# SpaPerb

Inspired by the transfer learning paradigm, we presented SpaPerb, a generative model that can learn the “content” and of the cell types from both the control and stimulus datasets, and transfer the style from the control dataset to the stimulus dataset . In the inference stage, given a specific cell type from the control dataset , and a randomly generated variable , SpaPerb will extract the cell type-related features , and project to latent space as the dataset-specific transfer latent variable , and reconstruct the pseudo-stimulus cell type based on and . We implement SpaPerb with the variational auto-encoder (VAE), with a content encoder $E\_^c(.) $ for cell-type awareness features and a style encoder for the dataset transformation, and a decoder to reconstruct the samples from the latent variables.

1. Encoders

We assumed the observations and from the control and stimulus 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 applied a shared weight encoder to extract the features:

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 SpaPerb, we applied a style encoder , which can project the into the latent space as the transformation variable to convert to :

1. Decoder

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

Note that our task was to perturb the cell types from the control dataset to the stimulus dataset, and therefore we didn’t reconstruct the samples from and .

1. Variational Inference

Based on our assumption, we have:

In the implementation, we assume follows the standard Gaussian distribution , and a linear transformation between and in the latent space. Therefore, we have:

We use the Bayesian formula to calculate the posterior distribution and used the variational inference to maximize the ELBO with the Kullback-Leibler divergence:

Where we can have

In our implementation, we used a linear layer to generate the and another linear layer to generate the :

1. Loss function

The objective functions will be combined with the reconstruction loss and the KL regulation terms. Additionally, we optimized the parameters in the style encoder by giving a constraint in the latent space:

Where SmoothL1Loss and KL divergence are:

# 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 the control and stimulus conditions within each cell type 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.

# Results and statistics

In SpaPerb, we evaluated the performance of our model 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 both the mean and variance of all genes, as well as for the top 100 Differentially Expressed Genes (DEGs).

To gain a visual understanding of the model’s results, 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 SpaPerb 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.