World Congress on Psychiatric Genetics, 2019
Education Day

Machine Learning in Psychiatric Genomics: Integrating Multiple Levels of Analysis Through Deep Network Models

Jonathan Warrell
Yale University

October 26th, 2019

Outline

I. Machine Learning and Statistics

Similarities/differences, New issues, Potential of ML

II. Recent Machine Learning Approaches and Neuroscience

Bayesian & Causal Networks, Random Forests, SVMs, Deep-Learning

III. Integrative Analysis: PsychENCODE case study

Multilevel modeling, Gene Regulatory Networks (QTLs, genomics, Hi-C),
WGCNA modules, Deep Neural Network integration,
Network Interpretation

Machine Learning and Statistics: Similarities/differences

- Machine learning as an extension of traditional statistics
- Machine learning and statistics address a common set of problems:
  - Classification/Regression (supervised learning)
  - Clustering/Density estimation (unsupervised learning)
  - Dimensionality reduction (unsupervised learning)
  - Time series analysis
- Machine learning also analyzes general learning problems, some of which have analogues in neuroscience/psychology
  - Semi-supervised learning
  - Reinforcement learning
  - Transfer-/meta-learning
  - …

Machine Learning and Statistics: New Issues

- Current machine learning methods characterized by several properties:
  - Complex function classes
  - Optimization landscapes with many local optima
  - Cross validation and generalization testing on hold-out data required
  - Interpretability methods required to analyze models
  - Black-box nature introduces issues of fairness

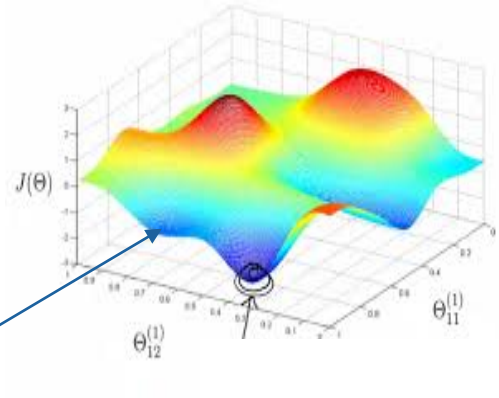
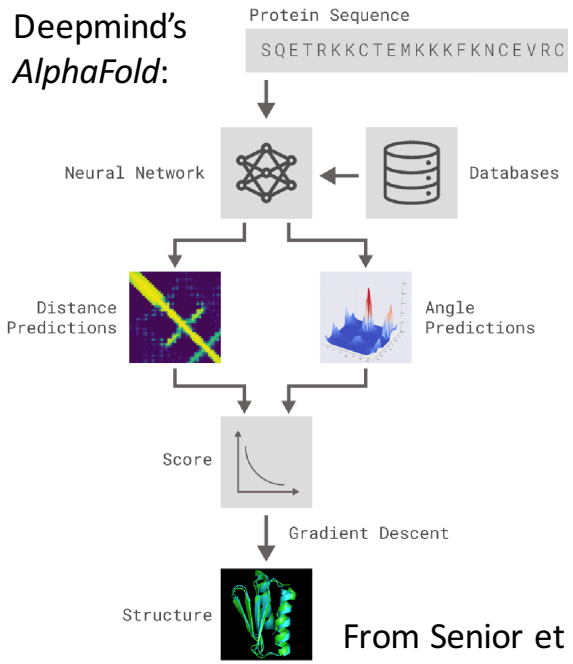


Figure from Arora, 2018

Machine Learning and Statistics: Potential of recent ML methods

- Factors motivating the use of recent machine-learning methods in numerous areas:
  - Deep-learning methods currently provide best (or comparable) predictive accuracy on a wide-range of problems
  - Examples in Genomics/neuroscience:
    - Epigenome annotation
    - Variant effect prediction
    - Protein folding prediction
    - Ancestry reconstruction
    - Polygenic risk scores
    - …

Deepmind's *AlphaFold*:



From Senior et al., 2018

Machine Learning and Statistics: Potential of recent ML methods

- Factors motivating the use of recent machine-learning methods in numerous areas:
  - Simpler models are known/suspected to be inadequate for many problems in genomics and neuroscience
  - Examples
    - In genetics: Problem of missing heritability may point towards substantial epistasis affecting many traits (particularly psychiatric, e.g. ASD, SCZ; Zuk et al., 2014)
    - In neuroscience: Neuronal dynamics are highly non-linear, and hand-tuned models require extensive coarse-graining for tractability (Mejias et al., 2018)

Machine Learning and Statistics: Potential of recent ML methods

- Factors motivating the use of recent machine-learning methods in numerous areas:
  - Interpretability methods are being developed to probe models for many problems of interest
  - Examples
    - In genomics: Deep-LIFT for variant prioritization (Shrikumar et al., 2017)
    - In neuroscience: Gradient-based methods for identifying causally relevant voxels in fMRI (Dezfouli et al., 2018)
    - In psychiatric genomics: Pathway/module-based annotation of hidden nodes in polygenic models (Wang et al., 2018)

Machine Learning and Statistics: Potential of recent ML methods

- Factors motivating the use of recent machine-learning methods in numerous areas:
  - Theoretical analysis of deep-learning has progressed substantially, clarifying its statistical properties (Arora, 2018)
  - Examples
    - Compressibility analysis: Networks learned on real data are highly stable and compressible
    - Information Bottleneck principle: DL automatically identifies relevant features

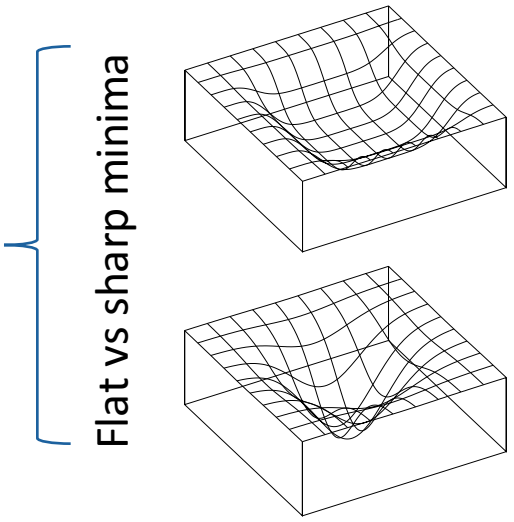


Figure from Arora, 2018

Recent Machine Learning Approaches and Neuroscience: Bayesian and Causal Networks

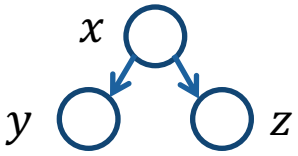
- Comparison of methods

Method	Flexible function forms	Global optimization	Compact models	Interpretability
Linear/logistic regression		X	X	X
Bayesian/causal networks	(X)	X	X	X
Random Forests	X			X
SVMs	X	X		
Deep Neural Networks	X		X	(X)

Recent Machine Learning Approaches and Neuroscience: Bayesian and Causal Networks

- Bayesian and Causal Networks
  - Any probability distribution can be factored: $p(x,y,z) = p(x)p(y|x)p(z|x,y)$
  - Bayesian networks model independence assumptions via a directed acyclic graph (DAG):

$$p(x,y,z) = p(x)p(y|x)p(z|x)$$



[e.g. gene expression levels]

  - Causal networks carry additional semantics

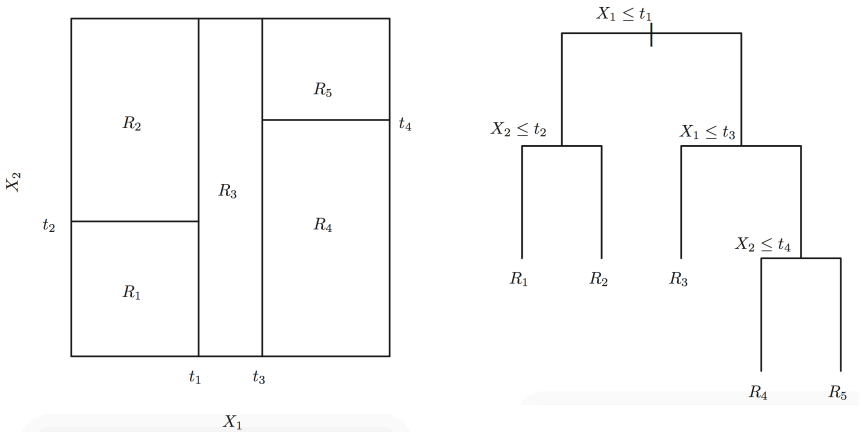
$$p(x,y,z) = p(x)p(y|x)p(z|y) \qquad p_{\mathrm{do}(y=y_0)}(x,y,z) = p(x)\delta(y=y_0)p(z|y_0)$$

- Applications in Genomics:
  - Genome annotation (ChromHMM, Ernst and Kellis, 2010)
  - Causal modeling of protein signaling networks (Sachs et al., 2005)
  - QTL modeling with hidden factors (PEER, Stegle et al., 2012)

Recent Machine Learning Approaches and Neuroscience: Random Forests and Support Vector Machines (SVMs)

- Random Forests and SVMs are flexible ML classification and regression models
- Both are universal approximators (can model any function)
- Random Forests:
  - Recursively split training data based on an error function such as the Gini-index or entropy

Left: A 2-dimensional parameter space is partitioned into 5 subspaces
Right: Tree representation of the resultant partition



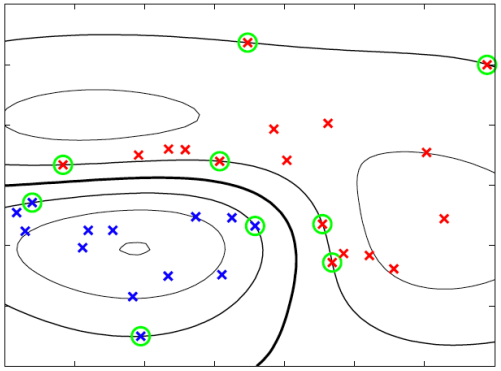
- Applications in Genomics/Neuroscience:
  - RFs: Brain cell-type discovery and prediction (Lake et al., 2014; Tasic et al., 2018)

Recent Machine Learning Approaches and Neuroscience: Random Forests and Support Vector Machines (SVMs)

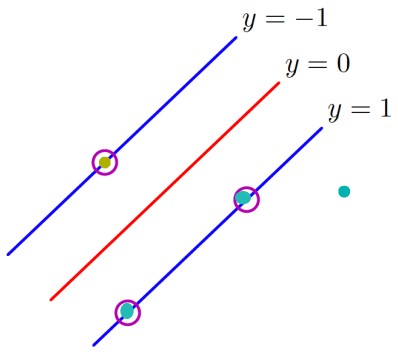
- Support Vector Machines:
  - Learn a linear classifier in a high-dimensional feature space
  - Kernel can be chosen based on problem; highly flexible
  - Global optimum can be found (convex), but model may be non-compact

SVM, with RBF kernel:

Input space:



Feature space:



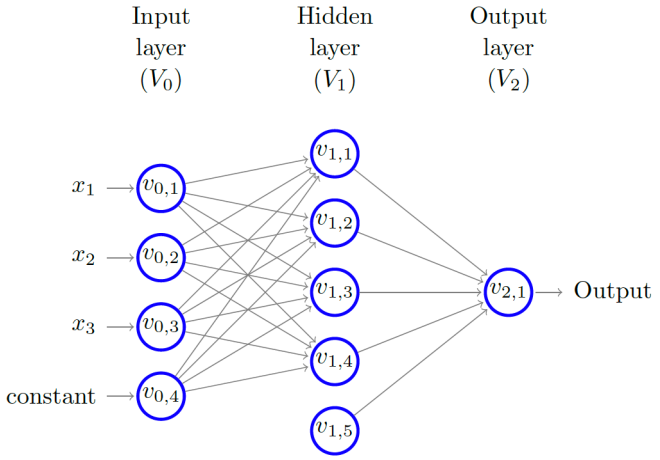
C. Bishop, 2006

- Applications in Genomics/Neuroscience:
  - SVMs: Binding site motif discovery (gkmSVM, Ghandi et al., 2014)
  - SVMs: Neuronal population code analysis (Schuck and Niv, 2018)

Recent Machine Learning Approaches and Neuroscience: Deep-Learning

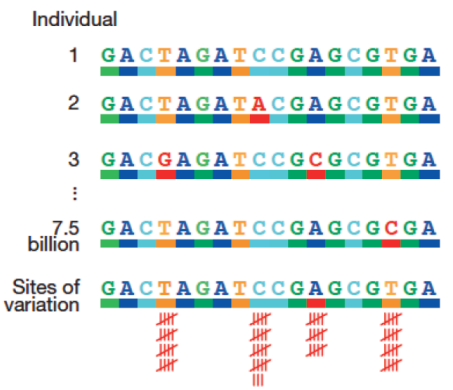
- Deep-Learning provides an alternative approach to supervised and unsupervised learning
- Uses highly non-linear models which can be efficiently optimized using back-propagation and stochastic gradient descent
  - NB optimization is non-global
- Hidden nodes discover features from training data (require interpreting)
- Function form: $y = \mathbf{W}_L \sigma(\mathbf{W}_{L-1} \sigma(\ldots \mathbf{W}_0 \mathbf{x}))$

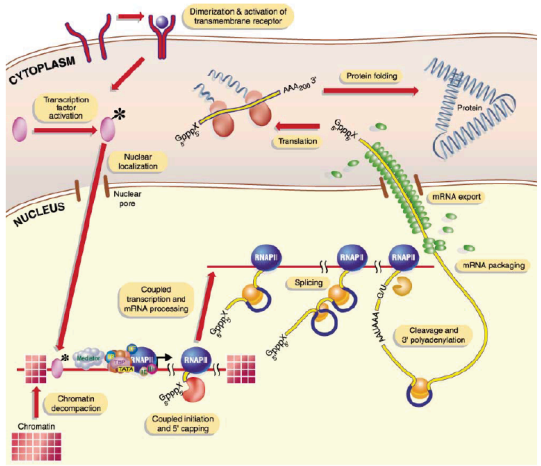
- Applications in Genomics/Neuroscience:
  - In silico prediction of variant effects on TF binding motifs (DeepSEA, Zhou et al., 2015)
  - Epigenome annotation (Avacado, Schreiber et al., 2018)
  - Non-linear PRSs (Tran et al., 2018)
  - Joint language/fMRI model (Jain et al., 2018)

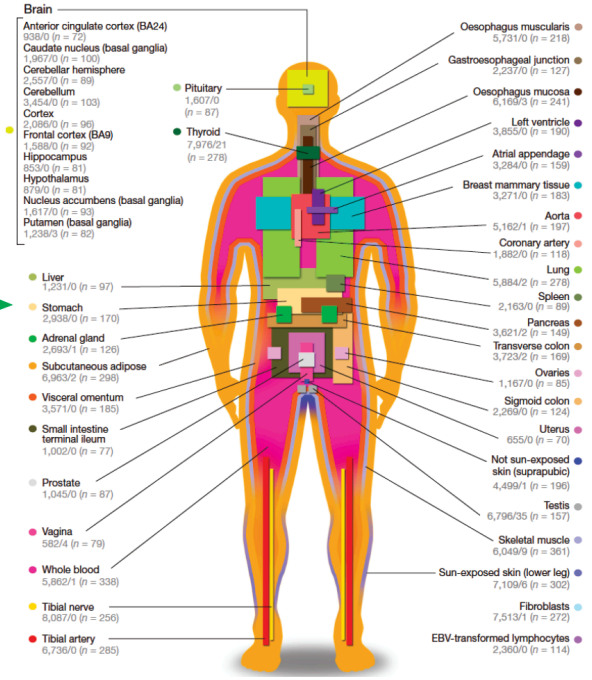


Integrative Analysis – PsychENCODE case study: Multilevel modeling

- For any trait, the effects of genetic variation arise through perturbations of processes at cellular and sub-cellular levels, which can be probed through genomics data.







Genetics

Shendure et al. **2017**, *Nat. Rev.* 550, P. 345-353

Genomics

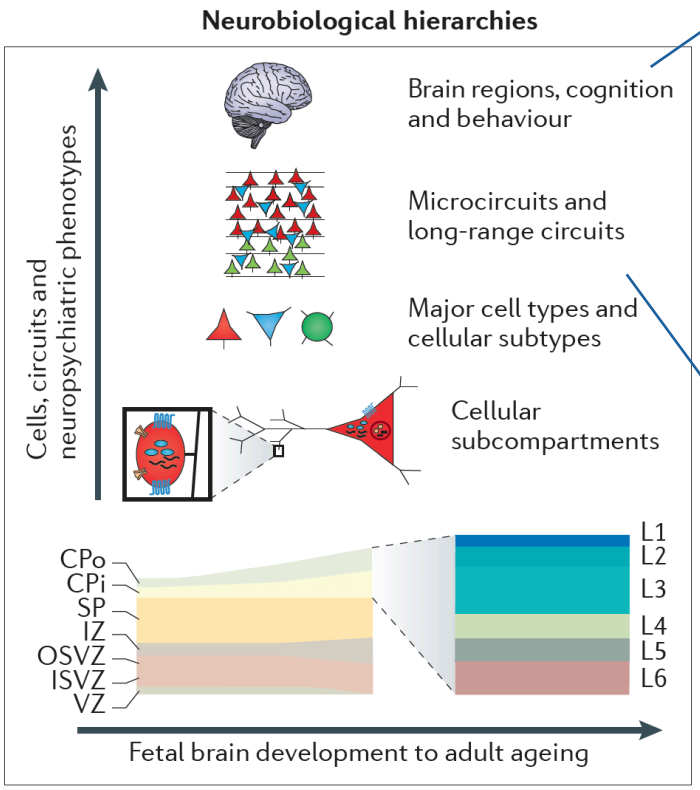
Orphanides and Reinberg **2002**, *Cell 108*, P. 439-451

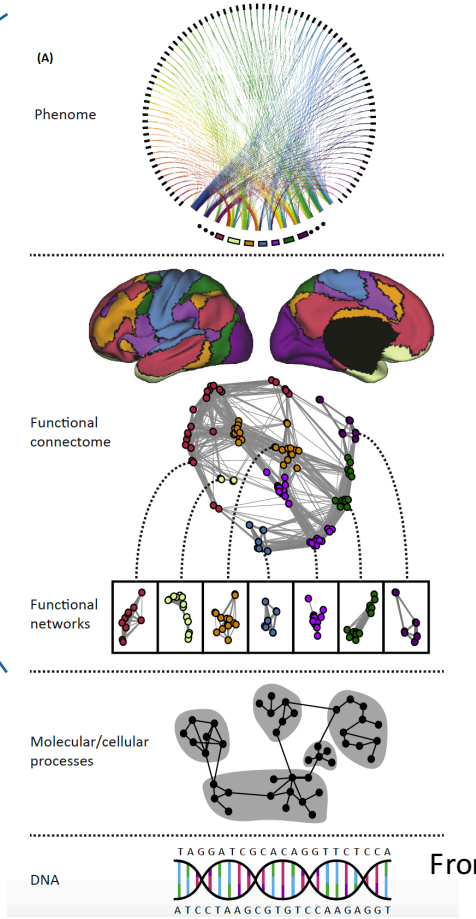
Clinical

GTEx Consortium **2017**, *Nature* 550, P. 204-213

Integrative Analysis – PsychENCODE case study: Multilevel modeling

- For psychiatric and brain-related traits, additional layers of complexity are relevant.



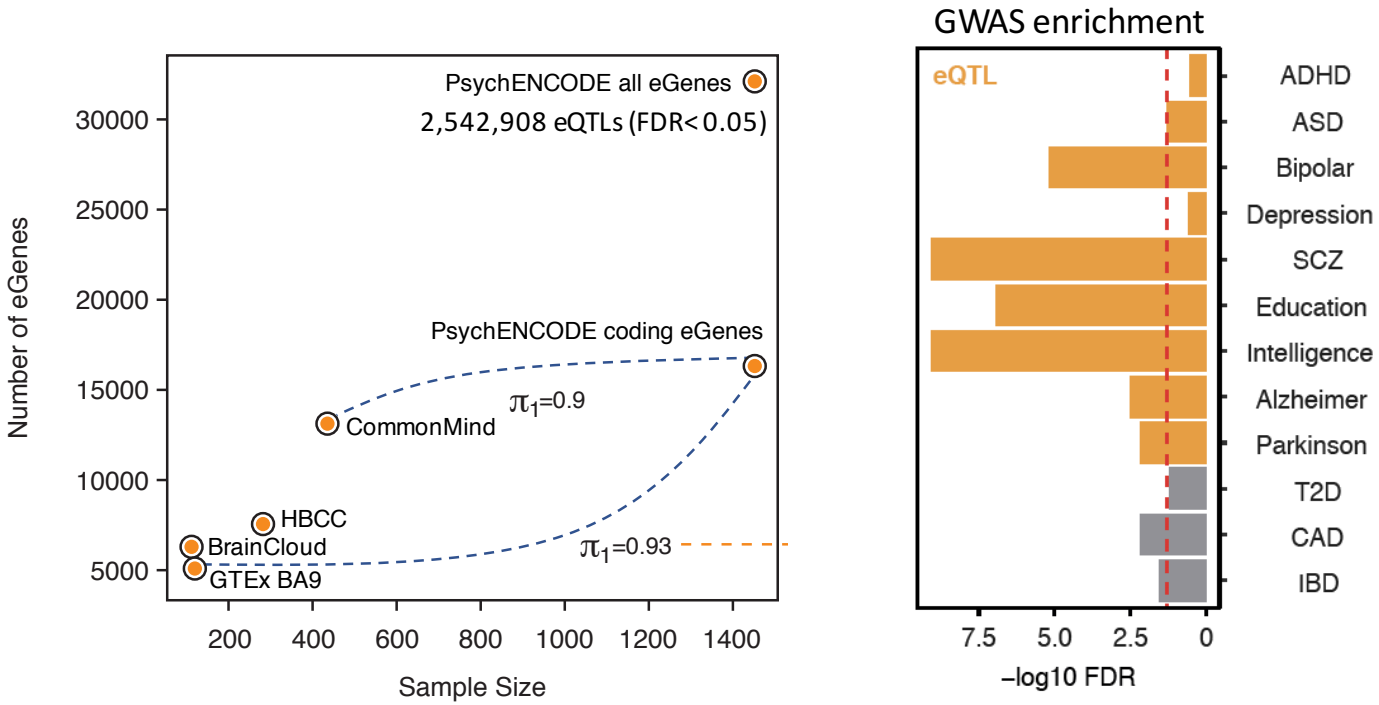


From: Parikshak et al., 2015
Holmes et al., 2018

Integrative Analysis – PsychENCODE case study: Multilevel modeling

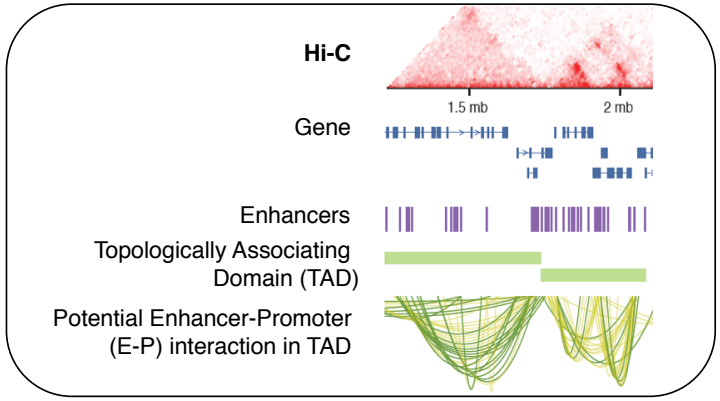
- For PsychENCODE integrative analysis (Wang et al., 2018), we introduced a deep-learning based predictive model for psychiatric conditions (Schizophrenia, Bipolar, Autism)
  - Learn to predict genomics variables from genetic variation
  - Model unobserved intermediate layers as latent factors
  - Predict psychiatric disorders from genetics by imputing genomics and latent factors
- Motivating assumptions:
  - Additive polygenic risk scores achieve low predictive power because of failure to model epistasis
  - Epistatic interactions can be captured through deep-learning by using genomics to embed known structure into the model
- Data:
  - 1039 control, 558 SCZ, 217 BPD and 44 ASD subjects
  - Post-mortem Prefrontal Cortex samples from all subjects
  - WGS genotype data from all subjects
  - RNA-seq data, all subjects
  - ChIP-seq (h3k27ac), and Hi-C data from a subset of subjects
  - Also integrate data from GTEx, Epigenetics Roadmap & CommonMind

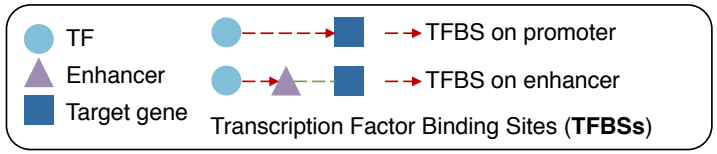
Integrative Analysis – PsychENCODE case study: Larger eQTL sets than previous brain studies

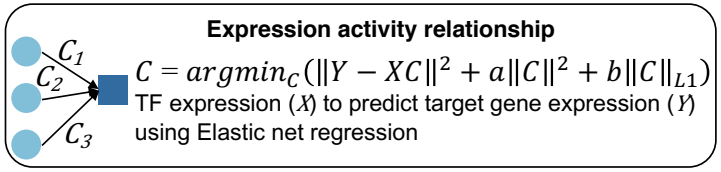


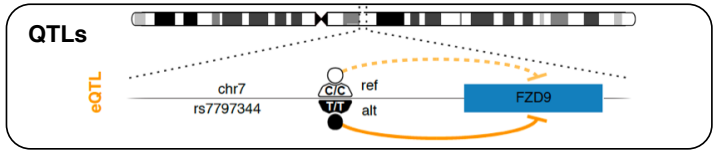
- Also, calculate ~8K chromatin QTLs (cQTLs), 1.6K cell fraction QTLs (fQTLS) and others

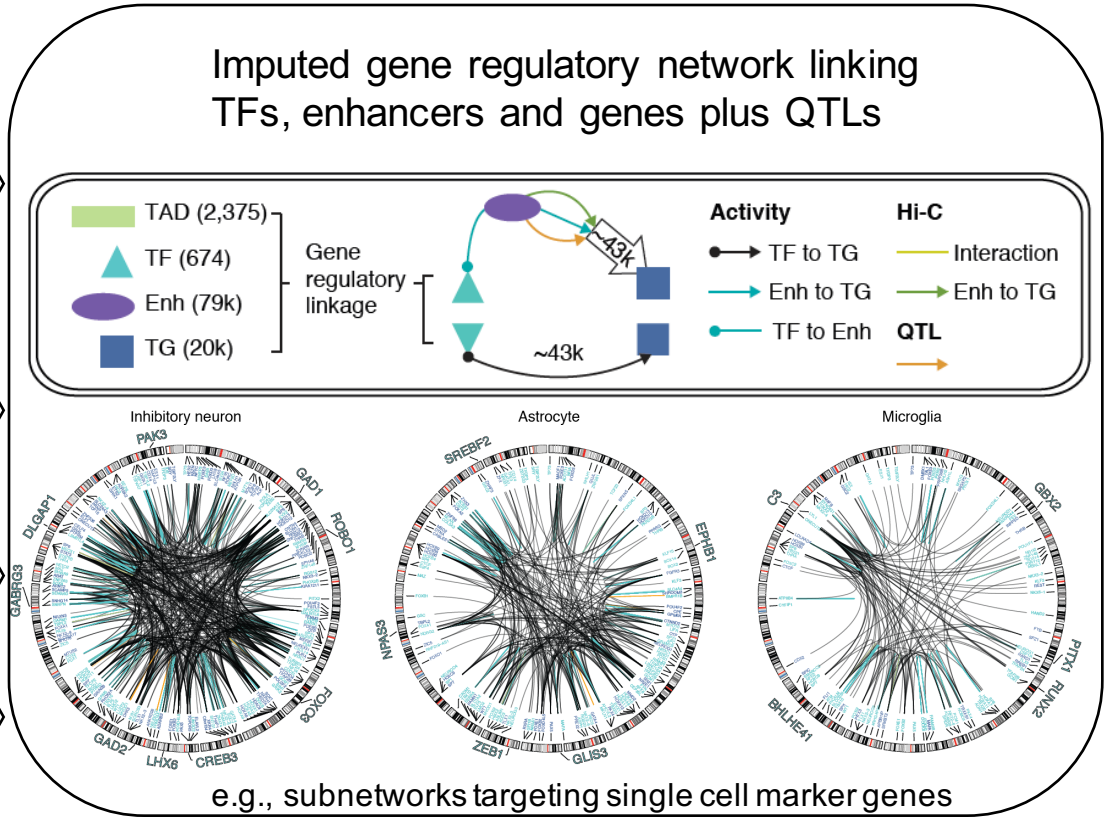
Integrative Analysis – PsychENCODE case study: Gene regulatory network inference









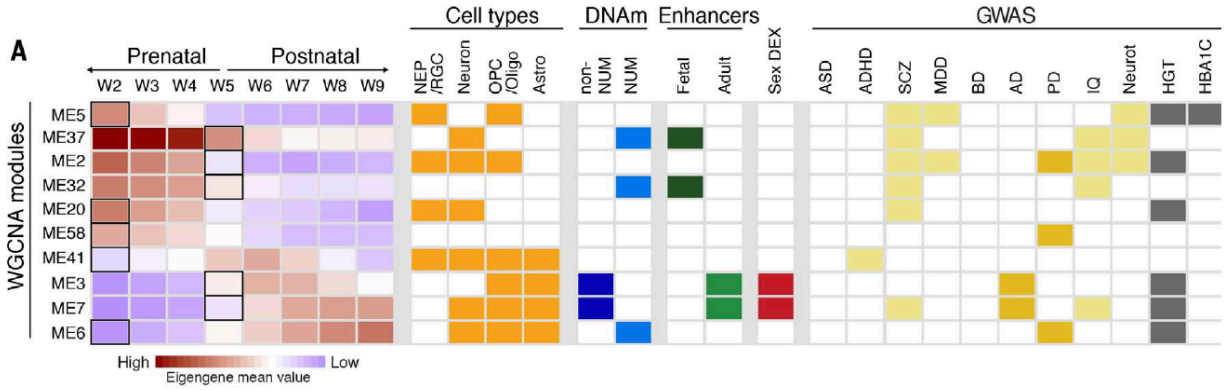


Integrative Analysis – PsychENCODE case study: Weighted Gene Co-expression Network Analysis (WGCNA)

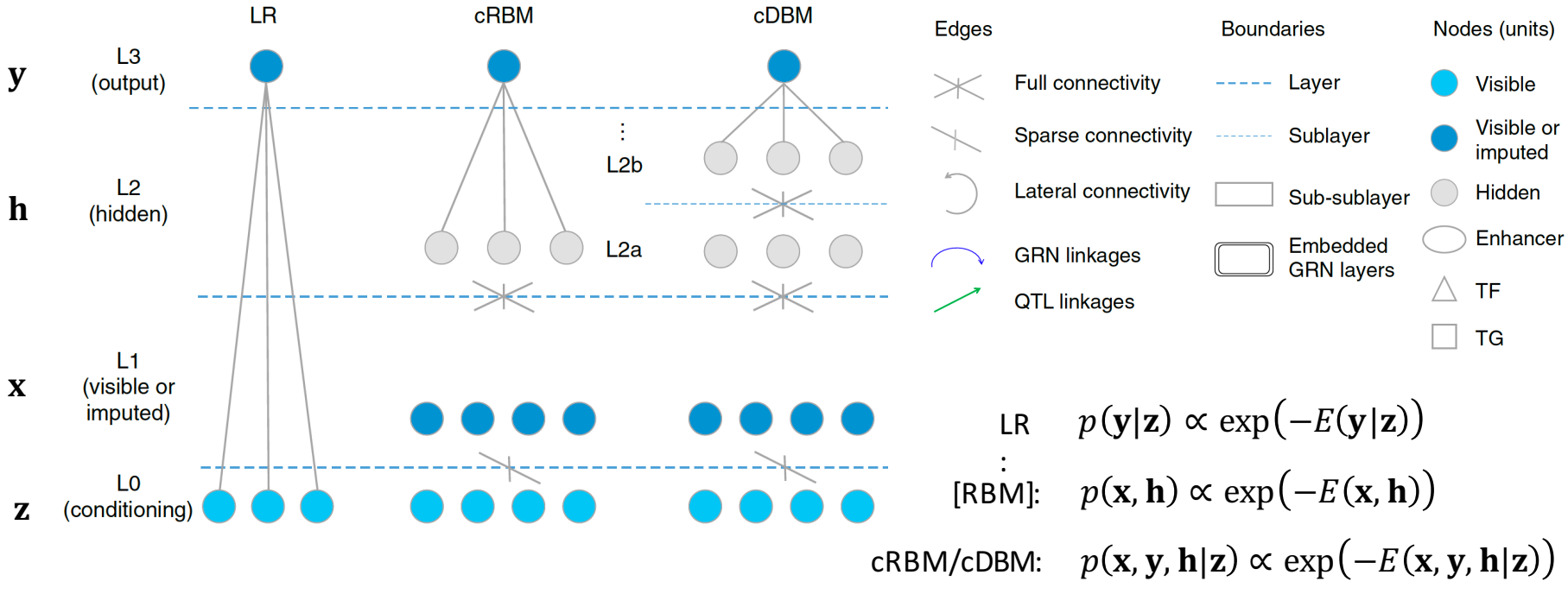
- WGCNA algorithm for finding network modules (Zhang et al., `05):

1. Start with a weighted network (correlation in expression values)
2. Compute the `Topological overlap' between all pairs of genes (proportional to # neighbors shared)
3. Build dendrogram using mean distance agglomerative clustering
4. Cut tree to produce final modules (Dynamic Tree Cut)

- We apply WGCNA to find 5024 modules
- Other PEC analyses find modules which change over development (Li et al., 2018), and with cross-disorder associations (Gandal et al., 2018)



Integrative Analysis – PsychENCODE case study: Conditional Boltzmann machine models



LR $p(\mathbf{y}|\mathbf{z}) \propto \exp\left(-E(\mathbf{y}|\mathbf{z})\right)$

:

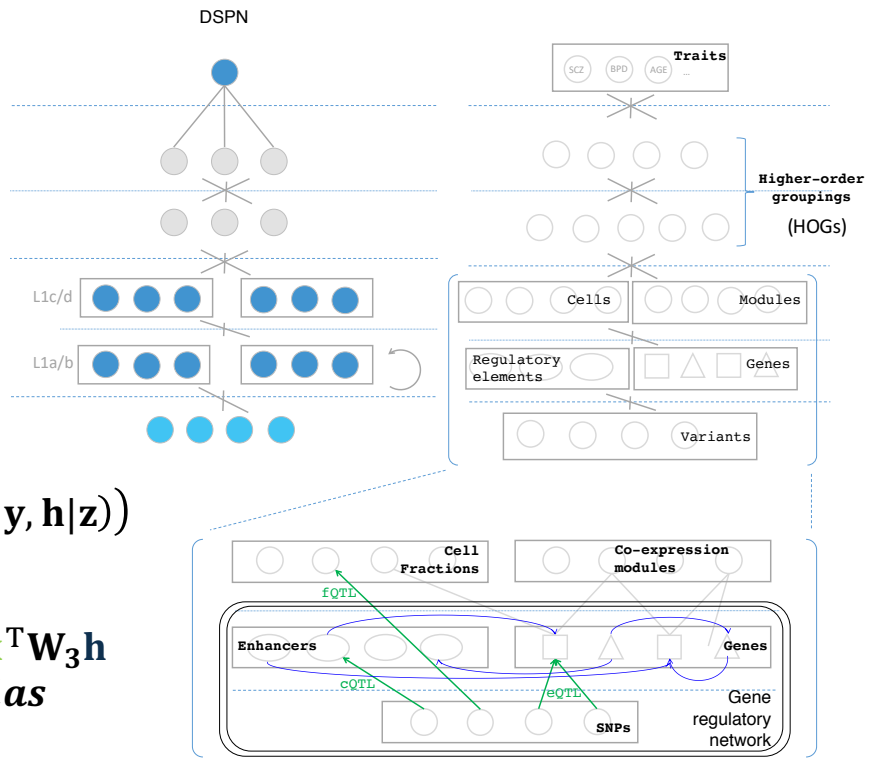
[RBM]: $p(\mathbf{x}, \mathbf{h}) \propto \exp\left(-E(\mathbf{x}, \mathbf{h})\right)$

cRBM/cDBM: $p(\mathbf{x}, \mathbf{y}, \mathbf{h}|\mathbf{z}) \propto \exp\left(-E(\mathbf{x}, \mathbf{y}, \mathbf{h}|\mathbf{z})\right)$

cRBM: V. Mnih, H. Larochelle, G. E. Hinton, Conditional Restricted Boltzmann Machines for Structured Output Prediction. *UAI*, 2012.
DBM: R. Salakhutdinov, G. Hinton, Deep Boltzmann Machines. *AISTATS*, 2009.

Integrative Analysis – PsychENCODE case study: Deep Structured Phenotype Network (DSPN)

GRN and QTL linkages embedded in DSPN



Energy model:

$$p(\mathbf{x}, \mathbf{y}, \mathbf{h}|\mathbf{z}) \propto \exp\left(-E(\mathbf{x}, \mathbf{y}, \mathbf{h}|\mathbf{z})\right)$$

$$E(\mathbf{x}, \mathbf{y}, \mathbf{h}|\mathbf{z}) = -\mathbf{z}^{\mathrm{T}}\mathbf{W}_1\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{W}_2\mathbf{x} - \mathbf{x}^{\mathrm{T}}\mathbf{W}_3\mathbf{h} - \mathbf{h}^{\mathrm{T}}\mathbf{W}_4\mathbf{h} - \mathbf{h}^{\mathrm{T}}\mathbf{W}_5\mathbf{y} - \boldsymbol{Bias}$$

Boltzmann machine

y: phenotypes (SCZ, BPD, ASD)

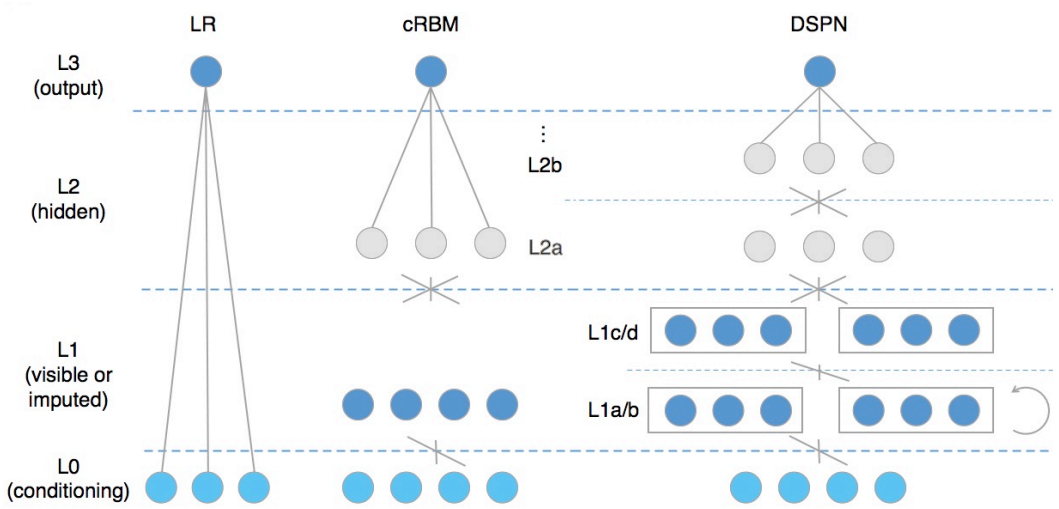
h: hidden units (e.g., circuits)

x: intermediate phenotypes (e.g., genes, enhancers)

z: genotypes (e.g. SNPs)

W: weights (e.g., regulatory network)

Integrative Analysis – PsychENCODE case study: DSPN improves disorder prediction over baseline PRS

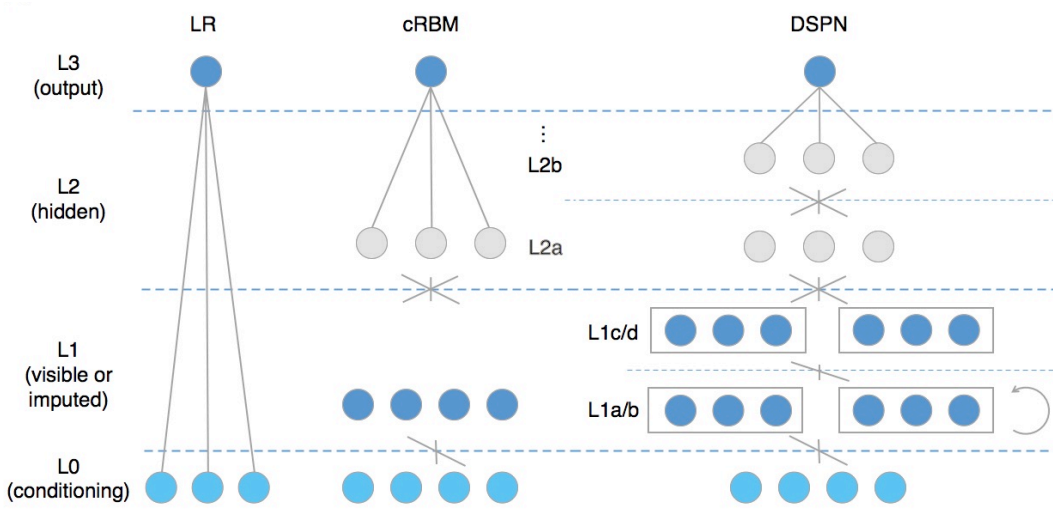


Method	LR-genotype	LR-transcriptome	cRBM	DSPN-imputation	DSPN-full
Schizophrenia	**54.6%**	63.0%	70.0%	59.0%	**73.6%**
Bipolar Disorder	**56.7%**	63.3%	71.1%	67.2%	**76.7%**
ASD	**50.0%**	51.7%	67.2%	62.5%	**68.3%**

X 6.0

Accuracy = chance to correctly predict disease/health

Integrative Analysis – PsychENCODE case study: Latent factors help prediction in DSPN

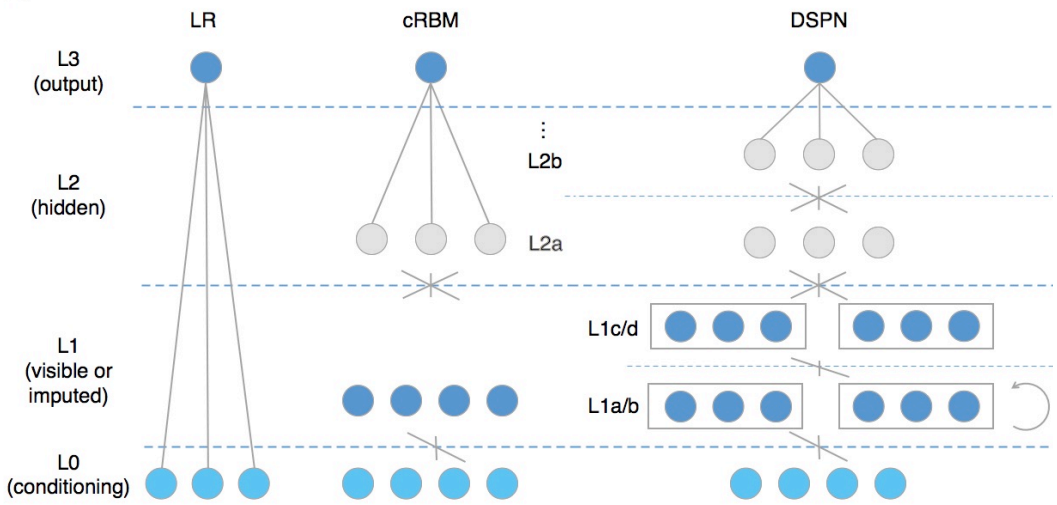


Method	LR-genotype	LR-transcriptome	cRBM	DSPN-imputation	DSPN-full
Schizophrenia	54.6%	63.0%	70.0%	59.0%	73.6%
Bipolar Disorder	56.7%	63.3%	71.1%	67.2%	76.7%
ASD	50.0%	51.7%	67.2%	62.5%	68.3%

X 2.5

Accuracy = chance to correctly predict disease/health

Integrative Analysis – PsychENCODE case study: Incorporating prior structure helps prediction

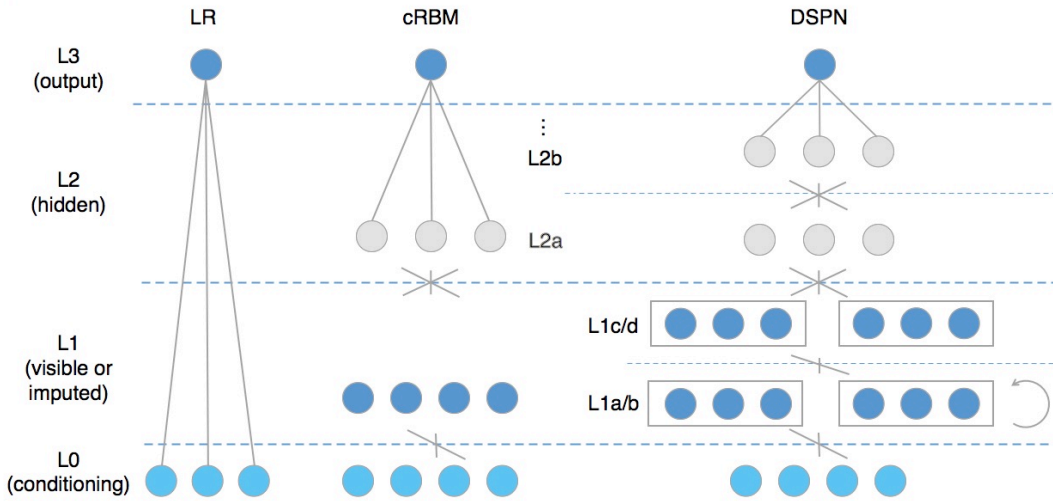


Method	LR-genotype	LR-transcriptome	cRBM	DSPN-imputation	DSPN-full
Schizophrenia	**54.6%**	63.0%	70.0%	**59.0%**	73.6%
Bipolar Disorder	**56.7%**	63.3%	71.1%	**67.2%**	76.7%
ASD	**50.0%**	51.7%	67.2%	**62.5%**	68.3%

X 3.1

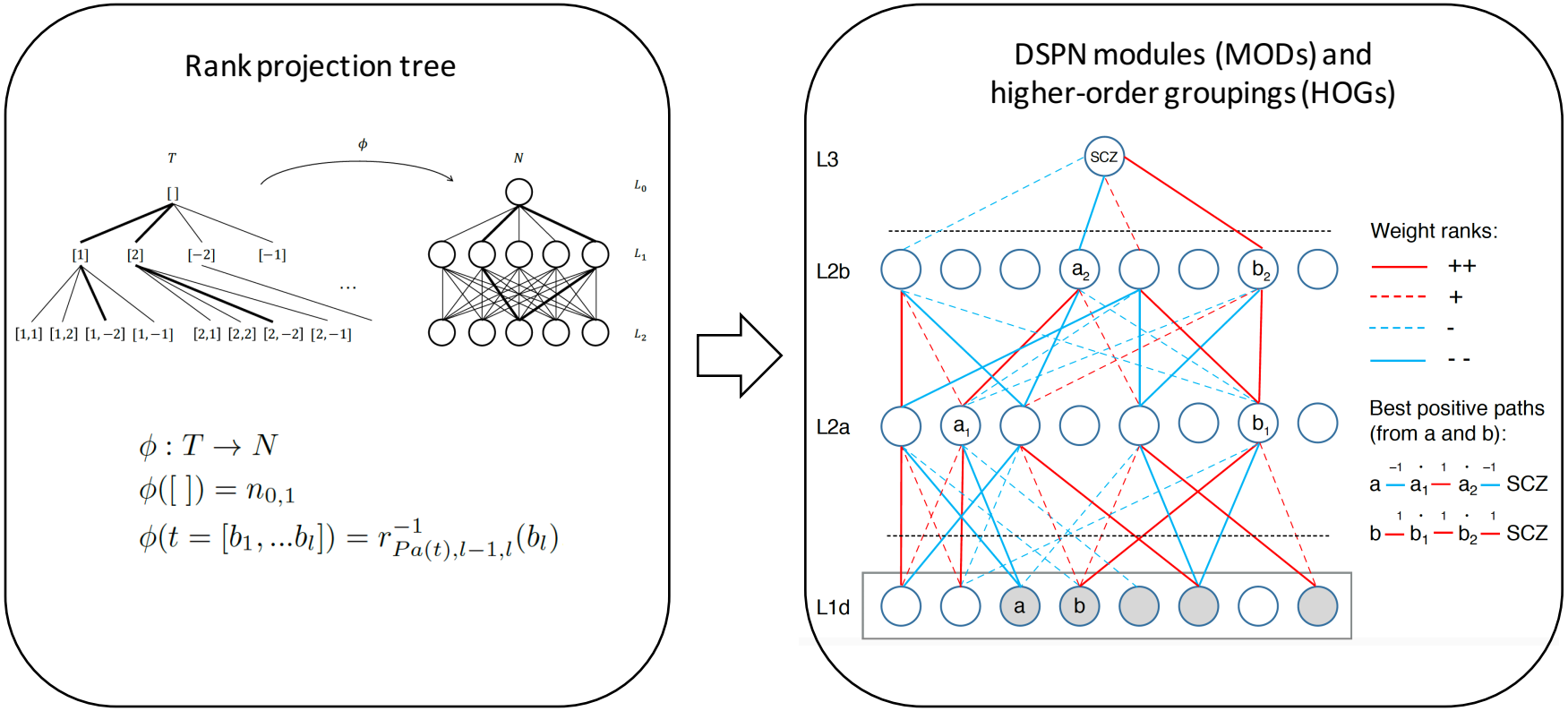
Accuracy = chance to correctly predict disease/health

Integrative Analysis – PsychENCODE case study: Converting predictive performance to liability scale



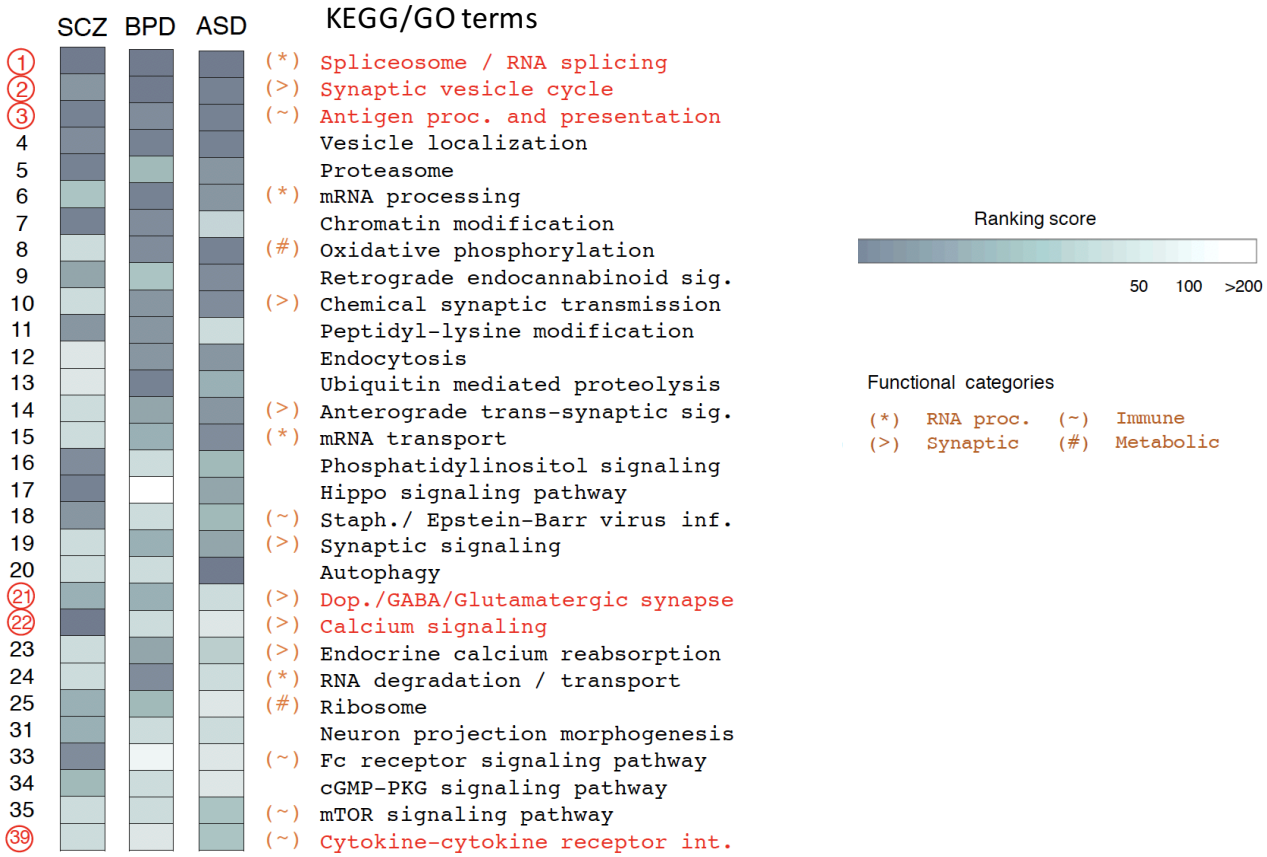
Method	LR-genotype	LR-transcriptome	cRBM	DSPN-imputation	DSPN-full
Schizophrenia	0.5%	4.8%	31.0%	1.8%	32.8%
Bipolar Disorder	2.5%	6.3%	22.6%	10.7%	37.4%
ASD	0.0%	1.8%	10.8%	3.2%	11.3%

Integrative Analysis – PsychENCODE case study: Multilevel network interpretation

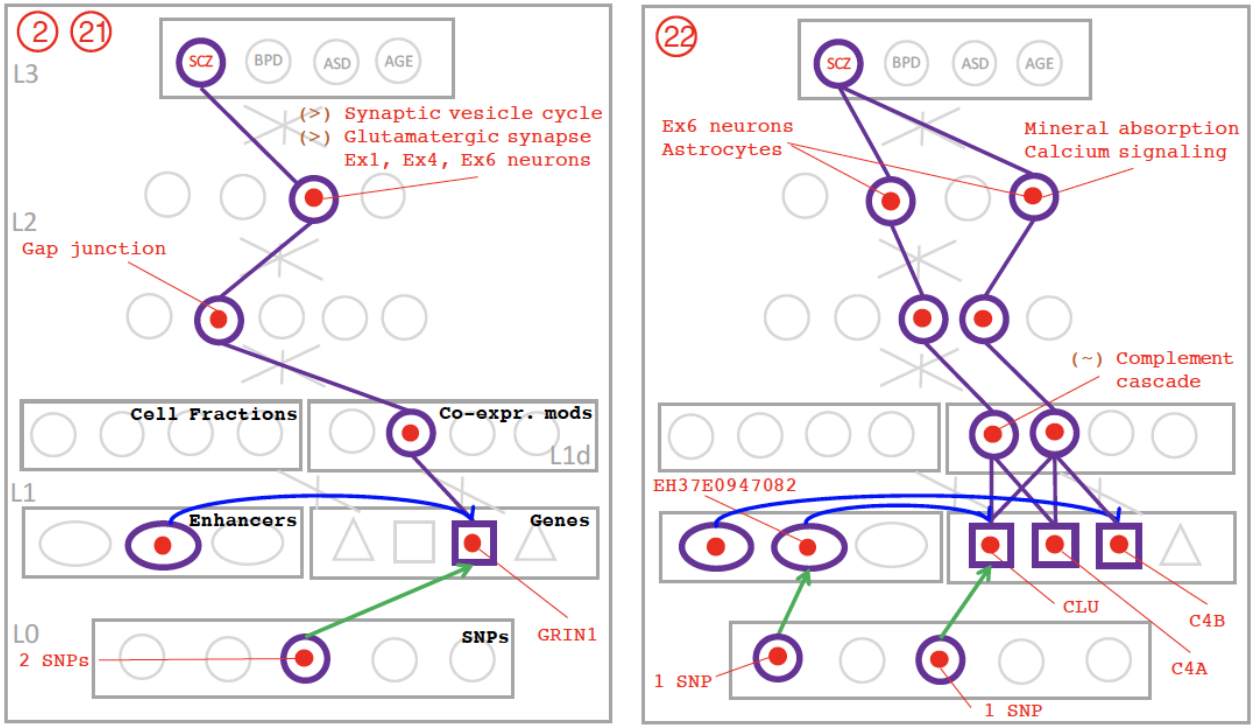


J. Warrell, H. Mohsen, and M. Gerstein. Rank projection trees for Multilevel Neural Network Interpretation. *NeurIPS ML4H workshop,* 2018

Integrative Analysis – PsychENCODE case study: Cross-disorder ranking of functional terms



Integrative Analysis – PsychENCODE case study: Linkage of eQTLs, cQTLs and enhancers to prioritized modules (SCZ)



Summary

- Current methods in machine learning lie on a spectrum with traditional statistics
- Recent methods such as Deep-Learning are highly flexible, and give high performance across multiple tasks
- However, require novel methods for model interpretation: many available and are being developed
- Many successes in genomics and neuroscience: tantalizing results may point to beginnings of a shared understanding of the brain and principles underlying AI

Thanks for your attention!

References

- Arora, S., 2018. Towards Theoretical Understanding of Deep Learning. *ICML'18 Tutorial*. http://unsupervised.cs.princeton.edu/deeplearningtutorial.html
- Dezfouli, A., Morris, R., Ramos, F.T., Dayan, P. and Balleine, B., 2018. Integrated accounts of behavioral and neuroimaging data using flexible recurrent neural network models. In *Advances in Neural Information Processing Systems* (pp. 4228-4237).
- Ernst, J. and Kellis, M., 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nature methods*, *9*(3), p.215.
- Gandal, M.J., Zhang, P., Hadjimichael, E., Walker, R.L., Chen, C., Liu, S., Won, H., Van Bakel, H., Varghese, M., Wang, Y. and Shieh, A.W., 2018. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science*, *362*(6420), p.eaat8127.
- Ghandi, M., Lee, D., Mohammad-Noori, M. and Beer, M.A., 2014. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS computational biology*, *10*(7), p.e1003711.
- GTEx Consortium, 2017. Genetic effects on gene expression across human tissues. *Nature*, *550*(7675), p.204.
- Hinton, G.E., 2012. A practical guide to training restricted Boltzmann machines. In *Neural networks: Tricks of the trade* (pp. 599-619). Springer, Berlin, Heidelberg.
- Holmes, A.J. and Patrick, L.M., 2018. The myth of optimality in clinical neuroscience. *Trends in cognitive sciences*, *22*(3), pp.241-257.

References

- Jain, S. and Huth, A., 2018. Incorporating context into language encoding models for fMRI. In *Advances in Neural Information Processing Systems* (pp. 6628-6637).
- Lake, B.B., Ai, R., Kaeser, G.E., Salathia, N.S., Yung, Y.C., Liu, R., Wildberg, A., Gao, D., Fung, H.L., Chen, S. and Vijayaraghavan, R., 2016. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*, *352*(6293), pp.1586-1590.
- Li, M., Santpere, G., Kawasawa, Y.I., Evgrafov, O.V., Gulden, F.O., Pochareddy, S., Sunkin, S.M., Li, Z., Shin, Y., Zhu, Y. and Sousa, A.M., 2018. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science*, *362*(6420), p.eaat7615.
- Mejias, J.F., Murray, J.D., Kennedy, H. and Wang, X.J., 2016. Feedforward and feedback frequency-dependent interactions in a large-scale laminar network of the primate cortex. *Science advances*, *2*(11), p.e1601335.
- Mnih, V., Larochelle, H. and Hinton, G.E., 2012. Conditional restricted boltzmann machines for structured output prediction. *arXiv preprint arXiv:1202.3748*.
- Orphanides, G. and Reinberg, D., 2002. A unified theory of gene expression. *Cell*, *108*(4), pp.439-451.
- Parikshak, N.N., Gandal, M.J. and Geschwind, D.H., 2015. Systems biology and gene networks in neurodevelopmental and neurodegenerative disorders. *Nature Reviews Genetics*, *16*(8), p.441.

References

- Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D.A. and Nolan, G.P., 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, *308*(5721), pp.523-529.
- Salakhutdinov, R. and Hinton, G., 2009, April. Deep boltzmann machines. In *Artificial intelligence and statistics* (pp. 448-455).
- Schuck, N.W. and Niv, Y., 2019. Sequential replay of nonspatial task states in the human hippocampus. *Science*, *364*(6447), p.eaaw5181.
- Schreiber, J., Durham, T., Bilmes, J. and Noble, W.S., 2019. Multi-scale deep tensor factorization learns a latent representation of the human epigenome. *BioRxiv*, p.364976.
- Senior et al., 2018. AlphaFold: Using AI for scientific discovery. https://deepmind.com/blog/article/alphafold
- Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A. and Waterston, R.H., 2017. DNA sequencing at 40: past, present and future. *Nature*, *550*(7676), p.345.
- Shrikumar, A., Greenside, P. and Kundaje, A., 2017, August. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (pp. 3145-3153). JMLR. org.
- Stegle, O., Parts, L., Durbin, R. and Winn, J., 2010. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS computational biology*, *6*(5), p.e1000770.

References

- Tasic, B., Yao, Z., Graybuck, L.T., Smith, K.A., Nguyen, T.N., Bertagnolli, D., Goldy, J., Garren, E., Economo, M.N., Viswanathan, S. and Penn, O., 2018. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, *563*(7729), p.72.
- Tran, D. and Blei, D.M., 2017. Implicit causal models for genome-wide association studies. *arXiv preprint arXiv:1710.10742*.
- Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F.C., Clarke, D., Gu, M., Emani, P., … and Gerstein, M. 2018. Comprehensive functional genomic resource and integrative model for the human brain. *Science*, *362*(6420), p.eaat8464.
- Warrell, J., Mohsen, H. and Gerstein, M., 2018. Rank Projection Trees for Multilevel Neural Network Interpretation. *arXiv preprint arXiv:1812.00172*.
- Zhang, B. and Horvath, S., 2005. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, *4*(1).
- Zhou, J. and Troyanskaya, O.G., 2015. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, *12*(10), p.931.
- Zuk, O., Hechter, E., Sunyaev, S.R. and Lander, E.S., 2012. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, *109*(4), pp.1193-1198.