Learning Causal Representations of Single Cells via Sparse Mechanism Shift Modeling

Author names withheld

Editor: Under Review for CLeaR 2023

Abstract

Latent variable models such as the Variational Auto-Encoder (VAE) have become a go-to tool for analyzing biological data, especially in the field of single-cell genomics. One remaining challenge is the interpretability of latent variables as biological processes that define a cell's identity. Outside of biological applications, this problem is commonly referred to as learning disentangled representations. Although several disentanglement-promoting variants of the VAE were introduced, and applied to single-cell genomics data, this task has been shown to be infeasible from independent and identically distributed measurements, without additional structure. Instead, recent methods propose to leverage non-stationary data, as well as the sparse mechanism shift assumption in order to learn disentangled representations, with a causal semantic. Here, we explore and extend the application of these methodological advances to the analysis of single-cell genomics data with genetic or chemical perturbations. More precisely, we propose a deep generative model of single-cell gene expression data for which each perturbation is treated as a stochastic intervention targeting an unknown, but sparse, subset of latent variables. We benchmark these methods on simulated single-cell data to evaluate their performance at latent units recovery, causal target identification and out-of-domain generalization. Finally, we apply those approaches to two real-world large-scale gene perturbation data sets and find that models that exploit the sparse mechanism shift hypothesis surpass contemporary methods on a transfer learning task. We implemented our new model and benchmarks using the scvi-tools library, and released it as open-source software at www.placeholder.com*.

Keywords: non-linear ICA; deep generative models; variational inference; single-cell genomics; perturbation biology

1. Introduction

Machine learning methods have been key to gaining insights from large, high-dimensional genomic datasets, especially in single cell genomics (Ching et al., 2018). Variational Auto-Encoders (VAEs) (Kingma and Welling, 2014; Rezende et al., 2014), a recent approach in inferring complex data generative processes, are often well-suited for these applications, because they allow for flexible model design, while keeping necessary changes to the inference procedure relatively minimal. Many generative models have been proposed for analyzing diverse biological data modalities, including gene (RNA) expression, chromatin accessibility and quantitative protein measurements (Yau and Campbell, 2019; Lopez et al., 2020).

However, VAEs suffer from a key disadvantage due to their lack of interpretability, reflected as the absence of direct (justified) correspondence between individual latent variables and biological processes (Way and Greene, 2018). While disentanglement-promoting VAEs (e.g., as in Higgins et al. (2017)) can help better relate the two in genomics (Eraslan et al., 2022), these methods often compromise the quality of the latent variables for downstream tasks (Kimmel, 2020). This is perhaps

∗. Code to be released upon publication.

© 2023 .

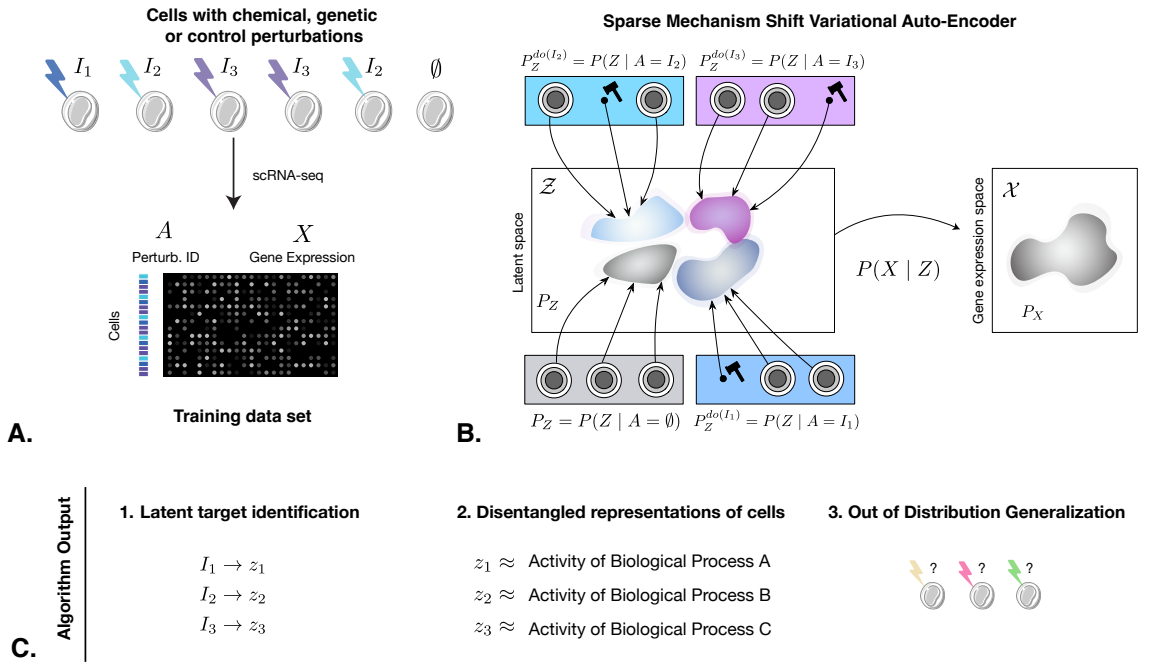

Figure 1: Overview of the sparse VAE framework applied to single-cell perturbation data. (A) Input data are gene expression profiles of cells under different genetic or chemical perturbations (colors), as well as the intervention label. (B) A schematic of the generative model, and the causal semantics of the sparse VAE (C) Three method outputs. (i) identification of target latent variables, encoded as a causal graph between the interventions and latent variables; (ii) a disentangled latent model for which individual latent variables are more likely to be interpreted as the activity of a relevant biological process; and (iii) the generalization of the generative model to unseen interventions (e.g., for latent target identification).

not surprising given recent theoretical developments showing that the recovery of ground truth latent variables is impossible from independent and identically distributed (i.i.d.) measurements (Locatello et al., 2019). In the remainder of this paper, we interchangably use the term disentanglement and latent variable recovery, following recent perspectives on non-linear Independant Component Analysis (ICA) (Locatello et al., 2019).

Recent efforts in disentanglement instead focus on the assumption of non-stationary data (Khemakhem et al., 2020), where data must be (i) observed in different regimes, with known pairing between data points and regimes, (ii) generated such that regimes are incurring changes in latent variables, and (iii) latent variables are conditionally independent given the regime. In this configuration, latent recovery with a conditional VAE (Sohn et al., 2015) is indeed theoretically possible. Follow-up work (Lachapelle et al., 2022) also draws connections to causal representation learning (Schölkopf, 2022), in which each regime may be modeled as an intervention targeting an unknown subset of latent variables.

Recent advances in biotechnology made non-stationary data increasingly available, especially in the context of genetic or chemical perturbation screens with single cell genomic profiles as a readout (Ji et al., 2021; Peidli et al., 2022). Recently, the Compositional Perturbation Autoencoder (CPA) (Lotfollahi et al., 2021) embedded perturbation profiles in latent space of an auto-encoder and predicted the effect of unseen combinations of single perturbations. However, this method neither exploits the non-stationary assumption in its probabilistic model nor refers to any identifiability guarantees for the latent space. Thus, to the best of our knowledge, there are so far no applications of the principles exposed in Khemakhem et al. (2020) and Lachapelle et al. (2022) to these new biological data types.

Here, we explore real-world applications of those principles (Figure 1). The promise is that causal models may eventually yield representations of perturbations and cells that are more mechanistically interpretable and more efficient for out-of-domain generalization. For example, the learned causal representations may lead to the delineation of biological processes (i.e., gene programs) that are affected by perturbations (Dixit et al., 2016). Furthermore, the model may also be used to project samples from unseen perturbations onto an existing atlas using transfer learning (Lotfollahi et al., 2022a).

After a brief introduction to disentangled representation learning and its intersection with causal inference (Section 2), we describe our motivation and effort to explore the learning of causal representations of perturbed cells profiled by single-cell RNA-seq (scRNA-seq) (Section 3). We also introduce sVAE+, a variant of the sVAE (Lachapelle et al., 2022) with a Bayesian approach for learning sparse mechanism shifts, and that requires minimal hyperparameter tuning. Then, we introduce a benchmarking tool for simulation of single-cell perturbation data and the evaluation of algorithms for latent units recovery, intervention target recovery, and transfer learning (Section 4). Finally, we present an application of the methods to two large-scale genetic screening experiments (Section 5). We show that models that exploit the sparse mechanism shift assumption outperform all methods by a significant margin at a transfer learning task. Our results suggest that causal inference is a promising paradigm for modeling the effects of perturbation in modern data sets from molecular biology.

2. Background

This paper is concerned with the recovery of latent variables that initially generated the data, as well as providing a causal semantic to those latent variables. Therefore, we briefly introduce non-linear Independent Component Analysis (ICA), one of the prominent methods for accomplishing this task, and its relationship to causal representation learning.

2.1. Non-linear Independent Component Analysis

ICA assumes that $x \in \mathbb{R}^d$ is generated using p independent latent variables $z = (z_1, \ldots, z_p)$, called *independent components* (Hyvärinen et al., 2002). More precisely, observations x are generated as $x = f(z) + \epsilon$ with f a mixing function and ϵ an exogenous noise variable. The ICA literature focuses on the identifiable case (e.g. if f is a linear function, then the original z may be recovered). In the general case of a non-linear *mixing function* f, however, the model is unidentifiable from i.i.d. observations of x (Hyvärinen and Pajunen, 1999). Given this negative result, several papers introduced identifiable forms of non-linear ICA models (Harmeling et al., 2003; Sprekeler et al., 2014; Hyvärinen and Morioka, 2016, 2017), based on the assumption that components $(z^i)_{i=1}^p$ are

conditionally independent given an *additional auxiliary* random variable $a \in \mathbb{R}^K$. Examples of auxiliary variables a include the past components in the case of time series analysis or some form of class label (Hyvärinen and Morioka, 2016). With the observation of auxiliary variables, latent recovery is possible up to a linear transformation under sufficient conditions (Hyvärinen and Morioka, 2016).

The iVAE framework (Khemakhem et al., 2020) proposes a VAE approach (Kingma and Welling, 2014; Rezende et al., 2014) for learning the parameters θ of a generative model $p_\theta(x \mid z, a)p_\theta(z \mid a)$, as well as ϕ, those of a variational approximation $q_\phi(z \mid x, a)$ to the posterior $p_\theta(z \mid x, a)$. The iVAE specifies $p_\theta(z \mid a) = \mathcal{N}(\mu_a, I)$ as a Gaussian location family with isotropic variance. As for the VAE, the parameters (θ, ϕ) of the iVAE are learned via maximization of the evidence lower bound (ELBO):

$$\log p_\theta(x \mid a) \geq \mathbb{E}_{q_\phi(z|x,a)} \log \frac{p_\theta(x, z \mid a)}{q_\phi(z \mid x, a)}. \tag{1}$$

2.2. Causal Inference from Unknown Interventions in Latent Space

Recent theoretical work (Lachapelle et al., 2022; Lachapelle and Lacoste-Julien, 2022) explores the assumption of sparse connections between the auxiliary variables $(a^l)_{l=1}^K$ and the latent components $(z^i)_{i=1}^p$, encoded in the form of a bipartite graph $G^a = ([K], [p], E)$, where E is the edge set. In the case where a describes a discrete data regime (i.e., via one-hot encoding), the sparsity pattern of G^a corresponds to the one of the mean vectors μ_a of $p_\theta(z \mid a)$.

The novel sparsity assumption allows for recovery of latent units up to a permutation, a much stronger result than the linear equivalence class recovered in Khemakhem et al. (2020). Perhaps as importantly, it also allows for the interpretation of the graph G^a from a causal perspective. More precisely, Lachapelle et al. (2022) considers the case where $a \in \{e_1, e_2, \ldots, e_K\}$, where each of e_l for $l \in [K]$ is a one-hot vector encoding the l-th intervention, and each intervention has unknown targets on the set of components of z. The unknown graph G^a describes which latent components are targeted by the intervention, that is $G^a_{i,l} = 1$ if and only if the l-th intervention targets z^i. In this context, the sparsity assumption corresponds precisely to the *sparse mechanism shift* hypothesis from Schölkopf (2022) i.e. that only a few mechanisms change at a time.

The VAE variant introduced in Lachapelle et al. (2022) (sparse VAE; sVAE) has an identical generative model as the one from the iVAE, except for the prior $p_\theta(z \mid a)$ for which a (stochastic) binary mask $\hat{G}^a_i \sim \text{Bernoulli}(\pi^a_i)$ is applied to the location parameter via element-wise product. The estimation procedure also relies on variational inference, with an addition of the regularization term $\alpha \|\pi^a_i\|_1$ to the ELBO, where α is a hyper parameter. To allow for gradient-based optimization of the objective function, sVAE uses the Gumbel-sigmoid distribution, a continuous relaxation of the Bernoulli distribution (Jang et al., 2017; Maddison et al., 2017).

3. A Sparse Mechanism Shift Model for Single Cell Measurements

For self-containment and context of this work, we first describe why interventions are a sensible model for single-cell pertubations, and then propose a new model for causal representation learning of those data.

3.1. Single-cell Perturbation Profiles as Interventional Data

Experimental advances in biology now allow us to actively intervene and change the properties of a single-cell by some genetic (Norman et al., 2019) or chemical (Srivatsan et al., 2020) perturbation and then simultaneously profile each individually perturbed cell for the identification of the perturbation and its molecular profile. A genetic perturbation may be induced by the delivery of a guide RNA in a cell expressing a CRISPR-associated protein (e.g., Cas9). Upon delivery, the CRISPR complex performs a genetic intervention (e.g., knock-out) at the location of the gene targeted by the RNA guide, altering the function of the gene, along with the associated protein. In the case of a chemical perturbation, cells are growth in the presence of a small molecule, which may enter the cell via the plasma membrane and interfere with one or several biochemical reactions. In both cases, interventions can affect (directly and indirectly) the expression of many genes (gene programs) that correspond to the activity of interpretable biological processes.

In such experiments, gene expression is profiled in each cell separately, after a fixed time, with single-cell RNA-seq (scRNA-seq) — a well-established technology used in diverse research areas of biology such as development (Semrau et al., 2017), autoimmunity (Gaublomme et al., 2015), and cancer (Patel et al., 2014). The measurements of a scRNA-seq experiment are summarized into a *gene expression matrix* $X = (X_1, \ldots, X_N) \in \mathbb{N}^{N \times d}$, with N the number of instances (cells), and d the number of genes. Individual entries of this matrix x_{ng} count the number of transcripts aligned to gene g in cell n. The same scRNA-seq assay also captures the identity of each intervention in each cells, often in the form of an RNA-expressed barcode sequenced alongside native gene expression levels. This information is summarized into an *intervention design matrix* $A \in \{0, 1\}^{N \times K}$, where K denotes the total number of treatments (Figure 1A).

3.2. Generative Model under Sparse Mechanism Shift

Perturbation Model We assume a fixed number p of latent variables $z = [z_1, \ldots, z_p]$, each representing the activity of a distinct biological process. Each intervention $a \in [K]$ targets latent variable $i \in [p]$ with probability

$$\pi_i^a \sim \text{Beta}\left(1, K\right),$$

such that the binary variable

$$\gamma_i^a \sim \text{Bernoulli}\left(\pi_i^a\right),$$

encodes whether latent variable i is targeted by the intervention. Then, latent variable z_i under intervention a is generated as a mixture distribution

$$z_i \mid a \sim \gamma_i^a \text{Normal}\left(\mu_i^a, 1\right) + (1 - \gamma_i^a)\text{Normal}\left(0, 1\right). \tag{2}$$

We note that the choice of the hyperpriors for the Beta distribution is sparsity inducing (Moran et al., 2022). Therefore, this generative model assumes that each perturbation should only affect a small subset of the latent variables. This hypothesis, inspired by Lachapelle et al. (2022), encourages the model to represent cells by the activity of biological processes that describe the effect of the perturbations.

Measurement Model For each single-cell $n \in [N]$, we measure the gene expression vector $x_n = [x_{n1}, \ldots, x_{ng}]$, as well as the perturbation information $a_n \in [K]$. Latent variables z_n are generated conditionally on a_n, following (2). The expected frequency of expression of gene g is calculated as

$$\rho_{ng} = f(z_n),$$

where f is a neural network with two hidden layers, 128 hidden units at each layer, ReLU non-linearity in between hidden layers. Notably, f also has a softmax non-linearity at its output, to allow for interpretation of its output as a frequency of expression. Finally, gene expression x_{ng} is generated as

$$x_{ng} \sim \text{NegativeBinomial}\left(l_n \rho_{ng}, \theta_g\right),$$

where l_n is the number of RNA transcript captured in cell n (referred to as library size, $l_n = \sum_g x_{ng}$), and θ_g are inverse-dispersion parameters. This measurement model is a simplication from the one described in Lopez et al. (2018), but still takes into account major technical factors of variation (count distribution, over-dispersion and library size).

Connection to prior work Existing non-linear ICA models such as the iVAE and the sVAE are not directly applicable to single-cell data, due to their unsuitable noise models. However, if we use the measurement model presented here, and change the prior on γ_i^a to be a point mass at 1, we obtain a model assimilated to the iVAE. Similarly, if we place a Laplace prior on π_i^a and perform MAP inference on γ_i^a, we obtain a model close to the sVAE. The model outlined here can be seen as a Bayesian treatment of the mechanism sparsity model in sVAE, and we therefore refer to it as sVAE+ in the remainder of this paper.

3.3. Variational Inference

The marginal probability of the data $p(x \mid a)$ is intractable. Therefore, we proceed to posterior approximation with variational inference to learn the model's parameters. We approximate the posterior distribution of each $\{\pi_i^a, \gamma_i^a, z_{ni}\}_{a \in [K], i \in [p], n \in [N]}$ with the mean-field variational distribution:

$$\bar{q} = \prod_{n \in [N]} q(z_n \mid x_n, a_n) \prod_{a \in [K], i \in [p]} q(\gamma_i^a) q(\pi_i^a). \tag{3}$$

As in VAEs, each $q(z_n \mid x_n, a_n)$ follows a Gaussian distribution with diagonal covariance matrix. The parameters of those distributions are encoded via neural networks. Each of the latent variables $q(\gamma_i^a)$ follows a Bernoulli distribution with free variational parameters. Finally, we use a point mass for each of $q(\pi_i^a) = \delta_{\psi_i^a}$, therefore performing MAP inference over this set of latent variables. We optimize the evidence lower bound (ELBO), derived as:

$$\mathbb{E}_{\bar{q}} \left[\sum_{n=1}^{N} \log \frac{p(x_n, z_n \mid \gamma_{a_n})}{q(z_n \mid x_n, a_n)} + \sum_{a \in [K], i \in [p]} \log \frac{p(\gamma_i^a, \pi_i^a)}{q(\gamma_i^a) q(\pi_i^a)} \right]. \tag{4}$$

This objective function is amenable to stochastic optimization, as in Kingma and Welling (2014). This allows us to sample a fixed number of data points at each iteration as well as from the variational distribution using the reparameterization trick and the Gumbel-sigmoid distribution for $q(\gamma_i^a)$. We provide further details for the derivation of this objective function in the Appendix A. We implemented sVAE+ as well as the other baselines within the `scvi-tools` codebase (Gayoso et al., 2022).

3.4. Downstream utilization of sVAE+

The sVAE+ model provides three main benefits (Figure 1C). First, the estimated graph G^a identifies which latent variables are affected by which perturbations. Edges in this graph are calculated by binarizing the matrix $(\pi_i^a)_{a\in[K],i\in[p]}$ at threshold 0.5. This helps in biological interpretation. For example, we can discover sets of perturbations that affect the same biological process. Second, the sparsity constraint on the latent space promises to help in disentanglement: identifying individual latent variables as distinct gene programs coordinately regulated by the perturbation, consistent with our understanding of the underlying organization of a cellular molecular circuits (Heimberg et al., 2016). Third, an important byproduct of the causal semantic is that one can reasonably expect the learned representations to perform better at downstream tasks such as transfer learning (Schölkopf, 2022), a growing use-case of deep generative models in single-cell genomics (Lotfollahi et al., 2022a). In this work, we present a synthetic case of holding out perturbations and performing target identification with a fixed generative model. However, we anticipate that transfer learning will be helpful in other concrete tasks, such that the projection of new cells onto perturbation atlases (Peidli et al., 2022) or the prediction of the outcome of perturbations across cellular contexts (e.g., different cell lines or tissues).

To the best of our knowledge, sVAE+ is the first proposed framework to explicitly model cellular perturbations as interventions on latent variables for the purpose of understanding causal mechanisms, while incorporating a model of experimental noise from single cell RNA-seq assays.

4. A Sandbox for Evaluation of Learned Representations

With sVAE+ being at the intersection of representation learning and single-cell biology, we consider it is important to present its alignment/compliance with both fields. To do so, we propose a sandbox for evaluation of current and future methods aimed at learning causal representations of single-cell data. Our sandbox consists of three main components: (i) simulation module for perturbed scRNA-seq data (ii) implementation of relevant baselines adapted for single-cell expression data and (iii) evaluation module including quantitative metrics for evaluation of learned representations.

4.1. Sandbox overview

Simulated data An important component of our sandbox is the simulation framework, that allows for systematic evaluation of the methods for all tasks on scRNA-seq data. We defer the details of the simulation framework in Appendix B, and briefly present here its main features. We simulated data from a deep generative model that takes into account common features of scRNA-seq data (count distribution, library size). We sampled cells from synthetic interventions on latent variables, targeting only a sparse subset of them. With these simulations, we have access to (i) ground truth latent variables, required to assess disentanglement (ii) true interventional targets in latent space for each perturbation, required to assess target identification and (iii) simulation of out-of-domain samples. In this last case, these simulations provide a way to report performance under more complex settings (different sparsity rate or effect size between train and test data).

Baselines Because practitioners rely on disentanglement methods with the hope for recovering more insightful representations of single cell data, we include as baselines the standard vanilla VAE (Kingma and Welling, 2014) and its popular derivative β-VAE (Higgins et al., 2017), both exploited by the biology community (Lopez et al., 2018; Eraslan et al., 2022). We also include the

Table 1: Mean and standard deviation per metric on simulations for $d = 15$. Best in bold.

	Disentanglement			Causal discovery			OOD
	Pearson MCC (↑)	**Spearman MCC** (↑)	$\mathbf{R^2}$ (↑)	**Precision** (↑)	**Recall** (↑)	**F1** (↑)	**Inter. NLL** (↓)
VAE	0.46 ± 0.02	0.40 ± 0.02	0.82 ± 0.00	-	-	-	327.10 ± 2.81
β-VAE	0.48 ± 0.02	0.42 ± 0.02	0.83 ± 0.00	-	-	-	392.68 ± 4.58
iVAE	0.47 ± 0.01	0.39 ± 0.00	0.85 ± 0.00	0.13 ± 0.01	$\mathbf{1.00 \pm 0.00}$	0.22 ± 0.00	323.26 ± 2.88
sVAE	0.72 ± 0.14	0.63 ± 0.13	0.84 ± 0.01	0.43 ± 0.19	0.39 ± 0.37	0.31 ± 0.13	318.26 ± 3.86
sVAE+	$\mathbf{0.88 \pm 0.04}$	$\mathbf{0.79 \pm 0.04}$	$\mathbf{0.86 \pm 0.01}$	$\mathbf{0.54 \pm 0.10}$	0.47 ± 0.09	$\mathbf{0.51 \pm 0.09}$	$\mathbf{315.43 \pm 2.46}$

iVAE (Khemakhem et al., 2020) as well as the sVAE (Lachapelle et al., 2022), both explicitly aiming at latent units recovery. All baselines were adjusted to account for single-cell readouts within the unified framework of `scvi-tools` (details appear in Appendix C).

Evaluation metrics Based on the ground truth provided by our simulation framework, we evaluate the performance of all methods in terms of disentanglement, causal structure learning, and transferability. For assessing disentanglement, we report the Mean Correlation Coefficient (MCC), a metric for permutation equivalence proposed by Khemakhem et al. (2020) that measures the average Pearson (or Spearman) correlation coefficients between pairs of ground truth and estimated latent variable, for the best possible permutation. A high MCC means that we successfully identified the true parameters and recovered the true sources. We also report R^2, a metric for assessing identifiability up to a linear transformation. To evaluate the learned causal structure, we report the precision, recall and F1 score of the learned adjacency matrix of $\hat{G}^a$ compared to the ground truth G^a, taking into account the permutation equivalence of z. Finally, to assess transferability, we report the negative log-likelihood of data points from holdout perturbations (Interventional NLL) after fine-tuning the embeddings, while keeping the generative model fixed (Gentzel et al., 2019). This likelihood is approximated using the importance weighted ELBO with $5,000$ particles (Burda et al., 2016).

4.2. Experiments with synthetic data

As part of the empirical evaluation, we explored the performance of all methods across several values of the latent space dimension ($d \in \{5, 10, 15, 20\}$). In all cases, we simulated data subject to $K = 100$ interventions, with 500 cells sampled per intervention. Measurements from 20 interventions were held-out from the training set, and used for transfer learning evaluation. All configurations were ran for 5 different random initialization. We selected the optimal hyper-parameters of each method (number of epochs, β, and sparsity penalty α) using the Unsupervised Disentanglement Ranking (UDR) framework (Duan et al., 2020) (hyper parameter grids appear in Appendix C). The number of latent variables of the generative model was fixed to the one used in the simulation.

We present the results for all methods across different metrics for $d = 15$ in Table 1. An extended, systematic overview for different setups can be found in Appendix D. All in all, our takeaways from the synthetic experiments are as follows:

- **Disentanglement** All methods achieve high R^2 score, which corresponds to a satisfactory recovery of the latent units, *up to a linear transformation*. However, the iVAE, sVAE and sVAE+ compare favorably to the standard VAE and β-VAE at the task of latent units recovery *up to a permutation*, with sVAE+ being the strongest performer.

- **Causal structure recovery** The correct structure of the causal graph can be best recovered with sVAE+, compared to sVAE and iVAE.
- **Out-of-distribution generalisation** sVAE and sVAE+ learning latent representations that better model data from held-out interventions, as evaluated by the interventional NLL.

To summarize, the proposed sandbox lets us better understand how different scenarios match with each baselines, which can be leveraged depending on the application a practitioner has in mind (whether the end goal is disentanglement, causal recovery or out-of-distribution generalization). In addition, we also notice sVAE+, that includes Bayesian treatment of the mechanism sparsity model, outperforms all VAE variants across all three objectives, without the need for tuning the sparsity hyper prior (Appendix D.4).

5. Empirical Evaluation on Real Datasets

We apply our benchmark models to real-world data from two recent large-scale Perturb-seq experiments (Norman et al., 2019; Replogle et al., 2022). Further details about the processing of these data sets appear in Appendix E.

5.1. In-domain Transferability and Interpretability on a Genetic Screen

In the dataset from Norman et al. (2019), $105,528$ cells from an erythrocytic leukemia cell line (K562) were profiled after interventions targeting one or two of 112 genes, including cell cycle regulators, transcription factors, kinases, phosphatases, and genes of unknown function. After quality control and data filtering, we retained $96,221$ cells of undergoing 212 different genetic interventions and $8,907$ unperturbed (control) cells. Due to experimental limitations (Dixit et al., 2016), we observe signal only for a subset of several thousand genes (here, $d = 3,000$). The goal of the experiment was to understand the mechanisms of genetic interactions and recover gene regulatory logics. In order to simulate a transfer learning scenario, we selected the top-30 interventions with the most significant effect on gene expression, as assessed by the maximum mean discrepency (Gretton et al., 2012) estimated with a linear kernel on a PCA with dimension 50, and held out the corresponding cells as a test set. Because we did not hold out interventions according to biological knowledge, but simply based on effect size across all interventions, we qualify this benchmark of "in-domain".

We applied each studied method to this dataset. Without ground truth, we use datapoints from held-out interventions to evaluate the models after transfer learning. For all models, we report the negative ELBO evaluated on a validation data set (val. nELBO; including additional cells with the same perturbations as in the training data), as well the Interventional NLL in Table 2. All of iVAE, sVAE and sVAE+ improve data fit compared to the VAE and β-VAE, as measured by the validation negative ELBO. However, while the iVAE provides the best fit to the validation set, with a thin margin, it fits the test set poorly compared to sVAE and sVAE+. This suggests that sVAE and sVAE+ have stronger transfer capabilities compared to other methods, and learn a more causal representation of the data.

Table 2: Results on the data set from Norman et al. (2019)

	val. nELBO	Interv. NLL
VAE	716.5	776.4
β-VAE	718.9	776.7
iVAE	**715.1**	777.8
sVAE	715.2	776.0
sVAE+	716.1	**775.1**

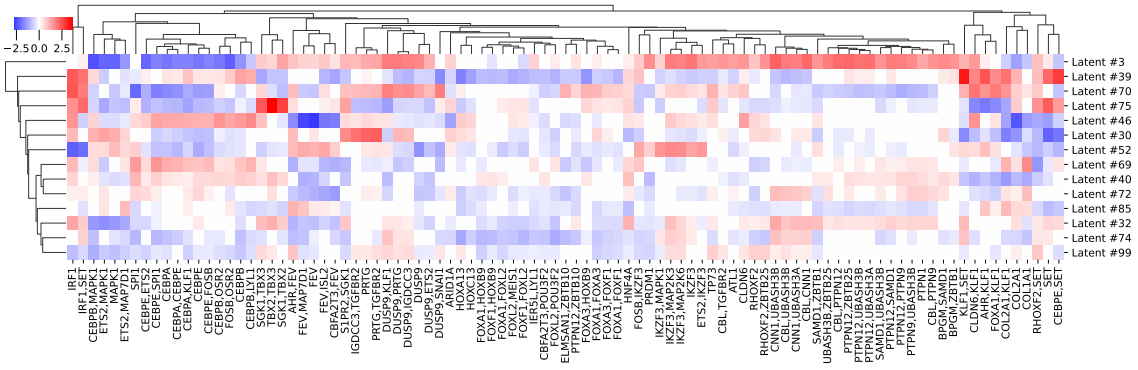

Figure 2: Perturbation effects on latent components (subset of perturbations, and components).

We also performed a preliminary biological interpretation of the sparse VAE model. We first visualized the effect of perturbations in latent space in the form of a weighted adjacency matrix W_{ij} for G^a, where the weight encodes the shift in the mean of the corresponding latent component z_i for perturbation j (Figure 2). For visual convenience we focused on a subset of perturbations and latent components to retain the most informative data. Briefly, many perturbations have similar effects on latent variables, as it has been observed previously at the level of individual genes and programs (Dixit et al., 2016). This can be interpreted as the perturbed genes being a part of the same pathway. Indeed, perturbations involving the same gene (in different combinations) grouped together by their shared effect on latent factors, as did those involving different genes from related pathways (e.g. FOXO and homeobox transcriptions factors affecting cell differentiation), whereas perturbations in genes from different pathways had different effects (e.g. IRF1 vs. CEBP family transcription factors). Additional biological interpretation of the model appears in Appendix F.

5.2. Out-of-Domain Transferability on a Genome-wide Genetic Screen

In order to assess more systematically the transferability of the generative model learned with our benchmark methods, across distinct biological pathways, we now consider the larger data set from Replogle et al. (2022). The original data has around two million cells, after interventions targeting one of around ten thousand genes. We focused on a subset of interventions with large effect on gene expression ($K = 683$), resulting in $N = 116,641$ cells profiled. As in the previous data set, we observe signal only for around a thousand genes (here, $d = 1,187$). Those interventions have been carefully annotated by experts into 63 clusters, of which 8 contained at least 20 distinct interventions and were matched to a known biological pathway. We applied our transfer learning benchmark, treating each pathway as a set of held-out perturbations (data splits are detailed in Appendix E), and report the interventional NLL in Table 3. In most scenarios, sVAE+ outperforms all methods with this metric. Again, this suggests that sVAE+ learns more causal representation of cells.

Table 3: Results on the data set from Replogle et al. (2022). Each entry reports the interventional likelihood on held-out interventions for a method (column) trained on the full data set with one pathway (row) hold-out.

	VAE	**β-VAE**	**iVAE**	**sVAE**	**sVAE+**
Exosome and mRNA turnover	1784.52	1786.02	1784.11	1784.41	**1782.48**
Spliceosome	1779.36	1782.87	1779.38	1779.43	**1777.16**
Mediator complex	1551.56	1553.17	**1550.90**	1551.50	1551.26
TFIIH/nucleotide excision repair	1653.27	1659.38	1653.25	1653.35	**1652.88**
39S ribosomal subunit, mitochondrial	1778.80	1780.09	1778.79	1788.08	**1777.61**
60S ribosomal subunit, cytoplasmic	**1889.73**	1892.89	1890.53	1890.39	1890.14
40S ribosomal subunit, cytoplasmic	1841.54	1844.55	1842.02	1841.95	**1840.03**
Mitochondrial protein translocation	1659.18	1661.01	1660.12	1659.85	**1658.31**

6. Related Work

Disentanglement in Variational Auto-Encoders In order to quantify, and potentially improve upon the poorly interpretable latent variables inferred by VAEs, Higgins et al. (2017) created (i) a dataset with ground truth factors, (ii) metrics to quantify the mutual information between the inferred and the true factors and (iii) a modification of the VAE that outperforms existing methods with respect to these metrics. The proposed modification, the β-VAE, consists in scaling the Kullback-Leibler divergence term in the evidence lower bound with a scalar $\beta > 1$. Notably, this line of work does not necessarily associate disentanglement with the exact recovery of latent variables (unlike this paper), but is more largely concerned by the conservation of coherent axes of variation in the data. Several research groups proposed novel algorithms, such as the β-TCVAE (Chen et al., 2018) or the Factor-VAE (Kim and Mnih, 2018), while others improved benchmarking and metrics (Eastwood and Williams, 2018; Duan et al., 2020).

Identifiable Models from Unsupervised Data Although deep generative models are in general unidentifiable, some recent work proposed different sets of assumptions for which recovery of latent variables is possible. For example, Moran et al. (2022) proposed a VAE model for which each feature may be generated from only a sparse set of latent factors. Such model is proven to be identifiable under an anchor assumption, that stipulates that for each latent factor, there exists at least two features depending only on that factor. This hypothesis is also relevant to the field of single-cell genomics, for which a few genes may be reasonably expected to be part of only one biological process. More recent work (Kivva et al., 2022) provides an identifiability result for unsupervised deep generative models under a mixture model prior.

Supervised Generative Models Both of the iVAE and the sVAE are particular instances of conditional deep generative models (Sohn et al., 2015). These models place the auxiliary variable a "upstream" of the latent factors z in the corresponding graphical model, in the form of a conditional prior $p(z \mid a)$. A different modeling choice would consist in placing the auxiliary variable "downstream" of z, an idea commonly exploited in supervised topic models (Mcauliffe and Blei, 2007). Such modeling hypotheses have limited causal semantic, but may still be useful for recovering

biological processes that are helpful for analyzing perturbation data, because the supervision may effectively guide inference towards topics that are more reflective of those perturbations.

Causal Structure Learning Although here we consider the setting of learning from interventional data with targets being latent variables (and therefore unknown), a related line of work is concerned with learning causal relationships at the level of features, possibly under interventions (Wang et al., 2017). These methods are in principle applicable to large-scale genetic screens (Lopez et al., 2022), because by design we know which genes are targeted by each intervention. However, current algorithms for causal structure learning are very limited in their ability to handle internal cellular states (confounding factors), a task for which latent variable models are more suitable. Therefore, this paper presents an alternative modeling choice for this type of interventional data.

7. Discussion

We propose to explicitly model perturbations in single-cell genomics as interventions on a latent space, with a causal semantic. This naturally leads to the application of the sVAE (Lachapelle et al., 2022) and iVAE (Khemakhem et al., 2020) framework, as well as our proposed sVAE+ model, to disentangle the latent space of single cell data by leveraging additional knowledge of perturbations. We provide a benchmarking framework for assessing the performance of the learned representations in terms of level of disentanglement, causal target identification, as well as transfer learning. In simulated data, both approaches outperform the β-VAE and the vanilla VAE, with a strong advantage for sVAE+, explicitly assuming sparsity in mechanism shifts for each perturbation. We also applied all methods to two real data sets. Our analysis suggests that sparsity may help in transfer learning and interpretation of latent variables. Importantly, we see in Figure 2 that multiple latent variables are affected by each intervention, which suggests more informative constraints to the model could be added to further improve its interpretability (Lotfollahi et al., 2022b).

The hypotheses from the sVAE+ model in its current state, however, may present a few limitations for biological applications. Importantly, although it may be reasonable to expect that genetic (and often chemical) interventions directly trigger a sparse subset of a cell's circuitry (e.g., blocking a single pathway (Dixit et al., 2016)), there are important molecular feedback mechanisms that can induce indirect downstream effects in other pathways, especially as increasing time passes from the initial perturbation (Freimer et al., 2022). Because many experiments measure gene expression from hours to days after intervention, sparsity may be a limiting assumption without resolving interactions between pathways, discussed at length in biological application of traditional causal discovery learning methods (Segal et al., 2005; Friedman et al., 2000; Pe'er et al., 2001). This issue could be resolved in the future with the potential availability of time-resolved measurements from single-cell perturbation experiments. We also discuss more technical assumptions of Lachapelle et al. (2022), such as covering of latent variable by the interventions, and sufficiently variability in Appendix G.

To conclude, we present a first attempt at leveraging the sparse mechanism shift assumption for the purpose of learning causal representations as well as interpretable models of single-cell perturbation data. Looking forward, we anticipate that this line of work may unlock new perspectives to reason about perturbations as interventions on the genetic and molecular circuits that govern a cell's identity (Wagner et al., 2016). This perspective is especially important for enhancing our understanding of biological processes leading to disease states, as well as proposing candidate drugs.

References

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877, 2020.

Yuri Burda, Roger B. Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *International Conference on Learning Representations*, 2016.

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.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141):20170387, 2018.

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.

Sunny Duan, Loic Matthey, Andre Saraiva, Nicholas Watters, Christopher P Burgess, Alexander Lerchner, and Irina Higgins. Unsupervised model selection for variational disentangled representation learning. In *International Conference on Learning Representations*, 2020.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

Gökcen Eraslan, Eugene Drokhlyansky, Shankara Anand, Evgenij Fiskin, Ayshwarya Subramanian, Michal Slyper, Jiali Wang, Nicholas Van Wittenberghe, John M Rouhana, Julia Waldman, et al. Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science*, 376(6594), 2022.

Jacob W Freimer, Oren Shaked, Sahin Naqvi, Nasa Sinnott-Armstrong, Arwa Kathiria, Christian M Garrido, Amy F Chen, Jessica T Cortez, William J Greenleaf, Jonathan K Pritchard, et al. Systematic discovery and perturbation of regulatory genes in human T cells reveals the architecture of immune networks. *Nature Genetics*, 54(8):1133–1144, 2022.

Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601–620, 2000.

Jellert T Gaublomme, Nir Yosef, Youjin Lee, Rona S Gertner, Li V Yang, Chuan Wu, Pier Paolo Pandolfi, Tak Mak, Rahul Satija, Alex K Shalek, et al. Single-cell genomics unveils critical regulators of Th17 cell pathogenicity. *Cell*, 163(6):1400–1412, 2015.

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.

Amanda Gentzel, Dan Garant, and David Jensen. The case for evaluating causal models using interventional measures and empirical data. In *Advances in Neural Information Processing Systems*, pages 11722–11732, 2019.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Stefan Harmeling, Andreas Ziehe, Motoaki Kawanabe, and Klaus-Robert Müller. Kernel-based nonlinear blind source separation. *Neural Computation*, 15(5):1089–1124, 2003.

Graham Heimberg, Rajat Bhatnagar, Hana El-Samad, and Matt Thomson. Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell Systems*, 2(4):239–250, 2016.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations*, 2017.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. *Advances in Neural Information Processing Systems*, 29, 2016.

Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469, 2017.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. Independent component analysis. *Studies in Informatics and Control*, 11(2):205–207, 2002.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-softmax. *International Conference on Learning Representations*, 2017.

Yuge Ji, Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. Machine learning for perturbational single-cell omics. *Cell Systems*, 12(6):522–537, 2021.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217, 2020.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658, 2018.

Jacob C Kimmel. Disentangling latent representations of single cell RNA-seq experiments. *bioRxiv*, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *International Conference on Learning Representations*, 2014.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models under mixture priors without auxiliary information. In *Advances in Neural Information Processing Systems*, 2022.

Sébastien Lachapelle and Simon Lacoste-Julien. Partial disentanglement via mechanism sparsity. *Conference on Uncertainty and Artificial Intelligence: Causal Representation Learning workshop*, 2022.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Conference on Causal Learning and Reasoning*, pages 428–484, 2022.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124, 2019.

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.

Romain Lopez, Adam Gayoso, and Nir Yosef. Enhancing scientific discoveries in molecular biology with deep generative models. *Molecular Systems Biology*, 16(9):e9198, 2020.

Romain Lopez, Jan-Christian Hütter, Jonathan K Pritchard, and Aviv Regev. Large-scale differentiable causal discovery of factor graphs. In *Advances in Neural Information Processing Systems*, 2022.

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Yuge Ji, Ignacio L Ibarra, F Alexander Wolf, Nafissa Yakubova, Fabian J Theis, and David Lopez-Paz. Compositional perturbation autoencoder for single-cell response modeling. *bioRxiv*, 2021.

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

Mohammad Lotfollahi, Sergei Rybakov, Karin Hrovatin, Soroor Hediyeh-zadeh, Carlos Talavera-López, Alexander Misharin, and Fabian J Theis. Biologically informed deep learning to infer gene program activity in single cells. *bioRxiv*, 2022b.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *International Conference on Learning Representations*, 2017.

Jon Mcauliffe and David Blei. Supervised topic models. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Gemma E. Moran, Dhanya Sridhar, Yixin Wang, and David M. Blei. Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*, 2022.

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.

Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, 2014.

Stefan Peidli, Tessa Durakis Green, Ciyue Shen, Torsten Gross, Joseph Min, Jake Taylor-King, Debora Marks, Augustin Luna, Nils Bluthgen, and Chris Sander. scPerturb: Information resource for harmonized single-cell perturbation data. *bioRxiv*, 2022.

Dana Pe'er, Aviv Regev, Gal Elidan, and Nir Friedman. Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, 17, 2001.

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*, 2022.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286, 2014.

Peter J Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

Bernhard Schölkopf. Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 765–804. Association for Computing Machinery, 2022.

Eran Segal, Dana Pe'er, Aviv Regev, Daphne Koller, and Nir Friedman. Learning module networks. *Journal of Machine Learning Research*, 6(4), 2005.

Stefan Semrau, Johanna E Goldmann, Magali Soumillon, Tarjei S Mikkelsen, Rudolf Jaenisch, and Alexander Van Oudenaarden. Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. *Nature Communications*, 8(1):1–16, 2017.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in Neural Information Processing Systems*, 28, 2015.

Henning Sprekeler, Tiziano Zito, and Laurenz Wiskott. An extension of slow feature analysis for nonlinear blind source separation. *The Journal of Machine Learning Research*, 15(1):921–947, 2014.

Sanjay R Srivatsan, José L McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A Pliner, Dana L Jackson, Riza M Daza, Lena Christiansen, et al. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328, 2017.

Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nature biotechnology*, 34(11):1145–1160, 2016.

Martin J Wainwright. *High-dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

Yuhao Wang, Liam Solus, Karren Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. *Advances in Neural Information Processing Systems*, 30, 2017.

Gregory P Way and Casey S Greene. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. In *Pacific Symposium on Biocomputing*, pages 80–91, 2018.

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

Christopher Yau and Kieran Campbell. Bayesian statistical learning for big data biology. *Biophysical Reviews*, 11(1):95–102, 2019.

Xiuwei Zhang, Chenling Xu, and Nir Yosef. Simulating multiple faceted variability in single cell rna sequencing. *Nature Communications*, 10(1):1–16, 2019.

Appendices

The appendices are organized as follows. In Appendix A, we provide the mathematical derivations for our extension of the sVAE framework, named sVAE+. In Appendix B, we detail our simulation framework. In Appendix C, we explain the baselines used in this study, and the hyperparameter grid used for reporting the experimental results. In Appendix D, we provide supplementary experimental results. In Appendix E, we report the preprocessing steps for the real-world data sets. In Appendix F, we provide biological interpretation of the sVAE+ results on the Norman data set. In Appendix G, we discuss the technical assumptions from the work of Lachapelle et al. (2022), and the implication it has in this study.

Appendix A. Derivation of the Evidence Lower Bound for sVAE+

We now derive the ELBO that is used as the objective function for inference within the sVAE+ framework. Let us note $\boldsymbol{X} = [x_n]_{n=1}^N$ and $\boldsymbol{A} = [a_n]_{n=1}^N$. We also adopt vector notation for the latent variables $\boldsymbol{\gamma} = [\gamma_i^a]_{a\in[K],i\in[p]}$, $\boldsymbol{\pi} = [\pi_i^a]_{a\in[K],i\in[p]}$ and $\boldsymbol{Z} = [z_n]_{n=1}^N$. We remind the reader that the likelihood of the data may be written as an intractable integral over the (latent) random variables:

$$\log p(\boldsymbol{X} \mid \boldsymbol{A}) = \log \prod_{n=1}^N p(x_n \mid a_n) \tag{5}$$

$$= \log \iiint \prod_{n=1}^N p(x_n, z_n, \boldsymbol{\gamma}, \boldsymbol{\pi} \mid a_n) d\boldsymbol{Z} d\boldsymbol{\gamma} d\boldsymbol{\pi} \tag{6}$$

$$= \log \iiint \prod_{n=1}^N p(x_n \mid z_n) p(z_n \mid \gamma_{a_n}) \prod_{a\in[K]} \prod_{i\in[p]} p(\gamma_i^a \mid \pi_i^a) p(\pi_i^a) d\boldsymbol{Z} d\boldsymbol{\gamma} d\boldsymbol{\pi}. \tag{7}$$

Now, we remind the reader of the variational distribution

$$\bar{q} = \prod_{n\in[N]} q(z_n \mid x_n, a_n) \prod_{a\in[K],i\in[p]} q(\gamma_i^a) q(\pi_i^a). \tag{8}$$

To derive the evidence lower bound, we start by weighting the integrand by the variational distribution as follows:

$$\log p(\boldsymbol{X} \mid \boldsymbol{A}) = \log \mathbb{E}_{\bar{q}} \left[\prod_{n=1}^N \frac{p(x_n \mid z_n) p(z_n \mid \gamma_{a_n})}{q(z_n \mid x_n, a_n)} \prod_{a\in[K]} \prod_{i\in[p]} \frac{p(\gamma_i^a \mid \pi_i^a) p(\pi_i^a)}{q(\gamma_i^a) q(\pi_i^a)} \right], \tag{9}$$

and then use the concavity of the natural logarithm to apply Jensen inequality:

$$\log p(\boldsymbol{X} \mid \boldsymbol{A}) \geq \mathbb{E}_{\bar{q}} \log \left[\prod_{n=1}^N \frac{p(x_n \mid z_n) p(z_n \mid \gamma_{a_n})}{q(z_n \mid x_n, a_n)} \prod_{a\in[K]} \prod_{i\in[p]} \frac{p(\gamma_i^a \mid \pi_i^a) p(\pi_i^a)}{q(\gamma_i^a) q(\pi_i^a)} \right]. \tag{10}$$

From this follows the celebrated evidence lower bound (ELBO):

$$\log p(\boldsymbol{X} \mid \boldsymbol{A}) \geq \mathbb{E}_{\bar{q}} \left[\sum_{n=1}^N \log \frac{p(x_n, z_n \mid \gamma_{a_n})}{q(z_n \mid x_n, a_n)} + \sum_{a\in[K],i\in[p]} \log \frac{p(\gamma_i^a \mid \pi_i^a) p(\pi_i^a)}{q(\gamma_i^a) q(\pi_i^a)} \right]. \tag{11}$$

Exploiting the analytical expressions for (i) the Kullback-Leibler divergence between two multivariate Gaussian distributions, (ii) the Kullback-Leibler divergence between two Bernoulli distribution and (iii) the simplification of the expression resulting from $q(\boldsymbol{\pi}) = \delta_{\boldsymbol{\psi}}$ being a Dirac distribution, we get the final objective function:

$$\log p(\boldsymbol{X} \mid \boldsymbol{A}) \geq \mathbb{E}_{\bar{q}} \left[\sum_{n=1}^{N} \log p(x_n \mid z_n) - D_{KL}(q(z_n \mid x_n, a_n) \parallel p(z_n \mid \gamma_{a_n})) \right] \tag{12}$$

$$- \sum_{a \in [K], i \in [p]} D_{KL}(q(\gamma_i^a) \parallel \text{Bernoulli}(\psi_i^a)) - \log \text{Beta}(\psi_i^a; 1, K). \tag{13}$$

The last technical challenge remains in applying the reparameterization trick to $q(\boldsymbol{\gamma})$, for which we use the Gumbel-sigmoid distribution (Maddison et al., 2017), as a continuous relaxation of the Bernoulli distribution. As written, the objective function may now be implemented in pytorch, and `scvi-tools` (Gayoso et al., 2022).

Appendix B. Simulation Details

We simulate single-cell gene expression profiles from perturbation experiments as follows. We assume we have measurements from N cells. Each of the cells, for example cell $n \in [N]$, has been exposed to a perturbation/intervention $a_n \in [K]$. We use $K = 100$ interventions in our simulations, and sample $N = 100,000$ cells in total. The first 80 interventions form the training set, and the last 20 form the test set. Each intervention $a \in [K]$ is represented by a sparse perturbation embedding $\mu_a \in \mathbb{R}^p$, where $p = 15$ is the dimension of the embedding.

For each intervention, we treat the number of affected latent dimensions $t_a = \{1, 2, 3\}$ uniformly at random. The indices of affected dimensions are also drawn without replacement from $[p]$, encoded into a binary vector $\beta_{a,\cdot} \in \{0, 1\}^p$. Finally, each component $\mu_{a,i}$ of $\mu_a = (\mu_{a,1}, \ldots, \mu_{a,p})$ for $i \in [p]$ is generated as:

$$\eta_{a,i} \sim \frac{1}{2}\text{Normal}(-e, 0.5) + \frac{1}{2}\text{Normal}(e, 0.5) \tag{14}$$

$$\mu_{a,i} \sim (1 - \beta_{a,i})\delta_0 + \beta_{a,i}\eta_{a,i}, \tag{15}$$

where e is a scalar quantifying the strength of the intervention, and δ_0 designates the Dirac delta distribution with mass at 0. For cell n, exposed to intervention a_n, latent variable z_n is generated as:

$$z_n \sim \text{Normal}(\mu_a, I), \tag{16}$$

where each individual component of z_n encodes the activity of a pathway, shifted by the intervention. Measurements x_{ng} from a single cell n and a gene g are generated from a count distribution:

$$x_{ng} \sim \text{Poisson}\left(l_n f^g(z_n)\right), \tag{17}$$

where l_n is the library size fixed to 10^5, and the mixing function f is a neural network with three hidden layers of 40 units, Leaky-ReLU activations with a negative slope of 0.2, and a softmax non-linearity on the last layer to convert the outputs to counts (Lopez et al., 2018). The weight matrices of f are sampled according to an isotropic Gaussian distribution, with orthogonal columns, to make

sure f is injective (Lachapelle et al., 2022). Although we used a Poisson distribution for simplicity, future work will investigate the use of more realistic count distributions such as Beta-Poisson, or negative binomial, that may be important to model over-dispersion of scRNA-seq data (Zhang et al., 2019).

Appendix C. Details for Empirical Evaluation of the Benchmark Methods

In this section, we provide the necessary details for reproducing the experiments in the paper.

C.1. Baseline and Metric Details

Shared components of the generative model All models rely on the generative model described in Section 3. For the VAE and β-VAE, $\boldsymbol{\gamma} = \mathbf{0}$ while for iVAE, $\boldsymbol{\gamma} = \mathbf{1}$. Similarly, the sVAE is implemented with a Laplace prior on $\boldsymbol{\pi}$ as in the original publication. The neural network architecture for f in the generative model is given in section B.

Shared components of the inference The mean and variance of the variational distribution for z are each obtained as the output of a neural network with two hidden layers, 128 hidden units at each layer and ReLU non-linearity in between hidden layers, taking as input the observed data x. We optimize the ELBO using the Adam optimizer (Kingma and Ba, 2015) with minibatches of size 128.

VAE and β-VAE Since both of these models do not have the ability to use the interventions while modeling the data, in our implementation we let them use all the available data and simply ignore the covariate a.

iVAE Unlike sVAE and sVAE+ which can estimate a causal graph, iVAE only outputs estimates for the parameters $\boldsymbol{\mu}$ (effect of each intervention on each latent variable). In order to compare those three baselines, we need a heuristic to build an adjacency matrix from the estimates of $\boldsymbol{\mu}$. In this benchmark, we use the outlier detection method from Rousseeuw and Driessen (1999), based on the idea that for each intervention, we seek to find the latent variables with most significant sensitivity to it. With such heuristic we ensure to have a sparse graph, based on the most meaningful factors.

Transfer Learning Details For the transfer learning experiments, we fix the parameters of the generative model, as well as the ones of the variational network. For VAE and β-VAE, we report the importance weighted ELBO (IWELBO) for the hold-out interventions. For iVAE, sVAE, and sVAE+, we proceed to a fine-tuning step with all of the above parameters frozen, except for the parameter $\boldsymbol{\mu}$. In this fine-tuning step, we fix the variables $\boldsymbol{\gamma}$ to $\mathbf{1}$ to avoid the sparsity prior to constrain the model in fitting the test set. The IWELBO (Burda et al., 2016) with T particles is calculated as:

$$\text{IWELBO}^T(\boldsymbol{X} \mid \boldsymbol{A}) = \sum_{n=1}^{N} \mathbb{E}_{z_n^1,\dots,z_n^T \sim q(z_n \mid x_n, a_n)} \left[\log \frac{1}{T} \sum_{t=1}^{T} \frac{p(x_n, z_n^t \mid \gamma_{a_n} = 1)}{q(z_n^t \mid x_n, a_n)} \right]. \tag{18}$$

C.2. Hyper parameters Grid for the Simulated Data

For all methods, we performed an exhaustive hyper parameter grid search. We used the Unsupervised Disentanglement Ranking (UDR) framework (Duan et al., 2020) for selecting the optimal hyper parameters. The complete hyper parameter search space for each algorithm is described in Table 4.

Table 4: Grid search spaces for each baseline.

	Hyperparameter space
VAE	$n_{\text{epochs}} \in \{300, 500\}$
β-**VAE**	$n_{\text{epochs}} \in \{300, 500\}$, $\beta \in \{8, 10, 30\}$
iVAE	$n_{\text{epochs}} \in \{300, 500\}$
sVAE	$n_{\text{epochs}} \in \{300, 500\}$, $\alpha \in \{10, 20, 40, 60, 80, 100\}$
sVAE+	$n_{\text{epochs}} \in \{300, 500\}$ $\alpha \in \{10, 20, 40, 60, 80, 100\}$

Appendix D. Additional Results from Synthetic Experiments

For quantitative comparison of the considered baseline methods, we leverage the availability of ground truth latent variables in simulated data sets. In the proposed sandbox in Section 4, we have control over: the size of the effect for interventions - e_a, the sparsity of the interventions - t_a, as well as the dimensionality of the latent and auxiliary variables. The aim of this paper is to evaluate the extent to which sparse, identifiable methods provide disentangled, or ultimately "generalizable/transferable" representations. To systematically assess the quality i.e. transferability of the learned representations, we propose the following scenarios:

- transfer to unseen, in-domain interventions (hold-out interventions form a data set with same level of sparsity and effect size);
- transfer to unseen out-of domain interventions (hold-out interventions form a data set with different level of sparsity or effect size).

D.1. Results for different latent dimensions

We present the disentanglement results for alternate number of latent dimensions in our simulations in Table 5. The conclusions in the main text are robust to this choice.

Table 5: Pearson MCC scores on hold out interventions with respect to the number of latent space dimensions.

	VAE	**iVAE**	**sVAE**	**sVAE+**	β-**VAE**
$p = 5$	0.59 ± 0.12	0.69 ± 0.09	0.71 ± 0.09	$\mathbf{0.94 \pm 0.03}$	0.63 ± 0.06
$p = 10$	0.55 ± 0.10	0.66 ± 0.13	0.72 ± 0.13	$\mathbf{0.91 \pm 0.03}$	0.58 ± 0.03
$p = 15$	0.56 ± 0.09	0.68 ± 0.11	0.72 ± 0.08	$\mathbf{0.88 \pm 0.04}$	0.58 ± 0.08
$p = 20$	0.47 ± 0.01	0.59 ± 0.03	0.70 ± 0.09	$\mathbf{0.76 \pm 0.02}$	0.55 ± 0.08

D.2. Unseen In-domain interventions

The results included in the main section of the paper correspond to the in-domain scenario. Namely, we generate a sample data set with sparse interventions, we train each baseline on cells affected by

only 80 of those interventions, and we evaluate on the hold out, not seen 20 interventions. Here, we extend this results to different dimensionality of the latent variable. In Table 6, we include the results for in-domain interventions with same levels of sparsity in both train and test sets. That is, we train a model on a data set where a fraction s' of all possible edges in G^a are included, and we test on a data set with same sparsity. In Table 7, we investigate how those results hold with different effect sizes (encoded by the absolute value of the mean of the Gaussian prior shift under intervention). Our experiments for the in-domain interventions provide the following takeaways:

- observing the interventional NLL scores, the denser the ground truth latent model is, the more difficult it is to fit it with the sparse baselines
- observing the MCC scores, disentanglement is easier to achieve in smaller dimensionality of the latent space and when sparse shifts are present

D.3. Unseen Out-of-domain interventions

Here, we use the flexibility of our sandbox to simulate different scenarios to mimic different interventions regarding change in effect size from train to test set, or change in sparsity of the bipartite graph G^a. In Table 8, we include the results for out-of-domain interventions with different levels of sparsity between train and test sets. That is, we train a model on a data set where s' of all possible edges in G^a are included, and we test on a data set with reduced sparsity, s'', s.t. $s'' \geq s'$. Similarly, we include results for different effect sizes in Table 9. We train all models on a smaller effect interventions e' and test on interventions with larger effects e''. Our experiments for the out-of-domain interventions provide the following takeaways:

- observing the interventional NLL scores, the more sparse the ground truth latent model in the transfer domain, the more difficult it is to fit it for all baselines
- observing the MCC scores, disentanglement, or identifiability of the true latent variables is not possible in the out-of-domain interventions.

D.4. Additional ablation analysis

Additionally, we explored the robustness of sVAE+ with regards to the hyper prior α defining the shape parameters $(1, \alpha)$ for the Beta distribution. We remind the reader that this distribution controls the level of sparsity of the causal graph G^a estimated by sVAE+. From Figure 3, we notice the robustness of sVAE+ with regards to different values α. In practice, we found that choosing $\alpha = K$ (the number of interventions) has competitive performance.

Appendix E. Preprocessing of real world single-cell data

E.1. Norman data set

The original data set from Norman et al. (2019) is publicly available from GEO (GSE133344). For this manuscript, we downloaded the data processed according to Lotfollahi et al. (2021). Cells with guide "NegCtrl1_NegCtrl0_NegCtrl1_NegCtrl0" were excluded. All unperturbed cells were merged as one control (observational) condition. $d = 3,000$ highly variable genes were selected

Table 6: Results for transferability for in-domain interventions at different levels of sparsity s of the adjacency matrix G^a. The effect size and latent dimension are kept fixed: $e^a = 5$ for all interventions and $d_z = 10$. Results in bold are best per metric.

	$s' = 0.2$		$s' = 0.5$		$s' = 0.9$	
	pMCC	**Interv. NLL**	**pMCC**	**Interv. NLL**	**pMCC**	**Interv. NLL**
iVAE	0.53	319.27	**0.58**	330.56	**0.60**	333.28
sVAE	0.79	318.69	**0.58**	329.68	0.52	333.98
sVAE+	**0.83**	**309.45**	0.57	**322.01**	0.59	**322.51**

Table 7: Results for transferability for in-domain interventions for different sizes of the shift effect e. The sparsity size and latent dimension are kept fixed: $s^a = 0.2$ for all interventions and $d_z = 5$. Results in bold are best per metric.

	$e' = 1$		$e' = 2$		$e' = 5$	
	pMCC	**Interv. NLL**	**pMCC**	**Interv. NLL**	**pMCC**	**Interv. NLL**
iVAE	**0.78**	341.08	0.68	356.84	0.67	381.62
sVAE	**0.78**	341.03	0.68	349.55	0.66	375.80
sVAE+	**0.78**	**338.78**	**0.73**	**339.80**	**0.92**	**341.40**

Table 8: Results for transferability for out-of-domain interventions at different levels of sparsity s of the adjacency matrix G^a. The effect size and latent dimension are kept fixed: $e^a = 5$ for all interventions and $d_z = 10$. Results in bold are best per metric.

	$s' = 0.2 \to s'' = 0.5$		$s' = 0.5 \to s'' = 0.7$		$s' = 0.7 \to s'' = 0.99$	
	pMCC	**Interv. NLL**	**pMCC**	**Interv. NLL**	**pMCC**	**Interv. NLL**
iVAE	**0.44**	667.77	0.38	721.55	0.44	762.35
sVAE	0.40	676.64	0.38	**720.33**	0.37	710.88
sVAE+	0.41	**656.61**	**0.42**	675.15	**0.49**	**681.25**

Table 9: Results for transferability for out-of-domain interventions for different sizes of the shift effect e. The sparsity size and latent dimension are kept fixed: $s^a = 0.2$ for all interventions and $d_z = 5$. Results in bold are best per metric.

	$e' = 1 \to e'' = 3$		$e' = 2 \to e'' = 4$		$e' = 5 \to e'' = 7$	
	pMCC	**Interv. NLL**	**pMCC**	**Interv. NLL**	**pMCC**	**Interv. NLL**
iVAE	**0.47**	1179.31	**0.59**	1153.24	**0.58**	1224.26
sVAE	**0.47**	1173.11	0.44	1168.79	0.43	1229.12
sVAE+	**0.47**	**1103.91**	0.51	**1104.70**	0.46	**1179.39**

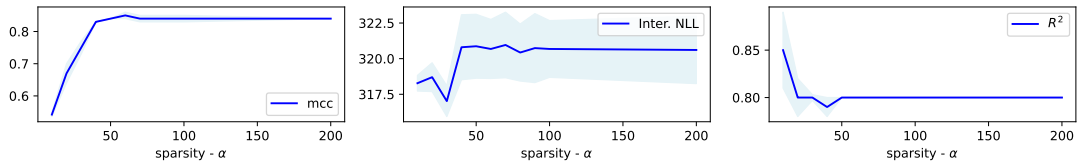

Figure 3: Pearson MCC, interventional NLL and R^2 score with respect to different values for the sparsity hyperparameter α in sVAE+.

using `scanpy` (Wolf et al., 2018) for training. For this data set, all models were run with $p = 50$ latent variables, and for $n_{\text{epochs}} = 300$ epochs. The remainder of the hyperparameters were selected via the evidence lower bound on a validation set, that contained the same interventions as the ones of the training set, as in Brouillard et al. (2020).

E.2. Replogle data set

Pre-filtered single-cell expression data of K562 cells from Replogle et al. (2022) was downloaded from http://gwps.wi.mit.edu. Selection of genes and interventions, as well as clustering of interventions into pathways was obtained from Supplementary Table 3 of the original paper. The number of interventions / cells used in the training and testing data sets are detailed for each data split in Table 10. For this data set, all models were run with $p = 50$ latent variables, and for $n_{\text{epochs}} = 100$ epochs. The remainder of the hyperparameters were selected via the evidence lower bound on a validation set, that contained the same interventions as the ones of the training set, as in Brouillard et al. (2020).

Table 10: Number of interventions used across data splits used for the Replogle dataset.

	K_{train}	K_{test}
Exosome and mRNA turnover	663	20
Spliceosome	648	35
Mediator complex	657	26
TFIIH/nucleotide excision repair	660	23
39S ribosomal subunit, mitochondrial	586	97
60S ribosomal subunit, cytoplasmic	640	43
40S ribosomal subunit, cytoplasmic	630	53
mitochondrial protein translocation	643	40

Appendix F. Biological interpretation of the sVAE+ model on the Norman dataset

We investigated how statistics of the number of perturbed latent variables and / or of the effect size was changing according to whether one or two genes were targeted in the cell (Figure 4a). The distribution of both statistics is significantly higher for double vs. single gene perturbations, as overall expected.

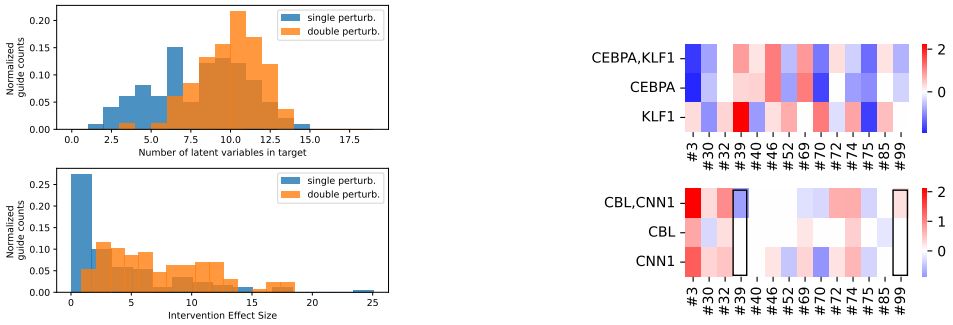

Figure 4: (A) Targets and effect size identified by the model. (B) Genetic interactions as identified from target and effect sizes with the model.

Finally, we sought to assess whether the learned latent variables are reflective of known patterns in genetic interactions. Two examples of genetic interactions are pointed out in Figure 4b. In the first one, we may notice that the latent shift for the perturbation that involved a combination of perturbations in CEBPA and KLF1 has a pattern mostly similar to the shift of a single gene perturbation CEBPA, as previously reported (Norman et al., 2019). This is an example of a dominant interaction, already visible in Figure 2, in other combinations (e.g., DUSP9, ETS2). In the second example (CLB, CNN1), the sparsity pattern identifies two latent variables (number 39 and 99; black rectangle) with a shift that did not appear in any of the individual perturbations. We applied Integrated Gradients (Sundararajan et al., 2017) to each of those two components of the encoder network to obtain a list of 50 most important genes, and used EnrichR (Chen et al., 2013) to obtain an associated gene signature. A positive change in latent variable 99 was associated with hemoglobin alpha binding, and hydrogen peroxyde metabolic process, both important in the context of erythocytes (the source cell line). Latent variable 39 was associated with RNA binding.

Appendix G. Discussion on identifiability (sVAE) assumptions in single-cell data

The original version of sVAE as proposed in Lachapelle et al. (2022) does not account for single-cell data. Hence we did modifications to adjust for that, as explained in Section 3. Additionally, for the identifiability theory of sVAE to hold, a number of assumptions are required, as stated in their Theorem 5. In what follows we discuss the main differences and modifications in the single-cell context.

G.1. Discrete observation space

The basic theory of Khemakhem et al. (2020), and Lachapelle et al. (2022) assumes that the data x is generated as $x = f(z) + \epsilon$ where components of ϵ are mutually independent, and all independent of z. It also assumes that f is a differentiable bijection with a differentiable inverse. Consequently, this theory may not be used to back up the claim that the decoder f of the generative model detailed in Section 3 is identifiable. Indeed, we assume that x follows a count distribution whose parameters depend on z, and this goes beyond the framework described by the aforementioned theoretical papers.

However, the appendices of Khemakhem et al. (2020) (as well as the Section 3 of this paper) describe encouraging empirical results that algorithms motivated by theory for data with linear Gaussian observation noise also improve performance on discrete observation spaces. The characterization of the theoretical conditions that are necessary and sufficient to guarantee identifiability with a discrete observation space remains an open problem in the field.

G.2. Sufficient variability

Both sVAE and iVAE require an "assumption of variability" that specify that the conditional distribution $p(z \mid a)$ must sufficiently vary with a. In both papers, this assumption is detailed as a geometric condition and cannot be validated in real data. Interestingly, this assumption is also not sufficient for guarantee correct estimation of the parameters in the case of finite sample sizes. This opens the way for new theoretical developments, based of non-asymptotic bounds (Wainwright, 2019). For example, we may wonder how many measurements are needed, under "sufficient variability", for estimating the latent variables when the interventional effect size becomes small.

From a more practical perspective, we accounted for this in our simulations by making sure the strength of the shift (encoded by parameter e in Appendix B) was large enough. Interestingly, as we see from the simulation, when the strength of the effect is low, methods have difficulties in disentangling the latent factors (Table 9).

This has also an important implication for real-world applications. In an earlier iteration of this project, we tried to apply iVAE and sVAE to the large-scale chemical screen presented in Srivatsan et al. (2020), but none of those baselines outperformed the vanilla VAE. After investigation, we attribute this to (i) a majority of chemicals having low effect on gene expression, (ii) a large fraction of the variance of the data being attributed to other signals, such as cell type variation, or cell cycles. In this case, all conditional VAEs may ignore information from the covariates and produce a similar result than the vanilla VAE.

G.3. Graphical criterion

The theoretical results of Lachapelle et al. (2022) hold under a precise condition on the structure of the ground truth graph G_a. For example, it is necessary (but not sufficient) for the intervention set to at least cover all of the latent variables in the model. If this is not the case, then we may have disentanglement only in a block of latent variables (Lachapelle and Lacoste-Julien, 2022). This assumption is particularly relevant because experiments may only focus on subsets of possible perturbations (due to cost and labor limitations) and not all latent variables may be impacted.

Importantly, perturbing all the latent variables is not sufficient. A complete graphical condition is exhibited the sVAE paper. However, this assumption may be difficult to validate or verify in practice. Still, according to the results of our simulation experiments (that did not take into account this condition while creating G_a, as well as the empirical evaluation in Appendix B.4. in Lachapelle et al. (2022), the causal graph can be approximate recovered with competitive performance compared to baselines even in the case when the graphical criterion is violated.