**Project Report: Exploring Gene Causality**

# Project Overview

The goal was to investigate gene-phenotype relationships through the analysis of causal and non-causal gene pair embeddings. The aim of was to find any signals that might aid in differentiating between causal and non-causal gene-phenotype pairs using vector analysis and unsupervised clustering techniques. The goal was to investigate possible separability between the two classes using the embeddings that were created from text descriptions of genes and phenotypes.

# Dataset Overview

The dataset used for this study consisted of 387 gene-phenotype pairs, including both causal and non-causal relationships. The embedding for each pair were pre-generated using the GPT-3.5 text-embedding-3-large model based on textual descriptions of the genes and phenotypes. Key statistics from the exploratory data analysis (EDA) are summarized below:

* Number of Unique Phenotypes: 215
* Number of Unique Causal Genes: 267
* Number of Unique Non-Causal Genes: 291

Top 5 Most Common Phenotypes:

* Type 2 diabetes (type II diabetes mellitus): 38 occurrences
* Breast carcinoma: 16 occurrences
* Type 2 diabetes (adjusted for BMI): 10 occurrences
* Hypertension: 9 occurrences
* Atrial fibrillation: 9 occurrences

The dataset also included information about causal and non-causal genes, with phenotypes linked to multiple gene symbols. For instance, Type 2 diabetes had the highest association with 25 causal genes.

Most frequently occurring Causal Genes:

* HMGCR: 9 occurrences
* F2: 9 occurrences
* PCSK9: 8 occurrences
* IL5: 8 occurrences
* IL12B: 7 occurrences

# Methodology

The study involved two primary analyses: vector analysis and clustering.

In Vector Analysis, the difference vectors between the embeddings of causal and non-causal gene-phenotype pairs were analysed, and the magnitude of these difference vectors was plotted.

Vector Analysis Observations:

* The magnitude of difference vectors between phenotype and gene embeddings was calculated for both causal and non-causal gene-phenotype pairs.
* The magnitude of difference vectors for causal and non-causal pairs was explored. Both distributions broadly overlap, but causal pairs tend to be more frequent in the 0.2 – 0.3 range, while non-causal pairs dominate the 0.3 – 0.5 range.
* There is no strong evidence of separability between causal and non-causal pairs based on the magnitude of difference vectors alone.

In Clustering approach, three methods were explored:

1. **K-Means Clustering**: A partition-based clustering method that tries to divide the dataset into distinct groups by minimizing the within-clustering variance.
2. **Hierarchical Clustering**: A tree-based method that builds a hierarchy of clusters based on the proximity of the data points.
3. **DBSCAN Clustering**: A density-based clustering approach that identifies clusters based on areas of high data point density.

Cluster Analysis Observations:

1. K-Means Clustering Observations:
   1. K-Means clustering showed a relatively homogeneous distribution with moderate separation between causal and non-causal pairs. However, most clusters overlapped, particularly in the centre of the PCA plot, suggesting that K-Means could not strongly differentiate between the two classes.
2. Hierarchical Clustering Observations:
   1. Compared to K-Means, hierarchical clustering produced clusters with more granularity, particularly evident in the distinct separation of certain clusters (example, purple and red clusters).
   2. These more distinct clusters may represent gene-phenotype pairs that have unique features, making them stand out more clearly than others.
   3. Hierarchical clustering’s ability to separate small, unique clusters may be more effective in capturing subtle variations between causal and non-causal relationships.
3. DBSCAN Clustering Observations:
   1. DBSCAN algorithm resulted in only one single cluster. This suggests that DBSCAN was unable to detect distinct clusters based on density, implying that the embeddings do not have significant density variations.
   2. The lack of clusters could indicate that the current embedding space does not possess natural separability between causal and non-causal gene-phenotype pairs based on density alone.
   3. This result emphasizes the homogeneity of the dataset and suggests that density-based clustering methods may not be suitable without further feature engineering or parameter tuning.

# Conclusion

Through the use of vector and cluster analysis, we observed that causal and non-causal gene-phenotype pairs do not exhibit clear separability based on their GPT-3.5-generated embeddings alone. All clustering methods K-Means, Hierarchical Clustering, and DBSCAN struggled to clearly differentiate between the two groups, though hierarchical clustering provided the most granular separation.

These findings highlight the complexity of the gene-phenotype relationships, suggesting that more advanced techniques, such as alternative embedding methods, feature engineering, or dimensionality reduction, may be required to improve separability.