# Appendix A: Mathematical and Topological Framework

## 1 Formal Problem Statement

### 1.1 Gene Enrichment Space

Let $\mathcal{G} = \{g_1, \ldots, g_N\}$ denote a set of $N$ genes ($N = 36$ in our analysis). For each gene $g_i$ and disorder $d \in \{\text{ADHD}, \text{ASD}\}$, we define an enrichment score:

$$E_d(g_i) = -\log_{10}(p_d(g_i)) \tag{1}$$

where $p_d(g_i)$ is the gene-level $p$-value from MAGMA analysis of disorder $d$.

### 1.2 Shared Enrichment Metric

The shared enrichment for gene $g_i$ across ADHD and autism is defined using the geometric mean:

$$E_{\text{shared}}(g_i) = \sqrt{E_{\text{ADHD}}(g_i) \cdot E_{\text{ASD}}(g_i)} \tag{2}$$

**Justification for geometric mean:**

1. **Balanced contribution**: For two values $a, b$ with $a < b$, the geometric mean $\sqrt{ab}$ is bounded by:
$$a \leq \sqrt{ab} \leq \frac{a+b}{2} \leq b \tag{3}$$
2. **Multiplicative interpretation**: Equivalent to arithmetic mean in log-space:
$$\sqrt{ab} = \exp\left(\frac{\log a + \log b}{2}\right) \tag{4}$$

3. **Penalizes imbalance**: Maximized when $a = b$:

$$\frac{\partial}{\partial a}\sqrt{ab}\bigg|_{a=b} = \frac{b}{2\sqrt{ab}} = \frac{1}{2}\sqrt{\frac{b}{a}} \to \infty \text{ as } a \to 0 \tag{5}$$

### 1.3 Feature Space Construction

Each gene $g_i$ is represented by a feature vector in pathway space:

$$\mathbf{f}(g_i) = \begin{bmatrix} E_{\text{DA}}(g_i) \\ E_{\text{5HT}}(g_i) \\ E_{\text{Glu}}(g_i) \\ E_{\text{GABA}}(g_i) \end{bmatrix} \in \mathbb{R}^4 \tag{6}$$

where subscripts denote dopaminergic (DA), serotonergic (5HT), glutamatergic (Glu), and GABAergic pathways.

**Standardization**: Features are z-score normalized:

$$\tilde{\mathbf{f}}(g_i) = \frac{\mathbf{f}(g_i) - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \tag{7}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are component-wise mean and standard deviation across all genes.

## 2 Topological Structure of Enrichment Space

### 2.1 Enrichment Manifold

The set of gene enrichment profiles forms a submanifold $\mathcal{M} \subset \mathbb{R}^4$:

$$\mathcal{M} = \{\tilde{\mathbf{f}}(g_i) : g_i \in \mathcal{G}\} \tag{8}$$

**Ambient space**: $\mathcal{M} \subset \mathbb{R}^4$ with standard Euclidean metric

**Intrinsic dimension**: Estimated via local PCA or persistent homology

### 2.2 Metric Structure

We endow $\mathcal{M}$ with the induced Euclidean metric:

$$d(\mathbf{f}(g_i), \mathbf{f}(g_j)) = \|\tilde{\mathbf{f}}(g_i) - \tilde{\mathbf{f}}(g_j)\|_2 \tag{9}$$

This metric satisfies:

1. **Positivity**: $d(\mathbf{f}_i, \mathbf{f}_j) \geq 0$ with equality iff $i = j$
2. **Symmetry**: $d(\mathbf{f}_i, \mathbf{f}_j) = d(\mathbf{f}_j, \mathbf{f}_i)$
3. **Triangle inequality**: $d(\mathbf{f}_i, \mathbf{f}_k) \leq d(\mathbf{f}_i, \mathbf{f}_j) + d(\mathbf{f}_j, \mathbf{f}_k)$

### 2.3 Stratification vs. Clustering

**Traditional clustering assumption** (NOT applicable here):

$$\mathcal{M} = \bigsqcup_{k=1}^{K} \mathcal{C}_k \tag{10}$$

where $\mathcal{C}_k$ are disjoint, well-separated components.

**Our model** (enrichment stratification):

$$\mathcal{M} = \bigcup_{k=1}^{K} \mathcal{S}_k \tag{11}$$

where $\mathcal{S}_k$ are overlapping high-density regions ("strata") along a continuum.

**Mathematical distinction**:

- Clustering: $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$ for $i \neq j$
- Stratification: $\mathcal{S}_i \cap \mathcal{S}_j \neq \emptyset$ possible

## 3 K-Means Clustering Algorithm

### 3.1 Objective Function

K-means minimizes within-cluster sum of squares (WCSS):

$$\arg\min_{\{\mathcal{C}_1,\ldots,\mathcal{C}_K\}} \sum_{k=1}^{K} \sum_{g_i \in \mathcal{C}_k} \|\tilde{\mathbf{f}}(g_i) - \boldsymbol{\mu}_k\|^2 \tag{12}$$

where $\boldsymbol{\mu}_k = \frac{1}{|\mathcal{C}_k|} \sum_{g_i \in \mathcal{C}_k} \tilde{\mathbf{f}}(g_i)$ is the centroid of cluster $k$.

### 3.2 Lloyd's Algorithm

**Initialization**: Random selection of $K$ initial centroids

**Iteration** (until convergence):

1. **Assignment step**: For each gene $g_i$, assign to nearest centroid:

$$c(g_i) = \arg\min_{k \in \{1,\ldots,K\}} \|\tilde{\mathbf{f}}(g_i) - \boldsymbol{\mu}_k\|^2 \tag{13}$$

2. **Update step**: Recompute centroids:

$$\boldsymbol{\mu}_k \leftarrow \frac{1}{|\mathcal{C}_k|} \sum_{g_i : c(g_i) = k} \tilde{\mathbf{f}}(g_i) \tag{14}$$

**Convergence**: Guaranteed to converge to a local minimum (not necessarily global).
**Complexity**: $O(NKdT)$ where $N =$ genes, $K =$ clusters, $d =$ dimensions, $T =$ iterations

### 3.3 Cluster Quality Metrics

**Silhouette coefficient** for gene $g_i$ in cluster $\mathcal{C}_k$:

$$s(g_i) = \frac{b(g_i) - a(g_i)}{\max\{a(g_i), b(g_i)\}} \tag{15}$$

where:

- $a(g_i) = \frac{1}{|\mathcal{C}_k| - 1} \sum_{g_j \in \mathcal{C}_k, j \neq i} d(g_i, g_j)$ (mean intra-cluster distance)
- $b(g_i) = \min_{l \neq k} \frac{1}{|\mathcal{C}_l|} \sum_{g_j \in \mathcal{C}_l} d(g_i, g_j)$ (mean nearest-cluster distance)

**Average silhouette**:

$$\bar{s} = \frac{1}{N} \sum_{i=1}^{N} s(g_i) \tag{16}$$

**Interpretation**:

- $s(g_i) \approx 1$: Well-clustered (close to own cluster, far from others)
- $s(g_i) \approx 0$: On cluster boundary
- $s(g_i) < 0$: Likely misassigned

## 4 Validation Framework

### 4.1 Permutation Test

**Null hypothesis** $H_0$: Observed clustering structure is no better than random.

**Test statistic**: $T = \bar{s}$ (average silhouette score)

**Procedure**:

1. Compute observed $T_{\text{obs}}$ from real data
2. For $b = 1, \ldots, B$ permutations:
   (a) Randomly shuffle enrichment values within each pathway
   (b) Re-cluster with same $K$
   (c) Compute $T^{(b)}$
3. Calculate $p$-value:

$$p = \frac{1 + \sum_{b=1}^{B} \mathbb{1}(T^{(b)} \geq T_{\text{obs}})}{B + 1} \tag{17}$$

**Our result**: $p = 0.974$ with $B = 1000$

**Interpretation**: Clustering not significantly better than random. However, this test assumes:

- Exchangeability of enrichment values (violated: pathway structure)
- Discrete well-separated clusters (violated: continuous stratification)
- Sufficient sample size (violated: $N = 36$)

## 4.2 Bootstrap Stability

**Measure**: Adjusted Rand Index (ARI) between clusterings of bootstrap samples

**Procedure**:

1. For $b = 1, \ldots, B$ bootstrap iterations:
   (a) Sample $N$ genes with replacement: $\mathcal{G}^{(b)}$
   (b) Cluster $\mathcal{G}^{(b)}$ with $K$ clusters
   (c) Record assignments $\mathbf{c}^{(b)}$
2. For each gene $g_i$, calculate stability:

$$\text{Stability}(g_i) = \frac{1}{B(B-1)/2} \sum_{b<b'} \mathbb{1}(c_i^{(b)} = c_i^{(b')}) \tag{18}$$

3. Average across genes:

$$\bar{\text{Stability}} = \frac{1}{N}\sum_{i=1}^{N}\text{Stability}(g_i) \tag{19}$$

**Our result**: $\bar{\text{Stability}} = 0.40$ with $B = 1000$

**Interpretation**: Only 40% stability (threshold: 0.75). Indicates:

- Gene assignments uncertain, especially borderline cases
- Polygenetic pattern (23 genes) contributes to instability
- Reflects continuous nature of enrichment distribution

## 4.3 Cross-Validation

**Leave-one-out cross-validation (LOOCV)**:

For each gene $g_i$:

1. Remove $g_i$ from dataset: $\mathcal{G}_{-i} = \mathcal{G} \setminus \{g_i\}$
2. Cluster $\mathcal{G}_{-i}$ with $K$ clusters
3. Compute silhouette score $\bar{s}_{-i}$

**Stability metric**:

$$\Delta_{\text{CV}} = \frac{1}{N}\sum_{i=1}^{N}|\bar{s} - \bar{s}_{-i}| \tag{20}$$

**Our result**: $\Delta_{\text{CV}} = 0.003$ (mean absolute change)

**Interpretation**: Clustering highly stable to individual gene removal. Not driven by outliers.

## 4.4 Independent Cross-Disorder Validation

**Correlation analysis** between original enrichment and independent signals:

For each gene $g_i$, let:

- $E_{\text{shared}}(g_i)$ = original shared enrichment
- $S_{\text{cross}}(g_i)$ = mean number of genome-wide significant SNPs across 11 cross-disorder studies

**Pearson correlation**:

$$r = \frac{\sum_{i=1}^{N}(E_{\text{shared}}(g_i) - \bar{E})(S_{\text{cross}}(g_i) - \bar{S})}{\sqrt{\sum_{i=1}^{N}(E_{\text{shared}}(g_i) - \bar{E})^2}\sqrt{\sum_{i=1}^{N}(S_{\text{cross}}(g_i) - \bar{S})^2}} \tag{21}$$

**Our result**: $r = 0.898$, $p = 1.06 \times 10^{-13}$ ($N = 36$ genes)

**Spearman rank correlation**:

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2 - 1)} \tag{22}$$

where $d_i$ is the difference in ranks for gene $i$.

**Our result**: $\rho = 0.782$, $p = 1.06 \times 10^{-13}$

**Interpretation**: Strong correlation with entirely independent data provides evidence for biological validity despite failed clustering tests.

# 5 Statistical Power Analysis

## 5.1 Power for Correlation Detection

Given $N = 36$ genes, the minimum detectable correlation at $\alpha = 0.05$, power $= 0.80$:

$$r_{\min} = \sqrt{\frac{(z_{\alpha/2} + z_{\beta})^2}{N - 3}} \tag{23}$$

where $z_{\alpha/2} = 1.96$ and $z_{\beta} = 0.84$.

**Calculation**: $r_{\min} = \sqrt{\frac{(1.96+0.84)^2}{35-3}} = \sqrt{\frac{7.84}{32}} \approx 0.495$

**Interpretation**: Our observed $r = 0.913$ far exceeds minimum detectable effect, indicating high statistical power.

## 5.2 Power for Clustering Validation

For permutation test with $B = 1000$ permutations:

**Minimum detectable effect size** (Cohen's $d$):

$$d_{\min} = \frac{z_{\alpha} + z_{\beta}}{\sqrt{N/2}} \tag{24}$$

For $N = 36$: $d_{\min} = \frac{1.96+0.84}{\sqrt{35/2}} \approx 0.67$

**Interpretation**: Medium-to-large effects detectable, but small $N$ limits power for subtle clustering patterns.

# 6 Computational Complexity

## 6.1 K-Means Algorithm

**Time complexity**: $O(NKdT)$

- $N = 35$ genes
- $K = 5$ clusters
- $d = 4$ dimensions (pathways)
- $T \approx 100$ iterations (typical convergence)

**Total operations**: $35 \times 5 \times 4 \times 100 = 70,000$

**Space complexity**: $O(Nd + K) = O(35 \times 4 + 5) = O(145)$

## 6.2 Validation Procedures

**Permutation test**: $O(BNKdT) = O(1000 \times 70,000) = O(7 \times 10^7)$ operations

**Bootstrap**: $O(BNKdT) = O(1000 \times 70,000) = O(7 \times 10^7)$ operations

**Cross-validation**: $O(N^2KdT) = O(35^2 \times 5 \times 4 \times 100) = O(2.45 \times 10^6)$ operations

**Total computational cost**: Approximately $10^8$ operations, feasible on standard hardware.

# 7 Manifold Learning Perspective

## 7.1 Intrinsic Dimensionality

The effective dimensionality of enrichment manifold $\mathcal{M}$ may be lower than ambient dimension $d = 4$.

**Local PCA estimate**: For each point $\mathbf{f}(g_i)$ and its $k$-nearest neighbors, compute local covariance matrix $\mathbf{C}_i$.

**Intrinsic dimension**:

$$\hat{d}_{\text{int}} = \arg\max_{d'} \left\{ \frac{\sum_{j=1}^{d'} \lambda_j}{\sum_{j=1}^{d} \lambda_j} \geq 0.95 \right\} \tag{25}$$

where $\lambda_1 \geq \ldots \geq \lambda_d$ are eigenvalues of $\mathbf{C}_i$.

### 7.2 Geodesic Distance

True distances on manifold may differ from Euclidean distances in $\mathbb{R}^4$:

**Geodesic distance** between $g_i$ and $g_j$:

$$d_{\mathcal{M}}(g_i, g_j) = \inf_{\gamma} \int_0^1 \|\gamma'(t)\| \, dt \tag{26}$$

where $\gamma : [0, 1] \to \mathcal{M}$ is a path connecting $\mathbf{f}(g_i)$ and $\mathbf{f}(g_j)$.

**Approximation**: Isomap or diffusion maps could estimate $d_{\mathcal{M}}$.

## 8 Alternative Geometric Interpretations

### 8.1 Mixture Model Perspective

Instead of hard k-means clustering, enrichment could arise from mixture of Gaussians:

$$p(\mathbf{f}(g_i)) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{f}(g_i) | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{27}$$

where $\pi_k$ are mixing proportions, $\boldsymbol{\mu}_k$ are means, $\boldsymbol{\Sigma}_k$ are covariances.

**Soft assignment** via posterior probability:

$$\gamma_{ik} = \frac{\pi_k \mathcal{N}(\mathbf{f}(g_i) | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{K} \pi_l \mathcal{N}(\mathbf{f}(g_i) | \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \tag{28}$$

**Interpretation**: Would better capture overlapping strata and uncertainty in assignments.

### 8.2 Density-Based Perspective

**Mode-seeking interpretation**: Enrichment patterns correspond to local maxima of density $p(\mathbf{f})$.

**Mean-shift algorithm**:

$$\mathbf{m}(\mathbf{x}) = \frac{\sum_{i=1}^{N} \mathbf{f}(g_i) K(\|\mathbf{x} - \mathbf{f}(g_i)\|)}{\sum_{i=1}^{N} K(\|\mathbf{x} - \mathbf{f}(g_i)\|)} \tag{29}$$

where $K(\cdot)$ is a kernel function.

**Advantage**: No assumption of spherical clusters or fixed number of patterns.

# 9 Limitations and Future Directions

## 9.1 Sample Size

With $N = 36$ genes, statistical power is limited for:

- Detecting subtle clustering structure
- Robustly estimating cluster boundaries
- Validating via resampling methods

**Recommendation**: Expand analysis to genome-wide scale ($N > 10,000$ genes).

## 9.2 Feature Engineering

Current features (pathway-level enrichment) may not capture:

- Gene-gene interactions
- Tissue-specific expression
- Developmental timing
- Regulatory networks

**Recommendation**: Integrate multi-omic data (expression, chromatin, protein-protein interactions).

## 9.3 Alternative Clustering Methods

K-means assumes:

- Spherical clusters
- Similar cluster sizes
- Linear separability

**Alternatives to explore**:

- Hierarchical clustering with linkage criteria

- DBSCAN for density-based patterns
- Gaussian mixture models for soft assignments
- Spectral clustering for non-convex shapes

## 9.4 Biological Validation

Mathematical framework requires experimental validation:

- Functional studies of genes within patterns
- Patient stratification based on genetic profiles
- Treatment response prediction
- Animal model phenotypes