What Makes a Representation Good for Single-Cell Perturbation Prediction?

Wenkang Jiang [1] Yuhang Liu[1,2] Yichao Cai [1] Erdun Gao [1,2] Jiayi Dong[3] Ehsan Abbasnejad [4] Lina Yao [5,6] Javen Shi [1,2]

[1]Australian Institute for Machine Learning, Adelaide University [2]Responsible AI Research Centre [3]Fudan University
[4]Monash University [5]University of New South Wales [6]Data61, CSIRO

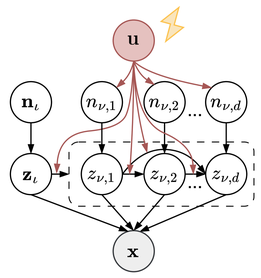
What Makes a Perturbation Representation Good?

Dominant invariant variation is a nuisance, not the target. It dominates reconstruction and representation learning, but does not imply biological priority.

The biological target is sparse perturbation response. The key variables are perturbation-responsive causal factors $\mathbf{z}_\nu$, induced by targeted genetic interventions.

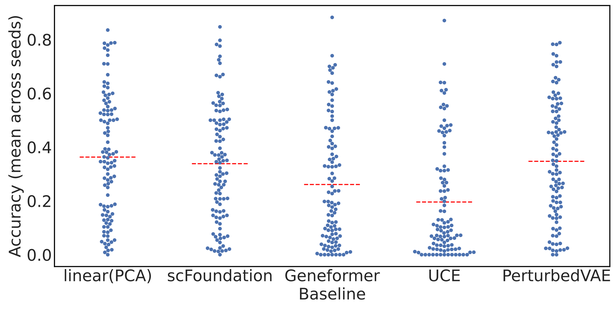
Our criterion. A good representation should be both identifiable and contrastively separated:

$$\text{Good representation} = \text{Identifiable } \mathbf{z}_\nu + \text{Contrastive block identifiable } \mathbf{z}_\iota.$$



PerturbedVAE. Contrastive alignment anchors invariant nuisance variation in $\mathbf{z}_\iota$, forcing perturbation-induced effects into $\mathbf{z}_\nu$, where they can be causally organized and identified.

Perturbation Suppression Hypothesis

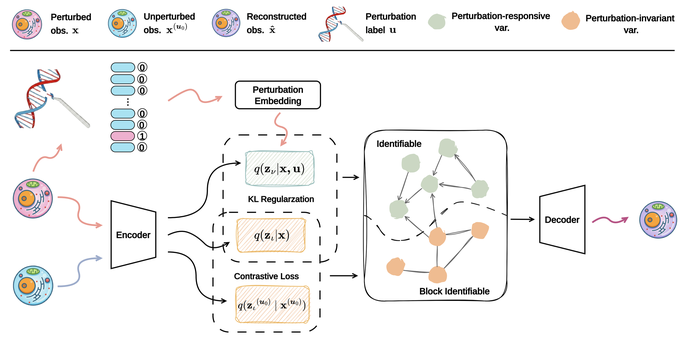


Diagnostic. Perturbation labels are less accessible in generic representations, suggesting that sparse perturbation-induced signals can be suppressed by dominant invariant variation.

Hypothesis. When perturbation-invariant variation dominates, representations that do not explicitly separate invariant and perturbation-specific information tend to fail systematically: they either entangle nuisance background with perturbation-related factors, or suppress sparse perturbation-induced signals. Importantly, dominant background variation is not the biological target; it is a disentanglement obstacle.

Implication. A good perturbation representation should isolate invariant nuisance variation and identify perturbation-responsive causal variables.

PerturbedVAE Framework



Framework. Perturbed cells learn the perturbation-responsive block $\mathbf{z}_\nu$, while unperturbed controls align the invariant nuisance block $\mathbf{z}_\iota$.

$$\mathbf{z} = (\mathbf{z}_\iota, \mathbf{z}_\nu), \qquad \mathbf{z}_\iota : \text{ invariant nuisance}, \quad \mathbf{z}_\nu : \text{ perturbation-response target}.$$

Good representation $\Longleftrightarrow$ identifiable $\mathbf{z}_\nu$ + contrastive separation from $\mathbf{z}_\iota$.

Identifiability-Guided Objective

Recovering latent variables from $(\mathbf{x}, \mathbf{u})$ alone is generally not identifiable. PerturbedVAE imposes a structured latent causal model:

$$\begin{aligned}
&\mathbf{z}_\iota := \boldsymbol{\lambda}_\iota \mathbf{z}_\iota + \mathbf{n}_\iota, \qquad \mathbf{n}_\iota \perp \mathbf{u}, \\
&\mathbf{z}_\nu := \boldsymbol{\lambda}_{\nu\iota}(\mathbf{u})\mathbf{z}_\iota + \boldsymbol{\lambda}_{\nu\nu}(\mathbf{u})\mathbf{z}_\nu + \mathbf{n}_\nu, \\
&\mathbf{n}_\nu \sim \mathcal{N}(\boldsymbol{\mu}_\nu(\mathbf{u}), \operatorname{diag} \boldsymbol{\beta}_\nu(\mathbf{u}))\,, \qquad \mathbf{x} := g(\mathbf{z}).
\end{aligned}$$

Here $\mathbf{z}_\iota$ captures invariant nuisance variation, while $\mathbf{z}_\nu$ captures perturbation-responsive causal variables whose mechanisms vary across environments.

Identifiability result.

$$\boxed{\mathbf{z}_\nu = \mathbf{P}_\nu \hat{\mathbf{z}}_\nu + \mathbf{c}_\nu} \qquad \boxed{\mathbf{z}_\iota = \mathbf{A}_\iota \hat{\mathbf{z}}_\iota + \mathbf{c}_\iota}$$

$$\mathbf{P}_\nu : \text{ permutation–scaling}, \qquad \mathbf{A}_\iota : \text{ invertible block transform}.$$

Thus, $\mathbf{z}_\nu$ is component-wise identifiable, while $\mathbf{z}_\iota$ is only block-identifiable.

Contrastive realization.

$$\boxed{\mathcal{L} = -\mathcal{L}_{\text{ELBO}} + \alpha \left\| \mathbf{z}_\iota^{(\mathbf{u})} - \mathbf{z}_\iota^{(\mathbf{u}_0)} \right\|_2^2}$$

Alignment prevents perturbation-dependent information from being explained by $\mathbf{z}_\iota$, forcing sparse perturbation responses into $\mathbf{z}_\nu$.

Double-Gene OOD Prediction

We evaluate zero-shot combinatorial prediction across different latent dimensions. Lower RMSE indicates better generalization.

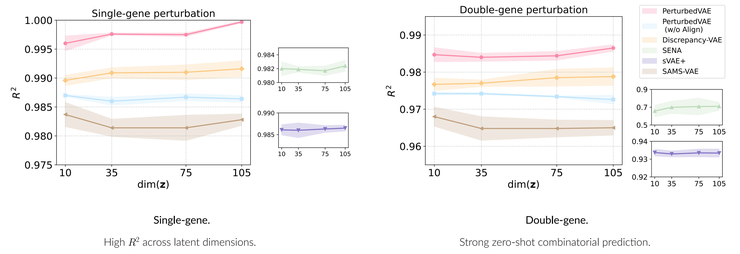
RMSE on double-gene perturbation prediction

Method	10	35	75	105
Discrepancy-VAE	$0.6084_{\pm 0.0045}$	$0.6037_{\pm 0.0025}$	$0.6075_{\pm 0.0072}$	$0.6082_{\pm 0.0045}$
SENA	$0.8573_{\pm 0.0205}$	$0.8514_{\pm 0.0248}$	$0.8507_{\pm 0.0396}$	$0.8483_{\pm 0.0248}$
sVAE+	$0.5663_{\pm 0.0009}$	$0.5667_{\pm 0.0008}$	$0.5665_{\pm 0.0011}$	$0.5664_{\pm 0.0012}$
SAMS-VAE	$0.4605_{\pm 0.0020}$	$0.4631_{\pm 0.0024}$	$0.4632_{\pm 0.0017}$	$0.4629_{\pm 0.0014}$
PerturbedVAE w/o Align	$0.4557_{\pm 0.0005}$	$0.4563_{\pm 0.0005}$	$0.4577_{\pm 0.0005}$	$0.4623_{\pm 0.0041}$
PerturbedVAE	$\mathbf{0.4493}_{\pm 0.0019}$	$\mathbf{0.4494}_{\pm 0.0008}$	$\mathbf{0.4489}_{\pm 0.0009}$	$\mathbf{0.4474}_{\pm 0.0007}$

PerturbedVAE consistently achieves the lowest RMSE across all latent dimensions. Removing alignment degrades double-gene OOD performance, supporting the role of contrastive separation in identifying $\mathbf{z}_\nu$.

Robustness across Latent Dimensions

We further compare R^2 across different latent dimensions for both single-gene and double-gene perturbation prediction.



Single-gene. High R^2 across latent dimensions.

Double-gene. Strong zero-shot combinatorial prediction.

Contributions & Scope

Contributions.

- We identify **perturbation suppression** as a key obstacle: sparse perturbation signals can be masked by invariant nuisance variation.
- We define a good perturbation representation as
  identifiable $\mathbf{z}_\nu$ + contrastive separation from $\mathbf{z}_\iota$.
- We propose **PerturbedVAE**, combining contrastive alignment with an environment-conditioned latent causal model for OOD perturbation prediction.

Scope. The identifiability guarantee relies on structured assumptions, including sufficient environment variation, optimal alignment, and an invertible generative map. The learned causal graph provides biologically plausible evidence, but is not a complete biological validation.

Correspondence to yuhang.liu01@adelaide.edu.au