# Supporting Information for

## Label-free SERS detection of proteins based on machine learning classification of chemo-structural determinants

Andrea Barucci,[#a] Cristiano D'Andrea,[#a] Edoardo Farnesi,[#a] Martina Banchelli,[a] Chiara Amicucci,[a] Marella de Angelis,[a] Byungil Hwang,[b] and Paolo Matteini,[*a]

*[a]Institute of Applied Physics "Nello Carrara", Italian National Research Council, via Madonna del Piano 10, Sesto Fiorentino, I-50019, Italy*
*[b]School of Integrative Engineering, Chung-Ang University, Seoul, 06974, Republic of Korea*
*#These Authors contributed equally to this work*
**p.matteini@ifac.cnr.it*

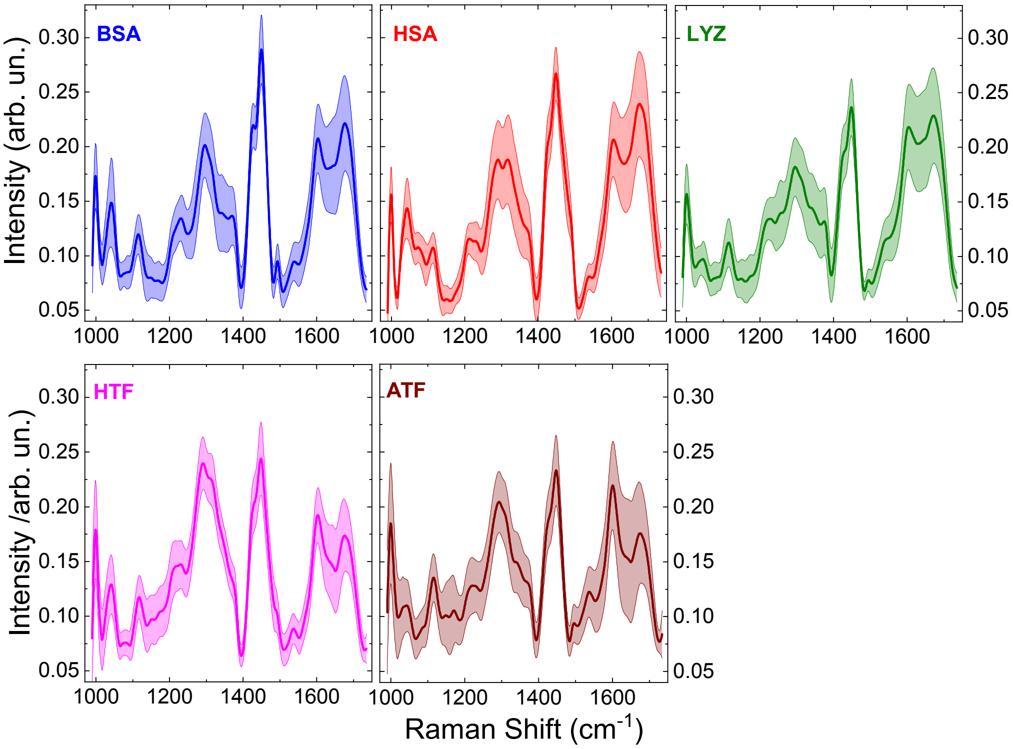

**Fig. S1.** SERS spectra with standard deviation of BSA, HSA, LYZ, HTF and ATF at 1x10$^{-6}$ M under a 785 nm excitation. SD estimated values are: 18 (BSA), 18 (HSA), 18 (LYZ), 17 (HTF) and 20 (ATF).

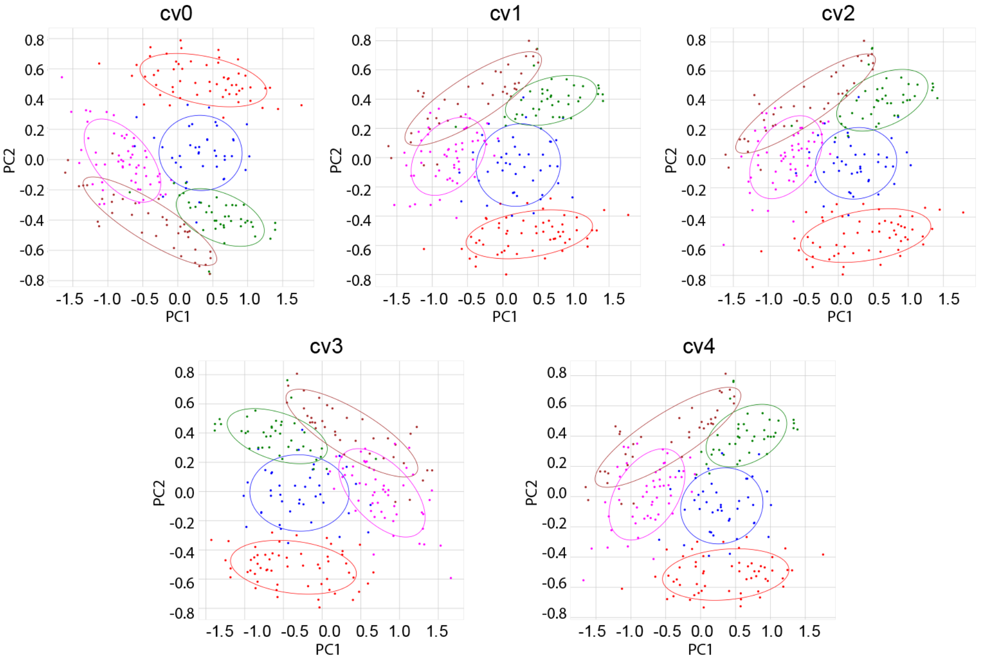

**Fig. S2.** Five-fold (cv0-cv4) cross-validated PC1vsPC2 score plots of SERS spectra of proteins. Data points of each protein are displayed in red for HSA, in blue for BSA, in green for LYZ, in purple for HTF, in brown for ATF along with their confidence ellipses.

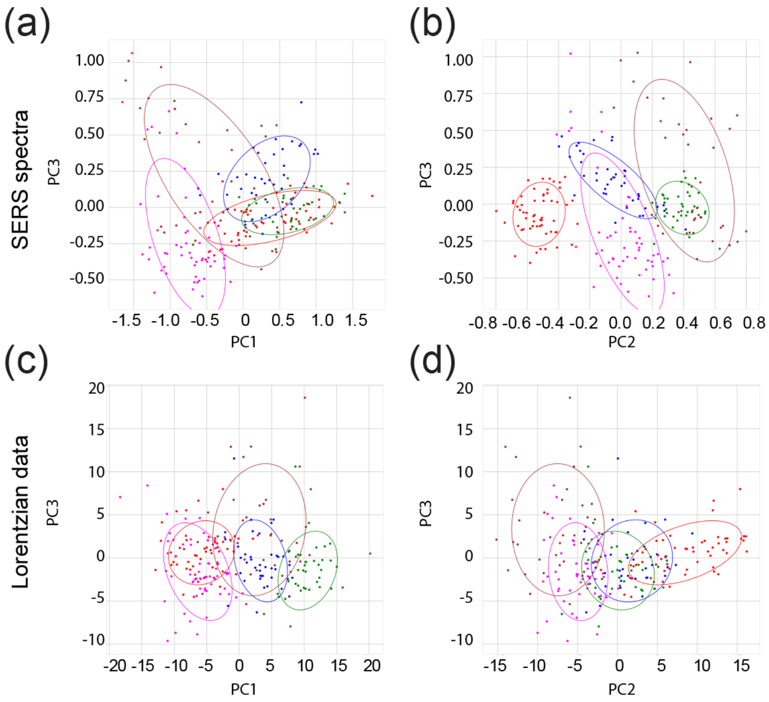

**Fig. S3.** Exemplary cross-validated PC1vsPC3 (a,c)) and PC2vsPC3 (b,d) score plots of SERS spectra (a,b) and of area intensity values of the Lorentzian curves (c,d) of proteins. Data points of each protein are displayed in red for HSA, in blue for BSA, in green for LYZ, in purple for HTF, in brown for ATF along with their confidence ellipses.

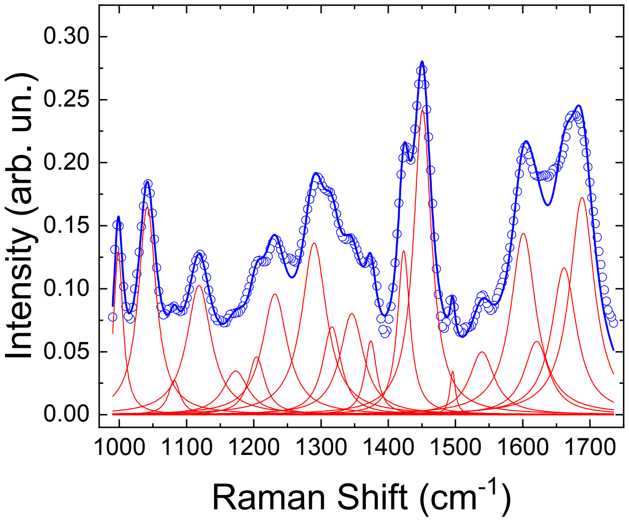

**Fig. S4.** Example of multipeak fit with Lorentzian curves of a SERS spectrum of BSA and according to 19 peaks as listed in **Table 2**.

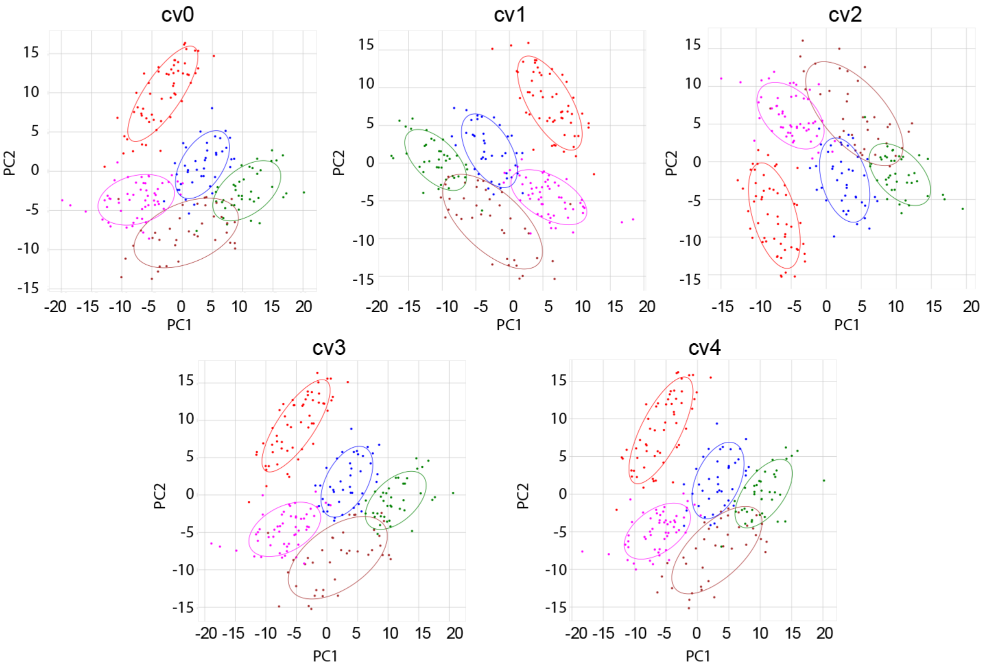

**Fig. S5.** Five-fold (cv0-cv4) cross-validated PC1vsPC2 score plots of SERS spectra of area intensity values of the Lorentzian curves of proteins. Data points of each protein are displayed in red for HSA, in blue for BSA, in green for LYZ, in purple for HTF, in brown for ATF along with their confidence ellipses.

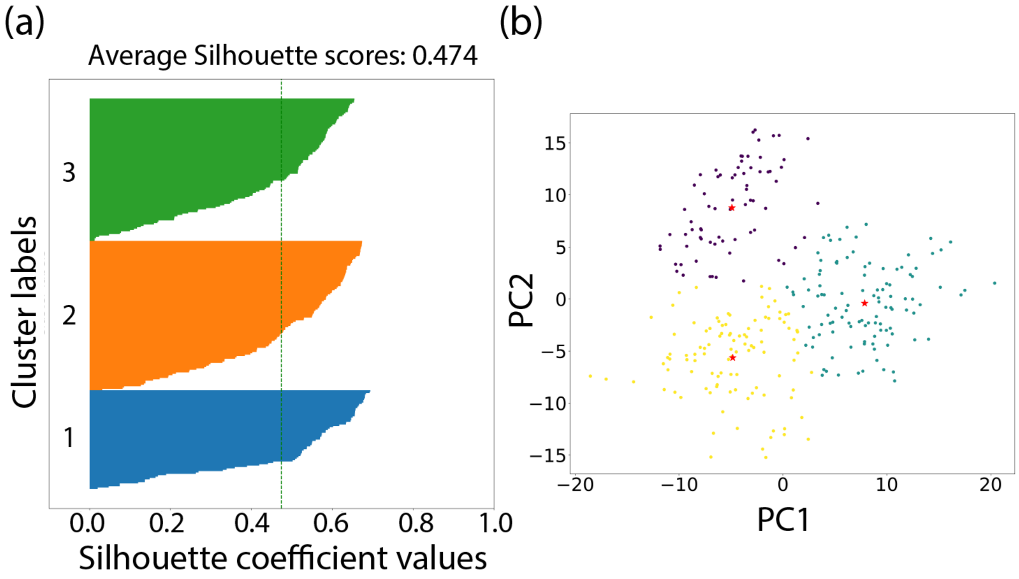

**Fig. S6.** Representation of the KM classification of proteins by Silhouette score metric as obtained by considering PC1 and PC2. A K clustering number = 3 as preferential output of the algorithm (according to **Table 3** with an average Silhouette score of 0.474) is displayed in (a) and imaged in the PC2vsPC1 score plot shown in (b). When compared with the PCA score plot of **Fig. 5(b)** we can notice a perfect superposition of the yellow scores with the HSA scores. The violet cluster is a merging of transferrin scores while the cyan cluster includes BSA, LYZ and part of the ATF scores.

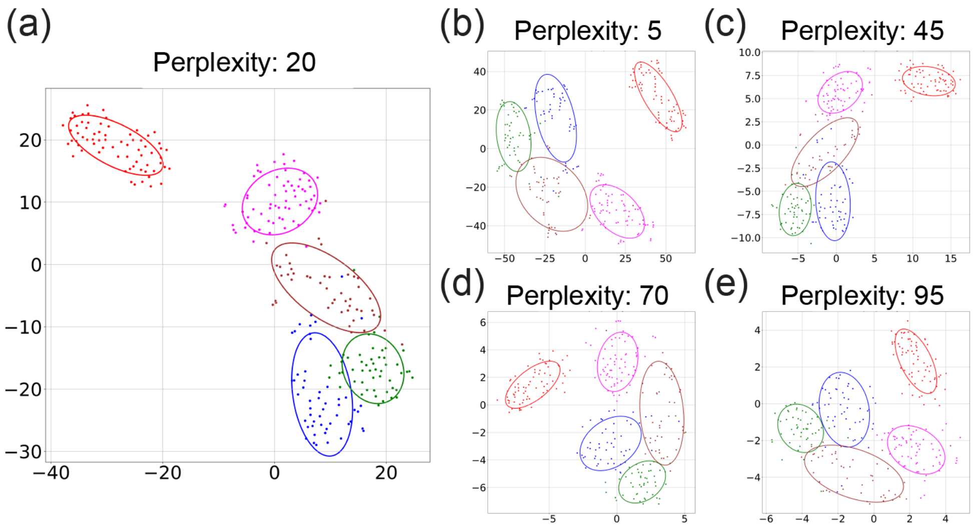

**Fig. S7.** t-SNE score plots as obtained by the application of t-SNE to the area intensity values of the Lorentzian curves once tuned the perplexity parameter to 20 (a), which corresponds to the highest level of separation among protein clusters, or to 5 (b), 45 (c), 70 (d), 95 (e). Data points of each protein are displayed in red for HSA, in blue for BSA, in green for LYZ, in purple for HTF, in brown for ATF along with their confidence ellipses.

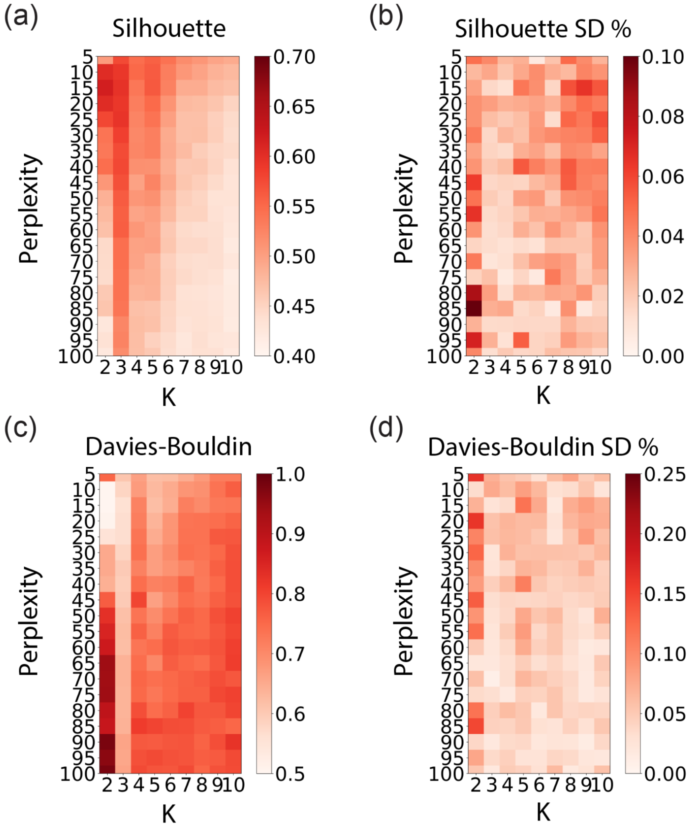

**Fig. S8.** Silhouette (a,b) and D-B (c,d) scores (a,c) and relative standard deviation values (b,d) visualized as variations in color intensity as a function of clustering number K (varying from 2 to 10) and of the perplexity parameter (varying from 5 to 100). The highest values of Silhouette and the lowest values of D-B represent the best indexes. A clustering K = 3 appears favored under both Silhouette and D-B whereas K = 5 ranks as a second option.

# Appendix 1

## 1. Multipeak fitting

Raman spectroscopy provides detailed information on chemical structures and molecular interactions, exploiting the inelastic scattering effect between the light and the chemical bonds of the sample under investigation. A typical Raman spectrum shows a number of peaks in which intensity and wavelength position correspond to a specific molecular bond vibration (i.e. OH, CH2, etc.) or group of bonds (i.e. Phe ring breathing). The accurate labeling of these peaks and the corresponding vibrations provide the chemical fingerprint of the analyte. The more complex the molecular structure of the analytes, like proteins in biological applications, the more information will be present in the Raman spectra, with peaks mixed to broad bands produced by overlapping vibrations related to specific amino acid residues or proteins secondary structure (i.e. Amide bands).

Multipeak fit with Lorentzian or Gaussian (or mixed) curves is a strategy for the analysis of Raman spectra and related plasmon-enhanced techniques such as SERS. It permits to decompose a complex spectrum in a number of main vibrations, typically found in literature (as reported in **Table 2**). Main data obtained by the fit (area, center, max intensity, width, etc.) allow an accurate assignment and weight of the Raman modes, making possible the unambiguous identification of the analyte or the molecular characterization (oxidation state, folding/unfolding state, etc.).

Working with a lower number of variables is considered advantageous in data analysis, because algorithm performances can be strongly related to the number of inputs and in its quality. Moreover, as general rule of thumb it is always a good choice to work with a number of variables lower than the number of data. Since a SERS spectrum is composed by a large number of data (typically above 1000 points), some of which are just noise or uncorrelated with the problem at hand, data analysis can be consequently distorted. Conversely, the use of fitted bands as obtained by a multipeak fit prevents the above issue by taking advantage of a smaller number of features (typically below 50).

## 2. Principal Component Analysis (PCA)

PCA is a fundamental technique in multivariate data analysis, evaluating the correlation between variables and their relevance in order to visualize objects (identification of outliers or of classes), to synthesize the description of data (noise and spurious information removal), to reduce data dimensionality and to investigate their principal properties [1]. Technically, PCA is an unsupervised learning method, very similar to clustering, producing a low-dimensional representation of the dataset that contains as much as possible of the original variability. This low-dimensional representation is obtained by geometrically projecting a high-dimensional dataset onto lower dimensions, called principal components (PCs). Aim of the method is to find the best data description limiting as much as possible the number of PCs used. Once PCs have been computed, their scoreplots are able to create a low-dimensional view of the observations, usually more interpretable than the original data while preserving as much information as possible. Such PCA scoreplots can be used to find potential clusters. The idea is that each of the n observations resides in a p-dimensional space, but not all of these dimensions are equally interesting. PCA seeks a small number of dimensions ($d < p$) that are as much informative as possible, where the concept of informative is assessed by the amount that the observations vary along each dimension. Each of the dimensions calculated by PCA is a linear combination of the p variables:

$PCj = a1jx1+a2jx2+...aMjxM \quad \forall j=1\ldots m$

PC1, …, PCM are ordered so that $Var(PC1) > Var(PC2) > Var(PCM)$

The m-PCs preserve all the original data variance, they are independent (without redundancy) and orthogonal.

## 3. Cross validation

One of the most important aspects in data analysis is reproducibility, i.e. the results should be independent from small perturbations of the dataset. In this work we verified that PCA as well as machine learning classifiers met this requirement. Accordingly, we implemented a 5-fold cross validation on PCA and t-SNE, checking the occurrence of highly reproducible results as actually demonstrated in **Figs. S2, S5 and S7.** Similarly, machine learning algorithms were trained and tested using a 5-fold cross-validation.

## 4. Classification algorithms

Computational models using data (experience) to improve their performance in terms of prediction and/or classification are usually defined as machine learning models/methods [2-4]. In this work we selected 4 machine learning algorithms for classification of our protein samples: linear discriminant analysis (LDA), support vector machines (SVM), K-nearest neighbors (KNN) and K-means (KM), using PCA-processed data obtained from SERS spectra and from area intensity values of the Lorentzian bands used for fitting the SERS spectra, respectively, as input. In order to prevent dependency of the results from a particular choice of the dataset, a 5-fold cross-validation was implemented. Moreover, since many algorithms depend on hyperparameters, that is parameters to tune in order to obtain the best algorithm performance, a grid search in the space of these hyperparamethers was performed.

### *4.1 Linear discriminant analysis - LDA*

LDA is a supervised machine learning technique which tries (taking labels into account) to project the sample onto a straight line so that the projection points of intraclass samples are as close as possible and the projection points of the interclass samples are as far apart as possible. In this work, PCA was used to project the entire dataset onto a different feature (sub)space followed by LDA aiming at finding the directions that maximize the separation (or discrimination) between different protein species.

### *4.2 Support vector machines - SVM*

SVM is probably among the most popular and talked about machine learning algorithms. Since its introduction in the 1990s, SVM continues to remain the go-to method for high performance algorithms and with few tuning parameters. SVM is primarily a classifier method that performs classification analysis by constructing hyperplanes in a multidimensional space by separating cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables.

### *4.3 K-Nearest Neighbor - KNN*

KNN approach is a non-parametric method, which assumes that “similar data” stay in close proximity in a certain space. The KNN algorithm treats the variables as coordinates in a multidimensional feature space. In this study we used KNN by setting K to 5 (i.e. as the number of proteins considered), looking for the best accuracy.

### *4.4 K-means (KM) and unsupervised clustering methods*

Clustering analysis represents a class of unsupervised learning techniques (no samples labeling required) for data classification, usually very interesting when dealing with samples composed e.g. of an unknown mixture of elements, that is without prior knowledge. Using these methods data elements are partitioned into groups (subsets), each one representing a collection of “similar elements” based on a intraclass/interclass distance function (different functions can be available). In this work an Euclidean distance function was used, and grid search was implemented to determine the optimum number of clustering, while Silhouette and Davies-Bouldin scores were used as a metric to measure algorithm performance [5, 6].

## 5. t-SNE

t-Distributed Stochastic Neighbor Embedding (t-SNE) is an unsupervised, non-linear technique originally used for visualizing high-dimensional data and allowing to project higher dimensional data in a 2D space without loss of information [7]. In simpler terms, t-SNE provides the perception of how data are arranged in a high-dimensional space. t-SNE algorithm calculates a similarity estimation between pairs of samples in the high-dimensional space and in the low-dimensional space, attempting to optimize these two similarity measures using a cost function. t-SNE differs from PCA in preserving only small pairwise distances or local similarities. Differently PCA is concerned with maintaining large pairwise distances in order to maximize variance. Although both PCA and t-SNE have their own pros and cons, some key differences between PCA and t-SNE can be noted as follows: 1) PCA is less computationally expensive than t-SNE (in fact t-SNE can take several hours on million-sample datasets where PCA will finish in seconds or minutes); 2) t-SNE is a probabilistic technique, while PCA is a mathematical one.

## References

1. J. Lever, M. Krzywinski, and N. Altman, "Points of significance: principal component analysis," Nat. Methods **14**, 641–642 (2017).
2. C. Bishop, *Pattern Recognition and Machine Learning* (Springer-Verlag, 2006).
3. A. Geron, *Hands-On Machine Learning With Scikit-Learn and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems* (Oreilly & Associates Inc 2017), p. 543.
4. S. Raschka, *Python Machine Learning - Second Edition: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow* (Packt Publishing, 2017).
5. P. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," J. Comp. Appl. Math. **20**, 53–65 (1987).
6. D. L. Davies and D. W. Bouldin, "A cluster separation measure," IEEE T. Pattern. Anal. **PAMI-1**, 224-227 (1979).
7. L. J. P. van der Maaten and G. E. Hinton, "Visualizing high-dimensional data using t-SNE," J. Mach. Learn. Res. **9**, 2579-2605 (2008).