GRAPH LAPLACIANS AND GRAPH CONVOLUTIONAL NETWORK FOR SINGLE-CELL DATA

Yuhan Hao
New York University
yh1970@nyu.edu

Yilun Kuang
New York University
yk2516@nyu.edu

1 Introduction

Single-cell RNA sequencing (scRNA-seq) technologies enable profiling transcriptome of thousands of individual cells (1). scRNA-seq is not only commonly used to identify new cell types and states in heterogeneous tissues, but also is applied in multiple disease contexts (2). However, identifying pathological-related biological signals from scRNA data is still challenging. The output of scRNA-seq data is typically a high-dimensional gene counts matrix that contains over 20,000 genes (features) across hundreds of thousands of cells. The variance of the scRNA dataset involves technical noise, cell type biological heterogeneity, and disease-related responses. Despite this complicated source of variance in scRNA data, this technology will reshape our understanding of the disease pathological process at the single-cell level.

To address the dramatic impact of the Coronavirus Disease of 2019 (COVID-19), caused by infection of SARS-CoV-2, the rapidly increasing studies have focused on peripheral blood mononuclear cells (PBMCs) using single-cell analytics (2, 3). The commonly unsupervised method in the analysis is to perform principal component analysis (PCA) analysis, define a k-nearest neighbor (KNN) graph, find clusters (cell type) from the graph (1). Then, to find responsive cell types and related genes or pathways, they performed differential analysis for clusters between healthy and COVID-19 samples. However, this analysis treated cells from cell type as one homogenous population and assumed that cells from one cluster all have the same responses towards the disease process. In reality, the responses of cells from one population may be heterogeneous. Though there are some supervised frameworks to analyze scRNA data (4), they treat each cell individually and ignored the cell type community relation in this dataset.

In recent years, graph convolutional networks (GCN) can be applied into three tasks: graph classification, link prediction, and node classification (5). The fundamental graph operation used in GCN is to perform convolution in the spectral domain using the eigenvector of the graph Laplacian operator. While the eigen-decomposition of the Laplacian matrix is computationally expensive, O(N^2), the Chebyshev polynomials can be used for the approximation (6).

In this work, we first explored the graph Laplacians in a scRNA-seq from 3K blood cells. We constructed the KNN graph and explore the eigenvector of its Laplacian matrix. We demonstrated each eigenvector represents a partition of different cell types. The eigenvector with a small eigenvalue represents a low "frequency" cell-type information. Next, to further analyze scRNA data from the COVID-19 project (3), we chose to implement the node classification task of the GCN work introduced by Kipf & Welling (2017) in the Deep Graph Library (6). This GCN model both utilities the cell type information in the adjacency graph and learned the pathological related variation by node classification. Lastly, we used the embeddings of hidden layers in GCN to identify and validate heterogeneous cell states in blood and explored how the cells in the human blood respond to SARS-CoV-2 infection differently.

2 Methods and dataset used

2.1 Graph fourier transform

The classical Fourier transform can be interpreted as a projection of the signal f to the eigenspace spanned by the eigenfunctions of the Laplace operator. With this intuition at hand, we can define the graph Laplacian for an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ as $L = D - A$, where the adjacency matrix $A \in \mathbb{R}^{n \times n}$ is defined as $A[i, j] = 1\{e_{ij} \in \mathcal{E}\}$ and the diagonal degree matrix $D \in \mathbb{R}^{n \times n}$ is defined as $D[i, i] = \sum_j A[i, j]$. Since the graph Laplacian is a real symmetric matrix, we can obtain the eigendecomposition of the graph Laplacian $L = U \Lambda U^\top$ (7).

Given $f : \mathcal{V} \to \mathbb{R}$, the graph Fourier transform $\hat{f}$ is the expansion of f in terms of the eigenvectors of the graph Laplacian. The spectral convolutions on graph between different layer of the GCN given by the signal $x \in \mathbb{R}^N$

multiplied with a filter $g_\theta = diag(\theta);\ \theta \in \mathbb{R}^N$ in the Fourier domain is: $g_\theta \circ x = U g_\theta U^\top x$, where $U^\top x$ is the graph Fourier transform and the run time complexity is $\mathcal{O}(N^2)$. The filter $g_\theta(\Lambda)$ can be approximated by the Chebyshev polynomials $T_k(x)$ up to the K^{th} order with complexity $\mathcal{O}(|\mathcal{E}|)$:

$$g_{\theta'}(\Lambda) \approx \sum_{k=0}^{K} \theta'_k T_k(\tilde{\Lambda}), \tag{1}$$

where $\tilde{\Lambda} = \frac{2}{\lambda_{\max}}\mathcal{L} - I_N$, $\lambda_{\max}$ is the largest eigenvalue of $\mathcal{L}$, $T_k(x) := 2xT_{k-1}(x) - T_{k-2}(x), T_0(x) := 1, T_1(x) := x$, and $\theta' \in \mathbb{R}^K$ is a vector of Chebyshev coefficients (Kipf & Welling, 2017) (6).

2.2 Graph convolutional network

In this work, we built a two-layer GCN for node classification (healthy or COVID-19) on an undirectly KNN graph. We first performed PCA analysis and constructed the KNN graph (A) for cells from both healthy and COVID-19 patients. Then, we normalized KNN graph A as $\tilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$. The forward model and the cross-entropy are defined in the same model from Kipf & Welling (2017) (6).

$$Z = f(X, A) = \text{softmax}(\tilde{A}\ \text{ReLU}(\tilde{A}XW^{(0)})W^{(1)}) \tag{2}$$

In this work, the input layer uses 5000 features and the first hidden layer has 50 dimensions. $W^{(0)} \in \mathbb{R}^{2000x50}$.

2.3 Datasets used

The blood scRNA-seq data contains 2,700 PBMC cells from 10X genomics. The COVID-19 scRNA dataset is from Su et al.,(2020) (3). It contains 550K blood cells from 254 COVID-19 patients and 16 healthy donors. In order to avoid the unbalacend dataset, we down-sampled cells from COVID-19 patients and healthy donors both to 10K cells. In this study, we used 20K cells in total and split 2K cells as validation and 2K cells as test.

3 Results

Before applying the GCN model in the single-cell analysis, we first explore what graph Laplacians represent in the scRNA data. In the visualization of 3K blood in the UMAP calculated from PCA embeddings, we can observe seven major cell types in the blood which are annotated by their marker genes (Figure 1A). In their KNN graph, an undirected adjacency matrix, different cell types form blocks due to the shared neighbors, and within similar cell types (CD8 Naïve, CD4 Naïve, CD4 Memory), the block boundaries are blurred with each other (Figure 1B). We performed graph Laplacian operator on this KNN graph and the derived eigenvectors clearly show their strong associations with cell types. For example, the second eigenvector is positive in Monocytes and Dendritic cells while T cell, NK cell, and B cells are negative along the second eigenvector (Figure 1C). The fifth eigenvector is distinguishable in the sub-population of monocytes and the 49th eigenvector shows the difference among very small populations. It is interesting that with the increase of eigenvalue, the eigenvector defines the smaller and smaller cell clusters. The result demonstrates that the eigenvalue of the Laplacian matrix is analogous to the frequency in the Fourier transform.

For the application of the GCN model, we use the scRNA-seq data from Su et al. (2020). We down-sampled to 20K cells in total, a half from COVID-19 and the other half from healthy donors. We first perform an unsupervised analysis and observe that cells from healthy and COVID-19 are generally mixed in most cell types (Figure 2A). It indicates that the majority variance in the dataset is the variance among different cell types rather than the cellular responses in the COVID-19 pathological process. Next, we train the GCN model with 1000 epochs, and Table 1 shows that training, validation, and test accuracy. The training accuracy is slightly higher than validation and test groups. When we stratified the accuracy into different cell types, we found that monocytes showed the highest accuracy and it matched with previously reported results that monocytes have the strongest response in the COVID19 patients (2, 3). Lastly, we further explore the GCN model we trained. We contract a UMAP visualization using the 50-dimensional latent space from the first hidden layer. We observe that, compared with UMAP from PCA space, healthy and COVID19 cells have a better separation in the UMAP and cell-type information are retained as well.

Taken together, we show the graph Laplacians in the scRNA represent the different sizes of the cell populations. Then, we implemented a GCN model to predict cells from healthy donors or COVID-19 patients and we demonstrated that the trained GCN model has learned cell-type-specific responses towards the COVID-19 process.

4 Reference

[1] E. Z. Macosko et al., "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets," (in eng), Cell, vol. 161, no. 5, pp. 1202-1214, May 2015, doi: 10.1016/j.cell.2015.05.002.

[2] A. J. Wilk et al., "A single-cell atlas of the peripheral immune response in patients with severe COVID-19," Nat Med, vol. 26, no. 7, pp. 1070-1076, Jul 2020, doi: 10.1038/s41591-020-0944-y.

[3] Y. Su et al., "Multi-Omics Resolves a Sharp Disease-State Shift between Mild and Moderate COVID-19," (in eng), Cell, vol. 183, no. 6, pp. 1479-1495.e20, 12 2020, doi: 10.1016/j.cell.2020.10.037.

[4] M. Lotfollahi, F. A. Wolf, and F. J. Theis, "scGen predicts single-cell perturbation responses," (in eng), Nat Methods, vol. 16, no. 8, pp. 715-721, 08 2019, doi: 10.1038/s41592-019-0494-8.

[5] J. Zhou et al., "Graph neural networks: A review of methods and applications," AI Open, vol. 1, pp. 57-81, 2020.

[6] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.

[7] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on graphs via spectral graph theory," Applied and Computational Harmonic Analysis, vol. 30, no. 2, pp. 129-150, 2011.

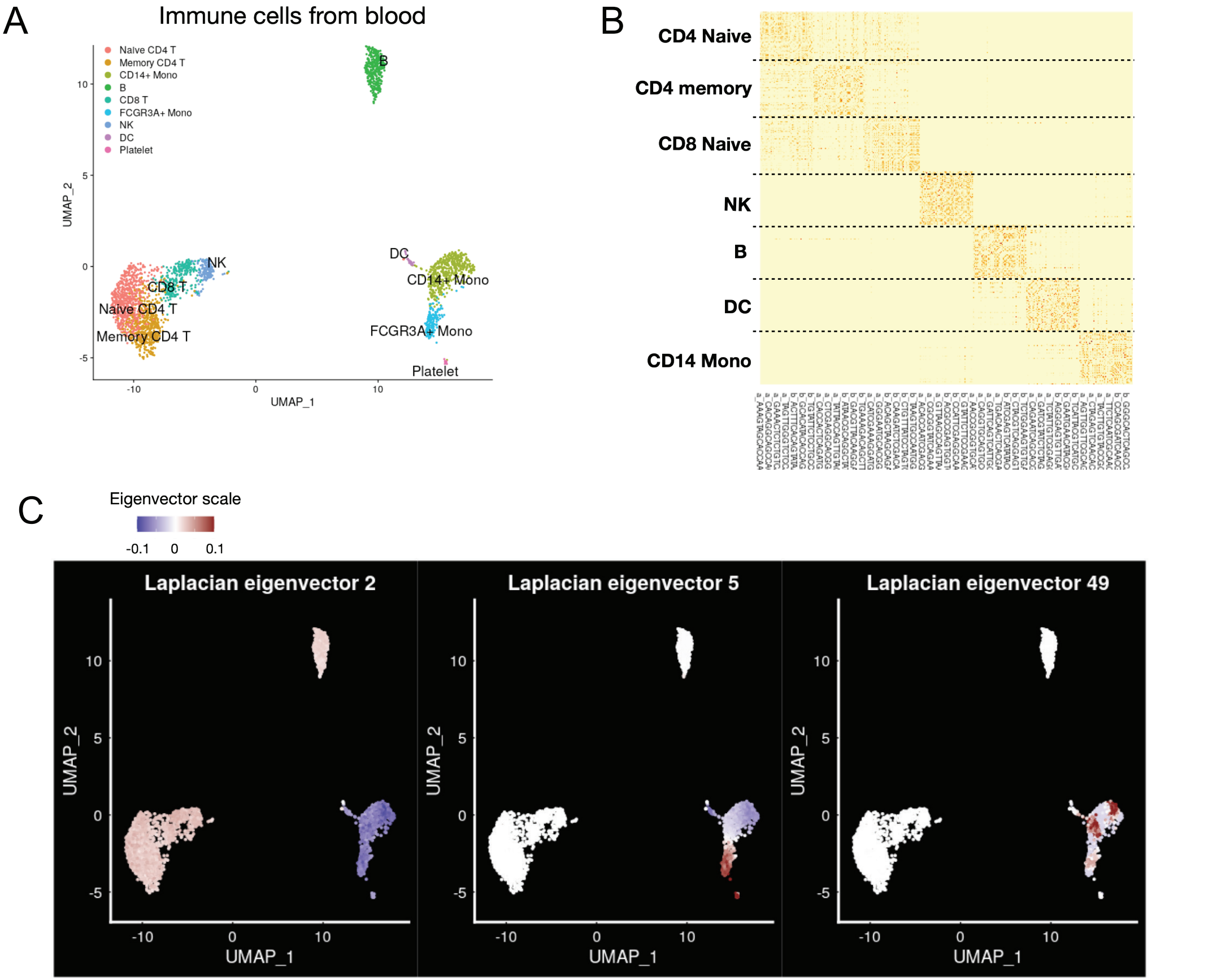

Figure 1: Evalution of graph laplacians in the blood 3k cells

(A) UMAP visulization of blood 3k cells. UMAP embeddings are calculated from PCA embeddings of this dataset. Cell annotations are their marker genes. **(B)** Single-cell KNN graph. The cells in the graph are uniformly downsampled from each celltype **(C)** Visulization of laplacian eigenvectors in the UMAP. The 2nd, 5th and 49th eigenvectors are plotted in the UMAP. The red color presents positive value, and blue color represents negative value.

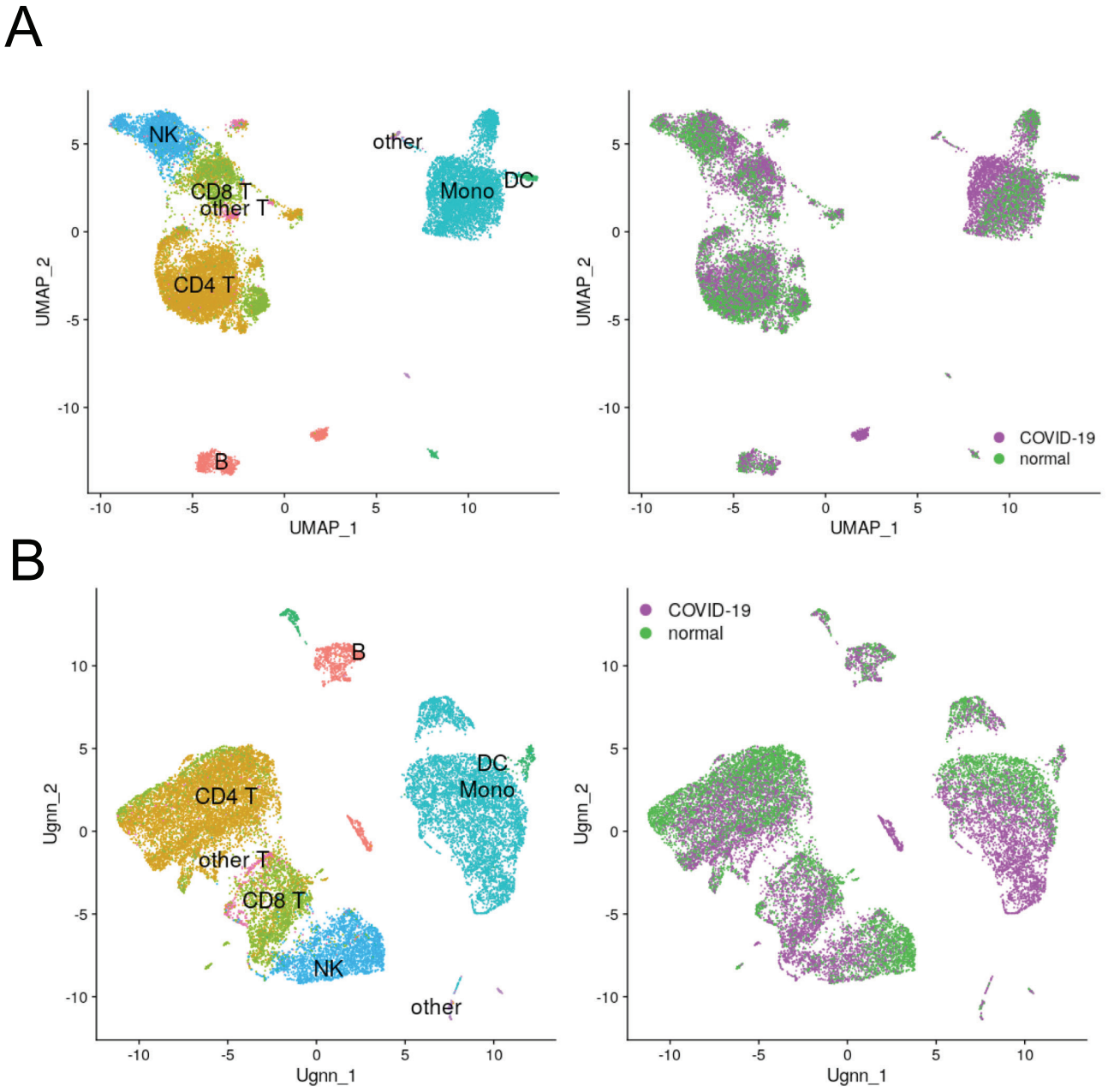

Figure 2: Visulization of 20K blood cells from COVID-19 and healthy cells

(A, B) UMAP visulization of cells in the PCA space and GCN hidden layer space. Cells are annotated by marker genes.The COVID-19 and normal status are true labels from the dataset.

Table 1: Global and cell-type-specific accuaracy of GCN model

	training	validation	test
global	0.79	0.765	0.768
CD4 T	0.766	0.765	0.744
CD8 T	0.796	0.699	0.781
other T	0.722	0.696	0.513
NK	0.799	0.775	0.795
B	0.752	0.773	0.72
Mono	**0.836**	**0.824**	**0.808**
DC	0.678	0.6	0.667
other	0.71	0.7	0.8