1. What is the difference between the two PBMC(peripheral blood mononuclear cell) datasets?

PBMC-Zheng and PBMC-Kang come from the same cell types but not the same tissue.

And they use different methods. PBMC-Zheng dataset uses the 10 \* genomics technique, one of the most widely use techniques for single-cell RNA sequencing in the world nowadays, while Kang and his team invented a new method of single-cell RNA sequencing.

1. How do you process your datasets?

I followed the data processing process of scGen to make a fair side-by-side comparison. We first filtered the cell with a minimum of 500 expressed genes. Then we removed the cell type with unstable labels (the megakaryocyte cells). After that, we log-transformed the data to make the training process more smoothly. Finally, we paired the control and the perturbed groups divided by each cell type for the style-transferring process.

1. What is ?

R^2 is the R^2 value between the model's prediction and the ground truth. The closer the R^2 is to 1, the closer the prediction is to the ground truth, so the better the model makes the prediction.

1. What method did you use to divide the training and testing dataset?

I randomly split the dataset into training sets and testing sets. The ratio between these two sets is 8:2.

1. **Did you do an ablation study?**

To be honest, scPerb is the lowest requirement to do the single-cell perturbation task. By that I mean, scPerb could do single cell perturbation prediction only when both VAE and style-transfer is used.

But we did some comparison with other models. For example, our model has a better performance than scGen, CVAE, stGAN, and sc-WGAN.

Also, we have designed other structures for scPerb. For example, the two encoder structure, the two decoder structure, and other losses such as L1 Loss, L2 Loss, Cos Similarity Loss, and Smooth L1 Loss.

Among all these structures, the model now we are using has the best result. While the other models only result in a R^2 value close to CVAE. So, we finally choose scPerb as our model.

Yes, I did some ablation study to find the best structure of the model.

1. Where do you get your data?

I get it from Zheng, Kang, and Haber. They tested RNA sequencing techniques in their paper and published these datasets. Download links can be found in their papers.

1. How to filter your data?

I will filter out the cells with a minium of 500 genes, and filter out the cells with minimum 3 expressed cells.

1. Why do you choose to filter your data in such a way? Will it filter out important data?
2. Could your model predict the mutation of genes?
3. Why do you participate in the Yau competition?
4. Is your code open source?
5. Why H.poly?

The data comes from pancreas cells and stomach cells, and they are infected by a bacteria called H.poly, or Helicobacter pylori.

1. How do you choose the coefficient of the loss function?

The coefficient of the loss function means to balance the scale of each loss. For example, the KL loss is much larger than Generated loss or style loss, then the coefficient of KL loss is much lower than that of style loss. The ratio between the coefficient of KL to Generated loss is about 1: 1000.