



## Low-content quantification in powders using Raman spectroscopy: a facile chemometric approach to sub 0.1% limits of detection.

Title	Low-content quantification in powders using Raman spectroscopy: a facile chemometric approach to sub 0.1% limits of detection.
Author(s)	Li, Boyan;Calvet, Amandine;Ryder, Alan G.;Casamayou-Boucau, Yannick;Morris, Cheryl
Publication Date	2015-02-23
Publisher	American Chemical Society
Repository DOI	<a href="https://doi.org/10.1021/ac504776m">10.1021/ac504776m</a>

DOI: [10.1021/ac504776m](https://doi.org/10.1021/ac504776m)

## **Low-content quantification in powders using Raman spectroscopy: *a facile chemometric approach to sub 0.1% limits of detection.***

Boyan Li, Amandine Calvet, Yannick Casamayou-Boucau, Cheryl Morris, and Alan G. Ryder\*

Nanoscale Biophotonics Laboratory, School of Chemistry, National University of Ireland, Galway, Galway, Ireland.

\* To whom correspondence should be addressed.

**Tel:** +3539149 2943 **Email:** alan.ryder@nuigalway.ie

### **ABSTRACT**

A robust and accurate analytical methodology for low-content (<0.1%) quantification in the solid-state using Raman spectroscopy, sub-sampling, and chemometrics was demonstrated using a piracetam–proline model. The method involved a 5-step process: collection of relatively large number of spectra (8410) from each sample by Raman mapping, meticulous data pretreatment to remove spectral artefacts, use of a 0–100% concentration range partial least squares (PLS) regression model to estimate concentration at each pixel, use of a more-accurate, reduced concentration range PLS model to accurately calculate analyte concentration at each pixel, and finally statistical analysis of all 8000+ concentration predictions to produce an accurate overall sample concentration. The relative prediction accuracy was ~2.4% for a 0.05~1.0% concentration range and the limit of detection was comparable to high performance liquid chromatography (0.03% *versus* 0.041%). For data pretreatment, we developed a unique cosmic ray removal method and used an automated baseline correction method, neither of which required subjective user intervention and thus were fully automatable. The method is applicable to systems, which cannot be easily analyzed chromatographically such as hydrate, polymorph, or solvate contamination.

### **INTRODUCTION**

Many active pharmaceutical ingredients (APIs) are solids, which are isolated and purified prior to incorporation into a dosage form like a tablet. The certification of solid API and excipient purity/quality is critical and the chemical contamination level must, by necessity be low and precisely specified. Chemical impurities are either solvents, unwanted side-products of the synthesis step, degradation products, unwanted polymorphs, solvates, hydrates, or foreign material introduced during handling. For each API, well-defined tests for specific contaminants are generated as part of the drug licensing process, and are often time-consuming or involve

considerable costs to the manufacturer. From a regulatory standpoint, it is also critical to quantify solid-state forms such as polymorph content in APIs.

High performance liquid chromatography (HPLC), the traditional means of quantitative analysis provides no information about polymorphic content. Powder X-ray diffraction (PXRD) is considered the definitive method for polymorph analysis, however, it is time-consuming, and it is not suitable for sub-1% quantification. Vibrational spectroscopy techniques are widely used in small molecule, solid-state API analysis.<sup>1-4</sup> These often require minimal sample preparation, and can provide much more chemical and structural information.<sup>5-6</sup> Solid-state forms like polymorphs, amorphous solids, salts, solvates, co-crystals can have different physicochemical properties due to the variations in their free energies and inter- and intra-molecular bonding,<sup>7-8</sup> thus accurate characterization and quantification of form or change is important.<sup>9</sup> Raman spectroscopy is particularly suited to the analysis of formulated APIs as they often have a higher Raman activity than the excipients.<sup>10</sup> Raman has also been used for relatively low-concentration quantification of polymorphs,<sup>11-12</sup> and impurities,<sup>13</sup> because the sharper and well-defined Raman spectra often enable easier discrimination in mixtures compared to NIR for example.<sup>12</sup>

Raman imaging/mapping has been used for contaminant detection and API/excipient distribution of solid-state materials.<sup>4, 14-19</sup> However, most studies focus on identification rather than quantification. These studies have revealed that numerous parameters are to be taken into account in order to achieve robust Raman based methods, such as spot size, number of data points per sample, sampled volume, sampling methods, particle size and homogeneity, among others.<sup>20-21</sup> In general, for most quantitative Raman methods, one key practical consideration is to increase the sampled area to reduce error associated with sub-sampling and sample heterogeneity. This can be achieved by sample rotation during measurement, or averaging of spectra collected at many different locations of the sample.<sup>21</sup> The use of macro spot sizes and transmission Raman are also solutions for quantification at relatively high concentrations (>1%w/w).<sup>10, 22</sup> The explicit use of sub-sampling and statistical analysis for the detection of specific components in low-content solid formulations was demonstrated by Šašić,<sup>12, 23</sup> however, it was not extended to the quantification of low-concentration contaminants.

The present work used Raman mapping with its high sampling rate to develop a robust and accurate analytical methodology for quantifying low-content (<0.1%) components in solid matrices using a piracetam in proline model where the piracetam signal was not obvious from visual spectral inspection. The novelty of the approach involved sub-sampling data collection to generate a large array of heterogeneous Raman spectra. If sufficiently large, and sampled from small enough physical regions of the solid sample, the heterogeneity of this Raman dataset should be related to fluctuations in local concentration, which in turn should be correlated to the concentration of the low-content analyte. To extract this information, the local concentration at each Raman map pixel was predicted using chemometric methods (Partial least squares, PLS<sup>24</sup>), and all the prediction data were then statistically analyzed to generate an accurate analyte concentration. The PLS calibrations for local concentration prediction were generated using a standard Raman mapping approach on “high concentration” samples spanning wide

concentration ranges where the spectra were averaged before building the PLS model. For the low-content analysis, Raman data had to be automatically corrected for measurement artefacts (baseline and cosmic ray events) that would otherwise impair quantification. The key goal was to generate an accuracy and limit of detection (LOD) comparable to HPLC. Finally, the PLS models were also used to generate Raman maps that more reliably located the low-level piracetam contaminant across the surface of homogeneous powder mixtures.

## MATERIALS AND METHODS

**Samples.** Piracetam (2-oxo-1-pyrrolidineacetamide, polymorphic form III) and L-proline ( $\geq 99\%$ ), were purchased from Sigma-Aldrich (Ireland) and used as received. 61 powder mixtures ( $\sim 3.0$  g each) were prepared in triplicate by weighting out appropriate amounts of piracetam and proline, and piracetam content varied from 0 to 100% (w/w%): 20 mixtures between 0.05 to 1.0% and 40 mixtures between 2.0 and 100%. To ensure sample homogeneity, components were individually ground first for 10 min, combined and mixed thoroughly using a vortex mixer for 2 min. Samples were placed in a stainless steel sample holder and tamped down to generate a flat surface. Hydrated proline was generated by exposure in a controlled humidity chamber (see Supporting Information, SI).

**Instrumentation, data collection and analysis.** Raman spectra were collected using a Raman RXN Systems Analyzer (Kaiser Optical Systems, Inc.) with 785 nm excitation and running the HoloGRAMS PhAT Version 4.1 software. This was fitted with a unique PhAT microprobe coupled into a microscope. This probe had 50 individually addressable optical fibers, which were input into the spectrometer and each fiber was focused/dispersed on different pixels rows on the charge-coupled device (CCD) detector, and the data were binned into 10 separate channels. A 35 mm focal length objective gave a 1 mm diameter spot size, and each fiber addressed a separate sample volume within this space. Raman spectra ( $200\sim 1896$   $\text{cm}^{-1}$ ) were collected from a  $29\times 29$  pixel grid, 1 mm spacing, which generated an 8410 spectra dataset from the 10 channels. Good quality spectra were obtained using 1 second exposure (50 minute acquisition time per sample). MATLAB was used for data pretreatment prior to quantitative analysis, and codes for ant colony optimization (ACO)<sup>25-26</sup> and calculating PLS limits of detection<sup>27</sup> were generously provided by Prof. A.C. Olivieri (Universidad Nacional de Rosario, Argentina).

ACO was required for two purposes: first, remove the influence of noisy or uninformative spectral variables, and second, reduce the variable-to-sample ratio (849/61) because high dimensionality in the spectral data often renders the prediction of a dependent variable unreliable with regression-based statistical estimators. ACO was implemented using  $\rho$  (rate of pheromone evaporation) = 0.65,  $N$  (number of ants) = 350,  $w$  (sensor width) = 2, a maximum number of time steps of 50, and 100 repeated Monte Carlo calculation cycles to build a histogram of variable selection probability. PLS model quality was assessed based on: root mean square error of

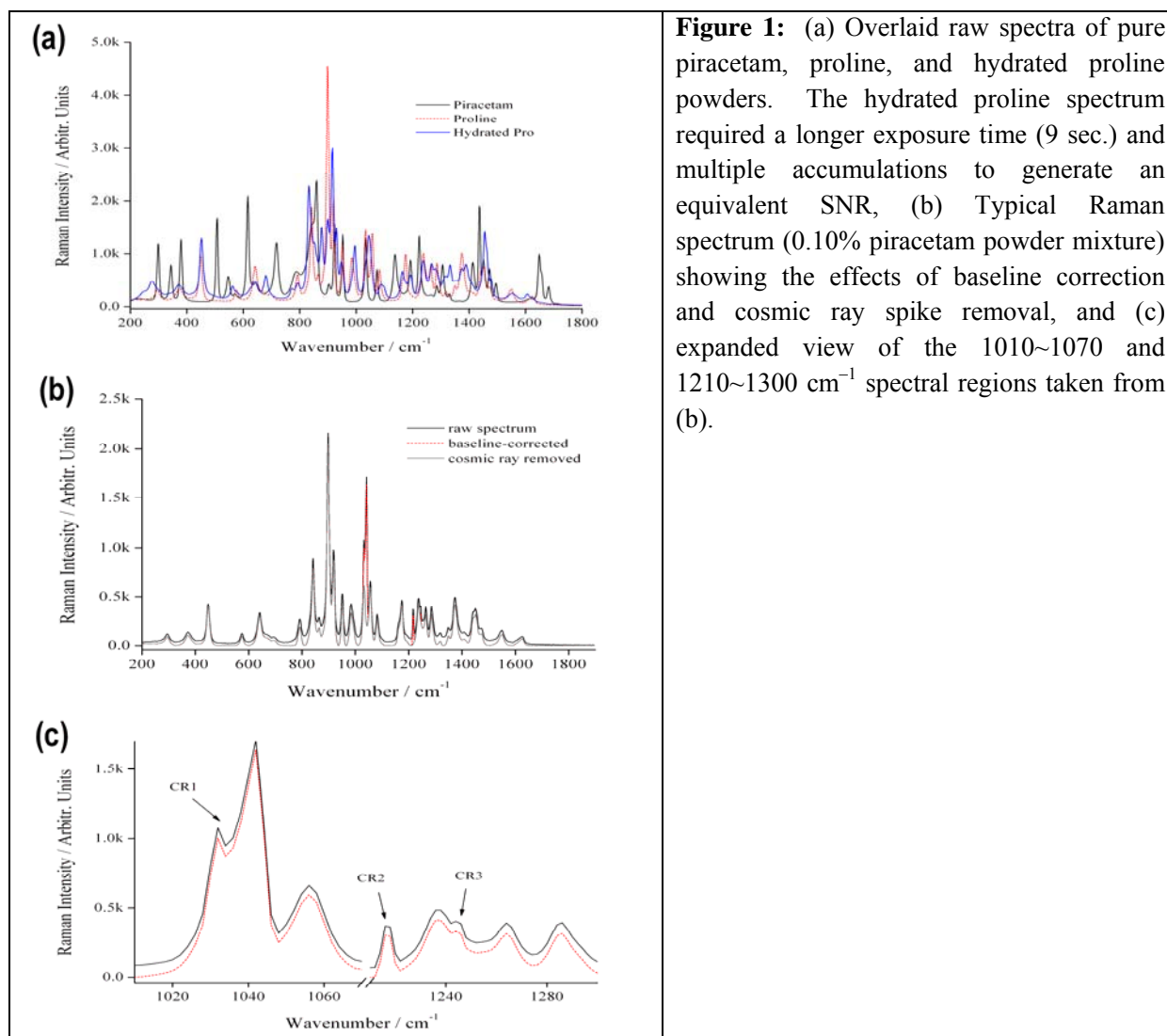
calibration (RMSEC), root mean square error from cross-validation (RMSECV), root mean square error of prediction (RMSEP), relative prediction accuracy ( $\text{REP}\% = 100 \times \text{RMSEP} / \bar{y}$ , where  $\bar{y}$  = mean value of measured piracetam concentration), and the square of the correlation coefficient ( $R^2$ ) between predicted and measured values.<sup>24, 28</sup> The relative model accuracy was described by  $\text{REC}\% = 100 \times \text{RMSEC} / \bar{y}$ , and  $\text{RECV}\% = 100 \times \text{RMSECV} / \bar{y}$ .

## RESULTS AND DISCUSSION

**Spectroscopy.** Each of the 50 fibres sampled a slightly different physical region of the solid within the objective field of view. Each of the 10 channels comprised different numbers of detector rows and thus the spectral intensity varied considerably at each sample point (Figure S-1, SI) with the best quality spectra from channels 5/6 and the lowest intensity spectra from channels 1/10. This resulted in variable signal-to-noise ratios (SNRs) for each channel. For a typical example, using the  $1650 \text{ cm}^{-1}$  piracetam band and  $1720\sim 1880 \text{ cm}^{-1}$  baseline region, SNRs (peak to peak) of 59.2, 83.5, 77.9, 89.8, **102.1**, **97.5**, 90.3, 87.9, 73.5, and 66.7 were calculated for channels 1 to 10 respectively. This variance and the fact that each channel represented a different physical sample location necessitated calculation of separate chemometric models for each channel. This constituted an almost independent, internal 10 replicate measurement at each map location.

Two very significant problems had to be addressed prior to chemometric analysis: variable baseline signals and cosmic ray induced spectral contamination. Both effects could generate larger variance than that produced by the low-content analyte, making accurate quantification impossible. However, because of the large numbers of spectra involved, automated methods, with no subjective user intervention were required.

**Baseline correction.** Baseline/background variance is a common issue in Raman spectroscopy caused by factors such as fluorescent contaminants, particle size variation, and instrumental factors. Baseline signal contributes unwanted shot noise, which obscures Raman signals from low-concentration analytes hindering identification and quantification. Experimental approaches to background/baseline correction are not suitable here.<sup>29-34</sup> A chemometric approach (of which there are many) was preferred because this could be more easily implemented on conventional Raman systems.<sup>35-50</sup>



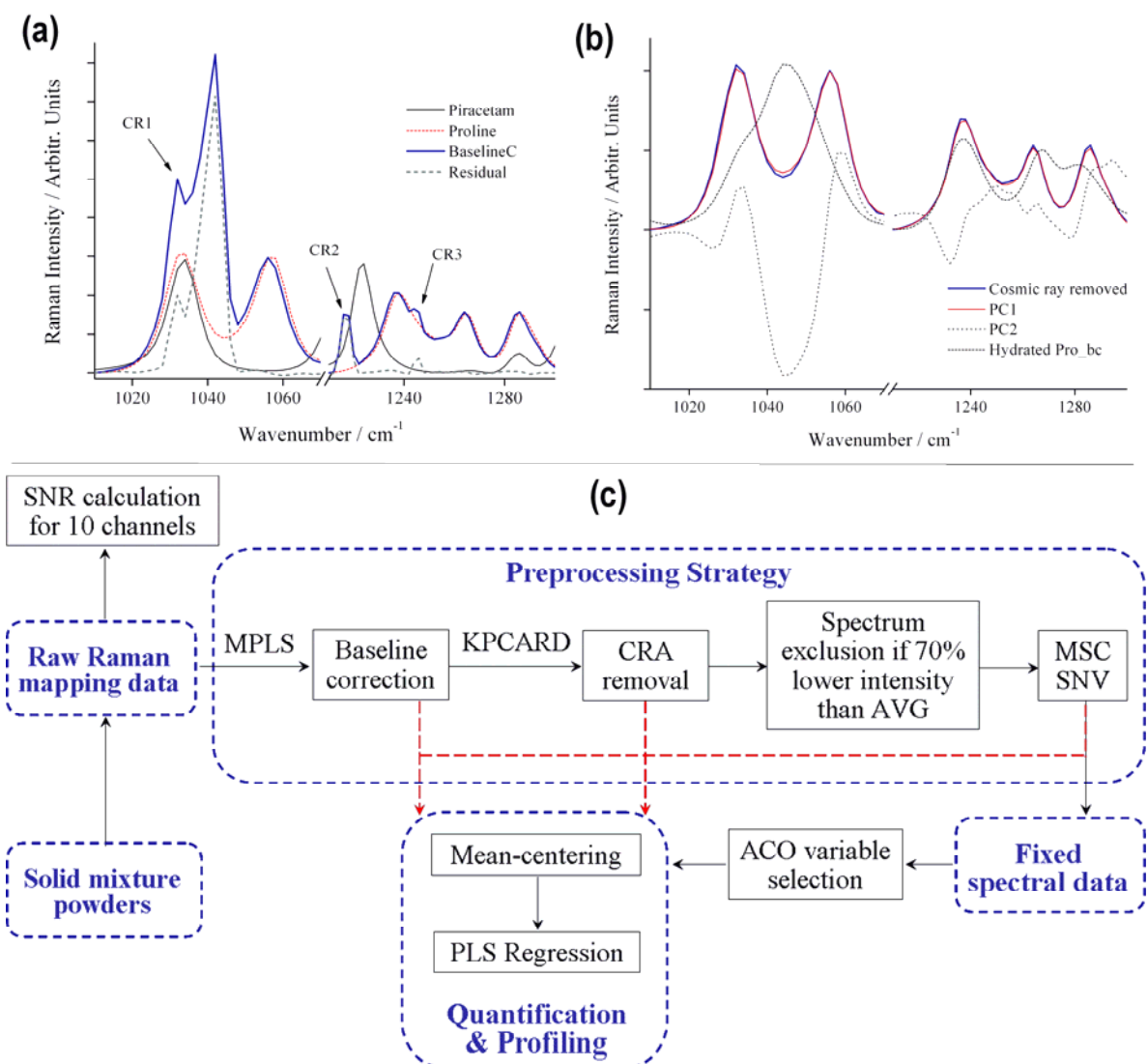
**Figure 1:** (a) Overlaid raw spectra of pure piracetam, proline, and hydrated proline powders. The hydrated proline spectrum required a longer exposure time (9 sec.) and multiple accumulations to generate an equivalent SNR, (b) Typical Raman spectrum (0.10% piracetam powder mixture) showing the effects of baseline correction and cosmic ray spike removal, and (c) expanded view of the 1010~1070 and 1210~1300  $\text{cm}^{-1}$  spectral regions taken from (b).

Morphological weighted penalized least squares (MPLS)<sup>49</sup> was used for baseline correction of Raman spectra because of its inherent simplicity, combined with its flexibility, suitability for automation, and effectiveness at mitigating baseline artefacts. MPLS required neither *a priori* knowledge nor subjective user intervention, and was reasonably efficient computationally (~5 minutes for 8410 spectra). Figure 1 shows typical raw and baseline corrected Raman spectra, and although not immediately obvious here, MPLS reduced spectral variance which had a significant effect on model quality (SI, sections S3/4, Tables S-2/S-3). The need for baseline correction is obviously greater for fluorescent samples or where there are diffuse reflectance artefacts.<sup>10</sup>

**Cosmic ray artefact (CRA) removal.** CRAs are problematic because they lead to random, positive, unidirectional, erroneous spikes in the spectra. The frequency and location of CRAs is random, and peak intensity and width can vary very significantly. CRAs, particularly when

overlapped with Raman bands of interest, complicate signal interpretation, increase unwanted signal variance, and degrade chemometric modelling accuracy.<sup>51-52</sup> Here, it was critical because CRAs increased spectral variance and thus non-analyte related noise, which negatively affected the statistical analysis of large Raman datasets with very weak analyte signal contributions. Various CRA correction methods have been described in the literature,<sup>51-65</sup> however, most are semi-automated requiring user intervention. Acquiring additional spectra at each sampling point and then discarding CRA contaminated spectra (by manual or automated assessment) was one option; however, this would result in an unsustainable increase in measurement time.

The need for a simple, fully automated method that could generate a minimal level of spectral distortion motivated us to develop a novel CRA removal method, *kernel principal component analysis residual diagnosis (KPCARD)*. KPCARD was based on the stochastic nature of CRAs; therefore, the most significant principal components (PCs) in PCA analysis of large Raman datasets should not contain CRAs. KPCARD comprised of two steps: First, CRA identification based on the statistics of the residual obtained at each wavenumber from running kernel PCA (kPCA) on the entire data of each Raman mapping measurement. Second, the nearest neighbor spectrum, most similar to the CRA contaminated spectrum and PCs obtained by kPCA were both used to generate a robust, best curve fit (*i.e.* the corrected spectrum)<sup>66</sup> to the CRA contaminated spectrum. This corrected spectrum was then used for modelling.



**Figure 2:** (a) Spectra of piracetam, proline, the baseline-corrected spectrum from a 0.1% sample and its residual profile after being submitted to the kPCA of the entire mapping data of one Raman measurement (0.10% piracetam powder mixture); (b) Overlap of the CRA corrected spectrum and the baseline-corrected spectrum of hydrated proline with PC1 and PC2 obtained from the kPCA of the entire Raman mapping data from the mixture. Explained variance was as follows: PC1 (99.9717%), PC2 (0.013%), PC3 (0.004%), and PC4 (0.004%); (c) Scheme showing all the steps in the quantification process for the piracetam analyte. Black solid lines show the methodology as implemented, red dashed lines show where iterative improvements were made during development.

Figure 2a/b shows KPCARD CRA in operation. CRAs were located at 1026~1048, 1212~1222, and 1242~1248 cm<sup>-1</sup> (CR1, 2, 3, respectively). Comparison with pure component spectra confirmed that these bands were not due to piracetam or proline (dry or hydrated). Since CRAs have varying bandwidth and intensity, they can have different spectral impacts and ultimately the quantification result. CR1 overlapped the 1034 cm<sup>-1</sup> band, common to piracetam

and proline, and give the appearance of another Raman band. CR2 was close to a real piracetam band and might contribute to a higher prediction of piracetam at this pixel. The narrow CR3 spike overlapped a proline band and so might cause a marginally smaller piracetam concentration to be calculated.

kPCA generated large residuals at the CR1, 2, and 3 locations. When compared to the residual statistics from all 8410 spectra, it indicated that cosmic rays gave rise to three large spikes and they were not real Raman bands. Two significant PCs, (Figure 2b), were obtained from kPCA of the entire Raman mapping data with explained variances of 99.97 and 0.01% respectively. PC1 was a composite profile of both proline and piracetam, whereas the PC2 profile resembled neither. Since proline was easily hydrated by exposure to the laboratory environment, we suspected that PC2 represented a hydration effect. Comparison of the dry/hydrated proline spectra with PC2 indicated that hydration was the largest contributor (Figure 1a and SI) but that there were also unresolved contributions from the natural spectral variance of the 8410 dataset spectra. This highlighted an issue with the use of PCs for image analysis because sometimes one cannot simply attribute an extracted PC with a specific chemical species present in a mixture sample, nor can the resultant PC scores be directly correlated with analyte concentration. This is particularly the case where the contaminant analyte is present in very low concentration. The corrected spectrum generated (Figure 2) was virtually identical in profile to PC1. Compared to existing CRA correction methods (data not shown),<sup>51-52</sup> KPCARD was conceptually simpler, computationally more efficient, and easier to implement. The method was able to automatically accomplish cosmic ray contaminant removal with minimal data distortion.<sup>67</sup> For example, CRA correction of a dataset (8410 spectra×849 variables) took ~2 minutes on a desktop computer (2.8 GHz CPU/6 GB RAM), and  $6.8 \times 10^{10}$  floating point operations (flops) were needed.

**Piracetam Quantification:** The first stage in the quantification methodology was to build the best possible large range PLS model, which was used for screening the data. The first quantification trial used PLS on the raw Raman spectra (from 27 mixtures, 0 to 100% piracetam concentration range) where the Raman data from each sample were averaged into 10 single spectra (corresponding to the 10 channels). The only spectra excluded from the averaging process were those with intensities 70% lower than the total average spectral intensity. Three step pre-processing was applied: multiplicative scatter correction (MSC),<sup>68-70</sup> to remove scattering artefacts; standard normal variate (SNV),<sup>68-70</sup> to remove from the average spectra of all the measurements some of the large amount of variability caused by scattering variations between measurements; and dataset mean-centering (MC).

PLS models were then built using the pre-processed spectra ( $200\sim 1896\text{ cm}^{-1}$ ) and performing leave-one-out cross-validation.<sup>71</sup> Only one PC was required because the samples should be binary mixtures. RMSEC and RMSECV for *channel 5* (the best SNR Raman data) were 2.32% and 2.62%, respectively. A second PLS model using the same sample set but with baseline-corrected spectra gave higher RMSEC (2.89%) and RMSECV (3.24%) values. This

might suggest that baseline correction was both unnecessary and detrimental; however, the final optimization step in the quantification modelling process required variable selection for which this pre-processing was required (see SI for details). This slight drop in PLS performance could be due to an increase in the relative noise contribution obscuring the weaker piracetam signals from the low content samples.

ACO selected 138 variables (Figure 3a), as being more descriptive for the piracetam analyte.<sup>25-26</sup> With these variables, and the 0~100% piracetam samples, new PLS models were recalculated for all 10 channels. ACO produced a *ca.* 30~45% improvement in terms of RMSEC/RMSECV errors (Table S-1, SI). While this was significant, it was still too large to obtain accurate predictions for the sub 1% piracetam containing samples. One cause of these relatively large errors in Model-1 was the very wide concentration range attempted which encompasses significant matrix changes. The next obvious stage was to implement a series of reduced range PLS models and see if that could improve accuracy by minimizing the effects of gross matrix changes. Samples were segmented piecewise into different concentration ranges: 0~2.5, 2.5~21.5, 21.5~85, 85~100% and four predictive models (*Model-2–Model-5*) using 21, 9, 12, and 10 samples respectively, were generated for every channel. Pre-processing was the same in every case apart from the 0~2.5% *Model-2* where orthogonal signal correction (OSC)<sup>72</sup> was implemented after SNV and before MC to minimize noise and contribution from the large proline signal.

There was a dramatic improvement in model quality (Table 1) and it was interesting to note that model performance was not related to channel SNR, which indicates that the pre-processing method was relatively robust. As expected, model performance was poorer for low-content samples where subtle spectral differences were convoluted with noise and small sampling variations. Further improvement in accuracy using the averaged Raman data, would require many additional calibration set samples, and this highlighted the fundamental difficulty with this approach. Attaining high accuracy for low analyte concentrations required large numbers of low-content calibration samples, which is technically challenging. First, there are practical difficulties with the manufacture of solid mixtures with precisely known homogeneously distributed contaminants. Second, %REP is also directly dependent on calibration sample numbers, for example, a REP of <5% required more than 200 samples.<sup>73</sup>

**Table 1:** RMSEC/RMSECV values (in w/w%) obtained for the final piracetam quantification PLS models obtained for each spectrometer channel. Model accuracy was assessed by relative REC% and RECV% for calibration and cross-validation respectively.

PLS model	Model-1	Model-2	Model-3	Model-4	Model-5
<b>Piracetam (w/w%)</b>	<b>0~100</b>	<b>0~2.5</b>	<b>2.5~21.5</b>	<b>21.5~85.0</b>	<b>85.0~100</b>
Channel1	1.63/1.86	0.027/0.029	0.14/0.16	0.84/0.99	0.20/0.30

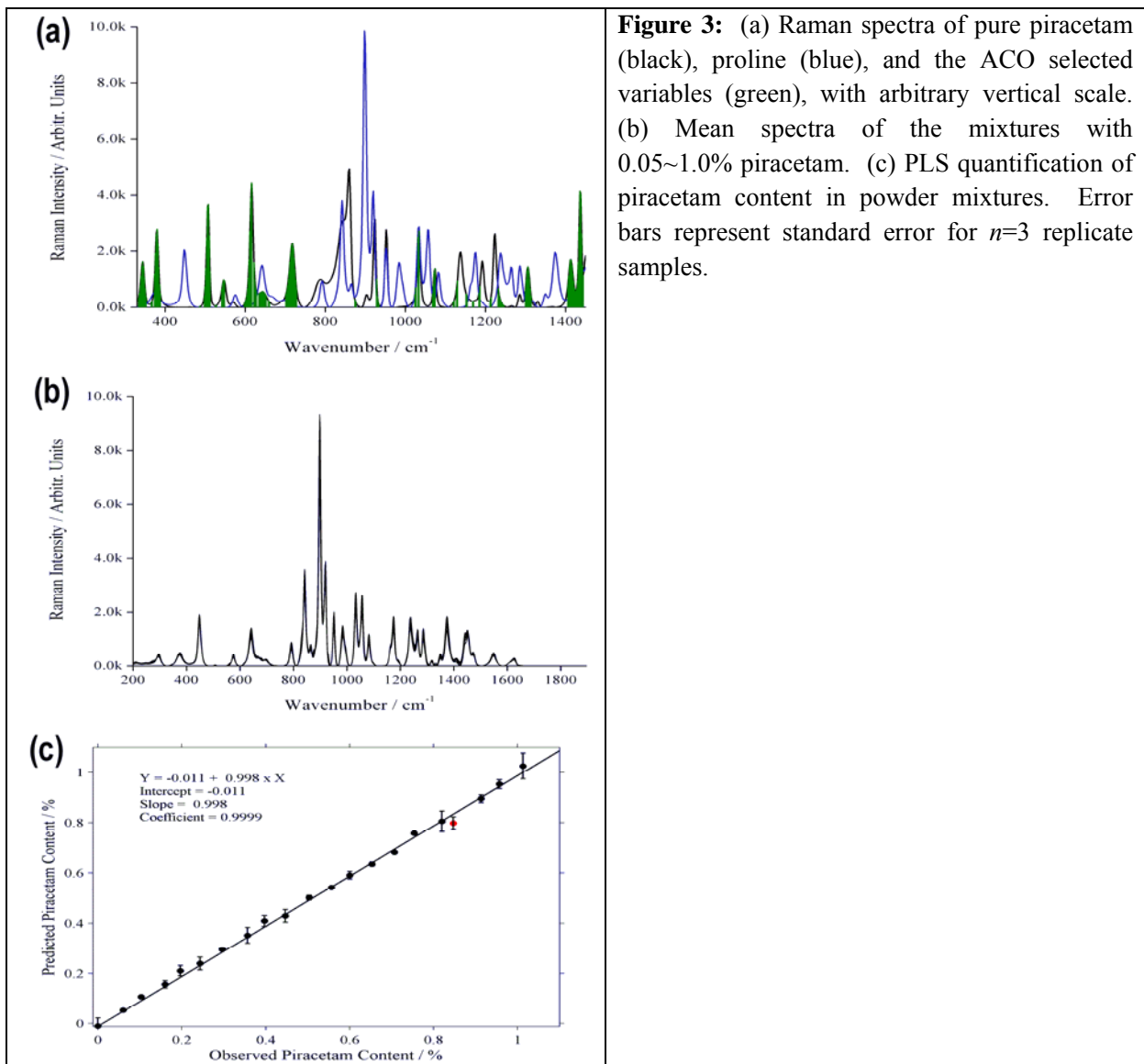
Channel2	1.65/1.86	0.039/0.041	0.14/0.16	0.90/1.04	0.18/0.27
Channel3	1.59/1.81	0.036/0.038	0.17/0.21	0.78/0.89	0.20/0.25
Channel4	1.63/1.85	0.036/0.039	0.15/0.18	0.86/0.99	0.19/0.26
Channel5	1.63/1.85	0.036/0.039	0.15/0.18	0.85/0.98	0.20/0.27
Channel6	1.60/1.81	0.034/0.036	0.15/0.17	0.84/0.96	0.22/0.36
Channel7	1.62/1.83	0.036/0.038	0.16/0.19	0.84/0.95	0.20/0.31
Channel8	1.70/1.91	0.042/0.045	0.15/0.17	0.95/1.09	0.19/0.27
Channel9	1.67/1.88	0.037/0.040	0.15/0.18	0.87/0.99	0.19/0.25
Channel10	1.66/1.88	0.031/0.033	0.16/0.18	0.86/0.99	0.16/0.30
<b>Mean value</b>	<b>1.64/1.85</b>	<b>0.035/0.038</b>	<b>0.15/0.18</b>	<b>0.86/0.99</b>	<b>0.19/0.28</b>
Standard Dev.	0.033/0.032	0.004/0.004	0.009/0.015	0.044/0.053	0.016/0.034
<b>REC%/RECV%</b>	<b>5.86/6.61</b>	<b>6.95/7.54</b>	<b>2.03/2.43</b>	<b>1.81/2.09</b>	<b>0.20/0.30</b>

Our alternate approach avoided these issues (Figure 2c). We first used the reasonably accurate 0-100% concentration range, individual channel PLS models (*Model-1*) to estimate a local piracetam concentration for each of the 841×10 sampling point-channel combinations (~8400 spectra) in each sample Raman map. Then according to this estimated value, the appropriate segmented concentration range PLS model was selected (*i.e.*, 0~2.5 *Model-2*, 2.5~21.5 *Model-3*, etc. Table 1) and piracetam content re-estimated more accurately at each point. Each Raman map thus generated ~8400 piracetam predictions, which were then averaged to give the final accurate piracetam content prediction. Therefore the final concentration prediction for piracetam was an average of the predictions produced by the 40 segmented range PLS models, which used the appropriately pre-processed data (See SI, section S4 for a step-by-step description of the process).

Using this approach, the calibration models were able to accurately quantify piracetam content over the full concentration range for all 61 samples (Figure 3c and Figures S-4, SI). Furthermore, since the final predictions were made on the constituent (*i.e.* independent) Raman spectra and, the calibration models-1~5 were built using averaged data ( $n > 8400$ ), we can consider the prediction samples as being a quasi-independent test set.

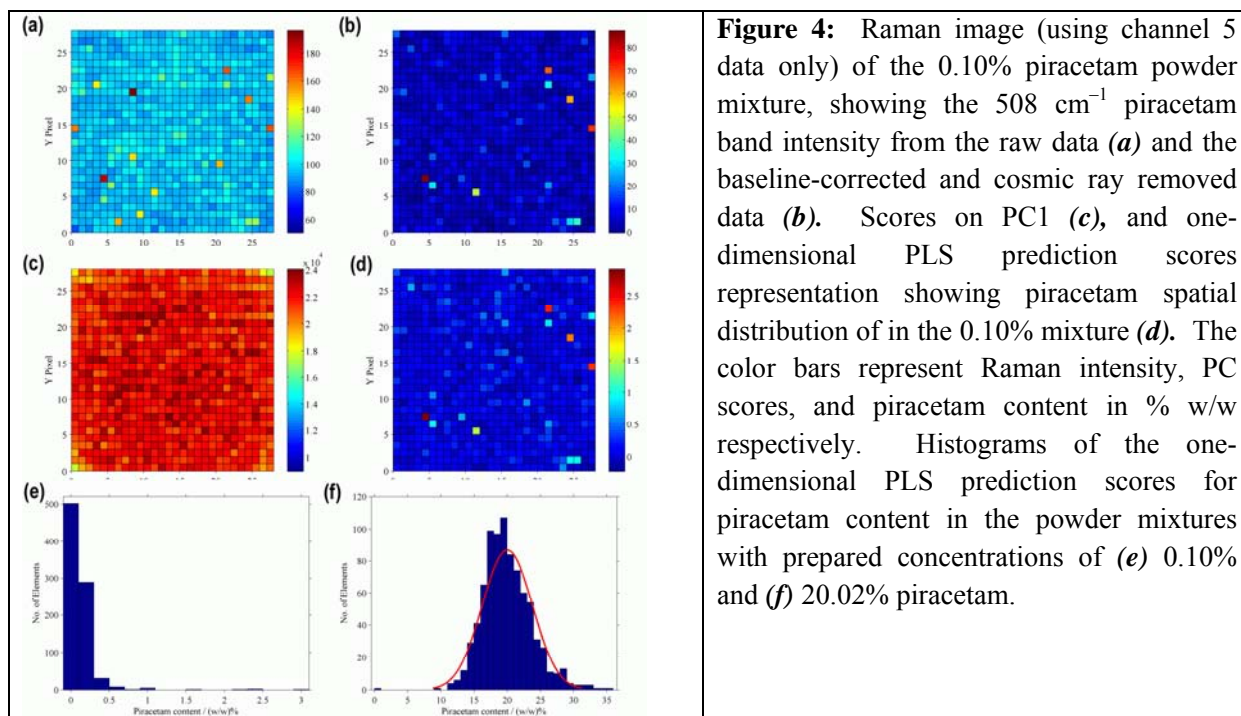
For the low analyte concentration (<1%) samples (Figure 3b-c), prediction accuracy for the triplicate measurements was particularly good. Only one sample (0.85% piracetam mixture) was mis-predicted, at 0.80% piracetam content (relative prediction error of 5.8%). This singular failure was caused by uncontrolled proline hydration (see SI, section S5). Excluding this sample, the RMSEP of the remaining samples with 0.05~1.0% piracetam content was 0.012%, and a very REP% of 2.43% was achieved. Moreover, prediction variations between triplicate Raman measurements were small (Figure 3b) and the  $R^2$  coefficient indicated an excellent model fit (Figure 3c). Prediction accuracy was far better than that attainable using the averaged spectra macro-PLS models (Table S-1), and just as important it covered the complete concentration range. For higher piracetam (>1%) concentration samples there was excellent agreement with HPLC validation measurements (SI, section S6). The LOD of this Raman method, calculated

from the triplicate spectra of the 0% piracetam sample (a total of 2523 physically sampled points), was calculated as spanning a range of 0.033~0.056%.<sup>27</sup> The HPLC method developed to validate the Raman method had an LOD of 0.0123 mg/100 mL, equivalent to 0.041% w/w. However, HPLC had several significant negative operational issues, namely significant day-to-day variability in terms of retention time variation, quantification reproducibility (see SI for details), and relatively extensive sample handling procedures.



**Figure 3:** (a) Raman spectra of pure piracetam (black), proline (blue), and the ACO selected variables (green), with arbitrary vertical scale. (b) Mean spectra of the mixtures with 0.05~1.0% piracetam. (c) PLS quantification of piracetam content in powder mixtures. Error bars represent standard error for  $n=3$  replicate samples.

**Characterization of spatial distribution.** In most common chemical imaging applications, one is looking for clear pixel-to-pixel spectral discrimination in order to identify the presence of impurities. One often depends on impurities being present as discrete particles (*i.e.*, of 100% composition). For low-content solid mixtures, this may not always be the case, such as for nearly homogeneous milled powders, where the impurity (or API) distribution will be relatively homogeneous on the macro-scale. However, if there is some variability on the micro-scale then one often needs to implement chemometric modelling to visualize the impurities and/or heterogeneity. Obviously, in the low-content samples here, the vast majority of the Raman image/map corresponded to the proline matrix and piracetam signal was correspondingly very weak. A common approach is to take a single unique analyte peak (here the unique piracetam  $508\text{ cm}^{-1}$  band) and plot its intensity for each pixel. However, for low-content measurements (here a 0.1% mixture) the result may not be very informative (Figure 4a). Application of automated spectral pre-processing (Figure 4b) improved discrimination dramatically, and gave a relatively homogeneous distribution with several clear hotspots of apparently high piracetam concentrations. One would next like to know if these hotspots represented discrete, pure piracetam particles (within the limits of the optical resolution of the Raman system) or if they represent localized regions of higher piracetam concentration arising from sample preparation heterogeneity.



Using simple PCA and then plotting the PC scores onto the image pixels,<sup>74</sup> (Figure 4c/d, PC2 not shown) did not generate true representations of piracetam spatial distribution. This was despite the fact that two PCs explained the spectral data variance, but with low-content mixtures nearly all the variance arose from the matrix material, here PC1 (Figure 2) explained 99.97% of data variance and was very similar to proline. PC2 (0.01% explained variance) was mostly hydrated proline rather than a pure piracetam signal (the higher PCs were essentially noise and very different to piracetam (data not shown)).

However, when the PLS prediction scores were plotted (Figure 4d), piracetam appeared to be evenly, but very sparsely distributed across the sample which verified that the sample preparation procedure generated reasonably homogeneous mixtures. Only five pixels (*i.e.*, hot spots) of clearly higher localized piracetam concentration could be identified, but there were another 12-20 pixels with slightly elevated concentrations. Statistical analysis of the PLS model predictions indicated that: no pixel contained pure piracetam, 791 of 841 pixels had a piracetam content below <0.3%, 31 pixels gave 0.3–0.5% piracetam content, and one pixel had a predicted concentration of 2.91% (Figure 4e). Another important consideration with Raman image analysis is that, due to particle size to sampled volume ratio, the percentage of pixels assigned to a specific component did not correspond exactly to the component weight percentage in the mixture. If however, one assumed otherwise then an inaccurate assessment of sample heterogeneity would result. This can be illustrated by looking at another mixture with higher piracetam content (20.018% prepared concentration), where the histogram (Figure 4f) of the one-dimensional PLS prediction scores was statistically centered at 20.02%. Most pixels had predicted piracetam concentrations in the 11~30% range, with individual outlier pixels of 0 and 35.95%. These results show that this quantification method can also be used to deliver rapid assessment of solid-state homogeneity, for example in the milling and/or freeze drying of low-content API formulations.

## CONCLUSIONS

We have demonstrated a Raman spectroscopy based method (Figure 2c) for the accurate and robust quantification of the minor analyte in solid mixtures by the use of a combination of sub-sampling, chemometric and statistical based approaches. The novelty of this approach to quantification is that it takes the opposite approach to sampling that is commonly implemented for quantitative analysis by Raman spectroscopy, namely that one seeks to average out the heterogeneity in the Raman signal.<sup>4, 75-76</sup> The method demonstrated high accuracy over a very wide concentration range (0.05~100%) with a LOD range of 0.033~0.056%. LODs in this range or close to this should be feasible with any combination of solid-state materials where there is some significant Raman spectral difference between analyte and host matrix. For example, we had previously demonstrated that it was feasible to quantify piracetam polymorph mixtures using conventional Raman and chemometrics with accuracies comparable to Model-1 above.<sup>10</sup> Implementation of this method therefore, would enable accurate polymorph quantification at

<0.1% w/w levels for example to assess contamination in seed crystal batches. These levels far exceed that possible with conventional powder X-ray diffraction.

This approach to low-content, solid-state quantification was competitive with HPLC in terms of LOD and accuracy, however, it required much less sample handling, user intervention, and was potentially fully automatable for high-throughput screening and quantification testing. In addition, it was significantly more robust since the Raman data collected over several months, showed no significant day-to-day variation whereas HPLC measurements had significant variation (retention times and reproducibility). Furthermore, HPLC cannot be used for polymorph quantification.

While we recognize that the method was computationally intensive, all steps can be implemented in a fully automated fashion. This enables the method to be applied with minimal subjective user intervention for both gross quantification and the assessment of solid-state heterogeneity to a wide variety of industrial process operations.

### Acknowledgements

Research undertaken as part of the Synthesis and Solid State Pharmaceutical Centre, funded by Science Foundation Ireland and industry partners, and Enterprise Ireland (Grant No: TC-2012-5106). Kaiser Optical Systems, Inc. (Ann Arbor, MI) and Mr. Harry Owen are thanked for the loan of the Raman instrumentation. We thank Ms. Caroline Janzen for sample preparation and data collection.

### Supporting Information Available

Additional information as noted in text. This material is available free of charge via the Internet at <http://pubs.acs.org>

### References

1. Shah, R. B., Tawakkul, M. A., Khan, M. A., *J. Pharm. Sci.* **2007**, *96*, 1356-1365.
2. Fini, G., *J. Raman Spectrosc.* **2004**, *35*, 335-337.
3. Reich, G., *Advanced Drug Delivery Reviews* **2005**, *57*, 1109-1143.
4. Gowen, A. A., O'Donnell, C. P., Cullen, P. J., Bell, S. E., *European Journal of Pharmaceutics and Biopharmaceutics* **2008**, *69*, 10-22.
5. Fevotte, G., *Chemical Engineering Research and Design* **2007**, *85*, 906-920.
6. Rantanen, J., *J. Pharm. Pharmacol.* **2007**, *59*, 171-177.
7. Vippagunta, S. R., Brittain, H. G., Grant, D. J. W., *Advanced Drug Delivery Reviews* **2001**, *48*, 3-26.
8. Bond, A. D., *Curr Opin Solid St M* **2009**, *13*, 91-97.
9. Bauer, J., Spanton, S., Henry, R., Quick, J., Dziki, W., Porter, W., Morris, J., *Pharmaceutical Research* **2001**, *18*, 859-866.
10. Hennigan, M. C., Ryder, A. G., *J. Pharm. Biomed. Anal.* **2013**, *72*, 163-171.

11. Croker, D. M., Hennigan, M. C., Maher, A., Hu, Y., Ryder, A. G., Hodnett, B. K., *J. Pharm. Biomed. Anal.* **2012**, *63*, 80-86.
12. Sasic, S., Mehrens, S., *Anal. Chem.* **2012**, *84*, 1019-1025.
13. Spencer, J. A., Kauffman, J. F., Reepmeyer, J. C., Gryniwicz, C. M., Ye, W., Toler, D. Y., Buhse, L. F., Westenberger, B. J., *J. Pharm. Sci.* **2009**, *98*, 3540-3547.
14. Zhang, L., Henson, M. J., Sekulic, S. S., *Anal. Chim. Acta* **2005**, *545*, 262-278.
15. Henson, M. J., Zhang, L., *Appl. Spectrosc.* **2006**, *60*, 1247-1255.
16. Sasic, S., Clark, D. A., *Appl. Spectrosc.* **2006**, *60*, 494-502.
17. Sasic, S., *Pharmaceutical Research* **2007**, *24*, 58-65.
18. Sasic, S., *Anal. Chim. Acta* **2008**, *611*, 73-79.
19. Lin, H. S., Marjanovic, O., Lennox, B., Sasic, S., Clegg, I. M., *Appl. Spectrosc.* **2012**, *66*, 272-281.
20. Rantanen, J., Wikstrom, H., Rhea, F. E., Taylor, L. S., *Appl. Spectrosc.* **2005**, *59*, 942-951.
21. Shin, K., Chung, H., *Analyst* **2013**, *138*, 3335-3346.
22. Bell, S. E. J., Beattie, J. R., McGarvey, J. J., Peters, K. L., Sirimuthu, N. M. S., Speers, S. J., *J. Raman Spectrosc.* **2004**, *35*, 409-417.
23. Sasic, S., Whitlock, M., *Appl. Spectrosc.* **2008**, *62*, 916-921.
24. Wold, S., Sjostrom, M., Eriksson, L., *Chemometr Intell Lab* **2001**, *58*, 109-130.
25. Shamsipur, M., Zare-Shahabadi, V., Hemmateenejad, B., Akhond, M., *J. Chemometr.* **2006**, *20*, 146-157.
26. Allegrini, F., Olivieri, A. C., *Anal Chim Acta* **2011**, *699*, 18-25.
27. Allegrini, F., Olivieri, A. C., *Anal Chem* **2014**, *86*, 7858-7866.
28. Geladi, P., *Spectrosc. Acta Pt. B-Atom. Spectr.* **2003**, *58*, 767-782.
29. de Faria, D. L. A., de Souza, M. A., *J. Raman Spectrosc.* **1999**, *30*, 169-171.
30. Millen, R. P., Temperini, M. L. A., de Faria, D. L. A., Batchelder, D. N., *J. Raman Spectrosc.* **1999**, *30*, 1027-1033.
31. Oshima, Y., Komachi, Y., Furihata, C., Tashiro, H., Sato, H., *Appl. Spectrosc.* **2006**, *60*, 964-970.
32. Osticioli, I., Zoppi, A., Castellucci, E. M., *J. Raman Spectrosc.* **2006**, *37*, 974-980.
33. Osticioli, I., Zoppi, A., Castellucci, E. M., *Appl. Spectrosc.* **2007**, *61*, 839-844.
34. P. Vandenberghe, *Practical Raman Spectroscopy: An Introduction*, . John Wiley & Sons: Chichester, 2013.
35. Jirasek, A., Schulze, G., Yu, M. M. L., Blades, M. W., Turner, R. F. B., *Appl. Spectrosc.* **2004**, *58*, 1488-1499.
36. Mazet, V., Carteret, C., Brie, D., Idier, J., Humbert, B., *Chemometr. Intell. Lab. Syst.* **2005**, *76*, 121-133.
37. Leger, M. N., Ryder, A. G., *Appl. Spectrosc.* **2006**, *60*, 182-193.
38. Zhao, J., Lui, H., McLean, D. I., Zeng, H., *Appl. Spectrosc.* **2007**, *61*, 1225-1232.
39. Gan, F., Ruan, G. H., Mo, J. Y., *Chemometr. Intell. Lab. Syst.* **2006**, *82*, 59-65.

40. Brown, C. D., Vega-Montoto, L., Wentzell, P. D., *Appl. Spectrosc.* **2000**, *54*, 1055-1068.
41. Liu, Y., Cai, W. S., Shao, X. G., *Chemometr. Intell. Lab. Syst.* **2013**, *125*, 11-17.
42. Zhang, Z.-M., Chen, S., Liang, Y.-Z., Liu, Z.-X., Zhang, Q.-M., Ding, L.-X., Ye, F., Zhou, H., *J. Raman Spectrosc.* **2010**, *41*, 659-669.
43. Hu, Y., Jiang, T., Shen, A., Li, W., Wang, X., Hu, J., *Chemometr. Intell. Lab. Syst.* **2007**, *85*, 94-101.
44. Camerlingo, C., Zenone, F., Gaeta, G. M., Riccio, R., Lepore, M., *Measurement Science & Technology* **2006**, *17*, 298-303.
45. Ramos, P. M., Ruisanchez, I., *J. Raman Spectrosc.* **2005**, *36*, 848-856.
46. O'Grady, A., Dennis, A. C., Denvir, D., McGarvey, J. J., Bell, S. E. J., *Anal. Chem.* **2001**, *73*, 2058-2065.
47. Zhang, D. M., Ben-Amotz, D., *Appl. Spectrosc.* **2000**, *54*, 1379-1383.
48. Zhang, Z.-M., Chen, S., Liang, Y.-Z., *Analyst* **2010**, *135*, 1138-1146.
49. Li, Z., Zhan, D.-J., Wang, J.-J., Huang, J., Xu, Q.-S., Zhang, Z.-M., Zheng, Y.-B., Liang, Y.-Z., Wang, H., *Analyst* **2013**, *138*, 4483-4492.
50. Cadusch, P. J., Hlaing, M. M., Wade, S. A., McArthur, S. L., Stoddart, P. R., *J. Raman Spectrosc.* **2013**, *44*, 1587-1595.
51. Schulze, H. G., Turner, R. F. B., *Appl. Spectrosc.* **2014**, *68*, 185-191.
52. Zhang, L., Henson, M. J., *Appl. Spectrosc.* **2007**, *61*, 1015-1020.
53. Phillips, G. R., Harris, J. M., *Anal. Chem.* **1990**, *62*, 2351-2357.
54. Hill, W., Rogalla, D., *Anal. Chem.* **1992**, *64*, 2575-2579.
55. Takeuchi, H., Hashimoto, S., Harada, I., *Appl. Spectrosc.* **1993**, *47*, 129-131.
56. Ehrentreich, F., Summchen, L., *Anal. Chem.* **2001**, *73*, 4364-4373.
57. Behrend, C. J., Tarnowski, C. P., Morris, M. D., *Appl. Spectrosc.* **2002**, *56*, 1458-1461.
58. Zhang, D., Hanna, J. D., Ben-Amotz, D., *Appl. Spectrosc.* **2003**, *57*, 1303-1305.
59. Katsumoto, Y., Ozaki, Y., *Appl. Spectrosc.* **2003**, *57*, 317-322.
60. Zhao, J., *Appl. Spectrosc.* **2003**, *57*, 1368-1375.
61. Cappel, U. B., Bell, I. M., Pickard, L. K., *Appl. Spectrosc.* **2010**, *64*, 195-200.
62. Chew, W., *J. Raman Spectrosc.* **2011**, *42*, 36-47.
63. Esmonde-White, F. W. L., Esmonde-White, K. A., Morris, M. D., *Appl. Spectrosc.* **2011**, *65*, 85-98.
64. Jones, H. D. T., Haaland, D. M., Sinclair, M. B., Melgaard, D. K., Collins, A. M., Timlin, J. A., *Chemometr. Intell. Lab. Syst.* **2012**, *117*, 149-158.
65. Li, S., Dai, L. K., *Appl. Spectrosc.* **2011**, *65*, 1300-1306.
66. Holland, P. W., Welsch, R. E., *Communications in Statistics Part a-Theory and Methods* **1977**, *6*, 813-827.
67. Wu, W., Massart, D. L., deJong, S., *Chemometr. Intell. Lab. Syst.* **1997**, *36*, 165-172.
68. Engel, J., Gerretzen, J., Szymanska, E., Jansen, J. J., Downey, G., Blanchet, L., Buydens, L. M. C., *Trac-Trends Anal. Chem.* **2013**, *50*, 96-106.
69. Vidal, M., Amigo, J. M., *Chemometr. Intell. Lab. Syst.* **2012**, *117*, 138-148.

70. Fearn, T., Riccioli, C., Garrido-Varo, A., Guerrero-Ginel, J. E., *Chemometr. Intell. Lab. Syst.* **2009**, *96*, 22-26.
71. Haaland, D. M., Thomas, E. V., *Anal Chem* **1988**, *60*, 1193-1202.
72. Wold, S., Antti, H., Lindgren, F., Ohman, J., *Chemometr. Intell. Lab. Syst.* **1998**, *44*, 175-185.
73. Martens, H. A., Dardenne, P., *Chemometr. Intell. Lab. Syst.* **1998**, *44*, 99-121.
74. Shinzawa, H., Awa, K., Kanematsu, W., Ozaki, Y., *J. Raman Spectrosc.* **2009**, *40*, 1720-1725.
75. Johansson, J., Sparen, A., Svensson, O., Folestad, S., Claybourn, M., *Appl. Spectrosc.* **2007**, *61*, 1211-1218.
76. Strachan, C. J., Rades, T., Gordon, K. C., Rantanen, J., *J. Pharm. Pharmacol.* **2007**, *59*, 179-192.