ISOLATING SALIENT VARIATIONS OF INTEREST IN SINGLE-CELL TRANSCRIPTOMIC DATA WITH CONTRASTIVEVI

Ethan Weinberger[*], **Chris Lin**[*] **& Su-In Lee**
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98195, USA
{ewein,clin25,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 four target-background dataset pairs, we apply contrastiveVI to perform a number of standard analysis tasks, including visualization, clustering, and differential expression testing, and we consistently achieve results that agree with known biological ground truths.

1 INTRODUCTION

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 (Wu et al., 2021), Alzheimer's disease (Grubman et al., 2019; Mathys et al., 2019), and COVID-19 (Wilk et al., 2020). In many settings, scRNA-seq data analysts are specifically interested in patterns enriched in one dataset, referred to as the *target*, as compared to a second related dataset, referred to as the *background*. 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- versus post-intervention groups form intuitive background and target pairs. Moreover, with the development of new technologies for measuring cellular responses to large numbers of perturbations in parallel, such as Perturb-Seq (Dixit et al., 2016) and MIX-Seq (McFarland et al., 2020), 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) (Zou et al., 2013; Abid et al., 2018; Jones et al., 2021; Li et al., 2020; Severson et al., 2019; Abid & Zou, 2019). While many recent studies have modeled scRNA-seq data by fitting probabilistic models and representing the data in a lower dimension (Lopez et al., 2018; Risso et al., 2018; Hao et al., 2021; Lotfollahi et al., 2021; 2019), few of these models are 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 (Abid et al., 2018). One recent study (Jones et al., 2021) designed a probabilistic model for analyzing scRNA-seq data in the CA setting. However, this method assumes that a linear model is sufficiently expressive to model the variations in scRNA-seq data,

[*] denotes equal contribution

even though previous work has demonstrated substantial improvements by using more expressive nonlinear methods (Lopez et al., 2018).

To address these limitations, we developed contrastiveVI, a deep generative model that enables analysis of scRNA-seq data in the CA setting. 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. Moreover, similar to previous work (Lopez et al., 2018), the full contrastiveVI probabilistic model accounts for the specific technical biases and noise characteristics of scRNA-seq data. contrastiveVI can be used for a number of analysis tasks, including dimensionality reduction, clustering, and differential gene expression testing. To highlight this functionality, we applied contrastiveVI to four publicly available background and target scRNA-seq dataset pairs and demonstrated strong performance on all of them compared to previously proposed methods.

2 THE CONTRASTIVEVI MODEL

contrastiveVI uses a probabilistic latent variable model to represent 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 whether each cell belongs to the background or target dataset (**Fig. 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.

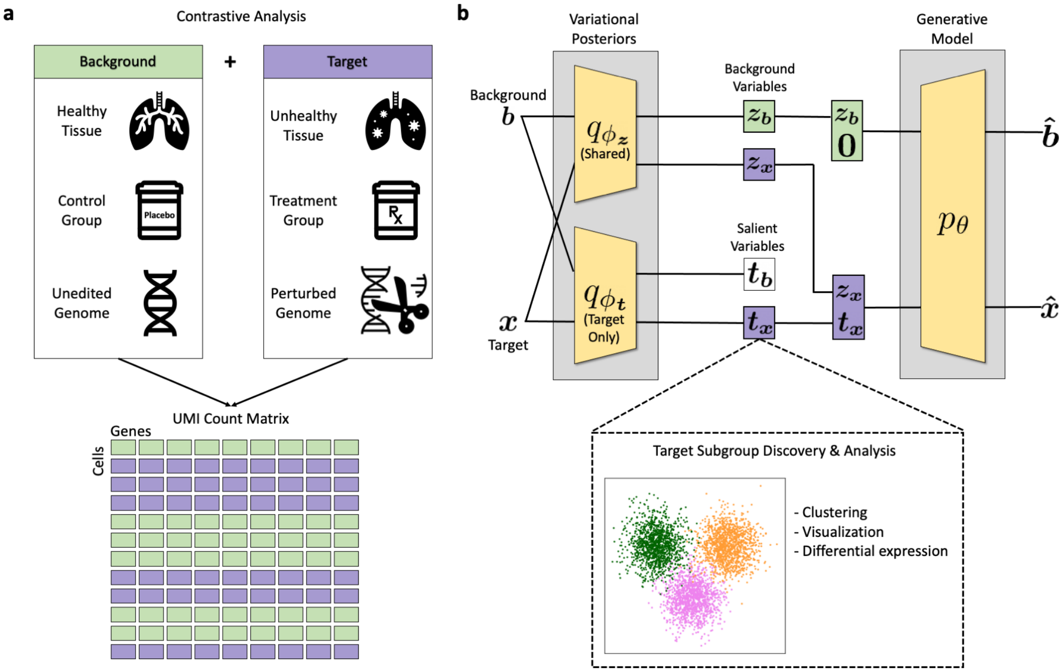


Figure 1: **Overview of contrastiveVI.** Given a reference background dataset and a second target dataset of interest, contrastiveVI separates the variations shared between the two datasets and the variations enriched in the target dataset. **a**, Example background and target data pairs. Samples from 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 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 assigned salient latent variable values; background data points are instead assigned a zero vector for the salient variables to represent their absence. 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 (**Supplementary Fig. 1)**. All distributions are parameterized by neural networks.

The contrastiveVI model is based on the variational autoencoder (VAE) framework (Kingma & Welling, 2013). 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 samples 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 **Appendix A**.

3 Results

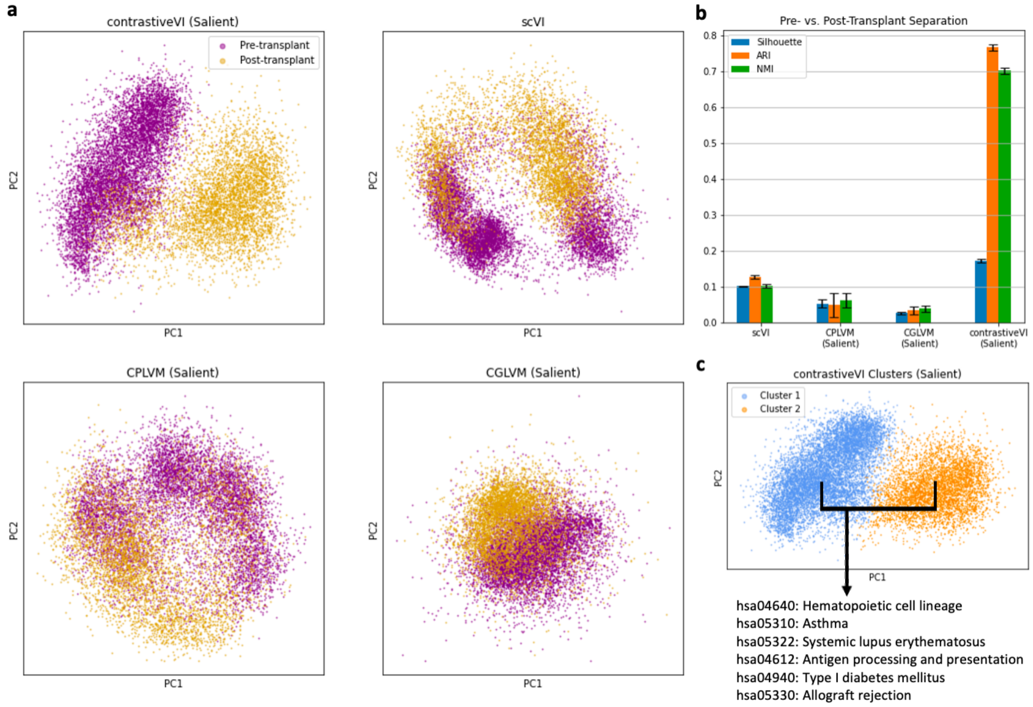
To evaluate contrastiveVI's performance, we rely on datasets with known ground truth variations in the target condition that are not present in the background condition. We benchmarked contrastiveVI's performance against that of three previously proposed methods for analyzing raw scRNA-seq count data. First, to demonstrate that our contrastive approach is necessary for isolating enriched variations in target datasets, we compared against scVI (Lopez et al., 2018). 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 capture salient variations in target samples. 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) (Jones et al., 2021). 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, CPLVM and CGLVM are the only existing contrastive methods for analyzing scRNA-seq count data.

3.1 Cancer treatment response

We first evaluated contrastiveVI on expression data from bone marrow mononuclear cells (BMMCs) from two patients with acute myeloid leukemia (AML). 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 (Zheng et al., 2017). 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. For background data we used measurements from two healthy control patients collected as part of the same study.

Qualitatively, pre- and post-transplant cells are well separated in the salient latent space learned by contrastiveVI (**Fig. 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 normalized mutual information (NMI; **Appendix F**). We find that contrastiveVI performs well on all of these metrics (**Fig. 2b**), indicating that it successfully recovers the variations enriched in the target dataset. On the other hand, we find qualitatively that none of the baseline models separate pre- and post-transplant cells as well as contrastiveVI. This finding is confirmed by our quantitative results (**Fig. 2b**). Across all three metrics, we find that contrastiveVI significantly outperforms baseline models, with especially large gains in the ARI and NMI. These results indicate that contrastiveVI recovered the variations enriched in the AML patient data far better than baseline models.

We also used this dataset to demonstrate 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 (**Fig. 2c**). As demonstrated by our quantitative results, the resulting two clusters exhibit strong agreement with the two ground-truth groups (ARI: 0.77 ± 0.01). To better understand the underlying biological phenomena associated with this separation, we obtained differentially expressed genes across the two clusters using Monte Carlo sampling of denoised, library size-normalized expressions from the contrastiveVI decoder (**Appendix A.3**). Pathway enrichment analysis (**Appendix B**) was then performed with these differentially expressed genes using the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database (Kanehisa & Goto, 2000). We find that the pathways enriched with the differentially expressed genes are related to immune response and graft rejection, aligning with known cellular state transitions of BMMCs before and after a transplant. We provide a full list of enriched pathways in **Supplementary Table 1**. These results illustrate how contrastiveVI can facilitate better understanding of variations specific to scRNA-seq target datasets.



hsa04640: Hematopoietic cell lineage
hsa05310: Asthma
hsa05322: Systemic lupus erythematosus
hsa04612: Antigen processing and presentation
hsa04940: Type I diabetes mellitus
hsa05330: Allograft rejection

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- vs. post-transplant cells, ARI is the adjusted Rand index, and NMI is the normalized 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, and pathway enrichment analysis was then performed on the differentially expressed genes between the two clusters.

3.2 Infection response

We also applied contrastiveVI to data collected in Haber et al. (2017). This dataset consists of gene expression measurements of intestinal epithelial cells from mice infected with either *Salmonella enterica* (*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 differences between cell types. We present our results in Figure 3.

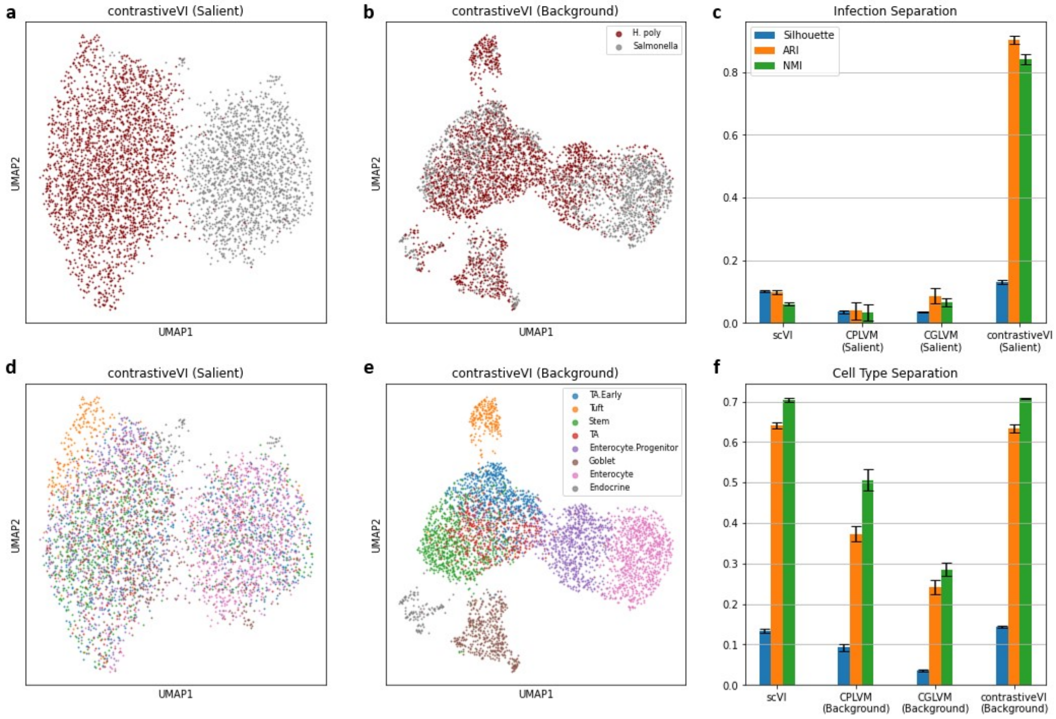


Figure 3: **contrastiveVI isolates responses to different infections in mouse intestinal epithelial cells. a,b,** UMAP plots of contrastiveVI's salient and background representations colored by infection type. Cells are correctly separated by infection type in the salient space, while they mix across 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 cells by infection type in its salient latent space (**Fig. 3a**). Moreover, we find that cells mix across infection types in the contrastiveVI background latent space as expected (**Fig. 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 (**Fig. 3c**).

Similar to our analysis for the BMMC dataset collected by Zheng et al. (2017), we applied k-means clustering to split the infected mouse epithelial cells into two groups based on their contrastiveVI salient latent embeddings and then used contrastiveVI to identify differentially expressed genes. We identified enriched KEGG pathways related to fat, cholesterol, and vitamin metabolism with the list of differentially expressed genes (**Supplementary Table 2**). These enriched pathways are consistent with previous findings that lipids and lipoproteins partake in innate immunity (Sviridov & Bukrinsky, 2014; Khovidhunkit et al., 2004) and that vitamins can alleviate or prevent infections (White, 2008; Hemilä, 2017). Particularly, it has been shown that active vitamin D may enhance the clearance of *Salmonella* via autophagy (Huang, 2016). Furthermore, six of the ten differentially expressed genes in the enriched pathways were found to have pathogen-specific expression in Haber et al. (2017) (e.g. *Apoc2* and *Fabp1*), while the other four genes belong to the same families as differentially expressed genes specific to *Salmonella* or *H. polygyrus* (e.g. *Apoc3* and *Fabp2*). These results show that contrastiveVI can be used to identify and interpret biologically relevant subgroups in target data.

For this dataset we further validated contrastiveVI's ability to disentangle target and background variations using ground truth cell type labels provided by Haber et al. (2017). In particular, we found strong mixing across cell types in contrastiveVI's salient latent space (**Fig. 3d**). This result agrees with the analysis in Haber et al. (2017), which found that responses to the two pathogens were mostly cell-type agnostic. On the other hand, cell types separated clearly in the background latent space (**Fig. 3e**). This result also agrees with prior biological knowledge, as we would expect the underlying factors of variation that distinguish cell types to be shared across healthy and infected cells. We quantified the degree of this cell-type separation in contrastiveVI's background latent space using our set of clustering metrics (**Fig. 3f**). We find that contrastiveVI's background latent space is far better at capturing differences between cell types than previously proposed contrastive methods' background latent spaces. Moreover, we find that contrastiveVI's background latent space separates cell types to a similar degree as the non-contrastive scVI's latent space.

Taken together, these results demonstrate that contrastiveVI successfully disentangles variations enriched in target data from shared variations, even when other methods struggle.

3.3 Small-molecule therapy response

We next applied contrastiveVI to a dataset collected using the recently developed MIX-Seq (McFarland et al., 2020) platform. MIX-Seq measures the transcriptional responses of up to hundreds of cancer cell lines in parallel after being treated with one or more small molecule compounds. Here our target dataset contains measurements from 24 cell lines treated with idasanutlin collected by McFarland et al. (2020). The small molecule idasanutlin is an antagonist of *MDM2*, a negative regulator of the tumor suppresor protein p53, hence offering cancer therapeutic opportunities (Vassilev et al., 2004). Based on the mechanism of action of idasanutlin, activation of the p53 pathway is observed in cell lines with wild type *TP53* and not in transcriptionally inactive mutant *TP53* cell lines (Vassilev et al., 2004). Our goal is thus to separate target cells by *TP53* mutation status. As the background dataset, we use measurements from the same cell lines treated with the control compound dimethyl sulfoxide (DMSO).

Qualitatively, contrastiveVI's salient latent space stratifies cells based on *TP53* mutation status (**Fig. 4a**). Our quantitative metrics also indicate that contrastiveVI separates the two classes of target cells more clearly than baseline methods (**Fig. 4b**). Moreover, we find that the clusters identified by applying k-means clustering to the contrastiveVI salient latent space have differentially expressed genes enriched in the p53 signaling pathway (**Fig. 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. These results demonstrate that contrastiveVI captures salient variations in the target samples treated with idasanutlin that specifically relate to the ground truth mechanism of idasanutlin perturbation.

We further evaluated contrastiveVI's performance on this dataset by embedding all cells, whether treated with DMSO or idasanutlin, into the model's background latent space. Ideally, contrastiveVI's background latent space would only capture variations that distinguish cell lines and not those related to treatment response. In particular, we would expect strong mixing between DMSO- and idasanutlin-treated cells even for cell lines with wild type *TP53*. We find that wild type *TP53* cell lines clearly separate by treatment type in the original data (**Supplementary Fig. 3**), whereas cells mix more strongly across treatment types (**Fig. 4d**) regardless of *TP53* mutation status (**Fig. 4e**) and instead separate primarily by cell line in the contrastiveVI background latent space (**Fig. 4f**). These results futher illustrate contrastiveVI's ability to disentangle shared and target-data-specific variations.

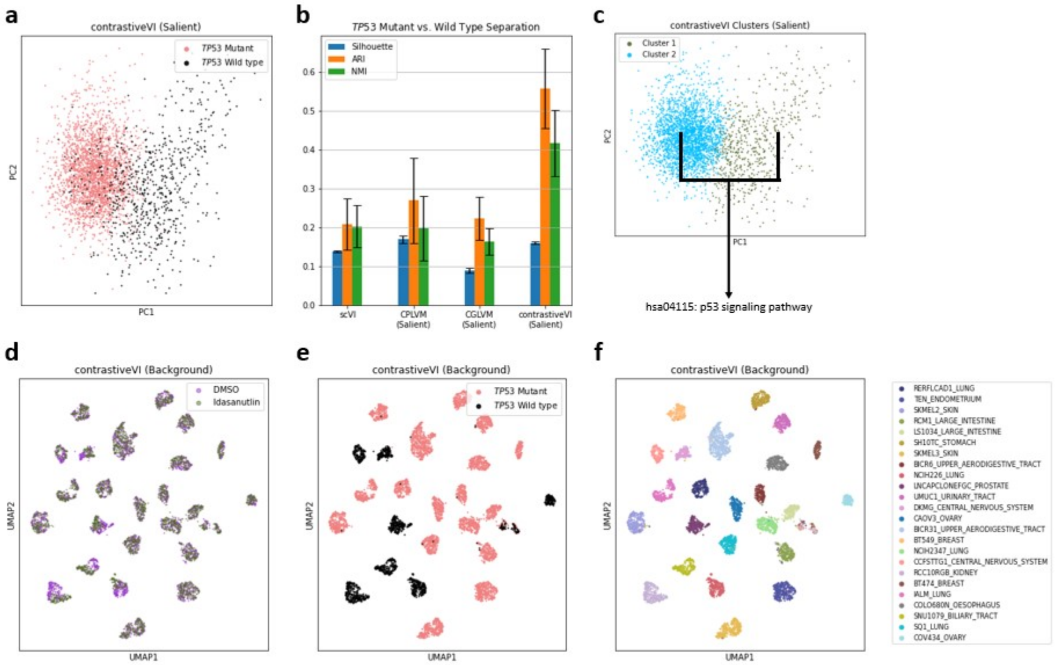


Figure 4: **contrastiveVI stratifies cancer cell lines by response to idasanutlin.** **a**, PC plot of contrastiveVI's salient latent representations for idasanutlin-treated cells from McFarland et al. (2020). **b**, The average silhouette width (silhouette), adjusted Rand Index (ARI) and normalized mutual information (NMI), with mean and standard error across five random trials plotted for each method. **c**, Two clusters identified by k-means clustering on contrastiveVI's salient latent representations of the idasanutlin-treated cells. Highly differentially expressed genes were identified from the two clusters, and these genes were used to perform pathway enrichment analysis. **d,e,f,** UMAP plots of contrastiveVI's background latent space colored by treatment type (**d**) *TP53* mutation status (**e**), and cell line (**f**).

3.4 CRISPR PERTURBATION RESPONSE

Finally, we applied contrastiveVI to data collected using the Perturb-Seq (Dixit et al., 2016; Adamson et al., 2016) platform. Perturb-Seq combines high-throughput scRNA-seq methods with barcoding of CRISPR-induced genomic perturbations, enabling the evaluation of such perturbations at single-cell resolution. Previous studies have successfully leveraged Perturb-Seq to better understand regulatory circuits related to innate immunity (Jaitin et al., 2016), the unfolded protein response pathway (Adamson et al., 2016), and the T cell receptor signaling pathway (Datlinger et al., 2017), among other applications. Despite these successes, recent work (Jones et al., 2021) has suggested that naive approaches for analyzing Perturb-Seq data may fail to capture subtle perturbation-induced transcriptomic changes due to the presence of intercellular variations unrelated to the perturbations. Thus, methods for isolating variations unique to the perturbed cells may unlock new biological insights missed by previous approaches.

In this experiment we applied contrastiveVI to a Perturb-Seq dataset from Norman et al. (2019). In this study, the authors assessed the effects of 284 different CRISPR-mediated perturbations on the growth of K562 cells, where each perturbation induced the overexpression of a single gene or a pair of genes. Here we focus on a subset of these perturbations for which the authors provided labels indicating a known gene program induced by the perturbation. We would expect cells to separate by these gene programs; however, in the latent space of an scVI model we observed significant mixing between cells with different gene program labels (**Fig. 5a**).

On the other hand, using cells treated with control guides as a background dataset, we find qualitatively that contrastiveVI better separates cells by gene program in its salient latent space (**Fig. 5b**).

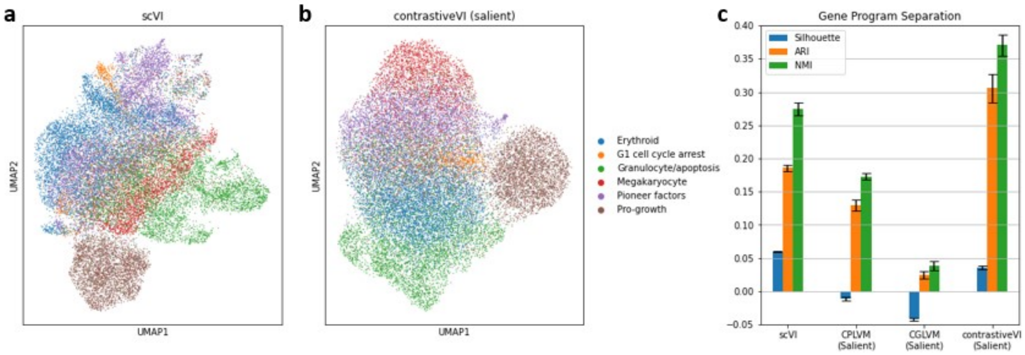


Figure 5: **contrastiveVI isolates CRISPR-perturbation-induced variations in a large-scale Perturb-Seq experiment. a,b**, UMAP plot of scVI's latent space (**a**) and contrastiveVI's salient latent space (**b**) colored by gene program. **c**, Clustering metrics quantifying how well cells separate by gene program label, with means and standard errors across five trials plotted.

Furthermore, we find that the relative positions of the different gene programs in the contrastiveVI salient latent space align with prior biological knowledge. For example, the group of cells labeled as overexpressing a pioneer-factor-related gene program abuts the groups of cells labelled as expressing erythroid and megakaryocyte gene programs. This positioning agrees with prior biological knowledge, as pioneer factors are known to play a role in cell type differentiation for these two cell types (Zaret & Carroll, 2011; Visvader et al., 1992; Kulessa et al., 1995). Similarly, pioneer factors have also been implicated in G1 cell cycle arrest (Zaret & Carroll, 2011; Zhang et al., 2011), the only other gene program that neighbors pioneer factors in the contrastiveVI salient latent space.

With our suite of metrics, we find that contrastiveVI again outperforms baseline methods (**Fig. 5c**). However, we note that our clustering metric values for this dataset are lower than those for previous datasets, potentially indicating that expression differences induced by the single- or double-gene CRISPR perturbations are more subtle than the clear separations found in previous datasets.

4 Discussion

In this work we introduce contrastiveVI, a deep generative model that explicitly disentangles enriched variations in a target scRNA-seq dataset from those shared with a related background dataset. contrastiveVI is the first method designed to analyze scRNA-seq count data in the contrastive analysis setting that both directly models the technical factors of variation in scRNA-seq data and takes advantage of the expressive power of deep generative modeling. Moreover, contrastiveVI includes a number of other capabilities relevant to scRNA-seq analysis out of the box, such as differential expression testing.

In four different contexts—response to cancer treatment, infection by different pathogens, exposure to small-molecule drug perturbations, and genomic perturbation via CRISPR guides—we find that contrastiveVI successfully isolates enriched variations in target cells while previously proposed methods struggle. With the recent development of new sequencing technologies for efficiently measuring transcriptomic responses to many perturbations in parallel, 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 (Gayoso et al., 2021a) Python library, thereby enabling seamless interoperability with the Scanpy (Wolf et al., 2018) and Seurat (Stuart et al., 2019) software ecosystems.

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 (Gayoso et al., 2021b), to better understand variations enriched in target datasets across different modalities of single-cell data. Moreover, recent work (Fortelny & Bock, 2020; Gut et al., 2021; Rybakov et al., 2020; Mao et al., 2019; Svensson et al., 2020) in learning more interpretable 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 model's salient and background latent spaces.

REFERENCES

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

Abubakar Abid and James Zou. Contrastive variational autoencoder enhances salient features. *arXiv preprint arXiv:1902.04601*, 2019.

Abubakar Abid, Martin J Zhang, Vivek K Bagaria, and James Zou. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature Communications*, 9(1):1–7, 2018.

Britt Adamson, Thomas M Norman, Marco Jost, Min Y Cho, James K Nuñez, Yuwen Chen, Jacqueline E Villalta, Luke A Gilbert, Max A Horlbeck, Marco Y Hein, et al. A multiplexed single-cell crispr screening platform enables systematic dissection of the unfolded protein response. *Cell*, 167(7):1867–1882, 2016.

Yoav Benjamini and Yosef 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.

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

Pierre Boyeau, Romain Lopez, Jeffrey Regier, Adam Gayoso, Michael I. Jordan, and Nir Yosef. Deep generative models for detecting differential expression in single cells. *Machine Learning in Computational Biology (MLCB)*, October 2019.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, et al. Api design for machine learning software: experiences from the scikit-learn project. *arXiv preprint arXiv:1309.0238*, 2013.

Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14(1):1–14, 2013.

Paul Datlinger, André F Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, Linda C Schuster, Amelie Kuchler, Donat Alpar, and Christoph Bock. Pooled crispr screening with single-cell transcriptome readout. *Nature methods*, 14(3):297–301, 2017.

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *Cell*, 167(7): 1853–1866, 2016.

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

Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Katherine Wu, Michael Jayasuriya, Edouard Melhman, Maxime Langevin, Yining Liu, Jules Samaran, Gabriel Misrachi, Achille Nazaret, Oscar Clivio, Chenling Xu, Tal Ashuach, Mohammad agha Lotfollahi, Valentine Svensson, Eduardo da Veiga Beltrame, Carlos Talavera-López, Lior Pachter, Fabian J Theis, Aaron M. Streets, Michael I. Jordan, Jeffrey Regier, and Nir Yosef. scvi-tools: a library for deep probabilistic analysis of single-cell omics data. *bioRxiv*, 2021a.

Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature Methods*, 18(3):272–282, 2021b.

Alexandra Grubman, Gabriel Chew, John F Ouyang, Guizhi Sun, Xin Yi Choo, Catriona McLean, Rebecca K Simmons, Sam Buckberry, Dulce B Vargas-Landin, Daniel 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.

Gilles Gut, Stefan G Stark, Gunnar Rätsch, and Natalie R Davidson. Pmvae: Learning interpretable single-cell representations with pathway modules. *bioRxiv*, 2021.

Adam L Haber, Moshe Biton, Noga Rogel, Rebecca H Herbst, Karthik Shekhar, Christopher Smillie, Grace Burgin, Toni M Delorey, Michael R Howitt, Yarden Katz, et al. A single-cell survey of the small intestinal epithelium. *Nature*, 551(7680):333–339, 2017.

Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck III, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, et al. Integrated analysis of multimodal single-cell data. *Cell*, 2021.

Harri Hemilä. Vitamin c and infections. *Nutrients*, 9(4):339, 2017.

Fu-Chen Huang. Vitamin d differentially regulates salmonella-induced intestine epithelial autophagy and interleukin-1β expression. *World Journal of Gastroenterology*, 22(47):10353, 2016.

Diego Adhemar Jaitin, Assaf Weiner, Ido Yofe, David Lara-Astiaso, Hadas Keren-Shaul, Eyal David, Tomer Meir Salame, Amos Tanay, Alexander van Oudenaarden, and Ido Amit. Dissecting immune circuits by linking crispr-pooled screens with single-cell rna-seq. *Cell*, 167(7):1883–1896, 2016.

Andrew Jones, F William Townes, Didong Li, and Barbara E Engelhardt. Contrastive latent variable modeling with application to case-control sequencing experiments. *arXiv preprint arXiv:2102.06731*, 2021.

Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

Purvesh Khatri, Marina Sirota, and Atul J Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Computational Biology*, 8(2):e1002375, 2012.

Weerapan Khovidhunkit, Min-Sun Kim, Riaz A Memon, Judy K Shigenaga, Arthur H Moser, Kenneth R Feingold, and Carl Grunfeld. Thematic review series: The pathogenesis of atherosclerosis. effects of infection and inflammation on lipid and lipoprotein metabolism mechanisms and consequences to the host1. *Journal of lipid research*, 45(7):1169–1196, 2004.

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

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Holger Kulessa, Jonathan Frampton, and Thomas Graf. Gata-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, and erythroblasts. *Genes & development*, 9(10):1250–1262, 1995.

Didong Li, Andrew Jones, and Barbara Engelhardt. Probabilistic contrastive principal component analysis. *arXiv preprint arXiv:2012.07977*, 2020.

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

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.

Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, Adam Gayoso, Nir Yosef, Marta Interlandi, et al. Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, pp. 1–10, 2021.

Weiguang Mao, Elena Zaslavsky, Boris M Hartmann, Stuart C Sealfon, and Maria Chikina. Pathway-level information extractor (plier) for gene expression data. *Nature methods*, 16(7): 607–610, 2019.

Hansruedi Mathys, Jose Davila-Velderrain, Zhuyu Peng, Fan Gao, Shahin Mohammadi, Jennie Z Young, Madhvi Menon, Liang He, Fatema Abdurrob, Xueqiao Jiang, et al. Single-cell transcriptomic analysis of alzheimer's disease. *Nature*, 570(7761):332–337, 2019.

James M McFarland, Brenton R Paolella, Allison Warren, Kathryn Geiger-Schuller, Tsukasa Shibue, Michael Rothberg, Olena Kuksenko, William N Colgan, Andrew Jones, Emily 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.

Thomas M Norman, Max A Horlbeck, Joseph M Replogle, Y Ge Alex, Albert Xu, Marco Jost, Luke A Gilbert, and Jonathan S Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019.

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

Sergei Rybakov, Mohammad Lotfollahi, Fabian J Theis, and F Alexander Wolf. Learning interpretable latent autoencoder representations with annotations of feature sets. *bioRxiv*, 2020.

Kristen A Severson, Soumya Ghosh, and Kenney Ng. Unsupervised learning with contrastive latent variable models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4862–4869, 2019.

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

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

Dmitri Sviridov and Michael Bukrinsky. Interaction of pathogens with host cholesterol metabolism. *Current opinion in lipidology*, 25(5):333, 2014.

Lyubomir T. Vassilev, Binh T. Vu, Bradford Graves, Daisy Carvajal, Frank Podlaski, Zoran Filipovic, Norman Kong, Ursula Kammlott, Christine Lukacs, Christian Klein, Nader Fotouhi, and Emily 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.

Jane E Visvader, Andrew G Elefanty, Andreas Strasser, and Jerry M Adams. Gata-1 but not scl induces megakaryocytic differentiation in an early myeloid line. *The EMBO journal*, 11(12): 4557–4564, 1992.

John H White. Vitamin d signaling, infectious diseases, and regulation of innate immunity. *Infection and immunity*, 76(9):3837–3843, 2008.

Aaron J Wilk, Arjun Rustagi, Nancy Q Zhao, Jonasel Roque, Giovanny J Martínez-Colón, Julia L McKechnie, Geoffrey T Ivison, Thanmayi Ranganath, Rosemary Vergara, Taylor 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.

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

Fengying Wu, Jue Fan, Yayi He, Anwen Xiong, Jia Yu, Yixin Li, Yan Zhang, Wencheng Zhao, Fei Zhou, Wei 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.

Kenneth S Zaret and Jason S Carroll. Pioneer transcription factors: establishing competence for gene expression. *Genes & development*, 25(21):2227–2241, 2011.

Chunpeng Zhang, Liguo Wang, Dayong Wu, Hongyan Chen, Zhong Chen, Jennifer M Thomas-Ahner, Debra L Zynger, Jérôme Eeckhoute, Jindan Yu, Jun Luo, et al. Definition of a foxa1 cistrome that is crucial for g1 to s-phase cell-cycle transit in castration-resistant prostate cancer. *Cancer Research*, 71(21):6738–6748, 2011.

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, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049, 2017.

James Y Zou, Daniel J Hsu, David C Parkes, and Ryan P Adams. Contrastive learning using spectral methods. *Advances in Neural Information Processing Systems*, 26:2238–2246, 2013.

A FURTHER DETAILS ON THE CONTRASTIVEVI PROBABILISTIC 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.

A.1 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:

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

In this process z_n and t_n 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 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 (Kingma & Welling, 2013). 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 $\ell_\sigma^2 \in \mathbb{R}_+^B$, where B denotes the cardinality of the categorical covariate, parameterize the prior for latent RNA library size scaling factor on a log scale, and s_n is a B-dimensional one-hot vector encoding categorical covariate index. 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.: $\mathbb{R}^d \times \{0,1\}^B \rightarrow \mathbb{R}^G$, where d is the size of the concatenated salient and background latent spaces. 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 (Lopez et al., 2018), 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 Fig. 1.**

A.2 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 (Blei et al., 2017). 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 (Kingma & Welling, 2013), 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 (**Appendix G**) and where ν denotes the parameters of our generative model. As with our approximate posteriors, we realize our generative model with deep neural networks. For Equation 1 we can derive (**Appendix H**) a corresponding variational lower bound:

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

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 target and background cells, 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 encourage 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:

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

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 I** for more details.

A.3 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. (Boyeau et al., 2019). 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 $r^g_{a_i,b_j} := \log_2(\rho^g_{a_i}) - \log_2(\rho^g_{b_j})$ 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 $p(\mathcal{M}_1^g | x_{a_i}, x_{b_j})$, which can be obtained via marginalization of the latent variables and categorical covariates:

$$p(\mathcal{M}_1^g | x_{a_i}, x_{b_j}) = \sum_s \int_{z_{a_i}, t_{a_i}, z_{b_j}, t_{b_j}} p(\mathcal{M}_1^g | z_{a_i}, t_{a_i}, z_{b_j}, t_{b_j}) p(s) dp(z_{a_i}, t_{a_i} | x_{a_i}, s) dp(z_{b_j}, t_{b_j} | x_{b_j}, s),$$

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

$$\int_{a,b} p(\mathcal{M}_1^g | x_a, x_b) dp(a) dp(b),$$

where we assume that the cells a and b are independently sampled $a \sim \mathcal{U}(a_1, ..., a_m)$ and $b \sim \mathcal{U}(b_1, ..., b_m)$. Computationally, this quantity can be estimated by a large random sample 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 (Gayoso et al., 2021a). Genes with group-level posterior probability of differential expression greater than 0.95 were considered for downstream pathway enrichment analysis.

B 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 (Khatri et al., 2012) for a review). In our analyses we use Enrichr (Chen et al., 2013), a pathway analysis tool for non-ranked gene lists based on Fisher's exact test, to find enriched pathways from the KEGG pathway database (Kanehisa & Goto, 2000). 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 (Benjamini & Hochberg, 1995)—are reported in this study.

C Baseline models

Because the choice of library size normalization method has been shown to drastically impact dimension reduction and subsequent clustering results of methods not designed to explicitly model

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

library sizes (Risso et al., 2018), 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 (Jones et al., 2021). We present a summary of previous work in CA in **Supplementary Table 6**. 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 (Lopez et al., 2018), 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 the background and target conditions are assumed to be captured by the shared latent variable values $\{z_i^b\}_{i=1}^n$ and $\{z_j^t\}_{j=1}^m$, and target condition-specific variations are captured by the salient latent variable values $\{t_j\}_{j=1}^m$, 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.

D Model optimization details

For all datasets, contrastiveVI models were trained with 80% of the background and target data; the remaining 20% of the data was 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 (Kingma & Ba, 2014) 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 (Jones et al., 2021). All models were trained with 10 salient and 10 background latent variables five times with different random weight initializations. To understand the impact of the size of the salient latent space on model performance, we also trained models with varying salient latent dimension sizes and obtained overall consistent results (**Supplementary Fig. 2**).

E 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 (Wolf et al., 2018). 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 condition 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 (10x Genomics, 2021). 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 the 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. Specifically, in this dataset, responses to the bacterium *Salmonella* and the parasite *H. polygyrus* were investigated. Here our target dataset included 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. The number of cells of each cell type can be found in **Supplementary Table 4**.

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. The number of cells for each cell line can be found in **Supplementary Table 5**.

Norman et al., 2019

This dataset (Gene Expression Omnibus accession number GSE133344) measured the effects of 284 different CRISPR-mediated perturbations on K562 cells, where each perturbation induced the overexpression of a single gene or a pair of genes. As done in the analysis from Norman et al. (2019), we excluded cells with the perturbation label `NegCtrl1_NegCtrl0__NegCtrl1_NegCtrl0` from our analysis. For our background dataset we used all remaining unperturbed cells, and for our target dataset we used all perturbed cells that had a gene program label provided by the authors.

F 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 (Buitinck et al., 2013).

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 closer to 1 indicate stronger agreement between the reference labels and labels assigned by a clustering algorithm.

Normalized mutual information

The normalized mutual information (NMI) measures the agreement between reference clustering labels and labels assigned by a clustering algorithm. The NMI is calculated as

$$\text{NMI} = \frac{I(P;T)}{\sqrt{\mathbb{H}(P)\mathbb{H}(T)}},$$

where P and T denote empirical distributions for the predicted and true clusterings, I denotes mutual information, and $\mathbb{H}$ the Shannon entropy.

G Integrating out contrastiveVI's latent variables

We first show that if

$$\begin{aligned} w &\sim \text{Gamma}(\rho, \theta) \\ y|w &\sim \text{Poisson}(\ell w) \end{aligned}$$

where $\rho, \theta \in \mathbb{R}_+$ are the mean and shape parameter of the gamma distribution, respectively, and $\ell \in \mathbb{R}_+$, then y follows a negative binomial distribution. We note that our analysis closely follows that of Lopez et al. (2018) and Gayoso et al. (2021b); we include it again here for completeness.

$$\begin{aligned} p(y) &= \int p(y|w)p(w)dw \\ &= \int \frac{\ell^y w^y e^{-\ell w}}{\Gamma(y+1)} \frac{\left(\frac{\theta}{\rho}\right)^\theta w^{\theta-1} e^{-\theta w/\rho}}{\Gamma(\theta)} dw \\ &= \frac{\ell^y \left(\frac{\theta}{\rho}\right)^\theta}{\Gamma(y+1)\Gamma(\theta)} \int w^{y+\theta-1} e^{-\left(\ell+\frac{\theta}{\rho}\right)w} dw \\ &= \frac{\ell^y \left(\frac{\theta}{\rho}\right)^\theta}{\Gamma(y+1)\Gamma(\theta)} \frac{\Gamma(y+\theta)}{\left(\ell+\frac{\theta}{\rho}\right)^{y+\theta}} \\ &= \frac{\Gamma(y+\theta)}{\Gamma(y+1)\Gamma(\theta)} \left(\frac{\theta}{\ell\rho+\theta}\right)^\theta \left(\frac{\ell\rho}{\ell\rho+\theta}\right)^y. \end{aligned}$$

The integral in the third line is evaluated by observing that the integrand is the unnormalized probability density function of a gamma distribution. The final line is exactly the probability mass function of a negative binomial distribution with mean $\ell\rho$ and inverse dispersion θ.

Next, we can incorporate multiplication of y by zero as a mixture between a point mass at zero and the original distribution of y. This enables us to write the probability mass function of $p(x_{ng}|z_n, t_n, \ell_n, s_n)$ as

$$\begin{cases} p(x_{ng} = 0|v_n, \ell_n) = f_h^g(v_n) + (1 - f_h^g(v_n))\left(\dfrac{\theta_g}{\ell_n f_w^g(v_n) + \theta_g}\right)^{\theta_g} \\ p(x_{ng} = y|v_n, \ell_n) = (1 - f_h^g(v_n))\dfrac{\Gamma(y + \theta_g)}{\Gamma(y + 1)\Gamma(\theta_g)}\left(\dfrac{\theta_g}{\ell_n f_w^g(v_n) + \theta_g}\right)^{\theta_g}\left(\dfrac{\ell_n f_w^g(v_n)}{\ell_n f_w^g(v_n) + \theta_g}\right)^{y}, \end{cases}$$

where $v_n = \{z_n, t_n, s_n\}$, and $y \in \mathbb{N}^+$. Letting f_w encode the mean of w and f_h the probability of technical dropout, this is exactly the probability mass function of a zero-inflated negative binomial (ZINB) distribution.

H contrastiveVI evidence lower bound derivation

Here we derive the variational lower bounds for contrastiveVI presented in the main text. For a given target cell x the contrastiveVI generative model's joint likelihood function factorizes as follows

$$p(x, z, t, \ell|s) = p(x|z, t, \ell, s)p(\ell|s)p(z)p(t)$$

Next, in order to perform variational inference we define the variational posterior as

$$q(z, t, \ell|x, s) = q(z|x, s)q(t|x, s)q(\ell|x, s)$$

Then we have

$$\begin{aligned} \log p(x|s) &= \log \int p(x, z, t, \ell|s) dz dt d\ell \\ &= \log \int \frac{p(x, z, t, \ell|s)q(z, t, \ell|x, s)}{q(z, t, \ell|x, s)} dz dt d\ell \\ &\geq \int q(z, t, \ell|x, s) \log \frac{p(x, z, t, \ell|s)}{q(z, t, \ell|x, s)} dz dt d\ell \\ &= \int q(z, t, \ell|x, s) \log \frac{p(x|z, t, \ell, s)p(z, t, \ell|s)}{q(z, t, \ell|x, s)} dz dt d\ell \\ &= \int \left(q(z, t, \ell|x, s) \log p(x|z, t, \ell, s) + q(z, t, \ell|x, s) \log \frac{p(z, t, \ell|s)}{q(z, t, \ell|x, s)}\right) dz dt d\ell \\ &= \mathbb{E}_{q(z,t,\ell|x,s)}[\log p(x|z, t, \ell, s)] - D_{KL}(q(z, t, \ell|x, s) \,||\, p(z, t, \ell|s)) \\ &= \mathbb{E}_{q(z,t,\ell|x,s)}[\log p(x|z, t, \ell, s)] - D_{KL}(q(z|x, s) \,||\, p(z)) \\ &\quad - D_{KL}(q(t|x, s) \,||\, p(t)) - D_{KL}(q(\ell|x, s) \,||\, p(\ell|s)) \end{aligned}$$

where we use Jensen's inequality in the third step and the independence of z, t, and ℓ to decompose the KL divergence term in the last step. Next, for a background point b we assume our generative process factorizes as

$$p(b, z, \ell|s) = p(b|z, \ell, s)p(\ell|s)p(z),$$

with a corresponding variational posterior of

$$q(z, \ell|b, s) = q(z|b, s)q(\ell|b, s).$$

We then have

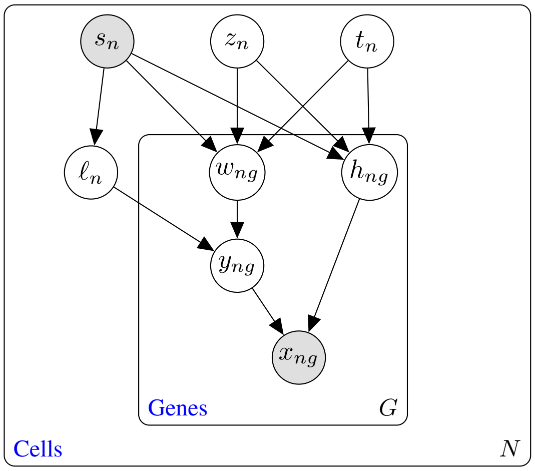
$$\begin{aligned}
\log p(b|s) &= \log \int p(b, z, \ell|s) dz d\ell \\
&= \log \int \frac{p(b, z, \ell|s) q(z, \ell|b, s)}{q(z, \ell|b, s)} dz d\ell \\
&\geq \int q(z, \ell|b, s) \log \frac{p(b, z, \ell|s)}{q(z, \ell|b, s)} dz d\ell \\
&= \int q(z, \ell|b, s) \log \frac{p(b|z, \ell, s) p(z, \ell|s)}{q(z, \ell|b, s)} dz d\ell \\
&= \int \left(q(z, \ell|b, s) \log p(b|z, \ell, s) + q(z, \ell|b, s) \log \frac{p(z, \ell|s)}{q(z, \ell|b, s)} \right) dz d\ell \\
&= \mathbb{E}_{q(z,\ell|b,s)}[\log p(b|z, \ell, s)] - D_{KL}(q(z, \ell|b, s) \,||\, p(z, \ell|s)) \\
&= \mathbb{E}_{q(z,\ell|b,s)}[\log p(b|z, \ell, s)] - D_{KL}(q(z, |b, s) \,||\, p(z)) - D_{KL}(q(\ell, |x, s) \,||\, p(\ell|s))
\end{aligned}$$

I FURTHER DETAILS ON THE CONTRASTIVEVI NETWORK ARCHITECTURE

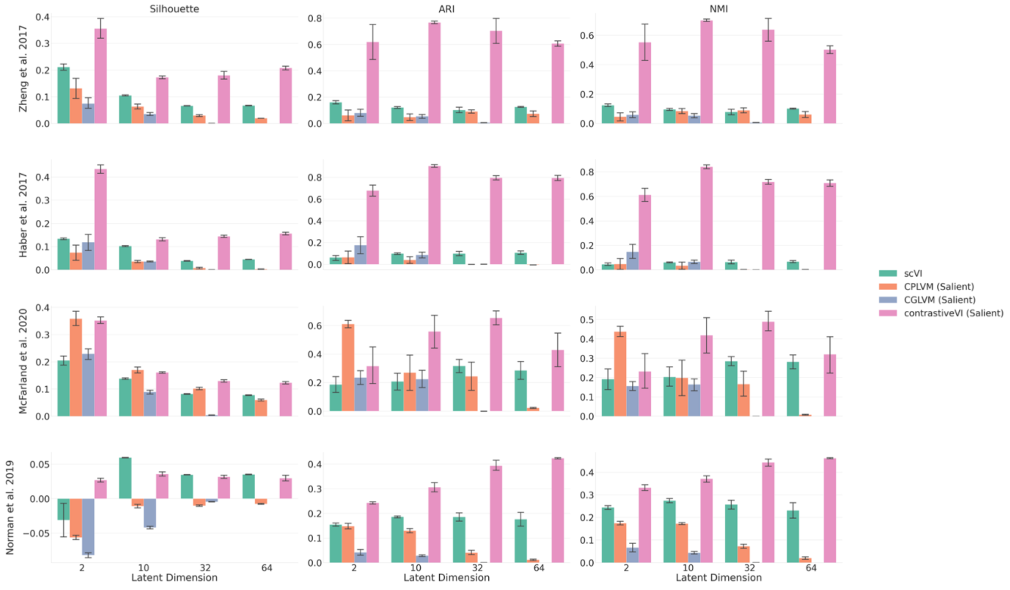
Three separate encoder neural networks were used to parameterize our approximate posterior distributions for z, t, and ℓ. Each network had a single hidden layer consisting of 128 nodes. This was followed by a batch normalization layer (Ioffe & Szegedy, 2015), a rectified linear unit (ReLU) activation function (Nair & Hinton, 2010), and then a dropout layer (Srivastava et al., 2014). During training the dropout probability was set to 0.1. The resulting 128 node values were then used as inputs for two linear layers that parameterized the given factor (e.g. for the encoder corresponding to $q(z|x, s)$, the linear layers parameterized the mean and variance of z). For results in the main text, we used 10-dimensional mean and variance parameters for z and t, and we used a 1-dimensional mean and shape parameter for ℓ.

Our decoder network began with a single hidden layer taking in values of our three latent factors (i.e., z, t and ℓ) with an output dimension of 128. This was followed by batch normalization, a ReLU activation function, and a dropout layer as described previously. The output of this sequence was then fed to three separate decoder layers, one for each of the three parameters of the ZINB distribution. To force the ZINB scale parameter to lie between 0 and 1, we applied a softmax activation function to its corresponding decoder's output. We note that similar decoding approaches have been successfully used by previous unsupervised modeling approaches for scRNA-seq data (Lopez et al., 2018; Gayoso et al., 2021b).

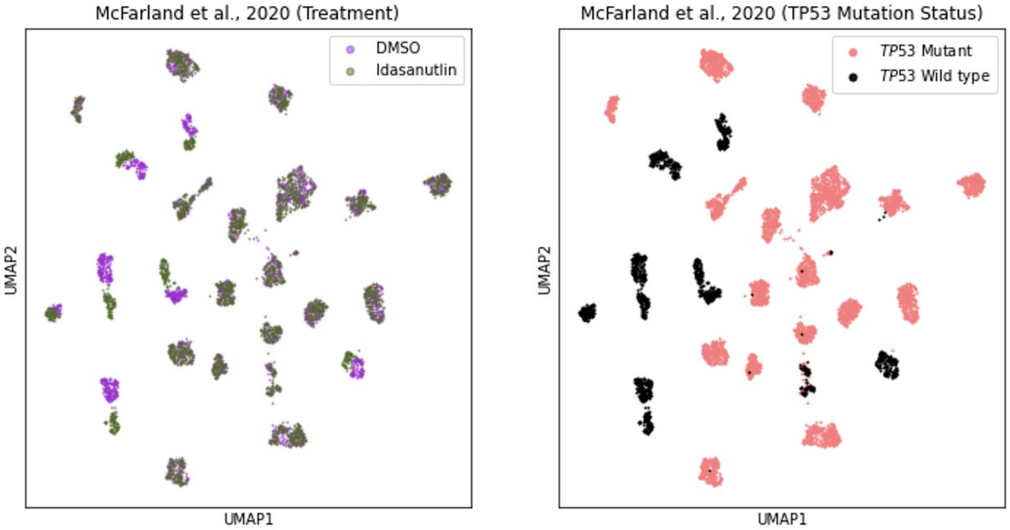
J SUPPLEMENTARY FIGURES



Supplementary Figure 1: **The contrastiveVI probabilistic graphical model.** Unshaded nodes represent latent variables, while shaded nodes represent observed variables. Edges denote conditional independence, while rectangles indicate independent replication.



Supplementary Figure 2: **Model performance with varying (salient) latent dimension.** Mean and standard error of average silhouette width (silhouette), adjusted Rand Index (ARI), and normalized mutual information (NMI) across five random model training trials are plotted for each method's (salient) latent variables at dimension $= 2, 10, 32, 64$ for all benchmark datasets. Results for CGLVM with dimension $= 64$ are not included due to numerical instabilities resulting in NaN values during optimization. Note y-axis scales vary in subplots.



Supplementary Figure 3: **Cell line separation by treatment type in McFarland et al., 2020.** UMAP plots of library-size-normalized and log-transformed data from McFarland et al. (2020) colored by treatment type (left) and *TP53* mutation status (right). Cells with wild type *TP53* clearly separate by treatment type.

K Supplementary Tables

Pathway Name	Pathway Entry	Adjusted p-value
Hematopoietic cell lineage	hsa04640	9.35e-10
Asthma	hsa05310	3.17e-08
Systemic lupus erythematosus	hsa05322	4.91e-05
Antigen processing and presentation	hsa04612	8.29e-05
Type I diabetes mellitus	hsa04940	1.03e-04
Allograft rejection	hsa05330	1.90e-04
Graft-versus-host disease	hsa05332	3.41e-04
Leishmaniasis	hsa05140	5.26e-04
Cell adhesion molecules	hsa04514	5.26e-04
Rheumatoid arthritis	hsa05323	1.09e-03
Chagas disease	hsa05142	1.33e-03
Toxoplasmosis	hsa05145	1.50e-03
Staphylococcus aureus infection	hsa05150	3.18e-03
Intestinal immune network for IgA production	hsa04672	4.04e-03
NF-kappa B signaling pathway	hsa04064	4.04e-03
Viral myocarditis	hsa05416	4.04e-03
Tuberculosis	hsa05152	5.88e-03
Autoimmune thyroid disease	hsa05320	7.14e-03
Inflammatory bowel disease	hsa05321	7.61e-03
Legionellosis	hsa05134	8.51e-03
Influenza A	hsa05164	1.04e-02
B cell receptor signaling pathway	hsa04662	1.59e-02
VEGF signaling pathway	hsa04370	1.59e-02
Glycine, serine and threonine metabolism	hsa00260	1.85e-02
Cytokine-cytokine receptor interaction	hsa04060	1.89e-02
HTLV-I infection	hsa05166	2.76e-02
Transcriptional misregulation in cancer	hsa05202	2.76e-02
Fc epsilon RI signaling pathway	hsa04664	2.81e-02
Apoptosis	hsa04210	3.48e-02
Primary immunodeficiency	hsa05340	4.57e-02
Pertussis	hsa05133	4.57e-02
Colorectal cancer	hsa05210	4.57e-02
Arachidonic acid metabolism	hsa00590	4.57e-02
Osteoclast differentiation	hsa04380	4.57e-02
Arginine and proline metabolism	hsa00330	4.63e-02
T cell receptor signaling pathway	hsa04660	4.69e-02

Supplementary Table 1: All pathways found to be enriched (false discovery rate < 0.05) based on the differentially expressed genes for the two clusters in contrastiveVI's salient latent space for the dataset collected by Zheng et al., 2017.

Pathway Name (Associated Differentially Expressed Genes)	Pathway Entry	Adjusted p-value
Fat digestion and absorption (*Apoa1*, *Apoa4*, *Fabp1*, *Fabp2*, *Pla2g3*)	mmu04975	2.35e-2
Vitamin digestion and absorption (*Apoa1*, *Apoa4*, *Cubn*, *Rbp2*)	mmu04977	2.35e-2
Cholesterol metabolism (*Apoa1*, *Apoa4*, *Apoc2*, *Apoc3*, *Apoh*)	mmu04979	2.93e-2

Supplementary Table 2: All pathways found to be enriched (false discovery rate < 0.05) based on the differentially expressed genes for the two clusters in contrastiveVI's salient latent space for the dataset collected in Haber et al. (2017).

Dataset	Num. Samples (background)	Num. Samples (target)	Platform
Zheng et al. 2017	4,457	12,399	GemCode
Haber et al., 2017	3,240	4,481	SMART-Seq2
McFarland et al., 2020	2,831	3,097	MIX-Seq
Norman et al., 2019	8,907	24,913	Perturb-Seq

Supplementary Table 3: Summary of datasets used.

Cell Type	Number of cells		
	Healthy (background)	*Salmonella* (target)	*H. polygyrus* (target)
Endocrine	112	69	82
Enterocyte	424	705	128
Enterocyte.Progenitor	545	229	586
Goblet	216	126	317
Stem	670	207	592
TA	421	112	353
TA.Early	792	300	436
Tuft	60	22	217

Supplementary Table 4: Number of cell types present in each condition for the dataset by Haber et al. (2017).

Cell Line	Number of cells	
	DMSO-treated (background)	Idasanutlin-treated (target)
BICR6_UPPER_AERODIGESTIVE_TRACT	82	111
BICR31_UPPER_AERODIGESTIVE_TRACT	245	277
BT474_BREAST	53	71
BT549_BREAST	100	131
CAOV3_OVARY	97	140
CCFSTTG1_CENTRAL_NERVOUS_SYSTEM	103	77
COLO680N_OESOPHAGUS	129	129
COV434_OVARY	60	75
DKMG_CENTRAL_NERVOUS_SYSTEM	103	93
IALM_LUNG	105	141
LNCAPCLONEFGC_PROSTATE	139	113
LS1034_LARGE_INTESTINE	72	118
NCIH226_LUNG	165	94
NCIH2347_LUNG	111	159
RCC10RGB_KIDNEY	172	114
RCM1_LARGE_INTESTINE	109	133
RERFLCAD1_LUNG	99	123
SH10TC_STOMACH	123	122
SKMEL2_SKIN	150	141
SKMEL3_SKIN	145	183
SNU1079_BILIARY_TRACT	101	105
SQ1_LUNG	113	150
TEN_ENDOMETRIUM	155	177
UMUC1_URINARY_TRACT	100	120

Supplementary Table 5: Number of cells by cell line present in each condition for the dataset by McFarland et al. (2020).

Model	Model Characteristics						Software Capabilities		
	Expressive	Count Distribution	Over-dispersion	Library size	Batch effects	Generative Model	Dimension Reduction	Differential Expression	Imputation
cPCA							✓		
PCPCA							✓		✓
CLVM							✓		
CPLVM		✓		✓		✓	✓	✓	
CGLVM		✓		✓		✓	✓	✓	
cVAE	✓					✓	✓		✓
contrastiveVI	✓	✓	✓	✓	✓	✓	✓	✓	✓

Supplementary Table 6: Comparison of contrastiveVI with previous methods for contrastive analysis. We compare across both features of the models as well as capabilities of their corresponding software packages. *Expressive*: Indicates whether the model can capture nonlinear relationships. *Count Distribution*: Denotes whether the distribution modeling the data has support in the integer set. *Over-dispersion*: Indicates whether the count-based distribution modeling the data accounts for variance being greater than the mean. *Library size*: Denotes whether the model corrects for library size differences. *Batch effects*: Whether the model can account for nuisance variations due to differences in experimental conditions. *Generative Model*: Indicates if the model supports sampling from a distribution.

Supplementary References

Adam Gayoso, Zoë Steier, Romain Lopez, Jeffrey Regier, Kristopher L Nazor, Aaron Streets, and Nir Yosef. Joint probabilistic modeling of single-cell multi-omic data with totalvi. *Nature Methods*, 18(3):272–282, 2021.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pp. 448–456. PMLR, 2015.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning*, 2010.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.