

# Automated Discovery of Pairwise Interactions from Unstructured Data

**Zuheng (David) Xu** \*

Department of Statistics  
University of British Columbia  
Vancouver, Canada  
zuheng.xu@stat.ubc.ca

**Moksh Jain**\*

Department of Computer Science  
Mila / Université de Montréal  
Montreal, Canada  
moksh.jain@mila.quebec

**Ali Denton**

Valence Labs  
Montreal, Canada  
ali@valencelabs.com

**Shawn Whitfield**

Valence Labs  
Montreal, Canada  
shawn@valencelabs.com

**Aniket Didolkar**\*

Department of Computer Science  
Mila / Université de Montréal  
Montreal, Canada  
aniket.didolkar@mila.quebec

**Berton Earnshaw**

Valence Labs & Recursion  
Salt Lake City, USA  
berton@valencelabs.com

**Jason Hartford**

Valence Labs  
London, UK  
jason@valencelabs.com

## Abstract

Pairwise interactions between perturbations to a system can provide evidence for the causal dependencies of the underlying mechanisms of a system. When observations are low dimensional, hand crafted measurements, detecting interactions amounts to simple statistical tests, but it is not obvious how to detect interactions between perturbations affecting latent variables. We derive two interaction tests that are based on pairwise interventions, and show how these tests can be integrated into an active learning pipeline to efficiently discover pairwise interactions between perturbations. We illustrate the value of these tests in the context of biology, where pairwise perturbation experiments are frequently used to reveal interactions that are not observable from any single perturbation. Our tests can be run on unstructured data, such as the pixels in an image, which enables a more general notion of interaction than typical cell viability experiments, and can be run on cheaper experimental assays. We validate on several synthetic and real biological experiments that our tests are able to identify interacting pairs effectively. We evaluate our approach on a real biological experiment where we knocked out 50 pairs of genes and measured the effect with microscopy images. We show that we are able to recover significantly more known biological interactions than random search and standard active learning baselines.

## 1 Introduction

Across the sciences, measurement of pairwise interaction between perturbations often reveals the existence of underlying mechanisms that single perturbations cannot. For example, quantum entanglement experiments support the counterintuitive predictions of quantum physics by demonstrating that entangled particles' spins which we would expect (under classical laws) to be independent, are in fact perfect anti-correlated once we make a measurement. In economics, people may be more (or less) willing to pay for goods when presented in a bundle than they are to pay for each good in isolation, which reveals complements (or substitutes) in the underlying consumer preferences. And

---

\*Work done during an internship at Valence Labs

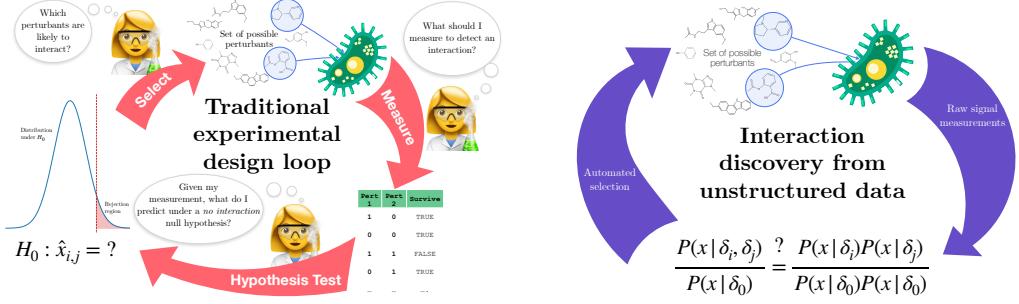


Figure 1: The traditional experimental design loop involves a number of human expert decisions: the expert need to *measure* a carefully chosen feature of the experimental outcome (e.g. did a cell survive a perturbation?), formulate a prediction for the behaviour of this feature under the null hypothesis to test for interactions, and select interacting pairs for the combinatorial space of possible perturbations. Our approach enables automated selection of perturbants and testing for interactions directly from raw signal data (e.g. images of cells under a microscope).

in biology—the application area on which we focus—the concept of *synthetic lethality* [Nijman, 2011] occurs when the simultaneous perturbation of two genes results in cell death while individual perturbations do not. Synthetic lethality reveals that the genes’ associated proteins play redundant roles in the underlying cellular mechanisms.

In order to demonstrate a pairwise interaction a scientist (the human expert) has to carefully design three steps of the experiment:

- Measurement:** the human expert decides on what specific properties of the system to measure in order to reveal the interaction. For example, measuring a particle’s spin may reveal entanglement, while the mass of a particle will not; similarly measuring cell viability may reveal synthetic lethality, while measuring the cell’s colour will not.
- Hypothesis testing:** an interaction is, by definition, a deviation from the effect that we expect under some null hypothesis which assumes each perturbation acts independently. The expert needs to specify the expected outcome assuming independence and compare this prediction to the actual outcome.
- Selection:** there are typically many variables that one could perturb (e.g. there are roughly 20 000 genes in the human genome), but only a small fraction of them will exhibit surprising interactions. The expert selects pairs of perturbations from this space of all possible pairs.

Measuring interactions is further complicated by the fact that measurement and hypothesis testing are interdependent. We can see this in synthetic lethality example, where if we choose to measure the fraction of cells that survived the respective perturbations, then in order to test independence we would need to test whether  $P(\text{survive}_a \cap \text{survive}_b) = P(\text{survive}_a)P(\text{survive}_b)$ . If we had instead measured the fraction of cells that died, we would have needed to test whether  $P(\text{die}_a \cup \text{die}_b) = P(\text{die}_a) + P(\text{die}_b) - P(\text{die}_a)P(\text{die}_b)$ . The state of the cells in the respective petri dishes is the same in both cases, but correctly testing (in)dependence depends on how the expert chose to measure that state. These three steps require significant expertise and knowledge to choose and conduct the measurements, which makes scientific discovery hard to automate and scale.

Modern high throughput screening platforms [Dove, 2007, Baillargeon et al., 2019, Blay et al., 2020, Bock et al., 2022, Fabio et al., 2023, Morris et al., 2023, Messner et al., 2023] allow us to run large scale perturbation experiments that collect information-rich unstructured measurements. For example, cell painting assays [Bra, 2016, Fay et al., 2023, Chandrasekaran et al., 2023], provide microscopy images of cells which capture the same cells that the human expert would have measured, but without pre-committing to measuring a particular property of the cell, such as whether it is alive or dead. This lets us *measure* all the potential properties of interest at once—albeit as unstructured signal data, such as raw pixel images or sensor measurements, rather than preprocessed properties—but it is not clear how to use these measurements for interaction *testing* and *selection* in order to efficiently discover pairwise interactions. This suggests the primary question that this paper seeks to address: if we dispense with prespecified structured measurements, can we automate the discovery of pairwise

interactions by designing testing and selection procedures that efficiently detect interactions from unstructured data?

To test for interactions, we show that pairwise perturbations are *separable* if the pairwise experiments provide no additional information beyond what is already known from the single perturbations. This separability property can be tested by comparing density ratios of the single and double perturbations to the base / control distribution. We then develop an analogous test for the case where pairs of perturbations affect disjoint subsets of the outcome space. For example, if two perturbations affect different organelles in the cell, each of which affect disjoint pixels in our images. This strong form of non-interaction is of interest because it gives sufficient conditions for when we can compose summary statistics of the form  $\mathbb{E}[h(x)]$  from single perturbations, to predict the corresponding statistics from double perturbations. In particular, this lets us compose embeddings of single perturbations to predict the embeddings of double perturbations in a manner analogous to the classic word-vector analogies [Mikolov et al., 2013].

With this notion of an interaction, we address the second task of selection via efficient experimental design. We can search the space of pairwise experiments by selecting pairs of perturbations that are likely to result in large test statistics. In doing so, we reduce the problem of finding interacting pairs of perturbations into an active matrix completion problem. By defining this notion of interaction, we avoid the need to characterize uncertainty over the pixel-level outcomes (or some embedding thereof). Instead we directly model the posterior over an unknown reward matrix defined by the test statistics and at every round select actions striking a balance between exploration and exploitation via information directed sampling [IDS; Russo and Van Roy, 2016, Xu et al., 2022].

We evaluated our approach on both synthetic and real biological experiments. On the synthetic tasks, we found that our tests were able to detect both forms of dependence, validating our theory. On a benchmark consisting of all pairs of gene knockouts for 50 genes in HUVEC cells, we found that our approach using IDS discovers pairs of genes with higher interaction scores significantly faster than baselines, resulting in 10% more known biological interactions being discovered and potentially far more novel interactions. The interactions we detected were also complementary to those which would have been discovered using existing cosine similarity-based approaches, and as a result, the two approaches can be combined to get a more detailed estimate of the relationships between genes from perturbation experiments.

In summary, we develop a system for discovering pairwise interactions through the following contributions,

- We show that two perturbations are separable if and only if, double perturbations provide no additional information than that which was revealed in single perturbations.
- We show that two perturbations are disjoint if and only if their density compose additively. This result also implies sufficient conditions for embeddings of perturbations to be composed additively to predict pairwise perturbations.
- We show that the test scores for detecting interactions can be used to efficiently search for pairs of perturbations that interact using active matrix completion.

## 2 Related work

**Causal representation learning** Our approach to the problem of detecting interactions builds on the modelling assumptions developed in nonlinear independent component analysis [Hyvärinen and Pajunen, 1999, Hyvärinen and Morioka, 2016, Hyvärinen et al., 2019] and causal representation learning [Schölkopf et al., 2021], where we assume we observe some nonlinear mixing function,  $g(\cdot)$ , of the underlying latent variables. The causal representation learning literature typically focuses on disentangling latent variables, whereas we look for testable implications without disentanglement. Our data generating process assumptions are most similar to the interventional setting [Ahuja et al., 2023, Squires et al., 2023, Buchholz et al., 2023, Varıcı et al., 2024]. If we could successfully disentangle latent variables, then our task would be straightforward, but disentanglement is challenging in practice because it is not possible to validate whether an algorithm has succeeded in disentangling latent variables without access to ground truth. Like us, Jiang and Aragam [2023] also attempt to learn latent dependence properties (they characterize the whole latent DAG, not just marginal dependence) without disentanglement, but they assume a fixed bipartite graph of dependence between latents and

observations, which does not apply to the pixel-level observations that we study. Zhang et al. [2023] develop causal representation learning techniques to disentangle latent variables and characterize conditions for extrapolation. Like us, they focus on biological applications, but they rely on stronger polynomial assumptions to achieve disentanglement. Their conditions for extrapolation are dual to our separability tests in that they argue that you can extrapolate when interventions affect *non-overlapping* latents, while we seek to discover when intervention affect *overlapping* latent. Finally, our separability test is closely related to, and inspired by Wang et al. [2023], but whereas they assume separability of concepts to manipulate generative models, we aim to test for an analogous notion of independence in experimental data.

**Representation learning for gene knockouts** Our experiments build on a number of recent works showing the effectiveness of embedding cells presented into a representation space Sypetkowski et al. [2023], Kraus et al. [2023], Xun et al. [2023]. These works show that you can infer that genes code for proteins that form part of complexes by finding embeddings that are highly cosine similar. This works because knocking out proteins on the same pathway will induce the same morphological effect, but this approach is limited to effects that are revealed by single perturbations. There have also been a number of papers that have attempted to learn disentangled representations of cells [e.g. Lotfollahi et al., 2023, Bereket and Karaletsos, 2023, Lopez et al., 2023], mostly from gene expression data.

**Design of Gene Knockout Experiments** There has been considerable work in the recent years to develop algorithms for the design of gene knockout experiments. Typically the problem is studied in the context of discovering single gene-knockouts [Mehrjou et al., 2021] which result in a particular phenotypic effect of interest. A variety of methods including bandits [Pacchiano et al., 2023] and traditional experimental design [Lyle et al., 2023] approaches have been studied in this context. Further, approaches which aim to learn predictors for the effects of unseen gene knockouts typically operate on RNA-seq data [Huang et al., 2023]. RNA-Seq data is far more structured than the image data we consider: the former is a high-dimensional vector of count data where each element corresponds to a particular gene’s expression, while the latter is just an unstructured collection of pixel intensities.

**Experimental design.** The task of designing of experiments that acquire the most information about the system efficiently can be formalized as sequential Bayesian optimal experimental design [BOED; Ryan et al., 2016, Foster, 2021, Rainforth et al., 2023], where the goal to design experiments  $x \in \mathcal{X}$  with outcomes  $y \in \mathcal{Y}$  governed by a generative process  $y \sim p(y | \gamma, x)$  with parameters  $\gamma$ . The experiments are performed sequentially  $(x_1, \dots, x_T)$ , with the objective of maximizing a measure of utility: the information gain [Lindley, 1956, Sebastiani and Wynn, 2000]. While this framework is elegant, it is challenging to apply when  $x$  is high-dimensional because information gain comparisons are very difficult to estimate in high dimensional data. We avoid this by defining the explicit task of detecting interactions.

### 3 Tests for pairwise interactions

In this section we develop tests for both *separable* and *disjoint* interventions. The separability test, Section 3.1, allows us to test whether two perturbations act on disjoint sets of latent variables. Disjointedness, Section 3.2, implies compositional generalization of summary statistics, allowing us to reduce the search space by predicting the outcome of experiments without explicitly running them.

**Setup** We have observations  $X$  in an *observation space*  $\mathcal{X}$  of unstructured measurements such as pixels in an image, and a finite set of perturbations  $\{T_i \in \mathcal{T}_i : i \in [n]\}$ . Although our theory accommodates generic perturbations, we restrict the discussion to binary perturbations for convenience, i.e.,  $\mathcal{T}_i := \{0, 1\}$  for all  $i \in [n]$ , and  $T_i = 1$  means perturbation  $i$  is applied. For all  $i, j \in [n]$ , denote the perturbation indicator as follows:

$$\delta_0 = \{T_{[n]} = 0\} \quad \delta_i = \{T_i = 1, T_{[n] \setminus \{i\}} = 0\} \quad \delta_{ij} = \{T_i = T_j = 1, T_{[n] \setminus \{i,j\}} = 0\}.$$

In this section, we assume that for a pair of perturbations  $T_i, T_j$ , we have access to experimental data from four distributions:  $p(x|\delta_0)$ ,  $p(x|\delta_i)$ ,  $p(x|\delta_j)$ , and  $p(x|\delta_{ij})$ ; when we discuss the active learning procedures in Section 4, we will assume that we have access to all single perturbations distributions,  $p(x|\delta_i)$ , and we will adaptively select the pairs,  $i, j$ , on which to collect samples from  $p(x|\delta_i, \delta_j)$ . Finally, we assume the existence of a set of latent variables  $\{Z_1, \dots, Z_L\} \subseteq \mathcal{Z}$  in some latent space,  $\mathcal{Z}$ , that capture all relevant information about the perturbation. We state this precisely as,

**Assumption 3.1.**  $X \perp\!\!\!\perp (T_1, \dots, T_n) | Z$ , or equivalently,

$$p(x|T_1, \dots, T_n) = \int_{\mathcal{Z}} p(x|z)p(z|T_1, \dots, T_n)dz.$$

Throughout, we assume all distributions have well-defined densities or probability mass functions with respect to some  $\sigma$ -finite base measure on their corresponding sample spaces. We use the same symbol to denote a distribution and its density. All proofs are deferred to Appendix A.

### 3.1 Separability testing

It would be trivial to verify whether two perturbations intervene on different latents if the underlying latent variables were observed directly. However, our observations are unstructured signal data, which we assume is generated by some deterministic mixing function.

**Assumption 3.2.** There exists a diffeomorphism<sup>2</sup>  $g : \mathcal{Z} \rightarrow \mathcal{X}$  such that  $X = g(Z)$ .

**Remark 3.3.** This diffeomorphism assumption between  $\mathcal{Z}$  and  $\mathcal{X}$  can be relaxed; our theory and methodology remain valid as long as the change of variable formula holds for the distributions of  $Z$  and  $X$ . For example, the distribution of  $X$  may have support on a low-dimensional manifold of  $\mathcal{X}$ , with the latent space  $\mathcal{Z}$  having a lower dimension than  $\mathcal{X}$ . See Krantz and Parks [2008, Lemma 5.1.4] for the generalized change of variable formula for non-bijective transformations.

The key observation that we leverage is that when latent variables are independent, the change of variable formula implies that the density ratio of a perturbed distribution to the original (control) distribution has a simple form that only involves the distribution of the intervened latent variable,

$$\frac{p(x|\delta_i)}{p(x|\delta_0)} = \frac{p_Z(g^{-1}(x)|\delta_i) |\det(J(g^{-1}(x)))|}{p_Z(g^{-1}(x)|\delta_0) |\det(J(g^{-1}(x)))|} = \frac{p_{Z_i}^\dagger(g^{-1}(x))}{p_{Z_i}(g^{-1}(x))}.$$

Here  $p_{Z_i}^\dagger$  denotes the perturbed distribution of the latent variable  $Z_i$  targeted by intervention  $i$ . The testable conclusion that we can derive from this observation, is that when two variables are independent, we can predict the density ratio of the double perturbation,  $\frac{p(x|\delta_{ij})}{p(x|\delta_0)}$ , as the product of the density ratios of the respective single perturbations,  $\frac{p(x|\delta_i)}{p(x|\delta_0)} \frac{p(x|\delta_j)}{p(x|\delta_0)}$ . To make this rigorous, we assume that there exists a causal factorization of the latent distribution, where each latent variable is conditionally independent of its non-descendents given its parents. We use  $\text{Pa}(Y)$  (resp.  $\text{Ch}(Y)$ ) to denote the parent (resp. children) of a random variable  $Y$ .

**Assumption 3.4.** The latent distribution  $p_Z(z)$  can be factorized into  $L$  latent factors:

$$p_Z(z|T_1, \dots, T_n) = \prod_{l=1}^L p_{Z_l}(z_l | \text{Pa}(z_l)),$$

where  $\forall l \in [L]$ ,  $\text{Pa}(Z_l) \subseteq T_{[n]} \cup Z_{[L]}$ , and  $\forall i \in [n]$ ,  $\text{Ch}(T_i) \subseteq Z_{[L]}$ . The random variable  $Z$  here should be interpreted as the concatenation of all latent factors  $\{Z_1, \dots, Z_L\}$ .

The observed effects of two non-interacting perturbations are attributed to distinct causal pathways, rather than being confounded by a shared latent factor.

**Definition 3.5.** Two perturbations  $\delta_i, \delta_j$  are **separable** if  $\text{Ch}(T_i) \cap \text{Ch}(T_j) = \emptyset$ .

When two perturbations act separably, the resulting density ratios of the observation distribution will have the predictable interactions that we derived above.

**Theorem 3.6.** Suppose that Assumptions 3.4 and 3.2 hold, and that  $\delta_i, \delta_j$  are separable. Then,

$$\frac{p(x|\delta_{ij})}{p(x|\delta_0)} = \frac{p(x|\delta_i)}{p(x|\delta_0)} \frac{p(x|\delta_j)}{p(x|\delta_0)}. \quad (1)$$

This provides a testable implication of separability, but to test it, we need to derive a real-valued test statistics. By taking logs and rearranging terms, we can rewrite Eq. (1), as testing,

$$\log p(x|\delta_{ij}) + \log p(x|\delta_0) - \log p(x|\delta_i) - \log p(x|\delta_j) \stackrel{?}{=} 0 \quad (2)$$

---

<sup>2</sup>A diffeomorphism is a differentiable bijection with a differentiable inverse.

If we instead take expectations of Eq. (2) with respect to  $p(x|\delta_0)$ , then this amounts to testing if

$$D_{KL}(p_0||p_i) + D_{KL}(p_0||p_j) \stackrel{?}{=} D_{KL}(p_0||p_{i,j}), \quad (3)$$

where  $p_i = p(x|\delta_i)$ ,  $p_0 = p(x|\delta_0)$  and  $D_{KL}(\cdot||\cdot)$  denotes the Kullback–Leibler (KL) divergence. Note that KL divergence measures the difference between two distributions; in our context, it quantifies the distribution shift resulting from interventions. Eq. (3) reflects the intuition that the influence of two separable perturbations, measured by KL divergence, is additive. We use the KL score,  $|D_{KL}(p||p_i) + D_{KL}(p||p_j) - D_{KL}(p||p_{i,j})|$ , to quantify the violation of the separability of  $\delta_i, \delta_j$ , where these KL divergences are estimated using samples.

In practice, we observe that the simple K-nearest neighbor-based (KNN) estimator of the KL divergence [Wang et al., 2009] suffices for low-dimensional problems. However, for high-dimensional data such as images, more dedicated estimation procedures are required [Belghazi et al., 2018, Song and Ermon, 2020, Ghimire et al., 2021]. We detail our estimation procedure in Appendix B.

### 3.2 Disjointedness testing

Our second testing procedure examines whether two perturbations operate on disjoint domains such that their effects are cumulative. For example, if the morphological changes from knocking out two non-interacting genes can be separated into distinct visual features, such that their measures sum, then they are disjoint. We can define disjointedness of two perturbations formally as,

**Definition 3.7.** Two perturbations  $\delta_i, \delta_j$  are **disjoint** if

$$p(x|\delta_{ij}) - p(x|\delta_0) = (p(x|\delta_i) - p(x|\delta_0)) + (p(x|\delta_j) - p(x|\delta_0)). \quad (4)$$

Disjointedness is important, because it implies that we can predict pairwise summary statistics of the distributions from individual perturbations. To see this, let  $h(\cdot)$  denote *any* feature map from the observation,  $X$ . If we have disjoint perturbations  $\delta_i, \delta_j$ , then,

$$\begin{aligned} \mathbb{E}[h(x)|\delta_{ij}] - \mathbb{E}[h(x)|\delta_0] &= \int h(x)[p(x|\delta_{ij}) - p(x|\delta_0)]dx \\ &= \int h(x)[(p(x|\delta_i) - p(x|\delta_0)) + (p(x|\delta_j) - p(x|\delta_0))]dx \\ &= \mathbb{E}[h(x)|\delta_i] - \mathbb{E}[h(x)|\delta_0] + \mathbb{E}[h(x)|\delta_j] - \mathbb{E}[h(x)|\delta_0]. \end{aligned} \quad (5)$$

This implies that we can define average centered embedding vectors,  $\vec{h}_i := \mathbb{E}[h(x)|\delta_i] - \mathbb{E}[h(x)|\delta_0]$  and  $\vec{h}_j$ , and accurately predict  $\vec{h}_{i,j} = \vec{h}_i + \vec{h}_j$  without running the experiments. It is worth noting that there is growing evidence in the literature [Lotfollahi et al., 2019, Gaudellet et al., 2024] that shows that these relationships often hold in real biological experiments (and our experiments support this). Disjointedness explains the sufficient conditions for this to hold. Appendix C.2 explains the choice of  $h$  we tested.

We can test whether Eq. (4) holds, by testing the following null hypothesis:

$$H_0 : \frac{1}{2}p(x|\delta_{ij}) + \frac{1}{2}p(x|\delta_0) = \frac{1}{2}p(x|\delta_i) + \frac{1}{2}p(x|\delta_j).$$

Given interventional data from  $p(x|\delta_0), p(x|\delta_i), p(x|\delta_j)$  and  $p(x|\delta_{ij})$ , we can frame this as a standard two-sample test problem. With balanced experiments, i.e.,  $p(\delta_0) = p(\delta_i) = p(\delta_j) = p(\delta_{ij})$ , we can create samples from the mixture  $\frac{1}{2}p(x|\delta_{ij}) + \frac{1}{2}p(x|\delta_0)$  by combining data from the controlled and doubly perturbed groups. Similarly, we can obtain samples from  $\frac{1}{2}p(x|\delta_i) + \frac{1}{2}p(x|\delta_j)$ . When experiments are heavily unbalanced, we can balance the datasets through downsampling or upsampling.

In practice, we employ the maximal mean discrepancy (MMD) based two-sample test [Gretton et al., 2012], which compares two distributions based on their embeddings in some reproducing kernel Hilbert space (RKHS). The estimated MMD serves as a measure of the extent to which the perturbations violate the disjointedness. Interestingly, MMD test amounts to test Eq. (5) on a “most discriminative” feature map  $h$  in the RKHS. We provide a self-contained introduction about MMD and kernel mean embedding of distributions in Appendix C.

---

**Algorithm 1** ASD for selecting perturbation pairs

---

```

1: Initialize  $H_1 = \{\}$ , batch size  $b$ 
2: for  $t = 1, \dots, T$  do
3:   Estimate posterior  $p(\mathbf{R} | H_t)$ 
4:   Compute information ratio  $\Psi_t$ 
5:   Pick batch  $(a_1, \dots, a_b)$  greedily which minimize  $\Psi_t$ 
6:   Perform experiments and compute  $\{\mathbf{R}_{a_1}, \dots, \mathbf{R}_{a_b}\}$ 
7:   Update  $H_{t+1} = H_t \cup \{(a_1, \mathbf{R}_{a_1}, \dots, (a_b, \mathbf{R}_{a_1})\}$ 
8: end for

```

---

Finally, it is helpful to consider a concrete generative process that yields disjoint perturbations. Unlike the separability formulation, which assumes the existence of the Markov factorization of the latent distribution (Assumption 3.4), the disjointedness models the latent distribution as a finite mixture. In this framework, non-interacting perturbations intervene on different mixing components.

**Assumption 3.8.** *The latent distribution  $p_Z(z)$  admits the form of an  $L$ -component mixture:*

$$p_Z(z|T_1, \dots, T_n) = \sum_{l=1}^L w_l \cdot p_{Z_l}(z|\text{Pa}(z_l)), \quad \text{where } (w_1, \dots, w_L) \in \Delta_L.$$

*In addition, perturbations do not intervene the mixing weights  $(w_1, \dots, w_L)$ .*

**Theorem 3.9.** *Suppose that Assumptions 3.1 and 3.8 hold. Then  $\delta_i, \delta_j$  are disjoint if  $\text{Ch}(T_i) \cap \text{Ch}(T_j) = \emptyset$ .*

## 4 Selecting perturbation pairs to efficiently discover interactions

With principled frameworks for testing separability and disjointedness in place, the natural next question is *how to select the experiments to run?* The space of possible pairs of perturbations is typically too large to perform experiments on all pairs as each experiment is costly. For example, pairwise knockouts of all 20 000 genes on the human genome would require approximately 200 million experiments (not including replicates). In this section we discuss how tools from experimental design and bandits can be used to efficiently select pairs of perturbations likely to have interactions.

**Selection of perturbation pairs as active matrix completion.** The testing frameworks discussed in Section 3 prescribe test statistics which can be used to detect pairwise interactions. The test statistics, however, require samples from  $p(x|\delta_{i,j})$ , which entails running pairwise perturbation experiments. We are thus interested in developing an approach for selecting pairs of perturbations which are *likely* to have high values for the test statistics and are thus likely to reveal pairwise interactions.

The test statistics for pairs of perturbations can be viewed as an (unknown) symmetric matrix  $\mathbf{R} \in \mathbb{R}^{n \times n}$ , where each entry  $\mathbf{R}_{i,j}$  contains the value of the test statistic that we will observe if we run perturbation  $\delta_{ij}$ .  $\mathbf{R}$  is unknown a priori but we are allowed to select entries to observe. This can be viewed as an *active matrix completion* problem [Chakraborty et al., 2013]. Active matrix completion is a variant of the standard matrix completion problem [Laurent, 2009] where values of entries of the matrix can be sequentially queried. Framing the problem of selecting perturbation pairs as an active matrix completion allows us to leverage existing efficient algorithms.

**Adaptive sampling for discovery.** In particular, we use the framework of *adaptive sampling for discovery* [ASD; Xu et al., 2022] which provides a bandit-based approach for active matrix completion. ASD involves using *information directed sampling* [IDS; Russo and Van Roy, 2016] in a “discovery” setting, where an action is only selected once. This is an instantiation of the general sleeping experts setting [Kanade et al., 2009], where the set of available actions shrinks every round.

Let  $\Delta$  denote the action space of possible experiment designs, which in this case is the set of perturbations  $\Delta := \{\delta_{ij} : i, j \in n, i > j\}$ , where  $n$  is the total number of distinct perturbations. To simplify notation, we denote the perturbation pair  $i, j$  selected at step  $k$  as  $a^{(k)} := (i^{(k)}, j^{(k)}) \in \Delta$ .  $\mathcal{D}(\Delta)$  is the set of possible (categorical) distributions defined over  $\Delta$ . Each time an action,  $a^{(k)}$ , is selected, the corresponding element of the (unknown) reward matrix,  $\mathbf{R}$ , is revealed to the agent. In our setting this reward matrix corresponds to the test statistic from either Section 3.1 or 3.2 for

each pair of perturbations. Let  $H_t = ((a^{(k)}, \mathbf{R}_{i^{(k)}, j^{(k)}}))_{k=1}^{t-1}$  denote the history of actions and their corresponding rewards until round  $t$ , and  $\Delta_t$  denotes the set of remaining actions at round  $t$ ; i.e. the pairs of perturbations that we have not yet tested experimentally. A policy  $\pi$  is defined as a map from  $H_t$  to  $\mathcal{D}(\Delta_t)$ . The IDS policy,  $\pi_{\text{IDS}}$  maintains a posterior distribution over  $\mathbf{R}$  given the data observed up to round  $t$ , which we denote  $p(\mathbf{R} | H_t)$ .

We can describe the sub-optimality of any action with respect to a set of beliefs by comparing the action's reward to that of the best action that could have been selected at time  $t$ , under the agent's current posterior over the reward matrix. Intuitively, we can evaluate this by sampling a plausible reward matrix from our posterior,  $\dot{\mathbf{R}} \sim (\mathbf{R}|H_t)$ , and then comparing the reward from action,  $a$ , to the reward an agent would have received from selecting the optimal action,  $a^* = \arg \max_{a \in \Delta_t} \dot{\mathbf{R}}(a)$ ; where  $\dot{\mathbf{R}}(a^{(k)}) := \dot{\mathbf{R}}_{i^{(k)}, j^{(k)}}$  for  $a^{(k)} = (i^{(k)}, j^{(k)})$ . This is known as the expected instantaneous regret incurred by an action and is defined as,

$$\Delta_t(a) = \mathbb{E}_{\dot{\mathbf{R}} \sim p(\mathbf{R}|H_t)} [\dot{\mathbf{R}}(a^*) - \dot{\mathbf{R}}(a)].$$

Additionally, we define the information gain about the top  $T - t + 1$  remaining actions as follows:

$$g(a) = MI(a_{t,1}^*, \dots, a_{t,T-t+1}^*; \mathbf{R}(a) | H_t, \delta_t = a).$$

Algorithm 1 summarizes the algorithmic procedure. The algorithm operates over a series of  $T$  rounds. In each round, the first step is to estimate the posterior distribution over  $\mathbf{R}$  given the data observed thus far. Next step involves selecting a batch of perturbations based on the information ratio. The IDS policy at round  $t$  can then be computed by minimizing the *information ratio*  $\Psi$ :

$$\pi_{\text{IDS}} \in \arg \min_{\pi \in \mathcal{D}(\Delta_t)} \Psi_{\pi,t} := \frac{(\Delta_t^\top \pi)^\lambda}{g_t^\top \pi}$$

where  $\lambda$  is a parameter controls the tradeoff between lower instant regret (exploitation) and higher information gain (exploration). As  $g_t(a)$  is intractable to compute in general, following Russo and Van Roy [2016], Xu et al. [2022] we use an approximation, replacing  $g_t$  with the conditional variance  $v_t(a) = \text{Var}_t(\mathbb{E}[\mathbf{R}_a | a_1^*, a])$ . The conditional variance is a lower bound on the information gain  $g_t(a) \geq v_t(a)$  and can thus replace the mutual information (since we are interested in maximizing it). After selecting the perturbation pair  $i, j$ , we compute the pairwise test statistic,  $\mathbf{R}_{i,j}$  using data from the experimental outcomes. We add these test statistics to our reward matrix, update our posteriors and then continue on to the next round.

Without any structural assumptions on  $\mathbf{R}$ , linear regret is unavoidable for any bandit algorithm. We thus make an assumption that  $\mathbf{R}$  is low-rank with a Gaussian prior on the columns. Under this assumption, ASD achieves sublinear regret [Xu et al., 2022].

**Batching.** In high-throughput experimental screens, it is possible to run multiple experiments in parallel at the cost of a single experiment. So instead of a single action  $\delta_t$ , we can select a set of actions  $\{\delta_t^1, \dots, \delta_t^b\}$  where  $b$  is the number of experiments we can run in parallel. However, as there are no efficient algorithms for combinatorial bandits in the discovery setting, we resort to a simple greedy scheme to select batches with ASD. Specifically, instead of picking a single action which minimizes the information ratio, we pick  $b$  actions with the lowest information ratio.

## 5 Experiments

Our experiments aim to address three objectives: (1) verifying that our theoretical claims about interactions are detectable on known synthetic tasks; (2) evaluating the test statistics' ability to recover known biological relationships from real pairwise perturbation experiments; and (3) assessing our active learning pipeline's efficiency in detecting interactions.

In all the experiments, our MMD-based tests used the RBF and Matern 2.5 kernels, and we chose bandwidth using median heuristics. Unless otherwise stated, we estimated the KL score by first learning the three log-density ratios  $\log \frac{p(x|\delta_{ij})}{p(x|\delta_0)}$ ,  $\log \frac{p(x|\delta_i)}{p(x|\delta_0)}$ ,  $\log \frac{p(x|\delta_j)}{p(x|\delta_0)}$  using contrastive learning [Hermans et al., 2020, NRE], and then obtaining the KL estimates via the smoothed mutual information “lower-bound” estimator (SMILE) [Song and Ermon, 2020] with clipping parameter  $\tau = 5$ . Detailed explanation about the NRE log-density ratio estimator and SMILE are provided in Appendix B. Additional experimental details are provided in Appendix D.

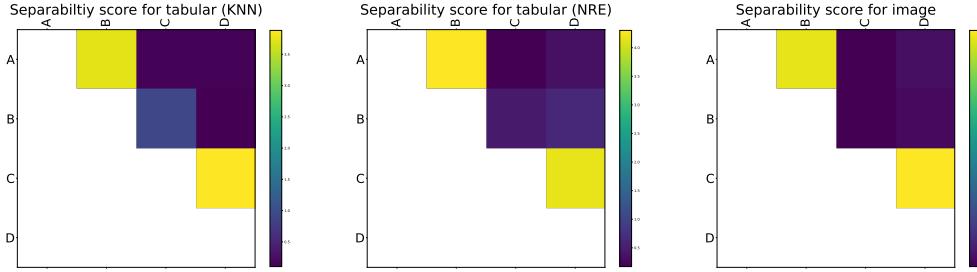


Figure 2: Separability testing on both the synthetic tabular data using KNN-based KL estimator (*left*) and NRE-based KL estimator (*middle*), and the synthetic images (*right*); brighter colors suggest stronger interactions. Ground truth interacting pairs for both examples are A-B and C-D, which are correctly identified.

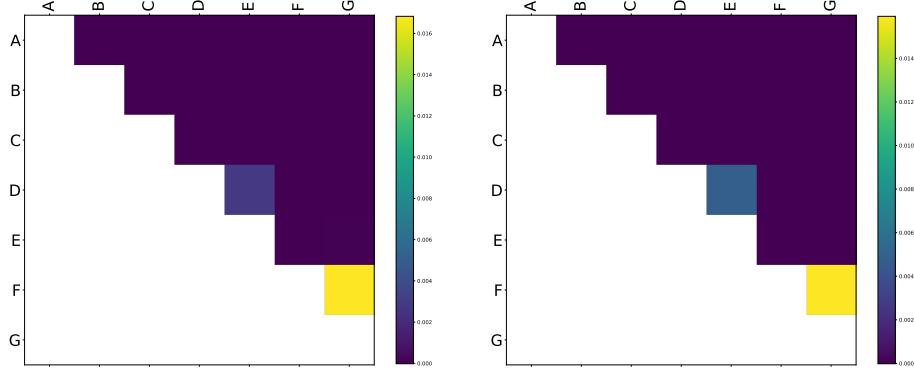


Figure 3: Disjointedness testing on synthetic example using the MMD-based statistics with a Matern 2.5 kernel (*left*) and an RBF kernel (*right*); brighter colors suggest stronger interactions. The ground truth interacting pairs are D-E and F-G, which are correctly identified by the test.

### 5.1 Testing on synthetic setting

We validated the separability test on two synthetic interventional examples: one with 3-dimensional tabular data and one with images (of size  $3 \times 128 \times 128$ ). We also validated the disjointedness test on an interventional tabular example. In each example, we generated observations by sampling from a latent distribution  $p_Z$  that obeys a DAG structure, followed by a mapping  $g(\cdot)$  that maps the latent samples to the observations. We then estimated the test statistics for the corresponding test for each pair of perturbations. Detailed descriptions of the data generating processes for the three examples, the DAG structures of the latent distributions  $p_Z$  and the mapping from latents to observations  $g(\cdot)$ , are provided in Appendix D.1.

Fig. 2 shows the synthetic results for the separability test on both the tabular and image data. The results indicate that, in both examples, our estimated KL score accurately characterized the separability relationships between perturbations: inseparable pairs result in large KL scores, while separable pairs result in small scores. The tests for disjointedness are shown in Fig. 3, demonstrating that the MMD test accurately identified failures of disjointedness and is relatively insensitive to the choice of kernel.

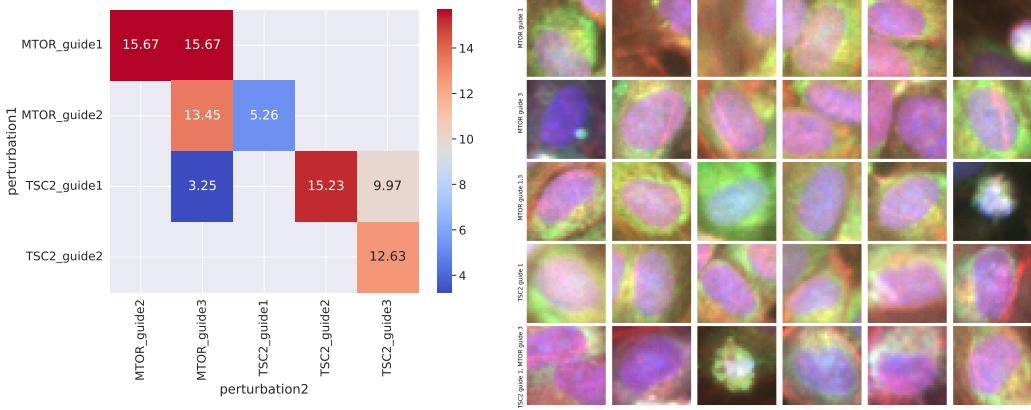


Figure 4: (left) Pairwise separability scores between different CRISPR guides of two genes, TSC2 and MTOR. Missing pairs means that the data for corresponding pairwise combination was not collected in our biological experiment. Guides targeting the same gene show high KL scores (red), while guides targeting different genes show low scores (blue). (right) Random samples of the actual single cell images used in these experiments. Note that detecting the presence or absence of an interaction is extremely difficult, even for trained experts.

## 5.2 Testing on Biological interactions

In order to evaluate whether our separability test can recover known biological interactions, we first ran the following test. In gene knockout experiments, one can target the same gene with multiple different CRISPR guides, each of which cuts the gene in different places, but (at least in theory) results in the same gene being knocked out. Intuitively, guides targeting the same gene should show high scores in the separability test because they are targeting the same latent variable, while guides targeting different genes should show lower scores (assuming the genes are on distinct pathways). While two different guides that target the same gene are not usually run in a single well, we have in our dataset a couple of examples of this data for two genes (TSC2 and MTOR). For this experiment, we used single-cell cell painting images and tested the separability between pairs of guides. The matrix of the separability scores displayed in Fig. 4 was consistent with what we would expect: we observed strong interaction scores between guides targeting the same gene (e.g. MTOR guide 3 with both MTOR guides 1 and 2), and much weaker scores for the interaction between the MTOR and TSC2 targeting guides. It is worth noting that both MTOR and TSC2 affect many systems within the cell, so we should expect some interaction, but the fact that the interaction was far smaller than the interaction score for guides targeting the same gene is very encouraging.

We then evaluated our testing approaches on a collection of 50 genes with the goal of detecting gene interactions. The dataset was collected by performing CRISPR knockouts on all pairs from a set of 50 selected genes, resulting in 1,225 gene pairs. The targeted genes have a bias towards known gene-gene interactions. We performed *in vitro* double gene knockout experiments on HUVEC cells using three CRISPR guides per gene. We label perturbations according to the targeted gene, aggregating the effects of the individual CRISPR guides. The experimental protocol and data preprocessing followed the procedures described by Fay et al. [2023] and Sypetkowski et al. [2023], respectively.

We computed the test statistics using 1024-dimensional embeddings of the original cell painting images extracted from a pre-trained masked autoencoder [He et al., 2022, Kraus et al., 2023]. Fig. 5 shows the matrices of the test statistics for all perturbation pairs; each entry represents the pairwise interaction scores. Qualitatively, both the disjointedness statistics and separability statistics effectively uncover plausible biological relationships. Many genes in the apoptosis pathway (programmed cell death, e.g., gene 2–3: BAX, BCL2L1) and the proteasome (protein degradation e.g., gene 28–30: PSMA1, PSMB2, PSMD1) show high scores that are expected from synthetic lethal relationships. Synthetic lethality (SL) is a complex phenomenon in which simultaneous inactivation of specific gene combinations leads to cell death or extreme sickness, while individual perturbations have little effect. Apoptosis, controlling cell survival, is tightly regulated; disruption of the anti-apoptotic BCL-2 family (BCL2, BCL2L1, MCL1) [Kale et al., 2018] interferes with critical barriers against cell death.

Proteasome function is similarly regulated, as cells must maintain a delicate balance of different proteins [Rousseau and Bertolotti, 2018]. Apoptosis and proteasome members are often found in SL screens in cancer cells [Li et al., 2020, Ge et al., 2024, Han et al., 2017, Cron et al., 2013, Steckel et al., 2012, Das et al., 2020] because cancer genomes accumulate mutations that overcome weakened cellular buffering capabilities.

As displayed in Fig. 5, the MMD-based statistics show strong signals of interactions with proteasome components; the proteasome helps control global protein levels so we would expect to see it interacting with many different pathways, some essential. The separability score provides less clear patterns but highlights several gene pairs that are known or expected to interact physically or genetically, e.g., BCL2L1-MCL1 (gene 3-8) [Shang et al., 2020, Carter et al., 2023], BAX-BCL2L1 (gene 2-3) [Lindqvist and Vaux, 2014], BCL2L1-PSMD1 (gene 3-30) [Craxton et al., 2012], and PSMB2-PSMD1 (gene 29-30) [Voutsadakis, 2017].

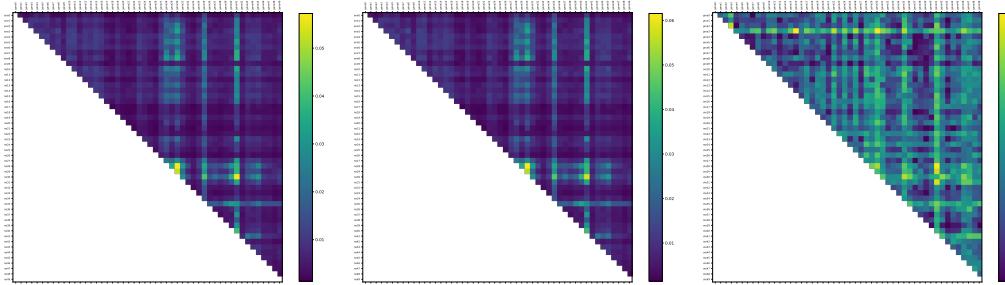


Figure 5: Pairwise interaction scores from a selected set of 50 genes using disjointedness test with Matern 2.5 kernel and RBF kernel (*left* and *middle*, respectively), and separability test (*right*); brighter colors suggest stronger interactions. Genes from the same pathway are ordered adjacently. The associated pathways of each selected gene are described in Table 1 of Appendix D.2.1.

### 5.3 Interaction discovery with active matrix completion

Finally, we showcase how to use our test statistics to adaptively select the pairwise experiment to run. As introduced in Section 4, the adaptive experimental design is framed as an active matrix completion problem (Algorithm 1). We applied Algorithm 1 on the gene-gene interaction detection example using the MMD-based scores.

**Baselines.** To evaluate the effectiveness of our automated selection method, we compare against selection with random policy, upper confidence bound (UCB), Thompson sampling (TS) and uncertainty sampling (US). To instantiate UCB in the discovery setting, we use the uncertainty from the low-rank matrix posterior in place of the counts used in the standard multi-armed bandit setting and mask the actions once they are selected. Similarly, in US we pick pairs based solely on the

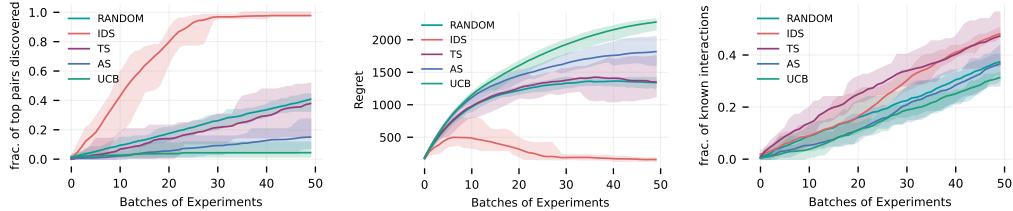


Figure 6: Empirical results on the active pairwise experiments selection for the gene-gene interaction detection example using MMD based test statistics. The solid lines represent the mean performance, whereas the shaded region represent all the runs (min-max). IDS (ours) outperforms all the baselines significantly in terms of top scoring pairs discovered as well as the regret. In terms of known interactions, IDS still outperforms the baselines by a small margin.

uncertainty of the posterior. TS is instantiated in the same way as IDS, but the pairs are selected to minimize only the instant regret. Further details are discussed in Appendix D.

**Evaluation metrics.** We evaluate each approach using three different metrics given a budget of 50 rounds, each with a batch size of 10 resulting in a total of 500 experiments covering 50% of all possible pairs of the gene set. First, we look at the fraction of the pairs with the top 5 percentile of scores recovered by the algorithm capturing the ability to explore the high scoring regions well. Next, we also evaluate the regret of each algorithm with respect to an optimal policy with access to the score matrix and acquires the highest scoring pairs at each round. Finally, we evaluate the number of known biological relations that appear in CORUM [Giurgiu et al., 2019], StringDB [Szklarczyk et al., 2021], Signor [Licata et al., 2020] and hu.MAP [Drew et al., 2021] to see how many each method is able to recover.

**Results.** Fig. 6 (and Fig. 8 in Appendix D.3) illustrate the empirical results. We observe that IDS discovers all the pairs with the top-5 percentile scores, whereas all the baselines barely recover half of the top pairs. Even in terms of the regret, IDS outperforms the baselines, with random performing the worst. This demonstrates that IDS is able to exploit the low-rank structure in the reward matrix effectively. Regarding known interactions discovered, IDS and TS outperform other methods by a slight margin. After 50 rounds, we observe that IDS and TS achieve between 12 and 15% improvement over the other baselines in the number of known biological interactions. The significant difference between the performance on the fraction of top pairs and the number of known relations prompts questions about the correlation between the prediction errors and the biological interactions.

## 6 Discussion

This paper presented a method for efficiently detecting interactions between perturbations. We present interactivity scores for both separability and disjointedness which both allow us to measure interactions between latent variables, and run experiments only on pairs of perturbations for which we will fail to compositionally generalize. From an active learning perspective, disjointedness is powerful: when we can learn a good posterior over this test statistic, we effectively have a “confidence score” for whether two perturbations are likely to compose additively. This allows us to dramatically reduce the number of experiments we run by only experimenting where embeddings are unpredictable. Separability gives a more intuitive notion of (in-)dependence in that we directly test for whether two perturbations interact in latent space. It will be interesting to explore to what extent this can be used to recover the target of interventions.

**Limitations** While many known interactions scored highly according to our test statistics, the overall correlation between known interactions and this metric was relatively low. It is not clear whether this is the result of a lack of specificity, or whether this is because we are in fact discovering real relationships that are not known to biology. To test this would require additional experimentation with orthogonal assays. Additionally, the results for the separability tests depended on the quality of the KL estimator. We used the SMILE estimator which is somewhat sensitive to the choice of the clipping parameter,  $\tau$ . In our experiments on images, a default of  $\tau = 5$  seemed to work well, but if we want this test to be applicable across a variety of modalities, we need a more robust method for choosing this hyper-parameter.

## 7 Acknowledgements

We are extremely grateful for the many discussions with colleagues at Recursion and Valence Labs, and external collaborators that lead to this work. The original design of the benchmark dataset was designed by Marta Fay and the experiments were managed and run by Jordan Finnell, Brandon Mendivil, Kate Brown and Vicky Chen. Thank you to Nathan Lazar for his help in preparing the data and helpful discussions throughout the project. The pairwise interaction testing section was strongly influenced by discussions with everyone at the Bellairs Causal Workshop; special thanks to Victor Veitch, Kartik Ahuja and Yixin Wang for their input. Finally, thank you to Cian Eastwood, Johnny Xi and Jana Osea for the helpful discussions and feedback on the work.

## References

- Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11(9):1757–1774, 2016.
- K. Ahuja, D. Mahajan, Y. Wang, and Y. Bengio. Interventional causal representation learning. In *International conference on machine learning*, pages 372–407. PMLR, 2023.
- P. Baillargeon, V. Fernandez-Vega, B. P. Sridharan, S. Brown, P. R. Griffin, H. Rosen, B. Cravatt, L. Scampavia, and T. P. Spicer. The scripps molecular screening center and translational research institute. *SLAS DISCOVERY: Advancing Life Sciences R&D*, 24(3):386–397, 2019.
- M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm. Mutual information neural estimation. In *International conference on machine learning*, 2018.
- M. Bereket and T. Karaletsos. Modelling cellular perturbations with the sparse additive mechanism shift variational autoencoder. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=DzaCE00jGV>.
- E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. A. Szerlip, P. Horsfall, and N. D. Goodman. Pyro: Deep universal probabilistic programming. *J. Mach. Learn. Res.*, 20:28:1–28:6, 2019. URL <http://jmlr.org/papers/v20/18-403.html>.
- V. Blay, B. Tolani, S. P. Ho, and M. R. Arkin. High-throughput screening: today’s biochemical and cell-based approaches. *Drug Discovery Today*, 25(10):1807–1821, 2020.
- C. Bock, P. Datlinger, F. Chardon, M. A. Coelho, M. B. Dong, K. A. Lawson, T. Lu, L. Maroc, T. M. Norman, B. Song, et al. High-content crispr screening. *Nature Reviews Methods Primers*, 2(1):1–23, 2022.
- J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL <http://github.com/google/jax>.
- S. Buchholz, G. Rajendran, E. Rosenfeld, B. Aragam, B. Schölkopf, and P. Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing, 2023.
- B. Z. Carter, P. Y. Mak, W. Tao, E. Ayoub, L. B. Ostermann, X. Huang, S. Loghavi, S. Boettcher, Y. Nishida, V. Ruvolo, P. E. Hughes, P. K. Morrow, T. Haferlach, S. Kornblau, M. Muftuoglu, and M. Andreeff. Combined inhibition of bcl-2 and mcl-1 overcomes bax deficiency-mediated resistance of tp53-mutant acute myeloid leukemia to individual bh3 mimetics. *Blood Cancer Journal*, 13(1):57, 2023. doi: 10.1038/s41408-023-00830-w. URL <https://doi.org/10.1038/s41408-023-00830-w>.
- S. Chakraborty, J. Zhou, V. Balasubramanian, S. Panchanathan, I. Davidson, and J. Ye. Active matrix completion. In *2013 IEEE 13th international conference on data mining*, pages 81–90. IEEE, 2013.
- S. N. Chandrasekaran, J. Ackerman, E. Alix, D. M. Ando, J. Arevalo, M. Bennion, N. Boisseau, A. Borowa, J. D. Boyd, L. Brino, P. J. Byrne, H. Ceulemans, C. Ch’ng, B. A. Cimini, D.-A. Clevert, N. Deflaux, J. G. Doench, T. Dorval, R. Doyonnas, V. Dragone, O. Engkvist, P. W. Faloon, B. Fritchman, F. Fuchs, S. Garg, T. J. Gilbert, D. Glazer, D. Gnutt, A. Goodale, J. Grignard, J. Guenther, Y. Han, Z. Hanifehlou, S. Hariharan, D. Hernandez, S. R. Horman, G. Hormel, M. Huntley, I. Icke, M. Iida, C. B. Jacob, S. Jaensch, J. Khetan, M. Kost-Alimova, T. Krawiec, D. Kuhn, C.-H. Lardeau, A. Lembke, F. Lin, K. D. Little, K. R. Lofstrom, S. Lotfi, D. J. Logan, Y. Luo, F. Madoux, P. A. M. Zapata, B. A. Marion, G. Martin, N. J. McCarthy, L. Mervin, L. Miller, H. Mohamed, T. Monteverde, E. Mouchet, B. Nicke, A. Ogier, A.-L. Ong, M. Osterland, M. Otrcka, P. J. Peeters, J. Pilling, S. Precht, C. Qian, K. Rataj, D. E. Root, S. K. Sakata, S. Scrace, H. Shimizu, D. Simon, P. Sommer, C. Spruiell, I. Sumia, S. E. Swalley, H. Terauchi, A. Thibaudeau, A. Unruh, J. V. de Waeter, M. V. Dyck, C. van Staden, M. Warchoł, E. Weisbart, A. Weiss, N. Wiest-Daessle, G. Williams, S. Yu, B. Zapiec, M. Żyła, S. Singh, and A. E. Carpenter. Jump cell painting dataset: morphological impact of 136,000 chemical and genetic perturbations. *bioRxiv*, 2023.

- A. Craxton, M. Butterworth, N. Harper, L. Fairall, J. Schwabe, A. Ciechanover, and G. M. Cohen. Noxa, a sensor of proteasome integrity, is degraded by 26s proteasomes by an ubiquitin-independent pathway that is blocked by mcl-1. *Cell Death Differ*, 19(9):1424–1434, Sep 2012. ISSN 1476-5403 (Electronic); 1350-9047 (Print); 1350-9047 (Linking). doi: 10.1038/cdd.2012.16.
- K. R. Cron, K. Zhu, D. S. Kushwaha, G. Hsieh, D. Merzon, J. Rameseder, C. C. Chen, A. D. D’Andrea, and D. Kozono. Proteasome inhibitors block dna repair and radiosensitize non-small cell lung cancer. *PLoS One*, 8(9):e73710, 2013. ISSN 1932-6203 (Electronic); 1932-6203 (Linking). doi: 10.1371/journal.pone.0073710.
- S. Das, X. Deng, K. Camphausen, and U. Shankavaram. Synthetic lethal drug combinations targeting proteasome and histone deacetylase inhibitors in tp53-mutated cancers. *Arch Cancer Biol Ther*, 1 (2):42–47, 2020.
- M. D. Donsker and S. S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on pure and applied mathematics*, 36(2):183–212, 1983.
- A. Dove. High-throughput screening goes to school. *Nature Methods*, 4(6):523–532, 2007.
- K. Drew, J. B. Wallingford, and E. M. Marcotte. hu. map 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies. *Molecular systems biology*, 17(5):e10016, 2021.
- S. Fabio, K. S. Pankaj, S. Kazem, M. Michela, L. Demetrio, and A. M. Michael. High throughput microscopy and single cell phenotypic image-based analysis in toxicology and drug discovery. *Biochemical Pharmacology*, page 115770, 2023.
- M. M. Fay, O. Kraus, M. Victors, L. Arumugam, K. Vuggumudi, J. Urbanik, K. Hansen, S. Celik, N. Cernek, G. Jagannathan, et al. Rxrx3: Phenomics map of biology. *bioRxiv*, pages 2023–02, 2023.
- A. Foster. *Variational, Monte Carlo and policy-based approaches to Bayesian experimental design*. PhD thesis, University of Oxford, 2021.
- T. Gaudelet, A. Del Vecchio, E. M. Carrami, J. Cudini, C.-A. Kapourani, C. Uhler, and L. Edwards. Season combinatorial intervention predictions with salt & peper. *arXiv preprint arXiv:2404.16907*, 2024.
- M. Ge, J. Luo, Y. Wu, G. Shen, and X. Kuang. The biological essence of synthetic lethality: Bringing new opportunities for cancer therapy. *MedComm – Oncology*, 3(1):e70, 2024. doi: <https://doi.org/10.1002/mog2.70>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/mog2.70>.
- S. Ghimire, A. Masoomi, and J. Dy. Reliable estimation of kl divergence using a discriminator in reproducing kernel hilbert space. In *Advances in Neural Information Processing Systems*, 2021.
- M. Giurgiu, J. Reinhard, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, C. Montrone, and A. Ruepp. Corum: the comprehensive resource of mammalian protein complexes—2019. *Nucleic acids research*, 47(D1):D559–D563, 2019.
- I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, 2014.
- A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.
- K. Han, E. E. Jeng, G. T. Hess, D. W. Morgens, A. Li, and M. C. Bassik. Synergistic drug combinations for cancer identified in a crispr screen for pairwise genetic interactions. *Nat Biotechnol*, 35(5): 463–474, May 2017. ISSN 1546-1696 (Electronic); 1087-0156 (Print); 1087-0156 (Linking). doi: 10.1038/nbt.3834.
- K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

- J. Hermans, V. Begy, and G. Louppe. Likelihood-free mcmc with amortized approximate ratio estimators. In *International conference on machine learning*, pages 4239–4248. PMLR, 2020.
- R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019.
- K. Huang, R. Lopez, J.-C. Hutter, T. Kudo, A. Rios, and A. Regev. Sequential optimal experimental design of perturbation screens guided by multi-modal priors. *bioRxiv*, pages 2023–12, 2023.
- A. Hyvarinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in Neural Information Processing Systems*, 29, 2016.
- A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999. ISSN 0893-6080. doi: [https://doi.org/10.1016/S0893-6080\(98\)00140-3](https://doi.org/10.1016/S0893-6080(98)00140-3). URL <https://www.sciencedirect.com/science/article/pii/S0893608098001403>.
- A. Hyvarinen, H. Sasaki, and R. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.
- Y. Jiang and B. Aragam. Learning nonparametric latent causal graphs with unknown interventions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=S8DFqgmEbe>.
- J. Kale, E. J. Osterlund, and D. W. Andrews. BCL-2 family proteins: changing partners in the dance towards death. *Cell Death & Differentiation*, 25(1):65–80, Jan. 2018. ISSN 1476-5403. doi: 10.1038/cdd.2017.186. URL <https://doi.org/10.1038/cdd.2017.186>.
- V. Kanade, H. B. McMahan, and B. Bryan. Sleeping experts and bandits with stochastic action availability and adversarial rewards. In *Artificial Intelligence and Statistics*, pages 272–279. PMLR, 2009.
- D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017. URL <https://arxiv.org/abs/1412.6980>.
- S. G. Krantz and H. R. Parks. *Geometric integration theory*. Springer Science & Business Media, 2008.
- O. Kraus, K. Kenyon-Dean, S. Saberian, M. Fallah, P. McLean, J. Leung, V. Sharma, A. Khan, J. Balakrishnan, S. Celik, et al. Masked autoencoders are scalable learners of cellular morphology. *arXiv preprint arXiv:2309.16064*, 2023.
- M. Laurent. Matrix completion problems. *Encyclopedia of Optimization*, 3:221–229, 2009.
- S. Li, W. Topatana, S. Juengpanich, J. Cao, J. Hu, B. Zhang, D. Ma, X. Cai, and M. Chen. Development of synthetic lethality in cancer: molecular and cellular classification. *Signal Transduction and Targeted Therapy*, 5(1):241, 2020. doi: 10.1038/s41392-020-00358-6. URL <https://doi.org/10.1038/s41392-020-00358-6>.
- L. Licata, P. Lo Surdo, M. Iannuccelli, A. Palma, E. Micarelli, L. Perfetto, D. Peluso, A. Calderone, L. Castagnoli, and G. Cesareni. Signor 2.0, the signaling network open resource 2.0: 2019 update. *Nucleic acids research*, 48(D1):D504–D510, 2020.
- D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.
- L. M. Lindqvist and D. L. Vaux. Bcl2 and related prosurvival proteins require bak1 and bax to affect autophagy. *Autophagy*, 10(8):1474–1475, Aug 2014. ISSN 1554-8635 (Electronic); 1554-8627 (Print); 1554-8627 (Linking). doi: 10.4161/auto.29639.

- R. Lopez, N. Tagasovska, S. Ra, K. Cho, J. Pritchard, and A. Regev. Learning causal representations of single cells via sparse mechanism shift modeling. In M. van der Schaar, C. Zhang, and D. Janzing, editors, *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pages 662–691. PMLR, 11–14 Apr 2023. URL <https://proceedings.mlr.press/v213/lopez23a.html>.
- M. Lotfollahi, F. A. Wolf, and F. J. Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019.
- M. Lotfollahi, A. Klimovskaia Susmelj, C. De Donno, L. Hetzel, Y. Ji, I. L. Ibarra, S. R. Srivatsan, M. Naghipourfar, R. M. Daza, B. Martin, J. Shendure, J. L. McFaline-Figueroa, P. Boyeau, F. A. Wolf, N. Yakubova, S. Günemann, C. Trapnell, D. Lopez-Paz, and F. J. Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023. doi: <https://doi.org/10.15252/msb.202211517>. URL <https://www.embopress.org/doi/abs/10.15252/msb.202211517>.
- C. Lyle, A. Mehrjou, P. Notin, A. Jesson, S. Bauer, Y. Gal, and P. Schwab. Discobax discovery of optimal intervention sets in genomic experiment design. In *International Conference on Machine Learning*, pages 23170–23189. PMLR, 2023.
- A. Mehrjou, A. Soleymani, A. Jesson, P. Notin, Y. Gal, S. Bauer, and P. Schwab. Genedisco: A benchmark for experimental design in drug discovery. In *International Conference on Learning Representations*, 2021.
- C. B. Messner, V. Demichev, Z. Wang, J. Hartl, G. Kustatscher, M. Mülleder, and M. Ralser. Mass spectrometry-based high-throughput proteomics and its role in biomedical studies and systems biology. *Proteomics*, 23(7-8):2200013, 2023.
- T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- J. A. Morris, J. S. Sun, and N. E. Sanjana. Next-generation forward genetic screens: uniting high-throughput perturbations with single-cell analysis. *Trends in Genetics*, 2023.
- K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.
- S. M. Nijman. Synthetic lethality: General principles, utility and detection using genetic screens in human cells. *FEBS Letters*, 585(1):1–6, 2011. ISSN 0014-5793.
- A. Pacchiano, D. Wulsin, R. A. Barton, and L. Voloch. Neural design for genetic perturbation experiments. In *The Eleventh International Conference on Learning Representations*, 2023.
- D. Phan, N. Pradhan, and M. Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*, 2019.
- T. Rainforth, A. Foster, D. R. Ivanova, and F. B. Smith. Modern bayesian experimental design. *arXiv preprint arXiv:2302.14545*, 2023.
- A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.
- R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR, 2014.
- G. Roeder, L. Metz, and D. Kingma. On linear identifiability of learned representations. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9030–9039. PMLR, 18–24 Jul 2021. URL <https://proceedings.mlr.press/v139/roeder21a.html>.
- A. Rousseau and A. Bertolotti. Regulation of proteasome assembly and activity in health and disease. *Nature Reviews Molecular Cell Biology*, 19(11):697–712, 2018. doi: 10.1038/s41580-018-0040-z. URL <https://doi.org/10.1038/s41580-018-0040-z>.

- D. Russo and B. Van Roy. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471, 2016.
- E. G. Ryan, C. C. Drovandi, J. M. McGree, and A. N. Pettitt. A review of modern computational algorithms for bayesian optimal design. *International Statistical Review*, 84(1):128–154, 2016.
- B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.
- P. Sebastiani and H. P. Wynn. Maximum entropy sampling and optimal bayesian experimental design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):145–157, 2000.
- E. Shang, T. T. T. Nguyen, C. Shu, M.-A. Westhoff, G. Karpel-Massler, and M. D. Siegelin. Epigenetic targeting of mcl-1 is synthetically lethal with bcl-xl/bcl-2 inhibition in model systems of glioblastoma. *Cancers (Basel)*, 12(8), Aug 2020. ISSN 2072-6694 (Print); 2072-6694 (Electronic); 2072-6694 (Linking). doi: 10.3390/cancers12082137.
- J. Song and S. Ermon. Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations*, 2020.
- C. Squires, A. Seigal, S. S. Bhate, and C. Uhler. Linear causal disentanglement via interventions. In *International Conference on Machine Learning*. PMLR, 2023.
- B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011.
- M. Steckel, M. Molina-Arcas, B. Weigelt, M. Marani, P. H. Warne, H. Kuznetsov, G. Kelly, B. Saunders, M. Howell, J. Downward, and D. C. Hancock. Determination of synthetic lethal interactions in kras oncogene-dependent cancer cells reveals novel therapeutic targeting strategies. *Cell Res*, 22(8):1227–1245, Aug 2012. ISSN 1748-7838 (Electronic); 1001-0602 (Print); 1001-0602 (Linking). doi: 10.1038/cr.2012.82.
- M. Sypetkowski, M. Rezanejad, S. Saberian, O. Kraus, J. Urbanik, J. Taylor, B. Mabey, M. Victors, J. Yosinski, A. R. Sereshkeh, et al. Rxrx1: A dataset for evaluating experimental batch correction methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4284–4293, 2023.
- D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, P. Bork, et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612, 2021.
- B. Varıcı, E. Acartürk, K. Shanmugam, A. Kumar, and A. Tajer. Score-based causal representation learning: Linear and general transformations, 2024.
- I. A. Voutsadakis. Proteasome expression and activity in cancer and cancer stem cells. *Tumor Biology*, 39(3):1010428317692248, 2017.
- Q. Wang, S. R. Kulkarni, and S. Verdú. Divergence estimation for multidimensional densities via  $k$ -nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405, 2009.
- Z. Wang, L. Gui, J. Negrea, and V. Veitch. Concept algebra for (score-based) text-controlled generative models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=SG1rCuwdsB>.
- D. Wingate and T. Weber. Automated variational inference in probabilistic programming. *arXiv preprint arXiv:1301.1299*, 2013.
- Z. Xu, E. Shim, A. Tewari, and P. Zimmerman. Adaptive sampling for discovery. *Advances in Neural Information Processing Systems*, 35:1114–1126, 2022.
- D. Xun, R. Wang, X. Zhang, and Y. Wang. Microsnoop: A generalized tool for unbiased representation of diverse microscopy images. *bioRxiv*, 2023. doi: 10.1101/2023.02.25.530004. URL <https://www.biorxiv.org/content/early/2023/05/06/2023.02.25.530004>.

J. Zhang, K. Greenewald, C. Squires, A. Srivastava, K. Shanmugam, and C. Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 2023.

# Appendix

## Table of Contents

<b>A Proofs</b>	<b>19</b>
<b>B Sample-based estimation of KL divergence</b>	<b>20</b>
B.1 Contrastive neural ratio estimator . . . . .	20
B.2 Smoothed Mutual Information “Lower-bound” Estimator (SMILE) . . . . .	21
<b>C MMD and kernel mean embedding</b>	<b>21</b>
C.1 Kernel mean embedding . . . . .	21
C.2 Compare distributions via a fixed feature map $h$ . . . . .	23
<b>D Experiment details</b>	<b>23</b>
D.1 Synthetic examples . . . . .	23
D.2 Real data examples . . . . .	25
D.3 Active learning . . . . .	27

---

## A Proofs

*Proof of Theorem 3.6.* By Assumptions 3.4 and 3.2, we can apply change of variable formula to obtain the log density of  $p_X(x|T_1, \dots, T_n)$ :

$$\log p(x|T_1, \dots, T_n) = \sum_{l=1}^L \log p_{z_l}([g^{-1}(x)]_l | \text{Pa}(z_l)) + \log |\det \nabla g^{-1}(x)|.$$

Here  $[ \cdot ]_l$  denote the projection operator that maps  $z \in \mathcal{Z}$  to the subspace on which  $z_l$  lies, i.e.,  $\forall z \in \mathcal{Z}, [z]_l = z_l$ . Recall that we interpret  $z$  as the concatenation of all latent factors  $(z_1, \dots, z_L)$ . If  $T_i, T_j$  are causally independent, perturbation  $\delta_i, \delta_j$  will intervene different terms in the above summand. Without lose of generality, suppose  $\delta_i$  and  $\delta_j$  intervene  $z_{l_i}$  and  $z_{l_j}$  respectively. Then,

$$\begin{aligned} \bar{\ell}_X(x|\delta_i) &= \log p(x|\delta_i) - \log p(x|\delta_0) \\ &= \log p_{z_{l_i}}([g^{-1}(x)]_{l_i}|\delta_i) - \log p_{z_{l_i}}([g^{-1}(x)]_{l_i}|\delta_0). \end{aligned}$$

where the equality follows from the modularity assumption that the intervention only affects  $l_i$ .

Similarly, we obtain that

$$\begin{aligned} \bar{\ell}_X(x|\delta_j) &= \log p_{z_{l_j}}([g^{-1}(x)]_{l_j}|\delta_j) - \log p_{z_{l_j}}([g^{-1}(x)]_{l_j}|\delta_0), \\ \bar{\ell}_X(x|\delta_{ij}) &= \log p_{z_{l_i}}([g^{-1}(x)]_{l_i}|\delta_i) - \log p_{z_{l_i}}([g^{-1}(x)]_{l_i}|\delta_0) \\ &\quad + \log p_{z_{l_j}}([g^{-1}(x)]_{l_j}|\delta_j) - \log p_{z_{l_j}}([g^{-1}(x)]_{l_j}|\delta_0), \end{aligned}$$

which completes the proof.  $\square$

*Proof of Theorem 3.9.* Provided with Assumption 3.1,  $\delta_i, \delta_j$  being disjoint (Definition 3.7) is equivalent to the same additivity in the intervened latent distributions, i.e.,

$$p_Z(z|\delta_{ij}) - p_Z(z|\delta_0) = p_Z(z|\delta_i) - p_Z(z|\delta_0) + p_Z(z|\delta_j) - p_Z(z|\delta_0). \quad (6)$$

Under the mixture model assumption Assumption 3.8, if  $T_i, T_j$  are causally independent, if  $\text{Ch}(T_i) \cap \text{Ch}(T_j) = \emptyset$ , perturbation  $\delta_i, \delta_j$  will intervene distinct mixing components. Without lose of generality,

suppose  $\delta_i$  and  $\delta_j$  intervene  $p_{Z_{l_i}}$  and  $p_{Z_{l_j}}$  respectively.

$$\begin{aligned} p_Z(z|\delta_i) - p_Z(z|\delta_0) &= w_{l_i} \cdot (p_{Z_{l_i}}(z|\delta_i) - p_{Z_{l_i}}(z|\delta_0)), \\ p_Z(z|\delta_j) - p_Z(z|\delta_0) &= w_{l_j} \cdot (p_{Z_{l_j}}(z|\delta_j) - p_{Z_{l_j}}(z|\delta_0)), \\ p_Z(z|\delta_{ij}) - p_Z(z|\delta_0) &= w_{l_i} \cdot (p_{Z_{l_i}}(z|\delta_i) - p_{Z_{l_i}}(z|\delta_0)) + w_{l_j} \cdot (p_{Z_{l_j}}(z|\delta_j) - p_{Z_{l_j}}(z|\delta_0)), \end{aligned}$$

which shows Eq. (6) and hence completes the proof.  $\square$

## B Sample-based estimation of KL divergence

Sample-based estimator of KL-divergence is a challenging task, particularly for high dimensional observations. In this section, we present the estimation procedure we used in our experiments. As mentioned in Section 5, to estimate the KL-divergence,

$$D_{\text{KL}}(p||p_i) = \mathbb{E} \left[ \log \frac{p(X|\delta_0)}{p(X|\delta_i)} \right], \quad X \sim p(X|\delta_0), \quad (7)$$

we used a two-step procedure:

1. We first estimate the log-density ratio

$$\log \frac{p(x|\delta_0)}{p(x|\delta_i)} \quad (8)$$

via a neural ratio estimator (NRE) based on a contrastive learning objective [Hermans et al., 2020, NRE].

2. Then, instead of taking naive Monte Carlo estimates of the KL-divergence, we adopt the SMILE estimator [Song and Ermon, 2020] using the learned log-density ratio.

Appendices B.1 and B.2 describe the two steps respectively.

### B.1 Contrastive neural ratio estimator

There are various methods to learn the log-density ratio between a single pair of distributions. For example, the optimal discriminator between two distributions is related to their density ratios [Goodfellow et al., 2014, Proposition 1]. However, a common structure in our applications involves a large set of perturbations, and training a classifier for each perturbation class (against the control group) would be very cumbersome and not data efficient. Therefore, we consider a contrastive learning model [Hermans et al., 2020, NRE], which trains a binary classifier to distinguish the joint data distribution  $p(x, c)$  from the product of the marginals  $p(x)p(c)$ , where  $c$  denotes the perturbation class.

The training objective is as follows:

$$\theta, w \in \arg \min_{\theta, w} -\frac{1}{2B} \left[ \sum_{b=1}^B \log(1 - \text{Sigmoid}(f_{\theta, w}(x^{(b)}, c^{(b)}))) + \sum_{b'=1}^B \log(\text{Sigmoid}(f_{\theta, w}(x^{(b')}, c^{(b')}))) \right],$$

where  $x^{(b)}, c^{(b)} \sim p(x)p(c)$  and  $x^{(b')}, c^{(b')} \sim p(x, c)$ , and

$$f_{\theta, W}(x, c) = \text{Encoder}_{\theta}(x)^T W_c.$$

Given infinite training samples and flexibility of the neural network  $f$ , the optimal  $f^*(x, c) = \log \frac{p(x, c)}{p(x)p(c)} = \log \frac{p(x|c)}{p(x)}$ . And we can then obtain Eq. (8) via  $f^*(x, \delta_i) - f^*(x, \delta_0)$ .

In our examples, we estimate our log-density ratios by training the NRE objective on all perturbation classes.

## B.2 Smoothed Mutual Information “Lower-bound” Estimator (SMILE)

After obtaining the log-density ratio estimator for Eq. (8), there are several options for estimating the KL-divergence. We provide a short review here on the various strategies [Ghimire et al., 2021, Belghazi et al., 2018, Hjelm et al., 2019, Song and Ermon, 2020], and explain why we opt for the SMILE estimator [Song and Ermon, 2020].

The most straightforward estimates of the KL-divergence Eq. (7) is by the Monte Carlo estimates based on samples from  $p(X|\delta_0)$ , i.e.,

$$D_{\text{KL}}(P(X|\delta_i) || P(X|\delta_0)) \approx \frac{1}{N} \sum_{i=1}^N f(X_i), \quad \text{where } f(\cdot) \text{ is the estimated } \log \frac{p(\cdot|\delta_i)}{p(\cdot|\delta_0)}.$$

However, it is noted that the variance of this estimator is often huge [Song and Ermon, 2020, Ghimire et al., 2021], making the estimated KL unreliable in practice.

Belghazi et al. [2018] proposed to estimate the KL-divergence based on its Donsker-Varadhan representation [Donsker and Varadhan, 1983], i.e.,

$$D_{\text{KL}}(p||q) = \sup_{f:\Omega \rightarrow \mathbb{R}} \mathbb{E}_p[f] - \log \mathbb{E}_q[\exp(f)],$$

where the supremum is taken over all functions  $f$  such that the two expectations are finite. Notice that the optimal  $f$  is indeed achieved as the log-density ratio between  $p$  and  $q$ . In practice, one can parameterize  $f$  using some neural network and maximize the above objective to approximate the KL [Belghazi et al., 2018]. However, the stochastic gradient estimator of the above objective is generally biased [Belghazi et al., 2018], making the optimization less stable. Therefore, it is also recommended to first learn the log-density ratio  $\log \frac{p}{q}$  as  $f$ , and then estimate the KL-divergence [Hjelm et al., 2019, Song and Ermon, 2020] via

$$\text{KL}_{\text{MILE}}(f) = \mathbb{E}_p[f] - \log \mathbb{E}_q[\exp(f)],$$

which technically gives a lower-bound of the KL divergence if the log-density ratio is not well-estimated. We would refer this estimator as the mutual information lower-bound estimator (MILE). MILE is often shown to have better performance than the naive Monte Carlo estimator [Belghazi et al., 2018, Hjelm et al., 2019, Song and Ermon, 2020].

However, MILE also suffers from high variance issues [Song and Ermon, 2020] particularly when the learned  $f$  has large values in the tail of  $q$ . Song and Ermon [2020] proposed a smoothed version of MILE, named as SMILE, by clipping the learned log-density ratios  $f$  between  $-\tau$  and  $\tau$ , i.e.,

$$\text{KL}_{\text{SMILE}}(f, \tau) = \mathbb{E}_p[f] - \log \mathbb{E}_q[\text{clip}(\exp(f), \exp(-\tau), \exp(\tau))].$$

$\text{KL}_{\text{SMILE}}(f, \tau)$  converges to  $\text{KL}_{\text{MILE}}(f)$  as  $\tau \rightarrow \infty$ , but smaller  $\tau$  significantly reduces the variance of the MILE estimator.

In our experiments, we use the SMILE estimator with the clipping parameter  $\tau$  set to be 5.

## C MMD and kernel mean embedding

### C.1 Kernel mean embedding

We provide here a minimal overview of the kernel mean embedding of probability measures [Muandet et al., 2017]. Given a feature map  $\phi : \mathcal{X} \rightarrow \mathcal{F}_\phi$ , where  $\mathcal{F}_\phi$  is some Hilbert space (sometimes called the feature space). This feature map  $\phi$  defines a kernel:

$$k_\phi : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}, \quad k_\phi(x, y) = \langle \phi(x), \phi(y) \rangle,$$

where  $\langle \cdot, \cdot \rangle$  denotes the inner product of  $\mathcal{F}_\phi$ . Kernel  $k_\phi$  defined this way induces a space of functions— $\mathcal{H}_\phi$ —from  $\mathcal{X}$  to  $\mathbb{R}$ , which is a *reproducing kernel Hilbert space* (RKHS). The name RKHS comes from a special property of  $\mathcal{H}_\phi$ , called the *producing property*:

$$\forall f \in \mathcal{H}_\phi, \quad \langle f, \phi(x) \rangle_{\mathcal{H}_\phi} = f(x).$$

We abuse the notation here by interpreting  $\phi(x)$  as a function in  $\mathcal{H}_\phi$  instead of a real value. A special instance of the reproducing property is that

$$\langle \phi(x), \phi(y) \rangle = k_\phi(x, y).$$

Precisely, we are using the following representation of  $\phi(x)$ :

$$\phi(x) : \mathcal{X} \rightarrow \mathcal{H}_\phi, \quad \phi(x)(\cdot) = k_\phi(x, \cdot) \in \mathcal{H}_\phi.$$

Here  $k_\phi(x, \cdot) \in \mathcal{H}_\phi$  is guaranteed by the definition of  $\mathcal{H}_\phi$ .

In what follows, we use  $\mathcal{M}_+^1(\mathcal{X})$  to denote the space of probability measures over  $\mathcal{X}$ . We can embed any probability measure to the RKHS.

**Definition C.1** (Kernel mean embedding of probability measure). *The kernel mean embedding of a probability measure  $\mathbb{P} \in \mathcal{H}_\phi$  is defined via the following mapping:*

$$\mu_\phi : \mathcal{M}_+^1(\mathcal{X}) \rightarrow \mathcal{H}_\phi, \quad \mu_\phi(\mathbb{P})(\cdot) = \int_{\mathcal{X}} \phi(x)(\cdot) \mathbb{P}(dx) = \mathbb{E}_{X \sim \mathbb{P}} [\phi(X)] (\cdot).$$

We denote the kernel mean embedding of  $\mathbb{P}$  by  $\mu_\phi(\mathbb{P})$ .

**Proposition C.2** (c.f Eq. (3.29) of Muandet et al. [2017]). *For all  $\mathbb{P}, \mathbb{Q} \in \mathcal{M}_+^1(\mathcal{X})$ ,*

$$\text{MMD}_{k_\phi}(\mathbb{P}, \mathbb{Q}) = \|\mathbb{E}_{X \sim \mathbb{P}} [\phi(X)] - \mathbb{E}_{Y \sim \mathbb{Q}} [\phi(Y)]\| = \|\mu_\phi(\mathbb{P}) - \mu_\phi(\mathbb{Q})\|. \quad (9)$$

*Proof of Proposition C.2.* The second equality of Eq. (9) follows directly from the mapping defined in Definition C.1. We then focus on proving the first equality.

$$\begin{aligned} & \text{MMD}_{k_\phi}(\mathbb{M}_1, \mathbb{M}_2) \\ &= \sup_{f \in \mathcal{H}_\phi: \|f\| \leq 1} \mathbb{E}_{X \sim \mathbb{P}} [f(X)] - \mathbb{E}_{Y \sim \mathbb{Q}} [f(Y)] \\ &= \sup_{f \in \mathcal{H}_\phi: \|f\| \leq 1} \mathbb{E}_{X \sim \mathbb{P}} [\langle f, \phi(X) \rangle] - \mathbb{E}_{Y \sim \mathbb{Q}} [\langle f, \phi(Y) \rangle] \quad (\text{reproducing property}) \\ &= \sup_{f \in \mathcal{H}_\phi: \|f\| \leq 1} \langle f, \mathbb{E}_{X \sim \mathbb{P}} [\phi(X)] \rangle - \langle f, \mathbb{E}_{Y \sim \mathbb{Q}} [\phi(Y)] \rangle \quad (\text{linearity of inner product}) \\ &= \sup_{f \in \mathcal{H}_\phi: \|f\| \leq 1} \langle f, \mathbb{E}_{X \sim \mathbb{P}} [\phi(X)] - \mathbb{E}_{Y \sim \mathbb{Q}} [\phi(Y)] \rangle \quad (\text{linearity of inner product}) \end{aligned}$$

Then the proof is completed by the observation that the inner product is maximized when

$$f = \{\mathbb{E}_{X \sim \mathbb{P}} [\phi(X)] - \mathbb{E}_{Y \sim \mathbb{Q}} [\phi(Y)]\} / \|\mathbb{E}_{X \sim \mathbb{P}} [\phi(X)] - \mathbb{E}_{Y \sim \mathbb{Q}} [\phi(Y)]\|.$$

□

Proposition C.2 essentially says that MMD between two distributions is indeed the distance of mean embeddings of features. It also tells us that  $\text{MMD}_{k_\phi}(\mathbb{P}, \mathbb{Q}) = 0$  if and only if  $\mu_\phi(\mathbb{P}) = \mu_\phi(\mathbb{Q})$ . To be able to separate any two distributions via MMD, the kernel mean embedding must be an injective map, in which case the feature map induces a *characteristic kernel*.

**Definition C.3** (Characteristic kernel).  *$k$  is said to be characteristic on  $\mathcal{M}_+^1(\mathcal{X})$  if the kernel mean embedding,*

$$\mu_k : \mathcal{M}_+^1(\mathcal{X}) \rightarrow \mathcal{H}_k, \quad \mu_k(\mathbb{P})(\cdot) = \int_{\mathcal{X}} \phi(x)(\cdot) \mathbb{P}(dx),$$

*is injective. In other words,  $k$  is a characteristic kernel if and only if*

$$\mu_k(\mathbb{P}, \mathbb{Q}) = 0 \iff \mu_k(\mathbb{P}) = \mu_k(\mathbb{Q}) \iff \mathbb{P} = \mathbb{Q}. \quad (10)$$

Here we change the subscript of  $\mu, \mathcal{H}$  from the feature map to the kernel, because from now on we do not necessarily work on explicit choice of feature maps; many characteristic kernels do not have tractable feature maps (and mostly infinite dimensional). Eq. (10) states that MMD is a metric on  $\mathcal{M}_+^1(\mathcal{X})$  if a characteristic kernel is used; otherwise, distinct distributions with the same kernel

mean embedding cannot be separated by MMD. Examples of characteristic kernels on  $\mathcal{M}_+^1(\mathbb{R}^d)$  are Gaussian kernels, Laplacian kernels, and the family of Matérn kernels. We refer readers to Sriperumbudur et al. [2011] for a comprehensive survey of the characteristic kernels.

Provided with a specified kernel, and samples from  $\mathbb{P}, \mathbb{Q}$ , one can obtain an unbiased estimate of the squared population MMD; see Gretton et al. [2012, Eq. (3)] for the detailed expression of the estimator. An interesting property of the estimates of MMD is that the convergence is dimension independent [Gretton et al., 2012, Theorem 7], although the penalization of the dimension in the context of kernel two sample test lies in the reduction of power [Ramdas et al., 2015].

## C.2 Compare distributions via a fixed feature map $h$

In some cases, one can compare  $\mathbb{P}, \mathbb{Q}$  by assessing on a fixed test function  $h$ . For example, the L2 norm between the expectation of  $h$ , i.e.,

$$\|\mathbb{E}_{X \sim \mathbb{P}}[h(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[h(Y)]\|_2 \quad (11)$$

can be a crude measure on how different  $\mathbb{P}, \mathbb{Q}$ . According to Proposition C.2,

$$\|\mathbb{E}_{X \sim \mathbb{P}}[h(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[h(Y)]\|_2 = \text{MMD}_{k_h}(\mathbb{P}, \mathbb{Q}).$$

The obvious limitation with a fixed choice of  $h$  is that it's typically not characteristic for all distributions, i.e.,

$$\|\mathbb{E}_{X \sim \mathbb{P}}[h(X)] - \mathbb{E}_{Y \sim \mathbb{Q}}[h(Y)]\|_2 = 0 \not\Rightarrow \mathbb{P} = \mathbb{Q}.$$

However, in practice, if the choice of  $h$  is sufficient to identify discriminate the collection of distributions that we care about, it is convenient to just examine Eq. (11).

In our specific application, interaction detection, we find the embedding vector that are constructed from the final hidden layer of a classifier works reliably. We assume that this classifier is trained optimally such that,  $\mathbb{P}(\delta_i | X) = \sigma(w_i^\top h(X))$ , where  $\sigma := \frac{\exp(x_i)}{\sum_j \exp(x_j)}$  is the softmax function, and that there are sufficiently diverse labels to ensure that this representation is identified up to a linear transformation; see Roeder et al. [2021] for details. Given a trained classifier,  $\bar{h}_i := \mathbb{E}[h(X) | \delta_i] - \mathbb{E}[h(X) | \delta_0]$  is then the average of the embeddings associated with a particular knockout centered around the control wells. We find in our empirical experiments that the metric  $\|\bar{h}_{ij} - \bar{h}_i - \bar{h}_j\|_2$  works well to identify highly interacting gene pairs.

We have not obtained theoretical arguments on justifying the use of optimal discriminator as the choice of test functions; we left this for future development.

## D Experiment details

For all separability tests, we trained the NRE model on all perturbation classes and then obtained the log-density ratios for each pair as introduced in Appendix B.1. For the NRE training, we evaluated multiple model architectures and optimizer step sizes, selecting the hyperparameter combination that yielded the best training accuracy. The selected model architectures and optimizer step sizes for each example are reported in the corresponding sections. We checkpointed the best model at the optimal training accuracy for further inference to obtain density ratio estimates and KL estimates. To optimize our models, we used ADAM [Kingma and Ba, 2017] with default hyperparameter settings.

All our experiments were run on NVIDIA H100 GPUs.

### D.1 Synthetic examples

For all the synthetic examples, we generated 20,000 i.i.d. samples for each perturbation class, including both single and double perturbations. The detailed data generation process for each example is provided below.

#### D.1.1 Synthetic tabular for separability test

The separability scores in this example are computed using two different KL-divergence estimators: the simple KNN-based estimator and the NRE-based estimation procedure described in Appendix B.

We used a 3-layer MLP with ReLU activation (hidden dimensions of 128 and 64) as the encoder for the NRE density ratio estimator. The NRE model was trained using the ADAM optimizer with a step size of 0.005 for 500 epochs and a batch size of 1024.

**Latent distribution** The latent variable  $Z$  consists of 3 independent one-dimensional variables  $P_1, P_2, P_3$ , i.e.,

$$P_Z(z_1, z_2, z_3) = P_1(z_1) \cdot P_2(z_2) \cdot P_3(z_3),$$

for which we consider 4 single perturbations in total, labelled as  $A, B, C, D$  respectively, resulting 6 pairwise perturbations. Perturbations are applied by changing the distribution of one or more of latent variables. In our setting, perturbations are classified as separable or inseparable. Double perturbations of separable ones will intervene in two distinct latent variables, while double perturbations of inseparable ones will intervene in the same latent variable. For each node, the control (unperturbed) distribution is  $\mathcal{N}(0, 1)$ , while the perturbed one—whether it's single perturbed or doubly perturbed—becomes  $\mathcal{N}(3, 1)$ .

The association between the perturbation class and corresponding intervened latent variable(s) is described as follows:

$$\begin{aligned} P_1 &\sim \begin{cases} \text{Normal}(0, 1) & \text{if unperturbed} \\ \text{Normal}(3, 1) & \text{if perturbed by (A) or/and (B)} \end{cases} \\ P_2 &\sim \begin{cases} \text{Normal}(0, 1) & \text{if unperturbed} \\ \text{Normal}(3, 1) & \text{if perturbed by (B)} \end{cases} \\ P_3 &\sim \begin{cases} \text{Normal}(0, 1) & \text{if unperturbed} \\ \text{Normal}(3, 1) & \text{if perturbed by (C) or/and (D)} \end{cases} \end{aligned} \quad (12)$$

As per Eq. (12), the only two inseparable pairs are A-B (both intervening  $P_1$ ) and C-D (both intervening  $P_3$ ).

**Sampling process** To generate observations for each perturbation class, we first generate latent samples from  $P_Z$  based on the Eq. (12), and then transform the latent samples via a diffeomorphism  $g(\cdot)$ . In this example, we chose  $g(\cdot)$  as a randomly initialized 7-layer Multi layer perceptron (MLP) with LeakyReLU activations.

### D.1.2 Synthetic image example for separability test

The synthetic images consist of three objects (three small colored balls) in different locations and with various backgrounds. The positions of the objects are encoded in a 3-dimensional latent variable  $Z$ , which follows the identical DAG structure described in Eq. (12). Each coordinate of  $Z$  corresponds to the location distribution of an object, determining the distribution of both the  $x$  and  $y$  coordinates of the object. Thus, a perturbation affecting the location distribution of one object will intervene in the distribution of both coordinates of that object. Background distortion is controlled by random noise. Scenes are generated using a rendering engine from PyGame, denoted as  $g(\cdot)$ . Example images are provided in Fig. 7.

For the NRE model, we used a 5-layer convolutional network (with channels 32, 54, 128, 256, and 512) featuring batch normalization and leaky ReLU activations as the encoder. The model was trained using the ADAM optimizer with a step size of 0.0002 for 200 epochs and a batch size of 1024.

### D.1.3 Synthetic tabular example for disjointedness test

The latent distribution is set to be a 6-component mixture distribution,

$$P_Z(z) = \frac{1}{8}P_0(z) + \frac{1}{8}P_1(z) + \frac{1}{8}P_2(z) + \frac{1}{8}P_3(z) + \frac{1}{4}P_4(z) + \frac{1}{4}P_5(z).$$

We define 7 single perturbations, labeled as A, B, C, D, E, F, G, respectively. The control distributions (unperturbed distribution) of all mixture component are  $\mathcal{N}(0, 1)$ . Perturbations are applied by changing the distribution of one or more of the mixture components. E.g., if perturbation  $D$  is applied,  $P_4 \sim \text{Normal}(0, 5)$ . In our setting, perturbations are classified as independent ones and interacting ones, where double perturbation of independent ones will intervene two distinct mixture components, while double perturbation of interacting ones will intervene the same mixture component. The

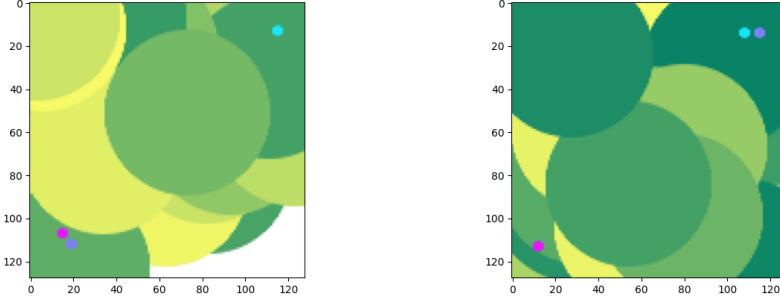


Figure 7: Example images of the interventional image data. 3 small balls with blue, purple, and red colors respectively are the targeted objects, whose locations are intervened by perturbations. Large balls with green and yellow colors form the background.

association of the perturbations and the corresponding component gets intervened is provided as follows:

The independent group consists of the mixture components  $P_0, P_1, P_2, P_3$ :

$$\begin{aligned} P_0 &\sim \text{Normal}(0, 1) && \text{(not intervened)} \\ P_1 &\sim \text{Normal}(5, 1) && \text{(A)} \\ P_2 &\sim \text{Normal}(10, 1) && \text{(B)} \\ P_3 &\sim \text{Normal}(-5, 5) && \text{(C)} \end{aligned}$$

The interacting group consists of the variables  $P_4, P_5$ :

$$\begin{aligned} P_4 &\sim \begin{cases} \text{Normal}(0, 5) & \text{(D)} \\ \text{Normal}(-10, 5) & \text{(E)} \\ \text{Normal}(10, 5) & \text{(DE)} \end{cases} \\ P_5 &\sim \begin{cases} \text{Cauchy}(15, 1) & \text{(F)} \\ \text{Cauchy}(-15, 1) & \text{(G)} \\ \text{Cauchy}(20, 1) & \text{(FG)} \end{cases} \end{aligned}$$

**Sampling process** To generate latent samples, we draw from the specified distributions and the constructed mixture model. Let  $n$  be the number of samples and  $d$  the dimension. The samples are generated as follows:

```

if no intervention:  $\mathbf{X} \sim P_0$ 
if intervention is specified:  $(\mathbf{P}_s, \mathbf{w}_s) = \text{build\_mixture(intervene)}$ 
draw categorical samples:  $\mathbf{z} \sim \text{Categorical}(\mathbf{w}_s)$ 
generate samples from each component:  $\mathbf{X}_i \sim \mathbf{P}_s[i]$ 

```

We generated latents for all 7 single perturbations, and all double perturbations, and then obtained the observations, which are used to perform tests on, by mapping the latent samples through a deterministic function  $g(\cdot)$ . We chose  $g(\cdot)$  as a randomly initialized 10-layer MLP with LeakyReLU activations.

## D.2 Real data examples

### D.2.1 Gene-gene interactions

For the separability test, we used a 3-layer MLP with ReLU activation (hidden dimensions of 2048 and 256) as the encoder for the NRE density ratio estimator. We trained the NRE model using the ADAM optimizer with a step size of 0.0001 for 2500 epochs and a batch size of 16384.

Table 1 describes the corresponding pathways for each selected gene.

Table 1: List of gene indices and their associated pathways

<b>Gene Index</b>	<b>Pathway</b>
gene0	Amino acid sensing (mTOR pathway)
gene1	
gene2	
gene3	
gene4	
gene5	Apoptosis
gene6	
gene7	
gene8	
gene9	
gene10	
gene11	
gene12	
gene13	
gene14	Autophagy
gene15	
gene16	
gene17	
gene18	
gene19	
gene20	
gene21	ERAD (protein folding)
gene22	
gene23	Integrated Stress Response
gene24	
gene25	Microtubule
gene26	
gene27	PI3K-Akt signaling
gene28	
gene29	Proteasome
gene30	
gene31	Protein translation
gene32	Protein translation (mTOR pathway)
gene33	
gene34	Ribosome
gene35	
gene36	Transcriptional regulation
gene37	
gene38	UPR (protein folding)
gene39	
gene40	
gene41	
gene42	
gene43	
gene44	mTOR signaling
gene45	
gene46	
gene47	
gene48	
gene49	p53 signaling

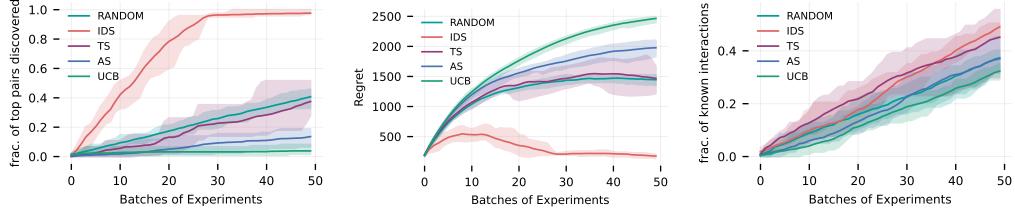


Figure 8: Empirical results on the active pairwise experiments selection for the gene-gene interaction detection example using MMD-based (using RBF kernel) test statistics. The solid lines represent the mean performance, whereas the shaded region represent all the runs (min-max). IDS (ours) outperforms all the baselines significantly in terms of top scoring pairs discovered as well as the regret. In terms of known interactions, IDS still outperforms the baselines by a small margin.

### D.2.2 Guide-guide interactions

The single-cell painting images are derived from multi-cell images, with each single-cell nucleus centered within a  $32 \times 32$  pixel box. The encoder of the NRE model maps single-cell images of shape  $(6, 32, 32)$  into a 128-dimensional feature vector. It consists of three convolutional blocks, each comprising a Conv2D layer with a  $3 \times 3$  kernel, BatchNorm2D, ReLU activation, and MaxPool2D, progressively increasing the number of channels from 6 to 32, 64, and 128 while halving the spatial dimensions at each max-pooling step. After the convolutional layers, the output tensor of shape  $(128, 4, 4)$  is flattened to  $(2048)$  and passed through two fully connected blocks, each with a Linear layer, ReLU activation, and Dropout (with a dropout rate of 0.3), transforming the feature size from 2048 to 256 and finally to 128. We train the NRE model with the ADAM optimizer, using a step size of 0.00005 for 5000 epochs and a batch size of 2048.

### D.3 Active learning

To obtain samples from the posterior distribution over the low-rank reward matrix (with rank  $m$ ), we use stochastic variational inference [Wingate and Weber, 2013, Ranganath et al., 2014], and specifically the implementation from `numpyro` [Bradbury et al., 2018, Bingham et al., 2019, Phan et al., 2019]. We train the variational posterior for 5000 epochs with a learning rate of 0.01 using the Adam optimizer. We then generate  $k$  samples from the posterior. For each algorithm, we run a sweep over all a set of hyperparameters. We then pick the best hyperparameters and run the experiment over 10 different seeds to get the final results. For IDS, we tune  $m \in \{3, 5, 7, 10, 12\}$ ,  $\lambda \in \{2, 3, 4, 5\}$  and  $k \in \{500, 750, 1000, 1500\}$ . For TS and US, we tune  $m \in \{3, 5, 7, 10, 12\}$  and  $k \in \{500, 750, 1000, 1500\}$ . For UCB we tune  $m \in \{3, 5, 7, 10, 12\}$ ,  $\beta \in \{0.01, 0.1, 0.2, 0.5, 1, 2, 5\}$  and  $k \in \{500, 750, 1000, 1500\}$ , where  $\beta$  controls the exploration.