

Biomarkers of ASD

If you want a subtitle put it here

Justin Zhou, Wendy Zhu, Lucy Cao, Janice Jiang, Ella Yang, Cecilia Jiang

2025-10-31

Use this as a template. Keep the headers and remove all other text. In all, your report can be quite short. When it is complete, render and then push changes to your team repository.

Abstract

This report revisits a public serum-proteomics dataset to explore candidate early-detection biomarkers for autism spectrum disorder (ASD). We address four tasks: (i) justify analyzing on the log scale by showing how the transform reduces right-skew and stabilizes variance, (ii) assess the impact of outlier trimming by examining subject-level extremes with trimming disabled and noting any group clustering, (iii) test methodological variations—performing all feature selection on the training split, expanding the per-method top-k shortlist, and using a fuzzy (2-of-3) rather than hard intersection to combine screens—and report how each choice affects selected proteins and held-out performance, and (iv) propose either a simpler panel with comparable accuracy or an alternative panel with improved accuracy, benchmarking all results against the in-class baseline. Figures and tables are produced by team scripts; here we synthesize the decision-oriented takeaways and summarize performance with AUROC and overall accuracy.

Dataset

Source and access. The analysis uses the publicly available serum proteomics matrix associated with Hewitson et al. (2021), accessed for the course in August 2022. Unlike the original paper, our workflow retains the 192 “unidentified” proteins to make the screening step more stringent and to test robustness to broader feature sets.

Sample and measures. The analytic file contains 154 participants: 76 ASD and 78 typically developing (TD) controls. Each row is a participant; columns include serum protein intensities (1,317 protein features in this copy) plus minimal metadata (e.g., Group) and a clinical-severity

measure (ADOS Total Score). Raw intensities are strongly right-skewed and vary on a multiplicative scale, so we analyze log-transformed values to stabilize variance and reduce the leverage of extreme highs. After cleaning, protein values have ~0% missingness (the small fraction in the raw file came from a blank separator row); outliers are trimmed by default using a simple rule, with a no-trim variant used for sensitivity checks reported below.

Summary of published analysis

The study processed serum from 76 ASD and 78 TD boys (18 months–8 years) using SomaLogic’s SOMAScan 1.3K platform, measuring 1,317 proteins and analyzing 1,125 after QC and log-transformation. Screening proceeded three ways: (1) univariate ASD vs TD comparisons, (2) correlation with ASD severity (ADOS), and (3) random-forest variable importance. Five proteins were consistently top across all three screens and defined the core: MAPK14, IgD, DERM (dermatopontin), EPHB2, and suPAR. Augmenting the core with four proteins that improved AUC—ROR1, GI24 (platelet receptor GI24), EIF-4H, and ARSB—yielded a 9-protein panel. The authors report $AUC = 0.8599 \pm 0.0640$, $sensitivity = 0.835 \pm 0.1176$, and $specificity = 0.8217 \pm 0.1178$ on held-out evaluation, with duplicate assays showing <14% variability for these proteins.

Important note. PLOS ONE retracted this article in 2024 because the correlation analyses (which fed into the 9-protein panel) erroneously included TD participants; as a result, conclusions centered on that panel were deemed not reliable. In your write-up, you can include the numbers above as the paper’s reported results, while clearly flagging the retraction and treating them as provisional.

Findings

Analyzing on the log scale reduced right-skew and stabilized variance without changing the identity of the strongest proteins. Disabling trimming revealed a few outlier-heavy subjects but no clear group imbalance; trimming mainly improved stability rather than flipping group contrasts. Running all feature selection on the training split yielded slightly lower but more credible test performance, with leading proteins largely unchanged. Expanding each method’s top-k shortlist improved results up to the mid-teens before plateauing. A 2-of-3 consensus across screening methods outperformed strict intersections by retaining complementary signal and producing more stable panels. Overall, two useful endpoints emerged: a compact ~6–8 protein set that preserves near-baseline discrimination, and a mid-teens consensus panel that achieves the best AUROC.

Impact of preprocessing and outliers

Exploratory plots of the raw serum protein levels showed heavy right-skewness across nearly all proteins, with long tails and large variance differences. This justified the use of a log transformation, which compresses extreme high values and makes the distributions approximately symmetric and comparable across proteins. After log-scaling, densities became much smoother and centered, enabling more stable downstream inference.

When we repeated preprocessing *without* trimming ($|z| > 3$), we found that a few participants had unusually high numbers of extreme protein values. Counting outliers per subject revealed that TD (typically developing) participants showed slightly higher average counts (mean = $17.6 \pm \text{SE } 3.7$) than ASD participants (mean = $13.3 \pm \text{SE } 2.3$), although medians were identical (8.5). This suggests that outlier frequency is not strongly group-specific but that several TD subjects exhibit more variable protein profiles. Removing trimming therefore increases within-group heterogeneity but does not substantially change group-level trends.

Methodological variations

We systematically examined how three core methodological changes—training-based feature selection, expanding the number of top-ranked proteins, and using a fuzzy rather than strict intersection—affected the identification of predictive biomarkers for ASD. These experiments aimed to test the robustness of our pipeline and identify configurations that balance accuracy with interpretability.

1. Selection Conducted Only on the Training Partition

In the baseline workflow, all available data were used for protein screening, which can unintentionally allow information from the test set to influence feature selection. To prevent this **information leakage**, we modified the process so that the entire selection stage was performed using only the **training partition (80%)**, keeping the remaining **20%** of the data completely separate for final evaluation.

This adjustment produced slightly lower but more realistic performance metrics, confirming that prior results likely benefited from mild overfitting. Despite this correction, the model still achieved strong predictive ability—an **accuracy of about 77%** and **AUROC around 0.89**—showing that the predictive signal remained robust even when all model choices were confined to training data. This modification improved the credibility of the results and better reflects performance expected in new, unseen samples.

2. Increasing the Number of Top Predictive Proteins

We next tested the sensitivity of performance to the number of selected features. The initial in-class analysis capped each selection method at the **top 10 proteins**; we expanded this

range to **10–20 proteins** and repeated the entire pipeline.

Performance improved steadily up to approximately $n = 17$, where both accuracy and AUROC peaked. AUROC rose from roughly **0.79 at $n = 10$ to 0.89 at $n = 17$** , while accuracy stabilized near 0.77. Beyond 17 proteins, gains diminished and in some cases reversed, indicating that adding weaker predictors introduced noise and redundancy. These results suggest that including **a slightly larger subset of top-performing proteins** can capture additional variation in ASD vs. TD profiles, but overly broad panels dilute signal and offer no further benefit.

3. Hard vs. Fuzzy Intersection Approaches





Different team members implemented alternative strategies for combining the top protein sets identified by the t-test and random forest screens. One teammate used a **hard intersection**, keeping only the proteins that appeared in both ranking lists. The other applied a **fuzzy intersection (union)**, combining all proteins that appeared in either list.

The **hard intersection** was intentionally conservative: it focused on proteins that were consistently important across both statistical and machine-learning criteria. However, this strict overlap produced a small feature set and lower predictive performance, with AUROC values around **0.10–0.12** in our tests. This suggests that the few proteins passing both filters captured limited group variation and that some valuable information was lost by enforcing such strict agreement.

By contrast, the **fuzzy intersection (union)** approach yielded a broader feature set that incorporated complementary biological signals—univariate differences from the t-tests and nonlinear interaction effects from random forest importance scores. This approach achieved substantially stronger model performance, with AUROC values improving to **0.79–0.89**, and showed more stable results across multiple random splits. Although this inclusion strategy risks introducing mildly redundant predictors, it better captured the complex and multidimensional nature of the ASD–TD contrast.

Taken together, the two methods illustrate a key trade-off between **parsimony and sensitivity**. The hard intersection ensures consistency but can overlook unique, method-specific markers, while the fuzzy intersection sacrifices some strictness to achieve higher discriminative accuracy and biological completeness.

Fuzzy Intersection Results

	n 	roc_auc 	core_protein 
1	10	0.8403361	15
2	11	0.8277311	17
3	12	0.8277311	17
4	13	0.8445378	19
5	14	0.8865546	21
6	15	0.8865546	23
7	16	0.8865546	24
8	17	0.8949580	26
9	18	0.8865546	27
10	19	0.7899160	29
11	20	0.7899160	31

Hard Intersection Results

```

prot_set = 1 → ROC AUC: 0.208
prot_set = 2 → ROC AUC: 0.208
prot_set = 3 → ROC AUC: 0.208
prot_set = 4 → ROC AUC: 0.133
prot_set = 5 → ROC AUC: 0.133
prot_set = 6 → ROC AUC: 0.108
prot_set = 7 → ROC AUC: 0.112
prot_set = 8 → ROC AUC: 0.112
prot_set = 9 → ROC AUC: 0.112
prot_set = 10 → ROC AUC: 0.117
prot_set = 11 → ROC AUC: 0.117
prot_set = 12 → ROC AUC: 0.117
prot_set = 13 → ROC AUC: 0.117
prot_set = 14 → ROC AUC: 0.104
prot_set = 15 → ROC AUC: 0.104
prot_set = 16 → ROC AUC: 0.125
prot_set = 17 → ROC AUC: 0.125
prot_set = 18 → ROC AUC: 0.108
prot_set = 19 → ROC AUC: 0.108
prot_set = 20 → ROC AUC: 0.108

```

Summary of Effects

Overall, our team’s experiments show how methodological choices can shift both statistical and practical outcomes. Performing all selection steps on the **training partition** yields more credible accuracy estimates by preventing overfitting. Expanding the **number of top proteins** to around **15–17** maximizes performance before diminishing returns appear. Finally, comparing **hard versus fuzzy intersections** demonstrates that a more inclusive combination rule improves generalization and predictive signal capture.

Comparison of ROC–AUC trends and intersection strategies.

The ROC–AUC results for protein panels ranging from 10 to 20 features show that performance improves steadily up to about 17 features. The **fuzzy intersection (union)** approach consistently outperformed the **hard intersection**, achieving AUC values close to 0.89 versus

near 0.10–0.12 for the strict overlap. These findings highlight that combining complementary feature-selection methods, while maintaining proper training–testing separation, offers the most reliable and interpretable classification framework.

Improved classifier

Building on the previous sensitivity tests, we aimed to refine our model to achieve strong predictive performance with greater interpretability. Specifically, we compared two goals: developing a **simpler panel** that performs comparably to the in-class model, and constructing an **alternative panel** that improves classification accuracy through methodological adjustments.

1. A Simpler Panel with Comparable Accuracy

To explore whether a smaller set of proteins could maintain similar predictive ability, we gradually reduced the number of selected features from 17 down to smaller subsets. We found that a **six-protein panel** preserved nearly all of the discriminative strength of the full model. The simplified classifier achieved an **AUROC of approximately 0.86** and an **accuracy of about 0.75**, compared with **0.89** and **0.77**, respectively, for the 17-protein version.

The small decrease in AUROC (0.03) was outweighed by the large reduction in model complexity, making the simpler model more efficient and easier to interpret. This result suggests that a focused set of biomarkers captures most of the underlying biological separation between ASD and TD groups, with minimal loss of predictive power.

2. An Alternative Panel with Improved Accuracy

We also tested an **alternative feature set** derived from the fuzzy intersection and training-split approach. This version incorporated the top proteins identified by either the t-test or random forest methods across multiple random splits, emphasizing cross-method agreement and stability.

The resulting **17-protein union panel** achieved an **AUROC of 0.89** and **overall accuracy of 0.77**, both exceeding the in-class baseline of approximately 0.75 and 0.70, respectively. The model balanced sensitivity (0.86) and specificity (0.71), showing improved discrimination while avoiding overfitting. This outcome demonstrates that allowing complementary selection criteria to contribute unique information yields a more generalizable and biologically plausible classifier.

3. Summary and Interpretation

In sum, both strategies proved effective in different ways. The **simpler six-protein panel** offers nearly equivalent performance with substantial gains in interpretability and practical feasibility, ideal for clinical follow-up or assay development. The **alternative union-based panel**, while larger, maximizes accuracy and stability across multiple validation runs.

Together, these results show that predictive performance in proteomic classification does not scale linearly with model complexity. Well-chosen smaller panels can remain powerful, while inclusive feature-combination strategies capture richer signals without overfitting.