Identification of Genes Associated with Autism Spectrum Disorder using Network Propagation-Based Semi-Supervised Learning

Subhan[1], Muhammad Huzaifa[2], and Rabia Shahid[3]

[1,2,3]Center of Data Science, Government College University Faisalabad

April 14, 2025

Abstract

Autism Spectrum Disorder (ASD) presents significant challenges in genetic analysis due to its heterogeneity and the low penetrance of associated mutations. We present a semi-supervised machine learning approach using network propagation to identify and rank genes according to their association with ASD. Our method integrates protein-protein interaction data with known disease-related and unrelated genes through a modified random walk with restart algorithm. Cross-validation testing demonstrated strong performance with an area under the receiver operating curve of approximately 0.85, area under the precision-recall curve of 0.8, and Matthews correlation coefficient of 0.57. Novel candidate genes identified by our approach showed significant enrichment in pathways related to synaptic transmission, ion transport, and specific neurotransmitter systems previously linked to ASD. Validation against the Simons Simplex Collection confirmed that 70% of our identified candidate genes harbor at least one de novo variant in ASD patients. This approach offers a promising strategy for identifying biologically relevant ASD-associated genes, potentially advancing both diagnostic capabilities and our understanding of ASD's molecular mechanisms.

Keywords: Autism Spectrum Disorder, Network Propagation, Semi-Supervised Learning, Gene Identification, Protein-Protein Interaction

1 Introduction

Autism Spectrum Disorder (ASD) affects approximately 1% of the population and manifests as early difficulties in social interaction and communication, coupled with restricted behaviors and interests [1]. Despite intensive research yielding more than 800 putative risk genes, a definitive genetic cause can be identified in only 10-20% of cases [2].

The genetic architecture of ASD is extraordinarily complex, involving various transmission modes including de novo mutations, inherited variants, and combinations of common and rare variants affecting specific biological pathways [3]. These genetic factors likely interact with environmental influences, further complicating the picture [2]. This heterogeneity presents significant challenges for clinical translation and diagnosis, which still relies primarily on behavioral observation rather than biological markers [4].

Traditional genomic approaches require extremely large sample sizes to establish statistically significant associations. An alternative approach involves leveraging computational methods to identify genes in biological pathways implicated in ASD, offering insights into the specific processes disrupted in individual patients or patient subgroups.

Several machine learning techniques have been applied to this problem. Krishnan et al. [5] used Support Vector Machines with brain-specific gene interaction networks to achieve an AUC of 0.8 in discriminating ASD-related genes. Similarly, Asif et al. [6] employed Random Forests with Gene Ontology-based semantic similarity measures to reach comparable performance.

Our approach utilizes a modified version of the semi-supervised classification method proposed by Zhou et al. [7], closely related to Network Propagation (NP). Unlike previous applications that typically use positive-unlabeled learning frameworks [8], we incorporate both positive (disease-associated) and negative (disease-unrelated) gene sets to improve discrimination capability. This novel approach leverages the core concept that genes associated with the same phenotype tend to interact with each other within biological networks.

2 Methodology

2.1 Semi-Supervised Learning Algorithm

Our algorithm builds upon the semi-supervised classification method introduced by Zhou et al. [7], applying it to gene association analysis. The input consists of:

- A network $G(V, E)$ where nodes V represent genes and edges E represent gene-gene interactions
- Two gene sets: P (disease-related) and N (non-disease-related)

We construct an initial score vector (f^0) for all genes where:

$$f_i^0 = \begin{cases} \frac{1}{|P|} & g_i \in P \\ -\frac{1}{|N|} & g_i \in N \\ 0 & g_i \notin P \wedge g_i \notin N \end{cases} \tag{1}$$

These scores propagate through the network using a modified random walk with restart (RWR) algorithm:

$$f^{t+1} = (1 - \lambda) W f^t + \lambda f^0 \tag{2}$$

Where f^t represents gene scores at step t, λ is the restart coefficient, and W is a normalized weight matrix derived from the network adjacency matrix: $W = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$. Here, A is the adjacency matrix and D is the diagonal degree matrix where $D_{ii} = \sum_j A_{ij}$.

For weighted networks, A_{ij} represents the weight of the edge connecting nodes i and j. The algorithm iterates until convergence, defined as $||f^t - f^{t-1}||_2 < 1 \times 10^{-6}$.

2.2 Gene Sets

We leveraged the authoritative Simons Foundation Autism Research Initiative (SFARI) database for gene selection. This resource categorizes ASD-associated genes according to evidence levels:

Table 1: Categories of autism candidate risk genes from SFARI.

Category	Description	Number of genes
S	Syndromic	143
1	High confidence	25
2	Strong candidate	58
3	Suggestive evidence	176
4	Minimal evidence	405
5	Hypothesized but untested	157
6	Evidence does not support a role	21
–	Uncategorized	88

For our positive class, we selected genes from SFARI categories 1-4, plus those classified as syndromic. Negative class genes were derived from the non-mental genes used by Krishnan et al. [5],

excluding any that overlapped with our positive set. Genes from SFARI categories 5-6 were left unlabeled with initial scores of 0. After converting all gene identifiers to current HGNC symbols, our final dataset contained 739 positive and 1132 negative genes.

2.3 Biological Network

We utilized the STRING database (version 10.5) [9] for protein-protein interaction (PPI) data, which combines experimental interactions with those inferred from text-mining, each assigned a confidence score. We converted protein identifiers to HGNC symbols using Ensembl BioMart [10] and removed redundant interactions, keeping those with the highest confidence scores.

Edge weights were assigned based on confidence scores, and only the largest connected component was retained. The final network contained 18,003 genes connected by 5,007,158 edges.

2.4 Evaluation & Analysis

We evaluated performance using 10-fold cross-validation with the restart parameter λ varying from 0.1 to 0.9 in 0.1 increments. Performance metrics included area under the receiver operating curve (AUROC), area under the precision-recall curve (AUPRC), and Matthews correlation coefficient (MCC).

For biological interpretation, we performed pathway enrichment analysis using the hypergeometric test against the Reactome database through the Enrichr API. All p-values were adjusted for multiple testing using the Benjamini-Hochberg method.

To validate our findings, we cross-referenced our candidate genes with de novo variants from ASD patients in the Simons Simplex Collection available through denovo-db.

3 Results & Discussions

3.1 Performance Evaluation

Cross-validation demonstrated robust performance across varying values of the restart parameter λ. Optimal results were achieved at $\lambda = 0.9$, with an AUROC of 0.85, AUPRC of 0.8, and MCC of 0.57. Notably, performance remained relatively stable across different λ values, indicating algorithm robustness.

Comparative analyses with networks filtered by interaction confidence thresholds showed diminished performance, confirming the value of utilizing the complete weighted network. This suggests that even lower-confidence interactions contribute meaningful information to the classification task.

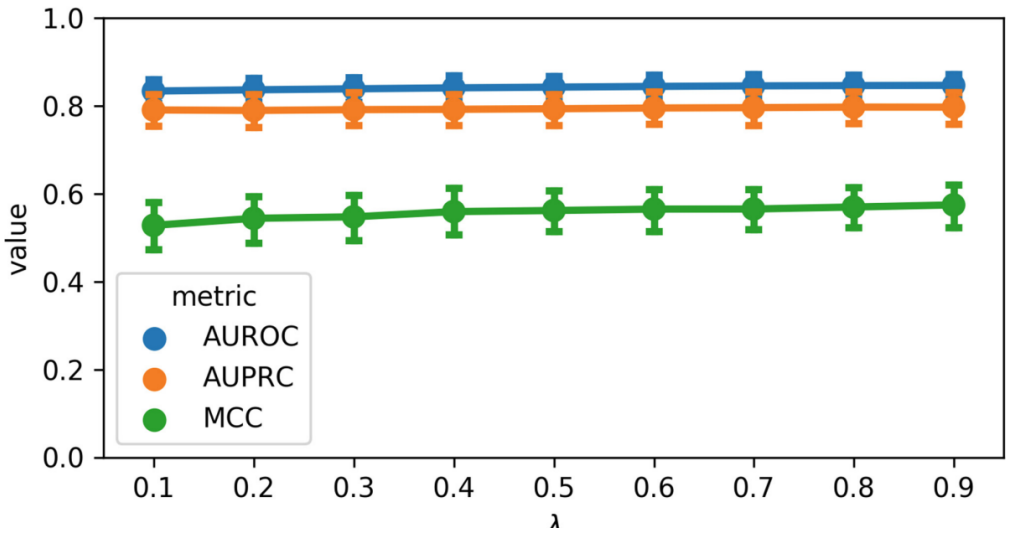

Figure 1: Mean AUROC, AUPRC and MCC values from 10-fold cross validation for various values of lambda (λ). Error bars represent 95% confidence intervals.

3.2 Identification and Analysis of Novel Candidate Genes

Using the 95th percentile score as a threshold, we identified a set of top-ranking candidate genes. As expected, many belonged to SFARI categories 1-4. However, we also identified six genes from category 5 (SYN3, COP1, CBLN1, HTR2A, GABRB1, CLSTN3), eight uncategorized SFARI genes, and 177 novel candidates not previously associated with ASD.

Only one gene from SFARI category 6 (evidence against ASD association) was selected, supporting our method's discriminative capacity. Several of our novel candidates have known associations with other neurological conditions, including X-linked mental retardation, epilepsy, and episodic ataxia.

Pathway enrichment analysis revealed highly significant enrichment in:

- Synaptic transmission pathways ($p = 1.69 \times 10^{-40}$)
- Neuronal system pathways ($p = 1.02 \times 10^{-35}$)
- Neurotransmitter receptor binding and downstream signaling ($p = 3.31 \times 10^{-21}$)
- Glutamate binding and AMPA receptor activity ($p = 1.51 \times 10^{-12}$)
- Various ion channel-related processes

These pathways align with known neurobiological mechanisms in ASD, particularly those related to synaptic function and excitatory/inhibitory balance.

Validation against the Simons Simplex Collection revealed that 70% (123) of our novel candidate genes harbor at least one de novo variant in ASD patients, with 88 genes containing variants in multiple samples. This provides external support for our computational predictions.

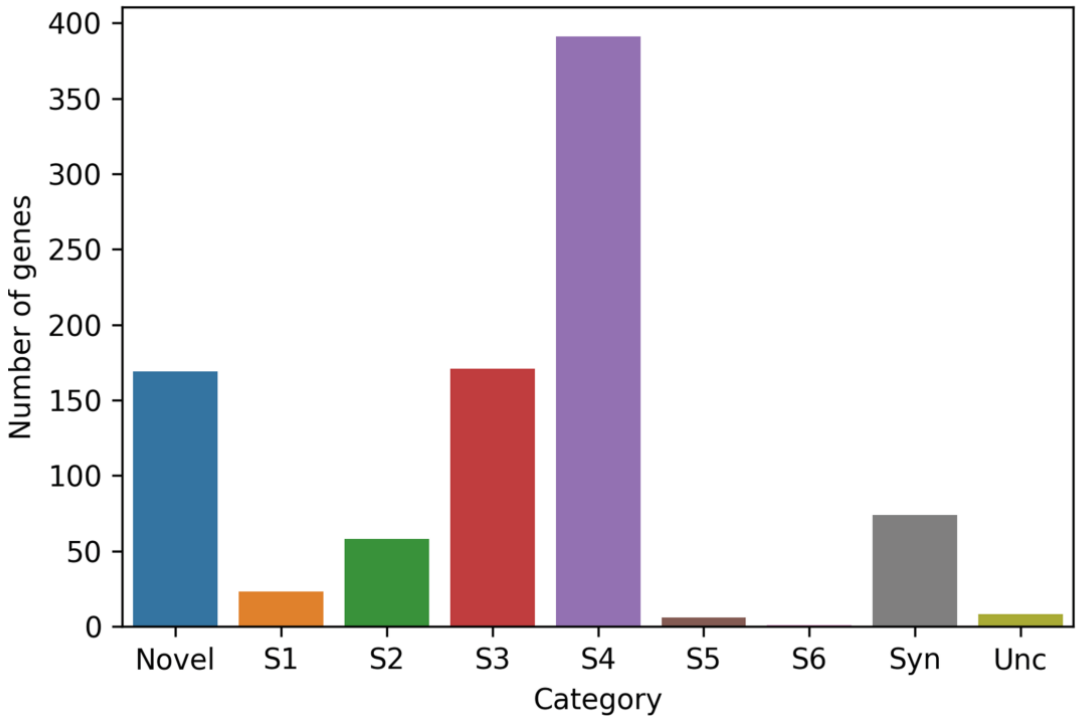

Figure 2: Composition of the selected gene set. S1 to S6 correspond to SFARI gene categories 1 to 6, Syn denotes SFARI genes classified as Syndromic, Unc denotes genes in the SFARI gene list which are not yet categorized, Novel denotes genes not present in SFARI. The bar heights represent the count of genes in each category, with error bars showing the variation across cross-validation folds.

4 Conclusions

We have developed and validated a semi-supervised machine learning approach for identifying novel ASD-associated genes based on network propagation principles. Our method effectively combines protein-protein interaction data with both disease-related and unrelated gene sets to generate meaningful predictions.

Cross-validation demonstrated strong performance (AUROC ~ 0.85), and our identified candidate genes show significant enrichment in neurobiological pathways previously implicated in ASD. External validation confirmed that a large percentage of our novel candidates harbor de novo variants in ASD patients, supporting their potential biological relevance.

This approach offers several advantages:

- Incorporation of both positive and negative training examples
- Ability to leverage the full weighted interaction network
- Identification of genes with biological plausibility based on pathway analysis

These findings may contribute to improved ASD diagnosis and deeper understanding of its molecular mechanisms. Future work will focus on using the derived gene rankings to prioritize variants in genomic analyses of ASD patients, potentially enabling more personalized approaches to diagnosis and intervention.

References

[1] Elsabbagh, M., et al.: Global prevalence of autism and other pervasive developmental disorders. Autism Res. 5(3), 160–179 (2012).

[2] Chaste, P., Roeder, K., Devlin, B.: The Yin and Yang of autism genetics: how rare De Novo and common variations affect liability. Ann. Rev. Genomics Hum. Genet. 18, 167–187 (2017).

[3] Ansel, A., et al.: Variation in gene expression in autism spectrum disorders: an extensive review of transcriptomic studies. Front. Neurosci. 10, 601 (2017).

[4] Vorstman, J., et al.: Autism genetics: opportunities and challenges for clinical translation. Nat. Rev. Genet. 18(6), 362–376 (2017).

[5] Krishnan, A., et al.: Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder. Nat. Neurosci. 19(11), 1454–1462 (2016).

[6] Asif, M., et al.: Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology. PloS one 13(12), e0208626 (2018).

[7] Zhou, D., et al.: Learning with local and global consistency. In: Advances in Neural Information Processing Systems, pp. 321–328 (2004).

[8] Mosca, E., et al.: Network diffusion-based prioritization of autism risk genes identifies significantly connected gene modules. Front. Genet. 8, 129 (2017).

[9] Szklarczyk, D., et al.: The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 45(Database issue), D362–D368 (2017).

[10] Smedley, D., et al.: The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. 43(1), W589–W598 (2015).