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, what is single cell perturbation? Sc-seq comes from the single-cell RNA sequencing technique, a newborn technique that 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 say 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 on 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 or an infection of another cell. For example, we can see a T-cell on this graph, and there’s also a tumor cell nearby. After being 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. For convenience, we denoted the cells before the stimulus as control cells, and the cells after the stimulus as the perturbed cells.

So, what is the importance of predicting perturbation? Perturbation prediction could provide insight into curing diseases. For example, imagine now we are testing whether a kind of new drug could kill cancer cells or not. However, before the clinical trial, researchers cannot collect large amounts of data because nearly all drugs need strict verification before being applied to the patients. This process is not only labor intensive, but also time-consuming. Inspired by the successful applications of machine learning techniques in many other fields, Researchers will now use model to predict the effect of this new drug to get sufficient data for further research. Our model could provide accurate predictions for such cases. And provide precious data for their future study.

Now, there are two mainstream ideas to solve this perturbation prediction. The first way is to use statistical methods such as linear regression, and the other way is to use neural networks such as GANs or VAEs.

But these methods have their own limitations. Statistics methods heavily rely 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 fails when applied to breast cancer.

On the contrary, the machine learning-based methods may have better generalization ability but require more data to finetune the parameters and may result 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 controlling the perturbed data. 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 add robust conditions to improve the model performance.

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

Our strongest competitor is scGen. It is a recent work published in Nature Methods for single-cell perturbation. it is based on VAE and uses a constant vector to represent the difference between the style of the control group and that of the perturbed group. However, such representation is not perfect because it only uses a constant vector for all cell types, while the style of the 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 version of GAN, which has promising potential in field the of 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 provides versatile results, and only a few of the styles could hit the style of the perturbed dataset.

Consequently, we are trying to add some conditions to guide the model to generate ideal results, and we turned to CVAE and conditional-GAN. CVAE could concatenate the label of the input data, telling the model whether the input is from control group or the perturb group and from which cells. But such learning process is not perfect, the problem of “vague” images still exists. stGAN could transfer the style from one image to another, for example, it could transfer the style of Van Gogh’s oil painting to a landscape photograph. 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 to try to decouple style-related features and the content-based features of a cell type and transfer the control style to the perturbed style in the latent space. We finally introduced the idea of style transfer into single-cell perturbation problems, and the experiments showed a significant 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 the part of style transfer. In this part, we are trying to generate the style difference between the control dataset and the perturbed dataset. The lower part of this graph is the VAE backbone, where we try to reconstruct the perturbed dataset. So we got our loss function. The first part of the function is style loss. Here, we are trying to minimize the generated style difference and real difference. And the last part generates a loss, meaning minimizing the difference between reconstructed a perturb result and the ground truth. The KL loss aims 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, the H.poly dataset and two PBMC datasets. The H.poly dataset has about 6000 cells with 7000 genes, and the 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 the ratio of training set and testing set is about 8:2.

Now let’s talk about the result of scPerb. We use R^2 as a criterion of accuracy. To be more specific, the R^2 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 we could also see two other facts on this graph. That is, all the methods that use VAE have better results than those using GAN, and all the methods that use style transfer are better than others.

This is the result of 5 models on the PBMC Kang dataset. And the graph 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 scPerb has better predictions. Here, we also introduce another criterion to further illustrate the accuracy of the models. We used R^2 of top 100 DEGs as a supplement. DEGs are the differently expressed genes, and the top DEG is the gene that changes the most during the perturbation process. These genes are usually the target genes that lead to illnesses such as cancer. Therefore, we introduced this criterion as a supplement. Under this criterion, scPerb still provides great prediction, resulting in a nearly 1.

This is the result of scPerb on the PBMC Zheng dataset. From the left graph, we can see the result of scPerb on seven different cell types on the PBMC Zheng dataset. For most of these cell types, scPerb got really great predictions, they resulted in an of 0.98. The right graph shows the performance of scPerb in three different situations. The first situation is that the gene expression of the control group is low, while the gene expression of the ground truth is high. The second situation is that the gene expression of both the control and perturb group are high. 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.

This is the result of 5 models on the H.poly dataset, and we can see five maps on this page. The left map is for a scPerb. The blue dot on this graph represents the control dataset, the orange dot represents the perturbed 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 the blue dots. This means that our model could divide the style of the control group and perturb the group very well.

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 perturbed group. This means that their methods couldn’t really divide the styles between the control datasets and the perturbed 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 perturbed 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 with some more fancy models such as VQVAE or the diffusion model. In computer vision applications, VQVAE and the diffusion models have proved their potential, and we believe the advantage of 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 at the sub-cellular level and also have 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, isolated after being injected with this drug. But with spatial transcriptomics, we could further predict what would react between the cells, say, will the immune cells start to recognize the human cells and kill them, or will the cancer cells grow bigger and affect other normal cells around them? In this way, our model would have a more promising future.

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