A pre-trained model for cellular responses to drug perturbations

Yuhao Tan[1,a], Shukai Wang[1,a], Ming Zhang[2] [1]Yuanpei School, Peking University, Beijing, China
[2]Department of Computer Science, School of EECS, Peking University, Beijing, China
tanyuhao,shukaiwang,mzhang_cs@pku.edu.cn

a These authors have contributed equally to this work.

Abstract. Measuring single-cell RNA sequencing (scRNA-seq) responses to drug perturbations can facilitate drug development. However, an exhaust exploration of countless drugs is experimentally unfeasible, so computational tools are necessary to predict perturbations of unmeasured drugs. Existing tools are not designed for predicting scRNA-seq of unseen drugs. Here, we present a pre-trained model, which combines BERT, a wildly used pre-trained model, with autoencoder. The model learns transcriptional drug responses from bulk RNA sequencing in pre-training stage and can adapt to scRNA-seq in finetuning stage. It can predict scRNA-seq responses to drugs unseen in finetuning stage and model the heterogeneity of cellular responses.

Keywords: pre-trained model · scRNA-seq · BERT · perturbation.

1 Introduction

Measuring and predicting cellular responses to perturbation could facilitate drug development[12] and help understand cell pathways[2]. Single-cell RNA sequencing (scRNA-seq) captures the cellular heterogeneity of responses and provides comprehensive phenotyping, which could not be identified with conventional methods[9].

While the development of high-throughput approaches lowers the experimental costs, it is still not easy to measure all perturbations in all cell lines due to the expensive library preparation[9] and countless conditions. Thus, several computational approaches have been developed to predict RNA sequencing (RNA-seq) responses to perturbations. Compositional perturbation autoencoder (CPA)[6] uses autoencoder and adversarial training to learn perturbation embeddings, which can be used to predict responses of new unseen drug, dose, and cell line combinations. Though achieving rather high performance, it cannot predict the responses of drugs that have not been seen in the training set, which can help augment responses data further. Drug structure has been used to predict bulk RNA-seq of drugs unseen in training set[13], but it has not been applied to scRNA-seq.

We present a pre-trained model to predict scRNA-seq responses to perturbation of unseen drugs by leveraging the existing massive amount of bulk RNA-seq data. We pre-train our model on bulk RNA-seq perturbation responses and finetune our model on single-cell RNA-seq data. Then we test our model on drugs that have only been seen in the pre-train step but not in finetune step. In this manner, we manage to transfer the knowledge learned from bulk RNA-seq to single-cell RNA-seq. Our model can predict the unseen drugs responses and capture the heterogeneity of scRNA-seq data.

2 Background

2.1 Transformer

The Transformer architecture is usually developed by stacking Transformer layers [11]. A Transformer layer operates on a sequence of vectors and outputs a new sequence of the same shape. The computation inside a layer is decomposed into two steps: the vectors first pass through a (multi-head) self-attention sub-layer and the output will be further put into a position-wise feed-forward network sub-layer. Residual connection [4] and layer normalization [1] are employed for both sub-layers.

2.2 Encoder-decoder architecture

As a basic algorithm in deep learning, encoder-decoder has been widely used in the analysis of scRNA-seq data. SAVER-X [quote] integrates the scRNA-seq expression data via autoencoder and statistical learning method. SAUCIE (Sparse Autoencoder for Clustering, Imputing, and Embedding) [quote] first explores the multitask learning based on the embeddings of autoencoder. scGen [quote] has tried to represent the sparse scRNA-seq data by some low-dimension probability distribution via variational autoencoder.

2.3 BERT

Recently, BERT [3] has been successfully used in various natural language processing tasks, such as textual entailment, name entity recognition, and machine reading comprehensions. With the progress in natural language processing (NLP), extracting valuable information from all kinds of literature has gained popularity among researchers, and deep learning has boosted the development of effective text mining models. However, in bioinformatics, although transcriptional drug responses prediction is similar to NLG(natural language generation) task, there are not many effective methods in this area. Instead of directly applying the advancements in NLP, we modified the original BERT model to make it suitable for our tasks.

3 Method

3.1 Problem Definition

Each of the data contains 4 properties: drug type d_i,cell type c_i, dosage t_i and gene expression series x_i, where i is the perturbation id. Our goal is to generate the unknown gene expression series for drugs unseen in training dataset. More precisely, we consider a transductive learning setting where the dataset V composes of V_{train} and V_{test}. V_{train} is the set of tuples that have both conditions and gene expression (d_i, c_i, t_i, x_i). V_{test} is the set of tuples that only have conditions (d_i, c_i, t_i). The goal of scRNA-seq generation is to generate x_i for the conditions $\in V_{test}$ according to the conditions and gene expression in the V_{train}.

3.2 Model

Our model generates scRNA-seq based on the perturbation conditions and gene expression using a encoder-decoder architecture with a pre-trained and a finetune stage. At the pre-training stage, we use BERT-liked model to calculate the gene expression embedding on a large dataset. At the finetuning stage, gene embedding is calculated by the transformer blocks encoder of the pre-trained BERT-liked model, and the conditional embedding is calculated by the conditional encoder. Then, the input embedding is the sum of the gene expression embedding and conditional embedding. Finally, Our model generates the gene expression x_i by using the embedding as the input to a Long short-term memory (LSTM) decoder [5].

Pre-training stage At the pre-training stage(Fig. 1), we modified the BERT base model, using a two-stage approach to train the gene expression embedding. At the first stage, we randomly mask a fixed ratio of the gene expression and add a linear layer to match the dimension of the input to the transformer latent space. At the second stage, we use four tasks to train the embedding. For the conditions, we propose cell classification, drug classification, and dosage prediction tasks to embed the cell, drug, and dosage conditions; for the gene expression, the model learns to predict the true value for the masked expression.

Finetuning stage The finetune model is shown in Fig. 2. First, we obtain condition embedding from (d_i, c_i, t_i) through conditional encoders composed of Multilayer Perception (MLP) and obtain gene expression embedding from x_i through transformer blocks encoder, respectively. Particularly, we follow the previous work [3] to calculate the input embedding. Then, we use a sequential decoder to decode the input embedding. We examined several sequential decoder in the experiments, and showed that LSTM performs best in this condition.

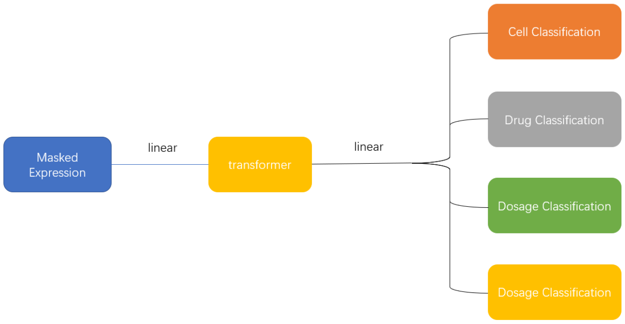

Fig. 1. Pre-training stage.

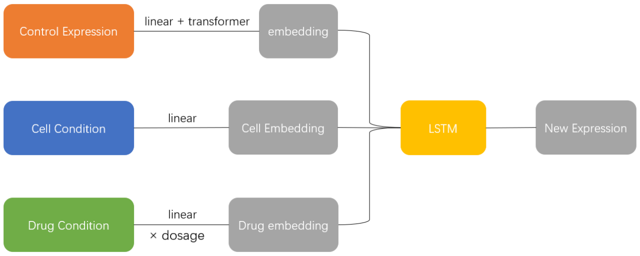

Fig. 2. Finetuning stage.

3.3 Evaluation of the distribution

Previous evaluation of the task often focuses on the correlation between real and predicted means and covariances[8, 7], with additional importance placed on differentially expressed (DE) genes. However, the metrics cannot evaluate the heterogeneity of scRNA-seq. We use $mean_i(max_j(pearsonr(x_i^t, x_j^p)))$ to evaluate the distribution between real expression $\{x_i^t\}$ and predicted expression $\{x_i^p\}$, where pearsonr means Pearson correlation. If the distribution of predicted expression can cover the distribution of true expression, i.e. $\{x_i^t\}_i \in \{x_i^p\}_i$, the metric will be high, demonstrating that the model predicts all true expression vectors successfully. Our metric considers the heterogeneity of cellular responses, which is a more suitable metric for scRNA-seq.

4 Results

4.1 pre-train

We pre-train our model on The Library of Integrated Network-Based Cellular Signatures (LINCS)[10]. We try two pre-train setting. In the first setting, we pre-train on four tasks: cell type classification, drug classification, dose prediction, and mask prediction. In the second setting, we only pre-train on mask prediction.

We pre-train on two datasets: the whole LINCS datasets and a curated subset of LINCS dataset. We test on multiple mask percentage. The cell accuracy, drug accuracy, dose root mean square error (RMSE), and mask RMSE are shown in Table 1.

Table 1. Metrics of pre-trained model. cell_acc means the predicted accuracy of cell, drug_acc means the predicted accuracy of drug, dose_rmse means the RMSE of dose prediction, and mask_rmse means the RMSE of masked gene expression. Item with '-' is not reported since the model is not trained for these tasks.

methods	mask_percentage	dataset	cell_acc	drug_acc	dose_rmse	mask_rmse
one-task	80	subset	-	-	-	1.07
one-task	15	whole	-	-	-	0.52
four-task	15	subset	95.49%	21.39%	1.55	0.64

4.2 Finetune

We finetune and test our model on sciplex3[9]. We use four baseline. First, we random select responses of five conditions to predict the responses (denoted as random). Second, we use five conditions that have the most similar gene expression in LINCS to predict the response (denoted as kNN). Third, we use CPA. Since CPA has not seen LINCS dataset, we try to make it comparable to our model. We train CPA on a small set of LINCS which only contains cell lines and drugs in sciplex3 dataset for 50 epochs, then we train it on sciplex3 (denoted as CPA_transfer). Fourth, we use conditional Variational autoencoders (cVAE).

We use BERT without pre-training (denoted as BERT), BERT trained on the small set (denoted as BERT_transfer), BERT pre-trained with whole dataset for one task (denoted as BERT_whole), and BERT pre-trained with the subset for four tasks (denoted as BERT_subset). Table 2 shows the metrics. Our pre-trained models achieve similar performance, and are better than model without pre-training on most metrics. Our model performs better than all baseline in terms of distribution, revealing that our model captures the heterogeneity of cellular responses in scRNA-seq.

We try to use MLP as decoder and find that LSTM outperform MLP (table 3), so we use LSTM as decoder. Our model is much faster than CPA: CPA needs 4.41 min for an epoch, while our model only needs 0.38 min for an epoch.

5 Conclusions and Future work

Perturbation response is both a fundamental biological question and a clinical question. Over 30,000 drugs have been tested for bulk RNA-seq responses, while only about 200 drugs have been tested for scRNA-seq. It is meaningful to transfer

Table 2. Pearson correlation of unseen drugs responses. Mean is the Pearson correlation between real and predicted means, and dis is distribution metrics. DE_50 means the top 50 DE genes, and DE_10 means the top 10 DE genes. Epoch means the epoch with highest performance in test dataset.

method	mean	mean_DE50	mean_DE10	dis	dis_DE50	dis_DE10	epoch
random	0.970	0.953	0.904	0.649	0.786	0.868	-
kNN	0.981	0.968	0.956	0.719	0.844	0.923	-
CPA	0.985	0.961	0.939	0.734	0.800	0.768	50
CPA_transfer	0.979	0.947	0.926	0.730	0.779	0.770	28
cVAE	0.968	0.953	0.915	0.716	0.781	0.719	3
BERT	0.920	0.907	0.879	0.740	0.814	0.855	1
BERT_transfer	0.920	0.907	0.884	0.738	0.813	0.862	1
BERT_whole	0.966	0.929	0.904	0.720	0.812	0.871	28
BERT_subset	0.964	0.931	0.910	0.719	0.818	0.879	31

Table 3. Results using MLP as decoder

method	mean	mean_DE50	mean_DE10	dis	dis_DE50	dis_DE10	epoch
BERT	0.958	0.923	0.877	0.716	0.807	0.854	14
BERT_whole	0.955	0.921	0.900	0.716	0.806	0.856	13
BERT_subset	0.965	0.929	0.897	0.703	0.809	0.865	36

knowledge from bulk responses to single-cell responses, as scRNA-seq has higher resolution and more accurate information.

We have introduced a new pre-trained model to make use of bulk responses for single-cell perturbation responses. Since our one-task pre-trained model and four-task pre-trained model have similar performance, we suspect that our model does not learn the drug similarity successfully, which is essential for our task. For our future work, we will try other architectures and pre-trained tasks to model drug similarities explicitly.

In most experiments, our pre-trained models either achieve higher performance or converge faster, implying that our model learns the basic gene expression logic. It implies that our model may work in other situations including cell type identification, where gene expression logic is more useful.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization (2016)
2. Chen, W., Zhou, X.: Drug effect prediction by integrating l1000 genomic and proteomic big data. In: Bioinformatics and Drug Discovery, pp. 287–297. Springer (2019)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
4. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks (2016)

5. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation **9**(8), 1735–1780 (11 1997). https://doi.org/10.1162/neco.1997.9.8.1735, https://doi.org/10.1162/neco.1997.9.8.1735
6. Lotfollahi, M., Klimovskaia, A., De Donno, C., Ji, Y., Ibarra, I.L., Wolf, F.A., Yakubova, N., Theis, F.J., Lopez-Paz, D.: Learning interpretable cellular responses to complex perturbations in high-throughput screens. bioRxiv (2021)
7. Lotfollahi, M., Naghipourfar, M., Theis, F.J., Wolf, F.A.: Conditional out-of-distribution generation for unpaired data using transfer vae. Bioinformatics **36**(Supplement_2), i610–i617 (2020)
8. Lotfollahi, M., Wolf, F.A., Theis, F.J.: scgen predicts single-cell perturbation responses. Nature methods **16**(8), 715–721 (2019)
9. Srivatsan, S.R., McFaline-Figueroa, J.L., Ramani, V., Saunders, L., Cao, J., Packer, J., Pliner, H.A., Jackson, D.L., Daza, R.M., Christiansen, L., et al.: Massively multiplex chemical transcriptomics at single-cell resolution. Science **367**(6473), 45–51 (2020)
10. Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K., et al.: A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell **171**(6), 1437–1452 (2017)
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2017)
12. Yofe, I., Dahan, R., Amit, I.: Single-cell genomic approaches for developing the next generation of immunotherapies. Nature medicine **26**(2), 171–177 (2020)
13. Zhu, J., Wang, J., Wang, X., Gao, M., Guo, B., Gao, M., Liu, J., Yu, Y., Wang, L., Kong, W., et al.: Prediction of drug efficacy from transcriptional profiles with deep learning. Nature Biotechnology pp. 1–9 (2021)