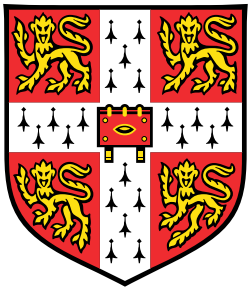
Statistical methods for the integrative analysis of single-cell multi-omics data



Ricard Argelaguet

European Bioinformatics Institute
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Robinson College

March 2020

Chapter 1

MOFA+: an improved framework for the comprehensive integration of structured single-cell data

In Chapter 2 we developed Multi-Omics Factor Analysis (MOFA), a statistical framework for the unsupervised integration of multi-modal data.
MOFA addresses key challenges in data integration, including overfitting, noise reduction, handling of missing values and improved interpretation of the model output. Hower, when applied to increasingly-large (single-cell) data sets, the inference scheme implemented in MOFA is still limited in scalability.
In addition to the increase in the number of cells, the increased experimental throughput has facilitated the study of larger numbers of experimental conditions. MOFA makes strong assumptions about the dependencies across samples and it hence has no principled way of modelling data sets where the samples are structured into multiple groups, where groups can be defined as batches, donors or different experiments. By pooling and contrasting information across studies or experimental conditions, it would be possible to obtain more comprehensive insights into the complexity underlying biological systems.

In this new Chapter we improve the first model formulation with the aim of performing integrative analysis of large-scale datasets simultaneously across multiple data modalities and across multiple groups.

1.1 Theoretical fundations

1.1.1 Gradient ascent

Gradient ascent is a first-order optimization algorithm for finding the (local) maximum of a function [**Bishop2006**, **Murphy**]. Formally, for a differentiable function $F(x)$, the iterative scheme of

gradient ascent is:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \rho^{(t)} \nabla F(\mathbf{x}^{(t)}) \tag{1.1}$$

Simply speaking, it works by taking steps proportional to the gradient ∇F evaluated at each iteration t. This leads to a monotonic sequence:

$$\mathbf{x}^0 \leq \mathbf{x}^1 \leq \mathbf{x}^1 \cdots$$

Importantly, the step size $\rho^{(t)}$ is typically adjusted at each iteration t such that it satisfies the Robbins-Monro conditions: $\sum_t \rho^{(t)} = \infty$ and $\sum_t (\rho^{(t)})^2 < \infty$. Then F is guaranteed to converge to the global maximum [**Robbins-Monro1951**] if the objective function is convex. If F is not convex, the algorithm is sensible to the initialisation $\mathbf{x}^{t=0}$ and can converge to local maxima instead of the global maximum.

Gradient ascent is appealing because of its simplicity, but

1.1.1.1 Stochastic gradient ascent

Gradient ascent becomes prohibitively slow with large datasets, mainly because of the computational cost involved in the iterative calculation of gradients [**Spall2003**].
A simple strategy to speed up gradient descent is to replace the actual gradient ∇F by an estimate $\hat{\nabla} F$ using a randomly selected subset of the data (minibatch). The iterative scheme is then defined in the same way as in standard gradient ascent:

$$\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} + \rho^{(t)} \hat{\nabla} F(\mathbf{x}^{(t)}) \tag{1.2}$$

1.1.2 Natural gradient ascent

Gradient descent becomes problematic when it comes to doing inference in probabilistic models.

Consider a probabilistic model with a hidden variable x and corresponding parameters θ, with a general objective function $\mathcal{L}(\theta)$. From the definition of a derivative:

$$\nabla \mathcal{L}(\theta) = \lim_{||h|| \to 0} \frac{\mathcal{L}(\theta + h) - \mathcal{L}(\theta)}{||h||}$$

where h represents an infinitesimally small positive step in the space of θ.
To find the direction of steepest descent, one would need to search over all possible directions d in an infinitely small distance h, and select the $\hat{d}$ that gives the largest gradient:

$$\nabla \mathcal{L}(\theta) = \lim_{h \to 0} \frac{1}{h} \underset{d\, s.t. \|d\| = h}{\arg\max}\ \mathcal{L}(\theta + d) - \mathcal{L}(\theta)$$

Importantly, this operation requires a distance metric to quantify what a *small* distance h means. In standard gradient descent, this is measured using an Euclidean norm, and the direction of steepest ascent is hence dependent on the Euclidean geometry of the θ space. Why is this problematic when

working with probability distributions?
The problem of using an Euclidean distance to optimise parameters of distributions is that it does not consider the uncertainity that underlies probability distributions. A small step from $\theta^{(t)}$ to $\theta^{(t+1)}$ does not guarantee an equivalently small change from $\mathcal{L}(\theta^{(t)})$ to $\mathcal{L}(\theta^{(t+1)})$.
To illustrate this, consider the following example of four random variables

$$\begin{aligned} \psi_1 &\sim \mathcal{N}(0 \,|\, 5) \qquad & \psi_3 &\sim \mathcal{N}(0 \,|\, 1) \\ \psi_2 &\sim \mathcal{N}(10 \,|\, 5) \qquad & \psi_4 &\sim \mathcal{N}(10 \,|\, 1) \end{aligned} \tag{1.3}$$

Using the Euclidean metric, the distance between ψ_1 and ψ_2 is the same as the distance between ψ_3 and ψ_4. However, the distance in distribution space (measured for example by the KL divergence) is much larger between ψ_1 and ψ_2 than between ψ_3 and ψ_4 (**??**).

This basic simulation suggests that replacing the Euclidean distance by the KL divergence as a distance metric may be more appropriate in the context of probabilistic modelling:

$$\nabla_{KL}\mathcal{L}(\theta) = \lim_{h \to 0} \frac{1}{h} \underset{d\, s.t. KL[p_\theta || p_{\theta+d}] = h}{\arg\max} \mathcal{L}(\theta + d) - \mathcal{L}(\theta)$$

The direction of steepest ascent measured by the KL divergence is called the natural gradient [**Amari1998**, **Martens2014**].
To find the optimal $\hat{d}_{KL}$, one needs to solve the following optimisation problem:

$$\arg\min_{d} \mathcal{L}(\theta + d) \qquad \text{subject to} \quad KL[p_\theta || p_{\theta+d}] < c$$

where c is an arbitrary constant. We will not derive the solution, but this can be solved by introducing Lagrange multipliers and Taylor expansions (see [**Amari1998**, **Kristiadi2019**]). The solution corresponds to the standard (Euclidean) gradient pre-multiplied by the inverse of the Fisher Information Matrix of $q(x|\theta)$:

$$\hat{d}_{KL} \propto \mathbf{F}^{-1}(\theta) \nabla_\theta \mathcal{L}(\theta) \tag{1.4}$$

where $\mathbf{F}(\theta)$ is defined as

$$\mathbf{F}(\theta) = \mathbb{E}_{q(x|\theta)}[(\nabla_\theta \log q(x|\theta))(\nabla_\theta \log q(x|\theta))^T]$$

In conclusion, while the standard gradient points to the direction of steepest ascent in Euclidean space, the natural gradient points to the direction of steepest ascent in a space where distances are defined by the KL divergence [**Kristiadi2019**, **Amari1998**, **Hoffman2012**].

1.2 Model description

In MOFA v2 we generalise the model to a disjoint set of M input views (i.e. groups of features) and G input groups (i.e. groups of samples).

The data is factorised according to the following model:

$$\mathbf{Y}_g^m = \mathbf{Z}_g \mathbf{W}^{mT} + \boldsymbol{\epsilon}_g^m \tag{1.5}$$

where $\mathbf{Z}_g \in \mathbb{R}^{N_g \times K}$ are a set of G matrices that contains the factor values for the g-th group and $\mathbf{W}^m \in \mathbb{R}^{D_m \times K}$ are a set of M matrices that define the feature weights for the m-th view. $\boldsymbol{\epsilon}_g^m \in \mathbb{R}^{D_m}$ captures the residuals, or the noise for each feature in ech group.

EXPLAIN INTUITION REGRESSED OUT GROUP EFFECT

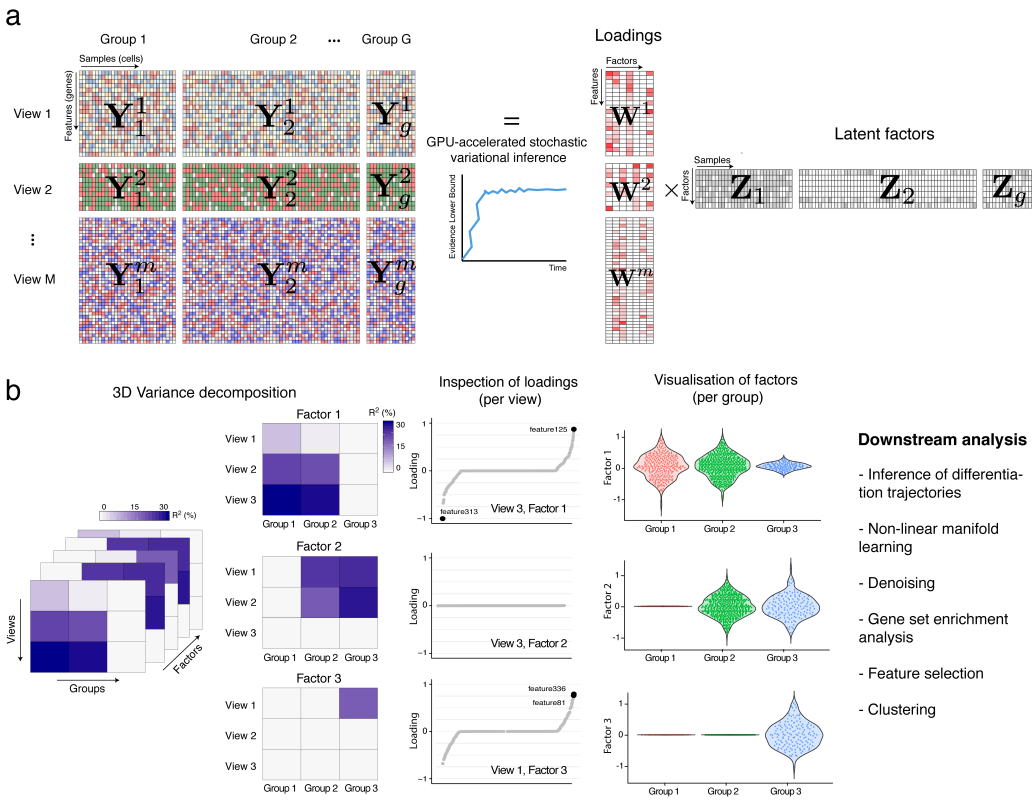


Figure 1.1:
Multi-Omics Factor Analysis v2 (MOFA+) provides an unsupervised framework for the integration of multi-group and multi-view single-cell data.
(a) Model overview: the input consists of multiple data sets structured into M views and G groups. Views consist of non-overlapping sets of features that can represent different assays. Analogously, groups consist of non-overlapping sets of samples that can represent different conditions or experiments. Missing values are allowed in the input data. MOFA+ exploits the dependencies between the features to learn a low-dimensional representation of the data (Z) defined by K latent factors that capture the global sources of molecular variability. For each factor, the weights (W) link the high-dimensional space with the low-dimensional manifold and provide a measure of feature importance. The sparsity-inducing priors on both the factors and the weights enable the model to disentangle variation that is unique to or shared across the different groups and views. Model inference is performed using GPU-accelerated stochastic variational inference.
(b) The trained MOFA+ model can be queried for a range of downstream analyses: 3D variance decomposition, quantifying the amount of variance explained by each factor in each group and view, inspection of feature weights, visualisation of factors and other applications such as clustering, inference of non-linear differentiation trajectories, denoising and feature selection.

1.2.1 Model priors and likelihood

1.2.1.1 Prior on the weights

This remains the same as in MOFA v1. We adopt a two-level sparsity prior with an Automatic Relevance Determination per factor and view, and a feature-wise spike-and-slab prior (reparametrised[**Titsias2011**]):

$$p(\hat{w}_{dk}^m, s_{dk}^m) = \mathcal{N}\left(\hat{w}_{dk}^m \,|\, 0, 1/\alpha_k^m\right) \mathrm{Ber}(s_{dk}^m \,|\, \theta_k^m) \tag{1.6}$$

with the corresponding conjugate priors for θ and α:

$$p(\theta_k^m) = \text{Beta}\left(\theta_k^m \,|\, a_0^\theta, b_0^\theta\right) \tag{1.7}$$

$$p(\alpha_k^m) = \mathcal{G}\left(\alpha_k^m \,|\, a_0^\alpha, b_0^\alpha\right) \tag{1.8}$$

The aim of the ARD prior is to disentangle the activity of factors to the different views, such that the weight vector $\mathbf{w}_{:,k}^m$ is shrunk to zero if the factor k does not explain any variation in view m. The aim of the spike-and-slab prior is to push individual weights to zero to yield a more interpretable solution.
For more details, we refer the reader to Chapter 2.

1.2.1.2 Prior on the factors

In MOFA v1 we adopted an isotropic Gaussian prior:

$$p(z_{nk}) = \mathcal{N}\left(z_{nk} \,|\, 0, 1\right) \tag{1.9}$$

which assumes *a priori* an unstructured latent space. This is the assumption that we want to break. Following the same logic as in the factor and view-wise ARD prior, the integration of multiple groups of samples requires introducing a *structured* prior that captures the existence of different groups, such that some factors are allowed to be active in different subsets of groups.

To formalise the intuition above we simply need to copy the double sparsity prior from the weights to the factors:

$$p(\hat{z}_{nk}^g, s_{nk}^g) = \mathcal{N}(\hat{z}_{nk}^g \,|\, 0, 1/\alpha_k^g)\,\text{Ber}(s_{nk}^g \,|\, \theta_k^g) \tag{1.10}$$

$$p(\theta_k^g) = \text{Beta}\left(\theta_k^g \,|\, a_0^\theta, b_0^\theta\right) \tag{1.11}$$

$$p(\alpha_k^g) = \mathcal{G}\left(\alpha_k^g \,|\, a_0^\alpha, b_0^\alpha\right), \tag{1.12}$$

where g is the index of the sample groups.
Notice that the spike-and-slab prior is introduced for completeness but is not necessarily required, and can be disabled by fixing $\mathbb{E}[\theta_k^g] = 1$.

1.2.1.3 Prior on the noise

The variable $\boldsymbol{\epsilon}$ captures the residuals, or the noise, which is assumed to be normally distributed and heteroskedastic. In MOFA v2 we generalise the noise to have an estimate per individual feature and per group:

$$p(\epsilon_g^m) = \mathcal{N}\left(\epsilon_g^m \,|\, 0, /\tau_g^m \mathbf{I}_{Dm}\right) \tag{1.13}$$

$$p(\tau_g^m) = \prod_{d=1}^{D_m} \mathcal{G}\left(\tau_g^m \,|\, a_0^\tau, b_0^\tau\right) \tag{1.14}$$

This formulation is important to capture the (realistic) events where a specific feature may be highly variable in one group but non-variable in another group.
In addition, as in MOFA v1, non-gaussian noise models can also be defined, but unless otherwise stated, we will always assume Gaussian residuals.

1.2.1.4 Graphical model

In summary, the updated model formulation introduces asymmetric sparsity prior in both the weights and the factors, which enables the model to simultaneously integrate multiple views as well as multiple groups of samples:

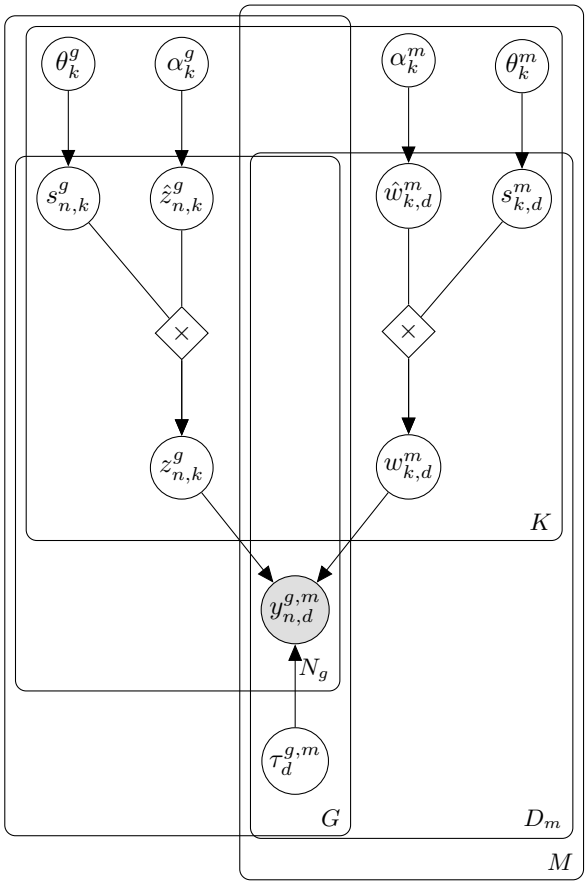


Figure 1.2: Graphical model for MOFA+$^{\text{oo}}$ .The white circles represent hidden variables that are inferred by the model, whereas the grey circles represent the observed variables. There are a total of five plates, each one representing a dimension of the model: M for the number of views, G for the number of groups, K for the number of factors, D_m for the number of features in view m and N_g for the number of samples in group g

1.2.2 Solving the rotational invariance problem

Conventional Factor Analysis is invariant to rotation in the latent space[**Zhao2009**]. To demonstrate this property, let us apply an arbitrary rotation to the loadings and the factors, specified by the

rotation matrix $\mathbf{R} \in \mathbb{R}^{K \times K}$:

$$\tilde{\mathbf{Z}} = \mathbf{Z}\mathbf{R}^{-1}$$
$$\tilde{\mathbf{W}} = \mathbf{R}\mathbf{W}$$

First, note that the model likelihood is unchanged by this rotation, irrespective of the prior distribution used.

$$p(\mathbf{Y}|\tilde{\mathbf{Z}}\tilde{\mathbf{W}}, \tau) = p(\mathbf{Y}|\mathbf{Z}\mathbf{R}^{-1}\mathbf{R}\mathbf{W}, \tau) = p(\mathbf{Y}|\mathbf{Z}\mathbf{W}, \tau)$$

However, the prior distributions of the factors and the loadings are only invariant to rotations when using isotropic Normal priors:

$$\ln p(\mathbf{W}) \propto \sum_{k=1}^{K}\sum_{d=1}^{D} w_{d,k}^2 = \mathrm{Tr}(\mathbf{W}^T\mathbf{W}) = \mathrm{Tr}(\mathbf{W}^T\mathbf{R}^{-1}\mathbf{R}\mathbf{W}) = \mathrm{Tr}(\tilde{\mathbf{W}}^T\tilde{\mathbf{W}})$$

where we have used the property $\mathbf{R}^T = \mathbf{R}^{-1}$ that applies to rotation matrices. The same derivation follows for the factors $\mathbf{Z}$.
In practice, this property renders conventional Factor Analysis unidentifiable, as shown using simulations in Figure SX (TO-FILL), hence limiting its interpretation and applicability.

Sparsity assumptions, however, partially address the rotational invariance problem [**Hore2015-thesis**]. When using independent identically distributed spike-and-slab priors the proof above cannot be applied, hence making the proposed factor analysis model not rotationally invariant.

It is important to remark that the factors are nonetheless invariant to permutations. This implies that under different initial conditions, the order of the factors is not necessarily the same in independent model fittings. To address this we manually sort factors *a posteriori* based on total variance explained.

1.2.3 GPU-accelerated stochastic variational inference

1.3 Model validation

We validated the new features of MOFA+ using simulated data drawn from its generative model.

1.3.1 Stochastic variational inference

We simulated data with varying sample sizes, with the other dimensions fixed to $M = 3$ views, $G = 3$ groups, $D = 1000$ features (per view), and $K = 25$ factors.

We trained a set of models with (deterministic) variational inference (VI) and a set of models with stochastic variational inference (SVI). Overall, we observe that SVI yields Evidence Lower Bounds

that matched those obtained from conventional inference across a range of batch sizes, learning rates and forgetting rates.

In terms of speed, GPU-accelerated SVI inference was up to ≈ 20x faster than VI, with speed differences becoming more pronounced with increasing number of cells. For completeness, we also compared the the convergence time estimates for SVI when using CPU versus GPU. We observe that for large sample sizes there is a speed improvement even when using CPUs, although these advantages become more prominent when using GPUs.

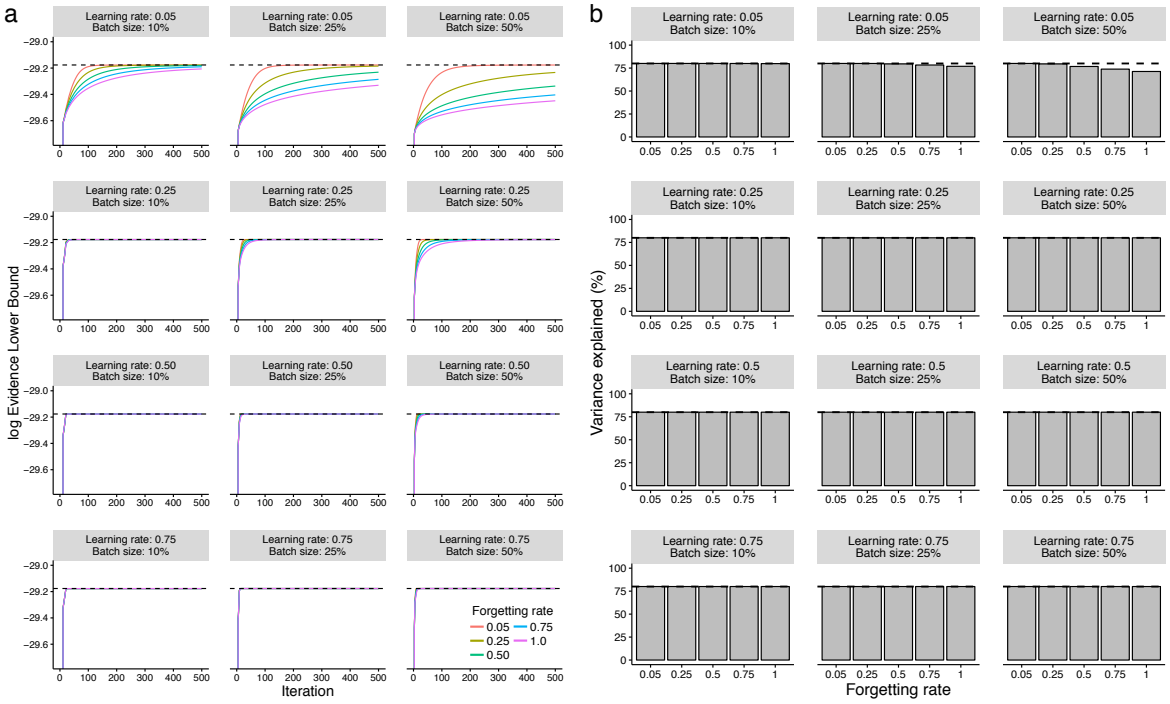


Figure 1.3: Validation of stochastic variational inference using simulated data.
(a) Line plots display the iteration number of the inference (x-axis) and the log- Evidence Lower Bound (ELBO) on the y-axis. Panels correspond to different values of batch sizes (10%, 25%, 50% of the data) and initial learning rates (0.05, 0.25, 0.5, 0.75). Colors correspond to different forgetting rates (0.05, 0.25, 0.5, 0.75, 1.0). The dashed horizontal line indicates the ELBO achieved using standard VI.
(b) Bar plots display the forgetting rate (x-axis) and the total variance explained (%) in the y-axis. Panels correspond to different values of batch sizes (10%, 25%, 50% of the data) and initial learning rates (0.05, 0.25, 0.5, 0.75). The dashed line indicates the variance explained achieved using standard VI.

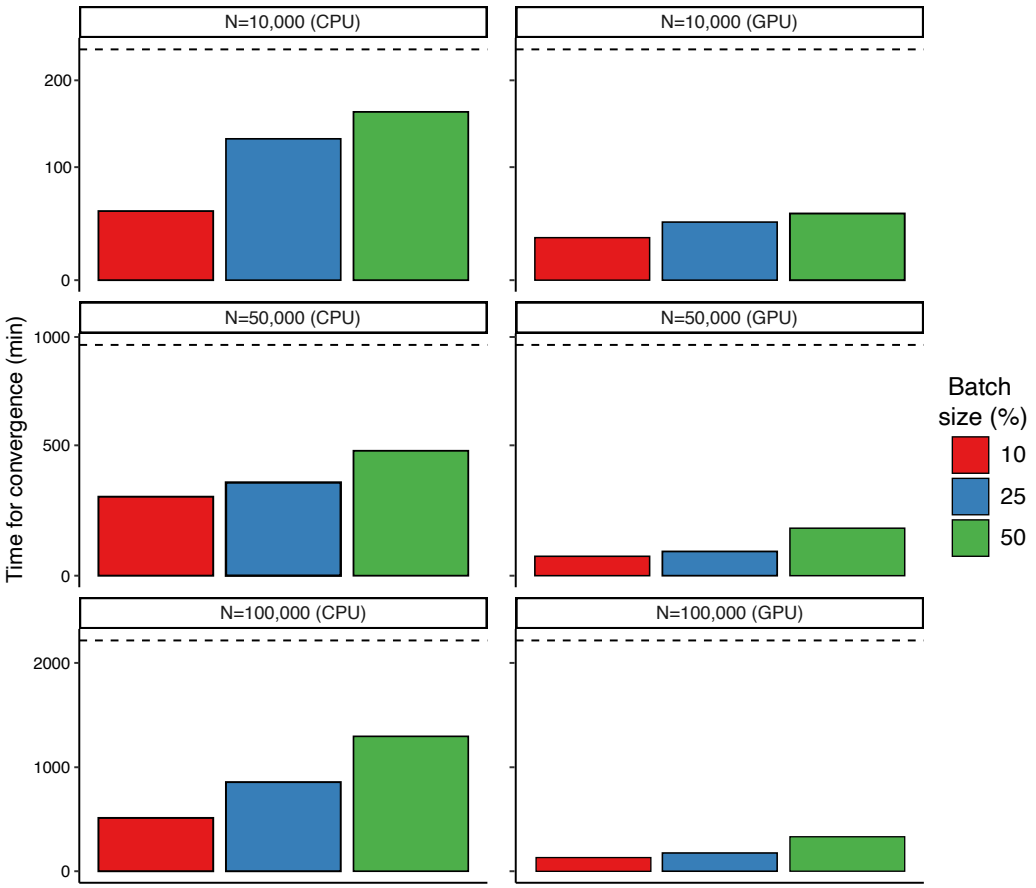


Figure 1.4: Evaluation of convergence speed for stochastic variational inference using simulated data.
Bar plots show the time elapsed for training MOFA+ models with stochastic variational inference (SVI). Colors represent different batch sizes (10%, 25% or 50%). The dashed line indicates the training time for standard VI.
VI models were trained using a single E5-2680v3 CPU. SVI models were trained either using a single E5-2680v3 CPU (first column) or using an Nvidia GTX 1080Ti GPU (second column).

1.3.2 Multi-group structure

Finally, we evaluated whether the double view and group-wise sparsity prior enables the detection of factors with simultaneous differential activity between groups and views.
We simulated data with the following parameters: $M = 2$ modalities, $G = 2$ groups, $D = 1000$ features, $N = 1000$ samples and $K = 10$ factors. Differential factor activities are incorporated in the simulation process by turning some factors off in random sets of modalities and groups (Figure 1.5, see ground truth). The task is to recover the true factor activity structure given a random initialisation.
We fit three models: Bayesian Factor Analysis (no sparsity priors), MOFA v1 (only view-wise sparsity prior) and MOFA+ (view-wise and group-wise sparsity prior). Indeed, we observe that when having factors that explain differing amounts of variance across groups and across views, MOFA+ was able to more accurately reconstruct the true factor activity patterns:

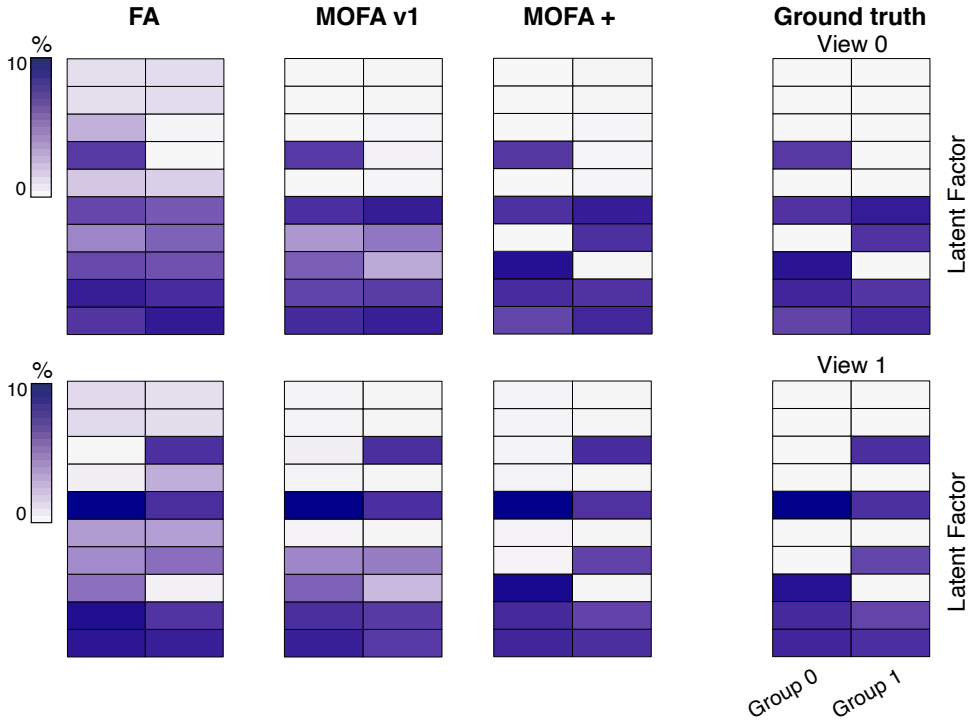


Figure 1.5: Validation of group-wise ARD prior in the factors using simulated data. Representative example of the resulting variance explained patterns. The first row of heatmaps correspond to modality 0 and the second row to modality 1. In each heatmap, the first column corresponds to group 0 and the second column to group 1. Rows correspond to the inferred factors. The colour scale displays the percentage of variance explained by a given factor in a given modality and group. The heatmaps displayed in columns one to three show the solutions yielded by different models (Bayesian Factor Analysis; MOFA; MOFA+). The ground truth is shown in the right panel.

1.4 Applications

1.4.1 Integration of a heterogeneous time-course single-cell RNA-seq dataset

To demonstrate the novel multi-group integration framework, we considered a time course scRNA-seq dataset comprising 16,152 cells that were isolated from a total of 8 mouse embryos from developmental stages E6.5, E7.0 and E7.25 (two biological replicates per stage), encompassing post-implantation and early gastrulation[**Pijuan-Sala2019**]. This data set, which has been introduced in Chapter 3, consists on a single view (RNA expression) but with a clear group structure where cells belongs to different biological replicates at different time points. Different embryos are expected to contain similar subpopulations of cells but also some differences due to developmental progression. As a proof of principle, we used MOFA+ to disentangle stage-specific transcriptional signatures from signatures that are shared across all stages.

MOFA+ identified 7 Factors that explain at least 1% of variance (across all groups). Notably, this latent representation captures between 35% and 55% of the total transcriptional heterogeneity per embryo:

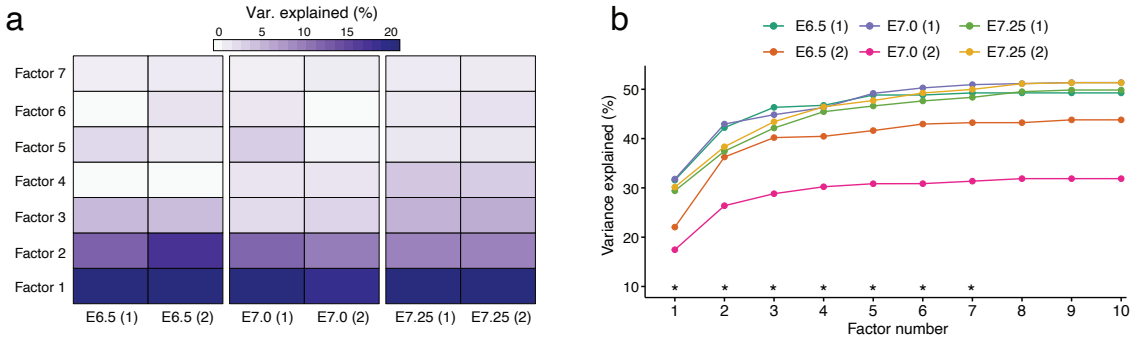


Figure 1.6: Variance explained estimates by MOFA+ applied to the gastrulation scRNA-seq atlas.
(a) Heatmap displays the variance explained (%) for each factor (rows) in each group (pool of mouse embryos at a specific developmental stage, columns). The bar plots show the variance explained per group with all factors.
(b) Cumulative variance explained (per group, y-axis) versus factor number (x-axis). Asterisks indicate the factors that are selected for downstream analysis (minimum of 1% variance explained).

1.4.1.1 Characterisation of individual factors

Some factors recover the existence of post-implantation developmental cell types, including extra-embryonic (ExE) tissue (Factor 1 and Factor 2), and the emergence of mesoderm cells from the primitive streak (Factor 4). Consistently, the top weights for these factors are enriched for lineage-specific gene expression markers, including *Ttr* and *Apoa1* for ExE endoderm Figure 1.7; *Rhox5* and *Bex3* for ExE ectoderm (not shown); *Mesp1* and *Phlda2* for nascent mesoderm Figure 1.8. Other factors captured technical variation due to metabolic stress that affects all batches in a similar fashion (Factor 3, Figure 1.9).
The characterisation of other factors is described in[**Argelaguet2020**] and is not reproduced here for simplicity.

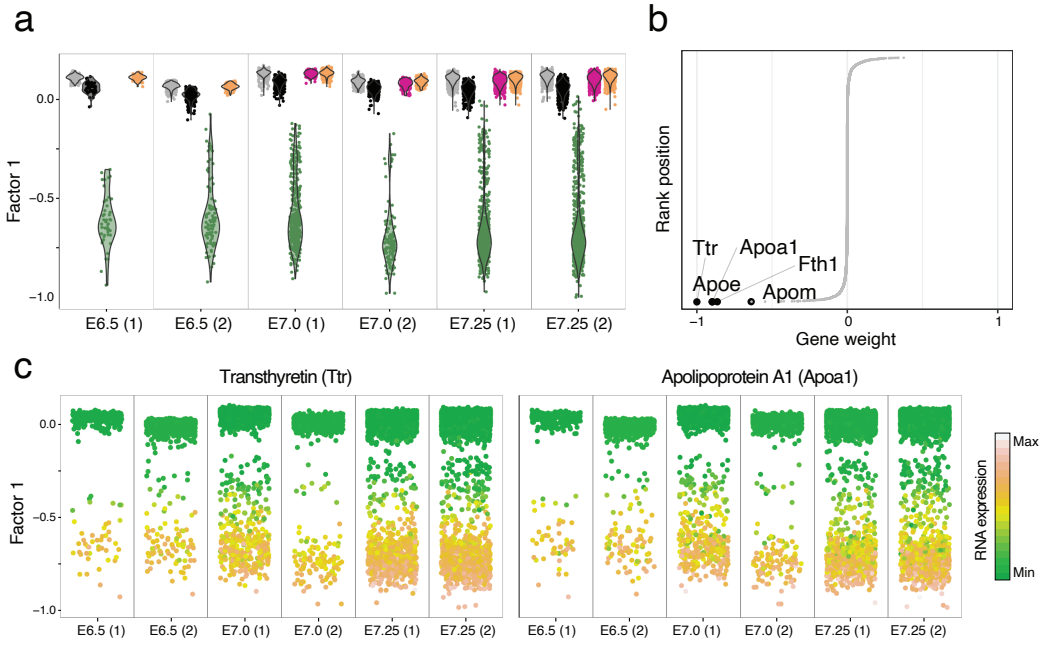


Figure 1.7: Characterisation of Factor 1 as extra-embryonic (ExE) endoderm formation.
(a) Beeswarm plot of Factor values for each group. Cells are grouped and coloured by cell type.
(b) Plot of gene weights. Highlighted are the top five genes with largest weight (in absolute values)
(c) Beeswarm plot of Factor values for each group. Cells are coloured by the expression of the two genes with largest weight (in absolute values).

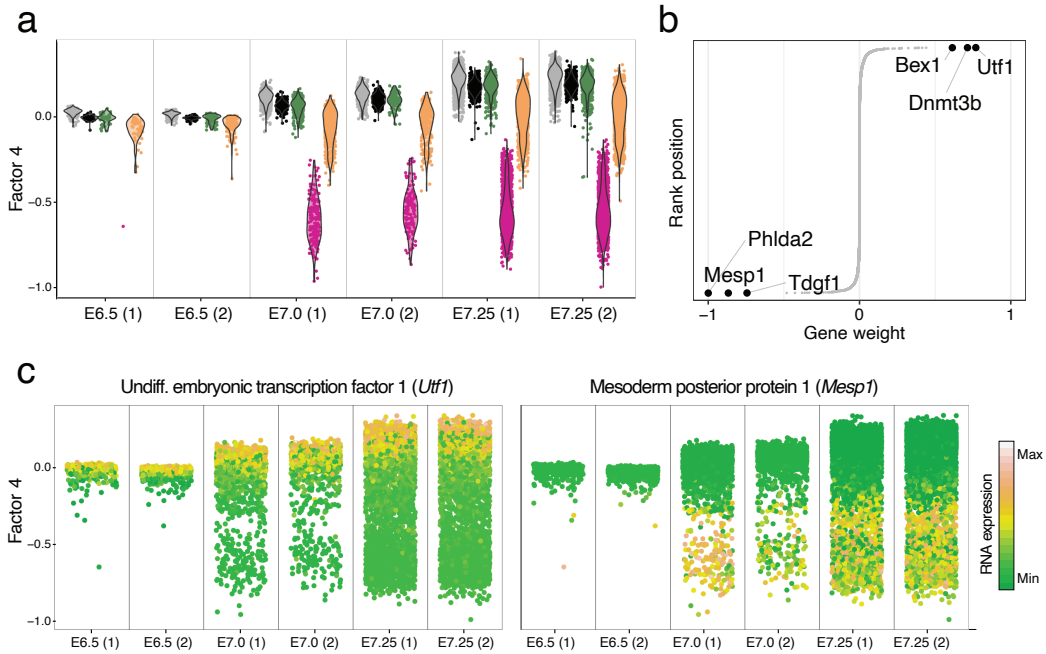


Figure 1.8:
Characterisation of Factor 4 as mesoderm commitment.
(a) Beeswarm plot of Factor values for each group. Cells are grouped and coloured by cell type.
(b) Plot of gene weights. Highlighted are the top five genes with largest weight (in absolute values)
(c) Beeswarm plot of Factor values for each group. Cells are coloured by the expression of the two genes with largest weight (in absolute values).

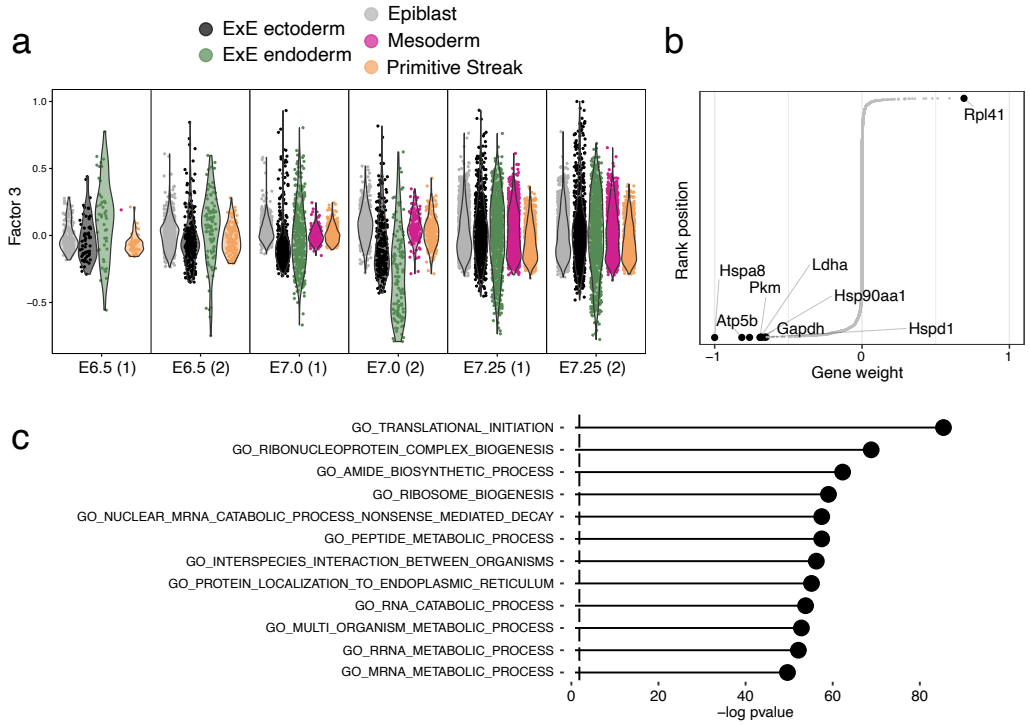


Figure 1.9:
Characterisation of Factor 3 as cell-to-cell differences in metabolic activity.
(a) Beeswarm plot of Factor values for each group. Cells are grouped and coloured by cell type.
(b) Plot of gene weights. Highlighted are the top seven genes with largest weight (in absolute values)
(c) Gene set enrichment analysis applied to the gene weights using the Reactome gene sets[**Fabregat2015**]. Significance is assessed via a parametric. Resulting p-values were adjusted for multiple testing using the Benjamini-Hochberg procedure.

Interestingly, Factors display differeent signatures of activity (variance explained) across developmental stages. For example, the variance explained by Factor 1 remains constant across developmental progression (Figure 1.6), indicating that commitment to ExE endoderm fate occurs early in the embryo and the proportion of this cell type remains relatively constant. In contrast, the activity of Factor 4 increases with developmental progression, consistent with a higher proportion of cells committing to mesoderm after ingression through the primitive streak.

In conclusion, this application shows how MOFA+ can identify biologically relevant structure in *structured* scRNA-seq datasets.

1.4.2 Identification of context-dependent methylation signatures associated with cellular diversity in the mammalian cortex

As a second use case, we considered how MOFA+ can be used to investigate cellular heterogeneity in epigenetic signatures between populations of neurons. This application illustrates how a multi-group and multi-view structure can be defined from seemingly uni-modal data.

We considered a data set of of 3,069 cells isolated from the frontal cortex of young adult mouse, where DNA methylation was profiled using single-cell bisulfite sequencing[**Luo2018**].
Some background to motivate our experimental design: in mammalian genomes, DNA methylation predominantly occurs at CpG dinucleotides (mCG), with more than 75% of CpG sites being methylated in differentiated cell types. By contrast non-CpG methylation (mCH) has been historically dismissed as methodological artifact of incomplete bisulfite conversion, until recent works have confirmed their existence in restricted cell types. Yet, evidence for a potential functional role remains controversial[**He2015**].

Here we used MOFA+ to dissect the cellular heterogeneity associated with mCH and mCG in the mouse frontal cortex. As input data we quantified mCH and mCG levels at gene bodies, promoters and putative enhancer elements. Each combination of genomic and sequence context was defined as a separate view.
As described in Chapters 1 and 3, methylation levels were calculated per cell and genomic feature using a binomial model where the number of successes correspond to the number of reads that support methylation (or accessibility) and the number of trials the total number of reads.
Finally, to explore the influence of the neuron's location we grouped cells according to their cortical layer: Deep, Middle or Superficial (**??**). Notably, the resulting data set is extremely sparse, which hampers the use of conventional dimensionality reduction techniques.The probabilistic framework underlying MOFA+ naturally enables the handling of missing values by ignoring the corresponding terms in the likelihood function.

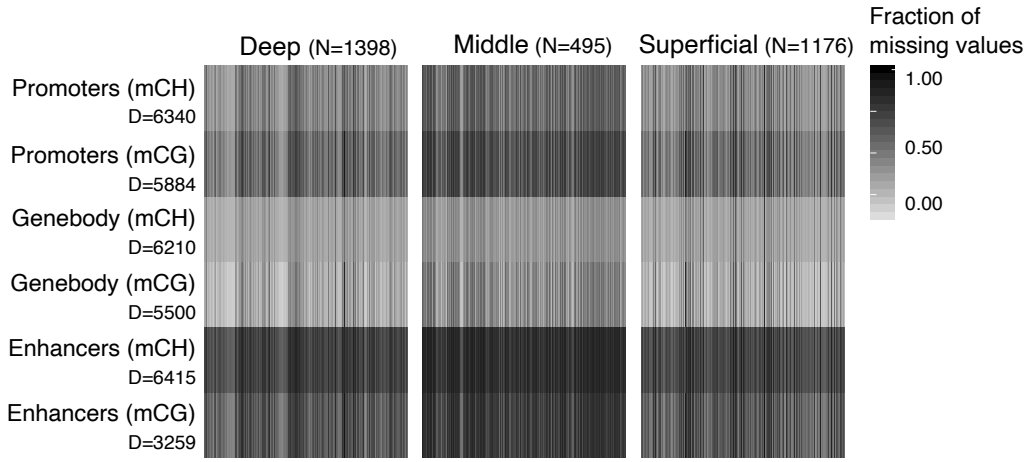


Figure 1.10

MOFA+ identifies 10 factors with a minimum variance explained of 1% in at least one data modality.

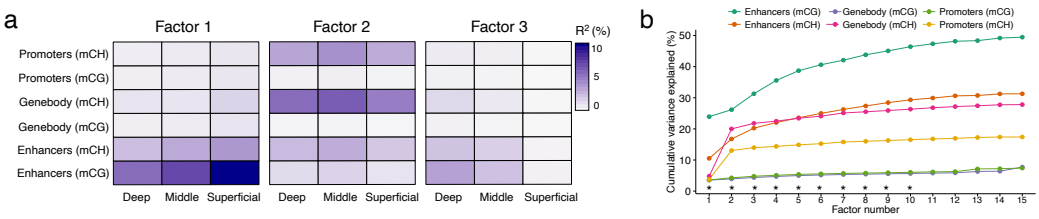


Figure 1.11

Factor 1, the major source of variation, is linked to the existence of inhibitory and excitatory neurons, the two major classes of neurons (**??**). This factor shows significant mCG activity across all cortical layers, mostly driven by coordinated changes in enhancer elements, but to some extent also gene bodies.

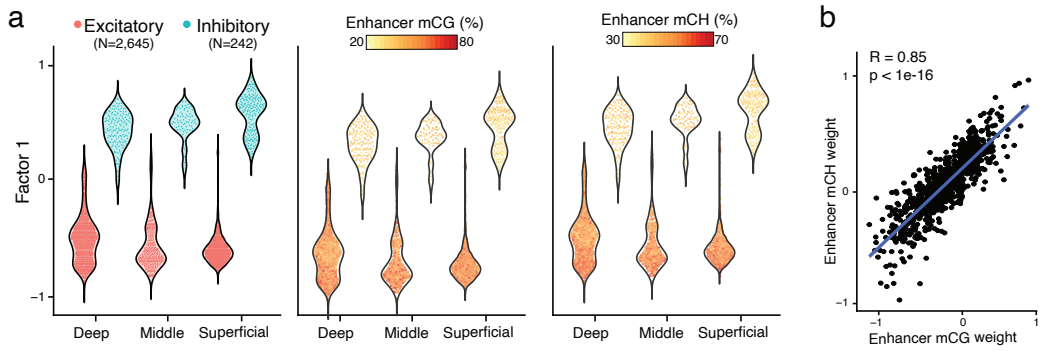


Figure 1.12

Factor 2 captures genome-wide differences in global mCH levels (R=0.99, not shown), most likely to be a technical source of variation.

Factor 3 captures heterogeneity linked to the increased cellular diversity along cortical depth, with the Deep layer displaying significantly more diversity of excitatory cell types than the Superficial layer (**??**).

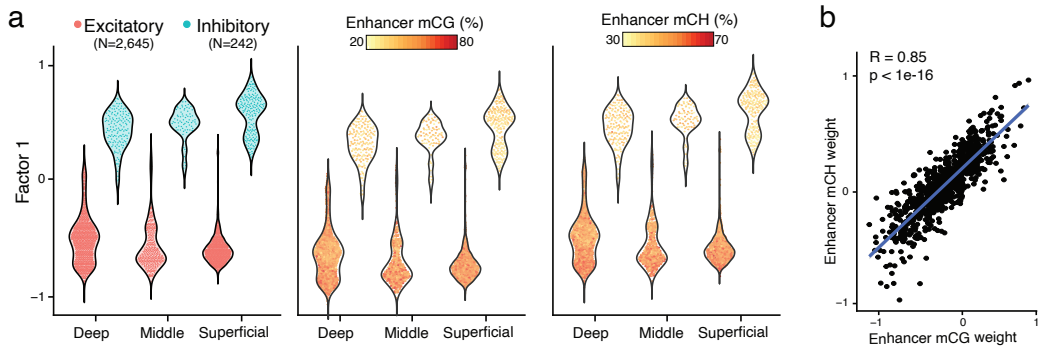


Figure 1.13

The (linear) MOFA factors can be combined by further non-linear dimensionality reduction algorithms such as UMAP or t-SNE. In this case, we show that the UMAP projections reveals the existence of multiple subpopulations of both excitatory and inhibitory cell types. Notably, the MOFA+ factors are significantly better at identifying these subpopulations than the conventional approach of using Principal Component Analysis with imputed measurements.

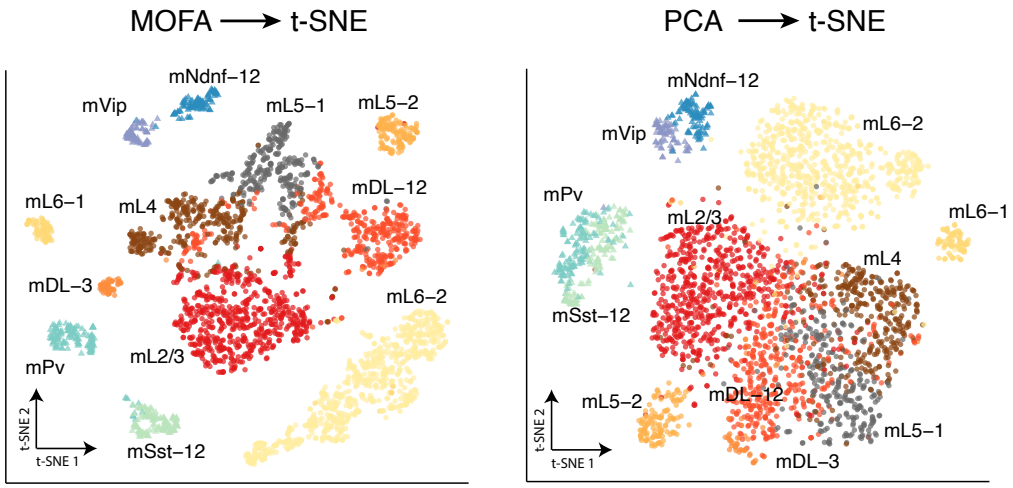

MOFA ⟶ t-SNE
mNdnf–12
mVip
mL5–1
mL5–2
mL6–1
mL4
mDL–12
mDL–3
mL6–2
mPv
mL2/3
mSst–12
t-SNE 2
t-SNE 1
PCA ⟶ t-SNE
mNdnf–12
mVip
mL6–2
mPv
mL2/3
mL6–1
mL4
mSst–12
mL5–2
mDL–12
mL5–1
mDL–3
t-SNE 2
t-SNE 1

Figure 1.14