



Spectral pre and post processing for Infrared and Raman spectroscopy of biological tissues and cells

Journal:	<i>Chemical Society Reviews</i>
Manuscript ID	CS-REV-05-2015-000440.R1
Article Type:	Review Article
Date Submitted by the Author:	03-Sep-2015
Complete List of Authors:	Byrne, Hugh; Focas Institute, Dublin Institute of Technology Knief, Peter; Royal College of Surgeons in Ireland, Department of Medical Physics and Physiology Keating, Mark; Dublin Institute of Technology, FOCAS Research Institute Bonnier, Franck; Université François-Rabelais de Tours, Nanomédicaments et Nanosondes



Journal Name

ARTICLE

Spectral pre and post processing for Infrared and Raman spectroscopy of biological tissues and cells

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Hugh J. Byrne,^{a†} Peter Knief^b, Mark E. Keating^{a,c} and Franck Bonnier^d

Vibrational Spectroscopy, both infrared absorption and Raman spectroscopy, have attracted increasing attention for biomedical applications, from *in vivo* and *ex vivo* disease diagnostics and screening, to *in vitro* screening of therapeutics. There remain, however, many challenges related to the accuracy of analysis of physically and chemically inhomogeneous samples, across heterogeneous sample sets. Data preprocessing is required to deal with variations in instrumental responses and intrinsic spectral backgrounds and distortions in order to extract reliable spectral data. Data postprocessing is required to extract the most reliable information from the sample sets, based on often very subtle changes in spectra associated with the targeted pathology or biochemical process. This review presents the current understanding of the factors influencing the quality of spectra recorded and the pre-processing steps commonly employed to improve on spectral quality. It further explores some of the most common techniques which have emerged for classification and analysis of the spectral data for biomedical applications. The importance of sample presentation and measurement conditions to yield the highest quality spectra in the first place is emphasised, as is the potential of model simulated datasets to validate both pre- and post- processing protocols.

Introduction

The potential of vibrational spectroscopy, both Infrared (IR) absorption and Raman scattering, for biomedical applications has been well established through many proof of concept studies over the past decades. Due to its unique chemical fingerprinting capability at the molecular level, vibrational spectroscopy can play a significant role in a new paradigm of histopathology, cytology, biopsy targeting, surgical targets, treatment monitoring and drug studies. However, translation into the clinical environment has been slow, and although the challenges facing the translation to realistic clinical applications are manifold, including those associated with large scale clinical trials, health economics and acceptance by the medical community [1], there remains a considerable amount of issues relating to the fundamental process of recording reliable spectra from complex, chemically and physically inhomogeneous samples and extracting reliable information from heterogeneous sample sets which may be influenced by a multitude of confounding factors.

The development of reliable data (pre- and post-) processing and data mining techniques has thus been identified as a rate determining step in the maturation of vibrational spectroscopic techniques towards real applications, *in vitro*, *ex vivo* and *in vivo*. Instrumental response functions and sample presentation can significantly impact on the quality of the data gathered. Both tissues and cells are physically and chemically inhomogeneous and can give rise to a number of artifacts which can distort and greatly reduce

the accuracy of a spectral measurement. IR absorption and Raman spectroscopy are very distinct physical processes [2], and therefore the spectral distortions can be of very different physical origin. Although many biomedical applications of the techniques have been targeted towards disease diagnostics, for which high specificity and sensitivity classification algorithms are desirable, more recent applications have been targeted at, for example, disease aetiology [3,4], radiation dosimetry [5], drug screening [6-10] and nanotoxicology [11-13], for which a range of other data mining and analysis protocols have been explored [14]. It is important, however, that these protocols are well validated, in order to progress the field with confidence.

This Special Issue Review attempts to summarise the current understanding of the underlying physical factors influencing the quality of spectra recorded, for both IR absorption and Raman spectroscopy, and the pre-processing steps commonly employed to improve on spectral quality, while emphasising the importance of sample presentation and measurement conditions to yield the highest quality spectra in the first instance. The review focusses on established Raman and IR techniques, rather than emerging techniques based on enhanced or stimulated effects [15]. It explores some of the most common techniques which have emerged for classification and analysis of the spectral data for biomedical applications. The potential of model simulated datasets to validate both pre- and post- processing protocols is highlighted.

Spectral Preprocessing: InfraRed absorption spectroscopy

In IR spectroscopy, commonly performed in transmission or transmission/reflection (transflection) mode, the absorption features are typically found to sit on top of a large, broad and undulating background. On the basis that the background is slowly varying with respect to the absorption features of interest, first and even second order differentiation of the spectra has been

^a FOCAS Research Institute, Dublin Institute of Technology, Kevin Street, Dublin 8, Ireland

^b Department of Medical Physics and Physiology, Royal College of Surgeons in Ireland, 123 St. Stephen's Green, Dublin 2, Ireland

^c School of Physics, Dublin Institute of Technology, Kevin Street, Dublin 8, Ireland.

^d Université François-Rabelais de Tours, Faculty of Pharmacy, EA 6295

Nanomédicaments et Nanosondes, 31 avenue Monge, 37200 Tours, France

† Corresponding author: Hugh.Byrne@dit.ie

employed to remove the background, accentuating the desired spectral features. Such an approach assumes that the features themselves are not impacted by the underlying physical effects, however, an assumption which cannot be assumed to be valid based on a better understanding of their origin.

Mie scattering of the incident radiation, of wavelength $\sim 2\text{-}10\mu\text{m}$, by cellular components of similar length-scale, has been identified as the source of the broad undulating background commonly observed [16]. The undulating background of FTIR absorption spectra can be satisfactorily removed using the so called Extended Multiplicative Scatter Correction (EMSC) protocol [17], derived from the multiplicative signal correction (MSC), which is an algorithm for removal of additive and multiplicative contributions from a given interfering signal. Offsets and baseline slopes are removed effectively, whilst the multiplicative part of the algorithm compensates for differences in optical path length, effectively normalising the spectra. This is done by taking a reference spectrum, which can be the mean spectrum of the sample data set or a spectrum with similar spectral features. The algorithm takes the reference spectrum and attempts to recreate the raw spectrum to be corrected by adding an offset, a slope and amplifying the reference by multiplication. However, it is apparent that the absorption features themselves are also distorted, an effect which is most obviously manifest by a dip in the absorption profile on the high wavenumber side of the amide I feature at $\sim 1675\text{cm}^{-1}$, the so-called "Dispersion Artifact" [16] (see for example Figure 1), and can cause a shift in the apparent maximum of all spectral features. Until recently, the physical origin of this has been unclear [16]. While Mie scattering is commonly considered to derive from the imaginary component of the refractive index of the scattering material, using simulated data and chemically homogeneous model systems, the artifact has been recently demonstrated to have origin in contributions of the real component of the refractive index in the form of resonant reflection (in transfection mode) [18] and resonant Mie scattering (in transmission mode), although it should be noted that both resonant reflection and scattering artefacts can contribute in both measurement geometries [19]. The real and imaginary components of the refractive index are related via the Kramer-Kronig relationship and thus the absorption and (resonant) scattering profiles are intimately related [18]. Understanding the physical origin of the effects has led to a refinement of the EMSC protocol (RMieS-EMSC), by incorporation of resonant contributions to the sample refractive index, to efficiently account for and correct spectra for the effects [20, 21] (figure 1). A full description of the EMSC and RMieS-EMSC protocols is provided in [21]. An alternative, more supervised approach to spectral correction using undistorted spectra as standards has been proposed by Bird et al. [22].

Notably, in the development of the correction algorithm, a better fundamental understanding of the physical origin of the effects was critical, and the use of model systems and simulated data sets was invaluable in validating the protocol. Independent of correction algorithm employed, the increased understanding of the physical origins of the spectral distortions implies that IR spectra as a representation of chemical variations within biological materials can now be interpreted with a significantly higher degree of confidence.

Despite the greater understanding of the physical origin of the so-called dispersion artefacts, there remains much debate on the questions of measurement geometry for IR absorption spectroscopy, and therefore optimum choice of substrates. In terms of cost, low-E, reflective slides appear most attractive, implying the use of a transfection measurement configuration. However,

questions have been raised concerning additional spectral artefacts which can result from the so called "Electric Field Standing Wave" effect in such measurements [23]. On reflection from any metallic surface, the electric field of the radiation is pinned with a node at the reflecting surface, and a standing wave above it. For varying frequencies, the electric field distribution is different, having nodes and anti-nodes at different heights above the surface. A sample of defined thickness on such a surface will experience different intensity distributions across the spectrum, and therefore the absorbance spectrum can be distorted. It has been argued, however, that the effects are diminished by thickness inhomogeneities, the range of sampling angles, and the source incoherence [24]. The alternative, transmission, geometry requires (at least partially) transparent substrates. It has been demonstrated that even glass substrates may provide transmission in a sufficiently broad (high wavenumber) region to provide diagnostic capabilities [25]. However, access to the broader spectrum is only provided by more costly polycrystalline substrates such as CaF_2 . In terms of translation to a clinical environment, choice of substrate may ultimately be dictated by cost, and therefore by sample throughput and by the target application. A full cost analysis is required to assess the relative demand and costs of applications for (i) near patient intra operative diagnostic (ii) postoperative histological and (iii) research purposes. Establishing the relative impact of the spectral distortions associated with the EFSW effect compared to the target intrinsic biochemical variations may be critical in reaching a decision over clinical applications of IR absorption for spectral histopathology.

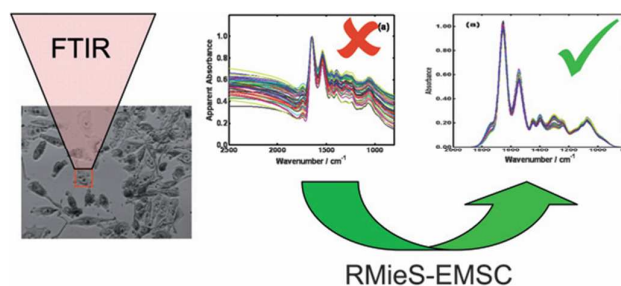


Figure 1: Correction of FTIR spectra for non and resonant Mie scattering (Reproduced from royal Society of Chemistry [20]).

In terms of sample presentation, fresh frozen sections are recommended as the tissue architecture and biochemistry is kept largely intact and, notably, the lipidic information can be accessed [26]. Furthermore, they are more amenable to combining immunohistochemistry, proteomics, and biospectroscopy. However, clinically, fresh tissue is normally only used for intraoperative work and stained, fixed sections are preferred for histopathology [27]. Therefore, standardised protocols for spectroscopic analysis of Formalin Fixed Paraffin Processed (FFPP) tissue samples are of paramount importance. Notably, analyses of archived tissue libraries may add much to understanding disease progression and patient prognosis.

It has been demonstrated that it is not necessary to remove the paraffin to obtain usable spectral information, particularly in the case of FTIR spectroscopy [28]. Standard tissue microarray protocols involve paraffin embedded tissue. Leaving the paraffin in place reduces scattering artefacts and effects of further variable removal of aromatic solvent soluble components [28-31]. The spectrally well defined paraffin contribution can be digitally

removed in a preprocessing step, involving Independent Component Analysis (ICA) and Non-Negatively Constrained Least Squares Analysis (NCLS), so called digital dewaxing [28]. However, it may be argued that greater consistency of spectral information is achieved when sections are deparaffinised. Deparaffinising also allows post-staining of the sections, although it has been demonstrated that the efficiency of the deparaffinisation process can depend on the tissue pathology [32]. Nevertheless, it can also be argued that, even for research purposes, protocols for such tissue processing should be maintained as close as possible to those currently employed in the clinical environment.

Spectral Preprocessing: Raman spectroscopy

In the case of Raman spectroscopy, as-recorded spectra can suffer from similarly additive and multiplicative contaminations as FTIR spectra. Background, substrate and instrumental contributions are routinely subtracted, whereas dealing with multiplicative contributions from resonant or otherwise enhanced scatterers can be more complex.

A similarly broad underlying spectral background is often observed in Raman spectra of biological samples. In the case where the sample contains fluorophores which are resonant at the source wavelength, the background contributions are often eradicated by spectral differentiation, as in the case of IR absorption spectra, although the relatively higher wealth of spectral information in Raman spectra means that the derivative spectra are difficult to interpret [33].

In many cases, however, the background to the Raman spectra can not easily be attributed to sample fluorescence, as it is unclear what fluorophores in tissue, or even raw proteins, can give rise to a broad fluorescence background, especially with near infrared sources [2]. More recently, it has been demonstrated that the background in pure proteins at 785nm is dependent on sample morphology and that measurement of tissue sections in water immersion effectively eliminates the background, suggesting that an origin in stray light from Mie scattering of the source laser line and even the Raman bands themselves [34]. The study demonstrates that an important route towards minimizing preprocessing is appropriate sample presentation. When measured in immersion, the water acts as an index matching fluid, visually reducing the scattering of the laser spot by the sample (Figure 2) and significantly enhancing the spectra (Figure 3), which can be preprocessed by subtraction of the instrument response and the water contribution. Measurement in water immersion also significantly reduces any photo damage in tissue samples, the water providing a heat sink to protect from photothermal effects [35].

In addition to background from the sample, in thin samples, the substrate itself can contribute significantly, and the contribution can be variable from sample to sample and difficult to remove routinely. Choice of substrate can minimize this contribution, and UV grade CaF₂ is widely accepted as an optimal choice [36], particularly as it is also compatible with IR transmission measurement. In terms of potential clinical applications, however, for example routine cytological screening, cost implications may dictate that substrate choice is limited to conventional glass microscope slides. In the development of Raman protocols for such applications, choice of substrate is intimately linked with choice of wavelength, and it has been demonstrated that, although conventional glass microscope slides have a substantial background at 785nm, this is greatly reduced at 532nm [2, 36].

Confocal operation, available in some Raman microspectrometers, can reduce the spectral contribution of the substrate, as previously described by Puppels et al. for single cells [37, 38]. Nevertheless, when performing measurements on optically thin single cells grown on a substrate, substrate contributions to the Raman spectra can still be significant. Furthermore, confocal operation is not available on many commercial spectrometers. Recent studies have demonstrated the benefits of using 3D collagen gels for Raman mapping of single live cells [39]. The substrate (collagen) contribution to the spectrum is shown to be negligible, reducing the requirement for substrate subtraction. Moreover the protein matrix reduces the toxicity of the substrate enabling prolonged studies of live cells even in full cell culture medium. Such extended mapping periods are ideally required for full cellular and sub cellular analysis to evaluate, for example, mechanisms of interaction of nanoparticles or chemotherapeutic agents.

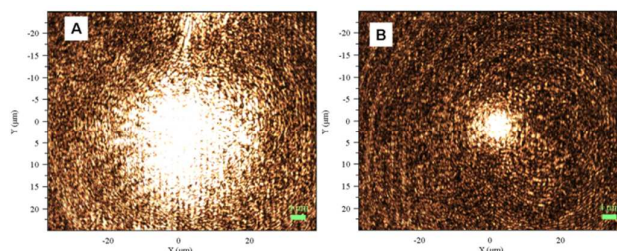


Figure 2: Laser spots obtained at the focus position on a skin sample after chemical dewaxing in the dry state (A) and immersion (B). (Reproduced from John Wiley and Sons [34])

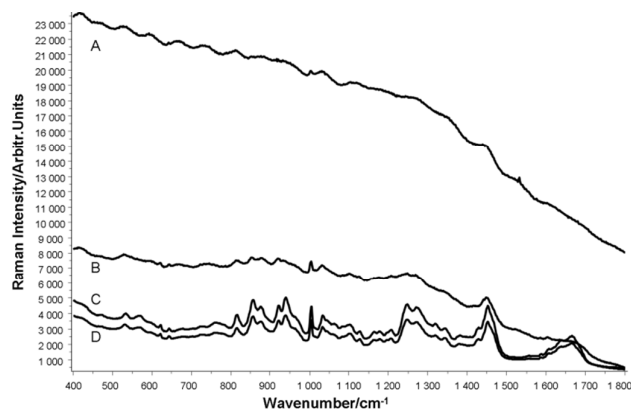


Figure 3: Example of mean spectra recorded on two different human skin samples from the thigh dry (A and B) and using the immersion objective (respectively C and D). The first sample has been dewaxed in hexane (A and D), the second using xylene (B and C). (Reproduced from John Wiley and Sons [34])

In cases where the spectral background cannot be completely eradicated experimentally, background removal is commonly performed, by one of numerous algorithms [40, 41] e.g. one automated approach is the implementation of the so-called EMSC (extended multiplicative scatter correction) algorithm, originally developed for IR spectroscopy [42]. Common semi-automated computational methods of background subtraction are the subtraction of a polynomial of certain order [43-45]. The so-called “rubber band” model is derived from the commercially available “Opus” software package commonly used on Bruker IR spectrometers [46]. However, it should be noted that such

subtraction of arbitrary backgrounds does not guarantee a reproduction of the true spectral features, as a spectral minimum between many overlapping Gaussian features may not be the true minimum of the signal. Prior knowledge of the contributing factors should be used, wherever possible. Accordingly, it is common to record and subtract “dark” and substrate spectra from the sample spectra before removing any arbitrary background. As the substrate spectrum is an independent spectral component, it can effectively be removed by ICA and NCLS, in a similar process to the so-called “digital dewaxing” of tissue spectra [28].

Further artifacts in Raman Spectroscopy include nonlinearity of axes and dark noise. In dispersive multichannel Raman spectrometers employing a CCD sensor, the data point spacing during acquisition can be irregular not only due to the use of different gratings. It may change from day to day due to different calibration settings and even drift over the course of a day due to variations in temperature and/or humidity. This leads to recordings with different abscissa and abscissa-linearity which therefore induces variance in the x-axes. This issue of “transfer calibration” has been recognised in a number of fields, notably in near IR spectroscopy in food science and a number of protocols have been developed [47-49]. A number of commercial instruments now incorporate a post recording calibration procedure, although it is not clear as yet whether a standard procedure has been adopted by the biospectroscopy community.

As Raman spectroscopy measures scattered intensity, as opposed to IR spectroscopy, which measures a transmission ratio, in addition to calibration of the spectral axis, intensity calibration is necessary in cases that the results from different instruments and laser sources are to be compared. To this end, the use of Standard Reference Materials (SRM) are desirable, e.g. those from the US National Institute of Standards, Boulder, Colorado, USA (NIST SRM 2243, 2242, 2241). The use of SRM also provides a means to correct Raman spectra for relative intensity on a day-to-day basis. The application of such standards requires measurements of its luminescence spectrum on the Raman instrument employed. Subsequent mathematical treatment of both the observed luminescence spectrum of the intensity standard and the observed Raman spectrum of the measured sample create the intended comparability of data between spectrometer and excitation lines. The relative intensities of measured Raman spectra are corrected for instrument specific response employing computational methods using a correction curve. These curves are generated with certified polynomials and pre-recorded fluorescence spectra of the SRM glass (a manganese doped borate matrix glass).

Electronic noise, consisting of flicker noise, shot noise and thermal noise is an unpredictable and constant occurrence primarily in the spectral intensity, and can have a huge impact on the quality of any signal [50] and Raman spectra are commonly additionally subjected to a noise reduction protocol to “smooth” the spectra, increasing the signal to noise and accentuating the true spectral features. A common smoothing algorithm incorporated in commercial instrument software is the Savitzky-Golay algorithm [51]. However, it is important to be aware of the potential influence on spectral features of order and window size of Savitzky-Golay filters for noise reduction. Recently a systematic study of data preprocessing techniques which considerably increase the confidence in the reliability of pre-processed Raman spectra has been carried out [46].

In both IR absorption and Raman spectroscopy, the measured spectra are also influenced by spot to spot variations of sample thickness and density, resulting in a variable overall signal intensity

reflective of the physical inhomogeneity rather than biochemical changes related to pathology or biochemical process [521]. For multivariate analysis, normalisation of variables into a relative variable space is recommended. Ideally, an internal standard or other pseudo-constant reference value to correct for the scaling effects should be employed [53]. In the absence of an independent internal standard, vector normalization is commonly employed. The process weights each spectrum according to its integrated intensity such that all samples contribute equally to the analysis model [54]. It is important that all background and/or baseline contributions are removed before normalisation.

Data Analysis

Analysis of the subtle changes in spectral profiles associated with, for example, disease, biological processes or the influence of external agents require the use of multivariate analytical techniques. Such multivariate methods have become invaluable to a wide range of fields, including geology, pharmaceutical science, pharmacology, astrophysics, imaging, and chemistry. Importantly, these methods allow for complicated and also in some instances very large datasets to be analysed and in effect they reduce the dimensionality and complexity of the data allowing for meaningful information to be extracted.

Specifically considering vibrational spectroscopic datasets, multivariate methods allow analysis of multiple spectra simultaneously and interdependently. This then allows for comparisons to be made between spectra and groups of spectra within a dataset and to identify trends these may contain e.g. spectral markers of disease in control and non-control patients, identification of nanoparticle containing spectra, response to external agents etc.. Although they derive from distinct physical processes, the approach to data processing is largely independent of the analytical technique (IR absorption or Raman spectroscopy) but is more dictated by the application (e.g. diagnostic classification, spectral mapping, progression analysis).

An important consideration in the application of multivariate methods to spectral data analysis is the requisite size of dataset. Statistical significance is a critical consideration, and when patient diagnosis is the outcome, misclassification has serious consequences. Beleites et al., have carried out a study examining the effects of sample size on multivariate classifier models for clinical biospectroscopy [55]. It is demonstrated that, while learning curves for dataset sizes common to small scale academic studies (2-25) can indicate acceptable performance, the model testing is itself limited by the dataset size and that datasets of 75-100 samples are required to produce “a good but not perfect classifier”.

In the construction of multivariate classifiers, robustness of the model should also be rigorously tested. A commonly used performance measure is the Standard Error of Prediction (SEP) [56]. If the data are normally distributed, the prediction errors are generally within the tolerance interval of $\pm 2 \times \text{SEP}$. However, evaluation of the performance of a model on the complete dataset acquired during a given experiment does not assess the potential performance of the model in a practical situation where it is presented with unseen data [56]. When the dataset is sufficiently large, individual spectra may be considered normally distributed and representative of the multivariate dataset that is likely to be encountered in reality [56]. In such situations, the dataset may be randomly sorted and split into calibration and test sets [56-59]. The calibration set is used to find the optimal model for that dataset, and the root-mean-squared error of calibration, (RMSEC), is

computed. The generalization performance of the model is determined by the root-mean-squared error of prediction (RMSEP) for the test dataset. A significant difference between RMSEC and RMSEP indicates that the model does not generalize well to unseen data. Often the evaluation of the models performance is conducted many times on different randomized splits of the calibration and test datasets to give a comprehensive assessment of the generalizability of the models [56]. Other model validation techniques commonly employed include Leave one out cross validation or up to 10-fold cross validation [60, 61]. Pérez-Guaita et al. [62] more recently evaluated the use of permutation testing, commonly used in metabolomics [63] and proteomics [64], which employs a random reallocation of class labels in order to establish the statistical significance of a cross-validation figure of merit of a classifier. Ultimately, however, the validation of the integrated techniques of spectroscopy and multivariate classifiers will have to comply with the rigours of the clinical environment, including large scale blind datasets and randomised trials [1].

In terms of post processing of multivariate spectral data, for diagnostics the primary emphasis to date has been on unsupervised classification. In this context, PCA has been used extensively. The use of PCA has been extended to include for example analysis of antibody activation of T-cells [65] and differentiation of embryonic stem cells [66], and also in the differentiation of spectra from different spatial locations within cells and tissues in spectral imaging [67,68]. It has also been used in conjunction with other multivariate methods such as Linear Discriminant Analysis to derive clinically relevant estimates of the sensitivities and specificities of the diagnostic protocols [69] and in comparative analyses of the spectral content in various types of human cancer cell lines [70]. Hierarchical cluster analysis is similarly often employed for example in the intercomparison of spectral classes in discrimination of malignant and non-malignant tissues [71], and for visualising the spatial distribution of identified components in spectral maps [72, 73]. Notably, however, for diagnostic applications, precise information on the chemical determinants of the differentiation of tissue types or sample regions is not required.

To extract information regarding the biochemical changes underpinning the spectral changes, a more sophisticated data processing toolbox is required. Supervised methods such as (Linear and Nonlinear) Partial Least Squares Regression (PLSR) can be employed to identify spectral variables which are specifically correlated with an external agent [5, 6] or indeed an observed physiological effect such as viability or proliferative capacity [7]. In such studies, correlation with accepted or “gold standard” assays can be used to guide and validate the interpretation of the vibrational spectroscopic results. As well as standard cytotoxicological assays, the spectroscopic results can be correlated with more precise biomarkers, such as protein over-expression in the case of HPV transfected cells [3, 4]. In advancing the applications of biospectroscopy, it is crucial to establish multivariate spectral equivalents of biomarkers which identify the action of for example chemotherapeutic agents or nanotoxicants. Regression coefficients can play a critical role in the validation of the model and provide indicators of the key contributions to the systematic spectral variations. Feature selection models can often aid in reducing the number of variables presented to the model. In PLSR, PLS Jack-knifing has been demonstrated to allow the reduction of the number of variables [74] avoiding over fitting and improving performance. PLS Jack-knifing produces results that are readily interpretable in terms of highlighting the systematic variation of important spectral features within the regression model, and allows

visualization of PLSR coefficients and their uncertainty, and their use in analysing spectroscopic responses associated with chemotherapeutic agents has been demonstrated [6, 7].

Whereas PCA and PLSR, for example, may be considered linear analytical methods, in that they attempt to describe the variability of the dataset according to single parameters, multiparameter or nonlinear statistical approaches can prove more powerful in applications to biospectroscopical datasets which have a high degree of variability. In this context, machine learning algorithms such as Artificial Neural Networks have received considerable attention [75]. Genetic algorithms (GA) and Support Vector Machines are other options which have been successfully utilised in multivariate regression problems [76] although direct visualisation and therefore interpretation of the identified spectral variables is not as simple.

A brief description of a selection of these multivariate methods employed in spectral analysis is given in the following sections.

K-Means Cluster Analysis

K-means clustering analysis (KMCA) is a statistical method which aims to partition data into clusters based on similarity. Firstly the method chooses a number of seed locations which serve as initial centroid locations in the dataset. Once a data point is assigned to one of the seed locations, it changes to a centroid which serves as a mean value of that cluster. The assignment of data points to clusters is often based on the Euclidean distance between data point and centroid, although other methods of calculating the distance also exist [77]. After each spectrum has been assigned to a centroid, the distance is then recalculated between each point and centroid to see if any points are closer to another centroid location, whereupon, if the point is closer to another cluster centroid, then it is reassigned and both cluster centroids are changed as a result. This process is completed for all data points until there is no movement between clusters.

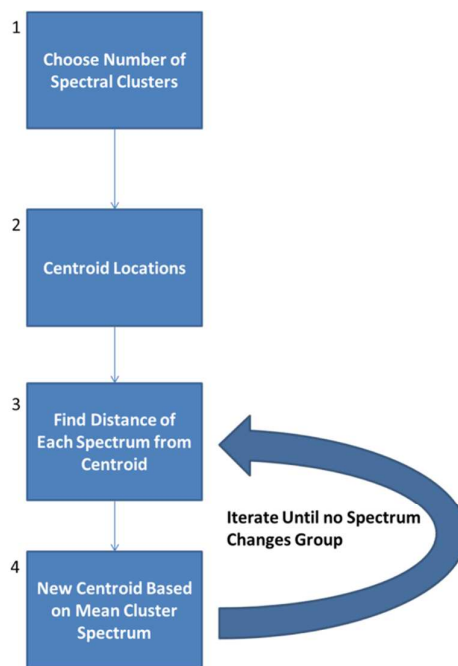
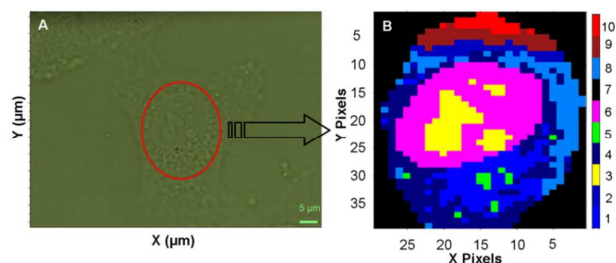


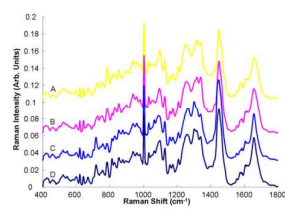
Figure 4. Schematic outlining the key steps in the K-Means Clustering Analysis algorithm.

From a spectroscopic imaging perspective, an initial number of seed locations is chosen. The spectra are then assigned to one of the seed locations. Once all spectra have been assigned, the mean spectrum or centroid is calculated and the distance between each spectrum and centroid is calculated. The spectra are then reassigned if necessary and the process is iterated until no spectra change groups. Figure 4 shows a diagram illustrating the main steps in the K-means clustering algorithm.

(I)



(II)



(III)

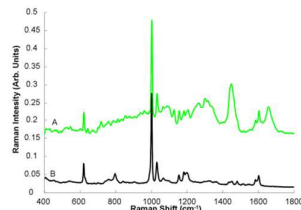


Figure 5: I; (A) Microscopic image of an A549 cell, showing the reduced area identified for spectral mapping. (B) K-means cluster map of the Raman profile of the same reduced area. II; K-means spectra of clusters 3 (A – representing nucleoli), 6 (B – representing nucleus), 1 and 4 (C and D, both from the cytoplasm). Spectra are offset for clarity. III; K-means spectrum of Cluster 5 (A), compared to the Raman spectrum polystyrene nanoparticles (B). Spectra are offset for clarity. (Reproduced from Royal Society of Chemistry [12])

In vibrational spectroscopy, KMCA has seen a number of uses to separate spectra into clusters based on spectral similarities. As an imaging tool, KMCA aims to separate each spectrum acquired in the image and assign it to a cluster. This assignment is termed 'hard' in that each spectrum is only assigned to one cluster. A good example of KMCA in Raman spectroscopy is shown in the study by Dorney et al. [12], in which it was used to identify regions in the Raman dataset which correspond to polystyrene nanoparticles, and differentiate them from neighbouring cytoplasm, as well as the nucleus and nucleolus (Figure 5). Many other examples of KMCA as a spectroscopic imaging reconstruction technique can be found in the investigation of a wide range of samples including tissue sections [78], cells [12, 79, 80] and in the analysis of human skin [26, 81].

While this method has been shown to be useful in partitioning spectra into clusters, it is important to highlight that the method is not without certain pitfalls. Firstly, as the initial choice of centroid location can be subjective, the reproducibility of the method can in some instances be called into question i.e. if the initial starting point of the analysis changes then it is possible to end up with

different results. It is important to ensure enough iterations (50–100) such that convergence is reached. Secondly, looking at the method to assess spectral imaging, each spectrum is assigned to only one cluster, and the cluster is represented by the average of all constituent spectra. As a spectrum may represent a number of different biological entities in differing quantities, KMCA may be correct in grouping a spectrum based on, for example, lipidic distribution, but it may misclassify a spectrum which also contains a small amount of another cluster's biochemistry. There is no weighting element introduced into the analysis so each spectrum must belong to only one cluster.

Hierarchical Cluster Analysis

Hierarchical clustering analysis is a multivariate method which is commonly used for clustering spectral data and generating images. There are two main forms of HCA, agglomerative and divisive. Agglomerative HCA is the more commonly used method. Briefly, this method starts out with each data point or spectrum in a separate group or cluster. The method then aims to group each data point together in an iterative process until there is only one cluster which contains all the data points. It is then possible to construct an image based on how these clusters are linked together. Often, the data can be represented using a two dimensional dendrogram which shows the linkage between each cluster. Divisive HCA on the other hand starts off with each spectrum in one cluster and then aims to separate each data point into one cluster.

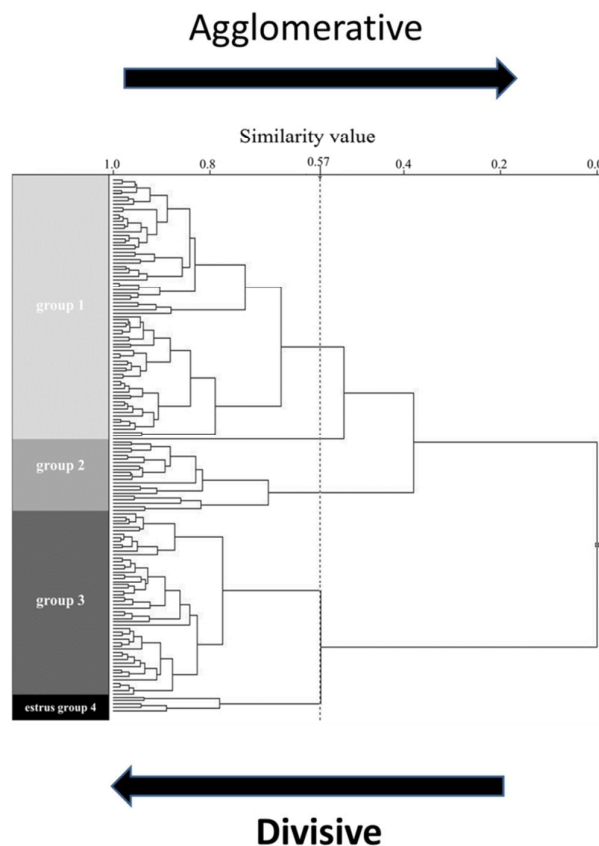


Figure 6: Showing a HCA dendrogram and both divisive and agglomerative clustering.

An example dendrogram is shown in Figure 6. An important point in relation to HCA is that, once a group of spectra has been assigned to a cluster or in the case of the agglomerative method merged into a cluster, the spectrum cannot be reassigned, unlike KMCA where the spectra can move clusters if closer to another centroid. This means that HCA results in a very definite grouping of spectra into clusters. In spectral datasets, it is assumed that all samples in a study (with the exception of outliers) belong to the overall source cluster, and individual spectra are clustered in an agglomerative approach. Assigning spectra to a cluster is based on the pair-wise similarity expressed as a matrix of correlation coefficients. The threshold of similarity is somewhat subjective, although algorithms are available to optimise performance [82, 83].

Although commonly performed in an unsupervised manner (UHCA), HCA can also be performed in a supervised or semi-supervised manner. UHCA is based on statistical similarity of the data but does not take any of the experimental variables such as treatment, phenotype, tissue, etc. into account while clustering. In contrast, supervised or semi-supervised clustering utilises such additional information to “guide” or “adjust” the clustering process [84].

HCA is like KMCA in that the method is deemed to be a hard clustering method with each spectrum being assigned to a specific group. From an image reconstruction perspective, this means that each pixel can only be assigned to one specific biochemical grouping, which may not be reflective of the actual dataset. HCA has been used in a classification method in number of studies which include cellular studies [85] as well as in the investigation of vibrational spectroscopy in diagnostics [86].

Vertex Component Analysis

Vertex component analysis (VCA) is another multivariate statistical method which is used in spectral analysis [87]. The algorithm makes an assumption that, contained within the dataset, are pure endmember spectra which in turn can be used to describe all the other spectra in the dataset. From this, abundance plots can be generated via a linear combination of endmember spectra and constructed into images which are described by the biochemical information contained in these endmember spectra.

Recently, VCA has seen a number of applications in hyperspectral imaging using both IR and Raman spectroscopy, with applications including Raman histopathological imaging and also cellular studies including nano-bio interactions [88]. Importantly, while this method can be used quite readily to reconstruct biochemical regions in the cell, like all methods it may be prone to error. Firstly, as highlighted by Chernenko et al. [88], endmember spectra may contain mixtures of different biochemical components and while this may be reflective of the actual nature of the sample, may lead to inaccuracies in interpretation. Additionally, the method makes a large assumption that the most extreme spectra in the dataset are the most reflective of pure component spectra, which may not be the case in complex biological spectra.

Fuzzy C – Means Clustering

Fuzzy C-means clustering (FCM) is a method which is similar to KMCA in that it also assigns spectra to centroids in the datasets. However, unlike KMCA, the method is a soft clustering method, whereby each point or spectrum in the dataset is assigned a value from 0 to 1, the value closest to 1 being representative of the cluster centre. Therefore, by analysing the C centroid spectrum it is possible to extract chemical information which describes each reconstructed image. FCM has seen some usage in Raman spectroscopy although primarily as an imaging method [79, 80].

Principal Components Analysis

Principal Components Analysis (PCA) is a method which aims to reduce the dimensionality of the data to describe the variation present in a dataset, whereby the first principal component is a description of the maximum variance present in the dataset, the second describes the second most variance, etc. The principal component scores can then be described by the loading vector which is an explanation of this variance. In a spectroscopic context, the scores represent values which correspond to a loading spectrum which contains peaks, both positive and negative which explains the spectral variation in the dataset.

This tool can be quite useful for providing a method to separate spectra into groups e.g. diseased and non-diseased [69]. It has also been used to reconstruct images [79,80] i.e. a variance plot based on the loadings plot. However, as these loadings plots may often contain a number of spectral features corresponding to different cellular biochemistry, interpretation can be difficult and it is quite possible to misinterpret.

Bonnier et al. have shown that pairwise PCA of clusters identified by KMCA can provide a clearer picture of the specific biochemical differences between regions [89]. In simulated mixtures of multiple components, the loadings can be seen to be weighted linear combinations of the spectra of the contributing spectra, as shown in Figure 7. As an example of the benefits of this approach, the dose dependent spectral changes induced by simulated solar radiation in the basal layer of a reconstructed skin model could be tracked according to the loading of the first principle component of the variation of the irradiated skin section compared to control [90].

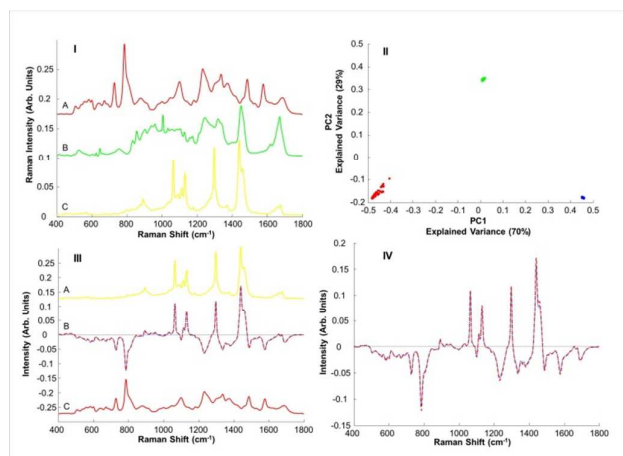


Figure 7: Mean Raman spectra recorded from RNA (A), histone (B) and ceramide (C) on CaF₂ windows. II: Scores plot of the 2 first principal components after PCA performed on Raman spectra recorded from RNA, histone and ceramide. III: Plot of the loadings of PC1 (blue dot line) compared with the difference spectrum calculated from the mean spectrum of ceramide minus the mean spectrum of RNA (red dash line). The loadings are compared with spectra recorded from ceramide (A) and RNA (C), both offset for clarity. IIIB: Plot of the loadings of PC2 (blue dot line) compared with the difference between the mean spectrum of histone minus the average mean spectrum of RNA and ceramide (red dash line). (Reproduced from Royal Society of Chemistry [89])

Classic Least Squares Analysis – supervised/Unsupervised

Classic Least Squares Analysis (CLSA) can be carried out in two different ways, either by generating spectral models using a factor analysis algorithm (unsupervised), or by manually inputting the component spectra (supervised).

The analysis method is based on a fit of a linear combination of reference component spectra to the spectra contained in the raw spectral map. There are two different ways to obtain the reference component spectra. The first way is to obtain a pure spectral reference from a compound or compounds which can then be fitted. The second method uses a factor analysis algorithm to generate the component spectra, the weighted sum of which is compared to the Raman spectral data set. Using the latter of the two methods, Zavaleta et al. demonstrated the power of the technique to quantify quantum dot accumulation in an *in-vivo* mouse model and to separate out the different spectral contributions from complex SERS signals in the same data set [91]. In a similar and different way, both approaches to CLSA were explored to extract spectra which contain polystyrene nanoparticles from a cellular Raman map and define other biochemical regions such as the RNA and lipid rich environments [13]. The example of the unsupervised CLSA analysis map is shown in Figure 8.

Notably, however, the unsupervised approach generates factors which are mixtures between different components. In the supervised case, the pure spectra are input into the analysis and a map of the relative contributions is similarly generated. However, for the same case of polystyrene, RNA and lipids in a subcellular environment, all individual components generated a background over the whole cell area and a thresholding of the data was required to produce a trustworthy image. Although such a thresholding procedure is somewhat arbitrary, the support of simulated datasets can add confidence to the procedure [13].

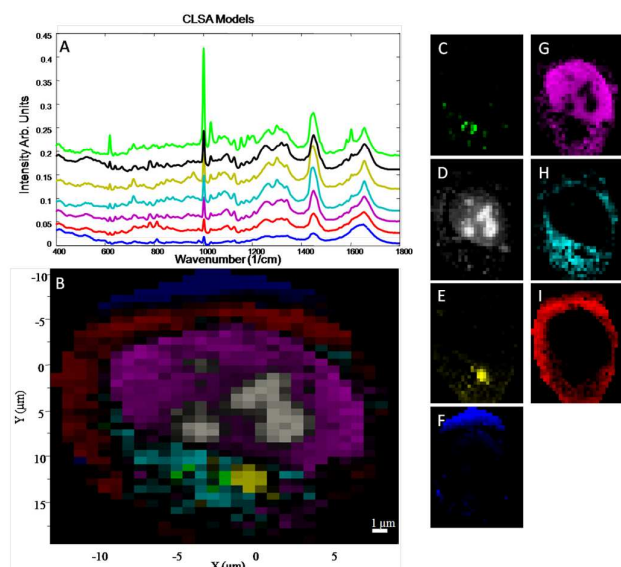


Figure 8: Clustering of spectra identified by unsupervised CLSA. (A) Spectral models generated from the analysis protocol and used to generate the clustered map shown in (B). The right panel (C-I) shows the distribution of each model created in the map. Of particular note, model 1(C), model 6(D) and model 7(H) have strong contributions of the spectra of polystyrene, RNA and lipid respectively. The spectra in (A) are colour coded and correspond to

images (B – F), with the exception of Model 6 which corresponds to the white image in (D). (Reproduced from Royal Society of Chemistry [13])

Spectral Cross Correlation Analysis

A novel analytical technique of Spectral Cross Correlation Analysis (SCCA) was also applied to the data set analysed in Figure 7 [13]. For SCCA, reference spectra from polystyrene, phosphatidylethanolamine and RNA (Figure 9) were used to screen the Raman spectral data set using a cross-correlation algorithm. The cross correlation function integrates the product of the two data series (spectra) at each point as they are shifted relative to each other along the x axis (wave number). The magnitude of the correlation quantifies the relative contribution of the component spectrum at that point in the cell, and an exact correlation occurs when the spectra are exactly matched (auto-correlation). In this way, it is possible to screen the map or spectra in the map and, based on the cross correlation function, cluster different biochemical regions of the cell based on the relative contributions of the reference spectrum used.

The results of such a SCCA of the subcellular localisation of polystyrene, RNA and 3-sn-phosphatidyl ethanolamine is shown in figure 9. Similar to the case of unsupervised CLSA, a considerable background of each component across the cell is observed initially, and a thresholding procedure was required to improve the precision of the technique. Simulated datasets generated for SCCA provided a good estimation of where this thresholding should take place and in combination with cellular data containing no nanoparticles it was possible to accurately reveal where the nanoparticles were located in the cell.

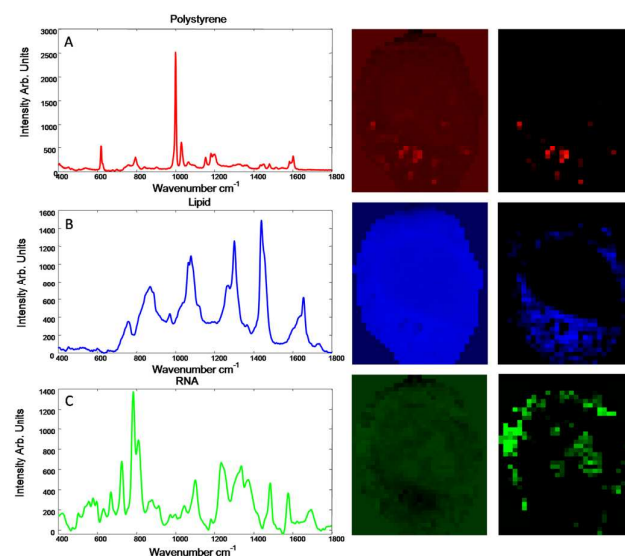


Figure 9: SCCA analysis using component spectra of polystyrene (A), 3-sn-phosphatidyl ethanolamine (B) and RNA (C). The spectrum of each pure component is shown on the left of the figure and the correlation maps for non-thresholded shown in the middle and thresholded on the right. (Reproduced from Royal Society of Chemistry [13])

Partial Least Squares Regression

Recently, regression modelling has seen a number of biomedical applications in both Raman and IR spectroscopies. Partial least

squares regression (PLSR) is an analytical technique which aims to match a test data set to a series of targets. The core idea of using this method is to investigate the spectral variability as a function of a systematic conditional change such as radiation dose [5] or viral infection [4]. PLSR can be employed to construct predictive models for spectral response as a function of the target variable. Therefore, an unknown dose or degree of infection can be determined from its spectrum, having obvious potential clinical applications. Furthermore, feature selection techniques such as PLSR coefficients, Jack-Knifing (JK) and genetic algorithms, amongst others [92], can be employed to identify the most statistically relevant spectral changes, such that the biological mechanisms underlying the spectral changes can be explored and understood.

A good practical example of this method in action in Raman spectral data is outlined in two studies by Nawaz et al., [6,7] in which, the aim was to investigate the capability of Raman as a technique to study drug interactions in cells and the physiological response. Looking specifically at cisplatin as an example chemotherapeutic drug, these studies were able to extract information relating to drug action in the cells via regression of the Raman dataset against cytotoxicological data and dose, extracting out features from the Raman spectra which correspond to changes to protein conformation and structural alterations of DNA [6]. A further study by Keating et al. demonstrated the validity of the technique, as well as potential pitfalls, resulting for example from cross contamination of targets for limited data ranges, using simulated datasets in which the systematically varied drug dose and cell viability dependent spectral variables were known [93].

Importantly, while these studies show the potential of Raman spectroscopy and PLSR as tools for studying drug interaction, PLSR used in this capacity is only relevant if the processes studied are in themselves linear. However, most pharmacological actions are non-linear processes and thus using a linear method to model a non-linear process may be subject to error. Thus, additional forms of validation of these methods in a spectral setting are necessary.

Artificial Neural Networks

Artificial neural networks (ANN's) are considered 'nonparametric nonlinear regression estimators' [75, 94] because of their ability to determine relationships between one or more input or 'independent' variables and one or more output or 'dependent' variables, regardless of the form of the function defining the relationship between the two sets of variables. Inspired by biological neural networks, their popularity stems from their general applicability to any problem [95], and they have seen applications in radiation science to dose-dependent models for flow cytometric analysis [96], for the prediction of depth dose in radiotherapy [97], and for neutron dosimetry [98]. Udelhoven et al. have demonstrated the use of artificial neural networks in conjunction with a hierarchical classification of FT-IR spectra for the identification of bacteria [99], while Lasch et al. have demonstrated the use of Artificial neural networks as supervised techniques for FT-IR microspectroscopic imaging [100]. ANN's have the potential to 'over-fit' noisy features within the input variables if the model is overly complex [75,94], and careful training and rigorous evaluation of the network is required to prevent this.

Genetic Algorithm

Calibration models are known to be greatly improved through the application of efficient feature selection methods, increasing the predictive ability and reducing model complexity. One such method is the adaptive search technique known as the genetic algorithm

(GA). A GA based variable selection procedure is used to reduce the original spectra to a subset of wavenumbers to correlate the spectra to response. The first generation for evaluation is a random population consisting of a number of individuals or "chromosomes", each containing a subset of the original variables. Each chromosome is composed of a vector of 1s and 0s, corresponding to the wavenumbers in the *X* matrix, (1 if selected and 0 if not) where each wavenumber is termed a "gene". The performance of models resulting from each chromosome is determined by means of a fitness function (e.g. the root mean square error of cross validation is used). Once each generation is evaluated, a new set of chromosomes is produced by retaining and "crossing" over the fittest individuals from the previous generation. "Mutations" are also produced which force the evaluation of new combinations avoiding saturation with similar sets of events and can further lower the number of variables and fitness values. The process continues until the difference in mean fitness level between successive generations is below a certain threshold, whereupon the GA is terminated to avoid over-training and avoid the risk of over fitting [101-103].

Genetic algorithms have found a number of applications in biospectroscopy, including in selection of preprocessing parameters for multivariate regression [104], as well as feature selection [76,105, 106].

Summary and Conclusions

Much progress has been made in recent years towards understanding the confounding factors which influence the integrity of vibrational spectra of biological tissues and cells. The potential of the techniques for biomedical and biochemical analysis applications lies in the ability to fingerprint the biochemical content, and changes to it, at a molecular level, in a label free manner. Such samples are intrinsically chemically and physically inhomogeneous, however, and it is important to differentiate the intrinsic biochemical profiles from confounding factors associated with sample scattering of instrumental responses.

In this context, a fundamental starting point is the quality of the recorded spectra, which can be influenced by sample presentation and measurement protocols. In IR absorption, wax embedded samples have been shown to exhibit minimal scattering artefacts, and measurement in attenuated total reflectance or using an integrated sphere can also reduce the associated spectral distortions. In Raman spectroscopy, large stray light backgrounds due to scattering of the source as well as the Raman bands themselves can be minimised by measurement in immersion, the water acting as an index matching fluid.

In both modalities, background can be minimised by appropriate choice of substrate. However, for applications in a clinical environment, cost implications may be a determining factor. In the case of IR for ex vivo clinical applications, a current imperative is to establish whether the spectral distortion associated with the EFSW effect are sufficient to dominate over the spectral variations associated with the pathological target, a consideration which may determine whether low cost, low-E slides can be employed for routine screening in the transfection mode, or whether more expensive transmissive slides are required.

The validation of the RMieEMSC correction algorithm using simulated datasets, in which known target spectra were distorted with various degrees of "dispersion artefacts" as well as other instrumental factors, demonstrates the importance of establishing confidence in the data processing protocol, based on knowing the

“right answer” [20]. A similar approach was taken by Keating et al., in exploring the relative merits of supervised and unsupervised CLS as well as SCCA in the biochemical profiling of subcellular spectral maps [13], as well as the validity of PLSR of spectral data against dose and cytotoxicity dependent targets to yield independent information regarding the direct chemical interaction of drugs within cells, and the subsequent, indirect cytological responses [93]. To date, a range of multivariate analytical techniques have been developed and applied to biospectroscopic datasets, in an academic context. Depending on the application, linear or nonlinear, supervised or unsupervised, may be most appropriate, although no extensive comparison of the range of techniques has been undertaken. Given the range of protocols for both pre and post processing, it would be of great benefit to the research community, and for the standardisation of protocols, towards clinical translation, to similarly validate all protocols against an established simulated dataset.

Regardless of the techniques applied or specific target application, it remains of critical importance that all details of data pre- and post- processing procedures are provided in publications. It remains similarly important to cross reference the spectral analysis with established biological or so-called “gold standard” clinical assays. Although clinical histology and cytology is primarily based on morphological changes, it is important to demonstrate that the analysis of the underlying biochemical changes can augment established practices, and ultimately improve on the understanding of disease onset and progression. Critically, while the research environment has demands on demonstrating the reproducibility of data, the demands of clinical deployment are substantially more rigorous. Studies must, therefore, be extended to use of a clinically appropriate scale and statistical analysis to be considered meaningful.

In terms of applications such as screening for modes of actions and efficacies of, for example, chemotherapeutic agents, correlation with established biochemical assays can help to establish spectral markers, which could lay the foundation for the development of high content, label free analysis based on spectral-omics.

- H.J. Byrne, M. Baranska, G.J. Puppels, N. Stone, B. Wood, K.M. Gough, P. Lasch, P. Heraud, J. Sulé-Suso, and G.D. Sockalingum, *Analyst*, 2015, **140**, 2066 - 2073
- H.J. Byrne, G.D. Sockalingum and N. Stone, in “*Biomedical Applications of Synchrotron Infrared Microspectroscopy: A Practical Approach*”, David Moss, (Editor), RSC Analytical Spectroscopy Monographs No. 11, 2011, ISBN: 978-0-85404-154-1
- K.M. Ostrowska, A. Malkin, A. Meade, J.J. O’Leary, C. Martin, C. Spillane, H.J. Byrne and F.M. Lyng, *Analyst*, 2010, **135**, 3087–3093
- K.M. Ostrowska, A. Garcia, A.D. Meade, A. Malkin, I. Okewumi, J.J. O’Leary, C. Martin, H.J. Byrne and F.M. Lyng, *Analyst*, 2011, **136**, 1365-1373
- A.D. Meade, C. Clarke, H.J. Byrne, F.M. Lyng, *Radiation Research*, 2010, **173**, 225-237
- H. Nawaz, F. Bonnier, P. Knief, O. Howe, F.M. Lyng, A.D. Meade and H.J. Byrne, *Analyst*, 2010, **135**, 3070-3076
- H. Nawaz, A. Garcia, A.D. Meade, F.M. Lyng, and H.J. Byrne, *Analyst*, 2013, **138**, 6177-6184
- A. V. Feofanov, A. I. Grichine, L. A. Shitova, T. A. Karmakova, R. I. Yakubovskaya, M. Egret-Charlier and P. Vigny, *Biophysical Journal*, 2000, **78**, 499-512.
- S.F. El-Mashtoly, D. Petersen, H.K. Yosef, A. Mosig, A. Reinacher-Schick, C. Kotting and K. Gerwert, *Analyst*, 2014, **139**, 1155-1161.
- Z. Farhane, F. Bonnier, A. Casey and H.J. Byrne, *Analyst*, 2015, DOI: 10.1039/C5AN00256G
- P. Knief, C. Clarke, E. Herzog, M. Davoren, F.M. Lyng, A.D. Meade and H.J. Byrne, *Analyst*, 2009, **134**, 1182-1191
- J. Dorney, F. Bonnier, A. Garcia, A. Casey, G. Chambers, and H.J. Byrne, *Analyst*, 2012, **137**, 1111-1119
- M. Keating, F. Bonnier and H.J. Byrne, *Analyst*, 2012, **137**, 5792-5802
- H.J. Byrne, K.M. Ostrowska, H. Nawaz, J. Dorney, A.D. Meade, F. Bonnier, F.M. Lyng, in “*Optical Spectroscopy and Computational Methods in Biology and Medicine*”, Challenges and Advances in Computational Chemistry and Physics, 2014, Volume 14, pp 355-399, M. Baranska (ed.), Springer Science+Business Media, Dordrecht ISBN 978-94-007-7831-3
- See for example, M. Keating and H.J. Byrne, *NanoMedicine*, 2013, **8**, 1335-1351
- M. Romeo and M. Diem, *Vibrational Spectroscopy*, 2005, **38**, 129–132
- A. Kohler, J. Sule-Suso, G. D. Sockalingum, M. Tobin, F. Bahrami, Y. Yang, J. Pijanka, P. Dumas, M. Cotte, D. G. van Pettius, G. Parkes and H. Martens, *Appl. Spectrosc.*, 2008, **62**, 259-266
- P. Bassan, H.J. Byrne, F. Bonnier, J. Lee, P. Dumas and P. Gardner, *Analyst*, 2009, **134**, 1586–1593
- P. Bassan, H.J. Byrne, J. Lee, F. Bonnier, C. Clarke, P. Dumas, E. Gazi, M.D. Brown, N.W. Clarke and P. Gardner, *Analyst*, 2009, **134**, 1171–1175
- P. Bassan, A. Kohler, H. Martens, J. Lee, H.J. Byrne, P. Dumas, E. Gazi, M. Brown, N. Clarke and P. Gardner, *Analyst*, 2010, **135**, 268 - 277
- M.J. Baker, C. Clarke, D. Demoulin, J.M. Nicholson, F. Lyng, H.J. Byrne, C.A. Hart, M.D. Brown, N.W. Clarke and P. Gardner, *Analyst*, 2010, **135**, 887-894
- B. Bird, M. Miljkovic’ and Max Diem, *J. Biophotonics*, 2010, **3**, 597–608
- P. Bassan, J. Lee, A. Sachdeva, J. Pissardini, K.M. Dorling, J.S. Fletcher, A. Henderson and P. Gardner, *Analyst*, 2013, **138**, 144-157
- T.P. Wrobel, B. Wajnchold, H.J. Byrne, and M. Baranska, *Vibrational Spectroscopy*, 2013, **69**, 84-92
- P. Bassan, J. Mellor, J. Shapiro, K.J. Williams, M. Lisanti and P. Gardner, *Analytical Chemistry*, 2014, **86**, 1648-1653
- S.M. Ali, F. Bonnier, A. Tfayli, H. Lambkin, K. Flynn, V. McDonagh, C. Healy, T.C. Lee, F.M. Lyng and H.J. Byrne, *Journal of Biomedical Optics*, 2013, **18**, 061202-1/12
- J.D. Bancroft and M. Gamble, *Theory and Practice of Histological Techniques*, 2002, Churchill Livingstone, London.
- A. Tfayli, C. Gobinet, V. Vrabie, R. Huez, M. Manfait and O. Piot, *Applied Spectroscopy*, 2009, **63**, 564-570

29. B. Bird, K. Bedrossian, N. Laver, M. Miljković, M. J. Romeo, and M. Diem, *Analyst*, 2009, **134**, 1067–1076
30. C.M. Krishna, G.D. Sockalingum, B.M. Vadhiraja, K. Maheedhar, A. C.K. Rao, L. Rao, L. Venteo, M. Pluot, D.J. Fernandes, M.S. Vidyasagar, V.B. Kartha, and M. Manfait, *Biopolymers*, 2007, **85**, 214–221
31. E. Ly, O. Piot, R. Wolthuis, A. Durlach, P. Bernard, and M. Manfait, *Analyst*, 2008, **133**, 197–205
32. L.M. Fullwood, D. Griffiths, K. Ashton, T. Dawson, R.W. Lea, C. Davis, F. Bonnier, H.J. Byrne and M.J. Baker, *Analyst*, 2014, **139**, 446–454
33. I. Marshall, S.D. Bruce, J. Higinbotham, A. MacLulich, J.M. Wardlaw, K.J. Ferguson, J. Seckl, *Magnetic Resonance in Medicine*, 2000, **44**, 646–649
34. F. Bonnier, A.Mehmood, P. Knief, A. Meade, W. Hornebeck, H. Lambkin, K. Flynn, V. McDonagh, C. Healy, T.C. Lee, F.M. Lyng, and H.J. Byrne, *Journal of Raman Spectroscopy*, 2011, **42**, 888–896
35. F. Bonnier, S.M. Ali, P. Knief, H. Lambkin, K. Flynn, V. McDonagh, C. Healy, T.C. Lee, F.M. Lyng, and H.J. Byrne, *Vibrational Spectroscopy*, 2012, **61**, 124–132
36. L. Kerr, H.J. Byrne and B.M. Hennelly, *Analytical Methods*, 2015, DOI: 10.1039/C5AY00327J
37. G.J. Puppels, F.F. de Mul, C. Otto, J. Greve, M. Robert-Nicoud, D.J. Arndt-Jovin and T.M. Jovin, *Nature*, 1990, **347**, 301–303
38. G.J. Puppels, W. Colier, J.H.F. Olminkhof, C. Otto, F.F.M. d. Mul and J. Greve, *Journal of Raman Spectroscopy*, 1991, **22**, 217–225
39. F. Bonnier, A.D. Meade, S. Merzha, P. Knief, K. Bhattacharya, F.M. Lyng and H.J. Byrne, *Analyst*, **135**, 3169–3177 (2010)
40. J.T. Bulmer, D.E. Irish, F.W. Grossman, G. Herriot, M. Tseng, and A.J. Weerheim, *Applied Spectroscopy*, 1975, **29**, 506–511.
41. S.M. Haight and D.T. Schwartz, *Applied Spectroscopy*, 1997, **51**, 930–938.
42. H. Martens, J.P. Nielsen and S.B. Engelsen, *Analytical Chemistry*, 2003, **75**, 394–404.
43. J.W. Chan, D.S. Taylor, T. Zwerdling, S.M. Lane, K. Ihara, and T. Huser, *Biophysics Journal*, 2006, **90**, 648–56.
44. J. Guo, W. Cai, B. Du, M. Qian, and Z. Sun, *Biophysical Chemistry*, 2009, **140**, 57–61.
45. Z.W. Huang, A. McWilliams, S. Lam, J. English, D.I. McLean, H. Lui, and H. Zeng, *International Journal of Oncology*, 2003, **23**, 649–655.
46. P. Knief, PhD Thesis, Dublin Institute of Technology, 2010
47. Z. Wang, D.J. Veltkamp and B.R. Kowalski, *Analytical Chemistry*, 1991, **63**, 2750–2756.
48. Z. Wang, T. Dean and B.R. Kowalski, *Analytical Chemistry*, 1995, **67**, 2379–2385.
49. C.M. Whelburg, D.M. Haaland, D.K. Melgaard, *Applied Spectroscopy*, 2002, **56**, 877–886.
50. R.J. McCreedy, *Raman Spectroscopy for Chemical Analysis*, ed. J.D. Winefordner, 2000, **157**, John Wiley and Sons, New York
51. A. Savitsky and M.J.E. Golay, *Analytical Chemistry*, 1964, **36**, 1627–1639.
52. H. Martens, T. Naes, *Multivariate Calibration*, 1992, John Wiley & Sons, New York
53. O. Maimon, L. Rokach, (Eds.), *Data Mining and Knowledge Discovery Handbook*, 2010, Springer US, Boston, MA.
54. N.K. Afseth, V.H. Segtnan, and J.P. Wold, *Applied Spectroscopy*, 2006, **60**, 1358–1367.
55. C. Beleites, U. Neugebauer, T. Bocklitz, Christoph Krafft, J. Popp, *Analytica Chimica Acta*, 2013, **760**, 25–33.
56. K. Varmuza and P. Filzmoser, *Introduction to Multivariate Statistical Analysis in Chemometrics*. 2009, CRC Press, Taylor and Francis Group, Boca Raton, FL
57. K. H. Esbensen, *Multivariate data analysis - in practice*. CAMO Software AS, Norway, 2006.
58. H. Martens and T. Naes, *Multivariate Calibration*. John Wiley and Sons, 1994.
59. M. Otto, *Chemometrics: Statistics and Computer Application in Analytical Chemistry*. Wiley VCH, Weinheim, 2007.
60. J. Riedl, S. Esslinger, C. Faul-Hassek and K. Kong, *Anal. Chim. Acta*, 2015, DOI: 10.1016/j.aca.2015.06.003.
61. B.A. Gutman, X. Hua, P. Rajagopalan, Y.-Y. Chou, Y. Wang, I. Yanovsky, A. W. Toga, C. R. Jack Jr., M. W. Weiner and P. M. Thompson, *NeuroImage*, 2013, **70**, 386–401.
62. D. Pérez-Guaita, J. Kuligowski, S. Garrigues, G. Quintásde and B.R. Wood, *Analyst*, 2015, **140**, 2422–2427
63. C. M. Rubingh, S. Bijlsma, E. P. P. A. Derks, I. Bobeldijk, E. R. Verheij, S. Kochhar and A. K. Smilde, *Metabolomics*, 2006, **2**, 53–61.
64. S. Smit, M. J. van Breemen, H. C. J. Hoefsloot, A. K. Smilde, J. M. F. G. Aerts and C. G. de Koster, *Anal. Chim. Acta*, 2007, **592**, 210–217
65. K. L. Brown, O.Y. Palyvoda, J.S. Thakur, S.L. Nehlsen-Cannarella, O.R. Fagoaga, S.A. Gruber and G.W. Auner, *J. Immunol. Methods*, 2009, **340**, 48–54
66. E. Gentleman, R.J. Swain, N.D. Evans, S. Boonrungsiman, G. Jell, M.D. Ball, T.A. Shean, M.L. Oyen, A. Porter and M.M. Stevens, *Nature Materials*, 2009, **8**, 763–770
67. D. C. Fernandez, R. Bhargava, S.M. Hewitt and I.W. Levin, *Nature Biotechnology*, 2005, **23**, 469–474
68. C. Krafft, T. Knetschke, R.H.W. Funk and R. Salzer, *Analytical Chemistry*, 2006, **78**, 4424–4429
69. F. Lyng, E. Ó Faoláin, J. Conroy, A. Meade, P. Knief, B. Duffy, M. Hunter, J. Byrne, P. Kelehan and H.J. Byrne, *Experimental and Molecular Pathology*, 2007, **82**, 121–129
70. P. Crow, B. Barrass, C. Kendall, M. Hart-Prieto, M. Wright, R. Persad and N. Stone, *British Journal of Cancer*, 2005, **92**, 2166–2170.
71. A. Tfayli, O. Piot, A. Durlach, P. Bernard and M. Manfait, *Biochim Biophys Acta*, 2005, **1724**, 262–269
72. C. Krafft, D. Codrich, G. Pelizzo and V. Sergo, *J Biophotonics*, 2008, **1**, 154–169
73. C. Krafft, D. Codrich, G. Pelizzo and V. Sergo, *Analyst*, 2008, **133**, 361–371
74. F. Westad and H. Martens, *Journal of Near Infrared Spectroscopy*, 2000, **8**, 117–124
75. F. Despaigne and D. L. Massart, *Analyst* 1998, **123**, 157R–178R
76. K. Varmuza and P. Filzmoser, *Introduction to Multivariate Statistical Analysis in Chemometrics*, 2009, CRC Press, Taylor and Francis Group, Boca Raton, FL,
77. A.C. Rencher, W.W. Christensen, *Methods of multivariate analysis*, 3rd ed., 2014, John Wiley and Sons, New York
78. N. Bergner, B.F.M. Romeike, R. Reichart, R. Kalff, C. Krafft and J. Popp, *Analyst*, 2013, **138**, 3983–90
79. M. Miljković, T. Chernenko, M.J. Romeo, B. Bird, C. Matthäus and M. Diem, *Analyst*, 2010, **135**, 2002–13.

80. M. Hedegaard, C. Matthäus, S. Hassing, C. Krafft, M. Diem and J. Popp, *Theoretical Chemistry Accounts*, 2011, **130**, 1249–1260
81. S.M. Ali, F. Bonnier, H. Lambkin, K. Flynn, V. McDonagh, C. Healy, T.C. Lee, F.M. Lyng, and H.J. Byrne, *Analytical Methods*, 2013, **5**, 2281
82. J.H. Ward, *J. Am. Stat. Assoc.*, 1963, **58**, 236–244.
83. C. Matthaus, T. Chernenko, J.A. Newmark, C.M. Warner, and M. Diem, *Biophysical Journal*, 2007, **93**, 668–673
84. Z. Lu and T.K. Leen, in *Advances in Neural Information Processing Systems 17*. MIT Press, 2005, pp. 849–856.
85. P. Bassan, A. Sachdeva, A. Kohler, C. Hughes, A. Henderson, J. Boyle, J.H. Shanks, M. Brown, N.W. Clarke and P. Gardner, *Analyst*, 2012, **137**, 1370–7
86. K. Kinoshita, M. Miyazaki, H. Morita, M. Vassileva, C. Tang, D. Li, O. Ishikawa, H. Kusunoki and R. Tsenkova, *Scientific Reports*, 2012, **2**, 856
87. JMP Nascimento, S. Member and JMB Dias, *IEEE Transactions on Geoscience and Remote Sensing*, 2005, **43**, 898–910
88. T. Chernenko, R.R. Sawant, M. Miljkovic, L. Quintero, M. Diem and V. Torchilin, *Molecular Pharmaceutics*, 2012, **9**, 930–6
89. F. Bonnier and H.J. Byrne, *Analyst*, 2012, **137**, 322–32
90. S. M. Ali, F. Bonnier, K. Ptasinski, H. Lambkin, K. Flynn, F.M. Lyng, H.J. Byrne, *Analyst*, 2013, **138**, 3946 - 3956
91. C.L. Zavaleta, B.R. Smith, I. Walton, W. Doering, G. Davis, B. Shojaeib, M.J. Natan and S.S. Gambhira, *Proceedings of the National Academy of Sciences of the United States of America*, 2009, **106**, 13511–6
92. R. M. Balabin and S. V Smirnov, *Anal. Chim. Acta*, 2011, **692**, 63–72
93. M.E. Keating, H. Nawaz, F. Bonnier and H.J. Byrne, *Analyst*, 2015, **140**, 2482–2492
94. J. Schmitt and T. Udelhoven, in *Infrared and Raman Spectroscopy of Biological Materials* (H. U. Gremlich and B. Yan, Eds.), pp. 379–420. Marcel Dekker, Basel, 2001.
95. A. Krogh, *Nat Biotechnol* 2008, **26**, 195–197
96. J. J. Liberda, K. Schnarr, P. Coulbaly and D. R. Boreham, *Int J Radiat Biol* 2005, **81**, 827–840
97. R. Mathieu, E. Martin, R. Gschwind, L. Makovicka, S. Contassot-Vivier and J. Bahi, *Physics in medicine and biology*, 2005, **50**, 1019–1028
98. H. R. Vega-Carrillo, V. M. Hernandez-Davila, E. Manzanares-Acuna, G. A. Mercado, E. Gallego, A. Lorente, W. A. Perales-Munoz and J. A. Robles-Rodriguez, *Radiation protection dosimetry*, 2006, **118**, 251–259
99. T. Udelhoven, D. Naumann, J. Schmitt, *Appl. Spectrosc* 2000; **54** 1471–1479.
100. P. Lasch, M. Diem, W. Hänsch and Dieter Naumann, *J. Chemom.* 2007, **20** 209–220
101. R. Leardi, *Journal of Chemometrics*, 2000, **14**, 643–655.
102. R. Leardi and A.L. Gonzalez, *Chemometrics and Intelligent Laboratory Systems*, 1998, **41**, 195–207.
103. D. Jouanrimbaud, D.L. Massart, R. Leardi, and O.E. Denoord, *Analytical Chemistry*, 1995, **67**, 4295–4301.
104. R.M. Jarvis, and R. Goodacre, *Bioinformatics*, 2005, **21**, 860–868
105. H. Yoshida, R. Leardi, K. Funatsu, and K. Varmuza, *Anal. Chim. Acta*, 2001, **446**, 485–494
106. A.D. Meade, C. Clarke, H.J. Byrne and F.M. Lyng, Proceedings of IEEE International Conference on BioInformatics BioMedicine, DOI: 10.1039/C4AN02167C, (2014)