1. **Introduction**

Single-cell RNA sequencing (scRNA-seq) is a revolutionary method to establish a pipeline to discover the heterogeneous gene expression for single-cell profiling. The technologies measure thousands of single cells transcripts from multiple biological samples under various conditions\cite[]. Such breakthrough technology has inspired biological tools including cell-type identifications\cite[这里需要帮忙找一些相关的reference], Oncology\cite[], Immunology\cite[] and Drug discovery and developments\cite[]. In this manuscript, we focus on the perturbation task, identifying the gene response to specific conditions such as a dose\cite[], a treatment\cite[], or a modification of genes.

Although the scRNA-seq has led to a remarkable growth of the single-cell data, it was still challenging to collect the pairs of control and stimulation sample slides for a particular dataset. Most of the data are missing the stimulation sample slides, and it raised an urgent need to leverage the limited existing data to generate a transferring approach to convert the control slides into reliable stimulation data. Since the data can be collected from different tissues\cite[], different platforms, and limited data size, the transferring approach need to have the following properties: (1) A robust methodology for various tissues, which had unique gene expression patterns for the specific diseases. (2) Generalization ability for data samples collected from multiple data platforms. (3) Limited data size.

Recently, data-driven algorithms have been proven to be reliable for multiple source types\cite[], and achieved state-of-the-art in computer vision task\cite[], Nature Language Processing (NLP) tasks\cite[] and also raise attention in the area of biomedical analysis\cite[]. The generative models are a good fit to fill up the blank of the missing pieces in the perturbation task. Generative Adversarial networks (GAN) and Variational Auto-Encoders (VAE) are the two mainstreams in generating the most associated data samples to fill up the blank in the missing data. GAN introduced a generator to construct the “fake” but realistic data and trained a discriminator to adversarial determine whether the “fake” data was good enough. The adversarial battle aimed to train a robust generator to infer high-quality data samples. The major drawback of GAN is that GAN is hard to balance the adversarial training which led to a useless collapsed generator that is very sensitive to the input noise. sc-WGAN [1] transferred the more stable WGAN to the single-cell perturbation and st-GAN [2] introduced the idea of style transferring that transfer multiple styles determined by the users to the generator. On the opposite, VAE is more elegant from the mathematic perspective, which assumed result can be sampled from a multivariate Gaussian distribution and used an encoder to estimate the mean and variance of the Gaussian distribution components of the original distributions, then generate new data observations from the estimated distribution using variational inference. The generated samples from VAE were more stable but usually blurred and related to the mean of the observation datasets. In particular, scGEN [3] assumed a fixed linear gap between the control cells and the stimulation cells, calculated the latent difference from both datasets and predicted the stimulation cell response using latent representation from control cells and the stimulated cells. Conditional Variational Auto-Encoder (CVAE) [4] introduced more constraints to the neural network, allowing the end-users to generate more desired reconstructed samples for customized demands.

In our manuscript, we presented a novel tool named “scPerb” (shown Fig.1), which was inspired by the VAE [5] and style-transfer GAN [2]. Suppose we have two datasets with different “styles” but the same cell types. We denoted to represent the ith cell from the control dataset, and for the jth cell from the stimulation dataset. scPerb decouples the perturbation task mentioned above into two steps: estimate the latent features of the cell types and a learnable dataset-related style transformation matrix. Inspired by the VAE architectures, scPerb first estimates the multi-variance normal distribution of the latent cell type feature c. Inspired by the style-transfer GAN [2], scPerb uses a neural network to learn the style transformation matrix from the dataset. Compared with scGEN, which used a fixed vector to transfer the latent features from the control cells to the stimulation cells, scPerb introduced more learnable parameters and allowed the neural network dynamically assign the weights of the “style-transfer” vector based on the data. Therefore, scPerb can better learn both the style and content difference between the control and stimulation datasets, and output a better prediction compared to scGEN.

1. **Methods**

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 Variational Auto-Encoder (VAE) [5] and the style-transfer GAN (stGAN) [2]. 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 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 pseudo-stimulus cell type based on and , a result of a neuro-network, learning the difference between and .

* 1. 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 datasets 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 [2], 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:

* 1. 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:

* 1. Loss function

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

1. **Datasets and preprocess**

We obtained the PBMC-Zheng dataset from Zheng et al. [6]. After removing the megakaryocyte cells, which have an uncertainty of assigned labels, we log-transformed and normalized the data and selected the top 7000 highly variable genes.

Kang et al. published a dataset from PBMCs including both control and stimulated cell types [7]. Among these data, we extracted the average of the top 20 cluster genes, which has 6998 genes in total, from seven cell types, respectively: B cells, CD4-T cells, CD8-T cells, CD14 Mono cells, Dendritic cells, FCGR3A Mono cells, and NK cells, the same cell types as the PBMC-Zheng dataset.

Harber et al. presented a dataset using the responses of epithelial cells infected by *Salmonella* and *H.poly* [8]. In this dataset, there are 1770 *Salmonella*-infected cells, 2711 *H.poly*-infected cells, and the rest 3240 control cells. The data were also normalized and log-transformed. The top 7000 highly variable genes were selected in this dataset.

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.

1. **Statistics and Reproduction**

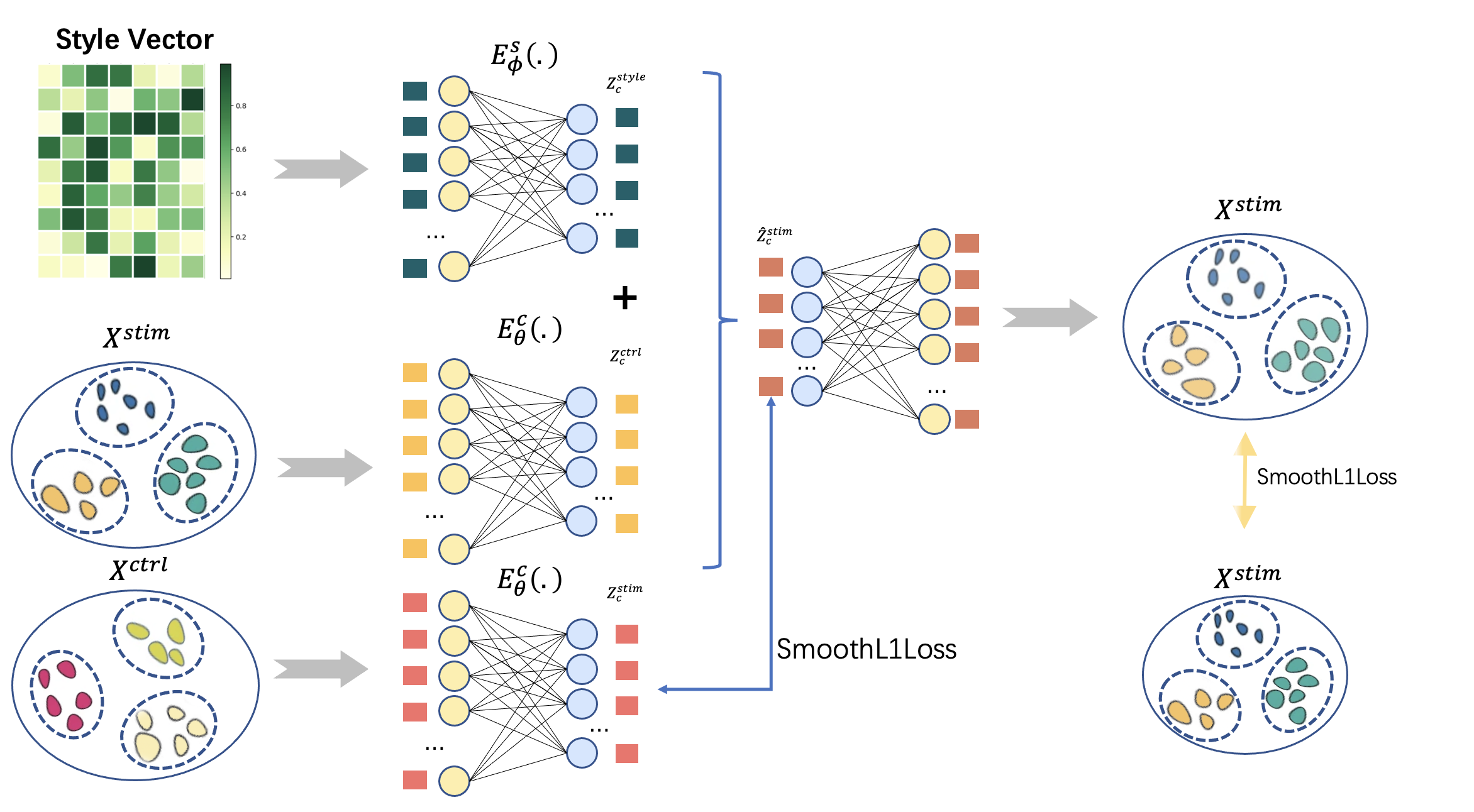
In scPerb, we evaluated the performance of our model under a fixed seed of 42 by using the square of the , which is calculated by the function [9]. This metric measures the correlation between the generated images and the ground truth data. We computed the values for all genes’ mean and variance and 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 [10] function, employing the Wilcoxon method [11].

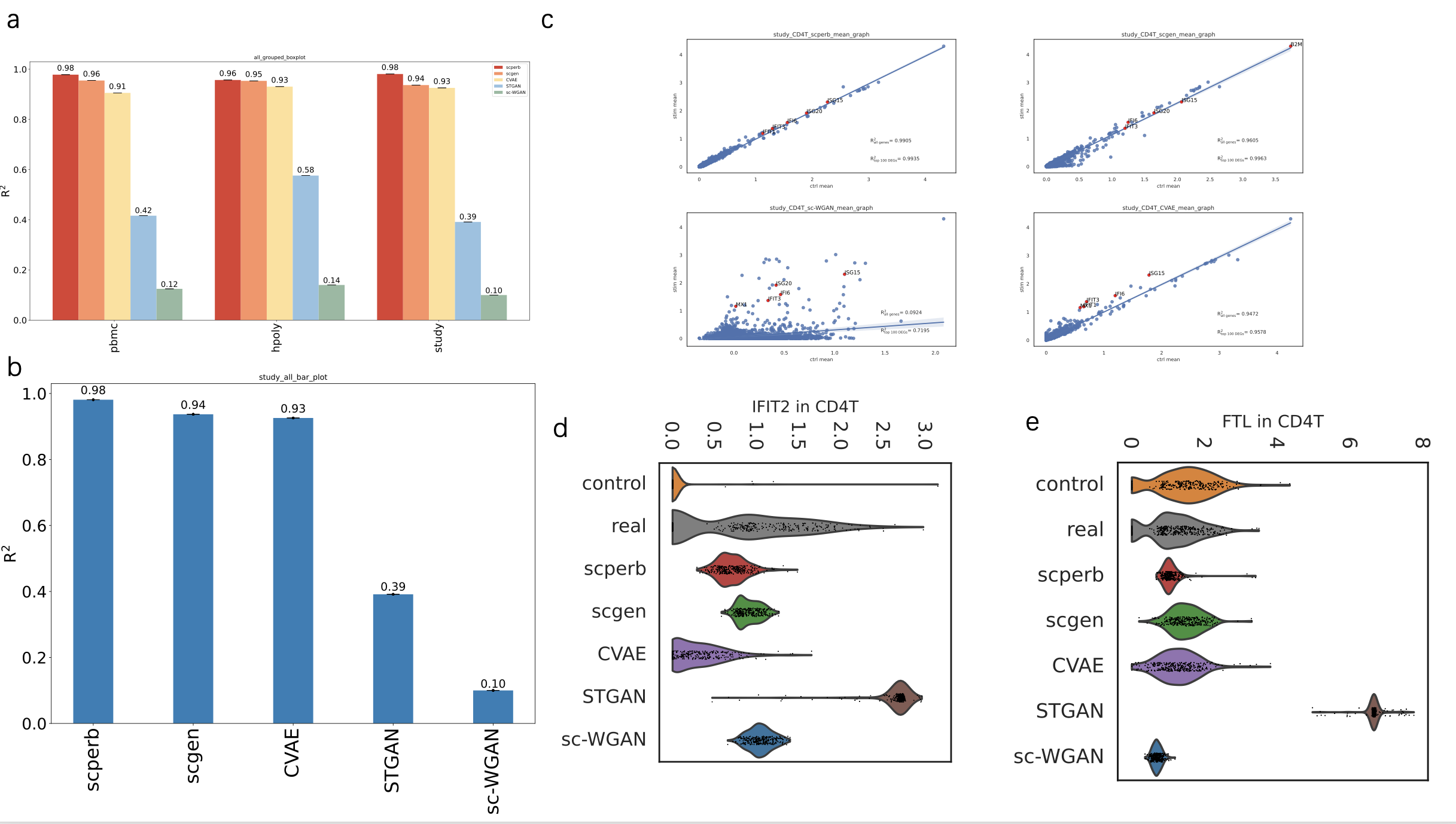
Through these analyses, we aimed to assess the accuracy and performance of our ScPerb model in generating realistic images based on the input gene expression data. The evaluation of values and the visualization of the scatter and violin plots provided valuable insights into the model’s capabilities and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

1. **Results**



5.1 scPerb is an innovative generative model that can accurately predict single-cell perturbation responses.

To demonstrate the performance of scPerb, we applied it to three datasets. Among these three datasets, two of them are two groups of published human peripheral blood mononuclear cells (PBMC) datasets [6, 7] stimulated with interferon () methods, and the rest is a group of intestinal epithelial cells fetched by parasitic helminth *H.poly* cells [8]. We fairly compared our proposed scPerb with other benchmarking papers [1-4]. In this process, we first run all the cell types of one dataset in each model and combine the results for further processing. Then, we compare the prediction of all the methods with the ground truth, and the stimulation cells in the dataset, and get a final score. Compared to all the other methods including scGEN, CVAE, style-transfer GAN, and sc-WGAN, scPerb has the best result in predicting the effect of perturbation. In the published human peripheral blood mononuclear cells (PBMC) dataset, scPerb gain a mean value of 0.98 on both datasets, while scGEN achieved 0.94 and 0.96 respectively on PBMC-Zheng and PBMC-Kang dataset. Meanwhile, CVAE only achieved 0.93 and 0.91, respectively. Moreover, both GAN-based methods style-transfer GAN and sc-WGAN poorly predicted the perturbation response, resulting in values of 0.39, 0.42 and 0.10, 0.12 accordingly. Furthermore, it is worth noting that scPerb outperforms all other methods in the *Hpoly* dataset, exhibiting a remarkable value of 0.96. (Fig. 2a). In conclusion, scPerb best correlates the stimulation cells among all the other benchmarking methods.

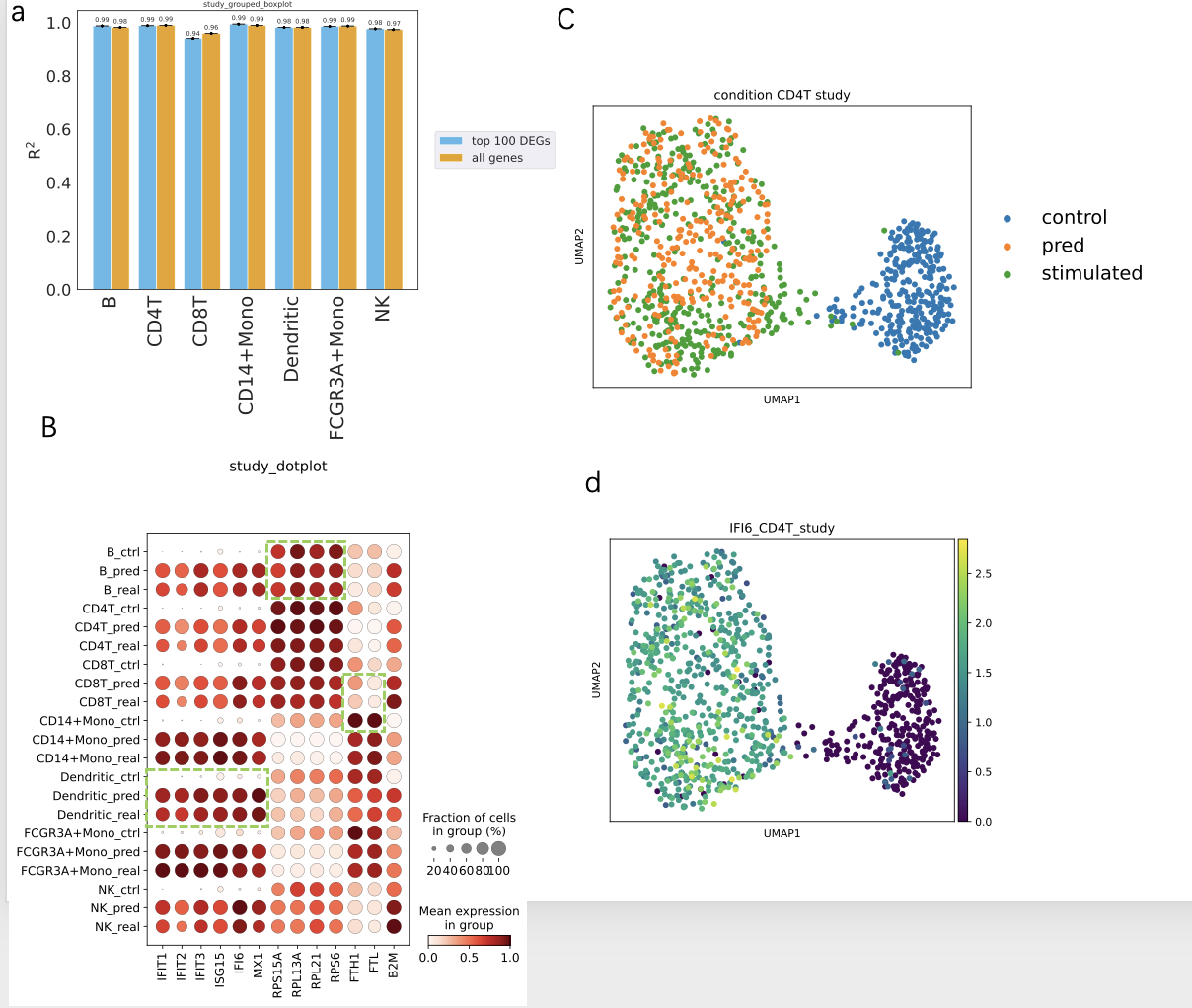


5.2 scPerb outperforms other benchmarks

In the PBMC-Zhang dataset (Fig. 2a), we measure the mean among all cell types. Each shows the specific correlation score between the prediction and the real stimulation cell type. scPerb achieved a mean of 0.98, which is higher than the second-best benchmarking scGEN () and the third-best benchmarking CVAE (). Surprisingly, the GAN-based methods had much worse performance. The stGAN and sc-WGAN only have and accordingly. When comparing the performance with a specific cell type , scPerb achieved =0.99, followed by scGEN and CVAE with and . stGAN and sc-wGAN had poor performance, with and .

Moreover, within the CD4T cell type, scPerb demonstrates superior performance over other benchmark methods, achieving a correlation value of 0.9905 across all genes. This stands in contrast to the values of 0.9605, 0.9472, and 0.0924 observed for scGEN, CVAE, and sc-WGAN, respectively. Additionally, scPerb achieves an R^2 value of 0.9935 when considering the top 100 differentially expressed genes (DEGs), showcasing a result that closely aligns with the ground truth. In particular, the top 5 DEGs are identified as IFIT1, IFIT3, IF16, ISG20, and ISG15. (可能需要增加DEG的作用的文献)

Furthermore, we inspected the correlation of a specific gene expression among the benchmarking papers. is **“the relation between FTL and CD4T, need refers 需要FTL和CD4T对应的参考文献”**. In our dataset (Fig. 2d), the gene expression of FTL in the control dataset is comparable to the stimulation dataset. In this case, only scPerb accuratly generated a prediction that exhibited a mean value comparable to the ground truth. In contrast, scGEN, and CVAE were more parallel to the gene expression in the control dataset, while st-GAN and sc-WGAN had relatively smaller gene expression. To support our observation, we also inspect another gene expression in the same cell type “这里需要IFIT2和CD4T对应的参考文献”. The gene expression of from the control dataset is full of zero values, while the gene expressions of in the stimulation dataset had higher values. (Fig. 2e) In this case, scPerb, scgen, and sc-WGAN can demonstrate the dataset difference, while CVAE and stGAN failed to transfer the gene expression from the control dataset to the stimulation dataset. Furthermore, we applied a Wilcoxon test [11] to examine whether the prediction and the real stimulation gene expression had significant differences. The p values showed that scPerb had 0.1763 in the prediction of and 0.0742 in the prediction of , which failed to reject the hypothesis that the prediction and the real stimulation gene expression had significantly difference. Meanwhile, in the rest of the benchmarks, the P values of scGEN were and , CVAE (pvalue=0.0307 for FTL and pvalue<0.0001 for IFIT2). For the GAN-base methods, all of the P values were smaller than 0.0001. Consequently, the P values from the benchmarking papers reject the hypothesis and indicate a significant difference between the prediction and real stimulation gene expressions.

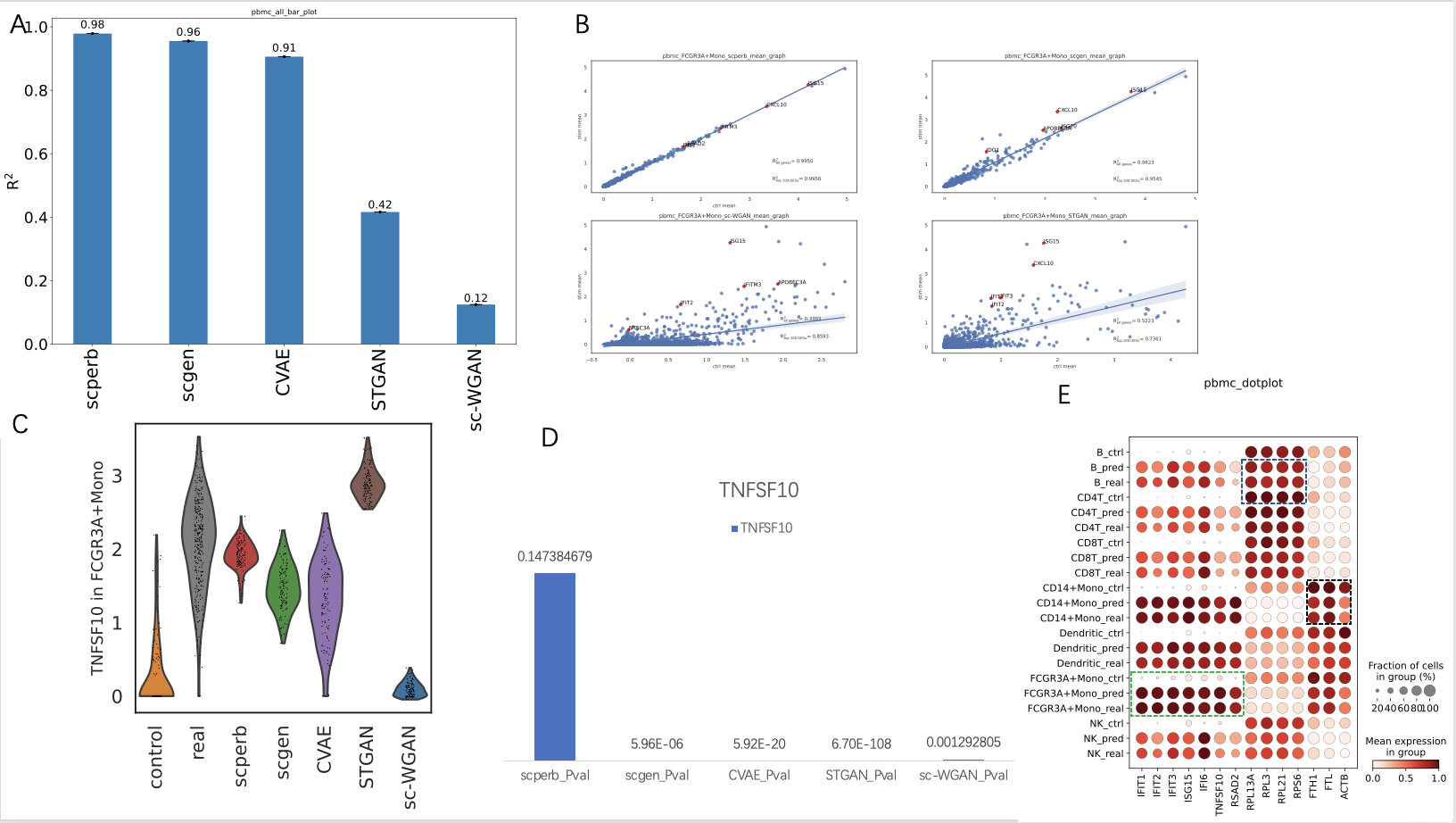


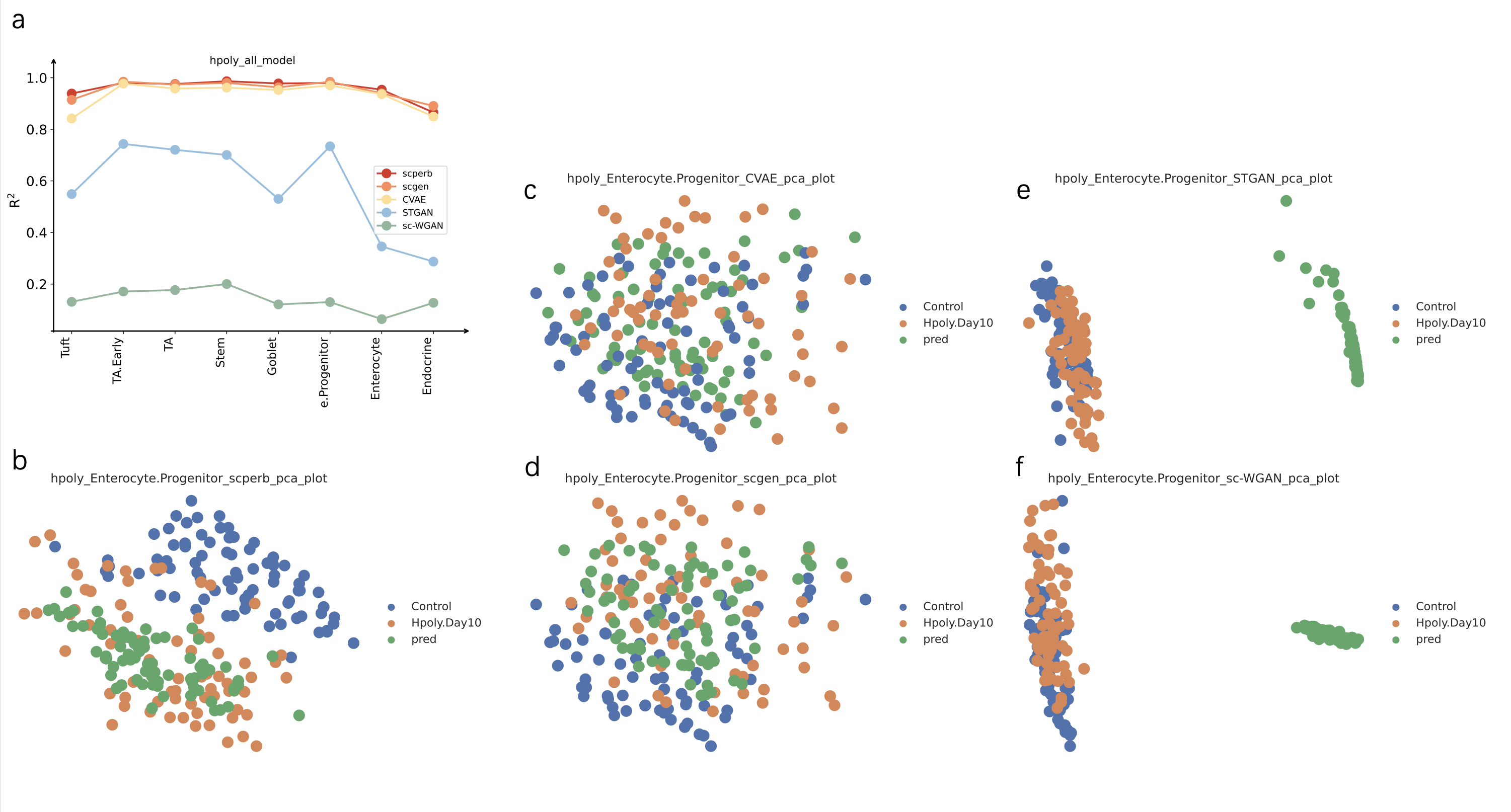
5.3 scPerb can accurately predict the perturbation of cells

We explored the performance of scPerb in a larger range of genes in more cell types. Figure 3a explored the among different cell types. Specifically in this study, we examined B cells, CD4-T cells, CD8-T cells, CD14 Mono cells, dendritic cells, FCGR3A+Mono cells, and NK cells. Remarkably, when considering all genes, B cells displayed an exceptional R² of 0.99, with the top 100 differentially expressed genes (DEGs) yielding a slightly lower yet highly significant value of 0.98. Meanwhile, CD4-T cells showcased an impressive consistency in prediction accuracy with R² scores of 0.99 and 0.99 for all genes and top 100 DEGs, while CD8-T cells demonstrated strong predictive power with R² values of 0.96 and 0.94. Moreover, NK cells exhibited robust R² scores of 0.97 and 0.98. Notably, CD14 Mono cells, dendritic cells, and FCGR3A+Mono cells collectively exhibited outstanding R² values of 0.99 and 0.99, further underscoring the reliability of predictions. This comprehensive evaluation delineates the remarkable predictive performance of the model across various cell types, shedding light on the intricacies of gene expression patterns and offering insights of potential significance in diverse biological contexts.

Figure 3b presents the dot plot of scPerb using the PBMC-Zhang dataset, revealing scPerb's adept predictive capabilities. Notably, this figure illustrated scPerb's ability to yield accurate predictions even when the mean gene expression of control cells remains low while that of stimulated cells is significantly higher. A compelling illustration of this can be observed in the case of IFI6, one of the most differentially expressed genes (DEGs) within the PBMC-Zhang dataset. In this case, the mean expression of the control cells is very close to 0, compared with that of over 0.6 in the stimulated cells. However, scPerb still delivers impressive predictions, as evidenced by the close alignment of its predictions with the actual stimulated cells, as depicted in Figure 3b.

To reinforce this point, the utilization of UMAP visualizations (depicted in Fig. 3c and Fig. 3d) further substantiates the argument. The UMAP projections distinctly reveal that control cells exhibit significantly lower mean expression levels compared to the mean expression levels observed in stimulated cells; but even in such scenarios, scPerb's predictions still exhibit a strong correlation with the stimulated cells, thereby underscoring its exceptional performance.





5.4 scPerb has robust results across different datasets.

scPerb has robust results in multiple datasets. In PBMC-Kang dataset [7], scPerb still outperforms other methods, achieving 0.98 in the mean of all the cell types, followed by scGen with a of 0.96, CVAE with 0.91, style transfer GAN with 0.42 and sc-WGAN with 0.12 (Fig. 4a). Moreover, scPerb precisely predicted the result of FCGR3A Mono cells, reaching of 0.9948 and 0.9978 respectively for all genes and its top 100 DEGs (Fig. 4b). Meanwhile, it is important to note that alternative benchmark methods, including scGEN, sc-WGAN, and style-transfer GAN, exhibited lower values in both the overall gene population and the top 100 differentially expressed genes (DEGs). To provide specific figures, scGEN yielded values of 0.9623 and 0.9545 for all genes and the top 100 DEGs, respectively, while sc-WGAN displayed values of 0.3303 and 0.8593 for the same categories, and style-transfer GAN yielded values of 0.5223 and 0.7361, respectively. This scatter plot further proved the strong prediction ability of scPerb. Moreover, in TNSF10, one of the top DEGs in FCGR3A Mono cells, which also has a control condition filled with zero values, scPerb made a better prediction than any other method, capturing the mean of the ground truth. In this case, the prediction of other methods barely captured the mean of the ground truth. (Fig. 4b) The Wilcoxon test can further explain the difference between the prediction and the real stimulated cells in the TNSF10 gene: only scPerb achieved a P value of 0.147, meaning that there is no statistically significant difference between the prediction of scPerb and the ground truth; however, all other methods including scGEN, CVAE, and both GAN-based methods resulted in a P value far less than 0.05, showing a significant difference between their predictions and the ground truth. (Fig. 4e) Besides, the dot plot (Fig. 4c) shows that scPerb can get robust prediction no matter whether the original control gene expression is lower (for example the IFIT1 gene), approximately the same (for example the RPL13A gene), or higher than (for example the FTH1 gene) the ground truth.

In alternative datasets, such as the Hpoly dataset, the robust predictive capacity of scPerb still remains. For most of the cell types in the Hpoly dataset, scPerb gained a higher or equal to 0.95, and has a significantly better result than GAN-based methods such as the sc-WGAN and the st-GAN (Fig. 5a). Moreover, scPerb made better predictions in this dataset, especially in Stem cells. In Figure 5b-5f, the UMAP of all five methods clearly showed this point. In these figures, the prediction of scPerb (the green dots in the graph) fitted the best to the stimulated data (the orange dots in the graph), while other VAE-based methods such as scGEN and CVAE had a prediction that has approximately the same distance between the control data and the simulated data, resulted in a worse prediction than scPerb. Moreover, both GAN-based methods style transfer GAN and sc-WGAN totally failed to predict the stimulated data, resulting in a prediction far away from the ground truth (Fig. 5b-5f).

**Discussion**

scPerb is a generative model that dynamically transfer the gene expression in the control dataset into the reliable stimulation dataset. The encoder of scPerb projected the raw control gene expression into the high-dimensional latent space, and aggregate with the dataset-specific styles to generate high-quality representation for the stimulation dataset. Based on the representation, the decoder from scPerb can reconstruct gene expression that are correlated with the mean of the stimulation dataset. The experiments demonstrated that scPerb can capture the latent content features and generate stable dataset-specific styles across different cell types and data from multiple studies. Moreover, quantitative evaluation indicated the performance of scPerb outperform five representative benchmarkings, having state-of-the-art results in three different datasets.

Compared with the traditional works [1-4], scPerb is a data-driven algorithm that can fully explore the gene expression in the raw dataset, and didn’t rely on solid domain priors. On the opposite, the traditional works extract the principle components and build up a graph-base model in the low-dimensional manifold. Such methods relied heavily on the experienced domain knowledge, and lack of generalization abilities. Compared with other data driven algorithms, scPerb incorporates the stableness from the VAE settings, and exploit the advantage from the GAN to generate high-quality samples.

**Reference**

1. Ghahramani, A., F.M. Watt, and N.M. Luscombe, *Generative adversarial networks uncover epidermal regulators and predict single cell perturbations.* bioRxiv, 2018: p. 262501.

2. Karras, T., S. Laine, and T. Aila. *A style-based generator architecture for generative adversarial networks*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

3. Lotfollahi, M., F.A. Wolf, and F.J. Theis, *scGen predicts single-cell perturbation responses.* Nature Methods, 2019. **16**(8): p. 715-721.

4. Cortes, C., et al. *Advances in neural information processing systems 28*. in *Proceedings of the 29th Annual Conference on Neural Information Processing Systems*. 2015.

5. Kingma, D.P. and M. Welling, *Auto-encoding variational bayes.* arXiv preprint arXiv:1312.6114, 2013.

6. Zheng, G.X., et al., *Massively parallel digital transcriptional profiling of single cells.* Nature communications, 2017. **8**(1): p. 14049.

7. Kang, H.M., et al., *Multiplexed droplet single-cell RNA-sequencing using natural genetic variation.* Nature biotechnology, 2018. **36**(1): p. 89-94.

8. Haber, A.L., et al., *A single-cell survey of the small intestinal epithelium.* Nature, 2017. **551**(7680): p. 333-339.

9. Virtanen, P., et al., *SciPy 1.0: fundamental algorithms for scientific computing in Python.* Nature methods, 2020. **17**(3): p. 261-272.

10. Wolf, F.A., P. Angerer, and F.J. Theis, *SCANPY: large-scale single-cell gene expression data analysis.* Genome biology, 2018. **19**: p. 1-5.

11. Cuzick, J., *A Wilcoxon‐type test for trend.* Statistics in medicine, 1985. **4**(1): p. 87-90.