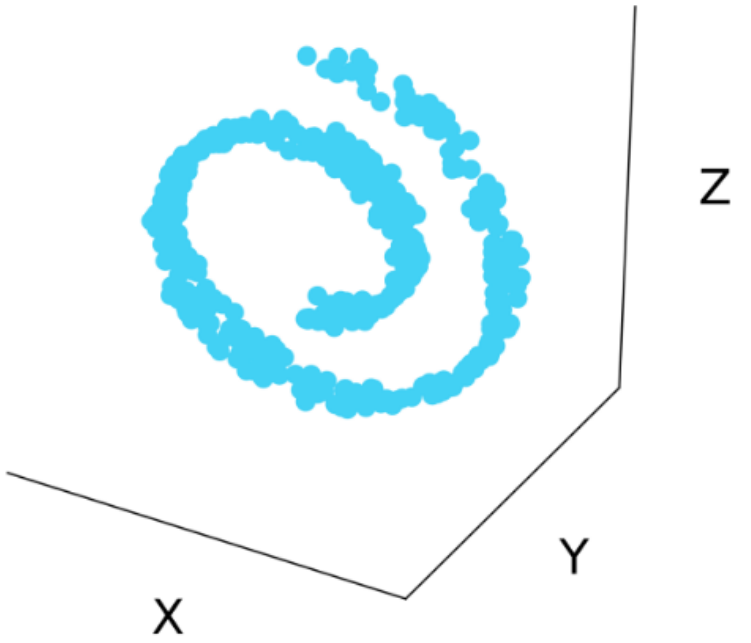
What is a passion you've discovered during quarantine?

Day 3: Denoising, Batch Correction, and Clustering

What is the intrinsic or latent dimensionality of this data?



0

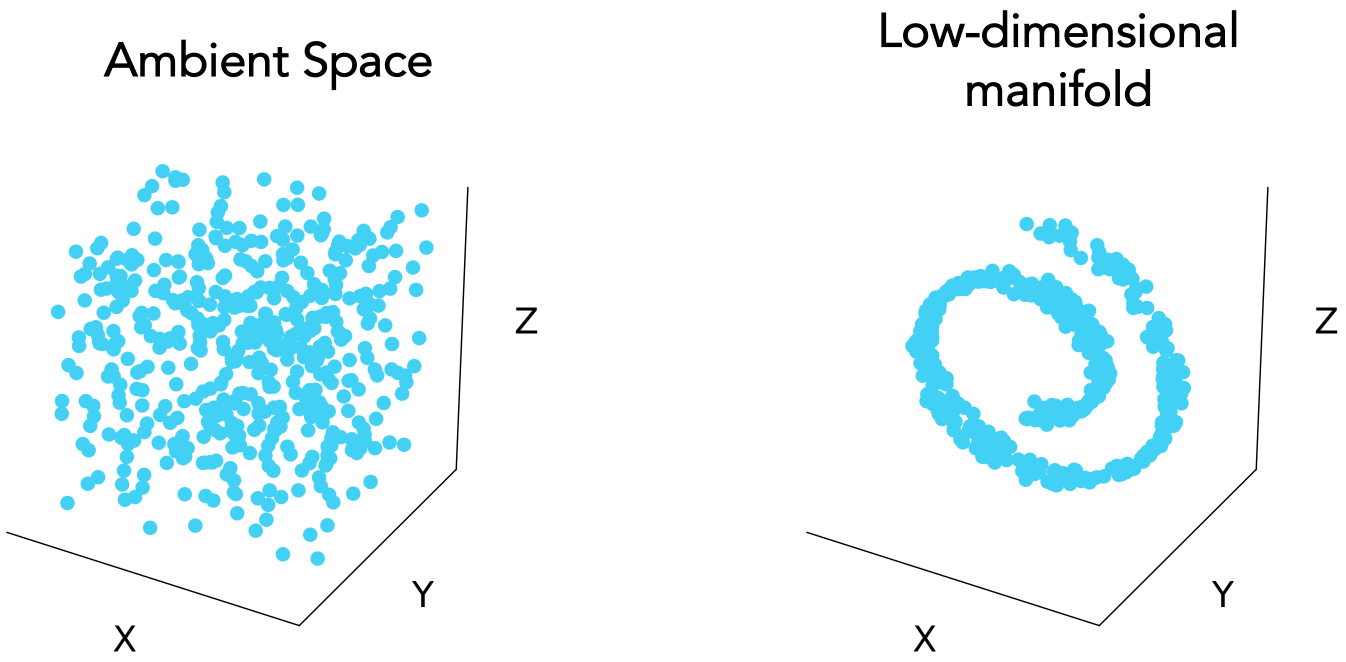
1

2

3

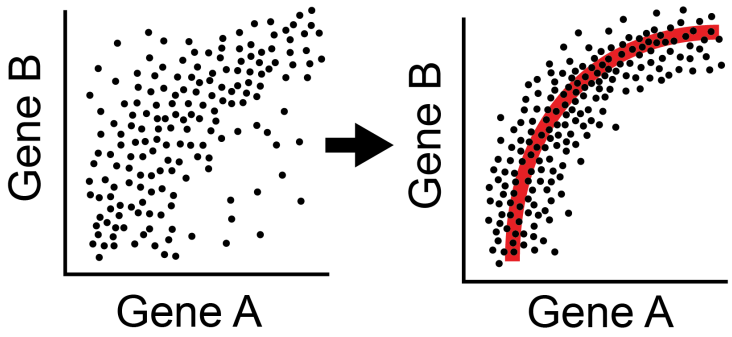
4

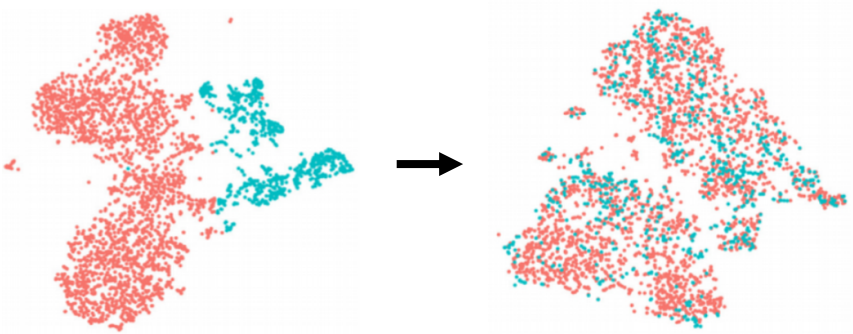
Latent structure in high dimensional data



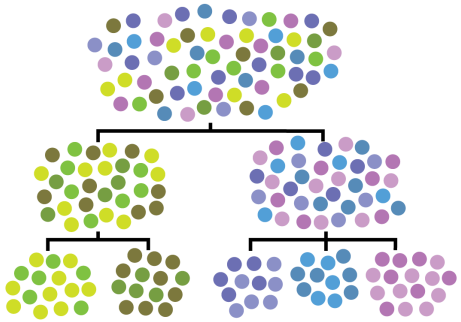
Other uses of affinity matrices, eigenvectors

Data denoising and batch normalization





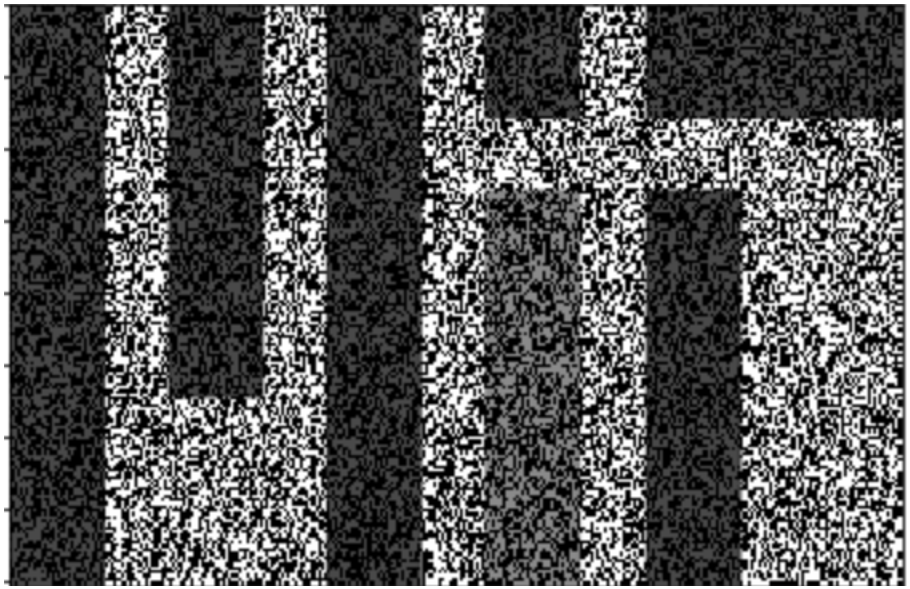
Clustering



Using the manifold model to denoise data

Denoising by eliminating dimensions

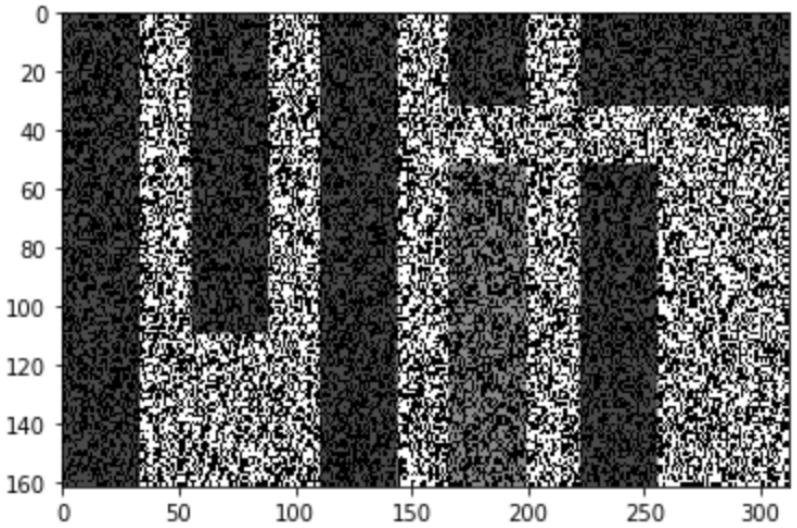
- The number of dimensions in a data that are independent are called RANK
- You can denoise data by lowering rank



Noisy image

Recreate the data without noise

- For this to work you have to have your dimensions be broken up into ***data dimensions and noise dimensions***
- Noisy image on the right is broken up into columns and rows instead
- What do we do?



When poll is active, respond at **PollEv.com/yaleml**

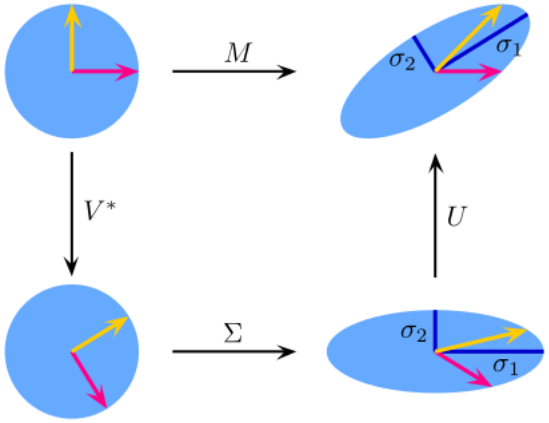
Text **YALEML** to **22333** once to join

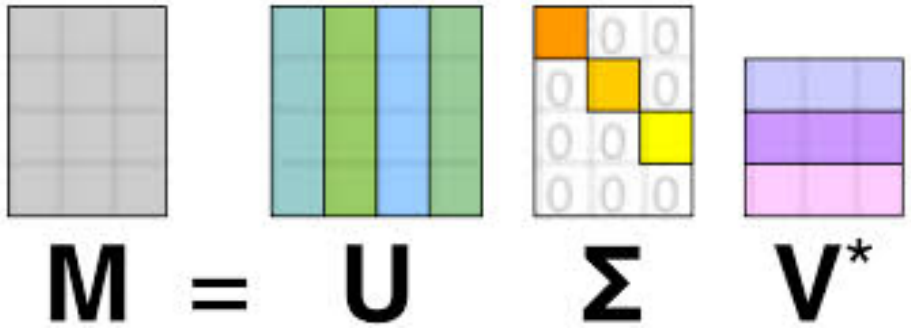
How do we get dimensions that correspond to data "signal" and "noise" ?

PCA splits signal and noise

- The axes with a lot of variation are likely to correspond to data
- Axes with little variation correspond to noise directions
- The amount of variation captured in each eigenvector is given by its eigenvalue
- PC1 will have the highest eigenvalue, captures the most variation ..
  - PC2 next most
    - And so on…

SVD: Process similar to Eigendecomposition on non-square matrices



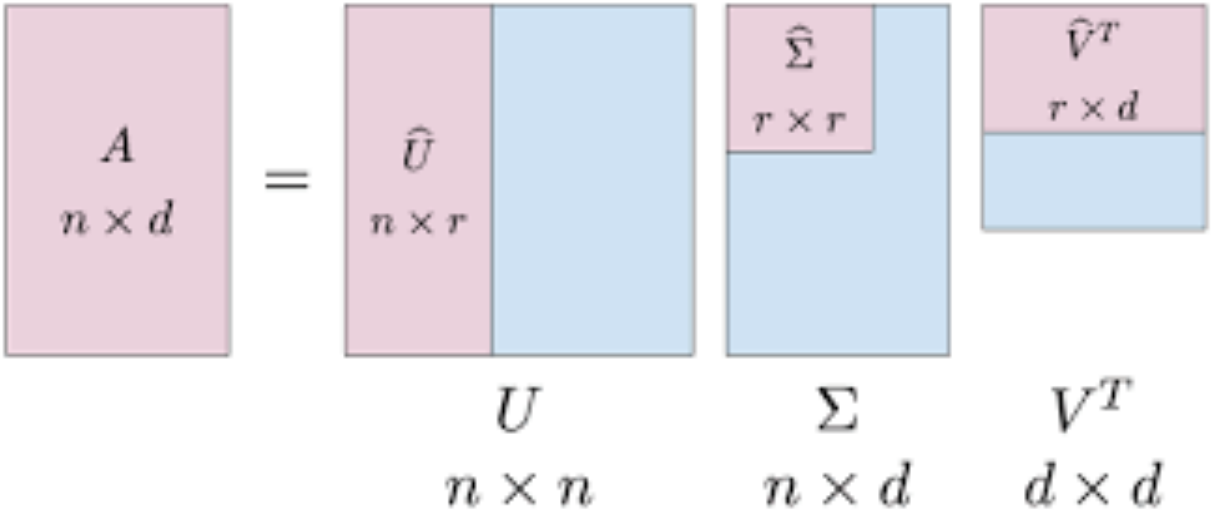


Singular values are eigenvalues of MM* or M*M
Singular vectors are eigenvectors of MM* or M*M

For a mean centered feature matrix the singular vectors U are also PCs

Low Rank approximation

- Eliminate the singular vectors with low singular values and recreate the matrix



Low Rank Approximation on MIT matrix

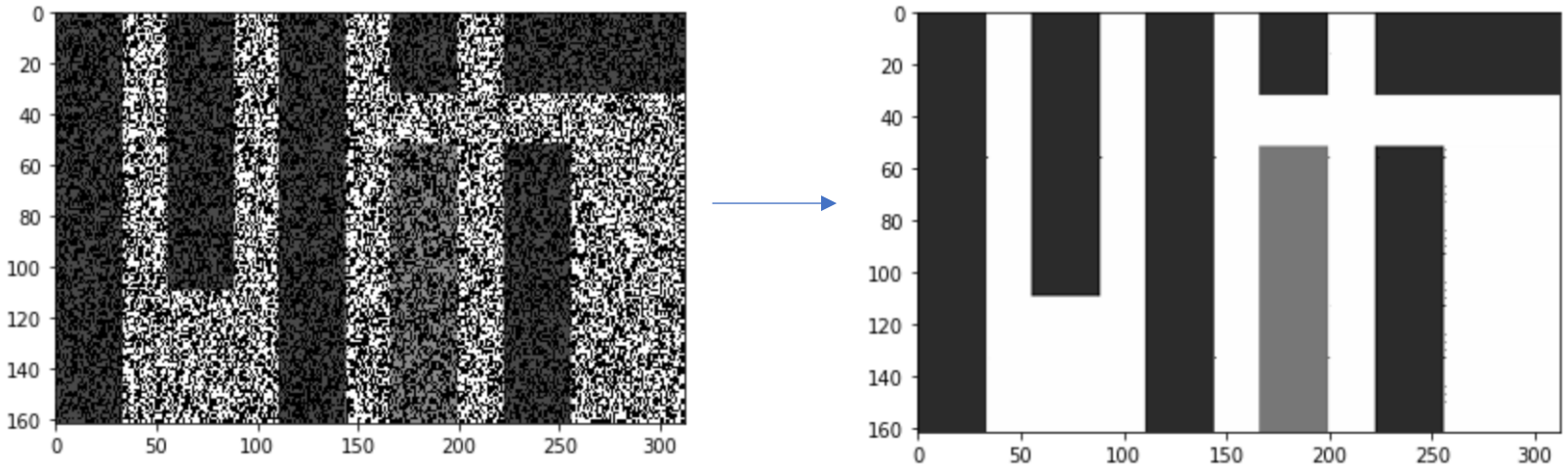
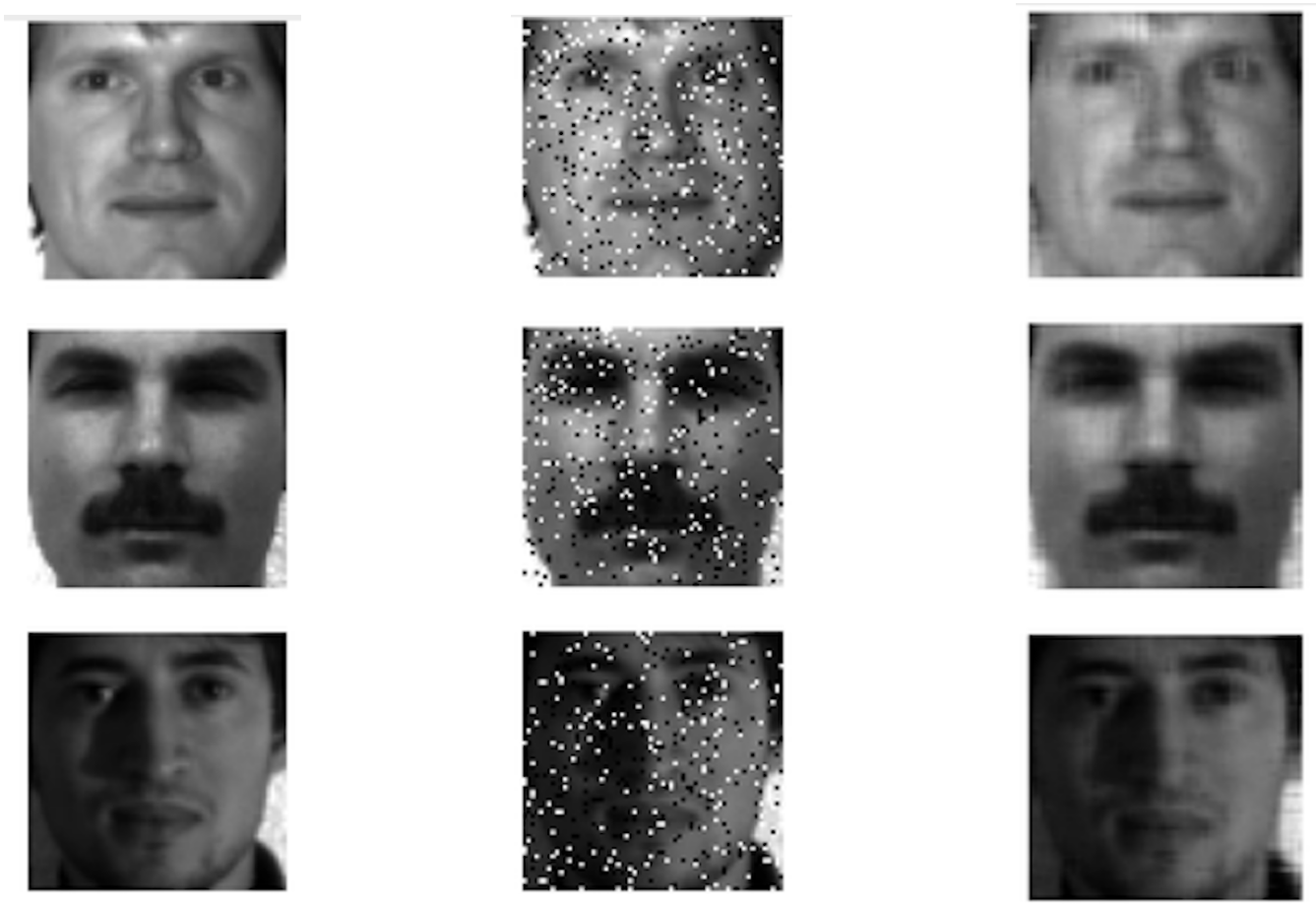


Image from https://medium.com/@amelie_yeh/

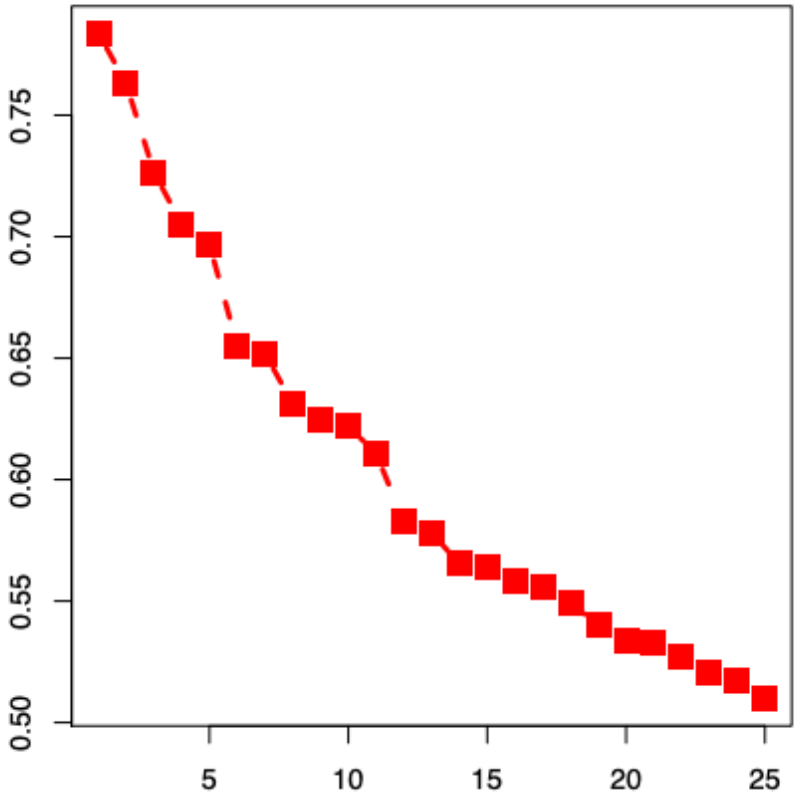
Data is smooth, Noise is jumpy



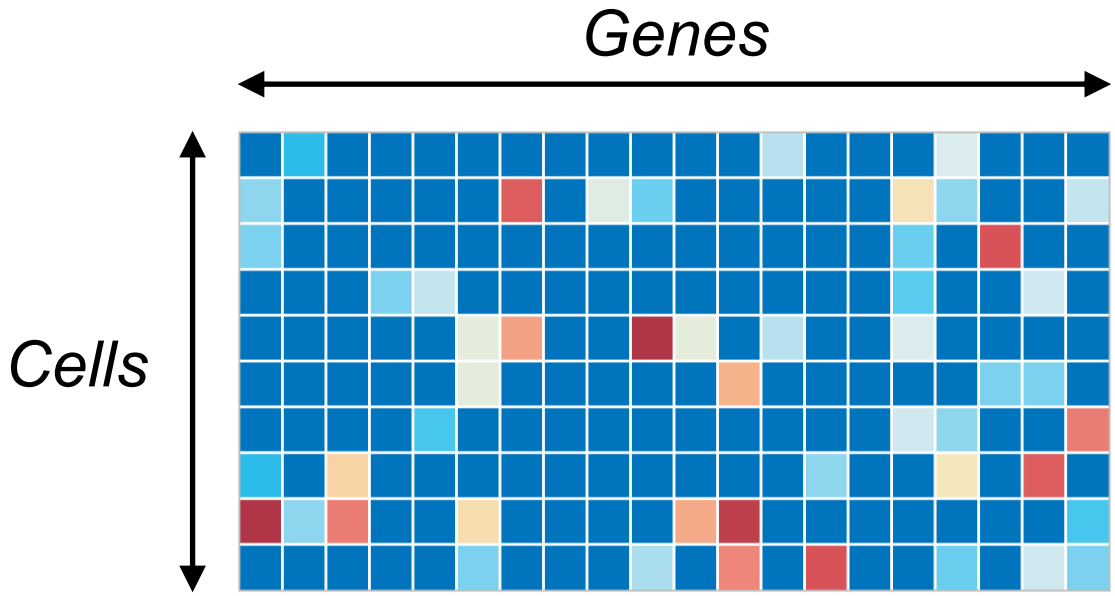
Can take off noise by taking off low-eigenvalued eigenvectors

How many eigenvectors should we use to reconstruct data? How many eigenvectors are noise? How do we know?

Eigengap



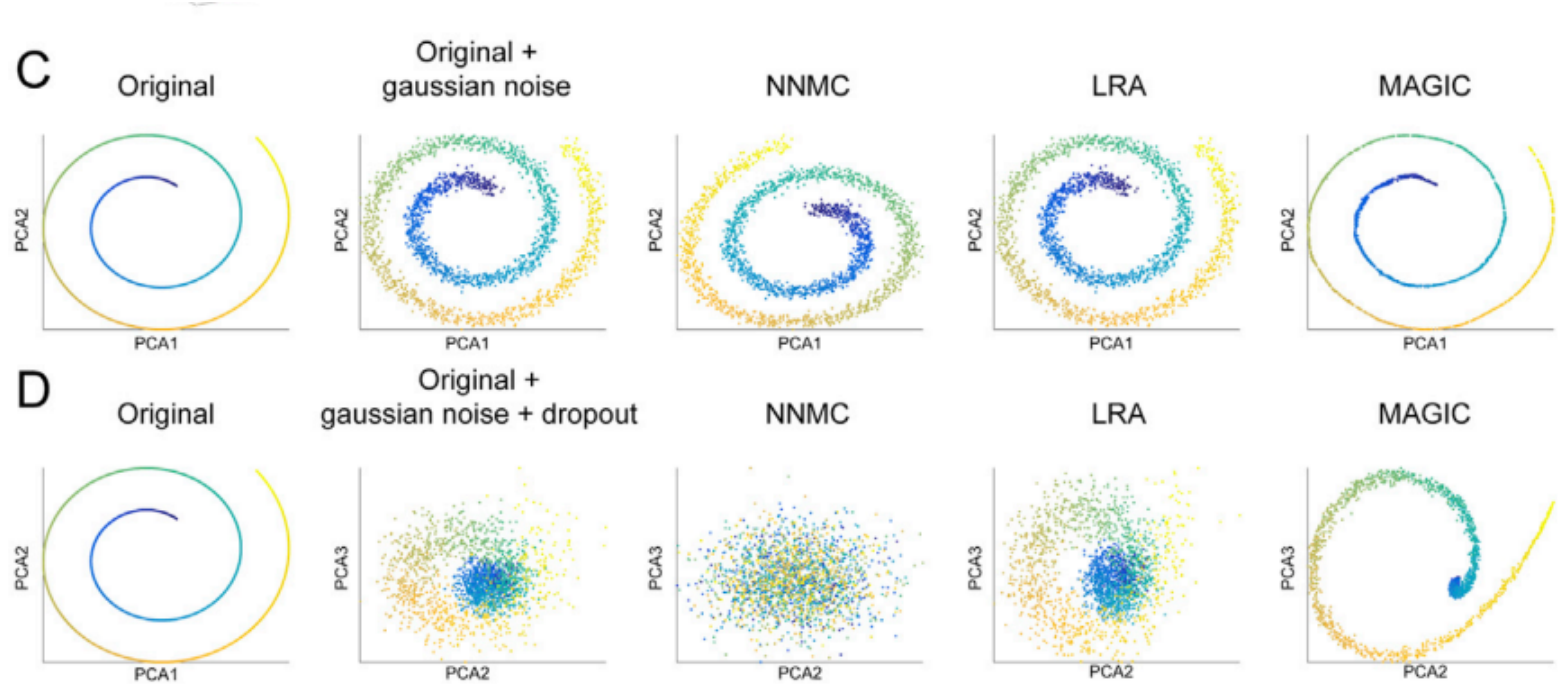
Jumps in eigenvalues
Might give you a clue



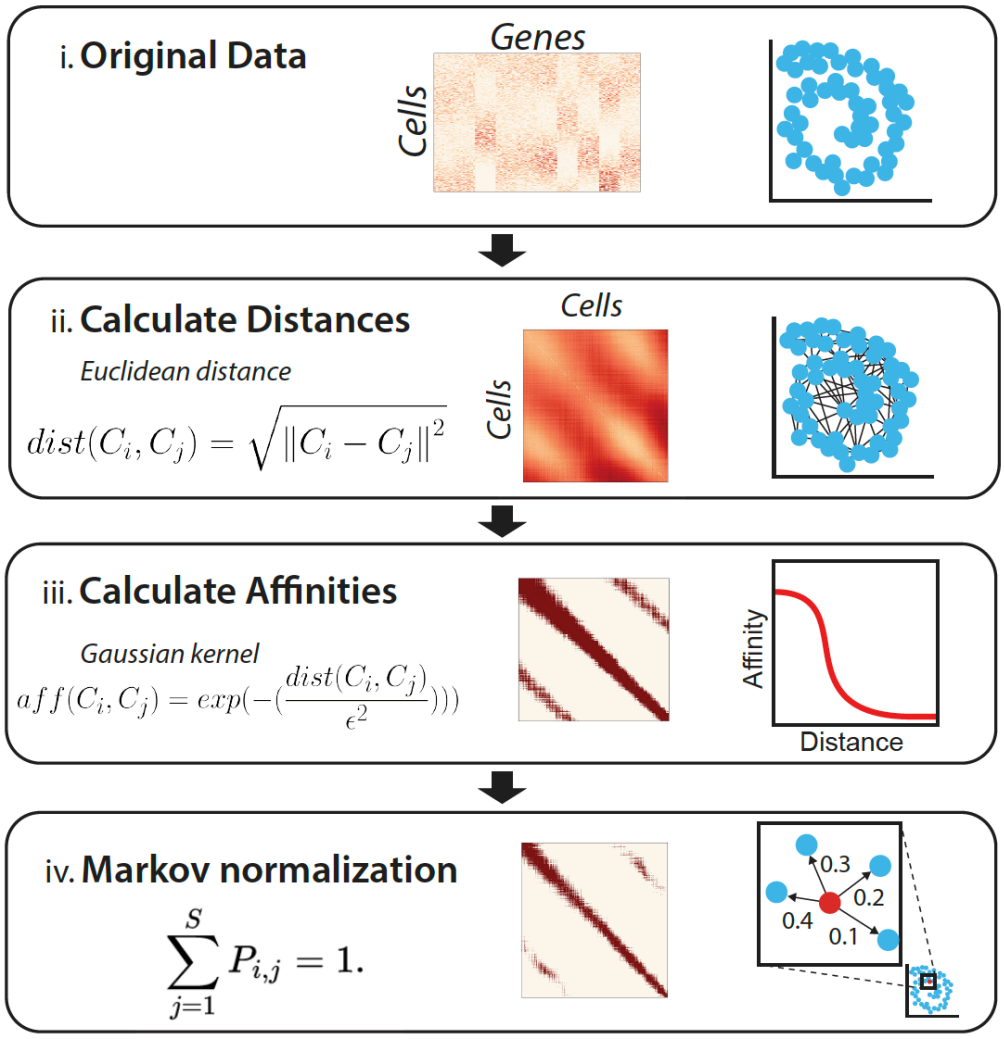
Data is noisy and sparse (scRNA-seq)

Can we use ideas from image denoising to denoise scRNA-seq data?

Low rank approximation for non-linear data

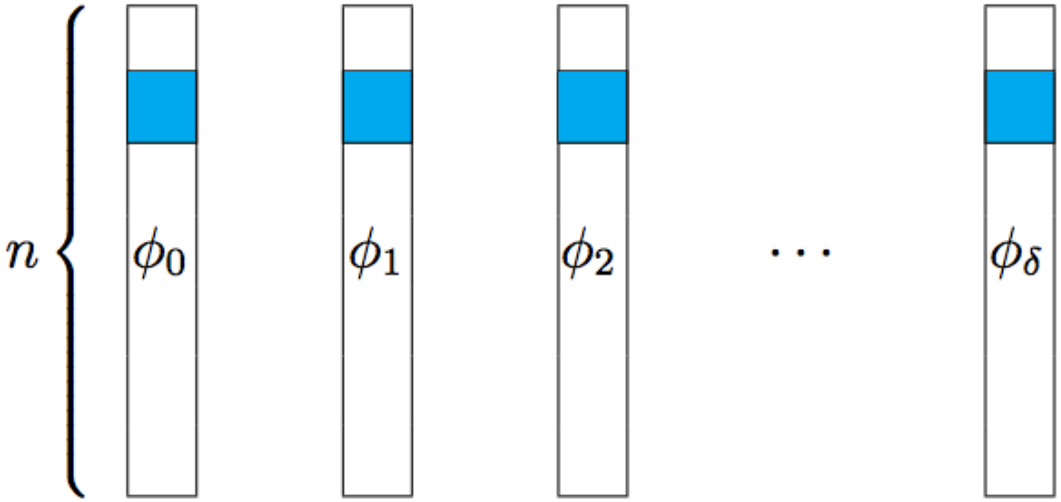


Problem with PCA/Single Value Decomposition (SVD): it takes off linear dimensions in data, noise can be along the non-linear data manifold


i. Original Data
Genes
Cells
ii. Calculate Distances
Euclidean distance
$dist(C_i, C_j) = \sqrt{\|C_i - C_j\|^2}$
Cells
Cells
iii. Calculate Affinities
Gaussian kernel
$aff(C_i, C_j) = exp(-(\frac{dist(C_i, C_j)}{\epsilon^2}))$
Affinity
Distance
iv. Markov normalization
$\sum_{j=1}^{S} P_{i,j} = 1.$
0.3
0.2
0.4
0.1

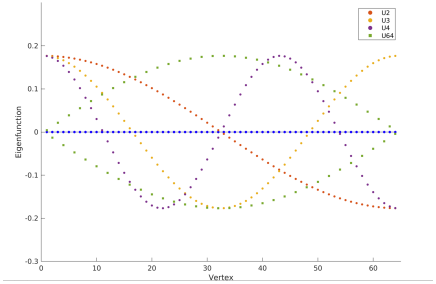
Eigenvectors of Affinity Matrix

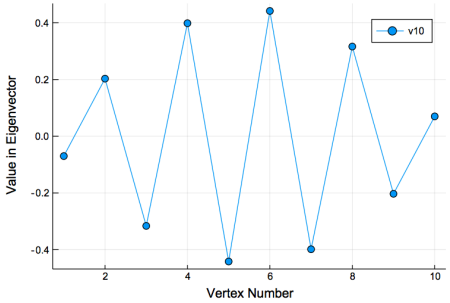
$$1 = \boxed{\lambda_0} \geq \boxed{\lambda_1} \geq \boxed{\lambda_2} \geq \cdots \geq \boxed{\lambda_\delta} > 0$$

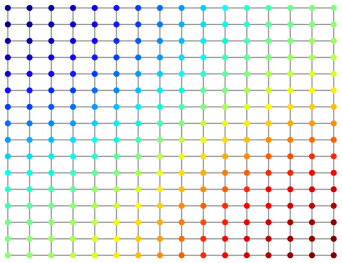


$$x \mapsto \Phi(x) \triangleq [\lambda_0\phi_0(x)\,,\ \lambda_1\phi_1(x)\,,\ \lambda_2\phi_2(x)\,,\ \ldots\,,\ \lambda_\delta\phi_\delta(x)]^T$$

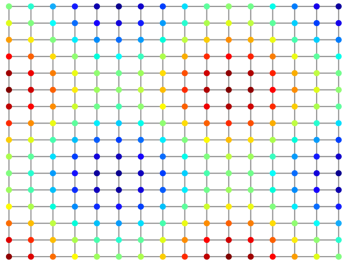
Eigenvectors are frequency harmonics



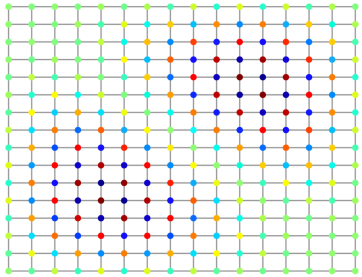




2nd Eigenvector

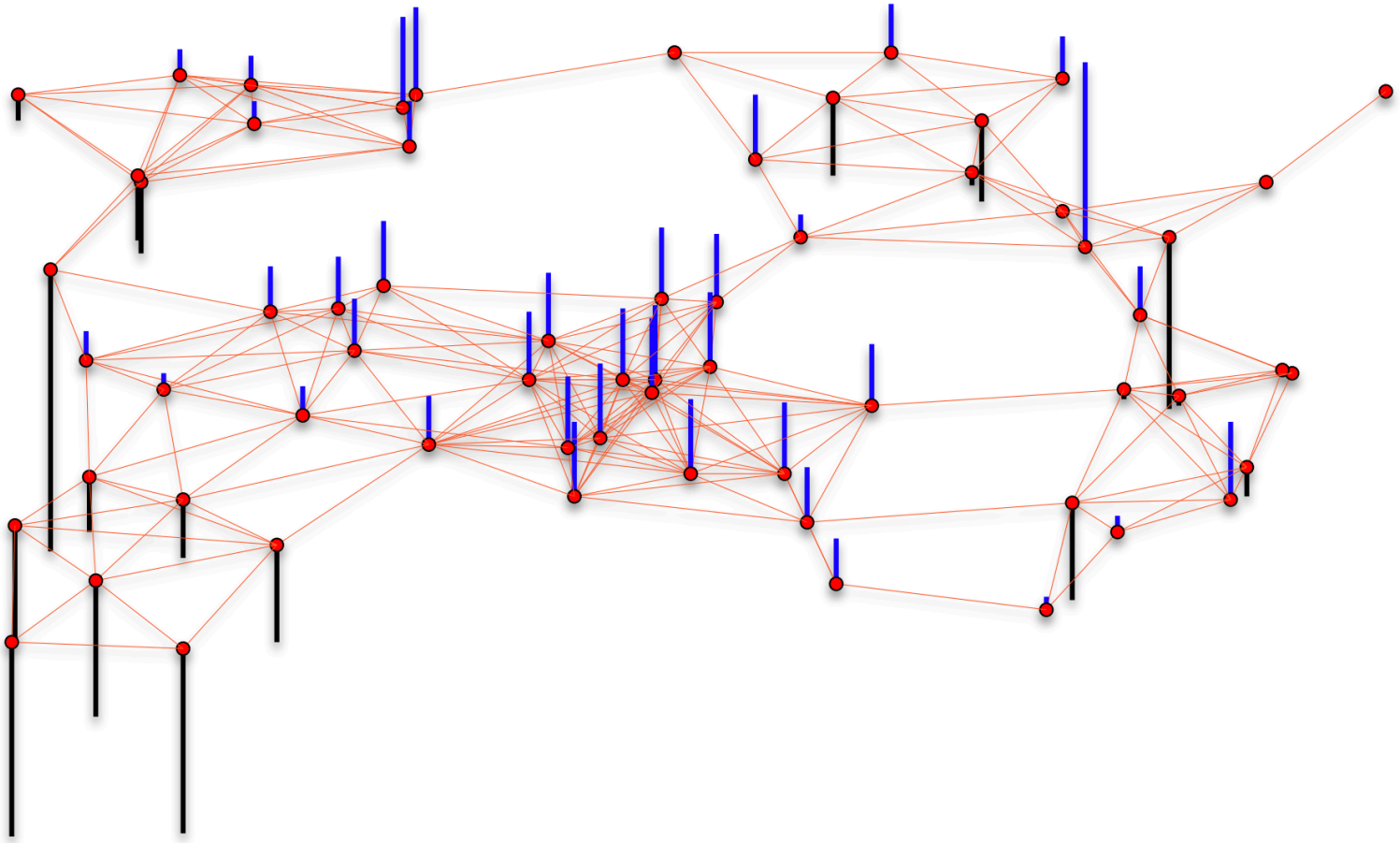


10th Eigenvector

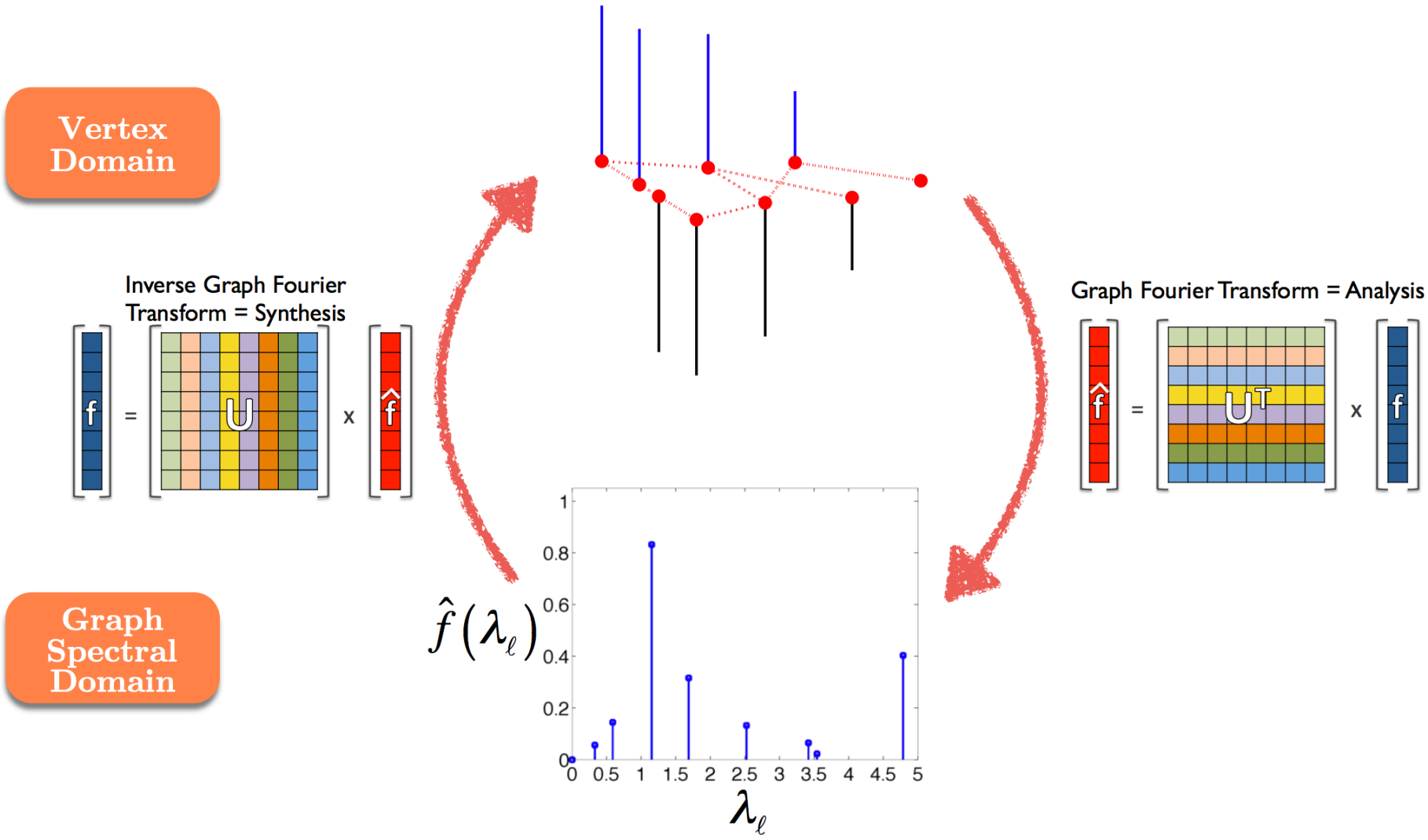


2nd to last eigenvector

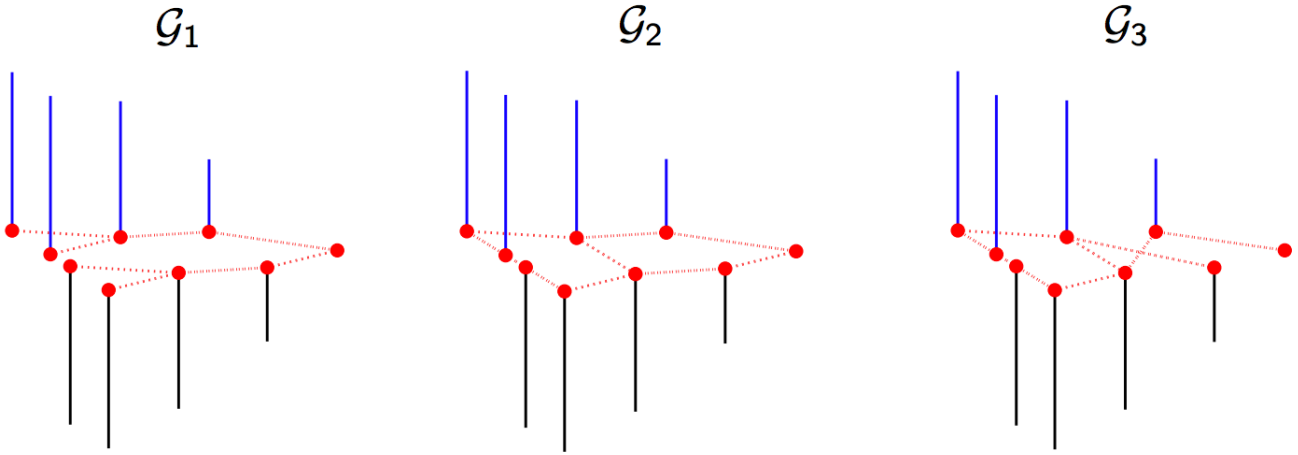
Cells are nodes, mRNA gene-counts are signals on a graph



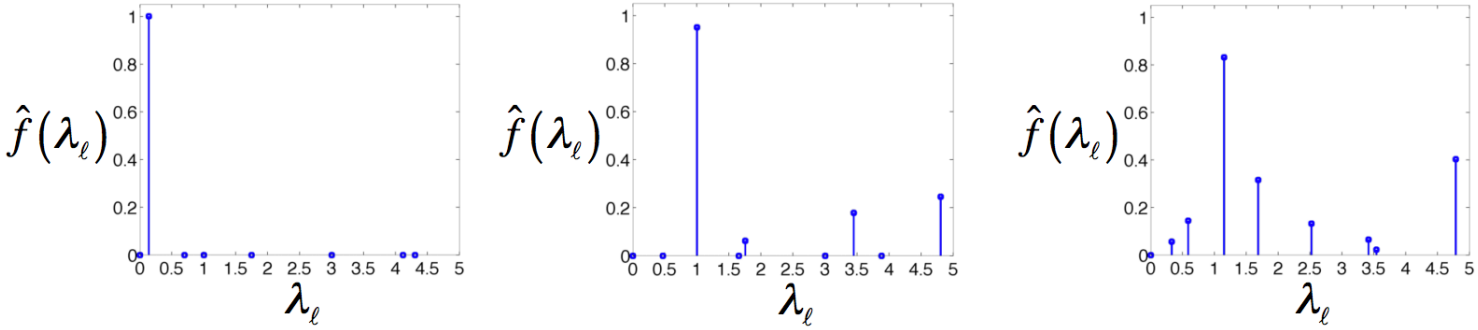
Graph Fourier Transform



Vertex Domain



Graph Spectral Domain

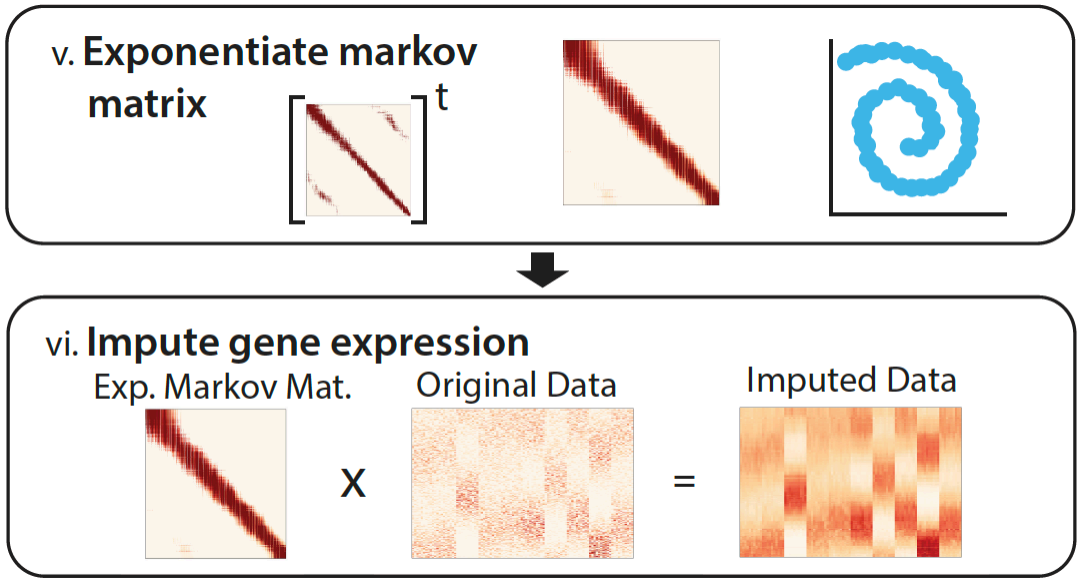


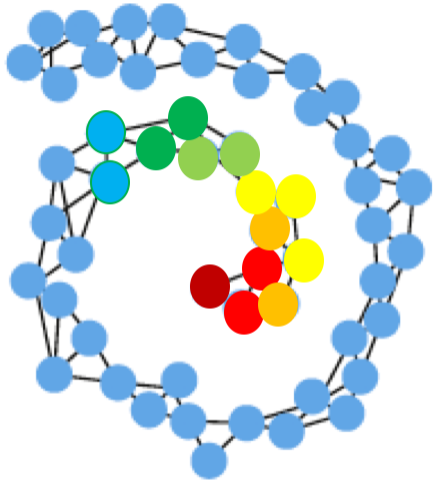
How many eigenvectors should we take off?

MAGIC

- Softly filters eigenvectors, down-weights them in a regular scheme rather than totally taking them off
- This is called low-pass filtering on the spectrum of the graph

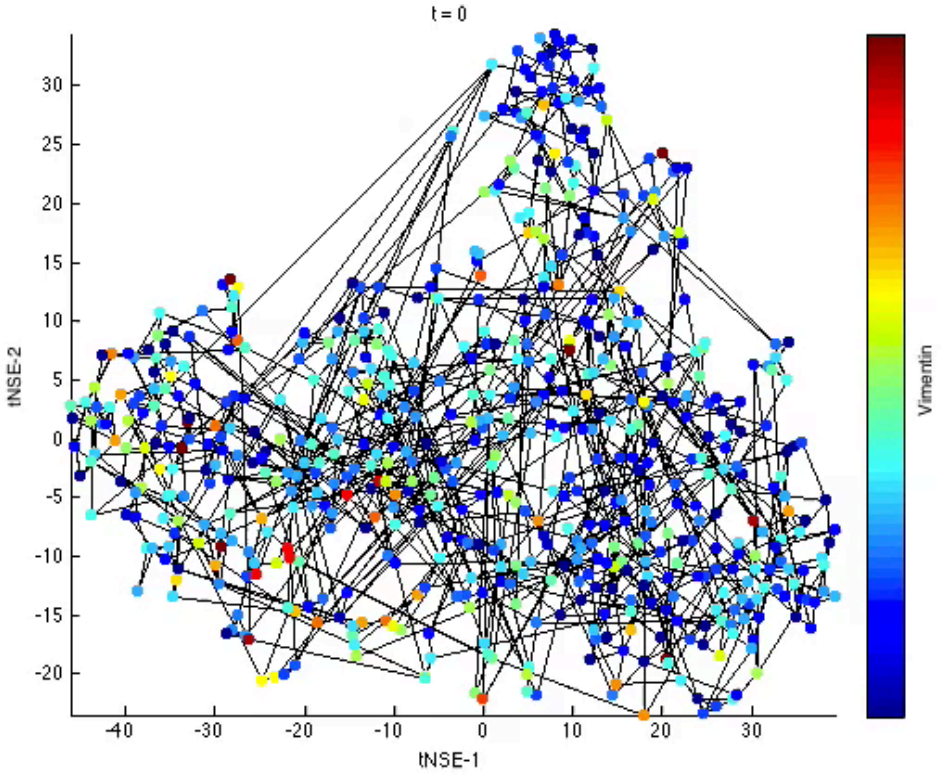
Imputation Step = Smoothing on Graph





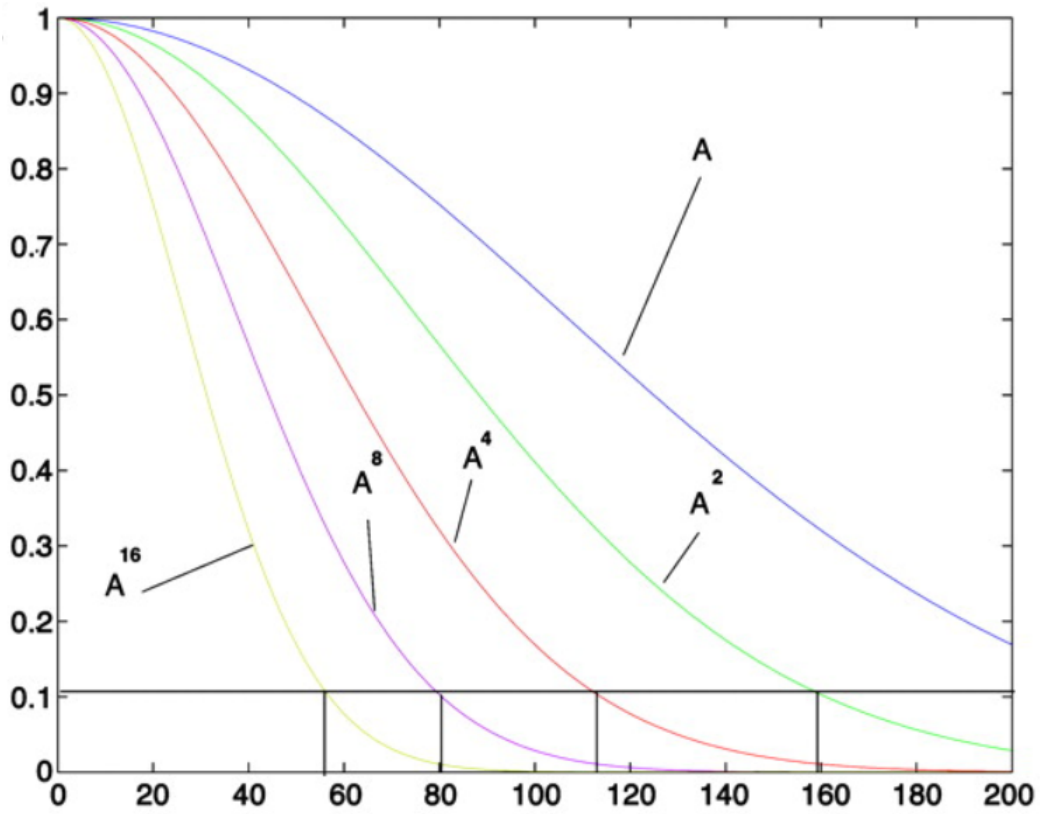
Vertex Domain

- Smooths signal on graph
- Takes weighted average of neighbor

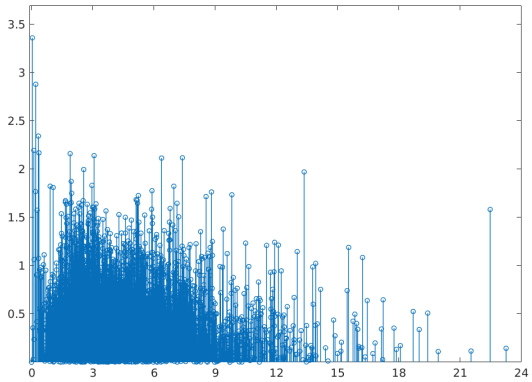


Low pass filter of Eigenvectors

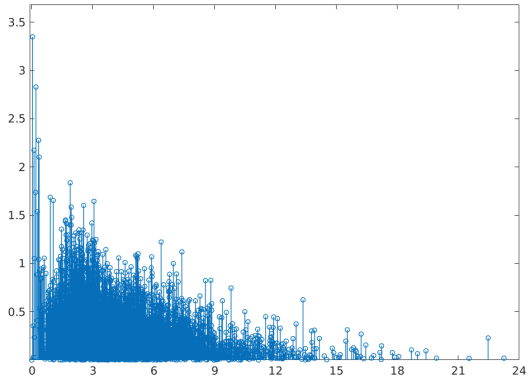
$$\boldsymbol{\Phi}_t(\boldsymbol{x}_i) : \boldsymbol{x}_i \longmapsto \left[\lambda_1^t \phi_1(i), \lambda_2^t \phi_2(i), \lambda_3^t \phi_3(i), ..., \lambda_{M-1}^t \phi_{M-1}(i)\right] \in \mathbb{R}^{M-1}$$



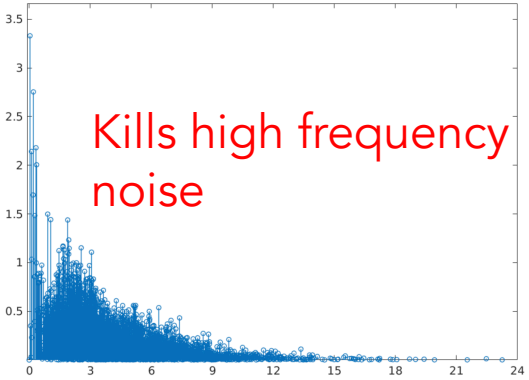
Frequency domain



No Smoothing

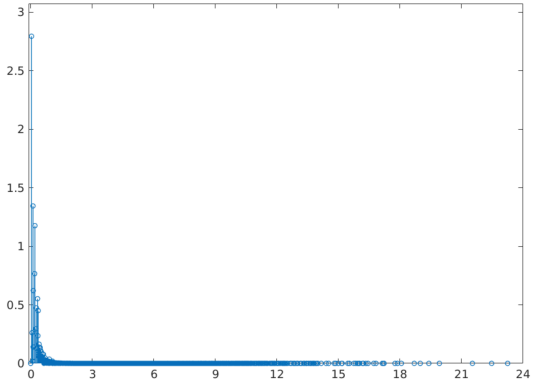


T=2

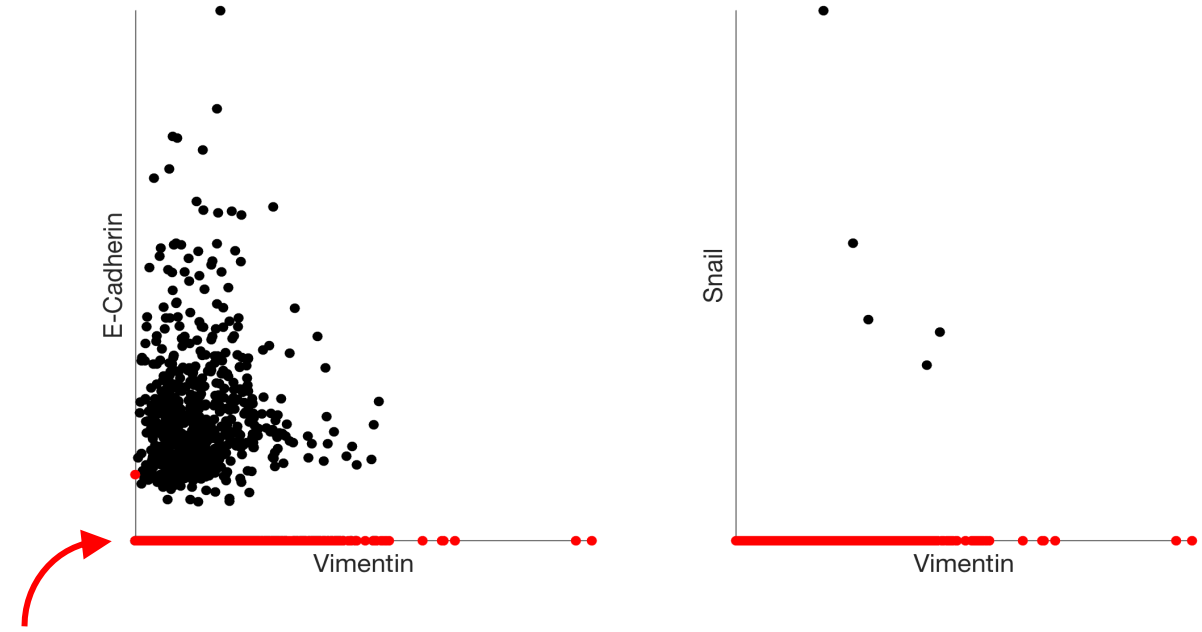


Kills high frequency noise

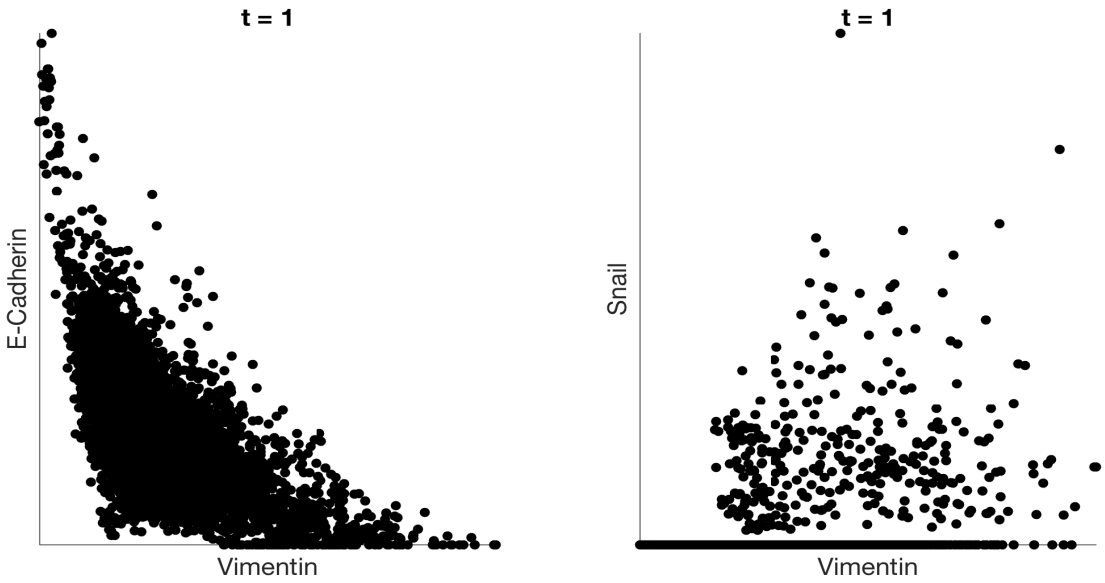
T=5



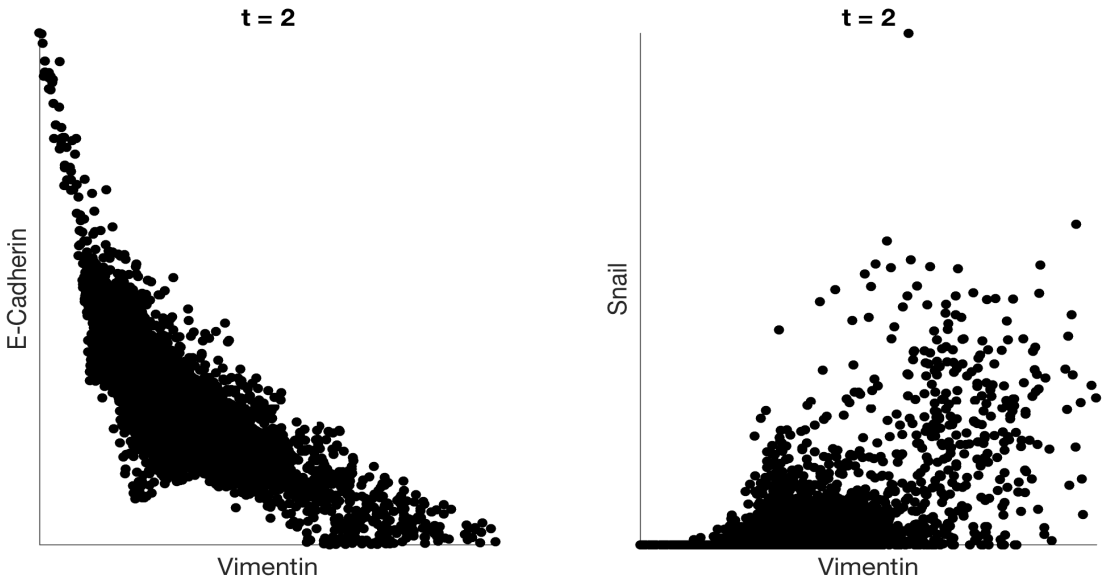
T=100



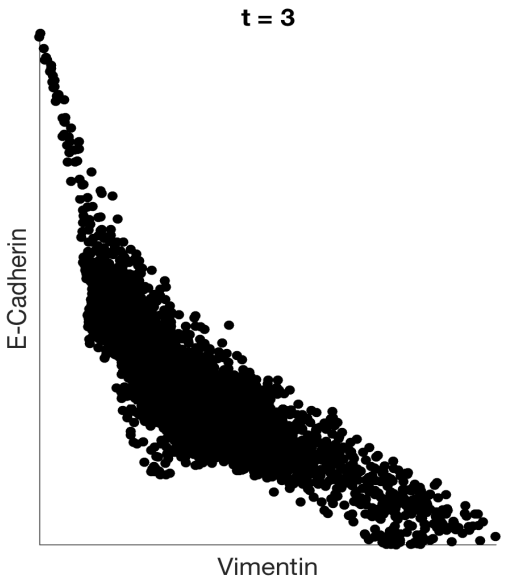
original data with dropout

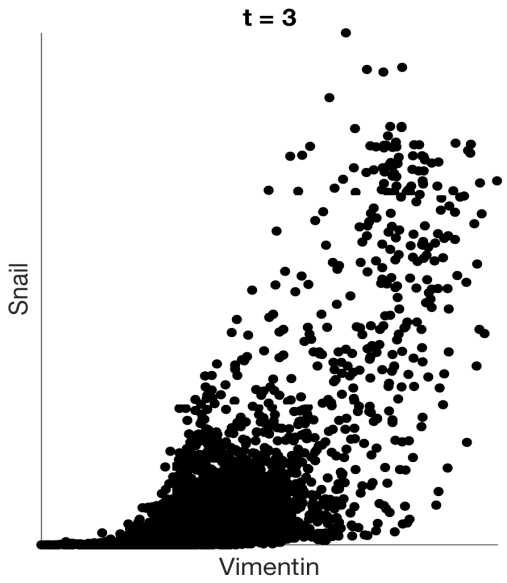


imputation with MAGIC

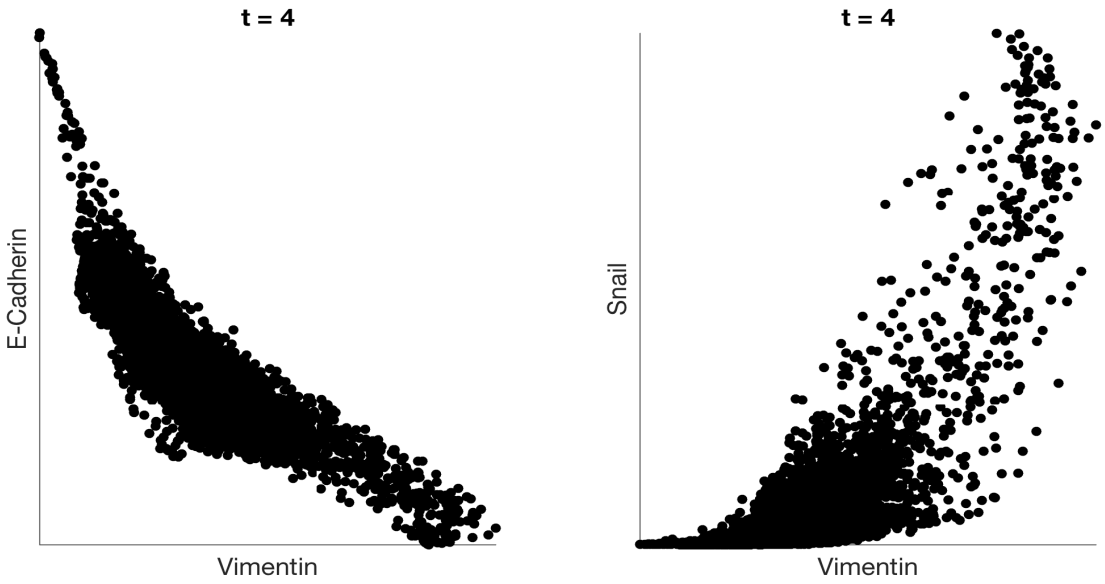


imputation with MAGIC

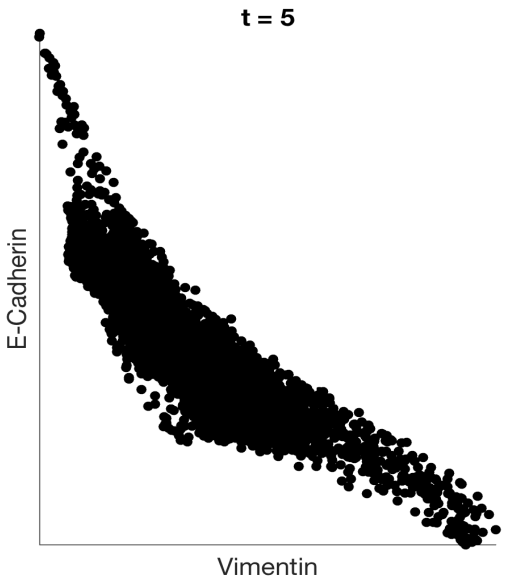


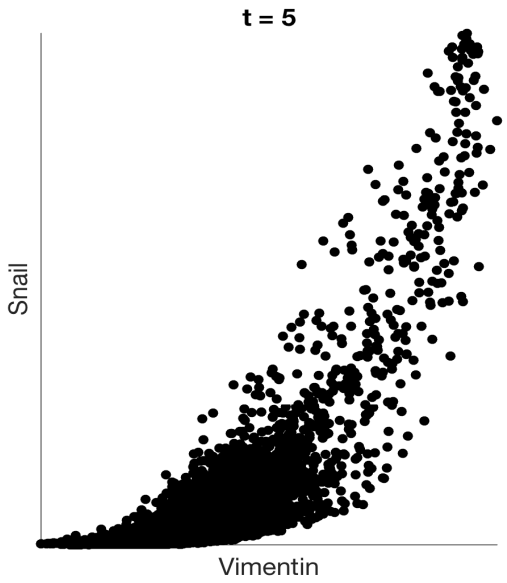


imputation with MAGIC

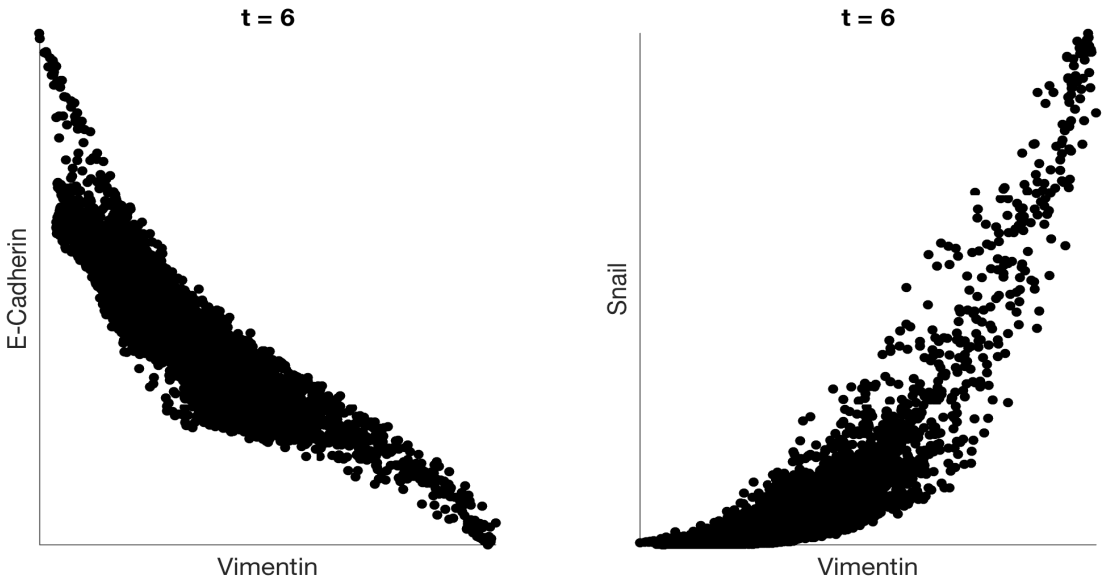


imputation with MAGIC



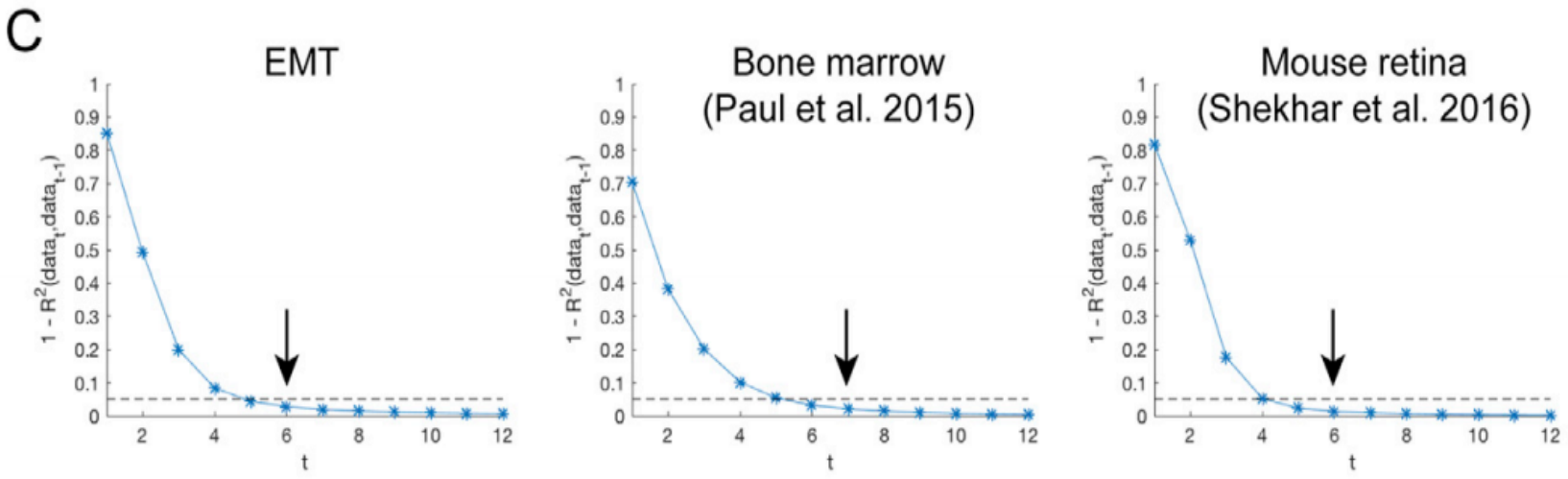


imputation with MAGIC



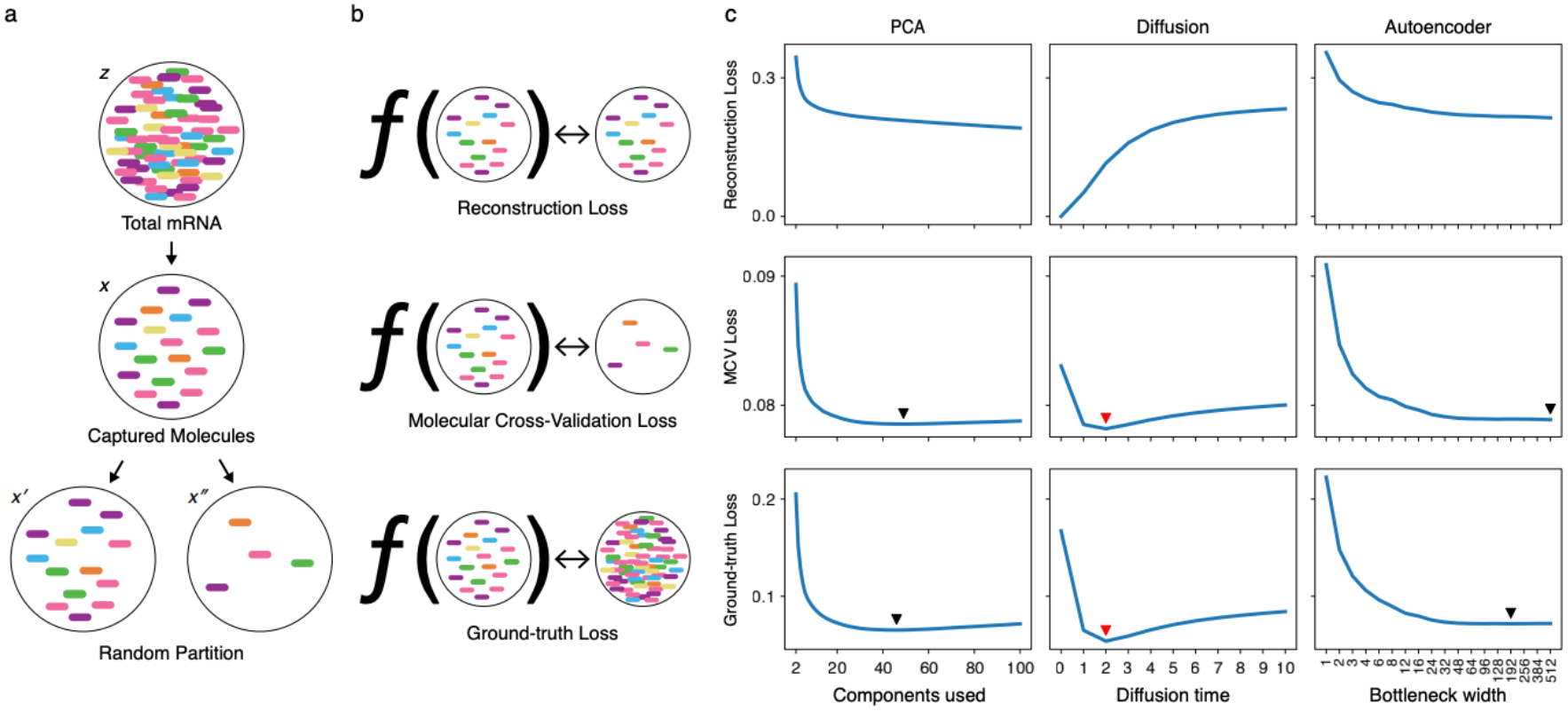
imputation with MAGIC

What t should we choose?



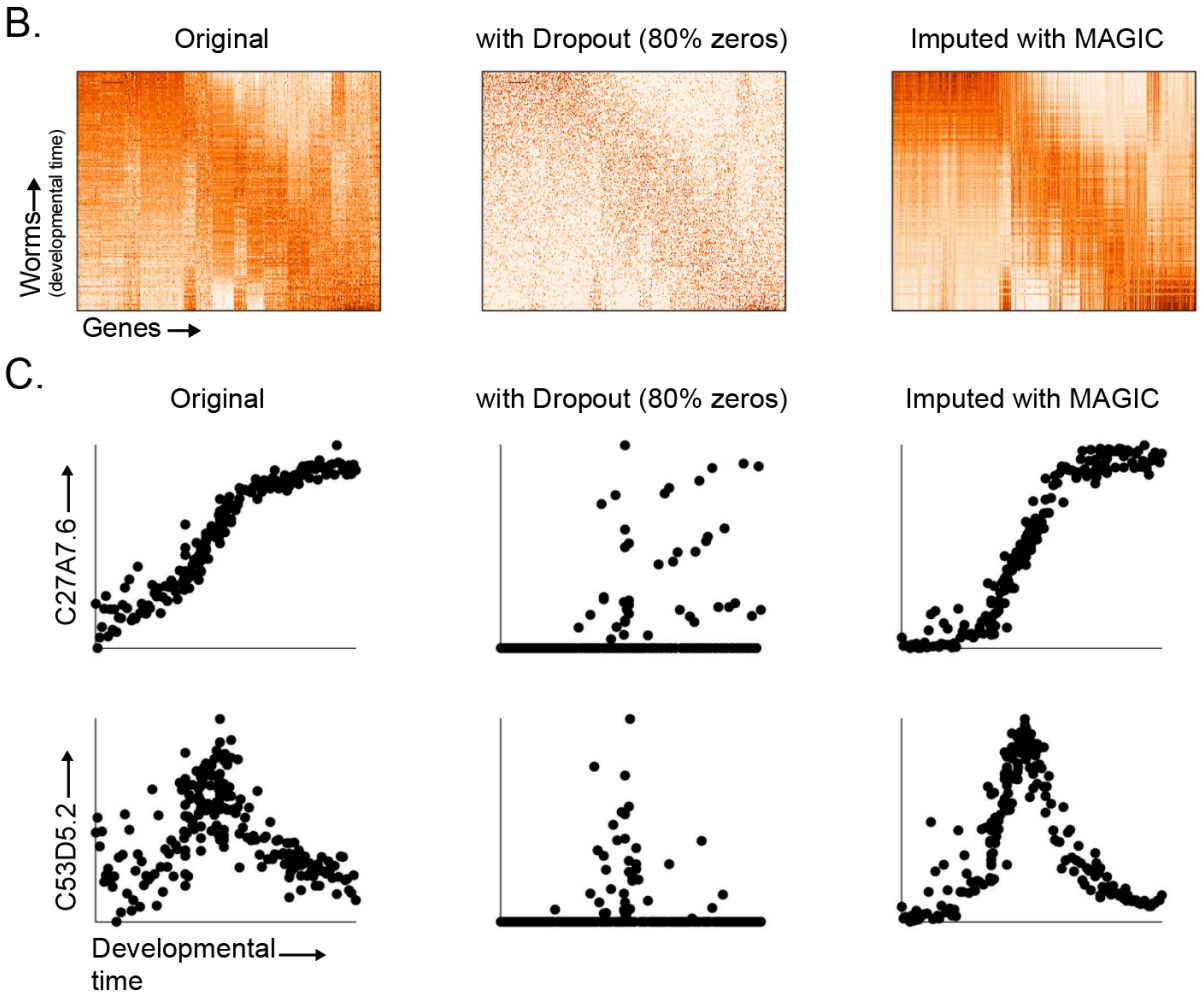
Separate “learning phase” from “stable phase”, quit when data stabilizes

Molecular Validation



https://www.biorxiv.org/content/10.1101/786269v1.abstract

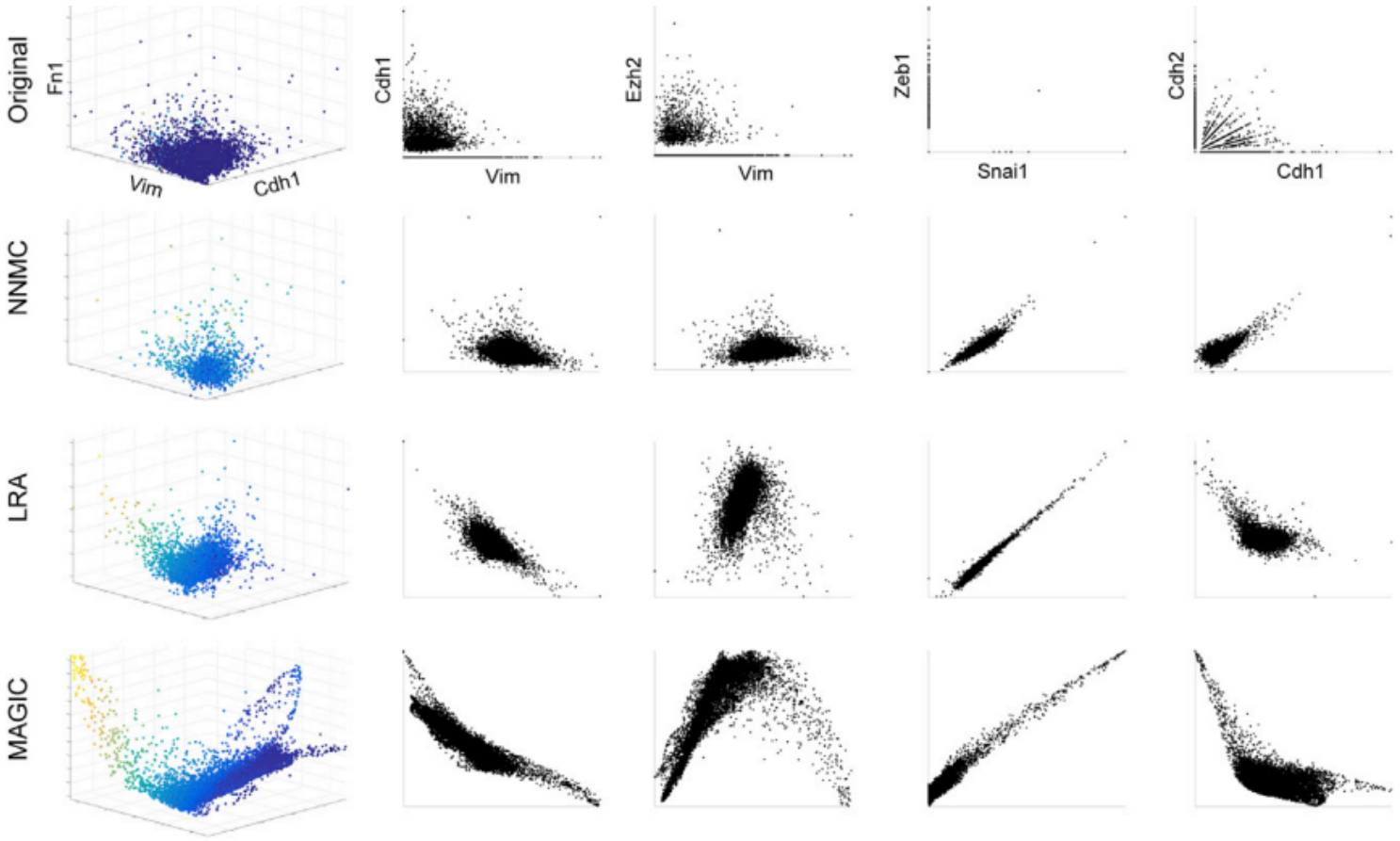
MAGIC recovers gene-gene relationships in an artificially dropped-out dataset



Denoising vs Missing value Imputation

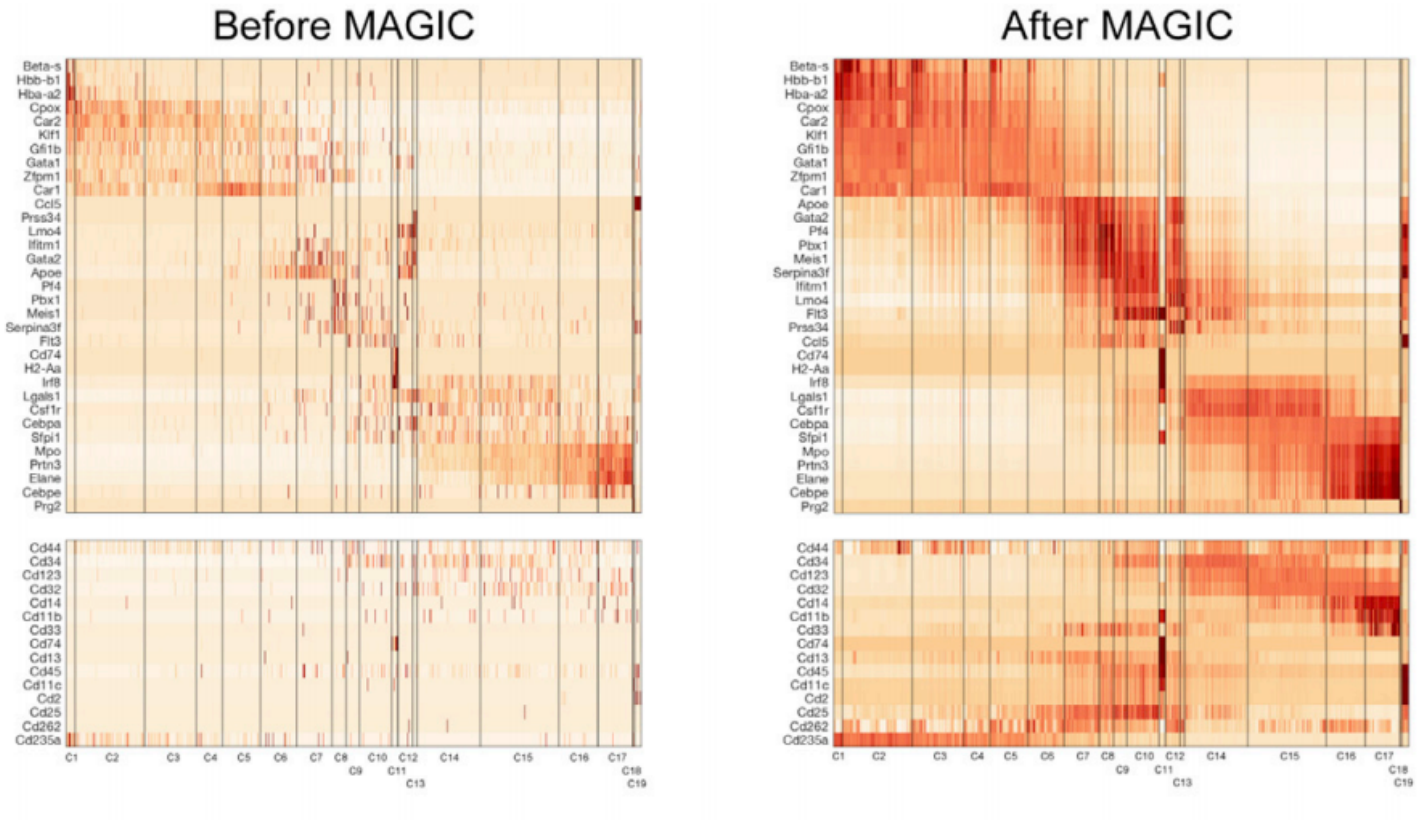
- Denoising assumes that data has some noise, all entries have noise added to them
- Missing value completion methods assume that the existing values are correct and that some values are missing (not measured)
- Matrix completion is a way of filling in missing values
- There is some literature in the single cell world trying to differentiate between true zeros and false zeros
  - People disagree on how this should be done
  - With a denoising viewpoint it is not necessary to distingsuish

Comparing MAGIC to LRA and matrix completion



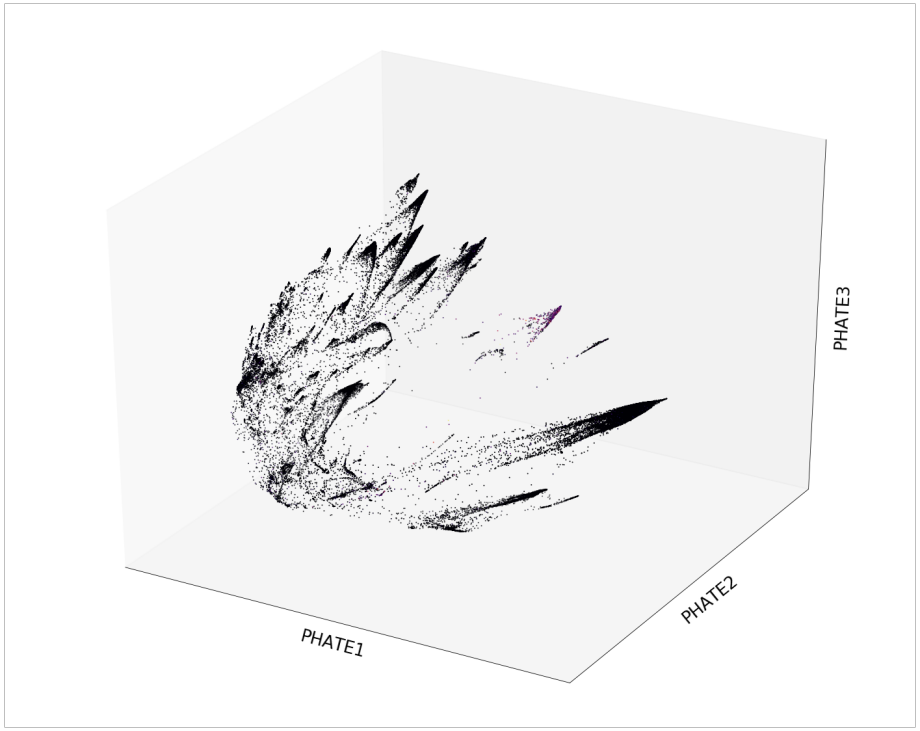
For what tasks is denoising necessary?

Restoring expressions of characteristic markers

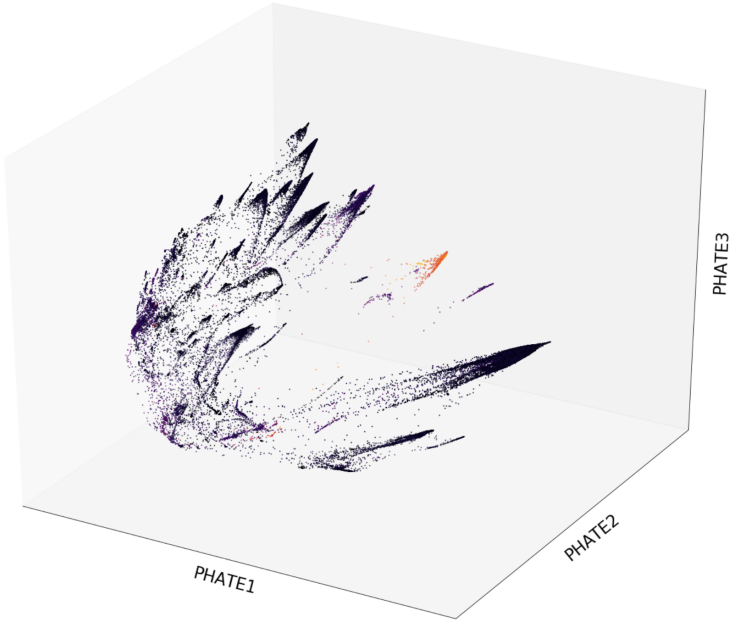


GFAP - marker for astrocytes

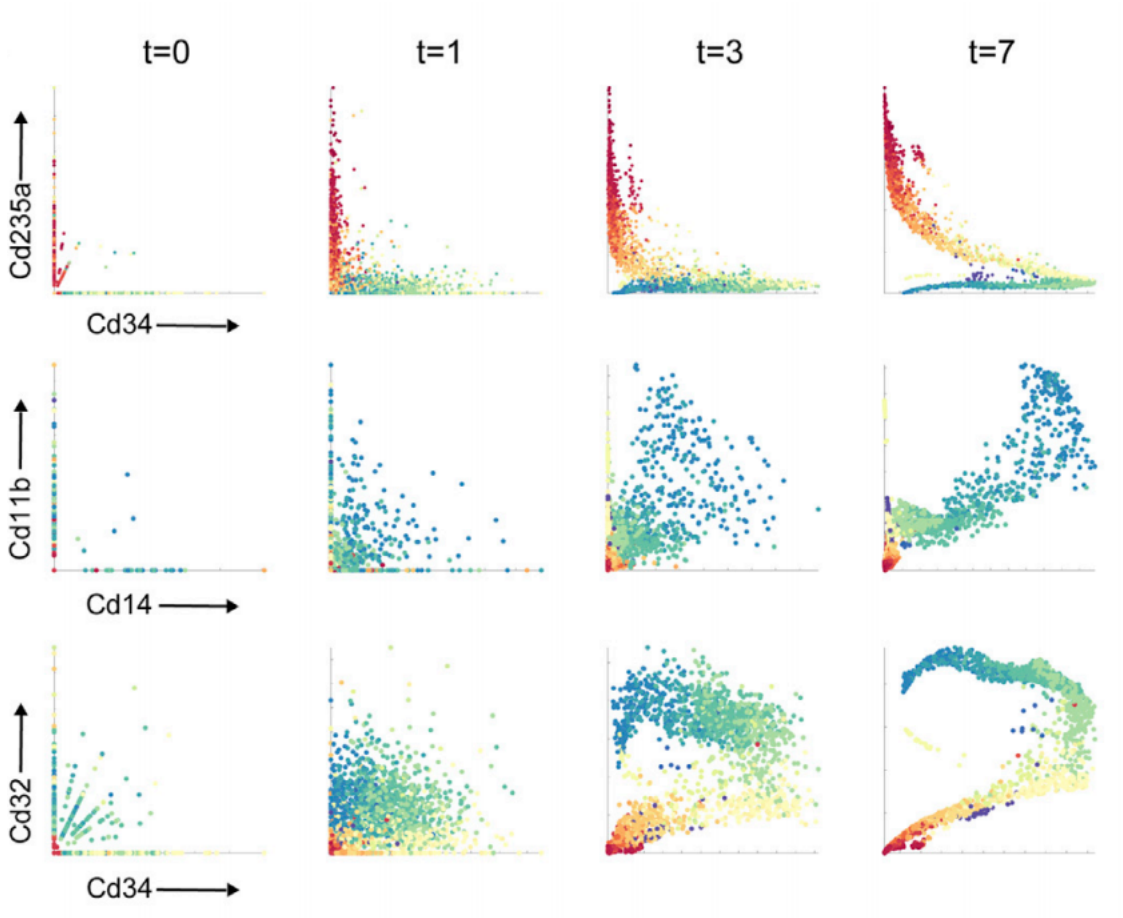
With MAGIC



Without MAGIC



Understanding Gene Gene Relationships



Summary of data denoising

- Diffusing or smoothing values over a graph can denoise data
- This kind of denoising is similar to averaging expression values across neighbors
- More diffusion = more denoising

What questions do you have?

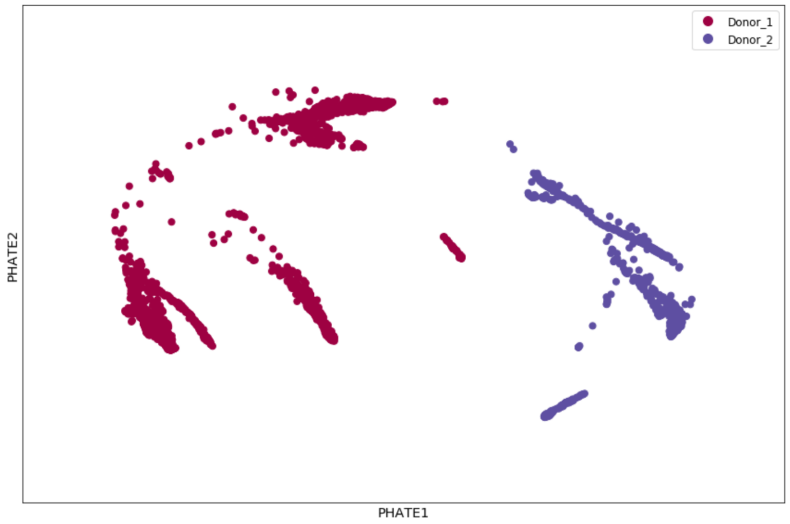
Please submit on Slack

Batch correction

Batch Effects in Single Cell

- **Systematic, non-biological differences between samples due to measurement conditions**
- Differences could be due to ambient conditions (temperature, humidity), machine calibration, differences in titration, bead/antibody batch, etc
- Sometimes refers to actually actual biological differences but those that are “uninteresting” (background demographics rather than immediate drug effect).

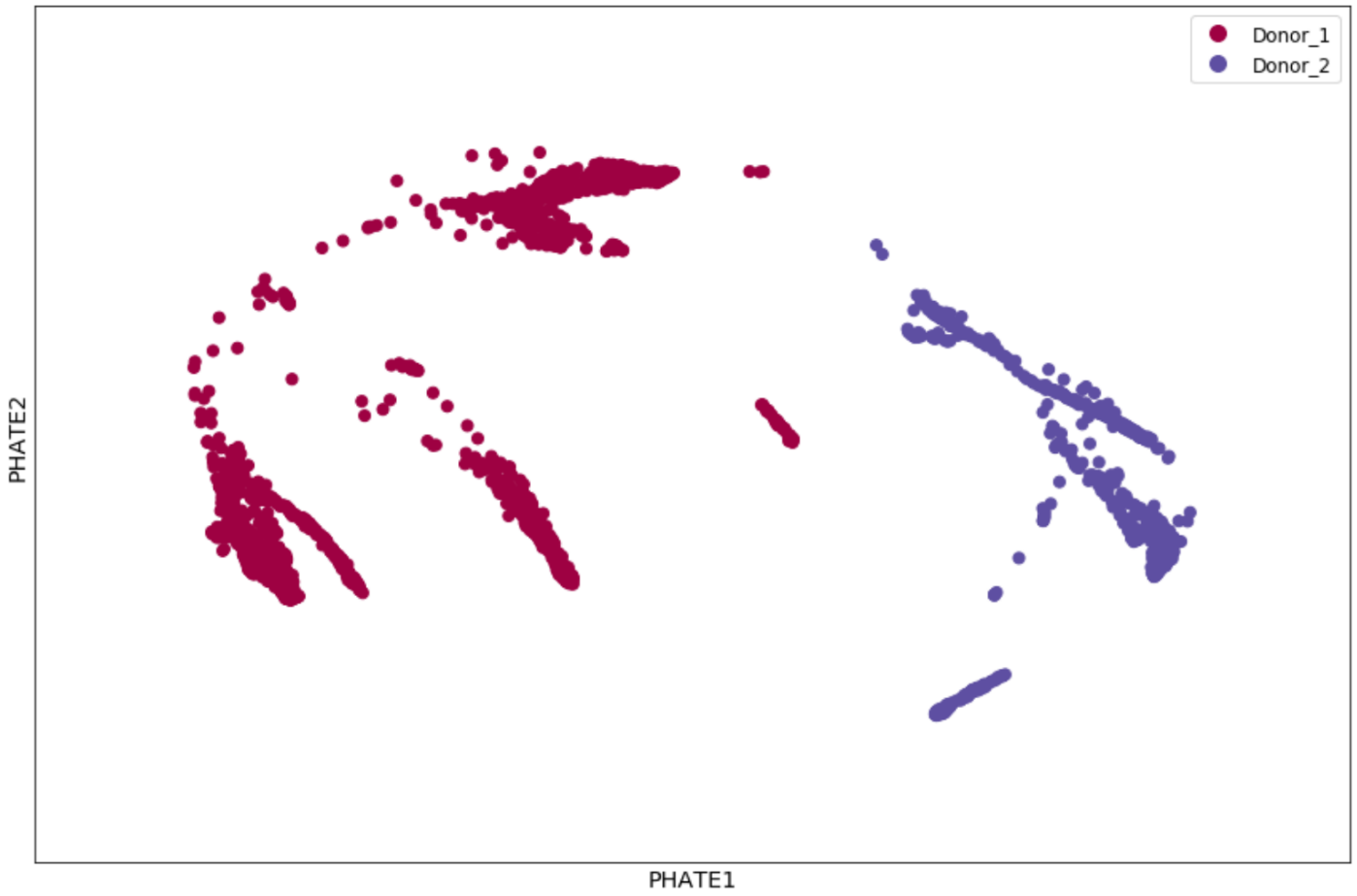
Problem with Batch Effect



Samples become hard to compare

All genes, cell types seem different!

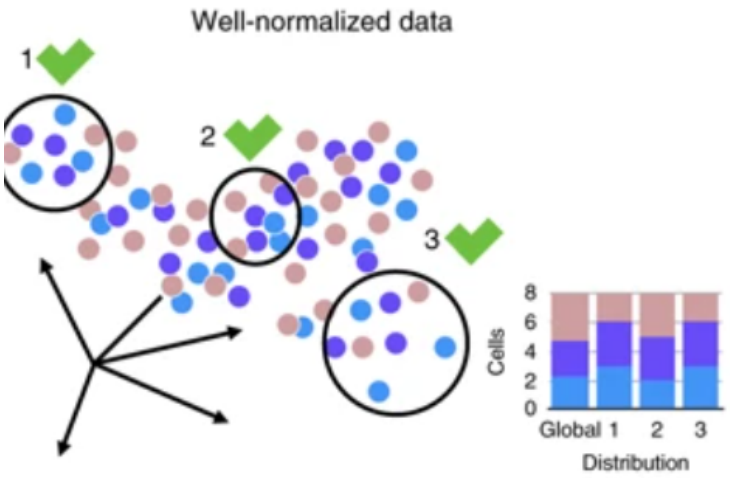
How do we detect batch effect?



Visualization!

Diversity Measures

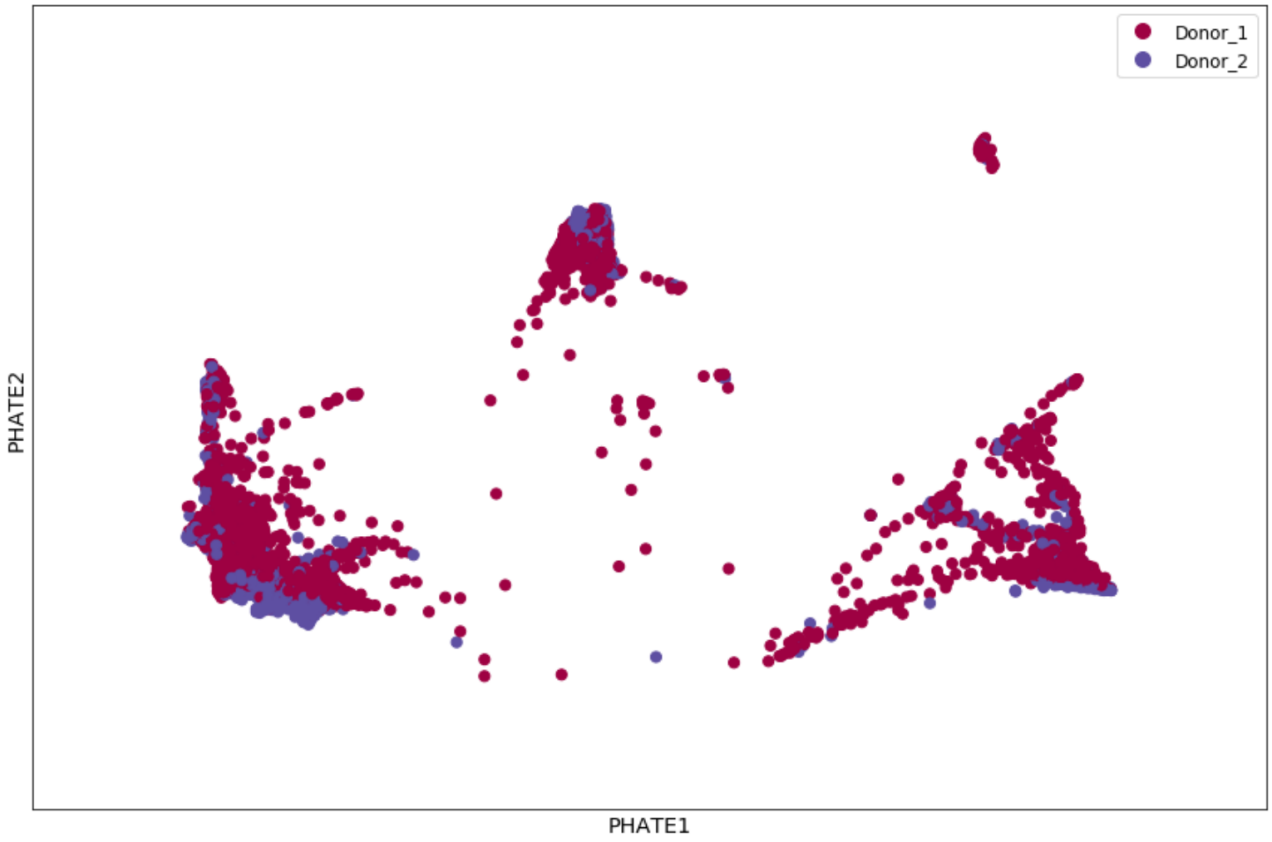
$$H' = -\sum_{i=1}^{R} p_i \ln p_i$$



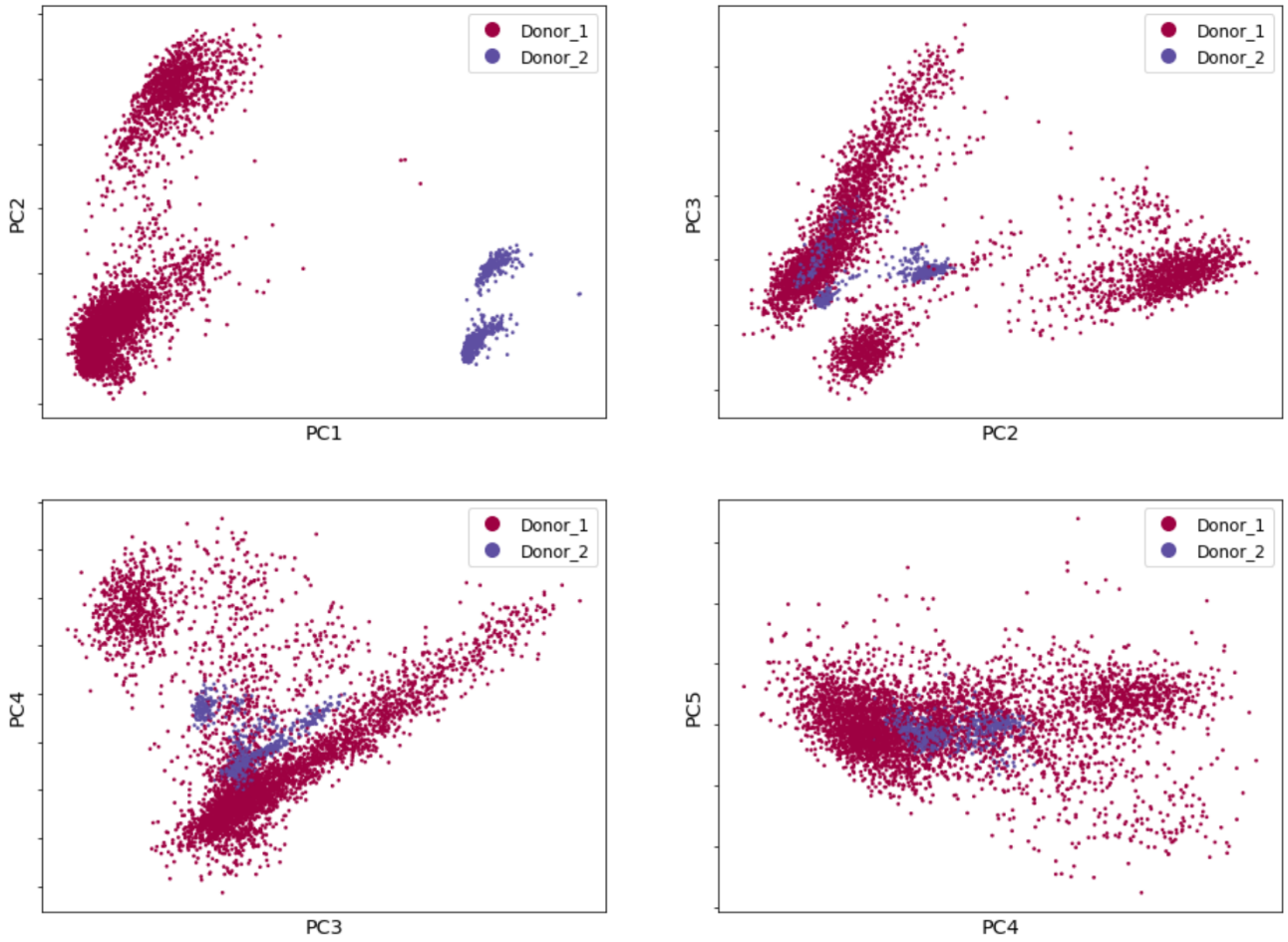
- **Diversity measures including Shannon entropy have been used to quantify the degree of "mixture" between samples**
- kBet [Buttner et al] compare global population proportions to local
- A priori, we don't know what the degree of mixture should be, but if batch correction works, it should increase

Batch Correction

Take off artifactual variation and keep biological variation!



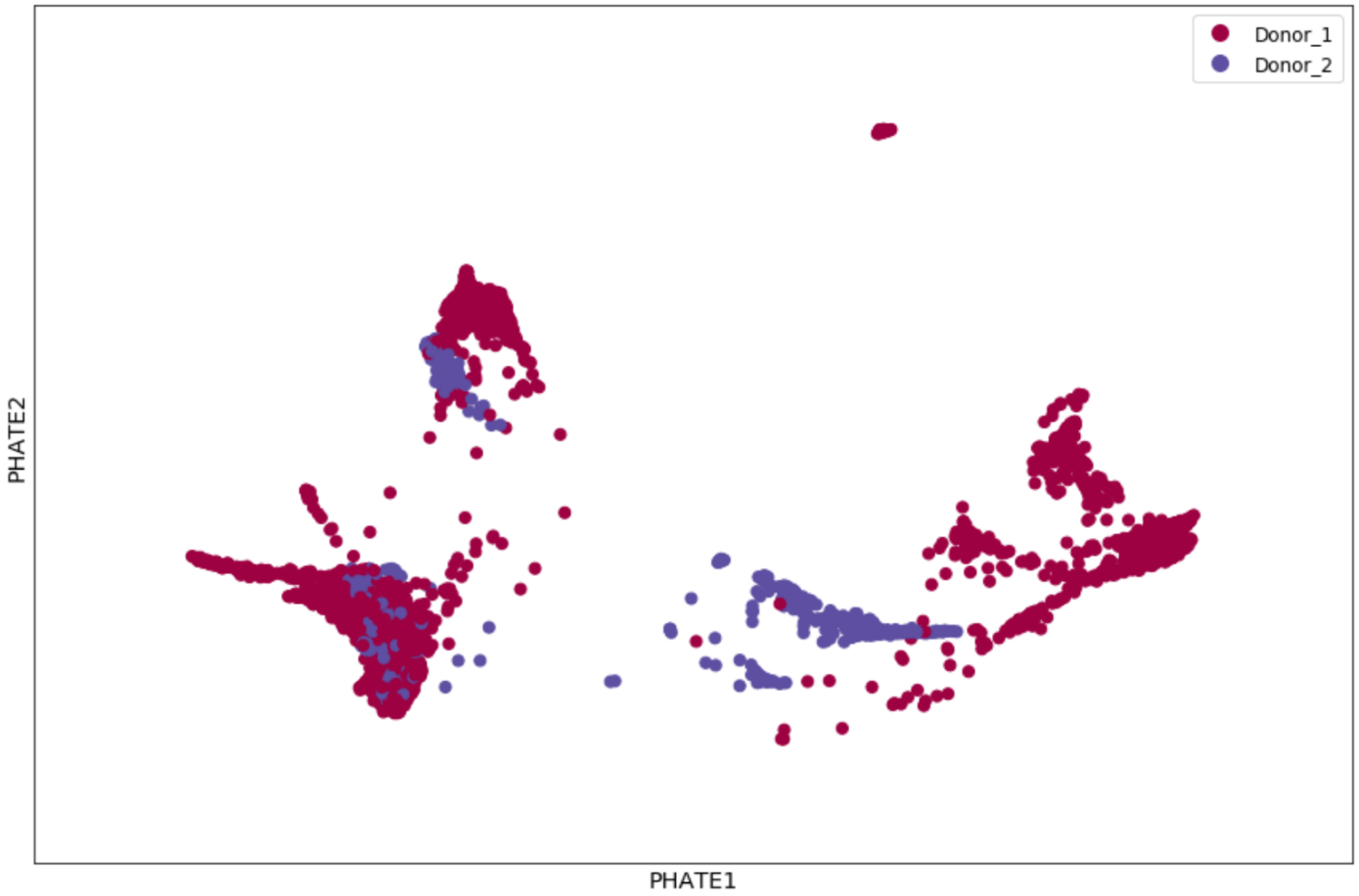
Can we use PCA to correct batch effect?



How could we use PCA or SVD to remove batch effect?

This noise is large scale

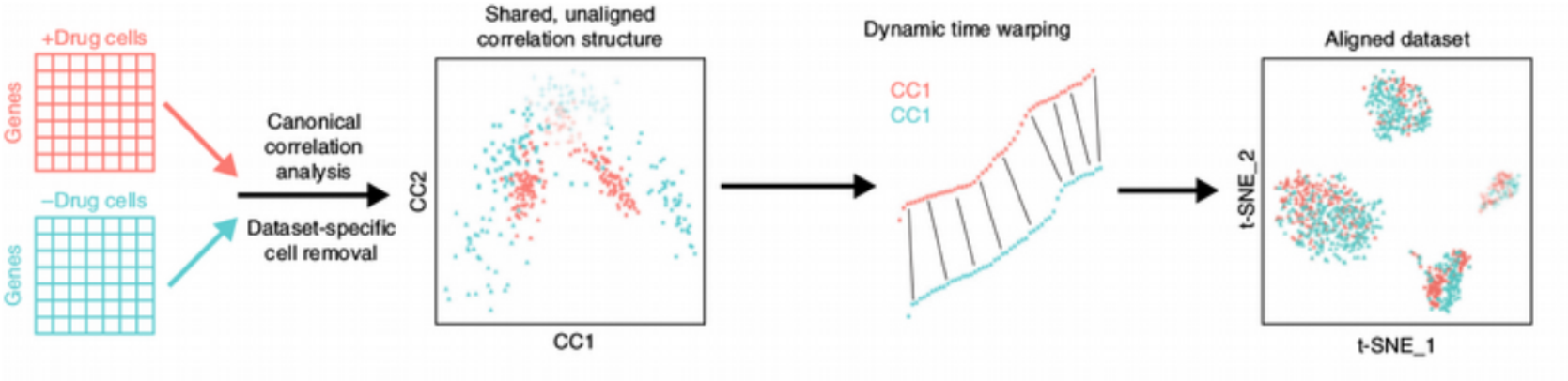
- Could be potentially addressed by removing first PC



Problem with PCA

- It can address a linear shift in the data
- But cannot handle more complex non-linear effects

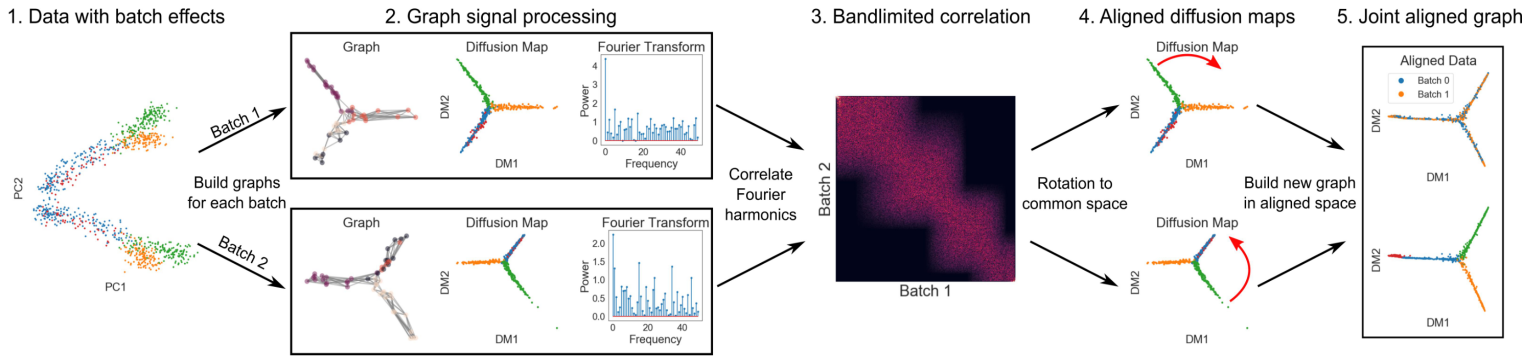
Cannonical Correlation Analysis



CCA : uses a variant of PCA that brings two sets of samples to a common space

- Can cluster or perform other analysis in common space but cannot go back to original space
- Finds linear axes of common variation
- Thrown off by non-matching populations

Harmonic Alignment

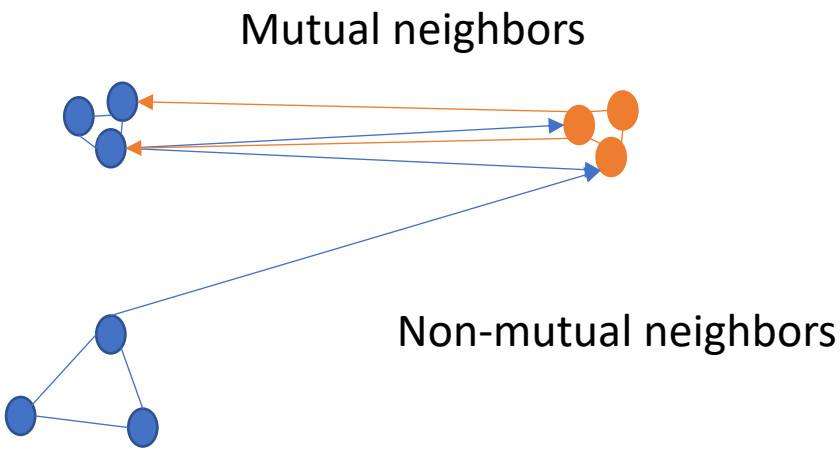


Aligns the diffusion components of two datasets using a rigid rotation that maximizes correlation of gene loadings onto eigenvectors

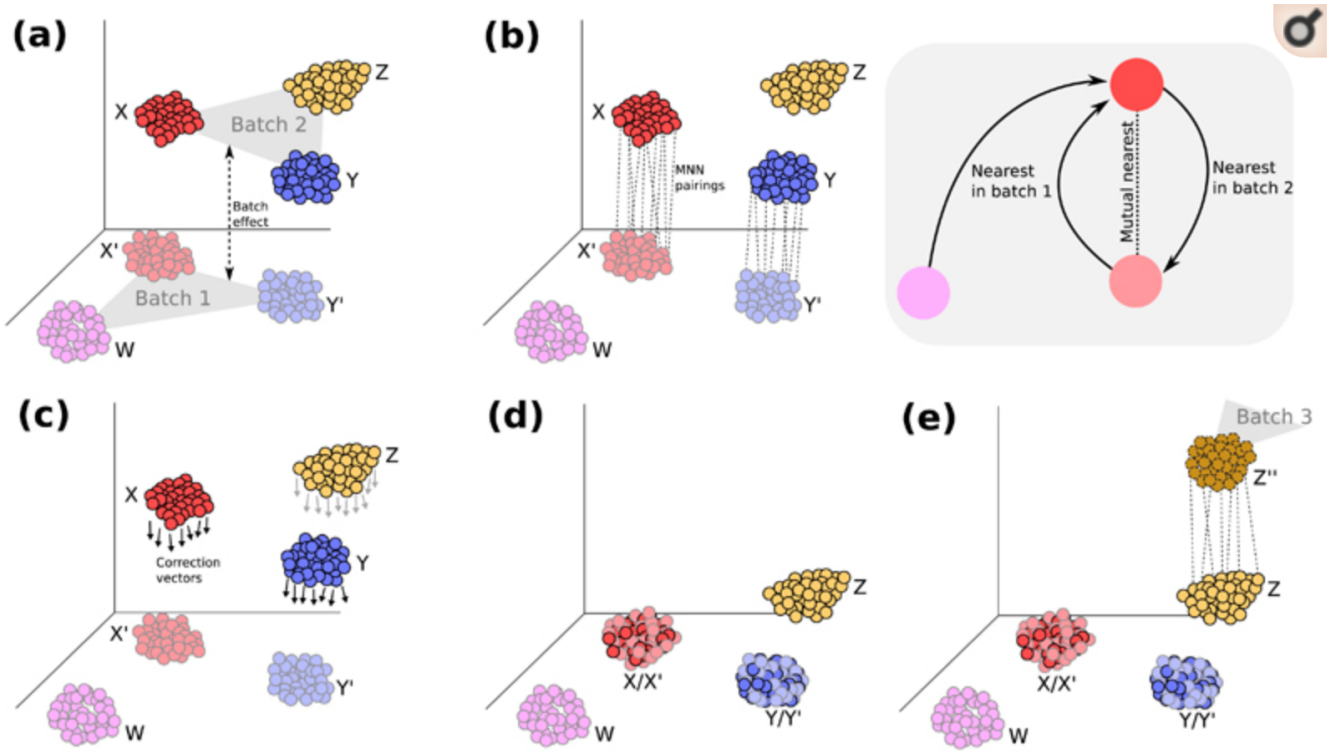
Simple rotation does not seem to handle all cases

Mutual Nearest Neighbors

- Creates a graph between two datasets
- Nearest neighbors in the other dataset could be “matching cells”
- But they have to be mutual!

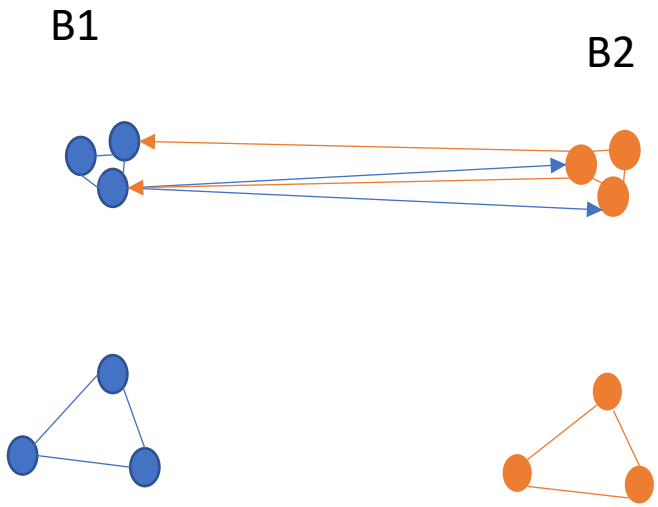


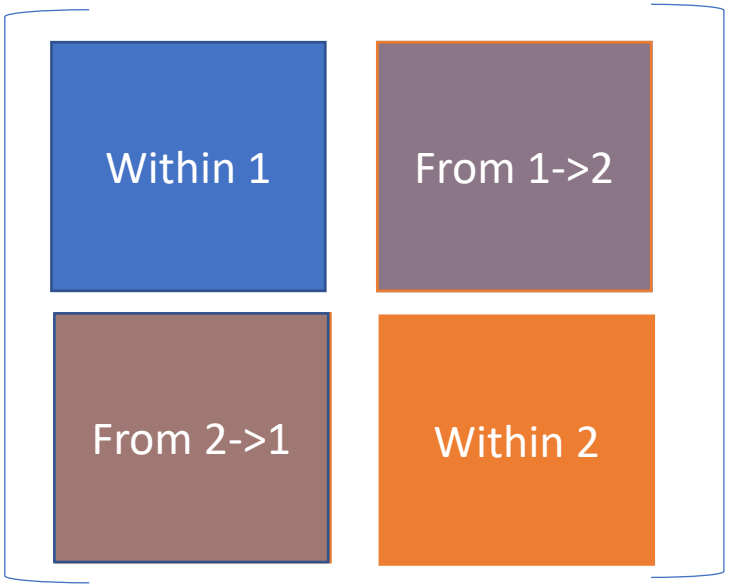
Mutual Nearest Neighbors



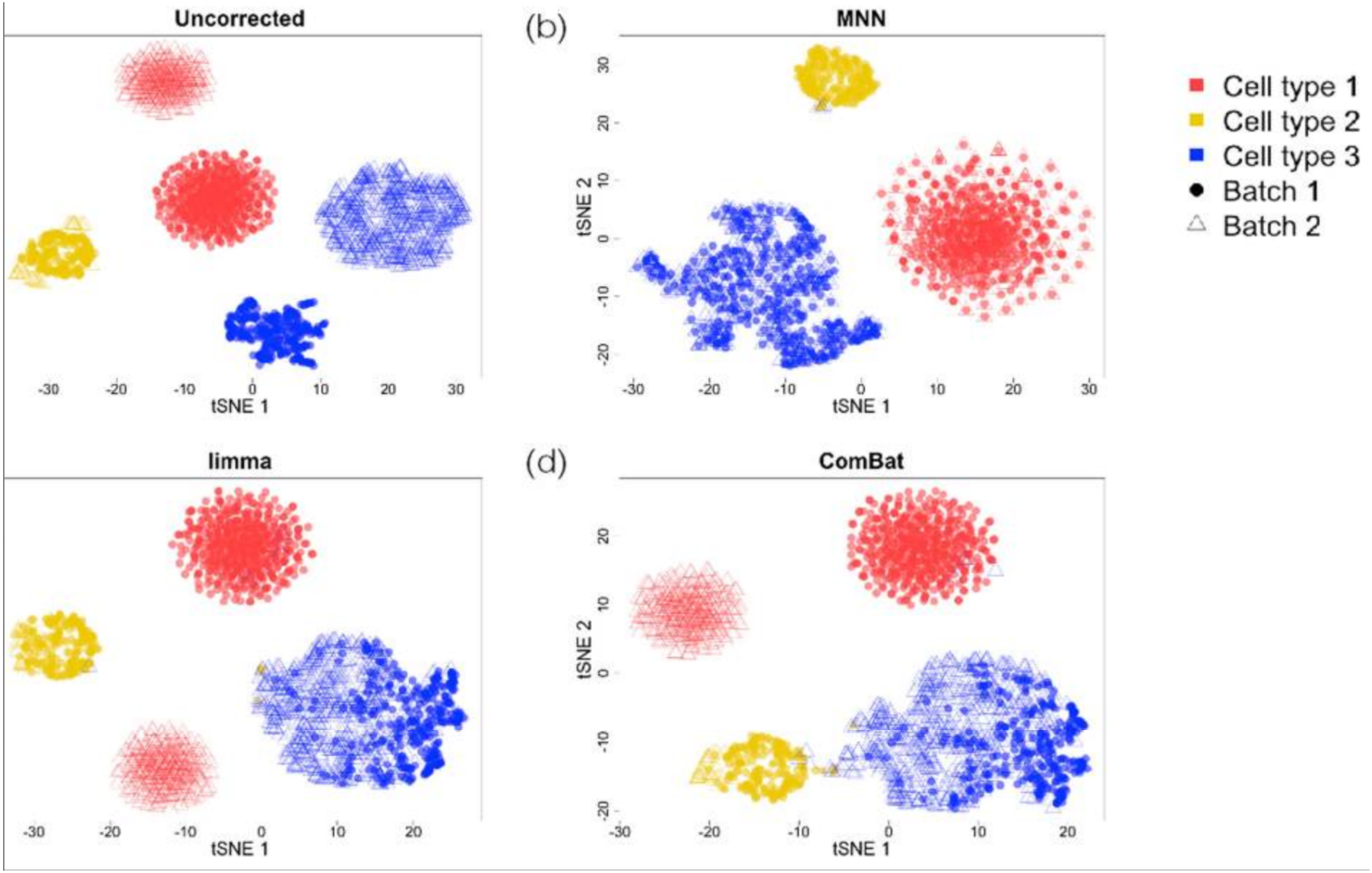
Use MNNs to find out the direction of the batch effect and correct it.

MNN with MAGIC

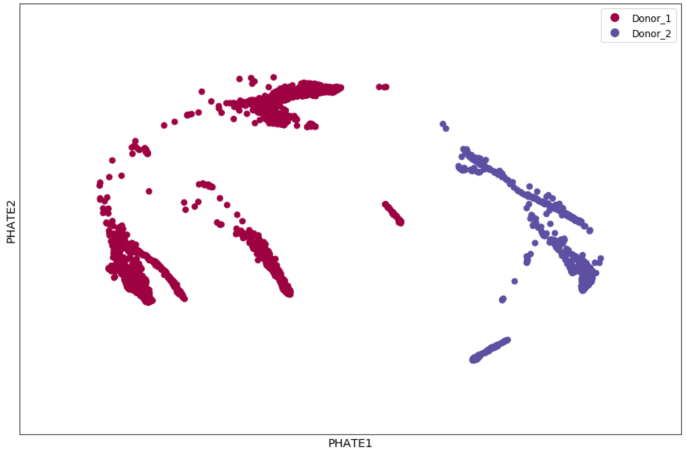


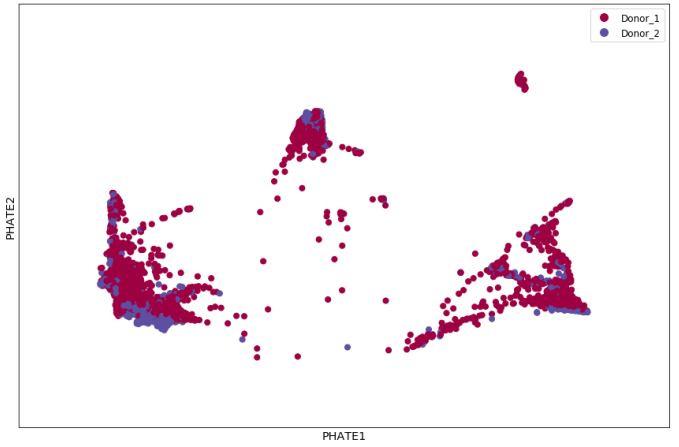


Uses softer affinity matrix and pulls the data in to correct the result


Uncorrected
(b)
MNN
limma
(d)
ComBat
tSNE 1
tSNE 2
Cell type 1
Cell type 2
Cell type 3
Batch 1
Batch 2

MNN with MAGIC Correction





Is mutual nearest neighbor (MNN) normalization linear or non-linear?

Linear

Non-linear

Summary of data denoising

- Batch effects are sample-specific changes in measurements
- The goal of batch-normalization is to align cells of the same “type” across samples
- Mutual nearest neighbors (MNN) normalized batches by matching cells that are both close to each other

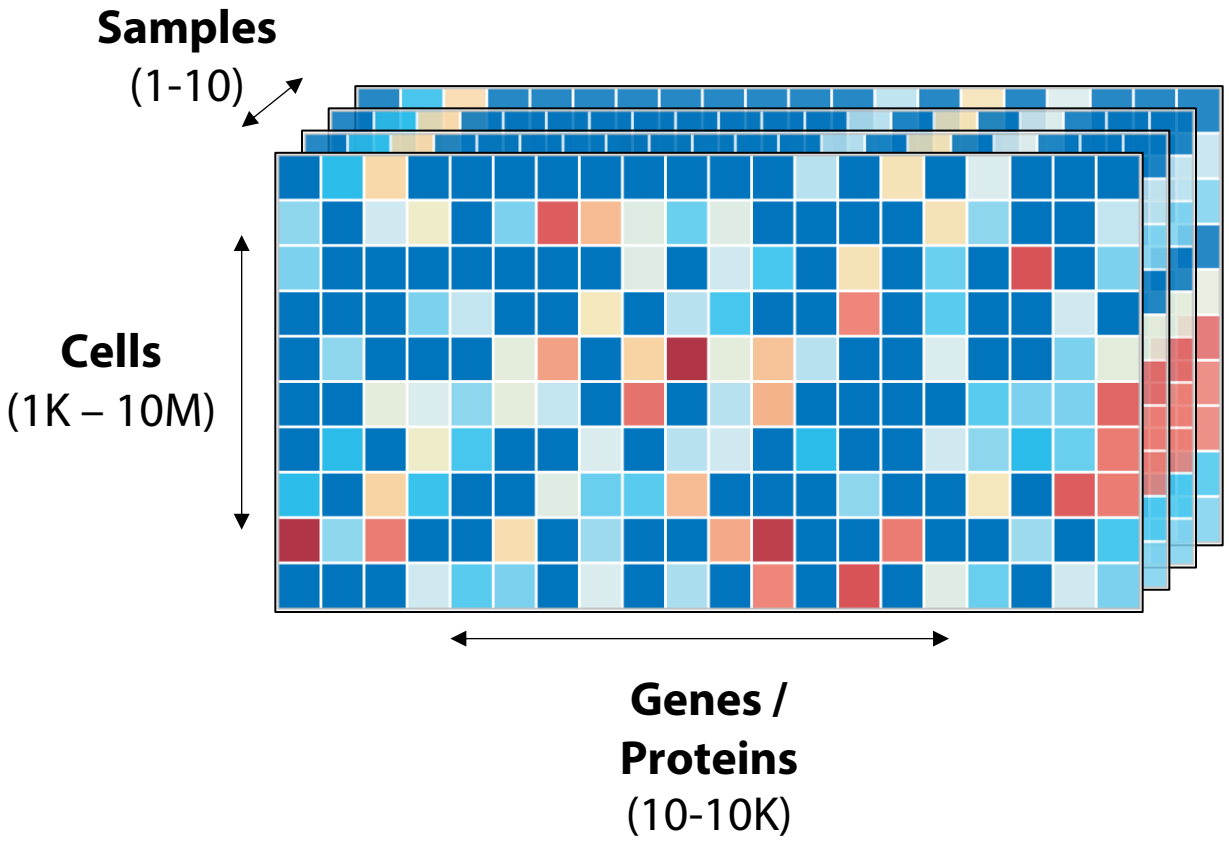
What questions do you have?

Please submit on Slack

Compositional Analysis

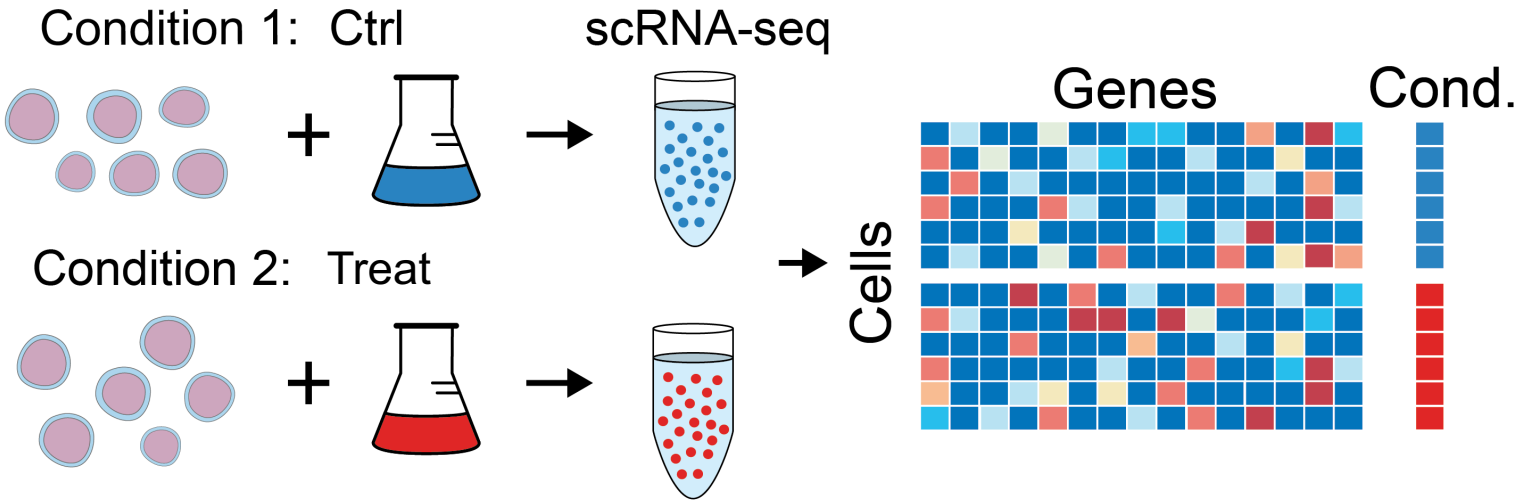
http://bit.ly/MELD_lecture

Single Cell Data

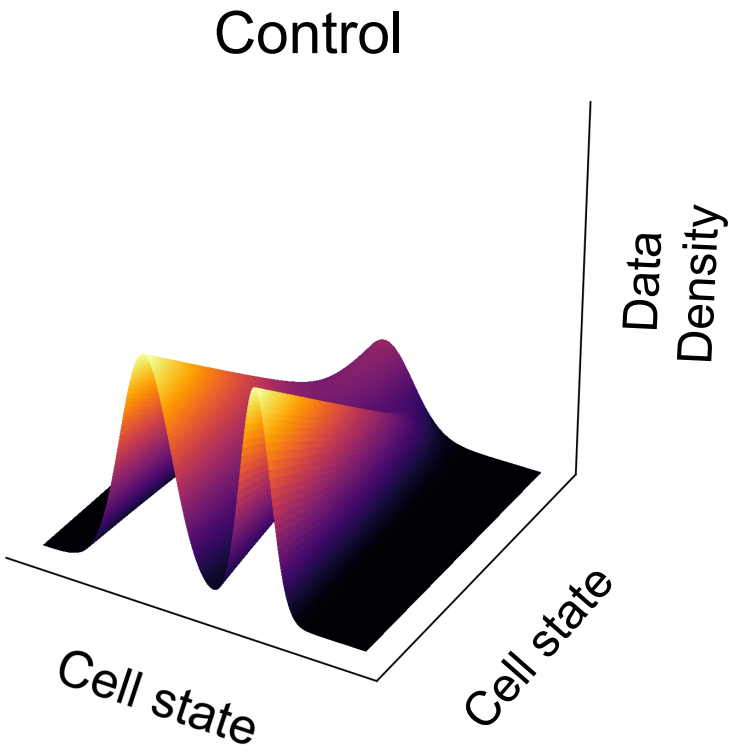


http://bit.ly/MELD_lecture

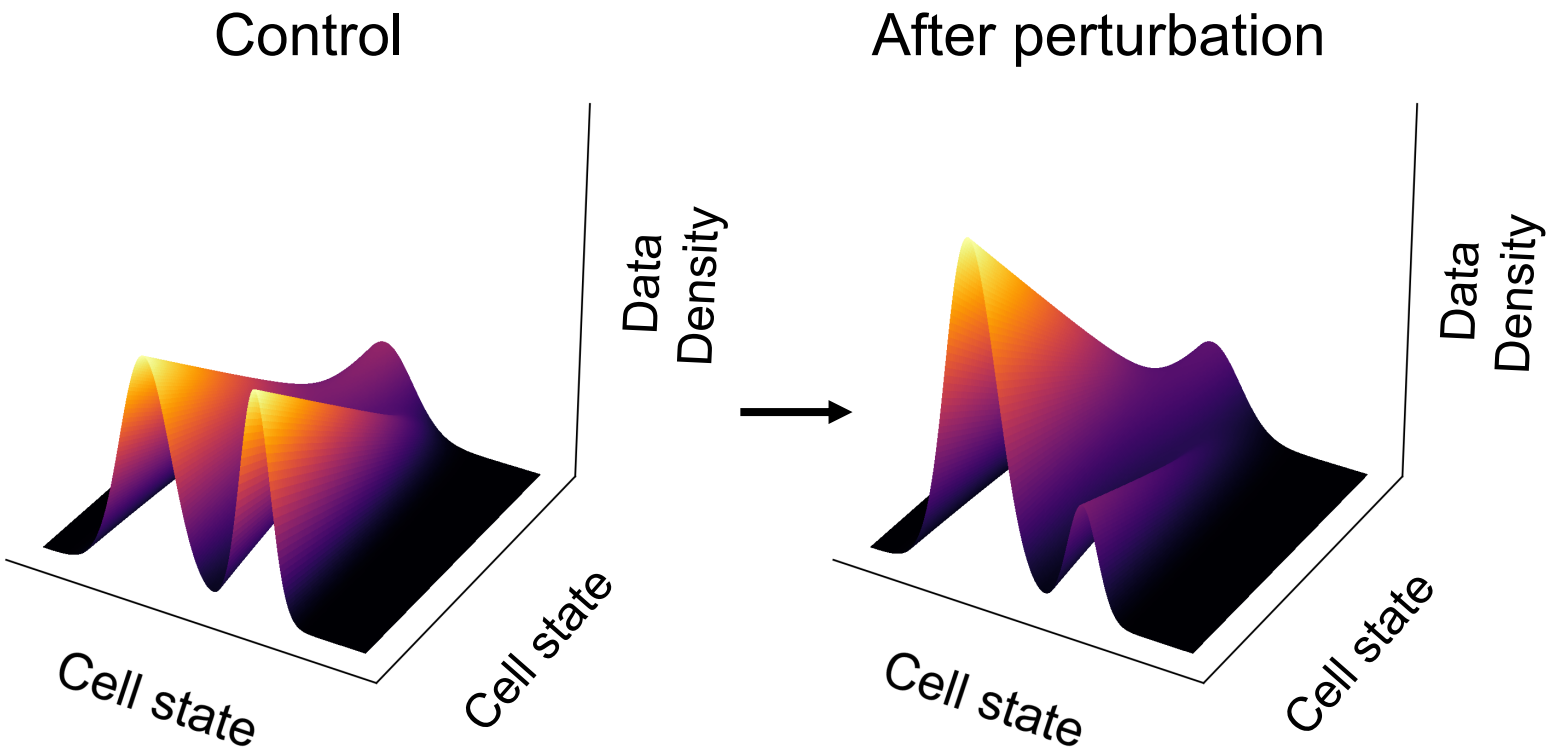
Common experimental setup



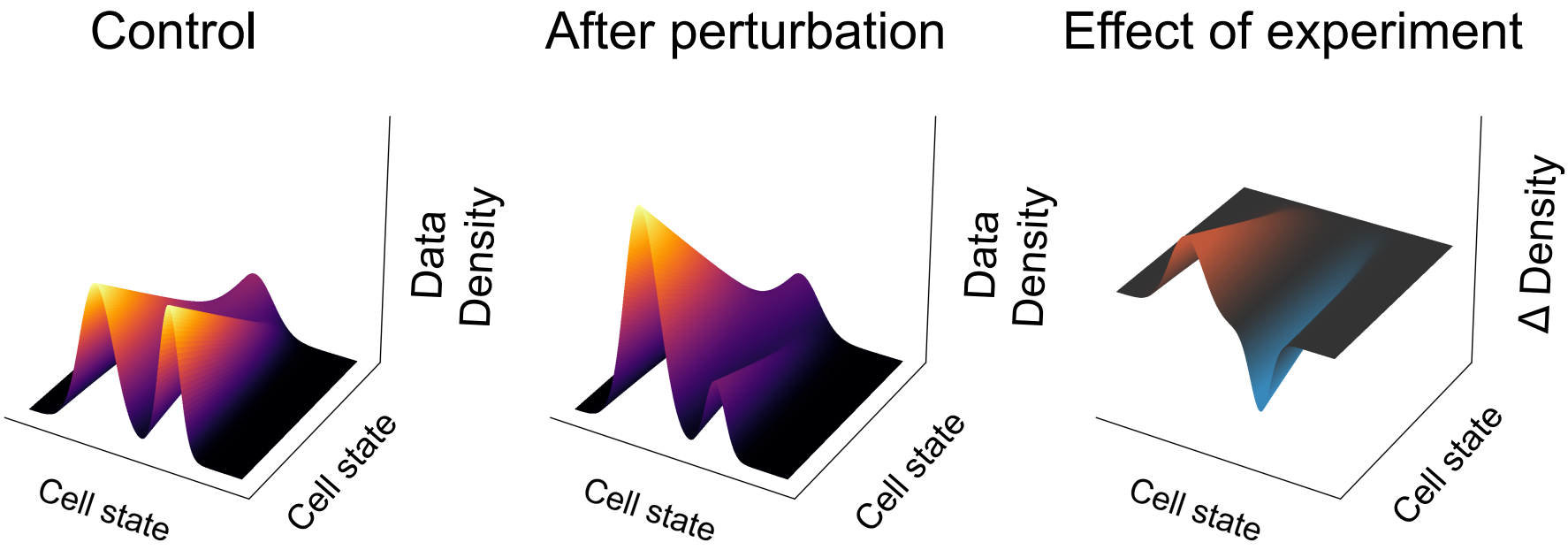
Goal: Quantify difference in data density



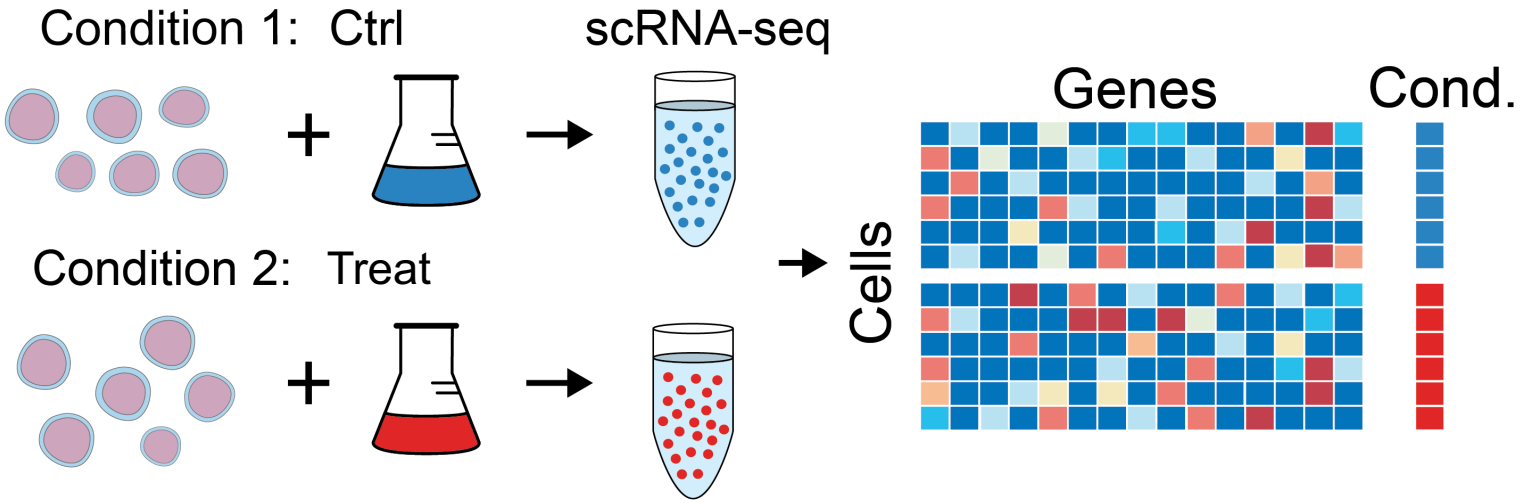
Goal: Quantify difference in data density



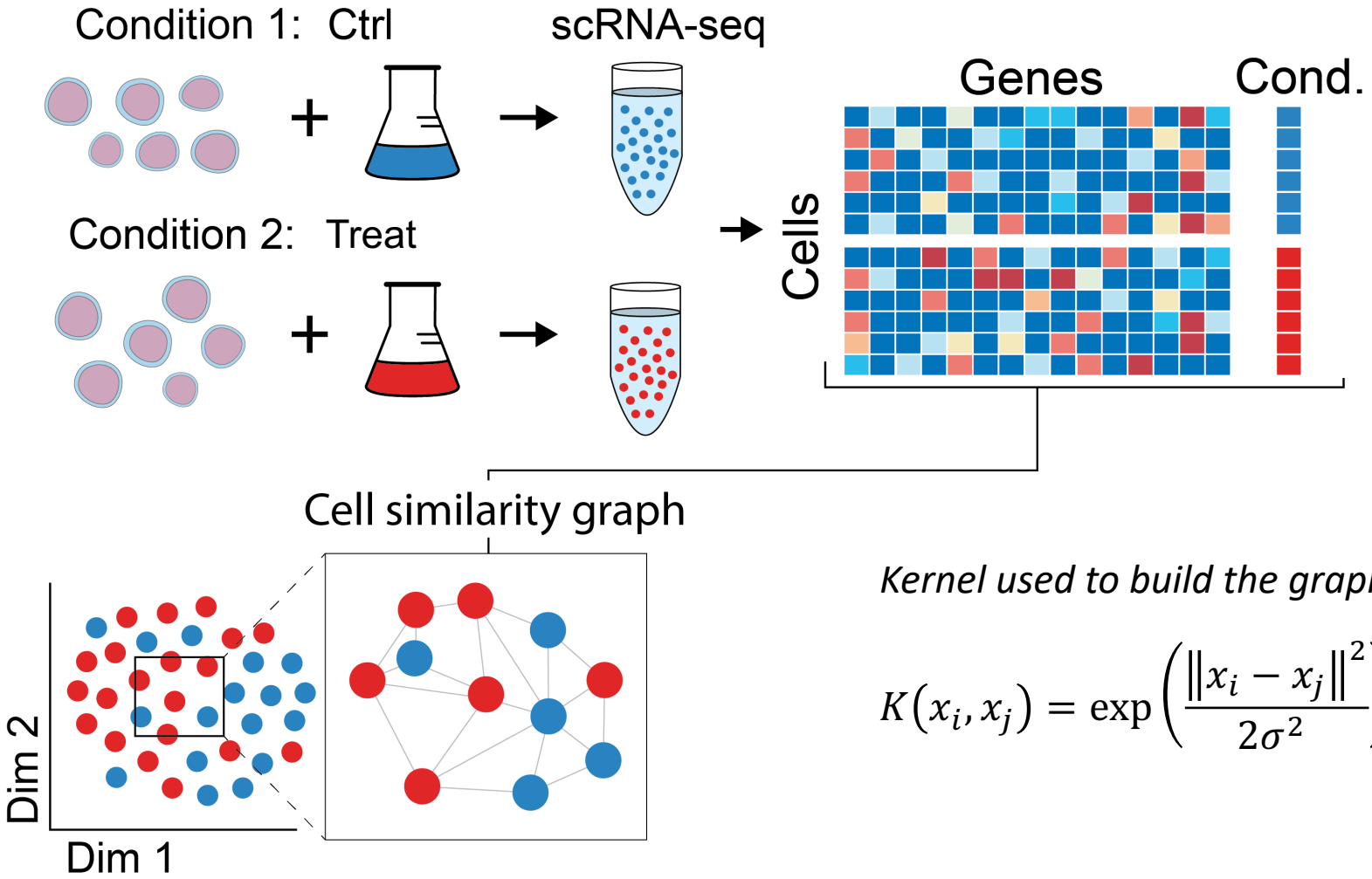
Goal: Quantify difference in data density



Estimating sample density over a graph



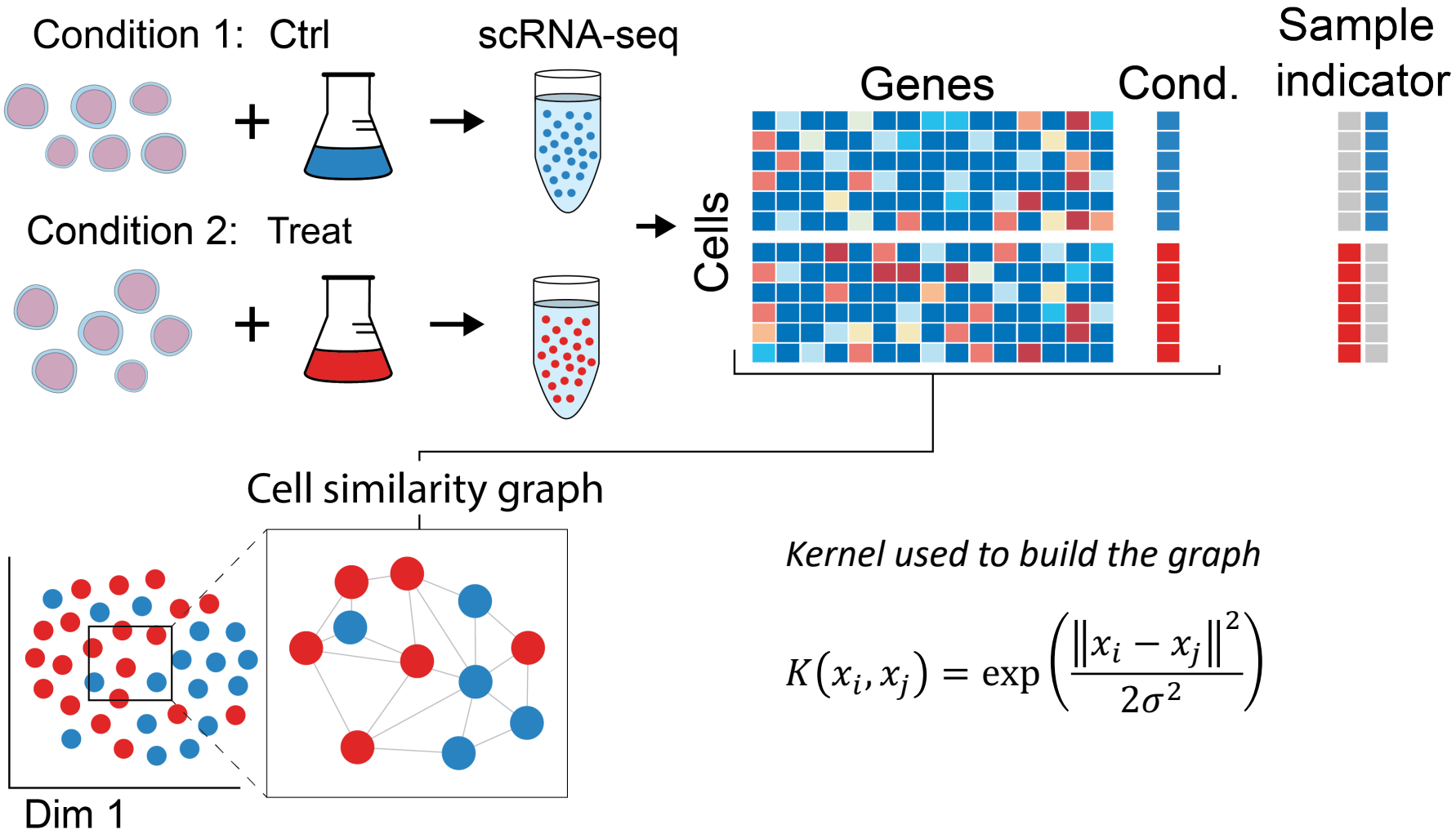
Estimating sample density over a graph



Kernel used to build the graph

$$K(x_i, x_j) = \exp\left(\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

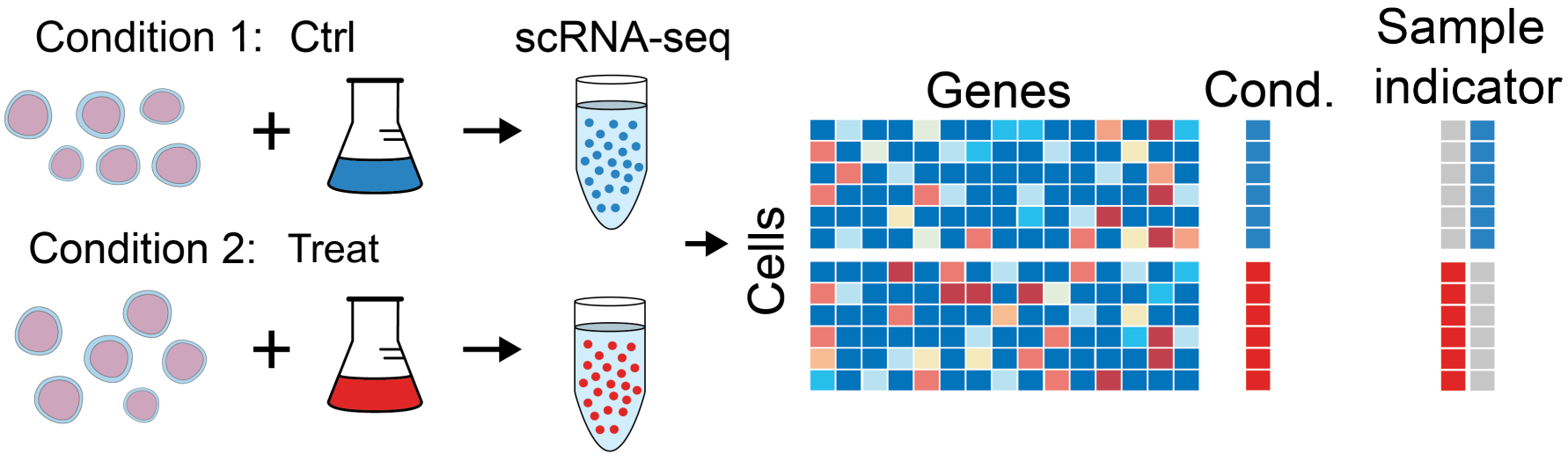
Estimating sample density over a graph



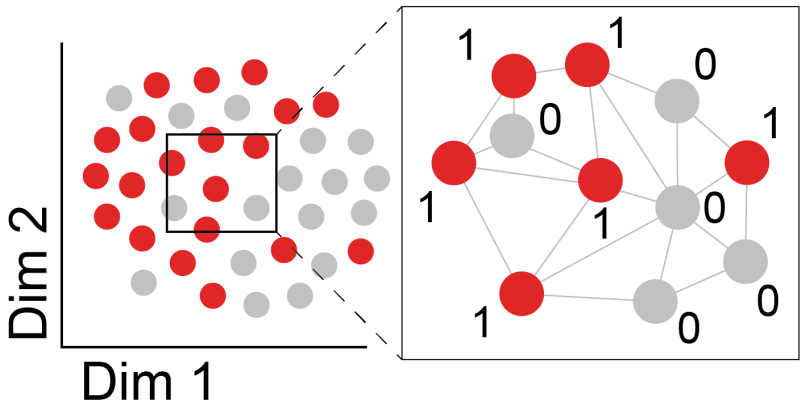
Kernel used to build the graph

$$K(x_i, x_j) = \exp\left(\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

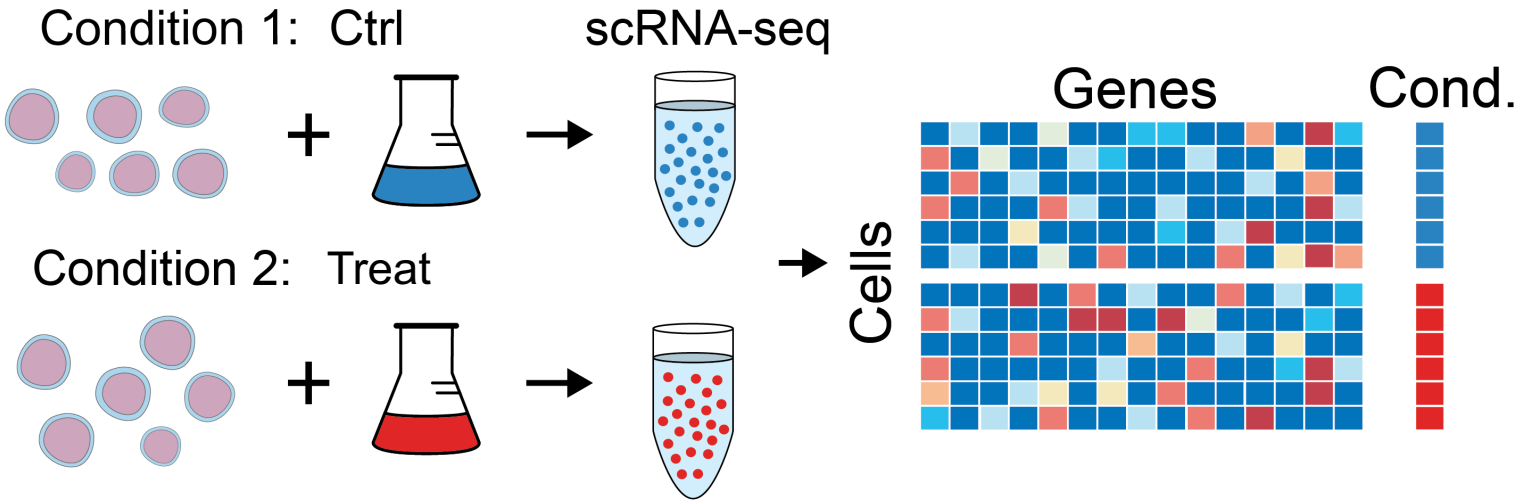
Estimating sample density over a graph



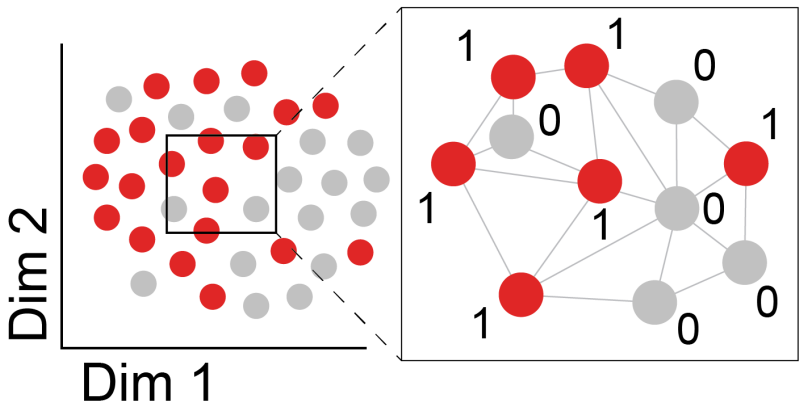
Treatment-associated indicator signal



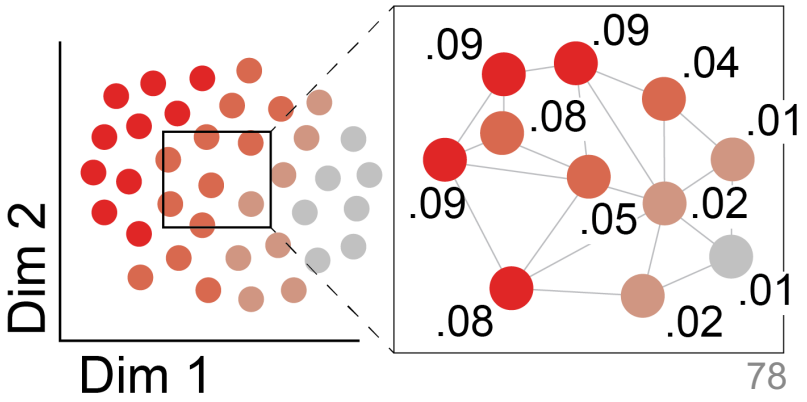
Estimating sample density over a graph

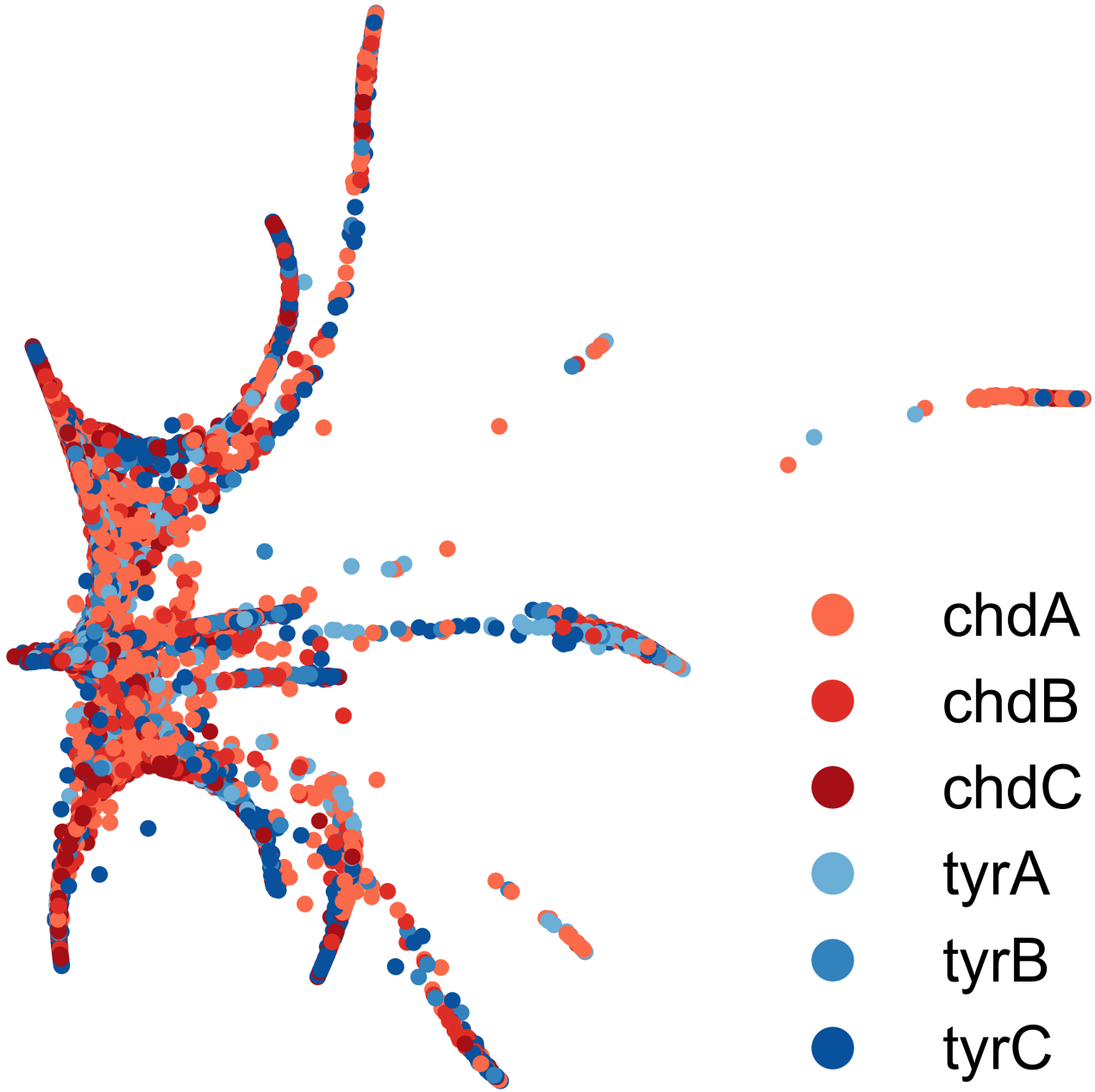


Treatment-associated indicator signal

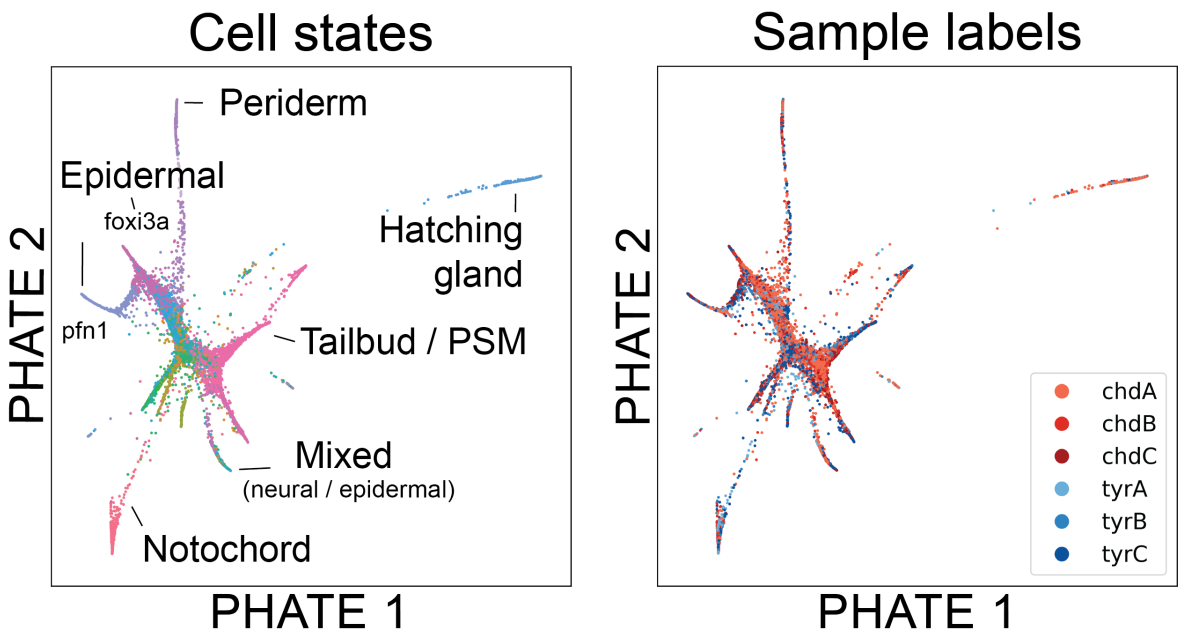


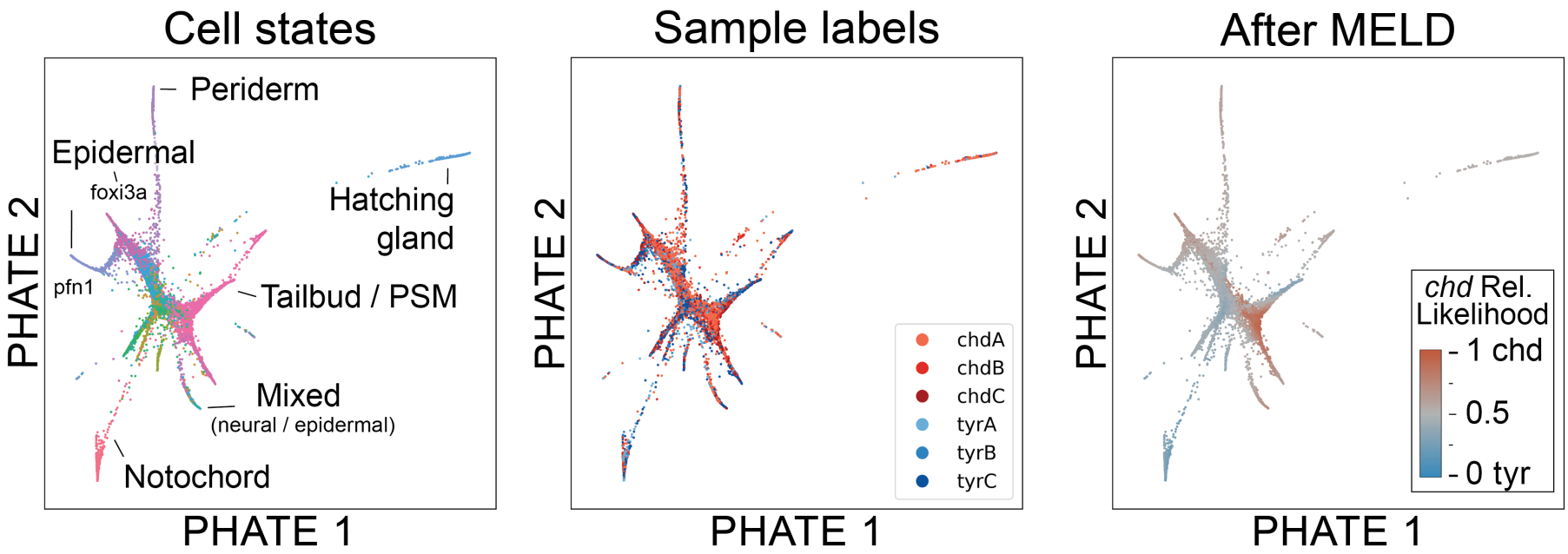
Treatment-associated sample density



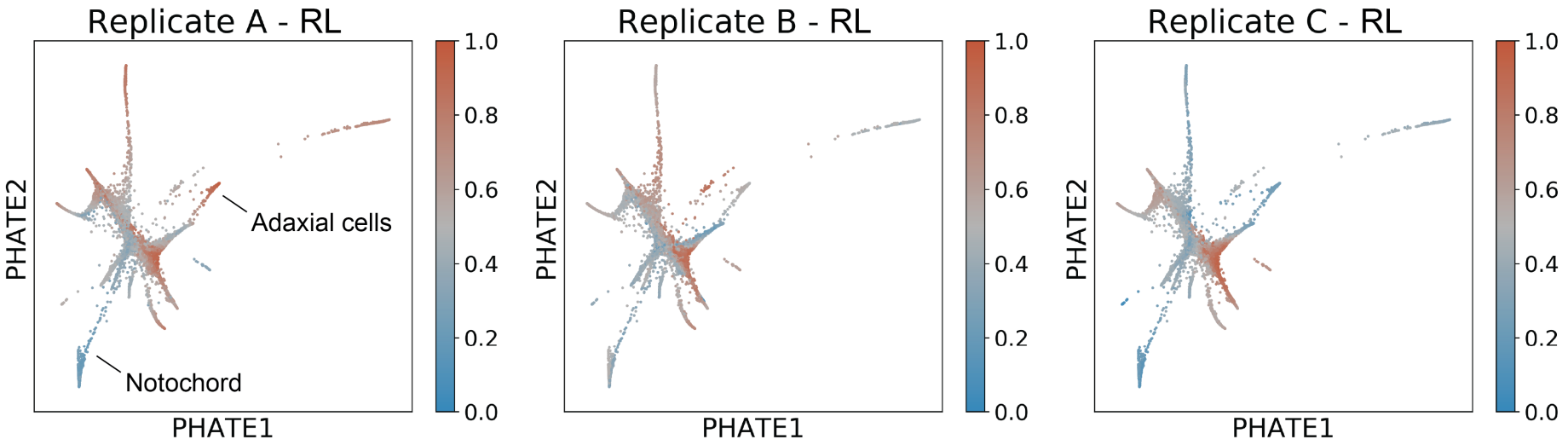


Wagner D. et al. *Science.* 2018.

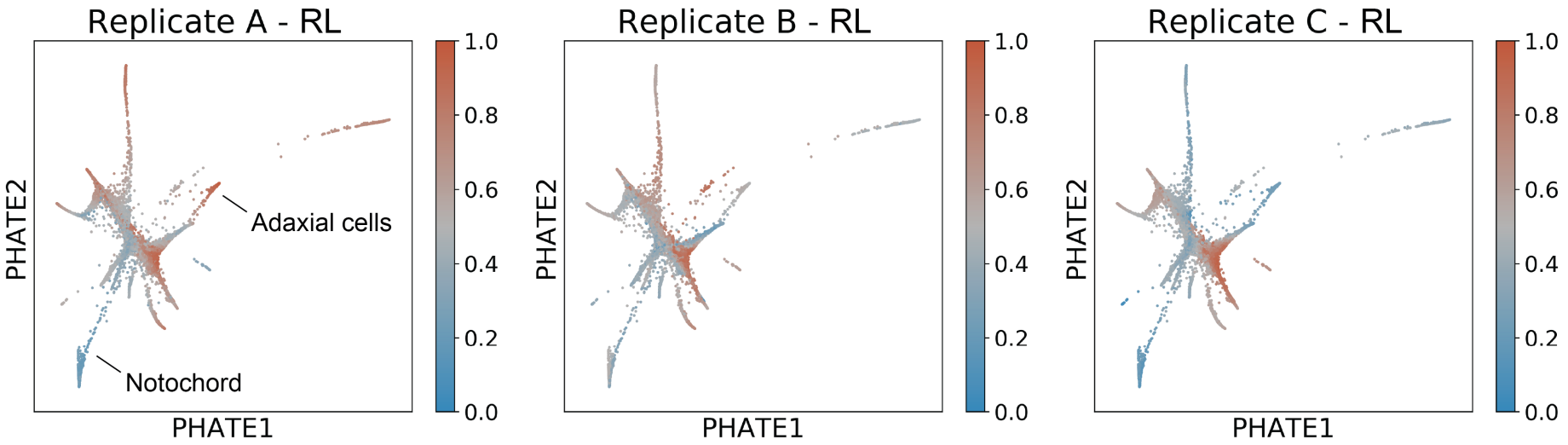

Cell states
Periderm
Epidermal
foxi3a
pfn1
Hatching
gland
Tailbud / PSM
Mixed
(neural / epidermal)
Notochord
PHATE 2
PHATE 1
Sample labels
chdA
chdB
chdC
tyrA
tyrB
tyrC
PHATE 2
PHATE 1


Cell states
Periderm
Epidermal
foxi3a
pfn1
Hatching gland
Tailbud / PSM
Mixed
(neural / epidermal)
Notochord
PHATE 2
PHATE 1
Sample labels
chdA
chdB
chdC
tyrA
tyrB
tyrC
PHATE 2
PHATE 1
After MELD
chd Rel. Likelihood
1 chd
0.5
0 tyr
PHATE 2
PHATE 1

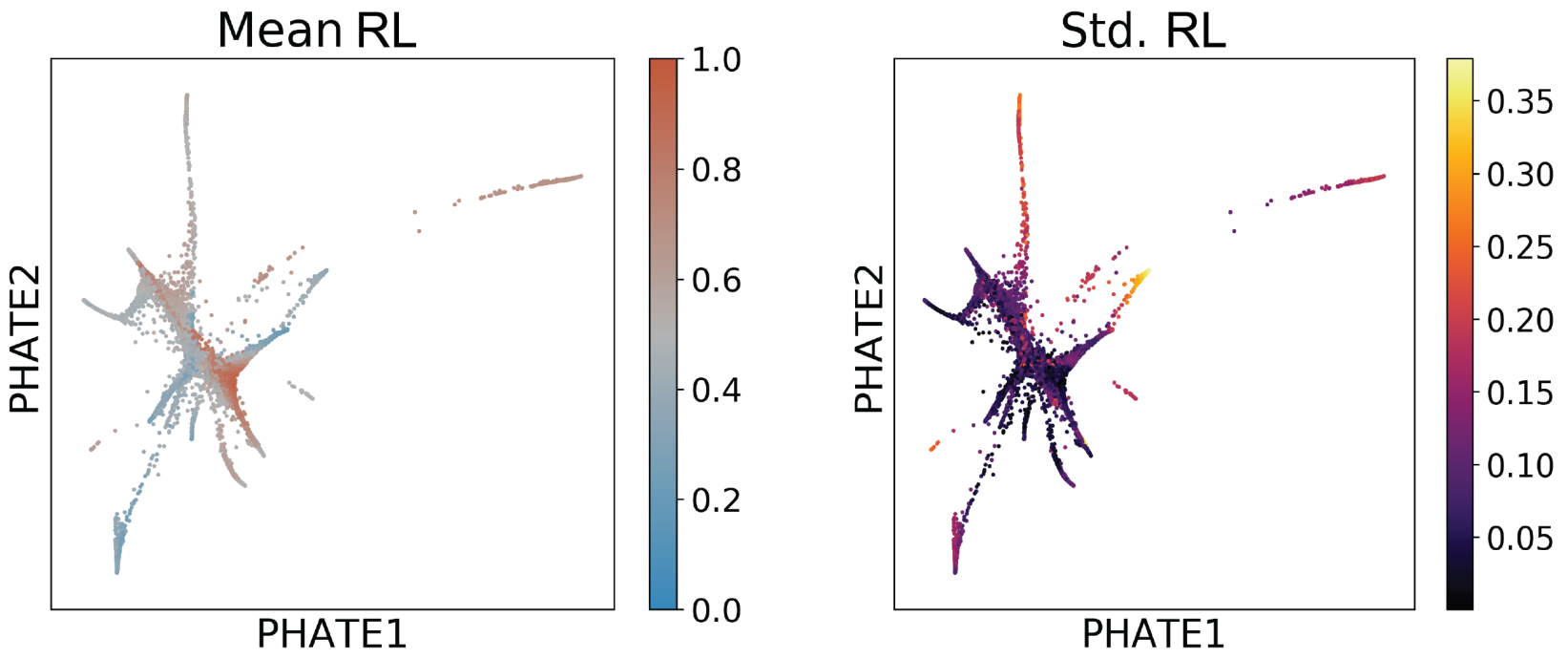
Analysis of replicates reveals heterogenous response

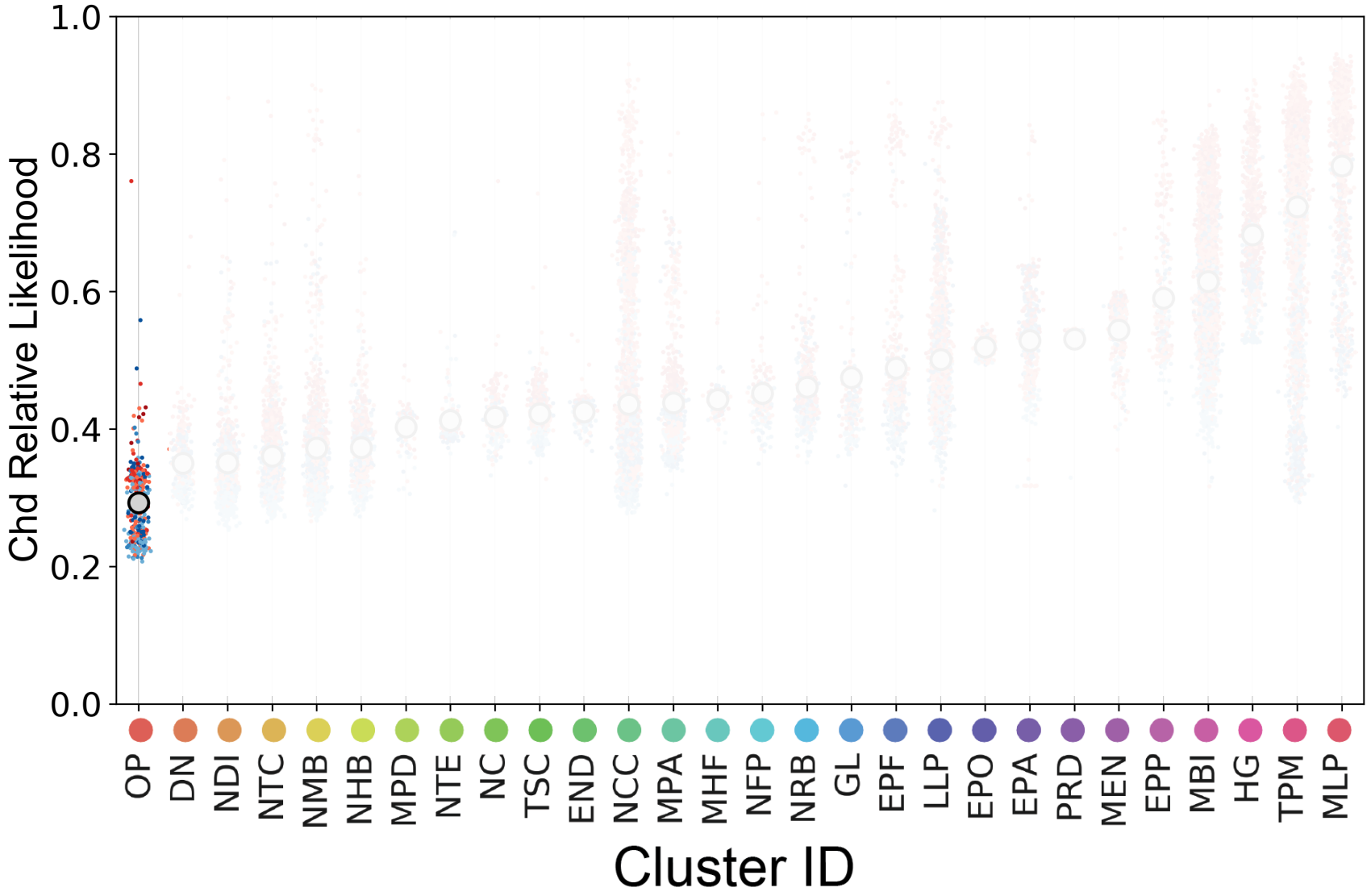


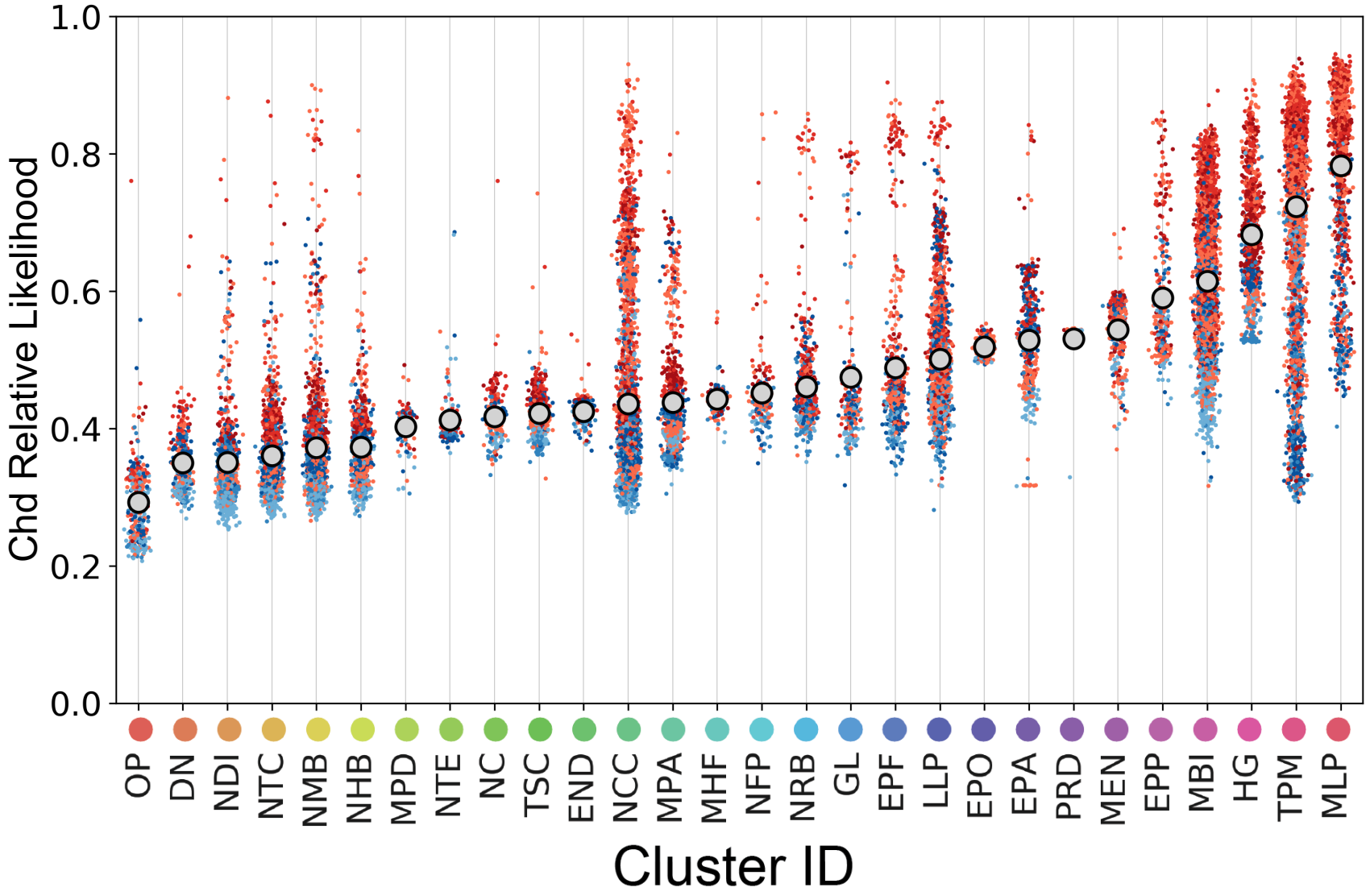
Analysis of replicates reveals heterogenous response

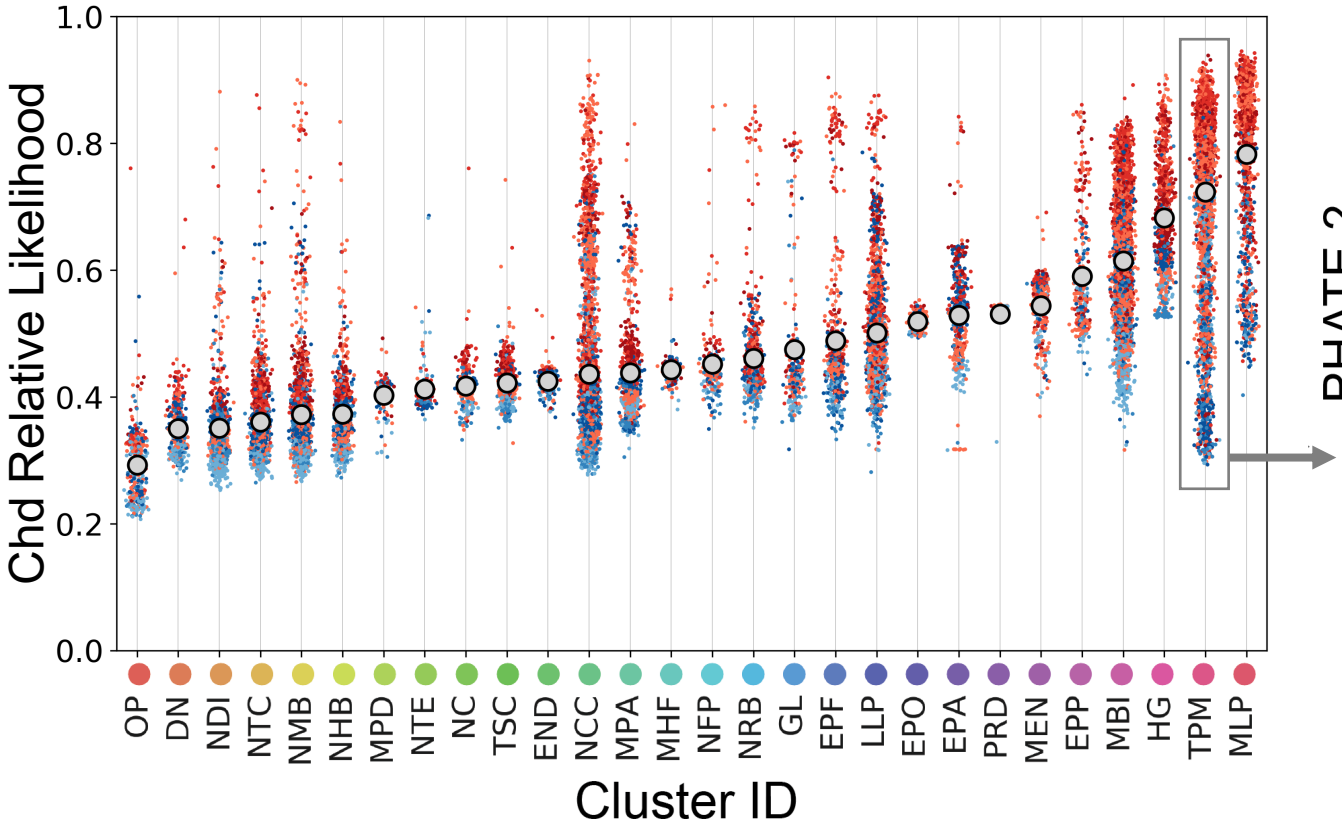


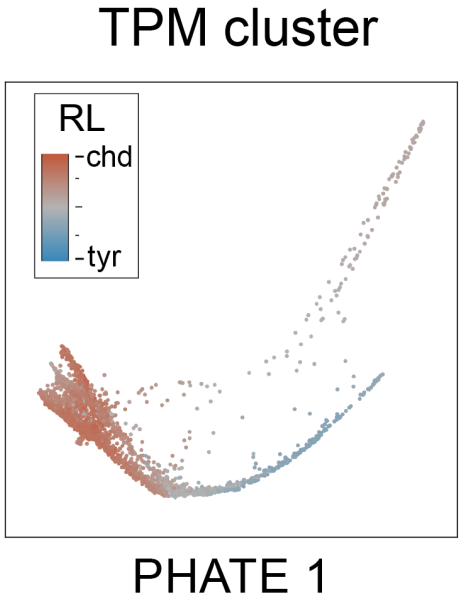
Analysis of replicates reveals heterogenous response

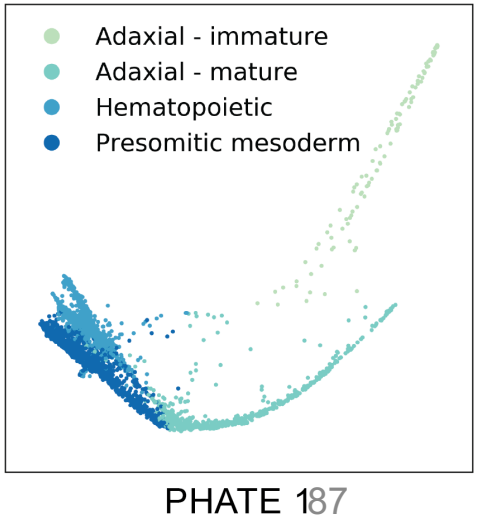


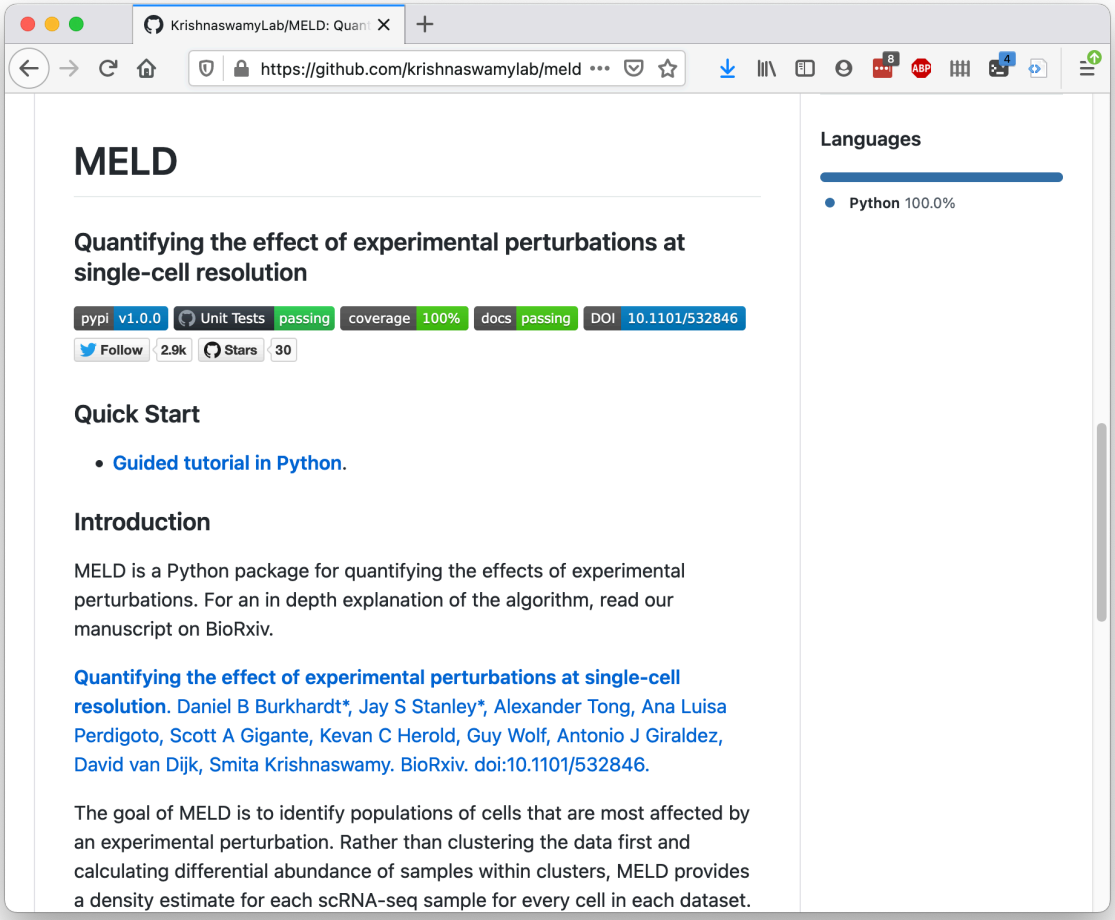

Chd Relative Likelihood
1.0
0.8
0.6
0.4
0.2
0.0
OP
DN
NDI
NTC
NMB
NHB
MPD
NTE
NC
TSC
END
NCC
MPA
MHF
NFP
NRB
GL
EPF
LLP
EPO
EPA
PRD
MEN
EPP
MBI
HG
TPM
MLP
Cluster ID


Chd Relative Likelihood
1.0
0.8
0.6
0.4
0.2
0.0
OP
DN
NDI
NTC
NMB
NHB
MPD
NTE
NC
TSC
END
NCC
MPA
MHF
NFP
NRB
GL
EPF
LLP
EPO
EPA
PRD
MEN
EPP
MBI
HG
TPM
MLP
Cluster ID


Chd Relative Likelihood
1.0
0.8
0.6
0.4
0.2
0.0
OP
DN
NDI
NTC
NMB
NHB
MPD
NTE
NC
TSC
END
NCC
MPA
MHF
NFP
NRB
GL
EPF
LLP
EPO
EPA
PRD
MEN
EPP
MBI
HG
TPM
MLP
Cluster ID


TPM cluster
RL
chd
tyr
PHATE 2
PHATE 1


Adaxial - immature
Adaxial - mature
Hematopoietic
Presomitic mesoderm
PHATE 187



bioRχiv bit.ly/quantsinglecell

github.com/KrishnaswamyLab/MELD

nature biotech *Coming soon!*