Deep generative modeling for quantifying sample-level heterogeneity in single-cell omics

Pierre Boyeau [1, *], Justin Hong [2, 6, *], Adam Gayoso [3], Michael I. Jordan [1, 3, 4], Elham Azizi [2, 5, 6], Nir Yosef [1,3,7 †]

[1] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley
[2] Department of Computer Science, Columbia University
[3] Center for Computational Biology, University of California, Berkeley
[4] Department of Statistics, University of California, Berkeley
[5] Department of Biomedical Engineering, Columbia University
[6] Irving Institute for Cancer Dynamics, Columbia University
[7] Department of Systems Immunology, Weizmann Institute of Science

* These authors contributed equally.
† Correspondence to `niryosef@berkeley.edu`

Abstract

Contemporary single-cell omics technologies have enabled complex experimental designs incorporating hundreds of samples accompanied by detailed information on sample-level conditions. Current approaches for analyzing condition-level heterogeneity in these experiments often rely on a simplification of the data such as an aggregation at the cell-type or cell-state-neighborhood level. Here we present MrVI, a deep generative model that provides sample-sample comparisons at a single-cell resolution, permitting the discovery of subtle sample-specific effects across cell populations. Additionally, the output of MrVI can be used to quantify the association between sample-level metadata and cell state variation. We benchmarked MrVI against conventional meta-analysis procedures on two synthetic datasets and one real dataset with a well-controlled experimental structure. This work introduces a novel approach to understanding sample-level heterogeneity while leveraging the full resolution of single-cell sequencing data.

1 Introduction

Technologies for single-cell omics readily allow multiplexing many samples via molecular labeling or genotype-based strategies [1, 2, 3]. Accordingly, these technologies permit experimental designs that provide single-cell readouts from hundreds of samples corresponding to different treatments, genetic perturbations, and/or individual donors [4, 5, 6, 7]. Beyond increasing the scale of individual experiments, substantial efforts have been made towards integrating many studies into atlases with an emphasis on refining characterizations of cellular phenotypes [8, 9, 10].

These large-scale datasets promise the ability to comprehensively identify cellular and molecular properties that distinguish or are common between samples. In this regime, a key challenge is to derive a metric that can be used to compare samples to one another at high resolution. Such a metric can be used to stratify the data for *exploratory* analysis, i.e., to reveal relevant covariates that are associated with particular phenotypes. For instance, it could help characterize disease sub-types, which in turn may inform treatment options [11]. A sample-sample distance metric can also be used in a *guided* manner, such as to identify subpopulations of cells that are enriched for a sample-level covariate of interest, like disease status.

A common approach for quantifying differences between single-cell samples relies on clustering cells into groups representing cell states/types and then quantifying sample-specific differences in the relative abundance of each group. This approach can be used to evaluate the distance between any pair of samples, and thus enables an *exploratory* analysis [6, 7, 12, 13, 14]. However, it also oversimplifies the task by reducing the high-dimensional omic information of every cell to a single, discrete label. This global comparison of samples may also reduce the power to detect sample effects at the cell-state-level; for example, many diseases and conditions (e.g. lupus [15], autism [16]) exhibit heterogeneous changes in gene expression across cell types and individuals. These issues have been somewhat addressed in a recent line of work on *guided* recovery and analysis of single-cell sample differences [17, 18, 19, 20]. These approaches operate on a graph representation of the data and seek to reveal subpopulations (neighborhoods, clusters, etc.) of cells that are enriched for particular covariates of interest (e.g., treatment). An underlying challenge for these approaches relates to learning a common metric space for cells in which distances capture biological (and not technical) variation. For this reason, all current approaches rely on single-cell integration methods as a preprocessing step, which have varying technical and scalability performance [21].

In this work, we reformulate the task of quantifying sample-level cellular heterogeneity as that of estimating a sample-sample distance matrix *per cell* (Figure 1a, b). This output enables both *exploratory* (sample stratification) and *guided* (identification of metadata-associated cell subpopulations) analyses to be conducted downstream at any desired level of resolution. To estimate the cell-level sample-sample distance matrices, we introduce a hierarchical probabilistic model, MrVI[1] (Multi-resolution Variational Inference), that posits cells as being generated from nested experimental designs such as multi-donor studies in which samples are collected from different clinics.

MrVI provides a normalized view of each cell at two levels. The first level is a low-dimensional stochastic embedding of each cell that is decoupled from its sample-of-origin and any additional known technical factors (e.g., which site the sample came from). This embedding space primarily reflects cell-state properties that are common across samples and can be used to identify biologically-coherent cell groups in complex study designs. The second level adds a latent effect corresponding to the sample-of-origin to each cell's first-level representation while still accounting for technical factors. As such, MrVI manifests as a hierarchical integration method, extending and leveraging concepts used by methods like scVI [22]. The functional relationship between MrVI's latent variable levels can also be used directly to estimate sample-sample distance matrices per cell. Finally, MrVI scales easily to millions of cells due to its reliance on variational inference, implemented with a hardware-accelerated and memory-efficient stochastic gradient descent training procedure within the scvi-tools package [23].

2 A generative model of single-cell transcriptomes in two-stage nested experimental designs

We focus on two-stage nested experimental designs [24] that are commonly used in single-cell omics. For example, multi-site, multi-donor studies nest donors into collection sites, while studies using Perturb-seq-like technologies [5, 25, 26, 27] nest genetic perturbations in multiple microfluidic chips or plates. Without loss of generality, we assume that single-cell sequencing libraries are prepared in B different sites (clinics, plates, chips, etc.), with S total samples collected. In this section, we describe a hierarchical model for these data, MrVI, that aims to learn sample-specific effects on single cells (Section 2.1) while also accounting for technical effects between sites (Section 2.2). The presentation here focuses on single-cell RNA-sequencing (scRNA-seq) as a readout given the vast amount of available data, but we emphasize the MrVI can be extended to other single-cell modalities by using appropriate noise models.

Notations We observe a collection of scRNA-seq experiments measuring transcriptomic profiles in S distinct samples, each sequenced in one of B possible sites. The sequencing provides cell-specific gene expression profiles $\{x_1 \dots x_N\}$, where $x_n \in \mathbb{N}^G$ is a vector of counts for cell n for the G observed genes. Let $s_n \in \{1 \dots S\}$ identify the biological sample cell n originates from and let b_n be the site it was sequenced in.

[1] Code available at `https://github.com/YosefLab/mrvi`

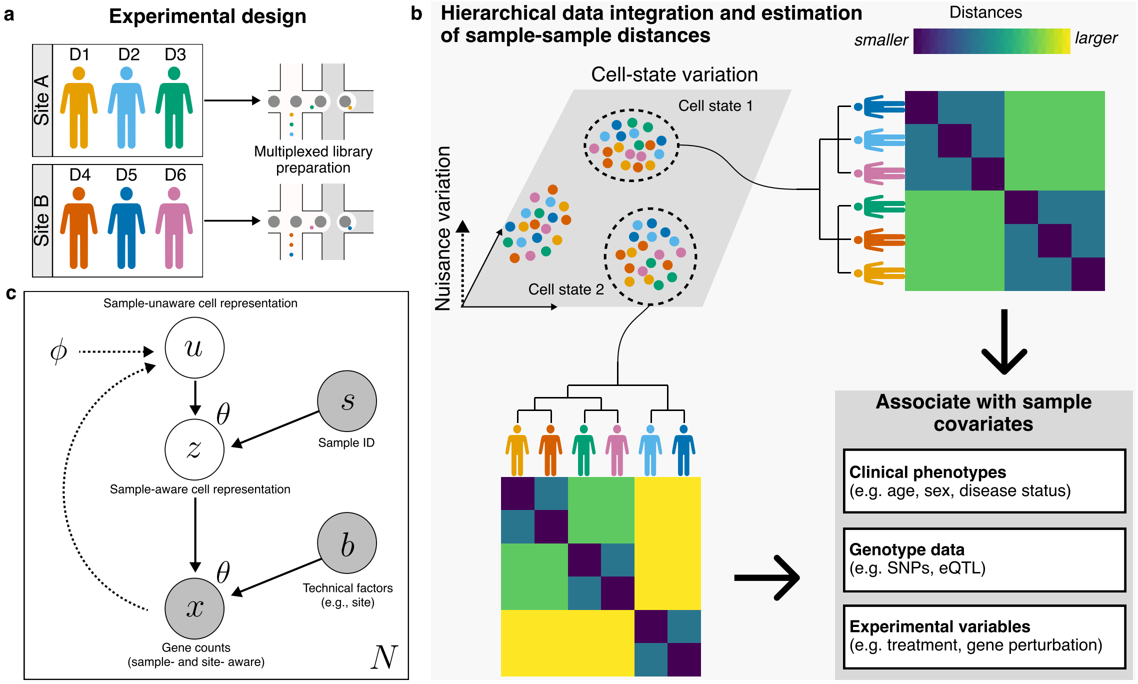

a
Experimental design
Site A
D1
D2
D3
Site B
D4
D5
D6
Multiplexed library preparation
b
Hierarchical data integration and estimation of sample-sample distances
Distances
smaller
larger
Cell-state variation
Cell state 1
Cell state 2
Nuisance variation
Associate with sample covariates
Clinical phenotypes
(e.g. age, sex, disease status)
Genotype data
(e.g. SNPs, eQTL)
Experimental variables
(e.g. treatment, gene perturbation)
c
Sample-unaware cell representation
ϕ
u
θ
z
s
Sample ID
Sample-aware cell representation
θ
b
Technical factors (e.g., site)
x
Gene counts (sample- and site- aware)
N

Figure 1: Overview of MrVI. **a.** We consider two-level, nested experimental designs. In the canonical case, we gather single-cell measurements from several donors, collected across several sites. **b.** In this work, we aim to infer the similarities between biological samples in a multi-site, multi-cohort scRNA-seq at the cell subpopulation level. **c.** MrVI's generative model. A sample-unaware cell representation captures shared type information. From the knowledge of this quantity, and of the sample of origin of the cell, we construct a sample-aware, cell-state representation of the cell, z. Last, we model gene expression as a function of this latent variable and of observed nuisance factors.

2.1 Isolating sample-specific effects on cell state

We consider a two-level hierarchical model of single cells, aiming at disentangling shared and sample-specific sources of variations (Figure 1c). The top-level latent variable u_n is a L-dimensional vector following a normal distribution, $\text{Normal}(0, I_L)$. Here L is a hyperparameter representing the dimension of the latent embedding. This latent variable captures sample-free sources of variation and aligns all samples in a common latent space, and hence carries information about the general identity of the cell. The latent variable at the next level, $z_n \in \mathbb{R}^L$, provides a more general characterization of the cell state, aware of sample-specific environmental factors that may shape its identity [28]. This variable, however, is still decoupled from other nuisance covariates (referred to here as site). The relationship between u_n and z_n is defined as follows:

$$z_n = u_n + f_\theta(u_n, s_n), \quad \text{where} \quad f_\theta(u_n, s_n) = A_{s_n} u_n + a_{s_n}. \tag{1}$$

Here, for any $s \in \{1 \ldots S\}$, the parameters $A_s \in \mathbb{R}^{L \times L}, a_s \in \mathbb{R}^L$ respectively characterize the effects of sample s that are type-specific and shared across types. While Equation 1 writes as a linear combination, it includes a interaction term between samples and states, $A_{s_n} u_n$, allowing to capture sample and state specific variations.

2.2 Accounting for technical variation from multiple sites

Assuming that z_n captures the unobserved biological state of cell n (i.e., representing underlying biological variation and sample-effects), we model normalized expression for gene g, denoted as h_{ng}, as a function of both z_n and the observed technical factors. In particular,

$$(C + C_{b_n})\, z_n + c_{d_n} =: \log h_n \in \mathbb{R}^G, \tag{2}$$

where log is the element-wise log operation. The matrix $C \in \mathbb{R}^{G \times L}$ maps the latent cell states to site-agnostic expression patterns, while for any site b, $C_b \in \mathbb{R}^{G \times L}$ and $c_b \in \mathbb{R}^G$ capture site-specific gene expression patterns.

Following existing approaches [22, 29, 30, 31], we model observed raw transcript counts under conditional independence assumption over the genes as $x_{ng} \sim \text{NegativeBinomial}\left(l_n h_{ng}, r_{ng}\right)$. The size factor l_n, is defined to be the total sum of counts of cell n and $r_{ng} \geq 0$ denotes the the inverse dispersion of the distribution learned during inference.

2.3 Model training using variational inference

The generative model is trained with the auto-encoding variational bayes (AEVB) paradigm [32]. In particular, we introduce a mean-field variational family $q_\phi(u \mid x)$, that lower-bounds the data log-evidence. The resulting evidence lower bound (ELBO) [32] is maximized over the generative model parameters θ and variational parameters ϕ using standard procedures [33, 34]. More details about the optimization procedure can be found in Supplement A.

3 Estimating sample-sample distances at the cellular level

We now describe how MrVI can be used to quantify sample effects on individual cells. We first introduce the computation of local sample-sample distances via counterfactual queries (Section 3.1) and outline how to construct type-specific sample-sample distance matrices via aggregation (Section 3.2). Critically, these procedures do not require sample-level metadata, and can be used for *exploratory* analyses. In cases where sample-level features of interest are available, we also describe a statistical procedure for retrospective *guided* analysis, finding type-specific associations between sample features and gene expression (Section 3.3).

3.1 Counterfactual comparison of cell states

We aim to predict the state of cell n collected in sample s, given that it had been a member of any other sample $s' \neq s$. This counterfactual cell state can help quantify the extent to which any particular cell's gene expression profile could have been explained by a different sample. To avoid technical factors from confounding our analysis, we make these predictions at the level of the z latent variable and denote the counterfactual cell state as $z_n^{s'}$. A straightforward approach to compute $z_n^{s'}$ consists of (i) inferring the sample-free latent representation u_n distribution based on gene expression profiles x_n using the variational posterior $q_\phi(u \mid x_n)$ and (ii) estimating the counterfactual cell state $z_n^{s'} := u_n + f_\theta(u_n, s')$.

Previous tangentially related methods have used a similar approach to model out-of-distribution outcomes when conditioning on categorical covariates [22, 35, 36]. These methods employ nonlinear decoders as part of conditional generative models to map some latent variable and observed covariates to gene expression. Due to the nonlinear aspect of these models, the distances in the respective latent spaces may not reflect relevant variations in gene expression [37, 38]. For instance, distinct latent codes can theoretically encode the same normalized expression profile.

However in the considered setup (Equations 1, 2), neighborhoods of cells in the z latent space characterize cells sharing similar gene expression profiles due to the linearity of the decoding function. Note indeed that for any two latent codes z^a, z^b, generating normalized gene expression profiles h^a, h^b, in the same site b, $\|\log h^a - \log h^b\|_2^2 = \left(z^a - z^b\right)\left(B + B_b\right)^T\left(B + B_b\right)\left(z^a - z^b\right)$. For this reason, we propose comparing cell states directly via the Euclidean distance between counterfactuals $z_n^{s_a}, z_n^{s_b}$ to quantify differences in gene expression across biological samples, as

$$d^{s_a, s_b}(n) := \|z_n^{s_a} - z_n^{s_b}\|_2 \tag{3}$$

These distances can then be used to construct the *sample-sample distance matrix* $D(n) = \left(d^{s_a, s_b}(n)\right)_{s_a, s_b \leq S} \in \left(\mathbb{R}^+\right)^{S \times S}$ for any cell n.

3.2 Aggregating local sample-sample distances

The aforementioned procedure can be extended to quantify sample-sample distances for any cell subpopulation $C \subset \{1, \ldots N\}$ without the need for additional model fitting. To smooth over the noise of individual cells, we consider the aggregate distance matrix, defined as

$$D(C) := \frac{1}{|C|} \sum_{x \in C} D(x). \tag{4}$$

It may occur that an observed sample s contains no cells in C. In such a case, the combination u, s and hence, z, may be unobserved in the training data and result in unreliably inferred distances for sample s in $D(C)$. To mitigate this issue, we discard any samples containing fewer than $k = 5$ observations in C. The matrix $D(C)$ can then be readily plugged in as a dissimiliarity measure for clustering the remaining samples via methods like k-means or hierarchical clustering.

3.3 Associating sample metadata with cell states

The previous approach assumes no knowledge of sample properties. In what we called the *guided* analysis approach, it is relevant to detect cells for which the inferred distances from Equation 3 induce a stratification correlating with a discrete sample characteristic, e.g., disease status. For this purpose, for each cell, we test whether the distributions of distances between samples of the same category are the same as the distances between samples of different categories, under a Kolmogorov-Smirnov test (multiple hypothesis correction is also performed [39]).

4 Experiments

We benchmarked MrVI against traditional methods for sample-sample distance estimation. The first set of methods we considered are useful for *exploratory* analyses and rely on counting the abundance of samples within cell types. For each cell type, we first clustered subpopulations of cells using the Leiden algorithm [40] on cell embeddings that are derived using either principal component analysis (PCA) or scVI. We then compared sample proportions within subclusters as a way to define distances between samples for each cell type. We refer to these baseline approaches as *CompositionPCA* and *CompositionSCVI*, respectively. Details about the implementation of these baselines can be found in Supplement C. For *guided* analyses, we considered Milo [18] for the detection of association between subpopulations and donor metadata. Milo is a statistical framework for differential abundance testing, aiming to detect cell neighborhoods enriched in certain sample groups based on a pre-computed cell-cell graph. In our case, we used the scVI latent space to construct the cell-cell neighbors graph.

We considered three scRNA-seq data sets, one synthetic, one semi-synthetic, and one real experiment in which sample statifications in the different cell subpopulations are either known or can be justified by the experimental design.

4.1 Sample stratification on synthetic data

We simulated, using SymSim [41], a scRNA-seq dataset mimicking a multi-site, multi-sample experiment. SymSim's model produces expression profiles simulating batch effects, measurement noise, biological variability, and transcriptional noise. This simulation reflects a dataset constructed from two batches, each containing 16 samples, and two cell types A and B, for a total of 20,000 cells ($\approx$ 625 cells per sample) and 2,000 genes. Each sample, which could represent donors or perturbations, is characterized by three underlying biological covariates that affect gene expression (Figure 2a). In this experiment, cell type B has uniformly distributed gene expression profiles across samples, but gene expression in cell type A has sample-specific biological variability, such that the correlations between sample-specific effects are known. More details about this dataset can be found in Supplement B.1.

We first assessed the ability of each method to capture the ground truth stratification for type A in an *exploratory* manner by comparing their estimated sample-sample distances (Figure 2b). MrVI provided a local representation of the sample-sample distances that qualitatively matched the ground-truth stratification. To quantify this, we performed hierarchical clustering of samples using the estimated distances and found that MrVI's stratification was the closest to the ground truth in terms of RF distance (Figure 2c).

The *guided* approach discussed in Section 3.3 accurately identified subpopulation stratification according to a predetermined metadata (Figure 2d). For the first two sample covariates, MrVI has more power while

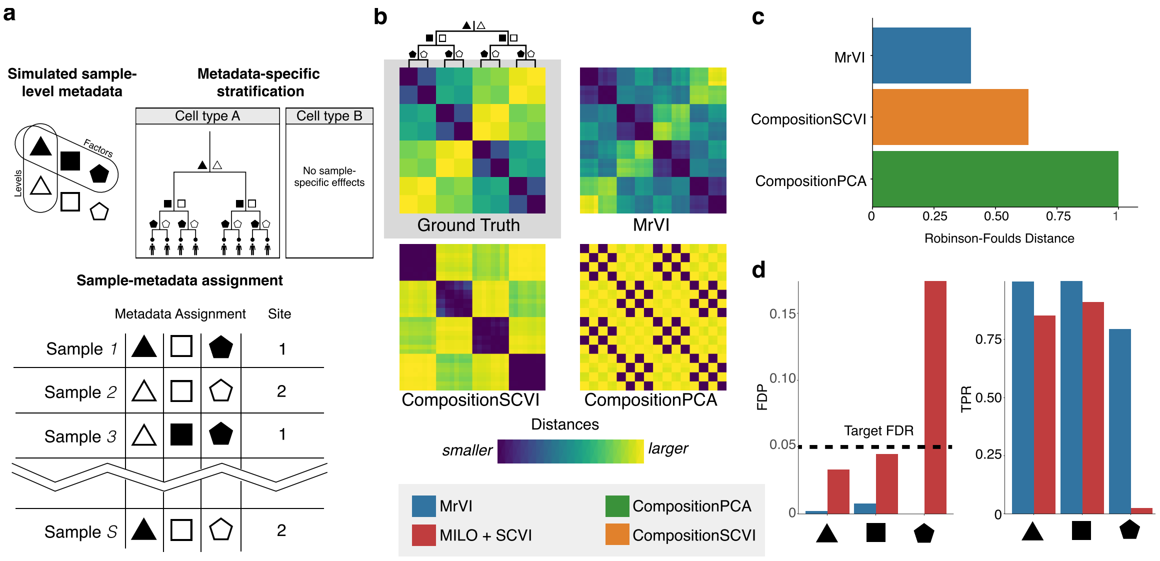


Figure 2: SymSim dataset. **a.** Dataset simulation procedure. We generated gene expression profiles for two cell types A and B, differently affected by metadata. In cell type A, gene expression is affected by three binary metadata. Gene expression for cell type B is uniform across all metadata assignments. Each cell was then assigned to a synthetic sample based on its metadata assignments (32 samples in total, the illustration shows the stratification for 8 samples for conciseness). We also introduced batch effects corresponding to two sites randomly assigned to each sample. **b.** *Exploratory analysis:* Heatmaps showing the distance matrices of the ground truth and of different approaches for cell type A. The samples are ordered by metadata assignments and batch. **c.** Bar plot showing the mean Robinson-Foulds (RF) distances between the ground truth hierarchical clustering and the clustering of different approaches for cell type A, averaged across five random runs of the algorithms (lower is better). MrVI's RF is significantly lower than both CompositionSCVI and CompositionPCA ($p \leq 10^{-5}$, under a one-sided t-test). **d.** *Guided analysis:* False Discovery Rate (FDR) and True Positive Rate (TPR) comparison for significance testing of the relevant sample metadata in cell type A for the different approaches, all aiming to control the FDR at target level $\alpha = 0.05$.

controlling the FDR. For the third covariate, however, only MrVI managed to detect cell subpopulations of type A, as hinted by the respectively large and low TPR of MrVI and MILO.

4.2 Capturing sample-sample distances on semi-synthetic data

We next generated a semi-synthetic dataset from a real scRNA-seq dataset of 12,000 human PBMCs [42] in which samples affect cells heterogeneously (Figure 3a). To do so, we randomly assigned cells to one of 32 synthetic samples, affecting the expression of the cell differently depending on the cell's type. In two cell types, referred to as cluster A and B, we introduced sample-specific *in silico* gene perturbations by modifying gene expression in selected subsets of genes such that the ground-truth sample-sample distances matrices in A and B were known (see Supplement B.2). The rest of the clusters were left unaffected by the sample assignments.

On this dataset, we observed that MrVI estimates local sample distances that more accurately represent the ground-truth stratification of the samples (Figure 3b), while the compositional approaches fail to capture relevant structure. This observation was validated by comparing the RF distances between true and inferred trees by the different algorithms (Figure 3c). We also assessed whether the p-values provided in the *guided* case by our framework are properly calibrated (Figure 3d). While MrVI reaches lower TPR levels than MILO, it better controls the FDR for the considered sample metadata.

4.3 Retrieving meaningful sample stratification from multi-technology single-cell data

Finally, we considered a real dataset [43], in which samples from four healthy human livers were sequenced with both scRNA-seq and single-nuclei RNA-sequencing (snRNA-seq) technologies for a total of 61,486 cells (Figure 4a), and a total of 3,000 genes, selected using highly variable gene selection using seurat [44]. We used the dataset structure in two ways to benchmark sample stratification. First, we evaluated performance in a

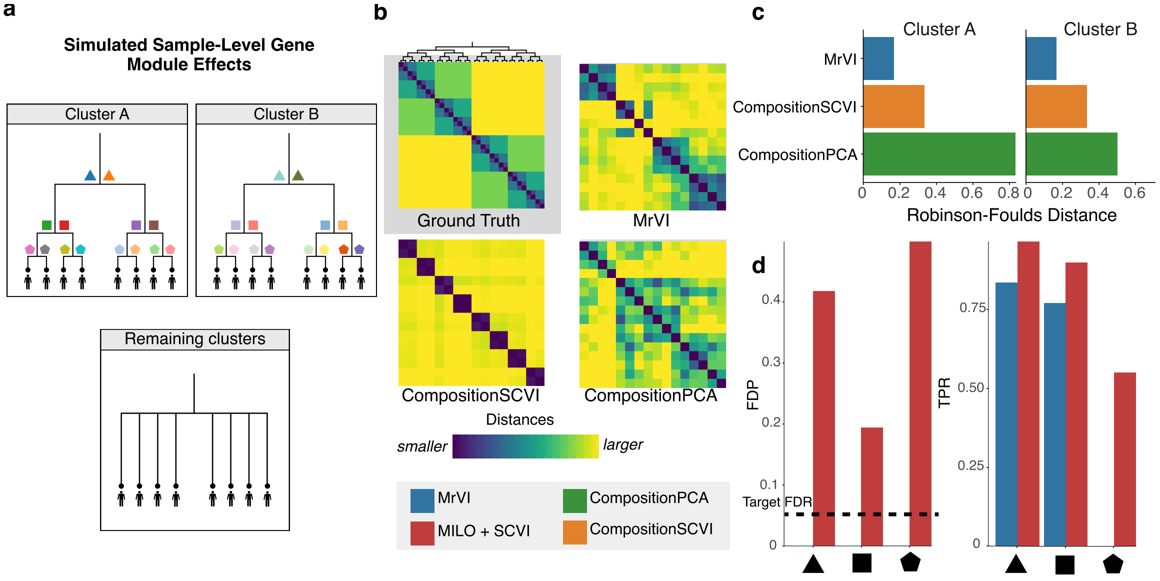


Figure 3: Semi-synthetic dataset. **a.** Dataset presentation. Starting from a scRNA-seq of PBMCs, we mimicked a multi-samples dataset of 32 samples. we generated sample-specific gene perturbations by modifying the gene expression matrix in two cell clusters, while leaving untouched the remaining clusters. the illustration shows the stratification for 8 samples for conciseness. **b.** Heatmaps showing the distance matrices for the ground truth and different approaches. **c.** Bar plots showing the RF distances between the ground truth hierarchical clustering and the clustering of different approaches for both clusters A and B, averaged across five random runs of the algorithms (lower is better). In each cluster, MrVI's RF is significantly lower than both CompositionSCVI and CompositionPCA ($p \leq 10^{-5}$, under a one-sided t-test). **d.** FDR comparison for significance testing of the three most relevant sample metadata in cluster A for the different approaches.

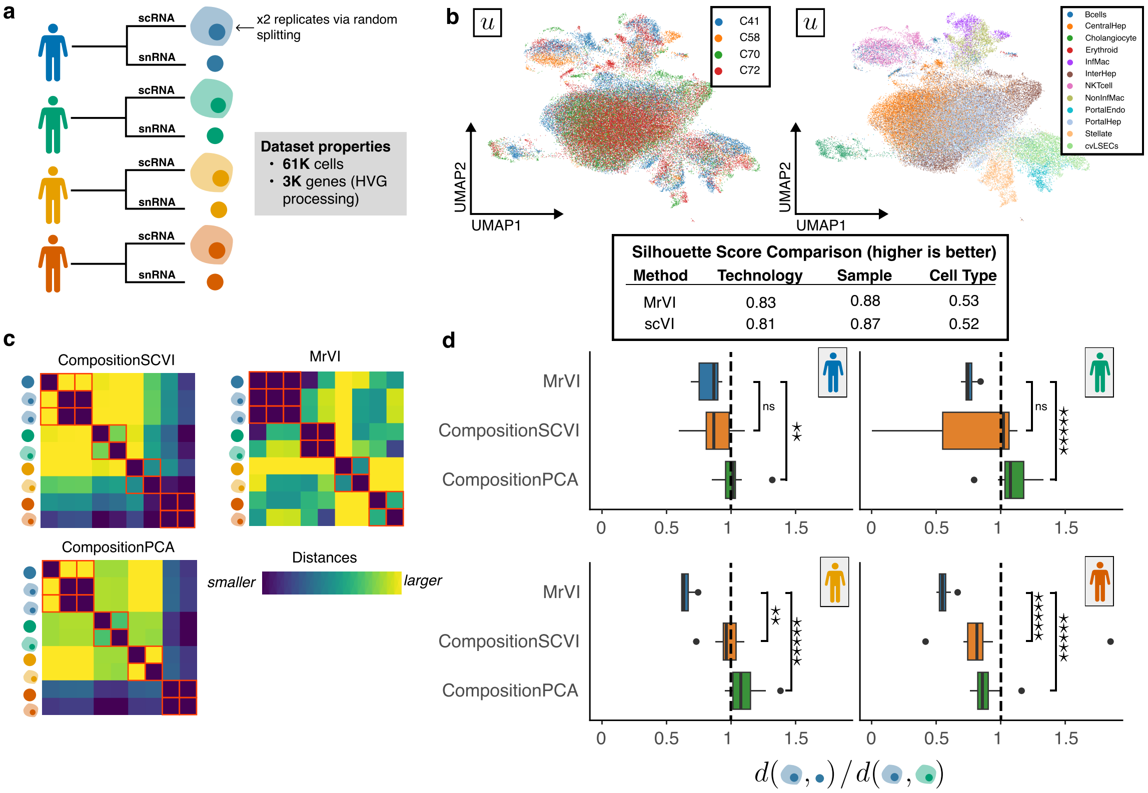
nested experimental design, where each combination of donor and technology represented a different sample. Here, the technology acts as a known nuisance factor, but the donor of each sample, which is shared with another sample measured with a different technology, is hidden from all methods. In this setup, an expected stratification will reflect the (unknown) donor origin of each sample, such that samples from the same donor will be grouped together. As an additional control, we randomly split one of the samples into two samples as another way to benchmark sample stratification inference.

MrVI successfully disentangled technology and donor effects from cell-type variation, providing an efficient way to integrate and annotate single-cell data (Figure 4b). To assess whether our latent representation does not exhibit unwanted variation, we used adjusted silhouette metrics [21] (see Supplement C for more details). In particular, we observed that the u latent representation has comparable silhouette scores (in terms of batch, sample, and author-provided cell annotations) to scVI, showing that the linear decoders from Equations 1, 2 compare favorably to the MLPs used by scVI.

We next visualized the sample-sample distances for the different approaches (Figure 4c). MrVI better captured the sample-sample structure corresponding to the first donor; in particular, it is the only approach successfully attributing low distances to all the relevant samples. To quantitatively evaluate how well the distance matrices aligned with our expectations, we compared ratios of average distances between samples from the same donor against samples from different donors but with the same technology (Figure 4d) over cell types. A ratio below one would indicate that the inferred sample-sample distances reflect samples from the same donor while properly removing technical noise corresponding to the sequencing technology differences. Here, MrVI provided much lower ratios for two of the four donors (almost two fold improvement for the orange and yellow donors), and overall provides significantly lower ratios than all other approaches (one-sided t-test; p-value < 0.001).

5 Discussion

Both the scale and complexity of single-cell omics data are growing, fueled by technological advancements and organized efforts to construct the Human Cell Atlas [45]. One of the major promises of multi-donor studies is



Silhouette Score Comparison (higher is better)

Method	Technology	Sample	Cell Type
MrVI	0.83	0.88	0.53
scVI	0.81	0.87	0.52

Figure 4: snRNA & scRNA experiment. **a.** Four human samples were sequenced with both scRNA- and snRNA-seq technologies. If we consider the sequencing technology as a nuisance factor, we expect cells coming from the same tissue to have similar profiles. **b.** *top*: Comparison of average batch silhouette width scores (using the sequencing technology and donor id as batches), as well as the cell-type silhouette width, using the original study's annotations (higher is better). *bottom*: MrVI's u latent space, colored by donor and cell-type annotations from the original study **c.** Heatmaps showing the sample distance matrices for the different algorithms. Red borders denote pairs of samples originating from the same tissue, and hence, expected to have small distances. **d.** Ratio of distance of same donor different technology over different donor same technology (lower is better). In the figure, *ns* denotes non-significant differences, $2 \times \star, 5 \times \star$ significant differences at respective levels $10^{-2}, 10^{-5}$ under one-side t-tests.

to identify previously unknown stratifications of the donor population that can lead to new discoveries such as disease sub-types with different prognostics and different targeted therapies. It also lends statistical power to identify molecular and cellular features associated with a-priori known stratifications, such as response to therapies. The dramatic increase in scale and resolution of the data enables these studies.

Here we presented MrVI, a deep generative model for single-cell omics that explicitly models nested experimental designs that are increasingly common in the field. MrVI uses hierarchical data integration to produce estimates of sample-sample distances at the resolution of single-cells. By doing so it generalizes concepts used in single-cell data integration [22, 46], as well as recent approaches for fine-grained differential abundance analysis.

In this work, we considered simple decoders to motivate the choice of Euclidean distances in the z latent space for cell state comparisons. For future work, we propose investigating alternatives for assessing sample-sample distances when using nonlinear decoders as well as determining whether they are necessary to improve integration. Removing intricate technical variation is an essential step to analyzing collections of disparate studies, sequenced with potentially different technologies, by different teams, and at different time points. While our work focuses on studies with a single observed nuisance factor, a natural extension of MrVI is to assess how to handle multiple, potentially continuous, and hierarchically-structured technical covariates. With the continued development of cell atlases, understanding how to deploy a model initially trained on such data to analyze a new experiment is a key challenge. Thus, extending MrVI to use transfer learning techniques like scArches [47] could further improve cell state characterization.

Acknowledgments and Disclosure of Funding

N.Y. is an advisor and/or has equity in Cellarity, Celsius Therapeutics, and Rheos Medicine.

References

[1] Christopher S McGinnis, David M Patterson, Juliane Winkler, Daniel N Conrad, Marco Y Hein, Vasudha Srivastava, Jennifer L Hu, Lyndsay M Murrow, Jonathan S Weissman, et al. “MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices”. en. In: *Nat. Methods* (2019).

[2] Vuong Tran, Efthymia Papalexi, Sarah Schroeder, Grace Kim, Ajay Sapre, Joey Pangallo, Alex Sova, Peter Matulich, Lauren Kenyon, et al. “High sensitivity single cell RNA sequencing with split pool barcoding”. en. 2022.

[3] Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, et al. “Multiplexed droplet single-cell RNA-sequencing using natural genetic variation”. en. In: *Nat. Biotechnol.* (2018).

[4] Seyhan Yazar, Jose Alquicira-Hernandez, Kristof Wing, Anne Senabouth, M Grace Gordon, Stacey Andersen, Qinyi Lu, Antonia Rowson, Thomas R P Taylor, et al. “Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease”. en. In: *Science* (2022).

[5] Sanjay R Srivatsan, José L McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A Pliner, Dana L Jackson, Riza M Daza, et al. “Massively multiplex chemical transcriptomics at single-cell resolution”. en. In: *Science* (2020).

[6] Christopher S Smillie, Moshe Biton, Jose Ordovas-Montanes, Keri M Sullivan, Grace Burgin, Daniel B Graham, Rebecca H Herbst, Noga Rogel, Michal Slyper, et al. “Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis”. en. In: *Cell* (2019).

[7] Emily Stephenson, Gary Reynolds, Rachel A Botting, Fernando J Calero-Nieto, Michael D Morgan, Zewen Kelvin Tuong, Karsten Bach, Waradon Sungnak, Kaylee B Worlock, et al. “Single-cell multi-omics analysis of the immune response in COVID-19”. en. In: *Nat. Med.* (2021).

[8] L Sikkema, D Strobl, L Zappia, E Madissoon, N S Markov, L Zaragosi, M Ansari, M Arguel, L Apperloo, et al. “An integrated cell atlas of the human lung in health and disease”. en. 2022.

[9] Chenqu Suo, Emma Dann, Issac Goh, Laura Jardine, Vitalii Kleshchevnikov, Jong-Eun Park, Rachel A Botting, Emily Stephenson, Justin Engelbert, et al. “Mapping the developing human immune system across organs”. en. In: *Science* (2022).

[10] Vinay S Swamy, Temesgen D Fufa, Robert B Hufnagel, and David M McGaughey. “Building the mega single-cell transcriptome ocular meta-atlas”. en. In: *Gigascience* (2021).

[11] Francisco Sanchez-Vega, Marco Mina, Joshua Armenia, Walid K Chatila, Augustin Luna, Konnor C La, Sofia Dimitriadoy, David L Liu, Havish S Kantheti, et al. “Oncogenic Signaling Pathways in The Cancer Genome Atlas”. en. In: *Cell* (2018).

[12] Gokcen Eraslan, Eugene Drokhlyansky, Shankara Anand, Ayshwarya Subramanian, Evgenij Fiskin, Michal Slyper, Jiali Wang, Nicholas Van Wittenberghe, John M Rouhana, et al. “Single-nucleus cross-tissue molecular reference maps to decipher disease gene function”. en. 2021.

[13] Stefan Salcher, Gregor Sturm, Lena Horvath, Gerold Untergasser, Georgios Fotakis, Elisa Panizzolo, Agnieszka Martowicz, Georg Pall, Gabriele Gamerith, et al. “High-resolution single-cell atlas reveals diversity and plasticity of tissue-resident neutrophils in non-small cell lung cancer”. en. 2022.

[14] Jonathan Mitchel, M Grace Gordon, Richard K Perez, Evan Biederstedt, Raymund Bueno, Chun Jimmie Ye, and Peter V Kharchenko. “Tensor decomposition reveals coordinated multicellular patterns of transcriptional variation that distinguish and stratify disease individuals”. en. 2022.

[15] Richard K Perez, M Grace Gordon, Meena Subramaniam, Min Cheol Kim, George C Hartoularos, Sasha Targ, Yang Sun, Anton Ogorodnikov, Raymund Bueno, et al. “Single-cell RNA-seq reveals cell type-specific molecular and genetic associations to lupus”. en. In: *Science* (2022).

[16] Dmitry Velmeshev, Lucas Schirmer, Diane Jung, Maximilian Haeussler, Yonatan Perez, Simone Mayer, Aparna Bhaduri, Nitasha Goyal, David H Rowitch, et al. “Single-cell genomics identifies cell type-specific molecular changes in autism”. en. In: *Science* (2019).

[17] Daniel B Burkhardt, Jay S Stanley 3rd, Alexander Tong, Ana Luisa Perdigoto, Scott A Gigante, Kevan C Herold, Guy Wolf, Antonio J Giraldez, David van Dijk, et al. “Quantifying the effect of experimental perturbations at single-cell resolution”. en. In: *Nat. Biotechnol.* (2021).

[18] Emma Dann, Neil C Henderson, Sarah A Teichmann, Michael D Morgan, and John C Marioni. "Differential abundance testing on single-cell data using k-nearest neighbor graphs". en. In: *Nat. Biotechnol.* (2022).

[19] Yakir A Reshef, Laurie Rumker, Joyce B Kang, Aparna Nathan, Ilya Korsunsky, Samira Asgari, Megan B Murray, D Branch Moody, and Soumya Raychaudhuri. "Co-varying neighborhood analysis identifies cell populations associated with phenotypes of interest from single-cell transcriptomics". en. In: *Nat. Biotechnol.* (2021).

[20] Jun Zhao, Ariel Jaffe, Henry Li, Ofir Lindenbaum, Esen Sefik, Ruaidhrı Jackson, Xiuyuan Cheng, Richard A Flavell, and Yuval Kluger. "Detection of differentially abundant cell subpopulations in scRNA-seq data". en. In: *Proc. Natl. Acad. Sci. U. S. A.* (2021).

[21] Malte D Luecken, M Büttner, K Chaichoompu, A Danese, M Interlandi, M F Mueller, D C Strobl, L Zappia, M Dugas, et al. "Benchmarking atlas-level data integration in single-cell genomics". en. In: *Nat. Methods* (2022).

[22] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef. "Deep generative modeling for single-cell transcriptomics". en. In: *Nat. Methods* (2018).

[23] Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, et al. "A Python library for probabilistic analysis of single-cell omics data". en. In: *Nat. Biotechnol.* (2022).

[24] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.

[25] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, et al. "Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens". en. In: *Cell* (2016).

[26] Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, et al. "Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq". en. In: *Cell* (2022).

[27] Paul Datlinger, André F Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, Linda C Schuster, Amelie Kuchler, Donat Alpar, et al. "Pooled CRISPR screening with single-cell transcriptome readout". en. In: *Nat. Methods* (2017).

[28] Allon Wagner, Aviv Regev, and Nir Yosef. "Revealing the vectors of cellular identity with single-cell genomics". en. In: *Nat. Biotechnol.* (2016).

[29] Christoph Hafemeister and Rahul Satija. "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression". en. In: *Genome Biol.* (2019).

[30] Valentine Svensson. "Droplet scRNA-seq is not zero-inflated". en. In: *Nat. Biotechnol.* (2020).

[31] F William Townes, Stephanie C Hicks, Martin J Aryee, and Rafael A Irizarry. "Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model". en. In: *Genome Biol.* (2019).

[32] Diederik P Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: (2013). arXiv: `1312.6114v10 [stat.ML]`.

[33] Diederik P Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: (2014). arXiv: `1412.6980 [cs.LG]`.

[34] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. "On Early Stopping in Gradient Descent Learning". en. In: *Constr. Approx.* (2007).

[35] Mohammad Lotfollahi, Mohsen Naghipourfar, Fabian J Theis, and F Alexander Wolf. "Conditional out-of-sample generation for unpaired data using trVAE". In: (2019). arXiv: `1910.01791 [cs.LG]`.

[36] Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Yuge Ji, Ignacio L Ibarra, F Alexander Wolf, Nafissa Yakubova, Fabian J Theis, and David Lopez-Paz. "Learning interpretable cellular responses to complex perturbations in high-throughput screens". en. 2021.

[37] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. "Latent Space Oddity: on the Curvature of Deep Generative Models". In: (2017). arXiv: `1710.11379 [stat.ML]`.

[38] Georgios Arvanitidis, Søren Hauberg, and Bernhard Schölkopf. "Geometrically Enriched Latent Spaces". In: (2020). arXiv: `2008.00565 [stat.ML]`.

[39] Y Benjamini and Y Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *J. R. Stat. Soc.* (1995).

[40] V A Traag, L Waltman, and N J van Eck. "From Louvain to Leiden: guaranteeing well-connected communities". en. In: *Sci. Rep.* (2019).

[41] Xiuwei Zhang, Chenling Xu, and Nir Yosef. "Simulating multiple faceted variability in single cell RNA sequencing". en. In: *Nat. Commun.* (2019).

[42] Grace X Y Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, et al. "Massively parallel digital transcriptional profiling of single cells". en. In: *Nat. Commun.* (2017).

[43] Tallulah S Andrews, Jawairia Atif, Jeff C Liu, Catia T Perciani, Xue-Zhong Ma, Cornelia Thoeni, Michal Slyper, Gökcen Eraslan, Asa Segerstolpe, et al. "Single-Cell, Single-Nucleus, and Spatial RNA Sequencing of the Human Liver Identifies Cholangiocyte and Mesenchymal Heterogeneity". en. In: *Hepatol Commun* (2022).

[44] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck 3rd, Yuhan Hao, Marlon Stoeckius, Peter Smibert, et al. "Comprehensive Integration of Single-Cell Data". en. In: *Cell* (2019).

[45] Regev, Teichmann, Lander, Amit, Benoist, et al. "Science forum: the human cell atlas". In: *Elife* (2017).

[46] Laleh Haghverdi, Aaron T L Lun, Michael D Morgan, and John C Marioni. "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors". en. In: *Nat. Biotechnol.* (2018).

[47] Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, Adam Gayoso, Nir Yosef, et al. "Mapping single-cell data to reference atlases by transfer learning". en. In: *Nat. Biotechnol.* (2021).

[48] Harm De Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C Courville. "Modulating early visual processing by language". In: *Adv. Neural Inf. Process. Syst.* (2017).

[49] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *J. Mach. Learn. Res.* (2014).

[50] Paszke, Gross, Massa, Lerer, et al. "Pytorch: An imperative style, high-performance deep learning library". In: *Adv. Neural Inf. Process. Syst.* (2019).

Supplement

A Model training

Objective function and training procedure We aim to find generative model parameters that maximize the conditional log-evidence of the data $\log p_\theta(x)$. Unfortunately, this requires to marginalize the latent variables of the model, which has no closed-form for this problem. For this reason, we resort to variational inference for model training. In particular, we introduce mean-field variational distributions $q_\phi(z_o \mid x, d, b) \sim \text{Normal}(\mu_\phi(x, d, b), \sigma_\phi(x, d, b)^2)$, whose parameters are the outputs of neural networks to construct an easier-to-optimize lower bound on the log-evidence:

$$\log p_\theta(x) \geq \mathbb{E}\left[\log \frac{p_\theta(x \mid z_o, d, b) p(z_o)}{q_\phi(z_0 \mid x, d, b)}\right] =: \mathcal{L}_{\text{ELBO}}$$

We optimize the algorithm using the Adam optimizer [33] using minibatches of size 256, with early-stopping.

Implementation details We use MLPs to parameterize the variational distributions means and log variances using ReLU activation functions and conditional batch normalization [48]. To avoid learning site-confounded z representations, we did not include the interactions terms $A_{s_n} u_n$ and $C_{b_n} z_n$, respectively appearing in Equations 1, 2, to compute gradients with respect to the latent variables. Additionally, we used dropout regularization [49] on the $C_{b_n} z_n$, which empirically improved performance. MrVI was implemented in PyTorch [50] using scvi-tools [23], and the same architecture was used for MrVI in all experiments, with $L = 10$.

B Data generation details

B.1 SymSim

SymSim [41] is a scRNA-seq simulation framework relying on a promoter kinetic model of gene expression. To model donor-specific variation, we simulated three underlying binary metadata associated with the generated cells. This was done by concatenating independently generated metadata associated-EVFs (external variability factors) in conjunction to the cell type-associated EVFs and non-differentially expressed EVFs.To model variation in the effect of metadata on gene expression, the three sets of EVFs were generated with different levels of covariance. The concatenated EVFs were then used to generated the observed gene counts with additionally batch effect between two batches. In the second cell type, the same number of EVFs were generated from an independent seed to simulate a single homogeneous population, irrespective of metadata.

Now with a cell-by-gene matrix with associated metadata, we assigned donor labels based on the ground truth cell metadata. For the first cell type, donor labels were assigned to the subset of cells with matching metadata. For example, for the donor exhibiting the metadata assignment (a, a, b), we would sample from the subset of cells generated with the respective metadata values. Although the cells in the second cell type were selected based on attached metadata, the distribution of the cell states was uniform across donors due to the EVF generation procedure.

B.2 PBMCs

To generate the semi-synthetic data set, we randomly assign one of 32 random donors to each cell of the real data [41]. Next, we introduce cell-type specific sample perturbation to two cell subpopulations, corresponding to two clusters obtained from Leiden clustering on scVI [22] latent representations of the cells in the data. In each of these subpopulations, we randomly generate a ground-truth sample similarity binary tree, whose leaves correspond to the donors. Next, we randomly partition the genes of the data set, such that a specific set of genes corresponds to each edge in the tree. For all cells of the given cluster and sample, we perturb all the genes associated to the root to sample leaf edges, by doubling gene expression for all the relevant genes.

Figure S1 illustrates the data generation scheme in the simpler case where only eight donors are considered. In particular, for sample $D2$, all cells belonging to cluster 1 will have doubled gene expression for genes belonging in $G_1 \cup G_2 \cup G_4$, and cells of cluster 2 doubled gene expression for genes in G'_1, G'_5, G'_6.

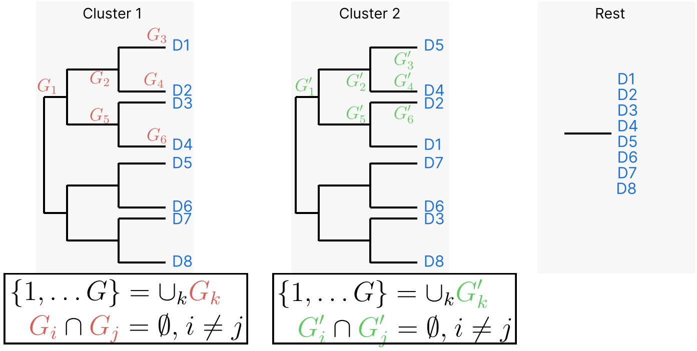


Figure S1: Semisynthetic data generation

C Baselines

C.1 Models

Differences in composition Assuming access to cell representations $\{z_n, n \leq N_C\}$ and to donor assignments $\{d_n, n \leq N_C\}$ in a given cluster C of cells, we use Leiden to partition the cluster into K clusters. For each donor, we then compute the proportion of cells coming from donor d, in cluster k, that we denote f_d^k. Finally, we compute the sample distance matrix from Euclidean distances between the proportions.

MILO To detect cell states affect by a sample characteristic, MILO performs differential abundance testing in the following way. It first construct a k-nearest neighbors graph based on precomputed cell representations, supposed site-agnostic. After inferring a set of characteristic neighborhoods in the graph, MILO compares the number of cells coming from each conditon via differential abundance testing that relies on a negative binomial regression model. We used scVI's cell representations as inputs for MILO.

C.2 Metrics

Cell-type silhouette scores We consider averaged silhouette width scores computed as in ref. [21] as a way to assess the relevance and the proper mixing of the latent representation u. To do so, we first compute the silhouette score with respect to author-provided cell-type annotations. For any cell n with cell representation $r(n)$, belonging to annotation C_o, let $d(n, C)$ denote the mean distance of $r(n)$ to representations of annotation C, excluding n if $C = C_o$. let $a(n)$ denote the average distance of $r(n)$ to cells of the same annotation, and $b(n)$ the smallest mean distance of $r(n)$. The silhouette score for cell n is computed as

$$s(n) = \frac{\min_{C, C \neq C_o} d(n, C) - d(n, C_o)}{\max\{\min_{C, C \neq C_o} d(n, C), d(n, C_o)\}}, \tag{5}$$

and the overall dataset silhouette score is the average of rescaled silhouette scores across all cells in the data. The rescaling, $\tilde{s}(n) = \frac{1}{2}(s(n) + 1)$, puts the dataset score in the range $(0, 1)$ This score assesses to what extent the data representations cluster according to the annotations. When the dataset score is 1, representations with the same annotation perfectly cluster together.

Batch silhouette scores We also used the silhouette to measure the extent to which batch IDs mix together in the latent space. To do so, we follow the procedure described in ref. [21], which consists of, for each previously-annotated cell type: (i) computing cell silhouette scores with respect to the batch assignments, (ii) rescaling these scores, such that $\hat{s}(n) = 1 - |s(n)|$, and (iii) computing an overall silhouette score computed as a weighted average of $\hat{s}(n)$, to ensure that that each cell type gets the same contribution.