Deep exponential families for single-cell data analysis

Pedro F. Ferreira[1,2], Jack Kuipers[1,2,*], Niko Beerenwinkel[1,2,*]

[1] *Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland*
[2] *SIB Swiss Institute of Bioinformatics, Basel, Switzerland*
[*] *Corresponding author*

Abstract

Single-cell gene expression data characterizes the complex heterogeneity of living systems. This heterogeneity is composed of various cells with diverse cell states driven by different sets of genes. Cell states can often belong to different categories according to biological hierarchies. Hierarchical clustering therefore not only improves functional interpretation, but can also be leveraged to ensure that gene signatures are less influenced by noise and batch effects. We present single-cell Deep Exponential Families (scDEF), a two-level Bayesian matrix factorization model which extracts hierarchical gene signatures from single-cell RNA-sequencing data. scDEF can additionally make use of prior information to guide the hierarchy. It can be used for dimensionality reduction, gene signature identification, and batch integration. We validate scDEF with simulations and multiple annotated real data sets, and show that scDEF recovers meaningful hierarchies in single- and multiple-batch scenarios.

1 Introduction

Single-cell RNA-sequencing (scRNA-seq) data generates snapshots of cell states in a high-throughput fashion by quantifying expression levels of all genes in the transcriptome, for thousands of individual cells [1]. These measurements permit the discovery of different cell populations within a tissue and the specific genes that are activated in each population. By aggregating scRNA-seq data from multiple donors, for example, cancer patients, these data also have the potential to reveal commonalities in cell subpopulations between different patients and to inform targeted therapy [2].

In recent years, a large number of methods to analyse scRNA-seq data have been developed, including dimensionality reduction [3], clustering [4], differential expression analysis [5], pseudo-time analysis, and trajectory inference [6, 7]. The analysis of scRNA-seq data typically involves the following steps: normalization, feature selection, dimensionality reduction, clustering, marker identification, and cluster annotation. Often, these steps rely on different computational methods or statistical models, each with their own assumptions. For example, the workflow described in [8] recommends dimensionality reduction with PCA on log-normalized read counts, followed by graph-based clustering, and differential expression between the clusters to identify the genes that drive each cluster using a Wilcoxon rank-sum test. However, this multi-step approach is prone to inconsistencies in modelling assumptions and to statistical issues, such as artificially low p-values in differential gene expression [9]. Other approaches, like scHPF [10], simultaneously perform dimensionality reduction, clustering and marker identification on the unnormalized read counts via a single probabilistic matrix factorization model, in which each factor describes a gene program. By relying on a model of the read counts, this method is able to produce gene signatures and cell groups in a single step.

However, none of these approaches explicitly models the inherent hierarchy of cell states. This hierarchical structure arises as some cell types may be further divided into subtypes (for example, regulatory or helper T cells) or into their specific functional roles (for example, cancer cells that promote angionesis or undergoing epithelial-mesenchymal transition) [11]. The hierarchy therefore reflects the specialization level at which cells perform their function, and is a natural feature to include in gene signature methods. Indeed, learning cell states at varying levels of resolution has been identified as one of the main challenges in single-cell data science [12]. Recently, nSBM [13] has been proposed to address hierarchical structure by learning different levels at which cells can be clustered. However, by working on a kNN graph instead of the gene expression matrix, the method does not provide the gene signatures that define those clusters.

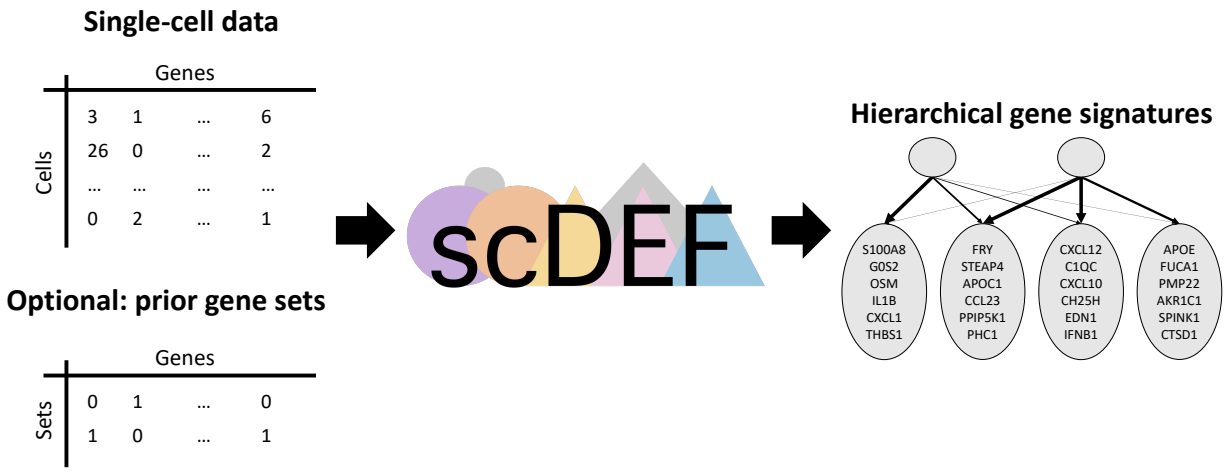


Figure 1: scDEF overview. scDEF takes as input a single-cell gene expression matrix containing the raw read counts and, optionally, prior gene sets. It then uses a stochastic variational inference algorithm to produce a set of gene signatures for the cell groups present in the sample and their hierarchical relationships. This enables an interpretable two-level clustering of cells and robust batch integration.

An additional challenge in scRNA-seq data analysis is the integration of data from different sources, such as human samples from different donors, for example [14]. Two main approaches exist to account for batch effects. One approach is late integration, in which cells from different samples are first projected onto a low-dimensional embedding and annotated separately, and then cells with common label across data sets (so-called anchor cells) are used to align those embeddings [15, 16, 17]. Alternatively, the data can be integrated early by appending data sets together and using indicator variables in a statistical model to remove the sample-specific signal. This is done, for example, in ZINB-WaVE [18] and the neural network-based approach scVI [19] and its interpretable counterpart LDVAE [20]. The main challenge in data integration is the ability to remove sample-specific technical noise while retaining the biological variability in the integration [21].

On the one hand, an unsupervised matrix factorization model for signature learning, like scHPF, has enough flexibility to learn technical batch effects. On the other hand, providing such a model with batch annotations, as in LDVAE, may lead to over-correction. Balancing these two extremes in a model that constrains the factors that it is able to learn batch effects without using annotations may help maintaining batch-specific biological variability while still being robust to technical variability. Here, we posit that using a hierarchical model of gene signatures may help in achieving this balance.

We therefore present scDEF (single-cell Deep Exponential Families), a Bayesian model designed to find the hierarchical structure in scRNA-seq data. scDEF consists of a deep exponential family model [22] tailored to single-cell data in order to cluster cells using multiple levels of abstraction, which can be mapped to different gene signature levels. By enforcing non-negativity, biasing towards sparsity and including hierarchical relationships among factors without using batch annotations, scDEF is a general tool for hierarchical gene signature identification in scRNA-seq data for both single- and multiple-batch scenarios (Figure 1).

In Section 2 we introduce the model in detail and we describe the inference procedure in Section 3. In Section 4 we apply scDEF to simulated data and to multiple extensively studied and annotated real scRNA-seq data sets. This illustrates the ability of our model to identify meaningful hierarchical representations of cell state, its robustness to batch effects without over-correction, and how it can use prior gene sets to simultaneously assign cells to known types and find cell type-specific clusters.

2 Model

scDEF models the gene expression heterogeneity of the cells of a tissue as a set of sparse factors containing gene signatures for different cell states. These factors are related to each other through higher-level factors that encode coarser relationships. scDEF extends DEFs [22] by including cell- and gene-specific scale factors similarly to scHPF [10], using automatic relevance determination (ARD) [23] to choose the number of factors.

More specifically, scDEF is a probabilistic non-negative matrix factorization model with Gamma-distributed factors in which each cell assigns weight to upper-level factors. These weights are propagated to the downstream factors through loading matrices that connect the different levels, which are also Gamma-distributed. We define the full model in Equation 4 below and describe its structure in the following paragraphs. We use the shape and rate parameterization of the Gamma distribution.

$$
\begin{aligned}
\zeta_n &\sim \text{Gamma}\left(1, a\right) \\
\eta_g &\sim \text{Gamma}\left(1, b\right) \\
\lambda_l^1 &\sim \text{Gamma}\left(10^{-2}, 10^{-2}\right) \\
w_{lk}^1 &\sim \text{Gamma}\left(0.1, \frac{1}{\lambda_l^1}\right) \\
z_{nl}^1 &\sim \text{Gamma}\left(\alpha, \frac{\alpha}{\lambda_l^1}\right) \\
\lambda_k^0 &\sim \text{Gamma}\left(10^{-2}, 10^{-2}\right) \\
w_{kg}^0 &\sim \text{Gamma}\left(0.1, \frac{\eta_g}{\lambda_k^0}\right) \\
z_{nk}^0 &\sim \text{Gamma}\left(\alpha, \frac{\alpha}{\mathbf{Z}_n^1 \mathbf{W}_k^1} \cdot \frac{\zeta_n}{\lambda_k^0}\right) \\
x_{ng} &\sim \text{Poisson}\left(\mathbf{Z}_n^0 \mathbf{W}_g^0\right)
\end{aligned}
\tag{1}
$$

We explain the model starting from the likelihood at the bottom and going upwards the generative process. The read counts of gene g at cell n are denoted x_{ng}, and we collect them in a vector $\mathbf{X}_n$ of size G genes. Each observation is drawn from a Poisson distribution with mean defined as a vector product between $\mathbf{Z}_n^0$ and $\mathbf{W}_g^0$, which is a rank K approximation to $\mathbf{X}_n$. Each cell has weight z_{nk}^0 for factor k, and factor k gives weight $w_k^0 g$ to gene g. To ensure non-negativity, z_{nk}^0 and w_{kg}^0 are drawn from Gamma distributions. This is the Poisson matrix factorization model [24], and it expresses the idea that each cell's gene expression profile is the product of multiple gene programs, each one of those containing a small set of genes.

In addition, $\mathbf{Z}_n^0$ is itself parameterized by a rank L approximation $\mathbf{Z}_n^1 \mathbf{W}_k^1$, with $L < K$. Whereas $\mathbf{W}^0$ groups the G genes into K factors, $\mathbf{W}^1$ groups the K factors into L top-level factors. Similarly, $\mathbf{Z}^1$ indicates which of the L factors best describe each cell n. The right panel in Figure 1 illustrates this: The lower K circles represent the factors, with the top 5 genes with highest weight $\mathbf{W}_k^0$ indicated inside each one. The upper L circles represent the upper-level factors, and the thickness of the edges connecting both layers represents the weights $\mathbf{W}^1$ that the upper-level factors attribute to each lower-level factor.

We include an ARD prior via λ_l^1 and λ_k^0 to capture the relative importance of the different (upper- and lower-level) factors. This enables us to ignore (upper- and lower-level) factors with small importance. In practice, this means we can run scDEF with large K and L ($L < K$) and remove the factors with small λ_l^1 and λ_k^0 after inference, instead of running many combinations of K and L and choosing the one with the best marginal likelihood. In practice, after inference we filter the (upper- and lower-level) factors by keeping only the ones with top 70% weight.

To account for cell- and gene-specific scale factors, which are quantities that may confound our analysis of gene signatures and cell states, we explicitly model them in $\boldsymbol{\zeta}$ and $\boldsymbol{\eta}$, respectively. Following [10], the hyperparameters a and b of the distributions of these random variables are set to preserve the empirical variance-to-mean ratio of the total molecules per cell or gene:

$$
a = \frac{\text{mean}(\sum_g \mathbf{X}_g)}{\text{variance}(\sum_g \mathbf{X}_g)}, \qquad b = \frac{\text{mean}(\sum_n \mathbf{X}_n)}{\text{variance}(\sum_n \mathbf{X}_n)}. \tag{2}
$$

To obtain gene signatures and the weight that each cell assigns to them, we use the size-corrected cell and gene scores, which we obtain after inference as

$$
\begin{aligned}
\text{CellScore}_{nk}^0 &= \mathbb{E}[z_{nk}^0] * \mathbb{E}[\zeta_n], \\
\text{CellScore}_{nl}^1 &= \mathbb{E}[z_{nl}^1], \\
\text{CellScore}_{kg} &= \mathbb{E}[w_{kg}^0] * \mathbb{E}[\eta_g].
\end{aligned}
\tag{3}
$$

scDEF can be seen as an extension of scHPF with an upper layer and ARD. In other words, scDEF reduces to scHPF by removing the upper layer, $\mathbf{Z}^1$ and $\mathbf{W}^1$, and the ARD priors, $\boldsymbol{\lambda}$.

2.1 Using prior information

scDEF is presented in Equation 4 in its most general form, where we do not make strong assumptions about the number of (upper- and lower-level) factors nor what type of gene programs they encode. However, we may wish to push the factorization towards some structure that we expect to find in the data. We can do this by removing the ARD prior, adding a connectivity matrix $\boldsymbol{\phi}$ defining which factors should be hierarchically related (i.e., influencing the prior distribution on the matrix $\mathbf{W}^1$) and a gene program matrix $\boldsymbol{\omega}$ indicating which genes should tend to be activated in each factor (i.e., in the matrix $\mathbf{W}^0$). This model is then defined as

$$\begin{aligned} \zeta_n &\sim \text{Gamma}\left(1, a\right) & w^0_{kg} &\sim \text{Gamma}\left(\omega_{kg}, \eta_g\right) \\ \eta_g &\sim \text{Gamma}\left(1, b\right) & z^0_{nk} &\sim \text{Gamma}\left(\alpha, \frac{\alpha}{\mathbf{Z}^1_n \mathbf{W}^1_k}\right) \\ w^1_{lk} &\sim \text{Gamma}\left(\phi_{lk}, 1\right) & x_{ng} &\sim \text{Poisson}\left(\mathbf{Z}^0_n \mathbf{W}^0_g\right). \\ z^1_{nl} &\sim \text{Gamma}\left(\alpha, \alpha \cdot \zeta_n\right) \end{aligned} \tag{4}$$

3 Inference

Given read count data $\mathbf{X}$, we wish to find the scDEF model defined in Equation 4, that best describes the data. We pursue a fully Bayesian approach and attempt to find the posterior distribution over all variables in the model, $P(\mathbf{Z}^0, \mathbf{Z}^1, \mathbf{W}^0, \mathbf{W}^1, \boldsymbol{\lambda}^0, \boldsymbol{\lambda}^1, \boldsymbol{\zeta}, \boldsymbol{\eta}|\mathbf{X})$. Since the model is non-conjugate, this distribution is not analytically tractable. In this case, the two main approaches to find the posterior distribution are sampling methods, such as Markov chain Monte Carlo (MCMC) [25], and variational inference [26]. Here we use variational methods, because they are often able to obtain an approximation faster than MCMC [27, 28].

In variational inference, we posit a simpler family of distributions over the model's latent variables, $Q(\mathbf{Z}^0, \mathbf{Z}^1, \mathbf{W}^0, \mathbf{W}^1, \boldsymbol{\lambda}^0, \boldsymbol{\lambda}^1, \boldsymbol{\zeta}, \boldsymbol{\eta})$, whose parameters are determined by minimizing the Kullback-Leibler divergence between $Q(\mathbf{Z}^0, \mathbf{Z}^1, \mathbf{W}^0, \mathbf{W}^1, \boldsymbol{\lambda}^0, \boldsymbol{\lambda}^1, \boldsymbol{\zeta}, \boldsymbol{\eta})$ and the true posterior $P(\mathbf{Z}^0, \mathbf{Z}^1, \mathbf{W}^0, \mathbf{W}^1, \boldsymbol{\lambda}^0, \boldsymbol{\lambda}^1, \boldsymbol{\zeta}, \boldsymbol{\eta}|\mathbf{X})$ or, equivalently, to maximize the evidence lower bound (ELBO),

$$Q^* = \arg\max_Q \text{ELBO}(Q), \tag{5}$$

with

$$\text{ELBO}(Q) = \mathbb{E}_Q\left[\log P(\mathbf{X}, \mathbf{Z}^0, \mathbf{Z}^1, \mathbf{W}^0, \mathbf{W}^1, \boldsymbol{\lambda}^0, \boldsymbol{\lambda}^1, \boldsymbol{\zeta}, \boldsymbol{\eta})\right] - \mathbb{E}_Q\left[\log Q(\mathbf{Z}^0, \mathbf{Z}^1, \mathbf{W}^0, \mathbf{W}^1, \boldsymbol{\lambda}^0, \boldsymbol{\lambda}^1, \boldsymbol{\zeta}, \boldsymbol{\eta})\right]. \tag{6}$$

We use a mean-field approximation for Q in which the posterior distribution over the latent variables factorizes completely, and we let each marginal variational distribution be in the same family as the prior of the corresponding latent variable (which in our model means that they are all approximated with Gamma distributions), with its own variational parameters:

$$\begin{aligned} Q(\mathbf{Z}^0, \mathbf{Z}^1, \mathbf{W}^0, \mathbf{W}^1, \boldsymbol{\lambda}^0, \boldsymbol{\lambda}^1, \boldsymbol{\zeta}, \boldsymbol{\eta}) = \prod_{n,g,k,l} & Q_{\zeta_n}(\zeta_n) Q_{\eta_g}(\eta_g) Q_{\lambda^0_k}(\lambda^0_k) Q_{\lambda^1_l}(\lambda^1_l) Q_{z^0_{nk}}(z^0_{nk}) \\ & \cdot Q_{z^1_{nl}}(z^1_{nl}) Q_{w^0_{kg}}(w^0_{kg}) Q_{w^1_{lk}}(w^1_{lk}). \end{aligned} \tag{7}$$

This approach is also used in [10]. However, since our model is not even conditionally conjugate due to the upper layer, we can not compute the ELBO analytically and may not use analytical updates for the variational parameters. We thus resort to black-box variational inference [29, 30], in which the variational parameters are updated using Monte Carlo approximations of their gradients with respect to the ELBO. Collapsing all the latent variables into $\boldsymbol{\theta}$ for readability, we then approximate the ELBO as

$$\text{ELBO}(Q) \approx \frac{1}{S} \sum_{s=1}^{S} \log P(\mathbf{X}, \boldsymbol{\theta}^s) - \log Q(\boldsymbol{\theta}^s), \qquad \boldsymbol{\theta}^s \sim Q(\boldsymbol{\theta}) \tag{8}$$

where S denotes the number of Monte Carlo samples from the variational distribution. We employ automatic differentiation with JAX [31] to compute the gradient of all the variational parameters with respect to the ELBO, and use the widely adopted gradient descent method Adam [32]. In practice, we run scDEF with $S = 3$, which enables a good

trade-off between speed and low gradient variance. To ensure scalability to data sets with large numbers of cells, we also make use of minibatch optimization, in which we sample, at each iteration, a subset of cells and optimize only their local variational parameters, and scale the gradient with respect to the global variational parameters accordingly. By default, we use a minibatch size of 256 cells.

4 Results

To assess the ability of scDEF to learn hierarchical gene signatures of cell state in single and multi-batch scenarios, we applied it to simulated data and multiple public scRNA-seq datasets.

4.1 Simulation study

We assess the ability of scDEF to associate its factors with the ground truth cell populations in single and multi-batch simulation settings, using Splatter as the scRNA-seq data simulation framework [33]. For each simulation, we generate eight cell groups (representing different types or states) with approximately the same number of cells each, to a total of 2000 cells with expression profiles over 800 genes, and repeat each setting ten times. To ensure the simulations are realistic, we estimate the global Splatter parameters using a public data set from 10x Genomics containing 3,000 peripheral blood mononuclear cells from a healthy donor [34].

We compare our method, scDEF, with two competing methods: scHPF, which is a single-layer version of scDEF without ARD, and which also does not use batch annotations, and LDVAE, which is a variational autoencoder with a linear decoder and relies on batch annotations for integration. Each method aims to provide interpretable factors that are associated with the underlying cell populations. scDEF automatically selects the number of factors. For both scHPF and LDVAE, we vary the number of factors from 5 to 15, and choose the model that yields the highest ELBO. To compare the methods, we define a metric that quantifies the extent to which each cell group is associated with mostly one factor, which indicates that the method learned interpretable factors. We compute the mean weight that the cells from each ground truth cell group attribute to each factor, leading to a matrix of factors by cell groups. We normalize this matrix row-wise and compute the average entropy across the rows, and then do the same procedure column-wise. If each factor in the model is strongly associated with a distinct cell group, the row- and column-wise entropies will be very low. We report the mean of the two quantities and use this as a metric to compare the three methods. As LDVAE produces both positive- and negative-valued factor weights, we take their absolute values before normalizing.

In the single-batch scenario (Figure 2 A), we vary the probability of genes being differentially expressed in each group (Splatter parameter "de.prob") from 0.3 to 0.7, where 0.3 leads to highly mixed cell groups and 0.7 leads to high separability. The results in this setting show that scDEF consistently outperforms scHPF and LDVAE. We observed that scHPF tended to underestimate the number of factors in the model, leading to cell groups being explained by the same factor, which explains its high entropy. While LDVAE was able to detect all populations, its use of both positive- and negative-valued quantities leads to dense and weakly interpretable factors. scDEF obtained the lowest entropy among all methods. By automatically detecting the appropriate number of factors, using non-negativity for interpretability, and relying on a hierarchical factorization for robustness to noise, scDEF obtained factors which strongly associated with the ground truth cell groups.

We also applied scDEF, scHPF and LDVAE to settings with two and four batches (Figure 2 B). In each setting, we vary the amount of simulated biological cell groups that are shared between batches: one case where each batch has its own set of populations, another where half of all the cell groups are shared and the rest are randomly chosen to belong to different batches, and another where all batches share all populations (corresponding to fractions of shared groups of 0, 0.5 and 1 in Figure 2 B). In both the two- and four-batches settings, scDEF outperforms scHPF and LDVAE when each batch has their own set of populations. In the two-batches setting, LDVAE outperforms scHPF and scDEF when all populations are present in both batches, and performs similarly to scDEF when half of the populations are shared. In the four batches setting, LDVAE again performs better than scHPF and scDEF when there are shared populations between batches. These results illustrate one of the core features of scDEF: by not using explicit batch annotations and relying on a hierarchical factorization, our method is better able to detect the relevant cell groups while being robust to technical batch effects when there is significant biological heterogeneity between batches. Nonetheless, using batch annotations may be a good option when there a significant amount of shared populations between batches.

4.2 scDEF reveals interpretable groups of gene signatures

We applied scDEF to the publicly available 3k PBMCs data set from 10x Genomics, which contains 2700 peripheral blood mononuclear cells from a healthy donor [34]. This data set has been generated and made publicly available by 10x Genomics as a high-quality example of their droplet-based single-cell RNA-sequencing protocol. It is often

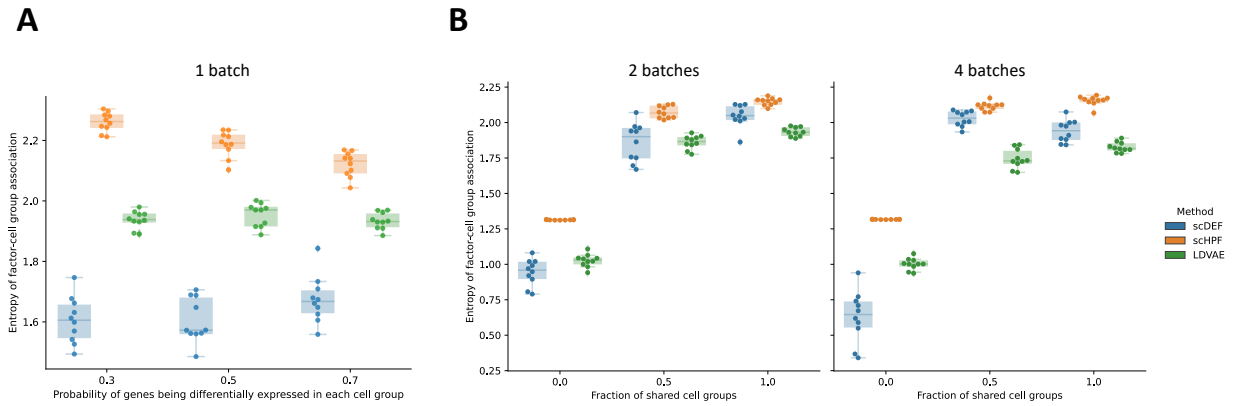


Figure 2: Simulation study. We generated noisy gene expression data for 2000 cells and 800 genes containing 8 cell groups, and compared the ability of our method scDEF (blue) and two alternative methods, scHPF (orange) and LDVAE (green) to recover the ground truth groups in single- and multi-batch scenarios. As a measure of comparison, we computed the entropy in the association between each model's latent factors, and the ground truth cell groups (lower is better). (A) We varied the distance between the cell groups in the single-batch case using Splatter's "de.prob" parameter, leading to low, medium and high group separability. (B) For simulated data with 2 and 4 batches, we varied the fraction of cell groups which are shared between the different batches, reflecting different levels of batch-specific biological heterogeneity (where 0 corresponds to no cell groups shared across batches and 1 to all cell groups being present in all batches).

used for benchmarking scRNA-seq workflows as it contains well-defined cell types for which marker genes have been derived [16, 35].

We downloaded the processed data using Seurat [16], which contains the cell type annotations obtained from the canonical workflow of dimensionality reduction, clustering, and differential gene expression analysis, as well as a UMAP [36] embedding of the data (Figure 3 A).

We ran scDEF in the standard mode (no prior gene sets and using ARD) with $K = 20$ and $L = 5$. After removing the least explanatory factors according to their relevance (Supplementary Figure S1 A, B), we obtained 8 factors related through 3 hierarchical factors. By assigning each cell to the (hierarchical) factor to which it attributes the most cell size-corrected weight (Section 2), we can inspect the correspondence between cell types and the scDEF factors (Figure 3 B, C). We found that the upper-level hierarchical factor 0 captured dentritic cells (DC) and monocytic cell populations, which are in turn captured by lower-level factors 4 (DCs), 0 (FCGCR3+ Mono) and 6 (CD14+ Mono) at the lower-level. By indicating the top 10 strongest genes in each factor (Figure 3 D), we can directly see which genes drive each population. Indeed, canonical marker genes are identified for most cell types, such as MS4A1 and CD79 for B cells (factor 5), CD2 for T cells (factor 3), and GNLY for Natural Killer cells (factor 1).

To obtain a more complete view of the associations between cell types and scDEF factors, we computed the mean score that cells of each type attributed to each factor (Figure 4 A) and hierarchical factor (Figure 4 B). Indeed, each cell type had an associated dominant factor, highlighting the ability of scDEF's factors to detect cell populations and their signatures. The upper layer of scDEF shows a higher-level description of the cell types, where Monocytes are grouped together in hierarchical factor 0, and Memory and Naive T cells associate with hierarchical factor 2. We additionally show the mean weight that cells of each type attributed to each latent dimension of LDVAE (Figure 4 C), which due to its lack of non-negativity and sparsity constraints does not show a strong cell type-factor association, in contrast to the high interpretability of scDEF.

scDEF infers cell- and gene-specific scale factors to account for library size and gene detection rate variation, and uses them to obtain normalized cell and gene weights in the matrix factorization model. The inferred cell and gene scale factors strongly correlate with their empirical values (Supplementary Figure S1 C, D), indicating that this nuisance variation is not captured by the factorization weights.

4.3 scDEF performs robust batch integration

One major advantage of scDEF over alternative batch integration methods is the fact that it does not explicitly aim to remove the batch effects, which avoids over-correcting and removing batch-specific biological variation. Instead,

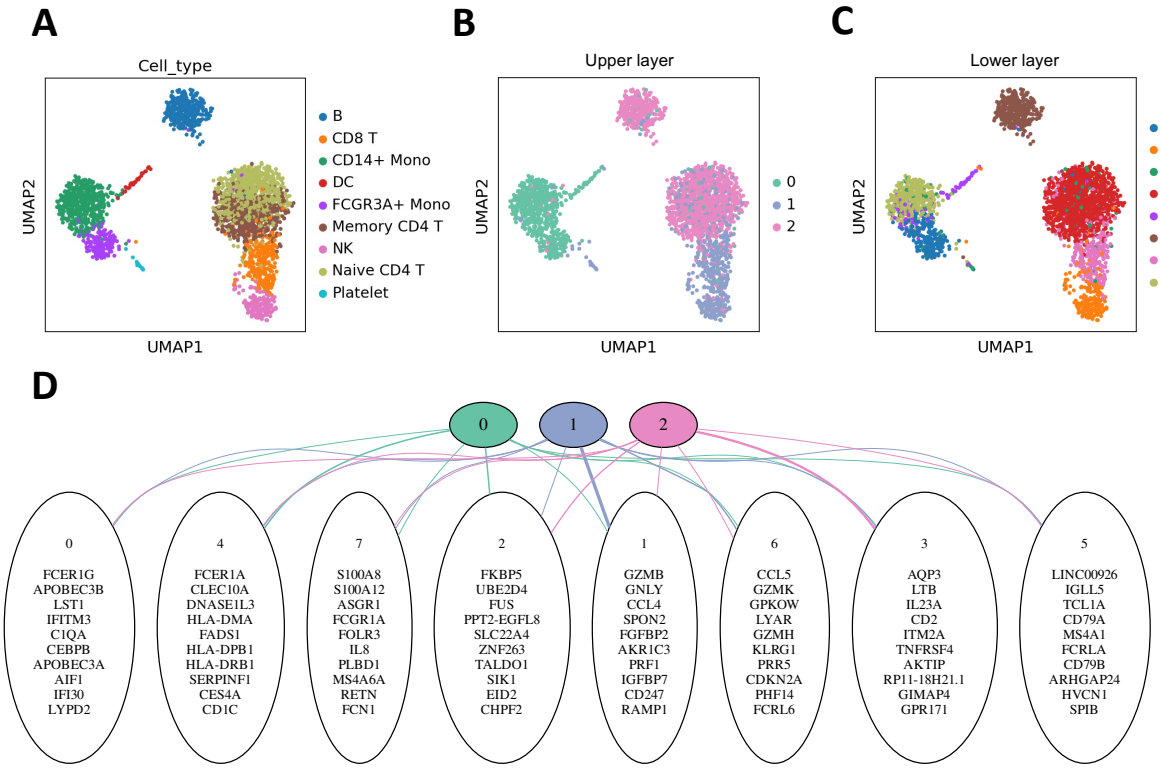


Figure 3: Application to the 3k PBMCs dataset from 10x Genomics. (A) UMAP embedding of cells after standard processing colored by known cell type. (B) UMAP embedding of cells colored by their assignment to scDEF's upper-level factors and to (C) eight factors in the lower layer. (D) Network representation of the learned scDEF. The upper colored circles denote the three upper-level factors, and the lower circles denote the eight lower-level factors. We show the top ten genes with the most weight in each lower-level factor.

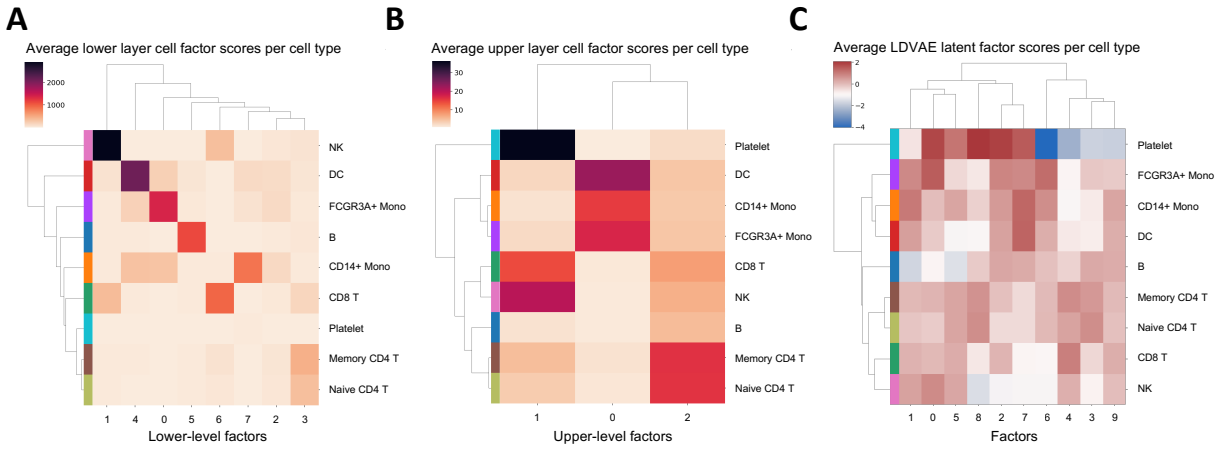


Figure 4: Association between factors and cell types in the 3k PBMCs dataset. (A) Mean cell score per cell type attributed to factors in scDEF's lower layer, showing a clear association between cell types and factors. (B) Mean cell score per cell type attributed to hierarchical factors in scDEF's upper layer. (C) Mean weight that cells of each type attributed to each latent dimension of LDVAE, which does not have non-negativity constraints.

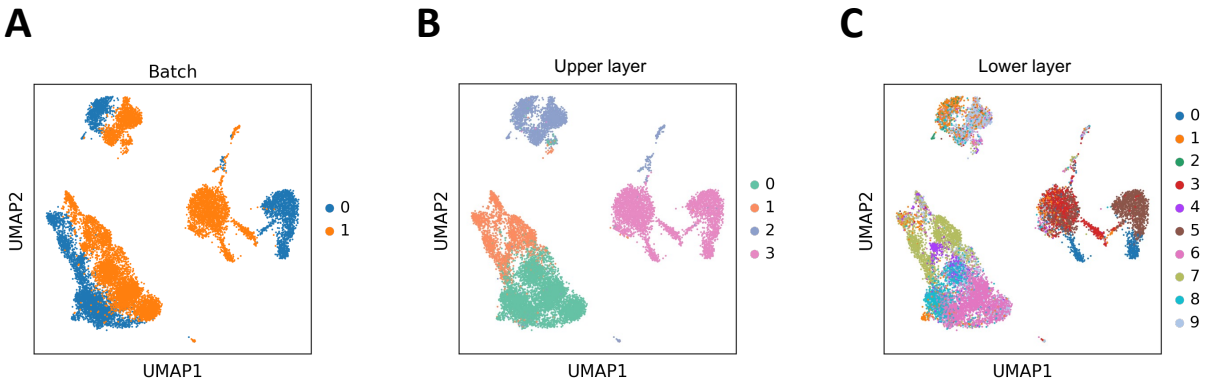


Figure 5: Application to 6k and 8k PBMCs from two healthy donors. (A) UMAP embedding of cells colored by their batch of origin, showing a clear separation between the two samples. (B) UMAP embedding of cells colored by assignment to scDEF's four hierarchical upper-level factors, where batch effects are not captured, and to (C) ten factors in the lower layer.

scDEF relies on (1) sparse factors, which tend to capture true biology (as opposed to dense factors [37]), and (2) non-independent factors, which encourages them to share information, instead of capturing non-shared noise.

4.3.1 Two batches, one condition

We first apply scDEF to 8k PBMCs [38] and 6k PBMCs [39] from 10x Genomics containing healthy peripheral mononuclear blood cells from two different donors. As both samples come from healthy donors, we expect nearly all batch-specific variation to be technical. Indeed, jointly performing a UMAP projection of all cells from both data sets reveals very strong batch effects (Figure 5 A).

We then ran scDEF and obtained cell-to-factor assignments, which is the matrix factorization counterpart to unsupervised cell clustering. The upper-level hierarchical factors show little to no batch-effects: cells from different batches are all assigned to the same factor, which correlates well with the main groups of cells visible in the UMAP space (Figure 5 B). The upper layer of scDEF therefore robustly captures the main cell populations irrespective of technical batch effects. Inspecting the lower layer factors reveals that while the factors are still largely batch-independent, there are a few instances in which they are not (for example, lower-level factors 1 and 9 are batch-specific), highlighting the fact that some factors may still be driven by noise in scDEF (Figure 5 C).

4.3.2 Two batches, two conditions

When the two batches come from different conditions, common batch integration methods will fail to resolve their differences, which may be biologically relevant. Here we apply scDEF to a dataset consisting of PBMCs from two lupus patients, one of which was treated with interferon, to examine cell type-specific responses to the stimulation. The data was generated in [40] using a droplet-based protocol and we downloaded it from the Seurat package. We expect that not all cell types react to the stimulus, which means that scDEF should be able to identify a factor indicating the genes that become active upon this stimulus and the cells do exhibit a response.

As in the previous example, a batch-uncorrected UMAP embedding indicates that batch effects are the dominating structure in the data (Figure 6 A, B). However, when running scDEF on this dataset and assigning cells to upper-level factors, we observed that while upper-level factors 1 and 3 correct for batch effects in B and T cells, upper-level factors 0 and 2 are still mostly driven by stimulus response, and contain mostly monocytes (Figure 6 C). We inspected the top genes in lower-level factor 4, which was the one with the strongest association (in terms of $\mathbf{W}^1$ weight) to upper-level factor 2, which appears to capture the stimulus response. We found that CXCL10 was among these genes. This is in accordance with the original analysis in [40], which concluded that there was a monocyte-specific response to the interferon stimulus.

To visualize the association between hierarchical factors to cell types in a non-discretized manner, we took the upper layer of the scDEF model and applied a UMAP projection to visualize it in two dimensions (Supplementary Figure S2). This representation highlights the meaning of the hierarchical structure found by scDEF: the major cell types are separated in this space, but not their subtypes (Supplementary Figure S2 A). Additionally, all cell types are well

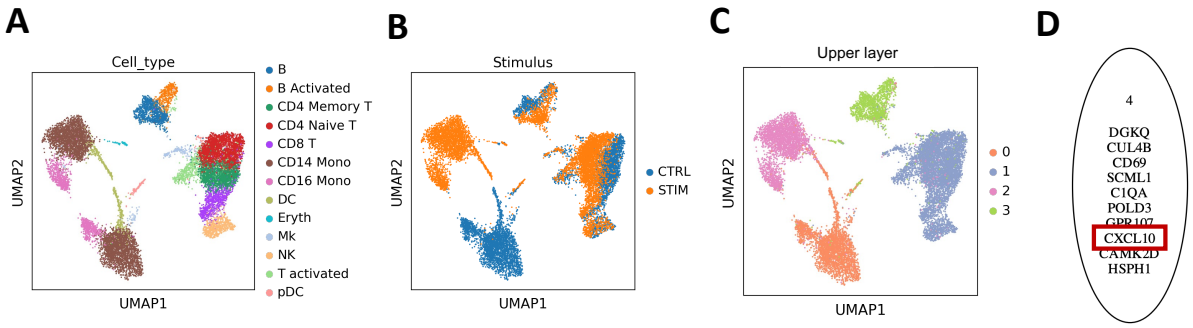


Figure 6: Application to normal and interferon-stimulated PBMCs. (A) UMAP projection of the complete dataset annotated by cell type, (B) stimulus and (C) assignment of cells to scDEF's hierarchical factors, where factors 0 and 2 capture stimulus state, unlike factors 1 and 3. (D) Top scoring genes in factor 4 in scDEF's lower layer, which was strongly associated with hierarchical factor 2 and captures stimulus response in monocytes.

mixed between the two conditions, except for monocytes (Supplementary Figure S2 B). As indicated previously, one of the genes driving the stimulus response in these cells is CXCL10, which this scDEF-based projection confirms (Supplementary Figure S2 C).

4.4 scDEF can be guided by prior hierarchies and gene sets

Finally, we show that scDEF can be combined with prior information to simultaneously assign cells to identities with known marker genes, and find a clustering structure within those identities. This is a complementary approach to the standard uninformed scDEF and may be useful when users know which major cell groups to expect in the data, instead of aiming to identify them *de novo*. To illustrate this feature, we use scRNA-seq data collected from both ovaries of a high-grade serous ovarian cancer patient [41]. These data have been previously annotated with eight major epithelial, stromal, and immune cell types (Figure S3 A, B).

We first applied scDEF in the uninformed setting, using $K = 30$, $L = 10$, and without any prior gene sets. This analysis revealed 6 hierarchical factors that did not follow the underlying cell type structure closely. For example, hierarchical factor 1 captured ovarian stromal cells, myofibroblasts, and vascular smooth muscle cells (Figure S3 C, D). The major structure identified by scDEF's upper layer was within epithelial cells from both ovaries, which were distributed across hierarchical factors 0, 3, and 4, which indicates both the presence of gene expression heterogeneity between and within epithelial cells in each ovary.

We then ran scDEF with prior gene sets corresponding to the cell types that we expect to find in a high-grade serous ovarian cancer sample (Figure 7), akin to the original approach in [41]. We used the same prior gene sets as in [41], where a list of expected marker genes for each major cell type is provided, without any information on their relative importance for each cell type. By using these gene sets in scDEF, we push the hierarchical factors towards the expected cell types, and increase the likelihood of finding cell type-specific clusters in scDEF's bottom layer, instead of letting them drive the structure in the upper layer. Indeed, the upper-level factors strongly associate with the cell type annotations (Figure 7 A, B), and detect the monocytes and the T, B, and endothelial cell population, which the uninformed scDEF failed to detect. This finding highlights the ability of the informed scDEF to obtain accurate assignments to known cell groups.

The lower layer directly provided cell type-specific structure that we can use to assess intra-cell type heterogeneity. scDEF strongly associated more than one factor in the lower layer with epithelial cells, and at most one factor for all the other cell types (Figure 7 C). We inspected the scDEF signatures associated with epithelial cells and found over-expression of glycolysis-associated gene ELF3 in the signature for epithelial-specific low-level factor 7, which consisted of cells in the right ovary (Figure 7. This is in concordance with the findings in [41], where unsupervised clustering within the epithelial cells revealed a right-ovary-specific cluster of cells with activated glycosis pathways.

This case study shows that the informed version of scDEF may serve as a powerful alternative to the usual analysis pipeline, in which users would first assign cells to known types and then perform unsupervised clustering. This was the approach followed in [41], and here we obtained consistent results using our approach of jointly assigning cells to their cell types and finding clusters within them. Additionally, the sensitivity of scDEF to cell-type specific substructure can be controlled via its hyperparameters (Section 2).

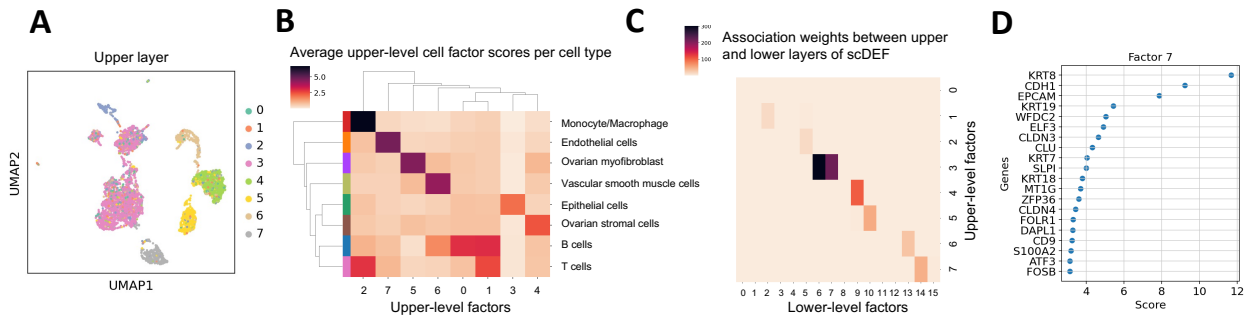


Figure 7: Application to a high-grade serous ovarian cancer sample with informed scDEF. (A) UMAP projection coloured by assignment of cells to marker-gene-informed hierarchical factors in scDEF's upper layer. (B) Mean score that cell types attribute to each of scDEF's upper-level factors, showing that the upper layer factors associate strongly with distinct cell types. In particular, epithelial cells are captured by upper-level factor 3. (C) The matrix indicating the weights between top and lower layers of scDEF shows that many lower layer factors are not associated with any upper-level factor, and that upper-level factor 3 contains two lower-level factors, 6 and 7. (D) Top 20 genes in the gene signature of factor 7 from informed scDEF's lower layer, which includes glycosis-associated genes.

5 Discussion

Identifying cell populations and their gene signatures is one of the key tasks in single-cell data analysis. Conventional approaches rely on a two-step procedure in which cells are clustered and differential expression tests are carried out between the clusters. These approaches have three main drawbacks: they lead to artificially low p-values, they ignore the fact that cell states are often arranged in a hierarchy, and they require extra normalization steps to perform this analysis on multiple batches jointly. Here we have followed a different approach based on a two-level probabilistic matrix factorization model which addresses both tasks jointly. The two-level factorization ensures not only that we are able to learn hierarchical and sparse gene signatures, but also that those gene signatures are more robust to noise than a single-level counterpart. Our approach is thus naturally more robust to batch effects without over-correcting.

The results show that scDEF is able to obtain representations of the data which are largely unaffected by technical batch effects, especially in the upper layer. Nevertheless, the lower layer may still show such effects. scDEF aims to balance the ability to capture all the relevant variation in the data by not using any batch annotations while still being constrained enough to avoid gene signatures being dominated by technical variation. Maintaining this balance fills a gap among batch integration methods, which use batch annotations in their integration with the risk of over-correction [42].

We have defined scDEF using a sparse gamma deep exponential family with cell- and gene-specific size factors and an ARD prior for automatic model selection. While the architecture of this model is akin to a two-layer neural network, by treating it as a generative model and performing approximate Bayesian inference we are able to include interpretable constraints such as non-negativity and sparsity in a probabilistic framework. This approach has advantages over models such as LDVAE [20], which is a variational autoencoder-based model without non-negativity and sparsity constraints, which leads to denser and less interpretable gene signatures, as well as not promoting a clustering of the cells. However, because we infer the posterior of all variables in our model, we are unable to make use of amortized inference techniques such as the ones employed by LDVAE. This makes scDEF generally slower than these methods.

Learning hierarchical representations of cell states in multiple tissues at scale is a particularly important task in cancer studies, where there is large cell-to-cell heterogeneity, and defining it in a hierarchical manner may help to interpret it. Tumor heterogeneity may be described as a hierarchy in which populations are defined by their cell type at one level and by phenotypic states at a more specific level [11]. Additionally, in analyses where multiple patients of a cancer are pooled together, the two-level factorization provided by scDEF provides a principled method to analyse gene signatures of cell populations that are specific or shared between patients, which may help in better defining treatments.

In the age of widespread single-cell data studies to discover, track, compare, and characterize cell populations, scDEF will be an important tool in gene signature learning robustly across data sets.

Code availability

scDEF is freely available as a Python package in `https://github.com/cbg-ethz/scDEF`.

References

[1] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):1–12, 2017.

[2] Paula Nieto, Marc Elosua-Bayes, Juan L Trincado, Domenica Marchese, Ramon Massoni-Badosa, Maria Salvany, Ana Henriques, Juan Nieto, Sergio Aguilar-Fernández, Elisabetta Mereu, et al. A single-cell tumor immune atlas for precision oncology. *Genome Research*, 31(10):1913–1926, 2021.

[3] Shiquan Sun, Jiaqiang Zhu, Ying Ma, and Xiang Zhou. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biology*, 20(1):1–21, 2019.

[4] Lijia Yu, Yue Cao, Jean YH Yang, and Pengyi Yang. Benchmarking clustering algorithms on estimating the number of cell types from single-cell RNA-sequencing data. *Genome Biology*, 23(1):1–21, 2022.

[5] Tianyu Wang, Boyang Li, Craig E Nelson, and Sheida Nabavi. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC bioinformatics*, 20(1):1–16, 2019.

[6] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. *Nature Biotechnology*, 32(4):381, 2014.

[7] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, et al. RNA velocity of single cells. *Nature*, 560(7719):494–498, 2018.

[8] Malte D Luecken and Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, 2019.

[9] Jesse M Zhang, Govinda M Kamath, and N Tse David. Valid post-clustering differential analysis for single-cell RNA-seq. *Cell systems*, 9(4):383–392, 2019.

[10] Hanna Mendes Levitin, Jinzhou Yuan, Yim Ling Cheng, Francisco JR Ruiz, Erin C Bush, Jeffrey N Bruce, Peter Canoll, Antonio Iavarone, Anna Lasorella, David M Blei, et al. De novo gene signature identification from single-cell RNA-seq with hierarchical poisson factorization. *Molecular Systems Biology*, 15(2):e8557, 2019.

[11] Eric A Smith and H Courtney Hodges. The spatial and genomic hierarchy of tumor ecosystems revealed by single-cell technologies. *Trends in cancer*, 5(7):411–425, 2019.

[12] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome Biology*, 21(1):1–35, 2020.

[13] Leonardo Morelli, Valentina Giansanti, and Davide Cittaro. Nested stochastic block models applied to the analysis of single cell data. *BMC bioinformatics*, 22(1):1–19, 2021.

[14] Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, 2022.

[15] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, 2019.

[16] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

[17] Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, Adam Gayoso, Nir Yosef, Marta Interlandi, et al. Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, 40(1):121–130, 2022.

[18] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nature Communications*, 9(1):1–17, 2018.

[19] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

[20] Valentine Svensson, Adam Gayoso, Nir Yosef, and Lior Pachter. Interpretable factor models of single-cell RNA-seq via variational autoencoders. *Bioinformatics*, 36(11):3418–3421, 2020.

[21] Jean Fan, Kamil Slowikowski, and Fan Zhang. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Experimental & Molecular Medicine*, 52(9):1452–1465, 2020.

[22] Rajesh Ranganath, Linpeng Tang, Laurent Charlin, and David Blei. Deep exponential families. In *Artificial Intelligence and Statistics*, pages 762–771. PMLR, 2015.

[23] Vincent YF Tan and Cédric Févotte. Automatic relevance determination in nonnegative matrix factorization with the beta-divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1592–1605, 2012.

[24] Prem Gopalan, Jake M Hofman, and David M Blei. Scalable recommendation with poisson factorization. *arXiv preprint arXiv:1311.1704*, 2013.

[25] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of markov chain monte carlo*. CRC press, 2011.

[26] Martin J Wainwright and Michael Irwin Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

[27] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[28] Mathieu Fourment, Andrew F Magee, Chris Whidden, Arman Bilge, Frederick A Matsen IV, and Vladimir N Minin. 19 dubious ways to compute the marginal likelihood of a phylogenetic tree topology. *Systematic biology*, 69(2):209–220, 2020.

[29] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR, 2014.

[30] Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *Journal of machine learning research*, 2017.

[31] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[33] Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell RNA sequencing data. *Genome Biology*, 18(1):1–15, 2017.

[34] 10x Genomics. 3k PBMCs, single cell gene expression dataset by Cell Ranger 1.1.0. May 2016.

[35] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):1–5, 2018.

[36] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[37] Florian Buettner, Naruemon Pratanwanich, Davis J McCarthy, John C Marioni, and Oliver Stegle. f-sclvm: scalable and versatile factor analysis for single-cell RNA-seq. *Genome Biology*, 18(1):1–13, 2017.

[38] 10x Genomics. 8k PBMCs, single cell gene expression dataset by Cell Ranger 2.1.0. November 2017.

[39] 10x Genomics. 6k PBMCs, single cell gene expression dataset by Cell Ranger 1.1.0. July 2016.

[40] Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, 36(1):89–94, 2018.

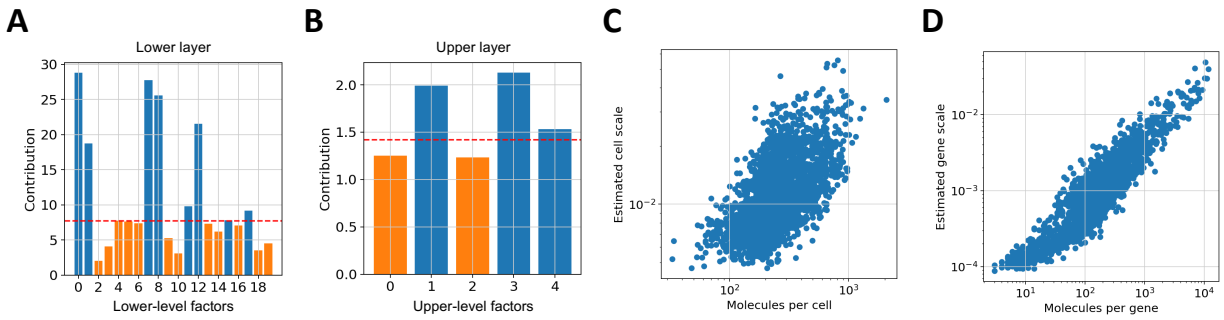
[41] Allen W Zhang, Ciara O'Flanagan, Elizabeth A Chavez, Jamie LP Lim, Nicholas Ceglia, Andrew McPherson, Matt Wiens, Pascale Walters, Tim Chan, Brittany Hewitson, et al. Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling. *Nature Methods*, 16(10):1007–1015, 2019.

[42] Hoa Thi Nhu Tran, Kok Siong Ang, Marion Chevrier, Xiaomeng Zhang, Nicole Yee Shin Lee, Michelle Goh, and Jinmiao Chen. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology*, 21(1):1–32, 2020.

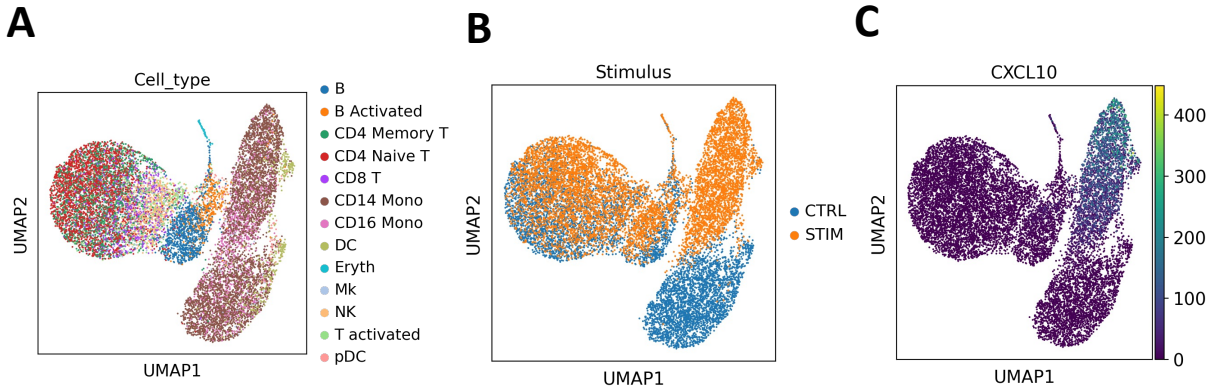
Deep exponential families for single-cell data analysis

Pedro F. Ferreira[1,2], Jack Kuipers[1,2,*], Niko Beerenwinkel[1,2,*]

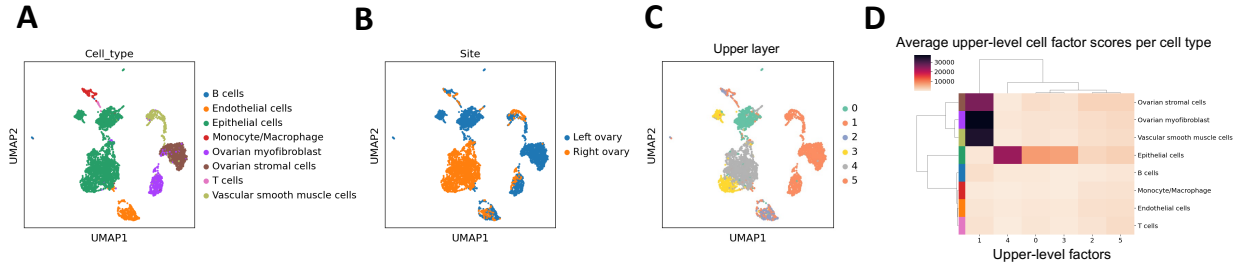
Supplementary figures



Supplementary Figure S1: scDEF diagnostics and ARD prior on the 3k PBMCs dataset. (A) scDEF's cell size factors versus empirical total number of molecules per cell and (B) gene size factors versus empirical total number of molecules per gene. (C) Contribution of five hierarchical factors to scDEF from the automatic relevance determination prior. The blue factors are kept for downstream analysis, the orange ones removed, and the red dashed line denotes the cut-off threshold. (D) Contribution of the bottom factors from scDEF.



Supplementary Figure S2: (A) UMAP projection of normal and interferon-stimulated data based on scDEF's top layer colored by cell type, showing cells clustering into the main cell types (e.g. T cells), but not their subtypes (CD4 Memory T, CD4 Naive T, CD8 T), highlighting the coarse level of representation enabled by scDEF's hierarchical factors. (B) scDEF's UMAP projection with cells colored by stimulus, where all cell types are well mixed except monocytes, indicating a monocyte-specific response. (C) scDEF's UMAP projection with cells colored by expression counts of gene CXCL10.



Supplementary Figure S3: Application to a high-grade serous ovarian cancer sample with uninformed scDEF. (A) UMAP embedding of the whole data set coloured by previously annotated cell types. (B) UMAP with cells colored by ovary of origin. (C) UMAP embedding colored by assignment of cells to hierarchical factors in scDEF's top layer. (D) Mean score that cell types attribute to each of scDEF's uninformed hierarchical factors, showing that the top layer factors do not follow closely the cell type structure.