**Exploratory Analysis of Phenotype-Gene Embeddings for Gene Causality Detection**

**1. Problem Statement**

The primary objective of this project was to determine if the combined phenotype and gene embeddings might indicate gene causality. Specifically, we aimed to explore whether the relationship between phenotype embeddings and gene embeddings could reveal signals that differentiate causal genes from non-causal ones for each phenotype. The challenge was to use exploratory data analysis (EDA), visualization, and clustering techniques to identify potential patterns indicating causality.

**2. Approach**

The approach involved mapping phenotypes to genes using embeddings and then analyzing the relationships between the phenotype and gene embeddings through clustering and dimensionality reduction techniques. This was done to investigate potential patterns or clusters that could hint at gene causality.

* **Step 1: Data Preparation**: The data used for this analysis included phenotype and gene embeddings. Each phenotype was associated with one causal gene and several non-causal genes.
* **Step 2: Feature Engineering**: Two types of embeddings, phenotype and gene embeddings, were retrieved, each containing 3072-dimensional feature vectors. We also created a "causal gene flag" to identify whether a gene was causal (1) or non-causal (0).
* **Step 3: Preprocessing**: Data preprocessing was applied to ensure that embeddings were properly formatted as numeric arrays. The combined dataset was created by horizontally stacking phenotype and gene embeddings to facilitate analysis across the two embeddings simultaneously.

**3. Dimensionality Reduction**

To better visualize and interpret the high-dimensional data, dimensionality reduction techniques such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) were applied:

* **PCA**: Applied initially to reduce the 3072-dimensional embeddings to 2 components for visualization. PCA helped in understanding the explained variance and capturing the maximum variance directions.
* **t-SNE**: This was also applied to both gene and phenotype embeddings, offering a nonlinear reduction that helped in better visualizing clusters and the distribution of points in a lower-dimensional space. This technique was especially useful to visualize more complex relationships in embeddings.

**4. Visualizations**

Various visualizations were conducted to understand the distributions and relationships of phenotype and gene embeddings:

* **Distribution Plots**: Count plots and pair plots were created to understand the distribution of causal and non-causal genes. This revealed a significant imbalance between causal and non-causal genes.
* **Correlation Heatmaps**: Correlation heatmaps for gene and phenotype embeddings were plotted to identify any highly correlated features across dimensions, which could indicate redundancy.
* **Scatter Plots Using PCA and t-SNE**: Reduced embeddings were visualized using scatter plots to identify separable patterns between causal and non-causal genes.

**5. Clustering**

We employed different clustering algorithms to explore relationships between phenotype-gene embeddings:

* **KMeans Clustering**: KMeans was used to create clusters based on combined phenotype-gene embeddings. The clusters were visualized using t-SNE reduced dimensions, providing well-separated clusters that were easy to interpret.
* **DBSCAN Clustering**: DBSCAN, a density-based clustering method, was applied to explore high-density regions and potential outliers. We ran DBSCAN both with and without scaling to analyze the effect of standardization. Without scaling, multiple smaller clusters emerged, indicating variations in density, while with scaling, the algorithm identified mostly noise points, highlighting the importance of scaling for distance-based clustering.

**6. Results and Findings**

* **Cluster Visualization and Interpretation**: KMeans provided distinct clusters across the combined phenotype-gene embeddings, suggesting that embedding features can potentially form coherent groupings. DBSCAN, however, identified smaller dense clusters and several noise points, which could indicate outlier embeddings or subtle patterns related to gene causality.
* **Comparison Between Techniques**: t-SNE provided a more nonlinear and informative visualization compared to PCA. It was better at separating points in the embedding space, and this helped in clearly understanding the clusters formed by KMeans and DBSCAN.

**7. Conclusion and Next Steps**

* **Conclusions**: The exploration revealed that clustering combined phenotype-gene embeddings could indicate signals for gene causality. KMeans clustering provided better and more distinct groupings, while DBSCAN highlighted areas of high density and potential outliers.
* **Next Steps**: Further analysis should involve evaluating the cluster quality using metrics like silhouette score and purity to determine the clustering effectiveness in identifying causal genes. Additionally, incorporating supervised models for validation, where cluster labels can be used as features, may provide insights into whether clustering results align with known causal relationships.