UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

CENTRE FOR ORGANISMAL STUDIES

VELTEN GROUP

HGS MathComp

Interdisciplinary Center for Scientific Computing

European Laboratory for Learning and Intelligent Systems

# MULTI-MODAL DATA ANALYSIS
## USING PROBABILISTIC FACTOR MODELS

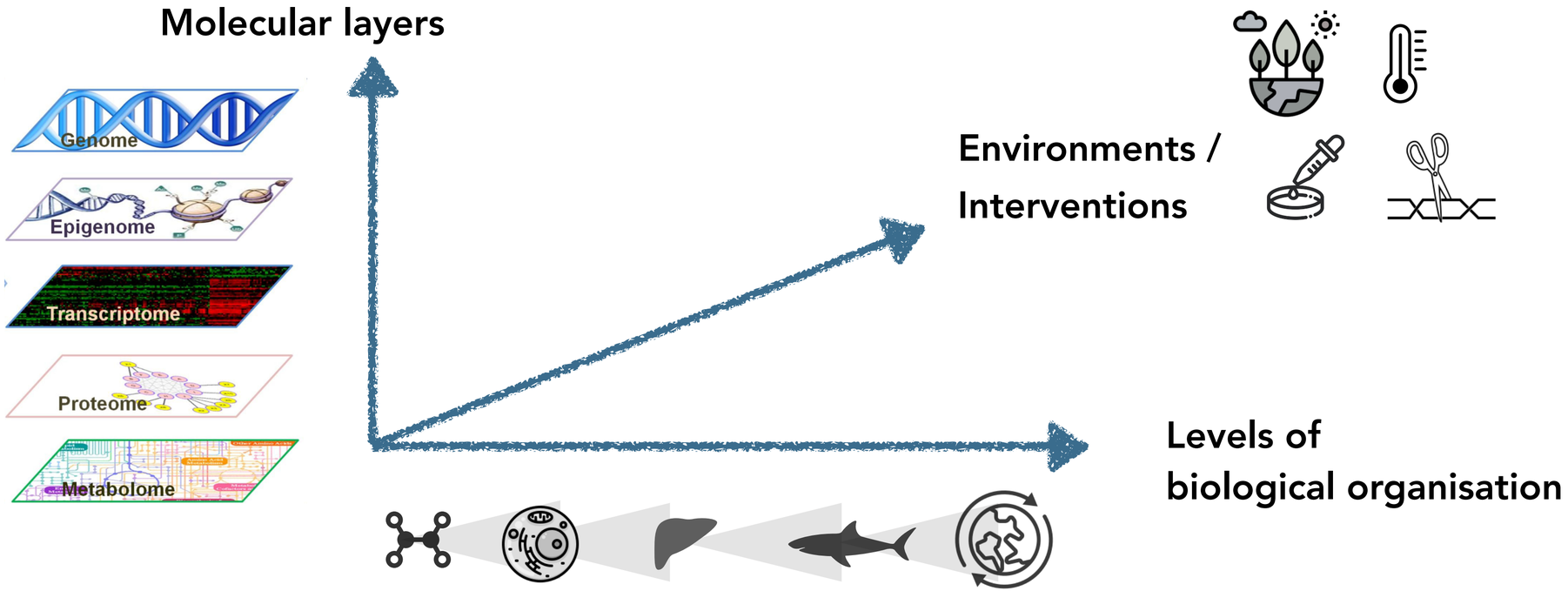
Francisca Gaspar Vieira

"Autumn School for Single Cell-ers"

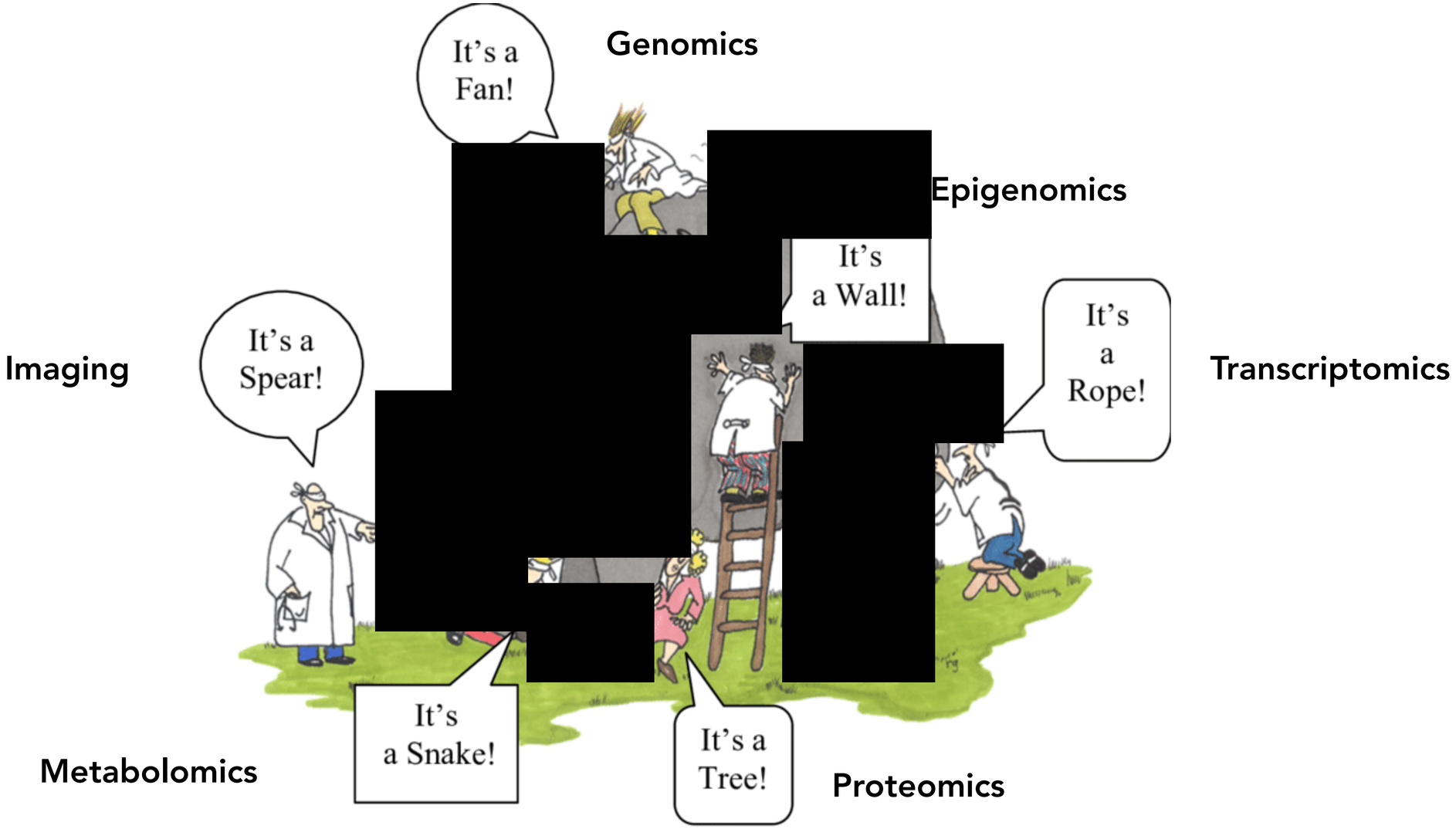
October 2025

1. Motivation
2. Intuition and core idea
3. The maths behind MOFA
4. Guide for factor interpretation
5. MEFISTO
6. Case studies

# A holistic understanding of biological processes requires multi-dimensional approaches

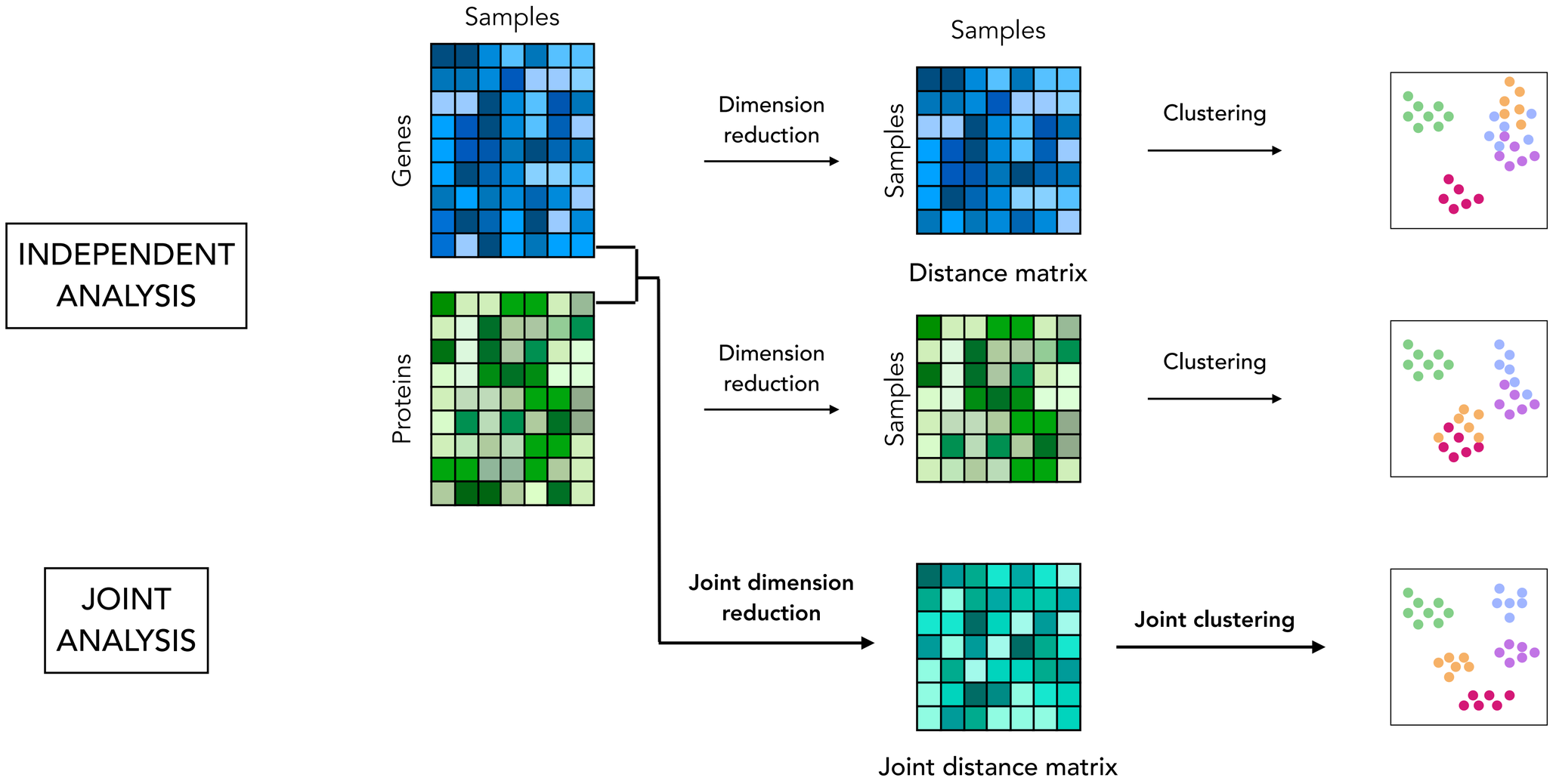


# Why do we (often) need to study different omic types together?




Adapted from Daigneault, Methodological Innovations Online 2013

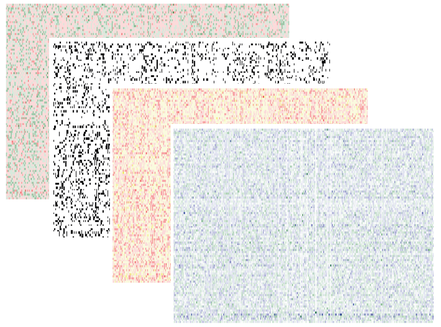
# Joint analysis can identify new sample subgroups and feature relationships



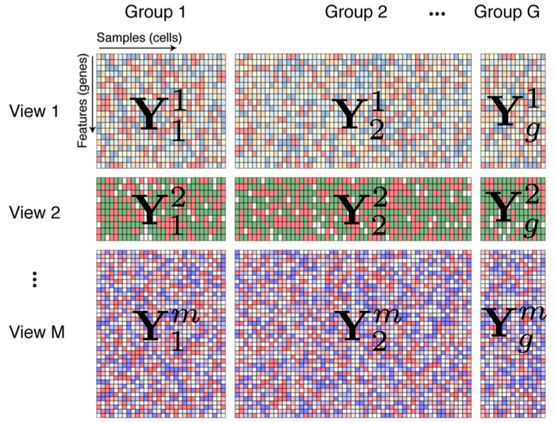

Adapted from Stuart & Satija, Nat Rev Gen 2019

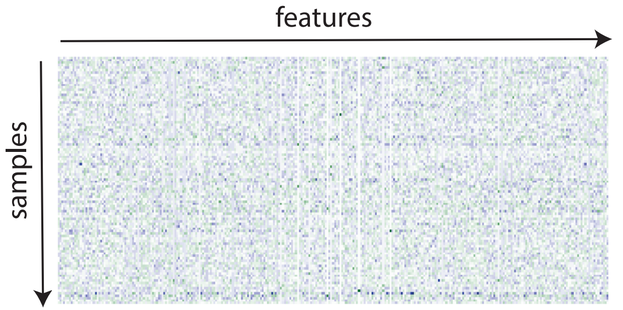
# Challenges in the integration of multi-modal data

- **Heterogeneous data** from different techniques come with distinct statistical properties and inherent structure
- Complex **correlation structures** and hidden confounders
- **High-dimensional** data requires appropriate **regularization** strategies
- Algorithms need to be **scalable** to large data sets
- Large amounts (and different patterns) of **missing values**



heterogeneous



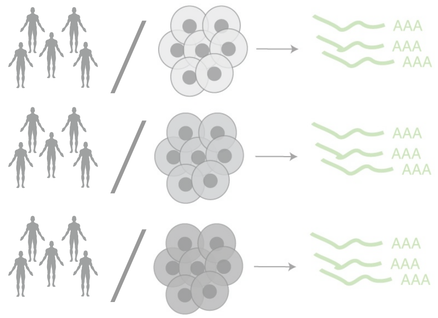



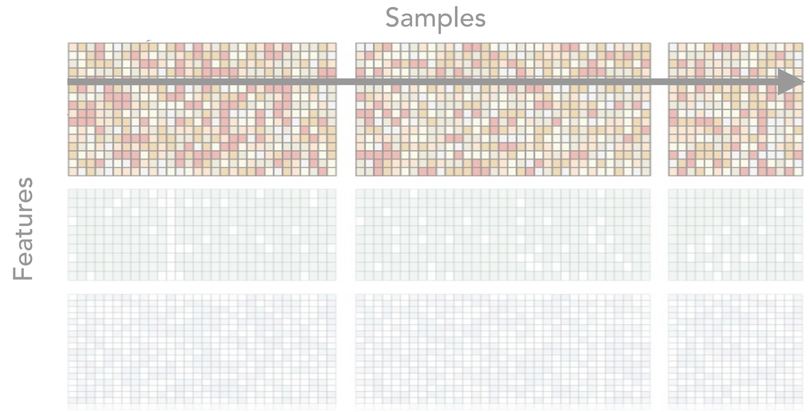

high-dimensional

Adapted from Stuart & Satija, Nat Rev Gen 2019

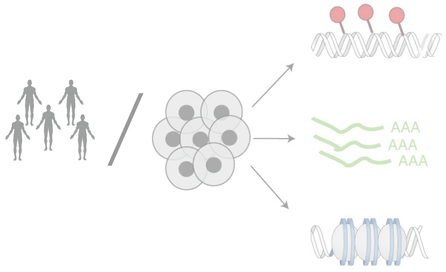
# How to integrate multi-modal data?

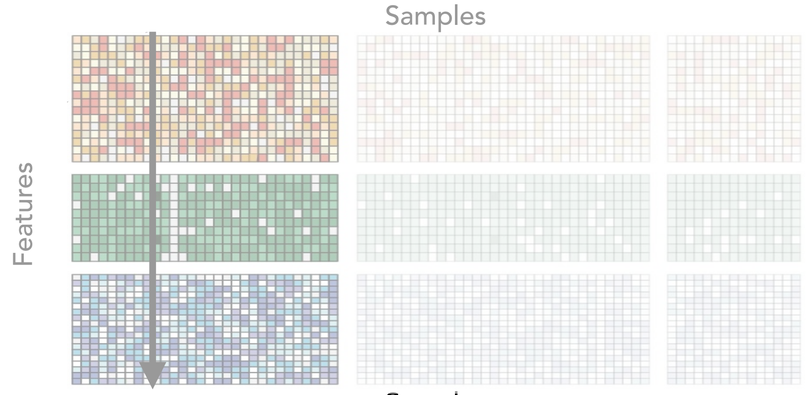
Horizontal integration
(features as anchors)



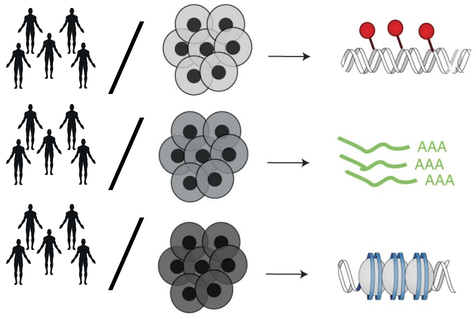



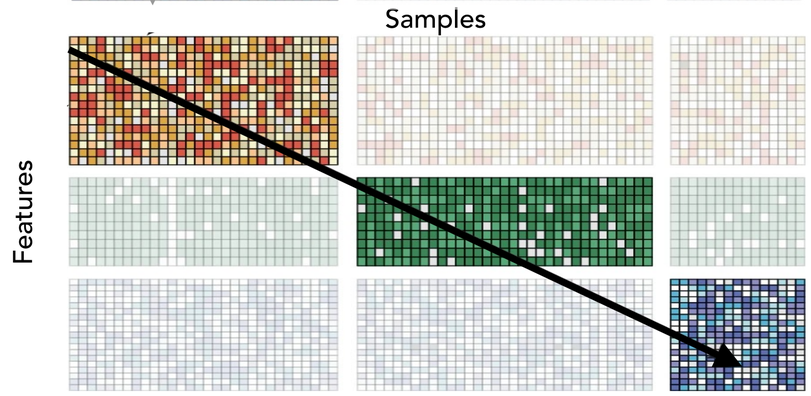

Vertical integration
(samples as anchors)




Diagonal integration
(no anchors)



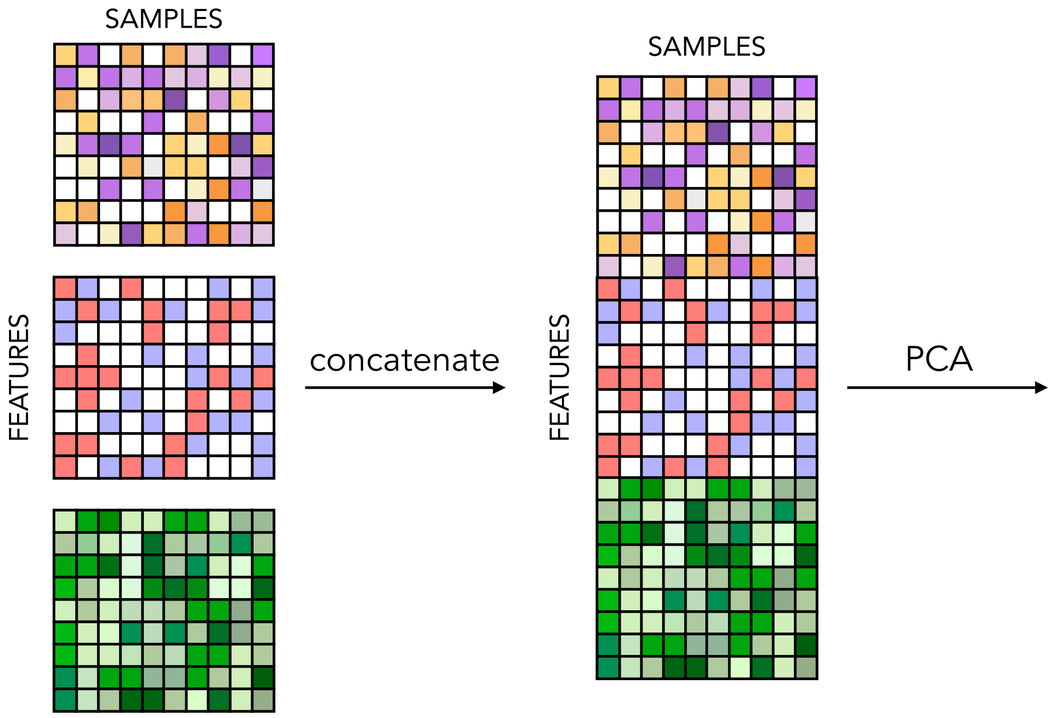



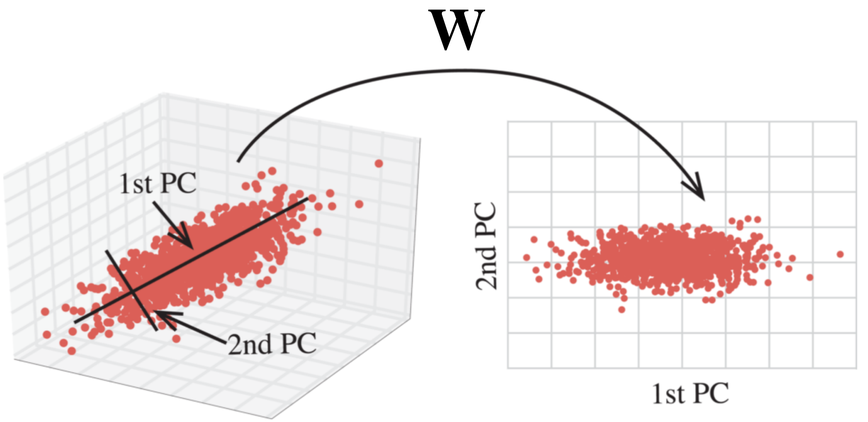

Adapted from Argelaguet*, Cuomo* et al. Nature Biotech 2021

# Latent variable models are a useful tool for global analysis and dimension reduction

PCA is widely used linear example that uses a linear mapping f that maps the observations $\mathbf{Y} \in \mathbb{R}^{N \times D}$ to latent representations $\mathbf{Z} \in \mathbb{R}^{N \times K}$ via an orthogonal weight matrix $\mathbf{W} \in \mathbb{R}^{D \times K}$.

Naive approach: Concatenation



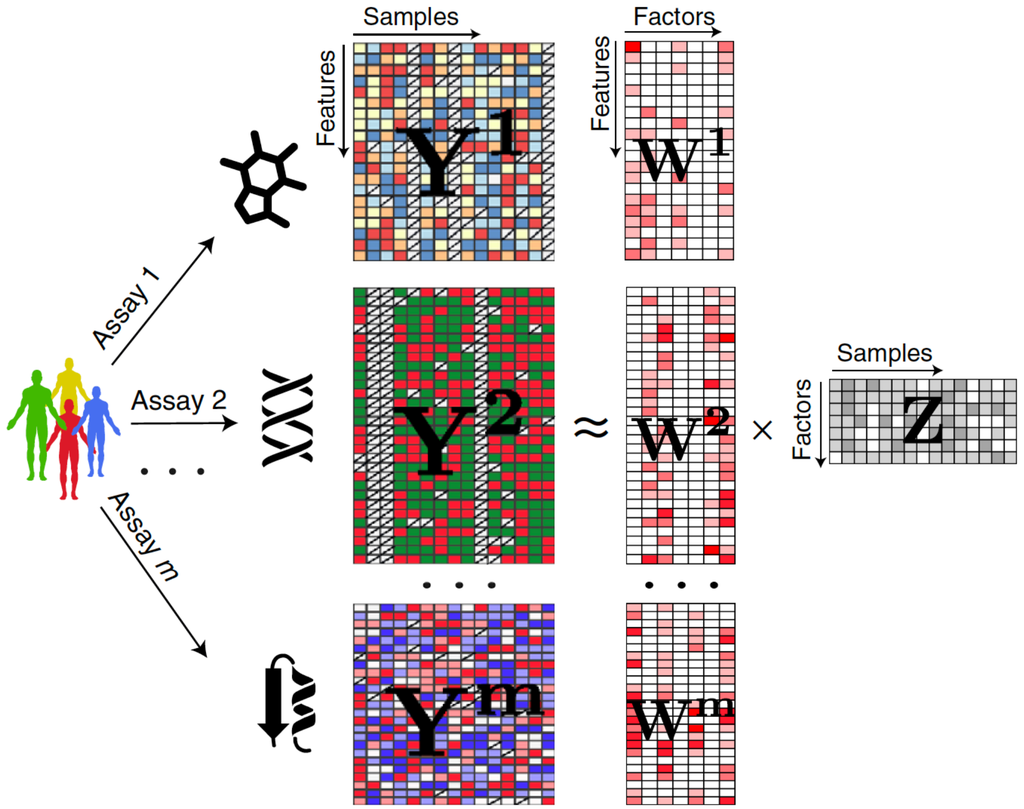




**Problems:**

- Does not respect heterogeneity of data modalities
- Cannot provide information on the relationships between data modalities (only between individual features)
- Data modalities with many features dominate the embedding

# MOFA extend PCA to multiple layers using *structured* matrix factorisation

**Bayesian factor model** to infer a joint low-dimensional representation of multi-modal data in terms of (hidden) factors representing the principal axes of variation



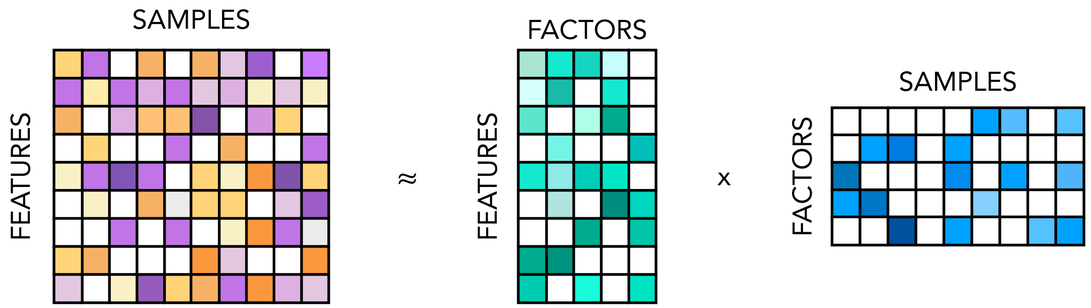

- Interpretable model using a **two-level sparsity priors**
  - *Which modalities are important for a factor?*
  - *Which features are important for a factor?*

- **Scalable inference** using (stochastic) variational Bayes

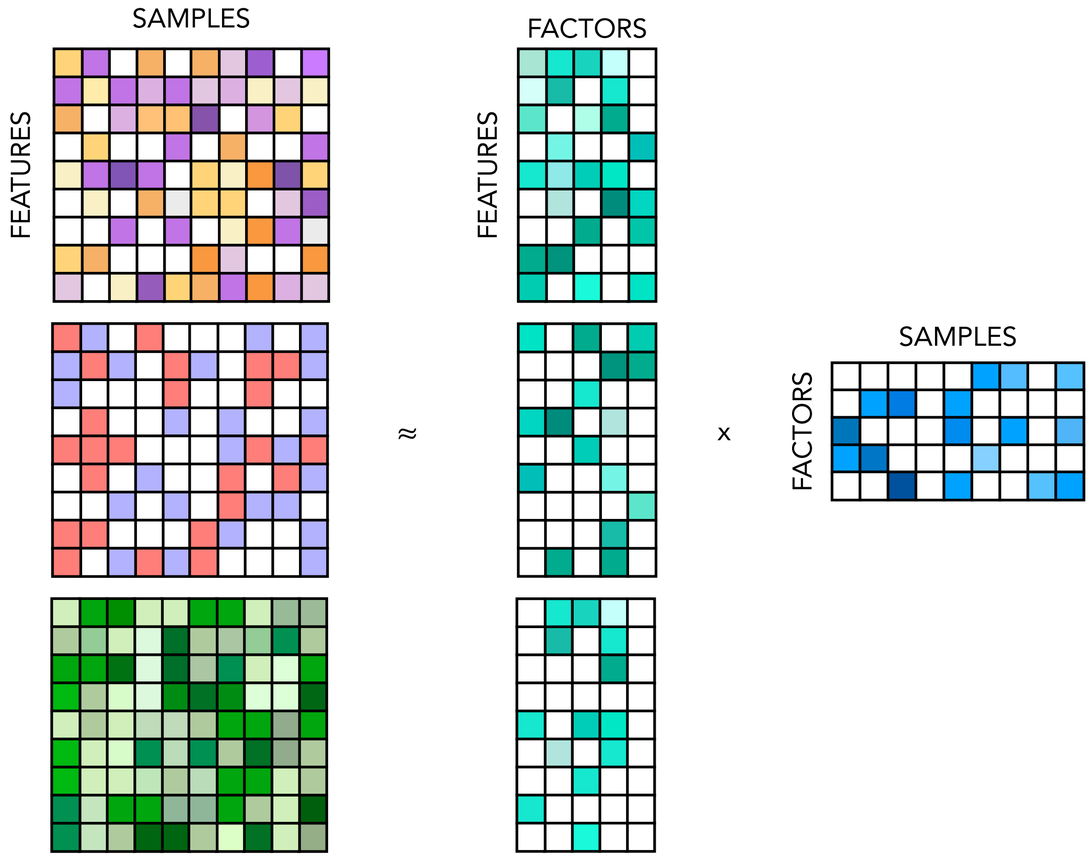
MOFA: Argelaguet* & Velten* et al *Mol Sys Bio*. 2018
MOFA+: Argelaguet*, Arnol*, Bredikhin* et al *Genome Biology*. 2020

1. Motivation
2. Intuition and core idea
3. The maths behind MOFA
4. Guide for factor interpretation
5. MEFISTO
6. Case studies

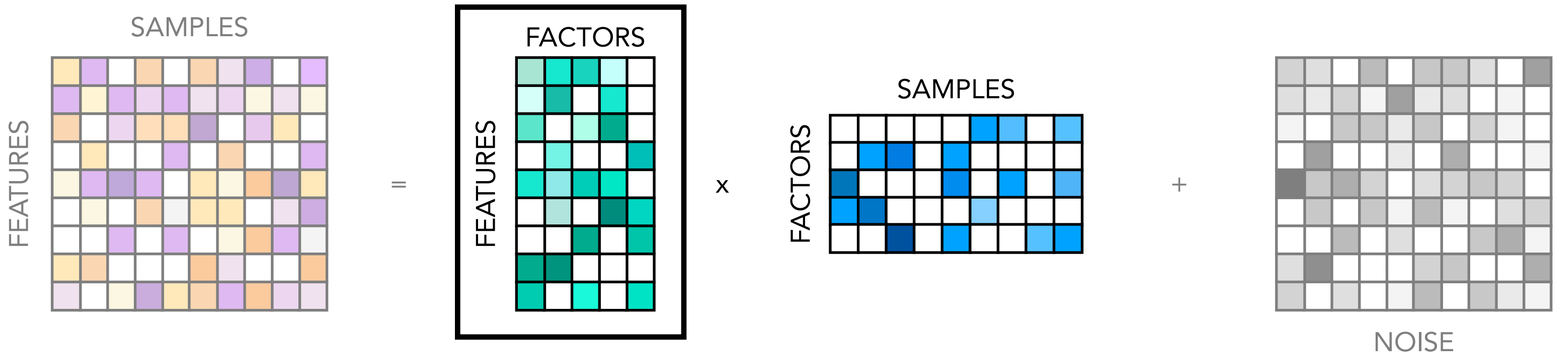
# MOFA is based on factor analysis

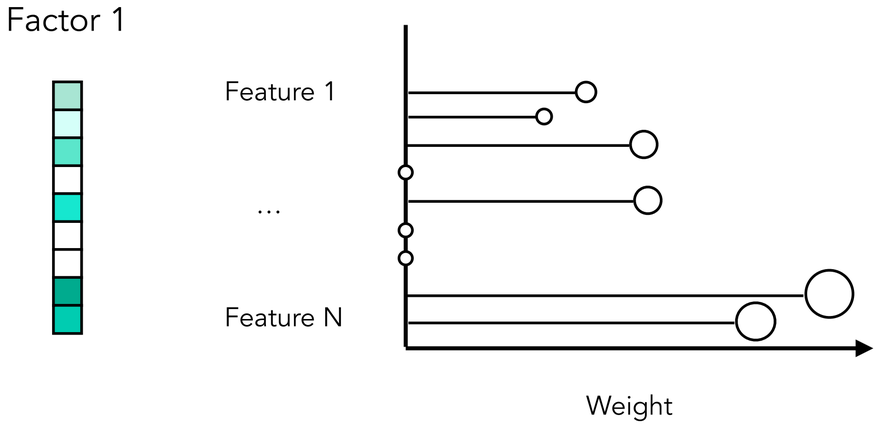


# Factor analysis can be extended to multi-view factor analysis

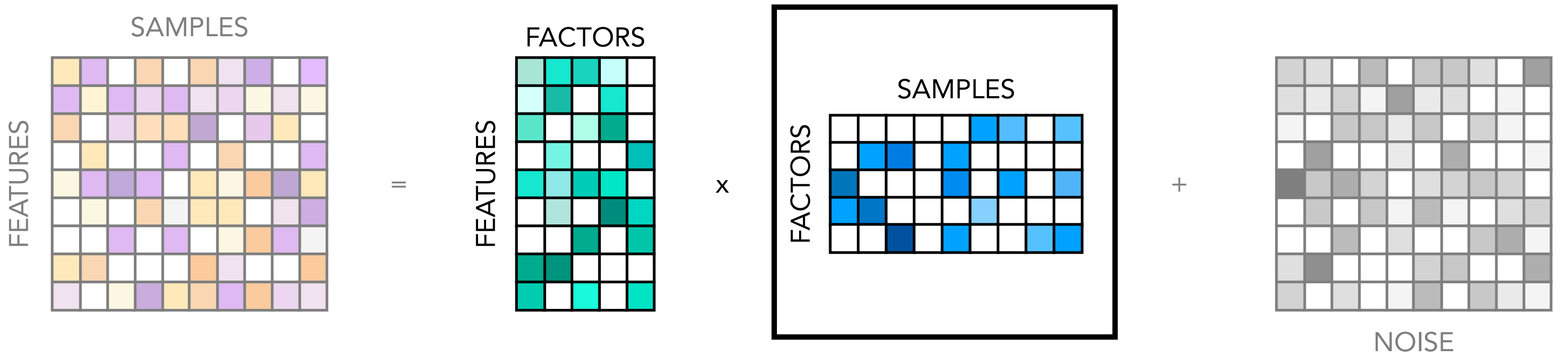


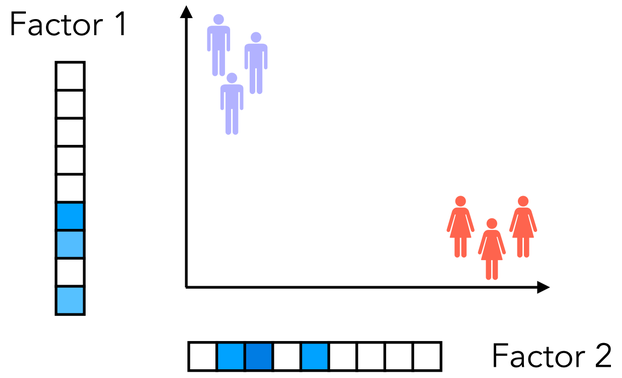
# What information can we get from the matrices?



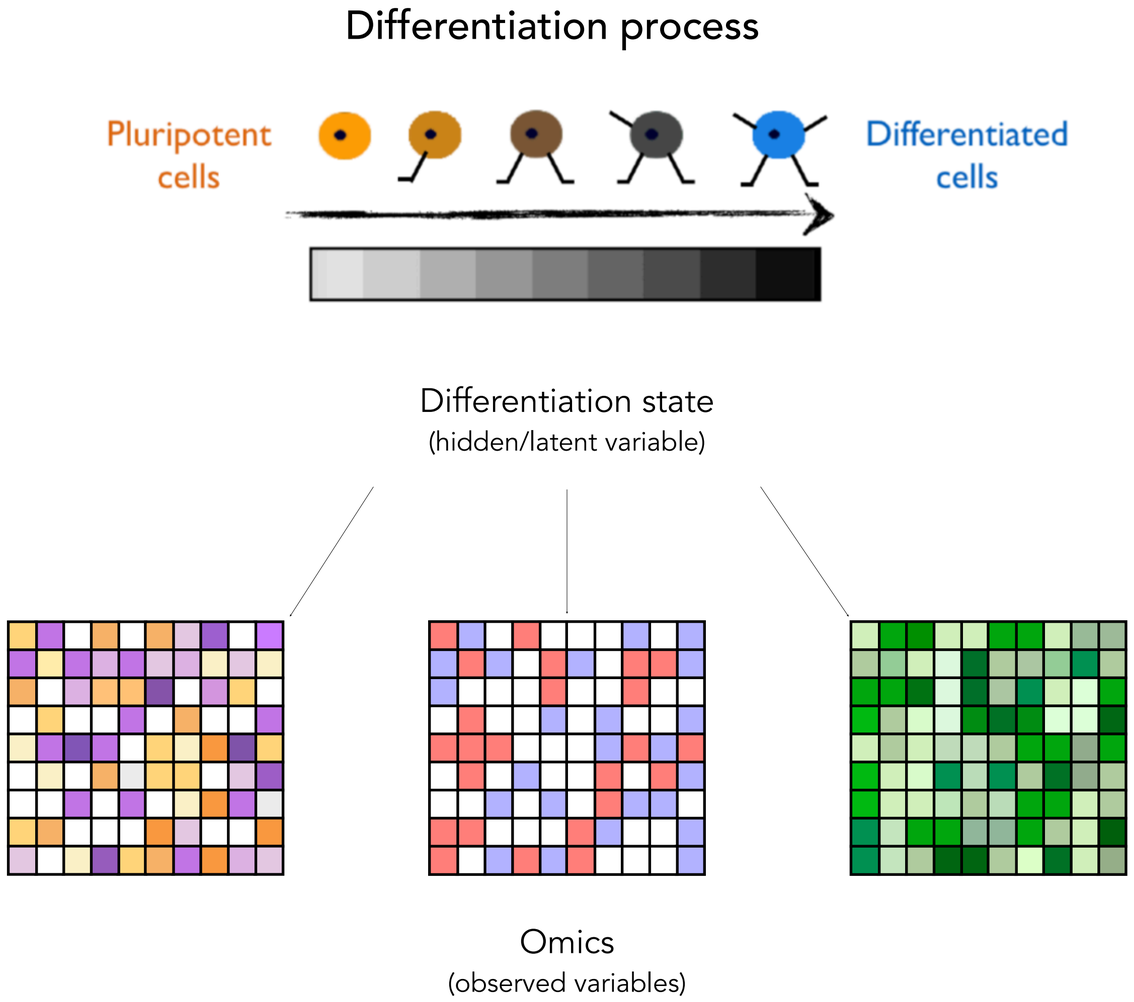



# What information can we get from the matrices?

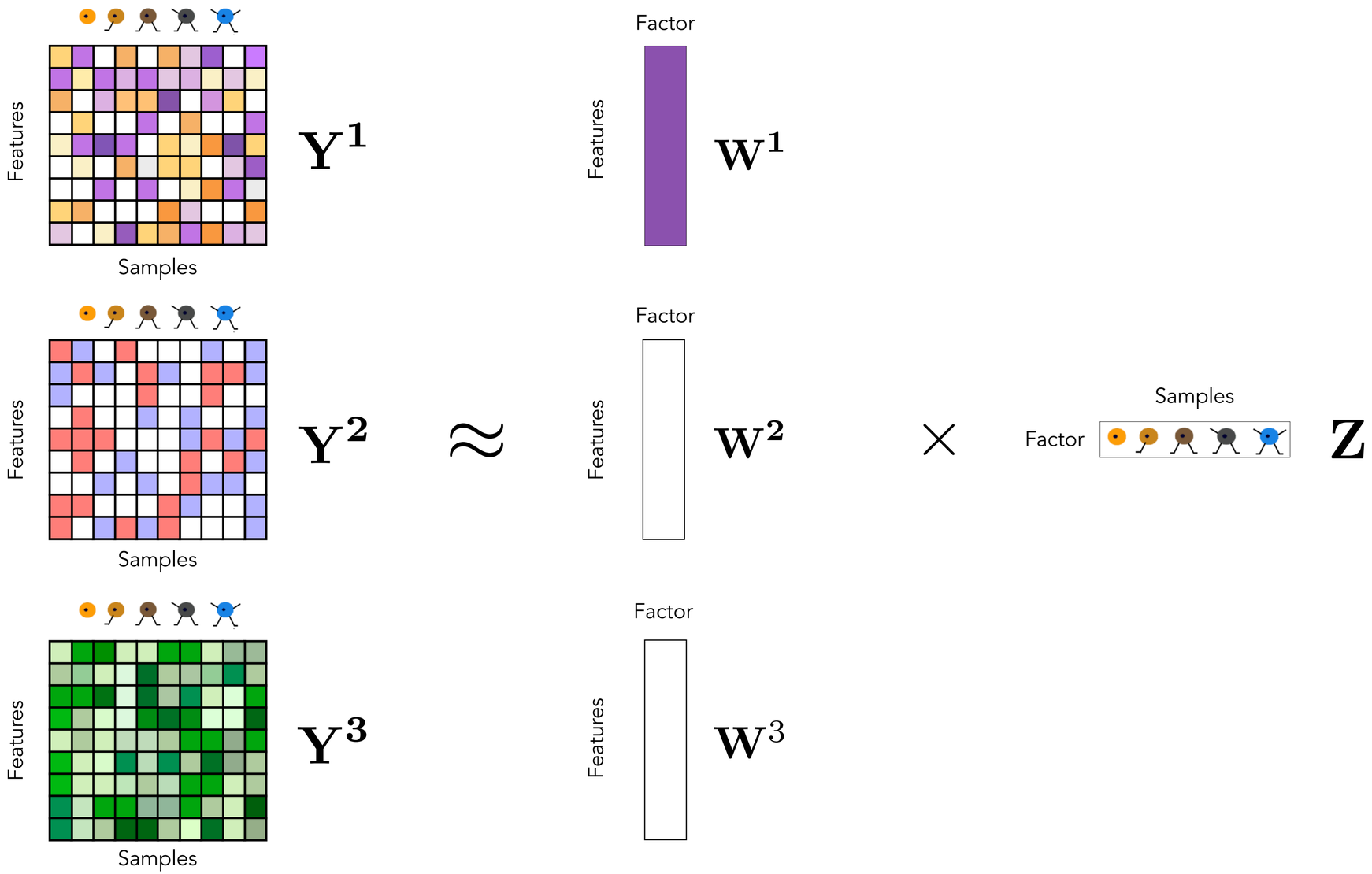


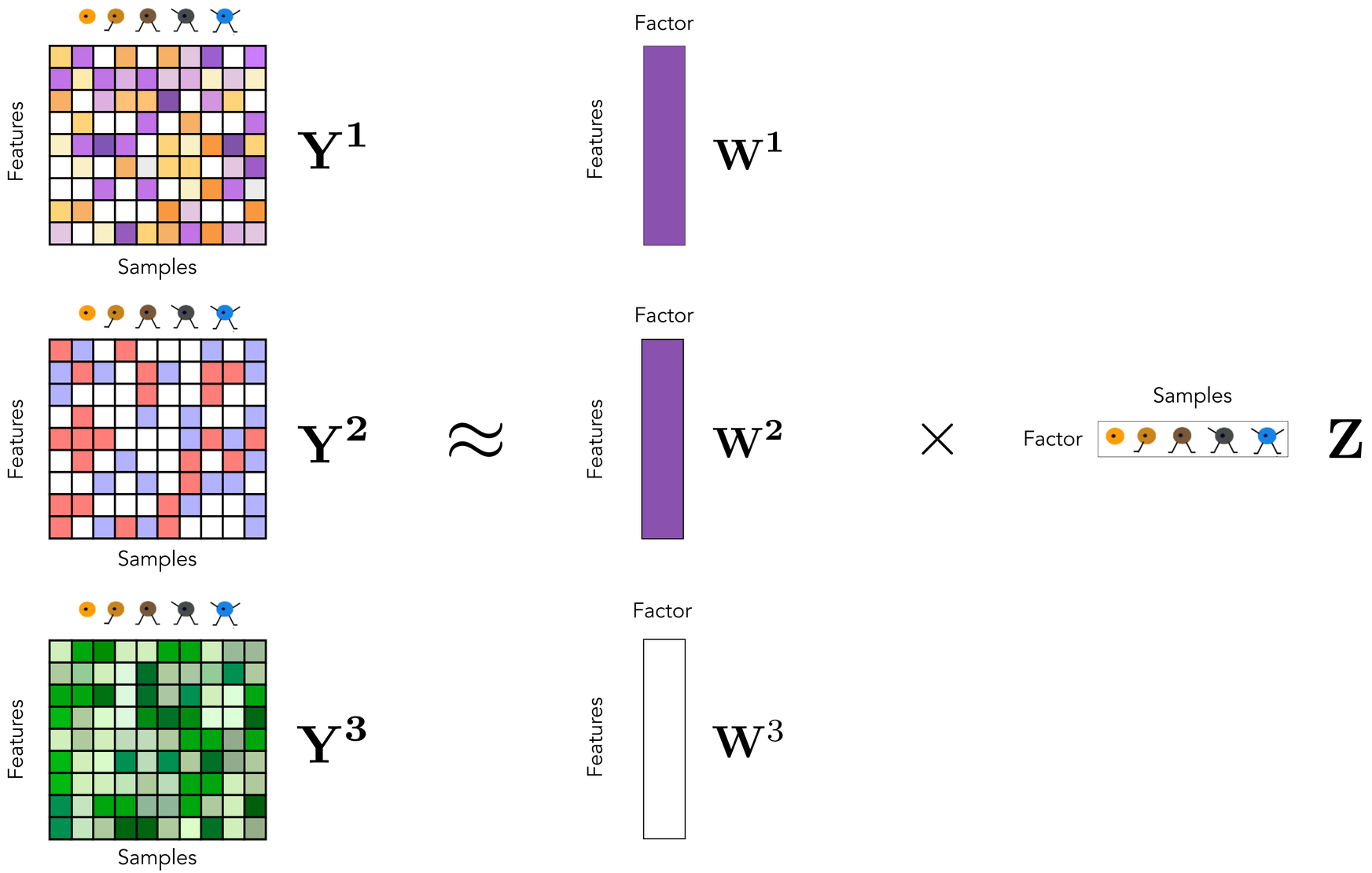
# Intuition behind MOFA



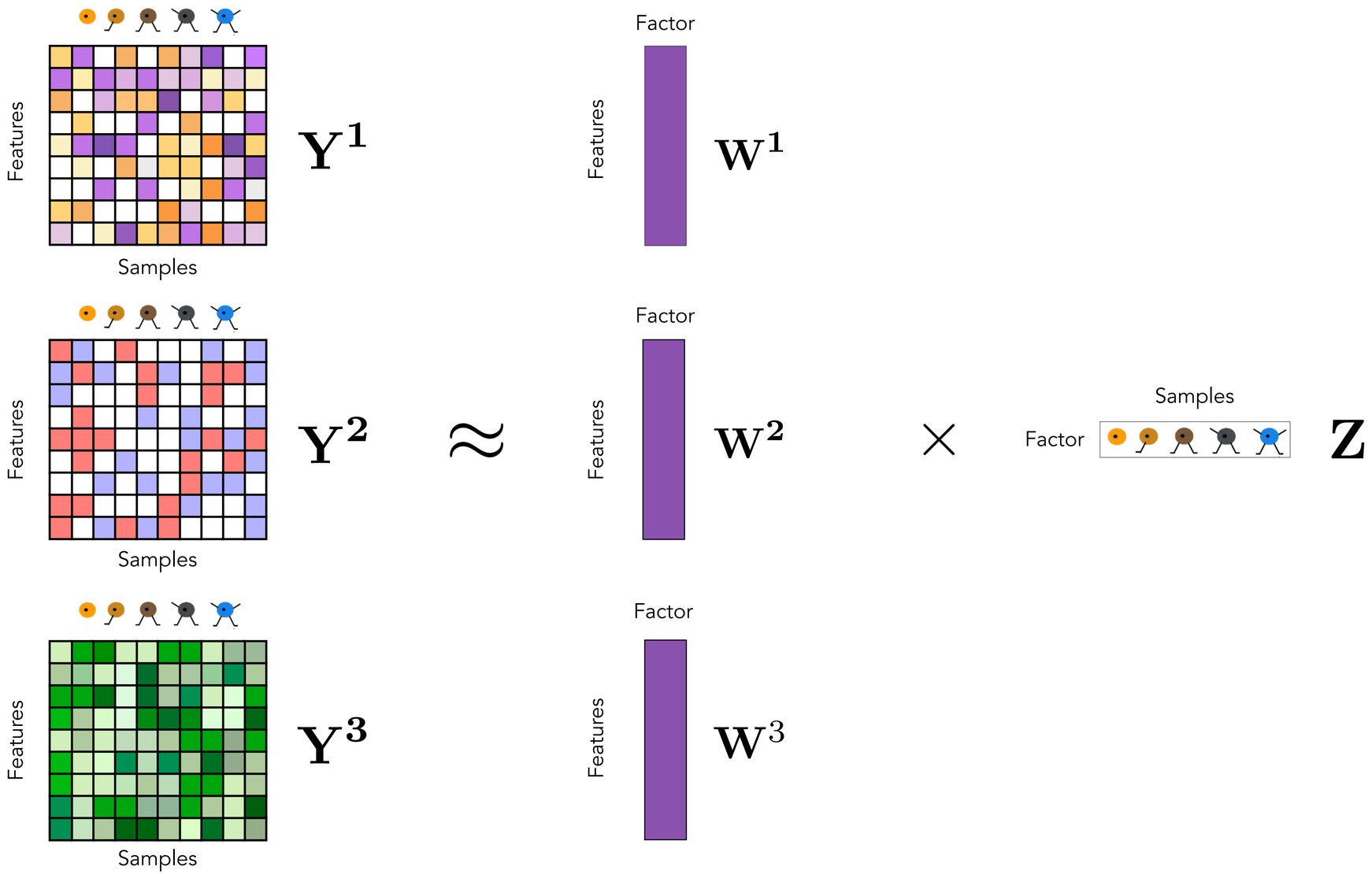
# The differentiation state is a driver of variation in transcriptomics only



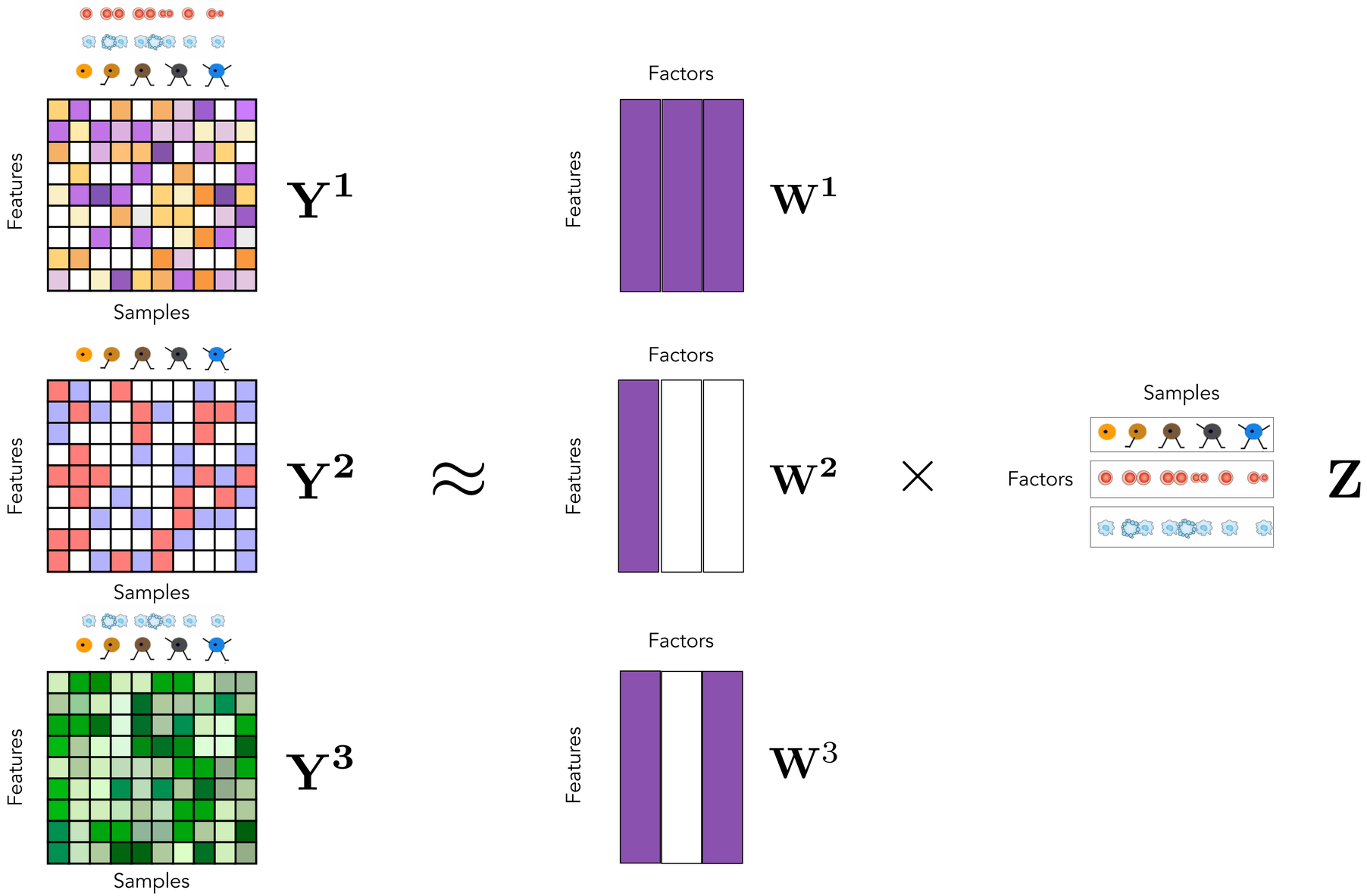
# The differentiation state is a driver of variation in transcriptomics and genomics



# The differentiation state is a driver of variation in all omics



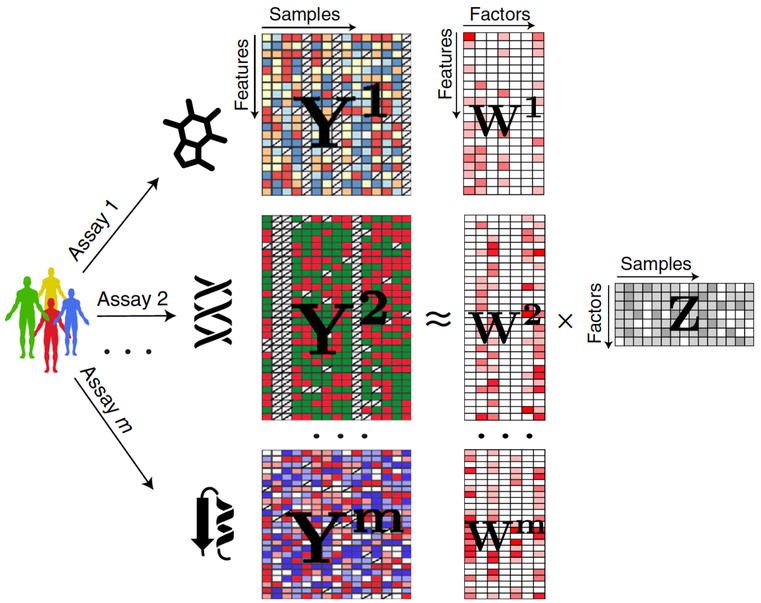
# In addition, there can be other sources of variation

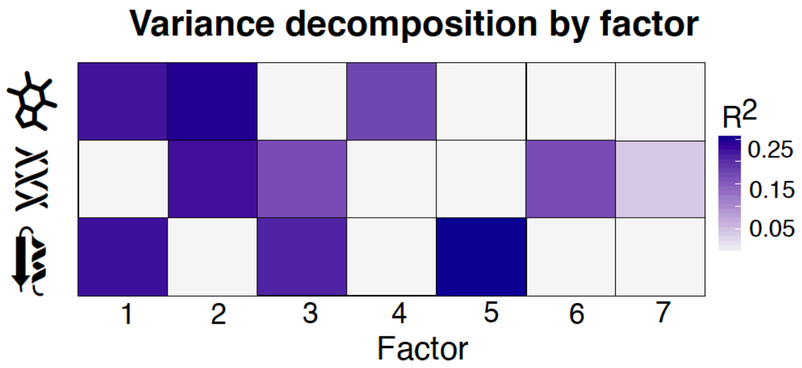


1. Motivation
2. Intuition and core idea
3. The maths behind MOFA
4. Guide for factor interpretation
5. MEFISTO
6. Case studies

# Downstream analysis: Variance decomposition

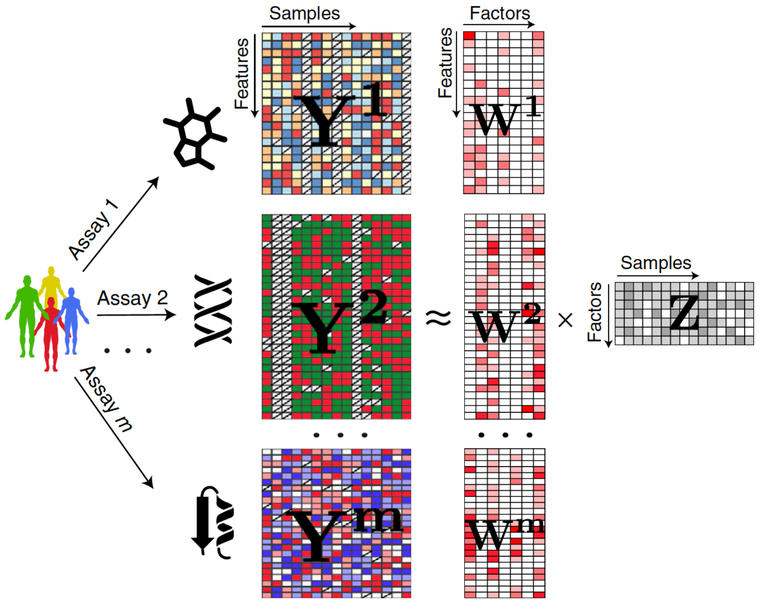
MOFA quantifies how much variance each factor explains in each view

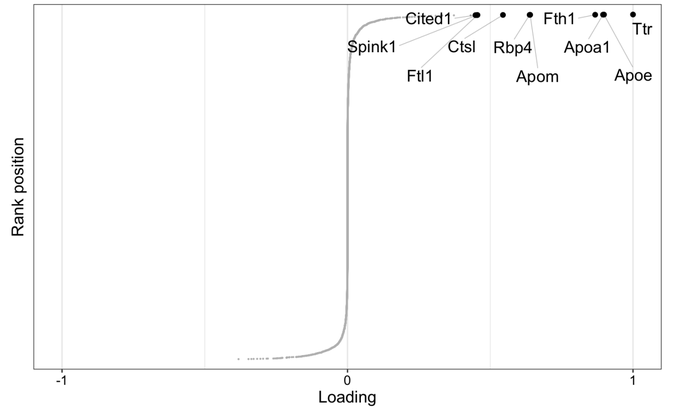


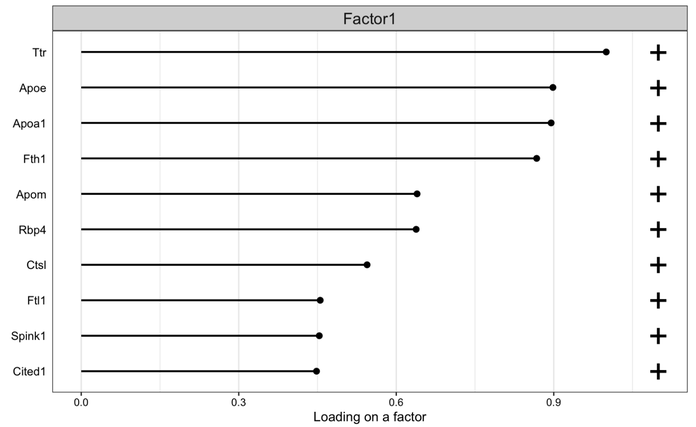



# Downstream analysis: Inspection of weights

Weights of a factor in each view can give insight into its molecular signature

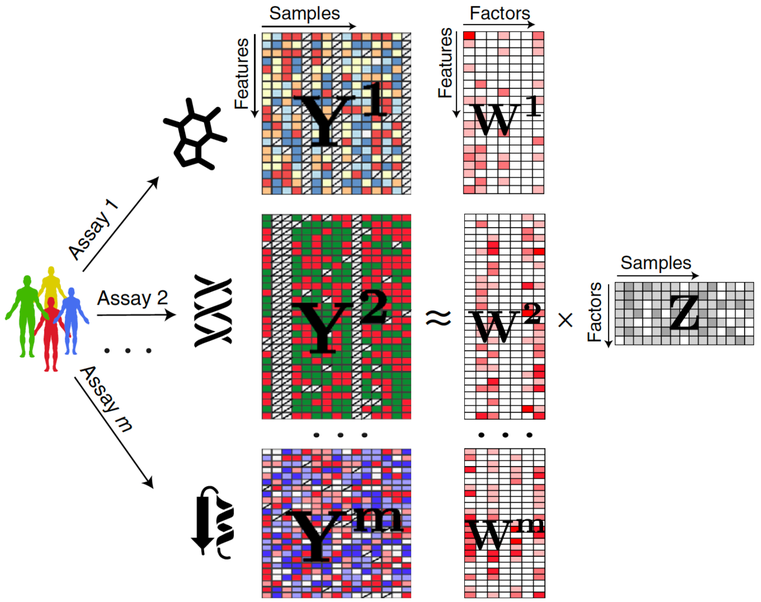


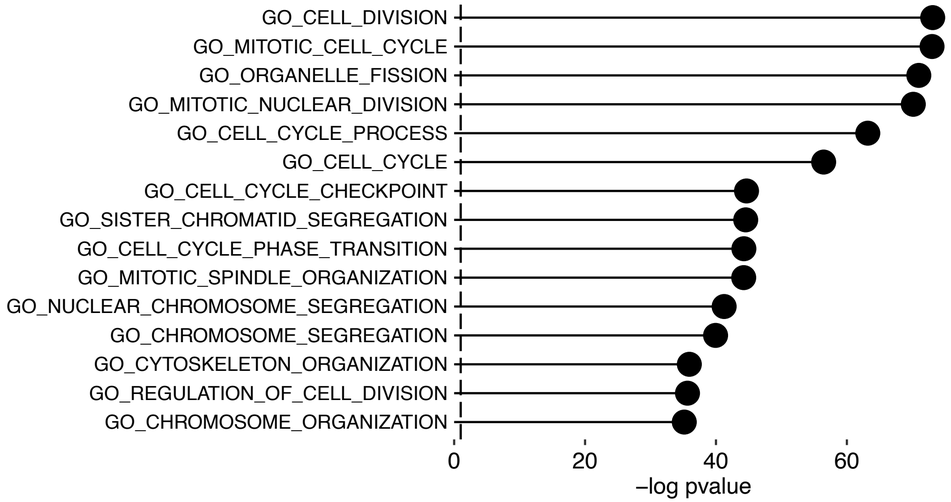



# Downstream analysis: Gene set enrichment analysis

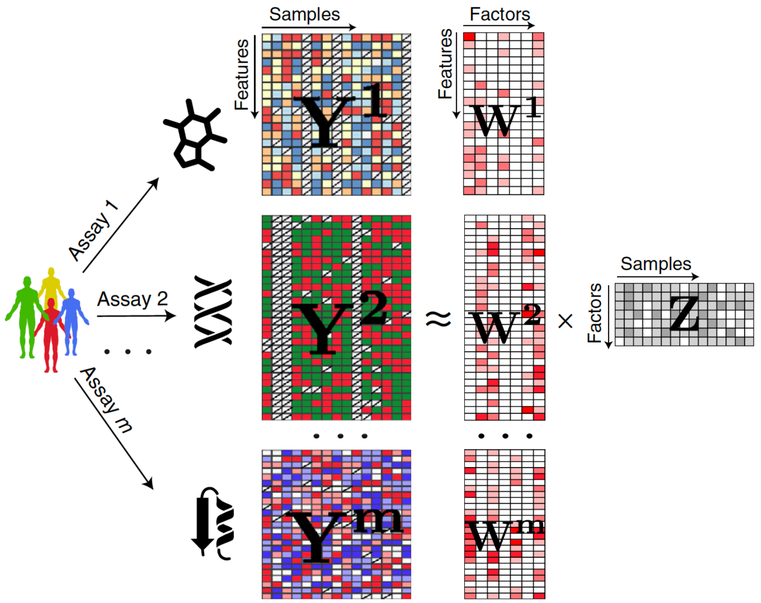
Enrichment analysis of the weights can be used to test for gene sets linked to a factor.

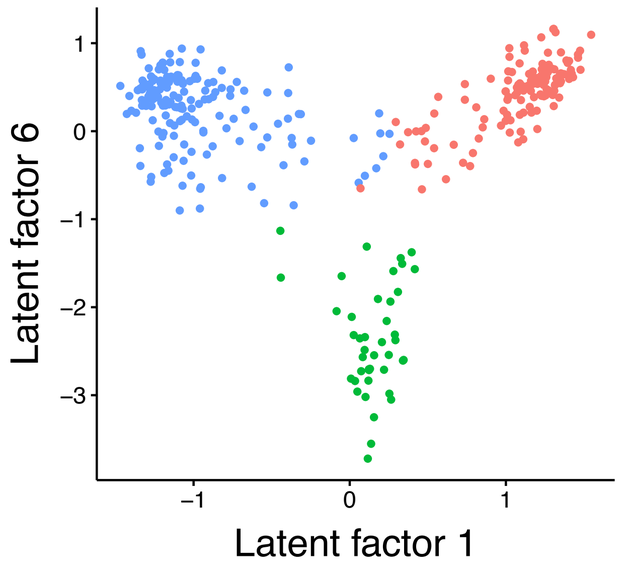


# Downstream analysis: Visualisation of samples in factor space

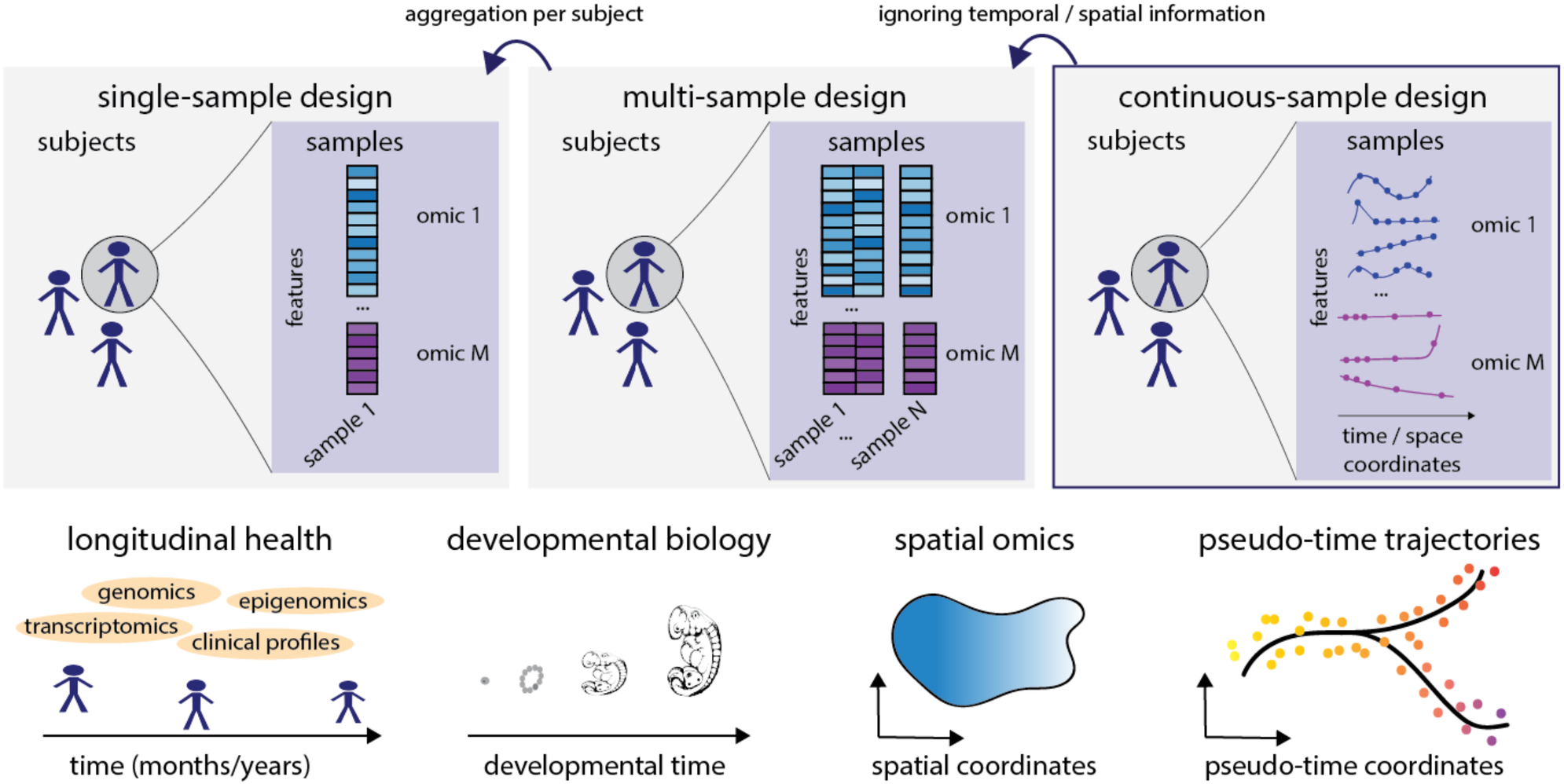
The factor space can be used to visualise or cluster samples or used as input for predictive models.



1. Motivation
2. Intuition and core idea
3. The maths behind MOFA
4. Guide for factor interpretation
5. MEFISTO
6. Case studies

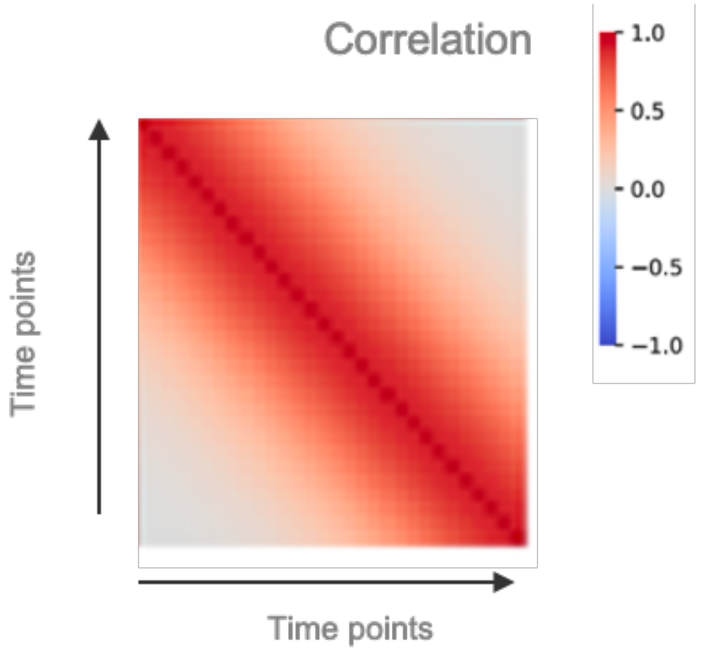
# Temporal and spatial resolution can give new insights into molecular dynamics and structures underlying biological processes

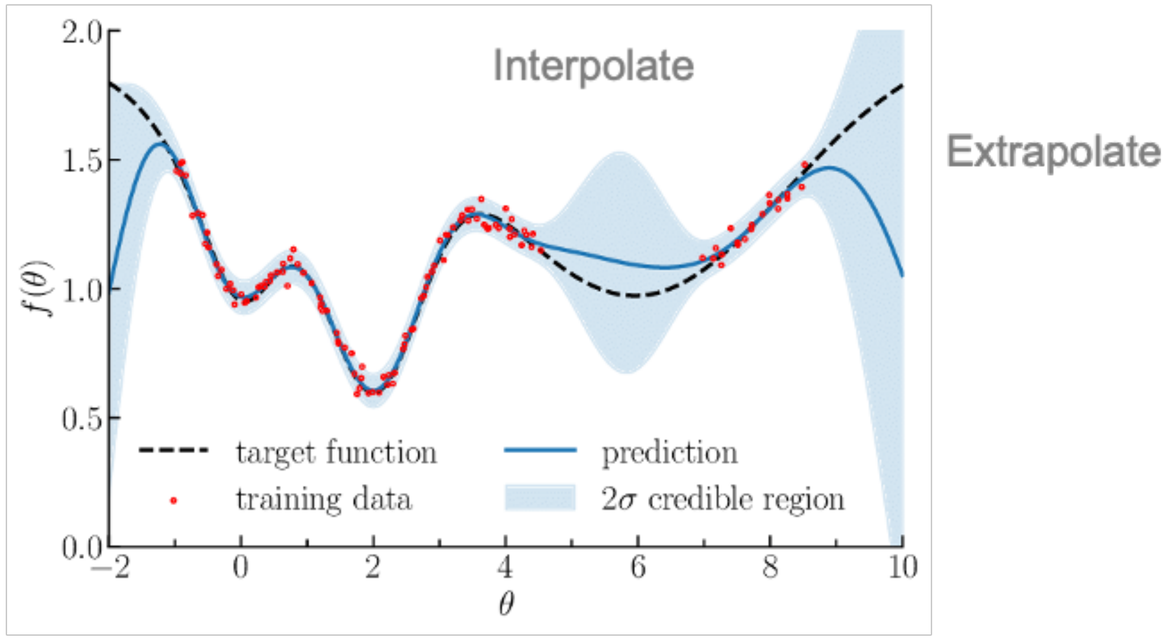



Velten & Stegle, Nath Methods 2023

# What is different in temporal and spatial data?

- Samples are **non-i.i.d**: Correlation across time points or spatial positions
- Assumptions of **smoothness** along time or space can help to
  - mitigate noise
  - interpolate or extrapolate to unmeasured time points or positions
- Need for **alignment** of temporal or spatial positions




Florent Leclercq, 2018

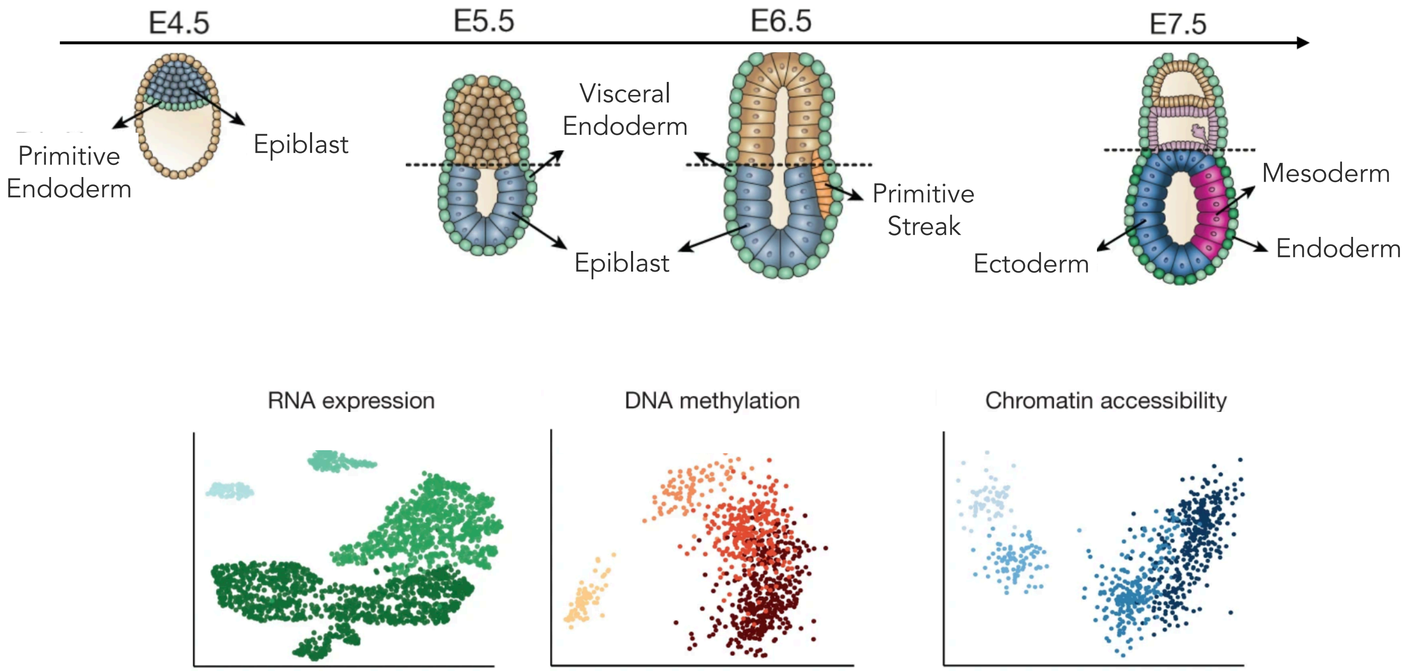
1. Motivation
2. Intuition and core idea
3. The maths behind MOFA
4. Guide for factor interpretation
5. MEFISTO
6. Case studies

Article | Published: 11 December 2019

# Multi-omics profiling of mouse gastrulation at single-cell resolution

Ricard Argelaguet, Stephen J. Clark ✉, Hisham Mohammed, L. Carine Stapel, Christel Krueger, Chantriolnt-Andreas Kapourani, Ivan Imaz-Rosshandler, Tim Lohoff, Yunlong Xiang, Courtney W. Hanna, Sebastien Smallwood, Ximena Ibarra-Soria, Florian Buettner, Guido Sanguinetti, Wei Xie, Felix Krueger, Berthold Göttgens, Peter J. Rugg-Gunn, Gavin Kelsey, Wendy Dean, Jennifer Nichols, Oliver Stegle ✉, John C. Marioni ✉ & Wolf Reik ✉

*Nature* **576**, 487–491 (2019) | Cite this article

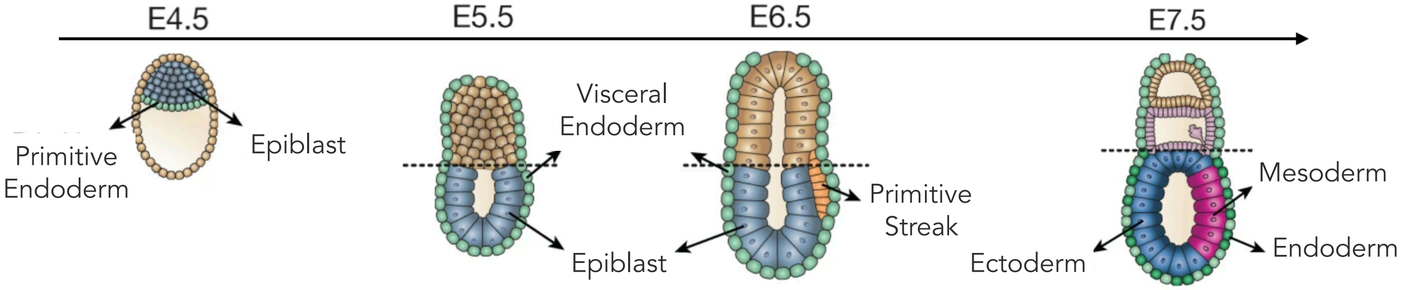


Article | Published: 11 December 2019

# Multi-omics profiling of mouse gastrulation at single-cell resolution

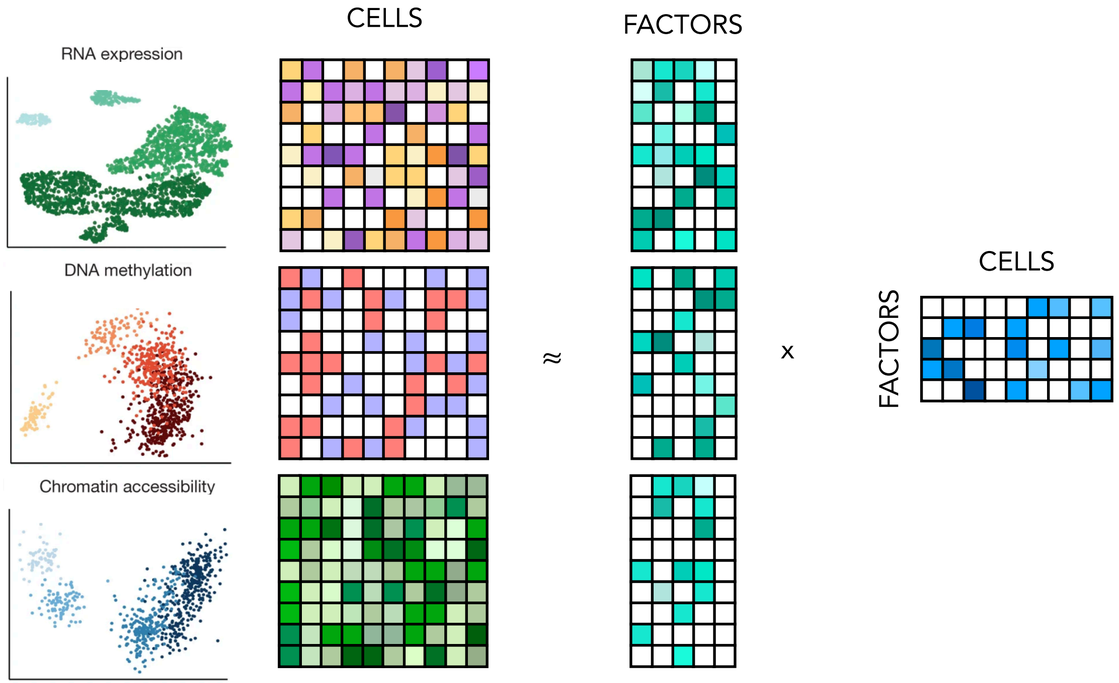
Ricard Argelaguet, Stephen J. Clark ✉, Hisham Mohammed, L. Carine Stapel, Christel Krueger, Chantriolnt-Andreas Kapourani, Ivan Imaz-Rosshandler, Tim Lohoff, Yunlong Xiang, Courtney W. Hanna, Sebastien Smallwood, Ximena Ibarra-Soria, Florian Buettner, Guido Sanguinetti, Wei Xie, Felix Krueger, Berthold Göttgens, Peter J. Rugg-Gunn, Gavin Kelsey, Wendy Dean, Jennifer Nichols, Oliver Stegle ✉, John C. Marioni ✉ & Wolf Reik ✉

*Nature* **576**, 487–491 (2019) | Cite this article




## 1. Multi-omics integration

**Goal**: uncover molecular coordination across omics and lineages



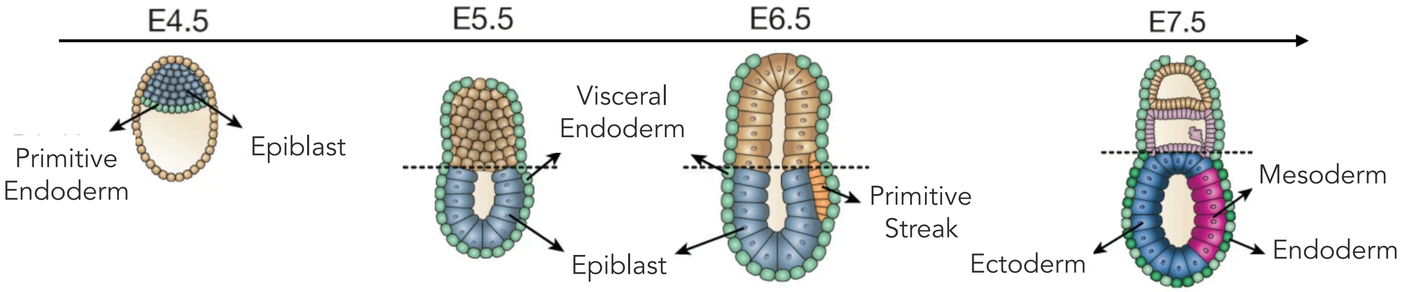

- Are certain factors shared across omics or specific to one modality?
- Which transcriptional programs are coordinated with epigenetic changes?
- Which cells are similar or distinct in their multi-omic profiles?

Article | Published: 11 December 2019

# Multi-omics profiling of mouse gastrulation at single-cell resolution

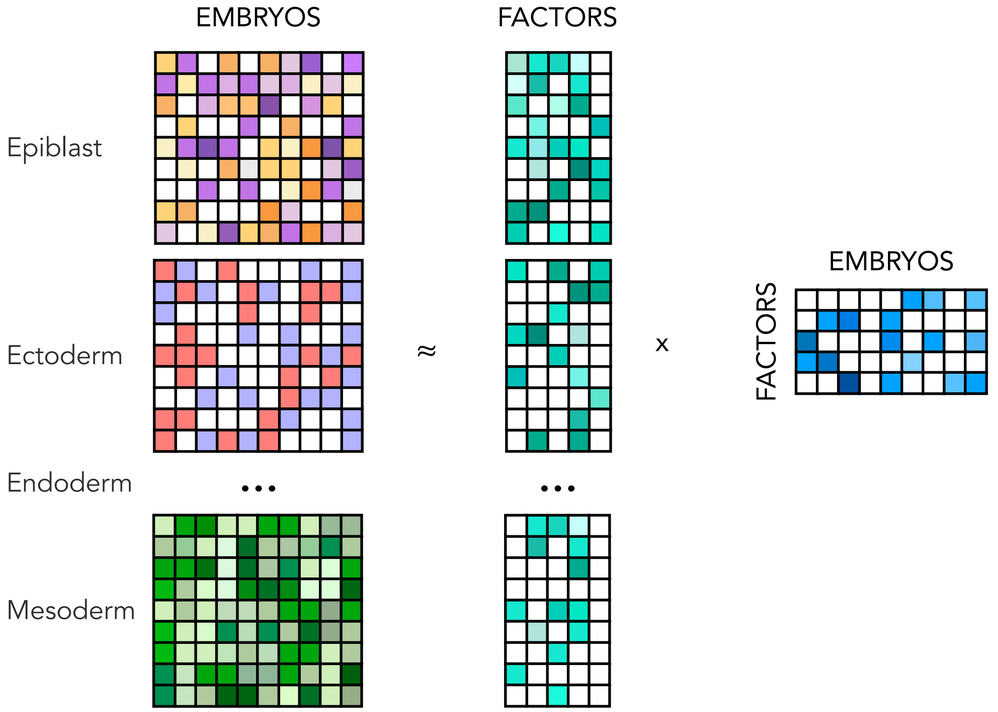
Ricard Argelaguet, Stephen J. Clark ✉, Hisham Mohammed, L. Carine Stapel, Christel Krueger, Chantriolnt-Andreas Kapourani, Ivan Imaz-Rosshandler, Tim Lohoff, Yunlong Xiang, Courtney W. Hanna, Sebastien Smallwood, Ximena Ibarra-Soria, Florian Buettner, Guido Sanguinetti, Wei Xie, Felix Krueger, Berthold Göttgens, Peter J. Rugg-Gunn, Gavin Kelsey, Wendy Dean, Jennifer Nichols, Oliver Stegle ✉, John C. Marioni ✉ & Wolf Reik ✉

*Nature* **576**, 487–491 (2019) | Cite this article




## 2. Cell lineage integration

**Goal**: uncover embryo-level programs across lineages: cross-lineage coordination



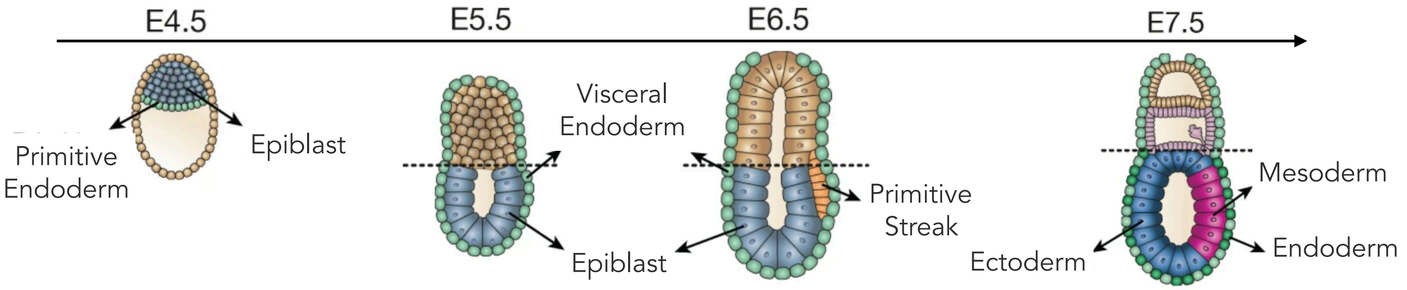

- Which embryo-level transcriptional programs are shared across multiple lineages?
- Which programs are lineage-specific, indicating unique differentiation pathways in particular lineages?
- How do these programs vary from embryo to embryo, and can we detect embryonic outliers or abnormal lineage patterns?

Article | Published: 11 December 2019

# Multi-omics profiling of mouse gastrulation at single-cell resolution

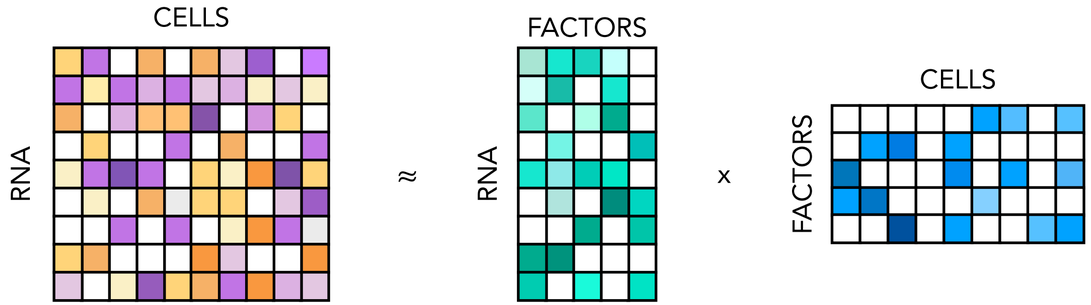
Ricard Argelaguet, Stephen J. Clark ✉, Hisham Mohammed, L. Carine Stapel, Christel Krueger, Chantriolnt-Andreas Kapourani, Ivan Imaz-Rosshandler, Tim Lohoff, Yunlong Xiang, Courtney W. Hanna, Sebastien Smallwood, Ximena Ibarra-Soria, Florian Buettner, Guido Sanguinetti, Wei Xie, Felix Krueger, Berthold Göttgens, Peter J. Rugg-Gunn, Gavin Kelsey, Wendy Dean, Jennifer Nichols, Oliver Stegle ✉, John C. Marioni ✉ & Wolf Reik ✉

*Nature* **576**, 487–491 (2019) | Cite this article




## 3. Stage modeling with MEFISTO

**Goal**: uncover smooth developmental programs across stages




- Which transcriptional programs evolve smoothly over development?
- Which genes drive each latent trajectory?