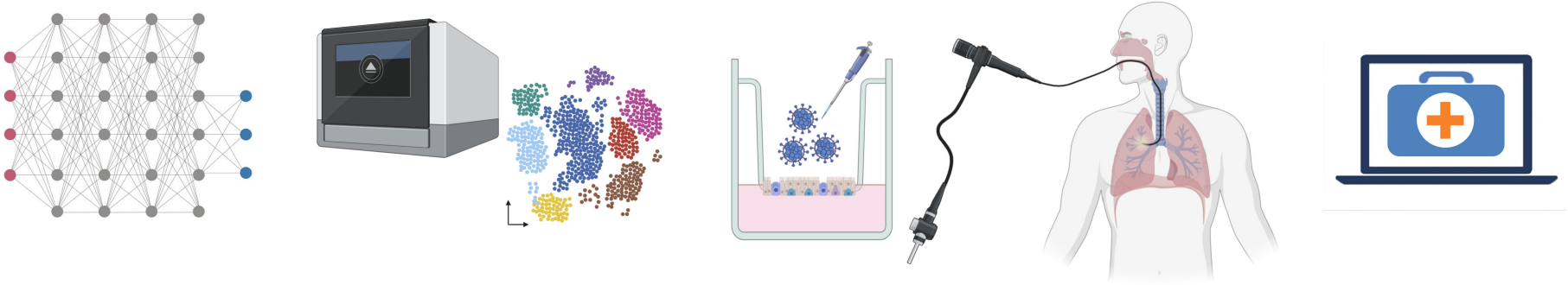
Leveraging geometric deep learning for knowledge discovery from single-cell data and explainable AI for clinical decision support tool development

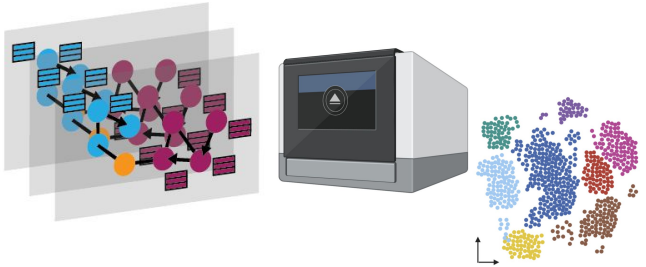


Neal G. Ravindra, Ph.D.
Machine Learning Postdoctoral Scholar at Stanford University
Dept. Of Biomedical Data Science, Anesthesiology, Pediatrics, & Stanford Artificial Intelligence Laboratory

Overview

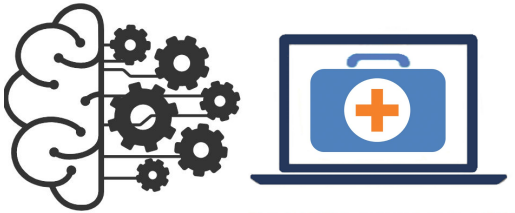
Interpretable ML to study molecular & cellular mechanisms of disease and cell state based on single-cell omics data

- Dynamical genes from landmark time-point data
- single-cell Graph Attention Networks (scGAT)



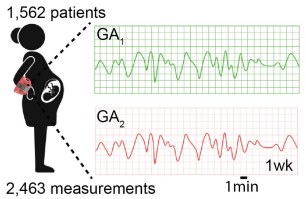
XAI to create clinically useful and parsimonious models

- qCSI from a COVID-19 Severity Index model for triaging patients in the emergency department



Translational research using relational reasoning and metric learning

- actigraphy2GA: sleep and activity disruptions and their relation to preterm birth
- sc2drug: perturbation modeling to align similar but disparate distributions

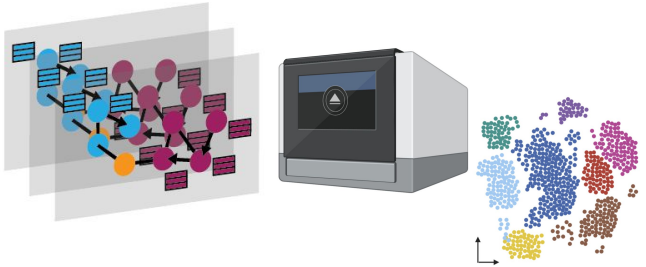


Overview

Interpretable ML to study molecular & cellular mechanisms of disease and cell state based on single-cell omics data

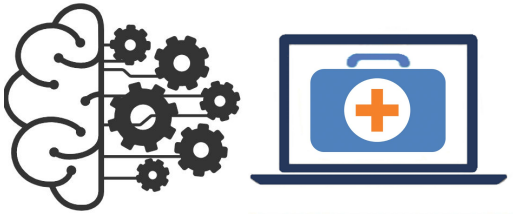
Dynamical genes from landmark time-point data

single-cell Graph Attention Networks (scGAT)



XAI to create clinically useful and parsimonious models

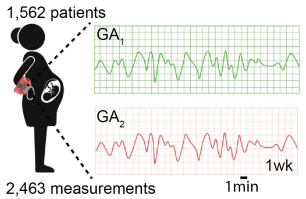
qCSI from a COVID-19 Severity Index model for triaging patients in the emergency department



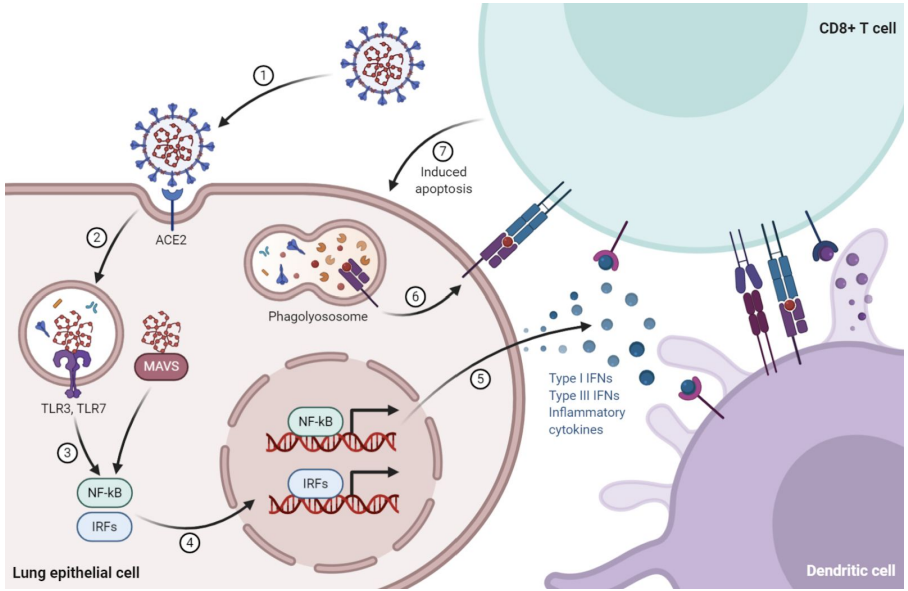
Translational research using relational reasoning and metric learning

actigraphy2GA: sleep and activity disruptions and their relation to preterm birth

sc2drug: perturbation modeling to align similar but disparate distributions



Immune response to SARS-CoV-2



Susceptibility:

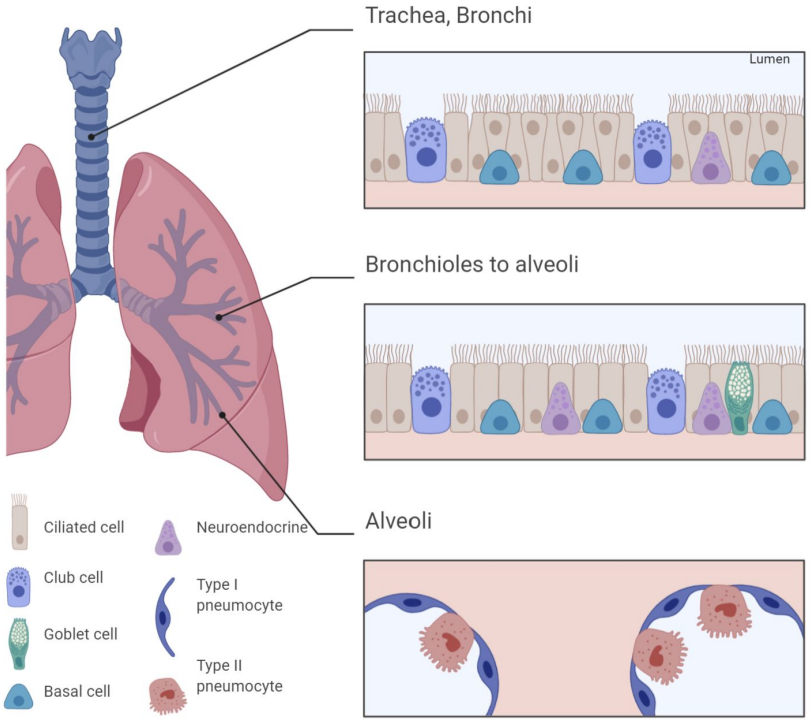
- Are there any transcriptional patterns that make a cell more likely to be infected?

Response:

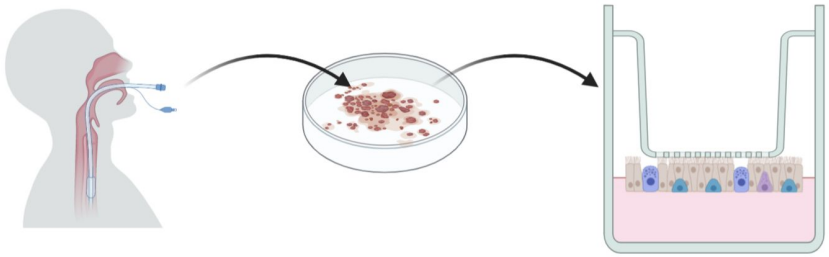
- How does the acute immune response vary across rare cell types?
- How does that evolve after controlled, initial exposure to the virus?

Problem because SARS-CoV-2 cell tropism and early dynamics of response were unknown

Human bronchial epithelium cell (HBEC) ogranoids



Air-liquid interface cultures (organoids)

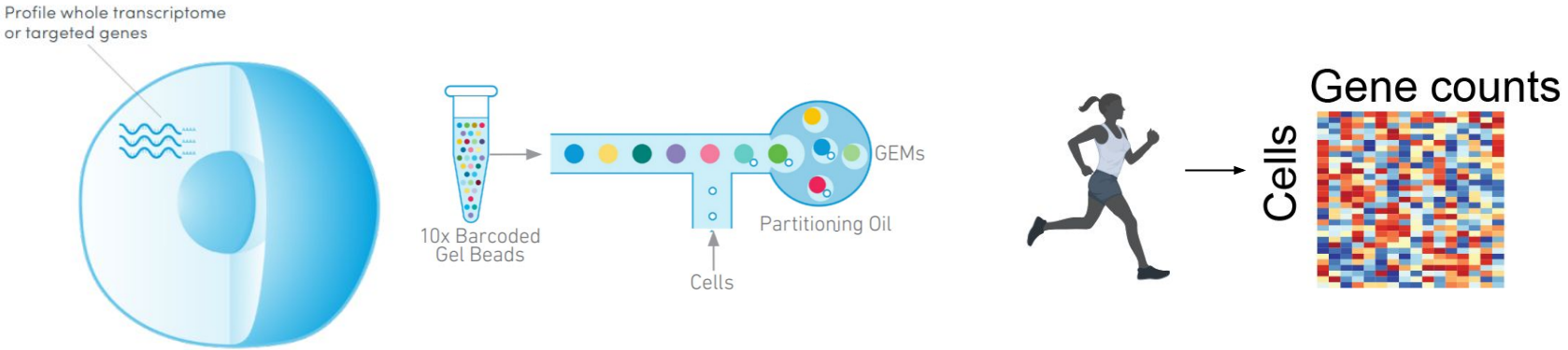


Availability of SARS-CoV-2 isolate + matured organoids → a timely solution:

- High-quality scRNA-seq data (many intact cells)
- Controlled time course

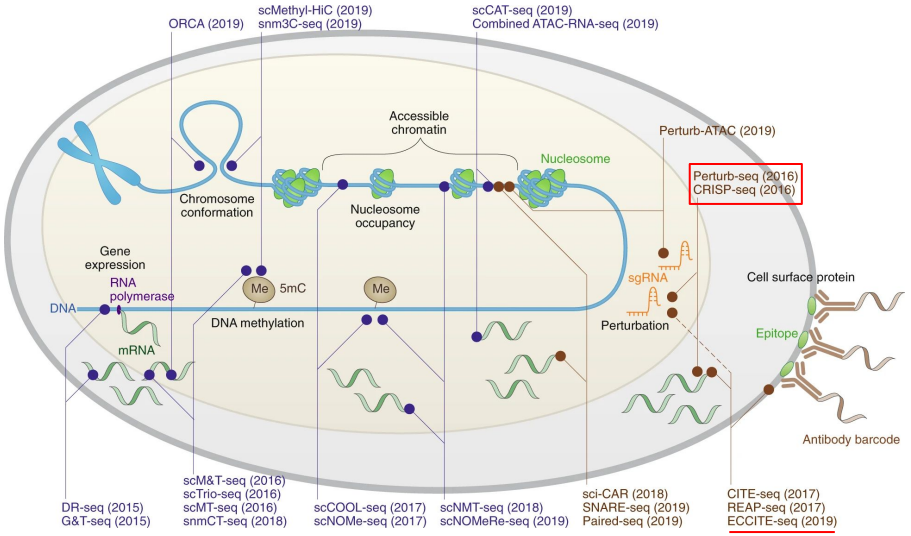
Single-cell omics

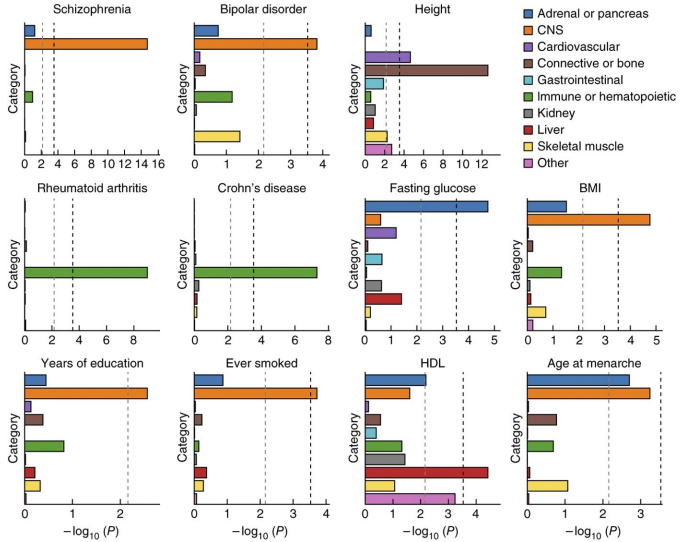
scRNA-seq: gene expression measurements for individual cells



Zheng et al. *Nat. Comm.*, 2017

Single-cell omics to study molecular pathophysiology

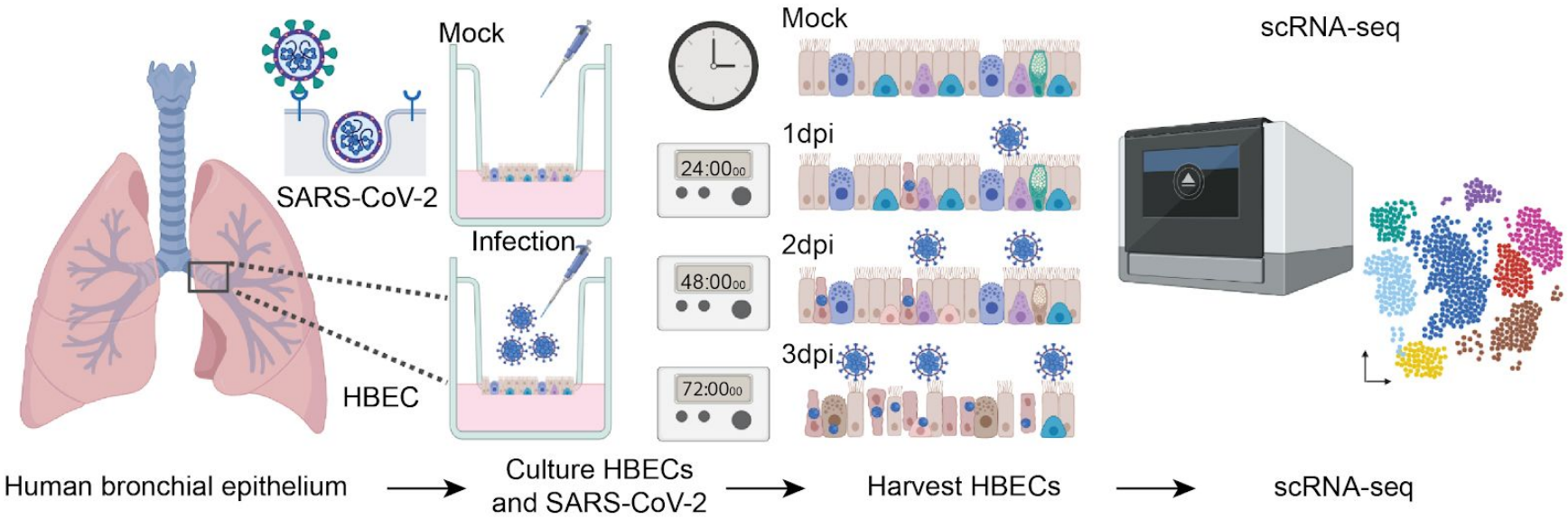




Need big data analysis/ML *and* even GWAS variants have diverse effects across cell types & map to noncoding sequences

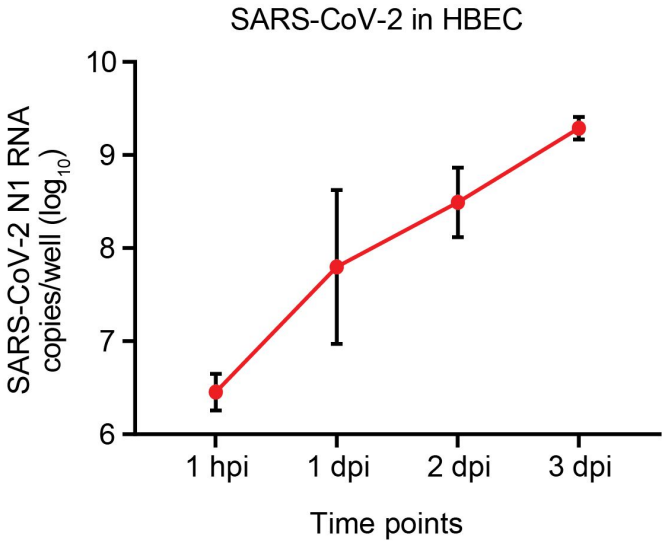
Zhu et al. *Nat. Meth.*, 2020; Finucane et al. *Nature Genetics*, 2015; Claussnitzer et al., Nature 2020; Bulik-Sullivan et al. Nat. Genet, 2015

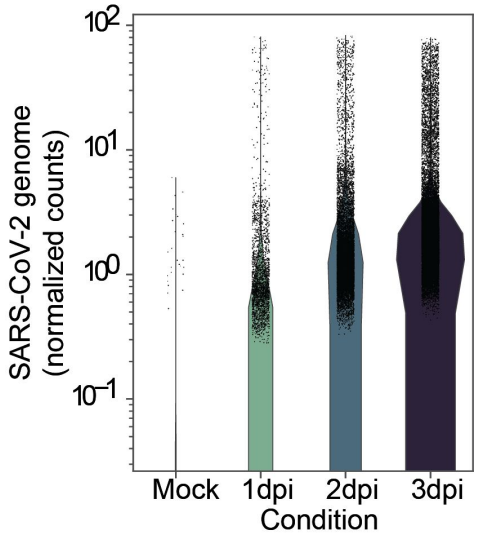
Longitudinal analysis of SARS-CoV-2 infection



Ravindra NG*, Alfajoro MM*, … van Dijk D, Wilen CB. *PLoS Biology*, 2021
Wei J, Alfajaro MM… **Ravindra NG**, … Wilen CB. *Cell*, 2021

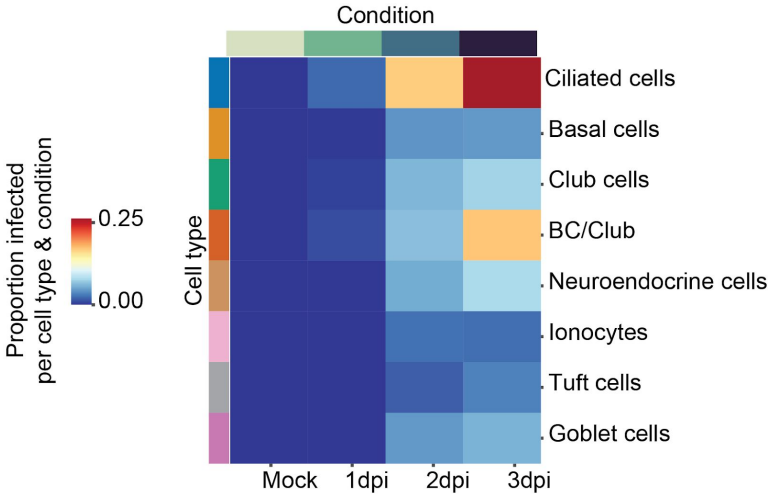
Bulk v. scRNA-seq counts of viral transcripts

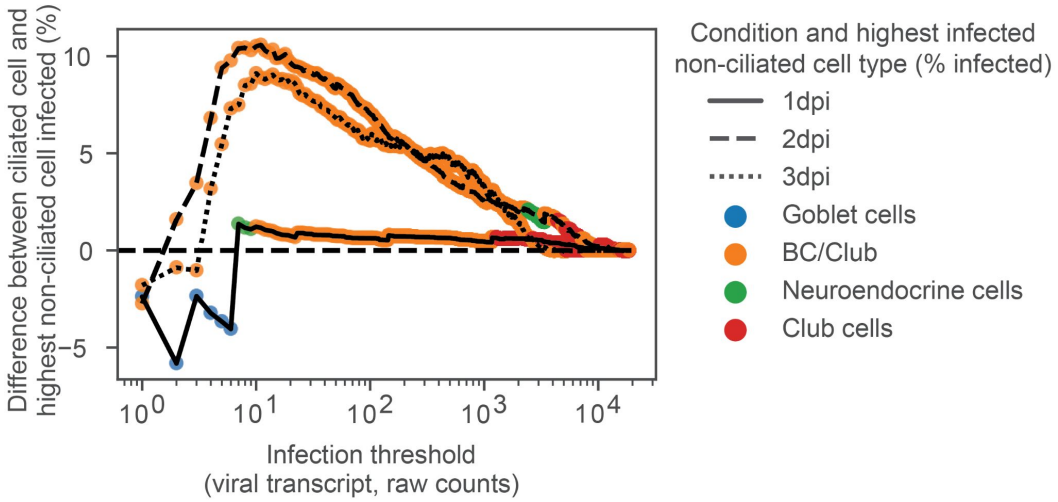




At a single-cell level, we can create a label denoting infection or bystander status

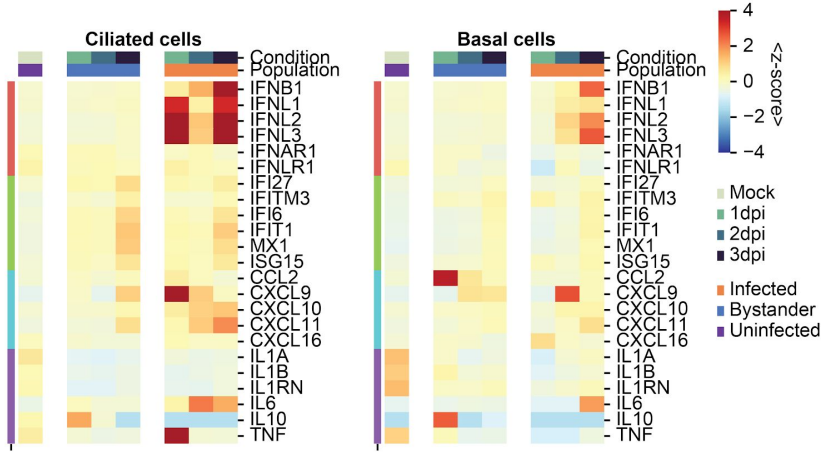
SARS-CoV-2 cell tropism in upper respiratory epithelium

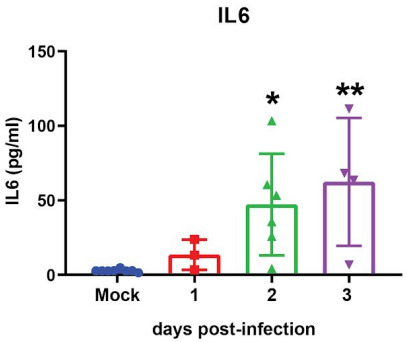


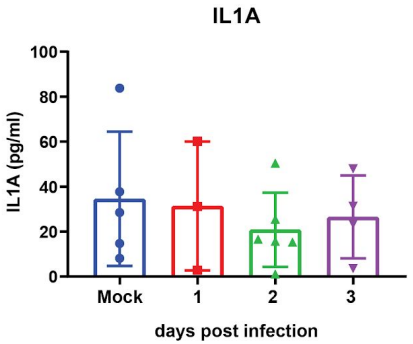


Ciliated cells are the major target of SARS-CoV-2 at the onset of infection and over the course infection, cell tropism expands to other epithelial cell types

Evolution of SARS-CoV-2 induced immune response

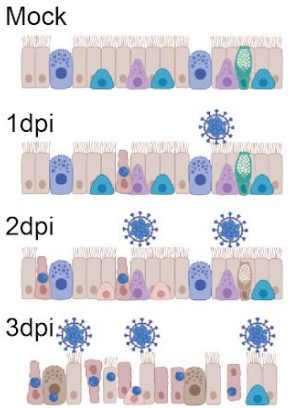


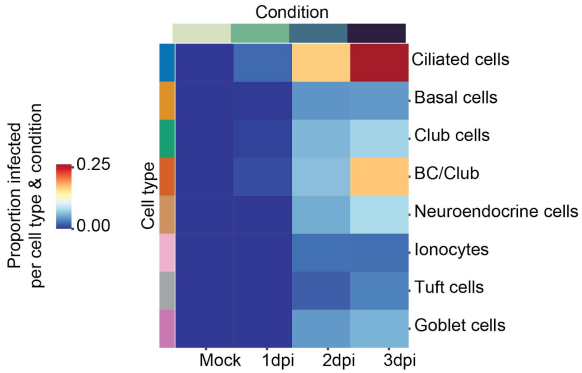


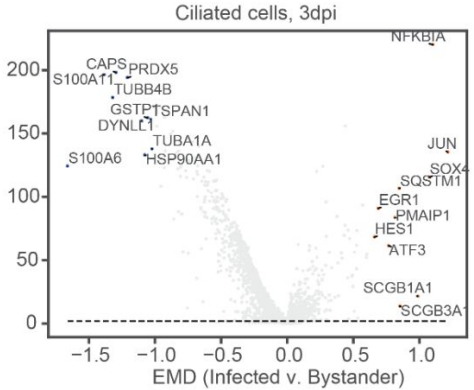


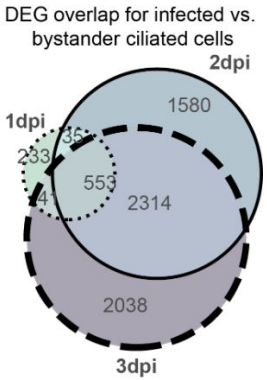
Infection induces cell-intrinsic expression of type I and III IFNs and IL-6 but not IL-1, resulting in expression of ISGs in both infected and bystander cells

Longitudinal scRNA-seq analysis of SARS-CoV-2 infection







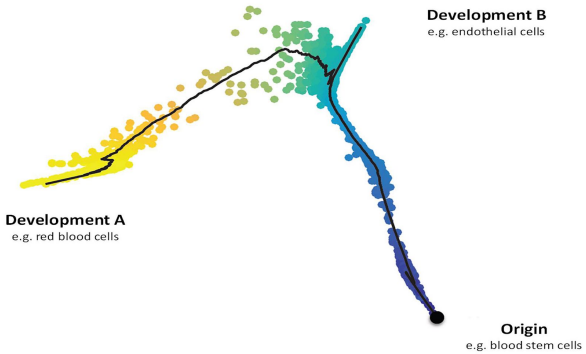


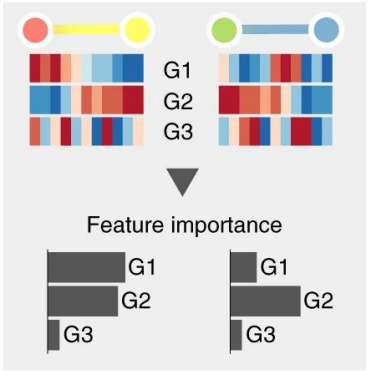
Identified:

- **ciliated cells as the primary initial target of SARS-CoV-2**
- **patterns of gene expression induced by the innate immune system and potential cell programs conferring resistance to infection**

Organoids dataset remains a useful resource

Cell “pseudotime” for un-labeled transitions and ordering





Hard to find *semantically meaningful* embeddings without supervision

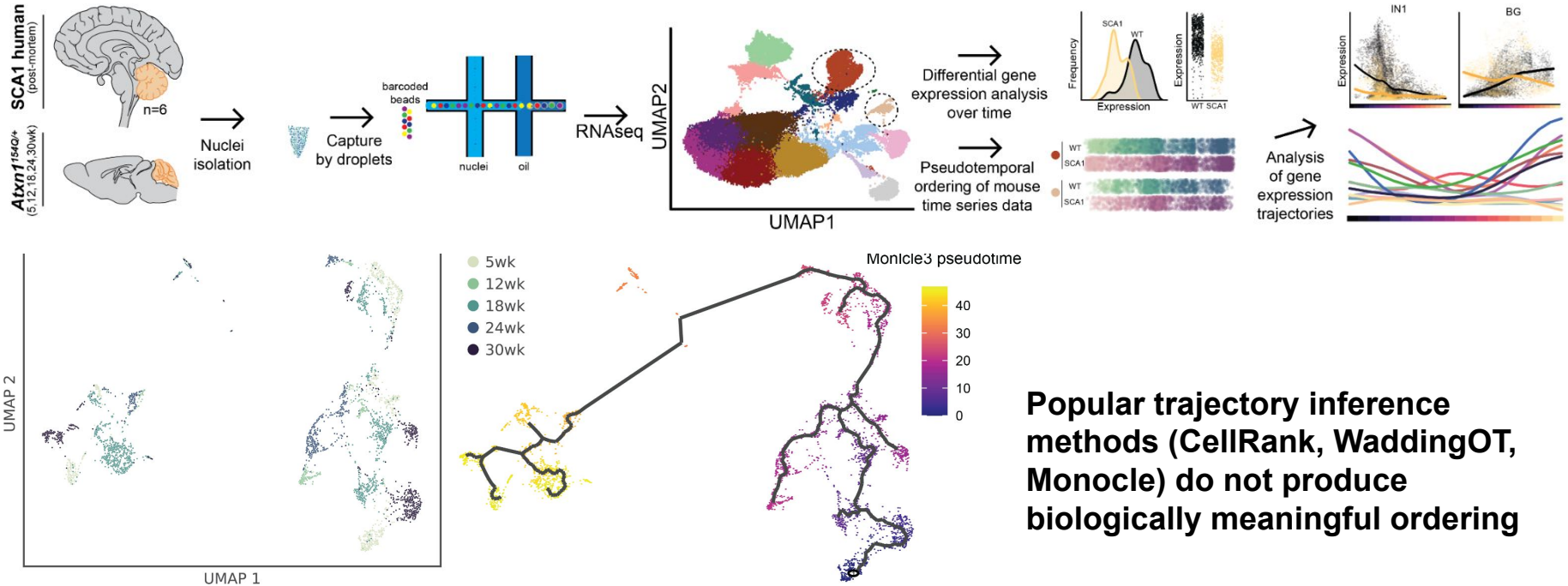
Manually select origin and end point, which is complicated for clusters with imperfect purity

Identifying factors that drive changes by correlation between gene expression and pseudotime

Problem because we want to study evolution of gene expression dynamics but often have landmark timepoints in controlled time-courses or model organisms at discrete developmental stages

Comparison of 45+ TI methods: Saelens et al. *Nat. Biotech.*, 2019

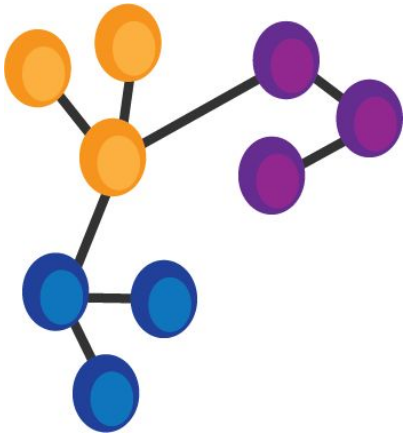
Longitudinal snRNA-seq analysis of neurodegeneration



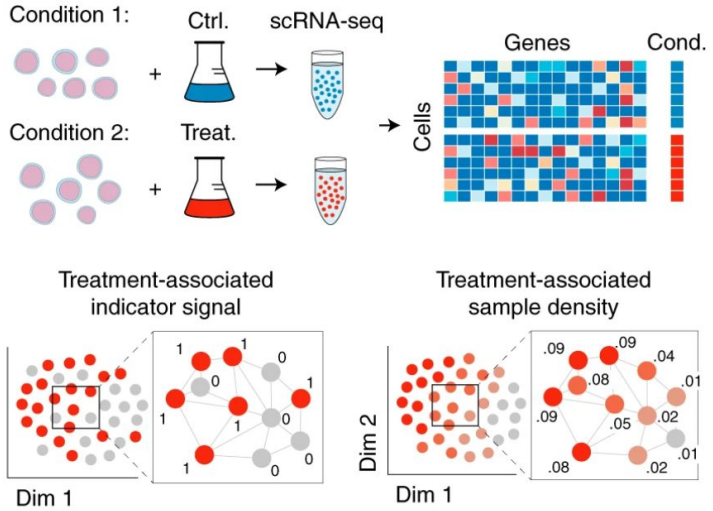
Popular trajectory inference methods (CellRank, WaddingOT, Monocle) do not produce biologically meaningful ordering

Tejwani L*, **Ravindra NG***, … van Dijk D, Lim J. *in revision* at *Cell*

Label smoothing based on cell-cell similarity



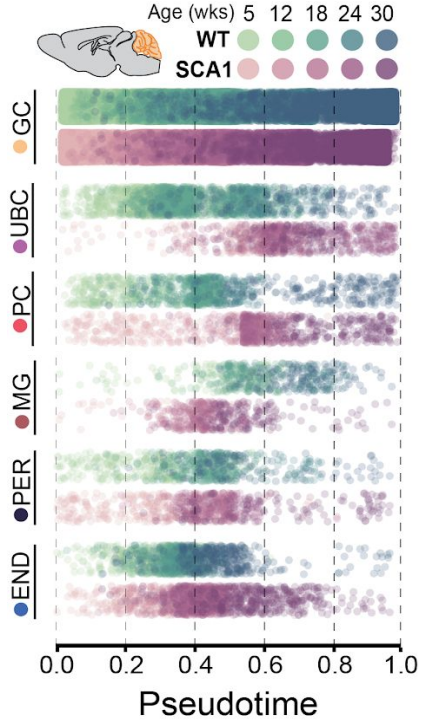
Cell graph
(cells are *nodes*,
edges have weights ~ distance in embedding)

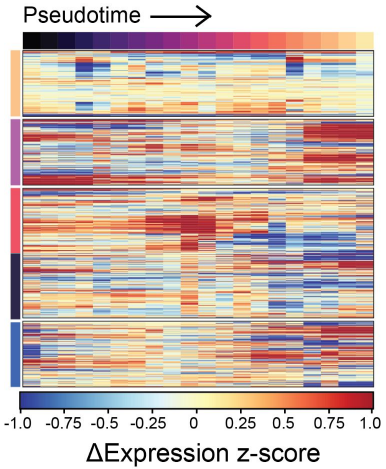


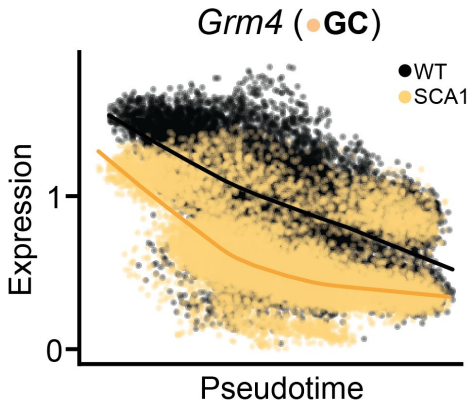
MELD: Burkhardt et al. *Nat. Biotech.*, 2021

Borrow from graph signal processing to diffuse known landmark timepoint according to cell-cell similarity, which may have impure modularity w.r.t. time

Label smoothing according to mouse developmental age

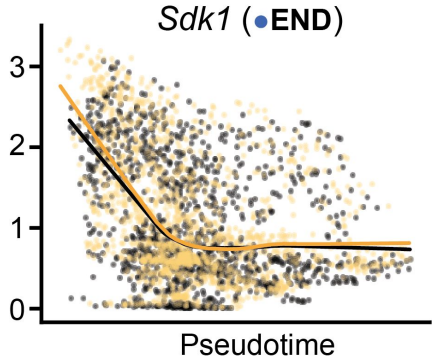


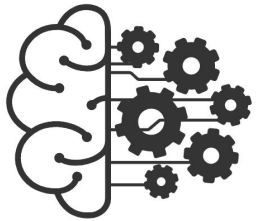




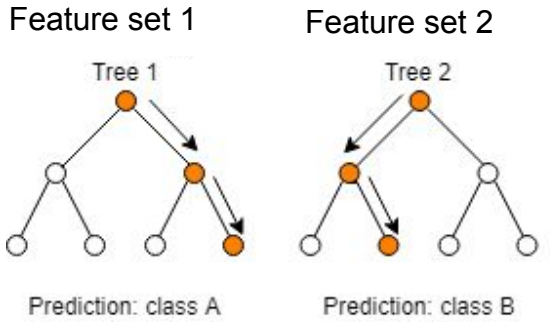
Label smoothing reduces confounding of discrete time-point groupings while respecting landmark time-point order

Regression model interpretability per pseudotime branch





Gradient boosting + MAGIC

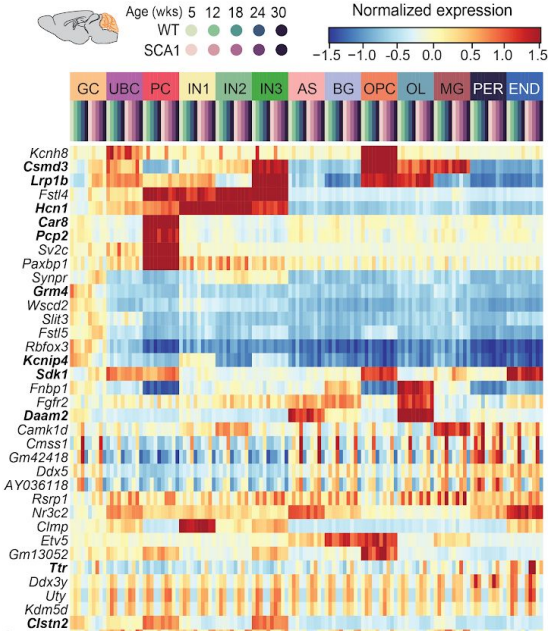


Importance ~ *gain* in accuracy when a feature is added to branch

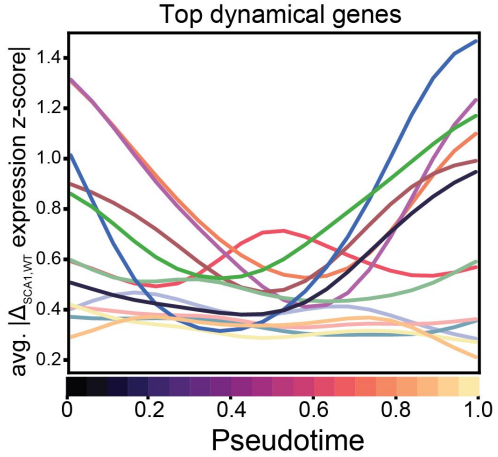
How to find genes with differential dynamics over progressive neurodegeration during mouse's life?

- Need an interpretable, non-linear model
- Fast because many features and sub-analyses by cell type

Gene and cell subsets driving neurodegeneration in mice



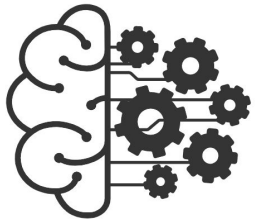
Aggregate by cell type to identify in which cell states have largest differential dynamics



Adapting graph signal processing for traditional single-cell analyses and combining with interpretable ML identifies drivers and signatures of progressive neurodegenerative decline

Finding interesting dynamics *without* supervision

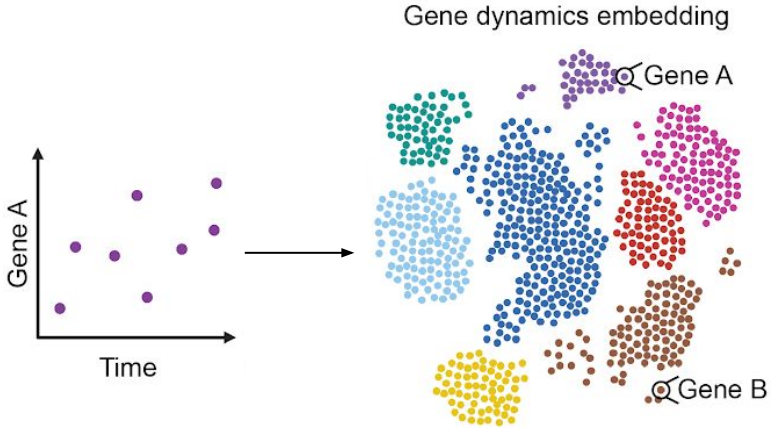
pseudotime ~ $\Delta_{c=1,c=2}$GEX



Gradient boosting + MAGIC

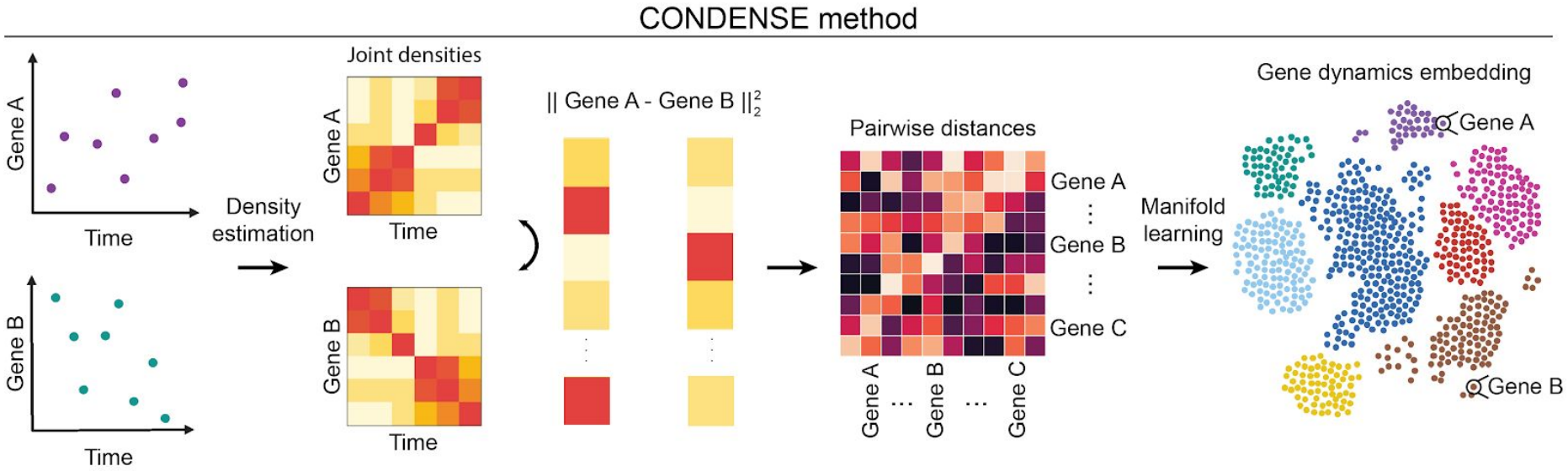
Face the problem of *missing dynamics* that are meaningful but too complex to model

Underlying pseudotime between conditions may not have *1-to-1 correspondence* as pseudotime is a rank order, adding bias



Want to cluster and phenotype genes by their dynamics

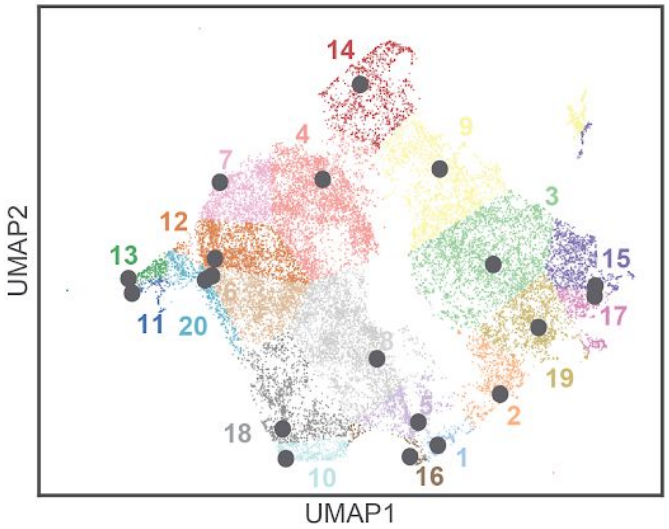
Unsupervised learning + label smoothing for dynamics



CONDENSE = CONditional DENSity Embedding

Ravindra NG*, Gasque V*, Alfajaro MM, Tejwani L, Lim J, Wilen CB, van Dijk D. (*in preparation*)

Archetypal analysis of gene dynamics embedding

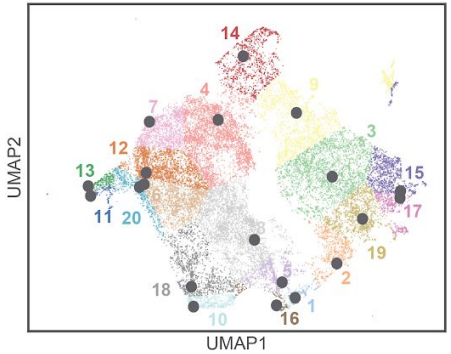


Gene dynamics vary across the HBEC data set

Archetypal analysis identifies *extremal* dynamics and min dist per gene assigns genes to artchetypes

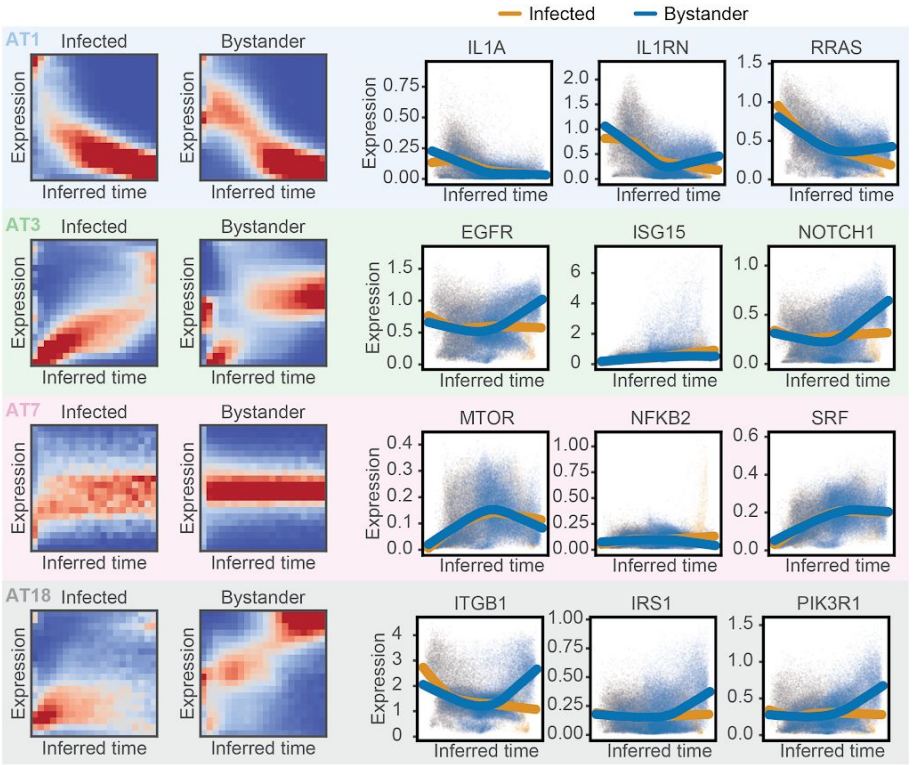
May have differential contribution to AT loading per condition

Separating archetype's conditional densities by group



Genes with signaling-related GO terms are suppressed in infected cells but up-regulated in bystander cells over time

CONDENSE identifies differential gene dynamics between infected and bystander cells' response to SARS-CoV-2

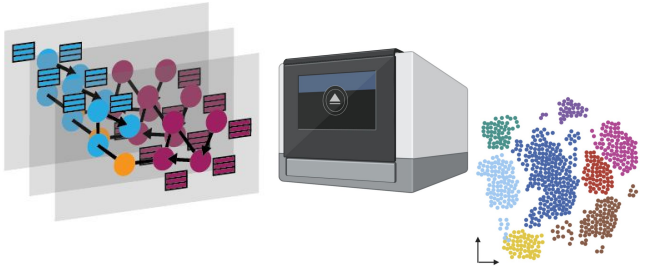


Overview

Interpretable ML to study molecular & cellular mechanisms of disease and cell state based on single-cell omics data

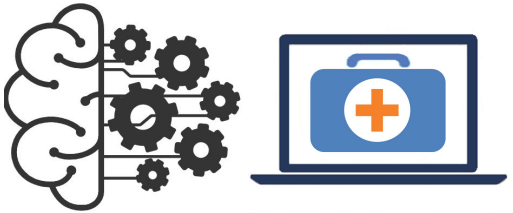
Dynamical genes from landmark time-point data

single-cell Graph Attention Networks (scGAT)



XAI to create clinically useful and parsimonious models

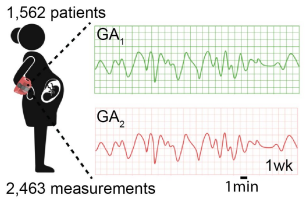
qCSI from a COVID-19 Severity Index model for triaging patients in the emergency department



Translational research using relational reasoning and metric learning

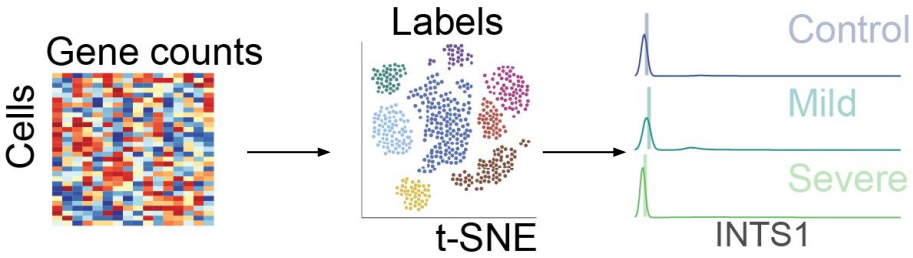
actigraphy2GA: sleep and activity disruptions and their relation to preterm birth

sc2drug: perturbation modeling to align similar but disparate distributions



Molecular mechanisms from single-cell data

scRNA-seq pipelines to study genetics and disease

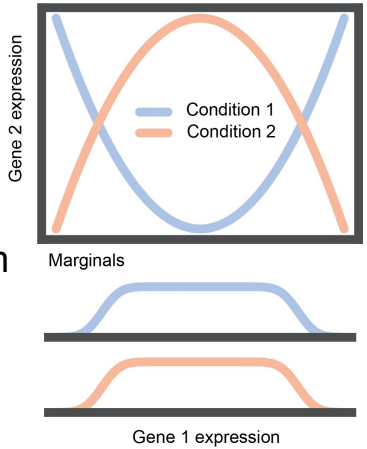


Filtering, normalization and clustering

Differential gene expression analysis

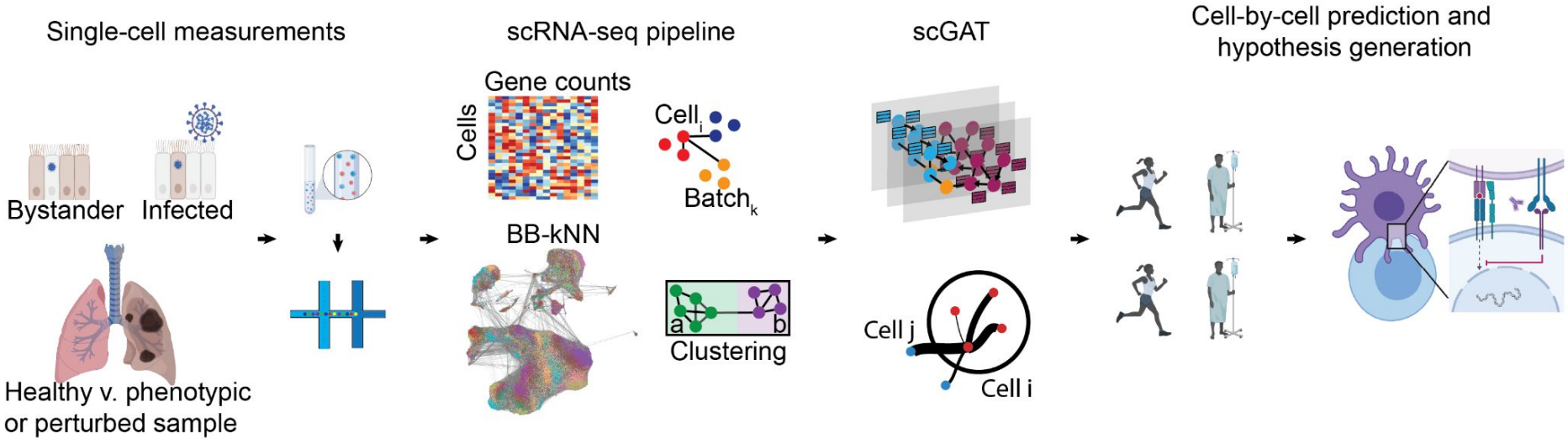
DGE is not necessarily *associated* with disease or cell state (“pipeline” error), does not allow for interactions between features, and the “most” differentially expressed genes do not yield causal structure

DGE can miss trivial non-linear examples; better to think of single-cell analysis as a supervised learning and interpretability problem



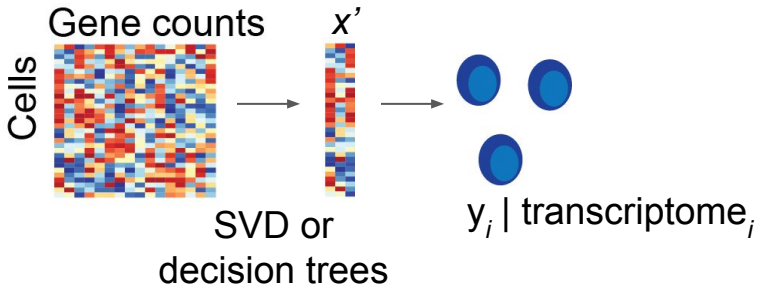
Want to *predict* cell biological state accurately, figure out *why* the model made those predictions, and propose targeted screens, biomarkers, and drugs or drug targets

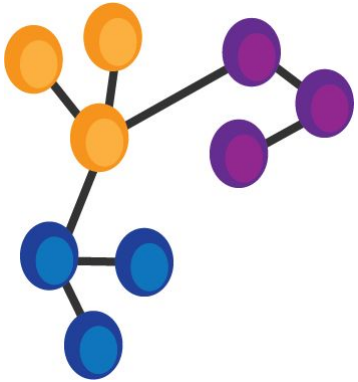
Geometric deep learning to represent single-cell data

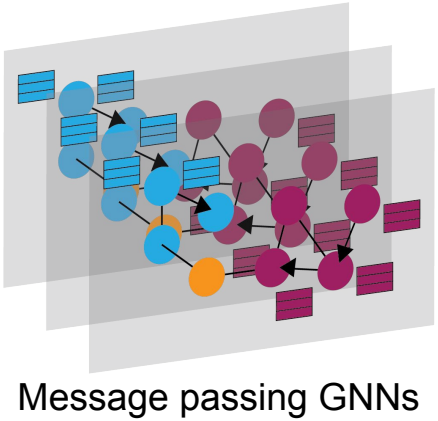


Ravindra NG*, Sehanobish A*, Pappalardo J, Hafler D, van Dijk D. *ACM CHIL*, 2020
Sehanobish A*, **Ravindra NG***, van Dijk D. *ICML'20 GRL+*
Sehanobish A*, **Ravindra NG***, van Dijk D. *AAAI'21*
Ravindra NG, Sehanobish A, Alfajaro MM, Wang B, Foxman EF, Wilen CB, van Dijk. (*in submission* at *Nature Methods*)
Ravindra NG, Sehanobish A, van Dijk D. (*in preparation for ICML'22 workshop*)

Geometric deep learning and attention mechanisms

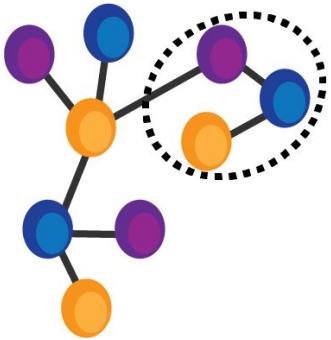



Cell graph


Message passing GNNs

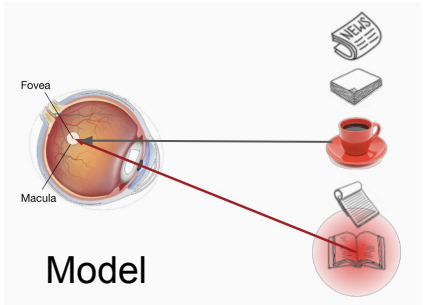
Single-cell data is heterogeneous, sparse, noisy, and # features ~ # cells so sub-sampling of features by classical ML is random--need inductive bias of self-*similarity*

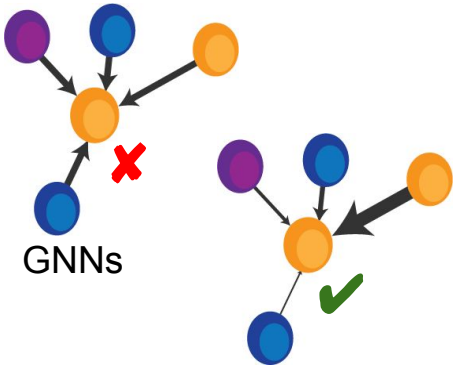
Geometric deep learning and attention mechanisms



Realistic cell graph

Attention





GNNs

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

GAT

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

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

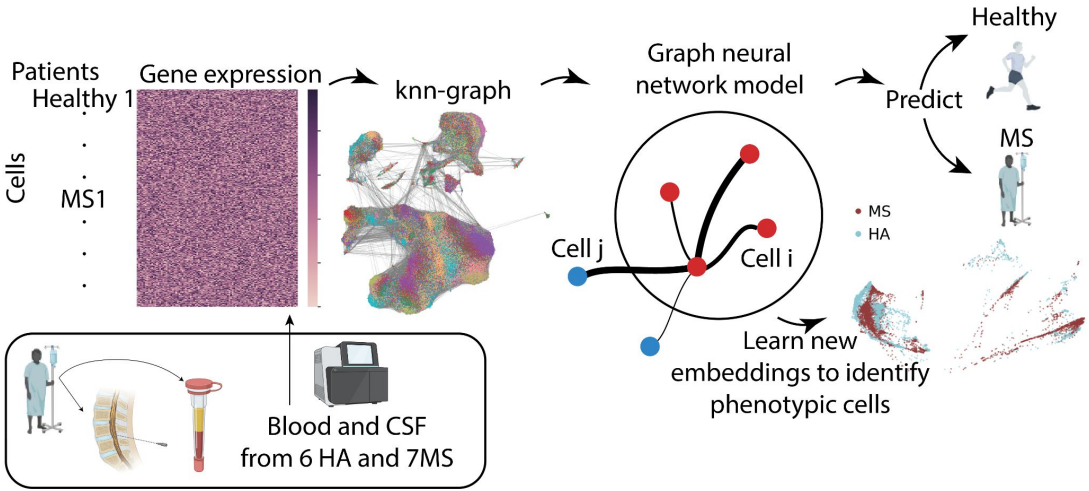
$$\alpha_{ij} = \mathrm{softmax}_j(a(\mathbb{W}h_i, \mathbb{W}h_j))$$

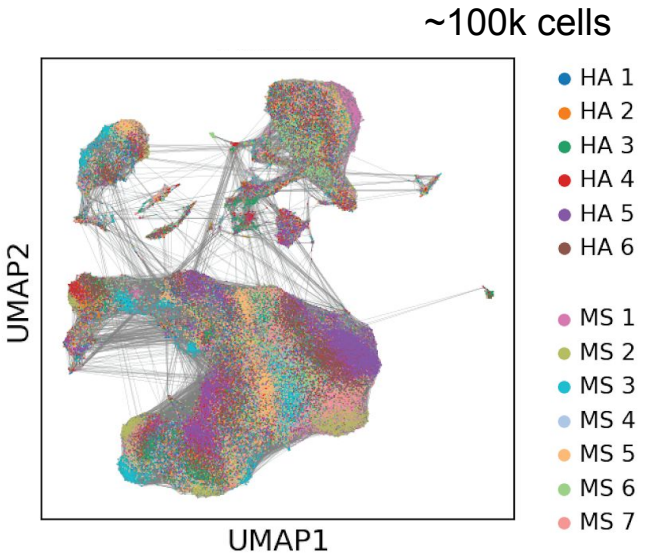
GAT or Graph attention networks
(axial attention)

Most GNNs equally weigh messages across edges, which is problematic with *bad* input cell graphs (e.g., high degree of heterophily)

Vaswani et al. *NIPS'17*; Lin et al. *ICLR'17*; Velickovic et al. *ICLR'18*

GAT supervised learning to study disease mechanisms





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

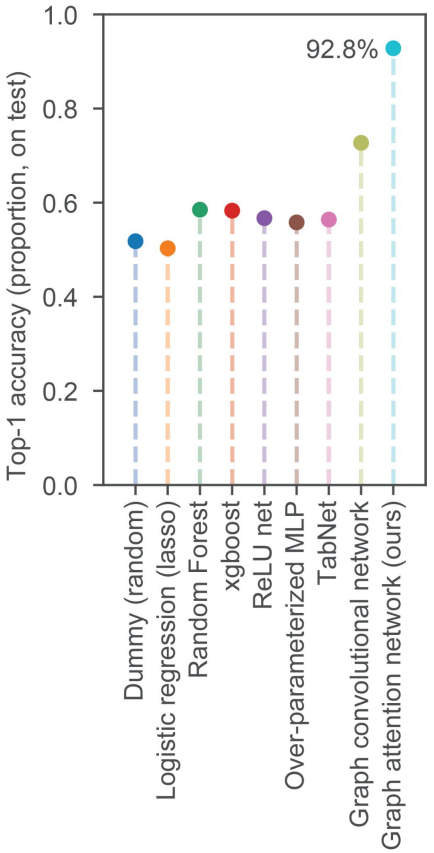
What factors are up-regulated in MS patients and what cell types/subsets are responsible

GATs learn to predict disease state from a transcriptome

Other common supervised learning approaches fail to learn how to predict each cell's disease state

GATs perform better relative to other popular GNN models

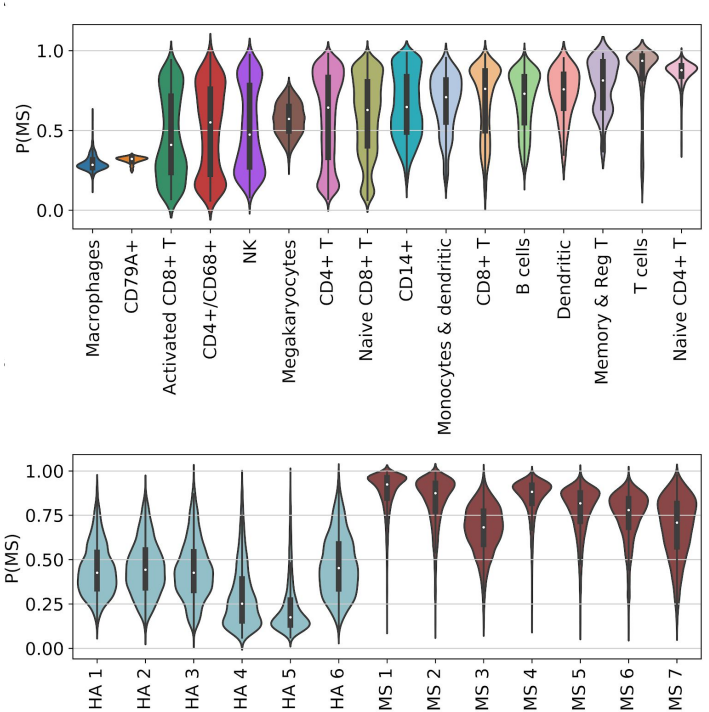
Task	Model	Accuracy
Inductive	Random	51.8
	MLP	56.7
	Random Forest	58.5
	Graph Convolutional Network	72.1
	Graph Attention Network(our)	**92.3** $\pm$ .7
Transductive	Graph Convolutional Network	82.91
	Graph Attention Network(our)	**86** $\pm$ .3



Simultaneous molecular and cellular interpretability

Aggregating predicted probabilities shows cell types important for predicting disease state

Variance of a patient's cells' probability of being in an MS state may indicate timing of flare-up



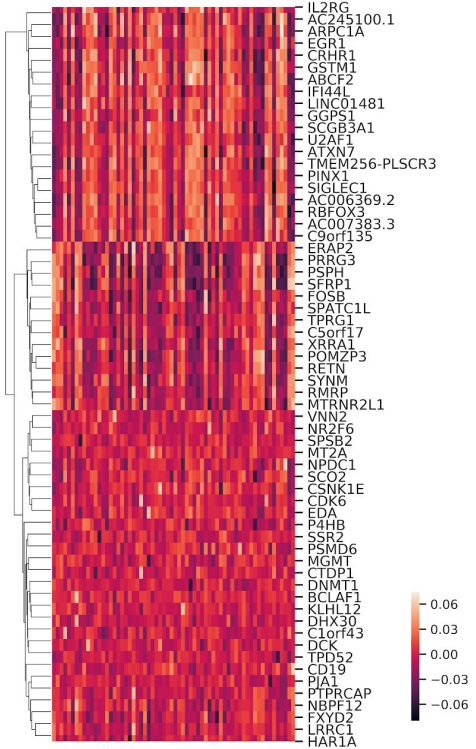
Finding genes important for predicting disease state

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

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

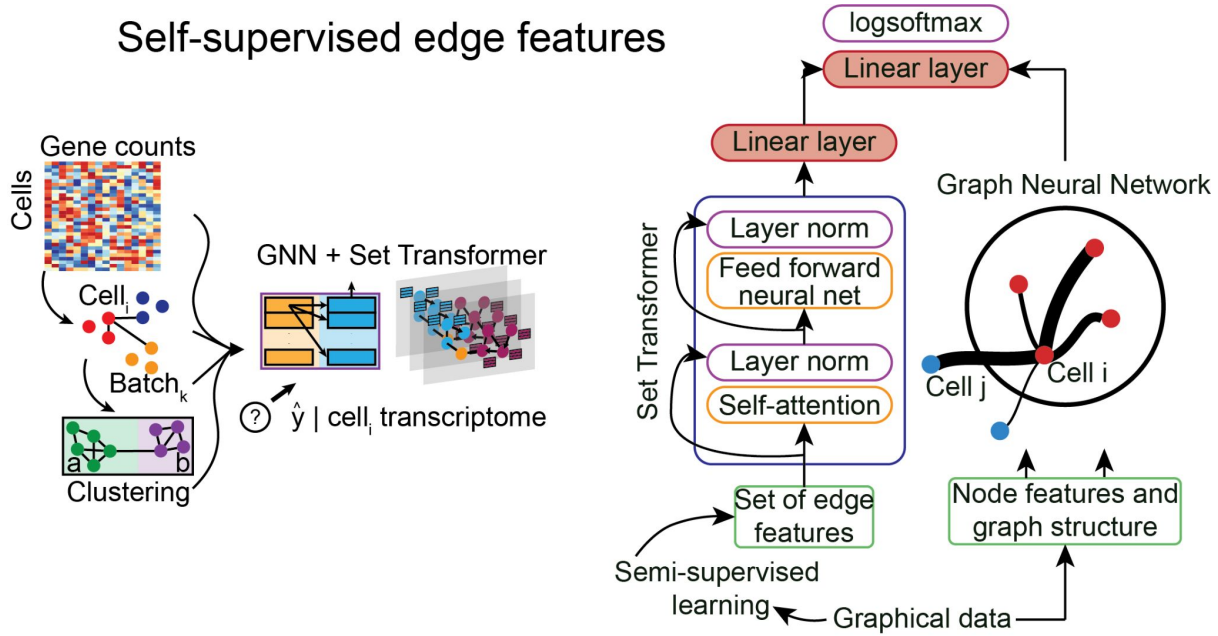
Marker for therapeutically targeted B cells (CD19) also among top features

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

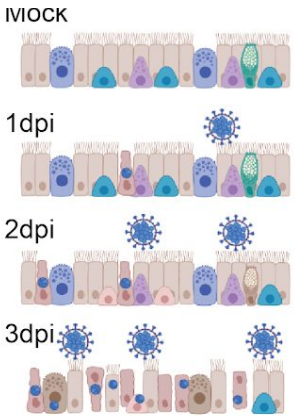


Integrating self-supervised learning and GATs

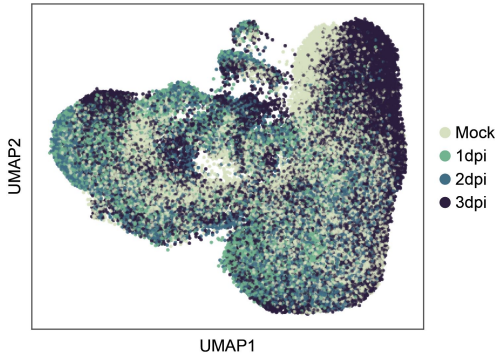
Goal to gain insight into SARS-CoV-2 infection and COVID-19 severity with additional controls for patient and sample source variability for robustness

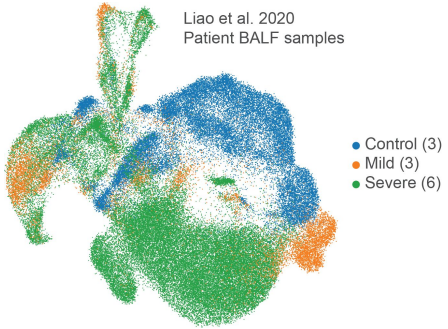


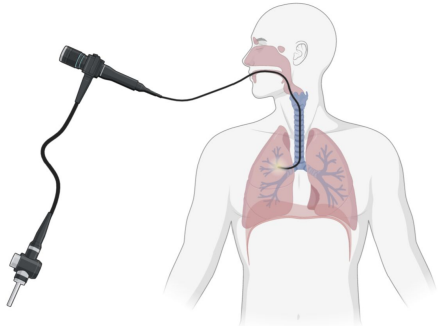
scGAT for HBEC & heterogeneous BALF samples



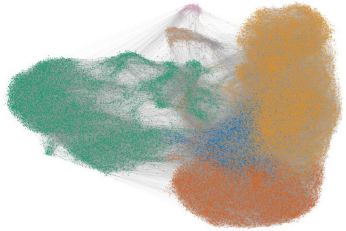
HBEC







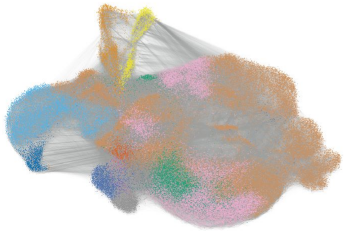
SARS-CoV-2 infected Bronchial Epithelium Organoids



Cell types

- BC/Club
- Basal cells
- Ciliated cells
- Club cells
- Goblet cells
- Ionocytes
- Neuroendocrine
- Tuft cells

COVID-19 Patient Samples



- Proliferating
- Ciliated
- Club
- Dendritic
- Eosinophil
- Macrophage
- Monocyte
- Neutrophil
- Progenitor
- T cells
- Tuft

~100k cell datasets: HBEC to study molecular and cellular determinants of early infection, BALF samples to study aberrant cell types and genes responsible for severe/critical COVID-19

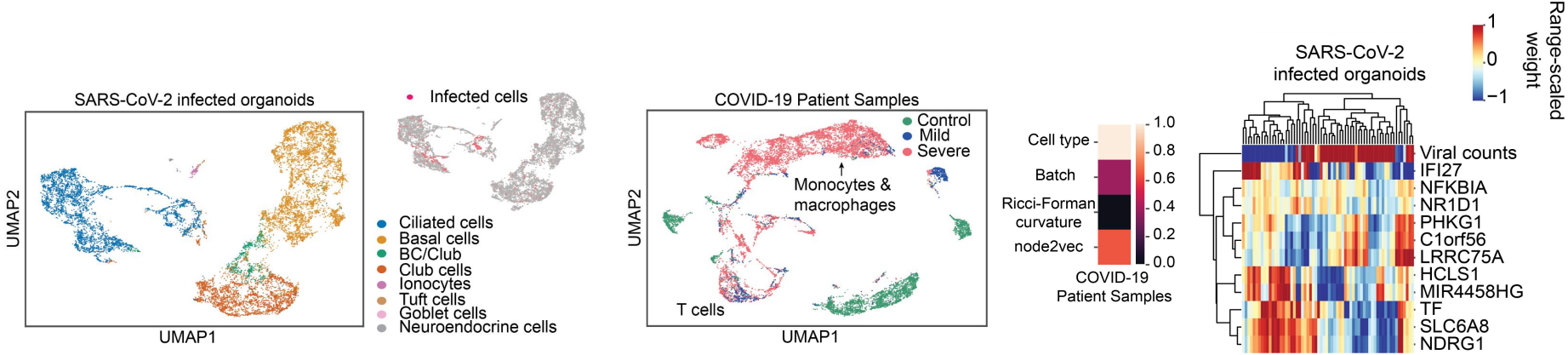
Our novel ingestion of *de novo* edge features using SSL

Models	SARS-CoV-2 infected organoids	COVID-19 patients
ClusterGCN	65.43 (65.21-65.65)	89.26 (89.06-89.47)
ClusterGCN + DeepSet	79.75 (78.75-80.75)	87.2 (87.02-87.38)
ClusterGCN + Set2Set	71.65 (69.89-73.42)	88.34 (87.89-88.79)
ClusterGCN + Set Transformer	81.61 (79.34-83.87)	92.84 (91.95-93.74)
GAT	73.10 (70.93-75.27)	92.25 (91.27-93.24)
GAT + DeepSet	79.45 (77.98-80.92)	75.99 (74.8-77.68)
GAT + Set2Set	82.95 (81.75-84.15)	92.87 (92.62-93.12)
GAT + Set Transformer (Ours)	**89.8 (88.89-91.71)**	**95.12 (94.02-96.22)**
GIN + EdgeConv[1]	63.36 (62.53-64.19	89.56 (88.54-90.58)
EdgeConditionedConvolution[1]	46.15 (34.72-57.59)	88.63 (86.07-91.20)

Our model significantly improves predictive performance over popular GNN methods, and controls for sample source and heterogeneity, giving high-confidence that it can be interpreted

Accuracy and 95% CIs for n=6 trials

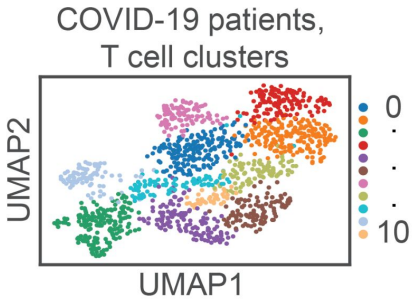
Subsetting cells important to COVID-19 severity using learned graphical representations

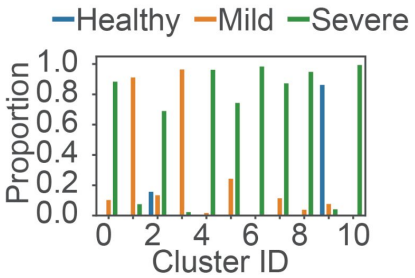


Visualizing attention using traditional unsupervised learning approaches shows model simultaneously discriminates by cell type and label

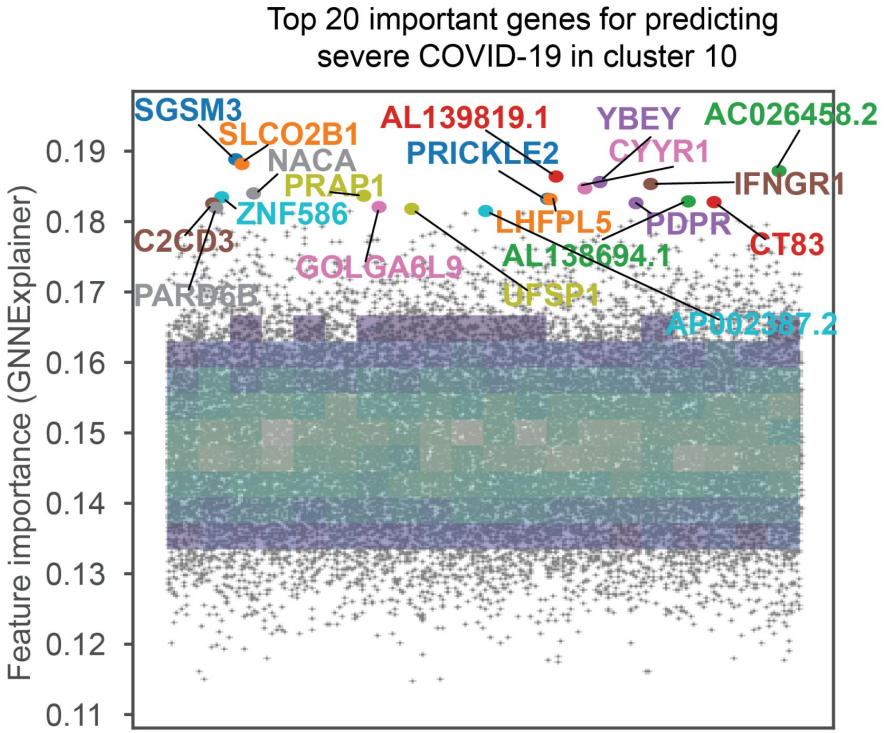
Model relies on genes involved in the innate immune system to discriminate between HBECs and cell types in heterogeneous clinical samples

Genes driving an active cluster of T cells in severe pts



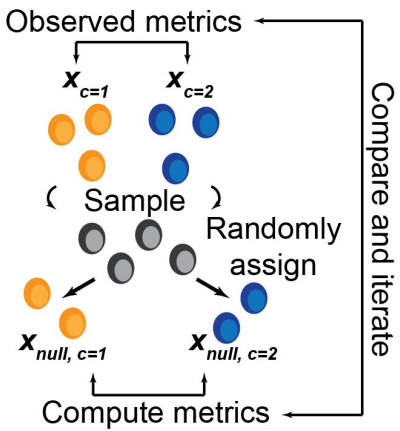


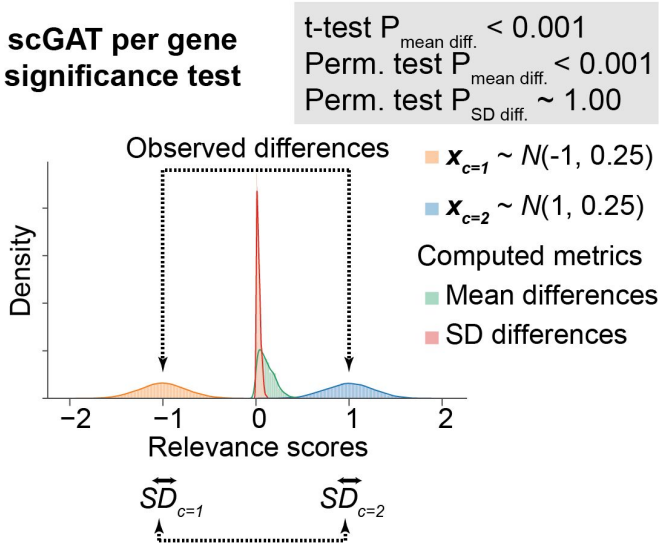
Combining interpretability via attention and feature attribution methods allows us to simultaneously gain insight into cells and genes driving model's association with label



Defining *significant* features associated with cell state

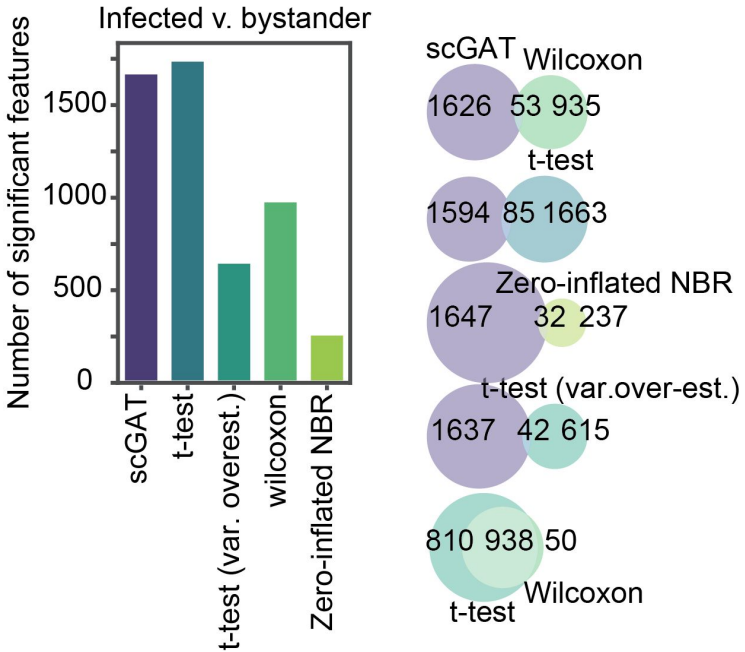
SmoothGrad scores

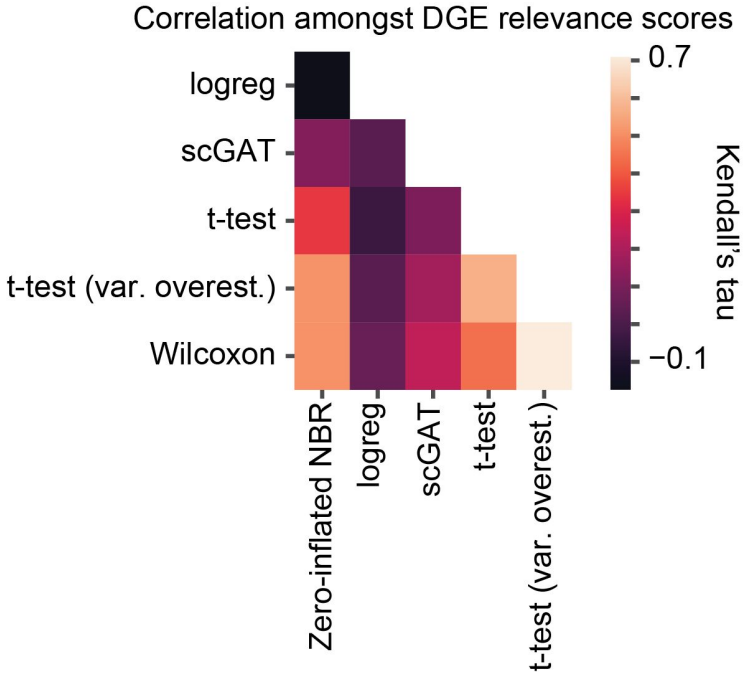




Goal: going beyond top-k features

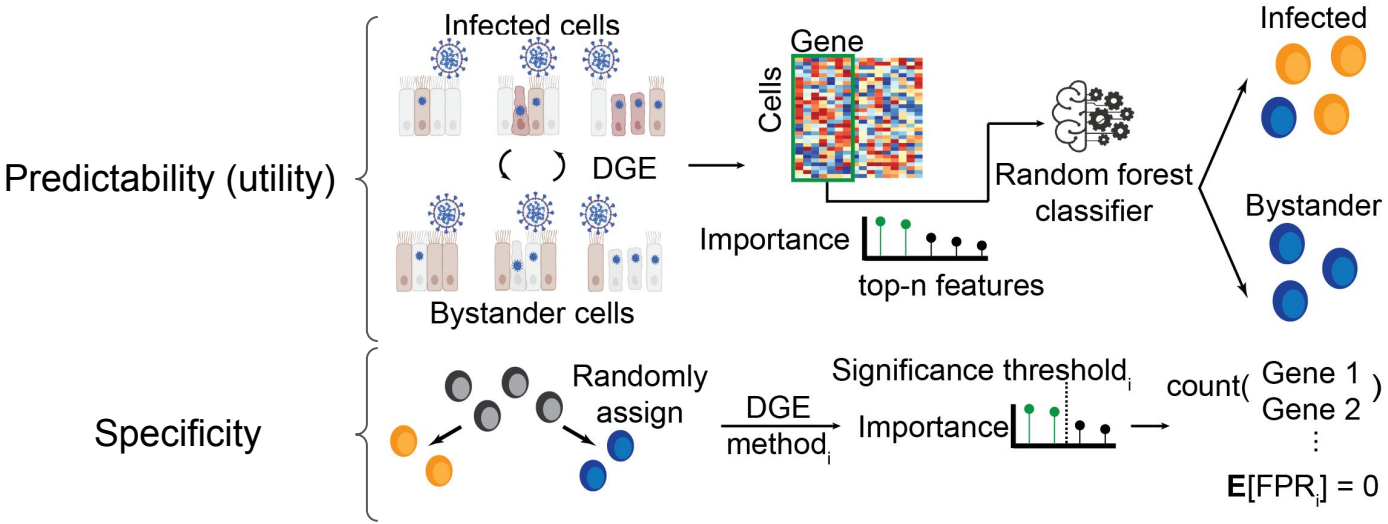
scGAT identifies unique infected v. bystander differences



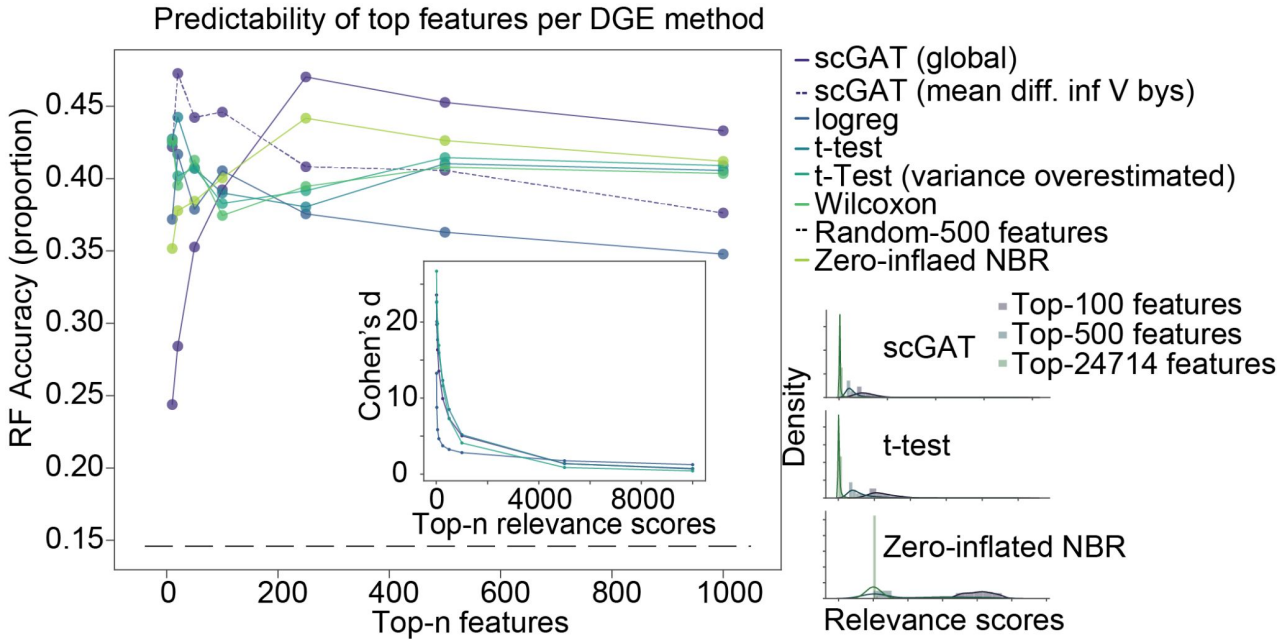


Lack of overlap with scGAT's *significantly* important features may suggest many genes are missed by standard DGE methods

Comparing scGAT features to other DGE methods

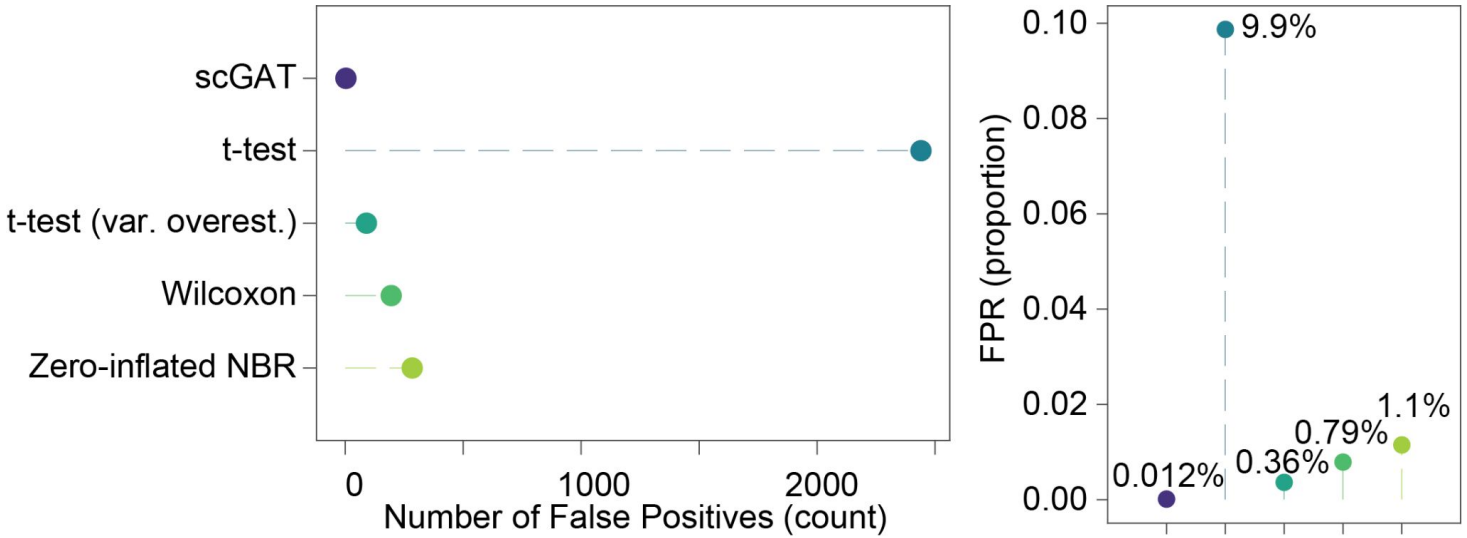


Predictability of top important features, scGAT v. DGE



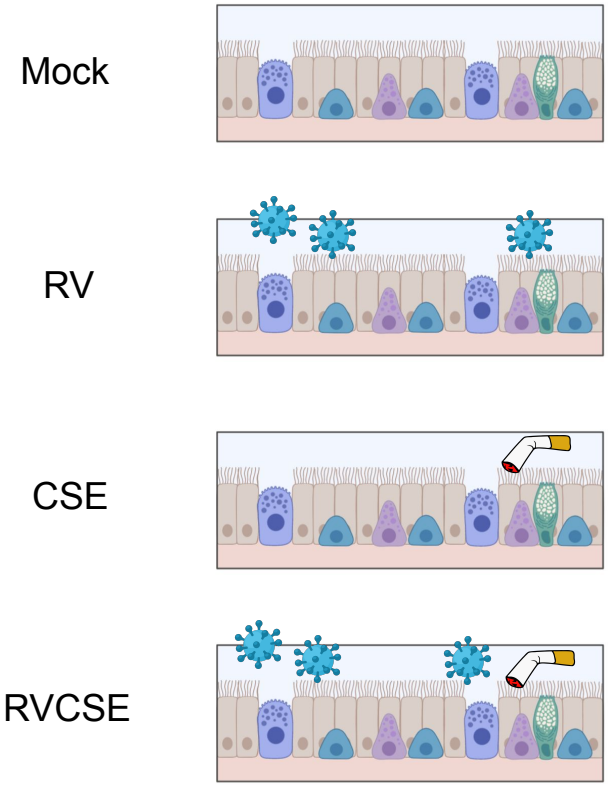
scGAT features are more predictable than features deemed "important" by standard single-cell DGE methods

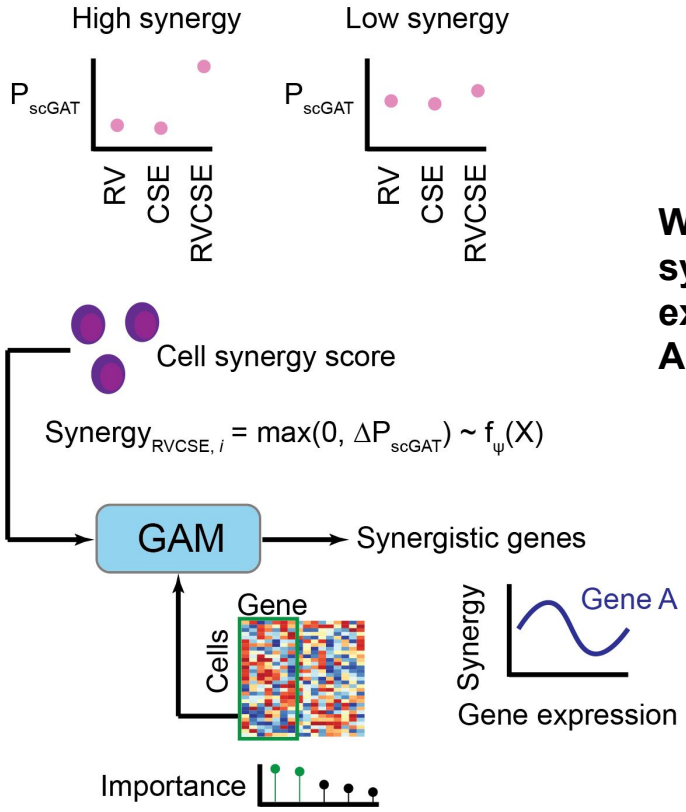
scGAT feature importance yields the lowest FPR



scGAT has a virtually 0 false-positive rate, which can be quite high for DGE methods, especially large datasets (post hoc inference tests use ~10k cells)

Identifying synergistic effects and associated genes





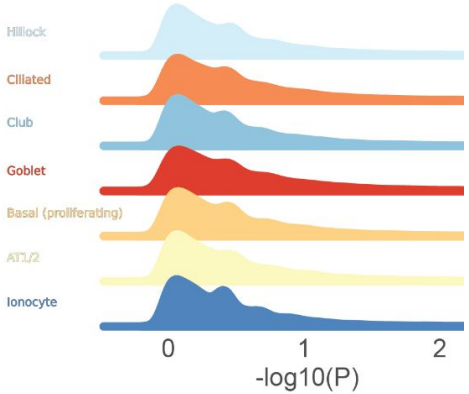
What genes/cells have synergism in combined exposure,
A v. B. v. AB?

Bao Wang and Ellen Foxman

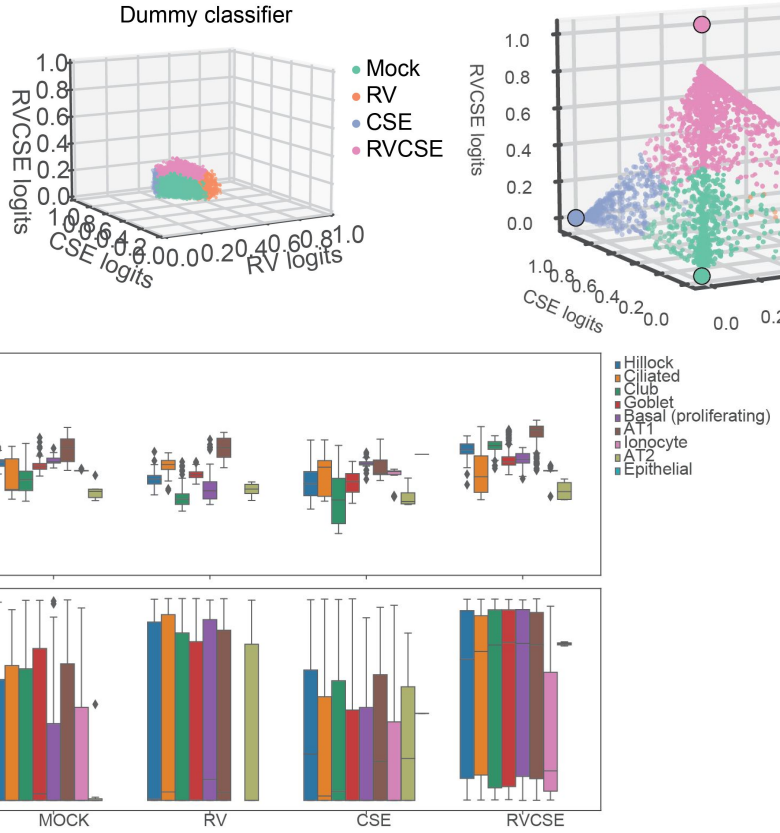
Using scGAT logits to define synergy scores per cell

Cell synergy score

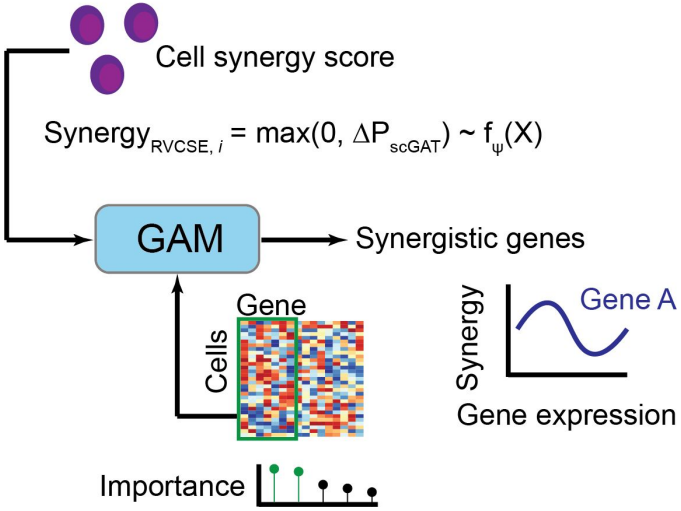
$$Synergy_{RVCSE,\,i} = \max(0, \Delta P_{scGAT}) \sim f_{\psi}(X)$$

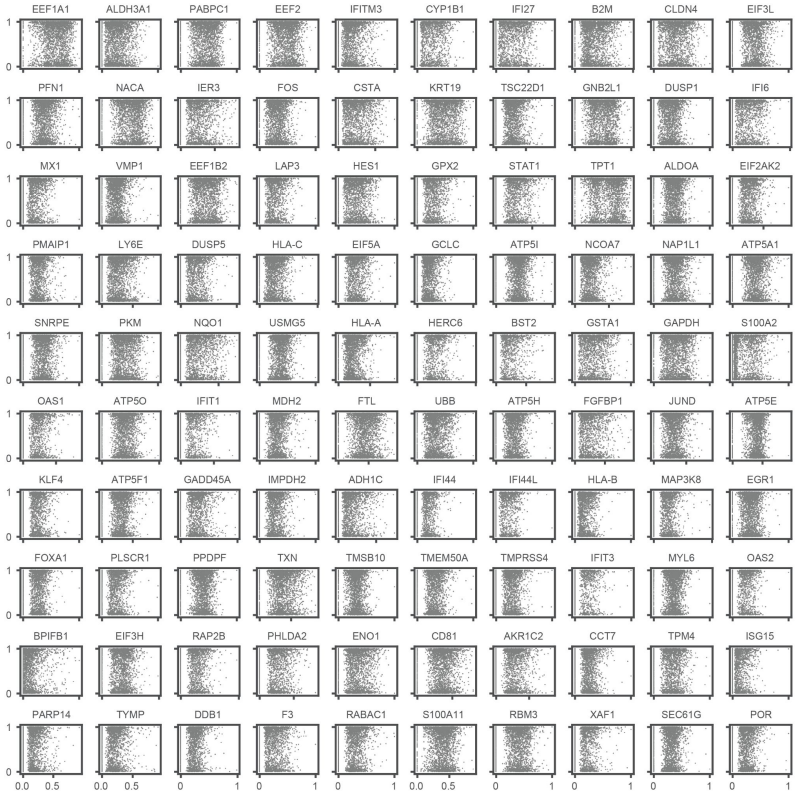


Flexible, non-linear, *learned* scoring



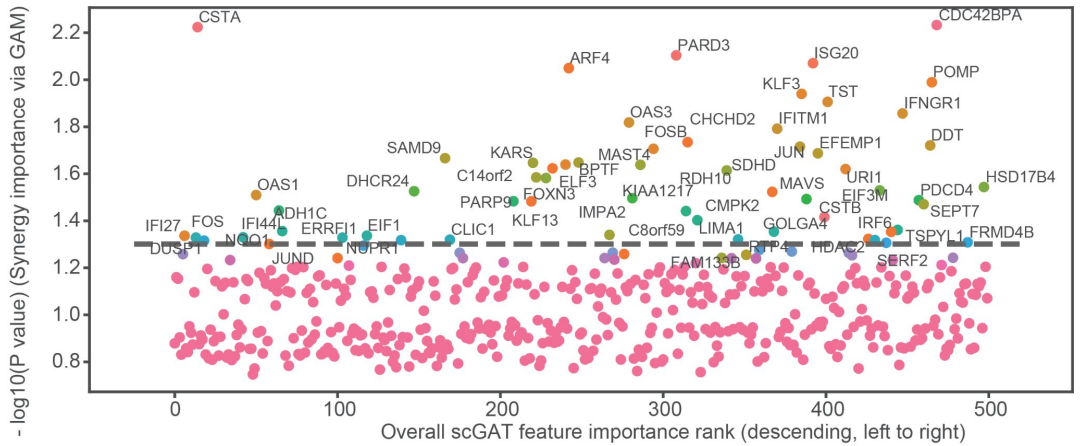
Synergistic genes: attributing genes to cell synergy score

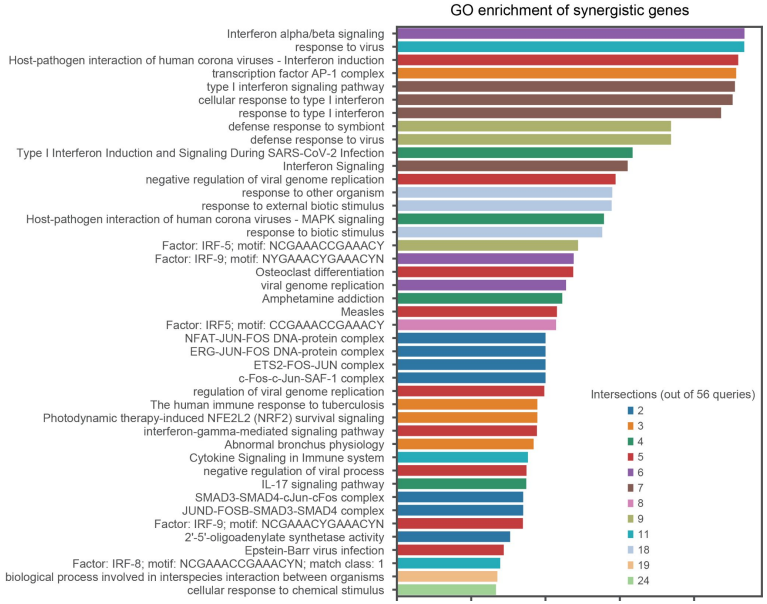




scGAT top features don't have easiest expression patterns for an interpretable, linear model

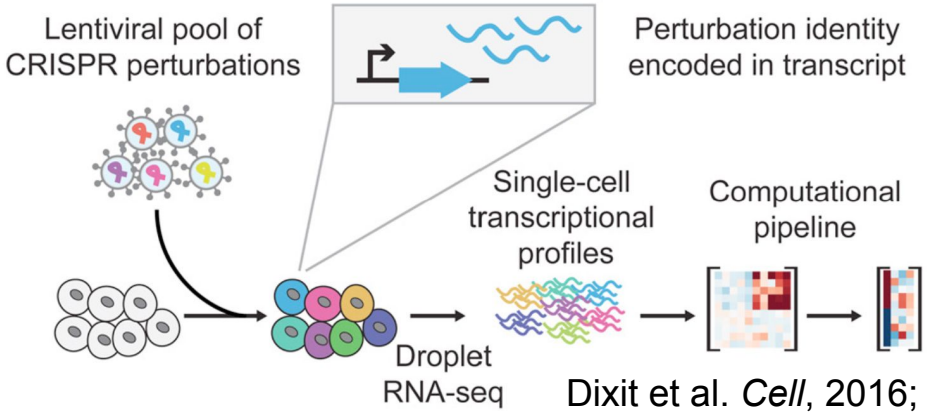
Biological processes of identified *synergistic genes*





scGAT can be used to identify semantically meaningful features associated with various complex and overlapping experimental designs or environmental contexts

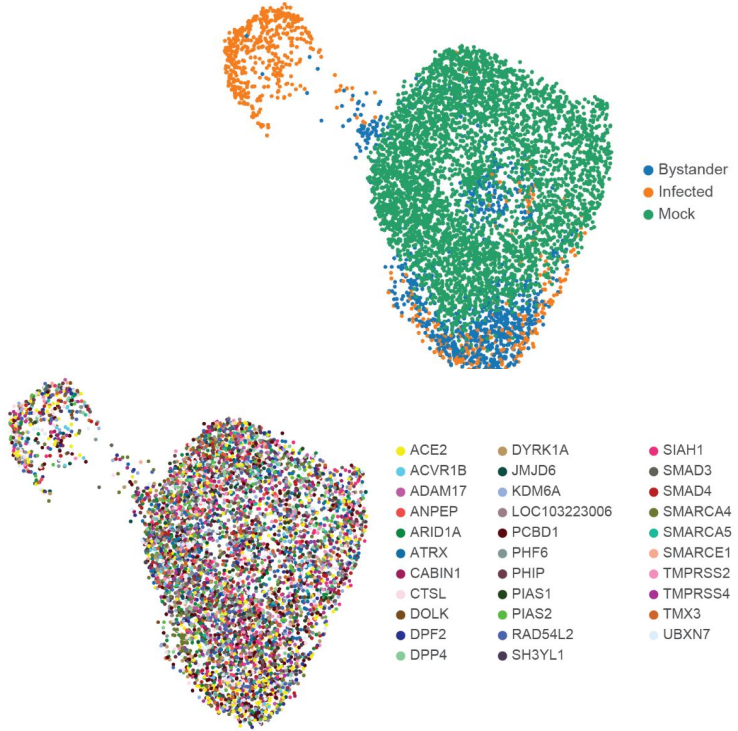
Single-cell Perturb-seq data in multiple contexts



Dixit et al. *Cell*, 2016;
Adamson et al. *Cell*, 2016

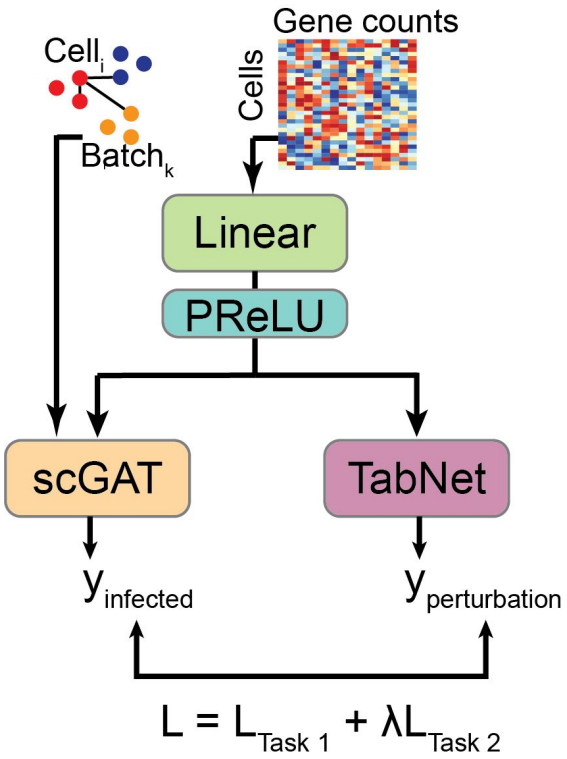
Want to find which genes *and* perturbations are associated with infection susceptibility or response to SARS-CoV-2

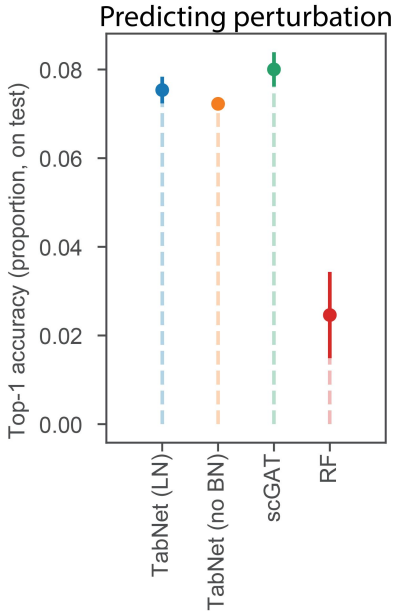
Want to know which genes respond to CRISR/Cas-9 mediated perturbation but only have cell-cell attention

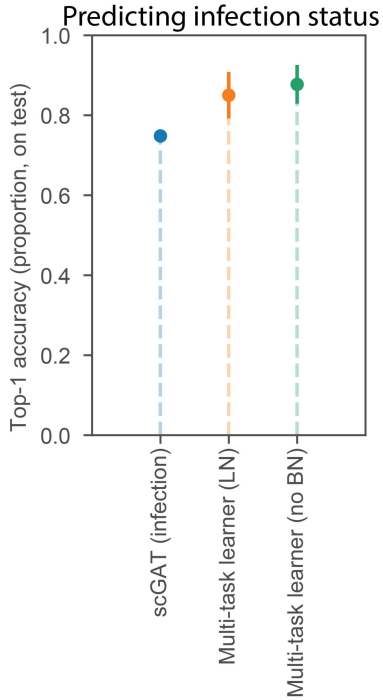


Vero E6 cells (monkey kidney), <10k cells, **MMA & CBW**

Multi-task learning to combine axial and feature attention



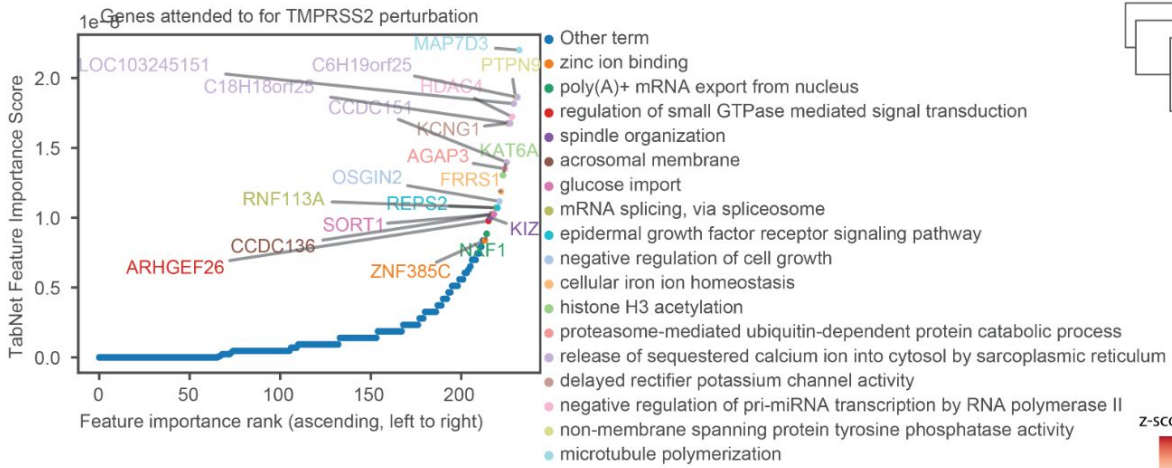


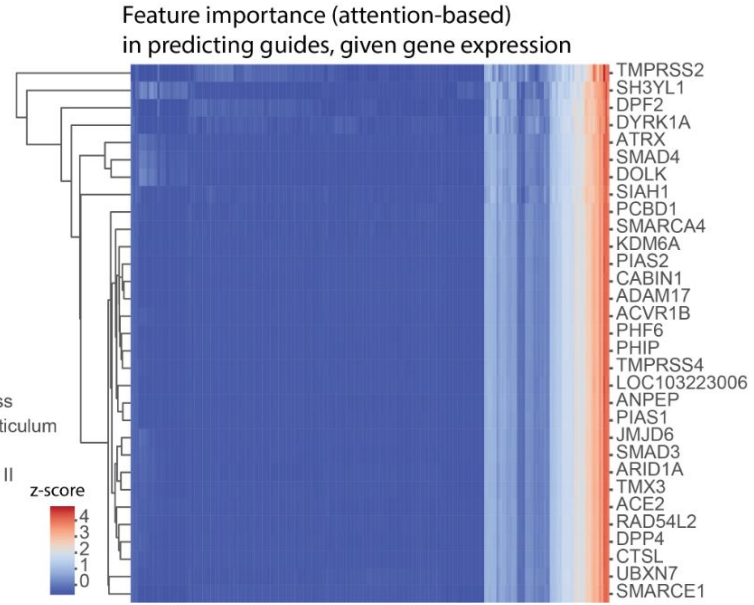


Using GATs for cell-cell attention and decision tree inspired feature masking for feature-wise attention boosts performance on primary task and aligns model and interpretability with application goal

Results | scGAT

Comparing feature-wise attention for infected & bystander





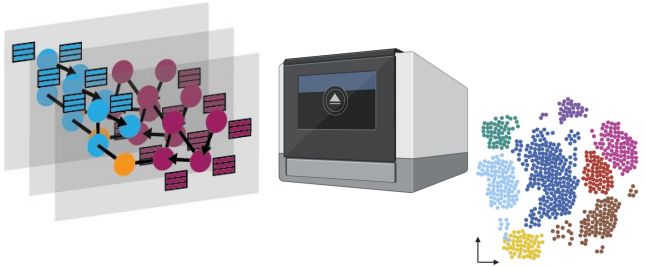
Multi-task, attention-based learning can provide insight into how perturbations differ between conditions, separating what responds to perturbation alone from what is responsive *and* differs between conditions

Overview

Interpretable ML to study molecular & cellular mechanisms of disease and cell state based on single-cell omics data

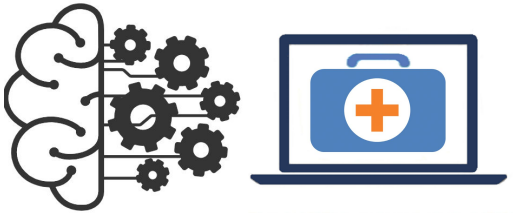
Dynamical genes from landmark time-point data

single-cell Graph Attention Networks (scGAT)



XAI to create clinically useful and parsimonious models

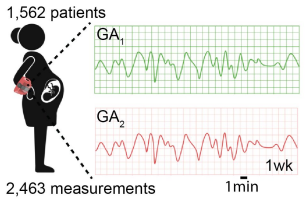
qCSI from a COVID-19 Severity Index model for triaging patients in the emergency department



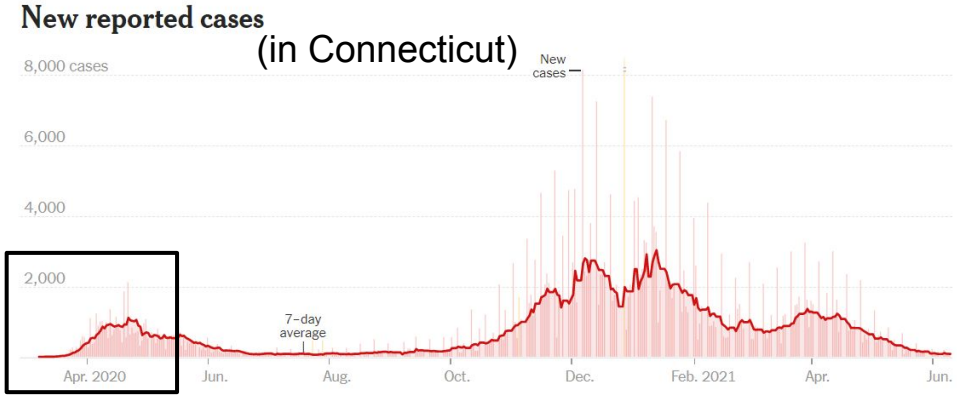
Translational research using relational reasoning and metric learning

actigraphy2GA: sleep and activity disruptions and their relation to preterm birth

sc2drug: perturbation modeling to align similar but disparate distributions



Prognostic challenge for the fast and perplexing onset of respiratory deterioration in SARS-CoV-2+ patients



Prior to YNHH “conversions” with no plateau in-sight

Resources: *inappropriate* inpatient dispositions lead to increased provider contacts and is associated with higher morbidity

Knew some patient risk factors for critical illness but scant evidence to guide safe dispositioning of SARS-CoV-2+ patients by providers

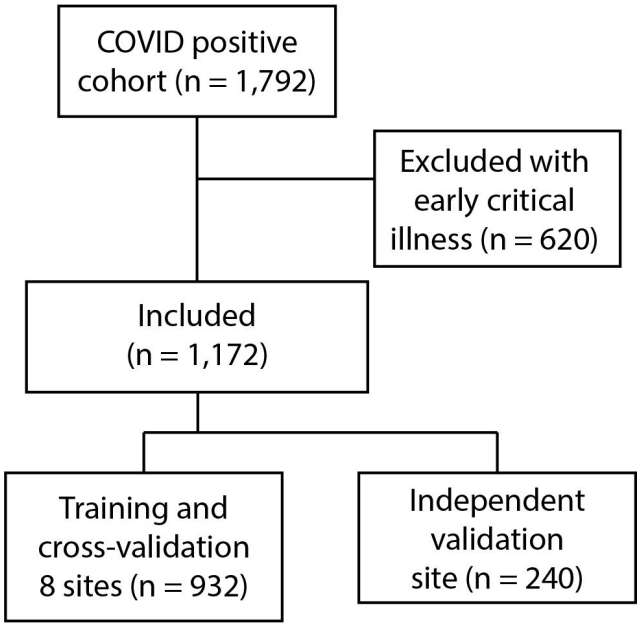
Already hints that quick Sequential [Sepsis-related] Organ Failure Assessment (qSOFA), Elixhauser, and CURB-65 score might be inadequate and definitions for severe COVID-19 were evolving

Relatively few patients (1,172 COVID-19+ across 9 EDs), high-dimensionality and messiness of EHR

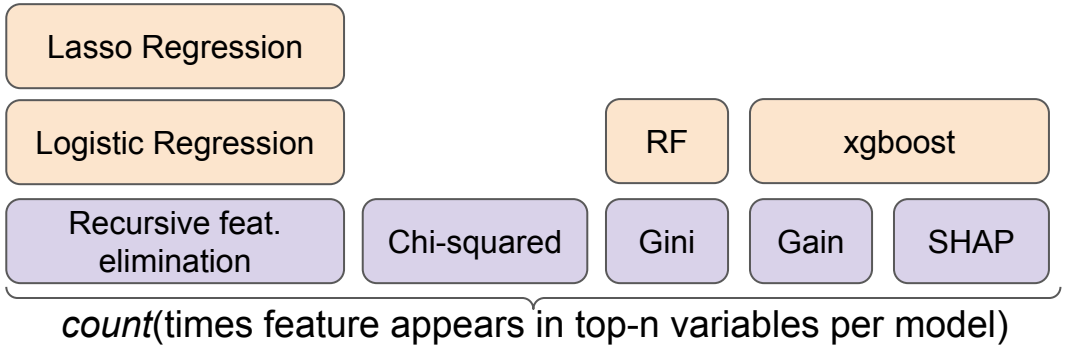
Adrian Haimovich, Andrew Taylor

Visual from *NYTimes*

Cohort selection, data preprocessing, and ML



Outcomes w/in 24h based on first 4h of data



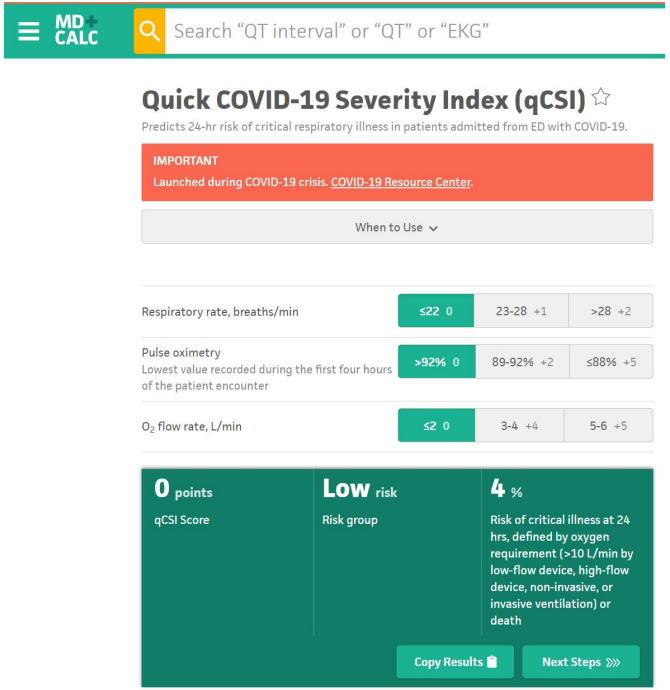
Ensemble approach: aggregating over multiple feature selection methods to overcome individual strengths and weaknesses

Leveraging interpretable ML to create a COVID-19 severity index for ED disposition decisions, w/easy entry

Developed a risk-stratification tool to predict **24h respiratory decompensation in admitted pts**

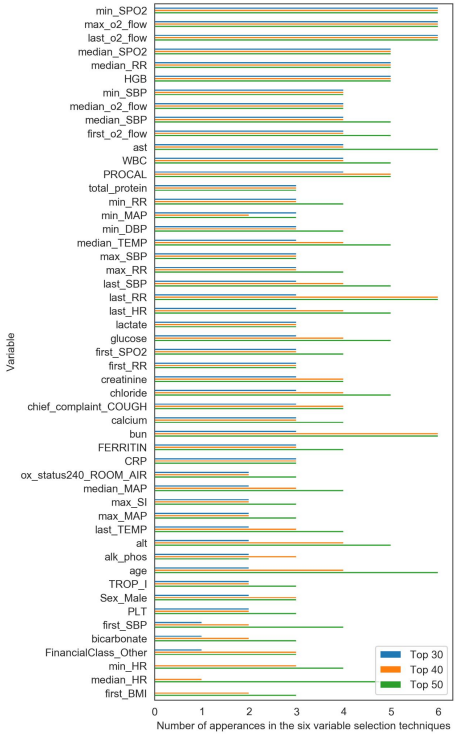
American College of Emergency Physicians official COVID-19 triage workflow

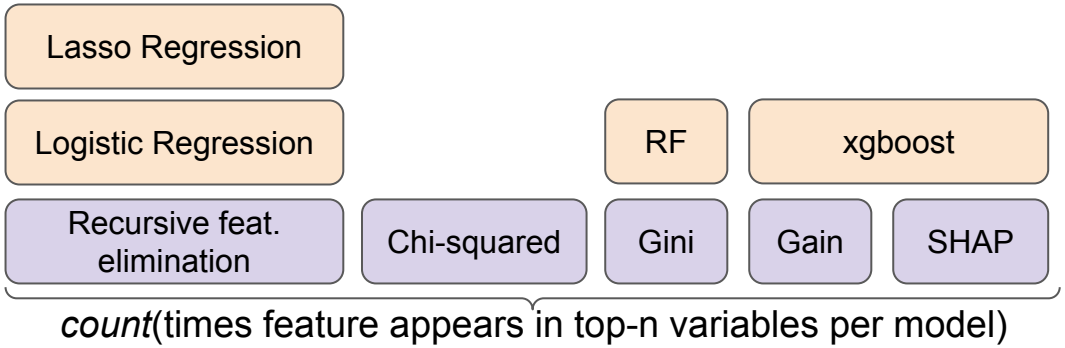
qCSI variable	Points	Additional CSI variables
Respiratory rate, breaths/min		Aspartate transaminase
≤22	0	Alanine transaminase
23–28	1	Ferritin
>28	2	Procalcitonin
Pulse oximetry, %*		Chloride
>92	0	C-reactive protein
89–92	2	Glucose
Oxygen flow rate, L/min		
≤88	5	Urea nitrogen WBC count
≤2	0	Age
3–4	4	
5–6	5	



Haimovich A*, **Ravindra NG***, … van Dijk D, Taylor RD. *Annal. Emer. Med.* 2020

Feature selection for qCSI and CSI





Preferentially selecting variables at bedside identifies 3 features consistently important: nasal cannula requirement, minimum pulse oximetry and respiratory rate for qCSI

+12 additional variables for a larger model, CSI

(q)CSI performance v. other clinical decision support tools

On external validation set (different hospital system):

	AU-ROC	Accuracy	Sensitivity	Specificity	AU-PRC	Brier Score	F1 Score	Average Precision
CSI	0.76 (0.65–0.86)*	0.79 (0.72–0.86)	0.73 (0.56–0.88)	0.81 (0.72–0.89)	0.38 (0.23–0.54)	0.25 (0.25–0.25)	0.47 (0.34–0.61)	0.40 (0.25–0.56)
CURB-65	0.50 (0.40–0.60)	0.64 (0.42–0.89)	0.57 (0.03–0.97)	0.52 (0.18–1.00)	0.18 (0.09–0.30)	0.12 (0.09–0.15)	0.13 (0.00–0.27)	0.16 (0.10–0.24)
Elixhauser	0.61 (0.51–0.70)	0.49 (0.26–0.74)	0.82 (0.45–1.00)	0.42 (0.15–0.78)	0.19 (0.11–0.29)	0.12 (0.09–0.15)	0.28 (0.20–0.37)	0.20 (0.13–0.30)
qCSI	0.81 (0.73–0.89)	0.82 (0.77–0.88)	0.79 (0.63–0.93)	0.79 (0.71–0.87)	0.47 (0.30–0.64)	0.10 (0.07–0.13)	0.49 (0.36–0.62)	0.44 (0.29–0.60)
qSOFA	0.59 (0.50–0.68)	0.83 (0.79–0.88)	0.47 (0.06–0.66)	0.72 (0.64–1.00)	0.22 (0.11–0.35)	0.12 (0.09–0.15)	0.08 (0.00–0.23)	0.20 (0.12–0.29)

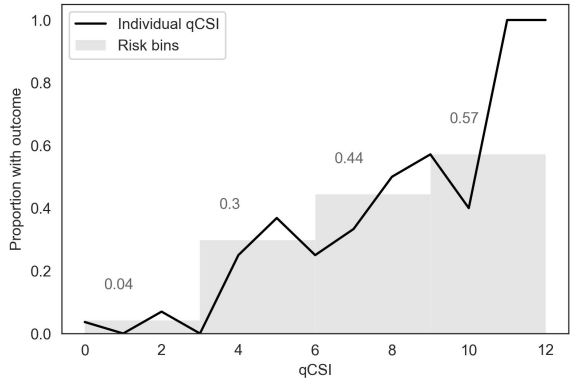
On cross-validation set:

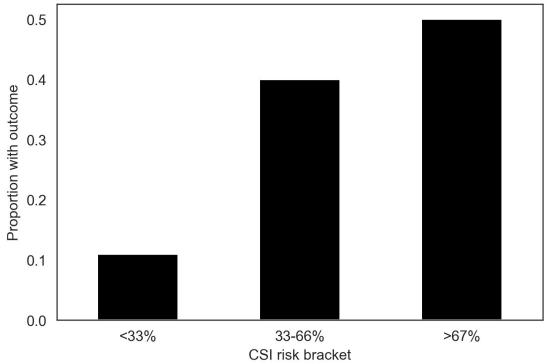
Model	AU-ROC	Accuracy	Sensitivity	Specificity	AU-PRC	Brier score	F1	Average Precision
CURB-65	0.66 (0.58,0.78)	0.79 (0.56,0.94)	0.67 (0.29,1.00)	0.62 (0.27,0.93)	0.26 (0.09,0.44)	0.10 (0.06,0.15)	0.20 (0.00,0.36)	0.20 (0.10,0.33)
qSOFA	0.76 (0.69,0.86)	**0.88 (0.82,0.95)**	0.79 (0.62,1.00)	0.70 (0.60,0.80)	0.35 (0.09,0.62)	0.09 (0.05,0.14)	0.21 (0.00,0.46)	0.26 (0.13,0.42)
Elixhauser	0.70 (0.62,0.80)	0.71 (0.40,0.86)	0.73 (0.47,1.00)	0.67 (0.33, 0.88)	0.20 (0.09, 0.36)	0.10 (0.06, 0.15)	0.30 (0.15,0.43)	0.22 (0.11, 0.36)
qCSI	0.90 (0.85,0.96)	0.84 (0.72,0.94)	0.90 (0.70,1.00)	0.79 (0.59,0.94)	0.54 (0.27,0.76)	**0.07 (0.04,0.11)**	0.49 (0.30,0.67)	0.52 (0.30,0.72)
CSI	**0.91 (0.86,0.97)**	0.83 (0.70,0.94)	**0.94 (0.77,1.00)**	**0.82 (0.67,0.95)**	**0.56 (0.25,0.80)**	0.25 (0.25,0.28)	**0.51 (0.29,0.70)**	**0.58 (0.31,0.81)**

Some overfitting but parsimony helps to reduce generalization error and qCSI outperforms other popular tools, without high FPR, which would reduce system burden

Calibrating the qCSI for use as a risk-estimation tool

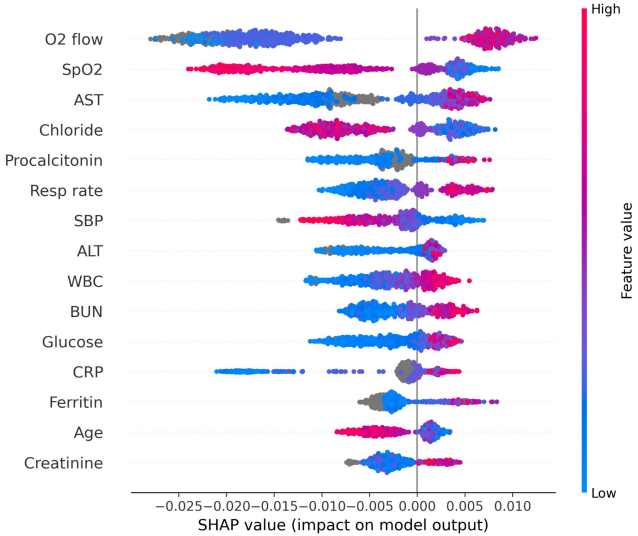
qCSI variable	Points	Additional CSI variables
Respiratory rate, breaths/min		Aspartate transaminase
≤22	0	Alanine transaminase
23–28	1	Ferritin
>28	2	Procalcitonin
Pulse oximetry, %*		Chloride
>92	0	C-reactive protein
89–92	2	Glucose
Oxygen flow rate, L/min		
≤88	5	Urea nitrogen WBC count
≤2	0	Age
3–4	4	
5–6	5	

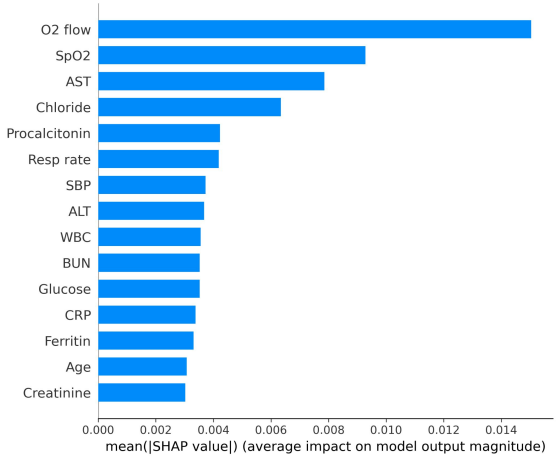


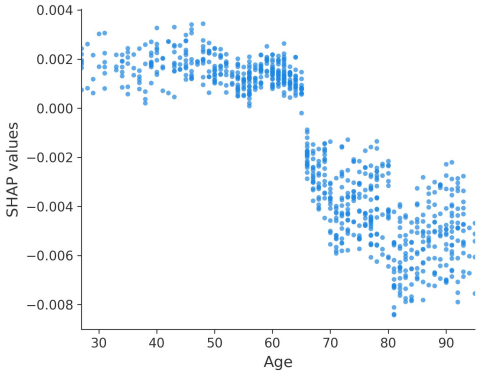


Akin to Platt scaling for calibration, used logistic regression to derive weights for the qCSI, then scaled ORs to an easy-to-use range. qCSI score of less than 3 has an resp. fail. outcome rate of ~4% w/in 24h

Interpreting CSI to gain insight into resp. decomp. course







Scale CSI model output by isotonic regression so that SHAP values reflect a relative weighting of contributions

Consistent finding that inflammatory markers are suggestive of clinical course and low oxygen flow rates and high pulse oximetry values are protective but, in these EDs, younger pts are at heightened risk

Lundberg et al. *ICML*, 2017; Niculescu-Mizil & Caruana. ICML, 2005

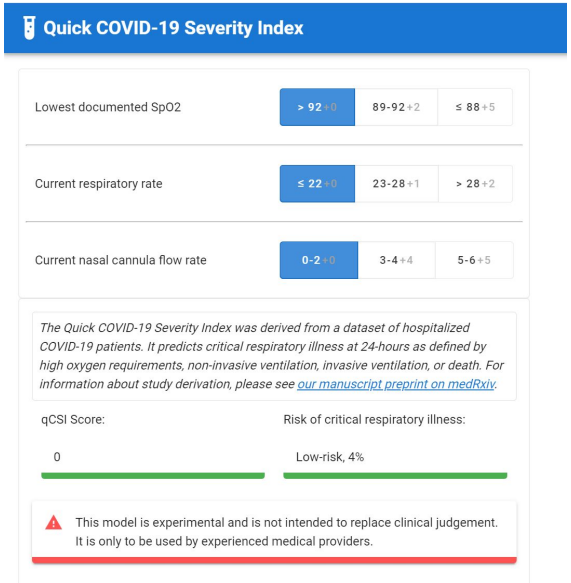
Rapid qCSI development to replace clinical decision support tools that were not optimized for COVID-19 triage

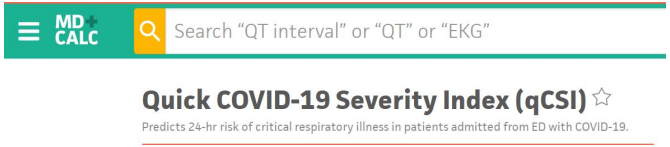
35% of admitted COVID-19 patients had respiratory failure w/in 24h; these events can be accurately predicted with simple bedside respiratory examination

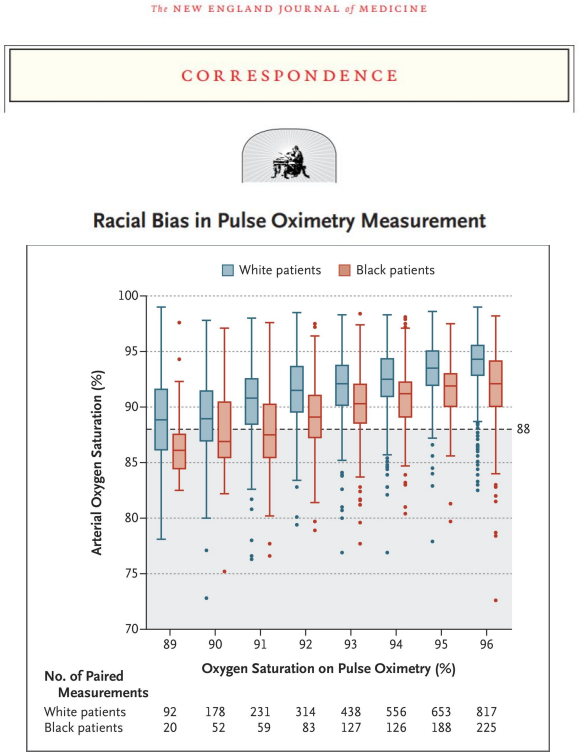
Interpretable ML can be used for feature selection but to also encourage providers to look out for hidden signs of risk, e.g., in younger patients and undetected aberrant liver chemistries

Further development could add stratification of patients for therapeutic interventions

See: https://covidseverityindex.org/







Sjoding et al. *NEJM*, Dec 17, 2020

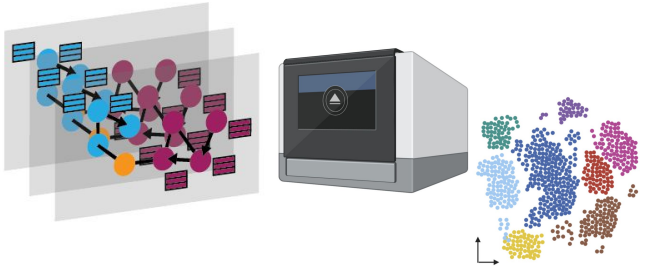
Yale SCHOOL OF MEDICINE

Overview

Interpretable ML to study molecular & cellular mechanisms of disease and cell state based on single-cell omics data

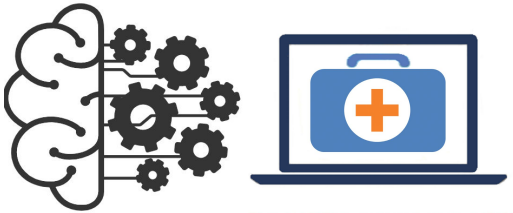
Dynamical genes from landmark time-point data

single-cell Graph Attention Networks (scGAT)



XAI to create clinically useful and parsimonious models

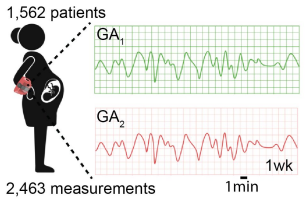
qCSI from a COVID-19 Severity Index model for triaging patients in the emergency department



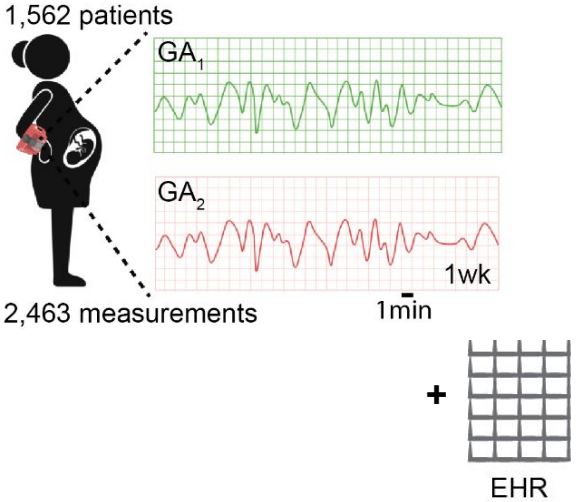
Translational research using relational reasoning and metric learning

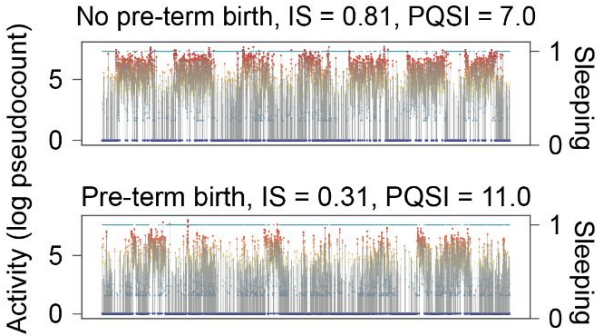
actigraphy2GA: sleep and activity disruptions and their relation to preterm birth

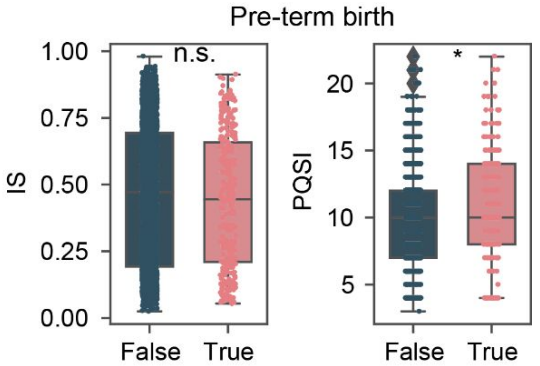
sc2drug: perturbation modeling to align similar but disparate distributions



Standard analyses fail to indicate risk of preterm birth





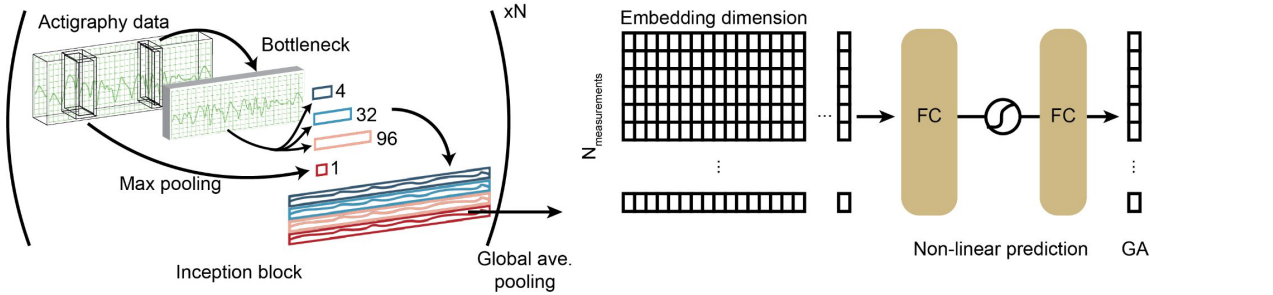


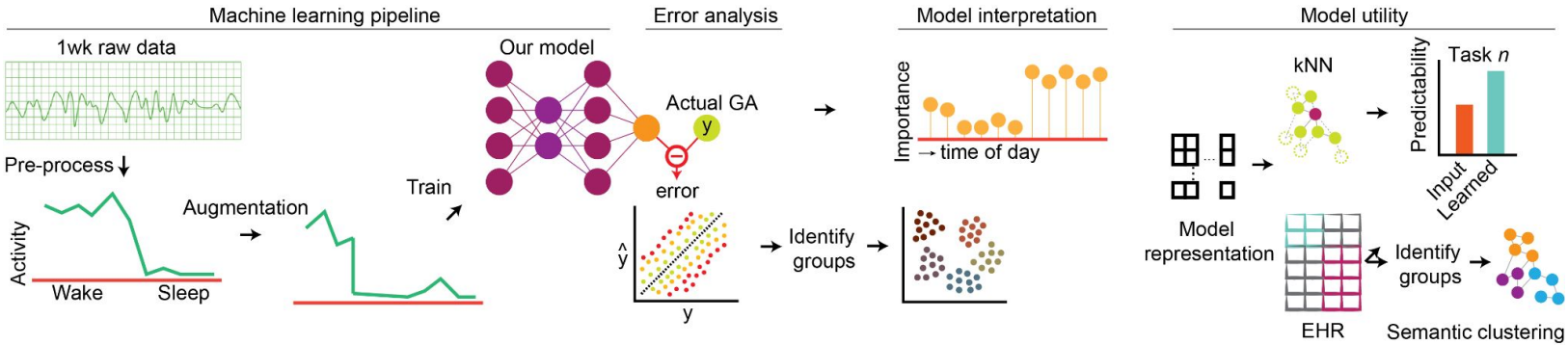
Ravindra, NG… Angst MS, Shaw GM, Stevenson D, Herzog E, Aghaeepour N. Deviations from Physical Activity and Sleep Patterns During Pregnancy Measured by a Wearable Device is Associated with Prematurity. 2022. (*in submission at Nat. Med.*)

Background | wearables

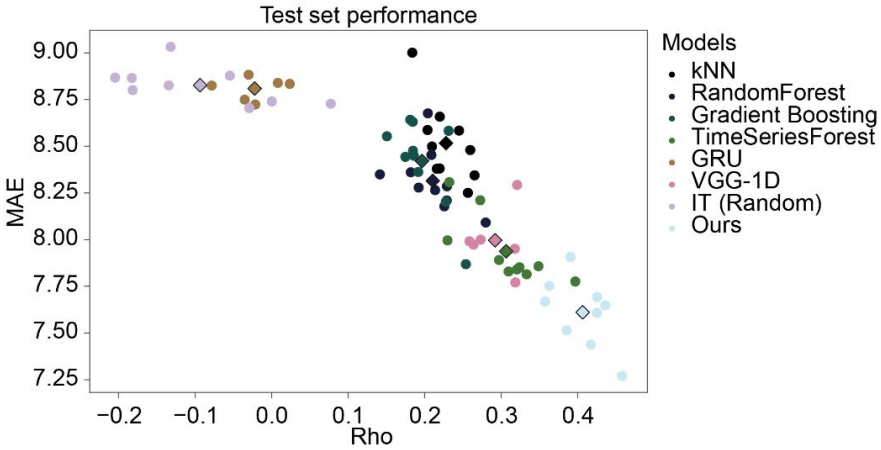
actigraphy2GA monitors pregnancy by applying deep learning to time series data monitoring activity and sleep

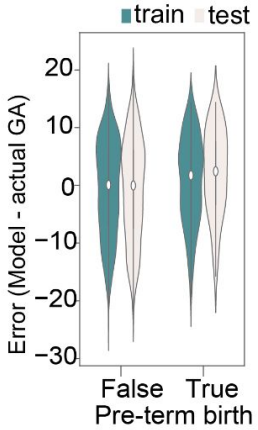
ResNet-inspired architecture

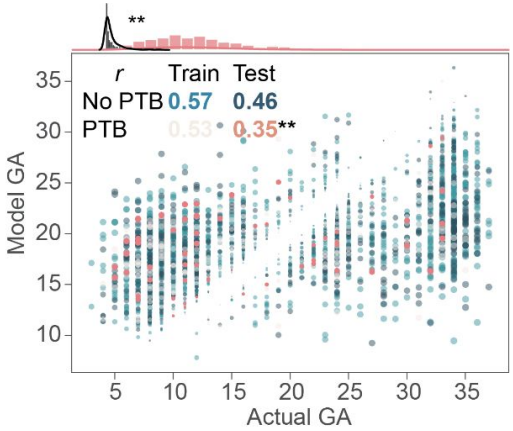




SOTA prediction of GA from 1wk of actigraphy data

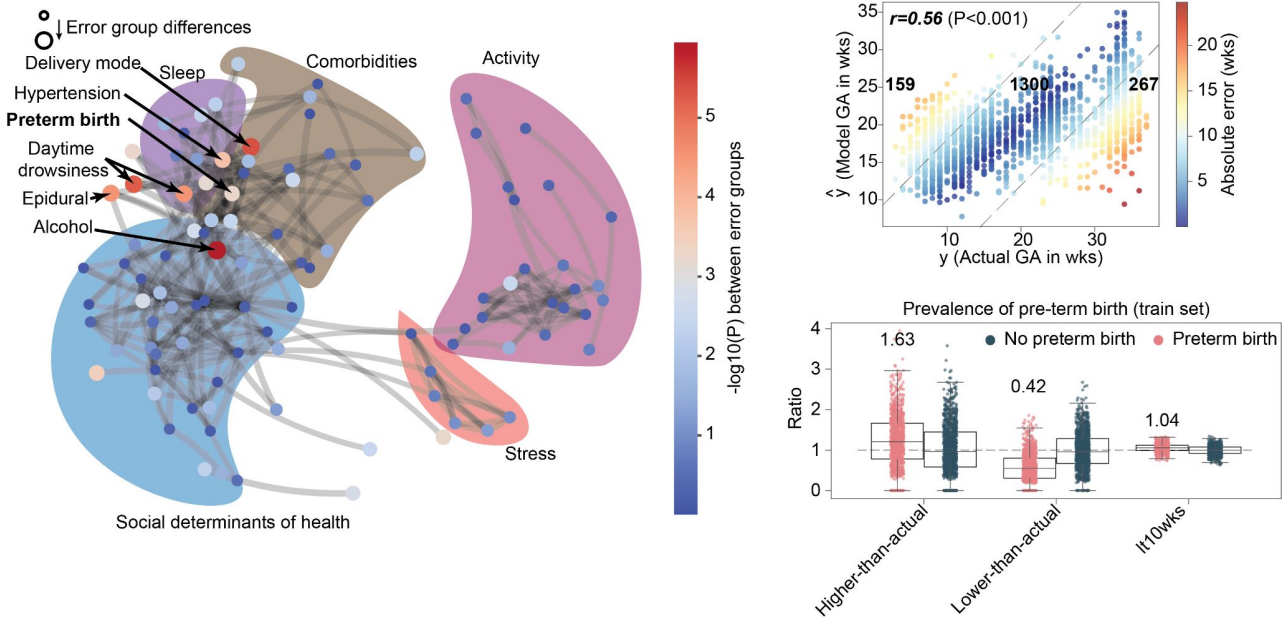




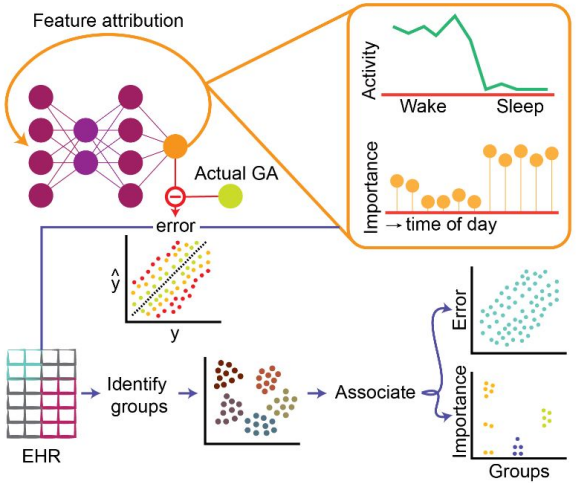


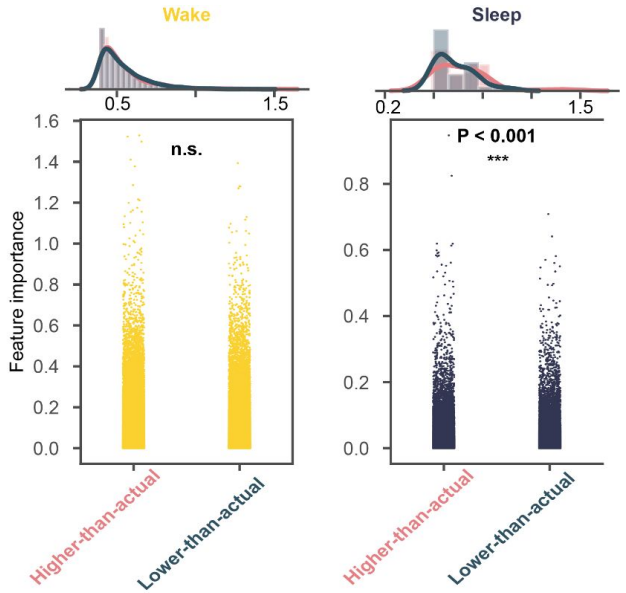
Results | wearables

Model error analyses reveals that actigraphy-GA indicates increased likelihood of adverse pregnancy outcomes

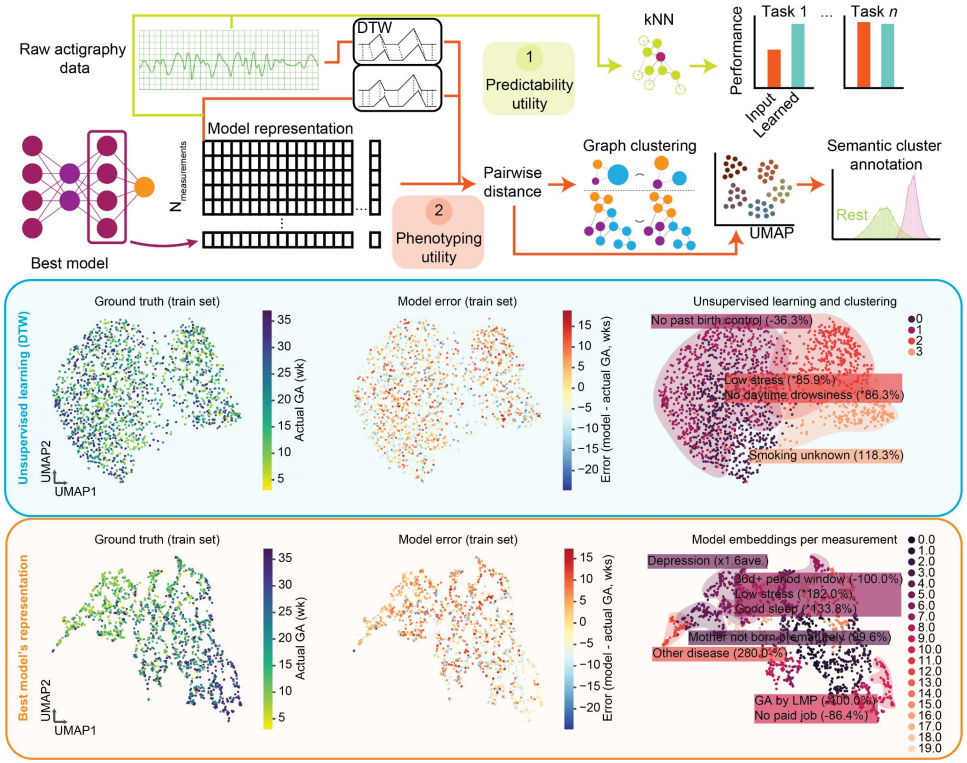


actigraphy2GA relies on deviations to sleep and activity in predicting higher- or lower-than actual GA





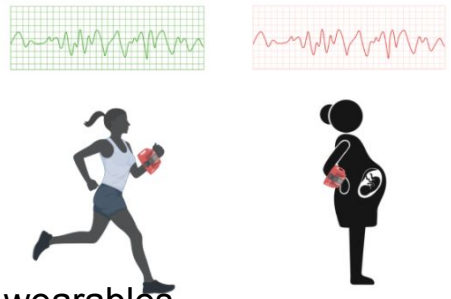
actigraphy2GA emeddings are useful for predicting ancillary tasks and semantic phenotyping



Characterizing biology of actigraphy-GA signaled patients

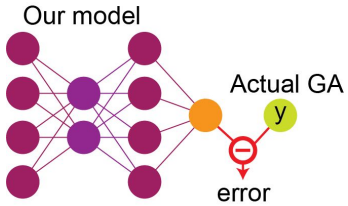
1

We can use actigraphy data to monitor pregnancy and identify risky behaviors



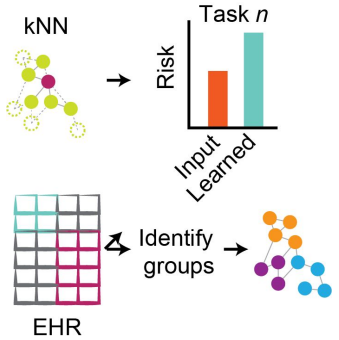
2

Sleep and activity disruptions cause the model to make errors



3

Interpreting the model shows we may be able to target inexpensive interventions



Attention on filters: transformer + ConvNet for time-series

Pre-training

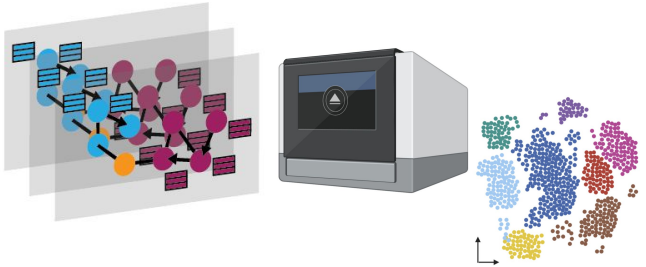
ConViT vs. TFT

Overview

Interpretable ML to study molecular & cellular mechanisms of disease and cell state based on single-cell omics data

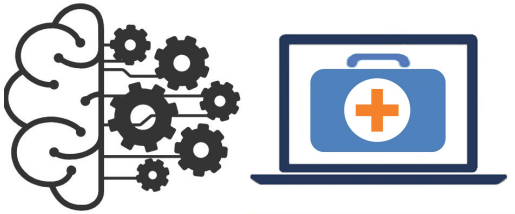
Dynamical genes from landmark time-point data

single-cell Graph Attention Networks (scGAT)



XAI to create clinically useful and parsimonious models

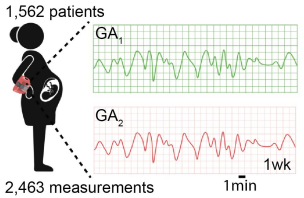
qCSI from a COVID-19 Severity Index model for triaging patients in the emergency department



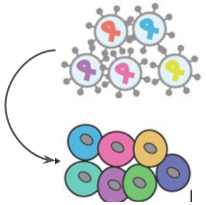
Translational research using relational reasoning and metric learning

actigraphy2GA: sleep and activity disruptions and their relation to preterm birth

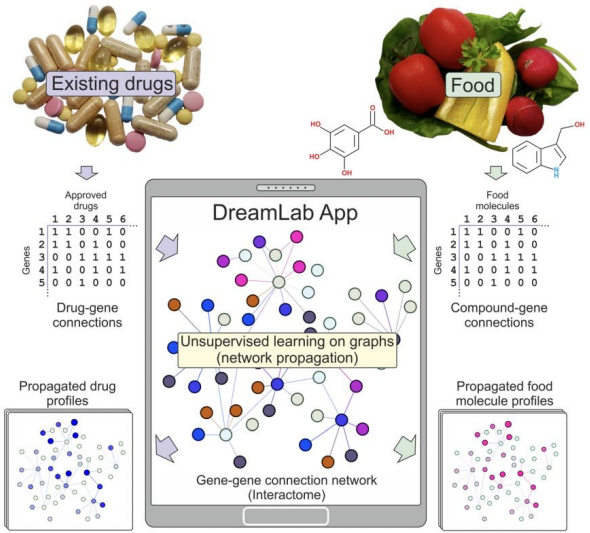
sc2drug: perturbation modeling to align similar but disparate distributions



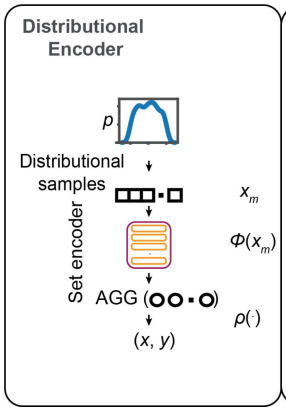
Preponderance of targets and determinants: combining recommender systems, representation and metric learning

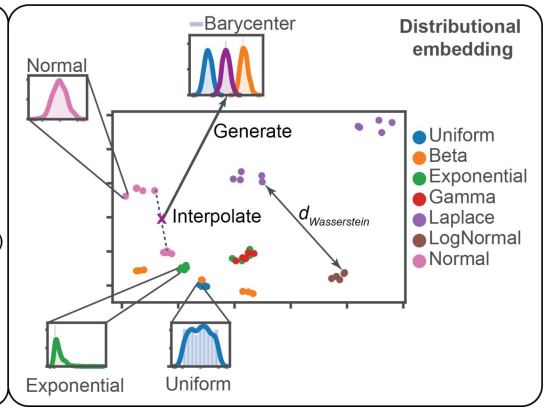


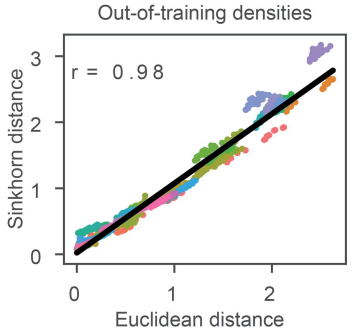
Perturbation modeling



Veselkov, K… Bronstein M et al. HyperFoods. *Sci. Rep.*, 2019

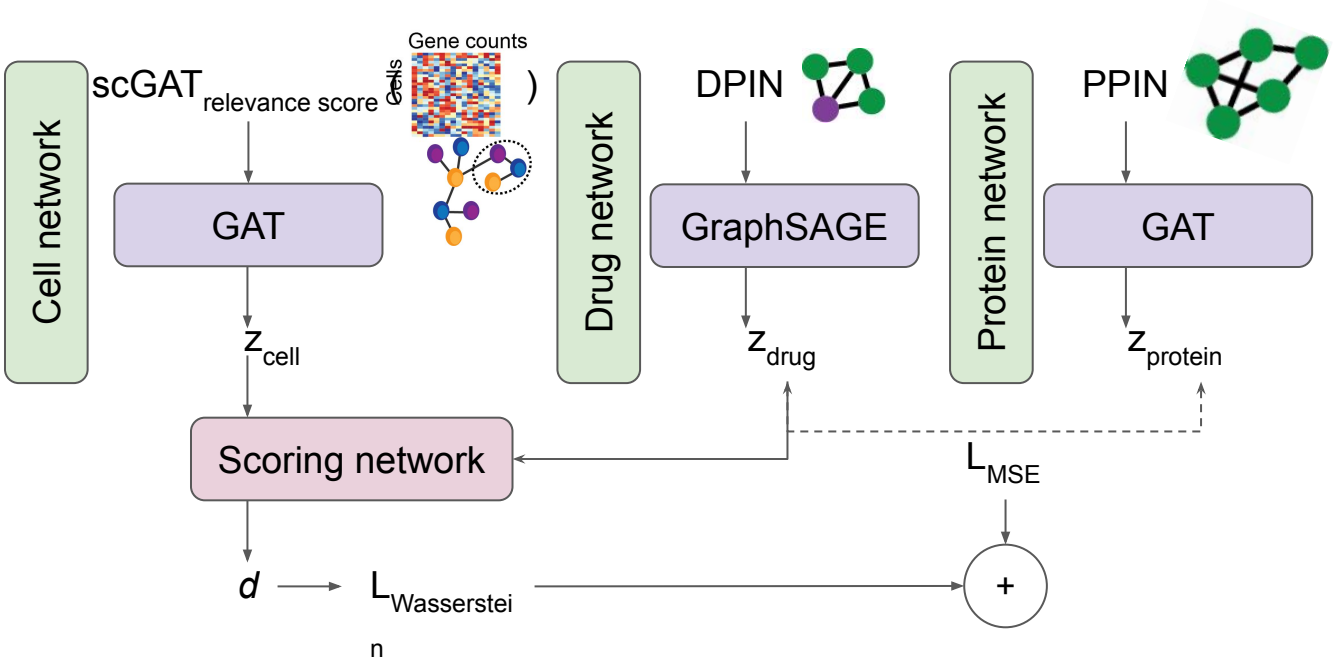






Sehanobish A*, **Ravindra NG***, van Dijk D. Permutation invariant networks to learn Wasserstein metrics. TDA workshop at *NeurIPS'20*

Constraining interpretability for omics-backed hypotheses



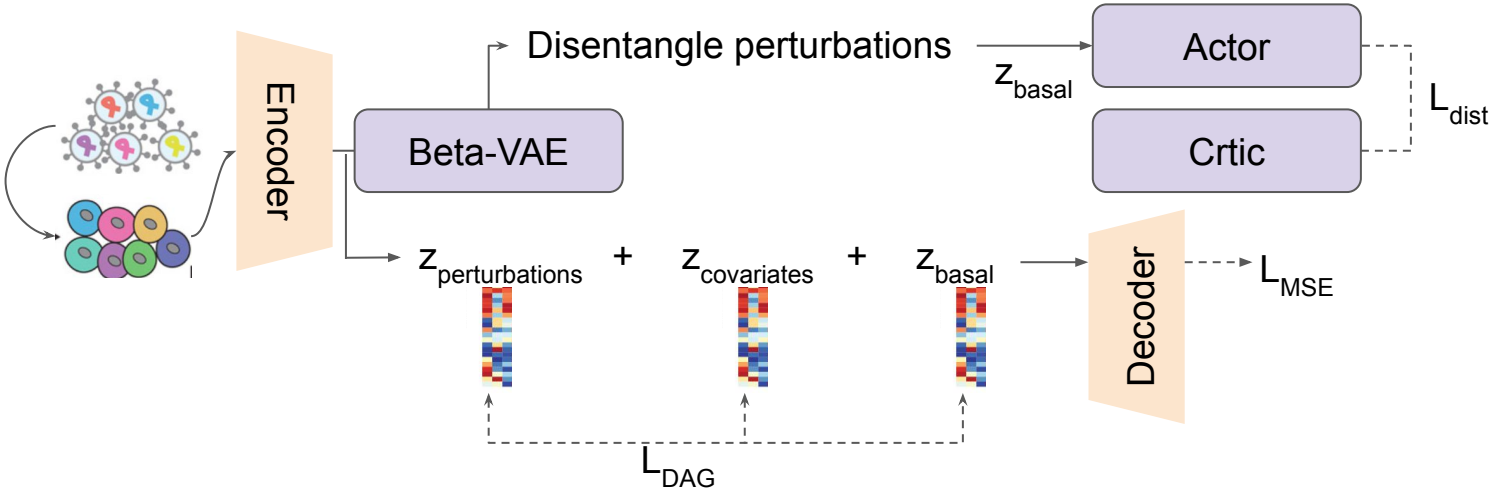

```
Rank 1: certoparin
Rank 2: dilazep
Rank 3: 2Q4
Rank 4: MHPP
Rank 5: chlordecone
Rank 6: AGN-PC-076G3F
Rank 7: acylsaligenin
Rank 8: Phenyl-1,2-Octanedione
Rank 9: adiphenine
Rank 10: AGN-PC-078YTI
Rank 11: nonachlor
Rank 12: ethoxyquin
```

?

Model	(1, 1)
sc2drug (DFA)	0.73 (0.40)
sc2drug (MSE loss)	0.34 (0.03)
sc2drug (Hinge loss)	0.34 (0.03)

Single-cell drug recommender systems: align representations of protein-protein, drug-protein, and cell-gene perturbations to suggest drugs to modulate genes varying across a particular cell subset; NLP to evaluate

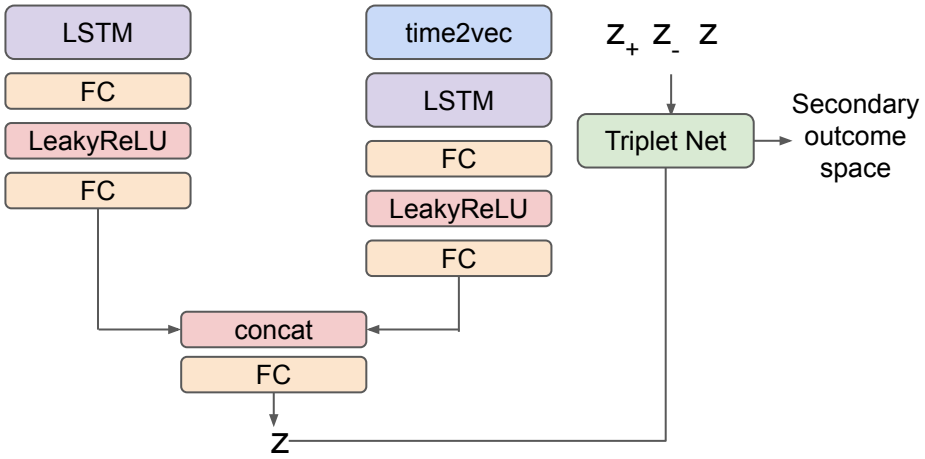
Constraining interpretability for omics-backed hypotheses

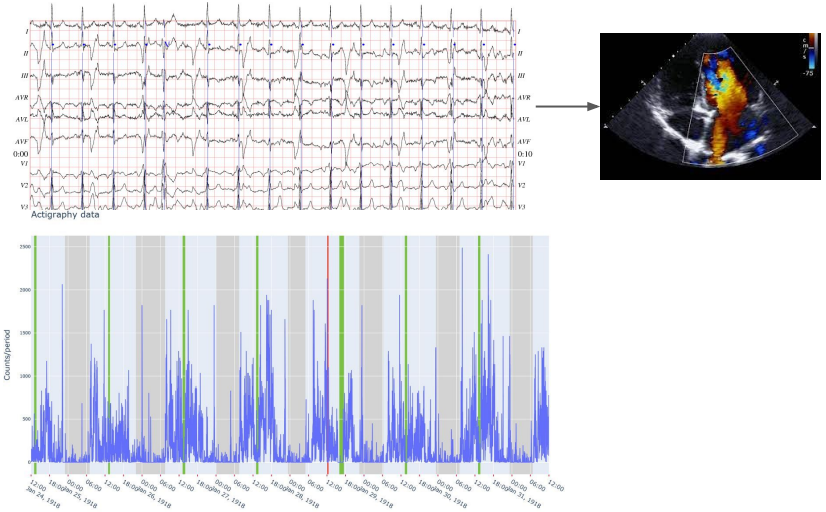


Build on recent developments in DAG learning to infer causal dependencies between perturbed genes, perturbations, and covariates

Future directions

Multi-modal data for phenotyping and substitute testing



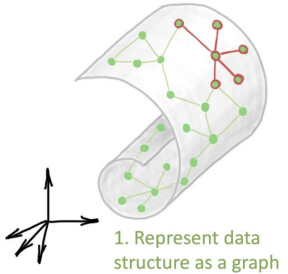


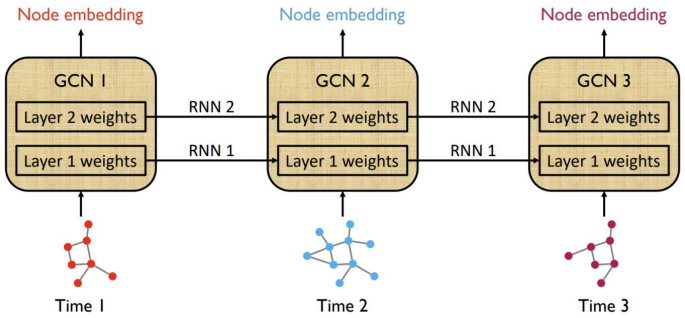
Contrastive pre-training + augmentation to learn representations of 12-lead EKG data that facilitates better prediction of lab values and echocardiogram features on small cohorts (esp. intra-op)

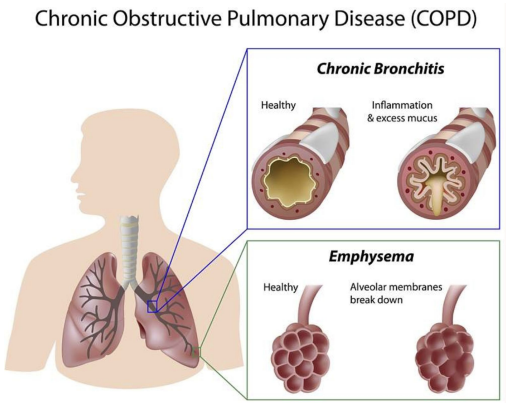
Wearables data to pseudolabel pregnancy, and study associations with metabolomics and immunomics to uncover biological basis for cheap, effective interventions

Future directions

End-to-end XAI for clinical and translational research







Link temporal graph representations with DAG learning to allow for causal feedback loops that are likely present within longitudinal omics datasets

Replaced fixed cell graphs with latent graph learning to improve networks' ability to learn simultaneously from disparate modalities, paired perturbation omics, or multi-dimensional imaging studies

Finding hidden signals: identifying pathophysiological mechanisms or pathways for diseases that have few known genetic risk factors but nonetheless have a familial component (e.g., COPD)

emed Primary Care; Michael Bronstein, *Medium*, 2021; Pareja et al. *AAAI'20*

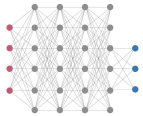
Acknowledgments

van Dijk Lab

Arijit Sehanobish
Victor Gasque
Jason Bishai
Antonio Fonesca
Aagam Shah
Yavuz Nuzumlali

Aghaeepour Lab

Eloise Berson
Joe T.P.
Davide de Francesco
Martin Becker
Shivam Saboo
Rishabh Gupta
Mingze Dong
Juanru Guo
David van Dijk



Nima Aghaeepour
Camilo Espinosa
Samson Mataraso
Ivana Maric

Collaborators

Craig B. Wilen, Yale
Mia Madel Alfajaro, Yale
Janghoo Lim, Yale
Leon Tejwani, Yale
Akiko Iwasaki, Yale
Adrian Haimovich, Yale
Andrew Taylor, Yale
Kristan Studemeyer, Stanford
Mike Snyder, Stanford
Jure Leskovec, Stanford
Trevor Hastie, Stanford
Stephanie C. Eisenbarth, Yale
Anna M. Pyle, Yale
Tamas L. Horvath, Yale
Bao C. Wang, Yale
Ellen F. Foxman, Yale
Richard W. Pierce, Yale
Tariq Ahmad, Yale
Nihar Desai, Yale
Erik Herzog, WashU
David Stevenson, Stanford
Gary Shaw, Stanford

Code: github.com/nealgravindra

NIH

Stanford Yale SCHOOL OF MEDICINE