Modeling variable guide efficiency in pooled CRISPR screens with ContrastiveVI+

Anonymous Author(s)
Affiliation
Address
email

Abstract

Genetic screens mediated via CRISPR-Cas9 combined with high-content readouts have emerged as powerful tools for biological discovery. However, computational analyses of these screens come with additional challenges beyond those found with standard scRNA-seq analyses. For example, perturbation-induced variations of interest may be subtle and masked by other dominant source of variation shared with controls, and variable guide efficiency results in some cells not undergoing genetic perturbation despite expressing a guide RNA. While a number of methods have been developed to address the former problem by explicitly disentangling perturbation-induced variations from those shared with controls, less attention has been paid to the latter problem of noisy perturbation labels. To address this issue, here we propose ContrastiveVI+, a generative modeling framework that both disentangles perturbation-induced from non-perturbation-related variations while also inferring whether cells truly underwent genomic edits. Applied to three large-scale Perturb-seq datasets, we find that ContrastiveVI+ better recovers known perturbation-induced variations compared to previous methods while successfully identifying cells that escaped the functional consequences of guide RNA expression.

1 Introduction

Advances in single-cell genomics technologies have enabled the profiling of molecular modalities across the central dogma at an unprecedented resolution. Moreover, recently developed genetic screening protocols combining CRISPR-Cas9-mediated genome editing with high-content single-cell readouts, such as Perturb-seq [1], hold major promise for identifying the genetic bases of functional phenotypes. Such screens are often conducted in a pooled fashion [2], in which a CRISPR guide RNA (gRNA) library is introduced in bulk to a cell population. Individual cells subsequently receive different gRNAs corresponding to different gene perturbations, and the specific perturbation induced in a cell can be determined by recovering the barcode sequence corresponding to an individual gRNA.

Despite the enormous potential of high-content genetic screens, computational analyses of these datasets are unfortunately beset by numerous challenges. Beyond confounding technical sources of noise present in scRNA-seq data generally, such as differences in sequencing depth and over-dispersion in RNA counts, analyses of pooled CRISPR datasets present an additional unique set of difficulties. For example, perturbation-induced variations in the data may be relatively subtle compared to those due to other biological processes, such as cell-cycle-related variations or those due to cellular stress responses [3]. Thus, standard single-cell analysis techniques, such as principal component analysis or generative modeling approaches (e.g. scVI [4]) may fail to capture perturbation effects, as these methods prioritize capturing factors with the highest variance across an entire dataset.

To work around this issue, a line of recent work [5, 6, 7, 8] has developed so-called contrastive latent variable models (cLVMs) based on the principle of contrastive analysis [9]. Such models explicitly disentangle perturbation-induced variations into a set of *salient* latent variables while factors of variation present in both control and perturbed samples are segregated into a second set of *background* variables. While cLVMs have shown promise for analyzing pooled CRISPR datasets [5, 6], their assumed generative processes disagree with the structure of pooled genetic screens in two important ways. First, while pooled screens measure the effects of many perturbations in a single experiment, standard cLVM models assume a single prior distribution over the salient variables for all perturbed samples. Such models thus may fail to discern perturbations with small effect sizes due to shrinkage toward the shared prior. Second, variable guide efficiency results in a subset of cells escaping the effects of perturbation and acting as control cells despite being labeled with a perturbation [3]. Thus, even when restricted to a single perturbation, the assumption of a single unimodal prior leads to nontrivial model misspecification.

To resolve these issues, here we introduce ContrastiveVI+, an extension of ContrastiveVI, a previously proposed cLVM for scRNA-seq data, that explicitly accounts for the additional structure in pooled genetic screening datasets. The remainder of this work proceeds as follows. In Section 2 we review cLVMs and related work. We then proceed to describe our proposed generative process (Section 3) and our corresponding inference procedure (Section 4). In Section 5 we apply our method to three pooled genetic screening datasets with scRNA-seq readouts, and we find that ContrastiveVI+ learns representations that exhibit better agreement with prior biological knowledge compared to baseline methods while also successfully identifying cells that escaped perturbation.

2 Background: Contrastive Latent Variable Models

Recall that the goal of CA is to disentangle novel perturbation-induced factors of variation from those shared with control samples. To formalize this idea, a number of recent works [7, 8, 6, 5] have developed contrastive latent variable models (cLVMs) using the following framework. Letting x_i denote a perturbed sample (e.g. a cell infected with a non-control gRNA), we assume that x_i is generated from a random process parameterized by θ and conditioned on two sets of latent variables z_i and t_i, i.e.,

$$x_i \sim p_\theta(x_i \mid z_i, t_i)$$

Here t_i denotes a set of salient latent variables capturing novel perturbation-induced variations, while z_i denotes a set of background latent variables capturing variations shared across control and perturbed samples and which are irrelevant our analysis. Without additional constraints, standard latent variable inference procedures are unlikely to naturally disentangle these two sets of latent variables. To address this issue, we leverage our control samples to impose an inductive bias on our model that implicitly encourages disentanglement. In particular, for a control sample $x_j^{\varnothing}$ we assume

$$x_j^{\varnothing} \sim p_\theta(x_j^{\varnothing} \mid z_j, 0).$$

In words, we assume that control samples are generated from the same process p_θ, but with the salient variables fixed at 0 to represent their absence. Equipped with this inductive bias, we may then apply standard inference techniques with appropriate modifications to learn these disentangled sets of variables, and previous works have leveraged this idea to explore high-content pooled CRISPR screening data with linear cLVMs [5] and non-linear cLVMs based on deep neural networks [6]. Yet, as discussed previously, standard cLVMs may not be ideal for modeling pooled CRISPR screens due to the assumption of a single unimodal prior over the salient latent variables shared across all perturbations. To resolve this issue, in the next section we propose a new generative model that extends the standard cLVM framework with a richer prior over the salient variables to reflect the additional structure present in pooled genetic screening datasets.

3 The ContrastiveVI+ Probabilistic Model

For a given cell i labelled with a non-control guide RNA c_i, let

$$z_i \sim \mathcal{N}(0, I_p)$$

denote a low-dimensional set of latent variables capturing factors of variation found in both perturbed cells as well as controls. Next, let

$$y_i \sim \text{Bern}(\alpha)$$

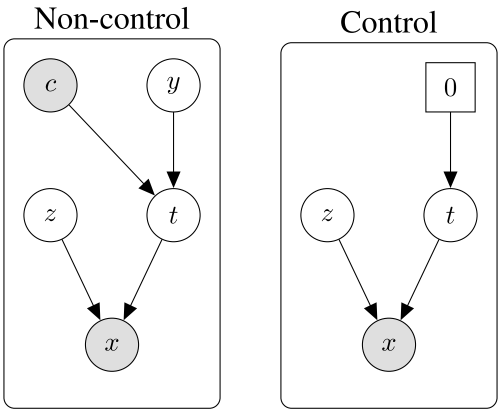

Figure 1: Graphical representation of the ContrastiveVI+ generative process for cells with non-control guides (left) and control guides (right).

denote a binary value indicating whether a cell expressing a gRNA successfully underwent a corresponding genetic perturbation ($y_i = 1$) or failed to do so ($y_i = 0$). For all of the experiments presented in this work we placed an uninformative prior on y_i with $\alpha = 0.5$. Conditioned on this value and the perturbation label c_i, we then draw

$$t_i \mid y_i, c_i \sim y_i \cdot \mathcal{N}(\mu_c, I_q) + (1 - y_i) \cdot \mathcal{N}(\mu_\emptyset, I_q).$$

That is, if a cell successfully underwent a genetic perturbation, we assume that its salient latent representation is drawn from a Gaussian centered at a perturbation-specific mean μ_c. For all results presented in this manuscript, we set our perturbation labels c_i as the gene targeted by the guide in cell i; however, in principle other labeling schemes to be used (e.g. c_i could denote the specific species of gRNA when multiple gRNAs targeting the same gene are used in an experiment). On the other hand, if the cell was not perturbed, we assume that its salient representation was drawn from a Gaussian with mean $\mu_\emptyset$ shared across all perturbations.

Letting f^η denote a neural network with a softmax non-linearity as the final layer, we then compute

$$\rho_i = f^\eta(z_i, t_i).$$

Analogous to scVI [4], this vector on the probability simplex represents the expected normalized expression frequency of each gene g. For a gene g we then assume that the observed gene expression x_{ig} in cell i is drawn

$$x_{ig} \sim \text{ZINB}(\ell_i \rho_{ig}, \theta_g, f^\nu(z_i, t_i)),$$

where ZINB denotes the zero-inflated negative binomial distribution, ℓ_i is the observed library size for cell i, θ_g is a gene-specific inverse dispersion parameter, and f^ν is a neural network whose outputs are interpreted as dropout probabilities.

For a cell j infected with a control gRNA, we assume the same generative process but with y_j fixed at 0. Thus, cell j's salient variables t_j are always drawn from a Gaussian centered at $\mu_\emptyset$, and the region of the salient latent space around $\mu_\emptyset$ thus semantically represents the absence of perturbation-induced variations. We depict our generative processes for cells with control and non-control gRNAs in graphical model form in **Fig. 1**.

4 Inference

Exact posterior inference for our model is intractable, so we instead resort to variational inference [10] via auto-encoding variational Bayes [11]. For cells with non-control guides, we assume that our variational distribution with parameters ϕ factorizes as follows

$$q_\phi(z_i, t_i, y_i, \mid x_i, c_i) = q_{\phi_z}(z_i \mid x_i) q_{\phi_t}(t_i \mid x_i) q_{\phi_y}(y_i \mid t_i),$$

where ϕ_z, ϕ_t, and ϕ_y denote parameters of inference networks for z, t, and y respectively. Here $q(z \mid x)$ and $q(t \mid x)$ take the form of Gaussian distributions, while $q(y \mid t)$ is a Bernoulli distribution. Our corresponding variational bound is then (derivation in Appendix A):

$$\begin{aligned}\mathcal{L}(x_i) = {} & \mathbb{E}_{q_{\phi_z}(z_i|x_i)q_{\phi_t}(t_i|x_i)}\left[p_\theta(x_i \mid z_i, t_i)\right] - D_{KL}(q_{\phi_z}(z_i \mid x_i) \parallel p(z_i)) \\ & - \mathbb{E}_{q_{\phi_t}(t_i|x_i)}\left[D_{KL}(q_{\phi_y}(y_i \mid t_i) \parallel p(y_i))\right] \\ & + \mathbb{E}_{q_{\phi_t}(t_i|x_i)}\left[\left(\sum_{y' \in \{0,1\}} q_{\phi_y}(y' \mid t_i)\left(\log p(t_i \mid y', c_i)\right)\right) - \log q_{\phi_t}(t_i \mid x_i)\right].\end{aligned} \tag{1}$$

For cells with non-targeting control (NTC) guides, we assume an alternative variational distribution incorporating our prior knowledge that factorizes as

$$q_{\phi_{NTC}}(z_j, t_j, y_j, \mid x_j) = q_{\phi_z}(z_j \mid x_j)\delta\{t_j = \mu_\emptyset\}\delta\{y_j = 0\},$$

That is, for control cells we assume that t_j and y_j are fixed at $\mu_\emptyset$ and 0, respectively to reflect the fact that cells with NTC guides are known to be unperturbed ($y_j = 0$) and that the salient variables t_j should not capture variations in the observed data for control cells. We also note that the same inference parameters ϕ_z are used as in the non-NTC case. We then derive a corresponding bound

$$\mathcal{L}_{NTC}(x_j^{\varnothing}) = \mathbb{E}_{q_{\phi_z}(z_j,|x_j^{\varnothing})}\left[p_\theta(x_j^{\varnothing}, |z_j, t_j = \mu_\emptyset)\right] - D_{KL}(q_{\phi_z}(z_j \mid x_j^{\varnothing}) \parallel p(z_j)). \tag{2}$$

By fixing $t_j = \mu_\emptyset$ for NTC cells during the inference procedure, we ensure that the salient variables t_j do not capture any sources of variation for cells with NTC guides. Thus, as the recognition network parameters ϕ_z are shared across NTC and non-NTC guide cells, the background variables z are implicitly encouraged to recover sources of variation shared across cells from both groups, while the salient variables t are then free to recover the additional variations only present in perturbed cells. As all cells are assumed to be generated independently, we may then perform inference by maximizing the sums of Equations 1 and 2 across all cells via minibatch gradient ascent similar to standard cLVMs [7, 5, 6, 8]. Perturbation-specific means μ_c along with $\mu_\emptyset$ are learned as point estimates and optimized along with our model's other parameters.

While similar implicit schemes have been successfully employed to encourage disentanglement in standard cLVMs, in initial experiments we found that additional regularization was required to ensure that our inference procedure respected the intended semantics of our more structured generative process. First, as the salient space recognition network q_{ϕ_t} does not learn from cells with NTC guides when optimizing Equation 2, we found that q_{ϕ_y} did not reliably associate nonperturbed cells to the region of the salient latent space around $\mu_\emptyset$. To remedy this issue, we added an additional KL penalty to $\mathcal{L}_{NTC}$ encouraging NTC cells to map to the null region of the salient space, yielding

$$\mathcal{L}^*_{NTC}(x_j^{\varnothing}) = \mathcal{L}_{NTC} - D_{KL}(q(t_j \mid x_j^{\varnothing}) \parallel \mathcal{N}(\mu_\emptyset, I_q)). \tag{3}$$

Second, for datasets with larger numbers of perturbations, we observed that perturbation-induced variations were sometimes undesirably captured in the model's background latent space. To discourage this behavior, we leveraged the maximum mean discrepancy (MMD) [12], a kernel-based test statistic for determining whether two groups of samples were drawn from the same distribution. In particular, letting Z^b denote a minibatch of posterior background latent space samples for cells with non-NTC guides and defining Z^b_{NTC} analogously for cells with control guides, we add the penalty

$$-\lambda \cdot \ell_{\text{MMD}}(Z^b, Z^b_{NTC}), \tag{4}$$

where ℓ_{MMD} denotes the empirical estimate of the MMD of Gretton et al. [12] and λ controls the regularization strength. For all experiments presented in this work, we tuned λ such that the MMD penalty was of similar magnitude to the KL regularization terms in our ELBOs, a strategy successfully employed in previous work [6]. With this penalty, ContrastiveVI+ 's background latent space was thus explicitly encouraged to capture variations shared across cells with NTC and non-NTC guides.

Combined with our more structured generative process, ContrastiveVI+'s inference procedure provides two major advantages over that of the original ContrastiveVI model. First, our richer prior on the salient latent space with perturbation-specific means may allow ContrastiveVI+ to recover subtle perturbation-induced effects in the salient latent space that would be shrunk to the shared unimodal prior in the standard cLVM setup. Second, post-training our approximate posterior $q_{\phi_y}(y_i \mid t_i)$ can be used to classify cells that escaped perturbation versus those that were successfully perturbed.

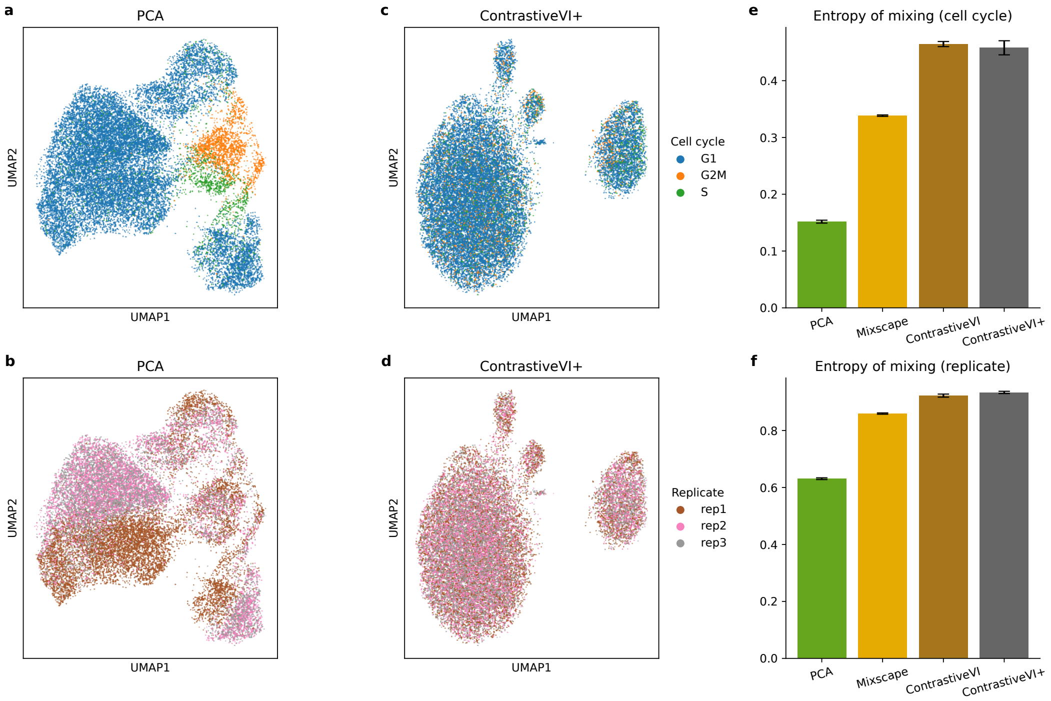

Figure 2: **a-b**, UMAP visualizations of PCA (**a**) applied to data from Papalexi et al. [3] colored by cell cycle (**a**) and replicate (**b**). **c-d** UMAP visualizations of ContrastiveVI+'s salient latent space colored by cell cycle (**c**) and replicate (**d**). **e-f**, Entropy of mixing for ContrastiveVI+ and baseline methods' representations with respect to cell cycle phase (**e**) and replicate identity (**f**).

5 Experiments

Overview. To evaluate our method, we applied it to three publicly available pooled genetic screening datasets [3, 13, 14]. Based on previous analyses, for each of these datasets we have known confounding sources of variation (e.g. cell cycle) and/or known perturbation-induced variations (e.g. common gene programs induced by groups of perturbations) that allowed us to assess the quality of ContrastiveVI+ 's learned representations. Moreover, we also used these datasets to assess the quality of ContrastiveVI+ 's predictions of escaping versus perturbed cells. Details regarding the preprocessing of these datasets can be found in Appendix B.

Baselines. For each dataset we benchmarked ContrastiveVI+ against two previously proposed methods for exploring perturbation-induced variations in pooled genetic screens. First, we considered the original ContrastiveVI model of Weinberger et al. [6] to assess whether our more structured generative process could better recover subtle perturbation-induced variations. Second, we considered the Mixscape method proposed in Papalexi et al. [3]. Mixscape provides both a procedure for isolating perturbation-induced variations, via a nearest-neighbors-based approach for computing so-called "perturbation signatures" for each cell, and a procedure for identifying escaping cells. Finally, as a naive baseline we also include results from simply applying principal component analysis (PCA) to normalized expression levels. Further details on our implementation of ContrastiveVI+ and baseline methods can be found in Appendix C.

5.1 Initial validation of ContrastiveVI+ on a small-scale ECCITE-seq dataset

We first applied ContrastiveVI+ to data originally presented in Papalexi et al. [3] collected using ECCITE-seq [15], a protocol that combines pooled CRISPR screening with single-cell transcriptomic as well as surface protein measurements. This dataset was collected from a human leukemia monocytic cell line (THP-1) after stimulation with interferon gamma (IFN-γ), with the goal of identifying immune checkpoint regulators. In this dataset a total of 25 genes were targeted for CRISPR knockout.

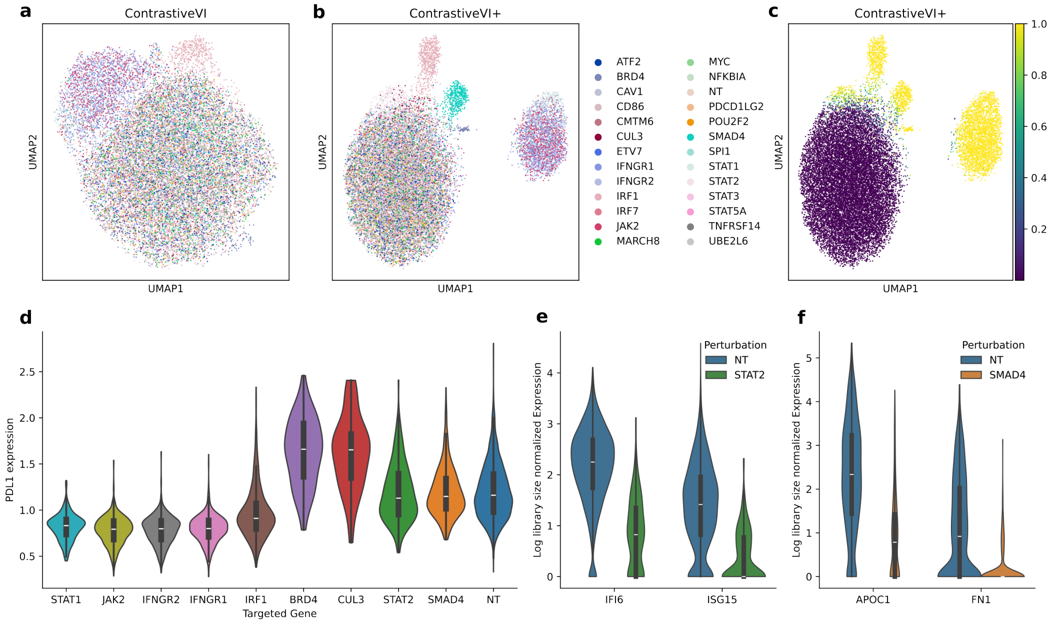

Figure 3: **a-b**, UMAP plots of ContrastiveVI and ContrastiveVI+'s salient latent representations colored by gene target. **c**, UMAP plot of ContrastiveVI+'s salient latent representations colored by inferred probability of perturbation. **d**, *PDL1* protein expression for gene perturbations highlighted in ContrastiveVI+'s salient latent space compared to control cells. **e-f**, Subset of transcriptomic changes for *STAT2*- and *SMAD4*-perturbed cells identified by ContrastiveVI+.

Beyond variations due to novel perturbation-induced phenotypes, in their original analyses Papalexi et al. [3] identified substantial confounding sources of variation shared with control cells due to cell cycle phase and batch effects (**Fig. 2a-b**). We thus began our experiments by assessing ContrastiveVI+ 's ability to remove these confounding sources of variation in its salient latent space. Qualitatively, we found that ContrastiveVI+ 's salient latent space was indeed invariant to these confounding variations (**Fig. 2c-d**). To quantify ContrastiveVI+ and baseline methods' performance on this task, we computed the entropy of mixing (Appendix D.1) for each method's representations with respect to cell cycle phase and replicate identity. While all methods resulted in better mixing compared to the naive PCA baseline, we found that ContrastiveVI and ContrastiveVI+ achieved stronger performance on this task for both confounding sources of variation (**Fig. 2e-f**) compared to Mixscape's perturbation signatures. This result suggests that the nearest-neighbors based approach employed by Mixscape is less effective at isolating perturbation-induced variations compared to deep generative models.

We next investigated whether ContrastiveVI+ 's richer model could highlight further trends compared to ContrastiveVI. We found that both methods (**Fig. 3a-b**) highlighted a cluster of cells with gRNA's corresponding to known upstream components of the IFN-γ pathway (*IFNGR1*, *IFNGR2*, *JAK2* and *STAT1*), a cluster of cells with gRNAs corresponding to the downtream IFN-γ mediator *IRF1*, and a third cluster containing containing cells from all perturbations (including non-targeting gRNAs). Beyond these clusters, ContrastiveVI+ additionally highlighted clusters of cells expressing gRNAs targeting *SMAD4*, *BRD4*, and *STAT2*. Moreover, when inspecting ContrastiveVI+ 's inferred values of y (**Fig. 3c**), we found that cells in mixed cluster shared with controls were assigned as non-perturbed ($y \approx 0$) while cells in the non-control clusters were classified as perturbed ($y \approx 1$), suggesting that ContrastiveVI+ successfully distinguished perturbed versus escaping cells.

We verified that these clusterings corresponded to meaningful perturbation effects by first inspecting *PDL1* surface protein expression levels (**Fig. 3d**) for cells predicted by ContrastiveVI+ as perturbed. For most genes, we found corresponding decreases (*IFNGR1*, *IFNGR2*, *JAK2*, *STAT1*, and *IRF1*) or increases (*BRD4*, *CUL3*) in *PDL1* expression compared to control cells. Moreover, while *STAT2* and *SMAD4* did not affect *PDL1* expression, we nevertheless found clear transcriptomic changes in cells predicted by ContrastiveVI+ as perturbed compared to controls (**Fig. 3e-f**). In particlar, *STAT2*-perturbed cells exhibited strong downregulation of interferon-induced genes (e.g. *IFI6*, *ISG15*)

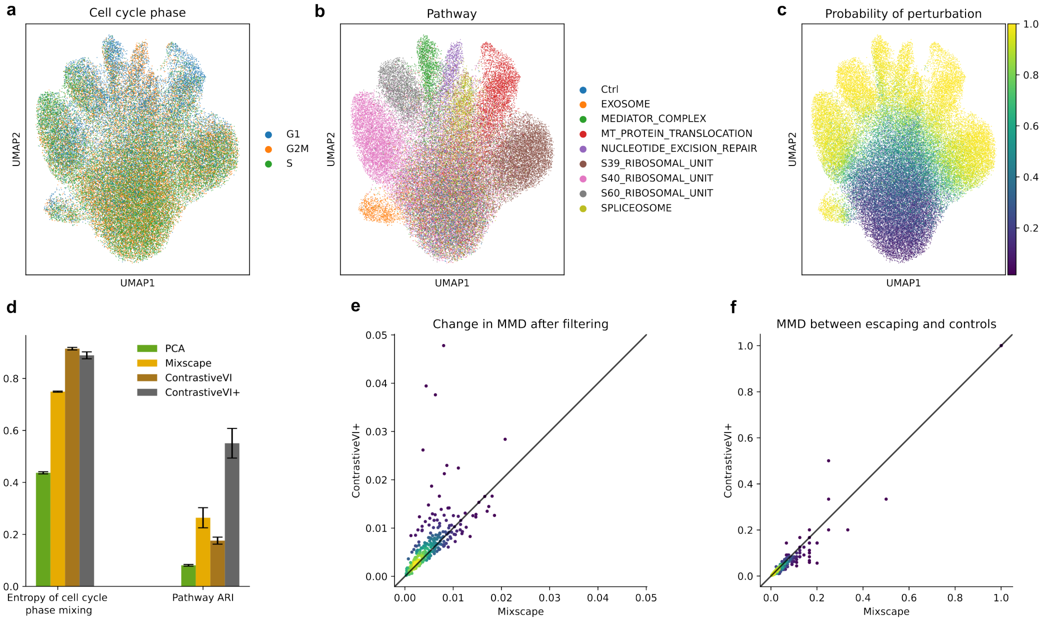

Figure 4: **a-c**, UMAP plots of ContrastiveVI+'s salient latent space for data from Replogle et al. [13] colored by cell cycle phase (**a**), pathway annotations (**b**) and inferred probability of perturbation (**c**). **d**, Quantitative assessments of invariance with respect to cell cycle phase (entropy of cell cycle phase mixing) and capturing of known perturbation-induced variations (pathway ARI) for ContrastiveVI+ and baseline method's salient representations. **e**, Change in MMD between cells labeled with gRNAs targeting a given gene versus cells with non-targeting control guides after filtering with ContrastiveVI+ (y-axis) or Mixscape (x-axis) compared to the MMD without filtering. **f**, MMD between cells labeled as escaping by ContrastiveVI+ (y-axis) or Mixscape (x-axis) versus control cells.

while *SMAD4*-perturbed cells demonstrated downregulation in inflammatory response genes (e.g. *APOC1*, *FN1*).

Taken together, these results illustrate that ContrastiveVI+ indeed may highlight additional structure in the model's salient latent space compared to standard cLVMs. Moreover, these results demonstrate that our inference procedure can identify cells exhibiting perturbation effects versus escaping cells.

5.2 Further validation on a larger-scale CRISPRi screen

We next applied ContrastiveVI+ to analyze data from a CRISPR interference (CRISPRi) Perturb-seq dataset presented in Replogle et al. [13]. Following previous work [16, 17], in our experiments we considered a subset of perturbations identified as having nontrivial effect sizes and which were labeled by the original authors of Replogle et al. [13] as affecting specific biological pathways. In total, we retained data from cells with guides targeting 336 genes as well as cells with NTC guides.

We began our analysis by assessing the quality of ContrastiveVI+'s salient representations. Previous analyses of this data [18] have identified cell cycle as a major confounding source of variation in this dataset shared with control cells. Thus, we would expect ContrastiveVI+ 's salient latent space to be invariant with respect to cell cycle phase. Moreover, we would expect cells to cluster based on the pathway labels assigned to perturbations by [13]. We found that ContrastiveVI+ 's salient space was indeed invariant to cell cycle (**Fig. 4a**), with clear clusters separating by pathway label (**Fig. 4b**). Moreover, cells from these distinct pathway clusters were inferred as truly perturbed, while cells in the remaining cluster mixed across pathways and control cells were inferred as escaping (**Fig. 4c**).

To compare ContrastiveVI+ and baselines' performance on this task, we employed the entropy of mixing to measure invariance with respect to cell cycle and used the adjusted rand index (ARI; Appendix D.2) to quantify separation based on pathway labels. When computing pathway ARI, for ContrastiveVI+ and Mixscape we restricted our attention to cells labeled by these methods as truly perturbed to mitigate the impact of escaping cells; for ContrastiveVI we used all cells as this method does not predict perturbed versus escaping cells. We found that ContrastiveVI+ and ContrastiveVI had the strongest performance on cell cycle mixing, while ContrastiveVI+ achieved substantially stronger performance on the pathway ARI metric compared to baselines. These results further demonstrate ContrastiveVI+ 's superior ability to isolate perturbation-induced variations.

Given its large number of distinct perturbations, we further used this dataset to benchmark ContrastiveVI+ and Mixscape's procedures for identifying perturbed versus escaping cells. Intuitively, cells classified as perturbed should have substantially different gene expression profiles compared to cells with NTC guides. To capture this idea, we used the MMD to measure the distance between populations of cells. Specifically, for each gene perturbation we computed the MMD between cells labeled with a gRNA targeting that gene versus cells with NTC guides. We then recalculated the MMD after filtering to cells labeled by ContrastiveVI+ or Mixscape as perturbed, and finally computed the change in MMD with filtering versus without filtering. We present our results for this experiment in **Fig. 4e**. Here a *higher* change in MMD indicates that cells classified as perturbed exhibit stronger differences from control cells and thus represents better performance. We found that for a majority of genes ContrastiveVI+ led to larger changes in MMD compared to Mixscape, with statistical significance confirmed with a binomial test assuming a null hypothesis of equal chance of either method achieving better performance for each gene ($p < 1 \cdot 10^{-5}$).

In isolation, such a metric could be maximized by only retaining cells with the most extreme changes compared to controls and erroneously labeling many truly-perturbed cells as escaping. To counteract this potential pathology, we thus also assessed whether cells labeled as escaping perturbation by each method were indeed similar to cells with NTC guides. To do so, we computed the MMD between cells labeled as escaping by each method versus cells with NTC guides, and we present our results in **Fig. 4f**. Here lower MMD values indicate that the cells flagged by a method as escaping are closer to true controls and thus represent better performance. We found that cells predicted as escaping by ContrastiveVI+ largely had lower MMDs compared to Mixscape ($p < 1 \cdot 10^{-5}$, binomial test).

Taken together, these results suggest ContrastiveVI+'s more expressive modeling procedure facilitates superior prediction of perturbed versus escaping cells compared to Mixscape.

5.3 Exploring the diversity in perturbation responses in a CRISPRa screen

As a final demonstration of ContrastiveVI+ 's capabilities, we applied it to explore a Perturb-seq dataset from Norman et al. [14]. In this dataset the authors assessed the effects of CRISPR activation (CRISPRa) perturbations on K562 cells. For our analysis, we focused on a subset of these perturbations labeled in Norman et al. [14] as inducing specific gene programs.

We began by confirming that ContrastiveVI+ successfully isolated perturbation-induced variations in its salient latent space. Based on the analysis of Norman et al. [14], we would expect cells to separate by gene program labels. Moreover, we would expect cells to mix across cell cycle phase, a known confounding source of variation shared with control cells in this dataset [6]. We found that cells indeed mixed across cell cycle phase (**Fig. 5a**), while separating by gene program labels (**Fig. 5b**), with cells in the separated gene program clusters being predicted by ContrastiveVI+ as perturbed (**Fig. 5c**). Moreover, ContrastiveVI+ achieved significantly better separation of gene programs compared to baseline methods as measured by ARI while also achieving strong performance at removing cell cycle effects as measured by entropy of mixing (**Fig. 5d**).

In addition to separation between the gene program clusters, in our analysis we also observed separation between substructures within these clusters. To highlight ContrastiveVI+'s potential to facilitate additional insights, we thus further inspected the cells with perturbations labeled as "granulocyte/apoptosis" and which were predicted by ContrastiveVI+ as perturbed (**Fig. 5e**). Notably, in Norman et al. [14], the authors of that work largely analyzed the relationships between perturbations at the pseudobulk level (i.e., by considering the mean expression profile for each perturbation). Thus, a particular focus of our analysis was to see if our single-cell-level model could uncover further relationships between perturbations beyond those discussed in Norman et al. [14].

First, to understand the genes driving separation in ContrastiveVI+'s latent space, we employed Hotspot [19], a tool for identifying informative genes in single-cell data by ranking genes in terms of spatial autocorrelation with respect to a given metric of cell–cell similarity (e.g., the latent space of a VAE). From the top genes returned by Hotspot, we found that separation in ContrastiveVI+'s salient space was strongly correlated with canonical granulocyte marker genes, such as *LST1*, *CSF3R*, and

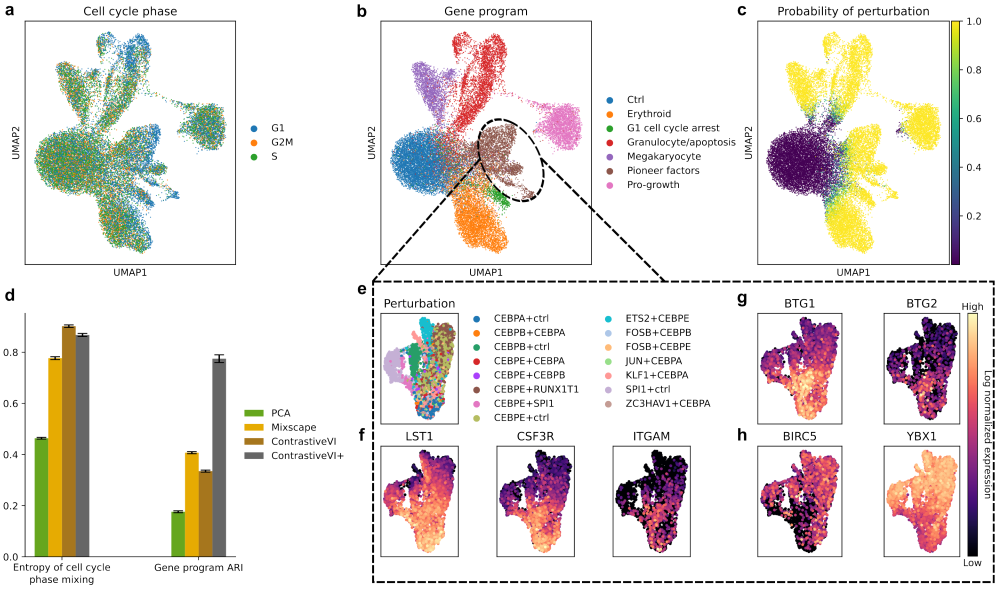

Figure 5: **a-c**, UMAP visualizations of ContrastiveVI+'s salient latent space for Norman et al. [14] colored by cell cycle phase (**a**), gene program labels provided by Norman et al. [14] (**b**), and inferred probability of perturbation (**c**). **d**, Quantitative assessments of ContrastiveVI+ and baseline method's salient representations. **e-h**, ContrastiveVI+'s salient latent representations for cells with perturbations labeled as "granulocyte/apoptosis" by Norman et al. [14]. Plots colored by perturbation labels (**e**), canonical granulocyte marker genes (**f**), the pro-apoptotic anti-proliferation factors *BTG1* and *BTG2* (**g**), and anti-apoptotic genes *BIRC5* (survivin) and *YBX1* (**h**).

ITGAM (**Fig. 5f**). Moreover, expression of the granulocyte markers was correlated with expression of the pro-apoptotic anti-proliferation factors *BTG1* and *BTG2* (**Fig. 5g**), and inversely correlated with the anti-apoptotic genes *BIRC5*, also known as survivin, and *YBX1* (**Fig. 5h**).

Furthermore, we observed clear heterogeneity in the responses induced by individual perturbations. For example, while some perturbations induced consistently strong upregulation of granulocyte markers (e.g. *CEBPA*+ctrl), other perturbations (e.g. *CEBPE+ctrl*) resulted in more variable responses. Notably, we often observed strong mixing between cells perturbed solely to activate *CEBPE* (i.e., *CEBPE*+ctrl) and cells perturbed to activate *CEBPE* along with a second gene (e.g. *CEBPE+RUNX1T1*, *CEBPE+CEBPA*, and *FOSB+CEBPE*). This phenomenon suggests that activation of *CEBPE* is sufficient to achieve a certain cellular state, with other perturbations not having an observable impact. Moreover, this behavior is consistent with *CEBPE*'s known function of strongly driving terminal differentiation for granulocytes [20]; in other words, due to *CEBPE* activation inducing cells to differentiate into granulocytes, the effects of additional perturbations may be muted.

Notably, these phenomena were not discussed in Norman et al. [14], and these results illustrate how the higher resolution of our single-cell-level modeling approach may facilitate insights into the diversity of perturbation responses beyond those possible from previous workflows.

6 Conclusion

Here we introduced ContrastiveVI+, a deep generative modeling framework for exploring perturbation-induced variations in pooled genetic screening datasets while explicitly accounting for variable guide efficiency and the diversity of responses induced by different perturbations. In experiments on three datasets with scRNA-seq readouts, we found that our model's additional structure resulted in substantially better recovery of known biological relationships compared to baseline methods while also successfully predicting truly perturbed versus escaping cells. Moreover, we found that our more structured modeling approach could reveal further biological insights beyond those provided by other analysis workflows. Future work will involve assessing ContrastiveVI+'s abilities on larger scale datasets and additional high-content screening modalities beyond scRNA-seq.

References

[1] 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.

[2] Christoph Bock, Paul Datlinger, Florence Chardon, Matthew A Coelho, Matthew B Dong, Keith A Lawson, Tian Lu, Laetitia Maroc, Thomas M Norman, Bicna Song, et al. High-content crispr screening. *Nature Reviews Methods Primers*, 2(1):1–23, 2022.

[3] Efthymia Papalexi, Eleni P Mimitou, Andrew W Butler, Samantha Foster, Bernadette Bracken, William M Mauck III, Hans-Hermann Wessels, Yuhan Hao, Bertrand Z Yeung, Peter Smibert, et al. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nature genetics*, 53(3):322–331, 2021.

[4] 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.

[5] Andrew Jones, F William Townes, Didong Li, and Barbara E Engelhardt. Contrastive latent variable modeling with application to case-control sequencing experiments. *The Annals of Applied Statistics*, 16(3):1268–1291, 2022.

[6] Ethan Weinberger, Chris Lin, and Su-In Lee. Isolating salient variations of interest in single-cell data with contrastivevi. *Nature Methods*, 20(9):1336–1345, 2023.

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

[8] 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, pages 4862–4869, 2019.

[9] 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, 2013.

[10] 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.

[11] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv e-prints*, pages arXiv–1312, 2013.

[12] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[13] Joseph M Replogle, Reuben A Saunders, Angela N Pogson, Jeffrey A Hussmann, Alexander Lenail, Alina Guna, Lauren Mascibroda, Eric J Wagner, Karen Adelman, Gila Lithwick-Yanai, et al. Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, 185(14):2559–2575, 2022.

[14] Thomas M Norman, Max A Horlbeck, Joseph M Replogle, Alex Y Ge, 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.

[15] Eleni P Mimitou, Anthony Cheng, Antonino Montalbano, Stephanie Hao, Marlon Stoeckius, Mateusz Legut, Timothy Roush, Alberto Herrera, Efthymia Papalexi, Zhengqing Ouyang, et al. Multiplexed detection of proteins, transcriptomes, clonotypes and crispr perturbations in single cells. *Nature methods*, 16(5):409–412, 2019.

[16] Michael Bereket and Theofanis Karaletsos. Modelling cellular perturbations with the sparse additive mechanism shift variational autoencoder. *Advances in Neural Information Processing Systems*, 36, 2024.

[17] Romain Lopez, Natasa Tagasovska, Stephen Ra, Kyunghyun Cho, Jonathan Pritchard, and Aviv Regev. Learning causal representations of single cells via sparse mechanism shift modeling. In *Conference on Causal Learning and Reasoning*, pages 662–691. PMLR, 2023.

[18] Xinming Tu, Jan-Christian Hutter, Zitong Jerry Wang, Takamasa Kudo, Aviv Regev, and Romain Lopez. A supervised contrastive framework for learning disentangled representations of cell perturbation data. *bioRxiv*, pages 2024–01, 2024.

[19] David DeTomaso and Nir Yosef. Hotspot identifies informative gene modules across modalities of single-cell genomics. *Cell systems*, 12(5):446–456, 2021.

[20] Kim Theilgaard-Mönch, Sachin Pundhir, Kristian Reckzeh, Jinyu Su, Marta Tapia, Benjamin Furtwängler, Johan Jendholm, Janus Schou Jakobsen, Marie Sigurd Hasemann, Kasper Jermiin Knudsen, et al. Transcription factor-driven coordination of cell cycle exit and lineage-specification in vivo during granulocytic differentiation: In memoriam professor niels borregaard. *Nature Communications*, 13(1):3595, 2022.

[21] Danila Bredikhin, Ilia Kats, and Oliver Stegle. Muon: multimodal omics analysis framework. *Genome biology*, 23(1):42, 2022.

[22] Adam Gayoso, Romain Lopez, Galen Xing, Pierre Boyeau, Valeh Valiollah Pour Amiri, Justin Hong, Katherine Wu, Michael Jayasuriya, Edouard Mehlman, Maxime Langevin, et al. A python library for probabilistic analysis of single-cell omics data. *Nature biotechnology*, 40(2): 163–166, 2022.

[23] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[24] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] Isaac Virshup, Danila Bredikhin, Lukas Heumos, Giovanni Palla, Gregor Sturm, Adam Gayoso, Ilia Kats, Mikaela Koutrouli, Bonnie Berger, et al. The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nature biotechnology*, 41(5):604–606, 2023.

[26] Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 36(5):421–427, 2018.

A Derivation of variational bounds for ContrastiveVI+

Here we present a full derivation of the variational bound for cells with non-NTC guides presented in Section 4. We begin by assuming that our variational distribution factorizes as

$$q_\phi(z_i, t_i, y_i, \mid x_i, c_i) = q_{\phi_z}(z_i \mid x_i) q_{\phi_t}(t_i \mid x_i) q_{\phi_y}(y_i \mid t_i),$$

where ϕ_z, ϕ_t, and ϕ_y denote parameters of inference networks for z, t, and y respectively. Leveraging our variational distribution as well as the generative process proposed in Section 3, we then proceed to derive a corresponding ELBO:

$$\begin{aligned}
&\mathcal{L}(x_i) = \mathbb{E}_{q_\phi(z_i, t_i, y_i \mid x_i)} \left[\log \frac{p(z_i, t_i, y_i, x_i \mid c_i)}{q_\phi(z_i, t_i, y_i \mid x_i)} \right] \\
&= \mathbb{E}_{q_{\phi_z}(z_i \mid x_i) q_{\phi_t}(t_i \mid x_i) q_{\phi_y}(y_i \mid t_i)} \left[\log \frac{p(z_i) p(t_i \mid y_i, c_i) p(y_i) p(x_i \mid z_i, t_i)}{q_{\phi_z}(z_i \mid x_i) q_{\phi_t}(t_i \mid x_i) q_{\phi_y}(y_i \mid t_i)} \right] \\
&= \mathbb{E}_{q_{\phi_z}(z_i \mid x_i) q_{\phi_t}(t_i \mid x_i) q_{\phi_y}(y_i \mid t_i)} \left[\log \frac{p(z_i)}{q_{\phi_z}(z_i \mid x_i)} + \log \frac{p(t_i \mid y_i, c_i)}{q_{\phi_t}(t_i \mid x_i)} + \log \frac{p(y_i)}{q_{\phi_y}(y_i \mid t_i)} + \right. \\
&\qquad \left. \log p(x_i \mid z_i, t_i) \right] \\
&= \mathbb{E}_{q_{\phi_z}(z_i \mid x_i) q_{\phi_t}(t_i \mid x_i)} \left[p(x_i \mid z_i, t_i) \right] + \mathbb{E}_{q_{\phi_z}(z_i \mid x_i)} \left[\log \frac{p(z_i)}{q_{\phi_z}(z_i \mid x_i)} \right] + \\
&\qquad \mathbb{E}_{q_{\phi_t}(t_i \mid x_i) q_{\phi_y}(y_i \mid t_i)} \left[\log \frac{p(t_i \mid y_i, c_i)}{q_{\phi_t}(t_i \mid x_i)} \right] + \mathbb{E}_{q_{\phi_t}(t_i \mid x_i) q_{\phi_y}(y_i \mid t_i)} \left[\log \frac{p(y_i)}{q_{\phi_y}(y_i \mid t_i)} \right] \\
&= \mathbb{E}_{q_{\phi_z}(z_i \mid x_i) q_{\phi_t}(t_i \mid x_i)} \left[p(x_i \mid z_i, t_i) \right] - D_{KL}(q_{\phi_z}(z_i \mid x_i) \parallel p(z_i)) \\
&\qquad - \mathbb{E}_{q_{\phi_t}(t_i \mid x_i)} \left[D_{KL}(q_{\phi_y}(y_i \mid t_i) \parallel p(y_i)) \right] \\
&\qquad + \mathbb{E}_{q_{\phi_t}(t_i \mid x_i)} \left[\left(\sum_{y' \in \{0,1\}} q_{\phi_y}(y' \mid t_i) \left(\log p(t_i \mid y', c_i) \right) \right) - \log q_{\phi_t}(t_i \mid x_i) \right]
\end{aligned}$$

For control cells infected with non-targeting control (NTC) guides we assume the following variational distribution:

$$q_{\phi_{NTC}}(z_j, t_j, y_j, \mid x_j^\varnothing) = q_{\phi_z}(z_j \mid x_j^\varnothing) \delta\{t_j = \mu_\emptyset\} \delta\{y_j = 0\}.$$

Our corresponding ELBO is then:

$$\begin{aligned}
&\mathcal{L}_{NTC}(x_j^\varnothing) = \mathbb{E}_{q(z_j, \mid x_j^\varnothing)} \left[\log \frac{p(z_j, x_j \mid t_j = \mu_\emptyset, y_j = 0)}{q(z_j \mid x_j^\varnothing)} \right] \\
&= \mathbb{E}_{q(z_j, \mid x_j^\varnothing)} \left[\log \frac{p(x_j^\varnothing, \mid z_j, t_j = \mu_\emptyset) p(z_j)}{q(z_j \mid x_j^\varnothing)} \right] \\
&= \mathbb{E}_{q(z_j, \mid x_j^\varnothing)} \left[p(x_j^\varnothing, \mid z_j, t_j = \mu_\emptyset) p(z_j) \right] - D_{KL}(q(z_j \mid x_j^\varnothing) \parallel p(z_j)).
\end{aligned}$$

B Dataset preprocessing

Here we provide descriptions of any preprocessing steps for the datasets considered in this work.

Papalexi et al. [3]. The cell by gene count matrix along with corresponding metadata for this dataset was obtained from the NIH gene expression omnibus entry GSE153056. For our analysis we considered the top 2,000 highly variable genes returned from the Scanpy `highly_variable_genes` function with `flavor=seurat_v3`. For the analysis presented in Section 5, normalized protein counts were computed using the centered log ratio transform as implemented in muon [21] with `axis=1` to match the original analysis of Papalexi et al. [3]. Based on the original analysis of Papalexi et al. [3], cells with perturbations identified as having trivial effects were all labeled as nonperturbed and considered as control cells.

Norman et al. [14]. The cell by gene count matrix for this dataset along with corresponding metadata was obtained from the NIH gene expression omnibus entry GSE133344. As done in the analysis of Norman et al. [14], cells with the perturbation label `NegCtrl1_NegCtrl0__NegCtrl1_NegCtrl0` were excluded from our analysis. Cells marked as doublets (i.e., a `number_of_cells` metadata value greater than 1.0) by Norman et al. [14] were also excluded from our analysis. For our experiments we retained all cells with control guides along with cells infected with non-control guides that were annotated with gene program labels by Norman et al. [14]. For our analysis we considered the top 2,000 highly variable genes returned from the Scanpy `highly_variable_genes` function with `flavor=seurat_v3`.

Replogle et al. [13]. For the analysis presented in this work we considered a filtered version of the original genome-wide data presented in Replogle et al. [13] provided by Bereket and Karaletsos [16] that retained data from perturbations with non-trivial effect sizes. Among this set of perturbations, for our experiments we considered the perturbations that had corresponding pathway labels provided by Replogle et al. [13], and we used the same set of highly variable genes considered in Bereket and Karaletsos [16].

C ContrastiveVI+ and baseline method implementation details

In the experiments presented in Section 5, we compared our proposed ContrastiveVI+ against two baselines: the original ContrastiveVI model of Weinberger et al. [6] and the Mixscape method of Papalexi et al. [3]. Here we provide a brief overview of these methods and their corresponding implementations used in this work.

C.1 ContrastiveVI+

Our implementation of ContrastiveVI+ was performed using the `scvi-tools` library [22]. Our variational distributions q_{ϕ_z} and q_{ϕ_t} were implemented as multilayer perceptrons with a single hidden layer of 128 units and ReLU activation functions [23]. For all experiments we set the dimensionality of both the background and salient latent spaces (i.e., z, and t) to 10. For q_{ϕ_y}, we found that using additional hidden layers led to more consistent performance across random initialization, with good stability achieved with three hidden layers. Thus, for all results presented in this manuscript we implemented q_{ϕ_y} as an MLP with three hidden layers with 128 units each. For our decoder network we used an MLP with a single hidden layer of 128 units. All ContrastiveVI+ models were optimized using Adam [24] with the default parameters in `scvi-tools`.

C.2 ContrastiveVI

The original ContrastiveVI model of Weinberger et al. [6] extends the scVI model of Lopez et al. [4] via the contrastive latent variable modeling framework described in Section 2. Specifically, for a cell i labelled with a non-control gRNA, ContrastiveVI assumes the following generative process. Let

$$z_i \sim \mathcal{N}(0, I)$$

denote a low-dimensional set of *background* latent variables capturing factors of variation found in both perturbed cells as well as controls. Next, let

$$t_i \sim \mathcal{N}(0, I)$$

denote a low-dimensional set of *salient* latent variables capturing novel perturbation-induced variations in cells labeled with non-NTC guides.

Letting f^η denote a neural network with a softmax non-linearity as the final layer, we then compute

$$\rho_i = f^\eta(z_i, t_i).$$

Analogous to scVI, this vector on the probability simplex represents the expected normalized expression frequency of each gene g. For a gene g we then assume that the observed gene expression x_{ig} in cell i is drawn

$$x_{ig} \sim \text{ZINB}(\ell_i \rho_{ig}, \theta_g, f^\nu(z_i, t_i)),$$

where ZINB denotes the zero-inflated negative binomial distribution, ℓ_i is the observed library size for cell i, θ_g is a gene-specific inverse dispersion parameter, and f^η is a neural network whose outputs are interpreted as dropout probabilities. For cells with NTC guides, ContrastiveVI assumes the same generative process but with the salient latent variables fixed at a constant zero vector.

For inference, ContrastiveVI posits a variational distribution with parameters ϕ that factorizes as

$$q_\phi(z_i, t_i, \mid x_i) = q_{\phi_z}(z_i \mid x_i) q_{\phi_t}(t_i \mid x_i)$$

for cells with non-NTC guides. For cells with control guides, the above variational distribution is modified to account for the assumption that the salient variables do not contribute to the generative process, yielding

$$q_\phi(z_i, t_i, \mid x_i) = q_{\phi_z}(z_i \mid x_i) \delta\{t_i = 0\}.$$

In other words, the salient variables t_n are simply fixed at 0 during inference for cells with control guides. We refer to Weinberger et al. [6] for derivation of corresponding evidence lower bounds.

In our experiments we used the scverse [25] compatible implementation of ContrastiveVI available in the `scvi-tools` [22] package. For our experiments we used the same model architecture and optimization hyperparameters as described in the ContrastiveVI paper [6].

C.3 Mixscape

The Mixscape procedure proposed in Papalexi et al. [3] begins by computing a so-called perturbation signature for each cell infected with a non-NTC guide. To do so, for a given cell with a non-NTC guide, that cell's nearest neighbors in the control population are identified. The gene expression profiles of these nearest control neighbors are then averaged together and substracted from the gene expression profile of the original given non-NTC guide cell. The result of this procedure is then defined as a cell's perturbation signature.

After computing cells' perturbation signature, Mixscape then classifies cells with non-NTC guides as perturbed or escaping perturbation. To do so, for each targeted gene Mixscape fits a mixture of Gaussian models with two components on the corresponding cells' perturbation signatures. As escaping cells' signatures are assumed to be similar to those of control cells, one component of each of these mixtures is constrained to be equal to of a unimodal Gaussian fit to control cells.

For all results presented in this work, we used the R implementation of Mixscape in the Seurat package with default parameters. When computing UMAP embeddings and metrics on representation quality (e.g. mixing across cell cycle phases), we used the principal components of cells' perturbation signatures as done in Papalexi et al. [3].

D Metrics

Here we provide details on the quantitative metrics used in this work to compare ContrastiveVI+ against baseline methods.

D.1 Entropy of mixing

For c groups (e.g. cell cycle phases.) the entropy of mixing [26] is defined as

$$\sum_{i=1}^{c} p_i \log p_i,$$

where p_i denotes the proportion of cells from group i in a given region, such that $\sum_{i=1}^{c} p_i = 1$. Next, let U denote a uniform random variable over the population of cells. Let B_U then denote the empirical proportions of cells' groups in the 50 nearest neighbors of cell U. We report the entropy of this variables averaged over 100 random cells U. Higher values of this metric indicate stronger mixing of the c groups.

D.2 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. In our experiments we used the k means clustering algorithm to assign cluster labels to cells with k equal to the true number of clusters.