DISENTANGLEMENT VIA MECHANISM SPARSITY BY REPLAYING REALIZATIONS OF THE PAST

Soroor Hediyeh-zadeh
School of Life Sciences
Technical University of Munich, Germany
Computational Health Center, Helmholtz Munich, Germany
soroor.hediyehzadeh@helmholtz-munich.de

Tom Fischer & Fabian J Theis
School of Computation, Information and Technology
Technical University of Munich, Germany
Computational Health Center, Helmholtz Munich, Germany
{tom.fischer,fabian.theis}@helmholtz-munich.de

ABSTRACT

Recent lines of work have proposed learning disentangled representations using observed auxiliary variables by *mechanism sparsity regularization*. These works assume that the pairing between the auxiliary variables and samples is known. Inspired by biological problems in controllable counterfactual generation and mechanism transportability for genomics explorations, this work combines mechanism sparsity regularization and methods from Continual Learning to introduce a representation learning method which applies when the auxiliary variables are not directly observed and the assignment between the latent auxiliary variables and samples is not known. Rather than requiring observed auxiliary variables for disentanglement, we propose to use realizations of the auxiliary variables of interest. We propose an estimation procedure based on variational autoencoders and demonstrate it on various synthetic and biological data in generating counterfactual instances of cell states or transcriptional signatures to achieve desired cell state shifts.

1 INTRODUCTION

The problem of disentanglement can be formulated as learning interpretable, semantically meaningful representations from high-dimensional observational data. Disentanglement via mechanism sparsity (Lachapelle et al., 2022) assumes that the latent factors of interest depend sparsely on observed auxiliary variables, such as time, environment index or auxiliary variables of the observation in the past, if there is a temporal structure. This idea of inducing disentanglement by assuming that *only a few mechanisms change at a time*, also known as the *sparse mechanism shift* hypothesis (Schölkopf et al., 2021), has already been used by several works on learning representations of single cells (Lopez et al., 2023; Bereket & Karaletsos, 2023). Disentanglement can additionally be viewed in terms of controllably generating counterfactual data. Komanduri et al. (2023), for example, defines controllable generation as modeling the causal process between known or unknown labels and data as a form of structural mechanism learning. Disentanglement is also related to *identifiability* of latent factors. Khemakhem et al. (2020) demonstrated that the latent factors learnt by deep generative models would be identifiable only when auxiliary variables are observed. Kivva et al. (2022) proved identifiability of deep latent variable models without auxiliary information.

Continual Learning is about learning incrementally between two or more domains and prevent forgetting the past data (Hadsell et al., 2020; Lopez-Paz & Ranzato, 2017; Van de Ven & Tolias, 2019). Common approaches include dynamic expansion of neural network architecture (Rao et al., 2019; Parascandolo et al., 2018; Yan et al., 2021), weight regularization (Kirkpatrick et al., 2017), and replay/rehearsal of past experiences (Shin et al., 2017), known as Generative Replay and Experience replay respectively. There is a natural connection between causal models and learning continually, as causal models are assumed to be invariant between domains, and the aim of continual learning is to learn across domains (Mundt et al., 2023). Here, we bridge the two fields of disentangled representation learning via mechanism sparsity and continual learning to address a special case of inducing disentanglement via mechanism sparsity regularization when the auxiliary information is not directly observed.

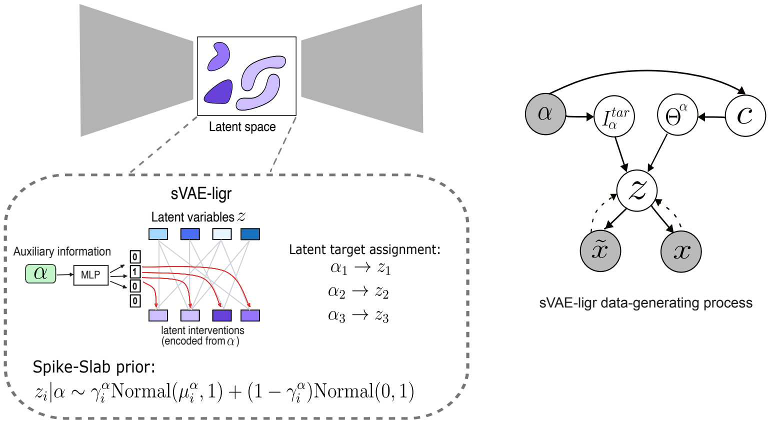
Latent space
sVAE-ligr
Latent variables z
Auxiliary information
α
MLP
0
1
0
0
latent interventions
(encoded from α)
Latent target assignment:
$\alpha_1 \rightarrow z_1$
$\alpha_2 \rightarrow z_2$
$\alpha_3 \rightarrow z_3$
Spike-Slab prior:
$z_i|\alpha \sim \gamma_i^\alpha \text{Normal}(\mu_i^\alpha, 1) + (1-\gamma_i^\alpha)\text{Normal}(0,1)$
α
I_α^{tar}
Θ^α
c
z
$\tilde{x}$
x
sVAE-ligr data-generating process

Figure 1: Schematic illustration of sVAE-ligr and the graphical model of the data generating process

We summarize our contributions as the followings:

- Inspired by biological applications, we propose a representation learning method based on mechanism sparsity for inducing partial disentanglement with respect to auxiliary variables that are not observed for the samples in the data. We propose to induce disentanglement using realizations of the auxiliary variables of interest. The realizations are created by Generative- or Experience Replay approaches in Continual Learning.
- We learn sparse dependencies between latent factors of the data and realizations of the auxiliary variables of interest. This enables us to learn the mechanisms that identify the auxiliary variable and generate counterfactual queries.
- We propose an estimation procedure based on variational autoencoders called sVAE-ligr (SpikeSlabVAE with learnable interventions by Generative Replay)

2 Problem formulation

In mechanism sparsity regularization frameworks, it is assumed that the latent factors of interest depend sparsely on observed auxiliary variables. These frameworks learn a graph G^a that models causal relations between the observed auxiliary variable, a, and latent factors, such that each auxiliary variable can be identified by targeting an unknown subset of latent factors. However, some biological applications involve modeling dependencies between latent factors and unobserved auxiliary variables, where the pairing between the variables (e.g. labels) and samples are not known.

For example, biologists may identify a disease-causing mechanism, captured by transcriptional signatures or changes in cell type composition, in mouse model systems and might be interested to see if the mechanism would translate in human data (to verify if the mechanism is also associated with an outcome in humans). Here, the biological mechanism of interest would be the auxiliary variable that is not observed in the target human data, but we aim to generate counterfactual instances of that therein.

In this work, we formulate this problem as finding sparse decencies between *abstraction of the biological mechanism to be transferred* and latent factors of the target data. Latent factors that identify a biological mechanism (equivalent to a in existing works) can be intervened upon to generate counterfactual instances of the biological mechanism. In other words, we define a prior over data points in the target data based on the auxiliary information that we aim to quantify in the target data. We propose to use Generative Replay or Experience Replay to create realizations of the information that is transferred. These realizations are additionally presented to the model (replayed) during training, which encourages learning shared representations between the training data and data that encode the auxiliary information.

Figure 1 illustrates the graphical model of the data generating process. Briefly, realizations are used to learn binary variables that define an unknown subset of latent factors that identify them. The realizations are also replayed along with training data. A latent variable probabilistically assigns training samples to auxiliary variable encoded by the realizations.

2.1 The generative process

We assume a spike-slab prior on the latent variables z_i, $i \in \{1, \cdots, m\}$:

$$\tau_a = NN(\tilde{x}_a)$$

$$\pi_i^a \sim \text{Beta}(1, \tau_a)$$

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

$$z_i|a \sim \gamma_i^a \text{Normal}(\mu_i^a, 1) + (1 - \gamma_i^a)\text{Normal}(0, 1)$$

where $\tilde{x}_a$ represents realizations of the auxiliary variable a from Experience- or Generative Replay and $NN(\cdot)$ is a neural network. We use the Gumbel-sigmoid distribution, a continuous relaxation of the Bernoulli distribution during optimization. Note that in Figure 1, I_a^{tar} represents γ^a (i.e. the binary mask) and θ^a represents μ^a in the prior formulation $(z|a)$. The binary interaction variables γ^a define a graph G^a over latent factors and a. The choice of this prior is closely related to the work by Lopez et al. (2023), with the difference being in the assumptions made about the pairing between a and data points x. In our work the assumption is that a is not directly observed for x, whereas Lopez et al. (2023) requires label-sample pairs to be known.

Each sample is probabilistically assigned to the most likely auxiliary variable through the latent assignment variable C:

$$C \sim \text{Cat}(\beta_1, \beta_2, \ldots, \beta_k),$$

$$\beta \sim Dir(\kappa_1, \kappa_2, \ldots, \kappa_k)$$

Note here $d_a = k$, and κ is the concentration parameter of the Dirichlet prior placed over cluster weights, β, from which cluster assignments are sampled according to a multinomial distribution.

2.2 Loss function

In practice, we implement the loss as follows:

$$\mathcal{L} = \mathcal{L}_{\text{sVAE}} \; + \lambda \, \mathcal{L}_{\text{sVAE}}^{R} + \mathbb{E}\left[D_{KL}(q(\beta) \,||\, \text{Dir}(\frac{1}{k}))\right] \; + \mathbb{E}\left[D_{KL}(p(z|a) \,||\, p(z|c))\right]$$

where $\mathcal{L}_{\text{sVAE}}$ is the loss from the SpikeSlab VAE model (Appendix B), $\mathcal{L}_{\text{sVAE}}^{R}$ is the loss with respect to replay (generative or experience) examples and λ is the importance weight for the replay loss. We add additional KL terms to ensure cluster assignments are learnt optimally.

3 Results

We first present a comparison of the proposed method with baselines in synthetic data and evaluate the performance on two tasks: disentanglement of the latent representations and causal discovery, that is the recovery of the true graph (G^a). We then assess disentanglement and counterfactual generation in data from single-cell Genetic Screens and Bulk sequencing of gene expression in primary tumors.

Table 1: Mean and standard deviation per metric on simulations for $d = 5$. Best is bold.

	Disentanglement			Causal discovery			
	Pearson MCC (↑)	**Spearman MCC(↑)**	**R^2(↑)**	**Precision (↑)**	**Recall(↑)**	**F1 (↑)**	**SHD (↓)**
VAE	0.511±0.04	0.497±0.04	0.718±0.02	NA	NA	NA	NA
betaVAE	0.479±0.01	0.460±0.01	0.698±0.03	NA	NA	NA	NA
iVAE	0.539±0.05	0.526±0.05	0.766±0.04	0.366±0.01	0.418±0.1	0.390±0.01	182.8±4.60
SpikeSlabVAE	0.521±0.03	0.504±0.03	**0.803±0.01**	**0.371±0.01**	0.371±0.02	0.371±0.01	**176.0±5.14**
sVAE-ligr (ours)	**0.554±0.06**	**0.542±0.07**	0.733±0.02	0.333±0.03	**0.480±0.05**	**0.393±0.04**	207.0±14.4

3.1 Comparison to baselines in synthetic data

We generated five synthetic data which differ in the number of generative factors (dimension of the latent variables), $d_z \in \{5, 10, 15, 20\}$. We experimented with five models: VAE (Kingma & Welling, 2013), betaVAE (Higgins et al., 2016), iVAE (Khemakhem et al., 2020) and SpikeSlabVAE (Lopez et al., 2023). The performance was assessed using a total of seven metrics. Among the aforementioned models, iVAE and SpikeSlabVAE use auxiliary information to learn conditionally independent latent variables given observed auxiliary variables a, and a causal graph G^a between those variables and the latent generative factors. Each model was ran with five different initialisation. More details on synthetic data generation is given in Appendix E. See Appendix D for description of evaluation metrics.

In Table 1, we present the results for experiments on synthetic data with $d = 5$ latent dimensions. We outperform all baselines in Mean Correlation Coefficient (MCC) of the learned latent variables with the ground truth latent variables. We are ranked third in R^2 score between learned and ground truth latent variables after SpikeSlabVAE and iVAE. As R^2 is a metric for linear identifiability (see Appendix D), this suggests that identifiability results are weaker for our approach, which, compared to SpikeSlabVAE, employs an indirect conditioning on the prior $p(z)$. In causal discovery metrics, we outperform the baselines in Recall and F1 score for the recovery of G^a. However, the Structural Hamming Distance (SHD) suggest that the graph learned by our model can be more different from the ground truth graph compared to other baselines that also use auxiliary variables to achieve conditionally independent latent variables, which can have implications for counterfactual generation and when the intention is to transport the mechanisms to other domains for generalisation. Overall, these results suggest that our proposed model competes closely in disentanglement and graph recovery with existing approaches.

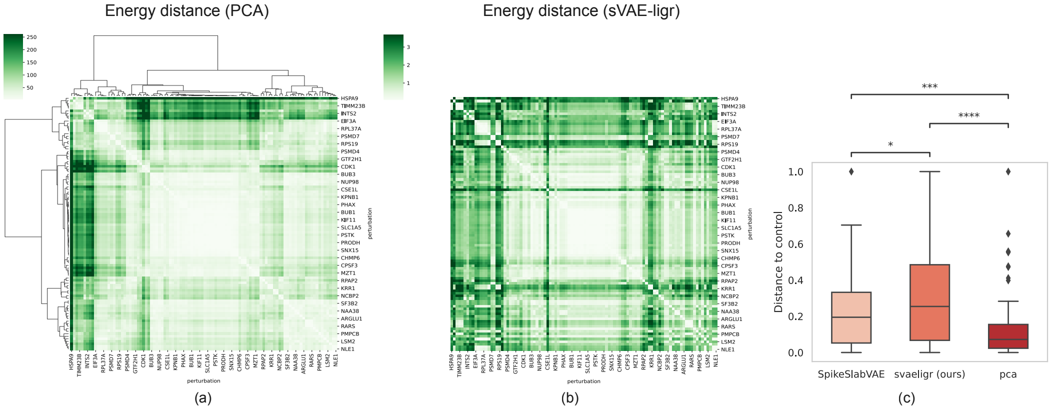

Figure 2: Disentanglement of perturbation effects in a Genetic Screen. (a) Energy distance between perturbations on PCA representations and (b) sVAE-ligr representations. (c) Energy distance of perturbations to the 'control' group. The larger the Energy distance the better.

3.2 Representation disentanglement in a Genetic Screen

Genetic perturbation screens with single-cell RNA-sequencing readouts, known as Perturb-seq, have empowered genome-scale mapping of gene functions (Replogle et al., 2022). However, the signal captured by this assay can be noisy and/or confounded by a variety of technical and biological factors, such as variability in the effectiveness of the perturbation and cell cycle (Jiang et al., 2024). Optimization of data analysis pipelines for improved detection of perturbation effects is an active area of research. For example, Lopez et al. (2023) and Tu et al. (2024) proposed DGMs for learning disentangled representations of single-cell Perturb-seq measurements, whereas Jiang et al. (2024) proposed classification-based approaches for optimal detection of perturbation effects.

Here, we compare disentanglement of the representations for the top 100 most effective perturbations in a Perturb-seq experiment in K562 cell line from Replogle et al. (2022) (See appendix A.1) between PCA representations, representations learnt by SpikeSlabVAE (Lopez et al., 2023), which models the effects of the perturbations on gene expression in cells by sparse mechanisms shift, and representations learned by our model (Figure 2). We computed the Energy Distance (Peidli et al., 2024) between pairs of perturbations, a metric that compares pairwise distances within and between sets of samples. We found that distances between perturbations are generally small in PCA representations, suggesting that PCA representations can not discern the effect of different perturbations (Figure 2a). However, in representations learned by sVAE-ligr, we observed larger distances between perturbations for a larger groups of perturbations (Figure 2b). Indeed, we could confirm with a two-sided Mann-Whitney-Wilcoxon test that the min-max normalised Energy Distances of the perturbations to the 'control' (unperturbed) group are significantly larger in sVAE-ligr cell representations compared to PCA and SpikeSlabVAE (Figure 2c, pvalue=5.721e-07 and =4.395e-02 respectively), implying that the effect of the perturbations are more discernible in representations learned by our model. This suggests that sVAE-ligr representations could be more disentangled compared to representations learnt by SpikeSlabVAE, which was also a finding of the experiments with synthetic data. An important fact to note is that while SpikeSlabVAE requires the pairing between perturbation labels and cells to be known, sVAE-ligr probabilistically assigns perturbation labels to cells, a modeling strategy that was also recently explored in Tu et al. (2024).

3.3 COUNTERFACTUAL GENERATION

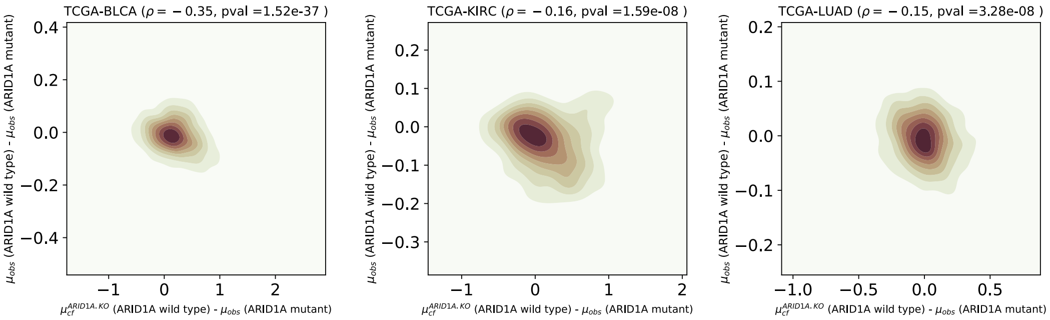

Figure 3: Counterfactual generation of ARIDA1 gene knockout transcriptional signature from a single-cell Perturb-seq study in primary tumors from TCGA shifts tumor state to druggable ARID1A mutant primary tumors.

Since the original motivation behind the development of this work is controllable counterfactual generation with applications in cross-modal cell state transfer, we sought to demonstrate applications in biological settings, specifically generating counterfactual instances of primary tumors (bulk RNA-seq) with cell states or gene knockouts (KO) from single-cell RNA-seq or Perturb-seq experiments.

We first describe the results of counterfactual generation of a gene knockout, ARID1A- measured in a single-cell perturbation screen in a AML cell line model- in primary tumors of three cancers. ARID1A is a subunit of the human chromatin remodeler BAF complex. BAF loss-of-function (LOF) mutations occur in 20%-25% of cancers (Wanior et al., 2021) and contribute to cancer initiation or progression. There are currently drugs to reduce viability of cells with BAF LOF (Otto et al., 2023). Patients with BAF LOF signatures can therefore take advantage of such drugs. Otto et al. (2023) performed a genetic knockout screen of this complex via Perturb-seq. We leveraged their data to generate counterfactual instances of ARID1A KO expression in ARID1A wild type (WT) primary tumors from three cancer types (see Appendix A.2).

Table 2: Cox Proportional Hazard model to test for the association of counterfactually generated cell states in colorectal cancer primary tumors (bulk RNA-seq) from TCGA to patient survival outcomes.

covariate	**coef**	**exp(coef)**	**se(coef)**	**z**	**p**
ajcc_pathologic_tT2	-0.948738	0.387229	0.868330	-1.093	0.27457
ajcc_pathologic_tT3	-1.407315	0.244800	0.787323	-1.787	0.07386
ajcc_pathologic_tT4	-0.850418	0.427236	0.814395	-1.044	0.29638
ajcc_pathologic_tT4a	-0.326012	0.721797	0.841378	-0.387	0.69841
ajcc_pathologic_tT4b	-0.842628	0.430577	0.961829	-0.876	0.38099
ajcc_pathologic_tTis	NA	NA	0.000e+00	NA	NA
cf_CyclingTA	0.007234	1.007260	0.004509	1.604	0.10863
cf_Stem	-0.001582	0.998420	0.003746	-0.422	0.67289
cf_TA2	0.003340	1.003346	0.001260	**2.650**	**0.00805**
cf_Immature Goblet	-0.009304	0.990739	0.005125	-1.815	0.06946

coxph model Likelihood ratio test=17.05 on 9 df, p=0.04788, n= 519, number of events= 115 (2 observations deleted due to missingness)

Let $\mu_{obs}(ARID1A\, wild\, type)$ denote the observational distribution of gene expression in ARID1A WT primary tumors. Let $\mu_{obs}(ARID1A\, mutant)$ denote the observational distribution of gene expression in ARID1A mutant primary tumors, and let $\mu_{cf}^{ARID1A.KO}(ARID1A\, wild\, type)$ denote the counterfactually generated gene expression of ARID1A KO (from Perturb-seq) in ARID1A WT primary tumors. In Figure 3, we demonstrate that for every unit increase in the difference between observed mean gene expression in ARID1A WT and ARID1A mutant, that is $\mu_{obs}(ARID1A\, wild\, type)$ - $\mu_{obs}(ARID1A\, mutant)$, the difference between counterfactually generated mean gene expression of ARID1A KO phenotype in ARID1A WT tumors and ARID1A mutant tumors, that is $\mu_{cf}^{ARID1A.KO}(ARID1A\, wild\, type)$ - $\mu_{obs}(ARID1A\, mutant)$, decreases. This trend, which is captured by negative correlation between the two axes, is observed at statistically significant levels in all three primary cancer tumors considered here (BLCA, KIRC and LUAD), suggesting that the proposed mechanism for generation of ARID1A KO phenotype has been effective.

Next, we generated counterfactual instances of Colorectal Cancer (CRC) primary tumors expressing cell states found to be enriched in a limited number of single cell RNA-seq profile of CRC patients (see Appendix A.3). Counterfactual generation of cell states in primary tumors could especially be helpful as bulk RNAseq is more affordable in clinical settings compared to single cell sequencing, enabling measurements of gene expression and recording rich patient metadata such as survival outcomes. Therefore, by generating counterfactual instances of tumors expressing specific cell states, we can test for association of the cell state with patient outcome.

In our CRC example described earlier, we found a group of Transit-Amplifying (TA) cells to be associated with poor survival (hazard ratio = 1.00, z-score = 2.650, pvalue = 0.008)(Table 2). Jones et al. (2023) reported the presence of colorectal cancer stem cell (CCSC)-like TA cells. The association between CCSCs and poor survival is already well known (Hervieu et al., 2021). Merlos-Suárez et al. (2011) found that TA cells are not associated with relapse but only at border-line significance, suggesting that the association could have been significant in datasets with larger statistical power. Collectively, these reports affirm the findings reported in Table 2.

4 Discussion

In this work, we presented a representation learning method based on mechanism sparsity for inducing partial disentanglement with respect to auxiliary variables that are not observed for the samples in the data. This model is inspired by biological problems related to counterfactual generation and mechanism transport. We propose to induce disentanglement using realizations of the auxiliary variables of interest. The realizations are created by Generative- or Experience Replay approaches in Continual Learning. We learn sparse dependencies between latent factors of the data and realizations of the auxiliary variables of interest. This enables us to learn the mechanisms that identify the auxiliary variable and generate counterfactual queries.

We demonstrated our method on synthetic and biological data. On biological data, we demonstrated that the proposed method can generate the desired counterfactual instances of cell states and transcriptional signatures in the target data. The simulation experiments flagged potential issues with the identifiability of the model (up to linear permutations). Lachapelle et al. (2024) recently proposed conditions under which quasi-linear identifiability results could be achieved in the absence of auxiliary variables, a similar case to our setting. We, hence, leave the assessment of model identifiability as the future direction of this work.

Code availability

The code and notebooks to reproduce the results of this study have been deposited to `https://zenodo.org/records/11072912` under DOI `10.5281/zenodo.11072911`.

Acknowledgments

We are grateful to Andrea Dittadi, Stefan Bauer, Paul Bertin and Simon Buchholz for helpful discussions. The authors are thankful to Julien Gagneur Lab and their ITS team for HPC support during early phases of the project.

References

Winston R Becker, Stephanie A Nevins, Derek C Chen, Roxanne Chiu, Aaron M Horning, Tuhin K Guha, Rozelle Laquindanum, Meredith Mills, Hassan Chaib, Uri Ladabaum, et al. Single-cell analyses define a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer. *Nature genetics*, 54(7):985–995, 2022.

Michael Bereket and Theofanis Karaletsos. Modelling cellular perturbations with the sparse additive mechanism shift variational autoencoder. *arXiv preprint arXiv:2311.02794*, 2023.

Emma Dann, Neil C Henderson, Sarah A Teichmann, Michael D Morgan, and John C Marioni. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nature Biotechnology*, 40(2):245–253, 2022.

Sunny Duan, Loic Matthey, Andre Saraiva, Nicholas Watters, Christopher P Burgess, Alexander Lerchner, and Irina Higgins. Unsupervised model selection for variational disentangled representation learning. *arXiv preprint arXiv:1905.12614*, 2019.

Raia Hadsell, Dushyant Rao, Andrei A Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences*, 24(12):1028–1040, 2020.

Céline Hervieu, Niki Christou, Serge Battu, and Muriel Mathonnet. The role of cancer stem cells in colorectal cancer: from the basics to novel clinical trials. *Cancers*, 13(5):1092, 2021.

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. In *International conference on learning representations*, 2016.

Longda Jiang, Carol Dalgarno, Efthymia Papalexi, Isabella Mascio, Hans-Hermann Wessels, Huiyoung Yun, Nika Iremadze, Gila Lithwick-Yanai, Doron Lipson, and Rahul Satija. Systematic reconstruction of molecular pathway signatures using scalable single-cell perturbation screens. *bioRxiv*, pp. 2024–01, 2024.

Josh Jones, Qiaojuan Shi, Rahul R Nath, and Ilana L Brito. Keystone pathobionts associated with colorectal cancer promote oncogenic reprograming. *bioRxiv*, pp. 2023–04, 2023.

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

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

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. *Advances in Neural Information Processing Systems*, 35:15687–15701, 2022.

Aneesh Komanduri, Xintao Wu, Yongkai Wu, and Feng Chen. From identifiable causal representations to controllable counterfactual generation: A survey on causal generative modeling. *arXiv preprint arXiv:2310.11011*, 2023.

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*, pp. 428–484. PMLR, 2022.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *arXiv preprint arXiv:2401.04890*, 2024.

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

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Anna Merlos-Suárez, Francisco M Barriga, Peter Jung, Mar Iglesias, María Virtudes Céspedes, David Rossell, Marta Sevillano, Xavier Hernando-Momblona, Victoria da Silva-Diz, Purificación Muñoz, et al. The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. *Cell stem cell*, 8(5):511–524, 2011.

Martin Mundt, Keiland W Cooper, Devendra Singh Dhami, Adéle Ribeiro, James Seale Smith, Alexis Bellot, and Tyler Hayes. Continual causality: A retrospective of the inaugural aaai-23 bridge program. In *AAAI Bridge Program on Continual Causality*, pp. 1–10. PMLR, 2023.

Jordan E Otto, Oana Ursu, Alexander P Wu, Evan B Winter, Michael S Cuoco, Sai Ma, Kristin Qian, Brittany C Michel, Jason D Buenrostro, Bonnie Berger, et al. Structural and functional properties of mswi/snf chromatin remodeling complexes revealed through single-cell perturbation screens. *Molecular Cell*, 83(8):1350–1367, 2023.

Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *International Conference on Machine Learning*, pp. 4036–4044. PMLR, 2018.

Stefan Peidli, Tessa D Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J Schumacher, Jake P Taylor-King, Debora S Marks, et al. scperturb: harmonized single-cell perturbation data. *Nature Methods*, pp. 1–10, 2024.

Sitara Persad, Zi-Ning Choo, Christine Dien, Noor Sohail, Ignas Masilionis, Ronan Chaligné, Tal Nawy, Chrysothemis C Brown, Roshan Sharma, Itsik Pe'er, et al. Seacells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nature Biotechnology*, pp. 1–12, 2023.

Dushyant Rao, Francesco Visin, Andrei Rusu, Razvan Pascanu, Yee Whye Teh, and Raia Hadsell. Continual unsupervised representation learning. *Advances in neural information processing systems*, 32, 2019.

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

Sandra Schick, André F Rendeiro, Kathrin Runggatscher, Anna Ringler, Bernd Boidol, Melanie Hinkel, Peter Májek, Loan Vulliard, Thomas Penz, Katja Parapatics, et al. Systematic characterization of baf mutations provides insights into intracomplex synthetic lethalities in human cancers. *Nature genetics*, 51(9):1399–1410, 2019.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.

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

Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

Marek Wanior, Andreas Krämer, Stefan Knapp, and Andreas C Joerger. Exploiting vulnerabilities of swi/snf chromatin remodelling complexes for cancer therapy. *Oncogene*, 40(21):3637–3654, 2021.

Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3014–3023, 2021.

Appendix

A Overview of datasets and biological analysis

A.1 Disentanglement of perturbation effects in a Genetic Screen

We used the K562 Perturb-Seq dataset from Replogle et al. (2022), which performs CRISPRi-mediated knock-down perturbations in chronic myeloid leukemia cell lines. This dataset has undergone some preliminary filtering, which, for example, has already ensured a minimum of 200 genes per cell and at least 3 cells per gene. In total, the dataset includes data from Perturb-seq screens targeting 2,057 common essential genes. For the purpose of refining our analysis, we followed Lopez et al. (2023) and applied a filtering process to keep only essential genes resulting in a narrowed focus on 1,187 genes.

Further refinement of our analytical dataset involved selecting the 100 most consequential perturbations, determined by cell proliferation outcomes as highlighted in recent findings by Tu et al. (2024).

We subsequently utilized the adpbulk library to aggregate samples of the top 100 perturbations into pseudobulk samples. These pseudobulk samples were then employed for the replay mechanism during the training phase.

The learnable interventions $d_a = 100$ correspond to the number of the most influential perturbations. For the latent space we took $d_z = 30$ as this provided good results before. Therefore G^a has the dimension $dim(G^a) = 100 \times 30$.

A.2 Counterfactual generation of a single-cell gene knock-out in pan-cancer primary tumors

The human SWI/SNF complexes are chromatin remodelers that regulate DNA accessibility in important cellular processes such as transcription, replication and DNA repair. The genes encoding the components of SWI/SNF complexes are mutated in 20%-25% human cancers (Wanior et al., 2021; Schick et al., 2019). BAF is one of the components of the SWI/SNF complexes. Loss-of-function mutations in BAF can contribute to cancer initiation and progression. There are currently a number of drugs that can reduce the viability of BAF-loss mutant cells (Otto et al., 2023). It, therefore, could be possible to intervene on cancer initiation or progression by characterising BAF-loss mutation signatures in transcriptomic profiles of patients.

In the study by Otto et al. (2023), 28 subunits of the SWI/SNF complexes were knocked-out in single cells from a MOLM13 cell line, a model system for human blood cancer (AML). Single-cell gene expression measurements of the perturbed cells were then acquired by single-cell perturbation screen, Perturb-seq. In this work we transferred ARID1A, a subunit of BAF complex frequently mutated in cancers, loss-of-function mutation transcriptional signature from the SWI/SNF Perturb-seq dataset to primary tumors in three cancer types to generate counterfactual instances of tumors under a ARID1A Knock-out (KO) intervention. We then compared counterfactually generated mean gene expression of ARID1A-KO in ARID1A wild type (WT) primary tumors, $\mu_{cf}^{ARID1A.KO}(ARID1A\,wild\,type)$, to the mean gene expression of primary tumors with known ARID1A mutation, $\mu_{obs}(ARID1A\,mutant)$) and assessed if the counterfactual generation shifted the gene expression to the desired state. The bulk RNA-seq profiles for the primary tumors were obtained from the Cancer Genome Atlas (TCGA) [1]. We selected three studies in the TCGA, BLCA (Bladder Urothelial Carcinoma), KIRC (Kidney renal clear cell carcinoma) and LUAD (Lung adenocarcinoma), where BAF mutations are known to be prevalent. We trained our model to learn latent targets of 6 perturbed subunits ($d_a = 6$, $d_z = 30$, $dim(G^a) = 6 \times 30$) from the BAF complex, that is 'ARID1A', 'SMARCA4', 'SMARCB1', 'SMARCC1', 'SMARCD2' and 'SMARCE1' in each of the TCGA studies and reported counterfactual generation for ARID1A-KO only due to its high mutation prevalence. The single cell RNA profiles of these perturbations were pseudobulked and replayed during model training akin to Experience Replay. Counterfactual generation is done by replacing the target latent variables in the encoding of bulk RNA-seq samples with that of encoded Experience Replay samples for each perturbation.

A.3 Counterfactual generation of colorectal cancer single-cell states in colon cancer primary tumors and association with patient survival

Colorectal Cancer is well known for extensive and heterogeneous genomic aberrations. The intra-tumor heterogeneity in CRC presents significant differences in prognoses and responses to treatment. This motivates the study of cellular states that are identifiable in single-cell RNA-sequencing of a limited number of patients, and transferring them to large patient cohorts where both gene expression and additional meta data on patient outcome is available to correlate with outcomes such as relapse, survival or response to treatment. An important note to make is that gene expression profiles of primary tumors are available predominantly in bulk RNA sequencing, where cellular-level information is lost, since a bulk measurement is, in a very overly simplified definition, average expression of genes over all cell types and states. The counterfactual generation of cell states in primary tumors could especially be helpful as bulk RNAseq is more affordable in clinical settings, and the transcriptomics readouts are often coupled with rich patient metadata such as survival outcomes. Therefore, by generating counterfactual instances of tumors expressing specific cell states, we can test for association of the cell state with patient outcome.

A.3.1 Overview of cell state identification in the colorectal cancer single cell data

Following Persad et al. (2023) and Dann et al. (2022), we define a cell state as a group of cells which have homogeneous transcriptional profiles. In this section, we describe the process of identifying cell states from single cell data, including generation of Generative Replay samples.

[1] https://www.cancer.gov/ccg/research/genome-sequencing/tcga

Becker et al. (2022) contains gene expression measurements from 82,6250 single cells collected from 72 sampels from 14 healthy individuals, patients with FAPs (Familial Adenomatous Polyposis), and sporadic CRC patients. We obtain latent representations of cancer and non-cancer single cells by a single cell-specialized (for single cell count gene expression data) VAE model, called SCVI (Lopez et al., 2018). We then built a neighbourhood graph on these lower-dimensional embedding and used Milo (Dann et al., 2022) to identify cells that are in homogeneous transcriptional state and statistically significantly abundant in cancer (single cell) samples, which define a cell state. Akin to Generative Replay, we created instances of the cell states enriched in CRC from the decoder of the SCVI model. These samples are logCPM- transformed before feeding into the model as the as.

Next, we generated counterfactual instances of these cellular states in colon cancer primary tumors from the TCGA COAD study. The COAD study in TCGA consists of 521 bulk RNA-seq samples (521 individuals) and their survival outcomes. We run the model to learn G^a, dimensions 8×30, and the latent interaction variables γ between cell states ($d_a = 8$) and latent representations ($d_z = 30$) in the primary tumors from TCGA COAD data. Counterfactual generation is done by replacing the target latent variables for each cell state in the encoding of bulk RNA-seq samples with that of encoded Generative Replay samples.

B The Spike-Slab VAE model

Lopez et al. (2023) places a spike-slab prior on the latent variables z_i, $i \in \{1, \cdots, m\}$ learned through a VAE to define a causal graph over latent variables and interventions (auxiliary variables) denoted by a. π_i^a denotes the probability with which the intervention a targets the latent variable i. γ_i^a encodes as a binary variable if the latent variable i is targeted by the intervention a.

$$\pi_i^a \sim \mathrm{Beta}(1, k)$$
$$\gamma_i^a \sim \mathrm{Bernoulli}(\pi_i^a)$$
$$z_i | a \sim \gamma_i^a \mathrm{Normal}(\mu_i^a, 1) + (1 - \gamma_i^a)\mathrm{Normal}(0, 1)$$

Note here $d_a = k$.

B.1 Evidence Lower Bound for Spike-Slab VAE

Since the marginal likelihood of the data $p(x|a)$ is intractable, the authors use variational inference to learn the parameters of the model. They approximate the posterior distribution by using the mean-field assumption. The mean-field variational distribution is defined as:

$$\overline{q} = \prod_{n \in [N]} q(z_n | x_n, a_n) \prod_{a_n \in [K], i \in [p]} q(\gamma_i^a) q(\pi_i^a)$$

where each $q(\pi_i^a) = \delta_{\psi_i^a}$ is represented by a Dirac distribution, each $q(\gamma_i^a)$ follows a Bernoulli distribution and each $q(z_n | x_n, a_n)$ follows a Gaussian distribution.

The ELBO is derived as:

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

Using the analytical expressions of the Kullback-Leibler divergence and the simplification of the Dirac distribution they arrive at:

$$\begin{aligned} \log p(X|A) \geq & \mathbb{E}_{\overline{q}} \left[\sum_{n=1}^{N} \log p(x_n | z_n) - D_{KL}(q(z_n | x_n, a_n) \,||\, p(z_n | \gamma_{a_n})) \right] \\ & - \sum_{a \in [K], i \in [p]} D_{KL}(q(\gamma_i^a) \,||\, \text{Bernoulli}(\psi_i^a)) - \log \text{Beta}(\psi_i^a; 1, K) \end{aligned}$$

Then they use the Gumbel-sigmoid distribution as a continuous relaxation of the Bernoulli distribution to apply the reparameterization trick to $q(\gamma)$.

C Baselines

We compare our model (sVAE-ligr) with comparable generative models: VAE (Kingma & Welling, 2013), β-VAE (Higgins et al., 2016), iVAE (Khemakhem et al., 2020) and SpikeSlabVAE (Lopez et al., 2023).

The architectures of the models used, differ noticeably in the choice of the prior $p(z)$. The standard VAE and β-VAE place a normal Gaussian prior on the latent space. The difference being that in β-VAE the Kullback-Leibler divergence term in the evidence lower bound is controlled with a scalar $\beta > 1$. For $\beta = 1$ we have exactly the case of the normal VAE. For $\beta > 1$ the emphasis lies on learning statistically independent latent factors. In contrast, iVAE uses a conditional prior $p(z|a)$. More specifically, it places a factorized exponential family prior on the latent variables, conditioned on an auxiliary random variable a to generate identifiable representations. The auxiliary random variable could be class label or a time index. SpikeSlabVAE follows a similar generative process as iVAE, but uses the spike-slab prior, which is also a conditional prior and can be seen in B. The difference to iVAE is that a stochastic binary mask $\hat{G}^a \sim \mathrm{Bernulli}(\pi_i^a)$ is applied to the location parameter via element wise product. If we interpret the mask $\hat{G}^a$ as a graph, we can identify which latent variables are affected by which intervention a. In addition, the Bernoulli distribution has a Beta distribution as a hyperprior for the parameters of the Bernoulli parameter π_i^a. The hyperprior controls the density of the sparsity of the binary mask.

In our model, denoted as sVAE-ligr, we incorporate the spike-slab prior with an innovative extension. Specifically, we introduce a NN for each cell state, tasked with encoding the generative replay samples, represented as $\tau_a = NN(\tilde{x})$. This encoded information subsequently informs the Beta hyperprior, parameterized as $\pi_i^a \sim \mathrm{Beta}(1, \tau_a)$.

D Evaluation metrics

In our evaluation, we use metrics to examine the disentanglement and the learned causal structures. Through our simulation, we obtain the ground truth data and the graph G^a, which maps the interventions to the latent space.

The metrics we use for disentanglement evaluation are the Mean Correlation Coefficient (MCC) and R^2. We use both, the mean Pearson and the mean Spearman coefficients for our analysis. The MCC is a metric for permutation equivalence and is calculated between pairs of ground truth and estimated latent space for the best possible permutation, while R^2 is a metric used to assess the identifiability of the latent space up to a linear transformation. The metrics precision, recall and F1-score compare the learned adjacency matrix $\hat{G}^a$ with the ground truth G^a , where all metrics consider the permutation equivalence of z. High precision means that most of the predicted causal relationships are accurate, minimizing false positives, whereas high recall means that the model captures a large proportion of the actual causal relationships, minimizing false negatives. The F1 score is the harmonic mean of both values. A high F1 score is an indication of high precision and recall. Additionally we also employ the Structural Hamming Distance (SHD). The SHD metric quantitatively assesses the discrepancy between the learned adjacency matrix, $\hat{G}^a$, and the ground truth, G^a , by counting the number of edge additions or deletions required to transform $\hat{G}^a$ into G^a. This evaluation takes the permutation equivalence of z into account, ensuring a fair comparison of graph structures.

E Experiments with synthetic data

Here we describe simulation of the synthetic data and grid search parameters used to benchmark our model against the baselines.

E.1 Simulation Details

To generate synthetic bulk gene expression count data, we start by creating single-cell gene expression count data based on simulated ground truth latent variables. The basic process is taken from Lopez et al. (2023). For each cell type ($n_{chem} = 80$) we generate one to three targets in the latent

Table 3: Grid search spaces for each baseline.

Hyperparameter space	
VAE	$n_{\text{epochs}} \in \{100, 300, 500\}$
betaVAE	$n_{\text{epochs}} \in \{100, 300, 500\}$, $\beta \in \{2, 4, 8, 10, 30\}$
iVAE	$n_{\text{epochs}} \in \{100, 300, 500\}$
SpikeSlabVAE	$n_{\text{epochs}} \in \{100, 300, 500\}$, $\alpha \in \{0.5, 1, 10, 60, 80, 100\}$
sVAE-ligr	$n_{\text{epochs}} \in \{100, 300, 500\}$, $\alpha \in \{0.5, 1, 10, 60, 80, 100\}$

space, where $p = 15$ is the dimension of the embedding, whereas the the affected dimensions are drawn without replacement from $[p]$ and stored as a binary vector $\beta_{a,\cdot} \in \{0,1\}^p$. Then we can calculate the sparse perturbation embedding $\mu_a = (\mu_{a,1}, \cdots, \mu_{a,p})$ and produce single-cell gene expression data of different cells through the following process:

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

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

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

$$x_{ng} \sim \text{Poisson}(l_n f^g(z_n))$$

e is a scalar quantifying the strength of the perturbation, δ_0 is the Dirac delta distribution with mass at 0, f^g is a decoder and l_n is the library size fixed to 10^5. $x_{n,g}$ is the gene expression of a single cell n and gene g from which we create the gene expression vector $x_n = [x_{n1}, \cdots, x_{n,g}]$.

For each cell, we record the ground truth latent z_n from which it originated. From these single-cell gene expressions we can subsequently create bulk gene expression count data. Each cell is randomly assigned a pseudobulk ID, and the cells are grouped based on this ID. We aggregate the expression data of the selected cells for each pseudobulk sample, summing up the expression levels of each gene across the cells included in the sample. Furthermore, we compute the mean latent representation for each pseudobulk sample by averaging the latent features of the cells included in the sample. At the end, each pseudobulk sample was randomly assigned one of the cell labels from n_{chem}.

For training our model, which necessitates replay data, we employ a method akin to that used for generating single-cell data, with one crucial distinction. Initially, we repurposed the previously calculated sparse perturbation embedding, μ_a, and follow the same procedure to sample z_n. However, to better resemble bulk samples, we adjust the library size within the Poisson distribution to 10^7. Subsequently, each replay sample is annotated with the appropriate cell label.

E.2 Hyper parameter Grid for simulated Data

We made a hyper parameter grid search for the validation of the models. The search range has been adapted for our simulated data set and can be found in **Table 3**. The parameter α denotes the sparse penalty. We also made a data set for each of the four latent space dimensions $d \in \{5, 10, 15, 20\}$ and tested our hyper parameter combinations on these data sets. The hyperparameters were selected using *unsupervised disentanglement ranking* UDR (Duan et al., 2019).