RPI

Towards the Development of a Diagnostic Test for Autism Spectrum Disorder: Data Science Meets Metabolomics

Juergen Hahn
Professor & Director, Center for Biotechnology and Interdisciplinary Studies
Department of Biomedical Engineering
Department of Chemical & Biological Engineering
Rensselaer Polytechnic Institute

ASD Background

Autism Spectrum Disorder (ASD) is a neurological and developmental disorder characterized by
(1) difficulties with social communication/interaction and
(2) expression of restricted repetitive behaviors and interests

Frequently Co-occurring Conditions:
- intellectual disability
- ADHD
- speech and language delays
- psychiatric diagnoses
- epilepsy
- sleep disorders
- gastrointestinal problems

autism prevalence

AS REPORTED BY THE CDC - ADDM REPORT: 2000-2022

Birth Year	Survey Year	Year Reported	Autism Rate
2014	2022	2025	1 in 31
2012	2020	2023	1 in 36
2010	2018	2021	1 in 44
2008	2018	2020	1 in 54
2006	2014	2018	1 in 59
2004	2012	2016	1 in 69
2002	2010	2014	1 in 68
2000	2008	2012	1 in 88
1998	2006	2009	1 in 110
1996	2004	2009	1 in 125
1994	2002	2007	1 in 150
1992	2000	2007	1 in 150

TACA

RPI

ASD Background

Autism Spectrum Disorder (ASD) encompasses a large group of early-onset neurological diseases characterized by

(1) difficulties with social communication/interaction and
(2) expression of restricted repetitive behaviors and interests

1 in 31	3x	95%
of 8 year-old children has ASD	more prevalent in boys	co-occurrence with other non-ASD developmental diagnoses

RPI

2. CDC. "Prevalence of Autism Spectrum Disorders — Autism and Developmental Disabilities Monitoring Network, 16 Sites, United States, 2022" 2025.

Cost of ASD

Nationally (US): ~ **$265 billion** USD per year to care for people with ASD [3]
Individual: $1.4MM over the lifetime ($2.4MM with Intellectual Disability) [4]

Major costs come from direct, nonmedical costs (e.g., intensive education) and indirect costs (e.g., loss in caregiver productivity) [2, 4]

$210 B	$143-266 B	$236 B	$245 B
Depression [5]	ADHD [6]	Alzheimer's [7]	Diabetes [8]

RPI

3. Leigh and Du. *J Autism Dev Disord.* 2015.
4. Buescher et al. *JAMA Pediatrics* 2014.
5. Greenberg et al. *Journal of Clinical Psychiatry* 2015.
6. Doshi et al. *Journal of the American Academy of Child & Adolescent Psychiatry* 2012.
7. CDC. "Alzheimer's Disease: Promoting Health and Independence for an Aging Population At A Glance" 2016.
8. ADA. Diabetes Care 2013.

ASD Diagnosis

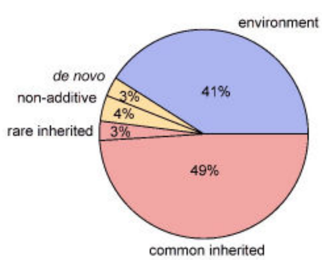
- Earlier diagnosis/treatment has shown to lead to better outcomes, however,
- ASD diagnosis is a clinical diagnosis
  - No lab test exists that can be used for general diagnostic purposes for ASD
  - It is difficult to determine if a 2 year-old is developing normally, albeit slower in some aspects, or has ASD
- Average age of ASD diagnosis is 4 years while the target for diagnosis is 18-24 months [2]

RPI

2. CDC. "Prevalence of Autism Spectrum Disorders — Autism and Developmental Disabilities Monitoring Network, 16 Sites, United States, 2022" 2025.

Current Understanding of ASD

Balance between genetics and environment



Many different body systems hypothesized to contribute

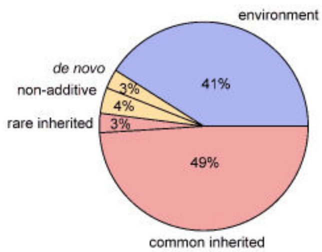
- Gut-brain axis
- Immune system
- Metabolic issues
- Mitochondria
- Oxidative Stress
- Epigenetics
- etc.

RPI

9. Gaugler et al. *Nature Genetics* 2015.

Current Understanding of ASD

Balance between genetics and environment



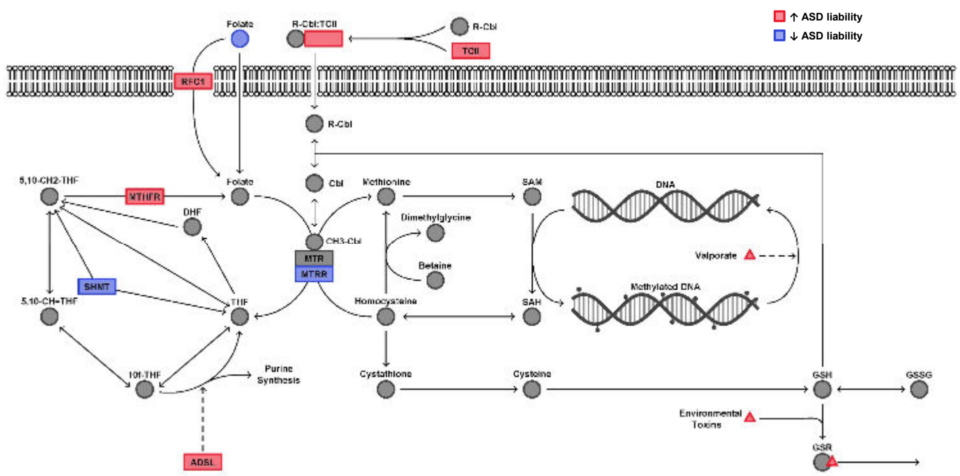
Many different body systems hypothesized to contribute

- Gut-brain axis
- Immune system
- Metabolic issues
- Mitochondria
- **Oxidative Stress**
- **Epigenetics**
- etc.

RPI

9. Gaugler et al. *Nature Genetics* 2015.

Folate-dependent One-carbon Metabolism and Transsulfuration



RPI

12. Howsmon et al. *PLoS Computational Biology* 2017.

Data Characteristics

Clinical Trial conducted at University of Arkansas for Medical Sciences[13]

Participants: **83 ASD** **47 SIB** **76 NEU**

Metabolites:

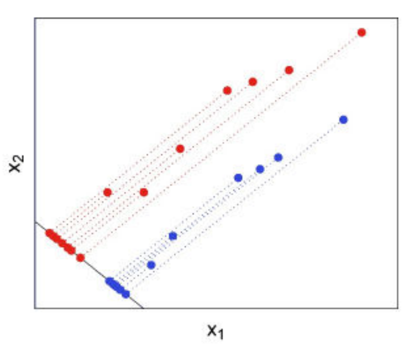
Methionine	SAM	SAH	SAM/SAH
% DNA methylation	8-OHG	Adenosine	Homocysteine
Cysteine	Glu.-Cys.	Cys.-Gly.	tGSH
fGSH	GSSG	fGSH/GSSG	tGSH/GSSG
Chlorotyrosine	Nitrotyrosine	Tyrosine	Tryptophan
fCystine	fCysteine	fCystine/fCysteine	% oxidized glutathione

ASD Behavior: Vineland Adaptive Behavior Composite

RPI

13. Melnyk et al. *Journal of Autism and Developmental Disorders* 2012.

Fisher Discriminant Analysis (FDA)



FDA: Latent variable technique that simultaneously maximizes between class scatter (S_B) and minimizes within class scatter (S_W)

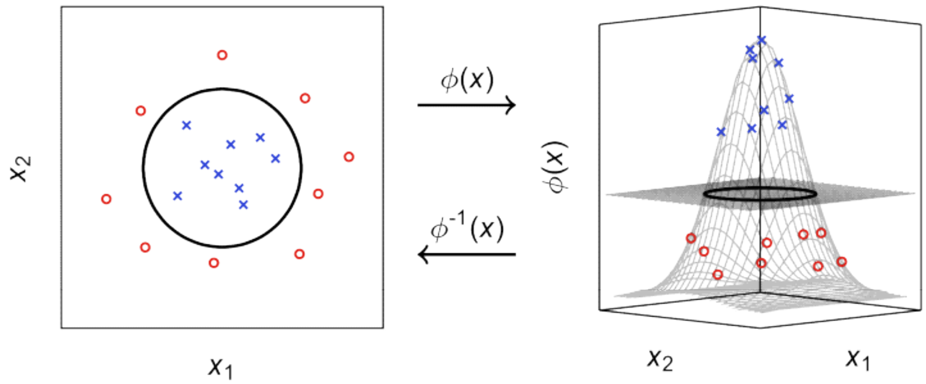
$$S_B = (\bar{x} - \bar{x})(\bar{x} - \bar{x})^T$$

$$S_W = \sum_{i \in RED} (x_i - \bar{x})(x_i - \bar{x})^T + \sum_{i \in BLUE} (x_j - \bar{x})(x_j - \bar{x})^T$$

Maximize $\quad J = \frac{w^T S_B w}{w S_W w^T}$

RPI

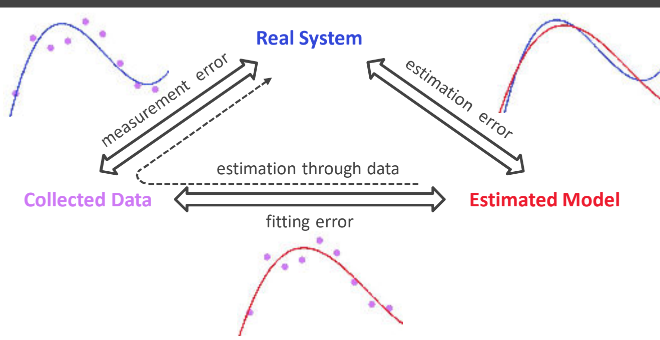
Kernel FDA



Nonlinear extensions are possible through the use of a nonlinear transformation

RPI

Parameter Estimation from Data



- Goal: reduce discrepancy between **estimated model** and **real system**
  - Implicitly attained by reducing difference between **estimated model predictions** and **collected data**

RPI

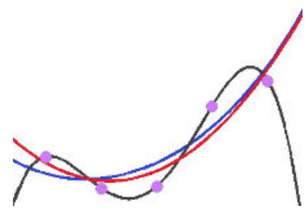
Over-fitting

- Good estimation balances retaining important information and filtering noise
- Estimating a complex model from scarce/noisy data can lead to over-fitting

true: $y = 0.1x^4 + 0.1x^3 + x^2 + \mathrm{x} + 1$

estimated: $y = \theta_4 x^4 + \theta_3 x^3 + \theta_2 x^2 + \theta_1 \mathrm{x} + \theta_0$

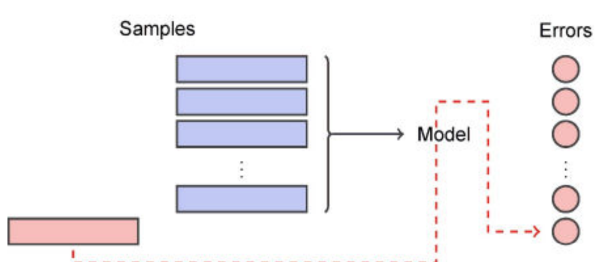
estimated: $y = \theta_2 x^2 + \theta_1 \mathrm{x} + \theta_0$



- Small fitting error can result in large estimation error
- Model should only have as many parameters as necessary

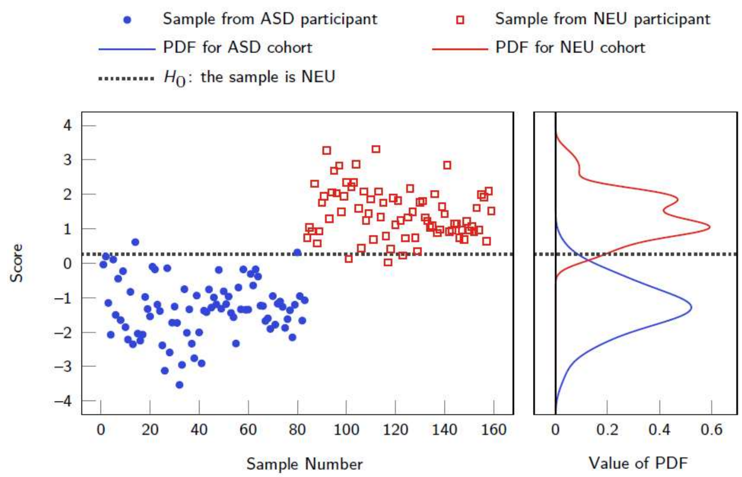
RPI

Leave-one-out Cross Validation



RPI

Analysis of Important Variables



Seven Variables:

DNA methylation,
8-OHG,
Glu.-Cys.,
fCystine/fCysteine,
% oxidized glutathione,
Chlorotyrosine, and
tGSH/GSSG

RPI

Classification Results Using Seven Inputs

Predicted

Actual	ASD	NEU	
ASD	TP 81	FP 2	TPR 0.964
NEU	FN 3	TN 73	FPR 0.026
	PPV 0.976	NPV 0.961	

RPI

Classification Summary

- Algorithm can predict with high accuracy if participant is in the ASD cohort or in the typically-developing cohort (81 of 83 ASD participants correctly categorized and 73 of 76 NEU participants correctly identified)
- Balance between Type I and Type II errors can be adjusted by modifying cut-offs
- Seven key inputs (metabolites or quantities computed from metabolite measurements) are required for classification
- Some of the key inputs are linked to oxidative stress and DNA methylation → our initial assumptions for looking at these pathways was a good one

RPI

Symptom Severity Prediction (Regression)

- Analysis has shown that classification is possible
- Is it possible to predict measures of ASD related symptoms from metabolite data?
- Regression can be performed between inputs (metabolites) and outputs (measures); regressed models need to be evaluated using cross-validation

- Note of caution: these types of studies usually have a large number of inputs compared to number of participants → overfitting is a concern that needs to be addressed

RPI

Regression: (Kernel) Partial Least Squares

Ordinary Least Squares

$$Y = XB \quad \Rightarrow \quad B = (X^T X)^{-1} X^T Y$$

- Does not account for correlations in X
- $(X^T X)^{-1}$ can be ill-conditioned

$$X \longrightarrow Y$$

Regression: (Kernel) Partial Least Squares

Ordinary Least Squares

$$Y = XB \quad \Rightarrow \quad B = (X^T X)^{-1} X^T Y$$

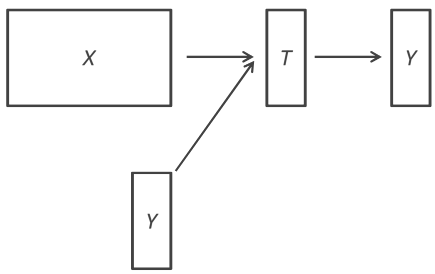
- Does not account for correlations in X
- $(X^T X)^{-1}$ can be ill-conditioned

Principal Component Regression

- Removes unimportant correlations in X
- Still ignores correlations in Y and correlations between X and Y

$$X \longrightarrow T \longrightarrow Y$$

Regression: (Kernel) Partial Least Squares



Ordinary Least Squares

$$Y = XB \quad \Rightarrow \quad B = (X^T X)^{-1} X^T Y$$

- Does not account for correlations in X
- $(X^T X)^{-1}$ can be ill-conditioned

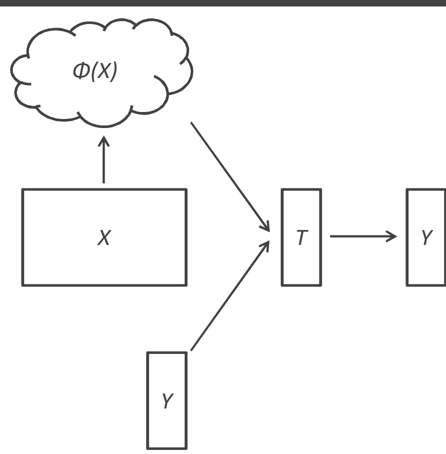
Principal Component Regression

- Removes unimportant correlations in X
- Still ignores correlations in Y and correlations between X and Y

Partial Least Squares

- Removes unimportant correlations in X and Y and between X and Y
- Robust even for relatively small data sets

Regression: (Kernel) Partial Least Squares



Ordinary Least Squares

$$Y = XB \quad \Rightarrow \quad B = (X^T X)^{-1} X^T Y$$

- Does not account for correlations in X
- $(X^T X)^{-1}$ can be ill-conditioned

Principal Component Regression

- Removes unimportant correlations in X
- Still ignores correlations in Y and correlations between X and Y

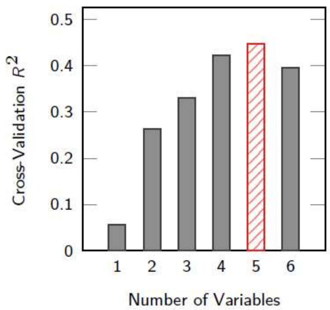
Partial Least Squares

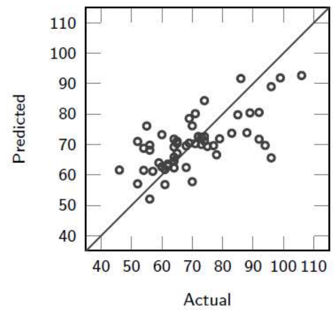
- Removes unimportant correlations in X and Y and between X and Y
- Robust even for relatively small data sets

Kernel Partial Least Squares

- Extends PLS to nonlinear transformations via a nonlinear transformation
- Consider Gaussian kernels in this work

KPLS: Predicting Vineland Adaptive Behavior Composite





Top-performing models always incorporated
1. nitrotyrosine,
2. tyrosine,
3. fGSH or tGSH/GSSG, and
4. fCysteine or fCystine/fCysteine

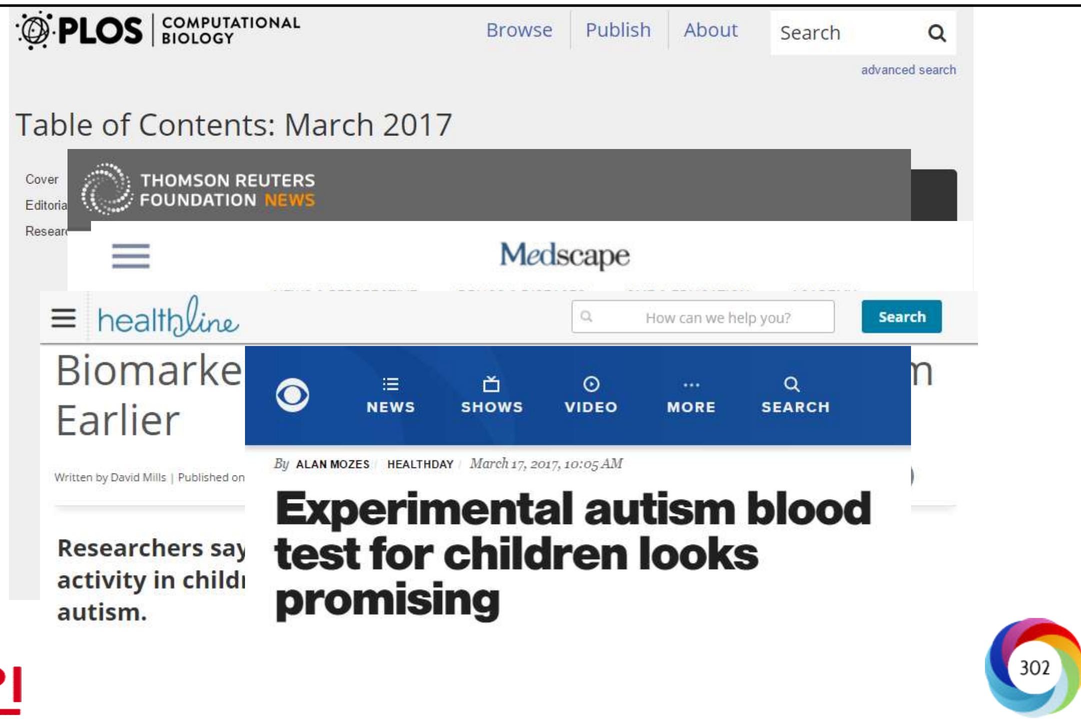
Good correlation can be found with cross-validatory R^2 of 0.45

RPI

Summary

- No reliable biomarker for ASD exists today
- Using data from one study, we have shown that
  - Metabolite concentration measurements can be used to predict if an individual has ASD or is typically-developing with a high degree of accuracy
  - Certain measures of ASD related symptoms can be predicted
- Significant potential impact of this work
- Next step: validate findings using data from another study

RPI



RPI

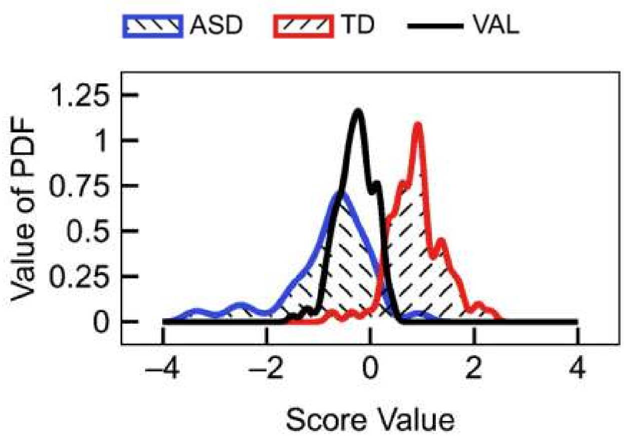
Validation Study

- Starting a new clinical trial is costly and time-consuming; while this will be necessary, we first want to make use of other data that are already available
- Study includes 154 participants with ASD but no typically developing peers
- Almost all metabolites from our previous study were measured, except % DNA methylation and 8-OHG
  - Since these two were part of the seven most important variables, validation "out of the box" cannot be used
  - Identified best model without % DNA methylation and 8-OHG from original study and then test classification on new data
  - Best model involves five variables: SAM/SAH, Glu-Cys, GSSG, fCystine/fCysteine, % oxidized glutathione

RPI

14. Frye et al. *Molecular Psychiatry* 2016.

Validation Study



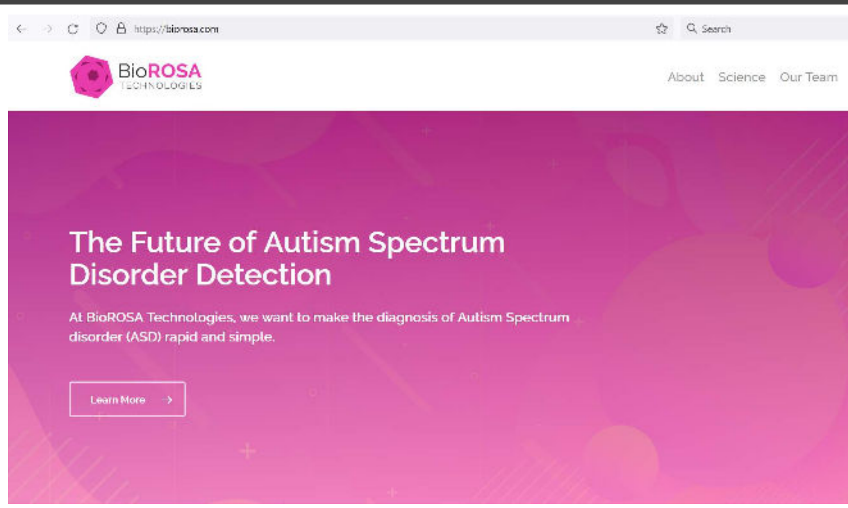
Results are promising

Note that the new data set is used only for validation and the model has no knowledge about this data set

135 out of 154 participants are correctly identified (88%) with five inputs

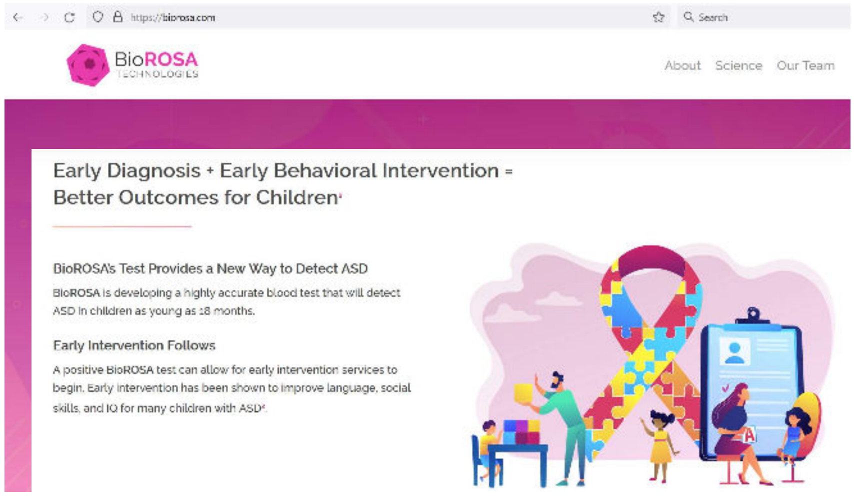
RPI

Commercialization



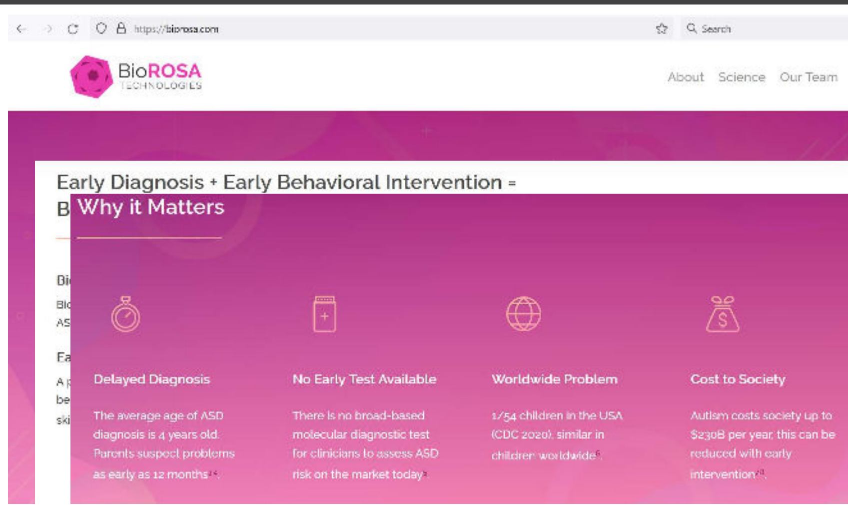
RPI

Commercialization



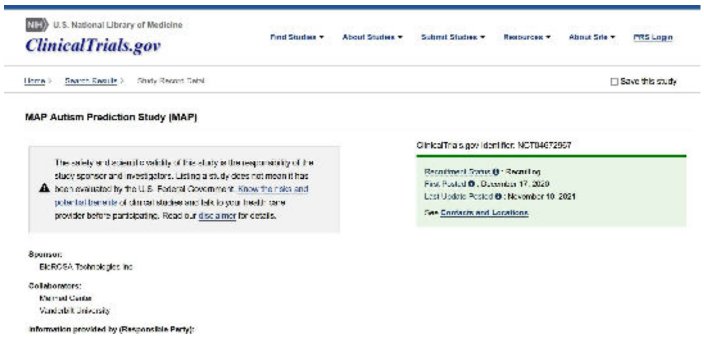
RPI

Commercialization



RPI

Commercialization



- Shift of focus to distinguish children with ASD from children with non-ASD related developmental delays (DD)

- Recruited 140 participants from developmental pediatric clinics
- Two-site trial (Vanderbilt University, TN; Melmed Center/Cortica, AZ)
- Prospective and blinded

RPI

Commercialization

- Goal was to obtain accuracy exceeding 80% to distinguish between ASD and DD cohorts
- Challenges:
  – Since Trial was prospective, number of ASD and DD cases was unknown at time of recruitment
  – # of participant was based on past % of children being diagnosed with ASD and DD in clinics
  – However, % of ASD cases has been increasing significantly over time → estimates based upon past data are only helpful to some degree
  – Goal was a 50% ASD, 50% DD split among trial participants, but ended up with a cohort that was predominantly ASD (81% ASD vs 19% DD)
  – Distinguishing ASD from DD is significantly more challenging than ASD vs TD

RPI

Commercialization

- Trial had 124 children with ASD and 26 with non-ASD related developmental delays (DD)
- Linear methods by themselves did not work
- Nonlinear (kernel-based) methods have data requirements that cannot be met with only 26 DD
- Solution:
  - Create artificial inputs from nonlinear combinations of existing inputs (feature engineering)
  - Use linear techniques for the actual classification
  - Select the best set of five inputs from 123 “engineered” inputs based upon the original FOCM/TS measurements

RPI

Commercialization

- Goal was to obtain accuracy exceeding 80% to distinguish between ASD and DD cohorts
- Achieved accuracy of 91.4% using five inputs and SVM

		Predicted		
		ASD	DD	
Actual	ASD	107	7	Sensitivity 93.86%
	DD	5	21	Specificity 80.77%
		Precision 95.54%	NPV 75.00%	Accuracy 91.43%

RPI

Prediction of ASD Before Symptoms Appear

- Study so far focused on clinical study results that included children as young as 3 years old
- ASD can be reliably diagnosed at 2 years of age
- One important question is <u>if you can predict if children will develop ASD before behavioral symptoms appear</u>
- New clinical study is required
  – Question is how early can you predict if children will develop ASD and how accurate those predictions are
- Last part of this presentation we will focus on going back to the time prior to birth, i.e., during pregnancy, to look at if predictions can be made

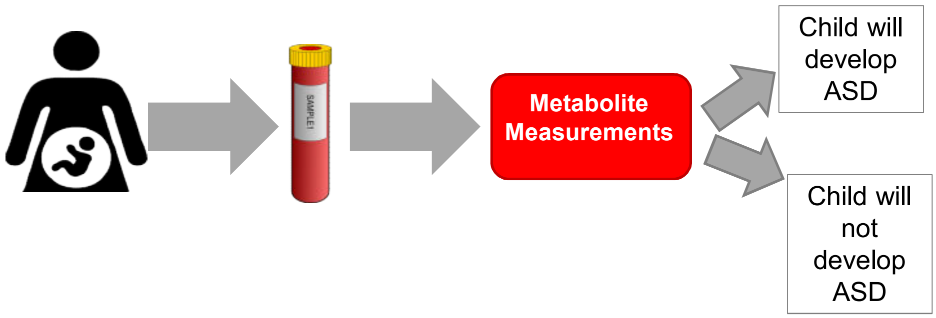
RPI

Prediction of ASD Before Symptoms Appear

- Challenge is ASD prevalence of ~3%:
  – Study with 200 participants will only have ~6 children develop ASD which is statistically not meaningful
  → Either dramatically increase number of participants or look at a higher-risk group
  – Large clinical studies are expensive and time consuming, ruling out this option at this point
- If a mother already had a child diagnosed with ASD then the probability of having another child with ASD is ~20%
- By collecting data only from mothers who have had a child with ASD, it is possible to obtain statistically meaningful results with 100+ participants

RPI

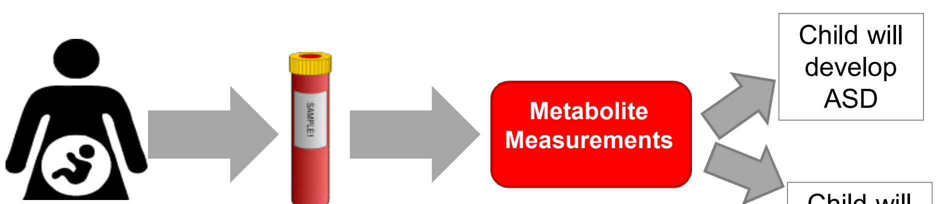
Prediction of ASD Before Symptoms Appear



- Data from 129 mothers were collected (107 had a child with ASD and 22 did not (not used yet))
- 20 FOCM/TS metabolites were measured during all three trimesters
- Follow up after 3 years to inquire about ASD diagnosis of child

RPI

Prediction of ASD Before Symptoms Appear



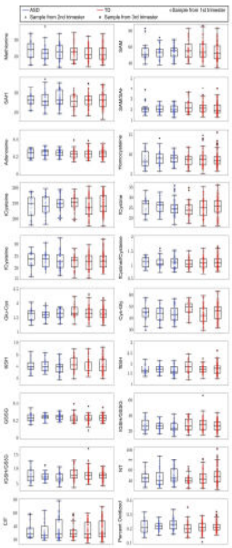
Groups	Time-points	Number of Mothers (separated by group)	Total Number of Mothers at each time point
High Risk Data	1st Trimester	**ASD:** 18 **TD:** 35	53
	2nd Trimester	**ASD:** 28 **TD:** 77	105
	3rd Trimester	**ASD:** 29 **TD:** 77	106
Control Data	1st Trimester	19	19
	2nd Trimester	22	22
	3rd Trimester	19	19

RPI

Univariate Analysis

- No significant difference can be found based upon outcome (ASD vs typically developing)

Metabolite	1st Trimester		2nd Trimester		3rd Trimester	
	Test	p-value	Test	p-value	Test	p-value
Methionine	t=	0.4358	t=	0.6386	MW	0.4719
SAM	MW	0.4930	MW	0.5818	MW	0.8096
SAH	t=	0.5816	t=	0.6727	t=	0.4222
SAM/SAH	MW	0.4139	MW	0.5475	MW	0.9915
Adenosine	t=	0.6067	t=	0.2123	t=	0.5608
Homocysteine	t=	0.6998	MW	0.4149	MW	0.3351
tCysteine	t≠	0.2515	t=	0.4941	MW	0.5305
fCystine	t=	0.1837	t=	0.3514	MW	0.4191
fCysteine	t=	0.5161	t=	0.2780	t=	0.9067
fCystine/fCysteine	t=	0.4748	MW	0.9192	MW	0.8399
Glu-Cys	MW	0.1481	t=	0.1578	t=	0.9812
Cys-Gly	t=	0.0645	t=	0.7281	t=	0.3198
tGSH	t=	0.3007	MW	0.8110	MW	0.2583
fGSH	t≠	0.0608	t=	0.8846	t=	0.0831
GSSG	MW	0.6253	MW	0.3650	t=	0.2305
tGSH/GSSG	t=	0.6238	MW	0.5621	t=	0.0911
fGSH/GSSG	t=	0.3852	MW	0.4642	t=	0.0327
NT	t=	0.1794	MW	0.7171	MW	0.8845
C↓T	MW	0.6795	MW	0.5793	MW	0.9379
Percent Oxidized	t=	0.3584	t=	0.4310	MW	0.0309

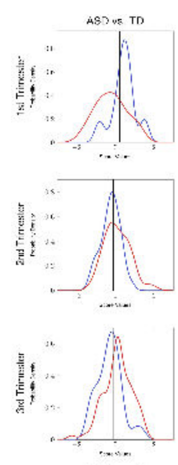


RPI

Multivariate Analysis

- Multivariate analysis also does not reveal differences

	ASD vs. TD	
	Errors	Metabolites
1st Trimester	Type I = 31.43% Type II = 38.89%	Homocysteine, fCysteine, Cys-Gly, fGSH, tGSH/GSSG
2nd Trimester	Type I = 41.56% Type II = 42.86%	SAH, Homocysteine, fCystine, Glu-Cys, tGSH
3rd Trimester	Type I = 38.96% Type II = 41.38%	fCystine, fCystine/fCysteine, Glu-Cys, Cys-Gly, Percent Oxidized



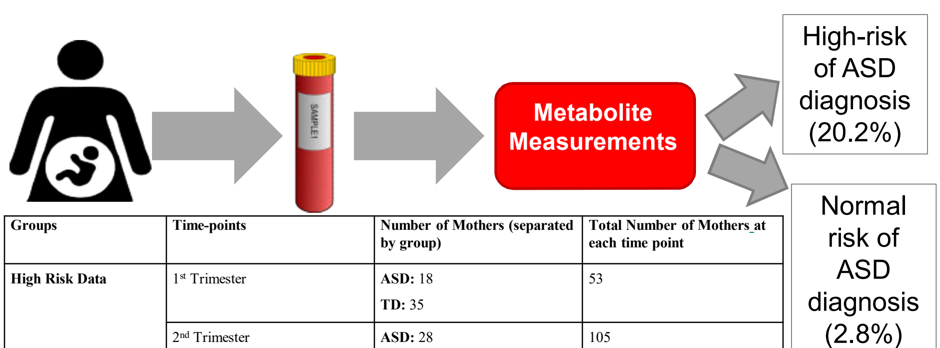
RPI

Summary I

- FOCM/TS metabolites collected from blood samples of a mother during pregnancy are not predictive for ASD diagnosis of a child in the future
  - Neither univariate nor multivariate analysis is able to return statistically significant differences between the ASD outcome at 3 years of age
- Next step: is there a difference between mothers in the “high-risk” group (20% of having a child with ASD) and the control group (3% as in the general population)?

RPI

Prediction of ASD Before Symptoms Appear



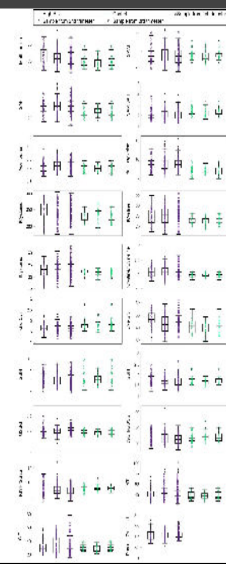
Groups	Time-points	Number of Mothers (separated by group)	Total Number of Mothers at each time point
High Risk Data	1st Trimester	**ASD:** 18 **TD:** 35	53
	2nd Trimester	**ASD:** 28 **TD:** 77	105
	3rd Trimester	**ASD:** 29 **TD:** 77	106
Control Data	1st Trimester	19	19
	2nd Trimester	22	22
	3rd Trimester	19	19

RPI

Univariate Analysis

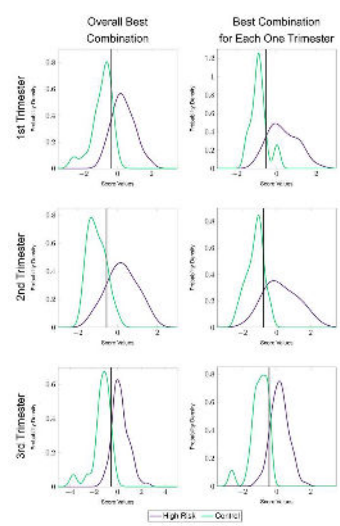
- Significant difference exist between High-risk and control group

Metabolite	1st Trimester		2nd Trimester		3rd Trimester	
	Test	p-value	Test	p-value	Test	p-value
Methionine	MW	0.0118	MW	0.0013	MW	0.0508
SAM	MW	0.7446	MW	0.4541	MW	0.5520
SAH	MW	0.0038	t≠	1.42E-04	MW	2.04E-04
SAM/SAH	MW	0.0280	MW	0.2502	MW	7.62E-04
Adenosine	t≠	0.0689	t≠	0.0104	t≠	0.0071
Homocysteine	t=	1.29E-04	MW	2.62E-06	MW	3.75E-06
tCysteine	MW	0.0040	MW	0.5324	t≠	6.14E-04
fCystine	t≠	0.0071	t≠	5.64E-06	t≠	4.44E-05
fCysteine	t≠	0.3609	t≠	0.0591	t≠	0.0875
fCystine/fCysteine	t≠	1.26E-04	t≠	1.67E-06	MW	0.0042
Glu-Cys	MW	0.0142	MW	0.0054	MW	4.59E-04
Cys-Gly	t=	9.75E-04	t≠	0.0088	t=	0.0446
tGSH	t=	0.9708	MW	0.8136	MW	0.6749
fGSH	MW	0.8983	t=	0.5139	MW	0.3750
GSSG	t≠	0.3396	MW	0.0022	t=	0.0421
tGSH/GSSG	MW	0.7205	MW	0.0782	t=	0.0703
fGSH/GSSG	t≠	0.6299	MW	0.0243	t≠	0.0057
NT	MW	0.2663	MW	0.0412	MW	0.0036
CIT	MW	0.2832	t≠*	0.0320	t≠*	0.0279



Multivariate Analysis

	Overall Best Combination		Best Combination for Each One Trimester	
	Errors	Metabolites	Errors	Metabolites
1st Trimester	Type I = 14.29% Type II = 7.55%	Homocysteine, fCystine/fCysteine, Glu-Cys, fGSH, NT	Type I = 7.14% Type II = 11.32%	Methionine, fCystine/fCysteine, Cys-Gly, tGSH/GSSG, NT
2nd Trimester	Type I = 13.64% Type II = 17.14%		Type I = 9.09% Type II = 13.33%	Methionine, Homocysteine, Glu-Cys, GSSG, NT
3rd Trimester	Type I = 10.53% Type II = 9.43%		Type I = 10.53% Type II = 6.60%	Methionine, SAM, Homocysteine, Glu-Cys, NT



Summary II

- FOCM/TS metabolites collected from blood samples of a mother during pregnancy are not predictive for ASD diagnosis of a child in the future
  - Neither univariate nor multivariate analysis is able to return statistically significant differences between the ASD outcome at 3 years of age

However

- FOCM/TS metabolites collected from blood samples of a mother during pregnancy are reasonably predictive of the likelihood (20% vs 3%) that the child will be diagnosed with ASD at age 3

RPI

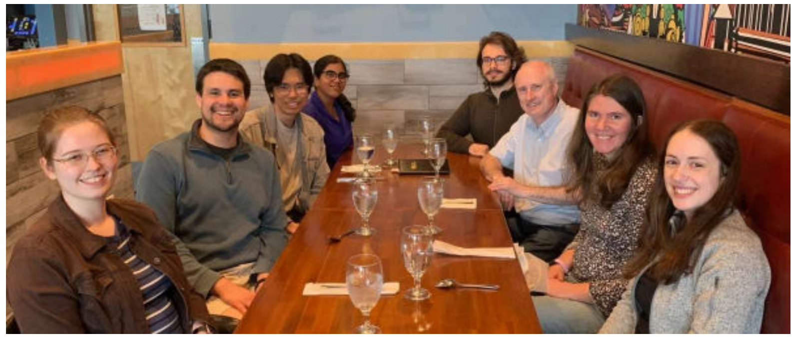
Overall Summary

- Metabolites of the FOCM/TS pathways show strong correlations with ASD in our studies
  - They can be used to support an ASD diagnosis in children
  - They are reasonably predictive of adaptive behavior
  - Metabolites from a mother during pregnancy can not be used to determine if a child will develop ASD, however, they are predictive of the likelihood of a mother of having a child with ASD
- One key aspect of this work is the systems approach as some of the study data had already been analyzed and published, however, results were inconclusive due to univariate statistics whereas more advanced analysis techniques allow for accurate predictions

RPI

Acknowledgments

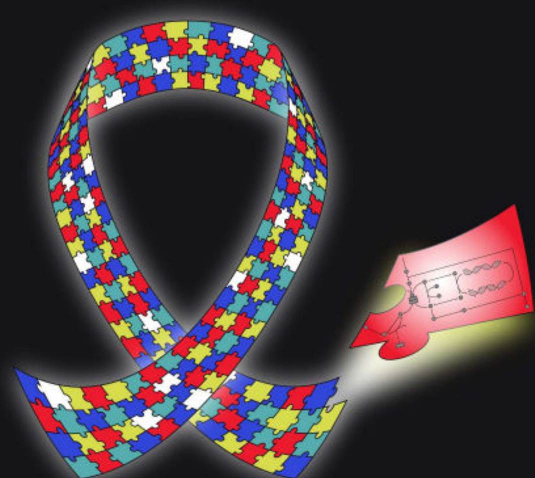
- Research Group
  - Halil Arici
  - Emma Dando
  - Jennifer Ferina
  - Hongbin Liu
  - Diego Machado



- Collaborators
  - Wayne Bequette, RPI
  - S. Jill James, UAMS
  - Richard Frye, PCH
  - James Adams, ASU
  - Arul Jayaraman, Texas A&M
- Funding Agencies
  - NIH, NSF, CDC, ACS-PRF, DOE, Rockwell, PSTC, N of One Foundation, TCFD, BRAIN Foundation, O'Sullivan Foundation, Bia-Echo Foundation

RPI

Think Beyond the Brain



RPI

References

1. Braun et al. *PLoS ONE* 2015. doi: 10.1371/journal.pone.0124120
2. CDC. “Prevalence of Autism Spectrum Disorders — Autism and Developmental Disabilities Monitoring Network, 14 Sites, United States, 2008” 2016. http://www.cdc.gov/mmwr/preview/mmwrhtml/ss6103a1.htm
3. Leigh and Du. *J Autism Dev Disord*. 2015. doi: 10.1007/s10803-015-2521-7
4. Buescher et al. *JAMA Pediatrics* 2014. doi: 10.1001/jamapediatrics.2014.210
5. Greenberg et al. *Journal of Clinical Psychiatry* 2015. doi: 10.4088/JCP.14m09298
6. Doshi et al. *Journal of the American Academy of Child & Adolescent Psychiatry* 2012. doi: 10.1016/j.jaac.2012.07.008
7. CDC. “Alzheimer’s Disease: Promoting Health and Independence for an Aging Population At A Glance” 2016. https://www.cdc.gov/chronicdisease/resources/publications/aag/alzheimers.htm
8. ADA. Diabetes Care 2013. doi: 10.2337/dc12-2625
9. Gaugler et al. *Nature Genetics* 2015. doi: 10.1038/ng.3039
10. Hallmeyer et al. *Arch Gen Psychiatry* 2011. doi: 10.1001/archgenpsychiatry.2011.76
11. Lichtenstein et al. *American Journal of Psychiatry* 2010. doi: 10.1176/appi.ajp.2010.10020223
12. Howsmon et al. *PLoS Computational Biology* 2017. doi: 10.1371/journal.pcbi.1005385
13. Melnyk et al. *Journal of Autism and Developmental Disorders* 2012. doi: 10.1007/s10803-011-1260-7
14. Frye et al. *Molecular Psychiatry* 2016. doi: 10.1038/mp.2016.168
15. Vargason et al. *Journal of Theoretical Biology* 2017. doi: 10.1016/j.jtbi.2016.12.021
16. Duncan et al. *Nutrients* 2013, 5(7): 2457-74.
17. Stabler et al. *Clin Chem* 2004, 50(2): 365-72.
18. Chwatko at al. *Clin Chem* 2005, 51(2): 408-15.
19. Refsum et al. Clin Chem 1985, 31(4): 624-8.
20. Hong et al. *J Korean Med Sci* 2005, 20(5): 721-6.

RPI

Questions?

RPI