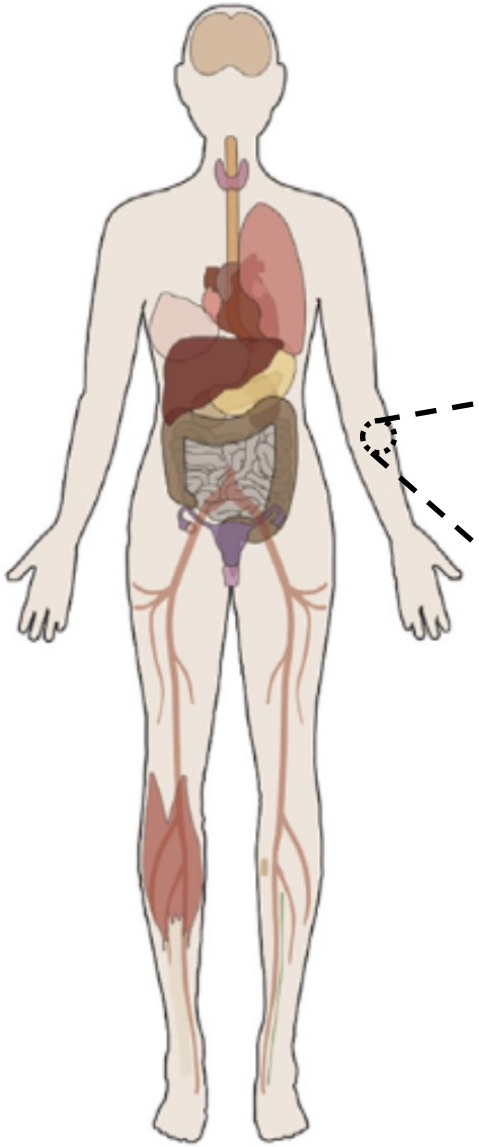
Wharton
University of Pennsylvania

Tilted-CCA: Quantifying common and distinct information in multi-modal single-cell data via matrix factorization

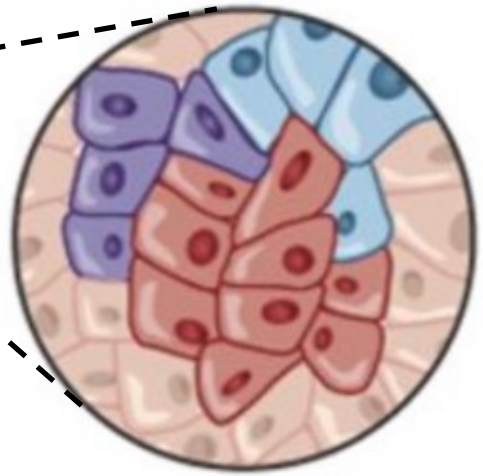
Kevin Lin
: linnykos

January 3, 2023

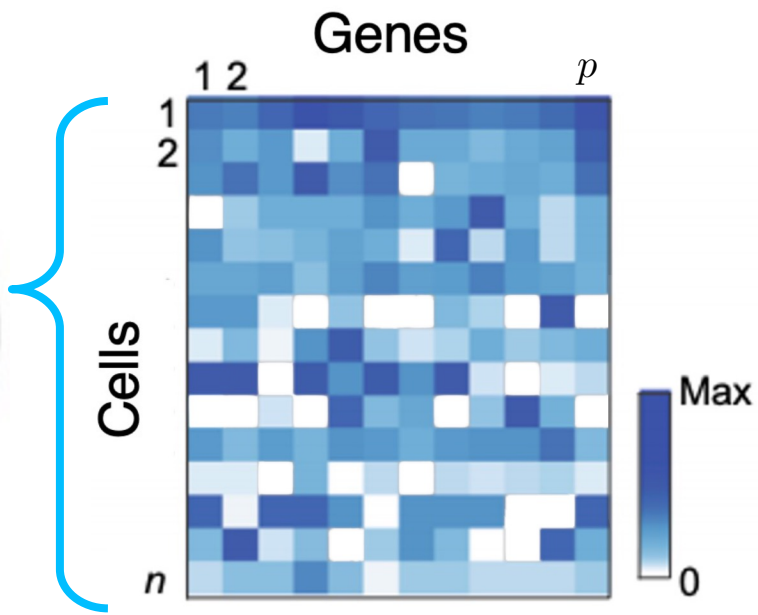
My body of work focuses on advancing statistical ideas that are inspired by advancements in biomedical technology.



Single-cell RNA-seq
(circa 2015)



RNA modality



Measuring “how active a gene is in a particular cell”

Previous statistical methods & theory I've developed for single-cell RNA-seq:

- Relevant for denoising single-cell data (Matrix factorization)
    - **K. Lin**, H. Liu, K. Roeder (JASA 2021)
    - **K. Lin**, J. Lei, K. Roeder (JASA 2022)

- Relevant for answering biological questions (Network methods and changepoint detection)
    - **K. Lin**, J. Sharpnack, A. Rinaldo, R.J. Tibshirani (Neurips 2017)
    - S. Hyun, **K. Lin**, M. G'Sell, R.J. Tibshirani (Biometrics 2021)
    - J. Lei, **K. Lin** (JASA 2022)

Previous statistical methods & theory I've developed for single-cell RNA-seq:

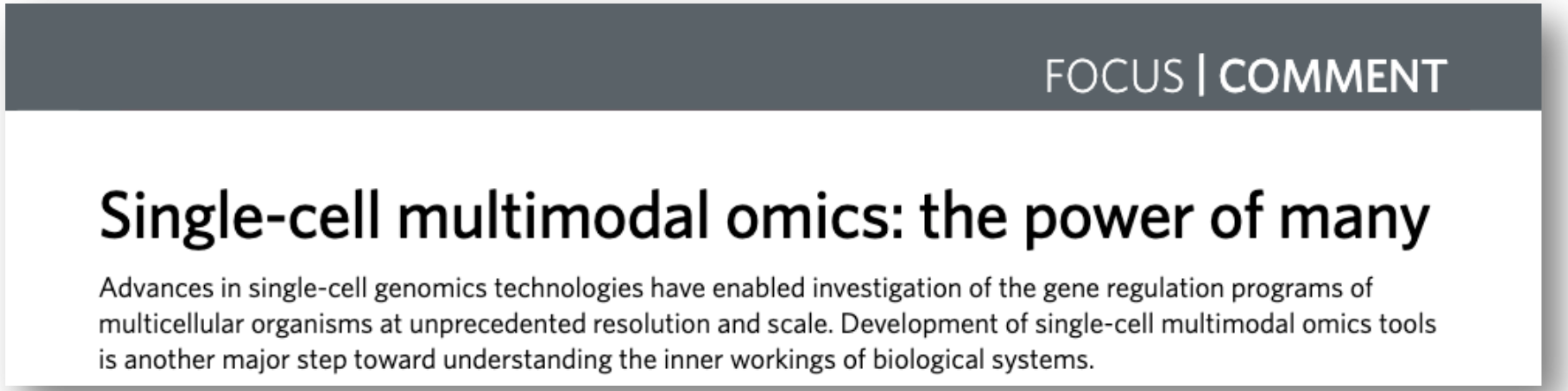
- Relevant for denoising single-cell data (Matrix factorization)
    - **K. Lin**, H. Liu, K. Roeder (JASA 2021)
    - **K. Lin**, J. Lei, K. Roeder (JASA 2022)

- Relevant for answering biological questions (Network methods and changepoint detection)
    - **K. Lin**, J. Sharpnack, A. Rinaldo, R.J. Tibshirani (Neurips 2017)
    - S. Hyun, **K. Lin**, M. G'Sell, R.J. Tibshirani (Biometrics 2021)
    - J. Lei, **K. Lin** (JASA 2022)

However, **our talk today** is inspired by newer biomedical technology.

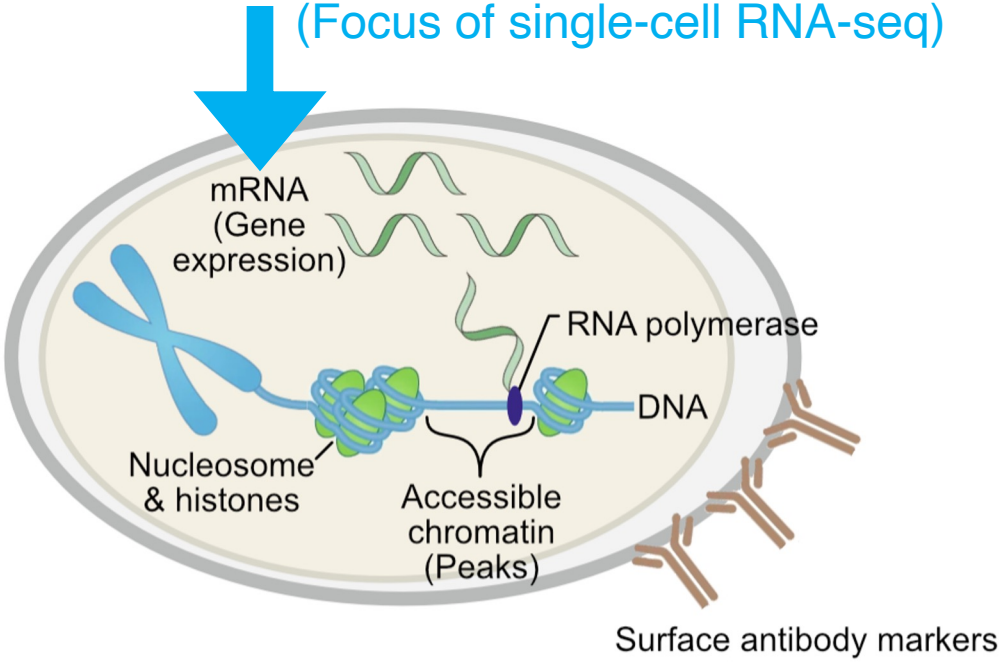
Our story today starts with a newer technology:

Teichmann, Efremova. Nature Methods (2020)

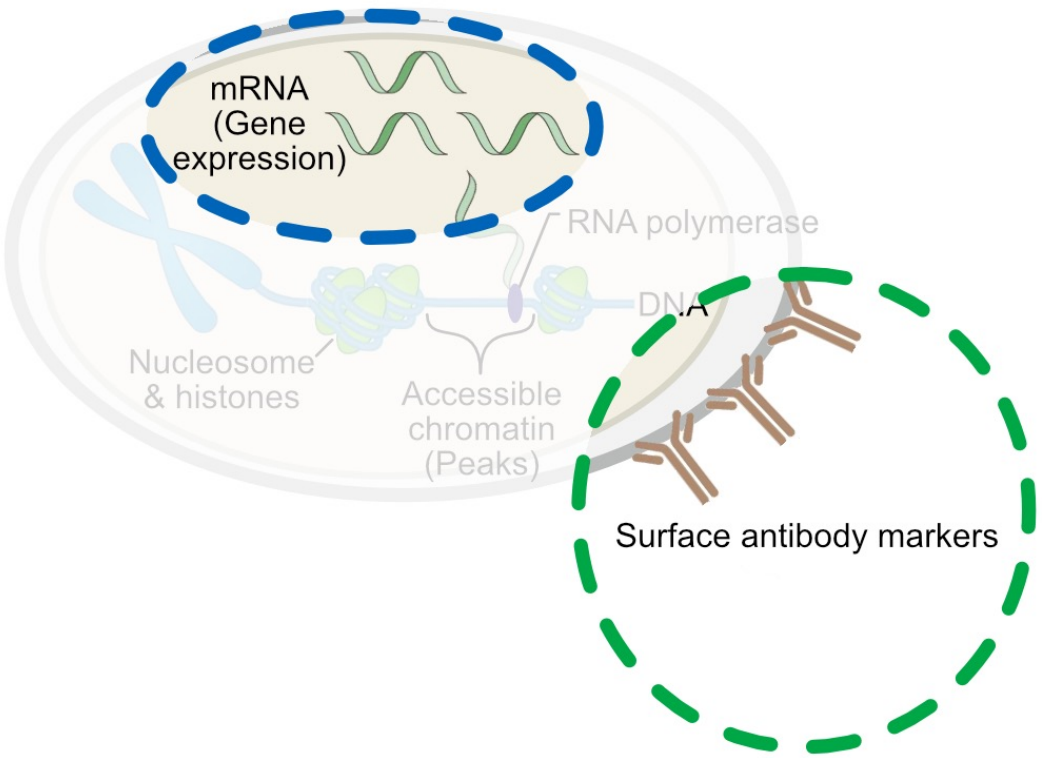


Zhu, Preissl, Ren. Nature Methods (2020)

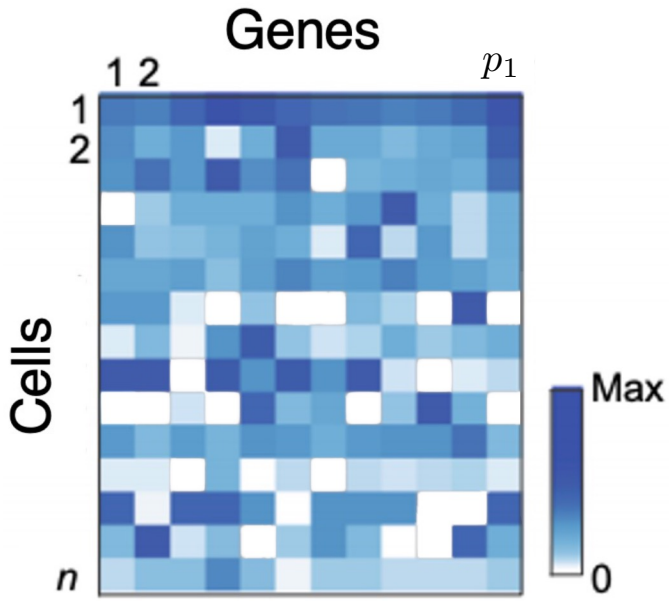
Our story today starts with a newer technology: Single-cell multi-modal (i.e., “multi-view”) sequencing (circa 2020).



Our story today starts with a newer technology: Single-cell multi-modal (i.e., "multi-view") sequencing (circa 2020).

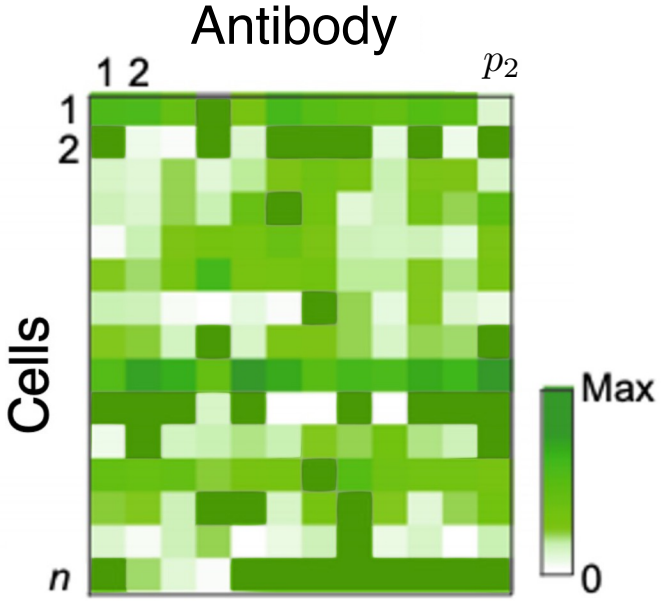


RNA modality



$$X^{(1)} \in \mathbb{R}^{n \times p_1}$$

Protein modality

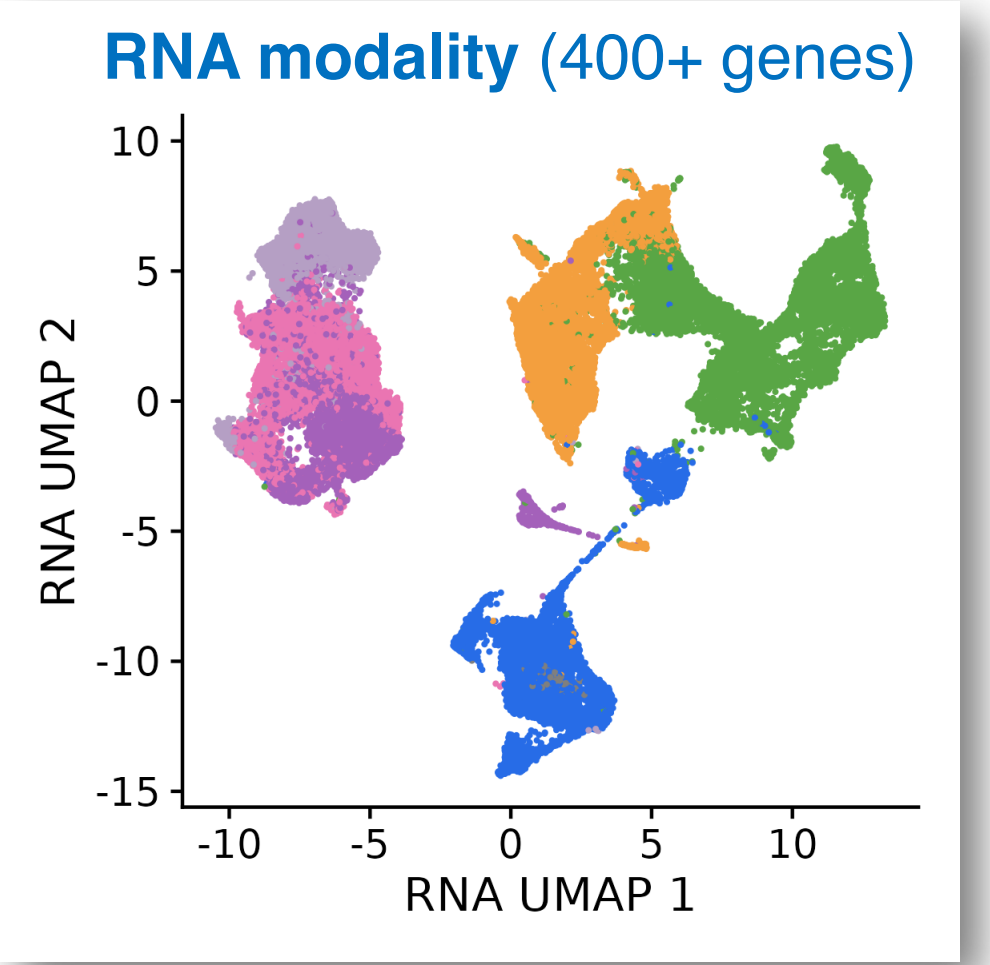


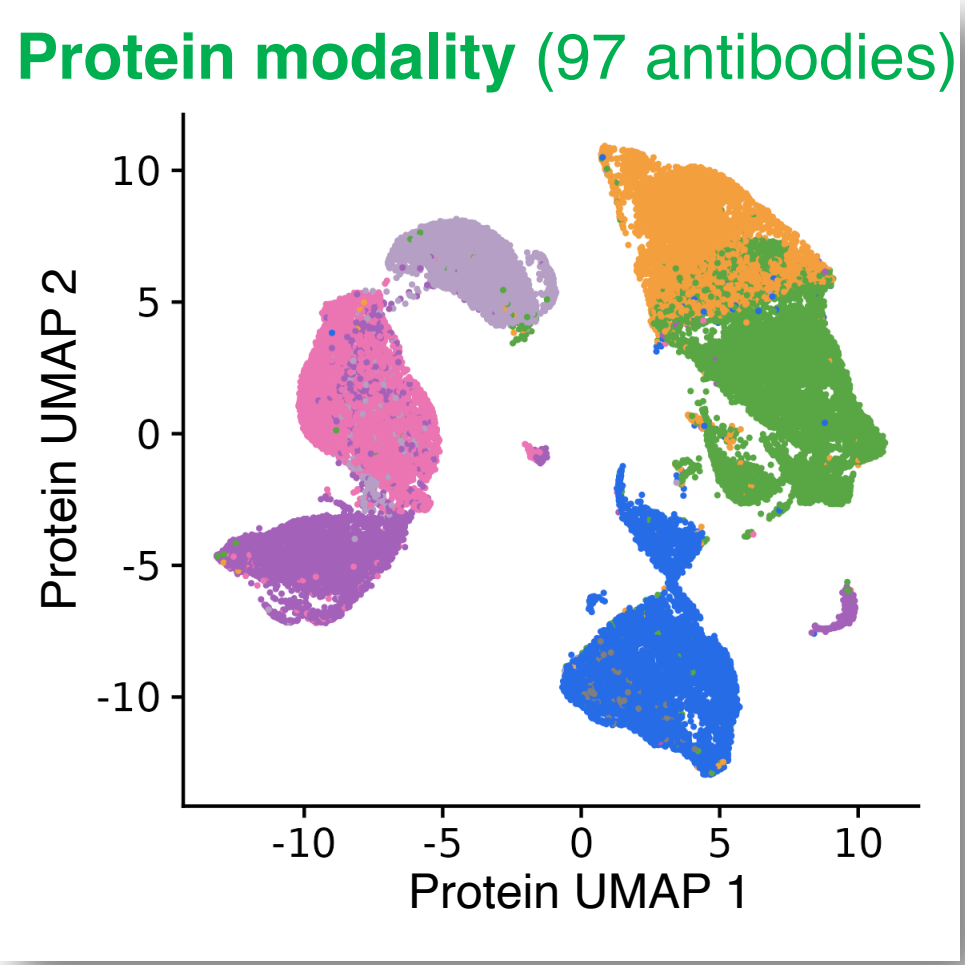
$$X^{(2)} \in \mathbb{R}^{n \times p_2}$$

Central question: What "information" is **unique** to a modality, or represents the **coordination** between both modalities?

What “information” is **unique** to a modality, or represents the **coordination** between both modalities?

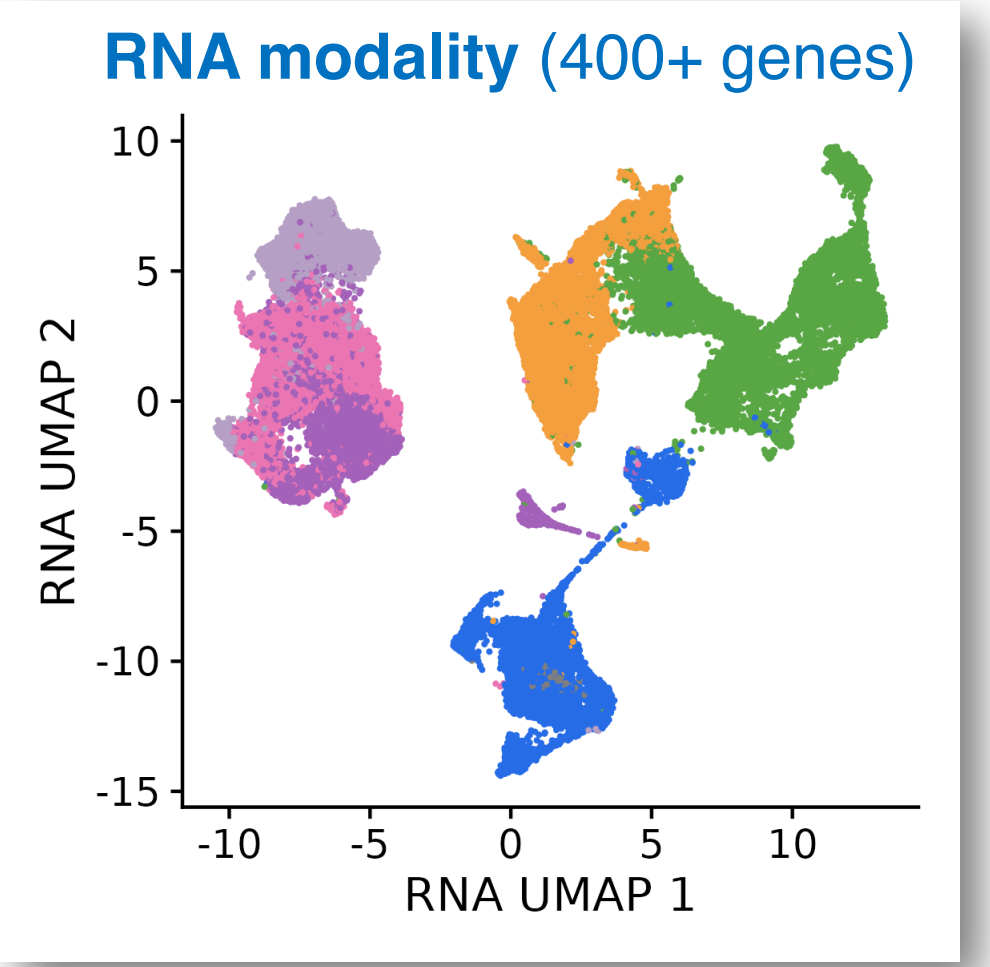
What “information” is **unique** to a modality, or represents the **coordination** between both modalities?

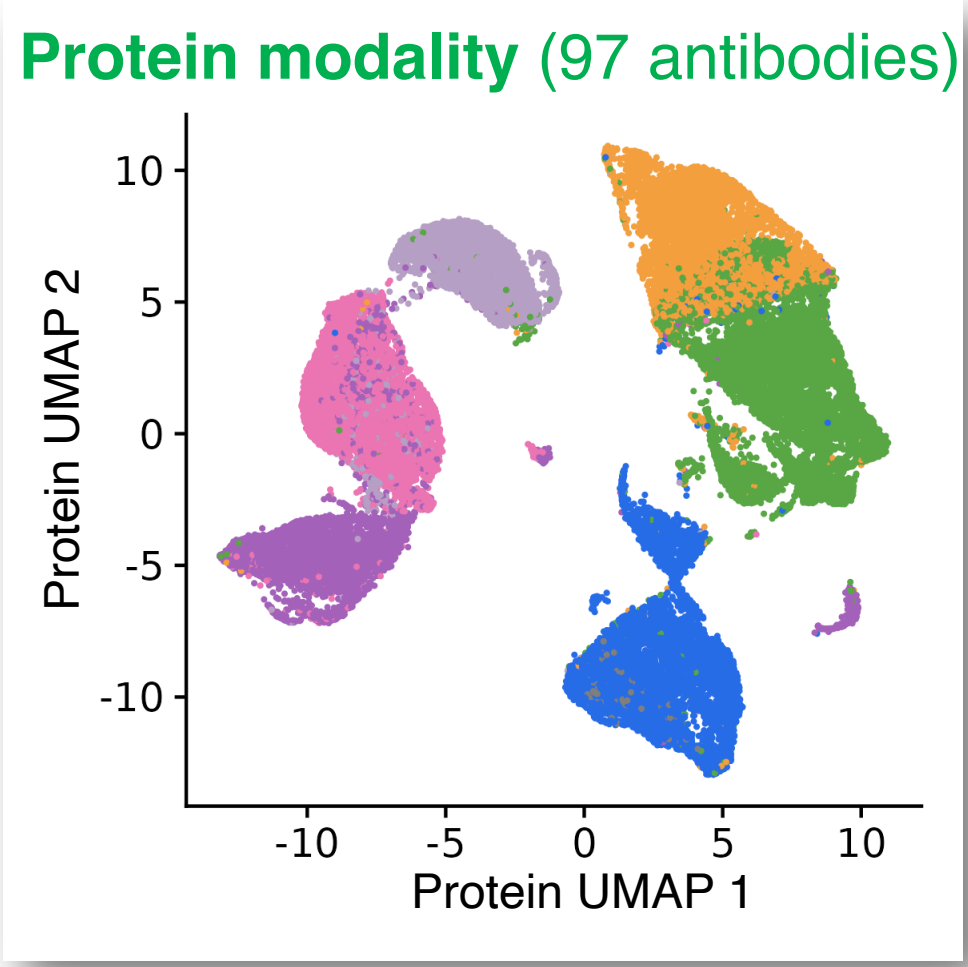




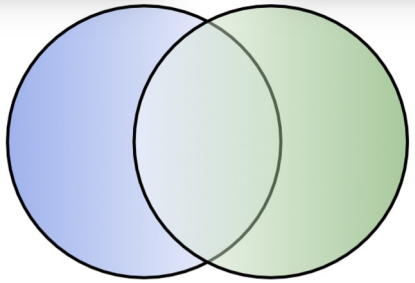
Human bone marrow (AbSeq, Triana et al., 2021), 49000+ cells (colored by annotated cell type)

What “information” is **unique** to a modality, or represents the **coordination** between both modalities?

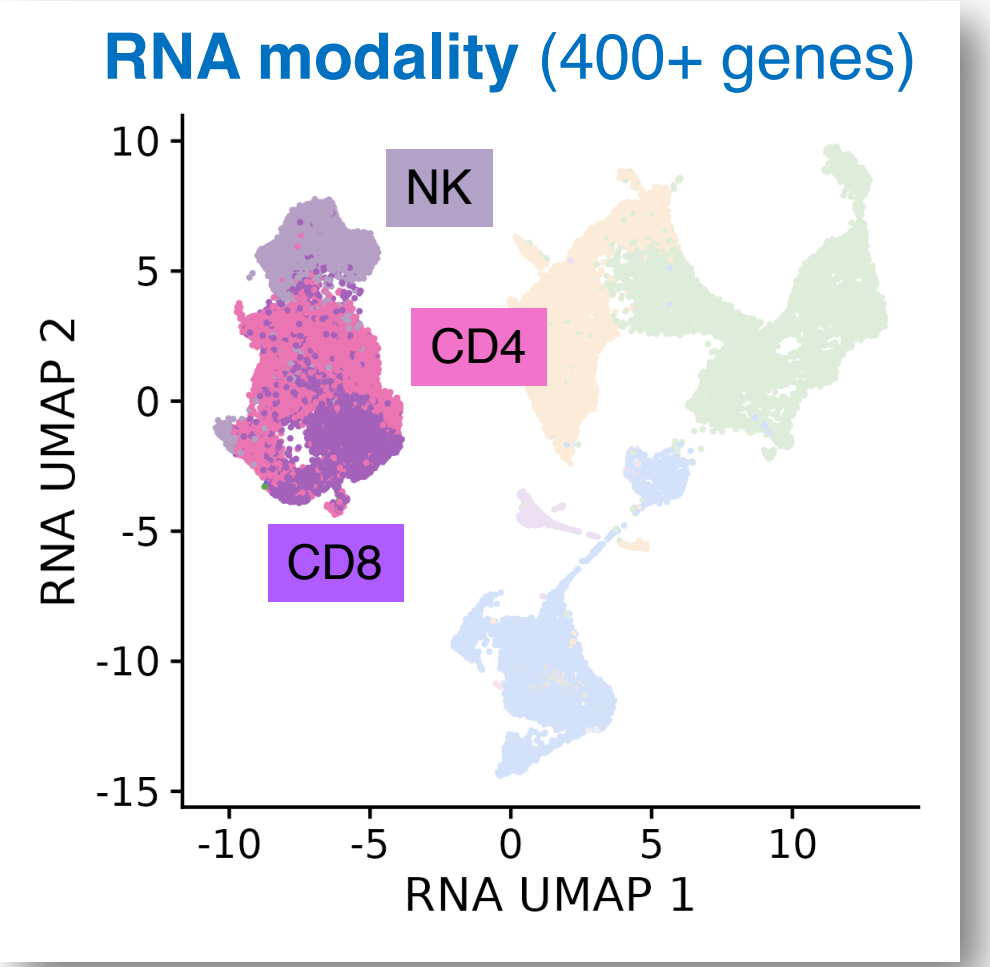


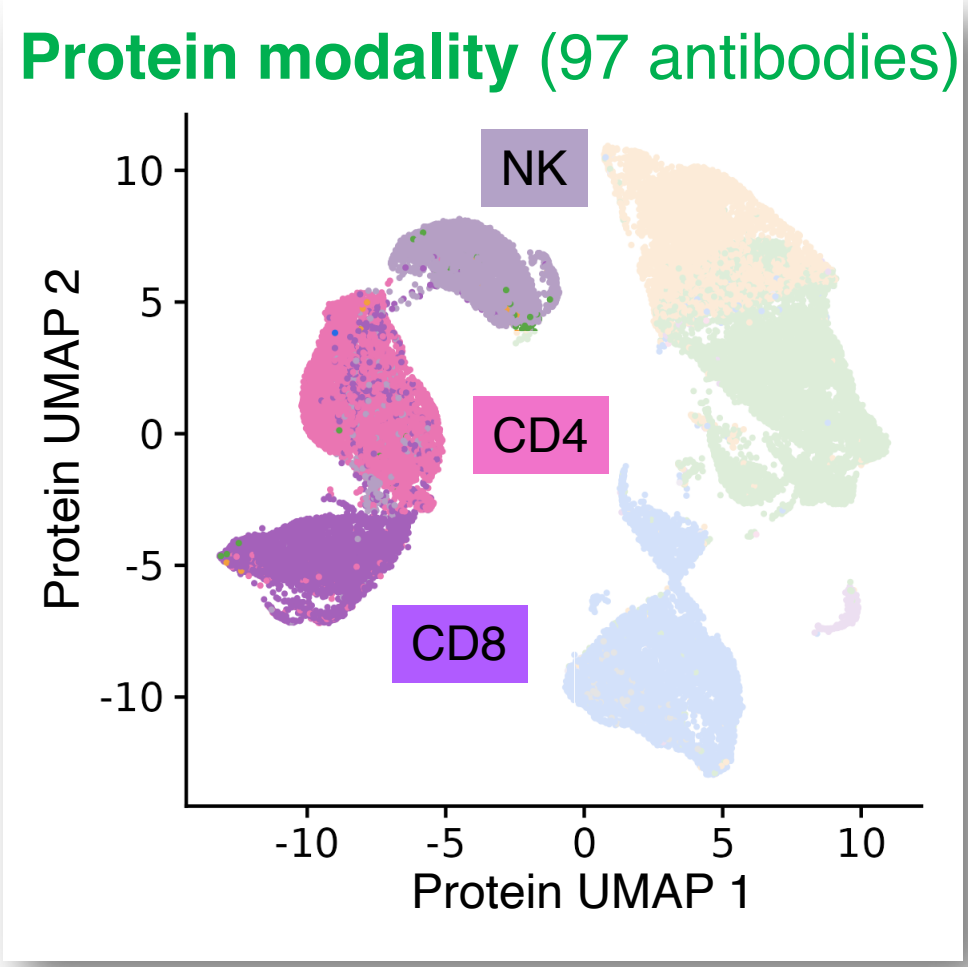


Venn diagram of geometry (“information”):

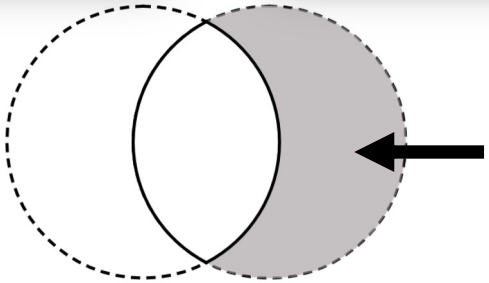


What “information” is **unique** to a modality, or represents the **coordination** between both modalities?



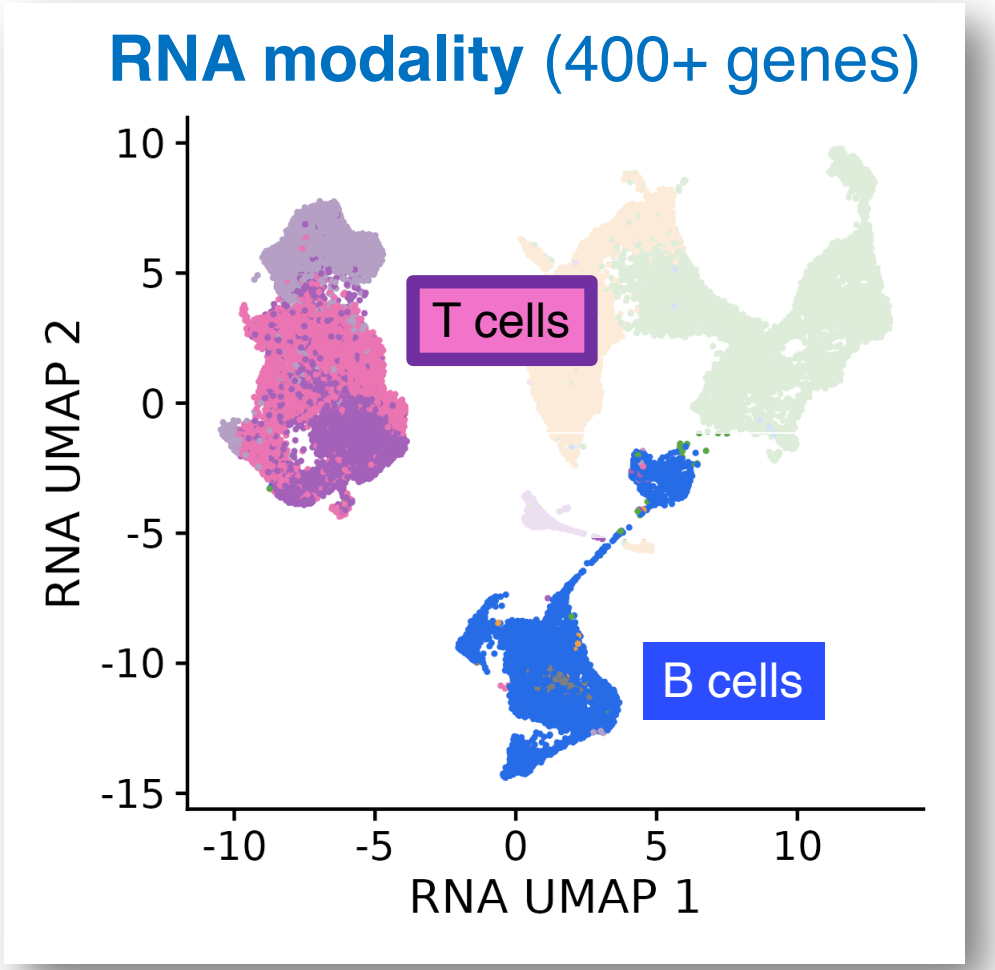


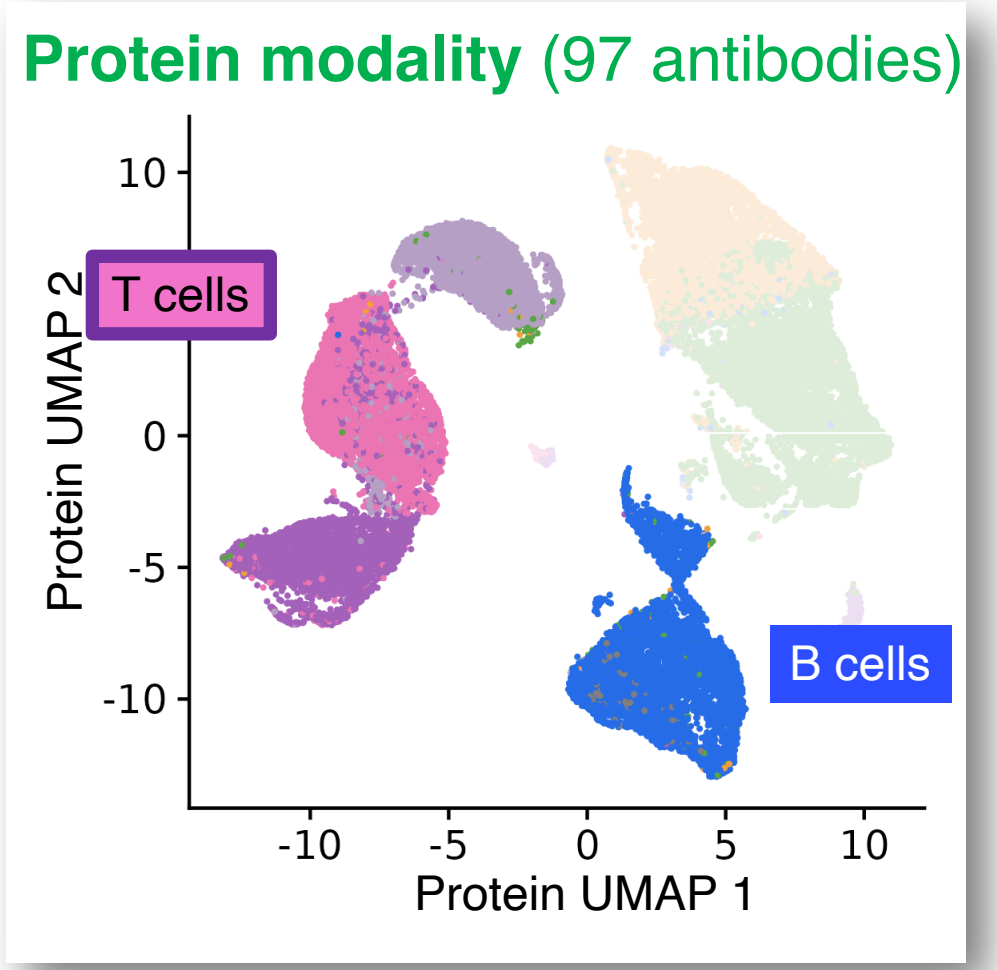
Venn diagram of geometry (“information”):



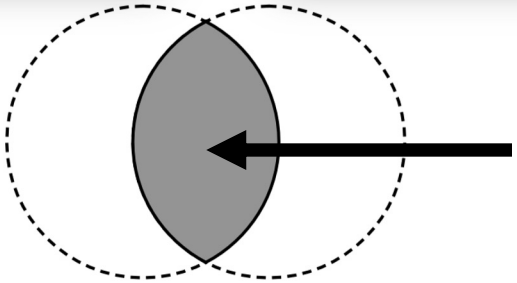
Cell-type separation unique to proteins

What “information” is **unique** to a modality, or represents the **coordination** between both modalities?



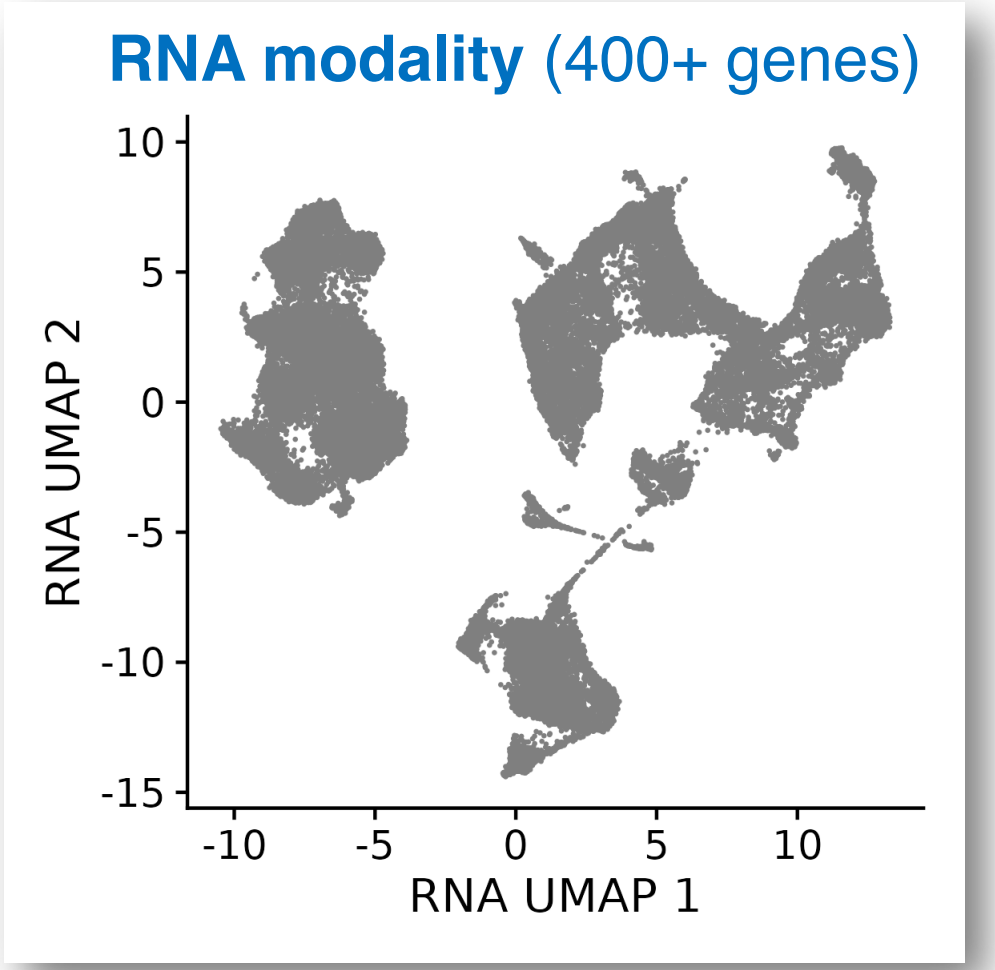


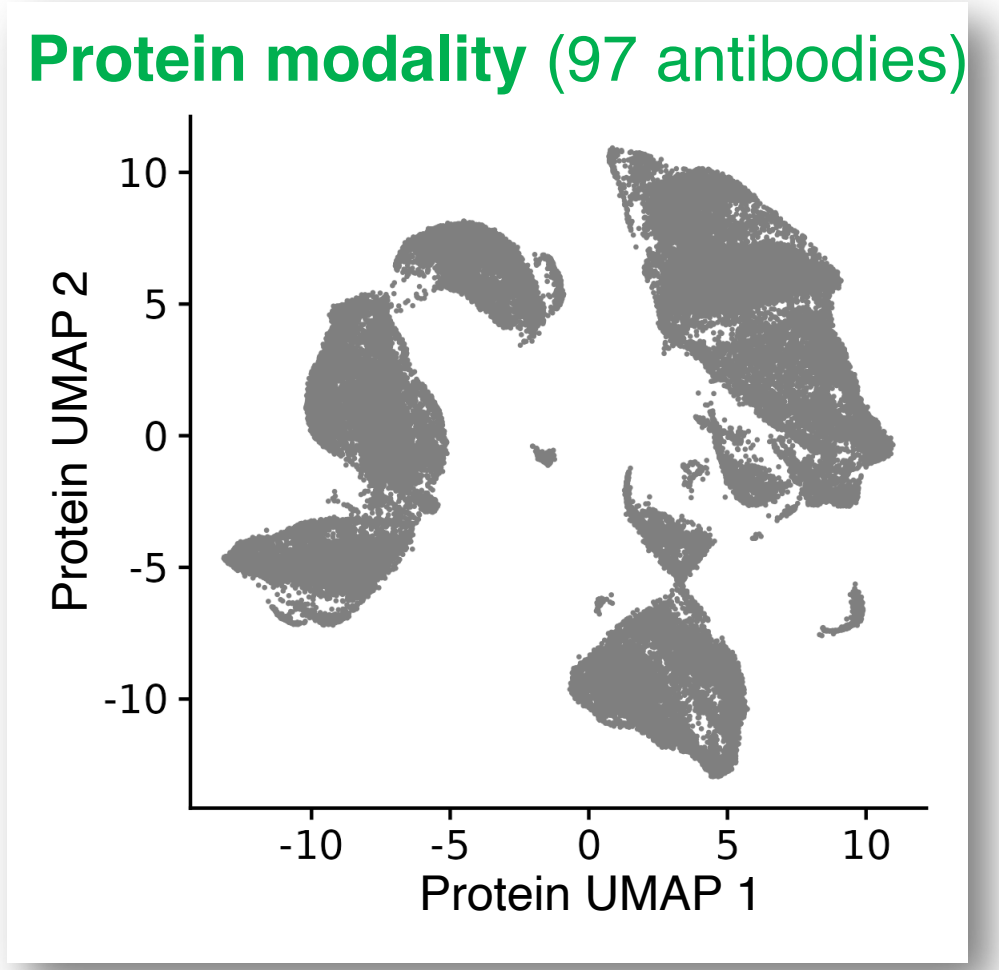
Venn diagram of geometry (“information”):



Between-modality coordination

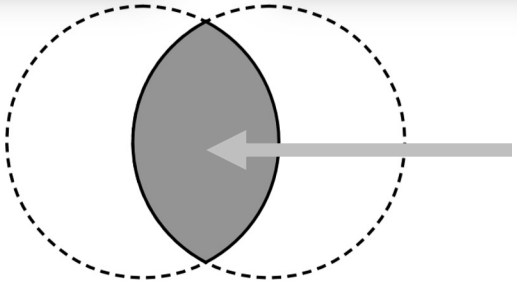
What “information” is **unique** to a modality, or represents the **coordination** between both modalities?





Venn diagram of geometry (“information”):

(No cell-type information)



Between-modality coordination

Statistical goal: Develop a new matrix factorization framework for multi-modal data based on shared/unique geometry to answer the following biological questions.

1. **(Experimental design):** Which pair of modalities should biologist sequence to have the most comprehensive understanding?

Statistical goal: Develop a new matrix factorization framework for multi-modal data based on shared/unique geometry to answer the following biological questions.

1. **(Experimental design):** Which pair of modalities should biologist sequence to have the most comprehensive understanding?

2. **(Variable selection):** For RNA-Protein data, how can we pick the antibodies that contribute the most additional information to the RNA modality?

Statistical goal: Develop a new matrix factorization framework for multi-modal data based on shared/unique geometry to answer the following biological questions.

1. **(Experimental design):** Which pair of modalities should biologist sequence to have the most comprehensive understanding?

2. **(Variable selection):** For RNA-Protein data, how can we pick the antibodies that contribute the most additional information to the RNA modality?

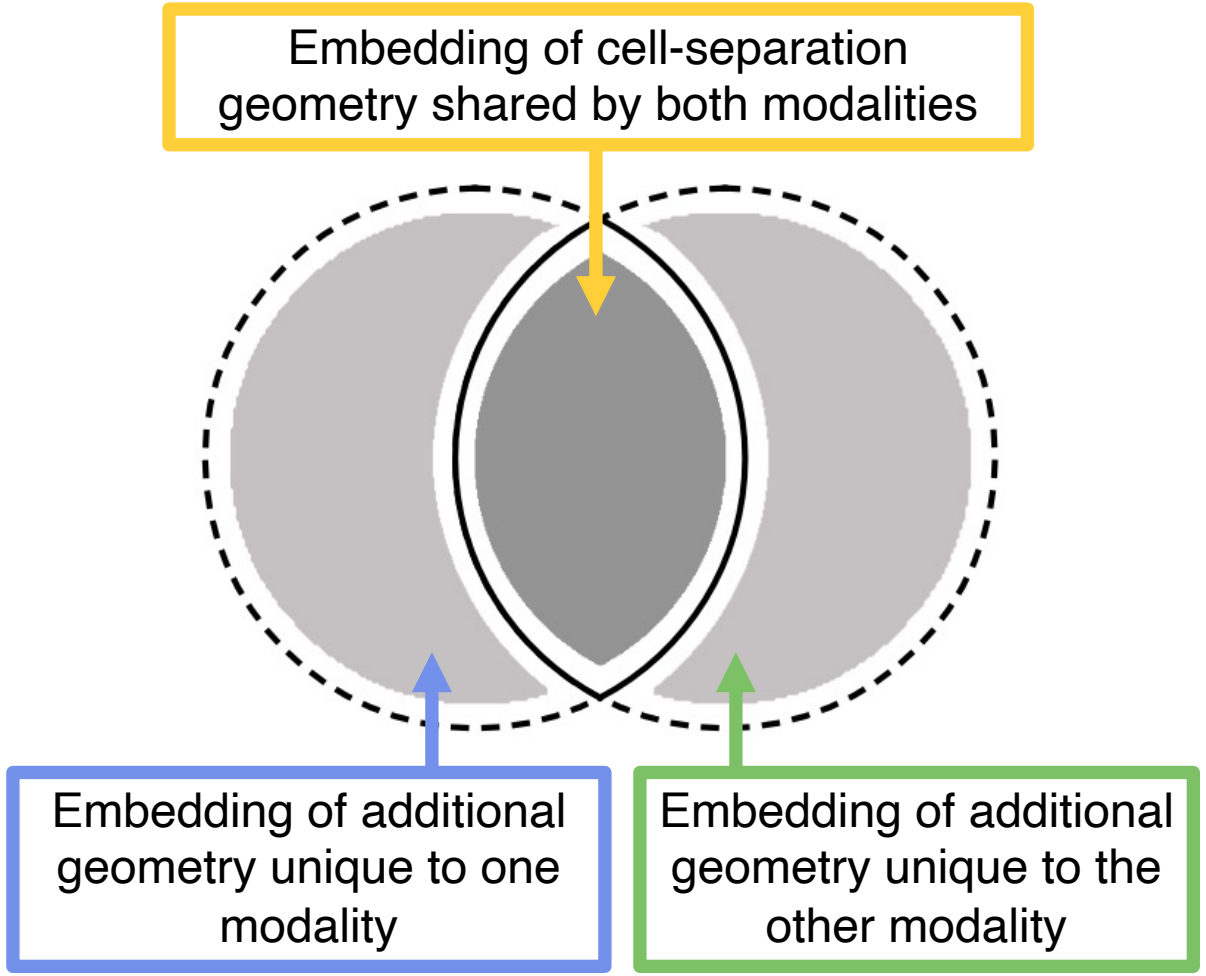
3. **(Developmental biology):** Can the amount of coordination between two modalities tell us if a cell in a steady-state or is undergoing development?

Statistical goal: Develop a new matrix factorization framework for multi-modal data based on shared/unique geometry to answer the following biological questions.

1. **(Experimental design):** Which pair of modalities should biologist sequence to have the most comprehensive understanding?

2. **(Variable selection):** For RNA-Protein data, how can we pick the antibodies that contribute the most additional information to the RNA modality?

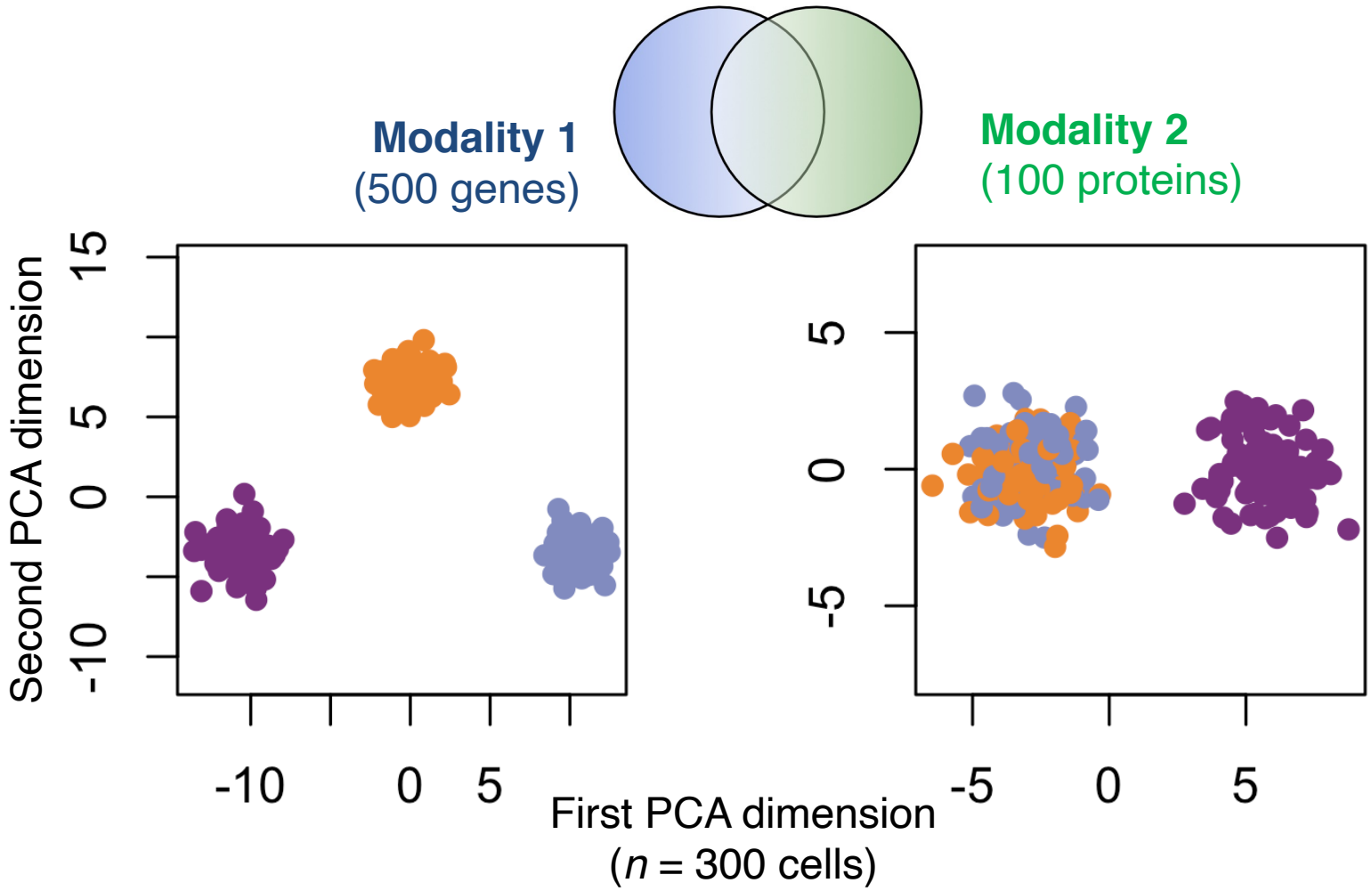
3. **(Developmental biology):** Can the amount of coordination between two modalities tell us if a cell in a steady-state or is undergoing development?



Our method inspires new theoretical questions and is applicable to <u>any</u> multi-modal dataset.

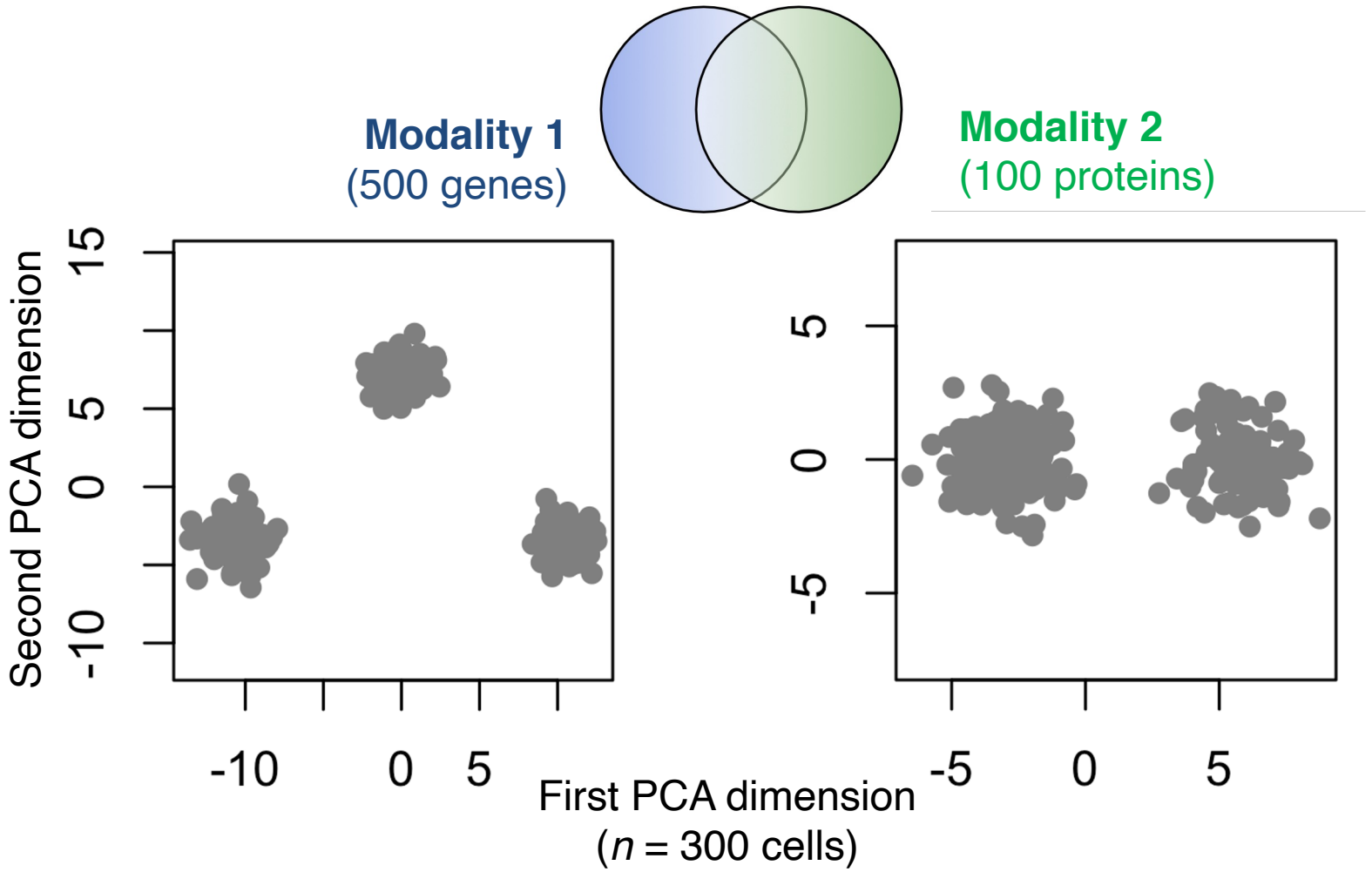
To start thinking about how to define geometry shared/unique to each modality, consider a toy example where one modality has more “information” than the other.

- Geometry = “Information” ≈ “Density clustering”



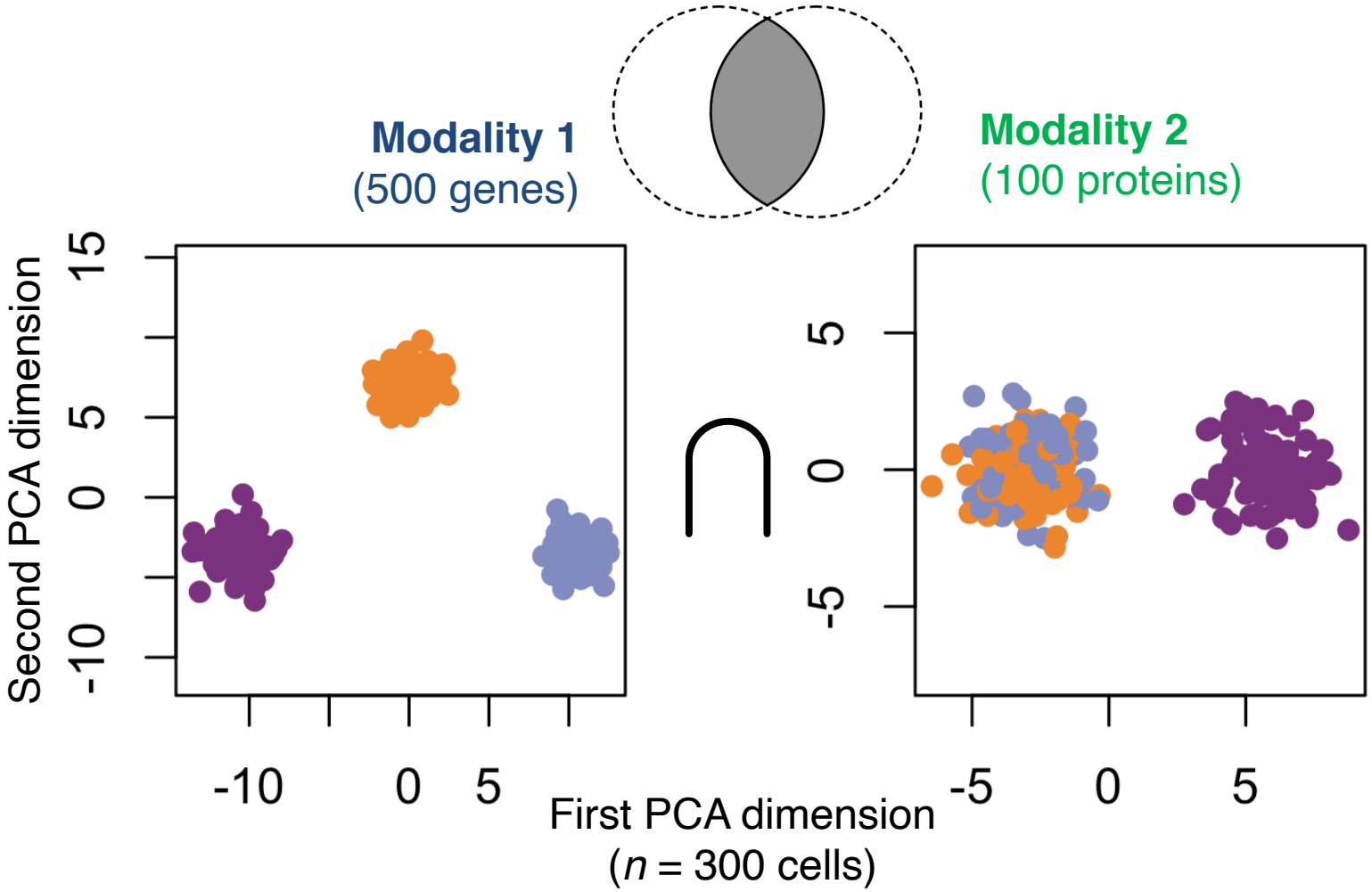
To start thinking about how to define geometry shared/unique to each modality, consider a toy example where one modality has more “information” than the other.

- Geometry = “Information” ≈ “Density clustering”



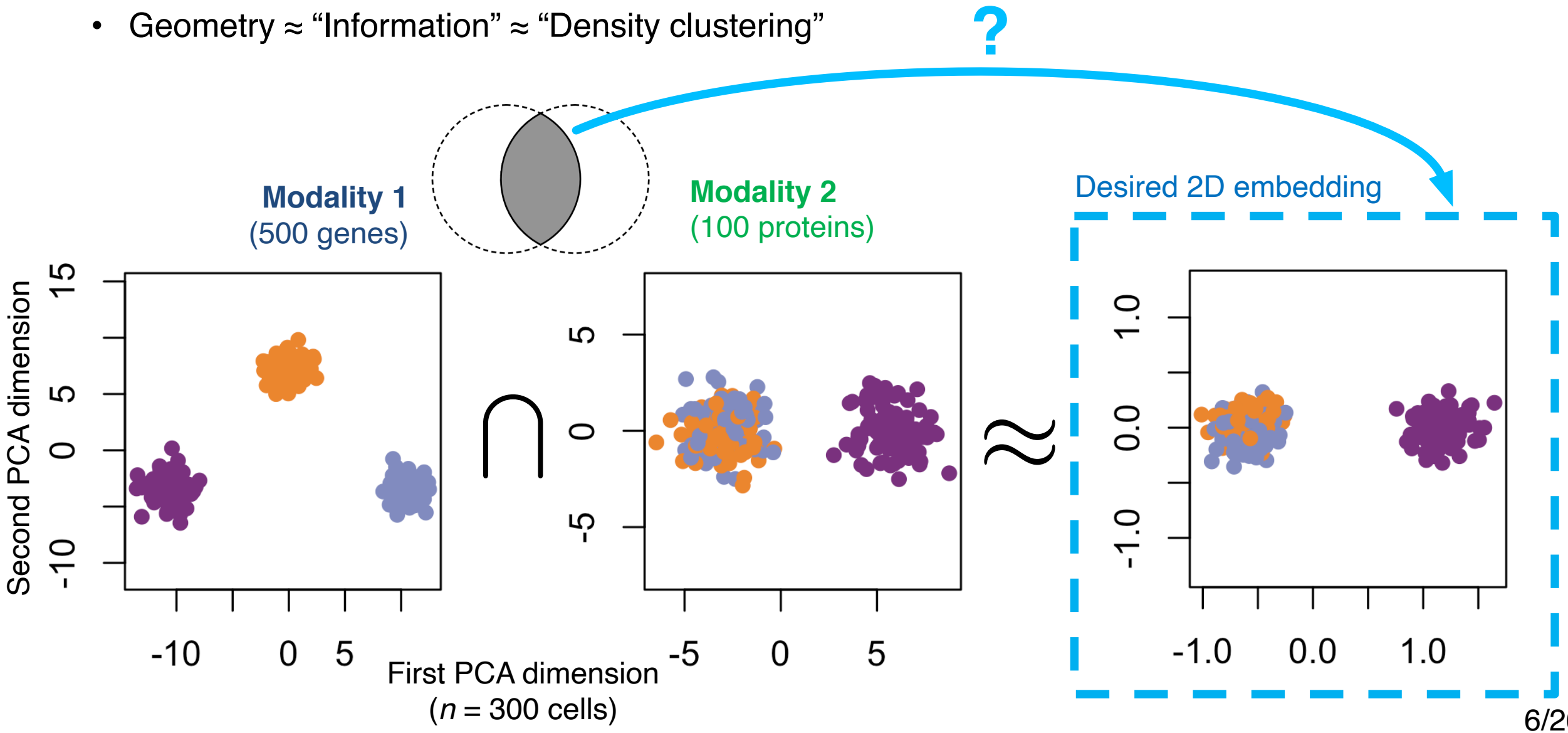
To start thinking about how to define geometry shared/unique to each modality, consider a toy example where one modality has more “information” than the other.

- Geometry ≈ “Information” ≈ “Density clustering”



To start thinking about how to define geometry shared/unique to each modality, consider a toy example where one modality has more "information" than the other.

- Geometry ≈ "Information" ≈ "Density clustering"



Many existing multi-modal dimension reduction methods focus minimizing reconstruction error, but these encapsulate the “union” of information.

Many existing multi-modal dimension reduction methods focus minimizing reconstruction error, but these encapsulate the "union" of information.

(a phrase we coined for multimodal matrix factorization)

Many existing multi-modal dimension reduction methods focus minimizing reconstruction error, but these encapsulate the “union” of information.

Consider Consensus PCA (Wold et al., 1987):

$$\widehat{L} = \underset{L \in \mathbb{R}^{(p_1+p_2)\times r}}{\arg\min} \left\| \left[X^{(1)} \; ; \; X^{(2)}\right] - \left[X^{(1)} \; ; \; X^{(2)}\right] LL^\top \right\|_F^2$$

Resulting embedding via SVD:

$$\left[X^{(1)} \; ; \; X^{(2)}\right] = UDV^\top$$

Many existing multi-modal dimension reduction methods focus minimizing reconstruction error, but these encapsulate the “union” of information.

Consider Consensus PCA (Wold et al., 1987):

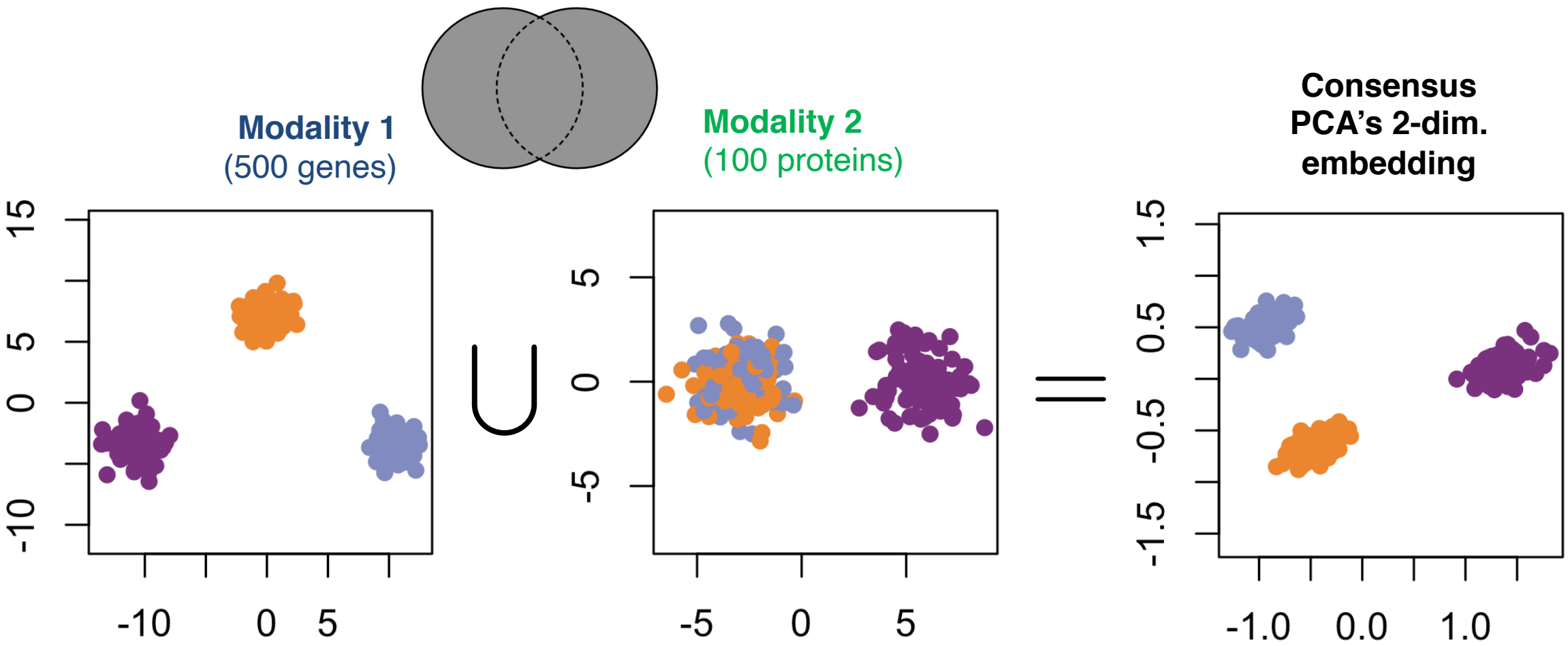
$$\widehat{L} = \underset{L \in \mathbb{R}^{(p_1+p_2)\times r}}{\arg\min} \left\| \left[X^{(1)} \; ; \; X^{(2)}\right] - \left[X^{(1)} \; ; \; X^{(2)}\right] L L^\top \right\|_F^2$$

Resulting embedding via SVD:

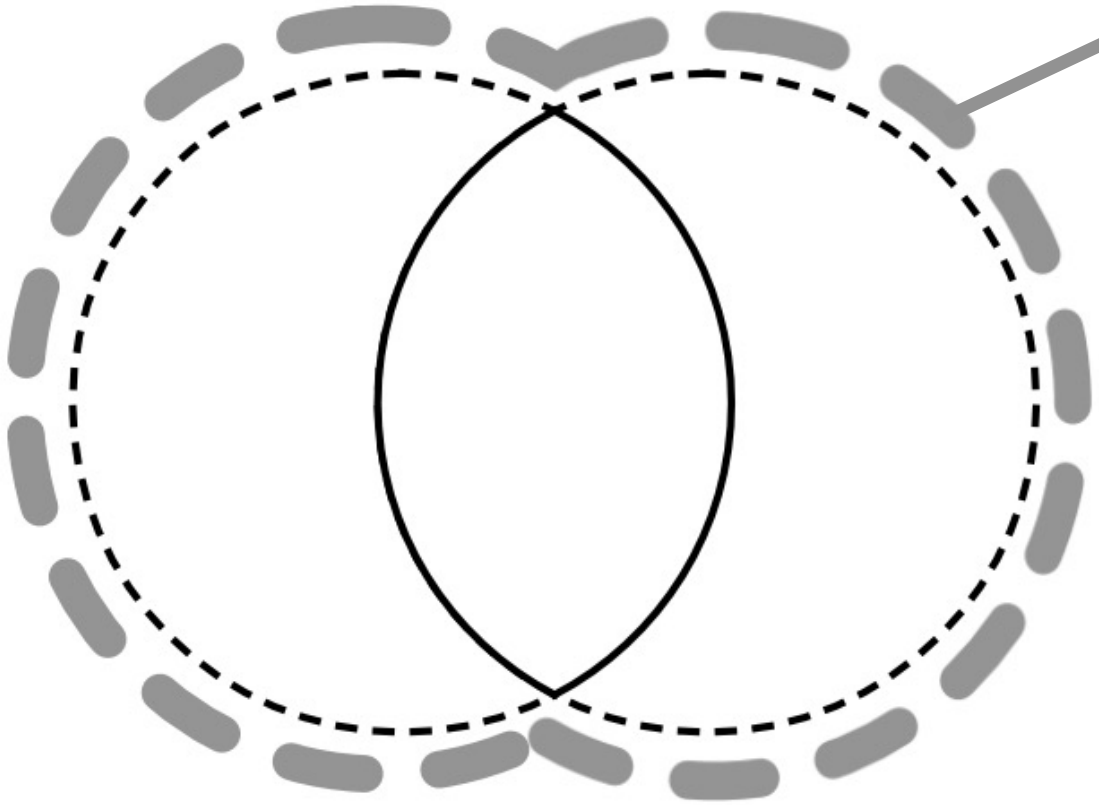
$$\left[X^{(1)} \; ; \; X^{(2)}\right] = \underbrace{UD}_{\text{Low-dimensional embedding}} V^\top$$

Observation: The embedding combines the axes of variation from each modality (i.e. the “union”).

Let’s see the “union of information” in toy example:



Our contribution: A matrix factorization for multi-modal data that separates of "intersection" and "unique" of information, as opposed many existing methods that represent the "union" of information.

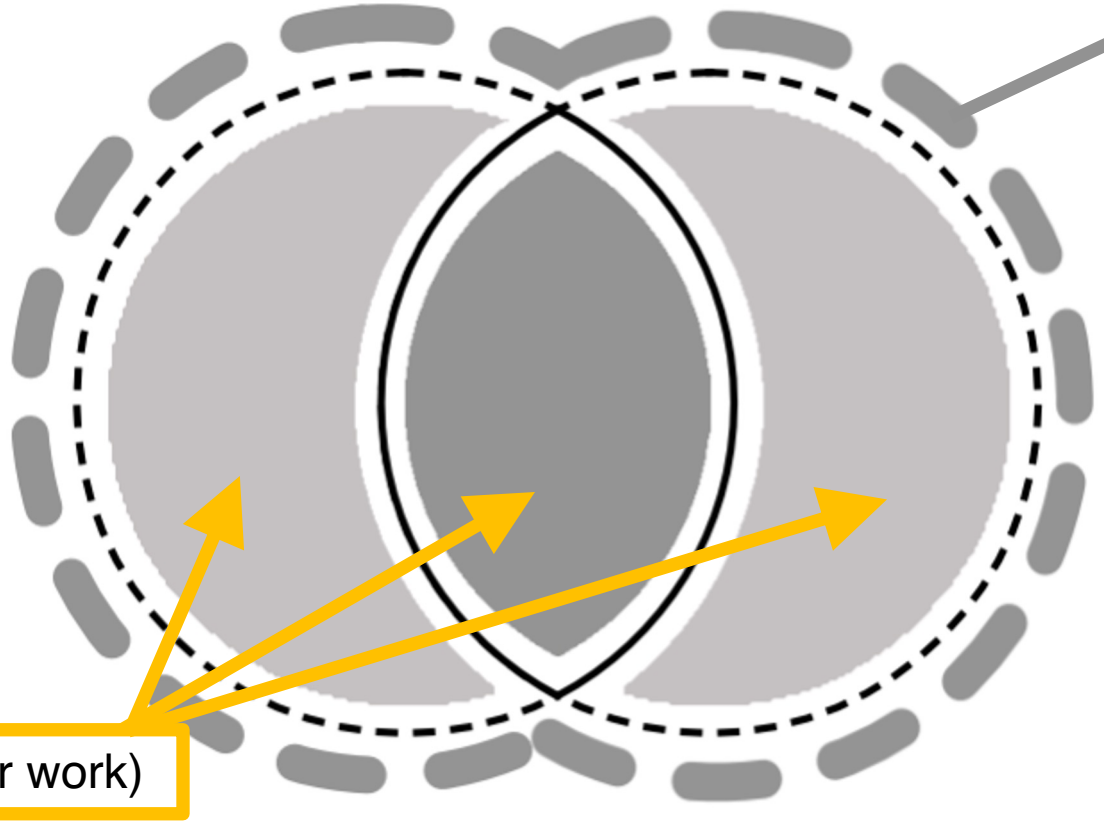


- Consensus PCA (Word et al., 1987)
- scAI (Nie et al., 2020)
- MOFA+ (Stegle et al., 2020)
- JSNMF (Ma et al., 2022)

- JIVE (Nobel et al., 2013)
- WNN (Satija et al., 2021)
- Cobalt (Purdom et al., 2021)

Cell embeddings for the "union" geometry: Useful for making an "atlas" across both modalities

Our contribution: A matrix factorization for multi-modal data that separates of "intersection" and "unique" of information, as opposed many existing methods that represent the "union" of information.



- Consensus PCA (Word et al., 1987)
- scAI (Nie et al., 2020)
- MOFA+ (Stegle et al., 2020)
- JSNMF (Ma et al., 2022)

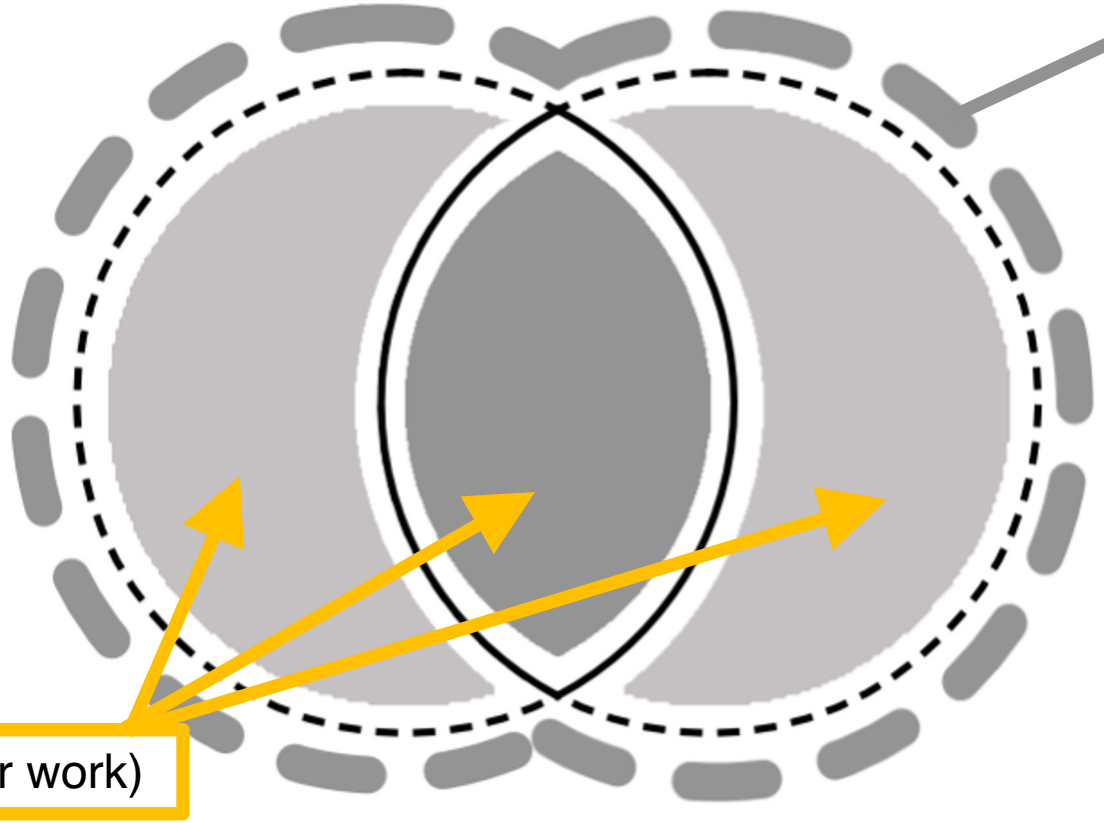
- JIVE (Nobel et al., 2013)
- WNN (Satija et al., 2021)
- Cobalt (Purdom et al., 2021)

Cell embeddings for the "union" geometry: Useful for making an "atlas" across both modalities

Tilted-CCA (Our work)

Cell embedding for the "intersection"/"unique" geometry: Useful to understand the coordination between modalities

Our contribution: A matrix factorization for multi-modal data that separates of "intersection" and "unique" of information, as opposed many existing methods that represent the "union" of information.



- Consensus PCA (Word et al., 1987)
- scAI (Nie et al., 2020)
- MOFA+ (Stegle et al., 2020)
- JSNMF (Ma et al., 2022)

- JIVE (Nobel et al., 2013)*
- WNN (Satija et al., 2021)*
- Cobalt (Purdom et al., 2021)*

Cell embeddings for the "union" geometry: Useful for making an "atlas" across both modalities

Tilted-CCA (Our work)

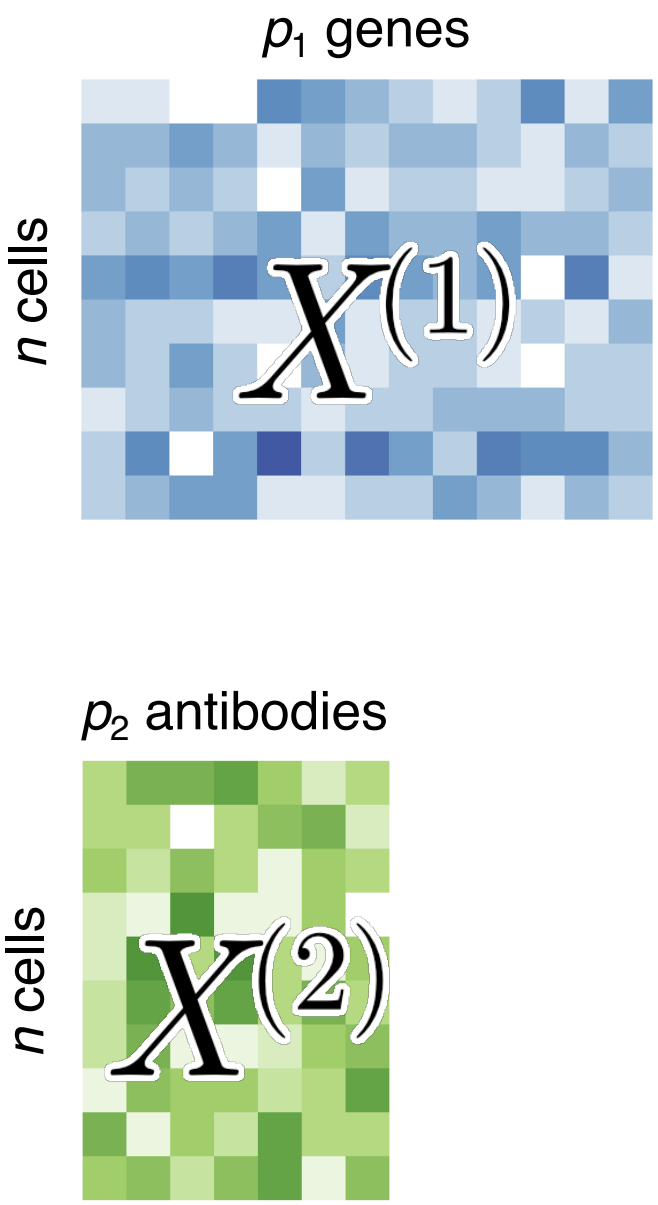
Cell embedding for the "intersection"/"unique" geometry: Useful to understand the coordination between modalities

* (Methods that don't cleanly fall in this dichotomy)

Statistical method:

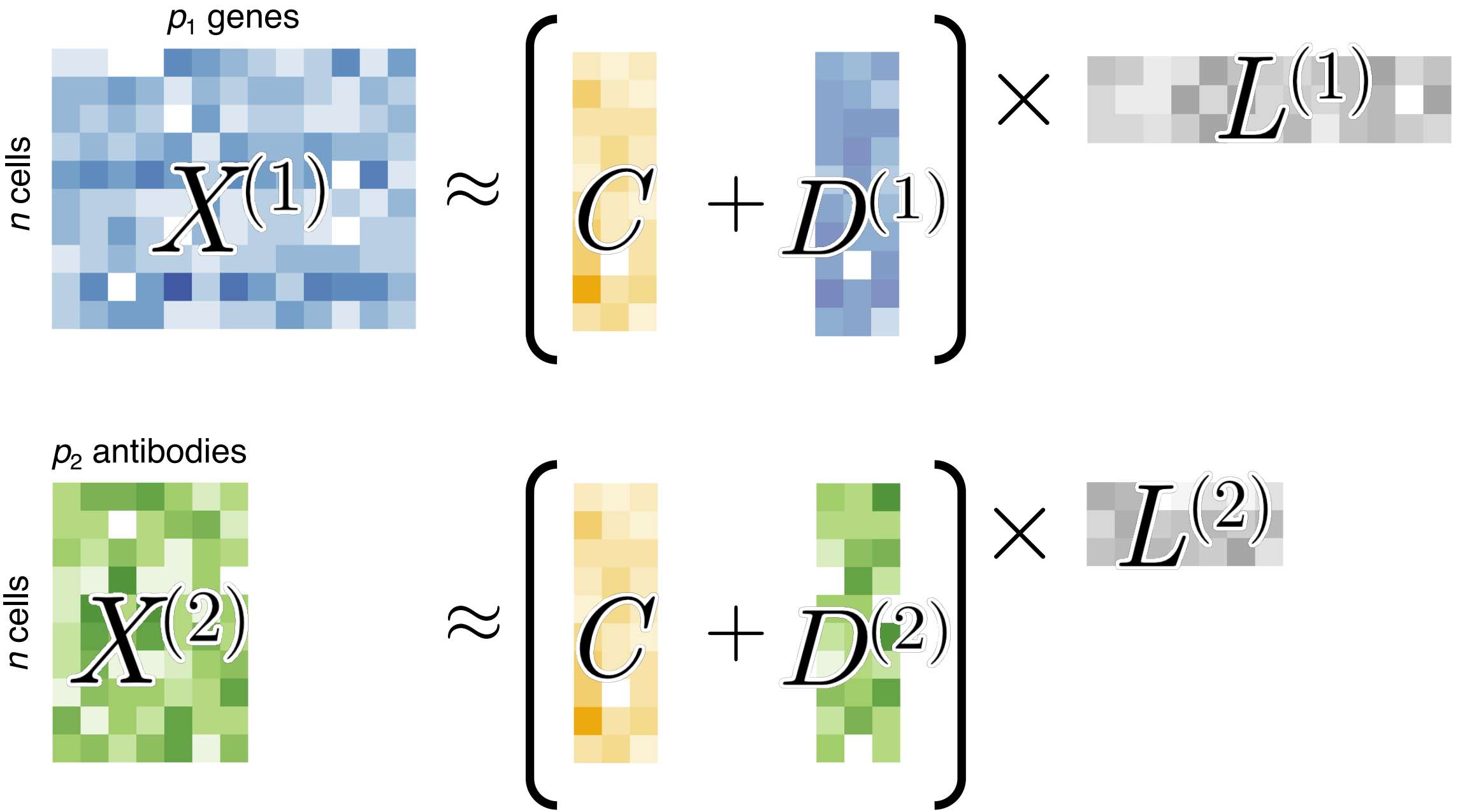
New dimension-reduction framework for multi-modal data,
not about denoising but instead the geometry

Proposed model for multi-modal data: Capturing the “shared” geometry

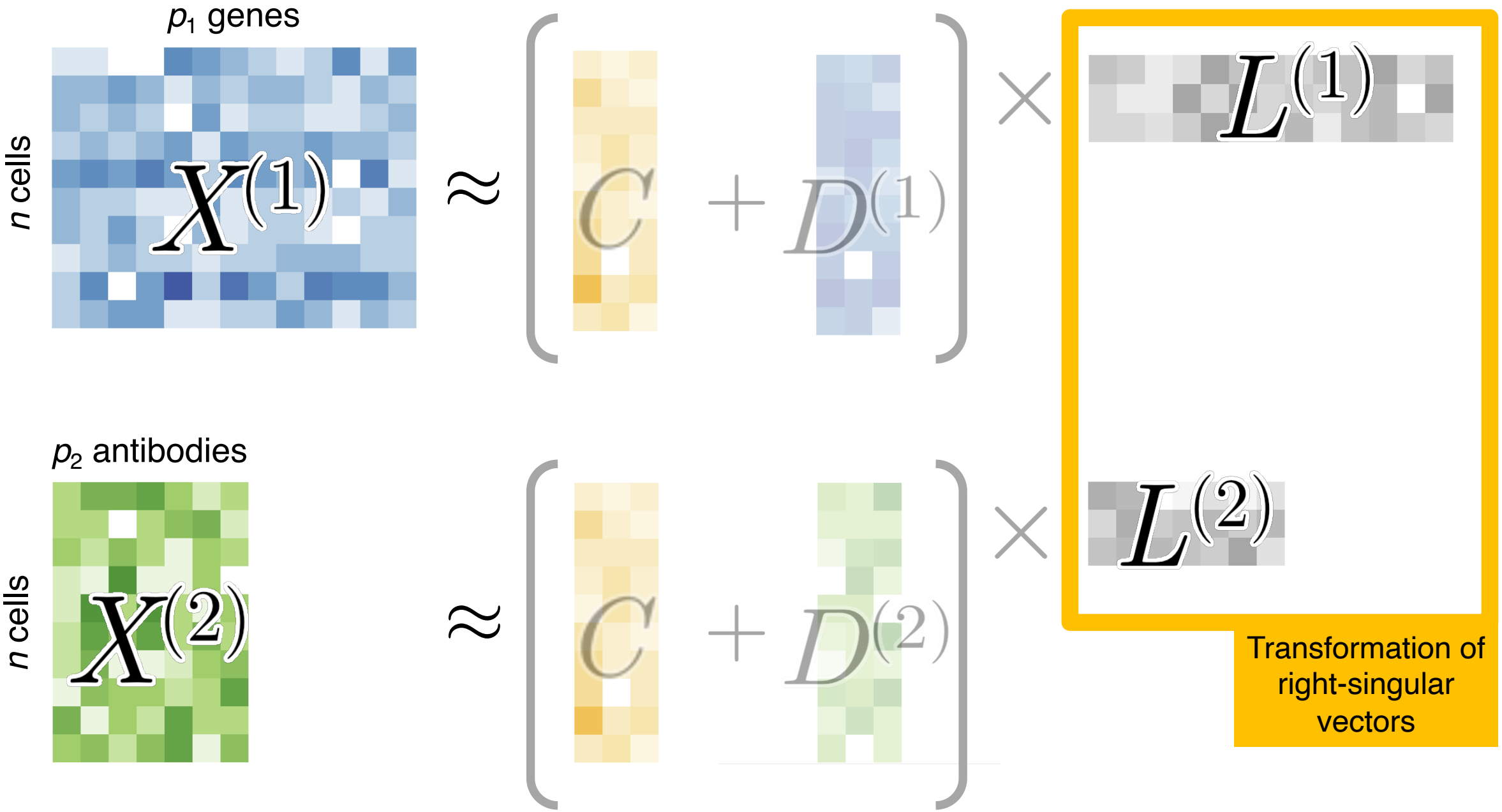


(Assumed to be sufficiently preprocessed)

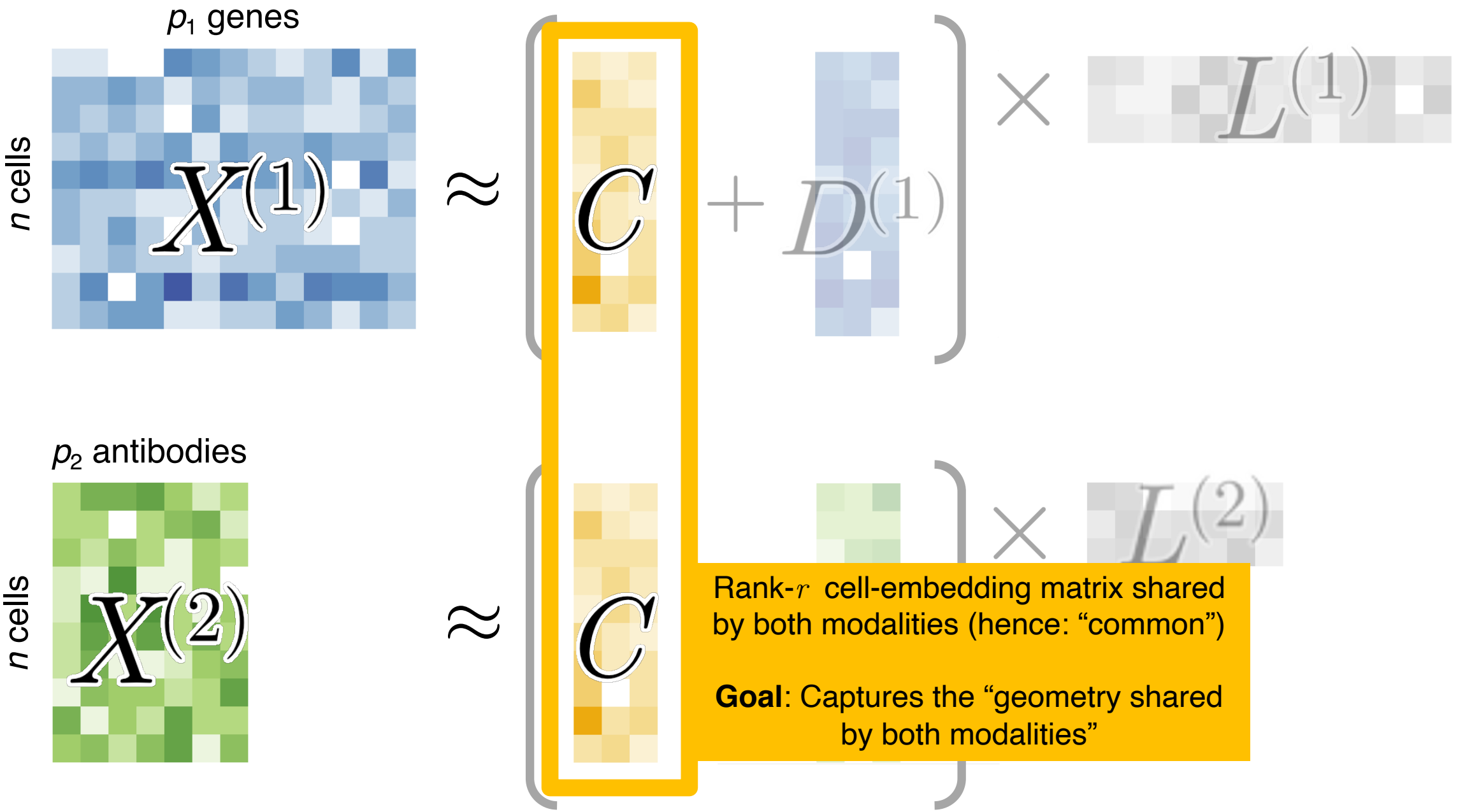
Proposed model for multi-modal data: Capturing the “shared” geometry



Proposed model for multi-modal data: Capturing the “shared” geometry



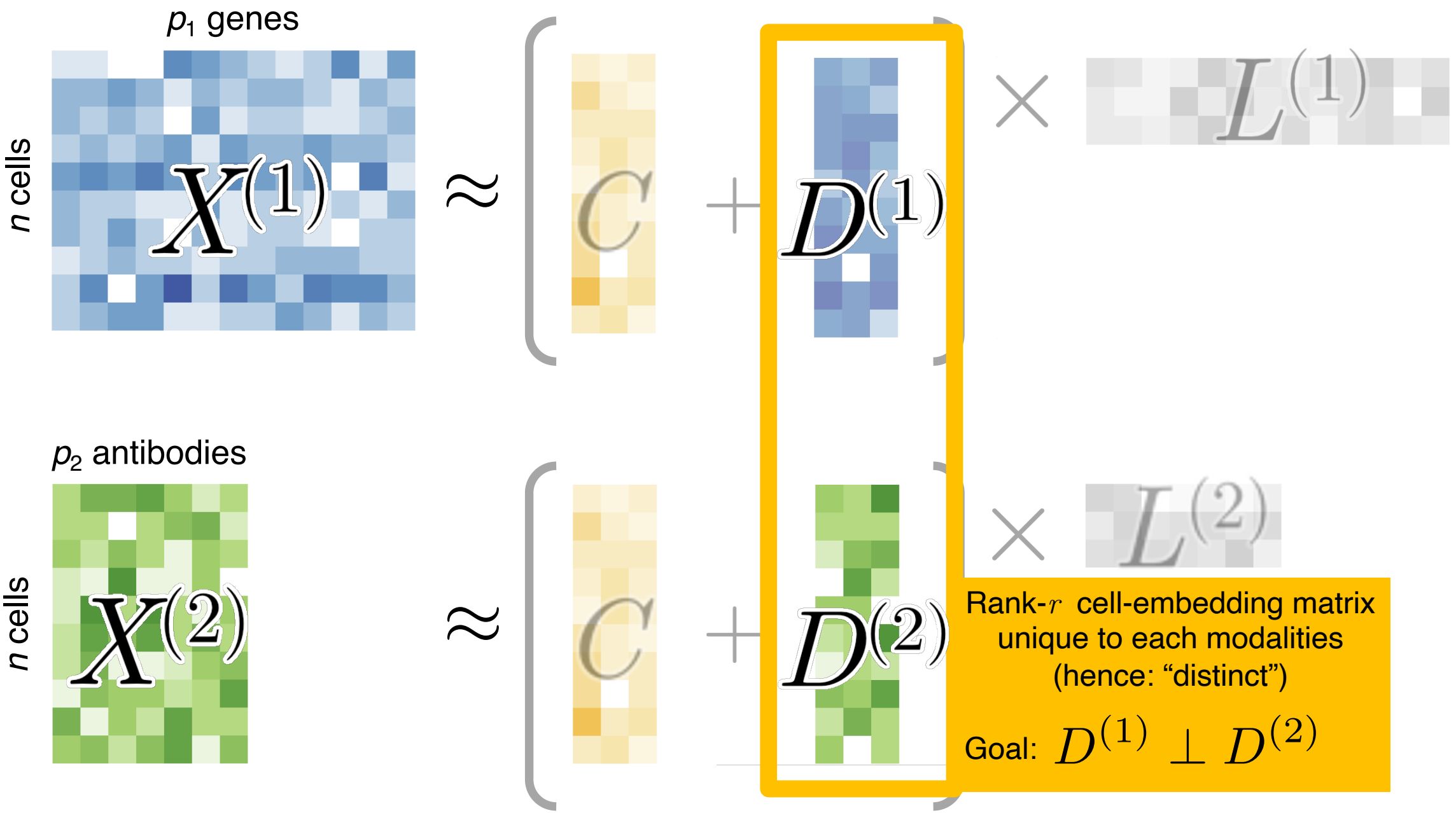
Proposed model for multi-modal data: Capturing the “shared” geometry



Rank-r cell-embedding matrix shared by both modalities (hence: “common”)

Goal: Captures the “geometry shared by both modalities”

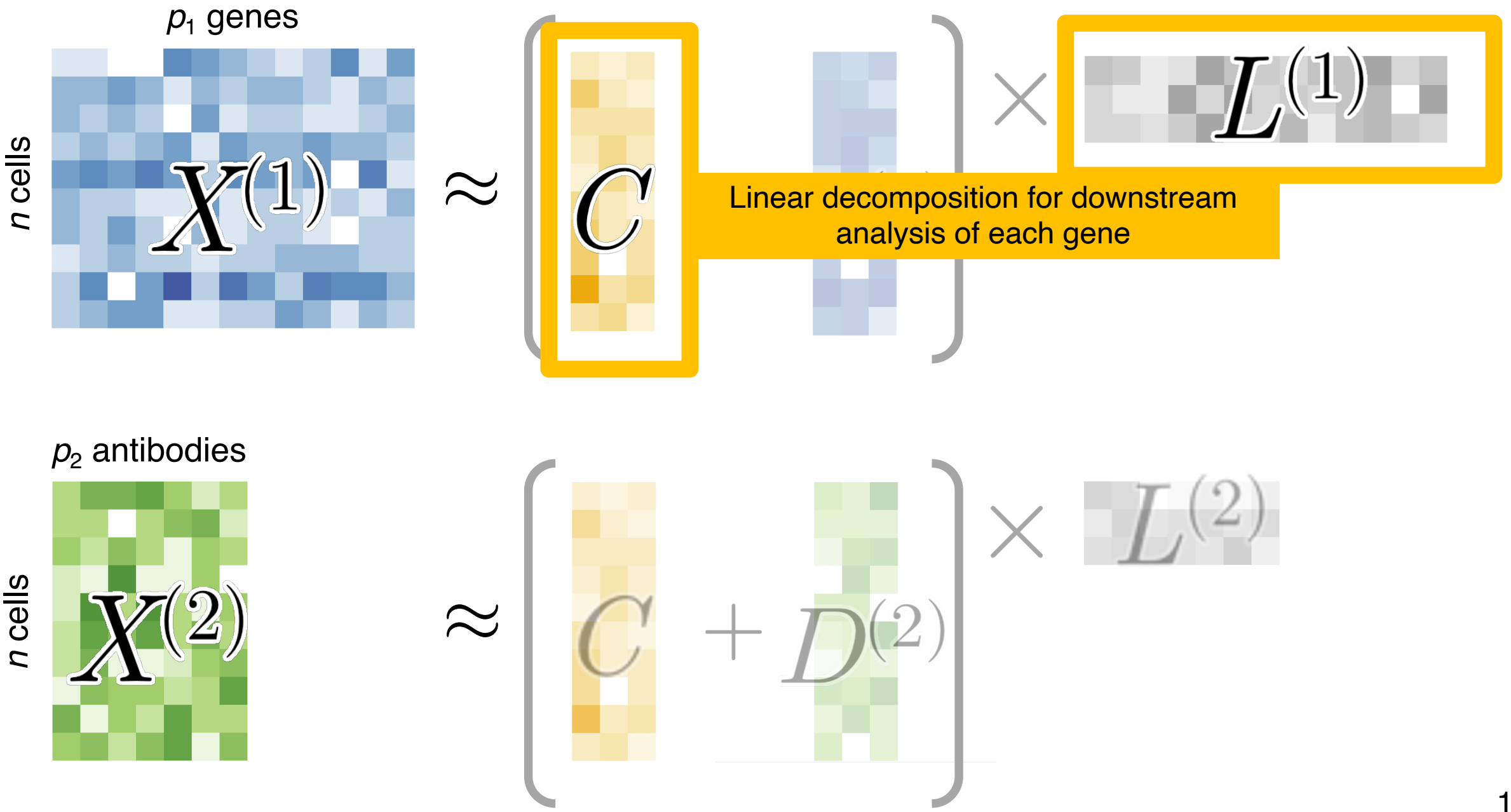
Proposed model for multi-modal data: Capturing the “shared” geometry



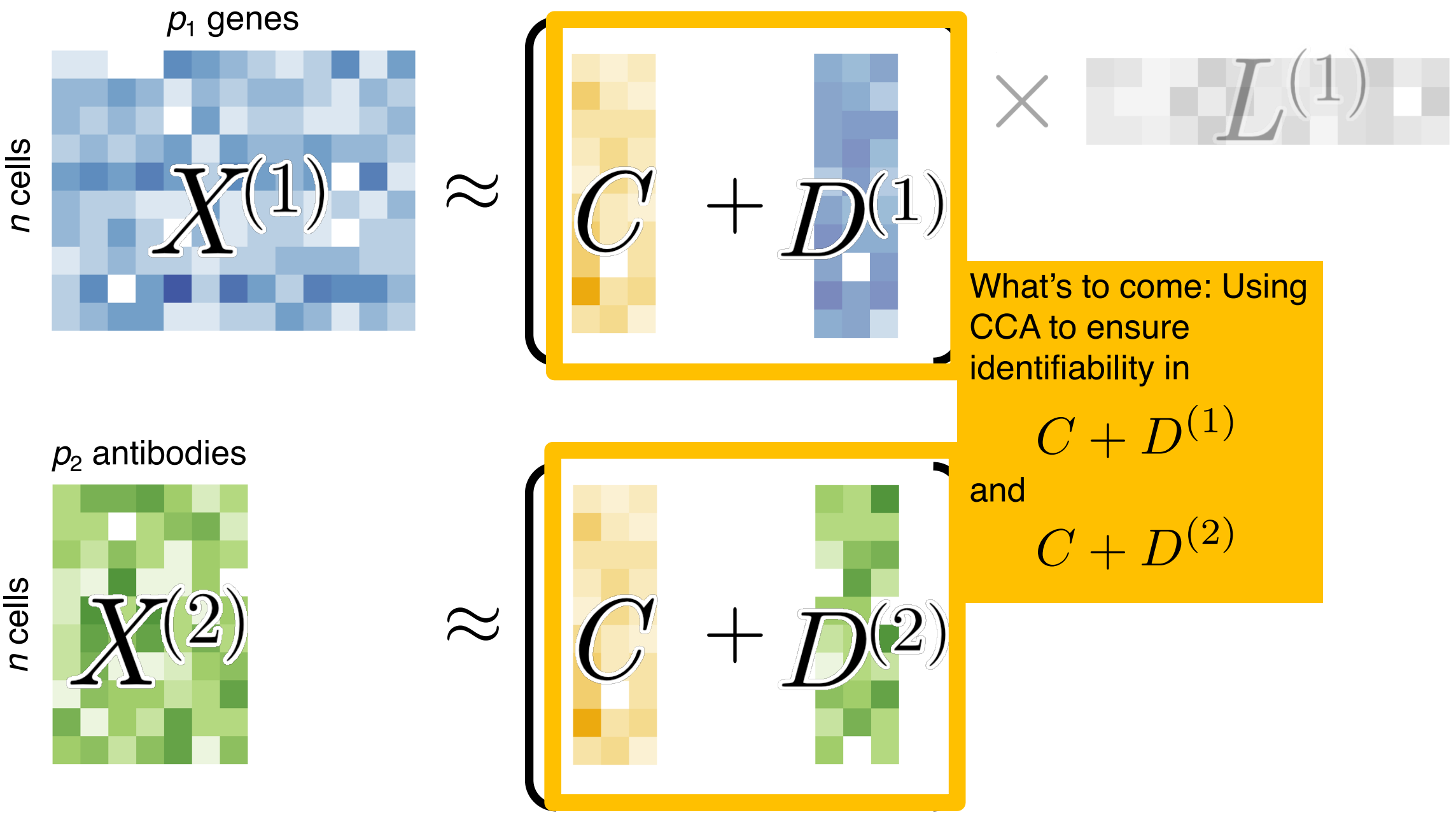
Rank-r cell-embedding matrix unique to each modalities (hence: “distinct”)

Goal: $D^{(1)} \perp D^{(2)}$

Proposed model for multi-modal data: Capturing the “shared” geometry



Proposed model for multi-modal data: Capturing the “shared” geometry



What’s to come: Using CCA to ensure identifiability in

$$C + D^{(1)}$$

and

$$C + D^{(2)}$$

Brief aside: Review of Canonical Correlation Analysis

CCA for the first latent dimension:

Given two modalities, find linear combination of variables that are highly correlated:

$$\{\widehat{a}, \widehat{b}\} = \underset{\substack{a \in \mathbb{R}^{p_1} \\ b \in \mathbb{R}^{p_2}}}{\arg\max} \, \mathsf{Corr}\Big(X^{(1)}a, X^{(2)}b\Big)$$

General multi-dimensional CCA:

$$\{\widehat{A}, \widehat{B}\} = \underset{\substack{A \in \mathbb{R}^{p_1 \times r} : A^\top \Sigma^{(1)} A = I_r \\ B \in \mathbb{R}^{p_2 \times r} : B^\top \Sigma^{(2)} B = I_r}}{\arg\max} \, \mathsf{Tr}\Big(A^\top \big(X^{(1)}\big)^\top X^{(2)} B\Big)$$

Brief aside: Review of Canonical Correlation Analysis

CCA for the first latent dimension:

Given two modalities, find linear combination of variables that are highly correlated:

$$\{\widehat{a}, \widehat{b}\} = \underset{\substack{a \in \mathbb{R}^{p_1} \\ b \in \mathbb{R}^{p_2}}}{\arg\max}\, \mathsf{Corr}\Big(\underbrace{X^{(1)}a, X^{(2)}b}_{\text{Pair of canonical score vectors}}\Big)$$

General multi-dimensional CCA:

$$\{\widehat{A}, \widehat{B}\} = \arg\max \mathsf{Tr}\Big(A^\top \big(X^{(1)}\big)^\top X^{(2)} B\Big)$$
$$A \in \mathbb{R}^{p_1 \times r} : A^\top \Sigma^{(1)} A = I_r$$
$$B \in \mathbb{R}^{p_2 \times r} : B^\top \Sigma^{(2)} B = I_r$$

Brief aside: Review of Canonical Correlation Analysis

CCA for the first latent dimension:

Given two modalities, find linear combination of variables that are highly correlated:

$$\{\widehat{a}, \widehat{b}\} = \underset{\substack{a \in \mathbb{R}^{p_1} \\ b \in \mathbb{R}^{p_2}}}{\arg\max} \, \mathsf{Corr}\Big(X^{(1)}a, X^{(2)}b\Big)$$

General multi-dimensional CCA:

$$\{\widehat{A}, \widehat{B}\} = \underset{\substack{A \in \mathbb{R}^{p_1 \times r} : A^\top \Sigma^{(1)} A = I_r \\ B \in \mathbb{R}^{p_2 \times r} : B^\top \Sigma^{(2)} B = I_r}}{\arg\max} \, \mathsf{Tr}\Big(A^\top \big(X^{(1)}\big)^\top X^{(2)} B\Big)$$

Orthogonality of canonical scores

Brief aside: Review of Canonical Correlation Analysis

CCA for the first latent dimension:

Given two modalities, find linear combination of variables that are highly correlated:

$$\{\widehat{a}, \widehat{b}\} = \underset{\substack{a \in \mathbb{R}^{p_1} \\ b \in \mathbb{R}^{p_2}}}{\arg\max} \text{Corr}\left(X^{(1)}a, X^{(2)}b\right)$$

General multi-dimensional CCA:

$$\{\widehat{A}, \widehat{B}\} = \arg\max \text{Tr}\left(A^\top \left(X^{(1)}\right)^\top X^{(2)} B\right)$$

$$A \in \mathbb{R}^{p_1 \times r} : A^\top \Sigma^{(1)} A = I_r$$

$$B \in \mathbb{R}^{p_2 \times r} : B^\top \Sigma^{(2)} B = I_r$$

At optimality: Orthogonality of canonical scores across modalities

Decomposition based on CCA:

Decomposition based on CCA:

$$\{\widehat{A}, \widehat{B}\} = \arg\max \mathsf{Tr}\left(A^\top \left(X^{(1)}\right)^\top X^{(2)} B\right)$$

$$A \in \mathbb{R}^{p_1 \times r} : A^\top \Sigma^{(1)} A = I_r$$

$$B \in \mathbb{R}^{p_2 \times r} : B^\top \Sigma^{(2)} B = I_r$$

Let: $Z^{(1)} = X^{(1)} \widehat{A}$

$$\underbrace{Z^{(2)}}_{\mathbb{R}^{n \times r}} = X^{(2)} \widehat{B}$$

Decomposition based on CCA:

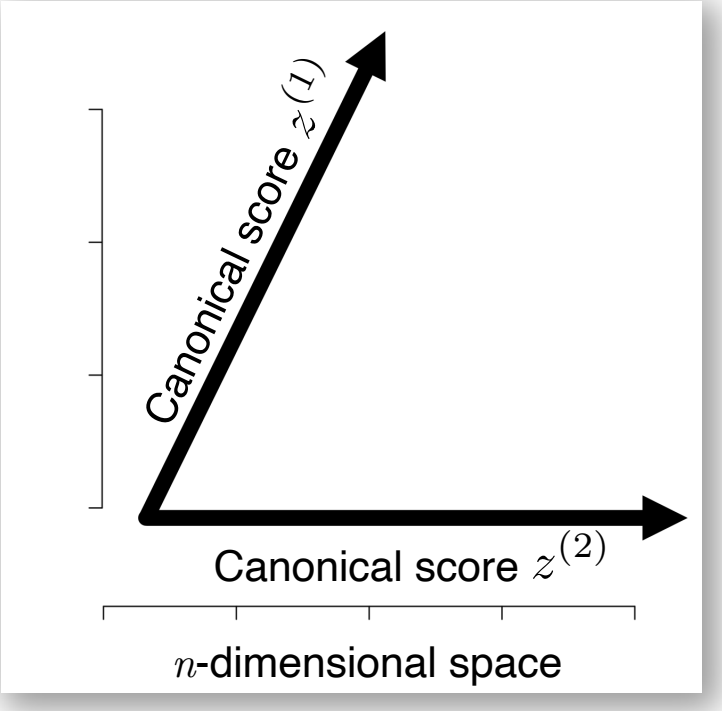
$$\{\widehat{A}, \widehat{B}\} = \arg\max \mathrm{Tr}\left(A^\top \left(X^{(1)}\right)^\top X^{(2)} B\right)$$

$$A \in \mathbb{R}^{p_1 \times r} : A^\top \Sigma^{(1)} A = I_r$$

$$B \in \mathbb{R}^{p_2 \times r} : B^\top \Sigma^{(2)} B = I_r$$

Let: $Z^{(1)} = X^{(1)} \widehat{A}$

$Z^{(2)} = X^{(2)} \widehat{B}$



The 2D hyperplane in *n*-dimensional space

Inspired by D-CCA (Zhu et al., 2020)

Decomposition based on CCA:

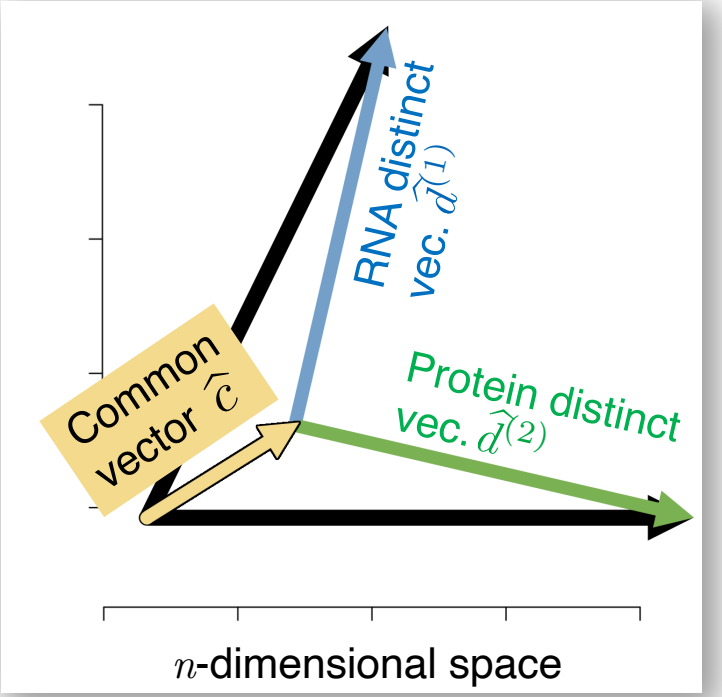
$$\{\widehat{A}, \widehat{B}\} = \arg\max \mathrm{Tr}\left(A^\top \left(X^{(1)}\right)^\top X^{(2)} B\right)$$

$$A \in \mathbb{R}^{p_1 \times r} : A^\top \Sigma^{(1)} A = I_r$$

$$B \in \mathbb{R}^{p_2 \times r} : B^\top \Sigma^{(2)} B = I_r$$

Let: $Z^{(1)} = X^{(1)} \widehat{A}$

$Z^{(2)} = X^{(2)} \widehat{B}$



Decomposition:

$$z^{(1)} = \widehat{c} + \widehat{d}^{(1)}$$

$$z^{(2)} = \widehat{c} + \widehat{d}^{(2)}$$

(quantifying what "common" means)

Inspired by D-CCA (Zhu et al., 2020)

Decomposition based on CCA:

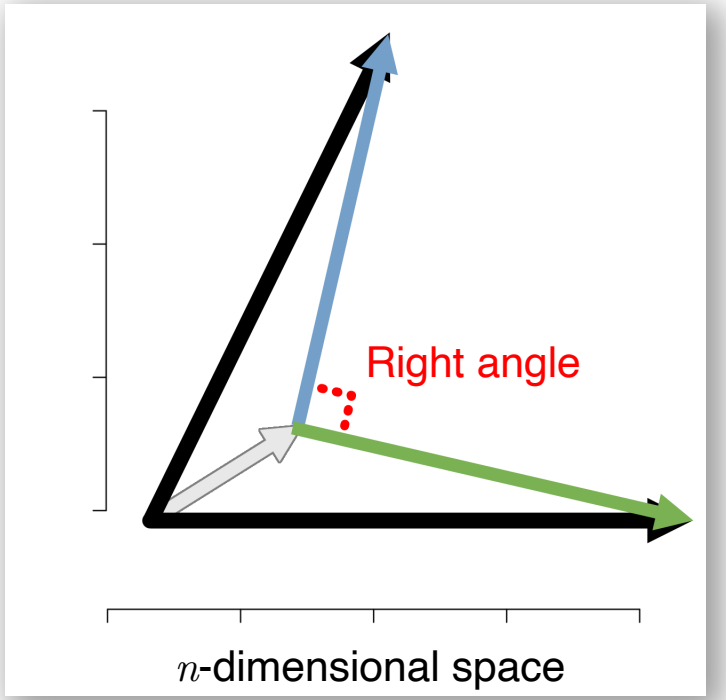
$$\{\widehat{A}, \widehat{B}\} = \arg\max \mathsf{Tr}\left(A^\top \left(X^{(1)}\right)^\top X^{(2)} B\right)$$

$$A \in \mathbb{R}^{p_1 \times r} : A^\top \Sigma^{(1)} A = I_r$$

$$B \in \mathbb{R}^{p_2 \times r} : B^\top \Sigma^{(2)} B = I_r$$

Let: $Z^{(1)} = X^{(1)} \widehat{A}$

$Z^{(2)} = X^{(2)} \widehat{B}$



Constrained to:

$\widehat{d}^{(1)} \perp \widehat{d}^{(2)}$

(quantifying what "distinct" means)

Inspired by D-CCA (Zhu et al., 2020)

Decomposition based on CCA:

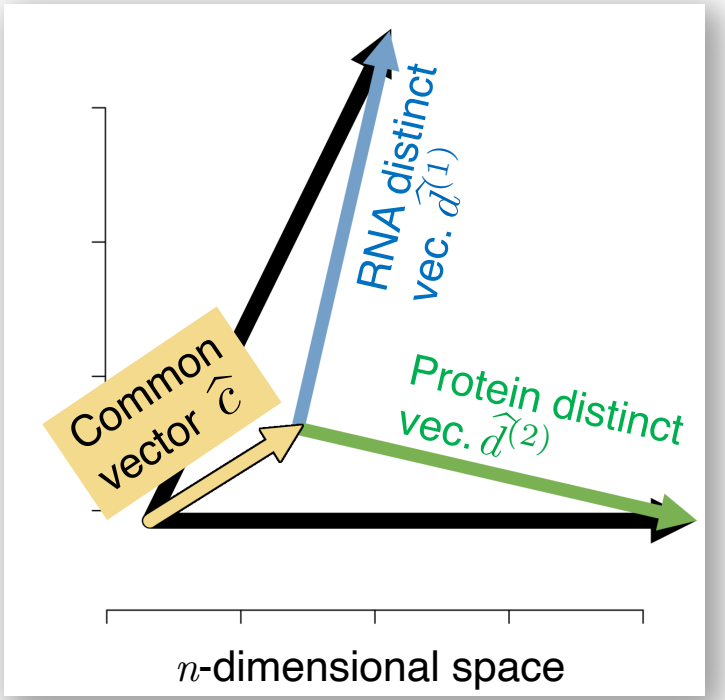
$$\{\widehat{A}, \widehat{B}\} = \arg\max \mathrm{Tr}\left(A^\top \left(X^{(1)}\right)^\top X^{(2)} B\right)$$

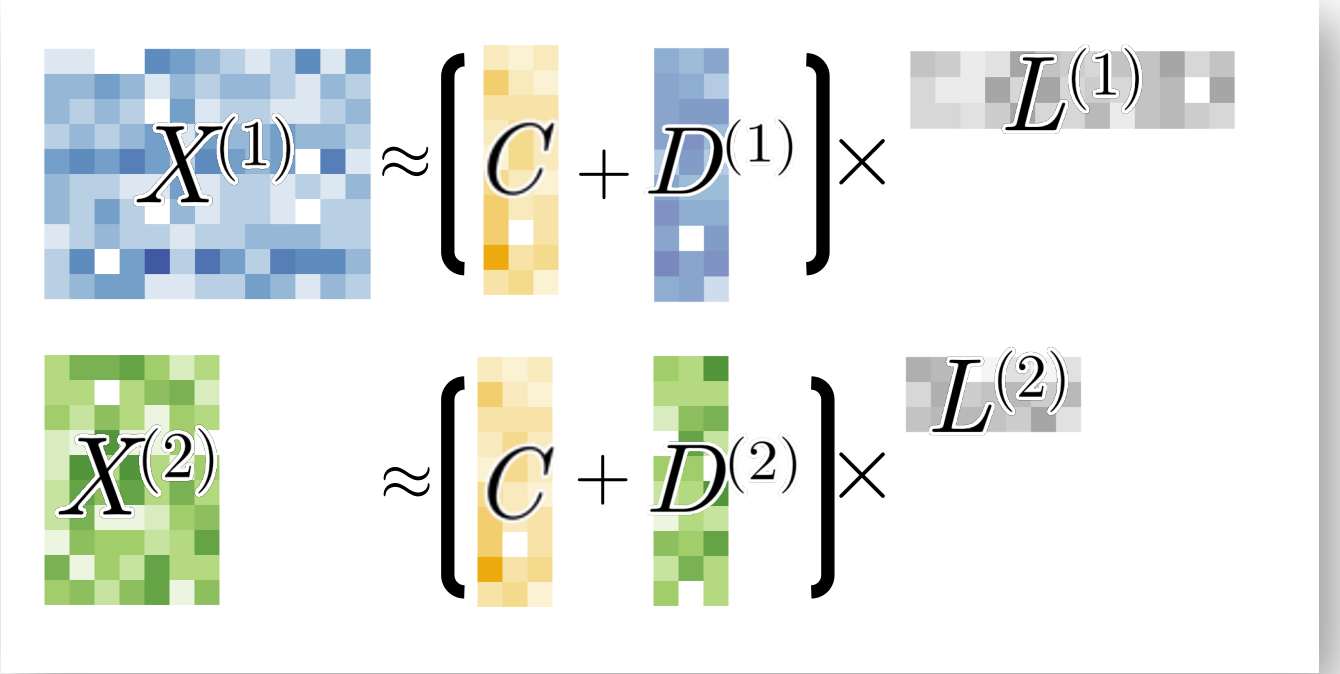
$$A \in \mathbb{R}^{p_1 \times r} : A^\top \Sigma^{(1)} A = I_r$$

$$B \in \mathbb{R}^{p_2 \times r} : B^\top \Sigma^{(2)} B = I_r$$

Let: $Z^{(1)} = X^{(1)} \widehat{A}$

$Z^{(2)} = X^{(2)} \widehat{B}$





Inspired by D-CCA (Zhu et al., 2020)

Decomposition based on CCA:

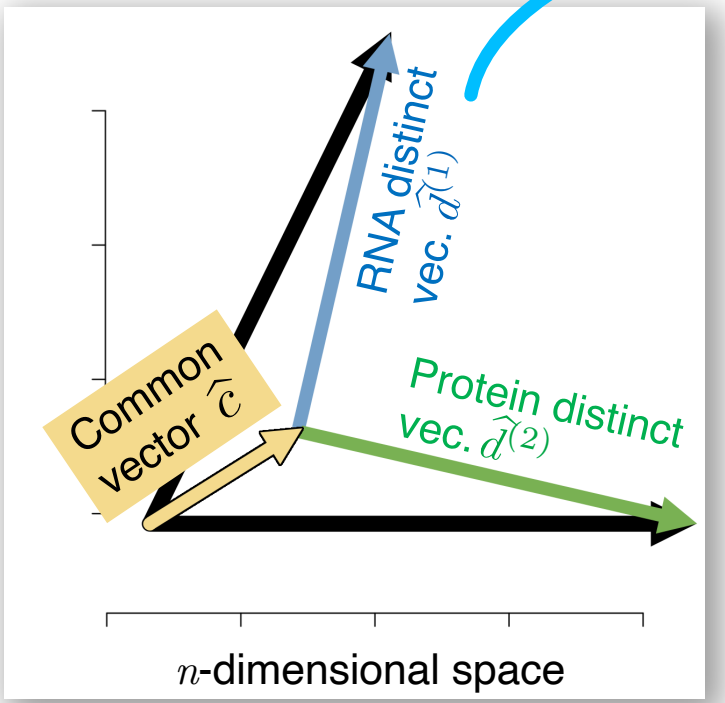
$$\{\widehat{A}, \widehat{B}\} = \arg\max \mathrm{Tr}\left(A^\top \left(X^{(1)}\right)^\top X^{(2)} B\right)$$

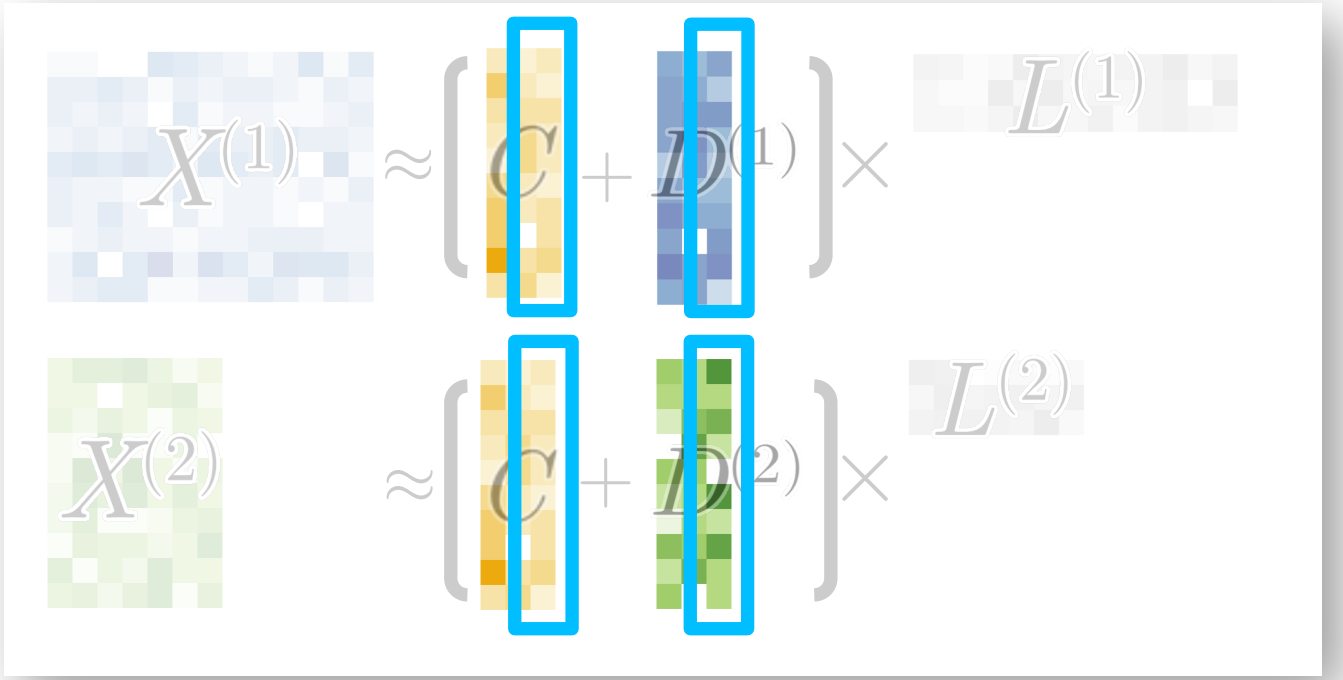
$$A \in \mathbb{R}^{p_1 \times r} : A^\top \Sigma^{(1)} A = I_r$$

$$B \in \mathbb{R}^{p_2 \times r} : B^\top \Sigma^{(2)} B = I_r$$

Let: $Z^{(1)} = X^{(1)} \widehat{A}$

$Z^{(2)} = X^{(2)} \widehat{B}$





Inspired by D-CCA (Zhu et al., 2020)

Decomposition based on CCA:

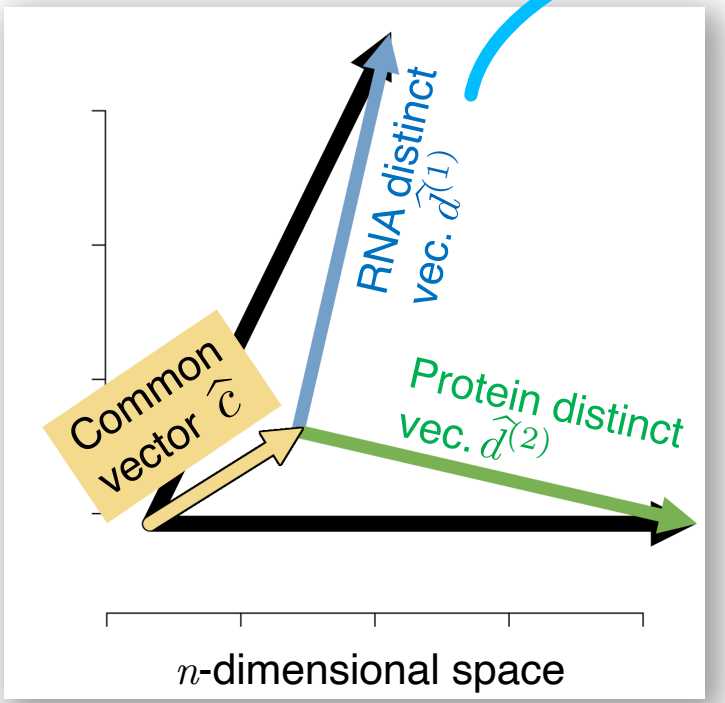
$$\{\widehat{A}, \widehat{B}\} = \arg\max \mathrm{Tr}\left(A^\top \left(X^{(1)}\right)^\top X^{(2)} B\right)$$

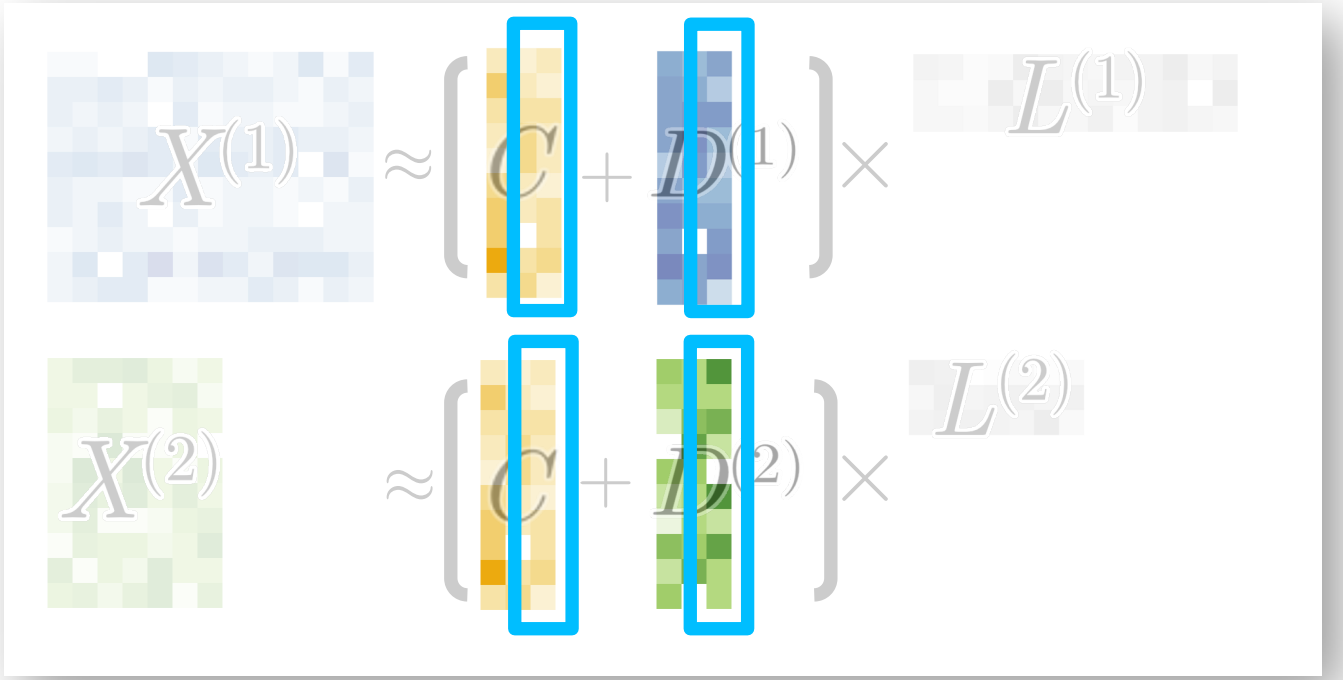
$$A \in \mathbb{R}^{p_1 \times r} : A^\top \Sigma^{(1)} A = I_r$$
$$B \in \mathbb{R}^{p_2 \times r} : B^\top \Sigma^{(2)} B = I_r$$

Identifiability ensured thanks to CCA's properties

Let: $Z^{(1)} = X^{(1)} \widehat{A}$

$Z^{(2)} = X^{(2)} \widehat{B}$





Inspired by D-CCA (Zhu et al., 2020)

Decomposition based on CCA

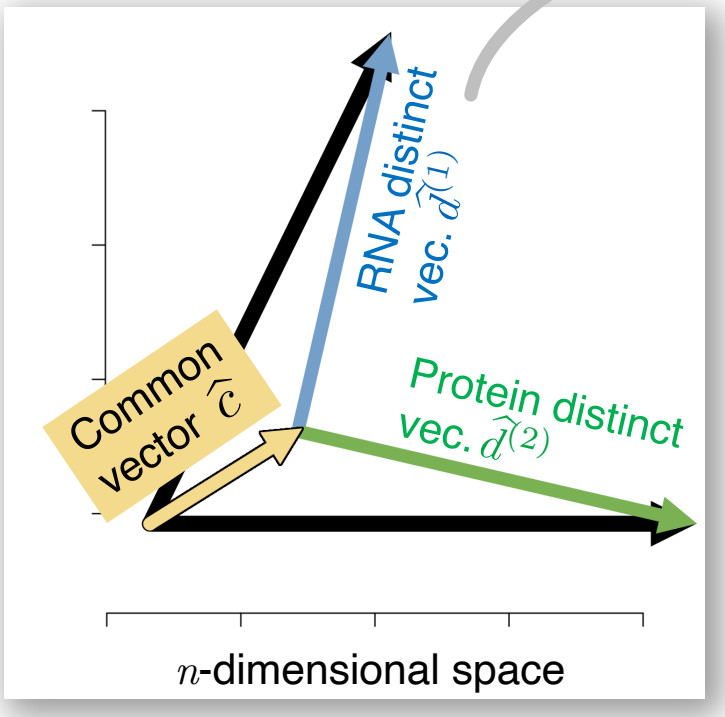
$$\{\widehat{A}, \widehat{B}\} = \arg\max \mathrm{Tr}\left(A^\top \left(X^{(1)}\right)^\top X^{(2)} B\right)$$

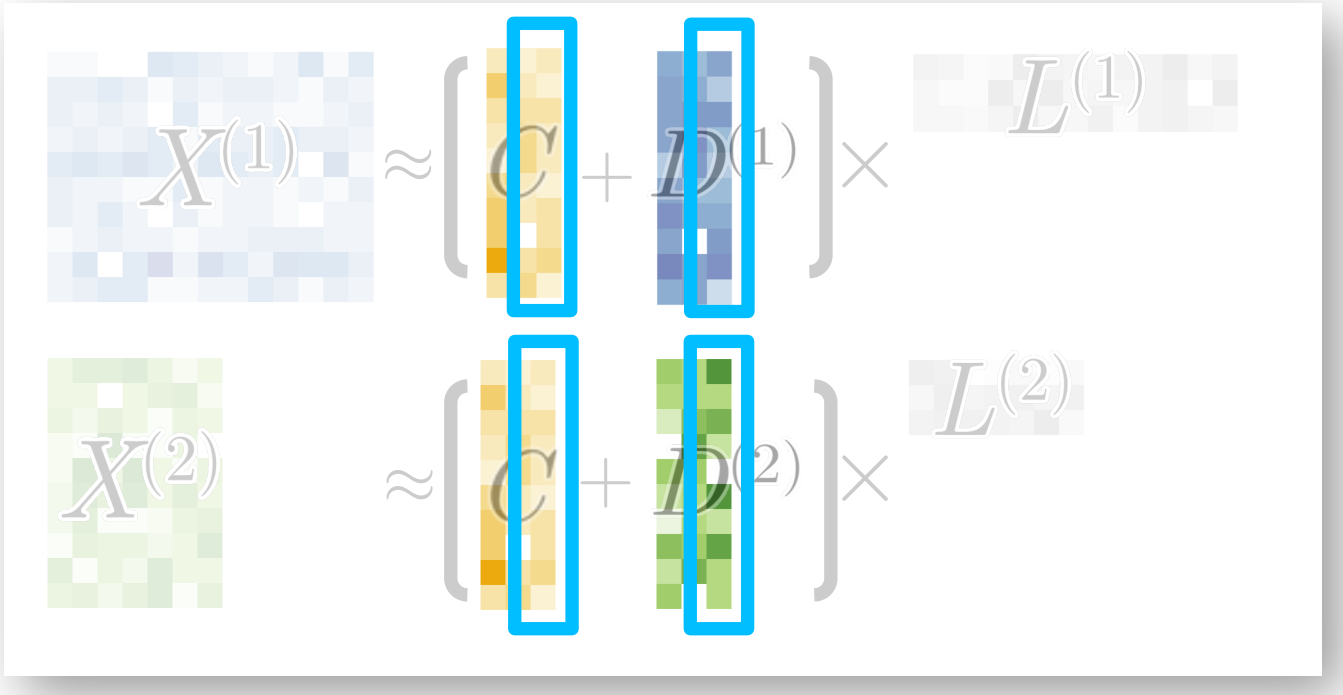
$$A \in \mathbb{R}^{p_1 \times r} : A^\top \Sigma^{(1)} A = I_r$$
$$B \in \mathbb{R}^{p_2 \times r} : B^\top \Sigma^{(2)} B = I_r$$

Identifiability ensured thanks to CCA's properties

Let: $Z^{(1)} = X^{(1)} \widehat{A}$

$Z^{(2)} = X^{(2)} \widehat{B}$





✓ Decomp. for each latent dim. ✓ Ortho. distinct vectors ? Common vector shared geometry

Decomposition based on CCA

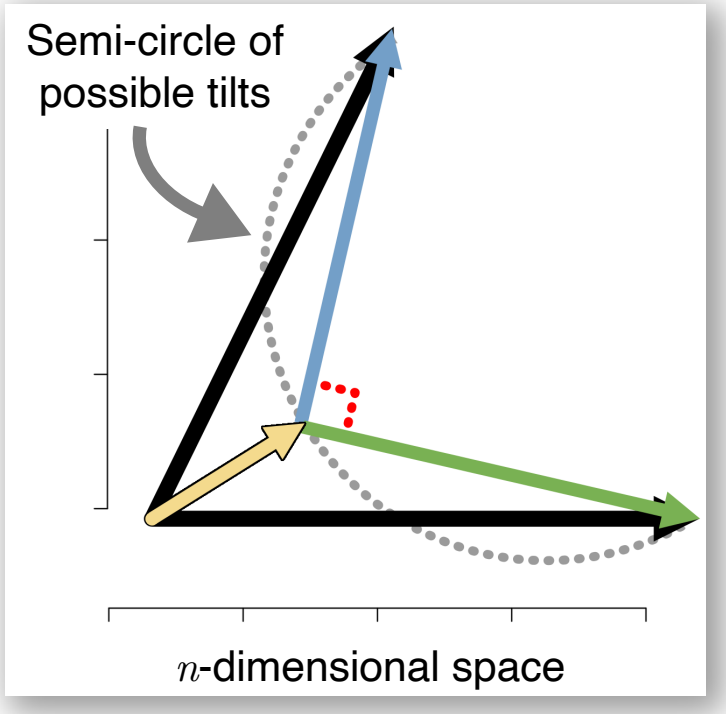
$$\{\widehat{A}, \widehat{B}\} = \arg\max \mathsf{Tr}\left(A^\top \left(X^{(1)}\right)^\top X^{(2)} B\right)$$

$$A \in \mathbb{R}^{p_1 \times r} : A^\top \Sigma^{(1)} A = I_r$$
$$B \in \mathbb{R}^{p_2 \times r} : B^\top \Sigma^{(2)} B = I_r$$

Identifiability ensured thanks to CCA's properties

Let: $Z^{(1)} = X^{(1)} \widehat{A}$

$Z^{(2)} = X^{(2)} \widehat{B}$

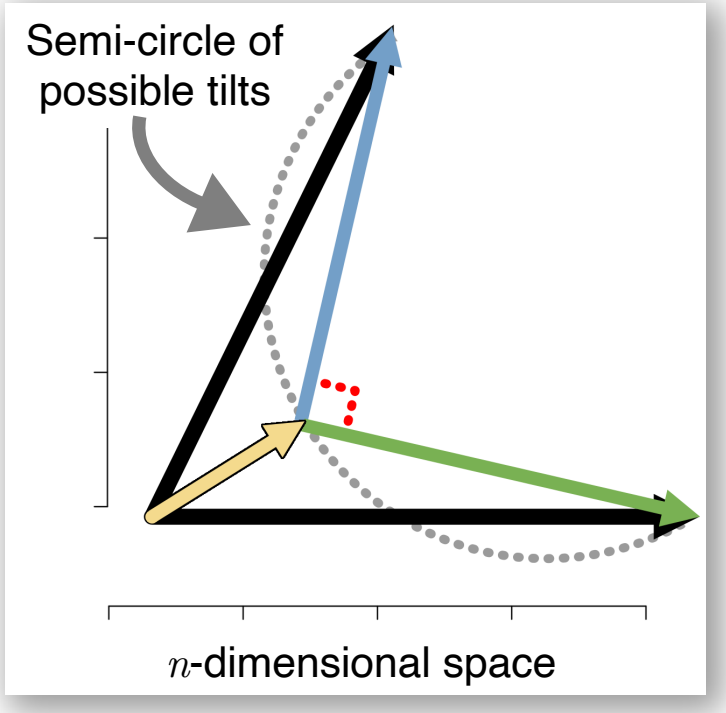


Our insight:

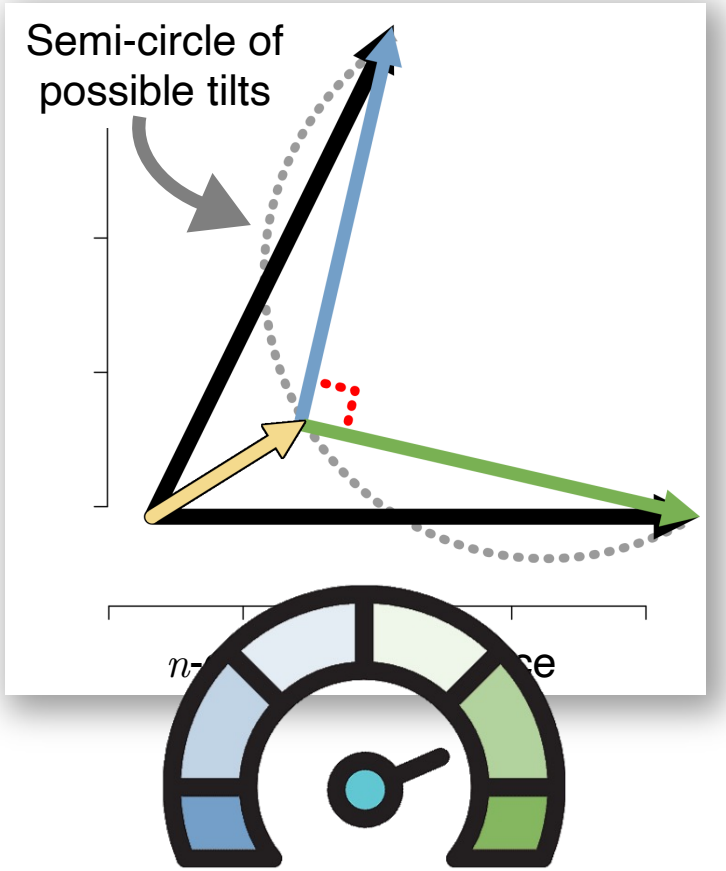
The common vector can fall anywhere along this **semi-circle** and still retain orthogonal distinct vectors.

(One extra degree of freedom per dimension)

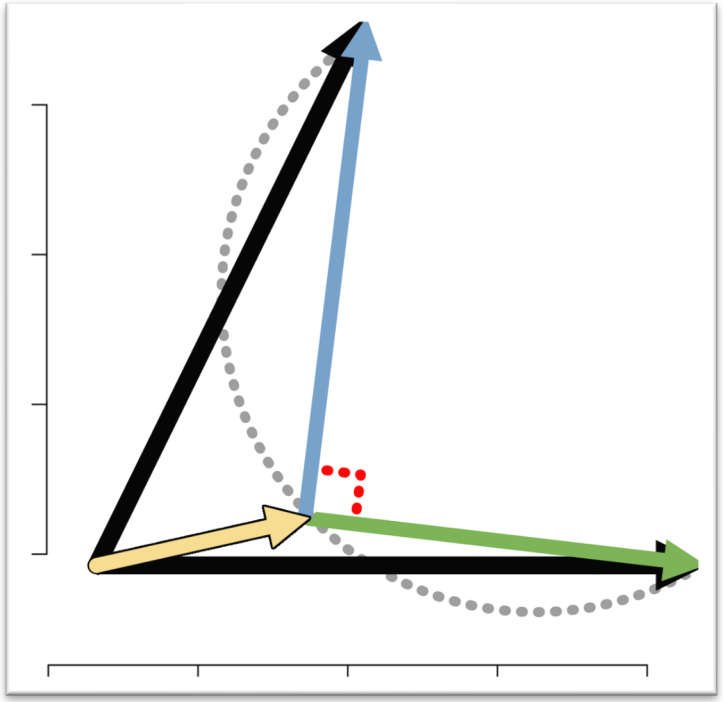
Our novelty (combining ideas in CCA with geometry): The tilt of the common vector (if chosen appropriately) allows the common embedding to reflect the shared geometry between both modalities.



Our novelty (combining ideas in CCA with geometry): The tilt of the common vector (if chosen appropriately) allows the common embedding to reflect the shared geometry between both modalities.



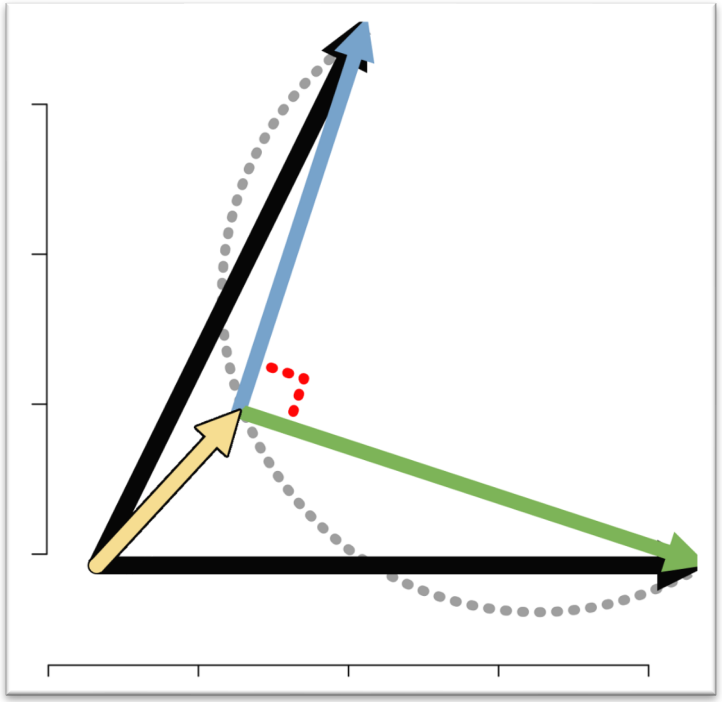
Tilt towards Modality 2



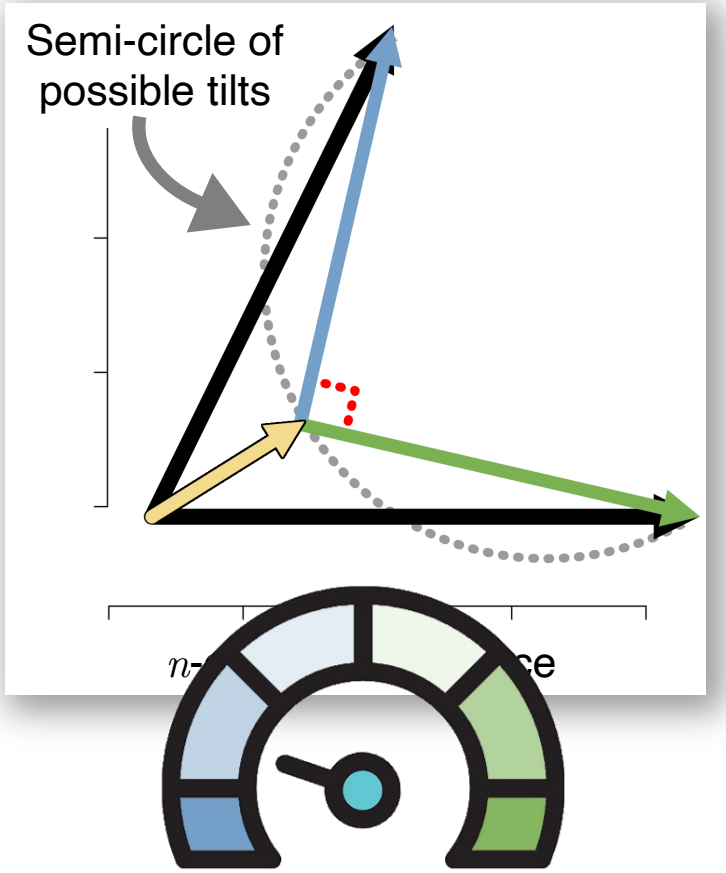
Common embedding has geometry similar to Modality 2

Our novelty (combining ideas in CCA with geometry): The tilt of the common vector (if chosen appropriately) allows the common embedding to reflect the shared geometry between both modalities.

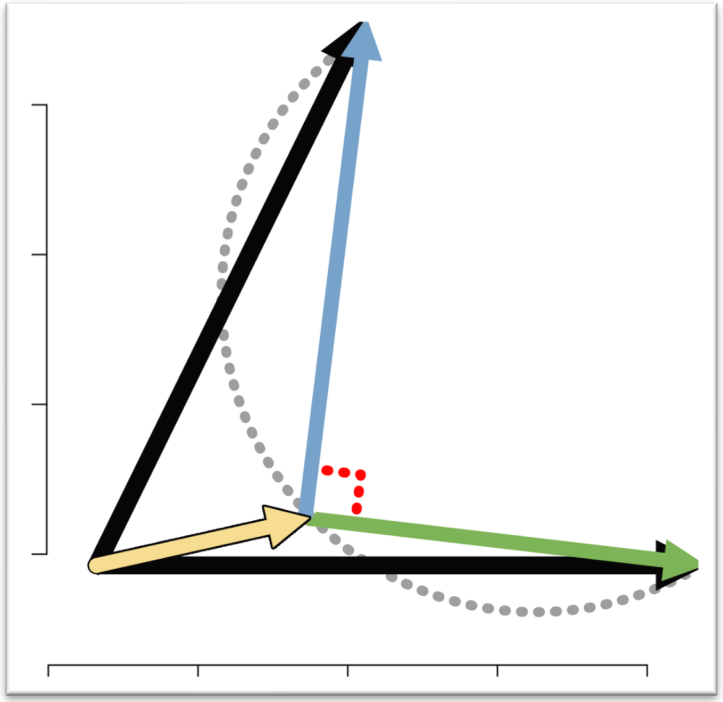
Tilt towards Modality 1



Common embedding has geometry similar to Modality 1



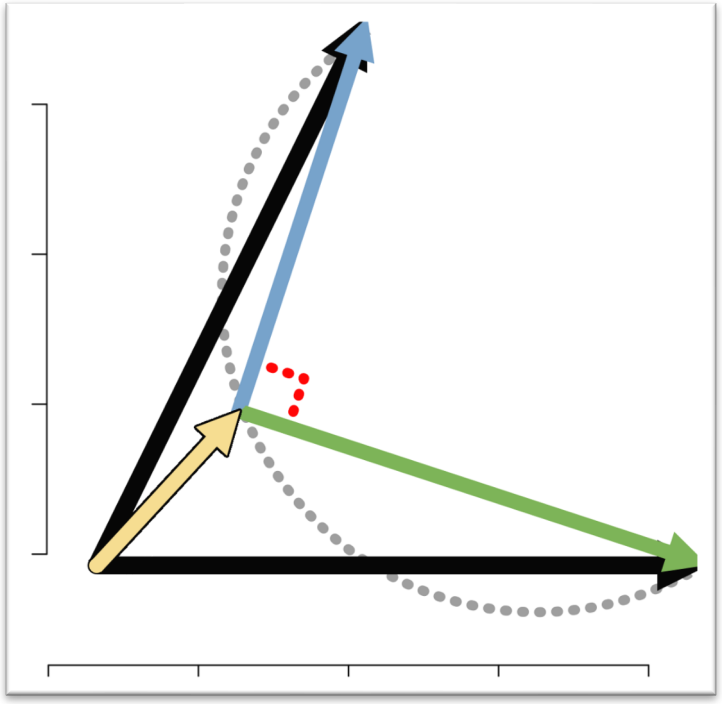
Tilt towards Modality 2

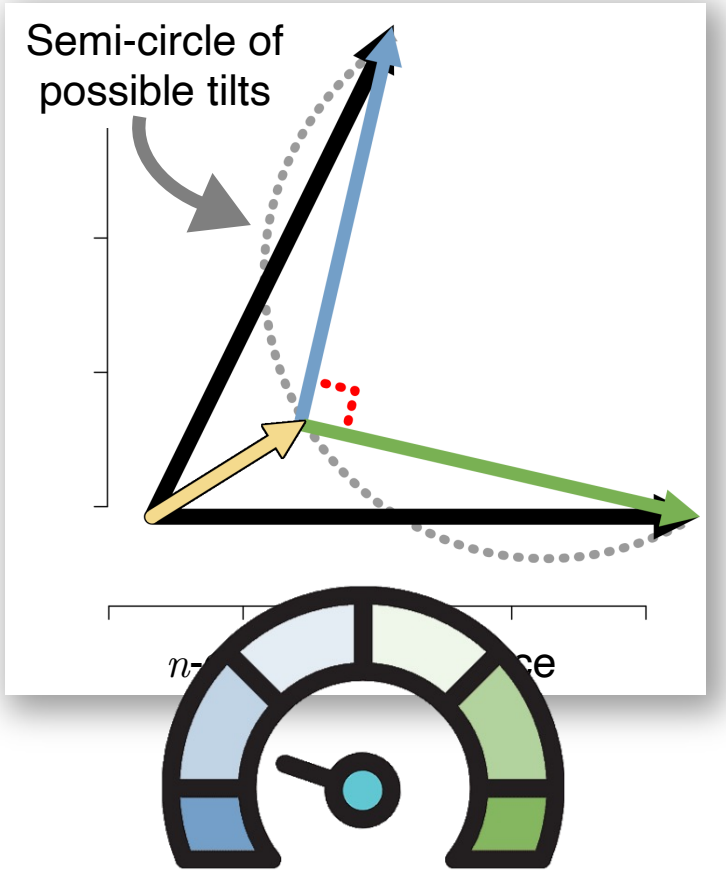


Common embedding has geometry similar to Modality 2

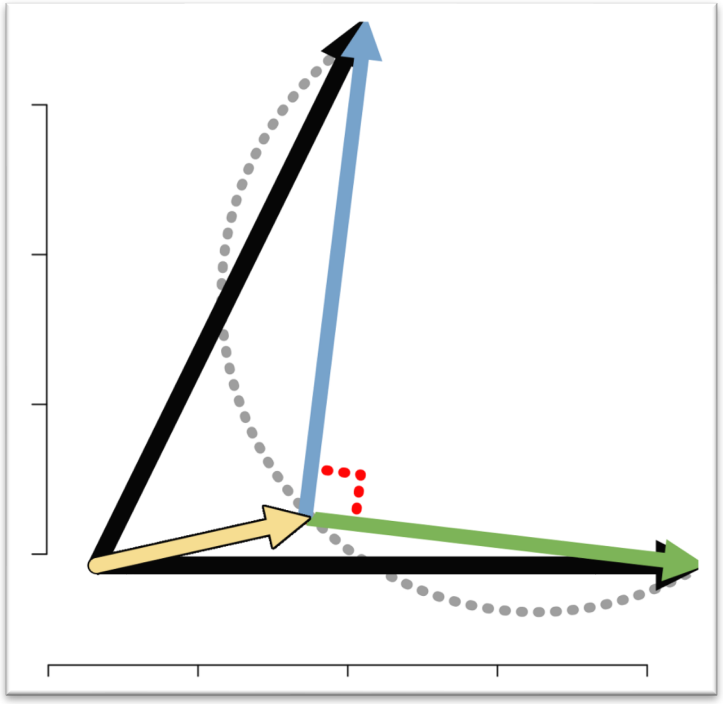
Our novelty (combining ideas in CCA with geometry): The tilt of the common vector (if chosen appropriately) allows the common embedding to reflect the shared geometry between both modalities.

Tilt towards Modality 1





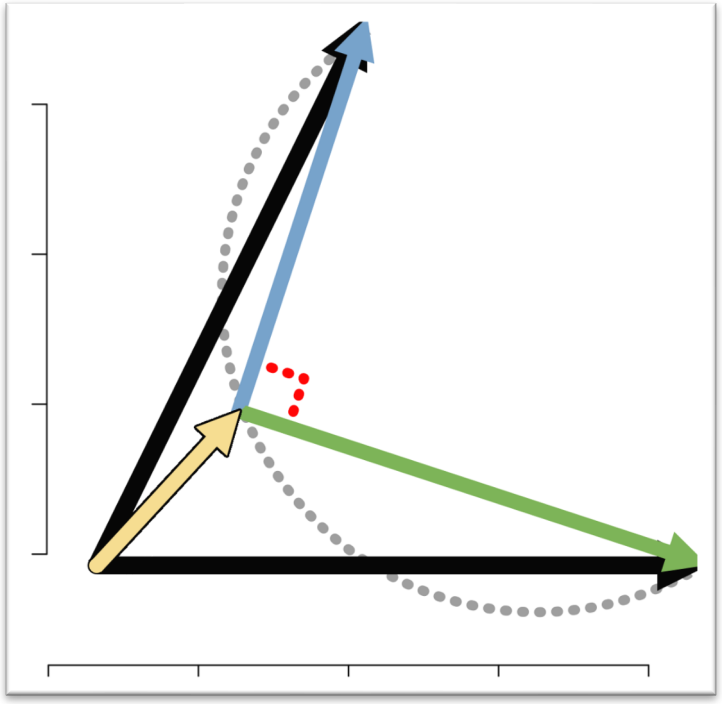
Tilt towards Modality 2

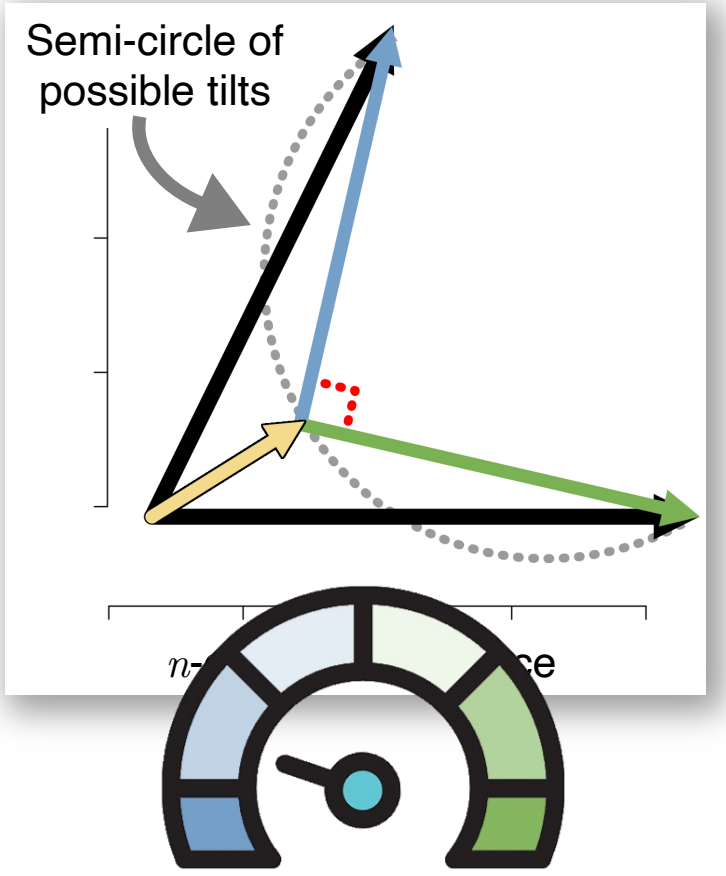


Main Goal: To estimate the tilt $\tau_j \in [1, 2]$ in each latent dimension j to capture the desired shared geometry

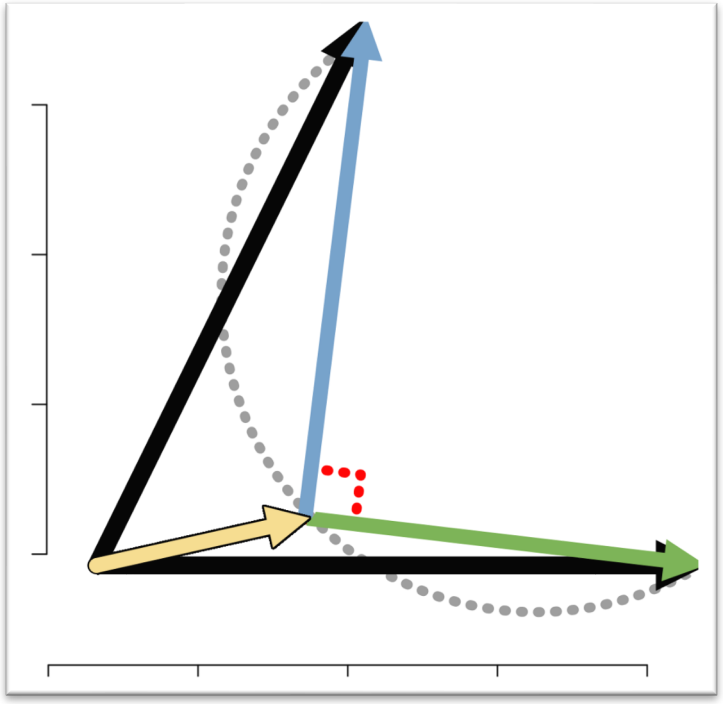
Our novelty (combining ideas in CCA with geometry): The tilt of the common vector (if chosen appropriately) allows the common embedding to reflect the shared geometry between both modalities.

Tilt towards Modality 1





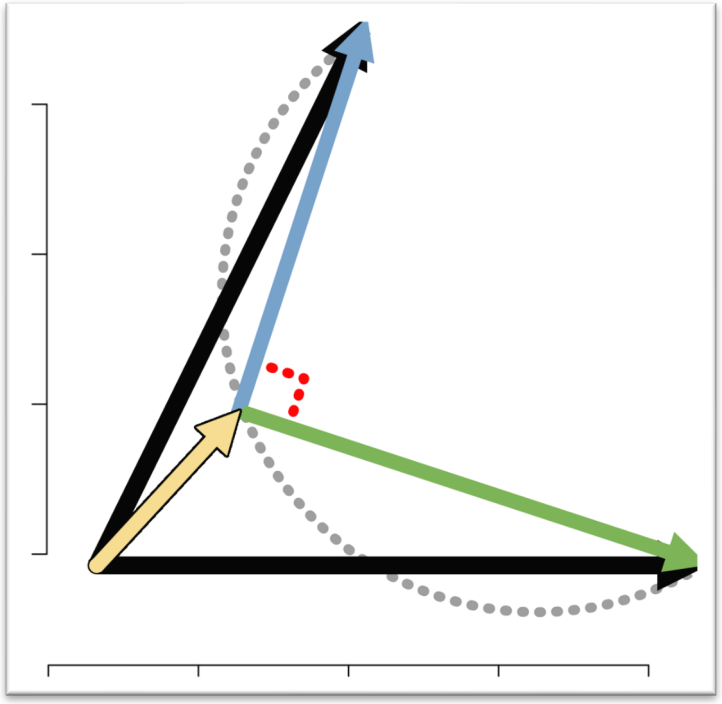
Tilt towards Modality 2

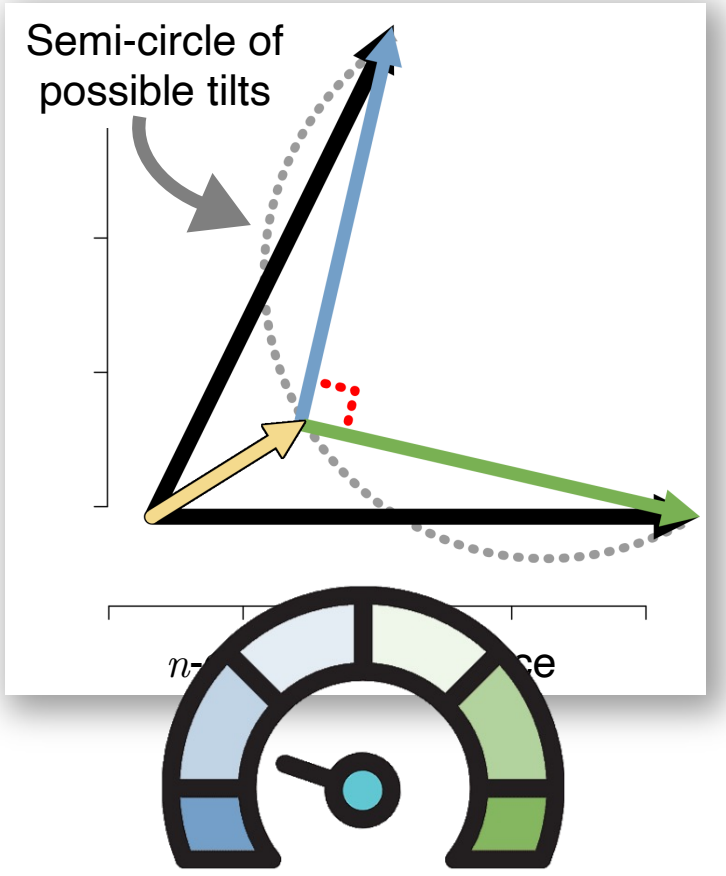


Main Goal: To estimate the tilt $\tau_j \in [1, 2]$ in each latent dimension j to capture the desired shared geometry

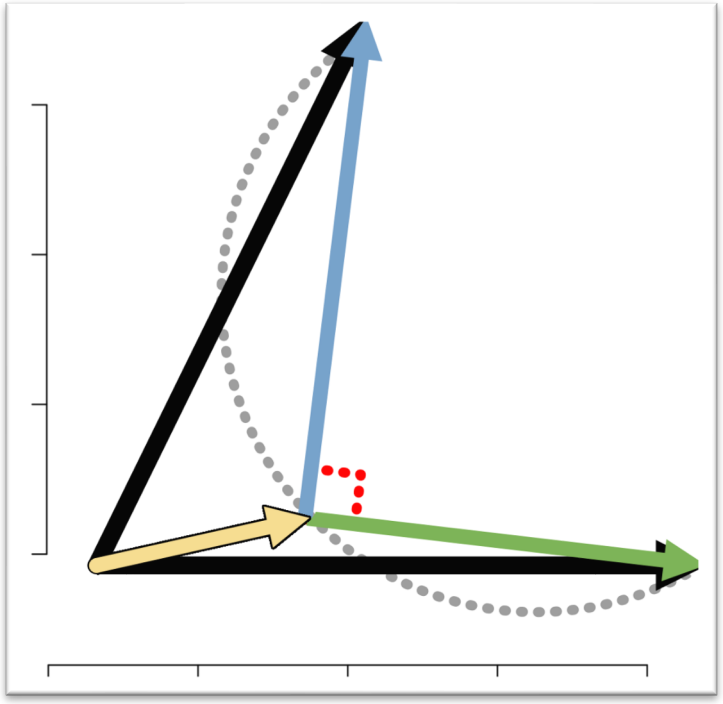
Our novelty (combining ideas in CCA with geometry): The tilt of the common vector (if chosen appropriately) allows the common embedding to reflect the shared geometry between both modalities.

Tilt towards Modality 1





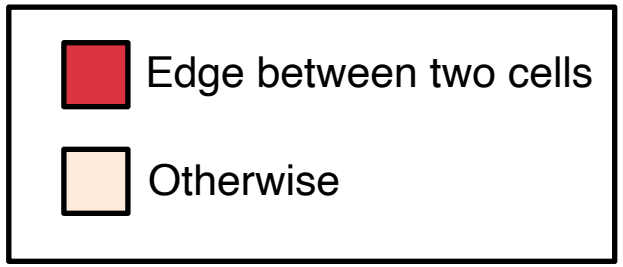
Tilt towards Modality 2



Main Goal: To estimate the tilt $\tau_j \in [1, 2]$ in each latent dimension j to capture the desired shared geometry

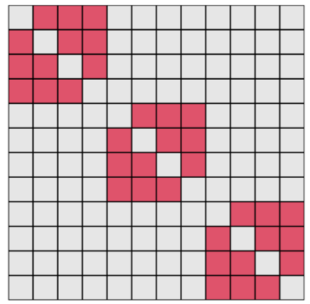
(Unsupervised) Estimation: Part 1 – Defining the shared geometry

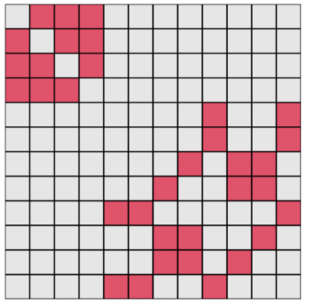
(Unsupervised) Estimation: Part 1 – Defining the shared geometry



1

Cell-by-cell nearest-neighbor graph (RNA modality)



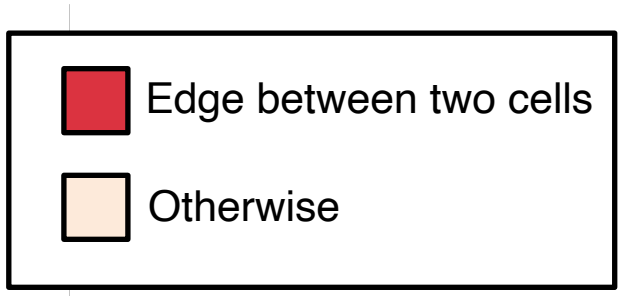


Cell-by-cell nearest-neighbor graph (Protein modality)

"3 clusters"

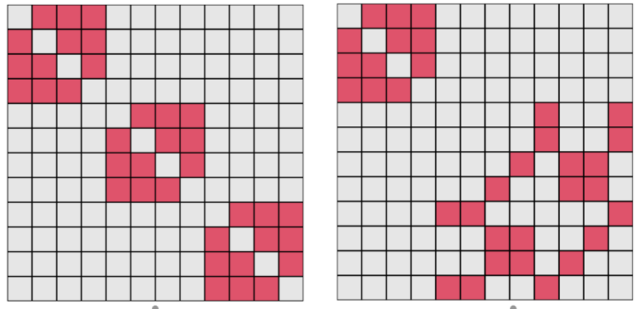
"2 clusters"

(Unsupervised) Estimation: Part 1 – Defining the shared geometry



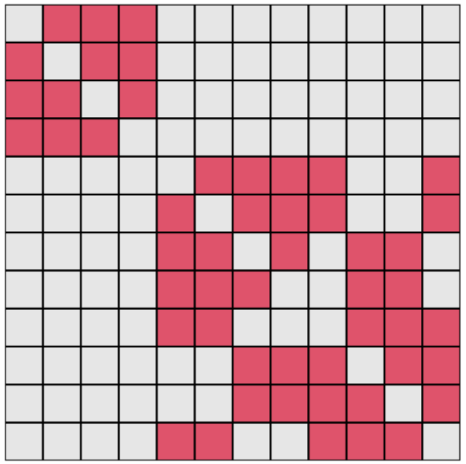
1

Cell-by-cell nearest-neighbor graph (RNA modality)



Cell-by-cell nearest-neighbor graph (Protein modality)

2 Desired shared geometry



“2 clusters”

Entry-wise operation: Aggregating the edges in both graphs

Intuition: Cell clusters unique to one modality should not be appear in the shared geometry.

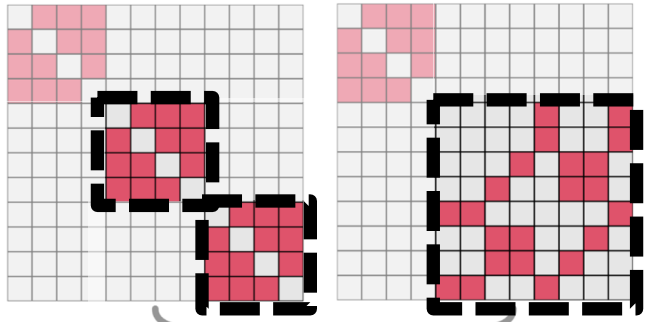
(Unsupervised) Estimation: Part 1 – Defining the shared geometry

Edge between two cells

Otherwise

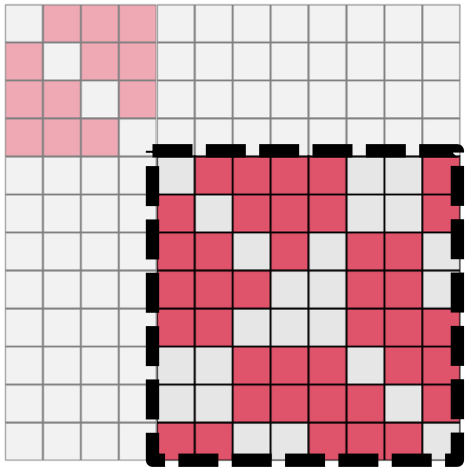
1

Cell-by-cell nearest-neighbor graph (RNA modality)



Cell-by-cell nearest-neighbor graph (Protein modality)

2 Desired shared geometry

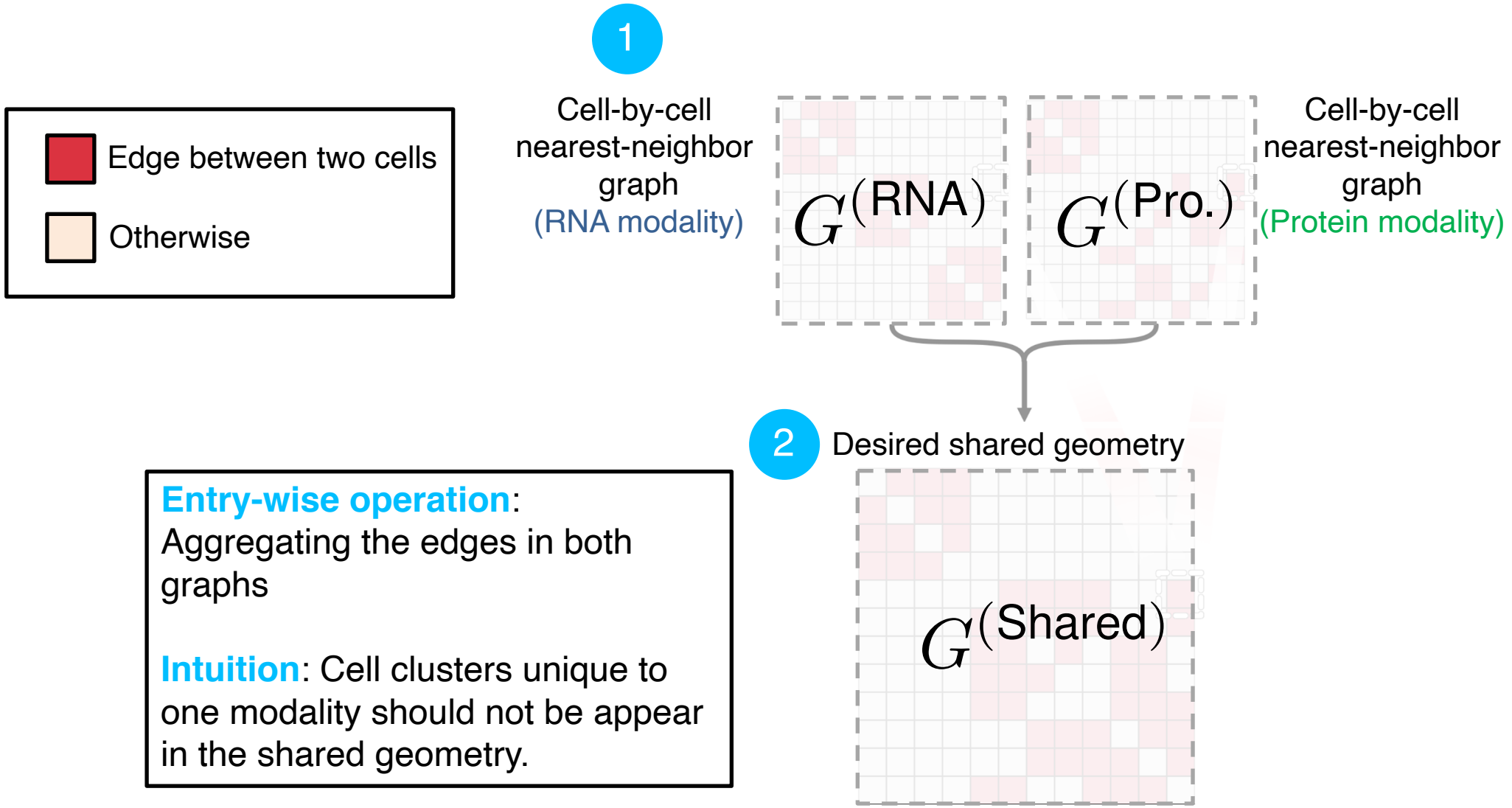


“2 clusters”

Entry-wise operation: Aggregating the edges in both graphs

Intuition: Cell clusters unique to one modality should not be appear in the shared geometry.

(Unsupervised) Estimation: Part 1 – Defining the shared geometry



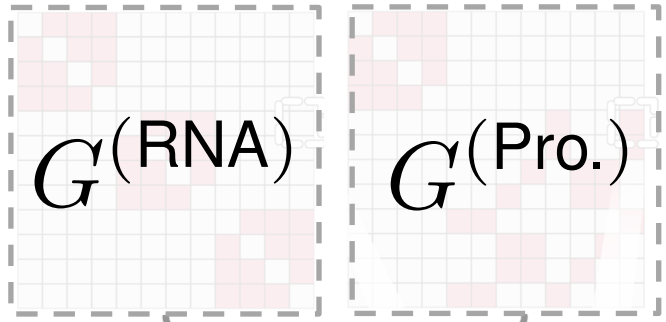
Entry-wise operation: Aggregating the edges in both graphs

Intuition: Cell clusters unique to one modality should not be appear in the shared geometry.

(Unsupervised) Estimation: Part 2 – Estimating the corresponding matrix factorization

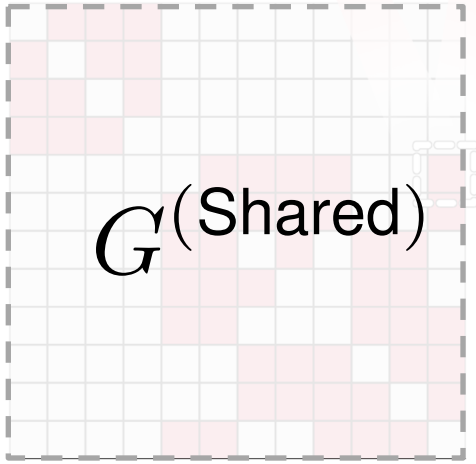
1

Cell-by-cell nearest-neighbor graph (RNA modality)



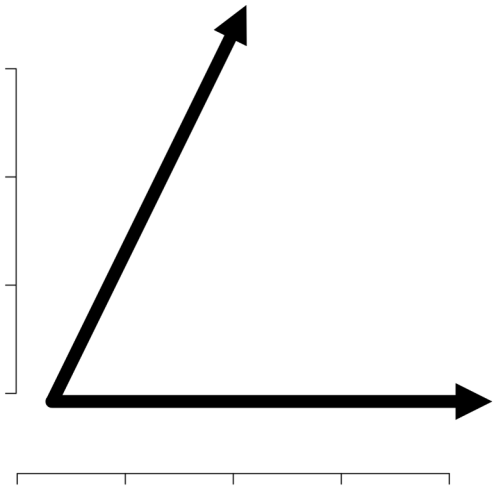
Cell-by-cell nearest-neighbor graph (Protein modality)

2 Desired shared geometry



3

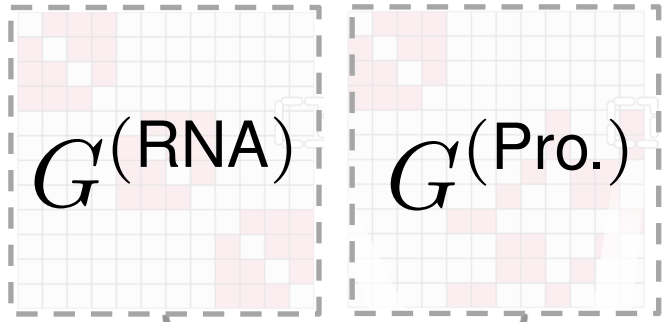
Perform CCA



(Unsupervised) Estimation: Part 2 – Estimating the corresponding matrix factorization

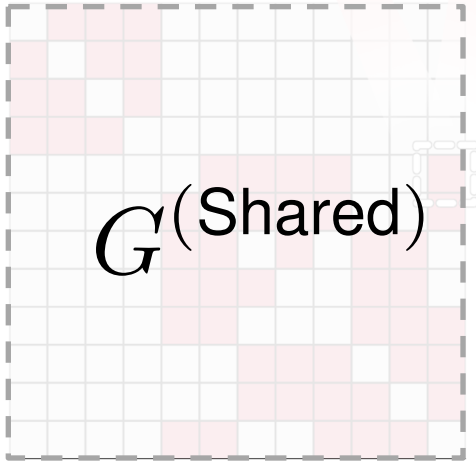
1

Cell-by-cell nearest-neighbor graph (RNA modality)

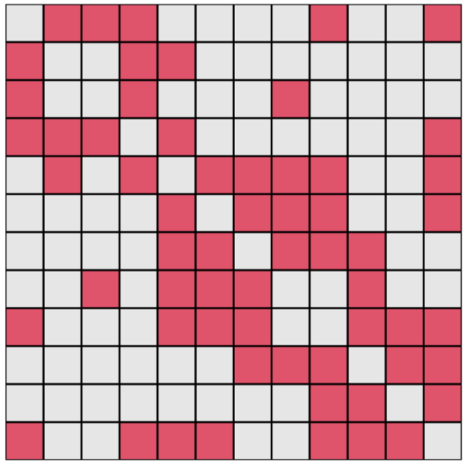


Cell-by-cell nearest-neighbor graph (Protein modality)

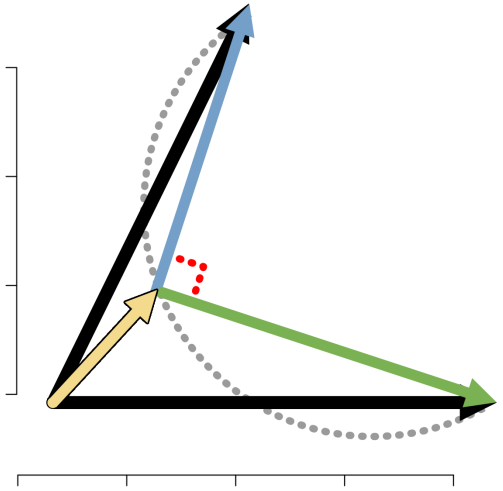
2 Desired shared geometry



… and its corresponding cell-by-cell nearest-neighbor graph for the common embedding



For a latent dim, posit a tilt of the common vector…



3 Perform CCA

(Unsupervised) Estimation: Part 2 – Estimating the corresponding matrix factorization

1

Cell-by-cell nearest-neighbor graph (RNA modality)

$G(\text{RNA})$

$G(\text{Pro.})$

Cell-by-cell nearest-neighbor graph (Protein modality)

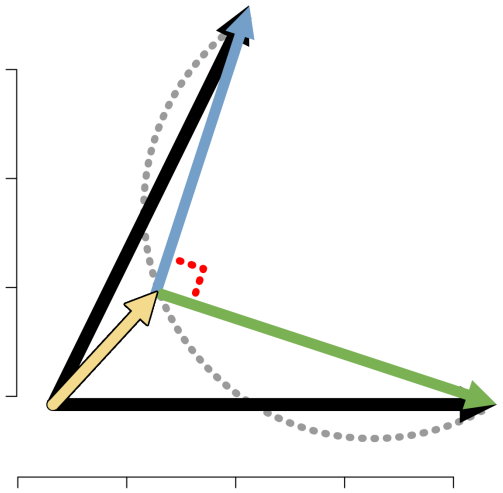
2 Desired shared geometry

$G(\text{Shared})$

… and its corresponding cell-by-cell nearest-neighbor graph for the common embedding

$G(\tau_j)$

For a latent dim, posit a tilt of the common vector…



3

Perform CCA

(Unsupervised) Estimation: Part 2 – Estimating the corresponding matrix factorization

1

Cell-by-cell nearest-neighbor graph (RNA modality)

$G(\text{RNA})$ $G(\text{Pro.})$

Cell-by-cell nearest-neighbor graph (Protein modality)

2 Desired shared geometry

… and its corresponding cell-by-cell nearest-neighbor graph for the common embedding

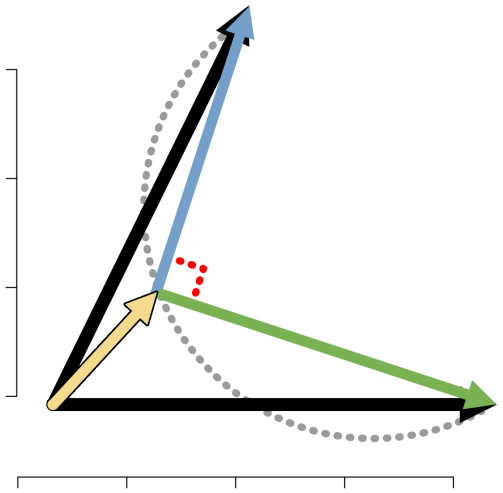
$G(\tau_j)$

Measure difference b/w graphs

(Grassmannian distance b/w Laplacian bases)

$G(\text{Shared})$

For a latent dim, posit a tilt of the common vector…



3 Perform CCA

4 Minimize via 0-order coordinate descent

(Unsupervised) Estimation: Part 2 – Estimating the corresponding matrix factorization

$$X^{(1)} \approx \left[C + D^{(1)}\right] \times L^{(1)}$$

$$X^{(2)} \approx \left[C + D^{(2)}\right] \times L^{(2)}$$

$$C + D^{(1)} = Z^{(1)},\ C + D^{(2)} = Z^{(2)}$$

1

Cell-by-cell nearest-neighbor graph (RNA modality)

$G(\text{RNA})$

$G(\text{Pro.})$

Cell-by-cell nearest-neighbor graph (Protein modality)

2 Desired shared geometry

$G(\text{Shared})$

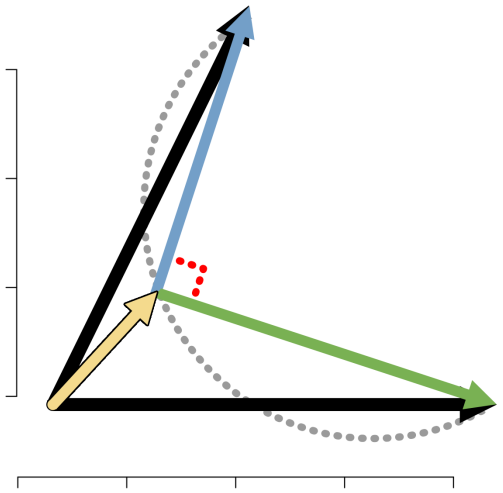
Measure difference b/w graphs

(Grassmannian distance b/w Laplacian bases)

… and its corresponding cell-by-cell nearest-neighbor graph for the common embedding

$G(\tau_j)$

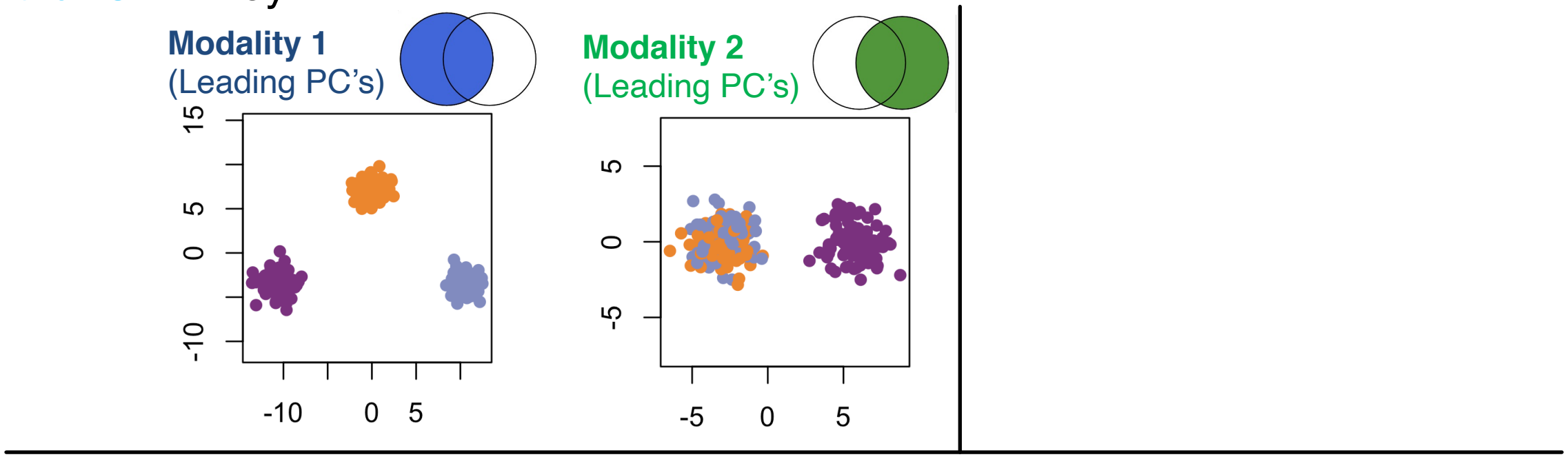
For a latent dim, posit a tilt of the common vector…



3 Perform CCA

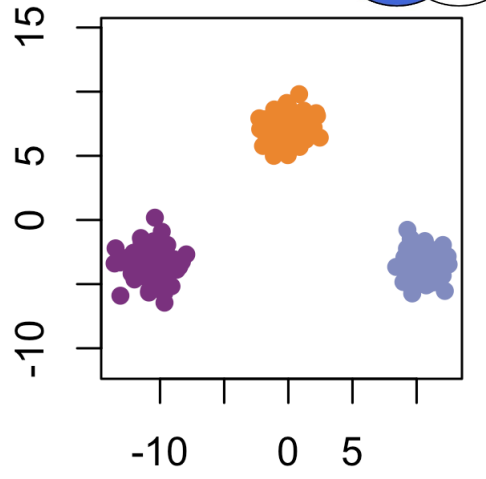
4 Minimize via 0-order coordinate descent

Simulation 1: Toy

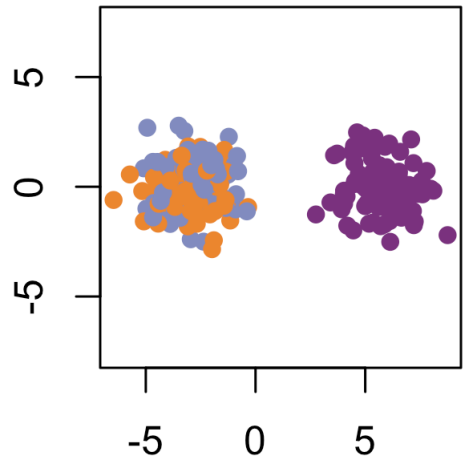


Simulation 1: Toy

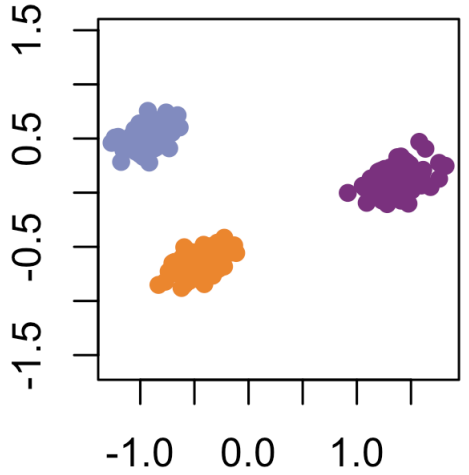
Modality 1
(Leading PC's)



Modality 2
(Leading PC's)

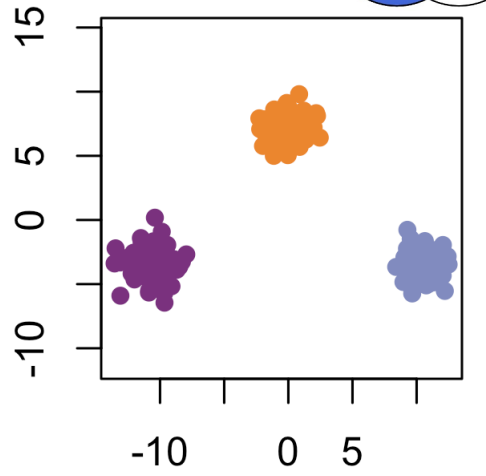


Consensus PCA
(2D: Union)

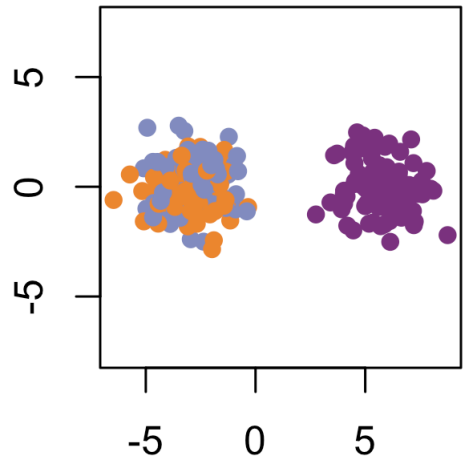


Simulation 1: Toy

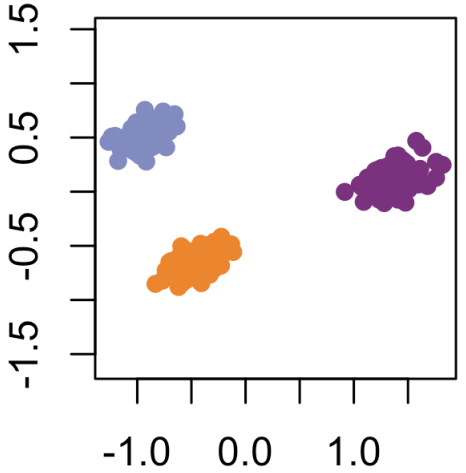
Modality 1
(Leading PC's)



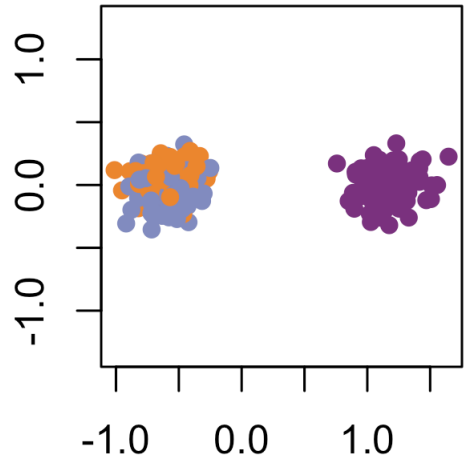
Modality 2
(Leading PC's)



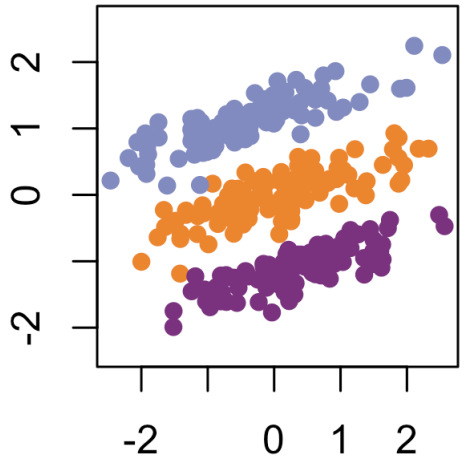
Consensus PCA
(2D: Union)



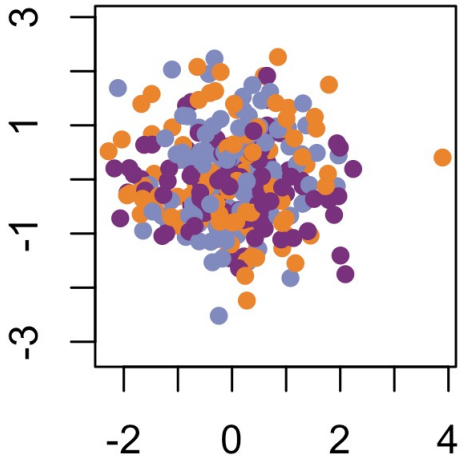
Tilted-CCA: Common
(2D: Intersection)



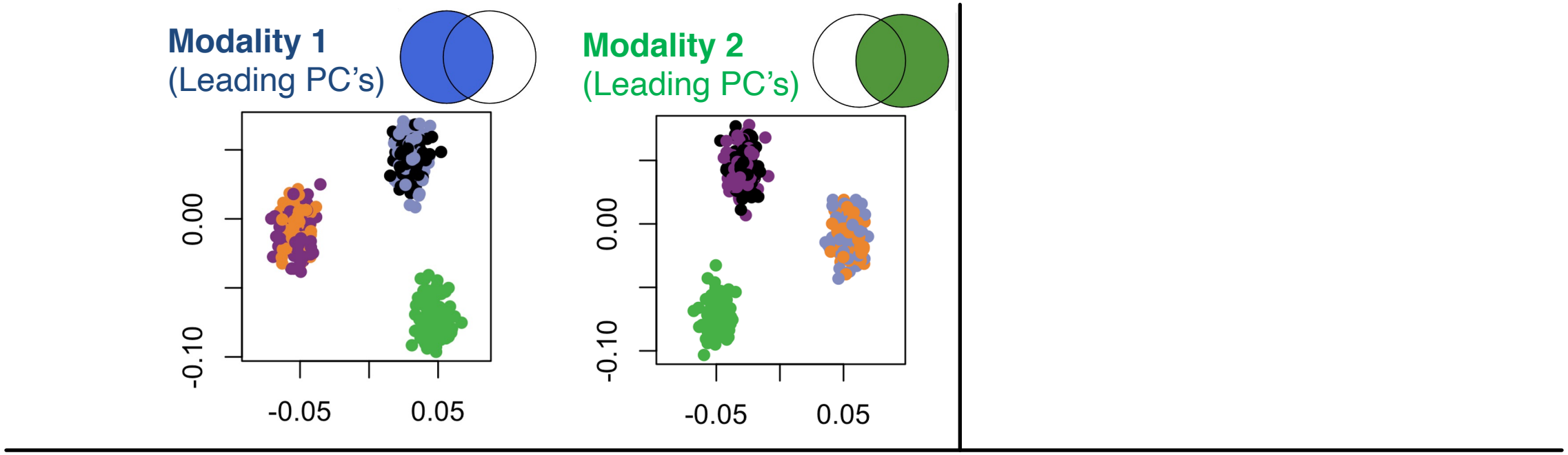
Distinct 1



Distinct 2

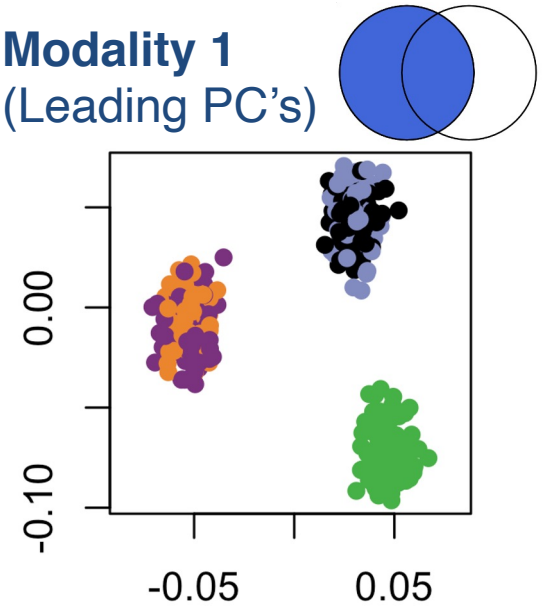


Simulation 2: Criss-cross

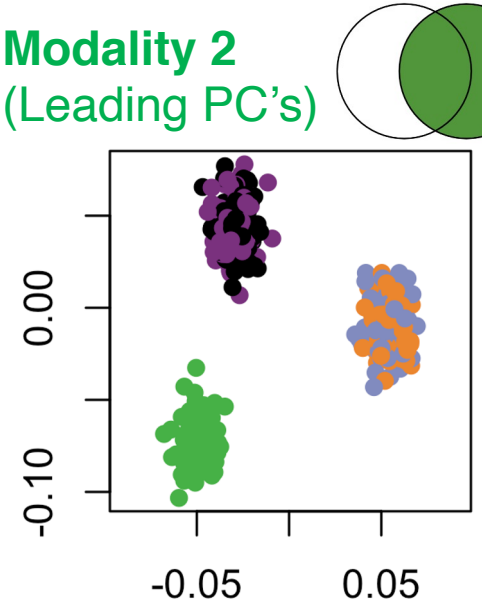


Simulation 2: Criss-cross

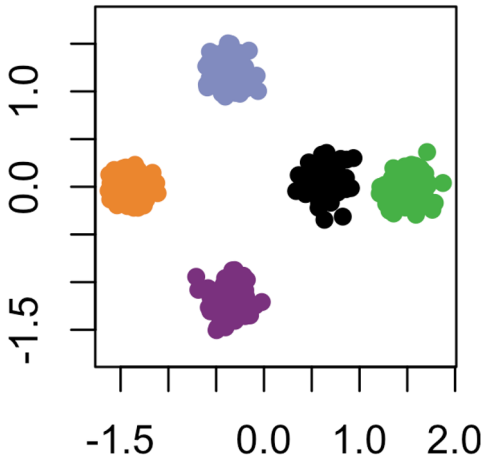
Modality 1
(Leading PC's)



Modality 2
(Leading PC's)

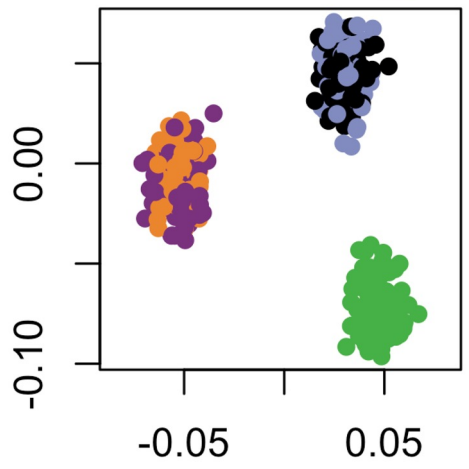


Consensus PCA
(2D: Union)

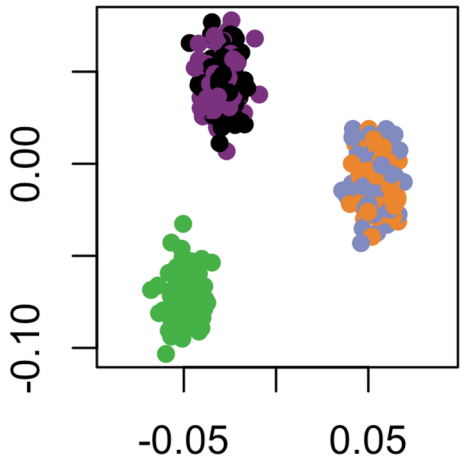


Simulation 2: Criss-cross

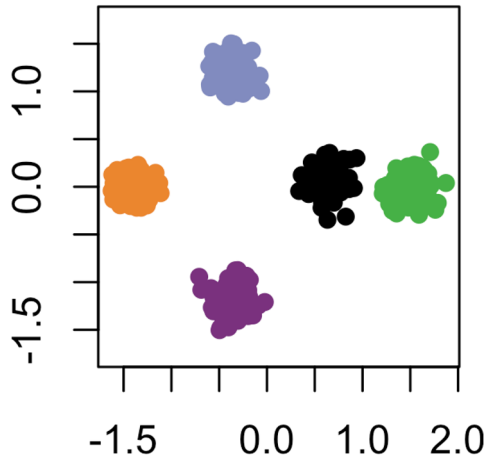
Modality 1
(Leading PC's)



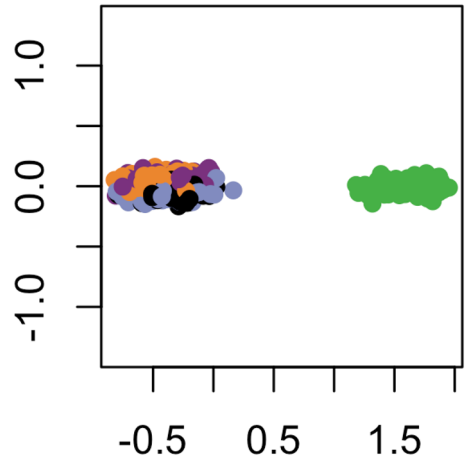
Modality 2
(Leading PC's)



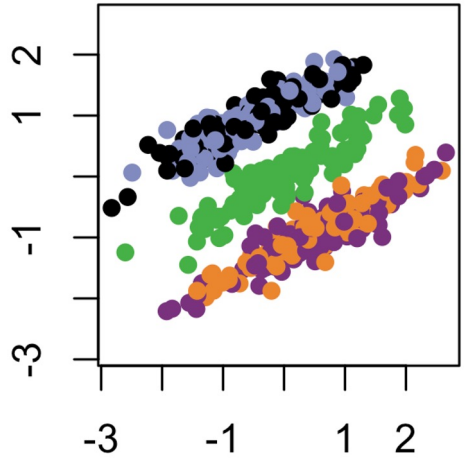
Consensus PCA
(2D: Union)



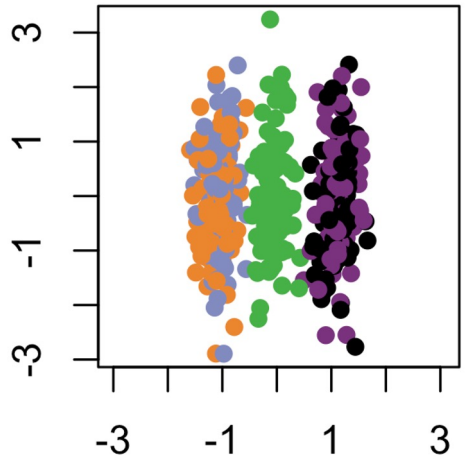
Tilted-CCA: Common
(2D: Intersection)



Distinct 1



Distinct 2



Single-cell investigation:

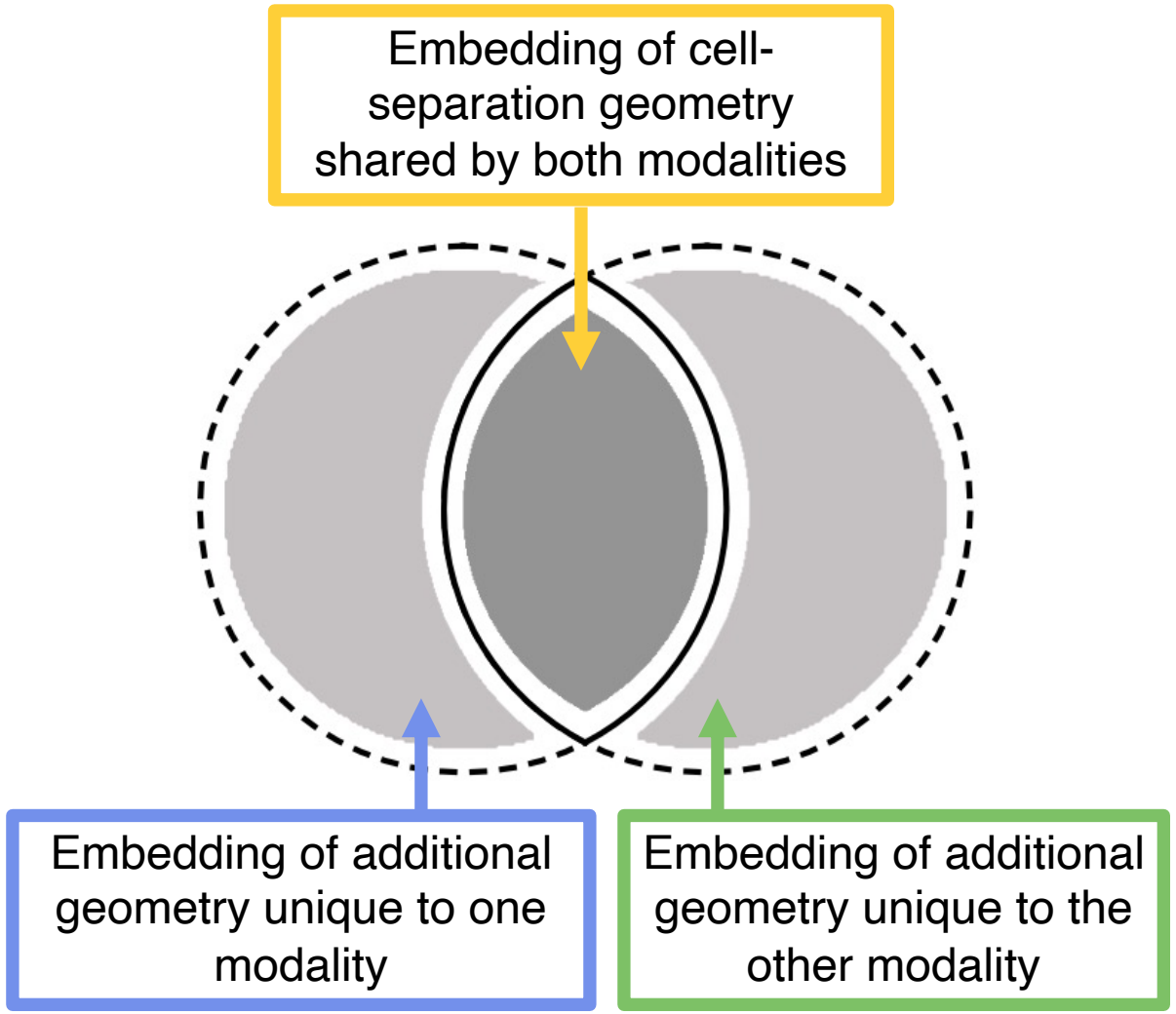
New perspective of developmental biology, now that we have estimated the shared/unique geometry

Recap of the biological goals:

1. **(Experimental design):** Which pair of modalities should biologist sequence to have the most comprehensive understanding?

2. **(Variable selection):** For RNA-Protein data, how can we pick the antibodies that contribute the most additional information to the RNA modality?

3. **(Developmental biology):** Can the amount of coordination between two modalities tell us if a cell in a steady-state or is undergoing development?



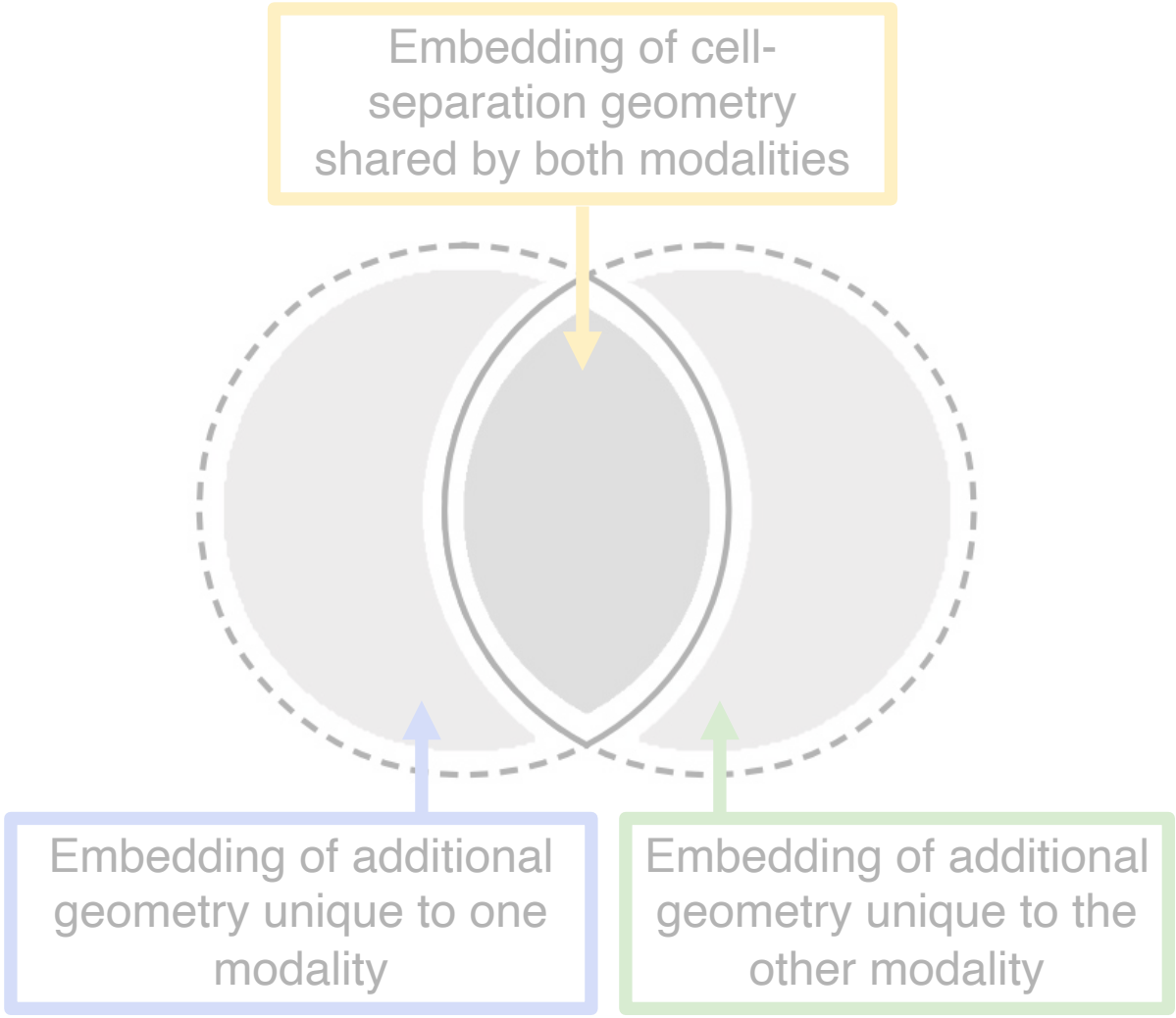
Recap of the biological goals:

1. **(Experimental design):** Which pair of modalities should biologist sequence to have the most comprehensive understanding?

2. **(Variable selection):** For RNA-Protein data, how can we pick the antibodies that contribute the most additional information to the RNA modality?

3. **(Developmental biology):** Can the amount of coordination between two modalities tell us if a cell in a steady-state or is undergoing development?

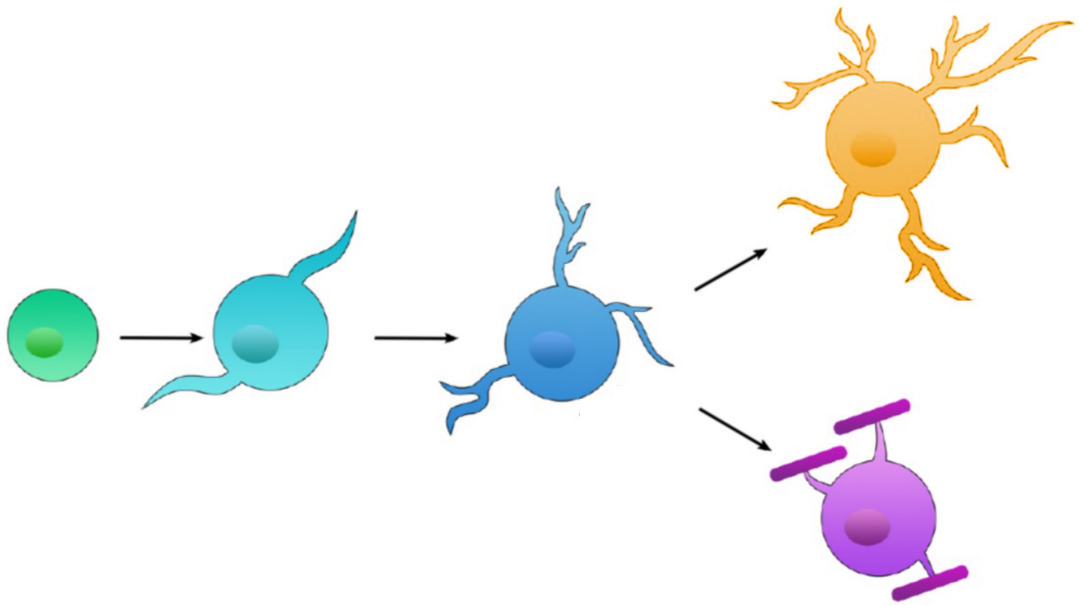
Focus of the remainder of the talk

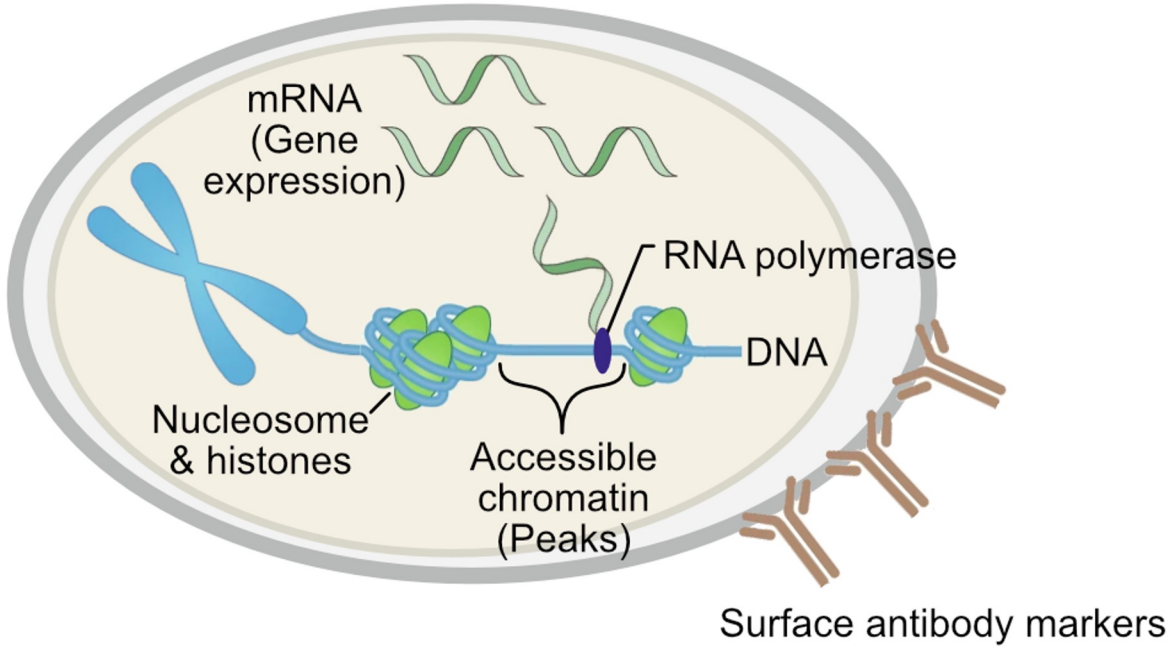


In developmental biology, we're interested in studying how cells continually specialize over time despite given static snapshots.

Progenitor cells
(Youngest)

Mature neurons
(Oldest)





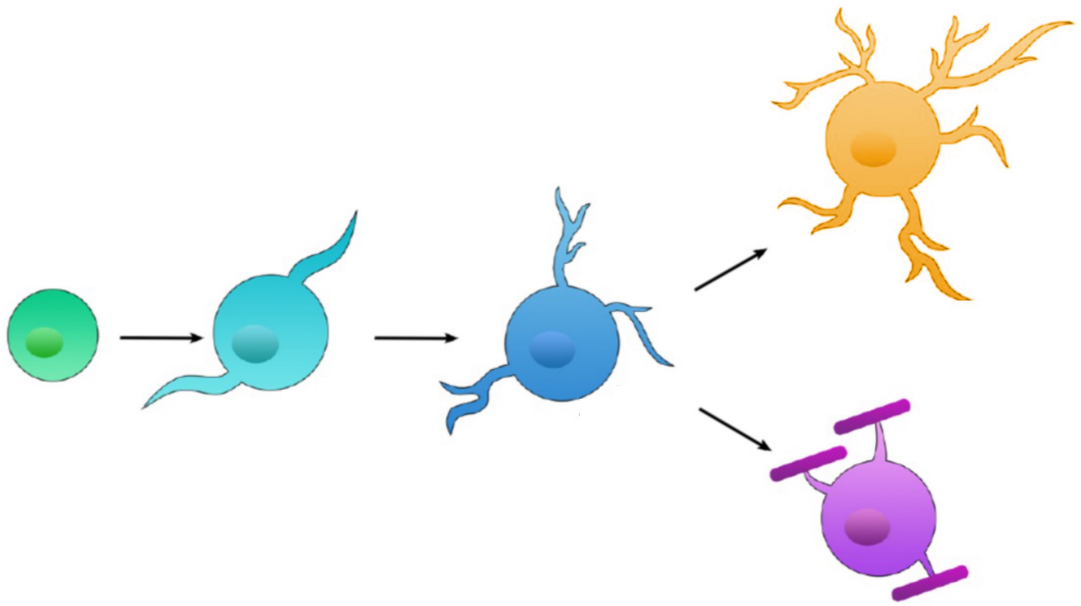
Existing methods:

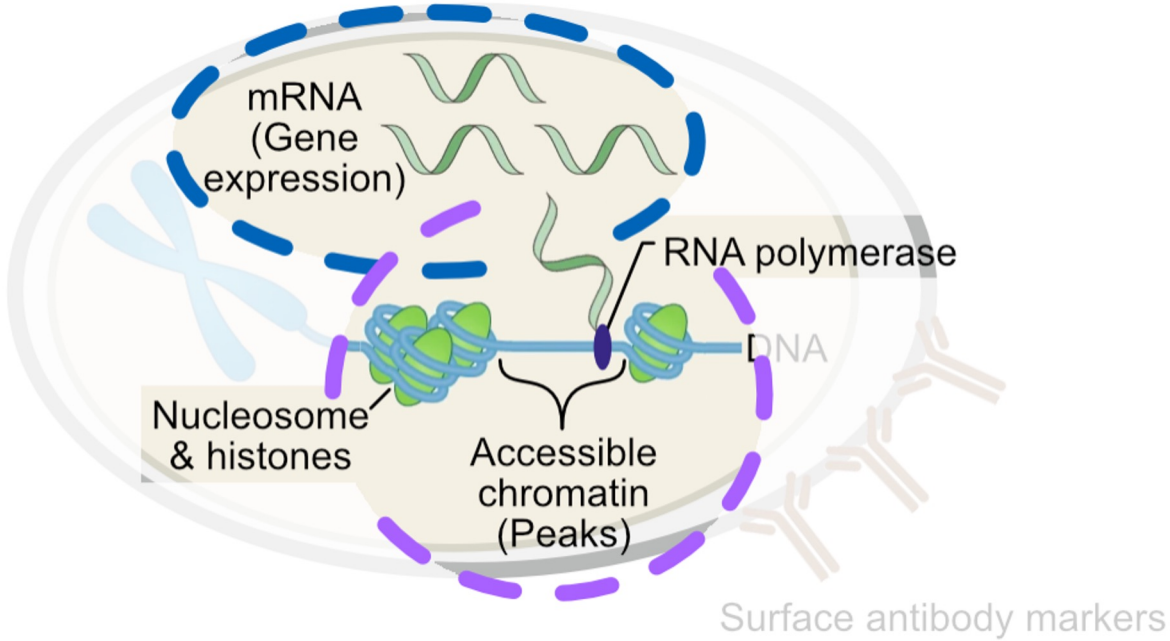
Monocle (Trapnell et al., Nature Biotech., 2014),
Slingshot (Dudoit et al., BMC Genomics, 2018),
RNA velocity (Kharchenko et al., Nature, 2018)

In developmental biology, we're interested in studying how cells continually specialize over time despite given static snapshots.

Progenitor cells
(Youngest)

Mature neurons
(Oldest)





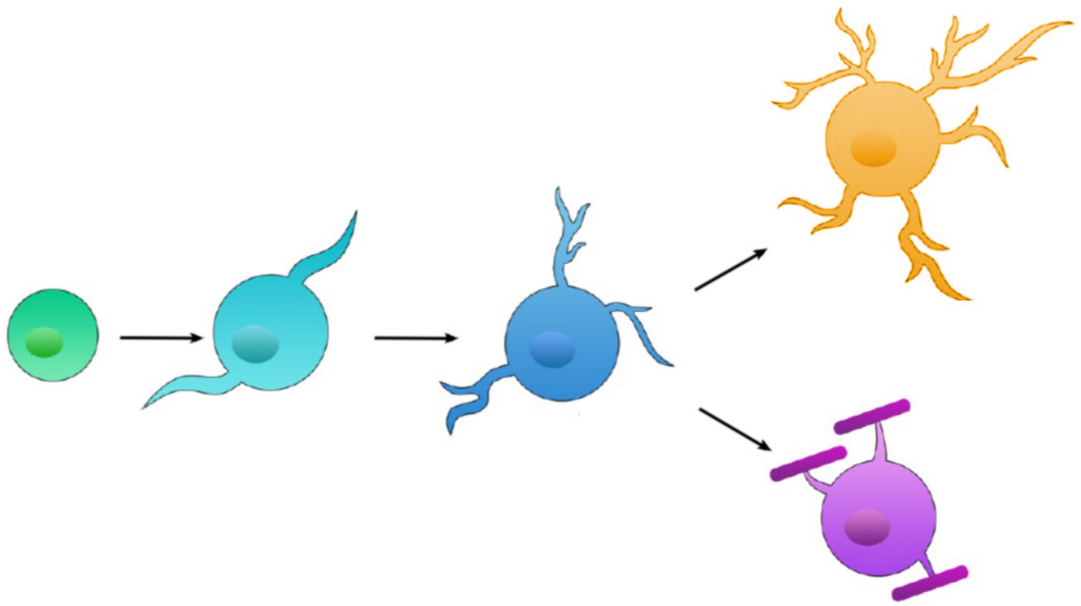
Existing methods:

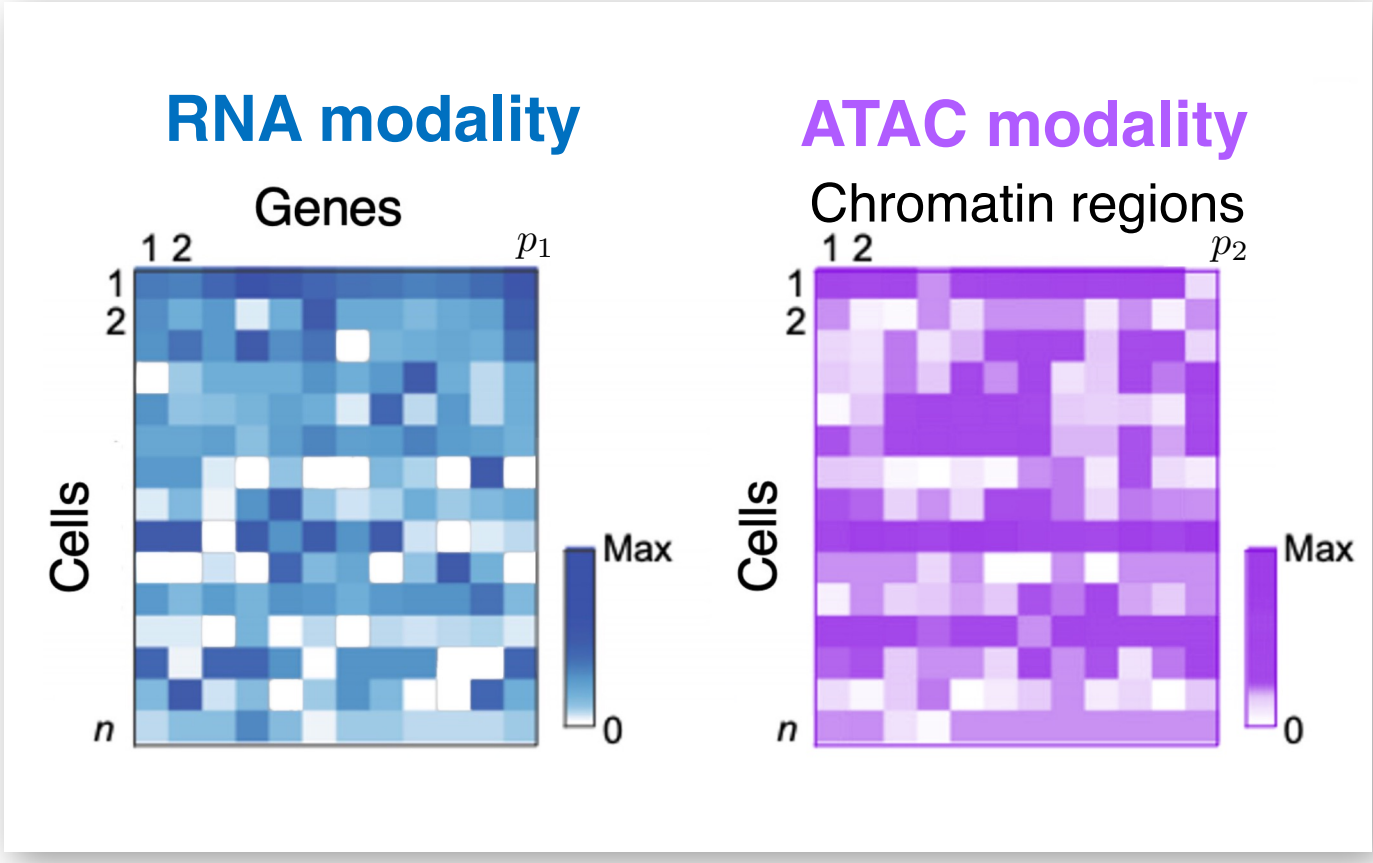
Monocle (Trapnell et al., Nature Biotech., 2014),
Slingshot (Dudoit et al., BMC Genomics, 2018),
RNA velocity (Kharchenko et al., Nature, 2018)

In developmental biology, we're interested in studying how cells continually specialize over time despite given static snapshots.

Progenitor cells (Youngest)

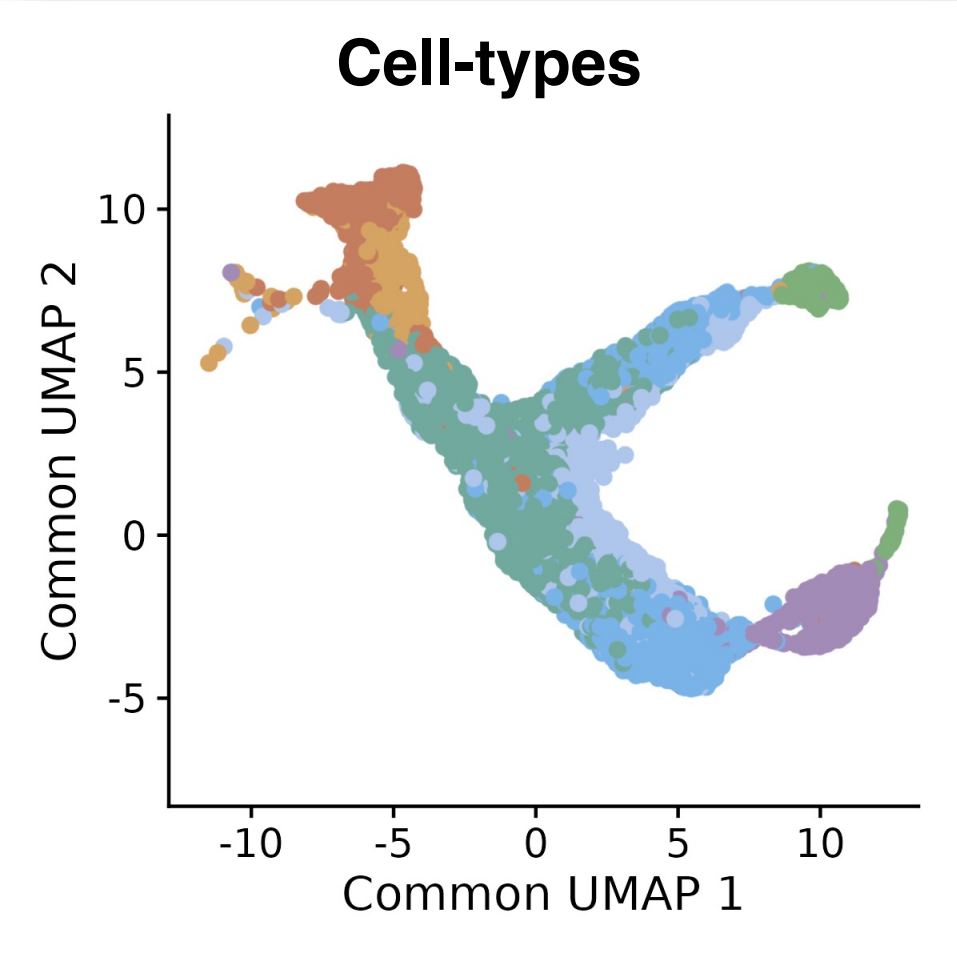
Mature neurons (Oldest)





While existing methods focusing on one modality tell us **how** cells develop ("layout of the highway"), we want to know **what** status the cell is in ("speed of traffic").

Knowing the developmental trajectories is **not the same** as knowing if a cell undergoing development or is in steady-state.



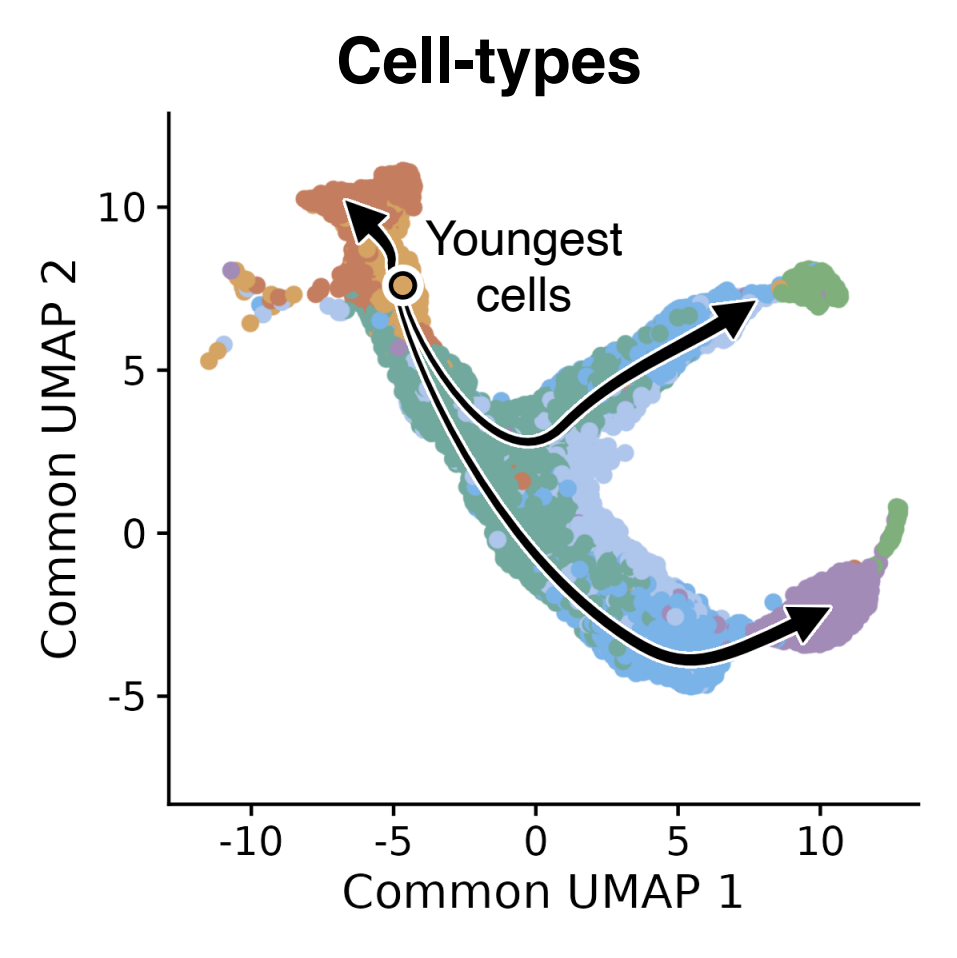
Cycling progenitors (Youngest cells)

Radial glia

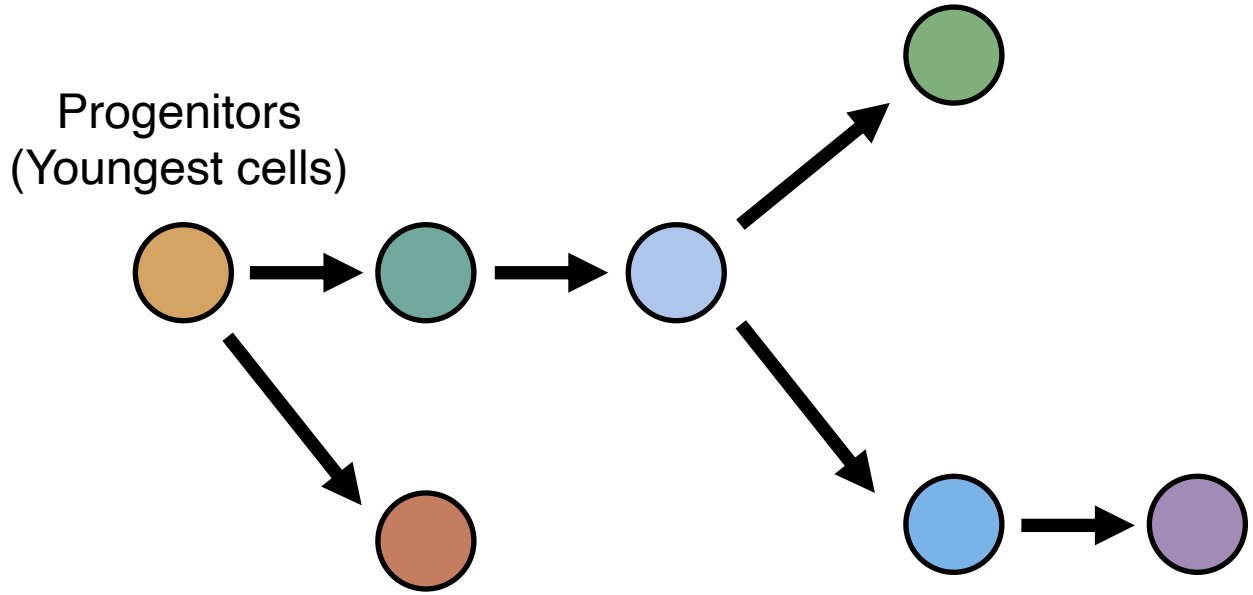
Different types of glutamatergic neurons

Human brain development (10x. RNA & ATAC. Greenleaf et al., Cell, 2021): 6000+ cells

Knowing the developmental trajectories is **not the same** as knowing if a cell undergoing development or is in steady-state.

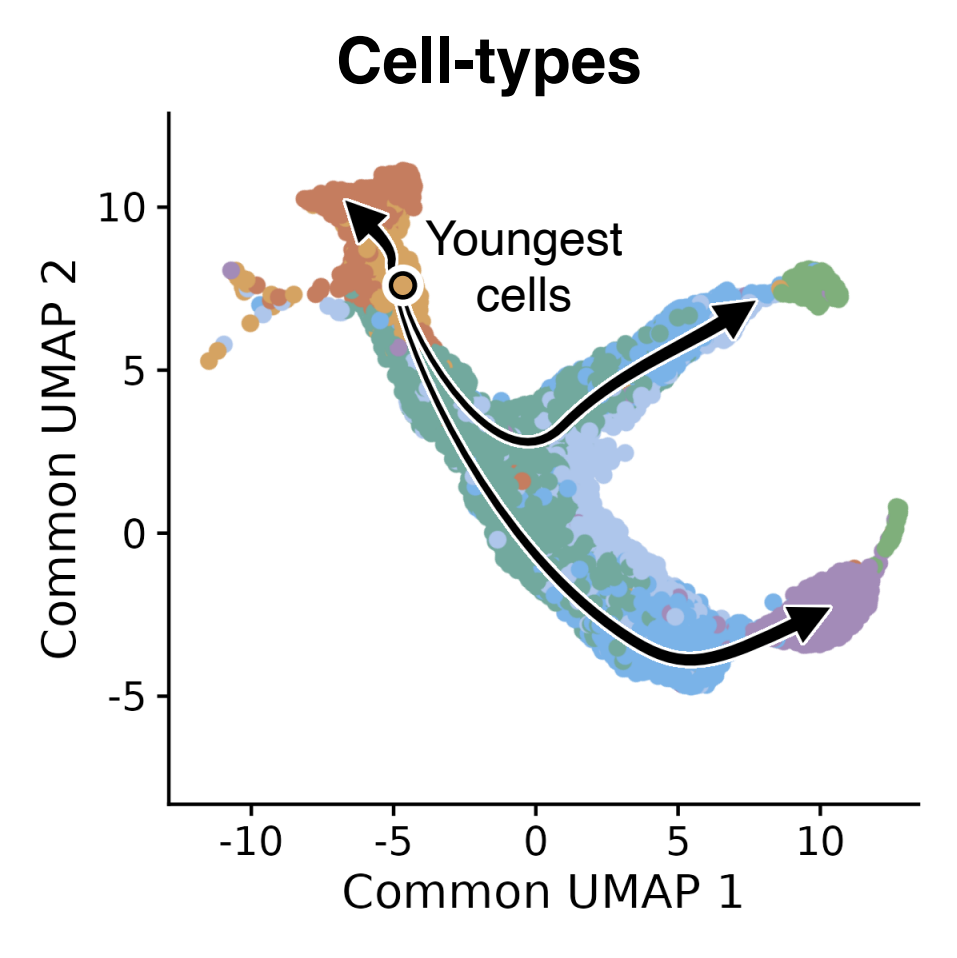


Human brain development (10x. RNA & ATAC. Greenleaf et al., Cell, 2021): 6000+ cells

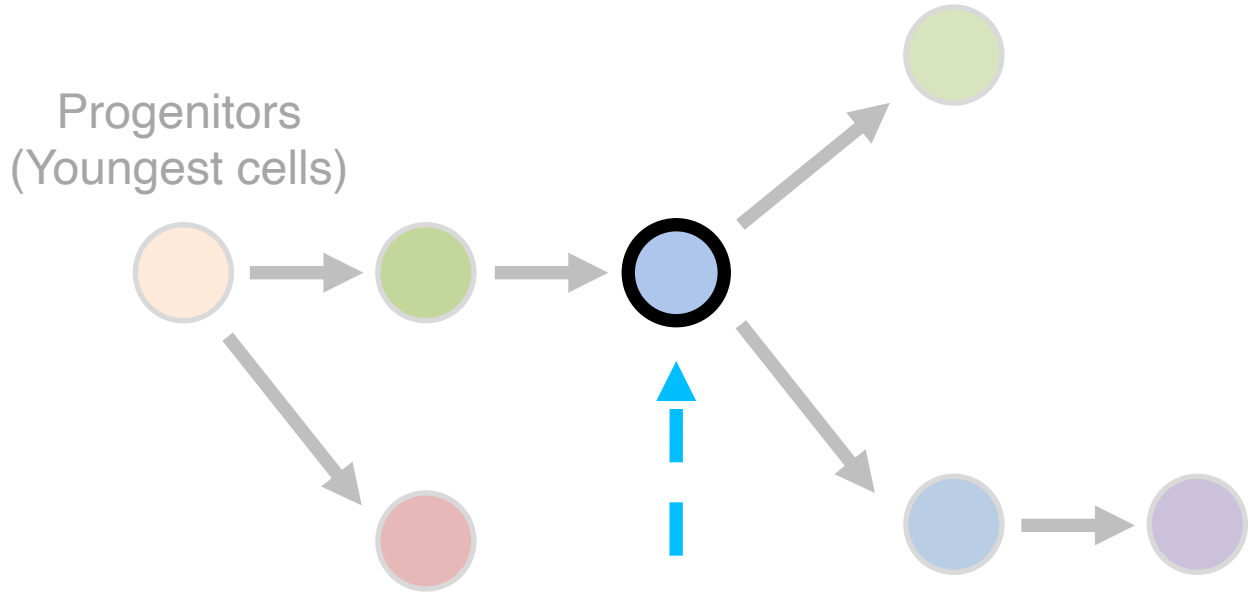


Existing methods:
"Layout of the highway"

Knowing the developmental trajectories is **not the same** as knowing if a cell undergoing development or is in steady-state.

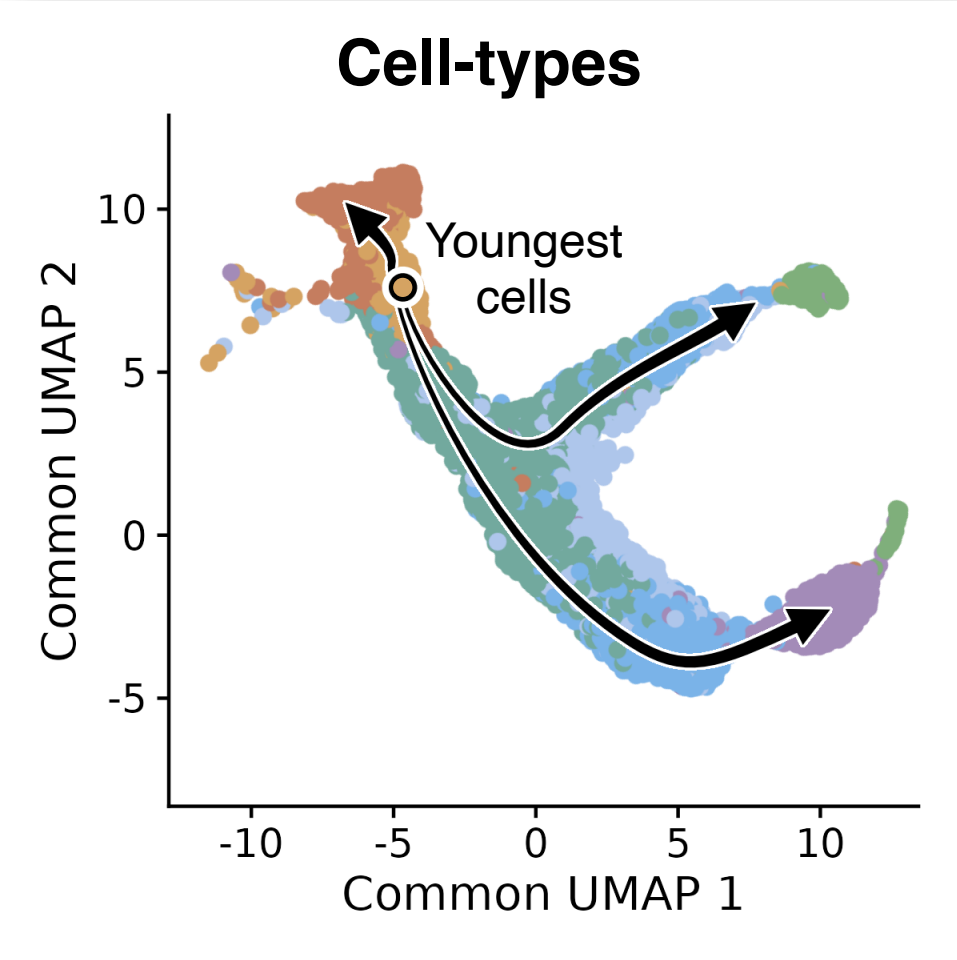


Human brain development (10x. RNA & ATAC. Greenleaf et al., Cell, 2021): 6000+ cells

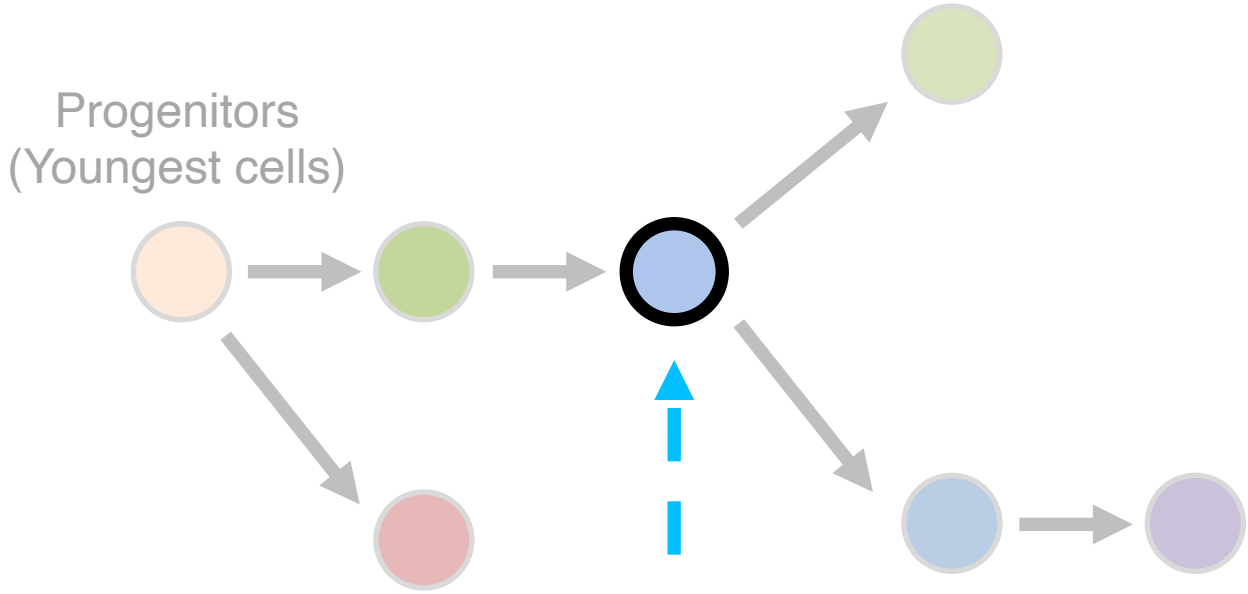


"Speed of traffic":
Are the neurons rapidly passing through **OR** idling around/waiting for an external stimulus?

Knowing the developmental trajectories is **not the same** as knowing if a cell undergoing development or is in steady-state.

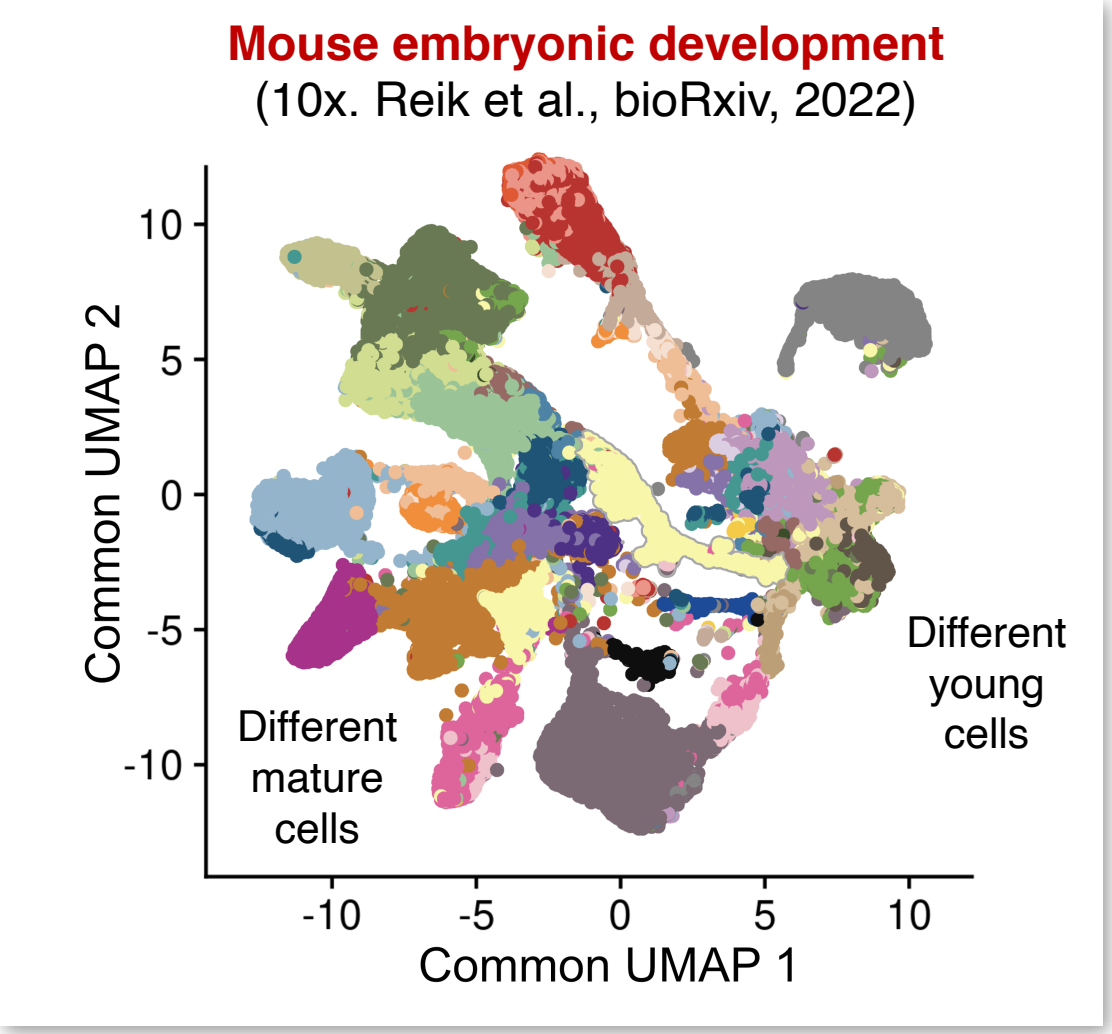


Human brain development (10x. RNA & ATAC. Greenleaf et al., Cell, 2021): 6000+ cells



"Speed of traffic":
Are the neurons in-development **OR** in steady-state?

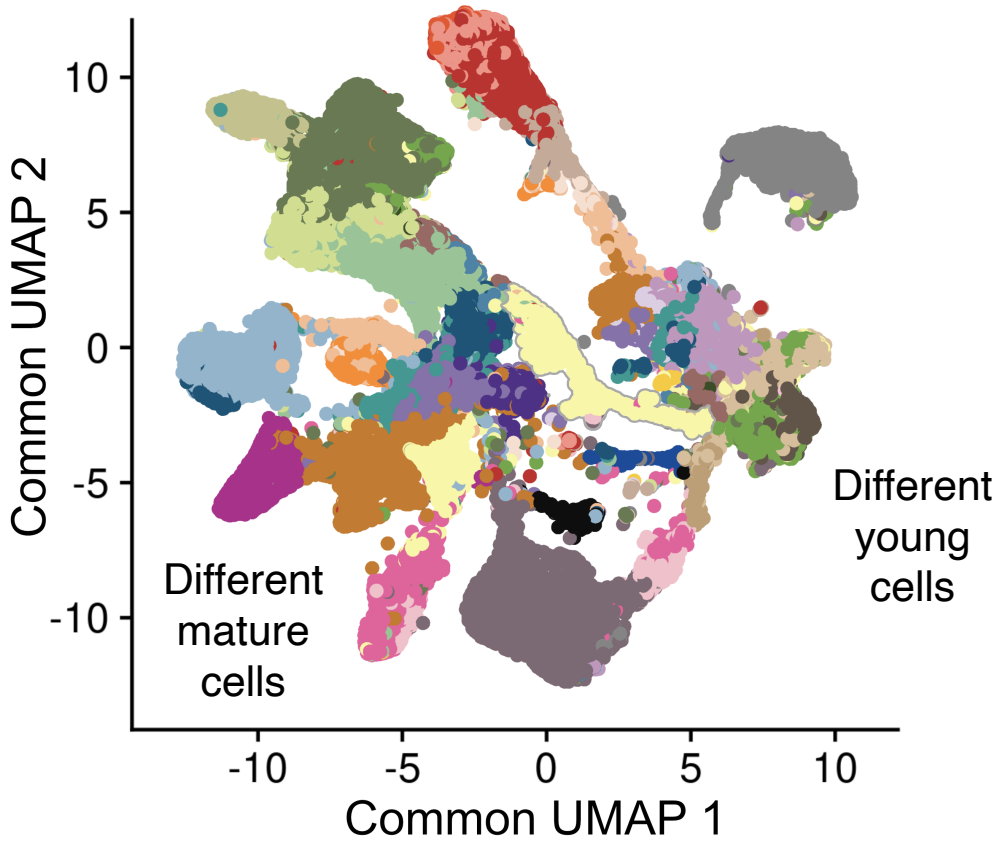
Brief aside: More complex systems



Difficult due to multiple trajectories

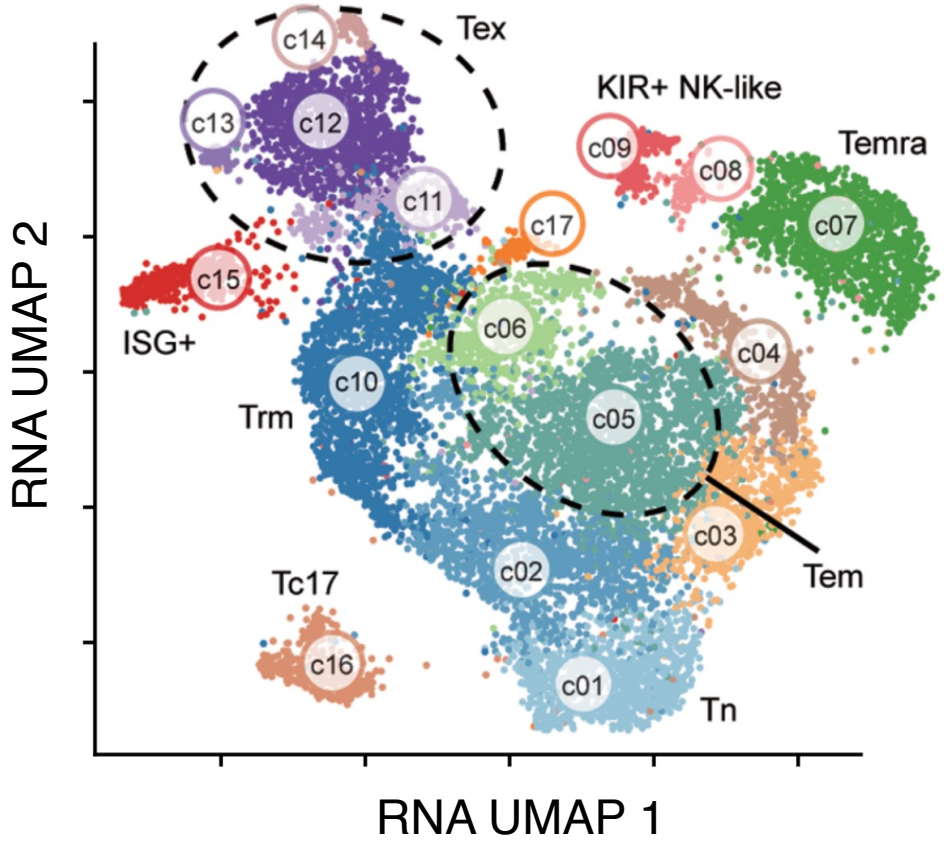
Brief aside: More complex systems

Mouse embryonic development
(10x. Reik et al., bioRxiv, 2022)



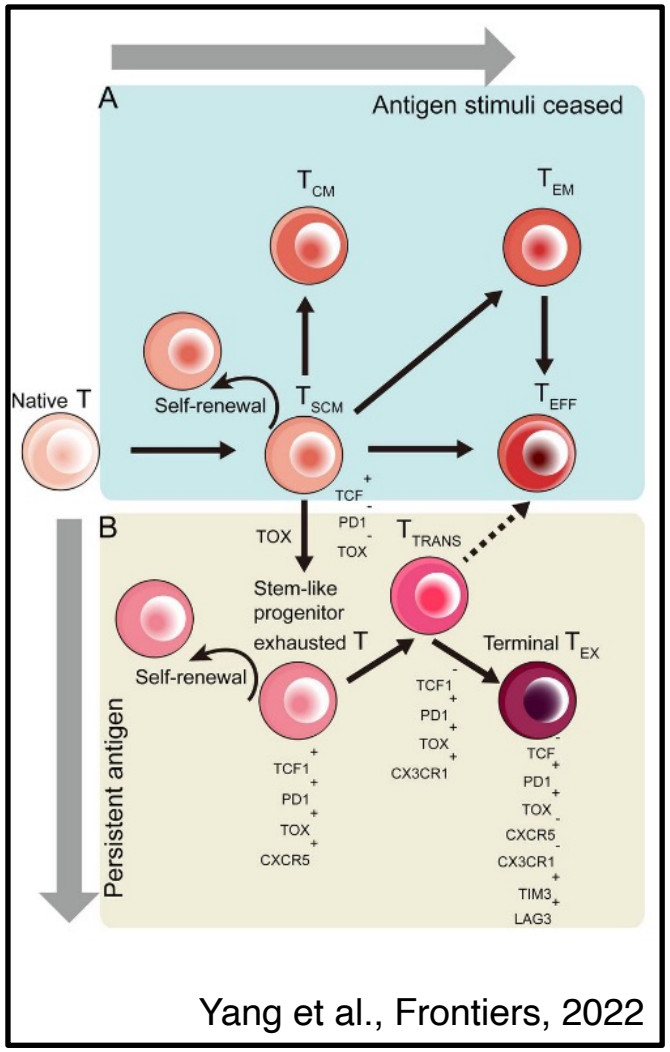
Difficult due to multiple trajectories

T-cell exhaustion
(Zhang et al., Science, 2021)

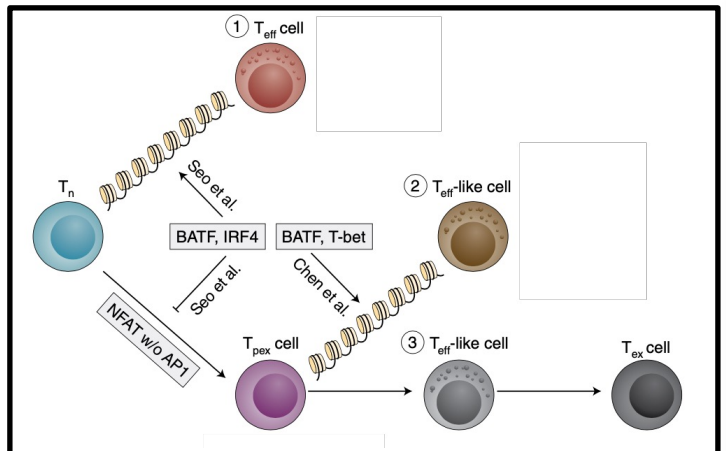


Even more difficult since the biology is still anyone's guess

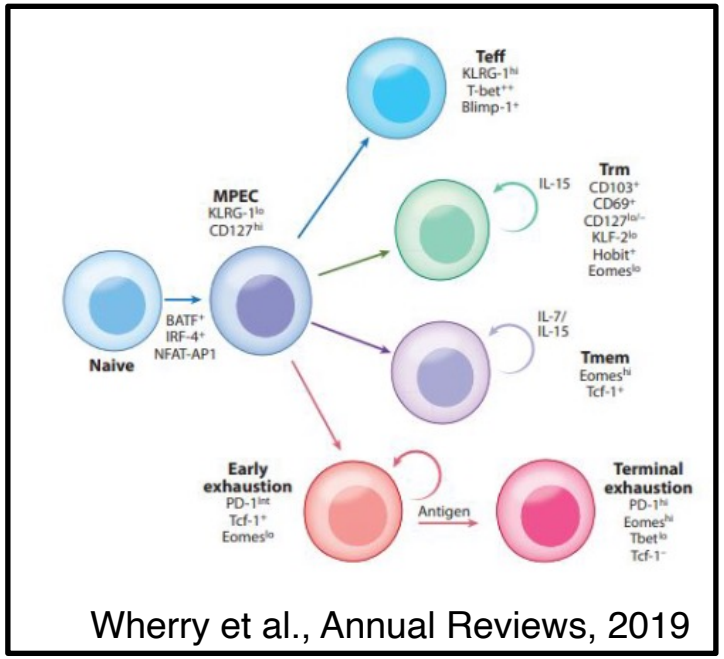
Brief aside: More complex systems



Yang et al., Frontiers, 2022

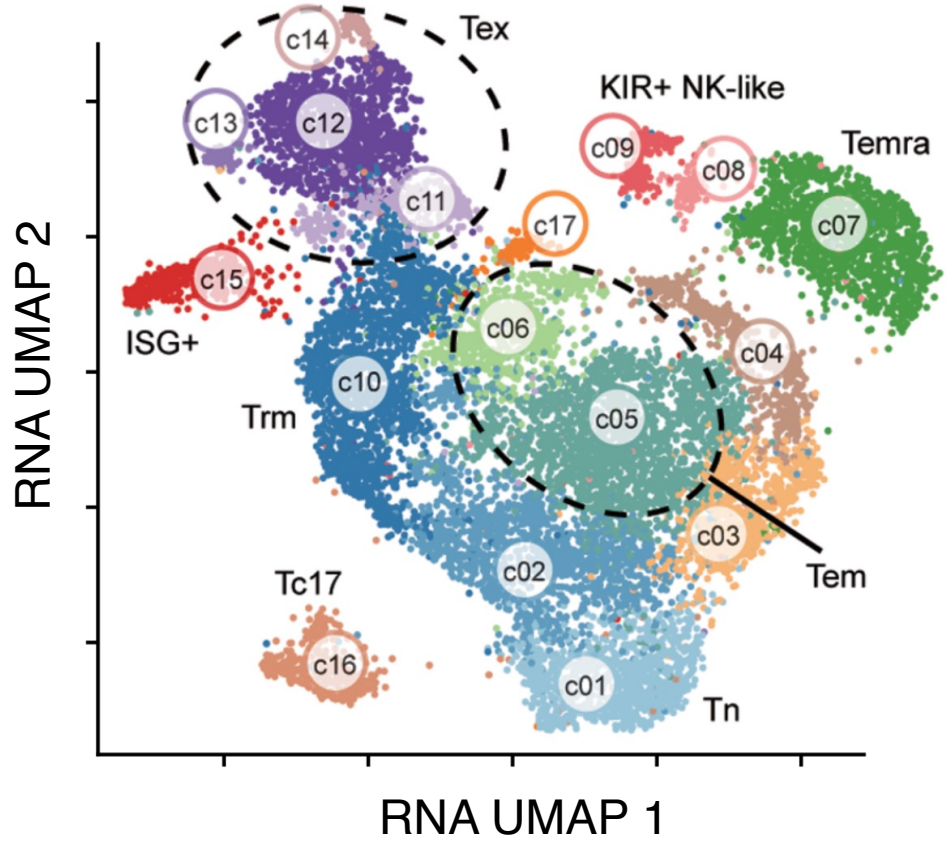


Youngblood et al., Nature Imm., 2021



Wherry et al., Annual Reviews, 2019

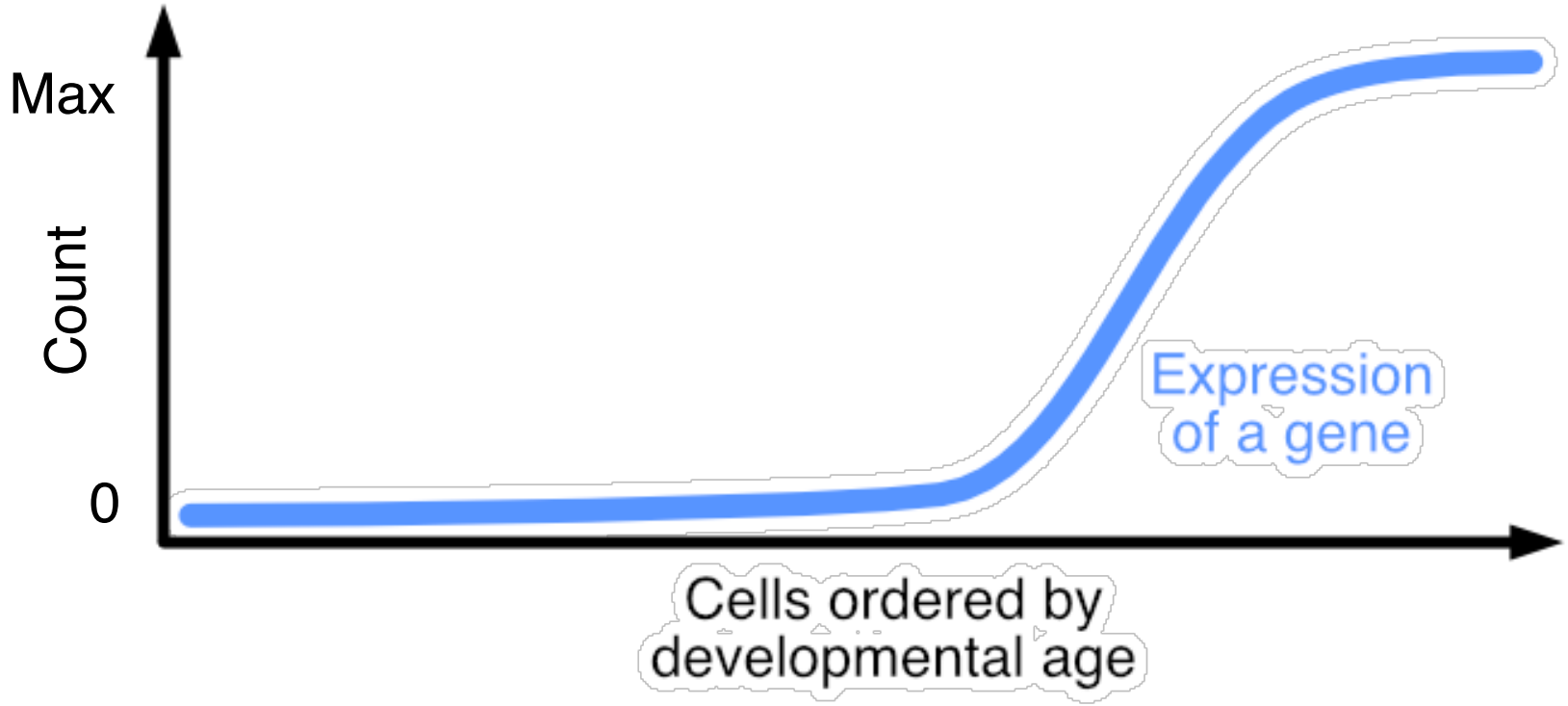
T-cell exhaustion
(Zhang et al., Science, 2021)



Even more difficult since the biology is still anyone's guess

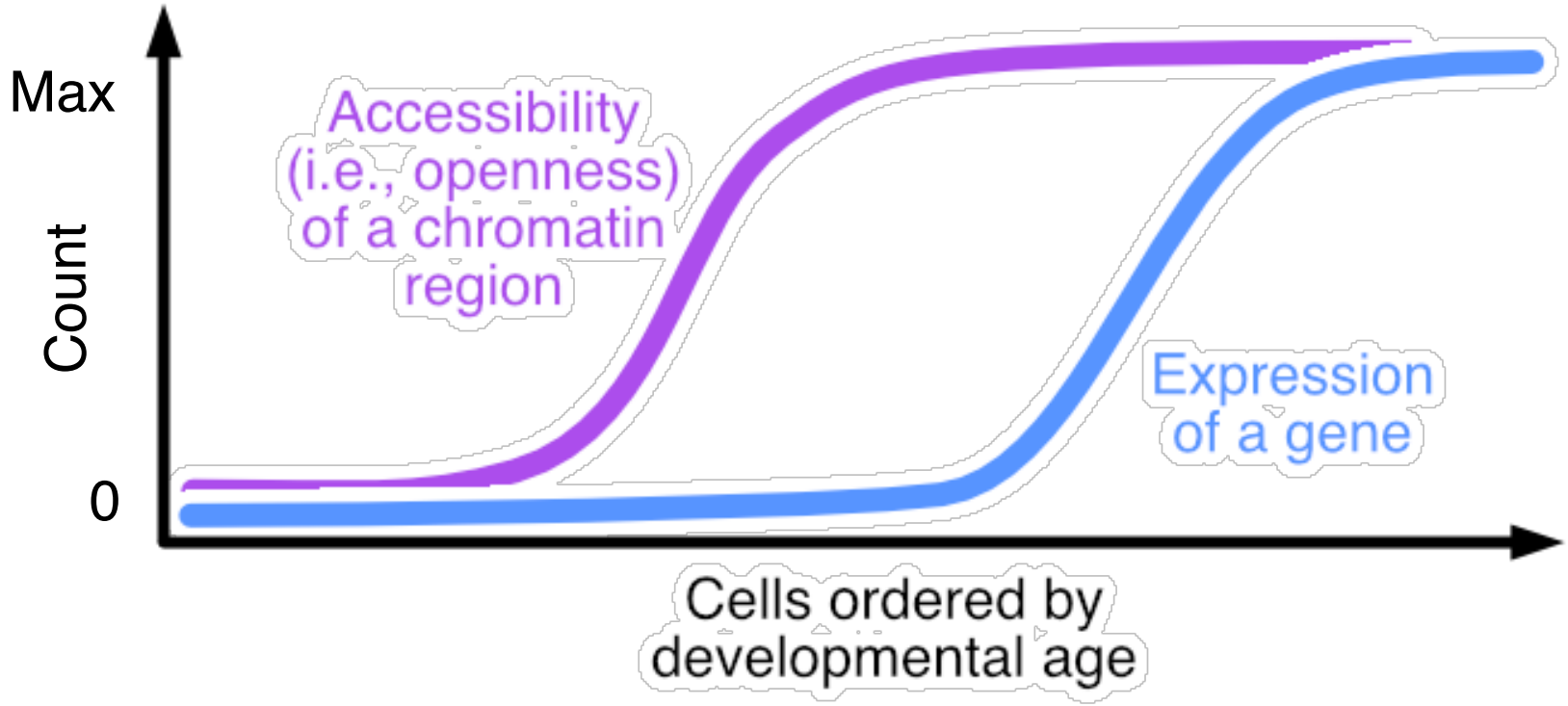
Tilted-CCA can learn which cells are in steady-state via RNA-ATAC relations thanks to principles discovered in previous work.

Schematic from SHARE-seq (Buenrostro et al., Cell, 2020)



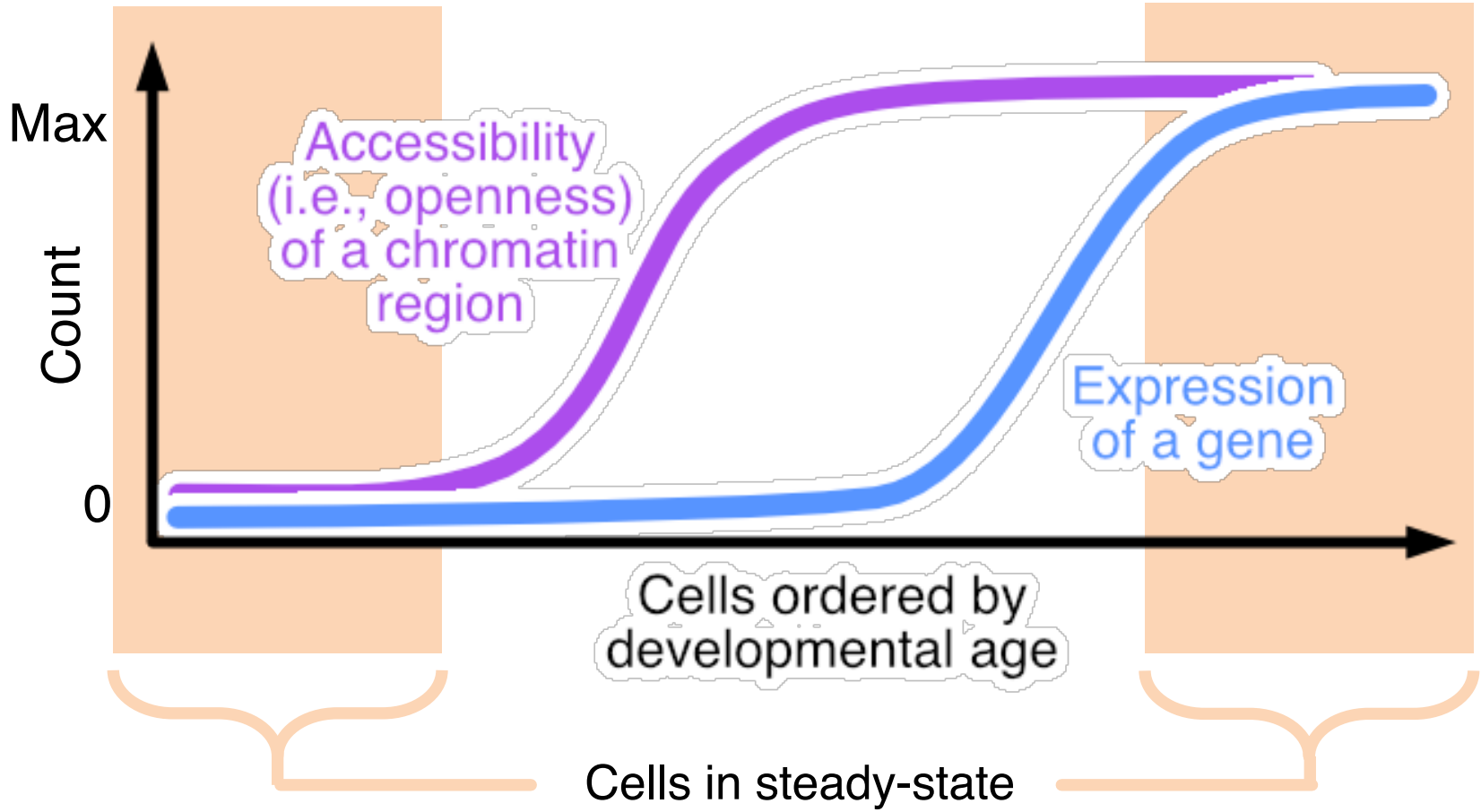
Tilted-CCA can learn which cells are in steady-state via RNA-ATAC relations thanks to principles discovered in previous work.

Schematic from SHARE-seq (Buenrostro et al., Cell, 2020)



Tilted-CCA can learn which cells are in steady-state via RNA-ATAC relations thanks to principles discovered in previous work.

Schematic from SHARE-seq (Buenrostro et al., Cell, 2020)



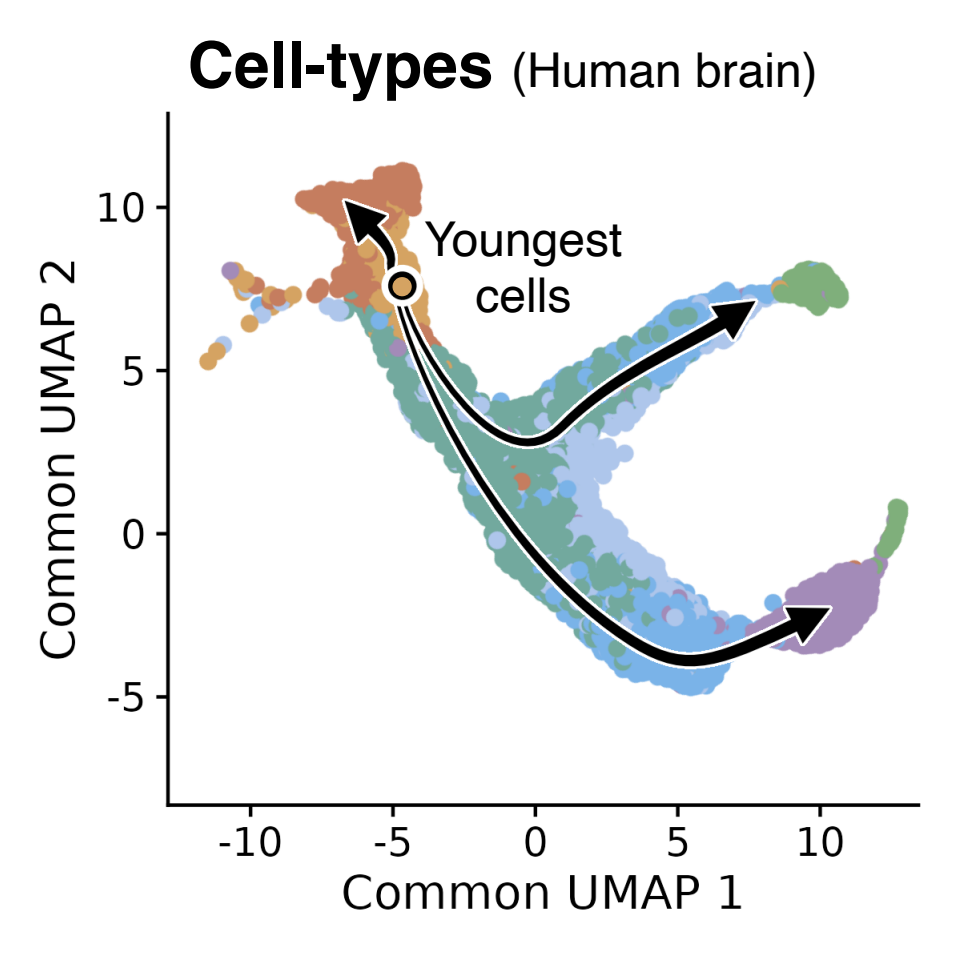
We want to learn which cells are in steady-state (w/o temporal ordering):
Local geometries similar in both modalities → Small distinct components

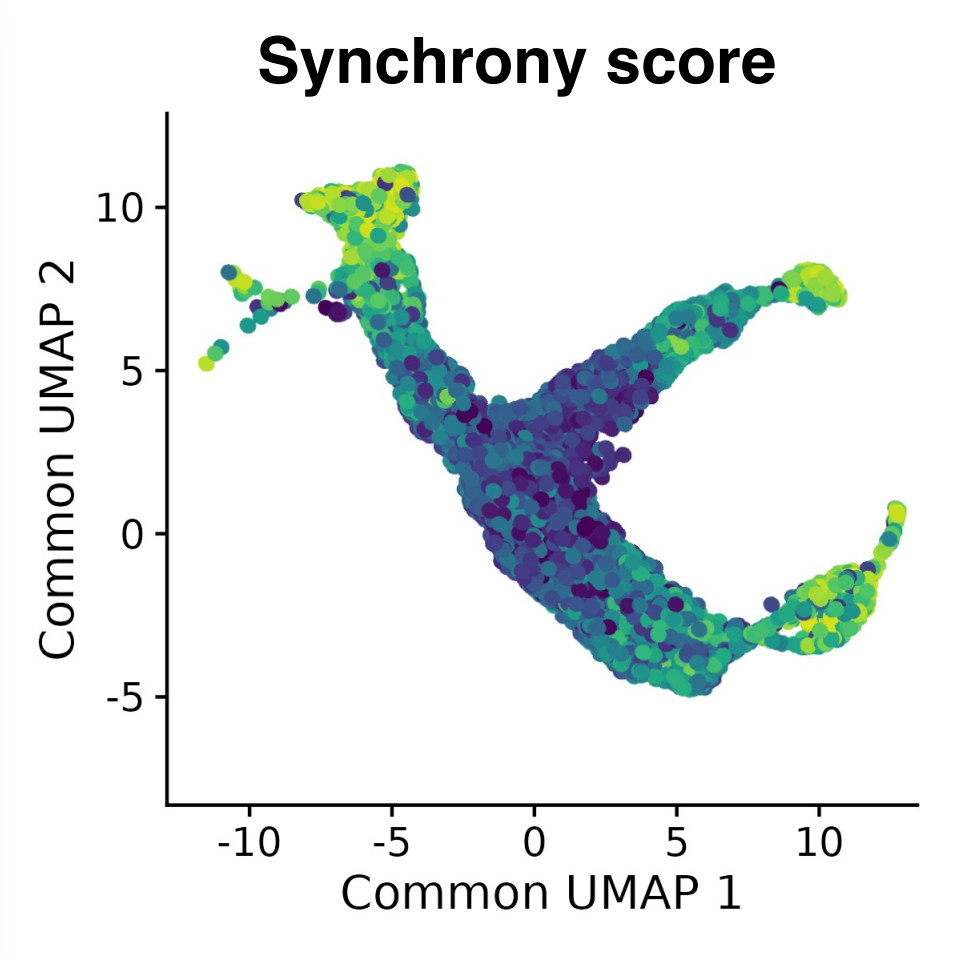
Our contribution: The **synchrony score**, which measures the coordination b/w genes and chromatin accessibility within a cell, to reveal if a cell is in steady-state

After fitting Tilted-CCA, for each cell i: Correlation $\left(\begin{array}{c}\text{common}\\ \text{component}\\ \text{among genes}\end{array}, \begin{array}{c}\text{common + distinct}\\ \text{components among}\\ \text{genes}\end{array}\right)$

Our contribution: The **synchrony score**, which measures the coordination b/w genes and chromatin accessibility within a cell, to reveal if a cell is in steady-state

After fitting Tilted-CCA, for each cell i: Correlation $\left(\begin{array}{c} \text{common} \\ \text{component} \\ \text{among genes} \end{array} , \begin{array}{c} \text{common + distinct} \\ \text{components among} \\ \text{genes} \end{array} \right)$



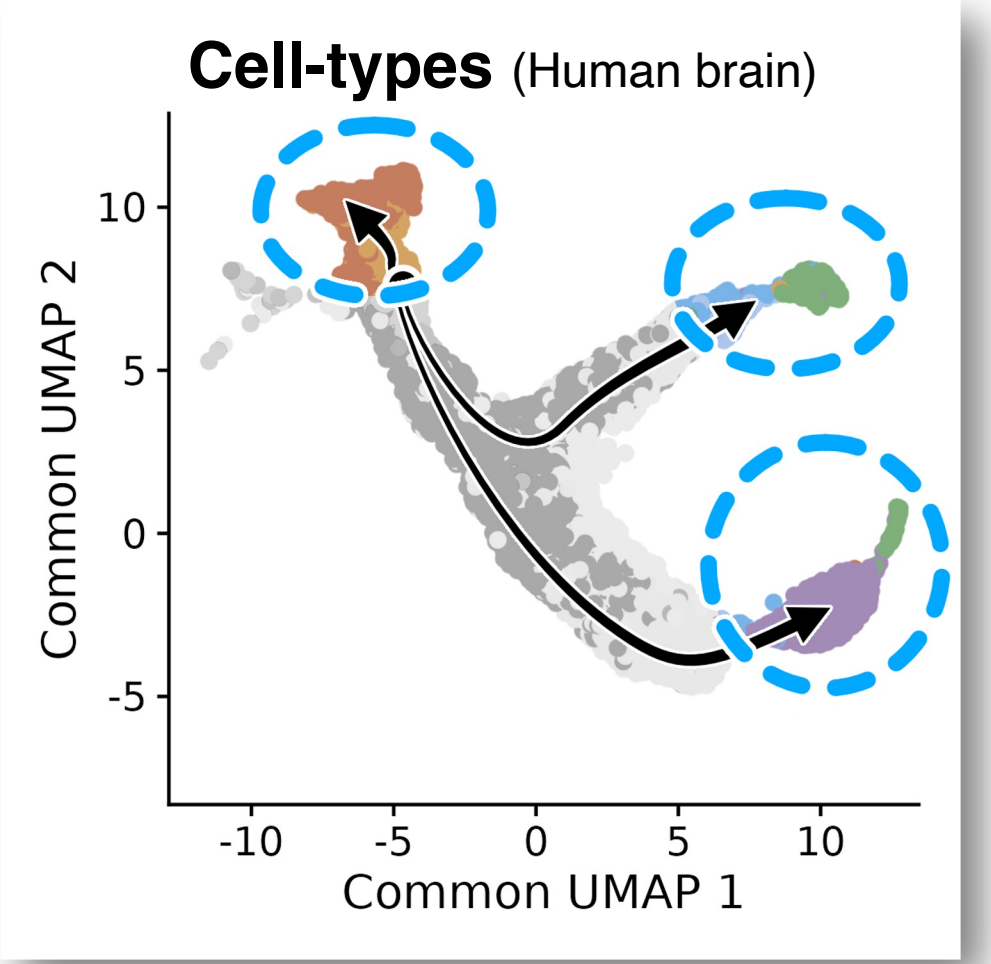


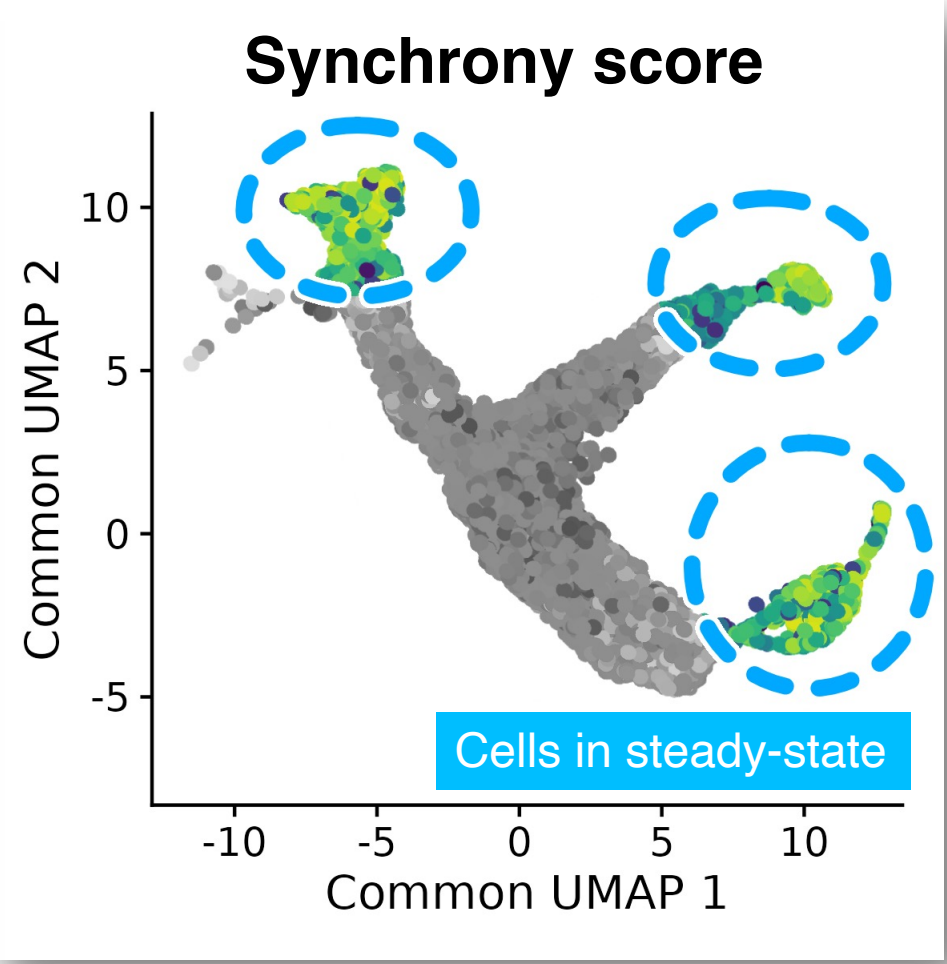
High synchrony score (i.e., steady-state)

Low synchrony score (i.e., undergoing development)

Our contribution: The **synchrony score**, which measures the coordination b/w genes and chromatin accessibility within a cell, to reveal if a cell is in steady-state

After fitting Tilted-CCA, for each cell i: Correlation $\left(\begin{array}{c} \text{common} \\ \text{component} \\ \text{among genes} \end{array} , \begin{array}{c} \text{common + distinct} \\ \text{components among} \\ \text{genes} \end{array} \right)$



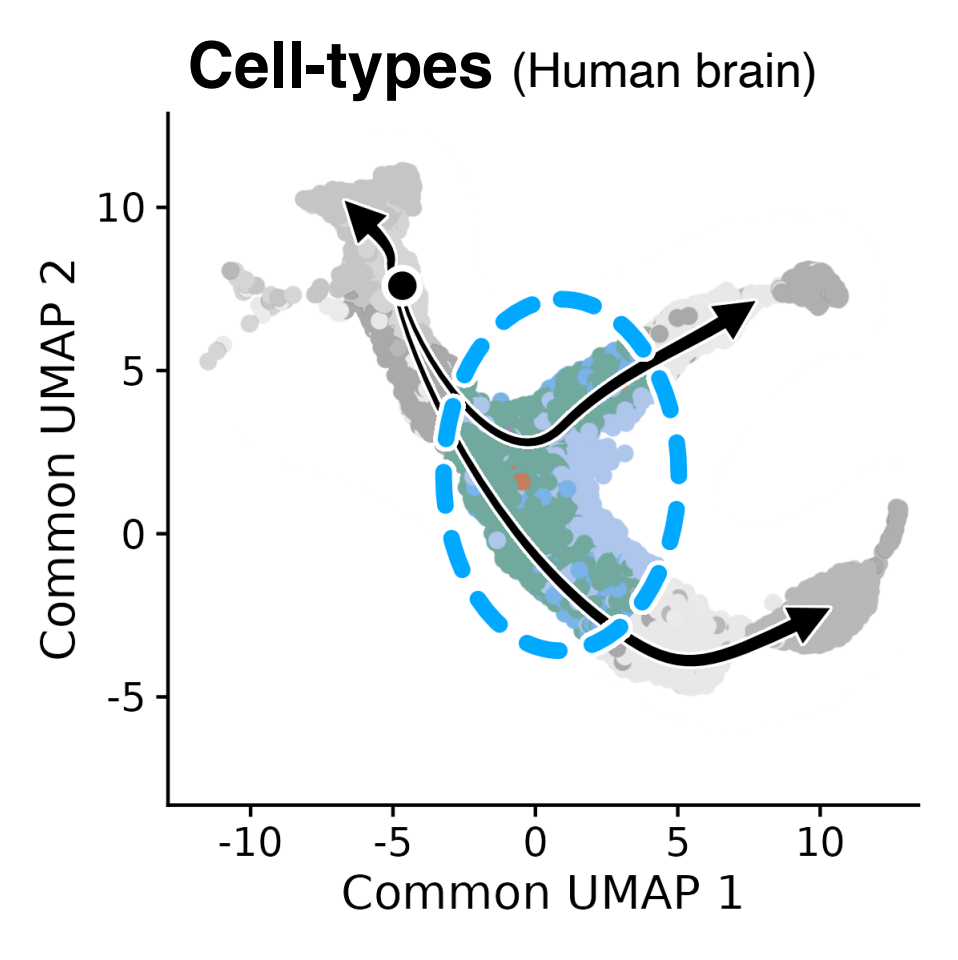


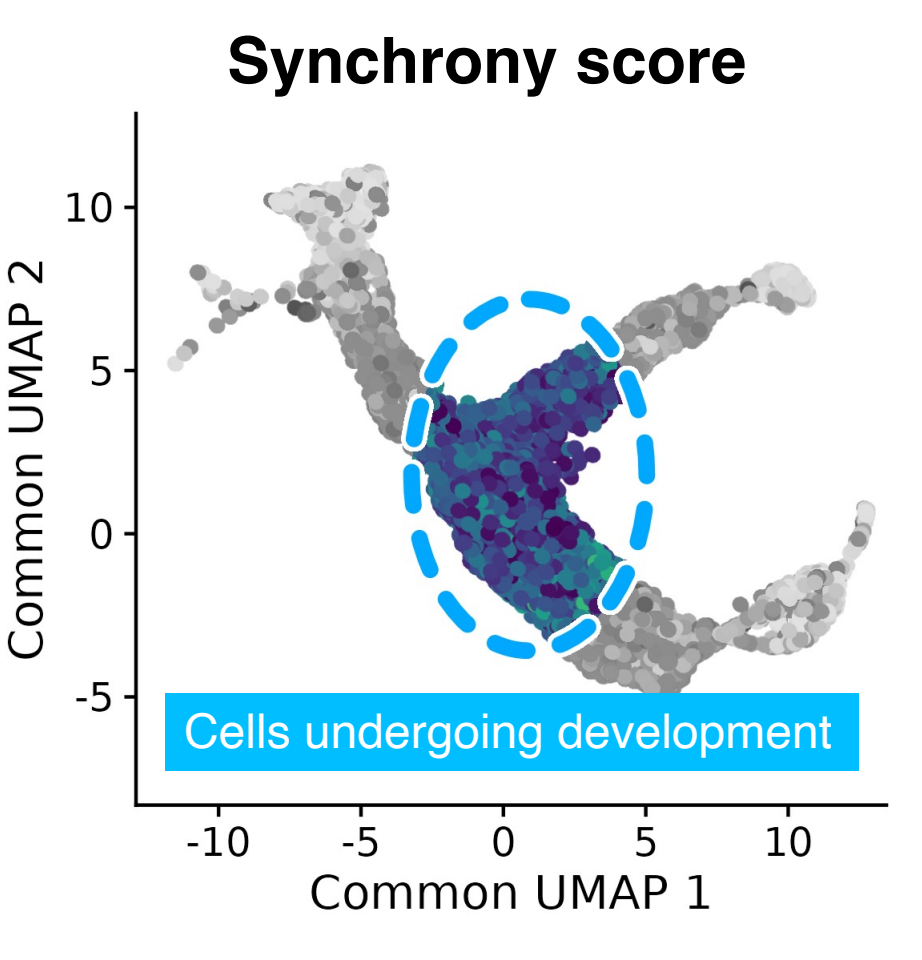
High synchrony score (i.e., steady-state)

Low synchrony score (i.e., undergoing development)

Our contribution: The **synchrony score**, which measures the coordination b/w genes and chromatin accessibility within a cell, to reveal if a cell is in steady-state

After fitting Tilted-CCA, for each cell i: Correlation $\left(\begin{array}{c} \text{common} \\ \text{component} \\ \text{among genes} \end{array} , \begin{array}{c} \text{common + distinct} \\ \text{components among} \\ \text{genes} \end{array} \right)$





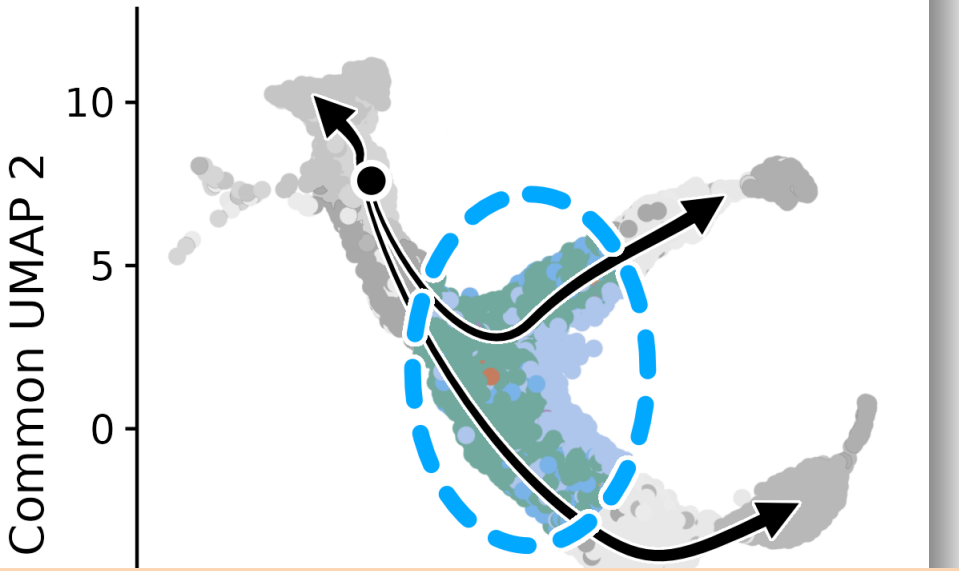
High synchrony score (i.e., steady-state)

Low synchrony score (i.e., undergoing development)

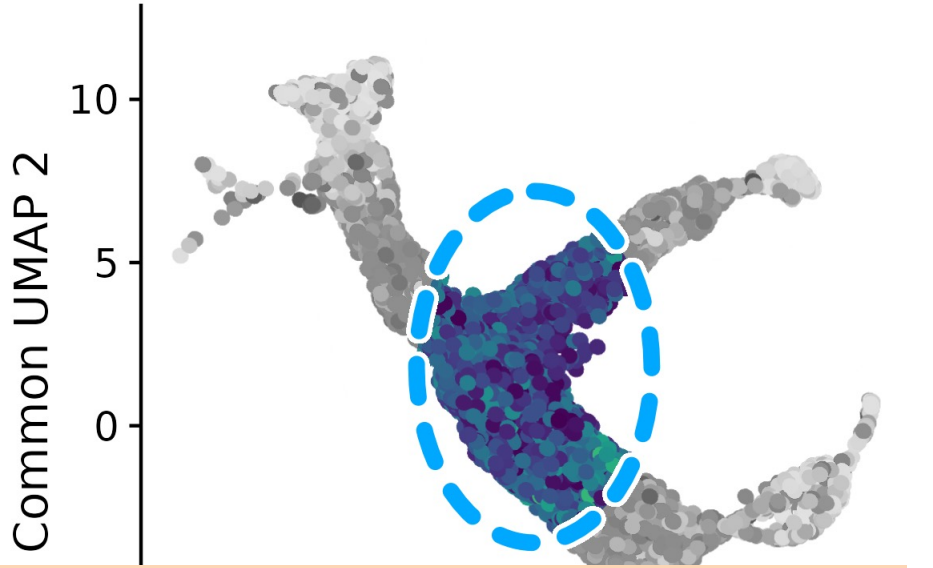
Our contribution: The **synchrony score**, which measures the coordination b/w genes and chromatin accessibility within a cell, to reveal if a cell is in steady-state

After fitting Tilted-CCA, for each cell i: Correlation $\left(\begin{array}{c} \text{common} \\ \text{component} \\ \text{among genes} \end{array} , \begin{array}{c} \text{common + distinct} \\ \text{components among} \\ \text{genes} \end{array} \right)$

Cell-types (Human brain)



Synchrony score



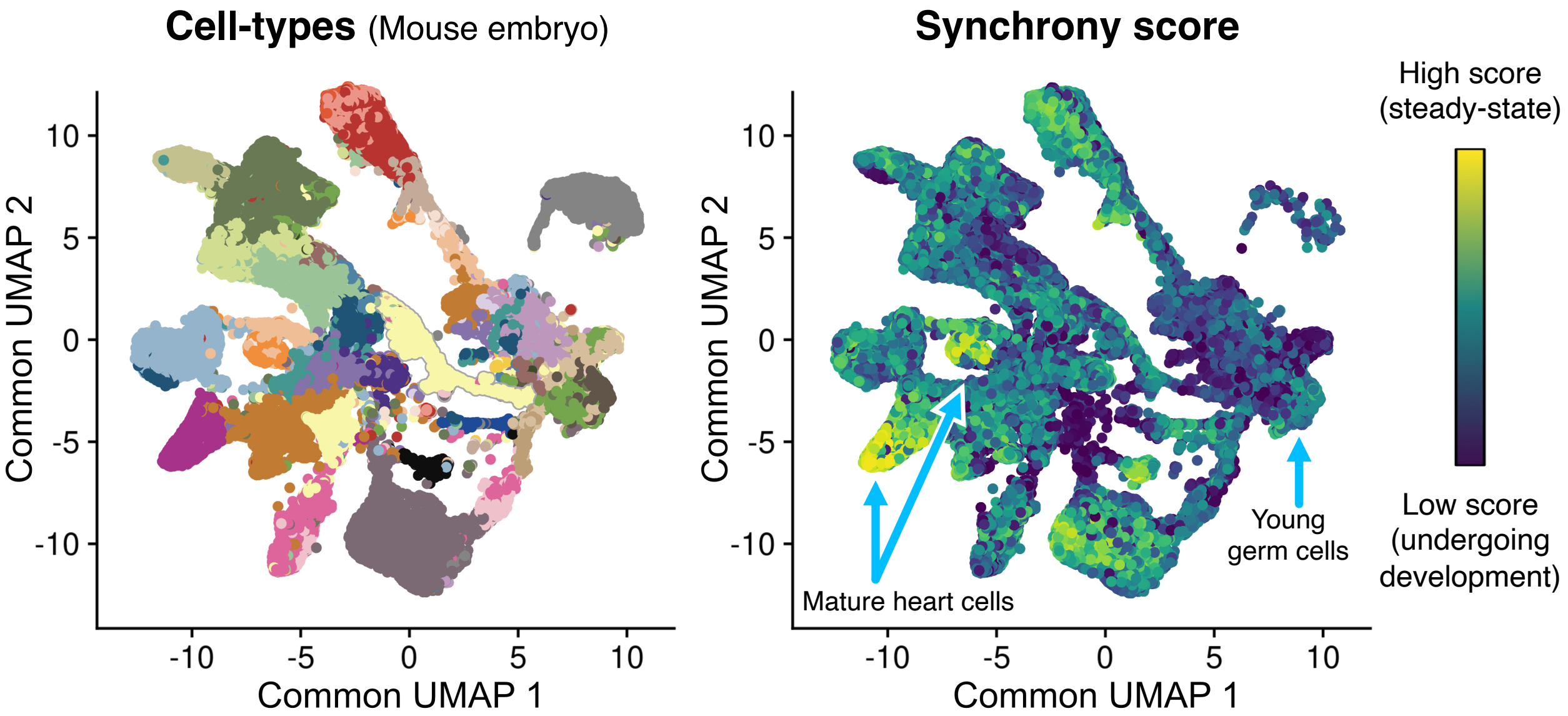
High synchrony score (i.e., steady-state)

Low synchrony score (i.e., undergoing development)

No trajectory estimation needed for the synchrony scores!
Learning the “highways” from “the speed” rely on different principles.

We demonstrate the utility of the synchrony scores on other harder systems.

We demonstrate the utility of the synchrony scores on other harder systems.

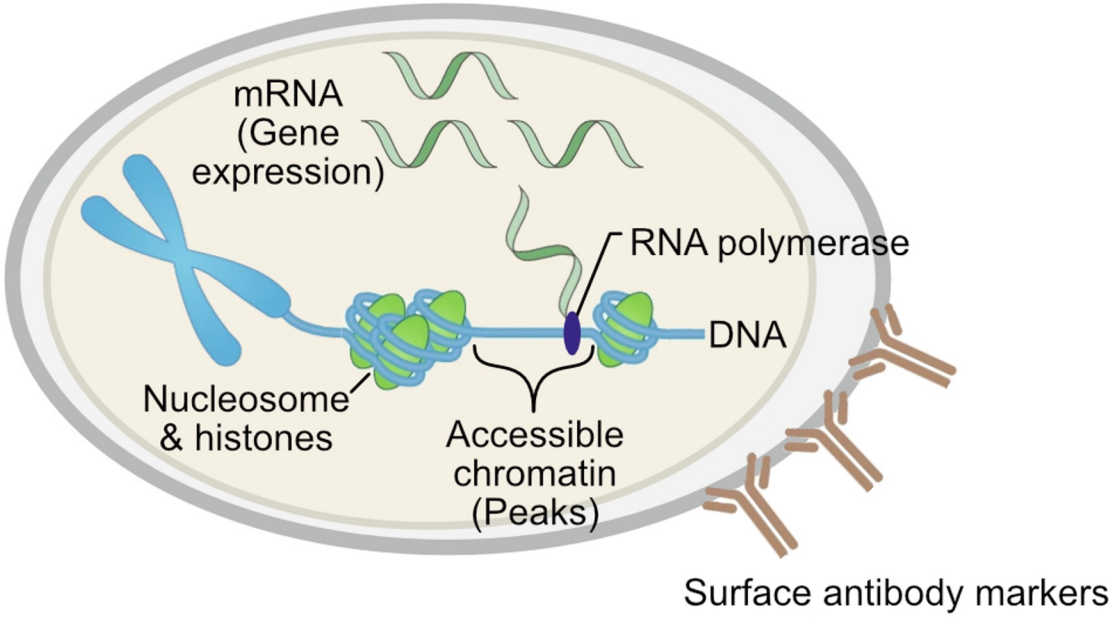


Recap: By matrix factorizing multi-modal data based on the common/distinct geometry, Tilted-CCA enables new perspectives to understand cell biology.

1. **(Experimental design):** Which pair of modalities should biologist sequence to have the most comprehensive understanding?

2. **(Variable selection):** For RNA-Protein data, how can we pick the antibodies that contribute the most additional information to the RNA modality?

3. **(Developmental biology):** Can the amount of coordination between two modalities tell us if a cell in a steady-state or is undergoing development?

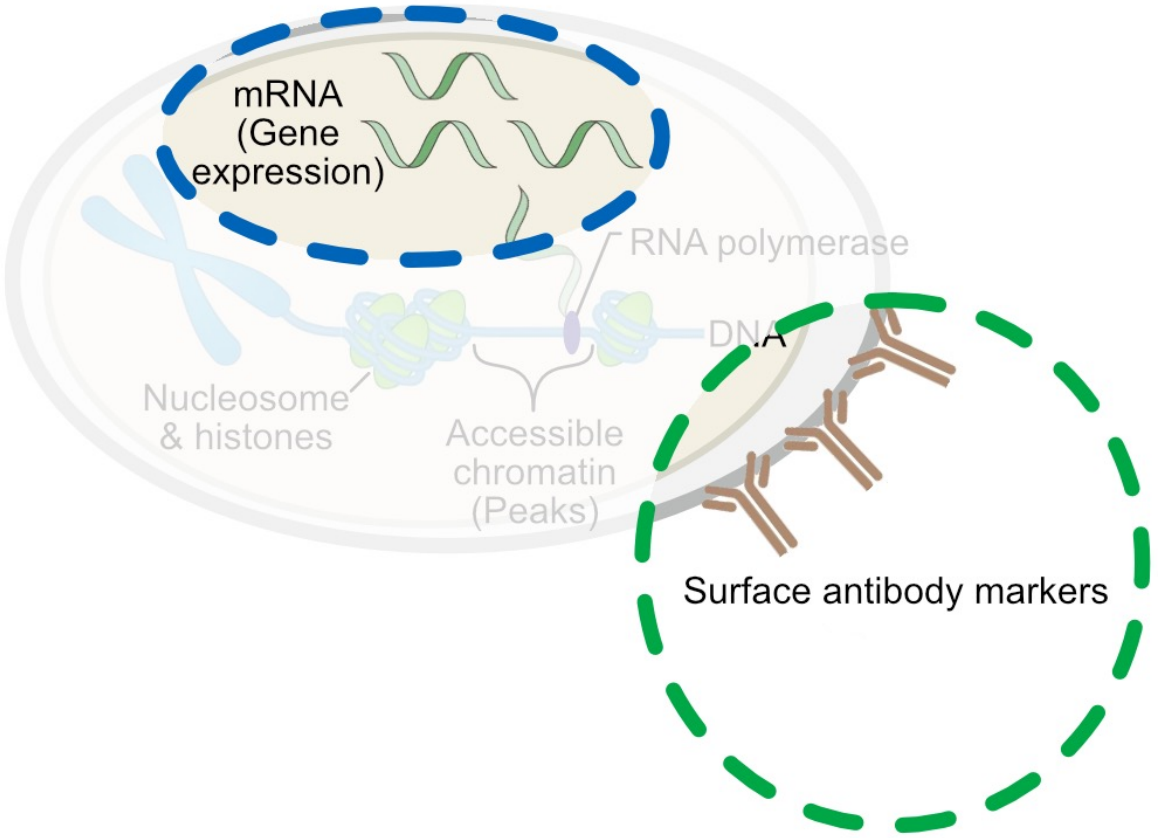


Recap: By matrix factorizing multi-modal data based on the common/distinct geometry, Tilted-CCA enables new perspectives to understand cell biology.

1. **(Experimental design):** Which pair of modalities should biologist sequence to have the most comprehensive understanding?

2. **(Variable selection):** For RNA-Protein data, how can we pick the antibodies that contribute the most additional information to the RNA modality?

3. **(Developmental biology):** Can the amount of coordination between two modalities tell us if a cell in a steady-state or is undergoing development?

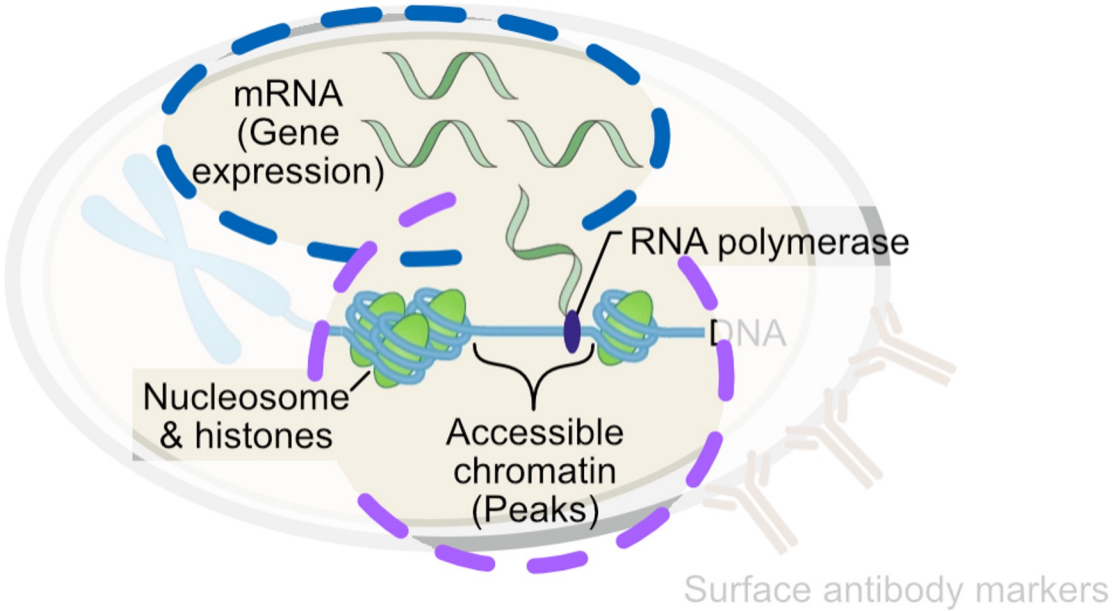


Recap: By matrix factorizing multi-modal data based on the common/distinct geometry, Tilted-CCA enables new perspectives to understand cell biology.

1. **(Experimental design):** Which pair of modalities should biologist sequence to have the most comprehensive understanding?

2. **(Variable selection):** For RNA-Protein data, how can we pick the antibodies that contribute the most additional information to the RNA modality?

3. **(Developmental biology):** Can the amount of coordination between two modalities tell us if a cell in a steady-state or is undergoing development?



Recap: By matrix factorizing multi-modal data based on the common/distinct geometry, Tilted-CCA enables new perspectives to understand cell biology.

1. **(Experimental design):** Which pair of modalities should biologist sequence to have the most comprehensive understanding?

2. **(Variable selection):** For RNA-Protein data, how can we pick the antibodies that contribute the most additional information to the RNA modality?

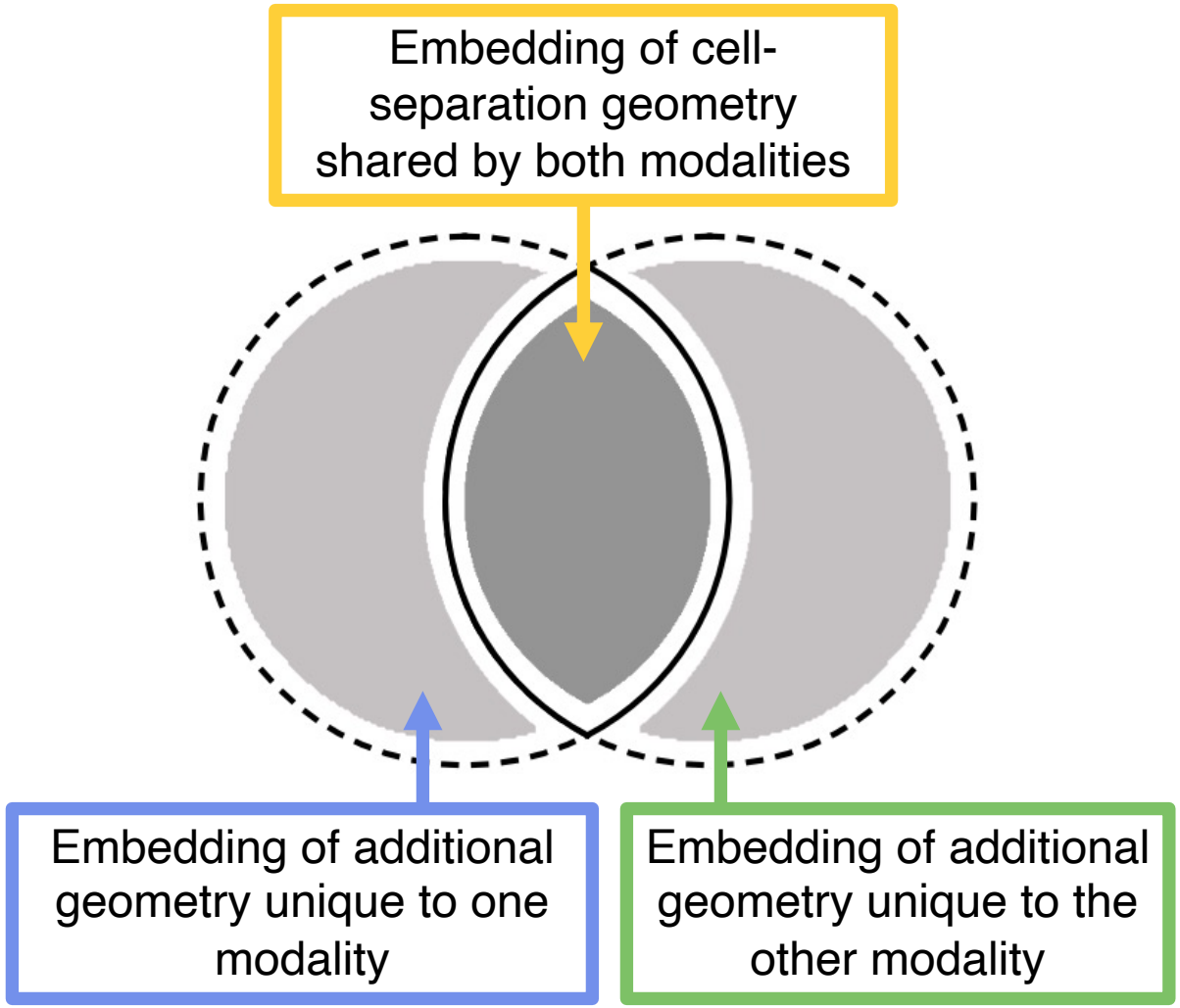
3. **(Developmental biology):** Can the amount of coordination between two modalities tell us if a cell in a steady-state or is undergoing development?

Multi-modal analyses will only become more prevalent in biomedical research, and this will raise new biological theories that will require new matrix factorization approaches to study.

Takeaways:

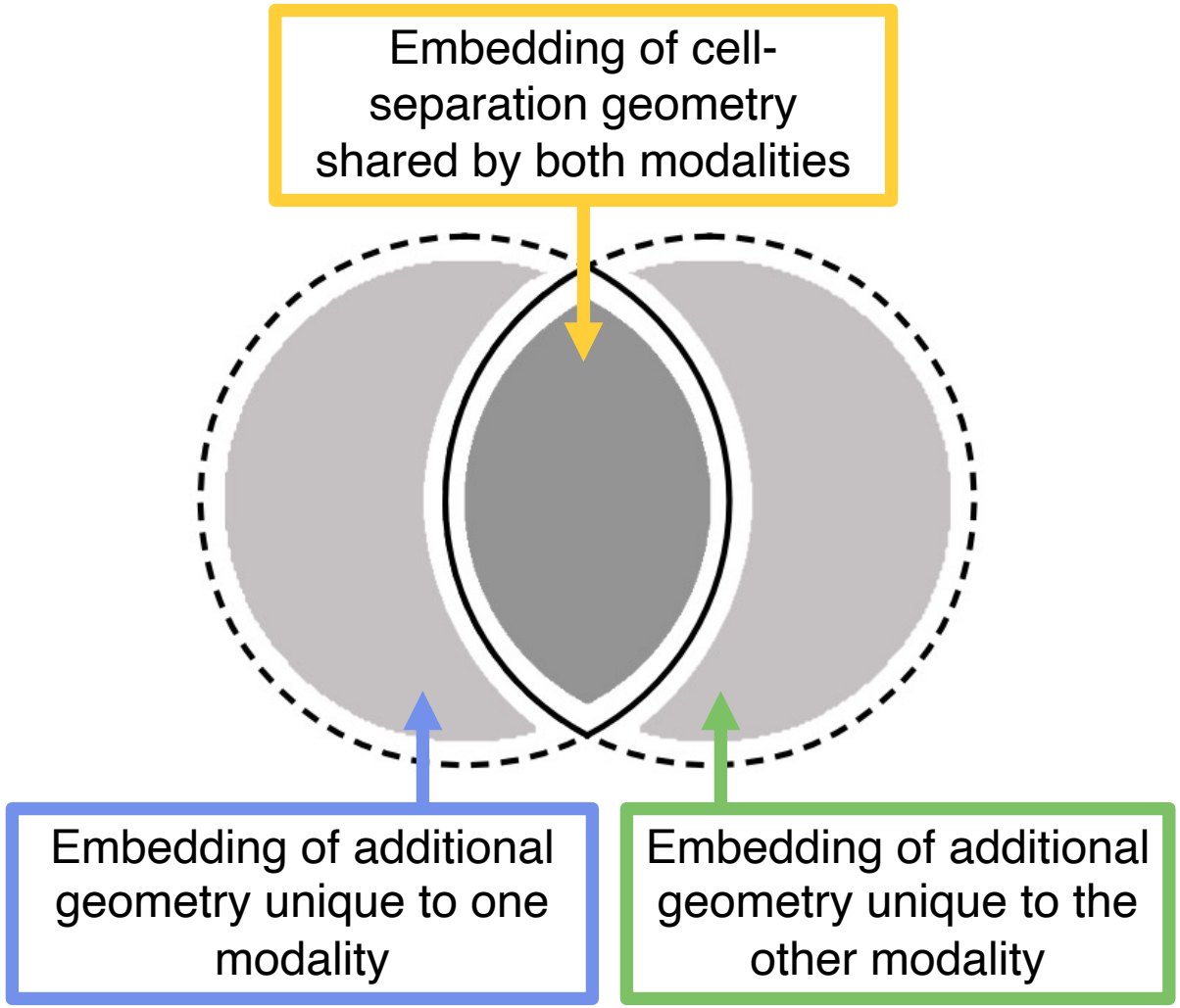
Takeaways:

- **Novelty:** Estimating the “intersection of information” in multi-modal data, as opposed many existing methods for “union of information”
- **Strategy:** Matrix-factorization to estimate embeddings of the common and distinct geometries
- **Specifically:** Build upon CCA by “tilting” of the common vector to approximate the desired shared geometry



Takeaways:

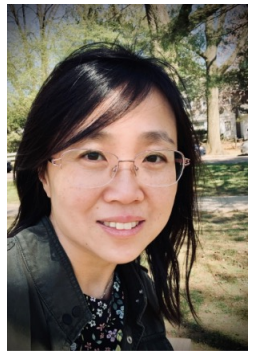
- **Novelty:** Estimating the “intersection of information” in multi-modal data, as opposed many existing methods for “union of information”
- **Strategy:** Matrix-factorization to estimate embeddings of the common and distinct geometries
- **Specifically:** Build upon CCA by “tilting” of the common vector to approximate the desired shared geometry
- **Future theory:** What is the “population” shared geometry?
- **Future biology:** Cancer biology, where we need to understand the specifics of cellular mechanisms



Thank you!

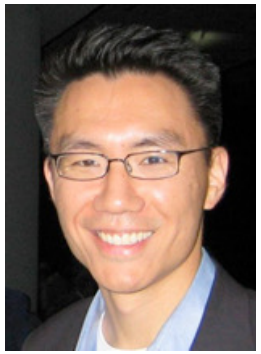
Tilted-CCA on Biorxiv: `2022.10.07.511320`

linnykos



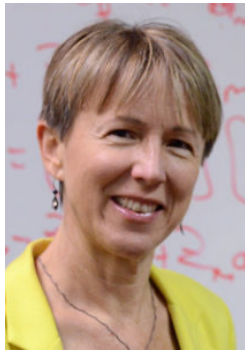
Nancy Zhang

Sydney Shaffer

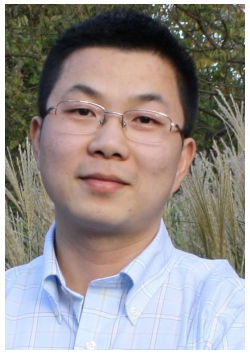


Andy Minn

E. John Wherry



Kathryn Roeder



Jing Lei

Ryan Tibshirani

Alessandro Rinaldo

Han Liu

Max G'Sell

James Sharpnack

Sangwon Hyun