Computers in Biology and Medicine 131 (2021) 104264

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

UMAP-assisted *K*-means clustering of large-scale SARS-CoV-2 mutation datasets

Yuta Hozumi [a], Rui Wang [a], Changchuan Yin [b], Guo-Wei Wei [a,c,d,*]

[a] Department of Mathematics, Michigan State University, MI, 48824, USA
[b] Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, IL, 60607, USA
[c] Department of Electrical and Computer Engineering, Michigan State University, MI, 48824, USA
[d] Department of Biochemistry and Molecular Biology, Michigan State University, MI, 48824, USA

ARTICLE INFO

Keywords:
PCA
t-SNE
UMAP
SARS-CoV-2
COVID-19

ABSTRACT

Coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has a worldwide devastating effect. Understanding the evolution and transmission of SARS-CoV-2 is of paramount importance for controlling, combating and preventing COVID-19. Due to the rapid growth in both the number of SARS-CoV-2 genome sequences and the number of unique mutations, the phylogenetic analysis of SARS-CoV-2 genome isolates faces an emergent large-data challenge. We introduce a dimension-reduced *K*-means clustering strategy to tackle this challenge. We examine the performance and effectiveness of three dimension-reduction algorithms: principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and uniform manifold approximation and projection (UMAP). By using four benchmark datasets, we found that UMAP is the best-suited technique due to its stable, reliable, and efficient performance, its ability to improve clustering accuracy, especially for large Jaccard distanced-based datasets, and its superior clustering visualization. The UMAP-assisted *K*-means clustering enables us to shed light on increasingly large datasets from SARS-CoV-2 genome isolates.

1. Introduction

Beginning in December 2019, coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has become one of the most deadly global pandemics in history. The COVID-19 infections in the United States (US) and other nations are still spiking. As of January 20, 2021, the World Health Organization (WHO) has reported 93,217,287 confirmed cases of COVID-19 and 2,014,957 confirmed deaths. The virus has spread to Africa, Americas, Eastern Mediterranean, Europe, South-East Asia and Western Pacific [1]. To prevent further damage to our livelihood, we must control its spread through testing, social distancing, tracking the spread, and developing effective vaccines, drugs, diagnostics, and treatments.

SARS-CoV-2 is a positive-sense single-strand RNA virus that belongs to the Nidovirales order, coronaviridae family and betacoronavirus genus [21]. To effectively track the virus, testing patients with suspected exposure to COVID-19 and sequencing the strand via PCR (polymerase chain reaction) are important. From sequencing, we can analyze patterns in mutation and predict transmission pathways. Without understanding such pathways, current efforts to find effective medicines and vaccines could become futile because mutations may change viral genome or lead to resistance. As of January 20, 2021, there are 203,344 available sequences with 26,844 unique single nucleotide polymorphisms (SNPs) with respect to the first SARS-CoV-2 sequence collected in December 2019 [36] according to our mutation tracker https://users.math.msu.edu/users/weig/SARS-CoV-2_Mutation_Tracker.html.

A popular method for understanding mutational trends is to perform phylogenetic analysis, where one clusters mutations to find evolution patterns and transmission pathways. Phylogenetic analysis has been done on the Nidovirales family [2,2,9,10,12,16] to understand genetic evolutionary pathways, protein level changes [6,12,31,32], large scale variants [31–33,35] and global trends [3,28,30]. Commonly used techniques for phylogenetic analysis include tree based methods [22] and *K*-means clustering. Both methods belong to unsupervised machine learning techniques, where ground truth is unavailable. These approaches provide valuable information for exploratory research. A main issue with phylogenetic tree analysis is that as the number of samples

* Corresponding author. Department of Mathematics, Michigan State University, MI, 48824, USA.
E-mail address: weig@msu.edu (G.-W. Wei).

https://doi.org/10.1016/j.compbiomed.2021.104264
Received 30 December 2020; Received in revised form 5 February 2021; Accepted 6 February 2021
Available online 22 February 2021

increase, its computation becomes unpractical, making it unsuitable for large genome datasets. In contrast, *K*-means scales well with sample size increase, but does not perform well when the sample size is too small. Jaccard distance is commonly used to compare genome sequences [37] because it offers a phylogenetic or topological difference between samples. However, the tradeoff to the Jaccard distance is that its feature dimension is the same as its number of samples, suggesting that for a large sample size, the number of features is also large. Since *K*-means clustering relies on computing the distance between the center of the clusters and each sample, having a large feature space can result in expensive computation, large memory requirement, and poor clustering performance. This become a significant problem as the number of SARS-CoV-2 genome isolates from patients has reached 200,000 at this point. There is a pressing need for efficient clustering methods for SARS-CoV-2 genome sequences.

One technique to address this challenge is to perform dimensional reduction on the *K*-means input dataset so that the task becomes manageable. Commonly used dimension reduction algorithms focus on two aspects: 1) the pairwise distance structure of all the data samples and 2) preservation of the local distances over the global distance. Techniques such as principal component analysis (PCA) [11], Sammon mapping [24], and multidimensional scaling (MDS) [8] aim to preserve the pairwise distance structure of the dataset. In contrast, the t-distributed stochastic neighbor embedding (t-SNE) [17,18], uniform manifold approximation and projection (UMAP) [4,19], Laplacian eigenmaps [5], and LargeVis [27] focus on the preservation of local distances. Among them, PCA, t-SNE, and UMAP are the most frequently used algorithms in the applications of cell biology, bioinformatics, and visualization [4].

PCA is a popular method used in exploratory studies, aiming to find the directions of the maximum variance in high-dimensional data and projecting them onto a new subspace to obtain low-dimensional feature spaces while preserving most of the variance. The principal components of the new subspace can be interpreted as the directions of the maximum variance, which makes the new feature axes orthogonal to each other. Although PCA is able to cover the maximum variance among features, it may lose some information if one chooses an inappropriate number of principal components. As a linear algorithm, PCA performs poorly on the features with nonlinear relationship. Therefore, in order to present high-dimensional data on low dimensional and nonlinear manifold, some nonlinear dimensional reduction algorithms such as t-SNE and UMAP are employed. T-SNE is a nonlinear method that can preserve the local and global structures of data. There are two main steps in t-SNE. First, it finds a probability distribution of the high dimensional dataset, where similar data points are given higher probability. Second, it finds a similar probability distribution in the lower dimension space, and the difference between the two distributions is minimized. However, t-SNE computes pairwise conditional probabilities for each pair of samples and involves hyperparameters that are not always easy to tune, which makes it computationally complex. UMAP is a novel manifold learning technique that also captures a nonlinear structure, which is competitive with t-SNE for visualization quality and maintains more of the global structure with superior run-time performance [19]. UMAP is built upon the mathematical work of Belkin and Niyogi on Laplacian eigenmaps, aiming to address the importance of uniform data distributions on manifolds via Riemannian geometry and the metric realization of fuzzy simplicial sets by David Spivak [26]. Similar to t-SNE, UMAP can optimize the embedded low-dimensional representation with respect to fuzzy set cross-entropy loss function by using stochastic gradient descent. The embedding is found by finding a low-dimensional projection of the data that closely matches the fuzzy topological structure of the original space. The error between two topological spaces will be minimized by optimizing the spectral layout of data in a low dimensional space.

The objective of this work is to explore efficient computational methods for the SARS-CoV-2 phylogenetic analysis of large volume of SARS-CoV-2 genome sequences. Specifically, we are interested in developing a dimension-reduction assisted clustering method. With the increase in available sequencing data, the SNP dataset of SARS-CoV-2 has run into large-data problem. By effectively analyzing clusters, we can find evolutionary trends, which will aid in finding effective medicines and vaccines. To this end, we compare the effectiveness and accuracy of PCA, t-SNE and UMAP for dimension reduction in association with the *K*-means clustering. To quantitatively evaluate the performance, we recast supervised classification problems with labels into a *K*-means clustering problems so that the accuracy of *K*-means clustering can be evaluated. As a result, the accuracy and performance of PCA, t-SNE and UMAP-assisted *K*-means clustering can be compared. By choosing the different dimensional reduction ratios, we examine the performance of these methods in *K*-means settings on four standard datasets. We found that UMAP is the most efficient, robust, reliable, and accurate algorithm. Based on this finding, we applied the UMAP-assisted *K*-means technique to large scale SARS-CoV-2 datasets generated from a Jaccard distance representation and a SNP position-based representation to further analyze its effectiveness, both in terms of speed and scalability. Our results are compared with those in the literature [32] to shed new light on SARS-CoV-2 phylogenetics.

2. Methods

2.1. Sequence and alignment

The SARS-CoV-2 sequences were obtained from GISAID databank (www.gisaid.com). Only complete genome sequences with collection date, high coverage, and without 'NNNNNN' in the sequences were considered. Each sequence was aligned to the reference sequence [36] using a multiple sequence alignment (MSA) package Clustal Omega [25]. A total of 203,344 complete SARS-CoV-2 sequences are analyzed in this work.

2.2. SNP position based features

Let N be the number of SNP profiles with respect to the SARS-CoV-2 reference genome sequence, and let M be the number of unique mutation sites. Denote V_i as the position based feature of the ith SNP profile.

$$V_i = \left[v_i^1, v_i^2, \ldots, v_i^M\right], \quad i = 1, 2, \ldots, N \tag{1}$$

is a $1 \times M$ vector. Here

$$v_i^j = \begin{cases} 1, & \text{mutation site} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

We compile this into an $N \times M$ position based feature,

$$S(i,j) = v_i^j \tag{3}$$

where each row represents a sample. Note that $S(i, j)$ is a binary representation of the position and is sparse.

2.3. Jaccard based representation

The Jaccard distance measures the dissimilarity between two sets. It is widely used in the phylogenetic studies of SNP profiles. In this work, we utilize Jaccard distance to compare SNP profiles of SARS-CoV-2 genome isolates.

Let A and B be two sets. Consider the Jaccard index between A and B, denoted $J(A, B)$, as the cardinality of the intersection divided by the cardinality of the union

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}. \tag{4}$$

The Jaccard distance between the two sets is defined by subtracting the Jaccard index from 1:

$$d_J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \tag{5}$$

We assume there are N SNP profiles or genome isolates that have been aligned to the reference SARS-CoV-2 genome. Let S_i, $i = 1, \ldots, N$, be the set with the position of the mutation of the ith sample. The Jaccard distance between two sets S_i and S_j is given by $d_J(S_i, S_j)$. Taking the pairwise distance between all the samples, we can construct the Jaccard based representation, resulting in an $N \times N$ distance matrix D

$$D(i,j) = d_J(S_i, S_j) \tag{6}$$

This distance defines a metric over the collections of all finite sets [15].

2.4. K-means clustering

K-means clustering is one of the most popular unsupervised learning methods in machine learning, where it aims to cluster or partition a data $\{x_1, \ldots, x_N\}$, $x_i \in \mathbb{R}^M$ into k clusters, $\{C_1, \ldots, C_k\}$, $k \leq N$.

K-means clustering begins with selecting k points as k cluster centers, or centroids. Then, each point in the dataset is assigned to the nearest centroid. The centroids are then updated by minimizing the within-cluster sum of squ ares (WCSS), which is defined as

$$\sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|_2^2. \tag{7}$$

Here, $\|\cdot\|_2$ denotes the l_2 norm and μ_j is the average of the data points in cluster j

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i. \tag{8}$$

This method, however, only finds the optimal centroid, given a fixed number of clusters j. In applications, we are interested in finding the optimal number of clusters as well. In order to obtain the best j clusters, elbow method was used. The optimal number of clusters can be determined via the elbow method by plotting the WCSS against the number of clusters, and choosing the inflection point position as the optimal number of clusters.

2.5. Principal component analysis

Principal component analysis (PCA) is one the most commonly used dimensional reduction techniques for the exploratory analysis of high-dimensional data [11]. Unlike other methods, there is no need for any assumptions in the data. Therefore, it is a useful method for new data, such as SARS-CoV-2 SNPs data. PCA is conducted by obtaining one component or vector at a time. The first component, termed the principal component, is the direction that maximizes the variance. The subsequent components are orthogonal to earlier ones.

Let $\{x_i\}_{i=1}^N$ be the input dataset, with N being the number of samples or data points. For each x_i, let $x_i \in \mathbb{R}^M$, where M is the number of features or data dimension. Then, we can cast the data as a matrix $X \in \mathbb{R}^{N \times M}$. PCA seeks to find a linear combination of the columns of X with maximum variance.

$$\sum_{j=1}^{n} a_j x_j = Xa, \tag{9}$$

where a_1, a_2, …, a_n are constants, and a is the vectorized a_1, a_2, …, a_n. The variance of this linear combination is defined as

$$\mathrm{var}(Xa) = a^T S a, \tag{10}$$

where S is the covariance matrix for the dataset. Note that we compute the eigenvalue of the covariance matrix. The maximum variance can be computed iteratively using Rayleigh's quotient

$$a_{(1)} = \underset{a}{\mathrm{argmax}} \frac{a^T X^T X a}{a^T a}. \tag{11}$$

The subsequent components can be computed by maximizing the variance of

$$\widehat{X}_k = X - \sum_{j=1}^{k-1} X a_j a_j^T \tag{12}$$

where k represents the kth principal component. Here, $k-1$ principal components are subtracted from the original matrix X. Therefore, the complexity of the method scales linearly with the number of components one seeks to find. In applications, we hope that the first few components give rise to a good PCA representation of the original data matrix X.

2.6. t-SNE

The t-distributed stochastic neighbor embedding (t-SNE) is a nonlinear dimensional reduction algorithm that is well suited for reducing high dimensional data into the two- or three-dimensional space. There are two main stages in t-SNE. First, it constructs a probability distribution over pairs of data such that a pair of near data points is assigned with a high probability, while a pair of farther away points is given a low probability. Second, t-SNE defines a probability distribution in the embedded space that is similar to that in the original high-dimensional space, and aims to minimize the Kullback-Leibler (KL) divergence between them [17].

Let $\{x_1, x_2, \ldots, x_N | x_i \in \mathbb{R}^M\}$ be a high dimensional input dataset. Our goal is to find an optimal low dimensional representation $\{y_1, \ldots, y_N | y_i \in \mathbb{R}^k\}$, such that $k \ll M$. The first step in t-SNE is to compute the pairwise distribution between x_i and x_j, defined as p_{ij}. However, we find the conditional probability of x_j, given x_i:

$$p_{j|i} = \frac{\exp\left(-x_i - x_j{}^2/2\sigma_i^2\right)}{\sum_{m \neq i} \exp\left(-x_i - x_m{}^2/2\sigma_i^2\right)}, \quad i \neq j, \tag{13}$$

setting $p_{i|i} = 0$, and the denominator normalizes the probability. Here, σ_i is the predefined hyperparameter called perplexity. A smaller σ_i is used for a denser dataset. Notice that this conditional probability is symmetric when the perplexity is fixed, i.e. $p_{i|j} = p_{j|i}$. Then, define the pairwise probability as

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}. \tag{14}$$

In the second step, we learn a k-dimensional embedding $\{y_1, \ldots, y_N | y_i \in \mathbb{R}^k\}$. To this end, t-SNE calculates a similar probability distribution q_{ij} defined as

$$q_{ij} = \frac{\frac{1}{1+y_i - y_j{}^2}}{\sum_m \sum_{l \neq m} \frac{1}{1+||y_m - y_l||^2}}, \quad i \neq j \tag{15}$$

and setting $q_{ii} = 0$. Finally, the low dimensional embedding $\{y_1, \ldots, y_N | y_i \in \mathbb{R}^k\}$ is found by minimizing the KL-divergence via a standard gradient descent method

$$\mathrm{KL}(P|Q) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}}, \tag{16}$$

where P and Q are the distributions for p_{ij} and q_{ij}, respectively. Note that the probability distributions in Eqs. (13) and (15) can be replaced by using many other delta sequence kernel of positive type [34].

2.7. UMAP

Uniform manifold approximation and projection (UMAP) is a nonlinear dimensional reduction method, utilizing three assumptions:

the data is uniformly distributed on Riemannian manifold, Riemannian metric is locally constant, and the manifold if locally connected. Unlike t-SNE which utilizes probabilistic model, UMAP is a graph-based algorithm. Its essentially idea is to create a predefined k-dimensional weighted UMAP graph representation of each of the original high-dimensional data point such that the edge-wise cross-entropy between the weighted graph and the original data is minimized. Finally, the k-dimensional eigenvectors of the UMAP graph are used to represent each of the original data point. In this section, a computational view of UMAP is presented. For a more theoretical account, the reader is referred to Ref. [19].

Similar to t-SNE, UMAP considers the input data $X = \{x_1, x_2, \ldots, x_N\}$, $x_i \in \mathbb{R}^M$ and look for an optimal low dimensional representation $\{y_1, \ldots, y_N | y_i \in \mathbb{R}^k\}$, such that $k < M$. The first stage is the construction of weighted k-neighbor graphs. Let define a metric $d : X \times X \rightarrow \mathbb{R}^+$. Let $k \ll M$ be a hyperparemeter, and compute the k-nearest neighbors of each x_i under a given metric d. For each x_i, let

$$\rho_i = \min\{d(x_i, x_j) | 1 \le j \le k, d(x_i, x_j) > 0\} \tag{17}$$

where σ_i is defined via

$$\sum_{j=1}^{k} \exp\left(\frac{-\max(0, d(x_i, x_j) - \rho_i)}{\sigma_i}\right) = \log_2 k. \tag{18}$$

One chooses ρ_i to ensure at least one data point is connected to x_i and having edge weight of 1, and set σ_i as a length scale parameter. One defines a weighted directed graph $\overline{G} = (V, E, \omega)$, where V is the set of vertices (in this case, the data X), E is the set of edges $E = \{(x_i, x_j) | 1 \le h \le k, 1 \le i \le N\}$, and ω is the weight for edges

$$\omega(x_i, x_j) = \exp\left(\frac{-\max(0, d(x_i, x_j) - \rho_i)}{\sigma_i}\right). \tag{19}$$

UMAP tries to define an undirected weighted graph G from directed graph $\overline{G}$ via symmetrization. Let A be the adjacency matrix of the graph $\overline{G}$. A symmetric matrix can be obtained

$$B = A + A^T - A \otimes A^T, \tag{20}$$

where T is the transpose and $\otimes$ denotes the Hadamard product. Then, the undirected weighted Laplacian G (the UMAP graph) is defined by its adjacency matrix B.

In its realization, UMAP evolves an equivalent weighted graph H with a set of points $\{y_i\}_{i=1,\cdots,N}$, utilizing attractive and repulsive forces. The attractive and repulsive forces at coordinate y_i and y_j are given by

$$\frac{-2ab||y_i - y_j||_2^{2(b-1)}}{1 + ||y_i - y_j||_2^2} w(x_i, x_j)(y_i - y_j), \text{ and} \tag{21}$$

$$\frac{2b}{\left(\varepsilon + ||y_i - y_j||_2^2\right)\left(1 + a y_i - y_{j2}^{2b}\right)} \left(1 - w(x_i, x_j)\right)(y_i - y_j) \tag{22}$$

where a, b are hyperparemeters, and ε is taken to be a small value such that the denominator does not become 0. The goal is to find the optimal low-dimensional coordinates $\{y_i\}_{i=1}^N$, $y_i \in \mathbb{R}^k$, that minimizes the edge-wise cross entropy with the original data at each point. The evolution of the UMAP graph Laplacian G can be regarded as a discrete approximation of the Laplace-Beltrami operator on a manifold defined by the data [7]. Implementation and further detail of UMAP can be found in Ref. [19].

UMAP may not work well if the data points is non-uniform. If part of the data points have k important neighbors while other part of the data points have $k' \gg k$ important neighbors, the k-dimensional UMAP will not work efficiently. Currently, there is no algorithm to automatically determine the critic minimal $k_{\min}$ for a given dataset. Additionally, weights $w(x_i, x_j)$ and force terms can be replaced by other functions that are easier to evaluate [34]. The metric d can be selected as Euclidean distance, Manhattan distance, Minkowski distance, and Chebyshev distance, depending on applications.

3. Validation

K-means clustering is one of the unsupervised learning algorithms, suggesting that neither the accuracy nor the root-mean-square error can be calculated to evaluate the performance of the K-means clustering explicitly. Additionally, K-means clustering can be problematic for high-dimensional large datasets. Dimension-reduced K-means clustering is an efficient approach. To evaluate its accuracy and performance, we convert supervised classification problems with known solutions into dimension-reduced K-means clustering problems. In doing so, we apply the K-means clustering to the classification dataset by setting the number of clusters equals to the number of the real categories. Next, in each cluster, we will take the data with the dominant label as the test for all samples and then calculate the K-means clustering accuracy for the whole dataset.

3.1. Validation data

In this work, we will consider the following classification datasets to test the performance of the clustering methods: Coil 20, Facebook large page-page network, MNIST, and Jaccard distanced-based MNIST. Previous work has been done on datasets using Euclidean and Minkowski distance for lower dimensions [17–19]. Here, we verify the result with higher reduction ratios, and tested the validity of using Jaccard distance as a metric.

- **Coil 20**: Coil 20 [20] is a dataset with 1440 gray scale images, consisting of 20 different objects, each with 72 orientation. Each image is of size 128 × 128, which was treated as a 16384 dimensional vector for dimensional reduction
- **Facebook Network**: Facebook large page-page network [23] is a page-page webgraph of verified Facebook sites. Each node represents a facebook page, and the links are the mutual links between sites. This is a binary dataset with 22,470 nodes; hence the sample size and feature size are both 22,470. Jaccard distance was computed between each nodes for the feature space.
- **MNIST**: MNIST [14] is a hand written digit dataset. Each image is a grey scale of size 28 × 28, which was treated as a 784 dimensional vector for the feature space, each with an integer value in [0, 255]. Standard normalization was used before performing dimensional reduction. There are 70,000 sample, with 10 different labels.
- **Jaccard distanced-based MNIST**: The above dataset was converted to a Jaccard distance-based dataset. This is to simulate position based mutational dataset, where 1 indicates a mutation in a particular position. Jaccard distance was used to construct the feature space, hence for each sample, the feature size is 70,000. This dataset can be viewed as an additional validation on our Jaccard distance representation.

3.2. Validation results

In the present work, we implement three popular dimensional reduction methods, PCA, UMAP, and t-SNE, for the dimension reduction and compare their performance in K-means clustering. For a uniform comparison, we reduce the dimensions of the samples by a set of ratios. The minimum between the number of features and the number of samples was taken as base of the reduction. For the Coil 20 dataset, since the numbers of samples and features were 1440 and 16384, respectively, dimension-reductions were based on 1440. For the Facebook Network, since the numbers of samples and features were both 22,470, dimension-reductions were based on 22,470. For the MNIST dataset, since the

numbers of samples and features were respectively 70,000 and 784, dimension-reductions were based on 784. Finally, for the Jaccard distanced-based MNIST dataset, since the numbers of samples and features were both 70,000, dimension-reductions were based on 70,000. Note that for the Jaccard distanced-based MNIST data, more aggressive ratios were used because the original feature size is huge, i.e., 70,000. The standard ratios of 2, 4, and 8, etc do not sufficiently reduce the dimension for effective *K*-means computation. For the purpose of visualization, two-dimensional reduction algorithms are applied to each reduction scheme. In order to validate PCA, UMAP, and t-SNE assisted *K*-means clustering, we observed their performance using labeled datasets. *K*-nearest neighbors (*K*-NN) was used to find the baseline of the reduction, which reveals how much information can be preserved in the feature after applying a dimensional reduction algorithm. For *k*-NN, 10 fold cross-validation was performed.

Notably, *K*-means clustering is an unsupervised learning algorithm, which does not have labels to evaluate the clustering performance explicitly. However, we can assess the *K*-means clustering accuracy via labeled datasets that has ground truth. In doing so, we choose the number of *K* as the original number of classes. Then, we can compared the *k*-means clustering results with the ground truth. Therefore, the accuracy can reveal the performance of the proposed dimension-reduction-assisted (*k*-means) clustering method. For the classification problem, we assume the training set is $\{(\mathbf{x}_i, y_i)|\mathbf{x}_i \in \mathbb{R}^m, y_i \in \mathbb{Z}\}_{i=1}^n$ with the $|\{y_i\}_{i=1}^n| = k$. Here n, m, and k represent the number of samples, the number of features $\{\mathbf{x}_i\}$, and the number of labels $\{y_i\}$, respectively. We set the number of clusters equals to the number of labels k. After applying the *K*-means clustering algorithm, we get k different clusters $\{\mathbf{c}_j\}_{j=1}^k$. In each cluster, we define the predictor of the *K*-means clustering in the cluster $\mathbf{c}_j$ to be:

$$\hat{\mathbf{y}}(\mathbf{c}_j) = \max\{F_j(y_1), \cdots, F_j(y_k)\}, \tag{23}$$

where $F_j(y_i)$, $\cdots$, $F_j(y_k)$ are the appearance frequencies of each label in the cluster $\mathbf{c}_j$. Then the clustering accuracy can be defined as:

$$\text{Accuracy} = \frac{\sum_i 1_{\{y_i = \hat{y}_i\}}}{n}, \tag{24}$$

where $\{\hat{y}_i\}$ are predicted labels. Moreover, other evaluation metrics such as precision, recall, and receiver operating characteristic (ROC) can also be defined accordingly.

3.2.1. Coil 20

Fig. 1 shows the performance of PCA-assisted, UMAP-assisted and t-SNE-assisted clustering of the Coil 20 dataset. For each case, the dataset were reduced to dimension 2 using default parameters, and the plots were colored with the ground truth of the Coil 20 dataset. It can be seen that PCA does not present good clustering, whereas UMAP and t-SNE show very good clusters.

Table 1 shows the accuracy of *k*-NN clustering of the Coil 20 dataset assisted by PCA, t-SNE, and UMAP with different dimensional reduction radio. The Coil 20 dataset has 1440 samples, 16,384 features, and 20 different labels. For PCA, the sklearn implementation on python was used with standard parameters. Note that for all methods, dimensions were reduced to 3 and 2 for a comparison. For t-SNE, Multicore-TSNE [29] was used because it offers up to 8 core processor, which is not available in the sklearn implementation, and it is the fastest performing t-SNE algorithm. For UMAP, we used standard parameters [19]. It can be seen that when we reduce the dimension to 3, t-SNE performs best. Moreover, when the dimensional reduction ratio is 1/100, PCA and UMAP also perform well. Notably, the *k*-NN accuracy for the data without applying any dimensional reduction algorithm is 0.956, indicating that UMAP does not provide the best clustering performance on the Coil 20 dataset. However, PCA and t-SNE will preserve the information of the original data with a dimensional reduction ratio larger than 1/100, and t-SNE even performs better for dimensional three on the Coil 20 dataset.

Table 2 describes the accuracy of *K*-means clustering of Coil 20 assisted by PCA, UMAP, and t-SNE with different dimensional reduction ratio. For consistency, we use the same set of standard parameters as *k*-NN. For the Coil 20 dataset, the accuracy of *K*-means clustering assisted by UMAP has the best performance. When the reduced dimension is 2048 (ratio 1/8), UMAP will result in a relatively high *K*-means accuracy (0.822). Moreover, although PCA performs best on *k*-NN accuracy, it performs poorly on the *K*-means accuracy, indicating that PCA is not a suitable dimensional reduction algorithm on the Coil 20 dataset. Furthermore, the highest accuracy of *K*-means clustering is 0.828, which is calculated from the t-SNE-assisted algorithm. However, the t-SNE-assisted accuracy under different reduction ratio changes dramatically. When the ratio is 1/64, the t-SNE-assisted accuracy is only 0.151, indicating that t-SNE is sensitive to the hyper-parameters settings. In contrast, the performance of UMAP is highly stable under all dimension-reduction ratios.

Note that dimension-reduced *k*-means clustering methods outperform the original *k*-means clustering. Therefore, the proposed dimension-reduced *k*-means clustering methods not only improve the *k*-means clustering efficiency, but also achieve better accuracy.

3.2.2. Facebook Network

Fig. 2 shows the visualization performance of PCA-assisted, UMAP-assisted, and t-SNE-assisted clustering of the Facebook Network. For each case, the dataset was reduced to dimension 2 using default parameters, and the plots were colored with the ground truth of the Facebook Network. Fig. 2 shows that the PCA-based data is located distributively, while the t-SNE- and UMAP-based data show clusters.

Table 3 shows the accuracy of *k*-NN clustering of the Facebook Network assisted by PCA, t-SNE, and UMAP with different dimensional reduction radio. The Facebook Network dataset has 22,470 samples

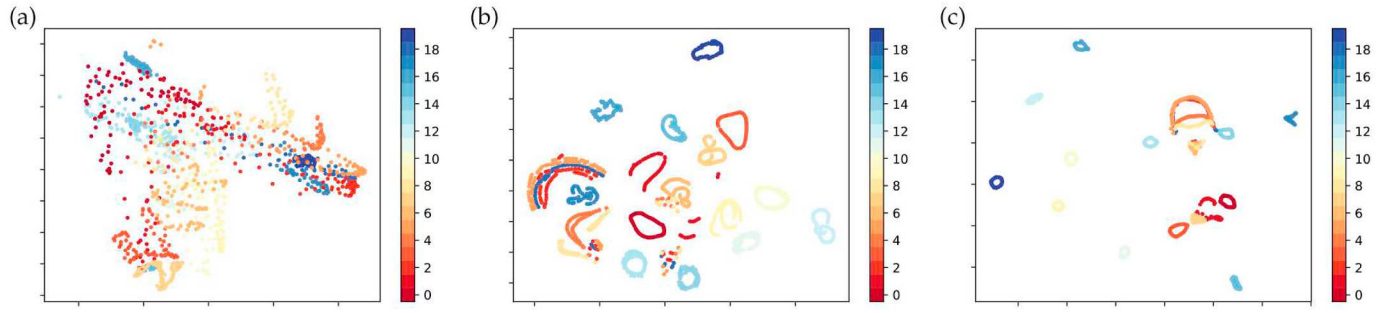


Fig. 1. Comparison of different dimensional reduction algorithms on Coil 20 dataset. Total 20 different labels are in the Coil 20 dataset, and we use the ground truth label to color each data points. (a) Feature size is reduced to dimension 2 by PCA. (b) Feature size is reduced to dimension 2 by t-SNE. (c) Feature size is reduced to dimension 2 by UMAP.

Table 1
Accuracy of *k*-NN of the Coil 20 dataset without applying any reduction algorithms, as well as the accuracy of *k*-NN assisted by PCA, UMAP and t-SNE with different dimensional reduction ratio. The sample size, feature size, and the number of labels of the Coil 20 dataset are 1440, 16384, and 20, respectively.

Dataset	*k*-NN accuracy w/o reduction	Reduced dimension	PCA accuracy	UMAP accuracy	t-SNE accuracy
Coil 20 (1440,16384,20)	0.956	720 (1/2)	0.955	0.668	0.850
		360 (1/4)	0.957	0.861	0.889
		180 (1/8)	0.973	0.867	0.881
		90 (1/16)	0.977	0.860	0.885
		45 (1/32)	0.980	0.861	0.875
		22 (1/64)	0.985	0.868	0.743
		14 (1/100)	0.730	0.851	0.878
		7 (1/200)	0.985	0.870	0.845
		3	0.850	0.863	0.959
		2	0.730	0.853	0.948

Table 2
Accuracy of *K*-means clustering of the Coil 20 dataset without applying any reduction algorithms, as well as the accuracy of *K*-means assisted by PCA, UMAP and t-SNE with different dimensional reduction ratio. The sample size, feature size, and the number of labels of the Coil 20 dataset are 1440, 16384, and 20, respectively.

Dataset	*K*-means accuracy w/o reduction	Reduced dimension	PCA accuracy	UMAP accuracy	t-SNE accuracy
Coil 20 (1440,16384,20)	0.626	720 (1/2)	0.64	0.301	0.798
		360 (1/4)	0.678	0.800	0.718
		180 (1/8)	0.633	0.822	0.648
		90 (1/16)	0.642	0.799	0.681
		45 (1/32)	0.666	0.800	0.615
		22 (1/64)	0.673	0.819	0.151
		14 (1/100)	0.631	0.817	0.154
		7 (1/200)	0.591	0.819	0.360
		3	0.561	0.800	0.780
		2	0.537	0.801	0.828

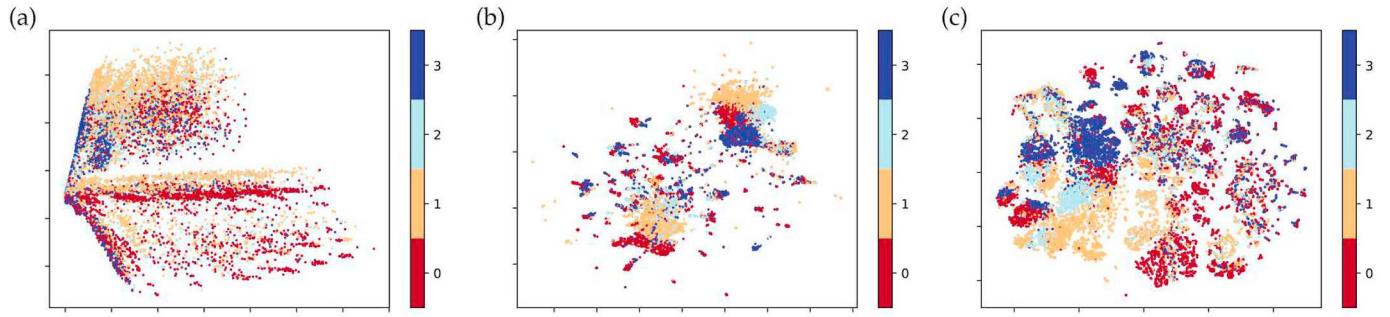


Fig. 2. Comparison of different dimensional reduction algorithms on the Facebook Network dataset. Total 4 different labels are in the Facebook Network dataset, and we use the ground truth label to color each data points. (a) Feature size is reduced to dimension 2 by PCA. (b) Feature size is reduced to dimension 2 by t-SNE. (c) Feature size is reduced to dimension 2 by UMAP.

with 4 different labels, and the feature size of the Facebook Network is also 22,470. For each algorithm, we use the same settings as the Coil 20 dataset. Without applying any dimensional reduction method, The Facebook Network has 0.755 *k*-NN accuracy. The reduced feature from PCA has the best *k*-NN performance when the reduction ratio is 1/2. UMAP has a better performance compared to PCA and t-SNE when the reduction ratio is smaller than 1/16.

Table 4 describes the accuracy of *K*-means clustering of the Facebook Network assisted by PCA, UMAP and t-SNE with different dimensional reduction ratio. PCA, UMAP, and t-SNE all have very poor performance, which may be caused by the smaller number of labels. The highest accuracy 0.427 is observed in the t-SNE-assistant algorithm with dimension 2.

Similar to the last case, UMAP-based and t-SNE-based dimension-reduced *k*-means clustering methods outperform the original *k*-means clustering with the full feature dimension. Therefore, it is useful to carry out dimension reduction before *k*-means clustering for large datasets.

Fig. 3 shows the performance of PCA-assisted, UMAP-assisted and t-SNE-assisted clustering of the MNIST dataset. The sample size of the MNIST dataset is 70000, which has 784 features with 10 different digit labels. For each case, the dataset was reduced to dimension 2 using default parameters, and the plots were colored with the ground truth of the MNIST dataset. In Fig. 3, by applying the UMAP algorithm, the clear clusters can be detected for the MNIST dataset. The t-SNE offers a reasonable clustering at dimension 2 too. However, the PCA does not provide a good clustering.

3.2.3. *MNIST*

Table 5 shows the accuracy of *k*-NN clustering of the MNIST dataset assisted by PCA, t-SNE, and UMAP with different dimensional reduction radios. For each algorithm, we use the same settings as the Coil 20 dataset. Without applying any dimensional reduction algorithms, the accuracy of *k*-NN is 0.948. By applying PCA/UMAP with the reduction ratio greater than 1/64, the accuracy of PCA/UMAP-assisted *k*-NN is at the same level without using any dimensional reduction algorithm. However, in contract with UMAP and t-SNE, when the reduced

Table 3
Accuracy of *k*-NN of the Facebook Network without applying any reduction algorithms, as well as the accuracy of *k*-NN assisted by PCA, UMAP and t-SNE with different dimensional reduction ratio. The sample size, feature size, and the number of labels of the Facebook Network are 22470, 22470, and 4, respectively.

Dataset	*K*-means accuracy w/ o reduction	Reduced dimension	PCA accuracy	UMAP accuracy	t-SNE accuracy
Facebook Network (22470, 22470, 4)	0.755	11235 (1/2)	0.756	0.360	0.307
		5617 (1/4)	0.755	0.669	0.316
		2808 (1/8)	0.754	0.754	0.355
		1404 (1/16)	0.751	0.816	0.707
		702 (1/32)	0.751	0.814	0.669
		351 (1/64)	0.746	0.815	0.690
		224 (1/100)	0.733	0.814	0.676
		112 (1/200)	0.721	0.819	0.633
		44 (1/500)	0.714	0.816	0.709
		22 (1/1000)	0.690	0.815	0.643
		3	0.552	0.801	0.741
		2	0.501	0.786	0.732

Table 4
Accuracy of *K*-means clustering of the Facebook Network without applying any reduction algorithms, as well as the accuracy of *K*-means assisted by PCA, UMAP and t-SNE with different dimensional reduction ratio. The sample size, feature size, and the number of labels of the Facebook Network are 22470, 22470, and 4, respectively.

Dataset	*K*-means accuracy w/ o reduction	Reduced dimension	PCA accuracy	UMAP accuracy	t-SNE accuracy
Facebook Network (22470, 22470, 4)	0.374	11235 (1/2)	0.331	0.306	0.306
		5617 (1/4)	0.331	0.307	0.299
		2808 (1/8)	0.331	0.411	0.314
		1404 (1/16)	0.331	0.397	0.313
		702 (1/32)	0.331	0.401	0.306
		351 (1/64)	0.331	0.400	0.308
		224 (1/100)	0.331	0.400	0.327
		112 (1/200)	0.331	0.400	0.306
		44 (1/500)	0.331	0.400	0.313
		22 (1/1000)	0.331	0.401	0.306
		3	0.332	0.351	0.344
		2	0.358	0.345	0.427

dimension is 2 or 3, PCA performs poorly. This indicates that the PCA may not be suitable for dimension-reduction for datasets with a large sample size.

Table 6 describes the accuracy of *K*-means clustering of the MNIST dataset assisted by PCA, UMAP, and t-SNE with different dimensional reduction ratios. By applying PCA, the accuracy of *K*-means is around 0.45. The t-SNE method performance is quite unstable, from very poor (0113) to the best (0.740), and to a relatively low value of 0.593. In contrast, we can see a stable and improved accuracy from using UMAP at various reduction ratios, indicating that the reduced feature generated by UMAP can better represent the clustering properties of the MNIST dataset compared to the PCA and t-SNE.

As observed early, the present UMAP and t-SNE-assisted *k*-means clustering methods also significantly out-perform the original *k*-means

Table 5
Accuracy of *k*-NN of the MNIST dataset without applying any reduction algorithms, as well as the accuracy of *k*-NN assisted by PCA, UMAP and t-SNE with different dimensional reduction ratio. The sample size, feature size, and the number of labels of the MNIST dataset are 70000, 784, and 10, respectively.

Dataset	*K*-means accuracy w/ o reduction	Reduced dimension	PCA accuracy	UMAP accuracy	t-SNE accuracy
MNIST (70000, 784, 10)	0.948	392 (1/2)	0.951	0.937	0.696
		196 (1/4)	0.956	0.938	0.846
		98 (1/8)	0.960	0.937	0.893
		49 (1/16)	0.961	0.937	0.886
		24 (1/32)	0.953	0.937	0.842
		12 (1/64)	0.926	0.937	0.676
		7 (1/100)	0.846	0.936	0.940
		3	0.513	0.929	0.938
		2	0.323	0.919	0.928

Table 6
Accuracy of *K*-means clustering of the MNIST dataset without applying any reduction algorithms, as well as the accuracy of *K*-means assisted by PCA, UMAP and t-SNE with different dimensional reduction ratio. The sample size, feature size, and the number of labels of the MNIST dataset are 70000, 784, and 10, respectively.

Dataset	*K*-means accuracy w/ o reduction	Reduced dimension	PCA accuracy	UMAP accuracy	t-SNE accuracy
MNIST (70000, 784, 10)	0.494	392 (1/2)	0.487	0.665	0.122
		196 (1/4)	0.492	0.667	0.113
		98 (1/8)	0.498	0.673	0.113
		49 (1/16)	0.496	0.718	0.113
		24 (1/32)	0.501	0.697	0.114
		12 (1/64)	0.489	0.682	0.138
		7 (1/100)	0.464	0.677	0.740
		3	0.365	0.727	0.537
		2	0.300	0.712	0.593

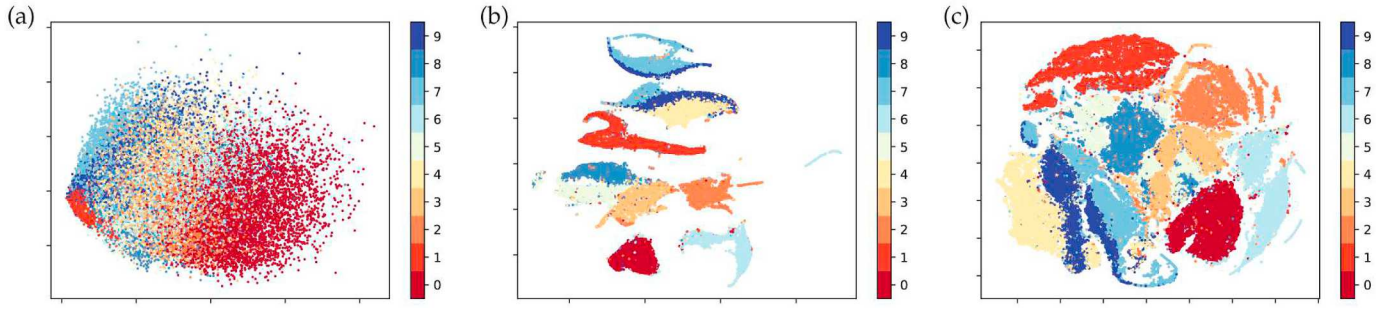


Fig. 3. Comparison of different dimensional reduction algorithms on the MNIST dataset. Total 10 different labels are in the MNIST dataset, and we use the ground truth label to color each data points. (a) Feature size is reduced to dimension 2 by PCA. (b) Feature size is reduced to dimension 2 by t-SNE. (c) Feature size is reduced to dimension 2 by UMAP.

clustering for this dataset.

3.2.4. Jaccard distanced-based MNIST

Our last validation dataset is Jaccard distanced-based MNIST. This dataset can be treated as a test on the Jaccard distance-based data representation. Fig. 4 shows the performance of PCA-assisted, UMAP-assisted, and t-SNE-assisted clustering of the Jaccard distanced-based MNIST dataset. The dataset was reduced to dimension 2 using default parameters for visualization, and the plots were colored with the ground truth of the Jaccard distanced-based MNIST dataset. From Fig. 4, we can see that UMAP provides the clearest clusters compared to PCA and t-SNE when the dimension is reduced to 2. The performance of t-SNE is reasonable while PCA does not give a good clustering.

Table 7 shows the accuracy of *k*-NN clustering of Jaccard distanced-based MNIST assisted by PCA, t-SNE, and UMAP with different dimensional reduction radios. For each algorithm, we use the same settings as the Coil 20 dataset. Notably, the *k*-NN accuracy for the data without applying any dimensional reduction algorithm is 0.958, which is at the same level as the PCA algorithm with a reduction ratio greater than 1/5000. Moreover, we can find that UMAP performs well compared to PCA and t-SNE, indicating that after applying UMAP, the reduced feature still preserves most of the valued information of the Jaccard distanced-based MNIST dataset. The stability and persistence of UMAP at various reduction ratios are the most important features.

Table 8 describes the accuracy of *K*-means clustering of the Jaccard distanced-based MNIST dataset assisted by PCA, UMAP, and t-SNE with different dimensional reduction ratio. For consistency, we will use the same standard parameters as *k*-NN. Similar to the MNIST dataset, the accuracy of *K*-means clustering assisted by UMAP still has the best performance. When the reduced dimension is 3, UMAP will result in the highest *K*-means accuracy 0.798. Noticeably, although PCA performs well on *k*-NN accuracy, it has the lowest *K*-mean accuracy, indicating that PCA is not a suitable dimensional reduction algorithm, especially for those datasets with a large number of samples. To be noted, the t-SNE accuracy at four reduced dimensions are not available due to the extremely long running time.

In a nutshell, PCA, UMAP, and t-SNE can all perform well for *k*-NN. However, for the Coil 20 dataset, UMAP performs slightly poorly, whereas the t-SNE performs well, which may be caused by a lack of data size. In order to train UMAP, it needs a suitable data size. The Coil 20 dataset has 20 labels, each with only 72 samples. This may not be enough to train UMAP properly. However, even in this case, UMAP performance is still very stable at various reduction ratios and is the best method in terms of reliability, which become the major advantages of UMAP. Another strength of UMAP comes from its dimension-reduction for *K*-means clustering. In most cases, UMAP can improve *K*-means clustering accuracy, especially for the Jaccard distanced-based MNIST dataset. Furthermore, UMAP can generate a very clear and elegant visualization of clusters with low dimensional reduction value such as 2. Additionally, UMAP performed better than PCA and t-SNE for a larger dataset (MNIST and Jaccard distanced-based MNIST). Especially for the Jaccard distanced-based MNIST data, where Jaccard distance was used as the metric, UMAP performed best, which indicates the merit of using UMAP for Jaccard distanced-based datasets, such as COVID-19 SNP datasets. Furthermore, the accuracies for *k*-NN classification and *K*-means clustering are both improved on the Jaccard distance-based MNIST dataset compared to the original MNIST dataset, which provides convincing evidence that the Jaccard distance representation will help improve the performance of the clustering on the SARS-CoV-2 mutation dataset in the following sections.

Table 7
Accuracy of *k*-NN of the Jaccard distanced-based MNIST dataset without applying any reduction algorithms, as well as the accuracy of *k*-NN assisted by PCA, UMAP and t-SNE with different dimensional reduction ratio. The sample size, feature size, and the number of labels of the Jaccard distanced-based MNIST dataset are 70000, 70000, and 10, respectively.

Dataset	*K*-means accuracy w/o reduction	Reduced dimension	PCA accuracy	UMAP accuracy	t-SNE accuracy
Jaccard distanced-based MNIST (70000, 70000, 10)	0.958	7000 (1/10)	0.958	0.958	0.588
		3500 (1/20)	0.958	0.966	0.601
		1750 (1/40)	0.958	0.967	0.725
		875 (1/80)	0.958	0.967	0.613
		437 (1/160)	0.958	0.968	0.718
		218 (1/320)	0.958	0.968	0.701
		109 (1/640)	0.958	0.968	0.873
		70 (1/1000)	0.958	0.968	0.915
		35 (1/2000)	0.956	0.968	0.872
		17 (1/5000)	0.938	0.968	0.916
		7 (1/10000)	0.867	0.967	0.942
		3	0.487	0.965	0.939
		2	0.313	0.960	0.924

3.3. Efficiency comparison

It is important to understand the computational time behaviors of various methods. To this end, we compare computational time for three dimension-reduction techniques. Fig. 5 depicts the computational time of three methods for the four datasets under various reduction ratios. The green, orange, and blue lines represent the computational time of t-SNE, UMAP, and PCA, respectively. Some points in green line of Fig. 5

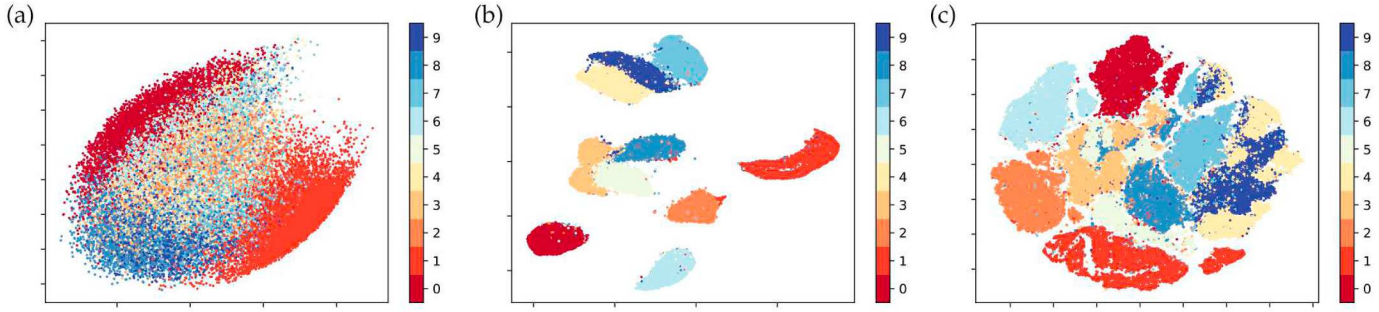


Fig. 4. Comparison of different dimensional reduction algorithms on the Jaccard distanced-based MNIST dataset. Total 10 different labels are in the Jaccard distanced-based MNIST dataset, and we use the ground truth label to color each data points. (a) Feature size is reduced to dimension 2 by PCA. (b) Feature size is reduced to dimension 2 by t-SNE. (c) Feature size is reduced to dimension 2 by UMAP.

Table 8
Accuracy of *K*-means clustering of the Jaccard distanced-based MNIST dataset without applying any reduction algorithms, as well as the accuracy of *K*-means assisted by PCA, UMAP and t-SNE with different dimensional reduction ratio. The sample size, feature size, and the number of labels of the Jaccard distanced-based MNIST dataset are 70000, 70000, and 10, respectively.

Dataset	*K*-means accuracy w/o reduction	Reduced dimension	PCA accuracy	UMAP accuracy	t-SNE accuracy
Jaccard distanced-based MNIST (70000, 70000, 10)	0.555	7000 (1/10)	0.436	0.329	0.119
		3500 (1/20)	0.436	0.693	0.120
		1750 (1/40)	0.436	0.792	0.112
		875 (1/80)	0.435	0.793	0.112
		437 (1/160)	0.435	0.793	0.114
		218 (1/320)	0.435	0.793	0.156
		109 (1/640)	0.435	0.794	0.114
		70 (1/1000)	0.436	0.793	0.113
		35 (1/2000)	0.435	0.794	0.116
		17 (1/5000)	0.436	0.793	0.113
		7 (1/10000)	0.431	0.793	0.737
		3	0.364	0.798	0.635
		2	0.261	0.791	0.635

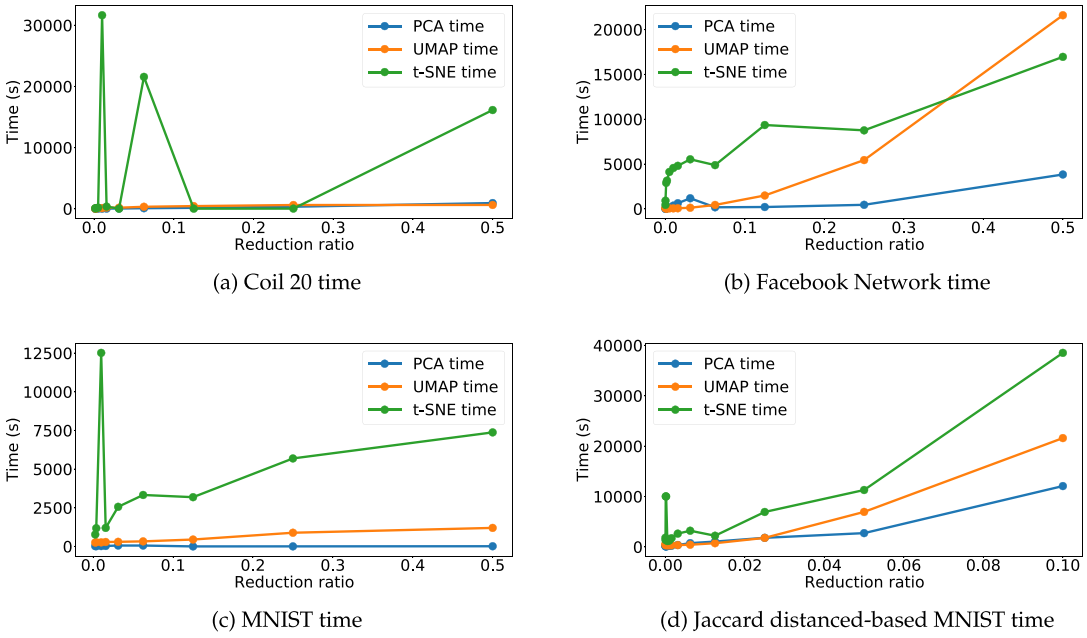
(d) are not available, which due to the extremely long running time. PCA performed best in most cases, except for the Coil 20 dataset, where UMAP had comparable computational time. This behavior is expected because PCA is a linear transformation, and its time should scale linearly with the number of components in the lower dimensional space. UMAP and t-SNE were slower than PCA, but it is evident from MNIST and Jaccard distanced-based MNIST datasets that UMAP scales better with the increase in the number of samples. Note that for Jaccard distanced-based MNIST, a higher dimension was not computed because the computational time was too long. For Facebook Network, UMAP is outperforming t-SNE; however, for higher dimensions, t-SNE computed faster. Nonetheless, from our baseline test Table 3, t-SNE does not perform well, indicating instability.

4. SARS-CoV-2 mutation clustering

4.1. World SARS-CoV-2 mutation clustering

We gather data submitted to GISAID up to January 20, 2021, and the total number of samples is 203,344. We first get the SNP information by applying the multiple sequence alignment, which leads to 26,844 unique SNPs. Next, we calculate the pairwise Jaccard distance of our dataset in order to generate the Jaccard distance-based features. Here, the number of rows is the number of samples (203,344), and the number of columns is the feature size (203,344). As we mentioned in Section 2.3, the Jaccard distance-based feature is a square matrix. However, due to the large size of samples and features, applying *K*-means clustering directly on the feature of the size of 203,344 × 203,344 is a very time-consuming process. Considering that UMAP outperforms the other two dimensional reduction algorithms (PCA and t-SNE) on the Jaccard distance-based MNIST dataset, we employ UMAP to reduce our original feature with the size of 203,344 × 203,344 to 203,344 × 203. To be noted, UMAP is a reliable and stable algorithm, which performs consistently in clustering at various reduction ratios. Therefore, there is no need to use the same reduction dimension of 203 and one can also choose a different reduction dimension value to generate similar results.

With the reduced dimension feature that has the size of 203,344 × 203, we split our SARS-CoV-2 dataset into different clusters by applying the *K*-means clustering methods. After comparing the WCSS under a different number of clusters, we find that there are 6 clusters forming within the SARS-CoV-2 population based on the elbow method (See Fig. 6), which can be determined from Fig. S1 in the Supporting



(a) Coil 20 time

(b) Facebook Network time

(c) MNIST time

(d) Jaccard distanced-based MNIST time

Fig. 5. Computational time of each reduction ratio. The green, orange and blue lines represent the computational time of t-SNE, UMAP, and PCA, respectively. Not surprisingly, PCA performs the best in the majority of cases, except for the Coil 20 dataset. UMAP and t-SNE perform worse than PCA, but UMAP scales better when there are more samples, as evident from MNIST and Jaccard distanced-based MNIST datasets. Note that for Jaccard distanced-based MNIST, the higher dimension was not computed because the computational time was too long.

Information. Table C1 in the Supporting Information shows the top 25 single mutations of each cluster. In order to understand the relationship, we also analyzed the co-mutation occurring in each cluster (Table 9). Here, we define a co-mutation as mutations that occur simultaneously in one SNP profile. For example, mutations occurring at position 241 and 3037 in a single SNP sample is a co-mutation [241, 3037]. From Table C1 in the Supporting Information and Table 9 we see the following:

Table C2 in the Supporting Information shows the cluster distributions of samples from 25 countries. Here, we use the ISO 3166–1 alpha-2 codes as the country code. The listed countries are the United Kingdom (UK), the United States (US), Australia (AU), India (IN), Switzerland (CH), Netherlands (NL), Canada (CA), France (FR), Belgium (BE), Singapore (SG), Spain (ES), Russia (RU), Portugal (PT), Denmark (DK), Sweden (SE), Austria (AT), Japan (JP), South Africa (ZA), Iceland (IS), Brazil (BR), Saudi Arabia (SA), Norway (NO), China (CN), Italy (IT), and Korea (KR). We can visualize the clusters on the world map from Fig. 7, which was visualized using Highcharts. The underlying color indicates the dominant cluster for each country. Furthermore, from table C2, we can see the following:

- SNP profiles from UK and DK are dominated in Clusters 5.
- Clusters 3's SNP profiles are predominantly found in AU. This may indicate that SARS-CoV-2 are mutating differently in AU.
- SNP profiles from the US are found mostly in Clusters 2 and 5.
- Most country's SNP profiles are found in Clusters 1,2,4,5 and 6, with some having slightly higher numbers.

Notably, in Table 9, Cluster 4 and 5 have the same co-mutations with relatively high frequencies, which indicates the Clusters 4 and 5 share the same "root". Clusters 1, 2, 3, and 6 shares the co-mutation as Clusters 4 and 5, indicating that Clusters 1, 2,3, and 6 may have branched from Cluster 4 and 5 in the 203-dimensional (203D) space. However, we cannot visualize the distribution of our reduced dataset in the 203D space. Therefore, benefit from the stable and reliable performance of UMAP at various reduction ratios, we reduce the dimension of our original dataset to 2, which enables us to observe the distribution of the dataset in the two-dimensional (2D) space. Fig. 7 visualizes the distribution of our dataset with 6 distinct clusters with 2D UMAP. It can be seen that Clusters 2, 3 and 4 share a same "root" in the middle. Clusters 3 and 6 are farther away from the center, indicating that they are a

Table 9
The frequency and occurrence percentage of SARS-CoV-2 co-mutations from each clusters in the world.

- Though Clusters 1 and 6 seem similar from the top 25 single mutations, the co-mutations tells a different story.
- Clusters 2 and 5 have high frequency of [241, 3037, 14408, 23403] mutations, but Cluster 5 has a clear co-mutation descendant with high frequency.
- Cluster 3 has a unique combination of mutation that is only popular in Cluster 3.
- Cluster 6 have high frequency of multiple co-mutations. Since it shares similarity with Clusters 4 and 5, it may be that Cluster 6 branched from Clusters 4 and 5.
- Cluster 6 has many co-mutations when compared to other clusters. As seen in table C2, the majority of the cases is found in Europe, including the United Kingdom (UK), Denmark (DK), Netherlands (NL), Switzerland (CH) and Luxemberg (LU).

Cluster	Co-mutations	Frequency	Occurrence percentage
Cluster 1	[241, 3037, 14408, 23403, 28881, 28882, 28883]	21802	0.926
Cluster 2	[241, 1059, 3037, 14408, 23403, 25563]	15008	0.660
Cluster 3	[241, 1163, 3037, 7540, 14408, 16647, 18555, 22992, 23401, 23403, 28881, 28882, 28883]	2089	0.606
Cluster 4	[241, 3037, 14408, 23403]	13387	0.936
Cluster 5	[241, 3037, 14408, 23403]	124290	0.915
Cluster 6	[241, 3037, 4543, 5629, 9526, 11497, 13993, 14408, 15766, 16889, 17019, 18877, 22992, 23403, 25563, 25710, 26735, 26876, 28975, 29399]	3279	0.940

descendants of the middle root. In addition, we looked specifically at the spike (S) protein because of its significance in viral infectivity. In all the clusters, 23403A > G (D614G) is present. Studies have shown that D614G increases the infectivity of SARS-CoV-2 [13], hence the high frequency in our data reflect such infectivity. In Clusters 1, 2 and 4, there are no significant co-mutations in the S protein. In Cluster 3, 100% of the variants contain the co-mutation [22992, 23401, 23403], which further supports its geographical isolation, where it is predominantly found in AU. Cluster 5 does not have a significant co-mutation, but the co-mutations [21614, 22227, 23403, 24334] occurred in 11290 SNP

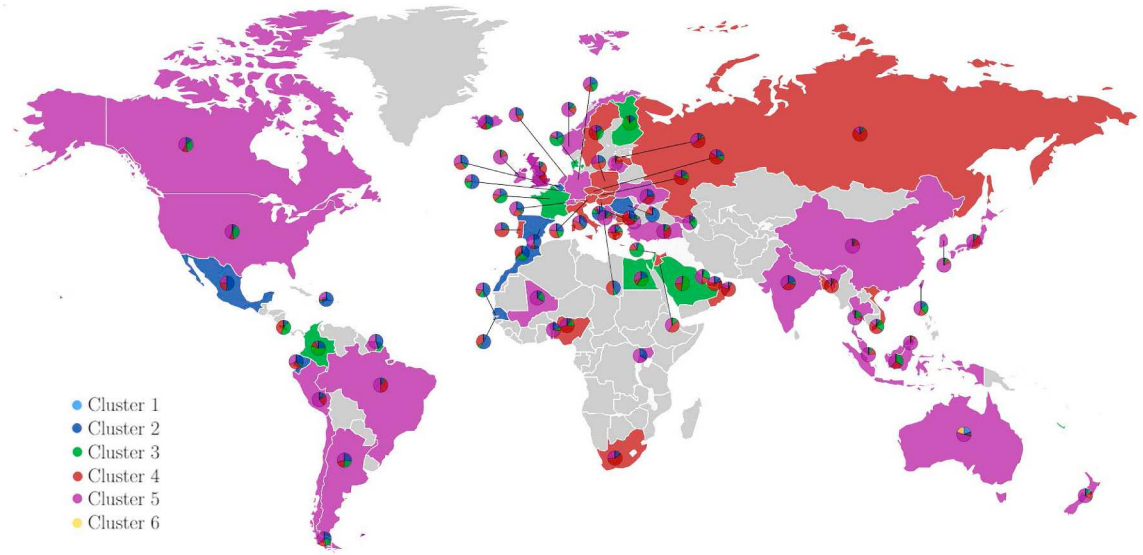


Fig. 6. Cluster distribution of the global SARS-CoV-2 mutation dataset. Using Highchart, the world map was colored, according to the dominant cluster. For example, United States have SNP profiles from all clusters, but Cluster 5 (purple) is the dominant type in the US. Only countries with more than 25 sequenced data available on GISAID were considered. Countries with fewer than 25 samples are labeled grayed.

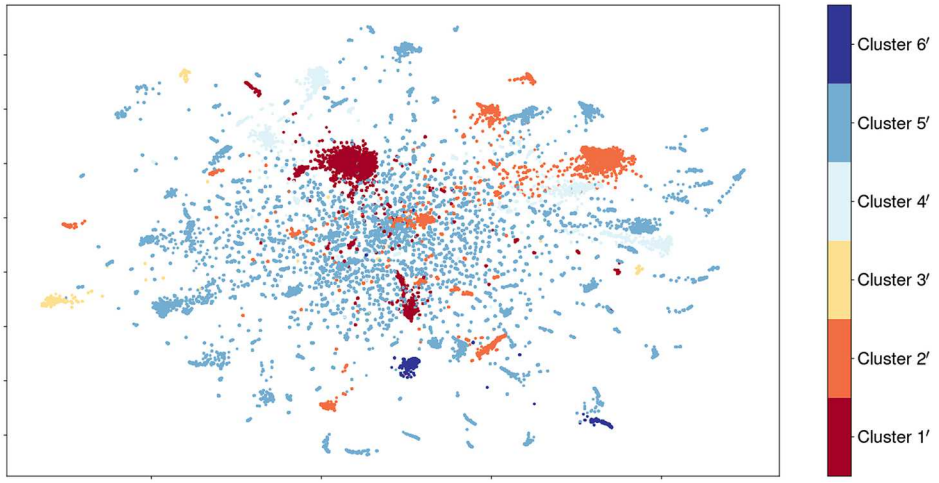


Fig. 7. 2D UMAP visualization of the world SARS-CoV-2 mutation dataset with 6 distinct clusters.

profiles (0.083). Cluster 6 has a pair of co-mutations [22992, 23403], which occurs in 99.7% of samples.

4.2. United States SARS-CoV-2 mutation clustering

In addition to analyzing the clustering in the world, SNP profiles of SARS-CoV-2 from the US were considered. In this section, the US dataset has 17164 unique single mutations and 43395 samples. Therefore, the dimension of the Jaccard distance-based dataset is 43395 × 43395. After applying the UMAP, we reduce the dimension of the original dataset to be 43395 × 216. Following the similar *K*-means clustering processes as we did for the world dataset, we find that using the elbow method, we can see from Fig. S2 in the Supporting Information that there are 6 predominant clusters forming in the United States. Fig. 8 show the US map with the cluster statistic. Here, Highchart was used to generate the plot with the pie chart. Each states were colored based on the dominant cluster.

Table C3 in the Supporting Information shows the top 25 mutations from each clusters in the United States. The cluster distribution of each states is listed in table C4. Table 10 shows the common occurring co-mutations, and we can observe the following:

- Cluster A has a high frequency of co-mutations [241, 1059, 3037, 14408, 23403, 25563], which is a descendant of common co-mutations of Cluster 2 [241, 1059, 3037, 14408, 23403, 25563] from table C3.

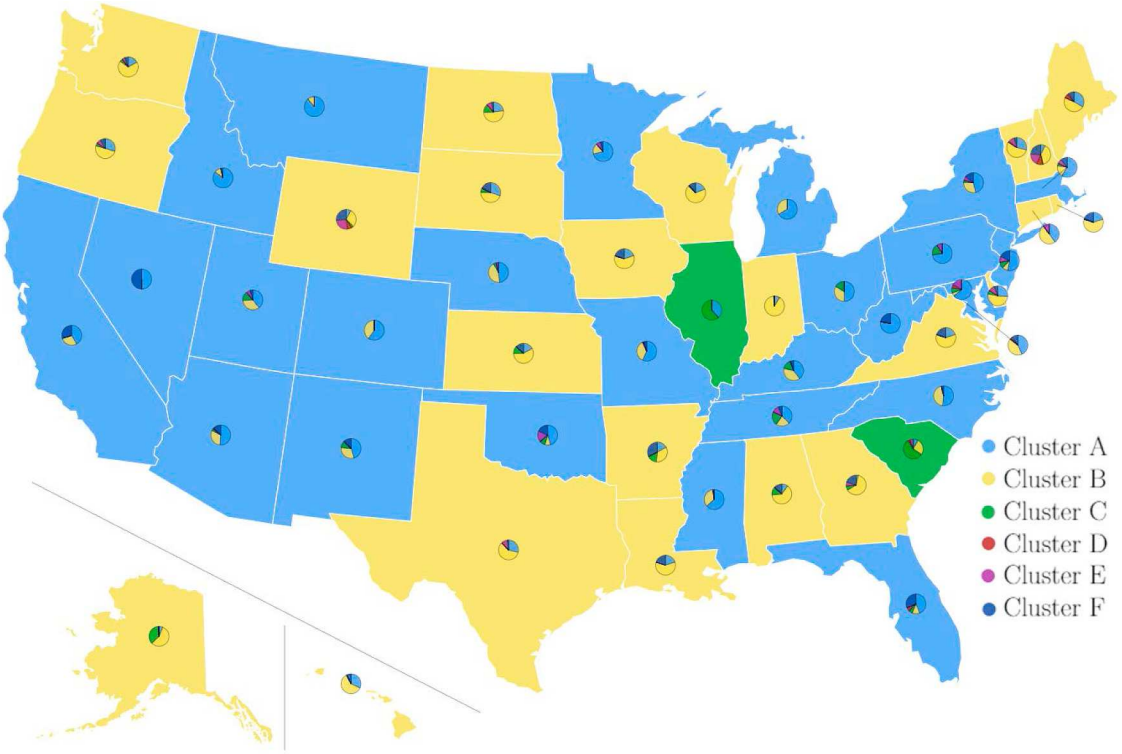


Fig. 8. Cluster distribution of United States SARS-CoV-2 mutation dataset. Using Highchart, the US map was colored, according to the dominant cluster. For example, United States have SNP profiles from all clusters, but Cluster E (purple) is the dominant type in the US. Only those countries that have more than 25 sequenced data available on GISAID were considered in the plot.

Table 10
The frequency and occurrence percentage of SARS-CoV-2 co-mutations from each clusters in US clusters.

Cluster	Co-mutations	Frequency	Occurrence percentage
Cluster A	[241, 1059, 3037, 14408, 23403, 25563]	6646	0.702
Cluster B	[241, 3037, 14408, 23403]	20442	0.932
Cluster C	[241, 3037, 14408, 23403, 28881, 28882, 28883]	4429	0.945
Cluster D	[241, 3037, 14408, 20268, 23403, 28854]	3276	0.643
Cluster E	[8782, 17747, 17858, 18060, 28144]	1183	0.744
Cluster F	[241, 1059, 3037, 11916, 14408, 18998, 23403, 25563, 29540]	501	0.789

- Cluster B has a high frequency of co-mutations [241, 3037, 14408, 23403], which is a descendant of common co-mutations of Cluster 4 and 5 [241, 3037, 14408, 23403].
- Cluster C have high frequency of co-mutations [241, 3037, 14408, 23403, 28881, 28882, 28883], which is a descendant of common co-mutations of Cluster 1 [241, 3037, 14408, 23403, 28881, 28882, 28883] from Table 10.
- Clusters D has high frequency of co-mutations [241, 3037, 14408, 20268, 23403, 28854], which is descendant of Clusters 4 and 5 [241, 3037, 14408, 23403]. US accounts for more than one third of mutations at site 23403 and half of mutations at site 28854
- Cluster E and F have a high frequency of co-mutations [8782, 17747, 17858, 18060, 28144] and [241, 1059, 3037, 11916, 14408, 18998, 23403, 25563, 29540], respectively, which are descendants of Cluster 4 and 5 [241, 3037, 14408, 23403].
- Cluster F has a high frequency of co-mutations [241, 1059, 3037, 11916, 14408, 18998, 23403, 25563, 29540], which is a descendant of Cluster 2's co-mutation [241,1059,3037,14408,23403,25563]

Notably, in Table 10, Cluster B has a co-mutation that is present in Clusters A, C, D and F, indicating that Clusters A, C, D, and E are descendants of Cluster B. Interestingly, Cluster E has a completely different set of co-mutations as the other clusters, indicating that they are a different strands of mutation. Considering the stability and reliability of UMAP at various reduction ratios, we employ UMAP to the original US dataset with reduced dimension 2, aiming to observe the distribution of the dataset in the 2D space. Fig. 9 illustrates the 2D visualization of the US dataset with 6 distinct clusters. We can see that there are 3 clusters (Clusters A′, B′, and C′) share the same "root" located in the middle of the figure, while the other 3 clusters (Clusters D′, E′, and F′) are not. Cluster E′ is quite distinct from other clusters. This confirms our deduction about why Clusters E′ has a high frequency of different co-mutations in Table 10. In addition Cluster D′ is located close to Cluster A', which may indicate that they have similar root that diverted.

In addition, we looked at co-mutations on the S protein. Every cluster, except for Cluster E, contains mutation 23403, which is expected due to its ability to increase the infectivity of SARS-CoV-2. Clusters A, C, and F does not have any significant co-mutation occurring in the S protein, aside from 23403. Cluster E does not have a significant co-mutation nor a significant mutation in the S protein. Cluster B has co-mutations [22255, 23403], which occur in 780 samples. Cluster D has co-mutations [23403, 23604, 24076] that occur in 892 samples.

5. Discussion

In this section, we compared our past results [32] with our new method to gain a different perspective in clustering with the SNP profiles of COVID-19. In our previous work, a total of 8309 unique single mutations are detected in 15,140 SARS-CoV-2 isolates. Here, we also calculate the pairwise distance among 15140 SNP profiles and set the number of clusters to be six. Table C5 shows the cluster distribution of samples from the 15 countries [32]. The listed countries are the United States (US), Canada (CA), Australia (AU), United Kingdom (UK), Germany (DE), France (FR), Italy (IT), Russia (RU), China (CN), Japan (JP), Korean (KR), India (IN), Spain (ES), Saudi Arabia (SA), and Turkey (TR), and we use Cluster I, II, III, IV, V, and VI to represent six clusters without applying any dimensional reduction algorithm. Table C6 lists the cluster distribution of samples from the same 15 countries, where we use I_p, II_p, III_p, IV_p, V_p, and VI_p to represent six clusters performed by PCA with the reduction ratio to be 1/160. Table C7 lists the cluster distribution of samples from the same 15 countries, where we use I_u, II_u, III_u, IV_u, V_u, and VI_u to represent six clusters performed by UMAP with the reduction ratio setting to be 1/160. Noticeably, the SNP profile is focused in Cluster I_u, whereas in the non-reduced version, the samples are more spread out. This may be caused by the large number of features, making computed distance between the centroid and each data too similar, and leading to samples being placed in incorrect clusters.

Not surprisingly, PCA and the original method for [32] has nearly identical result. It has been shown in Ref. [32] that PCA is the continuous solution of the cluster indicators in the *K*-means clustering method. On the other hand, UMAP shows a slightly different result. In the PCA method, the distribution is more spread out. In addition, the top occurrence for each country is higher for UMAP. On the other hand, we see that there are more samples in Cluster I_u for UMAP, which may

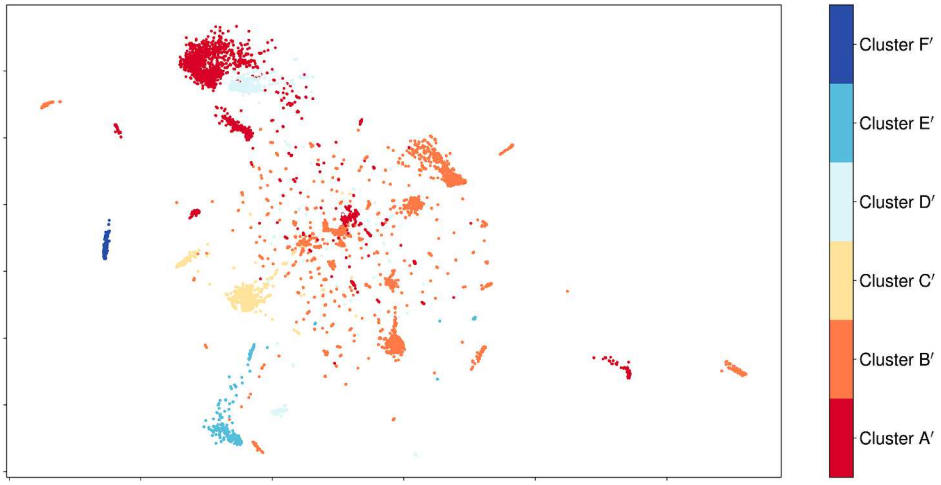


Fig. 9. The 2D UMAP visualization of the US SARS-CoV-2 mutation dataset with 6 distinct clusters.

indicate that mutations in Cluster I_u are the main strand.

Moreover, Fig. 10 illustrates the 2D visualizations of the US dataset up to June 01, 2020, with 6 distinct clusters by applying two different dimensional reduction algorithms. We can see that the data distribute disorderly under both PCA- and UMAP-assisted *K*-means clustering algorithms. Specifically, the PCA-assisted algorithm has a really poor clustering performance, while the UMAP-assisted algorithm forms more clear and better clusters than the PCA-assisted algorithm, which is consistent with our previous analysis in Section 3.1.

Table 11 shows co-mutations occurred in each cluster from the UMAP-assisted *K*-means from data collected up to June 01, 2020. Cluster III_u has 2 dominant co-mutations. Note that the dataset had 15,140 SARS-CoV-2 isolates, whereas our current dataset has over 200,000 isolates. Nonetheless, we can compare the clusters to see which clusters persists. Cluster 1's co-mutations are the same as those of Cluster V_u, indicating that Cluster 1 may have been derived from Cluster V_u. Cluster 2 shares the same co-mutations as those of Cluster II_u. Cluster 3's co-mutations are the descendants of Cluster V_u. Clusters 4 and 5 have the same co-mutations as those of Clusters III_u and VI_u, indicating Clusters 4 and 5 are derived from Cluster III_u and VI_u. Cluster 6's co-mutations are descendants of Clusters III_u and VI_u. Note that co-mutations of Cluster I_u and the second set of co-mutations of Cluster II_u ([8782, 28144]) are not predominant co-mutations in our dataset, which may indicate a weaker infectivity. For example, every co-mutation in Table 9 has mutation 23403A > G (D614G) in the spike protein, which has been shown to increase infectivity of COVID-19 [13]. It is not surprising to see a co-mutation group not being dominant in our current dataset. By comparing these co-mutations, we can see that co-mutations that are dominant in both datasets (up to June 01, 2020 and January 20, 2021) will most likely persist in the future.

Table 11
The frequency and occurrence percentage of SARS-CoV-2 co-mutations from each clusters collected from June 01, 2020.

Cluster	Co-mutations	Frequency	Occurrence percentage
Cluster I_u	[11083, 14805, 26144]	948	0.730
Cluster II_u	[241, 3037, 14408, 23403, 25563]	2800	0.893
Cluster III_u	[241, 3037, 14408, 23403]	1468	0.412
	[8782, 28144]	1475	0.414
Cluster IV_u	[241, 1059, 3037, 14408, 23403, 25563]	1318	0.621
Cluster V_u	[241, 3037, 14408, 23403, 28881, 28882, 28883]	1872	0.817
Cluster VI_u	[241, 3037, 14408, 23403]	2222	0.969

6. Conclusion

The rapid global spread of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has led to genetic mutation stimulated by genetic evolution and adaptation. Up to January 20, 2021, 203,344 complete SARS-CoV-2 sequences, and a total of 26,844 unique SNPs have been detected. Our previous work traced the COVID-19 transmission pathways and analyzed the distribution of the subtypes of SARS-CoV-2 across the world based on 15,140 complete SARS-CoV-2 sequences. The *K*-means clustering separated the sequences into six distinguished clusters. However, considering the tremendous increase in the number of available SARS-CoV-2 sequences, an efficient and reliable dimensional reduction method is urgently required. Therefore, the objective of the present work is to explore the best suited dimension reduction algorithm based on their performance and effectiveness. Here, a linear algorithm PCA and two non-linear algorithms, t-distributed stochastic neighbor embedding (t-SNE) and uniform manifold approximation and projection (UMAP), have been discussed. To evaluate the performance of dimension reduction techniques in clustering, which is an unsupervised problem, we first cast classification problems into clustering problems with labels. Next, by setting different reduction ratios, we test the effectiveness and accuracy of PCA, t-SNE, and UMAP for *k*-NN and *K*-means using four benchmark datasets. The results show that overall, UMAP outperforms other two algorithms. The major strengths of UMAP is that UMAP-assisted *k*-NN classification and UMAP-assisted *K*-means clustering at various dimension reduction ratios have a consistent performance in terms of accuracy, which proves that UMAP is a stable and reliable dimension reduction algorithm. Moreover, compared to the *K*-means clustering accuracy that does not involve any dimensional reduction, UMAP-assisted *K*-means clustering can improve the accuracy for most cases. Furthermore, when the dimension is reduced to two, the UMAP clustering visualization is clear and elegant. Additionally, UMAP is a relatively efficient algorithm compared to t-SNE. Although PCA is a faster algorithm, its major limitation is its poor performance in accuracy. To be noted, UMAP performs better than PCA and t-SNE for the dataset with a large number of samples, indicating it is the best suited dimensional reduction algorithm for our SARS-CoV-2 mutation dataset. Moreover, we apply the UMAP-assisted *K*-means clustering to the world SARS-CoV-2 mutation dataset (up to January 20, 2021), which displays six distinct clusters. Correspondingly, the same approaches are also applied to the United States SARS-CoV-2 mutation dataset (up to January 20, 2021), resulting in six different clusters as well. Furthermore, we provide a new perspective by utilizing UMAP-assisted *K*-means clustering to analyze our previous SARS-CoV-2 mutation datasets, and the 2D visualization of UMAP-assisted *K*-means clustering of our previous world SARS-CoV-2 mutation dataset (up to June 01, 2020) forms more clear clusters than the PCA-assisted *K*-means clustering. Finally,

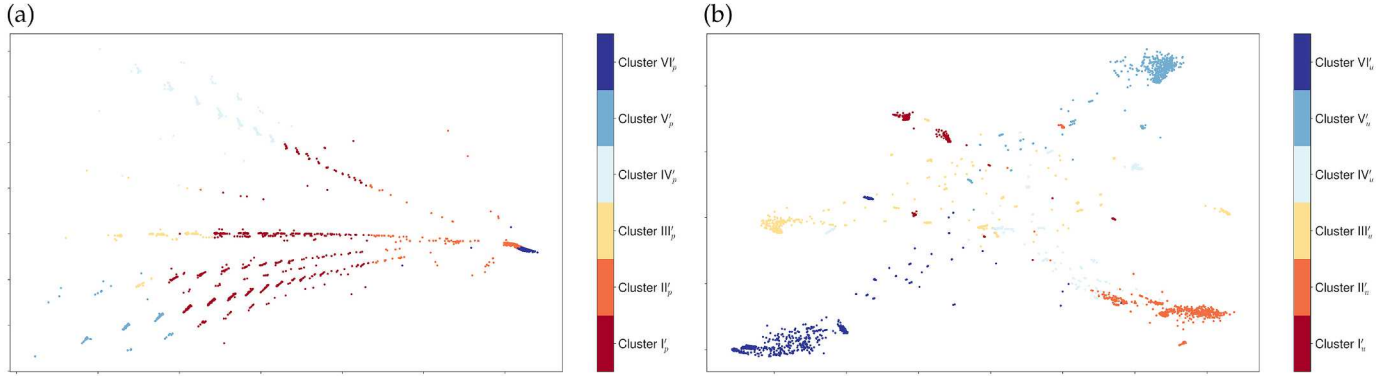


Fig. 10. 2D visualizations of the US SARS-CoV-2 mutation dataset up to June 01, 2020 with 6 distinct clusters by applying two different dimensional reduction algorithms. (a) 2D PCA visualization. (b) 2D UMAP visualization.

one of our four datasets was generated by the Jaccard distance representation, which improves both *k*-NN classification and *k*-means clustering accuracies on the original dataset.

Supporting Information

The Supporting Information is available for:

1. S1: *K*-means clustering
2. S2: Cluster distribution
3. S3: Supporting Tables

Declaration of competing interest

None Declared.

Acknowledgment

This work was supported in part by NIH grant GM126189, NSF Grants DMS-2052983, DMS-1761320, and IIS1900473, Michigan Economic Development Corporation, Bristol-Myers Squibb, and Pfizer. The authors thank The IBM TJ Watson Research Center, The COVID-19 High Performance Computing Consortium, and NVIDIA for computational assistance.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.104264.

References

[1] COVID19 Weekly Epidemiological Update, 19 January 2021, 2021.
[2] I. Alam, A.A. Kamau, M. Kulmanov, Ł. Jaremko, S.T. Arold, A. Pain, T. Gojobori, C. M. Duarte, Functional pangenome analysis shows key features of e protein are preserved in SARS and SARS-CoV-2, Front Cell Infect Microbiol 10 (2020) 405.
[3] Y. Bai, D. Jiang, J.R. Lon, X. Chen, M. Hu, S. Lin, Z. Chen, X. Wang, Y. Meng, H. Du, Comprehensive evolution and molecular characteristics of a large number of SARS-CoV-2 genomes reveal its epidemic trends, Int. J. Infect. Dis. (2020) 164–173, 100.
[4] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I.W. Kwok, L.G. Ng, F. Ginhoux, E. W. Newell, Dimensionality reduction for visualizing single-cell data using umap, Nat. Biotechnol. 37 (1) (2019) 38–44.
[5] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, Adv. Neural Inf. Process. Syst. 14 (2001) 585–591.
[6] J. Chen, R. Wang, M. Wang, G.-W. Wei, Mutations strengthened SARS-CoV-2 infectivity, J. Mol. Biol. 432 (2020) 5212–5226.
[7] J. Chen, R. Zhao, Y. Tong, G.-W. Wei, Evolutionary de rham-hodge method, arXiv preprint arXiv:1912.12388, 2019.
[8] M.A. Cox, T.F. Cox, Multidimensional scaling, in: Handbook of Data Visualization, Springer, 2008, pp. 315–347.
[9] P. Forster, L. Forster, C. Renfrew, M. Forster, Phylogenetic network analysis of sars-cov-2 genomes, in: Proceedings of the National Academy of Sciences, vol. 117, 2020, pp. 9241–9243, 17.
[10] Y.-N. Gong, K.-C. Tsao, M.-J. Hsiao, C.-G. Huang, P.-N. Huang, P.-W. Huang, K.-M. Lee, Y.-C. Liu, S.-L. Yang, R.-L. Kuo, et al., SARS-CoV-2 genomic surveillance in Taiwan revealed novel ORF8-deletion mutant and clade possibly associated with infections in Middle East, Emerg. Microb. Infect. 9 (1) (2020) 1457–1466.
[11] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Phil. Trans. Math. Phys. Eng. Sci. 374 (2016), 20150202, 2065.
[12] S.M. Kasibhatla, M. Kinikar, S. Limaye, M.M. Kale, U. Kulkarni-Kale, Understanding evolution of SARS-CoV-2: A perspective from analysis of genetic diversity of RdRp gene, J. Med. Virol. 92 (10) (2020).
[13] B. Korber, W.M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E.E. Giorgi, T. Bhattacharya, B. Foley, et al., Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus, Cell 182 (4) (2020) 812–827.
[14] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, in: Proceedings of the IEEE, vol. 86, 1998, pp. 2278–2324, 11.
[15] M. Levandowsky, D. Winter, Distance between sets, Nature 234 (5323) (1971) 34–35.
[16] X. Li, J. Zai, Q. Zhao, Q. Nie, Y. Li, B.T. Foley, A. Chaillon, Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2, J. Med. Virol. 92 (6) (2020) 602–611.
[17] G.C. Linderman, M. Rachh, J.G. Hoskins, S. Steinerberger, Y. Kluger, Fast interpolation-based t-sne for improved visualization of single-cell rna-seq data, Nat. Methods 16 (3) (2019) 243–245.
[18] L.v. d. Maaten, G. Hinton, Visualizing data using t-sne, J. Mach. Learn. Res. 9 (Nov) (2008) 2579–2605.
[19] L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, arXiv preprint arXiv:1802.03426, 2018.
[20] S.A. Nene, S.K. Nayar, H. Murase, Columbia Object Image Library (Coil-20), 1996.
[21] C.S.G. of the International, et al., The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2, Nature Microbiol 5 (4) (2020) 536.
[22] R.D. Page, Space, time, form: viewing the tree of life, Trends Ecol. Evol. 27 (2) (2012) 113–120.
[23] B. Rozemberczki, C. Allen, R. Sarkar, Multi-scale Attributed Node Embedding, 2019.
[24] J.W. Sammon, A nonlinear mapping for data structure analysis, IEEE Trans. Comput. 100 (5) (1969) 401–409.
[25] F. Sievers, A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, et al., Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Mol. Syst. Biol. 7 (1) (2011) 539.
[26] D.I. Spivak, Metric realization of fuzzy simplicial sets, Self Published Notes (2012) available online at: https://www.semanticscholar.org/paper/METRIC-REALIZATION-OF-FUZZY-SIMPLICIAL-SETS-Spivak/a73fb9d562a3850611d2615ac22c3a8687fa745e.
[27] J. Tang, J. Liu, M. Zhang, Q. Mei, Visualizing large-scale and high-dimensional data, in: Proceedings of the 25th International Conference on World Wide Web, 2016, pp. 287–297.
[28] Y. Toyoshima, K. Nemoto, S. Matsumoto, Y. Nakamura, K. Kiyotani, SARS-CoV-2 genomic variations associated with mortality rate of COVID-19, J. Hum. Genet. 1–8 (2020).
[29] D. Ulyanov, Multicore-tsne. https://github.com/DmitryUlyanov/Multicore-TSNE, 2016.
[30] L. van Dorp, M. Acman, D. Richard, L.P. Shaw, C.E. Ford, L. Ormond, C.J. Owen, J. Pang, C.C. Tan, F.A. Boshier, et al., Emergence of genomic diversity and recurrent mutations in SARS-CoV-2, Infect. Genet. Evol. 83 (2020), 104351.
[31] R. Wang, J. Chen, K. Gao, Y. Hozumi, C. Yin, G.-W. Wei, Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants, Commun. Biol. 4 (2021).
[32] R. Wang, J. Chen, Y. Hozumi, C. Yin, G.-W. Wei, Decoding asymptomatic COVID-19 infection and transmission, J. Phys. Chem. Lett. 11 (23) (2020) 10007–10015.
[33] R. Wang, Y. Hozumi, C. Yin, G.-W. Wei, Decoding SARS-CoV-2 transmission, evolution and ramification on COVID-19 diagnosis, vaccine, and medicine, J. Chem. Inf. Model. 60 (2020) 5853–5865.
[34] G. Wei, Wavelets generated by using discrete singular convolution kernels, J. Phys. Math. Gen. 33 (47) (2000) 8577.
[35] M. Worobey, J. Pekar, B.B. Larsen, M.I. Nelson, V. Hill, J.B. Joy, A. Rambaut, M. A. Suchard, J.O. Wertheim, P. Lemey, The emergence of SARS-CoV-2 in europe and north America, Science 370 (6516) (2020) 564–570.
[36] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, et al., A new coronavirus associated with human respiratory disease in China, Nature 579 (7798) (2020) 265–269.
[37] T. Zhou, K.C. Chan, Y. Pan, Z. Wang, An approach for determining evolutionary distance in network-based phylogenetic analysis, in: International Symposium on Bioinformatics Research and Applications, Springer, 2008, pp. 38–49.