Published in: Analytica Chimica Acta (20214), vol. 1155, n°April, 338361
DOI: 10.1016/j.aca.2021.338361
Status : Postprint (Author's version)

Pixel-based Raman hyperspectral identification of complex pharmaceutical formulations

Laureen Coic [a,*1], Pierre-Yves Sacre [a], Amandine Dispas [a,b], Charlotte De Bleye [a], Marianne Fillet [b], Cyril Ruckebusch [c], Philippe Hubert [a], Eric Ziemons [a]

[a] *University of Liege (ULiege), CIRM, Vibra-Sante Hub, Laboratory of Pharmaceutical Analytical Chemistry, Avenue Hippocrate 15, 4000, Liege, Belgium*

[b] *University of Liege (ULiege), CIRM, MaS-Sante Hub, Laboratory for the Analysis of Medicines, Avenue Hippocrate 15, 4000, Liege, Belgium*

[c] *University of Lille, CNRS, UMR 8516 LAboratoire de Spectroscopie pour les Interactions, la Reactivite et l'Environnement (LASIRE), F-59000, Lille, France*

KEYWORDS : Hyperspectral imaging; Spectral identification: Pixel selection: Essential spectral pixels; Falsified medicines

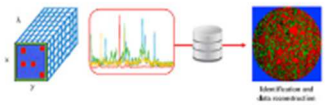
ABSTRACT

Hyperspectral imaging has been widely used for different kinds of applications and many chemometric tools have been developed to help identifying chemical compounds. However, most of those tools rely on factorial decomposition techniques that can be challenging for large data sets and/or in the presence of minor compounds. The present study proposes a pixel-based identification (PBI) approach that allows readily identifying spectral signatures in Raman hyperspectral imaging data. This strategy is based on the identification of essential spectral pixels (ESP), which can be found by convex hull calculation. As the corresponding set of spectra is largely reduced and encompasses the purest spectral signatures, direct database matching and identification can be reliably and rapidly performed. The efficiency of PBI was evaluated on both known and unknown samples,

1 * Corresponding author

considering genuine and falsified pharmaceutical tablets. We showed that it is possible to analyze a wide variety of pharmaceutical formulations of increasing complexity (from 5 to 0.1% (w/w) of polymorphic impurity detection) for medium (150 x 150 pixels) and big (1000 x 1000 pixels) map sizes in less than 2 min. Moreover, in the case of falsified medicines, it is demonstrated that the proposed approach allows the identification of all compounds, found in very different proportions and, sometimes, in trace amounts. Furthermore, the relevant spectral signatures for which no match is found in the reference database can be identified at a later stage and the nature of the corresponding compounds further investigated. Overall, the provided results show that Raman hyperspectral imaging combined with PBI enables rapid and reliable spectral identification of complex pharmaceutical formulations.

GRAPHICAL ABSTRACT



HIGHLIGHTS

- The pixel-based approach allowed to detect low-dose compound in less than 2 min.
- The elucidation of 1,000,000 spectra was possible with only 0.1% of the data.
- The ESP approach allowed to keep all relevant spectral information.

1. Introduction

Hyperspectral chemical imaging techniques are now widely approved as efficient analytical tools providing meaningful highquality data relatively rapidly. Depending on the spectroscopy used, both organic and inorganic compounds can be analyzed, opening a large range of applications [1-5]. Combined with appropriate data analysis, both chemical and spatial information is provided.

In the pharmaceutical field, analysis of tablets by Raman hyperspectral imaging is widely used for quality control purposes and has now been included in the general chapters of the European Pharmacopeia [6]. Indeed, it can be used for many applications as for the detection of impurities such as polymorphic forms [7], the evaluation of tablet homogeneity [8-10] or to analyze falsified medicines [11-13]. However, data analysis can be a challenge because pharmaceutical tablets are commonly or generally wide objects to observe, providing a huge amount of data to analyze.

Furthermore, the useful Raman information can be hindered by the fluorescence of some compounds (e.g. cellulose derivatives). In case of spectral unmixing requirement, the most appropriated algorithm to use is the multivariate curve resolution - alternating least squares (MCR-ALS) [14-16]. It is based on the assumption that the signal can be described as a weighted sum of the pure spectra. Nevertheless, this strategy requires a lot of user input, is very time consuming and hardly applicable for the analysis of large data. Indeed, it is well known that factorial decomposition methods can be difficult to apply on big data matrices, as well as in the presence of many constituents, and results may suffer from rotational ambiguity [17-19]. In case of falsified medicines, the use of sequential MCR-ALS [14,16,22] allowed to elucidate the composition of falsified medicines [11,23]. However, it requires a lot of computational time and, because of the rotational ambiguity, some spectra are hardly resolved.

Moreover, even when the composition is known, the MCR resolution can also be challenging. First of all, some low-variance sources can be diluted in the process of unmixing and can hardly be resolved without information on the expected sample composition [20,21]. Some minor compounds can also be present in a few pixels and missed in the MCR process. Finally, chemicals are sometimes mixed so intimately that the mathematical unmixing of their spectra is impossible and should be considered as a single MCR component.

However, the sample composition is generally known or partially known in the pharmaceutical field, and spectral identification of the raw materials can be attempted readily. Some studies have proposed strategies for database matching purpose using the correlation coefficient or some adaptations [24]. Others have used decision trees [25] or deep learning [26]. In all cases, many parameters have to be optimised and results can be rather different depending on the nature of the data and the pre-processing. For these reasons, the most commonly used database matching tool remains the spectral correlation coefficient [27] because it provides a good compromise between implementation, easiness and reliability of the results.

Regarding the compound identification for samples having a complex composition, a new strategy must be worked out to bypass the limitation associated to approaches based on factorial decomposition of the data matrix. One way can be to step back to the analysis of individual pixel. This, if it could be performed somehow exhaustively, i.e. for all pixels, would consist of the ultimate and most efficient method for database matching [28].

However, this is not always possible in practice because of the number of pixels measured. One has thus to rely on pixel selection and the analysis of a reduced data set of pixels. To select subsets of pixels, different approaches that rely on different priors exist. The easiest way is random selection (RS) of a predetermined proportion of pixels, but of course, the risk is to miss information, in particular e.g. information that would be only observed at a few pixels. Others would select pixels with a criterion in mind as, for instance, the Kennard-Stone (KS) [29] algorithm that aims to select a subset of samples which provides uniform coverage over the data set. Recently, it was also proposed to select Essential Spectral Pixels (ESP) based on archetype analysis. This algorithm allows getting free from the pre-stated issues by reducing drastically the number of pixels while preserving the most linearly dissimilar spectral information [30,31].

In this context, the objective of the present study is to develop a pixel-based identification (PBI) strategy that relies on the identification of ESP to elucidate chemical composition of Raman hyperspectral images of complex pharmaceutical formulations. We will first apply PBI on multi-component samples of different map size in presence of impurity at different concentrations in order to mimic samples having minor compounds. We will then evaluate the performance of our approach by the composition elucidation of falsified chloroquine tablets seized during the COVID-19 pandemic [23]. For the sake of comparison, the results obtained with RS and KS pixel selection will be provided. The last point will focus on the usefulness of the PBI approach to detect all relevant spectral signatures, including the ones that are not referenced in the spectral database.

2. Materials and methods

2.1. Samples

2.1.1. Raw materials

2.1.1.1. Carbamazepine polymorphs.

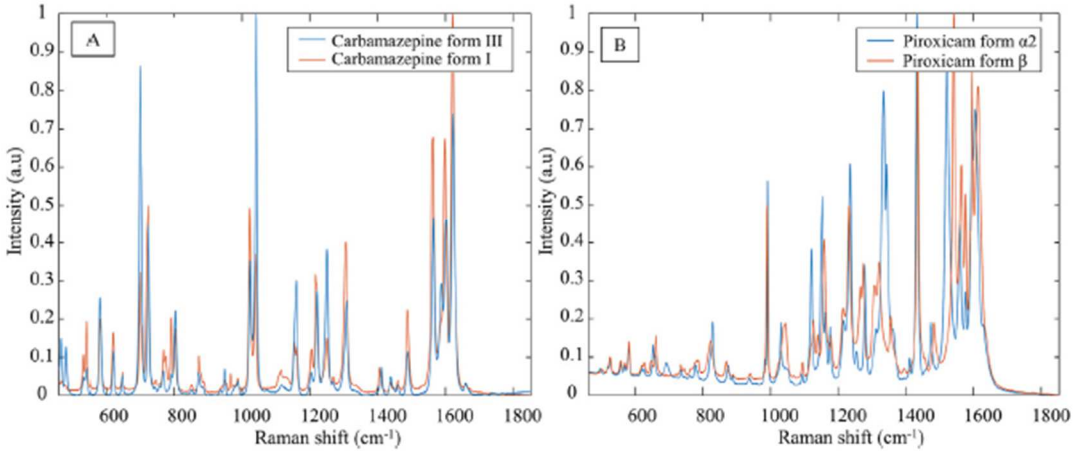
Carbamazepine formulations were made of carbamazepine form III (TCI, Belgium), carbamazepine form I and a fixed proportion of excipients (see Table S1): microcrystalline cellulose (Sigma-Aldrich, Belgium) (MCC), mannitol (Sigma-Aldrich, Belgium), sodium croscarmellose (Fagron, Belgium), aerosil (Certa, Belgium) and magnesium stearate (Certa, Belgium). The carbamazepine form I (triclinic) was obtained following the protocol described in Ref. [32]: carbamazepine form III is dehydrated during 3h at

65 C and then heated during 40 min at 175 C. The raw spectra of the two carbamazepine forms are presented in Fig. 1-A. They were acquired with the same Raman acquisition parameters than the ones used for the implementation of the in-house database (c.f. 2.2.1), which explains the nice resolved and smooth spectra. The characteristic peaks were compared with the literature [33]. Two levels of polymorphic contamination were realized: 1 and 5% (w/w). The mixtures were homogenized in a mortar with a pestle. Powders were then pressed with a manual hydraulic press (Specac, United Kingdom) using 12 mm dies, with 5-ton compression to obtain 200 mg tablets.

2.1.1.2. Piroxicam polymorphs. Piroxicam formulations were made of piroxicam form β (TCI, Belgium) and piroxicam form α2. A fixed proportion of lactose (Fagron, Belgium) was added to the mixture. Piroxicam form α2 was obtained following the protocol described in Ref. [7], by dissolving 1.5 g of piroxicam form β in 380 mL of ethanol at 60 C until complete dissolution. The solution was left for cooling overnight and filtered using Büchner apparatus. The powder obtained was dried under vacuum for one hour. The raw spectra of the two piroxicam forms are shown in Fig.1-B. The characteristic peaks were compared with a reference to check the effectiveness of the transformation [7]. Five levels of concentration of polymorph impurity were obtained (0.1, 0.5, 1, 1.5 and 2% (w/w)). The mixtures were homogenized in a mortar with a pestle. Powders were then pressed with a manual hydraulic press (Specac, United Kingdom) using 12 mm dies, with 5-ton compression to obtain 200 mg tablets.

2.1.1.3. Falsified medicines. The falsified chloroquine samples were seized by local authorities during the COVID-19 pandemic. They are described in details in the study [23]. The analyzed data are Raman hyperspectral images obtained from the authors.

***Figure 1**. Raman spectra of the raw materials. A) Spectra of the two polymorphic forms of carbamazepine. B) Spectra of the two polymorphic forms of piroxicam.*



2.2.Hyperspectral imaging techniques

2.2.1.RAMAN MICROSCOPY

Raman hyperspectral imaging analyses of the samples were performed with a Labram HR Evolution (Horiba scientific) equipped with an EMCCD detector (1600 X 200-pixel sensor) (Andor Technology Ltd.), a Leica 50x Fluotar LWD objective and a 785 nm laser with a power of 45 mW at sample (XTRA II single frequency diode laser, Toptica Photonics AG). The spectra were collected with the LabSpec 6 (Horiba Scientific) software.

For the carbamazepine samples a 300 gr/mm grating fixed at 1200 cm^{-1} (spectral range of 463-1853 cm^{-1}) was used to perform the mappings with two accumulations of 1 s. The confocal slit-hole was fixed at 200 mm. The whole tablet surface was analyzed with a 150 x 150 pixels mapping and a step size of 87 mm (total map size of ~13 X 13 mm^2). The data analysis time was 12.5 h.

For the piroxicam samples, a 600 gr/mm grating fixed at 1300 cm^{-1} (spectral range of 900-1700 cm^{-1}) was used to perform the mappings with a single acquisition of 0.05 s using the SWIFT™ mode. The confocal slit-hole was fixed at 200 mm. The middle of the tablet surface was mapped with a step size of 5.5 µm over a 5.5 X 5.5 mm^2, providing a 1000 x 1000 pixels mapping. The data analysis time was 22 h.

Regarding the falsified medicines tablets, they were glued on a microscope slide and their surface was milled using a Leica EM Rapid milling system equipped with a tungsten carbide miller (Leica Microsystems GmbH) before Raman mapping. The acquisition parameters were the same as the ones

used for the carbamazepine samples. A 150 x 150 pixels mapping (step size of ~50 µm) was done for each sample. The data analysis time was 12.5 h.

2.3. Data analysis

The aim of the study being to identify chemical composition from database matching of the spectral information corresponding to selected pixels. Different approaches were investigated as illustrated in Fig. 2. Details are provided in the following sub-sections.

2.3.1. Software and toolbox

The proposed algorithm was developed on a workstation with Intel® Core™ i7-7820X CPU @ 3.60 GHz, 8 cores with 128 Go of RAM. The Essential Spectral Pixels algorithm used is described in Ghaffari et al. [30]. For some key steps, the Matlab Parallel Computing Toolbox™ was used to improve the speed of algorithms. All computations were carried out with Matlab R2019b (The Mathworks) with the PLS Toolbox (version 8.6.2, Eigenvector Research Inc).

2.3.2. Preprocessing

The first step was to remove spikes from the original data with a one-dimensional median filter (using the medfilt1 Matlab function), with a degree 5 polynomial function. Then, a noise reduction was done by a Savitzky & Golay smoothing (polynomial order: 1, window size: 15). The baseline was finally removed by an Automatic Whittaker filter (λ 3.10^4, p = 1.10^{-5}).

For the identification of database missing spectra, the ESP were preprocessed by a Savitzky-Golay first derivative (polynomial order: 2, window size: 15) followed by a unit area normalization and mean centering.

2.3.3. Database matching

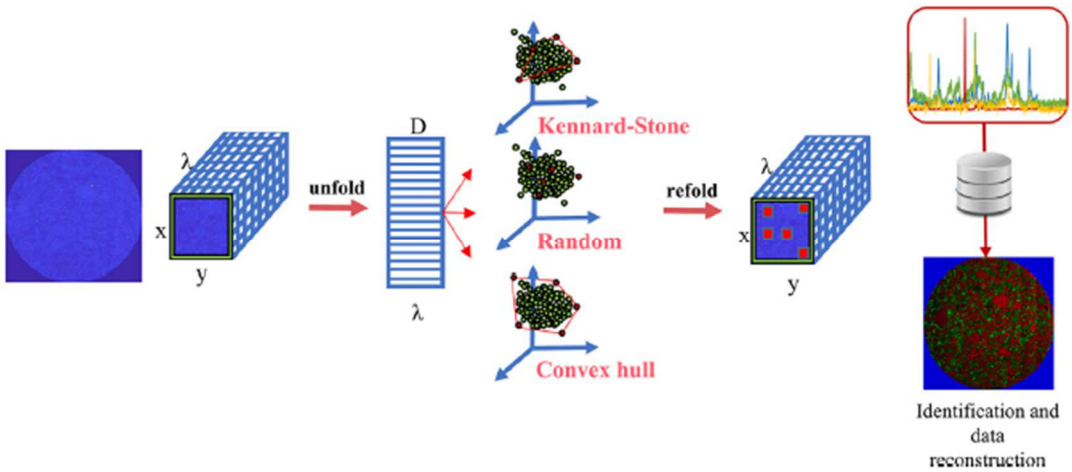
The database used is an in-house database comprising 169 Raman spectra of active pharmaceutical ingredients and excipients. Because the Raman shift scale was different for database and sample spectra, a homemade function was developed to do linear interpolation of spectral data to ensure a good correspondence between them. The database matching was performed with the correlation coefficient without supplementary preprocessing. Values inferior to 0.5 were not reported in the tables.

2.3.4. Pixels selection algorithm

2.3.4.1. Random selection. Random selection of pixels was performed to benchmark the proposed algorithms. For that purpose, κ random indices were generated n times, with κ corresponding to the number of ESP.

2.3.4.2. Kennard-Stone randomized. The KS [29] algorithm was evaluated because of its geometric properties. Indeed, the KS algorithm starts randomly choosing a point in the data cloud. It calculates the Euclidean distances (ED) between this point and its neighbors. The data point that has the highest ED is then selected. It continues iteratively until selecting the desired percentage or number of samples, allowing getting a uniform coverage over the data set. Thus, the idea was to generate κ random indices, to get random direction of data. Then, the KS was applied on this subspace, keeping 60% of the data. The selection was repeated n times in order to select κ pixels, keeping in mind that multiple selection of the same pixel should count as one. For simplicity, the algorithm will be denoted KSr in the manuscript.

Figure 2. *Workflow of the proposed strategy. The first step is to unfold the measured hyperspectral data cube. Then three different approaches were evaluated to select pixels which were subsequently matched with the in-house database. For further investigation, spectral pixels for which correlation coefficient values > 0.90 are gathered and can be used to do data-reconstruction by a least-squares approach.*



2.3.4.3. Essential spectral pixels. The ESP [31] approach enables to select the pixels that correspond to the most linearly dissimilar spectra. Geometrically, these spectra correspond to observations (points) that are found on the envelope of the

data cloud and, specifically, at its vertices, where the purest information is found. The ESPs can be obtained by calculating the convex hull of the data set in the row-space of the data set. This set of κ spectral pixels correspond to a highly reduced data set containing the most linearly relevant spectral features. We refer to Ref. [30] for more detailed explanations.

2.3.5. PIXEL-BASED ANALYSIS OF KNOWN SAMPLES

The approach taken for PBI of carbamazepine was the following. The three pixel-selection algorithms were applied on the data set corresponding to a Raman image covering the whole tablet surface and results were compared. The results were then compared to the ones obtained by matching all the measured spectra with the database.

The approach taken for the PBI of piroxicam analysis was slightly different because of the consequent data size. For the sake of comparison, a unique patch of 200 x 200 pixels was first evaluated to check the difference in results compared to the carbamazepine study, keeping the same order of magnitude data size. Then, the entire Raman image was composed of 200 x 200 pixels non-overlapping squares, resulting in 25 patches. Each pixel selection algorithm was subsequently applied for each strategy and results were compared.

2.3.6. PIXEL-BASED ANALYSIS OF UNKNOWN SAMPLE

Unknown samples corresponding to pharmaceutical tablets of complex composition were also investigated. In this case, the number of compounds was higher including potentially minor ones. The calculation was thus performed using a region of 40 x 40 pixels, in order to have the same order of magnitude as the carbamazepine study. The computation of ESP on unknown samples was done with a relatively high number of components to ensure that no spectral variability was missed. The ESP obtained for all the regions were matched with the database.

2.3.7. IDENTIFICATION OF UNKNOWN COMPOUNDS

The computation of the convex hull being dependent of the selected number of components, some ESP could potentially be redundant or noisy. Thus, each of them would not necessarily receive a significant match. Finally, because database matching is dependent of the in-house database, some spectra showing clear spectral features would remain unidentified. In order to select those spectra, an outlier detection approach was performed on the data set containing ESP that did not correspond to a significant match applying PBI (correlation coefficient < 0.5). For this purpose, principal component analysis (PCA) was applied and the samples with a high value of Q2 residuals

were identified. They correspond to those having clear specific spectral features. To check the validity of this approach, the reference spectra of three compounds were removed from the database and it was checked whether it was possible to identify these signatures using the proposed approach.

3. Results and discussion

3.1. Carbamazepine formulation

The first case was evaluated on two tablets with different proportions of carbamazepine polymorphic impurity (form I): 5% (w/ w) and 1% (w/w). The aim was to evaluate the relevance of a PBI to elucidate the entire tablet composition. Three different strategies for pixel selection were applied. The corresponding spectra were matched with the database and the significance was assessed using the correlation coefficient. For the sake of comparison, the correlation coefficient was also computed for each of the 22,500 spectra of the entire map ("Full" analysis). As one can see in Table 1, the results obtained with the pixels selected by ESP and with the "Full" approach are very close for all compounds of the Level 1. For the Level 2, the RS was not able to find the CBZ (form III) meanwhile both ESP and KSr strategy elucidate correctly the composition.

For both level of concentration, the results obtained with the RS and KSr pixel selection methods had worst correlation coefficient values. It can be explained by the inherent geometric properties of the convex hull, which will keep the edges of the data cloud in a PCA subspace, allowing having the most linearly dissimilar spectra. On the contrary, the KSr approach will select the most different spectra in the initial data space ignoring spectral redundancy and spectral correlation, inducing less relevant results. The results were obtained in 50 s, using only 8% of the initial data for each level, which is much faster than the database matching with all spectra (3 min in this case) and requires less computing memory.

Table 1 *- Comparison of the database matching results of the carbamazepine impurity detection. All results are correlation coefficient values.*

	Level 1 (5% (w/w))				Level 2 (1% (w/w))			
	ESP	KSr	Random	Full	ESP	KSr	Random	Full
CBZ (form I)	0.95	0.95	0.93	0.95	0.95	0.95	0.92	0.99
CBZ (form III)	0.90	0.88	NF	0.91	0.91	0.89	NF	0.95
Mannitol	0.92	0.88	0.86	0.94	0.94	0.93	0.92	0.94
MCC	0.86	0.76	0.81	0.89	0.9	0.86	0.82	0.96
MgSt	NF	NF	NF	0.58	NF	NF	NF	NF

Abbreviations: CBZ: carbamazepine; MCC: microcrystalline cellulose; MgSt: magnesium stearate; NF: not found.

3.2.Application to pixel selection for piroxicam beta impurity detection

After evaluating the different pixel selection strategies on medium mapping size, the second study was focused on the evaluation of the presented strategies for large amount of data (1000 x 1000 pixels maps). The analysis was performed on five tablets contaminated with different amounts of piroxicam β impurity. The major challenge of this study was to detect the impurity at low levels ranging from 2 to 0.1% (w/w). Because the map size was big, working with the whole map was not possible. That is why a region of 200 x 200 pixels was randomly chosen (patch) and compared to 25 patches of 200 x 200 pixels (grid). The results are provided in Table 2.

Table 2 *- Comparison of the piroxicam impurity detection for high dimensionality data for the three different pixel-selection strategies. Each of PBI were tested by using one patch and the entire grid.*

	Compound	KSr		Random		ESP	
		Patch	Grid	Patch	Grid	Patch	Grid
Impurity	Lactose	NF	NF	NF	0.60	0.60	0.71
2% (w/w)	Piroxicam α_2	0.96	0.98	0.96	0.95	0.98	0.98
	Piroxicam β	0.98	0.98	0.98	0.50	0.98	0.99
Impurity	Lactose	0.54	0.59	0.54	0.72	0.63	0.77
1.5% (w/w)	Piroxicam α_2	0.95	0.98	0.97	0.98	0.98	0.98
	Piroxicam β	NF	0.98	0.88	0.61	0.98	0.99
Impurity	Lactose	0.63	0.71	0.63	0.62	0.63	0.76
1% (w/w)	Piroxicam α_2	NF	0.98	NF	0.97	0.98	0.98
	Piroxicam β	0.85	0.98	0.89	NF	0.98	0.99
Impurity	Lactose	0.63	0.55	0.63	0.67	0.63	0.70
0.5% (w/w)	Piroxicam α_2	0.97	0.98	0.98	0.85	0.98	0.98
	Piroxicam β	0.99	0.96	0.90	NF	0.99	0.99
Impurity	Lactose	0.58	NF	0.58	NF	0.58	0.75
0.1% (w/w)	Piroxicam α_2	0.99	0.99	0.98	0.74	0.99	0.99
	Piroxicam β	NF	NF	NF	NF	NF	0.98

Abbreviation: NF: not found.

The elucidation was possible for all levels, until 0.1% (w/w) of impurity contamination by selecting one patch. In each case, the obtained correlation coefficient was lower for the random approach. Because the amount of impurity was smaller than the previous case of study, the use of equivalent size patch seemed to be more hazardous. Indeed, the results were significantly improved with the grid analysis. In addition, the analysis time between one patch and the entire grid, was 14 s vs 2 min respectively, which is still an acceptable computation time regarding the size of data. Consequently, it seems that the most appropriate approach for bigger size of data analysis should be the use ESP approach combined with a grid of patches, especially when the impurity level is very low (0.1%

(w/w)). For more details about the results obtained for each PBI approach, PCA representations are given in the supplementary materials (Fig. S1).

The best matching spectra of each compound were then gathered and a least square projection was performed to obtain the repartition of each compound along the tablet surface. As it can be seen in Fig. 3, the different chemical compounds were correctly identified. Additionally, if the aim was to go further in the chemical repartition evaluation, the ESP selected by the proposed strategy could be used as initial estimates for spectral unmixing algorithms.

Compared to the other pixel-selection approaches evaluated in this study, the ESP approach has shown the best results in terms of correlation coefficients but also the smallest analysis time. Indeed, the spectral identification was possible in a maximum of 50 s for the classical data size and 2 min for the big map size by using the grid. Moreover, the correlation coefficients gathered in the end were close to those obtained by the entire database matching strategy, which is much more time-consuming.

As a global conclusion for the analysis of known samples, the ESP strategy have shown a high ability to elucidate from 5 to 0.1% (w/w) of polymorphic impurity for both classical and huge amount of data. Thanks to the proposed strategy, it was possible to detect few contaminating pixels among a million of pixels. The strategy can be considered as validated. Consequently, it will be applied to elucidate the composition of unknown pharmaceutical tablets.

3.3.Investigation of unknown pharmaceutical tablet composition

The principal difficulty when analysing falsified medicines consists in the elucidation of a sample composition that is totally unknown. The Raman hyperspectral images analyzed here were acquired on falsified chloroquine phosphate tablet samples seized during the covid-19 pandemic during a previous study [23]. The PBI approach was applied on these samples to elucidate their chemical composition. The data size of these samples were similar to the carbamazepine ones (150 x 150 pixels). However, because the number of chemical compounds was unknown, the computation of the convex hull was tricky. That is why it has been decided to analyze them using the grid strategy used in the piroxicam study, but with a smaller patch size of 40 x 40 pixels. The results are provided in Table 3.

As it can be seen, it is possible to elucidate each sample with high correlation coefficient. The seven samples are different in terms of chemical composition and number of compounds. As noticed above, the computation of the convex hull was tricky because, depending on the sample, it would be

possible to miss some of them. The obtained results were very close to the ones obtained in the previous study [16] but in much less time. Indeed, because the MCR-ALS is a factorial method, it requires many computation steps to obtain the results. Thus, the improvement in terms of speed of the analysis was at least two-fold, 3 h for the MCR-ALS and 1 min for the proposed PBI. In addition, the applied strategy requiring fewer inputs may be automated.

A least squares projection was done using the best matching ESPs to obtain the distribution maps of the identified compounds. The proposed strategy was able to detect the chloramphenicol localised in a unique pixel, represented in blue in Fig. 4B-C. This observation is very interesting because it means that, with a simple pixel selection strategy, it was possible to detect a single pixel amongst 22,500 pixels, without any unmixing step easing considerably the interpretation and the repeatability of the results.

After evaluating the samples composition by database matching, the second objective was to evaluate if it was possible to detect unknown chemicals that are not yet in the database. Indeed, in the context of falsified medicine analysis, it is important to provide the most exhaustive results to evaluate the potential hazard of the medicine. To mimic the absence of chemical compounds in the database, three of them were removed (titanium dioxide, sodium bicarbonate and sodium sulfate) when doing the database matching. As explained in the material and method part, after removing all the matched ESP, the remaining ones were gathered and projected onto a new principal component space. This procedure was done to detect potential outliers present in the remaining ESP. For that purpose, the Q2 residual distance was observed, giving information about how well the sample are modelled by the PCA.

***Figure 3**. Representation of the results of the least squares prediction with the best matched ESP for 0.5% of impurity. A) Projection of the different chemical compounds along the map. B) Best matched ESP corresponding to the ones who served to do the least square projection. The different spectra are those for the lactose monohydrate, piroxicam a2 and piroxicam β in green, pink and yellow respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)*

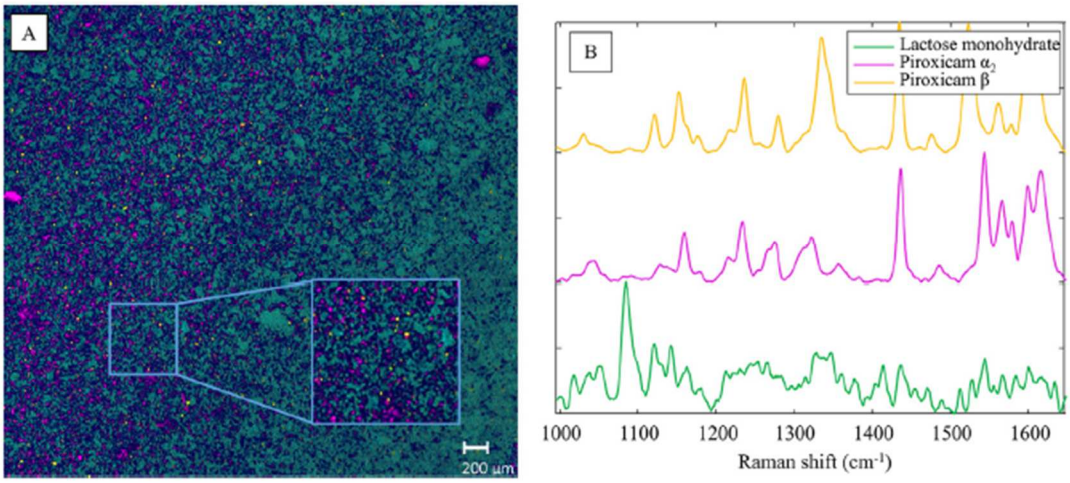


***Table 3** - Results of the seven samples spectral identification. The number given in brackets are the correlation coefficient obtained for the database matching.*

Sample	Identification
A	Metronidazole (0.99), starch (0.96), MgSt (0.92), $Ca_3(PO_4)_2$ (0.81)
B	Metronidazole (0.93), paracetamol (0.91), chloramphenicol (0.87), starch (0.94), MgSt (0.92), $Ca_3(PO_4)_2$ (0.86)
C	Metronidazole (0.97), starch (0.96), $CaCO_3$ (0.83), MgSt (0.97), $Ca_3(PO_4)_2$ (0.93)
D	Metronidazole (0.98), starch (0.94), MgSt (0.90), $Ca_3(PO_4)_2$ (0.77)
F	Chloroquine phosphate (0.93), starch (0.94), $CaCO_3$ (0.88), MgSt (0.90), $Ca_3(PO_4)_2$ (0.94), talc (0.77), MCC (0.78)
G	Chloroquine sulfate (0.96), paracetamol (0.82), starch (0.95), $CaCO_3$ (0.91), MgSt (0.94), $Ca_3(PO_4)_2$ (0.91), talc (0.85), MCC (0.79)
H	Chloroquine sulfate (0.97), starch (0.90), $CaCO_3$ (0.94), MgSt (0.62), talc (0.67), MCC (0.79), sodium bicarbonate (0.79), TiO_2 (0.96), sodium sulfate (0.88)

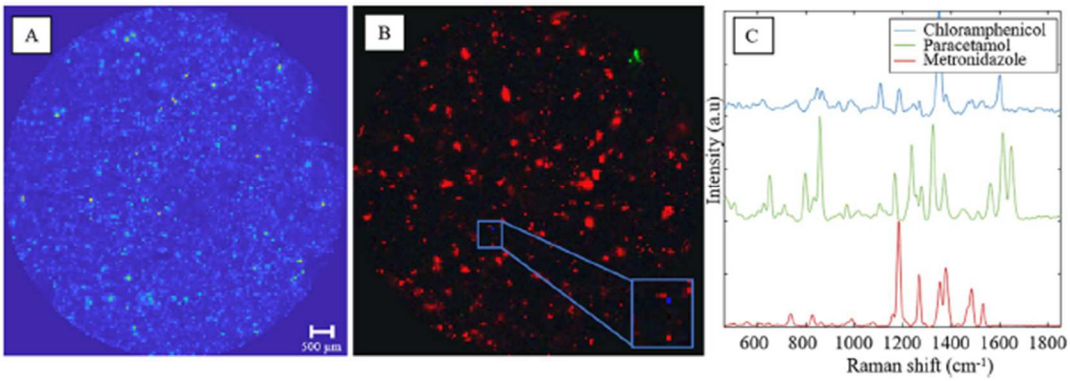
Abbreviation: MCC: microcrystalline cellulose; $CaCO_3$: calcium carbonate; TiO_2: titanium dioxide; MgSt: magnesium stearate; $Ca_3(PO_4)_2$: calcium phosphate.

As it can be seen on Fig. 5-A and Fig. 5-B, the outlying points represented in red, are the most well-resolved spectral signatures, providing more specific information than the ones in blue (Fig. 5-C). It is worth noting that, in ideal case, the ESP are those that either are pure spectra or correspond to the most linearly combined pixels. However, because the computation of the ESP was done with a relatively high number of components, the obtained ESP are noisier which explains in this case why pure spectra can be easily detected as outliers.

Hence, the spectra in red were gathered, the reference spectra were reinjected into the database and the database matching was reiterated. As expected, the results were unequivocal, the strategy

successfully found the three chemical compounds (Fig. 6) with high correlation coefficient (>0.80). Interestingly that the titanium dioxide spectrum was not a pure one, but the Raman scattering at 600 cm^{-1} [1] is a well-known large and intense band for this compound. An unmixing step could be therefore useful to obtain a pure spectrum.

***Figure 4**. Results of the least squares projection on initial data. A) Initial map. B) Representation of 3 compounds elucidated by the strategy. C) Representation of the different pure spectra obtained by the strategy. The spectrum in red, corresponds to the spectrum of metronidazole. The spectrum in green, corresponds to the paracetamol. The spectrum in blue, corresponds to the chloramphenicol. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)*



Therefore, two major conclusions can be drawn. First, the use of PBI based on the selection of ESP allowed elucidating the entire composition of the unknown samples without any ambiguity. Second, thanks to this strategy, it was possible to find information that was not included in the in-house database by using an outlier detection strategy. The next step would be to keep all of the unknown detected spectral signatures in order to, build an “unknown component database” or to check the identity in another database at a later stage. Indeed, even if the compound did not get a database matching, it can be interesting to keep it because it could represent a falsified-manufacturer signature that could be useful in eventual forensic analyses.

***Figure 5**. Results of the PCA done on the ESPs which had correlation coefficient <0.5. A) Representation of the ESP in the PCA space, with the correspondent distances Q2res and T2 Hotelling. B) Plot of the spectra which had Q2res values superior to the critical value obtained with alpha = 0.05. C) Plot of the spectra which had Q2res values inferior to the critical value obtained with alpha = 0.05.*

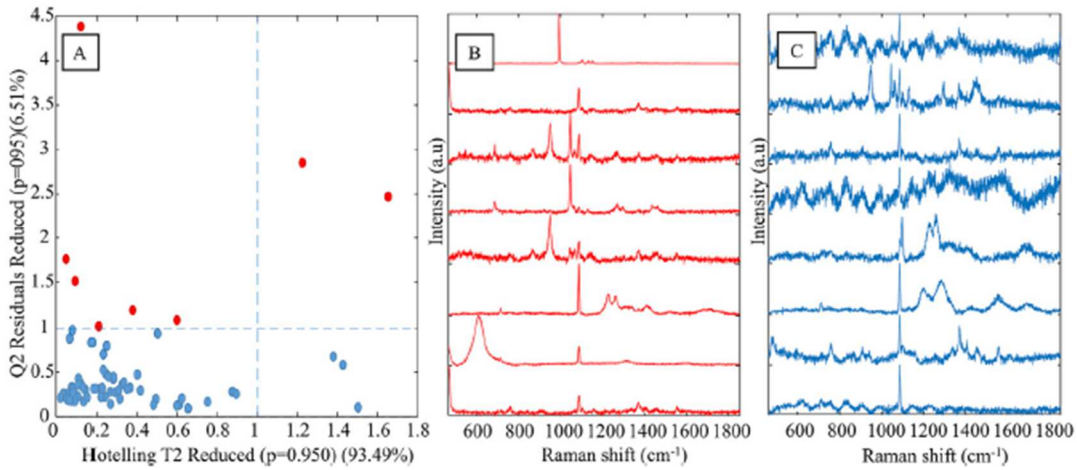
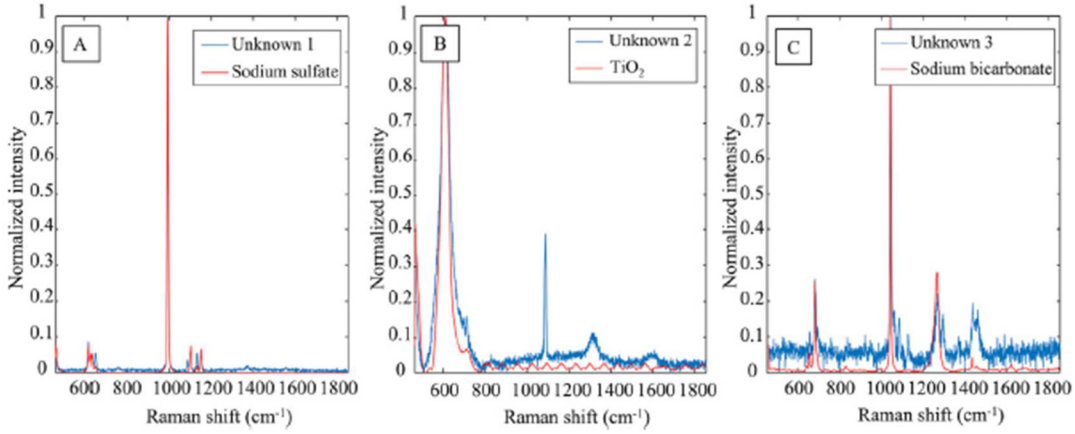


***Figure 6**. Database matching of the three removed compounds. A) Spectrum matching of the sodium sulfate. B) Spectrum of the titanium dioxide (TiO2). C) Spectrum matching of the sodium bicarbonate*.



4. Conclusion

The proposed study highlighted the potential of using essential pixel approach for chemical identification purposes. It has been shown that, for known samples, both tiny and huge amount of data can be analyzed without the need of the entire map, by selecting only a few percentage of pixels (8% of the initial data). However, when the analyte is present in very low amounts (0.1% (w/w)), it is better evaluating the ESP on the entire map by means of a grid. One major interest of the proposed

strategies is that the computational time did not exceed 2 min and provided the entire elucidation of the chemical compounds.

In addition, it has been shown that the analysis of unknown samples was possible. Indeed, thanks to the inherent properties of the algorithm, the spectra gathered in the end correspond to the most interesting ones (both pure and mixed spectra). On one side, the database matching was possible with good correlation coefficients. On the other side, thanks to the use of a simple statistical calculation, the Q2 residuals, unknown spectral signatures were identified. Thanks to the high correlation coefficient, it was possible to perform a classical least square to obtain the repartition of the chemicals over the tablet surface. This methodology could thus be applicable for seized falsified medicines analysis.

Overall, we highlighted the applicability of the propose methodology to other hyperspectral imaging techniques or other kind of matrices. Indeed, thanks to the inherent properties of the essential spectral pixel algorithm combined with the proposed methodology, the only requirement is to have at least one pure pixel by component. In case of mixed spectra or non-linear pixels, the ESP could be used as a pre-processing step, to reduce data dimensionality, which has been successfully demonstrated in other study [30] and could be easily implemented. Indeed, further investigation could be done by initializing the MCR-ALS with the obtained ESP.

CRediT authorship contribution statement

Laureen Coic: Writing - original draft, Writing - review & editing, Software, Formal analysis, Visualization. **Pierre-Yves Sacre**: Conceptualization, Writing - review & editing. **Charlotte De Bleye**: Writing - review & editing. **Marianne Fillet**: Conceptualization, Funding acquisition, Writing - review & editing. **Cyril Ruckebusch**: Conceptualization, Funding acquisition, Writing - review & editing. **Philippe Hubert**: Supervision, Funding acquisition, Project administration, Writing - review & editing. **Eric Ziemons**: Supervision, Funding acquisition, Project administration, Conceptualization, Writing - review & editing.

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

This project has been supported by the European funds of regional development (FEDER) and by Walloon Region of Belgium as part of the operational program “Walloon-2020.EU” (L. Coic and A. Dispas).

The financial support of this research by the Walloon Region of Belgium in the framework of the Vibra4Fake project (convention n:7517) is gratefully acknowledged.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aca.2021.338361.

References

[1] V. Castiglione, P.-Y. Sacre, E. Cavalier, P. Hubert, R. Gadisseur, E. Ziemons, Raman chemical imaging, a new tool in kidney stone structure analysis: casestudy and comparison to Fourier Transform Infrared spectroscopy, PloS One 13 (2018), e0201460, https://doi.org/10.1371/journal.pone.0201460.

[2] A. Käppler, F. Windrich, M.G.J. Löder, M. Malanin, D. Fischer, M. Labrenz, K.- J. Eichhorn, B. Voit, Identification of microplastics by FTIR and Raman microscopy: a novel silicon filter substrate opens the important spectral range below 1300 cm1 for FTIR transmission measurements, Anal. Bioanal. Chem. 407 (2015) 6791-6801, https://doi.org/10.1007/s00216-015-8850-8.

[3] B. Prats-Mateu, M. Felhofer, A. de Juan, N. Gierlinger, Multivariate unmixing approaches on Raman images of plant cell walls: new insights or overinterpretation of results? Plant Methods 14 (2018) 52, https://doi.org/ 10.1186/s13007-018-0320-9.

[4] X. Xu, J. Li, C. Wu, A. Plaza, Regional clustering-based spatial preprocessing for hyperspectral unmixing, Remote Sens. Environ. 204 (2018) 333-346, https:// doi.org/10.1016/J.RSE.2017.10.020.

[5] J. Bøtker, J.X. Wu, J. Rantanen, Hyperspectral imaging as a part of pharmaceutical product design, in: Data Handl. Sci. Technol, Elsevier Ltd, 2020, pp. 567-581, https://doi.org/10.1016/B978-0-444-63977-6.00022-5.

[6] EDQM - European directorate for the quality of medicines (n.d.), https://www. edqm.eu/. (Accessed 4 December 2020).

[7] J. Cailletaud, C. De Bleye, E. Dumont, P.-Y. Sacre, Y. Gut, L. Bultel, Y.-M. Ginot, P. Hubert, E. Ziemons, Towards a spray-coating method for the detection of low-dose compounds in pharmaceutical tablets using surface-enhanced Raman chemical imaging (SER-CI), Talanta 188 (2018) 584-592, https:// doi.org/10.1016/j.talanta.2018.06.037.

[8] C. De Bleye, P.Y. Sacre, E. Dumont, L. Netchacovitch, P.F. Chavez, G. Piel, P. Lebrun, P. Hubert, E. Ziemons, Development of a quantitative approach using surface-enhanced Raman chemical imaging: first step for the determination of an impurity in a pharmaceutical model, J. Pharmaceut. Biomed. Anal. 90 (2014) 111-118, https://doi.org/10.1016/j.jpba.2013.11.026.

[9] H. Mitsutake, M.D.G. Neves, D.N. Rutledge, R.J. Poppi, M.C. Breitkreitz, Extraction of information about structural changes in a semisolid pharmaceutical formulation from near-infrared and Raman images by multivariate curve resolutionealternating least squares and ComDim, J. Chemom. (2020), https://doi.org/10.1002/cem.3288.

[10] P.Y. Sacre, P. Lebrun, P.F. Chavez, C. De Bleye, L. Netchacovitch, E. Rozet, R. Klinkenberg, B. Streel, P. Hubert, E. Ziemons, A new criterion to assess distributional homogeneity in hyperspectral images of solid pharmaceutical dosage forms, Anal. Chim. Acta 818 (2014) 7-14, https://doi.org/10.1016/ j.aca.2014.02.014.

[11] L. Coic, P.-Y. Sacre, A. Dispas, A.K. Sakira, M. Fillet, R.D. Marini, P. Hubert, E. Ziemons, Comparison of hyperspectral imaging techniques for the elucidation of falsified medicines composition, Talanta 198 (2019) 457-463, https://doi.org/10.1016/j.talanta.2019.02.032.

[12] R. Deidda, P.-Y. Sacre, M. Clavaud, L. Coïc, H. Avohou, P. Hubert, E. Ziemons, Vibrational spectroscopy in analysis of pharmaceuticals: critical review of innovative portable and handheld NIR and Raman spectrophotometers, TrAC Trends Anal. Chem. (Reference Ed.) 114 (2019) 251-259, https://doi.org/10.1016/J.TRAC.2019.02.035.

[13] P.-Y. Sacre, E. Deconinck, L. Saerens, T. De Beer, P. Courselle, R. Vancauwenberghe, P. Chiap, J. Crommen, J.O. De Beer, Detection of counterfeit Viagra® by Raman microspectroscopy imaging and multivariate analysis, J. Pharmaceut. Biomed. Anal. 56 (2011) 454-461, https://doi.org/ 10.1016/j.jpba.2011.05.042.

[14] A. de Juan, R. Tauler, Multivariate curve resolution-alternating least squares for spectroscopic data, in: Data Handl. Sci. Technol, Elsevier Ltd, 2016, pp. 5-51, https://doi.org/10.1016/B978-0-444-63638-6.00002-4.

[15] A. de Juan, Multivariate curve resolution for hyperspectral image analysis, in: Data Handl. Sci. Technol, Elsevier Ltd, 2020, pp. 115-150, https://doi.org/ 10.1016/B978-0-444-63977-6.00007-9.

[16] R. Tauler, A. de Juan, Multivariate curve resolution, in: P. Gemperline (Ed.), Pract. Guid. To Chemom., CRC Press, 2006, pp. 417-467.

[17] H. Mitsutake, S.R. Castro, E. de Paula, R.J. Poppi, D.N. Rutledge, M.C. Breitkreitz, Comparison of different chemometric methods to extract chemical and physical information from Raman images of homogeneous and heterogeneous semi-solid pharmaceutical formulations, Int. J. Pharm. 552 (2018) 119-129, https://doi.org/10.1016/J.IJPHARM.2018.09.058.

[18] J. Jaumot, R. Tauler, MCR-BANDS: a user friendly MATLAB program for the evaluation of rotation ambiguities in Multivariate Curve Resolution, Chemometr. Intell. Lab. Syst. 103 (2010) 96-107, https://doi.org/10.1016/ j.chemolab.2010.05.020.

[19] A. Malik, R. Tauler, Ambiguities in multivariate curve resolution, in: Data Handl. Sci. Technol., Elsevier Ltd, 2016, pp. 101-133, https://doi.org/10.1016/ B978-0-444-63638-6.00004-8.

[20] M. Boiret, A. de Juan, N. Gorretta, Y.M. Ginot, J.M. Roger, Distribution of a low dose compound within pharmaceutical tablet by using multivariate curve resolution on Raman hyperspectral images, J. Pharmaceut. Biomed. Anal. 103 (2015) 35-43, https://doi.org/10.1016/j.jpba.2014.10.024.

[21] C. Fauteux-Lefebvre, F. Lavoie, R. Gosselin, A hierarchical multivariate curve resolution methodology to identify and map compounds in spectral images, Anal. Chem. 90 (2018) 13118-13125, https://doi.org/10.1021/ acs.analchem.8b04626.

[22] A. de Juan, Multivariate curve resolution for hyperspectral image analysis, in: Data Handl. Sci. Technol, Elsevier Ltd, 2020, pp. 115-150, https://doi.org/ 10.1016/B978-0-444-63977-6.00007-9.

[23] 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. Pharmaceut. Biomed. Anal. (2020) 113761, https://doi.org/10.1016/j.jpba.2020.113761.

[24] J.-K. Park, S. Lee, A. Park, S.-J. Baek, Adaptive hit-quality index for Raman spectrum identification, Anal. Chem. 92 (2020) 10291-10299, https://doi.org/ 10.1021/acs.analchem.0c00209.

[25] V. Sevetlidis, G. Pavlidis, Effective Raman spectra identification with treebased methods, J. Cult. Herit. 37 (2019) 121-128, https://doi.org/10.1016/ j.culher.2018.10.016.

[26] J.A. Fine, A.A. Rajasekar, K.P. Jethava, G. Chopra, Spectral deep learning for prediction and prospective validation of functional groups, Chem. Sci. 11 (2020) 4618-4630, https://doi.org/10.1039/c9sc06240h.

[27] A. Mahmood, S. Khan, Correlation-coefficient-based fast template matching through partial Elimination, IEEE Trans. Image Process. 21 (2012) 2099-2108, https://doi.org/10.1109/TIP.2011.2171696.

[28] M. Boiret, N. Gorretta, Y.M. Ginot, J.M. Roger, An iterative approach for compound detection in an unknown pharmaceutical drug product: application on Raman microscopy, J. Pharmaceut. Biomed. Anal. (2016), https:// doi.org/10.1016/j.jpba.2015.12.038.

[29] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137-148, https://doi.org/10.1080/00401706.1969. 10490666.

[30] 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.

[31] C. Ruckebusch, R. Vitale, M. Ghaffari, S. Hugelier, N. Omidikia, Perspective on essential information in multivariate curve resolution, TrAC Trends Anal.Chem. (Reference Ed.) 132 (2020) 116044,https://doi.org/10.1016/j.trac.2020.116044.

[32] C. Rustichelli, G. Gamberini, V. Ferioli, M.C. Gamberini, R. Ficarra, S. Tommasini, Solid-state study of polymorphic drugs: Carbamazepine, in: J. Pharm. Biomed. Anal, Elsevier, 2000, pp. 41-54, https://doi.org/10.1016/ S0731-7085(00)00262-4.

[33] K.M. Lutker, A.J. Matzger, Crystal polymorphism in a carbamazepine derivative: Oxcarbazepine, J. Pharmacol. Sci. 99 (2010) 794-803, https://doi.org/ 10.1002/jps.21873.