Chemometric Treatment of Raman Spectra in Biomedicine and Food Science

Ángel Sánchez-Illana[1], David Pérez-Guaita[1*]

1. Department of Analytical Chemistry, Chemistry Faculty, Universtitat de València, Doctor Moliner 50, 46100 Burjassot, Spain

** david.perez-guaita@uv.es*

1. Introduction

Biomedicine and food science are active fields of research with deep economic and social implications. Both fields require analytical methodologies which provide quantitative and qualitative information about the chemical components of the samples and thus, new analytical improvements are continuously developed to drive innovative lines of research as well as to perform routine analysis.

In this context, Raman spectroscopy (RS) is a versatile technique that provides wide range of tools to investigate the composition of biological samples, from high throughput microscopes for the analysis of cells and bacteria to handheld portable instruments to investigate the composition of bulk samples *in situ*. Although clinical and food sciences and industries are very distinct fields, both deal with complex biological samples containing several components such as water, carbohydrates, proteins, or lipids.

RS has many advantages when it comes to the analysis of food and clinical samples. The characteristic Raman spectrum of a biological sample constituent, caused by the unique vibrational modes of each Raman active molecule, provides enough selectivity to identify

biological molecules, pharmaceuticals, adulterants, and contaminants. Furthermore, since Raman intensities are proportional to the concentration of the molecules under study, quantification methodologies can be developed to predict the concentration of a nutrient or a clinical parameter. In addition, the weak Raman scattering of water facilitates the direct measurement on aqueous samples.

However, biological samples are diverse matrices containing a wide range of components, and their Raman spectrum is normally composed by a set of overlapped bands. This hampers the proper assignment of bands to specific molecular vibrations. Consider as an example the Raman shifts depicted in Figure 1 for the different moieties of lipids, nucleic acids, and proteins. The combination of these signals would result in a spectrum rich in chemical information, but it would be virtually impossible to extract this information with the naked eye. In practice, the experimental Raman spectra of biological samples are even more complicated because other molecules are also present [1–3].

Moreover, Raman spectra are acquired employing a variety of techniques including classical non-resonant Raman, resonant Raman (RR), surface-enhanced Raman scattering (SERS), surface-enhanced resonance Raman scattering (SERRS), and Fourier-transform Raman spectroscopy (FT-Raman). In these techniques, different experimental parameters are adjusted to optimize the analytical features. These technical parameters include the laser (e.g., wavelength, continuous/pulsed, power) and the chemistry of SERS substrates (e.g., metal composition, nanoparticle type, colloidal/solid) and cause different enhancements for each vibrational mode, which adds variability to the spectra. In summary, the extraction of information from the Raman spectra of clinical samples and foodstuffs is challenged by the intrinsic complexity

of the overlapped Raman bands and, in most of the cases, the univariate study of a single Raman variable is not enough. In this context, the use of chemometric approaches that can provide an integrated analysis of the whole spectra using multivariate approaches becomes indispensable.

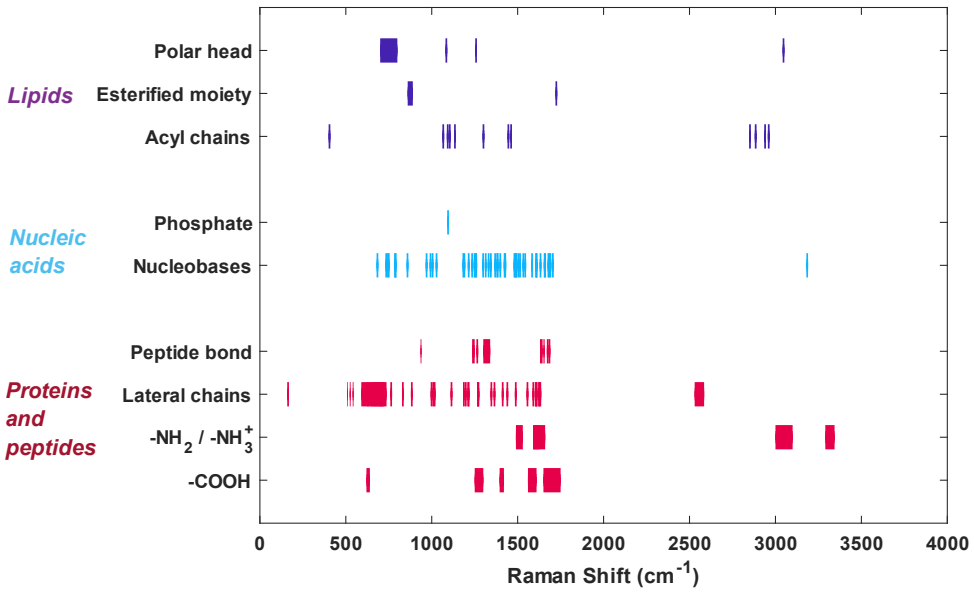

Figure 1. a) Raman shifts regions of different moieties of lipids, nucleic acids, and proteins. Data extracted from Niaura et al. [1].

A plethora of data analysis workflows employing multivariate statistics are currently being developed in the field of chemometrics, which has been influenced by emergence of machine learning, deep learning, and data mining. These methods have been applied to the analysis of Raman spectra of biological samples and foodstuffs and used for the discrimination (i.e. classification) or the determination of parameters (i.e. quantification). In this regard, comprehensive reviews and tutorials have been published in recent years, and the field aims to move towards the standardization of methodologies [4–7]. Here we provide a brief description

of the modelling process for the reader, and we refer to dedicated chapters in this book for further information.

The first step is preprocessing, which is used to reduce the noise and technical variations and enhance the signals related to the analyte or class of interest. This can include basic operations (e.g. spectral range selection, scaling) or more sophisticated routines such as baseline correction, random noise reduction or spectral differentiation [4,7]. Preprocessed spectra are then used in the model training, which is, in general, a function approximation [8]. Thus, any model considered is used to approximate a function that assign one element of response space $\mathcal{Y}$ from the spectra space $\mathcal{X}$. Given a spectrum matrix $\boldsymbol{X} \in \mathcal{X}$, and a model $f: \mathcal{X} \rightarrow \mathcal{Y}$, the response matrix $\boldsymbol{Y} \in \mathcal{Y}$ is approximated by the prediction of $\widehat{\boldsymbol{Y}} \in \mathcal{Y}$ as:

$$\boldsymbol{Y} \approx \widehat{\boldsymbol{Y}} = f(\boldsymbol{X}) \tag{1}$$

In the spectrum matrix $\boldsymbol{X}$, each x is a spectrum vector. Likewise, $\hat{y}$ is a predicted vector in the matrix $\widehat{\boldsymbol{Y}}$ of predicted responses (i.e., class or concentrations).

Then, validation approaches are used for optimizing the model fitting (i.e., optimizing the parameters of the model) and for a first evaluation of the results and significance of the model. For the former, different types of cross-validation are the most common methods. For the latter, resampling approaches are the most popular (e.g., a single evaluation set, leave-one-out, leave-more-out, single and double cross-validation, bootstrapping and permutation tests). Finally, independent testing (i.e., test set validation) is used to assess the actual predictive performance on new samples and it is the previous step towards the final implementation of the method in

routine analysis [4,7,9]. It should be noted that a proper test set validation, including samples as independent from the calibration sample set, is the only way to estimate error metrics [10].

In this book chapter we wanted to provide an overview about one of the most popular applications of the chemometric analysis of Raman data: the elucidation of chemical information contained in Raman spectra from biological matrices. To that end, we discuss state-of-the art chemometric workflows for most common scenarios in the fields of biomedicine and food science, assessing specific challenges and pitfalls and providing an outline about recent and future trends. Section 2 describes the use of chemometrics in the Raman analysis of food contamination, authentication, and in the prediction of nutritional parameters, while section 3 focuses on applications in diagnosis of diseases, prediction of clinical parameters, and analysis of clinical longitudinal studies as well as data fusion.

3. Analysis of Raman datasets in food science

When the first Raman analysis of foodstuffs was published in the pre-chemometrics era (i.e. early 70s), Raman bands were being integrated from the paper records using a classical planimeter. In this context, the implementation of early computers such as the IBM1800 for spectra processing was a breakthrough [11] that enabled faster analysis and minimized human related errors. Since then, the development of computing and chemometrics have allowed the study of an increasing number of food matrices qualitatively and quantitatively.

Different chemometric approaches have revealed the potential of Raman spectroscopy to the food and nutritional science community. However, these tools are still being used exclusively

for research purposes and they are rarely employed in official food safety laboratories. A lack of method harmonization of chemometric methods applied to Raman, and the need of technicians trained on advanced multivariate analysis could explain the marginal use of these tools in official institutions. This section presents state-of-the-art chemometrics methodologies for food Raman datasets covering their use, validation, and transferability to official laboratories in three main fields: 1) Analysis of food contamination, 2) Analysis of food authentication, and 3) Nutritional quality assessment.

For further clarification on food terminology, analytical guidelines, and limits we refer the readers to the publications made by Codex Alimentarius (FAO/WHO) (https://www.fao.org/fao-who-codexalimentarius).

3.1 Analysis of food contamination

Analysis of contaminants in foodstuffs allows the detection of harmful microbiological agents and the quantitative and qualitative evaluation of pesticide and pollutant residues. Due to the large variability among each class of contaminants and matrices, a wide range of methodologies based on the chemometric analysis of Raman data have been developed. Multivariate analysis of Raman bands can be used as a label-free approach which identifies spectral fingerprints of bacteria (i.e., composition of the outer surface of bacteria and the excretions of small molecules), fungi (i.e., excreted mycotoxins) or the contaminant itself (i.e., antibiotic, pesticide, or pollutant residues). Besides, multivariate analysis can also be used in labeled approaches for extracting the reporter signal within the matrix. Furthermore, more

complex recognition methods have also been developed combining the separation and enrichment of the analyte with label-free or label-based Raman approaches [12].

Table 1 exemplifies that, standard Raman and other related techniques such as SERS, Raman imaging or single cell Raman spectroscopy (SCRS) have been coupled with different chemometric workflows for the analysis of different types of contaminants and matrices. Among these chemometric approaches, classification models are the most common, but quantitative models have also been carried out. Preprocessing steps include baseline correction, Savitzky–Golay (SG) smoothing with or without derivatization, and standard normal variate (SNV) normalization. Regarding classification and regression, various methods have been proposed including principal component analysis (PCA), partial least squares (PLS), support vector machines (SVM) for both discriminant analysis (SVMDA) and regression (SVR), convolutional neural network (CNN), and recurrent neural network (RNN) among others.

For bacterial contamination analysis, SCRS offers an interesting alternative to traditional microbiological cultures, especially in terms of analysis time. However, Raman signals of bacteria are complex due to the low scattering efficiency of single cells, the high number of bands in their spectra, and the heterogeneity of bacteria individuals within the same strain. In order to overcome these challenges, models based on kernel principal component analysis-decision tree (KPCA-DT) and PCA-SVM were implemented by Yan et al. to classify among different genera and strain of bacteria cultures [13]. The main advantage of the applied methodologies was that, while standard PCA extracted only linear information, PCA-SVM or KPCA-DT were able to exploit non-linear features for the discrimination. Figure 2 shows the high discrimination performance of the KPCA-DT algorithm applied for the different bacteria strains.

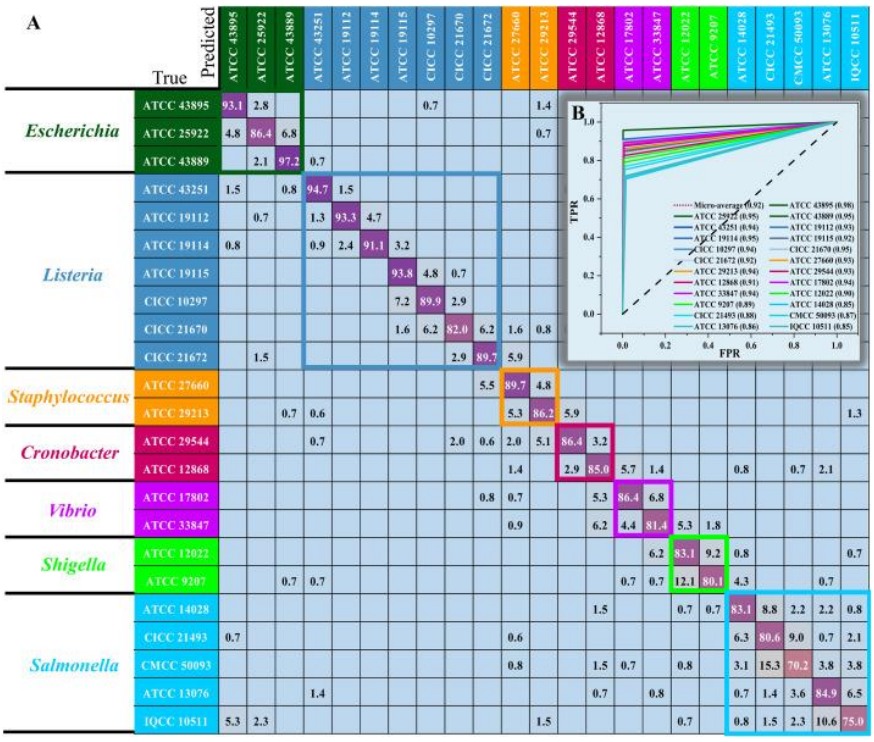

Figure 2. Discrimination performance of a KPCA-DT classification model for all bacterial strains analyzed with SCRS by Yan et al. (Reprinted from ref. [13] with permission from Elsevier).

Contamination	Analyte(s)	Type of sample	Raman technique	Chemometric technique(s)			Software	Ref.
				Pre-processing	*Training*	*Validation*		
Bacteria	*Bacteria Species*	Cultures	SCRS	SNV, S-G	KPCA-DT and PCA-SVM	kf-CV	Scikit-Learn (Python)	[13]
	Cyanobacteria	Spirulina	RS	Norm, S-G	PCA, SVMDA and SVR	Validation set	PLS toolbox (MATLAB)	[14]
Fungi	Aflatoxin B_1	Peanut oil	RS	ASN-SMOTE	CNN and RNN	Validation set	Tensorflow and Keras (Python)	[15]
	Zearalenone	Maize	RS	S-G	PLS, siPLS, ACO-PLS, siPLS-ACO	Validation set	MATLAB	[16]
Antibiotic residue	Chloramphenicol	Pork meat	SERS	RAW, SNV, MSC	PLS, VCPA-PLS, ACO-PLS, UVE-PLS, CARS-PLS	Validation set	MATLAB	[17]
	Chlortetracycline and oxytetracycline	Spiked animal feed	SERS	Norm, S-G	KNN, LDA, MLR, PLS	Validation set	SAS	[18]
Pesticide residue	Thiabendazole	Citrus	SERS	SNV, S-G, MSC, MC, AirPLS	PLS, GA-PLS, CARS-PLS, ACO-PLS	Validation set and reference method	MATLAB	[19]
Heavy metals	Mercury	Tea spiked samples	SERS	SNV, S-G	ACO-PLS	Validation set	MATLAB	[20]
Microplastics	Microplastics	Caryfish	RI	CR, Baseline corr., S-G	PVR*	-	WIRE and ImageJ	[21]

Table 1. Examples of recent publications employing Raman and chemometrics for contamination assessment.

Other methods based on machine learning for analyzing bacteria and mycotoxins include SVMDA, SVR, and deep learning neural networks such as CNN and RNN for both classification and quantification. Adejimi et al. [14] used SVMDA in spirulina samples for the classification of several species of cyanobacteria at concentration levels of 10^4 cells/ml. Furthermore, SVR allowed the quantification of bacteria within the 10^4 –10^8 range with acceptable root mean square error of

prediction (RMSEP) values. For hazardous mycotoxin analysis, CNN and RNN models were proposed by Deng et al. in peanut oil samples. [15] It was observed that with CNN and RNN the complex spectra preprocessing steps were not necessary, but the dataset had to be balanced (i.e., same elements of each class) to obtain proper results. To that aim, a synthetic minority class oversampling technique (ASN-SMOTE) was implemented to generate new similar data of the minority class. The methodology reported a perfect classification and a quantification with ratio of performance to deviation (RPD) values of 3.58 and 4.86 for CNN and RNN, respectively.

Variable selection strategies for PLS models have been applied to different contaminant analysis as it is shown in Table 1. For zearalenone mycotoxin detection, variable selection strategies were applied prior to PLS models (i.e., synergy interval-PLS [siPLS], ant colony optimization-PLS [ACO-PLS], and siPLS-ACO), and these models were compared to the PLS with the full spectral range [16]. The results suggested that siPLS-ACO model produced the best results among the four models with a R^2 of 0.9 and RMSEP of 87.9 µg/kg. However, the authors only proposed their method for screening approaches due to high limit of detection.

Application of variable selection approaches for PLS in SERS methods have also been published recently for antibiotic residues, pesticide residues, and heavy metals [17–20]. Moreover, Pan et al. also proposed the use of adaptive iteratively reweighted Penalized Least Squares (airPLS) for baseline correction.[19] Figure 3 (a-d) exemplifies the application of airPLS for preprocessing of SERS spectra of citrus samples. Genetic algorithm-partial least squares (GA-PLS) prediction was then applied for the quantification of thiabendazole (Figure 3 (a-b)). The method was validated by method comparison (i.e., comparison with HPLC reference method)

using a t-test which showed no significant differences and the limit of detection obtained meets the corresponding official standard.

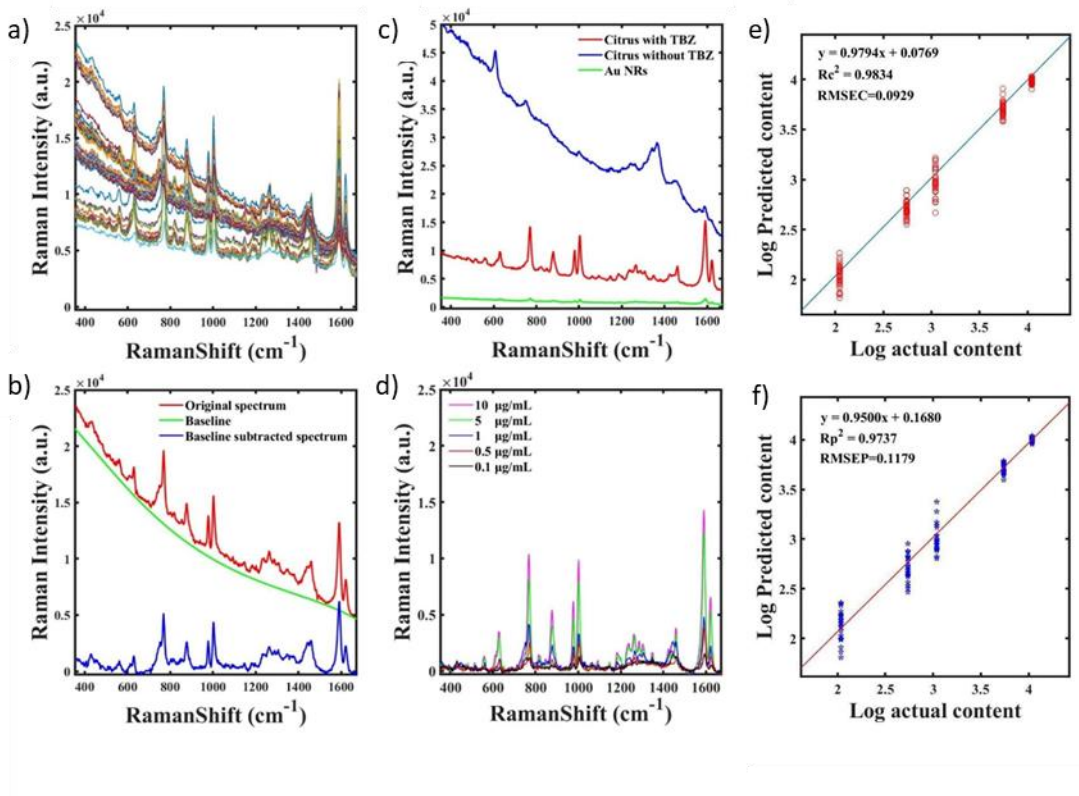

Figure 3. Raman spectra of thiabendazole (TBZ) residues with different concentrations in citrus (a), SERS spectra of gold nanorods (SERS substrate), citrus without TBZ and citrus with TBZ (b), the Air-PLS fitted baseline and the baseline subtracted spectra (c) and the average spectra of TBZ residues with different concentrations after baseline correction (d). Scatter plots of predicted versus experimental values by the GA-PLS model for the calibration (e) and prediction (f) sets (Reprinted from Ref. [19] with permission from Elsevier.)

3.2 Analysis of food authentication

International authorities commit significant resources to evaluate the inauthenticity of food [22] in order to prevent fraud and ensure food safety. Adulteration could compromise the human health as manifested by well-known foodborne illness and mass food poisoning such as the aniline adulteration in the toxic oil syndrome in 1981 in Spain (more than 20,000 people affected and 2,500 deaths) [23] or the scandal of the melamine in 2008 in China (more than 300,000 people affected and 6 infant deaths) [24]. In this regard, RS and chemometrics have emerged as valuable techniques providing several analytical methods, as it was recently reviewed by Xu et al. [25] and by Rui Hu et al. [26].

Table 2. summarizes recent publications of methodologies based on Raman and chemometrics for food authentication with potential effects in human health. Classification methods such as linear discriminant analysis (LDA), PCA, soft independent modelling of class analogies (SIMCA), KNN or partial least squares-discriminant analysis (PLSDA) are the most popular. Besides, PLS models are generally used for the quantification of certain adulterants.

Adulterations issue	**Type of sample(s)**	**Raman technique**	**Chemometric technique(s)**			**Software**	**Ref**
			Preprocessing	*Training*	*Validation*		
Grapefruit adulteration	Orange juice	RS	SNV	PCA, PC-DFA, PLSDA, PLS	CV	MATLAB	[27]
Eggs falsification	Eggs	RI	MSC, airPLS	PCA, PLSDA	-	MATLAB	[28]
Sudan dyes adulteration	Paprika powder	RS	IPF	PCA, PLS	Validation set	MATLAB and Unscrambler	[29]
Waste cooking oil	Edible oils spiked with waste oil	RS	Baseline corr., SG, SNV	iPLS, siPLS	Validation set	OptoTrace	[30]
Melamine adulteration	Powder milk	RS	SG, SNV	PCA, SIMCA	Validation set	SOLO	[31]

Table 2. Examples of recent publications employing Raman and chemometrics for food adulteration.

Grapefruit adulteration in juices is, surprisingly, a serious adulteration since this juice has dangerous interactions with some pharmaceuticals. Employing a portable handheld Raman spectrometer, Varnasseri et al. [27] developed differnt models based on PCA, principal component-discriminant function analysis (PC-DFA), PLSDA and PLS for the classification of adulterated juices and for the quantification of the adulterant. PCA and PC-DFA were used for exploratory analysis and for identifying sources of variation associated with the origin of the juice that could interfere with the quantitative model. This effect was mitigated via PLSDA/PLS hierarchical model to predict the juice origin followed by a regression to assess the adulteration degree.

Another important application recently developed is the detection of melamine adulteration in milk [31]. Figure 4. depicts the Raman spectra of melamine adulterated milk powder (MP) samples, acquired with two different portable Raman spectrometers, before and after spectra preprocessing (See insets). As it is shown, the derivatization of the spectra allowed the resolution of characteristic melamine bands (e.g., 676 cm^{-1}). Results also indicated that one of the portable devices (Figure 4(b)) was less sensible due to their poor lineal dependence with the melamine concentration. The chemometric approach proposed in this work was based on SIMCA classification model and allowed the classification of all control samples with no false negatives and the detection of adulterated milk samples (i.e., ≥ 1% of adulteration) with no false positives.

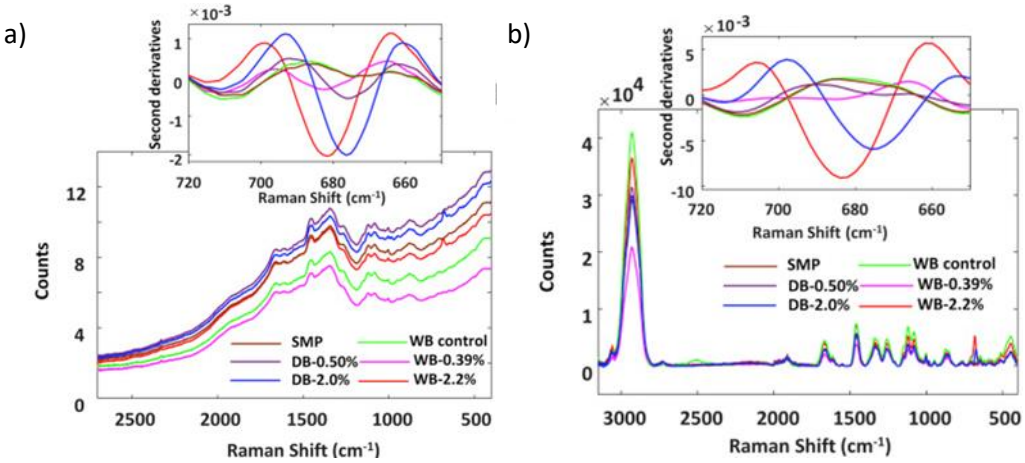

Figure 4. Raman spectra for two representative MP (labeled standard milk powder [SMP] and wet blend [WB]) samples and four MP blends measured with the TrueScan (a) and Bravo devices (b). WB and dry blend (DB) were measured with two different concentrations of melanine adulteration. Inset are the second derivative spectra for same samples. Inset are the second derivative spectra for the corresponding samples. (Reprinted from Ref. [31] with permission from Elsevier.)

Furthermore, Raman imaging has also been used in adulteration studies. Joshi et al. [28] have developed an interesting method for the detection of fake eggs, which is a dangerous falsified food that contains substances with potentially harmful effects on human health. Authors demonstrate that PCA of hyperspectral images of real and fake eggs allowed the detection of fake eggshells, albumens and yolks. Additionally, they also employed airPLS and a fluorescence correction algorithm to monitor a characteristic Raman shift found for the fake eggshell, enabling the prompt detection of fake eggs (See Figure 5). Finally, other studies have focused on edible oil adulteration [26]. In this regard, Li et al. [30] developed a method for analyzing waste oil (i.e., used cooking oil) contamination employing interval PLS (iPLS) and siPLS models. Results showed

that the use of siPLS enables the prediction of waste oil concentration with adequate RMSEP and a limit of detection of 0.5%.

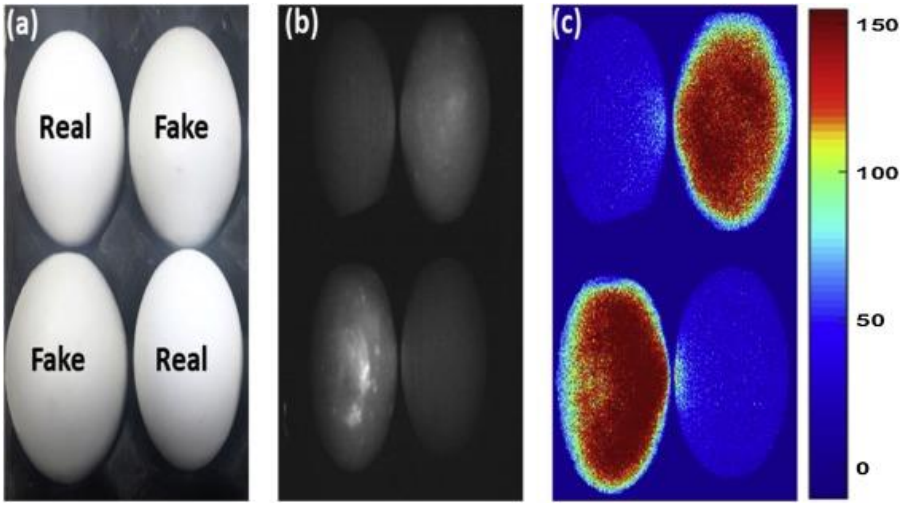

Figure 5. Original (a), selected waveband (b), and fluorescence corrected (c) Raman image at 1295 cm−1 for real and fake egg samples. (Reprinted from Ref. [28] with permission from Elsevier.)

3.3 Nutritional parameters assessment

Vibrational spectroscopy has been widely applied for the fast evaluation of nutritional parameters of food. Chemometrics has plaid a key role in the development of these methodologies, allowing the multicomponent quantification of parameters on complex liquid and solid matrices. Recent examples include *in situ* Fourier-transform infrared spectroscopy (FTIR) methods for quantitative profiling [32]. Some of them have been implemented into commercial analyzers such as the Miris HMA system for human milk (Miris Human Milk Analyzer, Miris AB, Uppsala, Sweden). In this context, RS has also emerged as a fast technique for analyzing nutritional profiles [33]. Table 3 summarizes different methods recently published which involved

the chemometric treatment of Raman spectra of food. Considering the need of fast and cost-effective approaches in the food industry, standard RS has been the most popular approach. Regarding the multivariate methodologies, conventional PCA, PLS, and PLSDA models [34,35] are normally applied, although other methods based on spectral deconvolution have been also proposed [36].

For example, nutritional studies feature the use of RS for the nutritional evaluation of extremely complex samples such as bee pollen or infant formula. In these approaches, the corresponding reference methods were firstly used for the analysis of the parameters. These values were used for the calibrations of PLS models and then, models were validated with external validation sets or cross-validation. In pollen analysis, the relative errors of prediction reported for macronutrients quantification (i.e., total protein, red. sugars, fats) were in the 1.9%–2.6% range comparable with the prediction errors obtained with FT-IR models of 1.7%–3.6% [34]. Similar results were obtained in the analysis of the nutritional profile of infant formula by Wang et al. [35], which employed the same approach for the prediction of preheat and storage temperatures, formula storage time, FAST index, soluble protein, fat content and surface free fat content. Similarly, Yazgan et al. [37] used PLSDA for the prediction of preheat treatment and animal origin of milk.

Parameters	Type of sample(s)	Raman technique	Chemometric technique(s)			Software	Ref
			Preprocessing	*Training*	*Validation*		
Amylose fraction, Phe fraction, Trp fraction, succ. substitution	Rice	RS	Voigtian deconvolution, FFT smoothing, SG	LS of integrated areas	Reference method correlation	LabSpec and Origin	[36]
Total protein, red. sugars, fat, total polyphenols, antioxidant activity, color	Bee pollen	RS	MSC, MC	PCA, PLS	Validation set	PLS Toolbox (MATLAB) and TQ Analyst	[34]
Preheat treatment, storage temp, storage time, FAST index, sol. protein, fat, SFF	Infant formula	RS	FC, CR, SNV	PCA, PLSDA, PLS	CV	OMNIC and Unscrambler	[35]
Preheat treatment and animal origin	Milk	RS	WF, MC	PLSDA	Validation set	SOLO	[37]

Table 3 Examples of recent publications employing Raman and chemometrics for nutritional parameters assessment.

Spectral deconvolution algorithms have also demonstrated their applicability to nutritional quantitative profiling. Recently, Pezzotti et al. [36] proposed voigtian deconvolution as a spectral processing step to obtain characteristic signals (i.e. deconvoluted bands) of nutritional parameters in rice samples. The authors carried out a detailed band assignment based on previous literature and density functional theory (DFT) calculations (See figure 5a). Then, the bands corresponding to each nutritional component were integrated and fitted via least-squares to the “actual values” obtained by the reference method. Moreover, an algorithm was developed to plot each deconvoluted profile in a barcode like plot as shown in Figure 5b-c. These types of representations can provide an appealing approach for fast quality control assessment and product matching with quality control purposes.

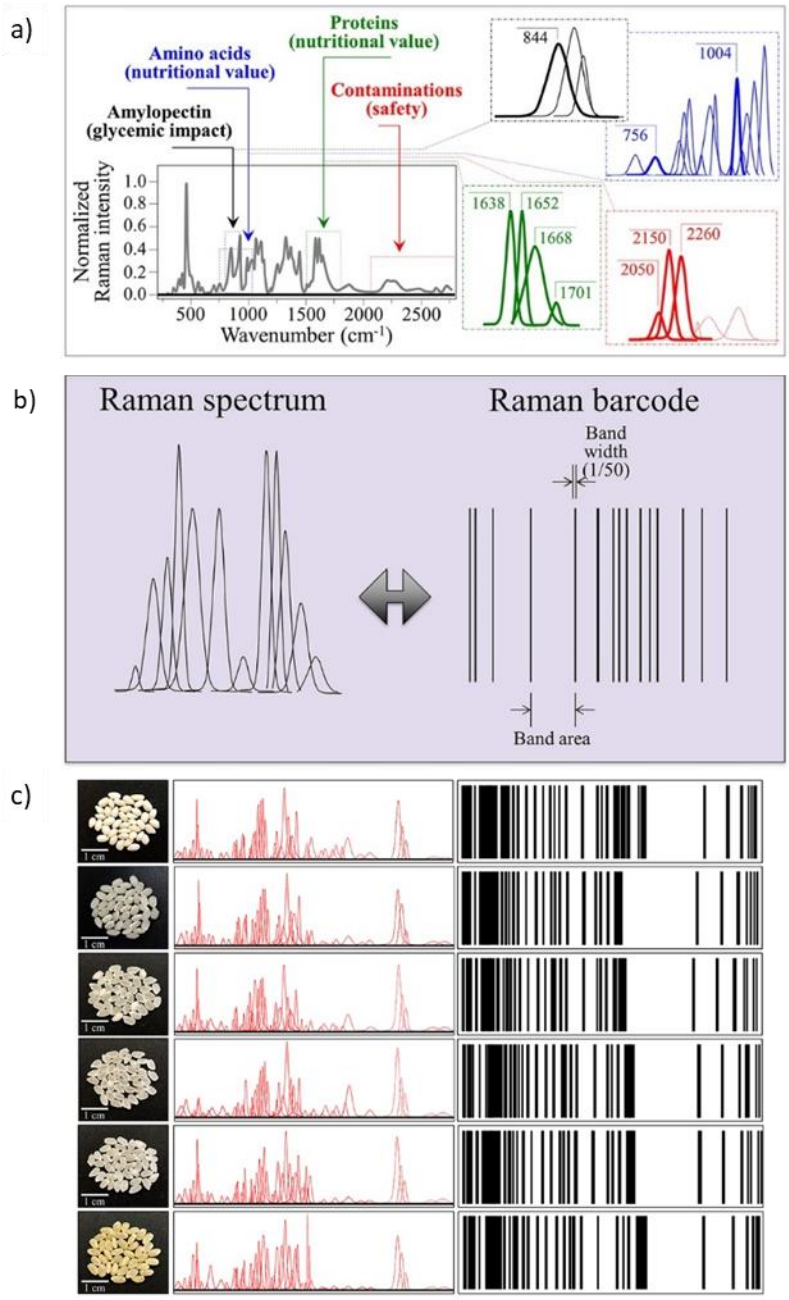

Figure 6. a) Schematic drafts locating different zones of the Raman spectrum representing different nutritional and quality characteristics and their deconvoluted sub-band components; and b), a proposed algorithm to convert the sub-band sequence into a linear barcode. c) images of different rice cultivars and associated Raman barcodes. (Modfied from Ref. [36], Creative Commons Attribution License, CC BY 4.0).

4. Analysis of Raman datasets in biomedicine and health science

4.1 Use of discriminant analysis for diagnosis

Effective disease treatment depends on a quick and correct diagnosis. If the illness causes detectable changes in the composition of a clinical sample, Raman bands can be used to determine if a disease is present. However, due to the complex nature of the Raman spectra of biological samples, the diagnosis is generally carried out by creating multivariate classification models using the Raman variables. To that end LDA, PLSDA, or SVMDA are common chemometrics discrimination methods which, are used to classify the data into "positive" and "negative" categories [38].

According to our analysis in the Web of Science database (November 2022) using the queries "Raman or SERS", "Diagnostic or Diagnosis", and "Chemometrics or Machine Learning", more than 500 papers have been published since 2010 dealing with the use of chemometrics and machine learning methodologies for the development of diagnostic tools. Figure 7(a) depicts the evolution of the number of papers published per year during the last decade, evidencing an exponential increasing of studies clearly driven by the irruption of machine learning. On the other hand, the term “chemometrics”, while it was the preferred keyword in the first five-year period, has experimented a more modest increasing followed by a plateau in the beginning of the new decade. Indeed, “Machine Learning” is considered a more appealing keyword and is used to attract interest from scientist from the clinical field (which are often not familiar with the term “chemometrics”). Furthermore, this change in nomenclature indicates that chemometricians are taking advantage of the artificial intelligence trend to adapt sophisticated machine learning

algorithms to treat RS signals [9]. The same analysis also showed that, among the different classification methodologies, LDA and SVM are the most popular techniques to create diagnostic models (See Figure 4.1(b)).

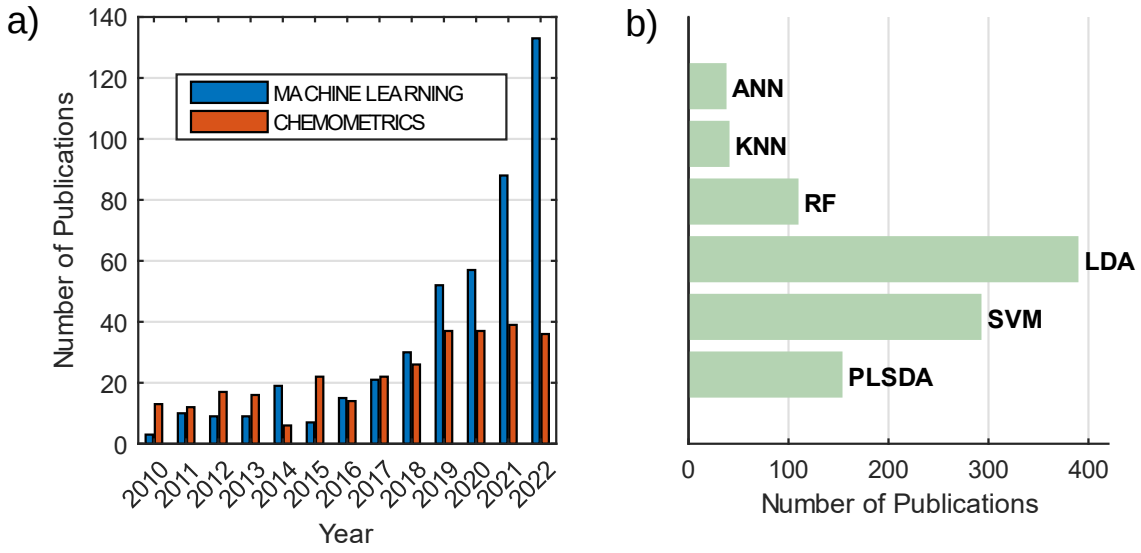

Figure 7. Results of a bibliographic analysis in the Web of science database (November 2022) searching the topics **"**Raman or SERS", "Diagnostic or Diagnosis" and "Chemometrics or Machine Learning" since 2010. (a) Evolution of the use "Machine Learning" and "Chemometrics". (b) Number of articles which included the terms "ANN", "KNN", "RF" and "LDA", "SVM" and "PLSDA".

The creation of diagnostic methods requires the consideration of specific aspects related to the clinical implantation of the models. The most important one, considering the strong impact of the treatment and decisions resulting from the analysis outcome, is the need of an accurate assessment of the true positive and true negative rates (sensitivity and specificity, respectively). In this sense, most of the output variables of discriminant models can be expressed as a

continuous score ($\hat{y}$ values in Equation 1) that indicates how close the sample is to the positive or negative classes. For example, Figure 4.2(a) depicts the predicted y score of a PLSDA model developed by Mancini et al. for leishmaniasis diagnosis in serum using SERS [39]. It can be seen that most of the positive samples hold positive values close to the reference $\hat{y}$ =1 value for positive class, while negative samples presented a predicted Y closer to the $\hat{y}$= 0 established for the negative samples. From these distributions of predicted $\hat{y}$ values, the probability of being either positive or negative can be calculated. Alternatively, a threshold value can be established to determine the sample class. In Figure 8(a) the threshold is depicted as a red dashed line which separates positive from negative samples, creating a few false negatives (green dots below the red line) and false positives (blue dots above the red line).

In principle, 0.5 is the midpoint between 0 and 1 and therefore it can be considered the most logical cut-off value. In contrast, authors used a Bayesian optimization technique to set the cut-off value slightly above 0.5. The optimal threshold can be indeed altered considering the requirements of a diagnostic methodology in terms of false positive and false negatives and, to that end, the so-called Receiver Operating Characteristic (ROC) curve is used. Figure 4.2(b) shows a ROC curve corresponding to the results of a logistic regression created by Webster et al. [40], which aimed to provide a screening method for oesophageal adenocarcinoma using biomarkers obtained from SERS analysis of blood (blue line). In the ROC, all the possible threshold values are considered, providing the corresponding true positive and true negative rates which are plotted as a curve in the sensitivity/1-specificity axes. Form this plot one can select the optimal threshold value according to the requirements of the diagnosis. Furthermore, the larger is the area under the curve (AUROC), the better is the performance of the classification, and therefore this value

can be used to compare the methodologies. For example, Figure 3b also sows the ROC curve for a logistic regression using clinical risk variables such as age and body mass index (red). It can be seen that the biomarker model performed much better with AUROC values of 0.830 and 0.930 for the red and blue ROCs, respectively.

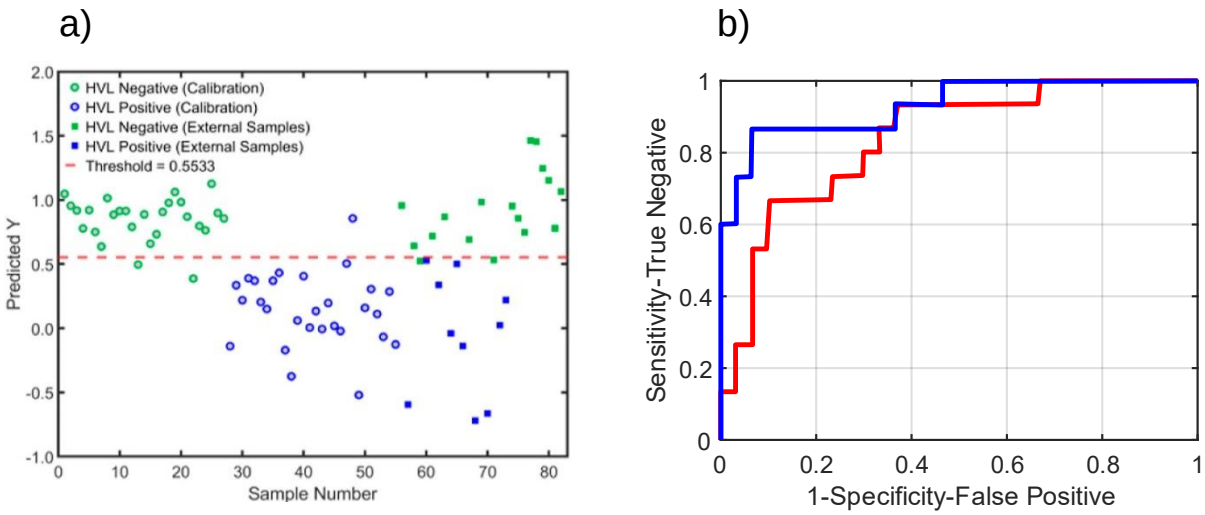

Figure 8. Typical plots used for the creation and evaluation of diagnostic models using Raman spectroscopy and Chemometrics. a) Predicted PLSDA Y-value for calibration and test samples obtained in a PLSDA model created for the diagnosis of leishmaniasis in serum using SERS. The dashed red line indicates the threshold selected in the study (Reprinted from Ref. [39] with permission from the American chemical Society). (b) ROC Curves obtained in logistic models for the screening of oesophageal adenocarcinoma using a biomarker panel (blue) and a clinical risk panel (Red). Adapted from Ref. [40] licensed under CC BY 4.0.

4.1.1 Single Point Measurements

One of the most popular Raman approaches employed in diagnostic methods is the measurement of a single spectrum representative of the general composition of a clinical biofluid or tissue. It can be performed in a straightforward and fast way by using inexpensive and portable instrumentation, offering methodologies very suitable for diagnosis within a Point of Care testing (PoC) context [41]. Furthermore, a single spectrum of specific regions of tissue can be recorded, serving as representative Raman variables which can be used for diagnostic purposes [42,43]. Similarly, a spectrum of a whole cell can be obtained and modelled to classify among different classes of cells [44] of bacterium [45].

Table 4.1 shows some examples of discrimination models based on the Raman spectra of different samples such as red blood cells, plasma, urine, and tissue. This exemplifies the versatility of RS, which is able to obtain information from a wide range of matrices. Regardless of the sample measured or disease under study, once the single spectrum is recorded, classification models previously created are used to directly predict the condition of the patient. In most cases, the disease under study affects several parts of the sample phenotype changing the concentration and/or internal structure of several sample biomolecules. This pattern of changes provokes shifts and alterations on the intensity of a lot Raman bands, which are used by the discriminant models to detect the presence of the disease. [46] In other cases, when the changes due to an individual biomarker dominates the spectrum, a more selective approach can be made by focusing on changes of the bands from the specific biomarker. For example, discriminant models created to identify malaria infected red blood cells focus on bands of hemozoin crystals, a byproduct of the digestion of hemoglobin by the plasmodium parasite [47,48].

Disease	Markers Targeted	PLS	Sample	Ref
Malaria	Hemozoin (Resonance)	PLSDA	Red Blood Cells	[47]
Leukemia	Multiple	PLSDA	Plasma	[46]
Bladder Cancer	Multiple	RF	Urine (SERS)	[49]
Oesophagus Cancer	Multiple	LDA	Oesophagus Tissue (Ex-vivo)	[50]
Nasopharyngeal Carcinoma	Multiple	CNN	Nasopharyngeal Tissue (In vivo)	[43]

Table 4. The versatility of diagnostic methods based on single point Raman Spectroscopy. Examples of methodologies proposed for the diagnosis of disease using chemometric models and Raman Spectra from different samples.

Finally, due to the tremendous advantages (e.g., Point of Care capability) and diagnostic potential of single point RS, the development of classification methods based on single point Raman is evolving rapidly with the implementation of more sophisticated algorithms. Shu et al. [43] has recently proposed a deep learning model featuring CNN for the identification of nasopharyngeal carcinoma from in-vivo Raman measurements of nasopharyngeal tissue. CNN models outperformed in most of the cases the use of classical PLS-LDA models, providing sensitivity values above 80% [43].

Nevertheless, one of the main challenges of using such sophisticated models is the need for a large number of samples for the training step. This is a notable limitation in the case of diagnostic studies, where it is hard to obtain a considerable number of patients with a proper diagnosis. In this sense, scientist can include several spectra from the same patient [43], or apply novel data augmentation methodologies [51]. Moreover, the validation can also be adapted to deal with an unbalanced number of data classes [52]. It must be noted that these strategies

should be implemented carefully because they can lead to over optimistic results when replicates of the same sample are included in calibration and testing, providing a “leaky” validation (also called replicate sample trap in the chemometric field) [53].

4.1.2 Imaging

Hyperspectral Raman images can also be used for creating classification models, which take into account both the composition of the sample (captured by the Raman spectra) and spatial features of the images for the diagnosis. In this case, however, the complexity of the models increases exponentially with the size of the images, as it requires to deal with a hypercube instead of a single spectrum per sample. In most of the cases, with the use simple Raman intensity maps or unsupervised analysis such as HCA or PCA, the submicron lateral resolution of Raman spectroscopy enables the spatial allocation of organelles within a cell and the assignment of different kinds of tissue on a tissue biopsy [54].

However, due to the high phenotypic variability among different patients, the creation of robust models that can predict the “positive” and “negative” nature of a sample from the hyperspectral image represents a more challenging task. It generally needs from more sophisticated algorithms and a large database of images for representative training and validation. Machine learning based techniques have been extensively used in regular histopathology for automatized diagnosis by analyzing stained images. This requires the segmentation of images with three channels corresponding to red, green, and blue intensities. The staining can be specific to a biomolecule (e.g. DNA or Lipids), providing chemical information which is used to identify spatial features (e.g. size and shape of the nucleus) that are used by

advanced algorithms to provide a diagnosis. In contrast, Raman imaging provides label-free chemical information as a set of thousands of variables, which can be included in computing intensive models that combine both spatial and chemical features in tissue [55,56] and cells [57].

4.2 Use of regression methodologies for the quantification of clinical parameters

The use of RS for the prediction of clinical parameters in biofluids has been proposed since the early 2000s [58]. Multivariate models aim to predict a continuous clinical parameter which can be used for a wide range of clinical applications including risk assessment, diagnosis, or prognosis. Figures of merit used in regression aim to assess the performance of the model by calculating the prediction error (i.e. the difference between the predicted and actual values). The most popular figure is the RMSEP, which can be normalized using the standard deviation, average or range of the actual values to compare among different validation sets. In the case of biomedical methodologies, it is recommended to always plot the actual versus predicted values to study the performance along the whole range of concentrations, putting special interest on the pathological values.

Again, the versatility of Raman spectroscopy has been exploited to develop quantifications of different parameters in different matrices (see examples in Table 5). In general, the complexity of the model depends on the analytical problem under study. For example, if SERS offers enough intrinsic selectivity for a specific analyte, quantifications can be carried out by simple models based on the ratio of bands, which use an internal standard to compensate for the irreproducibility of SERS signals [59]. In other cases, due to the large complexity of the matrices involved, SERS signals requires from machine learning methodologies. [60] Finally, deep learning

is also being implemented in biomarker detection for treating signals from multiplexed sensors with different SERS-active “nanorattles”. [61]

When the prediction of clinical parameters is performed using non-enhanced Raman spectroscopy, the regression becomes more challenging. The bands of other biomolecules from complex biological matrices overlap with the ones from the target analyte. In these cases, to create robust quantification models, extensive training and testing sets are required, which should be representative of a wide clinical range of both analytes and interferences contained in the sample. For example, a recent study used PLS and SVR for the multiparametric quantification of quality parameters of blood using space offset Raman spectroscopy (SORS) [62]. The quantification of glucose and lactate allowed the quality control of blood prior to donation through the blood bag, eliminating any potential contamination of blood, which is key in the transfusion process.

Analyte	Raman technique	Data analysis method	Sample	Ref
GSH	SERS	Ratio of Bands	Lysed RBC	[59]
Virus	SERS	SVR	Virus Solution	[60]
mRNA	SERS	CNN	Tissue	[61]
Total protein, glucose, urea, uric acid, cholesterol, triglycerides, HDL cholesterol, and LDL cholesterol.	Raman	PLS	Oesophagus Tissue (Ex-vivo)	[58]
Glucose and Lactate	SORS	PLS	Supernatant Mixtures of RBC	[62]

Table 5. Examples of recent publications of quantitative applications of clinical parameters employing Raman spectroscopy.

4.3 Analysis of longitudinal datasets

One of the key advantages of Raman spectroscopy is that, under well controlled laser power conditions, it does not destroy the sample. Therefore, it can be used to repeatedly analyze the same sample over time in longitudinal studies. Clinical researchers have taken advantage of this feature to characterize changes in the composition of cells in different time dependent biological processes [63]. Longitudinal experiments generally require the monitorization of the spectra of cells incubated in the presence of an external agent such as a drug, a toxic molecule or ultraviolet radiation. The outcome is a time dependent dataset which provides information about subtle changes in the molecular composition over time provoked by the presence of the agent, offers valuable insights into the underlying mechanisms of drug activity [64] or toxicity [65] involved.

The treatment of such datasets requires specific models that can identify spectral changes over time and relate these changes to specific biomolecules. The most straightforward approach is to apply PCA and look for principal components which capture the variance associated to the effect investigated. In this case, the score values can be used to study the evolution of the uptake of a drug or cellular response to a toxic over time. Furthermore, loading values can be used for the biological interpretation of the changes, for example by identifying bands of the drug or changes in bands of the cellular components caused by the cell response.

Unfortunately, in most of the cases, other sources of variation dominate the firsts principal components and supervised analysis is needed to investigate subtle changes associated to the inoculation time. To that end, regression models such as PLS can be applied by using inoculation time as a variable to predict ($\hat{y}$ values in Equation 1) and the Raman spectra as predictors. Then, the regression vector can be used to investigate spectral changes associated to the incubation

time [64]. One of the main limitations of using PLS is that it is based on linear assumptions and, in contrast, complex processes such as drug uptake or toxicity are not-linear with time [66]. Non-linear models such as SVR can be used, but unfortunately, they do not provide a regression vector for the investigation of spectral bands involved in the modelling.

Multivariate curve resolution-alternate least squares (MCR-ALS) has been proposed as a suitable alternative in this context [66,67]. MCR-ALS uses iterative optimization to resolve multiple component responses in mixtures by fitting them to a bilinear model, which relates the dataset to a set of components with specific concentration profiles and reference spectra. It requires the input of a few parameters, such as an estimation of the number of components, an initial guess of concentrations or spectra and a set of constraints to guide the iteration (e.g., non-negativity). Its main advantage is that, while the Raman intensities are still assumed to be linear with the concentration of the components, the evolution of the different components over time can follow any profile.

The differences between PLS and MCR-ALS have been studied using datasets which simulated Raman spectra of A549 cells incubated with doxorubicin (see Figure 9) [66]. The simulation included two different effects with two different temporal evolutions (see Figure 9(a)). The first one (blue) simulated the fast uptake of DOX by the cell, and the second one (green) simulated the subsequent slower cellular response. The simulated bands of the uptake included (see Figure 9(b)) bands of doxorubicin (orange spectrum) and changes in DNA conformation associated to the binding of the drug (blue spectrum). The simulated bands of the subsequent cellular response simulated changes in protein content (green spectrum). The simulated spectra and responses were used to create a dataset which was analyzed using PLS and MCR-ALS. Figure

9(e) depicts the actual incubation time values versus PLS predicted values and shows that the model was skewed due to the lack of linearity of the response with the time. In addition, the regression vector (See Figure 9(f)) showed the bands associated to the incubation time, but it was not able to discriminate between the two responses included in the simulation.

In contrast, MCR-ALS was able to identify the different evolution profiles for both responses in the estimated concentration matrix (see Figure 9(c)). The spectra matrix (see Figure 4.3(d)) showed the computed spectra of the uptake (blue), which included bands of the drug and bands of the binding, and spectra of the cell response (green). In both cases MCR-ALS was able to successfully predict the spectra and evolution of the responses. Furthermore, kinetic constrains can also be implemented to create hard-soft models which takes into account the pharmacokinetics of the drugs within the model, providing an estimation of the constant rates [68].

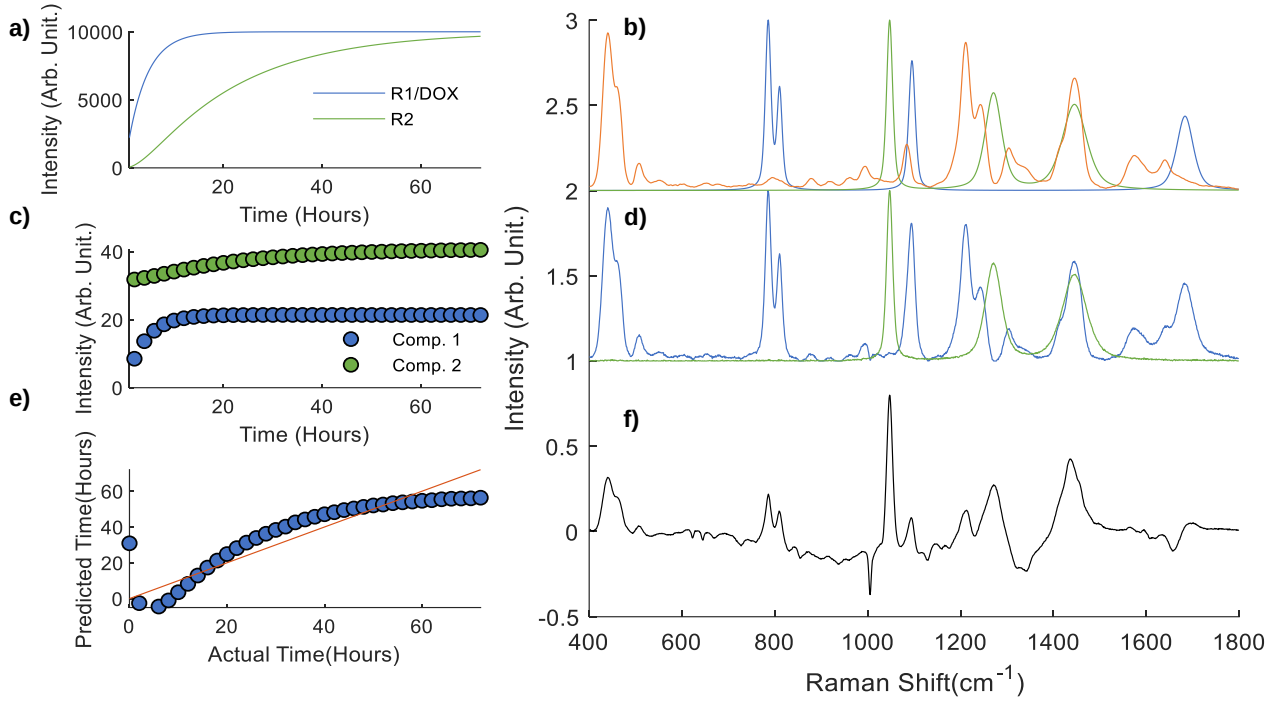

Figure 9. MCR-ALS and PLSR analysis of data. Simulated time evolution (a) and simulated spectral signatures (b) employed onused to build the creation of the Dataset II, simulating dataaset, which simulated the spectral signatures of the disruption (e.g. a drug) the doxorubicin bonding of the disruptionto DNA (Response 1) and subsequent cellular responses (Response 2). Concentration MCR-ALS estimation of the concentration (c) and spectral (d) profiles determined by the MCR-ALS. Actual time values. PLSR actual versus predicted (by values obtained by cross validation) time values (e) and first loading regression vector of the oPLSR orthogonal PLSR (f). (Reprinted from Ref. [66] with permission from Elsevier).

4.3 Data fusion and other applications

RS provides information about a wide range of Raman active compounds with clinical interest. Nevertheless, sometimes it is necessary to measure the same sample with other complementary techniques (e.g. infrared, X-ray fluorescence) to obtain a more comprehensive picture of the phenotype. The result of these multimodal experiments are extended data matrices (or extended hypercubes in the case of images), which are analyzed with data fusion methodologies that perform an integrated analysis of data obtained from different modalities. Considering the increasing number of spectroscopic modalities used nowadays, low-level or high-level data fusion methodologies have emerged as a very suitable tools to acquire information about the sample by investigating relationships among spectral markers from the different methodologies [69].

Since Raman and infrared active modes provide complementary information, their integrated analysis could assist in the assignments of bands. Statistical heterospectroscopy (HSY), for example, has been applied to study the covariance between Raman and IR signals measured across the same sample set [70]. Besides, 2D correlation analysis have also used to create plots of synchronous and asynchronous maps in longitudinal studies [67]. In addition, Raman spectra can be fused from data from other molecular techniques such as matrix-assisted laser desorption/ionization (MALDI) [71] or with elemental techniques such as Laser-induced breakdown spectroscopy (LIBS) [72].

A more troublesome task is the data fusion of hyperspectral images of cells and tissues, which requires accurate registration of the data cubes obtained by the different modalities. This

can be obtained by careful acquisition of the images and by using specific image registration software[73]. On top of that, lateral resolutions of different spectroscopies are generally not the same, requiring the unification of the pixel size before concatenating the images. However, multimodal imaging provides unmatched comprehensive information of the composition of spatial features of the samples. For example, Figure 10 shows an HCA of a multimodal image containing fused IR and Raman spectra of an algae from the genus *Micrasterias*. [70] The average spectra of IR (See Figure 10(d)) and Raman (See Figure 10(b)) show bands associated to the cell wall, algae lobes, chloroplasts and cytoplasm depicted in blue, yellow, violet and orange, respectively. It can be found interesting correlations among IR and Raman bands, such as the high intensity of C-H bands (2800-3000 cm^{-1}) for the lobe class in both modalities. In fact, the fusion of IR and Raman is nowadays gaining popularity considering the recent development of microscopes which combine optical-photothermal IR (oPTIR) and Raman, allowing for simultaneous acquisition of Raman and IR with sub-micron lateral resolution [74].

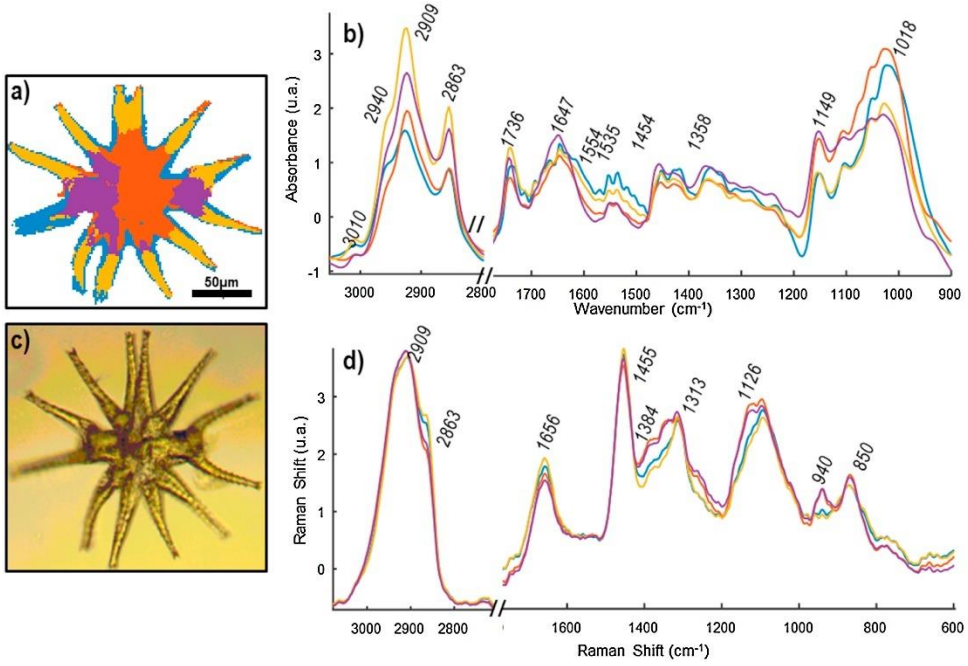

Figure 10. *Micrasterias* fused IR and Raman images analyzed by HCA. (a) Cluster image (b) Class averages spectra for Raman variables. (c) Visible image. (d) Class averages spectra for infrared variables. (Reprinted from Ref. [70] with permission from Elsevier)

References

1 Niaura, G. (2014) Raman Spectroscopy in Analysis of Biomolecules. In Encyclopedia of Analytical Chemistry, pp 1–34, John Wiley & Sons, Ltd.

2 De Gelder, J., De Gussem, K., Vandenabeele, P. and Moens, L. (2007) Reference database of Raman spectra of biological molecules. J. Raman Spectrosc. **38**, 1133–1147.

3 (2015) Appendix F: Infrared and Raman Spectra of Selected Cellular Components. In Modern Vibrational Spectroscopy and Micro-Spectroscopy, pp 399–404, John Wiley & Sons, Ltd.

4 Morais, C. L. M., Lima, K. M. G., Singh, M. and Martin, F. L. (2020) Tutorial: multivariate classification for vibrational spectroscopy in biological samples. Nat. Protoc., Nature Publishing Group **15**, 2143–2162.

5 Ellis, D. I., Brewster, V. L., Dunn, W. B., Allwood, J. W., Golovanov, A. P. and Goodacre, R. (2012) Fingerprinting food: current technologies for the detection of food adulteration and contamination. Chem. Soc. Rev., Royal Soc Chemistry, Cambridge **41**, 5706–5727.

6 Goodacre, R., Graham, D. and Faulds, K. (2018) Recent developments in quantitative SERS: Moving towards absolute quantification. TrAC Trends Anal. Chem. **102**, 359–368.

7 Wang, J., Chen, Q., Belwal, T., Lin, X. and Luo, Z. (2021) Insights into chemometric algorithms for quality attributes and hazards detection in foodstuffs using Raman/surface enhanced Raman spectroscopy. Compr. Rev. Food Sci. Food Saf. **20**, 2476–2507.

8 Abu-Mostafa, Y. S., Magdon-Ismail, M. and Lin, H.-T. (2012) Learning from data: a short course, AMLbook.com, S.l.

9 Houhou, R. and Bocklitz, T. (2021) Trends in artificial intelligence, machine learning, and chemometrics applied to chemical data. Anal. Sci. Adv. **2**, 128–141.

10 Brereton, R. G., Jansen, J., Lopes, J., Marini, F., Pomerantsev, A., Rodionova, O., Roger, J. M., Walczak, B. and Tauler, R. (2018) Chemometrics in analytical chemistry—part II: modeling, validation, and applications. Anal. Bioanal. Chem. **410**, 6691–6704.

11 Bailey, G. F. and Horvat, R. J. (1972) Raman spectroscopic analysis of the cis/trans isomer composition of edible vegetable oils. J. Am. Oil Chem. Soc. **49**, 494–498.

12 Wang, P., Sun, Y., Li, X., Wang, L., Xu, Y., He, L. and Li, G. (2021) Recent advances in dual recognition based surface enhanced Raman scattering for pathogenic bacteria detection: A review. Anal. Chim. Acta **1157**, 338279.

13 Yan, S., Wang, S., Qiu, J., Li, M., Li, D., Xu, D., Li, D. and Liu, Q. (2021) Raman spectroscopy combined with machine learning for rapid detection of food-borne pathogens at the single-cell level. Talanta, Elsevier, Amsterdam **226**, 122195.

14 Adejimi, O., Ignat, T., Sadhasivam, G., Zakin, V., Schmilovitch, Z. and Shapiro, O. (2022) Low-Resolution Raman Spectroscopy for the detection of contaminant species in algal bioreactors. Sci. Total Environ. **809**.

15 Deng, J., Zhang, X., Li, M., Jiang, H. and Chen, Q. (2022) Feasibility study on Raman spectra-based deep learning models for monitoring the contamination degree and level of aflatoxin B-1 in edible oil. Microchem. J. **180**.
16 Guo, Z., Wang, M., Wu, J., Tao, F., Chen, Q., Wang, Q., Ouyang, Q., Shi, J. and Zou, X. (2019) Quantitative assessment of zearalenone in maize using multivariate algorithms coupled to Raman spectroscopy. Food Chem. **286**, 282–288.
17 Li, H., Geng, W., Hassan, M. M., Zuo, M., Wei, W., Wu, X., Ouyang, Q. and Chen, Q. (2021) Rapid detection of chloramphenicol in food using SERS flexible sensor coupled artificial intelligent tools. Food Control **128**, 108186.
18 Lee, K., Yarbrough, D., Kozman, M., Herrman, T., Park, J., Wang, R. and Kurouski, D. (2020) Rapid detection and prediction of chlortetracycline and oxytetracycline in animal feed using surface-enhanced Raman spectroscopy (SERS). Food Control **114**.
19 Pan, H., Ahmad, W., Jiao, T., Zhu, A., Ouyang, Q. and Chen, Q. (2022) Label-free Au NRs-based SERS coupled with chemometrics for rapid quantitative detection of thiabendazole residues in citrus. Food Chem. **375**, 131681.
20 Guo, Z., Barimah, A., Guo, C., Agyekum, A., Annavaram, V., El-Seedi, H., Zou, X. and Chen, Q. (2020) Chemometrics coupled 4-Aminothiophenol labelled Ag-Au alloy SERS off-signal nanosensor for quantitative detection of mercury in black tea. Spectrochim. Acta Part -Mol. Biomol. Spectrosc. **242**.
21 Xiao, X., Liu, X., Mei, T., Xu, M., Lu, Z., Dai, H., Pi, F. and Wang, J. (2022) Estimation of contamination level in microplastic-exposed crayfish by laser confocal micro-Raman imaging. Food Chem. **397**, 133844.
22 Popping, B., Buck, N., Bánáti, D., Brereton, P., Gendel, S., Hristozova, N., Chaves, S. M., Saner, S., Spink, J., Willis, C., et al. (2022) Food inauthenticity: Authority activities, guidance for food operators, and mitigation tools. Compr. Rev. Food Sci. Food Saf. 1–36.
23 Valadés, P. S.-P., Paz, M. P. de la, Copa, P. de A., Ribota, O. G. and Borda, I. A. (2003) Toxic oil syndrome: Survival in the whole cohort between 1981 and 1995. J. Clin. Epidemiol. **56**, 701–708.
24 Sharma, K. and Paradakar, M. (2010) The melamine adulteration scandal. Food Secur. **2**, 97–107.
25 Xu, Y., Zhong, P., Jiang, A., Shen, X., Li, X., Xu, Z., Shen, Y., Sun, Y. and Lei, H. (2020) Raman spectroscopy coupled with chemometrics for food authentication: A review. Trac-Trends Anal. Chem., Elsevier Sci Ltd, Oxford **131**, 116017.
26 Hu, R., He, T., Zhang, Z., Yang, Y. and Liu, M. (2019) Safety analysis of edible oil products via Raman spectroscopy. Talanta **191**, 324–332.
27 Varnasseri, M., Xu, Y. and Goodacre, R. (2022) Rapid detection and quantification of the adulteration of orange juice with grapefruit juice using handheld Raman spectroscopy and multivariate analysis. Anal. Methods, The Royal Society of Chemistry **14**, 1663–1670.
28 Joshi, R., Lohumi, S., Joshi, R., Kim, M. S., Qin, J., Baek, I. and Cho, B.-K. (2020) Raman spectral analysis for non-invasive detection of external and internal parameters of fake eggs. Sens. Actuators B Chem. **303**, 127243.
29 Monago-Maraña, O., Eskildsen, C. E., Afseth, N. K., Galeano-Díaz, T., Muñoz de la Peña, A. and Wold, J. P. (2019) Non-destructive Raman spectroscopy as a tool for measuring ASTA color values and Sudan I content in paprika powder. Food Chem. **274**, 187–193.
30 Li, Y., Fang, T., Zhu, S., Huang, F., Chen, Z. and Wang, Y. (2018) Detection of olive oil adulteration with waste cooking oil via Raman spectroscopy combined with iPLS and SiPLS. Spectrochim. Acta. A. Mol. Biomol. Spectrosc. **189**, 37–43.
31 Karunathilaka, S. R., Yakes, B. J., He, K., Brückner, L. and Mossoba, M. M. (2018) First use of handheld Raman spectroscopic devices and on-board chemometric analysis for the detection of milk powder adulteration. Food Control **92**, 137–146.

32 Perez-Guaita, D., Richardson, Z., Rajendra, A., Byrne, H. J. and Wood, B. (2021) From bench to worktop: Rapid evaluation of nutritional parameters in liquid foodstuffs by IR spectroscopy. Food Chem. **365**, 130442.
33 Silva, M. G., de Paula, I. L., Stephani, R., Edwards, H. G. M. and de Oliveira, L. F. C. (2021) Raman spectroscopy in the quality analysis of dairy products: A literature review. J. Raman Spectrosc. **52**, 2444–2478.
34 Węglińska, M., Szostak, R., Kita, A., Nemś, A. and Mazurek, S. (2020) Determination of nutritional parameters of bee pollen by Raman and infrared spectroscopy. Talanta **212**, 120790.
35 Wang, X., Esquerre, C., Downey, G., Henihan, L., O'Callaghan, D. and O'Donnell, C. (2018) Assessment of infant formula quality and composition using Vis-NIR, MIR and Raman process analytical technologies. Talanta **183**, 320–328.
36 Pezzotti, G., Zhu, W., Chikaguchi, H., Marin, E., Boschetto, F., Masumura, T., Sato, Y.-I. and Nakazaki, T. (2021) Raman Molecular Fingerprints of Rice Nutritional Quality and the Concept of Raman Barcode. Front. Nutr. **8**.
37 Yazgan, N. N., Genis, H. E., Bulat, T., Topcu, A., Durna, S., Yetisemiyen, A. and Boyaci, I. H. (2020) Discrimination of milk species using Raman spectroscopy coupled with partial least squares discriminant analysis in raw and pasteurized milk. J. Sci. Food Agric. **100**, 4756–4765.
38 Ralbovsky, N. M. and Lednev, I. K. (2020) Towards development of a novel universal medical diagnostic method: Raman spectroscopy and machine learning. Chem. Soc. Rev., The Royal Society of Chemistry **49**, 7428–7453.
39 Mancini, R. S. N., Sabaine, A. E., Castro, C. E., Carnielli, J. B. T., Dietze, R., de Oliveira, V. L., Lanfredi, A. J. C., Kubota, L. T., Mamián-López, M. B. and Alves, W. A. (2022) Development and Validation of a SERS-Based Serological Test Combined with PLS-DA Method for Leishmaniasis Detection. ACS Appl. Electron. Mater., American Chemical Society **4**, 3997–4006.
40 Webster, J. A., Wuethrich, A., Shanmugasundaram, K. B., Richards, R. S., Zelek, W. M., Shah, A. K., Gordon, L. G., Kendall, B. J., Hartel, G., Morgan, B. P., et al. (2021) Development of EndoScreen Chip, a Microfluidic Pre-Endoscopy Triage Test for Esophageal Adenocarcinoma. Cancers **13**, 2865.
41 Pérez-Guaita, D. and de la Guardia, M. (2023) Introductory statement – PoC special issue – Microchemical journal. Microchem. J. **184**, 108153.
42 Kong, K., Kendall, C., Stone, N. and Notingher, I. (2015) Raman spectroscopy for medical diagnostics — From in-vitro biofluid assays to in-vivo cancer detection. Adv. Drug Deliv. Rev. **89**, 121–134.
43 Shu, C., Yan, H., Zheng, W., Lin, K., James, A., Selvarajan, S., Lim, C. M. and Huang, Z. (2021) Deep Learning-Guided Fiberoptic Raman Spectroscopy Enables Real-Time In Vivo Diagnosis and Assessment of Nasopharyngeal Carcinoma and Post-treatment Efficacy during Endoscopy. Anal. Chem., American Chemical Society **93**, 10898–10906.
44 Huang, W., Shang, Q., Xiao, X., Zhang, H., Gu, Y., Yang, L., Shi, G., Yang, Y., Hu, Y., Yuan, Y., et al. (2022) Raman spectroscopy and machine learning for the classification of esophageal squamous carcinoma. Spectrochim. Acta. A. Mol. Biomol. Spectrosc. **281**, 121654.
45 Kanno, N., Kato, S., Ohkuma, M., Matsui, M., Iwasaki, W. and Shigeto, S. (2021) Machine learning-assisted single-cell Raman fingerprinting for in situ and nondestructive classification of prokaryotes. iScience **24**, 102975.
46 da Silva, A. M., de Siqueira E Oliveira, F. S. A., de Brito, P. L. and Silveira, L. (2018) Spectral model for diagnosis of acute leukemias in whole blood and plasma through Raman spectroscopy. J. Biomed. Opt. **23**, 1–11.
47 Marzec, K. M., Perez-Guaita, D., De Veij, M., McNaughton, D., Baranska, M., Dixon, M. W., Tilley, L. and Wood, B. R. (2014) Red Blood Cells Polarize Green Laser Light Revealing Hemoglobin' s Enhanced Non-Fundamental Raman Modes. ChemPhysChem, WILEY-VCH Verlag Weinheim **15**, 3963–3968.

48 Perez-Guaita, D., Marzec, K. M., Hudson, A., Evans, C., Chernenko, T., Matthäus, C., Miljkovic, M., Diem, M., Heraud, P., Richards, J. S., et al. (2018) Parasites under the Spotlight: Applications of Vibrational Spectroscopy to Malaria Research. Chem. Rev. **118**, 5330–5358.
49 Moisoiu, T., Dragomir, M. P., Iancu, S. D., Schallenberg, S., Birolo, G., Ferrero, G., Burghelea, D., Stefancu, A., Cozan, R. G., Licarete, E., et al. (2022) Combined miRNA and SERS urine liquid biopsy for the point-of-care diagnosis and molecular stratification of bladder cancer. Mol. Med. **28**, 39.
50 Ishigaki, M., Maeda, Y., Taketani, A., Andriana, B. B., Ishihara, R., Wongravee, K., Ozaki, Y. and Sato, H. (2016) Diagnosis of early-stage esophageal cancer by Raman spectroscopy and chemometric techniques. Analyst, The Royal Society of Chemistry **141**, 1027–1033.
51 Zhang, X., Li, H., Tian, X., Chen, C., Su, Y., Li, M., Lv, J., Chen, C. and Lv, X. (2022) Application of spectral small-sample data combined with a method of spectral data augmentation fusion (SDA-Fusion) in cancer diagnosis. Chemom. Intell. Lab. Syst. **231**, 104681.
52 Song, H., Leng, H., Hou, Z., Gao, R., Chen, C., Meng, C., Sun, J., Li, C. and Ma, B. (2022) Grouped-sampling technique to deal with unbalance in Raman spectral data modeling. Photodiagnosis Photodyn. Ther. **40**, 103059.
53 Pérez-Guaita, D., Kuligowski, J., Lendl, B., Wood, B. R. and Quintás, G. (2018) Assessment of discriminant models in infrared imaging using constrained repeated random sampling - Cross validation. Anal. Chim. Acta **1033**, 156–164.
54 Dodo, K., Fujita, K. and Sodeoka, M. (2022) Raman Spectroscopy for Chemical Biology Research. J. Am. Chem. Soc., American Chemical Society **144**, 19651–19667.
55 Taieb, A., Berkovic, G., Haifler, M., Cheshnovsky, O. and Shaked, N. T. (2022) Classification of tissue biopsies by Raman spectroscopy guided by quantitative phase imaging and its application to bladder cancer. J. Biophotonics **15**, e202200009.
56 Fung, A. A. and Shi, L. (2020) Mammalian cell and tissue imaging using Raman and coherent Raman microscopy. Wiley Interdiscip. Rev. Syst. Biol. Med. **12**, e1501.
57 Yosef, H. K., Krauß, S. D., Lechtonen, T., Jütte, H., Tannapfel, A., Käfferlein, H. U., Brüning, T., Roghmann, F., Noldus, J., Mosig, A., et al. (2017) Noninvasive Diagnosis of High-Grade Urothelial Carcinoma in Urine by Raman Spectral Imaging. Anal. Chem., American Chemical Society **89**, 6893–6899.
58 Rohleder, D., Kiefer, W. and Petrich, W. (2004) Quantitative analysis of serum and serum ultrafiltrate by means of Raman spectroscopy. Analyst, The Royal Society of Chemistry **129**, 906–911.
59 Sánchez-Illana, Á., Mayr, F., Cuesta-García, D., Piñeiro-Ramos, J. D., Cantarero, A., Guardia, M. de la, Vento, M., Lendl, B., Quintás, G. and Kuligowski, J. (2018) On-Capillary Surface-Enhanced Raman Spectroscopy: Determination of Glutathione in Whole Blood Microsamples. Anal. Chem. **90**, 9093–9100.
60 Yang, Y., Xu, B., Murray, J., Haverstick, J., Chen, X., Tripp, R. A. and Zhao, Y. (2022) Rapid and quantitative detection of respiratory viruses using surface-enhanced Raman spectroscopy and machine learning. Biosens. Bioelectron. **217**, 114721.
61 Li, J. Q., Dukes, P. V., Lee, W., Sarkis, M. and Vo-Dinh, T. (2022) Machine learning using convolutional neural networks for SERS analysis of biomarkers in medical diagnostics. J. Raman Spectrosc. **53**, 2044–2057.
62 Dybas, J., Wajda, A., Alcicek, F. C., Kaczmarska, M., Bulat, K., Szczesny-Malysiak, E., Martyna, A., Perez-Guaita, D., Sacha, T. and Marzec, K. M. (2022) Label-free testing strategy to evaluate packed red blood cell quality before transfusion to leukemia patients. Sci. Rep., Nature Publishing Group **12**, 21849.
63 Farhane, Z., Nawaz, H., Bonnier, F. and Byrne, H. J. (2018) In vitro label-free screening of chemotherapeutic drugs using Raman microspectroscopy: Towards a new paradigm of spectralomics. J. Biophotonics **11**, e201700258.

64 Farhane, Z., Bonnier, F. and Byrne, H. J. (2017) Monitoring doxorubicin cellular uptake and trafficking using in vitro Raman microspectroscopy: short and long time exposure effects on lung cancer cell lines. Anal. Bioanal. Chem. **409**, 1333–1346.
65 Efeoglu, E., Casey, A. and Byrne, H. J. (2016) In vitro monitoring of time and dose dependent cytotoxicity of aminated nanoparticles using Raman spectroscopy. The Analyst **141**, 5417–5431.
66 Perez-Guaita, D., Quintas, G., Farhane, Z., Tauler, R. and Byrne, H. J. (2020) Data mining Raman microspectroscopic responses of cells to drugs in vitro using multivariate curve resolution-alternating least squares. Talanta, Elsevier **208**, 120386.
67 Perez-Guaita, D., Chrabaszcz, K., Malek, K. and Byrne, H. J. (2020) Multimodal vibrational studies of drug uptake in vitro: Is the whole greater than the sum of their parts? J. Biophotonics, WILEY-VCH Verlag GmbH & Co. KGaA Weinheim **13**, e202000264.
68 Pérez-Guaita, D., Quintás, G., Farhane, Z., Tauler, R. and Byrne, H. J. (2022) Combining Pharmacokinetics and Vibrational Spectroscopy: MCR-ALS Hard-and-Soft Modelling of Drug Uptake In Vitro Using Tailored Kinetic Constraints. Cells, MDPI **11**, 1555.
69 Azam, K. S. F., Ryabchykov, O. and Bocklitz, T. (2022) A Review on Data Fusion of Multidimensional Medical and Biomedical Data. Molecules, Multidisciplinary Digital Publishing Institute **27**, 7448.
70 Perez-Guaita, D., Kochan, K., Martin, M., Andrew, D. W., Heraud, P., Richards, J. S. and Wood, B. R. (2017) Multimodal vibrational imaging of cells. Vib. Spectrosc. **91**, 46–58.
71 Ryabchykov, O., Popp, J. and Bocklitz, T. (2018) Fusion of MALDI Spectrometric Imaging and Raman Spectroscopic Data for the Analysis of Biological Samples. Front. Chem., Frontiers Media Sa, Lausanne **6**, 257.
72 Teng, G., Wang, Q., Cui, X., Wei, K., Xiangli, W. and Chen, G. (2021) Feature-level fusion of laser-induced breakdown spectroscopy and Raman spectroscopy for improving support vector machine in clinical bacteria identification. J. Raman Spectrosc., Wiley, Hoboken **52**, 805–814.
73 Quintás, G., Wood, B. R., Byrne, H. J. and Perez-Guaita, D. (2021) Multiplexed Fourier Transform Infrared and Raman Imaging. In Multiplexed Imaging, pp 299–312, Humana, New York, NY.
74 Spadea, A., Denbigh, J., Lawrence, M. J., Kansiz, M. and Gardner, P. (2021) Analysis of Fixed and Live Single Cells Using Optical Photothermal Infrared with Concomitant Raman Spectroscopy. Anal. Chem., American Chemical Society **93**, 3938–3950.