# Biomarkers of ASD

Lucas Childs, Minu Pabbathi, Nathan Kim, Bahaar Ahuja, Anna Liang

2025-11-06

## Abstract

This report investigates blood biomarkers for autism spectrum disorder (ASD) with a focus on evaluating how methodological choices influence results. Using the dataset from Hewitson et al. (2021), we examine protein distributions and justify the use of log-transformations to stabilize variance and reduce skewness. We explore subject-level outliers and assess whether their frequency differs between ASD and typically developing (TD) groups. We then evaluate how modifications to protein selection procedures, such as varying the number of top predictive proteins, using training/test partitions, and applying fuzzy intersections, affect classification performance. Finally, we identify alternative protein panels that achieve comparable or improved classification accuracy and benchmark these results against the original in-class analysis. This work highlights the sensitivity of biomarker analyses to preprocessing and methodological decisions and provides guidance for more robust ASD protein panel selection.

## Aim

The aim of this report is to explore the sensitivity of ASD biomarker analyses to key methodological choices. Specifically, we examine the effects of log-transforming protein levels, assess the presence of subject-level outliers and their group distribution, and investigate how modifications to protein selection procedures (e.g., training/test splits, number of top proteins, fuzzy intersections) impact classification performance. Finally, we identify alternative protein panels that achieve comparable or improved classification accuracy and benchmark these results against the original in-class analysis.

## Dataset

The data consist of blood serum samples from 76 boys with autism spectrum disorder (ASD) and 78 typically developing (TD) boys, aged 18 months to 8 years. Proteomic analysis was

performed using SomaLogic's SOMAScan™ 1.3K platform, measuring levels of 1,317 proteins. Two additional variables, ADOS (ASD severity) and group (ASD vs. TD), were included.

Preprocessing involved removing missing values, applying a log transformation to stabilize variance and reduce skewness, centering and scaling each protein to standardize distributions, and trimming extreme outliers to prevent disproportionate influence. For exploratory analyses, we also considered an untrimmed dataset to examine subject-level outliers.

## Summary of published analysis

The original researchers used multiple $t$-tests to see whether there were significant differences in protein levels between the ASD group and the neurotypical group. Based on those $t$ values, the top 10 proteins were selected for the prediction model. Each protein was also correlated with ADOS scores and the top 10 highly correlated proteins were selected. The third approach was random forest, which involves using a random forest to predict whether a participant has ASD or is typically developing. By keeping track of which variables were used most to define splits, a variable importance score can be used to determine which predictors were most influential in prediction. Using this method, the top proteins were selected.

After running all three methods and determining the top 10 proteins for each method, 5 proteins that were common to all three methods were selected as the core proteins. Each of the other proteins were added one at a time to see their impact on the AUC, leading to about four additional proteins being classified as optimal proteins. The final 9 proteins were IgD, suPAR, MAPK14, EPHB2, DERM, ROR1, GI24, elF-4H, and ARSB. After all nine were combined, the AUC of the classifier was approximately 0.860.

## Findings

### Impact of preprocessing and outliers

After looking at the raw protein level distributions for a sample of 4 proteins in the dataset, it became clear that the scale of each protein level was very different across proteins. For example, one protein's mean was 17,164 while another's was 458. Furthermore, we found certain proteins had skewed distributions where one protein had a mean roughly 5,000 units larger than its median, indicating a right skew. Looking at each protein's density further revealed the differing scales of each protein level, large outliers, and skewed distributions apparent in 2 out of the 4 randomly selected proteins.

The log transformation of the protein levels helps to compress the scale of protein levels across the wide range of positive values that were encountered. Additionally, the logarithm helps with skewed data, helping to make data more symmetric and closer to a normal distribution (as confirmed with histograms and a QQ Plot).