Received: 9 October 2023 | Revised: 20 February 2024 | Accepted: 8 March 2024

DOI: 10.1002/cem.3544

RESEARCH ARTICLE

Journal of CHEMOMETRICS WILEY

Use of *t*-distributed stochastic neighbour embedding in vibrational spectroscopy

François Stevens[1] | **Beatriz Carrasco**[2] | **Vincent Baeten**[1] | **Juan A. Fernández Pierna**[1]

[1]Quality and Authentication of Agricultural Products Unit, Knowledge and Valorisation of Agricultural Products Department, Walloon Agricultural Research Centre (CRA-W), Gembloux, Belgium

[2]Mining4Quality, Murcia, Spain

Correspondence
François Stevens, Quality and Authentication of Agricultural Products Unit, Knowledge and Valorisation of Agricultural Products Department, Walloon Agricultural Research Centre (CRA-W), Gembloux, Belgium.
Email: f.stevens@cra.wallonie.be

Abstract

The *t-distributed stochastic neighbour embedding* algorithm or *t-SNE* is a non-linear dimension reduction method used to visualise multivariate data. It enables a high-dimensional dataset, such as a set of infrared spectra, to be represented on a single, typically two-dimensional graph, revealing its global and local structure. t-SNE is very popular in the machine learning community and has been applied in many fields, generally with the aim of visualising large datasets. In vibrational spectroscopy, t-SNE is gaining notoriety but principal component analysis (PCA) remains by far the reference method for exploratory analysis and dimension reduction. However, t-SNE may represent a real aid in the analysis of vibrational spectroscopic datasets. It provides an at-a-glance global view of the dataset allowing to distinguish the main factors influencing the spectral signal and the hierarchy between these factors, and gives an indication on the possibility of performing predictive modelling. It can also provide great support in the choice of the pre-processing, by comparing rapidly different general pre-processing approaches according to their effect on the variable of interest. Here we propose to illustrate these advantages using different datasets. We also propose an approach based on a synergy between the t-SNE and PCA methods, allowing respective advantages of each to be exploited.

KEYWORDS

data exploration, principal component analysis, *t*-distributed stochastic neighbour embedding, vibrational spectroscopy, visualisation

1 | INTRODUCTION

The *t*-distributed stochastic neighbour embedding or t-SNE algorithm[1–3] is a nonlinear dimensionality reduction technique used to visualise high-dimensional data. It works by constructing a probability distribution that measures the similarity between pairs of high-dimensional data points. Then, it optimises a lower-dimensional representation, aiming to minimise the divergence between the original and transformed distributions, effectively mapping the data points to a lower-dimensional space. t-SNE is able to reveal the organisation of the dataset locally and more globally. For instance, as not all the variation can be mapped in the lower dimension, a user-adjustable parameter called *perplexity* allows the algorithm to balance the preservation of local and global structures. t-SNE is commonly used to explore and visualise

complex datasets, revealing patterns, relationships and clusters that may not be easily discernible in the original high-dimensional space.

t-SNE is popular in a wide variety of disciplines. It has been used in different types of multivariate data to, among other things, process natural language,[4] aid in tumour identification,[5] explore patterns in music,[6] interpret geological data,[7,8] better understand ancient Egyptian paintings[9] or compare medical treatments.[10] In transcriptomics, t-SNE is considered as a cornerstone for the exploration and interpretation of single-cell RNA sequencing datasets.[3,11] In vibrational spectroscopy, however, according to our experience, the popularity of t-SNE remains limited. t-SNE was used, for example, to explore the patterns in a dataset of Vis-NIR hyperspectral images of different paper types,[12] to reduce the dimension of Near-InfraRed (NIR) datasets of agri-food products before applying a classification method,[13] to differentiate specific human cells measured by infrared microscopy and subjected to different treatments,[10] to verify the feasibility of using NIR Spectroscopy (NIRS) to assess chemical oxygen demand in water systems[14] or to evaluate the significance of the spectral difference between tea categories.[15]

t-SNE is a versatile method and is not dedicated to a single use. It was first proposed as a pure visualisation method,[1] that is, a visual aid in exploratory data analysis. Used as a preliminary step, it can help identifying relationships between the data matrix and other quantitative or categorical variables and even show the hierarchy of importance of these variables. It can therefore provide a first indication of the feasibility of further predicting these variables from the data matrix using regression or classification modelling techniques.[14–17] The strong dimensionality reduction capability of t-SNE can also help to explore the relationships between the original variables. For example, Linderman et al.[11] implemented a heatmap-style visualisation for scRNA-seq based on one-dimensional t-SNE for simultaneously visualise the expression patterns of thousands of genes. But t-SNE can also be used to visualise and explore the predictive model itself. For example, Hajibabaee et al.[18] applied different binary classification methods on t-SNE scores instead of the original dataset to represent decision surfaces and assess the effect of class imbalance in these models. Hoyt & Owen[19] used t-SNE to inspect different layers of a convolutional neural network, revealing the importance of each layer and how relevant non-linear information was captured in the features. Going a step further, attempts have also been made to use t-SNE as part of the predictive workflow itself. Poličar et al.[20] developed a method for embedding new samples into the t-SNE map constructed with reference samples, allowing them to be assigned a class. Shekhar et al.[21] successfully combined t-SNE with density-based partitioning to identify cell subpopulations from high-dimensional mass cytometry data. Regarding theoretical aspects, Linderman and Steinerberger[22] showed that t-SNE is often able to recover well-separated clusters and, under specific parameter constraints, is equivalent to spectral clustering, making it a legitimate method for partitioning.

In this paper, we propose to illustrate the opportunities and benefits of using t-SNE for the analysis of vibrational spectroscopic data. We present examples based on two datasets, one dataset of pesticide samples measured by Raman microscopy and one dataset of pure and adulterated oregano samples originating from two countries and measured by NIRS. Sometimes, the dominant factor(s) influencing the position of the points on the t-SNE map is (are) not the one(s) of interest to the analyst. This is the case, for example, when the objective is to determine the composition of a product and important batch effects are present, which may mask or make it difficult to observe the effect of the relevant factors. To address this issue, we also propose an approach that combines t-SNE and principal component analysis (PCA).

2 | MATERIAL AND METHODS

2.1 | t-SNE

Here we propose a brief theoretical summary of t-SNE. For a more complete description of the original algorithm and two faster variants, we refer to the original articles.[1,11,23] The method is also summarised in Figure 1. t-SNE has been implemented in different languages and software's.[24,25] The script combining PCA and t-SNE is available on request from the authors of this article.

Given a set of N objects represented in a high-dimensional space by the data points $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N$, t-SNE creates a low-dimensional space, called the t-SNE map, where each data point $\boldsymbol{x}_i$ is represented by its low-dimensional counterpart $\boldsymbol{y}_i$. In both spaces, a measure of similarity having the form of a mathematical probability is defined for each pair of data points. In order for the t-SNE map to represent as faithfully as possible the similarities between objects and thus the local and global structures of the dataset, the Kullback–Leibler (KL) divergence between the measures of similarity in the two spaces is minimised.

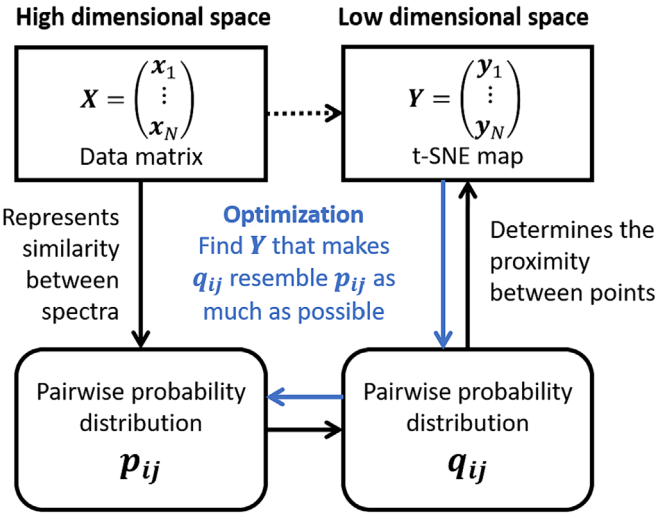

FIGURE 1 General outline of *t*-distributed stochastic neighbour embedding (t-SNE). Pairwise probability distributions are defined to represent, on the one hand, the similarity between the spectra in the high dimensional space and, on the other hand, the proximity between points in the low dimensional space. Then, gradient-descent optimization is performed to find a solution *Y* that maximises the resemblance, that is, minimises the Kullback–Leibler divergence between *p* and *q*.

In the high-dimensional space, the similarities between objects are first obtained by converting the Euclidean distances between data-points into conditional probabilities. In essence, the conditional probability $p_{j|i}$ represents the likelihood that the *i*th object would choose the *j*th object as its neighbour in the low-dimensional space, assuming that neighbours are selected in proportion to their probability density. The conditional probability $p_{j|i}$ is characterised by a Gaussian distribution centred at $\boldsymbol{x}_i$.

$$p_{j|i} = \frac{e^{-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2 / 2\sigma_i^2}}{\sum_{k \neq i} e^{-\|\boldsymbol{x}_i - \boldsymbol{x}_k\|^2 / 2\sigma_k^2}}$$

where, for each $\boldsymbol{x}_i$, the variance of the Gaussian, σ_i, is automatically adjusted to ensure that

$$2^{-\sum_j p_{j|i} log_2 p_{j|i}} = P$$

where the term at the left side of the equal sign represents the so-called perplexity and can be interpreted as an approximate measure of the effective number of neighbours while, on the right side, the perplexity value P is a scalar hyperparameter fixed by the user. t-SNE further incorporates a symmetrisation of the conditional probabilities by defining the probabilities p_{ij} as

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

In the low-dimensional space, a measure of similarity is also defined for each pair of data points, but this time using a naturally symmetric probability, based on the Student's *t* distribution with one degree of freedom, also called Cauchy distribution

$$q_{ij} = \frac{\left(1 + \left\|\mathbf{y}_i - \mathbf{y}_j\right\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2\right)^{-1}}$$

with $\boldsymbol{y}_i$ the coordinates of object i in the low-dimensional map. The use of this distribution with its very slowly decaying tail allows some flexibility in how large distances between data points in the high dimensional space are represented in the low-dimensional space.

Starting from random values, the t-SNE algorithm searches for the optimal values of the $\boldsymbol{y}_i$, the configuration of data points in the low-dimensional map, by minimising C, the Kullback-Leibler divergence between the joint probability distribution, P, in the high-dimensional space and the joint probability distribution, Q, in the low-dimensional space

$$C = KL(P|Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}$$

The algorithm converges towards a minimum (typically a local minimum) with a gradient descent procedure. Various tricks are used to obtain better quality solutions or to accelerate the convergence such as the 'early compression', the 'early exaggeration', the Barnes–Hut algorithm,[23] the FFT-accelerated Interpolation-based t-SNE or Fit-SNE[11] or the use of parallel or GPU-based computation schemes.[26]

2.2 | Datasets

In this study, two datasets were used to illustrate the applications and advantages of t-SNE. The dataset PESTICIDE consists of the 467 Raman spectra of pesticides of nine different formulations. Raman spectra were acquired with a Confocal Raman Microscope Senterra II spectrometer (Bruker Optics, Ettlingen, Germany) with a 100 mW, 785-nm diode laser and a thermoelectrically cooled CCD detector, operating at −65°C. For spectra collection, the products were first sprayed on the skin of organic apples. A swab previously soaked in a mixture of acetone and water (20/80) was used to retrieve the product directly from the apple. The product was then spread into a single aluminium plate, which was air dried during 10 min at room temperature (20°C). For each sample at least 20 spectra were acquired with an integration time of 10 s and 5 co-additions. Raman intensity was recorded between 50 and 3650 cm^{-1} with a spectral resolution of 4 cm^{-1}. The OPUS 7.8 Software (Bruker Optics, Ettlingen, Germany) was used for spectral data acquisition.

The dataset OREGANO is a database of NIR spectra of pure and adulterated oregano.[27] The experimental design consisted in measuring oregano samples originating from two different countries (one batch by country), pure or adulterated with four different products at five adulteration levels, with three replicates for each combination of factors. This results in a total of 126 spectra. The countries are Italy and Turkey, the levels of adulteration are 1%, 2%, 5%, 25% and 50% and the adulterants are olive leaves, cistus leaves, myrtle or sumac. The absorbance was measured with a FOSS XDS NIR Spectrometer (FOSS Analytics, Denmark) at every 2 nm between 400 and 2500 nm.

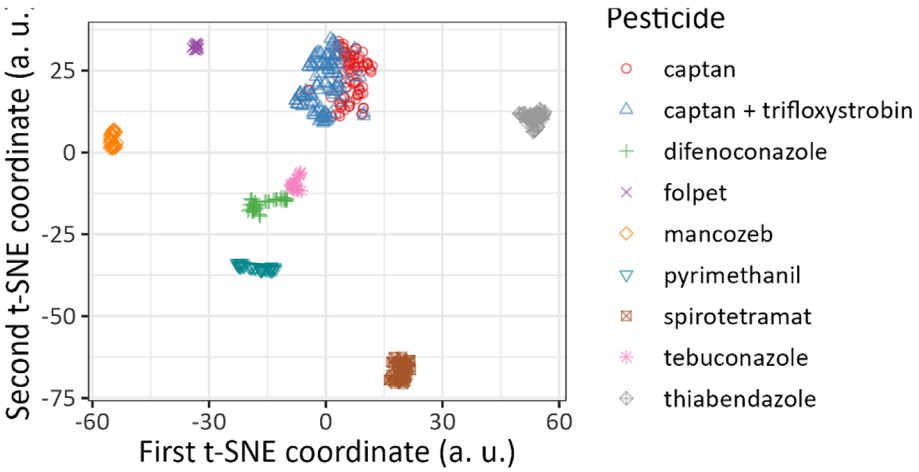

FIGURE 2 *t*-distributed stochastic neighbour embedding (t-SNE) coordinates for the PESTICIDE dataset using a perplexity of 25 and 1D-SNV pre-processing. Colour and symbol both represent the pesticide formula.

2.3 | Advantages and application of t-SNE in vibrational spectroscopy

In vibrational spectroscopy, PCA is the reference method for multivariate data exploration. However, t-SNE has significant advantages over PCA. While PCA is a linear method, t-SNE is also able to capture non-linear relationships and complex structures in the data, making it more suitable for visualising intricate patterns. t-SNE also preserves the local structure and clustering of data points, whereas PCA focuses on capturing global variances. Therefore, with PCA, local fine-grained relationships can be overlooked and, in addition, outliers can significantly impact the principal components and distort the resulting representation. Nevertheless, the most obvious advantage of t-SNE is the possibility to summarise and represent most of the relevant data variation and structure in a single graph, whereas a rigorous exploration of the data with PCA would require the inspection of a large number of score plots, and therefore much more time.

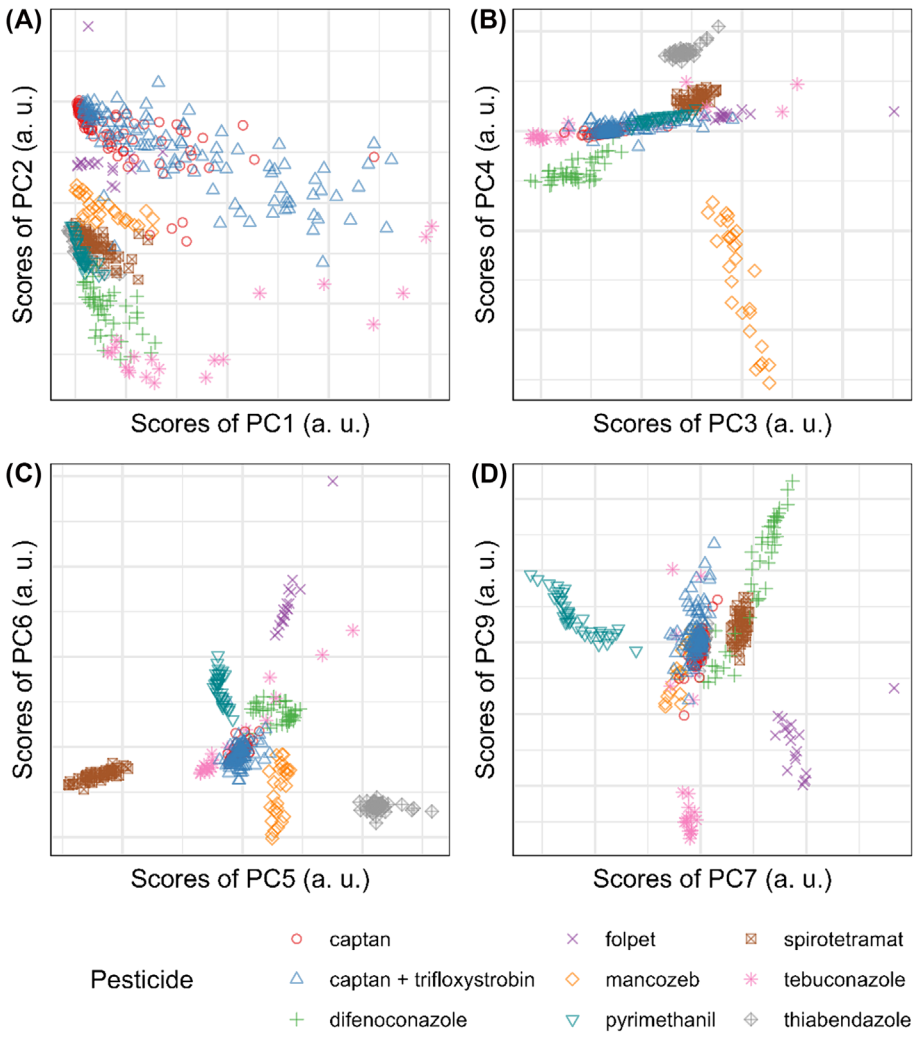

FIGURE 3 Scatterplots of the PCA scores of selected PCs, for the PESTICIDE dataset using a perplexity of 25 and 1D-SNV preprocessing. (A) The scores of PC1 and PC2 indicate that tebuconazole can be well discriminated from all other pesticides, except difenoconazole for which the separation of data points is not perfect. (B) The scores of PC3 and PC4 indicate that difenoconazole, mancozeb and thiabendazole can all be well discriminated from other pesticides. (C) The scores of PC5 and PC6 indicate that folpet, pyrimethanil, spirotetramat and thiabendazole can all be well discriminated from other pesticides. (D) The scores of PC7 and PC9 confirm this discrimination potential for folpet and pyrimethanil. This figure illustrates that using PCA to assess the potential of discrimination of different groups requires the visual control of multiple scatterplots of scores, as the information on separability is distributed among the different PCs or their combinations. In contrast, *t*-distributed stochastic neighbour embedding (t-SNE) can provide a relevant insight with a single figure (Figure 2).

Besides classical exploration of patterns, another attractive utilisation of t-SNE is as a tool to support pre-processing workflow selection. In vibrational spectroscopy, many different methods of pre-processing exist which can be combined in pre-processing workflows. For predictive modelling applications, the rigorous approach would be to include the choice of the pre-processing in the cross-validation framework. This means applying a cross-validation procedure where multiple pre-processing methods or combinations of them would be compared along with different values for the hyper-parameters of the prediction method. This procedure can be very time-consuming. Therefore, many practitioners rely instead on expert knowledge. An intermediate approach is to use t-SNE or an equivalent technique as a decision-support tool. In this perspective, some pre-processing workflows could be compared by simply applying t-SNE to the pre-processed dataset and inspecting the t-SNE plot. In doing so, it also makes sense to compare various values for the perplexity parameter, as this allows setting the focus on different scales in the t-SNE map and thus to better discern the actual dominant patterns in the dataset.

In the article, we propose to illustrate the use of t-SNE to choose the pre-processing workflow with an example based on the PESTICIDE dataset. To quantify the separability of groups by t-SNE with the different settings, the silhouette coefficient[28] was used on the t-SNE maps. The silhouette coefficient is a metric that measures the proximity of each data point to its own group compared to other groups. It is calculated by determining the average distance between a point and all other points within its own group (same pesticide) and the average distance between the point and all the points in the nearest neighbouring group, then subtracting the two values and dividing by the maximum of the two, resulting in a value between −1 and 1, where higher values indicate better separation of the groups. The silhouette coefficient of one group (pesticide) is the mean coefficient over all the objects of this group. In our example, both the minimum and the average values over groups were considered to judge the quality of the group separation.

2.4 | Combining t-SNE with PCA

As previously explained, sometimes, the dominant factor(s) influencing the position of the points on the t-SNE map is (are) not the one(s) of interest to the analyst. To address this issue, we propose an approach that combines t-SNE and PCA. Our use of PCA goes beyond the simple application of t-SNE on the scores of the first PCs to speed up computation time, as presented in the original article,[1] and really aims to offer new perspectives in dataset exploration. In a first step, PCA is applied on the dataset after applying a pre-processing workflow that has been identified as relevant, using

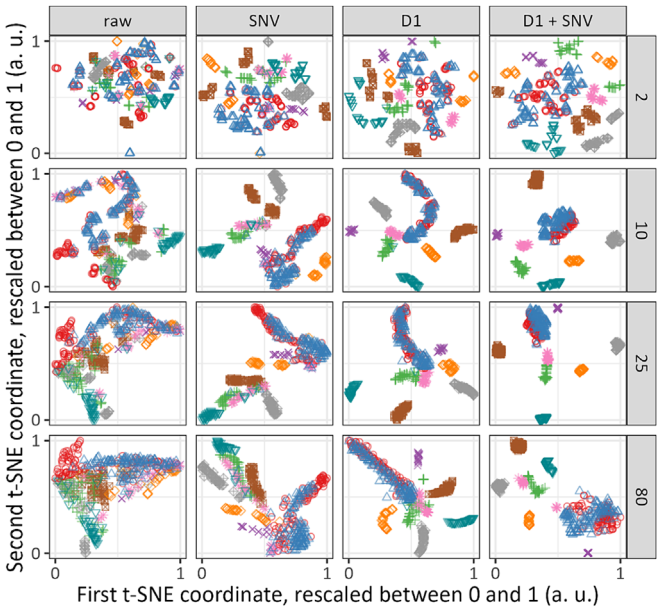

FIGURE 4 *t*-distributed stochastic neighbour embedding (t-SNE) maps for different pre-processing workflows (columns) and different values of the perplexity parameter (rows). The D1 + SNV pre-processing with perplexity values of 10 and 25 seem to be the best scenarios, as these are the only cases where all pesticides form separate clusters, except for the two formulas containing captan. For shape and colour legend, please refer to Figure 2 or 3.

a large number of PCs in order to capture most of the total variance. The PCs corresponding to the irrelevant but dominant variability factors are then identified. In general, due to their importance, these factors are found in the first PCs. Next, t-SNE is applied to the matrix of the scores for all the PCs except the ones of the irrelevant factors. Doing this, the new t-SNE map focuses on the secondary sources of variability and is able to more clearly display the potential influence of relevant factors. In this paper, we illustrate the benefits of the PCA-t-SNE approach on the OREGANO dataset.

3 | RESULTS

The following example uses the PESTICIDE dataset to illustrate the potential of t-SNE as a tool to explore a dataset, thanks to its ability to represent the dataset structure with a single figure. A scatterplot of the t-SNE scores for the two t-SNE coordinates and a shape and colour coding based on the ground truth for pesticide (Figure 2) shows clearly that most of the pesticides form strongly separated clusters, except for pure captan and the mixture containing captan, which form a single cluster with a relatively good separation between these two within it. These patterns indicate that the nature of the pesticides is the most important factor of variability within the Raman spectral dataset. This also suggests that it should be possible to build discrimination models between these different pesticides that perform well on similar data. In contrast, visualising the same dataset with the PCA method leads to a large number of possible scatterplots of scores to compare, and information on the separability of the different pesticides is scattered between them (Figure 3).

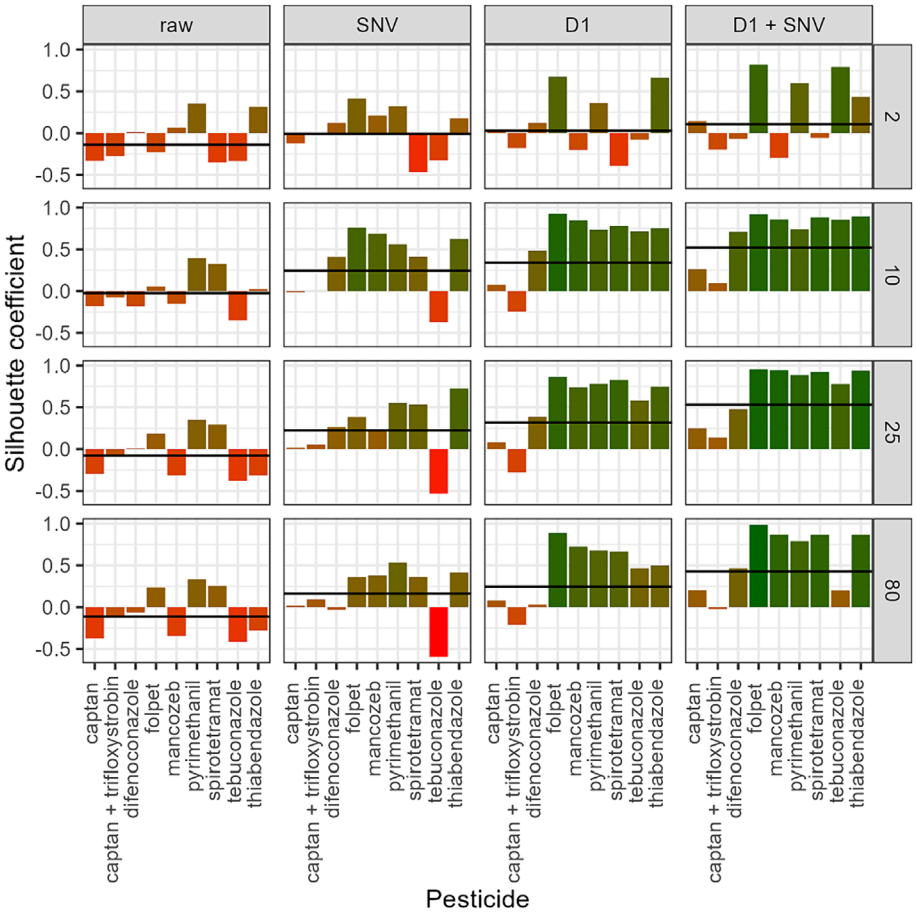

FIGURE 5 Silhouette coefficient for each pesticide calculated from the *t*-distributed stochastic neighbour embedding (t-SNE) scores and used to objectivize the quality of the clustering for the different pre-processing workflows (columns) and perplexity values (rows). The horizontal black line represents the mean silhouette coefficient over all the samples. The combination of pre-processing workflows D1 + SNV with perplexity 25 shows the best results, as it has both the highest value for the minimum silhouette coefficient (captan + trifloxystrobin) and the highest mean silhouette coefficient value.

To illustrate the use of t-SNE as a tool to support pre-processing workflow selection, we tested four pre-processing workflows and four different perplexity values between 2 and 80 in the pesticide dataset. The pre-processing workflows tested are 'none' (the raw spectra), the Standard Normal Variate (SNV), the Savitzky-Golay first derivative with a second-order polynomial and a window width of seven points (D1) and the combination of D1 followed by SNV. The parameters for D1 were set based on our experience with the instrument.

Figure 4 shows the resulting t-SNE maps for the different pre-processing workflows and the different values of the perplexity parameter. At first glance, the D1 + SNV pre-processing with perplexity values of 10 and 25 seem to be the best scenarios, as these are the only cases where all pesticides form separate clusters, except for the two formulas containing captan, which form a common cluster.

In order to objectivize the quality of the clustering, Figure 5 shows the silhouette coefficient calculated from the t-SNE scores for each pesticide. The combination of pre-processing workflows D1 + SNV and perplexity 25 shows the best results, as it has both the highest value for the minimum silhouette coefficient (captan + trifloxystrobin) and the highest mean silhouette coefficient value. These perplexity values are consistent with the recommended range 5–50,[1] the typical default value of 30 or the recommended value of the square root of the number of objects, with $\sqrt{467} = 21.6$.[29]

To illustrate the advantage of combining t-SNE with PCA, an example is shown based on the OREGANO dataset. Savitzky-Golay first derivative with second order polynomial and a window width of five points, followed by SNV and autoscaling were applied as pre-processing.

The scatterplot of PC1 and PC2 scores (Figure 6A) indicates that PC1 is strongly related to the country, with all Turkish samples having negative values and all Italian samples having positive values. The t-SNE map of the pre-processed spectra (Figure 6B) also shows that the signal is largely dominated by the effect of the country of origin of the oregano, with the two countries forming well separated clusters (except for one three replicates of highly adulterated Italian oregano). Other trends also seem to be present, but their visualisation is hampered by the importance of the country effect.

Applying t-SNE on the scores of all the PCs except PC1 allows the strong trend related to the country to be eliminated in the t-SNE map (Figure 6C). The influence of the level of adulteration is now clear, with the most adulterated

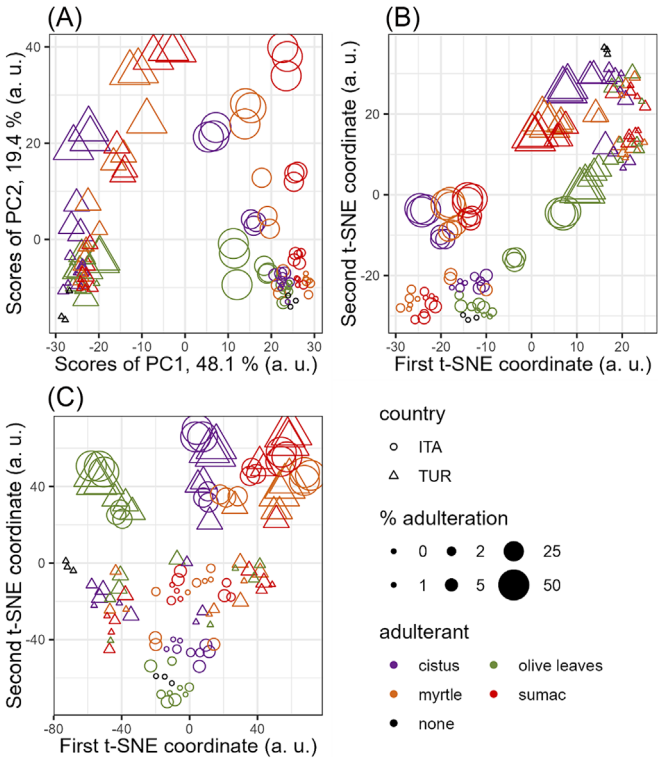

FIGURE 6 Results obtained with principal component analysis (PCA) and *t*-distributed stochastic neighbour embedding (t-SNE) applied on PCA scores, for the OREGANO dataset (a) scores of PC1 and PC2, (b) t-SNE using all scores and (c) t-SNE using all scores except the scores of PC1.

samples located at the top of the figure. In addition, differences between adulterants are also visible, with the samples with the larger levels of adulteration being grouped by adulterant rather than by country. The figure also shows a spectral resemblance between myrtle and sumac at the higher adulteration levels, while olive and cistus form more separated clusters. At the lower adulteration levels (bottom and centre), the country effect still dominates the effect of the adulterant, as the samples are rather clustered by country.

Figure 6C may also give first indications on the possibility to discriminate adulterated from non-adulterated samples. In any case, samples with 25% of more adulteration should be easy to discriminate from non-adulterated samples, as they form completely separate clusters. In the case of Turkish oregano, the three replicates of the non-adulterated sample are clearly separated from the adulterated samples, even with low levels of adulteration, indicating that discrimination should be possible. For Italian oregano, on the other hand, the non-adulterated samples fall into the group of samples adulterated with 1%–5% olive leaves.

4 | CONCLUSIONS

While PCA is a traditional method that has been incorporated into the typical arsenal of chemometrics for several decades, t-SNE is a more versatile approach which is still in the process of adoption, at least by a portion of community of chemical analysis and vibrational spectroscopy. t-SNE fits perfectly with the actual definition of chemometrics, 'the chemical discipline that uses mathematical and statistical methods, (a) to design or select optimal measurement procedures and experiments, and (b) to provide maximum chemical information by analysing chemical data'.[30] Indeed, t-SNE has mathematical foundations in statistics and information theory and the patterns revealed by a t-SNE analysis can help reorienting the experimental design or providing chemical insight.

In practice, t-SNE represents a real help in the analysis of vibrational spectroscopic datasets. It provides a global view of the dataset allowing to distinguish the main factors influencing the spectral signal and the hierarchy between these factors. It can also provide strong support in the choice of pre-processing, by comparing rapidly different pre-processing approaches according to their effect on the variable of interest. In addition, the PCA-t-SNE approach presented in this article allows for a finer exploration when non-relevant factors dominate the global variability. Therefore, we recommend using t-SNE or, if needed, PCA-t-SNE at the beginning of the analysis.

It is important to highlight that, despite its ease of use and the swiftness with which it delivers an interpretable result, t-SNE fundamentally diverges from linear factorization methods such as PCA, introducing unique constraints, restrictions, and limitations that demand careful consideration in its application. At our opinion, the main points to consider are the fact that t-SNE is not deterministic, not parametric and that the results are largely dependent on parameters that can be fixed by the user. We propose to detail these points below.

First, unlike PCA, t-SNE is indeed not deterministic. This is due to its random initialization and the fact that the optimization method does not guarantee finding the absolute minimum. Therefore, running the algorithm multiple times on the same data with the same settings may lead to different results, as the algorithm may converge to different local minima. This may be viewed with suspicion and may cause the user to question the validity of the solution. However, although the orientation and the positions and distances between individual points may vary, the overall patterns and relationships between data points are usually preserved across multiple runs.

The second limitation of t-SNE is linked to the fact that it is not parametric. Unlike PCA, there is not explicit transformation between the original variables and the new coordinates. Therefore, there is no simple way to interpret the new coordinates and relate them to the spectra, as the inspection of the loadings with PCA would allow. However, some approaches and tools have been developed to increase the potential insight provided by t-SNE.[31,32]

Third, it must be emphasised here that t-SNE relies on some parameters that may impact the resulting visualisation. For example, the *perplexity* is a parameter that influences the balance between local and global aspects of the resulting embedding. Different value choices have been proposed, such as 5–50,[1] the square root of the number of objects[29] or 30, the default value in most t-SNE implementations. Note that more advanced extensions of t-SNE allow automatic scale adjustment without the need to fix the perplexity, at the expense of theoretical complexity.[33–35] Another parameter to fix is the learning rate, which defines the step size in the gradient-descent optimization. However, according to our experience, software defaults or rule of thumbs values are sufficient to provide satisfying results for the different t-SNE parameters. In contrast, fine-tuning is usually recommended for large or very large datasets (1000–1 M objects).[2,36]

The possibility to visualise non-linearities is also a real advantage. With vibrational spectroscopy, the link between the target variable and the spectral signal is usually linear or nearly linear, due to the intrinsic linearity of the Beer-

Lambert law under favourable conditions. This justifies the use of a linear multivariate exploration methods such as PCA. However, many situations such as complex sample matrices, physical effects or instrumental limitations lead to non-linearities. In this context, t-SNE offers the possibility to visualise the structure of the dataset encompassing these non-linearities. In practice, in predictive modelling applications, when no good result is obtained using a linear prediction method, such as partial least-squares, the application of t-SNE can provide more information. If trends or clusters related to the variable of interest appear on the t-SNE map, this means that alternative non-linear predictive methods, such as support vector machine or neural network, might provide better results. In contrast, if nothing appears on the t-SNE map after having considered the different relevant pre-processing workflows and a wide range of perplexity values, and excluded the dominant and non-relevant variance factors using the presented PCA-t-SNE approach, then a predictive model with good performances is in principle out of reach.

ACKNOWLEDGEMENTS

The authors would like to thank Olivier Pigeon and Pierre Hucorne (Protection, control products and residues Unit of the CRA-W) for providing the samples of the PESTICIDE dataset and Quentin Arnould (Quality and authentication of agricultural products Unit of the CRA-W) for spectral acquisition as well as Jet Van De Steene (Ciboris) for providing the samples of the OREGANO dataset and Sandrine Mauro and Stéphane Brichard (Quality and authentication of agricultural products Unit of the CRA-W) for spectral acquisition. We would also like to thank Tom Fearn and Janet Riedl for their inspiring comments during the reflexion that led to the decision to write this article.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/cem.3544.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID


François Stevens https://orcid.org/0000-0002-9823-159X

REFERENCES

1. van der Maaten L, Hinton G. Visualizing Data Using T-SNE. *J Mach Learn Res*. 2008;9:2579-2605.
2. Belkina AC, Ciccolella CO, Anno R, Halpert R, Spidlen J, Snyder-Cappione JE. Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nat Commun*. 2019;10(1):5415. doi:10.1038/s41467-019-13055-y
3. Cieslak MC, Castelfranco AM, Roncalli V, Lenz PH, Hartline DK. T-distributed stochastic neighbor embedding (t-SNE): a tool for eco-physiological transcriptomic analysis. *Mar Genomics*. 2020;51(September):100723. doi:10.1016/j.margen.2019.100723
4. Heuer H. Text Comparison Using Word Vector Representations and Dimensionality Reduction. In *8th European Conference on Python in Science (Euroscipy 2015)*; 2016; 13–16.
5. Jamieson AR, Giger ML, Drukker K, Li H, Yuan Y, Bhooshan N. Exploring nonlinear feature space dimension reduction and data representation in breast CADx with Laplacian Eigenmaps and T-SNE. *Med Phys*. 2010;37(1):339-351. doi:10.1118/1.3267037
6. Hamel P, Eck D. Learning Features from Music Audio with Deep Belief Networks. *Proc. 11th Int. Soc. Music Inf. Retr. Conf. ISMIR 2010* **2010**, No. Ismir, 339–344.
7. Balamurali M, Silversides KL, Melkumyan A. A comparison of T-SNE, SOM and SPADE for identifying Material type domains in geological data. *Comput Geosci*. 2019;125:78-89. doi:10.1016/j.cageo.2019.01.011
8. Leung R, Balamurali M, Melkumyan A. Sample truncation strategies for outlier removal in geochemical data: the MCD robust distance approach versus t-SNE ensemble clustering. *Math Geosci*. 2021;53(1):105-130. doi:10.1007/s11004-019-09839-z
9. Alfeld M, Pedetti S, Martinez P, Walter P. Joint data treatment for Vis–NIR reflectance imaging spectroscopy and XRF imaging acquired in the Theban necropolis in Egypt by data fusion and t-SNE. *Comptes Rendus Phys*. 2018;19(7):625-635. doi:10.1016/j.crhy.2018.08.004
10. Xie B, Njoroge W, Dowling LM, Sulé-Suso J, Cinque G, Yang Y. Detection of lipid efflux from foam cell models using a label-free infrared method. *Analyst*. 2022;147(23):5372-5385. doi:10.1039/d2an01041k
11. Linderman GC, Rachh M, Hoskins JG, Steinerberger S, Kluger Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-Seq data. *Nat Methods*. 2019;16(3):243-245. doi:10.1038/s41592-018-0308-4
12. Devassy BM, George S, Nussbaum P. Unsupervised clustering of hyperspectral paper data using T-SNE. *J Imaging*. 2020;6(5). doi:10.3390/JIMAGING6050029

13. Luo N, Yang X, Sun C, Xing B, Han J, Zhao C. Visualization of vibrational spectroscopy for agro-food samples using t-distributed stochastic neighbor embedding. *Food Control*. 2021;126:107812. doi:10.1016/j.foodcont.2020.107812
14. Han X, Xie D, Song H, et al. Estimation of chemical oxygen demand in different water systems by near-infrared spectroscopy. *Ecotoxicol Environ Saf*. 2022;243(March):113964. doi:10.1016/j.ecoenv.2022.113964
15. Mishra P, Nordon A, Tschannerl J, Lian G, Redfern S, Marshall S. Near-infrared hyperspectral imaging for non-destructive classification of commercial tea products. *J Food Eng*. 2018;238:70-77. doi:10.1016/j.jfoodeng.2018.06.015
16. Kanchanatawan B, Sriswasdi S, Thika S, et al. Deficit schizophrenia is a discrete diagnostic category defined by neuro-immune and neurocognitive features: results of supervised machine learning. *Metab Brain Dis*. 2018;33(4):1053-1067. doi:10.1007/s11011-018-0208-4
17. Kessler N, Bonte A, Albaum SP, et al. Learning to classify organic and conventional wheat - a machine learning driven approach using the MeltDB 2.0 metabolomics analysis platform. *Front Bioeng Biotechnol*. 2015;3(MAR):35. doi:10.3389/fbioe.2015.00035
18. Hajibabaee P, Pourkamali-Anaraki F, Hariri-Ardebili MA. An Empirical Evaluation of the T-SNE Algorithm for Data Visualization in Structural Engineering. Proc. - 20th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2021 2021, 1674–1680. 10.1109/ICMLA52953.2021.00267.
19. Hoyt CR, Owen AB. Probing Neural Networks with T-SNE, Class-Specific Projections and a Guided Tour. 2021.
20. Poličar PG, Stražar M, Zupan B. Embedding to reference T-SNE space addresses batch effects in single-cell classification. *Mach Learn*. 2023;112(2):721-740. doi:10.1007/s10994-021-06043-1
21. Shekhar K, Brodin P, Davis MM, Chakraborty AK. Automatic classification of cellular expression by nonlinear stochastic embedding (ACCENSE). *Proc Natl Acad Sci U S a*. 2014;111(1):202-207. doi:10.1073/pnas.1321405111
22. Linderman GC, Steinerberger S. Clustering with t-SNE, provably. *SIMODS*. 2019;1(2):313-332. doi:10.1137/18M1216134
23. Van Der Maaten L. Accelerating T-SNE using tree-based algorithms. *J Mach Learn Res*. 2014;15:3221-3245.
24. van der Maaten L. List of t-SNE implementations https://lvdmaaten.github.io/tsne/#implementations (accessed Aug 2, 2023).
25. Wikipedia. t-distributed stochastic neighbor embedding https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding
26. Chan DM, Rao R, Huang F, Canny JF. T-SNE-CUDA: GPU-Accelerated T-SNE and Its Applications to Modern Data. Proc. - 2018 30th Int. Symp Comput Archit High Perform Comput SBAC-PAD 2018 2019, 330–338. 10.1109/CAHPC.2018.8645912.
27. Van De Steene J, Ruyssinck J, Fernandez-Pierna J-A, et al. Authenticity analysis of oregano: development, validation and fitness for use of several food fingerprinting techniques. *Food Res Int*. 2022;162(Pt A):111962. doi:10.1016/j.foodres.2022.111962
28. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20(C):53-65. doi:10.1016/0377-0427(87)90125-7
29. Oskolkov N. How to Tune Hyperparameters of TSNE 2020; 1–19.
30. Otto M. *Chemometrics: statistics and computer application in analytical chemistry*. Third ed. John Wiley & Sons; 2017. doi:10.1002/9783527699377
31. Rainer RJ, Mayr M, Himmelbauer J, Nikzad-Langerodi R. Opening the black-box of neighbor Embeddings with Hotelling's T2 statistic and Q-residuals. *Chemom Intel Lab Syst*. 2023;238:104840. doi:10.1016/j.chemolab.2023.104840
32. Chatzimparmpas A, Martins RM, Kerren A. T-ViSNE: interactive assessment and interpretation of t-SNE projections. *IEEE Trans vis Comput Graph*. 2020;26(8):2696-2714. doi:10.1109/TVCG.2020.2986996
33. De Bodt C, Mulders D, Verleysen M, Lee JA. Perplexity-Free t-SNE and Twice Student Tt-SNE. In *ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*.; Bruges (Belgium), 2018; i: 25–27.
34. Kobak D, Linderman GC. Initialization is critical for preserving global data structure in both T-SNE and UMAP. *Nat Biotechnol*. 2021;39(2):156-157. doi:10.1038/s41587-020-00809-z
35. Lee JA, Peluffo-Ordóñez DH, Verleysen M. Multi-scale similarities in stochastic neighbour embedding: reducing dimensionality while preserving both local and global structure. *Neurocomputing*. 2015;169:246-261. doi:10.1016/j.neucom.2014.12.095
36. Kobak D, Berens P. The art of using T-SNE for single-cell transcriptomics. *Nat Commun*. 2019;10(1):5416. doi:10.1038/s41467-019-13056-x

How to cite this article: Stevens F, Carrasco B, Baeten V, Fernández Pierna JA. Use of *t*-distributed stochastic neighbour embedding in vibrational spectroscopy. *Journal of Chemometrics*. 2024;1-11. doi:10.1002/cem.3544