VARMOLE: A BIOLOGICALLY DROP-CONNECT DEEP NEURAL NETWORK MODEL FOR PRIORITIZING DISEASE RISK VARIANTS AND GENES

RSG
WITH
DREAM
2020

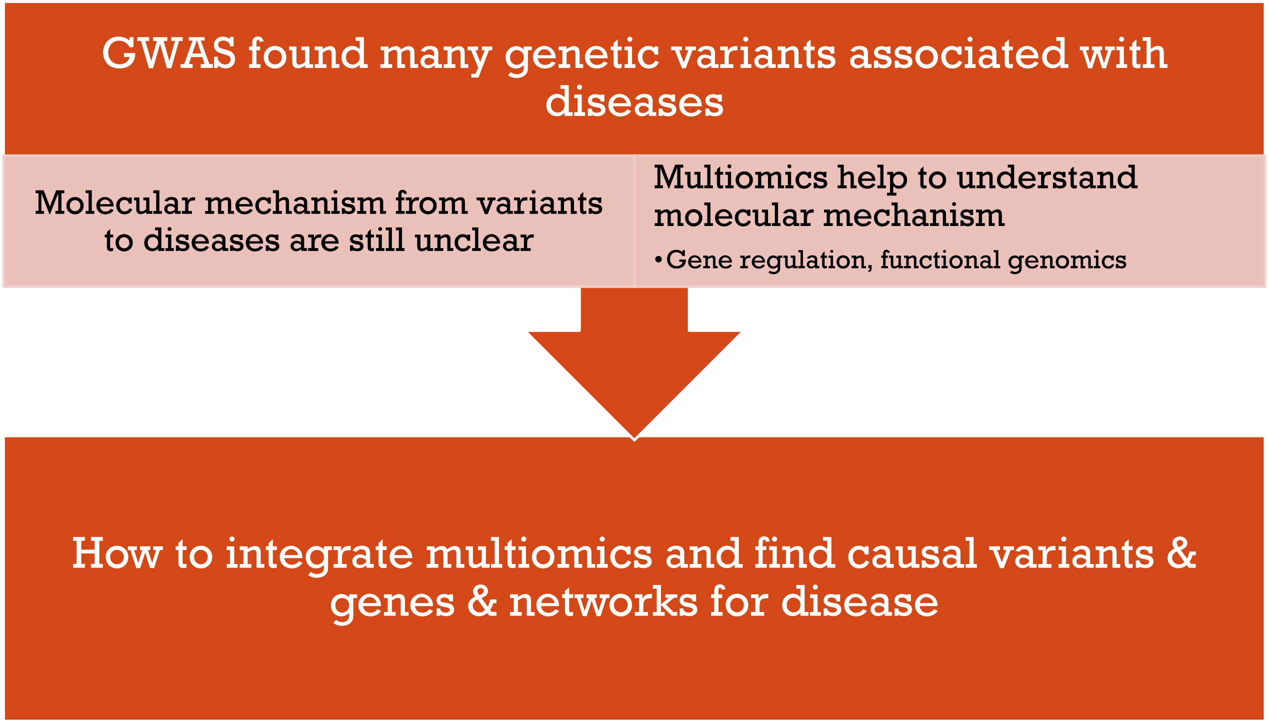
Nam D. Nguyen[1,3], Ting Jin[2,3], Daifeng Wang[2,3,*]

[1]Department of Computer Science, Stony Brook University, Stony Brook, NY 11794, USA

[2]Deparment of Biostatistics and Medical Informatics, University of Wisconsin-Madison, WI 53726, USA

[3]Waisman Center, University of Wisconsin-Madison, WI 53705, USA

MOTIVATION



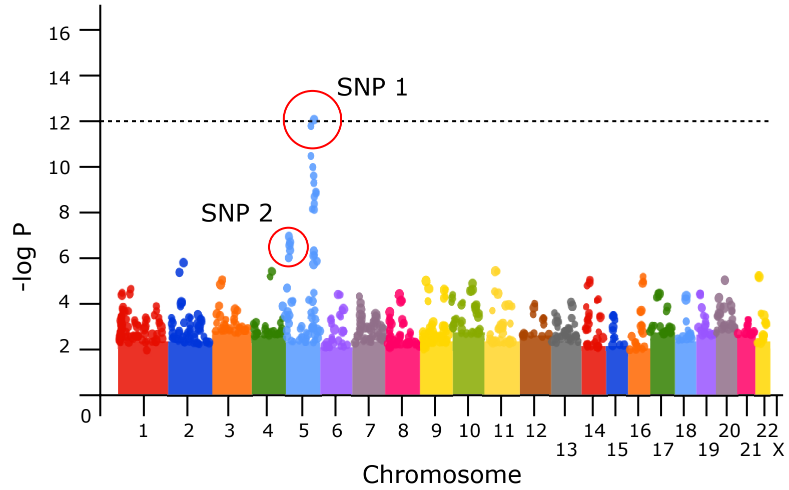
GWAS found many genetic variants associated with diseases

Molecular mechanism from variants to diseases are still unclear

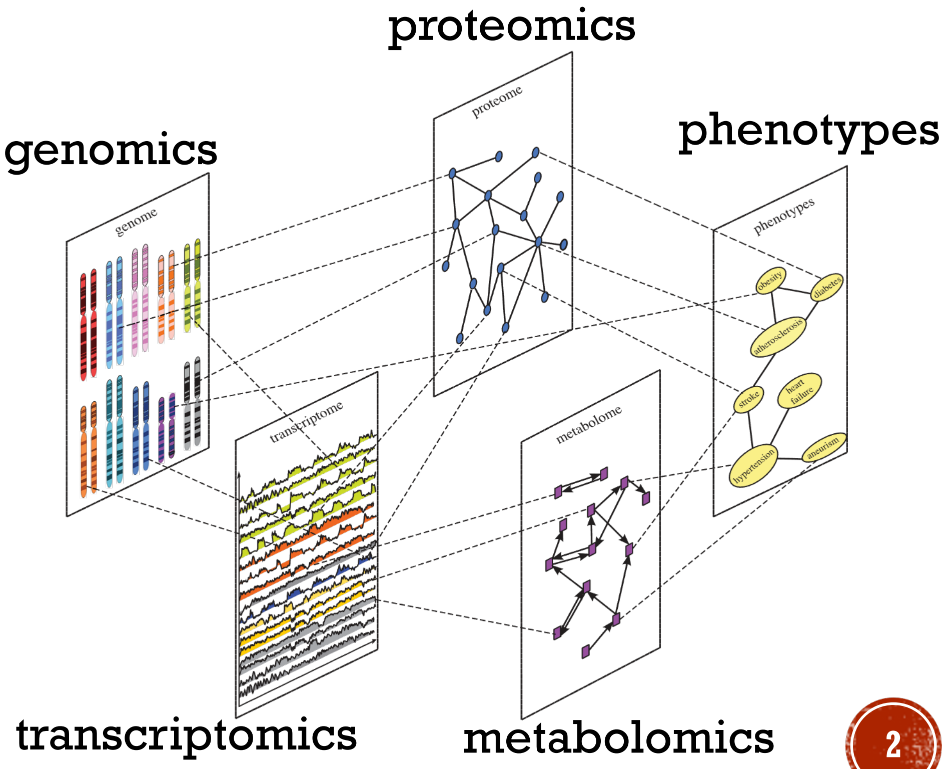
Multiomics help to understand molecular mechanism

- Gene regulation, functional genomics

How to integrate multiomics and find causal variants & genes & networks for disease

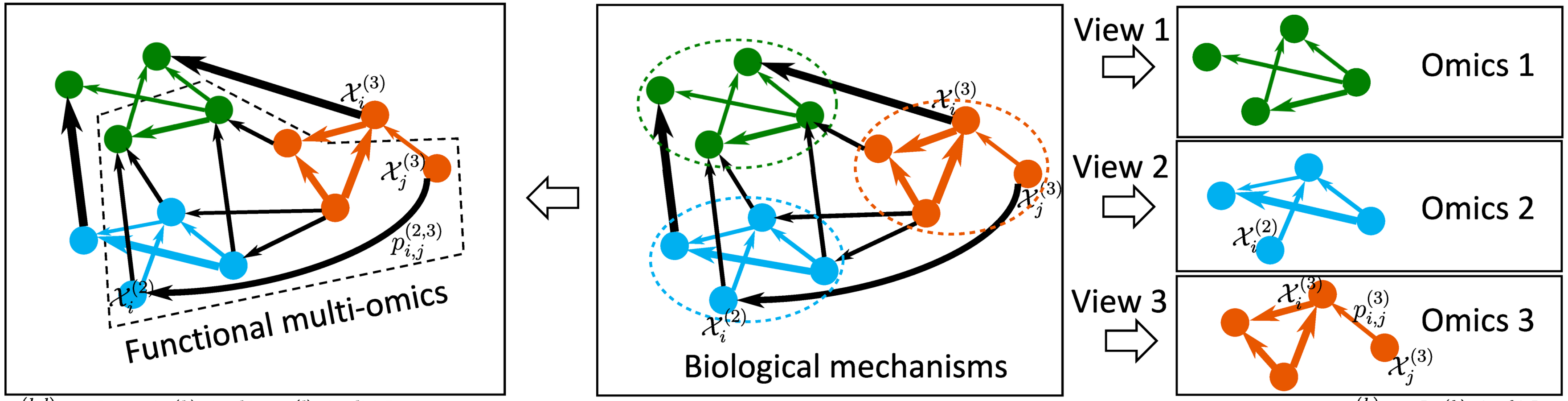


https://commons.wikimedia.org/wiki/File:GWAS-%C3%9Cbersicht.svg



Gligorijevic´ V, Prˇzulj N., J. R. Soc. Interface 12: 20150571.

MULTIVIEW LEARNING FOR UNDERSTANDING FUNCTIONAL MULTI-OMICS



$p_{i,j}^{(k,l)} = \left[\Omega_{co}\left(f^{(k)}\left(\mathcal{X}^k\right), f^{(l)}\left(\mathcal{X}^l\right)\right)\right]_{i,j}$ Multi-view learning

Interaction pattern across omics

Single-view learning $p_{i,j}^{(k)} = \left[f^{(k)}\left(\mathcal{X}^k\right)\right]_{i,j}$

Interaction pattern within an omic

- For example, gene regulation can relate to
  1. Genomics; e.g., SNPs
  2. Transcriptomics; e.g., genes
  3. Proteomics; e.g., transcription factors (TFs)

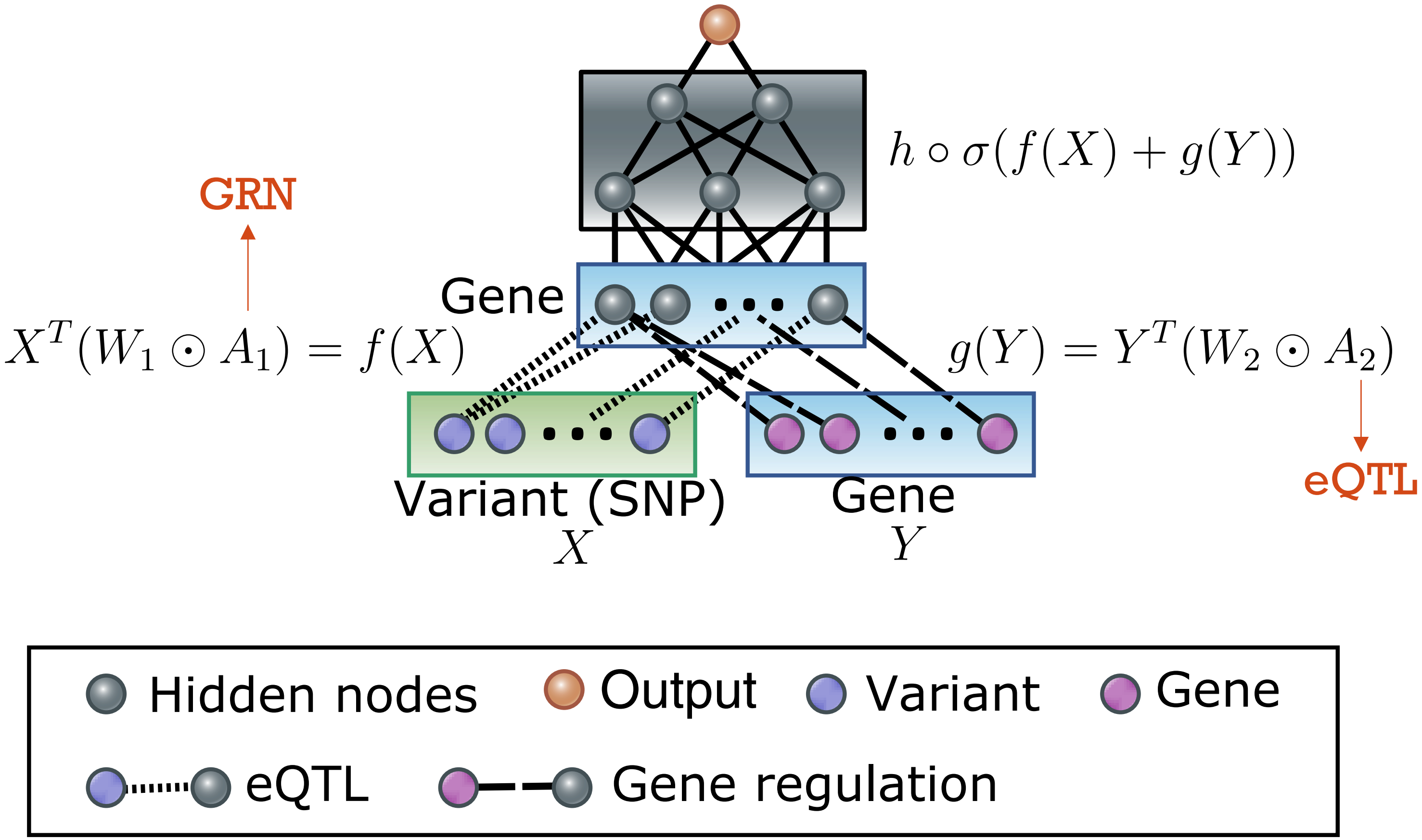
Cross-talk patterns:

$\Omega_{co}(f^{(1)}, f^{(3)})$: SNPs break TF binding sites

$\Omega_{co}(f^{(2)}, f^{(3)})$: TFs control gene expression

$\Omega_{co}(f^{(1)}, f^{(2)})$: SNPs associate with gene expression

Nguyen, Wang, PLoS Computational Biology, 2020


$h \circ \sigma(f(X) + g(Y))$
GRN
Gene
$X^T(W_1 \odot A_1) = f(X)$
$g(Y) = Y^T(W_2 \odot A_2)$
eQTL
Variant (SNP)
X
Gene
Y
Hidden nodes
Output
Variant
Gene
eQTL
Gene regulation

Nguyen, Jin, Wang, Bioinformatics, 2020

VARMOLE

- Input form 2 views, X, Y (SNPs & genes)
- First layer embed A_1 and A_2 – gene regularoty network (GRN) and eQTL

→ From variants (& gene regulations) to gene expression

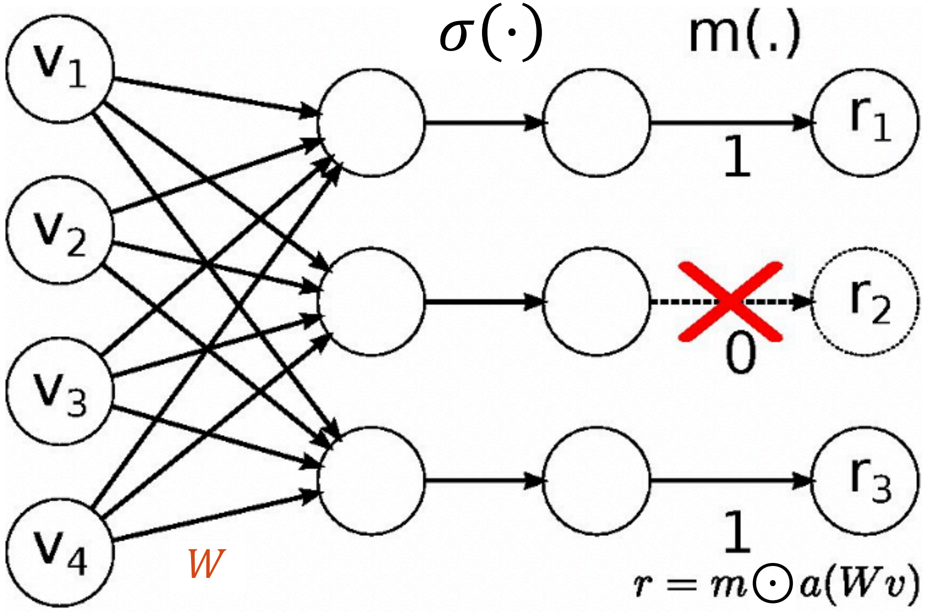
- Other fully connected hidden layers: h;

→ From gene expression to phenotypes

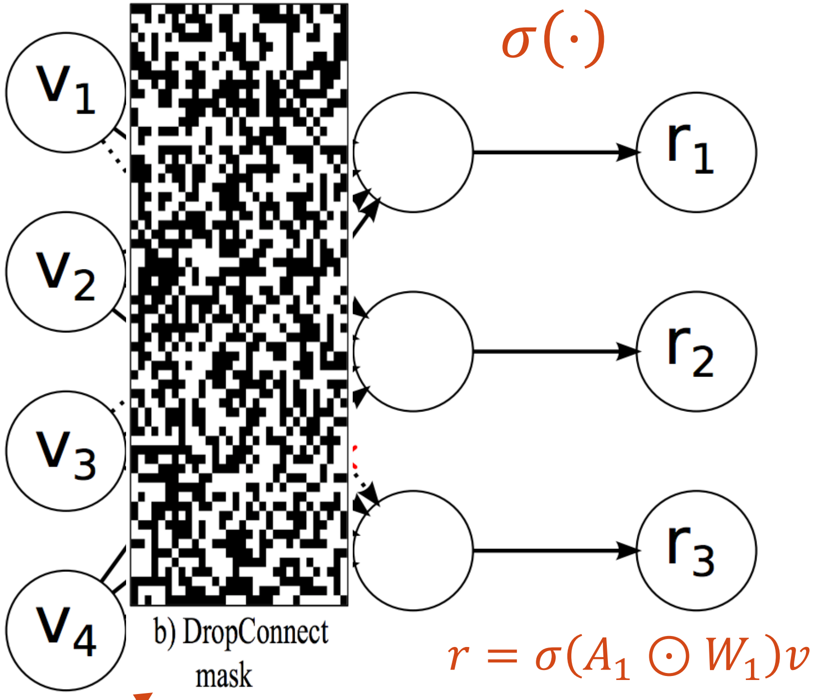
- Softmax classification layer: $o = \delta\left(h \circ \sigma\big(f(X) + g(Y)\big)\right)$;
- The Cross-Entropy: $L(o, \hat{o}) = -\frac{1}{n \sum_{i=1}^{n} y_i log(\hat{y}_i)}$
- Varmole: $\min\ L(o, \hat{o}) + \|W\|_1$

DROP-CONNECT

- Drop-out and drop-connect are 2 simple but effective regularization techniques



$r = m \odot a(Wv)$



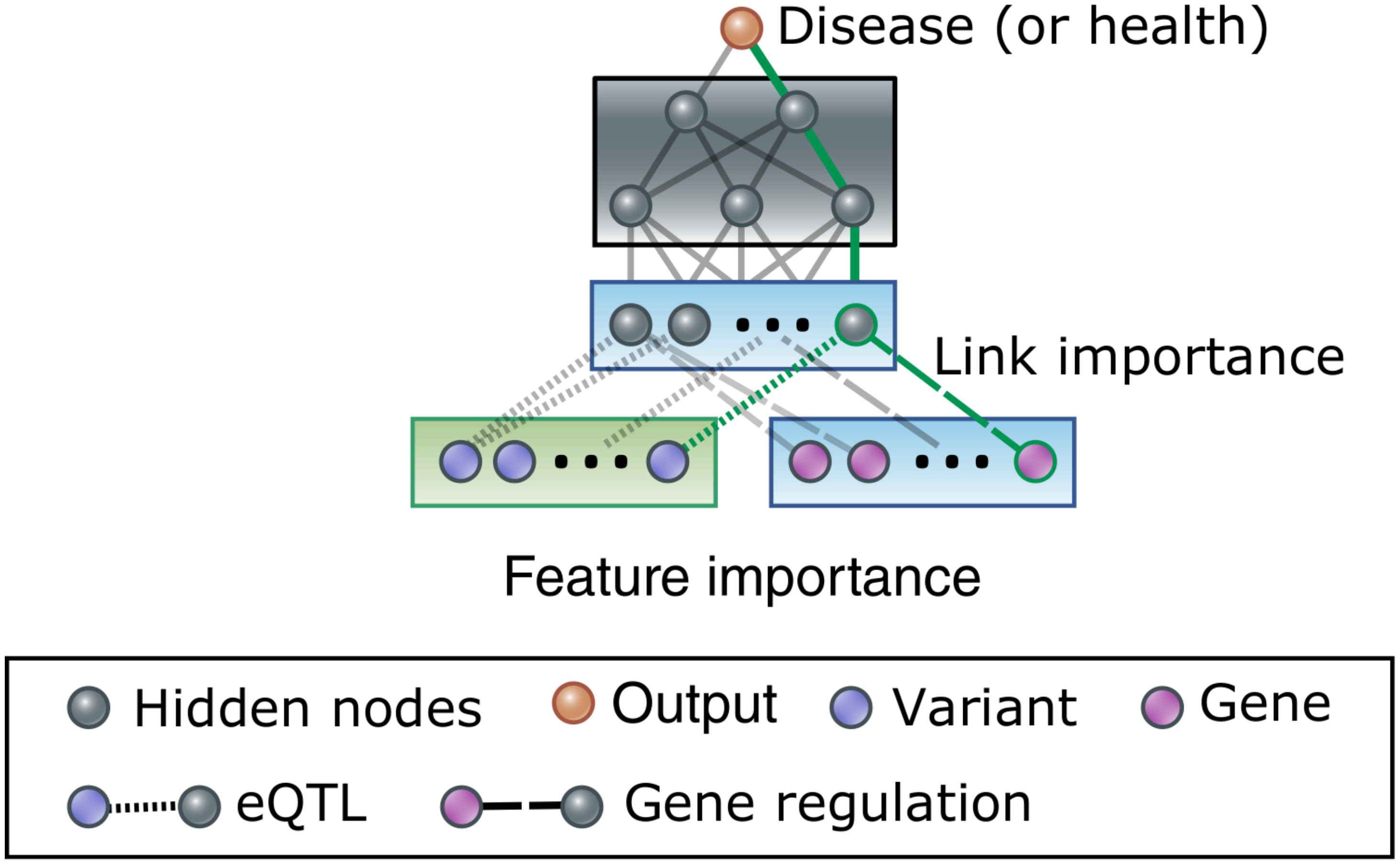
$r = \sigma(A_1 \odot W_1)v$

The drop-connect mask is GRN or eQTL (A_1 or A_2)

INTERPRETATION: PRIORITIZATION VIA INTEGRATED GRADIENTS

- Given a model F, an input x, and the output $F(x)$ of the model for input in question, an attribution methods returns the ‘relevance’ of each input feature i to the output

Interpret with Integrated gradient



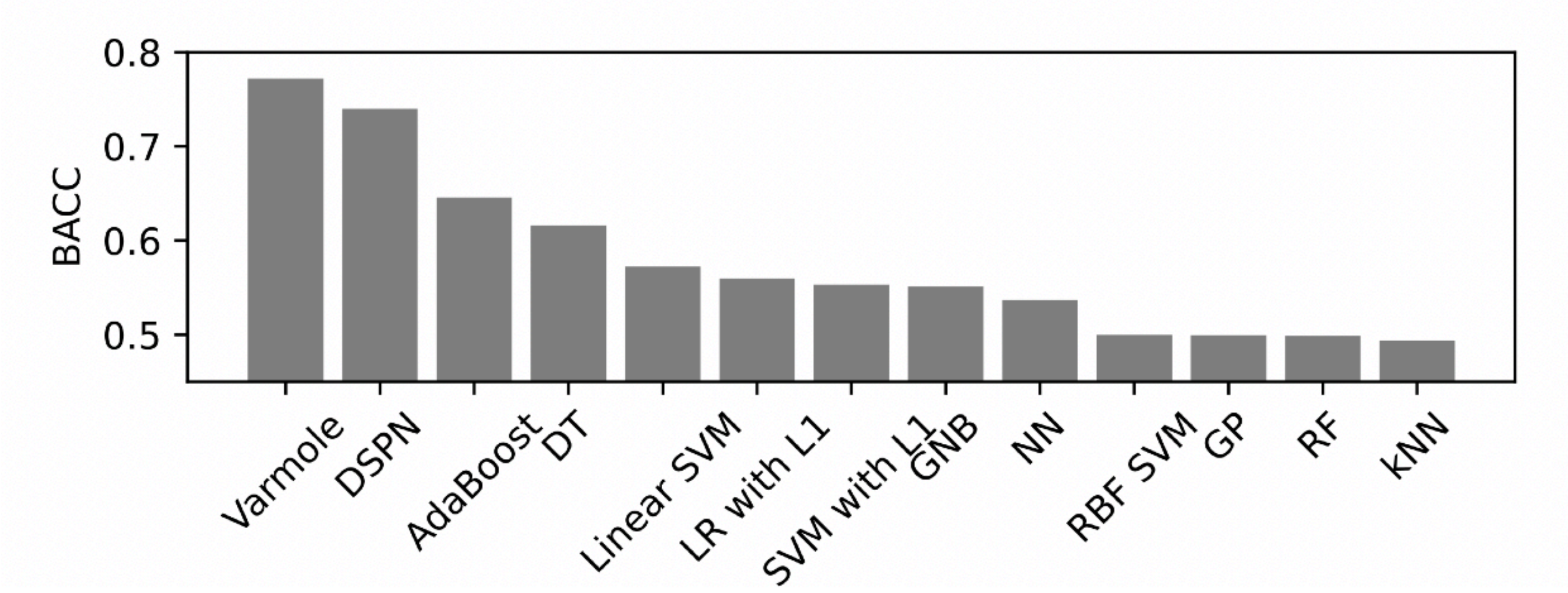
$$\boxed{\text{IntegratedGrads}_i(x)} ::= (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i}\, d\alpha$$

Importance score of feature i of input x

Sundararajan, M., Taly, A., & Yan, Q. (2017). *arXiv preprint arXiv:1703.01365.*

APPLICATION FOR SCHIZOPHRENIA

- Dataset:
  - RNA-seq gene expression & genotype data (dosage) for 487 schizophrenia (scz) vs. 891 non-scz human brain samples (front cortex)
  - Embedding GTEx eQTLs & PsychENCODE GRN for human brain front cortex
  - → 127304 SNPs, 2598 genes



PRIORITIZED GENE FUNCTIONS & REGULATORY LINKS FOR SCHIZOPHRENIA

- A list of enriched functions (FDR<0.05) from prioritized genes:
  - neuron development
  - axon guidance
  - cell adhesion
  - calcium signaling
  - response to external stimulus
  - NMDA receptor
  - insulin secretion
- Prioritized SNP-gene pairs
  - SNP-gene pairs on the interacting enhancers and promoters (Hi-C) have significantly higher importance scores ($p<5e-5$)
  - Potential regulatory roles of prioritized SNPs to genes via enhancers

FUTURE WORK

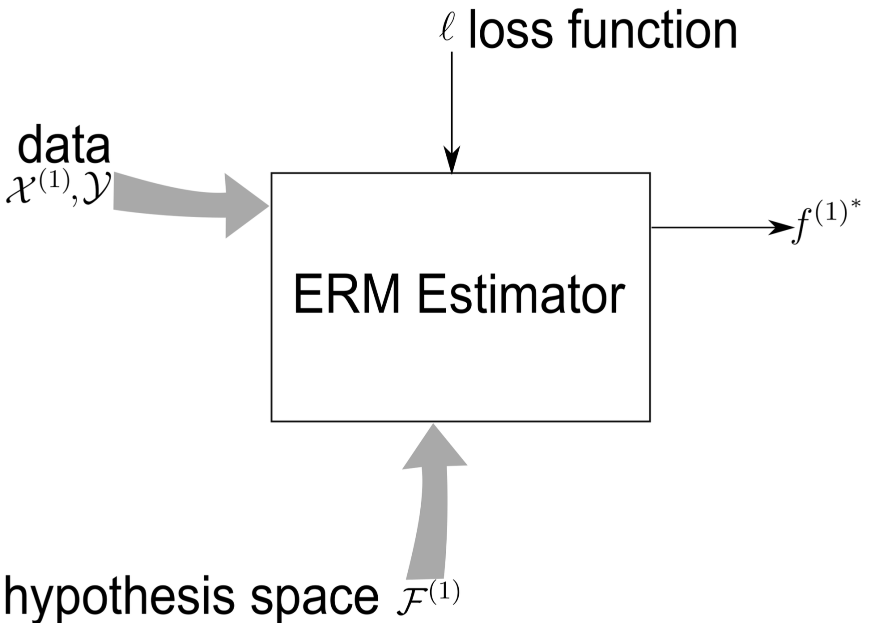
Single cell data integration	Cell-type gene regulatory networks
Additional omics	Epigenomics (e.g., ATAC-seq)
Deeper phenotypes	Imaging, behavior

THANK YOU!

This work has been supported by NIH grants R01AG067025, R21CA237955 and U01MH116492.

Q&A

EMPIRICAL RISK MINIMIZATION (ERM) FOR SINGLE-VIEW LEARNING


ℓ loss function
data
$\mathcal{X}^{(1)}, \mathcal{Y}$
ERM Estimator
$f^{(1)^*}$
hypothesis space $\mathcal{F}^{(1)}$

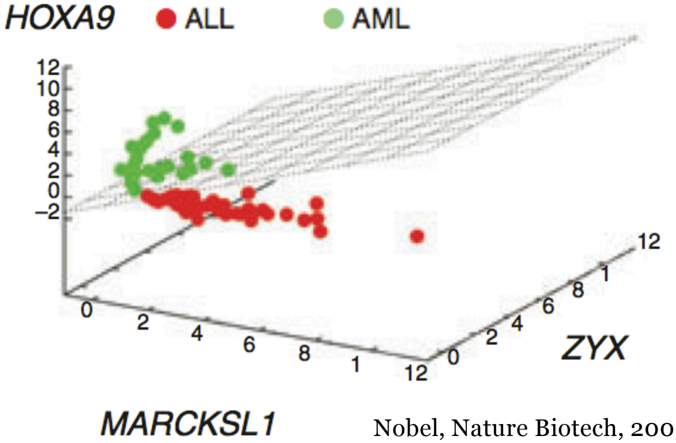
$$R(f) = \frac{1}{|S|}\sum_{(x_i,y_i)\in S} \ell(f(x_i), y_i)$$

$$f^* \in \underset{f}{\operatorname{argmin}}\{R(f) + \lambda\Omega(f)\}$$

Regularize f by biological knowledge Ω (e.g., rules)

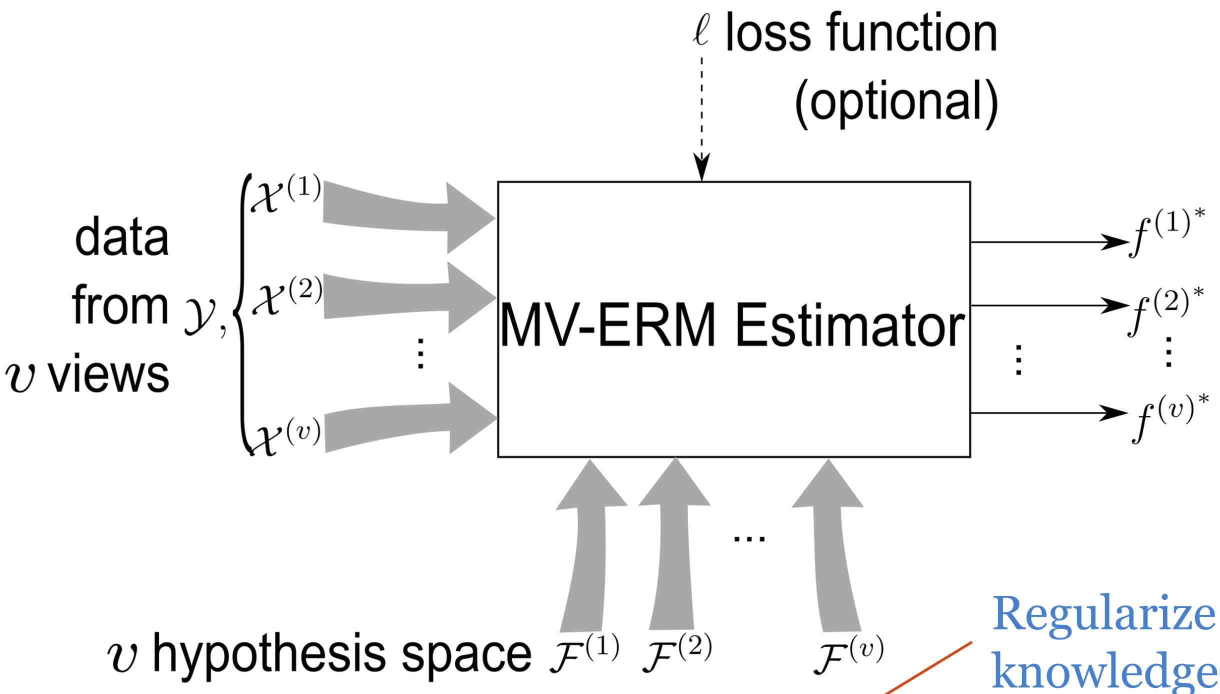
Nguyen, Wang, PLoS Computational Biology, 2020

- e.g., Leukemia patient classification
  - y_i: Acute lymphoblastic leukemia (ALL) vs. Acute myeloid leukemia (AML)
  - x_i: gene expression
  - f : SVM (with l is a hinge loss)



Nobel, Nature Biotech, 2006

EMPIRICAL RISK MINIMIZATION FOR MULTI-VIEW LEARNING (MV-ERM)



ℓ loss function (optional)
data from $\mathcal{Y}$, v views
$\mathcal{X}^{(1)}$
$\mathcal{X}^{(2)}$
⋮
$\mathcal{X}^{(v)}$
MV-ERM Estimator
$f^{(1)^*}$
$f^{(2)^*}$
⋮
$f^{(v)^*}$
v hypothesis space $\mathcal{F}^{(1)}$ $\mathcal{F}^{(2)}$ … $\mathcal{F}^{(v)}$

Regularize *f* by biological knowledge Ω from single omics

Regularize *f* by biological knowledge Ω_{co} across multi-omics

$$f^{(1)^*}, f^{(2)^*}, \cdots \in \underset{f^{(i)}\in\mathcal{F}^{(i)}}{\operatorname{argmin}} \left\{ \underbrace{\sum_i \ell\left(f^{(i)}\left(\mathcal{X}^{(i)}\right), \mathcal{Y}\right)}_{\text{optional}} + \underbrace{\lambda \sum_i \Omega\left(f^{(i)}\right)}_{\text{complementary}} + \underbrace{\lambda_{co}\sum_{i,j}\Omega_{co}\left(f^{(i)}, f^{(j)}\right)}_{\text{consensus}} \right\}$$

Nguyen, Wang, PLoS Computational Biology, 2020