2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313378

Unsupervised Identification of SARS-CoV-2 Target Cell Groups via Nonlinear Dimensionality Reduction on Single-cell RNA-Seq Data

Saiteja Danda
School of Computer Science
University of Windsor
Windsor, Canada
saiteja@uwindsor.ca

Akram Vasighizaker
School of Computer Science
University of Windsor
Windsor, Canada
vasighi@uwindsor.ca

Luis Rueda
School of Computer Science
University of Windsor
Windsor, Canada
lrueda@uwindsor.ca

***Abstract*—Recent emergence of a new coronavirus, SARS-CoV-2, has caused the disease COVID-19 and has been declared a worldwide pandemic. Identification of relevant modules such as target cells is a significant step for characterizing diseases and consequently leads to better diagnosis, treatment and prognosis. High-throughput single-cell RNA-Seq (scRNA-seq) technologies have advanced in recent years, enabling researchers to investigate cells individually and understand their biological mechanisms. Computational techniques such as data clustering, which are categorized via unsupervised learning methods, are the more suitable for the pre-processing step in scRNA-seq data analysis. They can be used to identify a group of genes that belong to a specific cell type based on similar gene expression patterns. However, due to the sparsity and high-dimensional nature of this type of data, classical clustering methods are not efficient. Therefore, the use of nonlinear dimensionality reduction techniques to improve clustering results is crucial. In this work, we aim to find representative clusters of SARS-CoV-2 target cell lung by combining dimensionality reduction and clustering techniques. We first perform upstream analysis on data, including normalization and filtering using quality control metrics. We then assess the impact of different dimensionality reduction techniques on the clustering results. Our results show that modified Locally Linear Embedding combined with Independent Component Analysis have a very positive impact on clustering large-scale COVID-19 scRNA-seq data. To validate our findings, we identified target cell types involved in immune system functionality and a list of overlapping marker genes among COVID-19, Influenza A and HSV-1 infection.**

***Index Terms*—non-linear dimensionality reduction, clustering, single-cell RNA sequencing, SARS-CoV-2 target cells, COVID-19**

I. Introduction

Single-cell sequencing is an emerging technology used to capture cell information at a single-nucleotide resolution and by which individual cell can be analyzed separately [1]. As of now, all available single-cell RNA-seq (scRNA-seq) data have been generated for different purposes [2]. However, these high-dimensional and sparse data lead to some analytical challenges. Analyzing scRNA-seq data can be divided into two main categories: at the cell level and gene level. Finding cell sub-networks or highly deferentially expressed tissue-specific gene lists is one of the common challenges at the cell level [3]. Arranging cells into clusters to find the heterogeneity in the data is arguably the most significant step of any scRNA-seq data downstream analysis. This step could be used to distinguish tissue-specific sub-networks based on identified gene sets. Indeed, cell clustering aims to identify cell sub-types based on the patterns embedded in gene expression without prior knowledge at the cell level. Since the number of genes that are profiled in scRNA-seq data is typically large, cells tend to be located close to each other following nonEuclidean, but a complex relationship in high-dimensional spaces [4]. Therefore, traditional clustering algorithms are unsuitable for this challenge, and hence, are not able to efficiently separate individual cell types. To alleviate this problem of the curse of dimensionality, several algorithms have been proposed to accurately cluster cells from scRNA-seq profiles.

Dimensionality reduction techniques have been widely used in several studies of large-scale scRNA-seq data processing [5]. Most of the previous studies use principal component analysis (PCA). However, there was no advantage in keeping the clustering performance after the changes in the data in lower dimensions [6]. Other works have also employed PCA as a pre-processing step to remove cell outliers for performing dimensionality reduction and visualization. Moreover, several studies have used unsupervised clustering models to identify rare novel cell types. For instance, the hierarchical clustering algorithm divides large clusters into smaller ones or merge each data points into larger clusters progressively. This algorithm has been employed to analyze scRNA-seq data by BackSPIN [7] and pcaReduce [8], through dimension reduction after each division or combination in an iterative manner. k-Means which is one of the most common clustering algorithms, has been employed in the Monocle, specifically for analyzing scRNA-seq data [9]. Also, the authors of [10] used the Louvain algorithm, which is based on community detection

This research work has been partially supported by the Natural Sciences and Engineering Research Council of Canada, NSERC. The work has also been partially supported by the Mitacs Internship Program (www.mitacs.ca) The authors would like to thank the University of Windsor, Office or Research Services and Innovation.

techniques to analyze complex networks [11]. However, to achieve acceptable clustering performance on scRNA-seq data, other comprehensive studies indicated that hybrid models, designed as a combination of clustering and dimensionality reduction techniques, tend to improve the clustering results [6]. They learned 20 different models using four dimensionality reduction method including PCA, non-negative matrix factorization (NMF), filter-based feature selection (FBFS), and Independent Component Analysis (ICA). They also used five clustering algorithms such as k-means, density-based spatial clustering of applications with noise (DBSCAN), fuzzy c-means, Louvain, and hierarchical clustering. Their experiments highlight the positive effect of hybrid models and showed that using feature-extraction methods could be a good way to improve clustering performance. Their experimental results indicate that Louvain combined with ICA performed well in small feature spaces.

In this paper, we proposed a model to obtain efficient and meaningful clusters of cells from large-scale COVID-19 scRNA-seq data. We focus on the combination of unsupervised dimensionality reduction followed by conventional clustering methods. We investigated different non-linear dimensionality reduction and manifold learning methods such as standard Locally Linear Embedding (LLE), modified LLE, and Laplacian eigenmaps. Also, ICA is employed to enhance visualization and clustering of the data, and combined with k-means clustering. Experimental results on a well-known scRNA-seq dataset show the power of modified LLE and ICA on clustering data in very low dimensions, providing very high accuracy and enhanced visualization.

II. Materials and Methods

The block diagram of our proposed approach is depicted in Fig. 1. Based on the main pipeline, the scRNA-seq data is pre-processed based on the number of cells and the number of genes. Filtered data is then normalized and scaled. Highly variable genes are extracted as part of the feature selection step, and linear regression is one of the most widely-used methods to correct technical artifacts present in the data based on the total counts per cell and mitochondrial percentage as discussed in [10] [14]. The data obtained at this point is then processed to reduce the dimensions of the feature space into two or three dimensions; afterwards, k-means clustering is applied. Besides, We performed ICA on the lower-dimensional data followed by k-means clustering to achieve meaningful clusters and enhanced visualization.

A. Dataset

The data used in this study is a gene expression profile dataset extracted from NCBI's Gene Expression Omnibus [12], accession number GSE148729 [13]. The data contains 27,072 gene expression profiles of 48,890 human lung cell lines, which were sequenced using Illumina NextSeq 500. In this dataset, different cell lines from lung tissue, which is one of the main cellular components in the immune system, were contaminated with SARS-CoV-1 and SARS-CoV-2 and sequenced at different time slots to study the impact of infection on immune system over time.

B. Data Pre-processing and Quality Control

This step includes filtering out genes and cells based on quality metrics, normalization and scaling, feature selection, and quality control. The Python package, Scanpy, is used to perform pre-processing and quality control. To this end, we follow the typical scRNA-seq analysis workflow, as described in [14]. As shown in Fig. 1, the first step of pre-processing is to filter poorly expressed genes. Low-quality cells that are dyed, degraded, or damaged during sequencing are represented by a low or large number of expressed genes. As such, we filtered out 6,066 genes expressed in less than three cells and cells with less than 200 expressed genes. Moreover, we removed a large number of mitochondrial genes, which are the result of damaged cells [15], [16]. To remove low-quality cells, we investigated the distribution of data to estimate quality control metrics. Based on Fig. 2, the number of expressed genes, i.e., the left plot (Fig. 2a) of the figure are mainly between 500 and 2,500 genes. Also, the distribution of the proportions of mitochondrial genes, i.e., the right plot (Fig. 2c) of the figure, contains very extreme values, above 0.05. We extracted the number of genes that are less than 2,500 and mitochondrial genes less than 5%. Plot in the middle (Fig. 2b) represents total number of samples per cell.

Then, we normalized the data using the Counts Per Million (CPM) normalization combined with logarithmic scaling on the data:

$$CPM = readsMappedToGene \times \frac{1}{totalReads} \times 10^6 \quad (1)$$

where $totalReads$ is the total number of mapped reads of a sample, and $readsMappedToGene$ is the number of reads mapped to a selected gene.

At this point, we extracted highly variable genes (HVGs) as a part of the feature selection step, aiming at minimizing the search space. We then removed any random noise and held genes that highlight relevant biological information. Highly-variable genes are those genes that are expressed more or less in some cells compared to other ones. Quality control makes sure that the differences occur because of biological differences and not technical noise. The simplest approach to compute such a variation is to quantify the variance of the expression values for each gene across all samples. Here, we use log-normalized data because we want to ensure having the same log-values in the clustering and dimensionality reduction follow a consistent analysis through all steps. To perform feature selection, a good mean-variance relationship is desired. Also, a good trade-off value would help select the subset of genes that keep useful biological knowledge, while removing noise. There are several widely-used approaches to find the best threshold. Based on Fig. 4, we used a minimum of 0.5 for normalized dispersion, a maximum mean of 3, and a minimum mean of 0.0125 to select relevant genes. Finally, we obtained

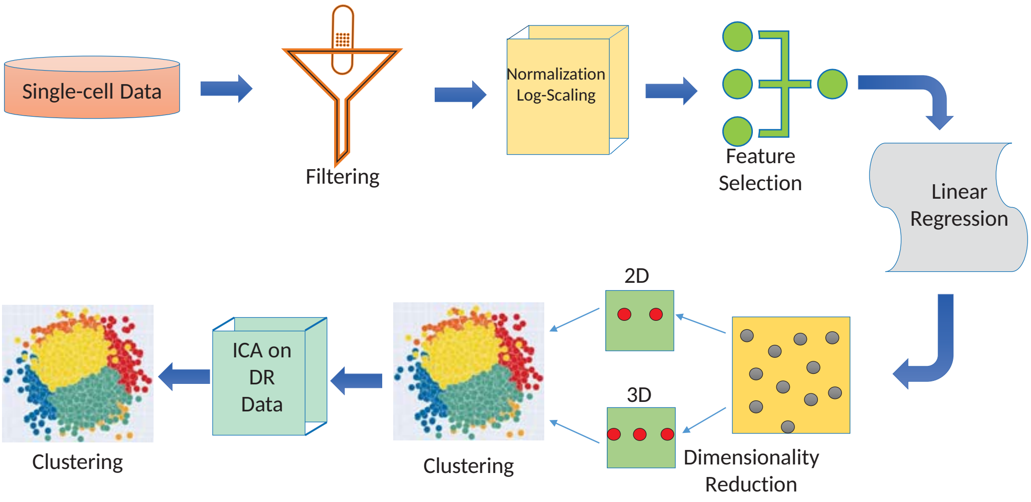


Fig. 1: Block diagram of the proposed approach.

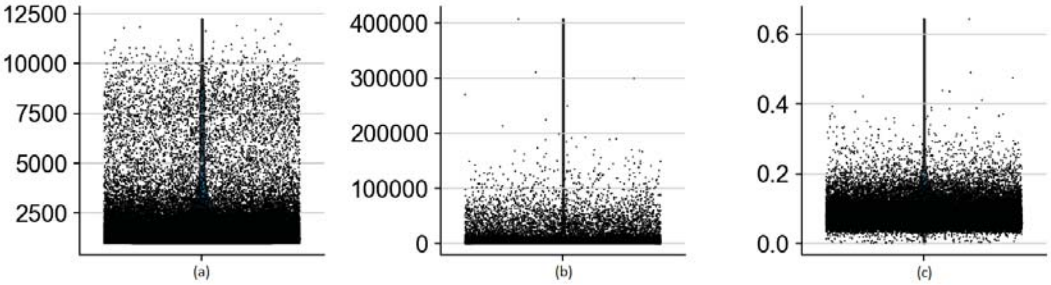


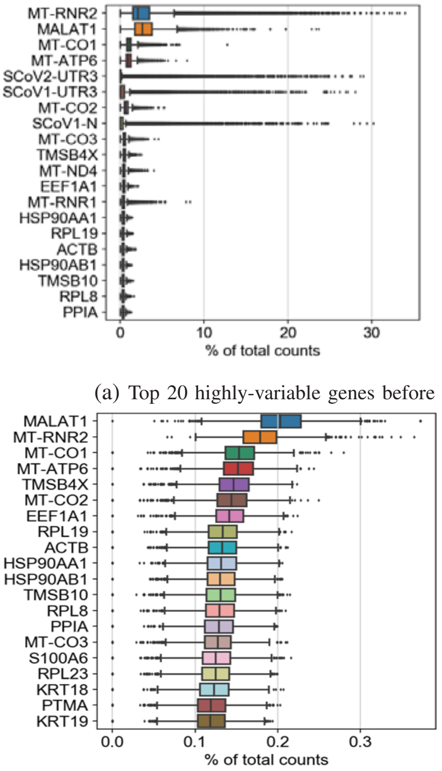
Fig. 2: (a) The number of expressed genes, (b) the total counts per cell, and (c) the percentage of mitochondrial genes.

2,194 genes with 3,791 cells for downstream analysis. The normalized dispersion is obtained by scaling the mean and standard deviation of the dispersion for genes falling into a given bin for the mean expression of genes. This means that for each bin of mean expression, highly-variable genes are selected. The 20 top genes extracted before and after normalization are shown in Fig. 3.

C. *Dimensionality Reduction*

High-dimensional gene expression data is unprecedentedly rich and should be well-explored. In a single-cell expression profile, each gene appears as a dimension of the data. As such, dimensionality reduction techniques tend to summarize biological features in fewer dimensions. With two genes, we can obtain two-dimensional points, each representing a cell. To reduce the number of individual dimensions, we aim to perform dimensionality reduction to obtain the most informative genes compressed into a smaller number of dimensions. As a result, we are able to perform the downstream analysis with less computational effort. In this regard, we used some of the dimensionality reduction and manifold learning techniques such as LLE, Laplacian eigenmaps, and ICA on this dataset. Here, high-dimensional data is reduced to two and three dimensions. As a result, we obtain the most informative components, which are further used for clustering.

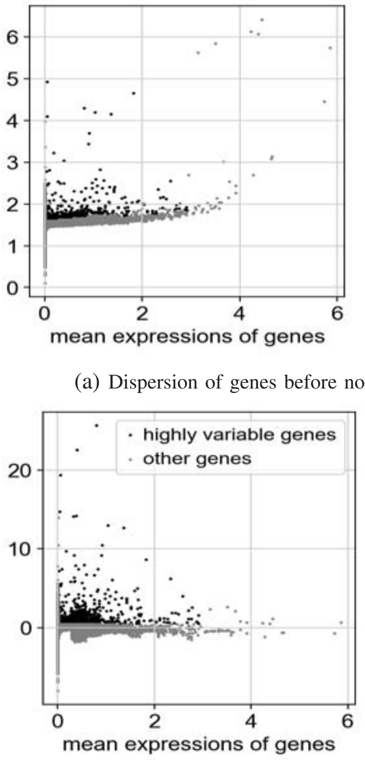
1) Locally Linear Embedding: LLE succeeds in discovering the underlying structure of the manifold when used for dimensionality reduction. This technique is empowered by preserving "locality" of the data, when reduced to lower dimensions. In addition, LLE is capable of generating highly nonlinear embeddings. Consider the sample points in a high-dimensional space, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, where $\{\mathbf{x}_j, j \in N\}$ and the weight matrix is represented by $\mathbf{W} = \{w_{ij}\}$. First, a directed graph $\mathcal{G} = (\mathbf{X}, \mathbf{E}, \mathbf{W})$ is constructed, where the edges of the graph, $\mathbf{E} = \{e_{ij}\}$, represent the neighbourhood relations among sample points, $\mathbf{X}$, in the high-dimensional space. Next,



(a) Top 20 highly-variable genes before normalization.

(b) Top 20 highly-variable genes after normalization.

Fig. 3: Comparison of the top 20 highly-variable genes before and after normalization.



(a) Dispersion of genes before normalization.

(b) Dispersion of genes after normalization.

Fig. 4: Comparison of dispersion of normalized and not normalized genes.

the weights $\mathbf{W} = \{w_{ij}\}$ are assigned to the edges of the graph. The optimal weights $\mathbf{W} = \{w_{ij}\}$ are computed by solving the following constrained least-squared problem [17]:

$$\min \ \mathbf{x}_i - \sum_{k \in K_n} w_{kn}\mathbf{x}_k \quad \text{s.t.} \sum_{k \in K_n} w_{kn} = 1\,. \qquad (2)$$

In the second step, the weights are assigned to each edge of the graph, and each sample is considered as a small linear patch of the sub-manifold. Finally, the weights are computed locally and linearly in the data by reconstructing each input pattern from its k-nearest neighbours, where the reconstruction error, ϵ_r, is calculated in terms of the mean squared error (MSE) as follows:

$$\epsilon_r = \sum_{i=1}^{n} \mathbf{x}_i - \sum_{k \in K_i} w_{ki}\mathbf{x}_k^{\ 2} \qquad (3)$$

Modified LLE (MLLE), is an enhanced version of standard LLE and has been shown to be closely related to Local Tangent Space Alignment (LTSA) [18]. MLLE attempts to exploit the dense relations that exist in the embedding space.

2) Other Dimensionality Reduction Methods: The Laplacian eigenmaps is a computationally effective approach to nonlinear dimensionality reduction that possesses locality-preserving properties and a natural connection to clustering [19]. Laplacian eigenmaps are similar to LLE. Given the input samples $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, the k nearest neighbours are computed as the first step of the algorithm.
Typically, the weights are constant, such as $w_{ij} = 1/k$ or $w_{ij} = e^{-(\frac{||x_i - x_j||^2}{s})}$ where s is the scalable parameter. Let $\mathbf{D} = \{d_{ij}\}$ be the diagonal matrix of elements $d_{ii} = \sum_{j=1}^{n} w_{ij}$. The final step is to minimize the reconstruction loss, ϵ_r, of the outputs, $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n, \}$.

$$\epsilon_r = \sum_{ij} \frac{w_{ij}\mathbf{y}_i - \mathbf{y}_j^{\ 2}}{\sqrt{\mathbf{d}_{ii}\mathbf{d}_{jj}}} \qquad (4)$$

With this function, nearby points are mapped to their nearest outputs by considering the weights $\mathbf{W}$. The minimum loss is computed from $m + 1$ eigenvectors of the matrix $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ corresponding to the smallest eigenvalues of $\mathbf{L}$. The matrix $\mathbf{L}$ is a symmetrical, normalized form of the Laplacian, given by $\mathbf{L} = \mathbf{D} - \mathbf{W}$. As in LLE, the eigenvectors corresponding to zero eigenvalues are discarded and the remaining n vectors are included to obtain the outputs $\mathbf{y}_i$ in $\mathbb{R}^n$

ICA is a dimensionality reduction method used to analyze multivariate data [20]. ICA learns an efficient linear transformation of the data and attempts to find the underlying components and sources present in the data by its simple statistical properties assumptions. Unlike other methods, the underlying vectors of the transformation are assumed to be independent of each other, and it uses a non-Gaussian structure of the data, which is important to retrieve the underlying components of the transformed data as follows:

$$\begin{aligned} \mathbf{r} &= \mathbf{A}\mathbf{s} \\ \mathbf{Y} &= \mathbf{A}\mathbf{X} \end{aligned} \tag{5}$$

where $\mathbf{r}$ and $\mathbf{s}$ are vectors and $\mathbf{A}$ is the matrix whose rows are orthogonal to each other. However, ICA assumes that the rows are linearly independent, and not necessarily orthogonal. As such, it leads to more informative components than PCA. Moreover, ICA does not require to know the output of the system to break the data into some measurements. The transformed data can then be used for cluster analysis to find a group of genes with similar expression patterns.

D. Cell Clustering

Clustering is done via k-means, which is the most popular clustering technique. This algorithm progressively finds a predetermined number of k cluster centers by minimizing the sum of the squared Euclidean distances between each center and its closest neighbour. The clusters can be denoted as $\mathbf{C} = \{\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_k\}$. This work includes a methodology that cooperatively considers ICA and k-means for clustering the cells.

E. Cluster Annotation

To annotate the cell clusters we obtained, we first extracted the top 25 differentially expressed genes as markers in each cluster using the Wilcoxon rank sum test. Then, we found the corresponding cell types of each group of marker genes in each cluster. CellKb is a search tool that collects curated cell types manually from the literature. Its knowledge base includes 403 manually curated publications from over 7,000 studies published between 2013 and 2020 to extract 1,802 different cell types. Specific marker genes of cell types in CellKb wer extracted directly from gene signature from the Human Protein Altas and MSig-db.

F. Parameter Optimization

With the aim of preserving locality, the number of neighbours used to construct the neighbourhood graph is a very important parameter in manifold learning techniques. In this work, this parameter has been learned by running the algorithm several times on the data, in a range from 4 to 16, and found 11 is the best number nearest neighbours for our experiments. Also, we use the Euclidean distance metric as the weights of the edges. Another critical step in any clustering algorithm is determining the number of clusters, k. Validity indices help measure how good the clustering is. For our dataset, we ran the validity of indices and the Silhouette score for a range of 4 to 14 and found 7 as the optimal number of clusters for this data [21].

G. Performance Evaluation

Generally speaking, the best clustering is the one that maintains high intra-cluster distance and gives the most compact clusters. In this work, we use the Silhouette coefficient, which is an evaluation metric that measures either the mean distance between a sample point and all other points in the same cluster or all other points in the next nearest neighbour cluster. Consider a set of clusters $\mathbf{C} = \{\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_k\}$, output by a clustering algorithm, k-means in our case. The Silhouette coefficient, SH, for the i^{th} sample point in cluster $\mathbf{C}_j$, where $j = 1, ..., k$, can be defined as follows:

$$SH(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{max(a(\mathbf{x}_i), b(\mathbf{x}_i))}\,, \tag{6}$$

where a is the mean distance between point $\mathbf{x}_i$ and all other points inside the cluster (intra-cluster distance), and b is the minimum mean value of the distance between a sample point $\mathbf{x}_i$ and the nearest neighbour cluster, and are calculated as:

$$\begin{aligned} a(\mathbf{x}_i) &= \frac{1}{|\mathbf{C}_k| - 1} \sum_{\mathbf{x}_j \in \mathbf{C}_k, i \neq j} d(\mathbf{x}_i, \mathbf{x}_j) \\ b(\mathbf{x}_i) &= \min_{k \neq i} \frac{1}{|\mathbf{C}_k|} \sum_{j=1}^{k} d(\mathbf{x}_i, \mathbf{x}_j)\,. \end{aligned} \tag{7}$$

We also used Calinski-Harabasz (CH) and Davies-Bouldin (DB) validity of indices to assess the clustering performance. Calinski-Harabasz score [22], is a score used to evaluate the model where a higher score tells better-defined clusters. CH score is the ratio of the sum of between-clusters dispersion and of inter-cluster dispersion for all clusters that is as follows:

$$CH = \frac{tr(\mathbf{S}_B)}{tr(\mathbf{S}_W)} \times \frac{n-k}{k-1} \tag{8}$$

in which n is size of input samples, $tr(\mathbf{S}_B)$ is the trace of the between-group dispersion matrix and $tr(\mathbf{S}_W)$ is the within-cluster dispersion.

Davies-Bouldin index [23] is another validity index defined as the average of the similarity measure of each cluster. DB is computed as follows:

$$DB = \frac{1}{k} \sum_{i=1}^{k} max_{i \neq j} s_{ij}\,, \tag{9}$$

where s_{ij} is the ratio between within-cluster distances and between cluster distances, and is calculated as $s_{ij} = \frac{w_i + w_j}{d_{ij}}$. The smaller DB value the better clustering, and as such, we aim to minimize Equation (9). Here, d_{ij} is the Euclidean distance between cluster centroids μ_i and μ_j, and w_i is the within-cluster distance of cluster $\mathbf{C}_k$.

Overall, we used the Silhouette score to evaluate the clustering performance whereas CH and DB indices used to verify and find the optimal parameters, namely the best number of clusters.

III. Results and Discussion

A. Clustering Results

After applying manifold learning techniques on the data for dimensionality reduction, we performed k-means. The results are depicted in Table I, where the clustering score ranges from 0 to 1. A score close to 1 represents good quality clustering, with 1 being the best, while a score near zero indicates that the clusters are not well defined. We observe that using MLLE the clusters are obtained with a score of 0.94 and that is the best clustering obtained from our experiments. As we can see in Fig.8, the cells are compactly bounded in their clusters and decent separation between the clusters. Also, two-dimensional ICA on three-dimensional MLLE data has been shown to provide the best visualization and clustering score of 0.943 because the three-dimensional representation is carried to two-dimensional and the clusters are well characterized as shown in Fig. 10.

TABLE I: Comparison of k-means clustering score using different dimensionality reduction techniques.

DR Technique	2D k-means	3D k-means
Standard LLE	0.623	0.683
Modified LLE	**0.938**	**0.937**
Laplacian eigenmap	0.700	0.782

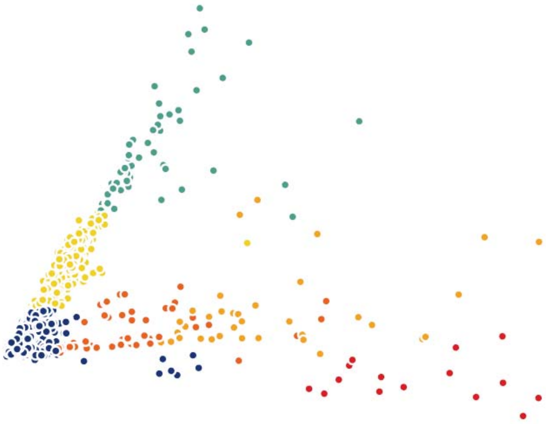


Fig. 5: k-means applied on two-dimensional Laplacian eigenmaps; outliers have been removed to enhance visualization.

More precisely, two- and three-dimensional Laplacian eigenmaps, which are depicted in Fig. 5 and 7, show good cluster separation and enhanced visualization of the data, with clustering scores of 0.70, and 0.782, respectively. We can see in Fig. 5 that cells are more scattered between the clusters using two-dimensional Laplacian eigenmaps and it is hard to capture cells to form compact clusters, whereas three-dimensional Laplacian eigenmaps give better clustering result. Also, when we applied only ICA with k-means, we obtained below-average results compared to other techniques as shown in Fig. 6 with clustering score 0.357. This is because ICA is limited to linear transformations, whereas manifold learning techniques consider data locality. As such, the latter can reveal complex relationships among the data points in higher-dimensional spaces. Therefore, we applied ICA on the dimensionally-reduced data because we observed interesting ”lines” or ”axes” in the three-dimensional data, and that led us to think that we could apply ICA to learn the linearly-independent, not necessarily orthogonal, components of the distribution of the data, and we witnessed slight improvement with clustering scores in MLLE and Standard LLE as it is displayed in Table II. Applying ICA revealed some hidden, complex relationships among the cells in the clusters which are not noticeable in three dimensions. As such, we observed a significant improvement of the clustering score using Laplacian eigenmaps since there is more dispersion of the clustering of cells in Fig. 9. We also note more compact clusters than those of the two and three-dimensional clustering whose scores are depicted in Table I.

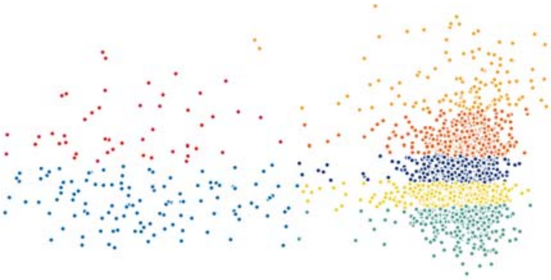


Fig. 6: k-means applied on two-dimensional ICA.

TABLE II: Results of manifold learning techniques followed by ICA and k-means clustering.

DR Technique	2D ICA-k-means on 2D DR data	2D ICA-k-means on 3D DR data.
Standard LLE	0.628	0.690
Modified LLE	0.930	**0.943**
Laplacian eigenmap	0.700	<u>0.826</u>

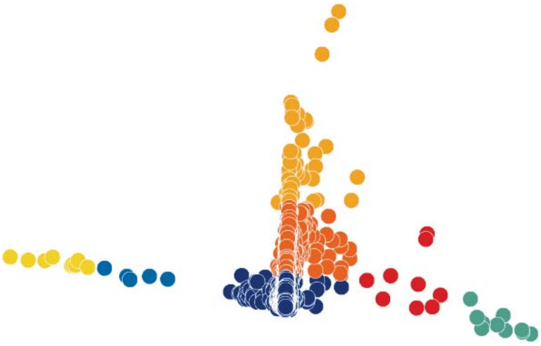


Fig. 7: k-means clustering on three-dimensional Laplacian eigenmaps.

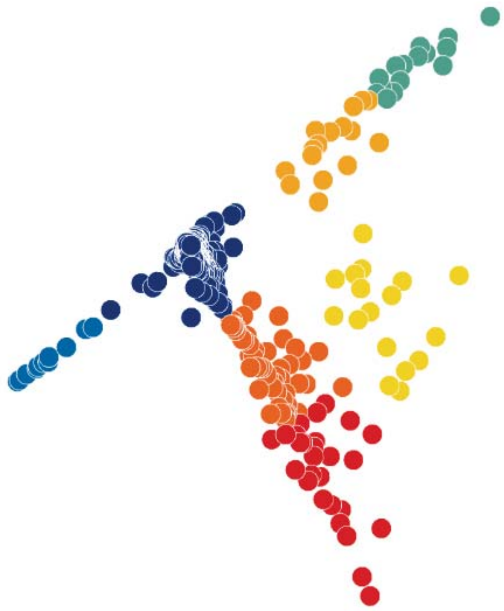


Fig. 8: k-means clustering on three-dimensional MLLE.

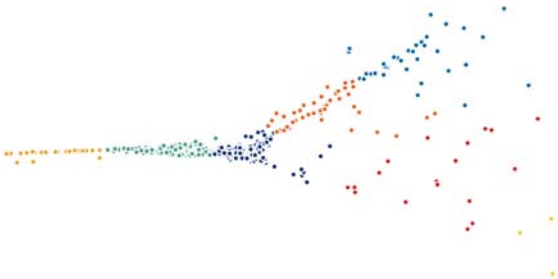


Fig. 9: Two-dimensional ICA + k-means clustering is performed on three-dimensional Laplacian eigenmaps data; outliers have been removed to enhance visualization.

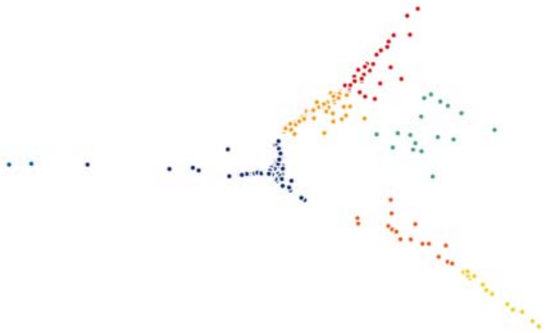


Fig. 10: Two-Dimensional ICA + k-means clustering performed on three-dimensional MLLE data; outlliers have been removed to enhance visualization.

B. Biological Assessment of the Results

The results obtained by CellKb [24] through finding overlapped genes in the literature are listed in Table III. The results show several cell types involved in immune system pathways. It is well-known that one of the main SARS-CoV2 targets is the immune system function. For example, CD4+ T cells are found on the surface of immune cells and are key cells in response to the viral infection [25]. Also, the results show that Club cells that are found in the small airways of the lungs are involved in the TAP2 binding pathway at a molecular level. TAP2 is a gene that encodes the protein antigen peptide transporter 2. In immunology, the presence of antigens in the body normally triggers an immune response. Moreover, the epithelial cells show enzyme inhibitor activity in the molecular function results. In addition, we obtained a list of overlapped marker genes that are involved in Herpes simplex virus 1 (HSV-1) infection and Influenza A pathway (Table IV). These results can be used for subsequent medical treatment or drug discovery through finding similar diseases in terms of functionality. Moreover, although numerous findings suggested potential links between HSV-1 and Alzheimer's disease (AD), a causal relation has not been demonstrated yet [26].

TABLE III: Cell types identified by our proposed method.

Cell Type
Proneural glioma stem-like cell
Th17/iTreg-stimulated CD4+ central memory T cell
Stem/Club/Hillock epithelial cell
Club cell

TABLE IV: Marker genes found in similar diseases.

Disease	Marker genes
Influenza A	RSAD2, IFIH1, MX1, STAT1 MX2, IRF7, TNFSF10, OAS1 DDX58,NFKBIA,OAS2 CXCL10,EIF2AK2,PML ICAM1,CXCL8,OAS3,STAT2
Herpes simplex virus 1 infection	IFIH1,HLA-B,STAT1,IRF7 TAP1,OAS1,DDX58,NFKBIA OAS2,STAT2,EIF2AK2,SP100 PML,HLA-E,B2M,OAS3,HLA-F

To summarize the results, performing ICA on transformed data after applying manifold learning techniques provides improved clustering results. Moreover, modified LLE combined with k-means leads to a more untangled view of the data and the corresponding clusters. Such non-linear dimensionality reduction methods have shown to be very powerful as they preserve the locality of the data from higher dimensions to lower dimensions. Evaluating the incidence of ICA as visualization and further reduction step shows even better results and the best possible clustering scores. As such, this trend leads to a research avenue that involves a combination of enhanced nonlinear manifold learning techniques such as MLLE, followed by linear methods such as ICA, which has shown to be more powerful than conventional, statistics-based methods such as PCA.

IV. Conclusion and Future Work

This work focuses on the identification of SARS-CoV-2 target cell groups using manifold learning and clustering techniques on unlabeled data. The use of clustering validity and performance measures helps to find the best clusters that are the result of combining dimensionality reduction and clustering techniques. Identifying similarities that may be a result of structural, functional, or evolutionary relationships among the genes is the main goal of clustering the cells. In our proposed two-step clustering method, we have demonstrated that genes in our dataset that have similar expression patterns were grouped in highly-scored clusters in lung tissue cell data, achieving more than 90% accuracy. Efficient nonlinear dimensionality reduction and manifold learning techniques help improve the clustering results significantly and enhance visualization in a reduced space. There are some potential applications for investigating scRNA-seq data, even beyond COVID-19. As a further analysis in the future, we aim to identify biomarker genes that are differentially expressed among different clusters of cells. Using multiple datasets with batch effect correction can improve the results as well. As such, this can lead to enhance the accuracy of classification of the cells, as a supervised learning technique, using gene expression patterns of each sub-network. Using sub-networks, we can take advantage of avoiding employing a considerable number of uninformative genes to classify the underlying cells. Moreover, performing gene set enrichment analysis to annotate a set of highly-variable genes obtaining from each cluster can reveal biomarker genes that are involved in different gene ontology terms related to COVID-19. This work attempts to highlight the power of combining linear methods such as ICA and manifold learning techniques such as MLLE for clustering to pave the way for further research in the future.

References

[1] D. Grün, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki,H. Clevers, and A. Van Oudenaarden, "Single-cell Messenger RNA Sequencing Reveals Rare Intestinal Cell Types," Nature, vol. 525,no. 7568, pp. 251–255, 2015.

[2] B. Hwang, J. H. Lee, and D. Bang, "Single-cell RNA Sequencing Technologies and Bioinformatics Pipelines," Experimental Molecular Medicine, vol. 50, no. 8, pp. 1–14, 2018.

[3] R. Sandberg, "Entering the era of Single-cell Transcriptomics in Biology and Medicine," Nature Methods, vol. 11, no. 1, pp. 22–24, 2014.

[4] V. Y. Kiselev, T. S. Andrews, and M. Hemberg, "Challenges in Unsupervised Clustering of Single-Cell RNA-Seq Data," Nature Reviews Genetics, vol. 20, no. 5, pp. 273–282, 2019.

[5] C. Dong, Y.-T. Jin, H.-L. Hua, Q.-F. Wen, S. Luo, W.-X. Zheng, and F.-B. Guo, "Comprehensive Review of the Identification of Essential Genes Using Computational Methods: Focusing on Feature Implementation and Assessment," Briefings in Bioinformatics, vol. 21, no. 1, pp. 171–181,2020.

[6] C. Feng, S. Liu, H. Zhang, R. Guan, D. Li, F. Zhou, Y. Liang, andX. Feng, "Dimension Reduction and Clustering Models for Single-Cell RNA Sequencing Data: A Comparative Study," International Journal of Molecular Sciences, vol. 21, no. 6, p. 2181, 2020.

[7] A. Zeisel, A. B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg,G. La Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz,et al., "Cell Types in the Mouse Cortex and Hippocampus Revealed by Single-cell RNA-seq," Science, vol. 347, no. 6226, pp. 1138–1142,2015.

[8] C. Yauet al., "pcaReduce: Hierarchical Clustering of Single-cell Transcriptional Profiles," BMC Bioinformatics, vol. 17, no. 1, p. 140, 2016.

[9] X. Qiu, A. Hill, J. Packer, D. Lin, Y.-A. Ma, and C. Trapnell, "Single-cell mRNA Quantification and Differential Analysis with Census," Nature Methods, vol. 14, no. 3, pp. 309–315, 2017.

[10] F. A. Wolf, P. Angerer, and F. J. Theis, "SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis," Genome Biology, vol. 19, no. 1,p. 15, 2018.

[11] M. Guerrero, F. G. Montoya, R. Baños, A. Alcayde, and C. Gil,"Adaptive Community Detection in Complex Networks Using Genetic Algorithms," Neurocomputing, vol. 266, pp. 101–113, 2017.

[12] T. Tatusova, M. DiCuccio, A. Badretdin, V. Chetvernin, E. P. Nawrocki,L. Zaslavsky, A. Lomsadze, K. D. Pruitt, M. Borodovsky, and J. Ostell,"Ncbi prokaryotic genome annotation pipeline," Nucleic Acids Research,vol. 44, no. 14, pp. 6614–6624, 2016.

[13] E. Wyler, K. Mösbauer, V. Franke, A. Diag, L. T. Gottula, R. Arsie,F. Klironomos, D. Koppstein, S. Ayoub, C. Buccitelli,et al., "Bulk and Single-cell Gene Expression Profiling of SARS-CoV-2 Infected Human Cell Lines Identifies Molecular Targets for Therapeutic Intervention," BioRxiv, 2020.

[14] M. D. Luecken and F. J. Theis, "Current Best Practices in Single-cell RNA-seq Analysis: a Tutorial," Molecular Systems Biology, vol. 15,no. 6, p. e8746, 2019.

[15] S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper,P. Lönnerberg, and S. Linnarsson, "Quantitative Single-cell RNA-seq with Unique Molecular Identifiers," Nature Methods, vol. 11, no. 2,p. 163, 2014.

[16] T. Ilicic, J. K. Kim, A. A. Kolodziejczyk, F. O. Bagger, D. J. McCarthy,J. C. Marioni, and S. A. Teichmann, "Classification of Low Quality Cells from Single-cell RNA-seq Data," Genome Biology, vol. 17, no. 1,pp. 1–15, 2016.

[17] Z. Zhang and J. Wang, "MLLE: Modified Locally Linear Embedding Using Multiple Weights," in Advances in Neural Information Processing Systems, pp. 1593–1600, 2007.

[18] J. Wang,Laplacian Eigenmaps, pp. 235–247. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[19] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," Neural Computation, vol. 15, no. 6,pp. 1373–1396, 2003.

[20] A. Hyvärinen, "Independent Component Analysis: Recent Advances, "Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 371, no. 1984, p. 20110534, 2013.

[21] P. J. Rousseeuw, "Silhouettes: a Graphical aid to the Interpretation and Validation of Cluster Analysis," Journal of computational and applied mathematics, vol. 20, pp. 53–65, 1987.

[22] T. Calinski and J. Harabasz, "A Dendrite Method for Cluster Analysis," Communications in Statistics-theory and Methods, vol. 3, no. 1, pp. 1–27, 1974.

[23] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," IEEE Transactions on Pattern Analysis and Machine Intelligence, no. 2,pp. 224–227, 1979

[24] https://www.cellkb.com/

[25] Cano-Gamez, Eddie et al. "Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4+ T cells to cytokines." Nature communications vol. 11,1 1801. 14 Apr. 2020.

[26] De Chiara, Giovanna, et al. "Recurrent herpes simplex virus-1 infection induces hallmarks of neurodegeneration and cognitive deficits in mice." PLoS pathogens 15.3. 2019.