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 .......................................................................................................... 1
S2. Description of chemometric methods ............................................................................... 2
S3. Raman mapping analysis scheme ..................................................................................... 6
S4. PLS quantification Model2 with non-OSC treated data ................................................... 6
S5. Piracetam prediction by the NAS-CLS method ................................................................. 8
S6. MCR analysis .................................................................................................................... 9
S7. PCA-CLS modelling .......................................................................................................... 11
S8. Maps of 0.103% piracetam powder mixture ................................................................... 12
S9. Selection of three specific spectral ranges ...................................................................... 12
S10. Rationale for piracetam/proline model system ............................................................... 13
S11. Samples used for study .................................................................................................. 15
S12. Effect of varying sample number on model accuracy .................................................... 16
S13. MCR Factors .................................................................................................................. 16
S14. References ..................................................................................................................... 17

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., $D_{X \times Y \times \lambda}$), 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 $t$), and lowercase italicized letters for scalars (as $f$). 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., $\|s\|_F$).

Each Raman mapping measurement generated a three-way array ($D_{X \times Y \times \lambda}$), where $X$ and $Y$ respectively denote the spatial co-ordinates of each pixel and $\lambda$ refers to the spectral pattern along the wavenumber axis for the $xy$th pixel [9]. These three-way datasets ($D_{X \times Y \times \lambda}$) were then unfolded into two-way matrices, $D_{XY \times \lambda}$, 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:

$$r_{pp} = \frac{I_{1652}}{I_{448}}$$

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 ($y_i$) in $n$ sample spectra, there is a unique shape-preserving piecewise cubic Hermite interpolating polynomial (PCHIP) [13-15]:

$$P(r_{pp,l}) = y_i, \quad l = 1, ..., n$$

This polynomial was able to satisfy interpolation conditions on the $l$th interval ($r_{pp,l+1} - r_{pp,l}$), hence either exactly producing or approximating the given data ($r_{pp,l}$, $y_i$). 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:

\[ y = Db_{PLS} + e \]  

(3)

through decomposing:

\[ D = TP^T + E \]  

(4)

\[ y = UQ^T + e^* \]  

(5)

where \( b_{PLS} \) is called the inner-relationship coefficients and can be specified by identifying \( r \) underlying factors (or latent variables, LVs) that explicitly maximize covariance (say \( D^Ty \)) between \( D \) and \( y \). Matrix \( T \) holds the factors (\( t_k, k=1,...,r \)), and \( U \) corresponds to scores (\( u_k, k=1,...,r \)) for \( y \), while \( P \) and \( Q \) individually have their loadings (\( p_k \) and \( q_k \)). \( e, E, \) and \( e^* \) 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 (\( t_k \) and \( u_k \)) and inner weight (\( w_k \)), (2) estimation of outer weights (\( q_k \)) and loadings (\( p_k \)), and path coefficients, and (3) obtainment of regression coefficients \( b_{PLS} \):

\[ b_{PLS} = W(P^TW)^{-1}Q^T \]  

(6)

where \( W \) stands for the weights (\( w_k, k=1,...,r \)).

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:

\[ d = cs^T + e \]  

(7)
where \( \mathbf{d} \) is one measured spectrum of the data matrix \( \mathbf{D} \), \( \mathbf{S} \) 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 (\( \lambda \)) by \( r \) components. \( \mathbf{c} \) stands for the concentration weights of components, representing the degree to which each component contributes to the overall measurement. Given the measurement \( \mathbf{d} \) and reference spectra \( \mathbf{S} \), the concentration \( \mathbf{c} \) can be estimated by the least square approximation,

\[
\hat{\mathbf{c}} = \mathbf{d} (\mathbf{S}^T \mathbf{S})^{-1}
\]

(8)

The critical is that all the spectrally-active components (\( \mathbf{S} \)) must be independent and available so that \( \mathbf{c} \) can be accurately estimated. However, in most practical cases where \( \mathbf{S} \) 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 \( \mathbf{d} \) 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 (\( \mathbf{d}_{\neg k} \)), which not only includes interference but also residual not explained by the model:

\[
\mathbf{d} = \mathbf{d}_k + \mathbf{d}_{\neg k}
\]

(9)

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

\[
\mathbf{D} = \mathbf{D}_k + \mathbf{D}_{\neg k} = \mathbf{c}_k \mathbf{s}_k^T + \mathbf{D}_{\neg k}
\]

(10)

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

If a NAS filtering \( \mathbf{F}_{\text{NAS}} \) can be defined, then \( \mathbf{D}_{\neg k} \) is able to be removed from \( \mathbf{D} \):

\[
\mathbf{F}_{\text{NAS}} = \mathbf{I} - \mathbf{D}_{\neg k}^T \mathbf{D}_{\neg k}
\]

(11)

\[
\mathbf{D} \mathbf{F}_{\text{NAS}} = (\mathbf{c}_k \mathbf{s}_k^T + \mathbf{D}_{\neg k}) \mathbf{F}_{\text{NAS}} = \mathbf{c}_k \mathbf{s}_k^T \mathbf{F}_{\text{NAS}} + \mathbf{E}
\]

(12)

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

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

(13)

Thus, the filter \( \mathbf{F}_{\text{NAS}} \) 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 \( \mathbf{D} \) can be decomposed to yield \( r \) orthogonal factors or principal components (PCs) in \( \mathbf{P} \). Then, with these \( r \) PCs, the reference spectra \( \mathbf{S} \) of the pure chemical components in Equation (8) were substituted to estimate the concentration \( \mathbf{C} \):

\[
\hat{\mathbf{C}} = \mathbf{D} \mathbf{S}(\mathbf{S}^T \mathbf{S})^{-1} = \mathbf{D} \mathbf{P}^T \mathbf{P}^{-1}
\]

(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 $\hat{C}$ 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 $\hat{C}$ 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 ($C$):

$$D = CS^T + E = \sum_{k=1}^{r} c_k s_k^T + E \quad (15)$$

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:

$$\% \text{LOF} = 100 \times \frac{\sum_{xy\lambda} e_{xy\lambda}^2}{\sum_{xy\lambda} d_{xy\lambda}^2} \quad (16)$$

where $d_{xy\lambda}$ and $e_{xy\lambda}$ respectively stand for the $xy\lambda$th element in the experimental spectra $D_{XY\lambda}$ and the residual associated with the reproduction of $d_{xy\lambda}$ by the MCR model. The explained variance can be indicated as $100-\%\text{LOF}$ 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:

$$f_k = \frac{\|c_k s_k^T\|}{\|CS^T\|} \quad (17)$$

where $f_k$ is a scalar value that defines the relative contribution of the $k$th particular component ($c_k s_k^T$) to the entire signal ($CS^T$) considering of the $r$ components in the measured mixture ($k = 1, ..., r$), 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 \( c_k \) and \( s_k \) 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 \( f_k^{\text{max}} \) and \( f_k^{\text{min}} \) respectively. Then, the extent of rotational ambiguity can be evaluated by the difference between \( f_k^{\text{max}} \) and \( f_k^{\text{min}} \) 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 \( \mathbf{S} \) is reliable. Therefore, the concentration profiles \( \mathbf{C} \) 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 \( \mathbf{C} \) 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.](image)

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.

<table>
<thead>
<tr>
<th>Piracetam in 0–2.5%</th>
<th>PLS Model2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSEC</td>
<td>RMSECV</td>
<td></td>
</tr>
<tr>
<td>Channel 1</td>
<td>0.063670</td>
<td>0.079980</td>
<td></td>
</tr>
<tr>
<td>Channel 2</td>
<td>0.058344</td>
<td>0.074055</td>
<td></td>
</tr>
<tr>
<td>Channel 3</td>
<td>0.058794</td>
<td>0.071869</td>
<td></td>
</tr>
<tr>
<td>Channel 4</td>
<td>0.052208</td>
<td>0.066314</td>
<td></td>
</tr>
<tr>
<td>Channel 5</td>
<td>0.063635</td>
<td>0.078263</td>
<td></td>
</tr>
</tbody>
</table>
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 (\(s_{1\text{incr}}\) and \(s_{2\text{incr}}\)) were obtained from the 0–100% piracetam content sample spectra for each channel. Both the cumulative variance and %LOF were also calculated.

<table>
<thead>
<tr>
<th>Spectral data (X)</th>
<th>% Variance captured by (s_{1\text{incr}})</th>
<th>% Variance captured by (s_{2\text{incr}})</th>
<th>Cumulative % variance</th>
<th>%LOF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel 1</td>
<td>56.1113</td>
<td>43.7195</td>
<td>99.8307</td>
<td>0.1693</td>
</tr>
<tr>
<td>Channel 2</td>
<td>55.0360</td>
<td>44.7875</td>
<td>99.8235</td>
<td>0.1765</td>
</tr>
<tr>
<td>Channel 3</td>
<td>55.1602</td>
<td>44.6612</td>
<td>99.8214</td>
<td>0.1786</td>
</tr>
<tr>
<td>Channel 4</td>
<td>55.2812</td>
<td>44.5401</td>
<td>99.8213</td>
<td>0.1787</td>
</tr>
<tr>
<td><strong>Channel 5</strong></td>
<td><strong>55.3274</strong></td>
<td><strong>44.4913</strong></td>
<td><strong>99.8186</strong></td>
<td><strong>0.1814</strong></td>
</tr>
<tr>
<td>Channel 6</td>
<td>55.2106</td>
<td>44.5980</td>
<td>99.8087</td>
<td>0.1913</td>
</tr>
<tr>
<td>Channel 7</td>
<td>55.1366</td>
<td>44.6797</td>
<td>99.8163</td>
<td>0.1837</td>
</tr>
<tr>
<td>Channel 8</td>
<td>54.9694</td>
<td>44.8519</td>
<td>99.8212</td>
<td>0.1788</td>
</tr>
<tr>
<td>Channel 9</td>
<td>54.9576</td>
<td>44.8629</td>
<td>99.8205</td>
<td>0.1795</td>
</tr>
<tr>
<td>Channel 10</td>
<td>54.9752</td>
<td>44.8487</td>
<td>99.8238</td>
<td>0.1762</td>
</tr>
</tbody>
</table>

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\(^{-1}\). \(f_{1,2}^{\text{max}}\) and \(f_{1,2}^{\text{min}}\) respectively signify two MCR components (\(s_{1\text{incr}}\) and \(s_{2\text{incr}}\)) 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 $f^\text{max}_{1,2} - f^\text{min}_{1,2}$ or $f^\text{max}_{2,1} - f^\text{min}_{2,1}$) between $f^\text{max}_{1,2}$ and $f^\text{min}_{1,2}$ were zero, indicating that the obtainment of MCR components was unique.

<table>
<thead>
<tr>
<th>Spectral data (X)</th>
<th>1st Component ($s_{1\text{mcr}}$)</th>
<th>2nd Component ($s_{2\text{mcr}}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel 1</td>
<td>0.7334 0.7334</td>
<td>0.6474 0.6474</td>
</tr>
<tr>
<td>Channel 2</td>
<td>0.7263 0.7263</td>
<td>0.6552 0.6552</td>
</tr>
<tr>
<td>Channel 3</td>
<td>0.7271 0.7271</td>
<td>0.6543 0.6543</td>
</tr>
<tr>
<td>Channel 4</td>
<td>0.7280 0.7280</td>
<td>0.6534 0.6534</td>
</tr>
<tr>
<td>Channel 5</td>
<td>0.7284 0.7284</td>
<td>0.6531 0.6531</td>
</tr>
<tr>
<td>Channel 6</td>
<td>0.7276 0.7276</td>
<td>0.6539 0.6539</td>
</tr>
<tr>
<td>Channel 7</td>
<td>0.7270 0.7270</td>
<td>0.6544 0.6544</td>
</tr>
<tr>
<td>Channel 8</td>
<td>0.7257 0.7257</td>
<td>0.6555 0.6555</td>
</tr>
<tr>
<td>Channel 9</td>
<td>0.7257 0.7257</td>
<td>0.6557 0.6557</td>
</tr>
<tr>
<td>Channel 10</td>
<td>0.7258 0.7258</td>
<td>0.6555 0.6555</td>
</tr>
</tbody>
</table>

**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

![Raman maps of the 0.103% piracetam powder mixture](image)

**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.

<table>
<thead>
<tr>
<th>Similarity</th>
<th>IntInt</th>
<th>PLS</th>
<th>NAS-CLS</th>
<th>MCR</th>
<th>PCA-CLS</th>
<th>BR-PCHIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>IntInt</td>
<td>1</td>
<td>0.048</td>
<td>0.212</td>
<td>0.097</td>
<td>0.009</td>
<td>0.061</td>
</tr>
<tr>
<td>PLS</td>
<td>0.048</td>
<td>1</td>
<td>0.737</td>
<td>0.780</td>
<td><strong>0.912</strong></td>
<td>0.645</td>
</tr>
<tr>
<td>NAS-CLS</td>
<td>0.212</td>
<td>0.737</td>
<td>1</td>
<td>0.781</td>
<td>0.596</td>
<td>0.521</td>
</tr>
<tr>
<td>MCR</td>
<td>0.097</td>
<td>0.780</td>
<td>0.781</td>
<td>1</td>
<td>0.764</td>
<td>0.572</td>
</tr>
<tr>
<td>PCA-CLS</td>
<td>0.009</td>
<td><strong>0.912</strong></td>
<td>0.596</td>
<td>0.764</td>
<td>1</td>
<td>0.634</td>
</tr>
<tr>
<td>BR-PCHIP</td>
<td>0.061</td>
<td>0.645</td>
<td>0.521</td>
<td>0.572</td>
<td>0.634</td>
<td>1</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Spectral ranges</th>
<th>PPSR</th>
<th>SOTAR</th>
<th>CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>200–1896 cm(^{-1})</td>
<td>100:94</td>
<td>0.59</td>
<td>0.63</td>
</tr>
<tr>
<td>480–830 cm(^{-1})</td>
<td>100:31</td>
<td>0.21</td>
<td>0.67</td>
</tr>
<tr>
<td>1040–1510 cm(^{-1})</td>
<td>100:104</td>
<td>0.36</td>
<td>0.35</td>
</tr>
<tr>
<td>1628–1740 cm(^{-1})</td>
<td>100:6</td>
<td>0.05</td>
<td>0.75</td>
</tr>
<tr>
<td>200–480 cm(^{-1})</td>
<td>100:89</td>
<td>0.62</td>
<td>0.69</td>
</tr>
<tr>
<td>830–1040 cm(^{-1})</td>
<td>100:148</td>
<td>0.76</td>
<td>0.51</td>
</tr>
<tr>
<td>1510–1628 cm(^{-1})</td>
<td>100:155</td>
<td>0.94</td>
<td>0.61</td>
</tr>
<tr>
<td>1740–1896 cm(^{-1})</td>
<td>100:117</td>
<td>0.99</td>
<td>0.85</td>
</tr>
</tbody>
</table>

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](image)

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

S11. Samples used for study

<table>
<thead>
<tr>
<th>No. of Samples</th>
<th>Low content</th>
<th>High content</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>20</td>
<td>30</td>
</tr>
<tr>
<td>Range</td>
<td>0–0.95%</td>
<td>1.0–100%</td>
</tr>
<tr>
<td>1</td>
<td>S0P000R1/2/3</td>
<td>S2P1R1/2/3</td>
</tr>
<tr>
<td>2</td>
<td>S1P005R1/2/3</td>
<td>S3P2R1/2/3</td>
</tr>
<tr>
<td>3</td>
<td>S2P010R1/2/3</td>
<td>S4P3R1/2/3</td>
</tr>
<tr>
<td>4</td>
<td>S3P015R1/2/3</td>
<td>S5P4R1/2/3</td>
</tr>
<tr>
<td>5</td>
<td>S4P020R1/2/3</td>
<td>S6P8R1/2/3</td>
</tr>
<tr>
<td>6</td>
<td>S5P025R1/2/3</td>
<td>S7P12R1/2/3</td>
</tr>
<tr>
<td>7</td>
<td>S6P030R1/2/3</td>
<td>S8P16R1/2/3</td>
</tr>
<tr>
<td>8</td>
<td>S7P035R1/2/3</td>
<td>S9P20R1/2/3</td>
</tr>
<tr>
<td>9</td>
<td>S8P040R1/2/3</td>
<td>S10P24R1/2/3</td>
</tr>
<tr>
<td>10</td>
<td>S9P045R1/2/3</td>
<td>S11P28R1/2/3</td>
</tr>
<tr>
<td>11</td>
<td>S10P050R1/2/3</td>
<td>S12P32R1/2/3</td>
</tr>
<tr>
<td>12</td>
<td>S11P055R1/2/3</td>
<td>S13P36R1/2/3</td>
</tr>
<tr>
<td>13</td>
<td>S12P060R1/2/3</td>
<td>S14P40R1/2/3</td>
</tr>
<tr>
<td>14</td>
<td>S13P065R1/2/3</td>
<td>S15P44R1/2/3</td>
</tr>
<tr>
<td>15</td>
<td>S14P070R1/2/3</td>
<td>S16P48R1/2/3</td>
</tr>
<tr>
<td>16</td>
<td>S15P075R1/2/3</td>
<td>S17P52R1/2/3</td>
</tr>
<tr>
<td>17</td>
<td>S16P080R1/2/3</td>
<td>S18P56R1/2/3</td>
</tr>
<tr>
<td>18</td>
<td>S17P085R1/2/3</td>
<td>S19P60R1/2/3</td>
</tr>
<tr>
<td>19</td>
<td>S18P090R1/2/3</td>
<td>S20P70R1/2/3</td>
</tr>
<tr>
<td>20</td>
<td>S19P095R1/2/3</td>
<td>S21P80R1/2/3</td>
</tr>
<tr>
<td>21</td>
<td>S21P84R1/2/3</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>S212P88R1/2/3</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>S22P90R1/2/3</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>S221P92R1/2/3</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>S222P96R1/2/3</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>S223P97R1/2/3</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>S224P98R1/2/3</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>S225P99R1/2/3</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>S225P995R1/2/3</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>S23P100R1/2/3</td>
<td></td>
</tr>
</tbody>
</table>

**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](image)

**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](image)

**Figure S-9:** Overlaid plot of the $S_{1mcr}$ and $S_{2mcr}$ 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


Electronic Supplementary Information

Electronic Supplementary Information