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 [1], a treatment [2, 3], or a modification of genes [4-6].

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 [7], Nature Language Processing (NLP) tasks [8] 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) [9] and Variational Auto-Encoders (VAE) [10] 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 [11] transferred the more stable WGAN to the single-cell perturbation and st-GAN [12] 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 [13] 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) [14] 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 [10] and style-transfer GAN [12]. 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 [12], 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) [10] and the style-transfer GAN (stGAN) [12]. 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 [12], 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. [15]. 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 [1]. 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* [2]. 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**

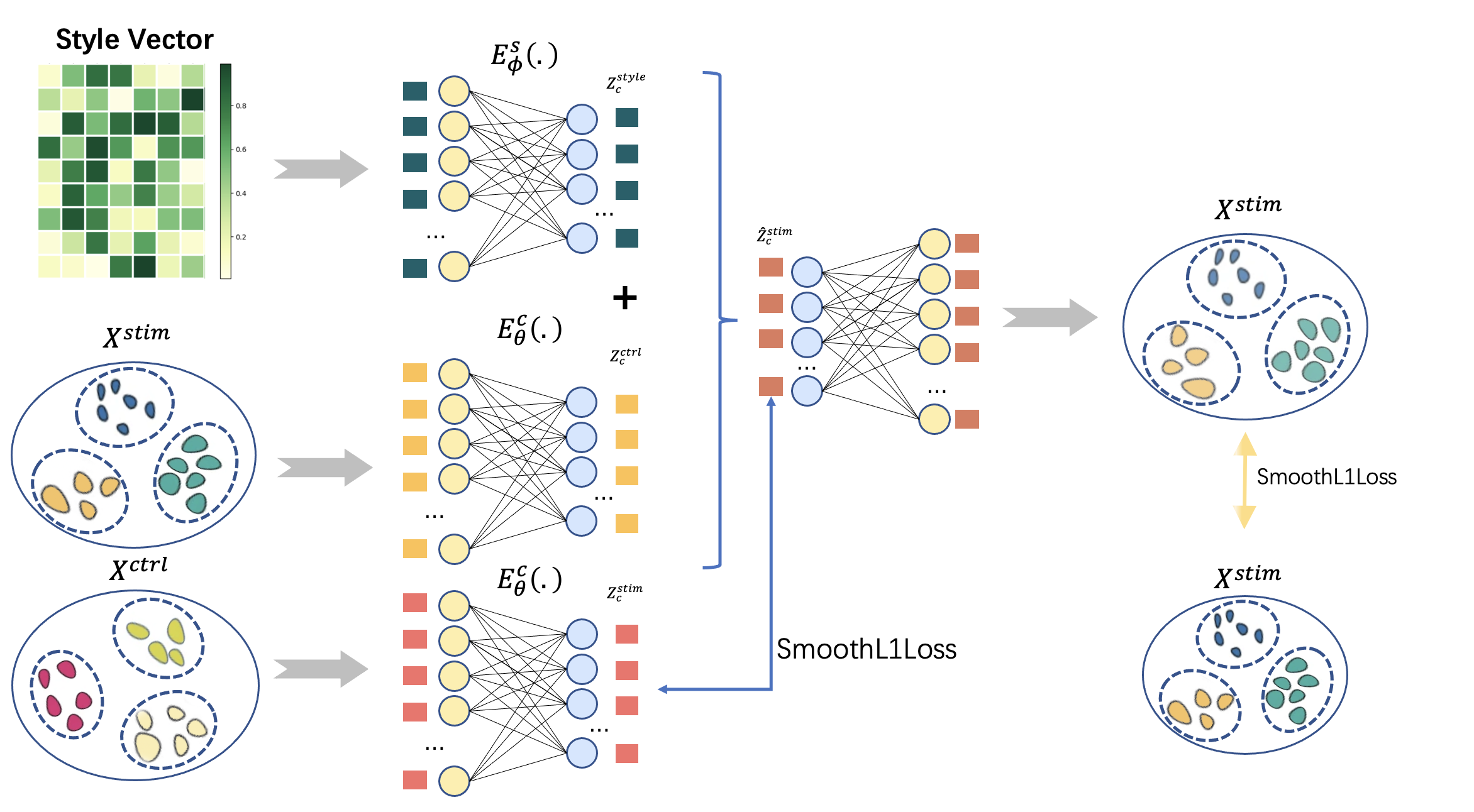
In scPerb, we evaluated the performance of our model under a fixed seed of 42 by using the square of the , which is calculated by the function [16]. This metric measures the correlation between the generated images and the ground truth data. We computed the values for all genes’ mean and variance and 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 [17] function, employing the Wilcoxon method [18].

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**



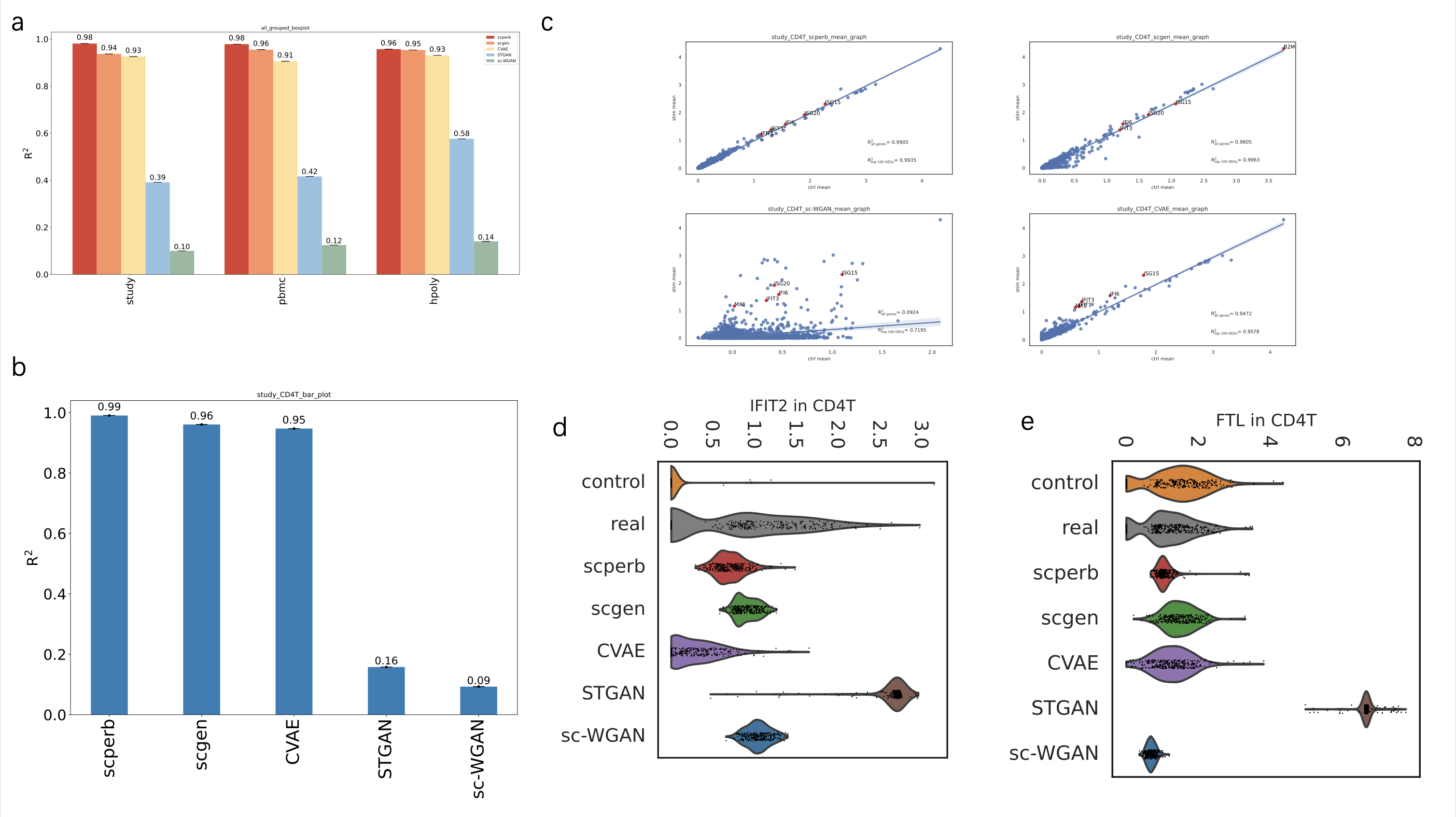
5.1 scPerb outperforms other benchmarks

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 [1, 15] stimulated with interferon () methods, and the rest is a group of intestinal epithelial cells fetched by parasitic helminth *H.poly* cells [2]. To differentiate the two PBMC datasets, for the rest of the manuscripts, we denote them as PBMC-Kang and PBMC-Zheng datasets, and PBMC-Zheng [15] was a mixed published dataset from two different studies.

We fairly compared our proposed scPerb with other benchmarking papers [11-14]. In this process, we select a target cell type and use the rest of the cell types to train a model to transfer the gene expression from the control dataset to the stimulation dataset. We evaluate the model performance using the between the prediction and the ground truth simulation dataset. We repeated such process for all cell types and presented the average of the in Fig. 2(a). In the study dataset, achieving the average score of 0.98, which was better than the performance of the competitors, including scGEN (average 0.94), CVAE (average = 0.93), st-GAN (average = 0.39) and sc-WGAN (average = 0.10). Meanwhile, in the PBMC dataset, our proposed method achieved the highest average score of 0.98, while the second-best and third-best approaches scGEN and CVAE had 0.96 and 0.91. Surprisingly, the GAN-based methods had much worse performance. The st-GAN only got an average score of 0.42 and sc-wGAN only had 0.12. Finally, we applied scPerb to the hpoly dataset, our proposed scPreb still got a 0.96 average score, followed by the scGEN, CVAE, st-GAN, and sc-wGAN with the average score of 0.95, 0.93, 0.58, 0.14. When comparing the result in a specific cell type, scPerb still outperforms other benchmarking methods (Fig. 2b). For example, in CD4-T, one of the most numerous cell types in the PBMC-Zheng dataset, scPerb achieved an outstanding score of 0.99, which is much than scGEN, CVAE, stGAN, and sc-WGAN, having score 0.96, 0.95, 0.16, and 0.09 respectively. (这里需要增加CD4-T的作用)

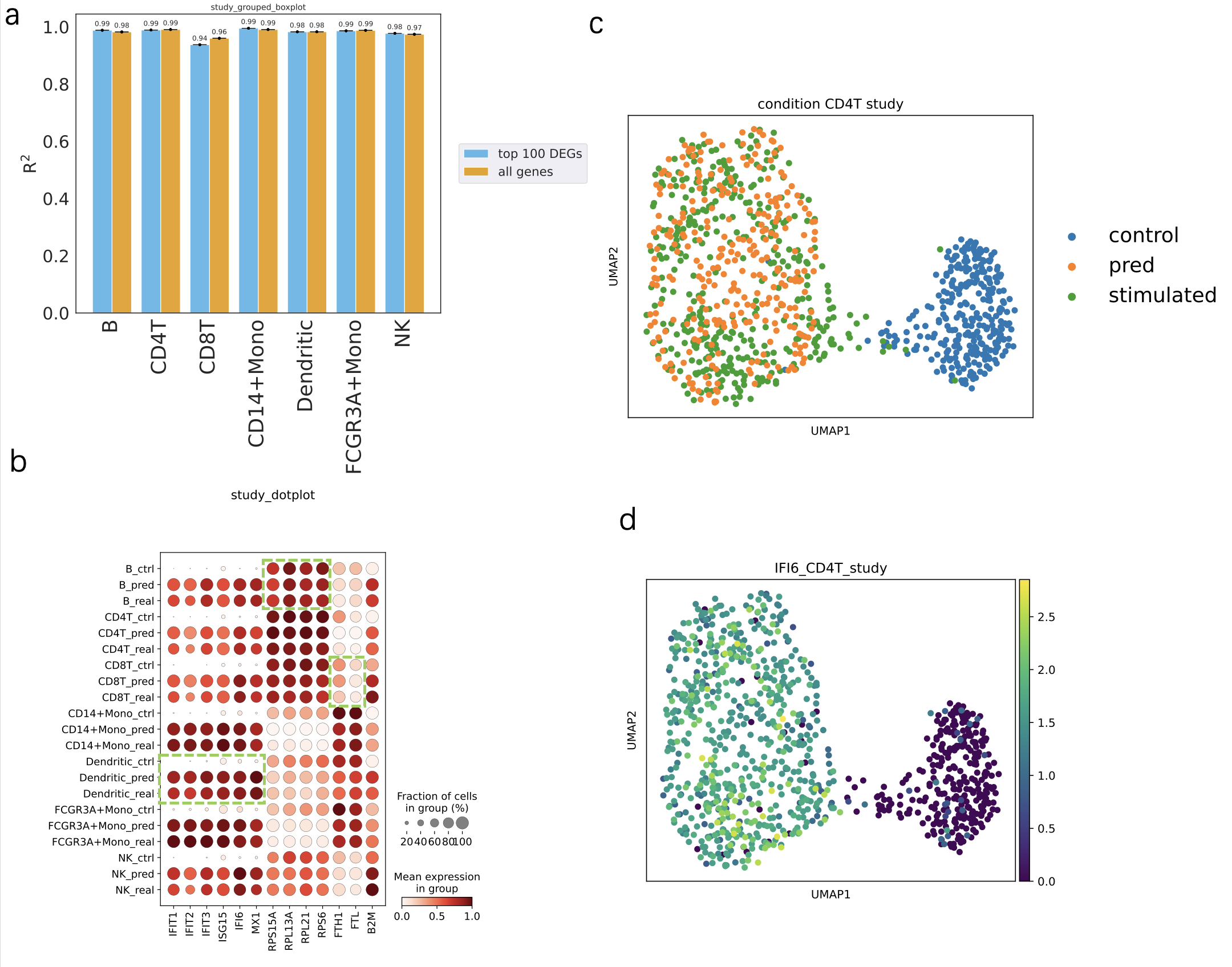
In particular, we evaluate the performance of the proposed scPerb and the benchmarking methods at the gene level. In Fig. 2c, we illustrated the prediction of our scPerb and the performance of the other three benchmarking methods in CD4T from the PBMC-Zheng dataset. The scatter plot demonstrated scPerb got the average score of 0.9905 when we used all the genes in this cell type. The performance can go up to 0.9935 when we only consider the top 100 DEGs. In comparison under the same setting, scGEN achieved the average score of 0.9605 over all genes and 0.9963 on the top 100 DEGs. Our scPerb can outperform CVAE (average score of all genes=0.9472, average score of top 100 DEGs =0.9578) and sc-WGAN (average score = 0.0924, average score=0.9578) on both the evaluation metrics. In particular, we annotated the top-5 best performance DEGs, which were IFIT1, IFIT3, IF16, ISG20, and ISG15. (可能需要增加DEG的作用的文献)

To be more specific, the violin plot in Fig. 2d and Fig. 2e indicated two representative scenarios in the dataset. (这里需要增加IFIT2 和 FTL 对于CD4T 作用的参考文献，说明为什么选择这两个基因) In Fig. 2d, the distribution of IFIT2 in the control dataset varied from the distribution in the simulation dataset. In the prediction results, our scPerb was closest to the mean of the stimulation dataset, with a relatively large value range between 0.5 to 3.0 after log transformation. The scGEN and sc-WGAN provided comparable predictions, but the mean of the predictions was slightly larger than the real stimulation data. In this particular gene, CVAE was more associated with the control data, and stGAN focused on the outliers with high gene expressions. In Fig. 2e, the distribution pattern of FTL in the control dataset was similar to the distribution in the real stimulation dataset. Under such a scenario, most of the predictions in scPerb were close to the mean of the simulation data, while the predictions from scgen and CVAE were expanded to a larger range(这里想说明scgen 和 cvae 的预测值更发散，而scperb 更集中). St-GAN responded to the high gene expressions while the predictions in sc-WGAN had lower gene expressions.



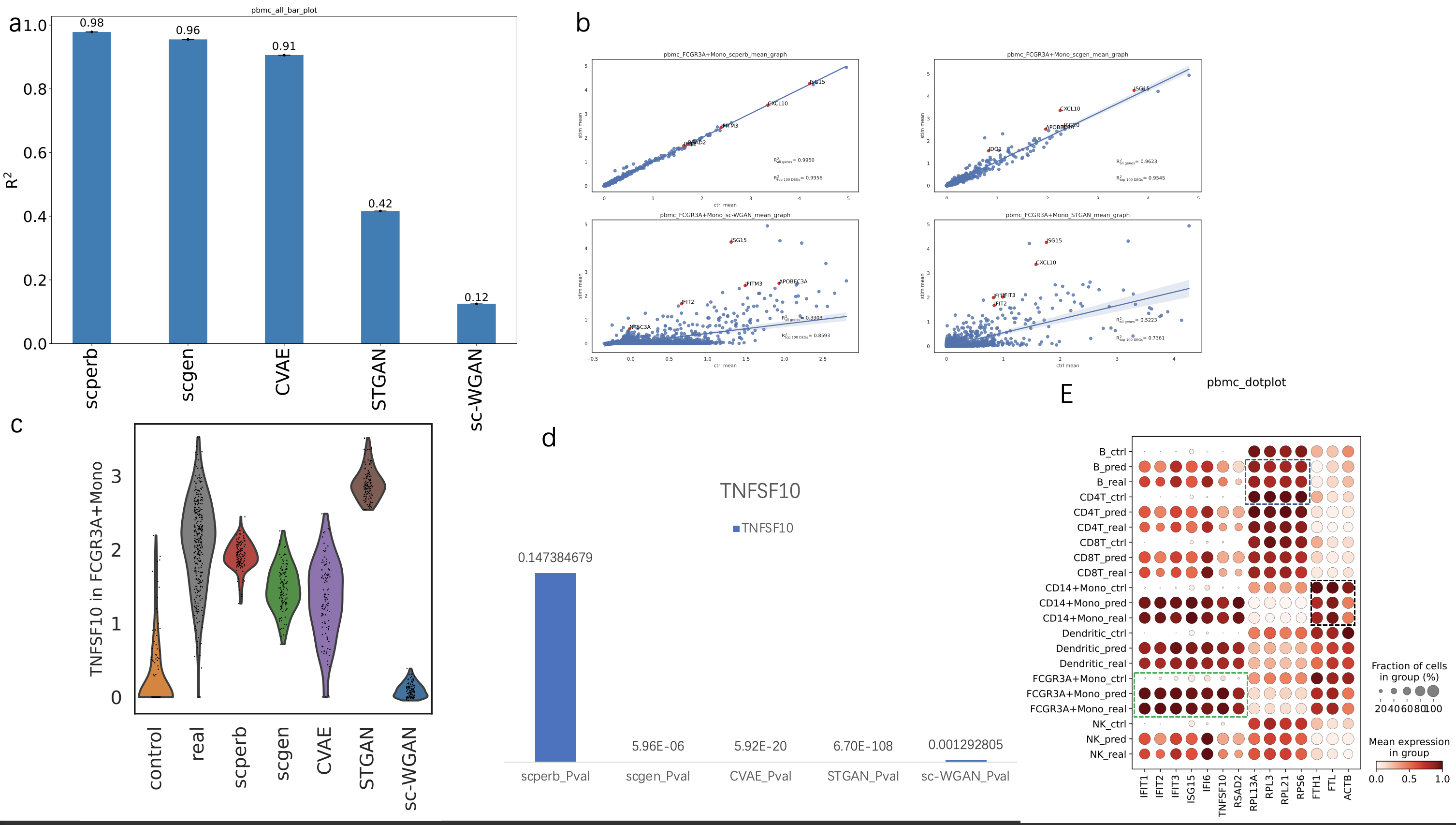
5.2 scPerb is an innovative generative model that can accurately predict single-cell perturbation responses.

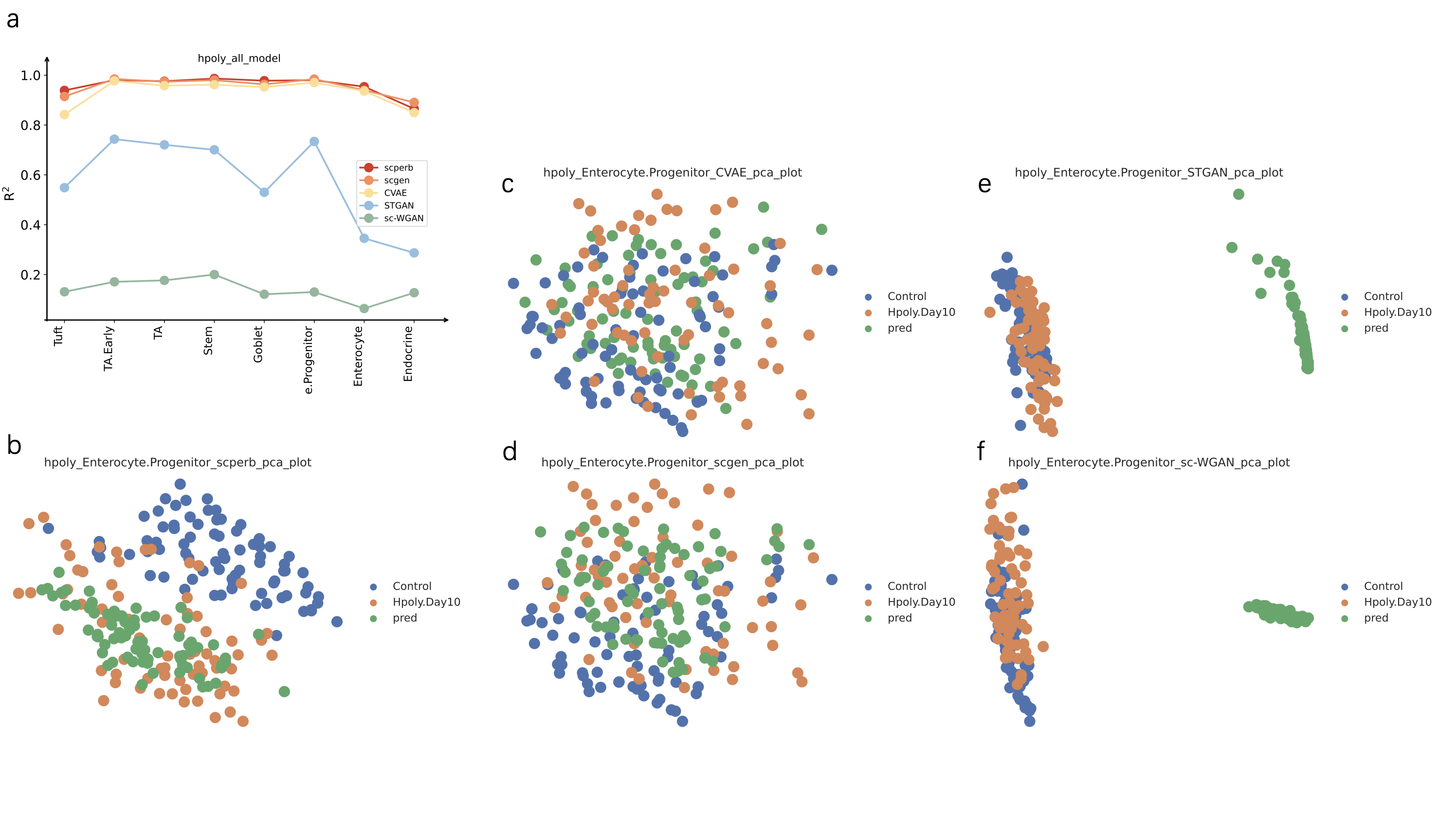
In this section, we aimed to show scPerb can accurately predict the single-cell perturbation responses for other cell types. Fig. 3a summarized the performance of scPerb over difference cell types. In CD4T, CD14+Mono, FCGR3A+Mono, scPerb can achieve average score=0.99 in both the top 100 DEGs and all gene expressions. In Dendritic, the average score was 0.98 and 0.98 accordingly. In B cells and NK cells, the performance of top 100 DEGs was slightly better than the performance on all genes, which was 0.99 vs. 0.98 and 0.98 vs. 0.97 respectively. We also observed in CD8T, the performance of top 100 DEGs was 0.94, which was slightly lower than the performance on all genes (average score=0.96). In Fig. 3b, the dot plot demonstrated the correlation of representative genes among different cell types. In half of the selected genes, the dot plot showed a strong difference between the gene expression and the real stimulation gene expression. On the other half of the selected genes, we presented similar gene patterns in both the control dataset and stimulation dataset. In the green dashed rectangle box, we highlighted the mean of the expression in the control, predicted and real stimulation dataset. Fig. 3b implied that the mean gene expression of B cells, CD8T, Dendritic in our scPerb prediction was associated with the mean gene expression in the real stimulation dataset. The umap in Fig. 3c showed that the predicted gene expression from scPerb in CD4T was correlated with the real stimulation gene expression in the latent space. In particular, for a specific gene IFI6, we also illustrated the consistent observation.



5.3 scPerb can accurately predict the perturbation of cells in PBMC dataset

scPerb has robust results in multiple datasets. In PBMC-Kang dataset [1], 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, st-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 R^2 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\cite[], 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 0.0001, 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. \cite[genes]





5.4 scPerb has robust results across different datasets.

In alternative datasets, such as the Hpoly dataset[2], 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, which is better than other VAE-based methods 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 Fig. 5b the distance between the prediction (green dot) and real stimulation data (orange dot) was closer than the distance between the stimulation dataset to the control dataset (blue dot). For the other benchmarks, the VAE-base methods, scGEN in Fig. 5c and CVAE in Fig. 5d can’t easily divide the control data samples from the prediction and stimulation data. And the GAN-based approaches, the st-GAN in Fig. 5e and sc-wGAN in Fig. 5f indicated the prediction was far away from both the control and stimulation datasets.

**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 [11-14], 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. Kang, H.M., et al., *Multiplexed droplet single-cell RNA-sequencing using natural genetic variation.* Nature biotechnology, 2018. **36**(1): p. 89-94.

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

3. Hagai, T., et al., *Gene expression variability across cells and species shapes innate immunity.* Nature, 2018. **563**(7730): p. 197-202.

4. Dixit, A., et al., *Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens.* cell, 2016. **167**(7): p. 1853-1866. e17.

5. Adamson, B., et al., *A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response.* Cell, 2016. **167**(7): p. 1867-1882. e21.

6. Datlinger, P., et al., *Pooled CRISPR screening with single-cell transcriptome readout.* Nature methods, 2017. **14**(3): p. 297-301.

7. He, K., et al. *Deep residual learning for image recognition*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

8. Vaswani, A., et al., *Attention is all you need.* Advances in neural information processing systems, 2017. **30**.

9. Goodfellow, I., et al., *Generative adversarial nets.* Advances in neural information processing systems, 2014. **27**.

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

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

12. 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.

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

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

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

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

17. 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.

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