Multi-Omics Factor Analysis

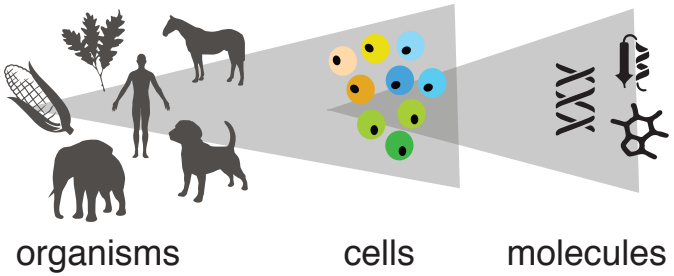
A probabilistic framework for scalable integration of multi-modal data

e-Rum 2020
Britta Velten, Postdoctoral Researcher
DKFZ - Computational Genomics and System Genetics

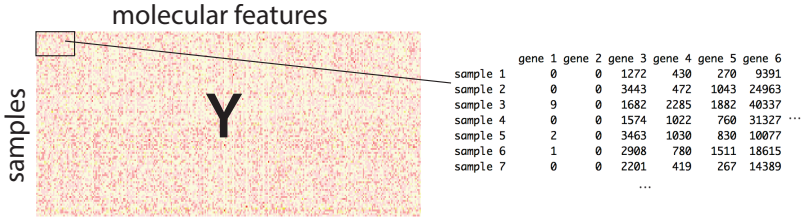
@BrittaVelten
bv2

Omics data to study the molecular underpinnings of life

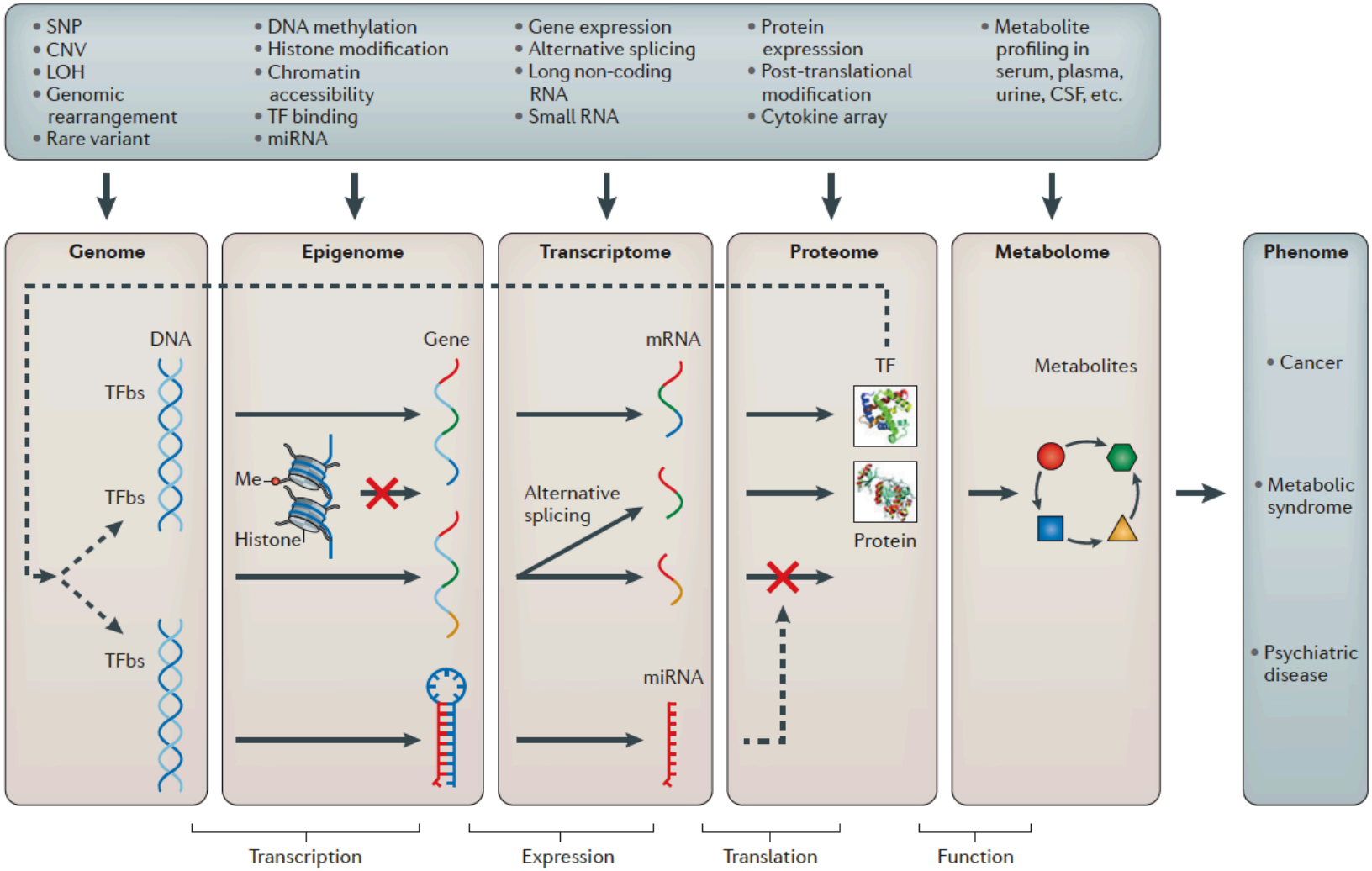
We aim to understand the molecular mechanisms underlying the functioning of an organism.



The term *omics* describes a comprehensive quantitative characterisation of a class of molecules in a given sample



Multi-omics assays study multiple molecular layers simultaneously

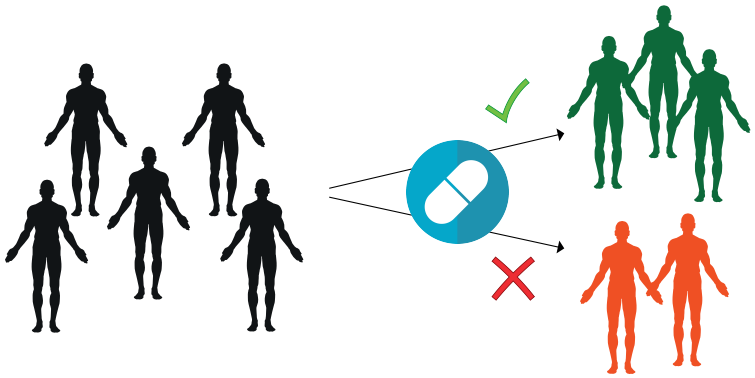


Ritchie et al, Nat Rev Gen 2015

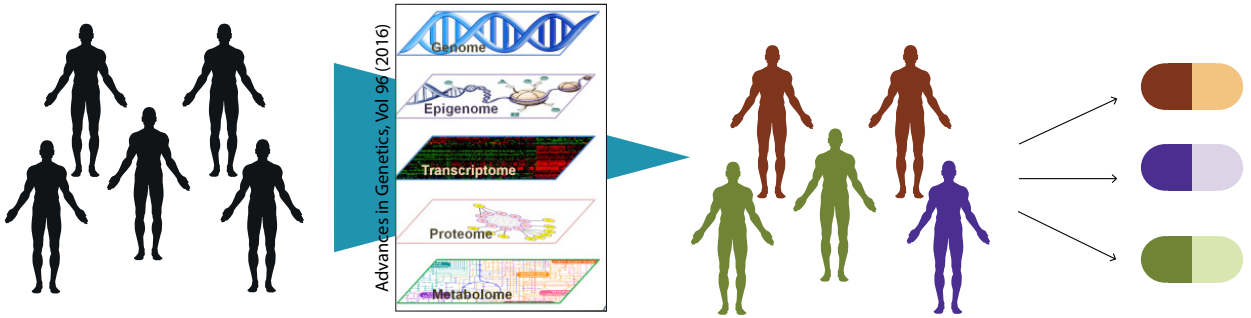
dkfz.

Motivation: Multi-omics for precision medicine

Heterogeneity in disease onset, progression and treatment outcome across patients makes it difficult to decide on the optimal treatment for a patient.

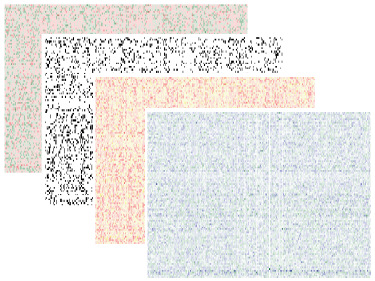


Aim: Gain better understanding of heterogeneity and eventually personalized treatment decisions on a molecular basis.

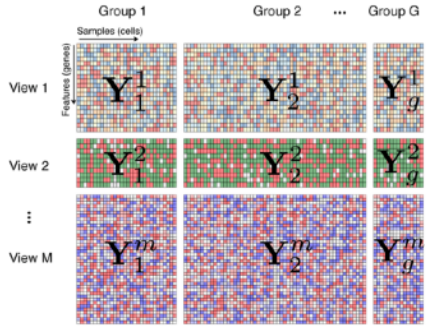


Challenges in the integration of multi-omic data

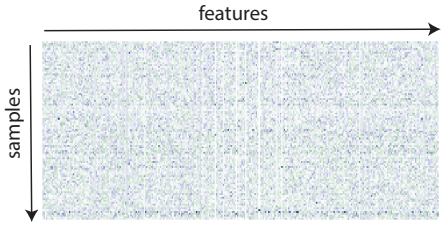
- **Heterogeneous data** from different techniques come with distinct statistical properties and inherent structure
- complex **correlation structures** and hidden confounders
- appropriate **regularization** strategies
- algorithms need to be **scalable** to large data sets
- large amounts (and different patterns) of **missing values**
- **interpretable** approaches for an unsupervised exploration



heterogeneous



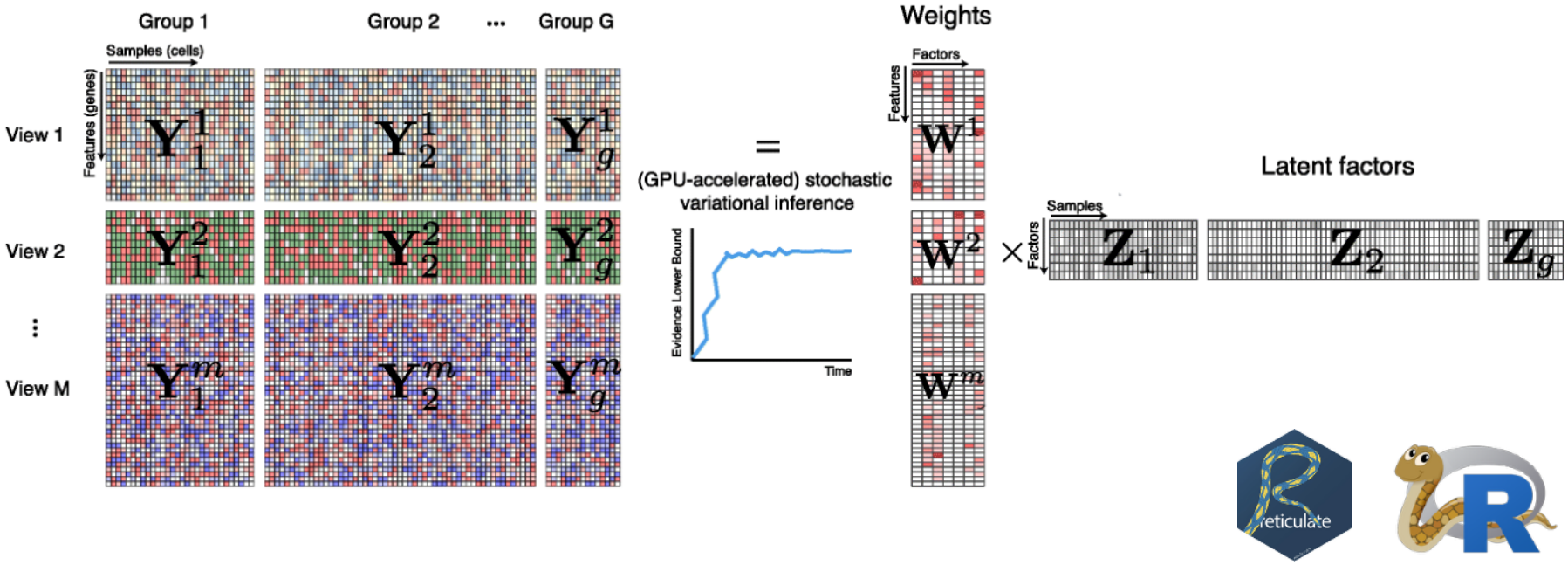
structured



high-dimensional

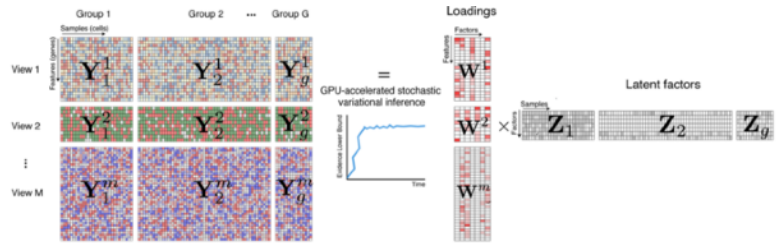
MOFA: A Bayesian model for unsupervised integration of multi-modal data

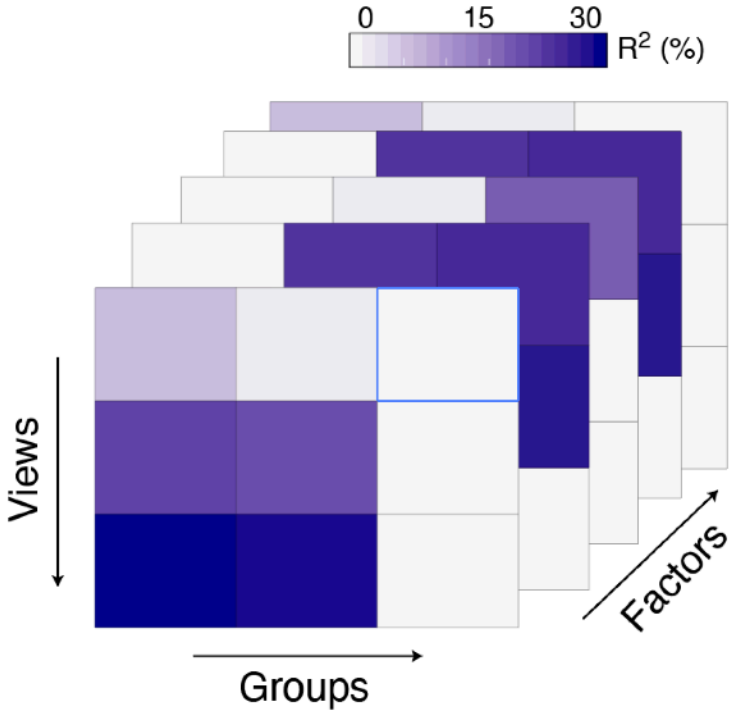
- MOFA performs *structured* matrix factorisation to infer a joint low-dimensional representation of multi-modal data
- different noise models an be used for each data modality
- sparsity priors enable automatic relevance determination of factors and feature weights
- Inference is performed using (stochastic) variational Bayes
- interfaces with Bioconductor classes such as *MultiAssayExperiment* or *Seurat*



Downstream analysis: Variance decomposition

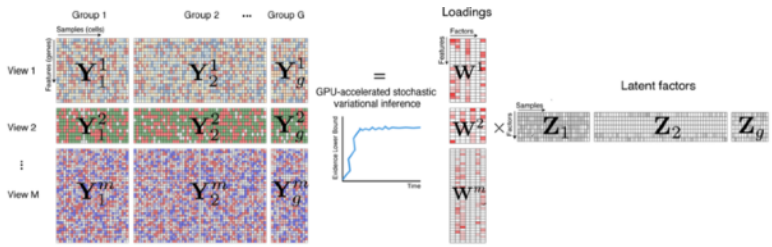
MOFA quantifies how much variance each factor explains in each group and/or view.

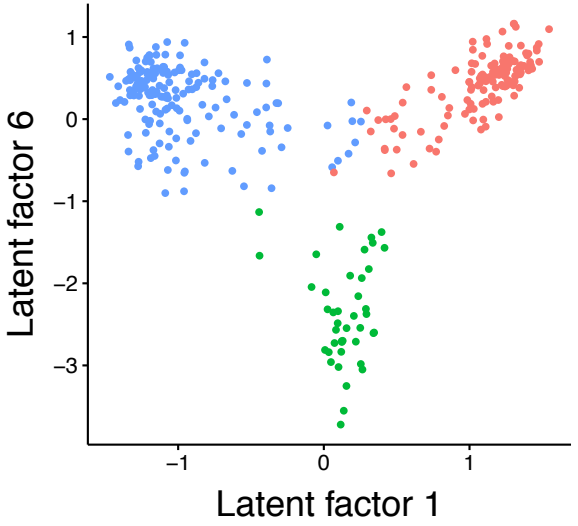




Downstream analysis: Visualisation of samples in factor space

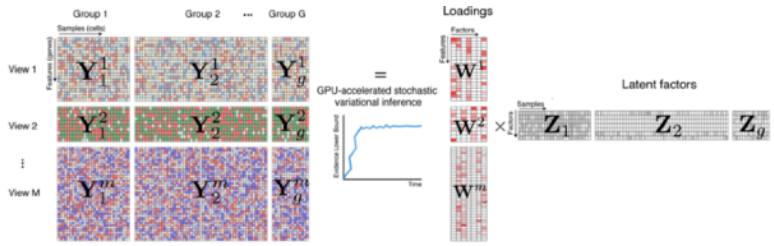
The factor space can be used to visualise or cluster samples or used as input for predictive models.

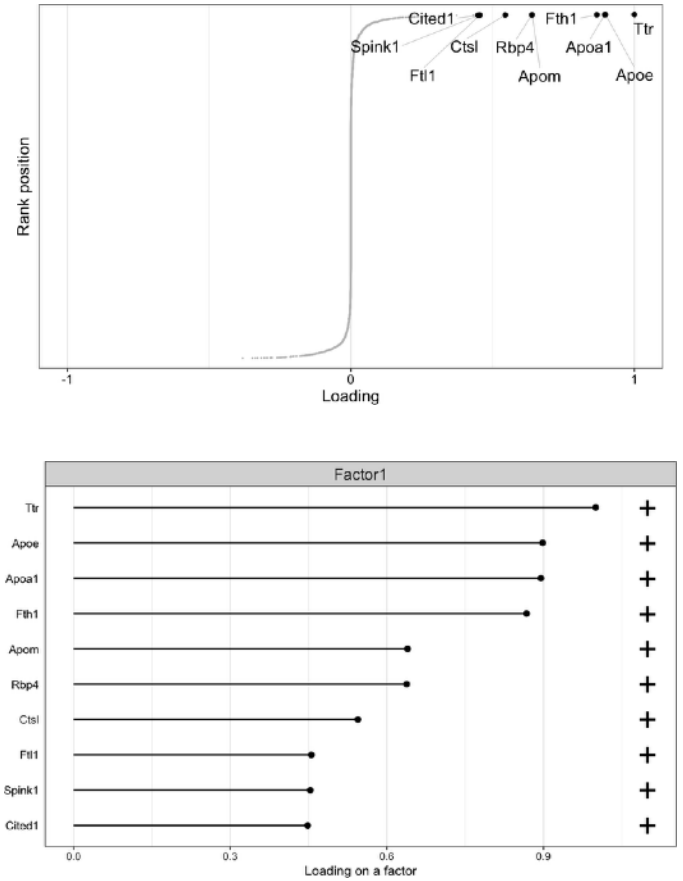




Downstream analysis: Inspection of weights

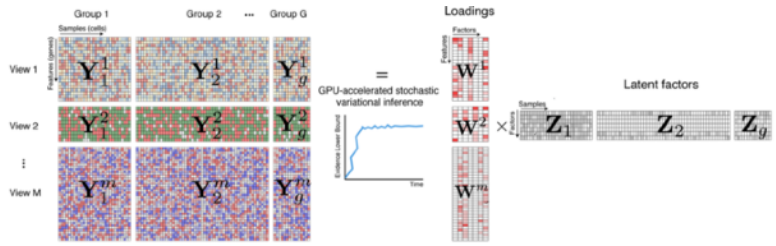
Weights of a factor in each view can give insight into its molecular signature.

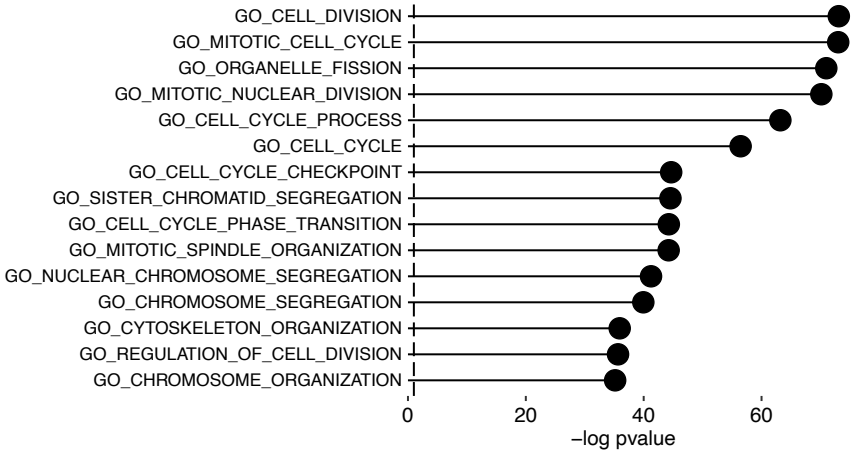




Downstream analysis: Gene set enrichment analysis

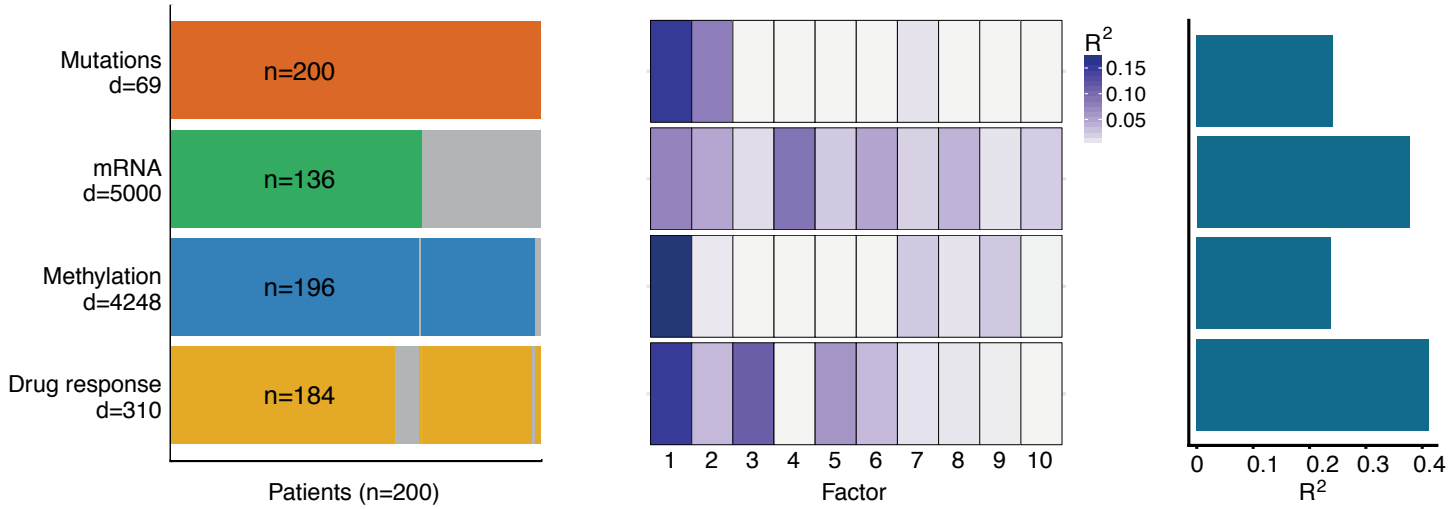
Enrichment analysis of the weights can be used to test for feature sets, e.g. gene sets, linked to a factor.





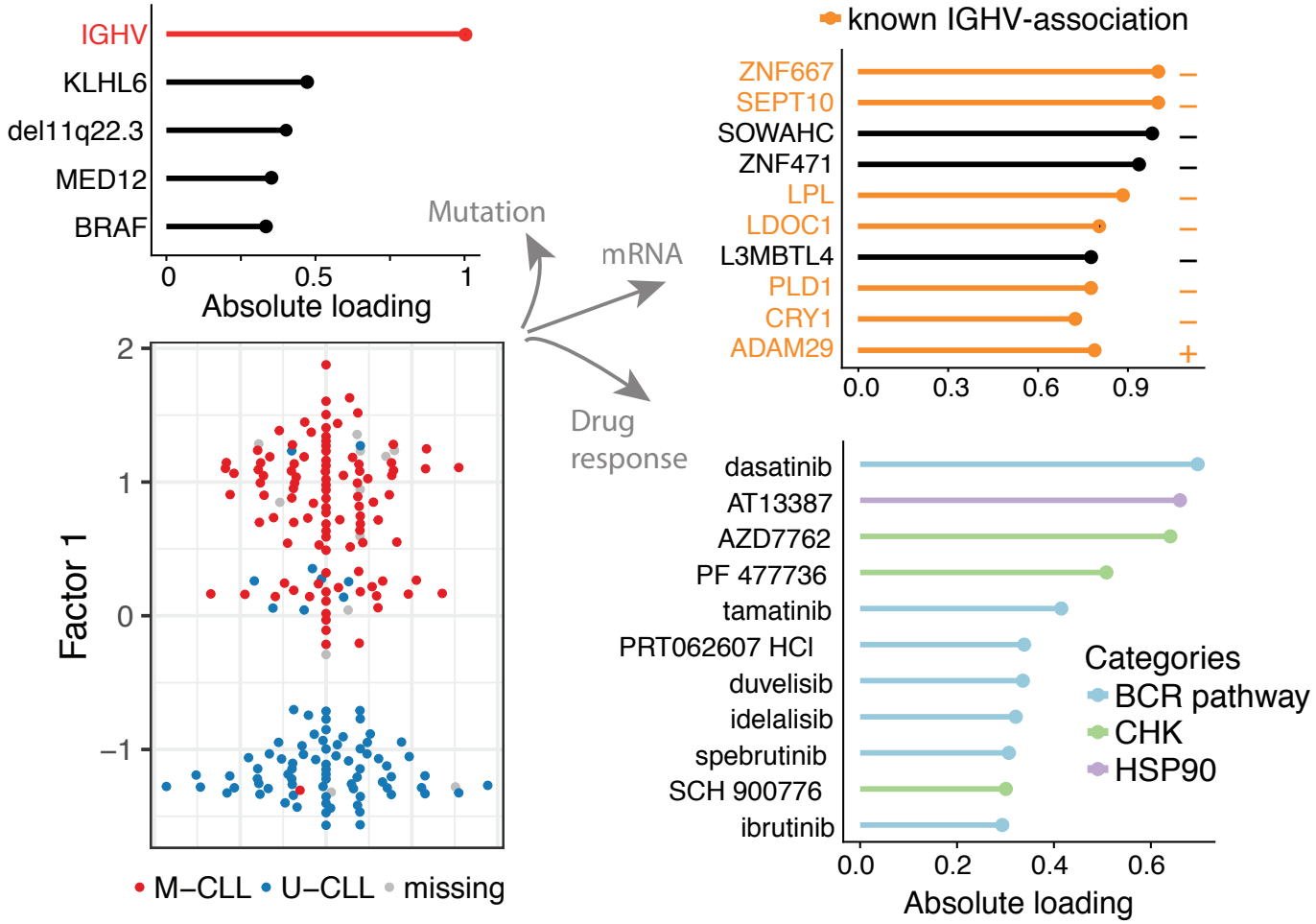
Application 1: Finding sources of heterogeneity in blood cancer

200 leukaemia samples (incompletely) characterized by genomic sequencing, RNAseq, methylation arrays and ex-vivo drug response assays



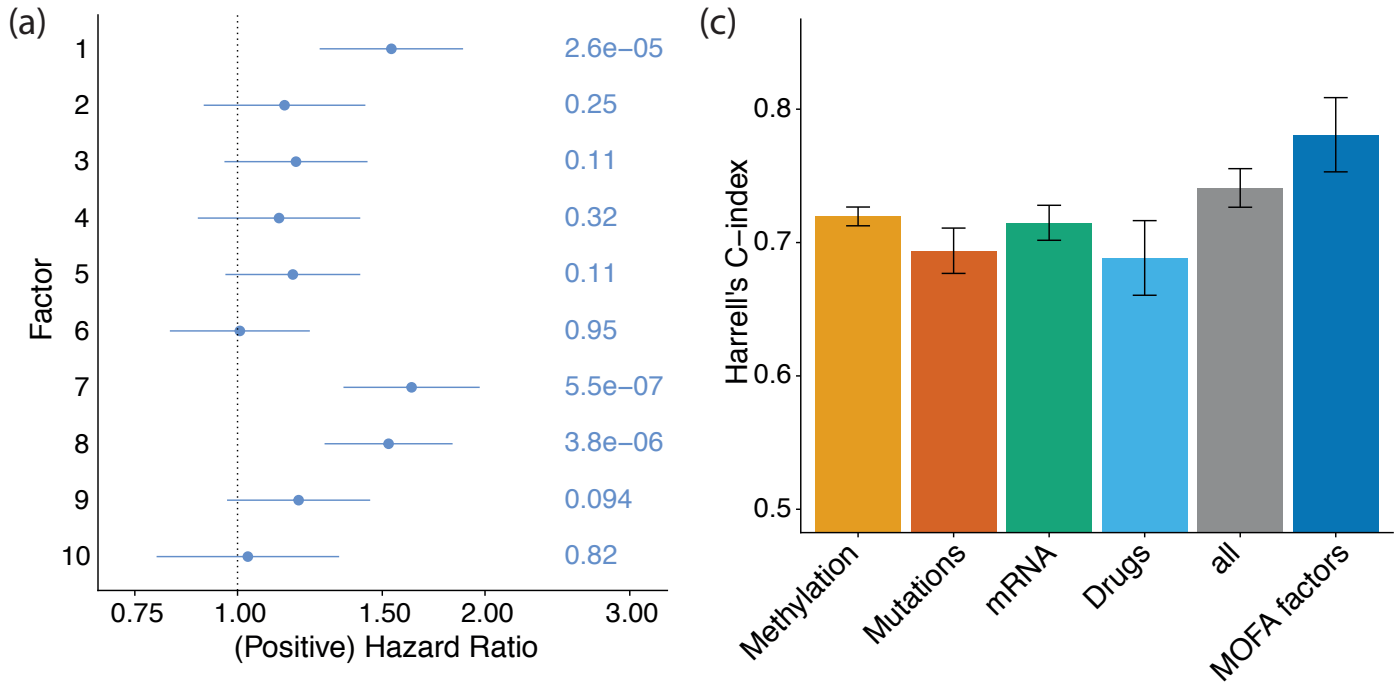
Factor 1 recovers and refines an important clinical marker

Weights link the factor to features from all molecular layers.



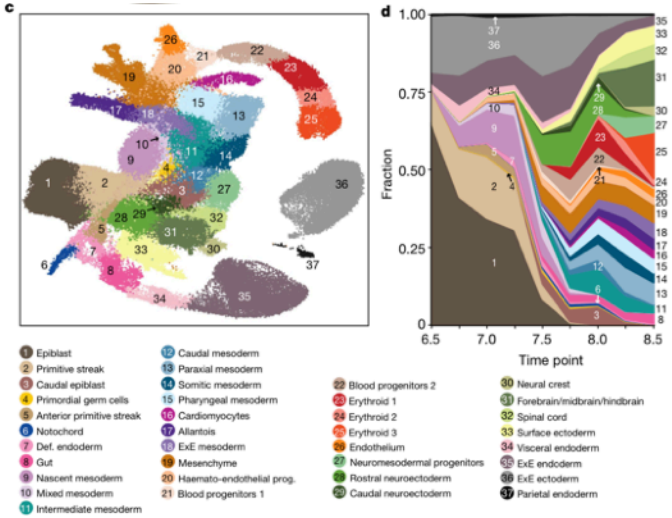
MOFA factors are predictive of clinical outcomes for patients

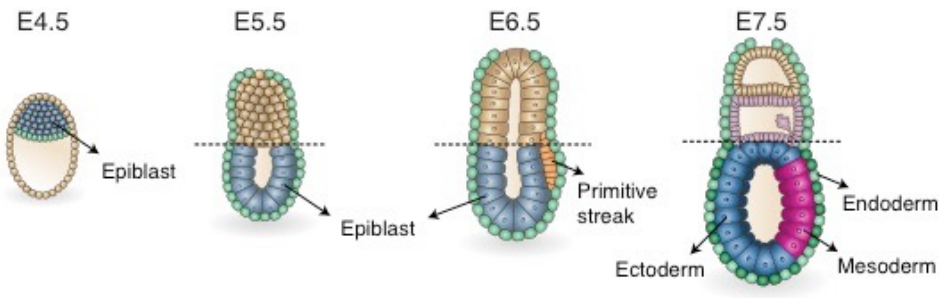
MOFA factors are associated with time to treatment and provide improved prediction compared to models relying on a single omic or concatenated data.



Application 2: Capturing lineage formation from time-course single cell RNA-seq

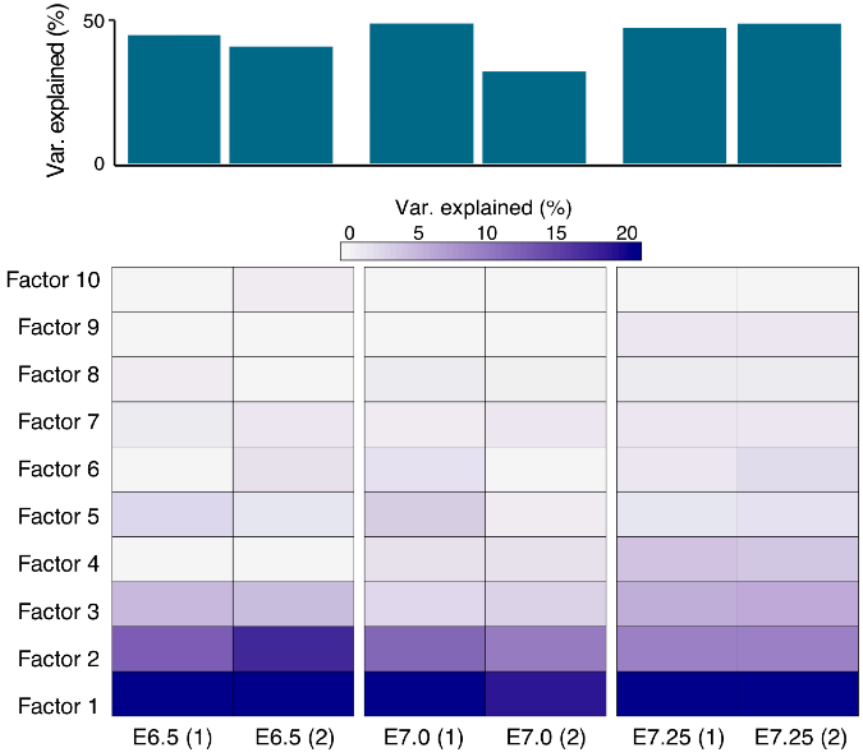
16,152 single cells from mouse embryos at three different developmental stages (E6.5, E7.0, and E7.25)

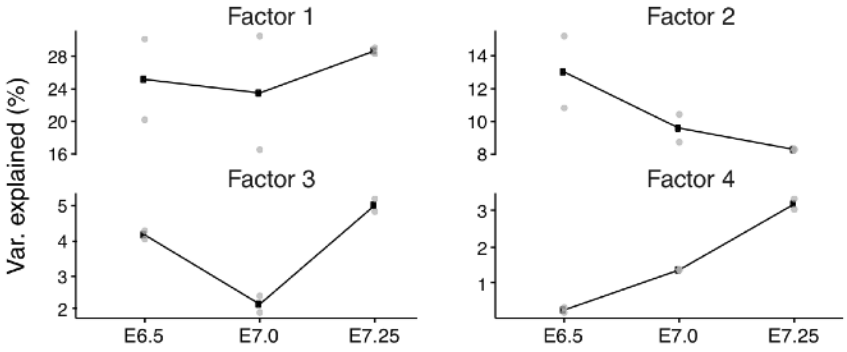




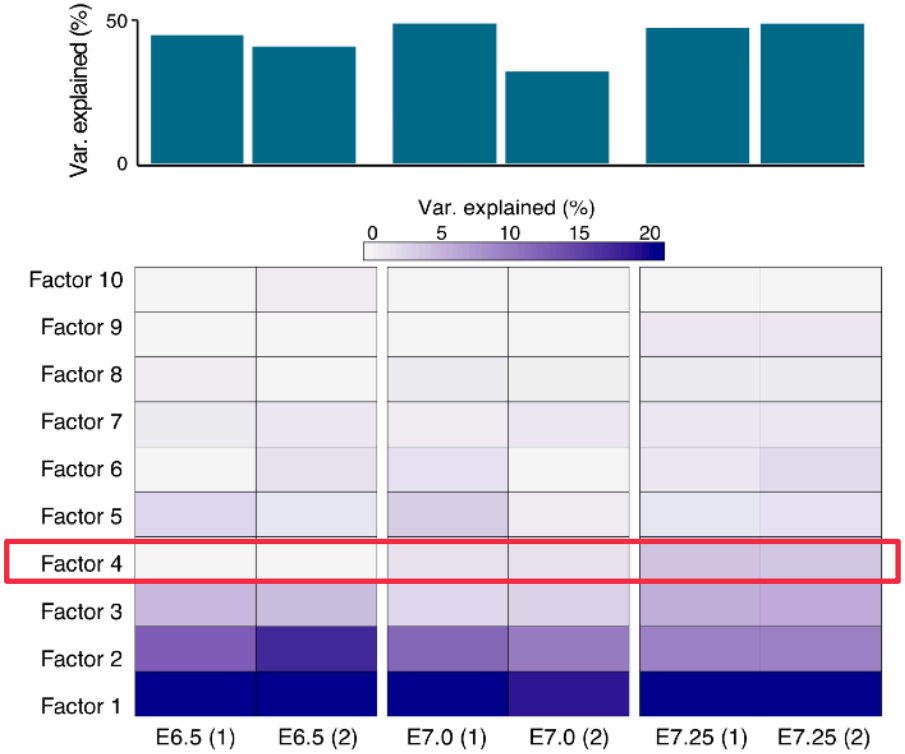
Which molecular processes underly the developmental decisions of a cell?

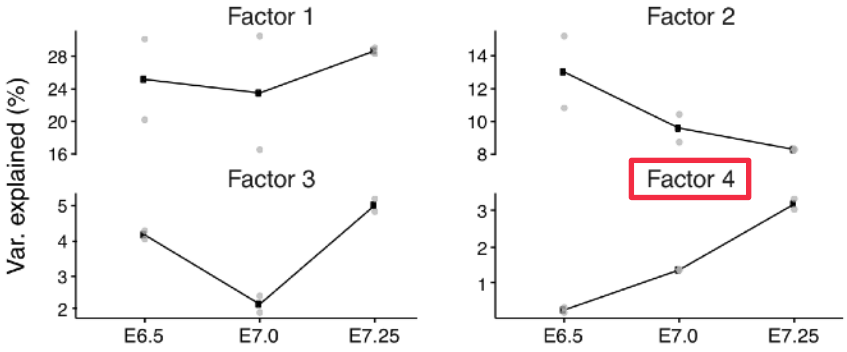
MOFA+ recovers latent factors with differential activity across developmental time



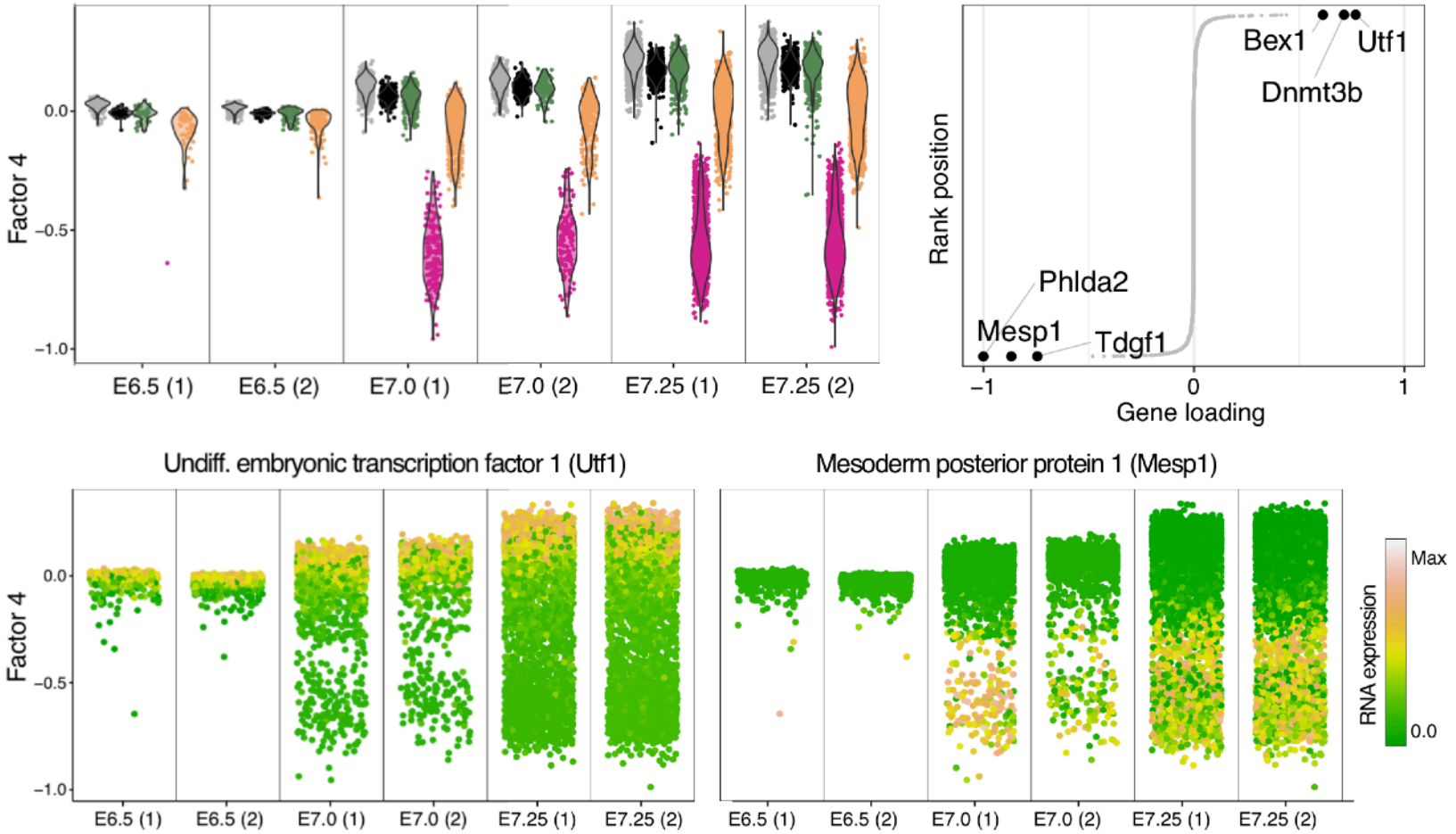


MOFA+ recovers latent factors with differential activity across developmental time



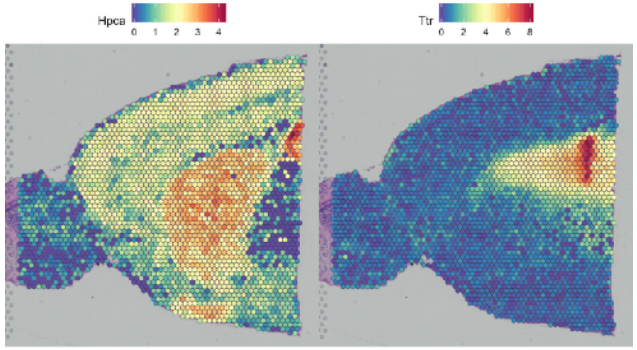


Factor 4 captures the emergence of the mesoderm lineage at E7.0



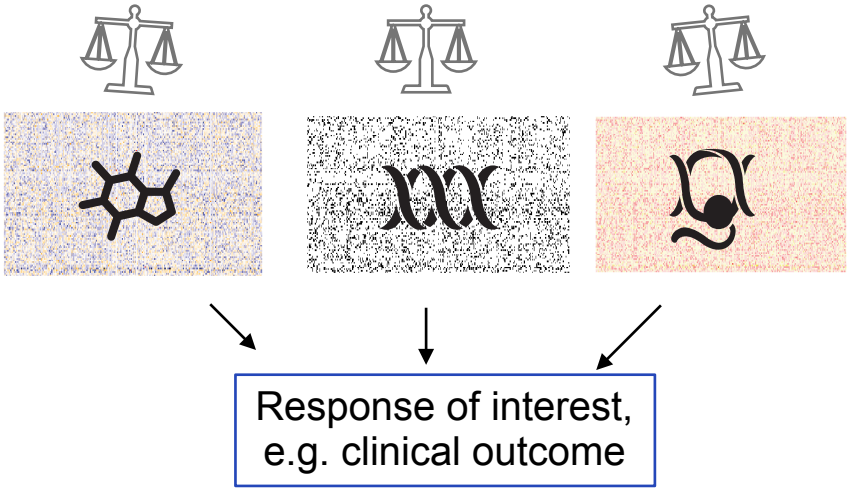
Related work and ongoing research

Encode other data structures
- temporal or spatial data
- networks



Non-linear extensions of MOFA

Supervised integration of multi-modal data: Bioconductor package **graper**



Adaptive penalization in high-dimensional regression and classification with external covariates using variational Bayes

BRITTA VELTEN*, WOLFGANG HUBER

Summary

- MOFA is a Bayesian factor analysis model to disentangle the sources of variation in multi-view and/or multi-group data
- MOFA copes with missing values, is scalable to 100,000's of samples and yields interpretable results by use of sparsity priors
- MOFA interfaces with R/Bioconductor classes, e.g. *MultiAssayExperiment* or *Seurat*
- For model training MOFA uses *reticulate* to interface with python
- Various functions for downstream analysis and a *Shiny App* to explore trained models in an interactive manner are provided

Software

- MOFA is available from Bioconductor
- MOFA2 is available from github.com/bioFAM/MOFA2
- Shiny App: http://www.ebi.ac.uk/shiny/mofa/

Acknowledgements

German Cancer Research Centre (DKFZ)
Oliver Stegle

National Center for Tumor Diseases and Heidelberg University Hospital
Thorsten Zenz
Sascha Dietrich

EMBL
Wolfgang Huber
Danila Bredikhin

EMBL-EBI
Ricard Argelaguet
Damien Arnol
Florian Buettner
John Marioni
Yonatan Deloro

Bundesministerium für Bildung und Forschung

dkfz. GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION

EMBL