Electronic Supplementary Information

Chemometric approaches to low-content quantification (LCQ) in solid-state mixtures using Raman mapping spectroscopy.

Boyan Li, Yannick Casamayou-Boucau, Amandine Calvet, and Alan G. Ryder*
Nanoscale Biophotonics Laboratory, School of Chemistry, NUI Galway, University Road, Galway, Ireland.
Email: alan.ryder@nuigalway.ie

SUPPLEMENTAL INFORMATION:

Contents

S1. Data pre-processing

Crucial and minimal data pre-processing was performed to facilitate the extraction of chemical information of the Raman mapping data, and thus to improve the model accuracy in the subsequent data analysis procedure. The pre-processing methods included: (i) morphological weighted penalized least-squares (MPLS) [1] for mitigating baseline artefact of spectra as baseline arising from particle size effects, sample surface roughness and unwanted shot noise could obscure the Raman signal identification and quantification of low-content analytes, (ii) kernel principal component analysis residual diagnosis (KPCARD) [2, 3] for removal of detrimental spikes caused by cosmic ray events, (iii) spectrum exclusion [2] to reject those with intensities 70% lower than the total average spectral intensity of each Raman mapping measurement because of mapping edge effect, (iv) multiplicative scatter correction (MSC) [4-6] and standard normal variate (SNV) [4-6] to reduce scattering variations between spectra or measurements for the PLS modelling procedure, (v) spectrum normalization to scale each individual spectrum in the predefined variable region to unit area under the spectrum curve, and (vi) variable selection using the ant colony optimization (ACO) method [7, 8].

S2. Description of chemometric methods

Conventional notation for variables has been adopted throughout this paper: underlined uppercase boldface letters for three-way arrays (*e.g.*, $\underline{\mathbf{D}}_{X\times Y\times\lambda}$), and uppercase boldface letters for two-way matrices (*e.g.*, **D**). Lowercase boldface characters denote vectors (*e.g.*, **s**), italicized subscript characters for vector index (as $\mathbf{t}_k$), and lowercase italicized letters for scalars (as f_k). Superscripts were assigned as follows: T, vector or matrix transpose; –1, matrix inverse; and +, pseudoinverse of the non-square matrix of an overdetermined system (in this case one in which the number of spectral variables are far in excess of the number of samples). Matrix and vector Frobenius norms (*i.e.*, 2-norms) are indicated by bracing the quantity (*e.g.*, $\left\|\mathbf{c}_k\mathbf{s}_k^{\mathrm{T}}\right\|$).

Each Raman mapping measurement generated a three-way array ($\underline{\mathbf{D}}_{X\times Y\times\lambda}$), where X and Y respectively denote the spatial co-ordinates of each pixel and λ refers to the spectral pattern along the wavenumber axis for the xy^{th} pixel [9]. These three-way datasets ($\underline{\mathbf{D}}_{X\times Y\times\lambda}$) were then unfolded into two-way matrices, $\mathbf{D}_{XY\times\lambda}$, to enable bilinear Beer-Lambert Law models to be developed. A brief theoretical background of the methods used in this study is given here, but readers could look for more information in the supplied references.

S2.1. BR-PCHIP

A simple univariate approach can be used to quantitatively analyze the constituent in spectroscopic measurements by means of the signal at selective wavenumbers, which is specific for the constituent, *e.g.*, the characteristic bands or the ratio between bands, and among others [10-12]. These characteristic bands and/or their ratio require to be free of interference from other constituents and to necessitate if and where the constituent is present in spectroscopic measurements. In the context of the investigated mixtures, the piracetam band at 1652 cm^{-1} and proline band at 448 cm^{-1} were both used for calculating their intensity ratios in each spectrum, as is:

$$r_{pp} = \frac{I_{1652}}{I_{448}} \qquad (1)$$

The 1652 and 448 cm^{-1} bands were selected because they were unique to the components, non-overlapping, and minimally affected by the baseline. The use of a band ratio, instead of an individual analyte band intensity, not only reduced the influence of measurement error but also mitigated multiplicative scattering effects between spectra.

Given n band ratios (r_{pp_l}, $l = 1, ..., n$) corresponding to the distinct piracetam content (y_l) in n sample spectra, there is a unique shape-preserving piecewise cubic Hermite interpolating polynomial (PCHIP) [13-15]:

$$P(r_{pp\_l}) = y_l,\ l = 1, ..., n \qquad (2)$$

This polynomial was able to satisfy interpolation conditions on the lth interval ($r_{pp\_l+1} - r_{pp\_l}$), hence either exactly producing or approximating the given data (r_{pp_l}, y_l). The band ratios not fulfilling a linear relationship with the piracetam content in the sample accounted for the deployment of PCHIP for the curve fitting purpose.

This univariate analysis method was simple, very computationally efficient, and reduced modelling complexity, however, it excluded a lot of the spectral information contained in the data. In more complex samples/situations, constituent-specific spectral bands free of interferences may not often be available and therefore, multivariate methods are instead applied, using regions of spectral data to analyze Raman mapping data through two-way matrix or three-way data array [2, 9, 16-24].

S2.2. PLS

The partial least squares (PLS) [25] method relates a spectral matrix (**D**) with the dependent response variables (**y**, *e.g.*, piracetam content) in an indirect linear formulation:

$$\mathbf{y} = \mathbf{D}\mathbf{b}_{PLS} + \mathbf{e} \qquad (3)$$

through decomposing:

$$\mathbf{D} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \qquad (4)$$

$$\mathbf{y} = \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{e}^{*} \qquad (5)$$

where $\mathbf{b}_{PLS}$ is called the inner-relationship coefficients and can be specified by identifying r underlying factors (or latent variables, LVs) that explicitly maximize covariance (say $\mathbf{D}^{\mathrm{T}}\mathbf{y}$) between **D** and **y**. Matrix **T** holds the factors ($\mathbf{t}_k$, $k=1,..., r$), and **U** corresponds to scores ($\mathbf{u}_k$, $k=1,..., r$) for **y**, while **P** and **Q** individually have their loadings ($\mathbf{p}_k$ and $\mathbf{q}_k$). **e**, **E**, and $\mathbf{e}^{*}$ are the model errors or residuals.

There are multiple algorithms available to extract PLS factors and these are all based on iterative calculations. For example, the eigenvalue decomposition algorithm extracts PLS factors from the first up to the *r*th ($k=1,..., r$) successively in three stages: (1) iterative estimation of factor scores ($\mathbf{t}_k$ and $\mathbf{u}_k$) and inner weight ($\mathbf{w}_k$), (2) estimation of outer weights ($\mathbf{q}_k$) and loadings ($\mathbf{p}_k$), and path coefficients, and (3) obtainment of regression coefficients $\mathbf{b}_{PLS}$:

$$\mathbf{b}_{PLS} = \mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{Q}^{\mathrm{T}} \qquad (6)$$

where **W** stands for the weights ($\mathbf{w}_k$, $k=1,..., r$).

Clearly, PLS calibration does not require that the individual spectra of each analyte and interference be known in advance. However, a sample set spanning an appropriate sample variance range (*e.g.*, concentration, physical properties, *etc.*) is required to build a calibration model, and then to predict for new samples. Unfortunately, when samples are complex, *e.g.*, multicomponent pharmaceutical formulations, it becomes difficult to obtain sufficiently comprehensive calibration sets, which limits the practical use PLS based methods.

S2.3. NAS-CLS

Classical least squares (CLS) [26, 27] is based on the Beer-Lambert law and assumes that a sample spectrum is made up of linearly independent signals weighted by the concentrations of individual spectrally-active components/analytes, and can be formulated as:

$$\mathbf{d} = \mathbf{c}\mathbf{S}^{\mathrm{T}} + \mathbf{e} \qquad (7)$$

where **d** is one measured spectrum of the data matrix **D**, **S** is the spectrum matrix consisting of independent untainted signals and cyclic noise of each pure chemical component corresponding to unit concentration, with dimensions of number of spectroscopic measurement domain (λ) by *r* components. **c** stands for the concentration weights of components, representing the degree to which each component contributes to the overall measurement. Given the measurement **d** and reference spectra **S**, the concentration **c** can be estimated by the least square approximation,

$$\hat{\mathbf{c}} = \mathbf{dS}(\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1} \qquad (8)$$

The critical is that all the spectrally-active components (**S)** must be independent and available so that **c** can be accurately estimated. However, in most practical cases where **S** is usually unknown, this becomes quite difficult or even impossible. Therefore, the net analyte signal (NAS)-based calibration has been applied for solving the problem.

NAS [28, 29] aims to discriminate the measured spectrum **d** into two different contributions: one stemming from the analyte *k* of interest ($\mathbf{d}_k$), and all the remaining information from other sources of variability ($\mathbf{d}_{-k}$), which not only includes interference but also residual not explained by the model:

$$\mathbf{d} = \mathbf{d}_k + \mathbf{d}_{-k} \qquad (9)$$

This equation can be written in a multi-wavelength and multi-sample embodiment in the matrix form:

$$\mathbf{D} = \mathbf{D}_k + \mathbf{D}_{-k} = \mathbf{c}_k\mathbf{s}_{\mathrm{k}}^{\mathrm{T}} + \mathbf{D}_{-k} \qquad (10)$$

$\mathbf{s}_k$ and $\mathbf{c}_k$ correspond to the sensitivity vector of the analyte *k* and its concentrations in each of spectra **D**, respectively. The signal contributions of all the other components, except for that from the analyte *k* ($\mathbf{D}_k$), give rise to $\mathbf{D}_{-k}$.

If a NAS filtering $\mathbf{F}_{NAS}$ can be defined, then $\mathbf{D}_{-k}$ is able to be removed from **D**:

$$\mathbf{F}_{NAS} = \mathbf{I} - \mathbf{D}_{-k}^{+}\mathbf{D}_{-k} \qquad (11)$$

$$\mathbf{DF}_{NAS} = (\mathbf{c}_k\mathbf{s}_k^{\mathrm{T}} + \mathbf{D}_{-k})\mathbf{F}_{NAS} = \mathbf{c}_k\mathbf{s}_k^{\mathrm{T}}\mathbf{F}_{NAS} + \mathbf{E} \qquad (12)$$

E indicates the model error, **I** means an identity matrix, and $\mathbf{D}_{-k}^{+}$ is the pseudoinverse of $\mathbf{D}_{-k}$. Owing to there is no *a priori* knowledge about $\mathbf{D}_{-k}$, a common alternative is to obtain a matrix that can account for as much of the possible variability in $\mathbf{D}_{-k}$:

$$\mathbf{D}_{-k} = \mathbf{D} - \mathbf{c}_k\mathbf{s}_k^{\mathrm{T}} = [\mathbf{I} - \mathbf{c}_k(\mathbf{c}_k^{\mathrm{T}}\mathbf{c}_k)^{-1}\mathbf{c}_k^{\mathrm{T}}]\mathbf{D} \qquad (13)$$

Thus, the filter $\mathbf{F}_{NAS}$ is computed for the NAS-CLS method, and the concentrations of the analyte *k* in new samples can be then predicted.

S2.4. PCA-CLS

Principal component analysis (PCA) [30] combined with CLS provided another route to perform indirect quantitative calibration. Based on Equation (4), the spectra **D** can be decomposed to yield *r* orthogonal factors or principal components (PCs) in **P**. Then, with these *r* PCs, the reference spectra **S** of the pure chemical components in Equation (8) were substituted to estimate the concentration **C**:

$$\hat{\mathbf{C}} = \mathbf{DS}(\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1} = \mathbf{DP}(\mathbf{P}^{\mathrm{T}}\mathbf{P})^{-1} \qquad (14)$$

These *r* PCs are not identical to the spectra of the pure chemical components in the sample, however, for a given analyte, its relative concentration distribution in the estimated concentration $\hat{\mathbf{C}}$ may be proportional to the actual distribution in the absolute concentration **C** associated with **S**. Using a set of samples with known concentrations, it was possible to perform least squares to calibrate the estimated $\hat{\mathbf{C}}$ for the analyte of interest.

S2.5. MCR-BANDS

Multivariate curve resolution (MCR) refers to a family of self-modeling mixture analysis methods [31-33]. In general, with no *prior* knowledge about the mixture sample, MCR makes use of the bilinearity of the experimental data matrix (**D**), resolving pure chemical components (**S**) and their contribution profiles (**C**):

$$\mathbf{D} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E} = \sum_{k=1}^{r} \mathbf{c}_k \mathbf{s}_k^{\mathrm{T}} + \mathbf{E} \tag{15}$$

However, for accurate MCR decomposition two intrinsic problems have to be dealt with, *i.e.*, rotational and intensity ambiguities [34]. To address these problems, some powerful strategies have been adopted to assist the data resolution. On one hand, the incorporation of additional information concerning the sample when available can make the MCR solutions more physical meaningful, for instance, initializing the spectral estimate with known pure chemical compositions. On the other hand, the application of certain constraints such as non-negativity, unimodality, closure, or local rank and selectivity, *etc.* to the solution can lead to results closer to the true sources of data variation [35].

For the MCR analysis, the lack of model fit (LOF) between the obtained results and original data can be described in relative percentage terms by the expression:

$$\%\,\mathrm{LOF} = 100 \times \frac{\sum_{xy,\lambda} e_{xy\lambda}^2}{\sum_{xy,\lambda} d_{xy\lambda}^2} \tag{16}$$

where $d_{xy\lambda}$ and $e_{xy\lambda}$ respectively stand for the $xy\lambda^{\mathrm{th}}$ element in the experimental spectra $\mathbf{D}_{XY\times\lambda}$ and the residual associated with the reproduction of $d_{xy\lambda}$ by the MCR model. The explained variance can be indicated as (100−%LOF) and both two parameters have been utilized to measure the quality of MCR model fit.

MCR-BANDS [35, 36] evaluated the degree of rotation ambiguities associated with the MCR solution, based on the calculation of the relative signal contribution of each component in the mixture:

$$f_k = \frac{\left\| \mathbf{c}_k \mathbf{s}_k^{\mathrm{T}} \right\|}{\left\| \mathbf{C}\mathbf{S}^{\mathrm{T}} \right\|} \tag{17}$$

where f_k is a scalar value that defines the relative contribution of the *k*th particular component ($\mathbf{c}_k \mathbf{s}_k^{\mathrm{T}}$) to the entire signal ($\mathbf{C}\mathbf{S}^{\mathrm{T}}$) considering of the *r* components in the measured mixture ($k = 1, \ldots, r$), by using the quotient of two Frobenius norms of them.

Under a set of constraints, the method looks for a particular set of profiles of $\mathbf{c}_k$ and $\mathbf{s}_k$ for the *k*th component within the band boundaries of feasible solutions and calculates the maximum and minimum values of the relative contribution function, giving f_k^{max} and f_k^{min} respectively. Then, the extent of rotational ambiguity can be evaluated by the difference between f_k^{max} and f_k^{min} values for each component ($k = 1, \ldots, r$). Appreciable difference would indicate the presence of rotational ambiguity. Otherwise, the MCR solutions are unique, and the obtainment of **S** is reliable. Therefore, the concentration profiles **C** can be used for quantitative calibration purpose with confidence, according to equation (14).

Unlike the PLS method that requires a calibration phase with a large set of comprehensive samples, MCR needs less calibration samples to scale the scores in the estimated concentration profiles **C** and thus may be preferable. This has to be taken into consideration in that a large set of calibration samples may not be always available due to sample complexity, overly capital and/or laborious cost, among others, particularly in the cases of pharmaceutical applications.

S3. Raman mapping analysis scheme

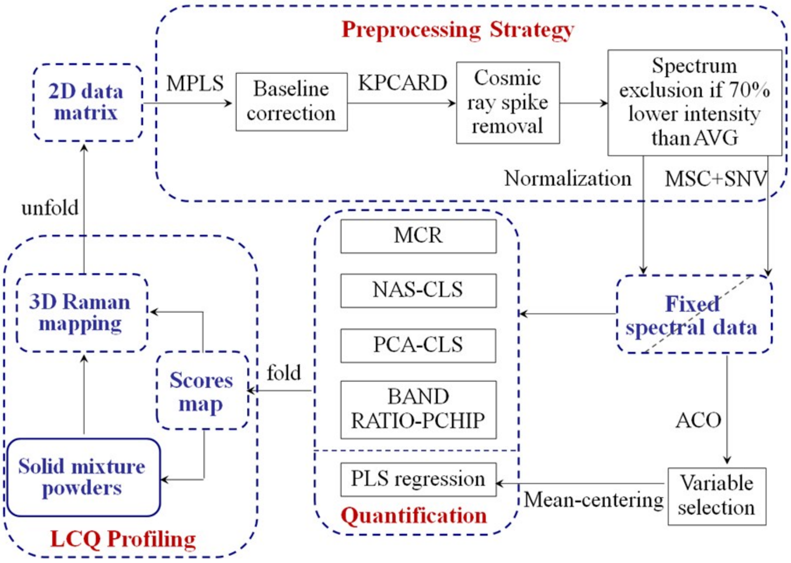


Figure S-1: Scheme of Raman mapping data analysis procedure in this study.

S4. PLS quantification Model2 with non-OSC treated data

PLS was implemented after assessing a variety of pre-processing methods, and combinations of methods. The best results were obtained when MSC, SNV, and ACO variable selection were applied. ACO was performed with the rate of pheromone evaporation = 0.65, ant number = 350, 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.[37] 142 variables were selected from the 200−1896 cm^{-1} range and using these variables, one PLS factor, and mean-centering (MC), 50 segmented piracetam concentration PLS quantification models were created for all ten spectral channels.

All channel-specific models gave quite similar RMSEC/RMSECV errors, even though they used Raman spectra with different SNRs [37]. This indicated that the pre-processing methods were suitable and optimized. For *Model2* (0−2.5%), orthogonal signal correction (OSC) [38] was also implemented after SNV and before MC because for the low-content samples, subtle spectral differences were convoluted with noise and small sampling variations. OSC was better able to minimize noise and contributions from the more intense proline signal in this case, leading to a more accurate low-content *Model2* with a mean REC% and RECV% of 6.94% and 7.52% respectively, which was a ~2-fold improvement compared to the non-OSC treated data (Table S-1, *SI*). In contrast to the best BR-PCHIP result, the LCQ accuracy was approximately 2.5-fold better.

The correlation coefficients between the pure spectra of piracetam, proline, and the PLS factors (LV$_{PLS}$) from *Model1* were also calculated. The results (taking channel 5 data as an example) showed that all the correlation coefficients between the piracetam spectrum and LV$_{PLS}$ for the 142 ACO-selected variables were equal or larger than 0.93. The comparison between the spectra of pure piracetam, proline, and the PLS factor in *Model1*, revealed that LV$_{PLS}$ mostly overlapped the piracetam spectrum. In contrast, the correlation coefficients between LV$_{PLS}$ and the proline spectrum were *ca.* −0.40 which proved that the selected variables were more descriptive of piracetam, and hence accurate piracetam prediction models were achieved.

Another important reason why PLS generated accurate LCQ was that there were a sufficiently large number of calibration samples available. Since predictive error is directly dependent on calibration set size, more and more samples are required to reduce prediction error [39], particularly for LCQ and this is not always feasible. There are also many practical difficulties with preparing calibration samples with precisely known low levels of analytes/contaminants. Therefore, we examined the feasibility of using NAS and MCR based approaches where smaller sized calibration sample sets might be employed, particularly for more complex mixtures with multiple low content components.

Table S-1: RMSEC/RMSECV values (in w/w%) obtained for the piracetam quantification models in the 0−2.5% piracetam content range by PLS method and non-OSC treated spectra for each spectrometer channel. Model accuracy was assessed by REC% and RECV% for calibration and cross-validation respectively.

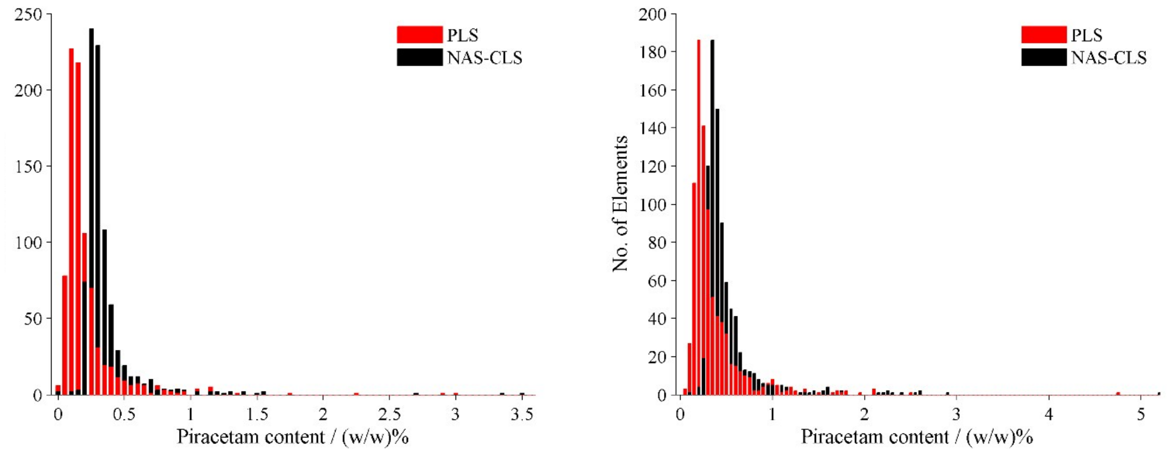
Piracetam	**PLS Model2**	
in 0−2.5%	RMSEC	RMSECV
Channel 1	0.063670	0.079980
Channel 2	0.058344	0.074055
Channel 3	0.058794	0.071869
Channel 4	0.052208	0.066314
Channel 5	0.063635	0.078263

Channel 6	0.068452	0.082948
Channel 7	0.057291	0.069536
Channel 8	0.060373	0.072486
Channel 9	0.058560	0.070503
Channel 10	0.057939	0.069599
mean value	0.060	0.074
standard dev.	0.004	0.005
REC%/RECV%	11.89	14.60

S5. Piracetam prediction by the NAS-CLS method.

The piracetam content in the mixtures was predicted by the NAS-CLS models. In the 0.05−1.0% range, the 0.197% and 0.357% samples were overestimated, giving large errors: the predicted piracetam concentrations were 0.354% and 0.502% respectively, which were highlighted with red solid circles. If excluding these two samples, then the RMSEP and REP became much smaller as were 0.03% and 7.06%, for the prediction of piracetam content in the 0.05−1.0% range.

The comparison of the piracetam content of these two mixture samples respectively predicted from the triplicate measurements for each sample by the NAS-CLS and PLS models shows that the NAS-CLS gave a higher prediction of piracetam content at almost every pixel. The correlation coefficients between the 841 NAS-CLS and PLS predictions of piracetam content were 0.977 for the 0.197% sample, and 0.984 for the 0.357% sample, respectively. The differences between the 841 NAS-CLS and PLS predictions of piracetam content were (0.144 ± 0.056)% for the 0.197% sample, and (0.149 ± 0.065)% for the 0.357% sample.



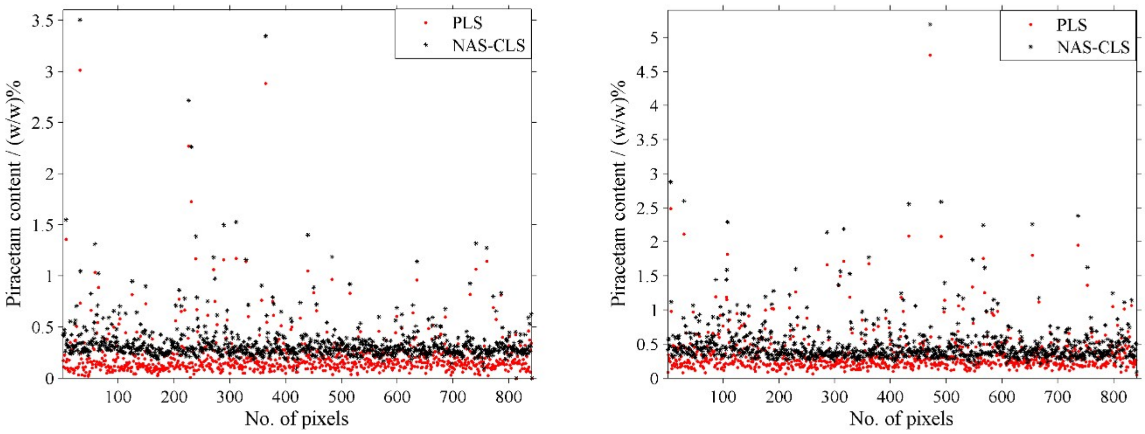


Figure S-2: Predictions of piracetam content at all the 841 pixels of Raman mapping measurement of (**left**) 0.197% and (**right**) 0.357% piracetam mixture samples by the PLS and NAS-CLS models, with respect to (top) the histograms, and (bottom) the predictive piracetam content at each pixel. The piracetam content was the mean value of triplicate measurements for each sample.

S6. MCR analysis

S6.1. Variance analysis of MCR factors

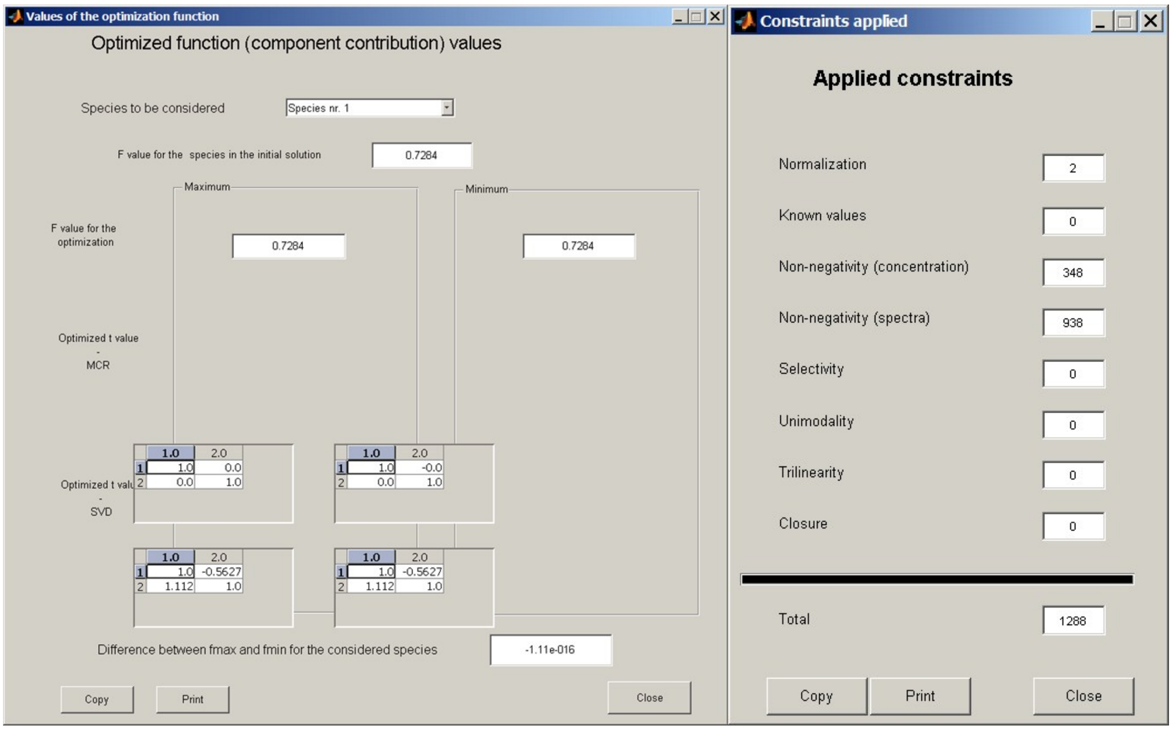
Table S-2: Spectral variance explained by MCR factors for the 10 spectrometer channel spectra using three specific ranges of 480−830, 1040−1510, and 1628−1740 cm^{-1}. Two MCR factors ($\mathbf{s}_{1mcr}$ and $\mathbf{s}_{2mcr}$) were obtained from the 0−100% piracetam content sample spectra for each channel. Both the cumulative variance and %LOF were also calculated.

Spectral data (X)	% Variance captured by $\mathbf{s}_{1mcr}$	% Variance captured by $\mathbf{s}_{2mcr}$	Cumulative % variance	%LOF
Channel 1	56.1113	43.7195	99.8307	0.1693
Channel 2	55.0360	44.7875	99.8235	0.1765
Channel 3	55.1602	44.6612	99.8214	0.1786
Channel 4	55.2812	44.5401	99.8213	0.1787
Channel 5	**55.3274**	**44.4913**	**99.8186**	**0.1814**
Channel 6	55.2106	44.5980	99.8087	0.1913
Channel 7	55.1366	44.6797	99.8163	0.1837
Channel 8	54.9694	44.8519	99.8212	0.1788
Channel 9	54.9576	44.8629	99.8205	0.1795
Channel 10	54.9752	44.8487	99.8238	0.1762

S6.2. MCR-BANDS optimization

Table S-3: Rotation ambiguity measured by MCR component relative contribution function maximum and minimum values, for the 10 spectrometer channel spectra using three specific ranges of 480−830, 1040−1510, and 1628−1740 cm^{-1}. $f_{1,2}^{max}$ and $f_{1,2}^{min}$ respectively signify two MCR components ($\mathbf{s}_{1mcr}$ and $\mathbf{s}_{2mcr}$) in terms of the maximum and minimum values in each channel case. Constrains of spectrum normalization and non-negativity of both spectra and concentrations were applied. The differences (either $f_1^{max} - f_1^{min}$ or $f_2^{max} - f_2^{min}$) between $f_{1,2}^{max}$ and $f_{1,2}^{min}$ were zero, indicating that the obtainment of MCR components was unique.

Spectral data (X)	**1st Component ($\mathbf{s}_{1mcr}$)**		**2nd Component ($\mathbf{s}_{2mcr}$)**	
	f_1^{max}	f_1^{min}	f_2^{max}	f_2^{min}
Channel 1	0.7334	0.7334	0.6474	0.6474
Channel 2	0.7263	0.7263	0.6552	0.6552
Channel 3	0.7271	0.7271	0.6543	0.6543
Channel 4	0.7280	0.7280	0.6534	0.6534
Channel 5	**0.7284**	**0.7284**	**0.6531**	**0.6531**
Channel 6	0.7276	0.7276	0.6539	0.6539
Channel 7	0.7270	0.7270	0.6544	0.6544
Channel 8	0.7257	0.7257	0.6555	0.6555
Channel 9	0.7257	0.7257	0.6557	0.6557
Channel 10	0.7258	0.7258	0.6555	0.6555



Values of the optimization function

Optimized function (component contribution) values

Species to be considered: Species nr. 1

F value for the species in the initial solution: 0.7284

Maximum

Minimum

F value for the optimization: 0.7284 0.7284

Optimized t value - MCR

Optimized t value - SVD

	1.0	2.0
1	1.0	0.0
2	0.0	1.0

	1.0	2.0
1	1.0	-0.0
2	0.0	1.0

	1.0	2.0
1	1.0	-0.5627
2	1.112	1.0

	1.0	2.0
1	1.0	-0.5627
2	1.112	1.0

Difference between fmax and fmin for the considered species: -1.11e-016

Constraints applied

Applied constraints

Normalization	2
Known values	0
Non-negativity (concentration)	348
Non-negativity (spectra)	938
Selectivity	0
Unimodality	0
Trilinearity	0
Closure	0
Total	1288

Figure S-3: Details about MCR-BANDS optimization results, using the spectra in three specific ranges of 480−830, 1040−1510, and 1628−1740 cm^{-1} obtained from the spectrometer channel 5 data.

S6.3. High-content quantification of piracetam

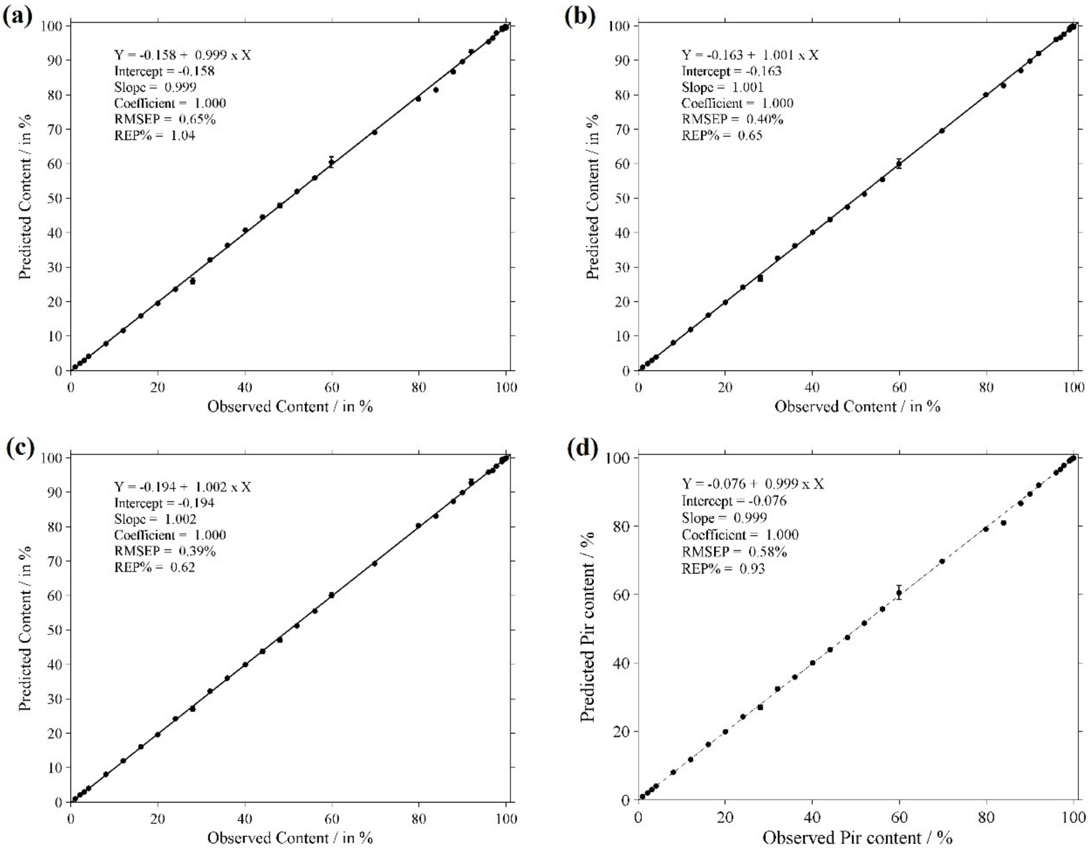


Figure S-4: High-content quantification of piracetam in mixture samples in the high concentration range of 1.0−100%, predicted by: (**a**) MCR, (**b**) NAS-CLS, (**c**) PCA-CLS, and (**d**) PLS models.

S7. PCA-CLS modelling

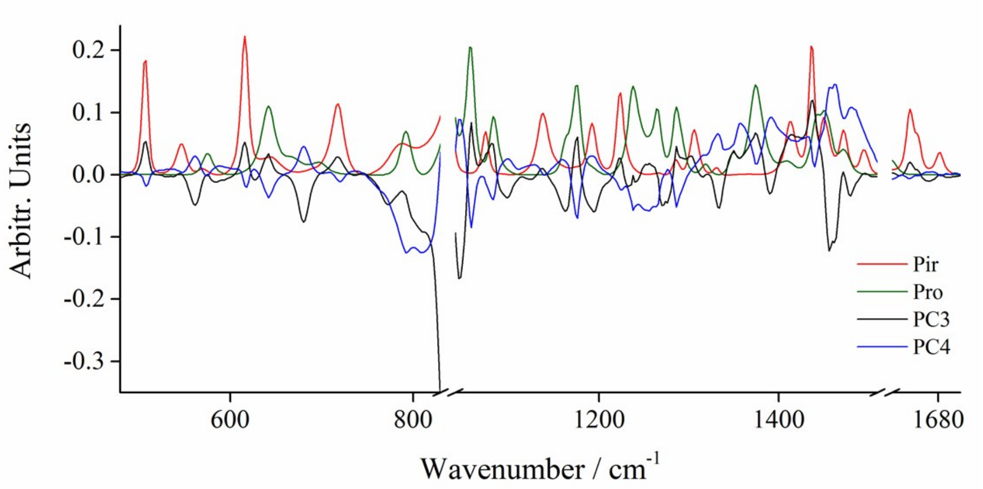


Figure S-5: PCA component loadings in the spectral ranges of 480−830, 1040−1510, and 1628−1740 cm^{-1}, compared with pure piracetam and proline spectra. They were obtained from *Model1* using spectrometer channel 5 spectra.

S8. Maps of 0.103% piracetam powder mixture

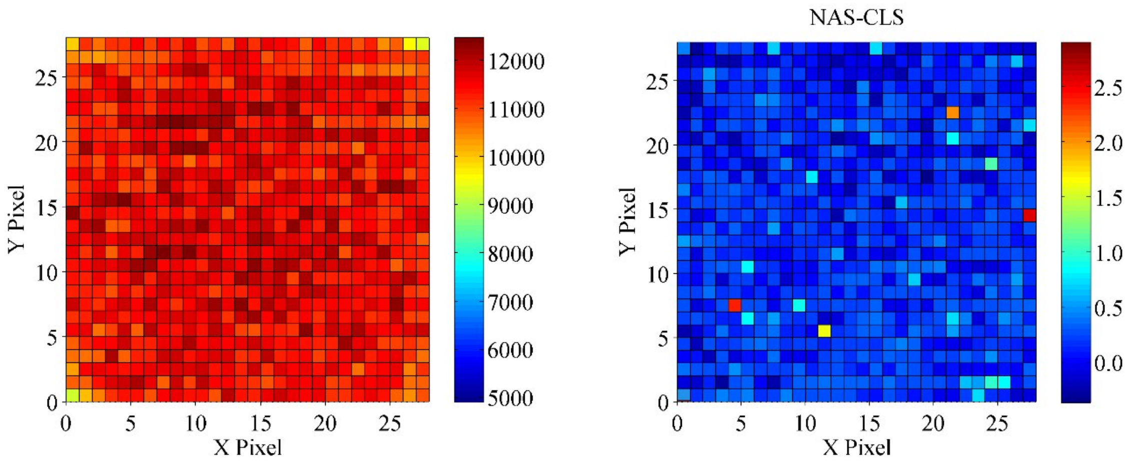


Figure S-6: Raman maps of the 0.103% piracetam powder mixture, showing the integrated intensity of 10 channel averaging spectra of (**a**) raw data, and (**b**) NAS-CLS prediction scores. Colour bars represent intensity and piracetam content in w/w%, respectively.

Table S-4: Similarity in terms of correlation coefficient between the maps represented by the integrated intensity (IntInt) of raw Raman spectra of the 0.103% piracetam powder mixture, and prediction scores obtained by PLS, NAS-CLS, MCR, PCA-CLS, and BR-PCHIP methods.

Similarity	IntInt	PLS	NAS-CLS	MCR	PCA-CLS	BR-PCHIP
IntInt	1	0.048	0.212	0.097	0.009	0.061
PLS	0.048	1	0.737	0.780	**0.912**	0.645
NAS-CLS	0.212	0.737	1	0.781	0.596	0.521
MCR	0.097	0.780	0.781	1	0.764	0.572
PCA-CLS	0.009	**0.912**	0.596	0.764	1	0.634
BR-PCHIP	0.061	0.645	0.521	0.572	0.634	1

S9. Selection of three specific spectral ranges

Three specific spectral ranges (480−830, 1040−1510, and 1628−1740 cm^{-1}) were selected for use in model development, based on their Raman scattering coefficients, the ratio of the spectrum overlap integral to the total spectral area, and the correlation coefficients between the spectra, as detailed below:

- Piracetam and proline had approximately equal Raman scattering coefficients, *i.e.*, the piracetam-to-proline ratio of their individual integrated spectra (PPSR) in the entire range of 200−1896 cm^{-1} was 100:94. This means that LOD is limited by such a ratio, with LOD decreasing as the relative scattering efficiency of the target analyte increases, compared to the matrix component. The ratio of the spectrum overlap integral to the total spectral area (SOTAR) of a constituent was 0.59 for piracetam and 0.63 for proline; in essence, both were close to 50%. The smaller this ratio (which can vary from 0 to 1), the easier it should be to quantify a low-content analyte in mixtures.

- In the 480−830 cm^{-1} spectral range: (1) the piracetam-to-proline ratio of the individual integrated spectra was 100:31; (2) the ratio of the spectrum overlap integral to the total spectral area was 0.21 for piracetam and 0.67 for proline; (3) the correlation coefficient (CC) between the two spectra was −0.024.
- In the 1040−1510 cm^{-1} range: (1) the piracetam-to-proline ratio of the individual integrated spectra was 100:104; (2) the ratio of the spectrum overlap integral to the total spectral area was 0.36 for piracetam and 0.35 for proline; (3) the correlation coefficient between the two spectra was −0.136.
- In the 1628−1740 cm^{-1} range: (1) the piracetam-to-proline ratio of the individual integrated spectra was 100:6; (2) the ratio of the spectrum overlap integral to the total spectral area was 0.05 for piracetam and 0.75 for proline; (3) the correlation coefficient between the two spectra was −0.084.

Table S-5: Comparison of the piracetam and proline Raman spectra in specific ranges.

Spectral ranges	PPSR	SOTAR		CC
		Pir	Pro	
200−1896 cm^{-1}	100:94	0.59	0.63	0.252
480−830 cm^{-1}	100:31	0.21	0.67	−0.024
1040−1510 cm^{-1}	100:104	0.36	0.35	−0.136
1628−1740 cm^{-1}	100:6	0.05	0.75	−0.084
200−480 cm^{-1}	100:89	0.62	0.69	0.085
830−1040 cm^{-1}	100:148	0.76	0.51	0.145
1510−1628 cm^{-1}	100:155	0.94	0.61	−0.230
1740−1896 cm^{-1}	100:117	0.99	0.85	0.964

Therefore, one could conclude that in these three ranges, the piracetam spectrum had high selectivity, and the selected Raman bands did not overlap with the proline spectrum too much, leading to more accurate models. In contrast, the piracetam and proline Raman spectra were strongly overlapped in the ranges of 200−480, 830−1040, 1510−1628, and 1740−1896 cm^{-1}, and so were not used.

S10. Rationale for piracetam/proline model system

Piracetam (API) and proline (excipient) were selected as a model system to develop a robust analytical methodology for several reasons:

(1) First, piracetam and proline have approximately equal Raman scattering coefficients. When we compared the integrated spectra (200~1896 cm^{-1} range) the *piracetam-to-proline-to-hydrated proline ratio* was 100:94:13. Obviously, the Limit of Detection (LOD) will be determined largely by this ratio, with LOD decreasing as the relative scattering efficiency of the target analyte increases compared to the matrix component.

(2) Second, the ratio of the spectrum overlap integral to the total spectral area of a constituent was 0.59 for piracetam and 0.63 for proline; in essence, both were close to 50%. The

smaller this ratio (which can vary from 0 to 1), the easier it should be to quantify a low-content analyte in mixtures. We selected this combination, as it was in the middle of the range and possibly more representative of the degree of spectral overlap encountered in real world applications.

(3) Third, to ensure facile HPLC validation of the Raman method, piracetam and proline could be easily separated and the low-content piracetam produced a quite strong peak facilitating accurate quantification by HPLC.

(4) Finally, the piracetam polymorph was stable while the proline matrix was sensitive to environmental factors, *e.g.*, water absorption leading to hydrate formation. This introduced another variable, which made the quantification of low-level analyte more complicated than a simple binary mixture model. Hydration is a common issue with solid-state matrix/formulation analysis, and this method needed to be able to identify samples that have been compromised.

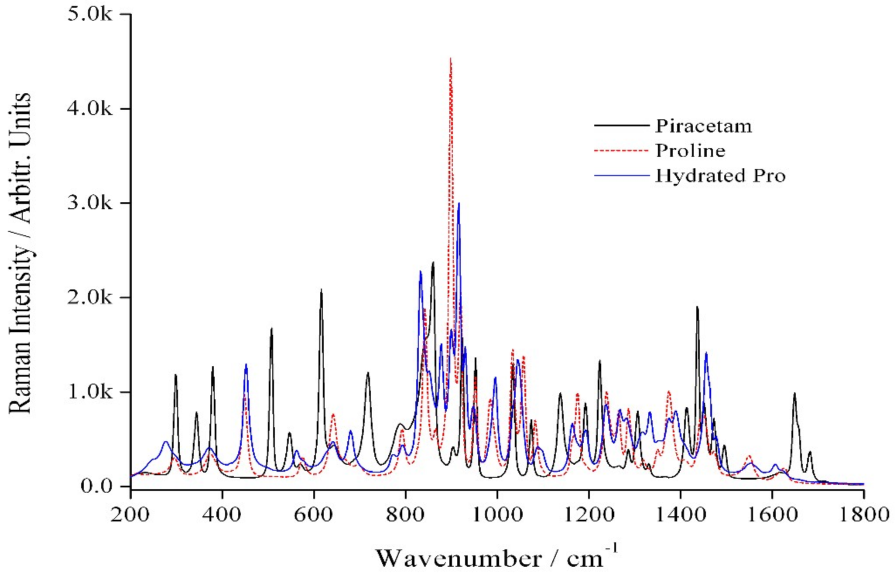


Figure S-7: Overlaid Raman spectra of piracetam, proline, and hydrated proline powders.

S11. Samples used for study

No. of Samples	Low content	High content
	20	30
Range	0~0.95%	1.0~100%
1	S0P000R1/2/3	S2P1R1/2/3
2	S1P005R1/2/3	S3P2R1/2/3
3	S2P010R1/2/3	S4P3R1/2/3
4	S3P015R1/2/3	S5P4R1/2/3
5	S4P020R1/2/3	S6P8R1/2/3
6	S5P025R1/2/3	S7P12R1/2/3
7	S6P030R1/2/3	S8P16R1/2/3
8	S7P035R1/2/3	S9P20R1/2/3
9	S8P040R1/2/3	S10P24R1/2/3
10	S9P045R1/2/3	S11P28R1/2/3
11	S10P050R1/2/3	S12P32R1/2/3
12	S11P055R1/2/3	S13P36R1/2/3
13	S12P060R1/2/3	S14P40R1/2/3
14	S13P065R1/2/3	S15P44R1/2/3
15	S14P070R1/2/3	S16P48R1/2/3
16	S15P075R1/2/3	S17P52R1/2/3
17	S16P080R1/2/3	S18P56R1/2/3
18	S17P085R1/2/3	S19P60R1/2/3
19	S18P090R1/2/3	S20P70R1/2/3
20	S19P095R1/2/3	S21P80R1/2/3
21		S211P84R1/2/3
22		S212P88R1/2/3
23		S22P90R1/2/3
24		S221P92R1/2/3
25		S222P96R1/2/3
26		S223P97R1/2/3
27		S224P98R1/2/3
28		S225P99R1/2/3
29		S2255P995R1/2/3
30		S23P100R1/2/3

Code: S____P_____R1/2/3

First digits after S were the sample number, and numbers after P were the piracetam concentration, 0 to 0.95% w/w for the low-content range, and 1.0 to 100% w/w for the high-content range. Number after R meant each sample was prepared in triplicate.

S12. Effect of varying sample number on model accuracy

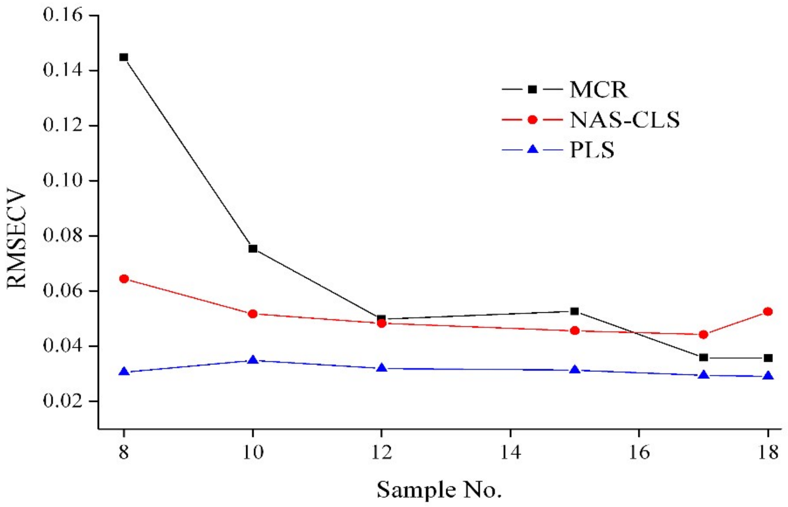


Figure S-8: Predictive errors showing the MCR, PLS and NAS-CLS model performance accuracy varied with the sample numbers. Leave-one-out (LOO) cross validation was used for calculating the model prediction errors. Data were from channel 5 spectra of the 0.05−1.0% piracetam content samples.

S13. MCR Factors.

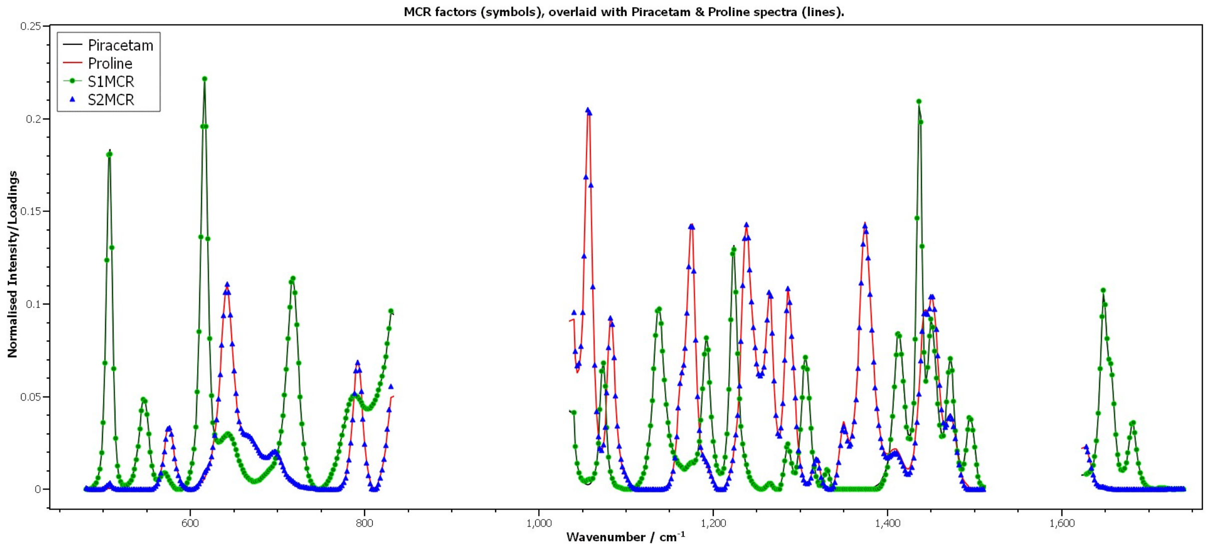


Figure S-9: Overlaid plot of the $\mathbf{S}_{1mcr}$ and $\mathbf{S}_{2mcr}$ factors from the MCR model with the Raman spectra of pure piracetam and proline showing the almost perfect agreement. [Channel 5 data only].

S14. References

[1] Z. Li, D.-J. Zhan, J.-J. Wang, J. Huang, Q.-S. Xu, Z.-M. Zhang, Y.-B. Zheng, Y.-Z. Liang, H. Wang, Morphological weighted penalized least squares for background correction, *Analyst*, 138 (2013) 4483-4492.

[2] B.Y. Li, A. Calvet, Y. Casamayou-Boucau, C. Morris, A.G. Ryder, Low-Content Quantification in Powders Using Raman Spectroscopy: A Facile Chemometric Approach to Sub 0.1% Limits of Detection, *Anal Chem*, 87 (2015) 3419-3428.

[3] B. Li, A. Calvet, Y. Casamayou-Boucau, A.G. Ryder, Kernel principal component analysis residual diagnosis (KPCARD): An automated method for cosmic ray artifact removal in Raman spectra, *Anal Chim Acta*, 913 (2016) 111-120.

[4] J. Engel, J. Gerretzen, E. Szymanska, J.J. Jansen, G. Downey, L. Blanchet, L.M.C. Buydens, Breaking with trends in pre-processing, *TrAC-Trend Anal Chem*, 50 (2013) 96-106.

[5] M. Vidal, J.M. Amigo, Pre-processing of hyperspectral images. Essential steps before image analysis, *Chemometr. Intell. Lab. Syst.*, 117 (2012) 138-148.

[6] T. Fearn, C. Riccioli, A. Garrido-Varo, J.E. Guerrero-Ginel, On the geometry of SNV and MSC, *Chemometr. Intell. Lab. Syst.*, 96 (2009) 22-26.

[7] M. Shamsipur, V. Zare-Shahabadi, B. Hemmateenejad, M. Akhond, Ant colony optimisation: a powerful tool for wavelength selection, J. Chemometr., 20 (2006) 146-157.

[8] F. Allegrini, A.C. Olivieri, A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy/partial least-squares analysis, *Anal Chim Acta*, 699 (2011) 18-25.

[9] L. Zhang, M.J. Henson, S.S. Sekulic, Multivariate data analysis for Raman imaging of a model pharmaceutical tablet, *Anal. Chim. Acta*, 545 (2005) 262-278.

[10] M. Fischer, C.D. Tran, Investigation of solid phase peptide synthesis by the near-infrared multispectral imaging technique: A detection method for combinatorial chemistry, *Anal Chem*, 71 (1999) 2255-2261.

[11] X.X. Han, Y. Xie, B. Zhao, Y. Ozaki, Highly Sensitive Protein Concentration Assay over a Wide Range via Surface-Enhanced Raman Scattering of Coomassie Brilliant Blue, *Anal Chem*, 82 (2010) 4325-4328.

[12] S. Sasic, D.A. Clark, J.C. Mitchell, M.J. Snowden, A comparison of Raman chemical images produced by univariate and multivariate data processing - a simulation with an example from pharmaceutical practice, *Analyst*, 129 (2004) 1001-1007.

[13] S. Pruess, Shape-preserving C-2 cubic spline interpolation, *Ima Journal of Numerical Analysis*, 13 (1993) 493-507.

[14] R.E. Carlson, F.N. Fritsch, An algorithm for monotone piecewise bicubic interpolation, *Siam Journal on Numerical Analysis*, 26 (1989) 230-238.

[15] C.M. David Kahaner, Stephen Nash *Numerical Methods and Software*, Prentice-Hall, Englewood Cliffs, New Jersey, 1989.

[16] A. de Juan, R. Tauler, R. Dyson, C. Marcolli, M. Rault, M. Maeder, Spectroscopic imaging and chemometrics: a powerful combination for global and local sample analysis, *TrAC-Trends Anal. Chem.*, 23 (2004) 70-79.

[17] K.C. Gordon, C.M. McGoverin, Raman mapping of pharmaceuticals, *International Journal of Pharmaceutics*, 417 (2011) 151-162.

[18] H. Shinzawa, K. Awa, W. Kanematsu, Y. Ozaki, Multivariate data analysis for Raman spectroscopic imaging, *J. Raman Spectrosc.*, 40 (2009) 1720-1725.

[19] P.G. Hans Grahn, *Techniques and Applications of Hyperspectral Image Analysis*, John Wiley & Sons Ltd., Chichester, 2007.

[20] B. Vajna, I. Farkas, A. Szabo, Z. Zsigmond, G. Marosi, Raman microscopic evaluation of technology dependent structural differences in tablets containing imipramine model drug, *Journal of Pharmaceutical and Biomedical Analysis*, 51 (2010) 30-38.

[21] J.M. Amigo, C. Ravn, Direct quantification and distribution assessment of major and minor components in pharmaceutical tablets by NIR-chemical imaging, *European Journal of Pharmaceutical Sciences*, 37 (2009) 76-82.

[22] C. Gendrin, Y. Roggo, C. Collet, Pharmaceutical applications of vibrational chemical imaging and chemometrics: A review, *J. Pharm. Biomed. Anal.*, 48 (2008) 533-553.

[23] W.F. de Carvalho Rocha, G.P. Sabin, P.H. Marco, R.J. Poppi, Quantitative analysis of piroxicam polymorphs pharmaceutical mixtures by hyperspectral imaging and chemometrics, *Chemometr. Intell. Lab. Syst.*, 106 (2011) 198-204.

[24] T.T. Lied, P. Geladi, K.H. Esbensen, Multivariate image regression (MIR): implementation of image PLSR-first forays, *J. Chemometr.*, 14 (2000) 585-598.

[25] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, *Chemometr. Intell. Lab. Syst.*, 58 (2001) 109.

[26] T.N. H. Martens, *Multivariate Calibration*, Wiley, New York, 1989.

[27] M.J. Pelletier, Quantitative analysis using Raman spectrometry, *Appl Spectrosc*, 57 (2003) 20A-42A.

[28] A. Lorber, K. Faber, B.R. Kowalski, Net Analyte Signal Calculation in Multivariate Calibration, *Anal. Chem.*, 69 (1997) 1620.

[29] J.H.K. J. Palmer, Net analyte signal (NAS) for selection of multivariate calibration models and development of NAS sample-wise target calibration model attributes, in: 40 Years of Chemometrics – From Bruce Kowalski to the Future, Chapter 9, 2015.

[30] I.T. Jolliffe, *Principal Component Analysis*, 2nd ed., Springer, New York, 2002.

[31] E.R. Malinowski, *Factor Analysis in Chemistry*, Wiley, New York, 2002.

[32] J.H. Jiang, Y.Z. Liang, Y. Ozaki, Principles and methodologies in self-modeling curve resolution, *Chemometr. Intell. Lab. Syst.*, 71 (2004) 1-12.

[33] A. de Juan, R. Tauler, Multivariate curve resolution (MCR) from 2000: Progress in concepts and applications, *Crit. Rev. Anal. Chem.*, 36 (2006) 163-176.

[34] E. Spjotvoll, H. Martens, R. Volden, Restricted least-squares estimation of the spectra and concentration of 2 unknown constituents available in mixtures, *Technometrics*, 24 (1982) 173-180.

[35] 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.

[36] R. Tauler, Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution, *J. Chemometr.*, 15 (2001) 627-646.

[37] B. Li, A. Calvet, Y. Casamayou-Boucau, C. Morris, A.G. Ryder, Low-content quantification in powders using Raman spectroscopy: a facile chemometric approach to sub 0.1% limits of detection, *Anal Chem*, 87 (2015) 3419-3428.

[38] S. Wold, H. Antti, F. Lindgren, J. Ohman, Orthogonal signal correction of near-infrared spectra, *Chemometr. Intell. Lab. Syst.*, 44 (1998) 175-185.

[39] H.A. Martens, P. Dardenne, Validation and verification of regression in small data sets, *Chemometr. Intell. Lab. Syst.*, 44 (1998) 99-121.