Disease State Prediction from Single-cell Data Using Graph Attention Networks

Neal G. Ravindra, Ph.D.
<neal.ravindra@yale.edu, https://github.com/nealgravindra>

Joint with Arijit Sehanobish*,
Jenna L. Pappalardo, David A. Hafler,
David van Dijk
*Equal contribution

ACM CHIL 2020
July 23-24

Yale University

Outline

1. Background
2. Experiments
3. Model interpretation
4. Implications

Single-cell transcriptomics

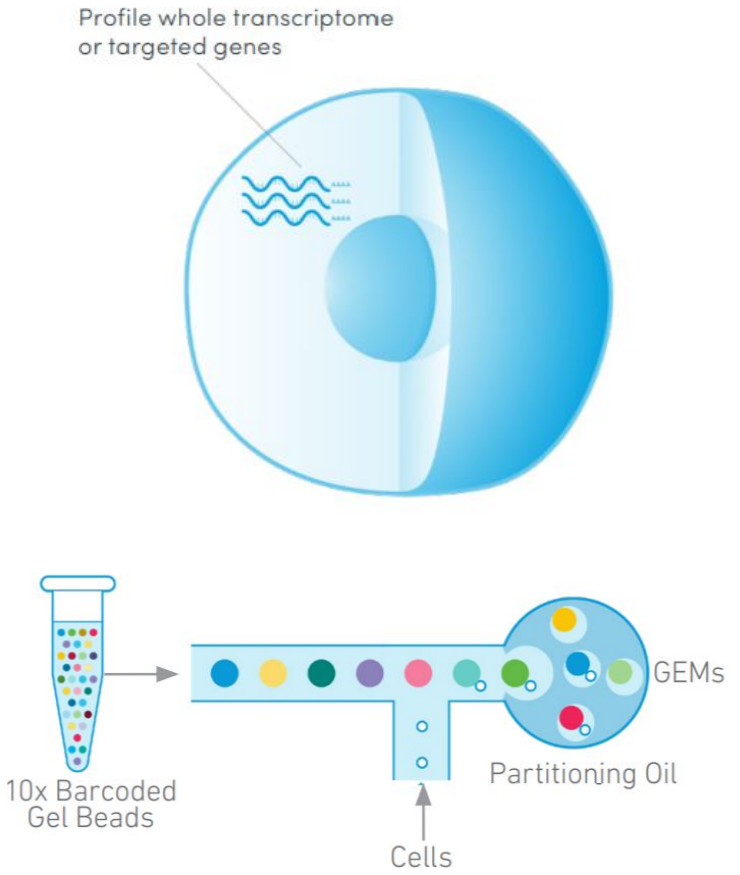
Gene expression of individual cells (scRNA-seq)

Broad-use across biomedical disciplines

Data is noisy, sparse, heterogeneous, and multidimensional

ML used for clustering, denoising, and imputation

Standard pre-processing yields large, sparse graphs



From 10x Genomics User Guide

Graph Neural Networks

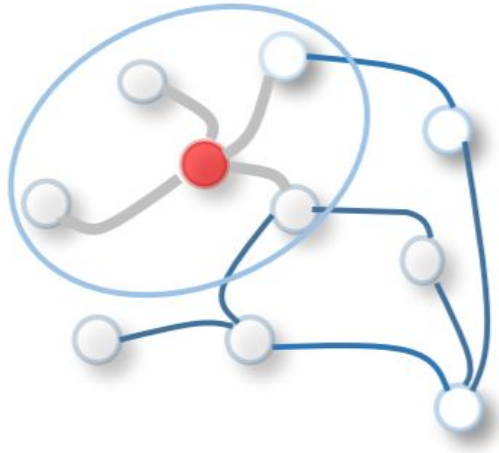
Used for tasks such as node classification, link prediction, and graph classification

Several models and research into understanding the expressive power of GNNs

Have been applied to EHR data, predicting protein-protein interactions, molecular structure & reactivity

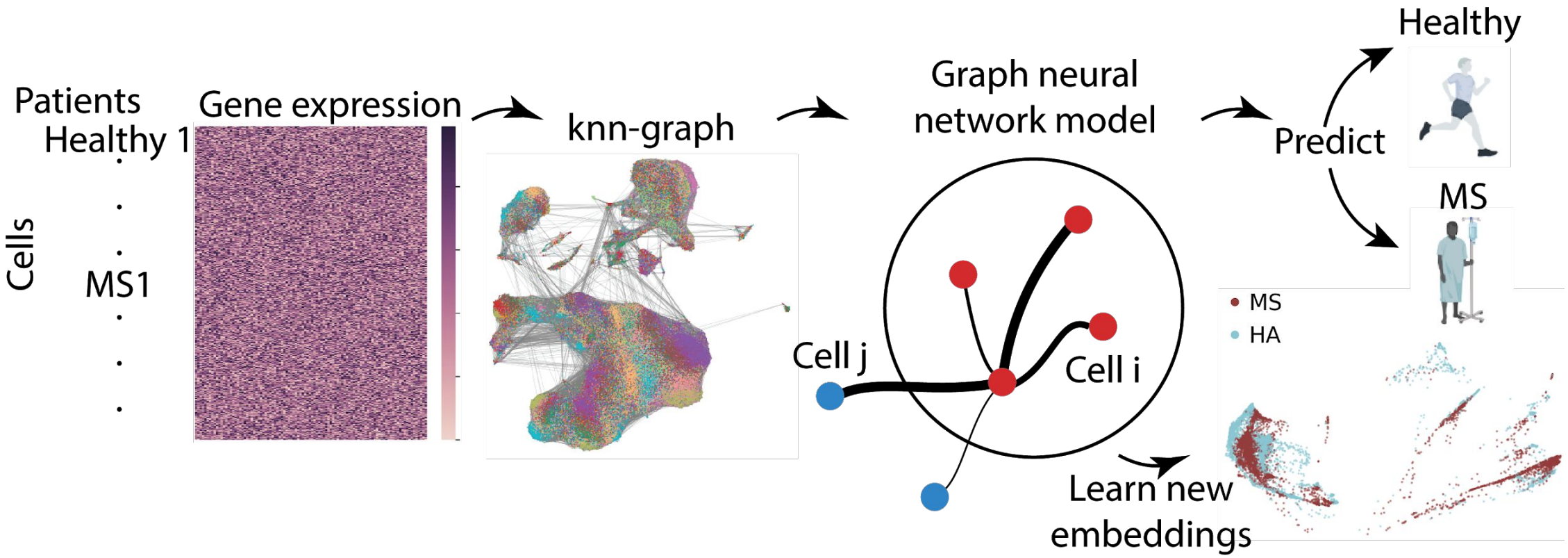
No prior work prior predicting disease states from single-cell data

Graph convolution to get hidden representation of red node



From Wu et al. survey, 2019

From count matrix to cell-by-cell disease state prediction

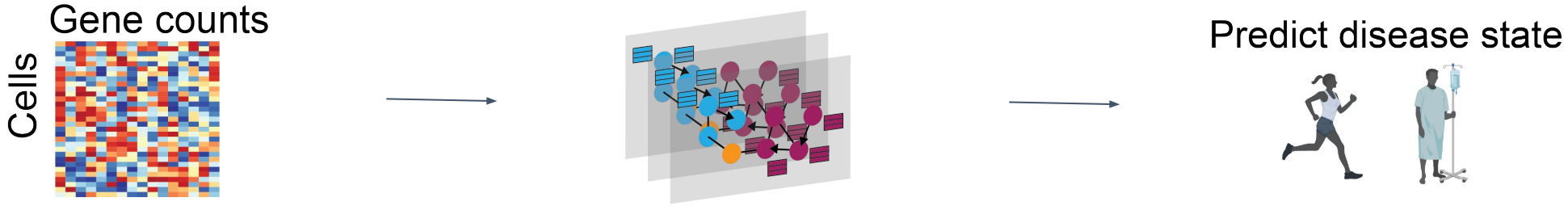


Our contributions

Use graph neural networks to predict disease state (healthy or MS) of each, individual cell from patient samples

Use an attention mechanism to enhance interpretability and learn new graphs to identify disease state clusters and “phenotypic” cells

Extract genes or identify novel biomarkers that are important to predicting disease state, given excellent predictive performance

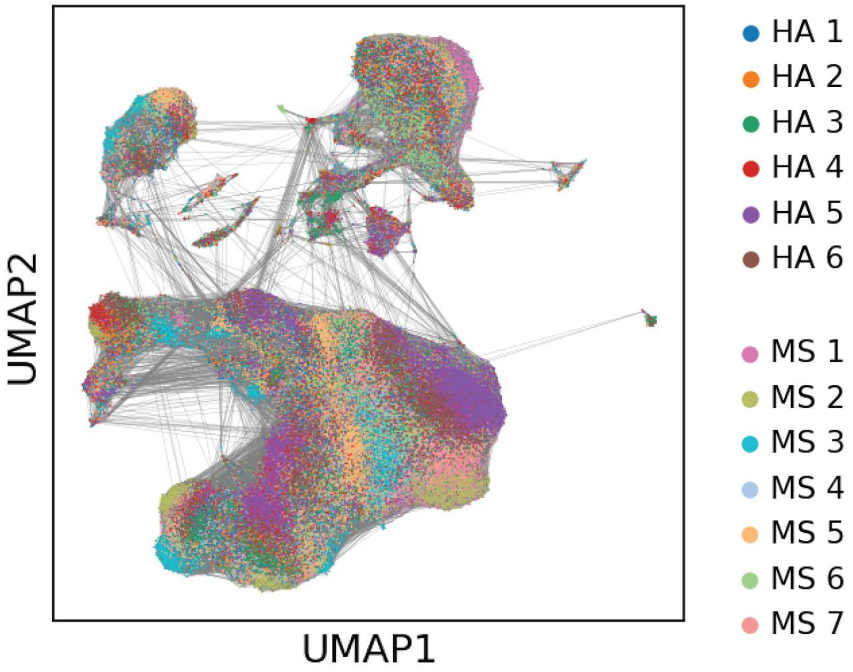


Graphs from scRNA-seq data

Multiple sclerosis (MS) is a disease of the nervous system, can be relapsing-remitting

Blood and cerebrospinal fluid samples from 7 MS patients and 6 healthy adults

Task		Train	Dev	Test
Inductive	# Nodes	43866	9686	13033
	# Edges	332398	73552	100715
	# Features	22005	22005	22005
	# Classes	2	2	2
	# Graphs	1	1	1
Transductive	# Nodes	54000	6000	6667
	# Features	22005	22005	22005
	# Classes	2	2	2
	# Edges	5007093		



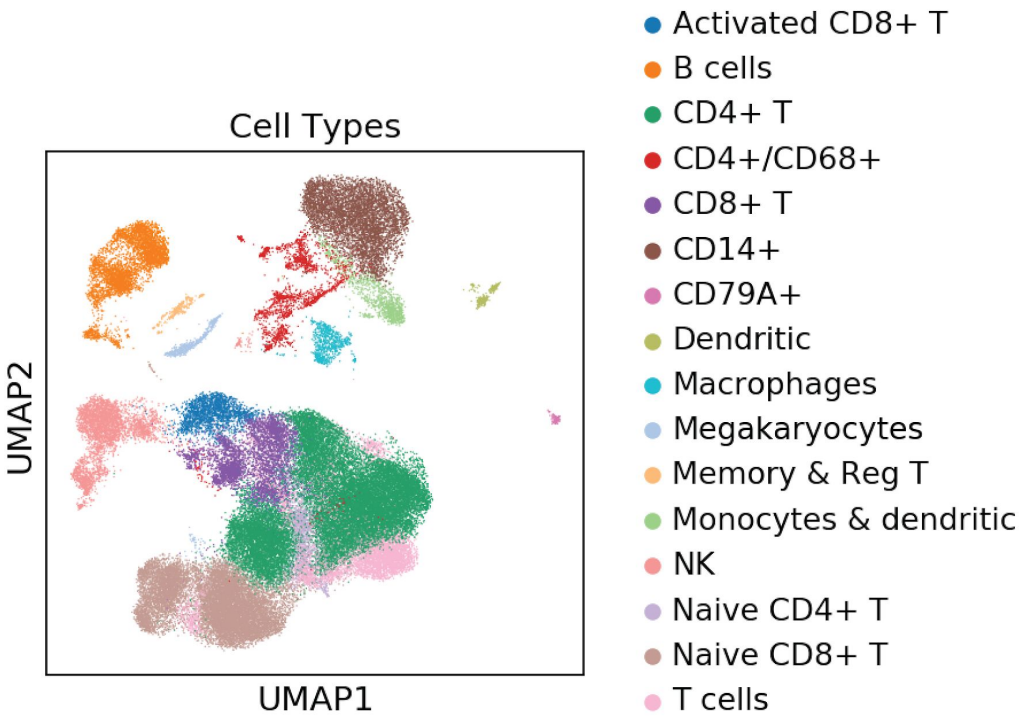
Cell type annotation for biological insight

"Cell type" (phenotypically distinct classes)

- Clustering a kNN graph with Louvain clustering for community detection

Value of bulk vs. single-cell RNA-seq

- Cell type v. cell-to-cell variability becomes resolved



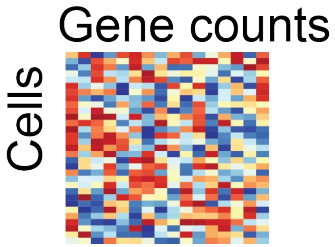
Our Graph Attention Network

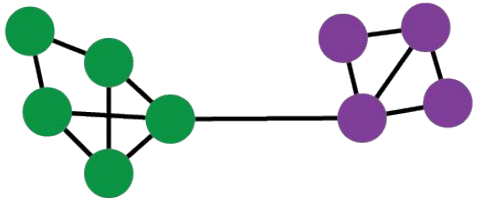
$$h = \{h_1, h_2, \ldots, h_N\}$$

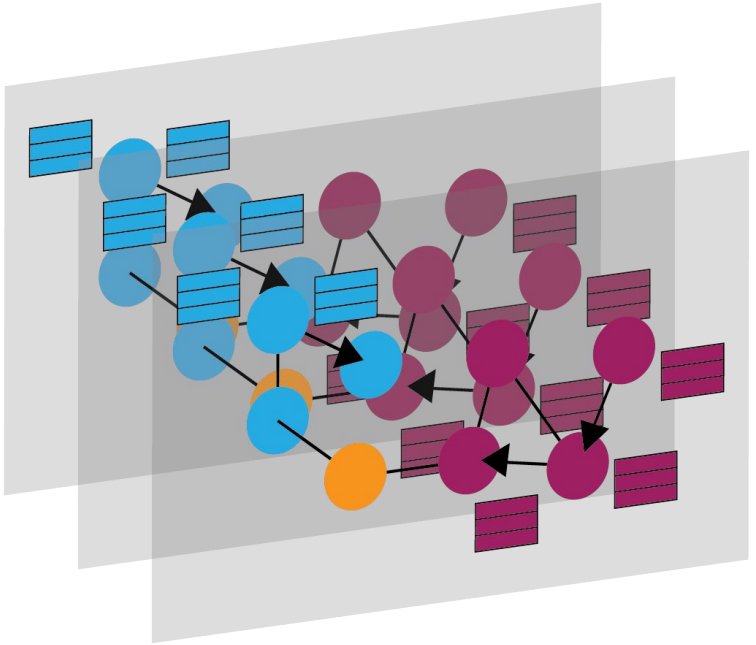
GAT layer

$$h'_i = \Big\|_{l=1}^{K} \sigma\Big(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^{l} \mathbb{W}^{l} h_j \Big)$$

$$h' = \{h'_1, h'_2, \ldots, h'_N\}$$



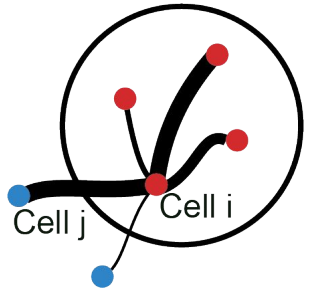




Message passing GNNs

Compute node self-attention to get multi-head edge attention coefficients over first-order neighborhood

$$\alpha_{ij} = \mathrm{softmax}_j(a(\mathbb{W}h_i, \mathbb{W}h_j))$$



Final layer averages over heads

Kipf & Welling, *ICLR* 2017 & Veličković et al., *ICLR* 2018

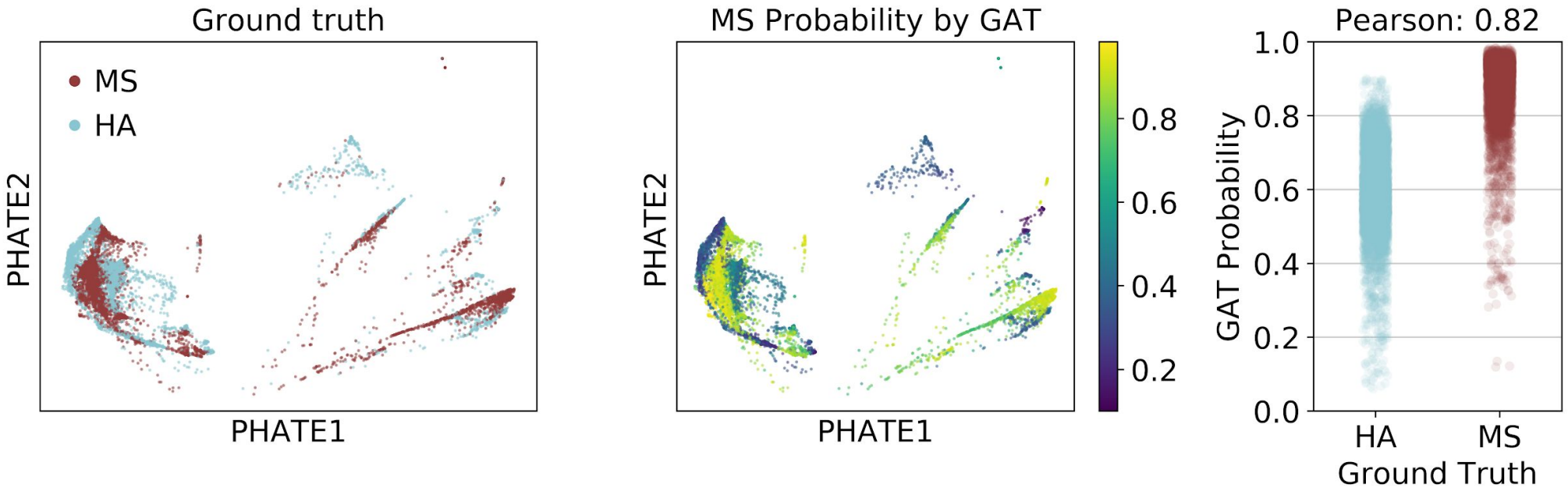
Large performance improvement with GATs

Compared our approach with other common ML models

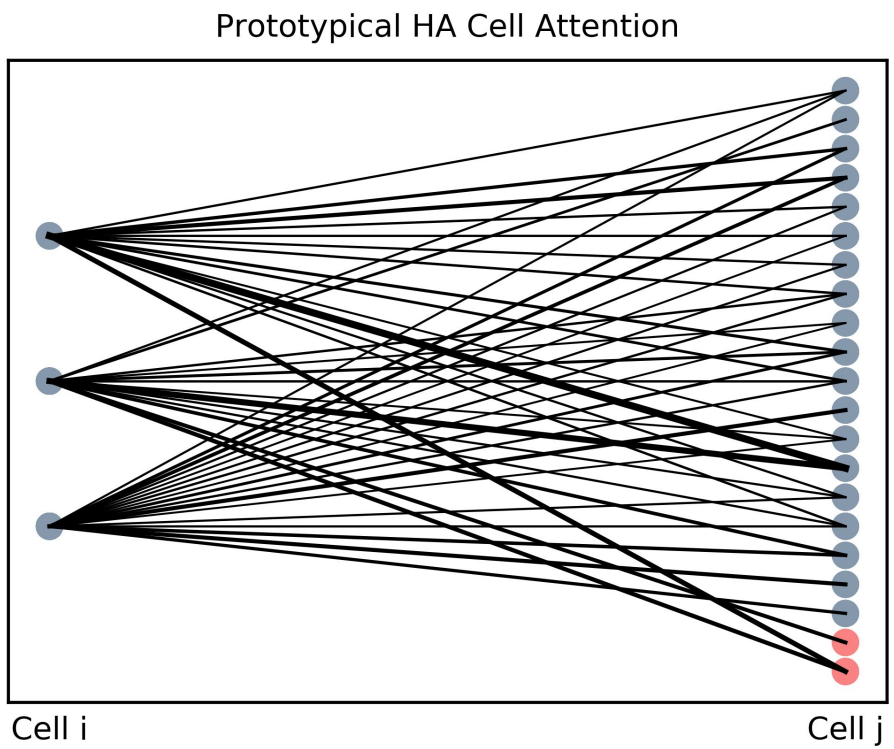
Compared our approach with another popular GNN

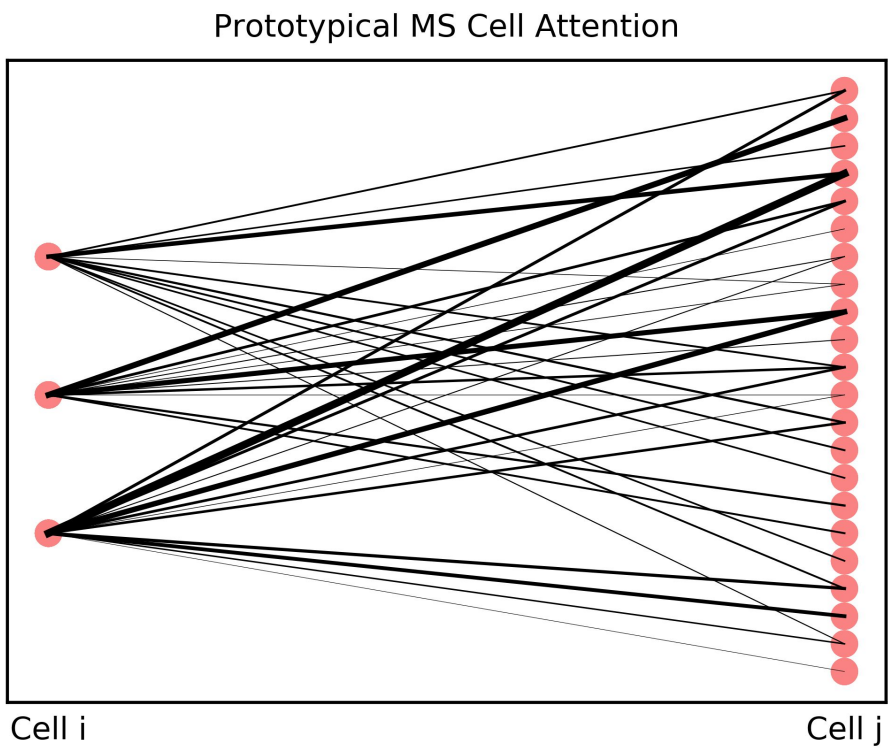
Task	Model	Accuracy
Inductive	Random	51.8
	MLP	56.7
	Random Forest	58.5
	Graph Convolutional Network	72.1
	Graph Attention Network(our)	**92.3** $\pm$ .7
Transductive	Graph Convolutional Network	82.91
	Graph Attention Network(our)	**86** $\pm$ .3

Model has good discriminability despite overlapping cell type

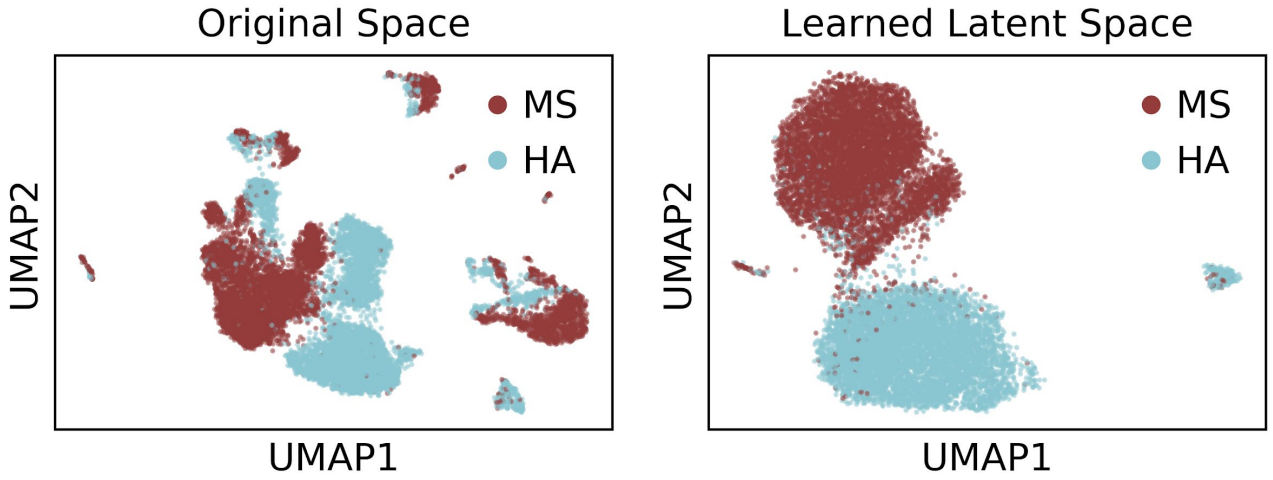


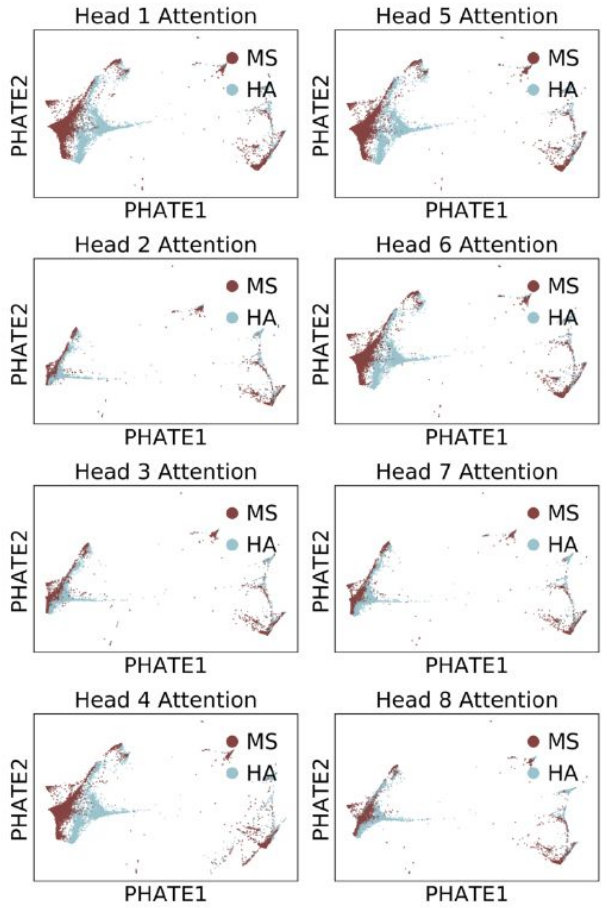
Classic attention diagrams don’t provide clear insights





New embeddings after learning new latent space and graphs



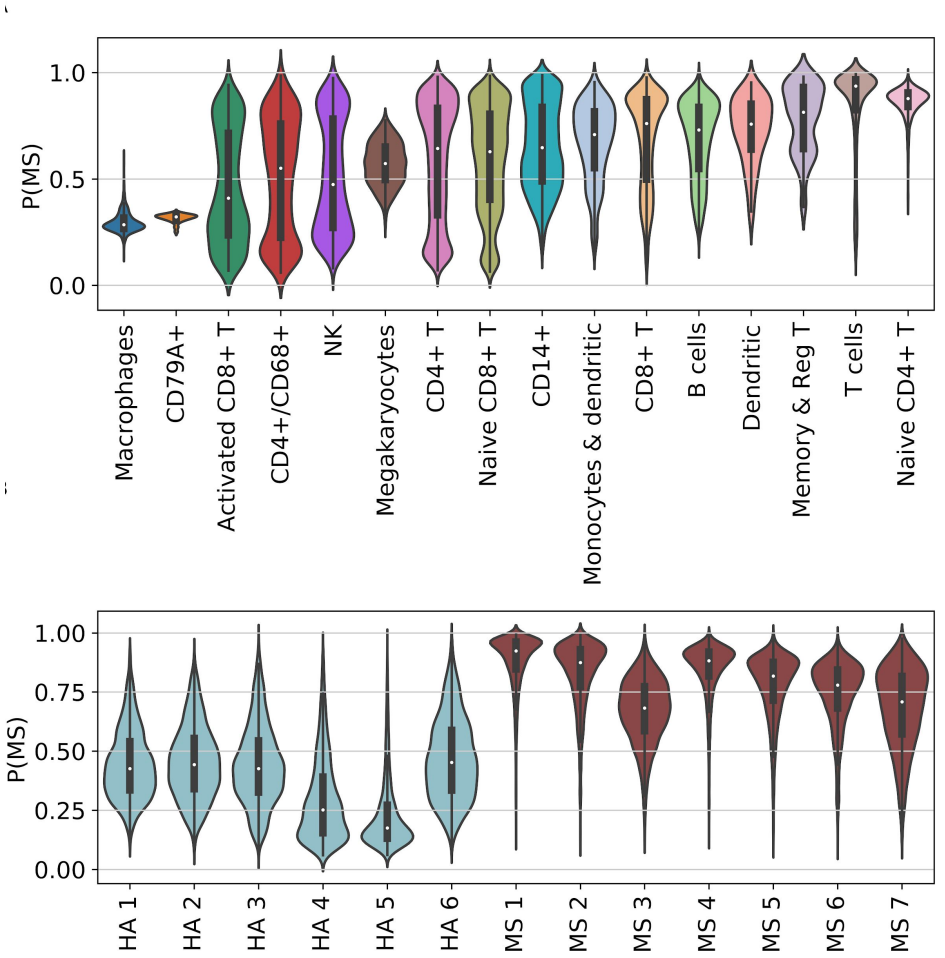


- May lead to identification of new "phenotypic" cells or unique disease state clusters

Contribution of cell types to MS and cells to patient's disease

Predicted probabilities from transductive task highlights important cell types for predicting disease state

Variance of cells' probability of MS across patient samples may indicate subtle disease states



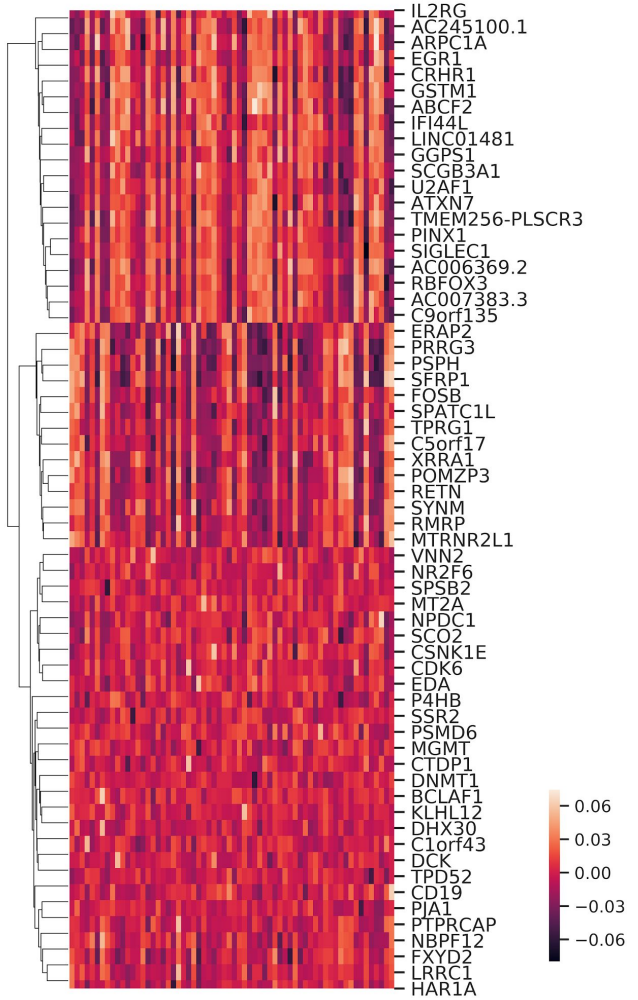
Genes that are important for predicting disease state

Per *k* head, $g_i^k = \max_j(|w_{ij}|)$

Interleukin-2 receptor subunit among top 10 predictive features per head

Marker for therapeutically targeted B cells (CD19) also among top features

Top predictive features regulate hormone secretion, nerve cell development, and lipid metabolism, suggesting relevant but novel hits

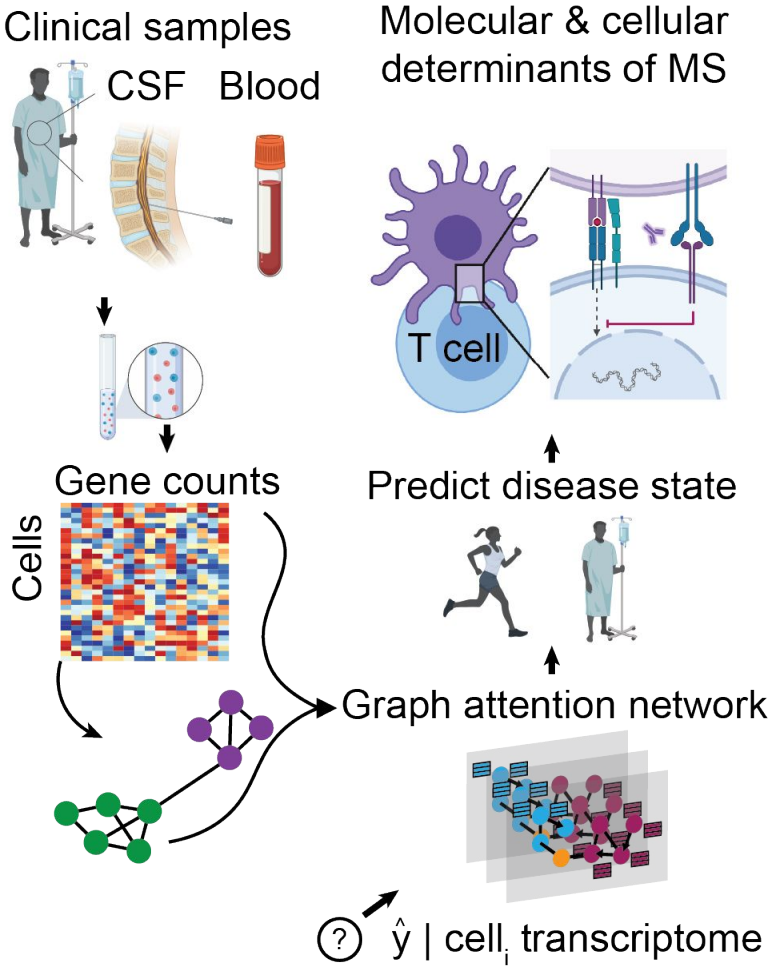


Disease state prediction from scRNA-seq data using GATs

For transductive & inductive tasks, achieve 86% and 92% accuracy in predicting whether a single cell is healthy or exhibits MS

Better performance compared to other ML approaches, which allows for interpretability of transcriptomic and genetic signatures of disease state

We identify subsets of cells using learned feature space and attention-weighted graphs, which may be of interest for further study



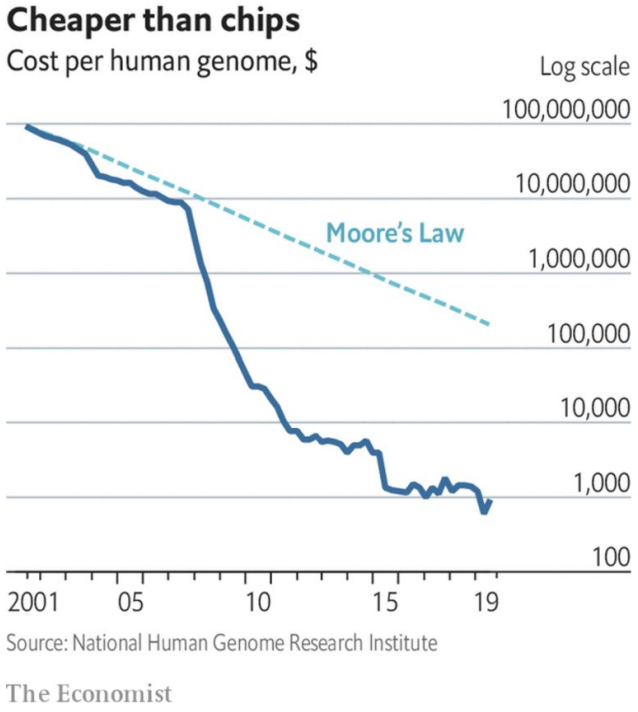
DL for insight into pathophysiology: disease state prediction

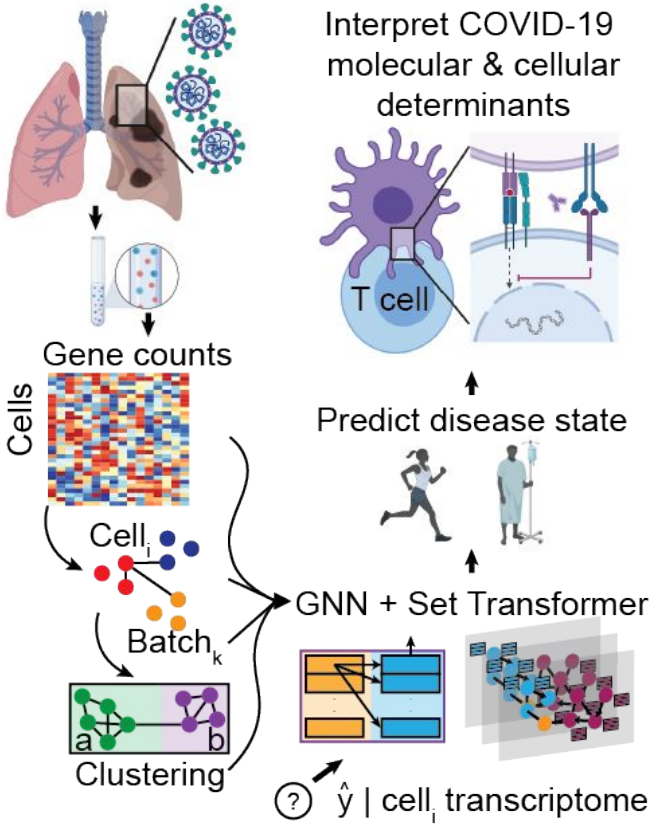
Interpretability of genes important for determining disease state

Visualization and identification of "phenotypic" populations in tissue

Transfer learning, esp. with CSF

Requiring fewer and fewer cells or features for diagnosis, prognosis, or treatment personalization





Sehanobish*, Ravindra*, & van Dijk, *ICML* 2020

Acknowledgments

van Dijk Lab @ Yale
*Arijit Sehanobish
Antonio Fonesca
Jason Bishai
Victor Gasque
Aagam Shah

Hafler Lab @ Yale
Jenna Pappalardo

Funding
NIH T32HL007950

Original paper

https://dl.acm.org/doi/10.1145/3368555.3384449

Follow-up work

https://grlplus.github.io/papers/69.pdf

https://arxiv.org/abs/2007.04777

For code/questions visit
vandijklab.org
github.com/vandijklab
github.com/nealgravindra

Or email
neal.ravindra@yale.edu
arijit.sehanobish@yale.edu
david.vandijk@yale.edu