, no. 7, pp. 493–497, 2015.

[39] M. G. Blanchard, M. H. Willemsen, J. B. Walker, S. D. Dib-Hajj, S. G. Waxman, M. C. Jongmans, T. Kleefstra, B. P. van de Warrenburg, P. Praamstra, J. Nicolai, *et al.*, "De novo gain-of-function and loss-of-function mutations of scn8a in patients with intellectual disabilities and epilepsy," *Journal of medical genetics*, vol. 52, no. 5, pp. 330–337, 2015.

[40] C. T. Myers, J. M. McMahon, A. L. Schneider, S. Petrovski, A. S. Allen, G. L. Carvill, M. Zemel, J. E. Saykally, A. J. LaCroix, E. L. Heinzen, *et al.*, "De novo mutations in slc1a2 and cacna1a are important causes of epileptic encephalopathies," *The American Journal of Human Genetics*, vol. 99, no. 2, pp. 287–298, 2016.

[41] X. He, S. J. Sanders, L. Liu, S. De Rubeis, E. T. Lim, J. S. Sutcliffe, G. D. Schellenberg, R. A. Gibbs, M. J. Daly, J. D. Buxbaum, *et al.*, "Integrated model of de novo and inherited genetic variants yields greater power to identify risk genes," *PLoS Genet*, vol. 9, no. 8, p. e1003671, 2013.

[42] M. T. Weirauch, *Gene Coexpression Networks for the Analysis of DNA Microarray Data*, ch. 11, pp. 215–250. John Wiley Sons, Ltd, 2011.

[43] B. Topp, K. Promislow, G. Devries, R. M. Miura, and D. T. Finegood, "A model of b-cell mass, insulin, and glucose kinetics: pathways to diabetes," *Journal of theoretical biology*, vol. 206, no. 4, p. 605, 2000.

[44] J. Downward, "Targeting ras signalling pathways in cancer therapy," *Nature Reviews Cancer*, vol. 3, no. 1, pp. 11–22, 2003.

[45] J. Taipale and P. A. Beachy, "The hedgehog and wnt signalling pathways in cancer," *nature*, vol. 411, no. 6835, p. 349, 2001.

[46] H. O. Kalkman, "A review of the evidence for the canonical wnt pathway in autism spectrum disorders," *Molecular autism*, vol. 3, no. 1, p. 10, 2012.

[47] D. Pinto, E. Delaby, D. Merico, M. Barbosa, A. Merikangas, L. Klei, B. Thiruvahindrapuram, X. Xu, R. Ziman, Z. Wang, *et al.*, "Convergence of genes and cellular pathways dysregulated in autism spectrum disorders," *The American Journal of Human Genetics*, vol. 94, no. 5, pp. 677–694, 2014.

[48] J. Cotney, R. A. Muhle, S. J. Sanders, L. Liu, A. J. Willsey, W. Niu, W. Liu, L. Klei, J. Lei, J. Yin, *et al.*, "The autism-associated chromatin modifier chd8 regulates other autism risk genes during human neurodevelopment," *Nature communications*, vol. 6, p. 6404, 2015.

[49] J. C. Darnell, S. J. Van Driesche, C. Zhang, K. Y. S. Hung, A. Mele, C. E. Fraser, E. F. Stone, C. Chen, J. J. Fak, S. W. Chi, *et al.*, "Fmrp stalls ribosomal translocation on mrnas linked to synaptic function and autism," *Cell*, vol. 146, no. 2, pp. 247–261, 2011.

[50] M. Ascano, N. Mukherjee, P. Bandaru, J. B. Miller, J. D. Nusbaum, D. L. Corcoran, C. Langlois, M. Munschauer, S. Dewell, M. Hafner, *et al.*, "Fmrp targets distinct mrna sequence elements to regulate protein expression," *Nature*, vol. 492, no. 7429, pp. 382–386, 2012.

[51] S. De Rubeis, X. He, A. P. Goldberg, C. S. Poultney, K. Samocha, A. E. Cicek, Y. Kou, L. Liu, M. Fromer, S. Walker, *et al.*, "Synaptic, transcriptional and chromatin genes disrupted in autism," *Nature*, vol. 515, no. 7526, pp. 209–215, 2014.

[52] I. Voineagu, X. Wang, P. Johnston, J. K. Lowe, Y. Tian, S. Horvath, J. Mill, R. M. Cantor, B. J. Blencowe, and D. H. Geschwind, "Transcriptomic analysis of autistic brain reveals convergent molecular pathology," *Nature*, vol. 474, no. 7351, pp. 380–384, 2011.

[53] S. M. Weyn-Vanhentenryck, A. Mele, Q. Yan, S. Sun, N. Farny, Z. Zhang, C. Xue, M. Herre, P. A. Silver, M. Q. Zhang, *et al.*, "Hits-clip and integrative modeling define the rbfox splicing-regulatory network linked to brain development and autism," *Cell reports*, vol. 6, no. 6, pp. 1139–1152, 2014.

[54] I. F. King, C. N. Yandava, A. M. Mabb, J. S. Hsiao, H.-S. Huang, B. L. Pearson, J. M. Calabrese, J. Starmer, J. S. Parker, T. Magnuson, *et al.*, "Topoisomerases facilitate transcription of long genes linked to autism," *Nature*, vol. 501, no. 7465, pp. 58–62, 2013.

[55] À. Bayés, L. N. Van De Lagemaat, M. O. Collins, M. D. Croning, I. R. Whittle, J. S. Choudhary, and S. G. Grant, "Characterization of the proteome, diseases and evolution of the human postsynaptic density," *Nature neuroscience*, vol. 14, no. 1, pp. 19–21, 2011.

[56] H. Y. Zoghbi and M. F. Bear, "Synaptic dysfunction in neurodevelopmental disorders associated with autism and intellectual disabilities," *Cold Spring Harbor perspectives in biology*, vol. 4, no. 3, p. a009886, 2012.

[57] S. Iwase, N. G. Bérubé, Z. Zhou, N. N. Kasri, E. Battaglioli, M. Scandaglia, and A. Barco, "Epigenetic etiology of intellectual disability," *Journal of Neuroscience*, vol. 37, no. 45, pp. 10773–10782, 2017.

[58] H. J. Kang, Y. I. Kawasawa, F. Cheng, Y. Zhu, X. Xu, M. Li, A. M. Sousa, M. Pletikos, K. A. Meyer, G. Sedmak, T. Guennel, Y. Shin, M. B. Johnson, Z. Krsnik, S. Mayer, S. Fertuzinhos, S. Umlauf, S. N. Lisgo, A. Vortmeyer, D. R. Weinberger, S. Mane, T. M. Hyde, A. Huttner, M. Reimers, J. E. Kleinman, and N. Sestan, "Spatio-temporal transcriptome of the human brain," *Nature*, vol. 478, 2011.

[59] C. Zhang and Y. Shen, "A cell type-specific expression signature predicts haploinsufficient autism-susceptibility genes," *Human mutation*, vol. 38, no. 2, pp. 204–215, 2017.

[60] O. Karakahya, "Predicting informative spatio-temporal neurodeveleopmental windows and gene risk for autism spectrum disorder," Master's thesis, Bilkent University, Ankara, Turkey, 2020.

[61] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[62] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," *arXiv preprint arXiv:1312.6203*, 2013.

[63] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams, "Convolutional networks on graphs for learning molecular fingerprints," in *Advances in neural information processing systems*, pp. 2224–2232, 2015.

[64] M. Henaff, J. Bruna, and Y. LeCun, "Deep convolutional networks on graph-structured data," *arXiv preprint arXiv:1506.05163*, 2015.

[65] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[66] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Advances in neural information processing systems*, pp. 3844–3852, 2016.

[67] S. Masoudnia and R. Ebrahimpour, "Mixture of experts: a literature survey," *Artificial Intelligence Review*, vol. 42, no. 2, pp. 275–293, 2014.

[68] B. McCann, N. S. Keskar, C. Xiong, and R. Socher, "The natural language decathlon: Multitask learning as question answering," *arXiv preprint arXiv:1806.08730*, 2018.

[69] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3712–3722, 2018.

[70] D. Grozeva, K. Carss, O. Spasic-Boskovic, M.-I. Tejada, J. Gecz, M. Shaw, M. Corbett, E. Haan, E. Thompson, K. Friend, *et al.*, "Targeted next-generation sequencing analysis of 1,000 individuals with intellectual disability," *Human mutation*, vol. 36, no. 12, pp. 1197–1204, 2015.

[71] L. E. Vissers, C. Gilissen, and J. A. Veltman, "Genetic studies in intellectual disability and related disorders," *Nature Reviews Genetics*, vol. 17, no. 1, p. 9, 2016.

[72] J. K. Inlow and L. L. Restifo, "Molecular and comparative genetics of mental retardation," *Genetics*, vol. 166, no. 2, pp. 835–881, 2004.

[73] P. Chiurazzi and F. Pirozzi, "Advances in understanding–genetic basis of intellectual disability," *F1000Research*, vol. 5, 2016.

[74] S. M. Sunkin, L. Ng, C. Lau, T. Dolbeare, T. L. Gilbert, C. L. Thompson, M. Hawrylycz, and C. Dang, "Allen brain atlas: an integrated spatio-temporal portal for exploring the central nervous system," *Nucleic acids research*, vol. 41, no. D1, pp. D996–D1008, 2012.

[75] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, 2010.

[76] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[77] R. M. Meredith, J. Dawitz, and I. Kramvis, "Sensitive time-windows for susceptibility in neurodevelopmental disorders," *Trends in neurosciences*, vol. 35, no. 6, pp. 335–344, 2012.

[78] T. Kroon, M. Sierksma, and R. M. Meredith, "Investigating mechanisms underlying neurodevelopmental phenotypes of autistic and intellectual disability disorders: a perspective," *Frontiers in systems neuroscience*, vol. 7, p. 75, 2013.

[79] E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, and A. Ma'ayan, "Enrichr: interactive and collaborative html5 gene list enrichment analysis tool," *BMC bioinformatics*, vol. 14, no. 1, p. 128, 2013.

[80] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G.

McDermott, C. D. Monteiro, G. W. Gundersen, and A. Ma'ayan, "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update," *Nucleic Acids Research*, vol. 44, no. W1, pp. W90–W97, 2016.

[81] A. Lachmann, H. Xu, J. Krishnan, S. I. Berger, A. R. Mazloom, and A. Ma'ayan, "Chea: transcription factor regulation inferred from integrating genome-wide chip-x experiments," *Bioinformatics*, vol. 26, no. 19, pp. 2438–2444, 2010.

[82] C. Ottolenghi, R. Veitia, L. Quintana-Murci, D. Torchard, L. Scapoli, N. Souleyreau-Therville, J. Beckmann, M. Fellous, and K. McElreavey, "The region on 9p associated with 46, xy sex reversal contains several transcripts expressed in the urogenital system and a novel doublesex-related domain," *Genomics*, vol. 64, no. 2, pp. 170–178, 2000.

[83] G. Vinci, S. Chantot-Bastaraud, B. El Houate, S. Lortat-Jacob, R. Brauner, and K. McElreavey, "Association of deletion 9p, 46, xy gonadal dysgenesis and autistic spectrum disorder," *Molecular human reproduction*, vol. 13, no. 9, pp. 685–689, 2007.

[84] P. Szatmari, A. D. Paterson, L. Zwaigenbaum, W. Roberts, J. Brian, X.-Q. Liu, J. B. Vincent, J. L. Skaug, A. P. Thompson, L. Senman, *et al.*, "Mapping autism risk loci using genetic linkage and chromosomal rearrangements," *Nature genetics*, vol. 39, no. 3, p. 319, 2007.

[85] Y. Yang, C. Wang, F. Wang, L. Zhu, H. Liu, and X. He, "Novel chromosomal translocation t (11; 9)(p15; p23) involving deletion and duplication of 9p in a girl associated with autism and mental retardation," *Gene*, vol. 502, no. 2, pp. 154–158, 2012.

[86] G. Zhou, O. Soufan, J. Ewald, R. E. Hancock, N. Basu, and J. Xia, "Networkanalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis," *Nucleic acids research*, vol. 47, no. W1, pp. W234–W241, 2019.

[87] I. Iossifov, B. J. O?Roak, S. J. Sanders, M. Ronemus, N. Krumm, D. Levy, H. A. Stessman, K. T. Witherspoon, L. Vives, K. E. Patterson, *et al.*, "The

contribution of de novo coding mutations to autism spectrum disorder," *Nature*, vol. 515, no. 7526, pp. 216–221, 2014.

[88] J. Halvardson, J. J. Zhao, A. Zaghlool, C. Wentzel, P. Georgii-Hemming, E. Månsson, H. Ederth Sävmarker, G. Brandberg, C. Soussi Zander, A.-C. Thuresson, and L. Feuk, "Mutations in hecw2 are associated with intellectual disability and epilepsy," *Journal of Medical Genetics*, vol. 53, no. 10, pp. 697–704, 2016.

[89] H. C. Mefford, G. M. Cooper, T. Zerr, J. D. Smith, C. Baker, N. Shafer, E. C. Thorland, C. Skinner, C. E. Schwartz, D. A. Nickerson, *et al.*, "A method for rapid, targeted cnv genotyping identifies rare variants associated with neurocognitive disease," *Genome research*, vol. 19, no. 9, pp. 1579–1585, 2009.

[90] P. D. Thomas, A. Kejariwal, M. J. Campbell, H. Mi, K. Diemer, N. Guo, I. Ladunga, B. Ulitsky-Lazareva, A. Muruganujan, S. Rabkin, *et al.*, "Panther: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification," *Nucleic acids research*, vol. 31, no. 1, pp. 334–341, 2003.

[91] M. A. Medina, V. M. Andrade, M. O. Caracci, M. E. Avila, D. A. Verdugo, M. F. Vargas, G. D. Ugarte, A. E. Reyes, C. Opazo, and G. V. De Ferrari, "Wnt/β-catenin signaling stimulates the expression and synaptic clustering of the autism-associated neuroligin 3 gene," *Translational psychiatry*, vol. 8, no. 1, pp. 1–12, 2018.

[92] S. Pletscher-Frankild, A. Pallejà, K. Tsafou, J. X. Binder, and L. J. Jensen, "Diseases: Text mining and data integration of disease–gene associations," *Methods*, vol. 74, pp. 83–89, 2015.

[93] M. C. Seabra, E. H. Mules, and A. N. Hume, "Rab gtpases, intracellular traffic and disease," *Trends in molecular medicine*, vol. 8, no. 1, pp. 23–30, 2002.

[94] I. Hwang, H. Oh, E. Santo, D.-Y. Kim, J. W. Chen, R. T. Bronson, J. W. Locasale, Y. Na, J. Lee, S. Reed, *et al.*, "Foxo protects against

age-progressive axonal degeneration," *Aging cell*, vol. 17, no. 1, p. e12701, 2018.

[95] B. S. Abrahams, D. E. Arking, D. B. Campbell, H. C. Mefford, E. M. Morrow, L. A. Weiss, I. Menashe, T. Wadkins, S. Banerjee-Basu, and A. Packer, "Sfari gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (asds)," *Molecular autism*, vol. 4, no. 1, p. 36, 2013.

[96] A. Bailey, P. Luthert, A. Dean, B. Harding, I. Janota, M. Montgomery, M. Rutter, and P. Lantos, "A clinicopathological study of autism.," *Brain: a journal of neurology*, vol. 121, no. 5, pp. 889–905, 1998.

[97] M. L. Bauman and T. L. Kemper, "Neuroanatomic observations of the brain in autism: a review and future directions," *International journal of developmental neuroscience*, vol. 23, no. 2-3, pp. 183–187, 2005.

[98] T. D. Rogers, E. McKimm, P. E. Dickson, D. Goldowitz, C. D. Blaha, and G. Mittleman, "Is autism a disease of the cerebellum? an integration of clinical and pre-clinical research," *Frontiers in systems neuroscience*, vol. 7, p. 15, 2013.

[99] N. Pankratz, W. C. Nichols, S. K. Uniacke, C. Halter, A. Rudolph, C. Shults, P. M. Conneally, T. Foroud, P. S. Group, *et al.*, "Genome screen to identify susceptibility genes for parkinson disease in a sample without parkin mutations," *The American Journal of Human Genetics*, vol. 71, no. 1, pp. 124–135, 2002.

[100] T. Wang, H. Guo, B. Xiong, H. A. Stessman, H. Wu, B. P. Coe, T. N. Turner, Y. Liu, W. Zhao, K. Hoekzema, *et al.*, "De novo genic mutations among a chinese autism spectrum disorder cohort," *Nature communications*, vol. 7, p. 13316, 2016.

[101] E. Cravedi, E. Deniau, M. Giannitelli, J. Xavier, A. Hartmann, and D. Cohen, "Tourette syndrome and other neurodevelopmental disorders: a comprehensive review," *Child and adolescent psychiatry and mental health*, vol. 11, no. 1, p. 59, 2017.

Appendix A

Supplementary Tables

Table A.1: Neurodevelopmental time periods proposed by Willsey *et al.* [3]. Sliding window approach is used to create 13 temporal windows which is presented in Figure 5.4(A).

Time Period	Age interval
1	Embryonic (4-8 pcw)
2	Early fetal (8-10 pcw)
3	Early fetal 2 (10-13 pcw)
4	Early mid-fetal (13-16 pcw)
5	Early mid-fetal 2 (16-19 pcw)
6	Late mid-fetal (19-24 pcw)
7	Late fetal (24-38 pcw)
8	Neonatal & early infacy (0 - 6 months)
9	Late infancy (6 - 12 months)
10	Early childhood (1 - 6 years)
11	Middle and late childhood (6 - 12 years)
12	Adolescence (12 - 20 years)
13	Young adulthood (20 - 40 years)
14	Middle adulthood (40 - 60 years)
15	Late adulthood (60+ years)

Table A.2: Brain region clusters formed by using hierarchical clustering based on transcriptional similarity of these brain regions (Willsey et al., 2013). The cross section of brain which shows these regions is presented in Figure 5.4(B).

Cluster Name	Brain Regions in Cluster
PFC-MSC	M1C (Primary motor cortex), S1C (Primary somatosensory cortex), VFC (ventral prefrontal cortex), MFC (medial prefrontal cortex), DFC (dorsal prefrontal cortex), OFC (orbital prefrontal cortex)
MDCBC	MD (mediodorsal nucleus of the thalamus), CBC (cerebellar cortex)
V1C-STC	V1C (primary visual cortex), ITC (inferior temporal cortex), IPC (posterior inferior parietal cortex), A1C (primary auditory cortex), STC (superior temporal cortex)
SHA	S (striatum), H (hippocampus), A (amygdala)

Table A.3: The positive ground truth genes for ASD

Gene	**Evidence**	**Gene**	**Evidence**	**Gene**	**Evidence**	**Gene**	**Evidence**	**Gene**	**Evidence**
NRXN1	E1	RELN	E1	SYNGAP1	E1	RIMS1	E2	AVPR1A	E2
KMT2A	E1	GRIN2B	E1	SHANK3	E1	KDM6B	E2	ATP10A	E2
ANK2	E1	CUL3	E1	CHD2	E1	PHF2	E2	GPHN	E2
SHANK2	E1	SETD5	E1	MET	E1	PRKCB	E2	CTTNBP2	E2
BCL11A	E1	ANKRD11	E1	SLC9A9	E1	KMT2E	E2	FOXP2	E2
CTNND2	E1	KDM5B	E1	KATNAL2	E1	DISC1	E2	PON1	E2
GABRB3	E1	PTEN	E1	MAGEL2	E1	SEMA5A	E2	LMX1B	E2
MECP2	E1	CHD8	E1	CACNA1H	E1	ASTN2	E2	MYO9B	E2
ADNP	E1	ASH1L	E1	POGZ	E1	SETBP1	E2	PAX5	E2
CNTNAP2	E1	KMT2C	E1	ASXL3	E1	NR3C2	E2	CEP41	E2
MED13L	E1	NLGN4X	E1	WAC	E2	CHRNA7	E2	NCKAP1	E2
SUV420H1	E1	PTCHD1	E1	TNRC6B	E2	TRIP12	E2	BCKDK	E2
MYT1L	E1	CNTN4	E1	RBFOX1	E2	SPAST	E2	DIP2A	E2
DYRK1A	E1	CACNA2D3	E1	GRIK2	E2	HMGN1	E2	OXTR	E2
DSCAM	E1	NLGN3	E1	WDFY3	E2	CDC42BPB	E2	OPHN1	E2
SCN2A	E1	DEAF1	E1	AUTS2	E2	TRPC6	E2	EHMT1	E2
FOXP1	E1	ARID1B	E1	CACNA1D	E2	MBD5	E2	SLC6A3	E2
TBR1	E1	GRIP1	E1	ZMYND11	E2	CNTNAP4	E2	SHANK1	E2

Gene	Evidence	Gene	Evidence	Gene	Evidence	Gene	Evidence	Gene	Evidence
CACNB2	E2	MACROD2	E2	SNAP25	E3	NTRK3	E3	GABBR2	E3
ATP2B2	E2	CTNNB1	E2	AFF2	E3	DPP6	E3	GRIA2	E3
CC2D1A	E2	TCF7L2	E2	SATB2	E3	DOCK3	E3	CDK14	E3
ITGB3	E2	CNR1	E3	PIK3R1	E3	DLG2	E3	CASK	E3
NAA15	E2	NTRK2	E3	GABRA5	E3	SLC1A1	E3	PAX6	E3
APH1A	E2	DCX	E3	GABRA1	E3	UBE3A	E3	VLDLR	E3
TMLHE	E2	CAMK2B	E3	NPAS2	E3	NF1	E3	TBL1XR1	E3
EFR3A	E2	HTR2C	E3	KCNMA1	E3	SORCS3	E3	NRXN2	E3
VIL1	E2	GRIN2A	E3	CDH8	E3	KCND2	E3	NPY	E3
ANXA1	E2	NRCAM	E3	ASIC2	E3	CADM1	E3	ARNT2	E3
ELP4	E2	PAFAH1B1	E3	CDH10	E3	APBA2	E3	SETD2	E3
LAMB1	E2	MAPT	E3	GRM5	E3	PTPRN2	E3	FABP7	E3
ADA	E2	GRIA3	E3	RIMS3	E3	MEF2C	E3	MAZ	E3
ETFB	E2	GAD2	E3	GABRA2	E3	FMR1	E3	RPS6KA2	E3
SLC38A10	E2	ERBB4	E3	GABBR1	E3	GABRA4	E3	MYO16	E3
TRIO	E2	DLGAP2	E3	APC	E3	CACNA1G	E3	HTR1E	E3
NRXN3	E2	YWHAB	E3	MARK1	E3	ADCY1	E3	NPY5R	E3
ZBTB20	E2	PCDH9	E3	YWHAZ	E3	SORCS1	E3	TAC1	E3
CTCF	E2	GAD1	E3	NOS1AP	E3	CALB1	E3	CSNK1E	E3
GIGYF1	E2	HTR2A	E3	NLGN1	E3	ATP6V0C	E3	AGAP1	E3

Gene	Evidence	Gene	Evidence	Gene	Evidence	Gene	Evidence	Gene	Evidence
NBEA	E3	RGS7	E3	RARB	E3	NTNG1	E3	IL1RAPL1	E3
PSD3	E3	GRM7	E3	SLC6A1	E3	STX1A	E3	TSC1	E3
BDNF	E3	NR2E1	E3	BZRAP1	E3	PPP4C	E3	SCN1A	E3
OMG	E3	TSPAN7	E3	GRIK3	E3	MIF	E3	EN2	E3
CSRNP3	E3	LRP2	E3	GABRG3	E3	RYR3	E3	DPP10	E3
RAB3A	E3	SEZ6L2	E3	RXRB	E3	SNRPN	E3	GABRA3	E3
NOS1	E3	NR4A2	E3	DCLK2	E3	ST8SIA2	E3	MED12	E3
EPC2	E3	OPRK1	E3	NXPH1	E3	TAOK2	E3	CDH22	E3
SOX5	E3	SYNE1	E3	SPON1	E3	NPTX2	E3	FBXO33	E3
PCDH19	E3	AKT1	E3	MOG	E3	GALNT13	E3	PHF8	E3
SCN3A	E3	GABRB1	E3	BAIAP2	E3	DOCK4	E3	ESR2	E3
KCNJ3	E3	NPY2R	E3	DOC2A	E3	VGF	E3	PTPRZ1	E3
DRD1	E3	NDEL1	E3	PIK3CA	E3	ESR1	E3	GTF2I	E3
RAPGEF4	E3	DRD5	E3	FEZF2	E3	TUB	E3	CUX2	E3
MAP2K1	E3	AHI1	E3	ICA1	E3	HCRTR2	E3	OPRM1	E3
SYN1	E3	CLOCK	E3	PRLR	E3	CCK	E3	H2AFY	E3
SLC6A2	E3	MIB1	E3	CPLX2	E3	AR	E3	CGA	E3
NRG1	E3	ING3	E3	ADRA2A	E3	PER2	E3	WNT2	E3
JMJD1C	E3	SCG5	E3	GRM8	E3	FMNL2	E3	DLX2	E3
PARK2	E3	NPY1R	E3	SYT17	E3	AGTR2	E3	HTR7	E3

Gene	Evidence	Gene	Evidence	Gene	Evidence	Gene	Evidence	Gene	Evidence
GRM1	E3	TBL1X	E3	KHDRBS2	E3	CARTPT	E3	HMGB1	E3
FOLH1	E3	MAP2K2	E3	PCDH10	E3	PPIG	E3	DLX1	E3
GHSR	E3	MC4R	E3	DDC	E3	SLC25A12	E3	MYO1D	E3
BCL2	E3	GABRG1	E3	MEST	E3	PIK3CG	E3	FAM120C	E3
PTK2	E3	TYR	E3	AGRN	E3	DBH	E3	PITX3	E3
SHFM1	E3	FOXO1	E3	OPRD1	E3	SMO	E3	SCT	E3
COBL	E3	FLT1	E3	ADIPOQ	E3	ASMT	E3	LHB	E3
NLGN4Y	E3	GRPR	E3	SULT2A1	E3	TDO2	E3	HLA-A	E3
NRP2	E3	CYP17A1	E3	GPC6	E3	ABCA13	E3	FRAXE	E3
HTR5A	E3	TSPO	E3	NGFR	E3	MBD4	E3	FRAXA	E3
FBXO40	E3	MTF1	E3	NGF	E3	HEPACAM	E3	TGFB1	E3
SCAMP5	E3	VEGFA	E3	EGR2	E3	FHIT	E3	CHRNB2	E3
MDGA2	E3	HTR6	E3	OPRL1	E3	OXT	E3	TFPI	E3
GLO1	E3	SLC6A4	E3	MTNR1B	E3	XIRP2	E3	SND1	E3
COL11A1	E3	FABP3	E3	MAOB	E3	LZTS2	E3	EIF1AY	E3
ALDOA	E3	ZNF385B	E3	HLA-B	E3	NOS2	E3	INPP1	E3
ABCB1	E3	TBC1D5	E3	MAPK3	E3	CACNA1F	E3	FASTK	E3
PLD5	E3	ABAT	E3	TPH1	E3	HCRT	E3	MTHFR	E3
IGF1	E3	DRD3	E3	HTR3C	E3	GPX1	E3	DRG1	E3
WNK3	E3	CCDC148	E3	PLN	E3	CNR2	E3	CASP9	E3

Gene	Evidence	Gene	Evidence	Gene	Evidence	Gene	Evidence	Gene	Evidence
CNTN5	E3	CALCRL	E3	CYP11B1	E3	KIR2DS4	E3	CD38	E3
RXRG	E3	PAH	E3	SLC13A1	E3	MSR1	E3	HSD17B2	E3
POU6F2	E3	ESRRB	E3	HTR3A	E3	STK38	E3	PTGIR	E3
RAB39B	E3	EIF4E	E3	CASP3	E3	LEP	E3	PITX1	E3
MCHR1	E3	CHRNA4	E3	PYY	E3	GCH1	E3	POMGNT1	E3
CTNNA3	E3	ZBBX	E3	UBL5	E3	DDX53	E3	REEP3	E3
NFIL3	E3	NEUROG1	E3	PTS	E3	TSPAN12	E3	ARX	E3
CADPS2	E3	BHMT	E3	AGBL4	E3	RXRA	E3	ITGA4	E3
VPS13B	E3	NTS	E3	C15orf43	E3	DMPK	E3	MUC1	E3
PRL	E3	EXT1	E3	DRD4	E3	C3orf58	E3	NDNL2	E3
IL1RAPL2	E3	KDR	E3	CPA2	E3	WFS1	E3	CYP1B1	E3
GABRA6	E3	MTNR1A	E3	TRPV1	E3	RETN	E3	KIAA1586	E3
UBE2A	E3	TH	E3	CHMP2B	E3	EGF	E3	IGF2	E3
RASGRP1	E3	ADRA1B	E3	KIR3DS1	E3	FSHB	E3	DAO	E3
LHCGR	E3	NLRP3	E3	CYP7A1	E3	HES6	E3	LEPR	E3
AQP4	E3	SRD5A1	E3	SLC1A3	E3	CTSD	E3	C4B	E3
PPP1R1B	E3	PER1	E3	HTR1D	E3	DYX1C1	E3	SETDB2	E3
LRFN5	E3	HOXD13	E3	MT1A	E3	HOXB1	E3	VWA8	E3
DLX6	E3	LASP1	E3	SLC6A8	E3	MLPH	E3	DHCR7	E3
KCTD13	E3	CBS	E3	CYP11A1	E3	CYP21A2	E3	GNB1L	E3

Gene	Evidence	Gene	Evidence	Gene	Evidence	Gene	Evidence	Gene	Evidence
VASH1	E3	CYP1A1	E3	HTR1B	E3	COMT	E3	ADRB2	E3
CDK5	E3	PTGS2	E3	RHOXF1	E3	FABP5	E3	HLA-DRB1	E3
STS	E3	HRAS	E3	MC3R	E3	FRK	E3	ALDH3A1	E3
MBD6	E3	HOXD11	E3	HOXD12	E3	GJA1	E3	IL1R2	E3
GRID2	E3	HOXA1	E3	SLC7A5	E3	SDC2	E3	APOE	E3
NTF3	E3	MKL2	E3	LRRC1	E3	NEDD4	E3	FIGNL1	E3
RAF1	E3	APAF1	E3	ADORA2A	E3	IMMP2L	E3	SLC25A13	E3
KCNN3	E3	POR	E3	NKX2-5	E3	NSD1	E3	GCHFR	E3
SNTG2	E3	ZNF778	E3	NTRK1	E3	HIRIP3	E3	PECAM1	E3
HTR1A	E3	CASC4	E3	AVPR1B	E3	QDPR	E3	RPL10	E3
EFHC2	E3	MBD1	E3	RARG	E3	NOSTRIN	E3	CYP2D6	E3
NDN	E3	HGF	E3	ARC	E3	NTF4	E3	DPYD	E3
HTR3B	E3	TAS2R1	E3	AGRP	E3	PRODH	E3	GSTM1	E3
HOMER1	E3	SRD5A2	E3	CA6	E3	SSTR5	E3	ADK	E3
PNOC	E3	AANAT	E3	COPG2	E3	ODF3L2	E3	AMT	E3
SETDB1	E3	MFSD6	E3	GSTP1	E3	POMC	E3	HSD17B4	E3
LHFPL3	E3	OR1C1	E3	KIR2DS5	E3	SLC6A14	E3	FOLR2	E3
OSBPL6	E3	TPH2	E3	HSD11B1	E3	NDUFA5	E3	SLC19A1	E3
CNTNAP5	E3	FOSB	E3	KIR2DS1	E3	AGMO	E3	HFE	E3
SLC30A4	E3	CCDC64	E3	RFC1	E3	TNF	E3	TCN2	E3

Gene	Evidence	Gene	Evidence	Gene	Evidence	Gene	Evidence	Gene	Evidence
CDH9	E3	TAF1C	E3	GHRL	E3	CLTCL1	E3	PCBD1	E3
CHRM5	E3	TMEM100	E3	VIPR1	E3	MBD3	E3	SPR	E3
RARA	E3	TBX1	E3	TSC2	E3	FEZF1	E3	SERPINE1	E3
HS3ST5	E3	UBE2H	E3	ST7	E3	ROBO3	E3	PEX7	E3
LAMB4	E3	HCRTR1	E3	SHBG	E3	MCPH1	E3	MTRR	E3
DRD2	E3	TBL1Y	E3	ROBO4	E3	CDKL4	E3	IL1RN	E3
CYP19A1	E3	UPP2	E3	DAOA	E3	TP53	E3	SHMT1	E3
STK39	E3	HES1	E3	TRIM32	E3	ST7-OT3	E3	MTR	E3
SLC4A10	E3	ATP1A1	E3	HSD17B3	E3	MAOA	E3	SCP2	E3
MCM7	E3	ADSL	E3	DHFR	E3				

Table A.4: The negative ground truth genes for ASD

ATXN1	NDP	PLEKHM1	GDF5	HSPB3	PDE6B	FAM161A	TNNI3	CACNA1S
NIPBL	PRPF8	SLC45A2	HSPD1	SEPT9	ZNF513	PRPF3	SLC11A2	PALB2
ATRX	SNCB	HR	MEN1	CHIC2	ABCC6	GHR	HFE2	TMC8
PLCB1	SOX10	LPP	ETV6	MAD1L1	ASPSCR1	BBS4	EVC2	TBP
KRAS	EDN3	FUS	LPL	PAX8	PANK2	USH1G	ALG12	IL10RA
KCNQ2	CA4	MSX1	SYNE2	NOTCH2	C2orf71	STRC	MKS1	WDPCP
REEP1	TTBK2	PHF6	TAT	SCN1B	GNB3	SLC19A3	TLR4	AP3B1
NEUROD1	MLLT10	NODAL	HMCN1	ALMS1	BEST1	DLD	CPS1	SLC4A11
TRIM33	ZIC2	SH3TC2	LRAT	TNNT2	CSTB	SLC2A4	IFNG	DSP
MAPK10	PDYN	ADAMTS17	GLIS2	ATXN3	MYOT	TAP1	BSND	JUP
CREBBP	AKAP9	ZFHX3	CYP7B1	RB1	RGR	TGFBR2	ESPN	SLC52A3
SCN3B	ACVR2B	PRICKLE2	PROM1	CCDC28B	GPC3	RP9	PROKR2	CENPJ
NEFL	HMGA1	TGIF1	LBR	GCGR	MYH9	CLN3	KCNQ4	PDE6G
LARGE	JAG1	NIPA1	PNKP	TUBA1A	STAT3	FIG4	SMARCB1	F13A1
OPCML	LZTS1	GATA2	HBA1	F13B	FGF8	SLC5A2	FLNB	MVK
SOX2	RNF139	SCN4B	SLC17A8	CIITA	NOBOX	SEMA4A	RPGR	MYH7
DCTN1	IDS	PARK7	TMEM67	MPZ	OPN1MW	EDAR	IDH2	BSCL2
FGFR2	MN1	OAT	ERCC4	SPTA1	TFR2	IL23R	AHCY	FKTN
CACNA1A	MXI1	LRRC8A	USH2A	PITPNM3	PITX2	KRT1	ABCD1	CAPN3

BMPR2	SPG7	ATXN8OS	MCM6	LIPI	RP1L1	F5	TGFB3	MYL3
PDE8B	COL4A4	MAP3K1	TNFSF11	TWIST1	NUP62	NYX	ICOS	ECE1
EP300	BCR	LDB3	TNFRSF11A	PSAT1	TBX20	SCN4A	CEP152	PSENEN
PDGFRA	CNGB1	MYO3A	ALOX12B	DNAI1	SPINK5	STRA6	PHKG2	GTF2H5
KIF5A	GNAS	PAX4	XPA	DYNC2H1	LRRK2	SPATA7	MTMR2	CDKN2A
DPM1	TRPM1	AMELX	NDUFS7	RSPH9	CABP4	IL1B	ERCC6	TRIP11
ARHGEF12	OCRL	IMPG2	BARD1	AVPR2	MYL2	GDF3	NHLRC1	ERCC8
ATN1	ZEB1	RLBP1	RPGRIP1	GJB1	TMPRSS3	GATA6	POMT1	TERT
ABCC8	HCN4	MSH6	RP1	IKBKAP	UCP3	CSRP3	IMPDH1	PKD2
SUMO1	OTX2	CLCNKB	CBX2	SNTA1	STXBP2	FLCN	AGT	VCL
CHD7	SPTBN2	PLA2G6	KLK4	FLG	CHM	ALPL	COX6B1	ALAS2
GPI	KCNJ1	OTOF	KRT83	FKRP	EYS	NDUFS3	STX11	L2HGDH
KCNA1	MMADHC	DFNB31	FHL1	REN	TMC1	BBS5	DFNA5	TMEM43
RUNX2	USH1C	FOXF1	PRNP	ERCC3	CEBPA	ACAN	AVP	HSD3B7
PPP2R2B	HTT	SIM1	HABP2	NPR2	PTPRJ	TSG101	NPHP3	DES
RET	PTHLH	PRPH2	GAA	PEX5	SLC26A2	HES7	GMPS	AAAS
KCNH2	SETX	PAX7	NOG	SGCG	KCNJ11	FBN1	OPN1LW	ADH1C
RB1CC1	EDNRB	GDAP1	TRIM24	IHH	NPPA	CEL	ODC1	OFD1
STX16	XYLT1	STOX1	ASPA	GNE	KCNE1	ROR2	NME1	PHKB
GABRG2	LEMD3	CDKL5	TMIE	KCNE2	GJB2	RNASEH2B	CD244	SURF1

ITPR1	CDHR1	UCP1	FASLG	PDZD7	SUFU	FLT4	FBXO7	XYLT2
MAPK8IP1	SLC3A1	ATM	AKR1D1	LIG4	SLC40A1	DFNB59	BBS10	LCAT
WNK1	KCNC3	NPHP4	GIF	SLC12A3	LRP5	GFI1	MYLK2	NCSTN
SNCA	NDUFA13	SLC30A8	PRICKLE1	RAB7A	PRCD	PC	GRK1	ABCC2
FGF14	ABCC9	ARHGAP26	UBR1	KCNK9	FGA	BRAF	POLG	PEX26
UCHL1	FZD4	ARFGEF2	FIGLA	POU4F3	IL10	FANCC	FOXC1	F12
TARDBP	SLC26A4	SMAD9	RPE65	ZNF592	IGF2BP2	CEP290	RYR1	UNC13D
ATP2A2	RNF6	FAM134B	ARL6	TGFBR1	MYOC	NPC1	PEX14	TNFRSF13B
PABPN1	ATP1A2	FOXE3	TTN	DNAH11	KRT86	SLC6A19	MYH6	KLHDC8B
TRPS1	BMP15	LCA5	COL4A5	ATG16L1	FGB	SCO2	APOA5	HLCS
KCNJ10	MSX2	TLL1	CRB1	GLI2	TTR	CLCN7	PPT1	KRT13
KIF1B	GIGYF2	SLC12A1	WWOX	GJB6	HGSNAT	NPC2	KRT81	CDK5RAP2
KCNA5	NR0B1	SAG	GARS	CYBA	KIF7	EPCAM	CDKAL1	HEXA
PPIB	BCOR	DMP1	TMPRSS6	NR5A1	CRX	MLH3	PMP22	NCF1
DICER1	GDF6	NRL	CCDC50	DLX3	HNF1B	BBS12	BBS1	KLF1
ZIC3	PTCH2	EDNRA	MTTP	CDH23	DBT	PFKM	SFTPA1	CA2
ACTN2	BMPR1B	FGF23	INSR	SLC34A2	PADI4	VHL	ACTA2	NR0B2
CACNA1C	EYA4	FOXE1	RXFP2	FGD4	CNGA1	MFN2	GRXCR1	SAMHD1
SOS1	BEAN1	AIP	CST3	EDARADD	BANK1	LMBR1	NRAS	SPTLC2
DMD	SGCD	AKT2	RSPH4A	DHH	PCCA	LOXHD1	UMOD	NOD2

LYST	CNGB3	ALOXE3	NF2	GUCY2D	TULP1	IRF1	MCOLN1	COA5
HPRT1	PRCC	AASS	STK11	CFC1	AGL	ACVRL1	DUOXA2	ABHD5
DLL3	PRKAR1A	FLT3	CHMP4B	FOXL2	HCCS	SGSH	OCLN	IDUA
ZFPM2	MYO15A	KLF11	F9	RDH12	UPK3A	DUOX2	RMRP	AMACR
ARHGEF9	GFAP	DNM2	MS4A1	ENAM	SHH	BLK	COMP	ERCC5
THRB	ATP7A	GLUD1	ACVR1	KARS	TERC	PRKCH	MAP3K8	TPRN
TOPORS	SIX3	PRKAG2	ENPP1	ACTG1	F2	EVC	MYH3	MMAA
EDA	ATXN7	SDC3	PHEX	LHFPL5	PINK1	HOGA1	MKKS	PRPF31
SBF2	EGFR	PTPN1	VAPB	COG1	POLR1D	NKX2-6	ESCO2	AFG3L2
RYR2	ATL1	AXIN2	COCH	LRTOMT	KIAA0196	PAX9	PHGDH	ITPKC
NR4A3	GRM6	FLNA	NME8	UROS	PEX10	AP5Z1	NT5C3A	RFX5
SOX9	PICALM	PRPH	SCN9A	MYBPC3	EPM2A	POMT2	RPS26	SGCB
PTCH1	SLC25A22	GDF1	ATP2C1	SLC5A5	APOA1	CD2AP	HNF1A	MANBA
PNPLA6	EWSR1	MPL	PORCN	KCNQ1	ARSB	AGXT	GALC	MED25
KIT	NEFH	SCNN1G	GLI3	PTPN11	INSL3	OTOA	INVS	ERCC2
PLP1	FGFR3	SLC16A2	VSX2	ADAMTS10	TLR2	LIPC	CERKL	ITGA2B
WHSC1L1	KCNJ2	SRY	AGTR1	COL10A1	CYP3A5	NPHP1	TGM1	SIX5
KLF6	RHAG	AXIN1	DIAPH2	PYGM	DSPP	CD81	GJC2	ASL
MAMLD1	GCK	OTC	DNAAF1	MMP13	MYO1A	ALG11	SCNN1B	MFSD8
PKD1	ACTN4	RAX	MSH2	EPB41	EPHA2	KRT10	NCOA4	EFHC1

EYA1	SOD1	WT1	BBS9	GNAS-AS1	ARG1	HBA2	POLH	PEX3
FGD1	RBM8A	TF	ALS2	MEFV	G6PC3	NR2E3	VWF	NPM1
PAX3	SNRNP200	FGF10	MITF	LFNG	CD19	GATM	GP1BB	PLA2G2A
ROBO2	SLC34A1	IFT80	TPO	PRX	FAM83H	ZNF469	FSCN2	PTPN22
GPR98	CYP2A6	CLRN1	PCDH15	PDX1	JAK2	GP9	YARS	BBS2
CNGA3	TYRP1	CYLD	IFT122	CCND1	D2HGDH	SPG20	SH2B3	CD79B
GUCA1A	THPO	CHN1	MYO6	SH3BP2	HGD	SIX1	PPARGC1B	CYBB
SPTAN1	CNBP	PROK2	IRF4	GJB4	WDR72	PTPRQ	CAV3	CFI
NDUFS8	CLCN5	GDNF	LITAF	MC1R	ABCB4	FOXP3	ELANE	DYSF
OCA2	NOTCH1	SCN5A	MMP20	PTGIS	WDR35	GATA4	CDC73	C5
IRS2	ADD1	GPR143	GLDC	KCNJ5	ITGA7	FKBP10	MPDU1	SLC16A1
COL4A3	SLC7A9	PPP1R3A	CLN8	ABCA4	TNFRSF13C	ACTC1	FERMT3	BBS7
TRIM37	TFAP2A	RDX	NDRG1	MYO7A	DSC2	PHKA2	RAX2	KLHL7
IRS1	ATP7B	ABCA1	PKHD1	G6PC	PKP2	CP	CLEC7A	INS
RHO	CDKN1C	SFTPA2	GRHL2	PDE6C	GNAT1	NLRP7	SLC25A4	TTC8
ATXN10	CFTR	F11	BMP4	LTBP3	SRD5A3	H19	TRPV4	RP2
FSHR	ADRB3	COL2A1	USP9Y	DNAI2	EMD	CLCNKA	MMAB	FBP1
ATXN2	POU3F4	BMP2	FOXI1	KCNQ1OT1	SEPN1	CFHR3	NBN	TCAP
RUNX1	PLAG1	AIPL1	EFNB1	RSPO4	NDUFA10	IL6	RPL35A	COL8A2
PRKCG	TECTA	RBM20	ZFYVE27	CD24	GPD1L	RFXAP	KCTD7	SPG11

PRPS1	TAPBP	SDCCAG8	PCM1	RD3	COL5A2	IGLL1	KCNE3	MARVELD2
PCCB	COX14	PDGFRL	MYH14	CTNS	MUTYH	IVD	ALG9	RPS10
COG4	GSN	PEX1	IQCB1	TNNI2	SLC37A4	ALG6	COL5A1	HEXB
FOXRED1	AK2	SDHC	TPP1	RPL5	HTRA1	SMPD1	GM2A	MLH1
COG5	CLN6	SUMF1	IGHM	B4GALT7	FASTKD2	FECH	RPS24	SEC23B
TNNC1	SDHB	PMM2	PEX13	RPS19	ANG	SERPING1	HADH	PNP
APOL1	SLC35A1	TPM1	KRT6B	ALG1	GALNS	ABCB7	DDB2	CRTAP
MPI	ADCK3	SLC25A15	OSTM1	FAS	RNASEH2A	MPLKIP	C3	OAS1
DPM3	KRT16	BCKDHB	SLC22A18	LOXL1	ASPM	PEX2	TDP1	C10orf2
MINPP1	SLC25A38	LMNA	WAS	SLC39A4	NDUFAF2	TCOF1	BTD	NDUFA9
NDUFAF6	SERPINB6	IRF5	ALG3	SMN2	TNXB	MCCC2	CAV1	FXN
ATR	TRIOBP	SLC17A5	UROD	WRN	CRYAB	ADAM9	GNPAT	LTBP2
CASP10	MMACHC	IL2RG	ALDH3A2	ENG	TAZ	CPOX	OPTN	FAH
MGAT2	COL1A1	NDUFA12	TAP2	AK1	GALE	COX15	ASAH1	DKC1
DNAAF2	C2	ASS1	BRCA1	ECM1	GOLGA5	EXT2	TPM2	ACAT1
SMN1	HAX1	HPS1	PLOD1	AGPAT2	LEPRE1	SLC35C1	FUCA1	ALG8
SNAI2	HSPB8	TMC6	CASP8	CHEK2	PDHA1	IDH3B	GOSR2	AGPS
WDR36	PRF1	SGCA	OGG1	DOLK	CTSC	NAGLU	IL10RB	RPS7
LMBRD1	TNNT3	COL6A1	BRCA2	HTRA2	EPHX1	GALK1	COG8	HMMR
GNS	SLC33A1	BLM	POLR1C	HMBS	COG7	ATP2A1	STIL	CD79A

BTK	CARD9	IRF6	NEXN	RNASEH2C	ARSA	KRT6A	FANCA	PPOX
BRIP1	SQSTM1	ALOX5AP	CDAN1	SLC22A5	MCCC1	HNF4A	AGA	TCIRG1
GCSH	CDH1	GP1BA	LYZ	FCGR2B	OSMR	HLA-DQB1	GLB1	BLNK
SDHA	ITGB2	EMG1	B4GALT1	DIAPH1	HBB	NCF2	LAMP2	HSPB1
HPGD	CYP27A1	CISD2	NDUFA2	MYC	MAN2B1	INF2	PHYH	CFHR1
KRT4	CFB	CLCN1	MESP2	SLC22A4	MYH11	UGT1A1	RAD51	GALT
HPD	GPD2	ERBB2	PEX19	COL1A2	SLC2A10	GRHPR	COL3A1	PHB
HAMP	GBA	GBE1	PPARG	RFXANK	LAMA2	DNASE1	SCARB2	PTRF
BCL10	NUP214	GLA	ALG10	COL11A2	RFT1	BCKDHA	KRT17	BUB1B
ELN	FBLN5	XPC	GATA1	RPS17	SPTLC1	TP63	TREX1	AURKA
SERPINA1	MUC5B	COL6A3	NDUFS4	RPL11	IKBKG	HLA-DQA1	XRCC3	CLN5
MOGS	RAD54L	CFH	PSPH	DPAGT1	GYG1	GLRX5	GUSB	CYP4V2
RDH5	COL6A2	GJB3	DSG2	RECQL4	PPP2R1B			

Table A.5: The positive ground truth genes for ID

Gene	**Evidence**	**Gene**	**Evidence**	**Gene**	**Evidence**	**Gene**	**Evidence**	**Gene**	**Evidence**
ADSL	E1	IDUA	E1	SLC6A8	E1	IDS	E1	CYB5R3	E2
AGA	E1	L1CAM	E1	SMS	E1	RPS6KA3	E1	DYNC1H1	E2
ATP7A	E1	LAMP2	E1	TAT	E1	ATP2A2	E2	DPAGT1	E2
ATRX	E1	MECP2	E1	TSC1	E1	ATR	E2	DPYD	E2
CREBBP	E1	MID1	E1	TSC2	E1	AUH	E2	CLN8	E2
DCX	E1	NDP	E1	UBE3A	E1	BBS1	E2	ERCC2	E2
DHCR7	E1	NEU1	E1	ALDH5A1	E1	BBS2	E2	ERCC3	E2
DKC1	E1	NF1	E1	IKBKG	E1	BBS4	E2	FBN1	E2
DMD	E1	OCRL	E1	ZEB2	E1	BCKDHA	E2	FGFR1	E2
DNMT3B	E1	OPHN1	E1	SLC12A6	E1	BCKDHB	E2	FGFR3	E2
ERCC6	E1	OTC	E1	IL1RAPL1	E1	BCS1L	E2	FGFR2	E2
FKTN	E1	PDHA1	E1	MLYCD	E1	BRAF	E2	FH	E2
FGD1	E1	PEPD	E1	NSD1	E1	CA2	E2	FOXG1	E2
FLNA	E1	PGK1	E1	FKRP	E1	CBS	E2	FUCA1	E2
FMR1	E1	PLP1	E1	PHF6	E1	CHD2	E2	GALE	E2
AFF2	E1	PRPS1	E1	ARX	E1	ERCC8	E2	GALT	E2
GCH1	E1	PTEN	E1	HPRT1	E1	COX10	E2	GAMT	E2
GK	E1	PTPN11	E1	ALDH18A1	E1	CTNNB1	E2	GATM	E2

Gene	Evidence	Gene	Evidence	Gene	Evidence	Gene	Evidence	Gene	Evidence
ACOX1	E2	ANK3	E2	ALDH3A2	E2	DBT	E2	GDI1	E2
AGTR2	E2	ASL	E2	AMT	E2	TIMM8A	E2	GLB1	E2
GLDC	E2	MANBA	E2	PEX7	E2	NKX2-1	E2	ADGRG1	E2
GLI3	E2	MAT1A	E2	POU1F1	E2	TSPAN7	E2	NRXN1	E2
GNAS	E2	MEF2C	E2	PPT1	E2	UBE2A	E2	EIF2AK3	E2
GNS	E2	MGAT2	E2	PEX5	E2	VLDLR	E2	MED12	E2
GRIA3	E2	NAGLU	E2	SCN8A	E2	ZIC2	E2	SCO2	E2
GRIN2B	E2	NDUFS1	E2	SDHA	E2	TUBA1A	E2	SLC25A15	E2
GSS	E2	NDUFS2	E2	SGSH	E2	SHOC2	E2	ALG3	E2
GUSB	E2	NDUFV1	E2	SIX3	E2	KMT2D	E2	SLC9A6	E2
HSD17B10	E2	NDUFS4	E2	SMARCA2	E2	MKKS	E2	ARFGEF2	E2
HCCS	E2	NDUFS8	E2	SMPD1	E2	KDM5C	E2	POMT1	E2
HCFC1	E2	NFIX	E2	SOS1	E2	GNPAT	E2	RAI1	E2
HLCS	E2	NHS	E2	SOX3	E2	CUL4B	E2	POLR3A	E2
HPD	E2	PNP	E2	SOX10	E2	OFD1	E2	PNKP	E2
HRAS	E2	NTRK1	E2	SPTAN1	E2	CASK	E2	RAB3GAP1	E2
IGF1	E2	PAFAH1B1	E2	CDKL5	E2	ALDH4A1	E2	IQSEC2	E2
KCNJ11	E2	PAH	E2	STXBP1	E2	SLC4A4	E2	ARHGEF9	E2
KRAS	E2	PAX6	E2	SUOX	E2	DPM1	E2	SATB2	E2
LAMA2	E2	PC	E2	SURF1	E2	SYNGAP1	E2	ZFYVE26	E2

Gene	Evidence	Gene	Evidence	Gene	Evidence	Gene	Evidence	Gene	Evidence
LRP2	E2	PCNT	E2	SYP	E2	ARHGEF7	E2	SETBP1	E2
MAN2B1	E2	PEX1	E2	TCF4	E2	LARGE1	E2	SLC17A5	E2
BSCL2	E2	POLR3B	E2	PORCN	E2	MMAA	E2	BCOR	E2
FOXP1	E2	MBD5	E2	UPF3B	E2	UBR1	E2	NSUN2	E2
MMADHC	E2	INPP5E	E2	ALG12	E2	ASPM	E2	POMGNT1	E2
ALG6	E2	MCCC1	E2	MCPH1	E2	KANSL1	E2	ABHD5	E2
NSDHL	E2	COQ8A	E2	EHMT1	E2	TUBB2B	E2	NDE1	E2
EXOSC3	E2	GJC2	E2	FRAS1	E2	KIF7	E2	MCOLN1	E2
SHANK3	E2	NLGN4X	E2	TRAPPC9	E2				

Table A.6: Genes discarded from the negative set for ID

ATRX	ALG6	ERCC8	CDKL5	FGD1	SDHA	FKRP	MGAT2	SGSH
KRAS	SMPD1	OFD1	ARFGEF2	SPTAN1	HPD	ERCC3	ADCK3	GATM
CREBBP	NAGLU	SURF1	FLNA	NDUFS8	BCKDHA	PEX5	PEX1	ALG12
LARGE	GNPAT	HLCS	OTC	NDP	NDUFS4	PTPN11	SLC25A15	ERCC6
FGFR2	FUCA1	CA2	TAT	SOX10	GUSB	KCNJ11	BCKDHB	POMT1
DPM1	PNP	ABHD5	PNKP	ZIC2	GALT	KIF7	SLC17A5	PPT1
ATP2A2	DKC1	IDUA	UBR1	IDS	MCCC1	DBT	ALG3	BBS1
SOS1	AGA	MMAA	PORCN	GNAS	MAN2B1	BBS4	ALDH3A2	MCOLN1
DMD	GLB1	MANBA	GLI3	OCRL	LAMA2	FBN1	ASPM	MKKS
HPRT1	LAMP2	ERCC2	GLDC	MMADHC	IKBKG	PC	GALE	GJC2
ARHGEF9	DPAGT1	ASL	CLN8	BCOR	PRPS1	BRAF	PDHA1	HCCS
PLP1	GNS	BBS2	TUBA1A	ATP7A	ATR	SCO2	FKTN	BSCL2
FGFR3	PHF6	SIX3						

Table A.7: DeepND ranking and posterior risk for selected genes located in 16p11.2

Gene Entrez ID	Gene Name	ASD Risk Probability	DeepND ASD Ranking	ASD Gold Standard	ID Risk Probability	DeepND ID Ranking	ID Gold Standard
10847	SRCAP	0.862	54	0	0.523	190	0
26470	SEZ6L2	0.563	475	E3-E4	0.123	10065	0
11151	CORO1A	0.524	617	0	0.092	16497	0
23214	XPO6	0.517	649	0	0.120	10533	0
9344	TAOK2	0.490	4126	E3-E4	0.204	5576	0
112755	STX1B	0.427	5041	0	0.086	18009	0
11273	ATXN2L	0.413	5537	0	0.087	17807	0
9810	RNF40	0.403	6104	0	0.098	14695	0
9739	SETD1A	0.399	6297	0	0.119	10629	0
2521	FUS	0.369	8084	-1	0.092	16411	-1
677813	SNORA30	0.368	8149	0	0.151	7663	0
253980	KCTD13	0.366	8247	E3-E4	0.107	12616	0
4150	MAZ	0.363	8464	E3-E4	0.134	8863	0
930	CD19	0.331	10663	-1	0.073	21718	-1
22928	SEPHS2	0.331	10685	0	0.355	631	0
8448	DOC2A	0.327	11020	E3-E4	0.127	9598	0
5595	MAPK3	0.309	13472	E3-E4	0.096	15377	0
8479	HIRIP3	0.295	15897	E3-E4	0.116	11060	0
6693	SPN	0.295	15906	0	0.100	14149	0
226	ALDOA	0.281	18315	E3-E4	0.092	16479	0
5531	PPP4C	0.269	19990	E3-E4	0.067	23007	0
10295	BCKDK	0.262	20875	E2	0.104	13332	0
1731	SEPTIN1	0.248	22013	0	0.128	9411	0

Table A.8: DeepND ranking and posterior risk for selected genes located in 15p11-13

Gene Entrez ID	Gene Name	ASD Risk Probability	DeepND ASD Ranking	ASD Gold Standard	ID Risk Probability	DeepND ID Ranking	ID Gold Standard
2562	GABRB3	0.782	108	E1	0.426	401	0
2567	GABRG3	0.656	250	E3-E4	0.112	11675	0
8123	PWAR5	0.513	1128	0	0.232	2126	0
123606	NIPA1	0.486	4158	-1	0.161	7192	-1
7337	UBE3A	0.423	5199	E3-E4	0.054	24908	E1
81109	OR11K1P	0.399	6278	0	0.135	8784	0
2558	GABRA5	0.385	7101	E3-E4	0.076	20894	0
503519	LINC00929	0.370	8002	0	0.148	7842	0
4692	NDN	0.339	10019	E3-E4	0.102	13775	0
54551	MAGEL2	0.336	10204	E1	0.119	10678	0
81614	NIPA2	0.315	12625	0	0.114	11393	0
6638	SNRPN	0.311	13216	E3-E4	0.068	22794	0
7681	MKRN3	0.304	14397	0	0.115	11249	0
8926	SNURF	0.301	14823	0	0.072	21917	0
4948	OCA2	0.298	15385	-1	0.113	11621	-1
114791	TUBGCP5	0.260	21091	0	0.117	10868	0
57194	ATP10A	0.228	23197	E2	0.078	20398	0

Table A.9: DeepND ranking and posterior risk for selected genes located in 15p13.3

Gene Entrez ID	Gene Name	ASD Risk Probability	DeepND ASD Ranking	ASD Gold Standard	ID Risk Probability	DeepND ID Ranking	ID Gold Standard
1139	CHRNA7	0.367	8172	E2	0.139	8496	0
4308	TRPM1	0.196	24521	-1	0.082	19113	-1
6263	RYR3	0.168	25265	E3-E4	0.089	17107	0
6447	SCG5	0.288	17168	E3-E4	0.101	14083	0
643699	GOLGA8N	0.352	9101	0	0.183	6290	0
653125	GOLGA8K	0.339	10013	0	0.293	938	0

Table A.10: DeepND ranking and posterior risk for selected genes located in 1q21.1

Gene Entrez ID	Gene Name	ASD Risk Probability	DeepND ASD Ranking	ASD Gold Standard	ID Risk Probability	DeepND ID Ranking	ID Gold Standard
149013	NBPF12	0.428	5028	0	0.479	273	0
647135	SRGAP2B	0.408	5799	0	0.524	187	0
9939	RBM8A	0.404	6044	-1	0.094	15852	-1
114814	GNRHR2	0.385	7115	0	0.426	403	0
171423	PDIA3P1	0.340	9924	0	0.392	494	0
148738	HJV	0.305	14261	-1	0.092	16456	-1

Table A.11: DeepND ranking and posterior risk for selected genes located in 22q11

Gene Entrez ID	Gene Name	ASD Risk Probability	DeepND ASD Ranking	ASD Gold Standard	ID Risk Probability	DeepND ID Ranking	ID Gold Standard
57553	MICAL3	0.608	352	0	0.105	13186	0
7152	TOP1P2	0.513	799	0	0.232	4626	0
30847	CECR9	0.485	4172	0	0.200	5713	0
723778	MIR650	0.484	4188	0	0.195	5913	0
7290	HIRA	0.483	4194	0	0.150	7758	0
55670	PEX26	0.323	11419	-1	0.111	11938	-1
2679	GGT3P	0.323	11489	0	0.082	19258	0
6899	TBX1	0.321	11678	E3-E4	0.087	17728	0
3543	IGLL1	0.312	13099	-1	0.072	21964	-1
2812	GP1BB	0.301	14969	-1	0.102	13819	-1
54584	GNB1L	0.296	15713	E3-E4	0.148	7848	0
91319	DERL3	0.296	15738	0	0.107	12751	0
1652	DDT	0.296	15827	0	0.088	17358	0
1415	CRYBB2	0.291	16620	0	0.075	21264	0
135	ADORA2A	0.281	18330	E3-E4	0.077	20657	0
4282	MIF	0.277	18924	E3-E4	0.180	6407	0
1312	COMT	0.258	21243	E3-E4	0.080	19737	0
8218	CLTCL1	0.210	24007	E3-E4	0.099	14635	0
5625	PRODH	0.148	25582	E3-E4	0.066	23301	0

Table A.12: DeepND ranking and posterior risk for selected genes located in 22q11.21-22

Gene Entrez ID	Gene Name	ASD Risk Probability	DeepND ASD Ranking	ASD Gold Standard	ID Risk Probability	DeepND ID Ranking	ID Gold Standard
2781	GNAZ	0.471	4313	0	0.238	1498	0
96610	BMS1P20	0.399	6287	0	0.424	407	0
91179	SCARF2	0.299	15268	0	0.098	14818	0
128989	TANGO2	0.297	15628	0	0.131	9138	0
64976	MRPL40	0.297	15646	0	0.109	12222	0
23532	PRAME	0.239	22582	0	0.055	24827	0

Table A.13: DeepND ranking and posterior risk for selected genes located in 17p11.2

Gene Entrez ID	Gene Name	ASD Risk Probability	DeepND ASD Ranking	ASD Gold Standard	ID Risk Probability	DeepND ID Ranking	ID Gold Standard
10743	RAI1	0.912	35	0	0.644	58	E2
23164	MPRIP	0.557	488	0	0.443	357	0
2310	FOXO3B	0.461	4422	0	0.645	57	0
201163	FLCN	0.384	7155	-1	0.076	21037	-1
257468	MEIS3P2	0.357	8805	0	0.312	816	0
51168	MYO15A	0.320	11894	-1	0.093	15989	-1
224	ALDH3A2	0.290	16847	-1	0.114	11373	E2
23495	TNFRSF13B	0.284	17888	-1	0.033	25768	-1
218	ALDH3A1	0.241	22451	E3-E4	0.082	19160	0
146802	SLC47A2	0.152	25533	0	0.044	25566	0
6470	SHMT1	0.138	25682	E3-E4	0.064	23561	0

Table A.14: DeepND ranking and posterior risk for selected genes located in 16p13

Gene Entrez ID	Gene Name	ASD Risk Probability	DeepND ASD Ranking	ASD Gold Standard	ID Risk Probability	DeepND ID Ranking	ID Gold Standard
1387	CREBBP	0.874	47	-1	0.581	111	E1
2903	GRIN2A	0.712	166	E3-E4	0.377	541	0
84656	GLYR1	0.676	213	0	0.185	6220	0
115	ADCY9	0.665	228	0	0.295	918	0
7874	USP7	0.631	302	0	0.429	386	0
57496	MRTFB	0.457	4479	E3-E4	0.075	21261	0
9665	MARF1	0.433	4893	0	0.140	8394	0
54715	RBFOX1	0.428	5017	E2	0.103	13627	0
1186	CLCN7	0.407	5824	-1	0.190	6075	-1
4629	MYH11	0.396	6541	-1	0.071	22325	-1
18	ABAT	0.361	8544	E3-E4	0.290	966	0
527	ATP6V0C	0.359	8688	E3-E4	0.116	10973	0
7249	TSC2	0.334	10397	E3-E4	0.079	20074	E1
8312	AXIN1	0.304	14308	-1	0.089	17267	-1
2072	ERCC4	0.298	15412	-1	0.162	7175	-1
6755	SSTR5	0.293	16268	E3-E4	0.091	16681	0
54820	NDE1	0.259	21130	0	0.083	18952	E2
56052	ALG1	0.258	21220	-1	0.106	12938	-1
5373	PMM2	0.257	21289	-1	0.152	7612	-1
8912	CACNA1H	0.247	22102	E1	0.088	17607	0
1773	DNASE1	0.235	22857	-1	0.117	10913	-1
5310	PKD1	0.213	23898	-1	0.056	24734	-1