# WHAT DO SINGLE-CELL MODELS ALREADY KNOW ABOUT PERTURBATIONS?

**Andreas Bjerregaard**[a,b*] **Vivek Das**[c] **Anders Krogh**[a,b]

[a]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
[b]Center for Health Data Science, University of Copenhagen, Copenhagen, Denmark
[c]Integrated Omics, AI and Analytics, Development, Novo Nordisk A/S, Søborg, Denmark

## ABSTRACT

Generative models implicitly learn underlying dynamics of data and can do more than just reconstruction. By leveraging output gradients with respect to the latent dimensions, we explore a simple approach to infer arbitrary perturbation effects which generates interpretive flow maps within high-dimensional biological datasets. By applying this method to several cases in single-cell RNA-sequencing, we demonstrate its use in inferring effects from knockdown, overexpression, toxin response and embryonic development. This approach can further add global structure to dimensionality reductions which normally only preserve local patterns. Needing only a decoder, our method simplifies analyses, is applicable to already trained models, and offers clearer insights into cellular dynamics without complex setups. In turn, this gives a more straightforward interpretation of results, making it easier to discern underlying biological pathways with easily understandable visual representations. Code available on `https://github.com/yhsure/perturbations`.

## 1 BACKGROUND

Modeling perturbation response in cells is essential for understanding gene function, regulatory effects, and drug response. Single-cell RNA sequencing (scRNA-seq) provides high-resolution snapshots of cellular states and captures implicit gene-gene interactions. Although knockout experiments coupled with scRNA-seq can reveal gene function, these experiments are prone to biases (Hicks et al., 2015) and costly across multiple conditions. Computational approaches can simulate perturbations, offering a scalable alternative for systematic analyses.

Here we study single-cell data to understand arbitrary perturbations for individual cells as well as populations, providing practical clues for regulatory mechanisms and gene function. Generative models are typically used for modeling this type of data, resulting in a latent space with a structure reflecting cell type differences (Lopez et al., 2020). Implicitly, such a model may learn underlying biological interactions between genes, cellular trajectories during development or disease, and responses to unseen perturbations. Recent works have introduced several ways of thinking about perturbations (Lotfollahi et al., 2019; 2020; Kamimoto et al., 2023; Bunne et al., 2023; Jiang et al., 2024; Klein et al., 2025), most commonly in a supervised fashion — we take a step backwards to explore what generative models have already learned.

We study a simple method for simulating perturbations based on decoder gradients as a first step of learning more from such generative models, and apply it to three diverse perturbation scenarios.

## 2 METHODS

### 2.1 DATA

The data utilized for this study comprises three single-cell RNA sequencing (scRNA-seq) datasets: *C. elegans* embryogenesis (Packer et al., 2019), *Irf8*-cKO mouse brains (Van Hove et al., 2019), and

[*] Correspondence to: Andreas Bjerregaard <anje@di.ku.dk>

cardiotoxin mouse injury (Takada et al., 2022). Preprocessing and filtering protocols are consistent across datasets and are detailed in Appendix A. Datasets were partitioned into 82% for training, 9% for validation, and 9% for testing. An overview of the used data is provided in Table 1.

| Dataset | Cells | Transcripts | HVG | Reference |
|---|---|---|---|---|
| *Irf8*-cKO mouse brains | $13,931$ | $14,581$ | $3,451$ | Van Hove et al. (2019) |
| Cardiotoxin mouse injury | $53,230$ | $21,809$ | $1,950$ | Takada et al. (2022) |
| C. elegans embryogenesis | $85,951$ | $17,711$ | $1,832$ | Packer et al. (2019) |

Table 1: Overview of scRNA-seq datasets applied in this study. HVG refers to highly variable genes.

### 2.2 Base model

Gene expression data was analyzed in an unsupervised approach with $\beta$-VAEs (Higgins et al., 2017). Using a negative binomial (NB) distribution, the posterior was modeled to account for overdispersion observed in such expression data (Robinson & Smyth, 2007; Oshlack et al., 2010; Grønbech et al., 2020). The NB is parameterized by the mean $m$ and the dispersion parameter $r$:

$$\mathrm{NB}(k; m, r) = \frac{\Gamma(k+r)}{k!\,\Gamma(r)} \left(\frac{m}{r+m}\right)^k \left(\frac{r}{r+m}\right)^r, \tag{1}$$

where $k$ is the observed count for a gene. The model outputs $m$ scaled by the mean count for the sample and $r$ is learned for each gene, i.e., sample independent.

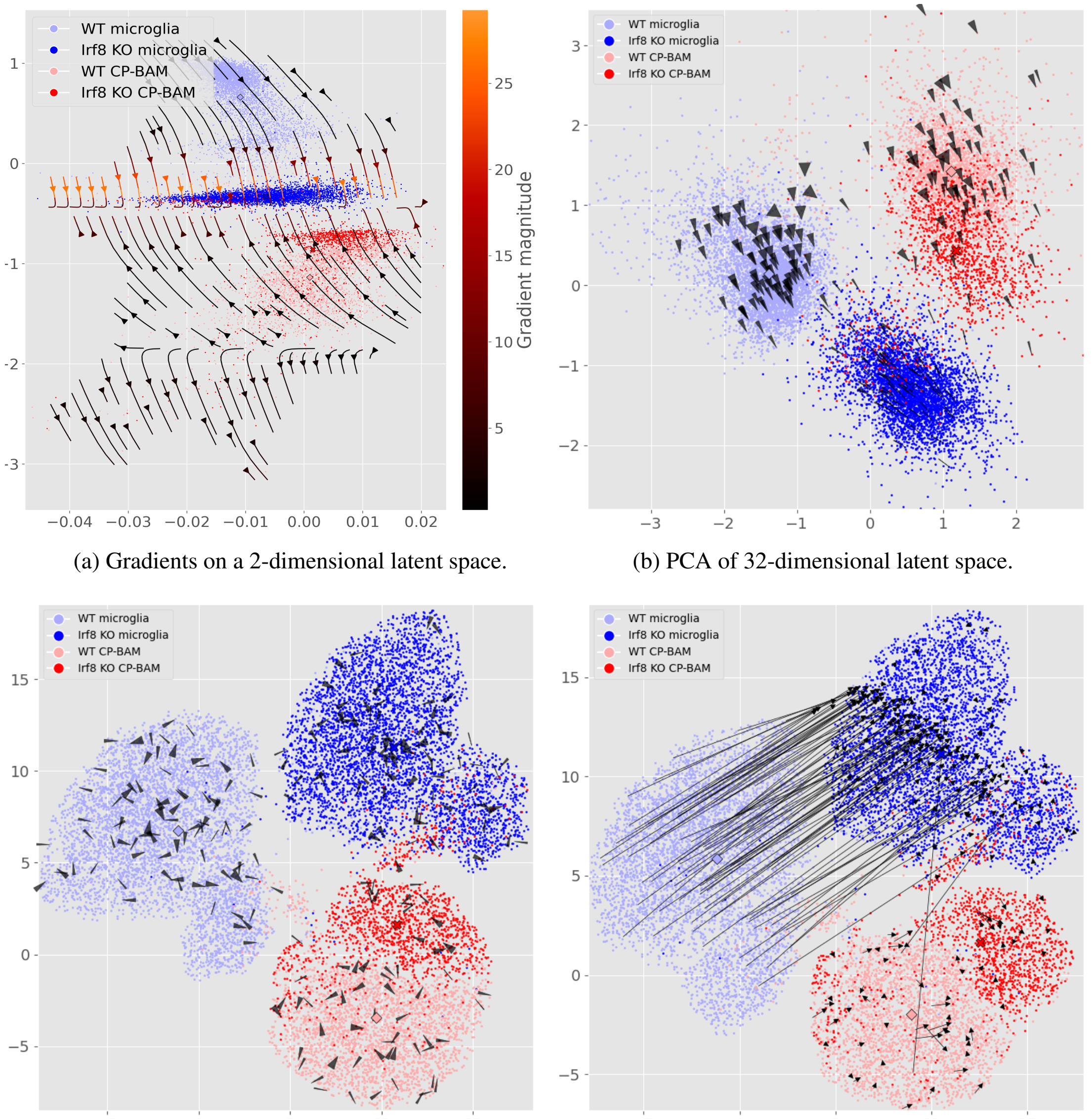
The training procedure maximizes the ELBO (with NB negative log likelihood as reconstruction error) and was implemented in Pytorch with early stopping to prevent overfitting. Compared to the NB approach, VAEs with other posteriors result in lower accuracy (Grønbech et al., 2020; Bjerregaard, 2023). The factor $\beta$ which scales the Kullback–Leibler divergence was annealed over 10 warmup epochs and stopped at a low $\beta$, resulting in almost an autoencoder. The encoder used ReLU activations, two hidden linear layers with sizes 512 and 256, and a latent dimensionality of 32 or 2. The $\beta$-VAE had a mirrored decoder and was trained on one dataset at a time with Adam (Kingma, 2014); refer to Appendix B. Note that the NB posterior accurately models raw count values for the decoder output but is non-trivial to design for the encoder, which in turn uses log-transformed mean-scaled counts. Pretrained models from hubs like `scvi-hub` (Ergen et al., 2024) are similar and can easily be utilized as base models.

### 2.3 Perturbation flows

The generative decoder learns how to interpret influences of gene expressions in relation to a latent cellular space. We simulate perturbations by following the gradient of gene expression from an initial latent representation. Specifically, for a latent sample $z_t$, the perturbed sample is given by $z_{t+1} = z_t + \delta \nabla y_i(z_t)$ where $\delta$ is the perturbation stepsize and $y_i(z)$ is the gene expression output of gene $i$. A negative $\delta$ thus simulates decreasing gene expression (knockdown), while a positive $\delta$ simulates overexpression.

Rather than selecting a specific starting cell, gradients are uniformly sampled across the latent space for visualization. Regions distant from training samples are masked away using morphological operations on a discretized grid.

Arbitrary perturbations can be constructed similarly by introducing an auxiliary output variable and a loss term in a multi-task setup. For treatment analysis, this could be a categorical or continuous variable indicating the treatment type or dosage. Existing models can be adapted either through finetuning or by adding a new linear layer. For higher dimensional latent spaces, subsampling existing data for starting points helps managing the exponential growth in latent volume. Dimensionality reduction is subsequently used to project samples and perturbation vectors. Here, PCA allows projection of the perturbation gradients directly while UMAP requires encoding perturbed endpoints into a new list, concatenating it to the data, and using this list to reconstruct the perturbation vector.



(a) Gradients on a 2-dimensional latent space.

(b) PCA of 32-dimensional latent space.

(c) UMAP of 32-dimensional latent space.

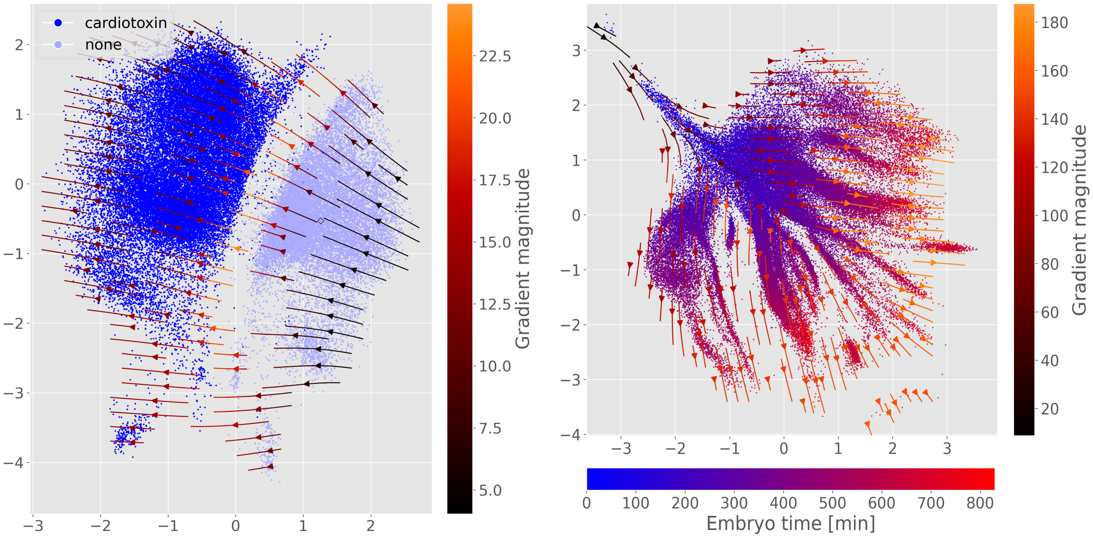
(d) UMAP when taking 400 gradient steps from sampled start points.

Figure 1: Perturbation flow maps for gene perturbations on the *Irf8*-cKO mouse brain dataset of Van Hove et al. (2019). Arrows illustrate directions of the negative mean gradient of a small subset of six genes inferred to be co-regulated with *Irf8*. Only 10% of cKO training samples were used.

## 3 Results

**Predicting knockout response** To evaluate the utility of the perturbation flows, a case study on the *Irf8*-cKO dataset (Van Hove et al., 2019) is performed. Visualizing the negative gradient of *Irf8*-expression in latent space shows the effect of gradual knockdown going from the wild type to the knockout population (Figure 1a), more evident for a higher dimensional latent space (Figure 1b). Similarly, effects of gene overexpression are successfully simulated (see Supplementary Figure S1).

**Predicting injury response** Next, we consider the dataset of cardiotoxin-induced mouse injury (Takada et al., 2022). A binary variable is added to the output features to indicate cardiotoxin injury. This output variable is included in the objective function with a scaled binary cross entropy loss term,



(a) Cardiotoxin-induced injury in mice.

(b) Time series data of *C. elegans* embryos.

Figure 2: Perturbation flow maps for more general perturbations. Gradients are computed from the output of cardiotoxin classification or embryo time regression.

$\alpha L_{\text{CTX}}$. Training with just 10% of injury labels, the model still achieves 99.0% accuracy in predicting the cardiotoxin label on the held out test set. Visualizing the gradient of the cardiotoxin prediction in latent space (Figure 2a) shows how toxin affects the latent samples, simulating changes in their expression profiles. As expected, perturbation vectors strictly go from wild type to experimentally perturbed samples.

**Predicting temporal dynamics** Dimensionality reduction on the *C. elegans* embryogenesis dataset was found to distinctly subcluster celltypes according to the age of the embryo sample (Packer et al., 2019). Adding this embryo time as a continuous output feature enables the inclusion of an additional L1 loss term $\alpha L_{\text{time}}$ in our objective function. Again training with just 10% of available time labels, Figure 2b shows that the gradient of time predictions can be used to infer how cells develop, and is well aligned with the observed sample times. The latent space further stays subdivided in distinct cell types (illustrated by Figure S2).

## 4 Discussion

Generative decoders can be queried to infer the effects of perturbations on gene expression. This is evidenced by the perturbation flow maps of Figures 1 and 2, demonstrating an intuitive and visual interpretation of these perturbation effects. For each dataset, the generative model converges with low reconstruction error (Supplementary Table S1). The decoder's knockdown predictions for the *Irf8*-cKO dataset align with findings by Van Hove et al. (2019) emphasizing *Irf8*'s significance in microglia. Flows point from WT samples to cKO samples when decreasing expression of *Irf8* (Figures 1a, 1b). Similarly, flows approximately reverse when considering the mean gradient of genes which are differentially overexpressed in the cKO set (Supplementary Figure S1). Further, the perturbation concept is easily generalized as demonstrated in the cases of cardiotoxin injury and embryonic development. In these contexts, the gradients highlight different cellular dynamics — e.g., transitions from control to cardiotoxin-altered states, and temporal patterns during development. Even small amounts of labeled data can be used to achieve a general understanding of the whole dataset — and thus simulate perturbation trajectories for new unlabeled cells. Differential expression analysis along the perturbation trajectory could lead to insights relevant for, e.g., drug design.

While more intuitive for 2-dimensional latent spaces, larger latent dimensionalities can be used. Figure 1b shows how a larger dimensionality could better encode the effect of *Irf8* for CP-BAM cells. Notably, the inferred perturbations also provide a means to visualize relationships between clusters in a UMAP-reduced space. As UMAP largely discards global stucture, this is an interesting direction for recovering cell dynamics between clusters.

This study shows that generative models can be utilized to a larger degree when considering, e.g., output gradients. Future work will compare the quality and fidelity of the in-silico results with other available tools, and explore the use of decoder gradients in inferring gene regulatory networks.

**Acknowledgements** The authors thank Viktoria Schuster, Iñigo Prada-Luengo, Yan Li, Adrián Sousa-Poza and Valentina Sora for helpful discussions. This work was supported by grants NNF20OC 0062606, NNF20OC0063268, and NNF20OC0059939 from the Novo Nordisk Foundation.

## REFERENCES

Andreas Bjerregaard. Save the mice: in-silico perturbation of genes in deep generative models. Master's thesis, University of Copenhagen, Copenhagen, Denmark, 5 2023.

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):1759–1768, 2023.

Can Ergen, Valeh Valiollah Pour Amiri, Martin Kim, Aaron Streets, Adam Gayoso, and Nir Yosef. scvi-hub: A flexible framework for reference enabled single-cell data analysis. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*, 2024.

Christopher Heje Grønbech, Maximillian Fornitz Vording, Pascal N Timshel, Casper Kaae Sønderby, Tune H Pers, and Ole Winther. scvae: variational auto-encoders for single-cell gene expression data. *Bioinformatics*, 36(16):4415–4422, 2020.

Stephanie C Hicks, Mingxiang Teng, Rafael A Irizarry, et al. On the widespread and critical impact of systematic bias and batch effects in single-cell rna-seq data. *BioRxiv*, 10:025528, 2015.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2017.

Qun Jiang, Shengquan Chen, Xiaoyang Chen, and Rui Jiang. scpram accurately predicts single-cell gene expression perturbation response based on attention mechanism. *Bioinformatics*, 40(5):btae265, 2024.

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.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Dominik Klein, Giovanni Palla, Marius Lange, Michal Klein, Zoe Piran, Manuel Gander, Laetitia Meng-Papaxanthos, Michael Sterr, Lama Saber, Changying Jing, et al. Mapping cells through time and space with moscot. *Nature*, pp. 1–11, 2025.

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

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

Mohammad Lotfollahi, Mohsen Naghipourfar, Fabian J Theis, and F Alexander Wolf. Conditional out-of-distribution generation for unpaired data using transfer vae. *Bioinformatics*, 36(Supplement_2):i610–i617, 2020.

Alicia Oshlack, Mark D Robinson, and Matthew D Young. From rna-seq reads to differential expression results. *Genome biology*, 11:1–10, 2010.

Jonathan S Packer, Qin Zhu, Chau Huynh, Priya Sivaramakrishnan, Elicia Preston, Hannah Dueck, Derek Stefanik, Kai Tan, Cole Trapnell, Junhyong Kim, et al. A lineage-resolved molecular atlas of c. elegans embryogenesis at single-cell resolution. *Science*, 365(6459):eaax1971, 2019.

Mark D Robinson and Gordon K Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.

Naoki Takada, Masaki Takasugi, Yoshiki Nonaka, Tomonori Kamiya, Kazuaki Takemura, Junko Satoh, Shinji Ito, Kosuke Fujimoto, Satoshi Uematsu, Kayo Yoshida, et al. Galectin-3 promotes the adipogenic differentiation of pdgfr$\alpha$+ cells and ectopic fat formation in regenerating muscle. *Development*, 149(3):dev199443, 2022.

Hannah Van Hove, Liesbet Martens, Isabelle Scheyltjens, Karen De Vlaminck, Ana Rita Pombo Antunes, Sofie De Prijck, Niels Vandamme, Sebastiaan De Schepper, Gert Van Isterdael, Charlotte L Scott, et al. A single-cell atlas of mouse brain macrophages reveals unique transcriptional identities shaped by ontogeny and tissue environment. *Nature neuroscience*, 22(6):1021–1035, 2019.

## A Filtering and Preprocessing

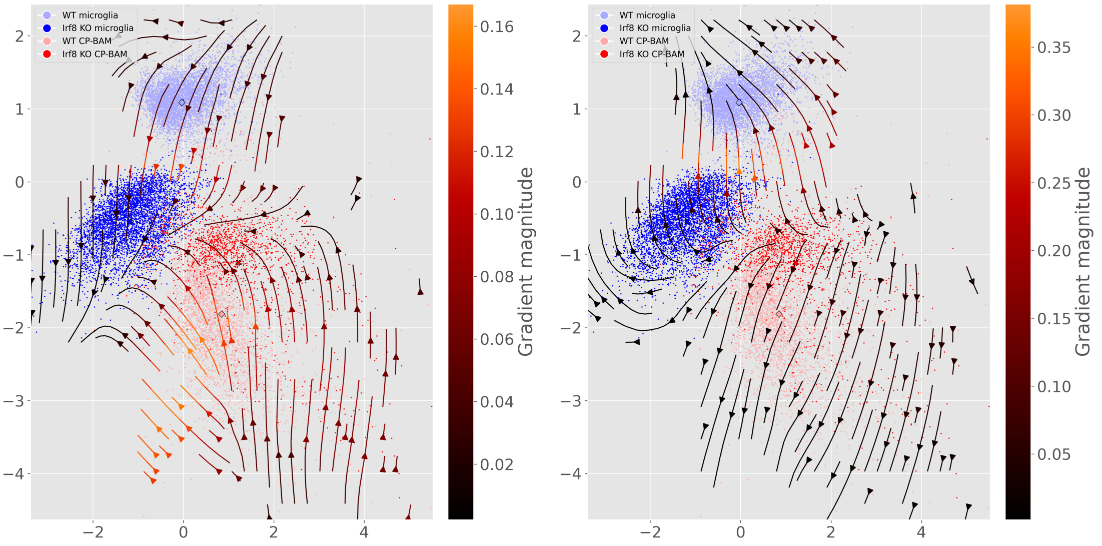
Uniformly across datasets, cells were retained only if they had 1) at least 200 non-zero genes 2) at least 500 counts across genes, and 3) less than 5000 non-zero genes. Transcripts expressed in fewer than 5 cells were removed. Raw counts were scaled by mean counts per cell and log-transformed to annotate highly variable genes with Scanpy, and then inversely transformed to recover the original raw counts. For the cKO mouse brains, the parameter `n_top_genes` for HVG selection was tuned via binary search to ensure inclusion of *Irf8*.

## B Supplementary Tables

| | **Train set** | | | | **Test set** | | | |
|---|---|---|---|---|---|---|---|---|
| **Instance** | **ARI** | **RMSE** | **MAE** | **Task** | **ARI** | **RMSE** | **MAE** | **Task** |
| *Irf8* cKO | $0.72 \pm 0.02$ | $0.19 \pm 0.00$ | $0.23 \pm 0.00$ | - | $0.74 \pm 0.02$ | $0.19 \pm 0.00$ | $0.23 \pm 0.00$ | - |
| 32D *Irf8* cKO | $0.87 \pm 0.02$ | $0.16 \pm 0.00$ | $0.21 \pm 0.00$ | - | $0.71 \pm 0.17$ | $0.17 \pm 0.00$ | $0.21 \pm 0.00$ | - |
| cardiotoxin | $0.77 \pm 0.05$ | $0.27 \pm 0.00$ | $0.31 \pm 0.00$ | $1.00 \pm 0.00$ | $0.76 \pm 0.05$ | $0.27 \pm 0.00$ | $0.31 \pm 0.00$ | $0.99 \pm 0.00$ |
| embryogenesis | $0.33 \pm 0.04$ | $0.31 \pm 0.01$ | $0.25 \pm 0.00$ | $11.37 \pm 3.81$ | $0.33 \pm 0.04$ | $0.31 \pm 0.01$ | $0.26 \pm 0.00$ | $32.04 \pm 1.75$ |

Table S1: Tabular overview of the trained models. Each row shows results based on 5 runs. Root mean squared error (RMSE) and mean absolute error (MAE) compare log-transformed model outputs to log-transformed mean-scaled counts. The task column denotes either decimal accuracy (cardiotoxin prediction task) or mean L1 norm (embryogenesis regression task). Adjusted rand index (ARI) is computed for sampling timepoint for the cardiotoxin dataset and for cell type for the other datasets, and relies on k-means clustering. The 32-dimensional latent variables of the high-dimensional model were reduced to 2D via UMAP before clustering.

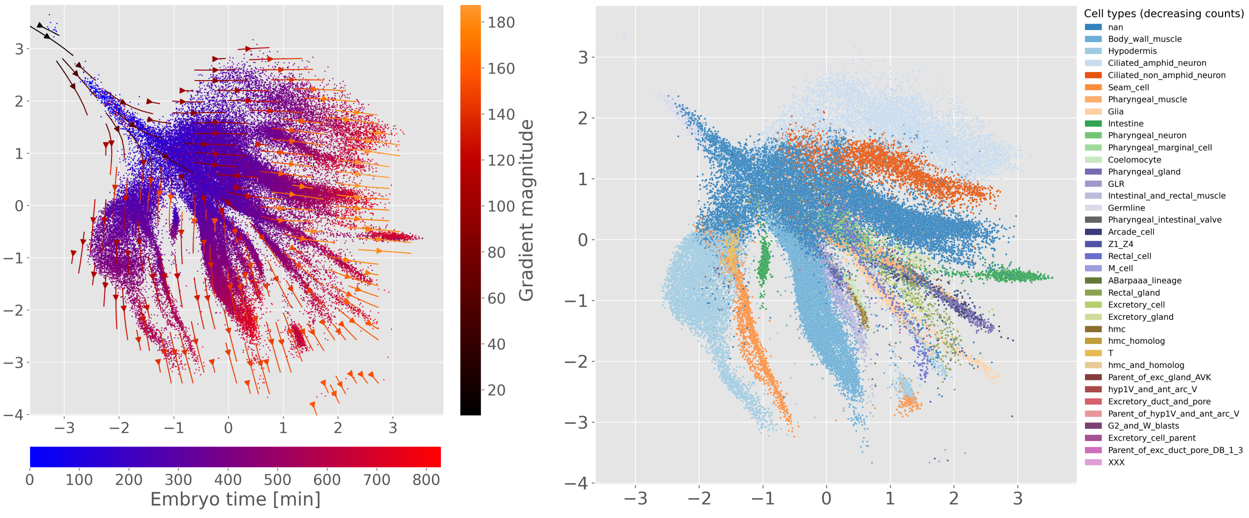
# C Supplementary Figures



(a) Mean negative gradient of differentially underexpressed genes.

(b) Mean negative gradient of differentially overexpressed genes.

Figure S1: Perturbation flow maps for gene perturbations on the *Irf8*-cKO dataset. Gradient is aggregated as the mean over genes which are differentially under- or overexpressed when comparing wild type and cKO populations.



(a) Re-print of Figure 2b for easier comparison.

(b) Cell type annotations for the *C. elegans* latent space.

Figure S2: Latent space representations for the *C. elegans* embryogenesis dataset. Latent variables are shown with two different labeling schemes: embryo measurement time and cell type. This illustrates how the model captures continuous development while maintaining distinct cellular populations.