Learning Causal Representations of Single Cells via Sparse Mechanism Shift Modeling

Romain Lopez[1,2,†], **Nataša Tagasovska**[1,†],
Stephen Ra[1], **Kyunghyun Cho**[1,4,5,6], **Jonathan K. Pritchard**[2,3], **Aviv Regev**[1]

[1] Division of Research and Early Development, Genentech
{lopez.romain, ra.stephen, cho.kyunghyun, regeva}@gene.com
natasa.tagasovska@roche.com

[2] Department of Genetics, Stanford University
pritch@stanford.edu

[3] Department of Biology, Stanford University

[4] Department of Computer Science, Courant Institute of Mathematical Sciences, New York University

[5] Center for Data Science, New York University

[6] CIFAR Fellow

[†] These authors contributed equally to this work.

Abstract

Latent variable models have become a go-to tool for analyzing biological data, especially in the field of single-cell genomics. One remaining challenge is the identification of individual latent variables related to biological pathways, more generally conceptualized as disentanglement. Although versions of variational autoencoders that explicitly promote disentanglement were introduced and applied to single-cell genomics data, the theoretical feasibility of disentanglement from independent and identically distributed measurements has been challenged. Recent methods propose instead to leverage non-stationary data, as well as the sparse mechanism assumption in order to learn disentangled representations, with a causal semantic. Here, we explore the application of these methodological advances in the analysis of single-cell genomics data with genetic or chemical perturbations. We benchmark these methods on simulated single cell expression data to evaluate their performance regarding disentanglement, causal target identification and out-of-domain generalisation. Finally, by applying the approaches to a large-scale gene perturbation dataset, we find that the model relying on the sparse mechanism shift hypothesis surpasses contemporary methods on a transfer learning task.[†]

1 Introduction

Machine learning methods have been key to gaining insights from large, high-dimensional genomic datasets, especially in single cell genomics [1]. Variational Auto-Encoders (VAEs, [2, 3]), 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

[†]An extended version of this manuscript is available at https://arxiv.org/abs/2211.03553.

NeurIPS 2022Workshop on Causal Machine Learning for Real-World Impact (CML4Impact 2022).

biological data modalities, including gene (RNA) expression, chromatin accessibility and quantitative protein measurements [4, 5].

However, VAEs suffer from a key disadvantage due to the lack of interpretability, reflected as the absence of direct correspondence between individual latent variables and biological processes [6]. While disentanglement-promoting VAEs (e.g., as in [7]) can help better relate the two in genomics [8], these methods often compromise the quality of the latent variables for downstream tasks [9]. This is in line with recent theoretical developments, showing that disentanglement, hereafter defined as the recovery of ground truth latent variables, is impossible from independent and identically distributed (i.i.d.) measurements [10].

Recent efforts in disentanglement instead focus on the assumption of non-stationary data [11], 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 [12] is theoretically possible. Follow-up work [13, 14] also draws connections to causal representation learning [15], in which regimes may be seen as interventions on latent variables with unknown targets.

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 [16, 17]. A recent Compositional Perturbation Autoencoder (CPA [18]) embedded perturbation profiles in latent space of an auto-encoder, focusing of the prediction the effect of unseen combinations of perturbations. However, this approach did not exploit the non-stationary assumption in its probabilistic model. Thus, to the best of our knowledge, there are so far no applications of the principles exposed in [11, 13] to these new data types.

Here, we propose to explore real-world applications of [11, 13]. The promise is that such models may eventually yield representations of perturbations and cells that have the benefits of a causal model, in being more mechanistically interpretable and more efficient for out-of-domain generalization. Biologically, this means being able to identify biological processes (i.e., gene programs) that are affected upon perturbations [19], to project new perturbation data onto an existing perturbation atlas [20] with transfer learning [21].

After a brief presentation of recent advances in disentanglement (Section 2), we describe our motivation and effort in applying the introduced causal representation framework to gene expression measurements from single-cell RNA-seq (scRNA-seq) experiments with perturbations (Section 3). Then, we introduce a benchmarking tool for simulation of single-cell data with perturbation and evaluation of algorithms for latent variable recovery (Section 4). Finally, we present an application of the methods to a large-scale genetic screening experiment (Section 5) in which we show that the model based on sparse mechanism shift assumption outperforms all methods by a significant margin in a transfer learning task. Such results suggest that causal inference is a promising paradigm for modeling the effects of perturbation in modern screenings data sets from molecular biology.

2 Background

The goal of representation learning for a given random vector $x \in \mathbb{R}^d$, is to learn a mapping to a lower dimensional (latent) space $z = (z^1, \dots, z^p) \in \mathbb{R}^p$, with $p \ll d$. This paper is concerned with recovering the latent variables that initially generated the data, and attributing a causal semantic to the perturbations. Therefore, we now introduce non-linear Independent Component Analysis (ICA) and its relationship to causal representation learning.

2.1 Non-linear Independent Component Analysis

ICA assumes that x is generated using p independent latent variables, called independent components [22]. More precisely, observations x are generated as $x = f(z) + \epsilon$ with f a mixing function and ϵ some exogenous noise. 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 [23].

Given this negative result, several papers introduced identifiable forms [24, 25, 26, 27] of non-linear ICA models, based on the assumption that components $(z^i)_{i=1}^p$ are conditionally independent given some *additional auxiliary* random variable $a \in \mathbb{R}^K$:

$$p(z \mid a) = \prod_{i=1}^{p} p\left(z^i \mid a\right). \tag{1}$$

Examples of auxiliary variables a include the past components in the case of time series analysis or some form of class label [26]. In this setting, latent recovery is possible up to a linear transformation under sufficient conditions [26].

The iVAE framework [11] proposes a VAE approach [2, 3] 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):

$$\log p_\theta(x \mid a) \geq \mathbb{E}_{q_\phi(z|x,a)} \log \frac{p_\theta(x, z \mid a)}{q_\phi(z \mid x, a)}. \tag{2}$$

2.2 Causal Inference from Unknown Interventions in Latent Space

Recent theoretical work [13, 14] proposed to investigate the assumption of sparse connections between the individual auxiliary variables $(a^l)_{l=1}^K$ and the latent components $(z^i)_{i=1}^p$, encoded in the form of a bipartite graph $G^a = (\{1, \dots K\}, \{1, \dots p\}, E)$, where the edge set E encodes those connections. In the case where a describes a discrete data regime (i.e., via one-hot encoding), the sparsity of G^a corresponds to the sparsity of the mean vectors μ_a of $p_\theta(z \mid a)$.

Compared to [11], the novel sparsity assumption allows for a new principle to achieve recovery of latent units up to a permutation, a much stronger result than the iVAE. Perhaps as importantly, it also allows for the interpretation of the graph G^a from a causal perspective. More precisely, [13] considers the case where $a \in \{e_1, e_2, \dots, e_K\}$, where each of e_l for $l \in \{1, \dots, K\}$ is a one-hot vector encoding the l-th intervention, and each intervention has unknown targets on the set of latent components of z. In that context, 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 [15] i.e. that only a few mechanisms change at a time.

The VAE variant introduced in [13] (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 we apply a (stochastic) binary mask $\hat{G}^a$ to the location parameter via element-wise product. The estimation procedure follows the one from iVAE with an addition of the regularization term $\alpha\|\hat{G}^a\|_1$ to the ELBO, where α is a hyper parameter. To allow gradient-based optimization, the binary masks are independently sampled from Bernoulli distributions with probabilities $\texttt{sigmoid}\,(\gamma_i^a)$. We optimize for parameters γ^a using the Gumbel-Softmax gradient estimator [28, 29].

3 A Sparse Mechanism Shift Model for Single Cell Measurements

Single-cell genomics data is of immense interest across diverse research areas of biology such as development [30], autoimmunity [31], and cancer [32]. Due to the high-dimensional nature of the measurements and the high level of noise, the proper modeling and interpretation of such data remains a topic of active research [33].

The gene expression data matrix $X = (X_1, \dots, X_N) \in \mathbb{R}^{N \times G}$ consists of i.i.d. observations of the random vector X, with N the number of instances (cells), and G the number of genes. Individual entries of this matrix x_{ng} count the number of transcripts aligned to gene g in cell n. Typically, latent variable models are learned from gene expression measurements [5], and cell-level latent variables z (embeddings) are later used in downstream tasks such as clustering, imputation, or differential expression analysis [34].

One significant advance is the ability to actively intervene and change the properties of a single-cell by some genetic [35] or chemical [36] perturbation. Each cell may be subjected to one or several (simultaneous) perturbations. Assays capture this information alongside gene expression levels (often with sequencing of intervention-specific DNA barcodes), and report a summary of this readout as a multiple treatment intervention design matrix $\mathbf{I} \in \{0,1\}^{N \times K}$, where K denotes the total number of treatments.

Our goal is to leverage the perturbation information $\mathbf{I}$ in order to disentangle the latent variables z and learn the sparse causal graph G^a that describes the (unknown) targets of each intervention. To do this, we modify the sVAE [13] (as well as the iVAE [11]) to allow for proper modeling of scRNA-seq data, following previous work [34]. Briefly, two important modifications are needed: (i) we choose a negative binomial distribution as the likelihood model for the observations, and (ii) we incorporate a scaling factor to the mean of the negative binomial for normalization purposes. The implementation is performed within the scvi-tools codebase [37].

An appealing aspect of the sVAE model is the interpretability of G^a, indicating which latent variables are affected by which perturbations. Ideally, this would help considering individual latent variables as pathways and / or gene modules coordinately regulated by the perturbation, consistent with the underlying organization of a cellular molecular circuits [38]. Another interesting benefit from the causal semantic and the stronger disentanglement guarantees is that one can reasonably expect the learned representations to perform better at downstream tasks such as transfer learning [15], a growing use-case of deep generative models in single-cell genomics [21]. The presented modification of the sVAE is to the best of our knowledge the first proposal 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

Simulations We simulate perturbation gene expression data 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 $\mu_a \in \mathbb{R}^K$, where $d_z = 15$ 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 $\mu_a = (\mu_{a,1}, \dots, \mu_{a,p})$ for $i \in [p]$ is generated as:

$$\eta_{a,i} \sim \frac{1}{2}\text{Normal}(-5, 0.5) + \frac{1}{2}\text{Normal}(5, 0.5) \tag{3}$$

$$\mu_{a,i} \sim (1 - \beta_{a,i})\delta_0 + \beta_{a,i}\eta_{a,i}, \tag{4}$$

where 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:

$$z_n \sim \text{Normal}(\mu_a, I), \tag{5}$$

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:

$$x_{ng} \sim \text{Poisson}\left(l_n f^g(z_n)\right), \tag{6}$$

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 [34]. The weight matrices of f are sampled according to an isotropic Gaussian distribution, with orthogonal columns, to make sure f is injective [13].

Metrics Based on the ground truth provided by our simulation framework, we may evaluate for disentanglement, causal structure learning, or transferability. In particular, we report the MCC, a metric for permutation equivalence proposed by [11] 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 metric for assessing linear equivalence. To evaluate for learned causal relationships, we report the precision, recall and F1 score of the learned adjacency matrix of $\hat{G}^a$ compared to the ground truth G^a. Finally, to assess transferability, we report the negative ELBO of data points from holdout perturbations after fine-tuning the embeddings, while keeping the generative model fixed [39].

	Pearson MCC	Spearman MCC	R^2	Precision	Recall	F1	val. ELBO	test ELBO
VAE	0.46	0.40	0.82	0.13	0.10	0.11	322.16	346.29
βVAE	0.49	0.42	0.82	0.12	0.10	0.22	314.67	335.30
iVAE	0.47	0.39	**0.85**	0.27	0.22	0.24	314.71	335.37
sVAE	**0.72**	**0.63**	0.84	**0.42**	**0.24**	**0.31**	**314.65**	**335.23**

Table 1: Results on simulations for $d_z = 15$.

Benchmark Methods Our set of benchmark methods consist of the sVAE, the iVAE, as well as the vanilla VAE and the β-VAE [7]. We tune the hyper parameters of each method (number of epochs and sparsity penalty) with UDR [40]. All experiments were run for 5 different random initialisations, for $d_z \in [5, 10, 15, 20]$, $d_a = 100$ and 500 cells per treatment.

Results We report the results for disentanglement, causal target identification and transferability in Table 1 for $d_z = 15$. All methods have high R^2, indicating linear disentanglement, but sVAE outperforms all methods in terms of permutation recovery of latent variables, evaluated via MCC. sVAE also performs favorably compared to other methods when evaluating the inferred causal graph $\hat{G}^a$. We include additional experimental results from our synthetic framework in Appendix A. In terms of transferability, sVAE performs as well as other methods, with a slight advantage.

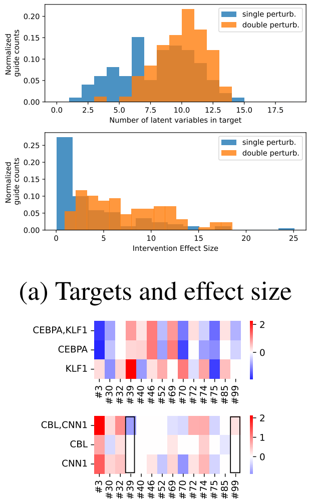
5 Transferability and Interpretability on a Large-scale Genetic Screen

We finally apply our benchmark models to real-world data, a recent large-scale Perturb-seq experiment [35] where $105,528$ 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. Because of experimental limitations [19], 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 an out-of-domain scenario, we selected the top-30 interventions with the most significant effect on gene expression, as assessed by the maximum mean discrepency [41] estimated with a linear kernel on a PCA with dimension 50, and held out the corresponding cells as a test set.

	val. ELBO	test ELBO
VAE	1202.1	1181.5
β-VAE	1206.0	1185.9
iVAE	**1149.6**	1154.0
sVAE	1149.7	**1149.9**

Table 2: Results on the Norman dataset.

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. We report the negative ELBO for all models, evaluated on a validation data set (including additional cells with the same perturbations as in the training data), as well on the held-out perturbations (test set) in Table 2. Both iVAE and sVAE are providing a large improvement in terms of data fit compared to VAE and β-VAE, for both metrics. While the iVAE provides the best fit in terms of the validation data set, with a thin margin, it fits the test set poorly compared to sVAE. This suggests that sVAE has stronger transfer capabilities than other methods, and learns a more causal representation of the data.



(a) Targets and effect size

(b) Genetic interactions

Figure 1: Model interpretation

We also performed a preliminary biological interpretation of the sVAE 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. First, many perturbations have similar effects on latent variables, as it has been observed previously

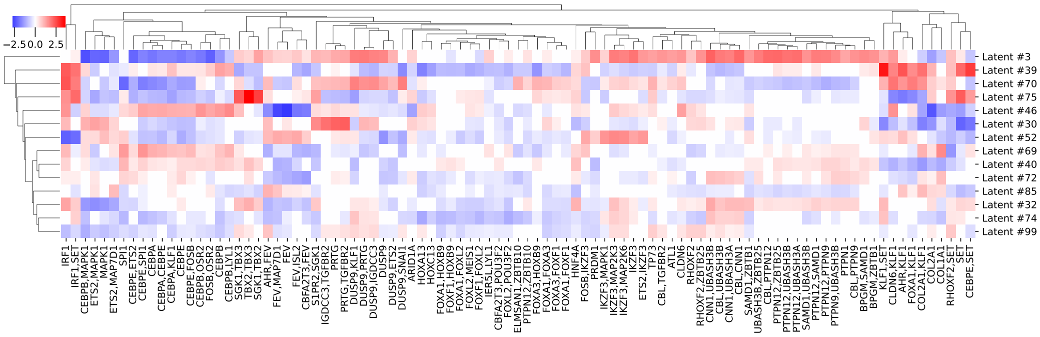


Figure 2: Perturbation effects on latent components (subset of perturbations, and components).

at the level of individual genes and programs [19]. 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 TFs affecting cell differentiation), whereas perturbations in genes from different pathways had different effects (e.g. IRF1 vs. CEBP family TFs). Second, 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 1a). The distribution of both statistics is significantly higher for double vs. single gene perturbations, as overall expected.

Third, 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 1b. 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 [35]. 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 [42] to each of those two components of the encoder network to obtain a list of 50 most important genes, and used EnrichR [43] 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.

6 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 our application of the sVAE [13] and iVAE [11] framework to disentangle the latent space of single cell data by leveraging additional knowledge of perturbations. We provided 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 a real dataset, and our preliminary analysis suggests that sparsity may help in transfer learning, interpretability, as well as capturing genetic interactions. 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 (e.g., [44]).

The hypotheses from the sVAE model in its current state present, however, two major limitations for biological applications. First, 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 [19]), there are important molecular feedback mechanisms that can induce indirect downstream effects in other pathways, especially as increasing time passes from the initial perturbation [45]. Because many experiments measure gene expression from hours to days after intervention, sparsity may be a limiting assumption without resolving interactions between pathways (e.g., as in more traditional causal discovery learning methods [46, 47, 48, 49]). This issue could be resolved in the future with the potential availability of time-resolved single-cell perturbation experiments. Second, perfect disentanglement requires the intervention to at least cover all of the latent variables in the model (see precise conditions in [13, 14]), but because experiments only focus on subsets of perturbations (due to cost and labor limitations), not all latent variables may be impacted. More comprehensive perturbation atlases, such as recent genome-wide screens [50] or combinatorial screens, may help mitigate this issue, as does the fact that many different interventions in biological systems converge on the same processes.

Acknowledgments and Disclosure of Funding

We thank Sébastien Lachapelle and Chandler Squires for insightful conversations about causal latent variable models. We acknowledge Kelvin Chen, Taka Kudo, Jan-Christian Huetter, Zia Khan, Basak Eraslan, and Anna Klimovskaia Susmelj for discussions about modeling single-cell perturbation data sets. Last but not least, we acknowledge the reviewers from the NeurIPS 2022 Workshop on Causal Machine Learning for Real-World Impact who provided very insightful feedback about this manuscript.

Disclosures: Romain Lopez, Natasa Tagasovska, Stephen Ra, and Kyunghyun Cho are employees of Genentech. Stephen Ra and Kyunghyun Cho have equity in Roche. Jonathan Pritchard acknowledges support from grant R01HG008140 from the National Human Genome Research Institute. Aviv Regev is a co-founder and equity holder of Celsius Therapeutics and an equity holder in Immunitas. She was an SAB member of ThermoFisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics, and Asimov until July 31st, 2020; she has been an employee of Genentech since August 1st, 2020, and has equity in Roche.

Code Availability Statement

We implement our new model and benchmarks using the scvi-tools library, and release it as open-source software at `https://github.com/Genentech/sVAE`.

References

[1] 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.

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

[3] 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.

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

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

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

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

[13] 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.

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

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] Yun Xiao, Yonghui Gong, Yanling Lv, Yujia Lan, Jing Hu, Feng Li, Jinyuan Xu, Jing Bai, Yulan Deng, Ling Liu, et al. Gene perturbation atlas (GPA): a single-gene perturbation repository for characterizing functional mechanisms of coding and non-coding genes. *Scientific reports*, 5(1):1–9, 2015.

[21] 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, 2022.

[22] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. Independent component analysis. *Studies in informatics and control*, 11(2):205–207, 2002.

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

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

[25] 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.

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

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

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome Biology*, 21(1):1–35, 2020.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

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

[43] 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.

[44] 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*, 2022.

[45] 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.

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

[47] 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.

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

[49] 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.

[50] 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.

Appendices

A 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).

A.1 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 4, 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.

pMCC	VAE	iVAE	sVAE	βVAE
$d_z = 5$	0.59 ± 0.12	0.69 ± 0.09	**0.71 ± 0.09**	0.63 ± 0.06
$d_z = 10$	0.55 ± 0.10	0.66 ± 0.13	**0.72 ± 0.13**	0.58 ± 0.03
$d_z = 15$	0.56 ± 0.09	0.68 ± 0.11	**0.72 ± 0.08**	0.58 ± 0.08
$d_z = 20$	0.47 ± 0.01	0.59 ± 0.03	**0.70 ± 0.09**	0.55 ± 0.08

Table 3: Pearson MCC scores on hold out interventions with respect to the number of latent space dimensions.

Our experiments for the in-domain interventions provide the following takeaways:

- observing the ELBO scores, the sparser the ground truth latent model is, the more difficult it is to fit it for all baselines
- observing the MCC scores, disentanglement is easier to achieve in smaller dimensionality of the latent space and when sparse shifts are present
- observing the area under the Precision recall curve (AUC PR), we notice that the identifiable methods iVAE and sVAE are more successful in recovering the causal structure model then their unsupervised counterparts

A.2 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 6, 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. We train all models on a smaller effect interventions e' and test on interventions with larger effects e''.

	$s' = 0.2$			$s' = 0.5$			$s' = 0.9$		
	pMCC	**ELBO**	**PR AUC**	**pMCC**	**ELBO**	**PR AUC**	**pMCC**	**ELBO**	**PR AUC**
VAE	0.59	310.32	0.15	0.54	309.14	**0.7**	0.61	306.51	0.72
bVAE	0.17	411.86	0.15	0.26	411.07	0.48	0.19	407.76	0.72
iVAE	0.69	308.55	0.28	0.6	**304.60**	0.62	**0.63**	**300.64**	0.78
sVAE	**0.78**	**308.33**	**0.42**	**0.63**	306.10	0.61	0.6	305.35	**0.82**

Table 4: 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.

	$e' = 1$			$e' = 2$			$e' = 5$		
	pMCC	**ELBO**	**PR AUC**	**pMCC**	**ELBO**	**PR AUC**	**pMCC**	**ELBO**	**PR AUC**
VAE	0.72	297.67	0.19	0.62	297.22	0.20	0.72	294.48	0.19
bVAE	0.62	335.95	0.19	**0.78**	335.12	0.19	**0.82**	325.48	**0.59**
iVAE	**0.74**	**297.09**	**0.59**	0.61	295.92	0.19	0.7	**291.10**	**0.59**
sVAE	0.73	297.24	**0.59**	0.62	295.95	**0.59**	0.7	**291.07**	**0.59**

Table 5: 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.

	$s' = 0.2 \to s'' = 0.5$			$s' = 0.5 \to s'' = 0.7$			$s' = 0.7 \to s'' = 0.99$		
	pMCC	**ELBO**	**PR AUC**	**pMCC**	**ELBO**	**PR AUC**	**pMCC**	**ELBO**	**PR AUC**
VAE	0.36	412,03	0.33	**0.28**	408.89	0.56	0.23	404.88	0.72
bVAE	0.17	411.86	0.33	0.26	411.07	0.56	**0.28**	392.47	0.82
iVAE	0.31	**405.58**	**0.41**	0.26	**397.08**	**0.64**	0.21	392.79	0.77
sVAE	0.33	406.11	0.4	0.25	397.18	**0.64**	0.27	**392.05**	**0.82**

Table 6: 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.

	$e' = 1 \to e'' = 3$			$e' = 2 \to e'' = 5$			$e' = 5 \to e'' = 7$		
	pMCC	**ELBO**	**PR AUC**	**pMCC**	**ELBO**	**PR AUC**	**pMCC**	**ELBO**	**PR AUC**
VAE	0.39	410.82	0.34	0.47	411.45	0.26	0.39	412.49	0.3
bVAE	**0.47**	422.86	0.34	0.46	424.96	0.26	0.29	422.07	0.3
iVAE	**0.47**	**409.19**	0.44	**0.52**	**407.95**	0.59	**0.33**	**403.81**	**0.6**
sVAE	0.46	409.48	**0.61**	**0.52**	408.06	**0.6**	0.32	403.47	**0.6**

Table 7: 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.

Our experiments for the out-of-domain interventions provide the following takeaways:

- observing the ELBO 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 ood interventions.
- observing the area under the Precision recall curve (AUC PR), we notice that the identifiable methods iVAE and sVAE are more successful in recovering the causal structure model then their unsupervised counterparts.
- it is possible to leverage representations/models learned on sparser interventions, to recover a causal structures in data sets with denser interventions (PR AUC, last column).