



## Rapid characterization and quality control of complex cell culture media solutions using Raman spectroscopy and chemometrics

Title	Rapid characterization and quality control of complex cell culture media solutions using Raman spectroscopy and chemometrics
Author(s)	Li, Boyan; Ryan, Paul W.; Ray, Bryan H.; Sirimuthu, Narayana M. S.; Ryder, Alan G.
Publication Date	2010
Repository DOI	<a href="https://doi.org/10.1002/bit.22813">DOI 10.1002/bit.22813</a>

## **RAPID CHARACTERISATION AND QUALITY CONTROL OF COMPLEX CELL CULTURE MEDIA SOLUTIONS USING RAMAN SPECTROSCOPY AND CHEMOMETRICS.**

Boyan Li,<sup>1,2</sup> Paul W. Ryan,<sup>1,2</sup> Bryan H. Ray,<sup>1,2</sup> Kirk J. Leister,<sup>3</sup> Narayana M. S. Sirimuthu,<sup>1,2</sup> and Alan G. Ryder.<sup>1,2\*</sup>

<sup>1</sup> Nanoscale Biophotonics Laboratory, School of Chemistry, National University of Ireland, Galway, Galway, Ireland.

<sup>2</sup> Centre for Bioanalytical Sciences, School of Chemistry, National University of Ireland, Galway, Galway, Ireland.

<sup>3</sup> Analytical Biochemistry Department, Bristol-Myers Squibb Company, East Syracuse, NY, USA.

\* To whom all correspondence should be addressed.

**Tel:** +353 91 49 2943 **Fax:** +353 91 49 4596 **Email:** alan.ryder@nuigalway.ie (A.G.R)

**Running Title:** Raman spectroscopic analysis of cell culture media.

### **Abstract:**

The use of Raman spectroscopy coupled with chemometrics for the rapid identification, characterisation, and quality assessment of complex cell culture media components used for industrial mammalian cell culture was investigated. Raman spectroscopy offers significant advantages for the analysis of complex, aqueous based materials used in biotechnology because there is no need for sample preparation and water is a weak Raman scatterer. We demonstrate the efficacy of the method for the routine analysis of dilute aqueous solution of five different chemically defined, commercial media components used in a Chinese Hamster Ovary (CHO) cell manufacturing process for recombinant proteins.

The chemometric processing of the Raman spectral data is the key factor in developing robust methods. Here we discuss the optimum methods for eliminating baseline drift, background fluctuations and other instrumentation artefacts to generate reproducible spectral data. Principal component analysis (PCA) and soft independent modelling of class analogy (SIMCA) were then employed in the development of a robust routine for both identification and quality evaluation of the five different media components. These methods have the potential to be extremely useful in an industrial context for “in house” sample handling, tracking and quality control.

**Keywords:** Raman spectroscopy, chemometrics, cell culture, chemically defined media, eRDF, PCA, SIMCA.

Final submitted version, not proof corrected.

## 1. Introduction

Due to the growth in biopharmaceutical production worldwide, the development of rapid spectroscopic analysis methods is of great immediate interest (Becker et al. 2007; Rhiel et al. 2002; Roychoudhury et al. 2006; Workman et al. 2007). The need for rapid, reliable, robust, and non-destructive analytical methods is of paramount importance to ensure efficient and reliable process control, which in turn results in consistent product quality and significant cost savings. In many cases, the raw materials, and culture media (basal and feed) used in cell culture, are complex mixtures and in many cases it is impossible to identify and quantify every individual component (Davis 2002; Newman and Cragg 2007; Pollard 1997). These aqueous media must provide all of the cell nutrients required for growth as well as providing an energy source, while maintaining pH, and osmolarity. There are two general types of cell culture media used in Fed-Batch culture: basal nutrient media and feed media. Both are complex mixtures of inorganic salts, sugars, amino acids, vitamins, organic acids, buffers, and other materials, the exact compositions vary significantly according to cell line and manufacturer (Davis 2002). Combinations are also found, for example RDF medium is a 1:1 mixture of RPMI 1640 and DMEM/F12 formulations (Murakami et al. 1984). Enriched basal RDF (eRDF) has increased levels (~two-fold) of amino acids and glucose compared to RDF. In mammalian cell culture, eRDF is more widely used to optimize cell culture growth, achieve rapid cell proliferation, and thereby increase product yield and is often supplemented with animal serum (Jayme et al. 1997). Chinese hamster ovary (CHO) cells are one of the most commonly used cells used for the production of recombinant proteins in suspension culture (Schroder et al. 2004). For the production of human use products, protein-free, chemically defined (CD) media are desirable.

Comprehensive analysis of these complex materials is time consuming, expensive, and impractical for screening of all incoming containers and/or lots of media. Thus for rapid and inexpensive identification and quality assessment of these materials there is a need to develop holistic analytical methods based on spectroscopic approaches. We have already demonstrated that fluorescence spectroscopy can be used to rapidly assess the quality and performance of aqueous feed media (Ryan et al. 2010), however fluorescence spectra do not yield much detail about specific compositional changes. For vibrational spectroscopy based analysis of complex, biologically relevant materials, near-infrared (NIR) and mid-infrared (MIR), (Hashimoto et al. 2005; Liu and Arnold 2009; Rhiel et al. 2002; Riley et al. 1997; Roychoudhury et al. 2006) have been used. However, water signals are very strong in NIR and MIR, so that for dilute aqueous media solutions, much of the analyte spectral detail is masked. This therefore represents a significant problem in biotechnology. Raman spectroscopy has a number of significant advantages because water has a very weak Raman signal and sample preparation is usually not required, permitting *in-situ* analysis (Clark et al. 1995; Tsuchihashi et al. 1997). Raman spectroscopy is generally implemented using visible or NIR excitation sources which allows for the use of fibre optic probes and compact spectrometers (Marteau et al. 1996; Utzinger and Richards-Kortum 2003). Raman analysis of complex materials is well established with an extensive literature on samples varying from proteins, to single cells, and human tissue (Edwards et al. 1997; Ellis and Goodacre 2006; Frank et al. 1994; Lawson et al. 1997; McCreery et al. 1995; Utzinger and Richards-Kortum

2003). The use of Raman spectroscopy for in-reactor bioprocess monitoring has also been demonstrated (Cannizzaro et al. 2003; Lee et al. 2004; McGovern et al. 2002) for the analysis of either product formation or usage of starting materials (Cannizzaro et al. 2003; Carey 1998). One of the key technologies driving the adoption of Raman based methods has been the use of chemometrics to extract useful quantitative and qualitative information from spectra (Vickers TJ 1991). Chemometrics can be defined as the sciences of relating measurements made on a chemical system or process to the state of the system via application of mathematical, statistical, computational, and informatics methods (Massart et al. 1997; Vandeginste et al. 1998). Chemometric methods can correlate observed spectral changes with properties such as concentration and have permitted the identification and quantification of various complex materials by Raman spectroscopy (Archibald et al. 1998; Cooper et al. 1997; Everall et al. 1994; Shimoyama et al. 1997; Sohn et al. 2005; Williams and Everall 1995). The potential of qualitative and quantitative chemometric based analytical methods is well documented (Lavine and Workman 2006; Lavine and Workman 2008; Leger and Ryder 2006; Scarff et al. 2006; Shinzawa et al. 2006). The most widely used technique is principal component analysis (PCA) (Geladi 2003; Jackson 1980; Jackson 1981a; Jackson 1981b). PCA captures the maximum variance in the data with the first and each subsequent principal component and these components are linear combinations of the original variables. Careful examination of the components (e.g. the loadings and scores) can lead to a better understanding of the different sources of variation in the data. In an analytical methodology, PCA is first used in the analysis of the spectral (or other) data to provide information on variance and measurement integrity prior to the implementation of classification, clustering or other multivariate techniques. While PCA is useful for identifying differences between samples and pinpointing the causes of variance, it is not good at classification tasks since the PCA model only attempts to describe the overall variation in the data. Soft independent modelling of class analogy (SIMCA) (Wold and Sjostrom 1977) combats this by using PCA in a way that utilizes sample information to help differentiate between sample classes. A SIMCA model consists of a collection of PCA models, one for each class in the data set where each class can have different numbers of principal components.

Since cell culture media are complex aqueous mixtures, understanding the spectra and extracting compositional and quality information requires the use of chemometric techniques. The aim of this work was to demonstrate the efficacy of coupling chemometrics and rapid Raman spectroscopy based methods to both identify and monitor the quality of aqueous solutions of chemically defined cell culture media components used in the formulation of cell culture media used in a current industrial process. This large scale production process uses final stage bioreactors in excess of 10,000 litre volumes and media usage is on the order of 10 kL plus per lot. The media used in this process were prepared using a variety of starting materials which are all complex mixtures. PCA and SIMCA are employed for both determination of sample quality and discrimination of five different aqueous chemically defined media components. The "quality control" discussed in this study means the identity test of complex media components in a quantitative manner using Raman characterisation and chemometric evaluation. That is, the proposed Raman-chemometric method can not only identify samples by the use of SIMCA calibration models derived from a set of defined

normal samples, but also quantitatively assess sample quality using the Hotelling's  $T^2$  and Q residual analysis with a 95% confidence level. Using a combination of the  $T^2$  and Q analysis and spectral/variance analysis, problematic samples/batches either can be rejected outright or selected for more comprehensive testing, thus providing an inexpensive first stage in monitoring and quality control of cell culture media. The generic methods developed in this work do not rely on any predetermined compositional or property information, thus allowing their application to different production systems and a much wider range of complex mixtures.

## 2. Materials and Methods:

### 2.1 Materials

A range of chemically defined (CD), cell culture media from a current industrial biopharmaceutical process (a CHO process making a recombinant glycoprotein product) were supplied by Bristol-Myers Squibb (Table 1). These CD media were supplied as liquids and were stored in a fridge at  $\sim 4$  °C prior to measurement. eRDF (a chemically defined media supplement) was in solid form and a solution was made up under aseptic conditions by dissolving 16.8 g/L in Millipore water (18 M $\Omega$  resistance), and were stored at  $-70$  °C. Dissolved solids (in most cases) were estimated to be in the 0.5~2% range. The exact compositional information of these materials is proprietary to Bristol-Myers Squibb and their suppliers. These CD media are blended together (with other components) to produce a feed media for the manufacturing process. All samples have to be stored at low temperatures as they are potentially labile and the composition can change with time. Raman measurements are largely insensitive to small temperature fluctuations, however, for complex aqueous based samples the solubility of their individual components can be temperature dependant, and this can be observed when they are removed from cold storage. In some cases, we observed precipitates and/or solid suspensions in defrosted samples and the presence of these suspended solids cause significant changes in the intensity of Raman bands and the spectral profile. Therefore it is critical to ensure complete dissolution prior to measurement.

### 2.2 Instrumentation, data collection and analysis.

Raman spectra were collected with 785 nm excitation using a RamanStation spectrometer (AVALON Instruments Ltd, Belfast, NI, now PerkinElmer). A laser power of  $\sim 70$  mW with an exposure time of  $2 \times 10$  seconds was generally used and spectra were collected at  $8 \text{ cm}^{-1}$  resolution from 250 to  $3311 \text{ cm}^{-1}$ . 100  $\mu\text{L}$  of each sample was pipetted directly into an empty well of an electropolished stainless steel, multiwell plate for measurement (Ryder et al. 2010). To minimise spectral variance, a  $3 \times 3$  sampling grid was used, and spectra were collected in triplicate, over a one month period. For the majority of the analyses undertaken, PLS\_Toolbox 4.0 (Eigenvector Research, Inc., US) and in-house-written code were used. All calculations were performed using MATLAB (The MathWorks, Inc., US) version 7.4. The chemometric methods are described in detail in the supplementary information. Raman data were subjected to a series of sequential pre-processing techniques prior to chemometric analysis. For each measurement (nine spectra), spectra containing cosmic interference were discarded prior to averaging and multiplicative (Martens and Naes 1989). Then an asymmetric weighted least squares algorithm (Haaland and Easterling 1982) was

implemented to automatically remove baseline offsets before the water signal was subtracted using an orthogonal projection procedure (Lorber 1986). The resulting data exhibit less extraneous sources of variance and more linear response of the variables. For quantitative analysis and SIMCA, first order derivative (Savitzky and Golay 1964) and normalisation to the band of maximum intensity was applied to all spectra to further improve the data for analysis by accentuating analyte signals while reducing instrumental effects.

### 3. Results and Discussion

#### 3.1 Spectral analysis.

The aqueous solutions of these CD media ingredients proved challenging to analyse because of the complexity (Table 1) and relatively low concentrations (1-2% dissolved solids). Any significant sample processing, such as pre-concentration or evaporation to dryness could potentially cause large changes, as well as introducing a significant time and handling penalty for high throughput analysis. Therefore, the analytical strategy was to analyse all samples as supplied using three different methods, NIR, MIR-ATR, and Raman spectroscopy. For NIR and MIR data, the very strong water signal and low analyte concentrations, ensured that the compositional information was almost fully obscured, hindering chemometric analysis (data not shown). Raman spectroscopy gave the best spectral data with which to implement robust chemometric analyses suitable for quality control purposes. Figure 1 shows the Raman spectra of the different sample types, and it is clear that the materials can be easily discriminated. There are significant difference in terms of spectral information, with CD-A1 and CD-S2 having highly detailed Raman spectra with relatively strong signals, while CD-A2, CD-S1, and eRDF solutions have much weaker and/or less detailed analyte bands. A second important factor about these spectra is the very significant baseline fluctuations and intensity variations which are evident for all samples. The likely causes of these effects are small changes in sample placement, focal distance, and other instrument variations that are difficult to eliminate. Contributory factors are the long exposures involved and the inherent weakness of the analyte signal relative to the water solvent. We are confident that these baseline variances are not due to intrinsic changes in the sample composition. These effects have to be removed from the data prior to undertaking chemometric analysis and we have determined that the optimal data pre-processing regime is baseline correction, followed by water elimination, then a first order derivative, and finally normalisation. The final important factor, about these spectra is the very significant contribution of water to the Raman spectra (Figure 1f). When the intensity of the water bending bands at  $\sim 1600\text{ cm}^{-1}$  are compared to the analyte band intensities, it is clear that they are often of greater magnitude. For chemometric data analysis it is necessary to remove the water signal using an orthogonal projection procedure. Figure 2 shows the Raman spectra after water-elimination and it is obvious that the analyte bands are more visible and easier to interpret which is important for analysing subtle changes in these complex mixtures.

Using several robust principal component analysis (ROBPCA) (Hubert et al. 2005) studies, the  $707\sim 1853\text{ cm}^{-1}$  spectral region was identified as containing the most useful information for these types of materials (Figure 2). The higher wavenumber region does not contain much useful information apart from the C–H and O–H stretching vibrations;

furthermore, the data has high levels of variance because of instrumental factors (e.g. detector quantum yield). Below  $\sim 700\text{ cm}^{-1}$ , there are contributions from scatter that shows up as large sloping baseline variances. Eliminating these regions leads to improved calibration models, more easily analysed spectra and smaller residual levels. One important point to note is that some samples have very little spectral detail (Figure 2b & 2c) while others have multiple bands, thus, for an integrated analytical solution that is suitable for all cell culture media components, one must be able to deal with both of these generic sample types, and their high-detail (HD) and low-detail (LD) spectra. A more detailed discussion of the spectral band assignments is provided in the attached supplemental information (*section S3.1*).

### 3.2 Reproducibility

One of the key factors which determine the efficacy of Raman based methods for the analysis of CD media is data reproducibility. The data were collected in triplicate for all samples over one month. Reproducibility was measured by PCA studies on each of the three separate Raman data sets and by comparing the PCs obtained from each dataset. In all cases reproducibility was excellent, as shown by the results from the CD-S2 and CD-A2 (Table 2) samples which have HD (S2) and LD (A2) spectra respectively. Figure 3 compares the PCA results from CD-S2 three runs of sample data. Individual comparison plots for each of the first three significant PCs are shown and their calculated correlation coefficients are presented in Table 2. One can see that these individual PCs from the three different data collection runs are nearly identical as the calculated coefficients equal or approach unity, indicating satisfactory reproducibility. For the lower quality CD-A2 spectra, the first three PCs obtained from the PCA study were compared (Table 2, and supplemental information) and the only mathematically observable difference between the different data collections was in PC3. However, this difference is not very significant because PC3 explains a mere 0.01% of the total variance in the data. This PC is therefore much more liable to include instrumental artifacts like minor baseline fluctuations and noise. Similar results were obtained for CD-A1, CD-S1, and eRDF (data not shown), thus Raman spectroscopy is suitable for sample quality and identification analysis by robust chemometric methods.

### 3.3 Analysis of spectral variance and sample quality testing.

The next step in an integrated analytical method is to study the sources of spectral variance, and then use this information to produce a sample quality testing model. Uncovering the sources of variance within the spectral fingerprint region of individual sample classes is key to achieving this. The CD-A1 samples (Figure 4a) are used as an example, to show how PCA can be used to determine sample quality parameters. PCA (Figure 4b) shows that only three PCs were required to explain most of the spectral variance. In the scores plot (Figure 4b), we find that three samples (#1, #3 and #4), are very different, and two samples, #17 & #30, are slightly different from the main group of samples. The spread of the main group samples also indicates that there is a significant level of inherent variability in these materials. This implies that there exists a significant level of product variation from lot to lot of the CD-A1 material and this may represent a sample quality issue. The Hotelling's  $T^2$  statistic within the PCA subspace and the Q residual to the PCA subspace calculated for each measurement were compared with the  $Q_\alpha$  and  $T_\alpha^2$  thresholds once the number of PCs had

been specified to yield an identification of the outliers (samples and inconsistent individual spectra). A 95% confidence level was utilised to determine the two critical thresholds:  $Q_\alpha=2.25 \times 10^{-2}$  and  $T^2_\alpha=8.30$ . The  $Q$  versus  $T^2$  statistics plot (Figure 4c) is now used to provide a quantitative assessment of the anomalous samples (outliers). Samples #1 and #3 possess large  $T^2$  values which exceed the 95% limit of  $T^2_\alpha$ , while sample #4 possesses a  $Q$  residual beyond the defined  $Q_\alpha$  threshold. In the context of CD media quality testing, this  $Q$  vs.  $T^2$  analysis can be used to either reject batches or select these samples for more comprehensive testing. In order to uncover why these samples were identified as abnormal, a closer examination was made in terms of the variance explained by each individual PC using a representative normal sample (#11) was used for comparison. Table 3 shows that the outlier samples (#1 & #3) have much higher explained variance in PC2 and PC3 compared to the normal sample #11. Outlier #4 is a little less different in terms of PC2 and 3. Figure 4d shows that the most significant features in PC2 and PC3 occur in the  $\sim 800$  to  $\sim 1000$   $\text{cm}^{-1}$  and  $1250$  to  $\sim 1850$   $\text{cm}^{-1}$  spectral regions.

Figure 5a shows a comparison between the Raman spectra (from run1) of sample #1 with #11, and it's clear that the differences are small, and relate predominantly to a higher baseline for the outlier, sample #1. Even in the normalised first derivative plot it is difficult to ascribe the spectral band differences to specific constituents. Even when we consider the significant PCs (Figure 5b) which describe the sample differences, it is not altogether obvious which specific components of the samples are causing the spectral differences. Figure 5b also includes the difference spectrum obtained by subtracting the spectrum of #11 from #1. This difference spectrum is somewhat similar to PC2 and PC3 as is evident from the correlation coefficients of 0.70 and 0.62, respectively. When combined, PC2 and PC3 account for 88.12% (namely, 49.24% and 38.88% respectively) of the variance in the difference spectrum. This highlights one of the very significant difficulties one faces when dealing with complex mixtures, in that it can be very difficult to ascribe clearly identities to the species causing the differences. In many cases we see the combination of spectral changes from multiple species all varying in relatively small amounts. The other outlier, #3, behaved much in the same way as sample #1, indicating a similar source of variation.

Figures 5c and 5d show the difference spectra between the representative normal sample #11 and the outliers #1 and #30 respectively. It is evident that sample #1 shows significant increases in intensity across the entire spectral range compared to sample #11. The greatest difference is the sharp peak at  $\sim 820$   $\text{cm}^{-1}$  which based on our earlier tentative peak assignments would appear to coincide with an increase in L-aspartic acid concentration (or some other amino acids with C-C skeletal vibrations at this wavenumber). The other alternative is tyrosine which has a band in this region, but the tyrosine vibration is a doublet which this peak clearly is not. Furthermore, we know that the L-aspartic acid concentration in these samples is relatively high. The other large positive differences in the  $1350$  to  $\sim 1500$   $\text{cm}^{-1}$  range ( $\text{CH}_2$  scissoring and  $\text{CH}_3$  deformation vibrations in aliphatic amino acids) and that at  $1730$   $\text{cm}^{-1}$  ( $\text{C}=\text{O}$  stretch) would tend to support the view that the concentrations of some amino acids are increased above the normal range in sample #1. In Figure 5b the spectral changes are relatively small which is unsurprising given the fact that sample #30 was only a more minor outlier (Figure 4c, shows that it's within the 95% confidence limits) due to its

combination of high PC3 and low PC2 scores (Figure 4b). In particular, for the difference spectrum, there is an increase in intensity at  $803\text{ cm}^{-1}$  but the overall trend is decrease in intensity over the  $800$  to  $\sim 1750\text{ cm}^{-1}$  range. These decreases can be ascribed as above to a possible variation in the relative concentrations of several amino acids in sample #30. The increase observed at  $803\text{ cm}^{-1}$  is unusual as there is no peak evident in the CD-A1 spectra at this wavenumber. Its presence is most likely due to increased concentration of some amino acid with C-C skeletal vibrations in this region coupled with a decrease in the relative concentrations of other amino acids whose vibrations in this region may have masked this peak in the spectra, e.g. L-tyrosine, L-aspartic acid.

Comparing the run 1 residuals for samples #4 and #11 (Figure 5e and Table 3) shows that there are considerable differences, and that there is still structure in the residuals unexplained by the PCA model. This indicates that there is still spectral information contained in these residuals, and that it is significantly higher (0.38% explained variance) for the outlier (#4) than for the representative normal sample (0.06% explained variance) and this is the reason for the discrimination. In sample #4 the source of this significant spectral difference originates from a combination of relatively higher concentrations (compared to specification) of L-aspartic acid, L-tyrosine, L-hydroxyproline, L-leucine, L-phenylalanine, L-proline, and L-valine, and relatively lower concentrations of L-asparagine, L-glutamic acid and L-lysine. In the #4-#11 difference spectrum (Figure 5f) one observes notable intensity increases at  $820\text{ cm}^{-1}$ , and over two ranges of  $850\sim 1100$  and  $1350\sim 1550\text{ cm}^{-1}$ . This is accompanied by a decrease in Raman intensity between  $1580$  and  $1800\text{ cm}^{-1}$ . Therefore it's clear that the outliers #1, #3, and #4 are very different, while samples (#17 & #30) are also significantly different from the main batch of samples. Identifying the specific source of variation from the Raman spectra is not always possible for these complex mixtures, but this unsupervised learning method does provide for the rapid identification of suspect samples, which can then be analysed more comprehensively using chromatographic methods.

Overall, we identified 87 (out of 423) outlying measurements by PCA and these were classified into two groups: Group 1 outliers (22 samples) are those sample measurements which proved to be repeatedly outlying (2 from 3 or 3 from 3), and were adjudged to result from real compositional changes, and Group 2 outliers consisted of samples which had 1 out of 3 replicate measurements was anomalous and as such are most due to instrumental or experimental effects (Table 4, & supplemental information). Thus only 84% of samples and 80% of spectra pass the initial screening step for outliers and inconsistent spectra and using these filtered sample sets, definitive PCA models for each sample class were generated for identification and quality assessment purposes (Table 4). The high rejection rate ( $\sim 1$  in 5 spectra/samples) is due to a combination of effects, the most significant of which variability in sample composition, which is the fundamental problem facing the manufacture of consistent complex cell culture media. For each model, the  $T^2$  and Q statistics of individual samples were compared with the  $Q_\alpha$  and  $T^2_\alpha$  thresholds for an evaluation of sample quality. In an industrial testing context, these models can now be used as the basis of an automated sample ID and quality classification method. The fact these samples originate from a current industrial manufacturing process, limits the dataset size. This is the reason why triplicate spectral data from the same sample were utilised in the PCA models. In particular cases

where the data are clustered or otherwise oddly distributed due to some non-normal cause,  $T^2$  statistics may break down. However, Q residual statistics are still well-behaved.

### 3.6 Combined classification and quality analysis for QC.

Two of the key requirements of good biopharmaceutical manufacturing practice are first the correct identification of raw materials and blended materials, and second the assessment of sample quality prior to use. The complexity of the materials used for cell culture media poses a distinct challenge for both these tasks. Ideally one requires a rapid, inexpensive method that requires minimal sample preparation to both confirm identity and check for product quality and/or consistency. The five PCA models (using only the 707~1853  $\text{cm}^{-1}$  spectral fingerprint region) generated in the preceding section can now be used as the basis of an integrated ID and quality check (see supplemental information for more details). Using a simple, supervised learning, SIMCA model (Figure 8a) based on the normalised, baseline corrected, background subtracted, and first order derivative Raman spectra, one can generate an automated ID method for these complex mixtures. Five clearly distinct classes were obtained and a variance analysis (Table 4, columns 6-10) shows the calculated within-class (WC) and between-class (BC) variances. The WC and BC variances vary greatly in terms of magnitude, with the former being quite small (all  $< 0.3$ ), whereas the latter are quite large (all  $> 2.9$ ). This indicates that SIMCA offers a reliable method for the identification of these complex mixtures using Raman spectroscopy. This also gives an indication about sample quality as the WC variance will be larger for the samples which have the largest compositional variation. Here, eRDF gives the largest degree of variation with a WC variance nearly 3.5 times that of the next worse sample type, CD-S1.

## 4. Conclusions

This study demonstrate that by using relatively simple and well established chemometric methods in conjunction with Raman spectroscopy one can generate a simple and effective identification and quality assessment methodology for the complex constituents of mammalian cell culture media. Raman spectroscopy proved to be highly advantageous for this purpose and spectral reproducibility was excellent once appropriate pre-processing techniques were applied. The Raman method also has the advantages of minimal sample preparation, the ability to analyse aqueous samples, availability of portable instrumentation, and the ability for aseptic sampling through transparent packing. In an industrial context, we have demonstrated that one can use a combination of both supervised and unsupervised learning methods for data analysis (Figure 8b). Unsupervised methods like PCA are first used to build up a picture of sample variances and determine which samples should be used for identification and quality assessment purposes. Second, once the correct samples have been selected, definitive PCA models can be constructed for use in a supervised learning method like SIMCA for the routine identification of incoming materials. Finally, a more detailed PCA model can be utilised to assess product quality before usage. This model can incorporate batch record data (product yield and performance) from the samples used in the model to predict whether or not the incoming media batches will generate good or bad product yield. In operational use, this combination of Raman spectroscopy and chemometrics

can be sequentially improved as more samples are analysed and performance data obtained, eventually one can build more refined PCA models suitable for process control. This has obvious advantages in “quality by design” by allowing process operators more control over the quality of the cell culture mixtures.

## 5. Acknowledgements

This work was funded by the Irish Industrial Development Authority and Bristol-Myers Squibb (BMS) under the Centre for Bioanalytical Sciences, a collaboration between the National University of Ireland, Galway, Dublin City University, and BMS. We thank Lindy Smith (BMS) for organising the sample transfer.

## References

- Archibald DD, Kays SE, Himmelsbach DS, Barton FE. 1998. Raman and NIR spectroscopic methods for determination of total dietary fiber in cereal foods: A comparative study. *Applied Spectroscopy* 52(1):22-31.
- Becker T, Hitzmann B, Muffler K, Portner R, Reardon KF, Stahl F, Ulber R. 2007. Future aspects of bioprocess monitoring. *White Biotechnology*. Berlin: Springer-Verlag Berlin. p 249-293.
- Cannizzaro C, Rhiel M, Marison I, von Stockar U. 2003. On-line monitoring of *Phaffia rhodozyma* fed-batch process with in situ dispersive Raman spectroscopy. *Biotechnology and Bioengineering* 83(6):668-680.
- Carey PR. 1998. Raman spectroscopy in enzymology: the first 25 years. *Journal of Raman Spectroscopy* 29(1):7-14.
- Castro JL, Montañez MA, Otero JC, Marcos JI. 1995. SERS and vibrational spectra of aspartic acid. *Journal of Molecular Structure* 349:113.
- Clark RJH, Cridland L, Kariuki BM, Harris KDM, Withnall R. 1995. Synthesis, Structural Characterization and Raman-Spectroscopy of the Inorganic Pigments Lead-Tin Yellow Type-I And Type-II and Lead Antimonate Yellow - Their Identification on Medieval Paintings and Manuscripts. *Journal of the Chemical Society-Dalton Transactions*(16):2577-2582.
- Cooper JB, Wise KL, Welch WT, Sumner MB, Wilt BK, Bledsoe RR. 1997. Comparison of near-IR, Raman, and mid-IR spectroscopies for the determination of BTEX in petroleum fuels. *Applied Spectroscopy* 51(11):1613-1620.
- Davis JM. 2002. *Basic Cell Culture*. J. Davis, editor: Oxford University Press. 416 p.
- De Gelder J, De Gussem K, Vandenabeele P, Moens L. 2007. Reference database of Raman spectra of biological molecules. *Journal of Raman Spectroscopy* 38(9):1133-1147.
- Edwards HGM, Lawson EE, Barry BW, Williams AC. 1997. Applications of FT-Raman spectroscopy to the study of healthy and diseased human skin tissue. *Spectroscopy of Biological Molecules: Modern Trends*:469-470.

Rapid Characterisation and Quality Control of Complex Cell Culture Media Solutions using Raman Spectroscopy and Chemometrics. B. Li, P.W. Ryan, B.H. Ray, K.J. Leister, N.M.S. Sirimuthu, and A.G. Ryder. *Biotechnology and Bioengineering*, 107(2), 290-301, (2010). DOI: [10.1002/bit.22813](https://doi.org/10.1002/bit.22813).

Ellis D, Goodacre R. 2006. Metabolic fingerprinting in disease diagnosis: biomedical applications of infrared and Raman spectroscopy. *ANALYST* 131(8):875-885.

Everall NJ, Chalmers JM, Ferwerda R, Vandermaas JH, Hendra PJ. 1994. Measurement of Poly(Aryl Ether Ether Ketone) Crystallinity in Isotropic and Uniaxial Samples Using Fourier-Transform Raman-Spectroscopy - A Comparison of Univariate And Partial Least-Squares Calibrations. *Journal of Raman Spectroscopy* 25(1):43-51.

Frank CJ, Redd DCB, Gansler TS, McCreery RL. 1994. Characterization Of Human Breast Biopsy Specimens With Near-Ir Raman-Spectroscopy. *Analytical Chemistry* 66(3):319-326.

Geladi P. 2003. Chemometrics in spectroscopy. Part 1. Classical chemometrics. *Spectrochimica Acta Part B-Atomic Spectroscopy* 58(5):767-782.

Haaland DM, Easterling RG. 1982. Application Of New Least-Squares Methods For The Quantitative Infrared-Analysis Of Multicomponent Samples. *Applied Spectroscopy* 36(6):665-673.

Hashimoto A, Yamanaka A, Kanou M, Nakanishi K, Kameoka T. 2005. Simple and rapid determination of metabolite content in plant cell culture medium using an FT-IR/ATR method. *Bioprocess and Biosystems Engineering* 27(2):115-123.

Hubert M, Rousseeuw PJ, Vanden Branden K. 2005. ROBPCA: A new approach to robust principal component analysis. *Technometrics* 47(1):64-79.

Jackson JE. 1980. Principal Components and Factor-Analysis .1. Principal Components. *Journal of Quality Technology* 12(4):201-213.

Jackson JE. 1981a. Principal Components and Factor-Analysis .2. Additional Topics Related to Principal Components. *Journal of Quality Technology* 13(1):46-58.

Jackson JE. 1981b. Principal Components and Factor-Analysis .3. What is Factor-Analysis. *Journal of Quality Technology* 13(2):125-130.

Jayme D, Watanabe T, Shimada T. 1997. Basal medium development for serum-free culture: A historical perspective. *Cytotechnology* 23(1-3):95-101.

Krishnan K, Plane RA. 2002. Raman spectra of ethylenediaminetetraacetic acid and its metal complexes. *Journal of the American Chemical Society* 90(12):3195.

Lavine B, Workman J. 2006. Chemometrics. *Analytical Chemistry* 78(12):4137-4145.

Lavine B, Workman J. 2008. Chemometrics. *Analytical Chemistry* 80(12):4519-4531.

Lawson EE, Barry BW, Williams AC, Edwards HGM. 1997. Biomedical applications of Raman spectroscopy. *Journal of Raman Spectroscopy* 28(2-3):111-117.

Lee H, Boccazzi P, Gorret N, Ram R, Sinskey A. 2004. In situ bioprocess monitoring of *Escherichia coli* bioreactions using Raman spectroscopy. *Vibrational Spectroscopy* 35(1-2):131-137.

Leger MN, Ryder AG. 2006. Comparison of derivative preprocessing and automated polynomial baseline correction method for classification and quantification of narcotics in solid mixtures. *Applied Spectroscopy* 60(2):182-193.

Rapid Characterisation and Quality Control of Complex Cell Culture Media Solutions using Raman Spectroscopy and Chemometrics. B. Li, P.W. Ryan, B.H. Ray, K.J. Leister, N.M.S. Sirimuthu, and A.G. Ryder. *Biotechnology and Bioengineering*, 107(2), 290-301, (2010). DOI: [10.1002/bit.22813](https://doi.org/10.1002/bit.22813).

Liu LZ, Arnold MA. 2009. Selectivity for glucose, glucose-6-phosphate, and pyruvate in ternary mixtures from the multivariate analysis of near-infrared spectra. *Analytical And Bioanalytical Chemistry* 393(2):669-677.

Lorber A. 1986. Error Propagation and Figures of Merit for Quantification by Solving Matrix Equations. *Analytical Chemistry* 58(6):1167-1172.

Marteau P, Adar F, ZanierSzydowski N. 1996. Application of remote Raman measurements to the monitoring and control of chemical processes. *American Laboratory* 28(16):H21-&.

Martens H, Naes T. 1989. *Multivariate Calibration*. Wiley, editor. New York.

Massart D, Vandeginste BGM, Buydens LMC, De Jong S, Lewi PJ, Smeyers-Verbeke J. 1997. *Handbook of Chemometrics and Qualimetrics: Part A*. Amsterdam: Elsevier.

McCreery RL, Frank CJ, Redd DCB. 1995. Raman spectroscopy of human biopsy specimens. *Advances in Fluorescence Sensing Technology II* 2388:90-98.

McGovern A, Broadhurst D, Taylor J, Kaderbhai N, Winson M, Small D, Rowland J, Kell D, Goodacre R. 2002. Monitoring of complex industrial bioprocesses for metabolite concentrations using modern spectroscopies and machine learning: Application to gibberellic acid production. *Biotechnology and Bioengineering* 78(5):527-538.

Mosier-Boss PA, Lieberman SH. 2000. Detection of Nitrate and Sulfate Anions by Normal Raman Spectroscopy and SERS of Cationic-Coated, Silver Substrates. *Applied Spectroscopy* 54(8):1126-1135.

Murakami H, Shimomura T, Nakamura T, Ohashi H, Shinohara K, Omura H. 1984. Development of a basal medium for serum-free cultivation of hybridoma cells in high density. *J. Agricult. Chem. Soc. Japan* 56:575-583.

Newman DJ, Cragg GM. 2007. Natural Products as Sources of New Drugs over the Last 25 Years. *Journal of Natural Products* 70(3):461-477.

Pollard JW. 1997. *Basic Cell Culture*. *Basic Cell Culture Protocols*. p 1.

Rhiel M, Ducommun P, Bolzonella I, Marison I, von Stockar U. 2002. Real-time in situ monitoring of freely suspended and immobilized cell cultures based on mid-infrared spectroscopic measurements. *Biotechnology and Bioengineering* 77(2):174-185.

Riley MR, Rhiel M, Zhou XJ, Arnold MA, Murhammer DW. 1997. Simultaneous measurement of glucose and glutamine in insect cell culture media by near infrared spectroscopy. *Biotechnology and Bioengineering* 55(1):11-15.

Roychoudhury P, Harvey LM, McNeil B. 2006. The potential of mid infrared spectroscopy (MIRS) for real time bioprocess monitoring. *Analytica Chimica Acta* 571(2):159.

Ryan PW, Li B, Shanahan M, Leister, K.J., Ryder AG. 2010. Prediction of Cell Culture Media performance using Fluorescence Spectroscopy. *Analytical Chemistry* 82(4):1311-1317.

Ryder AG, De Vincentis J, Li B, Ryan PW, Sirimuthu NMS, Leister KJ. 2010. A Stainless Steel Multi-Well Plate (SS-MWP) for High Throughput Raman Analysis of Dilute Solutions. *Journal of Raman Spectroscopy*.

Savitzky A, Golay MJE. 1964. Smoothing + Differentiation Of Data By Simplified Least Squares Procedures. *Analytical Chemistry* 36(8):1627-&.

Scarff M, Arnold SA, Harvey LM, McNeil B. 2006. Near Infrared Spectroscopy for bioprocess monitoring and control: Current status and future trends. *Critical Reviews In Biotechnology* 26(1):17-39.

Schroder M, Matischak K, Friedl P. 2004. Serum- and protein-free media formulations for the Chinese hamster ovary cell line DUKXB11. *Journal of Biotechnology* 108(3):279-292.

Shimoyama M, Maeda H, Matsukawa K, Inoue H, Ninomiya T, Ozaki Y. 1997. Discrimination of ethylene vinyl acetate copolymers with different composition and prediction of the vinyl acetate content in the copolymers using Fourier-transform Raman spectroscopy and multivariate data analysis. *Vibrational Spectroscopy* 14(2):253-259.

Shinzawa H, Morita S, Ozaki Y, Tsenkova R. 2006. New method for spectral data classification: Two-way moving window principal component analysis. *Applied Spectroscopy* 60(8):884-891.

Sohn M, Himmelsbach DS, Kays SE, Archibald DD, Barton FE. 2005. NIR-FT/Raman spectroscopy for nutritional classification of cereal foods. *Cereal Chemistry* 82(6):660-665.

Takeda M, Iavazzo RES, Garfinkel D, Scheinberg IH, Edsall JT. 1958. Raman Spectra of Amino Acids and Related Compounds. IX. Ionization and Deuterium Substitution in Glycine, Alanine and beta-Alanine. *Journal of the American Chemical Society* 80(15):3813.

Tsuchihashi H, Katagi M, Nishikawa M, Tatsuno M, Nishioka H, Nara A, Nishio E, Petty C. 1997. Determination of methamphetamine and its related compounds using Fourier transform Raman spectroscopy. *Applied Spectroscopy* 51(12):1796-1799.

Utzinger U, Richards-Kortum RR. 2003. Fiber optic probes for biomedical optical spectroscopy. *Journal of Biomedical Optics* 8(1):121-147.

Vandeginste BGM, Massart D, Buydens LMC, De Jong S, Lewi PJ, Smeyers-Verbeke J. 1998. *Handbook of Chemometrics and Qualimetrics: Part B*. Amsterdam, New York: Elsevier.

Vickers TJ MC. 1991. Quantitative analysis by Raman spectroscopy. . Bulkin BJ GJ, editors. , editor: New York: John Wiley and Sons, Inc. 107-135 p.

Williams KPJ, Everall NJ. 1995. Use of Micro-Raman Spectroscopy for the Quantitative-Determination of Polyethylene Density Using Partial Least-Squares Calibration. *Journal of Raman Spectroscopy* 26(6):427-433.

Wold S, Sjostrom M. 1977. SIMCA: A method for analyzing chemical data in terms of similarity and analogy. In: Kowalski BR, editor. *American Chemical Society Symposium Series 52*. Washington D.C.: American Chemical Society. p 243-282.

Workman J, Koch M, Veltkamp D. 2007. *Process Analytical Chemistry*. *Analytical Chemistry* 79(12):4345-4364.

**Tables:**

**Table 1:** Summary of aqueous cell culture media materials analysed. The number of components refers to the numbers of discrete molecular entities present in each sample class.\*

<b>Material</b>	<b># samples</b>	<b>Description</b>	<b>No. of components</b>
CD-A1	31	Mainly amino acids	23+
CD-A2	31	mainly inorganic salts	33+
CD-S1	28	inorganic and organic salts	7+
CD-S2	30	mostly organic	22+
eRDF	21	Diverse mixture of organic & inorganic materials	31+

\* The exact compositions of these media components cannot be revealed because of IP issues. However, the exact compositions are not required for the purposes of this application.

**Table 2:** Comparison of correlation coefficients and explained variances of the three most significant PCs for the three data collection runs on the CD-S2 and CD-A2 samples.

	<b>CD-S2</b>			<b>CD-A2</b>		
	<b>Run1</b>	<b>Run2</b>	<b>Run3</b>	<b>Run1</b>	<b>Run2</b>	<b>Run3</b>
<i>Coefficient</i> ( <i>Expl. Var.</i> )	<i>PC1</i> (99.94%)	<i>PC1</i> (99.95%)	<i>PC1</i> (99.95%)	<i>PC1</i> (99.92%)	<i>PC1</i> (99.93%)	<i>PC1</i> (99.93%)
PC1 from Run1	1.00	1.00	1.00	1.00	1.00	1.00
PC1 from Run2	1.00	1.00	1.00	1.00	1.00	1.00
PC1 from Run3	1.00	1.00	1.00	1.00	1.00	1.00
<i>Coefficient</i> ( <i>Expl. Var.</i> )	<i>PC2</i> (0.04%)	<i>PC2</i> (0.04%)	<i>PC2</i> (0.04%)	<i>PC2</i> (0.04%)	<i>PC2</i> (0.04%)	<i>PC2</i> (0.04%)
PC2 from Run1	1.00	0.976	0.991	1.00	0.987	0.990
PC2 from Run2	0.976	1.00	0.964	0.987	1.00	0.98
PC2 from Run3	0.991	0.964	1.00	0.990	0.98	1.00
<i>Coefficient</i> ( <i>Expl. Var.</i> )	<i>PC3</i> (0.01%)	<i>PC3</i> (0.01%)	<i>PC3</i> (0.01%)	<i>PC3</i> (0.01%)	<i>PC3</i> (0.01%)	<i>PC3</i> (0.01%)
PC3 from Run1	1.00	0.980	0.991	1.00	0.797	0.857
PC3 from Run2	0.983	1.00	0.975	0.797	1.00	0.956
PC3 from Run3	0.991	0.975	1.00	0.857	0.956	1.00

**Table 3:** Variance explained by the individual PCs with respect to the pre-processed data of six CD-A1 samples from run1, i.e. the main outliers (#1, #3, & #4), minor outliers (#17 & #30) and one representative normal sample (#11).

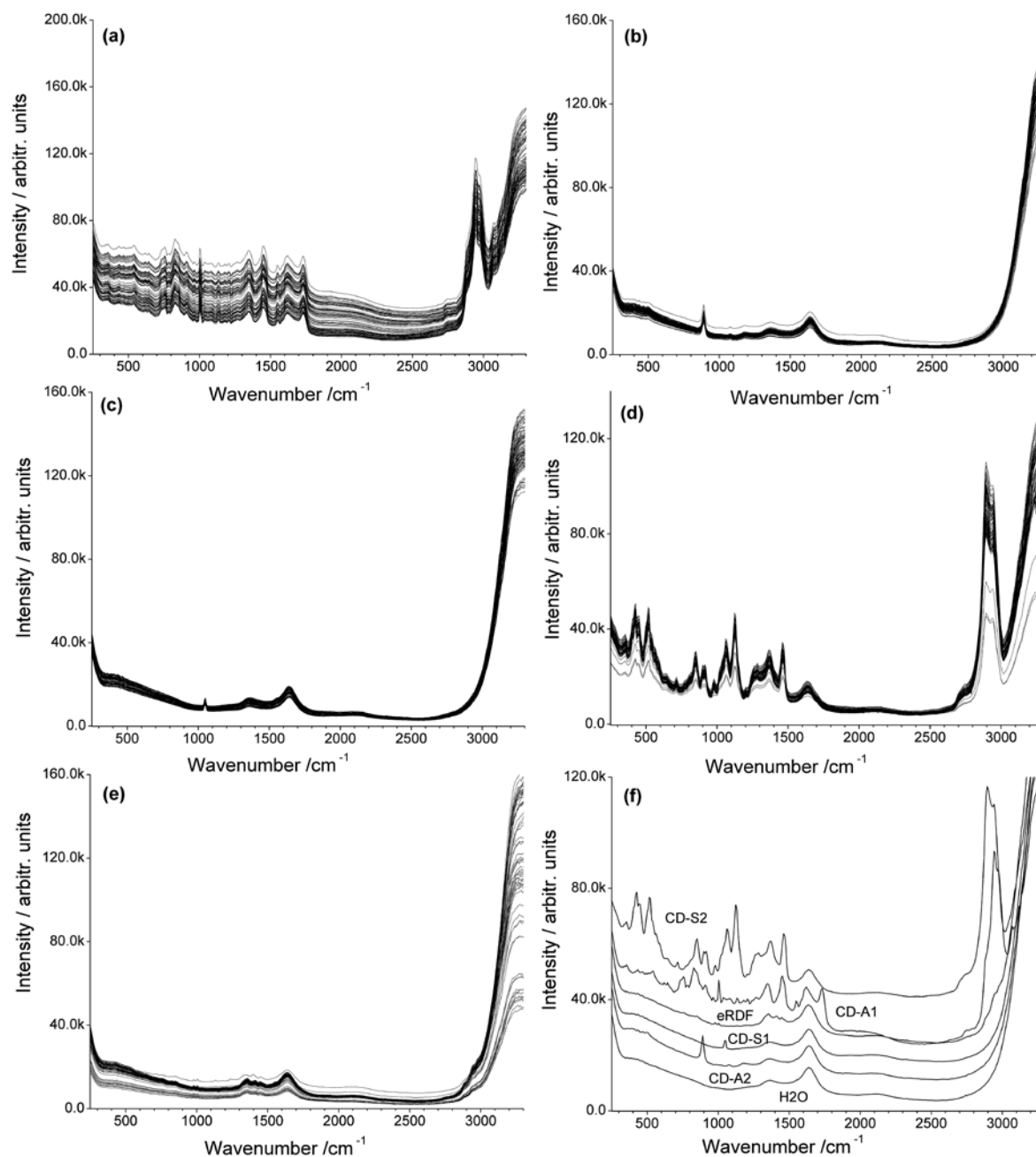
Variance explained (%)	by PC1	by PC2	by PC3	Cumulative
Sample #1	99.46	0.29	0.20	99.95
Sample #3	99.52	0.27	0.15	99.94
Sample #4	99.42	0.07	0.13	99.62
Sample #17	99.81	0.11	0.03	99.95
Sample #30	99.69	0.22	0.02	99.93
Sample #11	99.92	0.01	0.01	99.94

**Table 4:** Columns 2-5 give an overview of the PCA models developed for the CD-A1, CD-A2, CD-S1, CD-S2 and eRDF sample solutions. Columns 6-10 show the SIMCA model performance in terms of variances within and between the sample classes.

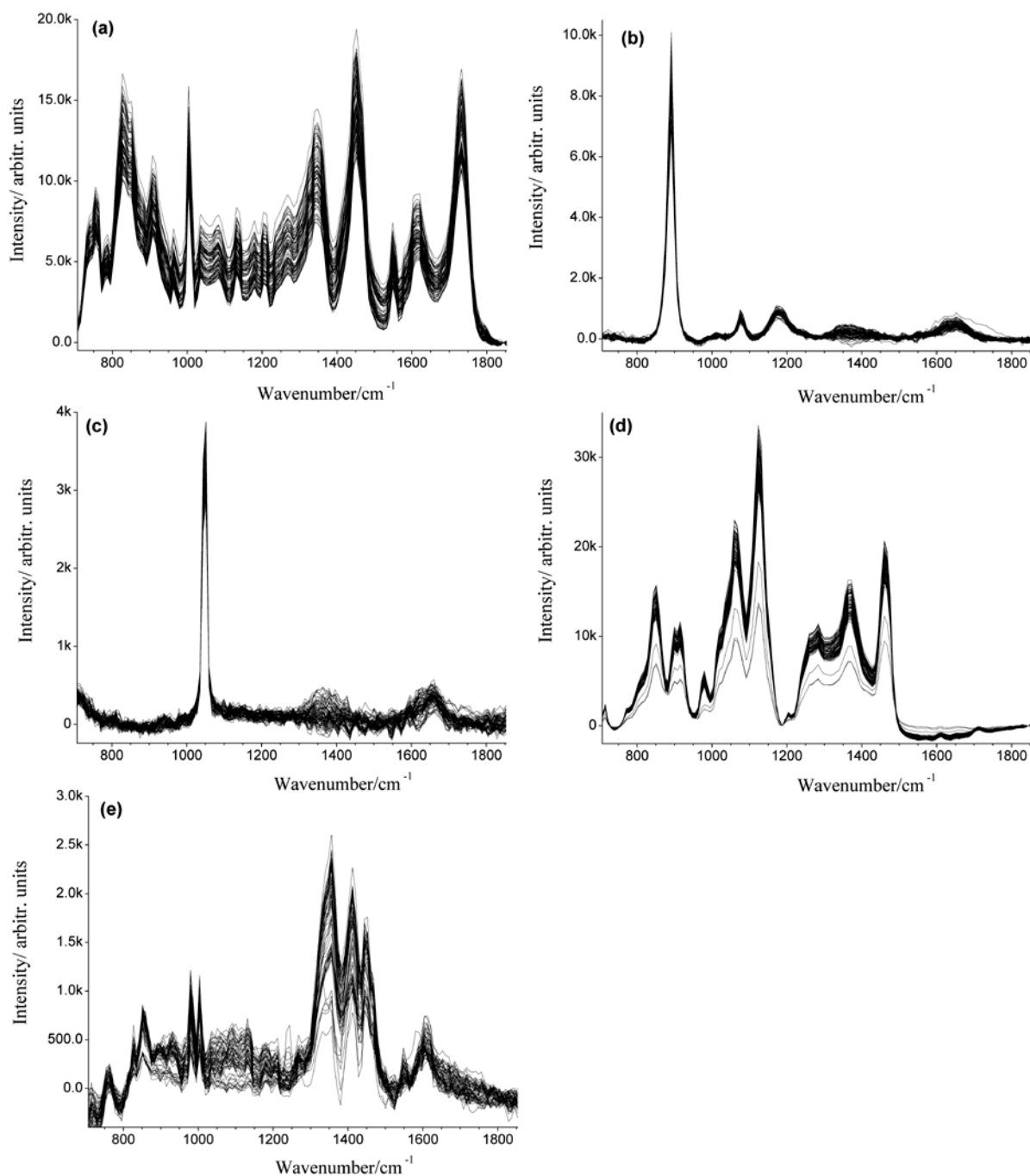
PCA model	# spectra	# samples	PCs	Variance (%)	CD-A1	CD-A2	CD-S1	CD-S2	eRDF
CD-A1	75	26	3	99.97	0.02*	6.45	6.44	4.78	3.22
CD-A2	75	26	3	99.98	6.45	0.01*	3.07	4.42	3.48
CD-S1	73	26	2	99.76	6.44	3.07	0.08*	2.97	3.60
CD-S2	71	25	2	99.99	4.78	4.42	2.97	0.001*	3.49
eRDF	42	16	3	99.33	3.22	3.48	3.60	3.49	0.28*

\* Note that these values correspond to within-class variance.

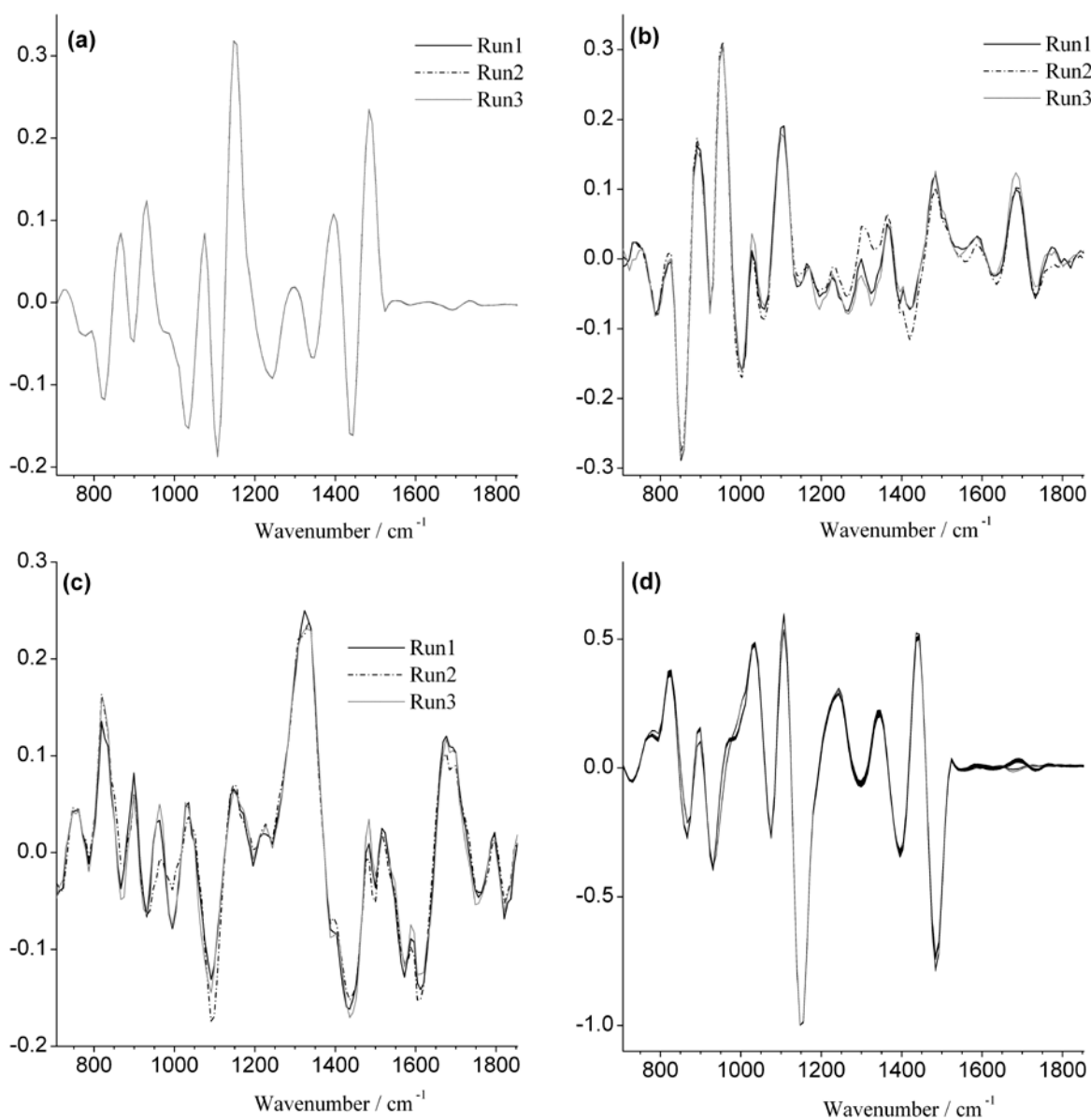
## Figures



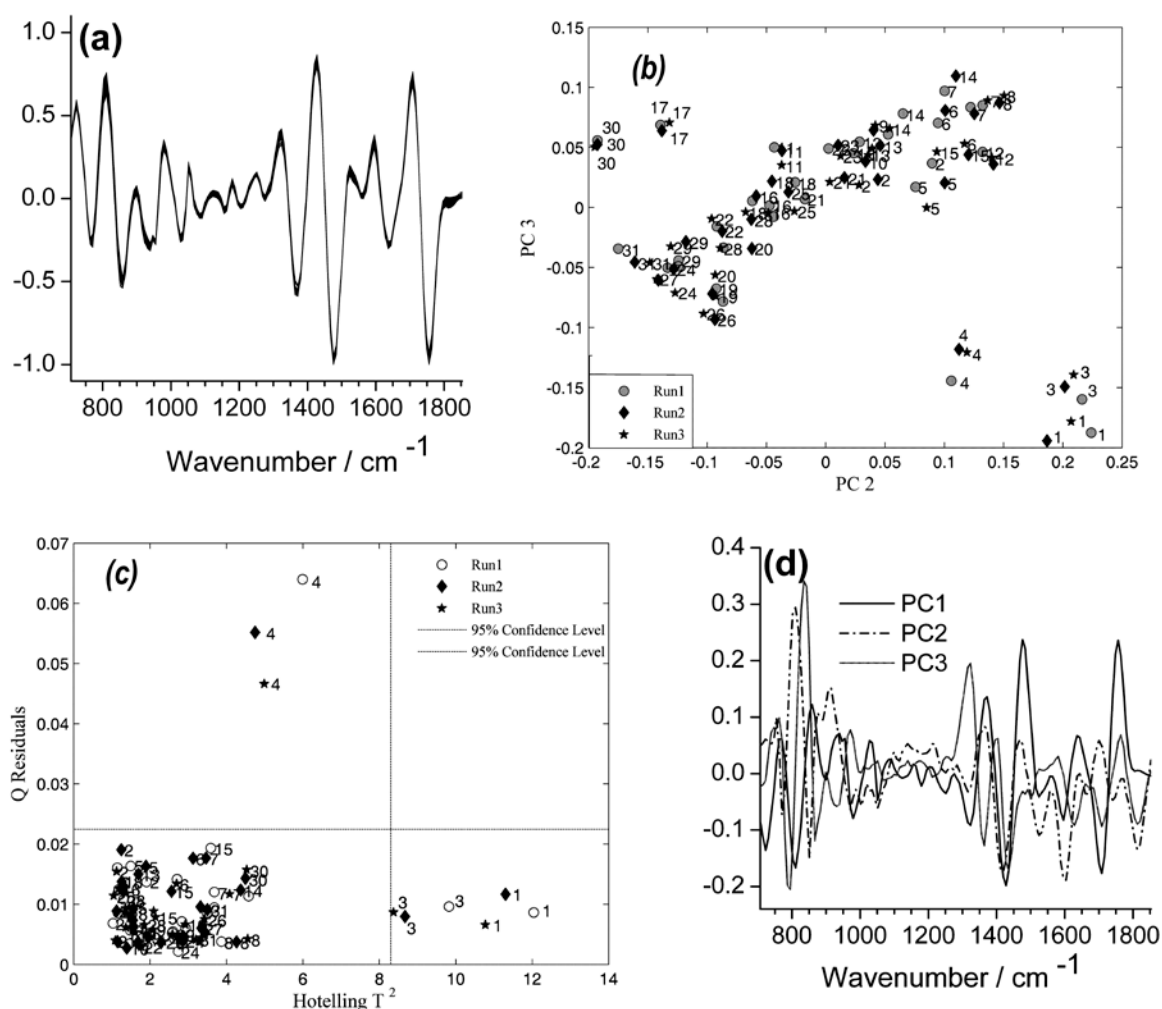
**Figure 1:** Raw Raman spectra of: (a) 93 CD-A1 measurements, (b) 93 CD-A2 measurements, (c) 84 CD-S1 measurements, (d) 90 CD-S2 measurements, (e) 63 eRDF measurements, and (f) spectral comparison of the five solutions. All the data shown were collected over a one month period.



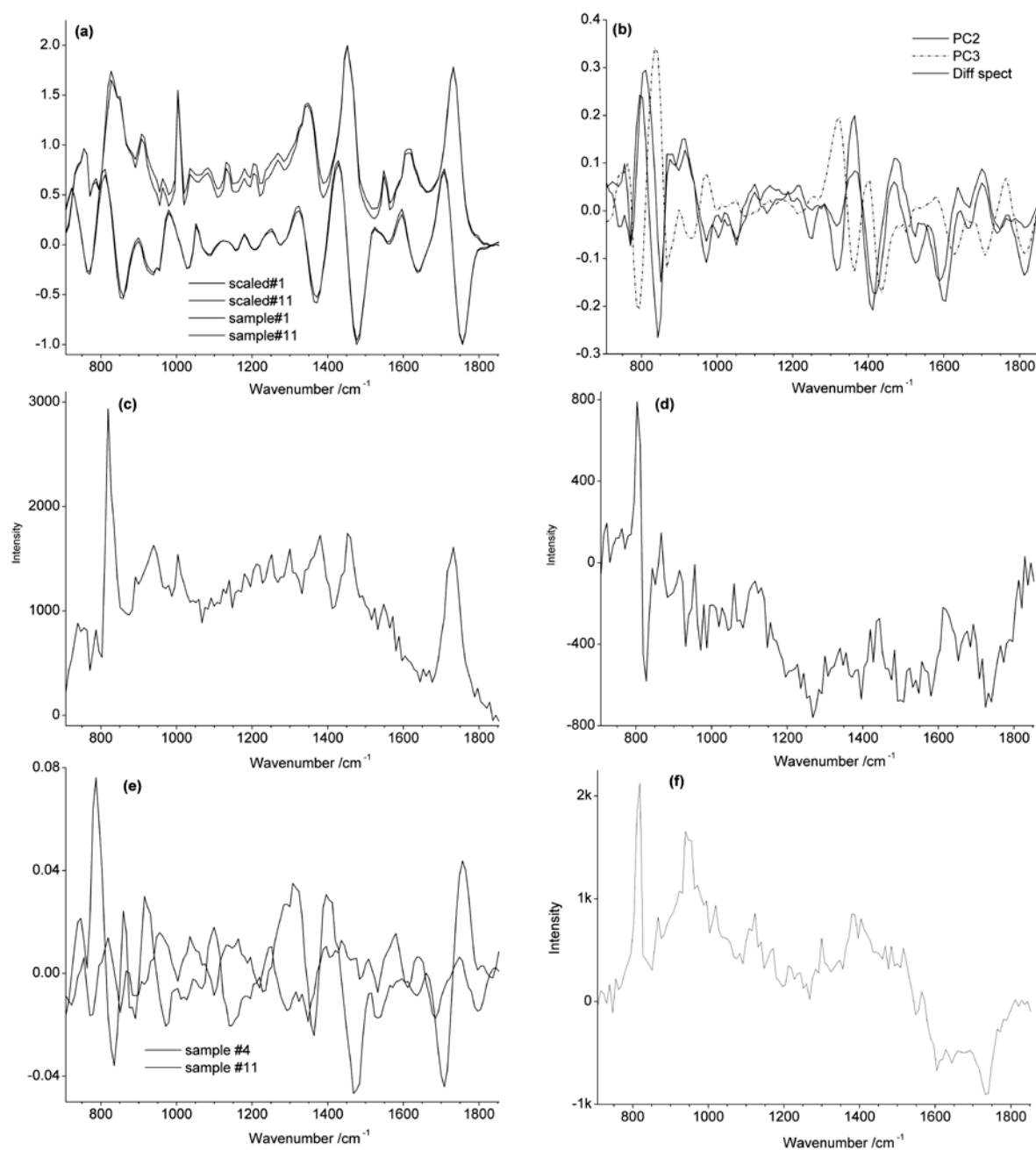
**Figure 2:** Overlaid water-eliminated Raman spectra in the fingerprint region (707 ~ 1853 cm<sup>-1</sup>) for aqueous solutions of: (a) CD-A1, (b) CD-A2, (c) CD-S1, (d) CD-S2, and (e) eRDF.



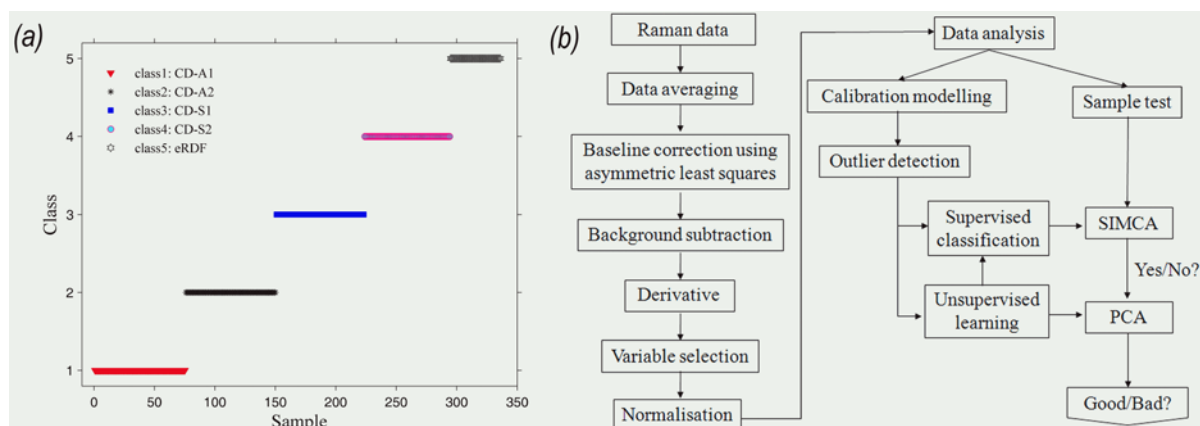
**Figure 3:** Comparison of principal component loadings plots for: (a) PC1, (b) PC2, and (c) PC3, which were obtained by performing PCA analyses on (d) the pre-processed data of 90 CD-S2 measurements from the three different data collection runs. Table 2 gives a numerical assessment of the differences in the PCs for the three data collections.



**Figure 4:** Overview of PCA model on 93 CD-A1 measurements from triplicate data collection runs: (a) the pre-processed data used for the model (707~1853 cm<sup>-1</sup> region), (b) PC2 vs PC3 scores plot, showing the significant outlier samples, (c) Plot of Q and T<sup>2</sup> statistics (dashed boundary defines a 95% confidence limit), again showing the outliers relative to the main group of samples, and (d) Overlay of PC loadings plots for the three major components.



**Figure 5:** Comparison of sample #1 with #11: (a) the scaled water-eliminated Raman spectra from run1 (top), and their first order derivatized-infinity normalised traces (bottom); (b) the resulting loadings plots of the PC2 and PC3 components, and the difference spectrum between first order derivatized-infinity normalised traces; (c) difference spectrum (background subtracted data) between samples #1 and #11; (d) difference spectrum (background subtracted data) between samples #30 and #11; (e) residuals of samples #4 and #11 from run1, showing the spectral variance not explained by the PCA model; and (f) difference spectrum (background subtracted data) between samples #4 and #11.



**Figure 6:** (a) SIMCA classification of the 336 sample measurements using the pre-processed Raman spectra of CD-A1, CD-A2, CD-S1, CD-S2, and eRDF samples. (b) Flow diagram summarising the integrated Raman-Chemometrics quality control methodology. This shows the sequence of steps for data collection and pre-processing, chemometric modelling, sample identification, sample quality assessment, and eventual output decision.

## Supplemental Information:

### SI.1 Chemometric methods of analysis.

A Principal Component Model (PCA) describes major trends in a dataset  $\mathbf{X}$  of  $m$  row observations by  $n$  column variables (e.g., Raman shift/wavenumbers), by using a set of fewer suboptimal principal components (PCs):

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T + \mathbf{E} \quad (1)$$

where  $\mathbf{U}$  is a  $m$ -by- $m$  matrix (i.e., scores),  $\mathbf{V}$  is a  $n$ -by- $n$  matrix (i.e., loadings), and  $\mathbf{\Sigma}$  is a  $m$ -by- $n$  matrix containing magnitude-decreasing singular values ( $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{\min(m,n)} \geq \dots \geq 0$ ), and zero off diagonal elements. The residual  $\mathbf{E}$  contains errors and the part of the data not explained by the PCA model. The superscript T denotes transpose. This process can be fulfilled by the singular value decomposition (SVD) of  $\mathbf{X}$ .

To optimally capture the variations of the data, the loading vectors corresponding to the  $k$  largest singular values are retained within the PCA representation. For each loading, the amount of variance explained by the model is calculated by comparing the estimates  $\sigma_i^2$  ( $i=1, 2, \dots, k$ ) computed by the model with the true variance,  $\Sigma\sigma^2$ :

$$\text{Explained variance (\%)} = \sigma_i^2 / \Sigma\sigma^2 \times 100 \quad (2)$$

Then, these first  $k$  loadings are stacked into a  $n$ -by- $k$  matrix  $\mathbf{P}$ . Using this data it is then possible to calculate two numerical values for the PCA model which can be used to assess sample variability. These are the Hotelling's  $T^2$  statistic, and  $Q$ , a lack-of-fit statistic, which are defined by (Jackson and Mudholkar 1979):

$$T_i^2 = \mathbf{x}_i^T \mathbf{P} \mathbf{\Sigma}_k^{-2} \mathbf{P}^T \mathbf{x}_i \quad (3)$$

$$Q_i = \mathbf{x}_i (\mathbf{I} - \mathbf{P} \mathbf{P}^T) \mathbf{x}_i^T \quad (4)$$

where  $\mathbf{x}_i$  is an observation corresponding to the  $i$ th measurement in the data. The  $T^2$  statistic is a measure of the variation in each sample within the PCA model, while the  $Q$  statistic is a measure of the variation (or difference/residual) of the data outside of the  $k$  PCs included in the model. The combination of both these values are used to quantify how well each sample conforms to the PCA model, and in the context of complex materials analysis, identify outliers.

Statistical confidence limits on  $T^2$  and  $Q$  statistics can be established as follows:

$$T_\alpha^2 = \frac{k(m-1)(m+1)}{m(m-k)} F_\alpha(k, m-k) \quad (5)$$

where,  $\alpha$  means a significant level tested for  $F$ -distribution, for instance,  $\alpha = 0.05$ .

$$Q_\alpha = \Theta_1 \left[ \frac{c_\alpha \sqrt{2\Theta_2 h_0^2}}{\Theta_1} + 1 + \frac{\Theta_2 h_0 (h_0 - 1)}{\Theta_1^2} \right]^{\frac{1}{h_0}} \quad (6)$$

where

$$\Theta_i = \sum_{j=k+1}^{\min(m,n)} \lambda_j^i \quad (i = 1, 2, 3) \quad (7)$$

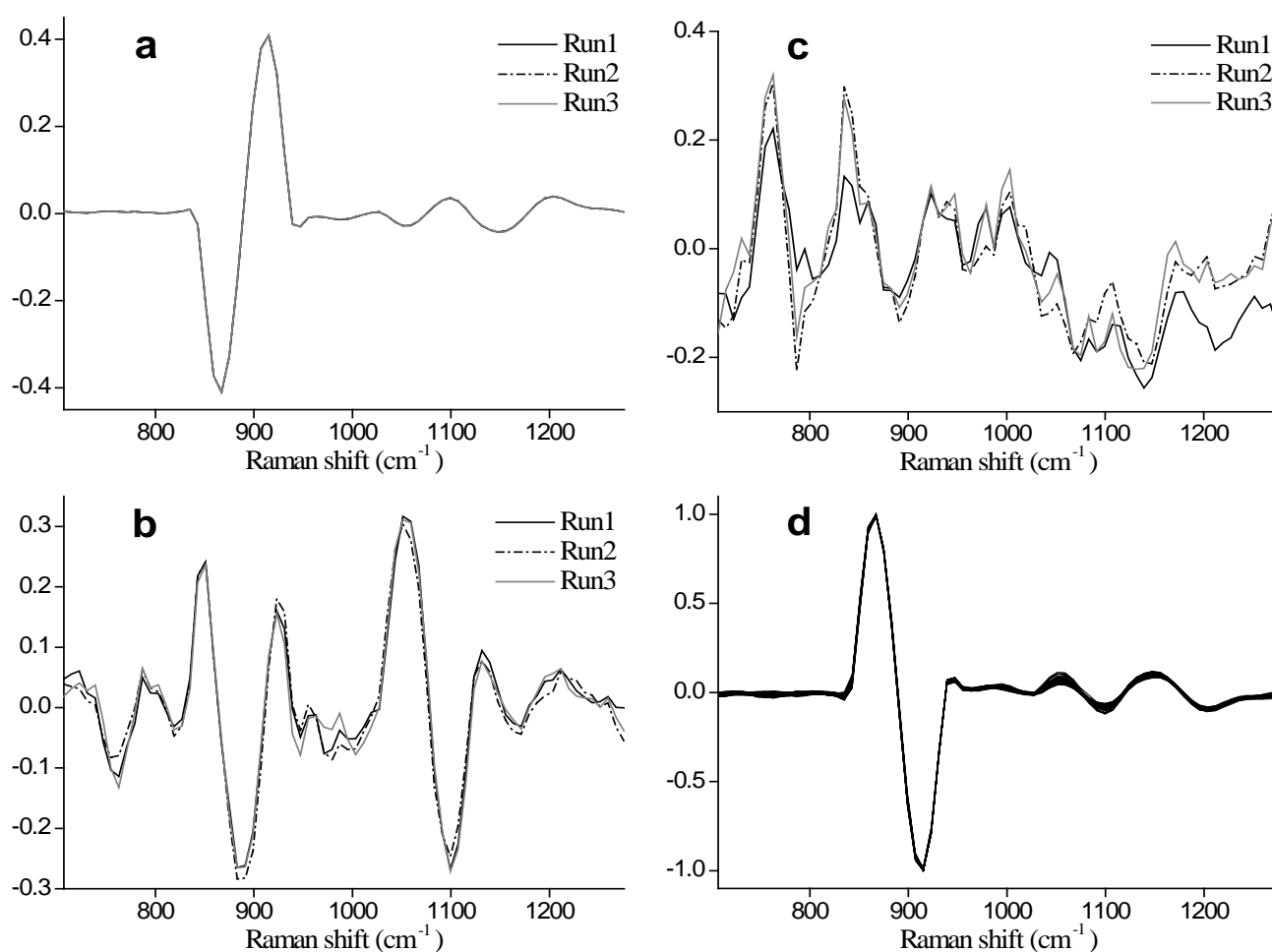
and

$$h_0 = 1 - \frac{2\Theta_1\Theta_3}{3\Theta_2^2} \quad (8)$$

$c_\alpha$  is the normal deviate corresponding to the upper  $(1 - \alpha)$  percentile. The  $T^2$  limit defines an ellipse on the hyperplane spanned by the  $k$  PCs within which the data normally project while the Q limit defines a distance off the hyperplane that is considered unusual based on the data used in the PCA representation. Both the  $T^2$  and Q statistics are calculated and utilised for qualitative evaluation of samples in the data. After identifying the outliers and the samples that should belong in each class model, standard SIMCA was then employed as the classification-identification method for the various media components and its performance was evaluated by means of variance analysis both within and between sample classes.

### *S2.1 Reproducibility Measurements:*

The CD-A2 spectra are not very detailed and possess one relatively strong peak at  $\sim 890 \text{ cm}^{-1}$  (probably from EDTA) together with two minor peaks at  $\sim 1076 \text{ cm}^{-1}$  (carbonate) and  $\sim 1180 \text{ cm}^{-1}$  (unknown). This is not unsurprising since the CD-A2 is a complex mixture of metal salts, most of which are present at very low concentrations ( $< 1 \text{ mg/L}$ ). For the reproducibility investigation, the first three PCs obtained from the PCA study were compared (Figure S-1 and Table 2, main manuscript) and the only mathematically observable difference between the different data collections was in PC3. However, this difference is not very significant because PC3 explains a mere 0.01% of the total variance in the data. This PC is therefore much more liable to include instrumental artifacts like minor baseline fluctuations and noise.



**Figure S-1:** Comparison plots of (a) PC1, (b) PC2, and (c) PC3, which were obtained by performing PCA analyses on (d) the pre-processed data of 93 CD-A2 measurements from the three different data collections.

### S3.1 Tentative spectral band assignments:

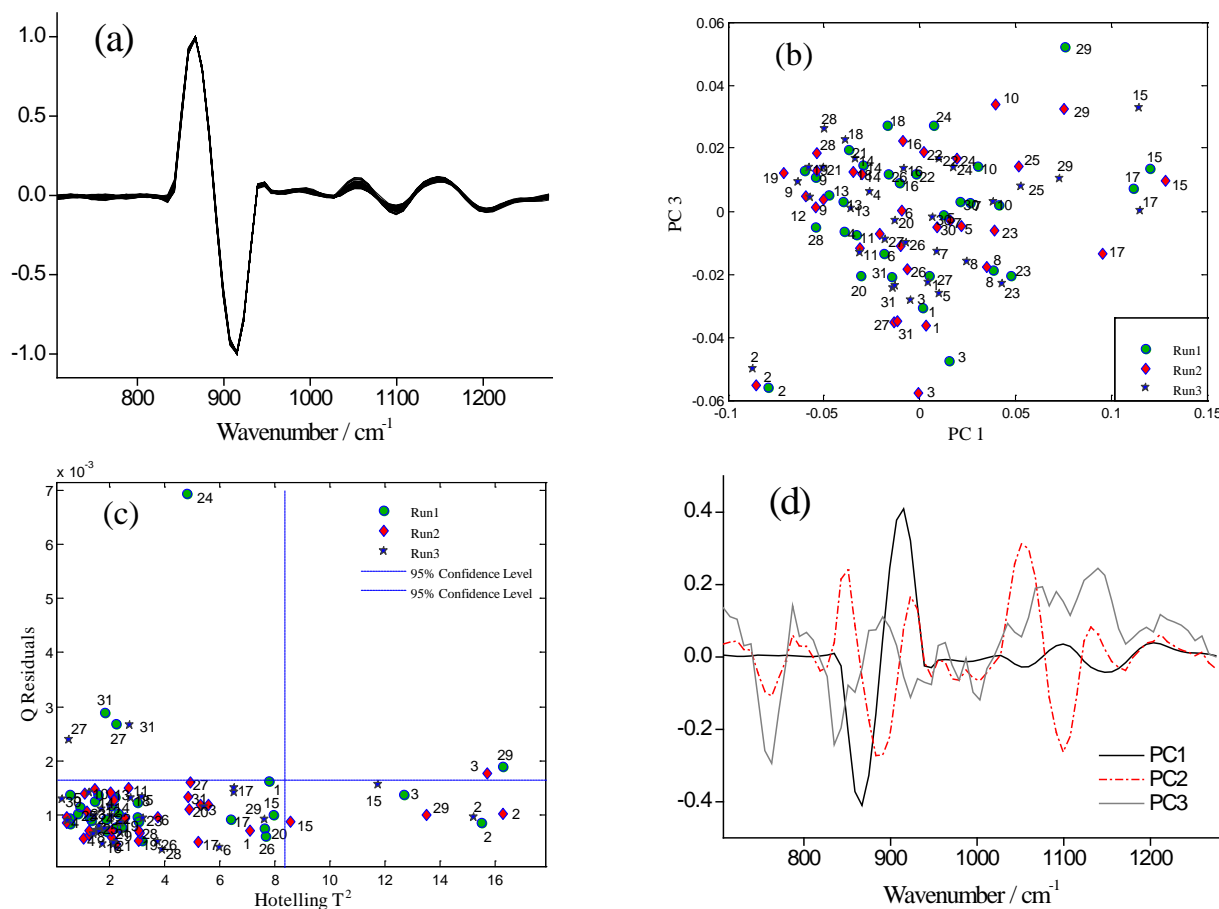
There are a variety of peaks present in the Raman spectra of CD-A1 samples (Figure 2a, main manuscript), with the peaks at 827/845, 1003, 1350, 1450, and 1730 cm<sup>-1</sup> being the most prominent. The CD-A1 samples are aqueous mixtures of various amino acids at low pH, and so the following tentative assignments can be made. The broad peak at 827/845 cm<sup>-1</sup> is most likely the characteristic doublet of L-tyrosine which is a result of Fermi resonance between ring breathing and a ring bending overtone. It is also probable that there is a contribution from C–C skeletal vibrations in this wavenumber range as sharp peaks attributed to such vibrations in the cationic form of amino acids such as L-aspartic acid, glycine and β-alanine have been observed previously in this range (Castro et al. 1995; Takeda et al. 1958), e.g. 822 cm<sup>-1</sup> in low pH aspartic acid solutions (Castro et al. 1995). The sharp peak at 1003 cm<sup>-1</sup> can be ascribed to symmetric ring C–C stretching in L-phenylalanine or alternatively indole ring breathing in L-tryptophan as both of these modes result in intense signals within this characteristic region. The broad peak at 1350 cm<sup>-1</sup> is more difficult to directly assign but L-tryptophan ring bending vibrations are commonly observed in this region, as are CH<sub>2</sub> scissors modes from aliphatic side chain amino acids (e.g. L-aspartic acid, L-asparagine and L-glutamic acid)

and CH<sub>3</sub> symmetric deformation modes from methyl group terminated amino acids including L-leucine, L-valine, and L-methionine among others. The large peak at 1450 cm<sup>-1</sup> can also be attributed to the CH<sub>2</sub> scissors mode from any number of aliphatic side chain amino acids. The magnitude of the 1350 and 1450 cm<sup>-1</sup> bands supports the above assignments when one considers the numerous potential contributors in the samples. The large broad band at 1730 cm<sup>-1</sup> (C=O stretching) may have contributions from any or all of the amino acids present especially when considering the low pH (0.5 to ~1.5) of these samples. In this pH range the protonated form of the carboxylate group in free amino acids is prevalent. All of the amino acids discussed are present in appreciable quantities with the CD-A1 samples.

The spectra of the CD-A2 samples have a strong 890 cm<sup>-1</sup> band (Figure 2b, main manuscript) which is very difficult to assign due to the complexity of the samples, but may arise from C–C stretching in the acetate group of some EDTA complex (Krishnan and Plane 2002) as both EDTA and a number of metal salts are present in these samples. CD-S1 spectra have a single band at 1050 cm<sup>-1</sup> (Figure 2c) which can be ascribed to NO<sub>3</sub><sup>-</sup> symmetric stretching as nitrate salts are present in these samples. As shown in Figure 2d (main manuscript), a series of major bands attributed to the carbohydrate sugars in CD-S2 samples arise in the C–O stretching region (950~1200 cm<sup>-1</sup>), C–C stretching region (1300~1500 cm<sup>-1</sup>) and the COH, CCH, and OCH side-group deformation region (700~950 cm<sup>-1</sup>). eRDF spectra have a number of weak Raman peaks at 850, 980, 1003, 1350, 1415 and 1450 cm<sup>-1</sup> (Figure 2e). The peak at 850 cm<sup>-1</sup> is difficult to directly assign given the numerous potential origins but might be a result of C–C stretching in aliphatic side chain amino acids or C–C–O stretching from alcoholic side chain amino acids among others. The peak at 980 cm<sup>-1</sup> might well be due to L-arginine which has a prominent stretching band in this region (De Gelder et al. 2007) or alternatively SO<sub>4</sub><sup>2-</sup> symmetric stretching (Mosier-Boss and Lieberman 2000) from sulphate ions. The 1003 cm<sup>-1</sup> peak is characteristic of symmetric ring C–C stretching in L-phenylalanine while the peaks in the 1350–1450 cm<sup>-1</sup> range are most likely a result of CH<sub>2</sub> scissors and CH<sub>3</sub> deformation modes from various amino acid contributors.

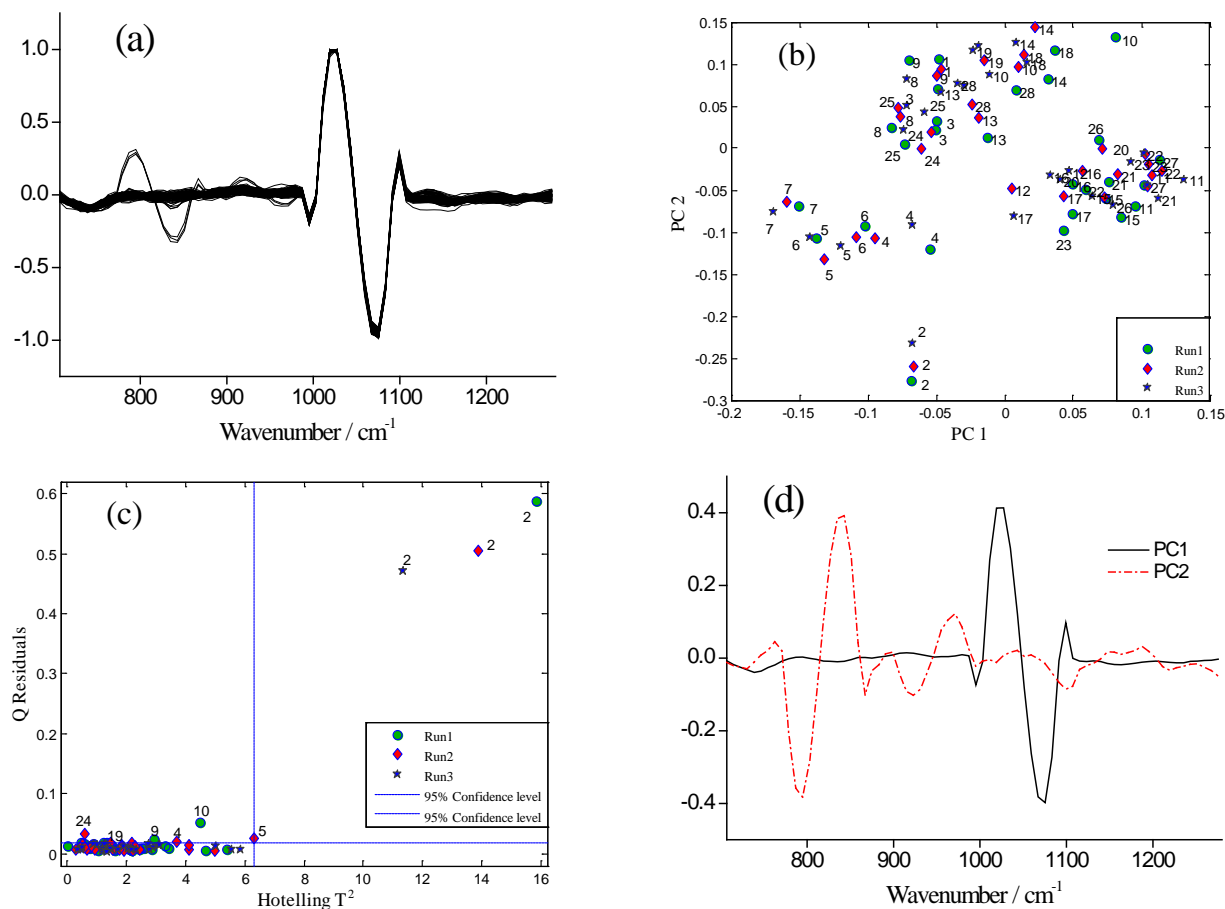
### S3.2 PCA models for each media class:

The PCA models used to select the spectra used for generating the SIMCA models (Table 4) are shown below for the various media types. In each case the Q and T<sup>2</sup> statistics plot is used to determine which samples are outliers and which individual spectra should be excluded. In general, samples were excluded from the SIMCA model if two or more of the replicate spectra were outside the boundaries. Individual spectra were excluded (but the sample and remaining spectra were used in the SIMCA model) if they fell outside the Q and T<sup>2</sup> boundaries.

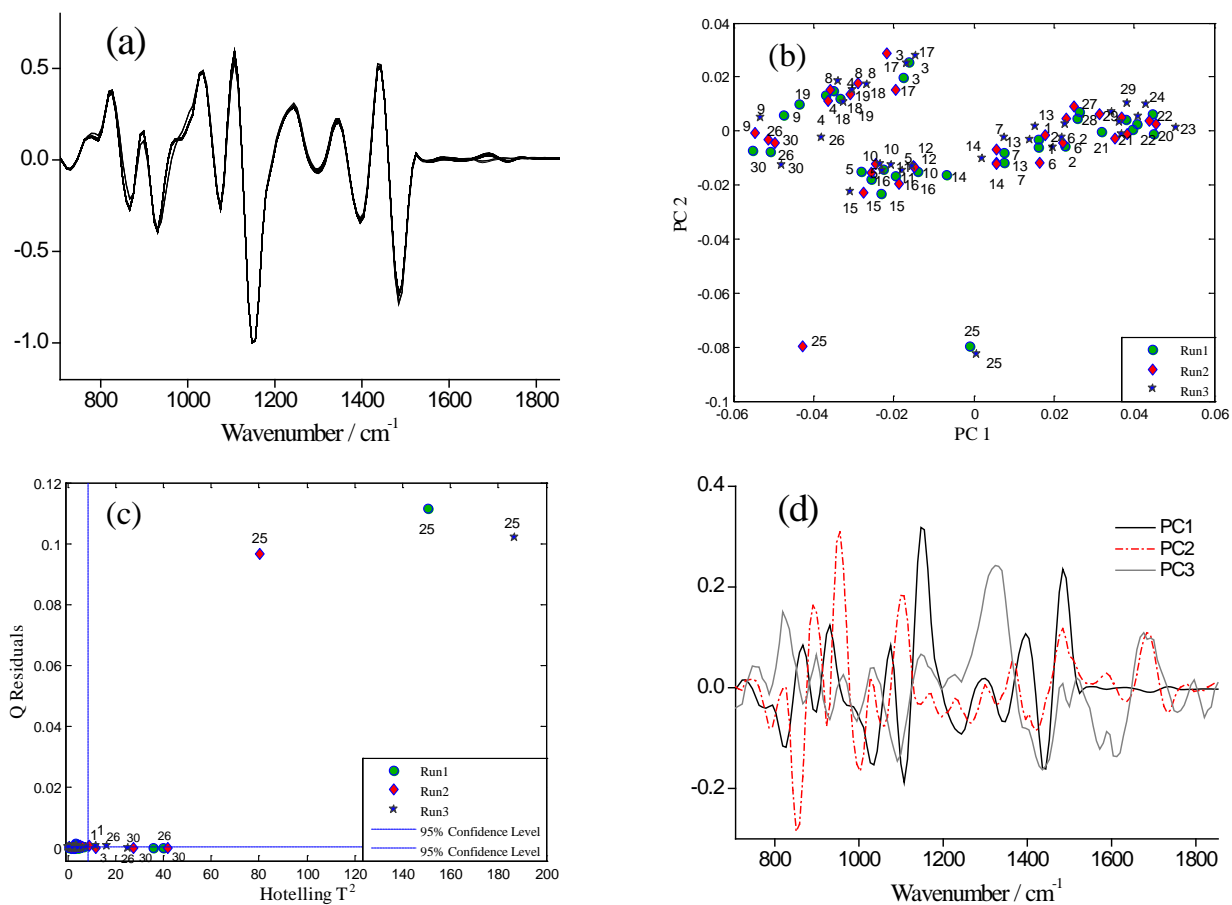


**Figure S-2:** Overview of PCA model on 93 CD-A2 measurements from triplicate data collection runs: (a) the pre-processed data used for the model (707~1276 cm<sup>-1</sup> region), (b) PC1 vs PC3 scores plot, showing the significant outlier samples, (c) Plot of Q and T<sup>2</sup> statistics (dashed boundary defines a 95% confidence limit), again showing the outliers relative to the main group of samples, and (d) Overlay of PC loadings plots for the three major components.

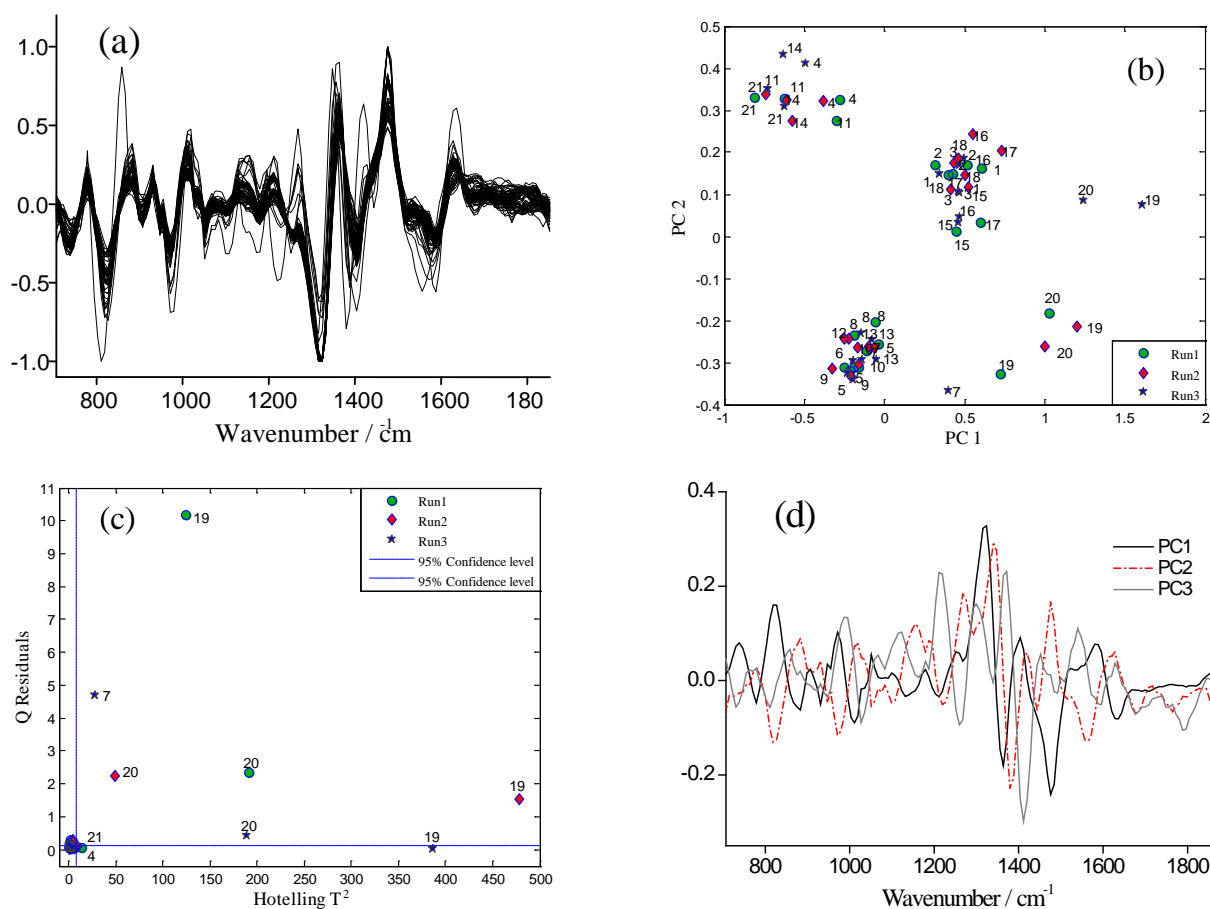
For the CD-A2 samples, we see from Figure S-2c that five samples (i.e., samples #2, #3, #27, #29, and #31) are repeatedly outlying (2 from 3 or 3 from 3 spectra are outside the boundaries). Thus we adjudge that the samples are outliers because of real compositional changes.



**Figure S-3:** Overview of PCA model on 84 CD-S1 measurements from triplicate data collection runs: (a) the pre-processed data used for the model (707~1276  $\text{cm}^{-1}$  region), (b) PC1 vs PC2 scores plot, showing the significant outlier samples, (c) Plot of Q and  $T^2$  statistics (dashed boundary defines a 95% confidence limit), again showing the outliers relative to the main group of samples, and (d) Overlay of PC loadings plots for the two major components.



**Figure S-4:** Overview of PCA model on 90 CD-S2 measurements from triplicate data collection runs: (a) the pre-processed data used for the model (707~1853 cm<sup>-1</sup> region), (b) PC1 vs PC2 scores plot, showing the significant outlier samples, (c) Plot of Q and T<sup>2</sup> statistics (dashed boundary defines a 95% confidence limit), again showing the outliers relative to the main group of samples, and (d) Overlay of PC loadings plots for the three major components.



**Figure S-5:** Overview of PCA model on 63 eRDF measurements from triplicate data collection runs: (a) the pre-processed data used for the model (707~1853  $\text{cm}^{-1}$  region), (b) PC1 vs PC2 scores plot, showing the significant outlier samples, (c) Plot of Q and  $T^2$  statistics (dashed boundary defines a 95% confidence limit), again showing the outliers relative to the main group of samples, and (d) Overlay of PC loadings plots for the three major components.

### References (supplemental):

- Castro JL, Montañez MA, Otero JC, Marcos JI. 1995. SERS and vibrational spectra of aspartic acid. *J Mol Struct* 349:113.
- De Gelder J, De Gussem K, Vandenabeele P, Moens L. 2007. Reference database of Raman spectra of biological molecules. *J Raman Spectrosc* 38(9):1133-1147.
- Jackson JE, Mudholkar GS. 1979. Control Procedures For Residuals Associated With Principal Component Analysis. *Technometrics* 21:341-349.
- Krishnan K, Plane RA. 2002. Raman spectra of ethylenediaminetetraacetic acid and its metal complexes. *J Am Chem Soc* 90:3195.
- Mosier-Boss PA, Lieberman SH. 2000. Detection of Nitrate and Sulfate Anions by Normal Raman Spectroscopy and SERS of Cationic-Coated, Silver Substrates. *Appl Spectrosc* 54:1126-1135.
- Takeda M, Iavazzo RES, Garfinkel D, Scheinberg IH, Edsall JT. 1958. Raman Spectra of Amino Acids and Related Compounds. IX. Ionization and Deuterium Substitution in Glycine, Alanine and beta-Alanine. *J Am Chem Soc* 80:3813-3818.