

Predicting single-cell perturbation responses with DL

Mehrshad Sadria

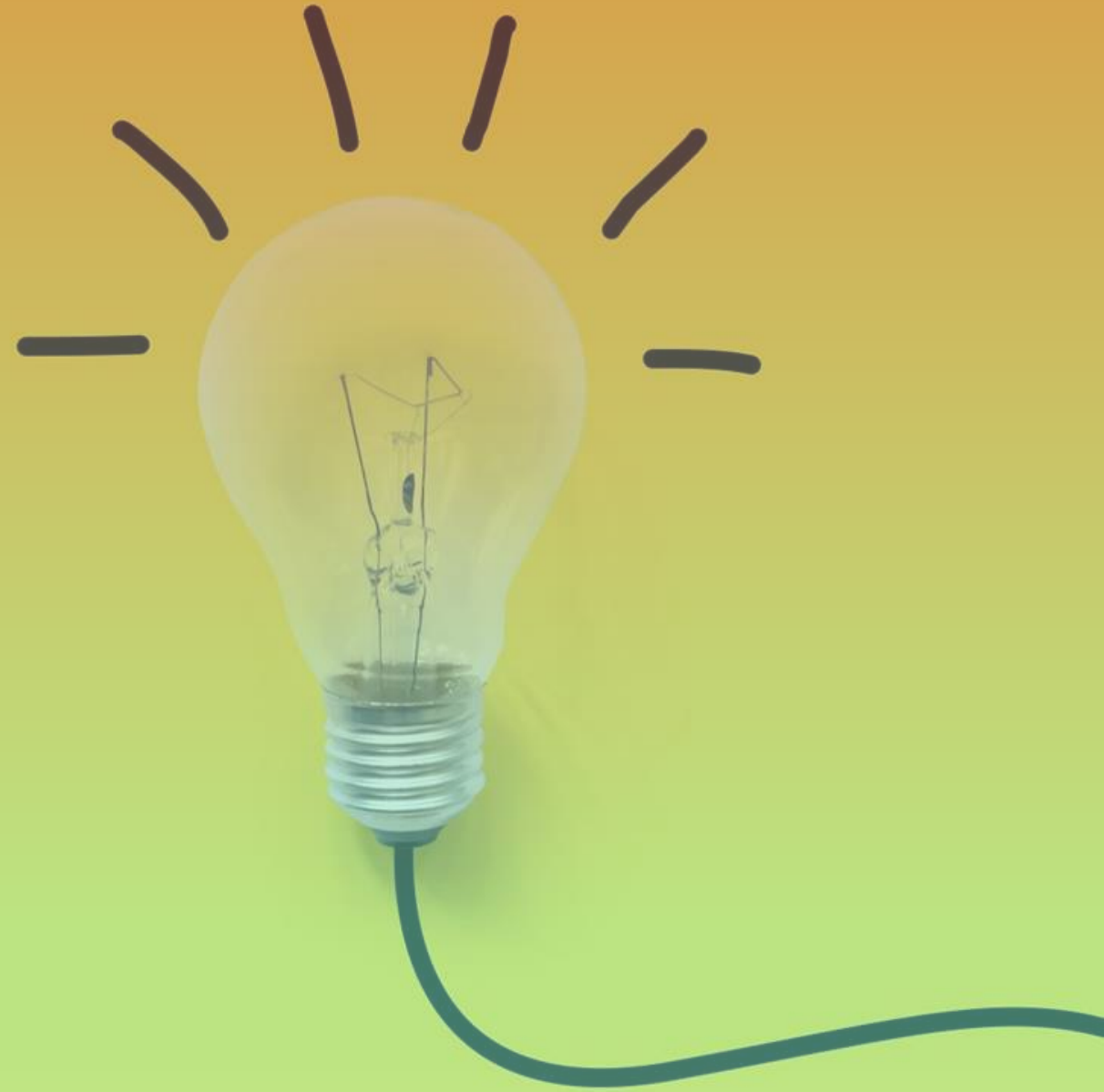
Deep learning for biotechnology

Spring 2022

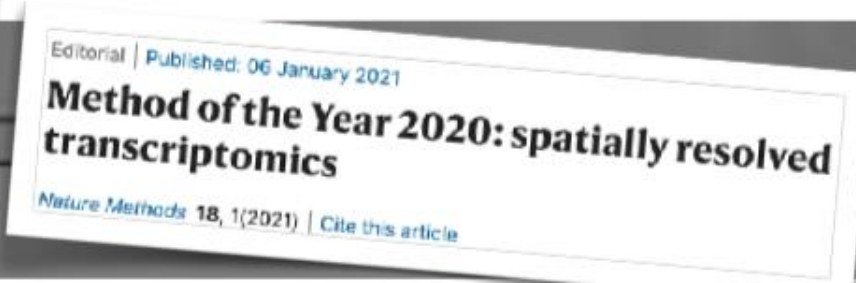
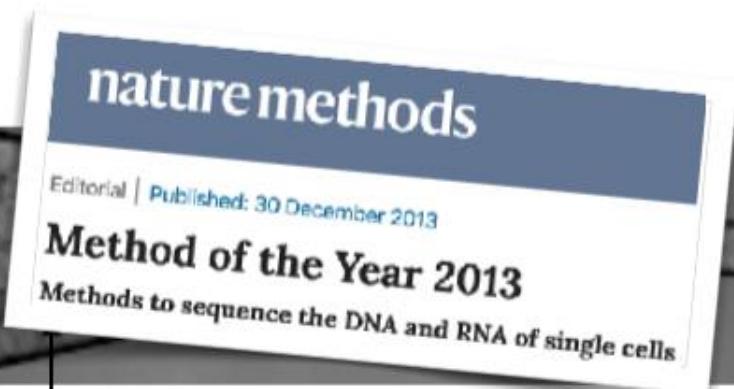
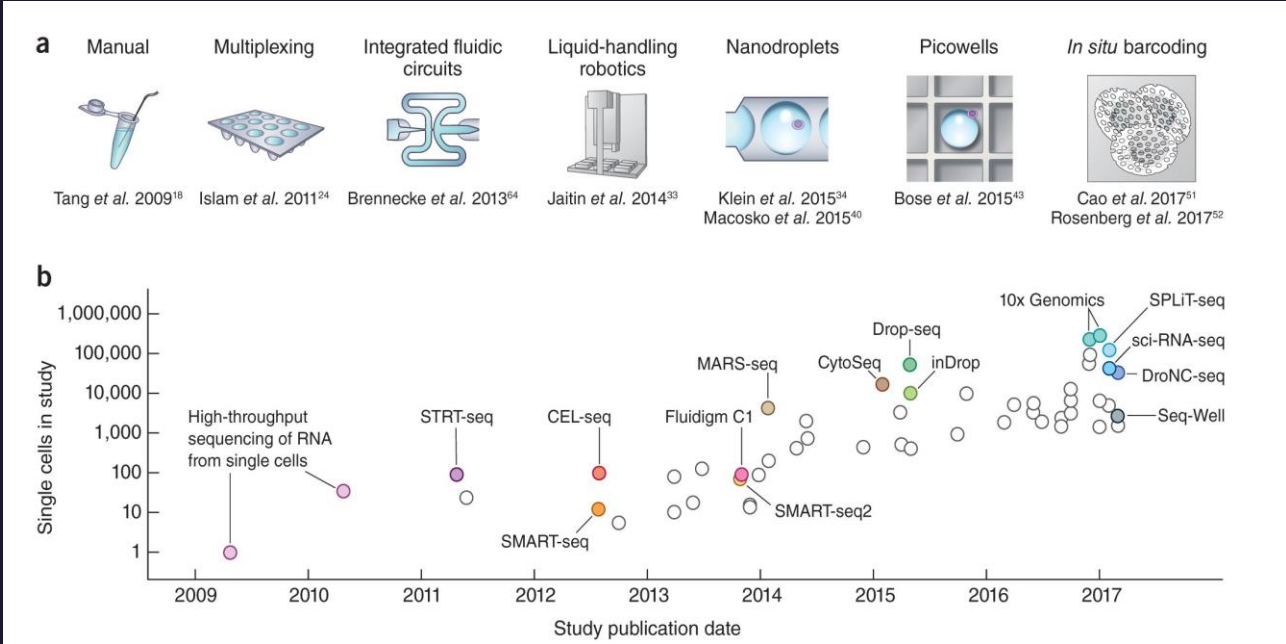


Overview

- Background & Introduction
- Method
- Results and performance
- Conclusion
- References

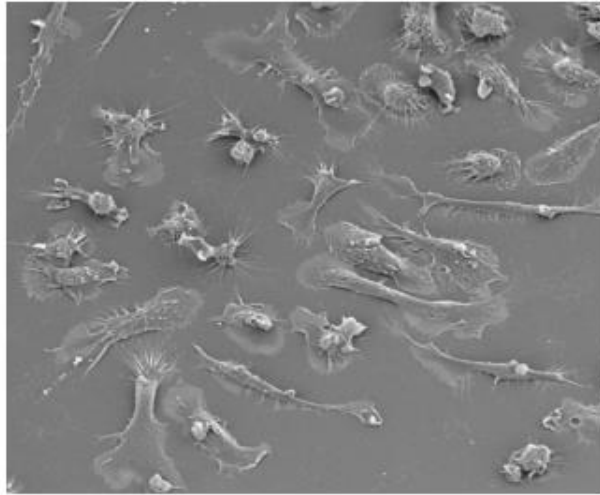


What is single cell technology?

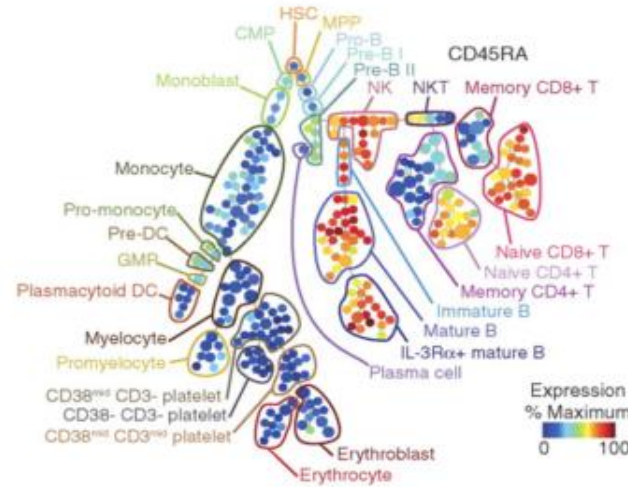


Why single cells

Cellular heterogeneity

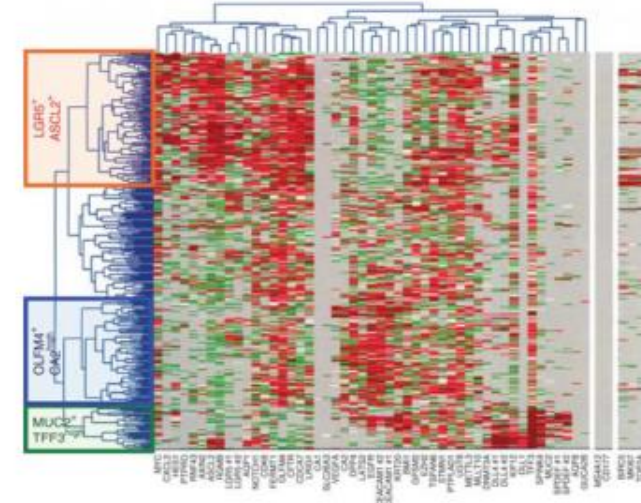


Differentiation trajectories



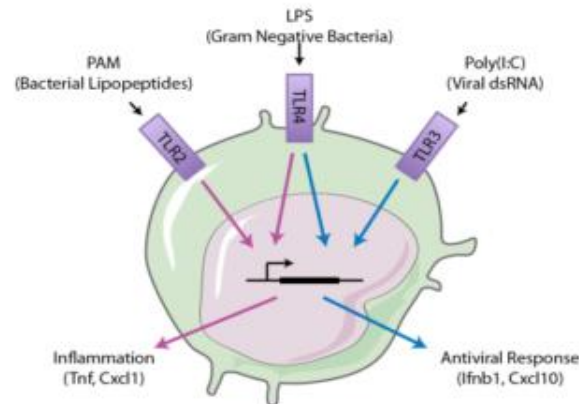
Bendall et al. (2011), Science

Within-cell-type differences

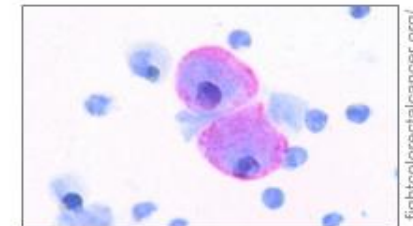
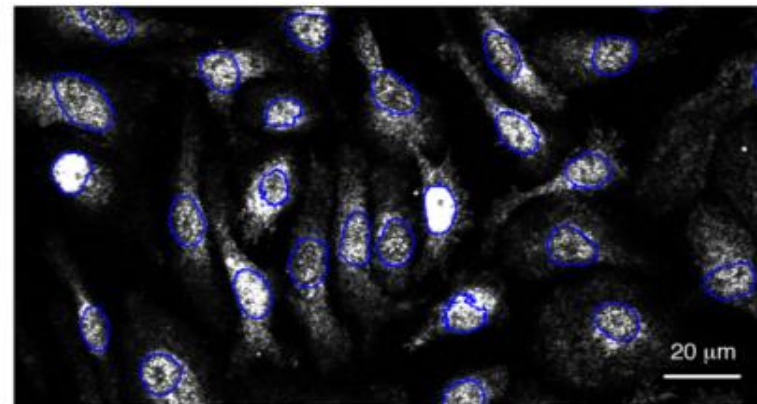


Dalerba et al. (2011), Nature Biotech

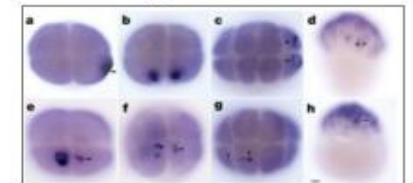
TLR Signaling



IRF3 Protein Levels - 4h LPS



Circulating Tumor Cells

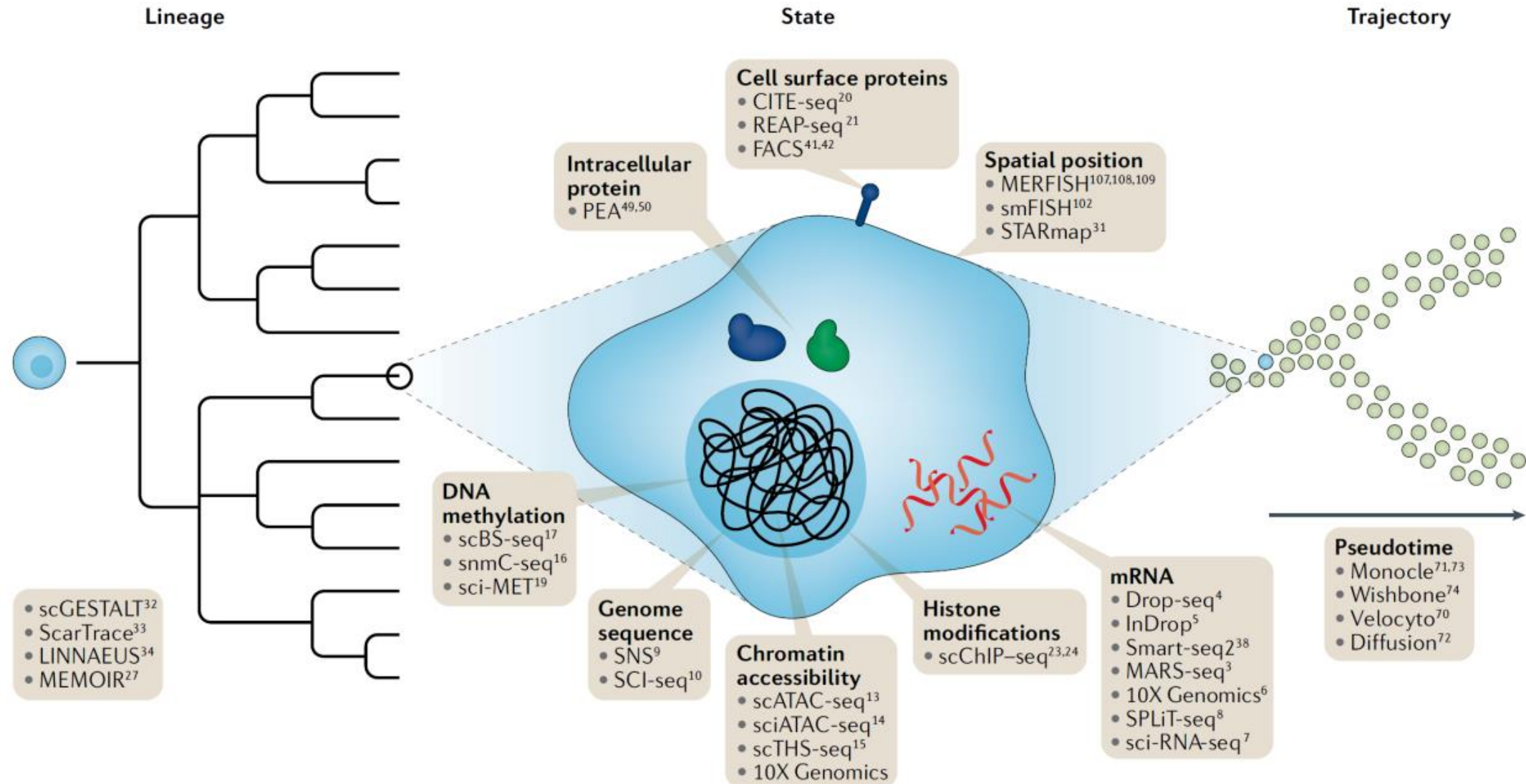


Zebrafish early embryo

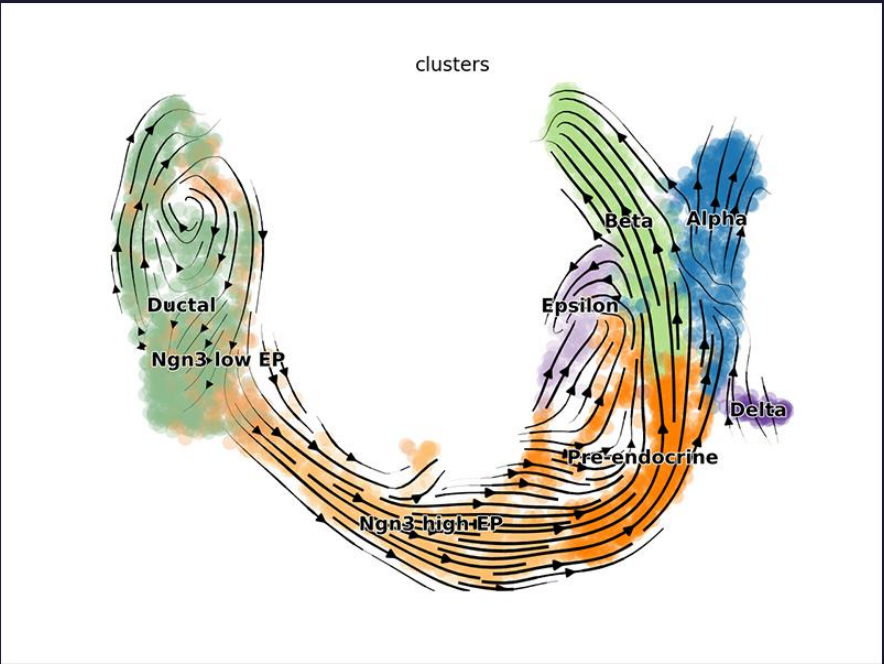
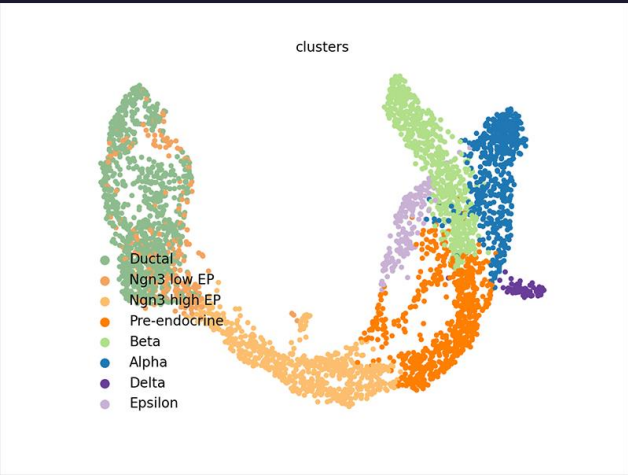
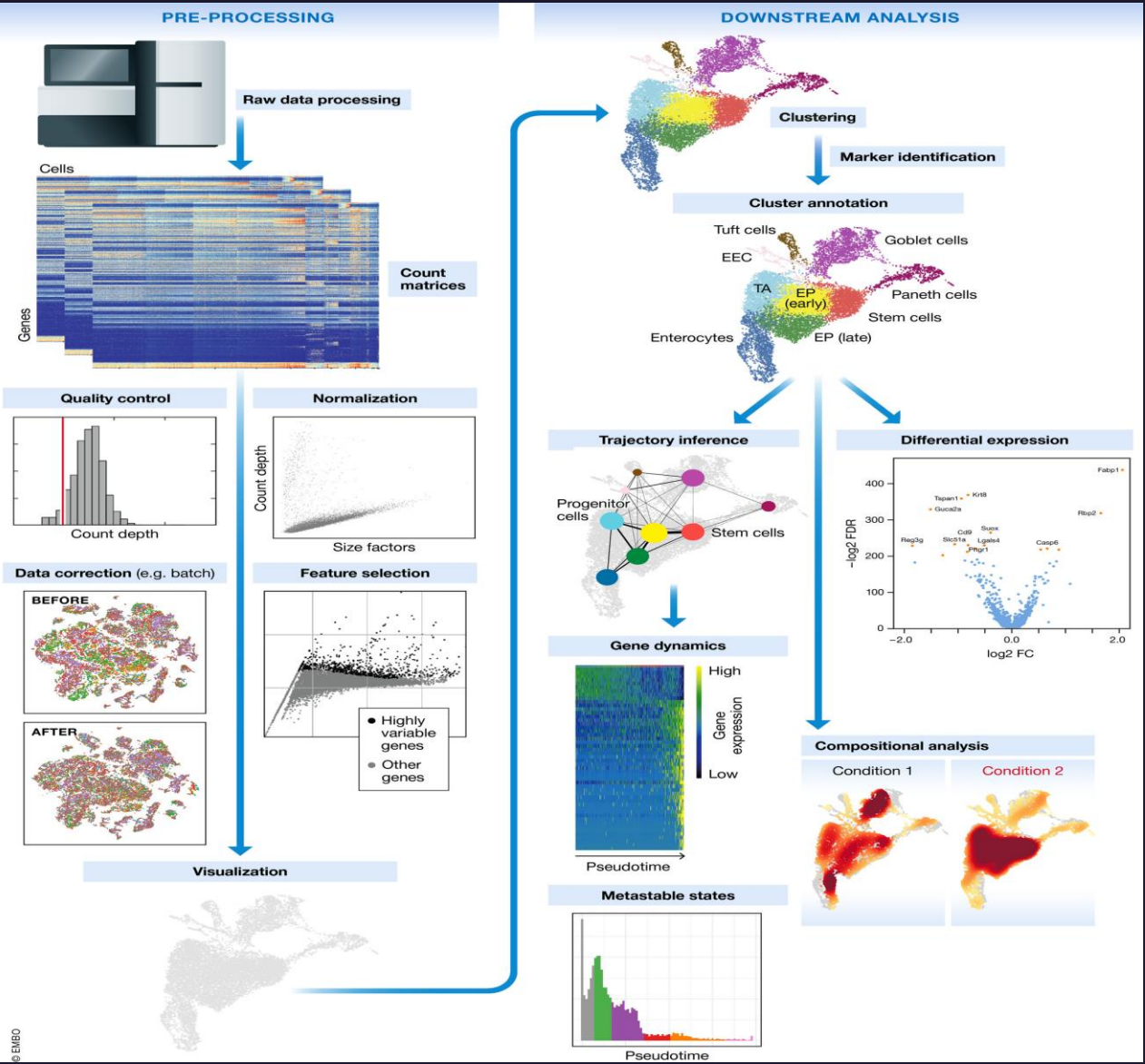
Cellular responses can vary substantially between “identical” cells.

Overcome low input

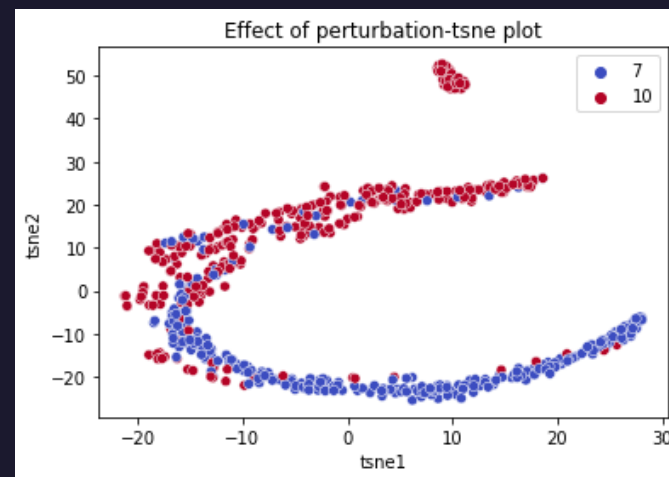
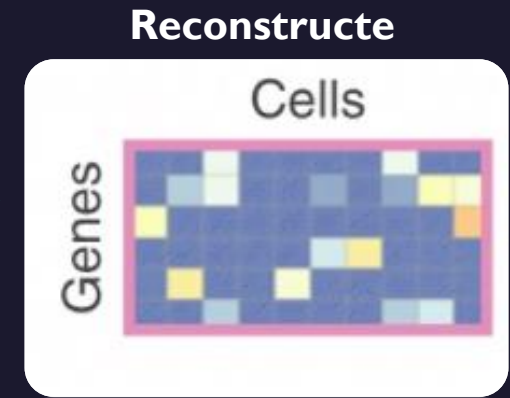
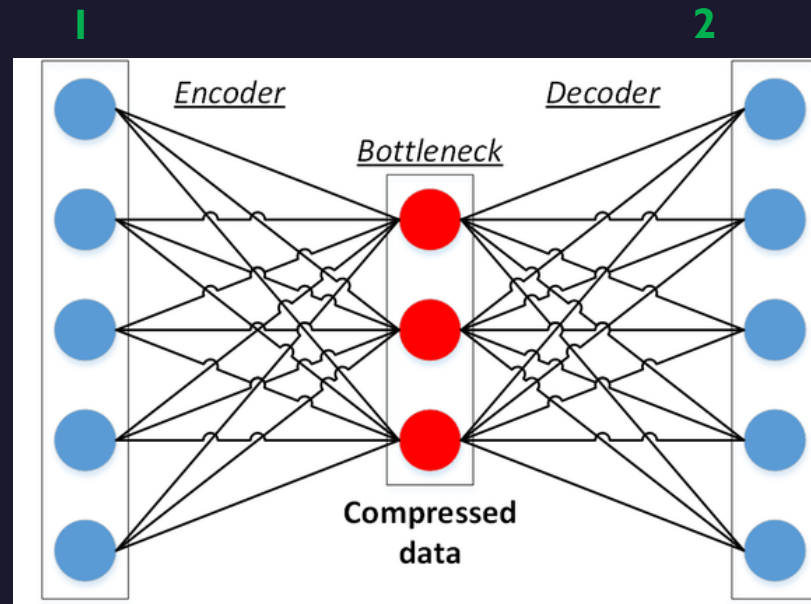
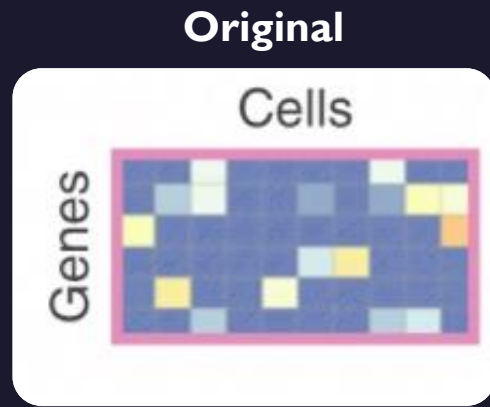
Diverse technologies for sc profiling



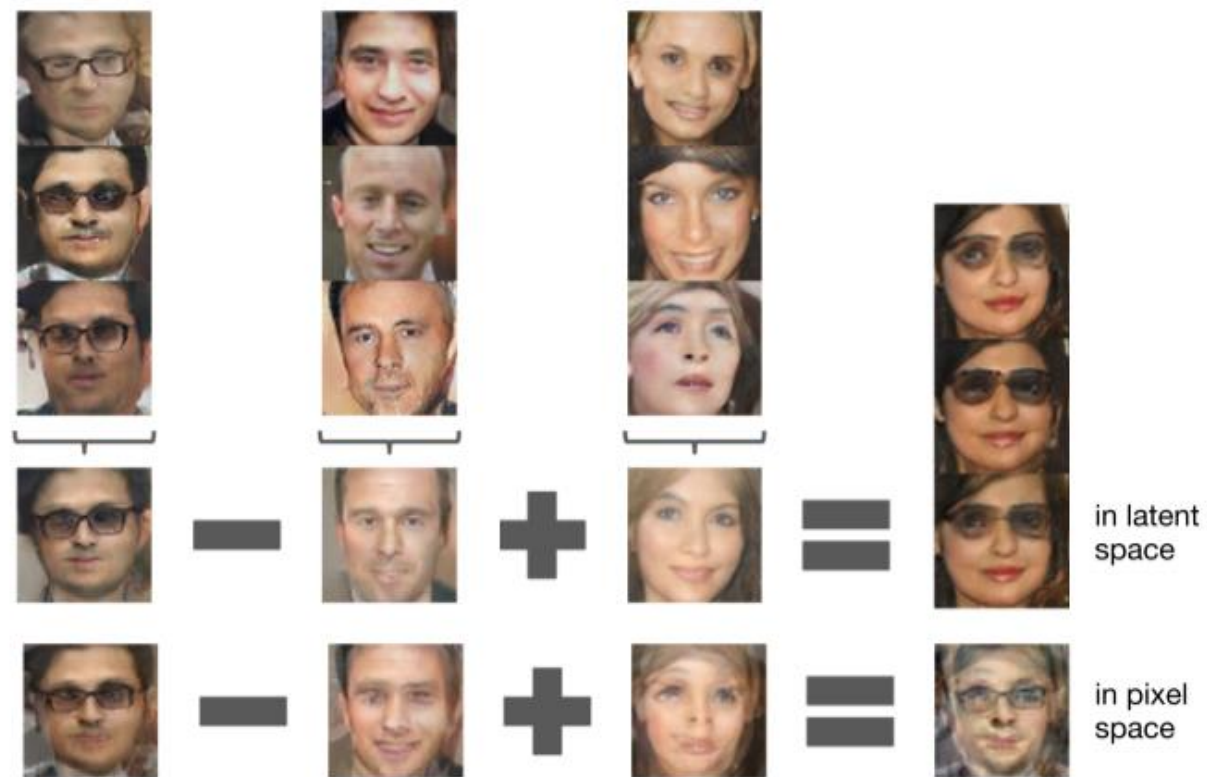
Typical single-cell RNA-seq analysis workflow



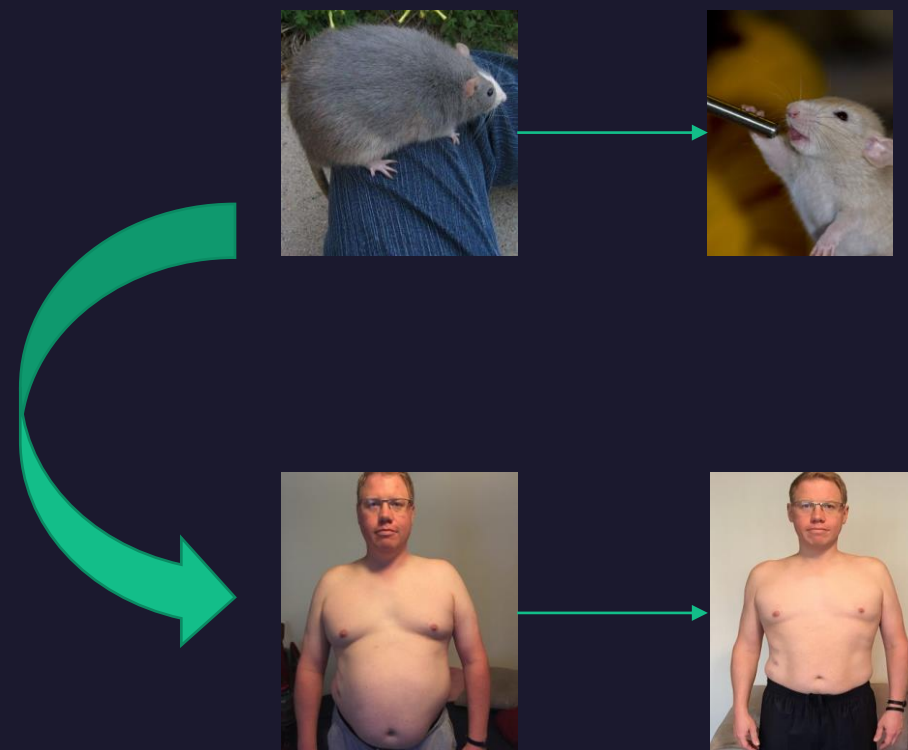
Loss: make 2 as close as 1




Style transfer & domain adaptation by generative neural networks



deep convolutional generative adversarial networks, Radford et al, ICLR 2016



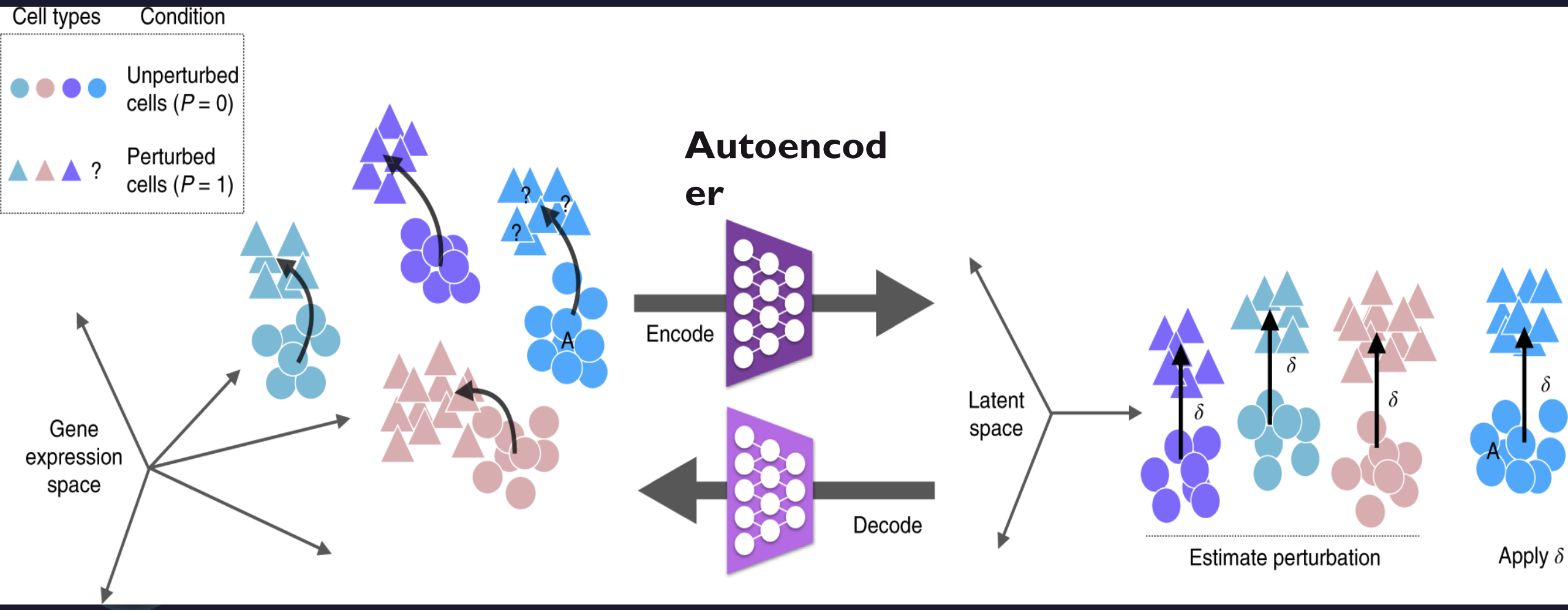
scGen predicts single-cell perturbation responses

Mohammad Lotfollahi ^{1,2}, F. Alexander Wolf ^{1*} and Fabian J. Theis ^{1,2,3*}

Accurately modeling cellular response to perturbations is a central goal of computational biology. While such modeling has been based on statistical, mechanistic and machine learning models in specific settings, no generalization of predictions to phenomena absent from training data (out-of-sample) has yet been demonstrated. Here, we present scGen (<https://github.com/theislab/scgen>), a model combining variational autoencoders and latent space vector arithmetics for high-dimensional single-cell gene expression data. We show that scGen accurately models perturbation and infection response of cells across cell types, studies and species. In particular, we demonstrate that scGen learns cell-type and species-specific responses implying that it captures features that distinguish responding from non-responding genes and cells. With the upcoming availability of large-scale atlases of organs in a healthy state, we envision scGen to become a tool for experimental design through in silico screening of perturbation response in the context of disease and drug treatment.

¹Helmholtz Zentrum München – German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Germany.

²School of Life Sciences Weihenstephan, Technical University of Munich, Munich, Germany. ³Department of Mathematics, Technical University of Munich, Munich, Germany. *e-mail: alex.wolf@helmholtz-muenchen.de; fabian.theis@helmholtz-muenchen.de



Method

$$P(x_i|\theta) = \int P(x_i|z_i; \theta) P(z_i|\theta) dz_i$$

Approximate the posterior distribution

$$\begin{aligned} & \text{KL}(Q(z_i|x_i, \phi) || P(z_i|x_i, \theta)) \\ &= E_{Q(z_i|x_i, \phi)} [\log Q(z_i|x_i, \phi) - \log P(z_i|x_i, \theta)] \end{aligned}$$

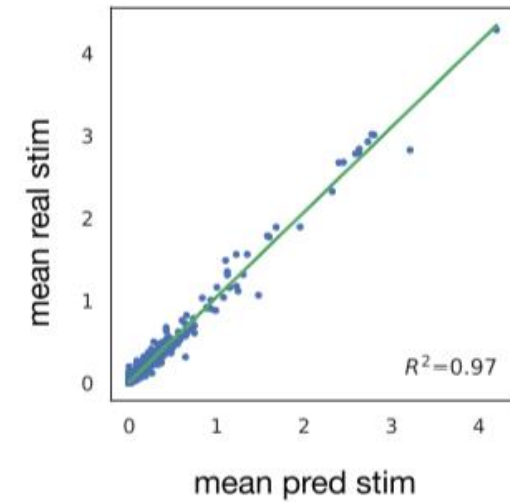
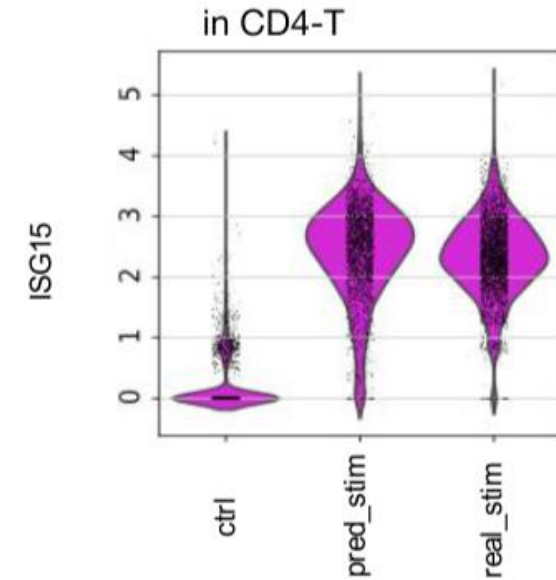
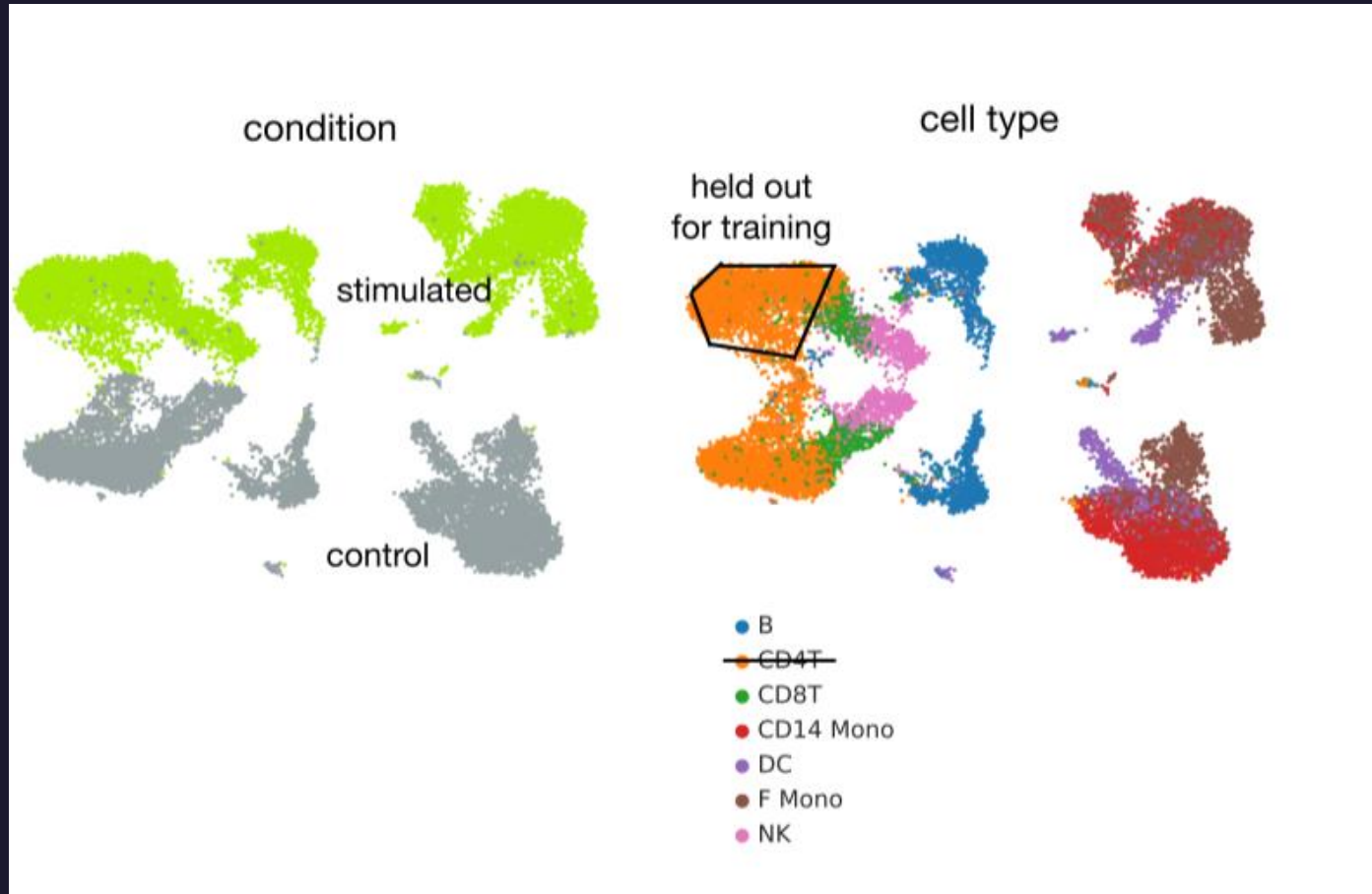
where θ is a model learnable parameters in our neural network and z_i is a latent variable, x_i is your sample.

$$\begin{aligned} & \log P(x_i|\theta) - \text{KL}(Q(z_i|x_i, \phi) || P(z_i|x_i, \theta)) \\ &= E_{Q(z_i|x_i, \phi)} [\log P(x_i|z_i, \theta)] - \text{KL}[Q(z_i|x_i, \phi) || P(z_i|\theta)] \end{aligned}$$

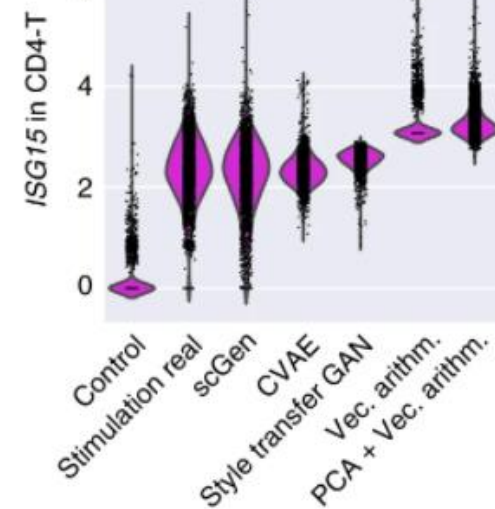
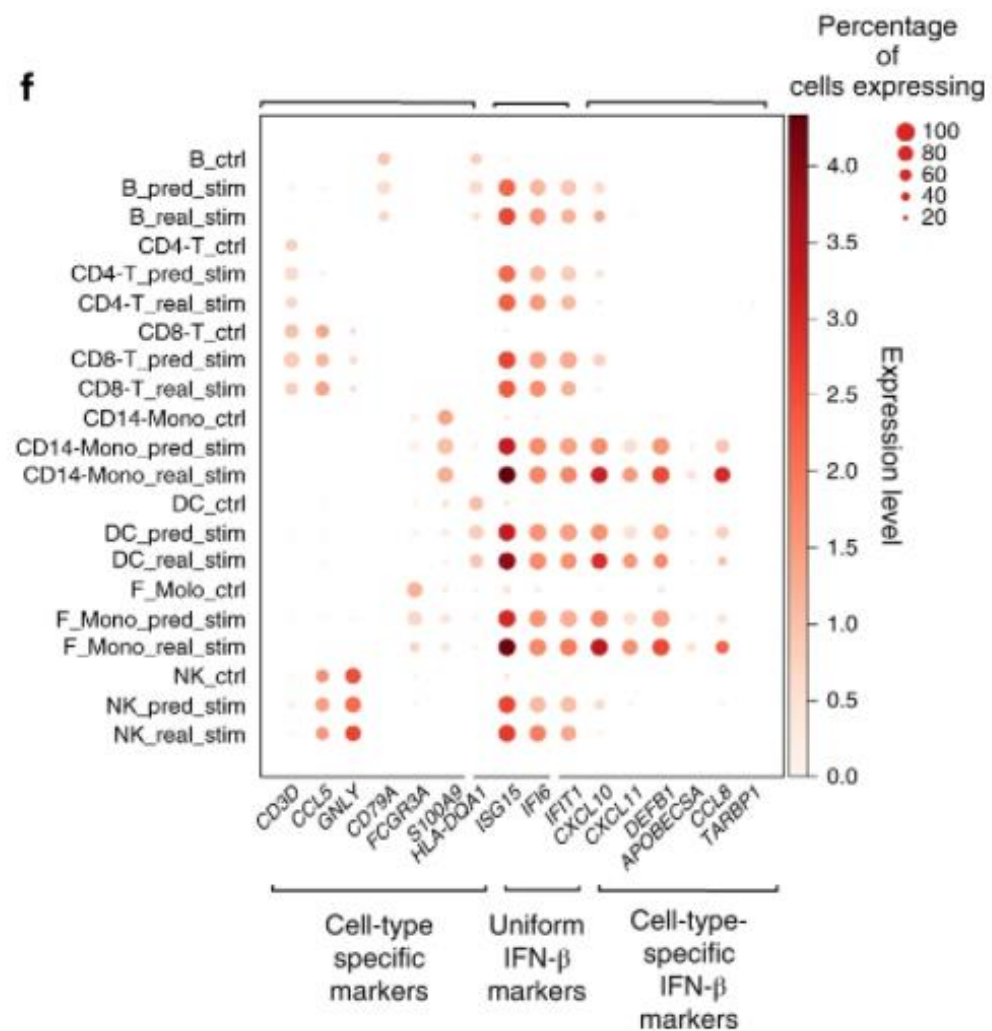


$$\text{Loss}(x_i) = \frac{1}{L} \sum_{l=1}^L \log P(x_i|z_{i,l}, \theta) - \alpha \text{KL}[Q(z_i|x_i, \phi) || P(z_i|\theta)]$$

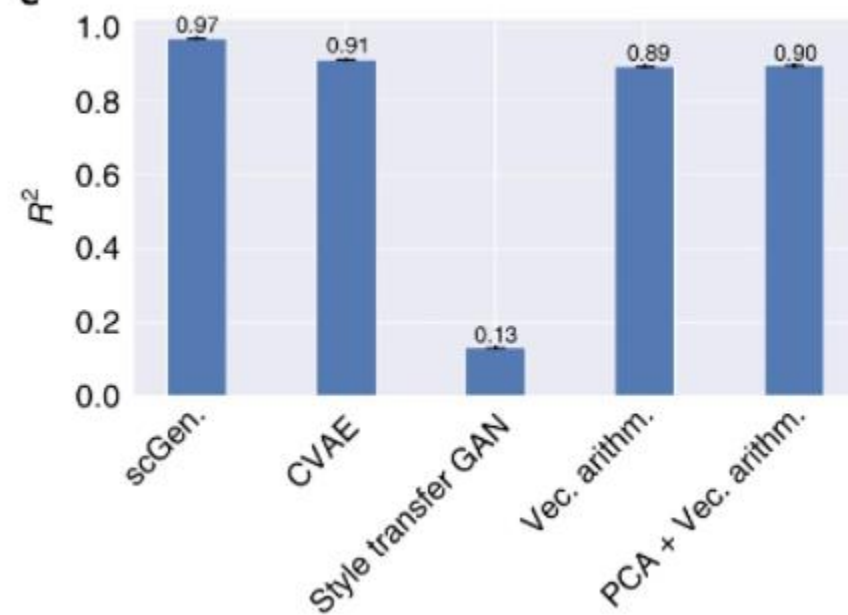
Results



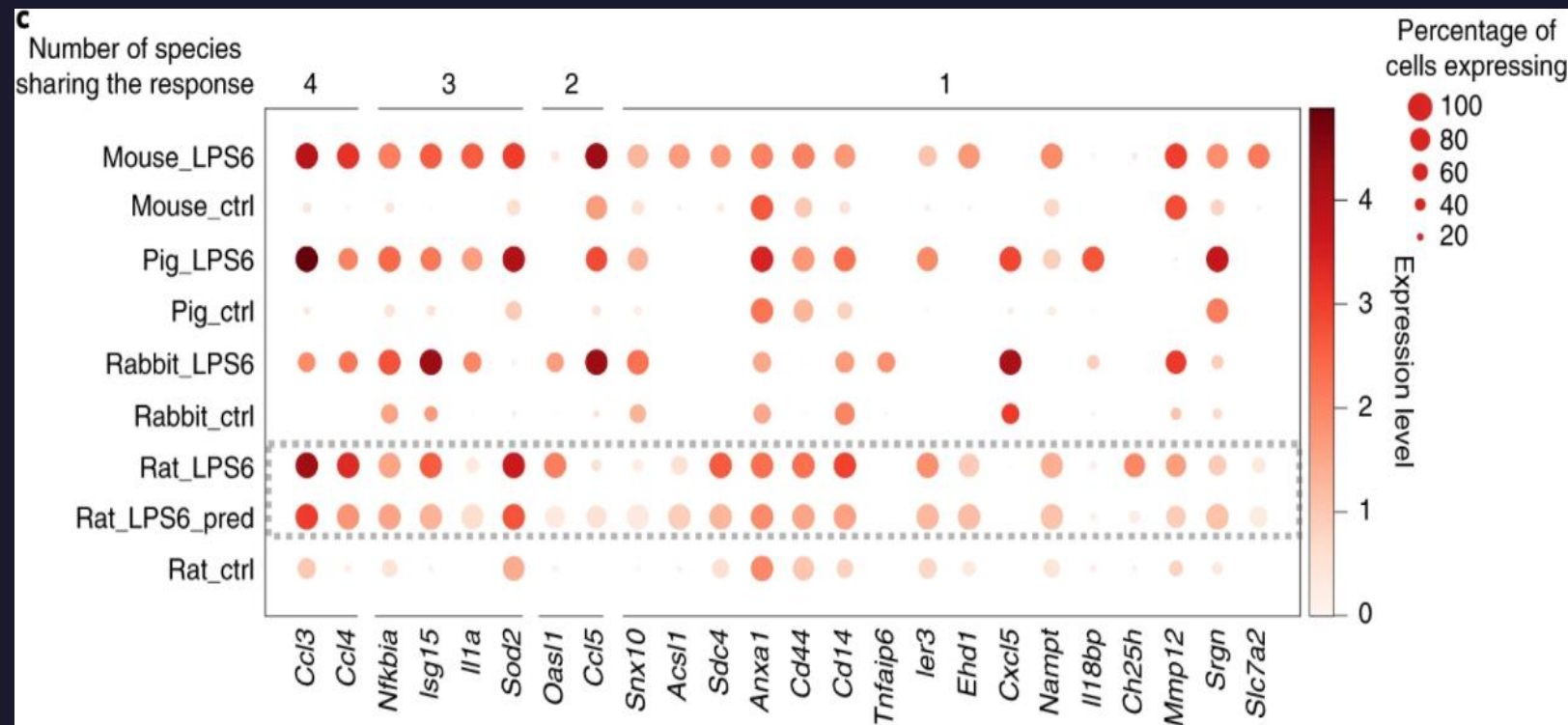
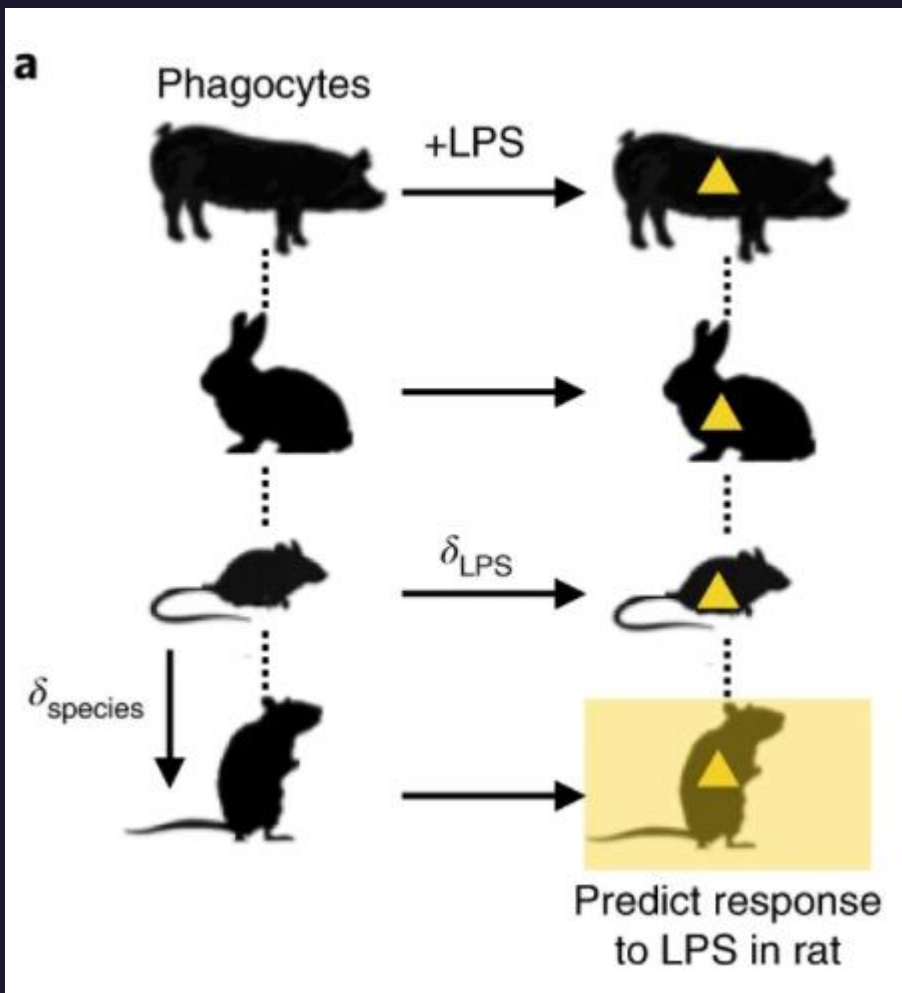
f



e



scGen predicts perturbation response across different species



$$z_{i,\text{rat,LPS}} = \frac{1}{2}(z_{i,\text{mouse,LPS}} + \delta_{\text{species}} + z_{i,\text{rat,control}} + \delta_{\text{LPS}})$$

Model interpretability

Types of DNN Interpretability

1) Example based methods

2) Attribution Methods

3) Data Generation

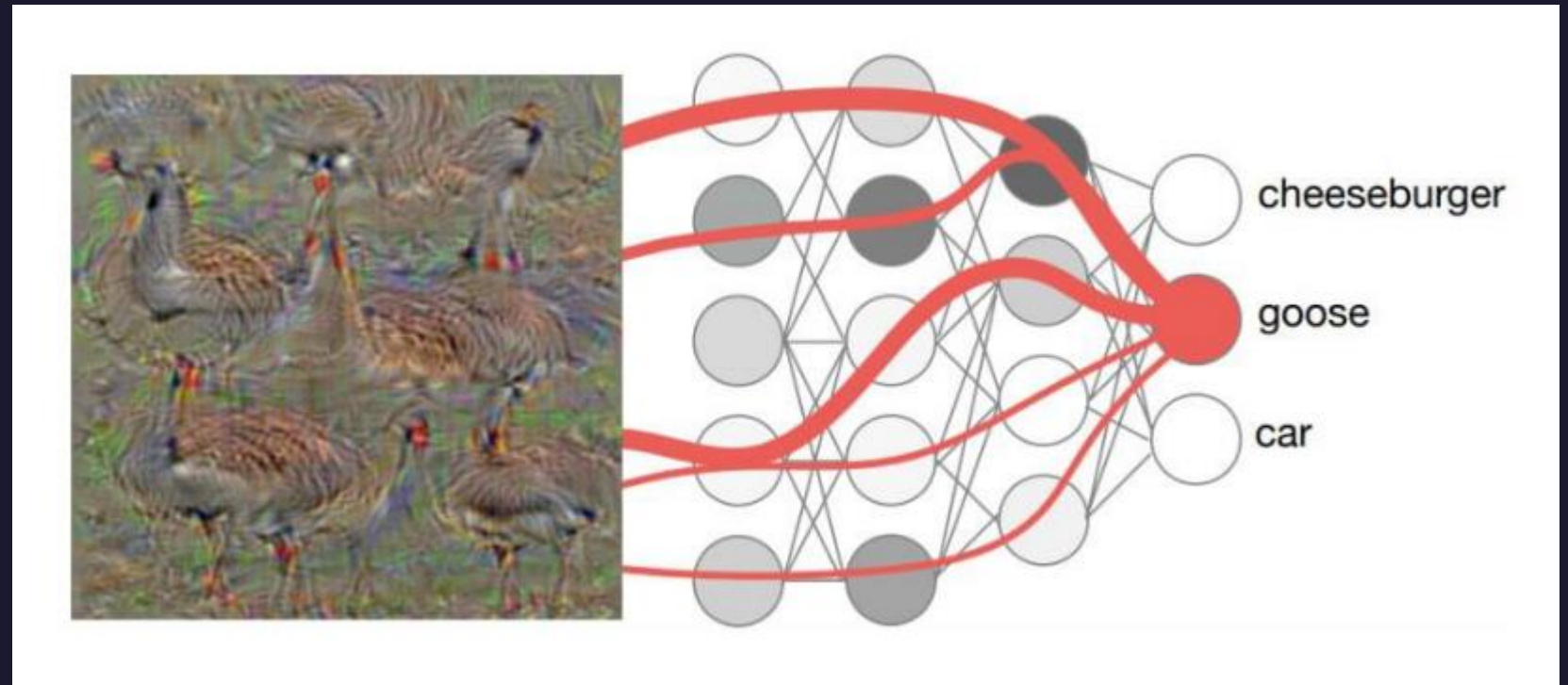


Activation maximization

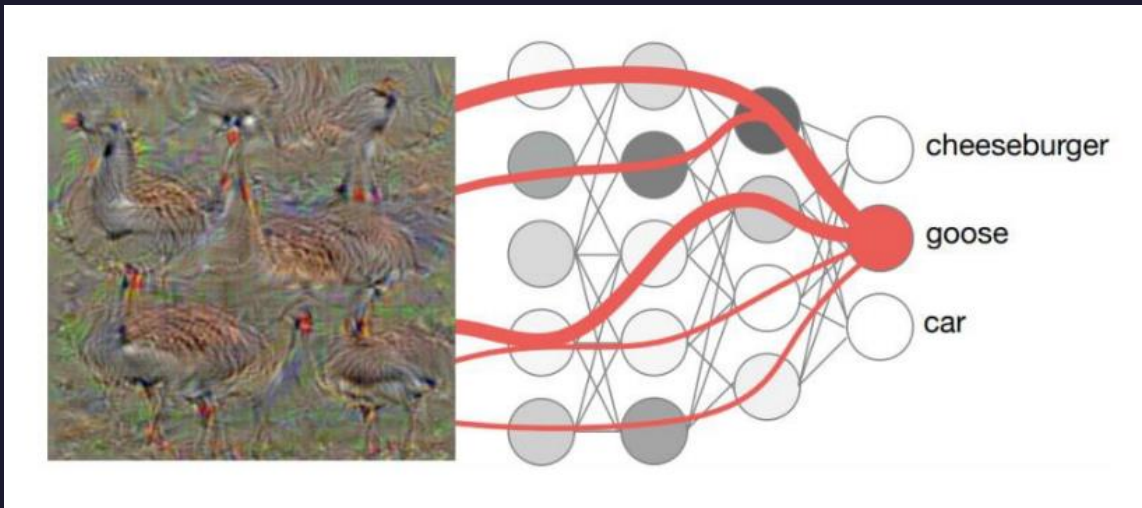
- Generates an image, which **maximizes the class score**.
- **The difference:** the optimization is performed with respect to the **input image**, while the weights are fixed to those found during the training stage.

$$\arg \max_I S_c(I) - \lambda \|I\|_2^2,$$

Find the **most likely** input pattern for a given class.

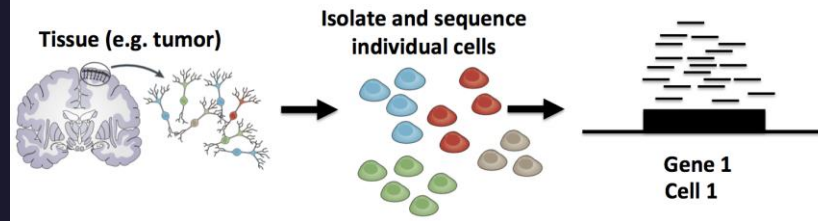


AM modification

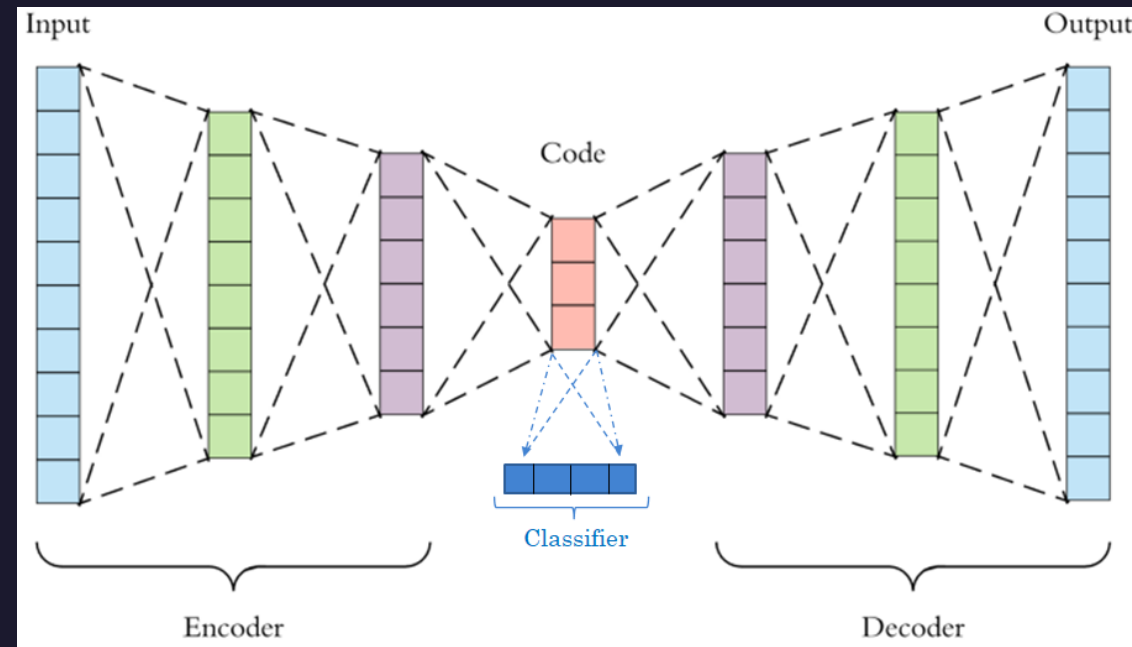
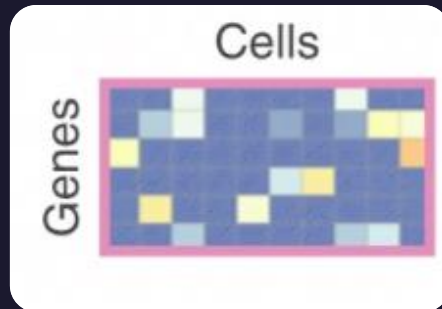


- Problem: The set of all possible images is so vast that it is possible to produce ‘fooling’ images that excite a neuron, but do not resemble the natural images that neuron has learned to detect.
- Image quality can be improved by using image priors.

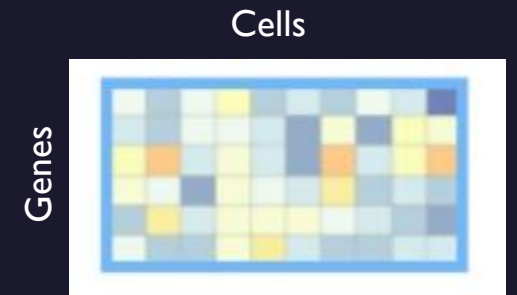
Single-cell RNA-Seq (scRNA-Seq)



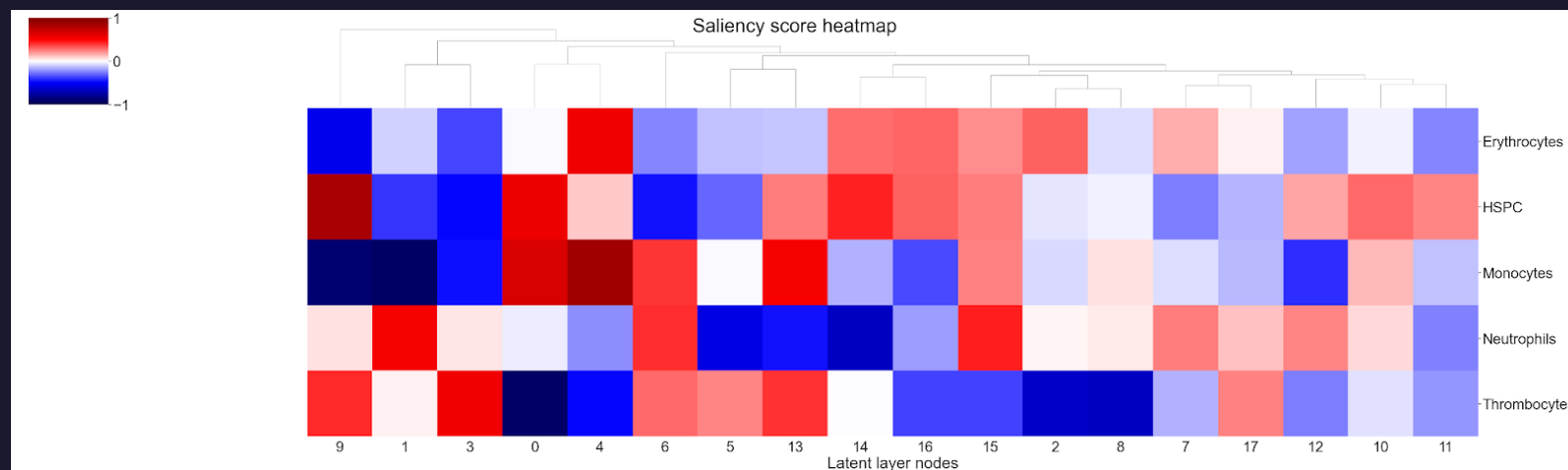
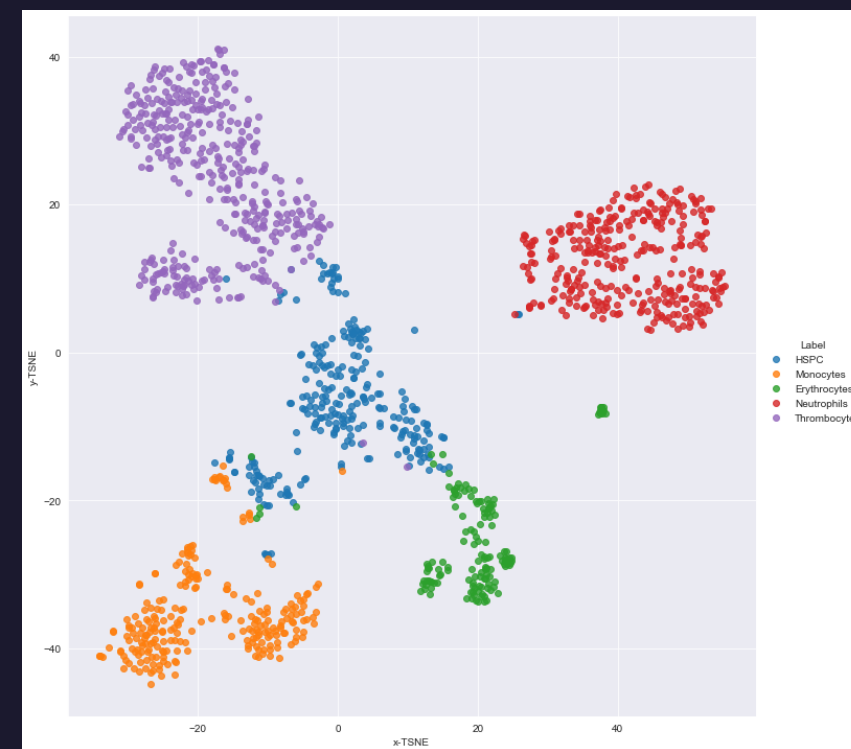
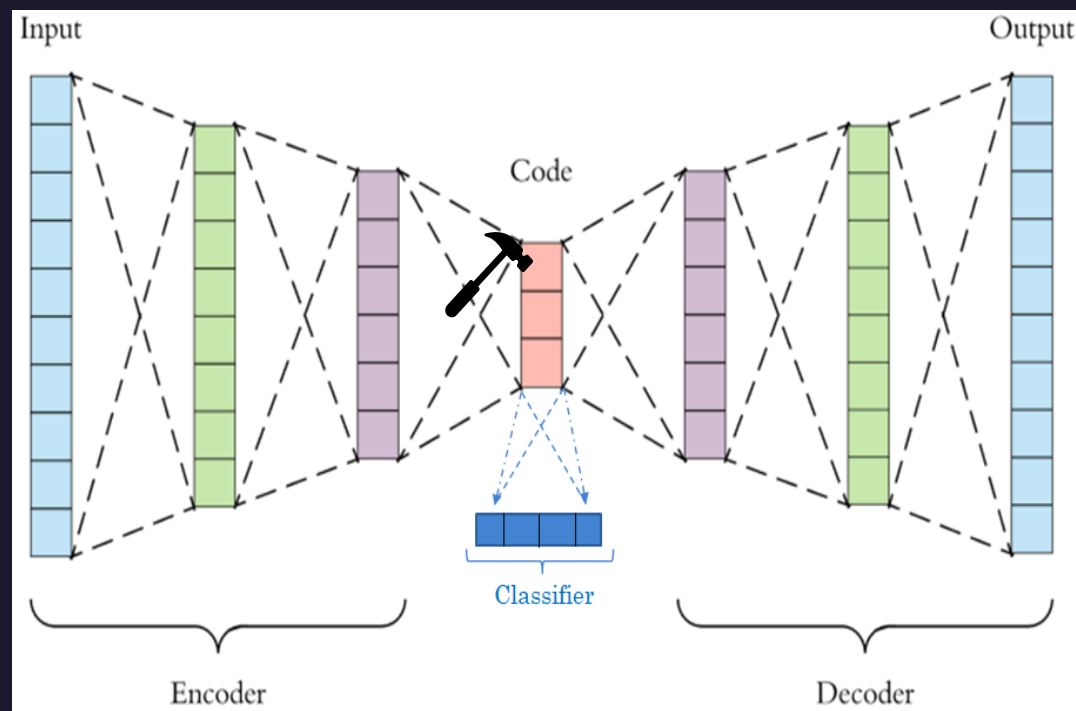
Original



Reconstructed



Zebrafish development



Summary

- The potential of single cell analysis.
- scGen can be used in several contexts including perturbation prediction response for unseen phenomena across cell types and species.
- scGen can be used to design in-silico experiments of gene perturbation.
- Interpretation methods can be used to understand the decision-making process in DL.
- Interpretation methods can help us in detecting master regulators of a biological process.



References

- Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. Nat Methods. 2019 Aug;16(8):715–21.
- Radford A, Metz L, Chintala S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv. 2015 Nov 19;
- Current best practices in single-cell RNA-seq analysis: a tutorial, Malte D Luecken, Fabian J Theis, doi.org/10.15252/msb.20188746
- Svensson V, Vento Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc. 2018 Apr;13(4):599–604



Example based methods

- Which training instance influenced the decision most?
- we need criticism to explain what are not captured by prototypes.
- Influential instances are the training data points that were the most influential for the parameters of a prediction model or the predictions themselves. Identifying and analyzing influential instances helps to find problems with the data, debug the model and understand the model's behavior better.
- Does not highlight which features are important.

