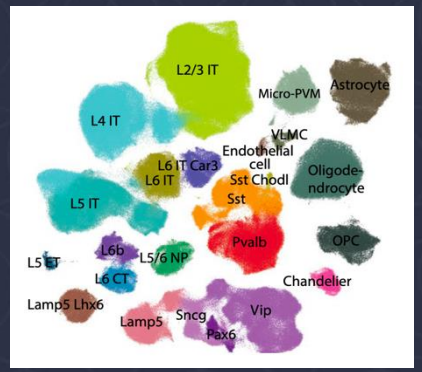
ALLEN INSTITUTE *for*
BRAIN SCIENCE

scVIP: personalized modeling of single-cell transcriptomes for Alzheimer's Disease phenotypes

Hsin-Yu (Jane) Lai

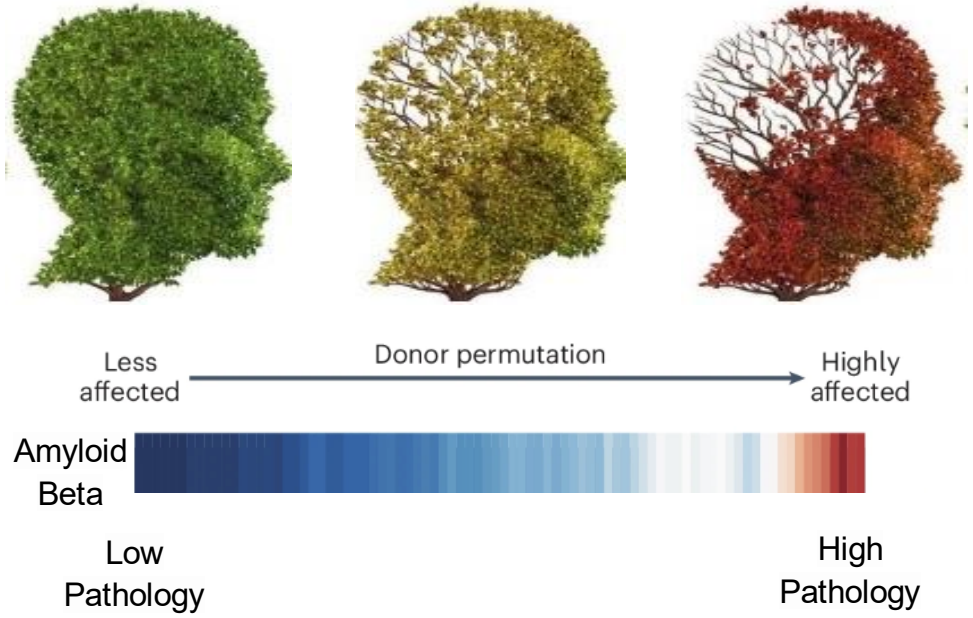
HBA-AI Team



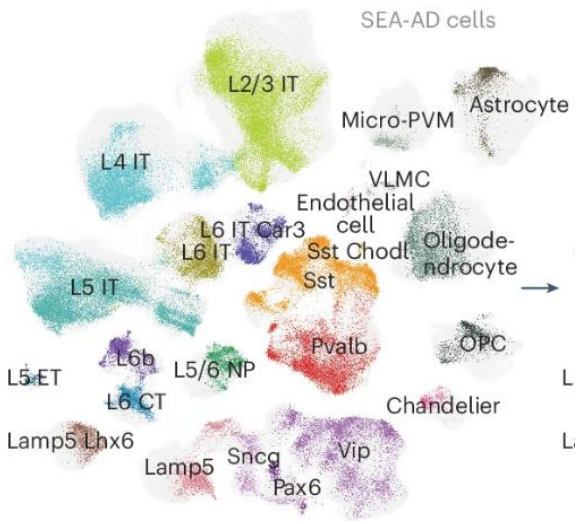


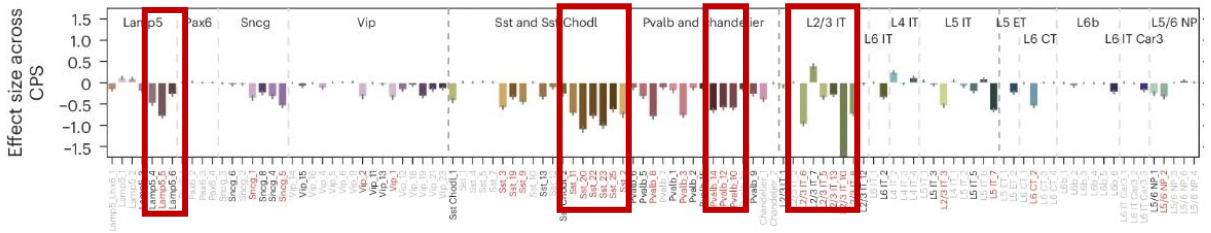


We study the cellular consequences of AD



Detailed Cellular Taxonomies





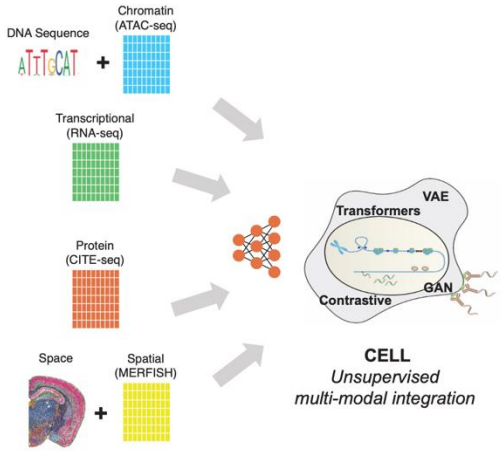
SEA-AD studies Alzheimer's Disease and has shifted our understanding of disease progression from a focus on neuropathology to cellular vulnerability.

Our regular analysis pipelines mostly pool data across subjects

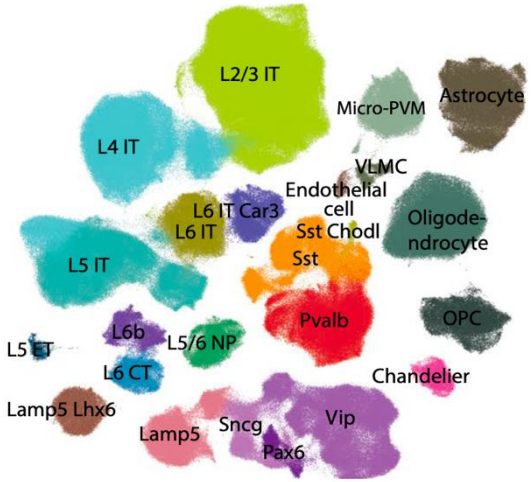
SEA-AD: 84 donors



Unsupervised models of cellular representation



Cell type annotations



Multiple downstream analyses of how cell types are differentially affected by diseases.

Heterogeneity in the data may be obscured.
Disease- and non-disease-associated cells may be mixed.

We built an ML model to satisfy the desired criteria

Desiderata		Next-generation ML model
Keeps donor heterogeneity	→	Personalized
Includes pathology information	→	Semi-supervised
Models cellular vulnerability	→	Cell-type aware
Is interpretable	→	Generative

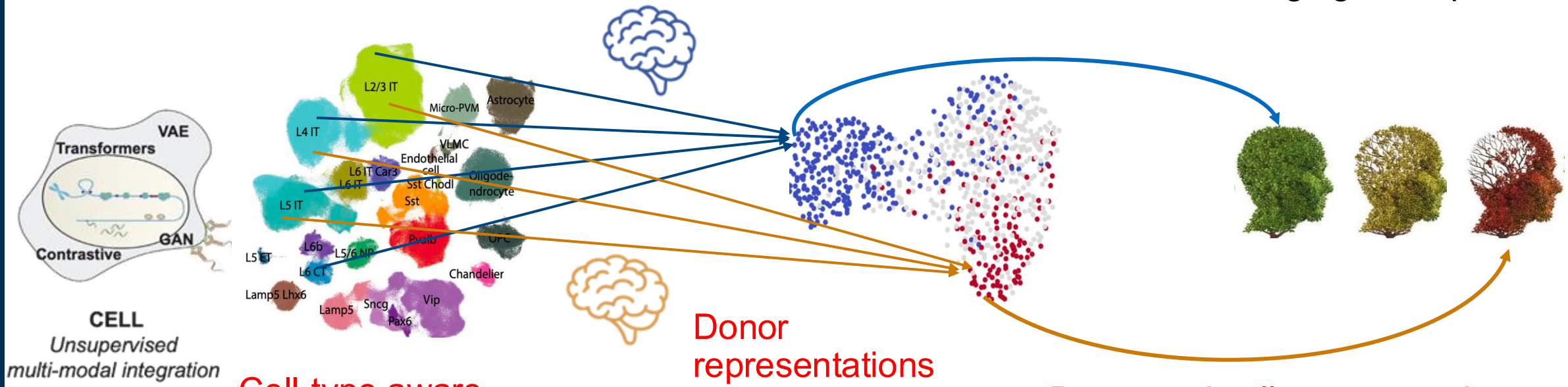
We developed an ML model that satisfies all desiderata

The statistical model follows biological insights (in a few slides)

Group cells from the same donor

Learn a donor representation

Predict disease staging/neuropath



Cell-type aware aggregation models cellular vulnerability (in a few slides)

Donor representations maintain donor heterogeneity

Donor and cell representations include pathology information

We developed an ML model that satisfies all desiderata



scVIP

scVIP achieves:

- Predictions of pathologies from scRNA-seq data
- Integrations and harmonizations across multiple AD cohorts
- Streamlined analyses of disease-associated cell types and genes

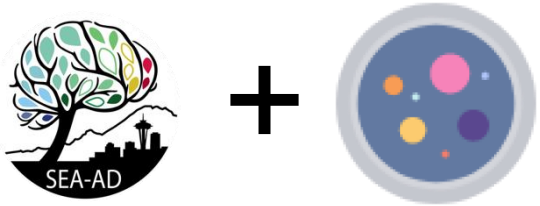
Outline



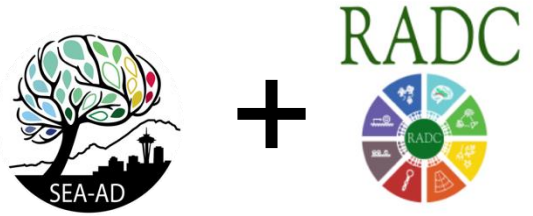
scVIP model



Debug: Mouse visual cortex development



SEA-AD + AMP-AD caudate analysis

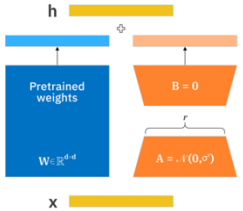


SEA-AD + ROSMAP multi-regional analysis

Bonus track



DREAM challenge



Enabling large-scale model training

Outline

Keeps donor heterogeneity

Models cellular vulnerability



scVIP model

Includes pathology information

Is interpretable



Debug: Mouse visual cortex development



SEA-AD + AMP-AD Caudate Analysis



SEA-AD + ROSMAP Multi-regional Analysis

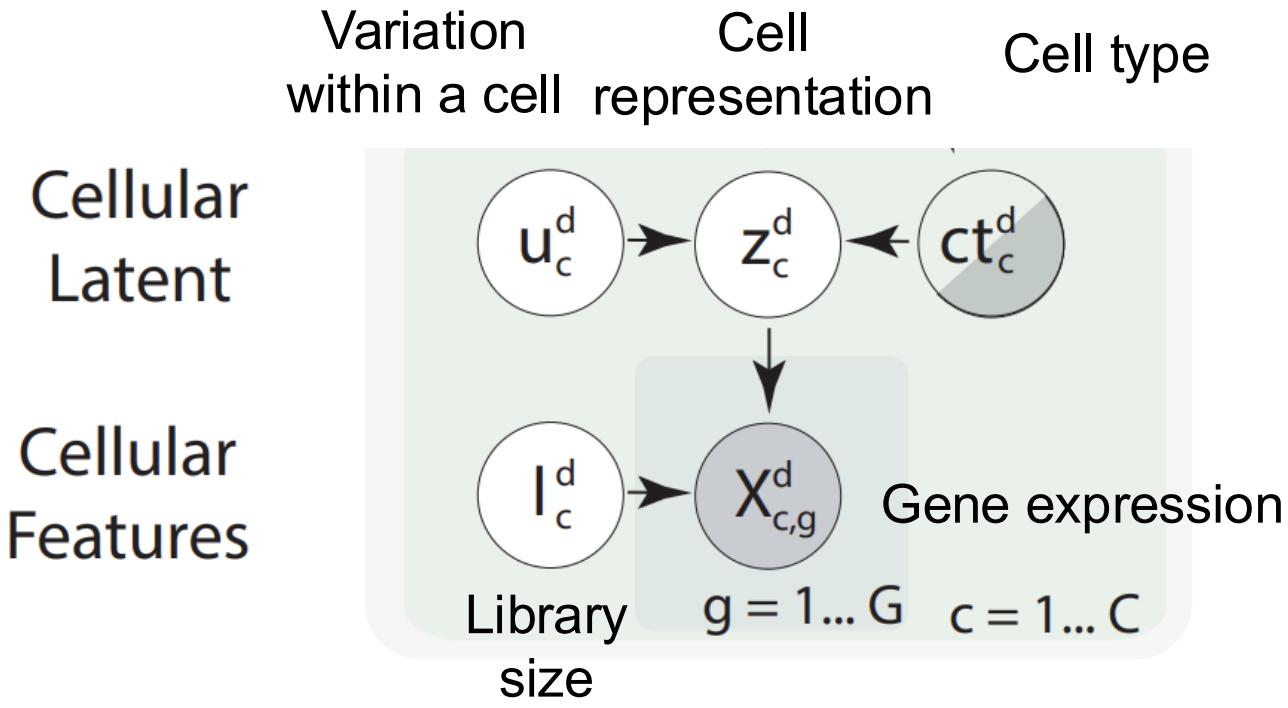
Bonus Track



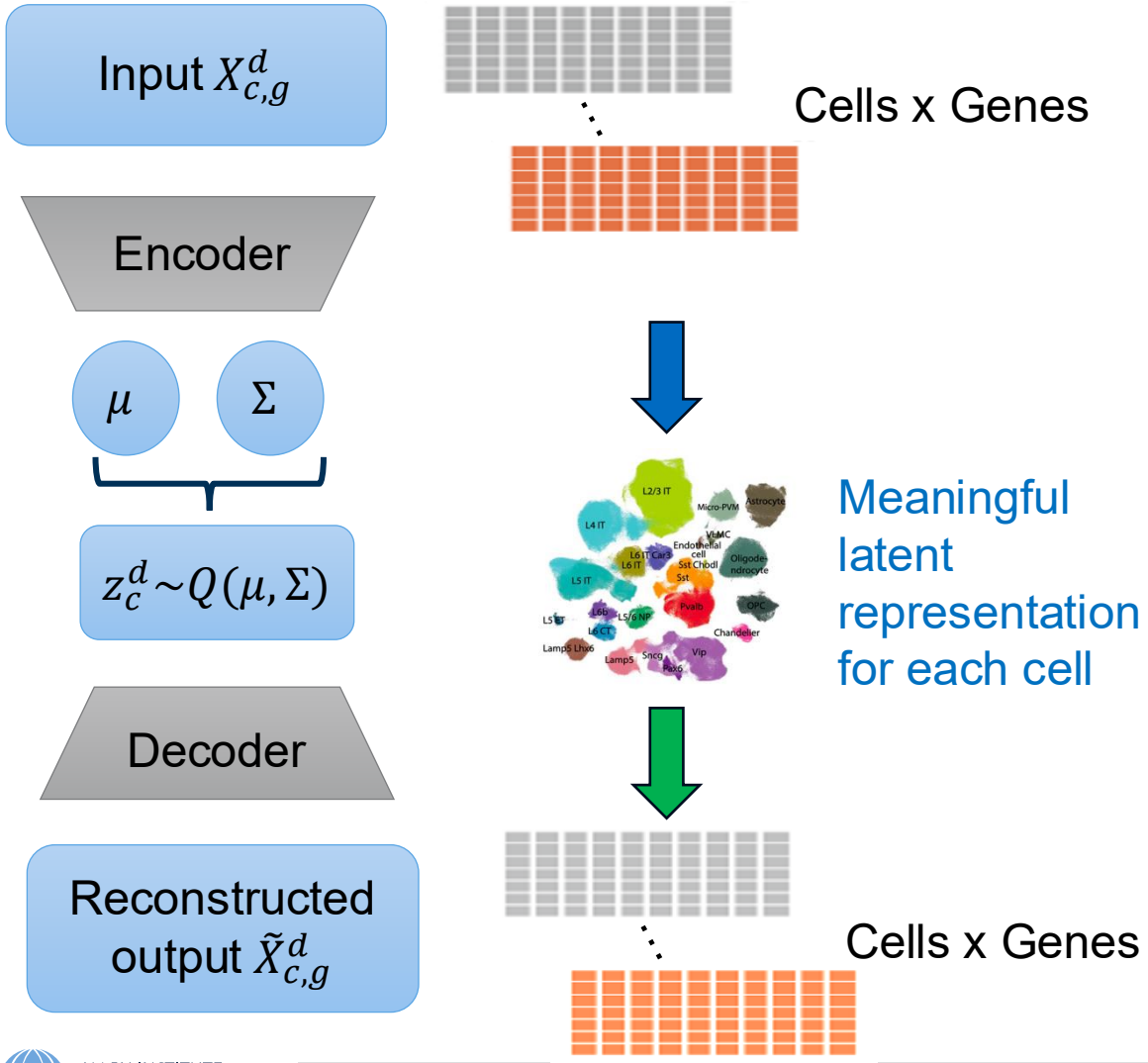
DREAM Challenge

Enabling Large-Scale Model Training

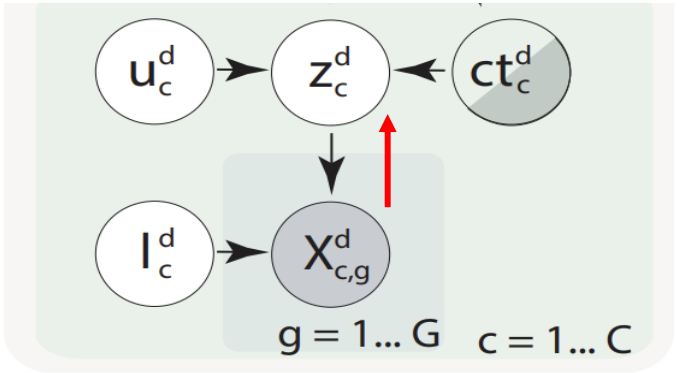
Existing interpretable cell models leverage a statistical model



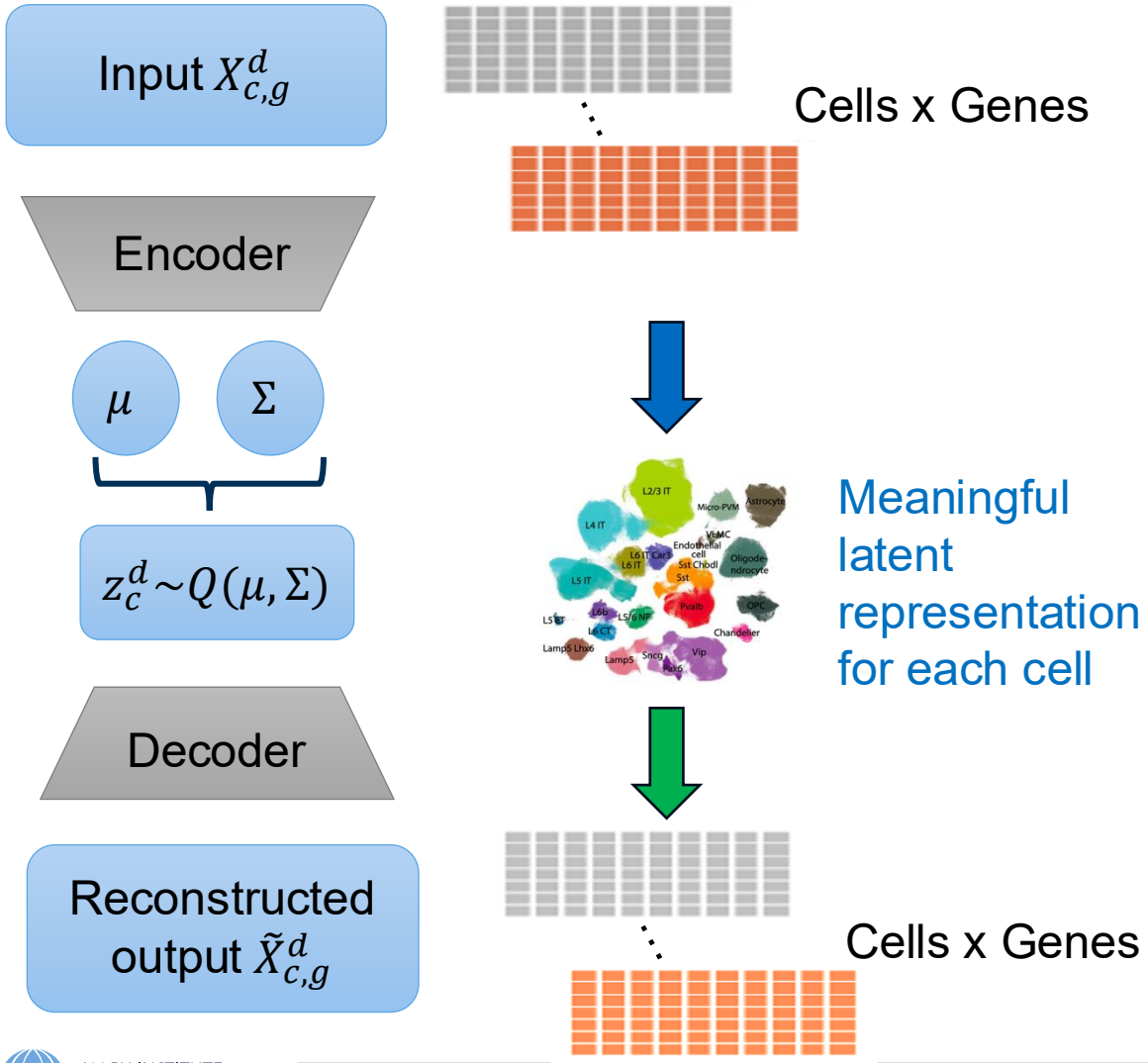
Variational autoencoder (VAE) is a powerful tool to learn a complex statistical model



Encoder aims to infer latent variables from observed variables using variational inference



Variational autoencoder (VAE) is a powerful tool to learn a complex statistical model



With a representation for each cell, we can…

- Annotate its cell type
- Impute missing modalities
- Understand the relationship between genotypes and cell types

We want to ask, can we learn a representation for each donor, so that we can

- Annotate a donor's disease staging
- Impute missing quantitative neuropathology measurements
- Understand the relationship among genotypes, cell types, and phenotypes

Extend existing model to include donor heterogeneity, pathology, and cellular vulnerability

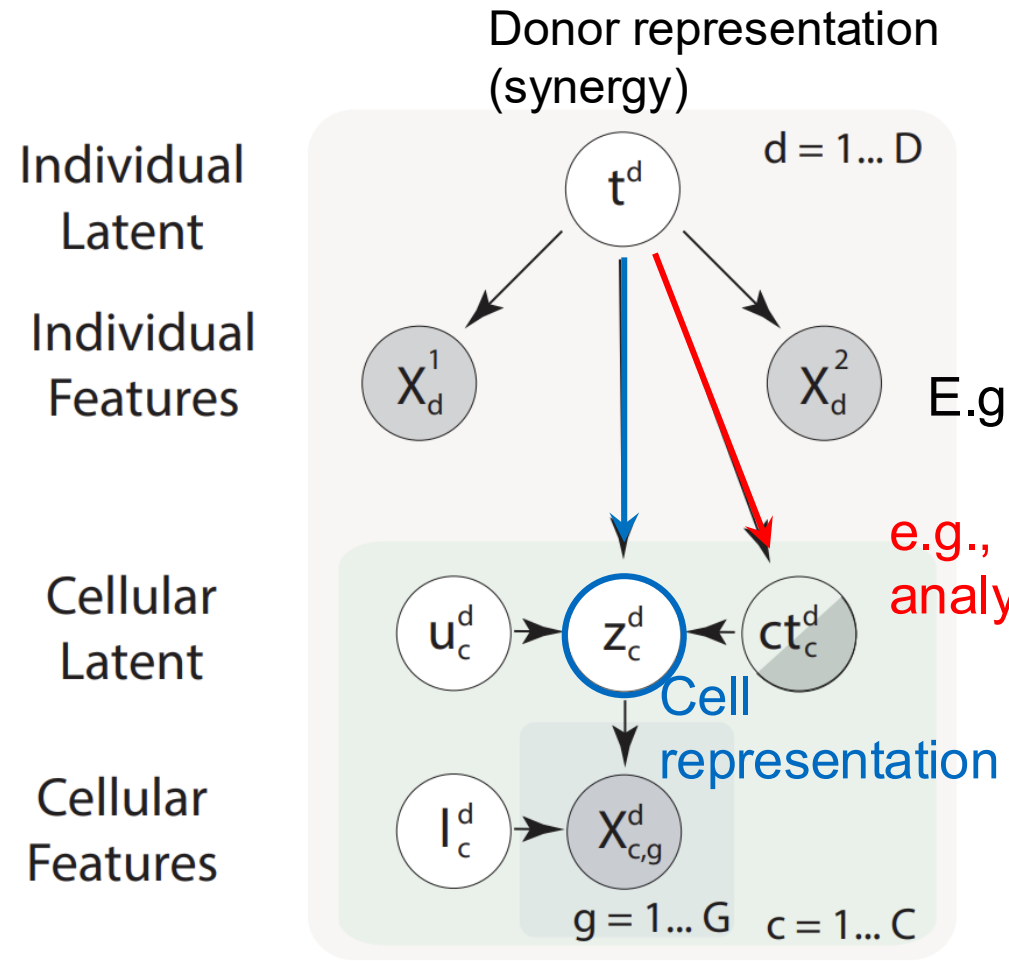
This model:

Keeps donor heterogeneity

Models cellular vulnerability

Includes pathology information

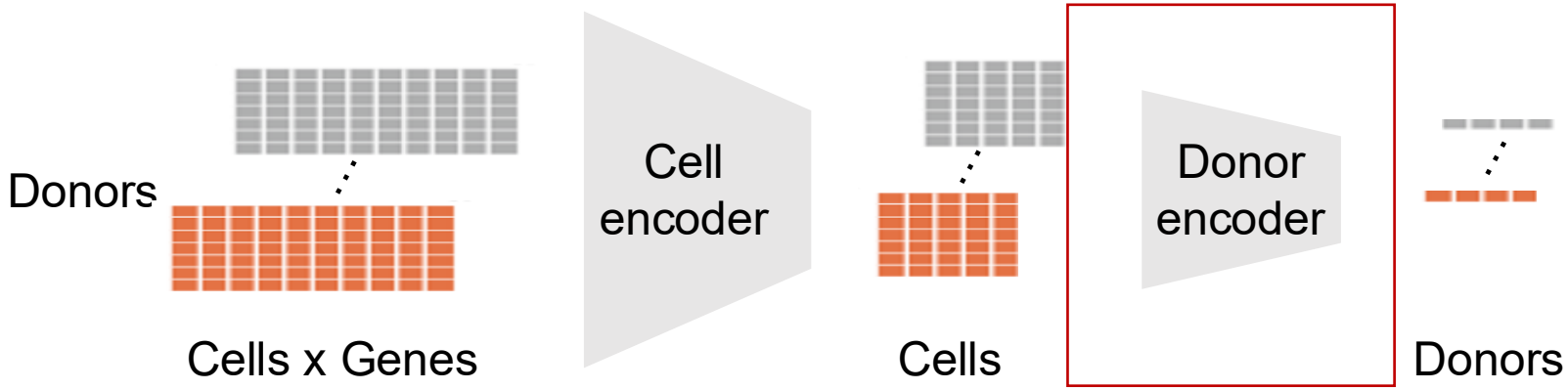
Is interpretable

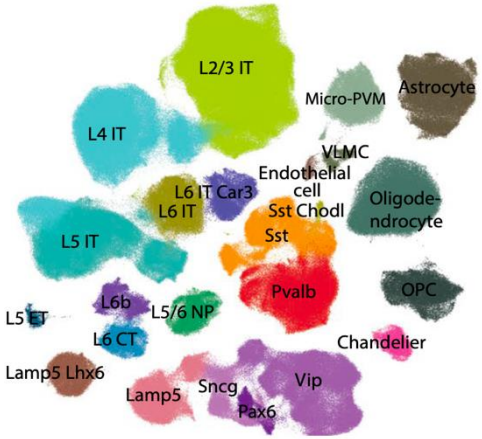


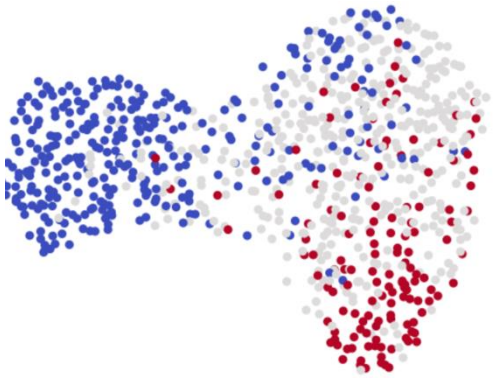
E.g., Staging, quantitative neuropath

e.g., scCODA, the model used to analyze SEA-AD MTG

The donor encoder uses cell-type-specific multi-head attention mechanisms to create one feature vector per donor



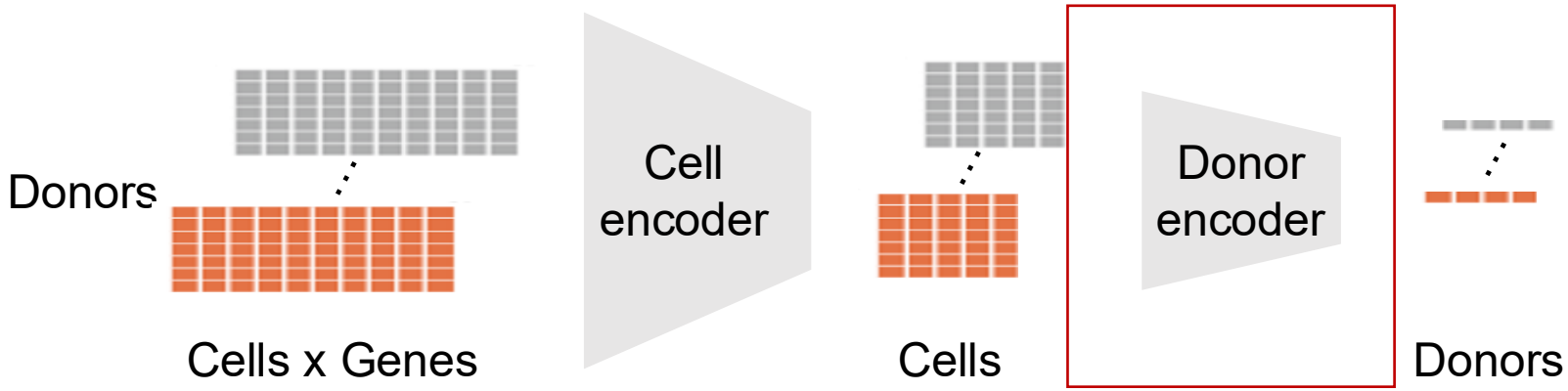




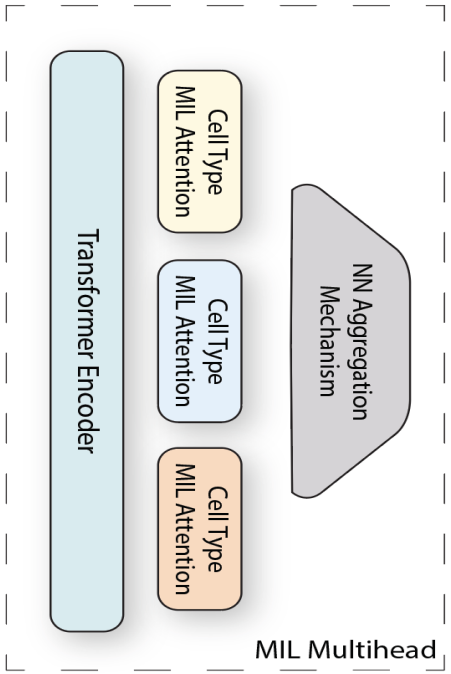
Cell representations
(Each dot is one cell. The colors are the different cell types.)

Donor representations
(Each dot is one donor. The colors are the ADNC scores.)

The donor encoder uses cell-type-specific multi-head attention mechanisms to create one feature vector per donor



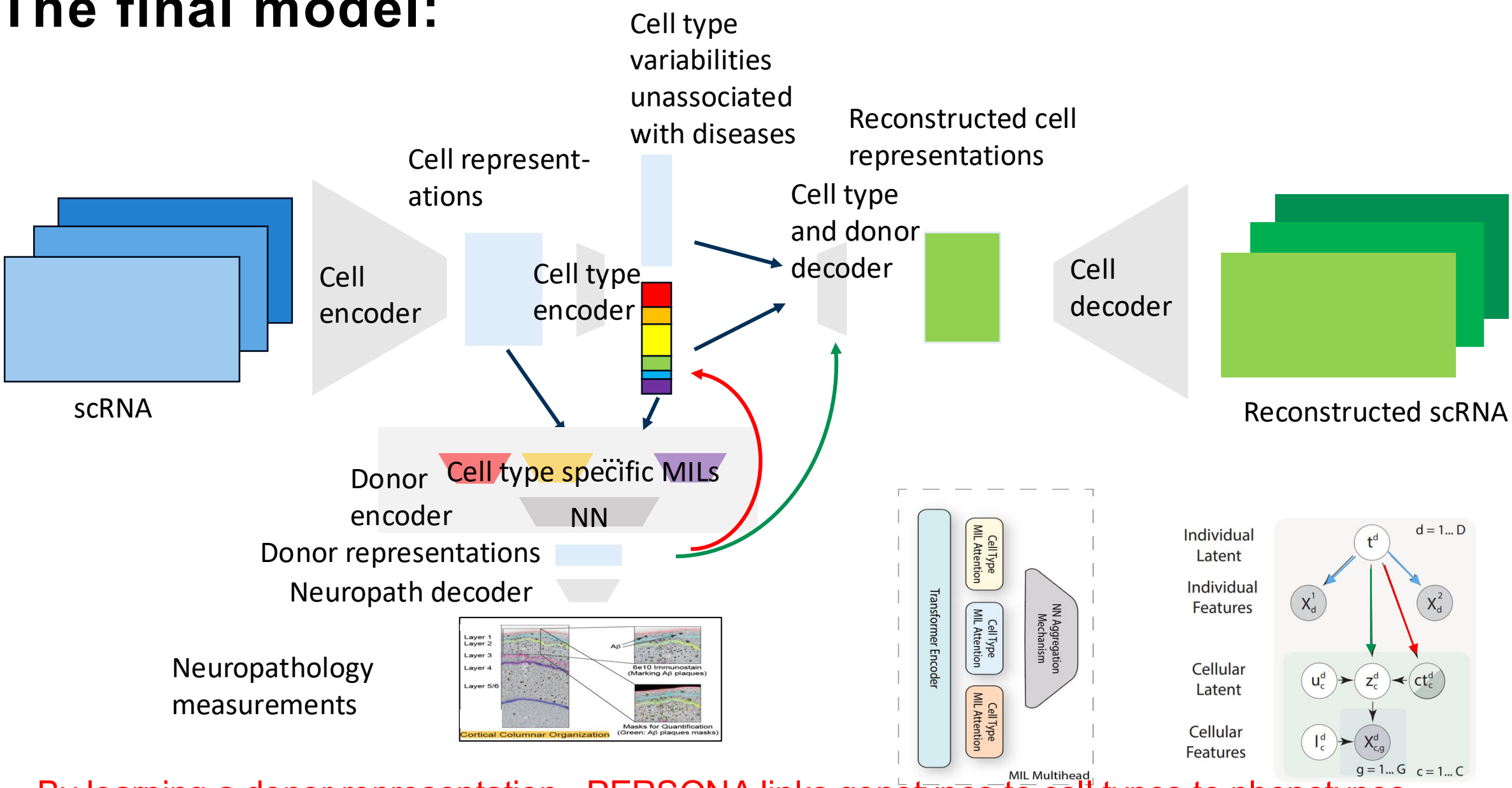
It further models cellular vulnerability.



Novel multihead self- attention

The encoder harnesses disease-associated changes in each cell type to learn a donor's latent representation

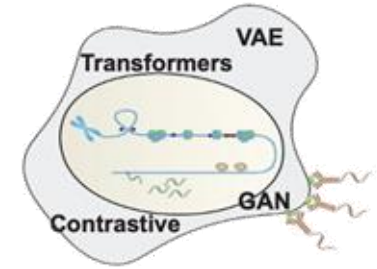
The final model:



By learning a donor representation, PERSONA links genotypes to cell types to phenotypes.

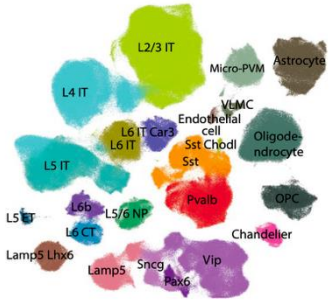
Training procedure of scVIP is simple and flexible

Train cell representations using any cell model



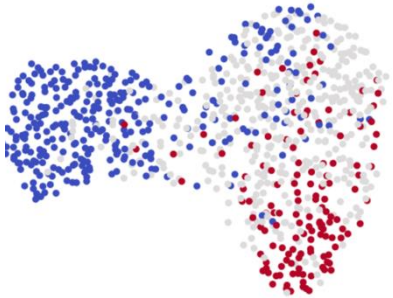
CELL

Unsupervised multi-modal integration



Finetuning the model with PERSONA and phenotypes





scVIP can be easily incorporated into the current analysis pipeline.

Outline



Debug: Mouse visual cortex development

Persona model

SEA-AD + AMP-AD Caudate Analysis

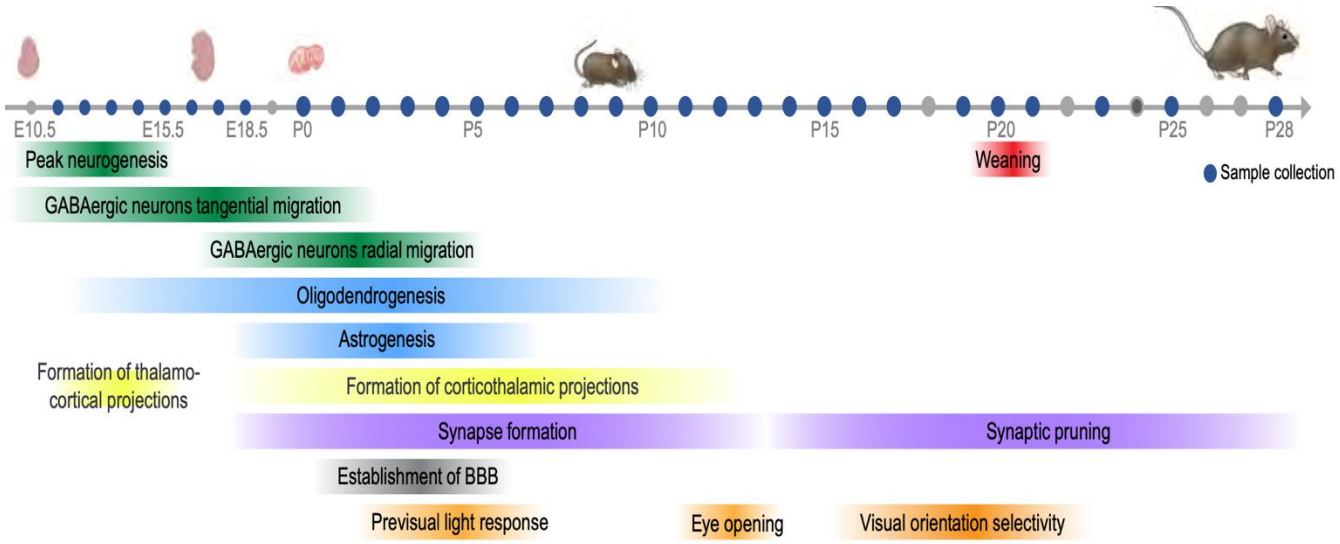
SEA-AD + ROSMAP Multi-regional Analysis

Bonus Track

DREAM Challenge

Enabling Large-Scale Model Training

Mouse development as a step to debug scVIP model and predictions



Subclass



Age



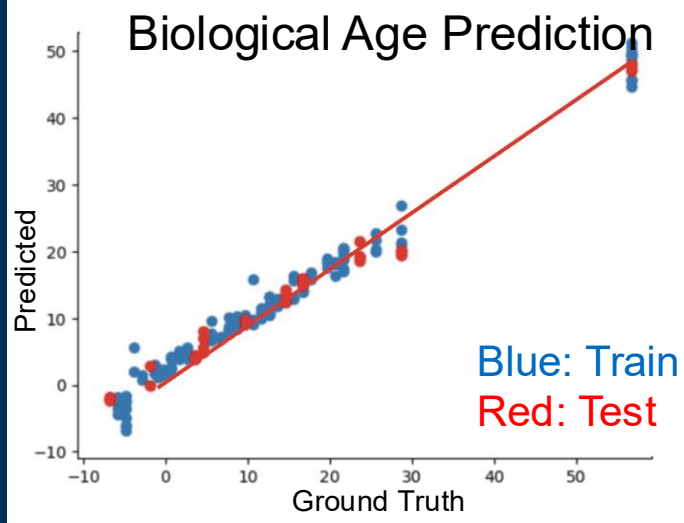
Data:
53 mice, 0.5M cells, 32k genes, embryonic day 11.5 to postnatal day 56

Phenotype: age

[1] Gao et al. “Continuous cell type diversification throughout the embryonic and postnatal mouse visual cortex development”

Goal: understand the model’s ability to learn phenotype, cell-type, genotype info

Phenotype: scVIP achieved comparable age predictions



Model comparisons (R2)

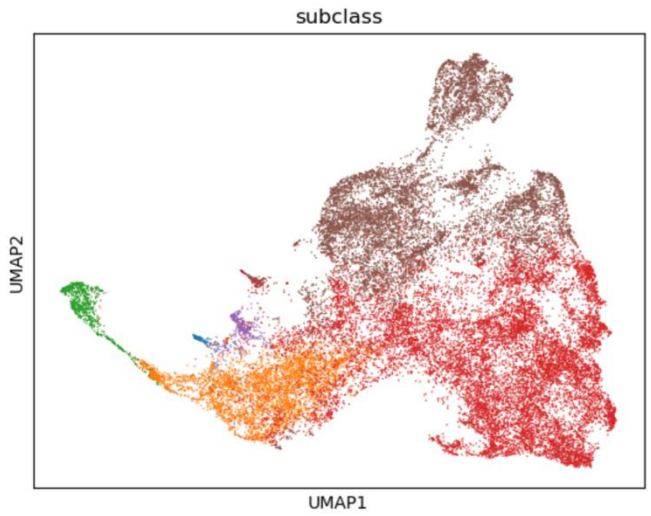
	Method A: From gene expression	Method B: From scVI latent rep	Method C: MIL without cell type specific heads	Method D: MIL with cell type specific heads
Test	0.91	0.94	0.94	0.98

MIL: Multi-instance learning; here MIL is referring to the architecture from cell representations to age prediction

Observations:

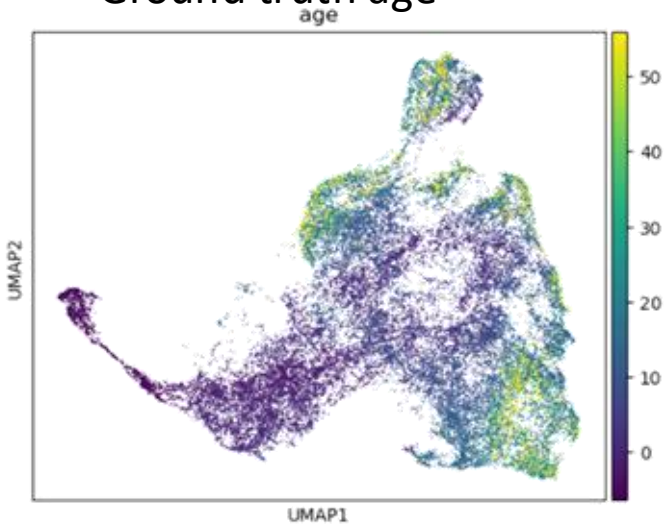
- B, C, D > A: Cell representations denoise irrelevant information in gene expressions
- C, D > B: Cell representations guided by age prediction can predict age better
- D > C: Cell-type-specific heads improve the result

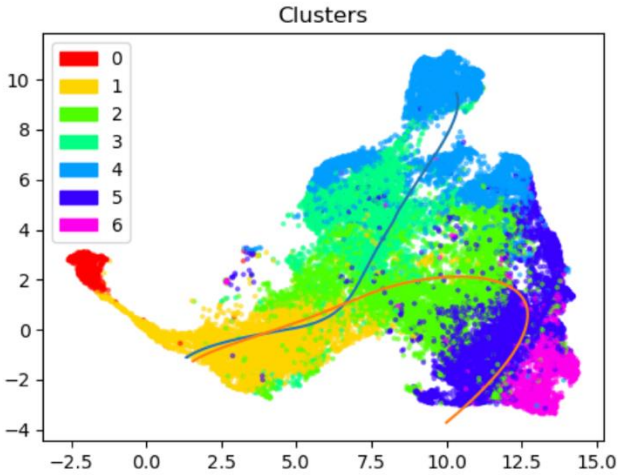
Cell type: The CTX-MGE GABA cell type UMAP shows cell representations in the developmental order

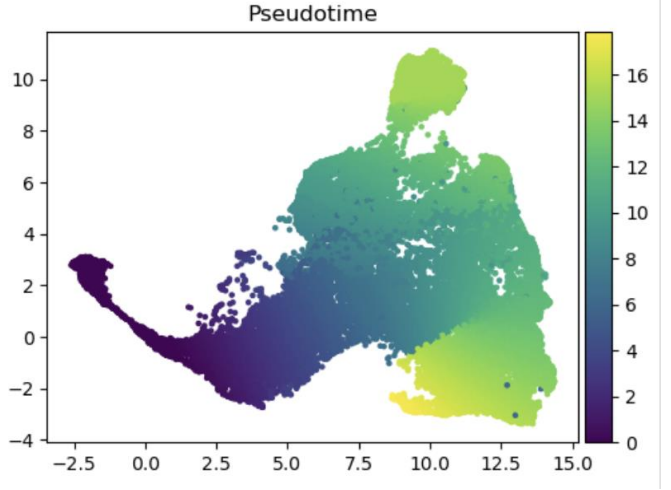




Ground truth age

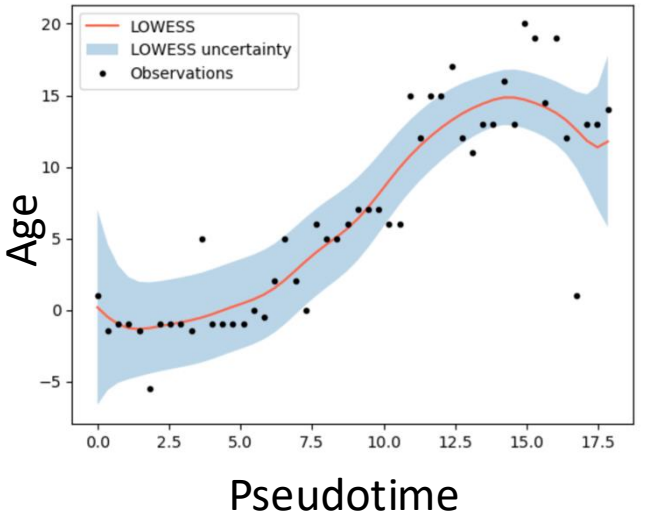






Persona arranges cells according to the ground truth age.

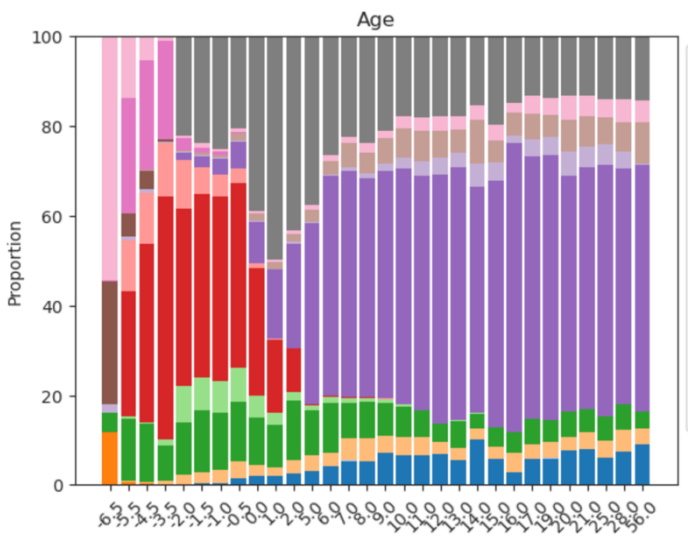
Age over pseudotime



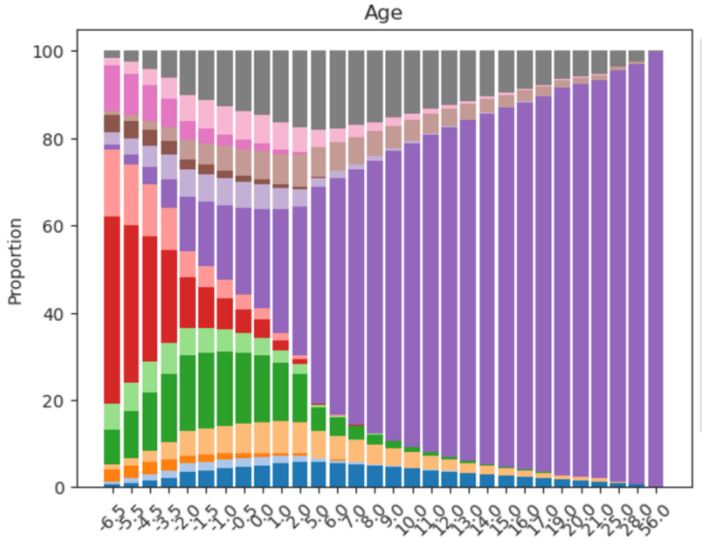
Pseudotime

Cell type: Similarity of cell type abundance over age between ground truth and learned by scVIP

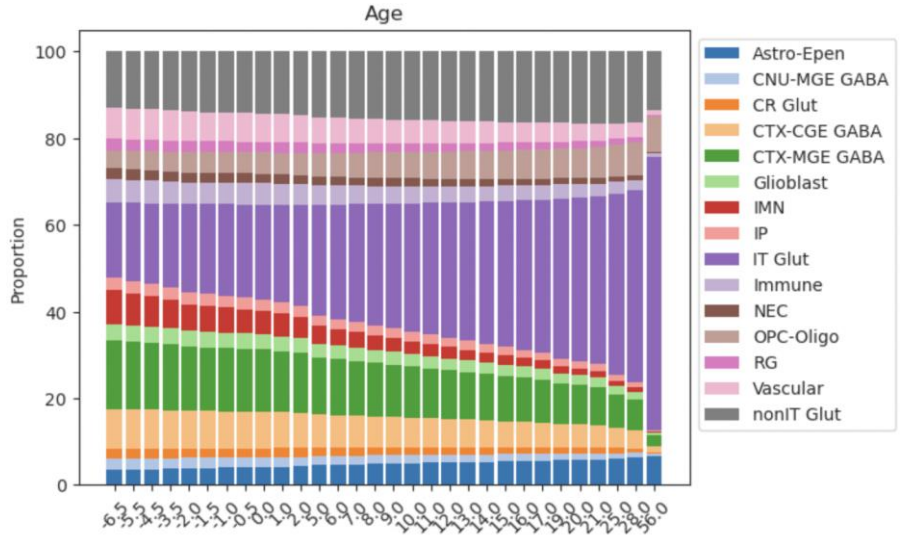
Ground truth



Predicted by our model



Linear model used in MTG-SEA-AD

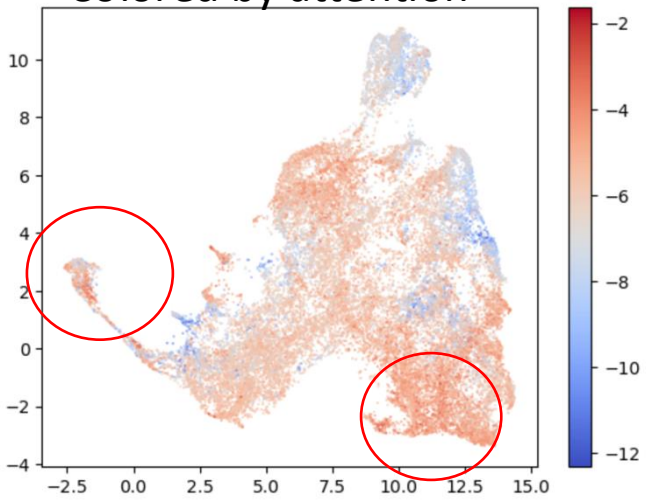


PERSONA can capture nonlinear trends in cell type abundance changes.

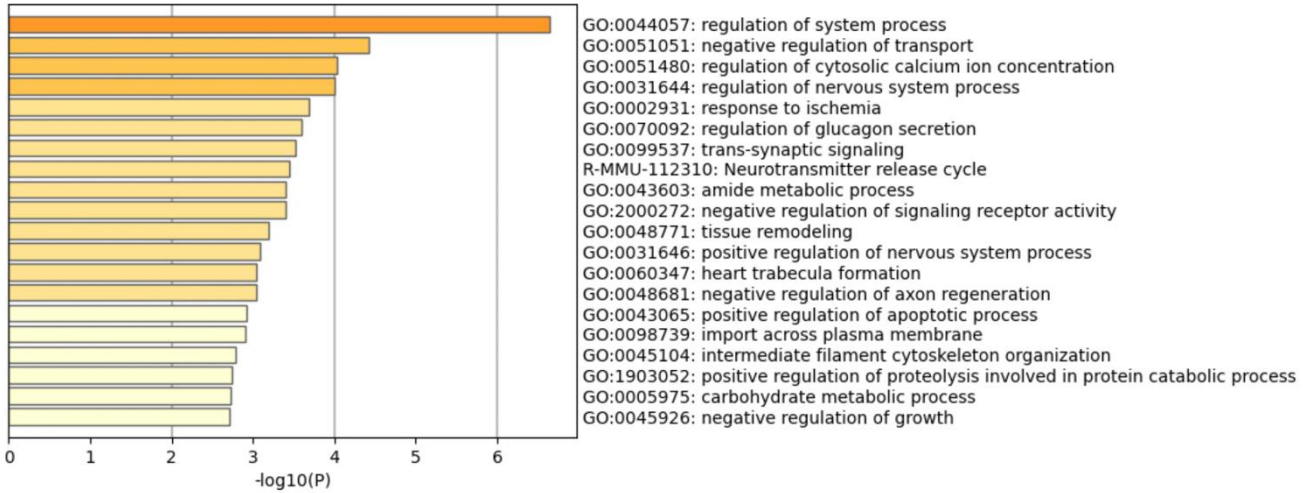
Genotype: Differential analysis between early and late cells with high attention shows upregulated genes versus downregulated genes

Identify two cell sets to run differential analysis

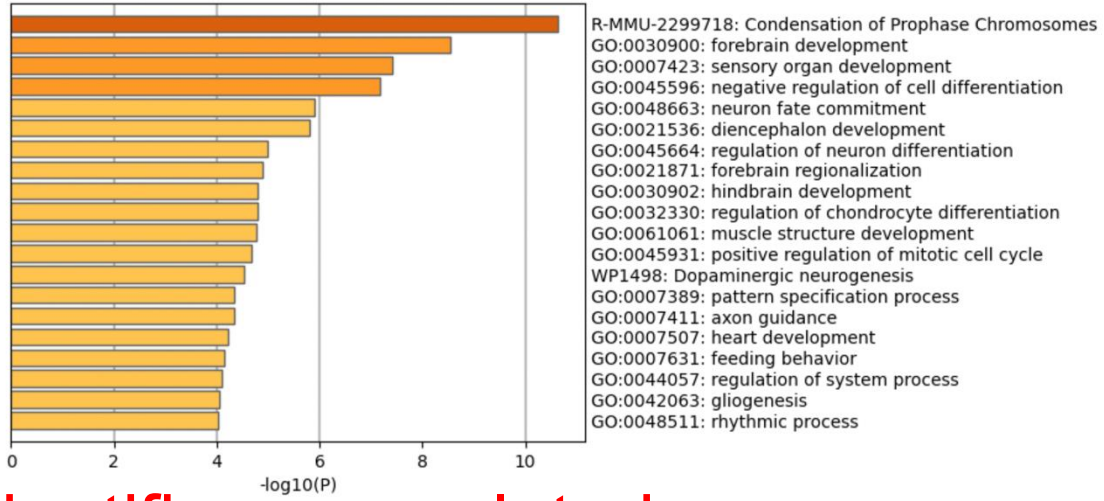
UMAP of CTX MGE-GABA
Colored by attention



Upregulated genes (neurotransmitting, synaptic signaling)



Downregulated genes (Development genes and cell differentiation)



scVIP identifies age-related genes

Outline

Persona model

Debug: Mouse visual cortex development

SEA-AD + ROSMAP
Multi-regional Analysis

SEA-AD + AMP-AD Caudate Analysis

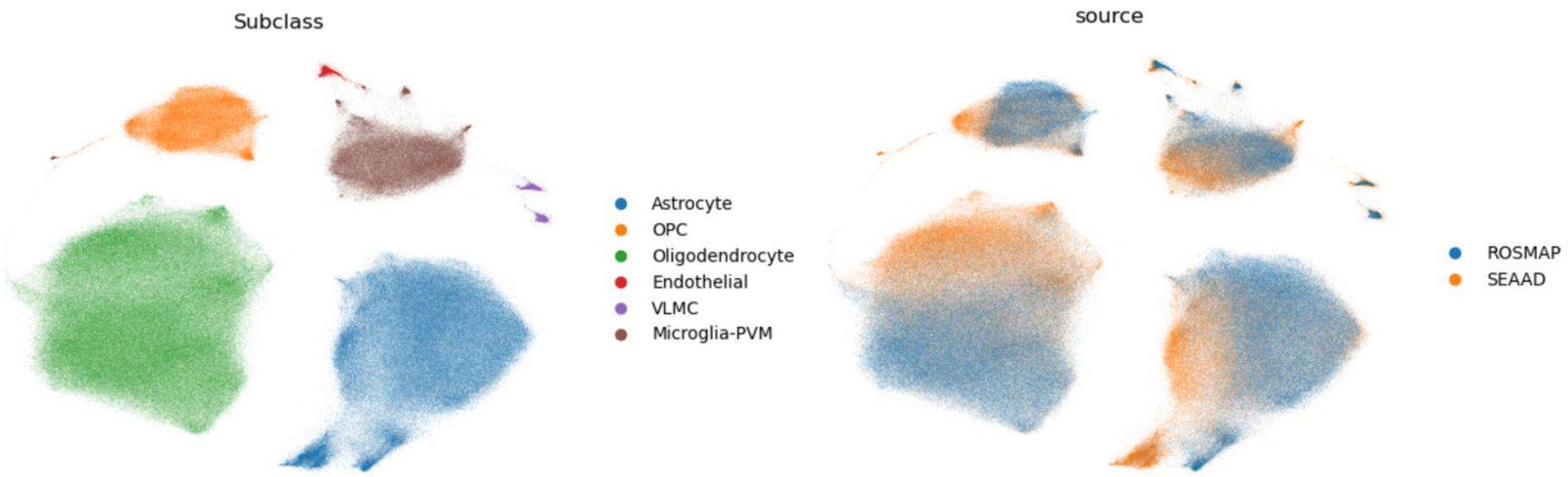
Bonus Track

DREAM Challenge

Enabling Large-Scale Model Training

SEA-AD + ROSMAP multi-region as a dataset to evaluate harmonization



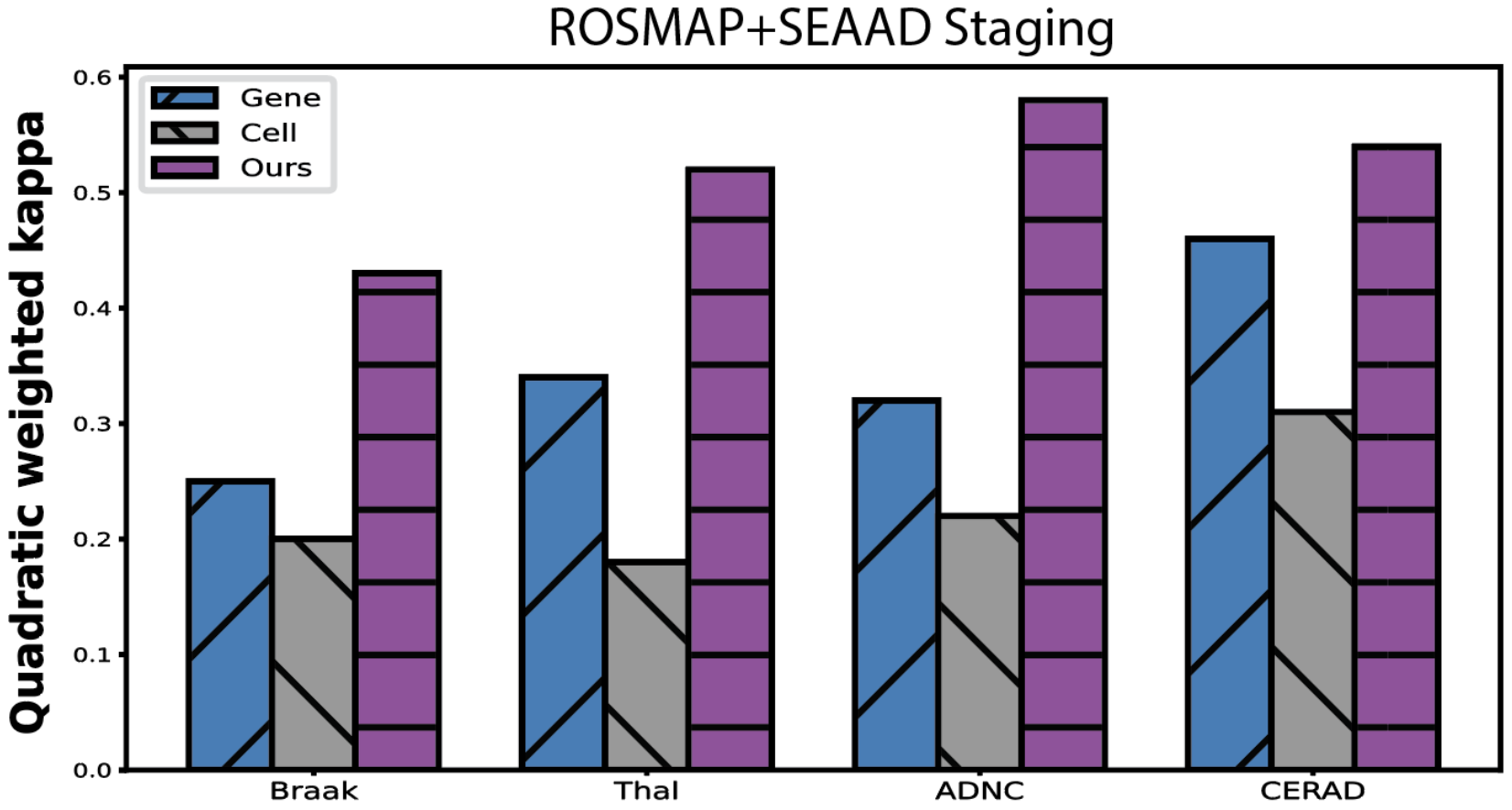


Datasets:
ROSMAP: 0.8M glia cells, 5k genes, 424 donors, DLPFC
SEA-AD: 0.2M glia cells, 5k genes, 84 donors, MTG + A9

Phenotypes:
- Shared: Braak, Thal, ADNC, CERAD
- ROSMAP: neuritic plaque, diffuse plaque, tau density
- SEA-AD: percent 6e10/ AT8

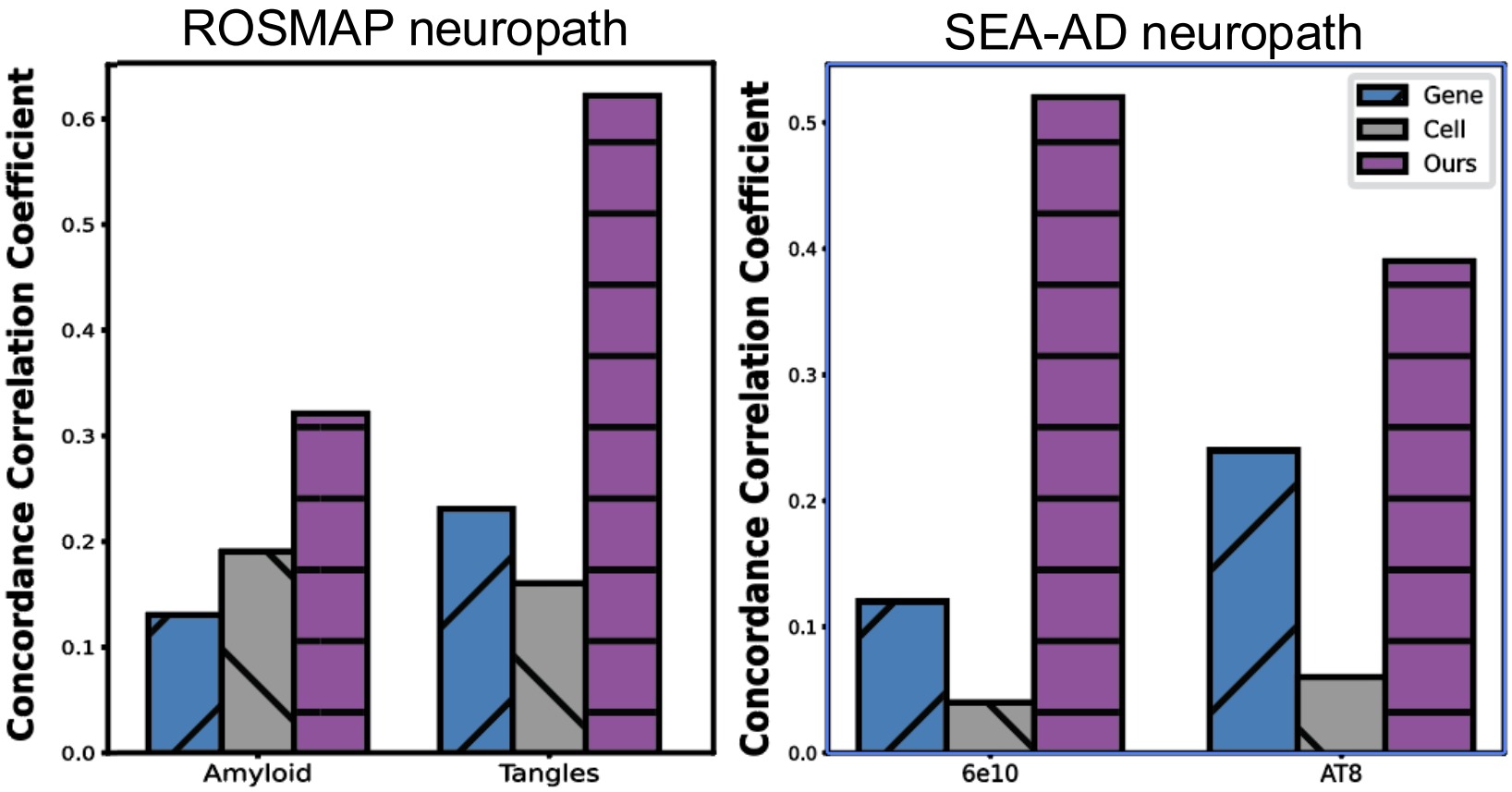
Goals:
- Understand the model's ability to learn phenotype, cell-type, genotype info
- Show harmonization efforts

Phenotype: Comparing scVIP predictions with existing methods



Quadratic weighted kappa (QWK): Measure agreement between two raters on an ordinal scale.

Phenotype: Comparing scVIP predictions with existing methods



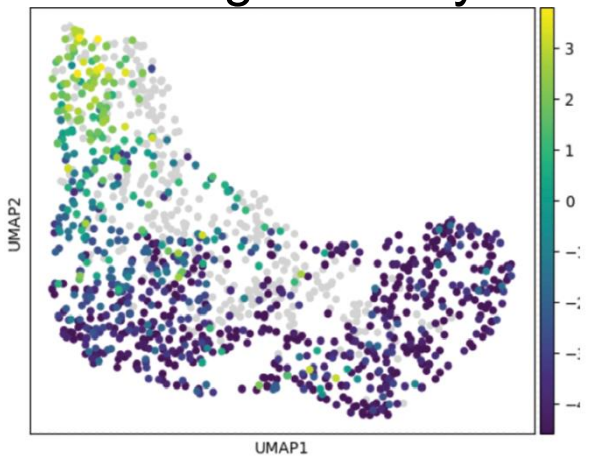
Concordance correlation coefficient (CCC): Measure agreement between two raters on a continuous scale.

scVIP outperforms existing methods.

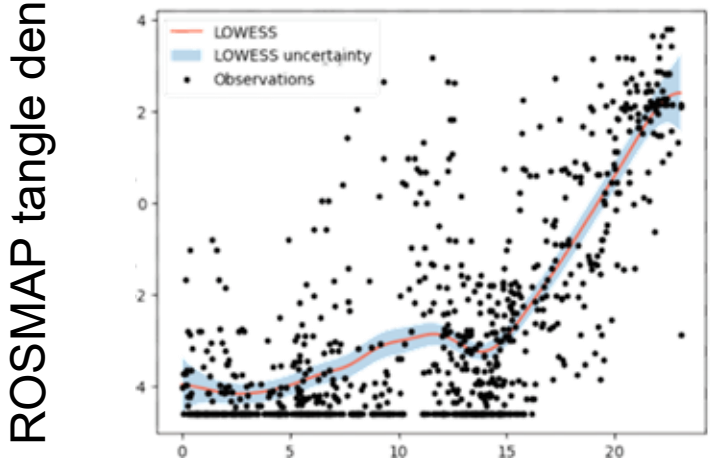
Phenotype: With the UMAP of donors from both ROSMAP and SEA-AD datasets

ROSMAP donors and measurements

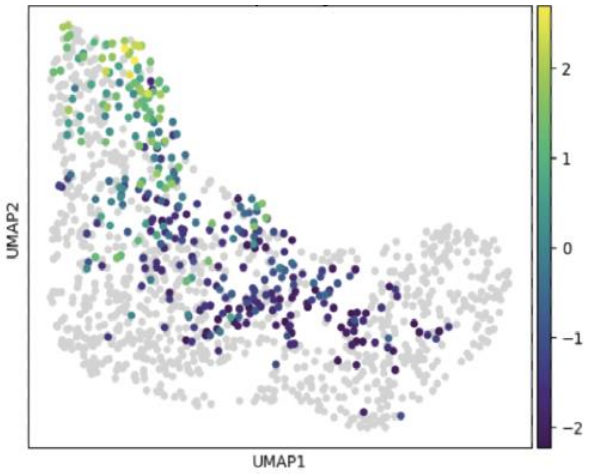
Donor UMAP colored with tangle density

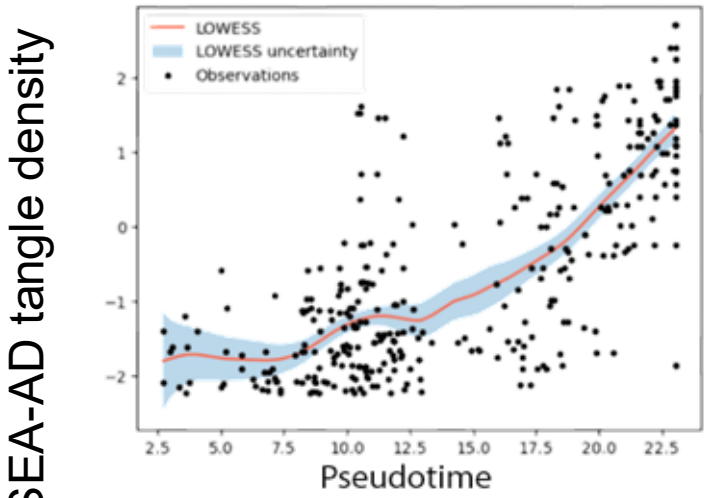


Tangle density over pseudotime learned from donor UMAP



SEA-AD donors and measurements

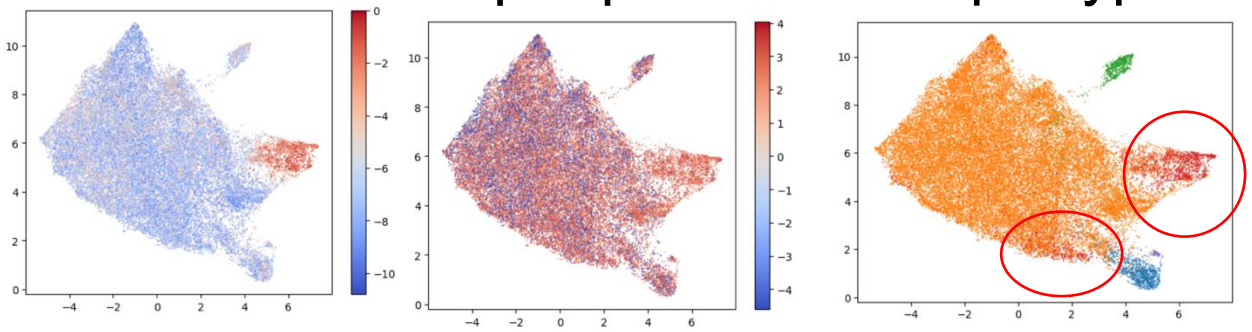




scVIP achieves harmonization

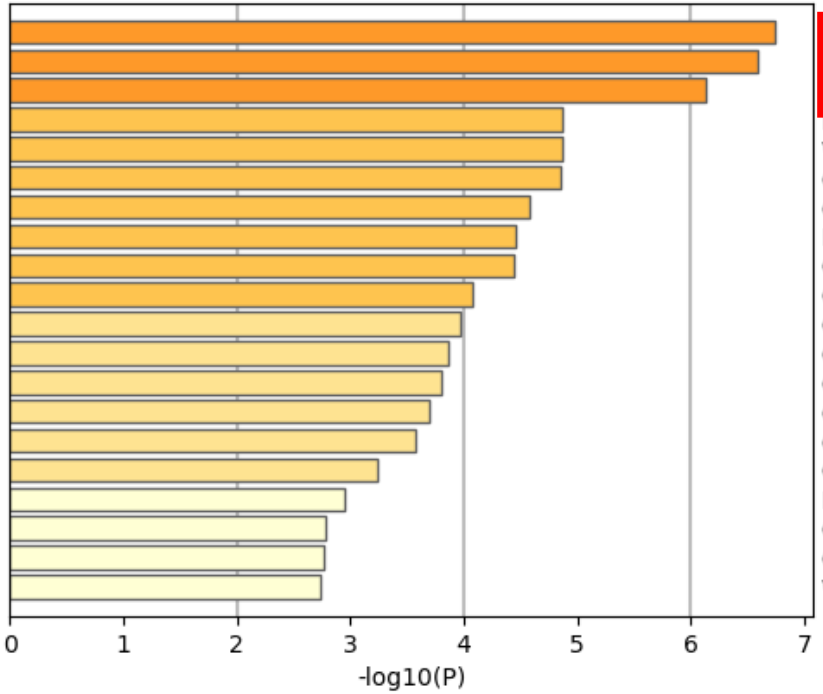
Cell type/Genotype: scVIP identifies micro-PVM_3 as a disease-associated cell type and identifies potentially involved genes

attention plaque supertype



All: Micro-PVM UMAPs
Red: Micro-PVM_3-SEAAD

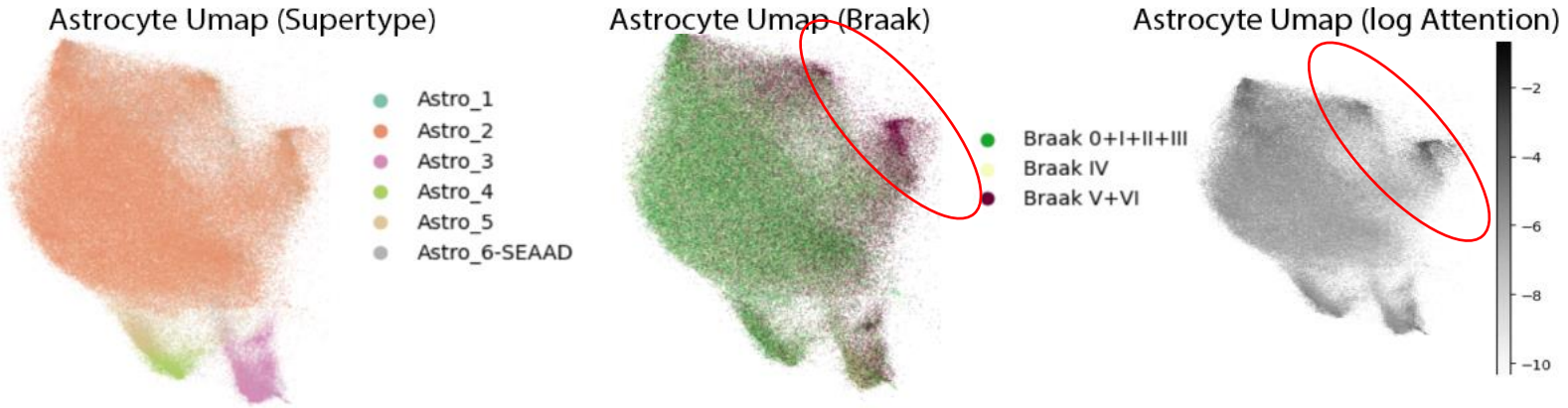
Perform Bayesian differential analysis on high-attention versus low attention in Micro-PVM 3-SEAAD

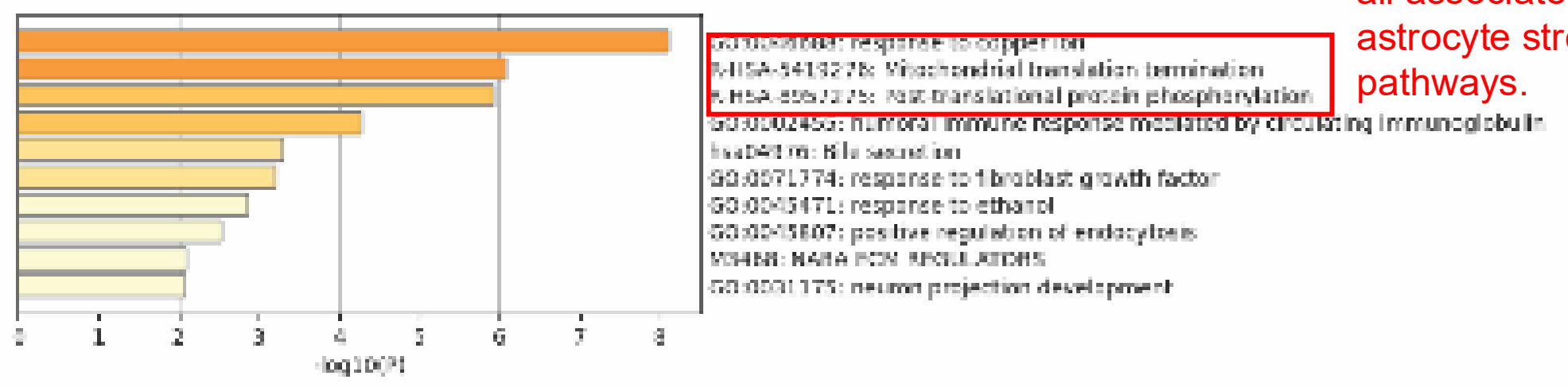


GO:0022407: regulation of cell-cell adhesion
GO:0071345: cellular response to cytokine stimulus
GO:1903034: regulation of response to wounding
R-HSA-1474244: Extracellular matrix organization
WP4522: Metabolic pathway of LDL HDL and TG including diseases
GO:0002764: immune response-regulating signaling pathway
GO:0007162: negative regulation of cell adhesion
M18: PID INTEGRIN1 PATHWAY
GO:0031349: positive regulation of defense response
GO:0006897: endocytosis
GO:0034103: regulation of tissue remodeling
GO:0071396: cellular response to lipid
GO:0001944: vasculature development
GO:0045766: positive regulation of angiogenesis
GO:0061024: membrane organization
GO:0007160: cell-matrix adhesion
M40: PID E2F PATHWAY
GO:0010951: negative regulation of endopeptidase activity
GO:0048008: platelet-derived growth factor receptor signaling pathway
WP2877: Vitamin D receptor pathway

These GO terms are all associated with disease-associated microglia.

Cell type/Genotype: scVIP identifies Astro_2 as a disease-affected cell type and identifies potentially involved genes





These GO terms are all associated with astrocyte stress pathways.

Outline

Persona model

Debug: Mouse visual cortex development

SEA-AD + ROSMAP Multi-regional Analysis

SEA-AD + AMP-AD Caudate Analysis

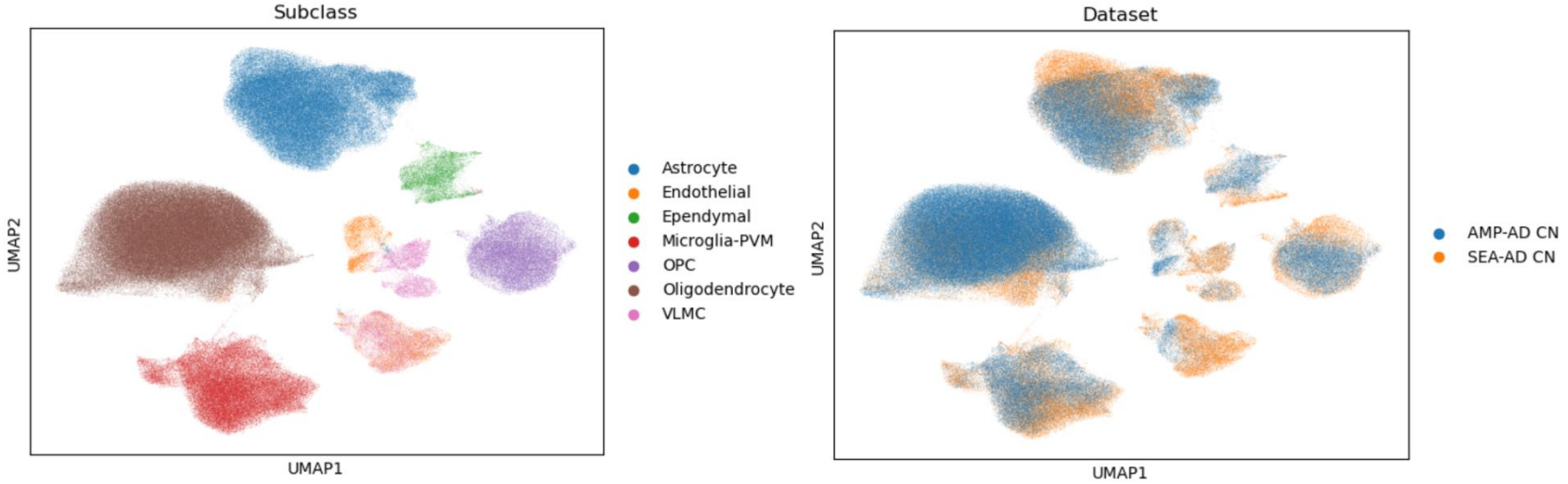
Bonus Track

DREAM Challenge

Enabling Large-Scale Model Training

Caudate is late in AD progression, and its signal is weaker and there are fewer donors than MTG



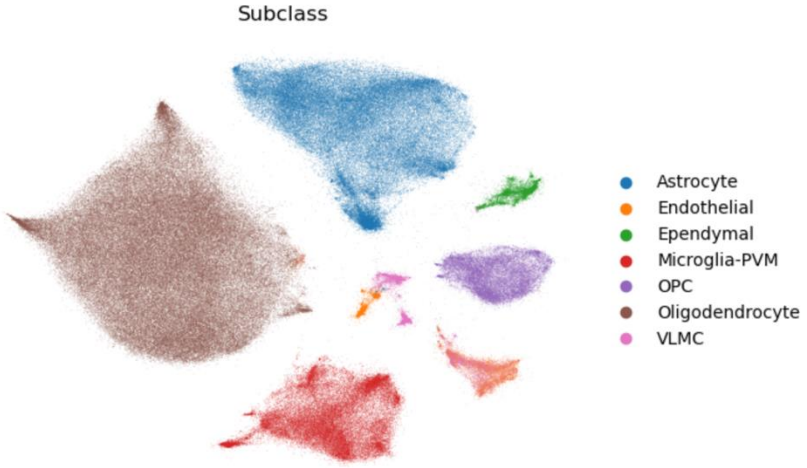


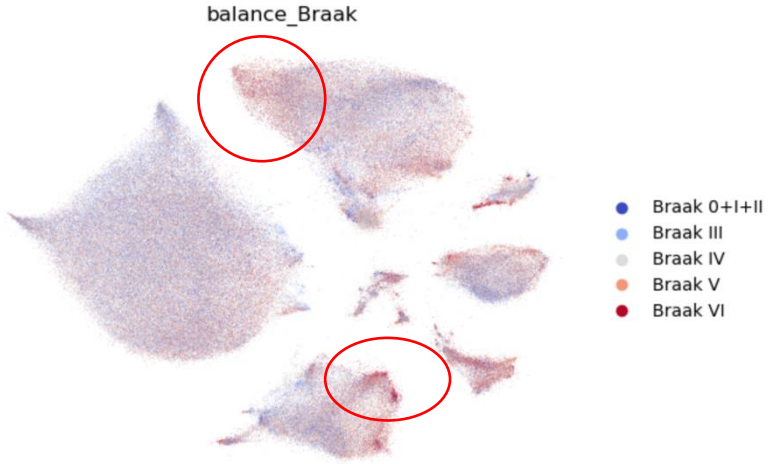
Datasets:
AMP-AD: 0.3M glia cells, 36k genes, 126 donors
SEA-AD: 0.2M glia cells, 36k genes, 42 donors

Phenotypes:
- Shared: Braak, ADNC, CERAD
- SEA-AD: percent AT8, percent 6e10

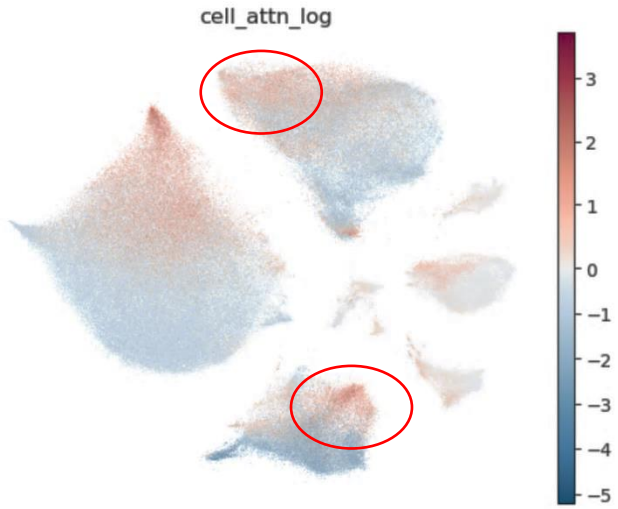
Goals:
- Understand the model's ability to learn phenotype, cell-type, genotype info
- Compare the result with the current analysis pipeline

Cell type: scVIP can identify weak signals that may be predictable of AD progression





With the labels of disease states, PERSONA can further pull cells with different staging apart.

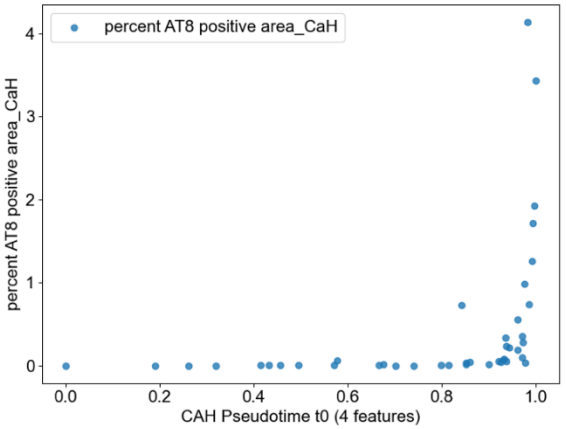


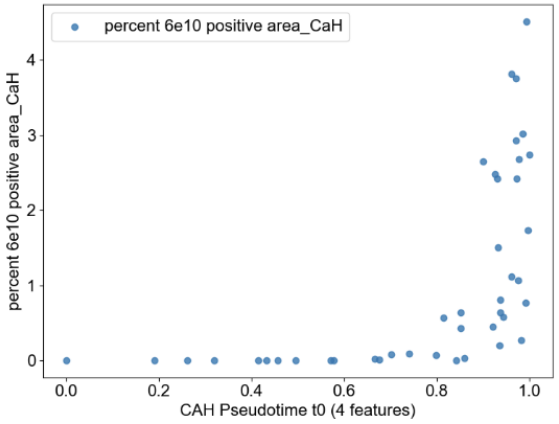
High-attention cells tend to be disease-associated

Genotype: Comparing scVIP's genotype result with the usual SEA-AD analysis pipeline



- Usual pipeline:
  - Calculate pseudoprogression using quantitative neuropathology measurements (n=42)
  - Identify genes changing significantly with respect to the pseudoprogression using frequentist methods such as Nebula





Nebula computes changes in gene expression in each cell, but regressors/covariate exist at the donor level

Pseudoprogression works at the donor level, ordering donors accordingly to pathological growth.

Usual procedures require calculations of the continuous pseudoprogression scores (CPS)

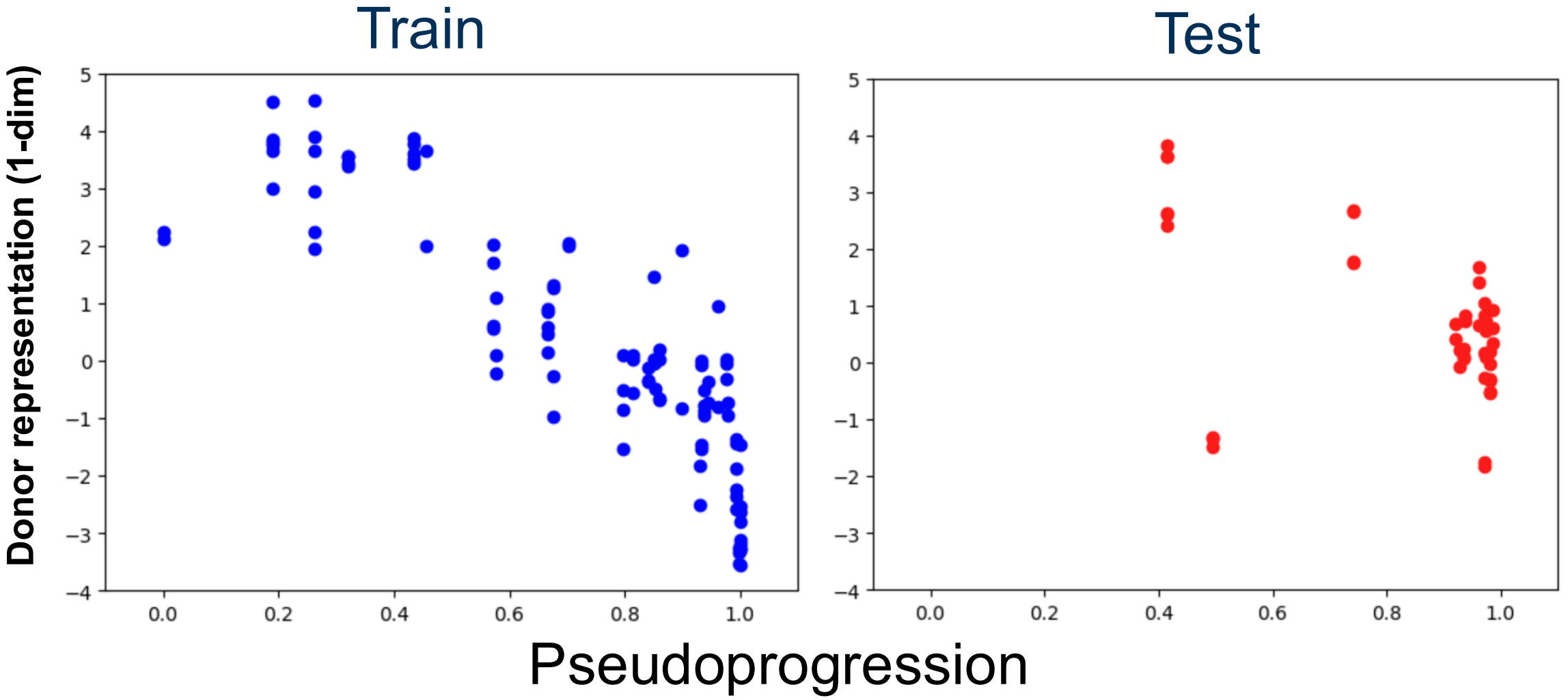
Genotype: Comparing scVIP's genotype result with the usual SEA-AD analysis pipeline

- scVIP:
  - Harmonize data across two datasets and use attention mechanisms to learn pseudoprogression (donor representation) (n=168)
  - Identify cells with high attention versus low attention
  - Run Bayesian differential analysis on these two sets of cells to identify differentially expressed genes

Benefits:

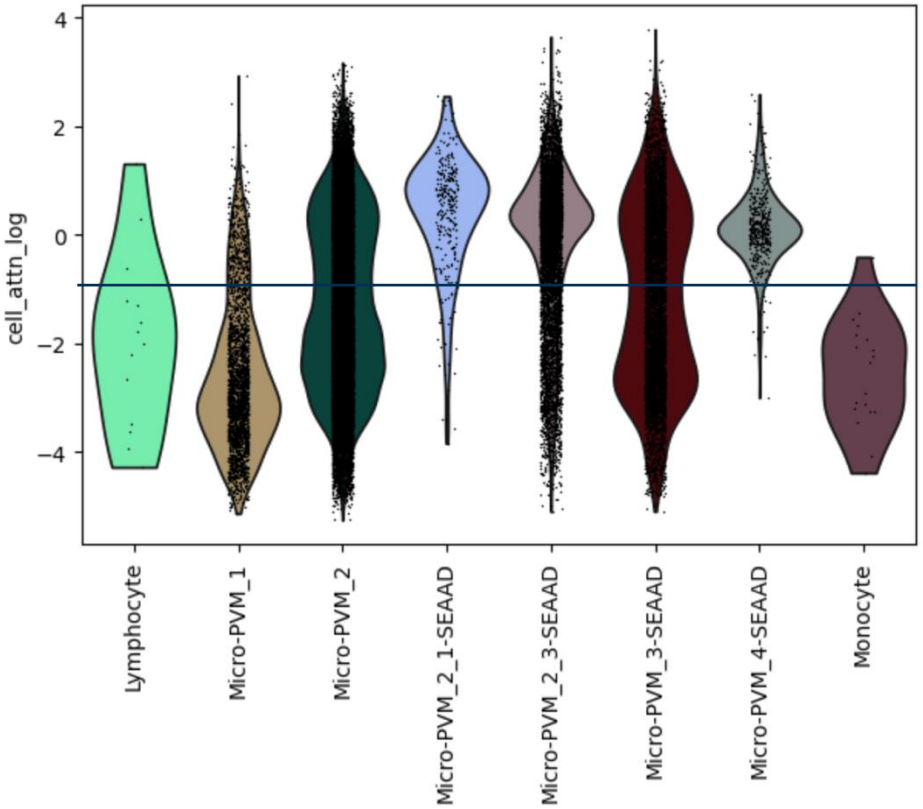
1. scVIP can utilize various data sources.
2. scVIP does not need to separately calculate CPS.
3. scVIP can separate disease-associated cells from non-disease-associated cells.

Donor representation learned by scVIP coincides with pseudoprogression



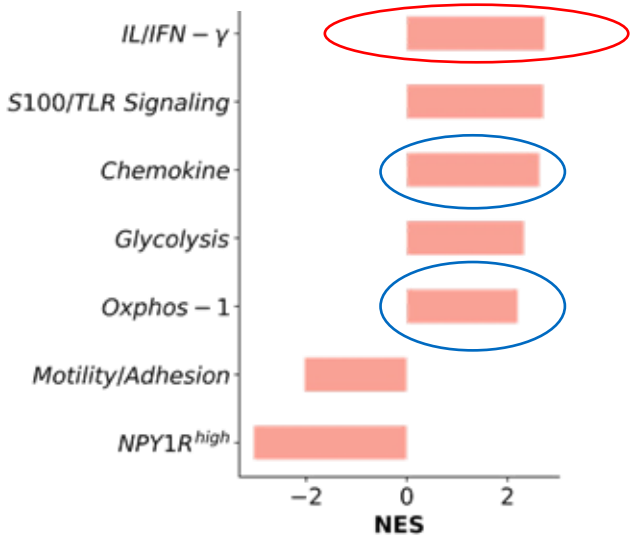
scVIP can identify disease pathways in microglia as from the usual analysis pipeline

Pseudoprogression + GEX changes in one go !



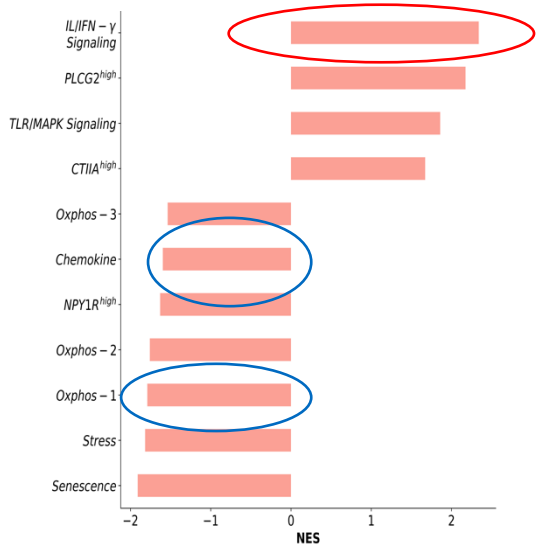
Lines: Cut-offs for differential expressions of high vs low attention @ uglia level

scVIP



Identify disease-associated microglia (disease-associated in the literature)

Nebula



Identify changes in microglia over AD progression (Tau-associated in the literature)

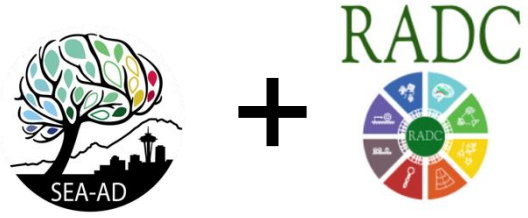
Outline



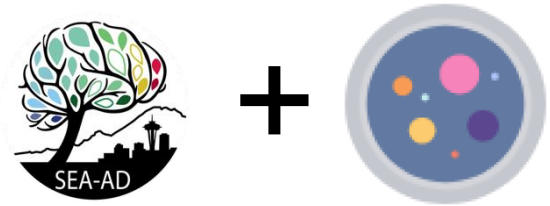
scVIP model



Debug: Mouse visual cortex development



SEA-AD + ROSMAP Multi-regional Analysis



SEA-AD + AMP-AD Caudate Analysis

Bonus Track



DREAM Challenge

Enabling Large-Scale Model Training

scVIP can:



Learn information from genotype to phenotype

Harmonize data across AD cohorts

Predict Donor Level Phenotypes from Omics Data sets

We envision to:

Transform our **analysis pipelines**: Run After Data collection to quickly explore disease heterogeneity

Create a **web product**: Researchers upload omics, and we predict disease severity (CPS) and neuropathology across cohorts

Outline

Persona model

Debug: Mouse visual cortex development

SEA-AD + AMP-AD Caudate Analysis

SEA-AD + ROSMAP Multi-regional Analysis

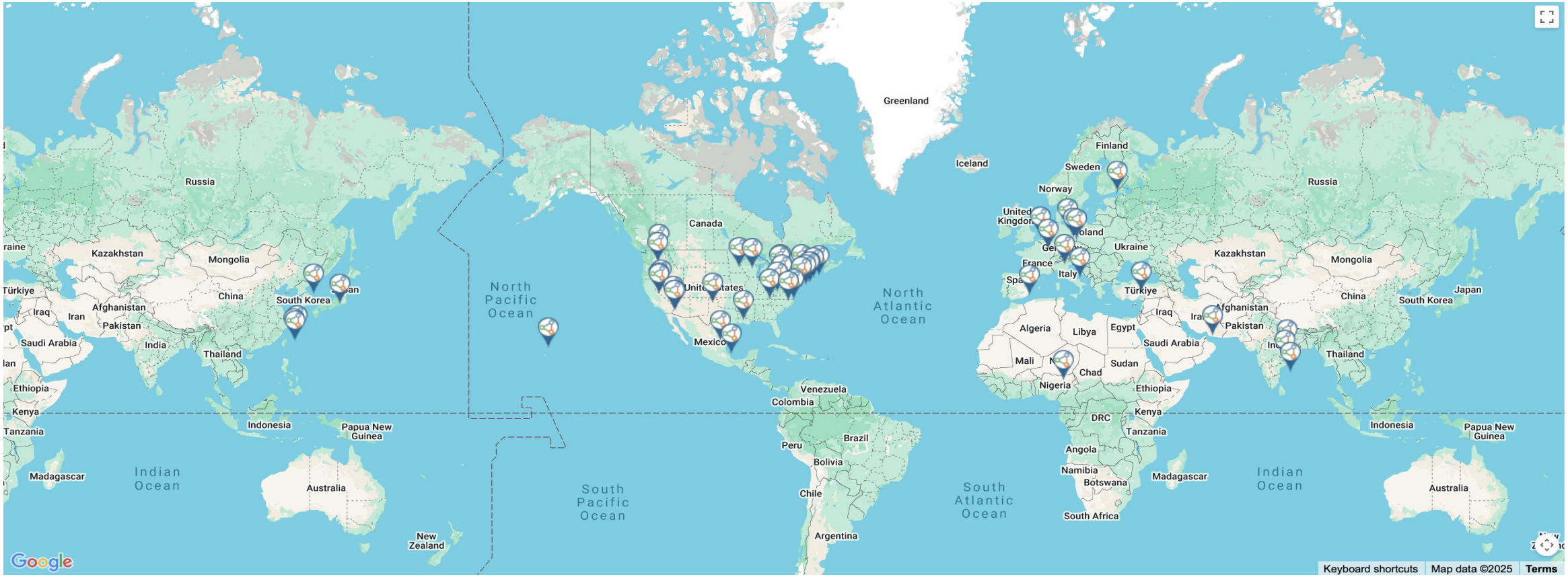
Bonus Track



DREAM Challenge

Enabling Large-Scale Model Training

- Co-organized by SEA-AD Allen Institute, Sage Bionetworks, DREAM, and UCSF.
- Goal: predict AD pathology from snRNA-seq data
- Data: SEA-AD MTG/A9/STG
- Tasks: 1) Predict ADNC/Braak/Thal/CERAD 2) Predict AT8/6e10 percentage



- scVIP is used as a baseline to provide proof that the challenge is valid.

Outline

Persona model

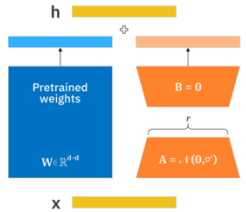
Debug: Mouse visual cortex development

SEA-AD + AMP-AD Caudate Analysis

SEA-AD + ROSMAP Multi-regional Analysis

Bonus Track

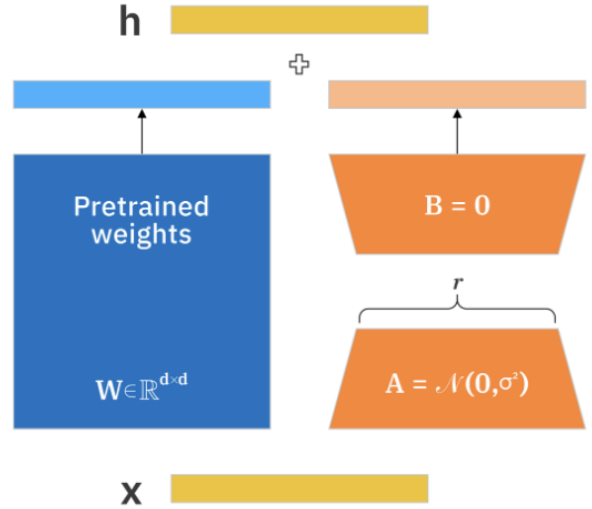
DREAM Challenge



Enabling large-scale model training

LoRA: Finetuning can be accomplished through low rank matrices

- Goal: When we have small fine-tuning data, we don't want the pretrained weights to be overfit to the small data.
- Key insight: Parameters for finetuning have low intrinsic dimensionality
- Solution: In any linear transformation, consider the updated weight as $W = W_0 + BA$, where $B \in \mathbb{R}^{d \times r}$, r is the rank, and W_0 is the pre-trained weight.
- Application to scVIP: Train scVI+scVIP where the parameters in scVI is replaced by LoRA



Comparing with and without LoRA

	Learning rate	Epochs	ROSMAP tau CCC	SEAAD tau CCC
No LoRA	0.0001	20	0.35	0.5
No LoRA	0.0001	30	0.45	0.54
No LoRA	0.0001	40	0.49	0.50
No LoRA	0.0005	20	0.43	0.44
No LoRA	0.0005	30	0.37	0.52
LoRA	0.0001	20	0.36	0.53
LoRA	0.0001	30	0.42	0.49
LoRA	0.0005	20	0.55	0.56
LoRA	0.0005	30	0.50	0.54

With LoRA implementation, scVIP's performance may improve, but learning rates and epochs need to be adjusted.

Mouse visual cortex dataset

Yuan Gao, Zizhen Yao, Cindy van Velthoven

Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD)

Acknowledgement



The Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD) consortium is supported by the National Institutes on Aging (NIA) grant 2U19AG060909.