IDENTIFICATION OF GENES ASSOCIATED WITH ASD USING RANDOM WALK WITH RESTART (RWR) ALGORITHM

CENTER OF DATA SCIENCE, GOVERNMENT COLLEGE UNIVERSITY FAISALABAD

GROUP MEMBERS

- **Subhan** (5121109)
- **Muhammad Huzaifa** (5121135)

ADVISORS

- **Rabia Shahid** (Supervisor)
- Dr. Tahir Ul Qamar, Dr. Usama Ahmed (Members)

WHAT IS ASD?

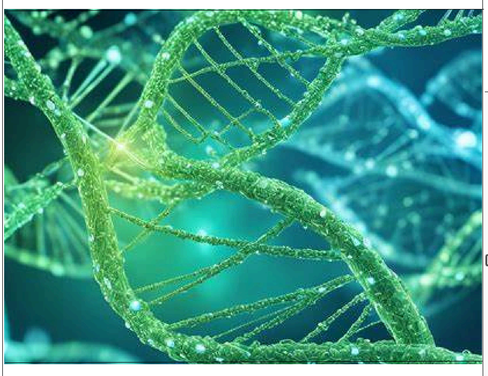
Autism Spectrum Disorder is a heterogeneous neurodevelopmental disorder affecting 1% of the population (78 Million People). Its symptoms are difficulties in Social Interaction, Restricted Behaviors, and Intellectual Disability

PROBLEM STATEMENT

- Current diagnosis relies on clinical observation, not **genetic biomarkers**.
- Existing methods (e.g., SVMs, Random Forests) use **arbitrary weighting** or lack biological network context.
- Only **10-20% of ASD cases** have identifiable genetic causes despite large genomic datasets.

OBJECTIVE

Develop a **Semi-Supervised Learning** model to identify and rank Autism Spectrum Disorder (ASD)-related genes using **Network Propagation** from real-time databases of human genes.



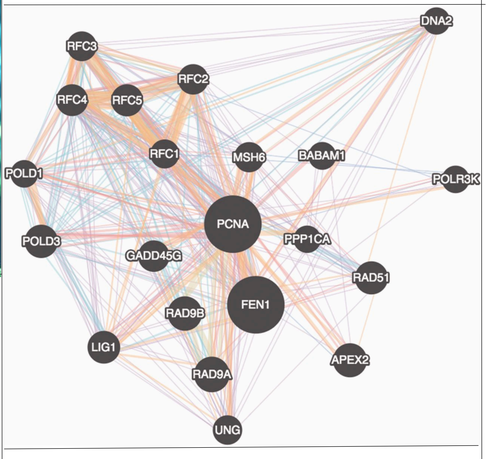
DATASET

PPI Interaction:

- Built from STRING database (v11.5, human proteins)
- Ensembl BioMart API due to the 128.7 GB Size of the file
- Contains 15,908 nodes (genes) and 241,262 edges (interactions)
- Mapped from Ensembl protein IDs to HGNC symbols

SFARI Database:

- Positive Genes are 1,123 genes (ASD-associated genes with scores 1-4 or syndromic)
- Negative genes are 2,954 non-ASD genes



RWR ALGORITHM

In a graph, a random walk proceeds by moving from the current node to one of its neighbors at each step. In Random Walk with Restart, the walker has a probability of jumping back to the starting node at each step, rather than continuing.

Score Initialization:
Assign a score of +1 to each positive seed and -1 to each negative seed; all other nodes start at 0.

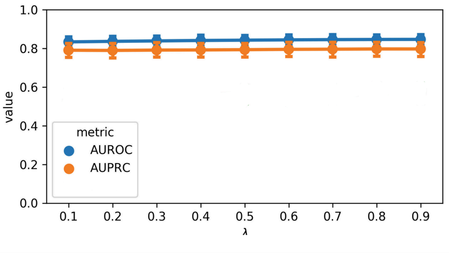
Formula:

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

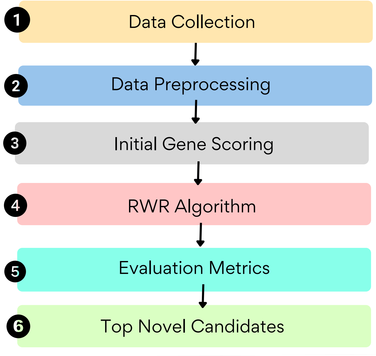
PERFORMANCE EVALUATION

Model Evaluation at **10-Fold** Cross-Validation

- AUROC (Area Under Receiver Operating Curve) = 0.85
- AUPRC (Area Under Precision-Recall Curve) = 0.8
  - Best performance at λ (restart parameter) = 0.9



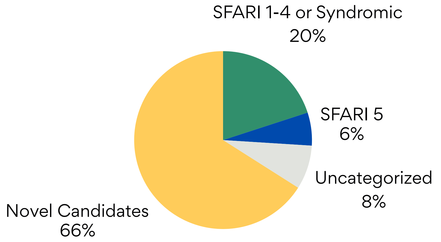
METHODOLOGY



TOP NOVEL CANDIDATES

Out of the top 5% scoring genes (240 genes):

- 63 known SFARI 1–4 + Syndromic Genes
- 6 SFARI 5 genes (e.g., HTR2A, GABRB1)
- 8 uncategorized SFARI genes (e.g., DLGAP3, GRM1)
- 177 novel candidates (70% carry de novo variants in SSC)



FUTURE WORK

- Explore additional datasets for broader applicability.
- Implementing the same strategy for Fragile X Syndrome
- Implementation with Graph Neural Network-based approach

CONCLUSION

This study presents a powerful semi-supervised machine learning approach to identify ASD-associated genes using protein-protein interaction networks. The results align with existing biological knowledge and highlight new genes for future investigation, advancing ASD genetic research.

REFERENCES

1. STRING Database (https://academic.oup.com/nar/article/45/D1/D362/2290901)
2. Ensembl BioMart API (https://grch37.ensembl.org/biomart/martview/bd96879d1ee3994ecb81895981aff072)
3. SFARI Gene Database (https://academic.oup.com/nar/article/45/D1/D362/2290901)
4. Krishnan et al. (Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder)

*All the source code can be found at:
https://github.com/subhan-liaqat/Identification-of-Genes-Associated-with-Autism-Spectrum-Disorder