EXPLORATION AND ANALYSIS OF GENE-DISEASE ASSOCIATIONS

Arun Agarwal*, Anna Yannakopoulos ⁑ , Dr. Arjun Krishnan ⁑
*Temple University, ⁑ Michigan State University

Introduction

- The trillions of cells in the human body are each sustained by the activities of thousands of genes in the genome
- Genetic mutations prevent one or more of these genes from working properly
- Critical gene mutations can disrupt normal development and cause a medical condition
- Many diseases, disorders, traits, and clinical or abnormal human phenotypes arise due to the contribution of disruptions in multiple genes
- To work towards cures for such complex diseases, the human genome has been studied through various experimental methods, each providing a different and sometimes unique understanding of the disease's underlying genetic cause.

Objectives

We hypothesize that:

The genomic basis of complex diseases can be understood on a holistic level through the exploration and combination of genetic data from distinct sources, experimental methods, and association types.

We try to answer:

1. Do different experimental methods provide the same view of the underlying biology or present different aspects of it?
2. Can models trained on one experimental method predict the genes associated to another?

About DisGeNET

In this project, we obtained these data from the **DisGeNET database**, which integrates **human gene-disease associations** and **variant-disease associations** from expert-curated repositories, GWAS catalogs, animal models, and scientific literature to create one of the largest publicly available collections of genes and variants mapping over **30,000 human diseases to 20,000+ genes**.

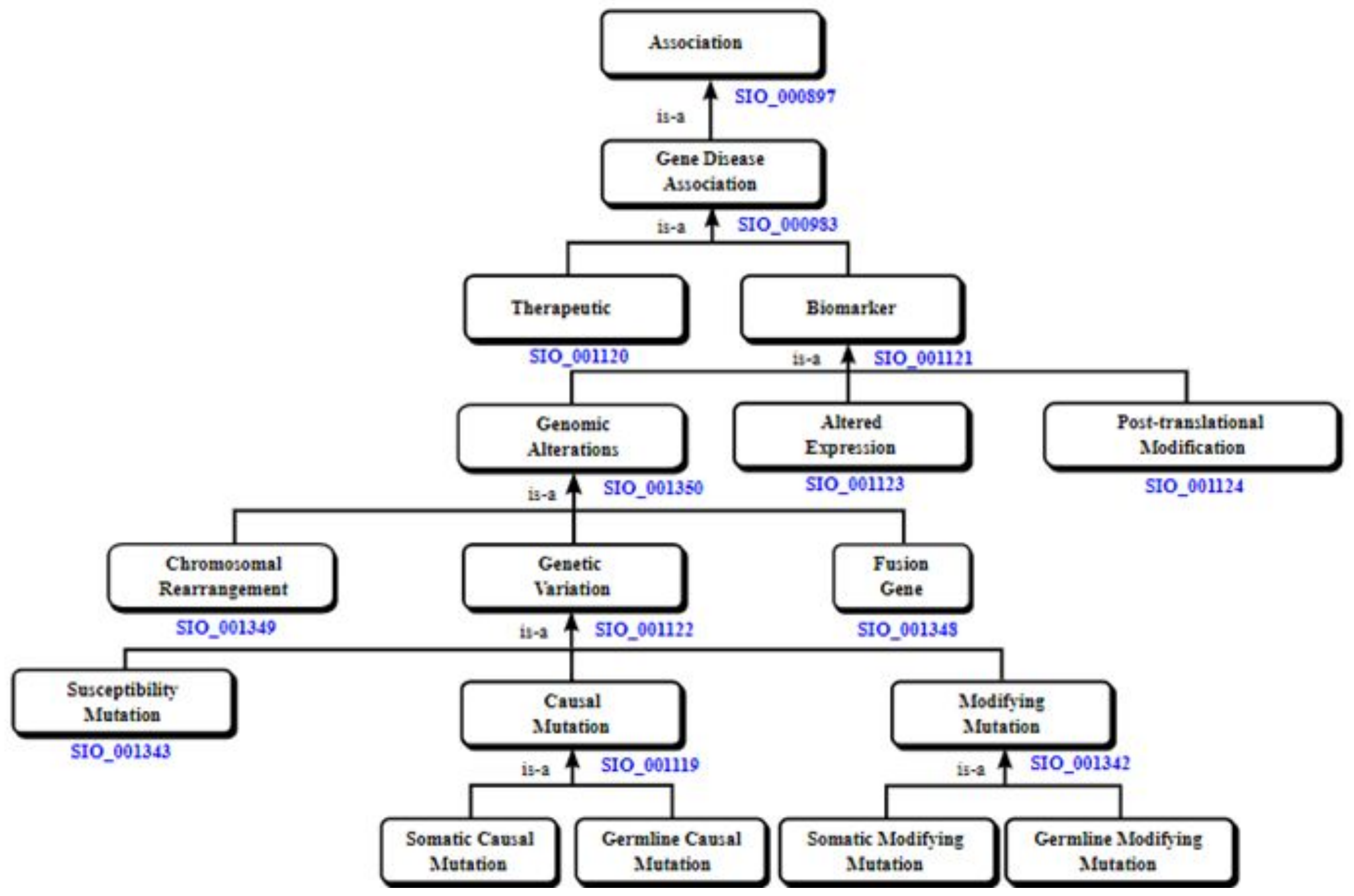

Figure 1: Association Type Hierarchy developed by the DisGeNET Database

Association Type Exploration

- Performed data exploration of the disease-gene data and the correlated association type hierarchy (Figure 1) to understand gene distribution per disease for each source
- **Propagated** the dataset: assigned gene-disease pairs to ancestor association types based on the hierarchy
- Recategorized the dataset into five association types: **Therapeutic**, **Genomic Alterations**, **Altered Expression**, **Post-translational Modification**, **Genetic Variation**, and **Biomarker** for more succinct plots
- Compared gene sets from the same disease—yet separate association types—to quantify the overlap between association types using **Jaccard Similarity Coefficient** (Figure 2)

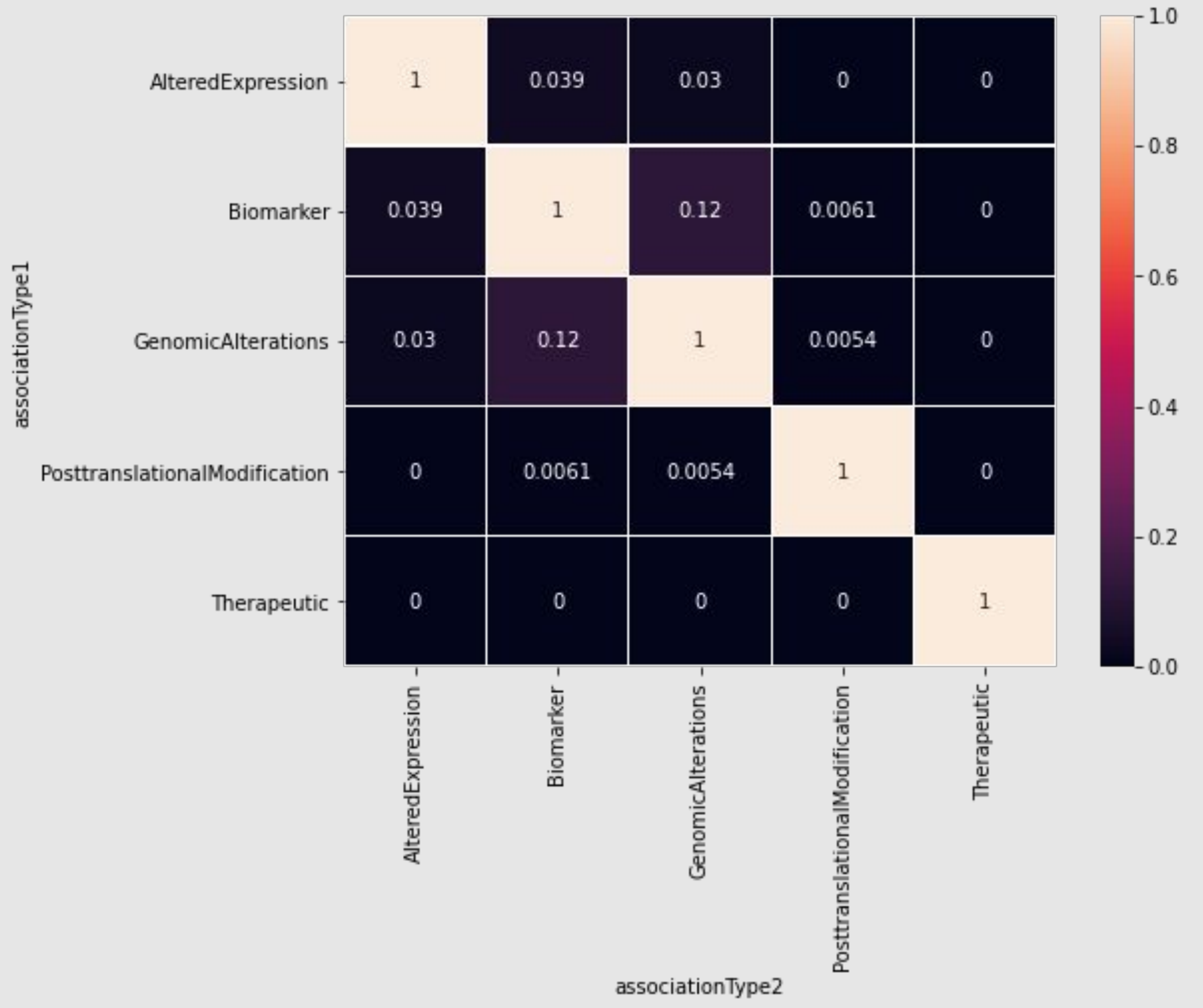

Figure 2: Heatmap of Median Correlation Between Gene Sets for Chosen Association Types

Sources Exploration

- Moved forward with a subset of the dataset containing only **curated data** that compares sources—**CGI**, **CTD-human**, **CLINGEN**, **GENOMICS-ENGLAND**, **ORPHANET**, **PSYGENET**, and **UNIPROT**—instead of association types
- Created another overlap heatmap (Figure 3) and based on all the perfect correlations between source pairs, decided to filter the data to meet conditions like a minimum gene set size

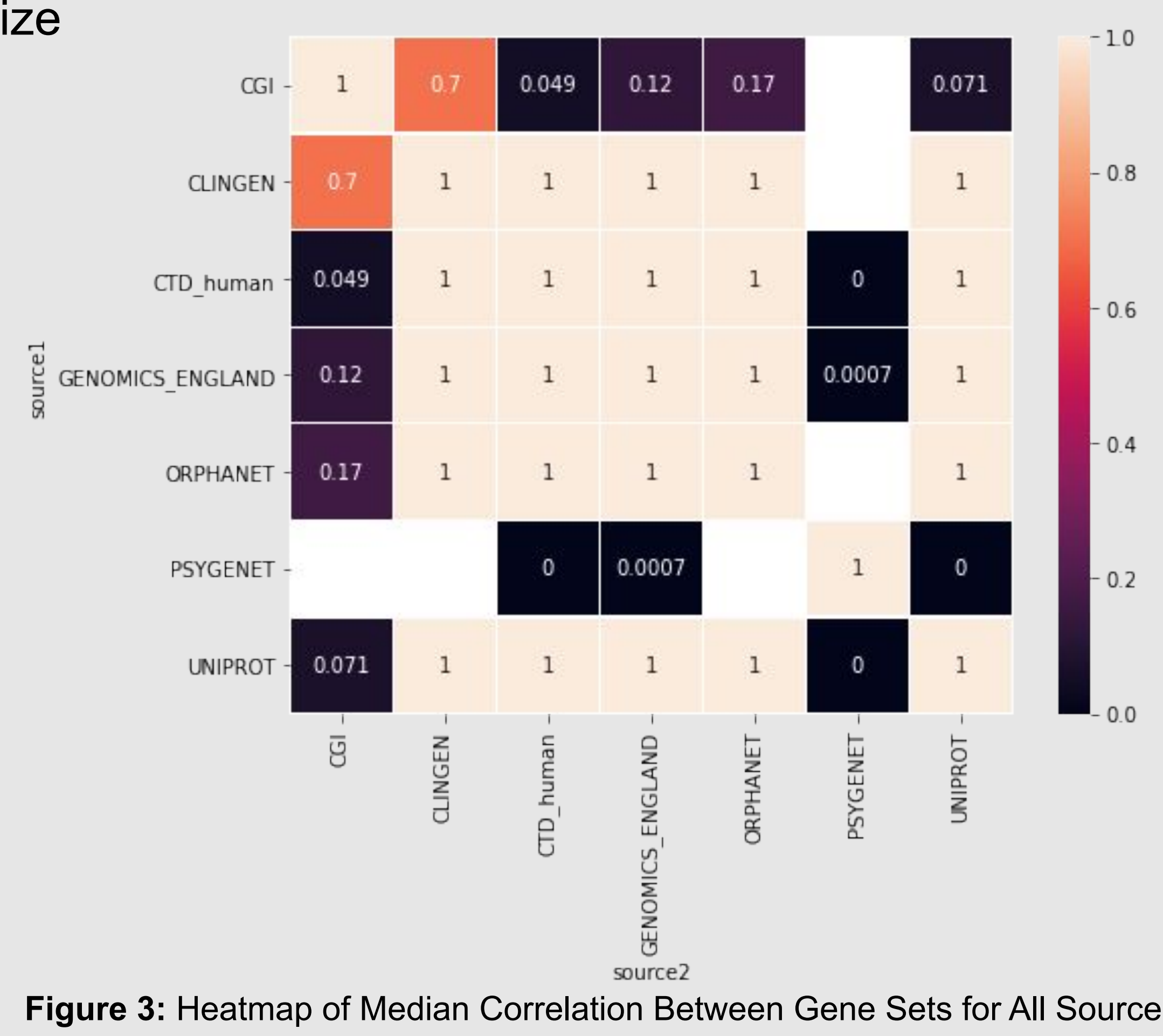

Figure 3: Heatmap of Median Correlation Between Gene Sets for All Sources

Machine Learning Model Approach

Filters on the curated data:

1. Extracted all diseases descending from one of **14 higher-level terms** using a disease ontology
2. Kept data with **at least 10 genes** from **at least 3 sources after propagation** to be able to draw interesting conclusions from multiple sources
3. Retained data with **at least 5 direct gene annotations** to ensure the dataset contains disease-gene relationships directly from DisGeNET and not only from propagation

- Applied a supervised machine learning approach to determine if, for any given disease, models trained on one source can predict genes associated to another source
- Visualized various performance metrics (Figure 4) from our disease-gene prediction model to highlight improvement of our novel framework over a random baseline

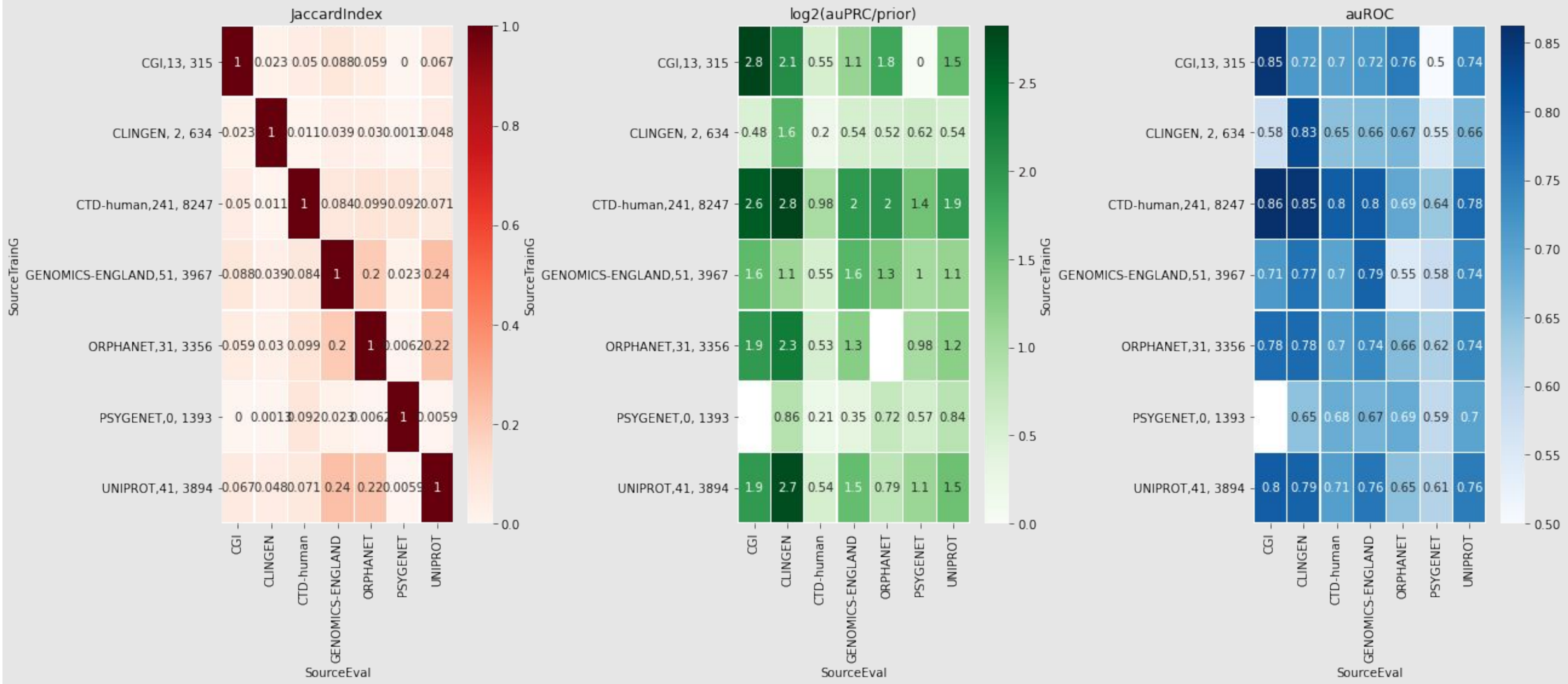

Figure 4: Heatmaps of Median Correlation Score Between Gene Sets, Log2(auPRC/prior) Score, and auROC Score for All Sources

Conclusion/Next Steps:

Overall, we determined some sources and association types demonstrate little correlation to others, represented by their low jaccard index values; however, a few provide similar views of the underlying biology. Furthermore, based on the performance metrics for certain sources trained or evaluated on, models trained on one experimental method can generally predict genes associated to another.

Future Work:

1. Combine the machine learning models from distinct sources into an **ensemble model** that discovers novel genes across the human genome associated with the disease
2. Make my work readily available to members of the Krishnan laboratory to allow for incorporation into further disease study and for continuation of the project

Such results can lead to a better understanding of the genomic basis of complex diseases and in the design of drugs that target and reverse the disease-related signals

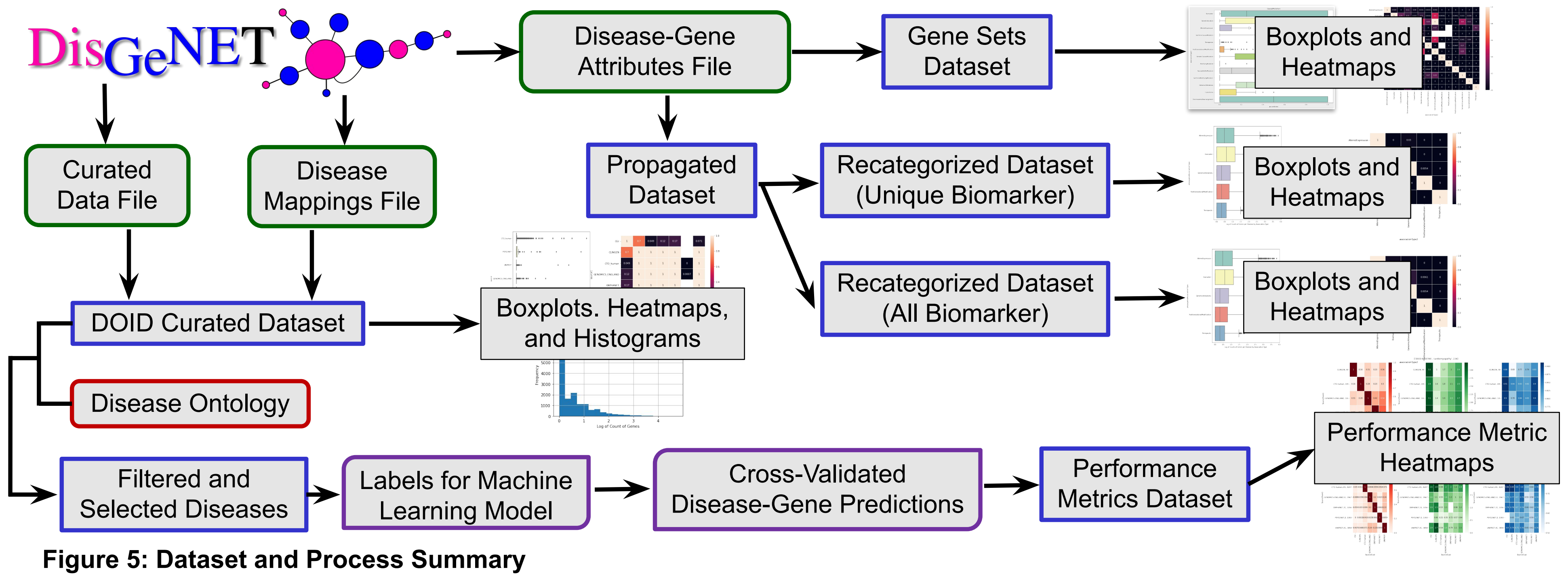
DisGeNET
Disease-Gene Attributes File
Gene Sets Dataset
Boxplots and Heatmaps
Curated Data File
Disease Mappings File
Propagated Dataset
Recategorized Dataset (Unique Biomarker)
Boxplots and Heatmaps
Recategorized Dataset (All Biomarker)
Boxplots and Heatmaps
DOID Curated Dataset
Boxplots, Heatmaps, and Histograms
Disease Ontology
Filtered and Selected Diseases
Labels for Machine Learning Model
Cross-Validated Disease-Gene Predictions
Performance Metrics Dataset
Performance Metric Heatmaps

Figure 5: Dataset and Process Summary

This project would not have been possible without the unwavering support from my mentors, the Krishnan Laboratory, and the ICER ACRES Program Staff.

Contact: arun.agarwal.aaa@gmail.com
https://www.linkedin.com/in/arunagarwal23/