Microchemical Journal 209 (2025) 112751

Contents lists available at ScienceDirect

Microchemical Journal

journal homepage: www.elsevier.com/locate/microc

Essential spectra to improve vibrational imaging of pharmaceutical samples

Laureen Coic [a], Yesid Roman Gomez [a], Pierre-Yves Sacré [b], Eric Ziemons [c], Raffaele Vitale [a], Cyril Ruckebusch [a,*]

[a] Univ. Lille, CNRS, LASIRE, Laboratoire Avancé de Spectroscopie pour les Interactions, la Réactivité et l'Environnement, F-59000 Lille, France
[b] University of Liege (ULiege), CIRM, Research Support Unit in Chemometrics, Department of Pharmacy, Liege, Belgium
[c] University of Liege (ULiege), CIRM, Laboratory of Pharmaceutical Analytical Chemistry, Department of Pharmacy, Liege, Belgium

ARTICLE INFO

Keywords:
Essential information
Fourier transform
Principal component analysis
Raman imaging
FTIR imaging
Linear spectral unmixing

ABSTRACT

Raman and infrared vibrational imaging are invaluable techniques for obtaining both spectral and spatial information from complex biological and medical samples. However, the direct analysis of hyperspectral imaging datasets is often hindered by the physical and chemical complexity of raw samples, resulting in vast amounts of data and numerous uncontrolled sources of variance. To address these challenges, the selection of essential spectra – comprising the most linearly distinctive rows of a data matrix – and their targeted analysis offer an efficient method to significantly enhance spectral unmixing. An estimation of the set of essential spectra can be performed by convex hull analysis of the normalized scores resulting from a truncated singular value decomposition. Alternatively, Fourier coefficients at selected harmonic frequencies may be used. Both approaches are valid, even though the data point cloud is represented in two different intrinsic coordinate systems, which translates into two different approximations. In this paper, we picture the use and efficiency of data reduction by essential spectra selection using both alternatives for the analysis of pharmaceutical samples by FTIR and Raman hyperspectral microimaging. These examples provide very large datasets and challenging analytical situations including the identification of very minor compounds and the presence of strong scattering components whose effect can mask the spectral information of other compounds. Our results clearly demonstrate the advantages of analyzing reduced datasets obtained by identifying essential spectra instead of full data in terms of speed, enabling a much faster determination of the drug composition. In addition, they show that the use of Fourier coefficients allows reaching better data reduction rates (down to 0.1% of the original number of measured spectra for the FTIR image investigated).

1. Introduction

The effectiveness of vibrational hyperspectral microimaging is well-established, with applications spanning diverse fields such as biomedicine [1,2], art conservation [3] or agrifood [4]. However, optimizing hyperspectral imaging requires careful balancing of multiple interconnected factors, including signal-to-noise ratio and acquisition time, to maximize the yield of useful information for subsequent data analysis. Various approaches have been proposed to address this challenge including instrumental advancements [5], real-time corrections for undesired physical effects [6,7], the application of data pre-processing techniques [8] and chemometric analysis [9]. In the analysis of hyperspectral vibrational microimaging data, multivariate curve resolution (MCR) [10] is frequently employed to decompose the spectral variation into a sum of relatively few individual components, each one characterized by a specific distribution map and a specific spectral signature. Despite these advancements, rapid acquisition and accurate analysis of vibrational hyperspectral microimaging data remain challenging, particularly when dealing with large samples [11], complex compositions [12] or in the presence of minor compounds [13] as in the characterization of pharmaceutical samples which often require to push the limits of current methodologies for both data acquisition and analysis.

The concept of essential information leverages data redundancy in bilinear spectroscopy datasets. Based on this concept, one can identify the most linearly dissimilar items within a data matrix [14–16]. This approach focuses on two key components: essential spectra (the most distinctive rows of a spectral dataset) and essential variables (the most distinctive columns of a spectral dataset) [17–19]. The identification of essential spectra and essential variables requires the application of an

[*] Corresponding author.
E-mail address: cyril.ruckebusch@univ-lille.fr (C. Ruckebusch).

https://doi.org/10.1016/j.microc.2025.112751
Received 4 September 2024; Received in revised form 23 December 2024; Accepted 10 January 2025
Available online 11 January 2025
0026-265X/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

appropriate data normalization to ensure a convex data geometry [20], wherein the most selective rows or columns correspond to the vertices of the resulting data structure. The selection of essential information to reduce bilinear spectroscopic datasets allows for more efficient analysis of complex data, whose volume can be overwhelming and redundancy considerable, but also potentially enhances results' accuracy permitting to focus only on the most meaningful and distinctive features of the data [21,22]. Recent applications cover near-infrared imaging for the identification of adulterants in food material [23] and plastic characterization [24]. The selection of essential information also paves the way for more efficient data processing and interpretation, ultimately leading to more robust and reliable analytical outcomes. As an example, a twofold acceleration of Raman confocal imaging by selective sampling, *i.e.* performing acquisition solely at pixel positions carrying essential spectra (ES), was recently demonstrated [25–27]. The idea of essential information-based compression was also extended to trilinear datasets to generate a reduced version of the data at hand [28].

A fundamental feature of data reduction approaches based on the selection of ES is that their performance does not depend on the spectral variation the data under study exhibit but on the geometrical structure of the data point cloud in an intrinsic coordinate system [16]. This enables to circumvent some of the issues mentioned above for the analysis of complex samples. ES can be identified through the computation of the convex hull that envelopes the data point cloud. This cloud is represented in the column-space of the data matrix for the identification of the most linearly distinctive spectra among those that have been measured. For subsequent data analysis, only these ES are considered. This data pruning process dramatically simplifies the analysis while retaining the most relevant information the considered measurements encode. To implement this selection practically, a linear transformation of the data is needed in order to perform convex hull calculations in a low-dimensional abstract column-space. In theory, any linear data transformation could be employed. In practice, two approaches were proposed in recent studies based on principal component analysis (PCA) and discrete Fourier Transform (DFT), respectively. With the former, ES are identified in the abstract PCA scores space [14] whereas with the latter, phasor plots of the most relevant harmonics are computed [25].

Both PCA and DFT approaches provide means to represent spectral data in an alternative coordinate system and both methods are valid to approximate the set of ES of a given dataset. However, they differ fundamentally. PCA is a "data-driven" method, in the sense that its basis vectors are oriented to capture the largest sources of variance of the dataset. Sample coordinates (scores) are then obtained by projecting the original data onto these basis vectors. In contrast, DFT is a signal processing technique that converts individual spectra from the wavelength domain to the frequency domain. This transformation results in Fourier coefficients, whose polar coordinates are the amplitude and phase of a complex sinusoidal component, and which serve as sample coordinates in a predetermined coordinate system, independent of the entire dataset itself [29,30].

In this paper, we report the results obtained from the analysis of two pharmaceutical samples using FTIR and Raman hyperspectral micro-imaging: i) a commercial tablet whose composition is publicly available and ii) a falsified formulation seized during the COVID-19 pandemic of unknown composition. These two challenging analytical scenarios are investigated to highlight the differences and complementarity between the PCA and DFT approaches for the selection of ES.

2. Materials and methods

2.1. 1 hyperspectral images of pharmaceutical specimens

Before analysis, samples were glued on a microscope slide and their surface was milled using a system equipped with a tungsten carbide miller (Leica Microsystems GmbH). FTIR imaging of a commercial Afebryl® tablet was performed using a Cary 670/620 Agilent series microscope (Agilent Technologies) equipped with a 15x infrared objective, with a numerical aperture (NA) of 0.62 and a FPA (64 × 64) detector, yielding a spatial resolution of 5.5 μm. Data acquisition was performed in reflection mode, with a resolution of 8 cm^{-1} over the spectral range 850–4000 cm^{-1} and 16 co-added scans. According to publicly available information, the active substances of Afebryl® are acetylsalicylic acid (300 mg), ascorbic acid (300 mg) and acetaminophen (paracetamol, 200 mg). Among other chemicals composing the drug, one may find sodium bicarbonate, anhydrous citric acid, sorbitol, lactose, sodium saccharin and essential oil of lemon. A 448 × 448 FTIR image was acquired resulting in an unfolded hyperspectral imaging dataset, $\mathbf{D}_{1,}$ of dimensions 200,704 × 293.

Raman imaging was performed with a Labram HR Evolution microscope (Horiba scientific) equipped with an EMCCD detector (1600 × 200, Andor Technology Ltd.), and a 50x Fluotar LWD objective (Leica). A 150 × 150 Raman image of a falsified chloroquine was recorded using a 785 nm laser with a reduced power of 45mW at sample [27]. The dwell-time was 2 s per pixel and the total image acquisition took 12 h and 50 min. The measurements provided an unfolded dataset, $\mathbf{D}_{2,}$ of dimensions 22,500 × 1600. The full composition of the sample is unknown, although it was partly elucidated in [13,31]. It contains metronidazole, chloramphenicol, magnesium stearate, acetaminophen, calcium phosphate and starch. Spectral baseline correction was performed by applying asymmetric least squares (parameters: $\lambda = 3.10^4$, $p = 1.10^{-5}$) [32].

The Raman and FTIR datasets investigated in this study can be downloaded from the institutional open research data sharing repository of the University of Liège (see [33]) and can be reused according to the FAIR data principles.

2.2. Assessment of essential spectra

Consider an unfolded hyperspectral imaging dataset **D** of dimensions $M \times N$, encompassing M spectral pixels recorded at N spectral channel. We assume that the matrix **D** is properly pretreated and that its elements d_{mn} are ≥ 0.

To fully comprehend the concept of ES, two key aspects must be emphasized. The first regards convex hull representation: ES correspond to the vertices of the convex hull encapsulating the data point cloud [14–16]. As such, they encode the most linearly distinctive information present within the dataset. Consequently, any measured spectrum can be approximated by a convex linear combination of these ES. Conversely, any spectrum that cannot be described as a combination of ES is by definition an essential spectrum itself. This property allows for significant data reduction through the elimination of redundant "non-essential" spectra which permits to effectively prune the dataset while retaining relevant information. The second crucial aspect relates to normalization: proper data normalization ensures that the most selective data points correspond to the vertices of the convex hull of the data point cloud [20,34], *i.e.* the purest measured spectra can be found among the ES.

To identify ES, one should compute the convex hull that envelopes the data point cloud defined by the projection of the original spectra in the column-space of **D**. For the sake of practicality, an approximation is required so as to perform the estimation of the convex hull in a low-dimensional abstract column-space of **D**. Among the linear transformations that could be applied, both PCA and DFT have been proposed. With PCA, convex hull analysis is performed on the normalized scores [14] whereas, with DFT, the data point coordinates are the complex Fourier coefficients [25].

The score-space **X** of the spectral data matrix **D** can be computed by applying singular value decomposition (SVD), as in Eq. (1):

$$\mathbf{D} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}} = \mathbf{X}\mathbf{V}^{\mathrm{T}} \tag{1}$$

where **U** and **V** contain the left and right singular vectors, respectively,

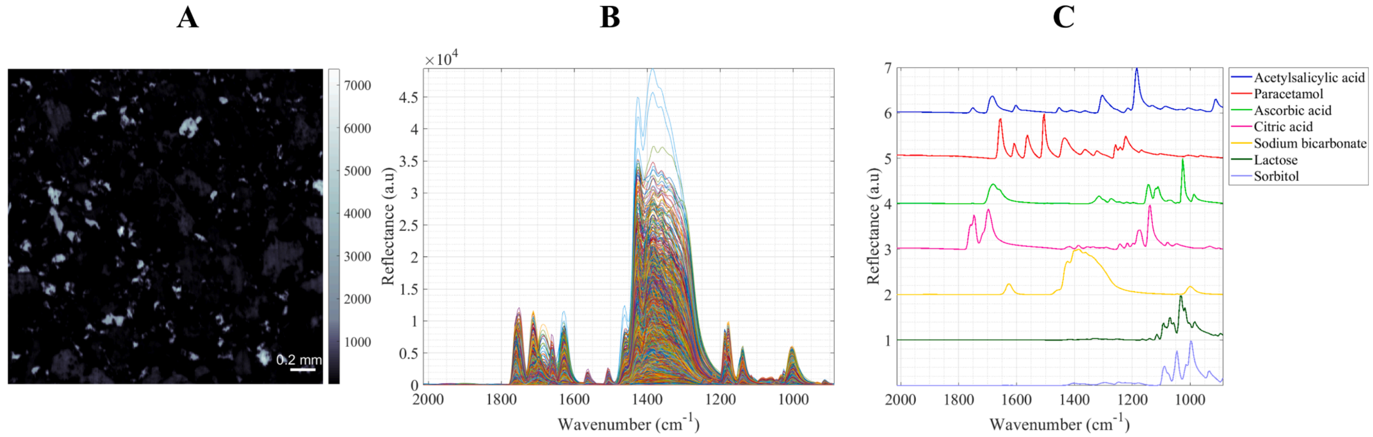


Fig. 1. FTIR imaging of Afebryl®. A) Original FTIR image (averaged across spectral channels); B) selection of FTIR spectra and C) FTIR spectra of the individual components constituting the tablet, extracted from an in-house database.

and **S** corresponds to the diagonal matrix of singular values. Setting all but the first K singular values to zero provides a reduced rank approximation of **D** and a truncation of the SVD model in Eq. (1). The dimensions of **U**, **V** (orthonormal matrices) and **S** are then $M\times K$, $N\times K$, and $K\times K$, respectively, and **X** ,of dimensions $M\times K$, carries an approximation of the coordinates of all the data points in the column-space of **D**.

As mentioned previously, the matrix **X** should be normalized such that the coefficients of the transformed vectors form a convex set with pure items to be located at the vertices. Among the possible options [20,33] to ensure this property, the normalization provided in Eq. (2) can be applied:

$$\breve{X} = X \oslash \mathbf{x}_1 1^T \qquad (2)$$

where $\mathbf{x}_1$ denotes the first column vector of **X**, the operator $\oslash$ represents the element-wise (Hadamard) division and 1 is a vector of ones of size $K\times 1$.

Convex hull calculations can then be performed on the normalized scores $\breve{X}$, resulting in the identification of a subset of P data points (P rows of **X)** corresponding to P ES (P rows of **D**). Full details on the algorithm for the identification of ES within the PCA subspace (ES-PCA) of a hyperspectral imaging data matrix can be found in [14].

Alternatively, the identification of ES can be carried out by the convex hull analysis of the Fourier coefficients derived from the application of DFT to the individual rows of **D** (ES-DFT) [25]. In this case, to ensure the convexity of the coefficients, the spectra in **D** are first normalized by their respective L$_1$-norm, *i.e.* each row vector of **D**, $\mathbf{d}_m^{\mathrm{T}}$, is divided by the sum of the absolute values of all its elements. DFT is then applied, as in Eq. (3):

$$\widetilde{\mathbf{d}}_m = DFT[\mathbf{d}_m] = \begin{bmatrix} \widetilde{d}_{m1} \\ \cdots \\ \widetilde{d}_{mj} \\ \cdots \\ \widetilde{d}_{mN} \end{bmatrix} \qquad (3)$$

with $\widetilde{d}_{mj}$ the complex-valued coefficient provided by the DFT of the m-th spectrum of **D** for the j-th harmonic frequency ($j = 1, \ldots, N$). Alternatively, DFT can be written in a matrix–vector form, as in Eq. (4):

$$\widetilde{\mathbf{d}}_m = \widetilde{\mathbf{W}}\mathbf{d}_m \qquad (4)$$

with $\widetilde{\mathbf{W}}$ a DFT matrix of dimensions $N\times N$ multiplying the N-dimensional signal $\mathbf{d}_m$. The coefficients associated to the real and imaginary part of $\widetilde{d}_{mj}$ correspond to the coordinates, $G_j(m)$ and $S_j(m)$, of the m-th spectrum in the two-dimensional phasor associated with the j-th harmonic frequency of its Fourier representation. For all j, these coordinates form a convex set of coefficients with the most selective data points located at the vertices of the resulting distribution. One can thus retrieve ES by computing the convex hull of the cloud of points with coordinates $[G_j(m), S_j(m)]$ – with $m = 1, \ldots, M$ – for some selected harmonics. We refer to [25] for additional details and results obtained on simulated datasets as well as for the MATLAB code utilized for ES-DFT.

Both ES-PCA and ES-DFT necessitate dimensionality reduction for convex hull calculations. For ES-PCA, this reduction is achieved by deciding on the number of singular values to be considered for the SVD truncation. With ES-DFT, instead, harmonic frequencies are sorted by decreasing energy and truncated accordingly. The two approaches are underlain by different approximation procedures and, therefore, in practice, result in the selection of different sets of ES. The choice of the dimensionality of the subspace used for convex hull calculations presents a critical trade-off between selection and approximation: the more accurate the approximation the larger the number of ES. This balance remains an open issue leading to user-dependent outcomes which can complicate direct comparisons between the results obtained by ES-PCA and ES-DFT.

In order to overcome this issue, we consider the following approach: we apply ES-PCA for a chosen K, performing convex hull calculations on the score matrix $\breve{X}$ of dimensions $M\times K$. This results in the selection of P ES out of the M measured spectra, with $P \ll M$. We denote then SR the selection ratio 100(P/M). We can thus compare the results obtained by applying ES-PCA with those yielded by ES-DFT considering the set of harmonic frequencies (sorted by decreasing energy) providing the same (or a very similar) SR value.

All computations were performed using MATLAB® 2016a (Mathworks, Natick, USA) and the PLS Toolbox/MIA Toolbox (version 8.6.2, Eigenvector Research Inc, Manson, USA).

3. Results and discussion

Fig. 1 illustrates the results obtained from the FTIR imaging analysis of the Afebryl® commercial tablet. Fig. 1A provides the representation of a 2.4 × 2.4 mm^2 area of this tablet. The corresponding hyperspectral microimaging dataset ($\mathbf{D}_1$) contains 200,704 spectra, some of which are plotted in Fig. 1B. Fig. 1C shows the FTIR spectra of the pure compounds known to be present in the commercial tablet. The tablet contains three active ingredients: acetylsalicylic acid, paracetamol (acetaminophen), and ascorbic acid. However, it is sodium bicarbonate, one of the excipients, that contributes the most to the spectral variation (as shown in

Table 1
Results obtained by applying ES-PCA and ES-DFT to dataset $\mathbf{D}_1$ considering data subspaces of decreasing dimensionality for convex hull calculations. For the scenario marked with a *, ES-DFT could not be applied given the original number of spectral data points.

ES-PCA				ES-DFT		
# PCs	Explained variance (%)	# ES	**SR (%)**	# harmonics	Energy (%)	# ES
10	99.6	6127	**2.78**	*	*	*
9	99.4	3350	**1.52**	*	*	*
8	99.3	1687	**0.76**	*	*	*
7	99.0	922	**0.42**	*	*	*
6	98.7	431	**0.21**	105	99.9	432
5	98.3	174	**0.09**	25	95.5	172
4	97.3	81	**0.04**	6	63.4	84
3	96.3	23	**0.01**	1	18.0	22

Fig. 1B).

Table 1 reports the results obtained by applying both ES-PCA and ES-DFT to dataset $\mathbf{D}_1$. For ES-PCA, we investigated various situations starting from a ten-component PCA model explaining 99.6 % of the data variance down to a three-component PCA model accounting for 96.3 % of the data variance. The corresponding number of ES ranges from 6127 (SR = 2.8 %) to 23 (SR = 0.01 %). For ES-DFT, the various sets of ES were estimated considering the number of harmonics needed to reach similar values of SR. The results displayed in Table 1 reveal two contrasted trends. The PCA models generally account for a high fraction of explained data variance (>96 %) even for low SR values whereas, for DFT, the energy value decreases significantly when less than 25 harmonics (SR = 0.09 %) are considered, a situation which corresponds to the identification of 172 ES.

Fig. 2 illustrates the ES selection results obtained by means of ES-PCA and ES-DFT. As expected, considering the fact that different data approximations are performed, the two approaches yield different ES sets

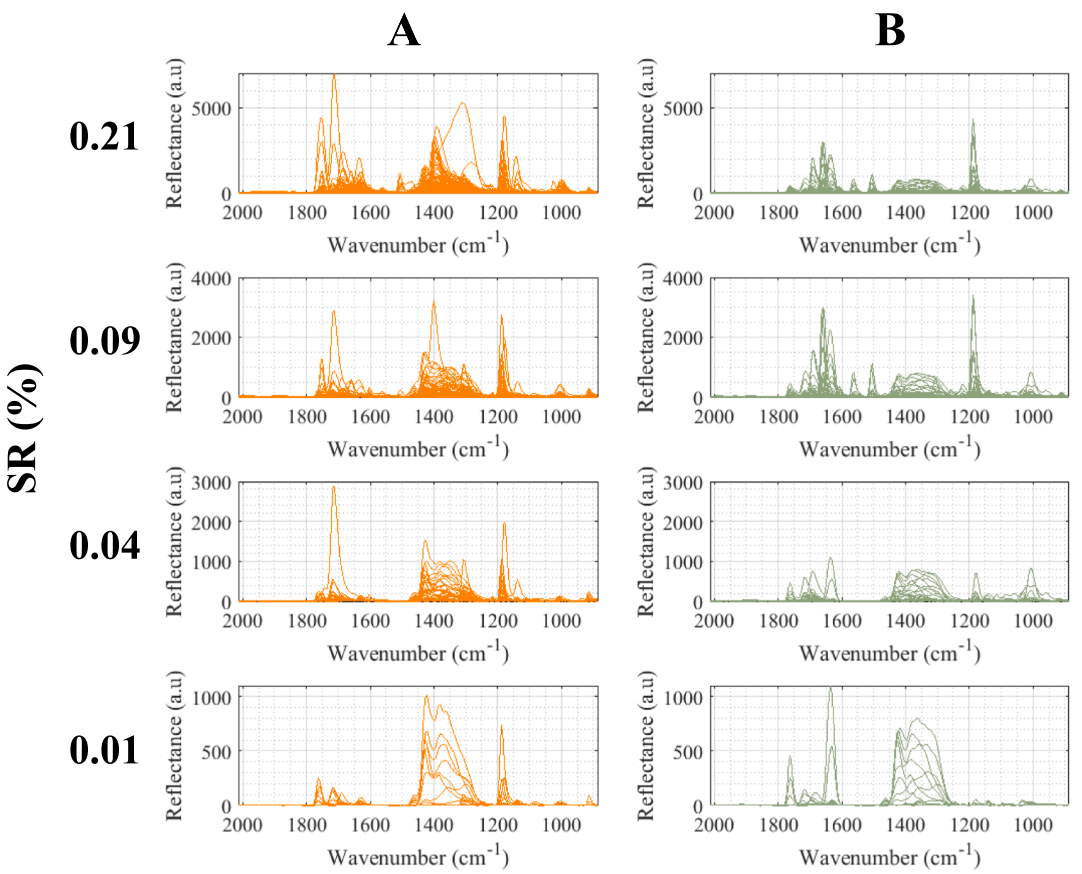


Fig. 2. Essential spectra of dataset $\mathbf{D}_1$ identified by ES-PCA (column A) and ES-DFT (column B) for different SR (%) values.

Table 2
Spectral matching results obtained for dataset $\mathbf{D}_1$. The table lists the best CCs found for the full dataset and for the reduced datasets resulting from the application of ES-PCA and ES-DFT.

			ES-PCA								ES-DFT			
	SR (%)	Full data	2.78	1.52	0.76	0.42	0.21	0.09	0.04	0.01	0.21	0.09	0.04	0.01
Target spectra	Acetylsalicylic acid	**0.97**	**0.9**	**0.90**	**0.90**	**0.90**	**0.89**	**0.81**	**0.81**	**0.78**	**0.87**	**0.81**	**0.72**	0.08
	Ascorbic acid	**0.97**	**0.96**	**0.95**	**0.82**	**0.63**	**0.63**	0.32	0.09	0.02	**0.84**	**0.61**	0.39	0.32
	Citric acid	**0.96**	**0.89**	**0.87**	**0.77**	**0.75**	**0.74**	**0.67**	0.49	0.02	**0.79**	**0.78**	0.55	0.04
	Lactose	**0.98**	**0.98**	**0.98**	**0.96**	**0.93**	0.54	0.21	0.02	0.02	**0.96**	**0.96**	**0.96**	**0.92**
	Acetaminophen	**1.00**	**0.95**	**0.92**	**0.91**	**0.90**	**0.80**	0.38	0.08	0.06	**0.92**	**0.92**	0.07	0.06
	Sodium bicarbonate	**0.98**	**0.98**	**0.98**	**0.98**	**0.98**	**0.97**	**0.97**	**0.97**	**0.97**	**0.96**	**0.96**	**0.96**	**0.95**
	Sorbitol	**0.99**	**0.98**	**0.98**	**0.97**	**0.97**	0.19	0.19	0.03	0.02	**0.91**	**0.91**	**0.60**	0.41

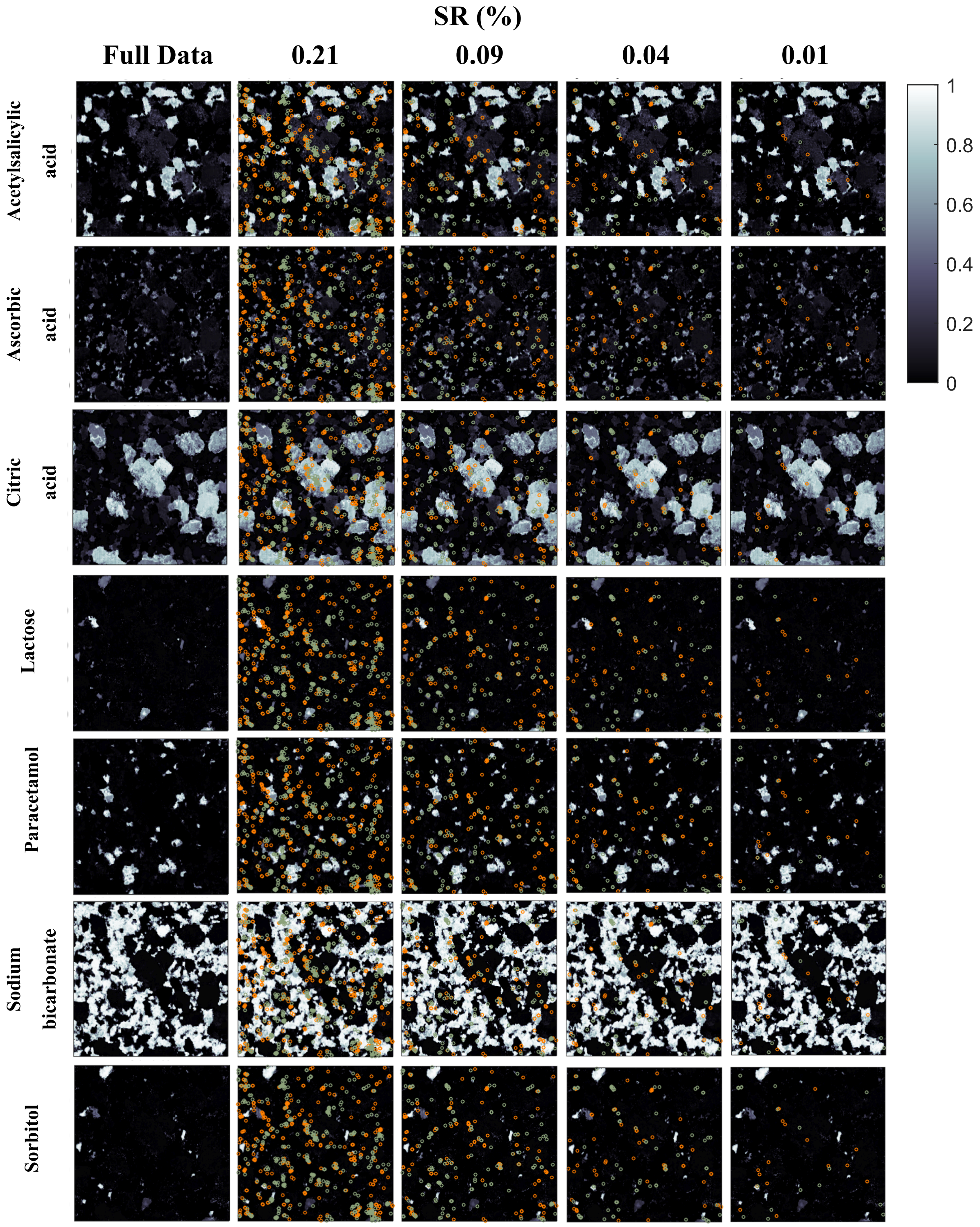


Fig. 3. Database matching results obtained for the full dataset $\mathbf{D}_1$ and for the reduced datasets (corresponding to different SR values) yielded by ES-PCA and ES-DFT. The scale bar reflects the values of the Pearson's CC estimated (%). Orange (green, resp.) circles mark the position of the ES identified by ES-PCA (ES-DFT, resp.). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

for the same SR. The most notable difference between these approaches lies in how they handle the spectral variation in the 1200–1400 cm^{-1} range. As mentioned previously, this region, attributed to sodium bicarbonate, contributes significantly to the overall data variance, as shown in Fig. 2 (column A).

Given the partially known composition of the commercial sample investigated, we performed spectral matching against the target spectra of individual compounds supposed to be present in the tablet and retrieved from an in-house database (see Fig. 1). Table 2 contains the results for the full dataset and the reduced datasets composed of the ES identified by ES-PCA and ES-DFT, respectively, for various SR values. As expected, the most comprehensive analysis – matching all the spectra of the full dataset – yields the best outcomes, with Pearson's correlation coefficient (CC) values exceeding 0.97 for the 7 target spectra. These

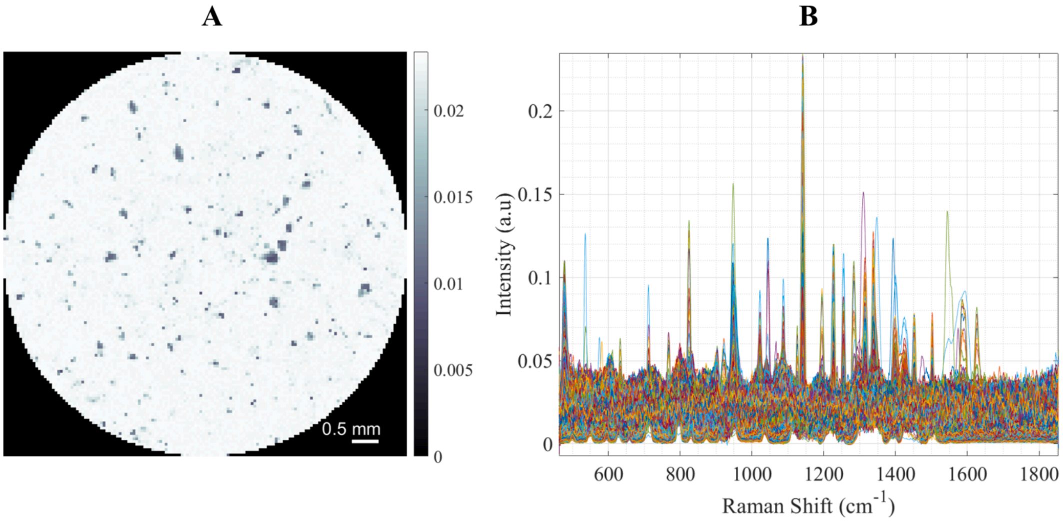


Fig. 4. Raman imaging of falsified chloroquine. A) Raman image (averaged across spectral channels); B) baseline-corrected Raman spectra.

high CC values indicate the presence of spectral pixels that closely match the databased spectra.

In addition, Fig. 3 provides images showing the spatial distribution of the CC values obtained from the processing of the full dataset for all the 7 target compounds. As previously noted, sodium bicarbonate appears to be a major component of the formulation, while sorbitol and lactose are present at lower concentrations. However, these results were obtained through an exhaustive spectral matching analysis, which required approximately 45 min of computation time on a standard computer.

The computation time can be significantly reduced by performing spectral matching with ES rather than with the full spectral dataset. Table 2 shows the results obtained by performing spectral matching on the various reduced datasets corresponding to different SR values. Both ES-PCA and ES-DFT are capable of retaining most of the collected spectral information, and, overall, ES-based spectral matching decreases processing time by several orders of magnitude compared to the aforementioned exhaustive analysis.

We begin our discussion with the results obtained at a SR of 0.42 %, for which ES-PCA returned 922 ES and provides outcomes nearly as accurate as those obtained from the full dataset analysis, enabling the clear identification of the 7 target compounds. The calculation time is here significantly reduced to approximately 2 min. We provide in Fig. S1 of the Supporting Information a representation of the normalized scores, highlighting the ones that were identified as corresponding to ES and, among them, those that correlate the most to the 7 databased spectra.

For ES-DFT, considering all 105 computed harmonics, 432 ES were found, which corresponds to a SR close to 0.21 %. Also here, the obtained CC values enable the clear identification of 7 target compounds. It should, however, be noticed that these values are significantly lower than those retrieved through the full dataset (see the following for explanation). In contrast, performing spectral matching using the 431 ES identified by ES-PCA at this same SR level does not permit to clearly identify sorbitol (CC = 0.19) and, to a lesser extent, lactose (CC = 0.54). This suggests that the data variance induced by sorbitol is minimal and not adequately captured by the reduced PCA model.

At a SR of 0.09 %, the number of ES decreases significantly for both approaches: 174 for ES-PCA and 172 for ES-DFT. At this SR, the two methods show clearly contrasted performance. While ES-DFT preserves its capability to identify all 7 target compounds from the reduced dataset, ES-PCA fails to clearly identify not only sorbitol and lactose, but also acetaminophen and ascorbic acid which are active ingredients. We provide in Fig. S2 of the Supporting Information a representation of the data point clouds of the 25 harmonics considered for the retrieval of ES at a SR of 0.09 %. This data reduction level is the best that can be achieved while still guaranteeing a full elucidation of the tablet composition. At a SR of 0.04 %, only 2 out of 7 compounds are identified, which can be explained in the light of the dramatic decrease of the energy percentage that drops down to 63.4 %.

The previous comparative study allowed us to illustrate the complementarity of ES-PCA and ES-DFT for ES selection in a practical case. Given its distinctive operational principles and the use of DFT, ES-DFT permits to reach a better SR while ensuring the identification of ES profiles that scarcely contribute to the overall data variance and that cannot be captured by ES-PCA. However, one may notice that at a SR of 0.21 %, when ES-DFT is concerned, the CC values returned for the 7 tablet ingredients are not as high as the ones obtained when considering the full dataset, despite the fact that all 105 harmonics are taken into account in the calculations. This can be explained by the fact that, in ES-DFT, convex hull calculations are performed sequentially on successive two-dimensional Fourier coordinate phasors [25]. As raised by reviewers, this is efficient but not optimal. An alternative could be to perform convex hull calculations on a multidimensional matrix obtained fusing the real and imaginary parts of the Fourier coefficients estimated for multiple harmonics. We provide in Fig. S3 of the Supporting Information the results obtained through a similar selection conducted on the coefficients of the three most energetic harmonics computed for dataset **D_1**, which yielded a total amount of 1251 ES and CC values of 0.87 for acetylsalicylic acid, 0.96 for ascorbic acid, 0.80 for citric acid, 0.98 for lactose, 0.92 for paracetamol, 0.97 for sodium bicarbonate and 0.98 for sorbitol. The more accurate characterization of the sample composition, though, came at the cost of a significantly longer processing time that can easily become prohibitive if more than 4 or 5 harmonic frequencies are simultaneously handled. Future studies will focus on a rigorous and

Table 3
Results obtained by applying both ES-PCA and ES-DFT to dataset D_2 for different SR (%).

ES-PCA				ES-DFT		
# PCs	Explained variance (%)	# ES	SR (%)	# harmonics	Energy (%)	# ES
11	97.8	2829	**12.57**	307	99.5	2830
10	97.7	1828	**8.12**	218	99.3	1822
9	97.6	1256	**5.58**	172	98.9	1254
8	97.5	780	**3.47**	133	98.2	779
7	97.3	406	**1.80**	96	95.6	405
6	97.2	210	**0.93**	66	90.5	210
5	97.0	95	**0.42**	26	69.1	94
4	96.8	32	**0.14**	5	29.4	33
3	96.1	11	**0.05**	1	11.3	10

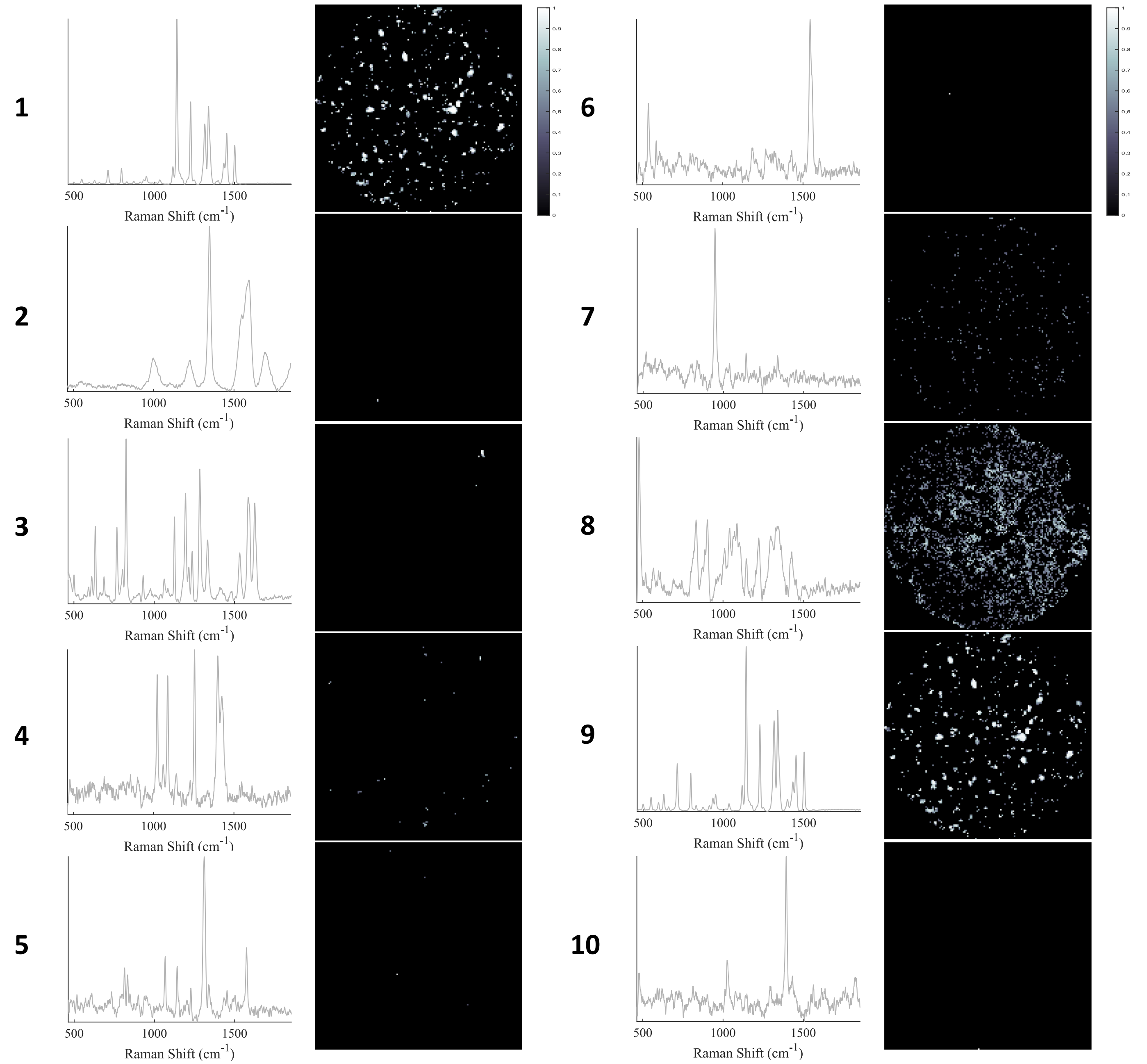


Fig. 5. Results obtained from the SIMPLISMA analysis of the full dataset $\mathbf{D}_2$. The scale bar reflects the values of the Pearson's CC estimated (%).

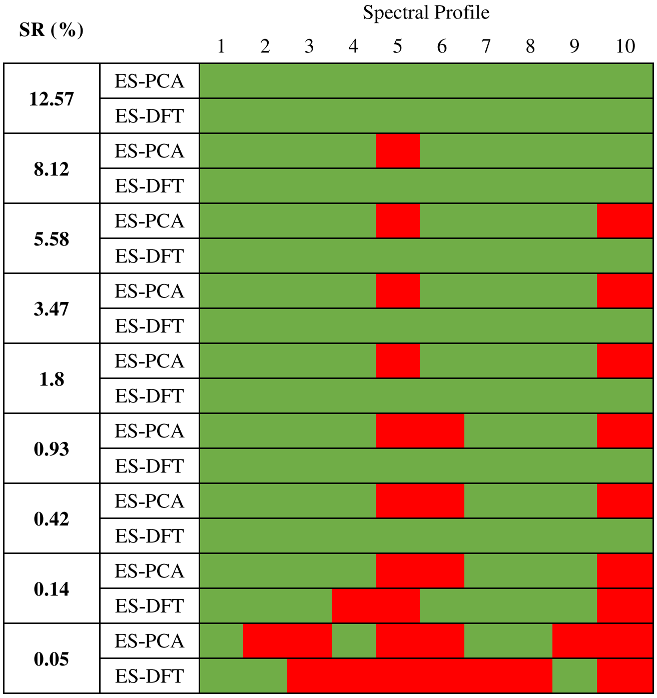
comprehensive comparison of these two variants of ES-DFT.

Through the following case-study, we will now extend our investigation to a scenario where a sample of fully unknown composition and higher complexity is coped with. Fig. 4 illustrates the results of Raman hyperspectral imaging conducted on a falsified chloroquine specimen, which was confiscated during the COVID-19 pandemic. The signal-to-noise ratio of the Raman spectra in Fig. 4B is evidently low. Indeed, the excitation laser power had to be reduced down to 10 % of the standard one to cope with the difficulties usually encountered when analyzing falsified medications, which commonly contain light-sensitive materials and are prone to degradation or burning during acquisition.

Table 3 reports the results obtained by applying both ES-PCA and ES-DFT to dataset $\mathbf{D}_2$. For ES-PCA, we investigated various situations starting from an eleven-component PCA model explaining 97.8 % of the data variance down to a three-component PCA model accounting for 96.1 % of the data variance. In this case, the number of ES ranged from 2829 (SR = 12.6 %) and 11 (SR = 0.05 %). The results obtained by applying ES-DFT considering the number of harmonics needed to reach similar SR values are also provided. As observed for the analysis of dataset $\mathbf{D}_1$, the variance explained by the PCA models remains very high even when considering only a few PCs whereas, for ES-DFT, the energy value starts clearly decreasing below a SR of 0.93 % which would correspond to 66 harmonic frequencies and to the identification of 210 ES.

Here, given the unknown composition of the analyzed sample, we applied SIMPLe-to-use Interactive Self-modeling Mixture Analysis (SIMPLISMA) [35] to the full dataset in order to extract its 10 most selective (purest) spectra. These spectra were then correlated to each individual spectrum of the full dataset which resulted in the CC maps represented in Fig. 5.

Components n. 1, 7, 8 and 9 seem to be distributed all over the sample. On the other hand, components n. 2, 5, 6 and 10 appear only



SR (%)		Spectral Profile									
		1	2	3	4	5	6	7	8	9	10
12.57	ES-PCA	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
	ES-DFT	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
8.12	ES-PCA	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩
	ES-DFT	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
5.58	ES-PCA	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟥
	ES-DFT	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
3.47	ES-PCA	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟥
	ES-DFT	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
1.8	ES-PCA	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟥
	ES-DFT	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
0.93	ES-PCA	🟩	🟩	🟩	🟩	🟥	🟥	🟩	🟩	🟩	🟥
	ES-DFT	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
0.42	ES-PCA	🟩	🟩	🟩	🟩	🟥	🟥	🟩	🟩	🟩	🟥
	ES-DFT	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
0.14	ES-PCA	🟩	🟩	🟩	🟩	🟥	🟥	🟩	🟩	🟩	🟥
	ES-DFT	🟩	🟩	🟩	🟥	🟥	🟩	🟩	🟩	🟩	🟥
0.05	ES-PCA	🟩	🟥	🟥	🟩	🟥	🟥	🟩	🟩	🟥	🟥
	ES-DFT	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟥

Fig. 6. Schematic representation of the outcomes obtained by applying SIMPLISMA to the reduced datasets resulting from the ES-PCA and ES-DFT analysis of $\mathbf{D}_2$ for various SR values (%). Green color indicates the correct identification of an individual component that was also detected in the full dataset. Red color denotes a component that was not successfully identified. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

localized at few (sometimes single) pixel locations. Few- or single-pixel components are generally encountered in low quality medicines, produced in uncontrolled environments and for which chemical contaminations are frequent. It should be noted that even trace amounts of contaminating active ingredients in falsified pharmaceutical tablets may represent a threat to the user's health. In addition, the corresponding spectra constitute forensic signatures that may be used for legal assessments and decision-making. These contaminants are generally detected by destructive techniques such as LC-MS [27], which however provide only partial information and destroy the sample. In contrast, Raman chemical imaging enables not only the detection of trace contaminants but also the determination of their spatial distribution over the sample.

We tested the same methodology on the datasets derived from the application of ES-PCA and ES-DFT to $\mathbf{D}_2$ (see Table 3). Fig. 6 provides a representation of the results obtained. Both ES-PCA and ES-DFT can be used to effectively reduce the full dataset while preserving the ability to conduct reliable spectral identification. However, ES-PCA requires a total number of 2829 ES for achieving the same degree of accuracy guaranteed by the full dataset. At lower SR values, ES-PCA's ability to ensure single-pixel component identification diminishes. This degradation is first observed for component n. 5 at SR equal to 8.12 %, then for component n. 10 at a SR equal to 5.58 %, and continues progressively. In contrast, ES-DFT yields satisfactory performance even at a SR of 0.42 %, which corresponds to the extraction of Fourier coefficients for 26 harmonic frequencies and to the selection of 94 ES only.

As pointed out before, the significant superiority of ES-DFT for data reduction and for minor component identification can be attributed to the fact that Fourier coordinates are calculated in a predetermined coordinate system, independent of the data themselves, contrarily to PCA that derives its basis vectors from the inherent variance of the investigated spectral data.

4. Concluding remarks

In this article, we demonstrate the use and efficacy of two alternative methods based on the selection of ES for analyzing pharmaceutical samples by FTIR and Raman hyperspectral imaging. The examples presented here encompass very large datasets and are representative of challenging analytical scenarios, *e.g.* the identification of very minor compounds and the presence of strong scattering components that can mask the spectral information of interest.

ES can be retrieved by convex hull analysis of the investigated data point cloud represented in an intrinsic coordinate system, using either normalized scores resulting from a truncated singular values decomposition or Fourier coefficients at selected harmonic frequencies. Our results demonstrate the two main advantages of analyzing reduced datasets composed of ES instead of full datasets: i) computational time reduced by several orders of magnitude and ii) improved outcome accuracy. Overall, these advantages enable a much faster determination of the composition of the analyzed drugs. Additionally, it was shown that using Fourier coefficients can allow a better assessment of drugs exhibiting very minor components.

CRediT authorship contribution statement

Laureen Coic: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yesid Roman Gomez:** Writing – review & editing, Validation, Investigation. **Pierre-Yves Sacré:** Writing – review & editing, Validation, Investigation. **Eric Ziemons:** Writing – review & editing, Supervision, Investigation. **Raffaele Vitale:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Cyril Ruckebusch:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

All authors acknowledge financial support from the ANR-21-CE29-0007 project (Agence Nationale de la Recherche).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.microc.2025.112751.

Data availability

Data will be made available on request.

References

[1] D. Cebeci, B.R. Mankani, D. Ben-Amotz, Recent Trends in Compressive Raman Spectroscopy Using DMD-Based Binary Detection, J. Imaging 2019, Vol. 5, Page 1. 5 (2018) 1. https://doi.org/10.3390/JIMAGING5010001.

[2] D. Cialla-May, C. Krafft, P. Rösch, T. Deckert-Gaudig, T. Frosch, I.J. Jahn, S. Pahlow, C. Stiebing, T. Meyer-Zedler, T. Bocklitz, I. Schie, V. Deckert, J. Popp, Raman Spectroscopy and Imaging in Bioanalytics, Anal. Chem. 94 (2022) 86–119, https://doi.org/10.1021/ACS.ANALCHEM.1C03235/ASSET/IMAGES/LARGE/AC1C03235_0010.JPEG.

[3] M. Kubik, Chapter 5 Hyperspectral Imaging: A New Technique for the Non-Invasive Study of Artworks, Phys. Tech. Study Art, Archaeol. Cult. Herit. 2 (2007) 199–259. https://doi.org/10.1016/S1871-1731(07)80007-8.

[4] W.H. Su, D.W. Sun, Fourier Transform Infrared and Raman and Hyperspectral Imaging Techniques for Quality Determinations of Powdery Foods: A Review, Compr. Rev. Food Sci. Food Saf. 17 (2018) 104–122, https://doi.org/10.1111/1541-4337.12314.

[5] M.G. Lizio, R. Boitor, I. Notingher, Selective-sampling Raman imaging techniques for: Ex vivo assessment of surgical margins in cancer surgery, Analyst 146 (2021) 3799–3809, https://doi.org/10.1039/D1AN00296A.

[6] R. Vitale, A. Zhyrova, J.F. Fortuna, O.E. de Noord, A. Ferrer, H. Martens, On-The-Fly Processing of continuous high-dimensional data streams, Chemom. Intell. Lab. Syst. 161 (2017) 118–129, https://doi.org/10.1016/J.CHEMOLAB.2016.11.003.

[7] J. Wu, M. Wu, H. Li, L. Li, L. Li, A Serverless-Based, On-the-Fly Computing Framework for Remote Sensing Image Collection, Remote Sens. 2022, Vol. 14, Page 1728. 14 (2022) 1728. https://doi.org/10.3390/RS14071728.

[8] M. Vidal, J.M. Amigo, Pre-processing of hyperspectral images. Essential steps before image analysis, Chemom. Intell. Lab. Syst. 117 (2012) 138–148, https://doi.org/10.1016/J.CHEMOLAB.2012.05.009.

[9] J. Burger, A. Gowen, Data handling in hyperspectral image analysis, Chemom. Intell. Lab. Syst. 108 (2011) 13–22, https://doi.org/10.1016/J.CHEMOLAB.2011.04.001.

[10] A. de Juan, R. Tauler, Multivariate Curve Resolution: 50 years addressing the mixture analysis problem – A review, Anal. Chim. Acta. 1145 (2021) 59–78, https://doi.org/10.1016/J.ACA.2020.10.051.

[11] L. Coic, P.Y. Sacré, A. Dispas, C. De Bleye, M. Fillet, C. Ruckebusch, P. Hubert, E. Ziemons, Pixel-based Raman hyperspectral identification of complex pharmaceutical formulations, Anal. Chim. Acta. 1155 (2021) 338361, https://doi.org/10.1016/j.aca.2021.338361.

[12] B. Fei, Hyperspectral imaging in medical applications, Data Handl, Sci. Technol. 32 (2019) 523–565, https://doi.org/10.1016/B978-0-444-63977-6.00021-3.

[13] L. Coic, P.Y. Sacré, A. Dispas, C. De Bleye, M. Fillet, C. Ruckebusch, P. Hubert, É. Ziemons, Selection of essential spectra to improve the multivariate curve resolution of minor compounds in complex pharmaceutical formulations, Anal. Chim. Acta. 1198 (2022), https://doi.org/10.1016/J.ACA.2022.339532.

[14] M. Ghaffari, N. Omidikia, C. Ruckebusch, Essential Spectral Pixels for Multivariate Curve Resolution of Chemical Images, Anal. Chem. 91 (2019) 10943–10948, https://doi.org/10.1021/acs.analchem.9b02890.

[15] M. Sawall, C. Kubis, H.g. Schröder, K. Neymeyr, Multivariate curve resolution methods and the design of experiments, J. Chemom. 34 (2019) e3159.

[16] C. Ruckebusch, R. Vitale, M. Ghaffari, S. Hugelier, N. Omidikia, Perspective on essential information in multivariate curve resolution, TrAC - Trends Anal. Chem. 132 (2020), https://doi.org/10.1016/j.trac.2020.116044.

[17] M. Ghaffari, N. Omidikia, C. Ruckebusch, Joint selection of essential pixels and essential variables across hyperspectral images, Anal. Chim. Acta. 1141 (2021) 36–46, https://doi.org/10.1016/J.ACA.2020.10.040.

[18] A. Olarini, M. Cocchi, L. Duponchel, C. Ruckebusch, Exploratory analysis of hyperspectral imaging data, Chemom. Intell. Lab. Syst. 252 (2024) 105174, https://doi.org/10.1016/j.chemolab.2024.105174.

[19] M. Sawall, C. Ruckebusch, M. Beese, R. Francke, A. Prudlik, K. Neymeyr, An active constraint approach to identify essential spectral information in noisy data, Anal. Chim. Acta 1233 (2022) 340448, https://doi.org/10.1016/j.aca.2022.340448.

[20] B.V. Grande, R. Manne, Use of convexity for finding pure variables in two-way data from mixtures, Chemom. Intell. Lab. Syst. 50 (2000) 19–33, https://doi.org/10.1016/S0169-7439(99)00041-6.

[21] R. Vitale, C. Ruckebusch, On a black hole effect in bilinear curve resolution based on least squares, J. Chemometr. 37 (2023) e3442.

[22] M. Ahmad, R. Vitale, C. Ruckebusch, Weighted multivariate curve resolution – alternating least squares based on sample relevance, J. Chemometr. 37 (2023) e3453. https://doi:10.1002/cem.3453.

[23] X.D. Qing, G.Y. Lu, X.H. Zang, Q.L. Chen, X.H. Zou, W. He, L. Xu, J. Zhang, Essential spectral pixel-based improvement of UMAP classifying data to idntify minor compounds in food matrix, Talanta 273 (2024) 125845. https://www.sciencedirect.com/science/article/pii/S0039914024002248.

[24] M. Ghaffari, M.C.J. Lukkien, N. Omidikia, G.H. Tinnevelt, M.C.P. van Eijk, S. Podchezertsev, J.J. Jansen, Systematic reduction of hyperspectral images for high- throughput plastic characterization, *Sci Rep* 13 (2023) 21591, https://doi.org/10.1038/s41598-023-49051-y.

[25] L. Coic, R. Vitale, M. Moreau, D. Rousseau, J.H. de M. Goulart, N. Dobigeon, C. Ruckebusch, Assessment of essential information in the Fourier domain to accelerate Raman hyperspectral microimaging, Anal. Chem. 95 (2023) 15497–15504. https://pubs.acs.org/doi/10.1021/acs.analchem.3c01383.

[26] V. Gilet, G. Mabilleau, M. Loumaigne, L. Coic, R. Vitale, T. Oberlin, J. Henrique de Morais Goulart, N. Dobigeon, C. Ruckebuch, D. Rousseau, Superpixels meet essential spectra for fast Raman hyperspectral imaging, Opt. Express 32 (2024) 932–948, https://doi.org/10.1364/OE.509736.

[27] H. Kouakou, J. Henrique de Morais Goulart, R. Vitale, T. Oberlin, D. Rousseau, C. Ruckebuch, N. Dobigeon, On-the-fly unmixing based on Kalman filtering, Chemom. Intell. Lab. Syst (2024).

[28] R. Vitale, A. Azizi, M. Ghaffari, N. Omidikia, C. Ruckebusch, Three-way data reduction based on essential information, J. Chemometr. (2024) e3617.

[29] A. Budak, Phasor Transformation, IEEE Trans. Educ. E-10 (1967) 48–49, https://doi.org/10.1109/TE.1967.4320215.

[30] M.A. Digman, V.R. Caiolfa, M. Zamai, E. Gratton, The phasor approach to fluorescence lifetime imaging analysis, Biophys. J. 94 (2008) L14–L16, https://doi.org/10.1529/biophysj.107.120154.

[31] C.A. Waffo Tchounga, P.Y. Sacre, P. Ciza, R. Ngono, E. Ziemons, P. Hubert, R. D. Marini, Composition analysis of falsified chloroquine phosphate samples seized during the COVID-19 pandemic, J. Pharm. Biomed. Anal. 194 (2021) 113761, https://doi.org/10.1016/j.jpba.2020.113761.

[32] J. Peng, S. Peng, A. Jiang, J. Wei, C. Li, J. Tan, Asymmetric least squares for multiple spectra baseline correction, Anal. Chim. Acta. 683 (2010) 63–68, https://doi.org/10.1016/J.ACA.2010.08.033.

[33] https://dataverse.uliege.be/dataset.xhtml?persistentId=doi:10.58119/ULG/QVOWTA, accessed December 3, 2024.

[34] R. Rajko, Studies on the adaptability of different Borgen norms applied in self-modeling curve resolution (SMCR) method, J. Chemometr. 37 (2009) 265–274. https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/cem.1221.

[35] W. Winding, D.A. Stephenson, Anal. Chem. 64 (1992) 2735.