# 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).

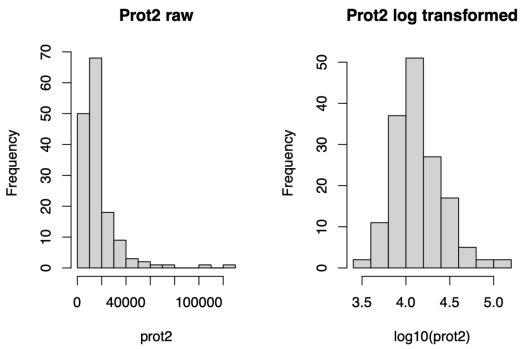


Figure 1: Question 1 comparison

From the density plot, `prot2` looks skewed (strongly right-skewed). It contain large outliers, and the scale of `prot2`'s protein level is very high. But, after log transforming, the right-skewed `prot2` now appears more symmetric and of a much smaller and more readable scale.

**Task 2:** After removing the original trimming, we flagged protein-level outliers at $|z| > 3$, where 99.7% of data is captured under a normal curve. Using the Tukey cutoff ($Q3{+}1.5{\cdot}IQR \approx 29.4$), the 154 test subjects were considered outliers if they had $\geq 30$ outlier proteins. This flagged 13 outlier subjects, distributed as $\frac{4}{76} \approx 5.3\%$ in the ASD group and $\frac{9}{78} \approx 11.5\%$ in the TD group. However, a two-proportion test found no statistically significant difference in this frequency of outlier subjects between ASD and TD ($p = 0.2668$), suggesting that extreme values are concentrated in a small number of individuals rather than representing a systematic group difference. These findings support the validity of the original trimming approach. The extreme values appear to be leverage points or noise rather than having significant impact, and the $\pm 3$ SD trimming in the initial dataset is appropriate. Thus, we can proceed with the original trimmed dataset for our analysis.

### Methodological variations

To evaluate the sensitivity of the analysis to different methodological choices, we first modified the workflow so that all feature selection was carried out exclusively on a training partition, while a separate test set was held out and used only once at the very end for evaluation. This change avoids data leakage and gives a more realistic estimate of model performance.

The original analysis selected 10 proteins. To improve the predictive ability, we extended this to 20 proteins. The top 20 proteins from the *t*-tests and random forest were used to fit a logistic regression model. This improved the sensitivity by 6%, specificity by 13%, accuracy by 9%, and AUC by 4%, likely because adding more predictors will explain more of the variability.

Additionally, instead of using a hard intersection between the top 10 proteins, we tested using a fuzzy intersection, meaning we allowed some overlap between proteins from both analyses. The new analysis decreased the sensitivity by 20%, specificity by 7%, accuracy by 10%, and AUC by 12%. This could be because proteins that were found to be important in both analysis likely had the highest predictive power. When using other proteins that are less explanative, the accuracy decreases significantly.

!Question 3 20 Protein vs 10 Protein Comparison

!Question 3 Fuzzy vs Hard Comparison

### Improved classifier

We explored two distinct alternative panel approaches for Task 4: a Pure LASSO approach and a Guided LASSO approach, benchmarking both against the original In-class Intersection model (a 5-protein panel). Our methods sought either a comparable simpler panel or an alternative panel with improved classification performance.The Pure LASSO approach used LASSO penalized regression for both feature selection and final classification, resulting in a 43-protein panel. Proteins identified through both multiple testing with t-tests and Random Forest importance (`proteins_sstar`) were assigned zero penalty to ensure these were included in the panel without significant shrinkage due to regularization. So, this prior validation was used since there was strong prior evidence that these proteins were relevant. The remaining proteins not selected by the two aforementioned methods received L1 regularization (LASSO regularization), allowing us to discover additional predictive biomarkers while controlling overfitting. We can think of this method from a Bayesian standpoint where the prior data (pre-selected panel of 5 proteins, `proteins_sstar`) informed our belief about protein importance. Overall, our method displayed increased classification accuracy, specifically highlighted by the metric: 'accuracy'. The metrics are depicted below:

| metric | estimator | estimate |
|---|---|---|
| sensitivity | binary | 0.750 |
| specificity | binary | 0.933 |
| accuracy | binary | 0.839 |
| roc auc | binary | 0.858 |

Compared to the benchmark results:

| metric | estimator | estimate |
|---|---|---|
| sensitivity | binary | 0.812 |
| specificity | binary | 0.733 |
| accuracy | binary | 0.774 |
| roc auc | binary | 0.883 |

This panel prioritized reducing false positives and resulted in a notable performance trade-off:

Table 1: Comparative Performance and Insights: Pure LASSO vs. In-class Baseline

| Metric | In-class Intersection | Pure LASSO | Change vs. Baseline | Key Insight |
|---|---|---|---|---|
| Accuracy | 0.774 | 0.767 | $\approx 1\%$ decrease | Comparable overall accuracy. |
| Sensitivity | 0.812 | 0.750 | 7.6% decrease | Reduced ability to correctly identify ASD cases. |
| Specificity | 0.733 | 0.933 | 27.3% increase | Vastly improved performance in ruling out controls. |
| AUROC | 0.883 | 0.858 | $\approx 2.8\%$ decrease | Lower overall discrimination ability. |

The most significant finding from the Pure LASSO model is the dramatic increase in Specificity (from 0.733 to 0.933). This trade-off is valuable, as it greatly reduces the rate of false positives, making the panel highly reliable for control classification, a desirable characteristic for preliminary diagnostic screening.

In addition to the Pure LASSO model, the Guided LASSO approach (using prior knowledge combined with LASSO, as detailed in the project narrative) resulted in a 35-protein panel. The proteins selected for the classifier are the following:

- CD59
- 4-1BB
- Dtk
- Cadherin-5
- HAI-1
- Kallikrein 11
- PAI-1
- Growth hormone receptor
- IGFBP-4
- MRC2
- CRDL1
- IL-17 RD
- TPSG1
- MP2K2
- ENPP7
- MFGM
- PCSK7
- ITI heavy chain H4
- IgD

- DBNL
- DERM
- Elafin
- RELT
- PPID
- Semaphorin 3E
- CD27
- CNDP1
- IL-17 RC
- SRCN1
- Epo
- GDNF
- 14-3-3 protein zeta/delta
- a-Synuclein
- CSRP3
- MIG

While the performance metrics for this model were similar to the baseline across all metrics except Specificity (Accuracy 0.770, Sensitivity 0.745, Specificity 0.733, AUROC 0.857), the overall strategy of combining multiple feature selection methods (t-tests, Random Forest, and Guided LASSO) provided a robust and statistically validated pathway to stable panel selection.