Isolating salient variations of interest in single-cell transcriptomic data with contrastiveVI

Ethan Weinberger[1,*], Chris Lin[1,*], and Su-In Lee[1,✉]

[1]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle

[*] denotes equal contribution.

[✉] corresponding: suinlee@cs.washington.edu

Abstract

Single-cell RNA sequencing (scRNA-seq) technologies enable a better understanding of previously unexplored biological diversity. Oftentimes, researchers are specifically interested in modeling the latent structures and variations enriched in one *target* scRNA-seq dataset as compared to another *background* dataset generated from sources of variation irrelevant to the task at hand. For example, we may wish to isolate factors of variation only present in measurements from patients with a given disease as opposed to those shared with data from healthy control subjects. Here we introduce Contrastive Variational Inference (contrastiveVI; `https://github.com/suinleelab/contrastiveVI`), a framework for end-to-end analysis of target scRNA-seq datasets that decomposes the variations into shared and target-specific factors of variation. On three target-background dataset pairs we demonstrate that contrastiveVI learns latent representations that recover known subgroups of target data points better than previous methods and finds differentially expressed genes that agree with known ground truths.

Main

Single-cell RNA sequencing (scRNA-seq) technologies have emerged as powerful tools for understanding previously unexplored biological diversity. Such technologies have enabled advances in our understanding of biological processes such as those underlying cancer [38],

Alzheimer's disease [13, 28], and COVID-19 [36]. In many settings, scRNA-seq data analysts are specifically interested in patterns that are enriched in one dataset, referred to as the *target*, as compared to a second related dataset, referred to as the *background*. Such target and background dataset pairs arise naturally in many biological research contexts. For example, data from healthy controls versus a diseased population or from pre-intervention and post-intervention groups form intuitive background and target pairs. Moreover, with the development of new technologies for measuring the effects of large numbers of perturbations in parallel, such as Perturb-Seq [9] and MIX-Seq [29], tools for better understanding variations unique to such perturbed cell lines compared to control populations will be critical.

Isolating salient variations present only in a target dataset is the subject of *contrastive analysis* (CA) [40, 3, 17, 22, 32, 2]. While many recent studies have modeled scRNA-seq data by fitting probabilistic models and representing the data in a lower dimension [23, 30, 16, 26, 24, 25], few were designed for CA. Such methods are thus unlikely to capture the enriched variations in a target dataset, which are often subtle compared to the overall variations in the data [3]. One recent study [17] designed a probabilistic model for analyzing scRNA-seq data in the CA setting. However, this method assumes that a generalized linear model is sufficiently expressive to model the variations in scRNA-seq data, even though previous work has demonstrated substantial improvements by using more expressive nonlinear methods [23].

To address these limitations, we developed contrastiveVI, a deep generative model that enables analysis of scRNA-seq data in the CA setting. contrastiveVI learns a probabilistic representation of the data that accounts for the specific technical biases and noise characteristics of scRNA-seq data as well as batch effects. Moreover, to handle CA tasks, contrastiveVI models the variations underlying scRNA-seq data using two sets of latent variables: the first, called the *background variables*, are shared across background and target cells while the second, called the *salient variables*, are used to model variations specific to target data. contrastiveVI can be used for a number of analysis tasks, including dimensionality reduction, target dataset subgroup discovery, and differential gene expression testing. To highlight this functionality, we applied contrastiveVI to three publicly available background and target scRNA-seq dataset pairs, and demonstrated strong performance on all of them.

Results

The contrastiveVI Model

contrastiveVI is a probabilistic latent variable model that represents the uncertainty in observed RNA counts as a combination of biological and technical factors. The input to

the contrastiveVI model consists of an RNA unique molecular identifier (UMI) count matrix along with labels denoting each cell as belonging to the background or target dataset (**Figure 1a**). Additional categorical covariates such as anonymized donor ID or experimental batch are optional inputs to the model that can be used to integrate datasets.

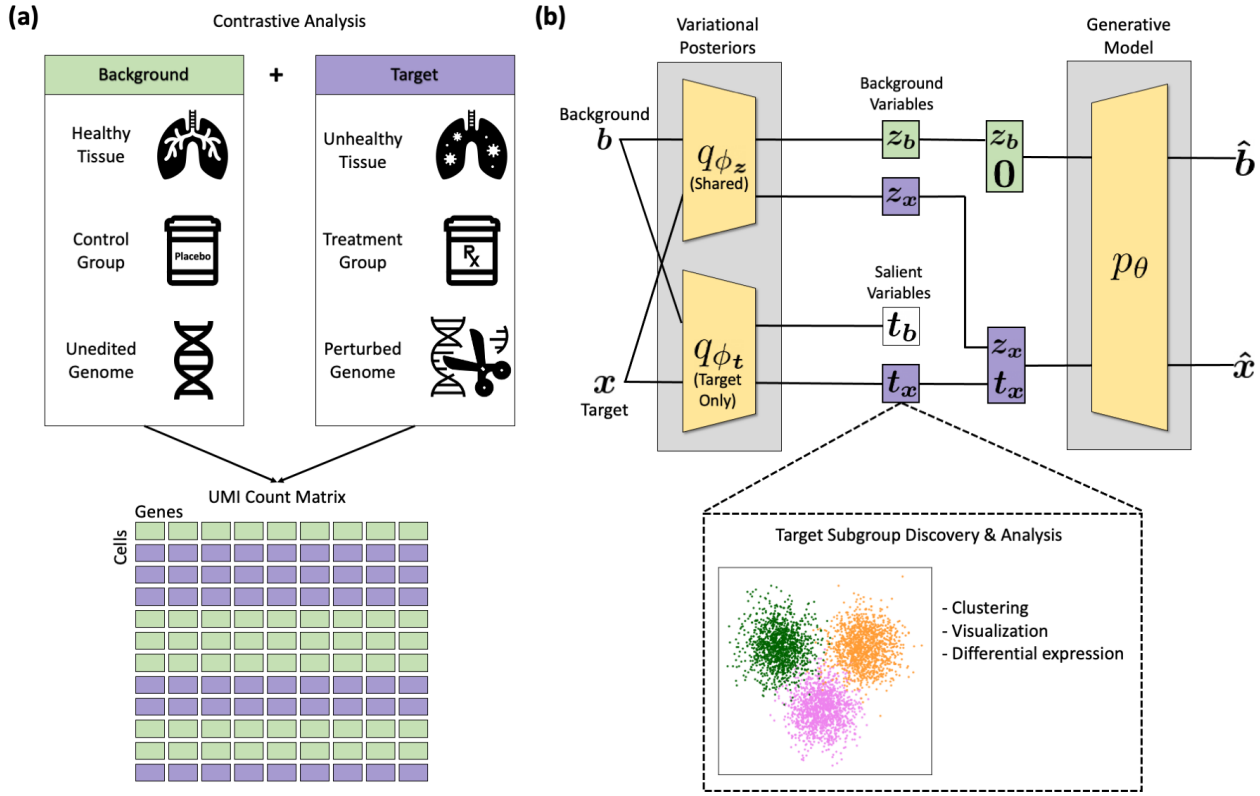

Figure 1: Overview of contrastiveVI. Given a reference background dataset and a second target dataset of interest, contrastiveVI separates the variations shared between the two datasets and the variations enriched in the target dataset. **a**, Example background and target data pairs. Samples from both conditions produce an RNA count matrix with each cell labeled as background or target. **b**, Schematic of the contrastiveVI model. A shared encoder network q_{ϕ_z} transforms a cell into the parameters of the posterior distribution for z, a low-dimensional set of latent factors shared across target and background data. For target data points only, a second encoder q_{ϕ_t} encodes target data points into the parameters of the posterior distribution for t, a second set of latent factors encoding variations enriched in the target dataset and not present in the background.

contrastiveVI encodes each cell as the parameters of a distribution in a low-dimensional latent space. This latent space is divided into two parts, each with its own encoding function. The first set of latent variables, called the background variables, capture factors of variation that are shared among background and target data. The second set of variables, denoted as the salient variables, capture variations unique to the target dataset. Only target data points are given salient latent variable values; background data points are instead assigned

a zero vector for these variables to represent their absence. As with scVI [23], contrastiveVI also provides a way to estimate the parameters of the distributions underlying the observed RNA measurements given a cell's latent representation. Such distributions explicitly account for technical factors in the observed data such as sequencing depth and batch effects. All distributions are parameterized by neural networks.

The contrastiveVI model is based on the variational autoencoder (VAE) framework [21]. As such, its parameters can be learned using efficient stochastic optimization techniques, easily scaling to large scRNA-seq datasets consisting of measurements from tens or hundreds of thousands of cells. Following optimization, we can make use of the different components of the contrastiveVI model for downstream analyses. For example, the salient latent representations of target data can be used as inputs to clustering or visualization algorithms to discover subgroups of target points. Moreover, the distributional parameters can be used for additional tasks such as imputation or differential gene expression analysis. A more detailed description of the contrastiveVI model can be found in **Methods**.

contrastiveVI isolates subtle variations in target cells

To evaluate the performance of contrastiveVI and other methods, we rely on datasets with known biological variations in the target condition that are not present in the background condition. One such dataset consists of expression data from bone marrow mononuclear cells (BMMCs) from two patients with acute myeloid leukemia (AML) and two healthy controls. The two patients underwent allogenic stem-cell transplants, and BMMC samples were collected before and after the transplant. It is known that gene expression profiles of BMMCs differ pre- and post-transplant [39]. Therefore, the known biological variations in this target dataset (AML patient BMMCs) correspond to pre- vs. post-transplant cellular states. A performant model should learn a salient latent space separating pre- vs. post-transplant status, while the latent space from a non-performant model does not make this distinction.

Qualitatively, pre- and post-transplant cells are well separated in the salient latent space learned by contrastiveVI (**Figure 2a**). We also quantified how well contrastiveVI's salient latent space separates the two groups of target cells using three metrics—the average silhouette width, adjusted Rand Index (ARI), and adjusted mutual information (AMI; **Methods**). We find that contrastiveVI performs well on all of these metrics (**Figure 2b**), indicating that it successfully recovers the variations enriched in the target dataset. Furthermore, we experimented with a workflow for using contrastiveVI for end-to-end biological discovery. After embedding the AML patient samples into the contrastiveVI salient latent space, we used

k-means clustering to divide the samples into two groups. Highly differentially expressed genes across the two clusters were then obtained by Monte Carlo sampling of denoised, library size-normalized expressions from the contrastiveVI decoder (**Methods**). Finally, pathway enrichment analysis (**Methods**) was performed with these differentially expressed genes using the Kyoto Encyclopedia of Genes and Genomes (KEGG) 2016 pathway database [18]. Based on our quantitative results, our two clusters exhibited strong agreement with the two ground-truth groups (ARI: 0.77 ± 0.01). Moreover, the pathways enriched by the differentially expressed genes between the two clusters are related to immune response and graft rejection (**Figure 2c**). We provide a full list of enriched pathways in **Supplementary Table 1**. These results align with known cellular state transitions of BMMCs before and after a transplant.

contrastiveVI outperforms other modeling approaches

To illustrate the advantages of contrastiveVI, we benchmarked its performance against that of three previously proposed methods for analyzing raw scRNA-seq count data. First, to demonstrate that our contrastive approach is necessary for capturing enriched variations in target datasets, we compared against scVI [23]. scVI has achieved state-of-the-art results on many tasks; however, it was not specifically designed for the CA setting and thus may struggle to isolate salient variations of interest. We also compared against two contrastive methods designed for analyzing scRNA-seq count data: contrastive Poisson latent variable model (CPLVM) and contrastive generalized latent variable model (CGLVM) [17]. While these methods are designed for the contrastive setting, they both make the strong assumption that linear models can accurately capture the complex variations in scRNA-seq data. To our knowledge, the CPLVM and CGLVM methods are the only existing contrastive methods for analyzing scRNA-seq count data.

Qualitatively (**Figure 2a**), we find that none of these baseline models are able to separate pre- and post-transplant cells as well as contrastiveVI can. This finding is further confirmed by quantitative results (**Figure 2b**). Across all of our metrics we find that contrastiveVI significantly outperforms baseline models, with especially large gains in the ARI and AMI. These results indicate that contrastiveVI recovered the variations enriched in the AML patient data far better than baseline models.

contrastiveVI separates intestinal epithelial cells by infection type

We next applied contrastiveVI to data collected in Haber et al. [15]. This data consists of gene expression measurements of intestinal epithelial cells from mice infected with either

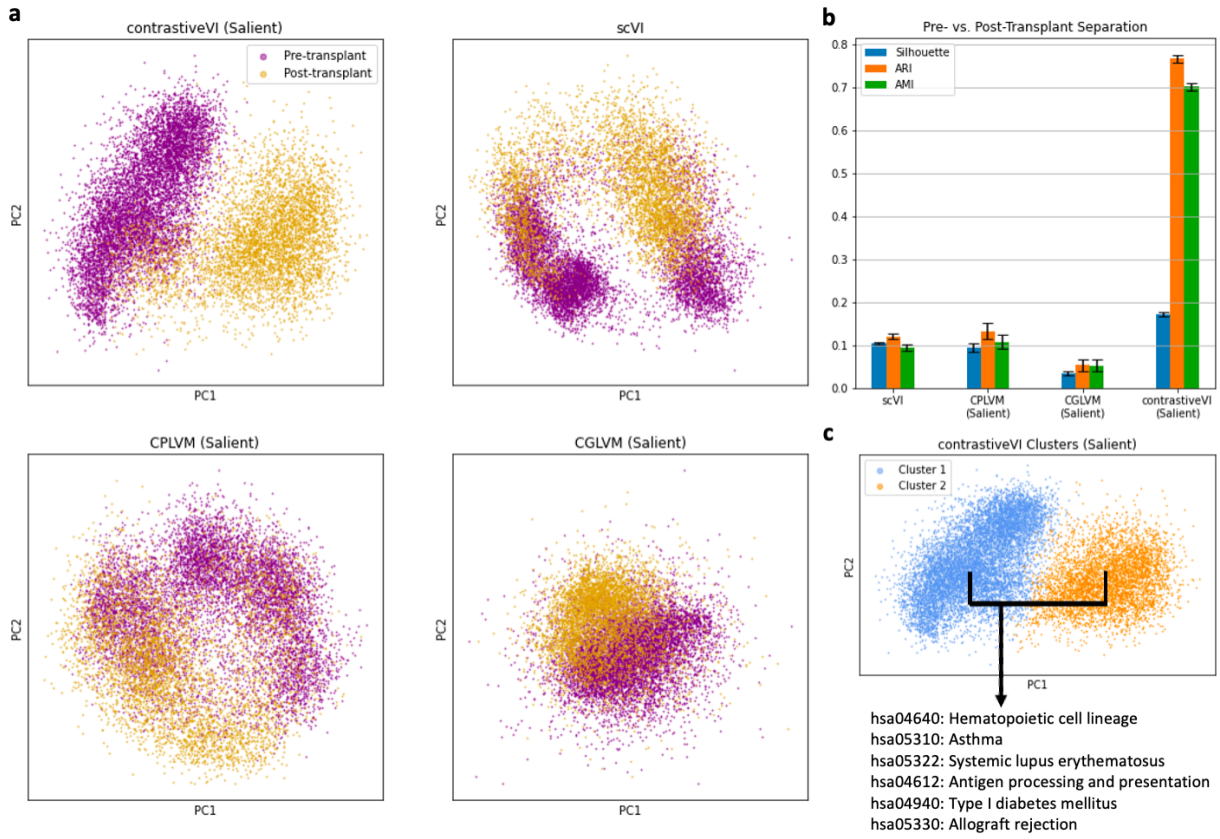

Figure 2: contrastiveVI successfully captures enriched variations in scRNA-seq data. a, Principal component (PC) plots of contrastiveVI and baseline models' latent representations. For scVI, the first two PCs of the model's single latent representations are plotted, while for contrastive methods the PCs from their salient latent representations are plotted. **b**, Quantitative measures of separation between pre- and post-transplant cells. Silhouette is the average silhouette width of pre-annotated subpopulations, ARI is the adjusted Rand index, and AMI is the adjusted mutual information. Higher values indicate better performance for all metrics. For each method, the mean and standard error across five random trials are plotted. **c**, contrastiveVI's salient latent representations of the target dataset were clustered into two groups. Pathway enrichment analysis was then performed on the differentially expressed genes between the two clusters.

Salmonella or *Heligmosomoides polygyrus (H. poly)*. As a background dataset we used measurements collected from healthy cells released by the same authors. Here our goal is to separate cells by infection type in the salient latent space. On the other hand, any separations in the background latent space should reflect variations shared between healthy and infected cells, such as those due to cell type differences. We present our results in Figure 3.

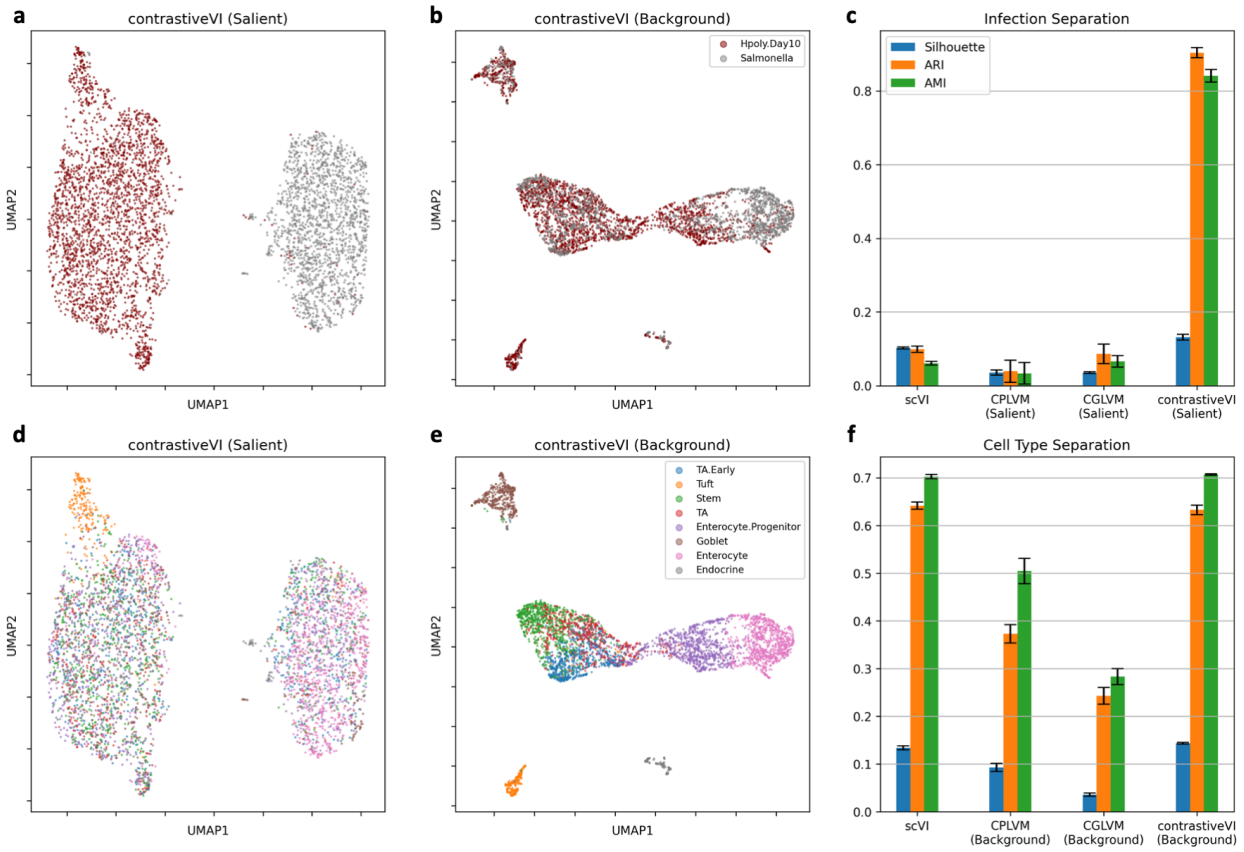

Figure 3: contrastiveVI isolates responses to different infections in intestinal epithelial cells. a,b, UMAP plots of contrastiveVI's salient and background representations colored by infection type. Cells are correctly separated by infection type in the salient space, while they mix across infection types in the background space. **c,** Clustering metrics quantify how well cells separate by infection type for scVI's single latent space and contrastive models' salient latent spaces, with means and standard errors across five random trials plotted. **d,e,** UMAP plots of contrastiveVI's salient and background representations colored by cell type. Cells separate well by cell type in the background space, while they mix across cell types in the salient space. **f**, Quantifying how well cells separate by cell type in scVI's single latent space and contrastive models' background latent spaces, with means and standard errors across five random trials for each method.

We find that contrastiveVI successfully separates the cells by infection type in its salient

latent space (**Figure 3a**). Moreover we find that cells mix across infection types in our background latent space as expected (**Figure 3b**). These results indicate that enriched variations due to infection response are correctly being relegated to the salient latent space. Once again we find that previously proposed methods fail to stratify the two classes of target samples in their salient latent spaces as demonstrated by a set of quantitative metrics (**Figure 3c**). For this dataset we were able to further validate contrastiveVI's separation of target and background variations using ground truth cell type labels provided by the authors (**Supplementary Table 2**). In particular, we found strong mixing across cell types in contrastiveVI's salient latent space (**Figure 3d**), while cell types separated clearly in the background latent space (**Figure 3e**). Our quantitative metrics indicate that contrastiveVI's background latent space is competitive with if not outright superior to other methods' at capturing variations between cell types (**Figure 3f**). Taken together, these results further indicate that contrastiveVI successfully disentangles variations enriched in target data from those shared across the target and background, even when other methods struggle.

contrastiveVI stratifies cells by response to molecular perturbations

In addition to studying transplant outcome and infection response, contrastiveVI can be applied to examine drug treatment response. We demonstrate this capability using cancer cell lines treated with vehicle control dimethyl sulfoxide (DMSO) or idasanutlin collected by McFarland et al. [29]. The small molecule idasanutlin is an antagonist of MDM2, a negative regulator of the tumor suppresor protein p53, hence offering cancer therapeutic opportunity [35]. In the CA context, DMSO-treated samples are considered the background dataset, and idasanutlin-treated samples the target dataset. Based on the mechanism of action of idasanutlin, activation of the p53 pathway is observed in cell lines with wildtype *TP53* (gene of p53) and not in transcriptionally inactive mutant *TP53* cell lines [35]. Therefore, unique variations in the target dataset should be related to *TP53* mutation status. This stratification of cell response based on *TP53* mutation is readily identified by the salient latent space of all methods (**Figure 4a** and **Figure 4b**). Notably, contrastiveVI outperforms other methods based on ARI and AMI, providing better separated clusters for downstream analyses. (**Figure 4b**). Particularly, the two clusters identified using the contrastiveVI salient latent space have differentially expressed genes enriched for the p53 signaling pathway (**Figure 4c**). It is worth noting that the p53 signaling pathway is the only statistically significant (under 0.05 false discovery rate) pathway identified by contrastiveVI. All these results show that contrastiveVI finds salient variations in the target samples treated with idasanutlin that specifically relate to the biological ground truth effect of idasanutlin perturbation.

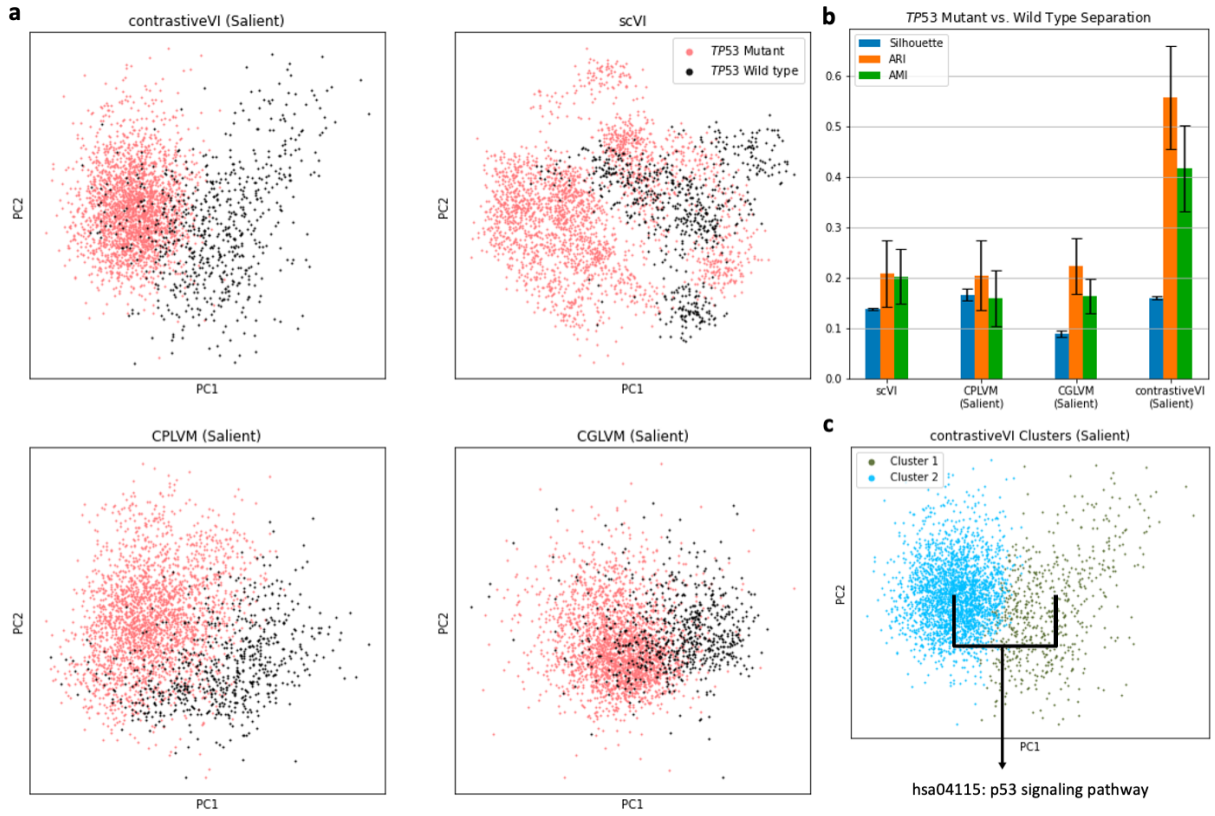

Figure 4: contrastiveVI stratifies cancer cell lines by response to idasanutlin. a, PC plots of target data latent representations from contrastiveVI and baseline models. The first two PCs of scVI's single latent space are plotted. For contrastive methods, the first two PCs of their salient latent space are plotted. **b**, The average silhouette (silhouette), adjusted Rand Index (ARI) and adjusted mutual information (AMI), with mean and standard error across five random trials plotted for each method. **c**, Two clusters identified by k-means clustering with contrastiveVI's salient latent representations of the target dataset. Highly differentially expressed genes were identified from the two clusters, and these genes were used to perform pathway enrichment analysis.

Discussion

In this work we introduced contrastiveVI, a scalable probabilistic framework for isolating enriched variations in a target scRNA-seq dataset as compared to a related background dataset. contrastiveVI is the first method designed to analyze scRNA-seq data in the contrastive analysis setting that both explicitly models the technical factors of variation in scRNA-seq data and takes advantage of the expressive power of deep generative modeling techniques. Moreover, contrastiveVI includes a number of other capabilities relevant to scRNA-seq analysis out of the box, such as batch effect correction and differential expression testing.

In three different contexts—response to cancer treatment, infection by different pathogens, and exposure to small-molecule drug perturbations—we demonstrated that contrastiveVI isolated enriched variations in target cells while other methods struggled. With the recent development of new sequencing technologies for efficiently measuring transcriptomic responses to various perturbations, such as Perturb-Seq and MIX-Seq, we expect contrastiveVI to be of immediate interest to the scRNA-seq research community. Moreover, contrastiveVI was implemented using the scvi-tools [11] Python library, thereby enabling interoperability with Scanpy [37] and Seurat [33] analysis pipelines.

The ideas behind contrastiveVI admit multiple potential directions for future work. Similar contrastive disentanglement techniques could be used to extend models that make use of multimodal data, such as totalVI [12], to better understand how variations enriched in target datasets are expressed across different modalities of single-cell data. Moreover, recent work [10, 14, 31, 27, 34] in learning biologically meaningful representations of gene expression data could be incorporated to better understand the different sources of variation learned by the model. For example, using a constrained architecture such that latent variables correspond to gene pathways could shed more light on the biological phenomena captured in the different latent spaces.

Methods

The contrastiveVI model

Here we present the contrastiveVI model in more detail. We begin by describing the model's generative process and then the model's inference procedure.

The contrastiveVI generative process

For a target data point x_n we assume that each expression value x_{ng} for sample n and gene g is generated through the following process:

$$
\begin{aligned}
z_n &\sim \text{Normal}(0, I) \\
t_n &\sim \text{Normal}(0, I) \\
\ell_n &\sim \text{log normal}(\ell_\mu, \ell_\sigma^2) \\
\rho_n &= f_w(z_n, t_n, s_n) \\
w_{ng} &\sim \text{Gamma}(\rho_{ng}, \theta_g) \\
y_{ng} &\sim \text{Poisson}(\ell_n w_{ng}) \\
h_{ng} &\sim \text{Bernoulli}\big(f_h^g(z_n, t_n, s_n)\big) \\
x_{ng} &= \begin{cases} y_{ng} & \text{if } h_{ng} = 0 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

In this process z_n and t_n both refer to sets of latent variables underlying variations in scRNA-seq expression data. Here z_n represents variables that are shared across background and target cells, while t_n represents variations unique to the target cells. We place a standard multivariate Gaussian prior on both sets of latent factors, as such a specification is computationally convenient for inference in the VAE framework [21]. To encourage the disentanglement of latent factors, for background data points b_n we assume the same generative process but instead set $t_n = \mathbf{0}$ to represent the absence of salient latent factors in the generative process. Categorical covariates such as experimental batches are represented by s_n.

ℓ_μ and $\ell_\sigma \in \mathbb{R}_+^B$, where B denotes the cardinality of the categorical covariate, parameterize the prior for latent RNA library size scaling factor on a log scale . For each category (e.g. experimental batch), ℓ_μ and ℓ_σ^2 are set to the empirical mean and variance of the log library size. The gamma distribution is parameterized by the mean $\rho_{ng} \in \mathbb{R}_+$ and shape $\theta_g \in \mathbb{R}_+$. Furthermore, following the generative process, θ_g is equivalent to a gene-specific inverse dispersion parameter for a negative binomial distribution, and $\theta \in \mathbb{R}_+^G$ is estimated via variational Bayesian inference. f_w and f_g in the generative process are neural networks that transform the latent space and batch annotations to the original gene space, i.e.: $\mathbb{R}^d \times \{0, 1\}^B \to \mathbb{R}^G$, where d is the latent dimension. The network f_w is constrained during inference to encode the mean proportion of transcripts expressed across all genes by

using a softmax activation function in the last layer. That is, letting $f_w^g(z_n, t_n, s_n)$ denote the entry in the output of f_w corresponding to gene g, we have $\sum_g f_w^g(z_n, t_n, s_n) = 1$. The neural network f_h encodes whether a particular gene's expression has dropped out in a cell due to technical factors.

Our generative process closely follows that of scVI [23], with the addition of the salient latent factors t_n. While scVI's modeling approach has been shown to excel at many scRNA-seq analysis tasks, our empirical results demonstrate that it is not suited for contrastive analysis (CA). By dividing the RNA latent factors into shared factors z_n and target-specific factors t_n, contrastiveVI successfully isolates variations enriched in target datasets missed by previous methods. We depict the full contrastiveVI generative process as a graphical model in **Supplementary Figure 1.**

Inference with contrastiveVI

We cannot compute the contrastiveVI posterior distribution using Bayes' rule as the integrals required to compute the model evidence $p(x_n|s_n)$ are analytically intractable. As such, we instead approximate our posterior distribution using variational inference [5]. For target data points we approximate our posterior with a distribution factorized as follows:

$$q_{\phi_x}(z_n, t_n, \ell_n|x_n, s_n) = q_{\phi_z}(z_n|x_n, s_n)q_{\phi_t}(t_n|x_n, s_n)q_{\phi_\ell}(\ell_n|x_n, s_n). \quad (1)$$

Here ϕ_x denotes a set of learned weights used to infer the parameters of our approximate posterior. Based on our factorization, we can divide ϕ_x into three disjoint sets ϕ_z, ϕ_t and ϕ_ℓ for inferring the parameters of the distributions of z, t and ℓ respectively. Following the VAE framework [21], we then approximate the posterior for each factor as a deep neural network that takes in expression levels as input and outputs the parameters of its corresponding approximate posterior distribution (e.g. mean and variance). Moreover, we note that each factor in the posterior approximation shares the same family as its respective prior distribution (e.g. $q(z_n|x_n, s_n)$ follows a normal distribution). We can simplify our likelihood by integrating out w_{ng}, h_{ng}, and y_{ng}, yielding $p_\nu(x_{ng}|z_n, t_n, s_n, \ell_n)$, which follows a zero-inflated negative binomial (ZINB) distribution (**Supplementary Note 1**) and where ν denotes the parameters of our generative model. As with our approximate posteriors, we realize our generative model as a deep neural network. For Equation 1 we can derive (**Supplementary Note 2**) a corresponding variational lower bound:

$$p(x|s) \geq \mathbb{E}_{q(z,t,\ell|x,s)} \log p(x|z,t,\ell,s) - D_{KL}(q(z|x,s)||p(z)) - D_{KL}(q(t|x,s)||p(t)) - D_{KL}(q(\ell|x,s)||p(\ell|s)). \quad (2)$$

Next, for background data points we approximate the posterior using the factorization:

$$q_{\phi_b}(z_n, \ell_n|b_n, s_n) = q_{\phi_z}(z_n|b_n, s_n) q_{\phi_\ell}(\ell_n|b_n, s_n), \quad (3)$$

where ϕ_b denotes a set of learned parameters use to infer the values of z_n and ℓ_n for background samples. Following our factorization, we divide ϕ_b into the disjoint sets ϕ_z and ϕ_ℓ. We note that ϕ_z and ϕ_ℓ are shared across target and background samples; this encourages the posterior distributions q_{ϕ_z} and q_{ϕ_ℓ} to capture variations shared across the datasets, while q_{ϕ_t} captures variations unique to the target data. Once again we can simplify our likelihood by integrating out w_{ng}, h_{ng}, and y_{ng} to obtain $p_\nu(x_{ng}|z_n, \mathbf{0}, s_n, \ell_n)$, which follows a ZINB distribution. We similarly note that the parameters of our generative model ν are shared across target and background points to encourag z to capture shared variations across target and background points while t captures target-specific variations. We then have the following variational lower bound for our background data points:

$$p(b|s) \geq \mathbb{E}_{q(z,\ell|x,s)} \log p(b|z,\ell,s) - D_{KL}(q(z|b,s)||p(z)) - D_{KL}(q(\ell|b,s)||p(\ell|s)). \quad (4)$$

We then jointly optimize the parameters of our generative model and inference networks using stochastic gradient descent to maximize the sum of these two bounds over our background and target data points. All neural networks used to implement the variational and generative distributions were feedforward and used standard activation functions. We used the same network architecture and hyperparameter values for all experiments, and we refer the reader to **Supplementary Note 3** for more details.

Differential gene expression analysis with contrastiveVI

For two cell groups $A = (a_1, a_2, ..., a_n)$ and $B = (b_1, b_2, ..., b_m)$ in the target dataset, the posterior probability of gene g being differentially expressed in the two groups can be obtained as proposed by Boyeau et al. [6]. For any arbitrary cell pair a_i, b_j, we have two mutually exclusive models

$$\mathcal{M}_1^g : |r^g_{a_i,b_j}| > \delta \text{ and } \mathcal{M}_0^g : |r^g_{a_i,b_j}| \leq \delta$$

where $r^g_{a_i,b_j} \coloneqq \log_2(\rho^g_{a_i}) - \log_2(\rho^g_{b_j})$ is the log fold change of the denoised, library size-normalized expression of gene g, and δ is a pre-defined threshold for log fold change magnitude to be considered biologically meaningful. The posterior probability of differential expression is therefore expressed as $p(\mathcal{M}^g_1|x_{a_i}, x_{b_j})$, which can be obtained via marginalization of the latent variables and categorical covariates:

$$p(\mathcal{M}^g_1|x_{a_i}, x_{b_j}) = \sum_s \int_{z_{a_i},t_{a_i},z_{b_j},t_{b_j}} p(\mathcal{M}^g_1|z_{a_i}, t_{a_i}, z_{b_j}, t_{b_j})p(s)dp(z_{a_i}, t_{a_i}|x_{a_i}, s)dp(z_{b_j}, t_{b_j}|x_{b_j}, s),$$

where $p(s)$ is the relative abundance of target cells in category s, and the integral can be computed via Monte Carlo sampling using the variational posteriors q_{ϕ_z}, q_{ϕ_t}. Finally, the group-level posterior probability of differential expression is

$$\int_{a,b} p(\mathcal{M}^g_1|x_a, x_b)dp(a)dp(b),$$

where, assuming that the cells are independent, $a \sim \mathcal{U}(a_1, ..., a_m)$ and $b \sim \mathcal{U}(b_1, ..., b_m)$. Computationally, this quantity can be estimated by a large random samples of pairs from the cell group A and B. In our experiments, 10,000 cell pairs were sampled, 100 Monte Carlo samples were obtained from the variational posteriors for each cell, and the δ threshold was set to 0.25, which is the default value recommended by the scvi-tools Python library [11]. Genes with group-level posterior probability of differential expression greater than 0.95 were considered for downstream pathway enrichment analysis.

Pathway enrichment analysis

Pathway enrichment analysis refers to a computational procedure for determining whether a predefined set of genes (i.e., a gene pathway) have statistically significant differences in expression between two biological states. Many tools exist for performing pathway enrichment analysis (see [19] for a review). In our analyses we use Enrichr [8], a pathway analysis tool for non-ranked gene lists based on Fisher's exact test, to find enriched pathways from the KEGG 2016 pathways database [18]. Specifically, the Enrichr wrapper implemented in the open-source GSEAPy[1] Python library was used for our analyses. Pathways enriched at false discovery rate smaller than 0.05 (adjusted by the Benjamini-Hochberg procedure [4]) are reported in this study.

[1] `https://gseapy.readthedocs.io/en/latest/`

Baseline models

Because the choice of library size normalization method tends to drastically impact dimension reduction and subsequent clustering results of methods not designed for modeling library sizes [30], we consider CA methods specifically tailored for scRNA-seq count data as baselines in this study. To our knowledge, CPLVM (contrastive Poisson latent variable model) and CGLVM (contrastive generalized latent variable model) are the only CA methods that explicitly model count-based scRNA-seq normalization [17]. We present a summary of previous work in CA in **Supplementary Table 4**. We also consider scVI, a deep generative model for UMI count data that takes batch effect, technical dropout, and varying library size into modeling considerations [23], to illustrate the need for models specifically designed for CA. Below we describe the CA methods CPLVM and CGLVM in more detail.

In CPLVM, variations shared between background and target condition are assumed to be captured by two sets of latent variables $\{z_i^b\}_{i=1}^n$ and $\{z_j^t\}_{j=1}^m$, and target condition-specific variations are described by latent variables $\{t_j\}_{j=1}^m$, where n, m are the number of background and target cells, respectively. Library size differences between the two conditions are modeled by $\{\alpha_i^b\}_{i=1}^n$ and $\{\alpha_j^t\}_{j=1}^m$, whereas gene-specific library sizes are parameterized by $\delta \in \mathbb{R}_+^G$, where G is the number of genes. Each data point is considered Poisson distributed, with rate parameter determined by $\alpha_i^b \delta \odot (S^\top z_i^b)$ for a background cell i and by $\alpha_j^t \delta \odot (S^\top z_j^t + W^\top t_j)$ for a target cell j, where S, W are model weights that linearly combine the latent variables, and $\odot$ represents an element-wise product. The model weights and latent variables are assumed to have Gamma priors, δ has a standard log-normal prior, and α_i^b, α_j^t have log-normal priors with parameters given by the empirical mean and variance of log total counts in each dataset. Posterior distributions are fitted using variational inference with mean-field approximation and log-normal variational distributions.

The CA modeling approaches of CGLVM and CPLVM are similar. In CGLVM, however, the relationships of latent factors are considered additive and relate to the Poisson rate parameter via an exponential link function (similar to a generalized linear modeling scheme). All the priors and variational distributions are Gaussian in CGLVM.

Model optimization details

For all datasets, contrastiveVI models were trained with 80% of the background and target data, and with 20% of the data reserved as a validation set for early stopping to determine the number of training epochs needed. Training was early stopped when the validation variational lower bound showed no improvement for 45 epochs, typically resulting in 127 to 500 epochs of training. All contrastiveVI models were trained with the Adam optimizer [20]

with $\varepsilon = 0.01$, learning rate at 0.001, and weight decay at 10^{-6}. The same hyperparameters and training scheme were used to optimize the scVI models using only target data, usually with 274 to 500 epochs of training based on the early stopping criterion. As in Jones et al., the CPLVMs were trained via variational inference using all background and target data for 2,000 epochs with the Adam optimizer with $\varepsilon = 10^{-8}$ and learning rate at 0.05, and the CGLVMs were similarly trained for 1,000 epochs and learning rate at 0.01 [17]. All models were trained with 10 salient and 10 background latent variables for five times with different random weight initializations. We also trained models with varying salient latent dimension sizes and obtained overall consistent results (**Supplementary Figure 2)**.

Datasets and preprocessing

Here we briefly describe all datasets used in this work along with any corresponding preprocessing steps. All preprocessing steps were performed using the Scanpy Python package [37]. All our code for downloading and preprocessing these datasets is publicly available at `https://github.com/suinleelab/contrastiveVI`. For all experiments we retained the top 2,000 most highly variable genes returned from the Scanpy `highly_variable_genes` function with the `flavor` parameter set to `seurat_v3`. For all datasets, the number of cells in the background vs. target can be found in **Supplementary Table 3**.

Zheng et al., 2017

This dataset consists of single-cell RNA expression levels of a mixture of bone marrow mononuclear cells (BMMCs) from 10x Genomics [1]. For our target dataset, we use samples taken from patients with acute myeloid leukemia (AML) before and after a stem cell transplant. For our background dataset, we use measurements taken from two healthy control patients released as part of the same study. All data is publicly available: files containing measurements from first patient pre- and post-transplant can be found here and here, respectively; from the second patient pre- and post-transplant here and here, respectively; and from the two healthy control patients here and here.

Haber et al., 2017

This dataset (Gene Expression Omnibus accession number GSE92332) used scRNA-seq measurements to investigate the responses of intestinal epithelial cells in mice to different pathogens. In particular, in this dataset responses to *Salmonella* and the parasite *H. polygyrus* were investigated. Here our target dataset consisted measurements of cells infected with *Salmonella* and from cells 10 days after being infected with *H. polygyrus*, while

our background consisted of measurements from healthy control cells released as part of the same study.

McFarland et al., 2020

This dataset measured cancer cell lines' transcriptional responses after being treated with various small-molecule therapies. For our target dataset, we used data from cells that were exposed to idasanutlin, and for our background we used data from cells that were exposed to a control solution of dimethyl sulfoxide (DMSO). *TP53* mutation status was determined by cross-referencing with a list of cell lines with mutations provided by the authors in the code repository accompanying the paper. The data was downloaded from the authors' Figshare repository.

Evaluation metrics

Here we describe the quantitative metrics used in this study. All metrics were computed using their corresponding implementations in the scikit-learn Python package [7].

Silhouette width

We calculate silhouette width using the latent representations returned by each method. For a given sample i, the sillhouete width $s(i)$ is defined as follows. Let $a(i)$ be the average distance between i and the other samples with the same ground truth label, and let $b(i)$ be the smallest average distance between i and all other samples with a different label. The silhouette score $s(i)$ is then

$$s(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)}.$$

A silhouette width close to one indicates that i is tightly clustered with cells with the same ground truth label, while a score close to -1 indicates that a cell has been grouped with cells with a different label.

Adjusted Rand index

The adjusted Rand index (ARI) measures agreement between reference clustering labels and labels assigned by a clustering algorithm. Given a set of n samples and two sets of clustering labels describing those cells, the overlap between clustering labels can be described using a contingency table, where each entry indicates the number of cells in common between the two sets of labels. Mathematically, the ARI is calculated as

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]\Big/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]\Big/\binom{n}{2}},$$

where n_{ij} is the number of cells assigned to cluster i based on the reference labels and cluster j based on a clustering algorithm, a_i is the number of cells assigned to cluster i in the reference set, and b_j is the number of cells assigned to cluster j by the clustering algorithm. ARI values close to 1 indicate agreement between the reference labels and labels assigned by a clustering algorithm.

Adjusted mutual information

The adjusted mutual information (AMI) is a corrected-for-chance version of the normalized mutual information, and it is another measure of the agreement between reference clustering labels and labels assigned by a clustering algorithm. For two clusterings U and V, we have

$$AMI(U,V) = \frac{I(U;V) - \mathbb{E}[I(U;V)]}{(H(U)+H(V))/2 - \mathbb{E}[I(U;V)]}$$

where I represents mutual information, and H represents entropy. AMI values closer to 1 indicate greater agreement between U and V.

References

[1] 10x Genomics. 10x genomics. support: single cell gene expression datasets. *https://support.10xgenomics.com/single-cell-gene-expression/datasets*, 2021.

[2] A. Abid and J. Zou. Contrastive variational autoencoder enhances salient features. *arXiv preprint arXiv:1902.04601*, 2019.

[3] A. Abid, M. J. Zhang, V. K. Bagaria, and J. Zou. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature Communications*, 9(1): 1–7, 2018.

[4] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

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

[6] P. Boyeau, R. Lopez, J. Regier, A. Gayoso, M. I. Jordan, and N. Yosef. Deep generative models for detecting differential expression in single cells. *Machine Learning in Computational Biology (MLCB)*, October 2019.

[7] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, et al. Api design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*, 2013.

[8] E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, and A. Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14(1):1–14, 2013.

[9] A. Dixit, O. Parnas, B. Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, 167(7):1853–1866, 2016.

[10] N. Fortelny and C. Bock. Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. *Genome biology*, 21(1):1–36, 2020.

[11] A. Gayoso, R. Lopez, G. Xing, P. Boyeau, K. Wu, M. Jayasuriya, E. Melhman, M. Langevin, Y. Liu, J. Samaran, G. Misrachi, A. Nazaret, O. Clivio, C. Xu, T. Ashuach, M. agha Lotfollahi, V. Svensson, E. da Veiga Beltrame, C. Talavera-López, L. Pachter, F. J. Theis, A. M. Streets, M. I. Jordan, J. Regier, and N. Yosef. scvi-tools: a library for deep probabilistic analysis of single-cell omics data. *bioRxiv*, 2021.

[12] A. Gayoso, Z. Steier, R. Lopez, J. Regier, K. L. Nazor, A. Streets, and N. Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature Methods*, 18 (3):272–282, 2021.

[13] A. Grubman, G. Chew, J. F. Ouyang, G. Sun, X. Y. Choo, C. McLean, R. K. Simmons, S. Buckberry, D. B. Vargas-Landin, D. Poppe, et al. A single-cell atlas of entorhinal cortex from individuals with alzheimer's disease reveals cell-type-specific gene expression regulation. *Nature neuroscience*, 22(12):2087–2097, 2019.

[14] G. Gut, S. G. Stark, G. Rätsch, and N. R. Davidson. Pmvae: Learning interpretable single-cell representations with pathway modules. *bioRxiv*, 2021.

[15] A. L. Haber, M. Biton, N. Rogel, R. H. Herbst, K. Shekhar, C. Smillie, G. Burgin, T. M. Delorey, M. R. Howitt, Y. Katz, et al. A single-cell survey of the small intestinal epithelium. *Nature*, 551(7680):333–339, 2017.

[16] Y. Hao, S. Hao, E. Andersen-Nissen, W. M. Mauck III, S. Zheng, A. Butler, M. J. Lee, A. J. Wilk, C. Darby, M. Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 2021.

[17] A. Jones, F. W. Townes, D. Li, and B. E. Engelhardt. Contrastive latent variable modeling with application to case-control sequencing experiments. *arXiv preprint arXiv:2102.06731*, 2021.

[18] M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[19] P. Khatri, M. Sirota, and A. J. Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Computational Biology*, 8(2):e1002375, 2012.

[20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[21] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[22] D. Li, A. Jones, and B. Engelhardt. Probabilistic contrastive principal component analysis. *arXiv preprint arXiv:2012.07977*, 2020.

[23] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, 2018.

[24] M. Lotfollahi, F. A. Wolf, and F. J. Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.

[25] M. Lotfollahi, A. Klimovskaia, C. De Donno, Y. Ji, I. L. Ibarra, F. A. Wolf, N. Yakubova, F. J. Theis, and D. Lopez-Paz. Learning interpretable cellular responses to complex perturbations in high-throughput screens. *bioRxiv*, 2021.

[26] M. Lotfollahi, M. Naghipourfar, M. D. Luecken, M. Khajavi, M. Büttner, M. Wagenstetter, Ž. Avsec, A. Gayoso, N. Yosef, M. Interlandi, et al. Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, pages 1–10, 2021.

[27] W. Mao, E. Zaslavsky, B. M. Hartmann, S. C. Sealfon, and M. Chikina. Pathway-level information extractor (plier) for gene expression data. *Nature methods*, 16(7):607–610, 2019.

[28] H. Mathys, J. Davila-Velderrain, Z. Peng, F. Gao, S. Mohammadi, J. Z. Young, M. Menon, L. He, F. Abdurrob, X. Jiang, et al. Single-cell transcriptomic analysis of alzheimer's disease. *Nature*, 570(7761):332–337, 2019.

[29] J. M. McFarland, B. R. Paolella, A. Warren, K. Geiger-Schuller, T. Shibue, M. Rothberg, O. Kuksenko, W. N. Colgan, A. Jones, E. Chambers, et al. Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nature Communications*, 11(1):1–15, 2020.

[30] D. Risso, F. Perraudeau, S. Gribkova, S. Dudoit, and J.-P. Vert. A general and flexible method for signal extraction from single-cell rna-seq data. *Nature Communications*, 9 (1):1–17, 2018.

[31] S. Rybakov, M. Lotfollahi, F. J. Theis, and F. A. Wolf. Learning interpretable latent autoencoder representations with annotations of feature sets. *bioRxiv*, 2020.

[32] K. A. Severson, S. Ghosh, and K. Ng. Unsupervised learning with contrastive latent variable models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4862–4869, 2019.

[33] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019.

[34] V. Svensson, A. Gayoso, N. Yosef, and L. Pachter. Interpretable factor models of single-cell rna-seq via variational autoencoders. *Bioinformatics*, 36(11):3418–3421, 2020.

[35] L. T. Vassilev, B. T. Vu, B. Graves, D. Carvajal, F. Podlaski, Z. Filipovic, N. Kong, U. Kammlott, C. Lukacs, C. Klein, N. Fotouhi, and E. A. Liu. In vivo activation of the p53 pathway by small-molecule antagonists of mdm2. *Science*, 303(5659):844–848, 2004. doi: 10.1126/science.1092472.

[36] A. J. Wilk, A. Rustagi, N. Q. Zhao, J. Roque, G. J. Martínez-Colón, J. L. McKechnie, G. T. Ivison, T. Ranganath, R. Vergara, T. Hollis, et al. A single-cell atlas of the peripheral immune response in patients with severe covid-19. *Nature Medicine*, 26(7): 1070–1076, 2020.

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

[38] F. Wu, J. Fan, Y. He, A. Xiong, J. Yu, Y. Li, Y. Zhang, W. Zhao, F. Zhou, W. Li, et al. Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nature Communications*, 12(1):1–11, 2021.

[39] G. X. Y. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049, 2017.

[40] J. Y. Zou, D. J. Hsu, D. C. Parkes, and R. P. Adams. Contrastive learning using spectral methods. *Advances in Neural Information Processing Systems*, 26:2238–2246, 2013.