UNIVERSITY OF CALIFORNIA SAN DIEGO

# Using emerging data analysis technics to improve pediatric disease diagnosis

A dissertation submitted in partial satisfaction
of the requirements for the Doctor of Philosophy

in

Bioinformatics and System Biology

by

Bokan Bao

Committee in charge:

Professor Nathan E. Lewis, Chair
Professor Eric Courchesne, Co-Chair
Professor Shamim Nemati
Professor Karen Pierce
Professor Debashis Sahoo

2022

The Dissertation of Bokan Bao is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

# DEDICATION

I am lucky to spend meaningful and lighthearted five years in my advisor Dr Nathan E. Lewis's lab to exploring not only my scientific but also lifetime interests.

# TABLE OF CONTENTS

# LIST OF FIGURES

Appendix Figures

## LIST OF TABLES

ACKNOWLEDGMENTS

Chapter 1, in full, is reprint of the material as it appears in Nat Communication. 2021 Bao, Bokan, Benjamin P Kellman, Austin WT Chiang, Yujie Zhang, James T Sorrentino, Austin K York, Mahmoud A Mohammad, Morey W Haymond, Lars Bode, Nathan E Lewis. Correcting for sparsity and interdependence in glycomics by accounting for glycan biosynthesis. Nat Commun. 2021;12(1):4988. The dissertation/thesis author was the primary investigator and author of the paper.

Chapter 2, in full, is reprint of the material as it appears in Molecular Psychatric. 2022. Bao, Bokan, Javad Zahiri, Vahid H Gazestani, Linda Lopez, Yaqiong Xiao, Raphael Kim, Teresa H Wen, Austin WT Chiang, Srinivasa Nalabolu, Karen Pierce, Kimberly Robasky, Tianyun Wang, Kendra Hoekzema, Evan E Eichler, Nathan E Lewis, Eric Courchesne. A predictive ensemble classifier for the gene expression diagnosis of ASD at ages 1 to 4 years. Molecular Psychiatry. 2022. The dissertation/thesis author was the primary investigator and author of the paper.

Chapter 3, in full, is reprint of the paper in preparation. Bao, Bokan, Yaqiong Xiao, Javad Zahiri, Charlene Andreason, Yakta Syed, Summer Zhu, Teresa H. Wen, Eric Courchesne, Nathan E. Lewis, Karen Pierce. Examination of automatic facial action unit measurement as a mechanism to differentiate ASD vs non-ASD toddlers. The dissertation/thesis author was the primary investigator and author of the paper.

# VITA

| | |
|---|---|
| 2015 | Bachelor of Science in Biological Science, and Biometry & Statistics, Cornell University |
| 2017-2018 | Teaching Assistant, University of California San Diego |
| 2017-2022 | Research Assistant, University of California San Diego |
| 2022 | Doctor of Philosophy in Bioinformatics, University of California San Diego |

# PUBLICATIONS

Bao, Bokan, Yaqiong Xiao, Javad Zahiri, Charlene Andreason, Yakta Syed, Summer Zhu, Teresa H. Wen, Eric Courchesne, Nathan E. Lewis, Karen Pierce. Examination of automatic facial action unit measurement as a mechanism to differentiate ASD vs non-ASD toddlers. In preparation.

Bao, Bokan, Javad Zahiri, Vahid H Gazestani, Linda Lopez, Yaqiong Xiao, Raphael Kim, Teresa H Wen, Austin WT Chiang, Srinivasa Nalabolu, Karen Pierce, Kimberly Robasky, Tianyun Wang, Kendra Hoekzema, Evan E Eichler, Nathan E Lewis, Eric Courchesne. A predictive ensemble classifier for the gene expression diagnosis of ASD at ages 1 to 4 years. Molecular Psychiatry. 2022.

Bao, Bokan, Benjamin P Kellman, Austin WT Chiang, Yujie Zhang, James T Sorrentino, Austin K York, Mahmoud A Mohammad, Morey W Haymond, Lars Bode, Nathan E Lewis. Correcting for sparsity and interdependence in glycomics by accounting for glycan biosynthesis. Nat Commun. 2021;12(1):4988.

Wen, Teresa H., Amanda Cheng, Charlene Andreason, Javad Zahiri, Yaqiong Xiao, Ronghui Xu, Bokan Bao et al. "Large scale validation of an early-age eye-tracking biomarker of an autism spectrum disorder subtype." *Scientific reports* 12, no. 1 (2022): 1-13.

Zhou, Jingtian, Jianzhu Ma, Yusi Chen, Chuankai Cheng, Bokan Bao, Jian Peng, Terrence J. Sejnowski, Jesse R. Dixon, and Joseph R. Ecker. "Robust single-cell Hi-C clustering by convolution-and random-walk–based imputation." *Proceedings of the National Academy of Sciences* 116, no. 28 (2019): 14011-14018.

Carlin, Daniel E., Samson H. Fong, Yue Qin, Tongqiu Jia, Justin K. Huang, Bokan Bao, Chao Zhang, and Trey Ideker. "A fast and flexible framework for network-assisted genomic association." *Iscience* 16 (2019): 155-161.

Kellman BP, Yujie Zhang, Emma Logomasini, Eric Meinhardt, Karla P Godinez-Macias, Austin WT Chiang, James T Sorrentino, Chenguang Liang, Bokan Bao, Yusen Zhou, Sachiko Akase, Isami Sogabe, Thukaa Kouka, Elizabeth A Winzeler, Iain BH Wilson, Matthew P Campbell, Sriram Neelamegham, Frederick J Krambeck, Kiyoko F Aoki-Kinoshita, Nathan E Lewis. A consensus-based and readable extension of Linear Code for Reaction Rules (LiCoRR). *Beilstein J Org Chem*. 2020;16:2645-2662.

Liang, Chenguang, Austin WT Chiang, Anders H. Hansen, Johnny Arnsdorf, Sanne Schoffelen, James T. Sorrentino, Benjamin P. Kellman, Bokan Bao, Bjørn G. Voldborg, and Nathan E. Lewis. "A Markov model of glycosylation elucidates isozyme specificity and glycosyltransferase interactions for glycoengineering." *Current research in biotechnology* 2 (2020): 22-36.

## FIELDS OF STUDY

Major Field:

Glycosylation,

Autism Genetics,

Applied Machine Learning

ABSTRACT OF THE DISSERTATION

Using emerging data analysis technics to improve pediatric disease diagnosis

by

Bokan Bao

Doctor of Philosophy in Bioinformatics

University of California San Diego, 2022

Professor Nathan E. Lewis, Chair

Professor Eric Courchesne, Co-Chair

Many data types are used in bioinformatics research, including genomics, transcriptomics, proteomics, pathway data, disease network, and gene ontology (GO) data, which are heavily studied in disease diagnosis or biomarker detection. The use of newer data types, such as glycomics, fMRI, and facial behavior data, is also growing and can provide unique perspectives for disease cell biology. These new data types have unique properties that require newly adapted algorithms for precise and granular characterization, which is essential before machine learning or statistical models can be confidently used to study disease mechanisms or identify biomarkers from large-scale datasets. The newly developed tools can then allow sophisticated evaluations and yield high-quality results. The first part of my thesis introduced GlyCompare, a powerful glycomics analysis pipeline. The pipeline corrects for the sparsity and non-independence in glycomics data by accounting for the shared biosynthetic network in the data. This new approach makes the downstream analyses more interpretable and better powered.

Then in the second part, a generalizable machine learning platform was developed with 42,840 models composed of 3570 gene expression feature sets and 12 classification methods. A gene expression ASD diagnostic classifier built with this platform had AUC-ROC ≥ 0.8 on both Training and Test sets. Our classifier is diagnostically predictive and replicable across different toddler ages, races, and ethnicities; outperforms the risk gene mutation classifier; and has potential for clinical translation.

In the last section, I developed a pipeline to evaluate facial behavior data from toddlers using state-of-the-art expression analysis software. In certain situations, emotional response is overly intense in ASD compared to other toddlers. Our action unit classifier had a sensitivity of 83.3% and a specificity of 67.5% in the test dataset (90.1% and 75% in the training dataset). We verified that our classifier was unbiased against common confounding factors (age, race, and ethnicity). By combining the action unit classifier and Geo-Pref non-social score, we achieved a specificity of 100% and sensitivity of 50% on the training and test datasets. The ensemble classifier maintained the high specificity while considerably increasing the sensitivity, which provides the potential for screening applications.

# INTRODUCTION

In the first two years, I mainly worked in the glyco-bioengineering field. Glycosylation is a complex post-translational modification and it decorates one-fifth to one-half of eukaryotic proteins(Khoury, Baliban, and Floudas 2011; Apweiler, Hermjakob, and Sharon 1999). The diversified glycans account for 12-25% of dry cell mass and have essential functional and pathological roles(RodrÍguez, Schetters, and van Kooyk 2018; Gutierrez et al. 2018). Despite their importance, glycans have complex structures that are difficult to study. The complex structures of glycans arise from a non-template-driven synthesis through a biosynthetic network involving dozens of enzymes. A simple change of a single intermediate glycan or glycosyltransferase will have cascading impacts on the final glycans obtained(Gabius et al. 2002; Spahn and Lewis 2014). Unfortunately, current data analysis approaches for glycoprofiling and glycomic data lack the critical systems perspective to decode the interdependence of glycans easily(Reiding et al. 2014, 2019; Doherty et al. 2018; Wohlschlager et al. 2018; Black et al. 2019; Ashwood et al. 2019). It is important to understand the network behind the glycoprofiles to understand the behavior of the process better.

New tools aiding in the acquisition and aggregation of glycoprofiles are emerging, making large-scale comparisons of glycoprofiles possible. Advances in mass spectrometry now enable the rapid generation of many glycoprofiles with detailed glycan composition and structure predictions(Reiding et al. 2014, 2019; Wohlschlager et al. 2018; Black et al. 2019; Ashwood et al. 2019; Maxwell et al. 2012; Hou et al. 2016; Kremkow and Lee 2018; Krambeck et al. 2017; Holst et al. 2017; Angel et al. 2017), exposing the complex and heterogeneous glycosylation patterns on lipids and proteins(Reiding et al. 2019; Doherty et al. 2018; Black et al. 2019; Cummings 2009; Holst et al. 2016; Čaval et al. 2018; Riley et al. 2019). Large glycoprofile datasets and supporting databases are also emerging, including GlyTouCan(Aoki-Kinoshita et al. 2015), UniCarb-DB(Campbell, Nguyen-Khuong, et al. 2014), GlyGen(York et al. 2019), and UniCarbKB(Campbell, Peterson, et al. 2014).

These technologies and databases facilitate efforts to associate glycans with disease and other phenotypes. However, the rapid and accurate comparison of glycoprofiles can be challenging with the size, sparsity and heterogeneity of such datasets(Reiding et al. 2019; Doherty et al. 2018; Black et al. 2019; Holst et al. 2016; Čaval et al. 2018; Riley et al. 2019; Yang et al. 2015). A glycoprofile provides glycan structure and abundance information, and each glycan is usually treated as an independent entity. Furthermore, in any one glycoprofile, only a tiny percentage of all possible glycans may be detected(Reiding et al. 2019; Doherty et al. 2018; Black et al. 2019; Holst et al. 2016; Yang et al. 2015). Thus, if there is a significant perturbation to glycosylation in a dataset, only a few glycans, if any, may overlap between samples. However, these non-overlapping glycans may only differ in their synthesis by as few as one enzymatic step. Thus, it requires deliberate manual coding to make them comparable(Reiding et al. 2014, 2019; Doherty et al. 2018; Wohlschlager et al. 2018; Black et al. 2019; Ashwood et al. 2019; Holst et al. 2016; Yang et al. 2015). These properties of glycomics data may not be problematic in the studies of individual glycans and their downstream effects on other biological processes. However, this may be a problem in determining the sources of changes in glycan abundance by using large amounts of data(Reiding et al. 2019; Doherty et al. 2018; Yang et al. 2015; Benedetti et al. 2017). Since many methods assume data independence (e.g., t-tests, ANOVA, etc.), their application to glycomics can lead to decreased statistical power or erroneous results.

Previous studies have investigated the similarities across glycans by using glycan motifs. Scientists are using glycan fingerprinting to describe glycan diversity in databases(Rademacher and Paulson 2012; Bojar et al. 2021), align glycan structures(Hosoda et al. 2018), identify glycan epitopes in glycoprofiles(Alocci et al. 2018) and lectin profiles(Khoury, Baliban, and Floudas 2011), deconstruct LC-MS data to quantify glycan abundance(Klein, Carvalho, and Zaia 2018), or compare glycans in glycoprofiles(Sharapov et al. 2018). These tools use information on glycan composition or epitopes. However, the accounting of shared biosynthetic steps could provide complete context to all glycan epitopes. That context includes connecting all glycans to the enzymes involved in their synthesis, the order of the enzyme reactions, and information on competition for glycan substrates. Thus, a generalized substructure approach could facilitate the study of large numbers of glycoprofiles by connecting them to the shared mechanisms involved in making each glycan.

In the first chapter, I presented the GlyCompare, a method enabling the rapid and scalable analysis and comparison of multiple glycoprofiles, while accounting for the biosynthetic similarities of each glycan. My colleague and I proposed glycan substructures, or intermediates, as the appropriate functional units for meaningful glycoprofile comparison since each substructure can capture one step in the complex process of glycan synthesis, which accounts for the shared dependencies across glycans. This approach addressed current challenges in sparsity and hidden interdependence across glycomic samples and will facilitate discovering mechanisms underlying the changes among glycoprofiles. My colleague and I demonstrated this approach's functionality and performance with various glycomic analyses, including recombinant erythropoietin (EPO) N-glycosylation, human milk oligosaccharides (HMOs), mucin-type O-glycans, gangliosides, and site-specific compositional data. Specifically, I analyzed sixteen MALDI-TOF glycoprofiles of EPO, where each EPO glycoprofile was produced in a different glycoengineered CHO cell line(Čaval et al. 2018; Yang et al. 2015). I also analyzed forty-eight HPLC glycoprofiles of HMO from six mothers(Mohammad, Hadsell, and Haymond 2012). By analyzing these glycoprofiles with GlyCompare, I quantified the abundance of important substructures, clustered the glycoprofiles of mutant cell lines, connected genotypes to unexpected changes in glycoprofiles, and associated a phenotype of interest with substructure abundance and flux. My colleagues and I further demonstrated that such analyses gain statistical power. Finally, we expanded our studies to include a tumor-normal comparison of mucin-type O-glycans, human retinal glycolipids, and site-specific N-glycan compositional data from the mouse brain. The analyses of the various N-type and O-type glycan datasets demonstrate that our framework presents a convenient and automated approach to elucidate insights into complex patterns in glycobiology.

After finishing the Glycompare, I started exploring how to apply machine learning techniques in the ASD diagnosis study and started closely working with Dr. Courchesne and Dr. Pierce.

ASD is a prenatal(Courchesne, Gazestani, and Lewis 2020; Courchesne et al. 2019; Gazestani et al. 2020; Courchesne et al. 2011; Marchetto et al. 2017; Courchesne and Pierce 2005; Willsey et al. 2013; Courchesne et al. 2007; Stoner et al. 2014; Parikshak et al. 2013; Packer 2016; Kaushik and Zarbalis 2016; Krishnan et al. 2016; Donovan and Basson 2017; Grove et al. 2019; Satterstrom et al. 2020), highly heritable disorder(Bai et al. 2019) that considerably impacts a child's ability to perceive and react to social information(Bal et al. 2019; Bacon et al. 2018, 2019). Despite this prenatal and strongly genetic beginning, robust and replicable early-age biological ASD diagnostic markers useful at the individual level have not been found. Indeed, ASD diagnosis remains behavior-based and the median age of the first diagnosis remains at ~52 months(Autism and Developmental Disabilities Monitoring Network Surveillance Year 2006 Principal Investigators and Centers for Disease Control and Prevention [CDC] 2009; Baio et al. 2018; Christensen et al. 2018; Maenner et al. 2020), which is nearly 5 years after its first trimester origin. The long delay between ASD's prenatal onset and eventual diagnosis is a missed opportunity for treatment. Moreover, the heterogeneity of ASD genetics and clinical characteristics impose barriers to identifying early-age molecular diagnostics that accurately diagnose the majority of those with this heterogeneous disorder(Lombardo, Lai, and Baron-Cohen 2019). Thus, there is a need for early-age molecular diagnostics of ASD that robustly surmounts this heterogeneity obstacle.

Since ASD's heritability is 81%(Bai et al. 2019), initial attempts have focused on genetics to develop clinically useful biomarkers for precision medicine and causal explanations for ASD pathogenesis. While syndromic risk mutations have been described for >200 genes in ASD(Satterstrom et al. 2020; Feliciano et al. 2019; "Human Gene Module" n.d.), each occurs only rarely in ASD. For 80-90% of patients, such mutations are not found. Thus, an estimated 80% or more of ASD individuals are considered 'idiopathic', wherein little is known about the genes and/or environmental factors causing their disorder. In this idiopathic majority of ASD, the risk is likely associated with many inherited common and rare risk variants in each individual child. Studies of polygenic ASD risk found that the combined effect of genetic risk variants in case-control studies accounts for less than 7.5% of the risk variance(Antaki et al. 2022); genetic ASD risk scores substantially overlap with controls(Robinson et al. 2016; Clarke et al. 2016; Klei et al. 2021; Aguilar-Lacasaña et al. 2022); and, because of this substantial overlap, polygenic risk scores are not clinically diagnostic or prognostic for individuals, nor are they explanatory for the majority of ASD. Thus, DNA-based mutations or polygenic risk scores may not yet be useful for the many idiopathic ASD subjects at the clinical diagnostic level.

RNA biomarkers have been sought using blood gene expression in >35 ASD studies since

2006(Pramparo, Lombardo, et al. 2015; Pramparo, Pierce, et al. 2015; Ch'ng et al. 2015; Diaz-Beltran, Esteban, and Wall 2016; Tylee et al. 2017; He et al. 2019; Lee et al. 2019; Kong et al. 2012; Gregg et al. 2008; Enstrom et al. 2009; Ansel et al. 2016), but many studies have been underpowered, older-aged, clinically heterogeneous, and/or lacking validation test datasets. Some early genetics researchers rejected blood-based biomarkers believing that ASD-relevant dysregulated gene expression must be restricted to the brain. Recent ASD genetics have reversed this view: The earliest prenatal drivers of deviant ASD development are, in fact, broadly expressed regulatory genes, a large percentage of which are active in non-brain organs and tissues such as blood leukocytes as well as in the prenatal brain(Courchesne, Gazestani, and Lewis 2020; Courchesne et al. 2019; Gazestani et al. 2020; Pramparo, Pierce, et al. 2015; Pramparo, Lombardo, et al. 2015; Tylee et al. 2017; Ansel et al. 2016; Hewitson et al. 2021; He et al. 2019). Broadly expressed genes that constitute most ASD risk genes are upregulated in early prenatal life and impact multiple stages of prenatal brain development from 1st and 2nd trimester proliferation and neurogenesis to neurite outgrowth and synaptogenesis in the 3rd trimester. These genes disrupt gene expression in signaling pathways such as PI3K-AKT, RAS-ERK, Wnt and insulin receptor pathways, which further disrupt prenatal functions(Courchesne, Gazestani, and Lewis 2020; Courchesne et al. 2019; Gazestani et al. 2020; Pramparo, Pierce, et al. 2015; Pramparo, Lombardo, et al. 2015; Tylee et al. 2017; Ansel et al. 2016; Hewitson et al. 2021; He et al. 2019). Thus, leukocyte gene expression holds the potential for the objective identification of molecular subtypes of ASD. In analyses of leukocyte gene co-expression, ASD-associated module eigengene values were significantly correlated with abnormal early brain growth and enriched in genes related to cell cycle, translation, and immune networks and pathways. These gene sets are very accurate classifiers of ASD vs. typically developing toddlers (TD)(Pramparo, Pierce, et al. 2015).

Leukocyte gene expression offers a non-invasive and clinically practicable avenue for understanding aspects of ASD cell biology, including those that could be ASD-relevant, ASD-specific, robust, and ASD-diagnostic or -prognostic. However, for the clinical translational potential of leukocyte transcriptomics to lead to robust and rigorous classifiers, high standards for verifying such classifiers should be implemented.

In the second chapter, I developed, operationalized, and tested a rigorous analytic pipeline to identify molecular diagnostic classifiers for ASD using leukocyte gene expression. Using additional clinical data, I verified that our composite gene expression classifier was unbiased against common confounding factors (age, race and ethnicity). Using this platform on leukocyte transcriptomics from male ASD and typically developing (TD) toddlers at ages 1-4, my colleague and I systematically analyzed the classification performance of 42,840 different models composed of 3,570 different feature selection sets and 12 commonly-used classification methods (Figure 2.1 and Appendix Figure 2.1). Through this, we developed a predictive ensemble diagnostic classifier of male ASD toddlers. Additionally, using targeted DNA sequencing of the coding regions for sets of ASD and neurodevelopmental disorder risk genes using single-molecule molecular inversion probes (smMIPs)(Wang et al. 2020; Stessman et al. 2017), my collaborators examined the diagnostic classifier value of presence or absence of a subset of ASD risk gene mutations in our ASD and TD subjects and whether toddlers with ASD risk gene mutations differ in classifier expression from those without such mutations.

In addition to the ensemble classifier using gene expression data, I also explored the potential diagnosis of ASD using facial emotional data. Since ASD was first identified in 1943, two stereotypes concerning the emotional lives of children affected by the disorder have prevailed: one in which negative emotions dominate and the other in which emotional expressions are muted, particularly positively valenced emotions(Cooper and Michels 1988; Harms, Martin, and Wallace 2010; Uljarevic and Hamilton 2013; Langdell 1978; Begeer et al. 2008; Kennedy and Adolphs 2012). Not surprisingly, autism research has focused on examining either a negative emotionality bias or an attenuation of positive emotion(Castelli 2005; Atkinson 2009; Philip et al. 2010). Recent evidence, however, shows that autistic individuals may not necessarily differ in expression intensity of emotions, nor have negative emotionality bias(Macari et al. 2018; Trevisan, Hoskyn, and Birmingham 2018; Press, Richardson, and Bird 2010; Deschamps et al. 2015; Weiss et al. 2019; Rozga et al. 2013). For example, in a recent study, evoked expressions in response to funny videos in ASD adults were rated as more intense, although less natural, than TD expressions(Faso, Sasson, and Pinkham 2015). The clustering based on evoked action unit intensity identified an ASD subgroup, proposed as an “over-responsive group”, that expresses more intense positive facial expressions than the TD group in response to the videos(Bangerter et al. 2020).

At the same time, research has shown that individuals with ASD sometimes exhibit emotions that are incongruent with real-world events, as evidenced by executing atypical expressions patterns(Carpenter et al. 2021; Brewer et al. 2016; Faso, Sasson, and Pinkham 2015; Weiss et al. 2019; Rozga et al. 2013). Indeed, children and adults with ASD exhibit reduced, atypical, or delayed spontaneous mimicry responses to photographs and videos of emotional facial expressions(Zampella, Bennetto, and Herrington 2020a; Rieffe, Meerum Terwogt, and Stockmann 2000). Specific facial behaviors, including eye contact, smiling, and eyebrow movements, can distinguish ASD subjects from control participants. Such changes are relevant to biological hypotheses about abnormalities in the medial prefrontal cortex and an optical network within the occipitotemporal cortex(Moore et al. 2018). It suggests that those differences in facial behavior could lead to potential phenotypic biomarkers of ASD.

In order to measure facial behavior objectively and quantitatively, automated facial analysis tools have been developed to empower the analysis in different parts of a range of disorders and conditions(Leo et al. 2018; LoBue and Thrasher 2014; Sariyanidi et al. 2020; Bangerter et al. 2020; Jacques et al. 2022). It enables scientists to measure the facial responses to emotional stimuli in an efficient, granular, and objective perspective(Bangerter et al. 2020; Pulido-Castro et al. 2021; Baltrusaitis et al. 2018). However, the efficacy of using the automatic facial expression test as an early screening tool for ASD remains underexplored(Jacques et al. 2022). Most of the established facial expression tests require the interaction between the psychologist and the child(Zampella, Bennetto, and Herrington 2020b). This limits researchers’ and clinicians’ ability to assess critical behaviors and measure differences across individuals, contexts, or time. Thus, there is a lack of established automatic methods for operationalizing toddlers’ emotional reciprocity objectively or granularly.

Preferential-looking paradigms have been successfully adopted to identify visual attention preferences in ASD(Kaliukhovich et al. 2021; Pierce et al. 2016; Wen et al. 2022). One such preferential-looking test, the GeoPref test, found that a subset of ASD toddlers strongly preferred geometric images when presented with social and geometric motion images(Pierce et al. 2016). The toddlers with a higher preference for geometric images demonstrated greater symptom severity and fewer gaze shifts at school age(Bacon et al. 2020). The success of the GeoPref Test as a symptom severity prognostic tool encourages us to study the toddler's facial emotional response to different movie scenes.

In the third chapter, I leveraged a new eye-tracking test called 'The Joint Attention Test' (Andreason et al., In preparation) that features a female speaking in a child-friendly, emotionally valent voice while engaging with various toys and objects. I utilized freely available software, Openface 2.0(Baltrusaitis et al. 2018) and Emonet(Toisoul et al. 2021), to analyze webcam images and measure faction action unit intensity (Figure 3.1). I then used the corresponding features to train a classifier to differentiate between ASD and non-ASD subjects. The classifier was unbiased against common confounding factors (age, race, and ethnicity). Further, I tested the combination of the classifiers with the GeoPref percent fixation(Wen et al. 2022) on a geometric image score shown to have high specificity and good PPV in predicting ASD diagnosis. The final unsupervised clustering analysis, including the classifier score, eye-tracking data, and social behavior data, provided further insight into the clinical behavior heterogeneity among different subgroups.

## References

Aguilar-Lacasaña, Sofía, Natàlia Vilor-Tejedor, Philip R. Jansen, Mònica López-Vicente, Mariona Bustamante, Miguel Burgaleta, Jordi Sunyer, and Silvia Alemany. 2022. "Polygenic Risk for ADHD and ASD and Their Relation with Cognitive Measures in School Children." *Psychological Medicine* 52 (7): 1356–64.

Alocci, Davide, Marie Ghraichy, Elena Barletta, Alessandra Gastaldello, Julien Mariethoz, and Frederique Lisacek. 2018. "Understanding the Glycome: An Interactive View of Glycosylation from Glycocompositions to Glycoepitopes." *Glycobiology* 28 (6): 349–62.

Angel, Peggi M., Anand Mehta, Kim Norris-Caneda, and Richard R. Drake. 2017. "MALDI Imaging Mass Spectrometry of N-Glycans and Tryptic Peptides from the Same Formalin-Fixed, Paraffin-Embedded Tissue Section." *Methods in Molecular Biology*. https://doi.org/10.1007/7651_2017_81.

Ansel, Ashley, Joshua P. Rosenzweig, Philip D. Zisman, Michal Melamed, and Benjamin Gesundheit. 2016. "Variation in Gene Expression in Autism Spectrum Disorders: An Extensive Review of Transcriptomic Studies." *Frontiers in Neuroscience* 10: 601.

Antaki, Danny, James Guevara, Adam X. Maihofer, Marieke Klein, Madhusudan Gujral, Jakob Grove, Caitlin E. Carey, et al. 2022. "Publisher Correction: A Phenotypic Spectrum of Autism Is Attributable to the Combined Effects of Rare Variants, Polygenic Risk and Sex." *Nature Genetics* 54 (8): 1259.

Aoki-Kinoshita, Kiyoko, Sanjay Agravat, Nobuyuki P. Aoki, Sena Arpinar, Richard D. Cummings, Akihiro Fujita, Noriaki Fujita, et al. 2015. "GlyTouCan 1.0--The International Glycan Structure Repository." *Nucleic Acids Research* 44 (D1): D1237–42.

Apweiler, R., H. Hermjakob, and N. Sharon. 1999. "On the Frequency of Protein Glycosylation, as Deduced from Analysis of the SWISS-PROT Database." *Biochimica et Biophysica Acta* 1473 (1): 4–8.

Ashwood, Christopher, Brian Pratt, Brendan X. MacLean, Rebekah L. Gundry, and Nicolle H. Packer. 2019. "Standardization of PGC-LC-MS-Based Glycomics for Sample Specific Glycotyping." *The Analyst* 144 (11): 3601–12.

Atkinson, Anthony P. 2009. "Impaired Recognition of Emotions from Body Movements Is Associated with

Elevated Motion Coherence Thresholds in Autism Spectrum Disorders." *Neuropsychologia* 47 (13): 3023–29.

Autism and Developmental Disabilities Monitoring Network Surveillance Year 2006 Principal Investigators, and Centers for Disease Control and Prevention (CDC). 2009. "Prevalence of Autism Spectrum Disorders - Autism and Developmental Disabilities Monitoring Network, United States, 2006." *Morbidity and Mortality Weekly Report. Surveillance Summaries* 58 (10): 1–20.

Bacon, Elizabeth C., Eric Courchesne, Cynthia Carter Barnes, Debra Cha, Sunny Pence, Laura Schreibman, Aubyn C. Stahmer, and Karen Pierce. 2018. "Rethinking the Idea of Late Autism Spectrum Disorder Onset." *Development and Psychopathology* 30 (2): 553–69.

Bacon, Elizabeth C., Adrienne Moore, Quimby Lee, Cynthia Carter Barnes, Eric Courchesne, and Karen Pierce. 2020. "Identifying Prognostic Markers in Autism Spectrum Disorder Using Eye Tracking." *Autism: The International Journal of Research and Practice* 24 (3): 658–69.

Bacon, Elizabeth C., Suzanna Osuna, Eric Courchesne, and Karen Pierce. 2019. "Naturalistic Language Sampling to Characterize the Language Abilities of 3-Year-Olds with Autism Spectrum Disorder." *Autism: The International Journal of Research and Practice* 23 (3): 699–712.

Bai, Dan, Benjamin Hon Kei Yip, Gayle C. Windham, Andre Sourander, Richard Francis, Rinat Yoffe, Emma Glasson, et al. 2019. "Association of Genetic and Environmental Factors With Autism in a 5-Country Cohort." *JAMA Psychiatry* 76 (10): 1035–43.

Baio, Jon, Lisa Wiggins, Deborah L. Christensen, Matthew J. Maenner, Julie Daniels, Zachary Warren, Margaret Kurzius-Spencer, et al. 2018. "Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014." *Morbidity and Mortality Weekly Report. Surveillance Summaries* 67 (6): 1–23.

Baltrusaitis, Tadas, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. "OpenFace 2.0: Facial Behavior Analysis Toolkit." In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 59–66.

Bal, Vanessa H., So-Hyun Kim, Megan Fok, and Catherine Lord. 2019. "Autism Spectrum Disorder

Symptoms from Ages 2 to 19 Years: Implications for Diagnosing Adolescents and Young Adults." *Autism Research: Official Journal of the International Society for Autism Research* 12 (1): 89–99.

Bangerter, Abigail, Meenakshi Chatterjee, Joseph Manfredonia, Nikolay V. Manyakov, Seth Ness, Matthew A. Boice, Andrew Skalkin, et al. 2020. "Automated Recognition of Spontaneous Facial Expression in Individuals with Autism Spectrum Disorder: Parsing Response Variability." *Molecular Autism* 11 (1): 31.

Begeer, Sander, Hans M. Koot, Carolien Rieffe, Mark Meerum Terwogt, and Hedy Stegge. 2008. "Emotional Competence in Children with Autism: Diagnostic Criteria and Empirical Evidence." *Developmental Review: DR* 28 (3): 342–69.

Black, Alyson P., Hongyan Liang, Connor A. West, Mengjun Wang, Harmin P. Herrera, Brian B. Haab, Peggi M. Angel, Richard R. Drake, and Anand S. Mehta. 2019. "A Novel Mass Spectrometry Platform for Multiplexed N-Glycoprotein Biomarker Discovery from Patient Biofluids by Antibody Panel Based N-Glycan Imaging." *Analytical Chemistry*. https://doi.org/10.1021/acs.analchem.9b01445.

Bojar, Daniel, Rani K. Powers, Diogo M. Camacho, and James J. Collins. 2021. "Deep-Learning Resources for Studying Glycan-Mediated Host-Microbe Interactions." *Cell Host & Microbe* 29 (1): 132–44.e3.

Brewer, Rebecca, Federica Biotti, Caroline Catmur, Clare Press, Francesca Happé, Richard Cook, and Geoffrey Bird. 2016. "Can Neurotypical Individuals Read Autistic Facial Expressions? Atypical Production of Emotional Facial Expressions in Autism Spectrum Disorders." *Autism Research: Official Journal of the International Society for Autism Research* 9 (2): 262–71.

Campbell, Matthew P., Terry Nguyen-Khuong, Catherine A. Hayes, Sarah A. Flowers, Kathirvel Alagesan, Daniel Kolarich, Nicolle H. Packer, and Niclas G. Karlsson. 2014. "Validation of the Curation Pipeline of UniCarb-DB: Building a Global Glycan Reference MS/MS Repository." *Biochimica et Biophysica Acta* 1844 (1 Pt A): 108–16.

Campbell, Matthew P., Robyn Peterson, Julien Mariethoz, Elisabeth Gasteiger, Yukie Akune, Kiyoko F. Aoki-Kinoshita, Frederique Lisacek, and Nicolle H. Packer. 2014. "UniCarbKB: Building a Knowledge Platform for Glycoproteomics." *Nucleic Acids Research* 42 (Database issue): D215–21.

Carpenter, Kimberly L. H., Jordan Hahemi, Kathleen Campbell, Steven J. Lippmann, Jeffrey P. Baker, Helen L. Egger, Steven Espinosa, Saritha Vermeer, Guillermo Sapiro, and Geraldine Dawson. 2021. "Digital Behavioral Phenotyping Detects Atypical Pattern of Facial Expression in Toddlers with Autism." *Autism Research: Official Journal of the International Society for Autism Research* 14 (3): 488–99.

Castelli, Fulvia. 2005. "Understanding Emotions from Standardized Facial Expressions in Autism and Normal Development." *Autism: The International Journal of Research and Practice* 9 (4): 428–49.

Čaval, Tomislav, Weihua Tian, Zhang Yang, Henrik Clausen, and Albert J. R. Heck. 2018. "Direct Quality Control of Glycoengineered Erythropoietin Variants." *Nature Communications* 9 (1): 3342.

Ch'ng, Carolyn, Willie Kwok, Sanja Rogic, and Paul Pavlidis. 2015. "Meta-Analysis of Gene Expression in Autism Spectrum Disorder." *Autism Research: Official Journal of the International Society for Autism Research* 8 (5): 593–608.

Christensen, Deborah L., Kim Van Naarden Braun, Jon Baio, Deborah Bilder, Jane Charles, John N. Constantino, Julie Daniels, et al. 2018. "Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2012." *Morbidity and Mortality Weekly Report. Surveillance Summaries* 65 (13): 1–23.

Clarke, T-K, M. K. Lupton, A. M. Fernandez-Pujals, J. Starr, G. Davies, S. Cox, A. Pattie, et al. 2016. "Common Polygenic Risk for Autism Spectrum Disorder (ASD) Is Associated with Cognitive Ability in the General Population." *Molecular Psychiatry* 21 (3): 419–25.

Cooper, Arnold M., and Robert Michels. 1988. "Diagnostic and Statistical Manual of Mental Disorders, 3rd Ed., Revised (DSM-III-R)." *American Journal of Psychiatry* 145 (10): 1300–1301.

Courchesne, Eric, Vahid H. Gazestani, and Nathan E. Lewis. 2020. "Prenatal Origins of ASD: The When, What, and How of ASD Development." *Trends in Neurosciences* 43 (5): 326–42.

Courchesne, Eric, Peter R. Mouton, Michael E. Calhoun, Katerina Semendeferi, Clelia Ahrens-Barbeau, Melodie J. Hallet, Cynthia Carter Barnes, and Karen Pierce. 2011. "Neuron Number and Size in

Prefrontal Cortex of Children with Autism." *JAMA: The Journal of the American Medical Association* 306 (18): 2001–10.

Courchesne, Eric, and Karen Pierce. 2005. "Why the Frontal Cortex in Autism Might Be Talking Only to Itself: Local over-Connectivity but Long-Distance Disconnection." *Current Opinion in Neurobiology* 15 (2): 225–30.

Courchesne, Eric, Karen Pierce, Cynthia M. Schumann, Elizabeth Redcay, Joseph A. Buckwalter, Daniel P. Kennedy, and John Morgan. 2007. "Mapping Early Brain Development in Autism." *Neuron* 56 (2): 399–413.

Courchesne, Eric, Tiziano Pramparo, Vahid H. Gazestani, Michael V. Lombardo, Karen Pierce, and Nathan E. Lewis. 2019. "The ASD Living Biology: From Cell Proliferation to Clinical Phenotype." *Molecular Psychiatry* 24 (1): 88–107.

Cummings, Richard D. 2009. "The Repertoire of Glycan Determinants in the Human Glycome." *Molecular bioSystems* 5 (10): 1087–1104.

Deschamps, P. K. H., L. Coppes, J. L. Kenemans, D. J. L. G. Schutter, and W. Matthys. 2015. "Electromyographic Responses to Emotional Facial Expressions in 6-7 Year Olds with Autism Spectrum Disorders." *Journal of Autism and Developmental Disorders* 45 (2): 354–62.

Diaz-Beltran, L., F. J. Esteban, and D. P. Wall. 2016. "A Common Molecular Signature in ASD Gene Expression: Following Root 66 to Autism." *Translational Psychiatry* 6 (January): e705.

Doherty, Margaret, Evropi Theodoratou, Ian Walsh, Barbara Adamczyk, Henning Stöckmann, Felix Agakov, Maria Timofeeva, et al. 2018. "Plasma N-Glycans in Colorectal Cancer Risk." *Scientific Reports*. https://doi.org/10.1038/s41598-018-26805-7.

Donovan, Alex P. A., and M. Albert Basson. 2017. "The Neuroanatomy of Autism - a Developmental Perspective." *Journal of Anatomy* 230 (1): 4–15.

Enstrom, Amanda M., Lisa Lit, Charity E. Onore, Jeff P. Gregg, Robin L. Hansen, Isaac N. Pessah, Irva Hertz-Picciotto, Judy A. Van de Water, Frank R. Sharp, and Paul Ashwood. 2009. "Altered Gene Expression and Function of Peripheral Blood Natural Killer Cells in Children with Autism." *Brain,* *Behavior, and Immunity* 23 (1): 124–33.

Faso, Daniel J., Noah J. Sasson, and Amy E. Pinkham. 2015. “Evaluating Posed and Evoked Facial Expressions of Emotion from Adults with Autism Spectrum Disorder.” *Journal of Autism and Developmental Disorders* 45 (1): 75–89.

Feliciano, Pamela, Xueya Zhou, Irina Astrovskaya, Tychele N. Turner, Tianyun Wang, Leo Brueggeman, Rebecca Barnard, et al. 2019. “Exome Sequencing of 457 Autism Families Recruited Online Provides Evidence for Autism Risk Genes.” *NPJ Genomic Medicine* 4 (August): 19.

Gabius, Hans-Joachim, Sabine André, Herbert Kaltner, and Hans-Christian Siebert. 2002. “The Sugar Code: Functional Lectinomics.” *Biochimica et Biophysica Acta (BBA) - General Subjects* 1572 (2): 165–77.

Gazestani, Vahid, Austin W. T. Chiang, E. Courchesne, and N. E. Lewis. 2020. “Autism Genetics Perturb Prenatal Neurodevelopment through a Hierarchy of Broadly-Expressed and Brain-Specific Genes.” *bioRxiv*.

Gregg, Jeffrey P., Lisa Lit, Colin A. Baron, Irva Hertz-Picciotto, Wynn Walker, Ryan A. Davis, Lisa A. Croen, et al. 2008. “Gene Expression Changes in Children with Autism.” *Genomics* 91 (1): 22–29.

Grove, Jakob, Stephan Ripke, Thomas D. Als, Manuel Mattheisen, Raymond K. Walters, Hyejung Won, Jonatan Pallesen, et al. 2019. “Identification of Common Genetic Risk Variants for Autism Spectrum Disorder.” *Nature Genetics* 51 (3): 431–44.

Gutierrez, Jahir M., Amir Feizi, Shangzhong Li, Thomas B. Kallehauge, Hooman Hefzi, Lise M. Grav, Daniel Ley, et al. 2018. “Genome-Scale Reconstructions of the Mammalian Secretory Pathway Predict Metabolic Costs and Limitations of Protein Secretion.” *bioRxiv*. https://doi.org/10.1101/351387.

Harms, Madeline B., Alex Martin, and Gregory L. Wallace. 2010. “Facial Emotion Recognition in Autism Spectrum Disorders: A Review of Behavioral and Neuroimaging Studies.” *Neuropsychology Review* 20 (3): 290–322.

Hewitson, Laura, Jeremy A. Mathews, Morgan Devlin, Claire Schutte, Jeon Lee, and Dwight C. German. 2021. “Blood Biomarker Discovery for Autism Spectrum Disorder: A Proteomic Analysis.” *PloS One*

16 (2): e0246581.

He, Yi, Yuan Zhou, Wei Ma, and Juan Wang. 2019. “An Integrated Transcriptomic Analysis of Autism Spectrum Disorder.” *Scientific Reports* 9 (1): 11818.

Holst, Stephanie, Anna J. M. Deuss, Gabi W. van Pelt, Sandra J. van Vliet, Juan J. Garcia-Vallejo, Carolien A. M. Koeleman, André M. Deelder, et al. 2016. “N-Glycosylation Profiling of Colorectal Cancer Cell Lines Reveals Association of Fucosylation with Differentiation and Caudal Type Homebox 1 (CDX1)/Villin mRNA Expression.” *Molecular & Cellular Proteomics: MCP* 15 (1): 124–40.

Holst, Stephanie, Gabi W. van Pelt, Wilma E. Mesker, Rob A. Tollenaar, Ana I. Belo, Irma van Die, Yoann Rombouts, and Manfred Wuhrer. 2017. “High-Throughput and High-Sensitivity Mass Spectrometry-Based N-Glycomics of Mammalian Cells.” *Methods in Molecular Biology*. https://doi.org/10.1007/978-1-4939-6493-2_14.

Hosoda, Masae, Yushi Takahashi, Masaaki Shiota, Daisuke Shinmachi, Renji Inomoto, Shinichi Higashimoto, and Kiyoko F. Aoki-Kinoshita. 2018. “MCAW-DB: A Glycan Profile Database Capturing the Ambiguity of Glycan Recognition Patterns.” *Carbohydrate Research* 464 (July): 44–56.

Hou, Wenpin, Yushan Qiu, Nobuyuki Hashimoto, Wai-Ki Ching, and Kiyoko F. Aoki-Kinoshita. 2016. “A Systematic Framework to Derive N-Glycan Biosynthesis Process and the Automated Construction of Glycosylation Networks.” *BMC Bioinformatics* 17 Suppl 7 (July): 240.

“Human Gene Module.” n.d. SFARI Gene. Accessed August 25, 2022. https://gene-archive.sfari.org/database/human-gene/.

Jacques, Claudine, Valérie Courchesne, Suzanne Mineau, Michelle Dawson, and Laurent Mottron. 2022. “Positive, Negative, Neutral-or Unknown? The Perceived Valence of Emotions Expressed by Young Autistic Children in a Novel Context Suited to Autism.” *Autism: The International Journal of Research and Practice*, February, 13623613211068221.

Kaliukhovich, Dzmitry A., Nikolay V. Manyakov, Abigail Bangerter, Seth Ness, Andrew Skalkin, Matthew Boice, Matthew S. Goodwin, et al. 2021. “Visual Preference for Biological Motion in Children and

Adults with Autism Spectrum Disorder: An Eye-Tracking Study." *Journal of Autism and Developmental Disorders* 51 (7): 2369–80.

Kaushik, Gaurav, and Konstantinos S. Zarbalis. 2016. "Prenatal Neurogenesis in Autism Spectrum Disorders." *Frontiers in Chemistry*. https://doi.org/10.3389/fchem.2016.00012.

Kennedy, Daniel P., and Ralph Adolphs. 2012. "Perception of Emotions from Facial Expressions in High-Functioning Adults with Autism." *Neuropsychologia* 50 (14): 3313–19.

Khoury, George A., Richard C. Baliban, and Christodoulos A. Floudas. 2011. "Proteome-Wide Post-Translational Modification Statistics: Frequency Analysis and Curation of the Swiss-Prot Database." *Scientific Reports* 1 (September). https://doi.org/10.1038/srep00090.

Klei, Lambertus, Lora Lee McClain, Behrang Mahjani, Klea Panayidou, Silvia De Rubeis, Anna-Carin Säll Grahnat, Gun Karlsson, et al. 2021. "How Rare and Common Risk Variation Jointly Affect Liability for Autism Spectrum Disorder." *Molecular Autism* 12 (1): 66.

Klein, Joshua, Luis Carvalho, and Joseph Zaia. 2018. "Application of Network Smoothing to Glycan LC-MS Profiling." *Bioinformatics* 34 (20): 3511–18.

Kong, Sek Won, Christin D. Collins, Yuko Shimizu-Motohashi, Ingrid A. Holm, Malcolm G. Campbell, In-Hee Lee, Stephanie J. Brewster, et al. 2012. "Characteristics and Predictive Value of Blood Transcriptome Signature in Males with Autism Spectrum Disorders." *PloS One* 7 (12): e49475.

Krambeck, Frederick J., Sandra V. Bennun, Mikael R. Andersen, and Michael J. Betenbaugh. 2017. "Model-Based Analysis of N-Glycosylation in Chinese Hamster Ovary Cells." *PLOS ONE*. https://doi.org/10.1371/journal.pone.0175376.

Kremkow, Benjamin G., and Kelvin H. Lee. 2018. "Glyco-Mapper: A Chinese Hamster Ovary (CHO) Genome-Specific Glycosylation Prediction Tool." *Metabolic Engineering* 47 (May): 134–42.

Krishnan, Arjun, Ran Zhang, Victoria Yao, Chandra L. Theesfeld, Aaron K. Wong, Alicja Tadych, Natalia Volfovsky, Alan Packer, Alex Lash, and Olga G. Troyanskaya. 2016. "Genome-Wide Prediction and Functional Characterization of the Genetic Basis of Autism Spectrum Disorder." *Nature Neuroscience* 19 (11): 1454–62.

Langdell, T. 1978. “Recognition of Faces: An Approach to the Study of Autism.” *Journal of Child Psychology and Psychiatry, and Allied Disciplines* 19 (3): 255–68.

Lee, Samuel C., Thomas P. Quinn, Jerry Lai, Sek Won Kong, Irva Hertz-Picciotto, Stephen J. Glatt, Tamsyn M. Crowley, Svetha Venkatesh, and Thin Nguyen. 2019. “Solving for X: Evidence for Sex-Specific Autism Biomarkers across Multiple Transcriptomic Studies.” *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics* 180 (6): 377–89.

Leo, Marco, Pierluigi Carcagnì, Cosimo Distante, Paolo Spagnolo, Pier Luigi Mazzeo, Anna Chiara Rosato, Serena Petrocchi, et al. 2018. “Computational Assessment of Facial Expression Production in ASD Children.” *Sensors* 18 (11).

LoBue, Vanessa, and Cat Thrasher. 2014. “The Child Affective Facial Expression (CAFE) Set: Validity and Reliability from Untrained Adults.” *Frontiers in Psychology* 5: 1532.

Lombardo, Michael V., Meng-Chuan Lai, and Simon Baron-Cohen. 2019. “Big Data Approaches to Decomposing Heterogeneity across the Autism Spectrum.” *Molecular Psychiatry* 24 (10): 1435–50.

Macari, Suzanne, Lauren DiNicola, Finola Kane-Grade, Emily Prince, Angelina Vernetti, Kelly Powell, Scuddy Fontenelle, and Katarzyna Chawarska. 2018. “Emotional Expressivity in Toddlers With Autism Spectrum Disorder.” *Journal of the American Academy of Child & Adolescent Psychiatry*. https://doi.org/10.1016/j.jaac.2018.07.872.

Maenner, Matthew J., Kelly A. Shaw, Jon Baio, EdS1, Anita Washington, Mary Patrick, Monica DiRienzo, et al. 2020. “Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2016.” *Morbidity and Mortality Weekly Report. Surveillance Summaries* 69 (4): 1–12.

Marchetto, Maria C., Haim Belinson, Yuan Tian, Beatriz C. Freitas, Chen Fu, Krishna Vadodaria, Patricia Beltrao-Braga, et al. 2017. “Altered Proliferation and Networks in Neural Cells Derived from Idiopathic Autistic Individuals.” *Molecular Psychiatry* 22 (6): 820–35.

Maxwell, Evan, Yan Tan, Yuxiang Tan, Han Hu, Gary Benson, Konstantin Aizikov, Shannon Conley, et al. 2012. “GlycReSoft: A Software Package for Automated Recognition of Glycans from LC/MS Data.” *PloS One* 7 (9): e45474.

Mohammad, Mahmoud A., Darryl L. Hadsell, and Morey W. Haymond. 2012. “Gene Regulation of UDP-Galactose Synthesis and Transport: Potential Rate-Limiting Processes in Initiation of Milk Production in Humans.” *American Journal of Physiology. Endocrinology and Metabolism* 303 (3): E365–76.

Moore, Adrienne, Madeline Wozniak, Andrew Yousef, Cindy Carter Barnes, Debra Cha, Eric Courchesne, and Karen Pierce. 2018. “The Geometric Preference Subtype in ASD: Identifying a Consistent, Early-Emerging Phenomenon through Eye Tracking.” *Molecular Autism* 9 (March): 19.

Packer, Alan. 2016. “Neocortical Neurogenesis and the Etiology of Autism Spectrum Disorder.” *Neuroscience and Biobehavioral Reviews* 64 (May): 185–95.

Parikshak, Neelroop N., Rui Luo, Alice Zhang, Hyejung Won, Jennifer K. Lowe, Vijayendran Chandran, Steve Horvath, and Daniel H. Geschwind. 2013. “Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism.” *Cell*. https://doi.org/10.1016/j.cell.2013.10.031.

Philip, R. C. M., H. C. Whalley, A. C. Stanfield, R. Sprengelmeyer, I. M. Santos, A. W. Young, A. P. Atkinson, et al. 2010. “Deficits in Facial, Body Movement and Vocal Emotional Processing in Autism Spectrum Disorders.” *Psychological Medicine* 40 (11): 1919–29.

Pierce, Karen, Steven Marinero, Roxana Hazin, Benjamin McKenna, Cynthia Carter Barnes, and Ajith Malige. 2016. “Eye Tracking Reveals Abnormal Visual Preference for Geometric Images as an Early Biomarker of an Autism Spectrum Disorder Subtype Associated With Increased Symptom Severity.” *Biological Psychiatry* 79 (8): 657–66.

Pramparo, Tiziano, Michael V. Lombardo, Kathleen Campbell, Cynthia Carter Barnes, Steven Marinero, Stephanie Solso, Julia Young, et al. 2015. “Cell Cycle Networks Link Gene Expression Dysregulation, Mutation, and Brain Maldevelopment in Autistic Toddlers.” *Molecular Systems Biology* 11 (12): 841.

Pramparo, Tiziano, Karen Pierce, Michael V. Lombardo, Cynthia Carter Barnes, Steven Marinero, Clelia Ahrens-Barbeau, Sarah S. Murray, Linda Lopez, Ronghui Xu, and Eric Courchesne. 2015. “Prediction of Autism by Translation and Immune/inflammation Coexpressed Genes in Toddlers from Pediatric

Community Practices." *JAMA Psychiatry* 72 (4): 386–94.

Press, Clare, Daniel Richardson, and Geoffrey Bird. 2010. "Intact Imitation of Emotional Facial Actions in Autism Spectrum Conditions." *Neuropsychologia* 48 (11): 3291–97.

Pulido-Castro, Sergio, Nubia Palacios-Quecan, Michelle P. Ballen-Cardenas, Sandra Cancino-Suárez, Alejandra Rizo-Arévalo, and Juan M. López López. 2021. "Ensemble of Machine Learning Models for an Improved Facial Emotion Recognition." In *2021 IEEE URUCON*, 512–16.

Rademacher, Christoph, and James C. Paulson. 2012. "Glycan Fingerprints: Calculating Diversity in Glycan Libraries." *ACS Chemical Biology* 7 (5): 829–34.

Reiding, Karli R., Dennis Blank, Dennis M. Kuijper, André M. Deelder, and Manfred Wuhrer. 2014. "High-Throughput Profiling of Protein N-Glycosylation by MALDI-TOF-MS Employing Linkage-Specific Sialic Acid Esterification." *Analytical Chemistry* 86 (12): 5784–93.

Reiding, Karli R., Albert Bondt, René Hennig, Richard A. Gardner, Roisin O'Flaherty, Irena Trbojević-Akmačić, Archana Shubhakar, et al. 2019. "High-Throughput Serum N-Glycomics: Method Comparison and Application to Study Rheumatoid Arthritis and Pregnancy-Associated Changes." *Molecular & Cellular Proteomics: MCP* 18 (1): 3–15.

Rieffe, C., M. Meerum Terwogt, and L. Stockmann. 2000. "Understanding Atypical Emotions among Children with Autism." *Journal of Autism and Developmental Disorders* 30 (3): 195–203.

Riley, Nicholas M., Alexander S. Hebert, Michael S. Westphall, and Joshua J. Coon. 2019. "Capturing Site-Specific Heterogeneity with Large-Scale N-Glycoproteome Analysis." *Nature Communications* 10 (1): 1311.

Robinson, Elise B., Beate St Pourcain, Verneri Anttila, Jack A. Kosmicki, Brendan Bulik-Sullivan, Jakob Grove, Julian Maller, et al. 2016. "Genetic Risk for Autism Spectrum Disorders and Neuropsychiatric Variation in the General Population." *Nature Genetics* 48 (5): 552–55.

RodrÍguez, Ernesto, Sjoerd T. T. Schetters, and Yvette van Kooyk. 2018. "The Tumour Glyco-Code as a Novel Immune Checkpoint for Immunotherapy." *Nature Reviews. Immunology* 18 (3): 204–11.

Rozga, Agata, Tricia Z. King, Richard W. Vuduc, and Diana L. Robins. 2013. "Undifferentiated Facial

Electromyography Responses to Dynamic, Audio-Visual Emotion Displays in Individuals with Autism Spectrum Disorders." *Developmental Science* 16 (4): 499–514.

Sariyanidi, Evangelos, Casey J. Zampella, Robert T. Schultz, and Birkan Tunc. 2020. "Can Facial Pose and Expression Be Separated with Weak Perspective Camera?" *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2020 (June): 7171–80.

Satterstrom, F. Kyle, Jack A. Kosmicki, Jiebiao Wang, Michael S. Breen, Silvia De Rubeis, Joon-Yong An, Minshi Peng, et al. 2020. "Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism." *Cell* 180 (3): 568–84.e23.

Sharapov, Sodbo, Yakov Tsepilov, Lucija Klaric, Massimo Mangino, Gaurav Thareja, Mirna Simurina, Concetta Dagostino, et al. 2018. "Defining the Genetic Control of Human Blood Plasma N-Glycome Using Genome-Wide Association Study." *bioRxiv*. https://doi.org/10.1101/365486.

Spahn, Philipp N., and Nathan E. Lewis. 2014. "Systems Glycobiology for Glycoengineering." *Current Opinion in Biotechnology* 30 (December): 218–24.

Stessman, Holly A. F., Bo Xiong, Bradley P. Coe, Tianyun Wang, Kendra Hoekzema, Michaela Fenckova, Malin Kvarnung, et al. 2017. "Targeted Sequencing Identifies 91 Neurodevelopmental-Disorder Risk Genes with Autism and Developmental-Disability Biases." *Nature Genetics* 49 (4): 515–26.

Stoner, Rich, Maggie L. Chow, Maureen P. Boyle, Susan M. Sunkin, Peter R. Mouton, Subhojit Roy, Anthony Wynshaw-Boris, Sophia A. Colamarino, Ed S. Lein, and Eric Courchesne. 2014. "Patches of Disorganization in the Neocortex of Children with Autism." *New England Journal of Medicine*. https://doi.org/10.1056/nejmoa1307491.

Toisoul, Antoine, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. 2021. "Estimation of Continuous Valence and Arousal Levels from Faces in Naturalistic Conditions." *Nature Machine Intelligence* 3 (1): 42–50.

Trevisan, Dominic A., Maureen Hoskyn, and Elina Birmingham. 2018. "Facial Expression Production in Autism: A Meta-Analysis." *Autism Research: Official Journal of the International Society for Autism*

*Research* 11 (12): 1586–1601.

Tylee, Daniel S., Jonathan L. Hess, Thomas P. Quinn, Rahul Barve, Hailiang Huang, Yanli Zhang-James, Jeffrey Chang, et al. 2017. “Blood Transcriptomic Comparison of Individuals with and without Autism Spectrum Disorder: A Combined-Samples Mega-Analysis.” *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics* 174 (3): 181–201.

Uljarevic, Mirko, and Antonia Hamilton. 2013. “Recognition of Emotions in Autism: A Formal Meta-Analysis.” *Journal of Autism and Developmental Disorders* 43 (7): 1517–26.

Wang, Tianyun, Kendra Hoekzema, Davide Vecchio, Huidan Wu, Arvis Sulovari, Bradley P. Coe, Madelyn A. Gillentine, et al. 2020. “Large-Scale Targeted Sequencing Identifies Risk Genes for Neurodevelopmental Disorders.” *Nature Communications* 11 (1): 4932.

Weiss, Elisabeth M., Christian Rominger, Ellen Hofer, Andreas Fink, and Ilona Papousek. 2019. “Less Differentiated Facial Responses to Naturalistic Films of Another Person’s Emotional Expressions in Adolescents and Adults with High-Functioning Autism Spectrum Disorder.” *Progress in Neuro-Psychopharmacology & Biological Psychiatry* 89 (March): 341–46.

Wen, Teresa H., Amanda Cheng, Charlene Andreason, Javad Zahiri, Yaqiong Xiao, Ronghui Xu, Bokan Bao, et al. 2022. “Large Scale Validation of an Early-Age Eye-Tracking Biomarker of an Autism Spectrum Disorder Subtype.” *Scientific Reports* 12 (1): 4253.

Willsey, A. Jeremy, Stephan J. Sanders, Mingfeng Li, Shan Dong, Andrew T. Tebbenkamp, Rebecca A. Muhle, Steven K. Reilly, et al. 2013. “Coexpression Networks Implicate Human Midfetal Deep Cortical Projection Neurons in the Pathogenesis of Autism.” *Cell* 155 (5): 997–1007.

Wohlschlager, Therese, Kai Scheffler, Ines C. Forstenlehner, Wolfgang Skala, Stefan Senn, Eugen Damoc, Johann Holzmann, and Christian G. Huber. 2018. “Native Mass Spectrometry Combined with Enzymatic Dissection Unravels Glycoform Heterogeneity of Biopharmaceuticals.” *Nature Communications*. https://doi.org/10.1038/s41467-018-04061-7.

Yang, Zhang, Shengjun Wang, Adnan Halim, Morten Alder Schulz, Morten Frodin, Shamim H. Rahman, Malene B. Vester-Christensen, et al. 2015. “Engineered CHO Cells for Production of Diverse, Homogeneous Glycoproteins.” *Nature Biotechnology* 33 (8): 842–44.

York, William S., Raja Mazumder, Rene Ranzinger, Nathan Edwards, Robel Kahsay, Kiyoko F. Aoki-Kinoshita, Matthew P. Campbell, et al. 2019. “GlyGen: Computational and Informatics Resources for Glycoscience.” *Glycobiology*. https://doi.org/10.1093/glycob/cwz080.

Zampella, Casey J., Loisa Bennetto, and John D. Herrington. 2020a. “Computer Vision Analysis of Reduced Interpersonal Affect Coordination in Youth With Autism Spectrum Disorder.” *Autism Research*. https://doi.org/10.1002/aur.2334.

# CHAPTER 1: Correcting for sparsity and interdependence in glycomics by accounting for glycan biosynthesis

## Abstract

Glycans are fundamental cellular building blocks, involved in many organismal functions. Advances in glycomics are elucidating the essential roles of glycans. Still, it remains challenging to properly analyze large glycomics datasets, since the abundance of each glycan is dependent on many other glycans that share many intermediate biosynthetic steps. Furthermore, the overlap of measured glycans can be low across samples. We address these challenges with GlyCompare, a glycomic data analysis approach that accounts for shared biosynthetic steps for all measured glycans to correct for sparsity and non-independence in glycomics, which enables direct comparison of different glycoprofiles and increases statistical power. Using GlyCompare, we study diverse N-glycan profiles from glycoengineered erythropoietin. We obtain biologically meaningful clustering of mutant cell glycoprofiles and identify knockout-specific effects of fucosyltransferase mutants on tetra-antennary structures. We further analyze human milk oligosaccharide profiles and find mother's fucosyltransferase-dependent secretor-status indirectly impact the sialylation. Finally, we apply our method on mucin-type O-glycans, gangliosides, and site-specific compositional glycosylation data to reveal tissues and disease-specific glycan presentations. Our substructure-oriented approach will enable researchers to take full advantage of the growing power and size of glycomics data.

## Introduction

Glycosylation is a complex post-translational modification and it decorates one-fifth to one-half of eukaryotic proteins(Khoury, Baliban, and Floudas 2011; Apweiler, Hermjakob, and Sharon 1999). The diversified glycans account for 12-25% of dry cell mass and have essential functional and pathological roles(RodrÍguez, Schetters, and van Kooyk 2018; Gutierrez et al. 2018). Despite their importance, glycans have complex structures that are difficult to study. The complex structures of glycans arise from a non-template-driven synthesis through a biosynthetic network involving dozens of enzymes. A simple change of a single intermediate glycan or glycosyltransferase will have cascading impacts on the final glycans obtained(Gabius et al. 2002; Spahn and Lewis 2014). Unfortunately, current data analysis approaches for glycoprofiling and glycomic data lack the critical systems perspective to decode the interdependence of glycans easily(Reiding et al. 2014, 2019; Doherty et al. 2018; Wohlschlager et al. 2018; Black et al. 2019; Ashwood et al. 2019). It is important to understand the network behind the glycoprofiles to understand the behavior of the process better.

New tools aiding in the acquisition and aggregation of glycoprofiles are emerging, making large-scale comparisons of glycoprofiles possible. Advances in mass spectrometry now enable the rapid generation of many glycoprofiles with detailed glycan composition and structure predictions(Reiding et al. 2014, 2019; Wohlschlager et al. 2018; Black et al. 2019; Ashwood et al. 2019; Maxwell et al. 2012; Hou et al. 2016; Kremkow and Lee 2018; Krambeck et al. 2017; Holst et al. 2017; Angel et al. 2017), exposing the complex and heterogeneous glycosylation patterns on lipids and proteins(Reiding et al. 2019; Doherty et al. 2018; Black et al. 2019; Cummings 2009; Holst et al. 2016; Čaval et al. 2018; Riley et al. 2019). Large glycoprofile datasets and supporting databases are also emerging, including GlyTouCan(Aoki-Kinoshita et al. 2015), UniCarb-DB(Campbell, Nguyen-Khuong, et al. 2014), GlyGen(York et al. 2019), and UniCarbKB(Campbell, Peterson, et al. 2014).

These technologies and databases facilitate efforts to associate glycans with disease and other phenotypes. However, the rapid and accurate comparison of glycoprofiles can be challenging with the size, sparsity and heterogeneity of such datasets(Reiding et al. 2019; Doherty et al. 2018; Black et al. 2019; Holst et al. 2016; Čaval et al. 2018; Riley et al. 2019; Yang et al. 2015). A glycoprofile provides glycan structure and abundance information, and each glycan is usually treated as an independent entity. Furthermore, in any one glycoprofile, only a tiny percentage of all possible glycans may be detected(Reiding et al. 2019; Doherty et al. 2018; Black et al. 2019; Holst et al. 2016; Yang et al. 2015). Thus, if there is a significant perturbation to glycosylation in a dataset, only a few glycans, if any, may overlap between samples. However, these non-overlapping glycans may only differ in their synthesis by as few as one enzymatic step. Thus, it requires deliberate manual coding to make them comparable(Reiding et al. 2014, 2019; Doherty et al. 2018; Wohlschlager et al. 2018; Black et al. 2019; Ashwood et al. 2019; Holst et al. 2016; Yang et al. 2015). These properties of glycomics data may not be problematic in the studies of individual glycans and their downstream effects on other biological processes. However, this may be a problem in determining the sources of changes in glycan abundance by using large amounts of data(Reiding et al. 2019; Doherty et al. 2018; Yang et al. 2015; Benedetti et al. 2017). Since many methods assume data independence (e.g., t-tests, ANOVA, etc.), their application to glycomics can lead to decreased statistical power or erroneous results.

Previous studies have investigated the similarities across glycans by using glycan motifs. Scientists are using glycan fingerprinting to describe glycan diversity in databases(Rademacher and Paulson 2012; Bojar et al. 2021), align glycan structures(Hosoda et al. 2018), identify glycan epitopes in glycoprofiles(Alocci et al. 2018) and lectin profiles(Khoury, Baliban, and Floudas 2011), deconstruct LC-MS data to quantify glycan abundance(Klein, Carvalho, and Zaia 2018), or compare glycans in glycoprofiles(Sharapov et al. 2018). These tools use information on glycan composition or epitopes. However, the accounting of shared biosynthetic steps could provide complete context to all glycan epitopes. That context includes connecting all glycans to the enzymes involved in their synthesis, the order of the enzyme reactions, and information on competition for glycan substrates. Thus, a generalized substructure approach could facilitate the study of large numbers of glycoprofiles by connecting them to the shared mechanisms involved in making each glycan.

In this work we present GlyCompare, a method enabling the rapid and scalable analysis and comparison of multiple glycoprofiles, while accounting for the biosynthetic similarities of each glycan. We propose glycan substructures, or intermediates, as the appropriate functional units for meaningful glycoprofile comparisons, since each substructure can capture one step in the complex process of glycan synthesis, which accounts for the shared dependencies across glycans. This approach addresses current challenges in sparsity and hidden interdependence across glycomic samples and will facilitate discovering mechanisms underlying the changes among glycoprofiles. We demonstrate the functionality and performance of this approach with a variety of glycomic analysis, including recombinant erythropoietin (EPO) N-glycosylation, human milk oligosaccharides (HMOs), mucin-type O-glycans, gangliosides, and site-specific compositional data. Specifically, we analyzed sixteen MALDI-TOF glycoprofiles of EPO, where each EPO glycoprofile was produced in a different glycoengineered CHO cell line(Čaval et al. 2018; Yang et al. 2015). We also analyze forty-eight HPLC glycoprofiles of HMO from six mothers(Mohammad, Hadsell, and Haymond 2012). By analyzing these glycoprofiles with GlyCompare, we quantify the abundance of important substructures, cluster the glycoprofiles of mutant cell lines, connect genotypes to unexpected changes in glycoprofiles, and associate a phenotype of interest with substructure abundance and flux. We further demonstrate that such analyses gain statistical power. Finally, we expand our studies to include a tumor-normal comparison of mucin-type O-glycans, human retinal glycolipids, and site-specific N-glycan compositional data from the mouse brain. The analyses of the various N and O-type glycan datasets demonstrate that our framework presents a convenient and automated approach to elucidate insights into complex patterns in glycobiology.

## Results

### Strikingly different glycoprofiles from small genetic changes can be compared with GlyCompare

Due to the sparsity and interdependence of glycans in each glycoprofile, comparing different glycoprofiles can be challenging(Ashwood et al. 2019; Doherty et al. 2018). We demonstrated the core idea with three diverse erythropoietin (EPO) profiles made by three glycoengineered CHO cell lines(Čaval et al. 2018; Yang et al. 2015). EPO produced in the wild type (WT) and two double glycosyltransferase knockout (Mgat4a/4b and St3gal3/6) CHO cell lines have very different glycoprofiles that do not share many detected glycans (Figure 1.1a). Efforts to identify primary and off-target effects of genetic modifications have limited success if relying only on overlapping glycans or on the presence/absence of a set of glycoforms. This would drastically limit their analytic power due to the sparsity of comparable consensus glycans (Figure 1.1a). The problem is that even glycans differing in only one single monosaccharide will be treated as two completely different glycans under conventional glycoprofile analysis methods(Reiding et al. 2019). In the end, the glycan abundance cannot be compared directly. This limited overlap between samples gets worse in analyzing large glycomics datasets. These challenges prompted us to develop GlyCompare, a substructure-based approach to glycan analysis. Glycoprofiles are first decomposed into a substructure network that encodes the shared biosynthetic pathways as well as the interdependence among glycans. Then, the substructure abundances are aggregated from all glycans to account for activities at each enzymatic step (Figure 1.1b). In essence, this shifts the focus of glycoprofile analysis from examining the increase/decrease of independent glycans to examining the increase/decrease of a series of glycan substructures (Figure 1.1c). This provides insightful information on a similar synthetic process and allows us to mitigate major statistical challenges of working with glycan-based glycoprofiles.

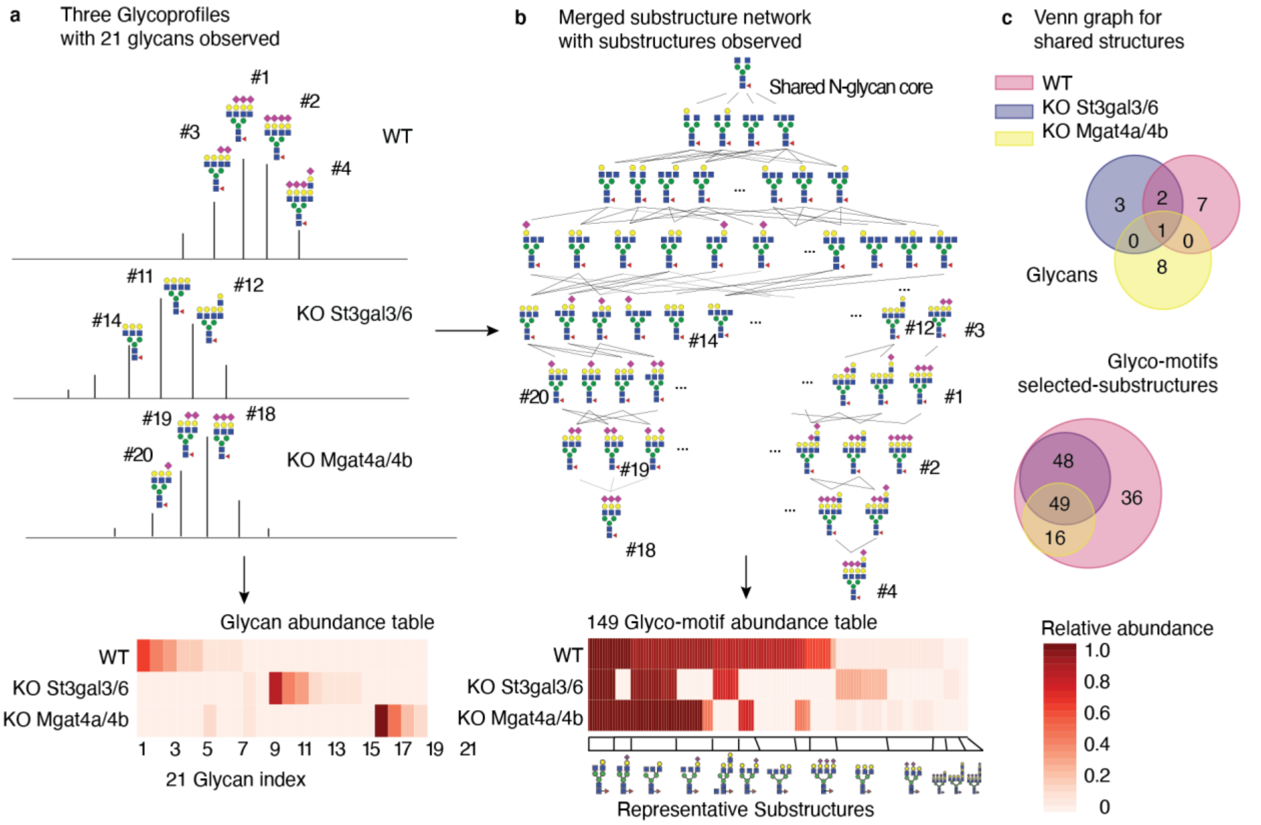


**Figure 1.1 The GlyCompare platform improves glycomics data analysis and interpretation by using glycan biosynthetic network data to account for glycan interdependence.** **a** Three example glycoprofiles (WT, Mgat4a/4b knockout, and St3gal4/6 knockout profiles), with annotated glycans and measured relative glycan abundances, show low overlap despite differing in only a few gene knockouts. **b** The low overlap can be rescued by propagating glycan substructures through the glycan biosynthetic network. Then, the glycoprofile is transformed into glyco-motif vectors. The representative substructure is generated to represent core glycan substructures of glycoprofiles (see Methods). **c** Venn diagrams show the imperfect overlap of glycans across samples (upper), which is rescued when using GlyCompare to analyze glyco-motif substructures (bottom). Source data are provided as a Source Data file.

### GlyCompare decomposes glycoprofiles to facilitate comparison

Glycoprofiles can be decomposed into abundances of intermediate substructures. The resulting substructure profile has richer information than whole glycan profiles and enables more precise comparison across conditions. Since glycan biosynthesis involves long, redundant pathways, the pathways can be collapsed to obtain a subset of substructures while preserving the information of all glycans in the dataset. We call this minimal set of substructures glyco-motifs. The GlyCompare workflow consists of several steps wherein glycoprofiles are annotated and decomposed, glyco-motifs are prioritized, and each glyco-motif is quantified for subsequent comparisons (Method). The specific workflow is described as follows.

First, to characterize one glycoprofile with substructures, all substructures in one glycoprofile are identified and occurrence per glycan is quantified (Figure 1.2a-b). Within a glycoprofile, a substructure's abundance is calculated by summing the abundance of all glycans containing the substructure. This transformation results in a substructure profile, which stores abundances for all glycan substructures (Figure 1.2b) in the given glycoprofile. The summation over similar structures asserts that they follow the same synthetic paths, which is appropriate for glycosylation wherein synthesis is hierarchical and acyclic(Spahn and Lewis 2014). Therefore, a substructure abundance is not simply a sum over similar structures but mirrors the activity of the enzymes through biosynthetic pathways.

Second, to identify the most informative substructures (i.e., glyco-motifs), substructures are prioritized using the substructure network. The substructure network is built by connecting all substructures with biosynthetic steps (Figure 1.2d and 1.3c). The network starts from a core structure. An additional network level represents one biosynthetic step, adding one of the monosaccharides to the previous level. The edges in the network represent enzymatic additions of each monosaccharide, which can be annotated with known reactions (Figure 1.4). Redundant substructures are identified when parent-child substructure abundances are the same (Figure 1.2d). Substructure network reduction proceeds by collapsing links with redundant substructures (connected with a solid arrow in Figure 1.2d) and only retaining the child substructure. The remaining substructures are called glyco-motifs (selected-substructures); they describe the variance enirely at the substructure level. The abundances of all glyco-motifs are then represented as a glyco-motif profile, the minimal subset of meaningful substructure abundances representing glycoprofiles (Figure 1.2e).

For larger datasets, it is necessary to summarize the structure difference and abundance changes by clustering glyco-motifs (Figure 1.5). After clustering glyco-motifs, the common structural features of a cluster are calculated using the average weight of each monosaccharide (Figure 1.2f, see Method). Monosaccharides with a weight larger than 51% are preserved, which illustrates the predominant structure in the cluster. This allows one to quickly evaluate the distinguishing structure features that vary across samples in any given dataset.

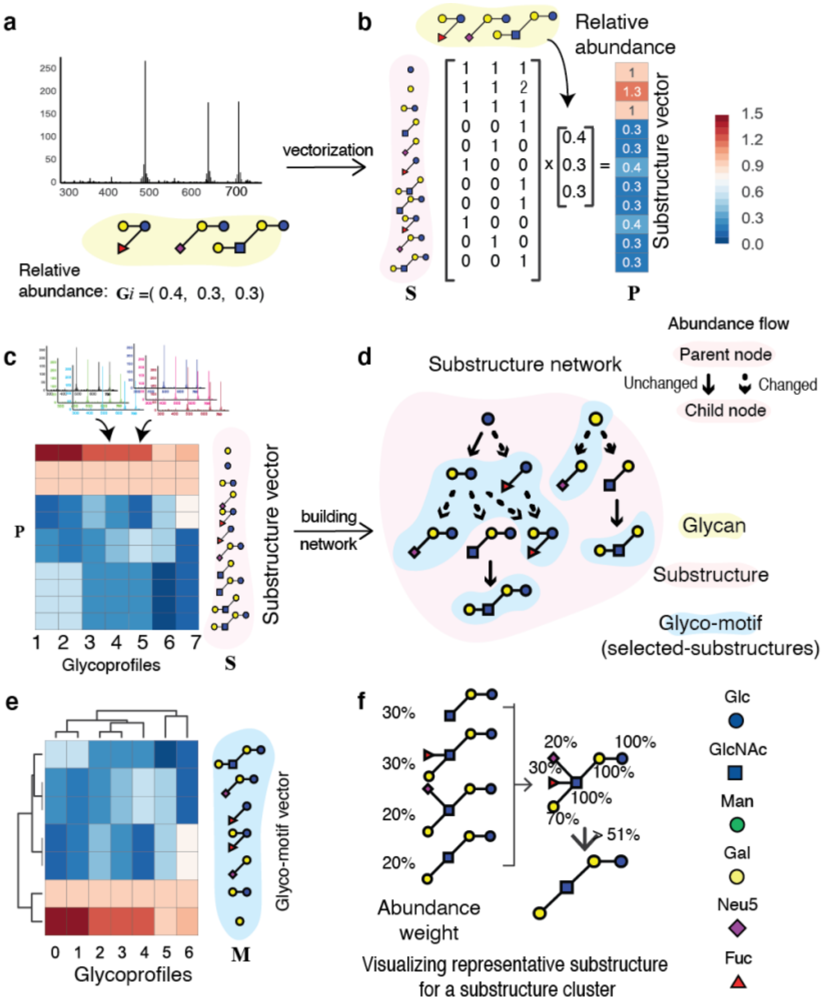
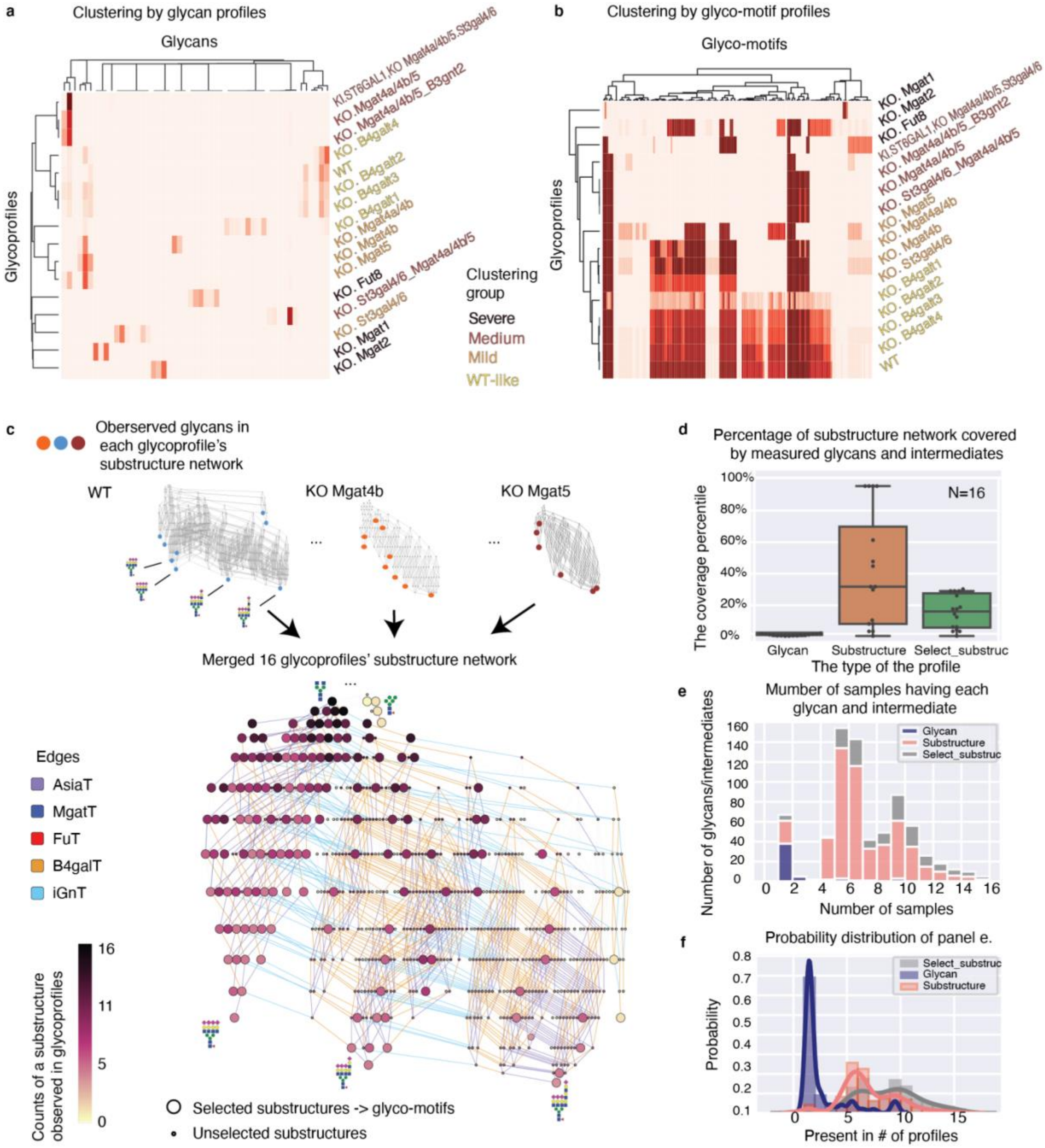


**Figure 1.2 The core methodology for transforming glycoprofiles to glyco-motif profiles and visualizing cluster-representative substructures using GlyCompare. a** and **b**, A glycoprofile with structure and relative abundance annotation **G** is obtained. The glycans are decomposed to a substructure set **S**, and the presence/absence of each substructure is recorded. Presence/absence vectors are weighted by the glycan abundance, and are summed into a substructure vector **P. c** Seven example glycoprofiles are transformed to substructure vectors as **a** and **b**. **d** A substructure network is constructed to identify the non-redundant glyco-motifs that change in abundance from their precursor substructures. **e** The glycoprofiles can then be compared by their glyco-motif vectors **M** to generate more meaningful clusters. Both glycoprofiles and substructures can be clustered for similarity analysis. **f** Core structure information can be visualized from a substructure cluster. For example, four substructures with different weights were aligned together, and the monosaccharides with a weight over 51% were preserved. Throughout the manuscript, glycan is referred to complete and secreted monosaccharide polymer; a glycan substructure is referred to a complete or incomplete monosaccharide polymer observable within at least one secreted glycan; a glycan motif (glyco-motif) is referred to an enriched functional glycan substructure for a dataset or biological process. Note that both glycan epitopes (typically terminal glycan substructures recognized by lectins) and glycan cores (biosynthetic glycan substructures common to select types (e.g. N- or O-glycosylation) or modes (e.g. complex or high-mannose) of biosynthesis) are glyco-motifs as they are biologically functional, interpretable and will be enriched in datasets selecting for specific glycan presentation of biosynthesis. Glycompare core methods are explained at length in the Methods section.

**Figure 1.3 Substructure-based profile comparison solves the glycan non-independence and sparsity challenges, enabling the use of hierarchical clustering on large glycomics datasets. a** Clustering of unprocessed glycoprofiles. Sixteen glycoprofiles from glycoengineered recombinant EPO were clustered based solely on their raw glycan abundances. **b** Clustering of glyco-motif profiles. The glyco-motif profiles, constructed using GlyCompare, were clustered based on the 151 glyco-motifs (see Methods). There are four different phenotypic glycoprofiles (based on the glycoengineered glycosylation changes relative to wild type): WT-like (yellow), Mild (orange), Medium (red), and Severe (brown). The clusters of glycoprofiles and glycan substructures are defined by distance threshold=0.5. In both cases, clustering was hierarchical clustering with a complete linkage and correlation-distance using seaborn 0.9.0. **c** The pan-network (516 intermediate substructures) that describes the synthesis of all glycans measured on the 16 glycoengineered recombinant EPO N-glycoprofiles. The glyco-motifs (in larger size) are the minimal set of 151 substructures selected by GlyCompare for further multi-glycoprofile comparison. The edges are colored by the enzyme family, AsiaT (purple), MgatT (blue), Fut (red), B4galT(orange), iGnt(cyan) and the node color according to frequency of existence in 16 glycoprofiles. **d** The coverage of the entire glycan synthetic pathway for 16 glycoprofiles using different structure types: glycan (purple, n=16, Min=0.00589, Q1=0.00786, Median=0.0128, Q3=0.0177, Max=0.0236), substructure (gray, n=16, Min=0.005894, Q1=0.082515, Median=0.318271, Q3=0.698428, Max=0.954813), and the selected-substructure (orange, n=16, Min=0.005894, Q1=0.058448, Median=0.161100, Q3=0.276031, Max=0.290766). **e** Proportion of samples containing a glycan, substructure, or glyco-motif in the 16 samples, and **f** The associated probability distribution. Source data are provided as a Source Data file.


a
Clustering by glycan profiles
Glycans
Glycoprofiles
KI.ST6GAL1,KO Mgat4a/4b/5.St3gal4/6
KO.Mgat4a/4b/5
KO. Mgat4a/4b/5_B3gnt2
KO. B4galt4
WT
KO. B4galt2
KO. B4galt3
KO. B4galt1
KO. Mgat4a/4b
KO. Mgat4b
KO. Mgat5
KO. Fut8
KO. St3gal4/6_Mgat4a/4b/5
KO. St3gal4/6
KO. Mgat1
KO. Mgat2
Clustering group
Severe
Medium
Mild
WT-like
b
Clustering by glyco-motif profiles
Glyco-motifs
Glycoprofiles
KO. Mgat1
KO. Mgat2
KO. Fut8
KI.ST6GAL1,KO Mgat4a/4b/5.St3gal4/6
KO. Mgat4a/4b/5_B3gnt2
KO.Mgat4a/4b/5
KO. St3gal4/6_Mgat4a/4b/5
KO. Mgat5
KO. Mgat4a/4b
KO. Mgat4b
KO. St3gal4/6
KO. B4galt1
KO. B4galt2
KO. B4galt3
KO. B4galt4
WT
c
Oberserved glycans in each glycoprofile's substructure network
WT
KO Mgat4b
KO Mgat5
...
...
Merged 16 glycoprofiles' substructure network
...
Edges
AsiaT
MgatT
FuT
B4galT
iGnT
Counts of a substructure observed in glycoprofiles
16
11
5
0
Selected substructures -> glyco-motifs
Unselected substructures
d
Percentage of substructure network covered by measured glycans and intermediates
The coverage percentile
100%
80%
60%
40%
20%
0%
N=16
Glycan
Substructure
Select_substruc
The type of the profile
e
Mumber of samples having each glycan and intermediate
Number of glycans/intermediates
160
140
120
100
80
60
40
20
0
Glycan
Substructure
Select_substruc
0
2
4
6
8
10
12
14
16
Number of samples
f
Probability distribution of panel e.
Probability
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Select_substruc
Glycan
Substructure
0
5
10
15
Present in # of profiles

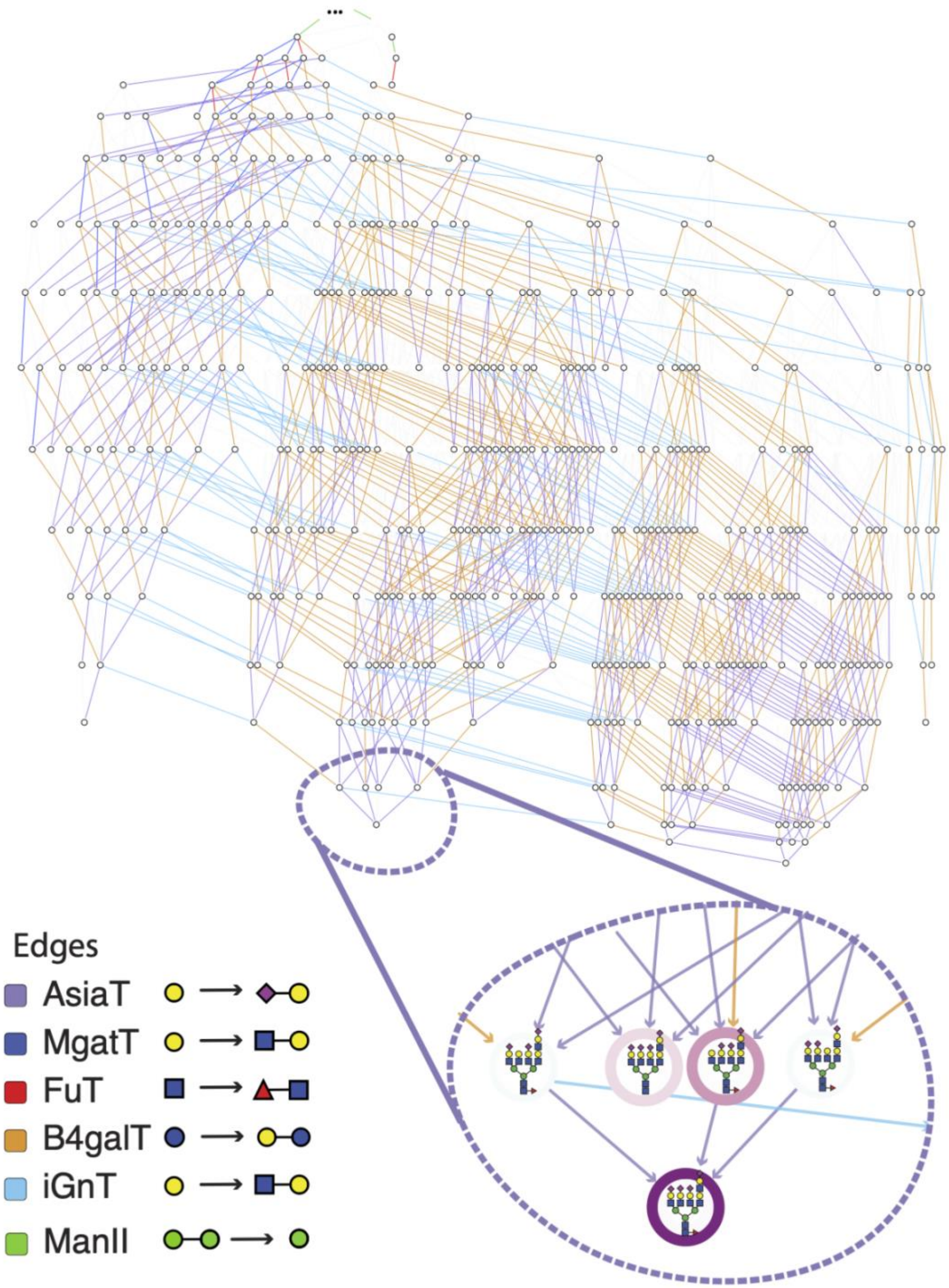


**Figure 1.4 The substructure network of EPO dataset.** The merged substructure network from 16 glycoprofiles contains 516 synthesizable glyco-substructures. The edges are colored with enzyme family, AsiaT (purple), MgatT (blue), Fut (red), B4galT (orange), iGnt (cyan), ManII (green) and the node color is according to the existence times in 16 glycoprofiles.

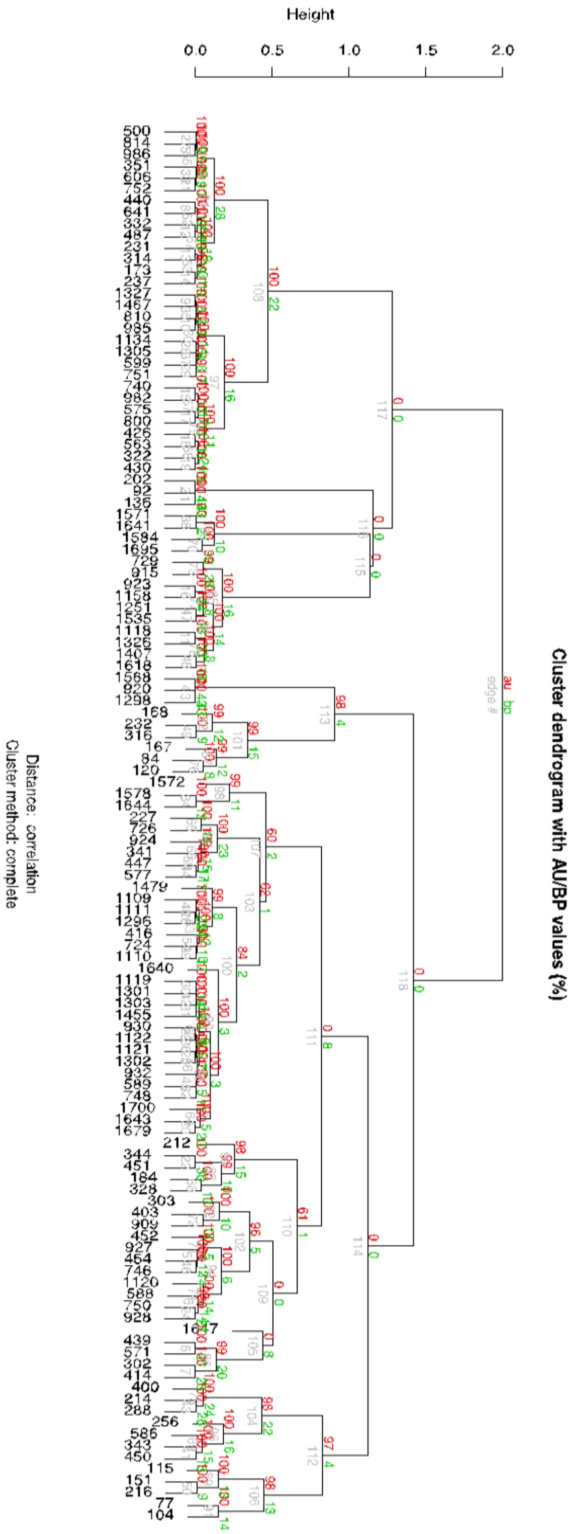


**Figure 1.5 Robustness of glyco-motifs clusters.** This is the cluster of glyco-motif vectors for EPO data. The robustness gives the criteria of how many substructure clusters should be generated[1]. The clusters are distinguished if AU (red)=100 (approximately unbiased probability value p<0.01) and then BP (green) (Bootstrap Probability) >15. The big block is further breakdown. We get 35 clusters in our EPO data.

The workflow described here will connect all glycoprofiles in a data set through their shared intermediate substructures, thus allowing robust analysis of the differences across glycomics samples and the evaluation of the associated genetic bases.

### GlyCompare accurately clusters glycoengineered EPO samples

We first apply GlyCompare on the dataset consisting of sixteen glycoprofiles coming from a panel of different Erythropoietin (EPO) glycoforms, each produced in different glycoengineered CHO cell lines. Clustering glycoprofiles did not adequately recapitulate the severity of glycosylation disruption, wherein many neighboring samples were not the most genetically similar mutants (Figure 1.3a and Figure 1.6). This inconsistency and poor clustering stem from the inherent sparseness of glycoprofiles, i.e., each glycoprofile only has a few observed glycans (Figure 1.3d), and most glycans appear only in a few glycoprofiles (Figure 1.3e-1.3f). Thus, the matrix of glycan abundances is sparse and incompatible with the glycan synthesis assumption. Since glycan composition is not utilized, the clustering is heavily affected by the categorical presence or absence of a glycan, rather than structural similarity.

GlyCompare addresses these problems by exposing hidden similarities between glycans after decomposing glycoprofiles to their composite substructures. The sixteen glycoprofiles with 52 glycans in total were decomposed into their constituent glycan substructures, resulting in a substructure network with 613 glycan substructures (Figure 1.3b,c). Furthermore, the known enzymatic rules are annotated to the edges and the network is collapsed to include 151 glyco-motifs (Figure 1.3c). By encoding the structure information, the glyco-motifs provide richer information than using glycans solely (Figure 1.3d-f). The glyco-motif clustering clearly distinguished the samples based on the structural patterns and separated profiles into groups more consistently than the raw glycan-based clusters (Figure 1.3b and Figure 1.6-1.8).

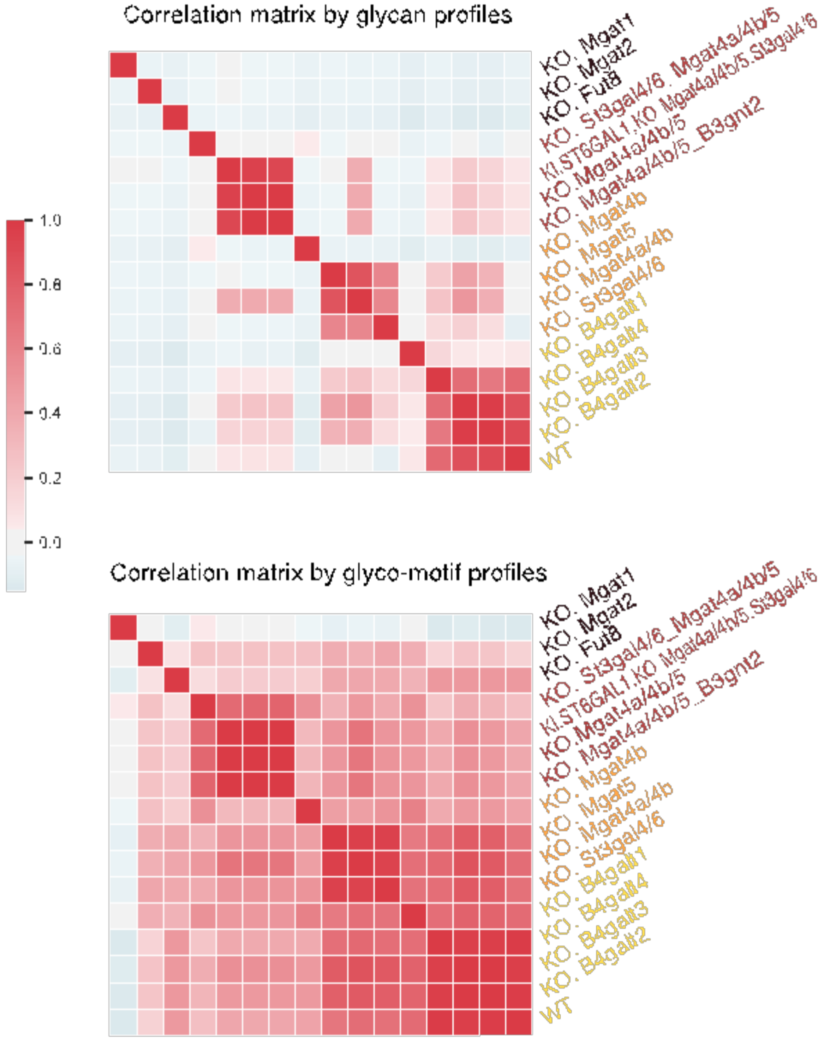
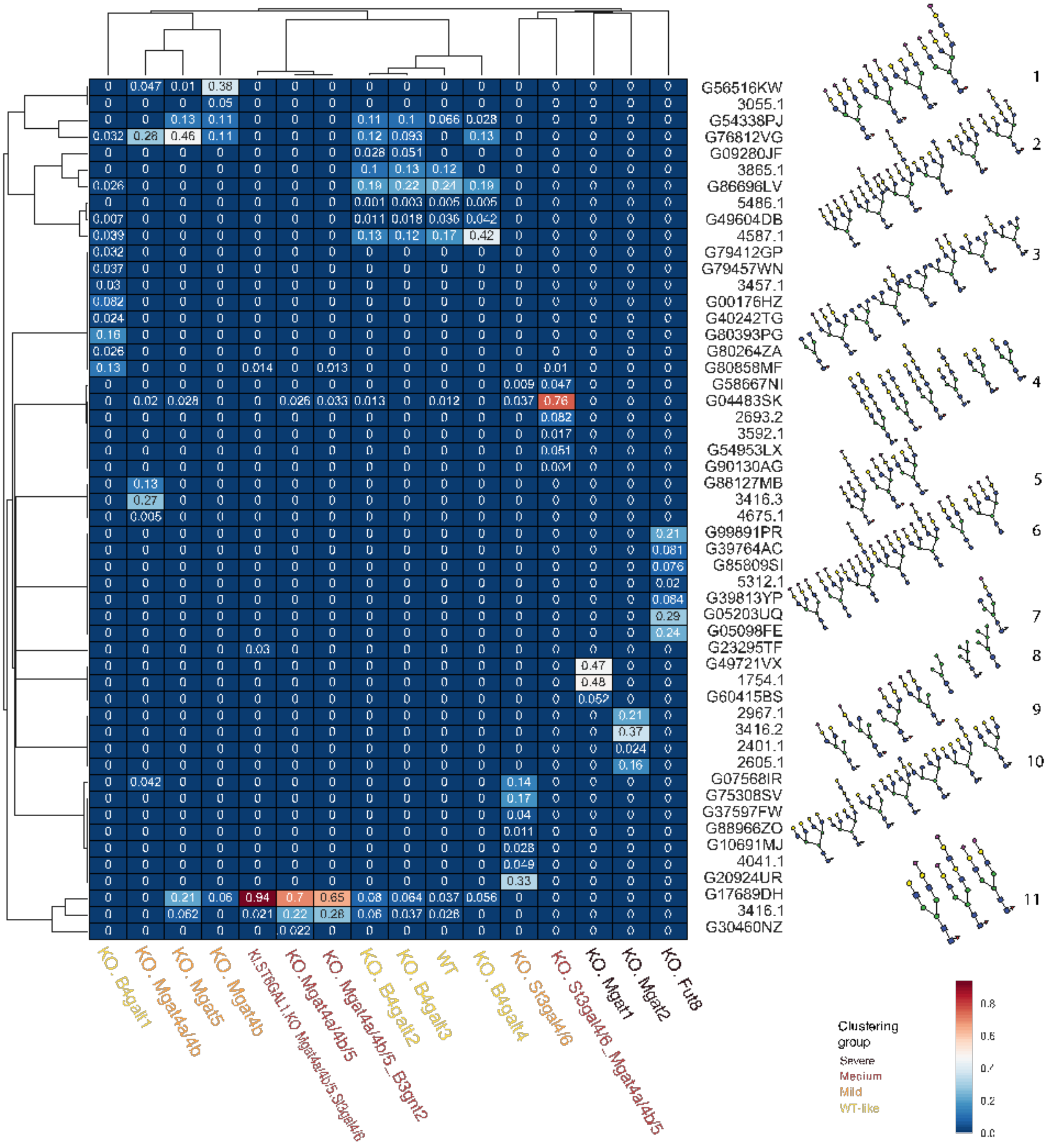
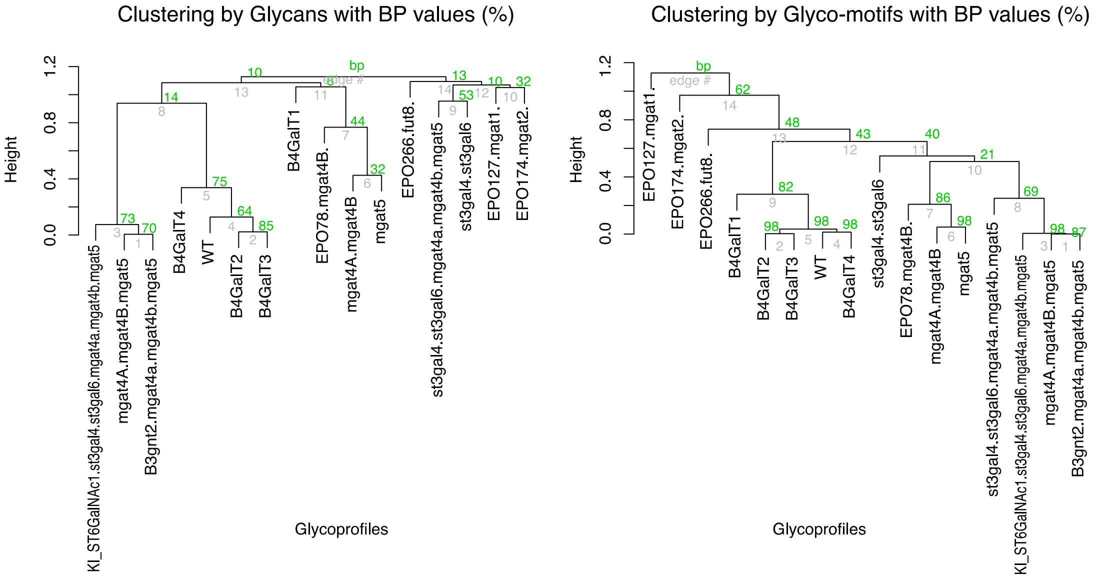


**Figure 1.6 The glycoprofile clustering table with the original glycans.** This is the clustering of sixteen glycoprofiles based on glycans. Since most of the glycans only exist in a few glycoprofiles, the clustering mainly focuses on the presence/absence of the glycans (clusters 3-10), which means the information of structural similarity tends to be ignored in the clustering. This would drastically limit their analytic power due to the sparsity of comparable consensus glycans. The correlation distance tables are also provided for both glycan profiles and glyco-motif profiles. Source data are provided as a Source Data file.

**Figure 1.6 The glycoprofile clustering table with the original glycans (continued).**



| | KO. B4galt1 | KO. Mgat4a/4b | KO. Mgat5 | KO. Mgat4b | KI.ST6GAL1.KO Mgat4a/4b/5.St3gal4/6 | KO.Mgat4a/4b/5 | KO. Mgat4a/4b/5_B3gnt2 | KO. B4galt2 | KO. B4galt3 | WT | KO. B4galt4 | KO. St3gal4/6 | KO. St3gal4/6_Mgat4a/4b/5 | KO. Mgat1 | KO. Mgat2 | KO. Fut8 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| G56516KW | 0 | 0.047 | 0.01 | 0.38 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3055.1 | 0 | 0 | 0 | 0.05 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G54338PJ | 0 | 0 | 0.13 | 0.11 | 0 | 0 | 0 | 0.11 | 0.1 | 0.066 | 0.028 | 0 | 0 | 0 | 0 | 0 |
| G76812VG | 0.032 | 0.28 | 0.46 | 0.11 | 0 | 0 | 0 | 0.12 | 0.093 | 0 | 0.13 | 0 | 0 | 0 | 0 | 0 |
| G09280JF | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.028 | 0.051 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3865.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.1 | 0.13 | 0.12 | 0 | 0 | 0 | 0 | 0 | 0 |
| G86696LV | 0.026 | 0 | 0 | 0 | 0 | 0 | 0 | 0.19 | 0.22 | 0.24 | 0.19 | 0 | 0 | 0 | 0 | 0 |
| 5486.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.001 | 0.003 | 0.005 | 0.005 | 0 | 0 | 0 | 0 | 0 |
| G49604DB | 0.007 | 0 | 0 | 0 | 0 | 0 | 0 | 0.011 | 0.018 | 0.036 | 0.042 | 0 | 0 | 0 | 0 | 0 |
| 4587.1 | 0.039 | 0 | 0 | 0 | 0 | 0 | 0 | 0.13 | 0.12 | 0.17 | 0.42 | 0 | 0 | 0 | 0 | 0 |
| G79412GP | 0.032 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G79457WN | 0.037 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3457.1 | 0.03 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G00176HZ | 0.082 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G40242TG | 0.024 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G80393PG | 0.16 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G80264ZA | 0.026 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G80858MF | 0.13 | 0 | 0 | 0 | 0.014 | 0 | 0.013 | 0 | 0 | 0 | 0 | 0 | 0.01 | 0 | 0 | 0 |
| G58667NI | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.009 | 0.047 | 0 | 0 | 0 |
| G04483SK | 0 | 0.02 | 0.028 | 0 | 0 | 0.026 | 0.033 | 0.013 | 0 | 0.012 | 0 | 0.037 | 0.76 | 0 | 0 | 0 |
| 2693.2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.082 | 0 | 0 | 0 |
| 3592.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.017 | 0 | 0 | 0 |
| G54953LX | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.051 | 0 | 0 | 0 |
| G90130AG | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.004 | 0 | 0 | 0 |
| G88127MB | 0 | 0.13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3416.3 | 0 | 0.27 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4675.1 | 0 | 0.005 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G99891PR | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.21 |
| G39764AC | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.081 |
| G85809SI | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.076 |
| 5312.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.02 |
| G39813YP | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.084 |
| G05203UQ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.29 |
| G05098FE | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.24 |
| G23295TF | 0 | 0 | 0 | 0 | 0.03 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G49721VX | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.47 | 0 | 0 |
| 1754.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.48 | 0 | 0 |
| G60415BS | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.052 | 0 | 0 |
| 2967.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.21 | 0 |
| 3416.2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.37 | 0 |
| 2401.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.024 | 0 |
| 2605.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.16 | 0 |
| G07568IR | 0 | 0.042 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.14 | 0 | 0 | 0 | 0 |
| G75308SV | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.17 | 0 | 0 | 0 | 0 |
| G37597FW | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.04 | 0 | 0 | 0 | 0 |
| G88966ZO | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.011 | 0 | 0 | 0 | 0 |
| G10691MJ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.028 | 0 | 0 | 0 | 0 |
| 4041.1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.049 | 0 | 0 | 0 | 0 |
| G20924UR | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.33 | 0 | 0 | 0 | 0 |
| G17689DH | 0 | 0 | 0.21 | 0.06 | 0.94 | 0.7 | 0.65 | 0.08 | 0.064 | 0.037 | 0.056 | 0 | 0 | 0 | 0 | 0 |
| 3416.1 | 0 | 0 | 0.062 | 0 | 0.021 | 0.22 | 0.28 | 0.06 | 0.037 | 0.026 | 0 | 0 | 0 | 0 | 0 | 0 |
| G30460NZ | 0 | 0 | 0 | 0 | 0 | 0.022 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |



Clustering by Glycans with BP values (%)

Clustering by Glyco-motifs with BP values (%)

**Figure 1.7 The clustering robustness.** The robustness is measured with BP (Bootstrap Probability). BP is a measure of the cluster robustness suggesting a significant similarity within clusters and thereby mitigating some challenges of clustering reproducibility. Glyco-motifs abundances showed higher BP than clustering with whole-glycan abundance profiles. In the whole-glycan profile clusters, wild-type (WT) glycoprofiles are closer to the double-knockouts with highly-perturbed glycoprofiles. Double-knockouts are predominantly determined to be strongly perturbed and therefore should not cluster with wild-types in a biologically meaningful clustering. As such, we believe the glycan clustering (which clusters WT with double knockouts) is less interpretable than the glyco-motif clustering which does not include the WT/double-knockout grouping.

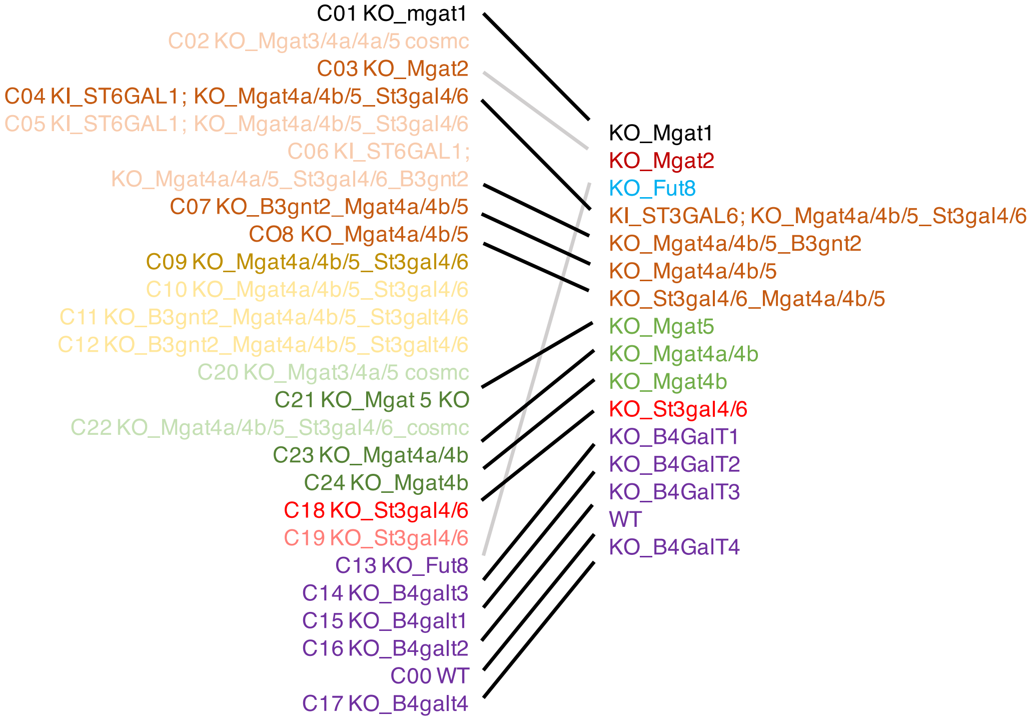


**Figure 1.8 Profile matching between the data from Čaval et al. 2018[2] and GlyCompare.** The lightened names are the knockouts that do not have MALDI-TOF data published[2]. The KO name with the same color (for example, in brown, yellow, green, red, purple) are the KO profiles that clustered together. While some glyco-motif clusters can be seen in the glycoprofile clusters, there are important differences, and the glyco-motif clusters provide more information and improved cluster stability. Furthermore, the clustering result based on the glyco-motif was consistent with the clustering based on the native mass spectrometry, except for the Mgat2 knockout and the Fut8 knockout, which considerably changed the glycoprofiles by removing many common glycans. The main reason is that GlyCompare accounts for structural differences caused by each glycosyltransferase. This allows us to evaluate the magnitude of differences between glycans, whether it be between glycans with the same mass but different structural topologies, or subtle structural variations due to single changes in monosaccharides. Therefore, we had a better interpretation of the glycan structure variants across multiple glycoprofiles. All these results demonstrated the excellent performance of our GlyCompare in assessing the structural similarity between different glycoprofiles.

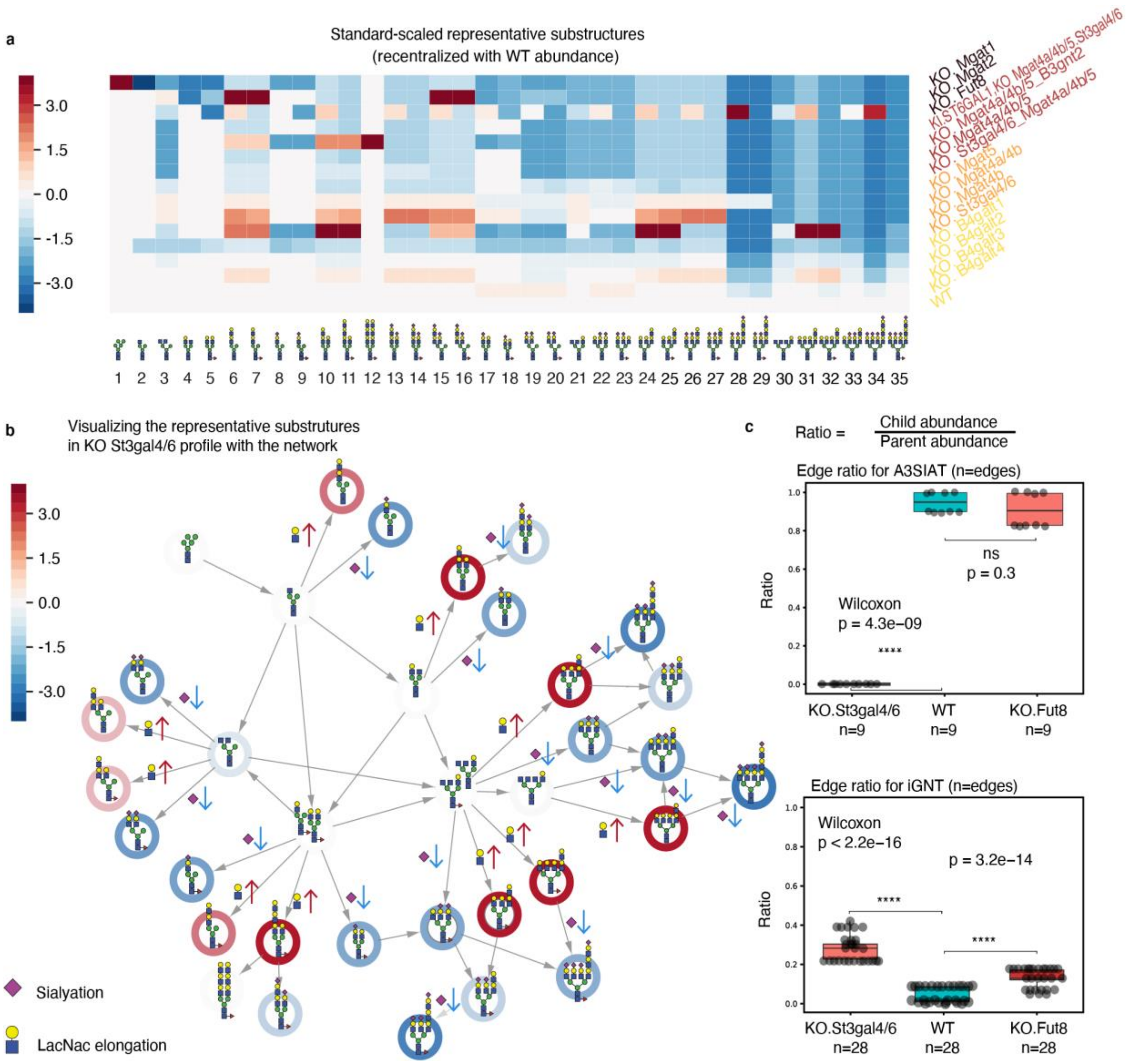
The sixteen glycoprofiles clustered into three groups with a few severely modified outliers (Figure 1.3b). The 151 glyco-motifs were clustered into thirty-five groups, each summarized by representative substructures Rep1 – Rep35 (Figure 1.9a and Figure 1.4). The clusters of glycoprofiles are consistent with the genetic similarities among the host cells. Specifically, the major substructure patterns cluster individual samples into four categories: ‘wild-type (WT)-like’, ‘mild’, ‘medium’ and ‘severe’. The WT-like category contains one group, WT and B4galt1/2/3/4 knockouts, which has most of the substructures seen in WT cells. The mild group includes the Mgat4b/4a, Mgat4b, and Mgat5 knockouts, where each loses the tetra-antennary structure, and a St3gal4/6 knockout, which loses the terminal sialylation. The medium category is a group that contains knockouts of St3gal4/6 and Mgat4a/4b/5, knockouts of Mgat4a/4b/5 and B3gnt2, knockouts of Mgat4a/4a/5 with a knock-in of human ST6GAL1, and knockouts of Mgat4a/4b/5 and St3gal4/6. The medium disruption category lost the tri-antennary structure. The ‘severe’ category includes three individual glycoprofiles with knockouts for Fut8, Mgat2, and Mgat1, each of which generates many glycans not detected in the WT-like, mild or medium categories. While some glyco-motif clusters can be seen in the glycoprofile clusters, there are important differences, and the glyco-motif clusters provide more information and improved cluster stability (Figure 1.9a, Figure 1.7, 1.8). These results demonstrate that standard methods are unfit to cluster glycan abundance from glycomics data in genetically diverse datasets; however, computing glyco-motif abundance accounts for the structural similarity of glycans between different glycoprofiles and allows one to use standard hierarchical clustering techniques reliably.

### GlyCompare summarizes structural changes across glycoprofiles

GlyCompare helps to more robustly group samples by accounting for the biosynthetic and structural similarities of glycans. Further analysis of the representative structures provides detailed insights into which structural features vary the most across samples. To accomplish this, we rescaled the representative structure abundances and identified significant changes between mutant cells and WT (Figure 1.9a, Figure 1.10). Analysis of the representative substructure network provides a more precise interpretation of the changes in the St3gal4/6 KO (Figure 1.9b) and the Fut8 KO profiles (Figure 1.11). This interpretation highlights the specific structural features of glycans that are impacted when glycoengineering recombinant EPO.

In-depth analysis showed, as expected, in the Mgat1 knockout glycoprofile, only high mannose N-glycans are seen. Also, in the Mgat2 knockout, the glycan substructure of bi-antennary on one mannose linkage significantly increases. The unique structure of bi-antennary LacNac elongated in the N-glycans emerges in the St3gal4/6 and Mgat4a/4b/5 knockouts. In the St3gal4/6 knockout profile, the abundance of structures with sialylation are zero, while the tetra-antennary and triantennary poly-LacNAc elongated N-glycan substructure without sialylation significantly increased (Rep24-25: p=$1.3 \times 10^{-3}$, Rep31-32: p=$2.3 \times 10^{-4}$) (Figure 1.9a-c). Along with expected changes in α-1,6 fucosylation in the Fut8 knockout glycoprofile, we also observed an increase in the tetra-antennary poly-LacNac elongated N-glycan without fucose, which has not been previously reported (One-sided one-sample Wilcoxon test, Rep28: p=$2.7 \times 10^{-4}$, Rep34: p=$2.0 \times 10^{-4}$) (Figure 1.9a). Both the St3gal4/6 and Fut8 knockout profiles have increased tri/tetra-antennary poly-LacNac elongated substructure (Rep24, Rep31). It is related to the increased conversion ratio of iGNT (Figure 1.9c). Finally, the Mgat4b, Mgat4a/4b and Mgat5 knockouts lose all core tetra-antennary substructures (Rep30-35: unscaled abundance=0) (Figure 1.10). While triantennary substructures with GlcNac elongation increased significantly for Mgat4b (Rep13-14, p=$2.6 \times 10^{-3}$; Rep26-27: p=$2.5 \times 10^{-4}$), the poly-LacNac elongation structure disappeared. Interestingly, while both the Mgat4b and Mgat5 knockouts do not have the tri-antennary poly-LacNac elongated N-glycan, the Mgat4a/4b mutant keeps a highly abundant poly-LacNac branch (Rep28-29: p= $2.4 \times 10^{-4}$). Thus, by using GlyCompare, we identified the specific glycan features impacted not only in individual glycoengineered cell lines but also in features shared by groups of related cell lines.

**Figure 1.9 Analysis of glycan abundance changes using representative substructures. a** The heatmap of normalized glycan abundance for the thirty-five substructure clusters from Figure 1.3b. The substructures are sorted based on the glycan structure complexity, followed by the number of branches, the degree of galactosylation, sialylation, and fucosylation. While comparing to WT, the weighted average abundance of each cluster is calculated then z-score standardized by each column. The color denotes the change of glycan abundance for the comparison of KO vs. WT of the indicated substructure. **b** The differential substructure representative network for the comparison between the St3gal4/6 knockout profile and the WT profile. The z-score rescaled substructure clusters' abundance in **a** are visualized on edges with a simplified network. The color is defined the same as in **a** for the changes of glycan abundance. The plot demonstrates the changes of the elongation and sialylation. **c** Differential enzyme activities of α-2,3-sialyltransferase (a3SiaT, reaction n=9) and β-1,3-N-acetylglucosaminyltransferase (iGNT, reaction n=28) for the knockout profiles (St3gal4/6 and Fut8) and wild type profile in terms of network edge ratio. Specifically, the network edge ratio is calculated on the reactions shared by three profiles. The 5 quartile boundaries of the a3SiaT table are KO.St3gal4/6, Min=0, Max=0; KO.Fut8, Min=0.795, Q1=0.795, Median=0.795, Q3=1, Max=1; WT, Min=0.871, Q1=0.871, Median=0.871, Q3=1, Max=1. The 5 quartile boundaries of the iGNT table are KO.St3gal4/6, Min=0.224, Q1=0.224, Median=0.285, Q3=0.314, Max=0.412; KO.Fut8, Min=0.096, Q1=0.104, Median=0.161, Q3=0.205, Max=0.205; WT, Min=0.0569, Q1=0.0637, Median=0.0709, Q3=0.129, Max=0.129. The one-sided Wilcox tests are performed. For a3SiaT table, KO.St3gal4/6 vs WT has $p=4.3e-09$, KO.Fut8 vs WT has $p=0.3$. For iGNT table, KO.St3gal4/6 vs WT has $p<2.2e-16$, KO.Fut8 vs WT has $p=3.2e-14$. Source data are provided as a Source Data file. Source data are provided as a Source Data file.


a
Standard-scaled representative substructures
(recentralized with WT abundance)
3.0
1.5
0.0
-1.5
-3.0
KO.Mgat1
KO.Mgat2
KO.Fut8
KI.ST6GAL1,KO.Mgat4a/4b/5,St3gal4/6
KO.Mgat4a/4b/5_B3gnt2
KO.Mgat4a/4b/5
KO.St3gal4/6_Mgat4a/4b/5
KO.Mgat5
KO.Mgat4a/4b
KO.Mgat4b
KO.St3gal4/6
KO.B4galt1
KO.B4galt2
KO.B4galt3
KO.B4galt4
WT
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35
b
Visualizing the representative substrutures
in KO St3gal4/6 profile with the network
3.0
1.5
0.0
-1.5
-3.0
Sialyation
LacNac elongation
c
Ratio = Child abundance / Parent abundance
Edge ratio for A3SIAT (n=edges)
Ratio
1.0
0.8
0.6
0.4
0.2
0.0
Wilcoxon
p = 4.3e−09
****
ns
p = 0.3
KO.St3gal4/6
n=9
WT
n=9
KO.Fut8
n=9
Edge ratio for iGNT (n=edges)
Ratio
1.0
0.8
0.6
0.4
0.2
0.0
Wilcoxon
p < 2.2e−16
p = 3.2e−14
****
****
KO.St3gal4/6
n=28
WT
n=28
KO.Fut8
n=28

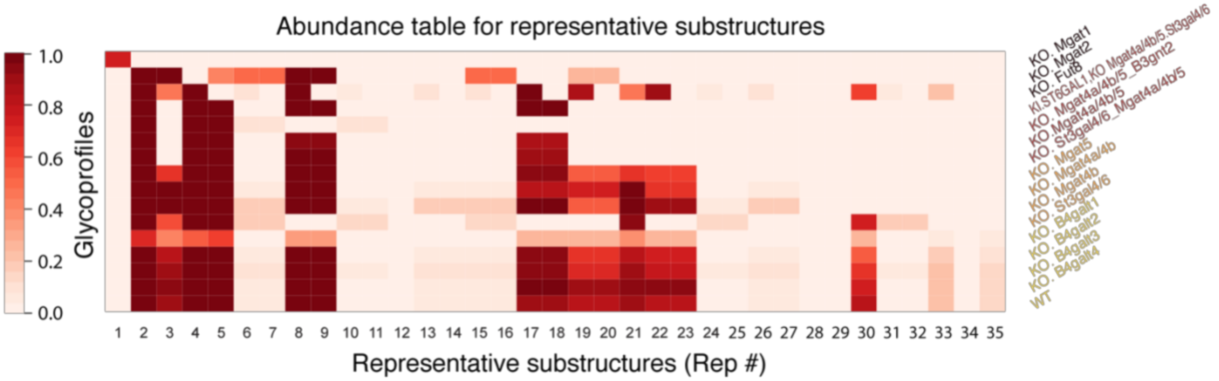


**Figure 1.10 The unscaled cluster abundance related to Figure 1.3a.** The heatmap of glycan abundance for the thirty-five substructure clusters. The substructures are sorted based on the glycan structure complexity, followed by the number of branches, the degree of galactosylation, sialylation, and fucosylation. While comparing to WT, the weighted average abundance of each cluster is calculated by each column. The color denotes the change of glycan abundance for the comparison of KO vs. WT of the indicated substructure. Source data are provided as a Source Data file.

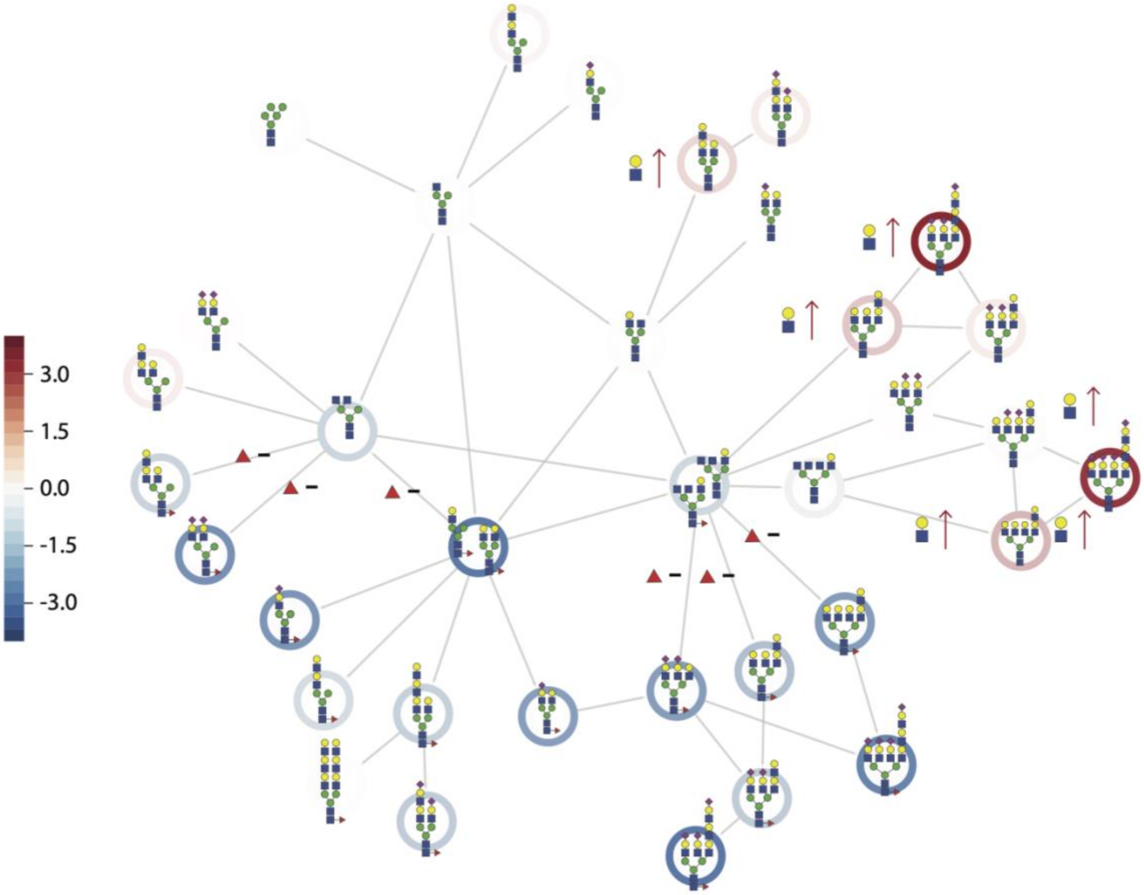


**Figure 1.11 Representative substructure network for KO.Fut8.** The differential substructure representative network for the comparison between the Fut8 knockout profile and the WT profile. The z-score rescaled substructure clusters' abundance in Figure 1.9a are visualized with a simplified network. The color is defined the same as in Figure 1.9a for the fold change of glycan abundance.

### GlyCompare reveals off-target changes in substructures invisible at the whole-glycan level

Many secreted and measured glycans are also precursors, or substructures, of larger glycans (Figure 1.12a). Thus, the secreted and observed abundance of one glycan may not equal the total amount synthesized. GlyCompare quantifies the total abundance of a glycan by combining the glycan abundance with the abundance of its products. To demonstrate this capability of GlyCompare, we analyzed HMO abundance, to test if maternal genetics underlying the secretor status has unexpected off-target effects on other HMO features. We obtained forty-seven HMO glycoprofiles from 6 mothers (1, 2, 3, 4, 7, 14, 28 and 42 days postpartum (DPP)), 4 "secretor" mothers with functioning FUT2 (α-1,2 fucosyltransferase), and 2 "non-secretor" mothers with non-functional FUT2. With GlyCompare addressing the interdependence of HMOs, we could use powerful statistical methods to study trends in HMO synthesis. Specifically, we used regression models to predict secretor status and DPP from substructure abundance.

**Figure 1.12 Analysis of intermediate substructures with GlyCompare elucidates unexpected associations in HMO abundance and reaction flux with secretor status, which are missed in the standard whole-glycan analysis. a-d** Over time (DPP), substructure X62, LSTb, DSLNT, and DSLNH show different trends for secretors and non-secretors. Furthermore, the abundance of aggregated X62 shows significant positive-correlation with secretor and negative-correlation with non-secretor. GEE models for each structure are visualized and approximated using a gaussian-link Generalized Linear Model with 95% confidence intervals; Odds Ratio (OR) significance (likelihood OR is non-zero) was measured with a two-sided Wald-test (**a** n=47, Coef=-1.37, p=4.3e-7; **b** n=47, Coef=-1.81, p=3.98e-13; **c** n=47, Coef=0.16, p=3.98e-13; **d** n=47, Coef=0.382, p=-0.23) **e** The substructure intermediates for four connected glycans are shown here. The synthesis of larger glycans must pass through intermediate substructures that are also observed glycans, where the substructures are as associated with measured glycans as follow X40=LNT, X62=LSTb, X106=DSLNT, X138=DSLNH. **f** and **g** Panels examine the product-substrate ratio for two reactions in panel **e**. X40, the LNT substructure, is a precursor to X62, the LSTb substructure, which is a precursor to X106, the DSLNT substructure. We estimate the flux of these conversions from X40 to X62 and X62 to X106 by examining the product-substrate ratio, i.e., the proportion of the total synthesized substrate converted to the product. LSTb/LNT substructure relative abundance ratios are not associated with secretor status while DSLNT/LSTb ratios are. Panels **f** and **g** show OR corresponding to the ratio association with secretor status. (**f.** n=47, OR=0.99, p=0.55; **g.** n=47, OR=0.95, p=0.018). See Table 1.1 for complete GEE statistics. Source data are provided as a Source Data file.

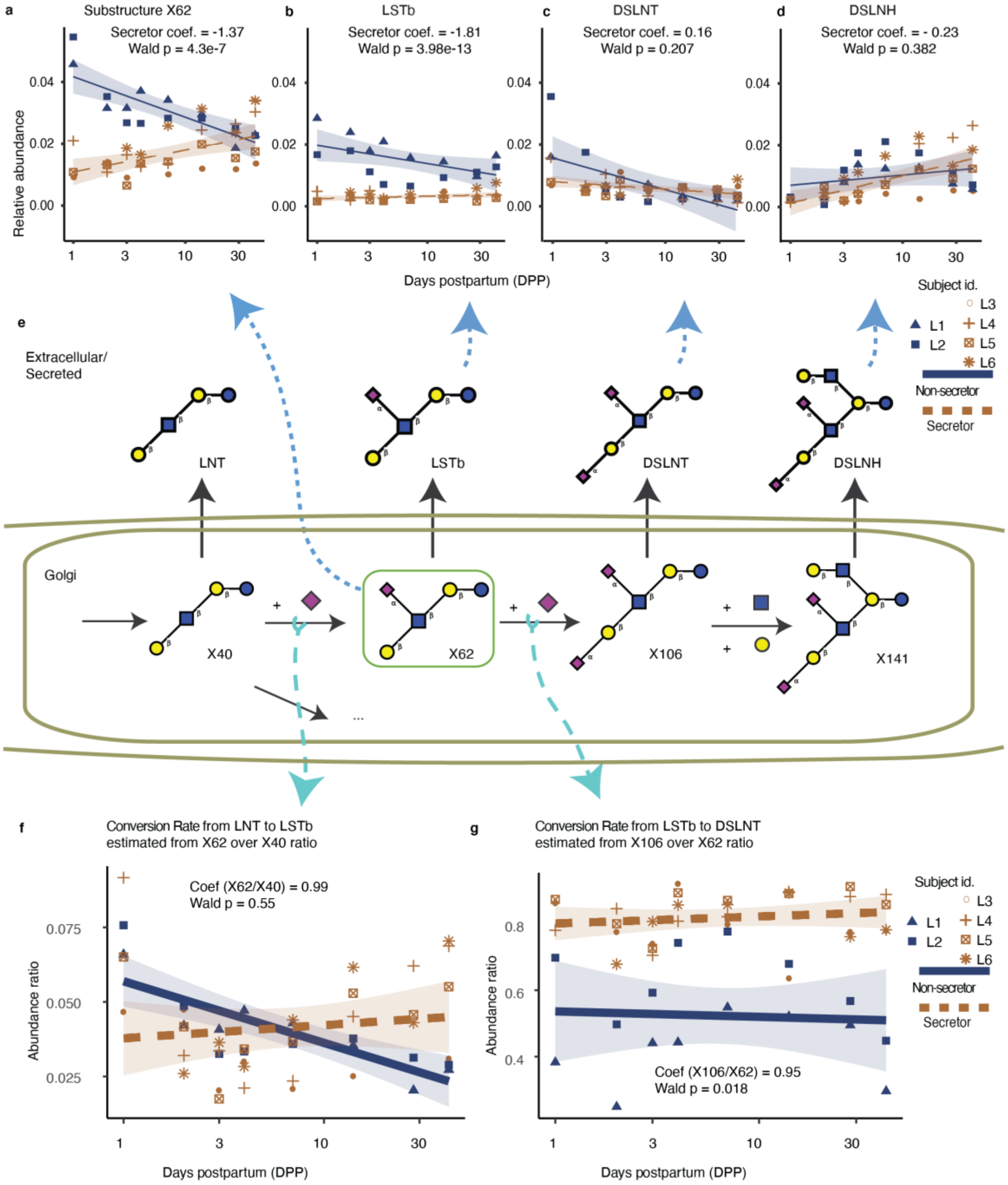
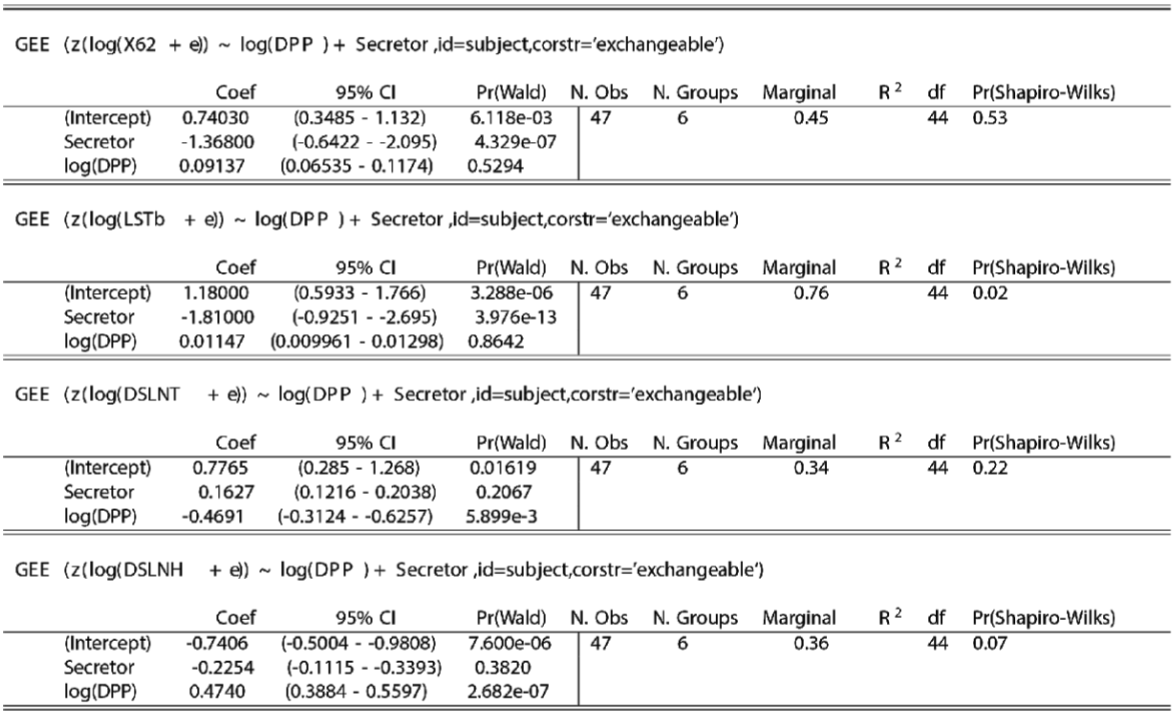

a
Substructure X62
Secretor coef. = -1.37
Wald p = 4.3e-7
b
LSTb
Secretor coef. = -1.81
Wald p = 3.98e-13
c
DSLNT
Secretor coef. = 0.16
Wald p = 0.207
d
DSLNH
Secretor coef. = - 0.23
Wald p = 0.382
Relative abundance
0.04
0.02
0.00
1
3
10
30
Days postpartum (DPP)
Subject id.
L1
L2
L3
L4
L5
L6
Non-secretor
Secretor
e
Extracellular/
Secreted
LNT
LSTb
DSLNT
DSLNH
Golgi
X40
X62
X106
X141
+
...
α
β
f
Conversion Rate from LNT to LSTb
estimated from X62 over X40 ratio
Coef (X62/X40) = 0.99
Wald p = 0.55
Abundance ratio
0.075
0.050
0.025
1
3
10
30
Days postpartum (DPP)
g
Conversion Rate from LSTb to DSLNT
estimated from X106 over X62 ratio
Coef (X106/X62) = 0.95
Wald p = 0.018
Abundance ratio
0.8
0.6
0.4
1
3
10
30
Days postpartum (DPP)
Subject id.
L1
L2
L3
L4
L5
L6
Non-secretor
Secretor

**Table 1.1 Complete information on generalized estimating equation models.**

The tables above specify the coefficient summary, confidence intervals, Wald test p-values. We report general model statistics including number of observations and groups, and degrees of freedom. We report effect size with marginal correlation for gaussian regressions and entropy for logistic regression. Finally, for gaussian regressions we report the Shapiro-Wilk's p-value for normality of a distribution. **a.** Gaussian GEE, predicting motif abundance from secretor status while controlling for DPP **b.** Gaussian GEE, predicting motif abundance from DPP split on secretor status **c.** Logistic GEE, predicting secretor status from estimated flux while controlling for DPP.

a



GEE (z(log(X62 + e)) ~ log(DPP ) + Secretor ,id=subject,corstr='exchangeable')

| | Coef | 95% CI | Pr(Wald) | N. Obs | N. Groups | Marginal | $R^2$ | df | Pr(Shapiro-Wilks) |
|---|---|---|---|---|---|---|---|---|---|
| (Intercept) | 0.74030 | (0.3485 - 1.132) | 6.118e-03 | 47 | 6 | 0.45 | | 44 | 0.53 |
| Secretor | -1.36800 | (-0.6422 - -2.095) | 4.329e-07 | | | | | | |
| log(DPP) | 0.09137 | (0.06535 - 0.1174) | 0.5294 | | | | | | |

GEE (z(log(LSTb + e)) ~ log(DPP ) + Secretor ,id=subject,corstr='exchangeable')

| | Coef | 95% CI | Pr(Wald) | N. Obs | N. Groups | Marginal | $R^2$ | df | Pr(Shapiro-Wilks) |
|---|---|---|---|---|---|---|---|---|---|
| (Intercept) | 1.18000 | (0.5933 - 1.766) | 3.288e-06 | 47 | 6 | 0.76 | | 44 | 0.02 |
| Secretor | -1.81000 | (-0.9251 - -2.695) | 3.976e-13 | | | | | | |
| log(DPP) | 0.01147 | (0.009961 - 0.01298) | 0.8642 | | | | | | |

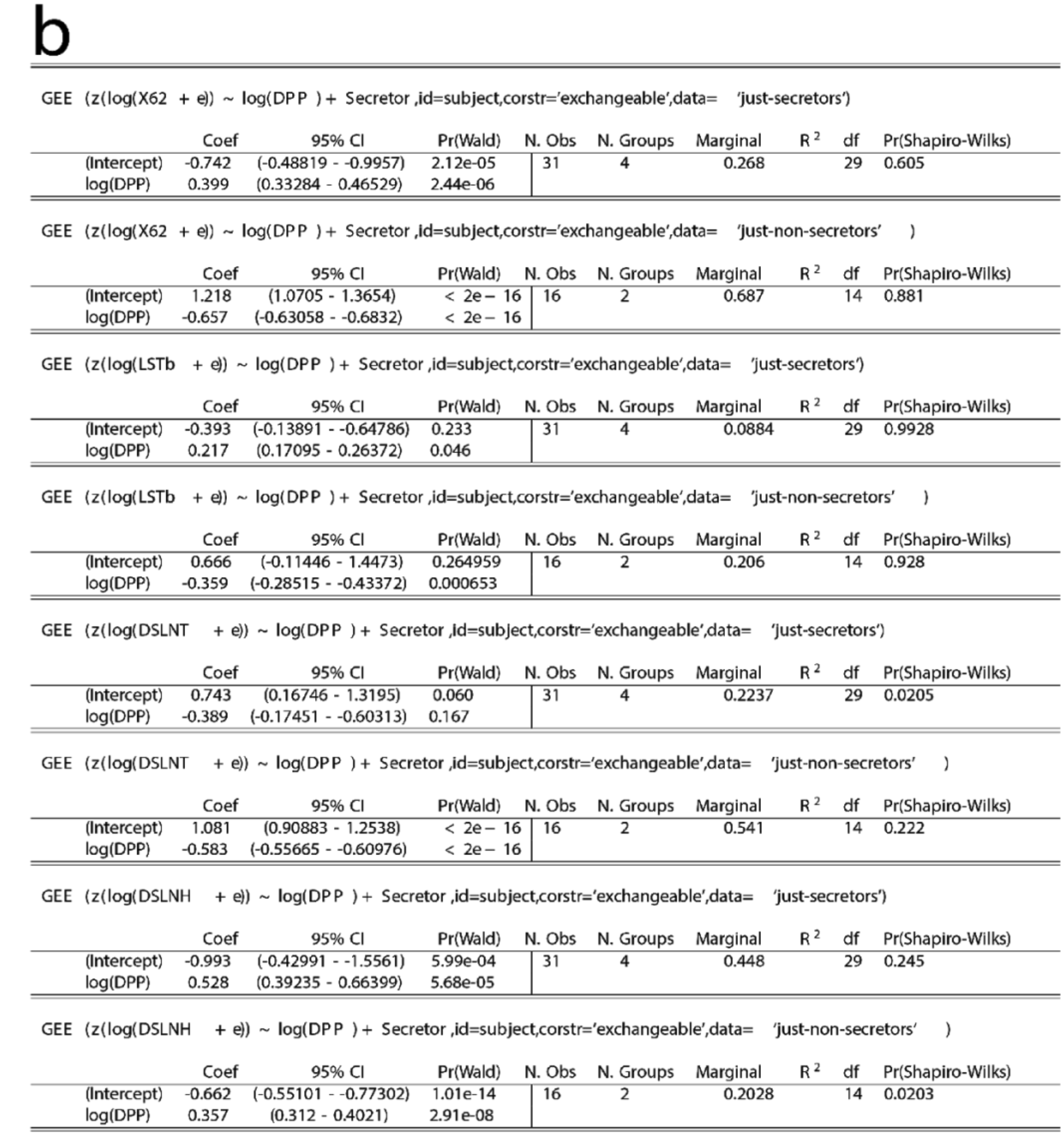
GEE (z(log(DSLNT + e)) ~ log(DPP ) + Secretor ,id=subject,corstr='exchangeable')

| | Coef | 95% CI | Pr(Wald) | N. Obs | N. Groups | Marginal | $R^2$ | df | Pr(Shapiro-Wilks) |
|---|---|---|---|---|---|---|---|---|---|
| (Intercept) | 0.7765 | (0.285 - 1.268) | 0.01619 | 47 | 6 | 0.34 | | 44 | 0.22 |
| Secretor | 0.1627 | (0.1216 - 0.2038) | 0.2067 | | | | | | |
| log(DPP) | -0.4691 | (-0.3124 - -0.6257) | 5.899e-3 | | | | | | |

GEE (z(log(DSLNH + e)) ~ log(DPP ) + Secretor ,id=subject,corstr='exchangeable')

| | Coef | 95% CI | Pr(Wald) | N. Obs | N. Groups | Marginal | $R^2$ | df | Pr(Shapiro-Wilks) |
|---|---|---|---|---|---|---|---|---|---|
| (Intercept) | -0.7406 | (-0.5004 - -0.9808) | 7.600e-06 | 47 | 6 | 0.36 | | 44 | 0.07 |
| Secretor | -0.2254 | (-0.1115 - -0.3393) | 0.3820 | | | | | | |
| log(DPP) | 0.4740 | (0.3884 - 0.5597) | 2.682e-07 | | | | | | |

**Table 1.1 Complete information on generalized estimating equation models (continued).**



GEE (z(log(X62 + e)) ~ log(DPP ) + Secretor ,id=subject,corstr=’exchangeable’,data= ‘just-secretors’)

| | Coef | 95% CI | Pr(Wald) | N. Obs | N. Groups | Marginal | $R^2$ | df | Pr(Shapiro-Wilks) |
|---|---|---|---|---|---|---|---|---|---|
| (Intercept) | -0.742 | (-0.48819 - -0.9957) | 2.12e-05 | 31 | 4 | 0.268 | | 29 | 0.605 |
| log(DPP) | 0.399 | (0.33284 - 0.46529) | 2.44e-06 | | | | | | |

GEE (z(log(X62 + e)) ~ log(DPP ) + Secretor ,id=subject,corstr=’exchangeable’,data= ‘just-non-secretors’ )

| | Coef | 95% CI | Pr(Wald) | N. Obs | N. Groups | Marginal | $R^2$ | df | Pr(Shapiro-Wilks) |
|---|---|---|---|---|---|---|---|---|---|
| (Intercept) | 1.218 | (1.0705 - 1.3654) | < 2e− 16 | 16 | 2 | 0.687 | | 14 | 0.881 |
| log(DPP) | -0.657 | (-0.63058 - -0.6832) | < 2e− 16 | | | | | | |

GEE (z(log(LSTb + e)) ~ log(DPP ) + Secretor ,id=subject,corstr=’exchangeable’,data= ‘just-secretors’)

| | Coef | 95% CI | Pr(Wald) | N. Obs | N. Groups | Marginal | $R^2$ | df | Pr(Shapiro-Wilks) |
|---|---|---|---|---|---|---|---|---|---|
| (Intercept) | -0.393 | (-0.13891 - -0.64786) | 0.233 | 31 | 4 | 0.0884 | | 29 | 0.9928 |
| log(DPP) | 0.217 | (0.17095 - 0.26372) | 0.046 | | | | | | |

GEE (z(log(LSTb + e)) ~ log(DPP ) + Secretor ,id=subject,corstr=’exchangeable’,data= ‘just-non-secretors’ )

| | Coef | 95% CI | Pr(Wald) | N. Obs | N. Groups | Marginal | $R^2$ | df | Pr(Shapiro-Wilks) |
|---|---|---|---|---|---|---|---|---|---|
| (Intercept) | 0.666 | (-0.11446 - 1.4473) | 0.264959 | 16 | 2 | 0.206 | | 14 | 0.928 |
| log(DPP) | -0.359 | (-0.28515 - -0.43372) | 0.000653 | | | | | | |

GEE (z(log(DSLNT + e)) ~ log(DPP ) + Secretor ,id=subject,corstr=’exchangeable’,data= ‘just-secretors’)

| | Coef | 95% CI | Pr(Wald) | N. Obs | N. Groups | Marginal | $R^2$ | df | Pr(Shapiro-Wilks) |
|---|---|---|---|---|---|---|---|---|---|
| (Intercept) | 0.743 | (0.16746 - 1.3195) | 0.060 | 31 | 4 | 0.2237 | | 29 | 0.0205 |
| log(DPP) | -0.389 | (-0.17451 - -0.60313) | 0.167 | | | | | | |

GEE (z(log(DSLNT + e)) ~ log(DPP ) + Secretor ,id=subject,corstr=’exchangeable’,data= ‘just-non-secretors’ )

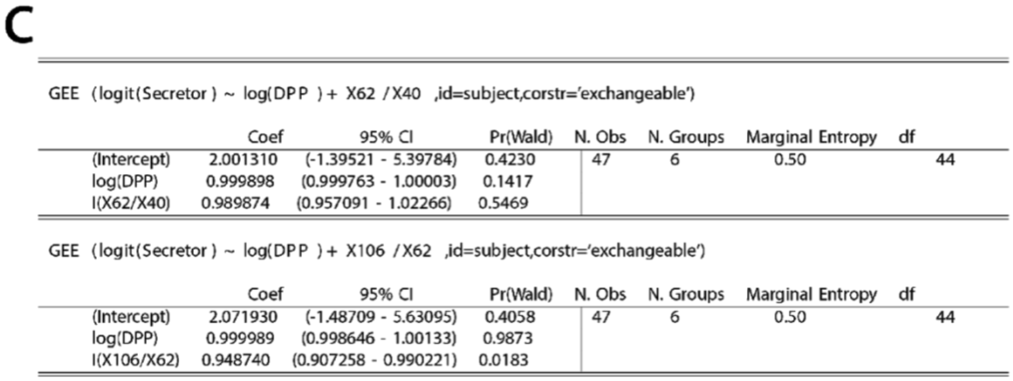
| | Coef | 95% CI | Pr(Wald) | N. Obs | N. Groups | Marginal | $R^2$ | df | Pr(Shapiro-Wilks) |
|---|---|---|---|---|---|---|---|---|---|
| (Intercept) | 1.081 | (0.90883 - 1.2538) | < 2e− 16 | 16 | 2 | 0.541 | | 14 | 0.222 |
| log(DPP) | -0.583 | (-0.55665 - -0.60976) | < 2e− 16 | | | | | | |

GEE (z(log(DSLNH + e)) ~ log(DPP ) + Secretor ,id=subject,corstr=’exchangeable’,data= ‘just-secretors’)

| | Coef | 95% CI | Pr(Wald) | N. Obs | N. Groups | Marginal | $R^2$ | df | Pr(Shapiro-Wilks) |
|---|---|---|---|---|---|---|---|---|---|
| (Intercept) | -0.993 | (-0.42991 - -1.5561) | 5.99e-04 | 31 | 4 | 0.448 | | 29 | 0.245 |
| log(DPP) | 0.528 | (0.39235 - 0.66399) | 5.68e-05 | | | | | | |

GEE (z(log(DSLNH + e)) ~ log(DPP ) + Secretor ,id=subject,corstr=’exchangeable’,data= ‘just-non-secretors’ )

| | Coef | 95% CI | Pr(Wald) | N. Obs | N. Groups | Marginal | $R^2$ | df | Pr(Shapiro-Wilks) |
|---|---|---|---|---|---|---|---|---|---|
| (Intercept) | -0.662 | (-0.55101 - -0.77302) | 1.01e-14 | 16 | 2 | 0.2028 | | 14 | 0.0203 |
| log(DPP) | 0.357 | (0.312 - 0.4021) | 2.91e-08 | | | | | | |



GEE (logit(Secretor ) ~ log(DPP ) + X62 / X40 ,id=subject,corstr=’exchangeable’)

| | Coef | 95% CI | Pr(Wald) | N. Obs | N. Groups | Marginal Entropy | df |
|---|---|---|---|---|---|---|---|
| (Intercept) | 2.001310 | (-1.39521 - 5.39784) | 0.4230 | 47 | 6 | 0.50 | 44 |
| log(DPP) | 0.999898 | (0.999763 - 1.00003) | 0.1417 | | | | |
| I(X62/X40) | 0.989874 | (0.957091 - 1.02266) | 0.5469 | | | | |

GEE (logit(Secretor ) ~ log(DPP ) + X106 / X62 ,id=subject,corstr=’exchangeable’)

| | Coef | 95% CI | Pr(Wald) | N. Obs | N. Groups | Marginal Entropy | df |
|---|---|---|---|---|---|---|---|
| (Intercept) | 2.071930 | (-1.48709 - 5.63095) | 0.4058 | 47 | 6 | 0.50 | 44 |
| log(DPP) | 0.999989 | (0.998646 - 1.00133) | 0.9873 | | | | |
| I(X106/X62) | 0.948740 | (0.907258 - 0.990221) | 0.0183 | | | | |

We first checked both the glycan-level and substructure-level clustering of the glycoprofile. Samples with same secretor status and days postpartum (DPP) were successfully grouped (Figure 1.13, 1.14). Further examination of the glyco-motif abundance (i.e., the total amount of substructure synthesized) revealed phenotype-related trends invisible on the glycan profile level. One observation of interest was that secretor status, defined by glycan fucosylation, significantly impacts the sialylation of non-fucosylated HMOs over time. While the relative abundance of both LSTb substructure (X62) and secreted LSTb was elevated in non-secretor milk (Wald p = $4 \times 10^{-7}$and Wald p = $3.98 \times 10^{-13}$; Figure 1.12a, b), only X62 showed a strong interaction between time and secretor status. At an adjusted sample size of 6, the time-dependent decrease in non-secretor X62 is significant (Wald p = 0.002). In contrast, the time-dependent decrease is only marginally significant for secreted LSTb (Wald p = 0.03). Previous work has already described an LSTb elevation at 3-4 months postpartum(Azad et al. 2018). Here, a substructure-analysis of X62 suggests that while the secreted LSTb is elevated in non-secretor milk, total LSTb produced (and consumed as the substrate for other sugars) may decrease over time.

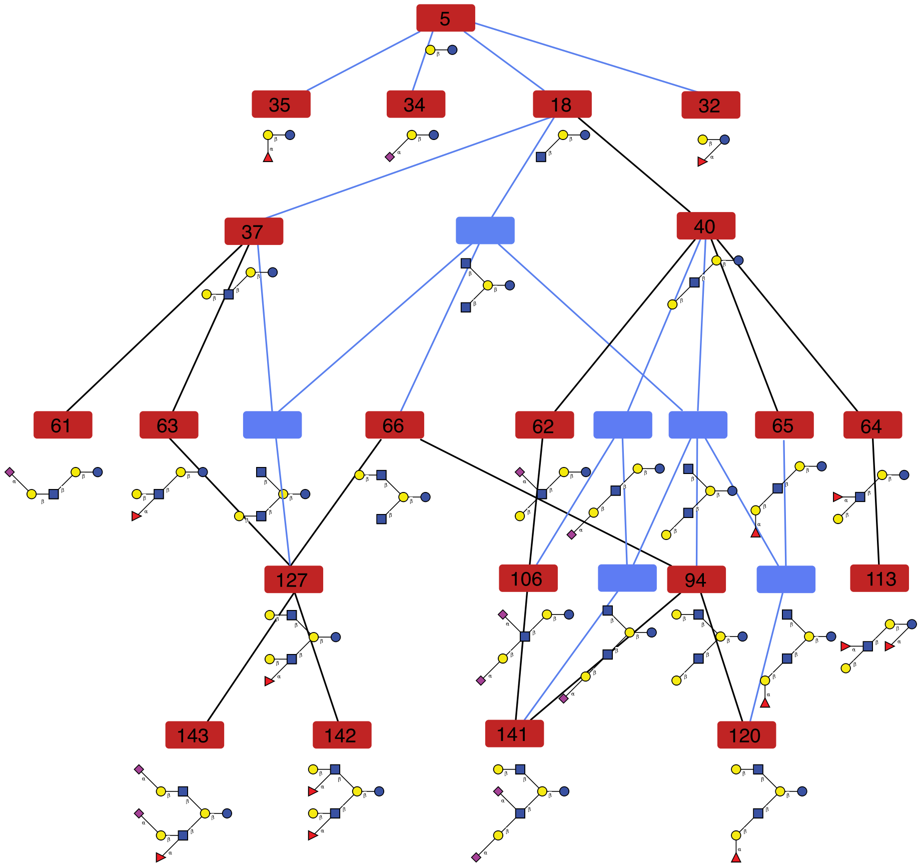


**Figure 1.13 HMO substructure network with dependent substructure removed.** All the glyco-motifs are shown, and redundant nodes are merged. This is a directed-acyclic-graph and the direction goes from top to bottom. An edge with black color is an important edge after merging that indicates the abundance changes. An edge with blue color is an edge that exists before merging that indicates the abundance variation between two substructures.

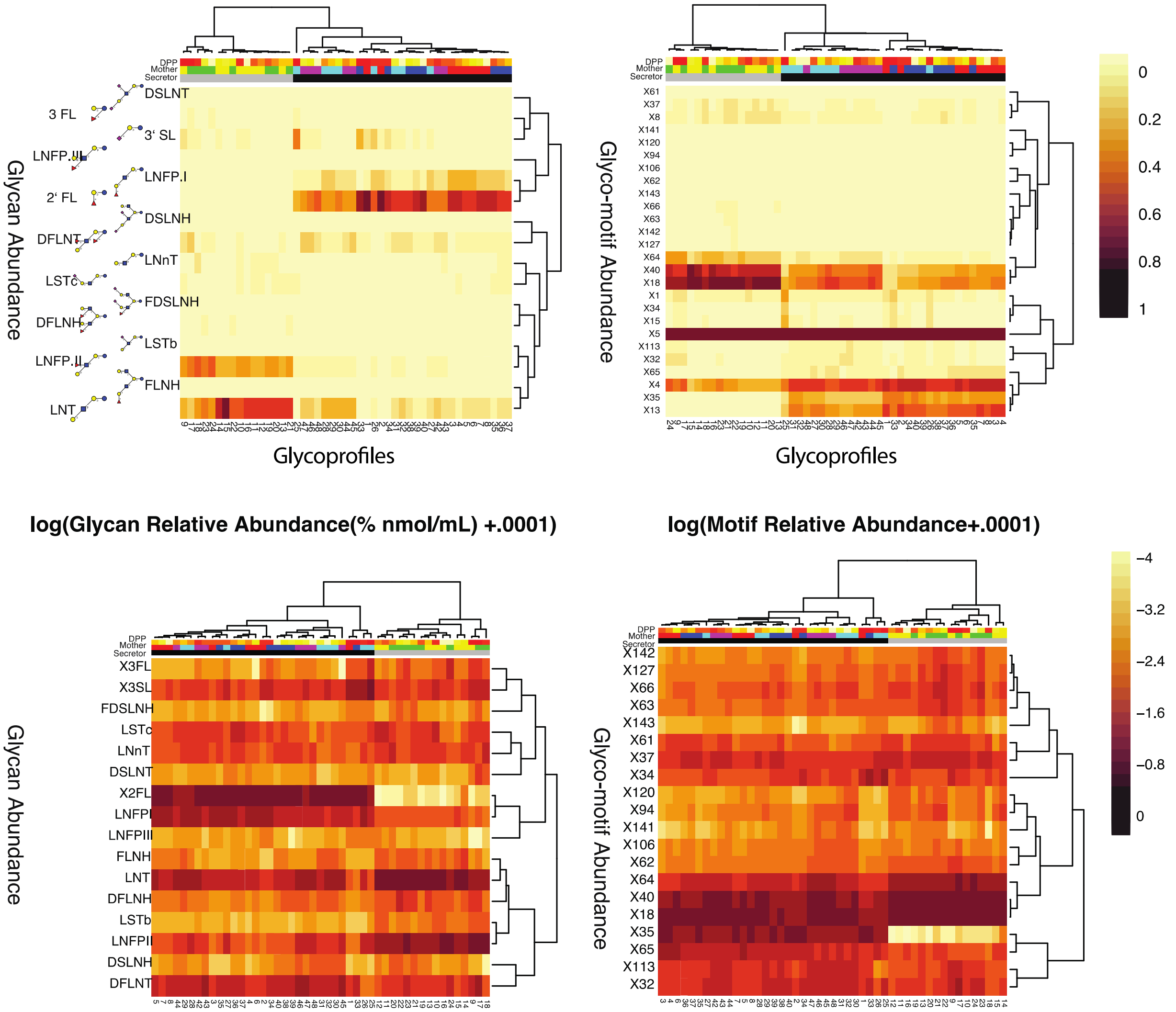


**Figure 1.14 HMO dataset, the clustering of HMO by glycan using Pearson correlation distance.** At the glycan-level, 2-fucosyllactose (2' FL) is the most abundant HMO in secretor mothers while Lacto-N-tetraose (LNT) and LNFPI are the most abundant HMOs in non-secretor mothers. The second major source of variance, DPP, shows a decrease in non-secretor LNFPI. At the substructure level, the clustering recapitulated the results from the raw HMO profiles and the α-1,2 fucosylated substructures were significantly associated with secretor status. The 2'FL substructure (X35) and the LNFPI substructure (X65) are significantly more abundant in secretor milk (Wald p=2.35x10$^{-25}$, Wald p=5.1x10$^{-12}$ respectively). The substructure abundance successfully reproduces the strongest effects known to be associated with secretor status. Source data are provided as a Source Data file.

Examining other secreted HMOs containing the X62 substructure (DSLNT and DSLNH), we see no significant secretor-status-dependent elevation (Wald $p > 0.2$; Figure 1.12a-d). Unlike X62, DSLNT shows no significant change over time (Coef=-0.39, Wald p = 0.17; Table 1.1a). Finally, DSLNH shows a significant increase over time (Wald $p = 2.91 \times 10^{-8}$ ; Table 1.1a). The secretor-specific trends in total LSTb are only clearly visible by examining the X62 substructure abundance (Figure 1.12a-d). Thus, while secretor status is expected to impact HMO fucosylation, GlyCompare reveals associations with non-fucosylated substructures. Viewing substructure abundance as total substructure synthesized provides a fundamental measure to the study of glycoprofiles (Figure 1.15); it also creates an opportunity to explore trends in synthesis.

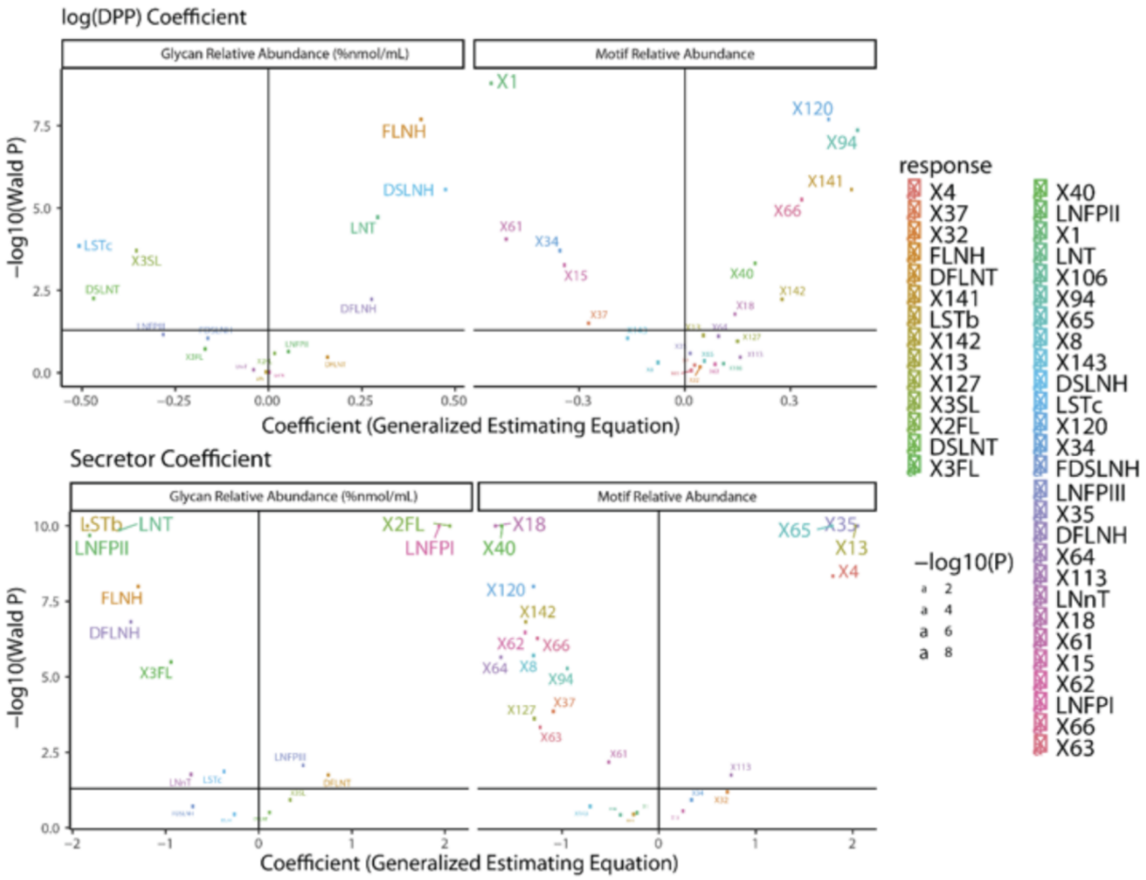


**Figure 1.15 HMO substructure general estimate equation coefficient and p-value plot.** Summary of regressions predicting either glycan or motif abundance from Days Postpartum (DPP) and Secretor status. The horizontal axis indicates the coefficient associating either DPP or secretor status with abundance and the vertical axis indicates the significance of that coefficient using the Wald test. Regression models were fitted using Generalized Estimating Equations (GEE) with an exchangeable covariance structure to control for dependency structures within mothers. Colors indicate the identification of the glycan or glycan motif, size indicates the significance of the coefficient and shape indicates if the coefficient was attributed to DPP or secretor status. Models fit to predict glycan abundance (left) were of the form: GEE(z(log(S+ $\epsilon$))) ~ DPP + secretor, while models to predict motif abundance were of the form: GEE(z(log(S+ $\epsilon$))) ~ DPP + secretor. Where, z(x) is a z-score normalization to center and standardize abundance and $\epsilon = 0.001$. This link function was chosen because it fit a normal distribution (Table 1.1a) and allowed for comparisons between regressions. There are some notable consistencies between the motif and glycan level results. As expected, 2’FL and its motif, X35, are both strongly and significantly enriched in secretor status. As are LNFPI and its motif, X65, are also strongly and significantly enriched with secretor status. Conversely, LSTb and X62 are negatively associated with secretor status. DPP has some significant but small negative associations with LSTc, 3’SL, and DSLNT. The 3’SL motif, X34 showed a consistent small negative significant association and the DSLNT motif. Most notably, X1, the sialic acid motif, was strongly negatively associated with DPP suggesting sialylation decreased in these samples over time. Source data are provided as a Source Data file.

**Flux estimation from GlyCompare identifies reaction responsible for an unexpected change in sialylation**

The identification of a non-fucosylated substructure that is associated with differences in secretor genotype raised the question of which reactions are responsible. Thus, we used GlyCompare to estimate enzyme fluxes to identify the reaction responsible for the unexpected change in HMOs . To do this, we estimate the flux for each biosynthetic reaction by quantifying the abundance ratio of products and substrates from parent-child pairs of glycan substructures. Thus, we could study changes in HMO synthesis through the systematic estimation of reaction flux across various conditions.

Although the fucosyltransferase-2 genotype defines secretor status, not all secretor-associated reactions were fucosylation reactions. We further explored the secretor-X62 association using the product-substrate ratio to estimate flux. Specifically, we examined the upstream reaction of LNT (X40) to LSTb (X62) and the downstream reaction of LSTb (X62) to DSLNT (X106) (Figure 1.12e). We estimated the flux of the upstream reaction of LNT converting to LSTb, using the X62/X40 ratio over time. However, no significant change was observed to secretor status (Figure 1.12f; Wald p=0.55). In the conversion of LSTb to DSLNT, we found a secretor-specific increase in reaction flux. Specifically, the X106/X62 ratio was significantly higher (Wald p=0.018) in secretor mothers (Figure 1.12g; Table 1.1c). In the average non-secretor mother, 52.3% (s.d. 15.1%) of LSTb is converted to DSLNT. Meanwhile, in secretors, on average, 81.8% (s.d. 7.2%) is converted. The LSTb to DSLNT conversion rate appears higher in secretors, while conversion from the LSTb precursor, LNT, appears unchanged (Figure 1.12f). Any changes in sialylation are intriguing, considering that secretor status is associated with genetic variation of a fucosyltransferase. A secretor-elevated conversion rate from LSTb to DSLNT is consistent with observing elevated X62 and secreted LSTb in non-secretor milk (Figure 1.12a-b)(Azad et al. 2018); if non-secretors consume less LSTb as a DSLNT substrate, more of the synthesized LSTb (X62) will remain LSTb through secretion. Examining the product-substrate ratio has revealed a phenotype-specific reaction propensity, thus providing insight into the condition-specific synthesis.

### GlyCompare increases the statistical power of glycomics data

GlyCompare successfully provides insights by accounting for shared biosynthetic routes of measured oligosaccharides. Since it includes information on the similarities between different glycans, we wondered how our approach impacts statistical power in glycan analysis. Thus, to quantify the benefit of the glyco-motif analysis, we constructed many regression models associating either glyco-motif abundance or glycan abundance, with a DPP and secretor status (see Methods). We found that regressions trained with glyco-motif abundance are more robust than those trained on whole glycan abundance, as indicated by the increased coefficient magnitude (Wilcoxon p = 0.0047, Figure 1.16a) and decreased standard error (Wilcoxon p = 0.033, Figure 1.16b). An increase in the stability of a statistic can result in an increased effect size. Consistent with the increased coefficient magnitude and decreased standard error, the effect size also increased, as measured by the marginal $R^2$ ($mR^2$) of glyco-motif-trained regressions (Wilcoxon p=0.04, Figure 1.16c). These effects were confirmed with a bootstrapping t-test; bootstrapping p-values were less than or equal to Wilcoxon p-values within 0.001. Increases in statistical magnitude, statistical stability, and effect size are all expected to increase analysis power. Using the median, 1st quartile, and 3rd quartile of observed $mR^2$, we estimated the expected power of glyco-motif-trained and glycan-trained regressions at various sample sizes. The expected power of a glyco-motif-trained regression reaches 0.8 at 36 samples and 0.9 at 57 samples. In contrast, a glycan-trained regression requires more than double the sample size to reach a comparable power (Figure 1.16d). GlyCompare provides additional power for discovering glycan-phenotypic associations.

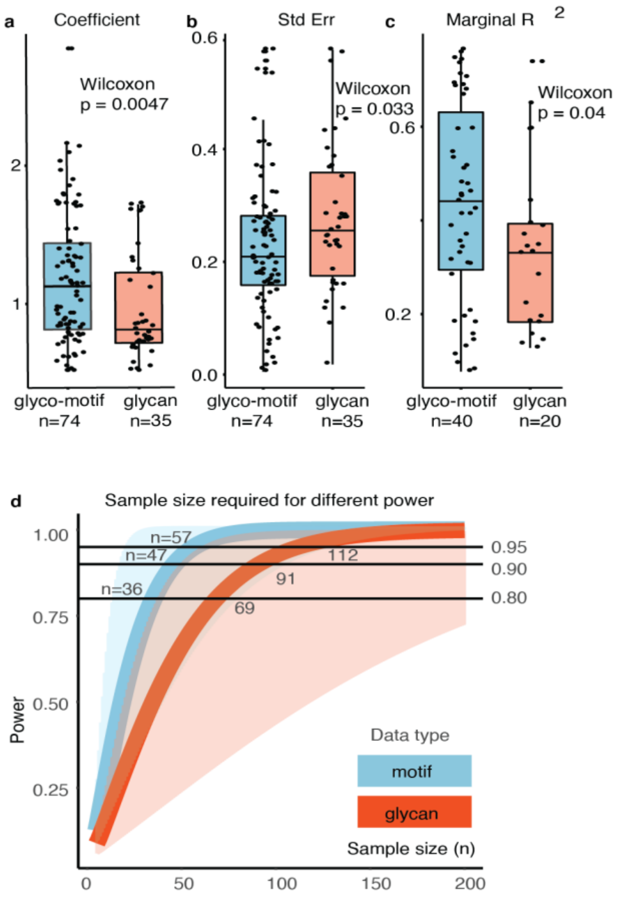


**Figure 1.16 Glyco-motif level statistics require half as many samples to reach the same level of statistical power as analysis with raw glycans. a, b** The use of glyco-motifs improves measures of regression robustness. The coefficient magnitude and Standard Error indicate the magnitude of the measured effect and the confidence with which a coefficient can be estimated. In **a,** the boxplot illustrates 25th, 50th, and 75th percentiles for regression coefficients using glycan data (Min=0.5094, Q1=0.7206, Median=0.8416, Q3=1.2706, Max=1.7166, n=35) or glyco-motif data (Min=0.5094, Q1=0.8365, Median=1.1403, Q3=1.5106, Max=2.8357, n=74). Distributions were compared using one-sided Wilcoxon tests (p=0.0047). In **b,** the boxplot again illustrates the 25th, 50th and 75th percentiles for regression Standard Error trained on glycan data (Min=0.0182, Q1=0.1631, Median=0.2446, Q3=0.2832, Max=0.4518, n=35) or glyco-motif data (Min=0.0053, Q1=0.1508, Median=0.2047, Q3=0.2747, Max=0.5398, n=74). Distributions were compared using a one-sided Wilcoxon test (p=0.033). **c** The $R^2$ describes the effect size of a regression; we used marginal $R^2$ ($mR^2$) because it was appropriate for the regression models used(Halekoh et al. 2006). Distributions for $mR^2$ of regression models trained on glycan data (Min=0.128, Q1=0.183, Median=0.331, Q3=0.441, Max=0.737, n=20) and glyco-motif data (Min=0.0949, Q1=0.3185, Median=0.46, Q3=0.686, Max=0.764, n=40) were compared using a one-sided Wilcoxon test (p=0.04). **d** We predicted power for a range of sample sizes (n=5-200) given the median effect size (solid line) within the interquartile range (shaded region) for glyco-motif-trained regressions ($mR^2$: Q1=0.31, Median=0.45, Q3=0.68) and the median effect size for glycan-trained regressions ($mR^2$: Q1=0.18, Median=0.33,Q3=0.44). Here, the use of GlyCompare and glyco-motif (grey-blue color) abundances required approximately half the number of samples to achieve equivalent power as standard glycan (red color) measures. Source data are provided as a Source Data file.

To further probe the increased statistical power, we compared our approach to another statistically-driven network approach. Benedetti et al. 2017 demonstrated that novel glycan biosynthetic reactions could be resolved using partial correlation(Benedetti et al. 2017). Using the Benedetti data, we computed partial correlation for glycan abundance and with GlyCompare-computed linkage-specified substructure abundance. We compared the partial correlation between glycans or substructures across true-positive, known reactions and false-positive, uncharacterized reactions (as specified in the Benedetti supplement). Partial correlations across known reactions between GlyCompare-computed substructures were significantly higher than partial correlations between corresponding glycan abundances (Figure 1.17). Partial correlation across known reactions was elevated for substructure abundance in all IgG isoforms (One-sided t-test, $p<0.0039$), and responses performed by B4GALT1 and ST6GAL1 (One-sided t-test, p<1.1x10-4). Interestingly, the lowest partial correlations across true-positive reactions between substructures were substantially higher than corresponding glycan correlations. The higher floor for substructure correlations suggests that substructure abundances may increase positive predictive value (Figure 1.17). Finally, while correlation increased between true-positive associated substructures, correlations across uncharacterized reactions were close to zero and indistinct from glycan correlations across the same reactions. Thus, using GlyCompare for glyco-motif-level analysis can substantially increase the robustness and statistical power in glycomics data analysis since it allows for comparing different glycans who share biosynthetic steps.

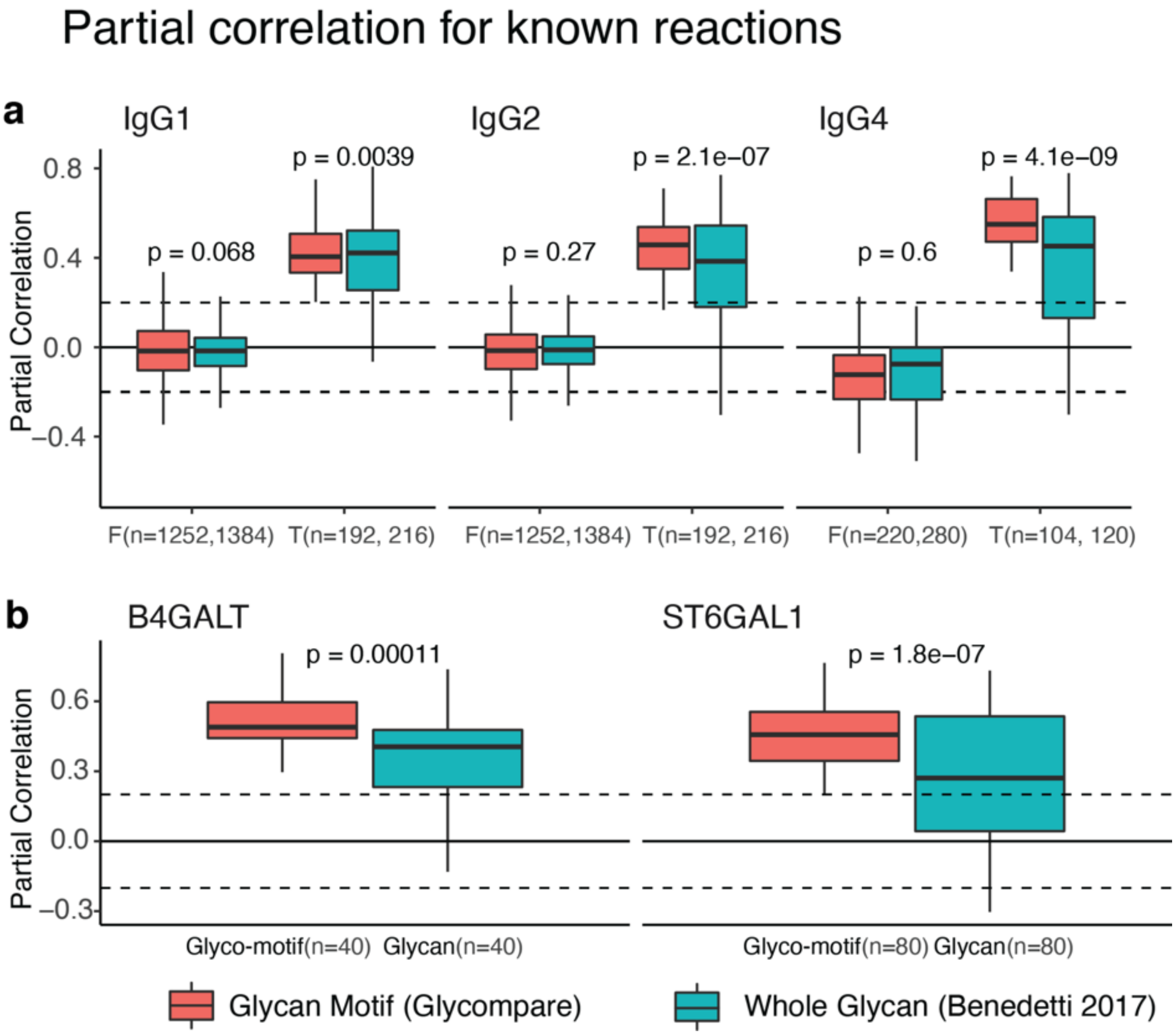


**Figure 1.17 Partial correlation between known and unknown biosynthesis reactions in N-glycosylation.** Glycan abundance data (MALDI-TOF) from Benedetti et. al. 2017[3] was realized to compute partial correlation between glycan abundance (as in the original paper, blue) and glycompare-computed linkage-specified substructure abundance (red). Partial correlations were stratified by prior knowledge, those known and previously characterized were designated the true-positive (T) reactions, while the other uncharacterized reactions were designated false-positive (F). The detailed information about the quartile boundary is provided in the Table 1.2. **a** A panel shows partial correlations that are split by IgG isoforms. The one-sided T.test are performed between the glycan abundance and substructure abundance (IgG1, F p=0.068, T p=0.0039; IgG2, F p=0.27, T p=2.1e-07; IgG4, F p=0.6, T p=4.1e-09); **b** A panel shows partial correlations split by related glycosyltransferases (B4GALT, T p=1.1e-04; IgG2, T p=1.8e-07). Source data are provided as a Source Data file.

**Table 1.2 Information for Figure 1.17.**

**a** provides the quartile information and sample size (n) of the Figure 1.17a and **b** provides the quartile information and sample size (n) of the Figure 1.17b.

| **a** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| IgG | truth | type | min | Q1 | median | Q3 | max | Sample(n) |
| 1 | F | motif | -0.5384 | -0.0983 | -0.0055 | 0.0887 | 1 | 1252 |
| 1 | F | glycan | -0.4874 | -0.0783 | -0.0109 | 0.0597 | 1 | 1384 |
| 1 | T | motif | 0.202 | 0.3331 | 0.4048 | 0.5074 | 0.8055 | 192 |
| 1 | T | glycan | -0.0646 | 0.255 | 0.4214 | 0.522 | 0.8079 | 216 |
| 2 | F | motif | -0.5248 | -0.0928 | -0.0077 | 0.0759 | 1 | 1252 |
| 2 | F | glycan | -0.6265 | -0.0729 | -0.008 | 0.0603 | 1 | 1384 |
| 2 | T | motif | 0.1666 | 0.3502 | 0.4574 | 0.5381 | 0.71 | 192 |
| 2 | T | glycan | -0.3031 | 0.1798 | 0.3843 | 0.5445 | 0.7704 | 216 |
| 4 | F | motif | -0.6441 | -0.1928 | -0.0998 | 0.1218 | 1 | 220 |
| 4 | F | glycan | -0.6354 | -0.2092 | -0.0582 | 0.0333 | 1 | 280 |
| 4 | T | motif | 0.3385 | 0.4713 | 0.5495 | 0.6628 | 0.7642 | 104 |
| 4 | T | glycan | -0.3013 | 0.1305 | 0.4522 | 0.5826 | 0.7783 | 120 |

| **b** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Enzyme | Truth | Type | min | Q1 | median | Q3 | max | Sample(n) |
| B4GALT | T | motif | 0.1666 | 0.4417 | 0.4889 | 0.5958 | 0.8055 | 40 |
| B4GALT | T | glycan | -0.1309 | 0.2322 | 0.4051 | 0.4771 | 0.7369 | 40 |
| ST6GAL1 | T | motif | 0.202 | 0.3443 | 0.4566 | 0.5547 | 0.7642 | 80 |
| ST6GAL1 | T | glycan | -0.3031 | 0.043 | 0.2707 | 0.5355 | 0.7315 | 80 |

### Additional statistical power reveals tumor-depleted mucin-type O-glycans

To explore the broad applicability of GlyCompare, we used our method to calculate substructure abundance for mucin-type O-glycans(Jin et al. 2017) (Figure 1.18), glycolipids(Sibille et al. 2016) (Figure 1.19), and site-specific compositional N-glycosylation(Riley et al. 2019) (Figure 1.20). In a re-examination of the mucin-type O-glycans from tumor and normal samples, glycan abundance and motif abundance were compared (Figure 1.18a, b). We found zero whole-glycan structures significantly distinguished between tumor and normal following multiple-test correction (FDR<0.1, Figure 1.18a). Yet, after substructure decomposition using Glycompare, we found five significantly depleted (FDR<0.1) mucin-type glycan motifs in gastrointestinal cancer (Figure 1.18b)(Jin et al. 2017). We found a substantial depletion in the tumor samples of five core 2 structures. These structures included three fucosylated and two with I-branches. The largest structures were over 30-fold depleted in tumors (FDR<0.03, Figure 1.18c). The core 2 depletion was noted as a nonsignificant trend in the original publication; we identified the specific core 2-type substructure depleted in tumors using substructure decomposition. Though this dataset contains few subjects and therefore may not be robustly generalizable, we demonstrate the increase in statistical power when using substructures. Additionally, a later study also found significant depletion of multiple bi-GlcNAc core 2 and I-branched structures(Jin et al. 2017). Also consistent with the decrease in bi-GlcNAc core-2 structures in gastric cancer, low expression of B3GNT3 in stomach cancer is significantly associated with decreased overall survival(Sibille et al. 2016). B3GNT3 is necessary for adding the second GlcNac to core 2 structure(Koda et al. 1996) and therefore upstream of all significantly depleted structures (Figure 1.18); B3GNT3 depletion could explain the observed differential glycosylation. The observation of significantly distinct substructures suggests GlyCompare provided increased statistical power to detect these distinguishing condition-enriched structures, and further showed continuity across similar structures was not evident in the original study.

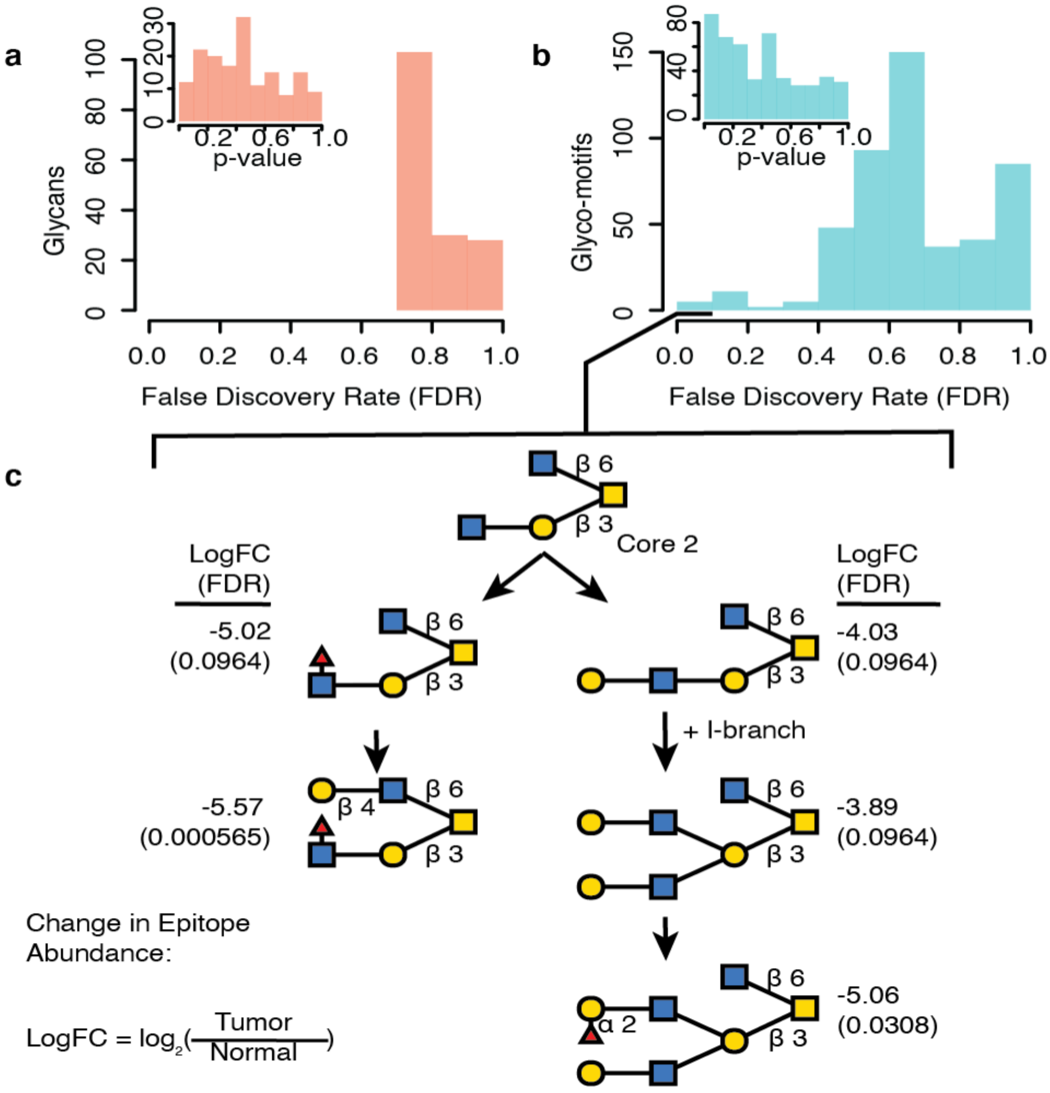


**Figure 1.18 Increased power for identifying diagnostic markers is shown through a re-analysis of mucin-type O-glycans from normal, tumor-proximal and gastrointestinal cancer biopsies, transformed to motif abundance. a** and **b** Welch Two Sample t-test P-value and False Discovery Rate (FDR) distributions for glycan abundance and glyco-motif abundance. **c** We found multiple core 2 substructures depleted in gastrointestinal cancer relative to normal tissue. Not all linkages are specified, only those relevant to the substructure definition. The information of log fold changes (logFC) and the FDR are presented next to each substructure. Source data are provided as a Source Data file.

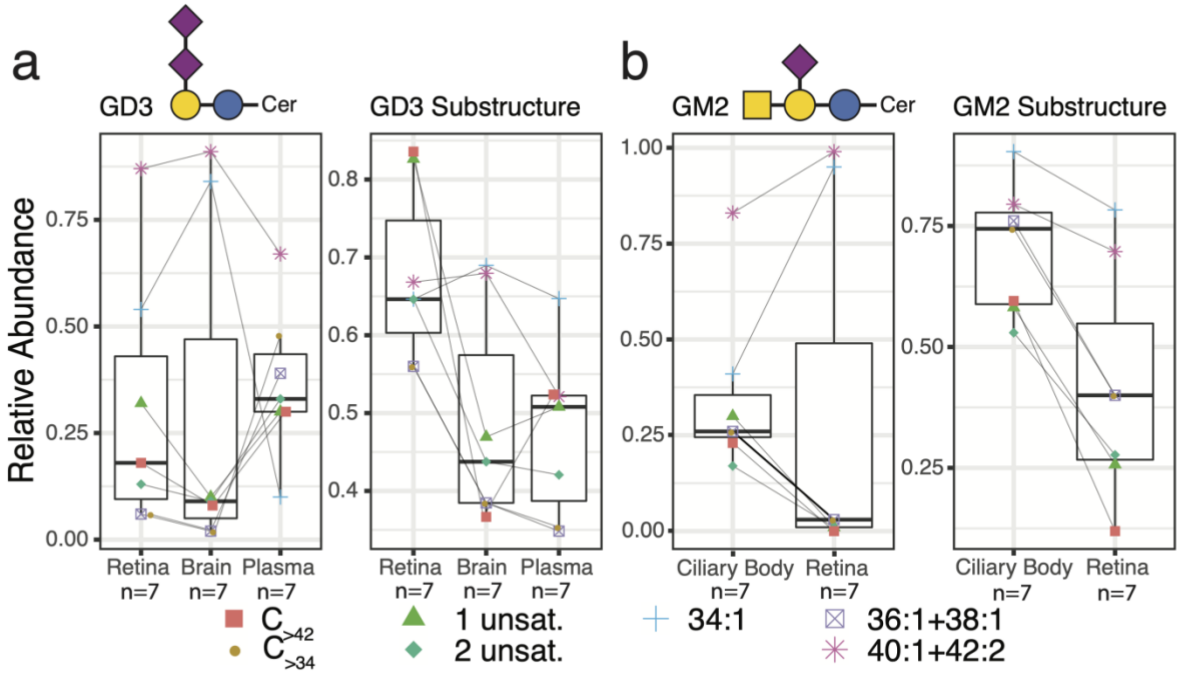


**Figure 1.19 A re-analysis of ganglioside glycolipid abundance pooled across various ceramide types.** Glycolipid and substructure abundance from a lactose root clarified distinct glycosylation in the retina[4]. **a,** Retinal GD3 substructure abundance is enriched in retina across nearly all ceramide species while the same effect is not visible at the whole glycan level. **b,** Retinal GM2 is depleted relative to the proximal ciliary body but the effect is only visible at the substructure level. Ceramide groupings include more than 42 or fewer than 35 Carbons ($C_{>42}$, $C_{<35}$), either 1 or 2 unsaturated bonds (1 unsat., 2 unsat), or groups of specific ceramides with X:Y carbons and unsaturated bonds (e.g. 34:1, (36:1+38:1), or (40:1+40:2). The subjects N=7 for all boxplots. The quartiles information is recorded on the Table 1.3. Source data are provided as a Source Data file.

**Table 1.3 Information for Figure 1.19.**

|  | Tissue | Min | Q1 | Median | Q3 | Max |
|---|---|---|---|---|---|---|
| GD3 | Retina | 0.06 | 0.095 | 0.18 | 0.43 | 0.87 |
|  | Brain | 0.02 | 0.05 | 0.09 | 0.47 | 0.91 |
|  | Plasma | 0.1 | 0.3 | 0.33 | 0.435 | 0.67 |
| GD3-Substr | Retina | 0.56 | 0.603 | 0.646 | 0.747 | 0.836 |
|  | Brain | 0.367 | 0.385 | 0.438 | 0.574 | 0.69 |
|  | Plasma | 0.348 | 0.387 | 0.508 | 0.522 | 0.647 |
| GM2 | Retina | 0 | 0.01 | 0.03 | 0.49 | 0.99 |
|  | Ciliary.Body | 0.17 | 0.245 | 0.26 | 0.355 | 0.83 |
| GM2-Substr | Retina | 0.119 | 0.267 | 0.4 | 0.549 | 0.783 |
|  | Ciliary.Body | 0.529 | 0.589 | 0.744 | 0.7778 | 0.904 |

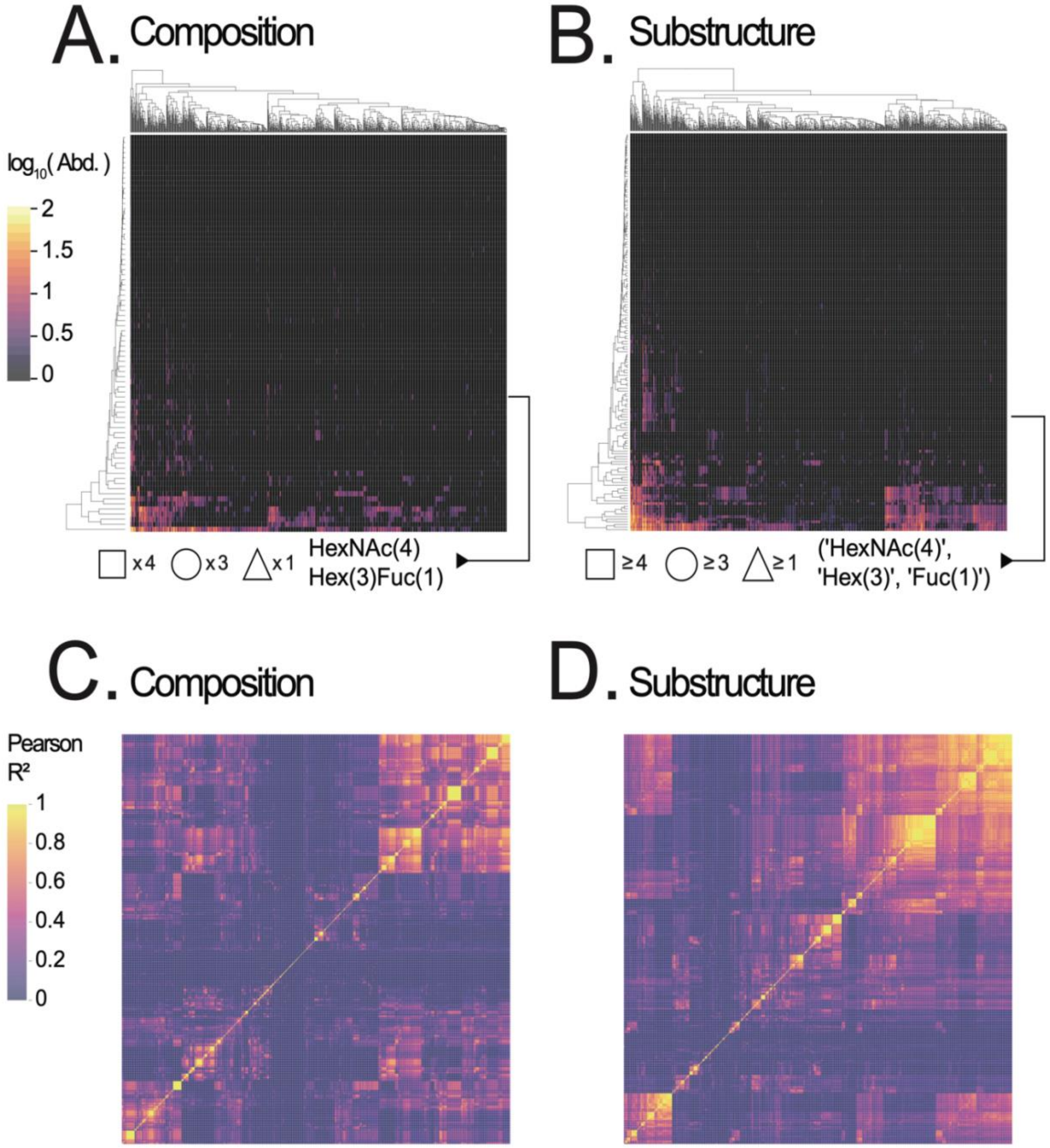


**Figure 1.20 A reanalysis of site-specific N-glycosylation in mouse brains.** Biclustered heatmaps of compositional **a** site-specific N-glycan data from mouse brain[5]. **b** The same compositional data was substructure-decomposed to calculate substructure abundances presented in another biclustered heatmap. **c, d** The Pearson correlation coefficient was calculated for the compositional and composition substructure abundance for each glycosylation site across proteins. Biclustered heatmaps of the resulting correlation coefficients are present as biclustered heatmaps. Biclustering used a complete agglomerative approach. See Appendix Method for detail.

### Re-analysis of ganglioside glycolipid abundance pooled across various ceramide types

When ganglioside and substructure abundance was pooled by ceramide types, we found the GD3 substructure enriched in retina relative to brain and plasma, while the GD3 ganglioside abundance showed no coherent effect (Figure 1.19a). Similarly, the GM2 substructure was enriched across several ceramide types in the Ciliary-body relative to the retina, while the GM2 ganglioside showed no coherent effect (Figure 1.19b). By aggregating over subtypes, we can account for confounding biosynthetic complexity thereby simplifying analyses and making crucial insights more accessible.

### Re-analysis of site-specific N-glycosylation in mouse brains

Examining site-specific N-glycan compositional data from rat brain, we found that the decomposition of composition abundance into composition substructure abundance reveals additional potential signal. As previously shown, the sparsity of the abundance matrix decreases, and the comparability of profiles is improved when glycan data is aggregated over substructures (Figure 1.20a, b). Further, the correlation structure of substructure aggregated abundance (Figure 1.20d) appears more robust than its compositional counterpart (Figure 1.20c); there are more clusters with clearer borders, multiple clear off-diagonal clusters and the median $R^2$ is approximately doubled. While it is possible that the higher correlation is indicative of an increased background, that is unlikely considering the increase in visible correlation is structured, not randomly distributed through the background.

### Discussion

Glycosylation has generally been studied from the whole-glycan perspective using mass spectrometry and other analytical methods. From this perspective, two glycans that differ by only one monosaccharide are distinct and are not directly comparable. Thus, the comparative study of glycoprofiles has been limited to changes between glycans shared by multiple glycoprofiles or small manually curated glycan substructures(Rademacher and Paulson 2012). GlyCompare sheds light on the hidden biosynthetic between glycans by integrating the structural similarity into the comparison. Glycoprofiles are converted to glyco-motif profiles; wherein each substructure abundance represents the cumulative abundance of all glycans containing that substructure. In another word, substructure abundance is automatically transformed from the upstream data; motif selection simply allows the user to focus on the least substructures necessary to understand their dataset. This quantification of substructures can be easily scaled up to compare more glycoprofiles in large datasets. Thus, it brings several advantages and different perspectives, with some important limitations, to enable the systematic study of glycomics data.

Like any analytical pipeline, GlyCompare is sensitive to upstream analysis (e.g., mass spectrometry methods measure the mass-to-charge ratios of glycans and their fragments, and thus require expert annotation to assign structures). Therefore, GlyCompare will continue to improve with advances in glycoprofile structure annotation quality. Going forward, we hope to include multiple methods for aggregating abundance over substructures, including aggregation using multiple functions (besides addition) over fully or partially specified biosynthetic networks. While summing abundance for all subsumed substructures works well, manual reaction specification can help avoid information loss when biosynthesis is not hierarchical and acyclic or glycans are not increasing in size. When these limitations are acknowledged, the current version of glycompare has demonstrated some exciting capabilities.

First, the GlyCompare platform computes a glyco-motif profile (i.e., the abundances of the minimal set of glycan substructures) that maintains the information of the original glycoprofiles, while exposing the shared intermediates of measured glycans. These glyco-motif profiles more accurately quantify similarities across glycoprofiles. This is made possible since glycans that share substructures also share many biosynthetic steps. If the glycan biosynthetic network is perturbed, all glycans synthesized will be impacted and the nearest substructures will directly highlight where the change occurred. For example, in EPO glycoprofiles studied here, the tetra-antennary structure is depleted in the Mgat4a/4b/5 knockout group and the downstream sialylated substructure depleted when St3gal4/6 were knocked out. Such structural patterns emerge in GlyCompare since the tool leverages shared intermediate substructures for clustering, thus identifying common features across diverse samples.

Second, trends in glycan biosynthetic flux become visible at the substructure level. For example, in the HMO data set, multiple glycans are made through a series of steps from LNT to DSLNH (Figure 1.9a). Only when the substructure abundances and product-substrate ratios are computed can we observe the secretor-dependent temporal differences in the abundance of the LSTb substructure, X62. This is particularly interesting. Though changes in α-1,2 fucosylation define secretor status, we see additional secretor-dependent effect on sialylated structures with no fucose. The biosynthetic interpretation of root-based substructures was applied to ocular gangliosides(Sibille et al. 2016) to identify tissue-specific glycolipid substructures (Figure 1.19). These are the systemic effects invisible without a systems-level perspective due to the interconnected nature of glycan synthesis; this disparity underlines the power of this method.

Third, the sparse nature of glycomic datasets and the synthetic connections between glycans make glycomic data unfit for many common statistical analyses. However, the translation of glycoprofiles into substructure abundance provides a framework for a more statistically powerful and robust analysis of glycomic datasets. These methods can enrich both structural (Figure 1.3a) and compositional (Figure 1.20) thereby increasing the interpretability and structure of the dataset. Single sample perturbations, such as the knockouts in the glycoengineered EPO, can be compared to wild-type; all substructure data can be normalized and rigorously distinguished from the control using a one-sample Wilcoxon-test. Furthermore, conditions or phenotypes with many glycoprofiles, such as the secretor status in the HMO dataset, can be compared using various statistical methods to evaluate the association between the phenotypes and glycosylation. For example, in HMO data, we revealed that the α-1,2 fucose substructure is enriched in secretor status, consistent with previous studies(Koda et al. 1996; Kudo et al. 1996; Viverge et al. 1990). Because the substructure approach includes comparisons of glycans that are not shared across the different samples but share intermediates, GlyCompare decreased sparsity and increased statistical power. We demonstrate the increase in statistical power and observable differences between HMO (Figure 1.16) and the tumor-proximal mucin-type O-glycan presentation (Figure 1.18). Thus, one can obtain richer glycan comparisons of representative substructures, total synthesized abundance, and flux.

Finally, in combination with the substructure network, we can systematically study glycan synthesis. The product-substrate ratio provides an estimation of flux through the glycan biosynthetic pathways. Using the HMO dataset, we demonstrated the power of this perspective by showing that more LSTb is converted to DSLNT in the secretor mother. The perspectives made available through GlyCompare are not limited to Wilcoxon-tests and regression models. Because the substructure-level perspective minimizes biosynthetic dependency between glycans, glyco-motif abundances can be used with nearly any statistical model or comparison demanded by a dataset. We have accommodated the sparse and non-independent nature of glycoprofiles, thereby making countless comparisons analyses possible.

## Methods

### The overview of the pipelines

The Figure 1.21 showed a summary of GlyCompare workflow. The GlyCompare workflow consists of several steps wherein glycoprofiles are annotated and decomposed, glyco-motifs are prioritized, and each glyco-motif is quantified for subsequent comparisons with or without specific phenotype data.

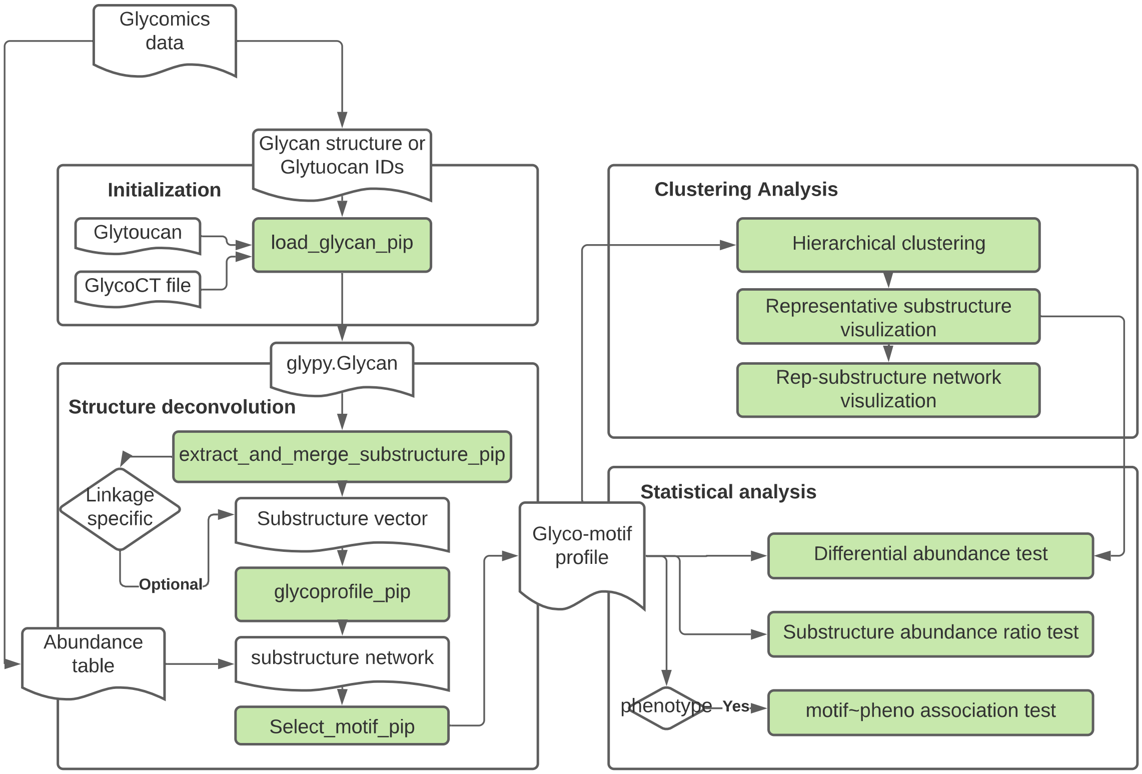


**Figure 1.21 The workflow of the pipeline.** This flowchart provides a basic overview of the glycompare platform. Glycomics data which contain the glycan structure (or compositional) information and abundance information, are fed into the pipeline. The green boxes are the main pipeline function in the glycompare platform. The glycan structures are loaded as glypy.Glycan object at the initialization step. Then, the glyco-motif vectors are generated with the help of the glycan abundance. After that, glycol-motif profiles are delivered to the clustering analysis and statistical analysis modules.

**N-glycosylation of EPO glycoprofile collection for re-analysis**

N-glycosylation data were previously published(Yang et al. 2015). Upon retrieving these data from the study, we picked 16 glycoprofiles that are used again in their follow-up study(Čaval et al. 2018) and further processed the data as follows. All measurements were taken from distinct samples.

Glycan substructures were extracted from the observed glycans. Substructure abundance was calculated from the glycan abundance of all glycans containing the substructure. The substructure network identifies a minimal set of 151 glyco-motif substructures to compare the mutants. Finally, representative substructures were extracted to pool abundance and summarize the structural changes across mutants. Each of these operations is further specified below.

**HMO glycoprofile collection and analysis**

HMOs were analyzed as de-identified samples previously for an independent study(Mohammad, Hadsell, and Haymond 2012; Mohammad and Haymond 2013) at Baylor College of Medicine. Following Institutional Review Board approval (Baylor College of Medicine, Houston, TX), lactating women provided written informed consent. Women with diabetes or impaired glucose tolerance, anemia, or renal or hepatic dysfunction were excluded from the study. Women were 18-35 years of age, had uncomplicated singleton pregnancies with vaginal delivery at term (>37 weeks) and pregnancy Body Mass Index (BMI) remained <26kg m$^{-2}$. Infants were healthy and exclusively breastfed. Forty-eight milk samples were collected from 6 human mothers (1, 2, 3, 4, 7, 14, 28, and 42 days postpartum (DPP)). More information on subject selection, exclusion, study design, and breast milk collection has already been published(Mohammad, Hadsell, and Haymond 2012; Mohammad and Haymond 2013).

Glycan composition and abundance were measured by high-performance liquid chromatography (HPLC) following fluorescent derivatization with 2-aminobenzamide (2AB, CID: 6942)(Bode et al. 2012; Alderete et al. 2015). Raffinose (CHEBI:16634, CID:439242), a non-HMO oligosaccharide, was added to each milk sample as an internal standard at the beginning of sample preparation to allow for absolute quantification. Of the 300-500 predicted HMO, the 16 most abundant HMO were detected based on retention time comparison with commercial standard oligosaccharides and mass spectrometry analysis, including 2-fucosyllactose (2'FL), 3-fucosyllactose (3'FL), 3-sialyllactose (3'SL), lacto-N-tetraose (LNT), lacto-N-neotetraose (LNnT), lacto-N-fucopentaose (LNFP1, LNFP2 and LNFP3), sialyl-LNT (LSTb and LSTc), difucosyl-LNT (DFLNT), disialyllacto-N-tetraose (DSLNT), fucosyl-lacto-N-hexaose (FLNH), difucosyl-lacto-N-hexaose (DFLNH), fucosyl-disialyl-lacto-N-hexaose (FDSLNH) and disialyl-lacto-N-hexaose (DSLNH). GlyTouCan IDs for each glycan are listed in Table 1.4.

**Table 1.4 HMO abbreviations.**

HMO abbreviations are specified in this table. Complete GlycoCT structures and GlyTouCan accession, for all HMO and EPO glycans used in this study, can be accessed at https://github.com/LewisLabUCSD/GlyCompare/blob/master/example_data/Glycan_Structures.complete.xlsx.

| **HMO** | **Abbreviation** | **GlyTouCan Accession** |
|---|---|---|
| LNT | Lacto-N-tetrose | G45827GY |
| LNnT | Lacto-N-neotetrose | G48059CD |
| 2’FL | 2’-fucosylactose | G10422IZ |
| 3FL | 3-fucosylactose | G06210XB |
| 3’SL | 3’-sialyllactose | G91237TK |
| LNFPI | Lacto-N-fucopentose I | G01650PH |
| LNFPII | Lacto-N-fucopentose II | G98173LG |
| LNFPIII | Lacto-N-fucopentose III | G83916HL |
| LSTb | LS-tetrasaccharide b | G19017MP |
| LSTc | LS-tetrasaccharide c | G72506RN |
| DSLNT | Disialyllactose-N-tetrose | G38710SX |
| DFLNT | Difucosyllacto-N-tetrose | G70115XG |
| FLNH | Fucosyllacto-N-hexose | G24504JY |
| DSLNH | Disialyllacto-N-hexaose | G47928KI |
| DFLNH | Difucosyllacto-N-hexaose | G63053GR |
| FDSLNH | Fucodisialyllacto-N-hexaose | N/A |

HMO measurements by HPLC were quantified using Chromeleon 7.2(Greco et al., n.d.). Technicians were blinded to metadata associated with each sample. No samples were excluded. The HPLC failed to quantify HMO in the day 1 sample collected from subject L6, therefore, no data from this sample could be included. Samples were analyzed in a random order to mitigate batch effects. In addition to absolute concentration of each glycan $w_i$, the proportion of each glycan per total glycan concentration (sum of all integrated glycans) was calculated and expressed as relative abundance (% of the total, $w_i/\Sigma w_*$). The presence of 2-FL defines secretor status. Absolute abundance of HMO is determined by a well-characterized low-noise method(Bode et al. 2012; Alderete et al. 2015) using HPLC analysis(McGuire et al. 2017). Therefore, no technical replicates were necessary.

HMO abundance profiles were treated similarly to the N-glycans. We identified and quantified 26 glyco-motifs from 121 substructures. We compared glyco-motif abundance and their abundance ratios directly to secretor status along with the log of days postpartum.

**Computing glycan substructure profiles from glycoprofiles**

Three procedures were used for preprocessing the studied glycoprofiles (Figure 1.1c). First, glycoprofiles are parsed into glycans with abundance. In each glycoprofile, the glycans are manually drawn and exported with GlycoCT format using the GlyTouCan Graphic Input tool(Aoki-Kinoshita et al. 2015). GlycoCT formatted glycans are loaded into Python (version 3+) and initialized as glypy.glycan objects using the Glypy (version 0.12.1)(Klein and Zaia 2019). Assuming we have a glycoprofile $i$, the corresponding abundance of each glycan $j$ in glycoprofile $i$ is represented by $g_{ij}$. For example, the relative m/z peak in the mass spectrum or the abundance value in an HPLC trace, is calculated relative to the total abundance of glycans in this glycoprofile $g_{ij}/\Sigma g_{i*}$. Glycans with ambiguous topologies are handled by assuming they belong to every possible structure with equal probability, thereby creating all possible $n$ structures, still, with $g_{ij}/n\Sigma g_{i*}$ abundance of each. Second, glycans are annotated with glycan substructure information, and this information is transformed into the substructure vector. Substructures within a glycan are exhaustively extracted by breaking down each linkage or a combination of linkages of the studied glycan. Note that this method cannot currently deal with cyclic glycans. All substructures extracted are merged into a substructure set **S**. Substructures are sorted by the number of monosaccharides and duplicates are removed. Then, each glycan is matched to the substructure set **S**, producing a binary glycan substructure presence (1) or absence (0) vector, $x_{ij}$. Lastly, a substructure (abundance) vector is calculated as $P_i = \Sigma x_{ij} g_{ij} / \Sigma g_{i*}$ representing the abundance of the substructures $s$ in this glycoprofile, where $P_i = (s_{1i}, \ldots, s_{ni})$. Third, a substructure network is built based on the substructure vectors. The substructure network is a directed acyclic graph wherein each node denotes a glycan substructure. Given the substructure set **S**, the root node starts from the monosaccharides or a defined root core structure, and a child node is a substructure with only one monosaccharide added to its parent node. We note that one child node might have multiple parent nodes and vice versa. The child node depends on its parent node(s) since it cannot exist without any parent node. The edges in the substructure network were annotated with known synthetic rules for further analysis. Substructure networks were visualized by networkx (version 2.1; https://networkx.org/). and cytoscape (version 3.8.2)(Shannon et al. 2003).

**Selecting glyco-motifs from the substructure network**

A larger subset of the substructure network is necessary to uniquely describe a more diverse set of glycoprofiles, while fewer substructures are needed to describe more similar glycoprofiles sufficiently. Comparisons become more focused when only examining these variable substructures. To simplify the substructure network, the parent/child substructure pair that have the same abundance can be merged without any information loss. As illustrated in Figure 1.2d, a parent-child substructure pair with the same abundance (solid arrow) can be merged. If they have the same abundance, we can conclude that the addition of the specific monosaccharide is not perturbed across all glycoprofiles, which means they carry the same information. Thus, the parent node can be pruned without information loss. All remaining nodes, namely, the glyco-motifs, are used to cluster the glycoprofiles.

After selecting the glyco-motifs (Figure 1.2d), we use the "monosaccharides weight" to track whose parent node is merged. All node weights are initialized as 1. When a node is removed, the weight is equally divided and distributed to child nodes that have the same abundance as the removed node. Since this method redistributes weight from the root to leaves, the descendant substructure node starting having different abundance from the parent node will gain the most weight. The weights **W** are used later for generating the representative substructures.

### Substructure based clustering of glycoprofiles

After generating the glyco-motifs, the Pearson correlation and 'complete' distance are used to cluster the glycoprofiles and substructures (Figure 1.2e). The elbow method is used to determine the cluster numbers. To identify the representative glycan substructures, a set of glycan substructures with weights **W** is first aligned (Figure 1.2f). Then, we calculate the sum of monosaccharide weights for each glycan substructure. The representative substructure is thus defined as the glycan substructures with their summed monosaccharide weights greater than 51% (a heuristic and flexible parameter to facilitate user-controlled clarity) of the total weight of glycan substructures. Lastly, the averaged abundances of the representative substructures are generated to assess their differential expressions between different glycoprofiles.

### Substructure cluster abundance comparison and network edge-based ratio comparison

We use the representative substructures to summarize and analyze the structural and quantitative changes across glycoprofiles. For the abundance of a representative substructure in a glyco-motif cluster, we combine the substructure abundance and the substructure monosaccharide weights to generate the weighted average of substructure abundance. Since the abundance range of representative substructures across different glycoprofiles is different, we re-centralized the representative substructure abundance based on WT and scaled them with standard deviation. There are many representative substructures significantly deviating from the WT's abundance. Since the abundance distributions are not normally distributed, we used a one-sided 1-sample Wilcoxon test to test if the abundance of a representative substructure in a glycoprofile is significantly divergent. Effect size, *r*, was calculated as z/sqrt(N)(Rosenthal and Rubin 1991). A Bonferroni correction (n=16) was used to correct for multiple testing, so p=0.0031 is used as criteria, and effect sizes are all above 0.68.

For those network edges annotated with enzyme information, we further test if an enzyme has the same efficacy in two glycoprofiles. Every edge has a parent/child abundance ratio. All edges annotated with the same enzyme consist of an abundance ratio distribution in one glycoprofile (Figure 1.3c). The Wilcoxon test is used to compare the ratio distribution for the same enzyme in two glycoprofiles.

To have a concise view of the representative substructure network, we further generate a simplified network. The nodes from the substructure network are merged based on the substructure clustering. The edges connecting the original nodes are merged to connect the new nodes. Lastly, the derived representative substructure network represents the merged nodes and the edges annotated by enzymatic rules (Figure 1.9b).

**Phenotype-associated substructure detection**

For revealing the phenotype-associated substructures, we estimated the influence of secretor status on glycan and glyco-motif abundance for revealing the phenotype-associated substructures using a generalized estimating equation (GEE, R3.6::geepack(Yan and Fine 2004; Halekoh et al. 2006)). GEE models account for resampling bias in longitudinal measurements(Zeger and Liang 1986); other regression models, like generalized linear models, overestimate the sample size and power by ignoring this bias. Unlike mixed effect models, which can account for resampling bias, GEE allows non-linear relations between the outcome and covariates, while accounting for correlation among repeated measurements from the same subject. Here we used GEE with an exchangeable correlation structure (assuming the within-subject correlation between two time-points is ρ). We log and z-score standardized each glycan and glyco-motif measurement to stabilize the variance and equalize the range. We also used the log of days postpartum (DPP) to linearize the relationship over time. The Wald test was used to measure the significance of Secretor status contribution. For additional information and diagnostic statistics for specific regressions, see Table 1.1a and 1.3b. All regression results can be found in Figure 1.15.

### Product-substrate ratio as a proxy for flux and estimating flux-phenotype associations

To further isolate glyco-motif-specific effects from biosynthetic biases, we explored methods to control for the product-substrate relations. First, we extract the relative abundance of parent-child pairs of glyco-motifs in the substructure network; these are product-substrate relations like LNT and LSTb (Figure 1.9e). Glyco-motif abundance represents the total substructure synthesized; therefore, when we examine the product-substrate ratio, we measure the total amount of the substrate substructure converted to the product substructure in the sample. Thus, the product-substrate ratio is a proxy for flux. Using logistic GEE regression modeling, similar to the approach used for testing substructure-phenotype associations, we can measure the influence of estimated flux between two glycans on secretor status; here we predicted secretor status from the estimated flux log(DPP). For additional information and diagnostic statistics, see Table 1.1c.

### Glyco-motif Abundance Robustness and Power Analysis

Similar to those used in Figure 1.14, GEE models were trained using either glyco-motif or whole glycan relative abundance. To stabilize the variance, equalize the range, and make the regressions comparable, we used a square root and z-score normalization on each glycan and glyco-motif measurement. Glyco-motif or relative glycan abundance was predicted from either DPP alone, Secretor status alone, DPP + Secretor status, or DPP + Secretor status + DPP:Secretor. To avoid biasing the analysis with misfit or uninformative models, models with small coefficients (|coef|<0.5) or non-normal abundance distributions (Shapiro-Wilks p < 0.001) were removed. Model robustness measures including, coefficient magnitude ($n_{glycan\text{-}stats}$=39, $n_{motif\text{-}stats}$=86), standard error ($n_{glycan\text{-}stats}$=39, $n_{motif\text{-}stats}$=86) and marginal $R^2$ ($n_{glycan\text{-}stats}$=21, $n_{motif\text{-}stats}$=47) were used to compare model performance. Robustness measures from glycan-trained and glyco-motif-trained models were compared using a one-sided Wilcoxon rank sum test with continuity correction. We validated these findings using a 10,000 iteration one-sided, two-sample bootstrapping t-tests (Rv3.6::nonpar::boot.t.test); bootstrapping p-values were less than or equal to Wilcoxon rank sum p-values within 0.001. Finally, using the Rv3.6::pwr::pwr.r.test v1.2.2 package, statistical power was predicted between n=5 and n=200 for the median and interquartile range of effect sizes observed in glyco-motif-trained and glycan-trained models.

### Substructure decomposition of published IgG N-glycosylation to distinguish known and unknown biosynthetic reactions

We re-analyized structural N-glycan data from IgG (Benadetti_2017)(Benedetti et al. 2017). IgG N-glycans were measured using liquid chromatography coupled with electrospray mass spectrometry (LC-ESI-MS). Pre-processing of these data was restricted to reformatting for input into Glycompare-compatible abundance matrix and structure annotation. Glycoprofiles were normalized to relative abundance. Substructure abundances and motif extraction were performed using an N-glycan thereby focusing analysis on biosynthetic motifs.

Using the IgG N-glycan data, we estimated partial correlation(Opgen-Rhein et al. 2007) between glycan abundances or between motif abundances. Previously, glycan abundance partial correlation was used to identify previously uncharacterized N-glycan biosynthetic reactions(Benedetti et al. 2017). Here, we used motif abundance partial correlation and compared predicted power. Edges (partial correlations between glycans or motifs) were filtered for direct relations (structures differing by only one monosaccharide), split into known (True) and unknown (False) reactions. Partial correlation distributions were stratified by prior knowledge (True vs False), structure type used for partial correlation (glycan vs motif), IgG isoform (1, 2, or 4), and reaction type (B4GALT or ST6GALT1; manually annotated). A one-sided t-test was used to determine if motif abundance calculated partial correlations were higher than those calculated from glycan abundance in either previously known or unknown reactions.

### Substructure decomposition of published mucin-type O-glycans to clarify tumor-specific glycan epitopes

We re-analyzed structural mucin-type O-glycan abundance (Table 1.5)(Jin et al. 2017). Mucin-type O-glycans were originally measured by Liquid Chromatography and Mass Spectrometry (LC-MS), structures were manually annotated using empirical masses from Unicarb-DB(Campbell, Nguyen-Khuong, et al. 2014). Pre-processing of these data was restricted to reformatting for input into Glycompare-compatible abundance matrix and structure annotation. Formatted data were normalized using probabilistic quotient normalization(Benedetti et al. 2020). Substructure abundances and motif extraction were performed using a monosaccharide core for thereby focusing analysis on epitope motifs.

Using the mucin-type O-glycan data, we examined both the original glycan abundance data and the motif-level abundance decomposition. Glycan and motif structure abundance was compared across cancer and non-cancer samples using two-sample t-tests; p-values were multiple-test corrected using False Discovery Rate(Benjamini and Hochberg 1995).

**Table 1.5 Glossary of analyses terms.**

Substructures types examined in each section.

| Results section | Substructure | Explanation |
|---|---|---|
| GlyCompare decomposes glycoprofiles to facilitate glycoprofile comparison | Glyco-motif | EPO clustering was done with glyco-motif abundance |
| GlyCompare decomposes glycoprofiles to facilitate glycoprofile comparison | All | Overview of methods discusses every substructure type. |
| GlyCompare accurately clusters glycoengineered EPO samples | Glyco-motif | EPO clustering was done with glyco-motif abundance |
| GlyCompare summarizes structural change across glycoprofiles | Representative substructure | EPO clusters were examined for enrichment and depletion of representative substructures |
| GlyCompare reveals phenotype-associated substructures and trends invisible at the whole glycan level | Substructure | All HMO substructures were used to avoid merging substructure matching known HMOs. This was necessary to allow comparison to know structures |
| GlyCompare identifies condition-specific synthesis dynamics | Substructure | All HMO substructures were used to avoid merging substructure matching known HMOs. This was necessary to allow comparison to know structures |
| GlyCompare increases statistical power of glycomics data | Glyco-motif | Just HMO glyco-motifs were used to avoid artificially overpowering the analysis. |

**Substructure decomposition of ganglioside glycolipids to compare abundance across tissues**

We re-anaylized structural ganglioside glycolipid abundance (Sibille et al. 2016). Published abundance was pooled (summation) within ceramide types, from mouse eye, brain and blood. Glycosides abundance was originally measured by Hydrophilic Interaction Liquid Chromatography stratified Mass Spectrometry (HILC-MS), and HPLC with glycoside standards for structural identification. Pre-processing of these data was restricted to reformatting for input into Glycompare-compatible abundance matrix and structure annotation. Formatted data were normalized using probabilistic quotient normalization(Benedetti et al. 2020). Substructure abundances and motif extraction were performed using a lactose core thereby focusing analysis on biosynthetic motifs.

We examined abundance from two gangliosides (GD3 and GM2) and their corresponding lactose-based substructure abundance. Ceramide groupings include more than 42 or fewer than 35 Carbons ($C_{>42}$, $C_{<35}$), either 1 or 2 unsaturated bonds (1 unsat., 2 unsat), or groups of specific ceramides with X:Y carbons and unsaturated bonds (e.g. 34:1, (36:1+38:1), or (40:1+40:2). Due to limited sample size, trends rather than formal statistics were used to compare abundance.

**Substructure decomposition of site-specific N-glycan compositions to enrich correlation structure**

We re-anaylized compositional site-specific N-glycan abundance (Table 1.5)(Riley et al. 2019). Intact site-specific N-glycan composition was measured using Activated-ion electron transfer dissociation (AI-ETD), the log of localized spectra count for each site-specific composition was used to represent abundance. Pre-processing of these data was restricted to reformatting for input into Glycompare-compatible abundance matrix and structure annotation. Formatted data were normalized using probabilistic quotient normalization(Benedetti et al. 2020). Substructure abundances and motif extraction were performed using compositional monosaccharides thereby focusing analysis on epitope motifs.

Examining site-specific N-glycan compositional data from rat brain, we used a slightly modified method to compute compositional substructure abundance from compositional abundance. To calculate compositional substructure, we sum over larger and subsuming structures in a compositional network. Consider the compositional abundance of a structure: HexNac(p)Hex(q)Fuc(r). Instead of abundance of HexNAc=p, Hex=q, and Fuc=r, we examine the compositional abundance for all measurements where HexNAc>=p, Hex>=q, and Fuc>=r. The network structure can be constrained to provide additional insight (e.g. Glyconnect Compozitor(Robin, Mariethoz, and Lisacek 2020)), currently, the aggregation criteria remains simple. In analyzing these data, we explored trends in correlation between observed compositional vs compositional-substructure abundance.

## Acknowledgments

The authors thank Joshua Klein for guidance on using glypy and Dr. Philip Gordts for insightful comments. We also thank Elisa Benedetti for helping reproduce and extend her prior work. This work was conducted with support from the Novo Nordisk Foundation provided to the Technical University of Denmark (NNF10CC1016517, NNF20SA0066621: N.E.L.), NIGMS (R35 GM119850: N.E.L.), NICHD (R21 HD080682: L.B.), and USDA (USDA/ARS 6250-6001; M.W.H). This work is a publication of the U.S. Department of Agriculture/Agricultural Research Service, Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine, Houston, Texas. The contents of this publication do not necessarily reflect the views or policies of the U.S. Department of Agriculture, nor does mention of trade names, commercial products, or organizations imply an endorsement from the U.S. government.

Chapter 1, in full, is reprint of the material as it appears in Nat Communication. 2021 Bao, Bokan, Benjamin P Kellman, Austin WT Chiang, Yujie Zhang, James T Sorrentino, Austin K York, Mahmoud A Mohammad, Morey W Haymond, Lars Bode, Nathan E Lewis. Correcting for sparsity and interdependence in glycomics by accounting for glycan biosynthesis. Nat Commun. 2021;12(1):4988. The dissertation/thesis author was the primary investigator and author of the paper.

## Data Availability

The EPO N-glycan, IgG, glycolipid, mucin, and site-specific N-glycan abundance data reformatted and re-analyized for this study as well as the HMO abundance data generated in this study have been deposited in the Zenodo database under accession code doi.org/10.5281/zenodo.5083029. The data supporting this work is made available under a CC-BY 4.0 licence.

## Code Availability

We provide the Glycompare python library (v1.1.3) described in this work and example code used to perform analysis and generate figures are available through Github: doi.org/10.5281/zenodo.5083029. In addition to the Glycompare python library, we provide jupyter notebooks to generate our figures and analysis. Finally, we give a dockerized environment that supports Glycompare and all EPO and HMO analyses in the manuscript: doi.org/10.24433/CO.9148600.v1. The glycompare python package and examples are made available under an MIT licence.

## References

Alderete, Tanya L., Chloe Autran, Benjamin E. Brekke, Rob Knight, Lars Bode, Michael I. Goran, and David A. Fields. 2015. "Associations between Human Milk Oligosaccharides and Infant Body Composition in the First 6 Mo of Life." *The American Journal of Clinical Nutrition* 102 (6): 1381–88.

Alocci, Davide, Marie Ghraichy, Elena Barletta, Alessandra Gastaldello, Julien Mariethoz, and Frederique Lisacek. 2018. "Understanding the Glycome: An Interactive View of Glycosylation from Glycocompositions to Glycoepitopes." *Glycobiology* 28 (6): 349–62.

Angel, Peggi M., Anand Mehta, Kim Norris-Caneda, and Richard R. Drake. 2017. "MALDI Imaging Mass Spectrometry of N-Glycans and Tryptic Peptides from the Same Formalin-Fixed, Paraffin-Embedded Tissue Section." *Methods in Molecular Biology*. https://doi.org/10.1007/7651_2017_81.

Aoki-Kinoshita, Kiyoko, Sanjay Agravat, Nobuyuki P. Aoki, Sena Arpinar, Richard D. Cummings, Akihiro Fujita, Noriaki Fujita, et al. 2015. "GlyTouCan 1.0--The International Glycan Structure Repository." *Nucleic Acids Research* 44 (D1): D1237–42.

Apweiler, R., H. Hermjakob, and N. Sharon. 1999. "On the Frequency of Protein Glycosylation, as Deduced from Analysis of the SWISS-PROT Database." *Biochimica et Biophysica Acta* 1473 (1): 4–8.

Ashwood, Christopher, Brian Pratt, Brendan X. MacLean, Rebekah L. Gundry, and Nicolle H. Packer. 2019. "Standardization of PGC-LC-MS-Based Glycomics for Sample Specific Glycotyping." *The Analyst* 144 (11): 3601–12.

Azad, Meghan B., Bianca Robertson, Faisal Atakora, Allan B. Becker, Padmaja Subbarao, Theo J. Moraes, Piushkumar J. Mandhane, et al. 2018. "Human Milk Oligosaccharide Concentrations Are Associated with Multiple Fixed and Modifiable Maternal Characteristics, Environmental Factors, and Feeding Practices." *The Journal of Nutrition* 148 (11): 1733–42.

Benedetti, Elisa, Nathalie Gerstner, Maja Pučić-Baković, Toma Keser, Karli R. Reiding, L. Renee Ruhaak, Tamara Štambuk, et al. 2020. "Systematic Evaluation of Normalization Methods for Glycomics Data Based on Performance of Network Inference." *Metabolites* 10 (7).

https://doi.org/10.3390/metabo10070271.

Benedetti, Elisa, Maja Pučić-Baković, Toma Keser, Annika Wahl, Antti Hassinen, Jeong-Yeh Yang, Lin Liu, et al. 2017. “Network Inference from Glycoproteomics Data Reveals New Reactions in the IgG Glycosylation Pathway.” *Nature Communications* 8 (1): 1483.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society* 57 (1): 289–300.

Black, Alyson P., Hongyan Liang, Connor A. West, Mengjun Wang, Harmin P. Herrera, Brian B. Haab, Peggi M. Angel, Richard R. Drake, and Anand S. Mehta. 2019. “A Novel Mass Spectrometry Platform for Multiplexed N-Glycoprotein Biomarker Discovery from Patient Biofluids by Antibody Panel Based N-Glycan Imaging.” *Analytical Chemistry*. https://doi.org/10.1021/acs.analchem.9b01445.

Bode, Lars, Louise Kuhn, Hae-Young Kim, Lauren Hsiao, Caroline Nissan, Moses Sinkala, Chipepo Kankasa, Mwiya Mwiya, Donald M. Thea, and Grace M. Aldrovandi. 2012. “Human Milk Oligosaccharide Concentration and Risk of Postnatal Transmission of HIV through Breastfeeding.” *The American Journal of Clinical Nutrition* 96 (4): 831–39.

Bojar, Daniel, Rani K. Powers, Diogo M. Camacho, and James J. Collins. 2021. “Deep-Learning Resources for Studying Glycan-Mediated Host-Microbe Interactions.” *Cell Host & Microbe* 29 (1): 132–44.e3.

Campbell, Matthew P., Terry Nguyen-Khuong, Catherine A. Hayes, Sarah A. Flowers, Kathirvel Alagesan, Daniel Kolarich, Nicolle H. Packer, and Niclas G. Karlsson. 2014. “Validation of the Curation Pipeline of UniCarb-DB: Building a Global Glycan Reference MS/MS Repository.” *Biochimica et Biophysica Acta* 1844 (1 Pt A): 108–16.

Campbell, Matthew P., Robyn Peterson, Julien Mariethoz, Elisabeth Gasteiger, Yukie Akune, Kiyoko F. Aoki-Kinoshita, Frederique Lisacek, and Nicolle H. Packer. 2014. “UniCarbKB: Building a Knowledge Platform for Glycoproteomics.” *Nucleic Acids Research* 42 (Database issue): D215–21.

Čaval, Tomislav, Weihua Tian, Zhang Yang, Henrik Clausen, and Albert J. R. Heck. 2018. “Direct Quality Control of Glycoengineered Erythropoietin Variants.” *Nature Communications* 9 (1): 3342.

Cummings, Richard D. 2009. “The Repertoire of Glycan Determinants in the Human Glycome.” *Molecular* *bioSystems* 5 (10): 1087–1104.

Doherty, Margaret, Evropi Theodoratou, Ian Walsh, Barbara Adamczyk, Henning Stöckmann, Felix Agakov, Maria Timofeeva, et al. 2018. “Plasma N-Glycans in Colorectal Cancer Risk.” *Scientific Reports*. https://doi.org/10.1038/s41598-018-26805-7.

Gabius, Hans-Joachim, Sabine André, Herbert Kaltner, and Hans-Christian Siebert. 2002. “The Sugar Code: Functional Lectinomics.” *Biochimica et Biophysica Acta (BBA) - General Subjects* 1572 (2): 165–77.

Greco, Giorgia, Darren Barrington-Light, Remco Swart, and U. K. Altrincham. n.d. “How to Realize LC-MS Quantitation with Chromeleon 7.2 CDS.” https://assets.thermofisher.cn/TFS-Assets/CMD/Technical-Notes/tn-167-cds-lc-ms-quantitation-tn71738-en.pdf.

Gutierrez, Jahir M., Amir Feizi, Shangzhong Li, Thomas B. Kallehauge, Hooman Hefzi, Lise M. Grav, Daniel Ley, et al. 2018. “Genome-Scale Reconstructions of the Mammalian Secretory Pathway Predict Metabolic Costs and Limitations of Protein Secretion.” *bioRxiv*. https://doi.org/10.1101/351387.

Halekoh, Ulrich, Søren Højsgaard, Jun Yan, and Others. 2006. “The R Package Geepack for Generalized Estimating Equations.” *Journal of Statistical Software* 15 (2): 1–11.

Holst, Stephanie, Anna J. M. Deuss, Gabi W. van Pelt, Sandra J. van Vliet, Juan J. Garcia-Vallejo, Carolien A. M. Koeleman, André M. Deelder, et al. 2016. “N-Glycosylation Profiling of Colorectal Cancer Cell Lines Reveals Association of Fucosylation with Differentiation and Caudal Type Homebox 1 (CDX1)/Villin mRNA Expression.” *Molecular & Cellular Proteomics: MCP* 15 (1): 124–40.

Holst, Stephanie, Gabi W. van Pelt, Wilma E. Mesker, Rob A. Tollenaar, Ana I. Belo, Irma van Die, Yoann Rombouts, and Manfred Wuhrer. 2017. “High-Throughput and High-Sensitivity Mass Spectrometry-Based N-Glycomics of Mammalian Cells.” *Methods in Molecular Biology*. https://doi.org/10.1007/978-1-4939-6493-2_14.

Hosoda, Masae, Yushi Takahashi, Masaaki Shiota, Daisuke Shinmachi, Renji Inomoto, Shinichi Higashimoto, and Kiyoko F. Aoki-Kinoshita. 2018. “MCAW-DB: A Glycan Profile Database Capturing the Ambiguity of Glycan Recognition Patterns.” *Carbohydrate Research* 464 (July): 44–56.

Hou, Wenpin, Yushan Qiu, Nobuyuki Hashimoto, Wai-Ki Ching, and Kiyoko F. Aoki-Kinoshita. 2016. “A

Systematic Framework to Derive N-Glycan Biosynthesis Process and the Automated Construction of Glycosylation Networks." *BMC Bioinformatics* 17 Suppl 7 (July): 240.

Jin, Chunsheng, Diarmuid T. Kenny, Emma C. Skoog, Médea Padra, Barbara Adamczyk, Varvara Vitizeva, Anders Thorell, Vignesh Venkatakrishnan, Sara K. Lindén, and Niclas G. Karlsson. 2017. "Structural Diversity of Human Gastric Mucin Glycans." *Molecular & Cellular Proteomics: MCP* 16 (5): 743–58.

Khoury, George A., Richard C. Baliban, and Christodoulos A. Floudas. 2011. "Proteome-Wide Post-Translational Modification Statistics: Frequency Analysis and Curation of the Swiss-Prot Database." *Scientific Reports* 1 (September). https://doi.org/10.1038/srep00090.

Klein, Joshua, Luis Carvalho, and Joseph Zaia. 2018. "Application of Network Smoothing to Glycan LC-MS Profiling." *Bioinformatics* 34 (20): 3511–18.

Klein, Joshua, and Joseph Zaia. 2019. "Glypy: An Open Source Glycoinformatics Library." *Journal of Proteome Research* 18 (9): 3532–37.

Koda, Y., M. Soejima, Y. Liu, and H. Kimura. 1996. "Molecular Basis for Secretor Type alpha(1,2)-Fucosyltransferase Gene Deficiency in a Japanese Population: A Fusion Gene Generated by Unequal Crossover Responsible for the Enzyme Deficiency." *American Journal of Human Genetics* 59 (2): 343–50.

Krambeck, Frederick J., Sandra V. Bennun, Mikael R. Andersen, and Michael J. Betenbaugh. 2017. "Model-Based Analysis of N-Glycosylation in Chinese Hamster Ovary Cells." *PLOS ONE*. https://doi.org/10.1371/journal.pone.0175376.

Kremkow, Benjamin G., and Kelvin H. Lee. 2018. "Glyco-Mapper: A Chinese Hamster Ovary (CHO) Genome-Specific Glycosylation Prediction Tool." *Metabolic Engineering* 47 (May): 134–42.

Kudo, T., H. Iwasaki, S. Nishihara, N. Shinya, T. Ando, I. Narimatsu, and H. Narimatsu. 1996. "Molecular Genetic Analysis of the Human Lewis Histo-Blood Group System. II. Secretor Gene Inactivation by a Novel Single Missense Mutation A385T in Japanese Nonsecretor Individuals." *The Journal of Biological Chemistry* 271 (16): 9830–37.

Maxwell, Evan, Yan Tan, Yuxiang Tan, Han Hu, Gary Benson, Konstantin Aizikov, Shannon Conley, et al. 2012. "GlycReSoft: A Software Package for Automated Recognition of Glycans from LC/MS Data." *PloS One* 7 (9): e45474.

McGuire, Michelle K., Courtney L. Meehan, Mark A. McGuire, Janet E. Williams, James Foster, Daniel W. Sellen, Elizabeth W. Kamau-Mbuthia, et al. 2017. "What's Normal? Oligosaccharide Concentrations and Profiles in Milk Produced by Healthy Women Vary Geographically." *The American Journal of Clinical Nutrition* 105 (5): 1086–1100.

Mohammad, Mahmoud A., Darryl L. Hadsell, and Morey W. Haymond. 2012. "Gene Regulation of UDP-Galactose Synthesis and Transport: Potential Rate-Limiting Processes in Initiation of Milk Production in Humans." *American Journal of Physiology. Endocrinology and Metabolism* 303 (3): E365–76.

Mohammad, Mahmoud A., and Morey W. Haymond. 2013. "Regulation of Lipid Synthesis Genes and Milk Fat Production in Human Mammary Epithelial Cells during Secretory Activation." *American Journal of Physiology. Endocrinology and Metabolism* 305 (6): E700–716.

Opgen-Rhein, Rainer, Juliane Schaefer, Korbinian Strimmer, and Maintainer Korbinian Strimmer. 2007. "The GeneNet Package." ftp://ftp.uni-bayreuth.de/pub/math/statlib/R/CRAN/doc/packages/GeneNet.pdf.

Rademacher, Christoph, and James C. Paulson. 2012. "Glycan Fingerprints: Calculating Diversity in Glycan Libraries." *ACS Chemical Biology* 7 (5): 829–34.

Reiding, Karli R., Dennis Blank, Dennis M. Kuijper, André M. Deelder, and Manfred Wuhrer. 2014. "High-Throughput Profiling of Protein N-Glycosylation by MALDI-TOF-MS Employing Linkage-Specific Sialic Acid Esterification." *Analytical Chemistry* 86 (12): 5784–93.

Reiding, Karli R., Albert Bondt, René Hennig, Richard A. Gardner, Roisin O'Flaherty, Irena Trbojević-Akmačić, Archana Shubhakar, et al. 2019. "High-Throughput Serum N-Glycomics: Method Comparison and Application to Study Rheumatoid Arthritis and Pregnancy-Associated Changes." *Molecular & Cellular Proteomics: MCP* 18 (1): 3–15.

Riley, Nicholas M., Alexander S. Hebert, Michael S. Westphall, and Joshua J. Coon. 2019. "Capturing Site-

Specific Heterogeneity with Large-Scale N-Glycoproteome Analysis." *Nature Communications* 10 (1): 1311.

Robin, Thibault, Julien Mariethoz, and Frédérique Lisacek. 2020. "Examining and Fine-Tuning the Selection of Glycan Compositions with GlyConnect Compozitor." *Molecular & Cellular Proteomics: MCP* 19 (10): 1602–18.

RodrÍguez, Ernesto, Sjoerd T. T. Schetters, and Yvette van Kooyk. 2018. "The Tumour Glyco-Code as a Novel Immune Checkpoint for Immunotherapy." *Nature Reviews. Immunology* 18 (3): 204–11.

Rosenthal, Robert, and Donald B. Rubin. 1991. "Further Issues in Effect Size Estimation for One-Sample Multiple-Choice-Type Data." *Psychological Bulletin*. https://doi.org/10.1037/0033-2909.109.2.351.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks." *Genome Research* 13 (11): 2498–2504.

Sharapov, Sodbo, Yakov Tsepilov, Lucija Klaric, Massimo Mangino, Gaurav Thareja, Mirna Simurina, Concetta Dagostino, et al. 2018. "Defining the Genetic Control of Human Blood Plasma N-Glycome Using Genome-Wide Association Study." *bioRxiv*. https://doi.org/10.1101/365486.

Sibille, Estelle, Olivier Berdeaux, Lucy Martine, Alain M. Bron, Catherine P. Creuzot-Garcher, Zhiguo He, Gilles Thuret, Lionel Bretillon, and Elodie A. Y. Masson. 2016. "Ganglioside Profiling of the Human Retina: Comparison with Other Ocular Structures, Brain and Plasma Reveals Tissue Specificities." *PloS One* 11 (12): e0168794.

Spahn, Philipp N., and Nathan E. Lewis. 2014. "Systems Glycobiology for Glycoengineering." *Current Opinion in Biotechnology* 30 (December): 218–24.

Viverge, D., L. Grimmonprez, G. Cassanas, L. Bardet, and M. Solere. 1990. "Discriminant Carbohydrate Components of Human Milk according to Donor Secretor Types." *Journal of Pediatric Gastroenterology and Nutrition* 11 (3): 365–70.

Wohlschlager, Therese, Kai Scheffler, Ines C. Forstenlehner, Wolfgang Skala, Stefan Senn, Eugen Damoc, Johann Holzmann, and Christian G. Huber. 2018. "Native Mass Spectrometry Combined with

Enzymatic Dissection Unravels Glycoform Heterogeneity of Biopharmaceuticals." *Nature Communications*. https://doi.org/10.1038/s41467-018-04061-7.

Yang, Zhang, Shengjun Wang, Adnan Halim, Morten Alder Schulz, Morten Frodin, Shamim H. Rahman, Malene B. Vester-Christensen, et al. 2015. "Engineered CHO Cells for Production of Diverse, Homogeneous Glycoproteins." *Nature Biotechnology* 33 (8): 842–44.

Yan, Jun, and Jason Fine. 2004. "Estimating Equations for Association Structures." *Statistics in Medicine* 23 (6): 859–74; discussion 875–77,879–80.

York, William S., Raja Mazumder, Rene Ranzinger, Nathan Edwards, Robel Kahsay, Kiyoko F. Aoki-Kinoshita, Matthew P. Campbell, et al. 2019. "GlyGen: Computational and Informatics Resources for Glycoscience." *Glycobiology*. https://doi.org/10.1093/glycob/cwz080.

Zeger, S. L., and K. Y. Liang. 1986. "Longitudinal Data Analysis for Discrete and Continuous Outcomes." *Biometrics* 42 (1): 121–30.

CHAPTER 2: A Predictive Ensemble Classifier for the Gene Expression Diagnosis of ASD at Ages 1 to 4 Years

**Abstract**

Autism Spectrum Disorder (ASD) diagnosis remains behavior-based and the median age of diagnosis is ~52 months, nearly 5 years after its first-trimester origin. Accurate and clinically-translatable early-age diagnostics do not exist due to ASD genetic and clinical heterogeneity. Here we collected clinical, diagnostic, and leukocyte RNA data from 240 ASD and typically developing (TD) toddlers (175 toddlers for training and 65 for test). To identify gene expression ASD diagnostic classifiers, we developed 42,840 models composed of 3,570 gene expression feature selection sets and 12 classification methods. We found that 742 models had AUC-ROC ≥ 0.8 on both Training and Test sets. Weighted Bayesian model averaging of these 742 models yielded an ensemble classifier model with accurate performance in Training and Test gene expression datasets with ASD diagnostic classification AUC-ROC scores of 85-89% and AUC-PR scores of 84-92%. ASD toddlers with ensemble scores above and below the overall ASD ensemble mean of 0.716 (on a scale of 0 to 1) had similar diagnostic and psychometric scores, but those below this ASD ensemble mean had more prenatal risk events than TD toddlers. Ensemble model feature genes were involved in cell cycle, inflammation/immune response, transcriptional gene regulation, cytokine response, and PI3K-AKT, RAS and Wnt signaling pathways. We additionally collected targeted DNA sequencing smMIPs data on a subset of ASD risk genes from 217 of the 240 ASD and TD toddlers. This DNA sequencing found about the same percentage of SFARI Level 1 and 2 ASD risk gene mutations in TD (12 of 105) as in ASD (13 of 112) toddlers, and classification based only on the presence of mutation in these risk genes performed at a chance level of 49%. By contrast, the leukocyte ensemble gene expression classifier correctly diagnostically classified 88% of TD and ASD toddlers with gene mutations. Our ensemble ASD gene expression classifier is diagnostically predictive and replicable across different toddler ages, races, and ethnicities; out-performs a risk gene mutation classifier; and has potential for clinical translation.

## Introduction

ASD is a prenatal(Courchesne, Gazestani, and Lewis 2020; Courchesne et al. 2019; V. Gazestani et al. 2020; Courchesne et al. 2011; Marchetto et al. 2017; Courchesne and Pierce 2005; Willsey et al. 2013; Courchesne et al. 2007; Stoner et al. 2014; Parikshak et al. 2013; Packer 2016; Kaushik and Zarbalis 2016; Krishnan et al. 2016; Donovan and Basson 2017; Grove et al. 2019; Satterstrom et al. 2020), highly heritable disorder(Bai et al. 2019) that considerably impacts a child's ability to perceive and react to social information(Bal et al. 2019; Bacon et al. 2018, 2019). Despite this prenatal and strongly genetic beginning, robust and replicable early-age biological ASD diagnostic markers useful at the individual level have not been found. Indeed, ASD diagnosis remains behavior-based and the median age of the first diagnosis remains at ~52 months(Autism and Developmental Disabilities Monitoring Network Surveillance Year 2006 Principal Investigators and Centers for Disease Control and Prevention (CDC) 2009; Baio et al. 2018; Christensen et al. 2018; Maenner et al. 2020), which is nearly 5 years after its first trimester origin. The long delay between ASD's prenatal onset and eventual diagnosis is a missed opportunity for treatment. Moreover, the heterogeneity of ASD genetics and clinical characteristics impose barriers to identifying early-age molecular diagnostics that accurately diagnose the majority of those with this heterogeneous disorder(Michael V. Lombardo, Lai, and Baron-Cohen 2019). Thus, there is a need for early-age molecular diagnostics of ASD that robustly surmount this heterogeneity obstacle.

Since ASD's heritability is 81%(Bai et al. 2019), initial attempts have focused on genetics to develop clinically useful biomarkers for precision medicine and causal explanations for ASD pathogenesis. While syndromic risk mutations have been described for >200 genes in ASD(Satterstrom et al. 2020; Feliciano et al. 2019; "Human Gene Module" n.d.), each occurs only rarely in ASD. For 80-90% of patients, such mutations are not found. Thus, an estimated 80% or more of ASD individuals are considered 'idiopathic', wherein little is known about the genes and/or environmental factors causing their disorder. In this idiopathic majority of ASD, the risk is likely associated with many inherited common and rare risk variants in each individual child. Studies of polygenic ASD risk found that the combined effect of genetic risk variants in case-control studies accounts for less than 7.5% of the risk variance(Antaki et al. 2022);

genetic ASD risk scores substantially overlap with controls(Robinson et al. 2016; Clarke et al. 2016; Klei et al. 2021; Aguilar-Lacasaña et al. 2022); and, because of this substantial overlap, polygenic risk scores are not clinically diagnostic or prognostic for individuals, nor are they explanatory for the majority of ASD. Thus, DNA-based mutations or polygenic risk scores may not yet be useful for the many idiopathic ASD subjects at the clinical diagnostic level.

RNA biomarkers have been sought using blood gene expression in >35 ASD studies since 2006(Pramparo, Lombardo, et al. 2015; Pramparo, Pierce, et al. 2015; Ch'ng et al. 2015; Diaz-Beltran, Esteban, and Wall 2016; Tylee et al. 2017; He et al. 2019; Lee et al. 2019; Kong et al. 2012; Gregg et al. 2008; Enstrom et al. 2009; Ansel et al. 2016), but many studies have been underpowered, older-aged, clinically heterogeneous, and/or lacking validation test datasets. Some early genetics researchers rejected blood-based biomarkers believing that ASD-relevant dysregulated gene expression must be restricted to the brain. Recent ASD genetics have reversed this view: The earliest prenatal drivers of deviant ASD development are, in fact, broadly expressed regulatory genes, a large percentage of which are active in non-brain organs and tissues such as blood leukocytes as well as in the prenatal brain(Courchesne, Gazestani, and Lewis 2020; Courchesne et al. 2019; V. Gazestani et al. 2020; Pramparo, Pierce, et al. 2015; Pramparo, Lombardo, et al. 2015; Tylee et al. 2017; Ansel et al. 2016; Hewitson et al. 2021; He et al. 2019). Broadly expressed genes that constitute most ASD risk genes are upregulated in early prenatal life and impact multiple stages of prenatal brain development from 1st and 2nd trimester proliferation and neurogenesis to neurite outgrowth and synaptogenesis in the 3rd trimester. These genes disrupt gene expression in signaling pathways such as PI3K-AKT, RAS-ERK, Wnt and insulin receptor pathways, which further disrupt prenatal functions(Courchesne, Gazestani, and Lewis 2020; Courchesne et al. 2019; V. Gazestani et al. 2020; Pramparo, Pierce, et al. 2015; Pramparo, Lombardo, et al. 2015; Tylee et al. 2017; Ansel et al. 2016; Hewitson et al. 2021; He et al. 2019).

In ASD 1 to 4 year-olds, leukocyte gene expression in these pathways is significantly dysregulated(V. H. Gazestani et al. 2019). The degree of pathway dysregulation was correlated with ASD social symptom severity and were validated in ASD neural progenitors and neurons(V. H. Gazestani et al.

2019). Broadly expressed genes in leukocytes from ASD toddlers are also associated with hypoactive brain responses to language and atypical cortical patterning, dysregulation of ASD and language relevant genes, and poor language outcomes(Michael V. Lombardo et al. 2018; M. V. Lombardo et al. 2021). Thus, leukocyte gene expression holds the potential for the objective identification of molecular subtypes of ASD. In analyses of leukocyte gene co-expression, ASD-associated module eigengene values were significantly correlated with abnormal early brain growth and enriched in genes related to cell cycle, translation, and immune networks and pathways. These gene sets are very accurate classifiers of ASD vs. typically developing toddlers (TD) (Pramparo, Pierce, et al. 2015). Studies and reviews of the ASD blood gene expression literature (Pramparo, Lombardo, et al. 2015; Pramparo, Pierce, et al. 2015; Ch'ng et al. 2015; Diaz-Beltran, Esteban, and Wall 2016; Tylee et al. 2017; He et al. 2019; Lee et al. 2019; Kong et al. 2012; Gregg et al. 2008; Enstrom et al. 2009; Ansel et al. 2016; V. H. Gazestani et al. 2019) show dysregulated gene expression in a number of pathways and processes, including PI3K-AKT-mTOR, RAS signaling pathways, ribosomal translation signal, cell cycle, neurogenesis, gastrointestinal disease, immune/inflammation, interferon signaling, and the KEGG natural killer cytotoxicity pathway.

Leukocyte gene expression offers a non-invasive and clinically practicable avenue for understanding aspects of ASD cell biology, including those that could be ASD-relevant, ASD-specific, robust, and ASD-diagnostic or -prognostic. However, for clinical translational potential of leukocyte transcriptomics to lead to robust and rigorous classifiers, high standards for verifying such classifiers should be implemented.

Thus, we developed, operationalized, and tested a rigorous analytic pipeline to identify molecular diagnostic classifiers for ASD using leukocyte gene expression. Using additional clinical data, we verified that our composite gene expression classifier was unbiased against common confounding factors (age, race and ethnicity). Using this platform on leukocyte transcriptomics from male ASD and typically developing (TD) toddlers at ages 1-4 years old, we systematically analyzed the classification performance of 42,840 different models composed of 3,570 different feature selection sets and 12 commonly-used classification methods (Figure 2.1 and Appendix Figure 2.1). Through this, we developed a predictive ensemble diagnostic classifier of male ASD toddlers. Additionally, using targeted DNA sequencing of the coding regions for sets of ASD and neurodevelopmental disorder risk genes using single-molecule molecular inversion probes (smMIPs) (Wang et al. 2020; Stessman et al. 2017), we examined the diagnostic classifier value of presence or absence of a subset of ASD risk gene mutations in our ASD and TD subjects and whether toddlers with ASD risk gene mutations differ in classifier expression from those without such mutations.

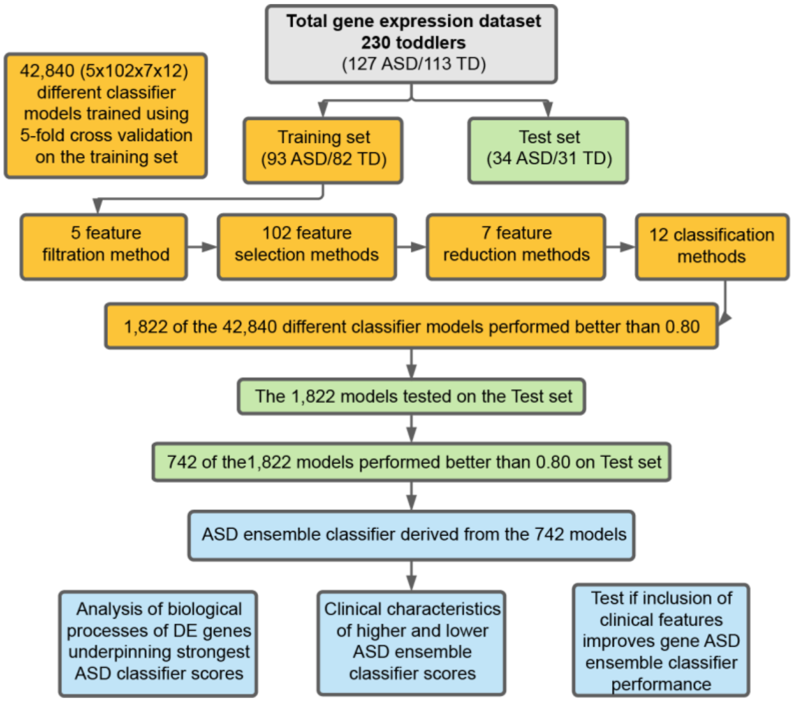

Total gene expression dataset
230 toddlers
(127 ASD/113 TD)
42,840 (5x102x7x12) different classifier models trained using 5-fold cross validation on the training set
Training set
(93 ASD/82 TD)
Test set
(34 ASD/31 TD)
5 feature filtration method
102 feature selection methods
7 feature reduction methods
12 classification methods
1,822 of the 42,840 different classifier models performed better than 0.80
The 1,822 models tested on the Test set
742 of the1,822 models performed better than 0.80 on Test set
ASD ensemble classifier derived from the 742 models
Analysis of biological processes of DE genes underpinning strongest ASD classifier scores
Clinical characteristics of higher and lower ASD ensemble classifier scores
Test if inclusion of clinical features improves gene ASD ensemble classifier performance

**Figure 2.1 Overview of the analysis platform.** The total gene expression dataset was split into a Training set with 175 subjects and a Test set with 65 subjects. Our platform tested 42,840 different models, with each model a combination of 1 feature filtration method, 1 feature selection method, 1 feature reduction method and 1 classification method (total different combinations = 5 x 102 x 7 x 12 = 42,840 models). Models processed the input datasets and returned classification scores. 742 models had classification scores $\geq$0.8 AUC-ROC in both Training and Test sets were used to build the final ensemble classifier model.

## Methods

### Participant recruitment and clinical evaluation

Participants in this study included 240 male toddlers ages 1 to 4 years (Table 2.1). About 70% of toddlers were recruited from the general population using an early screening, detection, and diagnosis strategy called the Get SET Early procedure(Pierce et al. 2021). Using this approach, toddlers who failed a broadband screen, i.e., the CSBS IT Checklist(Wetherby et al. 2002), at 12, 18 or 24 month well-baby visits in the general pediatric community settings, were referred to our center for a comprehensive diagnostic and psychometric evaluation. The remaining subjects were obtained by general community referrals and evaluated in the identical way. Median ages were ASD 2.3 years and TD 1.4 years. All toddlers received a battery of standardized psychometric tests by experienced Ph.D.-level psychologists, including the Autism Diagnostic Observation Schedule (ADOS; Module T, 1 or 2)(Lord 2012), the Mullen Scales of Early Learning(Mullen n.d.), and the Vineland Adaptive Behavior Scales(Sparrow, Balla, and Cicchetti, n.d.). Testing sessions routinely lasted 4 hours in one day or occurred across 2 separate days. Toddlers younger than 30 months upon initial clinical evaluation were followed longitudinally approximately every 9-12 months until final confirmation diagnosis at ages 2 to 4 years; Table 2.1 shows demographic and subject characteristics at final confirmation ages. 127 toddlers were diagnosed ASD, and 113 were TD. Research procedures were approved by the Institutional Review Board of the University of California, San Diego. Parents of subjects underwent Informed Consent Procedures with a psychologist or study coordinator at the time of their child’s enrollment.

**Table 2.1 Subjects demographics.**

ADOS, Autism Diagnostic Observation Schedule; ASD, autism spectrum disorder; CoSo, Communication Social Score; M/F, male/female; RRB, Restricted and Repetitive Behavior; SA, Social Affect.

| | Training set | | | Test set | | |
|---|---|---|---|---|---|---|
| | ASD | TD | ASD vs TD p Value | ASD | TD | ASD vs TD p Value |
| Number of Subjects | 93 | 82 | 0.406 | 34 | 31 | 0.71 |
| Age of last Visit in months | 50.8 ± 28.8 | 34.5 ± 8.4 | <0.001 | 47.4 ± 28.0 | 39.2 ± 15.3 | 0.144 |
| Mullen Scales of Early Learning | | | | | | |
| Visual Reception | 40.1 ± 13.3 | 60.3 ± 10.9 | <0.001 | 38.8 ± 15.4 | 55.1 ± 9.2 | <0.001 |
| Fine Motor | 34.8 ± 11.4 | 54.4 ± 9.5 | <0.001 | 36.9 ± 13.9 | 52.8 ± 8.9 | <0.001 |
| Receptive Language | 32.2 ± 13.1 | 53.3 ± 8.1 | <0.001 | 29.0 ± 16.6 | 52.4 ± 7.6 | <0.001 |
| Expressive Language | 30.7 ± 15.8 | 54.4 ± 9.6 | <0.001 | 28.5 ± 16.8 | 49.8 ± 8.1 | <0.001 |
| Early Learning Composite | 73.6 ± 18.5 | 111.1 ± 13.3 | <0.001 | 71.9 ± 21.0 | 105.0 ± 11.1 | <0.001 |
| Vineland Adaptive Behavior Scales | | | | | | |
| Communication | 82.0 ± 17.5 | 104.9 ± 10.5 | <0.001 | 79.5 ± 17.8 | 100.4 ± 9.6 | <0.001 |
| Daily Living | 83.7 ± 12.8 | 103.0 ± 10.2 | <0.001 | 84.0 ± 13.0 | 99.9 ± 10.6 | <0.001 |
| Socialization | 80.5 ± 13.0 | 106.3 ± 10.9 | <0.001 | 79.1 ± 9.8 | 99.4 ± 11.0 | <0.001 |
| Motor Skills | 87.8 ± 10.9 | 103.1 ± 10.4 | <0.001 | 88.5 ± 10.5 | 98.8 ± 8.5 | <0.001 |
| Adaptive Behavior | 80.8 ± 13.0 | 105.0 ± 9.7 | <0.001 | 79.9 ± 11.6 | 99.2 ± 10.3 | <0.001 |

**Table 2.1 Subjects demographics (continued).**

| | | | | | | |
|---|---|---|---|---|---|---|
| Autism Diagnostic Observation Schedule | | | | | | |
| ADOS SA/CoSo Score | 14.3 ± 3.4 | 2.2 ± 2.0 | <0.001 | 13.3 ± 4.3 | 2.8 ± 1.9 | <0.001 |
| ADOS RRB Score | 3.8 ± 1.5 | 0.3 ± 0.6 | <0.001 | 3.1 ± 1.5 | 0.5 ± 0.6 | <0.001 |
| ADOS Total Score | 18.1 ± 4.1 | 2.4 ± 2.1 | <0.001 | 16.4 ± 4.7 | 3.3 ± 2.2 | <0.001 |

### Targeted sequencing data from ASD and TD subjects

For 112 of the 127 ASD and 105 of the 113 TD study subjects, we also had targeted sequencing data by smMIPs from prior studies aimed at detecting rare severe mutations in autism and neurodevelopmental disorder risk genes; that study was from our Center's collaboration with the Eichler Lab (Wang et al. 2020; Stessman et al. 2017). Two sets of neurodevelopmental disorders and ASD risk genes were used for targeted sequencing. The ASD significant variants in our ASD toddlers had been previously reported, but here we additionally report ASD significant variants in our TD toddlers. More than 87% of the ASD toddlers (83 out of 93 and 29 out of 34 ASDs in the Training and Test datasets, respectively), and 92% of the TD toddlers (74 out of 82 and all 31 TDs in the Training and Test datasets, respectively) were tested for mutations. Rare (MAF < 0.01%) severe missense mutation with a combined annotation-dependent depletion (CADD) score ≥30 (MIS30) and likely gene-disruptive (LGD, including splicing donor or acceptor, frameshift, and stop-gained) mutations were considered for further analysis. Among the 105 TD toddlers, 12 had SFARI Level 1 or 2 ASD risk gene mutations and among the 112 ASD toddlers, 13 had such mutations. One of these ASD had two ASD risk gene mutations. Thus, among the 217 subjects, a total of 25 subjects carried ASD risk gene mutations (26 genes). The two-sided independent T-test was performed to test the ensemble score distribution difference between subjects with or without mutations.

### Blood sample collection for gene expression analyses

Blood samples were collected from each subject during clinical evaluation visits. To monitor health status, the temperature of each toddler was taken using an ear digital thermometer immediately preceding the blood draw. When the temperature was higher than 99 Fahrenheit, the blood draw was re-scheduled for a later visit. Moreover, the blood draw was not taken if a toddler had some illness (e.g., cold or flu), as observed by us or stated by parents. We collected four to six milliliters of blood into ethylenediaminetetraacetic-coated tubes from all toddlers. Leukocytes in the blood samples were captured and stabilized by LeukoLOCK filters (Ambion) and were immediately placed in a −20°C freezer. Total RNA was extracted following standard procedures and manufacturer's instructions (Ambion).

### Summary of main steps in design and analyses of the RNA data from the 240 study subjects

Figure 2.1 outlines the main design and analysis steps, and Appendix Figure 2.1 provided details of the feature engineering. The 240 subjects were divided into a Training dataset of 175 subjects and a Test set of 65 subjects. The training dataset was used to build gene expression classifiers and the Test set was held out and later used to test the classifiers. High-performing classifiers evaluated by the Test set were used to build a single, final ensemble classifier, which was a Bayesian averaging model of all top-performing classifiers. The performance of this ensemble classifier was then measured on Training and Test subjects; DE genes underlying its accurate performance were identified and pathway and process enrichment determined; and clinical characteristics across classifier scores were examined. Lastly, post hoc exploratory analyses were performed to test whether including specific social behavioral and prenatal features might improve overall performance.

### Microarray data processing

Gene expression of subject RNA samples was assayed using the Illumina HT-12 platform. Arrays were scanned with the Illumina BeadArray Reader and read into Illumina GenomeStudio software (version 1.1.1). Raw Illumina probe intensities were converted to expression values using the lumi package(Du, Kibbe, and Lin 2008). We employed a three-step procedure to filter for probes with reliable expression levels. First, we only retained probes that met the detection $p < 0.05$ cut-off threshold in at least 3 samples. Second, we required probes to have expression levels above the 95th percentile of negative probes in at least 50% of samples. The probes with detection $p > 0.1$ across all samples were selected as negative probes and their expression levels were pooled together to estimate the 95th percentile expression level. Third, for genes represented by multiple probes, we considered the probe with the highest mean expression level across our dataset, after quantile normalization of the data. These criteria led to the selection of 14,312 coding genes as expressed in our leukocyte transcriptome data, which highly overlaps with the reported estimate of 14,555 protein-coding genes (chosen based on unique Entrez gene IDs) for whole blood by the GTEx consortium(Ardlie et al. 2015).

**Building the classifier platform on the training dataset**

The pipeline ran five-fold cross-validations. At the beginning of each iteration, the pipeline held out 20% of samples and used the remaining 80% of samples for hyper-parameter selection, feature selection, and classifier training. In the first step (Appendix Figure 2.1), feature filtration, five methods were used, including no (no action), cov (remove 50% of features with the smaller coefficient of variation), var (remove 50% of features with smaller variance), cov_var (remove 50% of features with the smaller coefficient of variation and then remove 50% of features with smaller variance in the rest), varImportance (keep only the 25% of features with the highest variance).

The second step, feature selection, included 102 methods, which were composed of seven groups; although conceptually similar, each using different approaches. These seven groups are no (no action), grn(Meyer, Lafitte, and Bontempi 2008) (genetic regulatory network), z-score, selectV(Antonio Pedro Duarte Silva <psilva@porto.ucp.pt> 2015), svm("penalizedSVM: Feature Selection SVM Using Penalty Functions" n.d.), GSEA(Subramanian et al. 2005), DE-analysis(Ritchie et al. 2015) (see Appendix Method 1).

The third step was feature reduction. Seven methods were used: no (no feature reduction), WGCNA(Langfelder and Horvath 2008), logisticFwd, SIS(Saldana and Feng 2018), principal component regression (PCR)(Mevik and Wehrens 2015), partial least squares regression (PLSR)(Wehrens and Mevik 2007), canonical powered partial least squares (CPPLS)(Wehrens and Mevik 2007) (see Appendix Method 1). After three steps, up to 1320 gene routes were created that can be used in the classification step.

The classification step exploited 12 classifiers, including reg (linear model), logReg(Ripley 2002) (logistic regression), lda(Ripley 2002) (Linear Discriminant Analysis), qda(Ripley 2002) (Quadratic Discriminant Analysis), ridgeReg(Friedman, Hastie, and Tibshirani 2010) (GLM with ridge regularization), lassoReg(Friedman, Hastie, and Tibshirani 2010) (GLM with lasso regularization), ridgeLogReg(Friedman, Hastie, and Tibshirani 2010) (logistic regression with ridge regularization), lassoLogReg(Friedman, Hastie, and Tibshirani 2010) (logistic regression with lasso regularization), elasticNetLogReg(Friedman, Hastie, and Tibshirani 2010) (logistic regression with elastic net regularization), boosting(Ridgeway 2007) (Generalized Boosted Regression Modeling with Bernoulli distribution), randomForest(Liaw, Wiener, and Others 2002) (random forest) and bagging(Liaw, Wiener, and Others 2002) (random forests with bagging to reduce the complexity). After training a classifier, the diagnostic ability was evaluated by AUC-PR (precision-recall) curve and AUC-ROC (Receiver operating characteristic) curve(Grau, Grosse, and Keilwagen 2015; Robin et al. 2011).

For every possible combination of the 5 feature filtration, 102 feature selection, 7 feature reduction, and 12 classification routes, we made a total of 42,840 different classifier models.

### Label permutation data

To generate the randomized background, we shuffled the diagnostic label of the Training dataset and randomly separated the data into training/validation segments (85%/15%). Then we performed the 5-fold cross-validation on the permuted dataset.

### Bayesian model averaging to create a single transcriptomic ensemble classifier

The training models that had 0.80 or higher AUC-ROC scores were tested on the Test dataset. Then, the models that had an AUC-ROC ≥0.80 were used with Bayesian Model Averaging (BMA) to create a single ensemble classifier. The ensemble score was the sum of weighted predictions of selected models. The weight was the mathematical average of the square of (AUC-ROC value minus 0.7). In a model selection, we used training data D to select a good model M (according to a score) to predict a targeted outcome T of interest based on patient features X, namely, P(T | X, M). BMA was based on the notion of averaging over a set of possible models and weighting the prediction of each model according to its probability given training data D, as shown in equations.

- $\mathrm{p(T|X)} = \sum_{\mathrm{m_i}} \mathrm{p(M_i|X)p(T|X, M_i)}$

M is the model, T is the prediction and X is the data.

- $\mathrm{p(M_i|X)} = \frac{\mathrm{AUC\_ROC_i - 0.7}}{\sum_{\mathrm{j}}\mathrm{(AUC\_ROC_i - 0.7)}}$

The ensemble scores of the independent dataset are calculated based on the same model built. The scores are then rescaled to 0 and 1.

- $\mathrm{ensembleScore_i} = \frac{\mathrm{ensembleScore_i - min(ensembleScore)}}{\mathrm{max(ensembleScore) - min(ensembleScore)}}$

**Adjust the averaging weights using the adapted Thresholdout algorithm**

In addition to using AUC-ROC score as weights in BMA method for the ensemble model, an adapted Thresholdout weight(Dwork et al. 2015) is implemented to mitigate the overfitting issue for reusing the holdout Test dataset. Since we used 5-fold cross-validation to evaluate each model *i* on the Training set, we calculated the mean of the AUC-ROC score and related standard deviation. We calculated the mean of the standard deviation for all methods which is 0.041. Thus, we set T=0.041, $\zeta$ = 0.041.

```
$\mathrm{trainScore_i}$ = mean(trainAUC-ROC$_\mathrm{i}$)
$\mathrm{testScore_i}$ = testAUC-ROC$_\mathrm{i}$
If abs(testScore$_\mathrm{i}$-trainScore$_\mathrm{i}$)<=T
    adjustedWeights$_\mathrm{i}$ = trainScore$_\mathrm{i}$
If abs(testScore$_\mathrm{i}$-trainScore$_\mathrm{i}$)>=T
    adjustedWeights$_\mathrm{i}$ = testScore$_\mathrm{i}$ + laplacian(0, $\zeta$ )
adjustedWeights$_\mathrm{i}$ = pmax(0,pmin(1, adjustedWeights$_\mathrm{i}$))
```

We sampled the adjusted weights 1000 times and calculated the mean final AUC-ROC scores for the ensemble models.

**Biological processes enriched by differentially expressed genes**

We additionally conducted differential expression (DE) analysis on ASD subjects with ensemble scores below-the-mean vs. all TD subjects. The Limma package(Smyth 2005; Ritchie et al. 2015) was then applied on quantile-normalized data for differential expression analysis in which moderated t-statistics were calculated by robust empirical Bayes methods. We used adjusted p < 0.01 (Benjamin–Hochberg) and log Fold Change > 0.1 to select genes and generate the volcano plot. The Gene Ontology (GO) enrichment was conducted using g:Profiler(Raudvere et al. 2019) (https://biit.cs.ut.ee/gprofiler/gost) with 12695 protein-coding genes (12695/14132 gene features) as background (g:Profiler, advanced option/statistical domain scope: Custom; custom over annotated genes). We only checked the "GO biological process" and KEGG terms of size 15-1500 in the biological process. The threshold was "Significance threshold: B-H FDR < 0.1". Then the terms were clustered with REVIGO(Supek et al. 2011), ordered with p (http://revigo.irb.hr/). The connections across terms were visualized by the Cytoscape 3.8.2(Su et al. 2014).

**Post-hoc analysis on common confounding factors**

The post-hoc analysis further verified that the classifier scores were stable across different age groups. The optbin R package was used to determine optimal age breakpoints for ASD and TD groups; age bins were [0,20], [20,31], and [31,49]. Games-Howell test("Rstatix" n.d.) was performed to compare the classifier score between TD or ASD groups in each of the three age bins (FDR adjusted p-value <0.05).

The one-way ANOVA test(Chambers and Hastie, n.d.) was conducted to test if statistically significant differences existed across three ethnicities and seven race groups for ASD subjects. For ethnicity, toddlers from ASD and TD were labeled as 'Hispanic or Latino', 'Not Hispanic or Latino' and 'Unknown'. For races, toddlers from ASD and TD were labeled as 'Caucasian', 'Caucasian/Asian', 'African American', 'Asian', 'Pacific Islander', 'Other', 'Unknown'.

## Results

### ASD risk gene mutation-based diagnostic classification of ASD vs TD

Targeted sequencing by smMIPs was performed on 217 (112 ASD and 105 TD) out of the 240 (127 ASD and 113 TD) toddlers in this study (see Methods). Analyses found 12 TD toddlers with missense or LGD mutations in SFARI (https://gene.sfari.org/) Level 1 or 2 ASD risk genes including: *ANK3*, *CACNA2D3*, *CLCN4*, *CTTNBP2*, *CUL7*, *DIP2A*, *DLG4*, *HECTD4*, *LRP2*, *LZTR1*, *MYH9*, and *NAV2*. Analyses found 13 ASD toddlers with missense or LGD mutations in SFARI (https://gene.sfari.org/) Level 1 or 2 ASD risk genes: *CACNA2D3*, *CHD2*, *DIP2A*, *DSCAM*, *KATNAL2*, *LRP2*, *MYH9*, *NCKAP1*, *NTNG1*, *PHF2*, *RELN*, *STXBP5*, *UNC80*, and *ZC3H4* (one subject had two mutations). To assess the power of using this mutation information alone in discriminating ASD from TD, we did a classification according to the presence/absence of the missense or LGD mutations in SFARI Level 1 or 2 ASD risk genes. More precisely, ASD toddlers with and without mutations were considered as true positive and false negative, respectively; and TD toddlers with and without mutations were considered as false positive and true negative, respectively. This mutation-based classification performed at a chance level, 49% (50% being chance), with precision (positive predictive value) of 52%, and recall (sensitivity) of 10%. In this mutation-based classification, a small number of TDs were falsely called ASD and a large number of ASD toddlers were falsely called TD.

**Development of a robust transcriptomic classifier platform with diverse feature engineering and classification methods**

Next, we used blood transcriptomic data from the 240 ASD and TD study toddlers to develop a diagnostic classifier. To identify potential transcriptome biomarkers in a Training sample of 175 of the 240 ASD and TD toddlers (Table 2.1), we developed a platform that examined the classification power of the blood transcriptomic data by systematically exploring the performance of 42,840 possible models composed of 3570 different feature selection routes, followed by 12 classification methods (see Methods). The platform started with removing genes with low variation across samples. Next, features that differentiate between ASD and TD subjects at expression or co-expression levels were selected using a suite of 102 feature selection methods. Third, to avoid overfitting, we reduced the number of features by collapsing expression data from the correlating genes. Finally, we trained 12 different classifiers for each selected feature set. To evaluate the performance of each of the 3570 feature selection routes and the 12 classification methods, we iterated the process 5 times while holding out 20% of samples and using the remaining 80% of samples for hyper-parameter selection, feature selection, and classifier training. Thus, each of the 42,840 models started with a “route” that consisted of 1 filtration method, 1 selection method, 1 reduction method, and ended with 1 classification method, and all possible combinations of the 5 filtration, 102 selection, 7 reduction and 12 classification methods were used. The platform reports the average performance of each of the 42,840 models across the 5 held-out folds as measured by area under the receiver operating characteristic curve (AUC-ROC) and area under the precision-recall curve (AUC-PR).

**Diverse pipelines successfully classify ASD vs TD**

Since the feature selection methods depended on the characteristics of training transcriptome datasets, some routes were not able to find qualified features in all five iterations of the validation. Therefore, the platform successfully classified the data in 15,840 out of 42,840 different ways, including 1320 different routes out of 3570 for feature selection and 12 different classification methods (Appendix Figure 2.2). From 15,840 trained models, 1822 (11.5%) models showed classification AUC-ROC > 0.8 with the max AUC-ROC of 0.856. Moreover, 1508 of the 1822 models also exhibited an AUC-PR > 0.8.

These 1822 models performed well due to their feature routes and were robust to variations in the data or the model. For example, we observed a subset of 175 feature routes (colored with a brown band in Figure 2.2a) that performed consistently well across different classifiers with a mean AUC-ROC of 0.81. Additionally, these 1822 high-performing models worked similarly well across all five held-out datasets with a mean range of 0.13 and variance of 0.02 (Appendix Figure 2.3). Furthermore, different models that largely overlapped in their feature selection routes also worked well across different classifier methods (Appendix Figure 2.5).

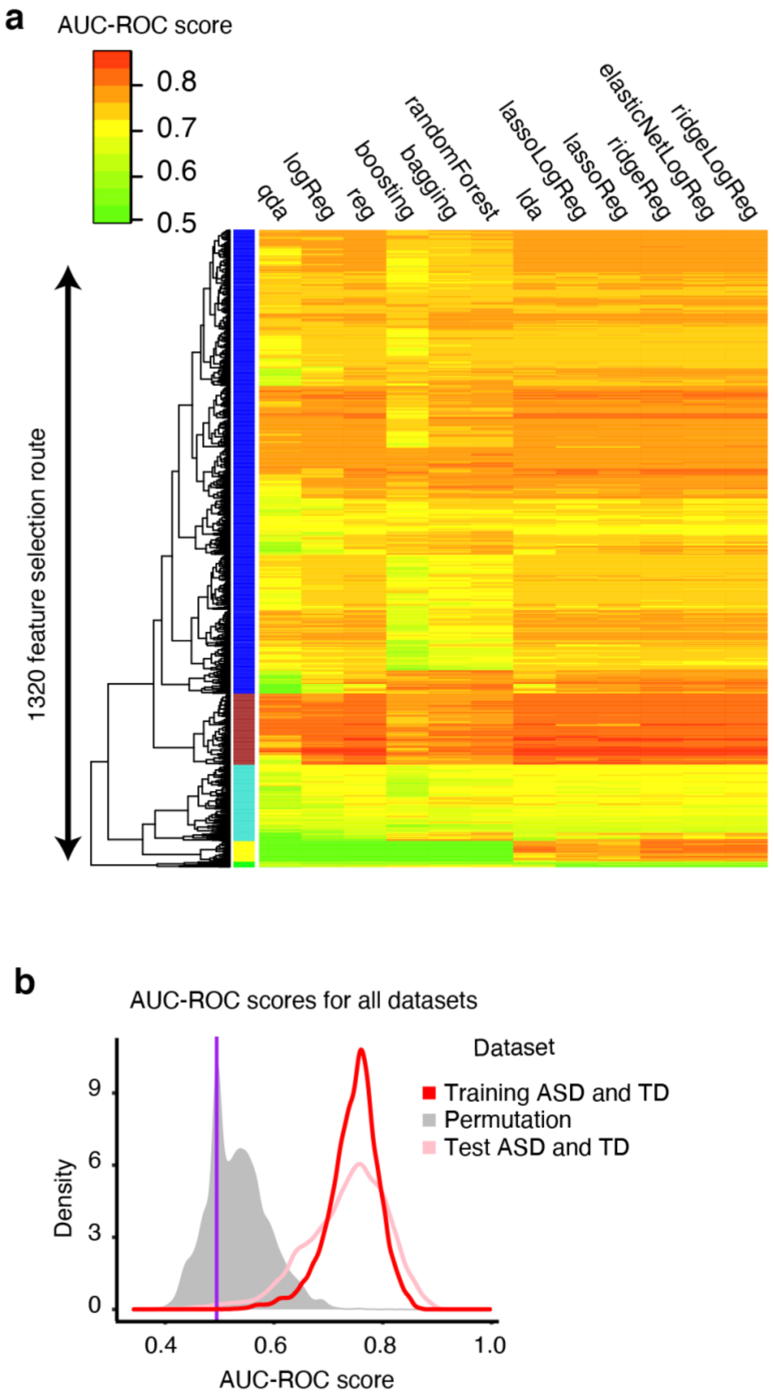


**Figure 2.2 A classification platform was developed to robustly identify the biomarkers for the early diagnosis of ASD.** (**a**) AUC-ROC classifier scores were computed for each of the 42,840 model results from the Training dataset. The AUC-ROC values were based on the average performance of each model across 5 iterations, with 20% of samples being held out each time. (**b**) 1,822 models with AUC-ROC scores ≥0.80 were then tested on the held out Test dataset. Permuting the sample labels (i.e., ASD and TD) further supported the validity of the signal.

To further verify that the performance of these 1822 models was not due to chance alone, we generated five separate randomized datasets by shuffling the sample labels (i.e., ASD or TD) from the Training dataset. We next ran the platform on each of the five datasets independently (see Methods). Importantly, the platform identified zero models out of 1822 with AUC-ROC and AUC-PR > 0.8 across the five datasets, respectively, suggesting that the accurate performance of the 1822 models was not due to chance.

We evaluated the performance of the 1822 high-performing models on the Test dataset of N=65 ASD and TD toddlers. Of the 1822 models with AUC-ROC > 0.8 in the Training dataset, 742 models (40%; Fisher's Exact Test $p < 2.2\text{x}10^{-16}$) also had an AUC-ROC > 0.8 for the Test dataset. These 742 heterogeneous predictive models involved 125 different feature routes and 2,721 gene features (Figure 2.3a, see Appendix Result 1-2).

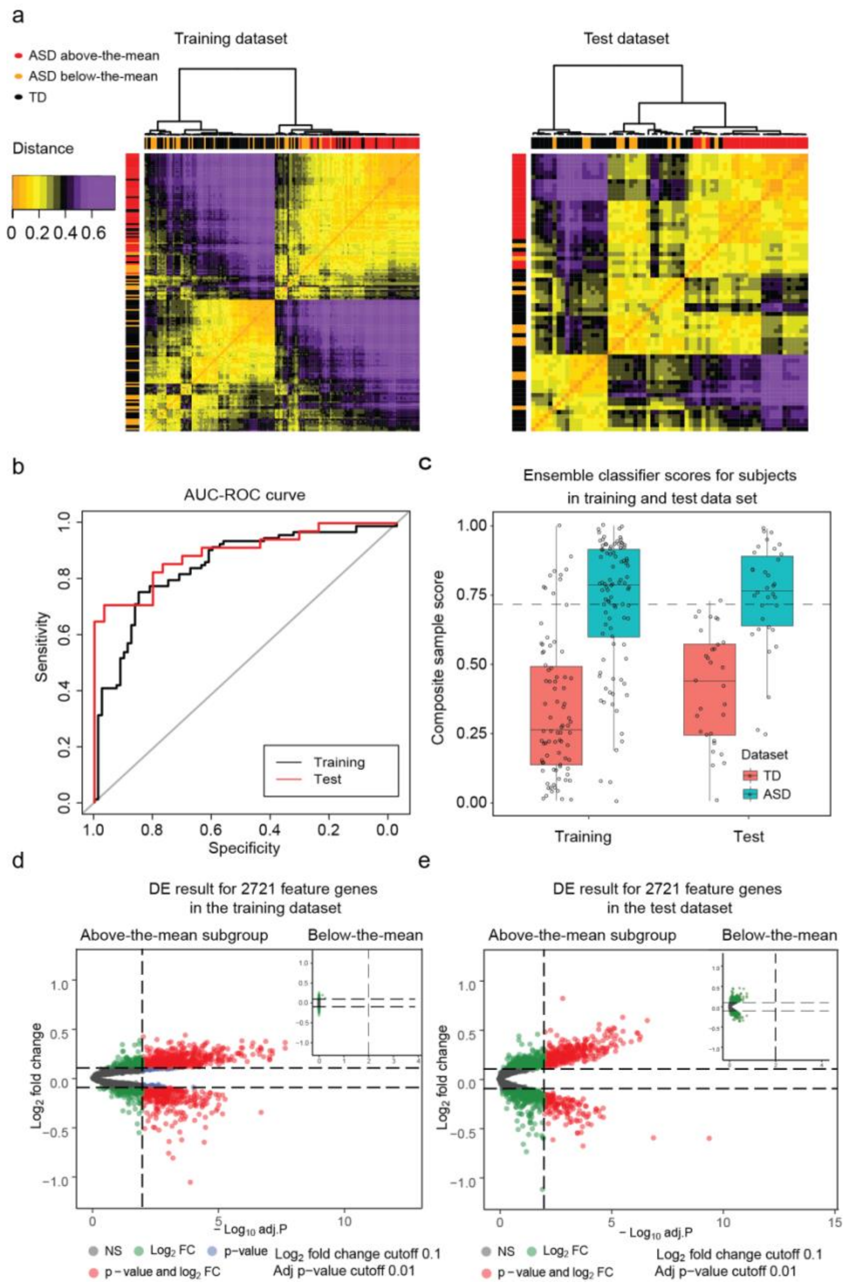


**Figure 2.3 Blood transcriptome ASD subtypes were identified by our classification platform.**
(**a**) The clustering table of subjects in Training and Test dataset based on the 742 models classification score similarity (distance='Euclidean' and method='ward.2d'). ASD and TD subjects showed distinct classification patterns. The red, orange and black bars on the sides represented above-the-mean ASD, below-the-mean ASD and TD subjects, respectively (the mean is the dashed line in Figure 2.3c). The orange and purple colors represented the gradient of dissimilarity between subjects based on their classification scores. (**b**) The AUC-ROC results on the ensemble classification model generated by the Bayesian model averaging approach. (**c)** Ensemble classifier model scores for ASD and TD individuals in Training and Test datasets. The ASD group mean was 0.723 and the TD group mean was 0.359. (**d**) and (**e**) The differential expression analysis of 2721 protein-coding feature genes. The volcano plots showed the adjusted p-value (cutoff=0.01) vs. log fold changes (cutoff=0.1) of genes in the above-the-mean to TD subjects and below-the-mean subjects to TD subjects in the Training dataset and Test dataset.

**Randomized data can be erroneously “classified” at reasonable AUC-ROC levels**

There were 1822 models that reached a high AUC-ROC value in the Training dataset. However, the question remained whether this range was significantly different from the AUC-ROC values that one could obtain from trying to classify subjects after randomizing their final diagnosis. To test this, we permuted the sample labels (i.e., ASD and TD) for all subjects in our Training set and ran the pipeline to test all feature engineering and classification methods. Importantly, we tested all 42,840 candidate models and found the median AUC-ROC score was 0.5101 with the 95th CI (0.42-0.65) on the randomized samples. As expected, only rare chance instances of good “classification” occurred. The fact that chance alone could lead to a rare “good classification” score for a single model, was a cautionary signal that literature reports of unvalidated and unreplicable single high-performance classifiers could be due to chance (see Methods, Figure 2.2b).

**Bayesian model averaging of the 742 predictive models to create a single transcriptomic ensemble classifier**

To build a single composite model that combined the 742 models that had AUC-ROC values of 0.80 on both Training and held-out Test sets, we used Bayesian model averaging (BMA). The ensemble model produced a single composite classification score by calculating weighted predictions from 742 models (see Methods). Scores ranged from 0 to 1 with 0 being the highest certainty in TD status and 1 being the highest certainty in ASD status. With this ensemble model, the AUC-ROC score was 84.67% and 89.18% for Training and Test datasets, respectively (Figure 2.3b) and AUC-PR was 84.33% and 92.11% for the Training and Test datasets, respectively. These values were significantly higher than the naive Random Forest baseline model (see Appendix Result 3) with 72.32% AUC-ROC (ROC.test $p < 10^{-44}$).

Since the model selection method mentioned above required repeated exposure to the hold-out Test test, we leveraged the idea from Thresholdout algorithm(Dwork et al. 2015; “How to Double Dip into Your Holdout Set” 2017) to adjust the weights on the AUC-ROC weighted prediction for 742 models (see Method). With sampled ensemble models, the mean AUC-ROC score was 82.47% and

81.16% for the Training and Test datasets, respectively, and AUC-PR was 83.73% and 82.11% for the Training and Test datasets, respectively. These values were significantly higher than the naive Random Forest baseline model (see Appendix Result 3) with 72.32% AUC-ROC (ROC.test $p < 10^{-44}$).

We calculated the mean of ensemble classification scores for all ASD toddlers in the Training and Test datasets. The overall ASD group median classifier score was 0.781 and the overall TD group median score was 0.303. To test for group differences in scores and possible age effects, we used multiple linear regression. The independent variables were diagnosed group, age and their interaction. The dependent variable was the ensemble classification score. Based on the coefficients in the model, we found a significant effect of group (coefficient, $p = 0.0011$) but non-significant effects of age (coefficient, $p = 0.056$) and group by age interaction (coefficient, group:age, $p = 0.76$).

**Classifier scores not significantly affected by age, ethnicity, race differences**

To further examine possible bias toward the age effects on group classification, we stratified subjects into three age bins ([0,20], [20,31], and [31,49]) and compared the classifier prediction performance on different bins (see Method, Figure 2.4). Game-Howell test("Rstatix" n.d.) showed there was no significant difference between classification scores for TD or ASD groups in each of the three age bins, and classification scores were significantly different only in the ASD vs. TD diagnostic group comparisons (FDR adjusted p-value < 0.05) (Figure 2.4). This further verified that potential confounding effects of age were excluded in the analysis.

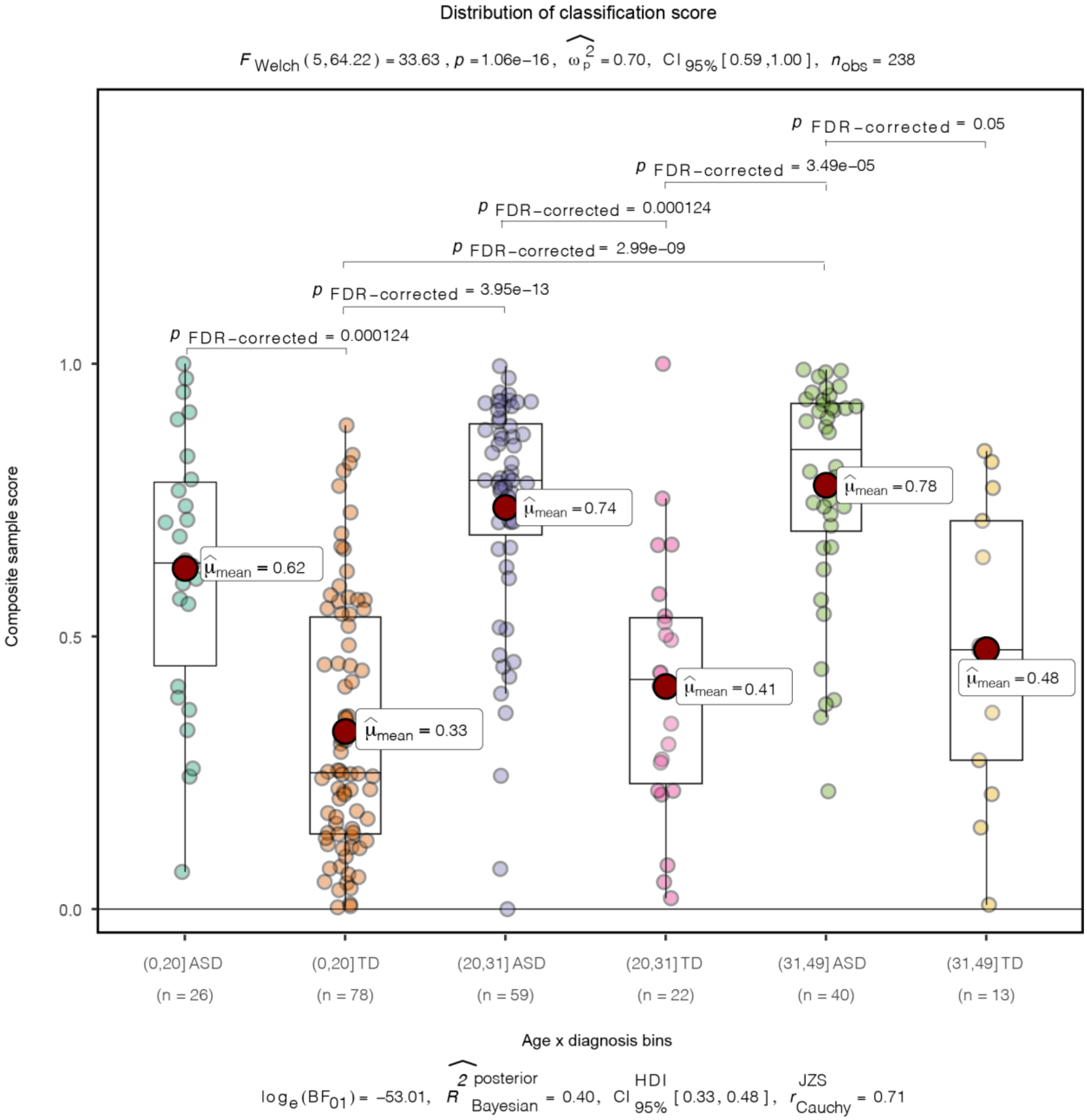


**Figure 2.4 The distribution of classification scores.** 240 subjects were partitioned into 6 groups. Games-Howell tests were performed to compare the group difference and only significant comparisons were shown. The classification scores were significantly different only in the ASD vs TD diagnostic group comparison.

Post-hoc examination of classifier scores in ASD groups showed there was no significant difference across three ethnicity groups (‘Hispanic or Latino’, ‘not Hispanic or Latino’, ‘unknown’; one-way ANOVA, F = 0.899, p = 0.409) (Table 2.2). However, the differences appeared in TD groups (F = 3.7, p = 0.021). The same analysis was also conducted on races. Toddlers were labeled as ‘Caucasian’, ‘Caucasian/Asian’, ‘African American’, ‘Asian’, ‘Pacific Islander’, ‘Other’, ‘Unknown’. No significant difference of means was found across all race groups (One-way ANOVA test, F = 1.151, p = 0.337) (Table 2.3). The differences appeared in the TD group (F = 5.25, p = 9.03e-05) and seemed likely due to the small number of individuals in different race categories. Both ethnicity and race analysis indicated that ASD molecular pathology is being consistently detected by our classifier.

**Table 2.2 Statistics of ASD and TD's classifier score for Hispanic/non-Hispanics.**

| Ethnicity | ASD | | TD | |
|---|---|---|---|---|
| | Size | Mean (Std) | Size | Mean (Std) |
| Not Hispanic Latino | 64 | 0.700 (0.253) | 82 | 0.396 (0.248) |
| Hispanic and Latino | 29 | 0.723 (0.197) | 18 | 0.273 (0.220) |
| Unknown | 34 | 0.765 (0.187) | 13 | 0.245 (0.138) |

**Table 2.3 Statistics of ASD and TD's classifier score for races.**

| Race | ASD | | TD | |
|---|---|---|---|---|
| | Size | Mean (Std) | Size | Mean (Std) |
| Caucasian | 65 | 0.675 (0.253) | 73 | 0.364 (0.240) |
| Unknown | 40 | 0.755 (0.180) | 19 | 0.284 (0.175) |
| Caucasian/Asian | 4 | 0.861 (0.150) | 5 | 0.358 (0.212) |
| African American | 5 | 0.799 (0.221) | 4 | 0.508 (0.288) |
| Asian | 8 | 0.789 (0.212) | 9 | 0.421 (0.314) |
| Pacific Islander | 3 | 0.795 (0.090) | 2 | 0.185 (0.088) |
| Other | 2 | 0.797 (0.005) | 1 | 0.592 (null) |

**Classifier scores not significantly affected by the presence or absence of ASD risk gene mutations**

There was no significant difference in the ensemble classifier scores between ASD toddlers with and without mutations (median = 0.738 vs 0.784, mean = 0.715 vs. 0.724, Welch t-test p = 0.875) (Figure 2.5); 11 of the 13 ASD toddlers with risk mutations were correctly classified by the ensemble model. However, there was a difference in the ensemble classifier scores between TD toddlers with and without ASD risk gene mutations; in fact, TDs with mutations had lower composite scores than the other TDs (median = 0.229 vs 0.340, mean = 0.223 vs. 0.375, Welch t-test p = 0.007) and robustly differed from the ASD composite score, median = 0.303 vs 0.781 (Figure 2.5). Thus, the presence of ASD risk gene mutations conferred no liability on the composite score of TD toddlers, and 11 out of 12 TDs with mutations in risk genes were correctly classified as typical by our gene expression classifier. The ensemble classifier correctly differentially diagnosed 88% of the ASD and TD toddlers with ASD risk gene mutations.

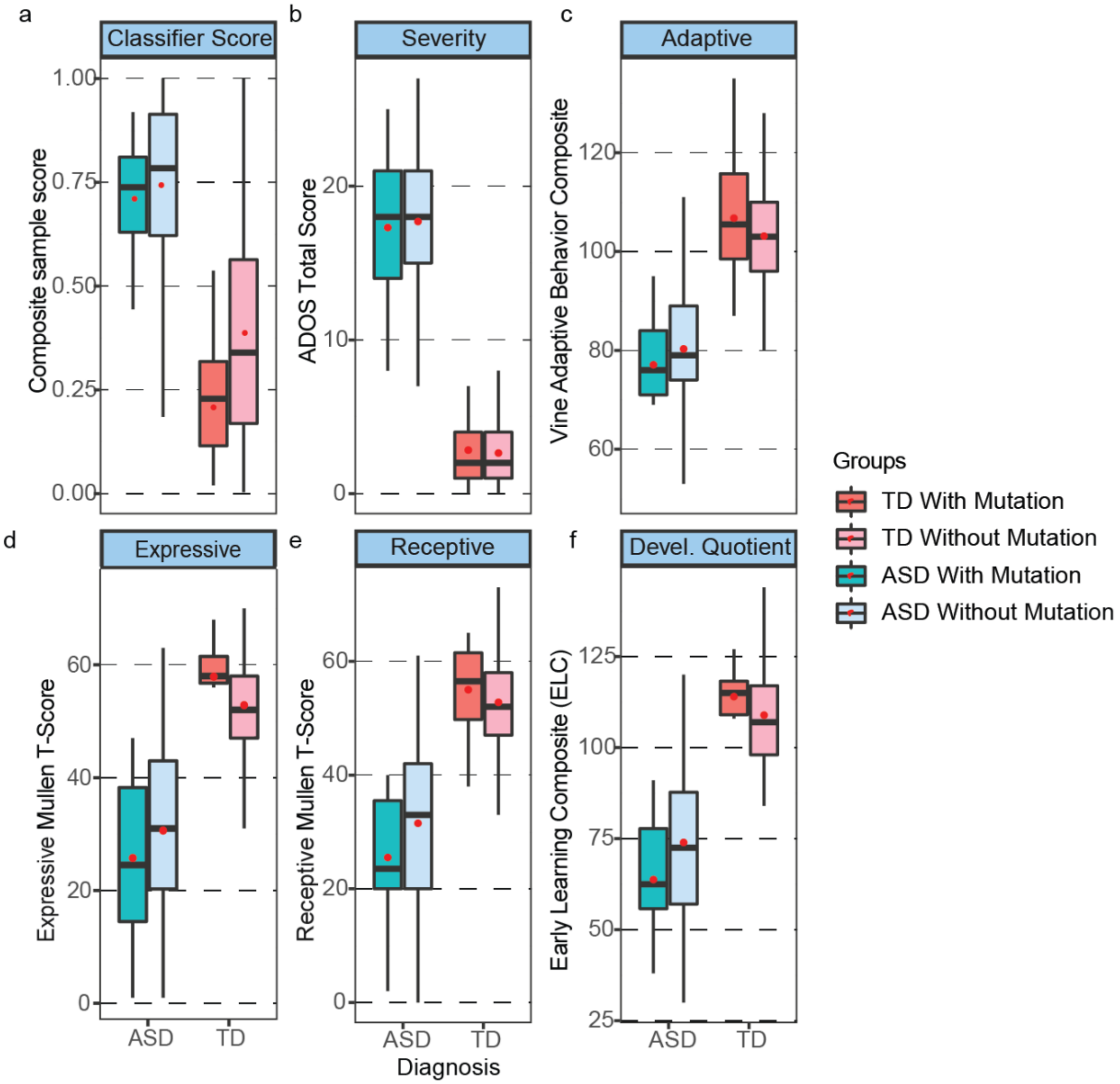


**Figure 2.5 Comparison of ASD with and without ASD risk gene mutations and TD with and without ASD risk gene mutations.** Shows (**a**) ensemble classifier gene expression scores, (**b**) ADOS scores (higher scores are more severe ASD symptoms), (**c**) Vineland adaptive behavior scores (average is 100 + 15), and (**d**) the Mullen T-scores for Expressive and (**e**) Receptive language (average is 50 + 10) as well as (**f**) the Mullen overall Developmental Quotient scores (average is 100 + 15). There were no significant differences in any scores between toddlers with and without SFARI Level 1 or 2 ASD risk gene mutations for ASD toddlers and for TD toddlers. Thus, the presence of an ASD risk gene mutation conferred no clinical liability or difference in gene expression diagnostic score for ASD toddlers. TD toddlers with ASD gene mutations had slightly better cognitive scores than other typically developing toddlers without mutations, but differences were not significant. Red dots are the means and dark lines medians.

In addition, TD subjects with and without SFARI Level 1 or 2 gene mutations did not differ significantly on any clinical test (ADOS, Vineland, Mullen), and, similarly, ASD subjects with and without gene mutations did not differ significantly on any clinical test (Figure 2.5; Appendix Figure 2.7).

### Biological processes enriched by differentially expressed (DE) genes in ASD with higher vs. lower ASD ensemble classifier scores

DE gene analyses (see Methods) found 1,186 DE genes for ASD toddlers with ensemble scores at or above the ASD group mean of 0.723, but no DE genes for those below the group mean (Figure 2.3d and e). Of the 1,186 DE genes, 394 were in the top 500 feature genes selected by the 125 feature routes, and 700 of the 1,186 DE genes were in the first 1000 feature genes. This indicated that DE genes were strong drivers of successful ASD classification. Enrichment analyses of GO biological processes (see Methods) of these 1,186 DE genes found Gene Ontology terms associated with mitotic cell cycle, inflammation/immune response, transcriptional gene regulation, and response to cytokine. Analyses of KEGG pathways using g:Profiler(Raudvere et al. 2019; Kanehisa and Goto 2000) of these 1,186 DE genes found significant pathways included cell cycle (KEGG:hsa04110), PI3K-AKT (KEGG:hsa04151), RAS signaling pathways (KEGG:hsa04014), and Wnt signaling pathways (KEGG:hsa04310), which was consistent with our previous finding(V. H. Gazestani et al. 2019).

### Clinical characteristics associated with higher vs. lower ASD ensemble classifier scores

We compared clinical scores on the ADOS, Mullen, and Vineland for ASD toddlers with ensemble classifier scores at or above the ASD mean of 0.723 to the ASD toddlers with classifier scores below that mean. Diagnostic and psychometric scores were not significantly different between ASD subjects above and below this mean (Appendix Figure 2.8).

Next, we stratified ASD toddlers based on ADOS CoSo Total symptom severity and Mullen scores. Ensemble scores for ASD subjects above vs. below the group average ADOS severity and the group average Mullen means were not practically different (p = 0.59).

We also performed analogous stratifications within the TD Training group and found no ADOS or Mullen differences between higher or lower than the TD mean ensemble classifier score, nor differences in the ensemble scores of TD toddlers with high vs. lower diagnostic and psychometric scores.

### Prenatal characteristics associated with higher vs. lower ASD ensemble classifier scores

Among 127 ASD subjects, 124 had complete prenatal records. We selected the “hospitalization during trimester”, “surgery during trimester” and “confined to bed during trimester” as the risk factors; “nausea during trimester”, “morning sickness during trimester” and “swelling during trimester” as the control prenatal events. Fisher’s t-tests were used to compare the prenatal risk factors across ASD toddlers with ensemble scores at or above the ASD group mean, below that mean, and TD toddlers. ASD toddlers with classifier scores at or above the ASD group mean of 0.718 had significantly fewer prenatal neurodevelopmental risk events, while ASD toddlers below the mean had disproportionately more prenatal risk scores than TD toddlers (Table 2.4 and 2.5). We tested if there was a different ratio of severe prenatal events that could potentially impact ASD development between these two ASD subgroups(Creagh et al. 2016; Atladóttir et al. 2010; Gardener, Spiegelman, and Buka 2009). We found a similar rate of prenatal events between TD subjects and above-the-mean ASD subjects (Odds Ratio: 0.88, Fisher's Exact Test p = 0.84). However, there was a significant enrichment of prenatal events among the below-the-mean ASD subjects compared to TD subjects (Odds Ratio: 2.78; Fisher's Exact Test p = 0.013). As a negative control, prenatal events that are unlikely to affect ASD development were not enriched among ASD subjects with below the ASD mean ensemble score(Gardener, Spiegelman, and Buka 2009). These results suggest the possible existence of different underlying etiological factors between ASD subjects with above vs. below the mean ASD ensemble classifier scores.

**Table 2.4 Distribution of Prenatal Events Among the Three Groups.**

ASD above-the-mean: autism spectrum disorder with ensemble score over 0.714; ASD below-the-mean: below 0.714; TD: typical development.

| Severe prenatal events | ASD Above-the-mean | ASD Below-the-mean | TD |
|---|---|---|---|
| Total subjects (n) | 79 | 45 | 107 |
| Hospitalizations during pregnancy | 5 | 5 | 6 |
| Surgery during pregnancy | 1 | 3 | 4 |
| Confinement to bed during pregnancy | 8 | 8 | 9 |
| General Anesthesia during delivery | 10 | 11 | 10 |
| Total (%) | 19 (24.1) | 20 (44.4) | 24 (22.4) |
| Negative control events | | | |
| Nausea | 4 | 2 | 12 |
| Morning sickness | 42 | 18 | 64 |
| Swelling | 23 | 10 | 26 |
| Total (%) | 52 (65.8) | 24 (53.3) | 70 (65.4) |

**Table 2.5 Statistical Differences between the Three Groups.**

P value is calculated by the two-sided Fisher's exact test.

| | Odds ratio | p value |
|---|---|---|
| ASD above-the-mean vs. TD | 0.914 | 0.861 |
| ASD below-the-mean vs. TD | 2.746 | 0.0102 |
| ASD below-the-mean vs. above | 2.506 | 0.027 |

In the post hoc exploratory analysis, we tested whether adding prenatal features and social behavior scores into models increases model performance. The Bayesian model AUC-ROC increased from 84.67% to 88.20% for the Training dataset, and increased from 89.18% to 91.48% for the Test dataset. (Appendix Result 4).

## Discussion

Despite its high heritability and prenatal beginnings(Courchesne, Gazestani, and Lewis 2020; Courchesne et al. 2019; V. Gazestani et al. 2020; Courchesne et al. 2011; Marchetto et al. 2017; Courchesne and Pierce 2005; Willsey et al. 2013; Courchesne et al. 2007; Stoner et al. 2014; Parikshak et al. 2013; Packer 2016; Kaushik and Zarbalis 2016; Krishnan et al. 2016; Donovan and Basson 2017; Grove et al. 2019; Satterstrom et al. 2020), ASD diagnosis remains behavior-based and the median age of the first diagnosis is about 52 months. Partially due to its genetic and clinical heterogeneity, no single genetic, behavioral or imaging diagnostic marker has been found that can accurately and reproducibly diagnose more than a small subset of affected children. Even among those capable of highly accurately diagnosing subsets of ASD infants and toddlers(Pierce et al. 2016), few have proven clinically useful, cost-effect, and/or practical at the ages when early detection and diagnosis are most needed and could be most important for the child and family.

To approach this dilemma, we addressed ASD genetic and clinical heterogeneity with classifier heterogeneity. That is, since we expected heterogeneity in classifier gene expression features, we designed a classifier pipeline using 42,840 models generated from 3,570 gene expression feature routes and 12 classification methods to classify ASD at ages 1 to 4 years, and applied it to both a Training sample and a held-out Test sample. Then, rather than selecting and reporting a single "best" performing model, we report there are hundreds of good to excellent models and that they can be combined using Bayesian model averaging to bring together 742 "heterogeneous" predictive models involving 125 different feature routes and 2,721 gene expression features. The smMIPs analyses detected 25 TD and ASD subjects with severe mutations in SFARI Level 1 or 2 ASD risk genes: mutation-based classification resulted in chance ASD detection performance, whereas the Bayesian gene expression model correctly classified 22 (88%) of those 25 subjects. The presence of ASD risk gene mutations in typically developing toddlers suggests that the mutations detected here in these specific SFARI genes are neither necessary nor sufficient to cause ASD, are not alone explanatory of autism, and apparently are not clinically diagnostically useful.

Post-hoc Game-Howell tests demonstrated the ensemble gene expression classifier is unbiased towards age differences. The one-way ANOVA test indicated the classifier scores for the ASD group were similar across Hispanic and non-Hispanic subjects and different races. This suggests the classifier is accurately detecting a gene expression pathology common across toddler ages, races and ethnicities in ASD, subjects with and without risk gene mutations, and thus points to common core molecular pathobiology in ASD.

This approach enabled the generation of a composite Bayesian “ensemble” model that is diagnostically predictive and replicable across different toddler ages, races, and ethnicities; performs accurately across the ASD spectrum from more affected to less affected; and has potential for clinical translation. Moreover, this composite ensemble model incorporates both differentially expressed (DE) genes and non-DE genes. This may be relevant to the known complexity of ASD genetics, which may involve common and rare variants and any one or more of >200 different ASD risk gene mutations in different individuals. Non-genetic heterogeneity was also detected here insofar as those with ASD classifier scores below the overall ASD mean tended to have more prenatal risk events in their history than those ASD toddlers with above the mean scores. This opens the important potential to utilize these ASD ensemble classifier scores in future research to identify ASD subtypes that are more driven by genetic versus subtypes more driven by a combination of non-genetic and genetic factors.

Our ensemble features include genes involved in PI3K-AKT, RAS-ERK, and Wnt signaling pathways, immune/inflammation, response to cytokines, transcriptional regulation, and mitotic cell cycle, which are among the pathways and processes found across diverse studies on ASD blood gene expression(Pramparo, Lombardo, et al. 2015; Pramparo, Pierce, et al. 2015; Ch'ng et al. 2015; Diaz-Beltran, Esteban, and Wall 2016; Tylee et al. 2017; He et al. 2019; Lee et al. 2019; Kong et al. 2012; Gregg et al. 2008; Enstrom et al. 2009; Ansel et al. 2016; V. H. Gazestani et al. 2019). This overlap is notable despite the fact that (1) some previous studies did not actively account for race- and ethnicity-related, age-related or clinical-symptom heterogeneity as moderating factors; (2) 84% of 35 previous ASD blood gene expression studies had fewer than 100 ASD subjects and averaged only 28 ASD subjects/study; and (3) many studies focused on older ASD children and adults and only few on ASD toddlers(Pramparo, Lombardo, et al. 2015; Pramparo, Pierce, et al. 2015; Ch'ng et al. 2015; Diaz-Beltran, Esteban, and Wall 2016; Tylee et al. 2017; He et al. 2019; Lee et al. 2019; Kong et al. 2012; Gregg et al. 2008; Enstrom et al. 2009; Ansel et al. 2016).

PI3K-AKT, RAS-ERK and Wnt signaling pathways may be pivotal to ASD prenatal neural maldevelopment. Recently, in a large sample study, we discovered that ASD toddlers had significant upregulation of PI3K-AKT, RAS-ERK and Wnt signaling pathways in both leukocytes and iPSC-derived prenatal neural progenitors and neurons(V. H. Gazestani et al. 2019). This leukocyte dysregulation in 1-4 year old ASD toddlers correlated with ASD social symptom severity(V. H. Gazestani et al. 2019; Michael V. Lombardo et al. 2018). Moreover, these pathways in leukocytes are downstream targets of regulatory risk ASD genes(V. H. Gazestani et al. 2019; V. Gazestani et al. 2020). Leukocyte gene expression also has an potential for understanding molecular correlates of brain size in ASD(Pramparo, Lombardo, et al. 2015) and of atypical cortical patterning subtypes in ASD toddlers with poor language outcome outcomes(Michael V. Lombardo et al. 2018; M. V. Lombardo et al. 2021). Leukocyte expression also relates to hypoactivation response to affective speech in ASD toddlers with poor language outcome(Michael V. Lombardo et al. 2018). Finally, multivariate leukocyte expression signatures can predict trajectories of response to early intervention treatment(Michael V. Lombardo et al. 2021), which underscores the mechanistic relevance of leukocytes to ASD and clinically important phenomena that can be individualized to specific patients. Thus, extensive literature, meta-analyses, and the predictive diagnostic discoveries in the present study, all point to the importance of leukocyte cell biology as clinically informative in ASD and show that ASD-relevant dysregulated gene expression is not restricted to the brain but is also present in other tissues and organs.

Here we developed an innovative and accurate ASD gene expression classifier in ASD toddlers with heterogeneous gene features designed to address early-age ASD genetic and clinical heterogeneity. This predictive classifier in ASD male toddlers aged 1 to 4-year-olds opens the possibility of further refining ASD molecular classifiers optimized for race, ethnicity, and age and with potential for clinical utility. It far outperformed a risk gene-mutation classifier tested in the same toddlers primarily because a significant proportion of TD toddlers have ASD risk gene mutations as well. The ensemble gene expression ASD classifier reported here is enriched in gene expression features involved in ASD prenatal and postnatal pathobiology, and as such, it appears to succeed because of this. Thus, it is more than a signature capable of ASD diagnostic prediction; it is additionally a marker of the underlying pathobiological bases of the disorder in a majority of affected toddlers. It has implications for future research targeting early-age ASD detection and treatment-relevant mechanisms.

## Acknowledgments

This work was supported by NIMH grant no. R01-MH110558 (E.C., N.E.L.), NIMH grant no. R01-MH080134 (K.P.), NIMH grant no. R01-MH104446 (K.P.), and NIMH grant no. P50-MH081755 (E.C.). This work was supported, in part, by a US National Institutes of Health (NIH) grant (R01MH101221) and a grant from the Simons Foundation (SFARI #608045) to E.E.E., and also by “the Fundamental Research Funds for the Central Universities” starting fund (BMU2022RCZX038) to T.W. E.E.E. is an investigator of the Howard Hughes Medical Institute.

Chapter 2, in full, is reprint of the material as it appears in Molecular Psychatric. 2022. Bao, Bokan, Javad Zahiri, Vahid H Gazestani, Linda Lopez, Yaqiong Xiao, Raphael Kim, Teresa H Wen, Austin WT Chiang, Srinivasa Nalabolu, Karen Pierce, Kimberly Robasky, Tianyun Wang, Kendra Hoekzema, Evan E Eichler, Nathan E Lewis, Eric Courchesne. A predictive ensemble classifier for the gene expression

diagnosis of ASD at ages 1 to 4 years. Molecular Psychiatry. 2022. The dissertation/thesis author was the primary investigator and author of the paper.

### Author Contributions

B.B., J.Z., V.H.G., R.K., A.W.T.C, K.P., K.R., N.E.L. and E.C. conceived the project and designed the experiments. L. L. and S.N. collected the samples and managed the data. K.H., T.W., and E.E.E. performed the smMIPs targeted sequencing and analyses. B.B., J.Z., V.H.G., T.W., R.K. and E.C. analyzed the data. E.C., B.B., J.Z., V.H.G., Y.X., R.K., K.P., and N.E.L. interpreted the results and wrote the manuscript. E.C. and N.E.L. supervised the project.

### Competing interests

Nathan Lewis, Eric Courchesne and Vahid H. Gazestani have two patents WO2014018774A1 and WO2020014620 relevant to this paper. E.E.E. is a scientific advisory board (SAB) member of Variant Bio, Inc.

### Code availability

We provide the code library in R and Python described in this work through Github: https://github.com/LewisLabUCSD/autism_classifier. We provide Jupyter notebooks in Python and R to generate our figures and analysis.

## References

Aguilar-Lacasaña, Sofía, Natàlia Vilor-Tejedor, Philip R. Jansen, Mònica López-Vicente, Mariona Bustamante, Miguel Burgaleta, Jordi Sunyer, and Silvia Alemany. 2022. "Polygenic Risk for ADHD and ASD and Their Relation with Cognitive Measures in School Children." *Psychological Medicine* 52 (7): 1356–64.

Ansel, Ashley, Joshua P. Rosenzweig, Philip D. Zisman, Michal Melamed, and Benjamin Gesundheit. 2016. "Variation in Gene Expression in Autism Spectrum Disorders: An Extensive Review of Transcriptomic Studies." *Frontiers in Neuroscience* 10: 601.

Antaki, Danny, James Guevara, Adam X. Maihofer, Marieke Klein, Madhusudan Gujral, Jakob Grove, Caitlin E. Carey, et al. 2022. "Publisher Correction: A Phenotypic Spectrum of Autism Is Attributable to the Combined Effects of Rare Variants, Polygenic Risk and Sex." *Nature Genetics* 54 (8): 1259.

Antonio Pedro Duarte Silva <psilva@porto.ucp.pt>. 2015. "SelectV: Variable Selection for High-Dimensional Supervised... In HiDimDA: High Dimensional Discriminant Analysis." October 19, 2015. https://rdrr.io/cran/HiDimDA/man/SelectV.html.

Ardlie, K. G., D. S. Deluca, A. V. Segre, T. J. Sullivan, T. R. Young, E. T. Gelfand, C. A. Trowbridge, et al. 2015. "The Genotype-Tissue Expression (GTEx) Pilot Analysis: Multitissue Gene Regulation in Humans." *Science* 348 (6235): 648–60.

Atladóttir, Hjördis O., Poul Thorsen, Lars Østergaard, Diana E. Schendel, Sanne Lemcke, Morsi Abdallah, and Erik T. Parner. 2010. "Maternal Infection Requiring Hospitalization during Pregnancy and Autism Spectrum Disorders." *Journal of Autism and Developmental Disorders* 40 (12): 1423–30.

Autism and Developmental Disabilities Monitoring Network Surveillance Year 2006 Principal Investigators, and Centers for Disease Control and Prevention (CDC). 2009. "Prevalence of Autism Spectrum Disorders - Autism and Developmental Disabilities Monitoring Network, United States, 2006." *Morbidity and Mortality Weekly Report. Surveillance Summaries* 58 (10): 1–20.

Bacon, Elizabeth C., Eric Courchesne, Cynthia Carter Barnes, Debra Cha, Sunny Pence, Laura Schreibman, Aubyn C. Stahmer, and Karen Pierce. 2018. “Rethinking the Idea of Late Autism Spectrum Disorder Onset.” *Development and Psychopathology* 30 (2): 553–69.

Bacon, Elizabeth C., Suzanna Osuna, Eric Courchesne, and Karen Pierce. 2019. “Naturalistic Language Sampling to Characterize the Language Abilities of 3-Year-Olds with Autism Spectrum Disorder.” *Autism: The International Journal of Research and Practice* 23 (3): 699–712.

Bai, Dan, Benjamin Hon Kei Yip, Gayle C. Windham, Andre Sourander, Richard Francis, Rinat Yoffe, Emma Glasson, et al. 2019. “Association of Genetic and Environmental Factors With Autism in a 5-Country Cohort.” *JAMA Psychiatry* 76 (10): 1035–43.

Baio, Jon, Lisa Wiggins, Deborah L. Christensen, Matthew J. Maenner, Julie Daniels, Zachary Warren, Margaret Kurzius-Spencer, et al. 2018. “Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014.” *Morbidity and Mortality Weekly Report. Surveillance Summaries* 67 (6): 1–23.

Bal, Vanessa H., So-Hyun Kim, Megan Fok, and Catherine Lord. 2019. “Autism Spectrum Disorder Symptoms from Ages 2 to 19 Years: Implications for Diagnosing Adolescents and Young Adults.” *Autism Research: Official Journal of the International Society for Autism Research* 12 (1): 89–99.

Chambers, and Hastie. n.d. “Statistical Models in S. Wadsworth & Brooks/Cole.” *Pacific Grove, CA*.

Ch’ng, Carolyn, Willie Kwok, Sanja Rogic, and Paul Pavlidis. 2015. “Meta-Analysis of Gene Expression in Autism Spectrum Disorder.” *Autism Research: Official Journal of the International Society for Autism Research* 8 (5): 593–608.

Christensen, Deborah L., Kim Van Naarden Braun, Jon Baio, Deborah Bilder, Jane Charles, John N. Constantino, Julie Daniels, et al. 2018. “Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2012.” *Morbidity and Mortality Weekly Report. Surveillance Summaries* 65 (13): 1–23.

Clarke, T-K, M. K. Lupton, A. M. Fernandez-Pujals, J. Starr, G. Davies, S. Cox, A. Pattie, et al. 2016. “Common Polygenic Risk for Autism Spectrum Disorder (ASD) Is Associated with Cognitive Ability in the General Population." *Molecular Psychiatry* 21 (3): 419–25.

Courchesne, Eric, Vahid H. Gazestani, and Nathan E. Lewis. 2020. "Prenatal Origins of ASD: The When, What, and How of ASD Development." *Trends in Neurosciences* 43 (5): 326–42.

Courchesne, Eric, Peter R. Mouton, Michael E. Calhoun, Katerina Semendeferi, Clelia Ahrens-Barbeau, Melodie J. Hallet, Cynthia Carter Barnes, and Karen Pierce. 2011. "Neuron Number and Size in Prefrontal Cortex of Children with Autism." *JAMA: The Journal of the American Medical Association* 306 (18): 2001–10.

Courchesne, Eric, and Karen Pierce. 2005. "Why the Frontal Cortex in Autism Might Be Talking Only to Itself: Local over-Connectivity but Long-Distance Disconnection." *Current Opinion in Neurobiology* 15 (2): 225–30.

Courchesne, Eric, Karen Pierce, Cynthia M. Schumann, Elizabeth Redcay, Joseph A. Buckwalter, Daniel P. Kennedy, and John Morgan. 2007. "Mapping Early Brain Development in Autism." *Neuron* 56 (2): 399–413.

Courchesne, Eric, Tiziano Pramparo, Vahid H. Gazestani, Michael V. Lombardo, Karen Pierce, and Nathan E. Lewis. 2019. "The ASD Living Biology: From Cell Proliferation to Clinical Phenotype." *Molecular Psychiatry* 24 (1): 88–107.

Creagh, O., H. Torres, K. Rivera, M. Morales-Franqui, G. Altieri-Acevedo, and D. Warner. 2016. "Previous Exposure to Anesthesia and Autism Spectrum Disorder (ASD): A Puerto Rican Population-Based Sibling Cohort Study." *Boletin de La Asociacion Medica de Puerto Rico* 108 (2): 73–80.

Diaz-Beltran, L., F. J. Esteban, and D. P. Wall. 2016. "A Common Molecular Signature in ASD Gene Expression: Following Root 66 to Autism." *Translational Psychiatry* 6 (January): e705.

Donovan, Alex P. A., and M. Albert Basson. 2017. "The Neuroanatomy of Autism - a Developmental Perspective." *Journal of Anatomy* 230 (1): 4–15.

Du, P., W. A. Kibbe, and S. M. Lin. 2008. "Lumi: A Pipeline for Processing Illumina Microarray." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btn224.

Enstrom, Amanda M., Lisa Lit, Charity E. Onore, Jeff P. Gregg, Robin L. Hansen, Isaac N. Pessah, Irva

Hertz-Picciotto, Judy A. Van de Water, Frank R. Sharp, and Paul Ashwood. 2009. “Altered Gene Expression and Function of Peripheral Blood Natural Killer Cells in Children with Autism.” *Brain, Behavior, and Immunity* 23 (1): 124–33.

Feliciano, Pamela, Xueya Zhou, Irina Astrovskaya, Tychele N. Turner, Tianyun Wang, Leo Brueggeman, Rebecca Barnard, et al. 2019. “Exome Sequencing of 457 Autism Families Recruited Online Provides Evidence for Autism Risk Genes.” *NPJ Genomic Medicine* 4 (August): 19.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2010. “Regularization Paths for Generalized Linear Models via Coordinate Descent.” *Journal of Statistical Software* 33 (1): 1–22.

Gardener, Hannah, Donna Spiegelman, and Stephen L. Buka. 2009. “Prenatal Risk Factors for Autism: Comprehensive Meta-Analysis.” *The British Journal of Psychiatry: The Journal of Mental Science* 195 (1): 7–14.

Gazestani, Vahid, Austin W. T. Chiang, E. Courchesne, and N. E. Lewis. 2020. “Autism Genetics Perturb Prenatal Neurodevelopment through a Hierarchy of Broadly-Expressed and Brain-Specific Genes.” *bioRxiv*.

Gazestani, Vahid H., Tiziano Pramparo, Srinivasa Nalabolu, Benjamin P. Kellman, Sarah Murray, Linda Lopez, Karen Pierce, Eric Courchesne, and Nathan E. Lewis. 2019. “A Perturbed Gene Network Containing PI3K-AKT, RAS-ERK and WNT-β-Catenin Pathways in Leukocytes Is Linked to ASD Genetics and Symptom Severity.” *Nature Neuroscience* 22 (10): 1624–34.

Grau, Jan, Ivo Grosse, and Jens Keilwagen. 2015. “PRROC: Computing and Visualizing Precision-Recall and Receiver Operating Characteristic Curves in R.” *Bioinformatics* 31 (15): 2595–97.

Gregg, Jeffrey P., Lisa Lit, Colin A. Baron, Irva Hertz-Picciotto, Wynn Walker, Ryan A. Davis, Lisa A. Croen, et al. 2008. “Gene Expression Changes in Children with Autism.” *Genomics* 91 (1): 22–29.

Grove, Jakob, Stephan Ripke, Thomas D. Als, Manuel Mattheisen, Raymond K. Walters, Hyejung Won, Jonatan Pallesen, et al. 2019. “Identification of Common Genetic Risk Variants for Autism Spectrum Disorder.” *Nature Genetics* 51 (3): 431–44.

Hewitson, Laura, Jeremy A. Mathews, Morgan Devlin, Claire Schutte, Jeon Lee, and Dwight C. German.

2021. “Blood Biomarker Discovery for Autism Spectrum Disorder: A Proteomic Analysis.” *PloS One* 16 (2): e0246581.

He, Yi, Yuan Zhou, Wei Ma, and Juan Wang. 2019. “An Integrated Transcriptomic Analysis of Autism Spectrum Disorder.” *Scientific Reports* 9 (1): 11818.

“Human Gene Module.” n.d. SFARI Gene. Accessed August 25, 2022. https://gene-archive.sfari.org/database/human-gene/.

Kanehisa, M., and S. Goto. 2000. “KEGG: Kyoto Encyclopedia of Genes and Genomes.” *Nucleic Acids Research* 28 (1): 27–30.

Kaushik, Gaurav, and Konstantinos S. Zarbalis. 2016. “Prenatal Neurogenesis in Autism Spectrum Disorders.” *Frontiers in Chemistry*. https://doi.org/10.3389/fchem.2016.00012.

Klei, Lambertus, Lora Lee McClain, Behrang Mahjani, Klea Panayidou, Silvia De Rubeis, Anna-Carin Säll Grahnat, Gun Karlsson, et al. 2021. “How Rare and Common Risk Variation Jointly Affect Liability for Autism Spectrum Disorder.” *Molecular Autism* 12 (1): 66.

Kong, Sek Won, Christin D. Collins, Yuko Shimizu-Motohashi, Ingrid A. Holm, Malcolm G. Campbell, In-Hee Lee, Stephanie J. Brewster, et al. 2012. “Characteristics and Predictive Value of Blood Transcriptome Signature in Males with Autism Spectrum Disorders.” *PloS One* 7 (12): e49475.

Krishnan, Arjun, Ran Zhang, Victoria Yao, Chandra L. Theesfeld, Aaron K. Wong, Alicja Tadych, Natalia Volfovsky, Alan Packer, Alex Lash, and Olga G. Troyanskaya. 2016. “Genome-Wide Prediction and Functional Characterization of the Genetic Basis of Autism Spectrum Disorder.” *Nature Neuroscience* 19 (11): 1454–62.

Langfelder, Peter, and Steve Horvath. 2008. “WGCNA: an R package for weighted correlation network analysis.” *BMC Bioinformatics* 9 (January): 559.

Lee, Samuel C., Thomas P. Quinn, Jerry Lai, Sek Won Kong, Irva Hertz-Picciotto, Stephen J. Glatt, Tamsyn M. Crowley, Svetha Venkatesh, and Thin Nguyen. 2019. “Solving for X: Evidence for Sex-Specific Autism Biomarkers across Multiple Transcriptomic Studies.” *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric*

*Genetics* 180 (6): 377–89.

Liaw, Andy, Matthew Wiener, and Others. 2002. “Classification and Regression by randomForest.” *R News* 2 (3): 18–22.

Lombardo, Michael V., Elena Maria Busuoli, Laura Schreibman, Aubyn C. Stahmer, Tiziano Pramparo, Isotta Landi, Veronica Mandelli, et al. 2021. “Pre-Treatment Clinical and Gene Expression Patterns Predict Developmental Change in Early Intervention in Autism.” *Molecular Psychiatry* 26 (12): 7641–51.

Lombardo, Michael V., Meng-Chuan Lai, and Simon Baron-Cohen. 2019. “Big Data Approaches to Decomposing Heterogeneity across the Autism Spectrum.” *Molecular Psychiatry* 24 (10): 1435–50.

Lombardo, Michael V., Tiziano Pramparo, Vahid Gazestani, Varun Warrier, Richard A. I. Bethlehem, Cynthia Carter Barnes, Linda Lopez, et al. 2018. “Large-Scale Associations between the Leukocyte Transcriptome and BOLD Responses to Speech Differ in Autism Early Language Outcome Subtypes.” *Nature Neuroscience* 21 (12): 1680–88.

Lombardo, M. V., L. Eyler, T. Pramparo, and V. H. Gazestani. 2021. “Atypical Genomic Cortical Patterning in Autism with Poor Early Language Outcome.” *bioRxiv*. https://www.biorxiv.org/content/10.1101/2020.08.18.253443v3.abstract.

Lord, Catherine. 2012. *Autism Diagnostic Observation Schedule, Second Edition (ADOS-2)*. WPS.

Maenner, Matthew J., Kelly A. Shaw, Jon Baio, EdS1, Anita Washington, Mary Patrick, Monica DiRienzo, et al. 2020. “Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2016.” *Morbidity and Mortality Weekly Report. Surveillance Summaries* 69 (4): 1–12.

Marchetto, Maria C., Haim Belinson, Yuan Tian, Beatriz C. Freitas, Chen Fu, Krishna Vadodaria, Patricia Beltrao-Braga, et al. 2017. “Altered Proliferation and Networks in Neural Cells Derived from Idiopathic Autistic Individuals.” *Molecular Psychiatry* 22 (6): 820–35.

Mevik, Bjørn-Helge, and Ron Wehrens. 2015. “Introduction to the Pls Package.” *Help Section of The “Pls” Package of R Studio Software; R Foundation for Statistical Computing: Vienna, Austria*, 1–23.

Meyer, Patrick E., Frédéric Lafitte, and Gianluca Bontempi. 2008. “Minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information.” *BMC Bioinformatics* 9 (October): 461.

Mullen, E. M. n.d. “Mullen Scales of Early Learning.” Accessed August 25, 2022. http://www.v-psyche.com/doc/special-cases/Mullen%20Scales%20of%20Early%20Learning.docx.

Packer, Alan. 2016. “Neocortical Neurogenesis and the Etiology of Autism Spectrum Disorder.” *Neuroscience and Biobehavioral Reviews* 64 (May): 185–95.

Parikshak, Neelroop N., Rui Luo, Alice Zhang, Hyejung Won, Jennifer K. Lowe, Vijayendran Chandran, Steve Horvath, and Daniel H. Geschwind. 2013. “Integrative Functional Genomic Analyses Implicate Specific Molecular Pathways and Circuits in Autism.” *Cell*. https://doi.org/10.1016/j.cell.2013.10.031.

“penalizedSVM: Feature Selection SVM Using Penalty Functions.” n.d. Accessed June 29, 2021. https://cran.r-project.org/web/packages/penalizedSVM/index.html.

Pierce, Karen, Vahid Gazestani, Elizabeth Bacon, Eric Courchesne, Amanda Cheng, Cynthia Carter Barnes, Srinivasa Nalabolu, et al. 2021. “Get SET Early to Identify and Treatment Refer Autism Spectrum Disorder at 1 Year and Discover Factors That Influence Early Diagnosis.” *The Journal of Pediatrics* 236 (September): 179–88.

Pierce, Karen, Steven Marinero, Roxana Hazin, Benjamin McKenna, Cynthia Carter Barnes, and Ajith Malige. 2016. “Eye Tracking Reveals Abnormal Visual Preference for Geometric Images as an Early Biomarker of an Autism Spectrum Disorder Subtype Associated With Increased Symptom Severity.” *Biological Psychiatry* 79 (8): 657–66.

Pramparo, Tiziano, Michael V. Lombardo, Kathleen Campbell, Cynthia Carter Barnes, Steven Marinero, Stephanie Solso, Julia Young, et al. 2015. “Cell Cycle Networks Link Gene Expression Dysregulation, Mutation, and Brain Maldevelopment in Autistic Toddlers.” *Molecular Systems Biology* 11 (12): 841.

Pramparo, Tiziano, Karen Pierce, Michael V. Lombardo, Cynthia Carter Barnes, Steven Marinero, Clelia Ahrens-Barbeau, Sarah S. Murray, Linda Lopez, Ronghui Xu, and Eric Courchesne. 2015. “Prediction of Autism by Translation and Immune/inflammation Coexpressed Genes in Toddlers from Pediatric

Community Practices." *JAMA Psychiatry* 72 (4): 386–94.

Raudvere, Uku, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. 2019. "g:Profiler: A Web Server for Functional Enrichment Analysis and Conversions of Gene Lists (2019 Update)." *Nucleic Acids Research* 47 (W1): W191–98.

Ridgeway, Greg. 2007. "Generalized Boosted Models: A Guide to the Gbm Package." *Update* 1 (1): 2007.

Ripley, Brian D. 2002. *Modern Applied Statistics with S*. springer.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. "Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkv007.

Robinson, Elise B., Beate St Pourcain, Verneri Anttila, Jack A. Kosmicki, Brendan Bulik-Sullivan, Jakob Grove, Julian Maller, et al. 2016. "Genetic Risk for Autism Spectrum Disorders and Neuropsychiatric Variation in the General Population." *Nature Genetics* 48 (5): 552–55.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. "pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves." *BMC Bioinformatics* 12 (March): 77.

"Rstatix." n.d. Accessed June 5, 2022. https://rpkgs.datanovia.com/rstatix/.

Saldana, Diego Franco, and Yang Feng. 2018. "SIS: An R Package for Sure Independence Screening in Ultrahigh-Dimensional Statistical Models." *Journal of Statistical Software, Articles* 83 (2): 1–25.

Satterstrom, F. Kyle, Jack A. Kosmicki, Jiebiao Wang, Michael S. Breen, Silvia De Rubeis, Joon-Yong An, Minshi Peng, et al. 2020. "Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism." *Cell* 180 (3): 568–84.e23.

Smyth, G. K. 2005. "Limma: Linear Models for Microarray Data." In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, edited by Robert Gentleman, Vincent J. Carey, Wolfgang Huber, Rafael A. Irizarry, and Sandrine Dudoit, 397–420. New York, NY: Springer New York.

Sparrow, Balla, and Cicchetti. n.d. "Vineland Scales of Adaptive Behavior, Survey Form Manual." *Circle*

*Pines, MN: American Guidance Service*.

Stessman, Holly A. F., Bo Xiong, Bradley P. Coe, Tianyun Wang, Kendra Hoekzema, Michaela Fenckova, Malin Kvarnung, et al. 2017. "Targeted Sequencing Identifies 91 Neurodevelopmental-Disorder Risk Genes with Autism and Developmental-Disability Biases." *Nature Genetics* 49 (4): 515–26.

Stoner, Rich, Maggie L. Chow, Maureen P. Boyle, Susan M. Sunkin, Peter R. Mouton, Subhojit Roy, Anthony Wynshaw-Boris, Sophia A. Colamarino, Ed S. Lein, and Eric Courchesne. 2014. "Patches of Disorganization in the Neocortex of Children with Autism." *New England Journal of Medicine*. https://doi.org/10.1056/nejmoa1307491.

Subramanian, Aravind, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, et al. 2005. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–50.

Su, Gang, John H. Morris, Barry Demchak, and Gary D. Bader. 2014. "Biological Network Exploration with Cytoscape 3." *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 47 (September): 8.13.1–24.

Supek, Fran, Matko Bošnjak, Nives Škunca, and Tomislav Šmuc. 2011. "REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms." *PloS One* 6 (7): e21800.

Tylee, Daniel S., Jonathan L. Hess, Thomas P. Quinn, Rahul Barve, Hailiang Huang, Yanli Zhang-James, Jeffrey Chang, et al. 2017. "Blood Transcriptomic Comparison of Individuals with and without Autism Spectrum Disorder: A Combined-Samples Mega-Analysis." *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics* 174 (3): 181–201.

Wang, Tianyun, Kendra Hoekzema, Davide Vecchio, Huidan Wu, Arvis Sulovari, Bradley P. Coe, Madelyn A. Gillentine, et al. 2020. "Large-Scale Targeted Sequencing Identifies Risk Genes for Neurodevelopmental Disorders." *Nature Communications* 11 (1): 4932.

Wehrens, Ron, and B-H Mevik. 2007. "The Pls Package: Principal Component and Partial Least Squares

Regression in R." https://repository.ubn.ru.nl/bitstream/handle/2066/36604/36604.pdf.

Wetherby, Amy M., Lori Allen, Julie Cleary, Kary Kublin, and Howard Goldstein. 2002. "Validity and Reliability of the Communication and Symbolic Behavior Scales Developmental Profile with Very Young Children." *Journal of Speech, Language, and Hearing Research: JSLHR* 45 (6): 1202–18.

Willsey, A. Jeremy, Stephan J. Sanders, Mingfeng Li, Shan Dong, Andrew T. Tebbenkamp, Rebecca A. Muhle, Steven K. Reilly, et al. 2013. "Coexpression Networks Implicate Human Midfetal Deep Cortical Projection Neurons in the Pathogenesis of Autism." *Cell* 155 (5): 997–1007.

# CHAPTER 3: Examination of automatic facial action unit measurement as a mechanism to differentiate ASD vs non-ASD toddlers

## Abstract

Historically, children with autism have been characterized as having challenges with emotional reactivity, most often under-reactivity, a profile that is central to both DSM-5 criteria as well as diagnostic tests such as ADOS. A less well-known finding is that in certain situations, emotional responding has been shown to be overly intense in ASD. The purpose of the current study was to determine if differences exist in emotional reactivity between toddlers with ASD and other toddlers using state of the art expression analysis software, and to determine if differences, if found, could be used as a diagnostic marker of ASD.

## Methods

A cohort of 184 toddlers (129 ASD, 55 non-ASD) watched an 87-second movie, 'The Joint Attention Test', in which a female is telling stories and playing with toys across several scenes. The point of gaze was collected using an eye-tracking machine. Toddlers' facial expressions were recorded using a standard webcam, and the intensity scores of the facial action unit (FACS) and the coexpression score were measured using OpenFace 2.0 and Emonet. A machine learning classifier was trained (n = 154 toddlers) and tested (n = 30 toddlers) to distinguish ASD from non-ASD toddlers.

## Results

Overall, children with ASD displayed more intense expressions in reaction to some portions of the video, particularly within the brow lowerer, chin raiser, and lip dimpler facial action units. Our action unit classifier had a sensitivity of 83.3% and a specificity of 67.5% in the test dataset (90.1% and 75% in the training dataset). We verified that our classifier was unbiased against common confounding factors (age, race, and ethnicity). By combining the action unit classifier and Geo-Pref non-social score, we achieved a specificity of 100% and sensitivity of 50% on the training and test datasets. The ensemble classifier

maintained the high specificity while considerably increasing the sensitivity, which provides the potential for screening applications.

## Introduction

ASD is a prenatal, highly heritable disorder that considerably affects a child's ability to perceive and react to social information(Courchesne, Gazestani, and Lewis 2020; Courchesne et al. 2019, 2011; Packer 2016; Kaushik and Zarbalis 2016; Bal et al. 2019; Bonnet-Brilhault et al. 2018). With early screening using parent report tools and detection programs such as Get SET Early (Pierce et al. 2021), toddlers can be quickly and inexpensively screened and many are diagnosed with ASD as early as 12 months, with stable diagnoses across toddler and early childhood(Pierce et al. 2011, 2019). Although reliable early diagnosis is possible, most ASD children are not diagnosed until around age 4(Maenner et al. 2021; Autism and Developmental Disabilities Monitoring Network Surveillance Year 2006 Principal Investigators and Centers for Disease Control and Prevention [CDC] 2009; Baio et al. 2018; Christensen et al. 2018) and thus misses valuable opportunities associated with early treatment engagement (Gabbay-Dizdar et al. 2021). As such, the discovery of objective, biologically-based markers of ASD that can increase the pace of diagnosis and reduce the requirement for highly-trained professionals (J. McPartland, Dawson, and Webb 2004; Frazier et al. 2021; J. C. McPartland et al. 2020) are needed. The development of computer vision technology is now supporting a surge in research that is designed to measure ASD subjects' facial behavior responding to social information in a subjective and quantitative way (Macari et al. 2018; Trevisan, Hoskyn, and Birmingham 2018; Press, Richardson, and Bird 2010; Deschamps et al. 2015; Weiss et al. 2019; Rozga et al. 2013; Faso, Sasson, and Pinkham 2015; Bangerter et al. 2020).

Since ASD was first identified in 1943, two stereotypes concerning the emotional lives of children affected by the disorder have prevailed: one in which negative emotions dominate and the other in which emotional expressions are muted, particularly positively valenced emotions (Cooper and Michels 1988; Harms, Martin, and Wallace 2010; Uljarevic and Hamilton 2013; Langdell 1978; Begeer et al. 2008; Kennedy and Adolphs 2012). Not surprisingly, research in autism has focused on examining either a negative emotionality bias or an attenuation of positive emotion.(Castelli 2005; Atkinson 2009; Philip et al. 2010). Recent evidence, however, shows that autistic individuals may not necessarily differ in expression intensity of emotions, nor have negative emotionality bias (Macari et al. 2018; Trevisan, Hoskyn, and

Birmingham 2018; Press, Richardson, and Bird 2010; Deschamps et al. 2015; Weiss et al. 2019; Rozga et al. 2013). For example, in a recent study evoked expressions in response to funny videos in ASD adults were rated as more intense, although less natural, than TD expressions(Faso, Sasson, and Pinkham 2015). The clustering based on evoked action unit intensity identified an ASD subgroup, proposed as an "over-responsive group," that expresses more intense positive facial expressions than the TD group in response to the videos (Bangerter et al. 2020).

At the same time, research has shown that individuals with ASD sometimes exhibit emotions that are incongruent with real-world events, as evidenced by executing atypical expressions patterns(Carpenter et al. 2021; Brewer et al. 2016; Faso, Sasson, and Pinkham 2015; Weiss et al. 2019; Rozga et al. 2013). Indeed, children and adults with ASD exhibit reduced, atypical, or delayed spontaneous mimicry responses to photographs and videos of emotional facial expressions (Zampella, Bennetto, and Herrington 2020a; Rieffe, Meerum Terwogt, and Stockmann 2000). Specific facial behaviors, including eye contact, smiling, and eyebrow movements, can distinguish ASD subjects from control participants. Such changes are relevant to biological hypotheses about abnormalities in the medial prefrontal cortex and a visual network within the occipitotemporal cortex (Moore et al. 2018). It suggests those differences in facial behavior could lead to potential phenotypic biomarkers of ASD.

In order to measure facial behavior objectively and quantitatively, automated facial analysis tools have been developed to empower the analysis in different parts of a range of disorders and conditions (Leo et al. 2018; LoBue and Thrasher 2014; Sariyanidi et al. 2020; Bangerter et al. 2020; Jacques et al. 2022). It enables scientists to measure the facial responses to emotional stimuli in an efficient, granular, and objective perspective (Bangerter et al. 2020; Pulido-Castro et al. 2021; Baltrusaitis et al. 2018). However, the efficacy of using the automatic facial expression test as an early screening tool for ASD remains underexplored (Jacques et al. 2022). Most of the established facial expression tests require the interaction between the psychologist and the child (Zampella, Bennetto, and Herrington 2020b). This limits researchers' and clinicians' ability to assess critical behaviors and measure differences across individuals, contexts, or time.

Thus, there is a lack of established automatic methods for operationalizing toddlers' emotional reciprocity objectively or granularly.

Preferential looking paradigms have been successfully adopted as a method for identifying visual attention preferences in ASD (Kaliukhovich et al. 2021; Pierce et al. 2016; Wen et al. 2022). One such preferential looking test, the GeoPref test found a subset of ASD toddlers strongly preferred geometric images when presented with social and geometric motion images (Pierce et al. 2016). The toddlers with a higher preference for geometric images demonstrated greater symptom severity and fewer gaze shifts at school age (Bacon et al. 2020). The success of the GeoPref Test as a symptom severity prognostic tool encourages us to study the toddler's facial emotional response to different movie scenes.

In this study, we leveraged a new eye-tracking test called 'The Joint Attention Test' (Andreason et al., In preparation) that features a female speaking in a child-friendly, emotionally valent voice while engaging with various toys and objects. We utilized freely available software, Openface 2.0 (Baltrusaitis et al. 2018) and Emonet (Toisoul et al. 2021), to analyze webcam images and measure faction action unit intensity (Figure 3.1). We then used the corresponding features to train a classifier to differentiate between ASD and non-ASD subjects. We verified that our classifier was unbiased against common confounding factors (age, race, and ethnicity). Further, we tested the combination of our classifiers with the GeoPref percent fixation (Wen et al. 2022) on geometric image score that has been shown to have high specificity and good PPV in predicting ASD diagnosis. The final unsupervised clustering analysis including the classifier score, eye-tracking data, and social behavior data provided further insight into the clinical behavior heterogeneity among different subgroups.

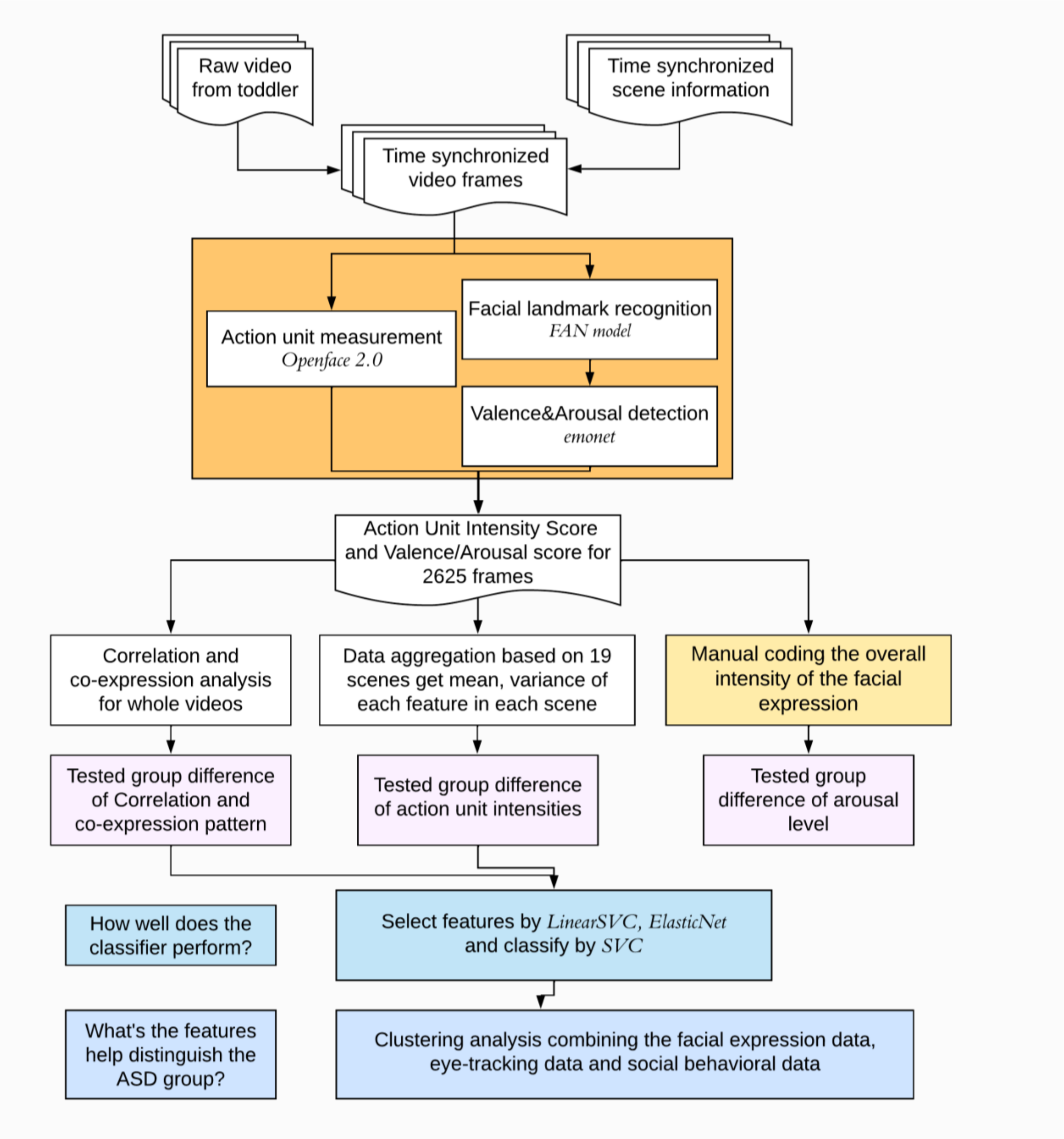

Raw video from toddler
Time synchronized scene information
Time synchronized video frames
Action unit measurement
Openface 2.0
Facial landmark recognition
FAN model
Valence&Arousal detection
emonet
Action Unit Intensity Score and Valence/Arousal score for 2625 frames
Correlation and co-expression analysis for whole videos
Data aggregation based on 19 scenes get mean, variance of each feature in each scene
Manual coding the overall intensity of the facial expression
Tested group difference of Correlation and co-expression pattern
Tested group difference of action unit intensities
Tested group difference of arousal level
How well does the classifier perform?
Select features by LinearSVC, ElasticNet and classify by SVC
What's the features help distinguish the ASD group?
Clustering analysis combining the facial expression data, eye-tracking data and social behavioral data

**Figure 3.1 An analysis workflow of this study.** After the videos were time-synchronized, the videos went through two open-source models for action unit detection and valence/arousal estimation. The manual coding validated the overall measurement result. The output data were first analyzed to quantify correlation and co-expression and then aggregated into 19 scenes to extract the features such as maximum, mean, and variance. The machine learning methods were then used to select features and train the final model.

## Methods

### Participants

Subjects were recruited through general community referral or through a population-based screening method known as Get SET Early(Pierce et al. 2021). Following the screening, toddlers were referred to the University of California, San Diego Autism Center of Excellence for an in-depth diagnostic evaluation and eye-tracking. Subjects were invited for repeat testing every ~12 months until age 3. Toddlers were assessed by licensed Ph.D.-level clinical psychologists blind to eye-tracking results using the Mullen Scales of Early Learning (Mullen n.d.), the Autism Diagnostic Observation Schedule (Module 1, or Module 2, or Toddler Modules) (Lord 2012), and the Vineland Adaptive Behavior Scales (Sparrow, Balla, and Cicchetti, n.d.). Parents were given diagnostic feedback and toddlers were referred for treatment as appropriate.

Only subjects with continuous, full-face visibility and high-quality video recordings were included. Among 423 toddlers who completed the eye tracking test, 209 recordings were excluded because the subjects' faces (the mouth area) were covered with hands for more than 60% of the time; subjects were wearing a face mask; the subjects were eating during testing. We further excluded 30 videos that had low eye-gazing time on screen, no face detected, or the subjects were age outliers. The final dataset used for analysis included 129 ASD (123 ASD, 6 ASD Features) and 55 non-ASD subjects (22 typically developing [TD], 16 language delay [LD], 10 Other, 4 typical toddlers [TD] with an ASD sibling [TypSibASD] and 3 developmental delay [GDD]).

### The Joint Attention Test

'The Joint Attention Test' is an 87.5-second-long video in which an actress sits at a table and speaks in an emotionally intense, child-friendly voice while engaging with various toys and objects (Figure 3.2). The actress uses a series of 19 joint attention bids to direct the child's attention to different items in the room. Each joint attention bid involves the actress using speech cues to direct the child's attention overtly (e.g., "Look at this comb"), directing the child to look toward it, and then looking and/or pointing at the object before finally interacting with it (Andreason et al., in preparation). This particular eye-tracking video was selected for the current study given its high emotional tone and the strong emotional response observed in toddlers during its development.

### Eye-tracking and facial expression data collection

Eye-gaze data was collected using the Tobii Pro Spectrum (Tobii, Stockholm, Sweden; www.tobii.com; 600 Hz sampling rate; 1280 × 1024) while toddlers watched 'The Joint Attention Test. To ensure that only the toddler's gaze was tracked and free from parent influence, standardized instructions were read out explicitly to parents prior to the eye-tracking test that requested that they looked at a dot placed ~3ft above the eye-tracking machine. A five-point calibration was then performed using animated cartoon ducks with sounds, and data were only used if calibration results, fell within manufacturer-reported parameters (accuracy, 0.5 degrees). For a subset of toddlers, a flashing star with a chime appeared for 5.00 sec prior to the start of the experiment to ensure toddlers was fixated on the screen.

Gaze data were processed in Tobii Pro Lab using a built-in fixation filter (Tobii IV-T fixation filter, velocity threshold: 30 degrees/second). Dynamic areas of interest (AOIs) were drawn in Tobii Pro Lab frame by frame and grouped on a scene-by-scene basis. Background AOIs included the white space that made up the background of every scene and the empty table space not occupied by objects present in the scene.

Gaze data were exported and analyzed offline. First, timestamp information obtained from Tobii Pro Lab "raw data exports" was used to match webcam participant recordings to gaze data processed by Pro Lab. After grouping AOIs into Social vs. NonSocial categories, all scenes were collapsed together and AOI hit information from raw data exports was used to determine percent fixation within Social and NonSocial AOIs across the entire duration of 'The Joint Attention Test'. For this, a value of -1 indicates the AOI was inactive; a 0 indicates the AOI was active but gaze data did not overlap with an AOI; and a 1 indicates an AOI was active and the toddler's gaze overlapped with the AOI coordinates. The Social AOI group included the Face AOI, while the NonSocial AOI group included all target and distractor objects present across all scenes except the Background_Wall AOI. For example, across all 2,625 frames, if a toddler spent 678 of those frames gazing at the Face_AOI, 110 of those frames gazing at the Background_Wall and Background_Table, and his total gazing captured by eye tracker is 2,300 frames, then the total Social percent fixation would be 33.9% (678/[2,300-110]) and the total non-Social percent fixation would be 60.6% ([2,300-110-678]/[2,300-110]).

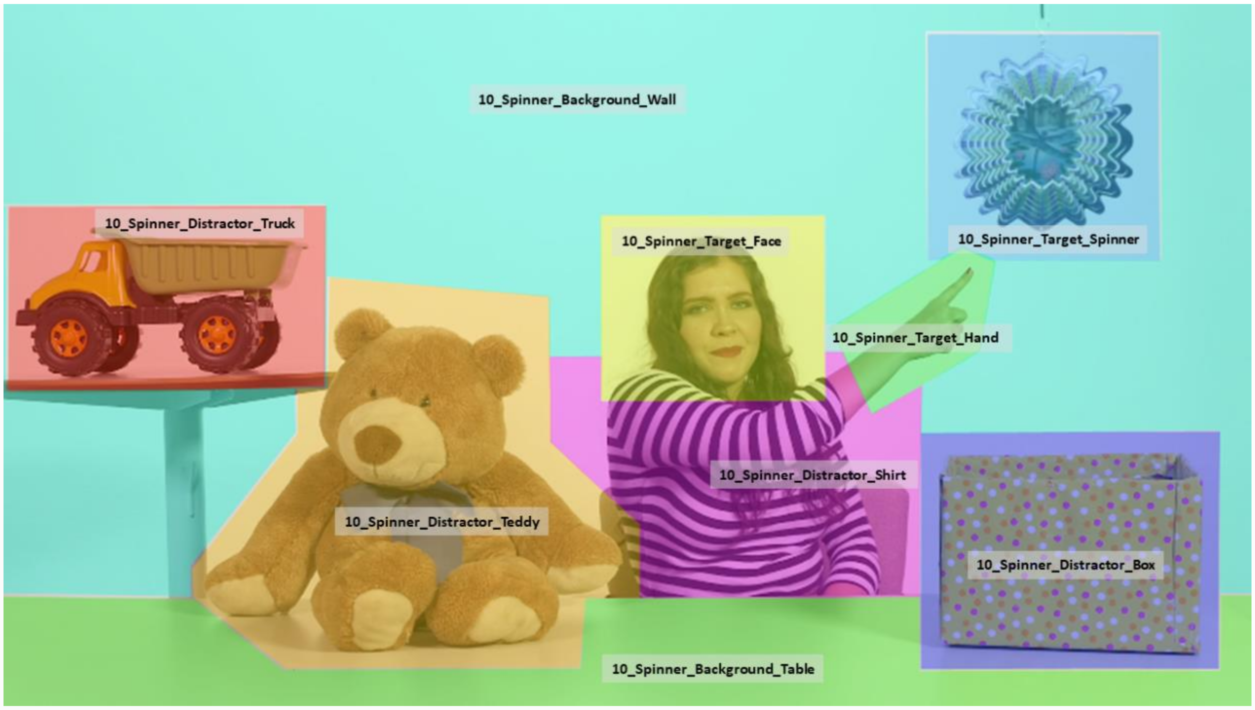

10_Spinner_Background_Wall
10_Spinner_Target_Spinner
10_Spinner_Distractor_Truck
10_Spinner_Target_Face
10_Spinner_Target_Hand
10_Spinner_Distractor_Shirt
10_Spinner_Distractor_Teddy
10_Spinner_Distractor_Box
10_Spinner_Background_Table

**Figure 3.2** A sample setup for the Spinner scene from the 87.6 sec (19 scenes) “The Joint Attention Test”.

**Facial expression data collection and video data processing**

A Logitech HD Pro Webcam C920 (1080p, Carl Zeiss, Newark, CA) was installed beneath the Tobii Pro Spectrum eye tracker and stimulus presentation screen was used to record toddler facial expressions throughout the duration of ‘The Joint Attention Test’ (webcam resolution: 1920x1080, 30 frames per second). While participant videos had a range of total frames (between 2700 – 3500), 2625 of the total frames make up the duration of ‘The Joint Attention Test’.

The 87.5-second long video contained 2,625 frames (30 frames per second). The length of raw videos recorded by Tobii Pro Eye tracking System ranges from 2,700 to 3,000 frames. The timestamp from the Tobii data file (exported from Tobii pro lab) is used to match the recorded face video with the test videos.

**Facial expression recognition**

Action unit detection was conducted using OpenFace 2.0 GUI (Baltrusaitis et al. 2018). We chose the Multi-task Convolutional Neural Network (MTCNN) (K. Zhang et al. 2016) for face detection and CE-CLM for action unit estimation (Baltrušaitis, Mahmoud, and Robinson 2015; Baltrusaitis et al. 2018). The absolute measurement of the action unit (static mode) and personal-adjusted measurement (dynamic mode) were both tested (Baltrušaitis, Mahmoud, and Robinson 2015; Baltrusaitis et al. 2018). For the arousal and valence index, the fan model recorded the facial landmarks (Bulat and Tzimiropoulos 2017) and fed them into emonet (Toisoul et al. 2021). The emonet model outputs the valence and arousal index. The index was further adjusted to remove the baseline (Haque n.d.; Z.-M. Zhang, Chen, and Liang 2010). Since a video has 2,625 frames, the raw data has 19 action units x 2,625 frames for each subject (Figure 3.3a).

**Eye-tracking data analysis**

The eye-gazing data were exported from Tobii-Prolab. The AOI was manually drawn and included a wall, table actress’ Face, Spinner, Box, Comb, Teddy, Scarf, Cup, Hand, and Bow. Excluding the background, the actress's face was categorized as the social AOI and the other subjects were categorized as the non-social AOI. The eye-gazing data were matched to 2,625 frames and the proportion of time spent on AOIs was calculated. Then, the time difference between social AOI and non-social AOI was calculated for a subject. The independent t-tests were performed to compare the time difference distribution (between social AOIs and non-social AOIs) between the two groups (Figure 3.4 a-d). 161 subjects have 'Geo-Pref' non-social fixation score and we compare them with non-social fixation time on 'The Joint Attention Test' (Figure 3.4e).

### Examination of levels of fixation within each AOI and action unit intensity score

For each subject, we calculated the average intensity of social and non-social AOIs for all action units. We find action units' intensity has no significant difference between social and non-social AOIs using the independent T-test.

### Manual Coding

To validate the expression intensity measured by the algorithms, 20 webcam recordings (13 ASD, 7 non-ASD) were randomly selected and manually coded for the overall facial muscle movement intensity. For the validation, we first examined if the algorithm captured the action unit movements against the neutral face baseline. Second, we tested if the action unit intensity value aligned with the overall expression level such as neutral, mild, normal, and extreme big expression coded by the coder.

We examined if the facial muscle was actively engaged and coded the arousal part of facial expression. Arousal (or intensity) is the level of autonomic activation that an event creates and ranges from calm 0 (or low) to excited 3 (or high). The neutral face is scored as 0; a mild arousal face as 1; a moderate arousal face as 2; an intense arousal face as 3. The independent coders watched the videos and recorded scores for every 5 seconds windows. The following action units were observed: cheek raiser, chin raiser, inner brow raiser, outer brow raiser, dimpler, lips part, lip corner depressor, lip corner puller, upper lip raiser, lip stretcher, lid tightener, and upper lid raiser (Farnsworth 2019).

A neutral face 0 is a neutral baseline that has no action units movement mentioned above. It is okay to ignore the mouth closeness (no matter whether closed or opened) since some neutral face conformations have the mouth open. A 1-score moderate face can be any sudden arousal expression. The clues are at least one "actively" involved action unit related to the chin, eyebrow, lip, and lid action units mentioned above. 'Actively' means one can see that part of the muscle is working and has increased the tension. The muscle movements in face 1 are gentle and fade quickly. For example, the subject might slightly pull the lip corner or stretch the lip. Another example can be a subject opening his mouth slightly larger than the baseline and then relaxing it or closing it without any other movement. Face 2's expression can be easily observed. The clue can be a single action unit activity or multiple action units that work together. For example, a child might pull the lip corner to make his lip stretch 1.5x the baseline or raise the chin to show happiness while the mouth is still closed. The signals can also be a combination of chin drop, eyebrow raiser, lip stretcher, and eyelid raise as mentioned above. Furthermore, a child who uses an eyelid and eyebrow muscle to frown can be associated with confusion. The big smiley/sad face above the normal expression is scored as 3. Many action units can work together with strong signals. For example, the big happy face might include a combination of an intense mouth open, lip corner puller, cheek raiser, and lid tightener.

### Comparing Open Face action unit scores to manual coding

Randomly selected 20 videos with hidden diagnosis information were manually coded by independent research assistants with the coding instructions mentioned above. Three assistants coded 8 videos. Reliability between coders was 82% within each of the coding scores. Then two assistants coded the rest of the 12 videos. The consistency between the two coders was 87%. The scores were compared with the action unit score to see if the trends of both matched. Further, we compared the mean coding scores between ASD and non-ASD subjects by independent T-tests.

### Correlation analysis

Spearman correlation coefficients (normality tests rejected) for 171 action unit pairs were calculated. The Mann-Whitney-U test (normality tests rejected) was used to test if the ASD (or non-ASD) group correlations were different from the zeros (B-H FDR correction α<0.05, p<0.05). The Mann-Whitney-U test (normality tests rejected) was used to compare if the correlations were different between the ‘ASD’ and ‘non-ASD’ groups.

### Aggregating the action unit intensity score

The videos included 19 scenes, "01_HiSweetie", "02_Teddy", "03_Bow", "04_Box", "05_Comb", "06_PointyComb", "07_MessyHair", "08_GreatHair", "09_Wind", "10_Spinner", "11_Scarf", "12_GreatScarf", "13_WarmCozy", "14_Thirsty", "15_DrankWhole", "16_GoHome", "17_LookTruck", "18_TeddyDrive", and "19_ByeBye". The length of scenes ranged from 66 frames to 363 frames. The 2,625-frame video was separated into 19 scenes (Figure 3.3a). The maximum values of each action unit were collected within each bin, and the variance was calculated. The subject has binned intensity data for 19 scenes (Figure 3.3c). For each action unit, the mean of maximum values between two groups was calculated in each scene (Figure 3.5). Then, the paired t-test (with B-H FDR correction) was used to compare the difference between the means (of maximum values) for 19 scenes.

### Co-expression analysis and correlation score

To test the correlation network activity in two groups, we selected 97 correlation pairs that have Pearson correlation significantly larger than zero (p-value>0.05) (Figure 3.6e). Among the non-ASD group, the mean of each correlation was used as the background. For each subject, the correlation of these 97 pairs was compared with the background distribution using the dependent t-test (normality test accepted) in the scipy package (Virtanen et al. 2020). The co-expression level was defined as the z-score from the dependent T-test. Positive scores imply that at least some interacting action unit pairs were significantly higher than chance and hence parts of the network were potentially active.

**Model building**

The 184 subjects were split into training (108 ASD and 46 non-ASD) and test datasets (21 ASD and 9 non-ASD). To check the age effect, the optbin R package (“Optbin: Optimal Binning of Data” n.d.) was used to find the optimal age breakpoint (Figure 3.3f). An Independent T-test was used to check if there was a difference between groups.

**Feature generation**

The first step was to remove the action units that had low signal detection. The action units with low signal and low variance were removed. The action unit signal with 66% of subjects not detected (score lower than 0.1) was thus removed. 25 features were selected that contained 17 action units and 5 variances ('var_1_Inner Brow Raiser', 'var_17_Chin Raiser', 'var_23_Lip Tightener', 'var_25_Lips part', ‘var_26_Jaw Drop') and 3 emotion statues (arousal, valence, and var_valence) across 19 scenes (19x25=475 features in total). The data were scaled by each feature with StandardScale from scikit-learn (Trappenberg 2019). Then, the 97 correlation features were incorporated (Figure 3.3e).

**Feature selection and cross-validation**

Five-fold cross-validation was performed on the training set with 154 samples (108 ASD and 46 non-ASD). The dataset was shuffled and separated into six folds. Five folds of data were used for training, and one fold was left out for the test. The feature selection methods include linearSVC (penalty=’l2’) and ElasticNet. The feature numbers were forced to be less than the sample size. SVM with the linear kernel was used as a final classifier. For each training fold, the ROC score and a set of coefficient weights for selected features were recorded. After five-fold training, the average ROC score was used to select the top model. Then, the recorded five sets of feature weights were accumulated, and the first 70% of ranked features were selected as the final feature set. We were able to obtain a 0.86-0.90 AUC-ROC score. We also validated with the LogisticRegression classifier that the final model reaches 0.85-0.88 AUC-ROC scores on the training set, demonstrating the model behavior's consistency.

## Results

### Data analysis overview

Figure 3.1 outlined the main design and analysis steps, and Figure 3.3 provided details of data engineering for machine learning classification. After the experiment, the data were exported from the Tobii and webcam software, and timestamps for 19 scenes were tabulated for each subject (Figure 3.3a). The action units were measured for 2,625 frames and had an individual baseline adjusted with OpenFace 2.0 (Baltrušaitis, Mahmoud, and Robinson 2015; Mavadati et al. 2013; Jeni et al. 2013; Baltrušaitis, Robinson, and Morency 2016). The valence and arousal index were measured with emonet (Toisoul et al. 2021) and also had baseline removed (Z.-M. Zhang, Chen, and Liang 2010) (see Methods). The correlation and co-expression of the eighteen action units and two valence/arousal scores were analyzed to generate correlation-related features (correlation and co-expression scores) (Figure 3.3b, d). Then, the action unit and valence/arousal were aggregated into 20 scenes and the means and variances were calculated (Figure 3.3c). We then selected 97 correlation-related and 475 intensity-related features and prepared these as subject features for feature selection and model training (Figure 3.3e). Meanwhile, the manual coding instruction was designed to validate if the automatic action unit detection could capture the action unit movement.

### Data quality

We tested first for group (129 ASD and 55 non-ASD) differences in age, gender, race (white and non-white), ethnicity (Hispanic/Latino and not Hispanic/Latino), and clinical scores. There were no significant group differences in age (two-sample independent T-test, t-statistic=1.606, p=0.110), reported gender (Fisher exact test, ratio=0.832, p=0.5969), reported race (Fisher exact test, ratio=0.821, p=0.617) or ethnicity (171/184 collected, Fisher exact test, ratio=1.303, p=0.501). (Table 3.1 and Table 3.2)

**Table 3.1 Demographic information table for age and clinical scores.**

| | ASD | non-ASD | T-statistic | p-value |
|---|---|---|---|---|
| Age | 28.6 (7.94) | 26.46 (9.05) | 1.606 | 0.109966 |
| Ados_CoSoTot | 13.55 (4.23) | 4.25 (3.65) | 14.203 | 2.69E-31 |
| Ados_RRTot | 5.57 (1.83) | 1.58 (1.77) | 13.659 | 1.07E-29 |
| Ados_CoSoTotRRTot | 19.12 (5.3) | 5.84 (4.98) | 15.833 | 4.59E-36 |
| Vine_ComTotal_DomStd | 78.31 (18.9) | 94.49 (10.1) | -5.989 | 1.1E-08 |
| Vine_DlyTotal_DomStd | 83.07 (14.11) | 94.15 (8.89) | -5.378 | 2.29E-07 |
| Vine_SocTotal_DomStd | 82.97 (14.81) | 94.73 (9.07) | -5.461 | 1.54E-07 |
| Vine_MtrTotal_DomStd | 91.12 (13.43) | 93.69 (20.8) | -0.998 | 0.319617 |
| Vine_AdapBehav_DomStd | 80.26 (13.39) | 92.67 (9.18) | -6.268 | 2.58E-09 |
| Vine_DomStdTotal | 335.24 (51.36) | 378.62 (36.99) | -5.665 | 5.66E-08 |
| Mullen_VRT | 35.17 (13.43) | 46.44 (12.73) | -5.289 | 3.51E-07 |
| Mullen_FMT | 35.16 (11.8) | 46.93 (11.56) | -6.227 | 3.21E-09 |
| Mullen_RLT | 29.37 (16.68) | 44.89 (13.28) | -6.119 | 5.63E-09 |
| Mullen_ELT | 26.16 (15.51) | 43.22 (15.85) | -6.787 | 1.56E-10 |
| Mullen_ELC_Std | 67.35 (22.18) | 91.91 (20.07) | -7.069 | 3.22E-11 |

**Table 3.2 Contingency table for gender, race and ethnicity.**

| | ASD | non-ASD | Ratio | p-value |
|---|---|---|---|---|
| Gender | | | | |
| Female | 35 | 17 | | |
| Male | 94 | 38 | 0.8323 | 0.5969 |
| Hispanian or Latino | | | | |
| HL | 51 | 19 | | |
| non-HL | 68 | 33 | 1.303 | 0.5006 |
| Race | | | | |
| White | 81 | 37 | | |
| non-white | 48 | 18 | 0.8209 | 0.6168 |

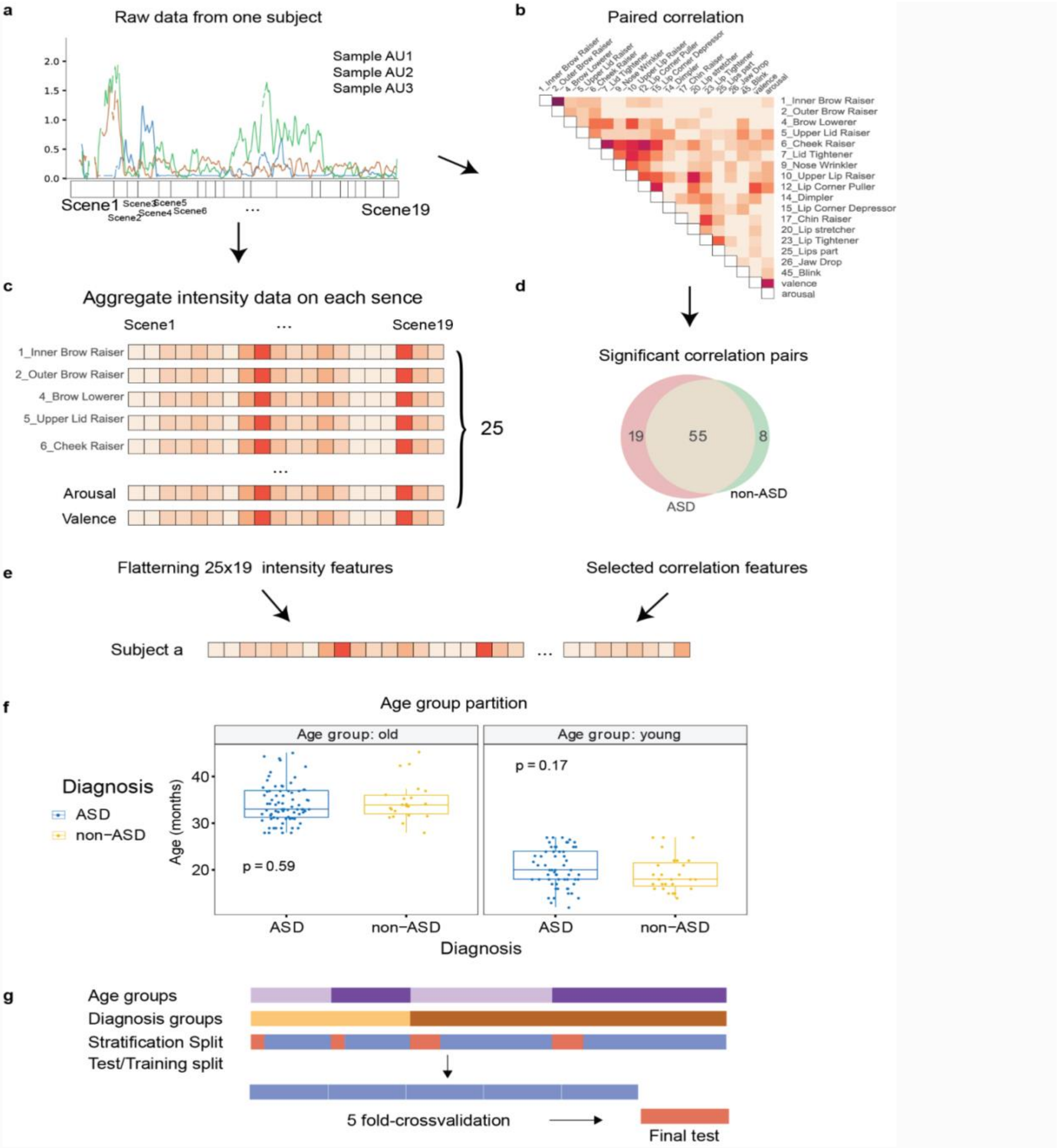


**Figure 3.3 Feature preparation pipelines. a** Sample raw data with scene timestamps annotated. **b** Correlations of each action unit pair through the whole experiment. **c** Raw action unit values were aggregated into 19 scenes. **d** Selected correlation pairs. **e** The input features consist of action unit intensity features and correlation features. **f** Subjects were partitioned into young and old age groups. **g** The stratification sampling based on age (<27.5 months, light purple;>27.5, dark purple) and diagnosis labels (non-ASD, cream; ASD, brown) helps to split the training (blue) and test (red) groups. The models were trained and selected based on the 5-fold cross-validation and finally evaluated by the final test group.

### Manually coded arousal muscle movement

An in-house manual coding system was designed to investigate if the tools used could capture facial muscle movement. Our coders are trained to code the overall arousal muscle movements in 4 levels (see Methods). 20 out of 184 subjects were randomly selected for manual coding (13 ASD and 7 non-ASD). Three different coders checked 8 out of 20 videos to guarantee consistency of the coding instruction—all three raters agreed upon 80% of codings. Two coders evaluated the remaining 12 of the 20 videos. The toddlers generally express happiness-related emotions but might express non-happy-related emotions, such as negative surprise, shock, and sadness. Based on the manual coding, the children have more action unit movement in the ASD group than the non-ASD group (two-sample independent t-test, p-value=0.014). The overlaps between manually coded facial expression intensity scores and computer-vision-based intensity peaks were matched. The results validated that the computer-vision-based algorithm can capture facial muscle movement from toddlers (See Methods).

### Visual preference on the area of interests (AOIs)

Although our analysis focused on facial emotional behavior, the eye-tracking data provide valuable information about subjects' visual behavior on quality control steps (See Methods). The subject's eye-gazing on 12 AOIs was recorded. The AOIs on the videos were categorized as social (actress's face), non-social (Table, Box, Shirt, Spinner, Truck, Teddy, Bow, Comb, Scarf, Cup and Hand), and background (Wall and Table) AOIs (Figure 3.2). In Figure 3.4a, non-ASD have more fixation time on the screen (social, non-social) (Failed normality test, p-value=1.012e-10; Wilcoxon rank-sum test, p-value=0.0102). The ASD group spent more time than the non-ASD group on the target AOI (Two-sample independent T-test, p-value=2.6e-06), while the non-ASD group spent more time on social AOI than the ASD group (Two-sample independent T-test, p-value=2.6e-06) (Figure 3.4 a,b). The paired comparison of social AOI and non-social AOI provided the same result (Two-sample independent T-test, p-value=2.6e-06, Cohen's d=0.8205) (Figure 3.4 d). The Games Howell Post-hoc Tests (“RPubs-Games-Howell Nonparametric Post-Hoc Test” n.d.) on differences of AOI were also conducted on the optimal partitioned age groups (Figure 3.3f) to validate that the age effect was insignificant over different age groups. The result was consistent with our prior findings on the ‘GeoPref’ experiment (See Method; Figure 3.4e, r=0.51). Part of ASD subjects preferred non-social objects to social objects(Wen et al. 2022).

### Group differences in action unit intensity

After the 2,625 frames were binned into 19 scenes (see Methods), within each bin, the maximum values of each action unit were collected and the variance was calculated. The descriptive statistics of maximum value were provided. Across 19 scenes, 16 out of 17 action units in the ASD group had higher mean intensity than non-ASD groups (two-sided dependent T-test and B-H FDR correction) (Table 3.3 and Figure 3.5). Additionally, the ASD group has higher valence (p=3.841e-08) and arousal (p=8.580e-09) scores (two-sided dependent T-test).

### Group differences in action units co-expression patterns

Video-based action units (including valence and arousal index) correlation was calculated and the correlation network activities were evaluated. 171 pair-wise correlation (named as c-pair) values were calculated for each subject (see Methods) (Figure 3.6a,b). 89 out of 171 c-pairs from the ASD group were significantly larger than zero (Figure 3.6c). 68 out of 171 c-pairs from the non-ASD group were significantly larger than zero (Figure 3.6d). 65 c-pairs were larger than zero in both ASD and non-ASD groups (Figure 3.6e). Within both groups, the action units such as Brow Lowerer, Cheek Raiser, Eyelid Tightener, Nose Wrinkler, Upper Lip Raiser, Lip Corner Puller, Dimpler, and Lip stretcher were correlated with each other and formed a hub in the correlation network. Correlation tables and volcano plots showed that, among the significant c-pairs, the correlation values from the ASD group were numerically larger than that of non-ASD groups (Figure 3.6c,d). Ten c-pairs have significant p-values (Mann-Whitteny-U with B-H FDR correction).

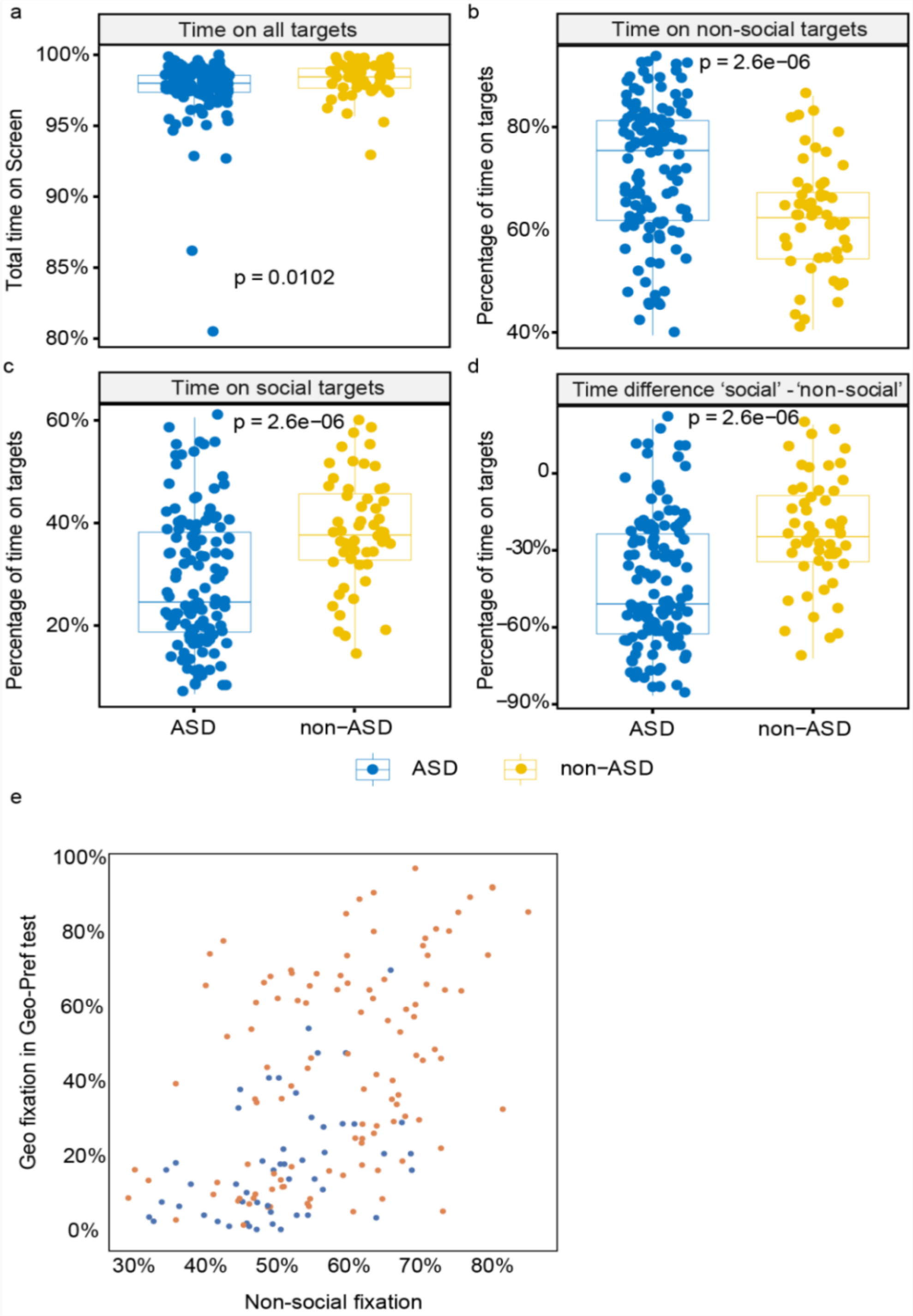


**Figure 3.4 Time spent on Area of Interests. a** The total time spent on the screen. **b-c** Proportion of the fixation time over the 'social', and 'non-social' targets. **d** The time difference between the 'social' and 'non-social' areas. **e** scatter plot for 'Geo-Pref' non-social fixation score vs non-social fixation time on 'The Joint Attention Test".

### Derived co-expression z-score among ASD and non-ASD groups

We further investigated the subjects' action units' co-expression patterns. We selected 97 c-pairs whose mean values were significantly larger than zero in either the ASD or non-ASD group. The 97 c-pair values from a subject were compared against the background's 97 c-pairs using the dependent T-test (normality test not rejected) (see Methods). The resulting z-score was then used to represent a subject's action unit co-expression level. The ASD group has a higher co-expression score than the non-ASD group (p=0.00834, Cohen's d=0.45) (Figure 3.5f).

### A machine learning model differentiates ASD vs. non-ASD toddlers

A classifier was built to differentiate the ASD and non-ASD groups. Four hundred and seventy-five action unit intensity features (19 scenes of 25 action unit intensity features) and 98 whole-video correlation features were used as the input feature for classification (Figure 3.3a-e). The 184 subjects were split into six batches. During the data splitting, the age effect was considered. The optimal two age bins were (12, 27.5) and (27.5, 48) (Figure 3.3f). The StratifiedKFold from scikit-learn(Pedregosa, Varoquaux, and Gramfort, n.d.) that considers age groups and diagnosis labels was used for data splitting (Figure 3.3g). The five-sixth of 184 subjects were used to train models with five-fold cross-validation, and the remaining subjects were isolated as an untouched test set to evaluate the model performance.

The final model used ElasticNet as the feature selection method and used SVM with the linear kernel (a=0.1) as the final classifier (See Methods). The model reached an average accuracy of 86% (0.91 AUC-ROC scores, 0.95 AUC-PR scores) in 5-fold cross-validation analysis (Figure 3.7a,b) and an accuracy of 77% on the test dataset (0.75 AUC-ROC scores, 0.78 AUC-PR ) in the test set. The prediction scores from both training and test datasets were provided (Figure 3.7e). There were 80 features selected as the final feature set. To evaluate the feature importance, these 80 features were binned into the action unit category and scenes category (Figure 3.6c, d). The top features that accounted for half of the weights were the inner brow raiser, lip tightener, the variance of the valence score, jaw drop, upper lid raiser, lip stretcher,

**Table 3.3 The statistical descriptive value of the maximum action unit intensity score across 19 scenes.**

| | ASD | non-ASD | stats | p-value | adj p-value | Cohen's d |
|---|---|---|---|---|---|---|
| 1_Inner Brow Raiser | 0.564 (0.279) | 0.467 (0.254) | 3.893 | 9.777E-04 | 1.161E-03 | 0.363 |
| 2_Outer Brow Raiser | 0.345 (0.235) | 0.257 (0.198) | 6.262 | 5.174E-06 | 9.831E-06 | 0.403 |
| 4_Brow Lowerer | 0.171 (0.08) | 0.076 (0.06) | 10.664 | 1.851E-09 | 1.172E-08 | 1.347 |
| 5_Upper Lid Raiser | 0.194 (0.135) | 0.184 (0.12) | 1.484 | 1.542E-01 | 1.542E-01 | 0.077 |
| 6_Cheek Raiser | 0.279 (0.127) | 0.206 (0.114) | 8.121 | 1.341E-07 | 3.640E-07 | 0.600 |
| 7_Lid Tightener | 0.485 (0.173) | 0.407 (0.166) | 6.492 | 3.208E-06 | 6.773E-06 | 0.462 |
| 9_Nose Wrinkler | 0.262 (0.161) | 0.223 (0.155) | 4.464 | 2.663E-04 | 3.373E-04 | 0.247 |
| 10_Upper Lip Raiser | 0.13 (0.097) | 0.083 (0.062) | 4.810 | 1.217E-04 | 1.652E-04 | 0.580 |
| 12_Lip Corner Puller | 0.265 (0.131) | 0.17 (0.112) | 8.803 | 3.935E-08 | 1.246E-07 | 0.779 |
| 14_Dimpler | 0.332 (0.154) | 0.176 (0.112) | 11.138 | 9.032E-10 | 8.580E-09 | 1.159 |
| 15_Lip Corner Depressor | 0.499 (0.24) | 0.393 (0.19) | 5.110 | 6.227E-05 | 9.859E-05 | 0.488 |
| 17_Chin Raiser | 0.938 (0.263) | 0.727 (0.191) | 9.601 | 1.011E-08 | 3.841E-08 | 0.917 |
| 20_Lip stretcher | 0.377 (0.192) | 0.311 (0.179) | 7.308 | 6.252E-07 | 1.485E-06 | 0.354 |
| 23_Lip Tightener | 0.465 (0.291) | 0.36 (0.239) | 3.734 | 1.407E-03 | 1.486E-03 | 0.394 |
| 25_Lips part | 0.805 (0.286) | 0.671 (0.241) | 5.240 | 4.665E-05 | 8.058E-05 | 0.506 |
| 26_Jaw Drop | 0.82 (0.284) | 0.688 (0.214) | 4.889 | 1.019E-04 | 1.489E-04 | 0.524 |
| 45_Blink | 0.551 (0.248) | 0.454 (0.2) | 3.833 | 1.121E-03 | 1.253E-03 | 0.429 |
| smoothed_valence_mean | 0.262 (0.07) | 0.206 (0.069) | 9.665 | 9.093E-09 | 3.841E-08 | 0.801 |
| smoothed_arousal_mean | 0.277 (0.073) | 0.227 (0.061) | 11.506 | 5.249E-10 | 8.580E-09 | 0.748 |

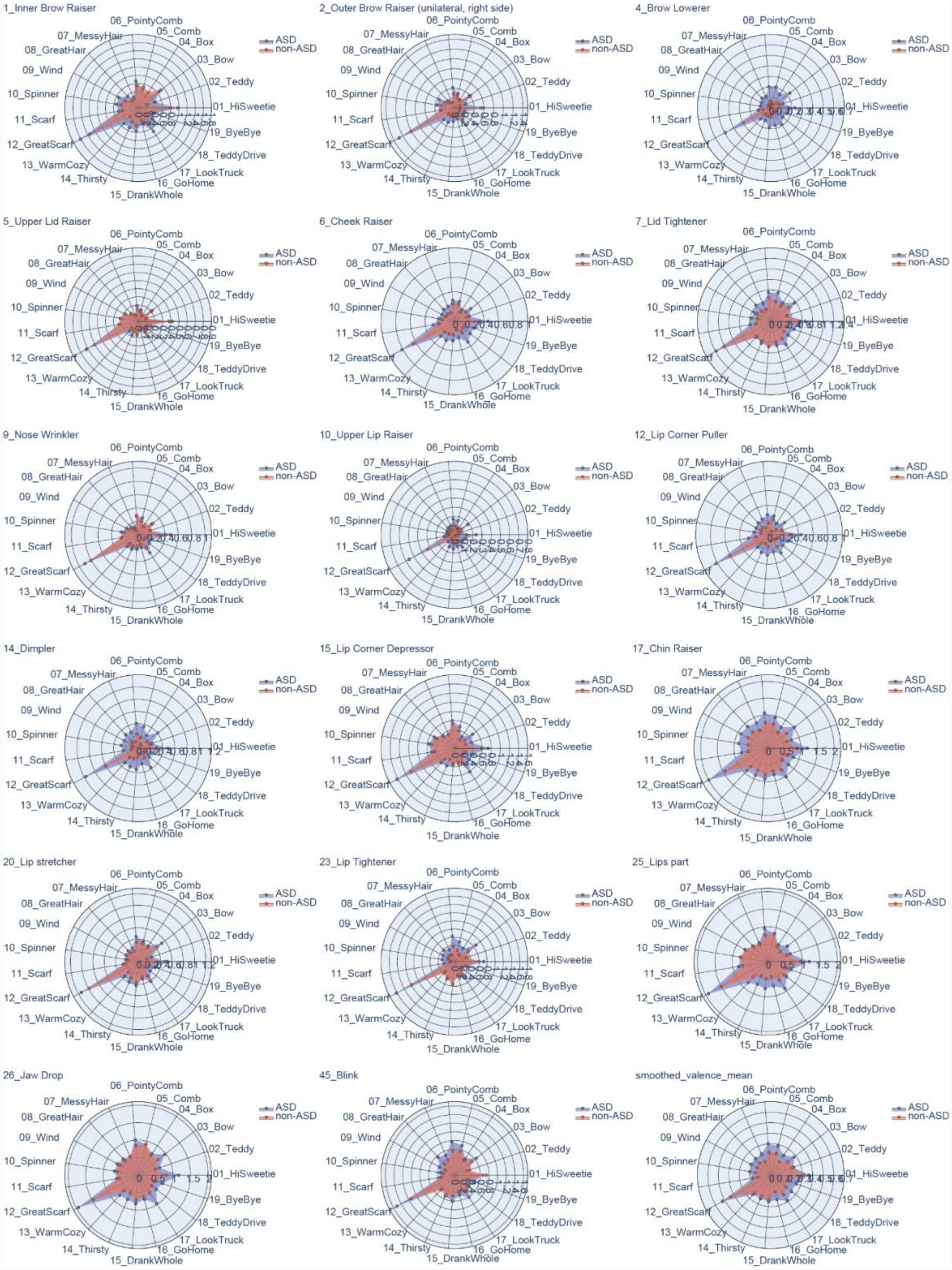


**Figure 3.5 Action unit intensities across 19 scenes.** Each radar plot showed one action unit's mean maximum intensity score for 19 scenes. The red one is the non-ASD group and the blue one is the ASD group.

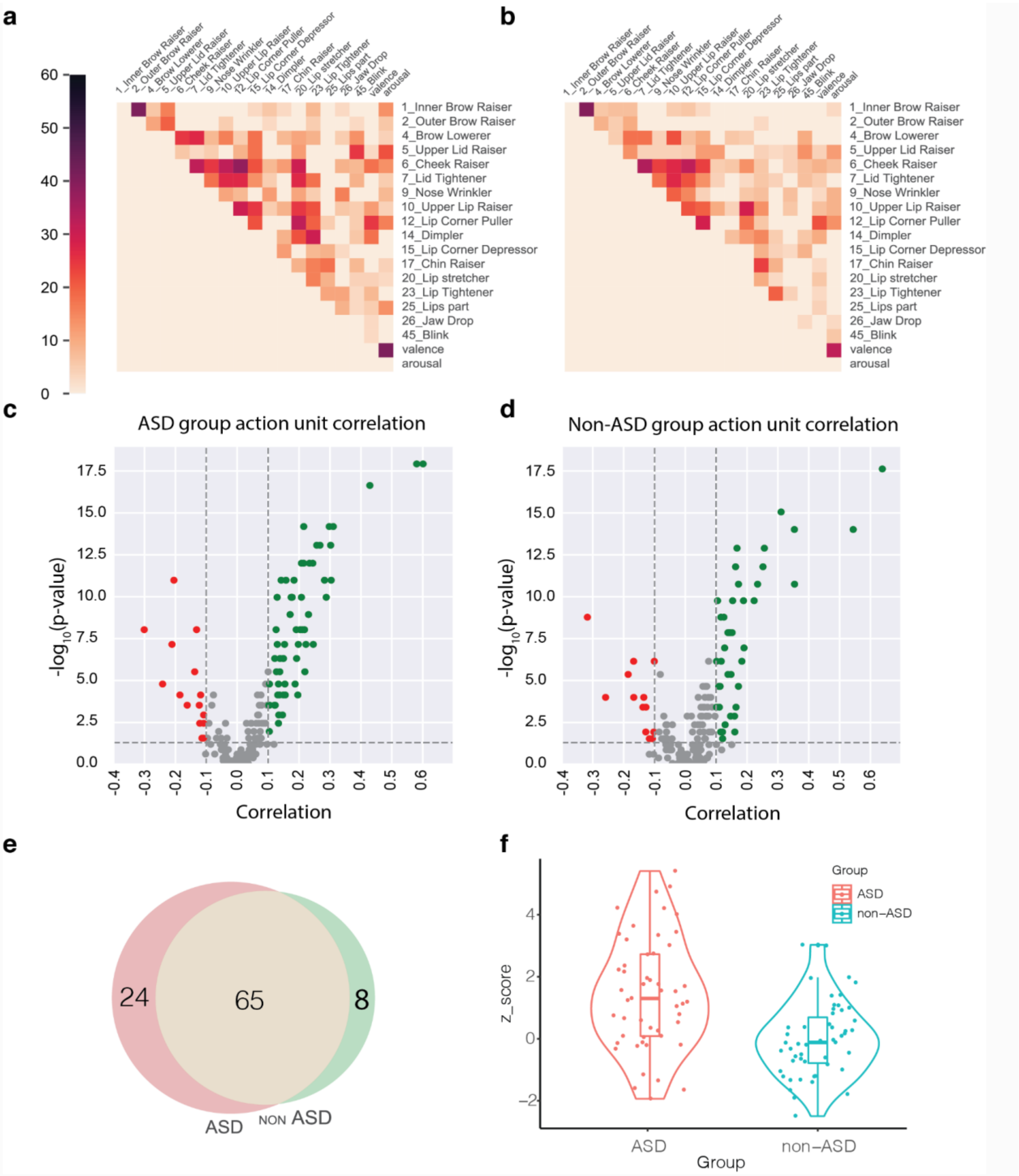


**Figure 3.6 Action unit correlations show co-varying features.** The correlation p-value table for **a** ASD and **b** non-ASD groups**.** The volcano plot of 171 correlation pairs for **c** ASD and **d** non-ASD groups. **e** Venn diagram for 97 correlation pairs positively correlated in either ASD or non-ASD group. **f** The distribution of derived correlation score (see Methods) between ASD and non-ASD groups.

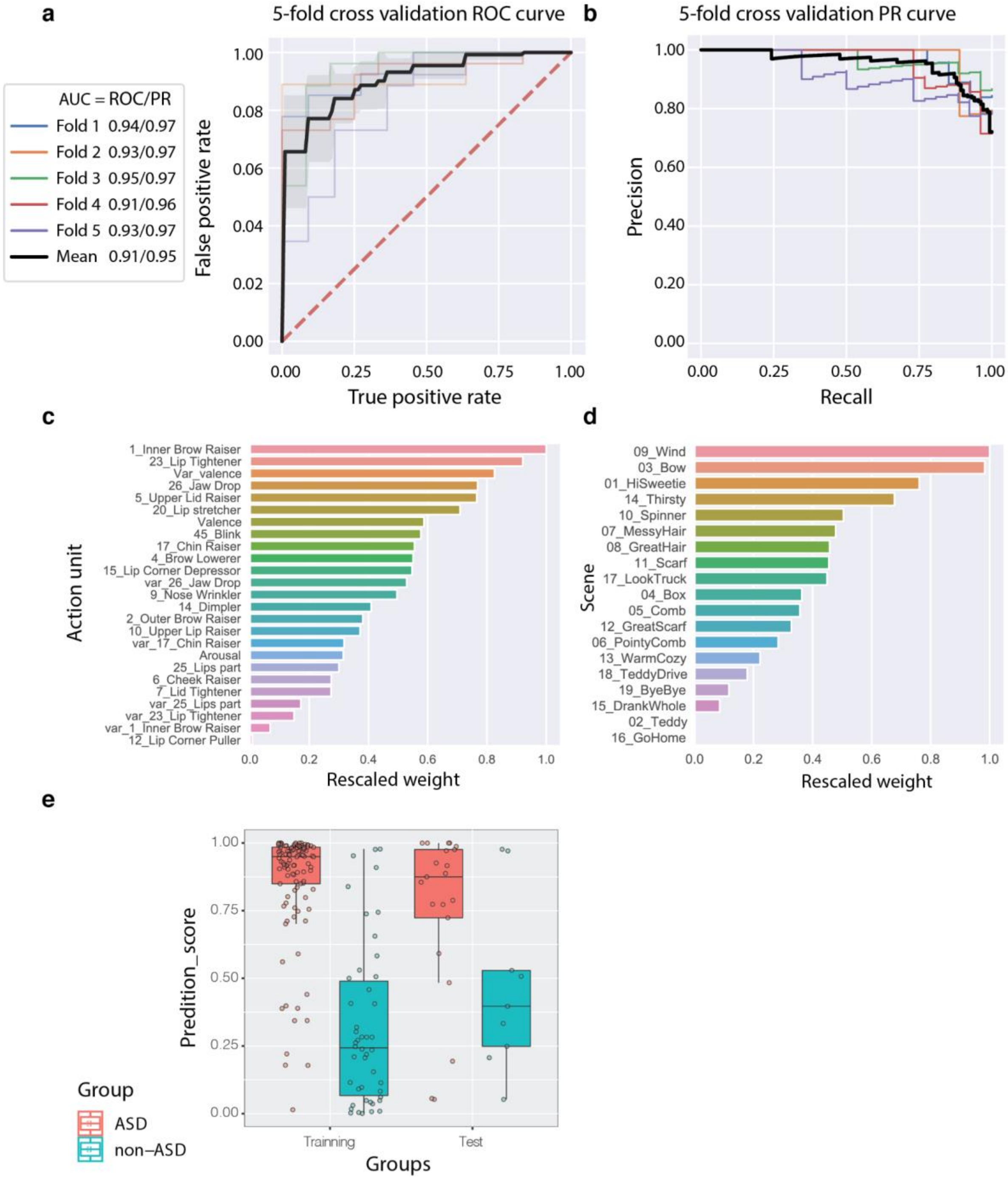


**Figure 3.7 The model result**. The **a** ROC and **b** PR curves of 5-fold cross validation. **c** The selected features' weights were aggregated by action units. **d** The selected features' weights were aggregated by senses. **e** The subject's predictive scores from training and test groups.

blink, and chin raiser. The top senses that accounted for half of the weight were 09_Wind, 03_Bow, 01_Hisweetie, 14_Thirsty, 10_Spinner, and 07_MessyHair.

**Classifier scores unaffected by age, ethnicity, and race differences**

To examine possible bias toward the age effects on group classification, we tested the Pearson correlation between age and classifier score on ASD and non-ASD subjects independently (ASD in Train, r = 0.174, p = 0.072; non-ASD in Train, r = 0.106, p = 0.481; ASD in Test, r = -0.040, p = 0.863; non-ASD in Test, r = 0.217, p=0.574). This further verified that potential confounding effects of age were excluded in the analysis.

The examination of classifier scores on ASD and non-ASD groups showed no significant difference across ethnicity groups (‘reported Hispanic and Latino’, ‘reported not Hispanic or Latino’; independent T-test) (Table 3.4). The same analysis was also conducted on gender. No significant difference was found between the male and female groups (independent T-test) (Table 3.5).

Post-hoc examination of age, ethnicity, and gender analysis indicated that the classifier has no bias against those common confounding factors.

**Ensemble classifier combining GeoPref non-social score significantly improved sensitivity**

We had 161 out of 184 subjects with GeoPref test non-social fixation score(Wen et al. 2022). The subjects who were identified as GeoPref subtype (Wen et al. 2022; Pierce et al. 2016) were 100% correctly labeled as ASD by the action unit classifier (Figure 3.8a). Taking ASD as the true label, our action unit classifier has 83.3% sensitivity and 67.5% specificity in the test dataset (90.1% and 75% in the training dataset). The identifier solely based on the GeoPref test had 95% specificity and 23% sensitivity. By combining the action unit classifier and GeoPref classifier, we adjusted the threshold (GeoPref score: 30% to 41% and action unit classifier predictive score=0.3) to achieve tentative 100% specificity and 45-50% sensitivity. The ensemble classifier preserves the high specificity while drastically increasing the sensitivity which provides a high potential for wide screening applications.

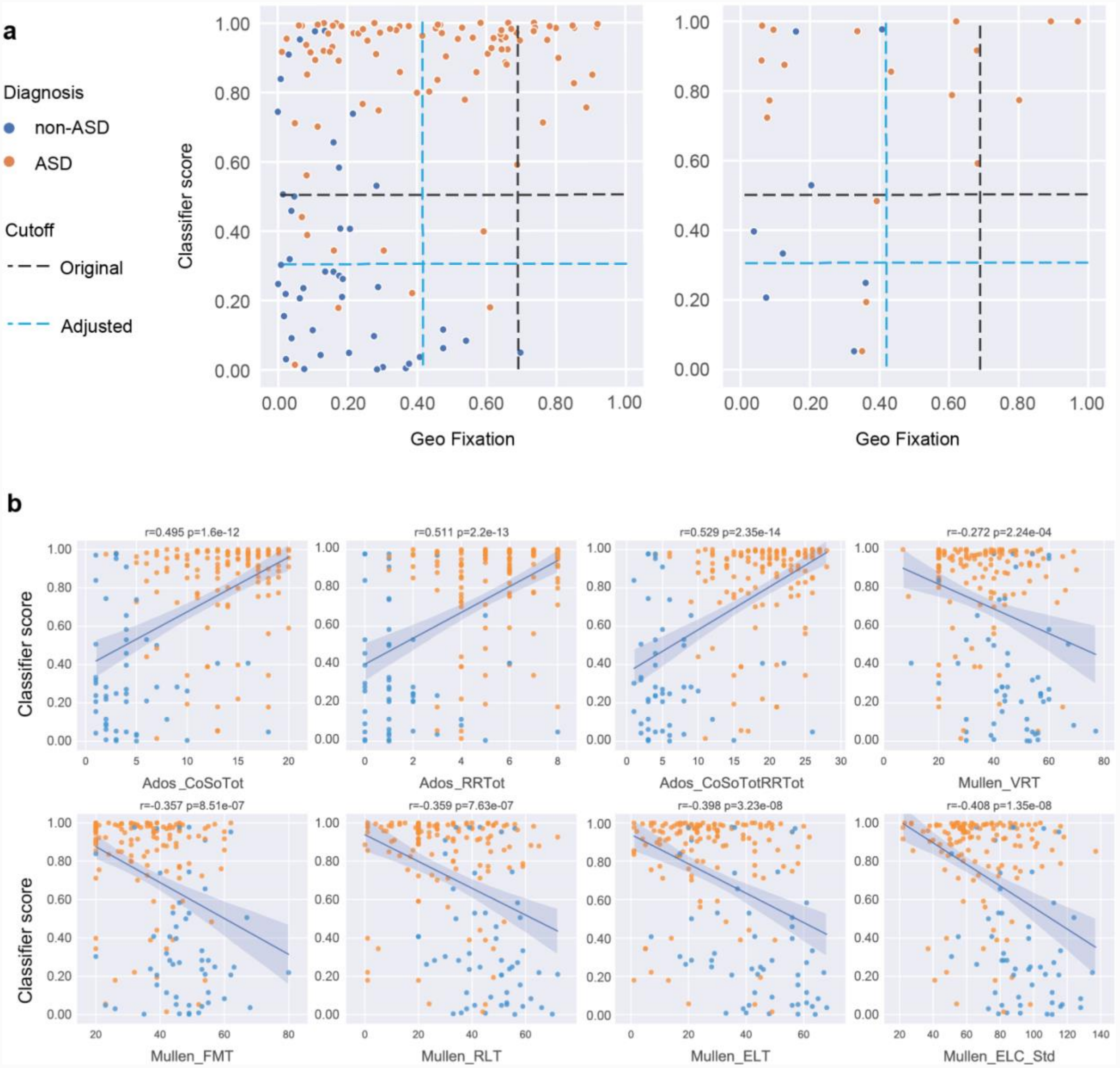


**Figure 3.8 analysis across classifier score, social/communication ability and GeoPref non-social test score. a** the classifier score vs GeoPref score. The left is the training dataset and the right is the test dataset. The lack of horizontal **b** the classifier score vs ADOS and Mullen scores.

**Table 3.4 The statistical descriptive value of gender and ethnicity separately.**

| Diagnosis | Group | Gender | Mean | Std | Count | T-statistic | P-value |
|---|---|---|---|---|---|---|---|
| ASD | Test | F | 0.8671 | 0.2160 | 5 | 0.9412 | 0.3584 |
| ASD | Test | M | 0.7186 | 0.3283 | 16 | | |
| ASD | Train | F | 0.8702 | 0.2011 | 30 | 0.2585 | 0.7965 |
| ASD | Train | M | 0.8586 | 0.2111 | 78 | | |
| non-ASD | Test | F | 0.5293 | | 1 | - | - |
| non-ASD | Test | M | 0.4617 | 0.3434 | 8 | | |
| non-ASD | Train | F | 0.3469 | 0.2721 | 16 | 0.5225 | 0.6039 |
| non-ASD | Train | M | 0.2986 | 0.3112 | 30 | | |

| Diagnosis | Group | Ethics | Mean | Std | Count | T-statistic | P-value |
|---|---|---|---|---|---|---|---|
| ASD | Test | HL | 0.6790 | 0.4073 | 5 | -0.7662 | 0.4555 |
| ASD | Test | non-HL | 0.8058 | 0.2671 | 12 | | |
| ASD | Train | HL | 0.8933 | 0.1582 | 46 | 1.5655 | 0.1206 |
| ASD | Train | non-HL | 0.8280 | 0.2439 | 56 | | |
| non-ASD | Test | HL | 0.7422 | 0.3320 | 2 | 1.4504 | 0.1902 |
| non-ASD | Test | non-HL | 0.3912 | 0.2965 | 7 | | |
| non-ASD | Train | HL | 0.3470 | 0.3276 | 17 | 0.4232 | 0.6744 |
| non-ASD | Train | non-HL | 0.3066 | 0.2917 | 26 | | |

**Table 3.5 The Pearson's correlation coefficient between the classifier score and social/communication ability.**

ADOS Communication Social Score (CoSoTot), restricted and repetitive behaviors severity scores (RRTot), Mullen visual reception (VRT), Fine motor (FMT), Receptive language (RLT), Expressive language (ELT), Early learning composite (ELC).

| | Training dataset (n=154) | | Test dataset (n=30) | |
|---|---|---|---|---|
| | r-value | p-value | r-value | p-value |
| Ados_CoSoTot | 0.5012 | 5.572E-11 | 0.4786 | 8.633E-03 |
| Ados_RRTot | 0.5164 | 1.142E-11 | 0.4809 | 8.272E-03 |
| Ados_CoSoTotRRTot | 0.5361 | 1.298E-12 | 0.4992 | 5.834E-03 |
| Mullen_VRT | -0.2292 | 4.646E-03 | -0.4561 | 1.289E-02 |
| Mullen_FMT | -0.3403 | 1.905E-05 | -0.4400 | 1.691E-02 |
| Mullen_RLT | -0.3541 | 8.161E-06 | -0.3749 | 4.507E-02 |
| Mullen_ELT | -0.3983 | 4.088E-07 | -0.3961 | 3.340E-02 |
| Mullen_ELC_Std | -0.3970 | 4.495E-07 | -0.4561 | 1.289E-02 |

### Post Hoc exploratory analysis across classifier score, social/communication ability and GeoPref non-social test score

We further investigated correlations between classifier score and social and communication abilities. The results showed significant correlations between the classification score and ADOS variables, including ADOS social affect, ADOS restricted and repetitive variables, and ADOS total. The results also showed significant correlations between the classification score and five Mullen variables, including Mullen fine motor, Mullen receptive language, Mullen visual reception, Mullen expressive language and Mullen early learning composite.

We selected the ADOS total, Mullen early learning composite, non-Social preference score and classifier score for the unsupervised clustering analysis. The hierarchical clustering (metric='euclidean', method='average') with distance=2.6 was used and we identified 4 subgroups (Figure 3.9a). The Games Howell Post-hoc Tests was conducted to derive descriptive statistics (Figure 3.9b-e). Cluster 1 (45 ASD) had the highest classifier score, highest ADOS total score and lowest Mullen early learning composite, and all of the GeoPref subtypes (Wen et al. 2022). Meanwhile, cluster 2 (51 ASD, 7 non-ASD) had the second highest classifier score, but lower ADOS total compared to Clusters 1 and 3. Cluster 3 (10 ASD, 4 non-ASD) with a classifier score of lower than 0.5 had no difference from cluster 1 in ADOS total and Mullen early learning composite scores. Cluster 4 (3 ASD and 41 non-ASD) with classifier scores lower than 0.5 had the lowest ADOS total scores and highest Mullen early learning composite scores.

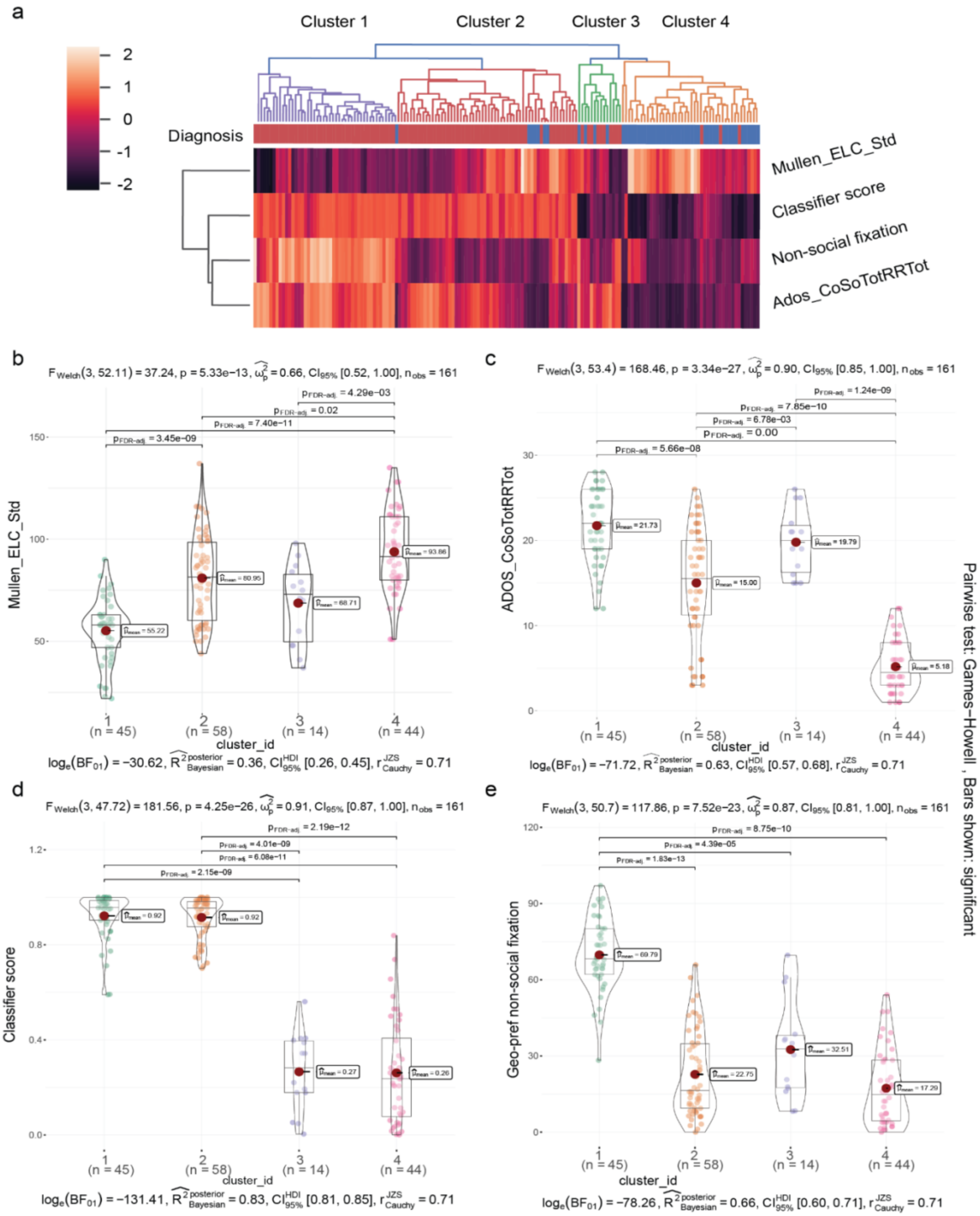


**Figure 3.9 Clustering analysis using classifier scores, ADOS_CoSoTotRRTot, Mullen_ELC_Std and GeoPref non-social test score. a** The hierarchical clustering (metric=’euclidean’, method=’average’) with distance=2.6. **b-e** The Games Howell Post-hoc Tests analysis on classifier scores, ADOS_CoSoTotRRTot, Mullen_ELC_Std and GeoPref non-social test score.

## Discussion

In this study, we determined that in certain situations, emotional responding is overly intense in ASD and the differences can be used as a potential diagnostic marker of ASD by using state of the art expression analysis software. We introduced a highly elicit experiment 'The Joint Attention Test' and tracked the toddler's natural facial expressions while watching the video. With the help of open-source analysis packages (OpenFace 2.0 and Emonet), we measured the action unit intensity on 2,625 frames of video and trained the model based on the action unit intensity. Instead of the raw action unit intensity, we used two sets of features. One set is the action unit intensity against the baseline and the other set is the correlation between action unit pairs across the whole experiment. The model reached an average accuracy of 86% (0.91 AUC-ROC, 0.95 AUC-PR) in the 5-fold cross-validation analysis and an accuracy of 77% on the test dataset (0.75 ROC score). Pearson correlation tests demonstrated the classifier was not correlated with age. The independent *t*-test indicated classifier scores did not differ between the Hispanic versus non-Hispanic subjects or between males and females within ASD and non-ASD groups independently, suggesting the classifier scores independent of toddlers' age, gender, and ethnicity.

Taking ASD as the true label, our action unit classifier has a sensitivity of 83.3% and a specificity of 67.5% in the test dataset (90.1% and 75% in the train dataset). The identifier solely based on the GeoPref test had a specificity of 95% and sensitivity of 23%. By combining the action unit classifier and GeoPref classifier, we adjusted the threshold (GeoPref score: 30% to 41% and action unit classifier predictive score=0.3) to achieve a tentative specificity of 100% and sensitivity of 45-50%. The ensemble classifier preserves the high specificity while drastically increasing the sensitivity, which provides high potential for wide screening application.

It is believed that the ASD and non-ASD groups had different facial behaviors(Carpenter et al. 2021; Brewer et al. 2016; Faso, Sasson, and Pinkham 2015; Weiss et al. 2019; Rozga et al. 2013; Zampella, Bennetto, and Herrington 2020b; Rieffe, Meerum Terwogt, and Stockmann 2000). Here, we found this difference can be captured by the action unit measurement quantitatively and qualitatively. Specifically, a subgroup of ASD subjects were facially over-responsive to the videos, whereas the majority of non-ASD subjects were under-responsive to the videos. This observation is consistent with manual coding. Additionally, the ASD group had no statistical difference in action unit intensity when they were looking at the social and non-social objectives. We also found the ASD group had a higher correlation among certain action unit pairs than non-ASD group (need more detail). It supports that the ASD group has heterogeneous facial expressions patterns and those facial expressions differences captured by the classifier as the top-ranked features had higher value among the ASD groups (need more details).

To gain a comprehensive understanding of social processing and emotional activity, we leveraged eye-tracking and clinical data. We first validated that the subject's non-social preference in the 'The Joint Attention Test' and the 'GeoPref' experiment was consistent (Figure 3.4e). Then, we found that facial expression was highly associated with social and language abilities (Figure 3.8b). Using the data-driven clustering approach, we identified 4 clusters. The predictive score plays an important role in separating clusters 1 and 2 versus 3 and 4. We found subjects in cluster 1 (45 ASD) had distinctive high non-social fixation scores that included all GeoPref subtypes (whose geo-pref non-social score was > 69%) while the rest of the clusters have lower non-social scores and no statistical difference. This might imply the GeoPref subtype has the abnormal facial expression. Cluster 1 also has the highest-responsive facial behavior and the worst social and language skills compared to toddlers in cluster 4 (41 non-ASD, 3 ASD) who have the least-responsive facial behavior and highest social and language skills. Clusters 2 and 3 are intermediate. Cluster 2 (51 ASD and 7 non-ASD) represents a group of toddlers that were highly responsive to the video but have low non-social preference phenotypes. However, we find that cluster 3 (10 ASD and 4 non-ASD) represents a subgroup of toddlers that has facial behavior under-responsive to video but has 2nd worst social and language skills. This cluster represents a very different ASD subgroup that cannot be correctly classified using facial expression detectors even if they have high social and language skills deficits. Overall, the action unit classifier score, eye-tracking score, and clinical score are all important features in clustering.

The clustering result provides a unique perspective to understand the variability within each group. This also gives us new evidence that there might be “over-responsive” and “under-responsive” ASD subgroups with different action unit intensity and clinical scores (Bangerter et al. 2020).

There are a few limitations worth noting. As this study was foundational and exploratory, future work will have to ascertain the reliability and reproducibility of the current results. Another limitation is that the sample sizes in ASD and non-ASD groups (including TD, DD, GDD, LD, etc.) are imbalanced. Since the data were collected during the COVID-19 pandemic period, we were not able to recruit enough TD toddlers in this study and thus included children with TD and delays (e.g., DD, GDD, LD, etc) as the non-ASD group. It is possible that the other neurodevelopmental disorders may be confounding the classification. In fact, we also had to remove more than half of the subjects as their facial movement data are unavailable (wearing masks, covered by hand, eating) during the experiment. The unequal sample size in ASD and non-ASD groups might bias the range of display of AUs associated with ASD. The limited sample size also hampered the model development. The raw data collected by video frames produced 2,625x19 features in a time series manner. However, since we only collected 184 subjects, we have to control the feature size to avoid the overfitting of our model. Although we used the OpenFace and Emonet packages to measure the action unit intensity—these open source and publicly well-known packages may help replicate the findings here, in the future, it would be necessary to develop more sophisticated toddler focused facial expression models to ensure the measurement quality. In addition, the experiment has some disadvantages, although the ‘The Joint Attention Test’ experiment is the most arousal experiment in our inventory. The social and non-social subjects are close to each other, and toddlers tend to shift their attention (eye gazing target) across multiple AOIs on a millisecond scale. Those fast AOI shits make spontaneous emotional analysis on AOI infeasible since the other studies might use a 3-7 seconds window to track the associated emotional response(Riehle, Kempkensteffen, and Lincoln 2017; Zampella, Bennetto, and Herrington 2020b). We also have a disproportion of social and non-social targets. Thus, we cannot conduct sophisticated analyses to identify whether the majority of under-responsive non-ASD is the result of the disproportionate-small social area (Wen et al. 2022).

In conclusion, the use of automatic facial recognition software enabled us to obtain data on facial expression from a relatively larger group of toddlers with ASD (n = 129) in an unobtrusive, accurate, and efficient way. Further, it allowed us to identify clusters or subgroups within the ASD group that differ from each other. These findings provide valuable evidence that the differences in facial expressions help delineate the heterogeneity of ASD with the help of the eye-tracking and clinical data.

Identifying subgroups within ASD may help explain some of the conflicting findings reported in previous studies (Faso, Sasson, and Pinkham 2015; Zane et al. 2018) that may display behavioral differences that are independent of the severity of the diagnosis. It may also provide a standardized and high-throughput way to parse some of the heterogeneity within ASD and enhance understanding of the complex relationship between differences in these subgroups and caregiver reported observations. Our results support the notion of multiple dimensions of observable behavior that contribute to the autism phenotype, and the need to look at behaviors that go beyond the diagnostic criterion to consider profiles of skills across dimensions. This could lead to the personalized interventions and a closer link to the causal pathways associated with ASD phenotypes.

## Acknowledgments

This work was supported by NIMH grant no. R01-MH110558 (E.C., N.E.L.), NIMH grant no. R01-MH080134 (K.P.), NIMH grant no. R01-MH104446 (K.P.), NIMH grant no. P50-MH081755 (E.C.). This work was supported, in part, by a US National Institutes of Health (NIH) grant (R01MH101221).

Chapter 3, in full, is reprint of the paper in preparation. Bao, Bokan, Yaqiong Xiao, Javad Zahiri, Charlene Andreason, Yakta Syed, Summer Zhu, Teresa H. Wen, Eric Courchesne, Nathan E. Lewis, Karen Pierce. Examination of automatic facial action unit measurement as a mechanism to differentiate ASD vs non-ASD toddlers. The dissertation/thesis author was the primary investigator and author of the paper.

## Author Contributions

B.B., N.E.L. and K.R. conceived the project and designed the experiments. C.A. and T.H.W. collected the samples and managed the data. B.B., C.A, Y.S., and S.Z. manually coded the videos. B.B., Y.X., analyzed the data. B.B., Y.X., J.Z., E.C., N.E.L. and K.P.. interpreted the results and wrote the manuscript. N.E.L. and K.P. supervised the project.

## Competing interests

The authors declare no competing interests.

## Code availability

We provide the code library Python described in this work through Github: https://github.com/LewisLabUCSD/actionunit_classifier. We provide Jupyter notebooks in Python to generate our figures and analysis.

## References

Atkinson, Anthony P. 2009. "Impaired Recognition of Emotions from Body Movements Is Associated with Elevated Motion Coherence Thresholds in Autism Spectrum Disorders." *Neuropsychologia* 47 (13): 3023–29.

Autism and Developmental Disabilities Monitoring Network Surveillance Year 2006 Principal Investigators, and Centers for Disease Control and Prevention (CDC). 2009. "Prevalence of Autism Spectrum Disorders - Autism and Developmental Disabilities Monitoring Network, United States, 2006." *Morbidity and Mortality Weekly Report. Surveillance Summaries* 58 (10): 1–20.

Bacon, Elizabeth C., Adrienne Moore, Quimby Lee, Cynthia Carter Barnes, Eric Courchesne, and Karen Pierce. 2020. "Identifying Prognostic Markers in Autism Spectrum Disorder Using Eye Tracking." *Autism: The International Journal of Research and Practice* 24 (3): 658–69.

Baio, Jon, Lisa Wiggins, Deborah L. Christensen, Matthew J. Maenner, Julie Daniels, Zachary Warren, Margaret Kurzius-Spencer, et al. 2018. "Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2014." *Morbidity and Mortality Weekly Report. Surveillance Summaries* 67 (6): 1–23.

Baltrušaitis, Tadas, Marwa Mahmoud, and Peter Robinson. 2015. "Cross-Dataset Learning and Person-Specific Normalisation for Automatic Action Unit Detection." In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 06:1–6.

Baltrušaitis, Tadas, Peter Robinson, and Louis-Philippe Morency. 2016. "OpenFace: An Open Source Facial Behavior Analysis Toolkit." In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–10.

Baltrusaitis, Tadas, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. "OpenFace 2.0: Facial Behavior Analysis Toolkit." In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 59–66.

Bal, Vanessa H., So-Hyun Kim, Megan Fok, and Catherine Lord. 2019. "Autism Spectrum Disorder

Symptoms from Ages 2 to 19 Years: Implications for Diagnosing Adolescents and Young Adults." *Autism Research: Official Journal of the International Society for Autism Research* 12 (1): 89–99.

Bangerter, Abigail, Meenakshi Chatterjee, Joseph Manfredonia, Nikolay V. Manyakov, Seth Ness, Matthew A. Boice, Andrew Skalkin, et al. 2020. "Automated Recognition of Spontaneous Facial Expression in Individuals with Autism Spectrum Disorder: Parsing Response Variability." *Molecular Autism* 11 (1): 31.

Begeer, Sander, Hans M. Koot, Carolien Rieffe, Mark Meerum Terwogt, and Hedy Stegge. 2008. "Emotional Competence in Children with Autism: Diagnostic Criteria and Empirical Evidence." *Developmental Review: DR* 28 (3): 342–69.

Bonnet-Brilhault, Fréderique, Toky A. Rajerison, Christian Paillet, Marie Guimard-Brunault, Agathe Saby, Laura Ponson, Gabriele Tripi, Joëlle Malvy, and Sylvie Roux. 2018. "Autism Is a Prenatal Disorder: Evidence from Late Gestation Brain Overgrowth." *Autism Research: Official Journal of the International Society for Autism Research* 11 (12): 1635–42.

Brewer, Rebecca, Federica Biotti, Caroline Catmur, Clare Press, Francesca Happé, Richard Cook, and Geoffrey Bird. 2016. "Can Neurotypical Individuals Read Autistic Facial Expressions? Atypical Production of Emotional Facial Expressions in Autism Spectrum Disorders." *Autism Research: Official Journal of the International Society for Autism Research* 9 (2): 262–71.

Bulat, Adrian, and Georgios Tzimiropoulos. 2017. "How Far Are We from Solving the 2d & 3d Face Alignment Problem?(and a Dataset of 230,000 3d Facial Landmarks)." In *Proceedings of the IEEE International Conference on Computer Vision*, 1021–30.

Carpenter, Kimberly L. H., Jordan Hahemi, Kathleen Campbell, Steven J. Lippmann, Jeffrey P. Baker, Helen L. Egger, Steven Espinosa, Saritha Vermeer, Guillermo Sapiro, and Geraldine Dawson. 2021. "Digital Behavioral Phenotyping Detects Atypical Pattern of Facial Expression in Toddlers with Autism." *Autism Research: Official Journal of the International Society for Autism Research* 14 (3): 488–99.

Castelli, Fulvia. 2005. "Understanding Emotions from Standardized Facial Expressions in Autism and

Normal Development." *Autism: The International Journal of Research and Practice* 9 (4): 428–49.

Christensen, Deborah L., Kim Van Naarden Braun, Jon Baio, Deborah Bilder, Jane Charles, John N. Constantino, Julie Daniels, et al. 2018. "Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2012." *Morbidity and Mortality Weekly Report. Surveillance Summaries* 65 (13): 1–23.

Cooper, Arnold M., and Robert Michels. 1988. "Diagnostic and Statistical Manual of Mental Disorders, 3rd Ed., Revised (DSM-III-R)." *American Journal of Psychiatry* 145 (10): 1300–1301.

Courchesne, Eric, Vahid H. Gazestani, and Nathan E. Lewis. 2020. "Prenatal Origins of ASD: The When, What, and How of ASD Development." *Trends in Neurosciences* 43 (5): 326–42.

Courchesne, Eric, Peter R. Mouton, Michael E. Calhoun, Katerina Semendeferi, Clelia Ahrens-Barbeau, Melodie J. Hallet, Cynthia Carter Barnes, and Karen Pierce. 2011. "Neuron Number and Size in Prefrontal Cortex of Children with Autism." *JAMA: The Journal of the American Medical Association* 306 (18): 2001–10.

Courchesne, Eric, Tiziano Pramparo, Vahid H. Gazestani, Michael V. Lombardo, Karen Pierce, and Nathan E. Lewis. 2019. "The ASD Living Biology: From Cell Proliferation to Clinical Phenotype." *Molecular Psychiatry* 24 (1): 88–107.

Deschamps, P. K. H., L. Coppes, J. L. Kenemans, D. J. L. G. Schutter, and W. Matthys. 2015. "Electromyographic Responses to Emotional Facial Expressions in 6-7 Year Olds with Autism Spectrum Disorders." *Journal of Autism and Developmental Disorders* 45 (2): 354–62.

Farnsworth, Bryn. 2019. "Facial Action Coding System (FACS) - A Visual Guidebook." Imotions. August 18, 2019. https://imotions.com/blog/facial-action-coding-system/.

Faso, Daniel J., Noah J. Sasson, and Amy E. Pinkham. 2015. "Evaluating Posed and Evoked Facial Expressions of Emotion from Adults with Autism Spectrum Disorder." *Journal of Autism and Developmental Disorders* 45 (1): 75–89.

Frazier, Thomas W., Daniel L. Coury, Kristin Sohl, Kayla E. Wagner, Richard Uhlig, Steven D. Hicks, and

Frank A. Middleton. 2021. "Evidence-Based Use of Scalable Biomarkers to Increase Diagnostic Efficiency and Decrease the Lifetime Costs of Autism." *Autism Research: Official Journal of the International Society for Autism Research* 14 (6): 1271–83.

Gabbay-Dizdar, Nitzan, Michal Ilan, Gal Meiri, Michal Faroy, Analya Michaelovski, Hagit Flusser, Idan Menashe, Judah Koller, Ditza A. Zachor, and Ilan Dinstein. 2021. "Early Diagnosis of Autism in the Community Is Associated with Marked Improvement in Social Symptoms within 1–2 Years." *Autism: The International Journal of Research and Practice*, October, 13623613211049011.

Haque, Md Azimul. n.d. *BaselineRemoval: Python Package Code Repo for BaselineRemoval. It Has 3 Methods for Baseline Removal from Spectra for Baseline Correction, Namely ModPoly, IModPoly and Zhang Fit. The Functions Will Return Baseline-Subtracted Spectrum*. Github. Accessed February 14, 2022. https://github.com/StatguyUser/BaselineRemoval.

Harms, Madeline B., Alex Martin, and Gregory L. Wallace. 2010. "Facial Emotion Recognition in Autism Spectrum Disorders: A Review of Behavioral and Neuroimaging Studies." *Neuropsychology Review* 20 (3): 290–322.

Jacques, Claudine, Valérie Courchesne, Suzanne Mineau, Michelle Dawson, and Laurent Mottron. 2022. "Positive, Negative, Neutral-or Unknown? The Perceived Valence of Emotions Expressed by Young Autistic Children in a Novel Context Suited to Autism." *Autism: The International Journal of Research and Practice*, February, 13623613211068221.

Jeni, László A., Jeffrey M. Girard, Jeffrey F. Cohn, and Fernando De La Torre. 2013. "Continuous AU Intensity Estimation Using Localized, Sparse Facial Feature Space." In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 1–7.

Kaliukhovich, Dzmitry A., Nikolay V. Manyakov, Abigail Bangerter, Seth Ness, Andrew Skalkin, Matthew Boice, Matthew S. Goodwin, et al. 2021. "Visual Preference for Biological Motion in Children and Adults with Autism Spectrum Disorder: An Eye-Tracking Study." *Journal of Autism and Developmental Disorders* 51 (7): 2369–80.

Kaushik, Gaurav, and Konstantinos S. Zarbalis. 2016. "Prenatal Neurogenesis in Autism Spectrum

Disorders." *Frontiers in Chemistry*. https://doi.org/10.3389/fchem.2016.00012.

Kennedy, Daniel P., and Ralph Adolphs. 2012. "Perception of Emotions from Facial Expressions in High-Functioning Adults with Autism." *Neuropsychologia* 50 (14): 3313–19.

Langdell, T. 1978. "Recognition of Faces: An Approach to the Study of Autism." *Journal of Child Psychology and Psychiatry, and Allied Disciplines* 19 (3): 255–68.

Leo, Marco, Pierluigi Carcagnì, Cosimo Distante, Paolo Spagnolo, Pier Luigi Mazzeo, Anna Chiara Rosato, Serena Petrocchi, et al. 2018. "Computational Assessment of Facial Expression Production in ASD Children." *Sensors* 18 (11). https://doi.org/10.3390/s18113993.

LoBue, Vanessa, and Cat Thrasher. 2014. "The Child Affective Facial Expression (CAFE) Set: Validity and Reliability from Untrained Adults." *Frontiers in Psychology* 5: 1532.

Lord, Catherine. 2012. *Autism Diagnostic Observation Schedule, Second Edition (ADOS-2)*. WPS.

Macari, Suzanne, Lauren DiNicola, Finola Kane-Grade, Emily Prince, Angelina Vernetti, Kelly Powell, Scuddy Fontenelle, and Katarzyna Chawarska. 2018. "Emotional Expressivity in Toddlers With Autism Spectrum Disorder." *Journal of the American Academy of Child & Adolescent Psychiatry*. https://doi.org/10.1016/j.jaac.2018.07.872.

Maenner, Matthew J., Kelly A. Shaw, Amanda V. Bakian, Deborah A. Bilder, Maureen S. Durkin, Amy Esler, Sarah M. Furnier, et al. 2021. "Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2018." *Morbidity and Mortality Weekly Report. Surveillance Summaries* 70 (11): 1–16.

Mavadati, S. Mohammad, Mohammad H. Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn. 2013. "DISFA: A Spontaneous Facial Action Intensity Database." *IEEE Transactions on Affective Computing* 4 (2): 151–60.

McPartland, James C., Raphael A. Bernier, Shafali S. Jeste, Geraldine Dawson, Charles A. Nelson, Katarzyna Chawarska, Rachel Earl, et al. 2020. "The Autism Biomarkers Consortium for Clinical Trials (ABC-CT): Scientific Context, Study Design, and Progress Toward Biomarker Qualification."

*Frontiers in Integrative Neuroscience* 14 (April): 16.

McPartland, J., G. Dawson, and S. J. Webb. 2004. “Event-related Brain Potentials Reveal Anomalies in Temporal Processing of Faces in Autism Spectrum Disorder.” *Journal of Child*. https://acamh.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-7610.2004.00318.x.

Moore, Adrienne, Madeline Wozniak, Andrew Yousef, Cindy Carter Barnes, Debra Cha, Eric Courchesne, and Karen Pierce. 2018. “The Geometric Preference Subtype in ASD: Identifying a Consistent, Early-Emerging Phenomenon through Eye Tracking.” *Molecular Autism* 9 (March): 19.

Mullen, E. M. n.d. “Mullen Scales of Early Learning.” Accessed August 25, 2022. http://www.v-psyche.com/doc/special-cases/Mullen%20Scales%20of%20Early%20Learning.docx.

“Optbin: Optimal Binning of Data.” n.d. Comprehensive R Archive Network (CRAN). Accessed October 19, 2022. https://cran.r-project.org/web/packages/optbin/index.html.

Packer, Alan. 2016. “Neocortical Neurogenesis and the Etiology of Autism Spectrum Disorder.” *Neuroscience and Biobehavioral Reviews* 64 (May): 185–95.

Pedregosa, Varoquaux, and Gramfort. n.d. “Scikit-Learn: Machine Learning in Python.” *The Journal of Machine*. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https://githubhelp.com.

Philip, R. C. M., H. C. Whalley, A. C. Stanfield, R. Sprengelmeyer, I. M. Santos, A. W. Young, A. P. Atkinson, et al. 2010. “Deficits in Facial, Body Movement and Vocal Emotional Processing in Autism Spectrum Disorders.” *Psychological Medicine* 40 (11): 1919–29.

Pierce, Karen, Cindy Carter, Melanie Weinfeld, Jamie Desmond, Roxana Hazin, Robert Bjork, and Nicole Gallagher. 2011. “Detecting, Studying, and Treating Autism Early: The One-Year Well-Baby Check-up Approach.” *The Journal of Pediatrics* 159 (3): 458–65.e1–6.

Pierce, Karen, Vahid Gazestani, Elizabeth Bacon, Eric Courchesne, Amanda Cheng, Cynthia Carter Barnes, Srinivasa Nalabolu, et al. 2021. “Get SET Early to Identify and Treatment Refer Autism Spectrum Disorder at 1 Year and Discover Factors That Influence Early Diagnosis.” *The Journal of Pediatrics* 236 (September): 179–88.

Pierce, Karen, Vahid H. Gazestani, Elizabeth Bacon, Cynthia Carter Barnes, Debra Cha, Srinivasa Nalabolu, Linda Lopez, Adrienne Moore, Sunny Pence-Stophaeros, and Eric Courchesne. 2019. "Evaluation of the Diagnostic Stability of the Early Autism Spectrum Disorder Phenotype in the General Population Starting at 12 Months." *JAMA Pediatrics* 173 (6): 578–87.

Pierce, Karen, Steven Marinero, Roxana Hazin, Benjamin McKenna, Cynthia Carter Barnes, and Ajith Malige. 2016. "Eye Tracking Reveals Abnormal Visual Preference for Geometric Images as an Early Biomarker of an Autism Spectrum Disorder Subtype Associated With Increased Symptom Severity." *Biological Psychiatry* 79 (8): 657–66.

Press, Clare, Daniel Richardson, and Geoffrey Bird. 2010. "Intact Imitation of Emotional Facial Actions in Autism Spectrum Conditions." *Neuropsychologia* 48 (11): 3291–97.

Pulido-Castro, Sergio, Nubia Palacios-Quecan, Michelle P. Ballen-Cardenas, Sandra Cancino-Suárez, Alejandra Rizo-Arévalo, and Juan M. López López. 2021. "Ensemble of Machine Learning Models for an Improved Facial Emotion Recognition." In *2021 IEEE URUCON*, 512–16.

Rieffe, C., M. Meerum Terwogt, and L. Stockmann. 2000. "Understanding Atypical Emotions among Children with Autism." *Journal of Autism and Developmental Disorders* 30 (3): 195–203.

Riehle, Marcel, Jürgen Kempkensteffen, and Tania M. Lincoln. 2017. "Quantifying Facial Expression Synchrony in Face-To-Face Dyadic Interactions: Temporal Dynamics of Simultaneously Recorded Facial EMG Signals." *Journal of Nonverbal Behavior* 41 (2): 85–102.

Rozga, Agata, Tricia Z. King, Richard W. Vuduc, and Diana L. Robins. 2013. "Undifferentiated Facial Electromyography Responses to Dynamic, Audio-Visual Emotion Displays in Individuals with Autism Spectrum Disorders." *Developmental Science* 16 (4): 499–514.

"RPubs - Games-Howell Nonparametric Post-Hoc Test." n.d. Accessed June 5, 2022. https://rpubs.com/aaronsc32/games-howell-test.

Sariyanidi, Evangelos, Casey J. Zampella, Robert T. Schultz, and Birkan Tunc. 2020. "Can Facial Pose and Expression Be Separated with Weak Perspective Camera?" *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference* *on Computer Vision and Pattern Recognition* 2020 (June): 7171–80.

Sparrow, Balla, and Cicchetti. n.d. “Vineland Scales of Adaptive Behavior, Survey Form Manual.” *Circle Pines, MN: American Guidance Service*.

Toisoul, Antoine, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. 2021. “Estimation of Continuous Valence and Arousal Levels from Faces in Naturalistic Conditions.” *Nature Machine Intelligence* 3 (1): 42–50.

Trappenberg, T. P. 2019. “Machine Learning with Sklearn.” *Fundamentals of Machine*. https://oxford.universitypressscholarship.com/downloadpdf/10.1093/oso/9780198828044.001.0001/oso-9780198828044-chapter-3.pdf.

Trevisan, Dominic A., Maureen Hoskyn, and Elina Birmingham. 2018. “Facial Expression Production in Autism: A Meta-Analysis.” *Autism Research: Official Journal of the International Society for Autism Research* 11 (12): 1586–1601.

Uljarevic, Mirko, and Antonia Hamilton. 2013. “Recognition of Emotions in Autism: A Formal Meta-Analysis.” *Journal of Autism and Developmental Disorders* 43 (7): 1517–26.

Virtanen, Pauli, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, et al. 2020. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python.” *Nature Methods* 17 (3): 261–72.

Weiss, Elisabeth M., Christian Rominger, Ellen Hofer, Andreas Fink, and Ilona Papousek. 2019. “Less Differentiated Facial Responses to Naturalistic Films of Another Person’s Emotional Expressions in Adolescents and Adults with High-Functioning Autism Spectrum Disorder.” *Progress in Neuro-Psychopharmacology & Biological Psychiatry* 89 (March): 341–46.

Wen, Teresa H., Amanda Cheng, Charlene Andreason, Javad Zahiri, Yaqiong Xiao, Ronghui Xu, Bokan Bao, et al. 2022. “Large Scale Validation of an Early-Age Eye-Tracking Biomarker of an Autism Spectrum Disorder Subtype.” *Scientific Reports* 12 (1): 4253.

Zampella, Casey J., Loisa Bennetto, and John D. Herrington. 2020a. “Computer Vision Analysis of Reduced Interpersonal Affect Coordination in Youth With Autism Spectrum Disorder.” *Autism*

*Research*. https://doi.org/10.1002/aur.2334.

Zampella, Casey J., Loisa Bennetto, and John D. Herrington. 2020b. "Computer Vision Analysis of Reduced Interpersonal Affect Coordination in Youth With Autism Spectrum Disorder." *Autism Research: Official Journal of the International Society for Autism Research* 13 (12): 2133–42.

Zane, Emily, Kayla Neumeyer, Julia Mertens, Amanda Chugg, and Ruth B. Grossman. 2018. "I Think We're Alone Now: Solitary Social Behaviors in Adolescents with Autism Spectrum Disorder." *Journal of Abnormal Child Psychology* 46 (5): 1111–20.

Zhang, Kaipeng, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks." *IEEE Signal Processing Letters*. https://doi.org/10.1109/lsp.2016.2603342.

Zhang, Zhi-Min, Shan Chen, and Yi-Zeng Liang. 2010. "Baseline Correction Using Adaptive Iteratively Reweighted Penalized Least Squares." *The Analyst* 135 (5): 1138–46.

## Appendix

### Appendix Result 1

**Sanity check for 1,822 high-performance model.**

First, we evaluated the performance of the 1,822 high-performing models on a Longitudinal dataset of 18 ASD and 15 TD samples from toddlers who were included in the Discovery dataset. Out of 1,822 models with AUC-ROC of above 0.8, 1,656 (90.8%, odds ratio 76.89514, 95% confidence interval [64.74134, 91.39221], two-sided Fisher's Exact Test $P < 2.2e\text{-}16$) of the models showed AUC-ROC above 0.8 in the Longitudinal dataset.

### Appendix Result 2

**Validation of Independent Replication dataset**

To test the sensitivity, the models were trained on the Discovery dataset (175 subjects together) and then tested on the dataset with 34 ASD and 31 TD. The AUC-ROC values were used to evaluate the performance of the selected 1,822 high-performing models. 1,076 models (59.0%) had an AUC-ROC value greater than 0.75. The two-sided Fisher's Exact Test calculated the odds ratio 5.133581 (95% confidence interval [4.688261, 5.620152]) with $P < 2.2e\text{-}16$.

In terms of the diagnostic specificity, the models were trained on the discovery dataset (175 subjects together) and then tested on the independent Replication dataset (55 subjects, 31 TD, and 24 LD). From 1,822 models with AUC-ROC or AUC-PR values greater than 0.8, none had an AUC-ROC value of above 0.8 in this LD vs TD dataset. The two-sided Fisher's Exact Test calculated the odds ratio 0 (95% confidence interval [0, 0.01765617]) with $P < 2.2e\text{-}16$. 26 methods had an AUC-ROC value of above 0.8 in this LD vs TD dataset. The two-sided Fisher's Exact Test calculated the odds ratio 0.1240666 (95% confidence interval [0.08055236, 0.18305634]) with $P < 2.2e\text{-}16$.

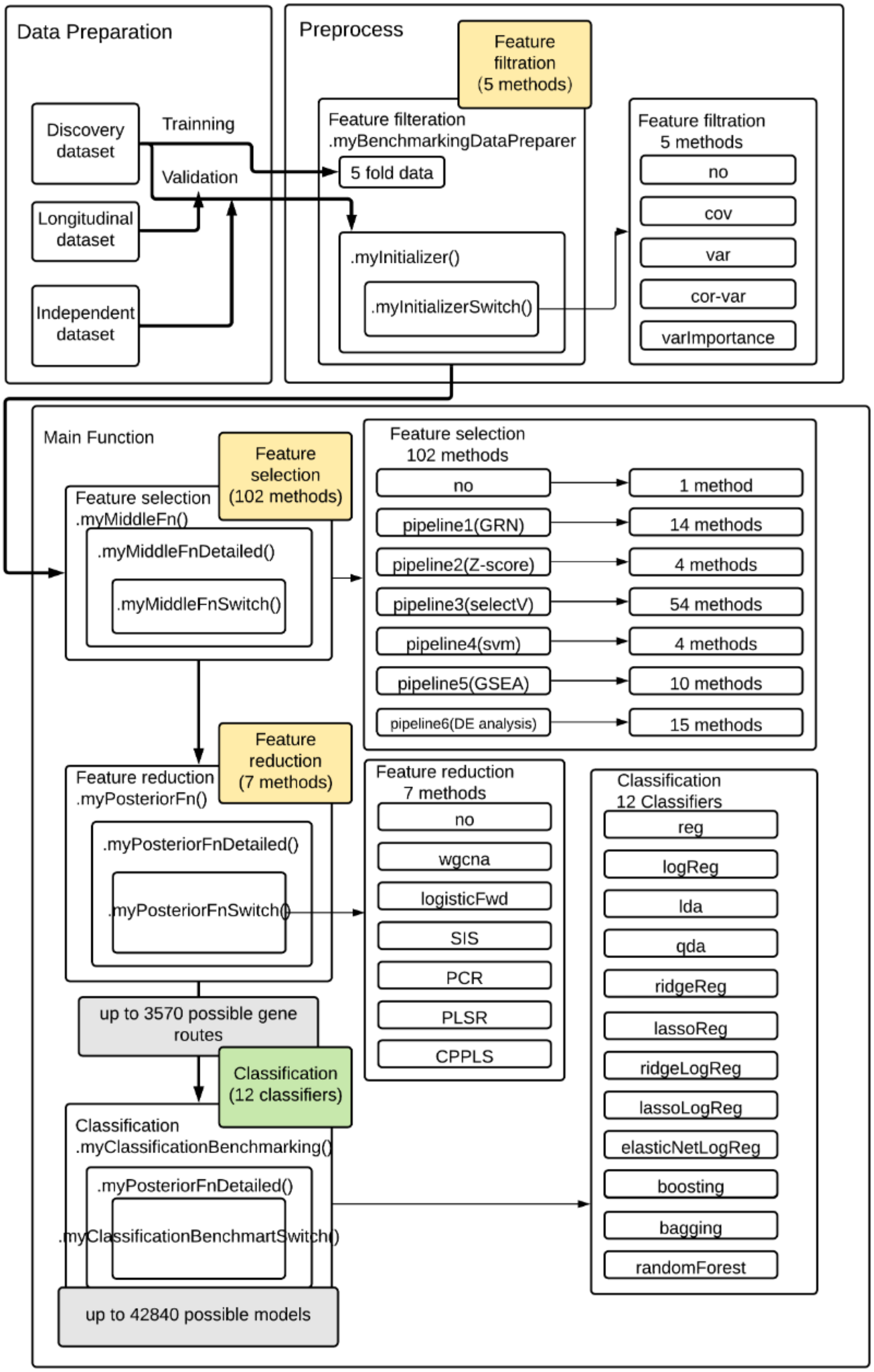
We also assessed the diagnostic specificity of the models by comparing performance of the 1,822 models in separating a Diagnostic Specificity dataset of 24 language delayed (LD) toddlers from the ASD samples. From 1,822 models with AUC-ROC or AUC-PR above 0.8, 26 (1.42%; odds ratio 0.1113861, 95%

confidence interval [0.07231532, 0.16436198]), two-sided Fisher's Exact Test *P* < 2.2e-16) have an AUC-ROC above 0.75 in this LD vs TD dataset.

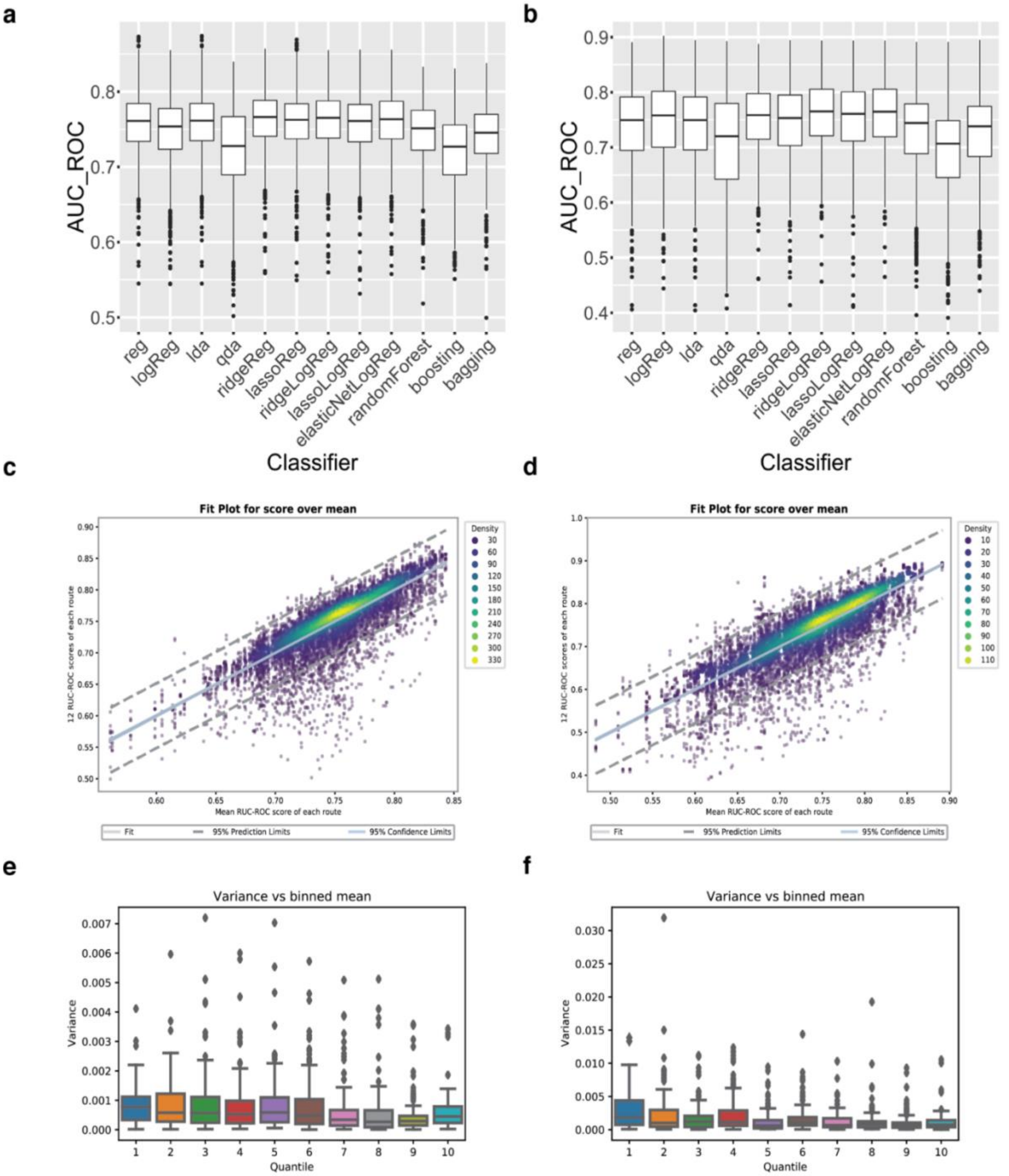
### Appendix Result 3

### Result and parameter setting on the baseline Random Forest model

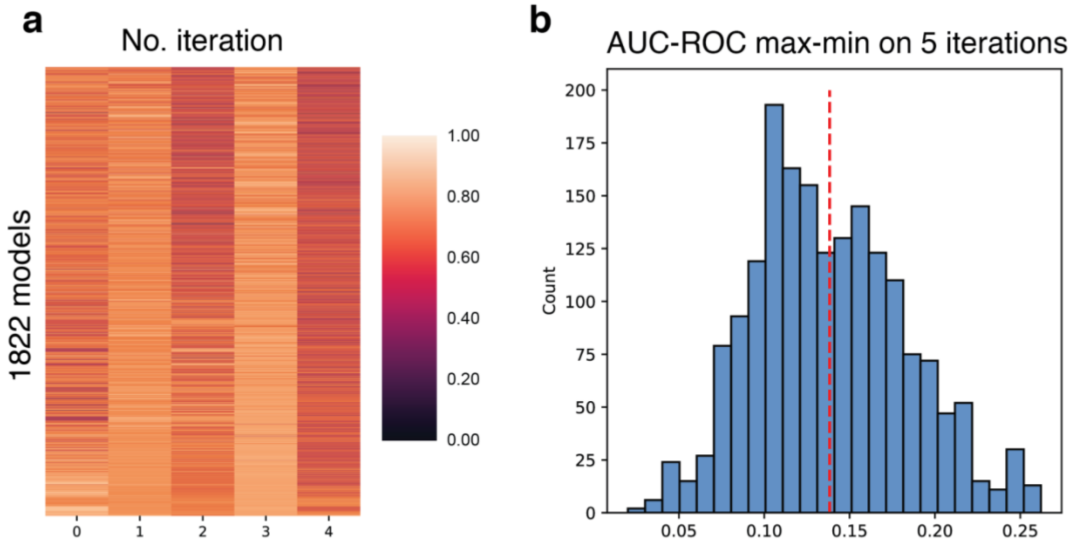
Number of top variables are chosen from results based on the validation data, over a grid of (10, 100, 500, 1,000). On this grid, 500 genes are chosen as optimal. Importances are generated from 100 rounds of evaluations of Random Forest, averaged. Final test results using these top 500 genes have the accuracy: 73.33%, sensitivity: 85.71%, specificity: 62.50%, AUC-ROC: 72.32%.



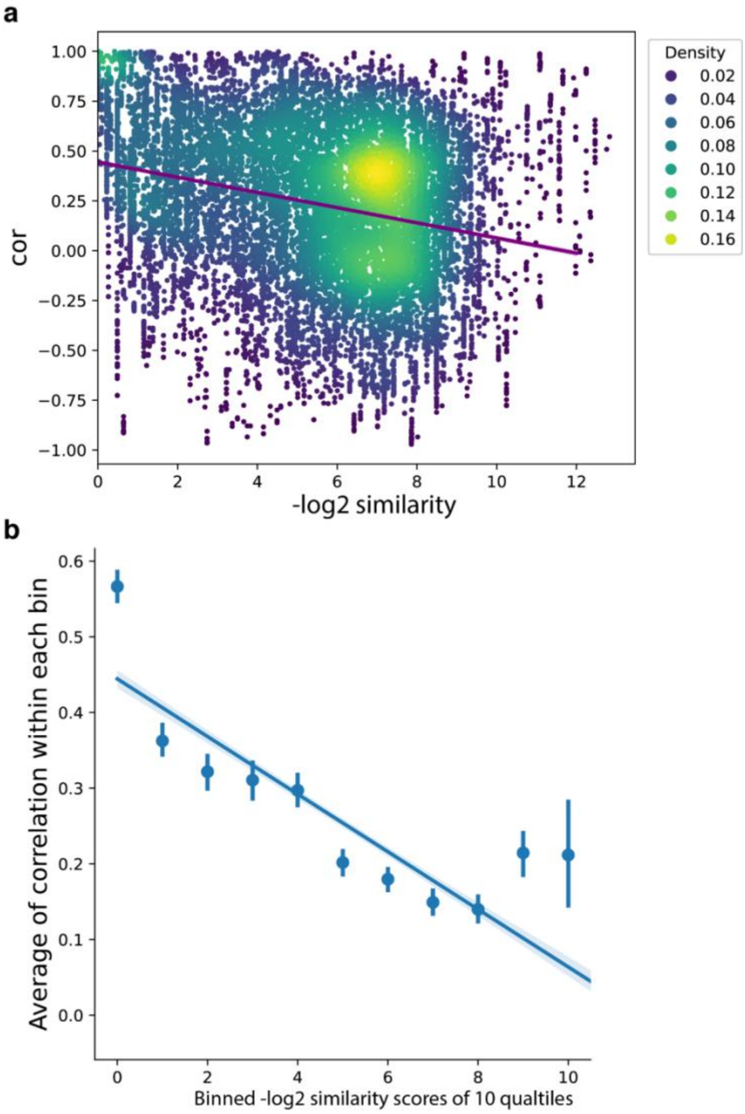
**Appendix Figure 1 The detail workflow of the feature engineering pipeline.**



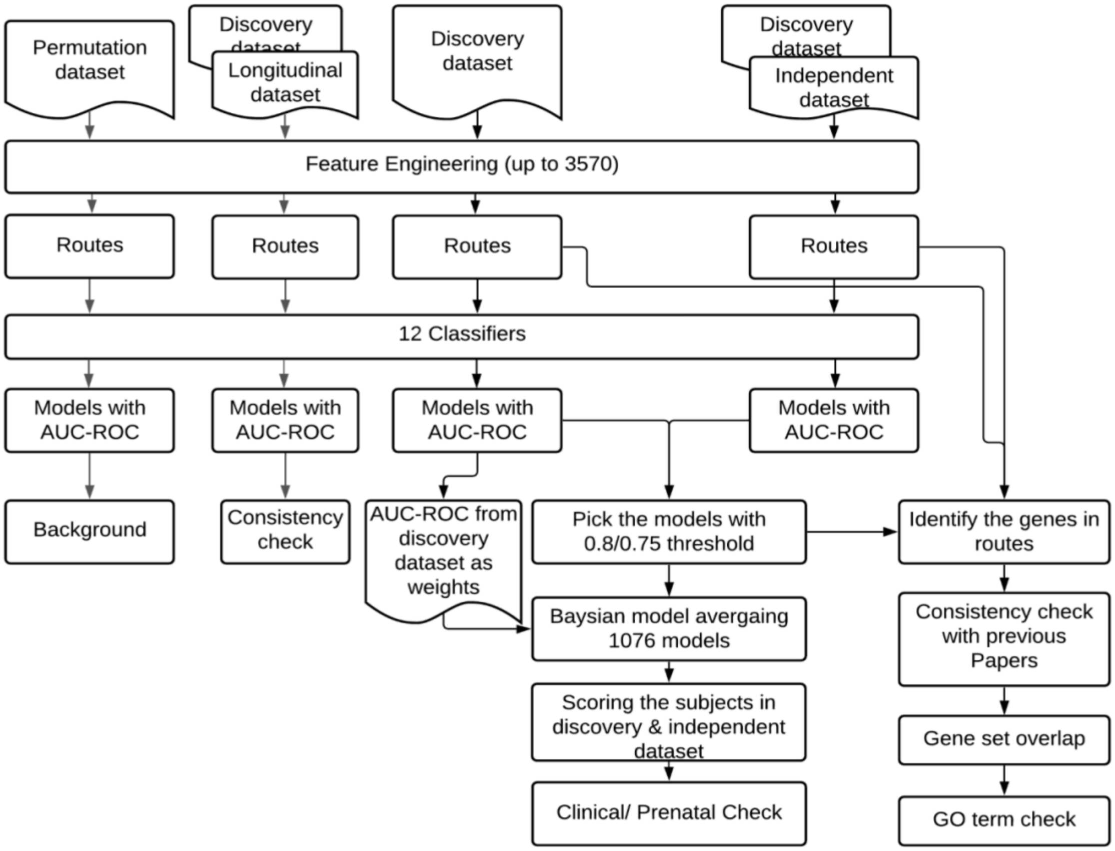
**Appendix Figure 2 The distribution of AUC-ROC scores of 12 classifiers against the AUC-ROC mean score of 1320 routes.** The distribution of AUC-ROC generated by 1,320 routes within each of 12 classifiers in **a.** the discovery dataset and **b.** the independent dataset. The distribution of AUC-ROC generated by 1,320 routes vs the mean of AUC-ROC of each route in **b.** the discovery dataset (with +- 0.0514 95% confidence interval) and **d.** the distribution in the independent dataset (with +- 0.0514 95% confidence interval). In the variance of AUC-ROCs generated by 12 classifiers of 1,320 routes, the x-axis is the mean of AUC-ROC of each route in **e.** the main dataset and **f.** in the test dataset.



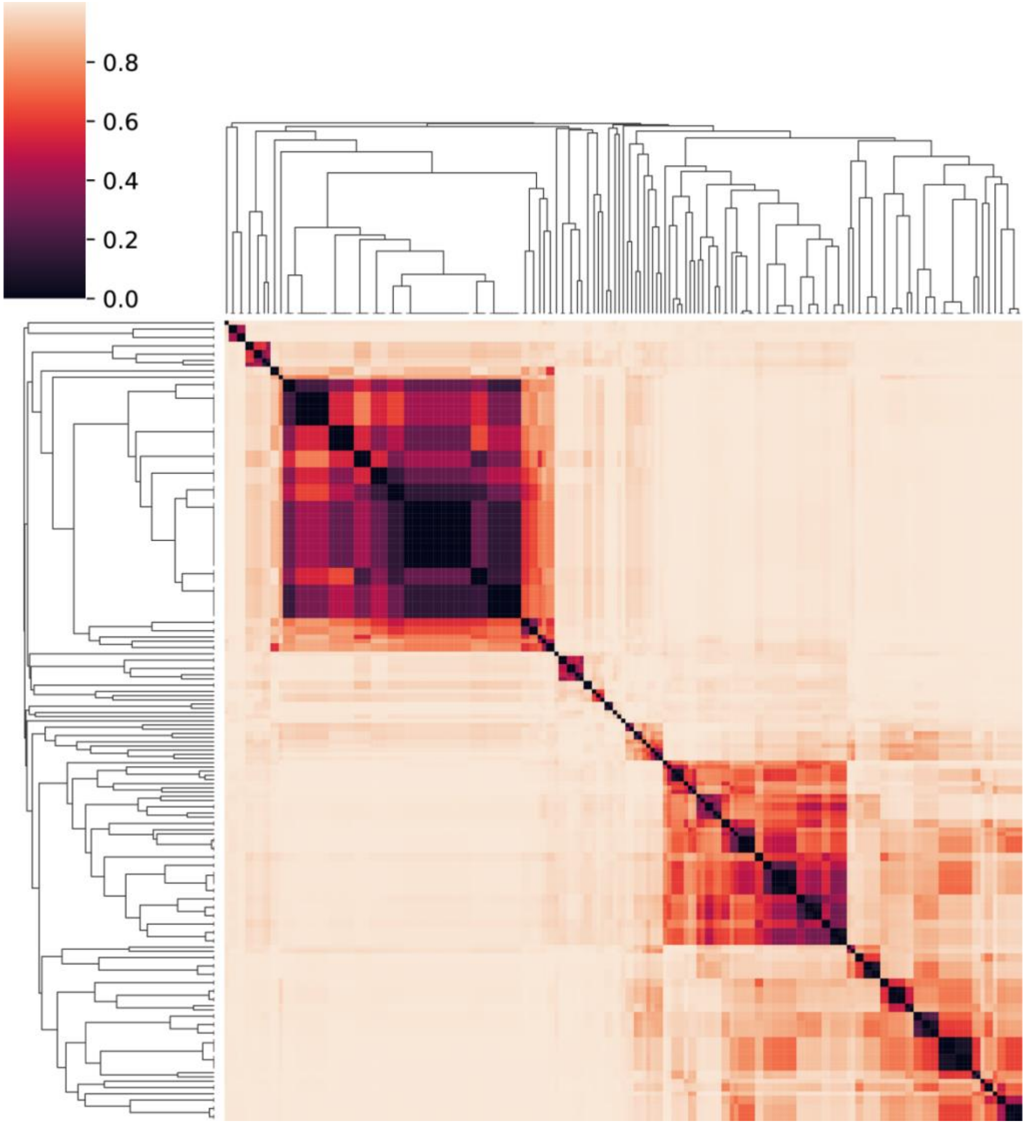
**Appendix Figure 3 The AUC-ROC score of 1,320 models trained on Discovery dataset. a.** The score distribution of 1,822 models that have AUC-ROC score above 0.8 in the discovery dataset in 5 iterations. **b.** The score-difference distribution of 1,822 models in 5 iterations.



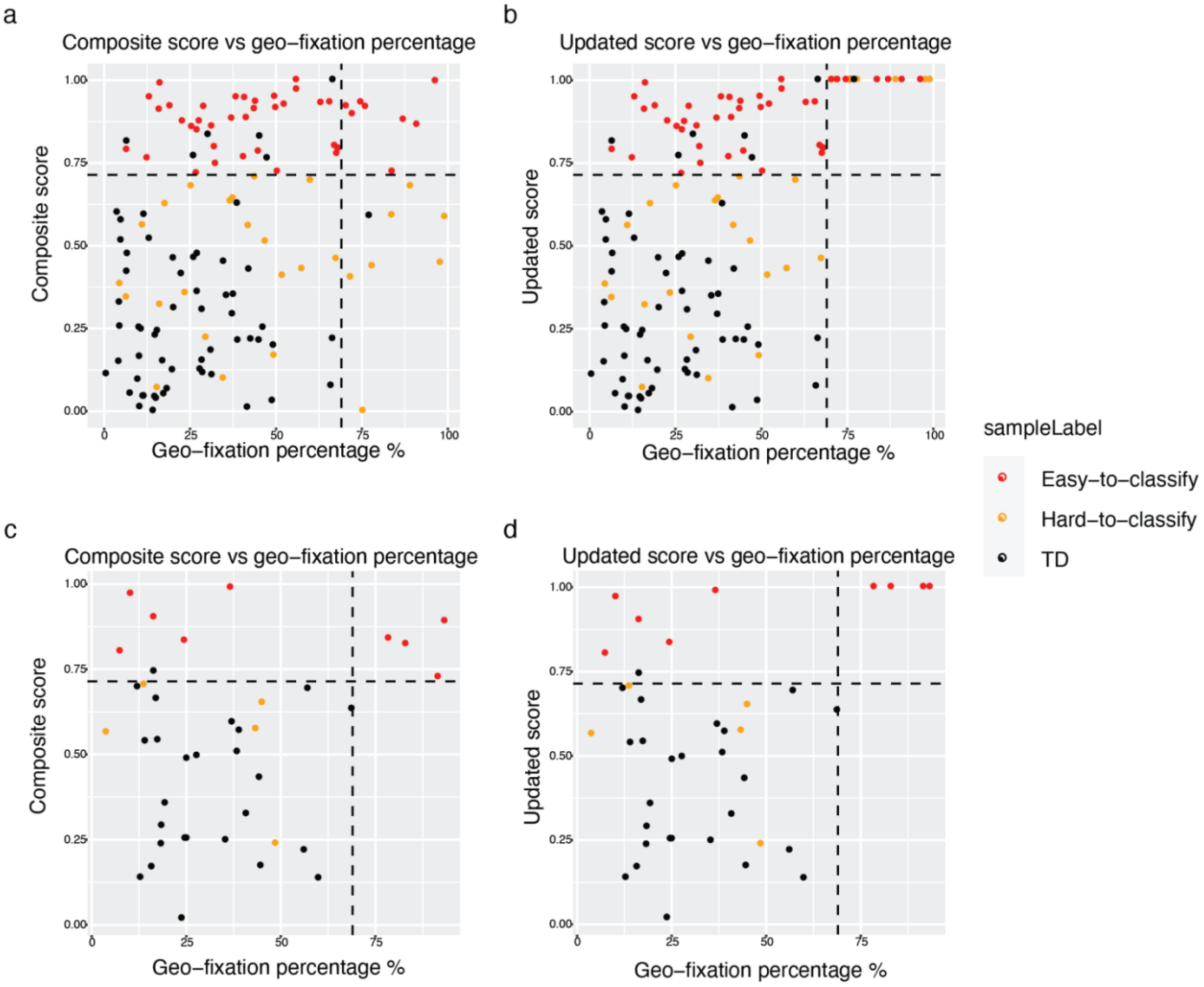
**Appendix Figure 4 The relation between the similarity of gene routes and the 12 classifiers behavior.** The correlation of 12-classifier score vs the distance between routes and models. The distance is measured by the -log2 of the geneset similarity on the first 500 genes (see **Methods**). The slope=-0.03806252570865483, intercept=0.4443827187728834, rvalue=-0.25294548575900266, pvalue=1.35416223411553e-265, and stderr=0.001075187710703305.



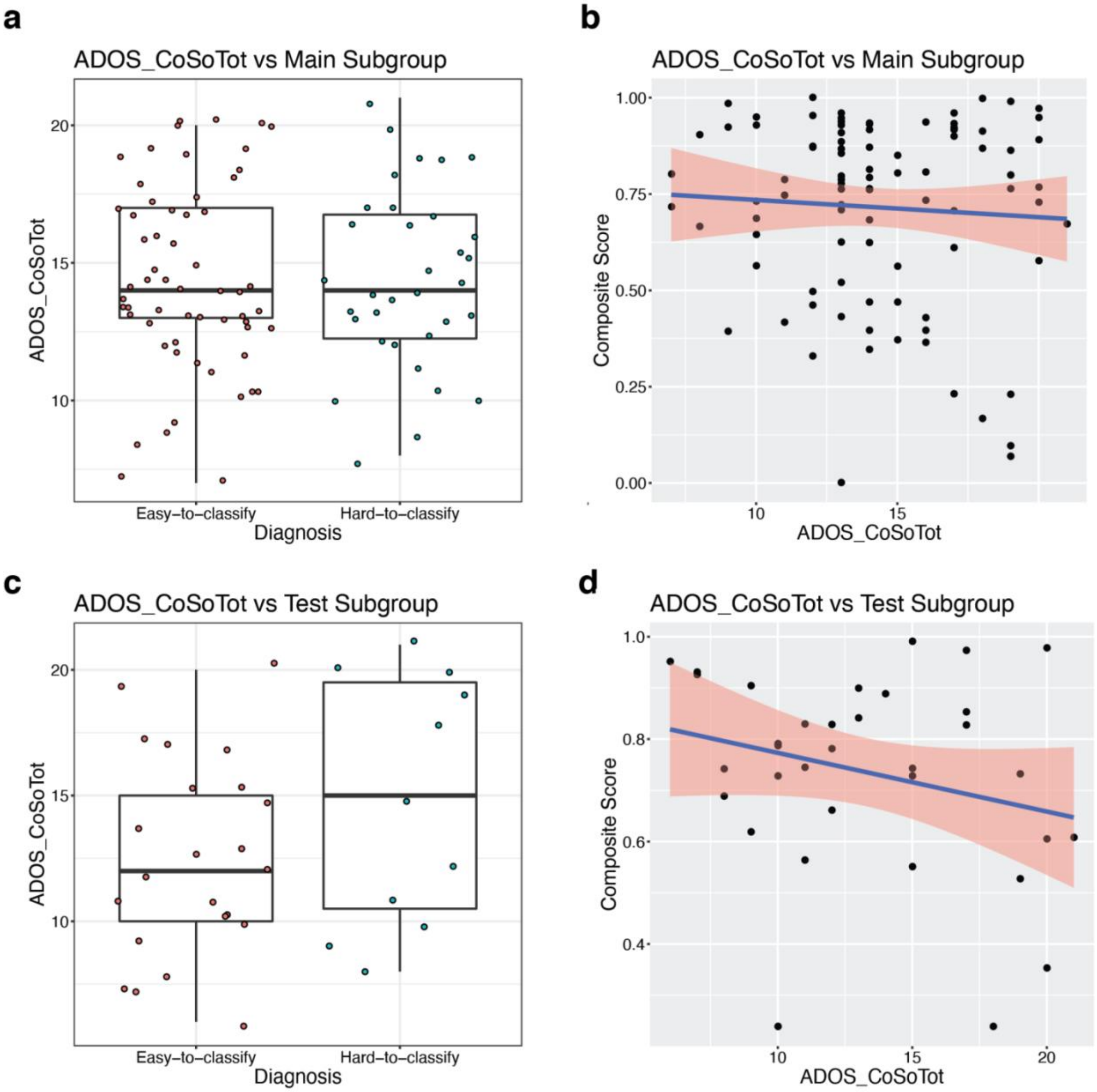
**Appendix Figure 5 The overall data processing workflow.** The workflow lists the processes of running main training test on Discovery dataset, validation test on independent Replication dataset, reproducibility test on longitudinal dataset and background test on randomized permutation dataset.



**Appendix Figure 6 The geneset similarity distance among 191 routes.** The distance score (see **Methods**) is measured by the top 500 genes that are used by 191 routes and then clustered by ‘average’ method. The data of the table is in eTable 6.



**Appendix Figure 7 Illustration of the updated model that combines the composite model with geo-fixation model.** The updated model was tested on 132 of 175 Discovery dataset subjects and 41of 65 Replication dataset subjects who had available Geo-Fixation data (e.g., moderate or good data quality, total looking time > 50%). By directly classifying the subjects who had percent fixation on non-social images >69% as ASD (GeoPref-subtype). **A**. The composite score vs geo-fixation percentage score in discovery dataset. **B**. The updated score vs geo-fixation percentage score for discovery dataset. **C**. The composite score vs geo-fixation percentage score for the independent dataset. **D**. The updated score vs geo-fixation percentage score for independent dataset.



**Appendix Figure 8 Diagnostic vs psychometric scores**. Diagnostic and psychometric scores were not significantly different between ASD toddlers above (easy-to-classify) and below the mean composite classifier scores (hard-to-classify). **a-b** discovery dataset r= 0.5310815, Two sided t.test P = 4.011e-14; **c-d** independent dataset r= 0.4871145, Two sided t.test P = 3.873e-05.

# CONCLUSION

As bioinformaticians, we keep leveraging new methodologies from wet and dry labs to increase our capabilities understanding complicated biological mechanisms.  Because of the inherent complication of biological processes, applying new analyses or evaluating new data is often full of assumptions both explicit and implicit. A better approach to presenting data usually brings new insight and makes data more interpretable. In the chapter 1, the glycomics data are transformed using the synthesis network knowledge which makes the interdependent relationship between individual glycans self-explanatory. In the chapter 3, the measurement of an action unit through whole videos (2,625 frames) produces a time series with 2,625 values. However, the maximum measure in 19 scenes is selected to represent data that help shrink the total feature space. Then, careful consideration of the methods applied to our data is paramount, as seemingly arbitrary choices in the hyperparameters of our analysis (e.g. the method for data transform, the minimum depth for clustering, the threshold for model selection, and precise cutoff frequencies of a filter) can have large impacts on the results and ultimate conclusions. For example, in chapter 2, the machine learning pipeline that analyzes gene expression data requires the specific consideration of the interaction between gene expression. Simply using the current AutoML pipeline which lacks the knowledge of gene-gene interaction, gene coexpression pattern, and gene pathway will result in a low discovery power to discriminate the positive signal from the control. Further, the possible confounding factors in clinical data require a strict examination to defend a diagnostic-related conclusion. In chapter 2 and 3, when I am splitting the data into training and test groups, I am required to ensure the patient's age (gender), ethnicity, and race profile are carefully balanced (or tested in post-hoc analysis) to remove the implicit biases caused by the general confounding factors. In-depth knowledge of the biological/clinical background enables us to test our model's result and assure the validity of our conclusions.

Specifically for chapter 3, the current analysis has limitations due to the foundational and exploratory nature of the work. The analysis is based on an established 'The Joint Attention Test' experiment. Future work will have to ascertain the reliability and reproducibility of the current results and there are a lot of room to make the experiment better. One limitation is that the sample sizes in ASD and non-ASD groups (including TD, DD, GDD, LD, etc.) are imbalanced. Since the data were collected during the COVID-19 pandemic period, we could not recruit enough TD toddlers in this study and thus included children with TD and delays (e.g., DD, GDD, LD, etc) as the non-ASD group. The other neurodevelopmental disorders may confound the classification. We also had to remove more than half of the subjects as their facial movement data were unavailable (wearing masks, covered by hand, eating) during the experiment, which was not a big problem before the facial emotional data became appreciated. In the future when the facial behavior data is pilled up, we are able to customize the facial model specifically for toddlers, instead of the OpenFace and Emonet packages, to give a better measurement of the action unit intensity. In addition, the experiment has some disadvantages, although the 'The Joint Attention Test' experiment is the most arousal experiment in our inventory. The social and non-social subjects are close to each other, and toddlers tend to shift their attention (eye gazing target) across multiple AOIs on a millisecond scale. We are not able to distinguish emotional behavior when toddlers look at Social and non-social AOI separately.

In conclusion, using emerging technics to facilitate the current study provides us the unique opportunity to make discoveries and regeneratively give us more feedback about the study as a whole (a spiral-up process). As a bioinformatician, it was a fun journey in my life to develop our data analysis technics and expand our knowledge in both glycan analysis and ASD diagnosis.