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

Tianxu Wang[a]; Yue Fan[b]; Xiuli Ma[a;*]

[a] Key Laboratory of Machine Perception(MOE), School of Intelligence Science and Technology, Peking University, Beijing, China

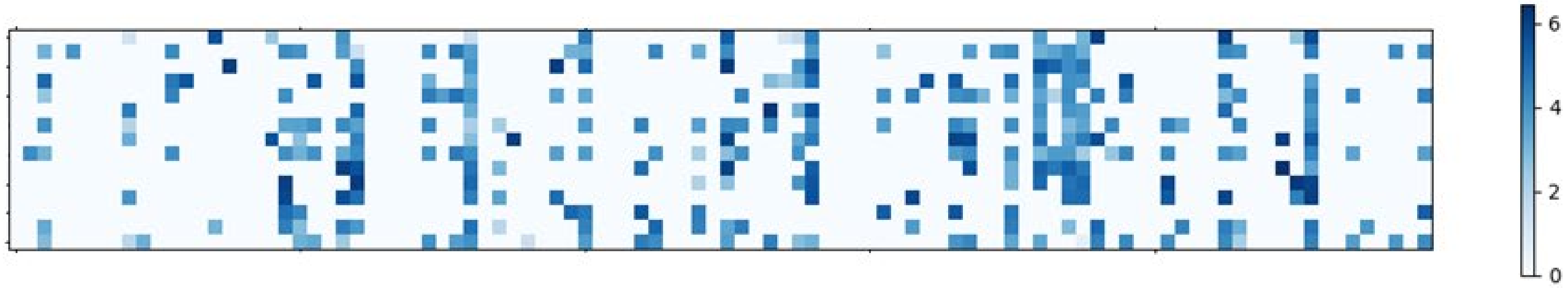
[b] Beijing Institute for General Artificial Intelligence, Beijing, China

Contents

- Introduction and Background
- Related Work
- Methods
- Experiments

Introduction and Background

- Traditional Bulk sequencing technology
  - Produces average gene expression values in mixtures of cells
- Single-Cell RNA-sequencing(scRNA-seq) technology
  - Profile the whole transcriptome of each cell
  - Measure the expression values of all genes in every single cell
- Example of single-cell RNA-seq data



Introduction and Background

- Challenges
    - High dimensional
        - Marker genes
        - House-keeping genes
    - Sparsity and Dropouts
        - Large percentage of zeros
        - True zeros: genes not expressed in cells
        - False zeros: genes expressed but fail to be detected by scRNA-seq technology

Introduction and Background

- Cell Type Classification
  - Fundamental analysis in bioinformatics
  - Recognize various cells in cancer microenvironment
  - Discover new cell types
  - Facilitate other downstream tasks
- Detailed Problem Statement
  - Input:
    - Matrix $M^{c \times g}$: expression values of g genes in c cells
    - Cells with cell type label in training set
  - Mission: predict cell type for cells without labels

Related Work

- Methods make use of prior knowledge
    - Reference Datasets
        - SingleR
        - Scmap
    - Marker genes
        - SCINA
- Challenges
    - Require high quality markers for every cell type
    - Rare cell types

Related Work

- Methods using neural network
  - More independent from prior knowledge
  - ACTINN
  - Cell BLAST
  - ScCapsNet
  - EpiAnno
- Challenges
  - Hard to effectively interpret the biological meanings hidden in the parameters

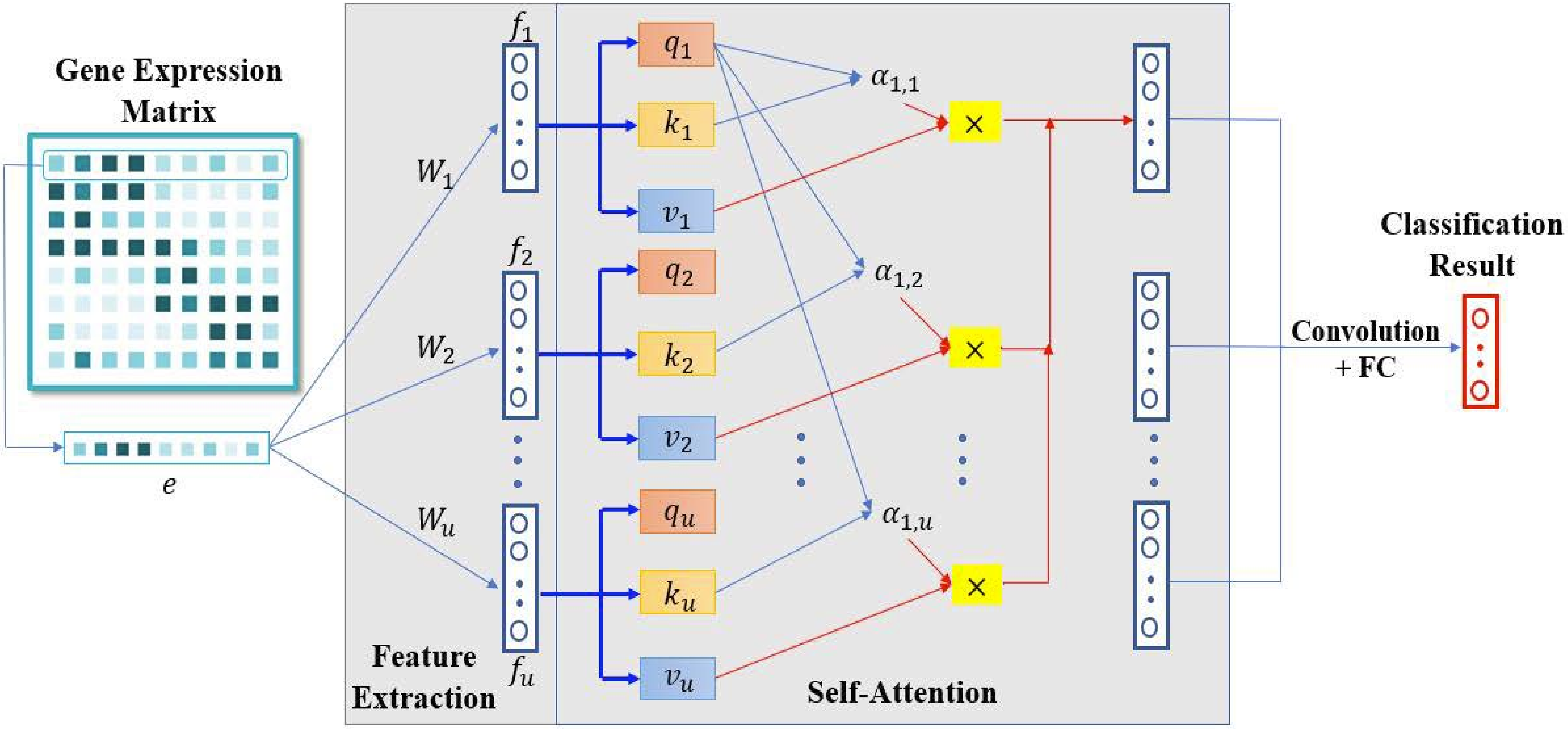
Methods

- Motivation
  - Both accurate and interpretable, providing insights into the underlying biological mechanism
  - Self-Attention Mechanism
    - Great success in diverse types of data
    - Attention weights serves as strong indicator of the affinities among tokens
  - Applying self-attention to cell type classification
    - Obstacles: tokens / features
    - Basic model: Cell Feature Attention Network (CFAN)
    - Further interpretability: Cell-Gene Representation Attention Network(CGRAN)

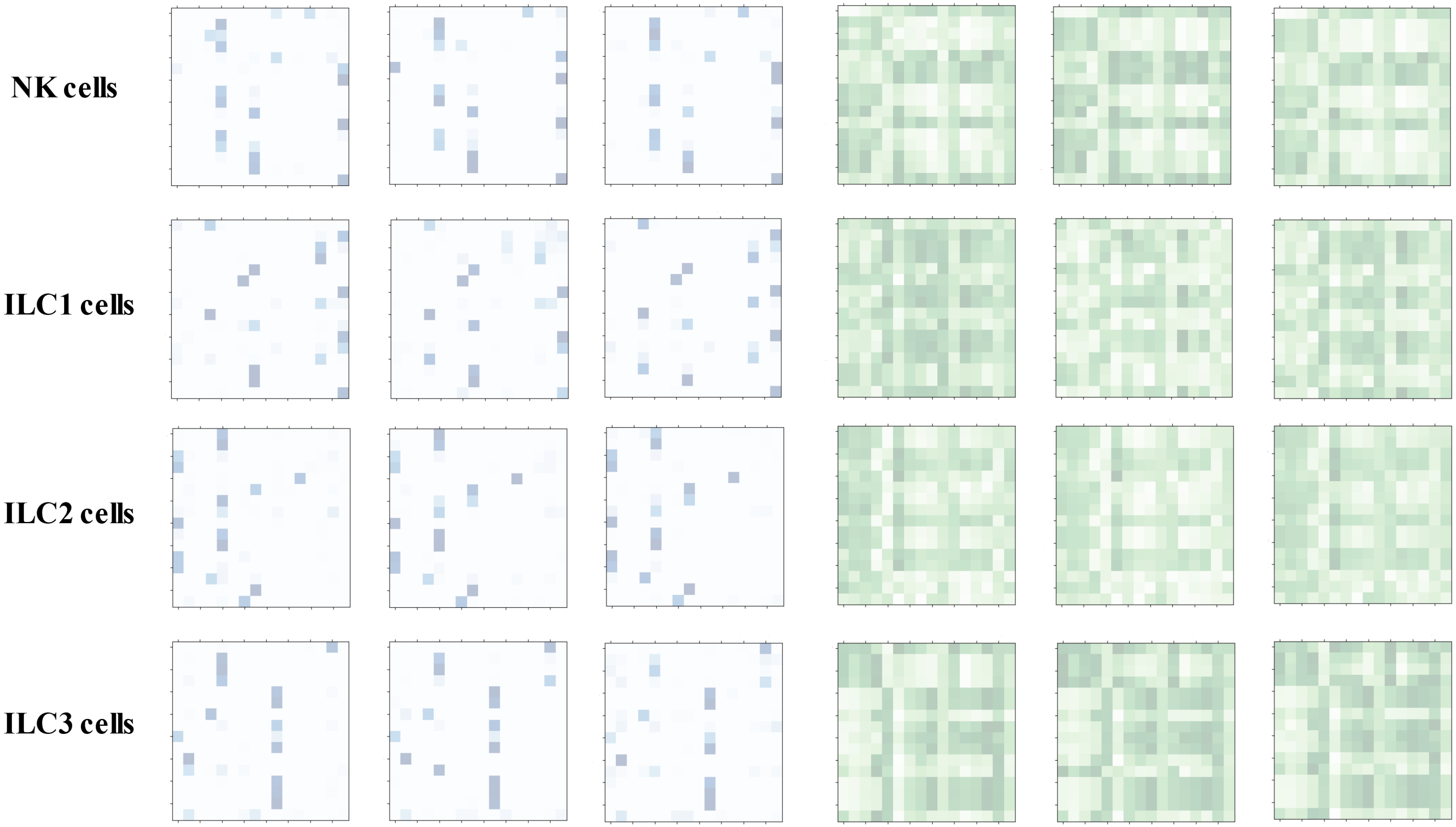
CFAN

- Cell Feature Extraction
  - Extract hidden features of a cell using u feature extractors (implemented as dense layer)
  - $f_i = Norm(ReLU(W_i e \ + b_i))$
- Renew Features
  - Single-head self-attention renew
  - $F = Norm(Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V)$
- Cell type classification
  - 1D convolution
  - Fully connected layer

CFAN Architecture



Insights from CFAN



- Visualizations of attention weights and output features produced by CFAN

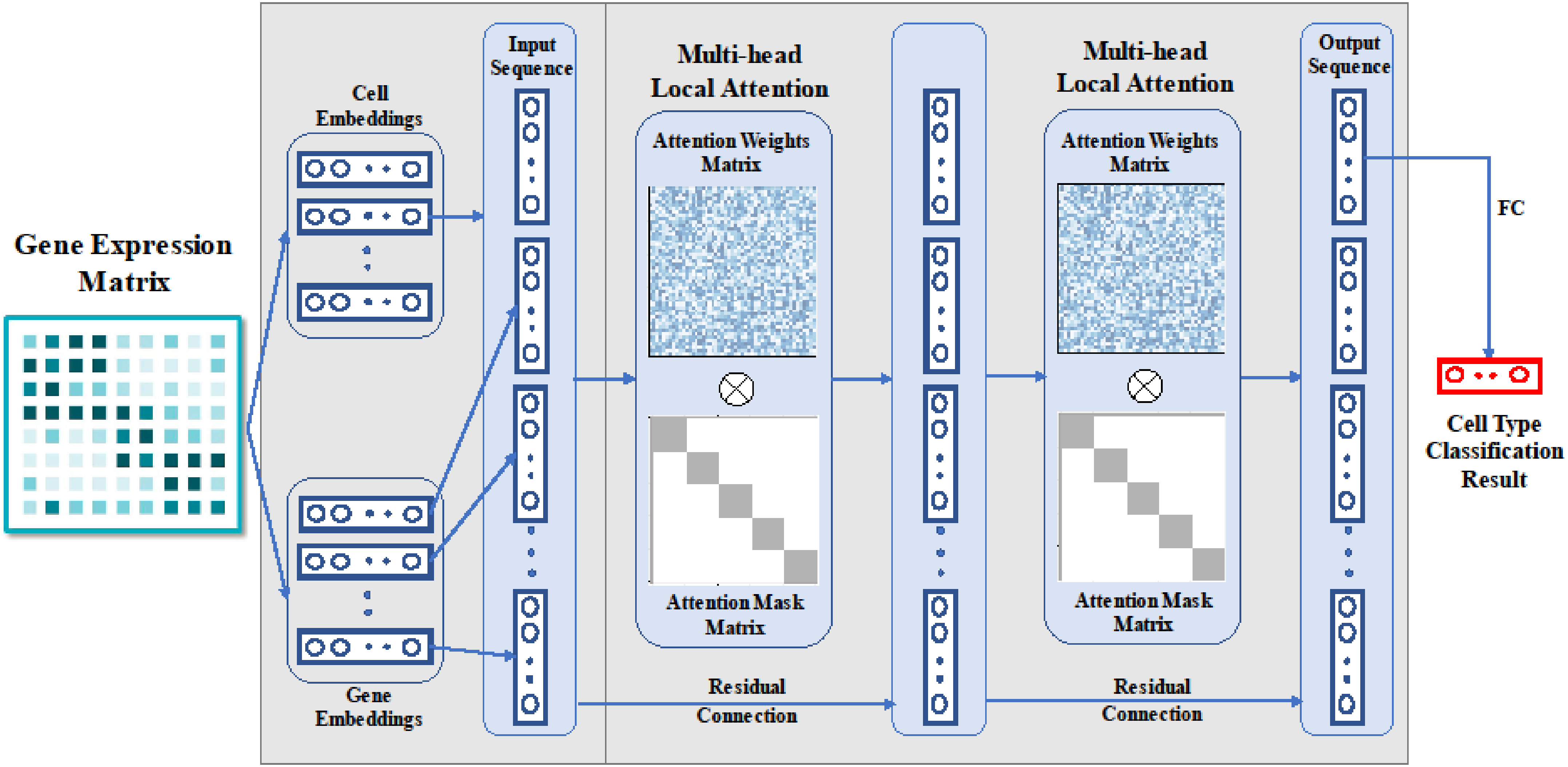
- Similar patterns in cells from same cell type

- Attention weights serves as indicator, however, not interpretable enough in CFAN

CGRAN

- Concretize every token as a biological entity, i.e. a cell or a gene
  - ‘Representation learning’ for cells and genes
    - Embedding vector for each cell and gene
    - Cell embeddings —— Classification
    - Attention scores among cells and genes —— contribution of the gene to a cell for cell type classification

CGRAN Architecture



CGRAN PART 1: Initialize Cell embeddings and Gene embedding

- Obtain initial Cell embedding vectors $A^{c\times m}$ and gene embedding vectors $B^{g\times m}$
- SVD decomposition：
  - $M = X \sum Y^T$
  - Cell embedding vectors： $\mathrm{A} = (X\sum_s)$:,:m
  - Gene embedding vectors： $\mathrm{B} = \left(\mathrm{Y}\sum_s{}^T\right)$:,:m
- Embedding vectors learned by gradient descent
  - $\underset{A,B}{\mathrm{argmin}}\, MSE(M, AB^T)$

CGRAN PART 2: Multi-head Local Attention

- Input:
  - For each cell i, input sequence $S = \{s_0, s_1, \dots s_g\}$
  - $s_0 = A_{i,:}, s_j = B_{j,:} + s_0, \ j \in \{1,2, \dots g\}$
- Two Sequential attention blocks
- Problems:
  - Fully attention on long sequence leads to low classification accuracy
  - Large g : most of the attention weight closed to zeros
- Introduce Local Attention
  - Genes are divided into several groups: only attention weights among genes from the same group are preserved
    - Uniform Grouping
    - Gene Cluster Grouping

Experiments

- Datasets Description

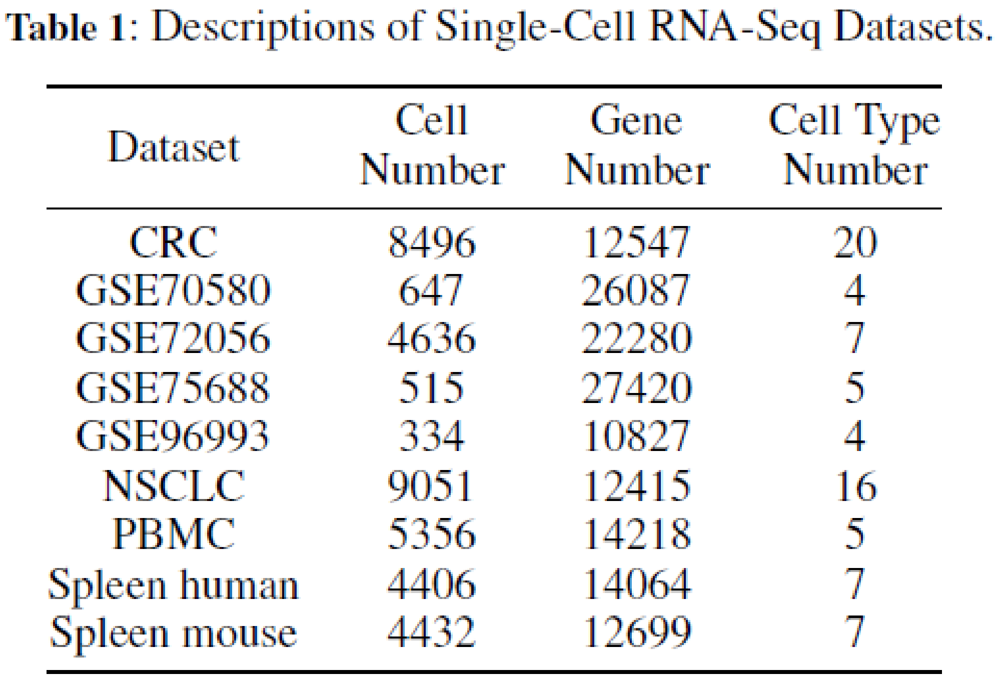
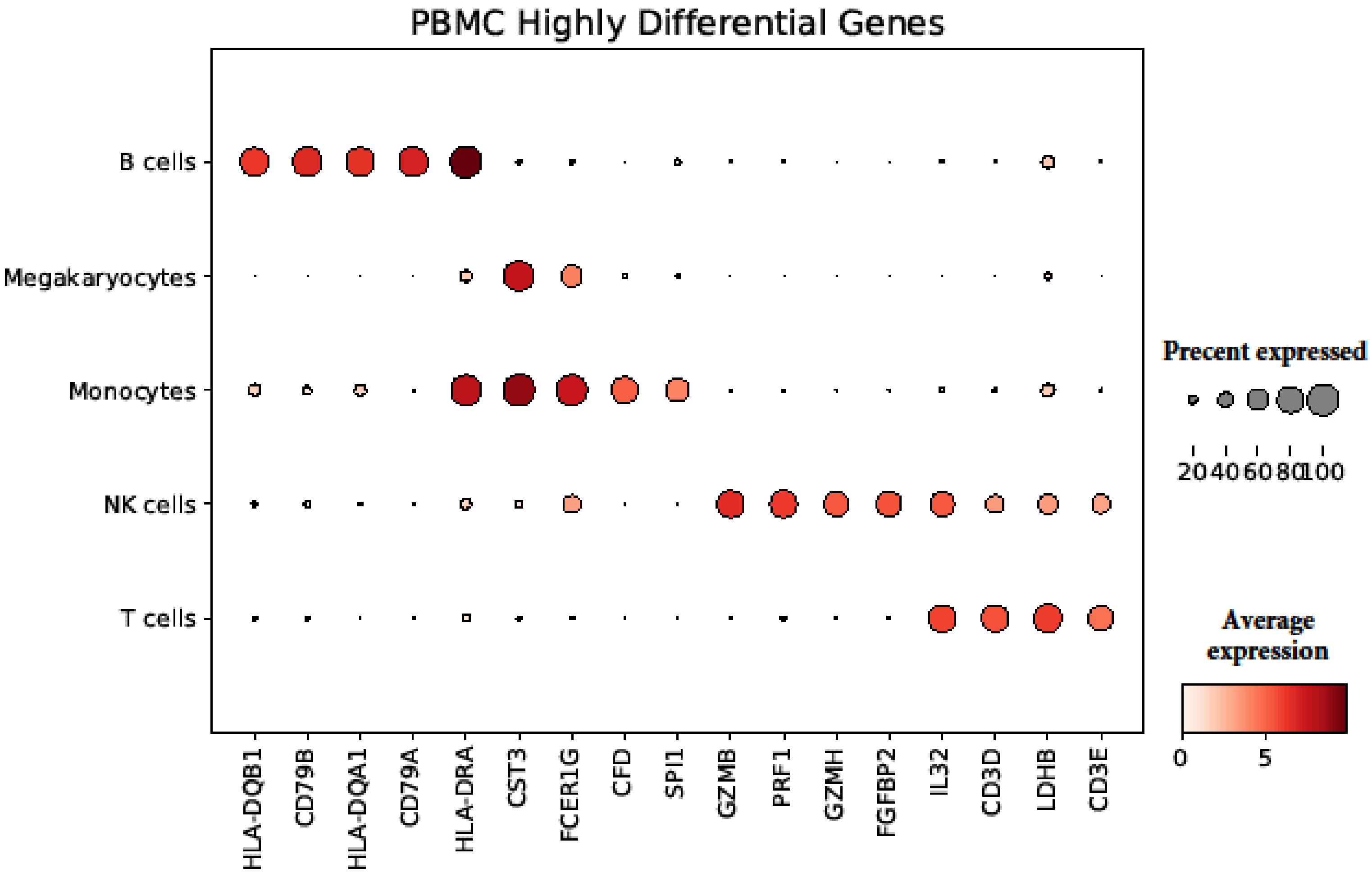


Table 1: Descriptions of Single-Cell RNA-Seq Datasets.

Dataset	Cell Number	Gene Number	Cell Type Number
CRC	8496	12547	20
GSE70580	647	26087	4
GSE72056	4636	22280	7
GSE75688	515	27420	5
GSE96993	334	10827	4
NSCLC	9051	12415	16
PBMC	5356	14218	5
Spleen human	4406	14064	7
Spleen mouse	4432	12699	7

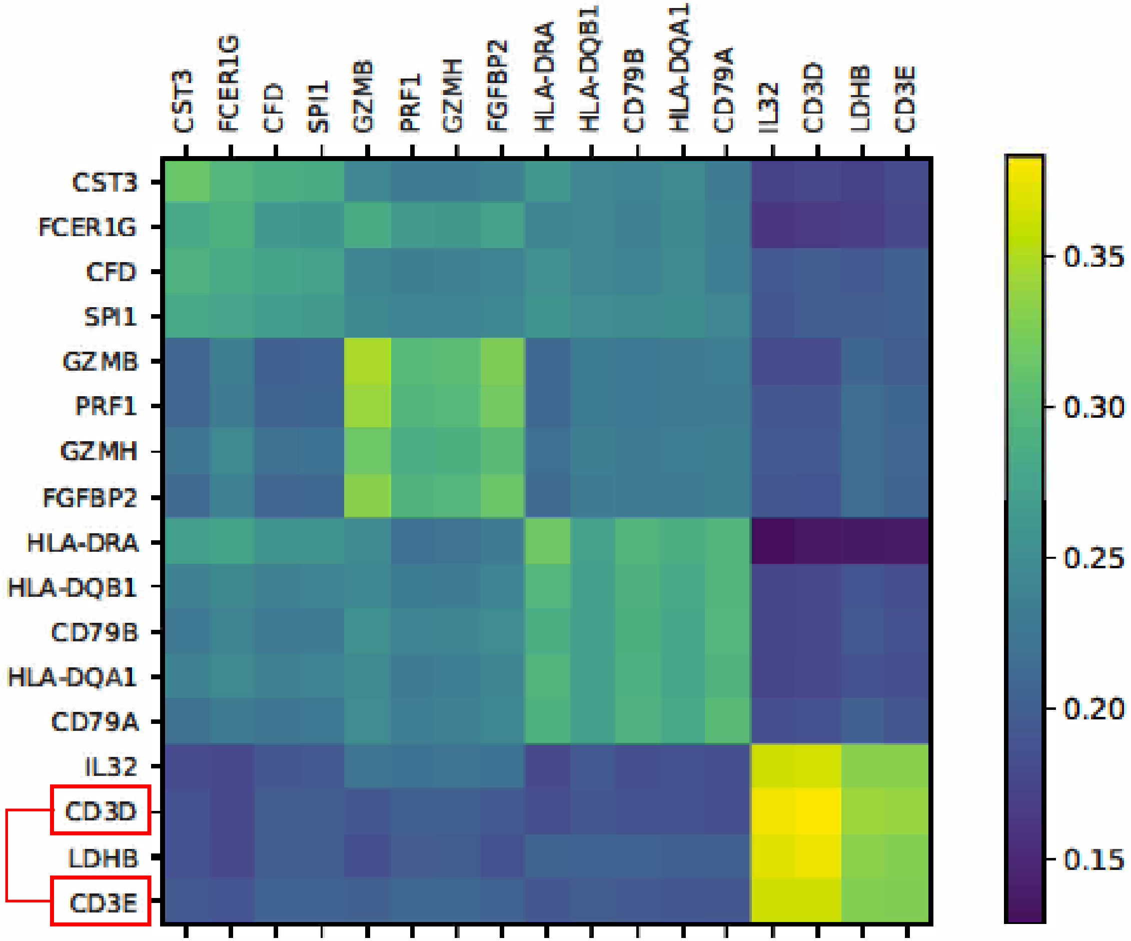
Interpretations of CGRAN(I)

- Identification of marker genes
  - For each cell and its input $S = \{s_0, s_1, \dots s_g\}$, consider attention weights of the gene tokens $\{s_1, \dots, s_g\}$ to the cell token s_0 in the first attention block
  - 'Picked' by a cell : if the gene is among the top-50 genes with highest attention weights
  - 'highly differential gene ': If a gene is 'picked' by most of the cells from a certain cell type exclusively



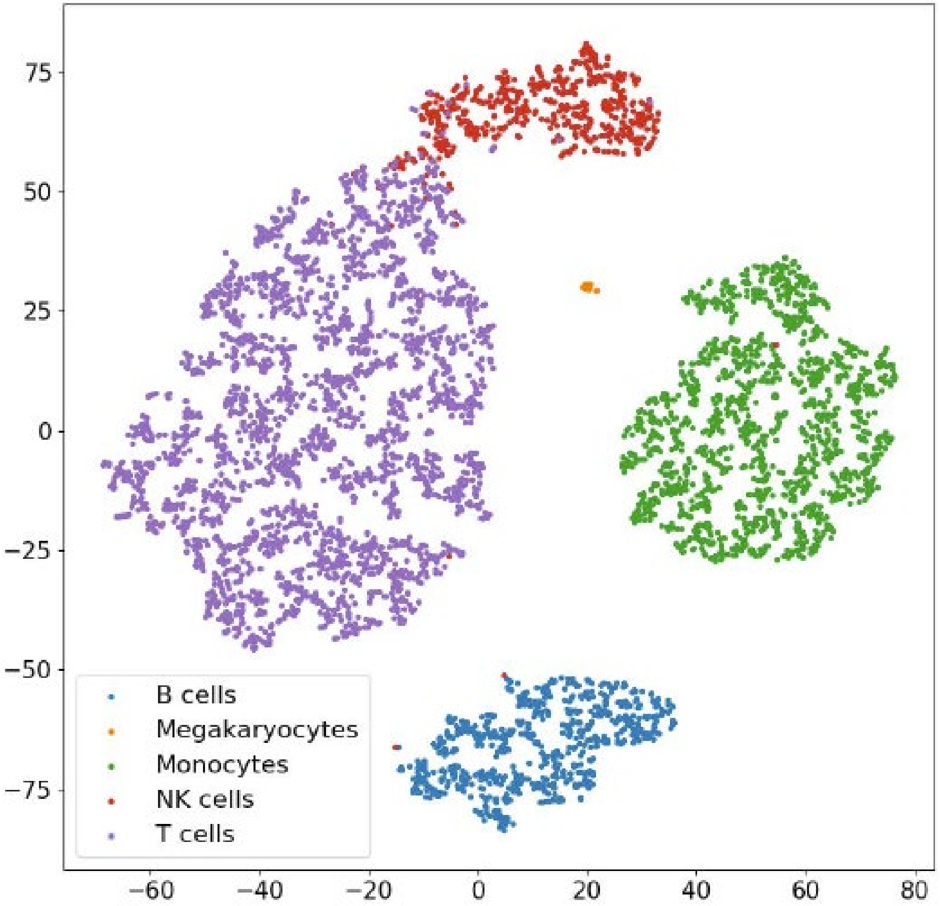
Interpretations of CGRAN(II)

- Analysis on ‘Gene Sets’
  - Gene with high attention weights
  - Illustrations of the attention weights among all highly differential genes identified by CGRAN in PBMC dataset



Interpretations of CGRAN(III)

- Classification-Friendly Embeddings
    - t-SNE visualizations of cell embeddings output by CGRAN



Classification Performance

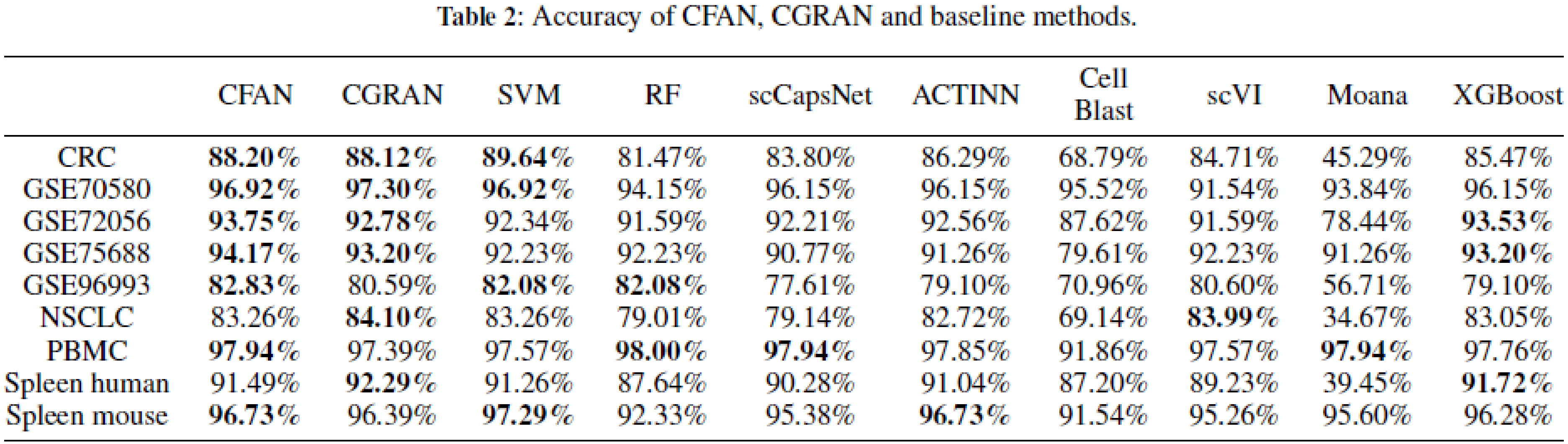


Table 2: Accuracy of CFAN, CGRAN and baseline methods.

	CFAN	CGRAN	SVM	RF	scCapsNet	ACTINN	Cell Blast	scVI	Moana	XGBoost
CRC	**88.20%**	**88.12%**	**89.64%**	81.47%	83.80%	86.29%	68.79%	84.71%	45.29%	85.47%
GSE70580	**96.92%**	**97.30%**	**96.92%**	94.15%	96.15%	96.15%	95.52%	91.54%	93.84%	96.15%
GSE72056	**93.75%**	**92.78%**	92.34%	91.59%	92.21%	92.56%	87.62%	91.59%	78.44%	**93.53%**
GSE75688	**94.17%**	**93.20%**	92.23%	92.23%	90.77%	91.26%	79.61%	92.23%	91.26%	**93.20%**
GSE96993	**82.83%**	80.59%	**82.08%**	**82.08%**	77.61%	79.10%	70.96%	80.60%	56.71%	79.10%
NSCLC	83.26%	**84.10%**	83.26%	79.01%	79.14%	82.72%	69.14%	**83.99%**	34.67%	83.05%
PBMC	**97.94%**	97.39%	97.57%	**98.00%**	**97.94%**	97.85%	91.86%	97.57%	**97.94%**	97.76%
Spleen human	91.49%	**92.29%**	91.26%	87.64%	90.28%	91.04%	87.20%	89.23%	39.45%	**91.72%**
Spleen mouse	**96.73%**	96.39%	**97.29%**	92.33%	95.38%	**96.73%**	91.54%	95.26%	95.60%	96.28%

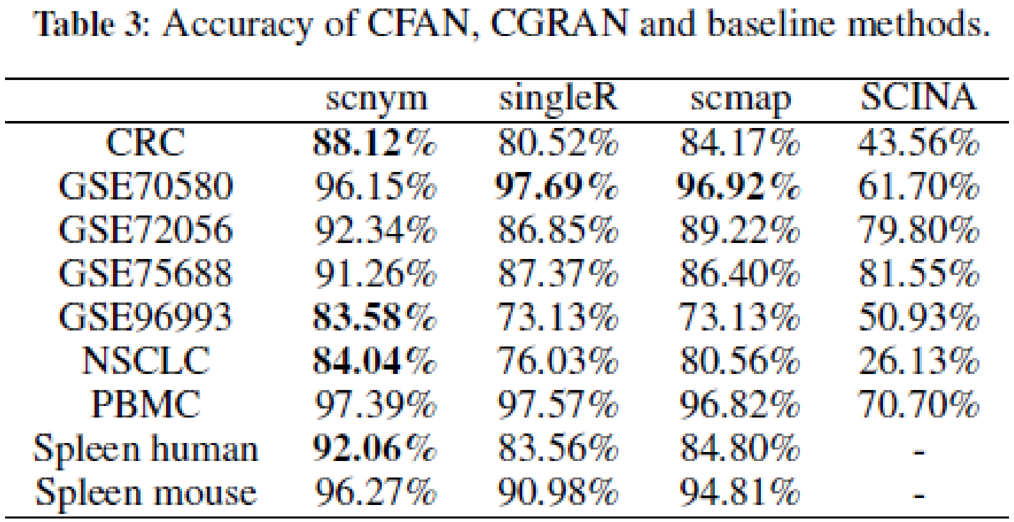


Table 3: Accuracy of CFAN, CGRAN and baseline methods.

	scnym	singleR	scmap	SCINA
CRC	**88.12%**	80.52%	84.17%	43.56%
GSE70580	96.15%	**97.69%**	**96.92%**	61.70%
GSE72056	92.34%	86.85%	89.22%	79.80%
GSE75688	91.26%	87.37%	86.40%	81.55%
GSE96993	**83.58%**	73.13%	73.13%	50.93%
NSCLC	**84.04%**	76.03%	80.56%	26.13%
PBMC	97.39%	97.57%	96.82%	70.70%
Spleen human	**92.06%**	83.56%	84.80%	-
Spleen mouse	96.27%	90.98%	94.81%	-

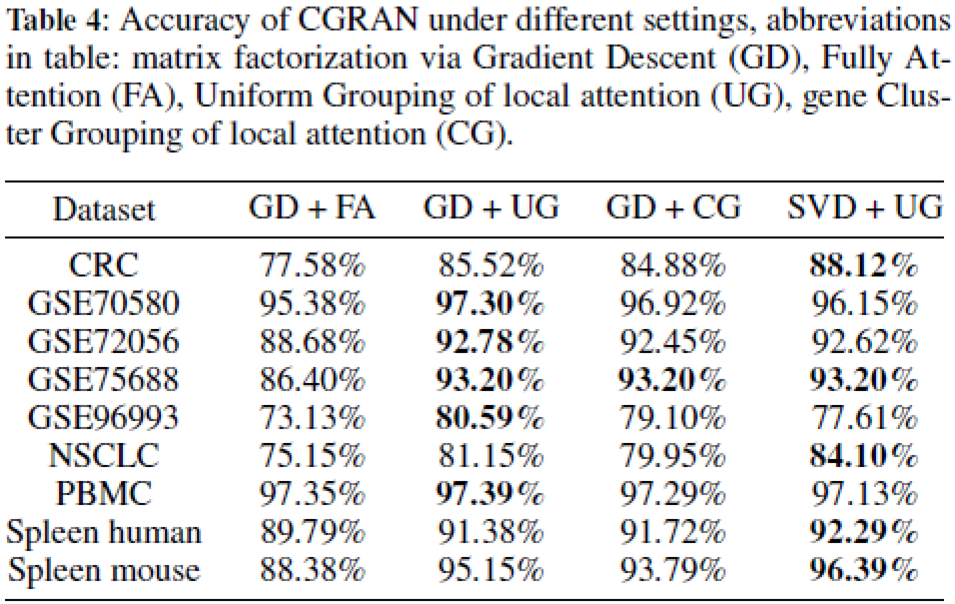
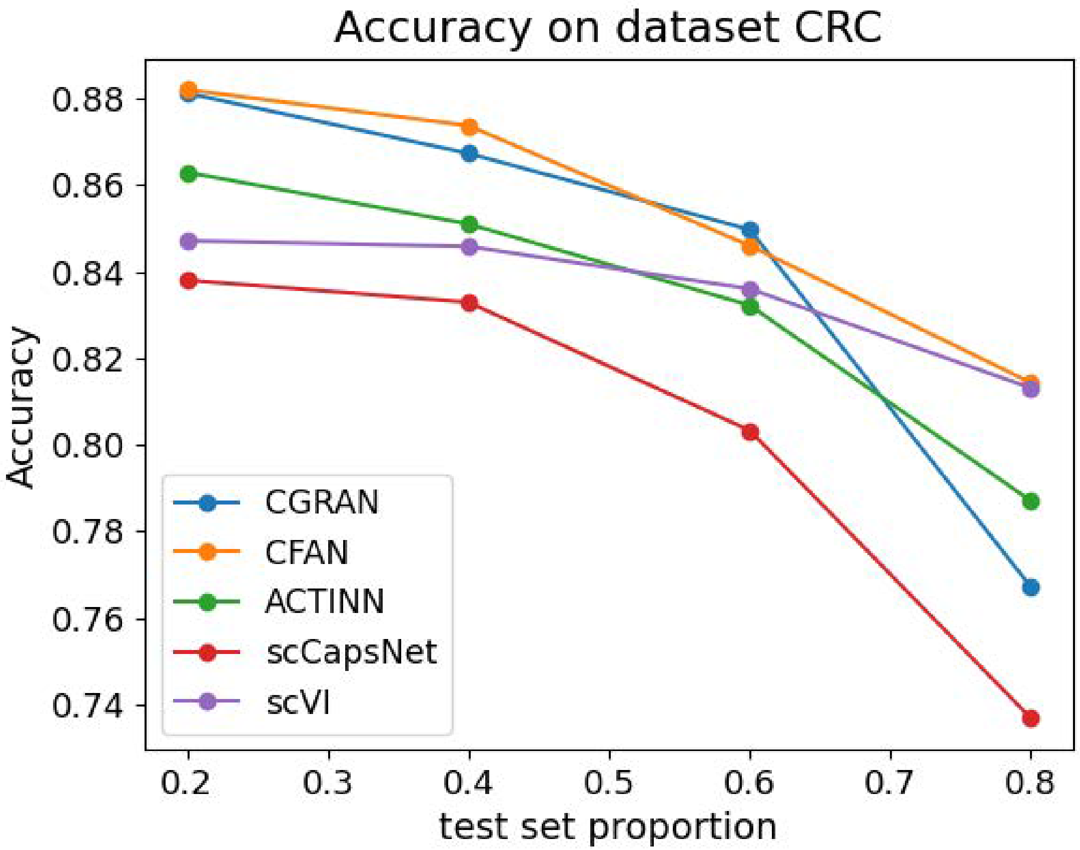
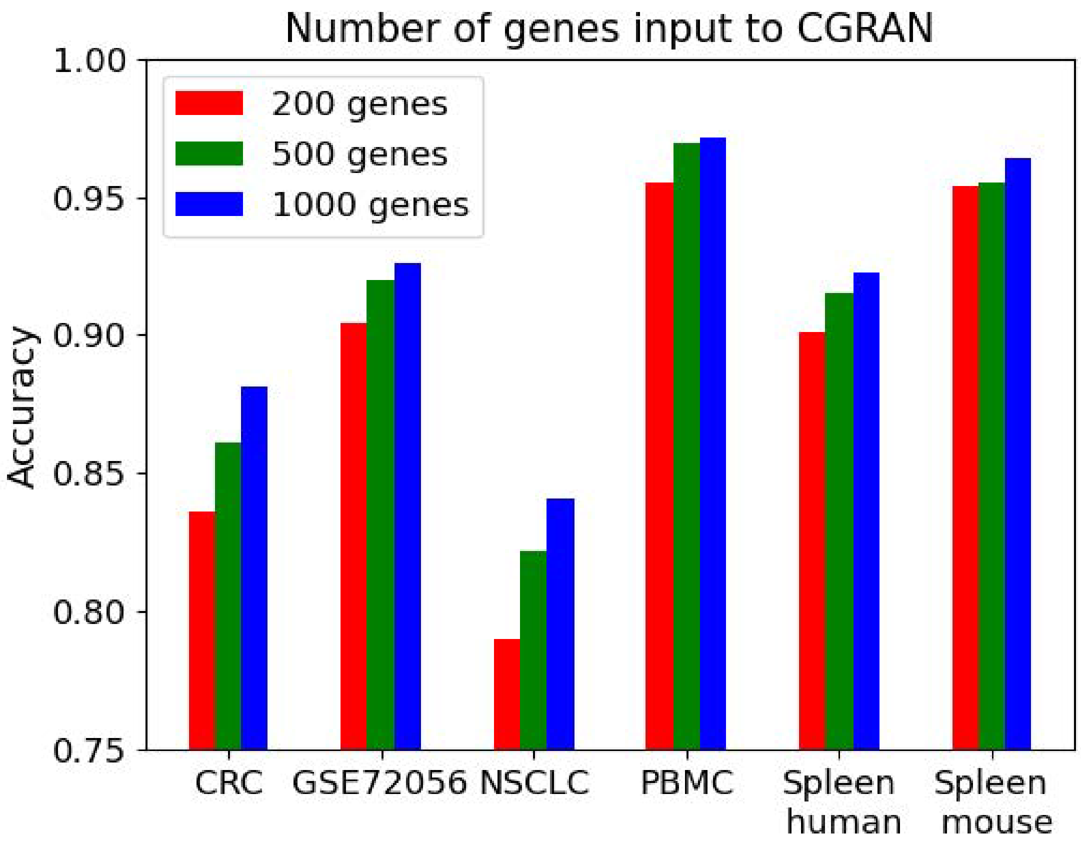


Table 4: Accuracy of CGRAN under different settings, abbreviations in table: matrix factorization via Gradient Descent (GD), Fully Attention (FA), Uniform Grouping of local attention (UG), gene Cluster Grouping of local attention (CG).

Dataset	GD + FA	GD + UG	GD + CG	SVD + UG
CRC	77.58%	85.52%	84.88%	**88.12%**
GSE70580	95.38%	**97.30%**	96.92%	96.15%
GSE72056	88.68%	**92.78%**	92.45%	92.62%
GSE75688	86.40%	**93.20%**	**93.20%**	**93.20%**
GSE96993	73.13%	**80.59%**	79.10%	77.61%
NSCLC	75.15%	81.15%	79.95%	**84.10%**
PBMC	97.35%	**97.39%**	97.29%	97.13%
Spleen human	89.79%	91.38%	91.72%	**92.29%**
Spleen mouse	88.38%	95.15%	93.79%	**96.39%**

Robustness Test





Experiments on Transferability of Models across Datasets

- From GSE72056 to PBMC
- Select top 1000 most variable genes in PBMC which appear in GSE72056
- Pretrained on GSE72056 on nine-cell-type classification

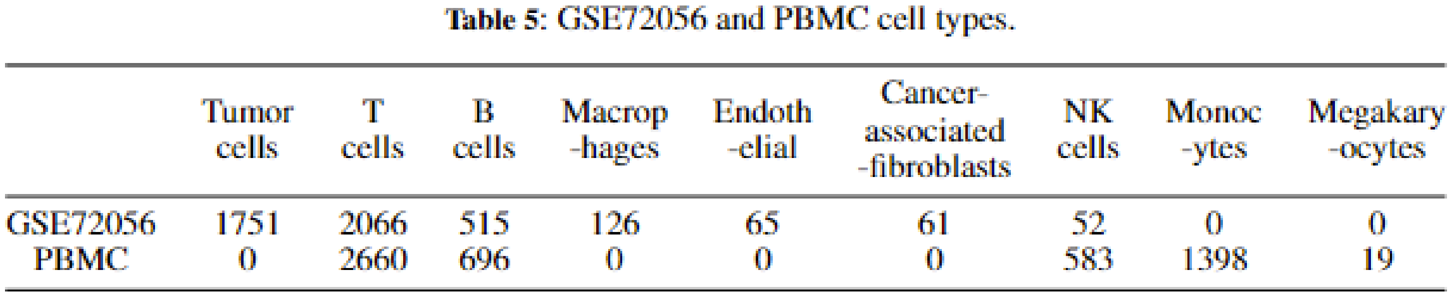


Table 5: GSE72056 and PBMC cell types.

	Tumor cells	T cells	B cells	Macrop -hages	Endoth -elial	Cancer-associated -fibroblasts	NK cells	Monoc -ytes	Megakary -ocytes
GSE72056	1751	2066	515	126	65	61	52	0	0
PBMC	0	2660	696	0	0	0	583	1398	19

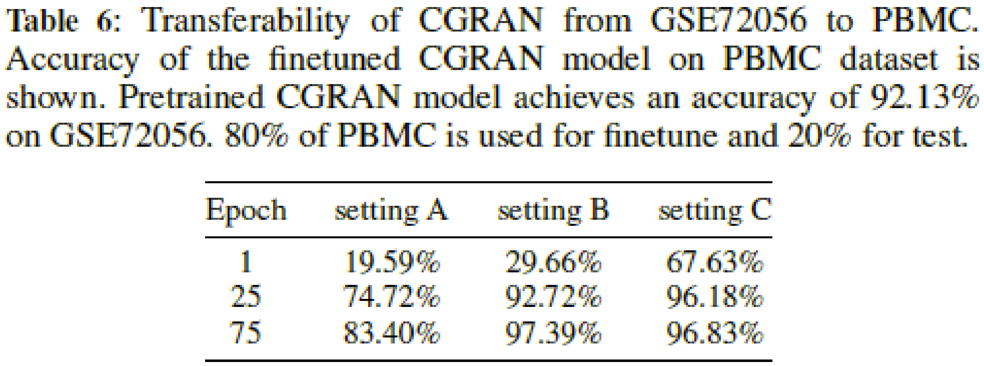
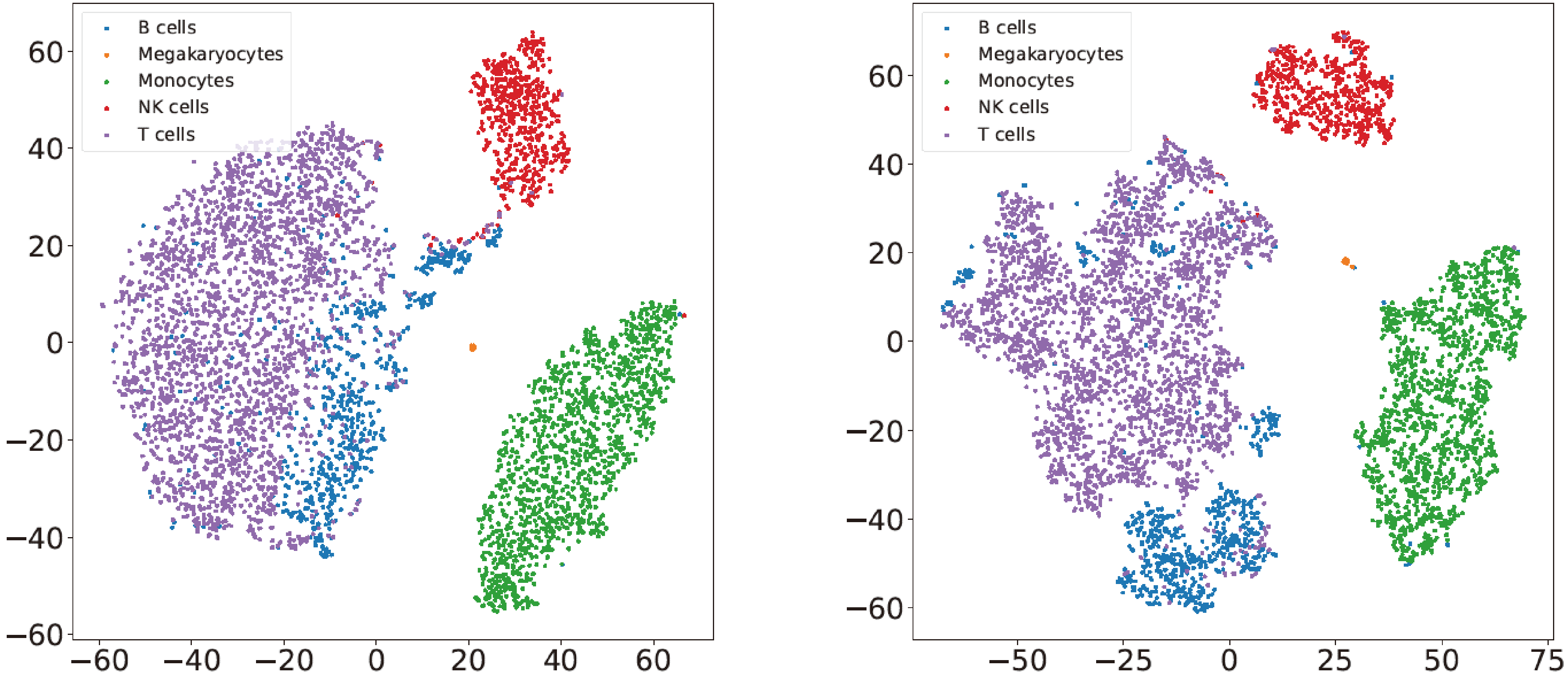


Table 6: Transferability of CGRAN from GSE72056 to PBMC. Accuracy of the finetuned CGRAN model on PBMC dataset is shown. Pretrained CGRAN model achieves an accuracy of 92.13% on GSE72056. 80% of PBMC is used for finetune and 20% for test.

Epoch	setting A	setting B	setting C
1	19.59%	29.66%	67.63%
25	74.72%	92.72%	96.18%
75	83.40%	97.39%	96.83%

Novel Cell Type Discovery

- Visualizations of cell embeddings produced by CFAN(left) and CGRAN(right)
- Blue cell types(B cells) masked during training



Conclusion and Future Work

- CFAN
    - Feature attention
- CGRAN
    - Cell-Gene attention
    - Interpretation
        - Marker genes discovery
        - Gene-Gene relationship
        - Visualization
    - Transfer
    - Novel Cell Type Discovery
- Future Work
    - Gene Embedding Initialization & Gene Groups

References

[1] Dvir Aran, Agnieszka P Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P Naikawadi, Paul J Wolters, Adam R Abate, et al., ‘Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage’, *Nature immunology*, **20**(2), 163–172, (2019).

[2] Zhi-Jie Cao, Lin Wei, Shen Lu, De-Chang Yang, and Ge Gao, ‘Searching large-scale scrna-seq databases via unbiased cell embedding with cell blast’, *Nature communications*, **11**(1), 1–13, (2020).

[3] Oscar Franzen, Li-Ming Gan, and Johan LM Bjorkegren, ‘Panglaodb: a web server for exploration of mouse and human single-cell rna sequencing data’, *Database*, **2019**, (2019).

[4] Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg, ‘scmap: projection of single-cell rna-seq data across data sets’, *Nature methods*, **15**(5), 359–362, (2018)

[5] Feiyang Ma and Matteo Pellegrini, ‘Actinn: automated identification of cell types in single cell rna sequencing’, *Bioinformatics*, **36**(2), 533– 538, (2020).

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, ‘Attention is all you need’, *Advances in neural information processing systems*, **30**, (2017).

[7] Lifei Wang, Rui Nie, Zeyang Yu, Ruyue Xin, Caihong Zheng, Zhang Zhang, Jiang Zhang, and Jun Cai, ‘An interpretable deep-learning architecture of capsule networks for identifying cell-type gene expression programs from single-cell rna-sequencing data’, *Nature Machine Intelligence*, **2**(11), 693–703, (2020).

[8] Xinxin Zhang, Yujia Lan, Jinyuan Xu, Fei Quan, Erjie Zhao, Chunyu Deng, Tao Luo, Liwen Xu, Gaoming Liao, Min Yan, et al., ‘Cellmarker: a manually curated resource of cell markers in human and mouse’, *Nucleic acids research*, **47**(D1), D721–D728, (2019).

Thanks For your Attention