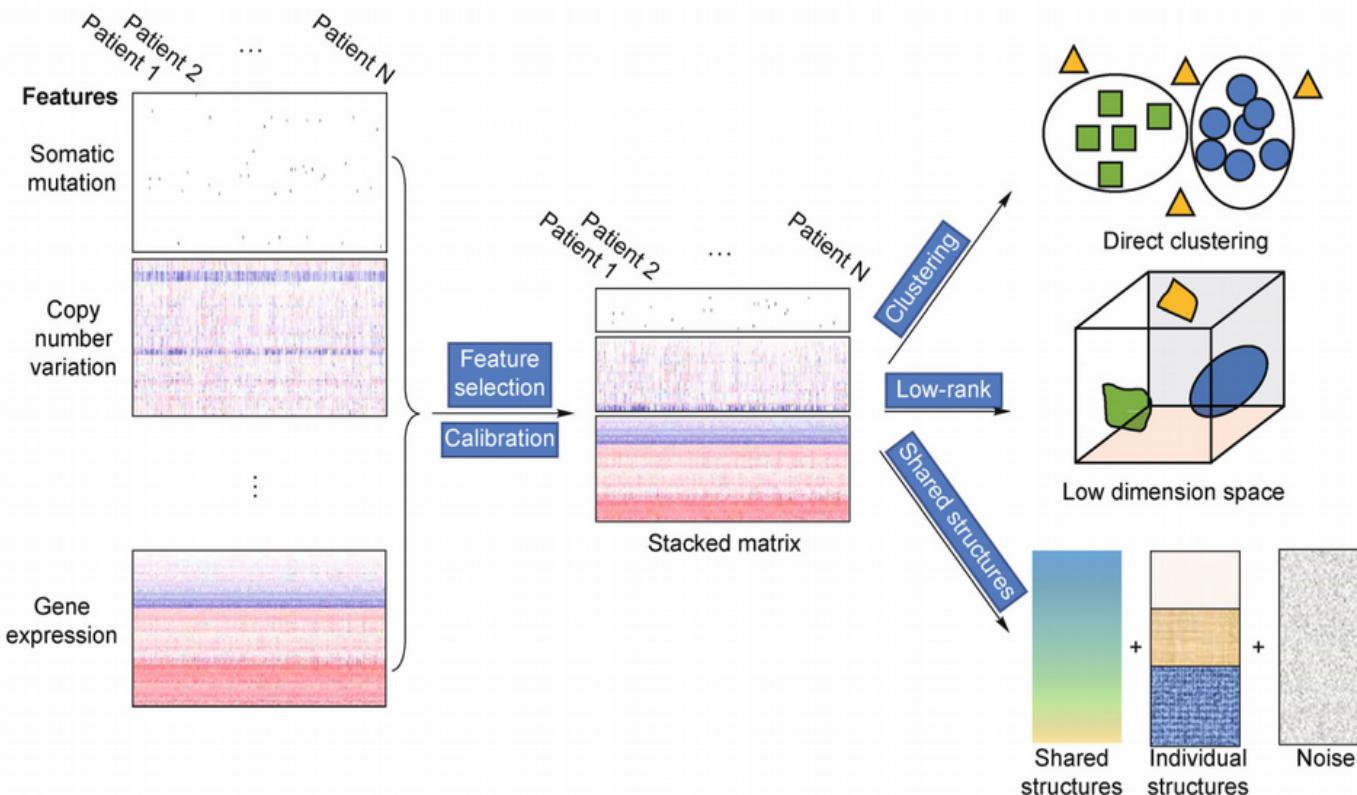
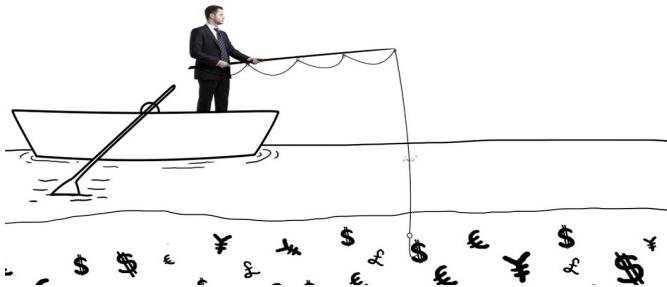


# Unsupervised Omics Integration

Omics Integration and Systems Biology NBIS SciLifeLab course  
Nikolay Oskolkov, Lund University, NBIS SciLifeLab, Sweden



## Fishing expedition



**BMC** Part of Springer Nature

**Genome Biology**

Home About Articles Submission Guidelines

Editorial | Open Access | Published: 03 September 2020

**A hypothesis is a liability**

Itai Yanai & Martin Lercher

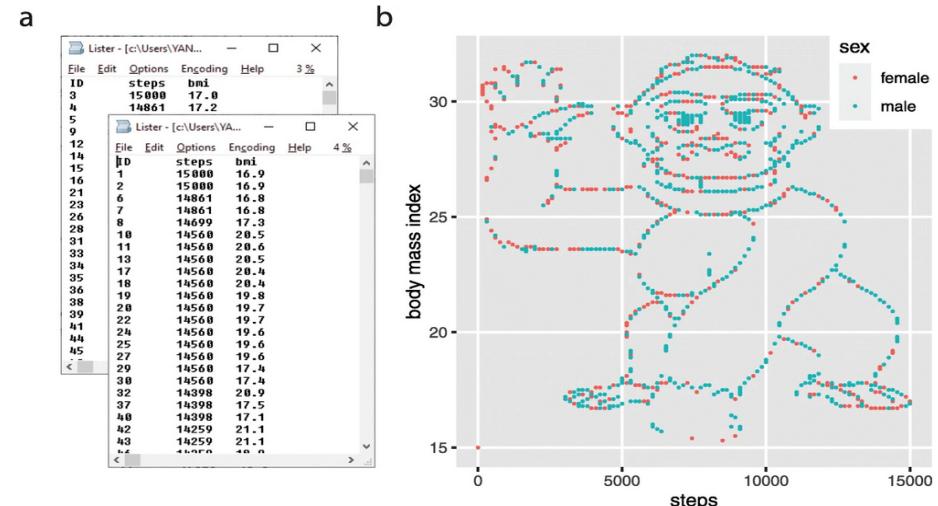
*Genome Biology* 21, Article number: 231 (2020) | Cite this article

12k Accesses | 619 Altmetric | Metrics

“‘When someone seeks,’ said Siddhartha, ‘then it easily happens that his eyes see only the thing that he seeks, and he is able to find nothing, to take in nothing. [...] Seeking means: having a goal. But finding means: being free, being open, having no goal.’” Hermann Hesse

There is a hidden cost to having a hypothesis. It arises from the relationship between night science and day science, the two very distinct modes of activity in which scientific ideas are generated and tested, respectively [1, 2]. With a hypothesis in hand, the impressive strengths of day science are unleashed, guiding us in designing tests, estimating parameters, and throwing out the hypothesis if it fails the tests. But when we analyze the results of an experiment, our mental focus on a specific hypothesis can prevent us from exploring other aspects of the data, effectively blinding us to new ideas. A hypothesis then becomes a liability for any night science explorations. The corresponding limitations on our creativity, self-imposed in hypothesis-driven research, are of particular concern in the context of modern biological datasets, which are often vast and likely to contain hints at multiple distinct and potentially exciting discoveries. Night science has its own liability though, generating many spurious relationships and false hypotheses. Fortunately, these are exposed by the light of day science, emphasizing the complementarity of the two modes, where each overcomes the

- I do not understand your biological hypothesis
- I do not have any



a An artificial dataset given to students with and without explicit hypotheses on the relationship between BMI and the steps taken on a particular day, for men and women. b A plot of the dataset. c The contingency table for students in the two groups (“hypothesis-focused,” “hypothesis-free”) that discovered the gorilla or not [6]

## Method

molecular  
systems' biology

# Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets

Ricard Argelaguet<sup>1,†</sup> , Britta Velten<sup>2,†</sup> , Damien Arnol<sup>1</sup> , Sascha Dietrich<sup>3</sup> , Thorsten Zenz<sup>3,4,5</sup> , John C Marioni<sup>1,6,7</sup> , Florian Buettner<sup>1,8,\*</sup> , Wolfgang Huber<sup>2,\*\*</sup> & Oliver Stegle<sup>1,2,\*\*\*</sup>

## Abstract

Multi-omics studies promise the improved characterization of biological processes across molecular layers. However, methods for the unsupervised integration of the resulting heterogeneous data sets are lacking. We present Multi-Omics Factor Analysis (MOFA), a computational method for discovering the principal sources of variation in multi-omics data sets. MOFA infers a set of (hidden) factors that capture biological and technical sources of variability. It disentangles axes of heterogeneity that are shared across multiple modalities and those specific to individual data modalities. The learnt factors enable a variety of downstream analyses, including identification of sample subgroups, data imputation and the detection of outlier samples. We applied MOFA to a cohort of 200 patient samples of chronic lymphocytic leukaemia, profiled for somatic mutations, RNA expression, DNA methylation and *ex vivo* drug responses. MOFA identified major dimensions of disease heterogeneity, including immunoglobulin heavy-chain variable region status, trisomy of chromosome 12 and previously underappreciated drivers, such as response to oxidative stress. In a second application, we used MOFA to analyse single-cell multi-omics data, identifying coordinated transcriptional and epigenetic changes along cell differentiation.

**Keywords** data integration; dimensionality reduction; multi-omics; personalized medicine; single-cell omics

**Subject Categories** Computational Biology; Genome-Scale & Integrative Biology; Methods & Resources

DOI 10.1525/nbs.20178124 | Received 27 November 2017 | Revised 28 May 2018 | Accepted 29 May 2018

Mol Syst Biol. (2018) 14: e8124

## Introduction

Technological advances increasingly enable multiple biological layers to be probed in parallel, ranging from genome, epigenome, transcriptome, proteome and metabolome to phenotype profiling (Hasin *et al.*, 2017). Integrative analyses that use information across these data modalities promise to deliver more comprehensive insights into the biological systems under study. Motivated by this, multi-omics profiling is increasingly applied across biological domains, including cancer biology (Gerstung *et al.*, 2015; Iorio *et al.*, 2016; Mertins *et al.*, 2016; Cancer Genome Atlas Research Network, 2017), regulatory genomics (Chen *et al.*, 2016), microbiology (Kim *et al.*, 2016) or host-pathogen biology (Soderholm *et al.*, 2016). Most recent technological advances have also enabled performing multi-omics analyses at the single-cell level (Macaulay *et al.*, 2015; Angermueller *et al.*, 2016; Guo *et al.*, 2017; Clark *et al.*, 2018; Colomé-Tatché & Theis, 2018). A common aim of such applications is to characterize heterogeneity between samples, as manifested in one or several of the data modalities (Ritchie *et al.*, 2015). Multi-omics profiling is particularly appealing if the relevant axes of variation are not known *a priori*, and hence may be missed by studies that consider a single data modality or targeted approaches.

A basic strategy for the integration of omics data is testing for marginal associations between different data modalities. A prominent example is molecular quantitative trait locus mapping, where large numbers of association tests are performed between individual genetic variants and gene expression levels (GTEx Consortium, 2015) or epigenetic marks (Chen *et al.*, 2016). While evidently useful for variant annotation, such association studies are inherently *local* and do not provide a coherent global map of the molecular differences between samples. A second strategy is the use of kernel- or graph-based methods to combine different

Method | Open Access | Published: 11 May 2020

## MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data

Ricard Argelaguet , Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C. Marioni & Oliver Stegle

*Genome Biology* 21, Article number: 111 (2020) | [Cite this article](#)

30k Accesses | 127 Citations | 119 Altmetric | [Metrics](#)

## Abstract

Technological advances have enabled the profiling of multiple molecular layers at single-cell resolution, assaying cells from multiple samples or conditions. Consequently, there is a growing need for computational strategies to analyze data from complex experimental designs that include multiple data modalities and multiple groups of samples. We present Multi-Omics Factor Analysis v2 (MOFA+), a statistical framework for the comprehensive and scalable integration of single-cell multi-modal data. MOFA+ reconstructs a low-dimensional representation of the data using computationally efficient variational inference and supports flexible sparsity constraints, allowing to jointly model variation across multiple sample groups and data modalities.

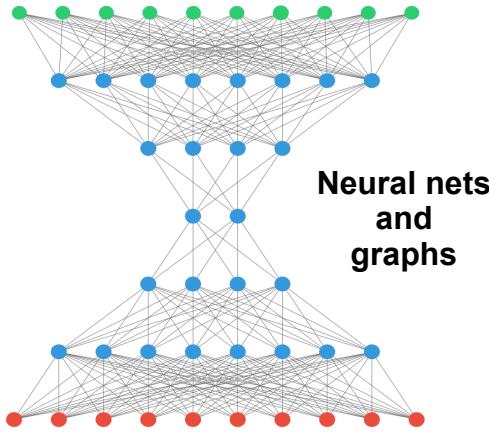
## Background

Single-cell methods have provided unprecedented opportunities to assay cellular heterogeneity. This is particularly important for studying complex biological processes, including the immune system, embryonic development, and cancer [1,2,3,4].

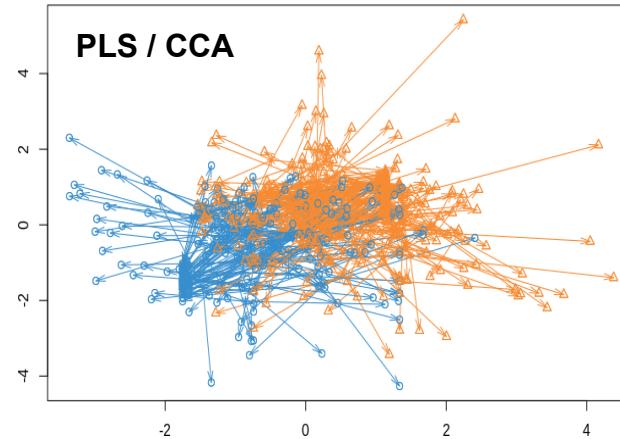
[Download PDF](#)

- [Sections](#)
- [Figures](#)
- [References](#)
- [Abstract](#)
- [Background](#)
- [Results](#)
- [Discussion](#)
- [Conclusions](#)
- [Methods](#)
- [Availability of data and materials](#)
- [References](#)
- [Acknowledgements](#)
- [Funding](#)
- [Author information](#)
- [Ethics declarations](#)
- [Additional information](#)
- [Supplementary information](#)
- [Rights and permissions](#)
- [About this article](#)

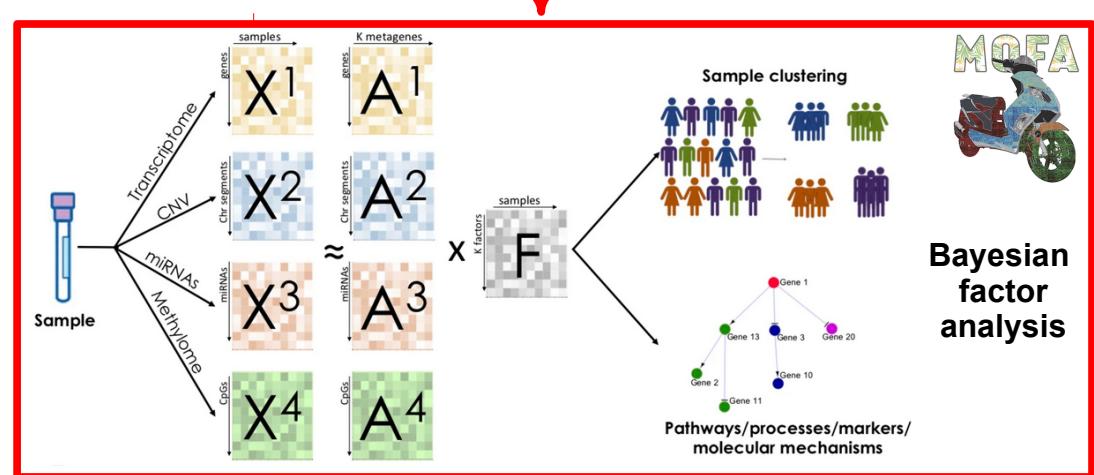
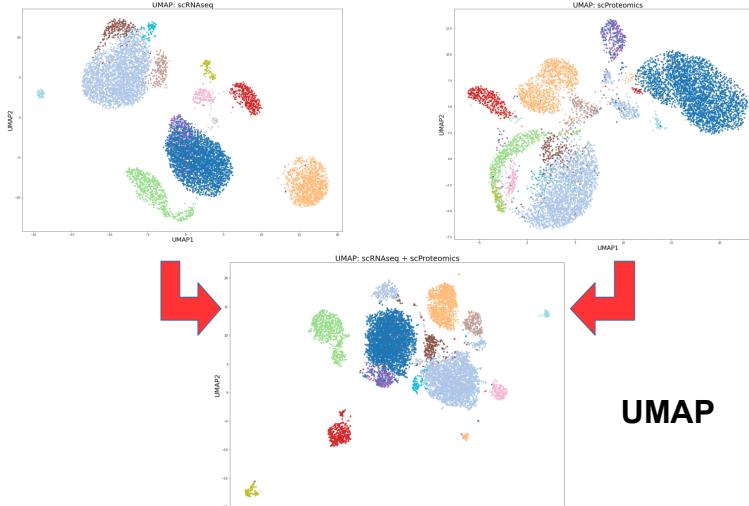
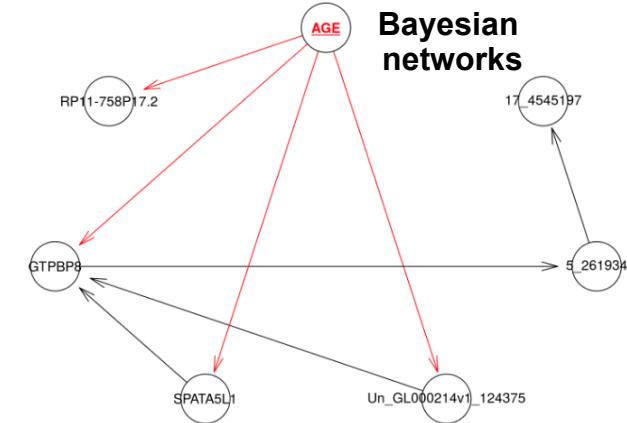
## Convert to common space

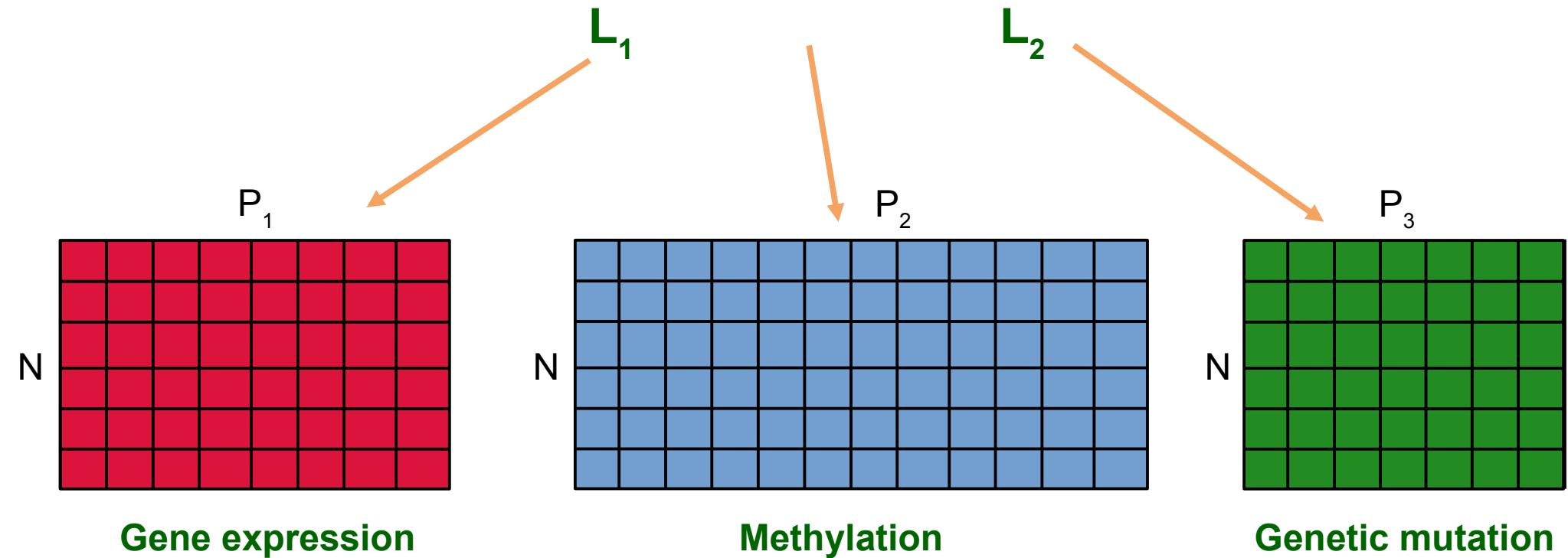
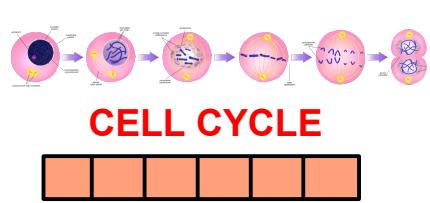
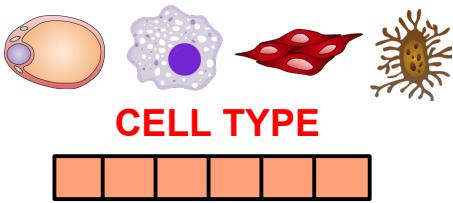


## Extract common variation

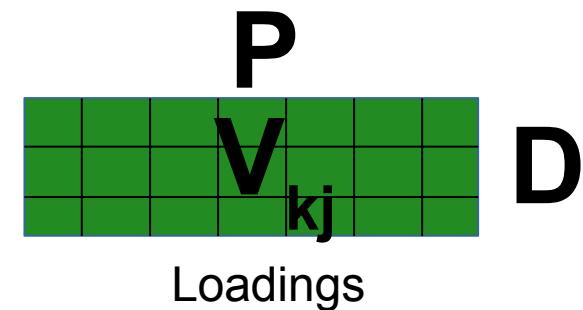
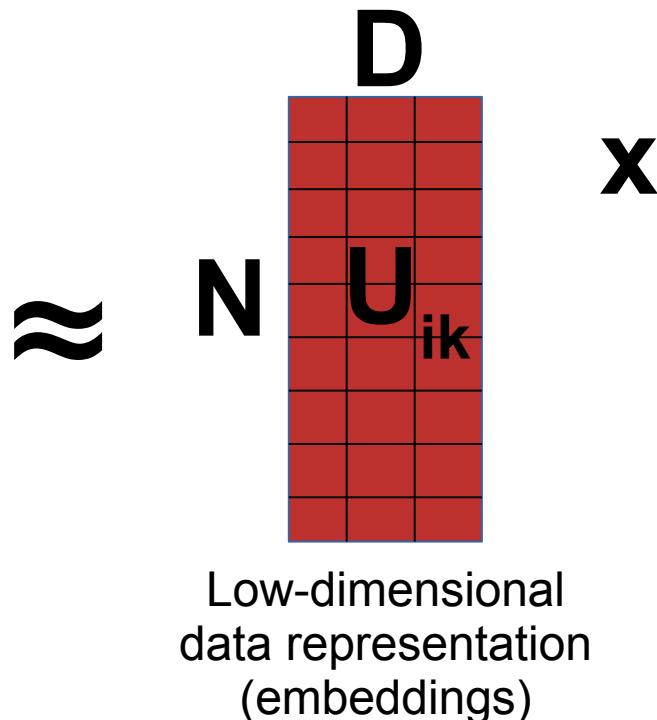
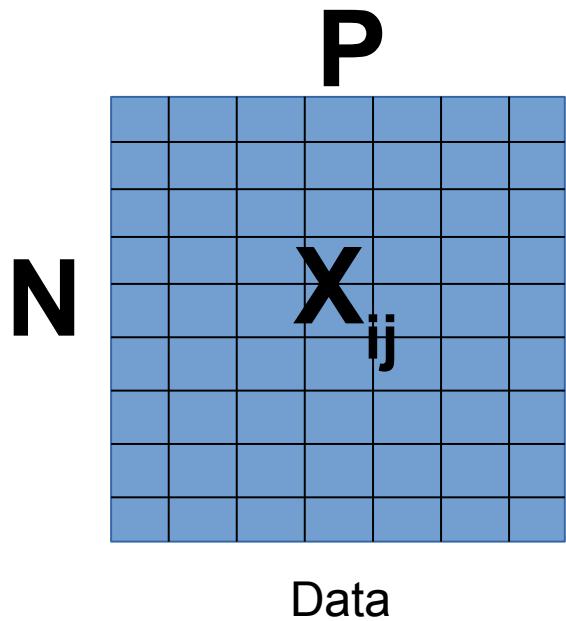


## Combine via Bayes rule

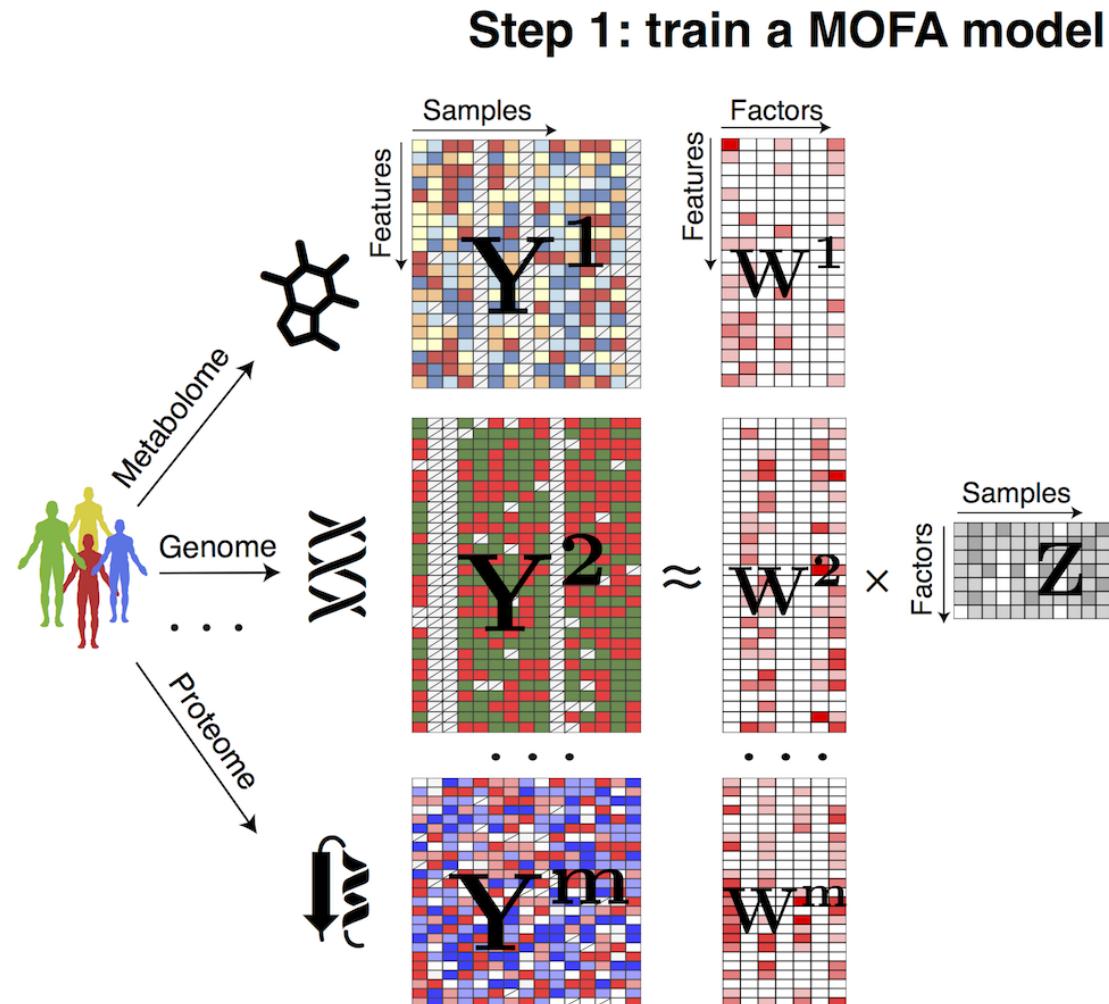
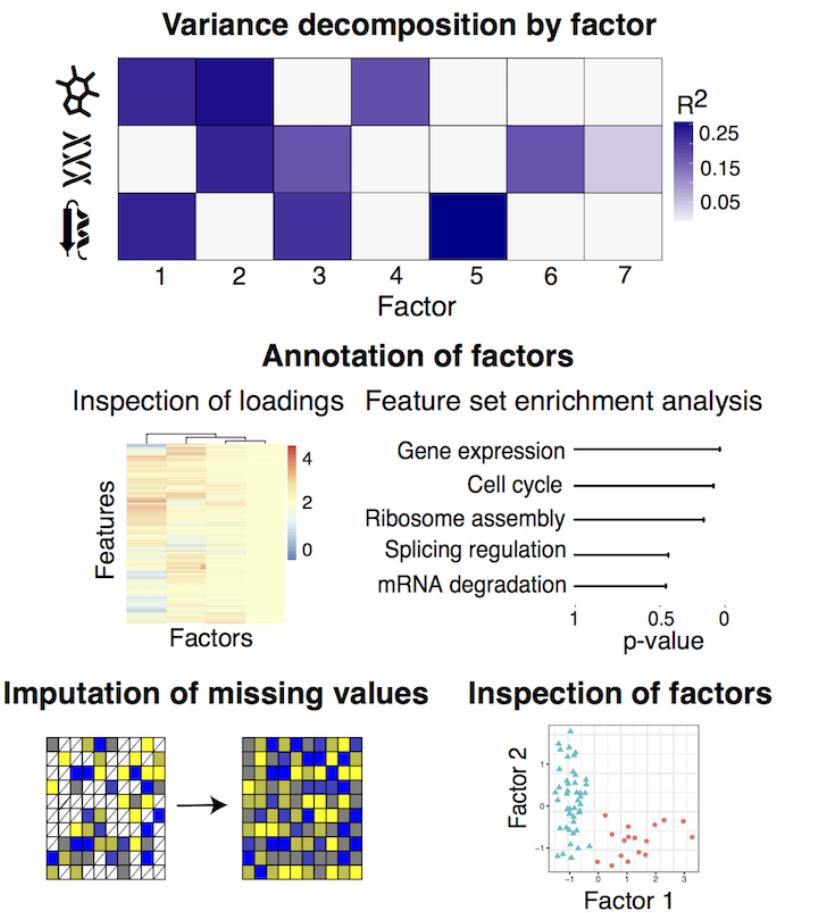


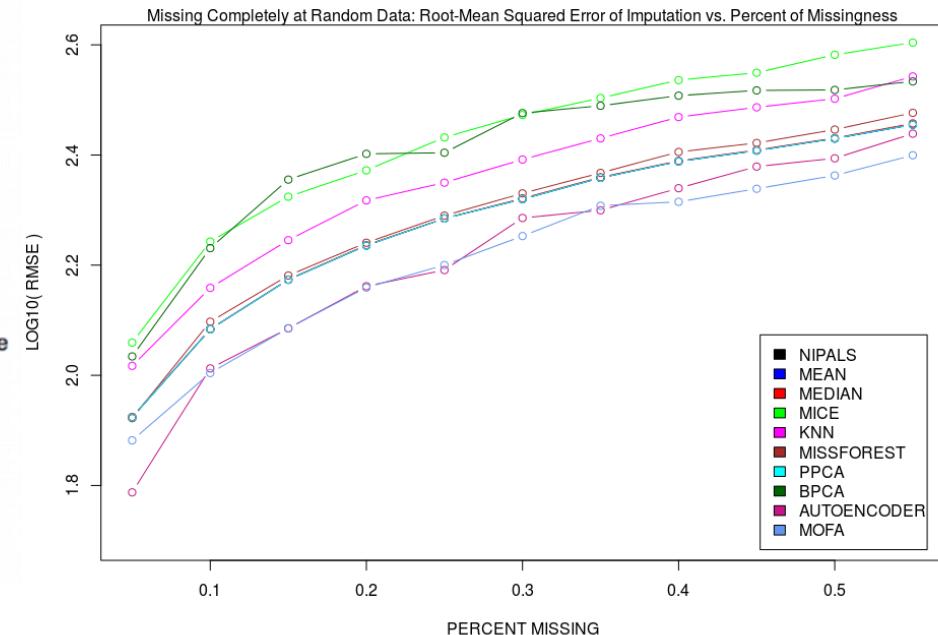
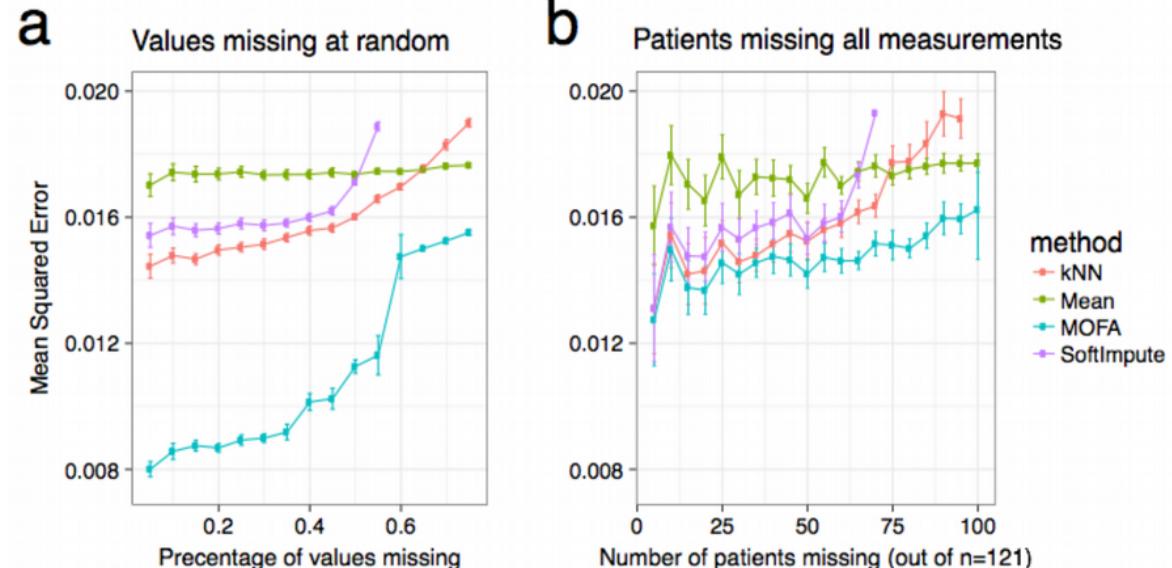


$$\mathbf{X}_{ij} \approx \mathbf{U}_{ik} \mathbf{V}_{kj}$$



$$\text{Loss} = \sum_{i=1}^N \sum_{j=1}^P (\mathbf{X}_{ij} - \mathbf{U}_{ik} \mathbf{V}_{kj})^2$$

**Step 2: downstream analysis**



Bayesian framework is insensitive to missing data, priors compensate for the lack of data

Factor analysis models, also called latent variable models, are a probabilistic modelling approach which aim to reduce the dimensionality of a (big) dataset into a small set of variables which are easier to interpret and visualise. More formally, given a dataset  $\mathbf{Y}$  of  $N$  samples and  $D$  features, latent variable models attempt to explain dependencies between the features by means of a potentially smaller set of  $K$  unobserved (latent) factors. MOFA is a generalisation of traditional Factor Analysis where the input data consists of  $M$  matrices  $\mathbf{Y}^m = [y_{nd}^m] \in \mathbb{R}^{N \times D_m}$  where each matrix  $m$  is called a view. Each view consists of non-overlapping features which usually, but not necessarily, represent different assays. The input data is then factorised as:

$$\mathbf{Y}^m = \mathbf{Z}\mathbf{W}^{mT} + \boldsymbol{\epsilon}^m, \quad (1)$$

where  $\mathbf{Z} = [z_{nk}] \in \mathbb{R}^{N \times K}$  is a single matrix that contains the low-dimensional latent variables,  $\mathbf{W}^m = [w_{dk}^m] \in \mathbb{R}^{D_m \times K}$  are loading matrices that relate the high-dimensional space to the low dimensional representation, and  $\boldsymbol{\epsilon}^m = [\epsilon_d^m] \in \mathbb{R}^{D_m}$  denotes residual noise. We start by assuming Gaussian residuals  $\boldsymbol{\epsilon}^m$ , similar to standard (group) factor analysis models, while allowing for heteroscedasticity across features:

$$p(\epsilon_d^m) = \mathcal{N}(\epsilon_d^m | 0, 1/\tau_d^m). \quad (2)$$

This results in the following normal likelihood (for extensions to non-Gaussian settings see section 4):

$$p(y_{nd}^m) = \mathcal{N}(y_{nd}^m | \mathbf{z}_{n,:}\mathbf{w}_{d,:}^{mT}, 1/\tau_d^m), \quad (3)$$

where  $\mathbf{w}_{d,:}^m$  denotes the  $d$ -th row of the loading matrix  $\mathbf{W}^m$  and  $\mathbf{z}_{n,:}$  the  $n$ -th row of the latent factor matrix  $\mathbf{Z}$ . For a fully probabilistic treatment we place prior distributions on the weights  $\mathbf{W}^m$ , the latent variables  $\mathbf{Z}$  as well as on the precision of the noise  $\tau^m$ . We use a standard Gaussian prior on the latent variables and a conjugate Gamma prior for the precision:

$$p(z_{n,k}) = \mathcal{N}(z_{n,k} | 0, 1), \quad (4)$$

$$p(\tau_d^m) = \mathcal{G}(\tau_d^m | a_0^\tau, b_0^\tau), \quad (5)$$

To ensure scalable inference we use a variational approach with a mean-field approximation [3]. Briefly, in variational inference the true intractable posterior distribution of the unobserved variables  $p(\mathbf{X}|\mathbf{Y})$  is approximated by a simpler distribution of factorized form  $q(\mathbf{X}) = \prod_i q(\mathbf{x}_i)$  that leads to an efficient inference scheme. Here,  $\mathbf{X}$  denotes all the hidden variables (including parameters) and  $\mathbf{Y}$  denotes all the observed variables.

Under this approximation, the true log marginal likelihood  $\log p(\mathbf{Y})$  is lower bounded by:

$$\begin{aligned} \mathcal{L}(\mathbf{X}) &= \int q(\mathbf{X}) \left( \log \frac{p(\mathbf{X}|\mathbf{Y})}{q(\mathbf{X})} + \log p(\mathbf{Y}) \right) d\mathbf{X} \\ &= \log p(\mathbf{Y}) - \text{KL}(q(\mathbf{X})||p(\mathbf{X}|\mathbf{Y})) \\ &\leq \log p(\mathbf{Y}) \end{aligned} \quad (11)$$

$\mathcal{L}(\mathbf{X})$  is called the Evidence Lower Bound (ELBO), which is equal to the sum of the model evidence and the negative KL-divergence between the true posterior and the variational distribution. The key observation here is that increasing the ELBO is equivalent to decreasing the KL-divergence between the two distributions.

Variational learning involves optimising the functional  $\mathcal{L}(\mathbf{X})$  with respect to the distribution  $q(\mathbf{X})$ . If we allow any possible choice of  $q(\mathbf{X})$ , then the maximum of the lower bound  $\mathcal{L}(\mathbf{X})$  will occur when the KL-divergence vanishes, which occurs when  $q(\mathbf{X})$  equals the true posterior distribution  $p(\mathbf{X}|\mathbf{Y})$ . Nevertheless, since the true posterior is intractable, this does not lead to any simplification of the problem. Instead, it is necessary to consider a restricted family of variational distributions that are tractable to compute and then seek the member of this family for which the KL divergence is minimised [2].

### Mean-field approximation

The most common type of variational Bayes, known as mean-field approach, assumes that the variational distribution factorises over  $M$  disjoint groups of variables:

$$q(\mathbf{X}) = \prod_{i=1}^M q(\mathbf{x}_i)$$

Evidently, this family of distributions does not usually contain the true posterior because the unobserved variables have dependencies, but this assumption allows the derivation of an analytical inference scheme



## ARTICLE

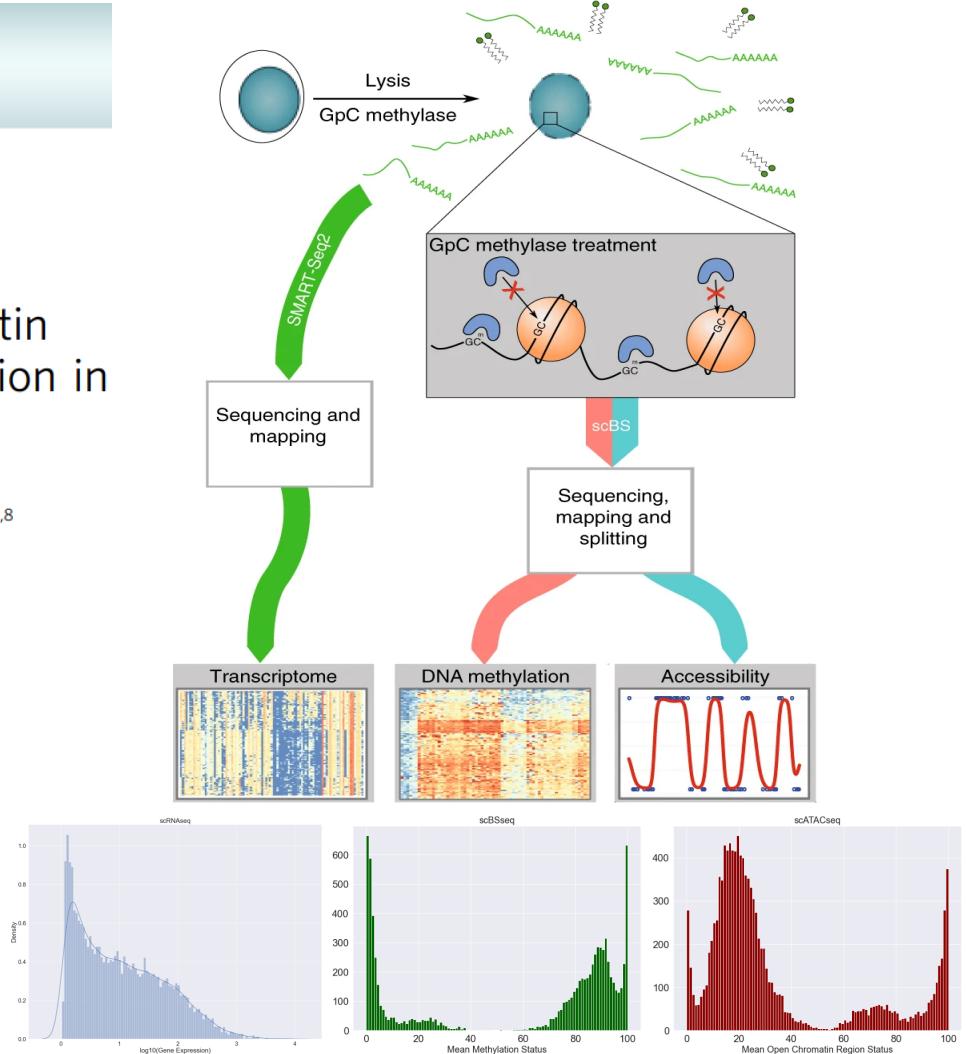
DOI: 10.1038/s41467-018-03149-4

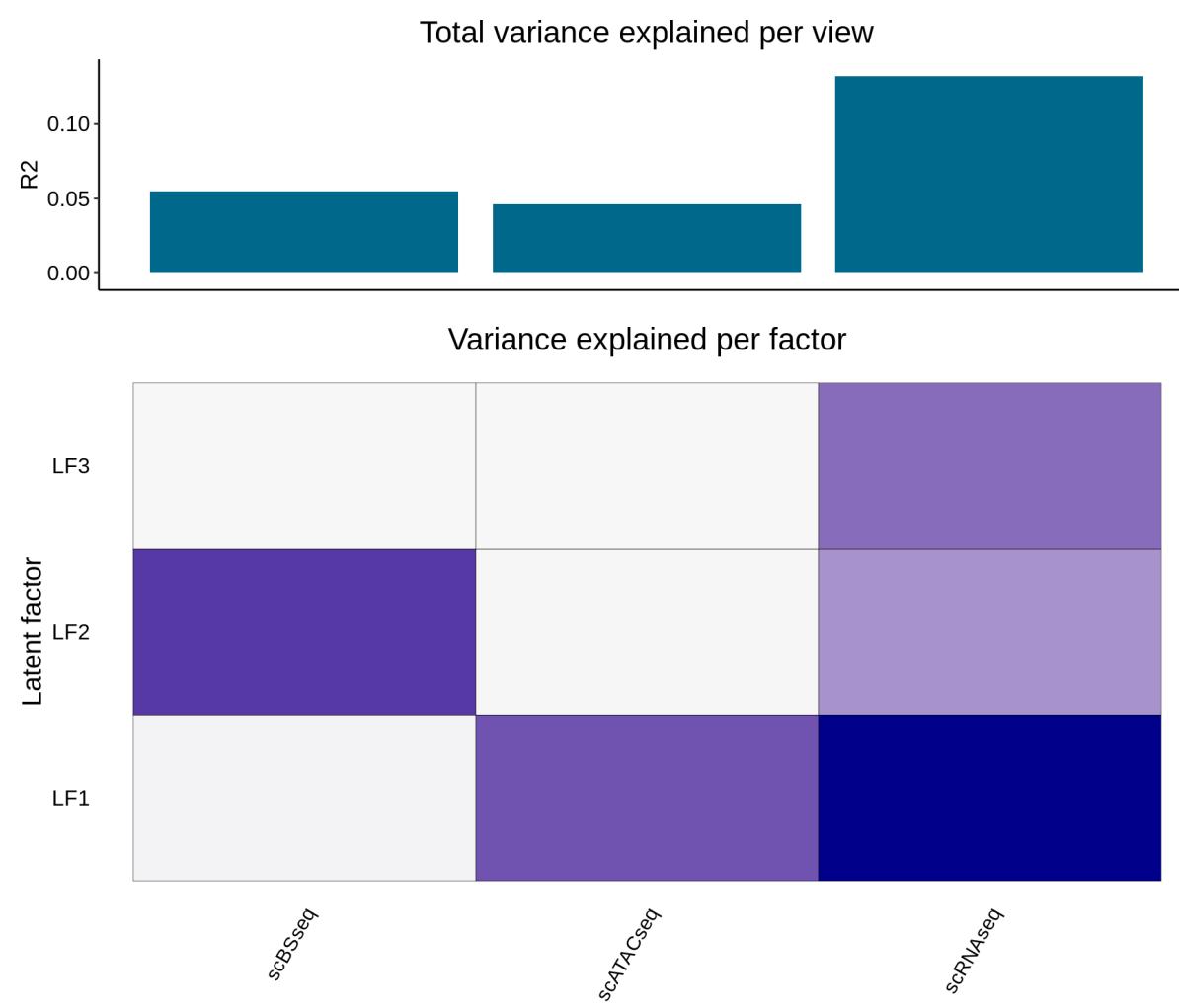
OPEN

## scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells

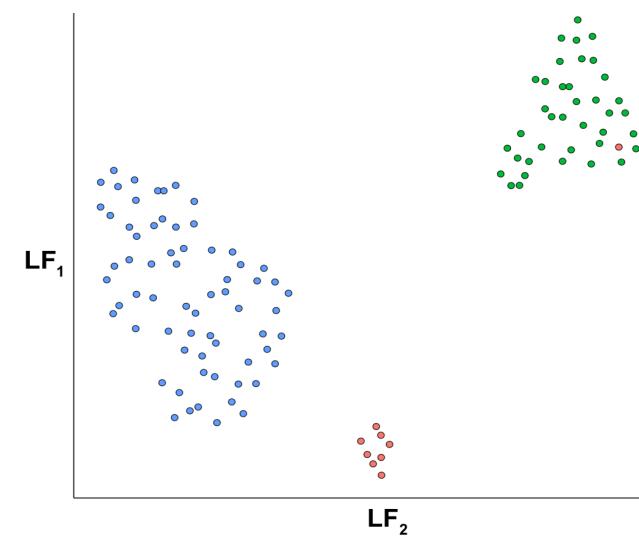
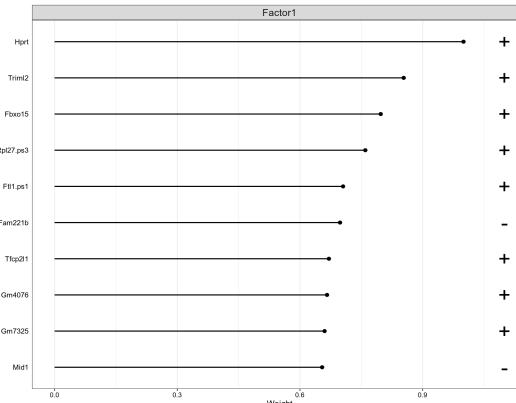
Stephen J. Clark<sup>1</sup>, Ricard Argelaguet<sup>2,3</sup>, Chantriont-Andreas Kapourani<sup>4</sup>, Thomas M. Stubbs<sup>1</sup>, Heather J. Lee<sup>1,5,6</sup>, Celia Alda-Catalinas<sup>1</sup>, Felix Krueger<sup>1</sup>, Guido Sanguinetti<sup>4</sup>, Gavin Kelsey<sup>1,8</sup>, John C. Marioni<sup>1,2,3,5</sup>, Oliver Stegle<sup>1</sup>, Wolf Reik<sup>1,5,8</sup>

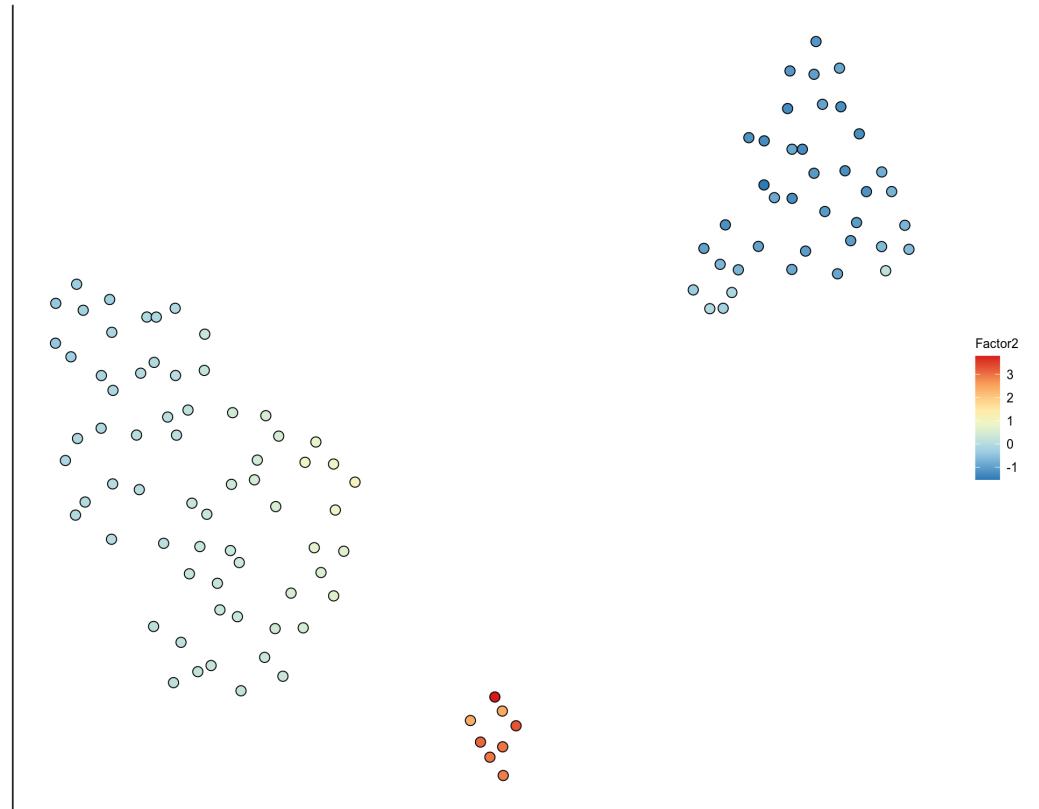
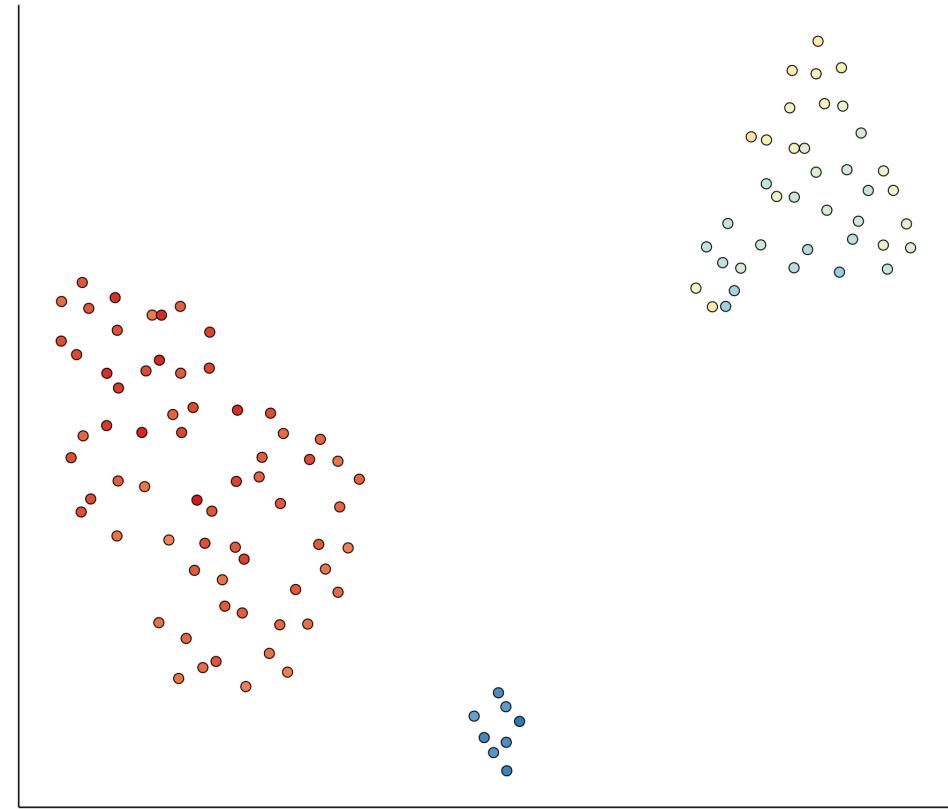
Parallel single-cell sequencing protocols represent powerful methods for investigating regulatory relationships, including epigenome-transcriptome interactions. Here, we report a single-cell method for parallel chromatin accessibility, DNA methylation and transcriptome profiling. scNMT-seq (single-cell nucleosome, methylation and transcription sequencing) uses a GpC methyltransferase to label open chromatin followed by bisulfite and RNA sequencing. We validate scNMT-seq by applying it to differentiating mouse embryonic stem cells, finding links between all three molecular layers and revealing dynamic coupling between epigenomic layers during differentiation.





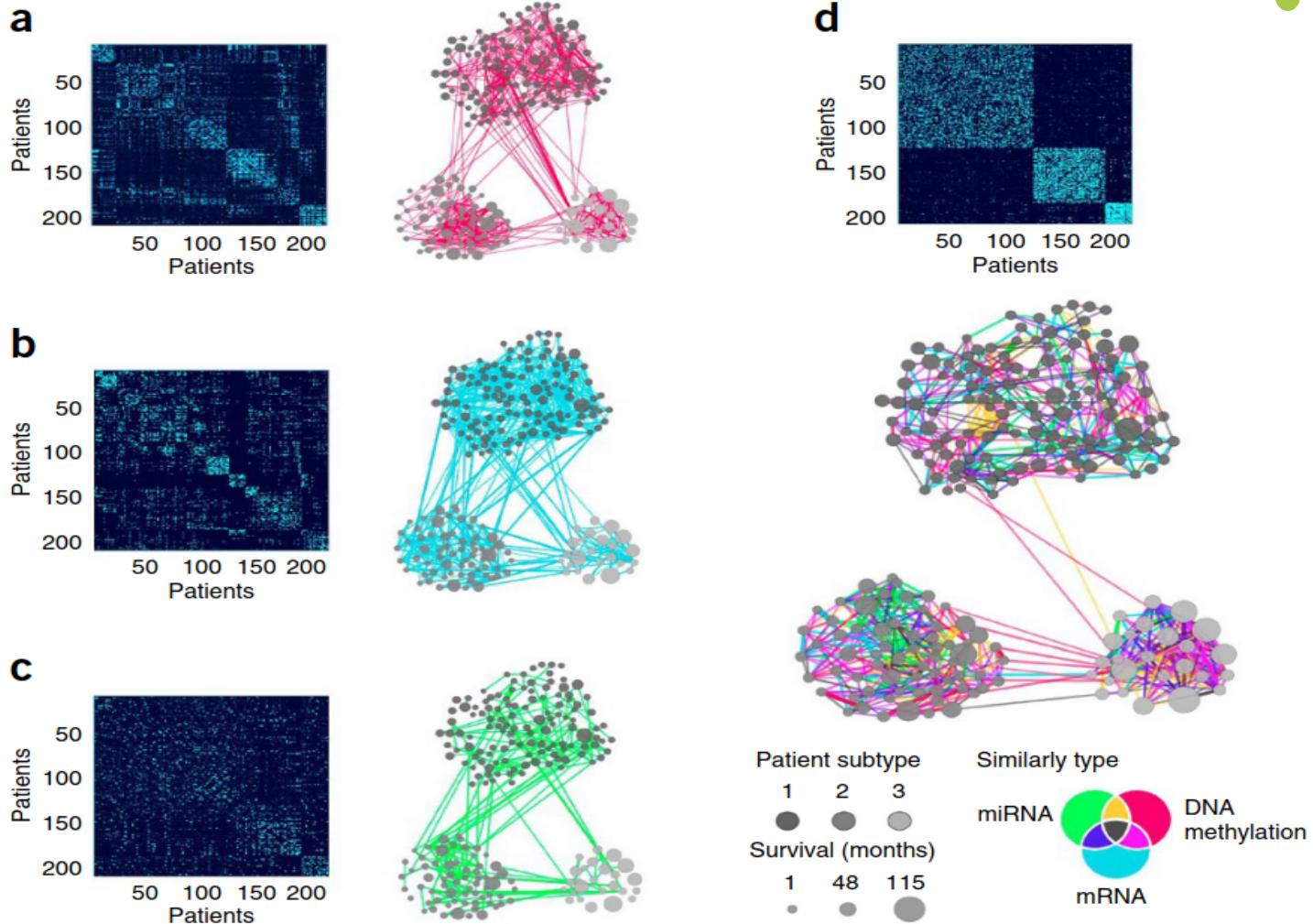
$$R_{m,k}^2 = 1 - \left( \sum_{n,d} y_{nd}^m - z_{nk} w_{kd}^m - \mu_d^m \right)^2 / \left( \sum_{n,d} y_{nd}^m - \mu_d^m \right)^2$$



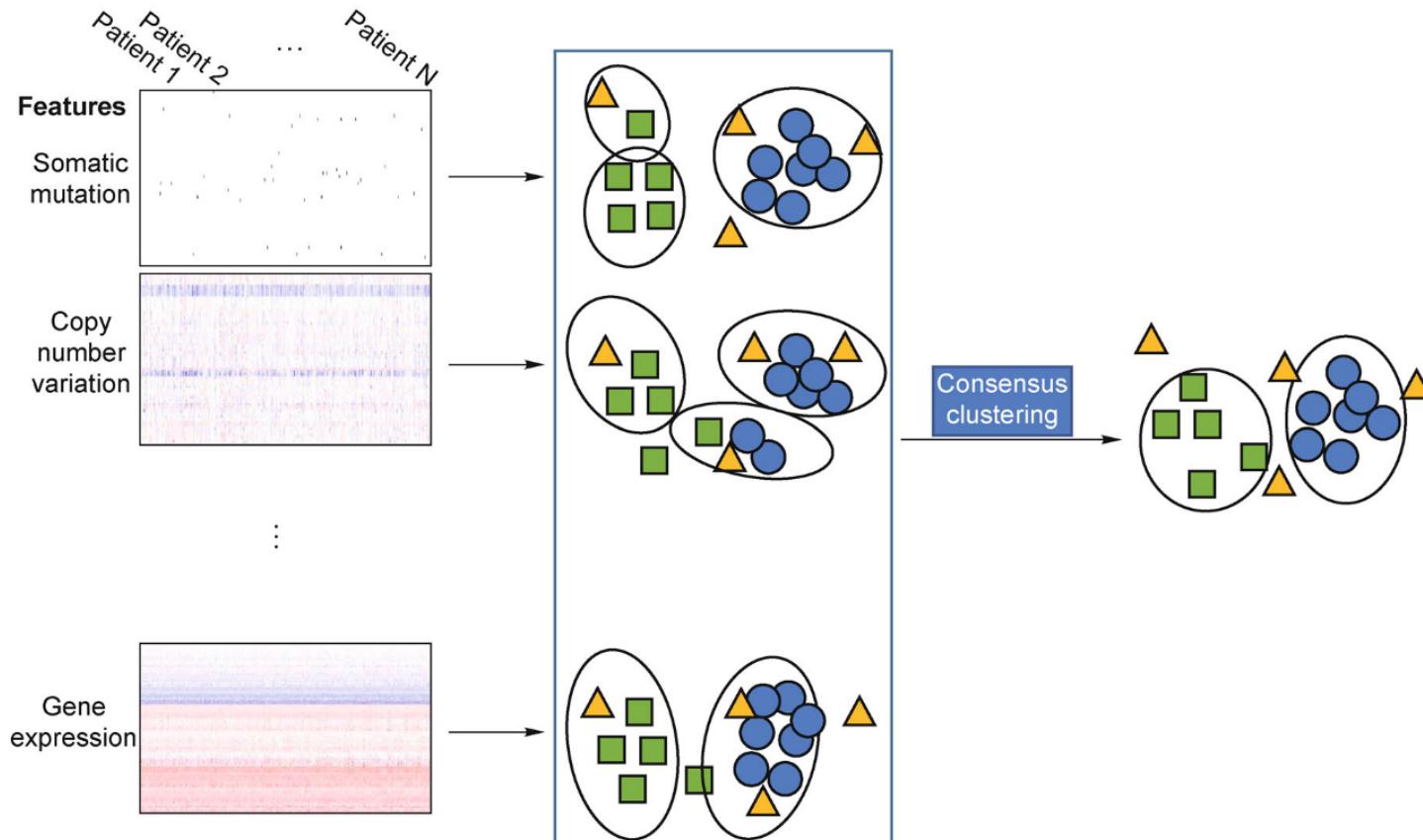


# Other unsupervised integrative Omics methods

## Similarity Network Fusion (SNF)



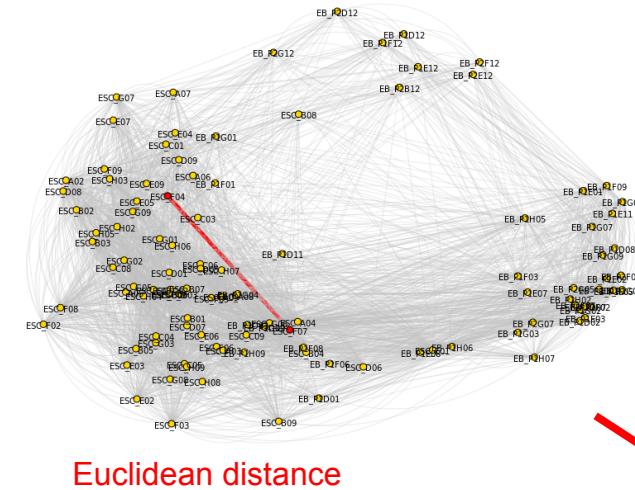
# Clustering of Clusters (COC)



**Figure 2. Clustering of clusters.** This kind of methods first clusters in every single omics dataset and then integrates the primary clustering results into final cluster assignments.

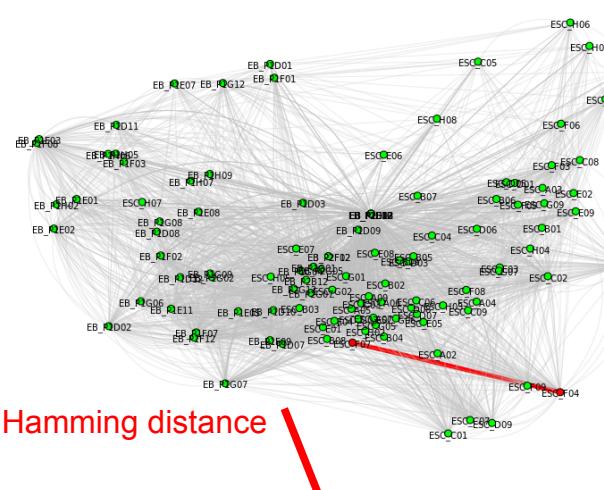
# Graph Intersection Method

scRNaseq KNN Graph



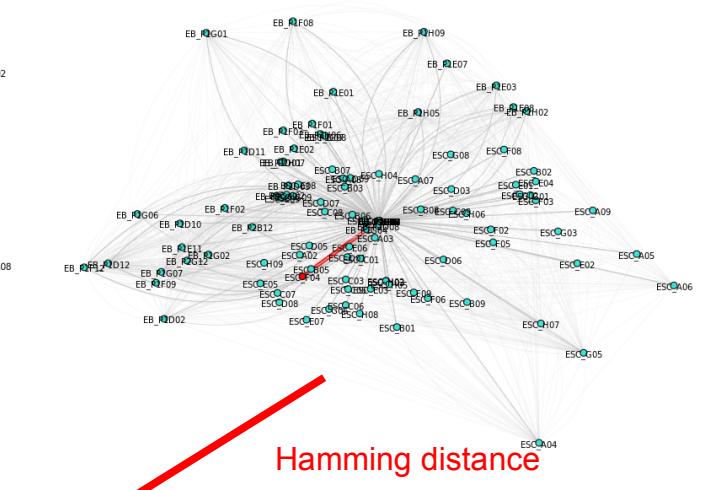
Euclidean distance

scBSseq KNN Graph



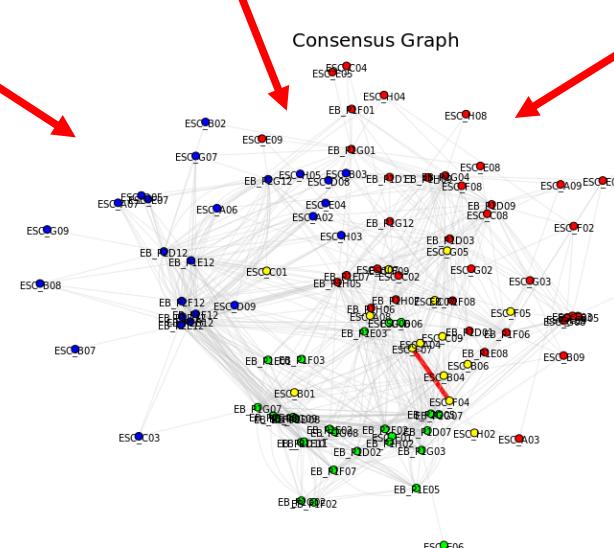
Hamming distance

scATACseq KNN Graph



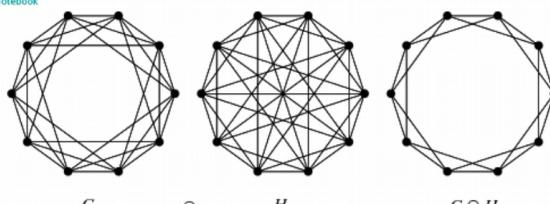
Hamming distance

Consensus Graph



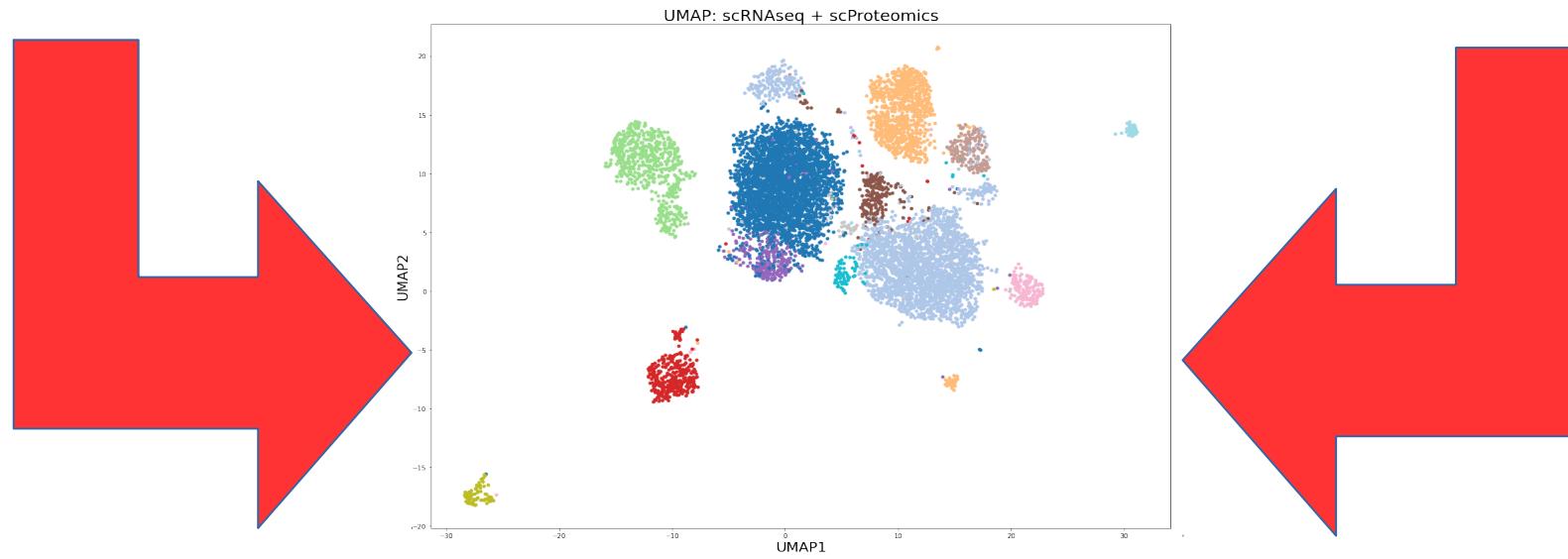
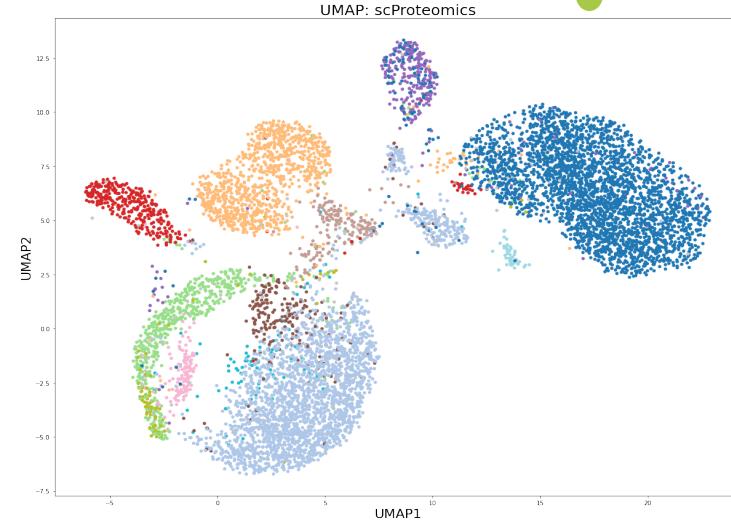
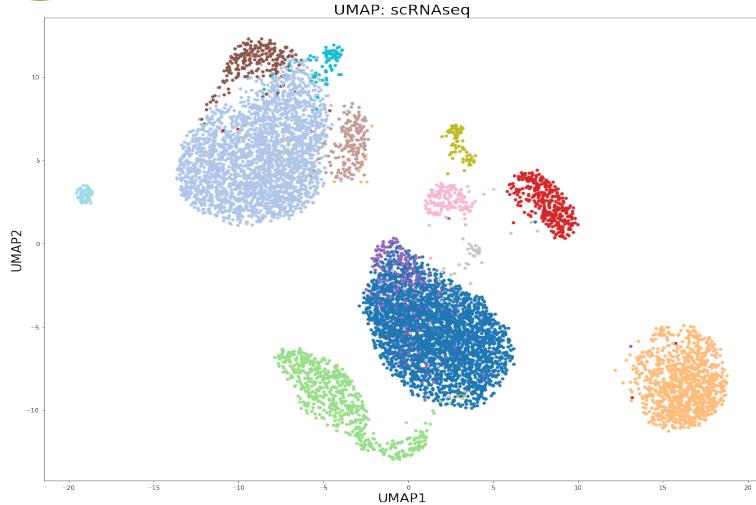
## Graph Intersection

[DOWNLOAD](#)  
Wolfram Notebook

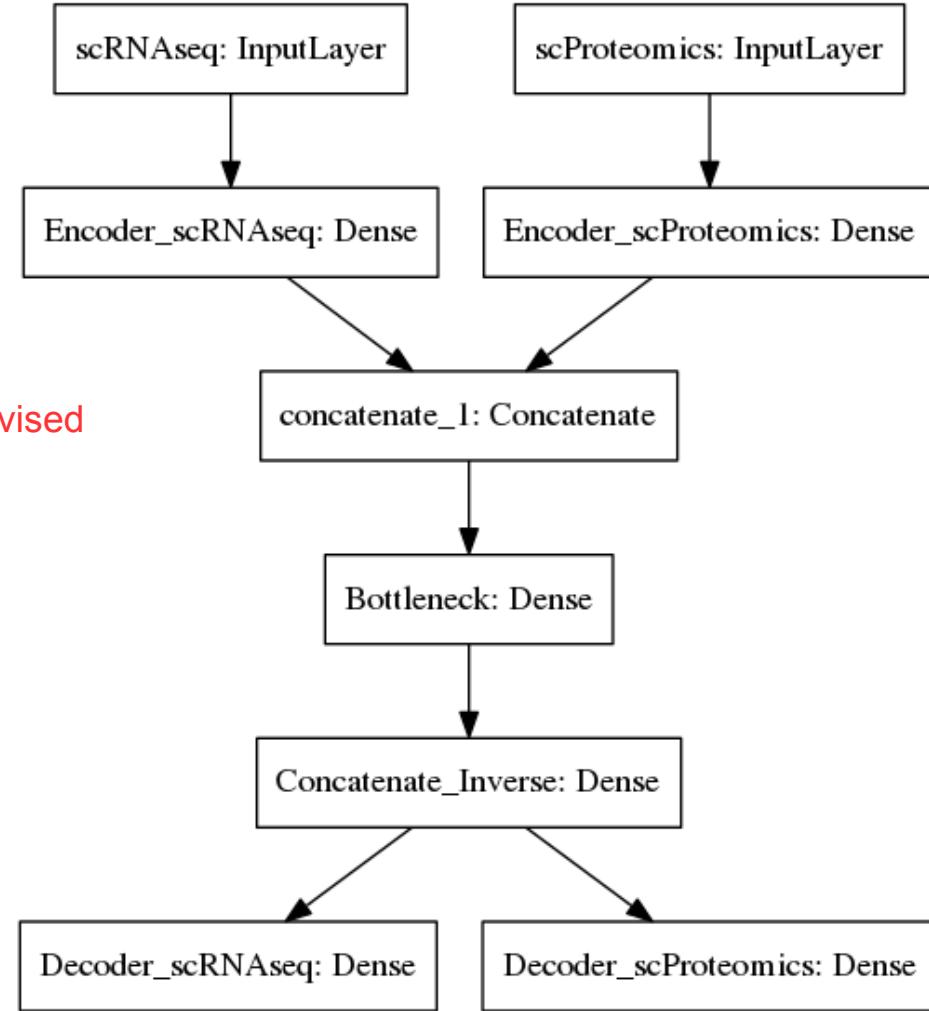
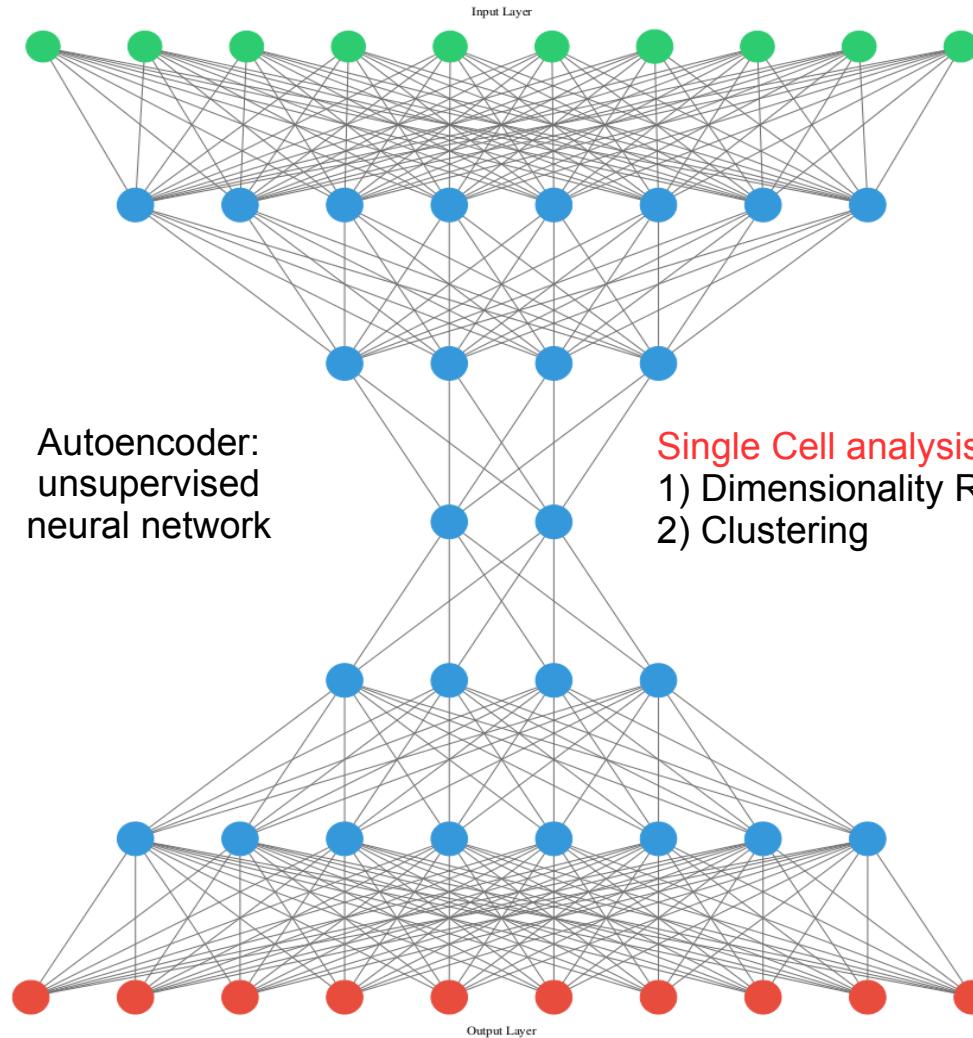


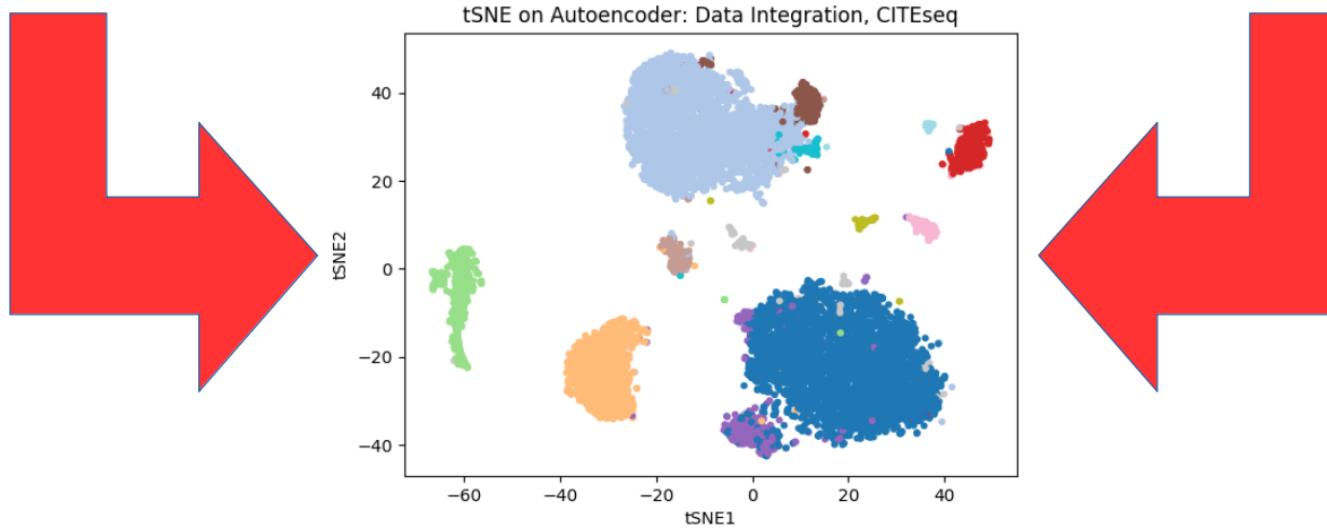
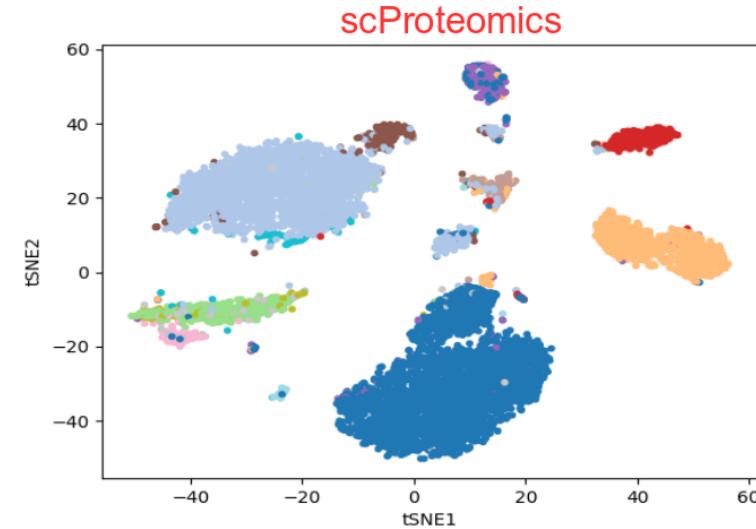
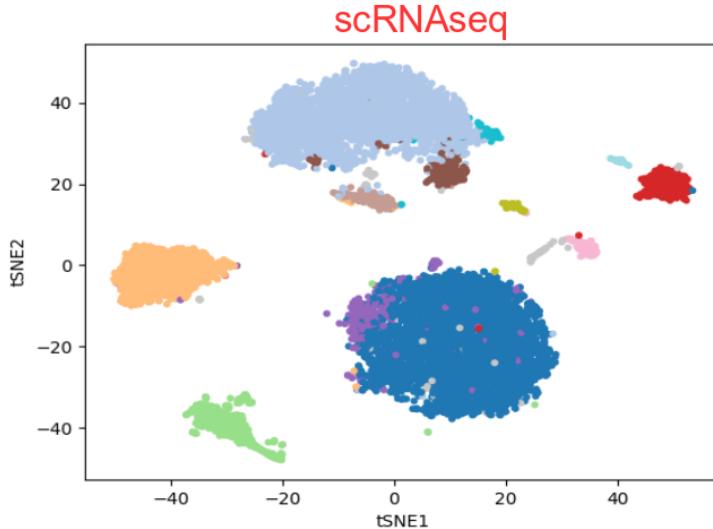
Let  $S$  be a set and  $F = \{S_1, \dots, S_p\}$  a nonempty family of distinct nonempty subsets of  $S$  whose union is  $\bigcup_{i=1}^p S_i = S$ . The intersection graph of  $F$  is denoted  $\Omega(F)$  and defined by  $V(\Omega(F)) = F$ , with  $S_i$  and  $S_j$  adjacent whenever  $i \neq j$  and  $S_i \cap S_j \neq \emptyset$ . Then a graph  $G$  is an intersection graph on  $S$  if there exists a family  $F$  of subsets for which  $G$  and  $\Omega(F)$  are isomorphic graphs (Harary 1994, p. 19). Graph intersections can be computed in the Wolfram Language using `GraphIntersection[g, h]`.

Keep edges consistently  
present across the Omics



# Autoencoder for data integration







*Knut och Alice  
Wallenbergs  
Stiftelse*



**LUNDS  
UNIVERSITET**