# scPerb

Inspired by the transfer learning paradigm, we presented scPerb, a generative model that can learn the “content” and of the cell types from both the control and stimulus datasets, where represented the “content features” of the cell types, and transfer the style from the control dataset to the stimulation dataset , where represented the “dataset styles”.

scPerb is inspired by Varational Auto-Encoder (VAE) and the style-transfer GAN (stGAN). We used the variational inference to estimate the distribution and of the “content features” in the latent space, and project style input vector into the latent space and learn the transformation from the control dataset to the stimulation dataset . For the rest of the descriptions, we denote $E\_^c(.) $ as a content encoder to learn the cell-type awareness features, to project the random-style input vectors to the latent space, and represented to the and estimation for the distribution of , and for the decoder to generate the stimulation data from the latent variables and .

In the inference stage, given a specific cell type from the control dataset , scPerb will extract the cell type-related features , and get the generated the pseudo-stimulus cell type based on and , a result of a neuro-network, learning the difference between and .

#### Encoders

We assumed the observations and from the control and stimulation datasets had two independent latent features: a cell type-related latent feature, denoted as “content” , and a dataset-specific feature, denoted as “style” . To extract the common cell type content feature, we first project the inputs into the latent space, then estimate the to represent the normal distribution of , and resample the latent variable based on the generated distribution:

We shared the projection weights between the two dataset and , and therefore we can have the latent representation of as:

In this manuscript, our task is to generate the pseudo-stimulus cell types from the same cell types in the control dataset. Therefore, instead of learning the dataset styles explicitly, we applied a light-wise network to learn the transformation in the latent space. Our idea was inspired by the style transfer learnings \cite[StyleGAN], where randomly sampled a noise and project the latent space as the styles. In ScPerb, we applied a style encoder , which can project the into the latent space as the transformation variable to convert to :

And therefore we have the following style loss and the KL regulations:

Where SmoothL1Loss and KL divergence are:

#### Decoder

We applied a decoder to generate the observations from the latent variables . Accordingly, the generated samples were denoted as:

Note that our task was to perturb the cell types from the control dataset to the stimulus dataset, instead of generating the samples from and , we use . Therefore, our Generated Loss is:

#### Loss function

The objective functions will be combined with the Generated loss, Style Loss, and the KL regulation terms.

# Datasets and preprocess

Mohammad et al. [] included three groups of control and stimulated cells: two groups of PBMC cells, and a group of HPOLY cells. Mohammad et al. preprocessed the data by removing megakaryocytic cells, filtering the cells with a minimum of 500 expressing cells, extracting the top 6998 cells, and log-transforming the original data. All the data are available on https://github.com/theislab/scgen-reproducibility.

In our model, we performed further data preprocessing to ensure consistency between control and stimulus cells within each cell type. Specifically, for each cell type, we randomly selected an equal number of cells from both the control and stimulated groups and used them to balance the dataset. This data preprocessing step helped us create a more robust and unbiased dataset, enabling accurate and fair comparisons between each cell type’s control and stimulus conditions during subsequent analyses. By doing such data processing, we guarantee that each pair of and have the same cell type, so the following style transfer process would be valid.

# Statistics and Reproduction

In ScPerb, we evaluated the performance of our model under a fixed seed of 42 by using the square of the , which is calculated by the function. This metric measures the correlation between the generated images and the ground truth data. We computed the values for all genes’ mean and variance and the top 100 Differentially Expressed Genes (DEGs).

To understand the model’s results visually, we created scatter plots comparing the generated images to the corresponding ground truth data. This graph allowed us to observe how well the model’s predictions aligned with the actual values.

Additionally, we investigated the differences between the generated images and the ground truth data for the top DEG using a violin plot. The DEGs were identified using the function, employing the Wilcoxon method.

Through these analyses, we aimed to assess the accuracy and performance of our ScPerb model in generating realistic images based on the input gene expression data. The evaluation of values and the visualization of the scatter and violin plots provided valuable insights into the model’s capabilities and highlighted any discrepancies between the generated and true data for further investigation.

# Results

Fig 1

Fig 2

Fig 3

Fig 4

Fig 5

# Discussion

1. Summarize:

* This is summarize.

1. Pros:

* This is the pros of our model.

1. Cons:

​ This is the cons of our model.

1. Future Improvement:

* In the future…