Deep Multitask Learning of Gene Risk for Comorbid Neurodevelopmental Disorders

A. Ercument Cicek

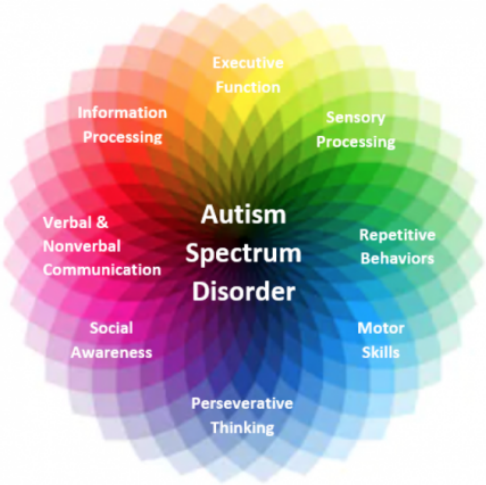
Assistant Professor, Computer Engineering Dept., Bilkent University

Adjunct Faculty, Computational Biology Dept., Carnegie Mellon University

CICEKLAB

Neurodevelopmental Disorders

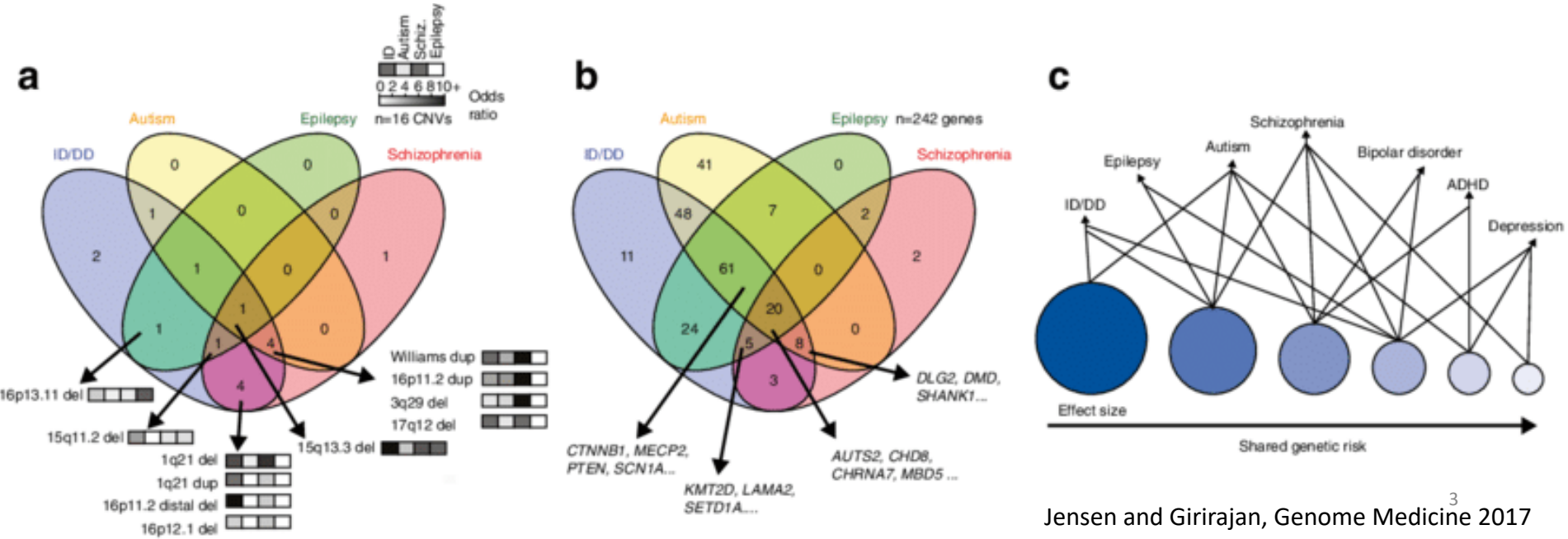
- Delay/disturbance in the skills : Social, Motor, Language, Cognition
- Heterogeneous phenotype
- Common: ASD 1 in every 54 children in the US
- Examples
  - Autism spectrum disorder (ASD)
  - Intellectual disability (ID)
  - Global developmental delay (GDD)
  - Attention Deficit Hyperactivity Disorder (ADHD)
  - Social Communication Disorder



https://carmenbpingree.com/blog/what-is-autism-spectrum-disorder/

Neurodevelopmental Disorders – cont'd

- Highly comorbid



Jensen and Girirajan, Genome Medicine 2017

Autism & Intellectual Disability As Comorbid Disorders

Autism Spectrum Disorder (ASD)

- restrictive and repetitive behaviors, dysfunctional reciprocal social behavior, and impaired communication abilities
- Seen in 2% school age children[1]

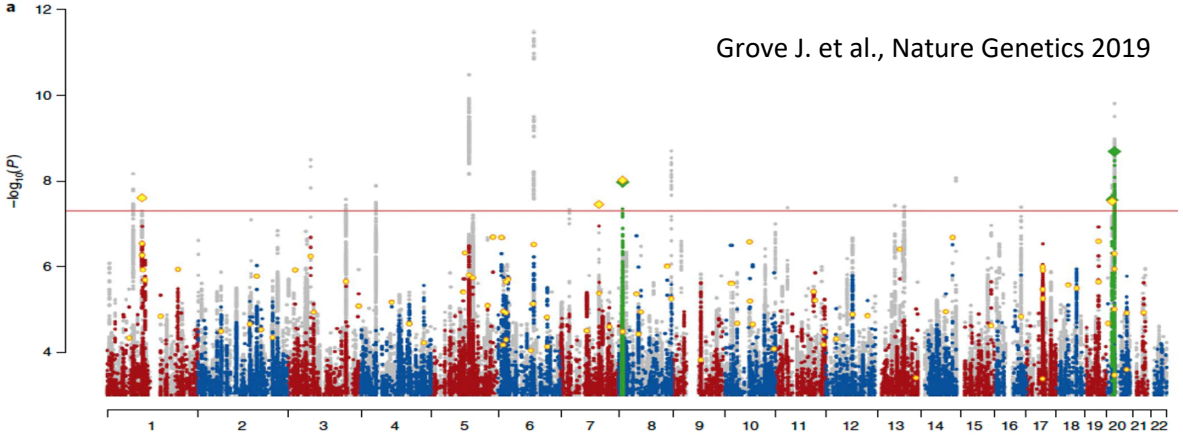
Intellectual Disability (ID)

- Characterized by below average intellectual functioning (IQ < 70) with significant limitations in adaptive functioning
- Affecting 1-3% of population[1]

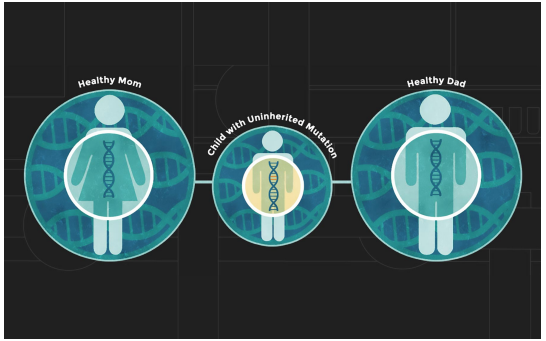
- 70% of children with ASD has also ID.
- Both conditions are heterogeneous.
- Both are associated with CNVs and single gene mutations.
- Evidence supports oligenic mode of inheritance for both.

[1]Srivastava, A. K., & Schwartz, C. E. (2014). *Neuroscience and biobehavioral reviews*, *46 Pt 2*, 161–174. https://doi.org/10.1016/j.neubiorev.2014.02.015

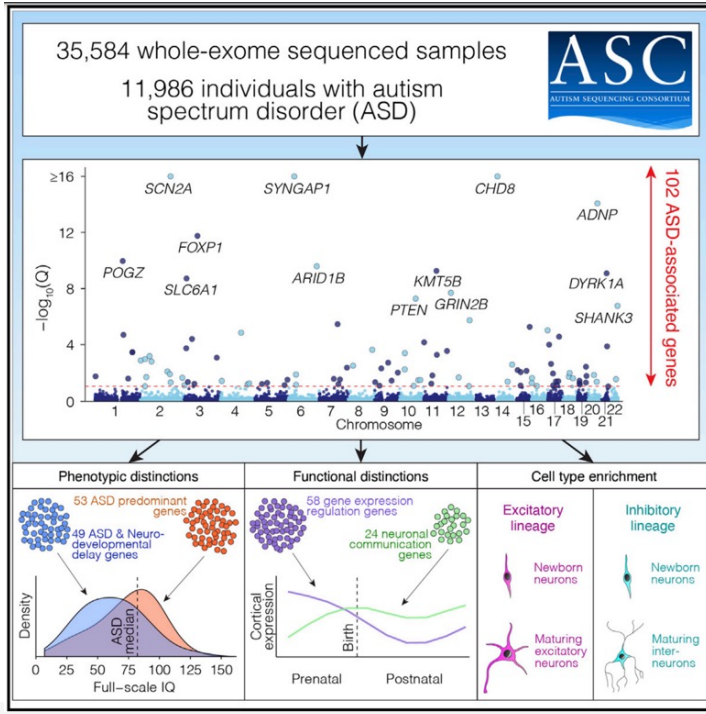
Risk Gene Discovery



Grove J. et al., Nature Genetics 2019



https://www.cshl.edu/autism-genetics-study-calls-attention-impaired-motor-skills-general-cognitive-impairment/



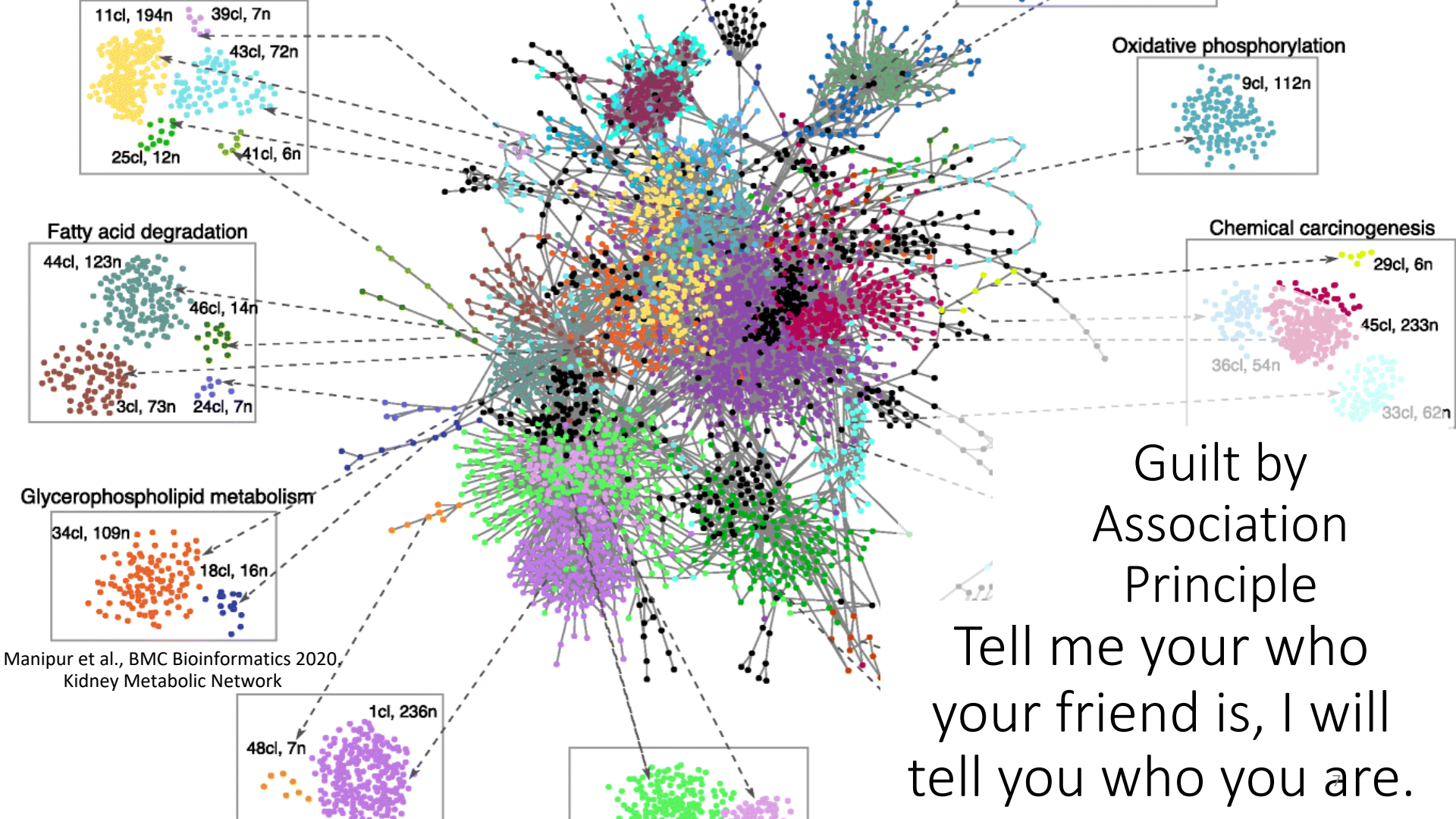
Highlights

- 102 genes implicated in risk for autism spectrum disorder (ASD genes, FDR ≤ 0.1)

Satterstrom et al., CELL 2020.

Risk Gene Discovery – cont’d

- For Autism, it is a large puzzle with
  - ~100 pieces known,
  - ~900 remaining,
  - ~20,000 possible pieces to choose from.
- Genes/Proteins are interacting in biochemical networks.
- Can we use the guilt by association principle to pinpoint connecting pieces?

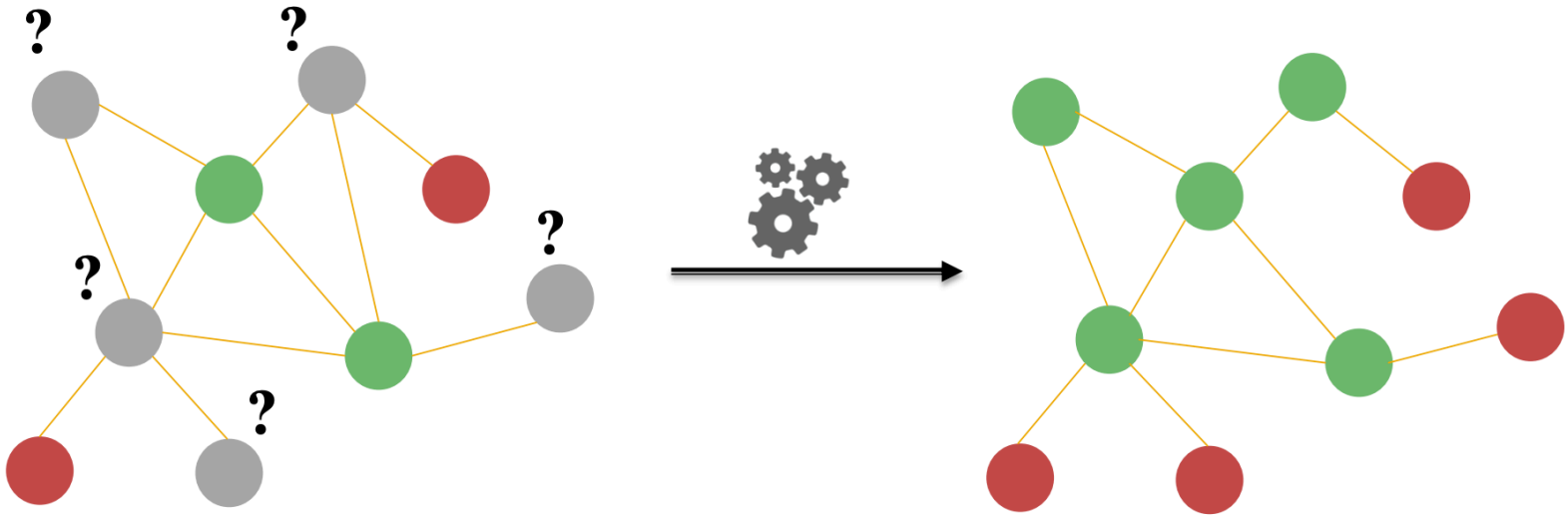


Manipur et al., BMC Bioinformatics 2020,
Kidney Metabolic Network

Guilt by Association Principle

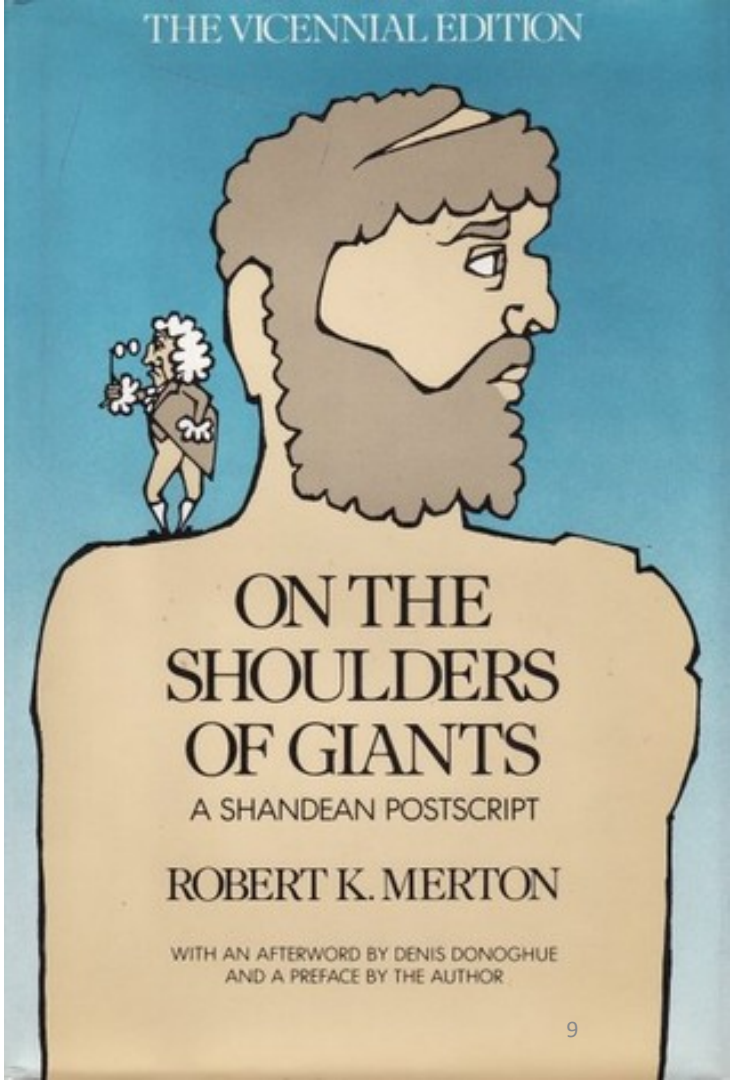
Tell me your who your friend is, I will tell you who you are.

Node Classification
A Semi-supervised Learning Problem



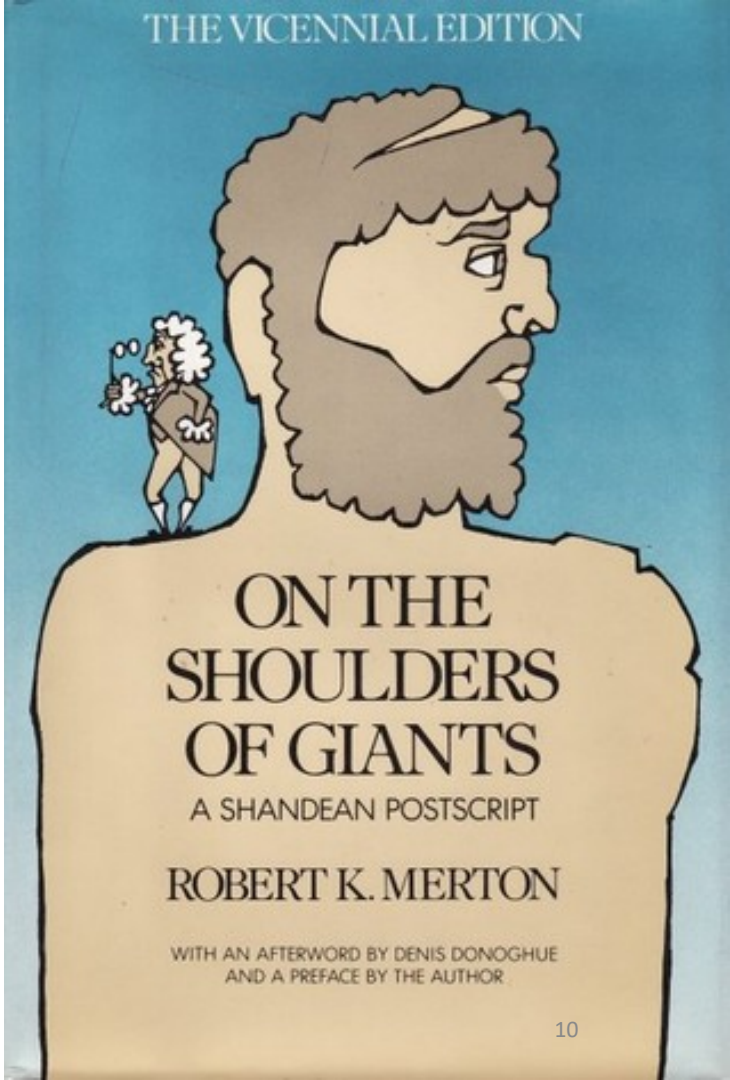
Risk Gene Discovery Algorithms

- NETBAG (Gilman et al., Neuron 2014)
- DAWN (Liu et al., Mol. Autism 2015)
- Evidence Weighted SVM (Krishnan et al., Nature Neuroscience 2016)
- Random Forest (Duda et al., Translational Psychiatry 2018)
- ST-Steiner (Norman and Cicek, Bioinformatics 2019)
- ForecASD (Brueggeman et al., Scientific Reports 2020)
- DeepND (Beyreli et al., bioRxiv 2021)



Risk Gene Discovery Algorithms

- NETBAG (Gilman et al., Neuron 2014)
- DAWN (Liu et al., Mol. Autism 2015)
- Evidence Weighted SVM (Krishnan et al., Nature Neuroscience 2016)
- Random Forest (Duda et al., Translational Psychiatry 2018)
- ST-Steiner (Norman and Cicek, Bioinformatics 2019)
- ForecASD (Brueggeman et al., Scientific Reports 2020)
- DeepND (Beyreli et al., bioRxiv 2021)



NETBAG

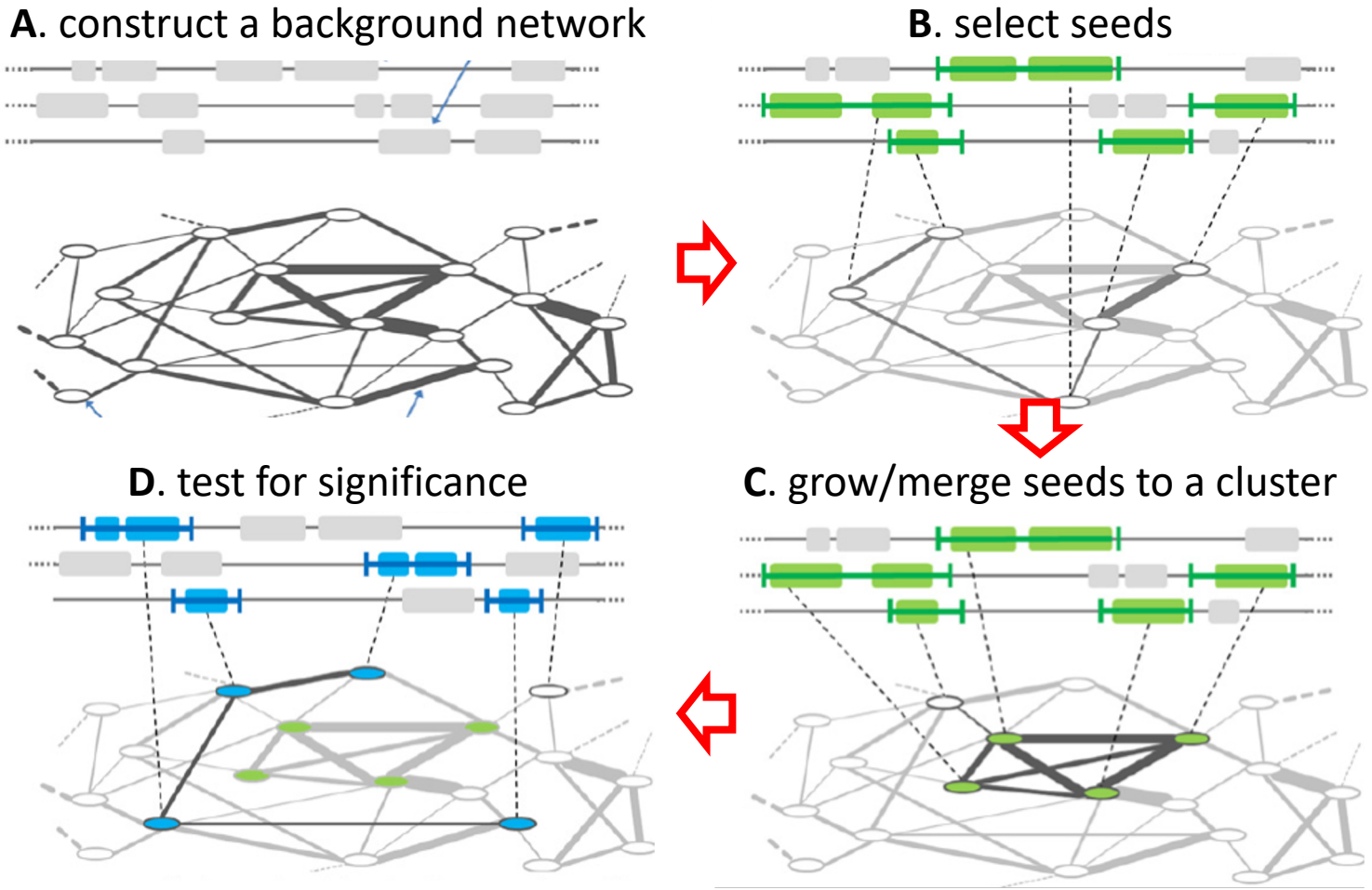
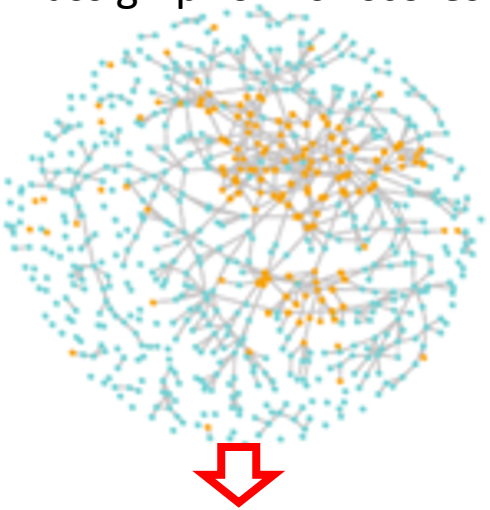


Figure from Gilman et al. 2011, “Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses”. *Neuron*.

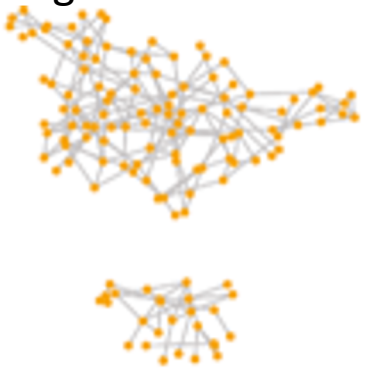
A. construct a coexpression network

DAWN

B. assign prior risk scores



C. high-scoring clusters of network genes



D. find risk genes among network genes

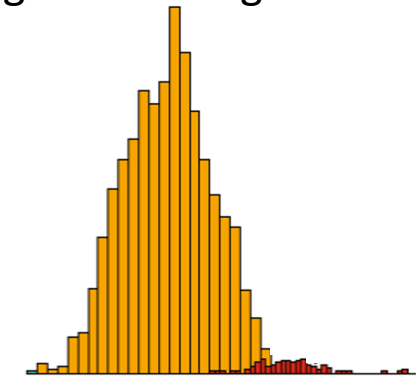
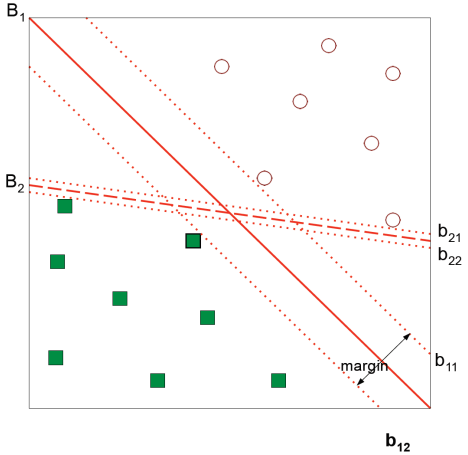


Figure from Liu et al. 2014, “DAWN: a framework to identify autism genes and subnetworks using gene expression and genetics”. *Molecular Autism*.

Evidence Based SVM



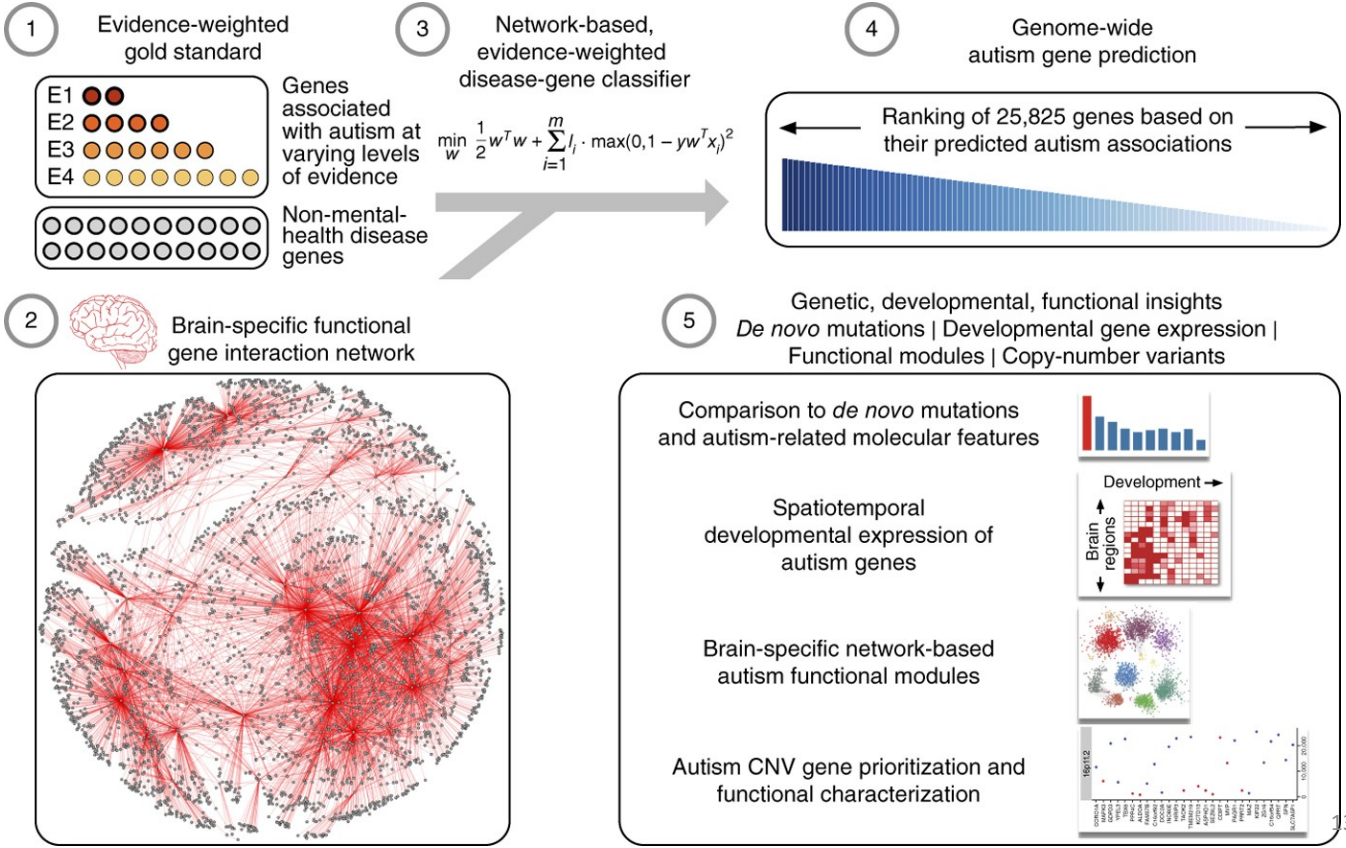


Figure from Krishnan et al. 2016, "Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder". *Nature Neuroscience*.

ST-Steiner

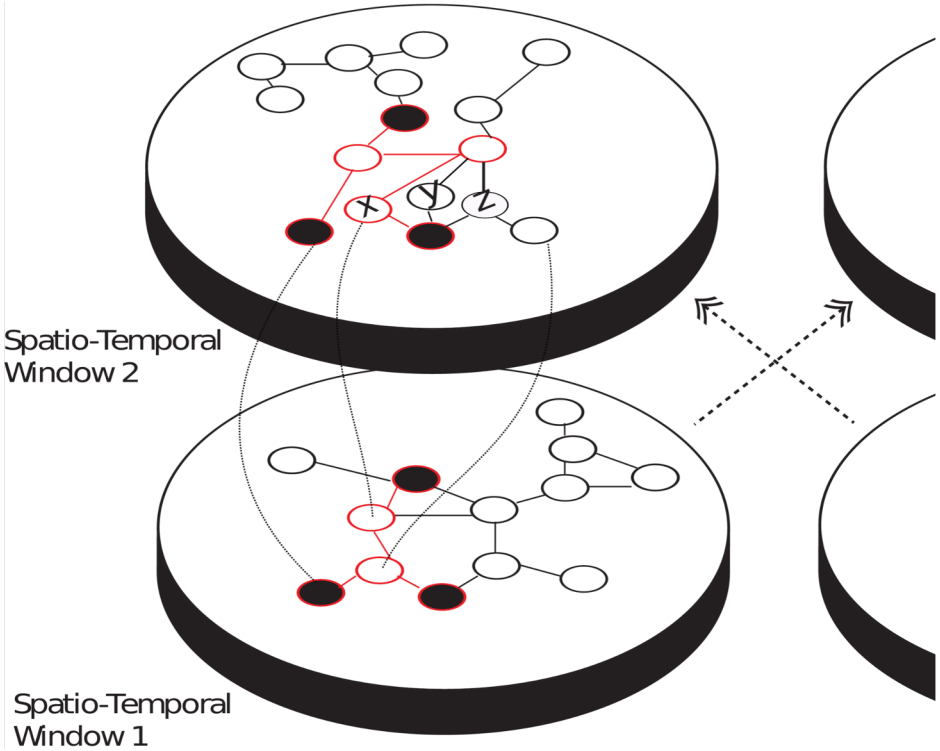
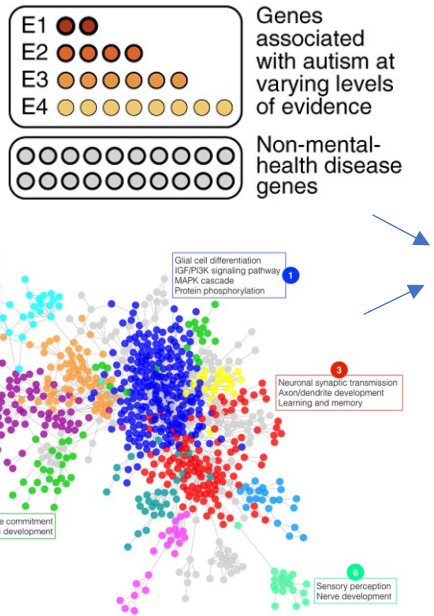


Figure from Norman and Cicek, 2019, “ST-Steiner: a spatio-temporal gene discovery algorithm”. Bioinformatics

Random Forest



Cross-species brain-specific functional relationship network

Random Forest

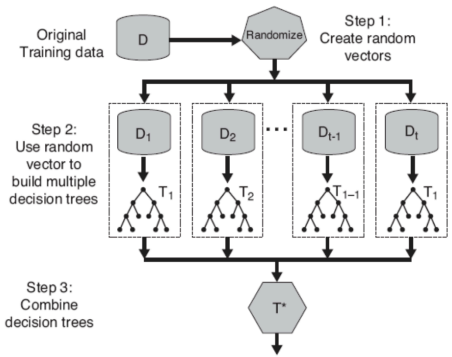
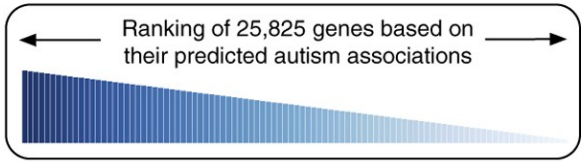


Figure 5.40. Random forests.

http://people.csail.mit.edu/dsontag/courses/ml13/slides/lecture13.pdf



Duda et al. 2018, “Brain-specific functional relationship networks inform autism spectrum disorder gene prediction”. *Translational Psychiatry*

forecASD

Figure from Brueggeman et al., 2020, “Forecasting risk gene discovery in autism with machine learning and genome-scale data”. Scientific Reports

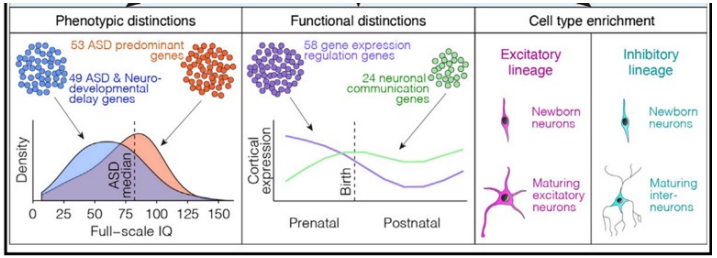
Shortcomings of the Literature

By design they are limited to work with a single disorder, shared genetic information is ignored.

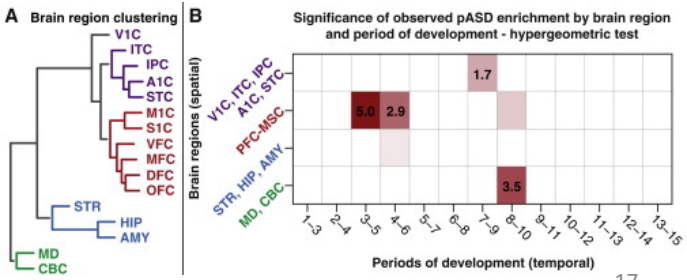
Bag mutational burden as if they are the same.	➡	Disorder specific features are lost.
Perform independent analysis per disorder and intersect results.	➡	Information coming from the shared genetic architecture is ignored.

Network-based gene discovery methods can work with at most a handful of integrated gene interaction networks.

Functional interaction networks (e.g., co-expression, protein interaction etc.) are disregarded.	➡	Cannot distinguish where the signal is coming from, often a separate downstream analysis



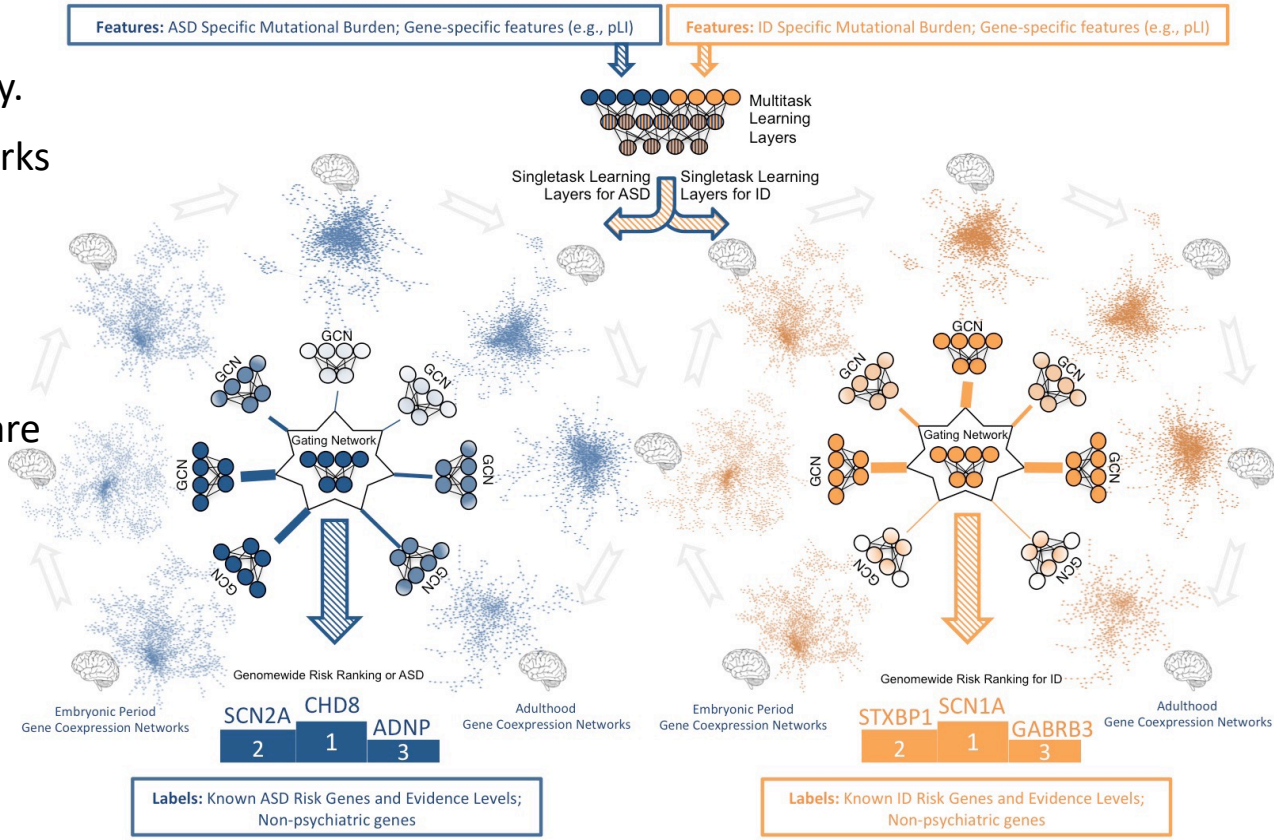
Satterstrom et al., CELL 2020.



Willsey et al., CELL 2013.

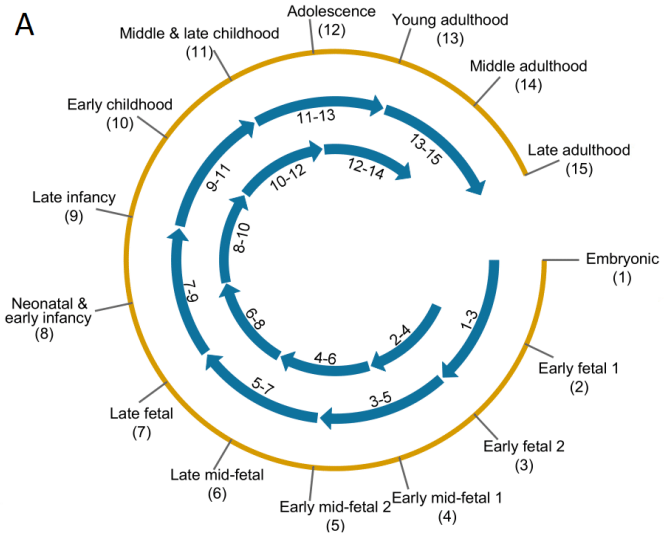
DeepND – Deep Neurodevelopmental Disorders

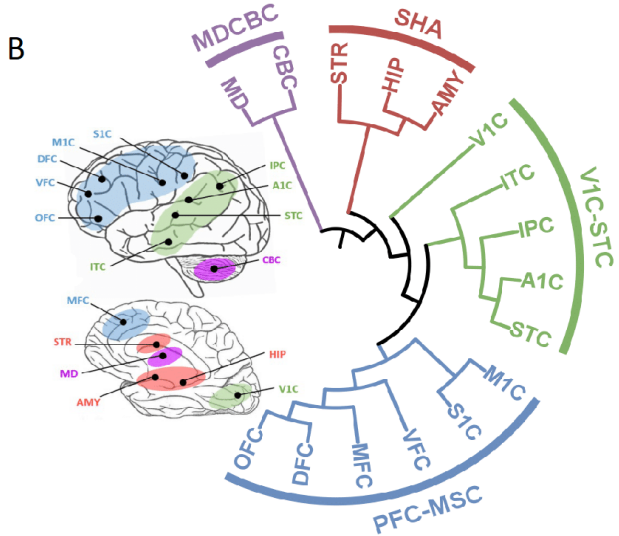
- Multi-task learning to analyze comorbid disorders simultaneously.
- Graph convolutional neural networks (GCNs).
- Analyze multiple gene interaction networks.
- Mixture-of-experts model learns which gene interaction networks are informative.

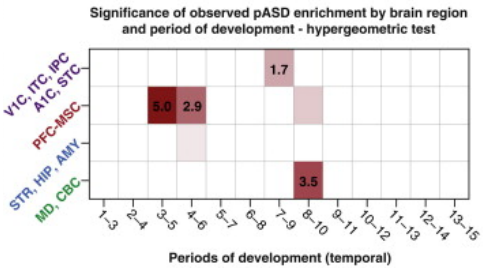


Gene Co-expression Networks & Features

- BrainSpan dataset of Allen Brain Atlas contains gene expression levels in samples from 16 regions of 57 postmortem brains.
- We constructed 52 spatio-temporal networks by partitioning the dataset into developmental periods and clusters of brain regions as also done by Willsey et al.[3].







Willsey, A. J., ... & Murtha, M. T. (2013). Coexpression networks implicate human midfetal deep cortical projection neurons in the pathogenesis of autism. *Cell*, *155*(5), 997-1007.

Features & Labels

- The only feature we use is pLI of the gene.
- Labels for ASD: SFARI gene scoring Cat I – III as positive ground truth genes and Krishnan et al.'s non-mental health genes as negative ground truth.
- Labels for ID: Investigating 5 review papers for positive labels, same negative set.

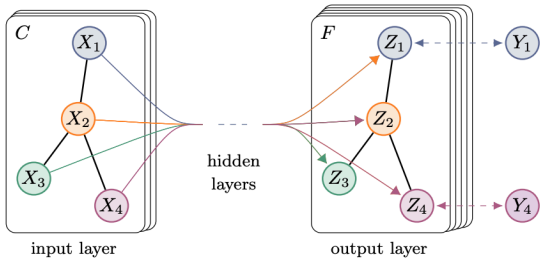
[1]Satterstrom, F. K., ... & Stevens, C. (2020). Large-scale exome sequencing study implicates both developmental and functional changes in the neurobiology of autism. *Cell*, *180*(3), 568-584.
[2]Nguyen, H. T., ... & Pinto, D. (2017). Integrated Bayesian analysis of rare exonic variants to identify risk genes for schizophrenia and neurodevelopmental disorders. *Genome medicine*, *9*(1), 114.

Graph Convolutional Neural Networks

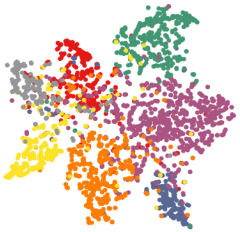
- Graph convolutional networks are used on arbitrarily structured data to extract patterns.
- Kipf and Welling proposed an efficient propagation rule which uses a localized first-order approximation of spectral graph convolutions.
- Each subsequent layer k of a GCN module used in DeepASD is defined as

$$H_k[i] = \sigma(\hat{D}^{-0.5}\hat{E}\hat{D}^{-0.5}H_{k-1}[i]W_{k-1})$$

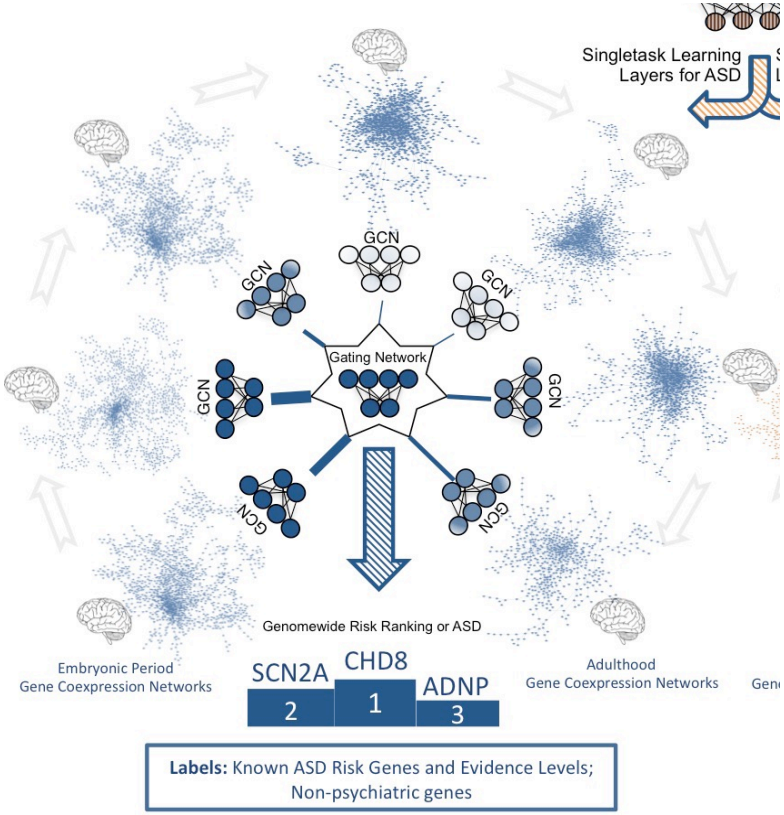
Kipf and Welling, https://arxiv.org/abs/1609.02907



(a) Graph Convolutional Network

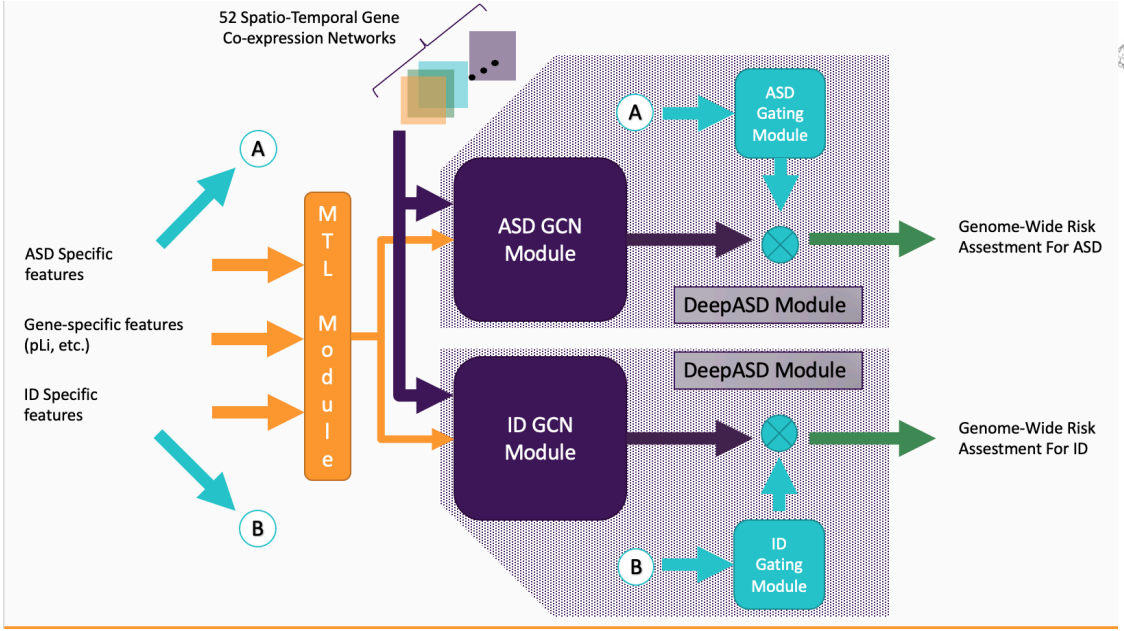


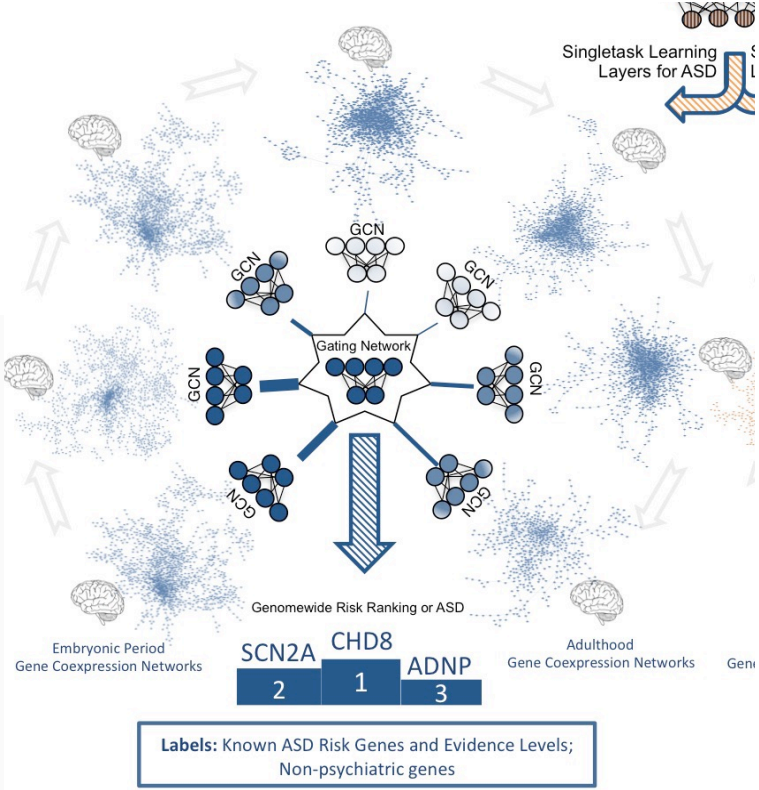
(b) Hidden layer activations



Mixture of Experts

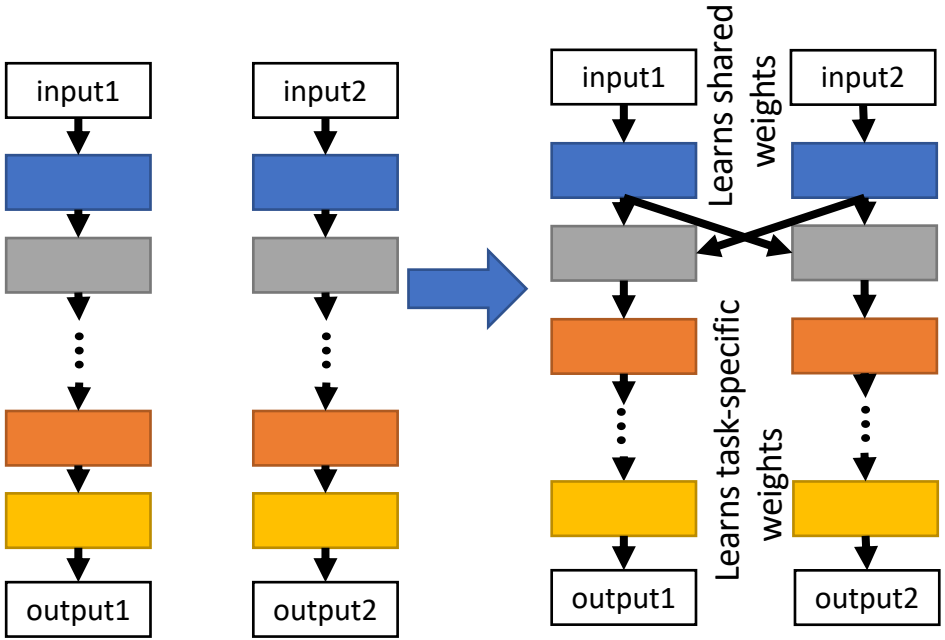
- Learn which GCNs are more informative (i.e., are better at predicting risk genes).
- Use raw input features to weigh each GCN which corresponds to a neurodevelopmental window.

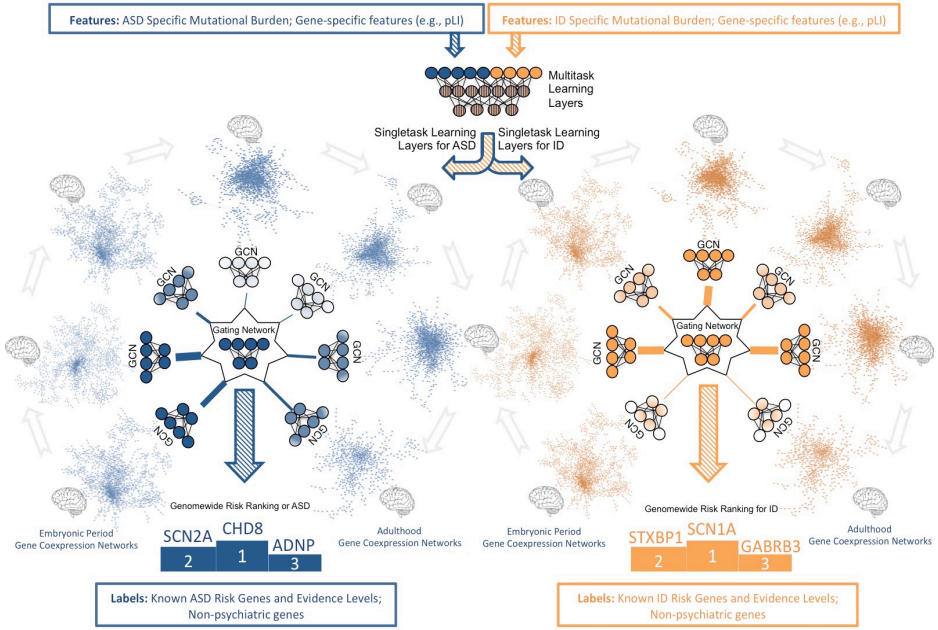




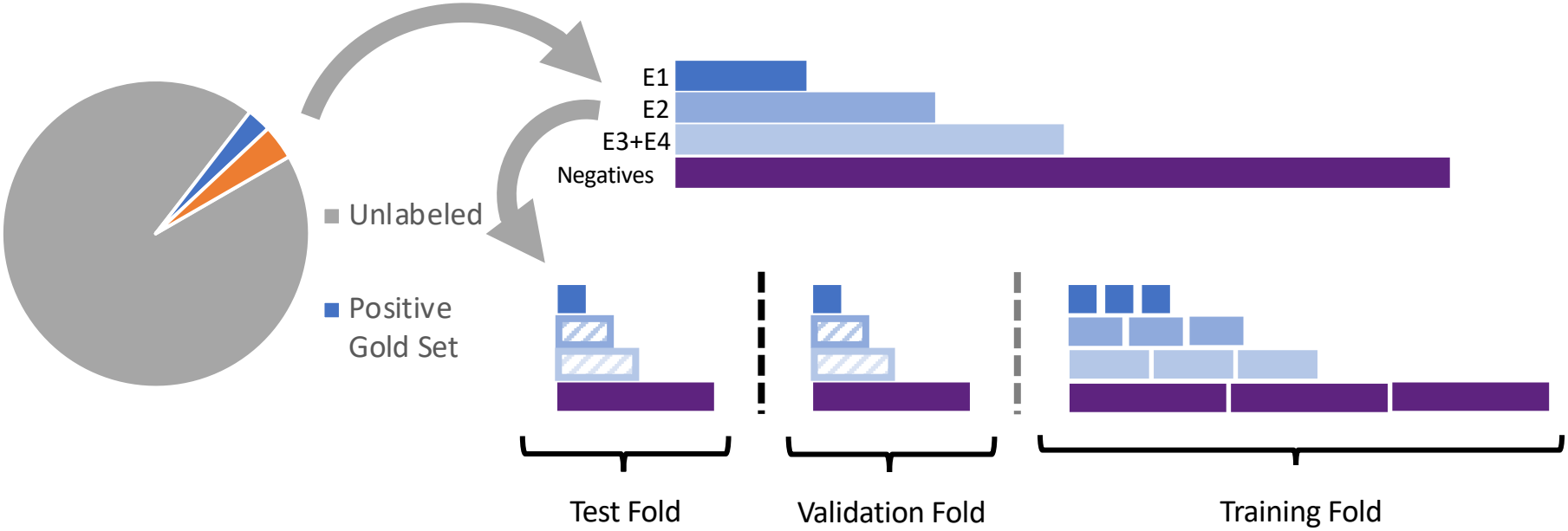
Multitask Learning

- In Multitask Learning (MTL), there are a set of general learning task, all or at least a subset of whom are assumed to be related to each other.
- Feature transformation approach is one of the MTL methods where the feed-forward network is trained to learn a common feature representation.



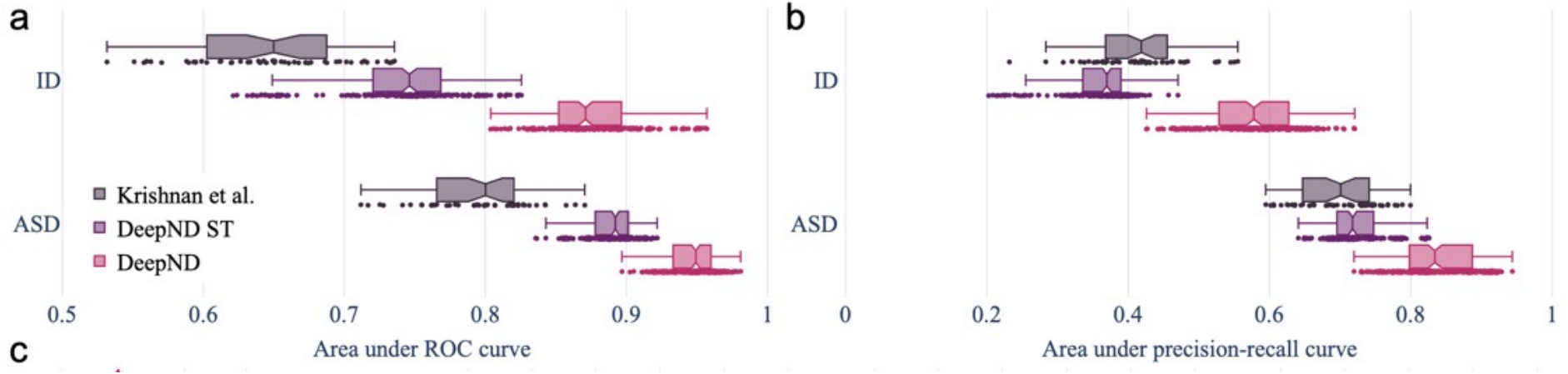


Experimental Setup: 3-1-1 Cross Validation



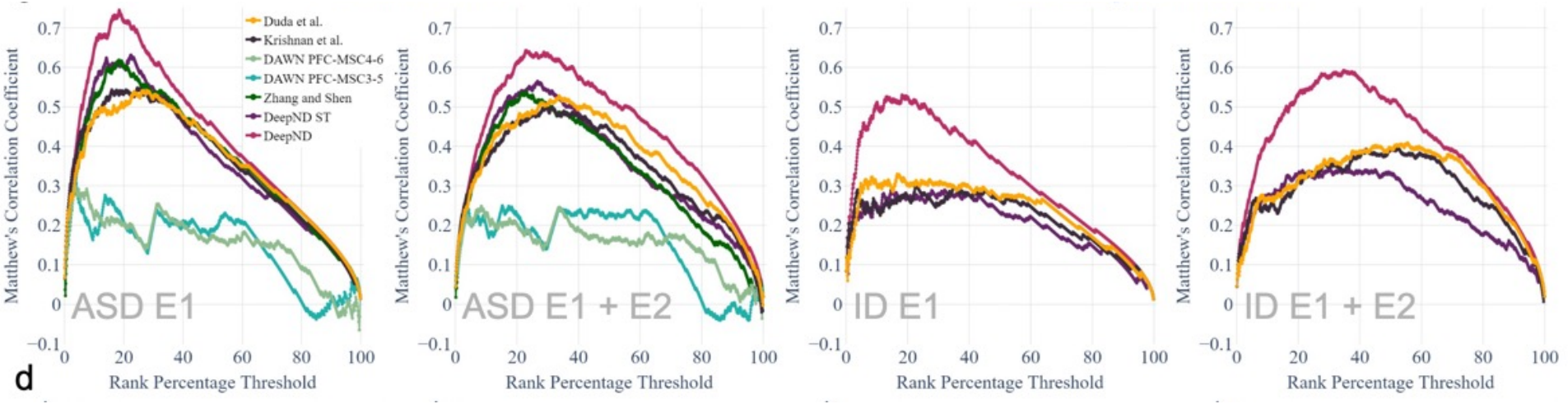
Results – Performance Comparison

AUROC & AUPR distributions



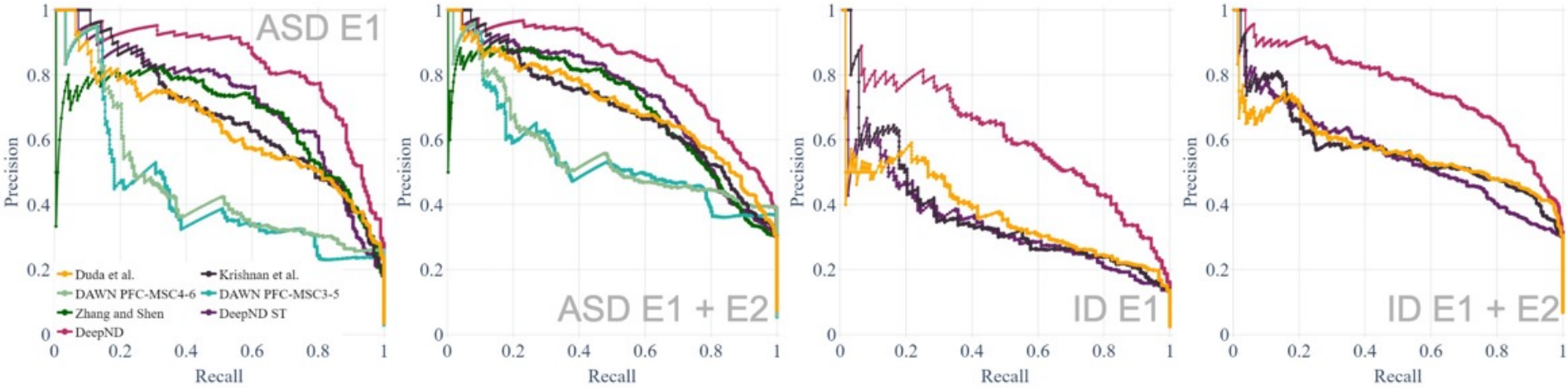
Results – Performance Comparison - cont'd

Matthew's Correlation Coefficient with respect to varying rank percentage thresholds

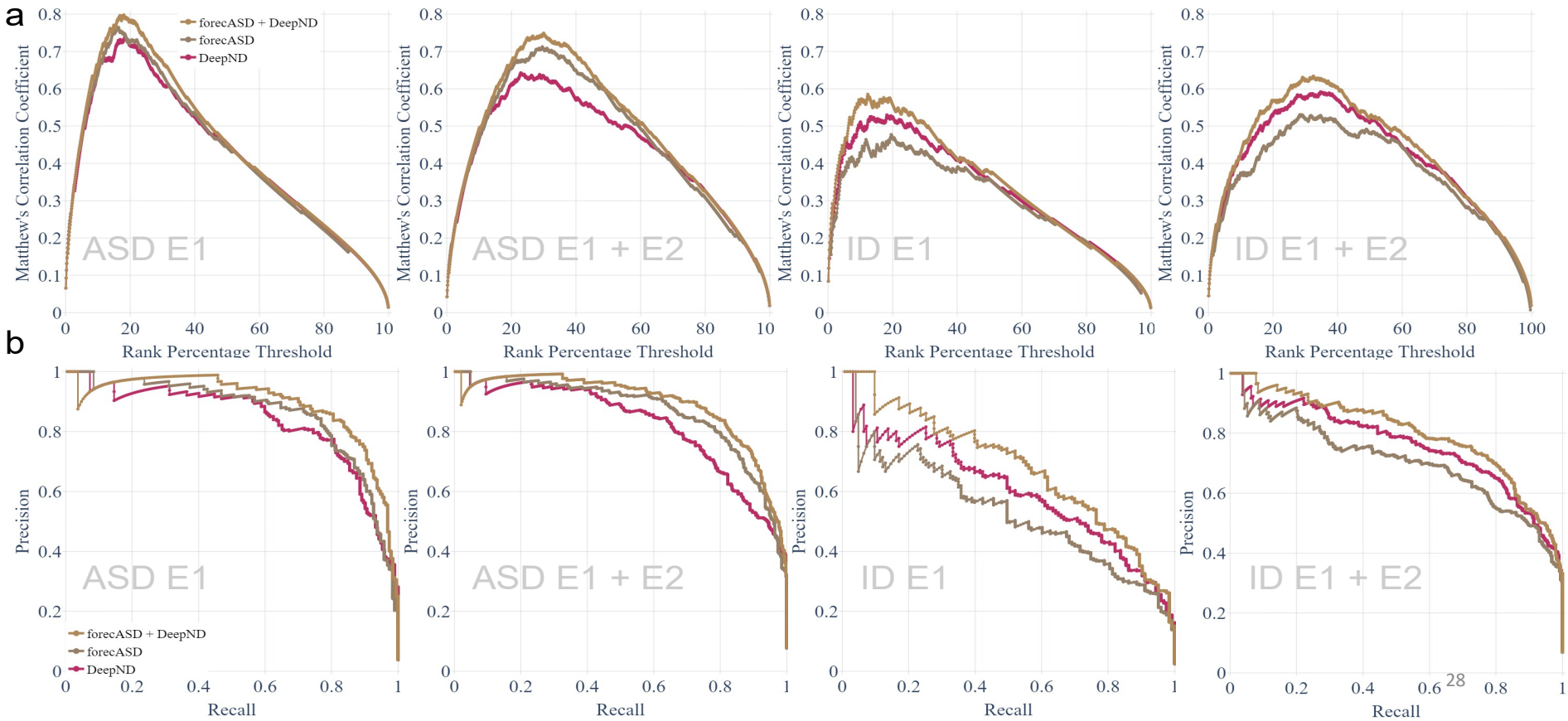


Results – Performance Comparison - cont’d

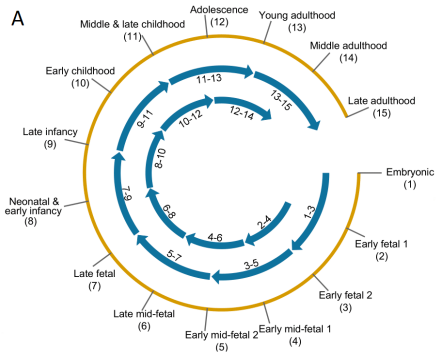
Precision- Recall Curve Comparisons over Final Rankings

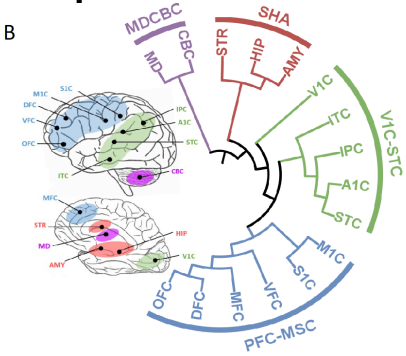


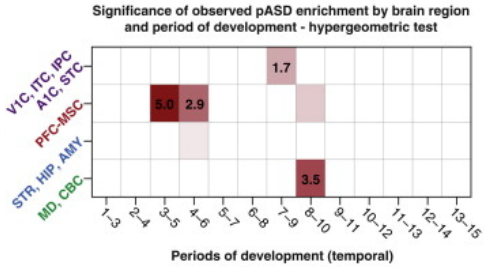
Comparison with forecASD

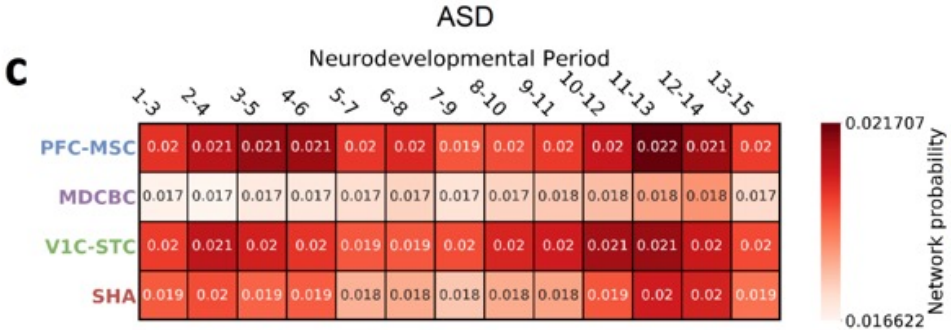


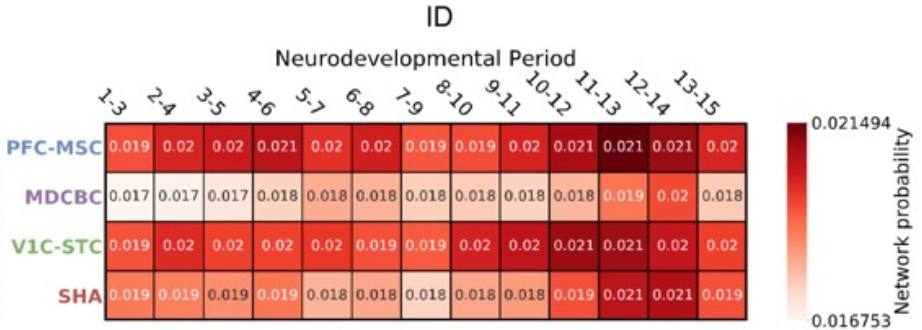
Informative Neurodevelopmental Windows



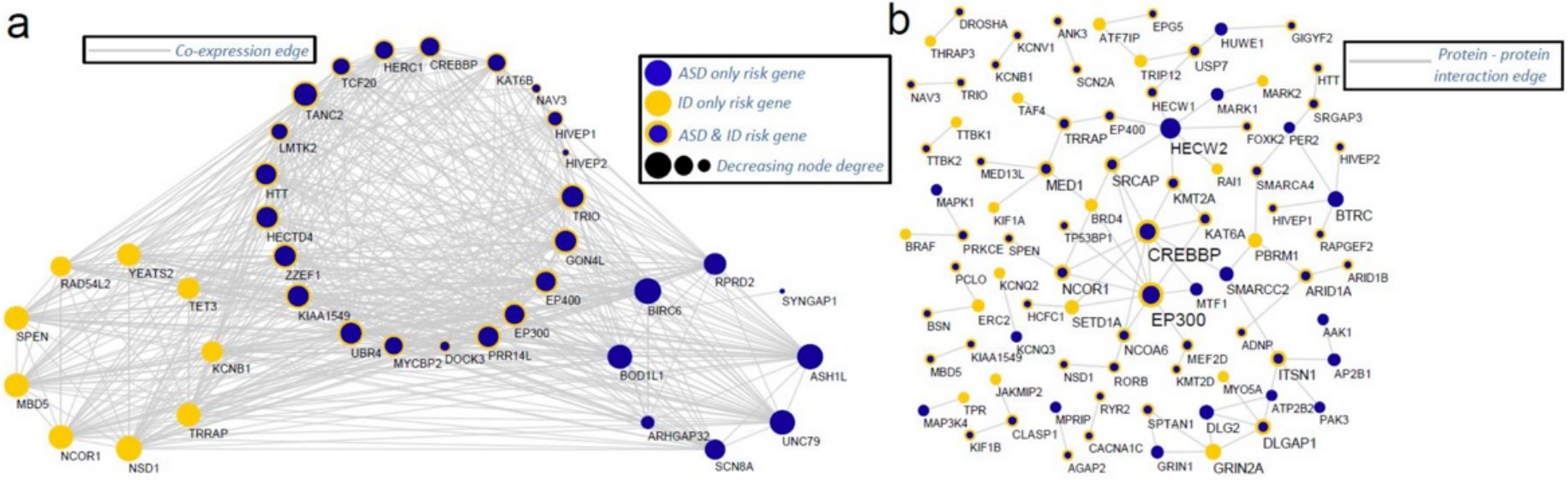








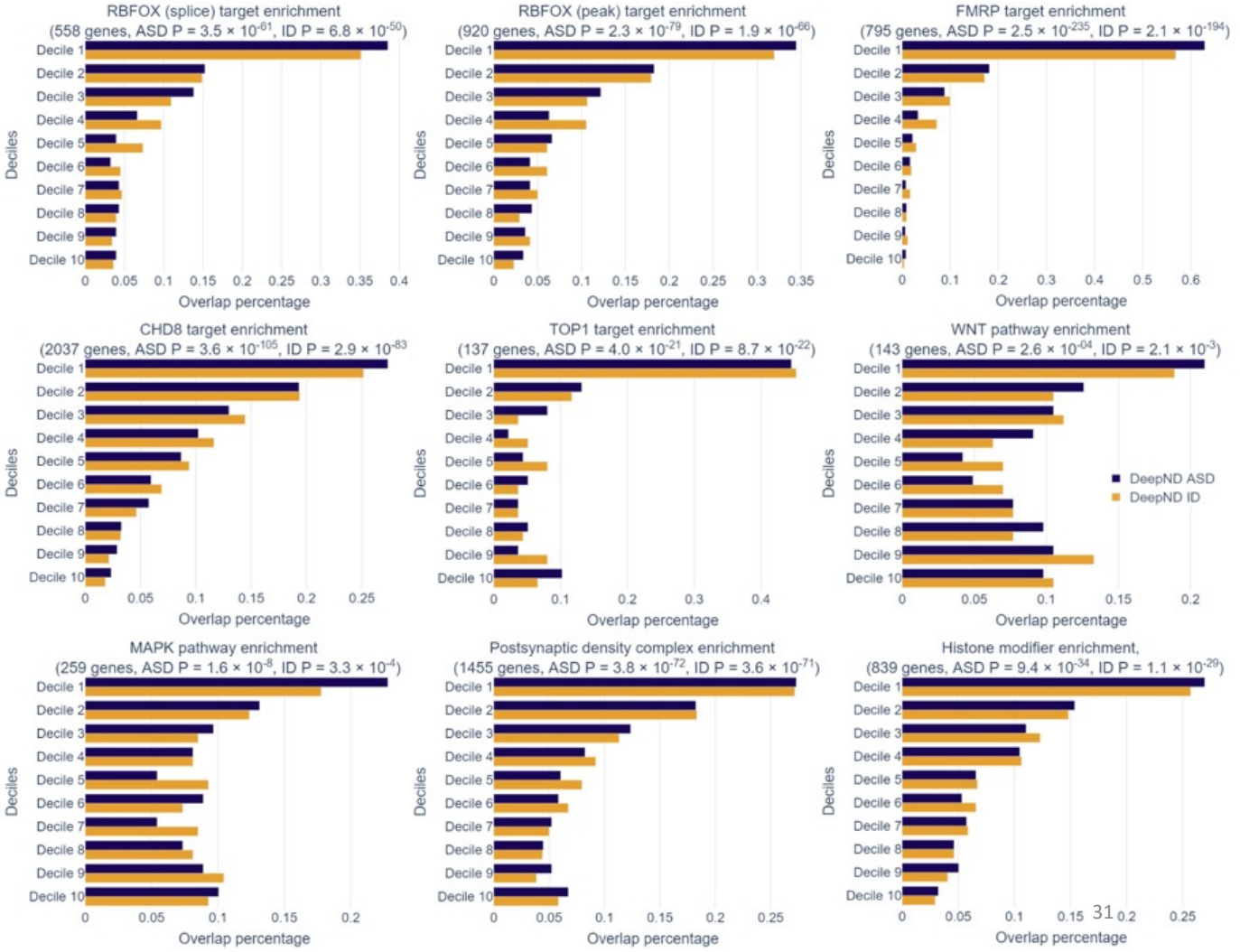
Network Analyses



PFC-MSC 4-6 connections of risk genes
75x more connected compared to MD-CBC 2-4

Brain specific PPI network:
HECW2 has very low prior signal yet is a hub

Enrichment Analyses



Novel Predictions in shared CNV regions

- *NIPA2*
  - is an ASD E3-E4 gene which encodes a magnesium transporter.
  - 5th decile for ASD, top decile for ID.
  - Its linkage to Prader-Willi Syndrome[1] which also suggests that NIPA2 might is an important candidate for ID rather than ASD.
- *MICAL3*
  - related to actin and Rab GTPase binding and cytoskeletal organization.
  - top decile for both ASD and ID.
  - low prior, no other gene discovery algorithm points to it.
  - ST-Steiner[2] and Satterstrom et al.[3] have pointed to the importance of cytoskeletal organization function in ASD.

1. Goytain et al., Journal of Physiology-Cell Physiology 2008.
2. Norman and Cicek, Bioinformatics 2019.
3. Satterstrom et al., CELL 2020.

Novel Predictions - cont'd

- *ZBTB20*
  - Transcriptional repressor, important for postnatal growth.
  - is an ASD E1 gene and CHD8 target.
  - 500th gene in Evidence-based SVM ranking.
  - Last decile DeepND for ASD but higher chance for ID.
  - Shown to be related to Primrose Syndrome[1,2] which is specifically characterized by intellectual disability.

1. Cordeddu et al., Nature genetics 2014.
2. Cleaver et al., American Journal of Medical Genetics Part A 2019.

Novel Predictions - cont'd

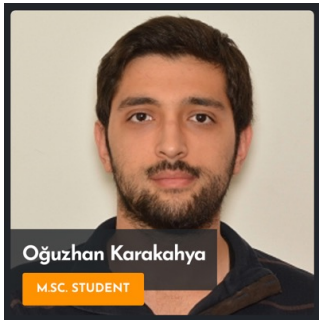
- *LMTK2*
  - nerve growth factor (NGF)-TrkA signaling and plays a role in spermatogenesis.
  - ranked 2nd for ASD and 7th for ID by DeepND.
  - Not in top 1000 for other algorithms.
  - Target of CHD9 and FMRP.

1. Goytain et al., Journal of Physiology-Cell Physiology 2008.
2. Norman and Cicek, Bioinformatics 2019.
3. Satterstrom et al., CELL 2020.

Conclusions

- DeepND is
  - the first multitask gene risk discovery algorithm which can work on comorbid disorders.
  - can utilize multiple networks and deconvolve the informativeness of each gene interaction network considered.
- Can be generalized to work with any combination of disorders/diseases with shared genetic architectures.
- Predicts several novel genes for ASD and ID and helps dissecting out ASD and ID specific genes.

Teşekkürler



CICEKLAB

http://ercumentcicek.com

@crashius

SFARI SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

BİLİM AKADEMİSİ

IEEE Turkey Section

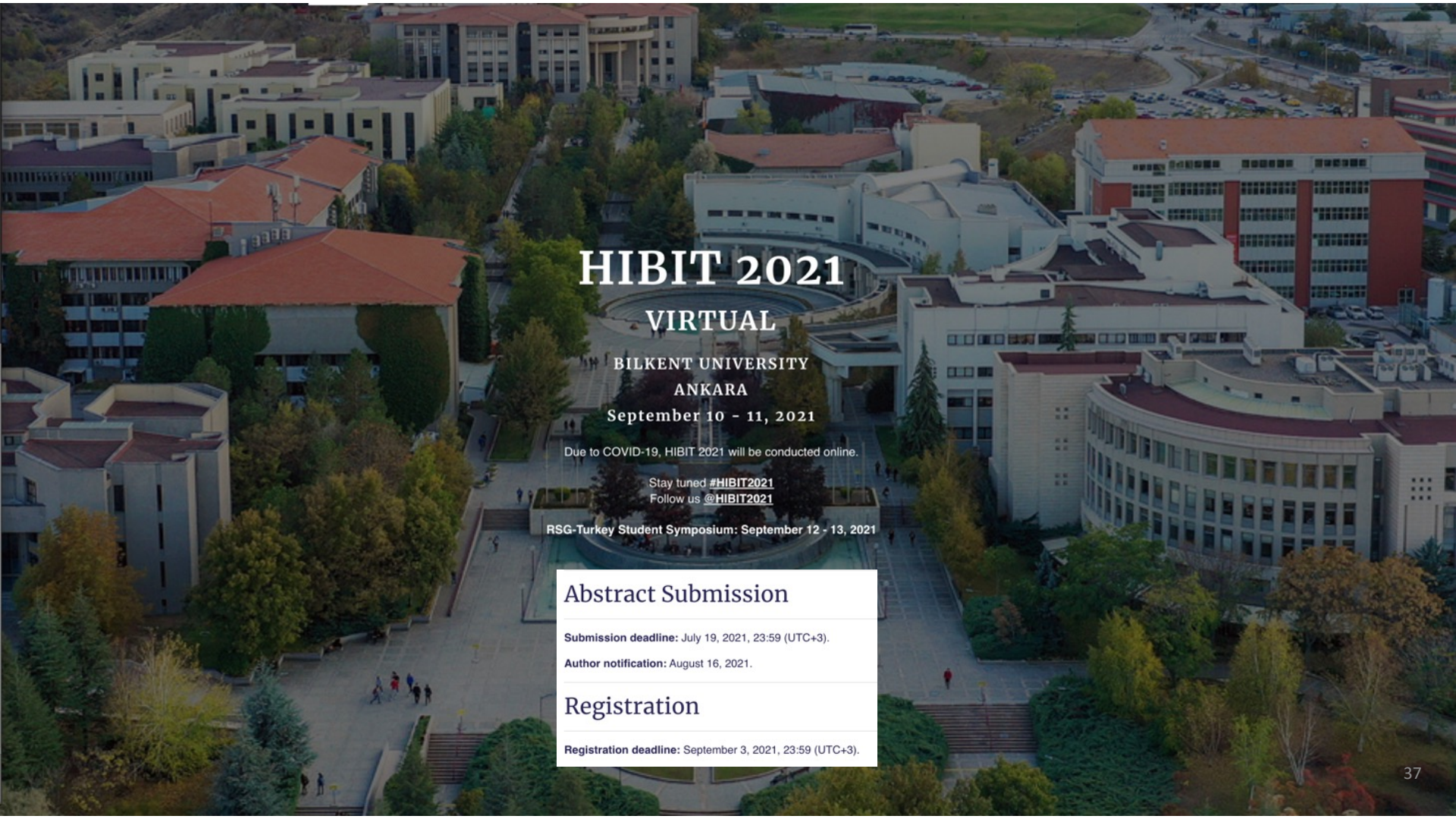
Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
AMBASSADE DE FRANCE EN TURQUIE

ODTÜ PROF. DR. MUSTAFA N. PARLAR EĞİTİM VE ARAŞTIRMA VAKFI

TÜBA

NIH

TÜSEB Türkiye Sağlık Enstitüleri Başkanlığı


HIBIT 2021
VIRTUAL
BILKENT UNIVERSITY
ANKARA
September 10 - 11, 2021
Due to COVID-19, HIBIT 2021 will be conducted online.
Stay tuned #HIBIT2021
Follow us @HIBIT2021
RSG-Turkey Student Symposium: September 12 - 13, 2021
Abstract Submission
Submission deadline: July 19, 2021, 23:59 (UTC+3).
Author notification: August 16, 2021.
Registration
Registration deadline: September 3, 2021, 23:59 (UTC+3).