Good morning, everyone. I am Tang Zijia from Guangdong experiment high school, and the title of my work is: single-cell perturbation via style transfer-based VAE.

First of all, let’s talk about single cell Sequencing techniques. Sc-seq comes from single cell RNA sequencing technique. Single cell sequencing technique is a newborn technique which can count the number of specific genes in one single cell. Compared with the traditional bulk method, single-cell RNA sequencing could calculate the gene expression of every single cell in a sample instead of just the mean gene expression of a mixed tissue. If we said that the traditional Bulk method is to calculate the mean concentration of sugar in a cup of fruit juice, then the single cell RNA sequencing technique is to calculate the concentration of sugar in every piece of fruit in the plate.

Now let’s talk about perturbation. Perturbation is the change of gene expression of cells after stimulus. A stimulus can be a treatment from a dose of a drug. For convenience, we denoted the gene expressions before stimulus as control cells, and the gene expressions after the stimulus as the perturbed cells. For example, we can see a T-cell on this graph, and there’s also a tumor cell nearby. After being a drug affected by the tumor cell, the T cell will produce some new proteins that may lead to the apoptosis of the tumor cell. This process of changing the gene expression is called perturbation.

So, what is the importance of predicting perturbation? Perturbation prediction could provide insight for curing diseases. For example, now we are testing whether a kind of new drug could kill cancer cells or not. However, before clinical trial, there were usually little data. And we cannot collect large amount of data because most tumor-related drugs need strictly verification before applying to the patients. In the research stage, we cannot directly collect the data, especially the perturbation data. Inspired by the successful applications of machine learning techniques in many other fields, Researchers can use a model to predict the effect of this new drug to get sufficient data for further research and processing. Our model could provide accurate predictions for such cases. And provide precious data for the future studying and medicine production process.

Now, there are two mainstream ideas to solve this perturbation prediction. The first way is to use statistic methods such as linear regression, and the other way is to use neural networks such as GANs or VAEs. What are the limitations of these ideas? Statistics methods heavily relied on precise features, and thus cannot be easily transferred to a different disease. For example, a method may predict the perturbation of lung cancer correctly, but failed when applying to breast cancer. On the contrary, the machine learning based methods may have better generalization ability but required more data to tuning the parameters and may results in unexpected bad performances. For example, GANs will provide versatile results, they will result in predictions with a wide range of styles instead of the styles of the perturbed dataset. And their training process is not stable. VAEs could result in vague results. That is, they couldn’t divide the style between control the perturbed data. And they may result in a prediction somewhere in between the control data and the perturbed data. Therefore, in my research, we are trying to keep the generalization ability from the machine learning-based methods and adding robust conditions to improve the model performance.

My research was inspired by the following works: scGen, sc WGAN, conditional-VAE and st-GAN.

Our strongest competitor is scGen. It is a recently work published on Nature Methods for single-cell perturbation. it is based on VAE and used a constant vector to represent the difference between the style of control group and that of perturbed group. However, such representation is not perfect because it only uses a constant vector for all cell type, while the style of control group and perturb group for different cells are different, so the conversion process can be improved.

Another attempt was sc-WGAN, a method based on WGAN. WGAN is a revised the version of GAN, which have promising potential in field the generative AI. Sc-WGAN introduced the GAN models in the perturbation problem, but the output from sc-WGAN was not stable. The issues may come from the GAN model, which providing versatile results, and only few of the styles could hit the style of the perturb dataset.

Consequently, we are trying to adding some conditions to guide the model to generate ideal results, and we turned to CVAE and conditional-GAN. One of the most popular conditional GAN was the style-transfer GAN. When given the style of Van Gogh, stGAN can transfer a realistic building into the style of oil painting. In the single-cell perturbation problem, we can condition the control cells and the perturbation cells as two different styles but share the same contents. So, our idea was trying to decouple to style related feature and the content-based features of a cell type, and transfer the control style to the perturbed style in the latent space. Therefore, we introduced the idea of style-transfer into single-cell perturbation problems, and the experiments showed a significantly improvement when compared with the previous works.

This is the overview of scPerb. On the upper part of this graph, the green style vector and the green style encoder, is the part of style transfer. In this part we are trying to generate the style difference of the control dataset and the perturb dataset. And the lower part of this graph is the VAE backbone, where we try to reconstruct the perturb dataset. So, we got our loss function. The first part of function is style loss. Here, we are trying to minimize the generated style difference and real are difference. And the last part is generated loss, meaning to minimize the difference between reconstructed a perturb result and the ground truth. And the KL loss aim to maintain the shape of the latent space. This could make our training process more stable and increase the accuracy of the model.

This is the dataset of scPerb. We use three datasets, H.poly dataset and two PBMC datasets. The H.poly dataset has about 6000 cells with 7000 genes, and PBMC dataset has about 16,000 to 18,000 cells with about 7000 genes.

In these datasets, the ratio of control to perturb is about 1:1, and ratio of training set and testing set is about 8:2.

Now let’s talk about the result of scPerb. We use as a criterion. To be more specific, the between the prediction and ground truth. The closer the value is closer to 1, the more accurate the model is. In this graph, the red bar represents the result of scPerb. As we can see on this graph, scPerb has the highest in all three datasets, meaning that scPerb has a great ability to predict perturbation. And also we could also see two other facts on this graph. That is, all the methods that use VAE has better results than that using GAN, and all the method that use style transfer are better than others.

This is the result of 5 models on PBMC Kang dataset. And the graph of at the left bottom is the graph of our model, scPerb. From this graph we can see that the points are closer to the best fit line than other methods, meaning that the as per has better predictions. Here, we also introduce other criteria to further illustrate the accuracy of the models. We used of top 100 DEGs as a supplement. DEGs are the differently expressed genes, and top DEG are the genes that changes the most during perturbation process. These genes are usually the target genes that leads to illnesses such as cancer. Therefore, we introduced these criteria as a supplement. Under this criterion, scPerb still provide great prediction, result in a nearly 1.

This is the result of scPerb on PBMC Zheng dataset. From the left graph, we could see the result of scPerb on seven different cell types on PBMC Zheng dataset. For most of these cell types, scPerb got really great predictions, they resulted in an of 0.98. And the right graph shows the performance of scPerb on three different situations. The first situation is that the gene expression of control group is low, while the gene expression of the ground truth is high. And the second situation is that the gene expression of both the control and perturb group are high. And the third situation is that the gene expression of both the control group and the ground truth are low. But in these three situations, our method still received a great prediction, as no matter the color or the size of the prediction is really close to the ground truth.

And this is the result of 5 models on H.poly dataset, and we could see five umaps on this page. The left umap is for a scPerb. The blue dot on this graph represents the control dataset, orange dot represents the perturb dataset, and the green dot represents the prediction of the models. For scPerb, we could see that the green dot and orange dot merge well with each other, and the green dot does not merge with blue dots. This means that our model could divide the style of control group and perturb group very well.

And for the VAE based method, say the CVAE and scGen, we could see that they merge three kinds of dots together, meaning that their predictions are somewhere in between the control group and the perturb group. This means that their methods couldn’t really divide the styles between the control datasets and perturb dataset. For the methods based on GAN, like the style-transfer GAN or the sc-WGAN, we could see that their predictions are far away from the control group and the perturb group, meaning that their predictions are not accurate at all.

So, what’s the next step of our model? Next, we are trying to replace the VAE into some more fancy models such as VQVAE or the diffusion model. In the computer vision applications, VQVAE and the diffusion models have proved their potentials, and we believe the advantage the more compilated and deliciated designed VAEs can also improve the performance in single-cell perturbation.

Moreover, we can also extend our work to the next-generation of single-cell gene expression datasets, which is the single-cell spatial data. The single-cell spatial data capture more detailed gene expression in sub-cellular level, and also have the high-resolution spatial images.

Imagine that we are trying to test whether a drug could kill cancer cells or not. Now we got a sample consisting of cancer cells, immune cells, and some other normal cells. For scPerb, what we could predict now is: what will happen to the cancer cells, the immune cells and normal cells, isolate after injected this drug. But with spatial transcriptomics, we could further predict what would react between the cells, say, will the immune cells started to recognize the human cells and kill it, or will the cancer cells grow bigger and affect other normal cells around it? In this way, our model would have a more promising future.

This is the reference sheet of the presentation, thanks for listening.