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

Tianxu Wang[a], **Yue Fan**[b] **and 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

Abstract. Cell type classification serves as one of the most fundamental analyses in bioinformatics. It helps recognizing various cells in cancer microenvironment, discovering new cell types and facilitating other downstream tasks. Single-cell RNA-sequencing (scRNA-seq) technology can profile the whole transcriptome of each cell, thus enabling cell type classification. However, high-dimensional scRNA-seq data pose serious challenges on cell type classification. Existing methods either classify the cells with reliance on the prior knowledge or by using neural networks whose massive parameters are hard to interpret. In this paper, we propose two novel attention-based models for cell type classification on single-cell RNA-seq data. The first model, Cell Feature Attention Network (CFAN), captures the features of a cell and performs attention model on them. To further improve interpretation, the second model, Cell-Gene Representation Attention Network (CGRAN), directly concretizes tokens as cells and genes and uses the cell representation renewed by self-attention over the cell and the genes to predict cell type. Both models show excellent performance in cell type classification; additionally, the key genes with high attention weights in CGRAN indicate and identify the marker genes of the cell types, thus proving the model's biological interpretation.

1 Introduction

Cell type classification can identify the cell type of a cell based on its gene expression. Single-cell RNA-sequencing (scRNA-seq) [25] can profile the whole transcriptome within individual cells, which brings unprecedented opportunities for cell type classification [19, 23, 4]. As being crucial in recognizing tumor cells in cancer microenvironment, discriminating diverse cell states in cell differentiation and facilitating many downstream tasks such as cell-cell communication [18], cell type classification plays a key role in understanding diseases and biological processes.

Despite the surging single-cell analytical methods, cell type classification remains challenging. Some methods heavily rely on the prior knowledge of reference datasets, such as scmap [17], or marker genes [35], which are the genes that have distinguishable expression levels in different cell types. However, defining marker genes for various cell types requires extensive biological experiments which are effort-consuming [1]. Some cell types' marker genes even remain unknown. Another challenge comes from the high dimensionality and high sparsity of the scRNA-seq data. The limitation of the sequencing technology usually causes a large percentage of zeros or dropouts, up to 95% for instance [24, 15], in the data, which easily misleads any analysis and makes it hard to discriminate different cells.

Neural networks are versatile in many tasks. For instance, neural networks with supervised learning can automatically extract useful patterns for classification, thus eliminating the need for prior knowledge of the 'marker genes'; representation learning can reduce the dimensionality and data sparsity. However, existing neural-network-based methods for cell type classification are difficult to interpret, making it hard to understand the knowledge gained from the models. Actually, biologists are also concerned about the underlying mechanisms for distinguishing two cell types besides simply correctly assigning all cells as ground-truth. A model that is both accurate and interpretable, providing insights into the underlying biological mechanisms, is in urgent demand.

The self-attention mechanism has achieved great success in handling diverse types of data including sentences [28, 11], graphs [29] and images [12] due to its outstanding performance and vivid interpretability. Usually, the self-attention layer takes a sequence of tokens as input, and renews every token's 'value' by an attention-weighted aggregation over all token values, where the attention weights are decided by the 'query' of the target token and the 'keys' of all tokens. The attention weights have been shown to be a strong indicator of the affinities among the tokens.

However, applying the self-attention mechanism to cell type classification still faces obstacles since it requires tokens of cell features. In this paper, we first propose Cell Feature Attention Network (CFAN) as a basic mechanism applying self-attention mechanism to classify cells. In CFAN, several feature extractors are first built to extract different cell features from the raw transcriptomic profile of the cell. By viewing the features as a set of tokens, we deploy a self-attention layer to update these features. Finally, a convolution operation is performed on the updated features to predict the cell type of the given cell. Experiments show that CFAN is among the best models in classification accuracy, while it is still hard to produce biological insights from the attention weights as the tokens lack concrete meanings.

To enhance interpretability further, the second model Cell-Gene Representation Attention Network (CGRAN) is developed by concretizing every token as a biological entity, specifically, a cell or a gene. In detail, CGRAN first generates embedding vectors for every gene and cell by factorizing the scRNA-seq matrix. Given a cell to be classified, a token sequence, with the first token being the cell and the following ones being the genes, is fed to self-attention layers. The cell vector renewed by the attention layers is used for the final clas-

* Corresponding Author. Email: xlma@pku.edu.cn

sification. Besides the good classification performance of CGRAN, it is also discovered that for a specific cell type, the genes having large attention weights to the cells match well with the marker genes. Furthermore, by analyzing the attention weights among the genes, one can gather genes with high attention weights into different 'gene sets', indicating their close associations. These insights into the attention weights demonstrate the strong interpretability of CGRAN.

The main contributions of our article are as follows. (1) The self-attention mechanism applied in CFAN learns cell-type-specific patterns from the features of the cell, which achieves a performance gain over existing methods. (2) By concretizing the input sequence tokens as biological entities, CGRAN learns the attention weights among the genes and the cell, which enables direct biological interpretation from the model's attention weights. (3) Local attention technique is proposed in CGRAN to make attention weights distinguishable, which helps to increase the classification accuracy. (4) Experiments show that both CFAN and CGRAN have ability to discover novel cell types and strong transferability across different datasets with common genes and cell types.

2 Related Work

Typical solutions for cell type classification make use of prior knowledge such as reference datasets or marker genes. SingleR [3] annotates cells by correlating single-cell transcriptomes with a reference dataset using Spearman coefficient and map cells to the cell types in reference dataset. Scmap [17] builds cluster in reference dataset and calculates three similarity measures (Pearson, Spearman and cosine) between cells in the target dataset and cluster centroids in reference datasets. Then it assigns cells in target dataset to the cell type with the highest similarity value. SCINA [35] makes use of marker genes and implement an expectation-maximization model for cell type classification. However, models based on reference datasets may fail to classify cells when some rare target cell types do not appear in reference datasets. Marker gene based methods require high quality markers for every cell type in the dataset, which may be diverse in different marker genes database. Also, when it comes to rare cell types, it is hard to provide sufficient marker genes.

As neural networks have shown strong ability in diverse areas, many methods deal with the noisy and high-dimensional single-cell RNA-seq data by using neural networks. ACTINN [22] employs a neural network with three hidden layers to predict cell type. EpiAnno [9] uses a Bayesian neural network to embed the cells into a latent space, where the cells follow a Gaussian mixture distribution. Cell BLAST [7] projects the cells from the high-dimensional transcriptomic space to a low-dimensional cell embedding space, and then search for similar cells. OnClass [32] embeds cell types into a low-dimensional space, maps each cell into the region of its cell type and classifies cells into different cell types in Cell Ontology. ScCapsNet [31] designs a deep learning architecture of capsule networks using dynamic routing and analyzes the internal weights among capsules, while it is hard to effectively interpret the biological meanings hidden in the parameters. By contrast, the parameters in our model CGRAN can be easily interpreted as rich biological information.

3 Methods

Given a sparse scRNA-seq matrix $\mathbf{M}^{c\times g}$ depicting the expression values of g genes in c cells, with some of the cells already annotated with its ground-truth cell type from the set $\mathcal{T} = \{t_1, t_2, ..., t_n\}$ containing n different cell types, cell type classification is to assign the remaining cells to their correct cell types. In this section, we elaborate on two attention based models, namely the Cell Feature Attention Network (CFAN) and Cell-Gene Representation Attention Network (CGRAN) for cell type classification on scRNA-seq data.

3.1 Cell Feature Attention Network

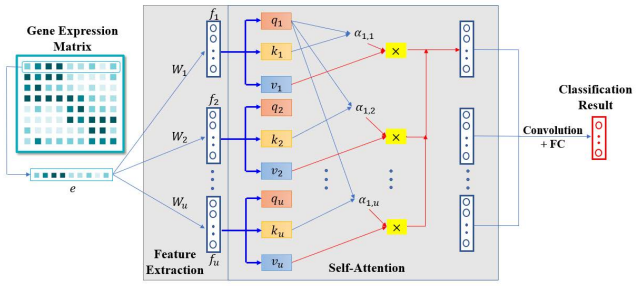

Figure 1: Cell Feature Attention Network.

The framework of Cell Feature Attention Network (CFAN) is shown in Figure 1. Given the vector $\boldsymbol{e} \in R^g$ of the raw gene expression values of a cell, CFAN extracts hidden features from $\boldsymbol{e}$ by using u feature extractors. Each feature extractor is a dense layer followed by ReLU activation and L2 layer normalization. Formally, the i-th feature $\boldsymbol{f}_i$ is formulated as:

$$\boldsymbol{f}_i = \text{LayerNorm}(\text{ReLU}(\mathbf{W}_i\boldsymbol{e} + \boldsymbol{b}_i)) \tag{1}$$

where $\mathbf{W}_i$ and $\boldsymbol{b}_i$ are the weight and the bias of the dense layer from the i-th feature extractor.

After obtaining u cell features $\{\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_u\}$ of the cell, CFAN utilizes a single-head self-attention layer to renew these features. It transforms every feature vector $\boldsymbol{f}_i$ into a query vector $\boldsymbol{q}_i$, a key vector $\boldsymbol{k}_i$ and a value vector $\boldsymbol{v}_i$. Let $\mathbf{Q}^{u\times d}$, $\mathbf{K}^{u\times d}$ and $\mathbf{V}^{u\times m}$ be the matrices stacked from $\{\boldsymbol{q}_i\}_{i=1}^u$, $\{\boldsymbol{k}_i\}_{i=1}^u$ and $\{\boldsymbol{v}_i\}_{i=1}^u$, with d being the dimension of the query and the key vectors, m being the dimension of the value vectors. Then the output matrix $\mathbf{F}^{u\times m}$ can be written as:

$$\mathbf{F} = \text{LayerNorm}(\text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V}) \tag{2}$$

At last, we input $\mathbf{F}$ to a 1D convolution layer followed by a fully connected layer, which finally generates the possibilities of a cell belonging to different cell types.

Here we examine the hidden parameters of CFAN by visualizing these parameters trained on the dataset GSE70580. The heatmaps of the attention weights and output features are illustrated in Figure 3. It can be seen that cells from the same cell type share a similar pattern in both attention weights and output feature matrix, and the patterns of different cell types are distinguishable from each other. It suggests that the hidden parameters, especially the attention weights, could serve as a reasonable indicator for biological knowledge, as discussed in our second model CGRAN.

3.2 Cell-Gene Representation Attention Network

Here we present Cell-Gene Representation Attention Network (CGRAN) that concretizes the attention tokens as the cells and the genes. The architecture of CGRAN is presented in Figure 2. CGRAN first learns the representation vectors for cells and genes through matrix factorization on the scRNA-seq matrix $\mathbf{M}$. Then, for each cell, a token sequence with the first one being the cell itself and the remaining ones being the genes are fed to the following multi-head local attention layers. The final output of the cell token is used for the classification. The details of CGRAN are as follows.

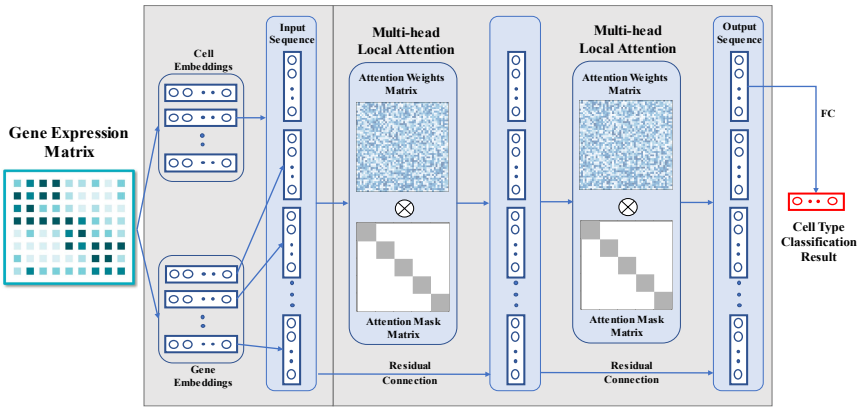

Figure 2: Cell-Gene Representation Attention Network.

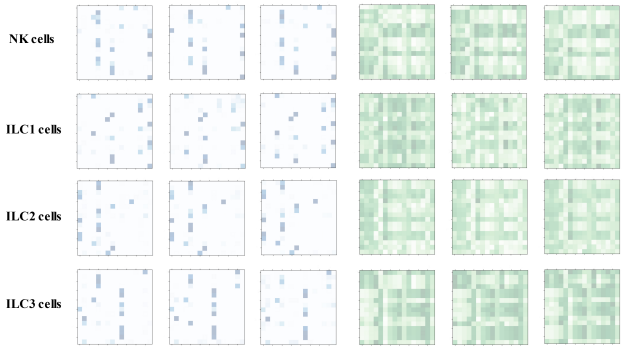

Figure 3: The attention weights and the output features produced by CFAN. On GSE70580, three random cells are sampled for each cell type, with their attention weights shown in the left three columns and the output features shown in the right three columns.

3.2.1 *The Embeddings of Cells and Genes*

The model first obtains the cell embedding vectors $\mathbf{A}^{c\times m}$ and gene embedding vectors $\mathbf{B}^{g\times m}$ by factorizing the scRNA-seq matrix $\mathbf{M}$, where m is the embedding dimension. Two different ways of matrix factorization have been evaluated.

The first way is Singular Value Decomposition (SVD), which factorizes the input scRNA-seq matrix as:

$$\mathbf{M} = \mathbf{X}\mathbf{\Sigma}\mathbf{Y}^T \tag{3}$$

where $\mathbf{X}^{c\times c}$ and $\mathbf{Y}^{g\times g}$ are real orthogonal matrices, and $\mathbf{\Sigma}^{c\times g}$ is a rectangular diagonal matrix with r non-negative real numbers in descending order on the diagonal, where r is the rank of $\mathbf{M}$. Here we denote $\mathbf{\Sigma}_s$ as the rectangular diagonal matrix of size $c \times g$ with the diagonal values being the square roots of the values in $\mathbf{\Sigma}$. Then, cell SVD vectors can be obtained as $\mathbf{X}\mathbf{\Sigma}_s$ and gene SVD vectors as $\mathbf{Y}{\mathbf{\Sigma}_s}^T$. We choose the first m dimensions from the SVD vectors since they correspond to top m biggest singular values. Therefore, the final embedding vectors for cells and genes are $\mathbf{A} = (\mathbf{X}\mathbf{\Sigma}_s)_{:,:m}, \mathbf{B} = (\mathbf{Y}{\mathbf{\Sigma}_s}^T)_{:,:m}$.

The second way of matrix factorization is by gradient descent. It first randomly initializes the cell embedding vectors $\mathbf{A}^{c\times m}$ and the gene embedding vectors $\mathbf{B}^{g\times m}$, and optimizes them by gradient descent using Mean Square Error (MSE) loss:

$$\arg\min_{\mathbf{A},\mathbf{B}} \text{MSE}(\mathbf{M}, \mathbf{A}\mathbf{B}^T) \tag{4}$$

Matrix Factorization using Gradient Descent (GD) is only adopted as comparison in Table 4, and the embeddings obtained by GD are normalized. As GD-based method may be influenced by the initializations of neural networks, we recommend SVD-based method for matrix factorization.

3.2.2 *Multi-Head Local Attention*

After obtaining the embedding vectors of cells and genes, an input sequence can be generated based on the learned embedding vectors. Specifically, given cell i in $\mathbf{M}$, the input embedding sequence $\mathcal{S} = \{\boldsymbol{s}_0, \boldsymbol{s}_1, ..., \boldsymbol{s}_g\}$ can be presented as:

$$\boldsymbol{s}_0 = \mathbf{A}_{i,:}, \quad \boldsymbol{s}_j = \mathbf{B}_{j,:} + \boldsymbol{s}_0,\ j \in \{1, 2, ..., g\}, \tag{5}$$

where $\boldsymbol{s}_0$ represents the cell token and $\{s_j\}_{j=1}^{g}$ represents the g gene tokens with regard to this cell.

The sequence is then fed to two sequential attention blocks. Each block consists of a multi-head attention layer, a residual connection using fully connected network and L2-normalization on the outputs. The attention blocks update these embedding vectors and more importantly, learn the hidden relationships among the cells and genes by renewing the attention weights among them.

However, the problem of over-fitting arises when using fully-attention on a long sequence, leading to a low classification accuracy. To be specific, let $\mathbf{Q}^{(g+1)\times d}$ and $\mathbf{K}^{(g+1)\times d}$ be the queries and the keys of $\{\boldsymbol{s}_0, \boldsymbol{s}_1, ..., \boldsymbol{s}_g\}$. Then the fully-attention weight matrix can be written as $\text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})$. Due to large g, the softmax operation will be performed on a large number of tokens, causing most of the attention weights very close to zero. Therefore, the fully attention mechanism may learn distorted relationships among the entities.

To solve this problem, local attention mechanism is introduced. The gene tokens are first divided into several groups, and only the attention weights among the genes from the same group are preserved, whereas those from different groups are neglected. Two different ways are proposed for the grouping. The first way is uniform grouping, where every hundred genes in $\mathbf{M}$ are treated as a group, which is simple for implementation. Besides, every group has the same number of genes, making the blocks of attention weights comparable with each other. The second way groups the genes by K-Means clustering algorithm and is named as gene cluster grouping. This grouping strategy is based on the similarities of genes' expression. However, it may lead to an imbalanced grouping since the size of the gene groups can be of great difference.

To implement the local attention mechanism, a mask matrix $\mathbf{H}^{(g+1)\times(g+1)}$ is applied to the attention weights $\text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})$.

Let Group(i,j), $i,j \in \{1,2,...,g\}$ be a Boolean function which equals 1 when the i-th gene token and the j-th gene token are from the same group, and 0 otherwise. Then, $\mathbf{H}$ can be written as:

$$\mathbf{H}_{i,j} = \begin{cases} 1, & \text{if } i=0 \text{ or } j=0 \\ \text{Group}(i,j), & \text{else} \end{cases} \tag{6}$$

The output $\mathbf{F}$ of a single-head local attention is formed as:

$$\mathbf{F} = \text{LayerNorm}(\mathbf{H} \odot \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}))\mathbf{V}, \tag{7}$$

where $\odot$ denotes Hadamard product, LayerNorm$(\cdot)$ is L2-normalization and $\mathbf{V}$ are the values of the tokens. The local attention preserves the weights between: 1) the cell and all the genes; 2) the genes from the same group. In CGRAN, multiple heads are used in each local attention block. The final output embedding of the cell token will be passed into a fully connected layer to generate the possibilities of a cell belonging to different cell types.

4 Experiments

In this section, we illustrate the interpretability of CGRAN. Then, the classification performance of CFAN, CGRAN and other baseline methods will be evaluated. Finally, transferability across diverse datasets and the ability to discover novel cell types of CFAN and CGRAN will be evaluated.

4.1 Datasets and Settings

Datasets. Nine scRNA-seq datasets are used for experiments, namely CRC [33], GSE70580 [5], GSE72056 [26], GSE75688 [10], GSE96993 [2], NSCLC [14], PBMC [1], Spleen human [6] and Spleen mouse [6]. Table 1 presents the details of the datasets. Every dataset is first preprocessed, and the top 1000 variable genes are considered as the input $\mathbf{M}$. We use five-fold cross validation on all of our datasets. 80% of the cells are used for training and validation, and the rest 20% for testing. Evaluations on classification are conducted for three times.

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

Settings. For CFAN, the number of feature extractors is set to $u = 16$. Each feature vector f_i has 128 dimensions. The dimension of the attention layer output embedding is $m = 16$. A dropout rate of 0.2 is used on all layers to prevent over-fitting. For CGRAN, the dimension of both cell embedding and the gene embedding is set to 128. Ten heads are used in both local attention blocks. The dimensions of output embeddings in the first and second block for each head are respectively 8 and 4. The dropout rate of all layers is set to 0.1. Cross entropy loss is used as loss function and adam optimizer is adopted with learning rate of 0.0001.

[1] https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc6k

4.2 The Interpretation of CGRAN

As mentioned, CGRAN is intended for providing biological insights from the learned attention weights among the cells and the genes. In this part, the crucial interpretability of CGRAN will be addressed from three perspectives as follows.

4.2.1 Identification of Marker Genes

It is found that for the cells from a specific cell type, there exist certain genes that have distinctively large attention weights with these cells, and match with the known marker genes of this cell type.

Specifically, for each cell and its input sequence $\mathcal{S} = \{\boldsymbol{s}_0, \boldsymbol{s}_1, ..., \boldsymbol{s}_g\}$, consider the attention weights of the gene tokens $\{\boldsymbol{s}_1, ..., \boldsymbol{s}_g\}$ to the cell tokens $\boldsymbol{s}_0$ in the first attention block of CGRAN (summing up attention weights of all attention heads). A gene is 'picked' by a cell if it is among the top-50 genes with the highest attention weights to the cell. If a gene is picked by most of the cells from a certain cell type exclusively, then the gene is called as a 'highly differential gene' of this cell type.

Figure 4(a) displays the highly differential genes found in PBMC dataset. Given a gene and a cell type, the size of the circle reflects the proportion of the cells in which the gene expresses in this cell type, while the color of the circle corresponds to the mean expression values of the gene in this cell type. As observed, for a gene exclusively picked by a cell type according to the high attention weights, its average expression level in this cell type tends to be high, indicating high probability of being marker gene of the cell type. Actually, most of the highly differential genes displayed in Figure 4(a) match with the ground-truth marker genes provided by CellMarker [34] and Panglao database [13], proving that CGRAN has a strong capability of identifying the marker genes by using the learned attention weights between the cells and genes.

To further illustrate the effectiveness of using attention scores among cell embeddings and gene embeddings for identification of marker genes, we also compare the interpretability of CGRAN with that of CFAN by using input perturbation. Given CFAN trained on PBMC, for each cell, the negative partial derivatives of the loss function on the expression input are calculated. Then for every cell type t_i, the partial derivatives of cells belonging to it are averaged to a vector f_i containing g elements, with each revealing the contribution of a certain gene to the accurate classification of t_i. Figure 4(b) visualizes the contribution of the genes to the classification of the five cell types on PBMC. Some of the genes, such as CD3D,CD3E for T cells, GZMB for NK cells revealed by lighter colors in Figure 4(b) are indeed marker genes for their cell types. However, Figure 4(a) indicates that CGRAN can provide more clear interpretations about potential markers, demonstrating the necessity of using attention scores to find markers.

4.2.2 Analysis on Gene Sets

The attention weights among the genes may provide rich information about the relationships of the genes. Specifically, genes with high attention weights to each other may have similar functions and close

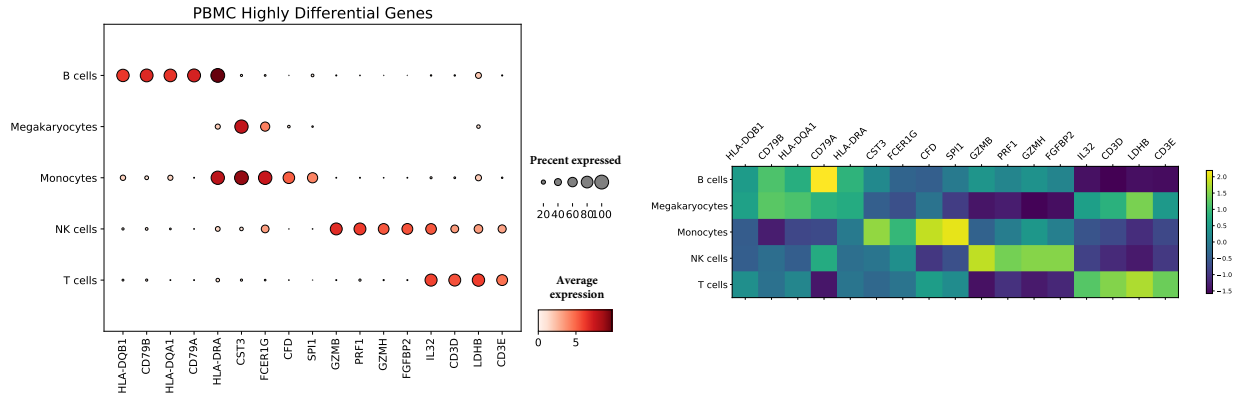

Figure 4: (a) The highly differential genes in PBMC chosen by the attention weights of CGRAN. (b) The contributions of the genes to the cell type classification on PBMC revealed by input perturbation on CFAN.

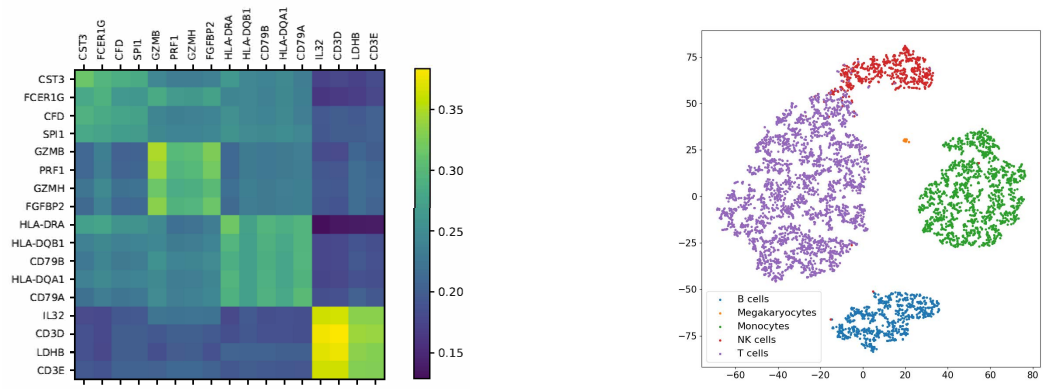

Figure 5: (a) The attention weights among the highly differential genes in PBMC. (b) The output embedding vectors of all cells visualized by t-SNE on PBMC.

associations. Figure 5(a) illustrates the attention weights among all highly differential genes identified by CGRAN in PBMC. The highly differential genes of the same cell type tend to have large attention weights to each other. Moreover, from literature search, we find that genes with high attention to each other tend to have similar functions and form a 'gene set'. For example, CD3D and CD3E are two highly differential genes of T cells. The proteins encoded by CD3D and CD3E are parts of the T-cell receptor/CD3 complex (TCR/CD3 complex) and are involved in T-cell development and signal transduction [20].

4.2.3 *Classification-Friendly Embedding*

In CGRAN, the output embedding vectors of the cells catch the characteristics of their cell types, which not only enable accurate classification, but may also serve as high quality representations of cells for other downstream tasks.

Figure 5(b) shows a t-SNE [27] visualization of the cell embeddings output by CGRAN. Every dot corresponds to a cell in the dataset and is colored with its ground-truth cell type. Cells from the same cell type are closely clustered, which accounts for the accurate classification performance mentioned in the following part. Therefore, CGRAN can effectively learn the classification-friendly embeddings for the cells.

4.3 *Classification Performance*

We first compare CFAN and CGRAN with baseline methods, including Support Vector Machine (SVM), Random Forest (RF), scCapsNet [31], ACTINN [22], Cell Blast [7], scVI [21], Moana [30], XGBoost [8], scnym [16], singleR [3], scmap [17] and SCINA [35], ranging from methods using prior knowledge to methods based on machine learning and deep learning algorithm.

Table 2 and 3 collectively present the evaluations of all methods, with top three ones on each dataset shown in bold. Each method is tuned to its best. CFAN and CGRAN are in the top three on most of the datasets, showing their robustness and excellent performance. For SVM, we adopt scikit-learn implementation. In fact, SVM implementation in scikit-learn adopts "one-versus-one" approach for multi-class classification. As a result, classification accuracy of SVM might be high due to the integration on all sub-models. For SCINA on Spleen human and Spleen mouse datasets, lacking sufficient marker genes for every cell type in these datasets leads to the unevaluated results.

Performances of CGRAN under different settings are listed in Table 4. From the table, we can see that SVD with Uniform Grouping (SVD + UG) and matrix factorization via Gradient Descent with Uniform Grouping (GD + UG) have better performance than the other two settings. CGRAN models with local attention have better performances than fully attention, which proves the effectiveness of local attention.

To further test robustness of different models, we split datasets according to different ratios of training set and test set and evaluate

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%

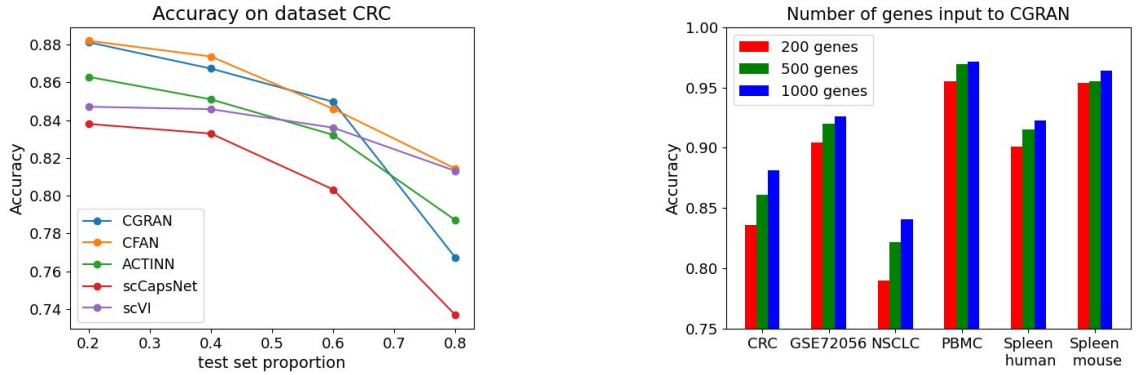

Figure 6: (a) Classification accuracy with different test set proportion on CRC dataset. (b) Classification accuracy with different number of genes.

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%**

their performance. As shown in Figure 6(a), both CFAN and CGRAN perform consistently well when test set proportion is smaller than 0.8 on CRC. Accuracy of CGRAN drops when the proportion of training set becomes smaller because of its deep architecture. More data are needed for training the embeddings of cells and genes as well as the attention weights. Also, it is not a common option to set training set proportion to 0.2 for cell type classification as cells from rare cell type may never appear in training set.

Moreover, the number of genes input to CGRAN can also affect its performance. Since the number of tokens is equal to the number of highly variable genes plus one, the more genes are selected, the more information of cells and genes are fed into the models. Due to computational cost and limitations of devices, we here provide experimental results with gene number varying from 200 to 1000. As shown in Figure 6(b), the more genes are involved, the higher classification accuracy becomes. In this paper, we use the top 1000 most variable genes for cell type classification. If computationally feasible, we expect CGRAN will achieve higher classification accuracy with more genes.

4.4 Experiments on Transferability of Models across Datasets

In this section, we demonstrate transferability of our models across datasets. GSE72056 and PBMC have genes and cell types in common, which makes transfer learning feasible across two datasets. Table 5 lists the common cell types (T cells, B cells and NK cells) between the two datasets, as well as the distinct cell types of each dataset. Here we mainly focus on transferability of CGRAN.

We select the top 1000 most variable genes in PBMC which also appear in GSE72056. Then CGRAN is pretrained on GSE72056 for a nine-cell-type classification task, which includes all the cell types

Table 5: GSE70256 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%

in the two datasets. With regard to whether finetuning on the last fully connected layer or the whole model, and whether initializing the embeddings from GSE72056 or PBMC, we define three different settings. **Setting A** takes embeddings using SVD on PBMC as input into the attention blocks and finetunes on the last fully connected layer; **setting B** takes the same embeddings as A while finetunes on the whole CGRAN; **setting C** takes the gene embeddings on GSE72056, whereas the cell embeddings are initialized by gradient descent with the goal to minimize MSE loss. We finetune the whole CGRAN model in setting C.

Table 6 shows the transferability of CGRAN from GSE72056 to PBMC. Finetuning on the whole model leads to higher accuracy than only finetuning the last FC layer. The finetuning process converges faster as test accuracy exceeds 90% using only 25 epochs when finetuning the whole model.

Furthermore, in setting C, the highly differential genes identified by the pretrained model on GSE72056 match with that of PBMC identified by the finetuned model. They both point out that CD3D, CD3E, CD2 and CD3G are highly differential genes for T cells; CD79A and CD79B are for B cells. These genes are well recognized marker genes for T cells and B cells. This not only indicates that CGRAN has strong transferability across datasets, but also proves that CGRAN provides reasonable interpretations and helps to discover marker genes.

4.5 Novel Cell Type Discovery

We provide two strategies for CFAN and CGRAN to discover novel cell types. The first strategy is visualizing the hidden cell embedding vectors. Figure 7 provides visualizations of the cell embeddings produced by CFAN and CGRAN trained on four cell types other than the type in blue. Although lacking the knowledge of the cell type in blue, both models generate distinguishing embedding vectors for the blue cell type that can be easily recognized and separated as a new cell type.

The second strategy is based on statistical analysis. For each cell, both CFAN and CGRAN output a softmax vector containing elements referring to the probabilities of the known cell types. A cell is called 'ambiguous' if the maximal element value of its softmax vector is below a threshold t. The proportion of the ambiguous cells in the sample of cells from known types is first calculated as p_{ref} for reference. When predicting the cell types for a new sample of cells, the proportion of the ambiguous cells in the new sample is calculated as p. Novel cell types may exist in the new sample if p is significantly higher than p_{ref}. This strategy is tested on GSE70580 with 4 cell types as shown in Table 7. In each setting, one cell type is masked, and both CFAN and CGRAN are trained on the remaining three cell types. The threshold t is set to 0.995 and 0.7 for CFAN and CGRAN (t can be set to a value that results in a small p_{ref}). p_{ref} and p are then calculated in the sample of the three known types and the unknown type respectively. Both models produce significantly higher p compared to p_{ref}, illustrating their effectiveness in detecting new cell types.

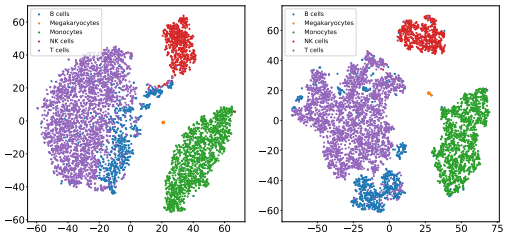

Figure 7: The t-SNE visualization of the PBMC cells produced by training CFAN (left) and CGRAN (right) with one cell type masked.

Table 7: p and p_{ref} of CFAN and CGRAN on GSE70580.

Settings	Masked Type / Masked Cells / Known Cells	NK 74 573	ILC1 126 521	ILC2 139 508	ILC3 308 339
CFAN	$p_{ref}(\%)$	5.24	5.95	6.50	43.07
	$p(\%)$	83.78	58.73	81.29	100.00
CGRAN	$p_{ref}(\%)$	26.18	17.47	39.57	4.72
	$p(\%)$	50.00	71.43	63.31	75.00

5 Conclusion

In this article, we propose two attention-based models for single-cell RNA-seq data cell type classification. Cell Feature Attention Network has higher classification accuracy compared with previous models. To further uncover the underlying mechanism behind the black box and figure out which genes make most contribution to cell type classification, we propose Cell-Gene Representation Attention Network. CGRAN learns embeddings for every cell and every gene as well as attention weights among cells and genes. With the help of local attention, CGRAN achieves satisfying classification performance on diverse datasets. By comparing the attention weights between genes and cells, CGRAN can find out highly differential genes for different cell types, which match well with the acknowledged marker genes. Moreover, by visualizing attention weights, we discover that highly differential genes from the same cell type have closer relationships, which may be an indication of their interactions in biological process. It is also worthy mentioning that CGRAN not only has outstanding classification performance and model interpretability, but also has strong transferability across different datasets and capability of discovering novel cell types, making it a strong tool for cell type classification on single-cell RNA-seq data.

References

[1] Tamim Abdelaal, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel JT Reinders, and Ahmed Mahfouz, 'A comparison of automatic cell identification methods for single-cell rna sequencing data', *Genome biology*, **20**(1), 1–19, (2019).

[2] Shaked Afik, Kathleen B Yates, Kevin Bi, Samuel Darko, Jernej Godec, Ulrike Gerdemann, Leo Swadling, Daniel C Douek, Paul Klenerman, Eleanor J Barnes, et al., 'Targeted reconstruction of t cell receptor sequence from single cell rna-seq links cdr3 length to t cell differentiation state', *Nucleic acids research*, **45**(16), e148–e148, (2017).

[3] 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).

[4] Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, et al., 'A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure', *Cell systems*, **3**(4), 346–360, (2016).

[5] Åsa K Björklund, Marianne Forkel, Simone Picelli, Viktoria Konya, Jakob Theorell, Danielle Friberg, Rickard Sandberg, and Jenny Mjösberg, 'The heterogeneity of human cd127+ innate lymphoid cells revealed by single-cell rna sequencing', *Nature immunology*, **17**(4), 451–460, (2016).

[6] Chrysothemis C Brown, Herman Gudjonson, Yuri Pritykin, Deeksha Deep, Vincent-Philippe Lavallée, Alejandra Mendoza, Rachel Fromme, Linas Mazutis, Charlotte Ariyan, Christina Leslie, et al., 'Transcriptional basis of mouse and human dendritic cell heterogeneity', *Cell*, **179**(4), 846–863, (2019).

[7] 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).

[8] Tianqi Chen and Carlos Guestrin, 'Xgboost: A scalable tree boosting system', in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, (2016).

[9] Xiaoyang Chen, Shengquan Chen, Shuang Song, Zijing Gao, Lin Hou, Xuegong Zhang, Hairong Lv, and Rui Jiang, 'Cell type annotation of single-cell chromatin accessibility data via supervised bayesian embedding', *Nature Machine Intelligence*, **4**(2), 116–126, (2022).

[10] Woosung Chung, Hye Hyeon Eum, Hae-Ock Lee, Kyung-Min Lee, Han-Byoel Lee, Kyu-Tae Kim, Han Suk Ryu, Sangmin Kim, Jeong Eon Lee, Yeon Hee Park, et al., 'Single-cell rna-seq enables comprehensive tumour and immune cell profiling in primary breast cancer', *Nature communications*, **8**(1), 1–12, (2017).

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'BERT: pre-training of deep bidirectional transformers for language understanding', in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, (2019).

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, 'An image is worth 16x16 words: Transformers for image recognition at scale', in *9th International Conference on Learning Representations, ICLR 2021*, (2021).

[13] Oscar Franzén, Li-Ming Gan, and Johan LM Björkegren, 'Panglaodb: a web server for exploration of mouse and human single-cell rna sequencing data', *Database*, **2019**, (2019).

[14] Xinyi Guo, Yuanyuan Zhang, Liangtao Zheng, Chunhong Zheng, Jintao Song, Qiming Zhang, Boxi Kang, Zhouzerui Liu, Liang Jin, Rui Xing, et al., 'Global characterization of t cells in non-small-cell lung cancer by single-cell sequencing', *Nature medicine*, **24**(7), 978–985, (2018).

[15] Mo Huang, Jingshu Wang, Eduardo Torre, Hannah Dueck, Sydney Shaffer, Roberto Bonasio, John I Murray, Arjun Raj, Mingyao Li, and Nancy R Zhang, 'Saver: gene expression recovery for single-cell rna sequencing', *Nature Methods*, **15**(7), 539–542, (2018).

[16] Jacob C Kimmel and David R Kelley, 'Semisupervised adversarial neural networks for single-cell classification', *Genome research*, **31**(10), 1781–1793, (2021).

[17] 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).

[18] Manu P Kumar, Jinyan Du, Georgia Lagoudas, Yang Jiao, Andrew Sawyer, Daryl C Drummond, Douglas A Lauffenburger, and Andreas Raue, 'Analysis of single-cell rna-seq identifies cell-cell communication associated with tumor characteristics', *Cell reports*, **25**(6), 1458–1468, (2018).

[19] Jialong Liang, Wanshi Cai, and Zhongsheng Sun, 'Single-cell sequencing technologies: current and future', *Journal of Genetics and Genomics*, **41**(10), 513–528, (2014).

[20] Maia Limbach, Mario Saare, Liina Tserel, Kai Kisand, Triin Eglit, Sascha Sauer, Tomas Axelsson, Ann-Christine Syvänen, Andres Metspalu, Lili Milani, et al., 'Epigenetic profiling in cd4+ and cd8+ t cells from graves' disease patients reveals changes in genes associated with t cell receptor signaling', *Journal of autoimmunity*, **67**, 46–56, (2016).

[21] Romain Lopez, Jeffrey Regier, Michael B Cole, Michael I Jordan, and Nir Yosef, 'Deep generative modeling for single-cell transcriptomics', *Nature Methods*, **15**(12), 1053–1058, (2018).

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

[23] Mauro J Muraro, Gitanjali Dharmadhikari, Dominic Grün, Nathalie Groen, Tim Dielen, Erik Jansen, Leon Van Gurp, Marten A Engelse, Francoise Carlotti, Eelco Jp De Koning, et al., 'A single-cell transcriptome atlas of the human pancreas', *Cell systems*, **3**(4), 385–394, (2016).

[24] Emma Pierson and Christopher Yau, 'Zifa: Dimensionality reduction for zero-inflated single-cell gene expression analysis', *Genome biology*, **16**(1), 1–10, (2015).

[25] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al., 'mrna-seq whole-transcriptome analysis of a single cell', *Nature Methods*, **6**(5), 377–382, (2009).

[26] Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, et al., 'Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq', *Science*, **352**(6282), 189–196, (2016).

[27] Laurens Van der Maaten and Geoffrey Hinton, 'Visualizing data using t-sne.', *Journal of machine learning research*, **9**(11), (2008).

[28] 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).

[29] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio, 'Graph Attention Networks', *International Conference on Learning Representations*, (2018).

[30] Florian Wagner and Itai Yanai, 'Moana: A robust and scalable cell type classification framework for single-cell rna-seq data', *bioRxiv*, (2018).

[31] 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).

[32] Sheng Wang, Angela Oliveira Pisco, Aaron McGeever, Maria Brbic, Marinka Zitnik, Spyros Darmanis, Jure Leskovec, Jim Karkanias, and Russ B Altman, 'Leveraging the cell ontology to classify unseen cell types', *Nature communications*, **12**(1), 1–11, (2021).

[33] Lei Zhang, Xin Yu, Liangtao Zheng, Yuanyuan Zhang, Yansen Li, Qiao Fang, Ranran Gao, Boxi Kang, Qiming Zhang, Julie Y Huang, et al., 'Lineage tracking reveals dynamic relationships of t cells in colorectal cancer', *Nature*, **564**(7735), 268–272, (2018).

[34] 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).

[35] Ze Zhang, Danni Luo, Xue Zhong, Jin Huk Choi, Yuanqing Ma, Stacy Wang, Elena Mahrt, Wei Guo, Eric W Stawiski, Zora Modrusan, et al., 'Scina: a semi-supervised subtyping algorithm of single cells and bulk samples', *Genes*, **10**(7), 531, (2019).