REPRESENTATION LEARNING FOR DISTRIBUTIONAL PERTURBATION EXTRAPOLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

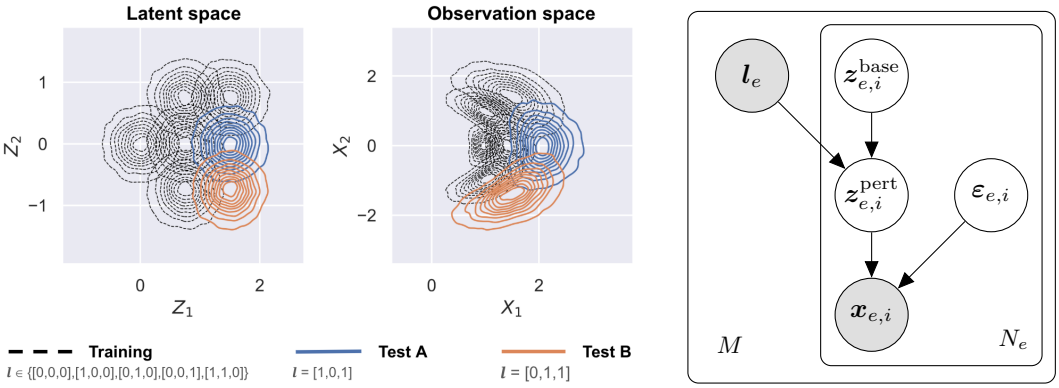
We consider the problem of modelling the effects of perturbations such as gene knockdowns or drug combinations on low-level measurements like RNA sequencing data. Specifically, given data collected under some perturbations, we aim to predict the distribution of measurements for new perturbations. To address this challenging extrapolation task, we posit that perturbations act additively in a suitable, unknown embedding space. More precisely, we formulate the generative process underlying the observed data as a latent variable model, in which perturbations amount to mean shifts in latent space. We prove that the representation and perturbation effects are identifiable up to affine transformation and use this to characterize the class of unseen perturbations for which we obtain extrapolation guarantees. To estimate the model from data, we propose the perturbation distribution autoencoder (PDAE) which is trained by maximising the distributional similarity between true and predicted perturbation distributions The trained model can then be used to predict previously unseen perturbation distributions. Preliminary empirical evidence suggests that PDAE compares favourably to CPA (Lotfollahi et al., 2023) and other baselines at predicting the effects of unseen perturbations.

1 INTRODUCTION

Due to technological progress, large-scale perturbation data is becoming more abundant across several scientific fields. This is particularly the case for single-cell biology, where advancements in gene editing, sequencing, and mass spectrometry have led to the collection of vast transcriptomic or proteomic databases for various drug and gene perturbations (Dixit et al., 2016; Jinek et al., 2012; Norman et al., 2019; Wang et al., 2009; Weinstein et al., 2013). However, the exponential number of possible combinations of perturbations renders exhaustive experimentation prohibitive. Observations are thus typically only available for a subset of perturbations of interest, e.g., some single and double gene knockdowns or certain dosages of drugs. This necessitates models capable of extrapolating to unseen combinations of perturbations, e.g., new multi-gene knockdowns or dosage combinations.

Prior Work. Several recent works leverage machine learning for biological perturbation modelling, e.g., to generalize to new cell types (Bunne et al., 2023; Lotfollahi et al., 2019), unseen combinations of perturbations (Lotfollahi et al., 2023), or entirely new perturbations by leveraging the molecular structure of the involved compounds (Hetzel et al., 2022; Qi et al., 2024; Yu & Welch, 2022) or prior knowledge about gene-gene interactions (Kamimoto et al., 2023; Roohani et al., 2024). A common theme is the use of representation learning techniques such as autoencoders (Hinton & Salakhutdinov, 2006; Kingma & Welling, 2014; Rumelhart et al., 1986) to embed observations in a latent space, in which the effects of perturbations are assumed to take on a simpler (e.g., additive) form. However, existing studies are purely empirical and lack theoretical underpinning. Despite promising results, the capabilities and fundamental limitations of existing methods thus remain poorly understood.

Overview and Contributions. In this work, we present a principled, theoretically-grounded approach for perturbation extrapolation. Given the unpaired nature of the available data (each cell is only measured under one experimental condition), we consider the task of predicting population-level effects of perturbations, which we formalize as a distributional regression problem (§ 2). We then postulate a generative model (§ 3) which, similar to prior works, assumes that perturbations act as mean shifts in a suitable latent space, see Fig. 1 for an overview. We analyse this model



(a) Simulated Multi-Domain Perturbation Data (b) Graphical Model Representation

Figure 1: **Task Description and Assumed Data Generating Process.** (a) During training, we are given $M = 5$ training data sets in observation space (right, grey), each of which is generated under a known combination of $K = 3$ elementary perturbations. The corresponding perturbation labels $\boldsymbol{l}_e$ are shown below the plots. During testing, we are given a new perturbation label and the task is to predict the corresponding distribution in observation space (right, blue and orange). We tackle this task by assuming that the effect of perturbations is linear additive in a suitable latent space (left). Both plots show kernel density estimates of the distributions. (b) Each dataset comprises a perturbation label $\boldsymbol{l}_e$ and N_e observations $\boldsymbol{x}_{e,i}$. Perturbations act as mean-shifts on a latent basal state, $\boldsymbol{z}^{\text{pert}}_{e,i} = \boldsymbol{z}^{\text{base}}_{e,i} + \boldsymbol{W}\boldsymbol{l}_e$. A stochastic nonlinear decoder with noise $\boldsymbol{\varepsilon}_{e,i}$ then yields the observed $\boldsymbol{x}_{e,i} = \boldsymbol{f}(\boldsymbol{z}^{\text{pert}}_{e,i}, \boldsymbol{\varepsilon}_{e,i})$. Shaded and white nodes indicate observed and unobserved/latent variables, respectively.

class theoretically (§ 4), proving that the latent representation and the relative training perturbation effects are identifiable up to affine transformation (Thm. 4.1). This result implies extrapolation guarantees for unseen perturbations that can be expressed as linear combinations of training perturbations (Thm. 4.6). Based on these insights, we devise an autoencoder-based estimation method (§ 5) which uses the energy score (Gneiting & Raftery, 2007) to assess the distributional similarity between predicted and ground-truth perturbation data. In preliminary simulations (§ 6), our approach compares favourably to the compositional perturbation autoencoder (CPA; Lotfollahi et al., 2023) in terms of mean prediction and distributional fit.

Notation. We write scalars as a, column-vectors as $\boldsymbol{a}$, and matrices as $\boldsymbol{A}$. We use uppercase for random variables and lowercase for their realizations. Equality in distribution is denoted by $\overset{d}{=}$ and the pushforward of a distribution $\mathbb{P}$ by a measurable function f is denoted by $f_{\#}\mathbb{P}$. The Euclidean (L2) norm is denoted by $\|\cdot\|$. Further, we use the shorthands $[n] = \{1, ..., n\}$ and $[n]_0 = [n] \cup \{0\}$.

2 Problem Setting: Distributional Perturbation Extrapolation

Let $\boldsymbol{x} \in \mathbb{R}^{d_X}$ be an observation (e.g., omics data) that is obtained under one of several possible experimental conditions. We model these conditions as combinations of K elementary perturbations, each of which we assume can be encoded by a real number. Further, let $\boldsymbol{l} \in \mathbb{R}^K$ be a perturbation label that indicates if, or how much of, each perturbation was applied before collecting the corresponding $\boldsymbol{x}$.

Example 2.1 (Gene perturbations). For data arising from gene knockouts, a perturbationcan be represented by a binary $\boldsymbol{l} \in \{0, 1\}^K$, where K denotes the number of potential targets and $l_k = 1$ if and only if target k was subject to a knockout experiment. For example, $\boldsymbol{l} = (1, 0, 1, 0, 0)$ indicates a multi-gene knockout on targets one and three.

Example 2.2 (Drug perturbations). For data arising from applying varying amounts of K different drugs, perturbations can be represented by continuous, non-negative labels $\boldsymbol{l} \in \mathbb{R}^K_+$, where l_k indicates the (relative or absolute) amount of drug k that was administered.

We have access to $M+1$ experimental datasets $\mathcal{D}_0, \mathcal{D}_1, ..., \mathcal{D}_M$, each comprising a sample of N_e observations $\boldsymbol{x}$ and a perturbation label $\boldsymbol{l}$, i.e., for all experiments or environments $e \in [M]_0$,

$$\mathcal{D}_e = \left(\{\boldsymbol{x}_{e,i}\}_{i=1}^{N_e}, \boldsymbol{l}_e \right) . \tag{2.1}$$

Given data in the form of (2.1), the task we consider is to predict the effects of new perturbations $\boldsymbol{l}_{\text{test}} \notin \{\boldsymbol{l}_0, \boldsymbol{l}_1, ..., \boldsymbol{l}_M\}$ without observing any data from this condition ("zero-shot"). In particular, we are interested in the distribution over observations resulting from $\boldsymbol{l}_{\text{test}}$. That is, we aim to leverage the training domains (2.1) to learn a map

$$\boldsymbol{l} \mapsto \mathbb{P}_{\boldsymbol{X}|\boldsymbol{l}} \tag{2.2}$$

which extrapolates beyond the training support of $\boldsymbol{l}$, i.e., the predictions should remain reliable for new inputs $\boldsymbol{l}_{\text{test}}$. Since (2.2) targets the full conditional distribution—rather than, say, the conditional mean $\mathbb{E}[\boldsymbol{X}|\boldsymbol{l}]$—it constitutes a (multi-variate) distributional regression task (Koenker, 2005; Koenker & Bassett Jr, 1978), also referred to as probabilistic forecasting (Gneiting & Raftery, 2007) or conditional generative modelling (Mirza, 2014; Sohn et al., 2015; Winkler et al., 2019). We therefore refer to our problem setting as *distributional perturbation extrapolation*.

Formally, extrapolation means that the value of the function (2.2) at $\boldsymbol{l}_{\text{test}}$ is determined by its values on the training support $\{\boldsymbol{l}_0, \boldsymbol{l}_1, ..., \boldsymbol{l}_M\}$. Intuitively, for this to be feasible, $\boldsymbol{l}_{\text{test}}$ must be somehow related to the training perturbations $\boldsymbol{l}_e$. For example, given data resulting from individual perturbations, predict the effects of combinations thereof. This type of extrapolation to new combinations of inputs is also called compositional generalization (Goyal & Bengio, 2022; Lake et al., 2017). It is known to be challenging (Montero et al., 2022; 2021; Schott et al., 2022) and requires assumptions that sufficiently constrain the model class (Brady et al., 2023; 2025; Dong & Ma, 2022; Lachapelle et al., 2023; Lippl & Stachenfeld, 2024; Wiedemer et al., 2024a;b).

3 Model: Perturbations as Mean Shifts in Latent Space

We now specify a generative process for the observed data in (2.1). In so doing, we aim to strike a balance between imposing sufficient structure on (2.2) to facilitate extrapolation, while remaining flexible enough to model the complicated, nonlinear effects which perturbations may have on the distribution of observations. Similar to Lotfollahi et al. (2023), we therefore model the effect of perturbations in a latent space with perturbation-relevant latent variables $\boldsymbol{z} \in \mathbb{R}^{d_Z}$, which are related to the observations $\boldsymbol{x}$ via a nonlinear (stochastic) mixing function or generator $\boldsymbol{f}$. The full generative process amounts to a hierarchical latent variable model, which additionally contains noise variables $\boldsymbol{\varepsilon}$ that capture other variation underlying the observations $\boldsymbol{x}$, and which is represented as a graphical model in Fig. 1b. Specifically, we posit for all $e \in [M]_0 \cup \{\text{test}\}$ and all $i \in [N_e]$:

$$\boldsymbol{z}_{e,i}^{\text{base}} \sim \mathbb{P}_{\boldsymbol{Z}}, \tag{3.1}$$

$$\boldsymbol{z}_{e,i}^{\text{pert}} := \boldsymbol{z}_{e,i}^{\text{base}} + \boldsymbol{W}\boldsymbol{l}_e, \tag{3.2}$$

$$\boldsymbol{\varepsilon}_{e,i} \sim \mathbb{Q}_{\boldsymbol{\varepsilon}}, \tag{3.3}$$

$$\boldsymbol{x}_{e,i} := \boldsymbol{f}\left(\boldsymbol{z}_{e,i}^{\text{pert}}, \boldsymbol{\varepsilon}_{e,i}\right), \tag{3.4}$$

where $(\boldsymbol{z}_{e,i}^{\text{base}})_{e\in[M]_0, i\in[N_e]}$ are independent and identically distributed (i.i.d.) according to $\mathbb{P}_{\boldsymbol{Z}}$, and $(\boldsymbol{\varepsilon}_{e,i}^{\text{base}})_{e\in[M]_0, i\in[N_e]}$ are i.i.d. according to $\mathbb{Q}_{\boldsymbol{\varepsilon}}$ and jointly independent of $(\boldsymbol{z}_{e,i}^{\text{base}})_{e\in[M]_0, i\in[N_e]}$.

The basal state $\boldsymbol{z}^{\text{base}}$ in (3.1) describes the unperturbed state of latent variables, which can, in principle, be affected by perturbations, and is distributed according to a base distribution $\mathbb{P}_{\boldsymbol{Z}}$. The perturbation matrix $\boldsymbol{W} \in \mathbb{R}^{d_Z \times K}$ in (3.2) captures the effect of the K elementary perturbations encoded in $\boldsymbol{l}$ on the latents and turns $\boldsymbol{z}^{\text{base}}$ into perturbed latents $\boldsymbol{z}^{\text{pert}}$. Since the same perturbation $\boldsymbol{l}_e$ is applied for all $i \in [N_e]$, all intra-dataset variability in $\boldsymbol{z}^{\text{pert}}$ is due to $\mathbb{P}_{\boldsymbol{Z}}$. The noise variables $\boldsymbol{\varepsilon} \in \mathbb{R}^{d_\epsilon}$ in (3.3) capture all other variation in the observed data that is unaffected by perturbations. It is distributed according to a fixed, uninformative distribution $\mathbb{Q}_{\boldsymbol{\varepsilon}}$ such as a standard isotropic Gaussian. The noise serves as an additional input to the (stochastic) mixing function or generator $\boldsymbol{f} : \mathbb{R}^{d_Z} \times \mathbb{R}^{d_\epsilon} \to \mathbb{R}^{d_X}$ in (3.4), which produces observations for the perturbed latent. This generative model allows us to model any conditional distribution $\mathbb{P}_{\boldsymbol{X}|\boldsymbol{z}}$ and is more flexible than, e.g., a Gaussian decoder with mean and covariance parametrised by $\boldsymbol{f}$ as used by Lotfollahi et al. (2023).

For a given e and $\boldsymbol{l}_e$, the generative process in (3.1)–(3.4) induces a distribution $\mathbb{P}_e$ over observations $\boldsymbol{x}$, which we also denote by $\mathbb{P}_{\boldsymbol{X}|\boldsymbol{l}_e}$, defined as the push-forward of $\mathbb{P}_{\boldsymbol{Z}}$ and $\mathbb{Q}_{\boldsymbol{\varepsilon}}$ through (3.2) and (3.4), such that:

$$\forall e \in [M]_0 : \qquad (\boldsymbol{x}_{e,i})_{i\in[N_e]} \overset{\text{i.i.d.}}{\sim} \mathbb{P}_e\,. \tag{3.5}$$

4 THEORY: IDENTIFIABILITY AND EXTRAPOLATION GUARANTEES

In this section, we present our theoretical analysis for the model class introduced in § 3.

Identifiability. We first study identifiability, that is, the question under what assumptions and up to what ambiguities certain parts of the postulated generative process can be provably recovered assuming access to the full distributions. As established by the following results, our model class is identifiable up to affine transformation, provided that the training perturbations are sufficiently diverse, the dimension of the latent space is known and some additional technical assumptions hold.

Theorem 4.1 (Affine identifiability for Gaussian latents)**.** *For $M \in \mathbb{Z}_{\geq 0}$, let $\boldsymbol{l}_0, ..., \boldsymbol{l}_M \in \mathbb{R}^K$ be $M+1$ perturbation labels. Let $\boldsymbol{f}, \widetilde{\boldsymbol{f}} : \mathbb{R}^{d_Z} \to \mathbb{R}^{d_X}$, $\boldsymbol{W}, \widetilde{\boldsymbol{W}} \in \mathbb{R}^{d_Z \times K}$, and $\mathbb{P}, \widetilde{\mathbb{P}}$ be distributions on $\mathbb{R}^{d_Z}$ such that the models $(\boldsymbol{f}, \boldsymbol{W}, \mathbb{P})$ and $(\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{W}}, \widetilde{\mathbb{P}})$ induce the same observed distributions, i.e.,*

$$\forall e \in [M]_0 : \qquad \boldsymbol{f}\left(\boldsymbol{Z} + \boldsymbol{W}\boldsymbol{l}_e\right) \stackrel{d}{=} \widetilde{\boldsymbol{f}}\left(\widetilde{\boldsymbol{Z}} + \widetilde{\boldsymbol{W}}\boldsymbol{l}_e\right), \quad \text{where} \quad \boldsymbol{Z} \sim \mathbb{P} \quad \text{and} \quad \widetilde{\boldsymbol{Z}} \sim \widetilde{\mathbb{P}}. \tag{4.1}$$

Assume further that:

(i) ***[invertibility]*** *$\boldsymbol{f}$ and $\widetilde{\boldsymbol{f}}$ are C^2-diffeomorphisms onto their respective images;*

(ii) ***[Gaussianity]*** *$\boldsymbol{Z}$ and $\widetilde{\boldsymbol{Z}}$ are non-degenerate multi-variate Gaussians, i.e., $\mathbb{P} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\widetilde{\mathbb{P}} = \mathcal{N}\left(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}\right)$ for some $\boldsymbol{\mu}, \widetilde{\boldsymbol{\mu}} \in \mathbb{R}^{d_Z}$ and positive-definite $\boldsymbol{\Sigma}, \widetilde{\boldsymbol{\Sigma}} \in \mathbb{R}^{d_Z \times d_Z}$;*

(iii) ***[sufficient diversity]*** *the matrix $\widetilde{\boldsymbol{W}}\boldsymbol{L} \in \mathbb{R}^{d_Z \times M}$, where $\boldsymbol{L} \in \mathbb{R}^{K \times M}$ is the matrix with columns $(\boldsymbol{l}_e - \boldsymbol{l}_0)$ for $e \in [M]$, has full row rank, i.e., $\operatorname{rank}(\widetilde{\boldsymbol{W}}\boldsymbol{L}) = d_Z$.*

Then the latent representation and the effects of the observed perturbation combinations relative to $\boldsymbol{l}_0$ (as captured by $\boldsymbol{W}\boldsymbol{L}$) are identifiable up to affine transformation in the following sense:

$$\forall \boldsymbol{z} : \qquad \widetilde{\boldsymbol{f}}^{-1} \circ \boldsymbol{f}(\boldsymbol{z}) = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{b}, \tag{4.2}$$

$$\widetilde{\boldsymbol{W}}\boldsymbol{L} = \boldsymbol{A}\boldsymbol{W}\boldsymbol{L}, \tag{4.3}$$

where $\boldsymbol{A} := \widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}} \boldsymbol{\Sigma}^{-\frac{1}{2}}$ and $\boldsymbol{b} := \widetilde{\boldsymbol{\mu}} - \boldsymbol{A}\boldsymbol{\mu} + (\widetilde{\boldsymbol{W}} - \boldsymbol{A}\boldsymbol{W})\boldsymbol{l}_0$.

Corollary 4.2 (Affine recovery of the perturbation matrix)**.** *If, in addition to the assumptions of Thm. 4.1, $\boldsymbol{L} \in \mathbb{R}^{K \times M}$ has full row rank (i.e., $\operatorname{rank}(\boldsymbol{L}) = K \leq M$), then the perturbation matrix $\boldsymbol{W}$ is identifiable up to affine transformation in the sense that*

$$\widetilde{\boldsymbol{W}} = \boldsymbol{A}\boldsymbol{W}, \tag{4.4}$$

for $\boldsymbol{A} := \widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}} \boldsymbol{\Sigma}^{-\frac{1}{2}}$. In this case, the expression for $\boldsymbol{b}$ in (4.2) simplifies to $\boldsymbol{b} = \widetilde{\boldsymbol{\mu}} - \boldsymbol{A}\boldsymbol{\mu}$.

The proofs of Thm. 4.1 and Cor. 4.2 are provided in Appx. B.1 and B.2.

Discussion. Thm. 4.1 can be interpreted as follows. Fix a set of perturbation labels $(\boldsymbol{l}_0, ..., \boldsymbol{l}_M)$ and a data generating process parametrised by $(\boldsymbol{f}, \boldsymbol{W}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then, for all $\widetilde{\boldsymbol{\mu}}$ and $\widetilde{\boldsymbol{\Sigma}}$, and for all $\widetilde{\boldsymbol{W}}$ such that (4.3) holds for $\boldsymbol{A} = \widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}} \boldsymbol{\Sigma}^{-\frac{1}{2}}$, there exists a unique $\widetilde{\boldsymbol{f}}$, characterized by (4.2), such that $(\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})$ gives rise to the same observed distributions of $\boldsymbol{X}_e$. At the same time, any $(\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{W}}, \widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})$ for which this holds for all $e \in [M]_0$ is of this form. In other words, the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ of the basal state are completely unidentifiable, but the mixing function $\boldsymbol{f}$ and the relative shift matrix $\boldsymbol{W}\boldsymbol{L}$ are identifiable up to affine transformation—provided that the observed training perturbation conditions are sufficiently diverse, as formalised by assumption *(iii)*.

Remark 4.3 (Sufficient diversity). The matrix product $\boldsymbol{W}\boldsymbol{L} \in \mathbb{R}^{d_Z \times M}$ captures the relative effects of the observed perturbations. Specifically, $(\boldsymbol{W}\boldsymbol{L})_{je}$ corresponds to the shift in the j^{th} latent Z_j resulting from $\boldsymbol{l}_e$, relative to a reference condition $\boldsymbol{l}_0$. Moreover, assumption *(iii)* of Thm. 4.1 implies

$$\min\left\{\operatorname{rank}\left(\widetilde{\boldsymbol{W}}\right), \operatorname{rank}(\boldsymbol{L})\right\} \geq \operatorname{rank}(\widetilde{\boldsymbol{W}}\boldsymbol{L}) = d_Z. \tag{4.5}$$

Hence, sufficient diversity requires at least d_Z elementary perturbations whose associated shift vectors $\boldsymbol{w}_k \in \mathbb{R}^{d_Z}$ are linearly independent, and we must observe at least d_Z perturbation conditions $\boldsymbol{l}_e$ other than $\boldsymbol{l}_0$ such that the relative perturbation vectors $(\boldsymbol{l}_e - \boldsymbol{l}_0) \in \mathbb{R}^K$ are linearly independent.

Remark 4.4 (Choice of reference). Since the environments are unordered, the choice of reference environment is arbitrary. Here, we choose $e = 0$ as reference without loss of generality. Intuitively, if a perturbation is always present (e.g., $l_{e,1} = 1$ for all e), then its effects cannot be discerned from the basal state. Therefore, only the effects of the relative perturbations $(\boldsymbol{l}_e - \boldsymbol{l}_0)$ can be recovered. In practice, we often have access to an unperturbed, purely observational control condition with $\boldsymbol{l}_0 = \mathbf{0}$.

Remark 4.5 (Deterministic vs noisy mixing.). The mixing function $\boldsymbol{f}$ in Thm. 4.1 is deterministic, i.e., does not take a separate noise variable $\boldsymbol{\varepsilon}$ as input, cf. (3.4). In principle, noise can be appended to $\boldsymbol{z}^{\text{pert}}$ as additional dimensions that are not influenced by perturbations. However, this increases d_Z and thus makes it harder to satisfy sufficient diversity (see Remark 4.3). Alternatively, the setting of additive noise, $\boldsymbol{f}(\boldsymbol{Z}) + \boldsymbol{\varepsilon}$, can be reduced to the noiseless case (Khemakhem et al., 2020).

From Identifiability to Extrapolation. Since we aim to make distributional predictions for new perturbations $\boldsymbol{l}_{\text{test}}$, identifiability is only of intermediary interest. The following result highlights the usefulness of the affine identifiability established in Thm. 4.1 for extrapolation. In particular, it shows that this allows us to uniquely predict the observable effects of certain unseen perturbations—specifically, those which can be expressed as linear combinations of the observed perturbations.

Theorem 4.6 (Extrapolation to span of relative perturbations)**.** *Under the same setting and assumptions as in Thm. 4.1, let $\boldsymbol{l}_{\text{test}} \in \mathbb{R}^K$ be an unseen perturbation label such that*

$$(\boldsymbol{l}_{\text{test}} - \boldsymbol{l}_0) \in \operatorname{span}\left(\{\boldsymbol{l}_e - \boldsymbol{l}_0\}_{e\in[M]}\right). \tag{4.6}$$

Then the effect of $\boldsymbol{l}_{\text{test}}$ is uniquely identifiable in the sense that

$$\boldsymbol{X}_{\text{test}} = \boldsymbol{f}(\boldsymbol{Z} + \boldsymbol{W}\boldsymbol{l}_{\text{test}}) \overset{d}{=} \widetilde{\boldsymbol{f}}\left(\widetilde{\boldsymbol{Z}} + \widetilde{\boldsymbol{W}}\boldsymbol{l}_{\text{test}}\right) = \widetilde{\boldsymbol{X}}_{\text{test}}. \tag{4.7}$$

The proof of Thm. 4.6 is provided in Appx. B.3.

Remark 4.7 (Additive vs linear perturbations). For our affine identifiability result (Thm. 4.1), it is not necessary that the mean shifts $\boldsymbol{W}\boldsymbol{l}_e$ are linear in $\boldsymbol{l}_e$. If we replace $\boldsymbol{W}\boldsymbol{l}_e$ and $\widetilde{\boldsymbol{W}}\boldsymbol{l}_e$ in (4.1) with arbitrary shift vectors $\boldsymbol{c}_e, \widetilde{\boldsymbol{c}}_e \in \mathbb{R}^{d_Z}$, the same result can be shown to hold with $\boldsymbol{W}\boldsymbol{L}$ and $\widetilde{\boldsymbol{W}}\boldsymbol{L}$ replaced by $\boldsymbol{C}$ and $\widetilde{\boldsymbol{C}}$, defined as the matrices with columns $(\boldsymbol{c}_e - \boldsymbol{c}_0)$ and $(\widetilde{\boldsymbol{c}}_e - \widetilde{\boldsymbol{c}}_0)$, respectively. That is, the relative shift vectors are identifiable up to affine transformation, regardless of whether they are linear in $\boldsymbol{l}$. This has implications, e.g., for the CPA model of Lotfollahi et al. (2023) which includes element-wise nonlinear dose-response functions applied to $\boldsymbol{l}$, see Appx. C.4 for details. However, linearity is leveraged in the proof of our extrapolation result (Thm. 4.6) where it is used in (4.6) to establish a link between $\boldsymbol{l}_{\text{test}}$ and the training perturbations. Since only $\boldsymbol{l}_{\text{test}}$ is observed at test time, the above argument thus cannot easily be extended to the extrapolation setting, as this would require establishing a link between $\boldsymbol{c}_{\text{test}}$ and the training shifts, all of which are unobserved.

5 Estimation Method: Perturbation Distribution Autoencoder

To leverage the extrapolation guarantees of Thm. 4.6, we seek to estimate the parts of the generative process that are identifiable according to Thm. 4.1 from the available data in (2.1). To this end, we build on an autoencoder framework and adapt it for multi-domain perturbation modelling and distributional regression. Our method, the *perturbation distribution autoencoder* (PDAE), comprises an encoder, a perturbation matrix, and a (stochastic) decoder, trained to maximise the similarity between pairs of true and simulated perturbation distributions, see Fig. 2 for an overview.

Encoder. The encoder $\widehat{\boldsymbol{g}} : \mathbb{R}^{d_X} \to \mathbb{R}^{d_Z}$ maps observations $\boldsymbol{x}$ to the space of perturbation-relevant latents $\boldsymbol{z}$. Ideally, it should invert the stochastic mixing function in (3.4) in the sense of recovering the perturbed latent state $\boldsymbol{z}_{e,i}^{\text{pert}}$ in (3.2) from observation $\boldsymbol{x}_{e,i}$. We therefore denote the encoder output by

$$\widehat{\boldsymbol{z}}_{e,i}^{\text{pert}} := \widehat{\boldsymbol{g}}(\boldsymbol{x}_{e,i}), \tag{5.1}$$

and refer to it as estimated perturbed latent. This is a key difference to the CPA method of Lotfollahi et al. (2023), which seeks an encoder that maps to the latent basal state, regardless of the domain e.

Perturbation Model. If the encoder recovers the perturbed latents up to affine transformation, the additivity of perturbation effects assumed in (3.2) allows us to map between the latent distributions

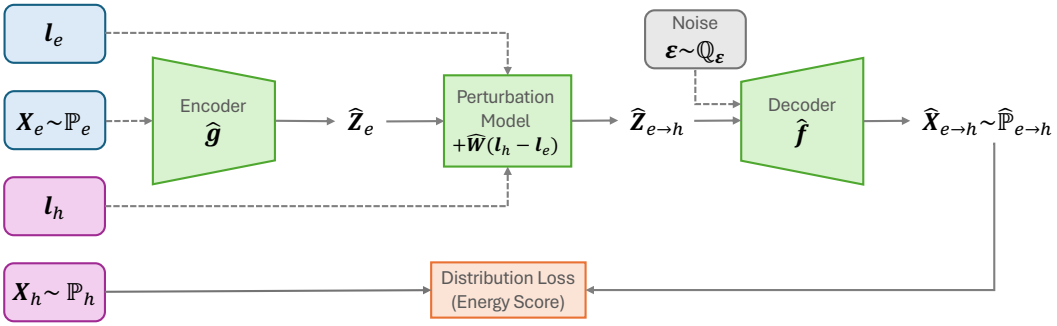


Figure 2: **Overview of the Perturbation Distribution Autoencoder (PDAE).** The distribution of a target perturbation condition h (purple) is simulated by encoding, perturbing, and decoding data from a source condition e (blue). Dashed arrows indicate model inputs and green boxes model components with learnable parameters, trained to maximise the similarity (orange) between the empirical true and simulated target distributions for all pairs of training domains (e, h). At test time, the target perturbation label $\boldsymbol{l}_h$ is replaced with an unseen $\boldsymbol{l}_{\text{test}}$.

underlying different perturbation conditions. Specifically, we use a perturbation model parametrised by a perturbation matrix $\widehat{\boldsymbol{W}} \in \mathbb{R}^{d_Z \times K}$ to create synthetic perturbed latents from domain h as:

$$\hat{\boldsymbol{z}}^{\text{pert}}_{e\to h,i} := \hat{\boldsymbol{z}}^{\text{pert}}_{e,i} + \widehat{\boldsymbol{W}}(\boldsymbol{l}_h - \boldsymbol{l}_e). \tag{5.2}$$

which can be interpreted as undoing the effects of perturbation $\boldsymbol{l}_e$ (i.e., mapping back to the latent basal state) and then simulating perturbation $\boldsymbol{l}_h$. (For $h = e$, this has no effect and $\boldsymbol{z}^{\text{pert}}_{e\to h,i} = \boldsymbol{z}^{\text{pert}}_{e,i}$.)

Decoder. The decoder $\hat{\boldsymbol{f}} : \mathbb{R}^{d_Z} \times \mathbb{R}^{d_\epsilon} \to \mathbb{R}^{d_X}$ maps estimated latents $\hat{\boldsymbol{z}}$ and noise $\boldsymbol{\varepsilon} \sim \mathbb{Q}_{\boldsymbol{\varepsilon}}$ back to observations. When viewed as a function of $\hat{\boldsymbol{z}}$ only, it is stochastic and induces the distribution $\hat{\boldsymbol{f}}(\hat{\boldsymbol{z}}, \cdot)_{\#}\mathbb{Q}_{\boldsymbol{\varepsilon}}$ from which we can sample synthetic observations,

$$\widehat{\boldsymbol{X}}_{e\to h,i} = \hat{\boldsymbol{f}}\left(\hat{\boldsymbol{z}}^{\text{pert}}_{e\to h,i}, \boldsymbol{\varepsilon}\right) \qquad \text{where} \qquad \boldsymbol{\varepsilon} \sim \mathbb{Q}_{\boldsymbol{\varepsilon}}\,. \tag{5.3}$$

Simulating Perturbation Distributions. Given a distribution $\mathbb{P}_e$ and the corresponding perturbation label $\boldsymbol{l}_e$, our model facilitates sampling synthetic observations for another perturbation condition with label $\boldsymbol{l}_h$. We denote the resulting distribution by $\hat{\mathbb{P}}_{e\to h}$, formally defined as the distribution of

$$\hat{\boldsymbol{f}}\left(\hat{\boldsymbol{g}}\left(\boldsymbol{X}_e\right) + \widehat{\boldsymbol{W}}\left(\boldsymbol{l}_h - \boldsymbol{l}_e\right), \boldsymbol{\varepsilon}\right) \qquad \text{where} \qquad \boldsymbol{X}_e \sim \mathbb{P}_e \qquad \text{and} \qquad \boldsymbol{\varepsilon} \sim \mathbb{Q}_{\boldsymbol{\varepsilon}}. \tag{5.4}$$

Learning Objective. To learn $\left(\hat{\boldsymbol{g}}, \hat{\boldsymbol{f}}, \widehat{\boldsymbol{W}}\right)$, we propose to minimise the pairwise distribution loss

$$\mathcal{L}\left(\hat{\boldsymbol{g}}, \hat{\boldsymbol{f}}, \widehat{\boldsymbol{W}}; \left\{(\mathbb{P}_e, \boldsymbol{l}_e)\right\}_{e\in[M]_0}\right) = \sum_{e,h\in[M]_0} d\left(\hat{\mathbb{P}}_{e\to h}, \mathbb{P}_h\right) \tag{5.5}$$

where $\hat{\mathbb{P}}_{e\to h}$ depends on $(\mathbb{P}_e, \boldsymbol{l}_e, \boldsymbol{l}_h)$ and the model parameters via (5.4); and d is a measure of dissimilarity between distributions. Here, we use the negative expected energy score for d, i.e.,

$$d\left(\hat{\mathbb{P}}_{e\to h}, \mathbb{P}_h\right) = -\mathbb{E}_{\boldsymbol{X}_h\sim\mathbb{P}_h}\left[\text{ES}_\beta\left(\hat{\mathbb{P}}_{e\to h}, \boldsymbol{X}_h\right)\right], \tag{5.6}$$

where ES_β denotes the energy-score (Gneiting & Raftery, 2007), defined for $\beta \in (0, 2)$ as

$$\text{ES}_\beta(\mathbb{P}, \boldsymbol{x}) = \frac{1}{2}\mathbb{E}_{\boldsymbol{X},\boldsymbol{X}'\overset{\text{i.i.d.}}{\sim}\mathbb{P}}\left\|\boldsymbol{X} - \boldsymbol{X}'\right\|^\beta - \mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}}\left\|\boldsymbol{X} - \boldsymbol{x}\right\|^\beta. \tag{5.7}$$

It is a strictly proper scoring rule, meaning that the expected energy score $\mathbb{E}_{\boldsymbol{X}}[\text{ES}_\beta(\mathbb{P}, \boldsymbol{X})]$ is maximised if and only if $\boldsymbol{X} \sim \mathbb{P}$, see Appx. C.1 for details on probabilistic forecasting and scoring rules. Combined with its computational simplicity, this property makes the negative expected energy score a popular loss function for distributional regression (Shen & Meinshausen, 2024a;b).

Corollary 5.1. *The objective in (5.5) is minimised if and only if* $\mathbb{P}_h = \hat{\mathbb{P}}_{e\to h}$ *for all* $e, h \in [M]_0$.

Training. Since we only have access to empirical distributions, we approximate the expectations in (5.6) and (5.7) with Monte Carlo samples based on the available data (2.1). Moreover, for a fixed encoder, the optimal perturbation matrix is available in closed form and given by the ordinary least squares solution to regressing the domain-specific means of $\widehat{\boldsymbol{z}}^{\text{pert}}_{e,i}$ on the corresponding perturbation labels $\boldsymbol{l}_e$. We, therefore, write $\widehat{\boldsymbol{W}}$ as a function of $\widehat{\boldsymbol{g}}$ and optimise (5.5) with respect to the parameters of $\widehat{\boldsymbol{g}}$ and $\widehat{\boldsymbol{f}}$ using stochastic gradient descent (Kingma & Ba, 2015).

Prediction. To simulate the distribution for a new perturbation label $\boldsymbol{l}_{\text{test}}$, we use our model to compute the synthetic perturbed test latents $\widehat{\boldsymbol{z}}^{\text{pert}}_{e\to\text{test},i}$ for all $e \in [M]_0$ and all $i \in [N_e]$ via (5.2), and then sample the corresponding synthetic test observations $\widehat{\boldsymbol{X}}_{e\to\text{test},i}$ according to (5.3). In other words, our estimate of $\mathbb{P}_{\text{test}}$ is the pooled version of the domain-specific empirical synthetic distributions,

$$\widehat{\mathbb{P}}_{\text{test}} = \frac{1}{M+1} \sum_{e\in[M]_0} \widehat{\mathbb{P}}_{e\to\text{test}} \tag{5.8}$$

6 Experiments

We present preliminary empirical evidence that our approach can outperform existing methods at distributional perturbation extrapolation. As we focus on a simple, controlled setting with synthetic data, our results should be viewed as proof of concept, rather than as comprehensive empirical study.

Data. For ease of visualisation, we consider $d_Z = d_X = 2$-dimensional latents and observations. The base distribution $\mathbb{P}_{\boldsymbol{Z}}$ is a zero-mean, isotropic Gaussian with standard deviation $\sigma = 0.25$. We consider $K = 3$ elementary perturbations with associated shift vectors $\boldsymbol{w}_1 = (1,0)^\top$, $\boldsymbol{w}_2 = (0,1)^\top$, and $\boldsymbol{w}_3 = (1,1)^\top$. We create $M+1 = 4$ training domains with labels $\boldsymbol{l}_0 = (0,0,0)^\top$, $\boldsymbol{l}_1 = (1,0,0)^\top$, $\boldsymbol{l}_2 = (0,1,0)^\top$, and $\boldsymbol{l}_3 = (0,0,1)^\top$, and test on $\boldsymbol{l}^{\text{ID}}_{\text{test}} = (1,1,0)^\top$ and $\boldsymbol{l}^{\text{OOD}}_{\text{test}} = (1,0,1)^\top$. By construction, $\boldsymbol{l}^{\text{ID}}_{\text{test}}$ results in the same mean shift of $(1,1)^\top$ as $\boldsymbol{l}_3$, whereas $\boldsymbol{l}^{\text{OOD}}_{\text{test}}$ results in a different shift not seen during training. We, therefore, refer to the respective test cases as in-distribution (ID) and out-of-distribution (OOD) relative to the decoder inputs seen during training. To generate observations, we use the complex exponential $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{z}) = \exp(z_1)(\cos z_2, \sin z_2)$ as a deterministic nonlinear mixing function, which was also used to generate Fig. 1a (where both test cases are partially OOD). The resulting datasets are in shown in Fig. 3 (left) in Appx. A.

Methods. We compare our approach with CPA (Lotfollahi et al., 2023) and the following baselines:

- *Pooled Mean*: pool all training observations, then output the mean;
- *Pseudobulked Mean*: pool only data arising from individual perturbations involved in the combination to be predicted (e.g., $\boldsymbol{l}_1$ and $\boldsymbol{l}_2$ for $\boldsymbol{l}^{\text{ID}}_{\text{test}}$), then output the mean;
- *Linear Regression*: linearly regress the environment-specific means of observations $\boldsymbol{\mu}^e_{\boldsymbol{x}}$ on $\boldsymbol{l}_e$ and use the resulting model to predict the test means $\boldsymbol{\mu}^{\text{test}}_{\boldsymbol{x}}$ from $\boldsymbol{l}_{\text{test}}$.

The former two were used by Lotfollahi et al. (2023); we propose the latter as an additional baseline.

Metrics. To assess distributional fit, we use the energy distance (ED; Székely & Rizzo, 2013), i.e., twice the normalized negative expected energy score, and the maximum mean discrepancy (MMD; Gretton et al., 2012) with Gaussian kernel and bandwidth chosen by the median heuristic, see Appx. C.2 for more on measures of distributional similarity. Since some methods only predict the mean, we also report the L2 norm of the difference between predicted and true mean, $\|\boldsymbol{\mu}_{\boldsymbol{x}} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}}\|$.

Experimental Details. We generate $N_e = 2^{14}$ observations for each domain. Both CPA and PDAE use 4-hidden layer MLPs with 64 hidden units as encoders and decoders and are trained for 2000 epochs using a batch size of 2^{12}. For CPA, all other hyperparameters are set to their default values. For PDAE, we set $\beta = 1$ in (5.7) and use a learning rate of 0.005.

Results. The quantitative results are summarized in Tab. 1, see also Fig. 3 in Appx. A for qualitative results. For the ID test setting, PDAE performs best, achieving near-perfect distributional fit. CPA outperforms linear regression but does substantially worse than PDAE at both mean and distributional prediction. For the OOD test setting, all methods perform much worse, with PDAE yielding the least bad performance. This failure for the OOD case is expected. Despite learning a good representation and perturbation model (as evident from the strong ID test performance), the decoder did not encounter inputs similar to the perturbed test latents during training and has thus not learnt which part of the observation space to map them to.

Table 1: **Results on Simulated Data.** For all metrics, lower is better. Best results highlighted in bold. Brackets indicate distributional similarities calculated using only the mean, i.e., a sample size of one.

Method	In-Distribution Test			Out-of-Distribution Test		
	ED	MMD	$\|\boldsymbol{\mu}_{\boldsymbol{x}} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}}\|_2$	ED	MMD	$\|\boldsymbol{\mu}_{\boldsymbol{x}} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{x}}\|_2$
Pooled Mean	(1.23)	(4.85)	0.82	(2.80)	(5.98)	1.74
Pseudobulked Mean	(1.49)	(5.43)	0.96	(2.45)	(5.56)	1.57
Linear Regression	(0.81)	(3.42)	0.60	(1.22)	(3.35)	0.88
CPA (Lotfollahi et al., 2023)	0.17	0.57	0.36	3.09	5.02	2.15
PDAE **(Ours)**	**0.001**	**0.005**	**0.03**	**0.45**	**1.33**	**0.61**

7 Discussion

Single-Cell vs Population-Level Effects. Since cells are typically destroyed during measurement, each unit is only observed for one perturbation condition. As a result, we do not have access to paired data, which would be required to establish ground-truth single-cell level effects. Instead, we can only observe the effects of perturbations at the population level. We, therefore, formulate the perturbation extrapolation task, our theoretical guarantees, and our learning objective in distributional terms. This contrasts with some prior works, which, despite a lack of ground truth training data, pursue the seemingly infeasible task of making counterfactual predictions at the single-cell level. While our model can, in principle, also make such predictions, we stress that they are not falsifiable from empirical data and that our guarantees do not extend to the single-cell resolution. For this reason, we consider distributional perturbation extrapolation a more meaningful and feasible task formulation.

Relation to Causal Models. In the field of causal inference, experimental data resulting from perturbations is modelled via interventions in an underlying causal model. For example, in the structural causal model (SCM; Pearl, 2009) framework, interventions modify a set of assignments, which determine each variable from its direct causes and unexplained noise. In general, our model for the effect of perturbations in latent space (§ 3) differs from how interventions are treated in SCMs. However, as detailed in Appx. D, if we restrict our attention to linear SCMs, then the class of shift interventions can be viewed as a special case of our model with $K = d_Z$ elementary perturbations and a particular choice of perturbation matrix $\boldsymbol{W}$. In this sense, our approach may be interpreted as causal representation learning (Schölkopf et al., 2021) with a linear latent causal model (Buchholz et al., 2024; Squires et al., 2023) and generalized shift interventions (Zhang et al., 2024).

Is Gaussianity Necessary? In Thm. 4.1, Gaussianity of $\mathbb{P}_{\boldsymbol{Z}}$ is assumed to prove identifiability. While this is a sufficient condition, it may not be necessary and can possibly be relaxed. Indeed, our method does not explicitly enforce any particular latent distribution. Moreover, preliminary empirical evidence suggests that training on simulated data generated from a Laplacian or Uniform basal state distribution also yields identifiable latent spaces and perturbation effects.

Open Problems. We consider the requirement for the decoder to generalize to new inputs the biggest open problem for distributional perturbation extrapolation. While this issue is absent from our theory, where $\mathbb{P}_{\boldsymbol{Z}}$ is Gaussian and thus has full support, it can pose major challenges in practice when learning from finite data. As discussed at the end of § 6, the issue of decoder extrapolation is orthogonal to learning the correct representation and perturbation model and thus appears fundamental. Besides addressing decoder extrapolation, future work should evaluate the proposed approach on more complex, noisy real-world data, extend our theoretical results to partially identifiable settings, and pursue extensions that incorporate covariates and nonlinear dose-response functions.

Meaningfulness Statement

We consider a "meaningful representation of life" an embedding of biological data that facilitates drawing non-trivial inferences, such as predicting the effects of new (combinations of) interventions or generalizing to new cell types or species. In the present work, we focus on a class of representations which we show to be provably meaningful in this sense and propose a principled estimation method to learn such representations from multi-domain perturbation data.

REFERENCES

Ludwig Baringhaus and Carsten Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206, 2004. [Cited on p. 16.]

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems*, volume 14, 2001. [Cited on p. 17.]

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. [Cited on p. 17.]

Jack Brady, Roland S Zimmermann, Yash Sharma, Bernhard Schölkopf, Julius von Kügelgen, and Wieland Brendel. Provably learning object-centric representations. In *International Conference on Machine Learning*, pp. 3038–3062. PMLR, 2023. [Cited on p. 3.]

Jack Brady, Julius von Kügelgen, Sebastien Lachapelle, Simon Buchholz, Thomas Kipf, and Wieland Brendel. Interaction asymmetry: A general principle for learning composable abstractions. In *International Conference on Learning Representations*, 2025. [Cited on p. 3.]

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In *Advances in Neural Information Processing Systems*, volume 36, 2024. [Cited on p. 8.]

Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia Del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature methods*, 20(11), 2023. [Cited on p. 1.]

Lawrence Cayton. Algorithms for manifold learning, 2005. Technical Report, University of California at San Diego. [Cited on p. 17.]

Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *cell*, 167(7):1853–1866, 2016. [Cited on p. 1.]

Kefan Dong and Tengyu Ma. First steps toward understanding the extrapolation of nonlinear models to unseen domains. In *International Conference on Learning Representations*, 2022. [Cited on p. 3.]

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. [Cited on p. 2, 3, 6, and 17.]

Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266):20210068, 2022. [Cited on p. 3.]

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. [Cited on p. 7 and 17.]

Leon Hetzel, Simon Boehm, Niki Kilbertus, Stephan Günnemann, Fabian Theis, et al. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26711–26722, 2022. [Cited on p. 1.]

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006. [Cited on p. 1 and 17.]

Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A Doudna, and Emmanuelle Charpentier. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *science*, 337(6096):816–821, 2012. [Cited on p. 1.]

Kenji Kamimoto, Blerta Stringa, Christy M Hoffmann, Kunal Jindal, Lilianna Solnica-Krezel, and Samantha A Morris. Dissecting cell identity via network inference and in silico gene perturbation. *Nature*, 614(7949):742–751, 2023. [Cited on p. 1.]

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217, 2020. [Cited on p. 5.]

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. [Cited on p. 7.]

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014. [Cited on p. 1.]

Roger Koenker. *Quantile regression*. Cambridge University Press, 2005. [Cited on p. 3.]

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: Journal of the Econometric Society*, pp. 33–50, 1978. [Cited on p. 3.]

Sébastien Lachapelle, Divyat Mahajan, Ioannis Mitliagkas, and Simon Lacoste-Julien. Additive decoders for latent variables identification and cartesian-product extrapolation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. [Cited on p. 3.]

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017. [Cited on p. 3.]

Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, volume 30, 2017. [Cited on p. 19.]

Samuel Lippl and Kim Stachenfeld. When does compositional structure yield compositional generalization? a kernel theory. *arXiv preprint arXiv:2405.16391*, 2024. [Cited on p. 3.]

Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. scgen predicts single-cell perturbation responses. *Nature methods*, 16(8):715–721, 2019. [Cited on p. 1.]

Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, Jay Shendure, Jose L McFaline-Figueroa, Pierre Boyeau, F Alexander Wolf, Nafissa Yakubova, Stephan Günnemann, Cole Trapnell, David Lopez-Paz, and Fabian J Theis. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular Systems Biology*, 19(6):e11517, 2023. [Cited on p. 1, 2, 3, 5, 7, 8, 18, and 19.]

James E Matheson and Robert L Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096, 1976. [Cited on p. 16.]

Mehdi Mirza. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. [Cited on p. 3.]

Milton L. Montero, Jeffrey Bowers, Rui Ponte Costa, Casimir JH Ludwig, and Gaurav Malhotra. Lost in latent space: Examining failures of disentangled models at combinatorial generalisation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. [Cited on p. 3.]

Milton Llera Montero, Casimir JH Ludwig, Rui Ponte Costa, Gaurav Malhotra, and Jeffrey Bowers. The role of disentanglement in generalisation. In *International Conference on Learning Representations*, 2021. [Cited on p. 3.]

Thomas M Norman, Max A Horlbeck, Joseph M Replogle, Alex Y Ge, Albert Xu, Marco Jost, Luke A Gilbert, and Jonathan S Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *science*, 365(6455):786–793, 2019. [Cited on p. 1.]

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2nd edition, 2009. [Cited on p. 8 and 19.]

Xiaoning Qi, Lianhe Zhao, Chenyu Tian, Yueyue Li, Zhen-Lin Chen, Peipei Huo, Runsheng Chen, Xiaodong Liu, Baoping Wan, Shengyong Yang, et al. Predicting transcriptional responses to novel chemical perturbations using deep generative model for drug discovery. *Nature Communications*, 15(1):1–19, 2024. [Cited on p. 1.]

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 42(6):927–935, 2024. [Cited on p. 1.]

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986. [Cited on p. 1 and 17.]

Lawrence K Saul and Sam T Roweis. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2003. [Cited on p. 17.]

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002. [Cited on p. 17.]

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. [Cited on p. 8.]

Lukas Schott, Julius von Kügelgen, Frederik Träuble, Peter Vincent Gehler, Chris Russell, Matthias Bethge, Bernhard Schölkopf, Francesco Locatello, and Wieland Brendel. Visual representation learning does not generalize strongly within the same domain. In *International Conference on Learning Representations*, 2022. [Cited on p. 3.]

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, pp. 2263–2291, 2013. [Cited on p. 17.]

Xinwei Shen and Nicolai Meinshausen. Distributional principal autoencoders. *arXiv preprint arXiv:2404.13649*, 2024a. [Cited on p. 6 and 18.]

Xinwei Shen and Nicolai Meinshausen. Engression: extrapolation through the lens of distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2024b. [Cited on p. 6.]

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, volume 28, 2015. [Cited on p. 3.]

Chandler Squires, Anna Seigal, Salil Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *40th International Conference on Machine Learning*, 2023. [Cited on p. 8.]

Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. [Cited on p. 7 and 17.]

Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000. [Cited on p. 17.]

Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009. [Cited on p. 1.]

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013. [Cited on p. 1.]

Thaddäus Wiedemer, Jack Brady, Alexander Panfilov, Attila Juhos, Matthias Bethge, and Wieland Brendel. Provable compositional generalization for object-centric learning. In *International Conference on Learning Representations*, 2024a. [Cited on p. 3.]

Thaddäus Wiedemer, Prasanna Mayilvahanan, Matthias Bethge, and Wieland Brendel. Compositional generalization from first principles. In *Advances in Neural Information Processing Systems*, volume 36, 2024b. [Cited on p. 3.]

Christina Winkler, Daniel Worrall, Emiel Hoogeboom, and Max Welling. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*, 2019. [Cited on p. 3.]

Hengshi Yu and Joshua D Welch. Perturbnet predicts single-cell responses to unseen chemical and genetic perturbations. *BioRxiv*, pp. 2022–07, 2022. [Cited on p. 1.]

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 36, 2024. [Cited on p. 8.]

Appendix

A Additional Results

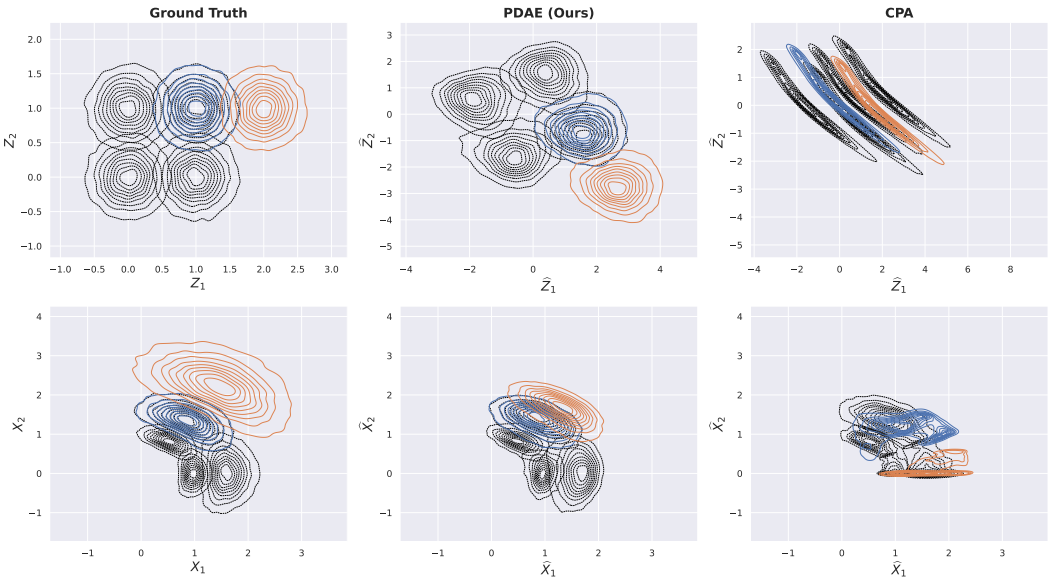


Figure 3: **Comparison of PDAE and CPA on Synthetic Data.** Shown are the results of our experiment described in § 6. Rows correspond to latent space (top) and observation space (bottom). Columns show the ground truth data (left), PDAE predictions (center), and CPA predictions (right). Training domains are shown in grey, the in-distribution (ID) test domain (which overlaps with one of the training domains) in blue, and the out-of-distribution (OOD) test domain in orange. All plots show kernel density estimates of the distributions. As can be seen, PDAE recovers an affine transformation of the true latents (top, center) leading to accurate distributional predictions for the training and ID test domain (bottom, center). However, the OOD test domain is mapped to a part of the latent space not seen during training (top, center). As a result, the corresponding decoder output does not accurately match the true OOD distribution (bottom left vs center). CPA appears to learn a latent space, in which all perturbed latent distributions are co-linear (top right), and the predicted distributions do not match the ground truth particularly well, particularly for the test conditions (bottom left vs right). However, recall that—unlike PDAE—CPA is not trained for distributional reconstruction, see Appx. C.4 for details.

B PROOFS

B.1 PROOF OF THM. 4.1

Theorem 4.1 (Affine identifiability for Gaussian latents)**.** *For $M \in \mathbb{Z}_{\geq 0}$, let $\boldsymbol{l}_0, ..., \boldsymbol{l}_M \in \mathbb{R}^K$ be $M+1$ perturbation labels. Let $\boldsymbol{f}, \widetilde{\boldsymbol{f}} : \mathbb{R}^{d_Z} \to \mathbb{R}^{d_X}$, $\boldsymbol{W}, \widetilde{\boldsymbol{W}} \in \mathbb{R}^{d_Z \times K}$, and $\mathbb{P}, \widetilde{\mathbb{P}}$ be distributions on $\mathbb{R}^{d_Z}$ such that the models $(\boldsymbol{f}, \boldsymbol{W}, \mathbb{P})$ and $(\widetilde{\boldsymbol{f}}, \widetilde{\boldsymbol{W}}, \widetilde{\mathbb{P}})$ induce the same observed distributions, i.e.,*

$$\forall e \in [M]_0 : \qquad \boldsymbol{f}\left(\boldsymbol{Z} + \boldsymbol{W}\boldsymbol{l}_e\right) \stackrel{d}{=} \widetilde{\boldsymbol{f}}\left(\widetilde{\boldsymbol{Z}} + \widetilde{\boldsymbol{W}}\boldsymbol{l}_e\right), \quad \text{where} \quad \boldsymbol{Z} \sim \mathbb{P} \quad \text{and} \quad \widetilde{\boldsymbol{Z}} \sim \widetilde{\mathbb{P}}. \tag{4.1}$$

Assume further that:

(i) ***[invertibility]*** *$\boldsymbol{f}$ and $\widetilde{\boldsymbol{f}}$ are C^2-diffeomorphisms onto their respective images;*

(ii) ***[Gaussianity]*** *$\boldsymbol{Z}$ and $\widetilde{\boldsymbol{Z}}$ are non-degenerate multi-variate Gaussians, i.e., $\mathbb{P} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\widetilde{\mathbb{P}} = \mathcal{N}\left(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}\right)$ for some $\boldsymbol{\mu}, \widetilde{\boldsymbol{\mu}} \in \mathbb{R}^{d_Z}$ and positive-definite $\boldsymbol{\Sigma}, \widetilde{\boldsymbol{\Sigma}} \in \mathbb{R}^{d_Z \times d_Z}$;*

(iii) ***[sufficient diversity]*** *the matrix $\widetilde{\boldsymbol{W}}\boldsymbol{L} \in \mathbb{R}^{d_Z \times M}$, where $\boldsymbol{L} \in \mathbb{R}^{K \times M}$ is the matrix with columns $(\boldsymbol{l}_e - \boldsymbol{l}_0)$ for $e \in [M]$, has full row rank, i.e., $\operatorname{rank}(\widetilde{\boldsymbol{W}}\boldsymbol{L}) = d_Z$.*

Then the latent representation and the effects of the observed perturbation combinations relative to $\boldsymbol{l}_0$ (as captured by $\boldsymbol{W}\boldsymbol{L}$) are identifiable up to affine transformation in the following sense:

$$\forall \boldsymbol{z} : \qquad \widetilde{\boldsymbol{f}}^{-1} \circ \boldsymbol{f}(\boldsymbol{z}) = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{b}, \tag{4.2}$$

$$\widetilde{\boldsymbol{W}}\boldsymbol{L} = \boldsymbol{A}\boldsymbol{W}\boldsymbol{L}, \tag{4.3}$$

where $\boldsymbol{A} := \widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}} \boldsymbol{\Sigma}^{-\frac{1}{2}}$ and $\boldsymbol{b} := \widetilde{\boldsymbol{\mu}} - \boldsymbol{A}\boldsymbol{\mu} + (\widetilde{\boldsymbol{W}} - \boldsymbol{A}\boldsymbol{W})\boldsymbol{l}_0$.

Proof. Let p_e and $\widetilde{p}_e$ denote the densities of

$$\boldsymbol{Z}_e := \boldsymbol{Z} + \boldsymbol{W}\boldsymbol{l}_e \tag{B.1}$$

and

$$\widetilde{\boldsymbol{Z}}_e := \widetilde{\boldsymbol{Z}} + \widetilde{\boldsymbol{W}}\boldsymbol{l}_e, \tag{B.2}$$

respectively. Due to (4.1), $\boldsymbol{f}$ and $\widetilde{\boldsymbol{f}}$ have the same image. Thus, by the invertibility assumption *(i)*, the function $\boldsymbol{h} := \widetilde{\boldsymbol{f}}^{-1} \circ \boldsymbol{f} : \mathbb{R}^{d_Z} \to \mathbb{R}^{d_Z}$ is a well-defined C^2-diffeomorphism. The change of variable formula applied to

$$\widetilde{\boldsymbol{Z}}_e \stackrel{d}{=} \boldsymbol{h}(\boldsymbol{Z}_e) \tag{B.3}$$

then yields for all e and all $\boldsymbol{z}$:

$$p_e(\boldsymbol{z}) = \widetilde{p}_e\left(\boldsymbol{h}(\boldsymbol{z})\right) \left|\det \boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{z})\right|, \tag{B.4}$$

where $\boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{z})$ denotes the Jacobian of $\boldsymbol{h}$. By taking logarithms of (B.4) and contrasting domain e with a reference domain with $e = 0$, the determinant terms cancel and we obtain for all e and all $\boldsymbol{z}$:

$$\log p_e(\boldsymbol{z}) - \log p_0(\boldsymbol{z}) = \log \widetilde{p}_e\left(\boldsymbol{h}(\boldsymbol{z})\right) - \log \widetilde{p}_0\left(\boldsymbol{h}(\boldsymbol{z})\right). \tag{B.5}$$

Next, denote the densities of $\boldsymbol{Z}$ and $\widetilde{\boldsymbol{Z}}$ by p and $\widetilde{p}$, respectively. From (B.1) and (B.2) it then follows that, for all e and all $\boldsymbol{z}$, p_e and $\widetilde{p}_e$ can respectively be expressed in terms of p and $\widetilde{p}$ as follows:

$$p_e(\boldsymbol{z}) = p\left(\boldsymbol{z} - \boldsymbol{W}\boldsymbol{l}_e\right), \tag{B.6}$$

$$\widetilde{p}_e(\boldsymbol{z}) = \widetilde{p}\left(\boldsymbol{z} - \widetilde{\boldsymbol{W}}\boldsymbol{l}_e\right). \tag{B.7}$$

By substituting these expressions in (B.5), we obtain for all e and all $\boldsymbol{z}$:

$$\log p(\boldsymbol{z} - \boldsymbol{W}\boldsymbol{l}_e) - \log p(\boldsymbol{z} - \boldsymbol{W}\boldsymbol{l}_0) = \log \widetilde{p}\left(\boldsymbol{h}(\boldsymbol{z}) - \widetilde{\boldsymbol{W}}\boldsymbol{l}_e\right) - \log \widetilde{p}\left(\boldsymbol{h}(\boldsymbol{z}) - \widetilde{\boldsymbol{W}}\boldsymbol{l}_0\right). \tag{B.8}$$

By the Gaussianity assumption *(ii)*, the contrast of log-densities in (B.8) takes the following form for all e and all $\boldsymbol{z}$:

$$(\boldsymbol{l}_e - \boldsymbol{l}_0)^\top \boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}) \qquad -\frac{1}{2}(\boldsymbol{l}_e - \boldsymbol{l}_0)^\top \boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{W}(\boldsymbol{l}_e + \boldsymbol{l}_0) \tag{B.9}$$

$$= (\boldsymbol{l}_e - \boldsymbol{l}_0)^\top \widetilde{\boldsymbol{W}}^\top \widetilde{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{h}(\boldsymbol{z}) - \widetilde{\boldsymbol{\mu}}\right) \qquad -\frac{1}{2}(\boldsymbol{l}_e - \boldsymbol{l}_0)^\top \widetilde{\boldsymbol{W}}^\top \widetilde{\boldsymbol{\Sigma}}^{-1}\widetilde{\boldsymbol{W}}(\boldsymbol{l}_e + \boldsymbol{l}_0) \tag{B.10}$$

We aim to show that the two representations are related by an affine transformation, i.e., that all second-order derivatives of $\boldsymbol{h}$ are zero everywhere. Taking gradients w.r.t. $\boldsymbol{z}$ yields for all e and all $\boldsymbol{z}$:

$$(\boldsymbol{l}_e - \boldsymbol{l}_0)^\top \boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} = (\boldsymbol{l}_e - \boldsymbol{l}_0)^\top \widetilde{\boldsymbol{W}}^\top \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{z}). \tag{B.11}$$

Let $\boldsymbol{L} \in \mathbb{R}^{K \times M}$ be the matrix with columns $\boldsymbol{l}_e - \boldsymbol{l}_0$ for $e \in [M]$. Then, for all $\boldsymbol{z}$:

$$\boldsymbol{L}^\top \boldsymbol{W}^\top \boldsymbol{\Sigma}^{-1} = \boldsymbol{L}^\top \widetilde{\boldsymbol{W}}^\top \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{J}_{\boldsymbol{h}}(\boldsymbol{z}). \tag{B.12}$$

Differentiating once more w.r.t. $\boldsymbol{z}$ yields for all $\boldsymbol{z}$:

$$\boldsymbol{0} = \boldsymbol{L}^\top \widetilde{\boldsymbol{W}}^\top \widetilde{\boldsymbol{\Sigma}}^{-1} \boldsymbol{H}_{\boldsymbol{h}}(\boldsymbol{z}), \tag{B.13}$$

where the 3-tensor $\boldsymbol{H}_{\boldsymbol{h}}(\boldsymbol{z}) \in \mathbb{R}^{d_Z \times d_Z \times d_Z}$ denotes the Hessian of $\boldsymbol{h}$, i.e., for all $i, j \in [d_Z]$ and all $\boldsymbol{z}$

$$\boldsymbol{0} = \boldsymbol{L}^\top \widetilde{\boldsymbol{W}}^\top \widetilde{\boldsymbol{\Sigma}}^{-1} \frac{\partial^2}{\partial z_i \partial z_j} \boldsymbol{h}(\boldsymbol{z}). \tag{B.14}$$

By assumption *(iii)*, the matrix $\boldsymbol{L}^\top \widetilde{\boldsymbol{W}}^\top$ has full column rank and thus a left inverse, i.e., there exists $\boldsymbol{V} \in \mathbb{R}^{d_Z \times M}$ such that $\boldsymbol{V}\boldsymbol{L}^\top \widetilde{\boldsymbol{W}}^\top = \boldsymbol{I}_{d_Z}$. Multiplying (B.14) on the left by $\widetilde{\boldsymbol{\Sigma}}\boldsymbol{V}$ then yields for all $i, j \in [d_Z]$ and all $\boldsymbol{z}$:

$$\frac{\partial^2}{\partial z_i \partial z_j} \boldsymbol{h}(\boldsymbol{z}) = \boldsymbol{0}. \tag{B.15}$$

This implies that $\boldsymbol{h}$ must be affine, i.e., there exist $\boldsymbol{A} \in \mathbb{R}^{d_Z \times d_Z}$ and $\boldsymbol{b} \in \mathbb{R}^{d_Z}$ such that for all $\boldsymbol{z}$:

$$\boldsymbol{h}(\boldsymbol{z}) = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{b}. \tag{B.16}$$

Further, since $\boldsymbol{h}$ is invertible, $\boldsymbol{A}$ must be invertible.

Recall that $\widetilde{\boldsymbol{Z}}_e \overset{d}{=} \boldsymbol{h}(\boldsymbol{Z}_e) = \boldsymbol{A}\boldsymbol{Z}_e + \boldsymbol{b}$. It follows from (B.1), (B.2) and assumption *(ii)* that for all e:

$$\mathcal{N}\left(\widetilde{\boldsymbol{\mu}} + \widetilde{\boldsymbol{W}}\boldsymbol{l}_e, \widetilde{\boldsymbol{\Sigma}}\right) = \mathcal{N}\left(\boldsymbol{A}(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{l}_e) + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top\right). \tag{B.17}$$

By equating the covariances, we find

$$\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top = \widetilde{\boldsymbol{\Sigma}} \qquad \Longrightarrow \qquad \boldsymbol{A} = \widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}} \boldsymbol{\Sigma}^{-\frac{1}{2}}, \tag{B.18}$$

where the matrix square roots exist and are unique since $\boldsymbol{\Sigma}$ and $\widetilde{\boldsymbol{\Sigma}}$ are symmetric and positive definite. By equating the means, we obtain for all e:

$$\widetilde{\boldsymbol{\mu}} + \widetilde{\boldsymbol{W}}\boldsymbol{l}_e = \boldsymbol{A}(\boldsymbol{\mu} + \boldsymbol{W}\boldsymbol{l}_e) + \boldsymbol{b} \qquad \Longleftrightarrow \qquad (\widetilde{\boldsymbol{W}} - \boldsymbol{A}\boldsymbol{W})\boldsymbol{l}_e = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b} - \widetilde{\boldsymbol{\mu}}. \tag{B.19}$$

By contrasting (B.19) for all $e \in [M]$ with $e = 0$ as before, we obtain

$$(\widetilde{\boldsymbol{W}} - \boldsymbol{A}\boldsymbol{W})\boldsymbol{L} = \boldsymbol{0} \qquad \Longleftrightarrow \qquad \widetilde{\boldsymbol{W}}\boldsymbol{L} = \boldsymbol{A}\boldsymbol{W}\boldsymbol{L}. \tag{B.20}$$

Finally, by choosing $e = 0$ in (B.19) we obtain the desired expression for $\boldsymbol{b}$,

$$\boldsymbol{b} = \widetilde{\boldsymbol{\mu}} - \boldsymbol{A}\boldsymbol{\mu} + (\widetilde{\boldsymbol{W}} - \boldsymbol{A}\boldsymbol{W})\boldsymbol{l}_0. \tag{B.21}$$

This completes the proof. □

B.2 Proof of Cor. 4.2

Corollary 4.2 (Affine recovery of the perturbation matrix)**.** *If, in addition to the assumptions of Thm. 4.1, $\boldsymbol{L} \in \mathbb{R}^{K \times M}$ has full row rank (i.e., $\operatorname{rank}(\boldsymbol{L}) = K \leq M$), then the perturbation matrix $\boldsymbol{W}$ is identifiable up to affine transformation in the sense that*

$$\widetilde{\boldsymbol{W}} = \boldsymbol{A}\boldsymbol{W}, \tag{4.4}$$

for $\boldsymbol{A} := \widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}} \boldsymbol{\Sigma}^{-\frac{1}{2}}$. In this case, the expression for $\boldsymbol{b}$ in (4.2) simplifies to $\boldsymbol{b} = \widetilde{\boldsymbol{\mu}} - \boldsymbol{A}\boldsymbol{\mu}$.

Proof. If $\operatorname{rank}(\boldsymbol{L}) = K$, then $\boldsymbol{L}$ has a right inverse, i.e., there exists $\boldsymbol{K} \in \mathbb{R}^{M \times K}$ such that $\boldsymbol{L}\boldsymbol{K} = \boldsymbol{I}_K$. Right multiplication of (B.20) by $\boldsymbol{K}$ then yields

$$\widetilde{\boldsymbol{W}} = \boldsymbol{A}\boldsymbol{W} \tag{B.22}$$

Finally, substitution of (B.22) into (B.21) yields

$$\boldsymbol{b} = \widetilde{\boldsymbol{\mu}} - \boldsymbol{A}\boldsymbol{\mu}. \tag{B.23}$$

□

B.3 Proof of Thm. 4.6

Theorem 4.6 (Extrapolation to span of relative perturbations)**.** *Under the same setting and assumptions as in Thm. 4.1, let $\boldsymbol{l}_{\text{test}} \in \mathbb{R}^K$ be an unseen perturbation label such that*

$$(\boldsymbol{l}_{\text{test}} - \boldsymbol{l}_0) \in \operatorname{span}\left(\{\boldsymbol{l}_e - \boldsymbol{l}_0\}_{e \in [M]}\right). \tag{4.6}$$

Then the effect of $\boldsymbol{l}_{\text{test}}$ is uniquely identifiable in the sense that

$$\boldsymbol{X}_{\text{test}} = \boldsymbol{f}(\boldsymbol{Z} + \boldsymbol{W}\boldsymbol{l}_{\text{test}}) \overset{d}{=} \widetilde{\boldsymbol{f}}\left(\widetilde{\boldsymbol{Z}} + \widetilde{\boldsymbol{W}}\boldsymbol{l}_{\text{test}}\right) = \widetilde{\boldsymbol{X}}_{\text{test}}. \tag{4.7}$$

Proof. With $\boldsymbol{h} = \widetilde{\boldsymbol{f}}^{-1} \circ \boldsymbol{f}$, the condition in (4.7) is equivalent to

$$\boldsymbol{h}(\boldsymbol{Z} + \boldsymbol{W}\boldsymbol{l}_{\text{test}}) \overset{d}{=} \widetilde{\boldsymbol{Z}} + \widetilde{\boldsymbol{W}}\boldsymbol{l}_{\text{test}}. \tag{B.24}$$

By Thm. 4.1, we have $\boldsymbol{h}(\boldsymbol{z}) = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{b}$ for

$$\boldsymbol{A} = \widetilde{\boldsymbol{\Sigma}}^{\frac{1}{2}} \boldsymbol{\Sigma}^{-\frac{1}{2}}, \tag{B.25}$$

$$\boldsymbol{b} = \widetilde{\boldsymbol{\mu}} - \boldsymbol{A}\boldsymbol{\mu} + (\widetilde{\boldsymbol{W}} - \boldsymbol{A}\boldsymbol{W})\boldsymbol{l}_0. \tag{B.26}$$

Together with $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, this lets us compute the distribution of the LHS of (B.24) as:

$$\boldsymbol{h}(\boldsymbol{Z} + \boldsymbol{W}\boldsymbol{l}_{\text{test}}) = \boldsymbol{A}(\boldsymbol{Z} + \boldsymbol{W}\boldsymbol{l}_{\text{test}}) + \boldsymbol{b} \sim \mathcal{N}\left(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{W}\boldsymbol{l}_{\text{test}} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top\right). \tag{B.27}$$

Similarly, since $\widetilde{\boldsymbol{Z}} \sim \mathcal{N}\left(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}\right)$, the distribution of the RHS of (B.24) is given by

$$\widetilde{\boldsymbol{Z}} + \widetilde{\boldsymbol{W}}\boldsymbol{l}_{\text{test}} \sim \mathcal{N}\left(\widetilde{\boldsymbol{\mu}} + \widetilde{\boldsymbol{W}}\boldsymbol{l}_{\text{test}}, \widetilde{\boldsymbol{\Sigma}}\right). \tag{B.28}$$

From (B.25), it follows directly that

$$\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top = \widetilde{\boldsymbol{\Sigma}}. \tag{B.29}$$

To complete the proof, it thus remains to show that the means of (B.27) and (B.28) are equal, i.e.,

$$\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{W}\boldsymbol{l}_{\text{test}} + \boldsymbol{b} = \widetilde{\boldsymbol{\mu}} + \widetilde{\boldsymbol{W}}\boldsymbol{l}_{\text{test}}. \tag{B.30}$$

Starting from the LHS of (B.30), by substituting the expression for $\boldsymbol{b}$ from (B.26) we obtain:

$$\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{W}\boldsymbol{l}_{\text{test}} + \boldsymbol{b} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{W}\boldsymbol{l}_{\text{test}} + \widetilde{\boldsymbol{\mu}} - \boldsymbol{A}\boldsymbol{\mu} + (\widetilde{\boldsymbol{W}} - \boldsymbol{A}\boldsymbol{W})\boldsymbol{l}_0 \tag{B.31}$$

$$= \widetilde{\boldsymbol{\mu}} + \boldsymbol{A}\boldsymbol{W}(\boldsymbol{l}_{\text{test}} - \boldsymbol{l}_0) + \widetilde{\boldsymbol{W}}\boldsymbol{l}_0 \tag{B.32}$$

Next, by (4.6), i.e., the assumption that $(\boldsymbol{l}_{\text{test}} - \boldsymbol{l}_0)$ lies in the span of $\{\boldsymbol{l}_e - \boldsymbol{l}_0\}_{e\in[M]}$, there exists $\boldsymbol{\alpha} \in \mathbb{R}^M$ such that

$$\boldsymbol{l}_{\text{test}} - \boldsymbol{l}_0 = \sum_{e\in[M]} \alpha_e(\boldsymbol{l}_e - \boldsymbol{l}_0) = \boldsymbol{L}\boldsymbol{\alpha} \tag{B.33}$$

where, as before, $\boldsymbol{L} \in \mathbb{R}^{K\times M}$ is the matrix with columns $(\boldsymbol{l}_e - \boldsymbol{l}_0)$. Substituting (B.33) into (B.32) yields

$$\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{W}\boldsymbol{l}_{\text{test}} + \boldsymbol{b} = \widetilde{\boldsymbol{\mu}} + \boldsymbol{A}\boldsymbol{W}\boldsymbol{L}\boldsymbol{\alpha} + \widetilde{\boldsymbol{W}}\boldsymbol{l}_0. \tag{B.34}$$

Finally, it follows from Thm. 4.1 that

$$\boldsymbol{A}\boldsymbol{W}\boldsymbol{L} = \widetilde{\boldsymbol{W}}\boldsymbol{L}. \tag{B.35}$$

which upon substituion into (B.34) yields the desired equality from (B.30)

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{W}\boldsymbol{l}_{\text{test}} + \boldsymbol{b} &= \widetilde{\boldsymbol{\mu}} + \widetilde{\boldsymbol{W}}\boldsymbol{L}\boldsymbol{\alpha} + \widetilde{\boldsymbol{W}}\boldsymbol{l}_0 && \text{(B.36)}\\
&= \widetilde{\boldsymbol{\mu}} + \widetilde{\boldsymbol{W}}(\boldsymbol{L}\boldsymbol{\alpha} + \boldsymbol{l}_0) && \text{(B.37)}\\
&= \widetilde{\boldsymbol{\mu}} + \widetilde{\boldsymbol{W}}\boldsymbol{l}_{\text{test}}. && \text{(B.38)}
\end{aligned}$$

This completes the proof. □

C Additional Background Material and Related Work

Since the problem of interest (§ 2) involves making distributional predictions for new perturbation conditions, we review some basics of probabilistic forecasting (Appx. C.1) and measuring the similarity between two distributions (Appx. C.2), in our case typically between an empirical distribution and its predicted counterpart. We then turn to representation learning with encoder-decoder architectures (Appx. C.3), in particular a recent approach that also targets distributional reconstruction. Finally, we cover some prior efforts on perturbation modelling and extrapolation (Appx. C.4), which we draw inspiration from.

C.1 Probabilistic Forecasting and Scoring Rules

Let Ω be a sample space, $\mathcal{A}$ a σ-algebra of subsets of Ω, and $\mathcal{P}$ a convex class of probability measures on $(\Omega, \mathcal{A})$. A *probabilistic prediction* or *probabilistic forecast* is a mapping into $\mathcal{P}$, which outputs predictive distributions $\mathbb{P}$ over over outcomes $x \in \Omega$. Probabilistic forecasting can thus be viewed as a distributional generalization point prediction (i.e., deterministic forecasting), which maps directly into Ω.

To evaluate, compare, or rank different forecasts, it is useful to assign them a numerical score reflecting their quality. A *scoring rule* is a function $S : \mathcal{P} \times \Omega \to \mathbb{R}$ that assigns a score $S(\mathbb{P}, x)$ to forecast $\mathbb{P}$ if event x materializes, with higher scores corresponding to better forecasts—akin to (negative) loss or cost functions for point predictions. If x is distributed according to $\mathbb{Q}$, we denote the *expected score* by $S(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{x\sim\mathbb{Q}}[S(\mathbb{P}, x)]$. The scoring rule is called *proper* if $S(\mathbb{Q}, \mathbb{Q}) \geq S(\mathbb{P}, \mathbb{Q})$ for all $\mathbb{P}$, and *strictly proper* if equality holds if and only if $\mathbb{P} = \mathbb{Q}$.

CRPS. For continuous scalar random variables (i.e., $\Omega = \mathbb{R}$), a popular scoring rule is the *continuous ranked probability score* (CRPS; Matheson & Winkler, 1976). When $\mathcal{P}$ is the space of Borel probability measures on $\mathbb{R}$ with finite first moment, it is strictly proper and given by:[1]

$$\text{CRPS}(\mathbb{P}, x) = \frac{1}{2}\mathbb{E}_{X,X'\overset{\text{i.i.d.}}{\sim}\mathbb{P}}\left|X - X'\right| - \mathbb{E}_{X\sim\mathbb{P}}\left|X - x\right|. \tag{C.1}$$

If $\mathbb{P}$ is a point mass, the negative CRPS reduces to the absolute error loss function; it can thus be viewed as a generalization thereof to probabilistic forecasts.

[1] The original definition by Matheson & Winkler (1976) is $\text{CRPS}(F, x) = -\int_{-\infty}^{\infty}(F(y) - 1\{y \geq x\})^2 \mathrm{d}y$ where F is the CDF of $\mathbb{P}$, but the simpler form in (C.1) has been shown to be equivalent (Baringhaus & Franz, 2004, Lemma 2.2).

Energy Score. Gneiting & Raftery (2007) propose the *energy score* as a multi-variate generalization of the CRPS for vector-valued $\boldsymbol{x} \in \Omega = \mathbb{R}^{d_X}$. For $\beta \in (0, 2)$, it is defined by

$$\mathrm{ES}_\beta(\mathbb{P}, \boldsymbol{x}) = \frac{1}{2}\mathbb{E}_{\boldsymbol{X}, \boldsymbol{X}' \overset{\text{i.i.d.}}{\sim} \mathbb{P}} \left\|\boldsymbol{X} - \boldsymbol{X}'\right\|_2^\beta - \mathbb{E}_{\boldsymbol{X} \sim \mathbb{P}} \left\|\boldsymbol{X} - \boldsymbol{x}\right\|_2^\beta , \tag{C.2}$$

where $\|\cdot\|_2$ denotes the Euclidean (L2) norm. ES_β is strictly proper w.r.t. $\mathcal{P}_\beta$, the set of Borel probability measures $\mathbb{P}$ for which $\mathbb{E}_{\boldsymbol{X} \sim \mathbb{P}} \|\boldsymbol{X}\|_2^\beta$ is finite (Gneiting & Raftery, 2007).

C.2 Assessing Distributional Similarity

Energy Distance. The expected energy score $\mathrm{ES}_\beta(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{\boldsymbol{Y} \sim \mathbb{Q}}[\mathrm{ES}_\beta(\mathbb{P}, \boldsymbol{Y})]$ consitutes a measure of similarity between $\mathbb{P}$ and $\mathbb{Q}$ and is closely linked to the *energy distance* (Székely & Rizzo, 2013):

$$\mathrm{ED}_\beta(\mathbb{P}, \mathbb{Q}) = 2\mathbb{E}_{\boldsymbol{X} \sim \mathbb{P}, \boldsymbol{Y} \sim \mathbb{Q}} \|\boldsymbol{X} - \boldsymbol{Y}\|_2^\beta - \mathbb{E}_{\boldsymbol{X}, \boldsymbol{X}' \overset{\text{iid}}{\sim} \mathbb{P}} \left\|\boldsymbol{X} - \boldsymbol{X}'\right\|_2^\beta - \mathbb{E}_{\boldsymbol{Y}, \boldsymbol{Y}' \overset{\text{iid}}{\sim} \mathbb{Q}} \left\|\boldsymbol{Y} - \boldsymbol{Y}'\right\|_2^\beta \tag{C.3}$$

$$= 2\left(\mathrm{ES}_\beta(\mathbb{Q}, \mathbb{Q}) - \mathrm{ES}_\beta(\mathbb{P}, \mathbb{Q})\right) \geq 0 \tag{C.4}$$

with equality if and only if $\mathbb{P} = \mathbb{Q}$, since ES_β is a strictly proper scoring rule.

Maximum Mean Discrepancy (MMD). Another well-known distance between probability measures that is rooted in kernel methods (Schölkopf & Smola, 2002) is the *maximum mean discrepancy* (MMD; Gretton et al., 2012), which for a positive definite kernel $k : \Omega \times \Omega \to \mathbb{R}$ is given by

$$\mathrm{MMD}_k^2(\mathbb{P}, \mathbb{Q}) = \mathbb{E}_{X, X' \overset{\text{iid}}{\sim} \mathbb{P}}[k(X, X')] - 2\mathbb{E}_{X \sim \mathbb{P}, Y \sim \mathbb{Q}}[k(X, Y)] + \mathbb{E}_{Y, Y' \overset{\text{iid}}{\sim} \mathbb{Q}}[k(Y, Y')] . \tag{C.5}$$

Energy Distance as a Special Case of MMD. As shown by Sejdinovic et al. (2013), if k in (C.5) is chosen to be the positive-definite *distance kernel*[2]

$$k_\beta(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{2}\left(\|\boldsymbol{X}\|_2^\beta + \|\boldsymbol{Y}\|_2^\beta - \|\boldsymbol{X} - \boldsymbol{Y}\|_2^\beta\right) \tag{C.6}$$

then the energy distance is recovered as a special case of MMD,

$$\mathrm{ED}_\beta(\mathbb{P}, \mathbb{Q}) = 2\,\mathrm{MMD}_{k_\beta}^2(\mathbb{P}, \mathbb{Q}) . \tag{C.7}$$

C.3 Representation Learning

Many modern data sources of interest contain high-dimensional and unstructured observations $\boldsymbol{x}$, such as audio, video, images, or text. Representation learning aims to transform such data into a more compact, lower-dimensional set of features $\boldsymbol{z}$ (the representation) which preserves most of the relevant information while making it more easily accessible, e.g., for use in downstream tasks (Bengio et al., 2013).[3] For example, a representation of images of multi-object scenes could be a list of the contained objects, along with their size, position, colour, etc. A key assumption underlying this endeavour is the so-called *manifold hypothesis* which posits that high-dimensional natural data tends to lie near a low-dimensional manifold embedded in the high-dimensional ambient space; this idea is also at the heart of several (nonlinear) dimension reduction techniques (Belkin & Niyogi, 2001; Cayton, 2005; Saul & Roweis, 2003; Tenenbaum et al., 2000).

Autoencoder (AE). An autoencoder (AE; Hinton & Salakhutdinov, 2006; Rumelhart et al., 1986) is a pair of functions $(\boldsymbol{g}, \boldsymbol{f})$, consisting of an *encoder* $\boldsymbol{g} : \mathbb{R}^{d_X} \to \mathbb{R}^{d_Z}$ mapping observations $\boldsymbol{X} \sim \mathbb{P}$ to their representation $\boldsymbol{Z} := \boldsymbol{g}(\boldsymbol{X})$, and a *decoder* $\boldsymbol{f} : \mathbb{R}^{d_Z} \to \mathbb{R}^{d_X}$ mapping representations $\boldsymbol{Z}$ to

[2] induced by the negative-definite semi-metric $\rho(\boldsymbol{X}, \boldsymbol{Y}) = \|\boldsymbol{X} - \boldsymbol{Y}\|_2^\beta$ on $\Omega = \mathbb{R}^{d_X}$ (centered at the origin),

[3] Representation learning is thus closely related to the classical task of dimension reduction. The former usually refers to nonlinear settings, involves some form of machine learning, and tends to be more focused on usefulness in terms of downstream tasks, rather than, say, explained variance.

their reconstructions in observation space, $\widehat{\boldsymbol{X}} := \boldsymbol{f}(\boldsymbol{Z}) = \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{X}))$. Typically, $d_Z < d_X$, such that there is a bottleneck and perfect reconstruction is not feasible. Autoencoders are usually trained with a (point-wise, mean) reconstruction objective, i.e., the aim is to minimise the mean squared error

$$\mathcal{L}_{\text{AE}}(\boldsymbol{g}, \boldsymbol{f}; \mathbb{P}) := \mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}} \left\| \boldsymbol{X} - \widehat{\boldsymbol{X}} \right\|_2^2 = \mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}} \left\| \boldsymbol{X} - \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{X})) \right\|_2^2 , \tag{C.8}$$

w.r.t. both $\boldsymbol{g}$ and $\boldsymbol{f}$. As a result, the optimal decoder $\boldsymbol{f}^*_{\text{AE}}$ for any fixed encoder choice $\boldsymbol{g}$ is given by the conditional mean

$$\boldsymbol{f}^*_{\text{AE}}(\boldsymbol{z}; \boldsymbol{g}, \mathbb{P}) = \mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}}[\boldsymbol{X} | \boldsymbol{g}(\boldsymbol{X}) = \boldsymbol{z}] . \tag{C.9}$$

Distributional Principal Autoencoder (DPA). Since the objective of a standard AE is mean reconstruction, the distribution of reconstructions $\widehat{\boldsymbol{X}}$ is typically not the same as the distribution of $\boldsymbol{X}$ (unless the encoder is invertible, i.e, the compression is lossless and perfect reconstruction is feasible, which is usually not the case in practice). To address this, Shen & Meinshausen (2024a) proposed the distributional principal autoencoder (DPA) which targets distributional (rather than mean) reconstruction. A DPA also consists of an encoder-decoder pair $(\boldsymbol{g}, \boldsymbol{f})$. However, unlike in a standard AE, the DPA decoder $\boldsymbol{f} : \mathbb{R}^{d_Z} \times \mathbb{R}^{d_\epsilon} \to \mathbb{R}^{d_X}$ is stochastic and receives an additional noise term $\boldsymbol{\varepsilon}$ as input, which is sampled from a fixed distribution $\mathbb{Q}_{\boldsymbol{\varepsilon}}$ such as a standard isotropic Gaussian. The DPA loss function is constructed such that for a fixed encoder $\boldsymbol{g}$, the optimal DPA decoder $\boldsymbol{f}^*_{\text{DPA}}$ maps a given latent embedding $\boldsymbol{z}$, to the distribution of $\boldsymbol{X}$, given $\boldsymbol{g}(\boldsymbol{X}) = \boldsymbol{z}$, i.e.,

$$\boldsymbol{f}^*_{\text{DPA}}(\boldsymbol{z}, \varepsilon; \boldsymbol{g}) \overset{d}{=} \left(\boldsymbol{X} | \boldsymbol{g}(\boldsymbol{X}) = \boldsymbol{z} \right) , \tag{C.10}$$

where $\overset{d}{=}$ denotes equality in distribution. This means that the decoder evaluated at $\boldsymbol{z}$ should match the distribution of all realizations of $\boldsymbol{X}$ that are mapped by the encoder to $\boldsymbol{z}$. At the same time, the DPA encoder aims to minimise the variability in the distributions in (C.10) by encoding the first d_Z "principal" components. As shown by Shen & Meinshausen (2024a), both of these goals are achieved by the following DPA objective,

$$\begin{aligned} \mathcal{L}_{\text{DPA}}(\boldsymbol{g}, \boldsymbol{f}; \mathbb{P}) &= \mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}, \boldsymbol{\varepsilon}\sim\mathbb{Q}_{\boldsymbol{\varepsilon}}} \left\| \boldsymbol{X} - \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{X}), \varepsilon) \right\|_2^\beta - \frac{1}{2} \mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}, \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}' \overset{\text{iid}}{\sim} \mathbb{Q}_{\boldsymbol{\varepsilon}}} \left\| \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{X}), \varepsilon) - \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{X}), \varepsilon') \right\|_2^\beta \\ &= -\mathbb{E}_{\boldsymbol{X}\sim\mathbb{P}} \left[\text{ES}_\beta \left(\boldsymbol{f}\left(\boldsymbol{g}(\boldsymbol{X}), \cdot\right)_{\#} \mathbb{Q}_{\boldsymbol{\varepsilon}}, \boldsymbol{X} \right) \right] , \end{aligned} \tag{C.11}$$

where $\boldsymbol{f}\left(\boldsymbol{z}, \cdot\right)_{\#} \mathbb{Q}_{\boldsymbol{\varepsilon}}$ denotes the pushforward distribution of $\mathbb{Q}_{\boldsymbol{\varepsilon}}$ through the function $\boldsymbol{f}(\boldsymbol{z}, \cdot)$, i.e., the distribution of $\boldsymbol{f}(\boldsymbol{z}, \varepsilon)$ for a fixed $\boldsymbol{z}$ when $\varepsilon \sim \mathbb{Q}_{\boldsymbol{\varepsilon}}$. In other words, a DPA minimizes the negative expected energy score between $\boldsymbol{X}$ and the corresponding (stochastic) decoder output, conditional on the encoding of $\boldsymbol{X}$. Due to this conditioning, the DPA objective differs from an energy distance by a "normalization constant" which depends on the encoder and encourages capturing principal (i.e., variation-minimizing) components, rather than random latent dimensions.

C.4 Perturbation Modelling

Compositional Perturbation Autoencoder (CPA). Lotfollahi et al. (2023) propose the compositional perturbation autoencoder (CPA) as a model for compositional extrapolation of perturbation data. Specifically, they assume the following model:

$$\boldsymbol{z}^{\text{pert}} = \boldsymbol{z}^{\text{base}} + \boldsymbol{W}^{\text{pert}} \begin{pmatrix} h_1(l_1) \\ \dots \\ h_K(l_K) \end{pmatrix} + \sum_{j=1}^{J} \boldsymbol{W}_j^{\text{cov}} \boldsymbol{c}_j , \tag{C.12}$$

where $\boldsymbol{z}^{\text{base}} \sim \mathbb{P}_{\boldsymbol{Z}}$ denotes an unperturbed basal state; the matrix $\boldsymbol{W}^{\text{pert}} \in \mathbb{R}^{d_Z \times K}$ encodes the additive effect of each elementary perturbation; $\{h_k : \mathbb{R} \to \mathbb{R}\}_{k\in[K]}$ are unknown, possibly non-linear dose-response curves; $\{\boldsymbol{c}_j \in \mathbb{R}^{K_j}\}_{j\in[J]}$ are observed one-hot vectors capturing J additional discrete covariates, such as cell-types or species; and the matrices $\{\boldsymbol{W}_j^{\text{cov}} \in \mathbb{R}^{d_Z \times K_j}\}_{j\in[J]}$ encode additive covariate-specific effects. Further, the basal state $\boldsymbol{z}^{\text{base}}$ is assumed to be independent of the perturbation labels $\boldsymbol{l}$ and covariates $\boldsymbol{C} = (\boldsymbol{c}_1, \dots, \boldsymbol{c}_J)$. Observations $\boldsymbol{x}$ are then drawn from a Gaussian whose mean and variance are determined by the perturbed latent state $\boldsymbol{z}^{\text{pert}}$,

$$\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{\mu}\left(\boldsymbol{z}^{\text{pert}}\right), \sigma^2\left(\boldsymbol{z}^{\text{pert}}\right) \boldsymbol{I} \right) . \tag{C.13}$$

To fit this model, Lotfollahi et al. (2023) employ an adversarial autoencoder (Lample et al., 2017). First, an encoder $\boldsymbol{g}$ estimates the basal state

$$\widehat{\boldsymbol{z}}^{\text{base}} = \boldsymbol{g}(\boldsymbol{x})\,. \tag{C.14}$$

The estimated perturbed latent state $\widehat{\boldsymbol{z}}^{\text{pert}}$ is then computed according to (C.12) using (C.14) and learnt estimates $\widehat{\boldsymbol{W}}^{\text{pert}}$, $\{\widehat{h}_k\}$, and $\{\widehat{\boldsymbol{W}}_l^{\text{cov}}\}$. Finally, a (deterministic) decoder $\boldsymbol{f}$ uses $\widehat{\boldsymbol{z}}^{\text{pert}}$ to compute estimates of the mean and variance in (C.13), i.e., $(\widehat{\boldsymbol{\mu}}, \widehat{\sigma}^2) = \boldsymbol{f}(\widehat{\boldsymbol{z}}^{\text{pert}})$. All learnable components of the model are trained by minimizing the (Gaussian) negative log-likelihood of the observed data $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{l}_i, \boldsymbol{C}_i)\}_{i\in[N]}$. To encourage the postulated independence of $\widehat{\boldsymbol{z}}^{\text{base}}$ and $(\boldsymbol{l}, \boldsymbol{C})$, an additional adversarial loss is used, which minimizes the predictability of the latter from the former.

D Relation to Causal Models

In the field of causal inference, experimental data resulting from perturbations is typically modelled via *interventions* in an underlying causal model. In the *structural causal model* (SCM) framework of Pearl (2009), interventions modify a set of assignments, which determine each variable as a function of its direct causes and unexplained noise.

Definition D.1 (Acyclic SCM)**.** An acyclic structural causal model (SCM) $\mathcal{M} = (\mathcal{S}, \mathbb{P}_{\boldsymbol{U}})$ with *endogenous* variables $\boldsymbol{V} = \{V_1, ..., V_n\}$ and *exogenous* variables $\boldsymbol{U} = \{U_1, ..., U_m\}$ consists of:

(i) a set of *structural equations*

$$\mathcal{S} = \left\{ V_i := f_i\left(\boldsymbol{V}_{\text{pa}(i)}, \boldsymbol{U}_i\right) \right\}_{i=1}^{n}, \tag{D.1}$$

where f_i are deterministic functions; $\boldsymbol{U}_i \subseteq \boldsymbol{U}$; and $\boldsymbol{V}_{\text{pa}(i)} \subseteq \boldsymbol{V} \setminus \{V_i\}$ is the set of *causal parents* of V_i, such that the *induced causal graph* with vertices $[n]$ and edges $j \to i$ iff. $j \in \text{pa}(i)$ is acyclic;

(ii) a joint distribution $\mathbb{P}_{\boldsymbol{U}}$ over the exogenous variables.

The *induced distribution* $\mathbb{P}_{\boldsymbol{V}}$ of $\mathcal{M}$ is given by the push-forward of $\mathbb{P}_{\boldsymbol{U}}$ via $\mathcal{S}$. An intervention replaces $\mathcal{S}$ by a new set of assignments $\mathcal{S}'$ such the graph induced by $\mathcal{S}'$ is acyclic. The *interventional distribution* is the induced distribution $\mathcal{M}' = (\mathcal{S}', \mathbb{P}_{\boldsymbol{U}})$.[4]

Our assumed generative process for the effect of perturbations (§ 3) is, in general, different from interventions in an SCM. However, as we show next, certain types of SCMs and interventions are recovered as a special case of our model.

D.1 Shift Interventions in Linear SCMs as a Special Case

Consider a linear SCM with d_Z endogenous variables $\boldsymbol{Z}$ and d_Z exogenous variables $\boldsymbol{U}$ of the form

$$\boldsymbol{Z} := \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{U}, \tag{D.2}$$

where $\boldsymbol{A}$ is a (lower-triangular) weighted adjacency matrix. The observational (i.e., non-intervened) distribution of $\boldsymbol{Z}$ induced by (D.2) is most easily understood via the *reduced form* expression

$$\boldsymbol{Z} = (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{U} \tag{D.3}$$

and thus given by $\mathbb{P}_{\boldsymbol{Z}} = (\boldsymbol{I} - \boldsymbol{A})^{-1}_{\#}\mathbb{P}_{\boldsymbol{U}}$. Now consider the class of *shift interventions* parametrised by constant shift vectors $\boldsymbol{c}_e \in \mathbb{R}^{d_Z}$, which modify the original SCM in (D.2) to

$$\boldsymbol{Z}_e := \boldsymbol{A}\boldsymbol{Z}_e + \boldsymbol{U} + \boldsymbol{c}_e. \tag{D.4}$$

Analogous to (D.3), the reduced form of (D.4) is given by

$$\begin{aligned} \boldsymbol{Z}_e &= (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{U} \quad &+(\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{c}_e \\ &= \quad \boldsymbol{Z} \quad &+(\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{c}_e, \end{aligned} \tag{D.5}$$

[4] Interventions can also introduce new sources of randomness $\boldsymbol{U}'$, which, for sake of simplicity, we do not consider here.

where the second equality follows from (D.3).

Thus, if we take the shift vectors as perturbation labels (i.e., $K = d_Z$ and $\boldsymbol{l}_e = \boldsymbol{c}_e$) and use a linear perturbation model of the form

$$\boldsymbol{Z}_e^{\text{pert}} = \boldsymbol{\phi}(\boldsymbol{Z}_e^{\text{base}}, \boldsymbol{l}_e) = \boldsymbol{Z}_e^{\text{base}} + \boldsymbol{W}\boldsymbol{l}_e \tag{D.6}$$

with $\boldsymbol{W} = (\boldsymbol{I} - \boldsymbol{A})^{-1}$, then our perturbation model captures shift interventions in a linear SCM with adjacency matrix $\boldsymbol{A}$.

Remark D.2. The above argument does not require causal sufficiency: it still holds if $\mathbb{P}_{\boldsymbol{U}}$ is not factorized (e.g., due to hidden confounding).

Remark D.3. When trained on data generated according to (D.4), the adjacency matrix associated with the learnt perturbation matrix $\widehat{\boldsymbol{W}}$ is given by $\widehat{\boldsymbol{A}} = \boldsymbol{I} - \widehat{\boldsymbol{W}}^{-1}$.

Remark D.4. The correspondence between mean shift perturbations and shift interventions appears to only hold for *linear* SCMs. Consider, for example, a nonlinear additive noise model,

$$\boldsymbol{Z} := \boldsymbol{f}(\boldsymbol{Z}) + \boldsymbol{U}, \tag{D.7}$$

with reduced form given by

$$\boldsymbol{Z} = \boldsymbol{g}(\boldsymbol{U}), \tag{D.8}$$

where $\boldsymbol{g}$ is the inverse of the mapping $\boldsymbol{z} \mapsto \boldsymbol{z} - \boldsymbol{f}(\boldsymbol{z})$. For shift interventions in (D.7) to match our perturbation model, we then must have

$$\boldsymbol{Z}_e = \boldsymbol{g}(\boldsymbol{U} + \boldsymbol{c}_e) = \boldsymbol{g}\left(\boldsymbol{g}^{-1}(\boldsymbol{Z}) + \boldsymbol{c}_e\right) = \boldsymbol{\phi}\left(\boldsymbol{Z}, \boldsymbol{l}_e\right) \tag{D.9}$$

for suitable $\boldsymbol{\phi}$ and $\boldsymbol{l}_e$. Thus, if $\boldsymbol{f}$ is nonlinear, so is $\boldsymbol{g}$ and therefore $\boldsymbol{\phi}$. In other words, shift interventions in a nonlinear SCM do not, in general, amount to mean shift perturbations (i.e., linear $\boldsymbol{\phi}$).