Cluster Identification in Metagenomics – A Novel Technique of Dimensionality Reduction through Autoencoders

Kalana Wijegunarathna[#1*],Uditha Maduranga[2*], Sadeep Weerasinghe[3*], Indika Perera[4*], and Anuradha Wickaramarachchi[5 †]

***Abstract*— Analysis of metagenomic data is not only challenging because they are acquired from a sample in their natural habitats but also because of the high volume and high dimensionality. The fact that no prior lab based cultivation is carried out in metagenomics makes the inference on the presence of numerous microorganisms all the more challenging, accentuating the need for an informative visualization of this data. In a successful visualization, the congruent reads of the sequences should appear in clusters depending on the diversity and taxonomy of the microorganisms in the sequenced sample. The metagenomic data represented by their oligonucleotide frequency vectors is inherently high dimensional and therefore impossible to visualize as is. This raises the need for a dimensionality reduction technique to convert these higher dimensional sequence data into lower dimensional data for visualization purposes. In this process, preservation of the genomic characteristics must be given highest priority. Currently, for dimensionality reduction purposes in metagenomics, Principal Component Analysis (PCA) which is a linear technique and t-distributed Stochastic Neighbor Embedding (t-SNE), a non-linear technique, are widely used. Albeit their wide use, these techniques are not exceptionally suited to the domain of metagenomics with certain shortcomings and weaknesses. Our research explores the possibility of using autoencoders, a deep learning technique, that has the potential to overcome the prevailing impediments of the existing dimensionality reduction techniques eventually leading to richer visualizations.**

***Keywords*— Metagenomic data visualizations, nonlinear dimensionality reduction, autoencoders, clustering**

I. Introduction

The field of metagenomics has shown popular interest among bioinformatics and computer science researchers in the recent years. It has opened up new pathways in many areas including population-level genomic diversity of the microbial organisms. Metagenomics [1] was coined, with the idea of performing analysis on similar in certain criteria, yet non identical, microorganisms which are extracted from diverse environmental samples or from the natural habitats in order to study the structure and functions of microorganisms. An earliest-known method for studying metagenomic DNA is the abundance of guanine-cytosine (%GC) content.

Correspondence: K. Wijethunga[#1](E-mail: kalanainduwara.16@cse.mrt.ac.lk)
Received: 20-12-2020 Revised: 25-02-2021 Accepted: 14-03-2021

This paper is an extended version of the paper "Dimensionality Reduction for Cluster Identification in Metagenomics using Auto encoders" presented at the ICTer 2020 conference.

K. Wijegunarathna#1*,U. Maduranga2*, S. Weerasinghe3*, I. Perera4* are from the Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka. {kalanainduwara.16, maduranga.16, sadeep.16, indika}@cse.mrt.ac.lk

Anuradha Wickramarachchi[5†] is from the Australian National University, Australia. (anuradha.wickramarachchi@anu.edu.au)

DOI: http://doi.org/10.4038/icter.v14i2.7224

%GC varying widely between species but remaining relatively constant within the species is proven. Acquiring oligonucleotide frequencies of the microbial organisms is a widely used method that identifies the nucleotide composition with much better accuracy and effectiveness, compared to %GC [2]. Contemporary studies have shown that the oligonucleotide frequencies as they appear in genomic sequences is unique for a given microorganism. Research on this which runs back to 1960s, showcase the fact that oligonucleotide frequencies having species-specific signatures [3]. Because of this, an array of all oligonucleotide frequencies for a given length provides genomic signatures for microorganisms.

Oligonucleotide frequencies can be represented as vectors in high dimensional Euclidean space. Visualization of metagenomic data, without prior taxonomic references using sequence fragments can use frequency vectors to be used as genomic signatures. An ideal visualization must be capable of capturing the authentic characteristics of the microorganisms in the sample, given a set of metagenomic sequence fragments, and display the alike species separated from the rest. Consequently, the visuals must be capable to display the taxonomic structure which is inherited by the original sequence data. Being self-explanatory and ability to carry out further analysis are few of the other characteristics that are expected of visualizations.

Visualization of metagenomic data is broadly twofold. The visualization of a single metagenome and the visualization of multiple metagenomes. While visualization of multiple metagenomes gives insight into the nature of the same or different types of species found in the two different environments and helps researchers gain insightful information on the environments, the study of single metagenomes focuses on the species richness and diversity in the particular environmental sample. Despite the high dimensionality and other challenges, metagenomic data can be visualized in various techniques as described by Sudarikov et al [4].

Conversely, taxonomic classification of metagenomic data can be broadly categorized into four categories. Sequence similarity based classification employs a database search on a database of reference sequences. This method has been successfully used in the identification of reads of length as short as 80 base pairs where most other methods have failed [5]. This method is usually slower and uses Basic Local Alignment Search Tool (BLAST) [6] to identify similarities. Classification based on sequence composition is another method of taxonomic classification. One way of doing this is by using nucleotide composition of the reads. The nucleotide composition of the reads are compared with the models built using the composition of reference genomes and the model that fits the composition of the reads best are chosen. The

main drawback of this method is that it fails to classify reads shorter than 1000 base pairs with reasonable accuracy [7]. The third classification method is a hybrid of the two previous methods that combines both approaches of read similarity and nucleotide composition. It is possible to increase the accuracy by taking an aggregation of the two matches for a better result. In this case the score from the read-sequence similarity and the score from the read composition and reference genome model are combined. Alternatively, it is possible to narrow down the database using composition matching as the initial step to apply the similarity search on the filtered, reduced database [8]. The last of the four approaches is the use of marker genes. The reads are classified according to the markers they hit. Although this method is faster and efficient, it induces a bias towards organisms with larger genomes since they generate a larger number of reads [7]. Due to variability in 16S rRNA (ribosomal Ribonucleic Acid) [9] copy number, Amplicon sequencing [10] also suffers with bias.

A key challenge in bioinformatics as well as metagenomics to date is the visualization of metagenomic fragment data without prior taxonomic identification. Usually for the visualization needs, the higher dimensional frequency vectors need to be embedded into lower dimensions (2D or 3D). Preservation of the inherited data from the genomic sequences is a deciding factor of the quality and accuracy of the visualizations when reducing into lower dimensions. Principal Component Analysis (PCA) [11] and t-distributed Stochastic Neighbor Embedding (t-SNE) [12] are two of the most widely used techniques for this purpose. But limitations in these techniques which hinder faithful visualizations prevail.

Massive volumes of genomic data can be produced in such an efficient manner using the advancements in the field of sequencing and with the introduction of latest sequencing technologies like Next-Generation Sequencing (NGS). It is evident that with the availability of large volumes of genomic data, optimizations and new analyzing techniques are becoming more crucial. Cutting-edge technologies that are being used throughout in computer science can be adopted in the field of bioinformatics as well. Deep learning techniques are built to handle the rapid rate of generation of data and for the intuitive and rapid analysis.

This research aimed at producing better visualizations in metagenomics for cluster identification by adopting an autoencoder based approach of dimensionality reduction. Autoencoders can be used in this context to convert higher dimensional metagenomic data into visualizable lower-dimensional data preserving the important characteristics of the original sequences. Since it is a deep learning technique it will be useful in a context like genomics. This allows better analysis on data-inherent taxonomic structure, free from alignments.

The rest of the paper is organized as follows. The related work on dimensionality reduction techniques and research that are carried out on using autoencoders in the genomics domain are discussed under Section 2. A broader introduction to autoencoders is given in Section 3. Dataset, methodology for dimensionality reduction and visualizations are discussed in Section 4 followed by the experimental results which are described and demonstrated in Section 5 in detail. Finally, Section 6 concludes the paper giving an overview of the future work.

II. Related Work

A. Dimensionality Reduction Techniques

Representing data visually can be considered as a major challenge in the field of genomics. Processes such as transforming, scaling, normalizing, color-encoding and clustering play major roles in visualizing genomic data. It is important not to hinder users' ability to interpret data while facilitating the users to carry out their analysis more conveniently and precisely. According to the studies done by Rall et al. [13], still number of challenges prevail in visualizing the genomic data. One of the leading challenges will be dealing with the dimensionality of the genomic sequence data. A plethora of research were carried out on lower-dimensional embedding techniques.

One of the dominant concerns is dimensionality reduction, in bioinformatics fields, which looks into analyzing sequence data. Genomic sequence data consists of extensive amounts of sequence data and features. Thus, it is essential to reduce the dimensionality of data to extract useful analysis and visualization by avoiding the curse of dimensionality. Transformation of high dimensional data to a lower number of dimensions is a major goal in dimensionality reduction, providing simple interpretations. In the ideal case, dimensionality of reduced representation must have the dimensionality that corresponds to the intrinsic data [14]. One of the related concerns is the dense preservation of information.

PCA, without a doubt, is the most commonly used linear technique in dimensionality reduction across multiple domains. It converts data in the higher dimension spaces to the lower-dimensional subspace making sure maximized variance in the projected data. Making sure that the maximum variance in the projected data also means PCA minimizes the squared reconstruction error. One of the leading drawbacks of PCA is its restriction with respect to linear transformations. Hauskrecht et al. [15] in in their work also displays the restrictions of PCA to guarantee high-quality features for discriminatory purposes because it is a totally unsupervised technique.

Stochastic Neighbor Embedding (SNE) [16] introduced by Hinton and Roweis is a non-linear technique to get lower dimensional representations. It uses a Gaussian distribution on each point of data in the higher dimension and defines a probability distribution for its neighbors. This unsupervised dimensionality reduction technique has been commonly used over many years. t-SNE [12] is a variant of the SNE for non-linear dimensionality reduction. This method is also used to produce lower dimensional representations of higher dimensions that can be visualized with ease, especially in scatterplots. t-SNE preserves the global structures of the data like clusters while capturing the local structure of the higher dimensions. t-SNE adopts a Gaussian kernel in order to identify the similarities between points in the higher dimensions. Lower dimensional points are plotted while providing similar probabilities to the points and they are usually configured in such a way that they will reduce the divergence between the two distributions. Importantly, t-SNE strongly advocates against using Gaussians to measure distances in lower dimensions. It will instead opt for the one-dimensional t distribution (i.e. the Cauchy Distribution). Thus, it has heavier tails and allows for more spread in the lower dimensional representation than Gaussian. Nevertheless, a

significant limitation of SNE as well as t-SNE is that their computational and memory complexity scale quadratically in the number of data objects N. Thus, SNE and its variants can only be used with limited number of points [17].

The standard implementation of t-SNE has a time complexity of $O(N^2)$, where N is the number of genomic fragments. Barnes-Hut-SNE (BH-SNE) introduced by Laurens van der Maaten [17] has a time complexity of O(N log N). Therefore, BH-SNE is significantly faster compared to t-SNE. Despite the efficiency, the results produced by BH-SNE are similar to that of t-SNE. Sparse similarities between each pair of points were obtained using vantage-point trees while the forces of the corresponding embedding were acquired using an enhanced version of Barns-Hut algorithm [18]. Due to the lower time complexity, BH-SNE can process more than a million data points within a reasonable time.

A 2009 paper [19] by Dick et al describes the use of Self-Organizing Maps [13] (SOMs) for reducing the dimensionality of tetranucleotide genomic signatures belonging to two acidophilic biofilm communities. The unsupervised learning technique SOM uses an artificial neural network to generate a two-dimensional representation of the high dimensional data. This technique was used by the researchers to bin tetranucleotide sequence fragments obtained from isolated genomes and metagenomic samples. Despite being neural networks based, SOM does not rely on error-correction learning. Instead, SOM depends on competitive learning to map the inputs to an output while preserving characteristics in the input space.

Gisbrecht et al. [21], in their research on "Nonlinear dimensionality reduction for cluster identification in metagenomic samples" has compared several currently used techniques for dimensionality reduction. The researchers have obtained oligonucleotide frequency vectors from a set of sequences generated by simulating metagenomic next-generation sequencing. These vectors have then been fed into dimensionality reduction algorithms. Researchers have used the effectiveness of these algorithms at clustering the output to compare the dimensionality reduction techniques. The techniques being compared are PCA, GTM (Generative Topographic Mapping) [22] and t-SNE. The research demonstrates that t-SNE outperforms the other techniques in terms of accuracy. The researchers have also introduced improvements to t-SNE to overcome its high complexity.

Datasets in bioinformatics are typically large. The quadratic time complexity of t-SNE does not scale well for the needs of the genomic datasets. This is a shortcoming that needs to be addressed. The use of autoencoders for this purpose has been explored by Wang and Gu in their paper on dimension reduction and visualization of single-cell RNA seq data [23].

B. Autoencoders in Genomics

As deep learning became mainstream, we have observed a rise of neural network based techniques that rival the traditional mathematical & probabilistic methods in their corresponding areas. In the case of dimensionality reduction, autoencoders can be named as the candidate, deep learning has to offer. The viability of autoencoders to replace t-SNE, and PCA in the context of bioinformatics has to be explored. Autoencoders, however, have already played a role in the bioinformatics domain in several instances.

A 2019 paper by Ersalan et al. [24] describes how they managed to denoise single-cell RNA sequencing (scRNA-seq) datasets using deep count autoencoders (DCA). They used an enhanced version of autoencoders although a number of other methods existed to perform this very task. The enhancement was in terms of the specialized loss functions which drive in the direction of denoised scRNA-seq data. In addition to that, DCA scales linearly with regard to the number of cells overcoming the limitation of limiting the datasets. Experiment results suggest that DCA has surpassed the existing methods for imputation by means of quality and speed.

Wang et al. [23] in the research on dimension reduction of single-cell RNA seq data, propose the use of variational autoencoders. Research experimented the use of variational autoencoders for single-cell RNA seq data because the chances of dropouts are higher when dealing with single-cell levels with higher transcriptional fluctuations. This experiment went on to show how to overcome the limitations of PCA and ZIFA (Zero Inflated Factor Analysis) [25] by testing on over 20 different datasets. Variational autoencoders gain a special edge over both PCA and ZIFA because of its ability to deal with complex non-linear relationships.

Wang and Wang [26] in their research have used variational autoencoders to study two subtypes of lung cancers, lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). Although the researchers were expected to capture underlying DNA methylation patterns of the different original subtypes separately, some LUSC samples were classified into a LUAD group, which may be an indication that some parts of LUSC tumor samples may have similar DNA methylation expression compared to LUAD tumor. It was evident that a biologically meaningful latent space can be extracted using variational autoencoders from the data consisting of two or more subtypes of the lung cancers.

III. AUTOENCODERS

Although autoencoders took the center stage in the late 2000s with the introduction of the deep architecture, the original concept goes back as far as 1980s when Rumelhart et al. [27] introduced a new learning procedure that recurrently adjusts the weights of the connections of the network in such a way that the disparity between the actual output and the expected output of the network is minimal. Autoencoders can be considered as a special type of neural networks, which consists of two symmetric components namely encoder and decoder. Encoder maps input to more compressed lower dimensions, in contrast to the decoder which does exactly the opposite. Autoencoders are an unsupervised learning technique that leverages reducing the dimensionality of the input vectors efficiently, such that it will preserve the important characteristics of the original data and then reconstructing the original data with minimum loss using compressed representation. L(x, x'), which represents the deviation between the original input (x) and the consequent reconstruction (x') should be diminished to get better reconstructions.

A general autoencoder (Fig. 1) can be expressed using the tuple X, Y, Φ, Ψ, X' where Φ denotes the encoder function and Ψ represents the decoder function of the autoencoder. X

and X' are the input and the output vectors of the autoencoder respectively.

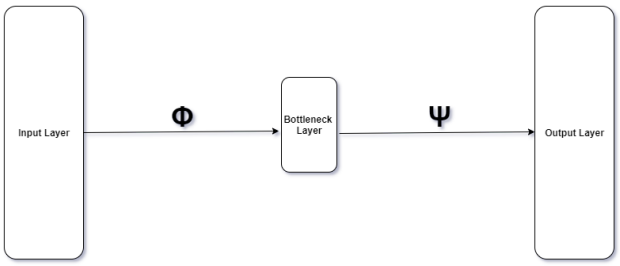

Fig. 1 Structure of a general autoencoder

$$\Phi : X \rightarrow Y \quad (1)$$

$$\Psi : Y \rightarrow X \quad (2)$$

$$\Phi, \Psi = . \arg_{\Phi,\Psi} \min (\text{loss}(x,x')) \quad (3)$$

Φ denotes the function that maps the original data X, to a latent space Y, which appears at the bottleneck and Ψ function does the opposite. Deep autoencoders can have multiple hidden layers.

Basically, when the layers of the neural network increased, those autoencoders are called deep autoencoders. Having advanced learnability will be essential to reconstruct the input data as it is if possible. A well trained autoencoder is capable of reconstructing the input with minimum loss. Although the reconstruction is done by the decoder, saving important characteristics of the input data plays a major role up to the bottle-neck layer of the autoencoder.

Autoencoders can be considered as the non-linear generalization of PCA that converts higher dimensional data into lower dimensional code using the encoder part [28]. Autoencoders are being used as a powerful tool for dimensionality reduction [29]. It is proven to be useful in fields which have extensive amounts of data to work with.

IV. Methodology

A. Dataset

For the experiment, we chose a subset of the genomic sequences used by Gisbrecht et al [21] as our dataset. The sequences were obtained from the NCBI (National Center for Biotechnology Information) microbial genomes database (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/). We first went on to reproduce the PCA, and t-SNE based results to work as the baseline for the research. To conduct the research the genomic sequences were downloaded in the FASTA file format. Complete genomes of 8 microbial species were used. Our chosen subset contained species that are taxonomically very close to each other, as well as species that are very different from each other. The intention was to explore how well the dimensional reduction process preserves the taxonomic similarities and dissimilarities.

B. Implementation

The DNA sequences that we obtained, must go through a multi-stage process before they are ready to be visualized. This does not mean that significant alterations are done to the underlying genomic features of the sequences. Two- or three-dimensional scatter plots are drawn based on the input sequences at the end of the following process.

- Processing raw microbial DNA sequences and extracting genomic metrics
- Converting the higher dimensional genomic metrics into a lower dimensional state
- Visualizing the lower dimensional state with scatter plots

First, we read contiguous blocks from each input DNA sequence. The blocks were taken from random locations within the sequence. The lengths of the sequences were chosen so that they form a normal distribution where the mean length is 5000 base pairs, and the standard deviation is 1000 base pairs. For a given k value, we calculated the number of the occurrences of each k-mer within each block. A k-mer and its reverse complement pairs were considered as a single k-mer; therefore, we had to take the sum of the occurrences of each k-mer in the pair and use that to represent that k-mer and its reverse complement. This method was proposed by Abe et al. [30]. For each block, we created a vector by taking all such k-mer frequencies. The resulting vector was then normalized.

After following the above process, what we end up with is a set of vectors, each having an equal dimensionality. The dimensionality depends on k. The high dimensionality of these vectors makes them impossible to visualize as they are. Transforming these high dimensional vectors to a lower dimensionality is done by the autoencoder. The output of the autoencoder is a set of lower dimensional vectors, typically the dimensionality of these vectors is less than 4. These vectors are then directly used for visualization by drawing scatter plots. Other techniques can also be used in the place of the autoencoder for the process of dimensionality reduction.

The autoencoders perform dimensionality reduction in 2 stages. In the first stage, we feed the high dimensional vectors to train the autoencoder. In this stage the input layer of the neural network is provided with the vectors, and the same vectors are expected from the output layer of the neural network. In other words, we train the neural network to produce the same vector as the input. Each of the vectors fed into the autoencoder corresponds to each block we read originally. Once the training is over, we send the same set of vectors as input and obtain the values produced in the bottleneck layer. The resulting values correspond to a lower dimensional representation of the input vector.

We experimented using various neuron counts for each layer and analyzed the results. As we identified the following configurations worked the best for respectively 3-mers and 4-mers,

1. {32, 16, 2, 16, 32}
2. {136, 64, 2, 64, 136}

The results we obtained using these autoencoders are presented in the paper. We used sigmoid as the activation function for neurons, while 'Adam' and mean square error were used as the optimizer and the loss function respectively. 1000 reads (blocks) were taken from each of the original genomes, each on average having a length of 5000 bp.

For comparison purposes, the results of PCA and t-SNE obtained using the same data fed to the autoencoder were visualized. The sequence lengths and standard deviation were kept constant across all three approaches. T-SNE and PCA

were both imported from the 'scikit-learn' [31] library. Seaborn data visualization library was used to visualize the data in its reduced 2D space. The availability of ground truth labels ensured comparability of the performances of the three techniques of dimensionality reduction using the clusters produced by the respective approaches.

Due to the use of a manually curated simulated metagenomic dataset, the ground truth label, i.e., the true identity of the microorganism to which the oligonucleotide frequency belonged to, is known. But this is rarely the case in the field of metagenomics, where a single metagenomic sample obtained from the environment may consist of large amounts of DNA fragments from a wide variety of species who may or may not be related to each other. Hence making the evaluation of latent space visualizations of real metagenomic datasets extremely challenging. Though this can be broadly named a clustering problem, several acute complications remain. While the creation of an isolated, lonely clusters consisting of data points from the right species is useful in the context of identifying the number of organisms in the sample, this sort of isolation can be costly in terms of taxonomic information. Not only is the creation of a cluster consisting of the same species important but the whole visualization must reflect the connections and similarities these different clusters bear with each other. Hence, mere arbitrary separation is not recommended.

In our experiment, we used DBSCAN (Density-based spatial clustering of applications with noise) DBSCAN [32] for clustering. Though ideal method for clustering low dimensional metagenomic data is still a point of debate in the research community, DBSCAN is a powerful algorithm that discovers clusters of arbitrary shapes and can efficiently deal with large datasets that are part and parcel of genomic problems. DBSCAN relies on a density-based notion of clusters and requires only one essential input parameter. The low dimensional (2D) coordinate pairs, obtained from the bottleneck layer of the autoencoder, of the oligonucleotide frequencies are then fed to the DBSCAN for clustering. These 2D coordinate pairs are then clustered by DBSCAN without knowledge of their ground truth labels into clusters solely based on their distances in the 2D coordinate plane. It is these clusters that are produced by DBSCAN that need evaluation. Clustering evaluation can be broadly categorized into two; intrinsic metrics and extrinsic metrics [33]. Extrinsic metrics, when calculating the quality of a cluster, considers the ground truth label of the data points in each of the clusters. Therefore, the use of an extrinsic metric demands ground truth labels, which we fortunately possess. Contrary to extrinsic metrics, intrinsic metrics only consider the intra-cluster closeness and inter-cluster distance, not taking into consideration the ground truth labels of the data. Intrinsic clustering evaluations evaluate the integrity of the clusters formed in the low dimension that meets the user's eyes.

The primitive logic that is used in the density-based clustering approach is derived from a human-intuitive clustering method. In DBSCAN the resultant clusters are the dense regions in the given dataset, separated by regions of the lower density of points. The ability to identify arbitrary shaped clusters and robustness towards the outliers make DBSCAN more effective with the datasets that have relatively similar densities in the clusters. This algorithm is based on connecting data points within certain distance thresholds. For any given set of data points, DBSCAN separates data points into three categories.

- Hub points - Points that are at the interior of a cluster (Centre).
- Edge points - Points that fall within the neighborhood of a hub point that is not a hub point.
- Noise points - Any point that is not a hub point or an edge point.

The most important factor to consider when using DBSCAN is getting appropriate values for the parameters of the algorithm. Although there are a few parameters that can be tuned, the eps parameter, and the min_points (min_samples) are crucial. Among those two, min_points is used to identify a dense region in the dataset by considering the number of neighboring points required for a point to be considered as a dense region.

The quality of clusters created by DBSCAN is governed by DBSCAN's vital input parameter, epsilon. Epsilon is the threshold for the maximum distance between two data points, above which distance, the two points will no longer be neighbors. For a fair estimation of epsilon in each of the DBSCAN clustering, the "knee point" [34] of the plot between the 10 nearest neighbors in sorted order vs. the distance was calculated to find the optimal epsilon. Scikit-learn's NearestNeighbors and "kneed's" KneeLocator was used for the calculations. To improve the visibility of the clusters, Convex Hull algorithm [35] was used to draw boundaries around them.

V. Results

For a formal evaluation two metrics were employed in addition to direct visual inspection. Distinct autoencoders were used with distinct number of layers and neurons for trinucleotide and tetranucleotide data to obtain separate visualizations. Clustering was then applied to the reduced dimensionality data obtained through the three dimensionality reduction approaches.

As was discussed earlier, a comprehensive evaluation of clustering must take into account an intrinsic measure as well as an extrinsic measure. There are important metrics which can be used to evaluate clustering. But most of these metrics alone will not make it an effective performance evaluation on clustering as they can be biased. In this experiment one intrinsic metric and one extrinsic metric will be used for the evaluations and comparisons. We chose V-measure and Silhouette coefficient to evaluate the performance of overall clustering with respect to all three techniques that will be analyzed.

Validity Measure (V-measure) [36] based on two other metrics called Homogeneity and Completeness, is an entropy based extrinsic clustering evaluation metric. It evaluates the successful clustering of data points with respect to their ground truth labels. Homogeneity and completeness are inversely proportional, and a good clustering should maintain a balance between those two metrics. Homogeneity evaluates how many data points in each cluster are with the same label. Maximum homogeneity is obtained by a clustering that has clusters that only have data points of the relevant class. This in turn results in zero entropy. Assume a clustering with N number of total data samples, C different class labels and K clusters. Assume also that the number of data points from class c in cluster k is $a_{c,k}$.

TABLE I
CLUSTERING USING TRINUCLEOTIDES

Data representation	Trinucleotides	
	V-measure	*Silhouette coefficient*
Autoencoder {32, 16, 2, 16, 32}	0.901	0.564
t-SNE	0.878	0.614
PCA	0.745	0.485

$$h = 1 - \frac{H(C,K)}{H(C)} \tag{4}$$

$$H(C,K) = -\sum_{k=1}^{K}\sum_{c=1}^{C}\frac{a_{c,k}}{N}\log\left(\frac{a_{c,k}}{\sum_{c=1}^{C}a_{c,k}}\right) \tag{5}$$

$$H(C) = -\sum_{c=1}^{C}\frac{\sum_{k=1}^{K}a_{c,k}}{C}\log\left(\frac{\sum_{k=1}^{K}a_{c,k}}{C}\right) \tag{6}$$

Completeness is a symmetrical metric to homogeneity. A clustering is complete when all the data points from the same class are clustered together in each cluster. In a clustering, completeness can be evaluated using the conditional entropy of the cluster distribution in comparison with the label given.

$$c = 1 - \frac{H(K,C)}{H(K)} \tag{7}$$

$$H(K,C) = -\sum_{c=1}^{C}\sum_{k=1}^{K}\frac{a_{c,k}}{N}\log\left(\frac{a_{c,k}}{\sum_{k=1}^{K}a_{c,k}}\right) \tag{8}$$

$$H(K) = -\sum_{k=1}^{K}\frac{\sum_{c=1}^{C}a_{c,k}}{C}\log\left(\frac{\sum_{c=1}^{C}a_{c,k}}{C}\right) \tag{9}$$

V-measure, the metric that is used can be computed using weighted harmonic mean of homogeneity and completeness.

$$V_{\beta} = 1 - \frac{(1+\beta)hc}{\beta h + c} \tag{10}$$

The second evaluation metric, silhouette coefficient [37] is a measure of distance. It gives a measure on how close each data point in one cluster is to other data points in the neighboring clusters. Silhouette coefficient ranges between -1 and 1. If the value of the coefficient is 0 that means the data point is on the inflection point of the two clusters.

$$S(i) = \frac{b(i) - a(i)}{max\{a(i),\ b(i)\}} \tag{11}$$

The tetranucleotide and trinucleotide reductions of autoencoders were separately compared and evaluated with the reductions of t-SNE and PCA. Autoencoder generated reductions colored with using the ground truth label can be seen in Fig. 2.

TABLE III
CLUSTERING USING TETRANUCLEOTIDES

Data representation	Tetranucleotides	
	V-measure	*Silhouette coefficient*
Autoencoder {136, 64, 2, 64, 136}	0.932	0.621
t-SNE	0.886	0.654
PCA	0.646	0.536

As visible, there are clear clusters and separations. The same 2D datapoints clustered using DBSCAN without taking their ground truth labels into consideration can be seen in Fig. 4. Worthy of notice is the close affinity of the two Clostridium bacteria, Clostridium phytofermentans and Clostridium beijerinckii.. As seen in Fig. 4., DBSCAN has managed to differentiate these two species to two different clusters but this is not always the case. In Fig. 3 and Fig. 5 with trinucleotide frequencies, the reduced dimensions do not distinctly classify the two Clostridium bacteria into two different clusters. But the close affinity of the two species on both occasions signals the preservation taxonomical information and relationships during dimensionality reduction when compared to t-SNE.

In Comparison, t-SNE's performance on the same 4-mer data can be found in Fig. 7 and Fig. 8. The distances between the clusters of the two Clostridium bacteria are arbitrary and conveys no information about the two species belonging to the same genus. Clostridium phytomenatas is in close affinity to all Microcytis, Mycobacterium, Erythrobacter and Rubrobacter as much as it is to Clostridium beijerinckii. Furthermore, DBSCAN seems to have identified separate clusters within the two Clostridium bacteria as well. PCA plots, Fig. 9 and Fig. 13 as seen to the naked eye and as also suggested by the evaluation metrics, are subpar. Well below the metrics for t-SNE and autoencoders. Note that a new optimized epsilon was calculated for each DBSCAN clustering, optimizing the algorithm to the data presented by each dimensionality reduction technique. Fig. 6. Gives the legend for plots colored using the ground truth labels.

The autoencoder outperforms both t-SNE and PCA in all cases with regards to the V-measure. This improvement reflects the relatively lower loss of information compared to the ground truth labels. The formation of string-like shapes in t-SNE is another reason for the relatively lesser V-Measures of t-SNE compared to autoencoders. The breakage of these strings has more than once led DBSCAN to identify the data points of the same species as few different clusters.

Silhouette coefficient, on the other hand, is highest in t-SNE. Compared to the autoencoder t-SNE's Silhouette coefficient is always slightly higher. This is due to the arbitrary yet clear separation of data points of different species. This arbitrary yet clear separation comes at a greater cost. The arbitrary distancing of the different clusters by t-SNE has led to a loss in taxonomic relationships among species. On the contrary, autoencoder's cluster distances are not arbitrary. The relative distances between clusters provide some useful insights into the taxonomic relationships in the real species the data points belong to.

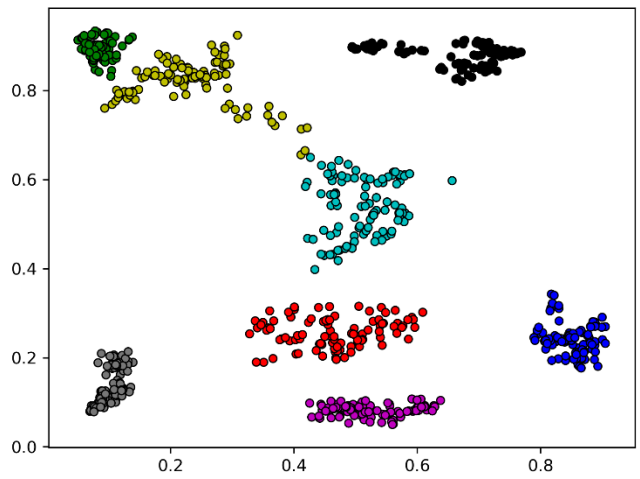

Fig. 2 Dimensionality reduction of tetranucleotide frequencies using autoencoder {136, 64, 2, 64, 136}. Coloured according to known true labels

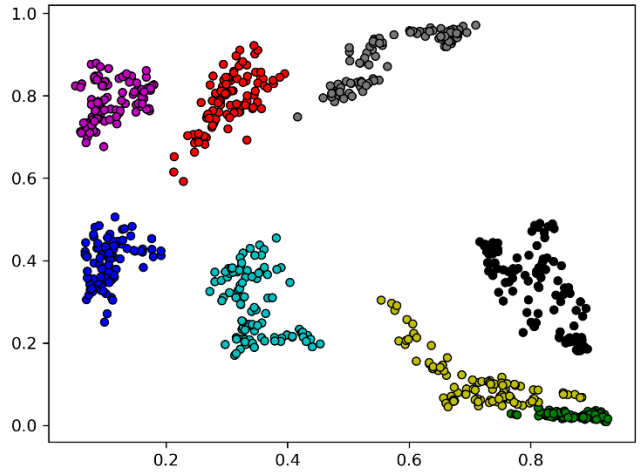

Fig. 3 Dimensionality reduction of trinucleotide frequencies using autoencoder {32, 16, 2, 16, 32}. Coloured according to known true labels

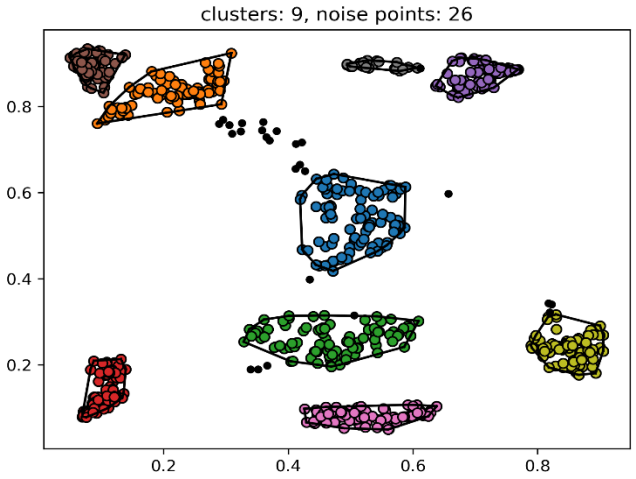

Fig. 4 DBSCAN cluster identification with convex hull for tetranucleotide frequencies in Fig. 2. Black dots show noise points.

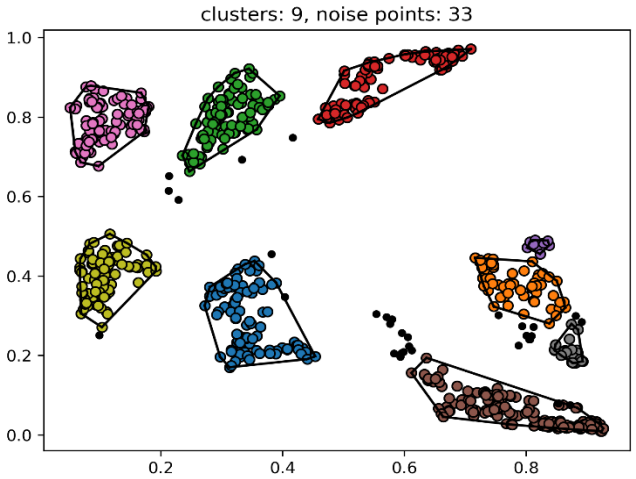

Fig. 5 DBSCAN cluster identification with convex hull for trinucleotide frequencies in Fig. 3. Black dots show noise points.

VI. Conclusion and Future Work

The results obtained from the research backed by the superior results obtained by autoencoders back the potential of using autoencoders in the field of metagenomics for dimensionality reduction and visualization of metagenomic reads. The systematically optimized DBSCAN clustering algorithm has always managed to identify a number of clusters that is quite close to the actual number of microorganisms present in the sample. This congruency of the lower dimensions with the information in the higher dimensions is reflected in the improved V-Measures. The significantly higher V-measure produced in comparison with the t-SNE and PCA dimensionality reductions prove Autoencoder's ability to preserve the intrinsic dimensionality of data in the process of dimensionality reduction. Silhouette coefficient does not consider whether the points in the same cluster actually belong to a single cluster in the higher dimensions. The autoencoder remains in close contention with t-SNE on the silhouette coefficient which is an intrinsic measure of clustering that only considers the visual integrity of the data in the lower dimensions.

Not only have autoencoders outperformed PCA and t-SNE on the metrics front but it has also managed to preserve taxonomic data by placing the species of the same genus in relatively closer affinity. Taxonomic data preservation ability of autoencoders standout in contrast to t-SNE's shortcoming of arbitrarily separating clusters giving no relevance to species' relationships. Additionally, unlike t-SNE's faltering quality with growing data volume, autoencoders, being a deep learning technique thrives with growing data volumes. These results demand a place for autoencoders in bioinformatics. Noisy reads can however plague the ultimate analysability and interpretability of visualizations. Noise can result in loss of important insights and false interpretations. The use of denoising autoencoders to denoise large volumes of sequence data is another potential avenue of research. Other forms of autoencoders like the denoising autoencoders and variational autoencoders are potential techniques that can be integrated to improve metagenomic analysis and further the use of deep learning in the wider domain of bioinformatics.

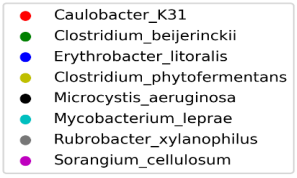

Fig. 6 Legend for Fig. 2, Fig 3, Fig 7, Fig 8, Fig 9 and Fig 13.

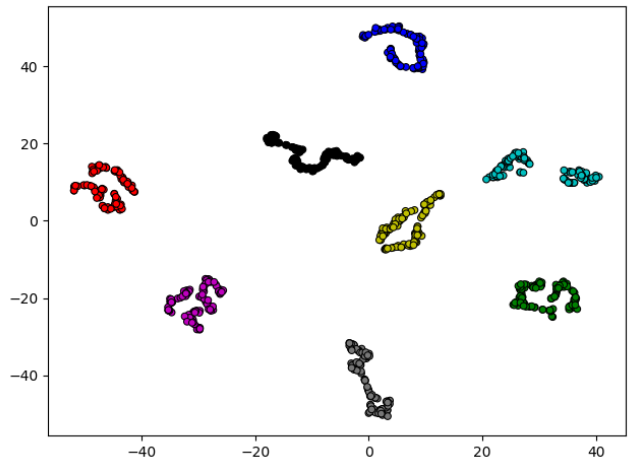

Fig. 7 Dimensionality reduction of tetranucleotide frequencies using t-SNE . Coloured according to known true labels

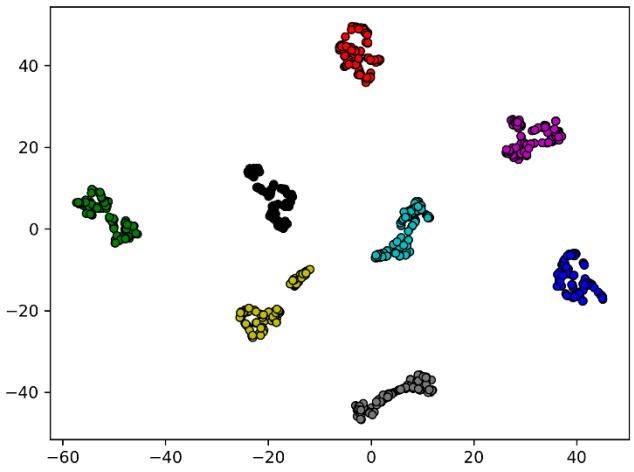

Fig. 8 Dimensionality reduction of trinucleotide frequencies using t-SNE. Coloured according to known true labels

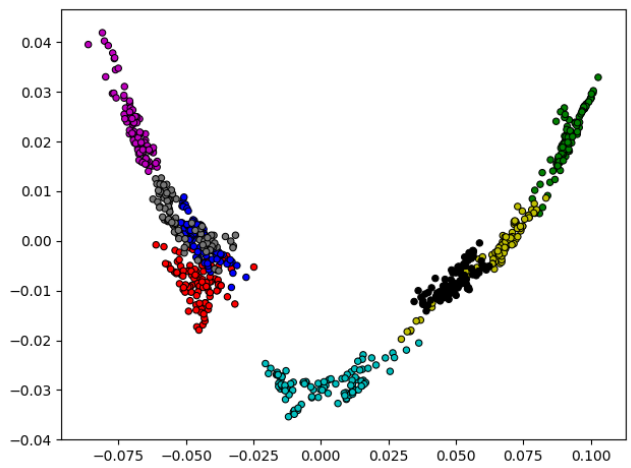

Fig. 9 Dimensionality reduction of tetranucleotide frequencies using PCA. Coloured according to known true labels

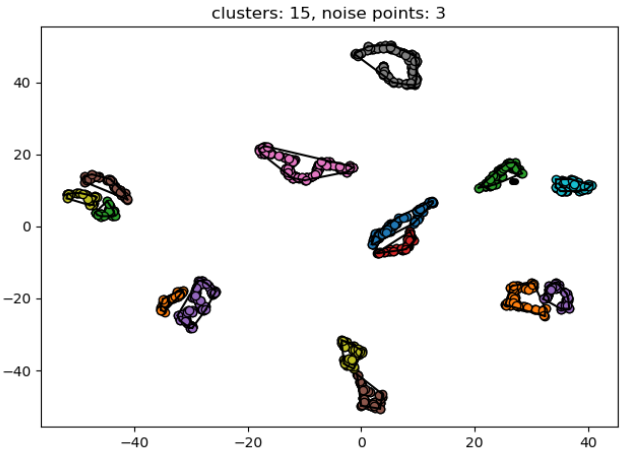

Fig. 10 DBSCAN cluster identification with convex hull for tetranucleotide frequencies in Fig. 7. Black dots show noise points.

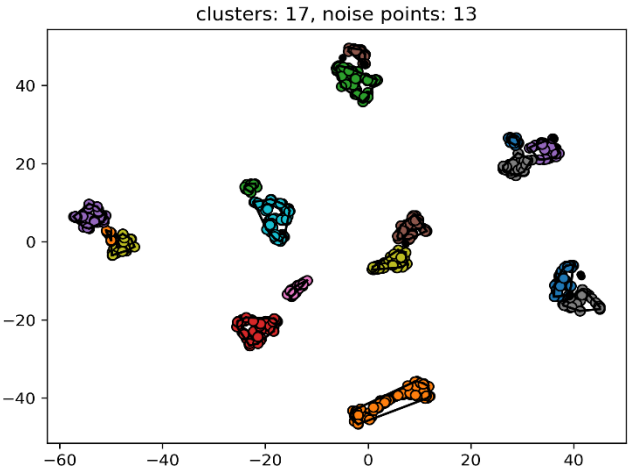

Fig. 11 DBSCAN cluster identification with convex hull for trinucleotide frequencies in Fig. 8. Black dots show noise points.

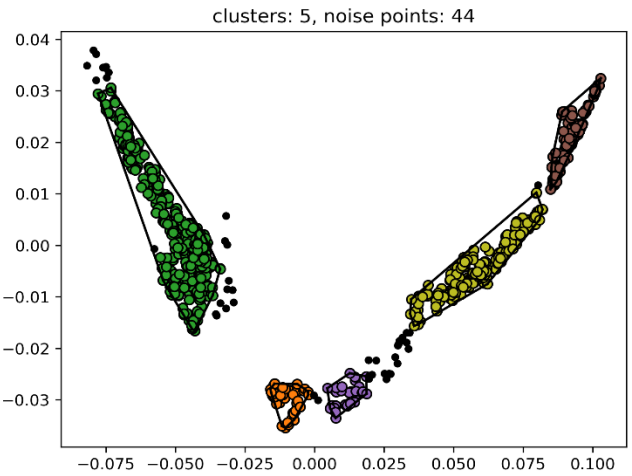

Fig. 12 DBSCAN cluster identification with convex hull for tetranucleotide frequencies in Fig. 9. Black dots show noise points.

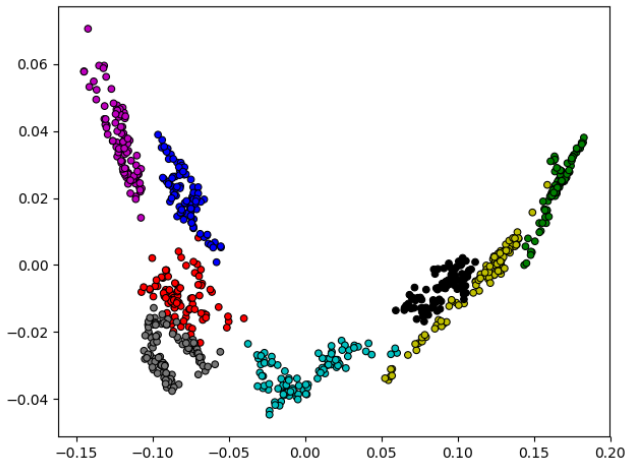

Fig. 13 Dimensionality reduction of trinucleotide frequencies using PCA. Coloured according to known true labels

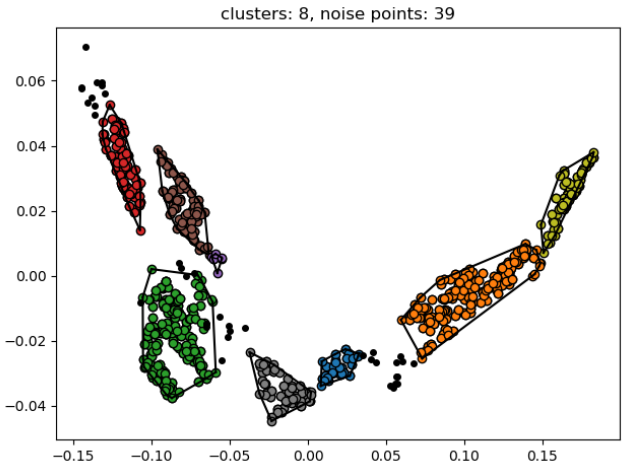

Fig. 14 DBSCAN cluster identification with convex hull for trinucleotide frequencies in Fig. 13. Black dots show noise points.

REFERENCES

[1] J. Handelsman, M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman. "Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products". *Chem. Biol.* 5:R245–R249, 1998.

[2] B. J. Baker, G. W. Tyson, R. I. Webb, J. Flanagan , P. Hugenholtz, E. E. Allen, J. F. Banfield, "Lineages of acidophilic archaea revealed by community genomic analysis". *Science* 2006, 314:1933-1935.

[3] C. C. Laczny, N. Pinel, N. Vlassis, and P. Wilmes, "Alignment-free visualization of metagenomic data by nonlinear dimension reduction," *Sci. Rep*., vol. 4, pp. 1–12, 2014.

[4] K. Sudarikov, A. Tyakht, and D. Alexeev, "Methods for the metagenomic data visualization and analysis," *Curr Issues Mol Biol*, vol. 24, pp. 37– 58, 2017.

[5] C. Ander, O. B. Schulz-Trieglaff, J. Stoye, and A. J. Cox, "metabeetl: high-throughput analysis of heterogeneous microbial populations from shotgun dna sequences," in *BMC bioinformatics*, vol. 14, p. S2, Springer, 2013.

[6] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990.

[7] M. A. Peabody, T. Van Rossum, R. Lo, and F. S. Brinkman, "Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities," *BMC bioinformatics*, vol. 16, no. 1, p. 362, 2015.

[8] M. H. Mohammed, T. S. Ghosh, N. K. Singh, and S. S. Mande, "Sphinx—an algorithm for taxonomic binning of metagenomic sequences," *Bioinformatics*, vol. 27, no. 1, pp. 22–30, 2011.

[9] J. M. Janda and S. L. Abbott, "16s rrna gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls," *Journal of clinical microbiology*, vol. 45, no. 9, pp. 2761–2764, 2007.

[10] N. A. Bokulich, S. Subramanian, J. J. Faith, D. Gevers, J. I. Gordon, R. Knight, D. A. Mills, and J. G. Caporaso, "Quality-filtering vastly improves diversity estimates from illumina amplicon sequencing," *Nature methods*, vol. 10, no. 1, pp. 57–59, 2013.

[11] S. Wold, K. I. M. Esbensen, and P. Geladi, "Principal Component Analysis," vol. 2, pp. 37–52, 1987.

[12] C. R. García-Alonso, L. M. Pérez-Naranjo, and J. C. Fernández-Caballero, "Multiobjective evolutionary algorithms to identify highly autocorrelated areas: The case of spatial distribution in financially compromised farms," *Ann. Oper. Res.,* vol. 219, no. 1, pp. 187–202, 2014.

[13] G. F. Rall, M. J. Schnell, B. M. Davis, "Tasks, Techniques, and Tools for Genomic Data Visualization," *Comput Graph Forum*., vol. 176, no. 1, pp. 139–148, 2019.

[14] L. J. P. Van Der Maaten, E. O. Postma, and H. J. Van Den Herik, "Dimensionality Reduction: A Comparative Review," *J. Mach. Learn. Res*., vol. 10, pp. 1–41, 2009.

[15] M. Hauskrecht, R. Pelikan, M. Valko, J. Lyons-Weiler, "Feature Selection and Dimensionality Reduction in Genomics and Proteomics," Fundamentals of Data Mining in Genomics and Proteomics, Springer, pp.149-172, 2006,

[16] G. Hinton and S. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems*, 2003.

[17] L. van der Maaten, "Barnes-Hut-SNE," *1st Int. Conf. Learn. Represent. ICLR 2013 - Conf. Track Proc*., pp. 1–11, 2013.

[18] J. Barnes and P. Hut., "A hierarchical O(N log N) force-calculation algorithm," *Nature*, 324(4):446–449, 1986.

[19] G. J. Dick et al., Community-wide analysis of microbial genome sequence signatures. *Genome Biol* 10, R85 (2009).

[20] T. Kohonen: Self-organizing maps. Volume 0. New York: *SpringerVerlag*; 1997.

[21] A. Gisbrecht, B. Hammer, B. Mokbel, and A. Sczyrba, "Nonlinear dimensionality reduction for cluster identification in metagenomic samples," *Proc. Int. Conf. Inf. Vis*., pp. 174–179, 2013.

[22] C. M. Bishop, M. Svensn, and C. K. I. Williams, "Gtm: The generative topographic mapping," *Neural Computation*, vol. 10, pp. 215–234, 1998.

[23] D. Wang and J. Gu, "VASC: Dimension Reduction and Visualization of Single-cell RNA-seq Data by Deep Variational Autoencoder," *Genomics, Proteomics Bioinforma*., vol. 16, no. 5, pp. 320–331, 2018.

[24] G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis, "Single-cell RNA-seq denoising using a deep count autoencoder," *Nat. Commun.*, vol. 10, no. 1, pp. 1–14, 2019.

[25] E. Pierson and C. Yau, "ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis," *Genome Biol*., vol. 16, no. 1, pp. 1–10, 2015.

[26] Z. Wang and Y. Wang, "Extracting a biologically latent space of lung cancer epigenetics with variational autoencoders," BMC *Bioinformatics*, vol. 20, no. Suppl 18, pp. 1–7, 2019.

[27] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[28] A. J. Holden et al., "Reducing the Dimensionality of Data with Neural Networks," *Science* (80-.)., vol. 313, no. July, pp. 504–507, 2006.

[29] W. Wang, Y. Huang, Y. Wang, and L. Wang, "Generalized Autoencoder: A Neural Network Framework for Dimensionality Reduction DeepVision: Deep Learning for Computer Vision 2014," *CVPR Work*., pp. 490–497, 2014.

[30] T. Abe, H. Sugawara, S. Kanaya, M. Kinouchi, and T. Ikemura, "Self-Organizing Map (SOM) unveils and visualizes hidden sequence characteristics of a wide range of eukaryote genomes," *Gene* 365, 27–34 (2006).

[31] G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller, "Scikit-learn," *GetMobile Mob. Comput. Commun*., vol. 19, no. 1, pp. 29–33, 2015.

[32] M. Daszykowski and B. Walczak, "Density-Based Clustering Methods," *Compr. Chemom*., vol. 2, pp. 635–654, 2009.

[33] E. Amigó, J. Gonzalo, J. Artiles, and F. Verdejo, "A comparison of extrinsic clustering evaluation metrics based on formal constraints," *Inf. Retr. Boston*., vol. 12, no. 4, pp. 461–486, 2009.

[34] B. Chazelle, "An optimal convex hull algorithm in any fixed dimension," *Discrete Comput. Geom*., vol. 10, no. 1, pp. 377–409, 1993.

[35] V. Satopää, J. Albrecht, D. Irwin, and B. Raghavan, "Finding a 'kneedle' in a haystack: Detecting knee points in system behavior," *Proc. - Int. Conf. Distrib. Comput. Syst*., pp. 166–171, 2011.

[36] A. Rosenberg and J. Hirschberg, "V-Measure: A conditional entropy-based external cluster evaluation measure," *EMNLP-CoNLL 2007 - Proc. 2007 Jt. Conf. Empir. Methods Nat. Lang. Process. Comput. Nat. Lang. Learn.*, no. June, pp. 410–420, 2007.

[37] M. Pant, T. Radha, and V. P. Singh, "Particle swarm optimization using Gaussian inertia weight," Proc. - *Int. Conf. Comput. Intell. Multimed. Appl. ICCIMA 2007*, vol. 1, pp. 97–102, 2008.