Attention Based Models for Cell Type Classification on Single-Cell RNA-Seq Data

1 More Descriptions of Datasets

Table 1 presents detailed descriptions of scRNA-seq datasets. Nine datasets cover cell samples from various cell types in human and mouse. Except for PBMC which is profiled from a single donor, all other datasets are profiled from multiple donors. Since the cells are randomly split into the training set and the test set, the labeled cells may come from different donors, which ensures the generalization ability of the model.

Table 1: Detailed Descriptions of Datasets

Dataset	Dataset Description	Number of Donors in Dataset
CRC	T cells from colorectal cancer patients	12
GSE70580	Human tonsil innate lymphoid cells	3
GSE72056	Melanoma tumor cells	19
GSE75688	Breast cancer cells	11
GSE96993	Human CD8+ T cells	human and mice
NSCLC	T cells from non-small cell lung cancer patients	14
PBMC	Peripheral blood mononuclear cells	1
Spleen human	Human splenic dendritic cells	2
Spleen mouse	Mouse splenic dendritic cells	2

2 Details of Data Preprocessing

For every cell in the data, we first divide the gene's expression level by the sum of gene expression levels in the cell. This makes genes' expression levels among cells comparable. Then, gene expression levels are multiplied by a constant 100000 and employed log transform in order to adjust them into a reasonable range. After that, we calculate the variances of genes and then pick out the top 1000 highest among them. Highly variable genes are more likely to help us distinguish different cells as they provide more information on cell heterogeneity. Also, there are tens of thousands genes in human and mouse genome. Choosing the top 1000 most variable genes also serves as dimension reduction of high dimensional scRNA-seq data.

3 Details of Hyperparameters' Settings for CFAN and CGRAN

For CFAN, the convolution layer's kernel size is set to 5 and output channel is set to 8. During the training process, we apply cross entropy loss as the loss function and train CFAN for 50 epochs. The activation function in CFAN is ReLU. Adam optimizer is adopted with the learning rate of 0.0001.

For GD-based Matrix Factorization of CGRAN, dropout is set to 0.1 and learning rate is set to 0.005. The initialization process takes 130 epochs. In both attention blocks, we set the head number of local attention layer to 10. In the first attention block, the output embedding's dimension for each head is 8. Then we concatenate every head's output, which is the input for the second attention block. In the second attention block, the output embedding's dimension for each head is 4, and we concatenate every head's output. As a result, the second attention block's output embedding's dimension is 40. We apply cross entropy loss as the loss function and train CGRAN for 75 epochs. Adam optimizer is adopted with the learning rate of 0.0001.

4 Experimental Settings for Other Methods

We choose $linear\_kernel$ and set $max\_iteration$ to 1000 for SVM using scikit-learn implementation.

For random forest and XGBoost algorithm, we also adopt scikit-learn implementation. Number of trees (namely $n\_estimators$) in random forest is set to 1000 and other parameters are set as default configurations. For XGBoost, $n\_estimators$ is set to 100 and other settings follow the default configuration as well.

For ACTINN and scCapsNet, we follow the implementation instructions mentioned in their articles. For Cell Blast, different hyperparameters have been tested on different datasets and best results are presented in the article. For scVI, we apply the implementation provided by its authors. $Hiddendimensionality$ is set to 256 and $latentdimensionality$ to 128. The number of layers is set to 2 and dropout probability to 0.3. For Moana, we conduct experiments on different hyperparameters for different datasets and the best results are presented. For singleR, we use the default parameters in its implementation. For scmap, we lower the threshold of assigning cells to "unassigned" category so that cells will be classified as the known cell types. For scnym, we set batchsize to 32 since some of the datasets only contain a few hundreds of cells. For SCINA, we provide the common marker genes found in marker gene databases for different cell types. Lacking sufficient marker genes for all cell types in datasets Spleen_human and Spleen_mouse, we fail to evaluate SCINA on these datasets.

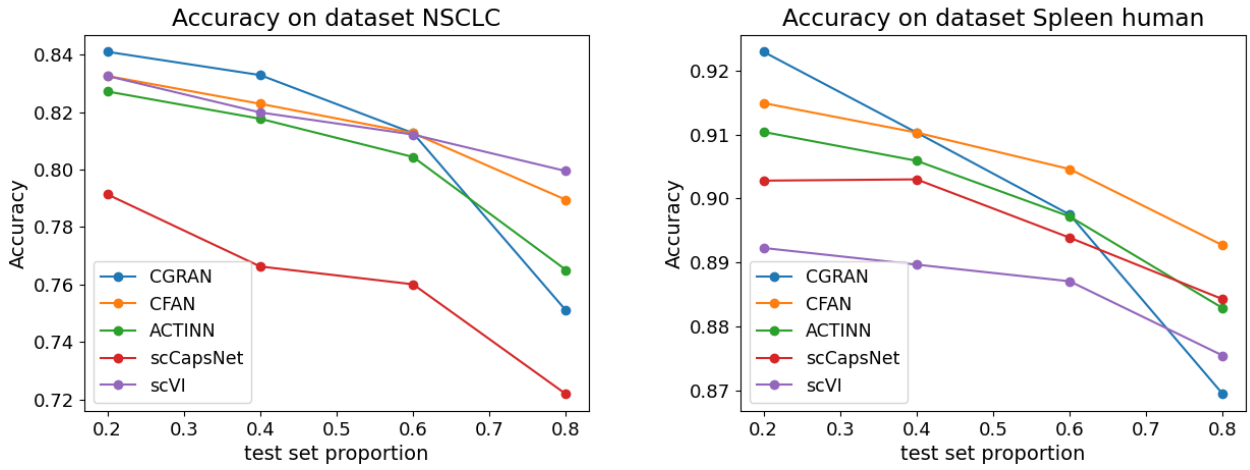

Figure 1: Classification accuracy with different test proportions on NSCLC, Spleen human

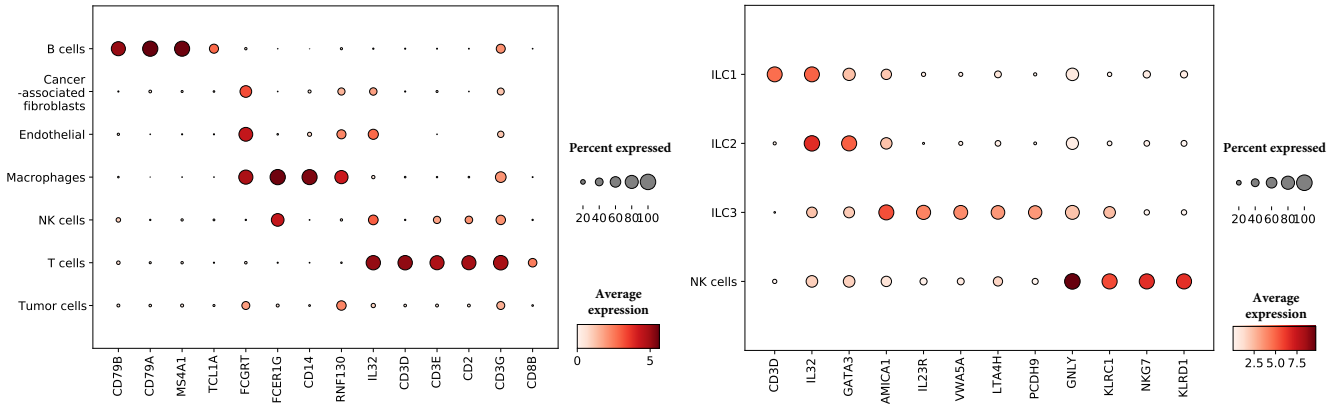

Figure 2: Highly differential genes chosen by attention weights of CGRAN in GSE72056(left) and GSE70580(right).

5 Different Splitting Ratio of Training and Test Set

In this section, we present experimental results on different ratio of training and test set on more datasets.

From Figure 1, we can see that both CFAN and CGRAN perform consistently well when test set proportion is smaller than 0.8. CGRAN's accuracy drops when the proportion of training set becomes smaller because of its deep architecture. As mentioned in the article, more data are needed for training appropriate cells and genes' embeddings as well as the attention weights among them. Also, it is not a common option to set training set proportion to a small value, such as 0.2, which will lead to the disappearance of cells from rare cell types in training set.

6 More About Interpretations of Models

6.1 Interpretations of CGRAN

6.1.1 Cell-Gene Attention and Discovery of Marker Genes

Here we illustrate the highly differential genes identified on GSE72056 and GSE70580 as well as their distinguishable expression levels in different cell types, as shown in Figure 2.

On dataset GSE72056, CD79B to TCL1A are highly differential genes of B cells. FCGRT is the highly differential gene for Endothelial. From FCER1G to RNF130 are highly differential genes for Macrophages. From IL32 to CD8B are highly differential genes for T cells.

In dataset GSE70580, ILC cells are cells that have close relationships with T cells. So some marker genes of ILC cells may be the same as T cells. CD3D is the highly differential gene for ILC1. IL32 and GATA3 are the highly differential genes for ILC2. From AMICA1 to PCDH9 are ILC3's highly differential genes. The rest are NK cells' highly differential genes.

These highly differential genes match well with the acknowledged marker genes.

6.1.2 Gene-Gene Attention Weights of CGRAN

More illustrations and analyses on attention weights among genes are provided in this section. Specifically, we select highly differential genes from PBMC and visualize their attention weights with the genes in the same group in uniform grouping. Figure 3 shows attention weights among highly differential genes for NK cells, B cells, T cells and the genes in the same group in PBMC dataset. It

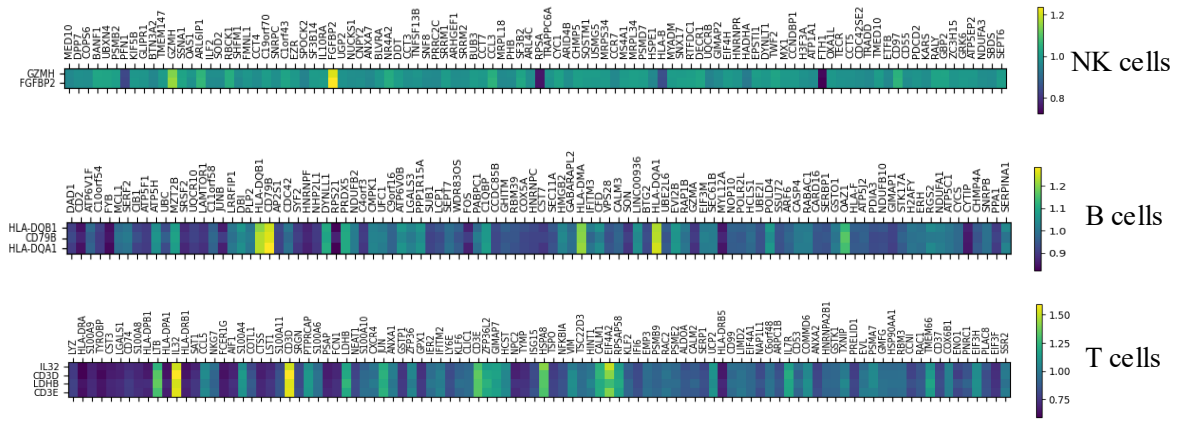

Figure 3: Attention weights of highly differential genes with genes in same group on PBMC dataset

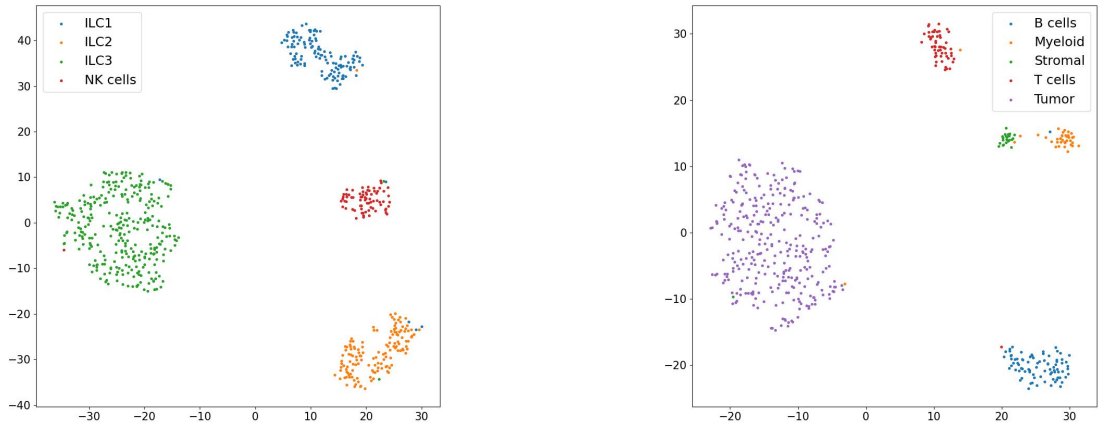

Figure 4: (a)Visualizations of GSE70580 cell embeddings colored by ground-truth cell types. (b)Visualizations of GSE75688 cell embeddings colored by ground-truth cell types.

is found that highly differential genes have larger attention weights among each other while smaller attention weights with most of the other genes in their group. This indicates that these highly differential genes form a 'gene set', binding tightly with each other. For example, HLA-DQB1 and HLA-DQA1 are closely associated with diseases such as Multiple sclerosis, Celiac disease[2, 3]; antibodies to CD79A and CD79B are used in the diagnosis of B-cell neoplasms, distinguishing from T-cell neoplasms or myeloid neoplasms[1]; GZMH and GZMB have similar regulatiion process in gene transcription[4].

6.1.3 Visualization on Cell Embeddings

More visualization results of the cell embeddings in CGRAN are presented here. We can see from Figure 4 that CGRAN is capable of separating most of the cells into their ground-truth cell types by boundaries, implying that CGRAN learns classification-friendly embeddings.

7 More about Transferability of CFAN and CGRAN

In this section, we demonstrate experiments of transferability of CFAN and CGRAN across GSE72056 and PBMC.

Table 2: Accuracy of CFAN and CGRAN on GSE72056 with regard to different cell type number

	3 labels	4 labels	9 labels
CFAN	98.67%	94.07%	92.02%
CGRAN	98.10%	94.07%	93.31%

We first present the baseline classification accuracy of CFAN and CGRAN on GSE72056. We calculate each gene's variance in GSE72056 and select top 1000 most variable genes that also appear in PBMC dataset. These 1000 genes are used for input for CFAN and CGRAN. As mentioned in the article, GSE72056 and PBMC have cell types in common (T cells, B cells, NK cells) as well as distinct cell types. Three classification tasks are used for evaluation on CFAN and CGRAN in GSE72056, as shown in Table 2. The first one performs classification on three common cell types. The second one per-

forms classification on four cell types: common cell types and 'other cell type'. The last one performs classification on nine cell types, including all the cell types in GSE72056 and PBMC. The results not only show the excellent performance of CFAN and CGRAN under different number of cell types, but also can be used for comparison with the following experiments results on transferability.

7.1 Transferability of CFAN

Using three pretrained CFAN models mentioned above, we finetune them on PBMC dataset. The result is shown in Table 3. 30% of PBMC is used for finetune and 70% for test. Only the last fully connected layer of CFAN is finetuned.

Table 3: Transferability of CFAN from GSE72056 to PBMC. Classification accuracy of finetuned CFAN on PBMC dataset is shown.

	3 labels	4 labels	9 labels
Epoch 1	87.38%	87.36%	61.31%
Epoch 50	90.72%	89.68%	89.20%

The result in Table 3 illustrates classification accuracy on PBMC after finetuning for 1 epoch and 50 epochs(finish finetuning). We can discover that CFAN has the ability to capture important features for cell type classification and transfer them to another dataset.

7.2 Transferability of CGRAN

More experiments on the transferability of CGRAN are conducted. Using the three CGRAN's pretrained model on GSE72056 mentioned above, we finetune them for 75 epochs on PBMC, as illustrated in Table 4. We adopt SVD matrix factorization directly on PBMC for initializations of cell embeddings and gene embeddings. Only the last fully connected layer of CGRAN is finetuned.

Table 4: Transferability of CGRAN from GSE72056 to PBMC. Accuracy of finetuned CGRAN model on PBMC is shown with regard to different finetuning proportion.

finetune : test	3 labels	4 labels	9 labels
2 : 8	54.95%	64.97%	72.11%
8 : 2	66.88%	72.20%	77.89%

From table 4, it is discovered that CGRAN can transfer knowledge from one dataset to another. Moreover, With more data provided for finetuning, CGRAN will achieve better classification accuracy.

At last, we compare CGRAN's transferability with regard to different genes' selection. we calculate every gene's variance in PBMC dataset and select the top 1000 most variable genes that appear in GSE72056. These genes are different from the genes used in Table 2. It still remains the same that only parameters in the last fully connected layer are being finetuned. We evaluate the transferability of CGRAN under different genes' choices on a nine cell type classification task, as presented in Table 5. 80% of PBMC is used for finetune, 20% for test.

Using different genes indeed has impact on transfer learning's accuracy on PBMC dataset. However, both gene choices come to high classification accuracy after finetuning for 75 epochs. It again proves the excellent transferability of CGRAN.

Table 5: Transferability of CGRAN from GSE72056 to PBMC with regard to different gene choices on nine cell types task.

	GSE72056 top 1000 common genes	PBMC top 1000 common genes
Epoch 1	38.53%	19.59%
Epoch 75	77.89%	83.40%

References

[1] Peiguo G Chu and Daniel A Arber, 'Cd79: a review', *Applied Immunohistochemistry & Molecular Morphology*, **9**(2), 97–106, (2001).

[2] Matthew R Lincoln, Sreeram V Ramagopalan, Michael J Chao, Blanca M Herrera, Gabriele C DeLuca, Sarah-Michelle Orton, David A Dyment, A Dessa Sadovnick, and George C Ebers, 'Epistasis among hla-drb1, hla-dqa1, and hla-dqb1 loci determines multiple sclerosis susceptibility', *Proceedings of the National Academy of Sciences*, **106**(18), 7542–7547, (2009).

[3] Francesca Megiorni and Antonio Pizzuti, 'Hla-dqa1 and hla-dqb1 in celiac disease predisposition: practical implications of the hla molecular typing', *Journal of biomedical science*, **19**, 1–5, (2012).

[4] Karin A Sedelies, Thomas J Sayers, Kirsten M Edwards, Weisan Chen, Daniel G Pellicci, Dale I Godfrey, and Joseph A Trapani, 'Discordant regulation of granzyme h and granzyme b expression in human lymphocytes', *Journal of biological chemistry*, **279**(25), 26581–26587, (2004).