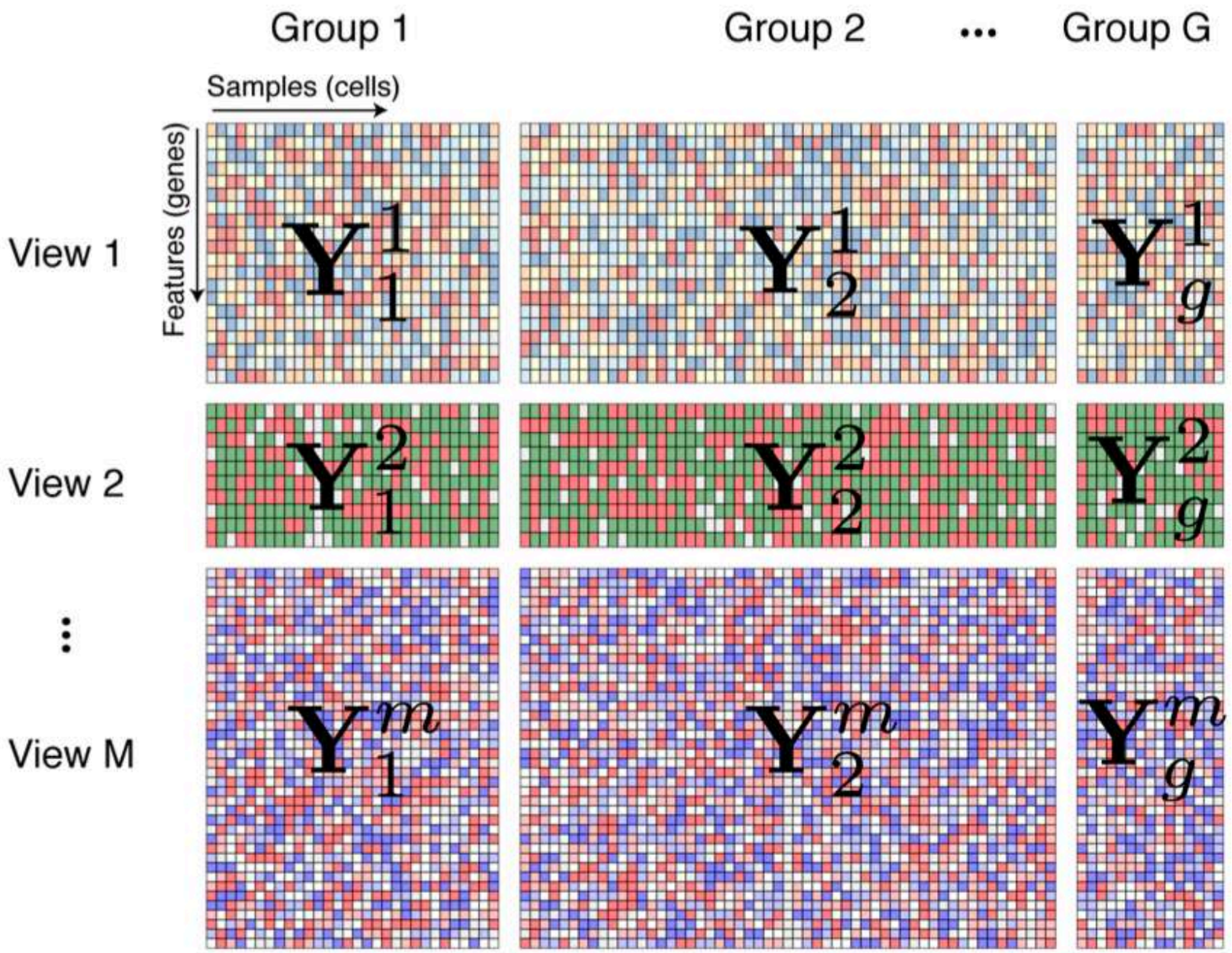
Multi-Omics Factor Analysis (MOFA)

A general framework for the unsupervised integration of multi-omic data sets

Ricard Argelaguet
PhD student at EMBL/EBI
(Marioni/Stegle Group)

ricard@ebi.ac.uk

Complex experimental designs yield large structured data sets



Challenges in multi-omics integration

- Data collected using different techniques (i.e. data modalities) generally exhibit heterogeneous statistical properties
- Large amounts (and different patterns) of missing values
- Undesired sources of heterogeneity
- Overfitting
- Complexity of the data requires unsupervised interpretable approaches

Latent variable models

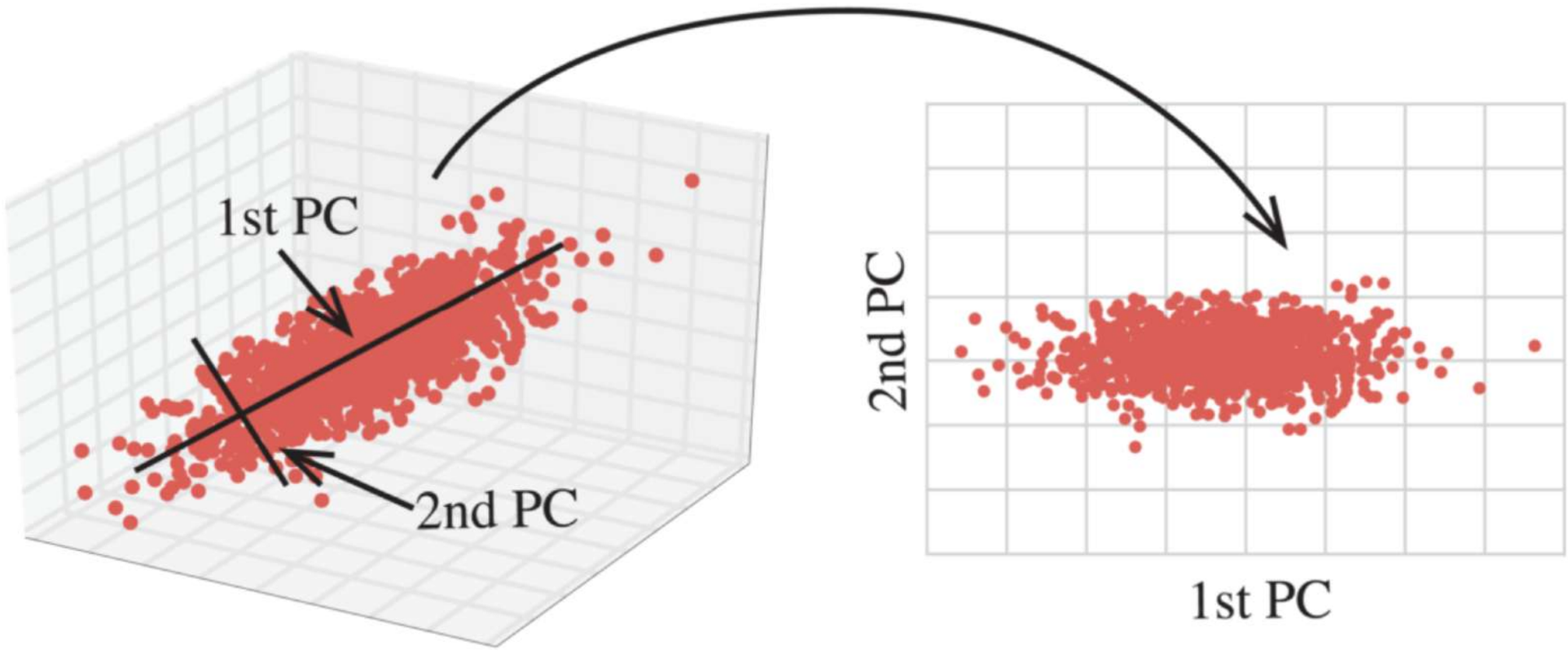
Y

Observed variables

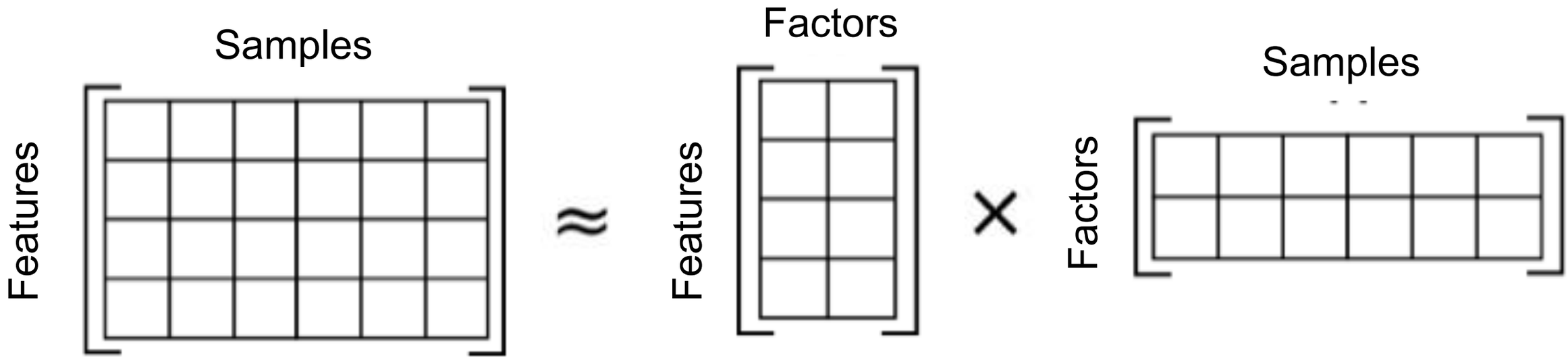
Z

Latent variables

f(x)



Matrix factorisation (MF)



Y **W** **Z**

Y are the observed measurements

W are the inferred feature weights

Z are the inferred latent factors

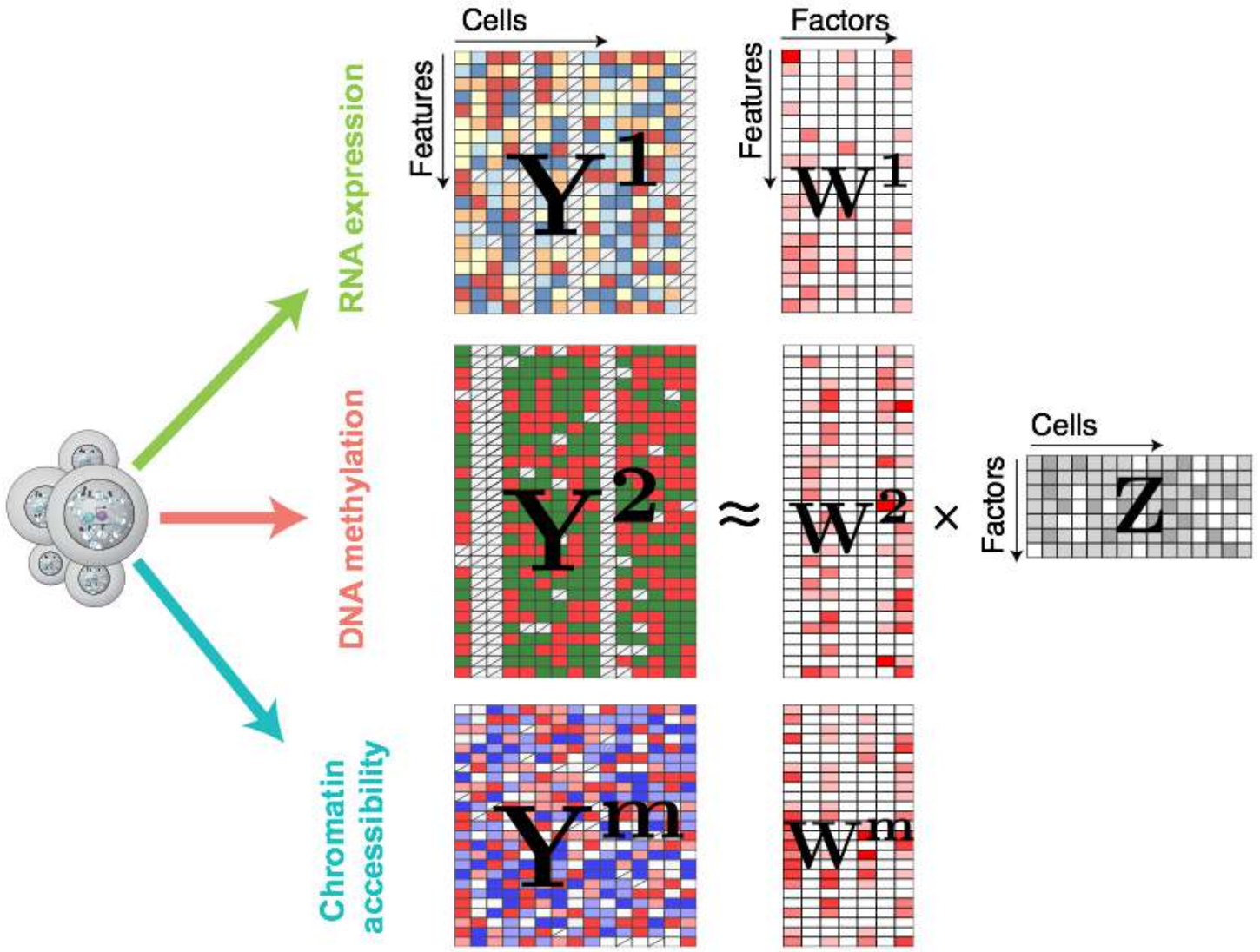
Principal Component Analysis is an instance of matrix factorisation!

Problems of using standard PCA in multi-omics data

PCA is a great method to understand the sources of variation in a single data modality, but it has problems in a multi-omics setting:

- No clear way to measure how much variance the PCs explain in each data modality
- No natural way to combine different data modalities (binary data with continuous data)
- Can’t handle missing values
- Non-sparse solutions: challenges in interpretability and risk of overfitting

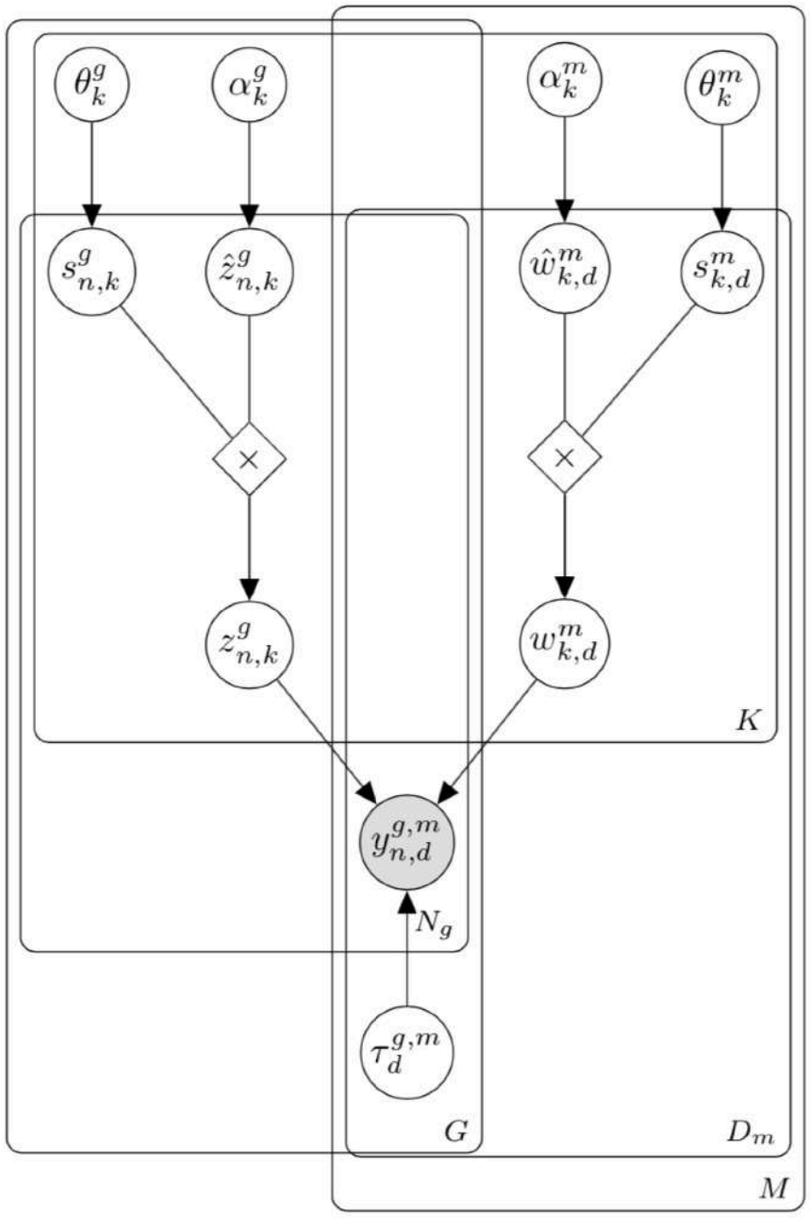
Multi-omics Factor Analysis (MOFA)



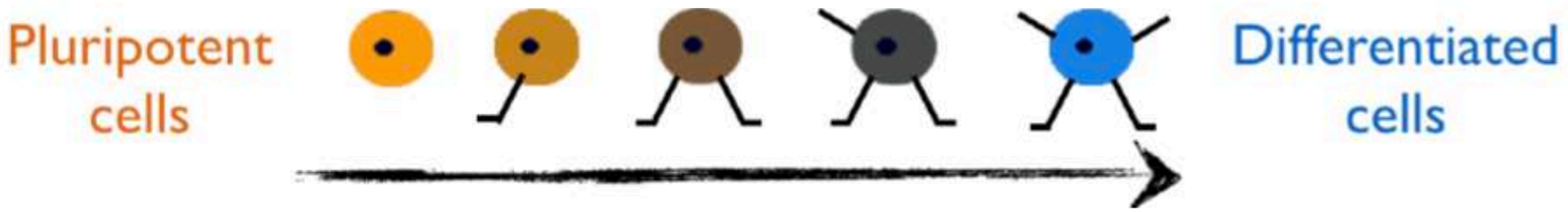
$$\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^{mT}$$

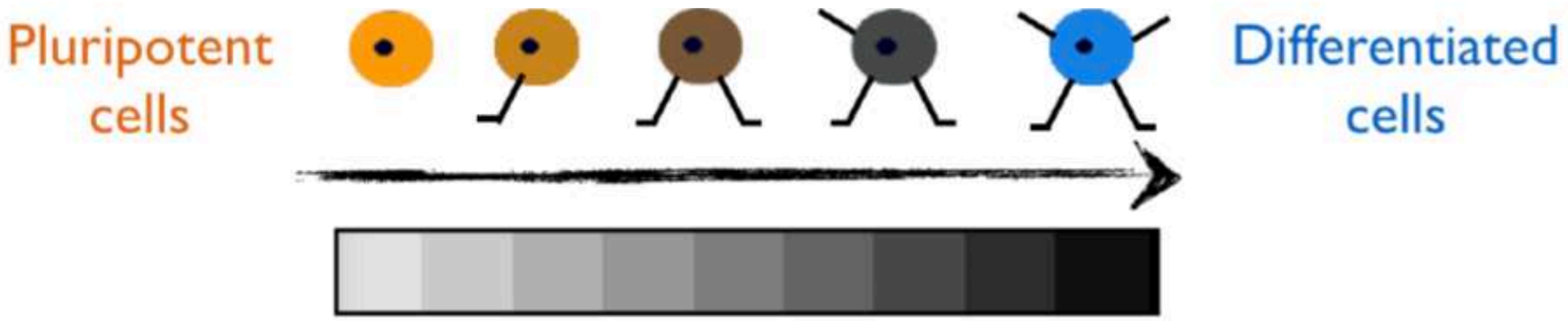
- The structure of the data is specified in the prior distributions of the Bayesian model
- The critical part of the model is the use sparsity priors, which enable automatic relevance determination of the factors
- Inference is performed using (fast) variational Bayes

Bayesian graphical model



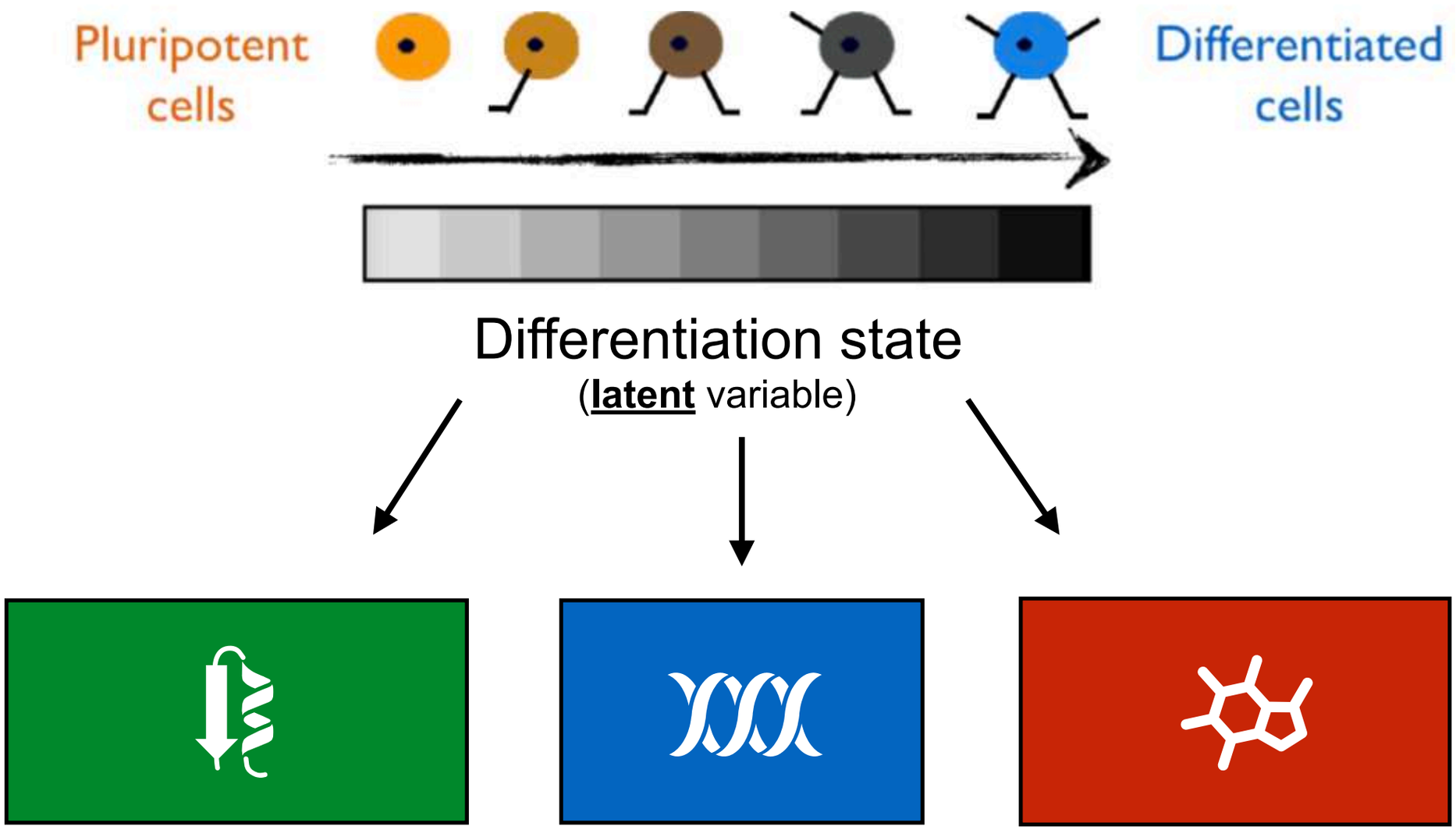
- Bayesian multi-view factor analysis framework
- Automatic Relevance Determination prior for the weights (view-wise)
- Spike and slab prior for the weights (feature-wise)


Pluripotent cells
Differentiated cells

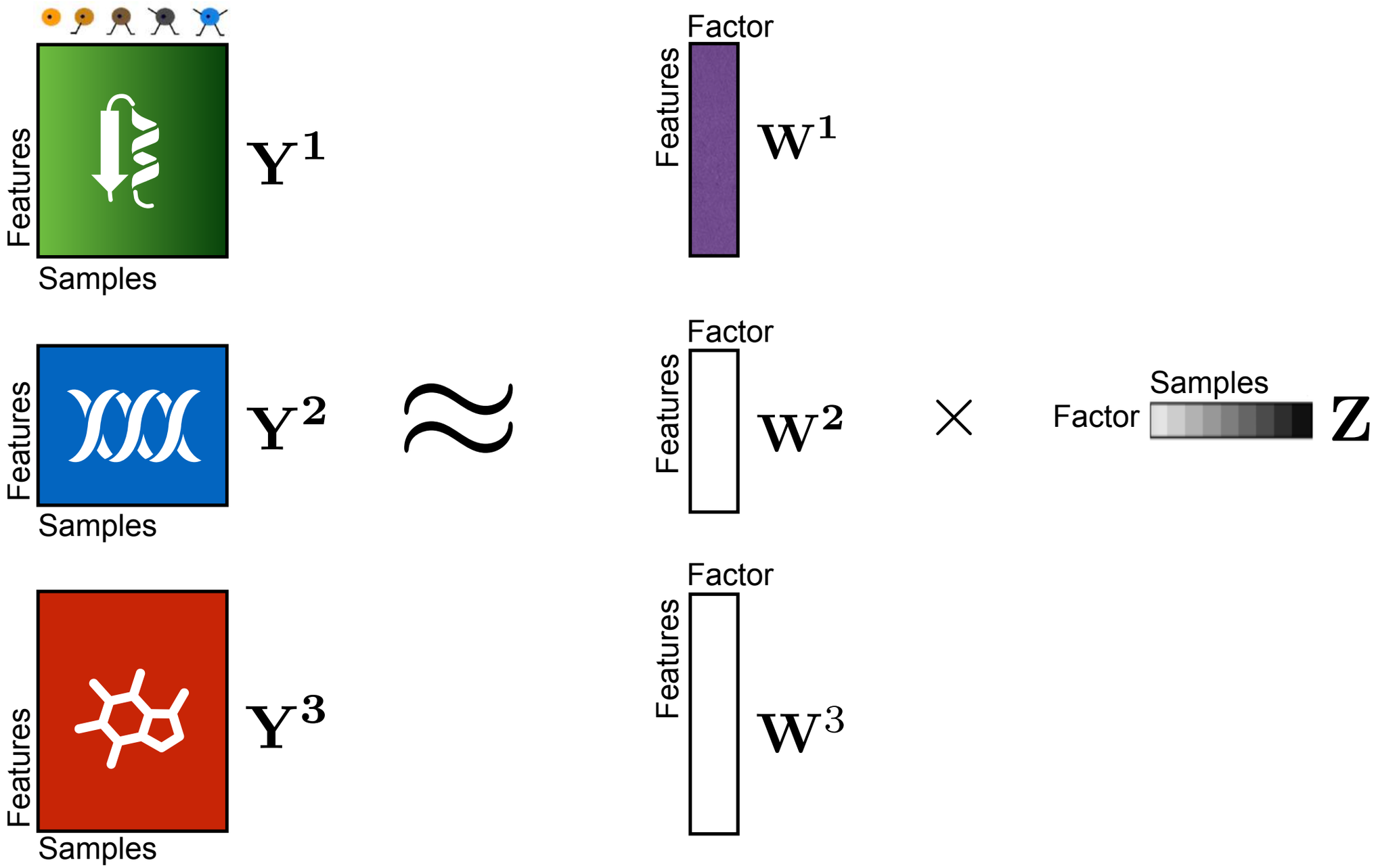


Differentiation state

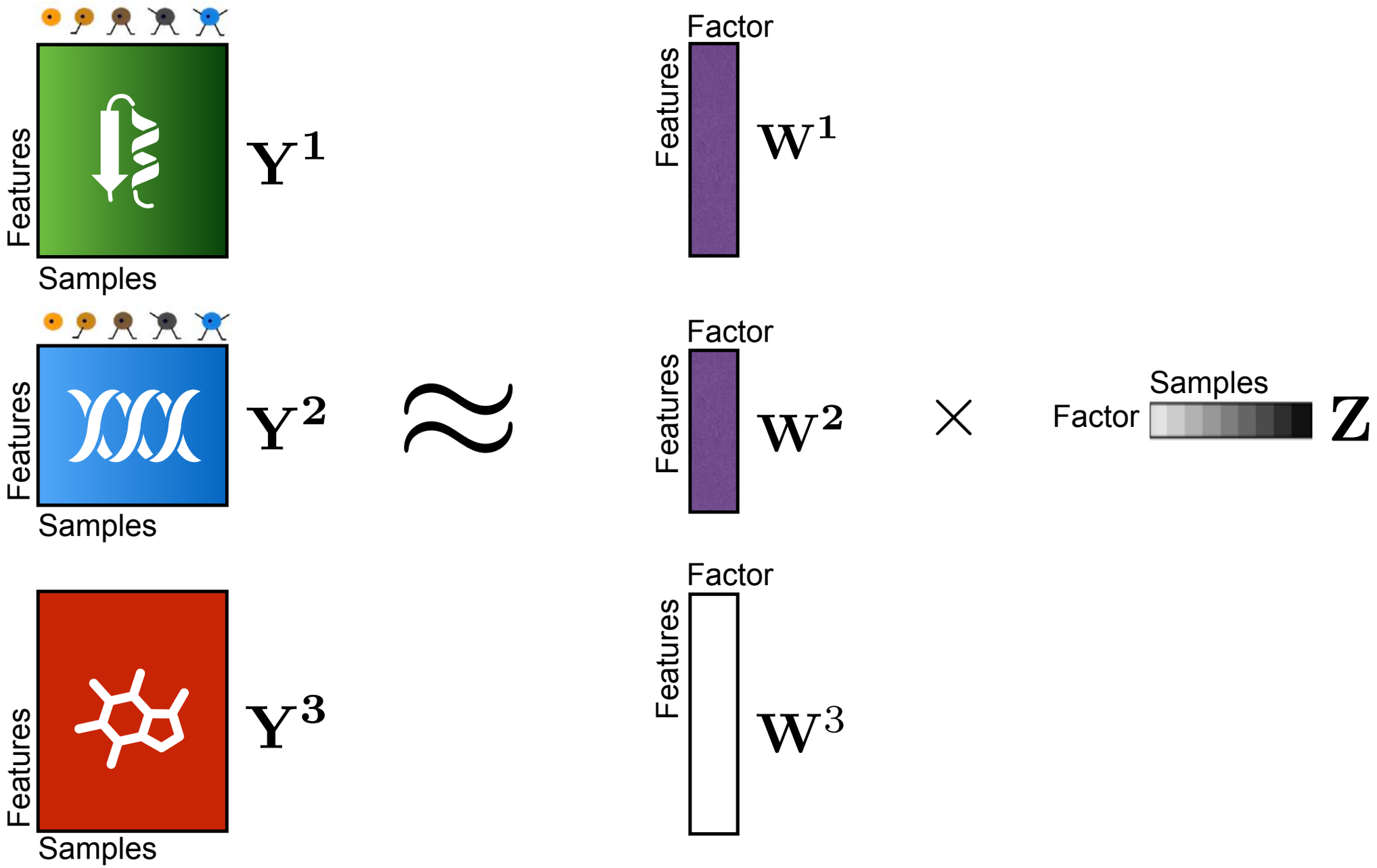
(hidden/latent variable)


Pluripotent cells
Differentiated cells
Differentiation state
(latent variable)
Omics
(observed variables)

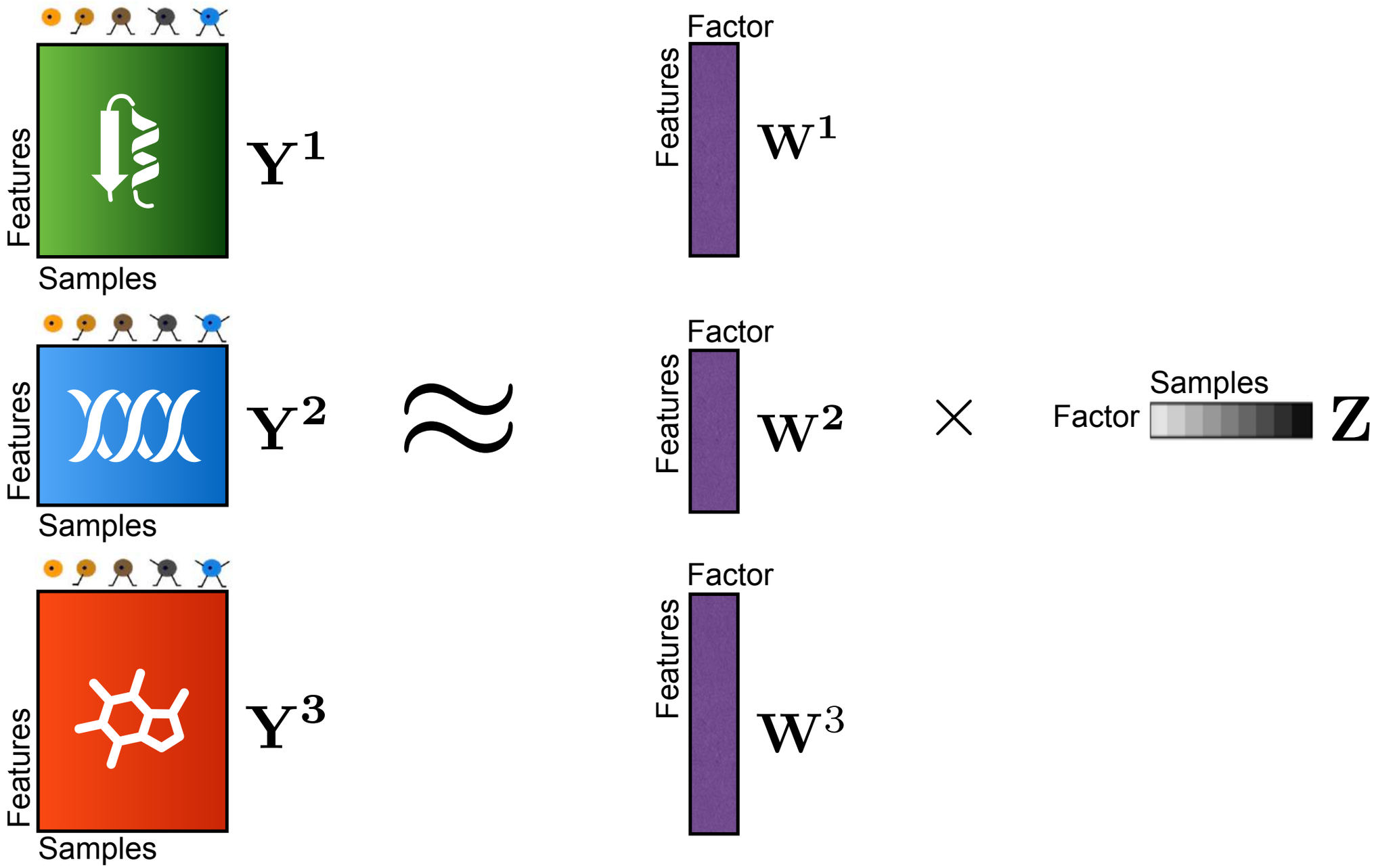
The differentiation state is the only driver of variation in **transcriptomics**



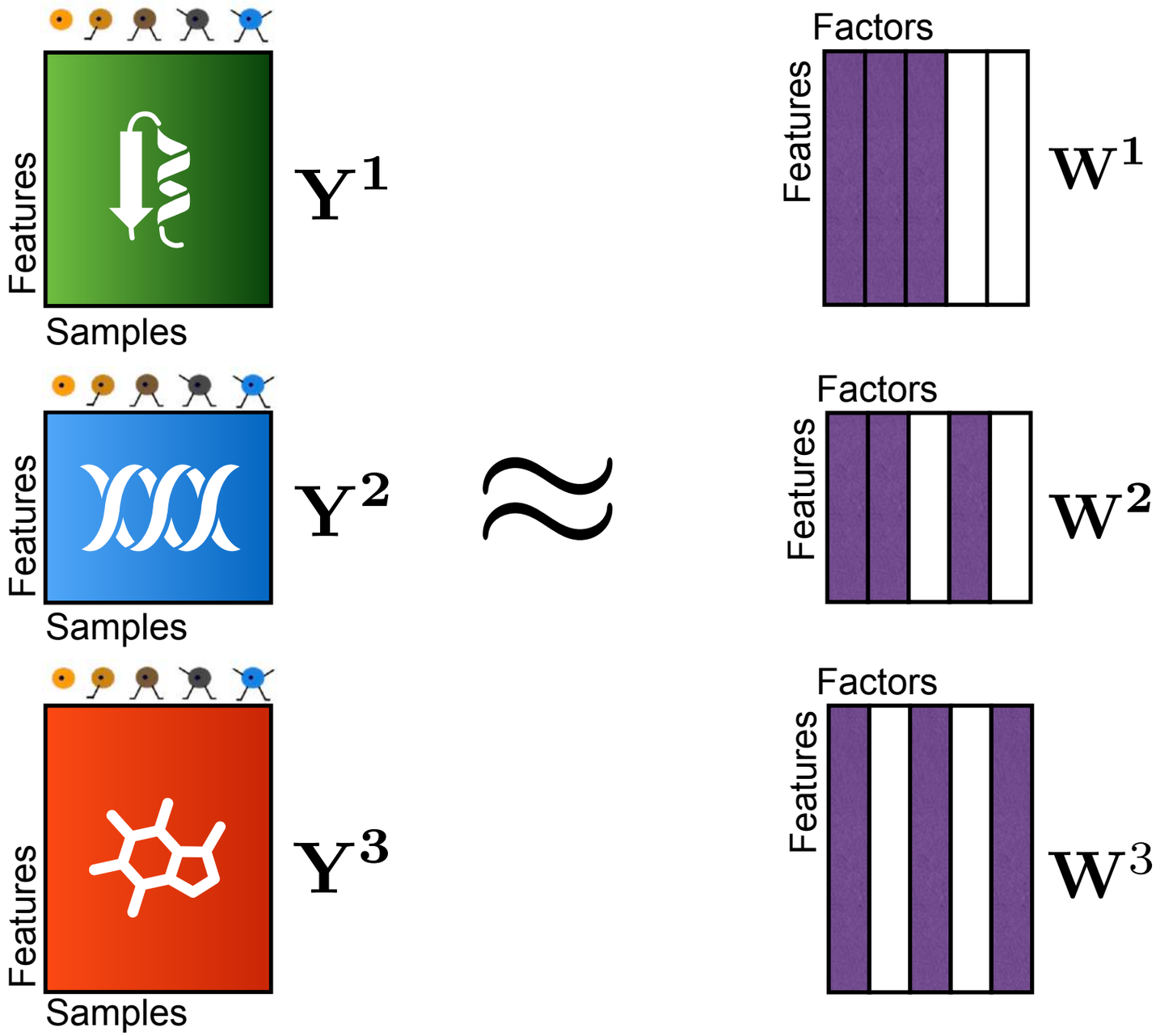
The differentiation state is the only driver of variation in **transcriptomics and genetics**



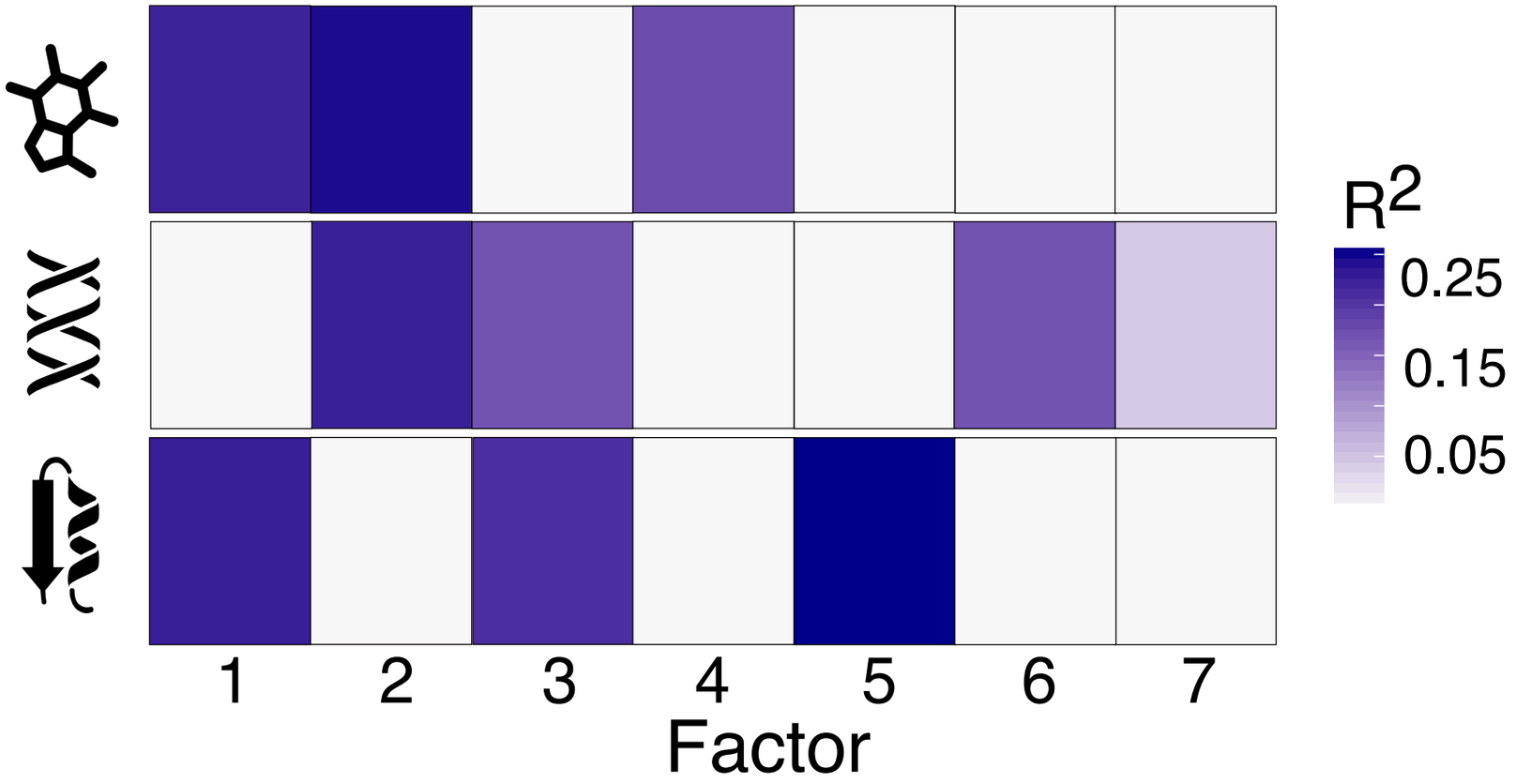
The differentiation state is the only driver of variation in **all omics**



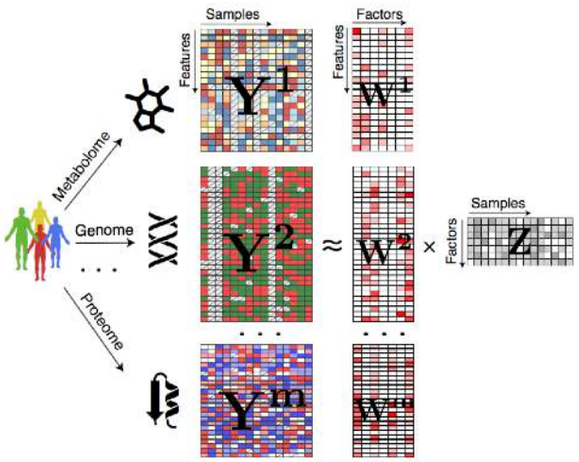
The differentiation state is the only driver of variation in **all omics**



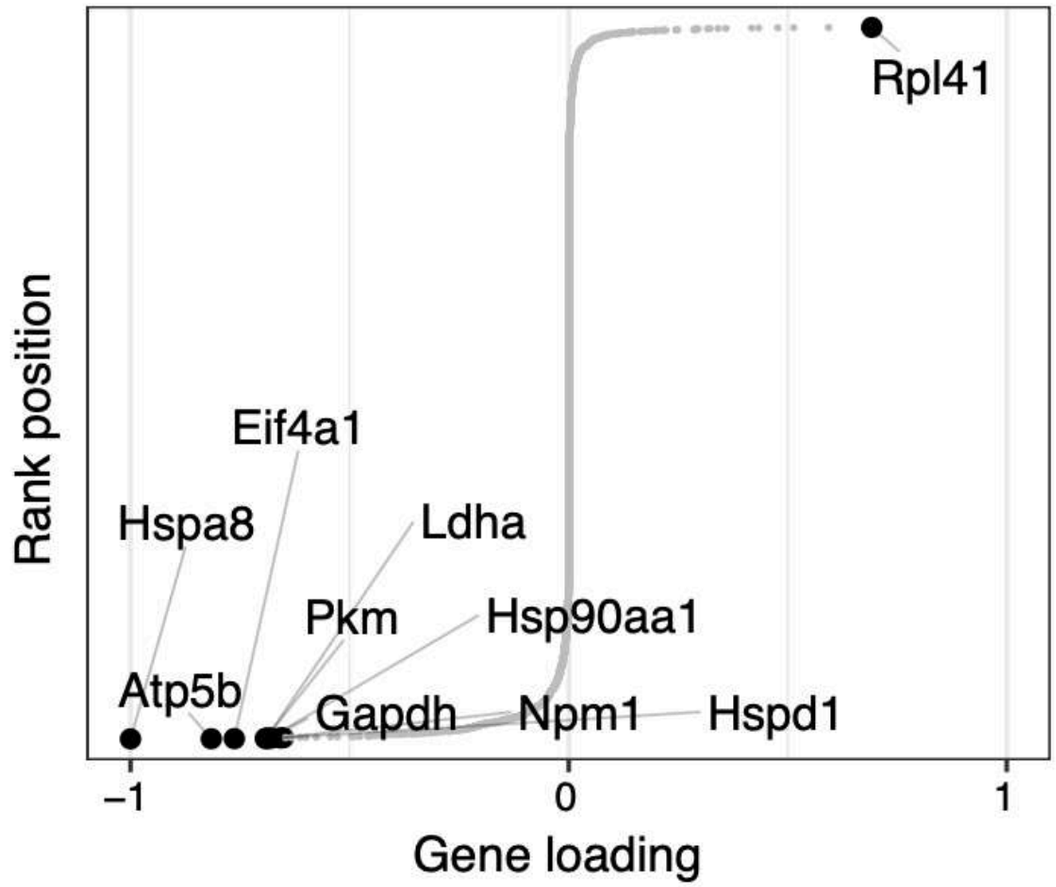
Variance decomposition by factor



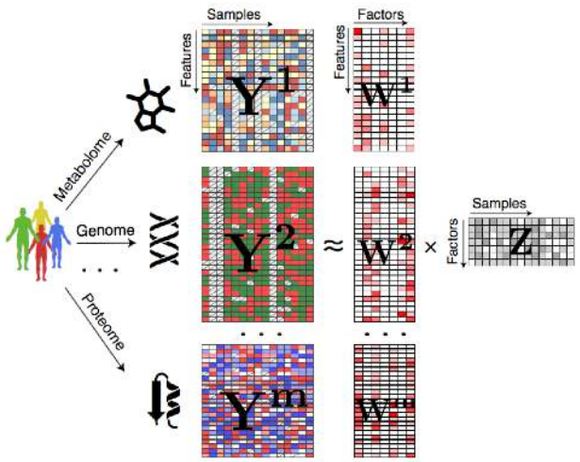
Downstream analysis

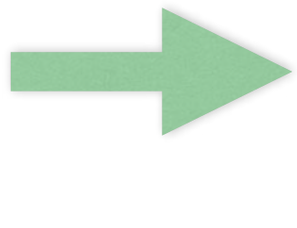


Inspection of feature weights

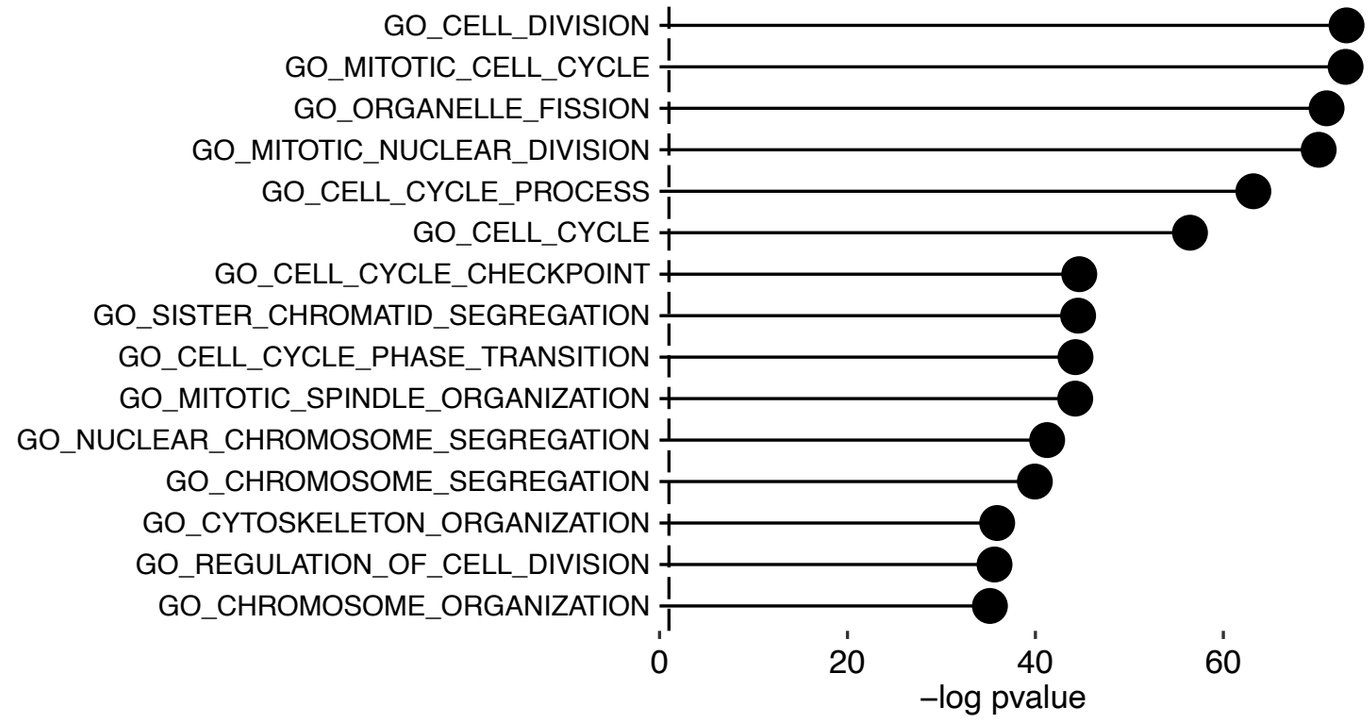


Downstream analysis

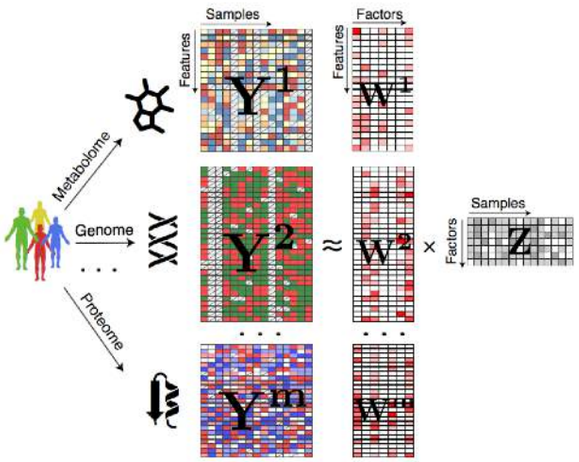


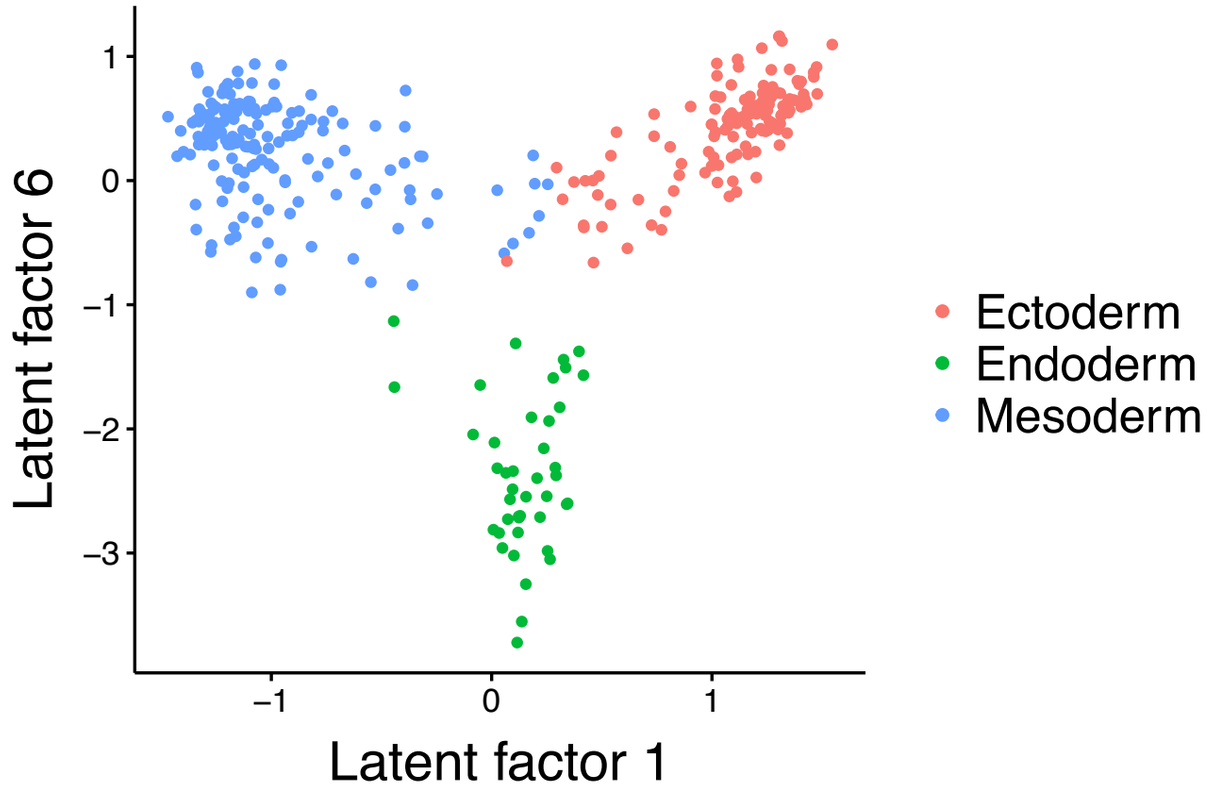


Gene set enrichment analysis

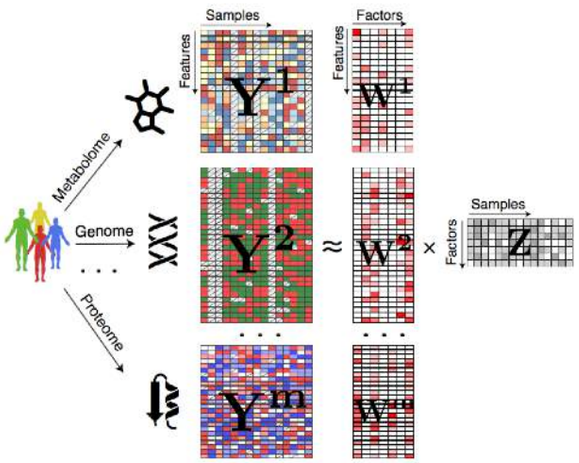


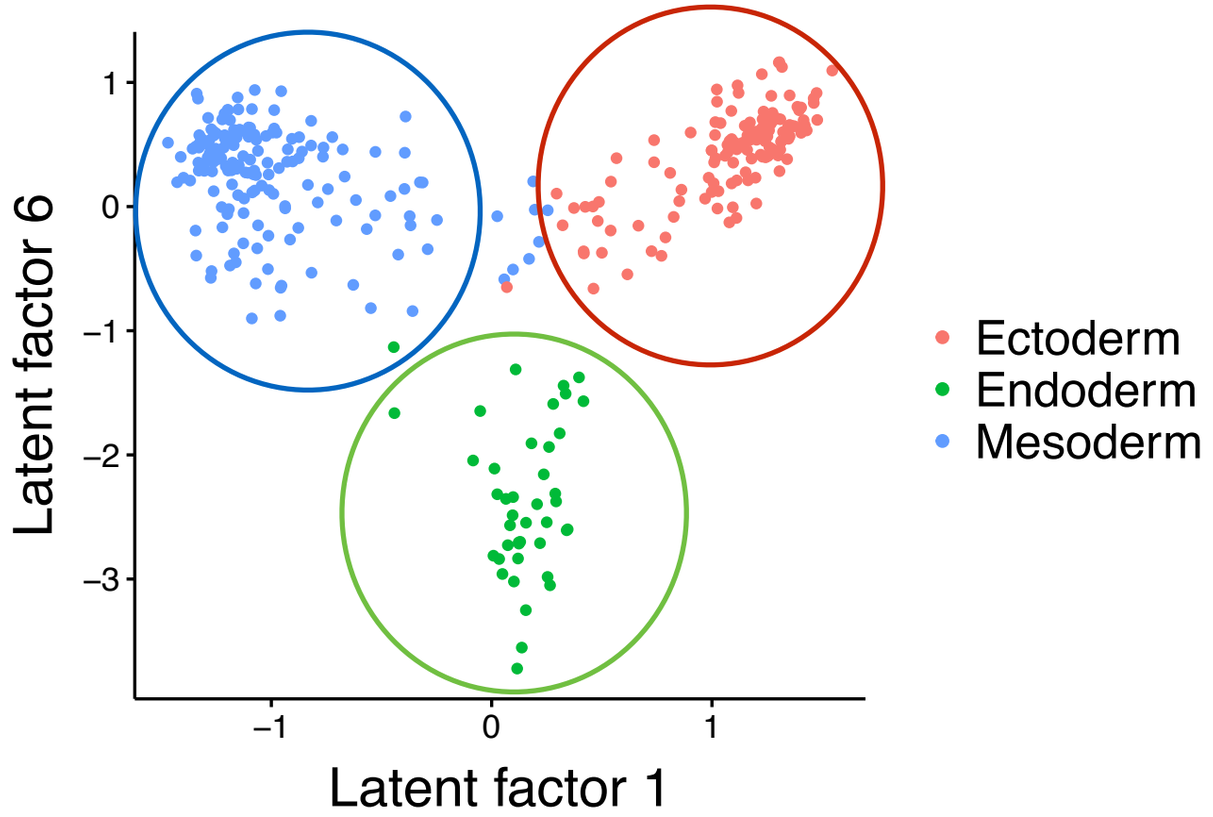
Downstream analysis



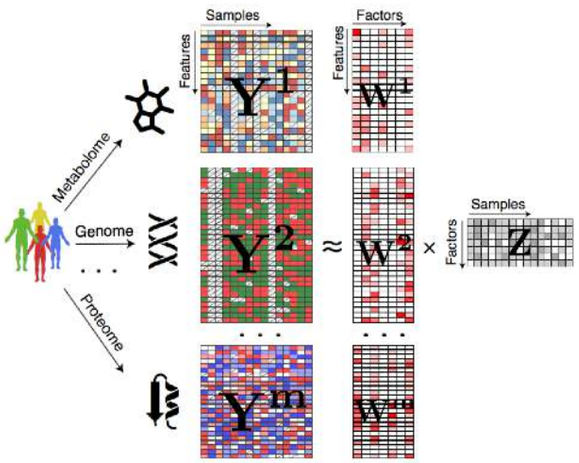


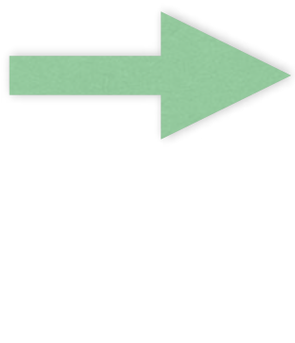
Downstream analysis



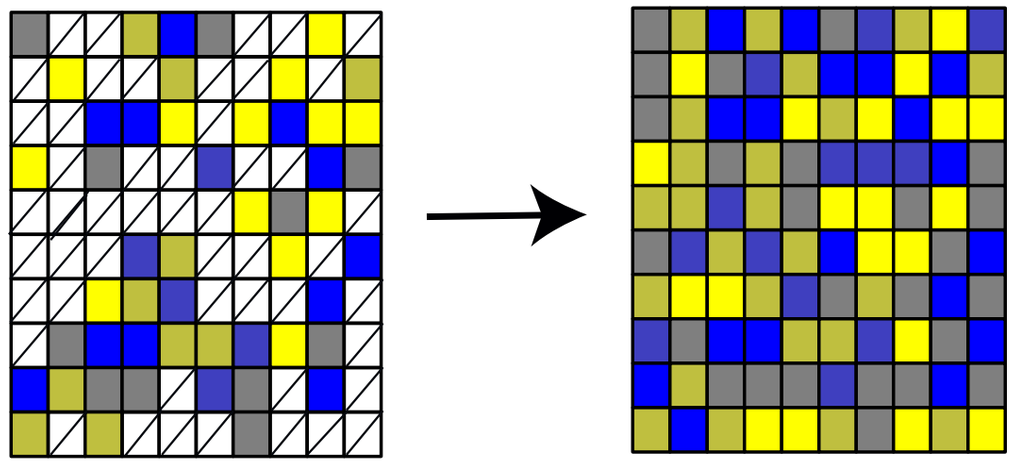


Downstream analysis



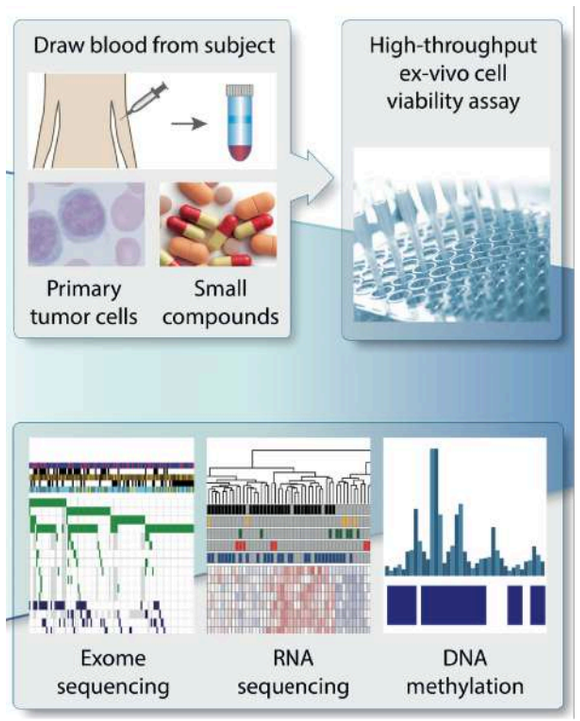


Imputation of missing values

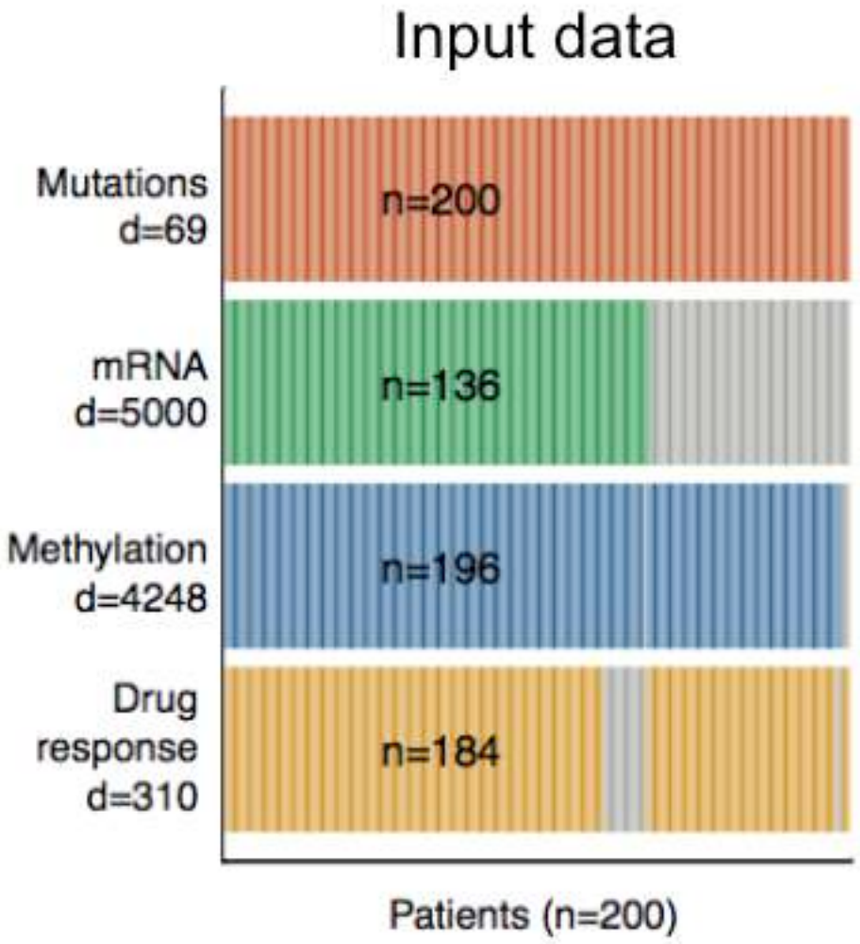


$$\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^{mT}$$

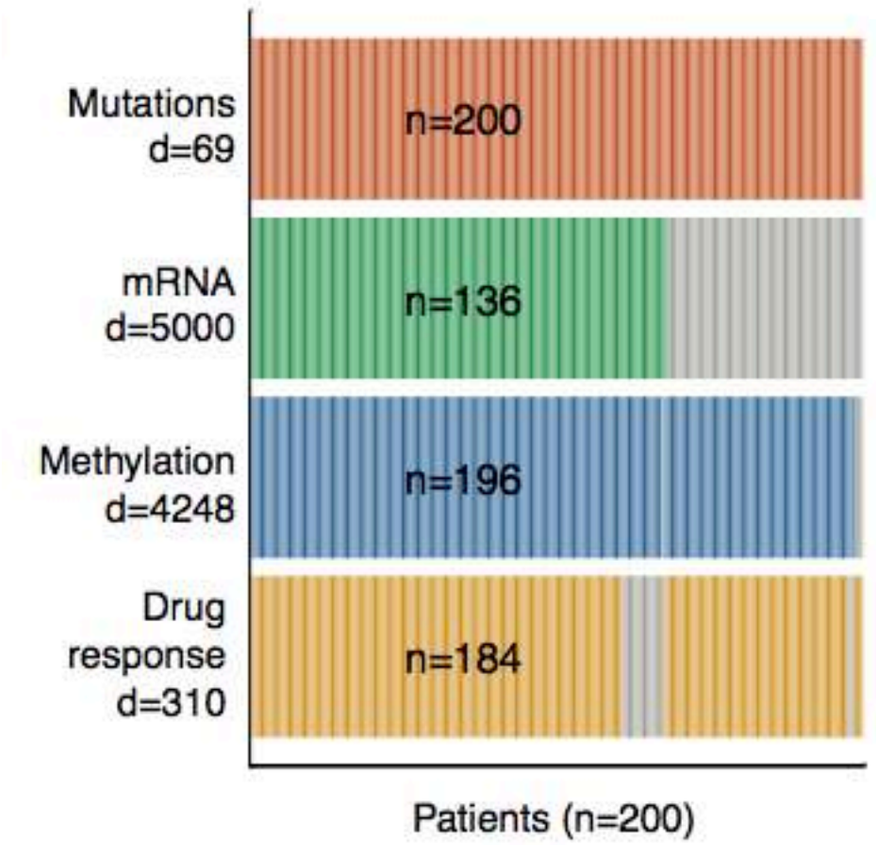
Application to Chronic Lymphocitic Leukaemia



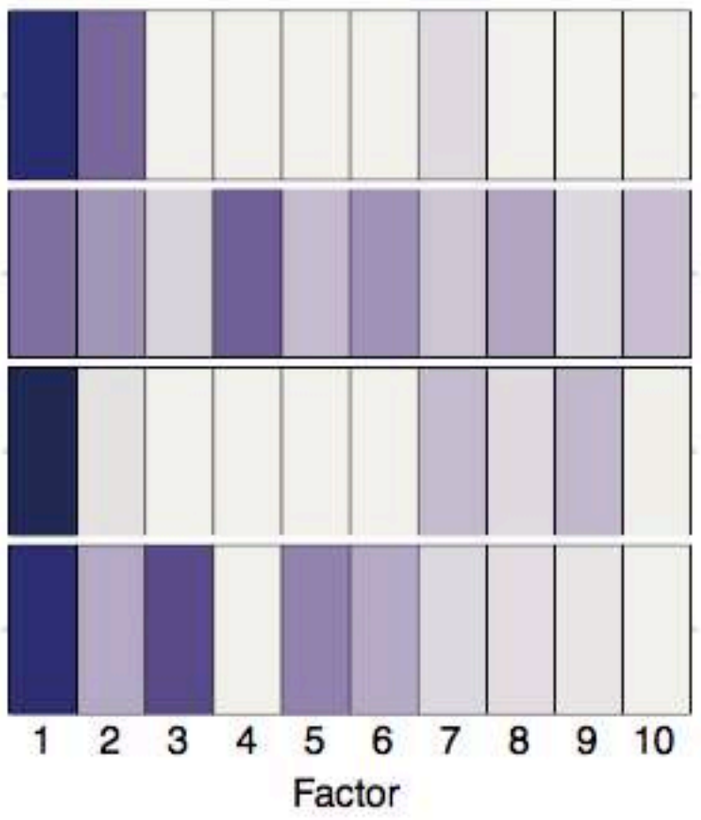
Thorsten Zenz group (Heidelberg)



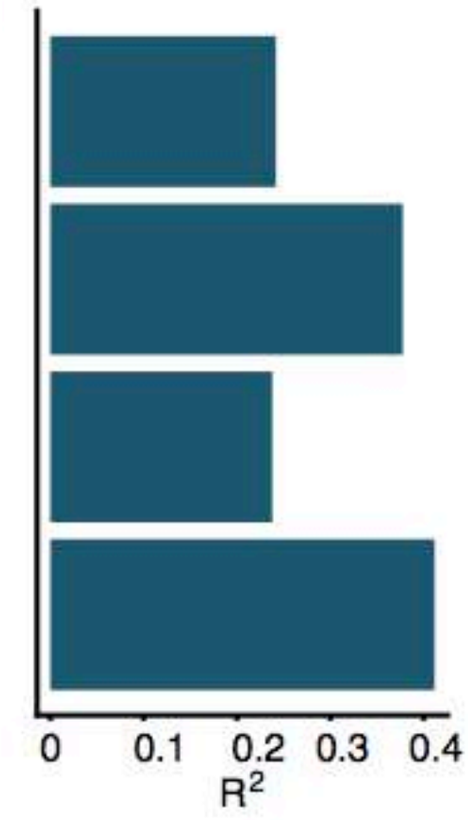
Application to Chronic Lymphocitic Leukaemia



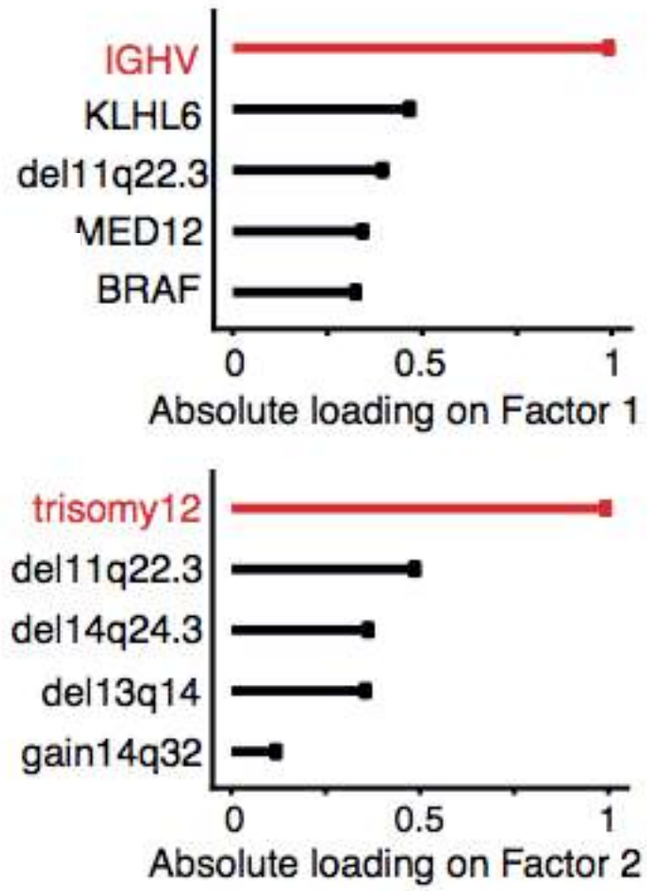
Variance explained per factor and view



Variance explained per view



Inspection of feature weights for Factors 1 and 2



American Journal of Hematology AJH

Explore this journal >

CLINICAL PEARLS IN BLOOD DISEASES

IGHV mutational status testing in chronic lymphocytic leukemia

Jennifer Crombie, Matthew S. Davids

Trisomy 12 chronic lymphocytic leukemia cells

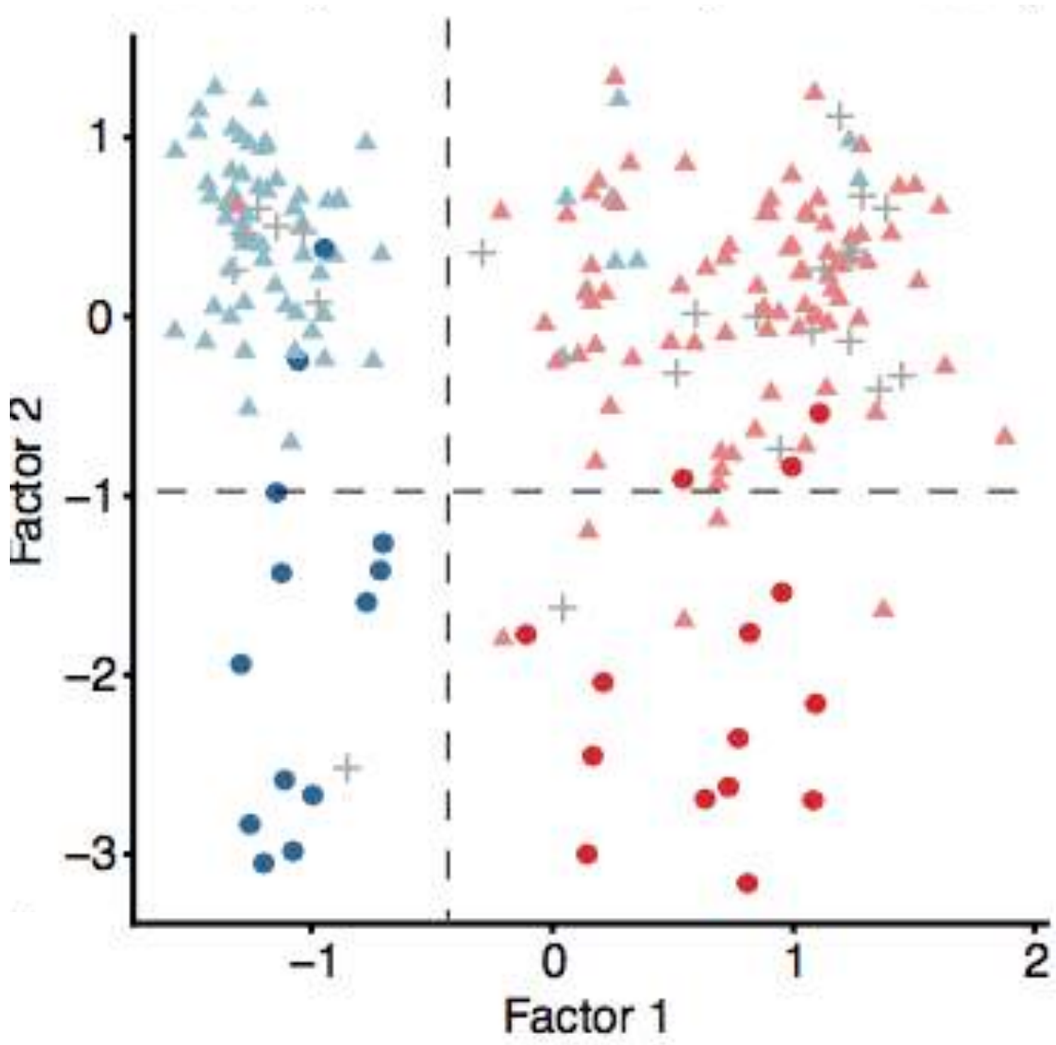
John C. Riches, Conor J. O'Donovan, Sarah J. Kingdon, Fabienne McClanahan, Andrew J. Clear, Laura Z. Rassenti, Thomas J. Kipps, and John G. Gribben

Blood 2014 123:4101-4110; doi: https://doi.org/10.1182/blood-2014-01-552307

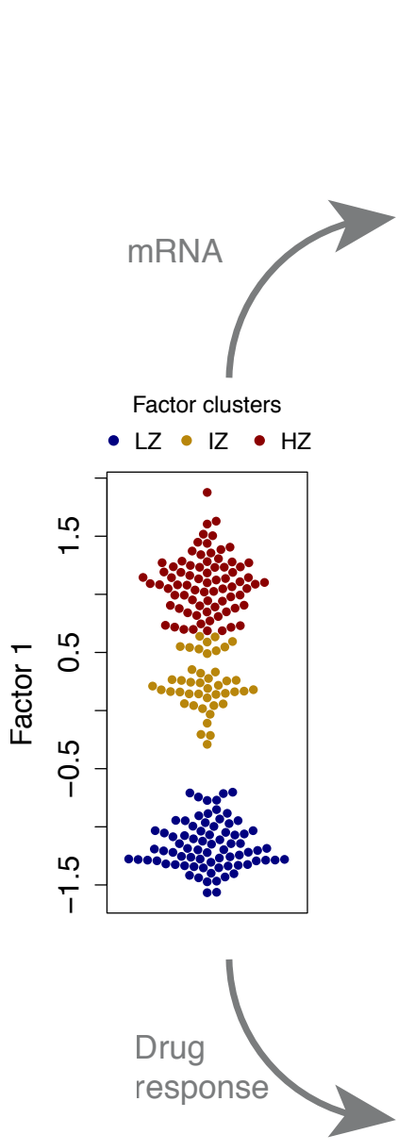
IGHV: *Immunoglobulin heavy chain variable region*

Visualisation of samples in the latent space

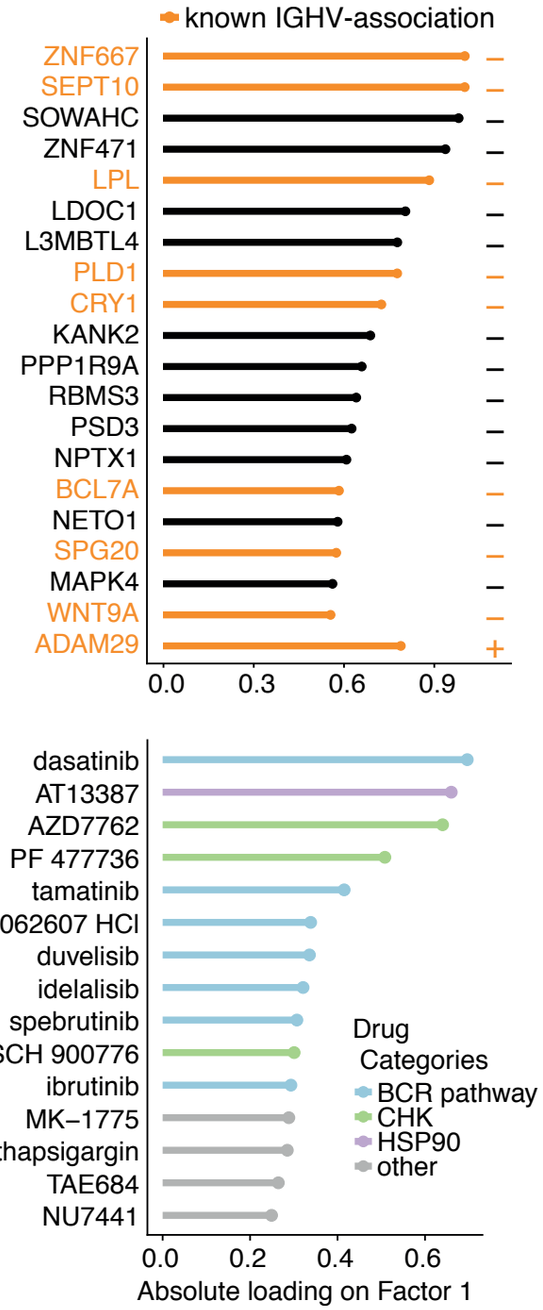
Factor 1: IGHV+ vs IGHV-
Factor 2: tr12+ (●) vs tr12- (▲)

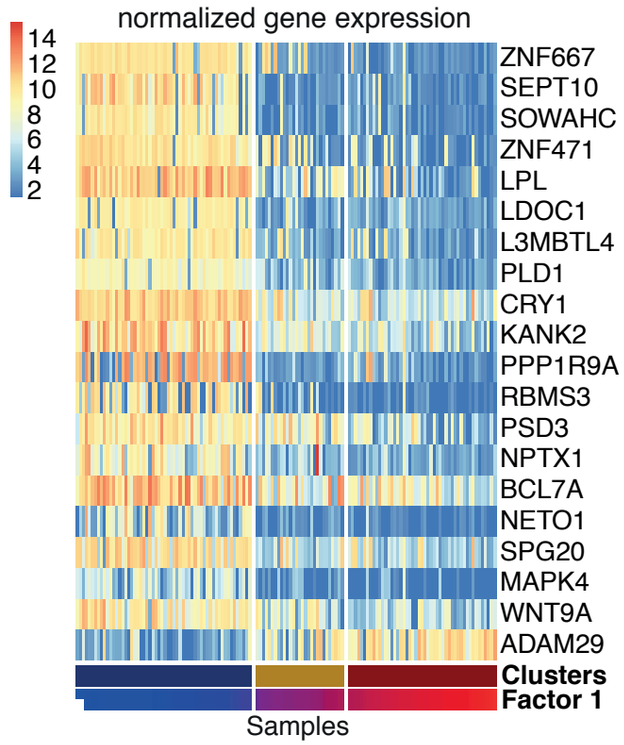


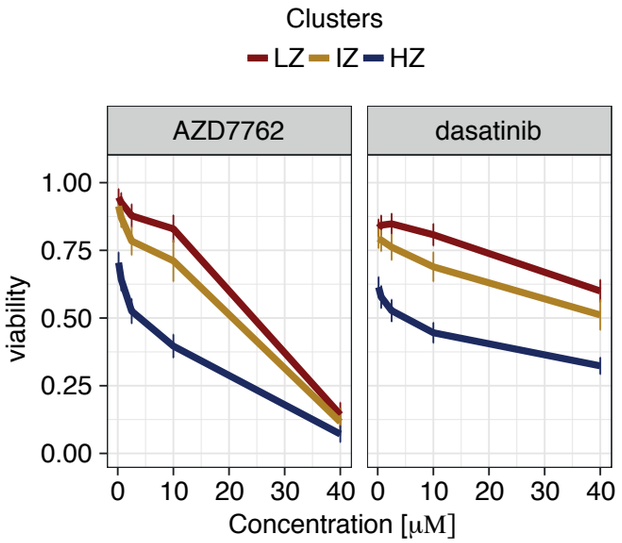
Further characterisation of Factor 1



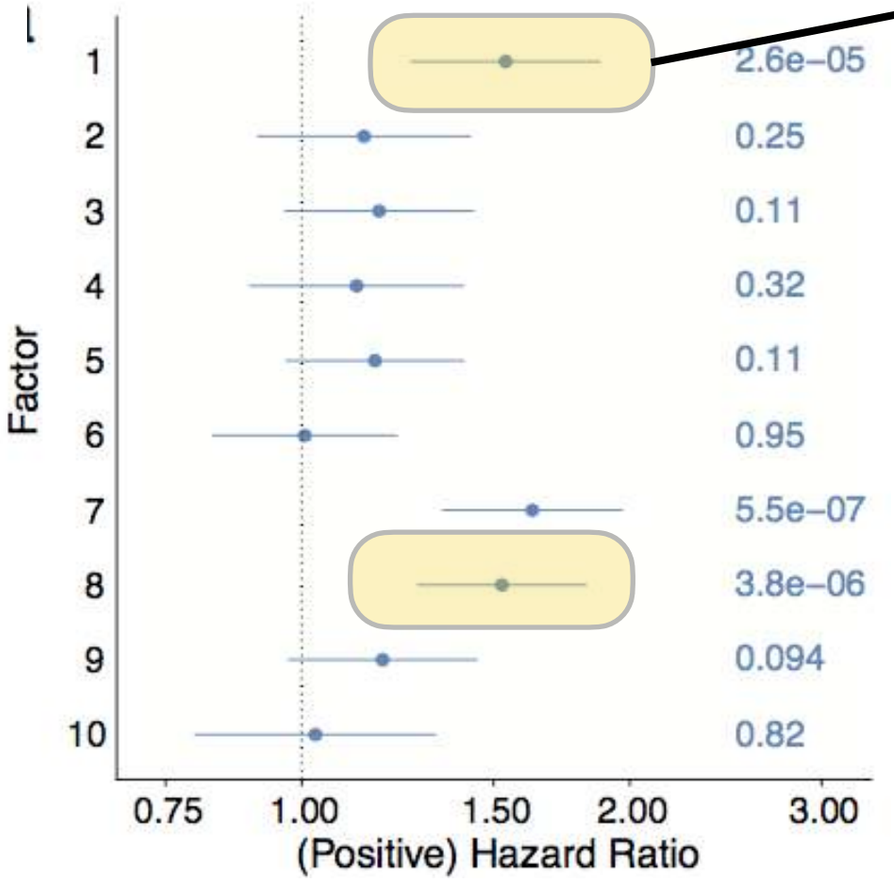
Drug response

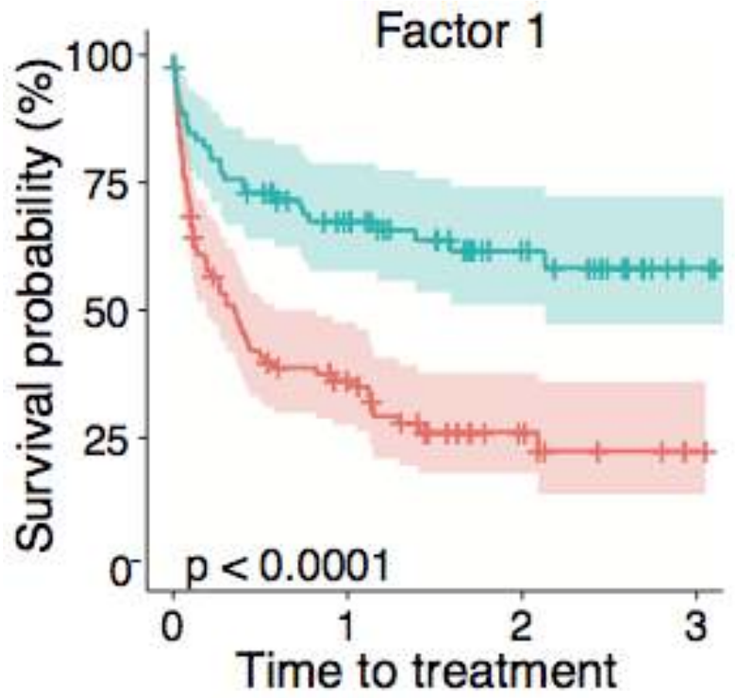






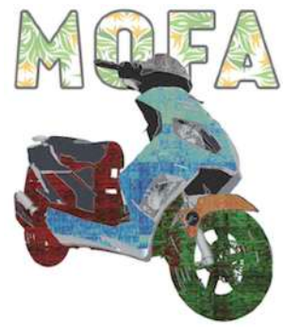
Factors are associated with clinical response





Summary of MOFA

- Unsupervised factor analysis model to disentangle the sources of variation in a multi-view data set
- Requires multi-omic measurements from the same sample
- No tunning of parameters
- Can cope with different data modalities: continuous, binary and counts
- Deals with missing values
- Fast
- Sparse
- Well-established workflow to characterise drivers of variation



Acknowledgements

UNIVERSITY OF CAMBRIDGE

EMBL Heidelberg

Britta Velten

Wolfgang Huber

EMBL-EBI

Oliver Stegle

John Marioni

Damien Arnol

Florian Buettner

EMBL-EBI