Gut Microbiota in children with Autism Spectrum Disorder

Ruchir R

Department of Health Technology, Technical University of Denmark

Exchange student from Indian Institure of Technology, Madras, India

May 12, 2025

Abstract

Performing Compositional Data Analysis (CoDa) on gut microbiome profiles from children with Autism Spectrum Disorder (ASD) and neurotypical controls. Using log-ratio transformations and statistical testing in Aitchison geometry, to identify key microbial differences.

1 Introduction

Autism Spectrum Disorder (ASD) has been linked to disruptions in the gut-brain axis. Metagenomic studies reveal microbiome shifts in ASD, but the compositional constraints of relative abundance data require specialized analysis. CoDa methods, grounded in Aitchison geometry, enable accurate interpretations of such data.

2 Data Overview and Preprocessing

2.1 Data Description

The dataset consists of microbial counts for genera from stool samples of children in two groups (initially not known which is which):

- Individuals diagnosed with ASD
- Neurotypical controls

Raw count data are normalized by read length. Below is Log plot of the arithmetic means with the number of zeroes next to the labels for intuitive grasp of the data.

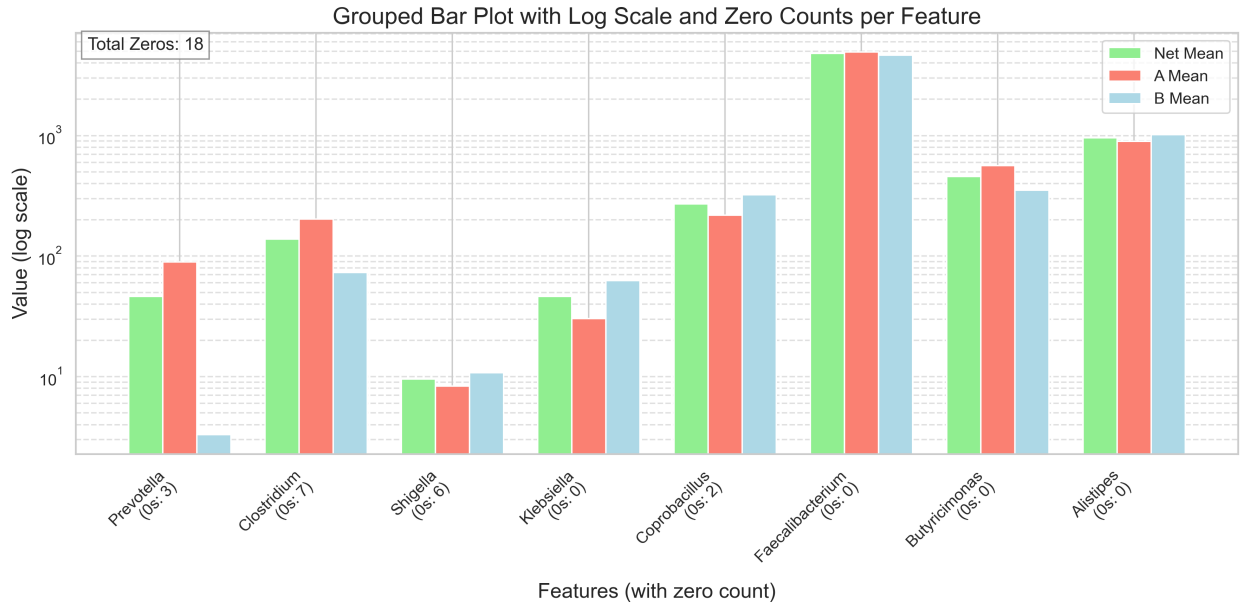

Figure 1: Grouped Bar plot with Log scale. Under the Feature labels the numbers of zeroes for that group in the data is given.

2.2 Zero Replacement

Zero values hinder log-ratio transformations, as these require strictly positive inputs. In metagenomic data, such zeros often stem from rounding or detection limits, especially after normalization. In this dataset, where zeros account for less than 10%, this is a plausible assumption.

First considering a simple imputation strategy, replacing zeros with a small constant ($\delta_i = 0.5$) and rescaling the non-zero parts:

$$x_i' = \begin{cases} \delta_i & \text{if } x_i = 0 \\ x_i \left(1 - \frac{1}{s} \sum_{k|x_k=0} \delta_k\right) & \text{if } x_i > 0 \end{cases} \tag{1}$$

Here, s is the total sum of the composition. While straightforward, this method may not reflect the true variability or uncertainty in low-abundance taxa.

Instead, a Bayesian zero-replacement method based on the sparse Dirichlet distribution is used, which models a distribution over proportions:

$$\text{Dirichlet}(\mathbf{r}, \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{D} r_i^{\alpha_i - 1}, \quad B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{D} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{D} \alpha_i\right)} \tag{2}$$

The Dirichlet distribution generates random compositions that sum to 1. The concentration parameters α_i determine how values are spread: small values (e.g., $\alpha_i = 0.5$) favor sparsity, producing vectors where a few components dominate and others are close to zero. This matches biological intuition, where some taxa dominate the microbial community.

Thus, by sampling from a Dirichlet distribution with $\alpha_i = 0.5$, zeros are replaced with small positive values that better reflect the uncertainty and distributional nature of the data while preserving the compositional structure.

Non-parametric replacement was used for descriptive statistical analysis and the visualizations on the simplex. The Bayesian equivalent was used for PCA and ANOVA analysis. This was decided based off trial and error, and is mostly arbitrary as they bothyield very similar results.

3 Descriptive statistics

3.1 Geometric Mean

Analogous to the arithmetic mean in Euclidian space, the geometric mean is the center of the sample of the compositions and is described by the following equation for n samples and a composition of D parts.

$$\hat{g}_j = \left(\prod_{i=1}^{n} x_{i,j} \right)^{1/n}, \quad j = 1, 2, \ldots, D \tag{3}$$

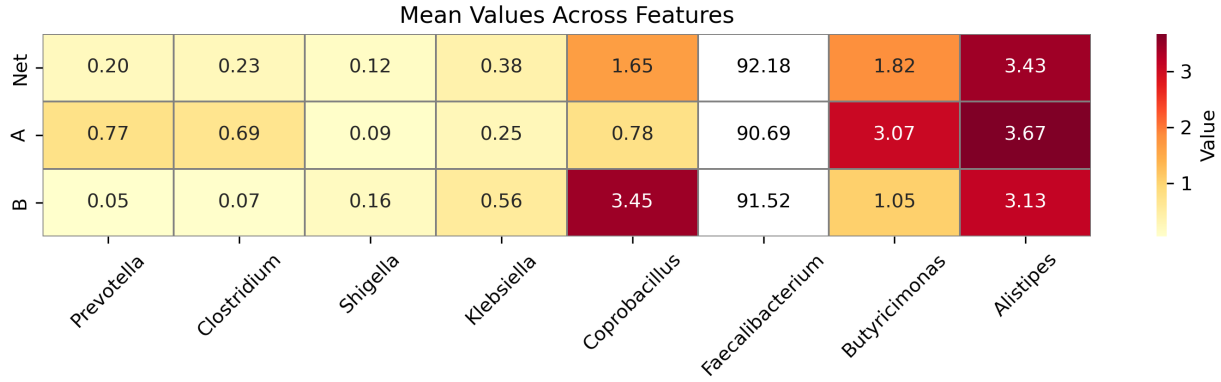

Figure 2: Geometric mean values with the color of Faecalibacterium masked out for readability.

3.2 Composition Variation Matrix

The pairwise log-ratio variance matrix highlights genera that co-vary across samples. Total variation provides an overall summary statistic.

$$T_{ij} = \mathrm{Var}\left[\ln\left(\frac{x_i}{x_j} \right) \right]$$

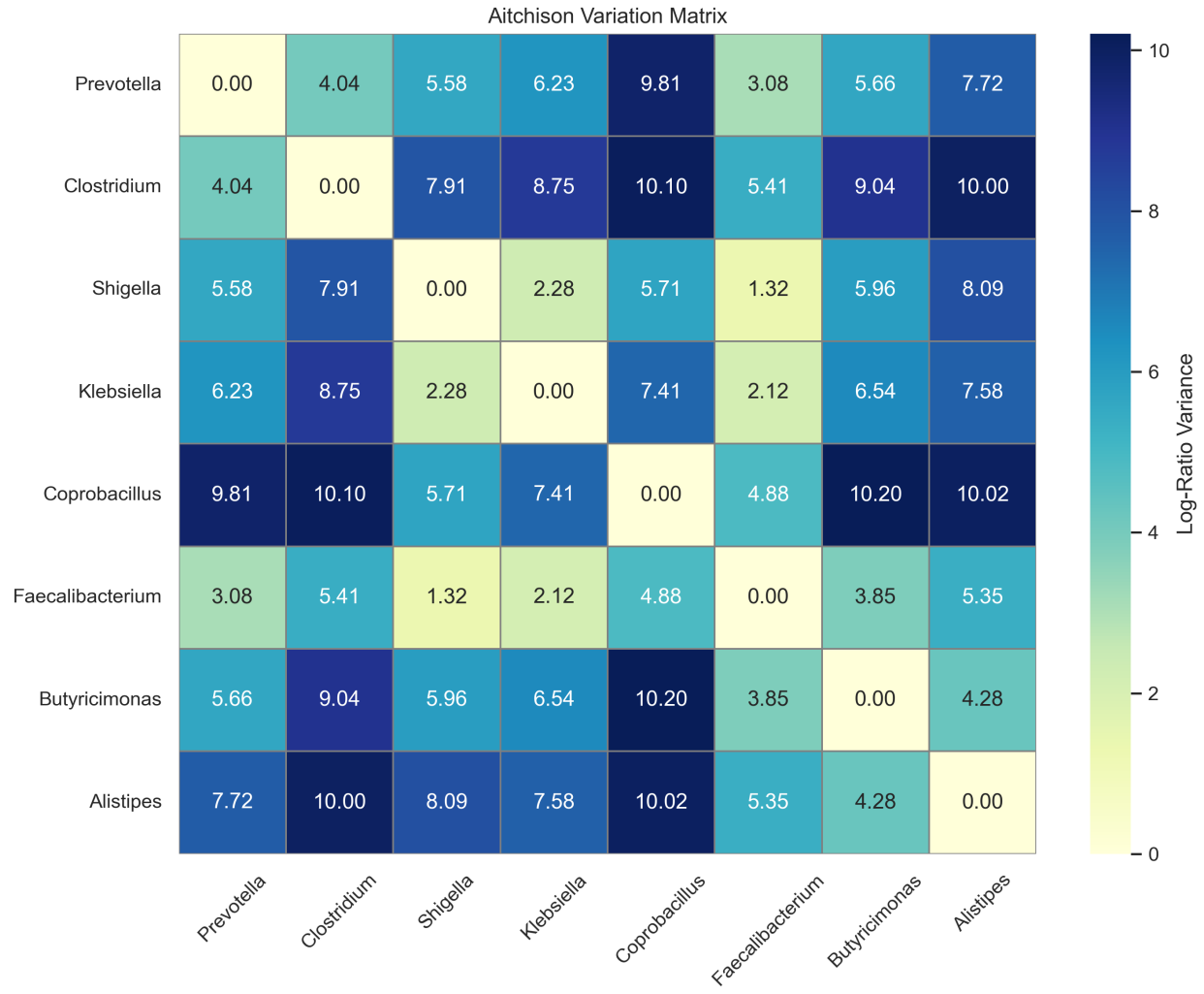

Figure 3: Variation matrix showing variability in log-ratio relationships between genera.

From these analysis we suspect *Prevotella, Clostridium, Coprobacillus* are significant groups showing relative variability between groups. We will later form our basis for ILR transforms using this hypothesis.

4 Principal Component Analysis (PCA)

4.1 Centered Log-Ratio (CLR) Transformation

To apply Euclidean tools, the CLR transform is used:

$$\mathrm{CLR}(x) = \left(\ln\left(\frac{x_1}{g(x)}\right), \ldots, \ln\left(\frac{x_D}{g(x)}\right) \right)$$

Where the $g(x)$ represents the geometric mean of the sample.

4.2 Plotting PCA BiPlot

We now perform PCA on CLR-transformed data to reduce dimensionality and visualize group separation.

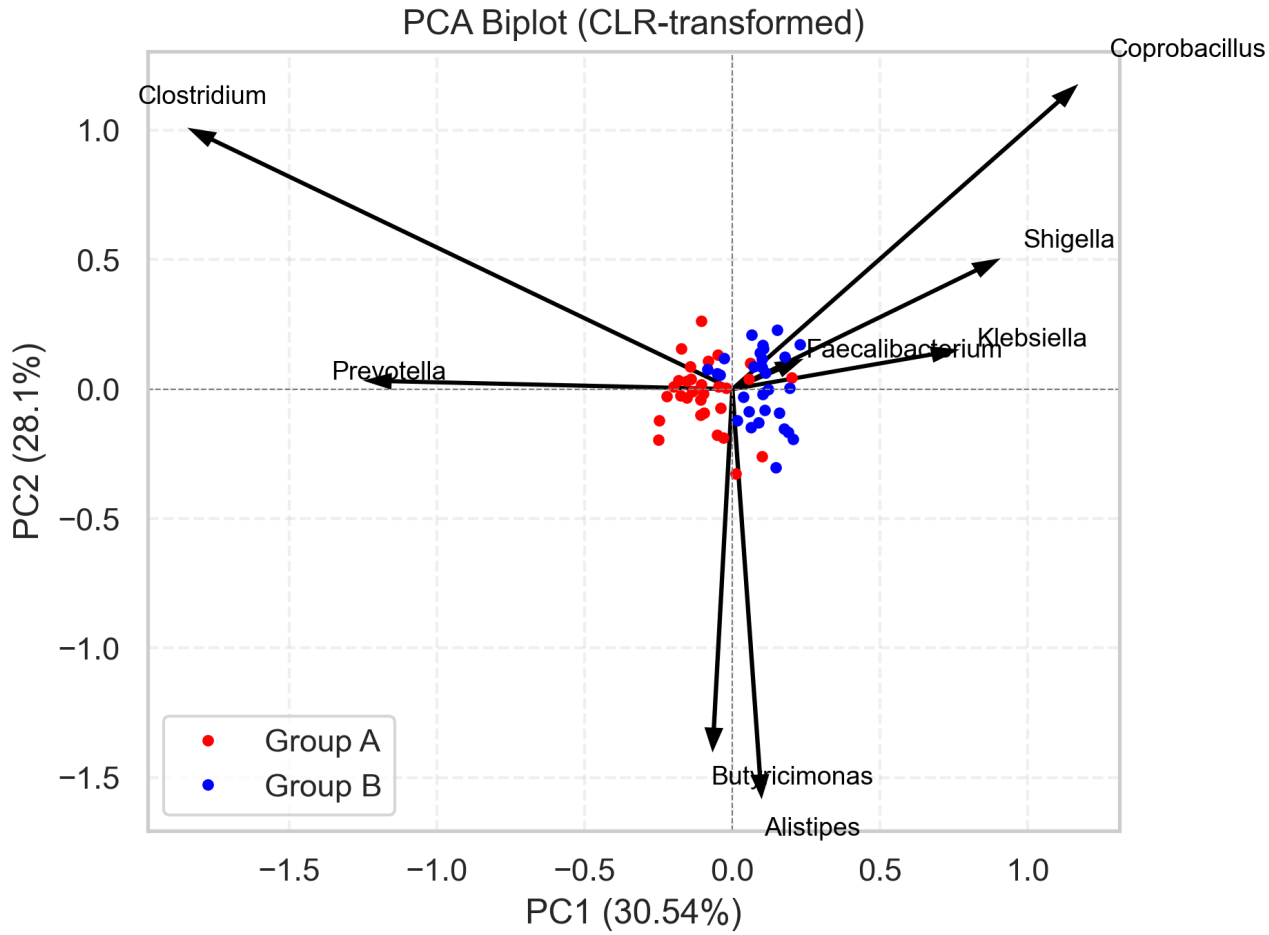

Figure 4: PCA biplot of CLR-transformed data.: Controls. Arrows show taxon loadings. The percentage of variance explained by the PC is given in brackets under the axes

Although only ∼60% of the variance is visible in this graph, we can draw some preliminary observations from this.

5 Verifying Hypothesis

From these we hypothesize that group A seems to cluster more towards *Prevotella, Clostridium*, while group B seems to cluster towards *Coprobacillus, Shigella and Klebsiella*. And *Butyricimonas, Alistipes and Faecalibacterium* seem to have a similar (neutral) effect between the two groups.
To evaluate this, we perform ANOVA on ILR-transformed data derived from selected balances.

5.1 ILR on Selected Balances

We define meaningful balances between specific genera - *Prevotella, Clostridium and Butyricimonas* vs *Coprobacillus, Shigella and Klebsiella* and *Faecalibacterium and Alistipes* as a neutral balance - using ILR coordinates derived from a sequential binary partitions based off our hypothesis.
Given a D-part composition $\mathbf{x} = (x_1, x_2, \ldots, x_D)$ and a contrast matrix $\Psi_{D-1,D}$ based on the basis $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$, ILR coordinates are given by:

$$\text{ILR}(\mathbf{x}) = \text{CLR}(\mathbf{x}) \cdot \Psi^T = \mathbf{x}^* \tag{4}$$

The balances and their interpretations are as follows:

- **Group 1 (G1)**: Prevotella (G1.1), Clostridium (G1.2), Butyricimonas (G1.3)
- **Group 2 (G2)**: Shigella (G2.1), Klebsiella (G2.2), Coprobacillus (G2.3)
- **Group 3 (G3)**: Faecalibacterium (G3.1), Alistipes (G3.2)

Balance	**G1.1**	**G1.2**	**G2.1**	**G2.2**	**G2.3**	**G3.1**	**G1.3**	**G3.2**	**Interpretation**
B1	1	1	-1	-1	-1	0	1	0	G1 vs G2
B2	1	-1	0	0	0	0	0	0	G1.1 vs G1.2
B3	1	1	1	1	1	-1	1	-1	G1, G2 vs G3
B4	0	0	1	-1	0	0	0	0	G2.1 vs G2.2
B5	0	0	1	1	-1	0	0	0	G2.1, G2.2 vs G2.3
B6	1	1	0	0	0	0	-1	0	G1.1, G1.2 vs G1.3
B7	0	0	0	0	0	1	0	-1	G3.1 vs G3.2

Table 1: Positive entries (+1) are numerator taxa, negative entries (–1) are denominator taxa, and 0 means the taxon is not involved in the balance.

5.2 ANOVA on Balances

We apply one-way ANOVA to test for group differences in selected ILR balances. Namely, the F-test and subsequent p-value conversions.

$\bar{Y}_i$ denotes the mean of the i^{th} group, which contains n_i samples. Here, K is the total number of groups, which is 2, and N is the total number of samples.

$$F = \frac{\sum_{i=1}^{2} n_i (\bar{Y}_i - \bar{\bar{Y}})^2 / (2-1)}{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 / (N-2)}$$

Where the numerator represents the between-group variability and the denominator represents the within-group variability and the p-value represents how uncertain (in %age) we are of our hypothesis

This proves our hypothesis with an error probability of the order of 10^{-11}.
We can conclusively say that our initial hypothesis is true.

Balance	F-statistic	p-value
B1	**67.07**	2.95×10^{-11}
B2	0.88	0.351
B3	1.85	0.179
B4	0.001	0.975
B5	1.34	0.251
B6	5.16	0.0268
B7	0.075	0.785

Table 2: ANOVA results on ILR-transformed balances.

5.3 Visualization of Subcomposition the Simplex

To also visualize this we make a ternary scatter plot on the simplex, to see the groups. Each group of the subcomposition was formed in a similar way as the system of balances. To bring the samples closer to the center of the simplex we perturb it with the inverses of the geometric means.

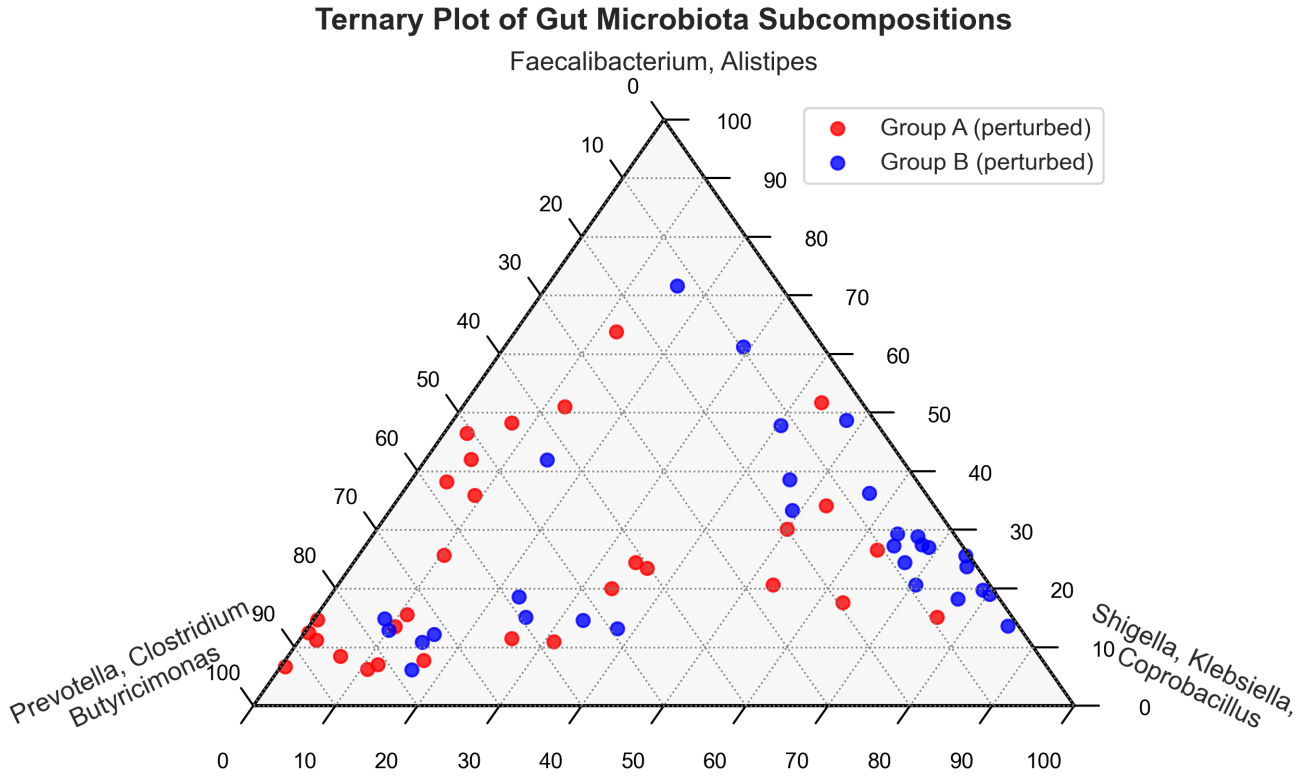

Figure 5: Ternary plot of sample compositions projected on selected 3-part subcomposition.

We can very clearly see two distinct clusters.

Literature Review and Conclusion

As the genera *Prevotella* and *Clostridium* were found to be less abundant in children with ASD, this finding is consistent with the study by Kang et al. (2013), titled Reduced incidence of Prevotella and other fermentors in intestinal microflora of autistic children, which reported significantly reduced levels of fermentative bacteria, particularly *Prevotella*, in the intestinal microbiota of autistic individuals.

Conversely, genera such as *Shigella*, *Klebsiella*, and *Coprobacillus* exhibited relatively higher abundances, potentially compensating for the reduction in fermentative taxa.

Thus, we can conclude Group A as the control group and Group B as the group with ASD.

These observations reinforce the notion that the gut microbiome in children with ASD exhibits distinct compositional differences from neurotypical controls, particularly a depletion in beneficial fermenters like *Prevotella*. Such alterations may have implications for gut-brain axis functioning and overall health. The elevated presence of opportunistic or potentially pathogenic taxa such as *Shigella* and *Klebsiella* may require further investigation.

Appendix

Code Repository: All scripts, data processing steps, and visualizations used in this analysis are available at: `https://github.com/Ruchir-r/CoDa`

Disclaimer: Some portions of the manuscript, including linguistic rephrasing and syntactic refinement, were assisted by OpenAI's ChatGPT. All scientific content, analyses, and interpretations are the sole responsibility of the author.