PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Classification and adaptive behavior prediction of children with autism spectrum disorder based upon multivariate data analysis of markers of oxidative stress and DNA methylation

Daniel P. Howsmon[1,2], Uwe Kruger[3], Stepan Melnyk[4], S. Jill James[4], Juergen Hahn[1,2,3]*

1 Department of Chemical and Biological Engineering, Rensselaer Polytechnic Institute, Troy, New York, United States of America, **2** Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, New York, United States of America, **3** Department of Biomedical Engineering, Rensselaer Polytechnic Institute, Troy, New York, United States of America, **4** Department of Pediatrics, University of Arkansas for Medical Sciences, Little Rock, Arkansas, United States of America

* hahnj@rpi.edu

OPEN ACCESS

Citation: Howsmon DP, Kruger U, Melnyk S, James SJ, Hahn J (2017) Classification and adaptive behavior prediction of children with autism spectrum disorder based upon multivariate data analysis of markers of oxidative stress and DNA methylation. PLoS Comput Biol 13(3): e1005385. doi:10.1371/journal.pcbi.1005385

Editor: Christos A. Ouzounis, Centre for Research and Technology-Hellas, GREECE

Received: August 12, 2016

Accepted: January 28, 2017

Published: March 16, 2017

Data Availability Statement: De-identified data are available in S1 Dataset.

Funding: DPH and JH gratefully acknowledge partial financial support from the National Institutes of Health (https://www.nih.gov/, Grant 1R01AI110642). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

The number of diagnosed cases of Autism Spectrum Disorders (ASD) has increased dramatically over the last four decades; however, there is still considerable debate regarding the underlying pathophysiology of ASD. This lack of biological knowledge restricts diagnoses to be made based on behavioral observations and psychometric tools. However, physiological measurements should support these behavioral diagnoses in the future in order to enable earlier and more accurate diagnoses. Stepping towards this goal of incorporating biochemical data into ASD diagnosis, this paper analyzes measurements of metabolite concentrations of the folate-dependent one-carbon metabolism and transulfuration pathways taken from blood samples of 83 participants with ASD and 76 age-matched neurotypical peers. Fisher Discriminant Analysis enables multivariate classification of the participants as on the spectrum or neurotypical which results in 96.1% of all neurotypical participants being correctly identified as such while still correctly identifying 97.6% of the ASD cohort. Furthermore, kernel partial least squares is used to predict adaptive behavior, as measured by the Vineland Adaptive Behavior Composite score, where measurement of five metabolites of the pathways was sufficient to predict the Vineland score with an R^2 of 0.45 after cross-validation. This level of accuracy for classification as well as severity prediction far exceeds any other approach in this field and is a strong indicator that the metabolites under consideration are strongly correlated with an ASD diagnosis but also that the statistical analysis used here offers tremendous potential for extracting important information from complex biochemical data sets.

Author summary

Autism spectrum disorder (ASD) encompasses a family of neurological disorders characterized by limited social interaction and restricted repetitive behaviors. The number of children diagnosed with ASD has grown exponentially over the last four decades and is now estimated to affect ∼1.5% of children. Although ASD is currently diagnosed and treated based solely on psychometric tools, a biochemical view applicable to at least a subset of ASD cases is emerging. Abnormalities in folate-dependent one carbon metabolism and transsulfuration pathways can summarize a large number of observations of genetic and environmental effects that increase ASD predisposition. However, these complex, highly interconnected pathways require more advanced statistical models than the typical univariate models presented in the literature. Therefore, we developed multivariate statistical models that classify participants based on their neurological status and predict adaptive behavior in ASD. We emphasize that these models are cross-validated, helping to ensure that the results will generalize to new samples. The models developed herein have much stronger predictability than any existing approaches from the scientific literature.

Introduction

Autism Spectrum Disorder (ASD) encompasses a large group of early-onset neurological diseases characterized by difficulties with social communication/interaction and expression of restricted repetitive behaviors and interests [1]. In addition to these defining behavioral symptoms, individuals with ASD frequently have one or more co-occurring conditions, including intellectual disability, ADHD, speech and language delays, psychiatric diagnoses, epilepsy, sleep disorders, and gastrointestinal problems [2–5]. ASD affects ∼1.5% of the population and affects males disproportionately [6–8]. It is associated with an impaired quality of life [9] and the lifetime cost of supporting an individual with ASD amounts to $1.4–2.4MM, depending on co-existing disorders [10].

It is generally acknowledged that ASD has a strong genetic component, but environmental effects have also recently emerged as important contributors to the etiology and pathophysiology of ASD in at least a subpopulation of cases. Early twin studies suggested that the heritability of ASD was 80–90% [11]; however, twin studies since 2010 suggest a lower heritability of only 37–55% [12, 13]. Despite this high genetic association, only 15% of ASD cases have a known genetic source [1]. Although genetic studies continue to provide new evidence for contributing factors to ASD etiology [14], environmental effects such as maternal/paternal age, toxic chemical exposure, maternal rubella infection, etc. are also emerging as key factors contributing to ASD liability [13].

No generally accepted biomarkers for the diagnosis or diagnosis of the severity of ASD exist to date. Instead, diagnostic evaluation involves a multi-disciplinary team of doctors usually including a pediatrician, psychologist, speech and language pathologist, and occupational therapist. Despite this current state of the art, work in identifying biomarkers that can support the diagnosis process is ongoing. In particular, abnormalities in folate-dependent one-carbon metabolism (FOCM) and transsulfuration (TS) likely contribute to the genetic and environmental predisposition to ASD [15]. FOCM contributes to epigenetic gene expression through DNA methylation and TS is the major contributor to intracellular redox status. An illustration of these pathways overlaid with genetic and environmental contributions to ASD predisposition is presented in Fig 1.

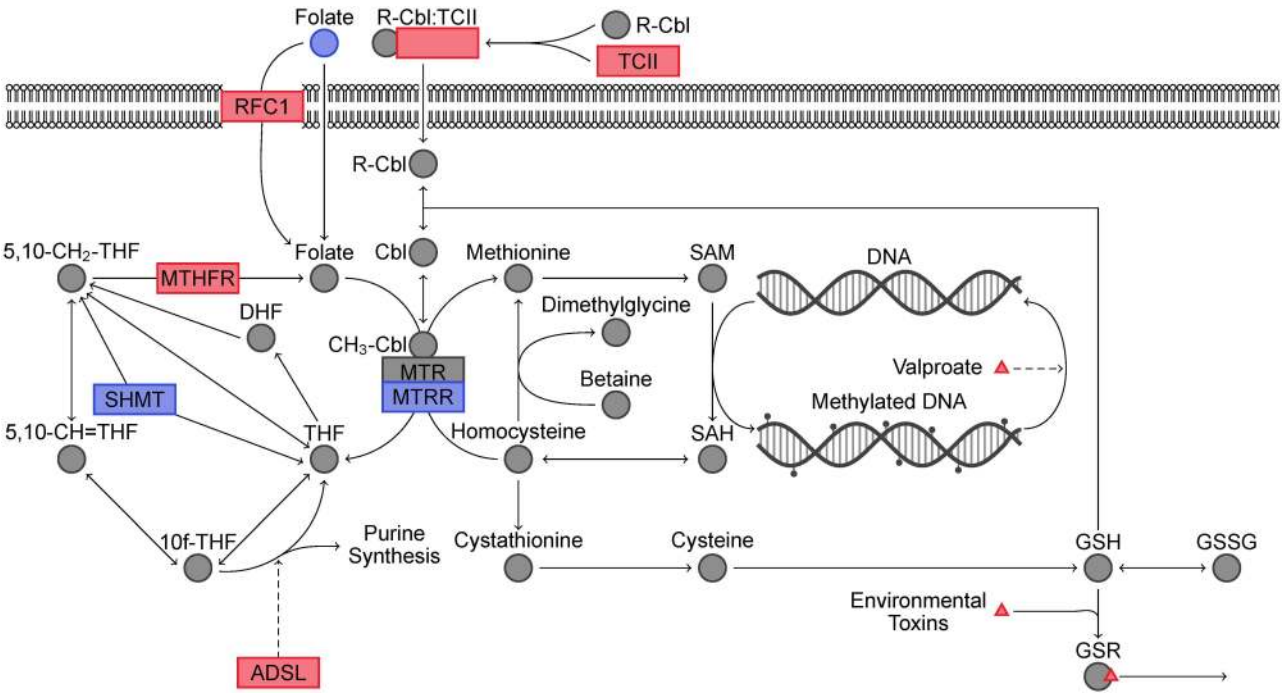

Fig 1. Illustration of folate-dependent one-carbon metabolism and transsulfuration pathways. Genetic and environmental effects that increase ASD predisposition are shown in red whereas those that decrease ASD liability are shown in blue.

doi:10.1371/journal.pcbi.1005385.g001

Mutations or altered expression levels of several genes in these pathways have been associated with increased risk of ASD. Adenylosuccinate lyase (ADSL) deficiency leads to a purely genetic form of autism by re-directing a large proportion of FOCM toward purine synthesis to compensate for a reduction in *de novo* purine synthesis [15, 16]. Methylenetetrahydrofolate reductase (MTHFR) is responsible for generating 5-methyltetrahydrofolate, which in turn is responsible for re-methylating homocysteine to methionine. In particular, the C677T polymorphism has been shown to increase ASD liability, especially in countries where prenatal folate supplementation is low [17]. Limited evidence linking mutations in reduced folate carrier (RFC1) [18, 19], transcobalamin II (TCII) [18], serine hydroxymethyltransferase I (SHMT1) [20], 5-methyltetrahydrofolate-homocysteine methyltransferase reductase (MTRR) [18, 20], and catechol-O-methyltransferase (COMT) [18, 21] to altered prevalence of ASD has also been presented, although these contributions to ASD liability are currently contested [22].

Evidence for the association between environmentally-rooted FOCM/TS dysfunction and ASD predisposition can be seen in prenatal valproate and toxic chemical exposure as well as lack of maternal folate supplementation. Maternal valproate use during pregnancy has been associated with higher incidence rates of ASD [23, 24] and *in utero* valproate exposure has been used to develop rodent models of autism [25]. Valproate exposure causes DNA hypomethylation [26, 27] in key neurodevelopmental processes that have been mitigated by folate supplementation [28]*in vitro*. Other chemicals such as heavy metals, ethyl alcohol, pesticides, phthalates, polychlorinated biphenyls, and traffic-related air pollution (TRAP) have also been shown to affect neurodevelopment and increase ASD liability [13, 29]. These organic toxins induce oxidative stress and heavy metals disrupt transsulfuration by binding glutathione, the major contributor to intracellular redox homeostasis [30]. Additionally, glutathione is an important regulator in the intracellular processing of methylcobalamin (vitamin B_{12}), an

essential cofactor for methionine synthase and the TS pathway [31]. Air dispersion models coupled with traffic patterns/roadway geometry, meteorological data, and vehicle emission data have been used to find a dose response between ASD prevalence and TRAP exposure [32]. Additionally, common organic pollutants have been associated with increased autism severity in children on the autism spectrum [33]. Two independent studies linked maternal folate supplementation to a reduced risk of having a child with ASD [34, 35]. This protective effect is usually attributed to the involvement of FOCM in early epigenetic regulation of neurodevelopment and neural tube formation [21, 36]. For a more complete description of the evidence for the potential contributions of FOCM/TS dysfunction to the ASD phenotype, see the excellent review by Deth *et al.* [15].

Although differences between FOCM and TS pathways in children with ASD versus neurotypical controls have been shown previously [18, 37, 38], investigators have struggled with identifying a single, predictive measurement of these pathways that separates individuals with ASD from neurotypical controls or that correlates well with ASD severity. However, in many complex problems one particular measurement may be insufficient and important information can only be extracted by using multivariate statistical analysis. Indeed, incorporating multiple measurements of environmental toxins has been shown to increase the separability of control and ASD participants [39] and better predict autism severity [33, 39].

Latent variable techniques enable the discovery of important multivariate interactions, leading to improved classification and regression performance. Furthermore, latent variable techniques allow assessing the importance of individual variables and are more robust to uninformative variables. One popular latent variable technique for classification problems is Fisher Discriminant Analysis (FDA), which achieves an optimal linear separability using a typically small set of latent variables that are linear combinations of the original variable set. FDA has a long history in biological classification problems and was first used by Rao in 1948 to interpret anthropological data [40]. Extensions of FDA, such as Kernel FDA (KFDA), exist which can take nonlinear relationships into account for classification [41]. Latent variable regression techniques include partial least squares (PLS) and its nonlinear counterpart kernel PLS (KPLS) [42, 43]. Using FDA for classification and KPLS for regression allow multivariate interactions to surface, which are often hidden when only univariate analysis is considered. To guarantee a statistically independent assessment of the multivariate classification and regression models, the presented study utilizes a cross-validatory approach, where the set of samples used for model identification does not contain samples to evaluate the performance of the identified models.

The presented work makes use of these advanced modeling and statistical analysis tools to examine metabolite data of the FOCM/TS pathway in neurotypical participants (NEU) and those on the autism spectrum (ASD) as well as their siblings (SIB). Using FDA, it is possible to clearly distinguish the participants on the spectrum from their neurotypical peers and KPLS unveils a strong correlation between metabolite concentrations of these pathways and adaptive behavior as measured by the Vineland Adaptive Behavior Composite. This work not only analyzes the largest data set of its kind of these pathways in the scientific literature [38], but also results in the strongest evidence to date of the association of FOCM/TS dysfunction with ASD.

Results

Classification into ASD, NEU, and SIB cohorts

Associating dysfunction of FOCM/TS pathways with ASD requires a distinction between or separation of ASD and NEU groups based on FOCM/TS metabolites. Therefore, cross-validatory FDA was performed using measurements of the FOCM/TS metabolites listed in

Table 1. FOCM/TS metabolites considered for analysis.

Methionine	SAM	SAH
SAM/SAH	% DNA methylation	8-OHG
Adenosine	Homocysteine	Cysteine
Glu.-Cys.	Cys.-Gly.	tGSH
fGSH	GSSG	fGSH/GSSG
tGSH/GSSG	Chlorotyrosine	Nitrotyrosine
Tyrosine	Tryptophane	fCystine
fCysteine	fCystine/fCysteine	% oxidized glutathione

doi:10.1371/journal.pcbi.1005385.t001

Table 1. A linear classifier based on these FDA scores is then used to classify ASD and NEU participants. FDA scores and estimated probability distribution functions (PDFs) are provided in Fig 2. The cross-validated misclassification rates of only 4.9% and 3.4% for the NEU and ASD samples, respectively, eliminated more complex, nonlinear KFDA analysis from consideration.

The performance of the classifier was then evaluated on the SIB cohort, a more challenging classification problem due to partially shared genetic and environmental effects with the ASD cohort. Using all measurements in Table 1, an FDA model was trained to separate the ASD and NEU cohorts. Then, the trained FDA model was used to evaluate the SIB cohort (which was not used for training). The resulting separation of ASD, NEU, and SIB presented in Fig 3 shows a slight increase in the overlap with the ASD cohort when compared with the performance of the ASD vs. NEU classification. Furthermore, the SIB PDF shows significantly more

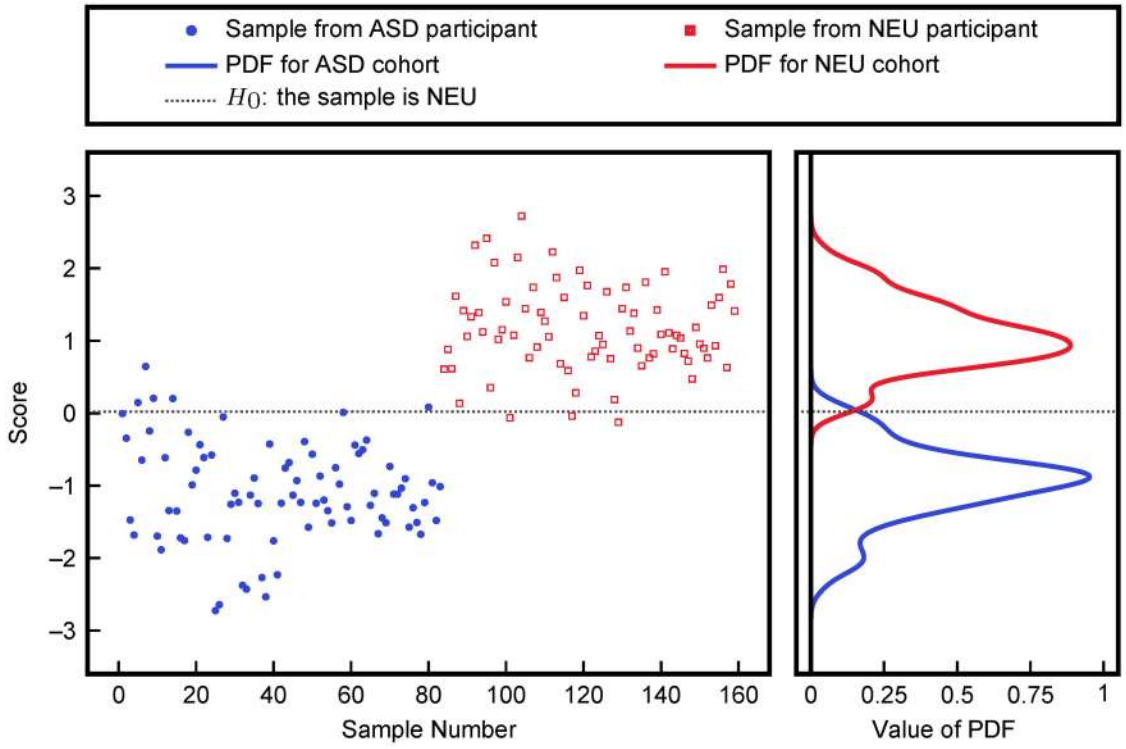

Fig 2. Classification into ASD and NEU cohorts using FDA on all FOCM/TS metabolites. The plotted scores were obtained via cross-validation and the probability distribution functions were obtained from fitting.

doi:10.1371/journal.pcbi.1005385.g002

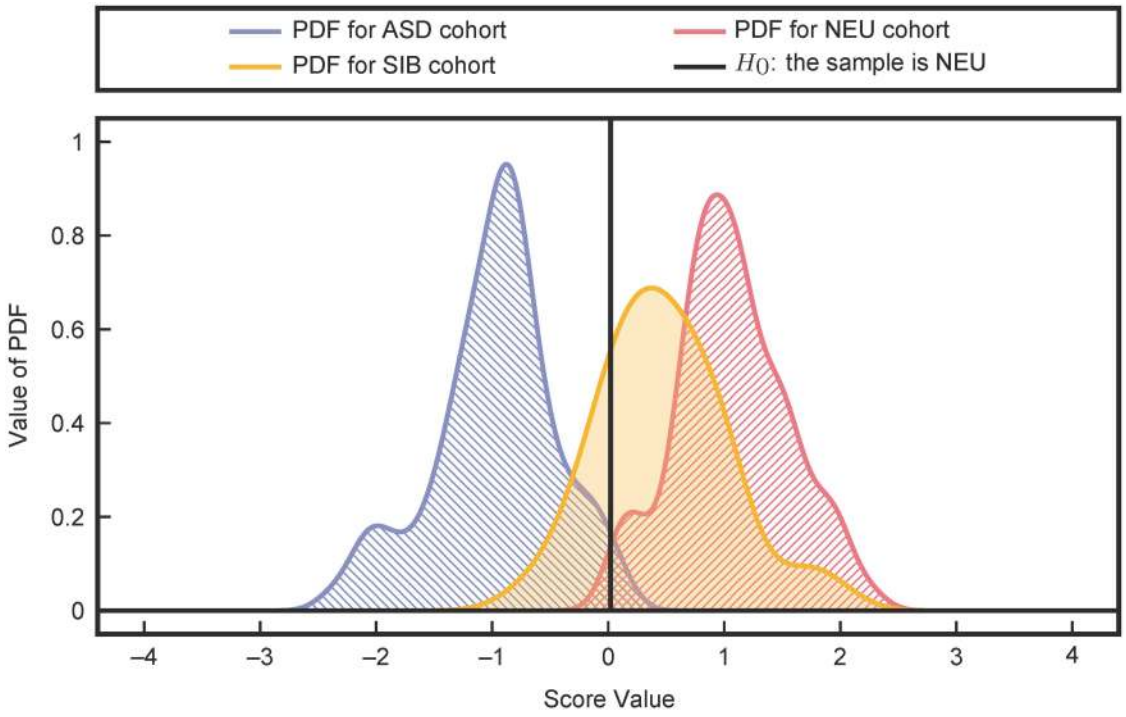

Fig 3. Classification performance on the SIB cohort (yellow). There is significantly more overlap of the SIB cohort with the NEU cohort (red) than with the ASD cohort (blue).

doi:10.1371/journal.pcbi.1005385.g003

overlap with the NEU PDF than the ASD PDF. These results support the hypothesis proposed by James *et al* [38] that the siblings of the participants on the spectrum have FOCM/TS metabolite profiles that are significantly more similar to their neurotypical peers than their siblings, even though genetically they are likely closer to their siblings than participants in the neurotypical control group.

Analysis of important metabolites for classification

The simultaneous use of multiple measurements promises to increase the separability of the cohorts; however, increasing the number of measurements increases the number of parameters in the projection vector w that maximizes the separability of the two groups (see Materials and methods). Although cross-validation can help mitigate these effects, the increased number of parameters can lead to over-fitting, which would indicate good performance for separation on the existing data set, but poor separation performance when the analysis results are translated to new data. These over-fitting problems can be further mitigated by selecting only the minimum number of variables required to adequately separate the two groups. Therefore, all combinations of up to six variables were evaluated for separability. Select combinations of higher numbers of variables were chosen in a greedy fashion to sequentially add measurements that best improve the separation of the best six variables. Cross-validatory FDA was performed on all variable combinations and probability distribution functions (PDFs) of the FDA scores of the two cohorts were estimated. A receiver-operating-characteristic (ROC) curve was generated based on these PDFs. The C-statistic of the ROC curve provides a measure of the ability of the classifier to separate into ASD and neurotypical groups. A C-statistic of 0.5 represents random classification and a C-statistic of 1.0 represents perfect classification. Fig 4 plots the

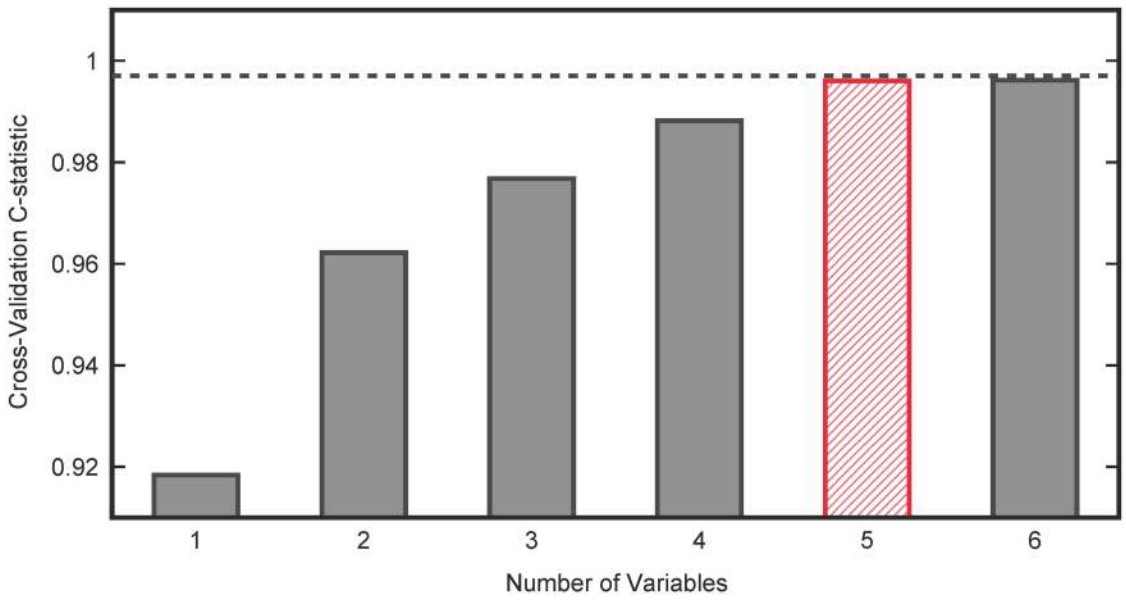

Fig 4. Selecting the Number of Variables for FDA based on C-statistic. Five variables were found to be sufficient for separating the ASD and NEU groups while an additional two variables (totaling seven variables) were incorporated to retain separation between ASD and SIB cohorts.

doi:10.1371/journal.pcbi.1005385.g004

maximum C-statistic for all combinations of a given number of variables. As the number of variables increases, the C-statistic increases, saturates at 0.997, and then slightly decreases when over-fitting occurs.

From these results, five variables (DNA methylation, 8-OHG, Glu.-Cys., fCystine/fCysteine, % oxidized glutathione) were considered for further analysis; however, it should be noted that select variable combinations distinct from this one provided similar performance for separating ASD and NEU participants. Chlorotyrosine and tGSH/GSSG were added to this set to improve separability of the ASD and SIB groups, increasing the number of metabolites under consideration to seven. The separability of the final minimal classifier based on these seven variables is presented in Fig 5 with Type I and Type II error plots in S1 Fig.

Prediction of adaptive behavior in ASD

In addition to separation into neurologically distinct cohorts, metabolites in the FOCM/TS pathway were investigated for predictability of adaptive behavior. Due to the inter-dependency of pathway metabolites and possible nonlinear effects on psychological outcomes, nonlinear regression via KPLS was used to evaluate the ability of pathway metabolites to predict adaptive behavior in ASD (as measured by the Vineland Adaptive Behavior Composite score). Just as was done in the FDA analysis, all combinations of a given number of variables were evaluated for predictability. The cross-validatory R^2 of the regression was then used to determine the optimal number of variables in the regression analysis. From the results in Fig 6, the R^2 begins to decrease when more than five variables are used in the KPLS analysis. The maximum cross-validatory R^2 was 0.45, corresponding to the KPLS model with the variable combination GSSG, tGSH/GSSG, Nitrotyrosine, Tyrosine, and fCysteine used as inputs. These regression results are plotted in Fig 6. (It is important to note that a few other variable combinations provided similar results, but only the best regression model is illustrated for clarity.) This strong

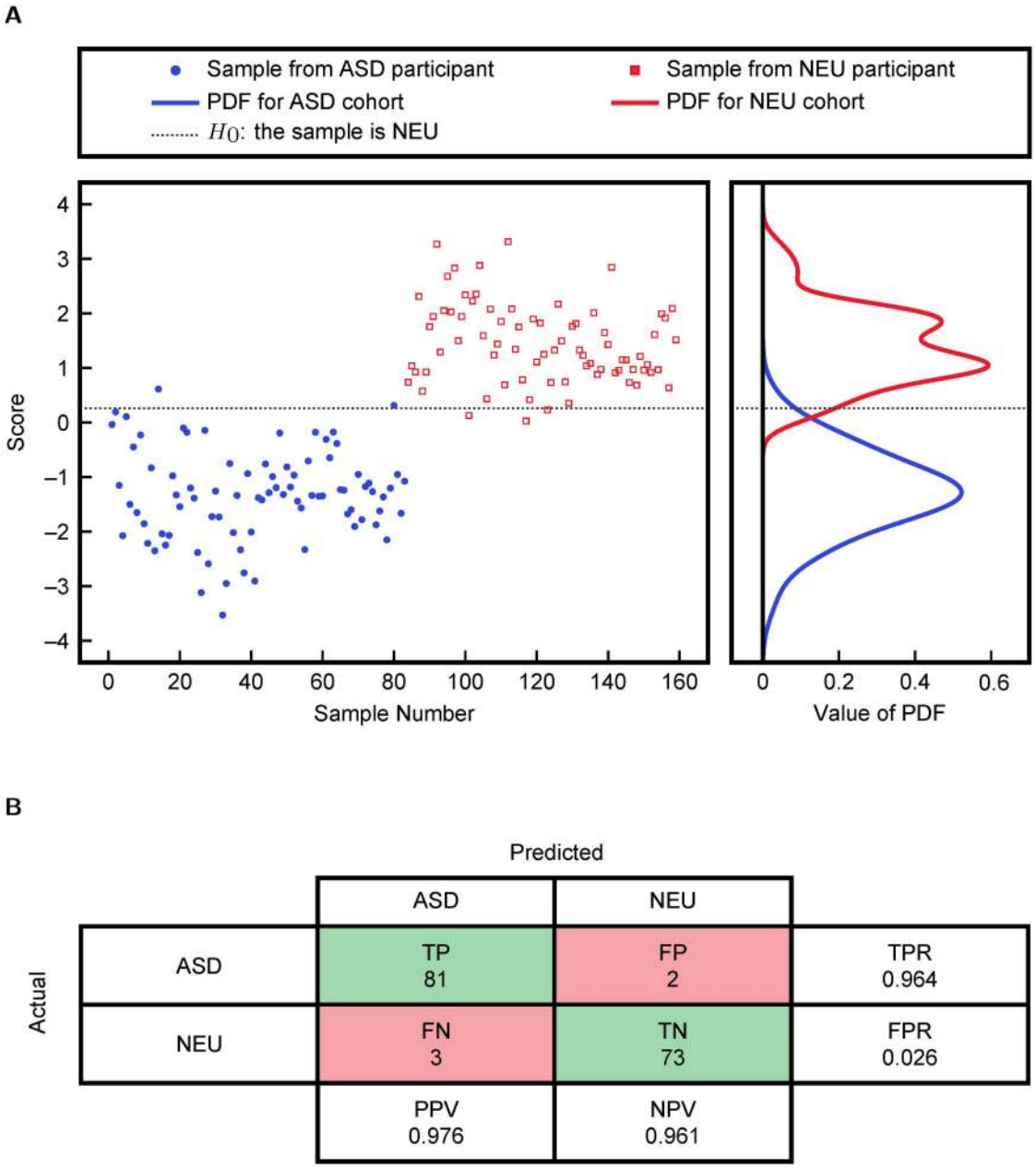

Fig 5. FDA analysis and binary classification using the variables DNA methylation, 8-OHG, Glu.-Cys., fCystine/fCysteine, % oxidized glutathione, Chlorotyrosine, and tGSH/GSSG. (a) individual cross-validated FDA scores and fitted probability distribution functions and (b) the cross-validated confusion matrix for separation of ASD and neurotypical (NEU) groups. TPR = TP/(TP + FN) is the True Positive Rate, FPR = FP/(FP + TN) is the False Positive Rate, PPV = TP/(TP + FP) is the Positive Predictive Value, and NPV = TN/(TN + FN) is the Negative Predictive Value.

doi:10.1371/journal.pcbi.1005385.g005

correlation even after cross-validation indicates the importance of FOCM/TS dysfunction in the pathophysiology of ASD.

Discussion

The multivariate statistical analysis presented herein provides unprecedented quantitative classification results for separating participants into ASD and NEU cohorts based solely on

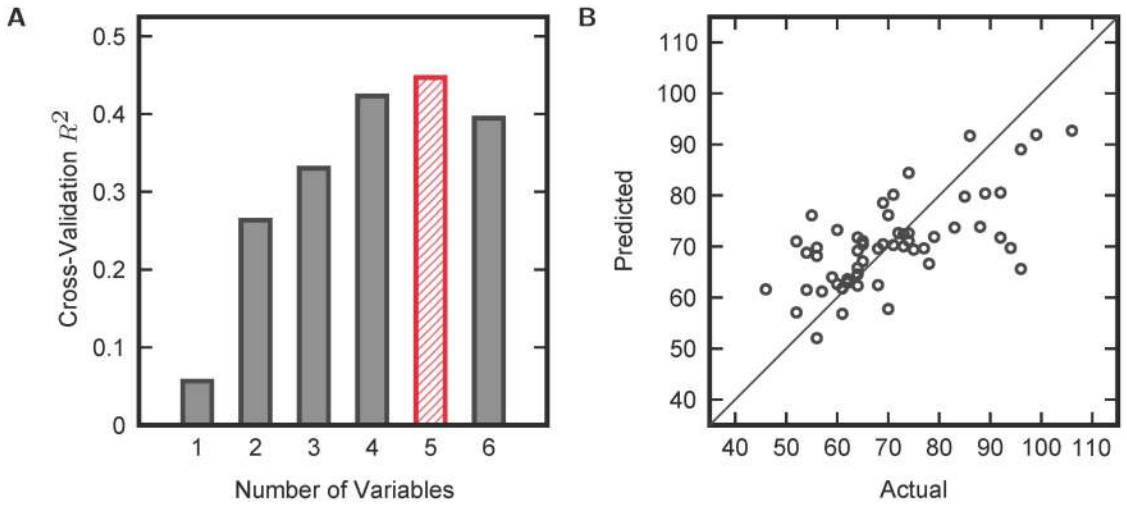

Fig 6. KPLS regression results. (a) maximum cross-validated R^2 for a given number of variables and (b) cross-validated model predictions versus actual data points for the best combination of five variables (GSSG, tGSH/GSSG, Nitrotyrosine, Tyrosine, and fCysteine).

doi:10.1371/journal.pcbi.1005385.g006

biochemical data. Existing analyses report differences in mean metabolite levels or provide qualitative illustrations of separating these two groups based on FOCM/TS metabolites [18, 37, 38]. However, these strategies are not designed for classification and thus fail to successfully classify participants. Here, FDA on seven metabolites allows sufficient separation such that a linear classifier can correctly resolve 96.9% of participants. Such low misclassification rates dissuaded the use of more complex, nonlinear methods such as KFDA. Although FOCM/TS dysfunction likely does not completely detail ASD etiology, this biochemical analysis approaches the accuracy needed for a clinical diagnostic tool.

Classification performance on the SIB group fortifies the argument for FOCM/TS involvement in ASD since the large degree of shared genetic and environmental effects with the ASD population only slightly worsens the separation. The sibling recurrence rate for ASD is estimated to be 6.9–18.7% [7, 44, 45] and many siblings perform behaviorally and/or cognitively at intermediate levels between those of ASD and NEU cohorts [45–47] or express traits characteristic of ASD [47–49]. Therefore, the classification performance placing the SIB group between the ASD and NEU groups, albeit much closer to the NEU group, is consistent with the broader scientific literature on psychometric analysis of siblings of people with ASD. Future work would benefit from assessing the SIB and NEU groups on measurements of the Broader Autism Phenotype to validate these hypotheses on mild FOCM/TS dysfunction in the SIB group.

Comparison or meta-analysis of regression analyses across studies is difficult due to differences in metabolites measured, origin of metabolites, available psychometric data, and metrics of model performance. It is emphasized that extreme caution should be used when evaluating fitted versus cross-validated metrics; for example, in [39], the best linear model can achieve a fitting R^2 of 0.296, while obtaining a cross-validated R^2 of only 0.192. In general, fitting results always surpass cross-validation results; nevertheless, the top-performing KPLS model in this study achieved a cross-validatory R^2 of 0.45 due to its ability to reflect nonlinear behaviors/interactions, which surpasses or compares with previous fitting [50, 51] and cross-validated results [39].

Nonlinear regression analysis of FOCM/TS metabolites enables prediction of key FOCM/TS metabolites that are associated with adaptive behavior in ASD. Based upon all variable combinations evaluated in the KPLS regression analysis, top-performing models always incorporated (1) nitrotyrosine, (2) tyrosine, (3) fGSH or tGSH/GSSG, and (4) fCysteine or fCystine/fCysteine. Interestingly, these variables are affected by high quality vitamin supplementation that also decreases ASD severity in at least a subset of cases [51–53]. While this forms an intriguing direction for future studies, it should be noted that these studies should be replicated and empirically tested on a wider scale before more definite conclusions can be drawn. Furthermore, this approach can be extended to include other psychometric instruments (e.g. the Autism Diagnostic Observation Schedule (ADOS) or Childhood Autism Rating Scales (CARS)) that are more appropriate for diagnosis of ASD.

Developmental pediatricians, psychologists and other professionals can effectively use the wealth of information provided by psychometric instruments to diagnose and evaluate patients with ASD. However, these tests can rarely diagnose children under two years old since they are based solely on behavioral assessment. As it is generally acknowledged that an earlier diagnosis can lead to a more favorable outcome in the long run [54], the identification of biomarkers which can be used in conjunction with psychometric measurements would be of significant importance for ASD diagnosis. Furthermore, identification of these biomarkers can facilitate the understanding of these complex disorders, which offers significant potential for developing intervention strategies targeted to normalize these biomarkers in the future. However, it is important to note that these biomarkers may not simply be measurements of certain metabolites but may require nonlinear statistical analysis of the measurements, as is done in this work.

Materials and methods

Description of data

The data used in this study comes from the Arkansas Children's Hospital Research Institute's autism IMAGE study [38]. The protocol was approved by the Institutional Review Board at the University of Arkansas for Medical Sciences and all parents signed informed consent. The interested reader is referred to [38] for detailed study design, including demographic information and inclusion/exclusion criteria. Briefly, children between the ages of 3 and 10 years were enrolled to assess levels of oxidative stress. ASD was defined by the *Diagnostic and Statistical Manual for Mental Disorders, Fourth Edition*, the Autism Diagnostic Observation Schedule (ADOS), and/or the Childhood Autism Rating Scales (CARS; score > 30). FOCM/TS metabolites from 83, 47, and 76 case (ASD), sibling (SIB), and age-matched control (NEU) children, respectively, were used for classification. The metabolites under investigation are tabulated in Table 1 and additional details of these measurements and derivations are presented in [38]. Of the 83 participants on the autism spectrum, 55 also had Vineland II Scores recorded for use in regression analysis (range 46–106). The Vineland Adaptive Behavior Composite evaluates adaptive skills across the domains of communication, socialization, daily living skills, and motor skills through a semi-structured caregiver interview [55]. Data are available in S1 Dataset.

Fisher Discriminant Analysis

Fisher Discriminant Analysis (FDA) is a dimensionality reduction tool that seeks to maximize differences between multiple classes. Specifically, for n samples of m measurements associated

with k different classes, the between cluster variability S_B is defined to be

$$S_B = \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})^{\mathrm{T}}$$

where $\bar{x}_i$ represents the mean vector of class i, $\bar{x}$ represents the mean vector of all samples, and n_i represents the number of samples in class i. The within cluster variation is defined as

$$S_W = \sum_{i=1}^{k} n_i \sum_{j \in i} (x_j - \bar{x}_i)(x_j - \bar{x}_i)^{\mathrm{T}}$$

where x_j represents an individual sample. FDA seeks to find at most $k - 1$ vectors that maximize

$$J(w) = \frac{w^{\mathrm{T}} S_B w}{w^{\mathrm{T}} S_W w}$$

In other words, FDA seeks to find linear combinations of variables that project samples in the same group close to each other and project samples in different groups far away from each other. The solution to this optimization problem is the generalized eigenvectors associated with the $k - 1$ largest generalized eigenvalues of $S_W^{-1} S_B$.

Kernel density estimation

Kernel density estimation attempts to determine the underlying probability distribution function from a set of reference samples. The main assumption is that additional samples are likely to be found near the reference samples [56–58]. Using a Gaussian kernel, this assumption is formulated into an algorithm by associating a kernel function

$$K\left(\frac{x - x_i}{\sigma}\right)$$

with each observation x_i. Here, x is the additional sample and σ is the kernel parameter that controls the shape of the distribution function. The estimated density function $\hat{f}(x)$ is then given by

$$\hat{f}(x) = \frac{1}{n\sigma} \sum_{i=1}^{n} K\left(\frac{x - x_i}{\sigma}\right)$$

where n is the number of reference samples. The kernel parameter σ is chosen to minimize the mean integrated squared error (MISE) between the unknown density function $f(x)$ and the estimated density function $\hat{f}(x)$:

$$\mathrm{MISE}(\sigma) = \int_{-\infty}^{\infty} \left(f(x) - \hat{f}(x)\right)^2$$

using a cross-validatory approach [56].

Kernel partial least squares

Kernel techniques provide general nonlinear extensions to the popular linear partial least squares (PLS) regression. The KPLS algorithm commences by defining a nonlinear transformation $f = \psi(x)$ on the predictor set x. In this work, $\psi(x)$ is a Gaussian kernel. Rather

than regression on x as in linear PLS, y is regressed onto the high dimensional feature space f [42, 43].

Cross-validation

To avoid over-fitting and over-stating results, leave-one-out cross validation is employed in both the FDA and KPLS analysis. The approach leaves out a single sample, fits an FDA or KPLS model, and evaluates the prediction of the sample left out. This scheme is repeated for each sample.

Supporting information

S1 Dataset. Biochemical and Adaptive Behavior Data from ASD, NEU, and SIB Participants.
(CSV)

S1 Fig. Type I and II Errors for the Final FDA Model. Cross-validated type I and type II errors for the FDA model using the variables DNA methylation, 8-OHG, Glu.-Cys., fCystine/fCysteine, % oxidized, Chlorotyrosine, and tGSH/GSSG.
(TIF)

Author Contributions

Conceptualization: DPH UK SJJ JH.

Data curation: SM SJJ.

Formal analysis: DPH UK.

Funding acquisition: JH.

Investigation: DPH SM UK.

Methodology: DPH UK SM SJJ JH.

Project administration: JH.

Resources: SM SJJ.

Software: DPH UK.

Supervision: SJJ JH.

Validation: DPH UK JH.

Visualization: DPH.

Writing – original draft: DPH.

Writing – review & editing: DPH UK SJJ JH.

References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders. 5th ed. American Psychiatric Association; 2013.
2. Levy SE, Giarelli E, Lee LC, Schieve LA, Kirby RS, Cunniff C, et al. Autism spectrum disorder and co-occurring developmental, psychiatric, and medical conditions among children in multiple populations of the United States. Journal of Developmental & Behavioral Pediatrics. 2010; 31(4):267–275.

3. Perrin JM, Coury DL, Hyman SL, Cole L, Reynolds AM, Clemons T. Complementary and Alternative Medicine Use in a Large Pediatric Autism Sample. Pediatrics. 2012; 130(Supplement 2):S77–S82. doi: 10.1542/peds.2012-0900E PMID: 23118257
4. Pulcini CD, Perrin JM, Houtrow AJ, Sargent J, Shui A, Kuhlthau K. Examining Trends and Coexisting Conditions Among Children Qualifying for SSI Under ADHD, ASD, and ID. Academic Pediatrics. 2015; 15(4):439–443. doi: 10.1016/j.acap.2015.05.002 PMID: 26142070
5. Saunders A, Kirk IJ, Waldie KE. Autism Spectrum Disorder and Co-Existing Conditions: A Lexical Decision Erp Study. Clin Exp Psychol. 2015; 1(001). doi: 10.4172/2471-2701.1000101
6. Centers for Disease Control and Prevention. Prevalence of Autism Spectrum Disorders—Autism and Developmental Disabilities Monitoring Network, 14 Sites, United States, 2008. Morbidity and Mortality Weekly Report. 2012; 61:1–19.
7. Grønborg TK, Schendel DE, Parner ET. Recurrence of autism spectrum disorders in full- and half-siblings and trends over time: a population-based cohort study. JAMA pediatrics. 2013; 167(10): 947–953. doi: 10.1001/jamapediatrics.2013.2259 PMID: 23959427
8. Maenner MJ, Rice CE, Arneson CL, Cunniff C, Schieve LA, Van Naarden Braun K, et al. Potential impact of DSM-5 criteria on autism spectrum disorder prevalence estimates. JAMA Psychiatry. 2014; 71(3):292–300. doi: 10.1001/jamapsychiatry.2013.3893 PMID: 24452504
9. van Heijst BF, Geurts HM. Quality of life in autism across the lifespan: A meta-analysis. Autism. 2015; 19(2):158–167. doi: 10.1177/1362361313517053 PMID: 24443331
10. Buescher AVS, Cidav Z, Knapp M, Mandell DS. Costs of Autism Spectrum Disorders in the United Kingdom and the United States. JAMA Pediatrics. 2014; 168(8):721–728. doi: 10.1001/jamapediatrics.2014.210 PMID: 24911948
11. Ronald A, Hoekstra RA. Autism spectrum disorders and autistic traits: A decade of new twin studies. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics. 2011; 156(3):255–274. doi: 10.1002/ajmg.b.31159
12. Gaugler T, Klei L, Sanders SJ, Bodea CA, Goldberg AP, Lee AB, et al. Most genetic risk for autism resides with common variation. Nature Genetics. 2014; 46(8):881–885. doi: 10.1038/ng.3039 PMID: 25038753
13. Mandy W, Lai MC. Annual Research Review: The role of the environment in the developmental psychopathology of autism spectrum condition. Journal of Child Psychology and Psychiatry. 2016; 57(3): 271–292. doi: 10.1111/jcpp.12501 PMID: 26782158
14. Scherer SW, Dawson G. Risk factors for autism: translating genomic discoveries into diagnostics. Human Genetics. 2011; 130(1):123–148. doi: 10.1007/s00439-011-1037-2 PMID: 21701786
15. Deth R, Muratore C, Benzecry J, Power-Charnitsky VA, Waly M. How environmental and genetic factors combine to cause autism: A redox/methylation hypothesis. NeuroToxicology. 2008; 29(1): 190–201. doi: 10.1016/j.neuro.2007.09.010 PMID: 18031821
16. Jurecka A, Zikanova M, Kmoch S, Tylki-Szymańska A. Adenylosuccinate lyase deficiency. Journal of Inherited Metabolic Disease. 2014; 38(2):231–242. doi: 10.1007/s10545-014-9755-y PMID: 25112391
17. Pu D, Shen Y, Wu J. Association between MTHFR Gene Polymorphisms and the Risk of Autism Spectrum Disorders: A Meta-Analysis. Autism Research. 2013; 6(5):384–392. doi: 10.1002/aur.1300 PMID: 23653228
18. James SJ, Melnyk S, Jernigan S, Cleves MA, Halsted CH, Wong DH, et al. Metabolic endophenotype and related genotypes are associated with oxidative stress in children with autism. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics. 2006; 141B(8):947–956. doi: 10.1002/ajmg.b.30366
19. James SJ, Melnyk S, Jernigan S, Pavliv O, Trusty T, Lehman S, et al. A functional polymorphism in the reduced folate carrier gene and DNA hypomethylation in mothers of children with autism. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics. 2010; 153B(6):1209–1220.
20. Mohammad NS, Jain JMN, Chintakindi KP, Singh RP, Naik U, Akella RRD. Aberrations in folate metabolic pathway and altered susceptibility to autism. Psychiatric Genetics. 2009; 19(4):171–176. PMID: 19440165
21. Schmidt RJ, Hansen RL, Hartiala J, Allayee H, Schmidt LC, Tancredi DJ, et al. Prenatal vitamins, one-carbon metabolism gene variants, and risk for autism. Epidemiology (Cambridge, Mass). 2011; 22(4): 476–485.
22. Schaevitz LR, Berger-Sweeney JE. Gene-Environment Interactions and Epigenetic Pathways in Autism: The Importance of One-Carbon Metabolism. ILAR Journal. 2012; 53(3–4):322–340. doi: 10.1093/ilar.53.3-4.322 PMID: 23744970

23. Bromley RL, Mawer GE, Briggs M, Cheyne C, Clayton-Smith J, García-Fiñana M, et al. The prevalence of neurodevelopmental disorders in children prenatally exposed to antiepileptic drugs. Journal of Neurology, Neurosurgery & Psychiatry. 2013; 84(6):637–643. doi: 10.1136/jnnp-2012-304270

24. Christensen J, Grønborg TK, Sørensen MJ, Schendel D, Parner ET, Pedersen LH, et al. Prenatal valproate exposure and risk of autism spectrum disorders and childhood autism. JAMA. 2013; 309(16): 1696–1703. doi: 10.1001/jama.2013.2270 PMID: 23613074

25. Roullet FI, Lai JKY, Foster JA. In utero exposure to valproic acid and autism—A current review of clinical and animal studies. Neurotoxicology and Teratology. 2013; 36:47–56. PMID: 23395807

26. Dong E, Chen Y, Gavin DP, Grayson DR, Guidotti A. Valproate induces DNA demethylation in nuclear extracts from adult mouse brain. Epigenetics. 2010; 5(8):730–735. doi: 10.4161/epi.5.8.13053 PMID: 20716949

27. Wang Z, Xu L, Zhu X, Cui W, Sun Y, Nishijo H, et al. Demethylation of Specific Wnt/β-Catenin Pathway Genes and its Upregulation in Rat Brain Induced by Prenatal Valproate Exposure. The Anatomical Record: Advances in Integrative Anatomy and Evolutionary Biology. 2010; 293(11):1947–1953. doi: 10.1002/ar.21232

28. Fuller LC, Cornelius SK, Murphy CW, Wiens DJ. Neural crest cell motility in valproic acid. Reproductive Toxicology. 2002; 16(6):825–839. doi: 10.1016/S0890-6238(02)00059-X PMID: 12401512

29. Landrigan PJ. What causes autism? Exploring the environmental contribution. Current Opinion in Pediatrics. 2010; 22(2):219–225. PMID: 20087185

30. Shelton JF, Hertz-Picciotto I, Pessah IN. Tipping the Balance of Autism Risk: Potential Mechanisms Linking Pesticides and Autism. Environmental Health Perspectives. 2012; 120(7):944–951. doi: 10.1289/ehp.1104553 PMID: 22534084

31. Kim J, Hannibal L, Gherasim C, Jacobsen DW, Banerjee R. A Human Vitamin B12 Trafficking Protein Uses Glutathione Transferase Activity for Processing Alkylcobalamins. Journal of Biological Chemistry. 2009; 284(48):33418–33424. doi: 10.1074/jbc.M109.057877 PMID: 19801555

32. Volk, Lurmann F, Penfold B, Hertz-Picciotto I, McConnell R. Traffic-related air pollution, particulate matter, and autism. JAMA Psychiatry. 2013; 70(1):71–77. doi: 10.1001/jamapsychiatry.2013.266 PMID: 23404082

33. Boggess A, Faber S, Kern J, Kingston HMS. Mean serum-level of common organic pollutants is predictive of behavioral severity in children with autism spectrum disorders. Scientific Reports. 2016; 6:26185. doi: 10.1038/srep26185 PMID: 27174041

34. Schmidt RJ, Tancredi DJ, Ozonoff S, Hansen RL, Hartiala J, Allayee H, et al. Maternal periconceptional folic acid intake and risk of autism spectrum disorders and developmental delay in the CHARGE (CHildhood Autism Risks from Genetics and Environment) case-control study. The American Journal of Clinical Nutrition. 2012; 96(1):80–89. doi: 10.3945/ajcn.110.004416 PMID: 22648721

35. Surén P, Roth C, Bresnahan M, Haugen M, Hornig M, Hirtz D, et al. Association between maternal use of folic acid supplements and risk of autism spectrum disorders in children. JAMA. 2013; 309(6): 570–577. doi: 10.1001/jama.2012.155925

36. Blom HJ. Folic acid, methylation and neural tube closure in humans. Birth Defects Research Part A: Clinical and Molecular Teratology. 2009; 85(4):295–302. doi: 10.1002/bdra.20581

37. James SJ, Cutler P, Melnyk S, Jernigan S, Janak L, Gaylor DW, et al. Metabolic biomarkers of increased oxidative stress and impaired methylation capacity in children with autism. The American Journal of Clinical Nutrition. 2004; 80(6):1611–1617. PMID: 15585776

38. Melnyk S, Fuchs GJ, Schulz E, Lopez M, Kahler SG, Fussell JJ, et al. Metabolic Imbalance Associated with Methylation Dysregulation and Oxidative Damage in Children with Autism. Journal of Autism and Developmental Disorders. 2012; 42(3):367–377. doi: 10.1007/s10803-011-1260-7 PMID: 21519954

39. Adams J, Howsmon DP, Kruger U, Geis E, Gehn E, Fimbres V, et al. Significant Association of Urinary Toxic Metals and Autism-Related Symptoms: A Nonlinear Statistical Analysis with Cross Validation. PLOS One. 2017; 12(1):e0169526. doi: 10.1371/journal.pone.0169526 PMID: 28068407

40. Rao CR. The utilization of multiple measurements in problems of biological classification. Journal of the Royal Statistical Society Series B (Methodological). 1948; 10(2):159–203.

41. Mika S, R atsch G, Weston J, Sch olkopf B, M uller KR. Fisher Discriminant Analysis with Kernels. In: Proceedings of the Neural Networks for Signal Processing IX Workshop; 1999. p. 41–48.

42. Rosipal R, Trejo LJ. Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space. J Mach Learn Res. 2002; 2:97–123.

43. Kim K, Lee JM, Lee IB. A novel multivariate regression approach based on kernel partial least squares with orthogonal signal correction. Chemometrics and Intelligent Laboratory Systems. 2005; 79(1–2): 22–30. doi: 10.1016/j.chemolab.2005.03.003

44. Ozonoff S, Young GS, Carter A, Messinger D, Yirmiya N, Zwaigenbaum L, et al. Recurrence Risk for Autism Spectrum Disorders: A Baby Siblings Research Consortium Study. Pediatrics. 2011; p. 2010–2825.
45. Constantino JN, Zhang Y, Frazier T, Abbacchi AM, Law P. Sibling recurrence and the genetic epidemiology of autism. The American journal of psychiatry. 2010; 167(11):1349–1356. doi: 10.1176/appi.ajp.2010.09101470 PMID: 20889652
46. Gizzonio V, Avanzini P, Fabbri-Destro M, Campi C, Rizzolatti G. Cognitive abilities in siblings of children with autism spectrum disorders. Experimental Brain Research. 2014; 232(7):2381–2390. doi: 10.1007/s00221-014-3935-8 PMID: 24710667
47. Ruzich E, Allison C, Smith P, Watson P, Auyeung B, Ring H, et al. Subgrouping siblings of people with autism: Identifying the broader autism phenotype. Autism Research. 2016; 9(6):658–665. doi: 10.1002/aur.1544 PMID: 26332889
48. Messinger D, Young GS, Ozonoff S, Dobkins K, Carter A, Zwaigenbaum L, et al. Beyond Autism: A Baby Siblings Research Consortium Study of High-Risk Children at Three Years of Age. Journal of the American Academy of Child and Adolescent Psychiatry. 2013; 52(3):300–308. doi: 10.1016/j.jaac.2012.12.011 PMID: 23452686
49. Pisula E, Ziegart-Sadowska K. Broader Autism Phenotype in Siblings of Children with ASD—A Review. International Journal of Molecular Sciences. 2015; 16(6):13217–13258. doi: 10.3390/ijms160613217 PMID: 26068453
50. Adams JB, Audhya T, McDonough-Means S, Rubin RA, Quig D, Geis E, et al. Nutritional and metabolic status of children with autism vs. neurotypical children, and the association with autism severity. Nutrition & Metabolism. 2011; 8:34. doi: 10.1186/1743-7075-8-34
51. Frye RE, Melnyk S, Fuchs G, Reid T, Jernigan S, Pavliv O, et al. Effectiveness of Methylcobalamin and Folinic Acid Treatment on Adaptive Behavior in Children with Autistic Disorder Is Related to Glutathione Redox Status. Autism Research and Treatment. 2013; 2013:e609705. doi: 10.1155/2013/609705
52. James SJ, Melnyk S, Fuchs G, Reid T, Jernigan S, Pavliv O, et al. Efficacy of methylcobalamin and folinic acid treatment on glutathione redox status in children with autism. The American Journal of Clinical Nutrition. 2009; 89(1):425–430. doi: 10.3945/ajcn.2008.26615 PMID: 19056591
53. Adams JB, Audhya T, McDonough-Means S, Rubin RA, Quig D, Geis E, et al. Effect of a vitamin/mineral supplement on children and adults with autism. BMC Pediatrics. 2011; 11:111. doi: 10.1186/1471-2431-11-111 PMID: 22151477
54. Zwaigenbaum L, Bryson S, Garon N. Early identification of autism spectrum disorders. Behavioural Brain Research. 2013; 251:133–146. doi: 10.1016/j.bbr.2013.04.004 PMID: 23588272
55. Sparrow SS, Cicchetti DV, Balla DA. Vineland Adaptive Behavior Scales. 2nd ed. Minneapolis, MN: Pearson Assessments; 2005.
56. Silverman BW. Density Estimation for Statistics and Data Analysis. CRC Press; 1986.
57. Chen Q, Kruger U, Leung AT. Regularised kernel density estimation for clustered process data. Control Engineering Practice. 2004; 12(3):267–274. doi: 10.1016/S0967-0661(03)00083-2
58. Scott DW. Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley & Sons; 2015.