FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Investigations on chemometric approaches for diagnostic applications utilizing various combinations of spectral and image data types

Dissertation
(Kumulative Dissertation)

zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat.)

vorgelegt dem Rat der Chemisch-Geowissenschaftlichen Fakultät
der Friedrich-Schiller-Universität Jena

von M.Sc. Oleg Ryabchykov
geboren am 6. Juli 1990 in Simferopol, die Ukraine

1. Gutachter: Prof. Dr. Jürgen Popp
Institut für Physikalische Chemie,
Friedrich-Schiller-Universität Jena

2. Gutachter: PD Dr. Thomas Bocklitz
Institut für Physikalische Chemie,
Friedrich-Schiller-Universität Jena

Tag der öffentlichen Verteidigung: 7. August 2019

Mathematical science shows what is. It is the language of unseen relations between things. But to use and apply that language, we must be able fully to appreciate, to feel, to seize the unseen, the unconscious.

Ada Lovelace (1815 – 1852)

Contents

Contents V

List of tables and figures VII

List of Abbreviations IX

1. Introduction 1

1.1. Tissue and cell-based diagnostics 2

1.2. Machine learning 5

1.3. Data fusion 6

1.4. Outline of thesis 9

2. Data fusion for tissue diagnostics 11

2.1. Raman spectroscopy 12

2.2. MALDI spectrometric imaging 18

2.3. Combined data analysis 20

2.4. Treating misaligned and incomplete data 23

3. Cell-based diagnostics 27

3.1. Image-based hemogram 28

3.2. Merging the predictions from morphological and Raman spectroscopic data 32

3.3. Combining scores obtained from Raman data and clinical values 34

4. Summary 39

5. Zusammenfassung 45

Bibliography 51

Publications 57

[P1] Automatization of spike correction in Raman spectra of biological samples 57

[P2] Leukocyte subtypes classification by means of image processing 65

[P3] Raman based molecular imaging and analytics: a magic bullet for biomedical applications!? 75

[P4] Toward food analytics: fast estimation of lycopene and β-carotene content in tomatoes based on surface enhanced Raman spectroscopy (SERS)........ 95

[P5] Raman spectroscopic investigation of the human liver stem cell line HepaRG 105

[P6] Fusion of MALDI spectrometric imaging and Raman spectroscopic data for the analysis of biological samples 115

[P7] UV-Raman spectroscopic identification of fungal spores important for respiratory diseases 127

[P8] Surface enhanced Raman spectroscopy-detection of the uptake of mannose-modified nanoparticles by macrophages in vitro: A model for detection of vulnerable atherosclerotic plaques 135

List of publications and conference contributions i

Acknowledgments iv

Erklärungen v

List of tables and figures

Table 1: SVM prediction of test data.. 32

Table 2: LBOCV prediction of inflammatory conditions by combined CPPLS model... 37

Figure 1: Types of data fusion architecture….. 7

Figure 2: Raman spectroscopy.. 12

Figure 3: Raman spectroscopic data processing workflow.................................. 14

Figure 4: Automated parameter selection...15

Figure 5: MALDI-TOF spectrometry..18

Figure 6: MALDI spectrometric data processing workflow................................ 19

Figure 7: PCA of separate and combined spectral data...................................... 22

Figure 8: PCA loading for separated and combined analysis............................. 23

Figure 9: Missing data imputation as a part of decentralized data fusion......... 25

Figure 10: Leave-one-batch-out performance of PCA-LDA prediction................26

Figure 11: Workflow of multi-module blood analysis..28

Figure 12: Microscopic cell image processing workflow...................................... 29

Figure 13: The visualization of pseudo-Zernike moments and the LDA loadings of a binary classification model..31

Figure 14: PCA-LDA classification between two major types of lymphocytes.. 34

Figure 15: Distributed data fusion workflow for sepsis diagnostics...................36

Figure 16: Scores of a combined CPPLS model and a cross-validated performance of disease severity prediction.. 37

Figure 17: Preprocessing workflows used for various data types........................39

Figure 18: A diagram that demonstrates the difference between data fusion approaches... 40

List of Abbreviations

AUC	area under curve
CCD	charge-coupled device
CPPLS	canonical powered partial least squares
CV	cross-validation
EM-PCA	expectation maximization principal component analysis
H&E	hematoxylin and eosin
HCC	hepatocellular carcinoma
ICA	independent component analysis
LBOCV	leave-batch-out-cross-validation
LDA	linear discriminant analysis
MALDI	matrix assisted laser desorption/ionization
MALDI-MS	MALDI mass spectrometry
MALDI-TOF	matrix assisted laser desorption/ionization time-of-flight
MCR-ALS	multivariate curve resolution alternating least squares
MS	mass spectrometric
NA	not available (missing value)
NMF	non-negative matrix factorization
PC	principal component
PC3	third principal component
PCA	principal component analysis

PLS	partial least squares
PZI	pseudo-Zernike invariants
PZM	pseudo-Zernike moments
ROC	receiver operating characteristic
RMS	root mean square
SIRS	systemic inflammatory response syndrome
SNIP	sensitive nonlinear iterative peak
SVM	support vector machines
TIC	total ion count
TOF	time-of-flight
WBC	white blood cells

1. Introduction

The diagnostics of diseases is one of the main challenges in modern medicine. Classical approaches such as interviewing a patient, screening body temperature, or measuring blood pressure provide valuable information about patients to physicians. These tests can be performed as a part of a general medical examination and may reveal a patient's state of health. More specific diagnostics require deeper insights into the patient's condition. In many cases, this additional information is obtained by means of anatomic and clinical pathology.

Anatomic pathology is associated with examination of surgical samples by trained pathologists. Clinical pathology is the discipline that focuses on the investigation and diagnostics of diseases by means of various routine laboratory tests, such as blood cell count, throat cultures, or urinalysis. One of the subsections of clinical pathology is chemical pathology, also called clinical chemistry. In contrast to other specialties of clinical pathology, clinical chemistry is focused on the measurement of concentrations of specific chemical substances in body fluids, cells and tissues. [1]

All the above-mentioned disciplines perform tests on samples taken from patients. These disciplines aim to improve the understanding of diseases. Deeper understanding can increase the effectiveness of diagnostics and lead to better patient outcomes. Within this work a combination of these disciplines, known as general pathology, will be referred as pathology.

The pathological investigation of tissues and cells reveals changes, which can be used as markers of a given disease. If some diseases are suspected due to the symptoms or the results of other tests, the final diagnostics in clinical practice are often done or confirmed by pathologists. To assist pathologists in their demanding work, a large number of physicians and scientists in the fields of biomedicine, bio-imaging, and physics are working together to develop new tools and to improve our understanding of diseases.

1.1. Tissue and cell-based diagnostics

Among various branches of pathology, we can find two major areas of application: the investigation of cells (cytopathology) and tissues (histopathology). Cytopathology mainly focuses on subcellular structures and the chemical composition of cells. It can provide additional information on diseases such as the presence of specific cell types, a shift in the cell count, or abnormal cell morphology. These factors can reveal details on a patient's condition to the physician. These details can be useful for infection and inflammation detection. On the other hand, histopathology is an investigation of whole tissue samples. It focuses on distinguishing specific structures within tissues and detecting various disorders related to specific organs. For such an investigation, a tissue sample taken from a specific organ should be analyzed. One of the typical examples of histopathological applications in clinical practice is the diagnostics of cancer.

A common approach for disease diagnostics in histopathology [2] and cytopathology [3] is the visual inspection of stained samples using an optical microscope. This approach is the "gold standard" for many diagnostic applications and for differentiation between cell types or tissue types. To allow such diagnostics, staining is utilized prior to the visual inspection of the sample. Staining reveals details of tissue morphology and subcellular structures [4]. After staining, the pathologist can assign tissue areas or cells within the field of view of the microscopic image.

The information obtained from white-light microscopy and staining techniques is limited for both conventional and immunohistochemical staining approaches. Conventional staining highlights only the distribution of broad classes of molecules, rather than specific molecules. The visualization of specific molecules is possible with immunohistochemical staining, but these staining techniques are costly, labor-some and complex. In contrast to the dramatic improvement in microscopic techniques, some conventional staining techniques remain unchanged for many decades or, as in the case of Hematoxylin and Eosin (H&E) stain, for about a century [5]. Even though these conventional staining techniques are quite old, they are widely applied because of their low cost and high diagnostic accuracy. H&E staining, as well as Kimura staining [6], are widely used in tissue and cell analyses. Unfortunately, these staining approaches often require manual investigation of stained samples, which is labor intensive and leads to an increase in the total analysis time. In many diagnostic applications, manual inspection of stained samples remains the “gold standard”.

Prior to sample preparation and staining, the tissue or cells need to be extracted from the human body and then reagents need to be added. Therefore, this procedure cannot be performed *in-vivo* [7]. On-site analysis of samples during surgery is also challenging and, if the analysis is performed manually, it may be impossible within the required time. Intra-operative sample inspection could be of critical importance since the additional information can increase efficiency during surgery. Even higher impact can be achieved by *in-vivo* analysis, because a precise detection of cancerous tissue margins during surgery decreases the possibility of local cancer re-occurrence. This low cancer re-occurrence results due to the complete removal of cancerous tissue [8] and preservation the surrounding tissue. These aspects are especially crucial in brain tumor surgery [9]. Therefore, stained frozen sections are analyzed during brain and head&neck tumor removal surgeries. Unfortunately, a certain trade-off between the required time and the quality of the diagnostics often needs to be made when staining approaches are applied. One strategy to achieve a better trade-off is to automate the analysis. Another strategy would be applying different measurements techniques in parallel to increase information content of the collected data. Furthermore, the development of optical probes may make *in-vivo* spectroscopic analysis possible in routine diagnostic.

The informational content used for diagnostics can be increased by applying optical measurement techniques, sensitive to chemical composition of the sample [2]. Due to the availability of other measurement techniques as alternatives to optical microscopy, additional information can be extracted, and a deeper understanding of tissue and cell structures can be obtained. In biomedical research tasks, various fluorescence, dark-field, and phase-contrast microscopic techniques, as well as spectrometric and spectroscopic imaging techniques, are widely used. Unfortunately, spectroscopic techniques are not often employed in routine diagnostic, because enormous amount of multidimensional data is collected. Manual analysis of these data sets is very time consuming, and spectral variations are often too small to be recognized by the naked eye. The mentioned issues make usage of spectroscopic techniques in clinical routine very challenging. Therefore, staining remains a main tool in pathology.

Pathological investigations of stained samples are usually performed in manual mode and require a decision made by pathologists. This decision is made based on a visual inspection of the sample and may be subjective [10]. The pathologist's experience [11] plays an important role in this diagnostic process. Experience is especially crucial in cases where the information content obtained by visual inspection of stained samples do not provide sufficient information for doubtless diagnostics. Extending the amount of information, which is available for pathologist, and representing it in a simple interpretable form, may improve efficiency of the diagnostics. To provide more information on the sample's chemical composition to pathologists and physicians, new spectroscopic tool may be introduced to the clinical routine.

Along with new measurement techniques, software tools for the automated analysis of the obtained spectral and image data need to be introduced. This software should extract diagnostic information and automatically convert this information into a simple interpretable form. The common challenges of the automated data analysis include simultaneous usage of multiple data types, large number of observations, and the multidimensionality of data. To overcome these issues, statistics and machine learning can be used.

Machine learning methods make it possible to obtain robust and unbiased results [12], but the data needs be standardized prior to the application of the machine learning methods [P3]. To standardize the data, a preprocessing pipeline must be established, which is specific to a given task and the measured data. At best, the data processing needs to be automated to decrease analysis time and avoid the introduction of a human bias.

Besides automation and optimization of the data processing pipeline, a researcher's expertise and a set of preliminary studies need to be employed to find an optimal measurement technique for every analytical task. For simple tasks, when the difference between the investigated groups among the samples is large and well understood, a suitable measurement technique can be easily chosen. For more complex diagnostic tasks, a single measurement technique may not provide all the required information. To investigate the complex samples to their full extent, multiple analytical approaches are often required.

Employing multiple measurement techniques instead of one single technique may increase the robustness of the analysis and the reliability of the diagnostics. Of a wide variety of spectroscopic and imaging techniques that are sensitive to and selective for different substances, only a few were considered within the frame of this work. Multiple data types provided by these techniques were merged for combined analyses by means of introducing various data fusion schemes in the data processing pipelines. For each demonstrated example, the data preprocessing, data fusion, and analysis by means of machine learning were implemented and adjusted to improve the efficiency of analysis.

1.2. Machine learning

As stated above, machine learning methods are needed to automate the analysis of large datasets. Automated data analysis in histological diagnostics makes it possible to overcome the bias of manual analysis [13] and the subjective decisions based on visual inspection of stained samples [14]. Additionally, an introduction of machine learning methods in clinical routine would improve the robustness of the diagnostics and prediction [15]. To efficiently apply machine learning methods, well-standardized data is needed. For a proper standardization, an automated data processing pipeline should be developed and optimized for each specific machine learning task.

Machine learning approaches can be either unsupervised or supervised. Unsupervised machine learning does not require additional metadata, such as labels, for establishing the model. The unsupervised machine learning methods reveal the common patterns and variations within the data. Dimension reduction and clustering are the common applications of unsupervised machine learning methods. On the other hand, supervised machine learning approaches, such as classification and regression, require labels ore other metadata to train the model. For instance, in diagnostic applications, sample types or the severity of a patient's condition can be predicted using supervised machine learning methods. Examples of supervised machine learning methods for multivariate data include artificial neural networks, linear discriminant analysis (LDA), and support vector machines (SVM). These methods are often used in a combination with dimension reduction techniques, such as principal component analysis (PCA) [P3]. Alternatively, supervised machine learning methods that do not require prior dimension reduction, such as partial least squares regression (PLS) [16], can be employed for regression, classification, and supervised unmixing or dimension reduction.

Proper implementation of the regression and classification methods described can provide the reliable prediction of diagnostic values. For this automated implementation of diagnostics, the data needs to be standardized using several preprocessing steps. Even images obtained by optical or holographic microscopy should be preprocessed before utilizing machine learning, chemometric, or classical statistical methods. Standardization is especially crucial for spectral data. Typical preprocessing routines for spectroscopic and spectrometric data usually include calibration [17, 18] and background correction. Another step that is often required prior to constructing models by means of machine learning is dimension reduction or feature extraction. This step is necessary to avoid model overfitting and to improve the performance of the diagnostic model on independent data.

1.3. Data fusion

Despite advances in measurement techniques and data analysis methods, a single measurement technique may not be suitable for investigation of the sample from different perspectives. To obtain a broader overview and improve the diagnostics, the data from available measurement techniques should be combined [19]. This approach of combining data from different sources is referred to as data fusion.

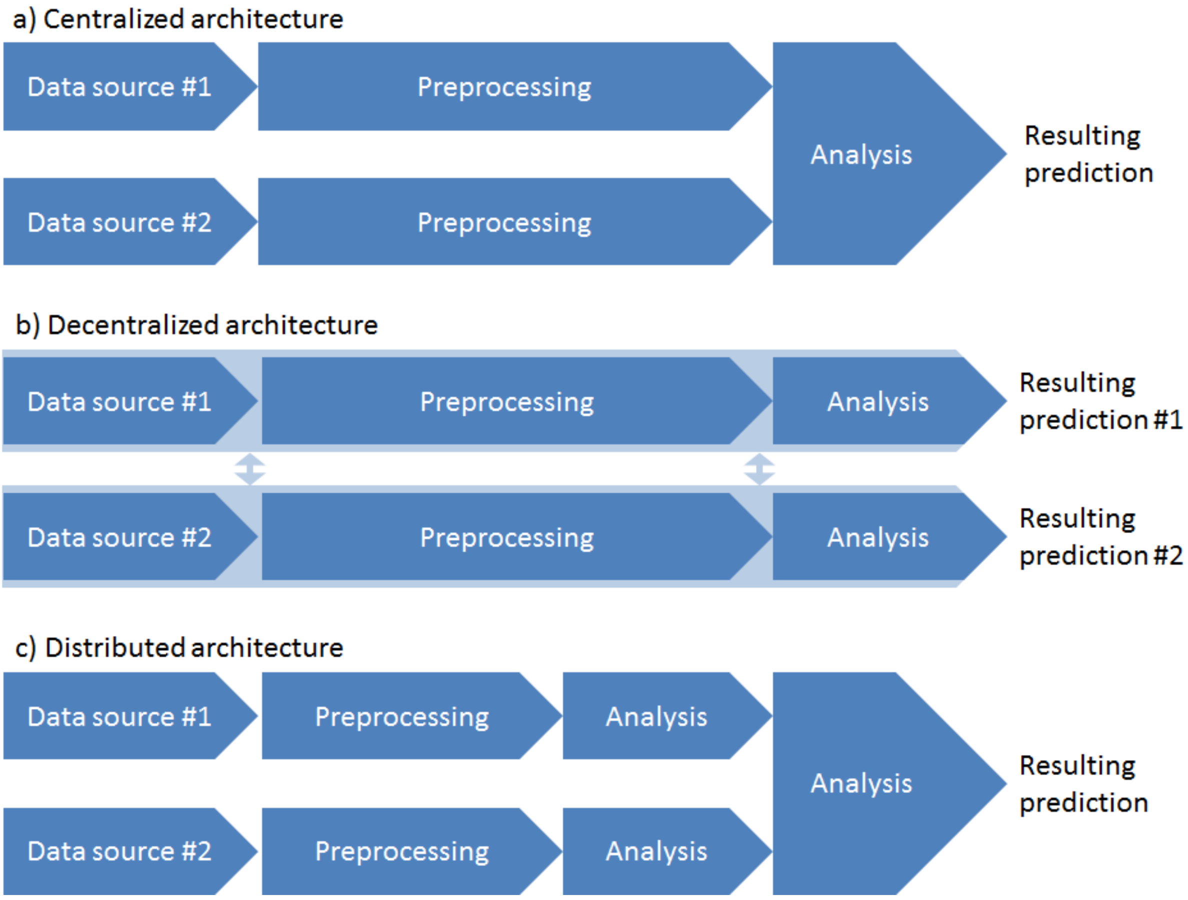


Figure 1: Types of data fusion architecture. Depending on the task, different fusion centers can be chosen within the processing workflow [20]. Centralized architecture (a) has a data fusion center directly after preprocessing. Decentralized architecture (b) allows interactions between data types at different stages of processing and may have more than one fusion center. In contrast to other approaches, distributed architecture (c) is aimed at performing data fusion during the last step of the analysis by combining the results obtained from the analysis of data from different sources.

Data fusion can be categorized in different ways [20]. In this work, the data fusion schemes are categorized based on the type of architecture (see Figure 1). When the data from different sources is combined using centralized or low-level architecture (see Figure 1a), the data fusion is performed directly after preprocessing and then the combined data analyzed together. In a decentralized architecture (see Figure 1b), the data fusion may be performed on different phases of the data processing pipeline for different data sources. Hierarchical architecture is a common example of the decentralized approach. In this architecture, the regions of interest are determined using one measurement technique, and then another measurement technique is employed to obtain a final prediction. In another type of architecture, referred to as distributed (see Figure 1c) – or high-level –

architecture, a data fusion center is located in the data processing workflow after the analysis for every data type is performed separately. According to the distributed data fusion approach, the predicted values or scores are combined instead of combining the data directly. The selection of data fusion architecture must be based on the analytical concept of the task and on the structure of data obtained by each measurement technique.

Correlated imaging, which uses the multiple measurements technique, is one approach. To utilize this approach, the sample needs to be measured in the imaging mode using different techniques. These measurements can be performed either simultaneously or sequentially. If the measurements are performed sequentially, the imaging data must be aligned pixel by pixel after acquiring the images or scans. One drawback of correlated imaging is that multiple technical and experimental challenges exist [21]. Despite these challenges, a combination of spectroscopic and mass spectrometric mapping techniques demonstrated high potential for biomedical applications. For example, combinations of spectroscopic and mass spectrometric imaging techniques have shown their potential for three-dimensional samples [22] and single cell analysis [23]. Data fusion of co-registered mass spectrometric and Raman spectroscopic data was successfully applied in a wide range of medical and biological applications, such as the investigation of biofilms [24, 25] and differentiation of bacteria [26] and fungi [27]. Correlated imaging using matrix assisted laser desorption/ionization (MALDI) and Raman spectroscopy has also brought advantages [28] in tissue research.

The combination of different measurement techniques can improve the investigation of biomedical samples. In addition, biomedical diagnostics can also be improved by adding clinical information to the analysis. In medical practice, physicians use the patient’s known symptoms, health condition, and history for reliable disease diagnostics. This clinical information can also strengthen the robustness and predictive efficiency of automated diagnostics using machine learning. Unfortunately, clinical information is often poorly structured, and some values may be missing, which leads to the exclusion of this data from automated data analysis. To overcome these issues, data imputation can be implemented within the data processing pipeline. The imputation of missing data allows the estimation of unknown values based on known values. One advantage of this approach is that the model can be trained on a complete set of variables even if some information or measurements are not available for all patients.

1.4. Outline of thesis

In the diagnostics of many diseases, cytopathology and histopathology play an important role [29]. To provide new tools and methods to the pathologist, different measurement techniques can be brought into clinical routine, and machine learning methods can be applied to increase the robustness and speed of the analysis. Nevertheless, a single measurement technique may not be sufficient for obtaining a complete overview of the sample. To overcome this issue, a few measurement techniques can be applied, and the recorded data can be combined by means of data fusion. In this work, the data obtained by various measurement techniques from diverse biological samples have been investigated by means of machine learning methods and data fusion.

As combining different data types increases complexity of data processing workflows, multiple challenges were faced in the combined analysis. In the second chapter, it is shown how to overcome the challenges of combining Raman spectroscopic and MALDI spectrometric data for tissue diagnostics. The chapter describes the measurement techniques, related preprocessing methods, and two different data fusion examples. First, an investigation of brain tissue using low-level data fusion is shown, in which an unbalanced contribution of different data types has been faced. To overcome this issue, two methods for balancing the data contribution in the resulting model were investigated: an identical normalization for both data types and a data weighting approach. In addition, a method for the comparison of different weighting schemes has been suggested. Another example of the data fusion of Raman and MALDI image data is shown for a cancer diagnostic task in which some measurements were missing. To avoid exclusion of incomplete observations, these missing values were computationally estimated during data fusion. This data imputation made a full set of variables available for model training and testing.

The third chapter of this work is focused on a cell-based diagnostic. In this chapter, the Raman spectra and morphological features of cells were used for white blood cell subtype classification. Another demonstrated example of cell-based diagnostics is the combined analysis of Raman spectroscopic data and biomarkers for the detection of sepsis and inflammation severity. These cell-based analyses were motivated by a multi-modal blood diagnostic device that should detect sepsis based on blood count, Raman spectral data, and biomarkers.

After presenting the examples of data fusion and the obtained results, the findings revealed in this work are summarized. The summary in English and German can be found in the fourth and fifth chapters, respectively.

2. Data fusion for tissue diagnostics

Robust clinical diagnostics may require performing specific tests on the sample. If specific biomarkers for diseases are known, then biomarker-specific staining procedures can be employed. However, if a specific biomarker is not known or the complexity of the disease does not allow diagnostics based on just a few biomarkers, then techniques with high chemical sensitivity, such as vibrational spectroscopy [30] or various "-omics" approaches, such as proteomics or metabolomics [31], can be applied.

Among a wide range of measurement techniques, Raman spectroscopic imaging and MALDI mass spectrometric imaging were considered for tissue investigations within this work. Both Raman spectroscopy and MALDI spectrometry were found suitable for biomedical imaging due to their high sensitivity. To gain deeper insight into the chemical composition of the sample, Raman spectroscopic and MALDI spectrometric data were fused and analyzed together. Fundamental differences between the working principles of the measurement techniques provided an overview of the chemical composition of the tissue samples from different perspectives. The working principle of the measurement techniques and the specifics of related data preprocessing are described in the first two sections of this chapter.

Following the description of the measurement techniques, two examples of data fusion approaches for tissue investigations are demonstrated in sections 2.3 and 2.4. In the first example, low-level data fusion was applied for the unsupervised analysis of a mouse brain tissue section. Within this analysis, a data-weighting approach was investigated to balance the contribution of different data types. Another example demonstrates the data fusion of Raman and MALDI spectral data for the detection of cancerous tissue in human liver samples. A high-level data fusion approach was implemented for this task. In this example, the challenge of handling missing data was met since Raman and MALDI measurements were performed independently, and not all regions of interest were measured by both measurement techniques.

2.1. Raman spectroscopy

One of the measurement techniques used for tissue investigations within this work is Raman spectroscopy. This measurement technique provides information on the vibrational states (see Figure 2a) of the molecules present in the sample. Raman spectroscopy is sensitive to the chemical content of the sample, and the preprocessed spectra characterize an overall chemical composition of the sample. Due to its high sensitivity, Raman spectroscopy is convenient for a wide range of biological [32] [P4] [P7] and biomedical [33] [P3] [P5] applications. The main advantages of this measurement technique include the possibility of non-invasive medical diagnostics [14] and real-time monitoring [34]. In combination with confocal microscopy, Raman spectroscopy can be utilized in the imaging mode, providing spectral scans with high spatial resolution [35]. Another advantage of Raman spectroscopy is that due to a relatively simple schematic of a basic Raman set up (see Figure 2b), easy-to-use and cost-effective instruments can be designed [36].

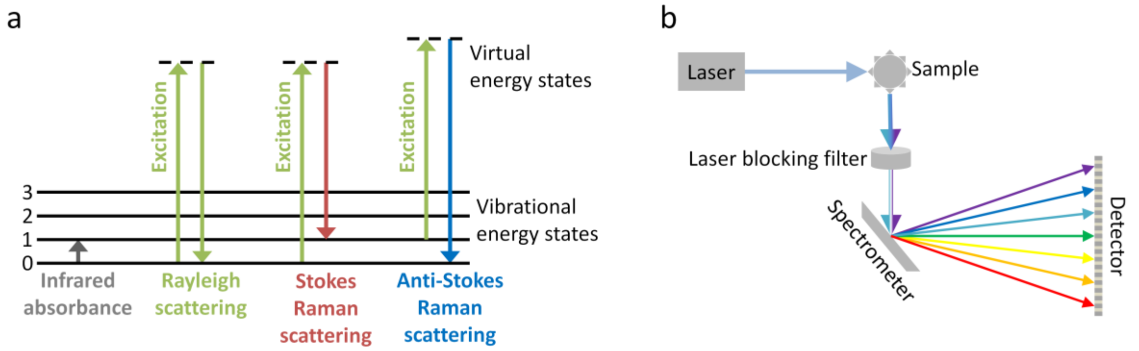


Figure 2: Raman spectroscopy. Energy diagram showing the energy states involved in the Raman scattering processes (a) and a simplified schematic of a Raman spectrometer with the key elements (b) are shown. In the Rayleigh scattering, the scattered photon features the excitation wavelength. On the other hand, the presence of the vibrational energy states gives rise to a small probability that the resulting vibrational state of the molecule differs from the incident vibrational state. Thus, the scattered photon will have an energy that is different from the energy of the excitation photon (a). This difference represents a certain vibrational energy state. By filtering out the frequency of the excitation laser and analyzing the spectrum of the Stokes (or anti-Stokes) Raman scattered light, the information about chemical composition of the sample can be obtained.

One drawback of analyzing biological samples using Raman spectroscopy is that the obtained data is extremely complex. This complexity of the data is further increased due to the corrupting effects that originate even from slight deviations in the sample preparation routine. Another typical example of a sample-related corrupting effect is fluorescence background, which impact is especially significant in the investigation of biological samples. The magnitude of such corrupting effects can be comparable to the level of the signal that represents actual biological information. To suppress these effects and achieve sample-to-sample comparability of such complex data, a set of standardization procedures must be applied during data preprocessing. This set of procedures was organized into a preprocessing pipeline [P3], and specific approaches were developed for treating each corrupting effect (see Figure 3).

Besides corrupting effect related to the samples, a number of non-sample-related corrupting effects can be found. One of these effects is a cosmic ray noise, which is not directly related to the sample or to the settings of the device. Other significant non-sample related corrupting effects are related to the differences in wavenumber axis and intensity responses of the measurement devices. Due to these inter-device differences, it may be extremely challenging to compare the data sets measured on different devices. Furthermore, deviations in the wavenumber axis can be observed even if the compared data sets are measured on the same device, but within a large time span. The preprocessing steps that aim to suppress these non-sample related corrupting effects will be discussed separately since these steps should not be optimized within the model optimization routine. Within the frame of this work, this group of preprocessing steps will be referred to as data pretreatment steps.

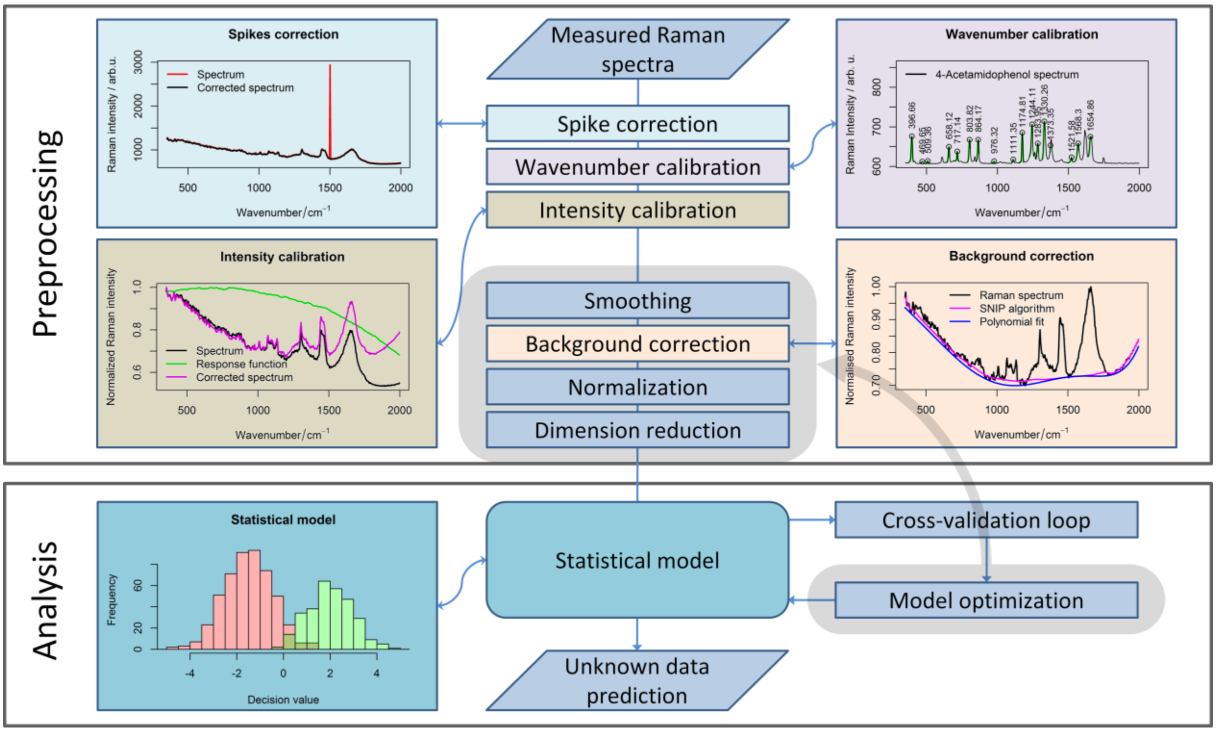


Figure 3: Raman spectroscopic data processing workflow. The illustrated pipeline includes the steps required for the Raman data analysis. It also shows the order and cross-interactions of the processing steps. The first three steps of preprocessing are aimed to eliminate non-sample related variations in data. The subsequent steps of the preprocessing further standardize the data and can be additionally tuned for a specific task along with a model optimization process. Adopted from [37] [P6].

The pretreatment steps for Raman spectroscopic data include cosmic ray spike correction and wavenumber and intensity calibration (see Figure 3). In the studies presented in this work, in-house written algorithms were employed for the pretreatment steps of spike correction [P1] and calibration [17].

The spike correction step aims to eliminate cosmic ray noise, which originates from high-energy particles hitting the charge-coupled device (CCD). This effect appears at random spectral positions and cannot be described with statistical distributions utilized for additive or shot noise effects. As in the imaging mode, many spectra are acquired, and spikes need to be located and removed in a fully automated mode. This can be achieved by setting constant parameters in the algorithm or by defining the parameters through the properties of the data. The cosmic ray noise removal can be more efficient if the spikes are detected considering intensities and sharpness of the peaks in the data subset. Thus, a marker was developed, and spectra with spikes were detected based on the distribution (see Figure 4) of this parameter R [P1]:

$$R_i = \max\left(\frac{sd\left((\Delta S)_{i,1}, \dots, (\Delta S)_{i,j-5}, (\Delta S)_{i,j+5}, \dots, (\Delta S)_{i,m}\right)}{(\Delta S)_{i,j}}\right), \tag{1}$$

where S is a matrix with spectra in rows and ΔS is a matrix with the second derivative of the spectra. In the case of time series measurements, the matrix ΔS can be a two-dimensional Laplacian of the matrix S, which further improves the performance of the algorithm.

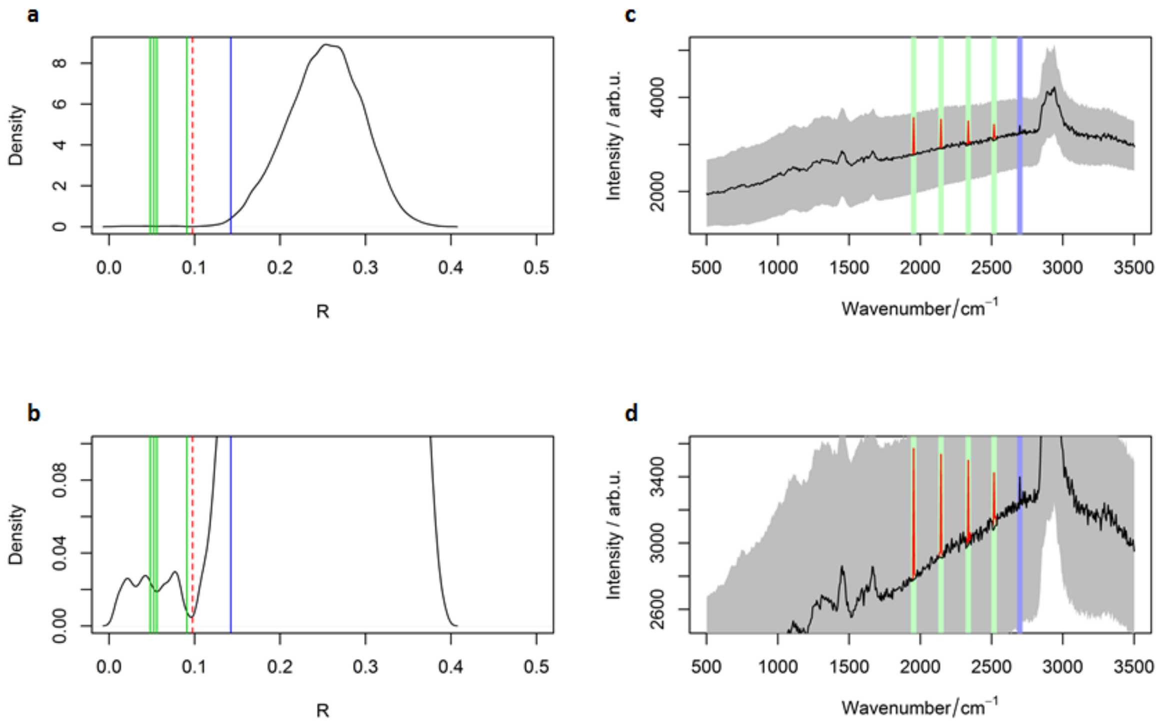


Figure 4: Automated parameter selection. On the left side (a, b), the distribution of the parameter R calculated for all spectra in the scan at the first iteration of spike correction is shown. This parameter is defined by formula(1). The red dashed line depicts a threshold, selected automatically from the distribution. On the right side (c, d), the plot demonstrating a typical spectrum from the data set with artificially induced spikes (red solid line) and the corrected spectrum (black solid line) are shown. The gray area depicts the standard deviation of the spectral values within the data set. The positions of the induced spikes, which were corrected by the proposed method, are depicted in green vertical lines. The undetected spike's position is depicted in blue. The values of the parameter R on different iterations are depicted with the same color codes (a, b). The bottom plots (b, d) depict the same as the upper plots (a, c), but with the range of ordinate that makes it possible to see more details. From the distribution of the parameter R, calculated for each spectrum in the scan, spectra with artifacts can be found in the minor peak of the distribution.

After spike correction, calibration steps should be performed. The wavenumber calibration can be carried out according to a measured wavenumber standard. This calibration is necessary for the comparability of spectra measured within different time frames or with different excitation wavelengths [17]. The purpose of the subsequent intensity calibration is that it should correct for the instrument response function, which is represented as a wavelength dependent function. These calibration procedures are necessary for obtaining comparable peak intensities between devices or conditions [38]. Although the wavenumber calibration is necessary for Raman data pretreatment, the intensity calibration may be skipped if all spectra are measured by a single instrument.

The data pretreatment steps described suppress non-sample-dependent corrupting effects. Nevertheless, for further standardization, Raman spectral data requires additional preprocessing steps such as noise reduction, baseline removal, and spectral normalization [17] [P3]. Although the smoothing of the spectrum is an optional step, it can be applied to decrease the influence of the signal-to-noise ratio on baseline correction and normalization. It may even be used for smoothing spectra prior to calculating the background, which is then subtracted from non-smoothed spectra.

Background removal is one of the most crucial preprocessing steps for the standardization of Raman spectra from biological samples [39]. The frequently-used approach for background fitting is a polynomial fit algorithm [40]. The main disadvantage of this method is the appearance of the so-called Runge's phenomenon, in which the error of the fitting is larger at the edges than at the center of the wavenumber range. Large errors at the edges result in insufficient fluorescence background removal, and a loss of information may occur. Another method applied for Raman spectral analysis is a sensitive nonlinear iterative peak (SNIP) clipping algorithm [41]. The baseline estimation using SNIP is more stable on the edges of the wavenumber region, but an optimal number of iterations may vary depending on a specific signal-to-noise ratio, spectral resolution, wavenumber range, and the width of Raman bands in the data set.

The next step in Raman spectral data preprocessing is normalization. This step allows decreasing variations in signal intensity that occur due to optical focusing during the acquisition of Raman spectra. Normalization and the smoothing step are optional, especially in applications where false-color images are generated by the integration of spectral bands.

Besides standardization steps, feature extraction or dimension reduction may be required for Raman spectroscopic data. These steps are applied if the analysis itself is not performed with methods that include an intrinsic dimension reduction. Since these preprocessing steps can be optimized within the model optimization routine, it is quite difficult to separate preprocessing from the analysis clearly. Thus, dimension reduction is often considered as part of the analysis stage rather than preprocessing and can even be used as an analytical technique. One example of such a procedure is the use of PCA for data analysis and visualization, which is described in section “2.3. Combined data analysis.”

The preprocessing routine described above allows the conversion of raw data into standardized Raman spectra suitable for biological and biomedical applications. These spectra provide complex information on the chemical composition with a high scattering cross section for lipids and proteins. This Raman spectra processing (see Figure 3) was applied for a different Raman spectroscopic task, such as the investigation of human hepatic stem cells [P5] and fungal spores [P7]. The same preprocessing approach was also used to investigate the content of carotenes in tomatoes [P4] and to detect the uptake of nanoparticles by macrophages [P8] using surface-enhanced Raman spectroscopy.

Despite the wide range of applications and high chemical sensitivity, Raman spectroscopic imaging is not the best measurement technique to differentiate proteins from each other or to differentiate lipids from each other, because different molecules of the same class may have very similar Raman spectroscopic signature and the concentrations of specific substances may be too low. Thus, it is beneficial to combine Raman spectroscopy with a measurement technique which provides specific information on the lipid (or protein) content of the sample. One such analyte-specific sensitive measurement technique is MALDI mass spectrometry, which can also be applied in the imaging mode.

2.2. MALDI spectrometric imaging

Matrix-assisted laser desorption/ionization (MALDI) is a soft ionization technique, which is typically coupled with a mass spectrometric detection technique. MALDI mass spectrometry (MALDI-MS) is used for the investigation of a certain class of chemical compounds in a sample. This selectivity is achieved by choosing the laser energy-absorbing matrix that is the most suitable for co-crystallization with a specific class of molecules [42]. After the matrix is applied to the sample, a mixture of the matrix and the analyte is illuminated by a short laser pulse (see Figure 5). The energy of this laser pulse activates the co-crystallization, desorption, and ionization of the matrix and the co-crystallized analyte. Subsequently, the detection of ion masses is performed. One conventional approach to ion mass detection in MALDI-MS is time-of-flight (TOF) detection. According to this detection approach, ions are accelerated by an electric field and fly freely through the field-free area to the detector. As the acceleration depends on the mass of the ions and their charge, the detected time of flight is proportional to the mass-to-charge ratio of the ion. For imaging applications, the process is performed in scanning mode, so spectra are collected from the sample area in a point-by-point manner.

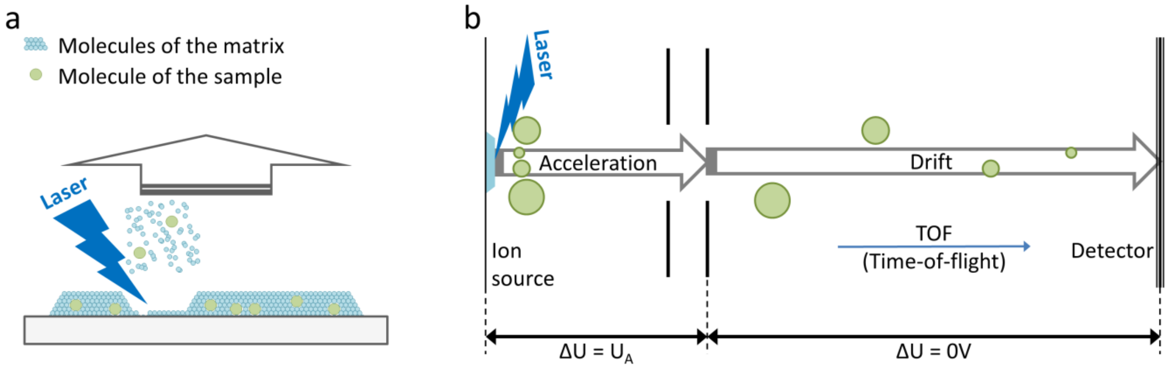


Figure 5: MALDI-TOF spectrometry. A diagram of the MALDI measurement (a) and a simplified schematic of the MALDI-TOF (time-of-flight) spectrometric system (b) with the key elements are shown. The co-crystallized matrix and analyte molecules absorb the energy of a laser pulse. This energy triggers desorption and ionization of the molecules (a). Then, the generated ions are accelerated by an electric field (b). Lighter ions are accelerated to higher velocities compared to heavier ions of the same charge. Thus, the time of flight through the field-free area is related to the mass-to-charge ratio of the ions.

Often MALDI-MS is used to obtain detailed information about the lipid or protein content of biological samples. MALDI imaging has been suggested as a suitable measurement technique in various biological applications including the imaging of kidney biopsies [43], rat brain slices [44], and the subcellular imaging of pancreatic islet cells [45]. Unfortunately, various corrupting effects introduce non-sample related variations in the spectra. The presence of these corrupting effects needs to be considered during data processing to obtain robust results.

By analogy with Raman spectroscopy, the data processing pipeline of MALDI data can be divided into preprocessing and analysis. Although both measurement techniques have similar analysis strategies, the preprocessing shows some significant differences. These differences appear as early as the first steps of the MALDI data processing pipeline (see Figure 6): the spike correction is not needed since particles are detected and not photons. However, calibration, or warping, is extremely crucial for mass spectral data preprocessing. The baseline correction is also needed, but the baseline can be subtracted efficiently without parameter optimization by the sensitive nonlinear iterative peak (SNIP) algorithm.

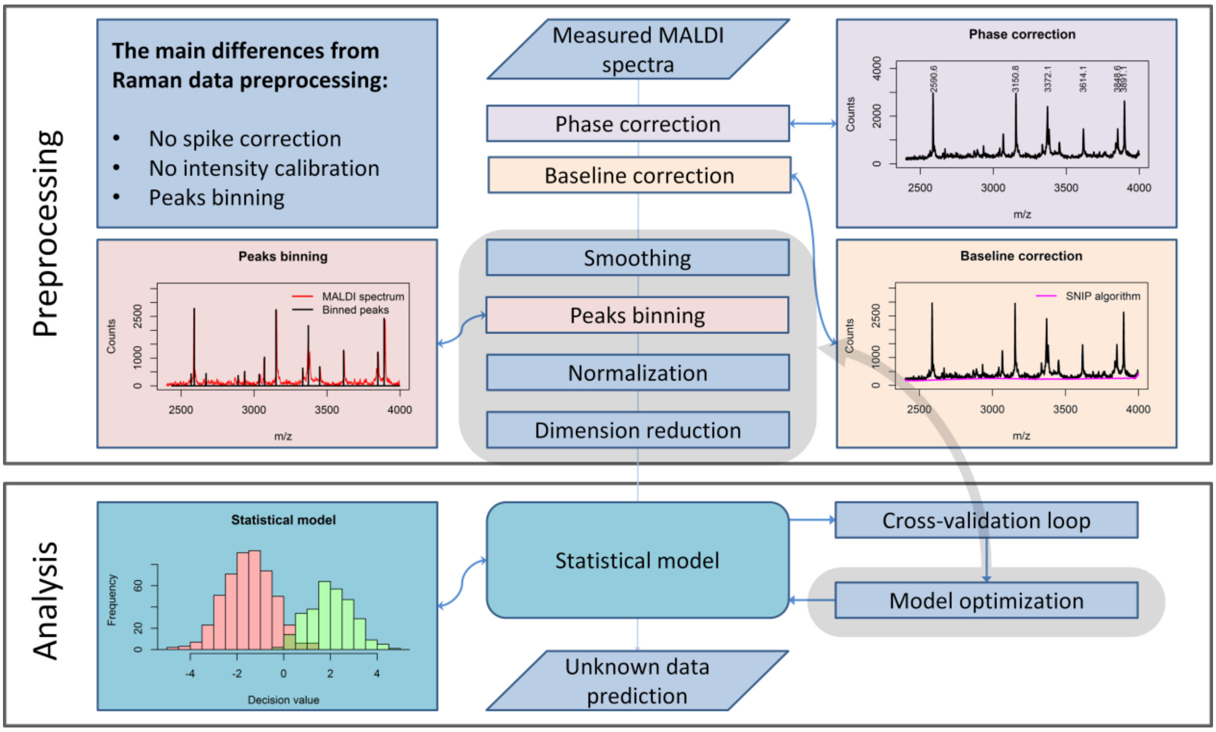


Figure 6: MALDI spectrometric data processing workflow. Although the workflow is divided into preprocessing and analysis parts, some preprocessing steps should be optimized to improve the robustness of the analysis. In addition to the order and interactions between the processing steps of the MALDI processing routine, the main differences in the Raman spectrometric data processing workflow (see Figure 3) are pointed out. Adopted from [37] [P6]

Besides baseline correction, the smoothing of the mass spectrometric data can be useful for improving the performance of peak picking methods; however, due to the high sharpness of peaks, it is important to use smoothing procedures carefully and only when absolutely required. Another drawback of the sharp MALDI peaks is that even after calibration, peaks might not be correctly aligned. To equalize the mass values of nearby peaks, a binning procedure must be applied. The peak binning procedure considers peaks within a certain window of masses as the same. The width of the window should be set indirectly through a parameter called tolerance, which specifies a relationship between the width of the binning window and the mass values. The tolerance parameter is used for peak binning instead of a fixed window size because peak width and mass detection precision differ for light and heavy ions. The peak binning procedure combines the data from different measurements in a single data matrix where every column represents the peak intensity of a specific mass value.

After background removal and peak binning, normalization can be performed. The normalization approaches of the MALDI data correspond to Raman spectroscopic data normalization. The most common normalization approaches include median, total ion count (TIC), and root mean square (RMS) normalization. Moreover, normalization to a spectrum noise level or to a peak value (as well as its square root or logarithm) can be useful in some applications.

The preprocessed MALDI spectra have much lower dimensionality than the raw data but still contain hundreds of variables. Since the data set is represented as a matrix with spectra in rows and variables (mass values) in columns, the dimension reduction for further analysis can be performed by means of component extraction methods (PCA, NMF, ICA, MCR-ALS) or by means of variable selection methods. Variable selection and dimension reduction can also be applied sequentially. One example of such a sequential use of dimension reduction and variable selection methods for MALDI data is shown in the next section.

2.3. Combined data analysis

To determine if a combination of Raman spectral imaging and MALDI mass spectrometric imaging leads to an improved analysis outcome, both measurement techniques were applied on the same mouse brain tissue section. A matrix that is specific for lipids was applied during the MALDI spectrometric investigation of the sample. Before both data types could be analyzed together, the Raman spectra were co-registered to the MALDI measurement grid. As a Raman spectroscopic scan had a higher resolution than a MALDI spectrometric scan, Raman spectra taken from the positions affiliated with a single MALDI measurement were averaged. After performing standard preprocessing for each type of spectra, the data matrices with Raman and MALDI spectra were combined into a single matrix. Each row contained a vector with normalized MALDI counts and normalized Raman intensities related to the same point on the sample. Subsequently, a PCA was performed on the Raman, MALDI, and combined data. As mentioned in the introduction ("1.3 Data fusion"), such a data combination approach that interacts with the preprocessed data directly is referred to as centralized data fusion (see Figure 1a).

Due to the use of a centralized data fusion, the problem of an unbalanced contribution of the two data types in the overall data variance was encountered. To overcome this issue, the PCA using the combined data matrix was performed after introducing a weighting coefficient between the Raman and MALDI data. The coefficient was selected to equalize the sums of the absolute values of the two spectral data matrices. This weighting approach was validated by investigating the cumulative proportion of variance explained by a given number of PCs. For this investigation, the PCA was performed for the data combined using different weighting coefficients. The weighting coefficient, which led to the slowest increase in cumulative variance, was considered optimal for extracting more valuable information (e.g., factors) [P6].

The first three principal components (PC) were visualized for separate and combined data analyses (see Figure 7). The PC-score false-color image obtained from the combined analysis demonstrates a better distinction of spatial features in comparison to both separate data analyses. A closer inspection of the PCA loadings (see Figure 8) showed that the first two PCs did not differ significantly from the combined and separated data analyses, but the third principal component (PC3) changed notably.

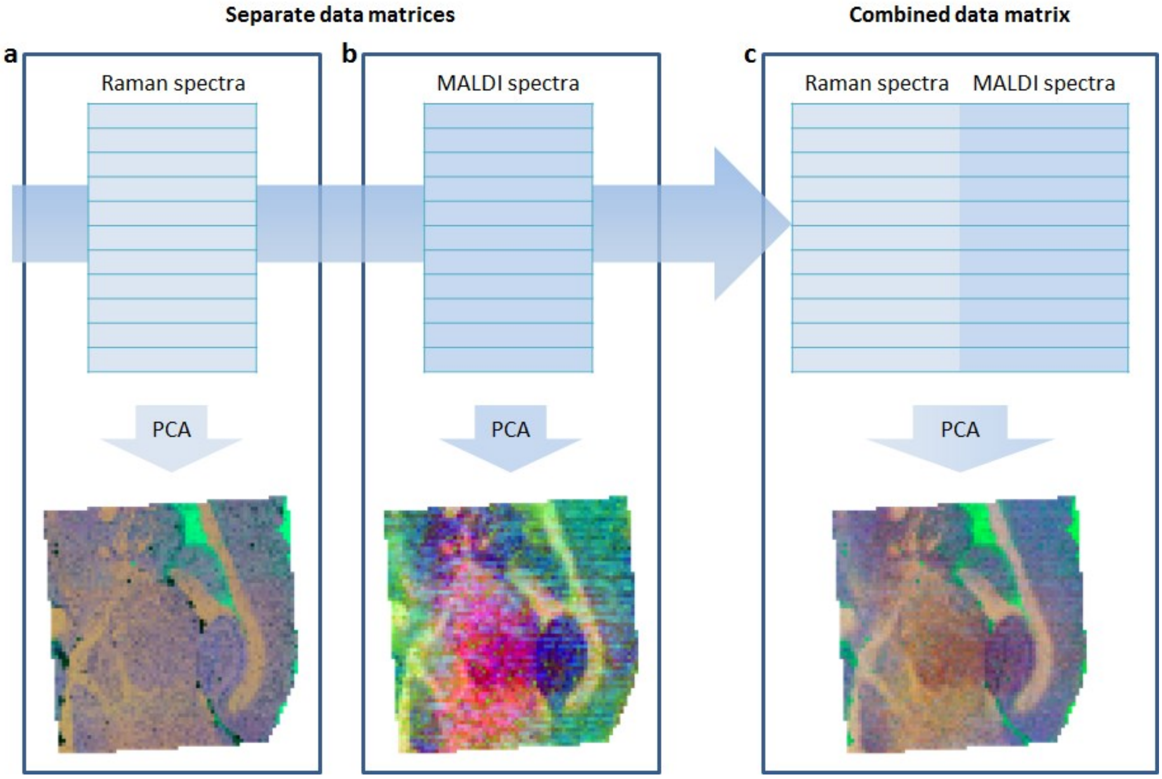


Figure 7: PCA of separate and combined spectral data. Separate analyses were performed utilizing a PCA of preprocessed normalized Raman (a) and MALDI (b) spectra. The combined data matrix (c) was created by merging Raman and MALDI data matrices in a way that each row contained both normalized Raman intensities and normalized MALDI counts. The false color images in (a), (b), and (c) were obtained by visualization of the principal components of Raman, MALDI, and combined data. In these images, the first three principal components are visualized in red, green, and blue, respectively.

Without data fusion, PC3 of the MALDI data contains coefficients with opposite signs related to the isotopes of the same ions. This can be caused by the difference in signal-to-noise ratio between spectra, rather than by the lipid content of the sample. PC3 of the Raman data alone is also hard to interpret. On the other hand, PC3 of the combined data contains MALDI and Raman parts at the same time. It depicts differences in the lipid content, according to the MALDI data, and differences in the protein-to-lipid ratio, according to the Raman data. This additional interpretable feature was revealed, and the data visualization quality was increased by implementing a weighted data fusion for Raman spectroscopic and MALDI spectrometric imaging data [P6].

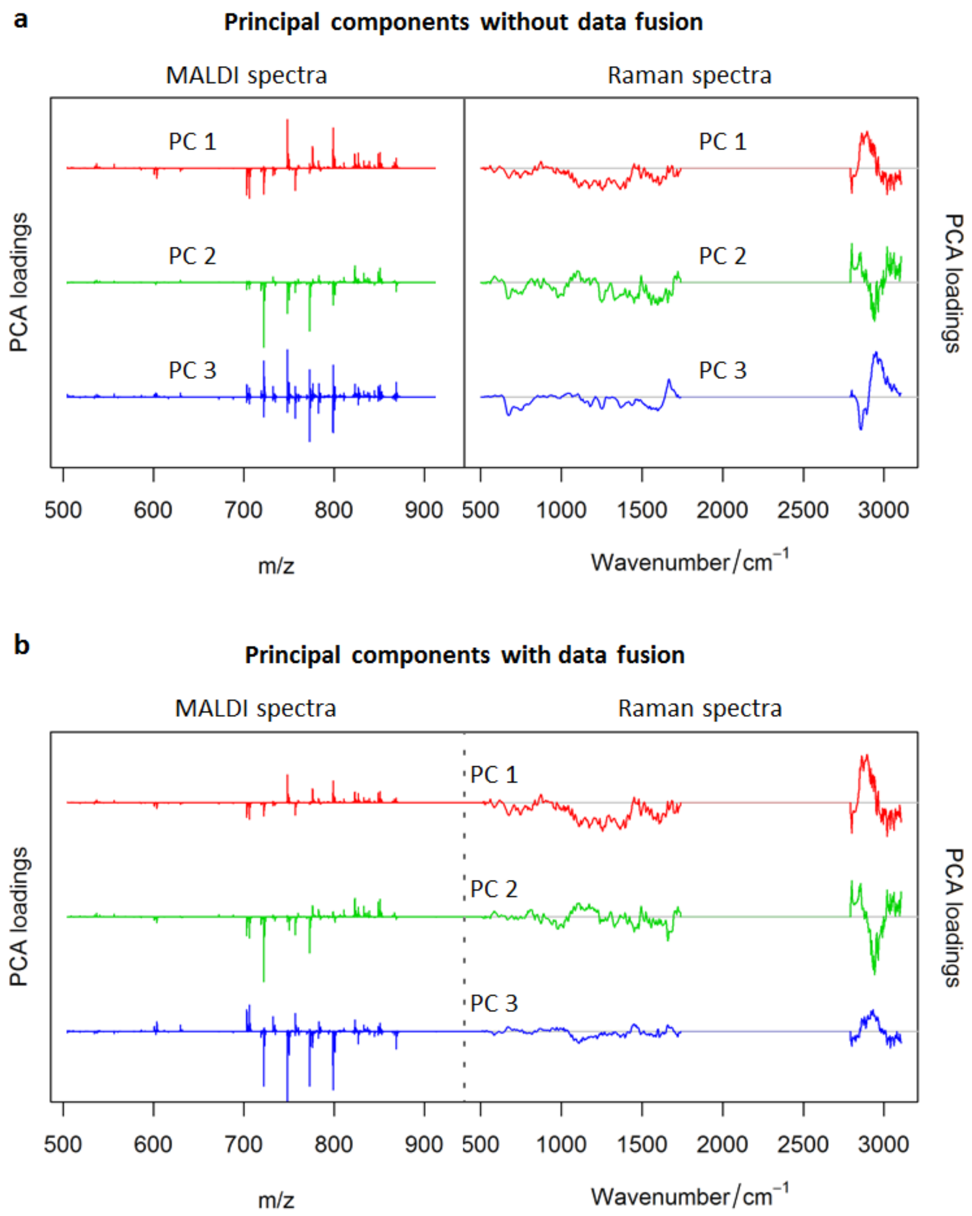


Figure 8: PCA loading for separated (a) and combined (b) analysis. The colors of the spectra are the same as in the composite false-color images of the PC in Figure 7. The significant change in the third principal component can be seen when a PCA is applied to the combined data: the behavior of the isotopes of the same ions is stable for the MALDI part of PC3 (b).

2.4. *Treating misaligned and incomplete data*

In research tasks that deal with multiple measurement techniques, the experiments should be designed to avoid missing values or incomplete observations. This planning increases the robustness and reliability of the data analysis from multiple data sources. Pathologists also rely on multiple diagnostic tools in clinical practice. However, in extreme cases physicians need to perform diagnostics even if not all tests were completed. Unfortunately, dealing with incomplete observations is challenging for machine learning algorithms. This complicates the transfer of a developed diagnostic pipeline to the clinic. Additional complications arise for correlated imaging [21] when the data obtained from several measurement techniques have to be aligned pixel by pixel. This alignment may even be impossible if the spectral imaging by various techniques is performed independently or on different areas of the sample. To overcome this issue, a data processing pipeline that allows analysis of such misaligned incomplete data was constructed.

As an example of data fusion for incomplete data, Raman and MALDI spectral scans of hepatic tissue samples were used. These samples were obtained from patients with hepatocellular carcinoma (HCC), and the spectra were acquired from areas that were characterized as HCC or fibrosis. As the Raman and MALDI measurements were not used in a correlated manner, the direct alignment of data points was not possible. In addition, for some samples only one measurement (Raman or MALDI) was performed. To handle such misaligned data, every area measured by one of the measurement techniques was divided into fifteen subareas. From these subareas, average spectra were extracted. After averaging, fifteen Raman spectra and/or fifteen MALDI spectra were obtained for each area. Subsequently, principal components (PCs) were calculated separately based on the obtained MALDI and Raman spectra.

The number of PCs was optimized separately for Raman and MALDI data by means of a leave-batch-out cross validation (LBOCV) of an LDA model. Next, the PCA scores from two data types were merged into a single matrix. In this matrix, a large portion of the observations had missing data because some areas were investigated only by one of the measurement techniques (see Figure 9). Removing incomplete data from the data set would lead to a significantly decreased dataset size. To avoid data exclusion, the missing data points were imputed using the iterative expectation maximization principal component analysis (EM-PCA) [46]. Finally, an LDA model was built for the combined data with imputed missing values.

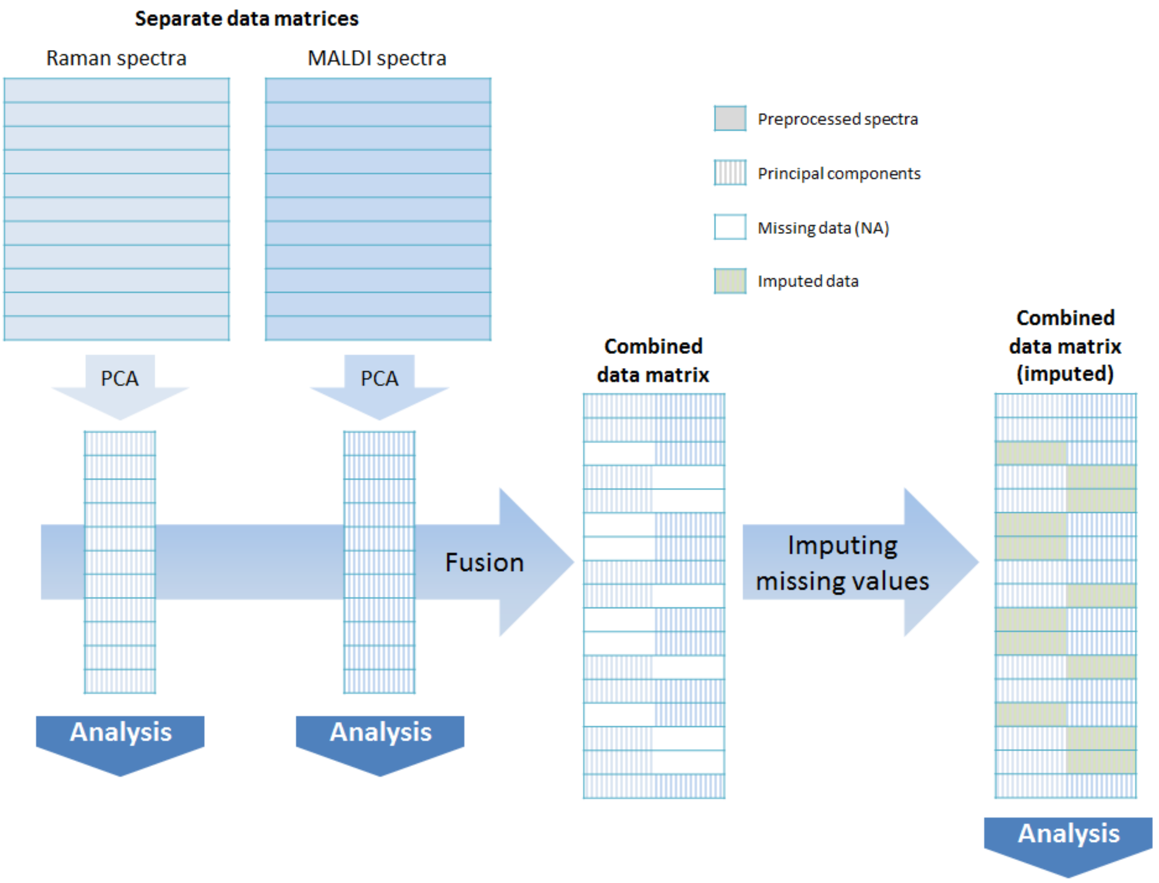
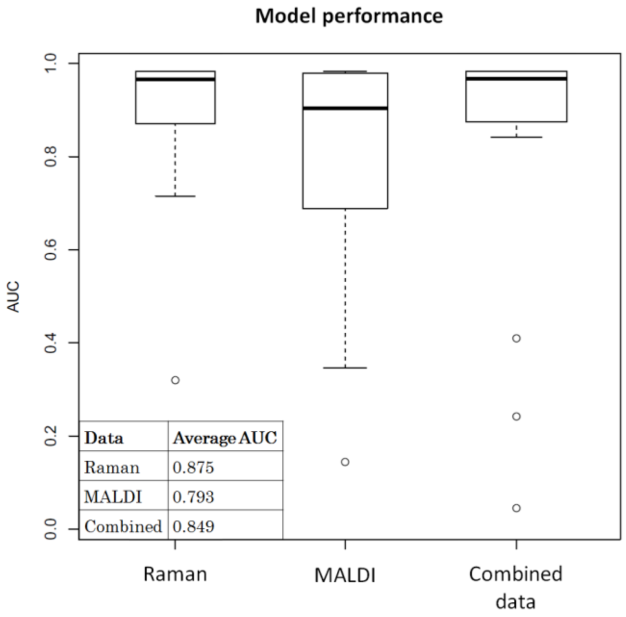


Figure 9: Missing data imputation as a part of decentralized data fusion. Decentralized data fusion allows interaction between different data types and the influence of data from one source to data from another source. After performing data alignment, cells with missing data were revealed and filled using EM-PCA imputation. The prediction for each sample can be achieved according to the type of available data (Raman or MALDI) but also by the model, which includes all the training data, unrelated to the type of data.

To estimate the performance of each LDA model, an LBOCV approach was utilized. Within the LBOCV loop, receiver operating characteristic (ROC) curves were built for the LDA scores of each patient. Areas under ROC curves (AUC) were used as the performance metric. The results of the data cross validation for separate and combined analyses are shown in Figure 10. The results show that different measurement techniques make it possible to achieve separation between two groups with different performance. So, Raman spectroscopic imaging significantly outperforms MALDI spectrometric imaging in HCC detection. The combined data analysis made it possible to achieve the efficiency of Raman spectral data analysis, even though the Raman spectra were not available for all the samples.



Data	Average AUC
Raman	0.875
MALDI	0.793
Combined	0.849

Figure 10: Leave-one-batch-out performance of PCA-LDA prediction. For each patient ROC curves were built for predicted decision values of three different LDA models. The figure shows boxplots for AUC of these LDA models. The Raman data analysis demonstrated better performance than the MALDI imaging data analysis, and the combined data with imputed missing values demonstrated good performance even though Raman spectra were not available for some samples.

3. Cell-based diagnostics

Cell analysis is a routine procedure in clinical diagnostics. Commonly, cell samples are analyzed to diagnose cancer, infectious diseases, and inflammatory conditions. This analysis is performed within a branch of pathology called cytopathology. The information obtained about cell count, cell morphology, and chemical composition can indicate various disorders and provides valuable information about the patient's condition to clinicians.

Cell morphology and chemical composition can be investigated by various measurement techniques to improve understanding of diseases. The techniques applied for this purpose include microscopy, spectroscopic techniques, mass spectrometry, and "-omics" techniques. These measurement techniques produce large amounts of data that cannot be analyzed in manual mode due to the large amount of time needed for manual inspection. To convert the measured multivariate data into a small set of interpretable scores, machine learning methods can be applied.

When machine learning is applied to each type of data separately, a prediction can be made based on each measurement technique. Although each of these predictions may be interpretable by itself, the full set of the results obtained from multiple techniques can be controversial. To overcome this issue and perform a reliable combined analysis of data from multiple sources, the data fusion approaches previously mentioned in "1.3 Data fusion" can be employed.

In this chapter, the use of data fusion for cell-based diagnostics is demonstrated. Such types of data as Raman spectra and microscopic images were employed for WBC subtype identification. In another example, Raman spectra and biomarker data were analyzed in a combined manner for rapid sepsis detection using a minimal amount of the patient's blood [47]. According to the suggested concept, a patient's blood sample should be separated by microfluidics into three parts (see Figure 11), which are further investigated using three different measurement techniques: holographic imaging [48], Raman spectroscopy [49], and fluorescence-based biomarker detection [50]. The suggested workflow promises higher objectivity and shorter analysis time than conventional methods of sepsis diagnostics. When a rapid diagnosis needs to be made, this three-module blood investigation device may become a useful analytical tool for a physician.

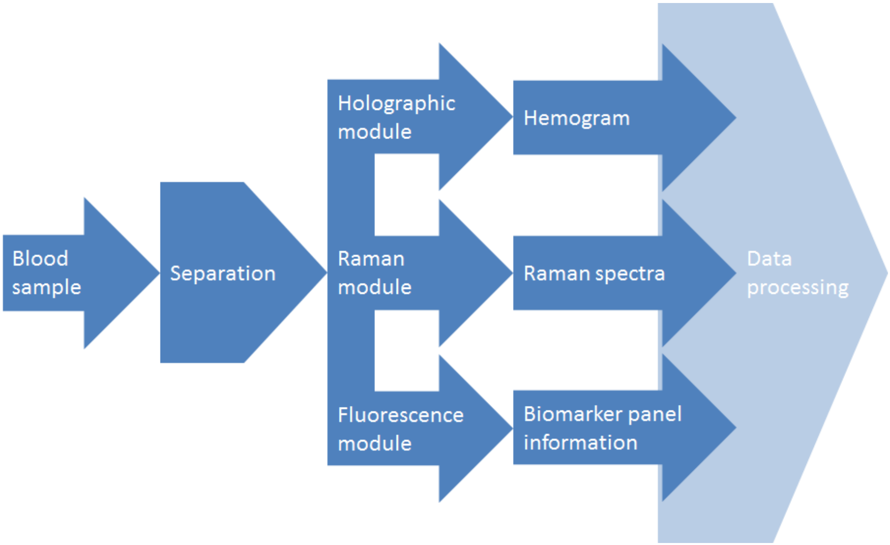


Figure 11: Workflow of multi-module blood analysis. The concept is based on a parallel analysis of a blood sample using three different measurement techniques. A single blood sample is divided inside the device into three subsamples. Subsequently, every subsample is analyzed by different measurement techniques, which provide specific information about the blood sample.

The chapter is organized as follows. Sections 3.1 and 3.2 demonstrate that image-based analysis alone and in combination with Raman spectroscopic data can provide information about the cell subtypes. At the end of this chapter (section 3.3), the improvement of the inflammation severity prediction based on the combined analysis of biomarker data and Raman spectra is shown.

3.1. Image-based hemogram

Despite multiple disadvantages highlighted in the introduction (“1.1 Tissue and cell-based diagnostics”), the visual inspection of microscopic cell images by a pathologist remains the “gold standard” in clinical cytopathology. As an alternative to manual inspection of the stained cell images, automated analysis methods for stained cell image classification can be used. These automated methods are often based on deep learning and convolutional neural networks. A typical data set size for training an image classification model by such deep learning methods is usually estimated above 1000 independent observations per class because these methods tend to overfit on smaller data sets. Since the images must be independent (collected from different patients) and the variations between the measurements are quite high, an enormous number of experiments need to be performed to obtain a training data set. This would be necessary for a stable image-based cell classification by deep learning. On the other hand, with well-standardized data, chemometric approaches can provide fast and objective analysis without large data sets [51]. Digital image processing was proven to allow blood cell analysis [52], specifically for blood cell identification [53], even with a limited amount of training data [54] [P2] and in the presence of low-resolution images [55]. Due to the ability of classic machine learning approaches to construct a robust model with a limited amount of data, they are more suitable for small data sets than deep learning methods. One drawback of classic machine learning methods is that a complex preprocessing and feature extraction pipeline needs to be implemented to obtain image features that can be statistically analyzed (see Figure 12).

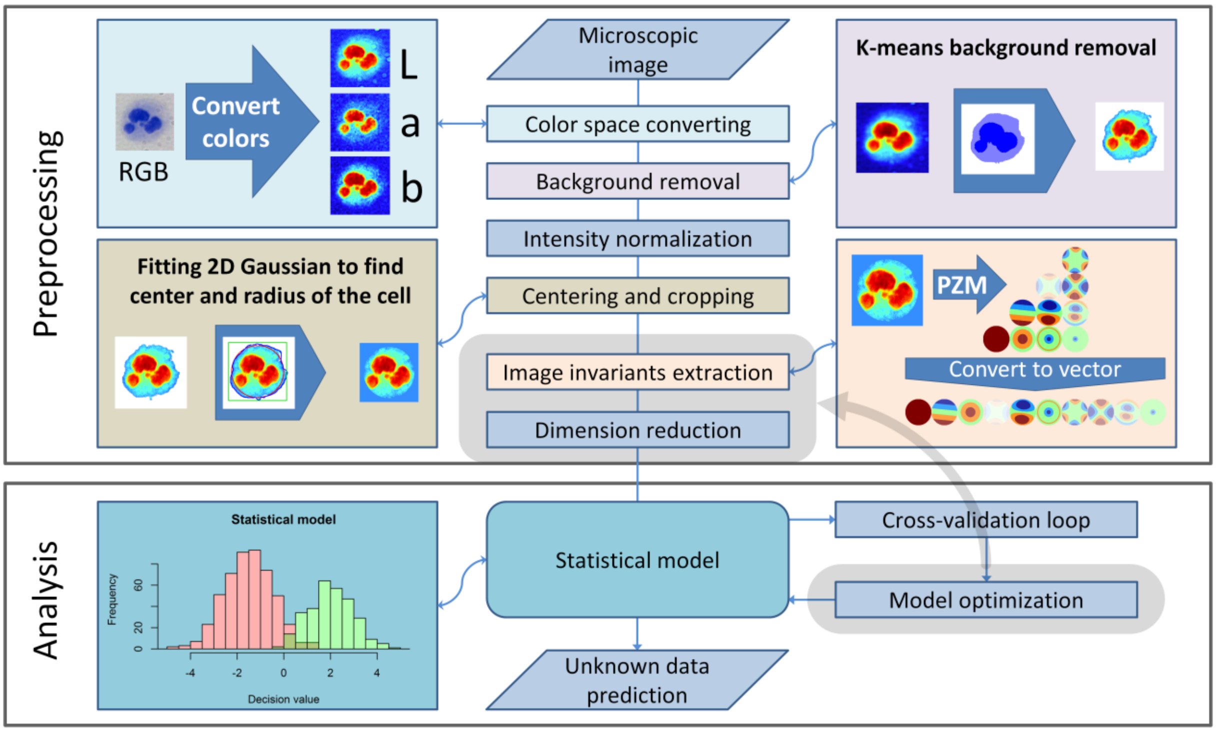


Figure 12: Microscopic cell image processing workflow. The processing pipeline is shown in analogy to the processing pipelines for Raman and MALDI data shown in Figure 3 and Figure 6. Despite the differences in preprocessing, the analysis part and its interaction with the preprocessing part are similar to spectral data processing. The first four preprocessing steps of the workflow pretreat the microscopic images to standardize them, and then the image invariants extraction and dimension reduction are performed. The number of

extracted image invariants and the number of PCs used for the analysis are optimized in the CV loop to obtain an optimized model.

We applied classic machine learning on microscopic images to differentiate between the WBC subtypes. The microscopic images were obtained from cells stained with a Kimura stain that colors the cell nuclei in blue. As the images of Kimura-stained cells are almost monochromatic, the variations of the color shades do not provide additional information. Therefore, only the lightness of the images was used in the further steps of the analysis. An additional reason to analyze monochromatic images was the possibility of transferring the developed image processing pipeline to other types of cell images, such as false color Raman images [56] or digital holographic images, acquired within the three-module sepsis diagnostic setup (see Figure 11).

After extracting the lightness values from the images, the background was removed from each image by a k-means clustering, and each cell image was cropped and centered according to the estimated radius and center of the cell. The radius and center were estimated for every cell by fitting a two-dimensional Gaussian function to the lightness values. Then, the histogram of the lightness values within the cropped cell area was equalized to obtain a uniform distribution for every cell.

In addition to preprocessing, which aims to decrease the variations that are related to the measurement and not to the cell or its subtype, an extraction of spatial morphological features from the images is necessary before analysis. To make the prediction independent from spatial orientation of the cell, these morphological features need to be independent from rotation and mirroring. To fulfill these requirements, pseudo-Zernike moments (PZM) were employed for feature extraction. Although these moments are not rotationally invariant, their absolute values, referred to as pseudo-Zernike invariants (PZI), are independent of mirroring and rotation of the image. Mathematically, the PZM can be expressed in the following way [57]:

$$ {}_{nl} = \frac{\boldsymbol{n+1}}{\boldsymbol{\pi}} \int_{0}^{2\pi} \int_{0}^{1} \boldsymbol{r}\, [\boldsymbol{V}_{\boldsymbol{nl}}(\boldsymbol{r}\cos\boldsymbol{\theta}, \boldsymbol{r}\sin\boldsymbol{\theta})]^{*} \boldsymbol{f}(\boldsymbol{r}\cos\boldsymbol{\theta}, \boldsymbol{r}\sin\boldsymbol{\theta})\, \boldsymbol{dr d\theta}\,. \tag{2} $$

As shown in equation (2), the PZM are defined through an orthogonal set of complex-valued pseudo-Zernike polynomials $\boldsymbol{V_{nl}}$ with order $\boldsymbol{n} = \boldsymbol{0}, \ldots, \infty$ and tion $\boldsymbol{l} = \boldsymbol{0} \leq \boldsymbol{n}$. In an oversimplified way, the order $\boldsymbol{n}$ can be described as spatial frequency along the radial coordinate $\boldsymbol{r}$, and the repetition $\boldsymbol{l}$ can be interpreted as a spatial frequency along the angular coordinate $\boldsymbol{\theta}$.

To visualize the principle of image decomposition into features based on the PZM, the functions described by a single PZM were generated and shown in Figure 13a. The relationship of these features to cell type can be seen from the LDA loadings, which are projected back to the PZI space in Figure 13b. This figure shows that the moments of high repetition (a spatial frequency along the angular coordinate $\boldsymbol{\theta}$) are related to neutrophils and the moments with low repetition are characteristic for lymphocytes.

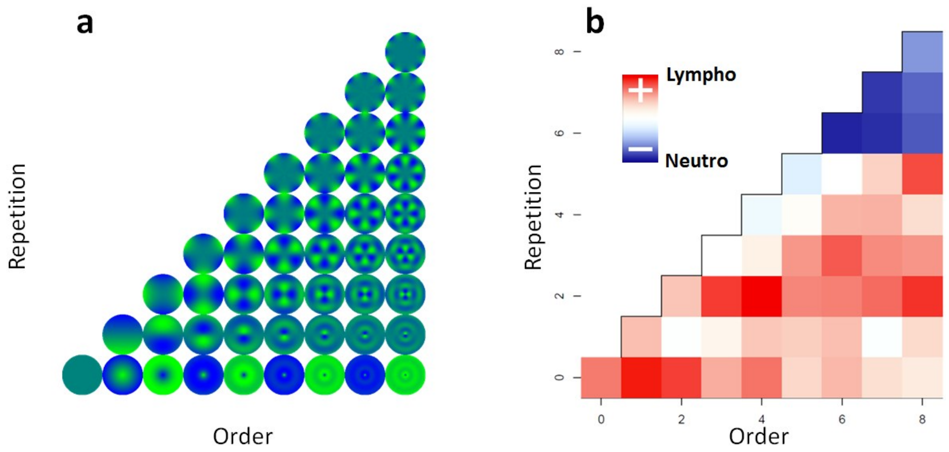


Figure 13: The visualization of the functions that can be described by a single pseudo-Zernike moment (a) and the LDA loadings of a binary classification between neutrophils and lymphocytes (b). For the visualization, the loadings are projected from the LDA space to the PCA space and then to the PZI space.

The LDA model shown in Figure 13b has the advantage that the loadings can be easily interpreted. One weak point regarding LDA is the assumption that the independent variables are normally distributed, which may not be the case for PZI. To increase the classification performance, a support vector machine (SVM) method was utilized, which does not assume normal distribution of variables [58].

The model was optimized by leave-batch-out cross validation (LBOCV) to achieve the best mean sensitivity. In addition to LBOCV, the model testing was performed on an independent data set. The test data classification results are shown in the form of a confusion matrix (see Table 1) with original and preprocessed images shown for all three misclassified cells. The accuracy of the prediction of an independent test data set for the optimized SVM model was 97% [P2].

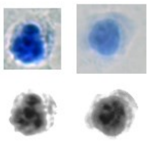
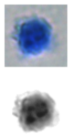
		Predicted (by statistical model)	
		Lymphocytes	Neutrophils
True (assorted by pathologist)	Lymphocytes	17	
	Neutrophils		101

Table 1: SVM prediction of test data. The correctly predicted cells are indicated by numbers. All three wrong predicted cells are shown as images in two variants: original images (top rows) and preprocessed images (bottom rows, grayscale). Along with the labels of true values, their schematic diagrams are shown.

3.2. Merging the predictions from morphological and Raman spectroscopic data

Raman micro-spectroscopy can be efficiently used both for imaging purposes and for differentiation between cell types. In the imaging mode, a Raman micro-spectrometer provides complete spectral information and generates false-color images of the investigated cell. Both representations of the measured data were used to illustrate the effect of combining Raman spectroscopic data with the morphological features of the cells [56].

The morphological features were extracted from the Raman false-color images generated from the spectral scans. To do so, a spectral band at 782 cm^{-1} (from 765 to 798 cm^{-1}), which is related to the ring breathing mode of nucleotides (cytosine, thymine, and uracil), was integrated [59]. Subsequently, the images were cropped to the cell area, centered, and scaled as described in the previous section (see Figure 12).

In addition to the analysis of Raman false-color images, Kimura-stained microscopic cell images were analyzed in combination with Raman data. Preprocessing of the Kimura-stained images was also performed as described in the previous section. The morphological features were extracted from the Raman false-color images and from the microscopic images by means of calculating pseudo-Zernike invariants. These morphological features extracted from both image types were further analyzed by PCA-LDA models.

Besides analyzing morphological features, the analysis of averaged Raman spectra extracted for each cell was carried out. Preprocessing was performed in the same manner as described in section "2.1 Raman spectroscopy" (see Figure 3), and a PCA-LDA model for the leukocyte subtype was constructed. To ensure a reliable validation of the results, the data set was divided into independent batches that were used as folds in the CV routine of the analysis. To keep the folds independent, the samples collected from the same donor or measured on the same day were combined to a batch. This CV approach was utilized because it estimates how the models will perform for an independently measured test data set.

After the separate models were evaluated, the morphological features were reduced via a PCA. Thereafter, the respective PC scores were combined with the PC scores extracted from the mean Raman spectra of the WBC. The combined data were further analyzed via an LDA. To decrease the number of free parameters, the number of PCs for the Raman spectroscopic data was fixed to achieve the best identification of leukocyte subtypes according to LBOCV. For the optimization of the number of PCs obtained from the morphological features, two-level cross validation has been used. In this CV, each batch was predicted by a model with an optimized PC number. This optimization was performed for every batch by an LBOCV of the data without the current batch. Two-level CV provided the best parameters for the model and estimated how accurate the predictive model will perform in practice. To get an overview on how the order of pseudo-Zernike invariants influences the model performance, the order of PZI was left as a free parameter (see Figure 14).

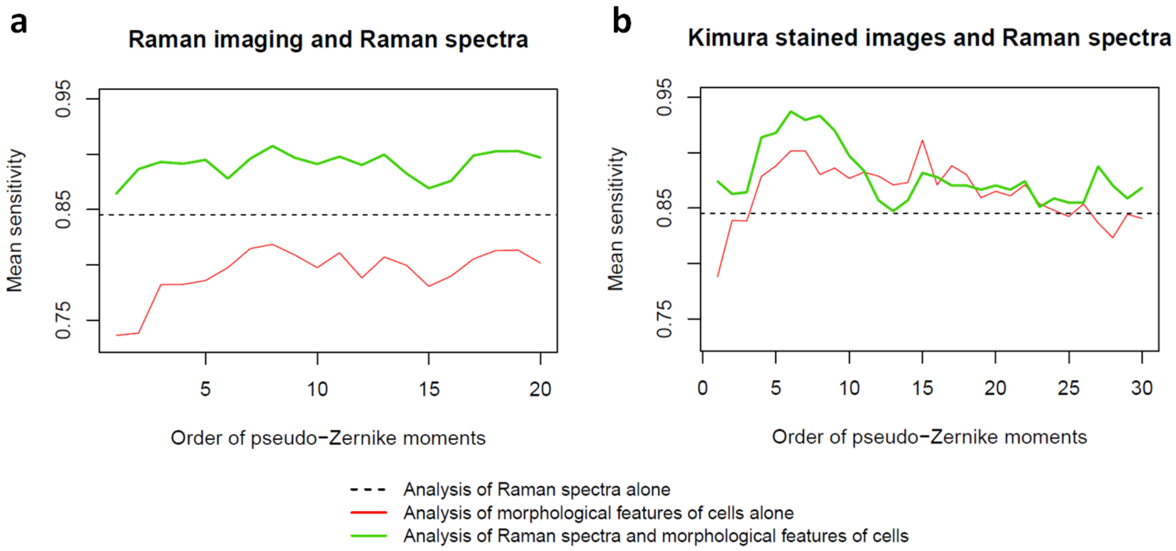


Figure 14: PCA-LDA classification between two major types of lymphocytes. Prediction efficiency of the model is visualized as a function of the included moments order. The dotted line is related to classification by Raman spectra without any morphological features [56]. The improvement of the classification by adding Raman data to the morphological features is clearly seen for low-resolution false-color Raman images (a). The analysis of Kimura-stained cell images also benefits from adding the Raman spectral data to the analysis (b) but to a smaller extent. Despite the smaller improvement in the second case, the optimized combined model with six orders of PZI (b) showed the best performance in this analysis.

As expected, the classification of high-resolution microscopic images provided better results than the classification of the low-resolution false-color Raman images. In the previous section, the high-resolution images were proven to be classified with even higher accuracy, up to 97%. This performance was reached by analyzing morphological features of a larger number of cells using a PCA-SVM model instead of a PCA-LDA. Nevertheless, Figure 14 demonstrates that Raman data significantly contributes to the prediction efficiency when the Raman data is combined with morphological features extracted from either type of images.

3.3. Combining scores obtained from Raman data and clinical values

In the following example of the automated sepsis diagnostic pipeline [60], a set of input variables was chosen considering the further implementation of the approach in the clinical routine. To construct the model, Raman spectroscopic data and inflammatory biomarkers were used. The chosen biomarkers were C-reactive protein (CRP), procalcitonin (PCT), interleukin-6 (IL-6), and soluble urokinase plasminogen activator receptor (suPAR). These biomarkers were used as the input because they are proven markers for bacteremia in patients with systemic inflammatory response syndrome (SIRS) [61]. For the prediction, three classes of severity were considered: patients with SIRS, patients with SIRS and an infection, and patients with sepsis (SIRS with an infection and organ dysfunction).

Prior to the analysis, the biomarkers and the averaged preprocessed Raman spectra for every patient were merged into a single data matrix. As the multiple cell spectra were recorded per patient, averaging also made it possible to increase the signal-to-noise ratio of the spectra, which improved the classification robustness compared to a model that was built on single cell spectra.

To establish a classification model using the combined data, canonical powered partial least squares (CPPLS) [62] discriminant analysis was utilized. This method performs weighted supervised dimension reduction and provides a good fit with just a few components. Another advantage of the CPPLS model is that additional response values can be included during model training. These additional responses incorporate additional information into a model (e.g., information that is not related to inflammation severity). In the constructed model, different demographical characteristics like age, sex, weight, and height of the patients were used in CPPLS as additional responses.

Although CPPLS does not require additional dimension reduction, limiting the number of components used for the prediction plays an important role. As only four biomarkers are available, less than four components must be used for a reliable interpretation of the loadings. Because low-level data fusion was utilized for the Raman and biomarker data, additional weighting between the data types was necessary. In order to balance the effect of the contributions of Raman spectral data and biomarker data on the analysis (see Figure 15), the weighting was performed using an unsupervised approach of the weighting coefficient optimization described in “2.3 Combined data analysis” [P6].

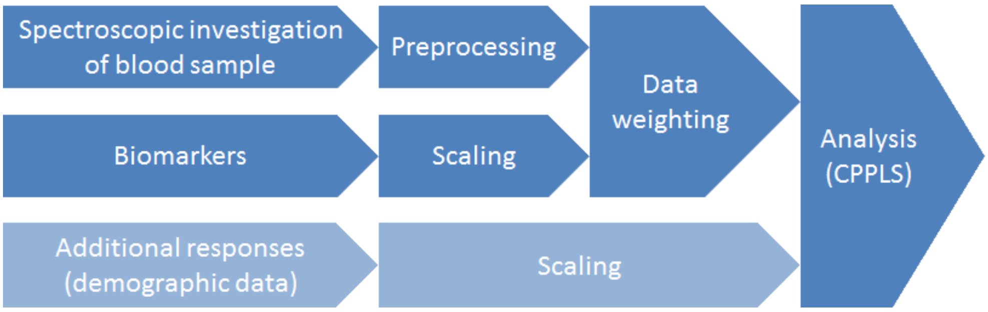

Spectroscopic investigation of blood sample
Preprocessing
Data weighting
Analysis (CPPLS)
Biomarkers
Scaling
Additional responses (demographic data)
Scaling

Figure 15: Distributed data fusion workflow for sepsis diagnostics. Three different branches of the workflow represent different data types. The Raman spectral data, shown in the first branch, requires a multi-step preprocessing procedure (see Figure 3). The other two data types do not require advanced preprocessing but need to be scaled as they contain variables of different measurement units and ranges. After data processing, the measured data (Raman data and biomarkers) is weighted and analyzed considering additional responses by a CPPLS model.

The weighted data set was utilized to construct a classification model using CPPLS discriminant analysis. A number of CPPLS components in the model was optimized by a leave-batch-out cross validation (LBOCV) to take a batch effect [63] into account. This validation approach also estimates how the model will perform on an independent test data set and helps to avoid model overfitting, which is a crucial issue for many advanced machine learning methods. For a CPPLS model, overfitting can occur with a small number of components because the method was designed to minimize the number of components needed to fit the data. In the CPPLS prediction of the inflammation severity, overfitting occurred with a third component. Therefore, two components were found optimal for the combined analysis of Raman spectra and biomarkers (see Figure 16b).

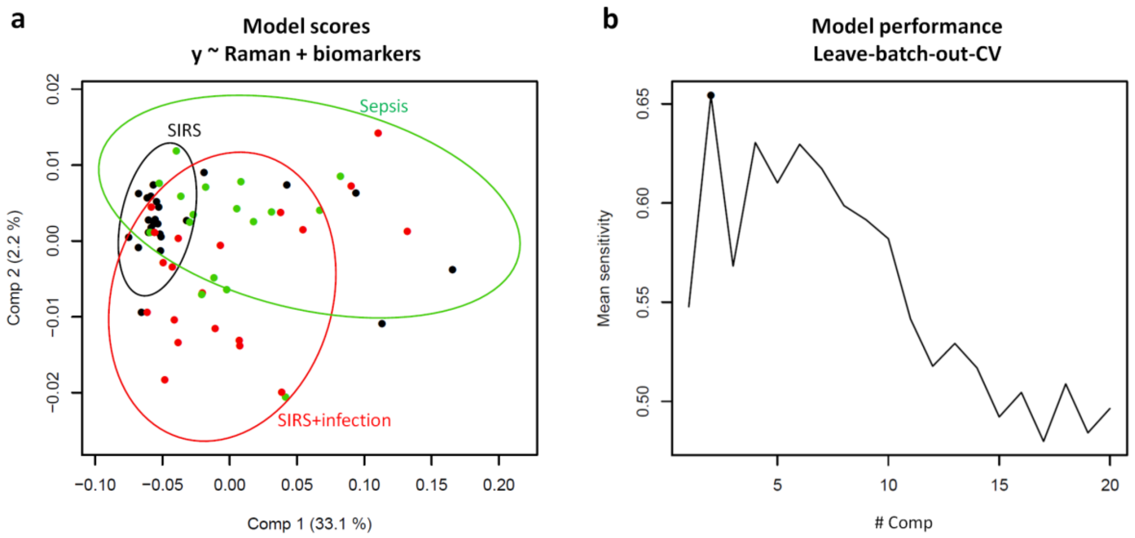


Figure 16: Scores of a combined CPPLS model (a) and a cross-validated performance of disease severity prediction (b). The score plot (a) demonstrates that the first component separates SIRS and sepsis groups, and the second component improves the discrimination of the intermediate state (SIRS with infection group). The mean sensitivity plot (b) demonstrates that two CPPLS components are optimal for prediction according to the LBOCV of the model.

Although the resulting mean sensitivity of the optimized model is only about 65%, it is important to keep in mind that this three-class prediction (see Table 2) corresponds directly to real-life prediction efficiency and can be further improved by extending the data set size and adding further clinical data. It was demonstrated on this small data set that data fusion can improve diagnostic efficiency.

		Predicted					
		1.	2.	3.	*Accuracy*	*Sensitivity*	*Specificity*
True	1. SIRS	**18**	1	0	0.66	0.95	0.71
	2. SIRS+infection	7	**10**	3		0.5	0.88
	3. Sepsis	4	3	**7**		0.5	0.90

Table 2: LBOCV prediction of inflammatory conditions by combined CPPLS model. The three-class model was constructed using a combination of biomarkers and Raman spectra of leukocytes. Although the overall accuracy of the model is 66%, the model provides high sensitivity for discrimination between non-infection inflammation (SIRS) and infection (sepsis and SIRS with infection) and high specificities for the infection groups.

The data fusion improved the mean sensitivity of the model from 52% (biomarker data) and 59% (Raman data) to 65%. To reveal the possible reason for the improvement, the combined model loadings and fitted scores of CPPLS were investigated. The inspection of the loadings revealed that the first CPPLS component is mainly influenced by the biomarkers, and the second component provided an additional separation according to Raman spectral data and improved the discrimination of the "SIRS with infection" group from other groups (see Figure 16a).

With the help of data fusion, not only the diagnostic performance but also understanding the changes that characterize the disease can be improved. In the latter example, it was shown that variations in the chemical composition of cells measured by Raman spectroscopy and the biomarker data are complementary to each other within the frame of inflammatory condition diagnostics. A similar observation was made in section "2.3 Combined data analysis", where low-level data fusion applied to Raman spectroscopic imaging and MALDI spectrometric imaging revealed an additional component in the tissue sample analysis.

4. Summary

In the presented work, several data fusion and machine learning approaches were explored within the frame of the data combination for various measurement techniques in biomedical applications. The overall goal of using data fusion with machine learning approaches was to improve the robustness of diagnostic applications. For each of the measurement techniques used in this work, the data was analyzed by means of machine learning. Prior to applying these machine learning algorithms, a specific preprocessing pipeline for each type of data had to be established. In these pipelines different preprocessing procedures, which were specific for the measurement techniques, had to be applied sequentially. The sequences of the preprocessing steps used for MALDI spectrometric data, Raman spectroscopic data, and microscopic images are briefly summarized in Figure 17. These pipelines made it possible to standardize the data and to decrease sample-to-sample variations which originate from the instability of devices or small deviations in the sample preparation or measurement routine.

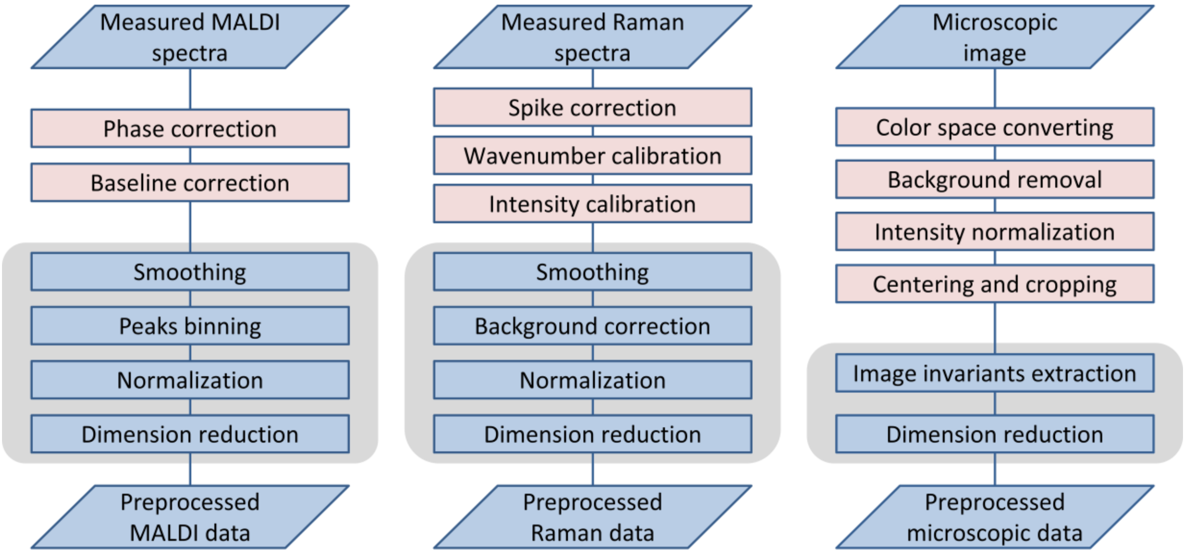


Figure 17: Preprocessing workflows used for various data types. Data obtained from each technique requires specific preprocessing. Some preprocessing steps (depicted in red) can be defined without optimization, but parameters for other preprocessing steps (depicted within gray boxes) can be optimized within the cross validation and can be included either in the preprocessing or in the analysis pipelines.

After preprocessing, the analysis models were established and optimized. Along with the model optimization, the preprocessing steps that deeply influence the outcome of the analysis (gray areas in Figure 17) were optimized, too. This optimization was carried out to establish the most robust preprocessing strategy.

The preprocessed data sets were used for various analyses of biological samples. Separate data analyses were performed for microscopic images [P2], Raman spectra [P5] [P7], and SERS data [P4] [P8]. However, this work mainly focused on the application of data fusion methods for the analysis of biological tissues and cells. To do so, different data fusion pipelines were constructed for each task, depending on the data structure. Both low-level (centralized) and high-level (distributed) data fusion approaches were tested and investigated within in this work. In the examples of centralized data fusion, the data types were combined for the analysis directly after the preprocessing. In the distributed data fusion pipeline, each type of data was preprocessed and analyzed separately, and then the scores obtained or predictions for each type of data were combined for further analysis. Schematic workflows of these two data fusion pipelines are shown in Figure 18.

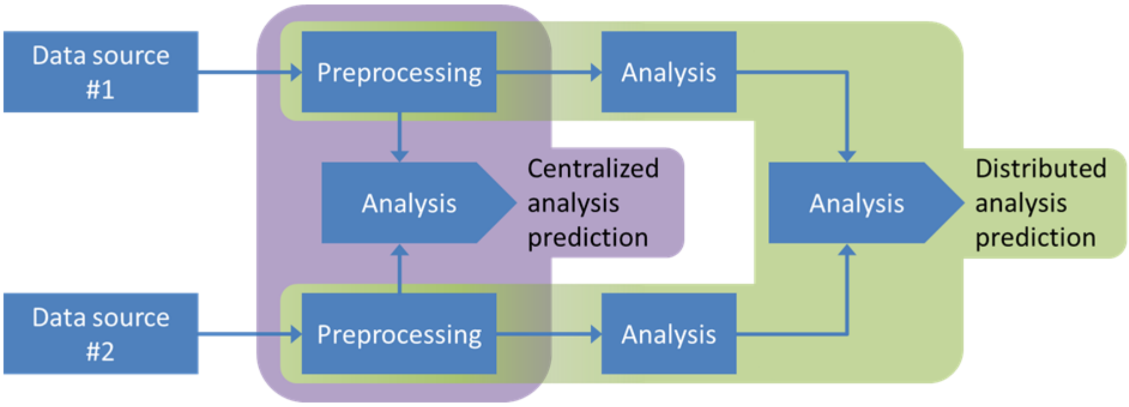


Figure 18: A diagram that demonstrates the difference between data fusion approaches. The data fusion may be performed on a low or high level by means of implementing a centralized or distributed data fusion approach, respectively. These approaches are highlighted in purple and green, respectively. The "preprocessing" boxes are related to the preprocessing workflows shown in Figure 17. However, depending on the data specifics, the dimension reduction can be considered either as a part of the preprocessing or as a part of the analysis.

To demonstrate centralized and distributed data fusion, two examples were implemented for tissue investigation. In both examples, a combination of Raman spectroscopic and MALDI spectrometric data were analyzed. One example demonstrated centralized data fusion for the analysis of the chemical composition of a mouse brain section, and the other example employed distributed data fusion for liver cancer detection.

For the analysis of mouse brain tissue, Raman spectroscopic and MALDI spectrometric imaging data were collected from the same brain section. After aligning Raman data to the grid of the MALDI scan, the spectra of both types were preprocessed and combined. Then an unsupervised analysis was performed by means of PCA. The unsupervised analysis was applied because assignments were not available for sample areas, and the study was focused on obtaining an overview of the observable chemical variations within the sample.

While analyzing the PCA results, it was found that a weighting coefficient needs to be introduced to balance the contribution of MALDI and Raman data within the low-level data fusion pipeline. Several weighting approaches were tested, but the optimal weighting approach was selected based on the cumulative proportion of variance explained by PCs. After applying optimal weighting, an increase in visualization quality was observed, and additional interpretable features could be extracted from the data [P6].

In addition to the unsupervised analysis of the brain tissue section, a combination of Raman spectroscopic and MALDI spectrometric data was used for the classification of hepatocellular carcinoma. It was found that the Raman spectroscopic data provided better separation between HCC and fibrosis than the MALDI spectrometric data. The combined analysis of these two data types was also performed, but it was complicated by observations with incomplete data. To overcome this issue, the missing values were estimated by EM-PCA. Imputation of estimated values to complete the observations extended the data set used in the analysis and increased stability of the model. Even though the significant number of the samples was measured using only MALDI spectrometry, a similar level of separation has been observed for the better performing technique (Raman spectroscopy) and for the combined data.

Other data fusion examples were demonstrated for cell-based analysis. It was demonstrated that leukocyte cell subtype identification can be improved by a centralized data fusion of Raman spectroscopic data and morphological features obtained from microscopic images of stained cells. Moreover, a similar trend was observed when Raman spectral information was combined with morphological features extracted from false-color images generated by integrating a DNA band (782 cm^{-1}) in the Raman spectral scans. This analysis confirmed that both spectroscopic data and microscopic images perform well for WBC subtype prediction, but the combined approach can further improve the model outcome.

The last example presented in this work demonstrated a sepsis diagnostic pipeline based on the combination of Raman spectroscopic data and biomarkers. The biomarkers utilized for this study were C-reactive protein (CRP), procalcitonin (PCT), interleukin-6 (IL-6), and soluble urokinase plasminogen activator receptor (suPAR). This example of a combined data analysis is a proof-of-concept study and aims at future implementation of a three-module-blood investigation device for inflammatory severity diagnostics (see chapter 3). According to the concept, the biomarkers should be measured by a fluorescence module, and the blood cell count should be obtained by a holographic microscopy module. These biomarkers and blood cell count values should be analyzed together with Raman spectra of white blood cells. The output of the complete analytical workflow should predict the severity of a patient's inflammatory condition. Besides the measured values, the demographic information of the patient was included in the analysis process for considering non-disease-related variations. This information was included in the training phase as additional responses for a CPPLS model.

These additional responses and the biomarkers used for model construction were represented as small sets of independent variables and did not require advanced preprocessing and dimension reduction prior to the analysis. The only necessary standardization step was the scaling because the variables had different dynamic ranges. In addition to the scaling of biomarkers, the contribution of Raman spectral data and biomarker data had to be balanced prior to the low-level data fusion. To balance the data, a weighting coefficient was introduced in the same manner as in the Raman-MALDI data fusion for mouse brain tissue analysis [P6]. In comparison to the single models, the combined model improved the sepsis prediction efficiency. This improvement was achieved by combining inflammatory biomarkers with Raman spectral data collected from leukocytes and introducing the additional responses at the phase of model training.

Finally, I would like to highlight that on the way to a combined data analysis, or data fusion, a number of challenges can be met, but the improvement of the results show that it is worth tackling these challenges. During the construction of data fusion pipelines, such issues as unbalanced data contribution, missing values, and variations that are not related to the investigated responses were faced. To resolve these issues, data weighting, missing data imputation, and the introduction of additional responses were employed. For further improvement of analysis reliability, the data fusion pipelines and data processing routine were adjusted for each study in this work. In doing so, researchers' knowledge was used at every step of the analysis process. As a result, the most suitable data fusion approach was found for every example, and a combination of the machine learning methods with data fusion approaches was demonstrated as a powerful tool for data analysis in biomedical applications.

5. Zusammenfassung

In der vorliegenden Arbeit wurden mehrere Datenfusionsverfahren und maschinelle Lernansätze im Rahmen einer Datenkombination verschiedener Messtechniken für biomedizinische Anwendungen untersucht. Das übergeordnete Ziel einer kombinierten Verwendung von Datenfusionsverfahren mit maschinellen Lernansätzen war es, die Robustheit von Diagnoseverfahren zu verbessern. Für jede, der in dieser Arbeit verwendeten, Messtechniken wurden die Daten mittels maschinellen Lernens analysiert. Vor der Anwendung dieser maschinellen Lernalgorithmen musste für jede Art von Daten eine spezifische Daten-Vorverarbeitungspipeline erforscht werden. In diesen Vorverarbeitungspipelines mussten verschiedene, spezifische Vorverarbeitungsverfahren nacheinander angewendet werden. Die Sequenzen der Vorverarbeitungsschritte für MALDI-spektrometrische Daten, Raman-spektroskopische Daten und mikroskopische Bilder sind in Abbildung 17 kurz zusammengefasst. Diese Datenpipelines ermöglichten es, die Daten zu standardisieren und Proben-zu-Proben-Variationen zu reduzieren, welche durch die Instabilität der Geräte oder kleine Abweichungen in der Probenvorbereitung oder der Messroutine entstehen können.

Nach der Datenvorverarbeitung wurden Analysemodelle konstruiert und optimiert. Neben der Modelloptimierung wurden auch die Vorverarbeitungsschritte optimiert, da diese Vorverarbeitungsschritte das Ergebnis der Analyse stark beeinflussen (siehe grau hinterlegte Vorverarbeitungsschritte in Abbildung 17). Diese Optimierung der Vorverarbeitungsparameter wurde durchgeführt, um die robusteste Vorverarbeitungsstrategie zu etablieren.

Die vorverarbeiteten Datensätze wurden für verschiedene Analysen von biologischen Proben verwendet. Eine separate Datenanalyse wurde für mikroskopische Bilder [P2], Ramanspektren [P5] [P7] und SERS-Daten [P4] [P8] durchgeführt. Allerdings konzentrierte sich diese Arbeit hauptsächlich auf die Anwendung von Datenfusionsmethoden für die Analyse von biologischen Geweben und Zellen. Dazu wurden für jede Analyseaufgabe und für verschiedene Datenstrukturen unterschiedliche Datenfusionsdatenpipelines konstruiert. Sowohl zentralisierte

(*Low-Level*) als auch verteilte (*High-Level*) Datenfusionsansätze wurden in dieser Arbeit getestet und untersucht. In der zentralen Datenfusion werden die Datentypen direkt nach der Vorverarbeitung kombiniert und dann kombiniert analysiert. In der verteilten Datenfusionspipeline wurde jede Art von Daten separat vorverarbeitet und analysiert. Dann werden die erhaltenen Ergebnisse oder Vorhersagen zur weiteren Analyse kombiniert. Schematische Arbeitsabläufe dieser beiden Datenfusionspipelines sind in Abbildung 18 dargestellt.

Um die zentralisierte und verteilte Datenfusion zu demonstrieren, wurden als Beispiele zwei Gewebeanalysen untersucht. In beiden Beispielen wurde eine Kombination aus Raman-spektroskopischen und MALDI-spektrometrischen Daten kombiniert. Im ersten Beispiel wurde eine zentralisierte Datenfusion für die Analyse der chemischen Zusammensetzung eines Maus-Hirnschnitts implementiert. Im zweiten Beispiel wurde eine verteilte Datenfusion für eine Lebertumor-Vorhersage verwendet.

Für die Analyse des Maus-Hirngewebes wurden Raman-spektroskopische und MALDI-spektrometrische Bilddaten vom gleichen Gehirnschnitt aufgenommen. Nachdem die Raman-Daten auf das Raster des MALDI-Scans interpoliert wurden, konnten die Spektren beider Datentypen vorverarbeitet und kombiniert werden. Anschließend wurde eine unüberwachte Datenanalyse mittels einer PCA durchgeführt. Dieses unüberwachte Analyseverfahren wurde angewandt, da für die Probenbereiche keine Gewebe-Annotation verfügbar waren. Daher konzentrierte sich die Studie darauf einen Überblick über die chemische Zusammensetzung der Probe zu generieren.

Bei der Interpretation der PCA-Ergebnisse wurde festgestellt, dass ein Gewichtungskoeffizient eingeführt werden muss, um den Beitrag von MALDI- und Raman-Daten innerhalb der *Low-Level*-Datenfusionspipeline anzugleichen. Es wurden mehrere Gewichtungsansätze getestet, wobei der optimale Gewichtungsansatz basierend auf dem kumulativen Anteil der Varianz, der durch verschiedene PC-Anzahlen beschrieben wird, gewählt wurde. Nachdem diese optimale Gewichtung gefunden wurde, konnte eine verbesserte Visualisierung des Scans beobachtet werden [P6]. Des Weiteren konnten zusätzliche spektrale Merkmale extrahiert werden, die interpretiert werden konnten [P6].

Neben dieser unüberwachten Analyse wurde eine Kombination aus Raman-spektroskopischen und MALDI-spektrometrischen Daten für die Klassifikation von hepatozellulären Karzinomen (HCC) verwendet. Es wurde festgestellt, dass die Raman-spektroskopischen Daten eine bessere Trennung zwischen HCC und Fibrose ermöglichen als die MALDI-spektrometrische Daten. Die kombinierte Analyse dieser beiden Datentypen wurde ebenfalls durchgeführt, aber sie wurde durch unvollständige Messdaten erschwert. Um dieses Problem zu lösen, wurden die fehlenden Werte mittels einer EM-PCA geschätzt. Diese Berechnung von Schätzwerten zur Vervollständigung des Datensatzes erweitert den in der Analyse verwendeten Datensatz und erhöht damit die Stabilität des Modells. Obwohl eine signifikante Anzahl der Proben nur mit der MALDI-Spektrometrie vermessen wurden, konnte eine ähnliche Klassifikationsgenauigkeit für die kombinierten Daten wie für die leistungsstärkere Raman-Spektroskopie beobachtet werden.

Weitere Beispiele für die Datenfusion wurden für zellbasierte Anwendungen demonstriert. Es konnte gezeigt werden, dass die Identifizierung des Leukozyten-Subtyps durch eine zentralisierte Datenfusion von Raman-spektroskopischen Daten und morphologischen Merkmalen aus mikroskopischen Bildern von gefärbten Zellen verbessert werden kann. Darüber hinaus wurde ein ähnlicher Trend beobachtet, wenn Raman-spektroskopische Informationen mit morphologischen Merkmalen kombiniert wurden, die aus Falschfarbenbildern einer DNA-Bande (782 cm^{-1}) erzeugt wurden. Diese Analyse bestätigte, dass sowohl spektroskopische Daten als auch mikroskopische Bilder für die Vorhersage des Leukozyten-Subtyps genutzt werden können. Durch einen Datenfusionsansatz kann das Modellergebnis weiter verbessert werden.

Das letzte Beispiel, das in dieser Arbeit vorgestellt wurde, zeigte eine Sepsis-Diagnosepipeline, die auf der Kombination von Raman-spektroskopischen Daten und Biomarkern basiert. Die für diese Studie verwendeten Biomarker waren C-reaktives Protein (CRP), Procalcitonin (PCT), Interleukin-6 (IL-6) und löslicher

Urokinase Plasminogenaktivator-Rezeptor (suPAR). Dieses Beispiel einer kombinierten Datenanalyse ist eine *Proof-of-Concept*-Studie und zielt auf die zukünftige Implementierung eines drei-moduligen Blutuntersuchungsgerätes zur Entzündungsdiagnostik ab (siehe Kapitel 3). Gemäß dem Konzept sollten die Biomarker mit einem Fluoreszenzmodul gemessen und die Leukozytensubtypen-Anzahl mit einem holographischen Mikroskopie-Modul ermittelt werden. Diese Biomarker und Leukozytensubtypen-Anzahlen sollen dann zusammen mit Raman-Spektren weißer Blutzellen analysiert werden. Die Ergebnisse des gesamten Verfahrens soll dann genutzt werden, um den Entzündungszustand eines Patienten vorherzusagen. Neben den Messwerten wurden auch die demographischen Informationen der Patienten in den Analyseprozess zur Berücksichtigung nicht krankheitsbedingter Schwankungen einbezogen. Diese Informationen wurden in der Trainingsphase als zusätzliche Informationen für ein CPPLS-Modell mit aufgenommen.

Diese zusätzlichen Informationen und die für den Modellaufbau verwendeten Biomarker wurden als kleine Mengen unabhängiger Variablen genutzt und erforderten vor der Analyse keine erweiterte Vorverarbeitung und Dimensionsreduktion. Der einzige notwendige Standardisierungsschritt war die Skalierung, da die Variablen unterschiedliche Dynamikbereiche aufwiesen. Neben der Skalierung der Biomarker-Werte musste der Beitrag der Raman-spektroskopischen Daten und der Biomarker-Daten vor der Low-Level-Datenfusion angeglichen werden. Um die Daten anzugleichen, wurde ein Gewichtungskoeffizient eingeführt, ähnlich dem Gewichtungskoeffizient bei der Raman-MALDI-Datenfusion für die Hirngewebeanalyse [P6]. Im Vergleich zu den Einzelmodellen verbesserte das kombinierte Modell die Effizienz der Sepsis-Vorhersage. Diese Verbesserung wurde durch die Kombination von inflammatorischen Biomarkern mit Raman-Spektren von Leukozyten und die Einführung von zusätzlichen Informationen in der Modelltrainingsphase erreicht.

Abschließend soll hervorgehoben werden, dass es auf dem Weg zu einer kombinierten Datenanalyse oder Datenfusion eine Reihe von Herausforderungen bewältigt werden müssen. Die Verbesserung der Ergebnisse zeigt aber, dass es sich lohnt, diese Herausforderungen anzugehen. Bei der Konstruktion von Datenfusionsdatenpipelines entstehen verschiedene Herausforderungen, wie das es zu einem unausgewogenen Beitrag verschiedener Datentypen im Modell kommt, das fehlende Messwerte existieren und das unkorrelierte Variationen auftreten. Um diese Probleme zu lösen, wurden Datengewichtung, Datenschätzungsverfahren und die Einführung zusätzlicher Informationen in die Trainingsphase eingesetzt. Zur weiteren Verbesserung der Robustheit der Analyseverfahren wurden die Datenfusionspipelines und die Datenverarbeitungsroutinen für jede Studie in dieser Arbeit individual angepasst und optimiert. Für diese Anpassung wurde Expertenwissen genutzt um jedem Schritt des Analyseprozesses zu verbessern. Dadurch wurde für jedes der präsentierten Beispiele in dieser Arbeit der am besten geeignetste Datenfusionsansatz gefunden und es konnte die Kombination von maschinellen Lernmethoden mit Datenfusionsansätzen als leistungsfähiges Werkzeug zur Datenanalyse in biomedizinischen Anwendungen demonstriert werden.

Bibliography

1. White, D., et al., *Laboratory medicine—an introduction*, in *Clinical Chemistry*. 2016, Garland Science. p. 17-30.
2. Gurcan, M.N., et al., *Histopathological Image Analysis: A Review.* IEEE Reviews in Biomedical Engineering, 2009. **2**: p. 147-171.
3. Khalbuss, W.E., L. Pantanowitz, and A.V. Parwani, *Digital Imaging in Cytopathology.* Pathology Research International, 2011. **2011**: p. 10.
4. Ghaznavi, F., et al., *Digital Imaging in Pathology: Whole-Slide Imaging and Beyond.* Annual Review of Pathology: Mechanisms of Disease, 2013. **8**(1): p. 331-359.
5. Gal, A.A., *In search of the origins of modern surgical pathology.* Advances in anatomic pathology, 2001. **8**(1): p. 1-13.
6. Kimura, I., Y. Moritani, and Y. Tanizaki, *Basophils in bronchial asthma with reference to reagin-type allergy.* Clinical & Experimental Allergy, 1973. **3**(2): p. 195-202.
7. Mireskandari, M. and I. Petersen, *Clinical Pathology*, in *Ex-vivo and In-vivo Optical Molecular Pathology*. 2014, Wiley-VCH Verlag GmbH & Co. KGaA. p. 1-26.
8. Moran, M.S., et al., *Society of Surgical Oncology–American Society for Radiation Oncology Consensus Guideline on Margins for Breast-Conserving Surgery With Whole-Breast Irradiation in Stages I and II Invasive Breast Cancer.* Annals of Surgical Oncology, 2014. **21**(3): p. 704-716.
9. Stummer, W., et al., *Counterbalancing risks and gains from extended resections in malignant glioma surgery: a supplemental analysis from the randomized 5-aminolevulinic acid glioma resection study.* Journal of neurosurgery, 2011. **114**(3): p. 613-623.
10. Avorn, J., *The Psychology of Clinical Decision Making — Implications for Medication Use.* New England Journal of Medicine, 2018. **378**(8): p. 689-691.

11. Shoo, B.A., R.W. Sagebiel, and M. Kashani-Sabet, *Discordance in the histopathologic diagnosis of melanoma at a melanoma referral center.* Journal of the American Academy of Dermatology, 2010. **62**(5): p. 751-756.
12. Frey, C.B. and M.A. Osborne, *The future of employment: How susceptible are jobs to computerisation?* Technological Forecasting and Social Change, 2017. **114**: p. 254-280.
13. Al-Janabi, S., A. Huisman, and P.J. Van Diest, *Digital pathology: current status and future perspectives.* Histopathology, 2012. **61**(1): p. 1-9.
14. Matousek, P. and N. Stone, *Recent advances in the development of Raman spectroscopy for deep non-invasive medical diagnosis.* Journal of Biophotonics, 2013. **6**(1): p. 7-19.
15. Garg, A.X., et al., *Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review.* JAMA, 2005. **293**(10): p. 1223-1238.
16. Wold, S., M. Sjöström, and L. Eriksson, *PLS-regression: a basic tool of chemometrics.* Chemometrics and Intelligent Laboratory Systems, 2001. **58**(2): p. 109-130.
17. Bocklitz, T.W., et al., *Spectrometer calibration protocol for Raman spectra recorded with different excitation wavelengths.* Spectrochim Acta A Mol Biomol Spectrosc, 2015. **149**(0): p. 544-9.
18. Gu, M., et al., *Accurate mass filtering of ion chromatograms for metabolite identification using a unit mass resolution liquid chromatography/mass spectrometry system.* Rapid Communications in Mass Spectrometry, 2006. **20**(5): p. 764-770.
19. Yang, W. and J.R. Liu. *Research and development of medical image fusion.* in *2013 IEEE International Conference on Medical Imaging Physics and Engineering*. 2013.
20. Castanedo, F., *A Review of Data Fusion Techniques.* The Scientific World Journal, 2013. **2013**: p. 19 Pages.
21. Masyuko, R., et al., *Correlated imaging - a grand challenge in chemical analysis.* Analyst, 2013. **138**(7): p. 1924-1939.

22. Ahlf, D.R., et al., *Correlated mass spectrometry imaging and confocal Raman microscopy for studies of three-dimensional cell culture sections.* Analyst, 2014. **139**(18): p. 4578-4585.
23. Fagerer, S.R., et al., *Analysis of single algal cells by combining mass spectrometry with Raman and fluorescence mapping.* Analyst, 2013. **138**(22): p. 6732-6736.
24. Masyuko, R.N., et al., *Spatial organization of Pseudomonas aeruginosa biofilms probed by combined matrix-assisted laser desorption ionization mass spectrometry and confocal Raman microscopy.* Analyst, 2014. **139**(22): p. 5700-5708.
25. Lanni, E.J., et al., *Correlated Imaging with C60-SIMS and Confocal Raman Microscopy: Visualization of Cell-Scale Molecular Distributions in Bacterial Biofilms.* Analytical Chemistry, 2014. **86**(21): p. 10885-10891.
26. Muhamadali, H., et al., *Chicken, beams, and Campylobacter: rapid differentiation of foodborne bacteria via vibrational spectroscopy and MALDI-mass spectrometry.* Analyst, 2016. **141**(1): p. 111-122.
27. Verwer, P.E.B., et al., *Discrimination of Aspergillus lentulus from Aspergillus fumigatus by Raman spectroscopy and MALDI-TOF MS.* European Journal of Clinical Microbiology & Infectious Diseases, 2014. **33**(2): p. 245-251.
28. Bocklitz, T.W., et al., *Deeper Understanding of Biological Tissue: Quantitative Correlation of MALDI-TOF and Raman Imaging.* Analytical Chemistry, 2013. **85**(22): p. 10829-10834.
29. Kierszenbaum, A.L. and L. Tres, *Histology and Cell Biology: An Introduction to Pathology E-Book.* 2015: Elsevier Health Sciences.
30. Naumann, D., *Vibrational spectroscopy in microbiology and medical diagnostics.* Biomedical Vibrational Spectroscopy. Hoboken, New Jersey, USA: John Wiley & Sons, 2008: p. 1-8.
31. Gowda, G.N., et al., *Metabolomics-based methods for early disease diagnostics.* Expert review of molecular diagnostics, 2008. **8**(5): p. 617-633.
32. Butler, H.J., et al., *Using Raman spectroscopy to characterize biological materials.* Nat. Protocols, 2016. **11**(4): p. 664-687.

33. Bocklitz, T.W., et al., *Raman Based Molecular Imaging and Analytics: A Magic Bullet for Biomedical Applications!?* Anal Chem, 2016. **88**(1): p. 133-51.

34. Kocks, M., et al., *Real-time monitoring of lycopene content in tomato-derived products during processing: implementation of a novel double-slit Raman spectrometer.* Appl Spectrosc, 2013. **67**(6): p. 681-7.

35. Grosse, C., et al., *Label-free imaging and spectroscopic analysis of intracellular bacterial infections.* Anal Chem, 2015. **87**(4): p. 2137-42.

36. Franzen, L. and M. Windbergs, *Applications of Raman spectroscopy in skin research — From skin physiology and diagnosis up to risk assessment and dermal drug delivery.* Advanced Drug Delivery Reviews, 2015. **89**: p. 91-104.

37. Ryabchykov, O., J. Popp, and T. Bocklitz, *Fusion of MALDI Spectrometric Imaging and Raman Spectroscopic Data for the Analysis of Biological Samples.* Front Chem, 2018. **6**: p. 257.

38. Dörfer, T., et al., *Checking and improving calibration of Raman spectra using chemometric approaches.* Zeitschrift Fur Physikalische Chemie-International Journal of Research in Physical Chemistry and Chemical Physics, 2011. **225**(6-7): p. 753-764.

39. Guo, S., T. Bocklitz, and J. Popp, *Optimization of Raman-spectrum baseline correction in biological application.* Analyst, 2016. **141**(8): p. 2396-2404.

40. Lieber, C.A. and A. Mahadevan-Jansen, *Automated method for subtraction of fluorescence from biological Raman spectra.* Applied spectroscopy, 2003. **57**(11): p. 1363-1367.

41. Ryan, C.G., et al., *SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications.* Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms, 1988. **34**(3): p. 396-402.

42. Fitzgerald, M.C., G.R. Parr, and L.M. Smith, *Basic matrices for the matrix-assisted laser desorption/ionization mass spectrometry of proteins and oligonucleotides.* Analytical chemistry, 1993. **65**(22): p. 3204-3211.

43. Gessel, M.M., J.L. Norris, and R.M. Caprioli, *MALDI imaging mass spectrometry: Spatial molecular analysis to enable a new age of discovery.* Journal of Proteomics, 2014. **107**: p. 71-82.

44. Nicklay, J.J., et al., *MALDI Imaging and in Situ Identification of Integral Membrane Proteins from Rat Brain Tissue Sections.* Analytical Chemistry, 2013. **85**(15): p. 7191-7196.

45. Thiery-Lavenant, G., A.I. Zavalin, and R.M. Caprioli, *Targeted Multiplex Imaging Mass Spectrometry in Transmission Geometry for Subcellular Spatial Resolution.* Journal of The American Society for Mass Spectrometry, 2013. **24**(4): p. 609-614.

46. Josse, J. and F. Husson, *Handling missing values in exploratory multivariate data analysis methods.* Journal de la Société Française de Statistique, 2012. **153**(2): p. 79-99.

47. *HemoSpec project*. Available from: hemospec.eu.

48. Kanka, M. and R. Riesenberg, *Lens-free inline holographic microscopy with numerical correction of layers with different refractive index.* Optics Letters, 2015. **40**(5): p. 752-755.

49. Neugebauer, U., et al., *Raman-Spectroscopy Based Cell Identification on a Microhole Array Chip.* Micromachines, 2014. **5**(2): p. 204.

50. Tombelli, S., et al. *{Optical heterogeneous bioassay for the detection of the inflammatory biomarker suPAR}*. in *{SPIE BiOS}*. 2015. SPIE.

51. Bocklitz, T., M. Schmitt, and J.E. Popp, *Image Processing -- Chemometric Approaches to Analyze Optical Molecular Images*, in *Ex-vivo and In-vivo Optical Molecular Pathology*. 2014. p. 215-248.

52. Bhamare, M.M.G. and D.S. Patil. *Automatic blood cell analysis by using digital image processing: A preliminary study*. in *International Journal of Engineering Research and Technology*. 2013. ESRSA Publications.

53. Putzu, L. and C. Di Ruberto. *White blood cells identification and counting from microscopic blood image*. in *Proceedings of World Academy of Science, Engineering and Technology*. 2013.

54. Ryabchykov, O., et al. *Leukocyte subtypes classification by means of image processing*. in *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*. 2016.

55. Habibzadeh, M., A. Krzyżak, and T. Fevens, *Comparative study of feature selection for white blood cell differential counts in low resolution images*, in *Artificial Neural Networks in Pattern Recognition*. 2014, Springer International Publishing. p. 216-227.

56. Ryabchykov, O., et al., *Combination of image processing and Raman spectroscopy for automated white blood cell classification.* Poster Presentation - 8th International Conference on Advanced Vibrational Spectroscopy, 2015.

57. Teh, C.H. and R.T. Chin, *On Image-Analysis by the Methods of Moments.* Ieee Transactions on Pattern Analysis and Machine Intelligence, 1988. **10**(4): p. 496-513.

58. Theodoridis, S., *Chapter 15-clustering algorithms iv In: Theodoridis S, Koutroumbas K, editors.* Pattern Recognition (Fourth Edition). Boston: Academic Press, 2009: p. 765-862.

59. Notingher, I. and L.L. Hench, *Raman microspectroscopy: a noninvasive tool for studies of individual living cells in vitro.* Expert Review of Medical Devices, 2006. **3**(2): p. 215-234.

60. Ranzani, O.T., et al., *New Sepsis Definition (Sepsis-3) and Community-acquired Pneumonia Mortality. A Validation and Clinical Decision-Making Study.* Am J Respir Crit Care Med, 2017. **196**(10): p. 1287-1297.

61. Hoenigl, M., et al., *Diagnostic accuracy of soluble urokinase plasminogen activator receptor (suPAR) for prediction of bacteremia in patients with systemic inflammatory response syndrome.* Clinical Biochemistry, 2013. **46**(3): p. 225-229.

62. Indahl, U.G., K.H. Liland, and T. Næs, *Canonical partial least squares—a unified PLS approach to classification and regression problems.* Journal of Chemometrics, 2009. **23**(9): p. 495-504.

63. Soneson, C., S. Gerster, and M. Delorenzi, *Batch Effect Confounding Leads to Strong Bias in Performance Estimates Obtained by Cross-Validation.* PLOS ONE, 2014. **9**(6): p. 1-13.

Publications

[P1] Automatization of spike correction in Raman spectra of biological samples

Copyright 2016 Elsevier. Reprinted, with permission, from [Ryabchykov, O., Bocklitz, T., Ramoji, A., Neugebauer, U., Foerster, M., Kroegel, C., Bauer, M., Kiehntopf, M. and Popp, J., 2016. Automatization of spike correction in Raman spectra of biological samples. *Chemometrics and Intelligent Laboratory Systems*, *155*, pp.1-6]

Erklärungen zu den Eigenanteilen des Promovenden sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an der Publikation.

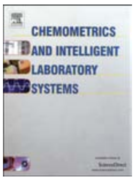
[1]Ryabchykov, O., [2]Bocklitz, T., [3]Ramoji, A., [4]Neugebauer, U., [5]Foerster, M., [6]Kroegel, C., [7]Bauer, M., [8]Kiehntopf, M., [9]Popp, J., 2016. Automatization of spike correction in Raman spectra of biological samples. *Chemometrics and Intelligent Laboratory Systems*, *155*, pp.1-6									
Beteiligt an (Zutreffendes ankreuzen)									
	1	2	3	4	5	6	7	8	9
Konzeption des Forschungsansatzes	**X**	X							X
Planung der Untersuchungen	**X**			X		X	X	X	X
Datenerhebung			X	X	X	X			
Datenanalyse und Interpretation	**X**	X							
Schreiben des Manuskripts	**X**	X							X
Vorschlag Anrechnung Publikationsäquivalent	1.0								

Chemometrics and Intelligent Laboratory Systems 155 (2016) 1–6

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab



Automatization of spike correction in Raman spectra of biological samples

Oleg Ryabchykov [a,b], Thomas Bocklitz [a,b,*], Anuradha Ramoji [a,c], Ute Neugebauer [a,b,c], Martin Foerster [d], Claus Kroegel [d], Michael Bauer [c], Michael Kiehntopf [e], Juergen Popp [a,b,c]

[a] Leibniz Institute of Photonic Technology, Albert-Einstein-Straße 9, 07745 Jena, Germany
[b] Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich Schiller University Jena, Helmholtzweg 4, 07743 Jena, Germany
[c] Center for Sepsis Control and Care (CSCC), Jena University Hospital, Erlanger Allee 101, 07747 Jena, Germany
[d] Clinic for Internal Medicine I, Department of Pneumology and Allergy/Immunology, Jena University Hospital, Erlanger Allee 101, 07747 Jena, Germany
[e] Institute for Clinical Chemistry and Laboratory Diagnostics, Jena University Hospital, Erlanger Allee 101, 07747 Jena, Germany

ARTICLE INFO

Article history:
Received 22 January 2016
Received in revised form 17 March 2016
Accepted 19 March 2016
Available online 31 March 2016

Keywords:
Cosmic ray noise
Raman spectroscopy
Pre-processing
Automation
Laplacian

ABSTRACT

Raman spectroscopy as a technique has high potential for biological applications, e.g. cell and tissue analysis. In these applications, large data sets are normally recorded which require automated analysis. Unfortunately, a lot of disturbing external influences exist, which negatively affect the analysis of Raman spectra. A problematic corrupting effect in big data sets is cosmic ray noise, which usually produces intense spikes within the Raman spectra. In order to exploit Raman spectroscopy in real world applications, detection and removing of spikes should be stable, data-independent and performed without manual control. Herein, an automatic algorithm for cosmic ray noise correction is presented. The algorithm distinguishes spikes from spectra based on their response to a Laplacian, e.g. their sharpness. Manual rating of the spike presence was used as a benchmark for algorithm validation. The algorithm's sensitivity was estimated to be above 99%.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

With Raman spectra a label-free molecular characterization of biological samples, such as prokaryotic and eukaryotic cells or tissue specimen, can be carried out [1–3]. For example, the analysis of leukocytes by means of Raman spectroscopy offers a high potential for future application as a Raman spectroscopic hemogram [4]. Such spectroscopic hemogram can be utilized along with conventional hemogram for the diagnosis of infections [5] and for routine medical examination [6]. However, the analysis of Raman spectra of biological specimens requires sophisticated statistical data analysis methods as the biochemical changes occurring are subtle.

Prior to an application of these statistical methods, it is important to pretreat the Raman data [7]. This includes standardization and correction procedures for dealing with corrupting effects, which might mask the useful Raman information. The pre-processing should always start with a quality control of the Raman spectra, in order to verify that they contain useful information [8]. Thereafter, correction procedures, such as wavenumber correction, noise reduction and background removal, have to be performed. For each task, specialized and adapted correction algorithms have to be applied for a stable and reliable analysis [9].

Spikes are usually sharp and intense features within a measured Raman spectrum originating from cosmic ray noise on a CCD camera. They result from high energy particles which constantly bombard the earth and hit the CCD detector. When hitting the CCD detector, the particle generates a large number of electrons. If the amount of generated electrons is much larger than the charge packet of the CCD, then blooming and smear effects can appear and bright pixels will be observed in two or more consecutive frames or pixels until all electrons are transferred or leaked [10,11].

In Raman spectra, cosmic ray noise is represented as spikes — high intensity sharp peaks. The intensities, position along the wavenumber axis and the frequency of spikes occurrence are random. Only in the case of a blooming effect, spikes can be found on the same position with decreasing intensities in a few consecutive Raman spectra or in neighboring pixels [12]. Cosmic ray noise influences the analysis of large data negatively, because it affects the outcome of normalization procedure and analysis methods. Therefore, spike correction has to be performed during the pre-processing before the normalization is carried out. Here, we present a methodology for a cosmic-ray spikes detection, removal and data interpolation, where no manual optimization by an operator is required. There are some published methods for cosmic ray spikes correction and these methods differ in respect to different

---

* Corresponding author.
E-mail address: thomas.bocklitz@uni-jena.de (T. Bocklitz).

http://dx.doi.org/10.1016/j.chemolab.2016.03.024
0169-7439/© 2016 Elsevier B.V. All rights reserved.

tasks. For some biological issues an algorithm could be used, which compares each data point with its nearest neighbors within the Raman spectra to determine presence of cosmic ray noise with a bandwidth of only 1–2 pixels [13]. Wavelet transform [14,15], polynomial filters or other smoothing methods [16] are also used for spikes correction along with noise reduction. Smoothed spectra can be used for further analysis or the difference between smoothed and original Raman spectra can be used for spikes detection [17]. The possibility to use these methods on a single Raman spectrum is an advantage of these methods. However, all of them require preset parameters, and the efficiency of the methods depends strongly on the parameter selection. For Raman spectral scans of biological samples this becomes a particular challenge, because of the big number of Raman spectra, which are often in the order of hundreds or thousands of spectra per scan. There are some algorithms that allow for a correction based on the similarity of spectra in a data set. Such methods as the Upper-Bound spectrum data matrices, which is a combination of the Upper-Bound Spectrum method with PCA [18], or the Nearest Neighbor Comparison method for Raman maps, which is based on calculation of correlation coefficients of the spectrum and each of its neighbors [19], are proven to be effective. However, these algorithms also require extensive processing time and adaptation of parameters to a specific data set.

Another important step for a spike correction method is the interpolation of the data points after removing of the spikes. Excluding data without interpolation is also possible, but can produce problems in further steps of analysis, due to missing data. The most reliable way for real-time systems is to repeat the measurement [13]. Different polynomials and linear approximations are most often used, but for the consecutive Raman spectra the average of neighboring Raman spectra can be also used.

In this contribution an automated optimization of the spike correction is achieved, which allows the researcher to work with the large data sets necessary for Raman spectroscopic investigations of complex biological samples. Appropriate criteria, obtained from the feedback of the correction procedure, were used for the purpose of automation.

2. Materials and methods

2.1. Sample preparation

White blood cells (WBCs) were isolated from five healthy volunteers' blood with informed consent according to the Ethics Committee of the Jena University Hospital (Ethic vote 4004–02/14). Briefly, ~100 μl blood from fingertip was obtained using lancet and collected in ethylenediaminetetraacetic acid (EDTA) capillary tube. Red blood cell lysis was carried out by mixing the blood with an ammonium chloride solution such that the total volume of the diluted blood is 1 ml. After 5 min of incubation at room temperature (RT) the mixture is centrifuged for 10 min at 400 g at RT. The WBCs pellet at the bottom of the Eppendorf tube was collected by discarding the supernatant and suspended in phosphate buffer saline solution (PBS, Biochrom AG, Berlin, Germany). The WBCs were chemically fixed with 4% formaldehyde (Carl Roth GmbH & Co. KG, Karlsruhe, Germany) for 10 min, followed by washing the cells successively with PBS and 0.9% NaCl (Carl Roth GmbH & Co. KG, Karlsruhe, Germany). For Raman spectroscopy the WBCs (~1×10^6 cells) were suspended in 100 μL of 0.9% NaCl prepared in distilled water and coated onto CaF_2 slides (Crystal GmbH, Berlin, Germany) by means of cytospin (Shandon Cytospin3 Cytocentrifuge, ThermoScientific, Waltham, USA, 6 min, 300 g). To ensure immobilization of the WBCs the CaF_2 slides were precoated for 10 min at RT, with 0.2% gelatin (Sigma-Aldrich, Darmstadt, Germany) solution prepared using distilled water and sterilized by heating up to 121 °C.

2.2. Raman spectra acquisition

Raman spectra of WBCs were measured with an upright micro-Raman setup (CRM 300, WITec GmbH, Germany) equipped with a 300 g/mm grating (spectral resolution about 7 cm^{-1}) and a Deep Depletion CCD camera (DU401 BR-DD, ANDOR, 1024 × 127 pixels) cooled to −75 °C. An excitation wavelength of 785 nm (diode laser, Toptica Photonics, Germany) was utilized. The laser was focused through a Zeiss 100× objective (NA 0.9) onto the cells giving 75 mW of power in the object plane. Raman images of leukocytes were recorded in the scanning mode with a step size of 0.3 μm and integration time of 1 s per spectrum. The investigated data set, in total 30 Raman spectral scans of cells from 5 donors, featured 53.235 Raman spectra and 1024 wavenumbers positions ranging from 249 cm^{-1} to 3452 cm^{-1}.

3. Calculations

3.1. Computer system

All calculations were carried out using R (version 3.0.2) [20] running on a Windows 7 Professional 64-Bit system. (Intel® Core™ i5–4570 CPU @ 3.20 GHz 2.70 GHz with 8 GB RAM). The used packages were 'Peaks' [21] and 'e1071' [22]. All Raman scans, were imported into R and arranged into a matrix with Raman spectra in the same order as they were measured. Therefore, this matrix can be considered as a 2-dimensional data set with wavenumbers in one dimension and time in the other. Prior to spike correction, fluorescence background was removed by the SNIP algorithm [23]. The wavenumbers region between 249 cm^{-1} and 415 cm^{-1} was excluded from analysis, because of the presence of a CaF_2 band that originated from the substrate.

3.2. Algorithm

There are three important steps within a correction procedure for spikes removal: identification of spikes, choosing a threshold to separate spikes from other peaks and the interpolation of data after spike removal. Several basic characteristics of spikes may be helpful for their identification: spikes usually feature a high intensity, sharpness and random position in the Raman spectrum. A high number of spikes have intensities much higher than the intensity of Raman spectral bands, but some of them are comparable with Raman peaks and cannot be recognized by their intensity alone. Therefore, a marker has to be developed to better differentiate between spikes and Raman spectral bands. This marker should take the different spike characteristics into account. Spikes are usually represented by one or few pixels within the Raman spectra that are more intense than previous and subsequent points. However, in a real Raman spectrum of biological samples the change of intensity from point to point is not as abrupt as for spikes. A mathematical formula of this idea is the discrete Laplace operator D_x^2 (Eq. (1)), which corresponds to the sharpness of spectral features.

$$D_x^2 = [-1 \ \ 2 \ \ -1] \tag{1}$$

The discrete Laplace operator response of a sharp thin spectral feature, like a spike is much higher compared to the response of a wider peak with same maximal intensity. On each side of the response to a sharp spectral feature two minima are occurring, which exhibit high absolute values. As it is shown in Fig. 1, the usage of a Laplacian operator enhances the separation between Raman peaks and spikes. In this way not only the intensity is used for separation but also the Spikes sharpness.

Another typical characteristic of spikes, but not for Raman bands, is their random position. In the case of biological Raman scans, the change between Raman spectra of nearby scanning positions are usually small. If a spike appears in one of the Raman spectra, then the intensity change at the spike position from previous and subsequent Raman spectra is significantly higher than in other spectral regions.

This feature can be used to further enhance the Laplacian response for spikes. If the matrix with Raman spectra consists of consecutive spectra, a similar method as for sharpness can be used to achieve a high response for unexpected deviations of intensity within the current

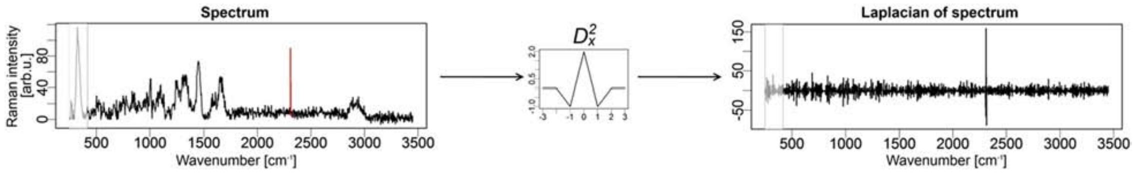


Fig. 1. Single Raman spectrum (left) and its one-dimensional Laplace operator response (right) is visualized. The response improves separation between Raman peaks and spikes.

spectrum in a comparison to previous and next rows of the spectral matrix. By applying operator D_y^2 (Eq. (2)) that is similar to D_x^2 (Eq. (1)) for columns of the matrix a higher response from peaks with random position can be obtained.

$$D_y^2 = \begin{bmatrix} -1 \\ 2 \\ -1 \end{bmatrix} \tag{2}$$

pt?>Each of these two operators gives an enhanced response for spikes. Combining Eqs. (1) and (2) a typical representations of a 2-dimensional Laplacian matrix can be constructed:

$$D_{xy}^2 = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}. \tag{3}$$

Applying the 2-dimensional Laplacian operator D_{xy}^2 (Eq. (3)) to a matrix of Raman spectra produces response matrix ΔS. Single element of this matrix can be described by the following formula:

$$(\Delta S)_{i,j} = 4\,S_{i,j} - S_{i-1,j} - S_{i+1,j} - S_{i,j-1} - S_{i,j+1}. \tag{4}$$

Location of spikes positions within the matrix of Raman spectra (Fig. 2a) can be performed by applying a preset threshold to the 2-dimensional Laplacian response (Fig. 2b).

However, for real-time or real-world application a spike correction applying a manual threshold is an unsuitable solution. For such applications it is important to find a way that allows the detection of spikes without additional manipulations by a researcher. The easiest way to determine a threshold automatically is to investigate the distribution of absolute values of the response. A difficulty occurring within this framework is that scanning regions differ widely in signal-to-noise ratio and contain bands with different sharpness and intensities. Therefore, the response of each Raman spectrum has to be normalized to account for this irregularity. One possible solution is the normalization to the standard deviation. The maximal value and a small window around it can be excluded from standard deviation estimation. This procedure minimizes the influence of the spike presence to normalization value.

As a marker indicating the presence of a spike in a Raman spectrum, the maximal value of the normalized Laplacian response was applied. The normalization was carried out with respect to the standard deviation of the Laplacian response excluding a window around the maximal value. To make this spike marker more representative it was inverted:

$$R_i = \max\left(\frac{sd\left((\Delta S)_{i,1}, \ldots, (\Delta S)_{i,j-5}, (\Delta S)_{i,j+5}, \ldots, (\Delta S)_{i,m}\right)}{(\Delta S)_{i,j}}\right). \tag{5}$$

In this equation the ratio R_i is calculated for the ith spectrum within the matrix of spectra (S). The Laplacian (response) of matrix S is denoted as ΔS and m represents the number of columns in ΔS. For large data sets the distribution of this value has two maxima. The maximum with the lower R value is much smaller compared to the maximum with higher R values and corresponds to Raman spectra with spikes. Therefore, we can determine the threshold ratio as a first local minimum on the left side from the absolute maximum of the distribution. Thereafter, all spectra with R below the threshold are considered as spectra with spikes.

For Raman spectra which are selected as spectra with spikes, the wavenumbers, related to spike positions, can be found in the areas around highest response within each of those spectra. After detecting the presence and position of a spike in a Raman spectrum it can be easily excluded. However, using data with empty values leads to a lot of additional problems and increased complexity of further analysis procedures. The simplest approach to replace empty values is an interpolation by linear functions between the previous and subsequent spectral points. This easy method has the disadvantage that it introduces artifacts if a spike is found on the top of a Raman band. Interpolation can be done using nonlinear functions within each spectrum or within the area around missing data points in the 2-dimensional matrix of the Raman spectra. This type of correction procedure is more efficient and produces better results. Nevertheless, choosing good parameters for this interpolation is a complex task because of the difference in noise levels of the Raman spectra and differences in intensities and sharpness of Raman bands in different wavenumber regions. Therefore, we suggest an interpolation procedure that iteratively replaces missing data points by the medians of more than three nearest neighbors in the

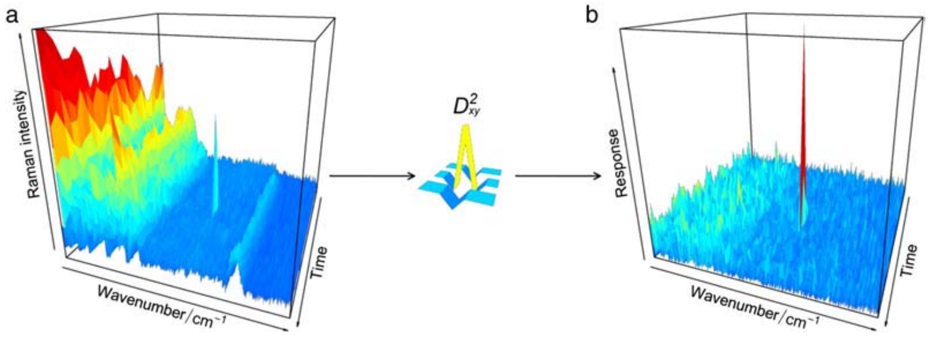


Fig. 2. Two-dimensional representation of spectra with low-intensity spike (a) and absolute response $|\Delta S|$ to the Laplace operator (b) are plotted. A threshold can be used to identify the spike within the Raman spectra.

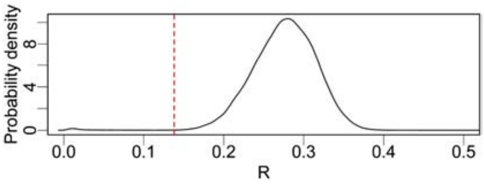


Fig. 3. Density distribution of the ratio *R* for 53.235 Raman spectra is visualized. The automatically detected threshold is indicated by red dotted line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

matrix of Raman spectra. This method takes Raman intensities around the current wavenumber position within the current Raman spectrum and in previous and subsequent Raman spectra into account. It can be compared with a wound healing, because "damaged" data is not interpolated in one step, but the area of missing data points is iteratively decreased step by step. In more detail this process is shown on Fig. 4. To protect the algorithm from an endless loop (in case of missing points in the corner of the matrix) the number of requested neighbors in the current iteration is decreased by one if none of the missing data points have enough neighbors with known values. The method is similar to running a median filter that makes the procedure less sensitive to noise.

However, with a single run of the described algorithm it is possible to detect only one spike within the spectrum. To detect spectra with more than one spike, the entire algorithm should be repeated iteratively until no spectra with spikes are detected or a fixed limit of iterations is reached.

4. Results

After defining the spike correction procedure a validation should be carried out. It is difficult to validate the results using real data set, because there is no available information about the presence of spikes in the original data. That is why it is not possible to make standard calculations for sensitivity and specificity of results. In order to evaluate the results of our correction procedure, we applied our algorithm to the test data set. The algorithm predicted 273 Raman spectra out of 53.235 Raman spectra to feature a spike. In order to estimate these results, all 273 Raman spectra with detected spikes and the same number of Raman spectra without identified spikes near to the threshold of 0.138 detected as a local minimum from density distribution of R_i (Fig. 3) were checked visually for the presence of spikes. (See Fig. 5.)

Only 5 Raman spectra of the 546 Raman spectra were wrongly predicted. 21 Raman spectra which were predicted contain a spike, had an artifact that could be affected by cosmic rays (Table 1). The judgment is not possible as the noise in these Raman spectra was so high that a differentiation between Raman signals, spikes and other types of noise was not possible. These Raman spectra are presented in the confusion table as "uncertain", because the "true value" for these spectra cannot be estimated. In Fig. 3 these Raman spectra have ratios between 0.1 and 0.138, which was the automatically determined threshold. The range of the ratios for wrong predictions is much larger, but their number is too small to generalize the statement.

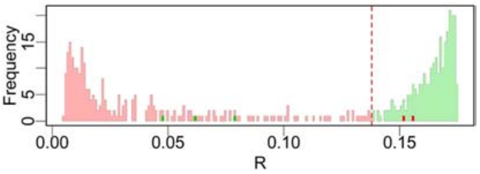


Fig. 5. Histogram of the ratio *R* for the manually checked Raman spectra is depicted. In total 546 spectra with spikes (red) and without spikes (green) are visualized. The threshold, which is determined automatically from probability density (Fig. 3), is shown by red dashed line. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Confusion table for the spike detection.

		Predicted	
		Spike	No spike
True	Spike	249	2
	No spike	3	271
	Uncertain	21	

However, one of the two Raman spectra, in which spikes were wrongly undetected, had a low-intensity spike and the other Raman spectrum featured a spike, which was untypically wide. Raman spectra with wrongly predicted spikes featured on a map as a small area surrounded by substrate only. Within these Raman spectra (Fig. 6) carbon occurred probably due to a burning effect. The atypical shape of the spectral features within these wrongly predicted spectra might be a reason for the wrong detection.

An example of a falsely detected spectrum is shown in Fig. 6. The figure shows that the detection method features some limitations. In case of a small number of spectra featuring unusual spectral features, the 1-dimensional operator D_x^2 (Eq. (1)) should be applied rather than the 2-dimensional D_{xy}^2 (Eq. (3)), as it produces robust results. In our experimental setup the Raman spectra were derived from a scanning experiment, where the spectra are collected sequentially. Therefore, the two dimensional scanning grid is rastered by the excitation laser, which allowed us to apply the presented algorithm without any change to the sequentially recorded spectra. However, if the recording of the scan is not made in such a manner, a 3d Laplacian operator should be applied in three dimensions: a wavenumber dimension and two dimensions related to the position within the scan. This application is straight forward and doesn't require changes in the idea of the algorithm. On the other side, it requires higher computational cost and is not necessary due to high efficiency of 2-dimensional approach.

5. Discussion

5.1. Spike detection limit

To illustrate the detection limit for the algorithm, artificial spikes were induced into 100 random spectra in the data set. In each of the

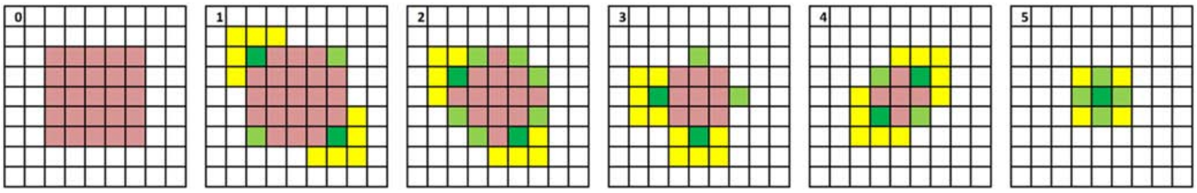


Fig. 4. The interpolation scheme is sketched, which is similar to a wound healing process. In such a process the new tissue starts to grow in the corners, which was adapted here. The first sub-image (0) shows a matrix before first iteration with marked spike positions (red). The spectra in these positions are not further used. In the next image (1) the spectra on the green positions are interpolated with the help of the spectra on the yellow positions. The yellow points are only shown for the darker green spots for clarity reasons and left away for the light green spots. At the last sub-image (5) the excluded data is filled up with interpolated spectra from the surrounding. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

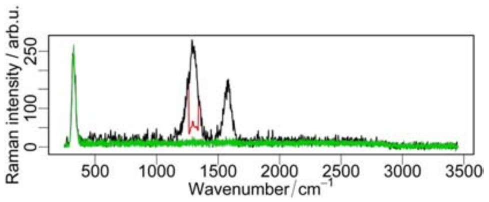


Fig. 6. An example of a wrongly predicted spectrum containing a spike (black line) is visualized. The green spectrum represents the mean of six nearby spectra and the red inset shows the corruption caused by our algorithm. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

selected spectra an intensity value for one wavenumbers between fingerprint region and CH-band was set to higher intensity. Thereafter, the method, proposed in this contribution was applied to this modified data set. By changing the intensity of the artificial spikes and tracking if induced spikes were detected, we could determine a detection limit. For a one-pixel spike with automatic threshold selection this detection limit was at spike intensity of around 40% of the highest peak in the Raman spectrum (Fig. 7). This corresponds to twelve times the standard deviation of the noise within the Raman spectra. The detection limit can be decreased by a manual change of the threshold. However, in this case the balance between sensitivity and specificity will be ruined and a large number of bands not reflecting spikes will be changed by the algorithm in a similar way as outlined for the wrongly predicted spectrum (Fig. 6). For the noise estimation, the difference between original Raman spectrum and Raman spectrum smoothed by fast Fourier transform was utilized. The maximal intensity of the Raman band is approximately 30 times higher than standard deviation of noise that corresponds to peak signal-to-noise ratio (Eq. (6)) around 29 dB. The peak signal-to-noise ratio (PSNR) is defined as

$$PSNR = 10{*}\,log_{10}\left(\frac{MAX_I^2}{MSE}\right) = 20{*}\,log_{10}(MAX_I) - 10{*}\,log_{10}(MSE). \qquad (6)$$

Here MAX_I represents the maximal intensity of the Raman signals and *MSE* corresponds to the mean variation of background noise. With this definition the ratio of lowest detectable spike-to-noise is 20 dB.

5.2. Non-consequence data sets

The use of the two-dimensional Laplace operators is possible for data from scanning spectroscopy. However, in the case of data sets with independent measurements or large distances between points within the Raman scan it is more appropriate to use the one-dimensional operator. Usage of the one-dimensional algorithm on the same data set provides a spike detection limit of about 22 dB or 50% intensity of highest peak of the spectrum (Fig. 8). Therefore, this method is less sensitive than the two-dimensional approach. The reason for this behavior may be associated with a reduction of data points covered by the operator and an increased response range from other types of noise. In this case, the Laplacian response enhancement occurs only by sharpness of the spike without taking its random position into account. However, a decrease in detection efficiency of typical high intensity spikes has not been noticed.

6. Conclusion

In this work, an algorithm for cosmic ray noise correction in large data sets of Raman spectra such as Raman scans is presented. It includes spike detection, removal and replacement by interpolated values. The detection is realized by an automatically selected threshold from a distribution of marker values, which is proportional to a maximal normalized response on the Laplace operator calculated for each Raman spectrum. This method detects spikes based on their three main characteristics: high intensity, sharpness and random position in Raman spectra. The algorithm is adapted to consecutive data sets, but a one-dimensional approach can be applied to a large data set as well. Testing of the algorithm on a real data set showed an estimated sensitivity of about 99%. The detection limit for simulated spikes was estimated to be around 40% of the highest Raman band intensity or 20 dB of the peak signal-to-noise ratio. The developed method can be used as part of an automated analysis system, which processes large data sets without the need of a manual control. Automated selection of the threshold for cosmic ray noise detection allows the application of the algorithm to heterogeneous Raman spectral data sets, which differ in Raman intensities and signal-to-noise ratios. The presented automated method for spike threshold estimation allows a spike correction without preselecting parameters. Therefore, it can be used in a fully automated analysis and pre-processing routine. However, for a fully-automated routine it is also important to construct similar automatic systems for fluorescence background correction methods and other steps of the pre-processing. To do so an optimization of the corresponding parameters has to be achieved with a certain feedback from the correction procedures. Only with a full automation of the pre-processing and analysis the application of Raman spectroscopy to real-world applications, like, for example, in diagnostics in a clinical setting, is possible.

Conflict of interest

The authors declare that there is no conflict of interest.

Acknowledgments

Financial support of the BMBF via the integrated research and treatment center "Center for Sepsis Control and Care" (FKZ 01EO1502) and the project "PhotoSkin" (FKZ 13N13243) are gratefully acknowledged. Funding from the EU via the project "HemoSpec" (FP 7, CN 611682) and cofounding of the project "PhotoSkin" via the 7th framework program of the EU are highly acknowledged.

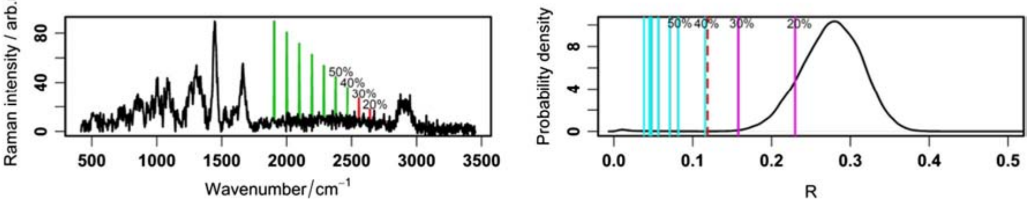


Fig. 7. The influence of artificial spike intensity relative to highest Raman peak in the spectrum is shown. The ratio (R_i) used for spike detection is derived from a two-dimensional Laplace operator. The green spikes were detected while the red spikes were not found. The ratio for undetectable spikes is plotted in magenta and the ratio for detected spikes is shown in cyan. The dashed red line represents the automatically detected threshold. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

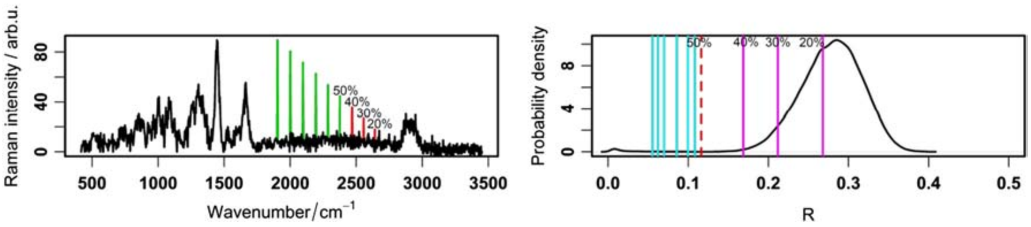


Fig. 8. Influence of artificial spike intensity in comparison with the highest peak in the Raman spectrum is visualized. The ratio (R_i) is calculated for the artificial spike spectra using the response of a one-dimensional Laplace operator. The green spikes were found and featured a ratio value, which is plotted in cyan. The red spikes were not detected and featured a ratio value (magenta) above the automatic threshold. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

A R-script for the described algorithm can be requested from Thomas Bocklitz (thomas.bocklitz@uni-jena.de) and Jürgen Popp (juergen.popp@ipht-jena.de).

References

[1] U. Neugebauer, J.H. Clement, T. Bocklitz, C. Krafft, J. Popp, Identification and differentiation of single cells from peripheral blood by Raman spectroscopic imaging, J. Biophotonics 3 (2010) 579–587.

[2] C. Bielecki, T.W. Bocklitz, M. Schmitt, C. Krafft, C. Marquardt, A. Gharbi, T. Knosel, A. Stallmach, J. Popp, Classification of inflammatory bowel diseases by means of Raman spectroscopic imaging of epithelium cells, J. Biomed. Opt. 17 (2012) 0760301–0760308.

[3] C. Grosse, N. Bergner, J. Dellith, R. Heller, M. Bauer, A. Mellmann, J. Popp, U. Neugebauer, Label-free imaging and spectroscopic analysis of intracellular bacterial infections, Anal. Chem. 87 (2015) 2137–2142.

[4] A. Ramoji, U. Neugebauer, T. Bocklitz, M. Foerster, M. Kiehntopf, M. Bauer, J. Popp, Toward a spectroscopic hemogram: Raman spectroscopic differentiation of the two most abundant leukocytes from peripheral blood, Anal. Chem. 84 (2012) 5335–5342.

[5] A.C. Muller Kobold, J.E. Tulleken, J.G. Zijlstra, W. Sluiter, J. Hermans, C.G.M. Kallenberg, J.W. Cohen Tervaert, Leukocyte activation in sepsis; correlations with disease state and mortality, Intensive Care Med. 26 (2000) 883–892.

[6] M.P. Weijenberg, E.J. Feskens, D. Kromhout, White blood cell count and the risk of coronary heart disease and all-cause mortality in elderly men, Arterioscler. Thromb. Vasc. Biol. 16 (1996) 499–503.

[7] T.W. Bocklitz, T. Dörfer, R. Heinke, M. Schmitt, J. Popp, Spectrometer calibration protocol for Raman spectra recorded with different excitation wavelengths, Spectrochim. Acta A Mol. Biomol. Spectrosc. 149 (2015) 544–549.

[8] T. Bocklitz, M. Schmitt, J.E. Popp, Image processing — chemometric approaches to analyze optical molecular images, Ex-vivo and In-vivo Optical Molecular Pathology 2014, pp. 215–248.

[9] T. Bocklitz, A. Walter, K. Hartmann, P. Rosch, J. Popp, How to pre-process Raman spectra for reliable and stable models? Anal. Chim. Acta 704 (2011) 47–56.

[10] G.E. Healey, R. Kondepudy, Radiometric CCD camera calibration and noise estimation, Pattern Anal. Mach. Intell. IEEE Trans. 16 (1994) 267–276.

[11] H. Young Seok, C. Euncheol, K. Moon Gi, Smear removal algorithm using the optical black region for CCD imaging sensors, Consum. Electron. IEEE Trans. 55 (2009) 2287–2293.

[12] C. Chi-Wai, C. Chung-Yen, C. Shih-Hao, Enhancement of signal performance in LED visible light communications using mobile phone camera, Photonics J. IEEE 7 (2015) 1–7.

[13] J. Zhao, H. Lui, D.I. McLean, H. Zeng, Integrated real-time Raman system for clinical in vivo skin analysis, Skin Res. Technol. 14 (2008) 484–492.

[14] F. Ehrentreich, L. Summchen, Spike removal and denoising of Raman spectra by wavelet transform methods, Anal. Chem. 73 (2001) 4364–4373.

[15] D. Chen, Z. Chen, E. Grant, Adaptive wavelet transform suppresses background and noise for quantitative analysis by Raman spectrometry, Anal. Bioanal. Chem. 400 (2011) 625–634.

[16] G.R. Phillips, J.M. Harris, Polynomial filters for data sets with outlying or missing observations — application to charge-coupled-device-detected Raman-spectra contaminated by cosmic-rays, Anal. Chem. 62 (1990) 2351–2357.

[17] L. Zhang, M.J. Henson, A practical algorithm to remove cosmic spikes in Raman imaging data for pharmaceutical applications, Appl. Spectrosc. 61 (2007) 1015–1020.

[18] D. Zhang, D. Ben-Amotz, Removal of cosmic spikes from hyper-spectral images using a hybrid upper-bound spectrum method, Appl. Spectrosc. 56 (2002) 91–98.

[19] U.B. Cappel, I.M. Bell, L.K. Pickard, Removing cosmic ray features from Raman map data by a refined nearest neighbor comparison method as a precursor for chemometric analysis, Appl. Spectrosc. 64 (2010) 195–200.

[20] R Core Team, R, A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2014.

[21] M. Morhac, Peaks: Peaks, 2012.

[22] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, e1071: Misc Functions of the Department of Statistics (e1071), TU Wien, 2014.

[23] C.G. Ryan, E. Clayton, W.L. Griffin, S.H. Sie, D.R. Cousens, SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications, Nucl. Instrum. Methods Phys. Res., Sect. B 34 (1988) 396–402.

[P2] Leukocyte subtypes classification by means of image processing

Copyright 2016 IEEE. Reprinted, with permission, from [Ryabchykov, O., Ramoji, A., Bocklitz, T., Foerster, M., Hagel, S., Kroegel, C., Bauer, M., Neugebauer, U. and Popp, J., 2016, September. Leukocyte subtypes classification by means of image processing. In *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on* (pp. 309-316). IEEE]

Erklärungen zu den Eigenanteilen des Promovenden sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an der Publikation.

[1]Ryabchykov, O., [2]Ramoji, A., [3]Bocklitz, T., [4]Foerster, M., [5]Hagel, S., [6]Kroegel, C., [7]Bauer, M., [8]Neugebauer, U., [9]Popp, J., 2016, September. Leukocyte subtypes classification by means of image processing. In *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on* (pp. 309-316). IEEE									
Beteiligt an (Zutreffendes ankreuzen)									
	1	2	3	4	5	6	7	8	9
Konzeption des Forschungsansatzes	**X**	X	X					X	X
Planung der Untersuchungen	**X**				X		X	X	X
Datenerhebung				X		X		X	
Datenanalyse und Interpretation	**X**	X	X					X	
Schreiben des Manuskripts	**X**		X						X
Vorschlag Anrechnung Publikationsäquivalent	**1.0**								

Proceedings of the Federated Conference on Computer Science and Information Systems pp. 309–316
DOI: 10.15439/2016F80
ACSIS, Vol. 8. ISSN 2300-5963

Leukocyte subtypes classification by means of image processing

Oleg Ryabchykov*†, Anuradha Ramoji*‡, Thomas Bocklitz*†, Martin Foerster§ Stefan Hagel‡¶,
Claus Kroegel§, Michael Bauer‡, Ute Neugebauer*†‡, Juergen Popp*†‡

Emails: oleg.ryabchykov@uni-jena.com, thomas.bocklitz@uni-jena.de, juergen.popp@leibniz-ipht.de
* Leibniz Institute of Photonic Technology, Albert-Einstein-Str. 9, 07745 Jena, Germany
† Institute of Physical Chemistry and Abbe Center of Photonics, FSU Jena, Helmholtzweg 4, 07743 Jena, Germany
‡ Center for Sepsis Control and Care (CSCC), Jena University Hospital, Erlanger Allee 101, 07747 Jena, Germany
§ Clinic for Internal Medicine I, Department of Pneumology and Allergy/Immunology, Jena University Hospital,
Erlanger Allee 101, 07747 Jena, Germany
¶ Center for Infectious Diseases and Infection Control, Jena University Hospital,
Erlanger Allee 101, 07747 Jena, Germany

Abstract—**The classification of leukocyte subtypes is a routine method to diagnose many diseases, infections, and inflammations. By applying an automated cell counting procedure, it is possible to decrease analysis time and increase the number of analyzed cells per patient, thereby making the analysis more robust. Here we propose a method, which automatically differentiate between two white blood cell subtypes, which are present in blood in the highest fractions. We apply generalized pseudo-Zernike moments to transfer morphological information of the cells to features and subsequently to a classification model. The first results indicate that information from the morphology can be used to obtain efficient automatic classification, which was demonstrated for the leukocyte subtype classification of neutrophils and lymphocytes. The approach can be extended to other imaging modalities, like different types of staining, spectroscopic techniques, dark field or phase contrast microscopy.**

I. INTRODUCTION

WHITE blood cells (WBCs) are also called leukocytes. These cells protect the body from infections caused by viruses and other foreign invaders like bacteria or fungi, which make WBCs an important part of the immune system. Leukocytes are produced and derived from the bone marrow and circulate through the bloodstream. A change of the number of different WBC subtypes in the blood is utilized as marker for various diseases. Therefore a blood cell count is often utilized for a routine health examination or diagnosis of specific conditions of a patient. There are five major subtypes of WBCs [1], [2]:

- neutrophils (50-70%);
- lymphocytes (25-30%);
- monocytes (3-9%);
- eosinophils (0-5%);
- basophils (0-1%).

The ranges within the brackets display the percentage of the corresponding cell subtypes in the blood, which are typical ratios for a healthy person. There are various classification approaches, which can be roughly divided into manual and automated methods of cell classification.

The manual classification is performed by a pathologist through the subjective recognition of cell subtypes on microscopic images of stained cells. This type of analysis does not require complex equipment or highly specialized chemical reagents. To simplify the identification, cells are usually stained with the Kimura stain, which colors cell nuclei in blue. Manual differentiation between varying subtypes is accomplished based on characteristics of the cell morphology, like cell size, transparency, granularity, and the shape of the cell nucleus, which are the major differences between the subtypes. Manual classification is widely used in some specific cases of diagnosis and as a “gold standard” for scientific purposes. However, variation of cell morphology within the same cell subtype is very high, and manual classification efficiency is dependent on the pathologist’s qualification and experience.

On the other side, there are various automated classification methods, based on different physical and chemical characteristics of the cells. The main advantage of the automated devices is that they efficiently analyze large number of cells in a short time. Unfortunately, their analyzing workflows include very specific combinations of chemical and physical processes. The complexity of the analysis does not allow the design of a simple portable device. Therefore, automated blood cell counting machines are usually big and expensive.

An alternative approach is an automatic image analysis of microscopic images of stained cells. In a combination with a small camera this method can become a useful tool for doctors, providing them an instant access to the information about WBCs population at bedside of a patient. There are some studies that show efficient leukocyte identification [3], [4] and segmentation [5], [6] within microscopic images. However, these studies are focused on the leukocyte count without the classification of the leukocytes into subtypes. That leads to the loss of important information about the proportions of each cell subtype. In distinction to the mentioned studies, the current manuscript describes an algorithm for the classification

978-83-60810-90-3/$25.00©2016, IEEE

of WBCs, focusing on the textural features analysis of single cell images.

The concept of the work is to extract quantitative features related to the cell morphology from the microscopic images. Subsequently, these features are used to train and evaluate a statistical model for cell subtype identification. Moreover, the same type of images as for manual classification is used, therefore, this approach allows a direct comparison to the "gold standard". In order to use these images for an automated image analysis, standardization and preprocessing have to be carried out. However, during the pretreatment step, it is important to eliminate corrupting effects, such as uniformities in staining and lighting, but to keep the morphological information for further analysis steps.

The textural information extraction from preprocessed images can be carried out by various methods [7], [8]. However, image description by means of pseudo-Zernike (PZ) moments [9] was chosen for the cell subtype identification because it was proven to be a reliable method for the recognition of shapes [10], characters [11], [12], faces [13], [14], [15], and viruses [16]. An advantage of the representation by PZ-moments is that their absolute values are independent from image rotation, which is necessary due to random orientation of the cells on a microscopic slide. The PZ-moments are derived from PZ-polynomials, which are orthogonal to each other and can be used in further statistical analysis, thus an automated classification technique can be established.

The proposed automated cell classification method is aimed to combine the simplicity of the manual classification and the advantages of automatization. The approach is based on the analysis of images, which are similar to the images used for manual "gold standard" method and are produced by common microscopy from a blood sample after non-complicated preparation. On the other side, due to automatization, extremely short classification times and objectivity, comparable with a human observer, can be achieved.

II. MATERIALS AND METHODS

A. Sample preparation

Leukocytes were isolated from the venous blood of patients admitted to the intensive care unit with informed consent according to the Ethics Committee of the Jena University Hospital (Ethic vote n 4004-02/14). Briefly, 2.7ml of blood in ethylenediaminetetraacetic acid (EDTA) was drawn freshly from an existing catheter using the BD monovettes. In case of healthy donor, blood (about 100μl) was collected from fingertip using lancet. Red blood cell lysis was carried out by mixing the blood with an ammonium chloride solution with a ratio of 1:5 in a 50ml falcon tube. After 5 minutes of incubation at room temperature (RT), the mixture was centrifuged for 10 minutes at 400g at RT. The WBC pellet at the bottom of the falcon tube was collected by discarding the supernatant and suspending it in a phosphate buffer solution (PBS). The WBCs were chemically fixed with 4% formaldehyde for 10 minutes, followed by washing the cells successively with PBS and 0.9% NaCl. The cells were coated on slides using cytospin and stained with a Kimura staining solution (which stains only the cellular nucleus) and washed with distilled water. The slides were dried at RT and stored at $4\,^{\circ}\mathrm{C}$ for maximum one hour until further use. The Kimura stained images of the WBCs (Fig. 1 *a,b*) were captured with an upright epifluorescence microscope (Axioplan 2, Carl Zeiss, Germany) equipped with an AxioCam HRc camera (Carl Zeiss, Germany). Images were acquired using Zeiss Axio Vert software (Carl Zeiss, Germany).

B. Calculations

All calculations reported in this work were carried out in Gnu R (version 3.0.2) [17] running on a Windows 7 Professional 64-bit system (Intel® Core$^{\mathrm{TM}}$ i5-4570 CPU @ 3.20 GHz 2.70 GHz with 8GB RAM). In addition to the base R package, which contains the input/output, basic programming support, and arithmetic functions, some more specific algorithms were utilized from other packages. For orthogonal moment analysis the "IM" package [18] was used. A support vector machine (SVM) classification model was built with the "e1071" package [19]. Parallel computing was obtained by functions from "foreach" [20] and "doParallel" [21] package. K-means clustering from the "stats" package [17] was utilized for the background removal. The functions for principal component analysis (PCA), nonlinear least squares estimation, and the fast Fourier transform (FFT) are all contained in the base package [17]. JPEG files were loaded into the R environment via the "jpeg" package [22].

Prior to analysis, each image was converted from *sRGB* color space to *Lab* color space, one of the most common color spaces for image analysis applications. It was chosen due to the fact that, unlike additive or subtractive color models (for example *RGB* or *CMYK*), it is not optimized for image representation on a screen or for printing, but is adapted to cover the entire range of colors distinguishable by the human eye and to match the perception of these colors. In this color space, a and b components are related to chromatic color values. The L component of *Lab* color space closely matches the human perception of lightness, which allows to expect that in this representation cell subtypes can be identified based on their morphology. The conversion of the color space was performed by base R function "convertColor".

Subsequently to the color space conversion, other steps, such as noise reduction, background removal and intensity normalization were performed. The details of these preprocessing steps are described in the "Results and discussion" section.

C. Pseudo-Zernike (PZ) Moments

As mentioned previously, PZ-moments were chosen for feature extraction from the images. These orthogonal, complex-valued moments are defined on a unit disk and are widely used for pattern recognition. The PZ-moments can describe a 2-dimensional function on the unit circle. However, the function $f(x, y)$ can represent an image if two arguments, x and y, are related to a pixel position and the function value is related to

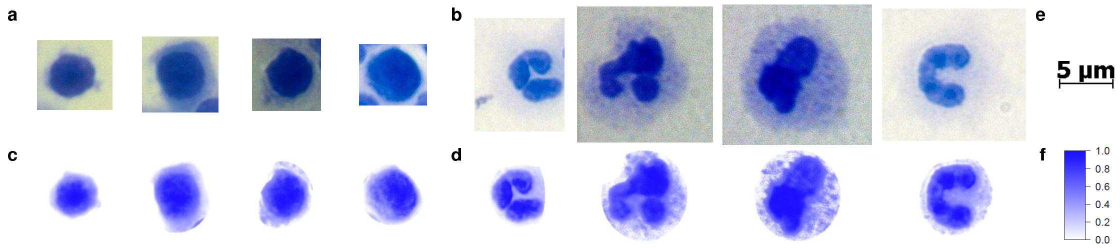


Fig. 1. Original images of two Kimura stained cell subtypes from the patients are displayed in the first row: lymphocytes (a), which are characterized by deep staining of the nuclei and a relatively small amount of cytoplasm, and neutrophils (b), which are the most common subtype that normally contain a nuclei divided into 2-5 lobes. All images are sized according to the scale (e). At the bottom preprocessed false-color equivalents of the presented images (c, d) normalized to the unit scale (f) are shown.

lightness or another color component in that pixel. The PZ-moments (A_{nl}) of an image on a unit disk are defined in radial coordinates by [23]:

$$A_{nl} = \frac{n+1}{\pi} \int_0^{2\pi} \int_0^1 \left[V_{nl} \left(r\cos\theta, r\sin\theta \right) \right]^* f\left(r\cos\theta, r\sin\theta \right) r dr d\theta \ .$$

In this equation $n = 0, \ldots, \infty$ represents the order, the repetition is denoted by $l \leq n$, and f is the value related to the current pixel position: $0 \leq r \leq 1$ and $0 \leq \theta \leq 2\pi$ (polar coordinates of the pixel). V_{nl} is the orthogonal set of complex-valued PZ-polynomials, which can be written as:

$$V_{nl}(r,\theta) = R_{nl}\left(r\right) e^{jl\theta} \ ,$$

where R_{nl} represents radial polynomials with integer coefficients $D_{n,m,s}$:

$$R_{nl}(r) = \sum_{s=0}^{n-|l|} D_{n,|l|,s} r^{(n-s)} \ ,$$

$$D_{n,m,s} = \frac{(-1)^s (2n+1-s)!}{s!(n-m-s)!(n+m-s+1)!} \ .$$

Both the order n and the repetition l are related to the spatial frequencies of the image. However, the order n represents the spatial frequency along the unit disk's radius, while the repetition l represents the spatial frequency along the unit disk's angular coordinate. Moreover, by clarifying the idea behind order n and repetition l, the respective moments can be interpreted. Therefore, the classification model can be checked, analyzed and the morphological differences between the cell subtypes can be examined.

As it is seen from the formulas, the angular coordinate is included in the PZ-moments only within the multiplier $e^{jl\theta}$, which is related to the phase of the complex value [9], [10], [12]. Due to this fact, the absolute values of moments are independent from a rotation of the coordinate system. Thus, they are independent from the spatial alignment of the cell within the image and from the orientation on the microscopic slide. Other advantages of these particular moments are their low sensitivity to noise [10] and that the PZ-moments are orthogonal to each other.

III. RESULTS AND DISCUSSION

A. *Data set*

Taking into account the extremely low number of monocytes, eosinophils, and basophils in the data, only two major subtypes could be investigated in the current study. These both subtypes represent about 90% of WBCs in the blood and were included in the statistical evaluation. Thus, the training data included 28 lymphocytes and 45 neutrophils from 6 patients which were showing signs of inflammation. On the other side, the test data included 128 cells from two healthy volunteers. Unlike the training set, where some cell subtypes were sorted out, the test data included randomly selected cells without presorting or labeling according to their subtypes.

The cell subtypes included in the training data are different in sizes and cell nuclei morphology (see Fig. 1). Most notable is that the neutrophils are relatively big and have multi-lobed nuclei, while lymphocytes have almost round nuclei and are smaller. Other WBC subtypes, which were not included in the training data, are characterized by their granularity and the following properties of the cell nuclei: monocytes have kidney shaped nuclei, eosinophils have relatively small bi-lobed nuclei, and basophils have bi-lobed or tri-lobed nuclei. Although each subtype has a typical average cell size and other specific characteristics, each single cell varies from that average characteristics, which make some of its parameters dissimilar to the typical characteristics of its subtype.

B. *Workflow*

To obtain a stable and efficient analytical system, an image processing workflow was developed and optimized for the specific task of leukocyte subtype classification. The data was loaded, preprocessed, and represented as a set of pseudo-Zernike moments based invariants for further analysis. The workflow is presented in more detail in Fig. 2.

Important and nontrivial steps are the image preprocessing and standardization, which have to be optimized. These procedures should reduce the variations of brightness and color tones between the images of cells within the same sample and occasional appearing variations caused by the sample preparation routine for images taken from different samples. If the workflow presented here is applied to other imaging modalities, like holographic imaging and phase contrast microscopy, these variations are expected to be less significant. Therefore, the preprocessing procedure has to be modified individually for each microscopic imaging technique and classification task.

For the construction of the classification model based on image analysis, the measured cells were labeled according to the classification made by the pathologist. The labeled and preprocessed training data were subsequently divided into three batches for cross-validation of the model. This step of the workflow was of enormous importance for setting model parameters and estimating the model quality. Thereafter, it should not be underestimated.

Leave-batch-out-cross-validation of SVM classification was performed on the training data with different combinations of input variables. This cross-validation procedure was designed to avoid any relations between different batches of cells. Therefore, the data splitting into three batches was arranged so, that the batch reflect the measurement dates and patient's origin. Thus, the generalization performance for the prediction of an independent dataset is well estimated by the leave-one-patient-out-cross-validation. Consequently, classification models with various numbers of PZ-moments' orders and principal components were compared. The variable selection was carried out according to the highest sensitivity for cross-validation of SVM classification model. The model with highest sensitivity was chosen as an optimal one and further used for the test data prediction.

Besides high identification efficiency, the proposed algorithm has to be suitable for real-life applications. Therefore, the workflow was optimized by parallelization of each single image loading, preprocessing, and calculation of the moments. Thereby, the parallelization on hardware with a multi-core processor should decrease the calculation time for a large amount of data roughly by a factor related to the number of calculation units. We chose the number of clusters for parallel calculation as one less than the number of processor cores, which was three for the PC on which the analysis was performed. During the preliminary study stage, the amount of data was relatively small, and thus, the parallelization of calculations had a negligible effect. However, despite the insignificant improvement on a small data set, parallelization is highly important for further applications and implementation of the algorithm, especially for the case if the number of analyzed cells is on the order of thousands.

C. Preprocessing

Examples of WBC images are shown in Fig. 1 *a,b*. As it can be seen by naked eye, differences between some images, which are not related to the cell's morphology, occur. These fluctuations originate from the sample preparation procedure, which is simple and standardized. There are some systematic deviations between the cells of different patients, but also the images of cells from the same patient can differ due to the spatial alignment of the cells and non-uniform coloring of samples along microscopic slides. Moreover, parts of other cells are visible within some images and, additionally, other influences on the brightness, contrast, and tone are present on the microscopic images. To reduce the discussed corrupting effects, an advanced preprocessing has to be carried out before the feature extraction procedure.

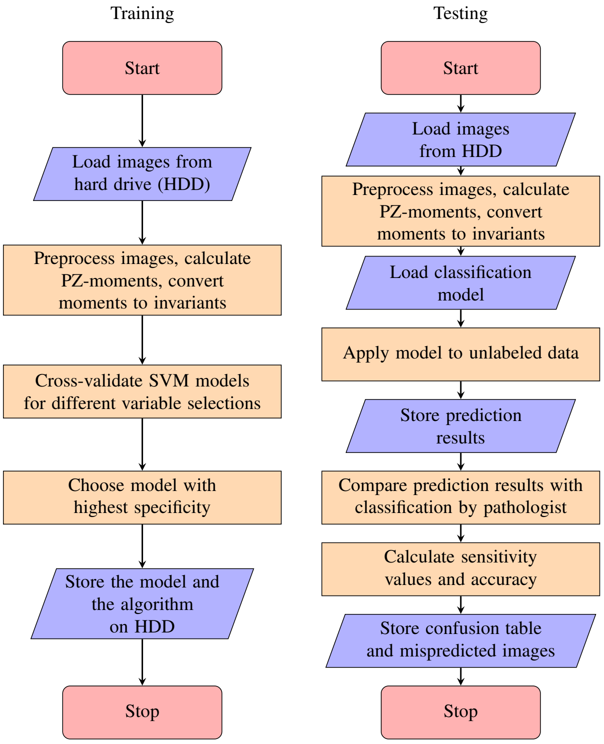

Training
Start
Load images from hard drive (HDD)
Preprocess images, calculate PZ-moments, convert moments to invariants
Cross-validate SVM models for different variable selections
Choose model with highest specificity
Store the model and the algorithm on HDD
Stop
Testing
Start
Load images from HDD
Preprocess images, calculate PZ-moments, convert moments to invariants
Load classification model
Apply model to unlabeled data
Store prediction results
Compare prediction results with classification by pathologist
Calculate sensitivity values and accuracy
Store confusion table and mispredicted images
Stop

Fig. 2. Schematic workflow of the presented algorithm and the model validation.

According to the chosen concept of the analysis, it was important to keep the morphological features which can be distinguished visually. The automation of the preprocessing procedure took an important part in the development of the algorithm. The original images were stored in the standard *sRGB* representation, which is designed to display images in electronic systems, such as a computer's screens. However, analysis of the color channels separately from each other can be problematic and leads to a high complexity of the classification model. Switching to a single component can be circumvented by applying a more convenient color space. As it was mentioned in "Materials and methods", the lightness L of *Lab* color space is closely related to the human visual perception of images. In order to keep the features used for manual classification, the *Lab* color space was used in the further analysis. Moreover, the cells used for analysis were colored by Kimura staining, which highlights the cell nucleus in blue. Due to monochromatic coloring, all variations of the chromatic values are only related to the deviations of the sample preparation process and staining. Thus, related color components (a and b) were skipped and only the lightness

L was analyzed. However, for staining procedures which stain different cell organelles or cytoplasm in different colors, normalized a and b components should be also included in the analysis.

Due to high variations between different images, even for the single L-component, the automation of preprocessing took an important part in the analysis development. Pretreatment was aimed to decrease deviations of the features extracted from images within the same cell subtype and to increase the overall identification accuracy. Consequently, the background, or non-cell area of the images, was cut off via the unsupervised k-means clustering of lightness values within each image. In order to improve the background removal, an FFT-filter was applied to the images prior to the clustering. After the background removal, the lightness distribution within each cell was standardized by means of normalization to the unit interval and equalization of the histogram.

Subsequently to the lightness standardization of the images, a 2-dimensional Gaussian function was fitted to each cell image using nonlinear least squares. Based on the coefficients of the fitted function, centers and estimated radii were determined for each cell. As the next step, background-free images of the single cells were cropped according to the estimated cells' radii. This procedure was performed, to preserve the full region of the stained nucleus with a cytoplasm area and to exclude regions of other cells, non-cell area, or unexpected artifacts which were present in some images outside of the cell area. After cropping, images were placed on frames with a determined preset size, which was chosen to fit the biggest cell expected among the analyzed cell subtypes: 13x13 μm, which was equivalent to 200x200 pixels. On this step the centers of the cells were also matched to the centers of the frames. Pretreated images are shown in Fig. 1 *c,d*.

D. Features extraction

As quantitative features which can be used to describe the morphology of cell images, the complex-valued pseudo-Zernike moments where chosen. However, the position of each individual cell on a slide is random and it is necessary to operate with rotationally independent features. Since the phase of the moment is related to the angular coordinate within the image plane, complex-valued moments were converted to absolute PZ-moments and then normalized to the zero-order moment. Therefore, invariants, which are not dependent on the image rotation and scale, were produced. These invariants skip all information about the phase (angular coordinate), and thus, the obtained variables are independent of the image rotation.

Unfortunately, as it is shown above in the "Materials and methods" section, the calculation of PZ-moments requires a double integration of a two-dimensional function which is a costly CPU process. Because the pre-computed images were transferred to a frame with a preset size, the algorithm for the PZ- moment calculation can be simplified. Instead of the integration, the sum of a scalar product of the image with a pre-computed complex matrix can be used. The matrices related to each moment can be generated once and then stored on a hard disk drive for the further use.

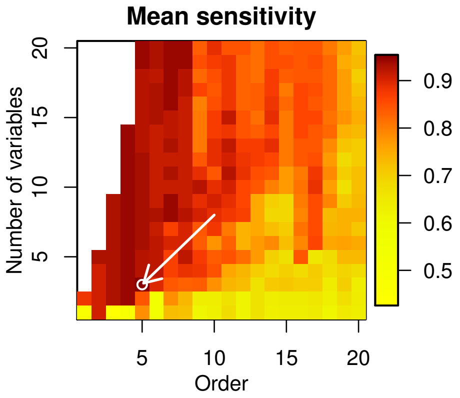


Fig. 3. Mean sensitivity of SVM leave-batch-out-cross-validation of training data. Classification models were created for a different number of selected orders of moments and for a different number of principal components (called variables in the image). The maximum value, which is related to the optimal model, is indicated with a white arrow.

TABLE I
CONFUSION TABLE FOR THE LEAVE-BATCH-OUT-CROSS-VALIDATION OF THE SVM MODEL WITH OPTIMAL VARIABLE SELECTION.

		Predicted		
		Lymphocytes	**Neutrophils**	**Sensitivity**
True	**Lymphocytes**	26	2	0.893
	Neutrophils	1	44	0.978

E. Statistical model establishment and evaluation

To avoid an overfitting of the statistical model, the dimensionality of the data was reduced. A dimension reduction was obtained via a principal component analysis (PCA). The dimensionality of the retaining data set was optimized based on a leave-batch-out-cross-validation of the training data set. The parameter intervals checked for the feature extraction was 1 to 20 for orders, while repetition was chosen maximal. The score dimension of the PCA was evaluated from 1 to 20. For each parameter set the model performance was estimated based on the mean sensitivity. These values are summarized in plot Fig 3. The maximal sensitivity is marked on the plot with an arrow. This parameter set defines the optimal combination of input variables (3 principal components, based on PZ-moments up to 5th order). The model trained with these parameters was further analyzed and visualized. In table I a confusion table of training data cross-validation is given. In Fig. 4 a histogram of its probability scores, which represents SVM decision values rescaled to the unit range, is plotted.

F. Blind prediction

Model validation was performed by applying the established model to the independent data, which contained 163 microscopic images of stained WBCs. All preprocessing and feature extracting steps were performed on these unlabeled

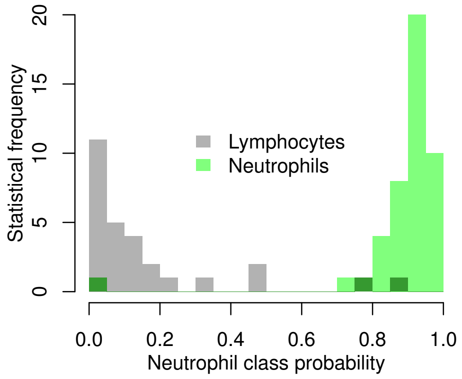


Fig. 4. Histogram for SVM posterior probabilities calculated by a leave-batch-out-cross-validation of the training data with the optimal number of variables are shown. Classification was performed between lymphocytes (gray bars) and neutrophils (green bars). The overlap of the groups is indicated within the histogram by dark green bins.

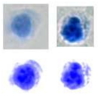
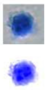
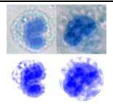
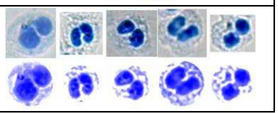
images in the same way as for the training data. In order to avoid the influence of the operator's subjectivity, a double blind prediction was carried out. Images were classified in manual mode by an experienced pathologist independently from the automated prediction. Subsequently, the statistical predictions were compared with the manual classification results. A summary of the results is visualized by a confusion table (see table II). Another representation of the classification performance is shown by means of a ROC curve in Fig 5. This curve, built for the threshold of the SVM decision values of the test data prediction, illustrates the high performance of the prediction. Moreover, the area under ROC curve (AUC) is about 0.984, which indicates an almost perfect classification. A perfect binary classification is characterized by an AUC equal to 1. Among 155 cells, which were classified as lymphocytes or neutrophils in manual mode, three images were wrongly identified by the statistical model. Such a low misclassification rate of independent test data corresponds to a high accuracy of the 2-class prediction. This accuracy was higher as 97%. Additionally, cells of the subtypes, which were not included in the training set, were present in the test data. These cells (five eosinophils and two monocytes), were predicted within the same class as neutrophils. This behavior was expected, since they feature a similar morphology as neutrophils compared to lymphocytes. Additionally, neutrophils, eosinophils, and monocytes feature a higher biological similarity and higher subjective similarity of the images. These classification results of the eosinophils and monocytes indicate that an extension of the presented model may be possible. A hierarchic layout of the classification seems optimal to incorporate eosinophils and monocytes.

IV. CONCLUSION

In this work, we presented an algorithm for a highly efficient classification between two dominant subtypes of leukocytes. The special feature of the proposed method is that by means

TABLE II

CONFUSION TABLE FOR THE PREDICTION OF THE UNLABELED TESTING DATA. CORRECT PREDICTED CELLS ARE SPECIFIED ONLY WITH THE QUANTITY OF THE IDENTIFIED CELLS. ALL INCORRECTLY PREDICTED CELLS AND CELLS, THAT RELATE TO OTHER SUBTYPES, WHICH WERE NOT INCLUDED IN THE TRAINING DATA, ARE SHOWN IN THE TABLE AS UNTREATED MICROSCOPIC (UPPER ROWS) AND PREPROCESSED (BOTTOM ROWS) IMAGES.

		Predicted (assorted by statistical model)	
		Lymphocytes	Neutrophils
True (assorted by pathologist)	Lymphocytes	17	
	Neutrophils		101
	Monocytes	0	
	Eosinophils	0	

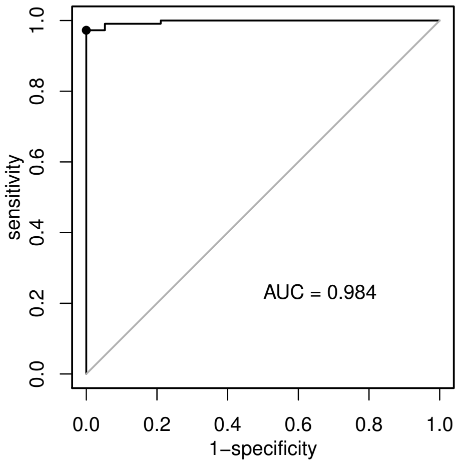


Fig. 5. The ROC curve and area under the curve (AUC) illustrate the high performance of the SVM prediction of the binary classification model between two WBC subtypes (lymphocytes and neutrophils) for independent unlabeled testing data.

of PZ-invariants the cell morphology is represented as a quantitative marker for the cell subtypes. Therefore, a combination of such common statistical methods as principal component analysis and support vector machine classification was applied to build the classification model. This approach showed a high stability against patient to patient and sample to sample variations. Moreover, an advanced image preprocessing made a further contribution to the robustness of the model. The standardization of the images decreased deviations, which occur between samples due to the sample preparation routine. Additionally, the automated framing and centering of the analyzed images of cells led to the replacement of the double numerical integration, performed for PZ-moment calculation, with a matrix product. This simplification of the calculation procedure resulted in the reduction of computation time and allowed the analysis to be performed in real-time. The classification results showed that WBCs subtypes as monocytes and eosinophils (which were not included in the model due to their low quantity in the training data) were predicted within the same class. Due to this fact, it can be assumed that the classification can be improved and extended to other cell types by a multilevel model. However, that requires a statistically significant amount of microscopic images for each leukocyte subtype in the training data set. The described approach can be applied for microscopy images taken of other staining types. Only important is that the images display the cell morphology. The method presented here may be also applied to images obtained with techniques such as fluorescence, dark field, or phase contrast microscopy.

ACKNOWLEDGMENT

Financial support of the BMBF via the integrated research and treatment center “Center for Sepsis Control and Care” (FKZ 01EO1502) and from the EU via the project “Hemo-Spec” (FP 7, CN 611682) is highly acknowledged. The authors thank Katrin Ludewig and Frank Brunkhorst for data collection and their support.

REFERENCES

[1] A. Kratz, M. Ferraro, P. M. Sluss, and K. B. Lewandrowski, “Normal reference laboratory values,” *New England Journal of Medicine*, vol. 351, no. 15, pp. 1548–1563, 2004. doi: 10.1056/NEJMcpc049016 PMID: 15470219. [Online]. Available: http://dx.doi.org/10.1056/NEJMcpc049016

[2] A. Ramoji, U. Neugebauer, T. Bocklitz, M. Foerster, M. Kiehntopf, M. Bauer, and J. Popp, “Toward a spectroscopic hemogram: Raman spectroscopic differentiation of the two most abundant leukocytes from peripheral blood,” *Analytical Chemistry*, vol. 84, no. 12, pp. 5335–5342, 2012. doi: 10.1021/ac3007363 PMID: 22721427. [Online]. Available: http://dx.doi.org/10.1021/ac3007363

[3] S. Khan, A. Khan, F. S. Khattak, and A. Naseem, “An accurate and cost effective approach to blood cell count,” *International Journal of Computer Applications*, vol. 50, no. 1, 2012. doi: 10.5120/7734-0682. [Online]. Available: http://dx.doi.org/10.5120/7734-0682

[4] L. Putzu and C. Di Ruberto, “White blood cells identification and counting from microscopic blood image,” *International Journal of Medical, Health, Biomedical, Bioengineering and Pharmaceutical Engineering*, vol. 7, no. 1, pp. 20 – 27, 2013. [Online]. Available: http://waset.org/Publications?p=73

[5] M. M. G. Bhamare and D. Patil, “Automatic blood cell analysis by using digital image processing: A preliminary study,” in *International Journal of Engineering Research and Technology*, vol. 2, no. 9, ESRSA Publications. ESRSA Publications, 2013. [Online]. Available: http://www.ijert.org/view-pdf/5460/

[6] F. Sadeghian, Z. Seman, A. R. Ramli, B. A. Kahar, and M.-I. Saripan, “A framework for white blood cell segmentation in microscopic blood images using digital image processing,” *Biological procedures online*, vol. 11, no. 1, pp. 196–206, 2009. doi: 10.1007/s12575-009-9011-2. [Online]. Available: http://dx.doi.org/10.1007/s12575-009-9011-2

[7] R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification,” *Systems, Man and Cybernetics, IEEE Transactions on*, vol. SMC-3, no. 6, pp. 610–621, 1973. doi: 10.1109/TSMC.1973.4309314. [Online]. Available: http://dx.doi.org/10.1109/TSMC.1973.4309314

[8] M. Habibzadeh, A. Krzyżak, and T. Fevens, *Comparative study of feature selection for white blood cell differential counts in low resolution images*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2014, vol. 8774, book section 20, pp. 216–227. ISBN 978-3-319-11655-6. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-11656-3_20

[9] T. Xia, H. Zhu, H. Shu, P. Haigron, and L. Luo, “Image description with generalized pseudo-Zernike moments,” *Journal of the Optical Society of America A*, vol. 24, no. 1, pp. 50–59, 2007. doi: 10.1364/JOSAA.24.000050. [Online]. Available: http://josaa.osa.org/abstract.cfm?URI=josaa-24-1-50

[10] S. O. Belkasim, M. Shridhar, and M. Ahmadi, “Pattern recognition with moment invariants: A comparative study and new results,” *Pattern Recognition*, vol. 24, no. 12, pp. 1117–1138, 1991. doi: 10.1016/0031-3203(91)90140-Z. [Online]. Available: http://www.sciencedirect.com/science/article/pii/003132039190140Z

[11] C. Kan and M. D. Srinath, “Invariant character recognition with Zernike and orthogonal Fourier-Mellin moments,” *Pattern Recognition*, vol. 35, no. 1, pp. 143–154, 2002. doi: 10.1016/S0031-3203(00)00179-5. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320300001795

[12] C. W. Chong, P. Raveendran, and R. Mukundan, “The scale invariants of pseudo-Zernike moments,” *Pattern Analysis and Applications*, vol. 6, no. 3, pp. 176–184, 2003. doi: 10.1007/s10044-002-0183-5. [Online]. Available: http://dx.doi.org/10.1007/s10044-002-0183-5

[13] Y.-H. Pang, A. T. B. J, and D. N. C. L, “Enhanced pseudo Zernike moments in face recognition,” *IEICE Electronics Express*, vol. 2, no. 3, pp. 70–75, 2005. doi: 10.1587/elex.2.70. [Online]. Available: http://dx.doi.org/10.1587/elex.2.70

[14] E. Walia, C. Singh, and N. Mittal, “Discriminative Zernike and pseudo Zernike moments for face recognition,” *Int. J. Comput. Vis. Image Process.*, vol. 2, no. 2, pp. 12–35, 2012. doi: 10.4018/ijcvip.2012040102. [Online]. Available: http://dx.doi.org/10.4018/ijcvip.2012040102

[15] J. Haddadnia, M. Ahmadi, and K. Faez, “An efficient feature extraction method with pseudo-Zernike moment in rbf neural network-based human face recognition system,” *EURASIP Journal on Advances in Signal Processing*, vol. 2003, pp. 890–901, 2003. doi: 10.1155/s1110865703305128. [Online]. Available: http://dx.doi.org/10.1155/s1110865703305128

[16] T. Bocklitz, E. Kämmer, S. Stöckel, D. Cialla-May, K. Weber, R. Zell, V. Deckert, and J. Popp, “Single virus detection by means of atomic force microscopy in combination with advanced image analysis,” *Journal of Structural Biology*, vol. 188, no. 1, pp. 30 – 38, 2014. doi: 10.1016/j.jsb.2014.08.008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1047847714001841

[17] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: https://www.R-project.org/

[18] B. Rajwa, M. Dundar, A. Irvine, and T. Dang, *IM: Orthogonal Moment Analysis*, 2013, R package version 1.0. [Online]. Available: https://CRAN.R-project.org/package=IM

[19] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2015, R package version 1.6-7. [Online]. Available: https://CRAN.R-project.org/package=e1071

[20] Revolution Analytics and S. Weston, *foreach: Provides Foreach Looping Construct for R*, 2015, R package version 1.4.3. [Online]. Available: https://CRAN.R-project.org/package=foreach

[21] ——, *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*, 2015, R package version 1.0.10. [Online]. Available: https://CRAN.R-project.org/package=doParallel

[22] S. Urbanek, *jpeg: Read and write JPEG images*, 2014, R package version 0.1-8. [Online]. Available: https://CRAN.R-project.org/package=jpeg

[23] C. H. Teh and R. T. Chin, "On image analysis by the methods of moments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, no. 4, pp. 496–513, Jul 1988. doi: 10.1109/34.3913. [Online]. Available: http://dx.doi.org/10.1109/34.3913

[P3] Raman based molecular imaging and analytics: a magic bullet for biomedical applications!?

Reprinted with permission from [Bocklitz, T.W., Guo, S., Ryabchykov, O., Vogler, N. and Popp, J., 2015. Raman based molecular imaging and analytics: a magic bullet for biomedical applications!?. *Analytical chemistry*, *88*(1), pp.133-151] Copyright 2016 American Chemical Society.

Erklärungen zu den Eigenanteilen des Promovenden sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an der Publikation.

[1]Bocklitz, T.W., [2]Guo, S., [3]Ryabchykov, O., [4]Vogler, N., [5]Popp, J., 2015. Raman based molecular imaging and analytics: a magic bullet for biomedical applications!?. *Analytical chemistry*, *88*(1), pp.133-151					
Beteiligt an (Zutreffendes ankreuzen)					
	1	2	**3**	4	5
Konzeption des Forschungsansatzes					
Planung der Untersuchungen					
Datenerhebung					
Datenanalyse und Interpretation					
Schreiben des Manuskripts	X	X	**X**	X	X
Vorschlag Anrechnung Publikationsäquivalent			**0.5**		

analytical chemistry
Review
pubs.acs.org/ac

Raman Based Molecular Imaging and Analytics: A Magic Bullet for Biomedical Applications!?

Thomas W. Bocklitz,*[†,‡] Shuxia Guo,[†,‡,§] Oleg Ryabchykov,[†,‡,§] Nadine Vogler,[†,‡,§]
and Jürgen Popp*[†,‡,§]

[†]Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich Schiller University Jena, Helmholtzweg 4, 07743 Jena, Germany

[‡]Leibniz Institute of Photonic Technology (IPHT), Albert-Einstein-Strasse 9, 07745 Jena, Germany

[§]InfectoGnostics Forschungscampus Jena e.V., Zentrum für Angewandte Forschung, Philosophenweg 7, 07743 Jena, Germany

CONTENTS

Instrumentation 134
Computational Methods 135
  Pretreatment Methods 135
  Analysis Techniques 137
    Clustering 138
    Classification and Regression Methods 139
    Unmixing: Determining the Mixture Composition 140
Applications 140
  Tissue Imaging and Diagnostics 140
    Tissue Imaging 140
    Tissue Diagnostics 141
    New Trends and Developments 141
  Cell Imaging and Diagnostics 144
    Cell Imaging 144
    Classification of Cells 144
    Cellular Drug Response 145
    Recent Developments for Cell Studies 146
Summary and Conclusion 146
Author Information 147
  Corresponding Authors 147
  Notes 147
  Biographies 147
Acknowledgments 147
References 147

The Raman effect was predicted by Schmekal[1] in 1923 and independently discovered in 1928 by two Indian physicists, Raman and Krishna.[2,3] In principle, monochromatic light is inelastically scattered at a quantified structure like the vibrational states of a molecule. The occurring energy shifts are an indirect representation of the vibrational states of the molecule and, thus, are molecule specific. If this principle is spectroscopically used, an ensemble of molecules is measured and the result is called a Stokes-Raman spectrum, or shorter a Raman spectrum. The Stokes-Raman spectrum is the part of inelastically scattered light, which is shifted to lower energies.[4,5] This is the dominant effect at room temperatures, which is the reason for skipping the attribute. Because of the ensemble mixing, the Raman spectrum is not representing the vibrational states of one molecule but of a mixture of molecules. Thus, the Raman spectrum is a superposition of Raman spectra of substances within the excitation focus. Because the unmixing of this superposition is only possible for limited cases, the Raman spectrum is used as a vibrational fingerprint. This fingerprint is either interpreted with a certain set of reference Raman spectra or evaluated by means of statistical methods. The latter procedure is often applied, if heterogonous mixtures like cells or tissues are investigated, while the former method is used, if pure substances or easy mixtures are studied. As investigations on biological samples, like cells or tissue are the topic of the review, we will focus on biological samples in the following. Therefore, a Raman spectrum is used as vibrational fingerprint.

In the same manner like for the human fingerprint, databases are essential in order to carry out the interpretation. In contrast to the specific human fingerprint, the specificity of the Raman fingerprint is reduced due to the ensemble mixing and other corrupting effects. These both factors complicate the creation of Raman spectroscopic databases for biological samples. In order to deal with corrupting effects, advanced computational methods can be applied, which correct for the influence of the measurement device or background contributions within the Raman spectra.[4] After, these corrections are carried out, the databases are not constructed directly but with the help of multivariate statistical methods[6] or other chemometrical techniques.[7] These computational methods showed better performance compared with a pure database search.

Beside the positive properties of the Raman effect, like the discussed molecule specificity, the linearity with the concentration and laser power and the insensitivity to water, it took almost 50 years that Raman spectroscopy was applied to biological questions. The reason for this long time is that the Raman effect is a rather weak effect, which had to be compensated by a development of the Raman spectroscopic instrumentation and analysis software. In order to apply Raman spectroscopy for real world application, like analysis of cells and tissue, the most important of these developments are the laser (light amplification by stimulated emission of radiation),[8] which provided high excitation intensities, the charge-coupled-device (CCD),[9,10] which allowed a simultaneous detection of a certain wavenumber region, the personal computer (PC), which allowed for the analysis of larger data sets, and the development

Special Issue: Fundamental and Applied Reviews in Analytical Chemistry 2016

Published: December 10, 2015

ACS Publications © 2015 American Chemical Society

DOI: 10.1021/acs.analchem.5b04665
Anal. Chem. 2016, 88, 133−151

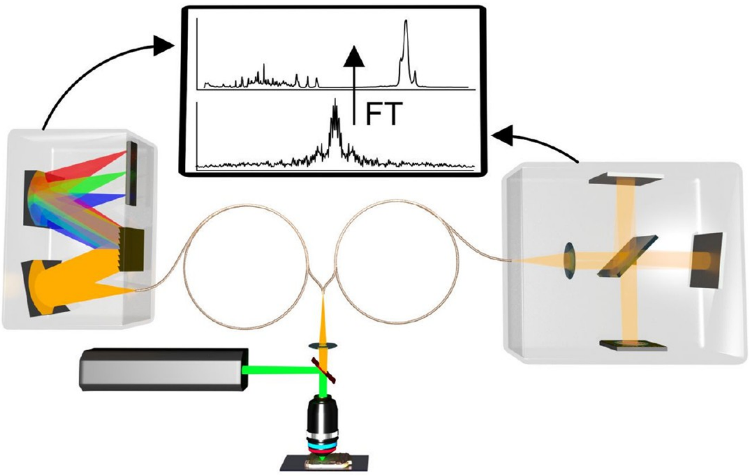


Figure 1. Schematic sketch of a dispersive (left gray box) and nondispersive (right gray box) Raman instrument. Monochromatic light is focused onto a sample, collected via the objective, filtered to exclude the excitation light, and detected. On the left, the light is dispersed using a grating and detected with a camera. The Raman spectrum is directly obtained. On the right, the light passes a Michelson interferometer and the interferogram is detected with a camera. The spectrum has to be recalculated from the interferogram using a (fast) Fourier transformation. Copyright 2015, Sandro Heuke/IPHT.

of interference filters[11,12] to suppress the Rayleigh scattered light.

Around 1970, these developments allowed the scientists to apply Raman spectroscopy in order to investigate biomolecules like proteins, lipids, and nucleic acids. Since these first investigations, a subdiscipline was created, which deals with the application of Raman spectroscopy for biological or medical tasks. The so-called biomedical Raman spectroscopy is fast growing and further developing. In 1977, the first review articles about protein analytic based on Raman spectroscopy was published.[13] In the 80s a first review of the applications of Raman spectroscopy to biological tasks was published,[14] in the 90s biomedical relevant Raman based analytics was carried out,[15] and at the end of this queue of review articles ref 16 showed a large number of publications dealing with Raman based cancer or cancer cell diagnostics.

Since then the field of biomedical Raman spectroscopy is further emerging and the current developments are the subject of this review. Within this review we like to emphasize this recent development and we will focus on the past 5 years. Nevertheless, we also included books and older literature, where we think these publications are an appropriate source of background information. So this review can be used as an introduction to the wide field of biomedical Raman spectroscopy but also as a review of the current development in the field. Within this contribution we will focus on linear Raman spectroscopy. Where it make sense, the link to other methods like coherent Raman techniques is marked and reviews in the respective field are cited.

This review article is structured as follows. In the beginning the instrumentation of typical Raman spectroscopic devices is summarized and the recent developments are set together. In the next section, the necessary spectral pretreatment and the used statistical methods are reviewed. Thereafter, a section for the application of Raman spectroscopy to biomedical tasks follows. This section is further separated into cell imaging and diagnostics and tissue imaging and diagnostics. At the end the article will be summarized and the future developments and applications are outlined and discussed.

■ INSTRUMENTATION

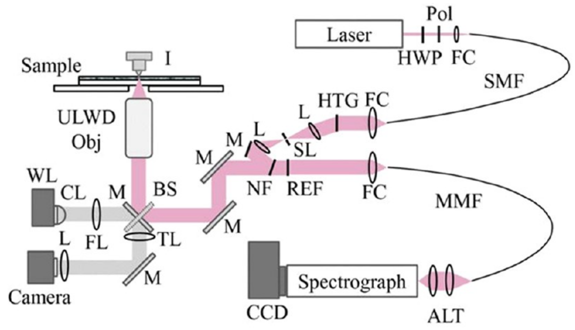
Raman spectroscopic devices typically share a set of components because the setup has to deliver light of a certain wavelength to the sample, the scattered light has to be collected from the sample after the interaction, and finally, the scattered light has to be detected. There is a huge variety of instruments available, starting with commercial systems, home-built laboratory solutions, as well as integrated systems combining Raman spectroscopy with other imaging techniques. Depending on the samples to be analyzed, the Raman spectroscopic system has to be carefully chosen. In the following, some of the criteria will be discussed in more detail, providing a brief overview of available systems with their respective advantages and disadvantages.

A foremost consideration involves the excitation source, which in the case of Raman spectroscopy is a highly monochromatic light source providing sufficient power to detect the Raman spectrum in a reasonable time frame. The most common laser types and excitation wavelengths used for Raman spectroscopy cover the visible and near-infrared range including Ar^+ (e.g., 488 and 514.5 nm), He−Ne (632.8 nm), Nd:YAG (1064 and 532 nm), as well as diode lasers. The latter laser type operates in a range of more than 200 nm starting from about 630 nm.[17,18] As mentioned above the sample dictates the wavelength range for the experiment based on its fluorescence properties. However, not only the sample might exhibit fluorescence, the sample matrix or the substrate might also contribute to the fluorescence background disturbing or masking the Raman spectrum of the sample. Several approaches have been introduced to minimize or avoid fluorescence contributions including excitation in the near-infrared, enhancement of the Raman signal (e.g., resonance Raman, surface or tip-enhanced Raman scattering[19]), temporal gating to separate the Raman signal from the fluorescence, and many others.[20,21] In general, continuous-wave and pulsed lasers are used. The latter are usually implemented in combination with a gated detection[22] or time-resolved investigations.[23]

DOI: 10.1021/acs.analchem.5b04665
Anal. Chem. 2016, 88, 133−151

The selection of a suitable detection system is another crucial point in Raman spectroscopy. The spectral resolution achieved by the whole spectroscopic system depends on spectrometer parameters such as the focal length of the spectrometer, the diffraction grating (in dispersive Raman systems), and the pixel size of the detector.[18] A huge step forward has been done with the invention of the CCD camera in 1970,[9] which allowed for higher sensitivities and faster acquisition times. In principle, Raman spectrometer can be divided into dispersive and nondispersive (FT-Raman) systems, which are both schematically sketched in Figure 1. Both systems require a light source and a collection system indicated by the microscope objective. A filter blocks the excitation light, and the Raman signal is guided to the detection system after interaction with the sample. In Figure 1, we used fibers to represent the optical setup. On the left side, a dispersive Raman spectrometer is shown. The light is dispersed on a grating, and the accordingly separated wavelengths can be detected.[24] Using a multidimensional camera every wavelength can be detected in a separate pixel/channel, whereas a single detector such as a photomultiplier tube or photodiode would require a scanning of the dispersed light. With this system, the Raman spectrum is obtained directly. In contrast, on the right side of Figure 1 a FT-Raman system is illustrated. FT-Raman spectrometers require an interferometer: in the figure a Michelson interferometer has been used. This type of interferometer is equipped with a moving mirror changing the path length between the two beams and therefore causing constructive and destructive interference. Suppliers, however, invented a number of interferometers to minimize the setups and optimize the performance of their devices. The resulting interferogram is detected, usually with a camera, and the spectrum is calculated using a (fast) Fourier transformation. Compared to a dispersive instrument the frequency can be detected with higher precision. In addition, FT-Raman spectrometers are preferably used in the NIR due to the comparably low or completely avoided fluorescence in that spectral region and the low performance of the detectors in dispersive instruments when using this excitation wavelength. Dispersive instruments always force a trade-off between spectral coverage and resolution with a varying resolution along the spectrum. In contrast, FT-Raman spectra provide the full spectral coverage with a constant resolution. However, this might also be a disadvantage depending on the measurement task. Concerning SNR, dispersive instruments show a higher signal-to-noise ratio compared with nondispersive systems.[25−28]

By now the excitation source has been discussed as well as the detection side. In between the light has to be guided to the sample and from the sample to the detector. A common concept to achieve this while reducing unwanted signals and at the same time increasing the axial resolution is Raman microspectroscopy using confocal microscopes.[29] Confocal setups are characterized by focusing the laser beam onto basically two small apertures. The first aperture is placed directly after the light source. This creates a diffraction-limited light source. The second aperture is placed directly in front of the detector. The light can then pass the entrance slit of the detector with minimal loss and can be split into the different wavelengths to generate the spectrum. To focus the light onto the sample and collect the backscattered light a high numerical aperture (NA) objective is used in order to achieve small focal volumes and on the other hand to collect as much of the backscattered light as possible. Figure 2 shows a Raman microspectroscopic setup which has been modified to allow for in situ studies of Raman active transparent bulk materials, thin films, or fibers which underwent mechanical deformation.[30] The multimode fiber (MMF) acts as the second aperture in front of the detector. The first aperture is present as a slit (SL). As this device allowed for the acquisition of point spectra only, Gerbig and co-workers presented an advancement by coupling the device with a laser scanning microscope.[31] The authors could follow the evolution of strain fields and were able to observe changes in the phase distributions of a material while they performed a compression.



ALT … achromatic lens telescope, BS … beam splitter, CCD … charge couple device, CL … collecting lens, FC … fiber coupler, FL … field lens, HTG … holographic transmission grating, HWP … half-wave plate, I … indenter, L … lens, M … mirror, MMF … multimode fiber, NF … notch filter, Pol … polarizer, REF … Raman edge filter, SL … slit, SMF … single-mode fiber, TL … tube lens, YULWD Obj … ultra long working distance objective, WL … white light source

Figure 2. Schematic view of the optical part of the Raman instrument used for Raman microprobing is plotted.[30] The system also allows for white light imaging. Reprinted with permission from Gerbig, Y. B.; Michaels, C. A.; Forster, A. M.; Hettenhouser, J. W.; Byrd, W. E.; Morris, D. J.; Cook, R. F. *Rev. Sci. Instrum.* **2012**, *83*, 125106 (ref 30). Copyright 2012, AIP Publishing LLC.

Many technical developments have been made in the meanwhile, which cannot be discussed in this manuscript. This includes the developments in sensor elements, e.g., for time-of-flight experiments,[32] which might be adapted in the near future. Further a miniaturization takes place while maintaining all functionalities of a large systems.[33] Investigations on the optimization of existing instruments are ongoing in order to use different wavelengths.[34] Also other setup changes and variations such as a wide-field imaging possibilities were described recently.[35]

■ COMPUTATIONAL METHODS

The data pipeline to analyze Raman spectra is sketched in Figure 3, and the necessary procedures can be grouped into data pretreatment and analytical modeling. The pretreatment aims at the correction of certain artifacts and side effects. This group of methods consists of spike correction, spectrometer calibration, and preprocessing, like baseline correction, smoothing, normalization, and dimension reduction. These methods are described in the subsection Pretreatment Methods. The analytical model and its evaluation will be described in the subsection Analysis Techniques.

Pretreatment Methods. After the Raman spectra of biological samples are measured by Raman spectroscopy chemometrical steps are required in order to pretreat the Raman spectra. This is necessary as besides the Raman

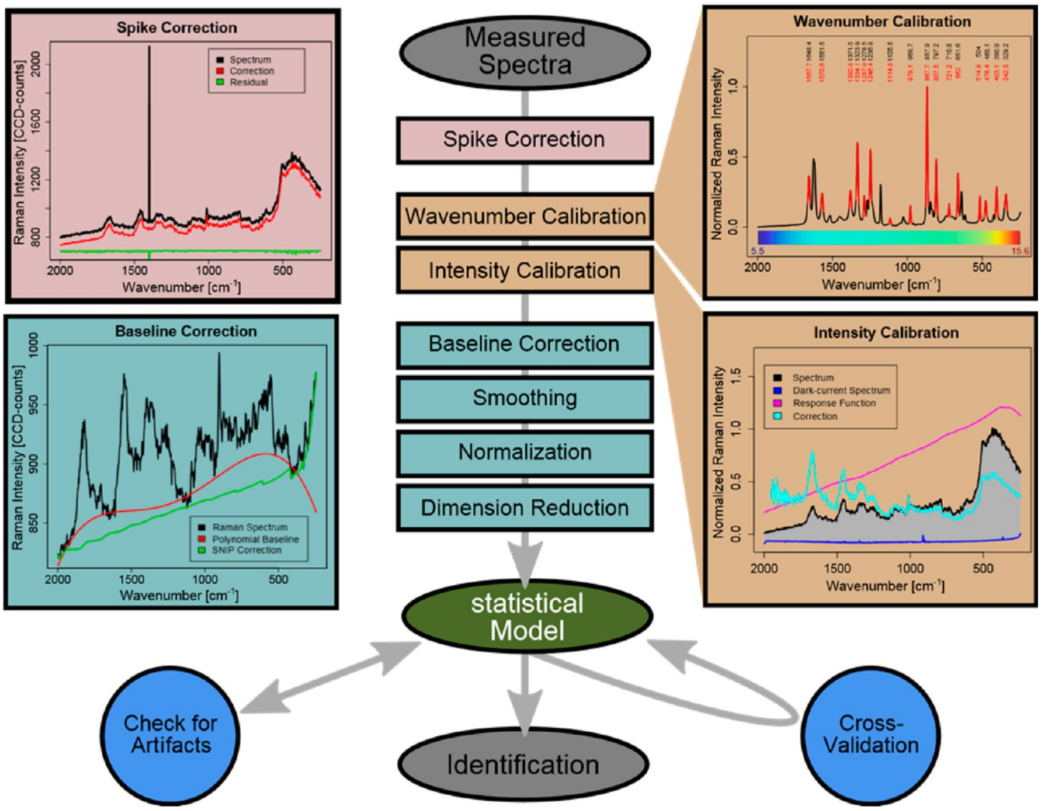


Figure 3. Data pipeline for the analysis of the Raman spectra is sketched. The measured Raman spectra are filtered for cosmic noise, spectrometer calibration procedures are applied, and the spectra are preprocessed. This preprocessing consists of a baseline correction, smoothing, normalization, and dimension reduction and reflects the composition of Raman spectra (see Figure 4). With these pretreated spectra, a statistical model is constructed, which is evaluated and checked for artifacts, before it is used to predict independent data. Copyright 2015, Thomas Bocklitz/Friedrich-Schiller-University Jena.

spectrum other contributions are measured. These contributions corrupt and often overwhelm the useful Raman information and a robust analysis is not possible. Thus, corrections for these contributions have to be carried out. The most disturbing contributions within Raman spectra originate from cosmic spikes, the fluorescence background, Gaussian or Poisson noise, and other contributions caused by experimental parameters (Figure 4). In order to allow for further statistical analysis, preprocessing, like spike correction, wavenumber and intensity calibration, baseline correction, and normalization, have to be carried out to obtain uncontaminated data.[36−39]

Like sketched in Figure 3, the pretreatment starts with a cosmic spike removal. One method often applied for the spike correction is done by collecting two Raman spectra of the sample or position during the experiment followed by a pixel-to-pixel comparison keeping the smaller intensity count.[40] Besides this approach, several mathematical methods have been explored and widely applied.[40,41] Filtering, such as polynomial and median filters, can be used in cases where the spike is sharper compared with real Raman bands. Particularly, for Raman spectroscopic imaging, information from unaffected neighbor pixels can be used. Accordingly, refined nearest neighbor comparison methods and upper bound spectrum method have been developed.[40,42,43] Principal component analysis (PCA) and wavelet transform are also used for this purpose.[40,44] Recently developed algorithms determine the spike position based on the evaluation of the second derivative.[45]

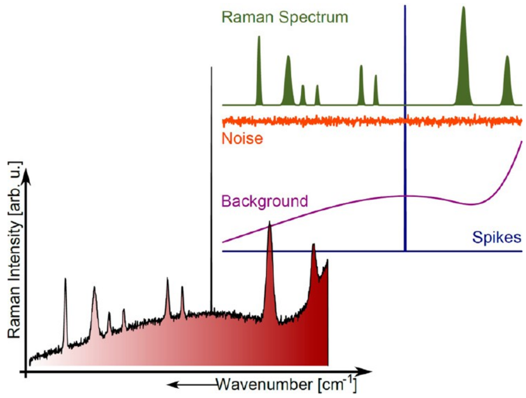


Figure 4. Composition of Raman spectra. The measured Raman spectra are suffering from different side effects, like fluorescence background, cosmic spikes and white noise. All contributions have to be rejected prior the analysis. Copyright 2015, Thomas Bocklitz/FriedrichSchiller-University Jena.

The next steps within the pretreatment of Raman spectra are the wavenumber, wavelength, and spectrometer calibration. These calibration steps aim to correct for wavenumber shifts caused by a drifting excitation wavelength and intensity changes resulting from environmental influences such as changes in the devices and temperature.[46−48] For wavenumber axis calibration, both absolute and relative wavenumber calibration can be applied. In both cases, a standard sample with known Raman bands or wavelength positions together with the exact excitation wavelength are required. The calibrated wavenumber axis is calculated according to the differences between true and reported Raman bands of the standard sample[46,47,49] in the case of a relative wavenumber calibration. In the case of an absolute calibration, the differences of true and measured wavelength positions are recalculated with the excitation wavelength to receive a calibrated wavenumber axis. In a recent study the quantum efficiency of the process itself was incorporated.[48] Besides these physical related methods, a method based on evolution theory was proposed for wavenumber alignment.[50] For intensity calibration, a standard sample with known response over wavelength or wavenumber is required. The intensity response curve of the instrument is computed as the ratio of measured and reported (Raman) intensity of the standard sample or calibration lamp. Afterward, the intensity values of the spectrometer can be recalculated.[47,51] This can be seen as unit transfer from electrons counted to photons counted. The intensity calibration is not widely used, but important, if the calculated model has to be applied to data of another spectrometer. In references 47 and 52, the comparison of different spectrometer devices have been investigated. It turned out that a fully set up independent Raman spectrum is hard to achieve and such a calibration procedure remains the scope of ongoing research.

Baseline removal is a preprocessing step, which has huge influence on the further analysis, because the baseline is often a few orders of magnitude more intense compared to Raman bands. It may easily hinder further data analysis.[53] Quite a number of mathematical methods exists, all try to estimate a part of the spectra, which varies slowly as a function of wavenumber assuming that this is the background contribution. These methods are implemented using different mathematical procedures. In references 54 and 55, a wavelet transform is applied, while in refs 56 and 57, morphological operations are utilized. Beak et al.[58] propose a combination of peak detection and interpolation to estimate the background contribution. Other approaches may involve frequency-domain filtering[59] or polynomial fitting.[53,60] Recent developments involve asymmetric cost functions for baseline fitting.[61,62] However, none of the procedures or methods work equally well for all data sets. Thus, all methods should be used with caution and spectroscopic knowledge is important to verify the baseline estimation result. As demonstrated by Emry et al., an unreasonable baseline correction can introduce errors and lead to misidentifications of substances.[63] To tackle this issue, methods for evaluating the quality of baseline estimations are needed. Emry et al. therefore proposed a method to judge the quality of a polynomial-based baseline correction algorithm.[63]

Smoothing is a step which can be performed before baseline estimation, after the baseline estimation step or it is implicitly done within the dimension reduction. If applied before the baseline estimation, the idea is that the baseline estimation is not depending on the noise level. For smoothing a lot of algorithms exist, like Savitzky−Golay filtration, finite impulse response modeling, wavelet transform, and Fourier transform based algorithms.[64] In reference 65, a new type of algorithm is developed, which is based on a spectral reconstruction by Wiener estimation. It is worth to mention that a noise reduction is also achieved using a factor based dimension reduction technique.

After the baseline estimation, smoothing is carried out. An intensity normalization is essential to remove the effect from different sample preparations and varying collection parameters. This is usually done by dividing each intensity value of the Raman spectrum by a constant value. Peak and vector normalization are widely used. In the former the maximal intensity of a defined peak or the integrated peak area are utilized as a normalization constant. In vector normalization, the 2-norm of the Raman spectrum is applied as the normalization constant.[36] Notably, it is demonstrated that different choices of Raman bands used for peak normalization, lead to rather different results. No single choice of Raman bands ensures an optimal result for all applications. Instead, analysts have to decide, which specific Raman band for normalization is reasonable for their application. Usually, a substrate or a solvent peak is used for normalization purposes.[66]

Finally, it was demonstrated in reference 37 that a high number of preprocessing combinations lead to a worse performance of a further analysis compared with doing no preprocessing. Thus, it is necessary to evaluate the quality of the preprocessing, which is far from straightforward. The "trial and error" method and a quality parameter based method are two possibilities for this purpose.[67] The former approach is often called model based and usually achieved by testing a subsequent model for its performance. The output of the model can be its sensitivity, specificity, or accuracy, if a classification model is tested, or the Root-Means-Squared Error of Prediction (RMSEP) for a regression model.[37] For the quality parameter based approach, there are several quality parameters existing for chromatographic, nuclear magnetic resonance (NMR) spectroscopy, and near-infrared (NIR) spectroscopy.[67,68] To the author's best knowledge, there are no such parameters applied in Raman spectral preprocessing.

At the end of this section, it is worth to mention that analysts have to be careful about the sequence of these aforementioned preprocessing steps. Currently, different sequences are being used depending on the idea a researcher has for the data at hand. There are no standardized preprocessing protocols, in order to compare the data output from different laboratories or devices. However, some sequences are definitely inappropriate, for example, applying the normalization before a baseline correction.[69] Developing Standard Operating Procedures (SOP) for Raman spectra of biological samples or for the Raman spectral analysis of a certain task or type of spectra should be investigated in the future.

Analysis Techniques. After the pretreatment of Raman spectra is carried out, statistical models are applied in order to extract relevant information, like concentrations of substances or disease markers. These methods try to translate the physical measured Raman spectra into higher level information, which can be further used by chemists, biologists, and physicians. Most of these methods have a statistical background, that is why they will be called statistical models within this contribution. The models, which are widely applied, do not use the spatial arrangement of the spectra. They only analyze the spectral information. Research about the incorporation of the spatial distribution within the analysis is in progress but should not be reviewed here. Instead we like to focus on the well-established methods.[6]

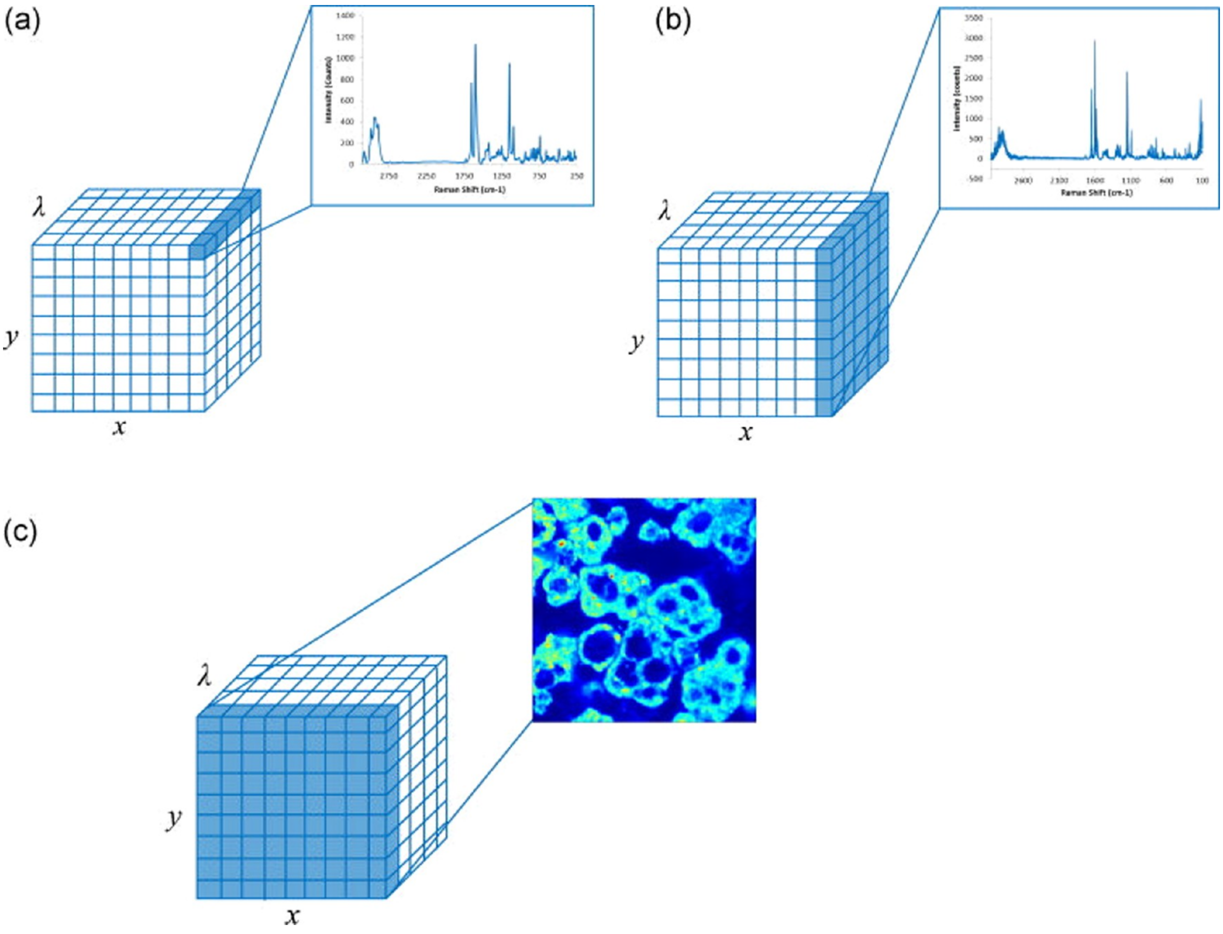


Figure 5. Hyperspectral data cube, which is a typical result of a Raman imaging experiment.[70] Such a hypercube can be analyzed in a number of ways. The easiest are given here. While panel a indicates the location and orientation of a spectrum within the data cube, panel b shows the spectral mean over a line in the *y*-direction. Panel c shows the distribution of a single intensity value over the *x*−*y*-plane. This procedure is the easiest way to generate an image from a hyperspectral data cube. Reprinted from *Journal of Pharmaceutical and Biomedical Analysis*, *101*, P.-Y. Sacré, C. De Bleye, P.-F. Chavez, L. Netchacovitch, Ph. Hubert, E. Ziemons, Data processing of vibrational chemical imaging for pharmaceutical applications, 123−140 (ref 70), Copyright (2014), with permission from Elsevier.

The statistical methods applied for the analysis of Raman spectra can be grouped according to different properties. Here we assort them into the groups, which are mostly used in the analysis of Raman spectra. We will introduce clustering algorithms, regression and classification models and unmixing models. From an abstract point of view, this separation is not consistent: The unmixing models can be seen as a fuzzy version of the clustering algorithms and some unmixing models share similarities with regression and classification models. However, from an application point of view, this separation is appropriate.

In order to understand advanced statistical methods for image generation and analysis of Raman spectra, we will start with the description of the easiest possibility to generate an image from a hyperspectral Raman data cube. To start with the description of the image generation techniques, the arrangement within a hyperspectral data cube has to be taken into account. The arrangement is visualized in Figure 5. While the spatial dimension are represented by the x and y direction, the wavelength/wavenumber dimension is drawn in the z direction. With that arrangement, a Raman spectra on position x,y is represented by the z direction (Figure 5a). A mean spectrum of a line in y can be generated by calculating the mean in an upward direction in Figure 5b. A false color image over one specific intensity values is a z-slice of the data cube. This is the simplest way to visualize the distribution of a species, where the intensity value can be attributed to. The visualized value can also be extracted by a peak fitting of a Raman peak of the species under investigation or the integration (sum) of a certain wavenumber region. These easy methodologies feature an advantage, they are computationally inexpensive and the interpretation is clear as expert knowledge was used to generate the model.

All further discussed methods are either statistically or mathematically motivated, and the computational complexity is increased compared with the easy univariate models described above. Nevertheless, also these models extract information, like group information or pseudoconcentrations, from every Raman spectrum and using these values for image generation. In the following, a short introduction together with a review of recent publications is given.

Clustering. Clustering algorithms can be divided into two major types: hard clustering and fuzzy cluster methods. As

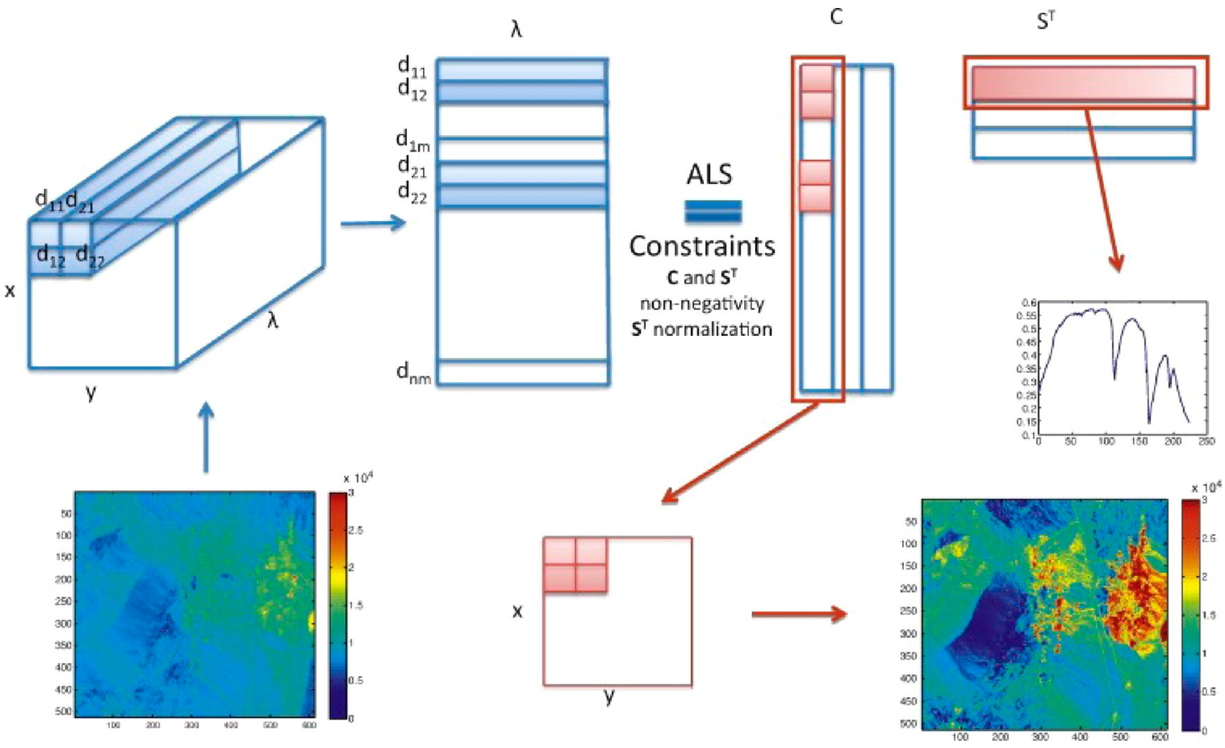


Figure 6. MCR-ALS algorithm is sketched.[116] The hyperspectral data cube is reoriented into a matrix and this matrix is decomposed into estimated component spectra and pseudoconcentrations. Additional knowledge can be incorporated by constraints to these matrices. The pseudoconcentrations can be used to do chemical imaging. Reprinted from *Analytica Chimica Acta*, *762*, Xin Zhang, Romá Tauler, Application of Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) to remote sensing hyperspectral imaging, 25−38 (ref 116), Copyright (2013), with permission from Elsevier.

fuzzy clustering will be discussed in the unmixing section, we only refer to hard cluster methods within this section. These hard cluster methods are widely applied in Raman spectroscopy in order to test if certain groups are reflected within the data or for imaging purposes. The reason is that these clustering procedures are simple and reliable methods for spectral analysis. Especially for imaging purposes, an unsupervised learning algorithm is ideal to produce an overview. Methods, which are commonly applied for imaging, are k-means clustering[71−73] and hierarchical clustering.[74−76] The first algorithm starts with a random cluster distribution of k clusters and then iteratively resorts the spectra according to their minimal distance to the mean spectra of the clusters. This is carried out until a stable arrangement is achieved. It is advisible to run the algorithm 10−100 times and then use the best cluster distribution. The hierarchical clustering exists in two versions: agglomerative and divisive clustering. While in the former, Raman spectra are merged to cluster until only one cluster exists at the end in the latter cluster are split until every spectrum is in its own cluster. The drawback of these methods are that they are not computational efficient and therefore a high processing power is required.

However, for big Raman spectral data sets, the complexity of high-dimensional clustering can be avoided by a subsequent unsupervised dimension reduction. For this purpose all feature extraction procedures or factor methods can be utilized.[6] The most often applied factor method is the principal component analysis (PCA),[77,78] but also other methods like partial least-squares regression (PLS) or similar methods are applied for dimension reduction. Besides these general algorithms, some algorithms are specifically adapted for large data sets and spectral clustering,[79,80] which includes the dimension reduction step as well.[81]

Classification and Regression Methods. If high level information should be extracted from the data, supervised machine learning algorithms, like regression or classification models, have to be applied. Among these, the linear classification methods and regression procedures are widely used by virtue of their simplicity and robustness.[78,82−85] Despite the efficiency of linear models, in many cases more powerful tools, like kernel support vector machines (SVMs) and (deep) artificial neural networks are required. SVMs are supervised learning models, which are basically linear classifier and regression models. With various kernels the SVM can be converted to a nonlinear classification and regression model.[86−93] Another example of powerful classification and regression algorithms are random forests (RF), which is an ensemble based method. RFs are based on the idea, which is related to automatic generated decision trees:[94−96] A predefined number of random decision trees are constructed and for prediction every tree is allowed to predict. The output of the whole random forest is generated by a voting procedure at the end. However, the high dimensionality of spectral data sets can cause problems due to long computational time and collinearity of variables. Like for clustering algorithms, these problems can be avoided by applying a dimension reduction, like a PCA before training a RF.

While the SVMs can be used intrinsically as a classification and regression model, pure regression models have to be converted to classification models by modeling pseudoconcentrations. Often applied regression models are principal component regression (PCR), which is based on PCA, and partial least-squares regression.[77,92] Despite the fact that these approaches produce higher errors compared with a SVM or RF, these models can be applied to avoid an additional step of dimension reduction and decrease model optimization time.[87] Common utilized partial least-squares regression methods are PLS[78,87,97−100] and interval PLS (iPLS), which is a modification of PLS that does not only decrease the data dimension during the analysis but also selects specific variables which give better prediction compared with the usage of all variables.[101]

Recently published methods for the supervised analysis of Raman spectra include methods, which combine certain analysis steps, incorporate preprocessing steps into the analytical model, or optimize the model in some regard. One approach is the localized feature selection, which uses an optimized feature selection for each spatial region of the sample[102] in order to describe the spatial distribution in an optimal way. Another feature selection based method is related to the Q-statistics.[103] A number of studies incorporate certain test statistics into the variable selection. In reference 104, the Kruskal−Wallis and Conover−Inman tests are incorporated into the feature selection and an optimized regression is carried out. Dual classification (DuC)[105] introduced by Lin et al. is insensitive to the noise and the presence of outliers in the training data set. The Constrained Optimization method based Extreme Learning Machine for Regression (CO-ELM-R)[106] is another method to optimize the result of a regression model. Besides improving feature selection and insensitivity to noise, hybridization of well-established methods is another way to optimize classification or regression models.[107]

Unmixing: Determining the Mixture Composition. An important task that cannot be easily covered by the described clustering, classification, and regression methods is the determination of mixture compositions, especially if no further information is available. If an appropriate training data set is existent, regression models can be applied. Nevertheless, for biological specimen this is often not the case. For this kind of analysis task unmixing methods are the ideal tool. One possibility to carry out unmixing is the application of a fuzzy version of the above-described clustering methods, for example c-means-clustering.[108] This approach will result in a cluster membership value for every spectrum, which can be interpreted together with the mean Raman spectrum of every cluster. Nevertheless, often a decomposition into pure spectra is desired. For this task the so-called end-member extraction methods are developed. They do not extract pure spectra but the most extreme spectra in a certain sense. Methods, which are commonly applied for end-member extraction, are N-FINDR[109,110] and Vertex Component Analysis (VCA).[109,111] Another technique, which allows the estimation of constituent spectra and pseudoconcentrations, is the multivariate curve resolution-alternating least-squares (MCR-ALS) method. This technique allows the use of additional knowledge and incorporates this information into the modeling. It estimates the chemical constituents or species of an unresolved mixture and then decomposes a spectrum of a mixture with respect to these estimated component spectra in an iterative manner.[82,112−115] In this iterative process constraints on the estimated component spectra and pseudoconcentrations can be incorporated. The MCR-ALS method is sketched in Figure 6.

■ APPLICATIONS

After we have reviewed the instrumentation and the analytical procedures necessary to obtain Raman spectra and Raman scans of biological specimens, we will focus in the next section on the application. A full comprehensive review of the application of Raman spectroscopy goes far beyond the scope of this review. Therefore, we restricted the topic to cell and tissue imaging and diagnostics based on Raman spectroscopy. In the section Cell Imaging and Diagnostics we will review recent studies on imaging and diagnostics of bacterial and eukaryotic cells. In the section Tissue Imaging and Diagnostics, we will focus on tissue based studies. As both topics are sometimes hard to separate, these sections are not meant to be a sharp separation.

Tissue Imaging and Diagnostics. Tissues provide abundant information about diseases and alterations. Therefore, the study of tissue by means of histological and immune stains is the basis of pathology and histology. As the application of Raman spectroscopy to study tissues is straightforward, it was already envisioned in the 70s and the Raman based analysis of tissue complement classical histo-pathology. In comparison to the classical histopathology, Raman spectroscopic tissue diagnostics uses a different type of information to separate morphological structures and alterations. While in the classical histopathology morphological information is employed in Raman spectroscopic based diagnostics, the biochemical composition is utilized to characterize abnormalities and tissue alterations in this case. This is realized with the help of Raman spectral imaging, either by raster scanning the tissue point by point or by line illumination. In both cases, a hyperspectral data cube is constructed. To this data cube the pretreatment and analytical methods described in the section Computational Methods are applied. With the help of these computational methods, maps are created, which allow for chemical imaging. Depending on the algorithm applied, the spatial distribution and arrangement of constituents[117] or groups reflected in the measurement can be visualized. Thanks to the development and optimization of measurement instruments and the increase of computational power, high spatial and temporal resolution can be obtained.[118−123]

Tissue Imaging. Over the past decade, a large number of tissue types have been studied by Raman spectral imaging methods, such as bone, breast, brain, tongue, larynx, spinal, bladder, cervix, and colon tissue. Literately all parts of the body were studied by Raman spectroscopy. Beside the visualization of certain compounds or morphological structures in the tissue, a diagnosis of a disease or diseases is desired. In principle, all diseases and alterations, which lead to a biochemical change, can be detected. Nevertheless, almost all Raman spectroscopic studies involve a cancer detection[96,117,124−140] and only a few studies work on the diagnosis for other types of diseases, like inflammatory bowel diseases (IBDs)[141,142] or bone and skin diseases. By now, various cancers, including brain, breast, lung, etc., and other diseases of bone and skins, can be detected by Raman spectral imaging.[96,117,124−140] Recently, a widefield Raman imaging technique has been reported for detection of Heterotopic Ossification (HO),[136] which means inappropriate bone growth in soft tissue. Moreover, Raman spectral imaging has found its great potential to investigate bone tissues, especially osteoblasts.[134,135] In situ detection of osteoblastic

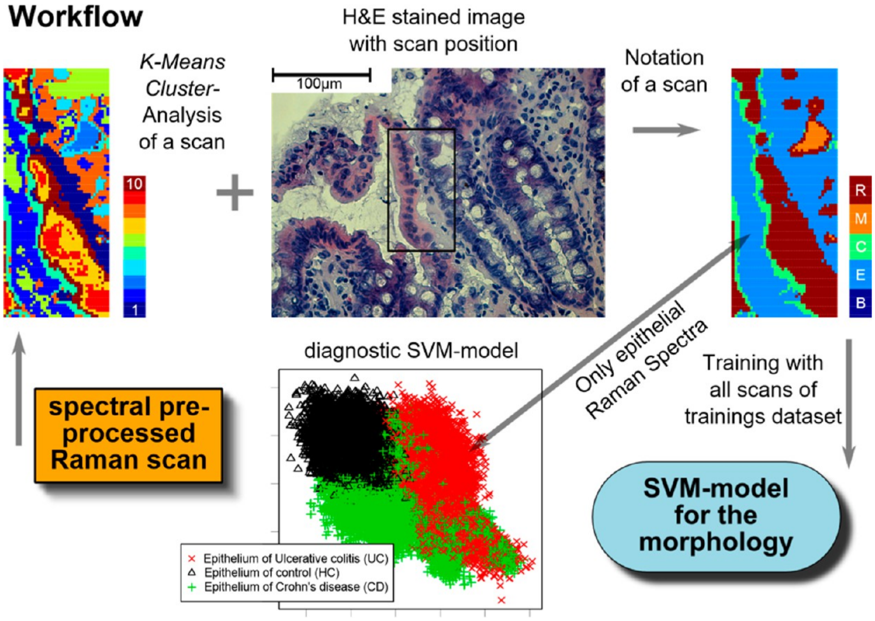

Workflow
K-Means Cluster-Analysis of a scan
10
1
+
H&E stained image with scan position
100µm
Notation of a scan
R
M
C
E
B
spectral pre-processed Raman scan
diagnostic SVM-model
Only epithelial Raman Spectra
Training with all scans of trainings dataset
SVM-model for the morphology
× Epithelium of Ulcerative colitis (UC)
△ Epithelium of control (HC)
+ Epithelium of Crohn's disease (CD)

Figure 7. Spectral histopathology workflow is sketched. The recorded Raman scans are pretreated and a k-means cluster analysis ($k = 10$) is carried out. The corresponding false-color images are subsequently compared with a HE-stained version of the section, on which a pathologist is doing his diagnosis. On the basis of the comparison, the cluster can be attributed to medically relevant groups, which are in that case blood, epithelial tissue, connective tissue, mucus, and a rest group. This training set can be used to construct a classification system for the morphology and a certain group or subset of groups can be utilized to construct a diagnostic classification model. In that case the epithelium of the both diseases Ulcerative colitis and Crohns disease is distinguished from the epithelium of healthy controls. Reproduced with permission from Christiane Bielecki; Thomas W. Bocklitz; Michael Schmitt; Christoph Krafft; Claudio Marquardt; Akram Gharbi; Thomas Knsel; Andreas Stallmach and Jrgen Popp, Classification of inflammatory bowel diseases by means of Raman spectroscopic imaging of epithelium cells, *J. Biomed. Opt. 17*, 076030 (2012) (ref 142). Copyright 2012 Society of Photo Optical Instrumentation Engineers.

mineralization has been achieved by detecting the hydroxyapatite (HA) distribution in bone tissues.[126] Great efforts have been carried out to investigate breast cancer diagnosis. By measuring a Raman spectrum, a real-time biopsy diagnosis can be achieved by Raman spectral imaging.[117,125,137,139] Brain abnormality is another topic of Raman spectral imaging, such as detection of injuries,[140] cancer,[131] and metastases.[130] Furthermore, Raman spectral imaging has been widely employed to diagnose dysplastic tissue or cancer of other tissues, for instance, cervical,[128] lung,[127,129,133] bladder,[138] skin,[132] and liver.[96]

Tissue Diagnostics. In order to compare the Raman based prediction for tissue diagnostics with standard histopathology, a special workflow is needed. This workflow is called spectral histo-pathology (SHP).[143,144] The workflow is sketched in Figure 7 using the study of Bielecki et al.[142] as an example. This workflow starts with the generation of a cluster analysis and the resulting false-color image is compared with a stained version of the measured section. A pathologist is conducting a diagnosis based on the stained image, but without knowing the spectral falsecolor image, otherwise the study would be biased. By doing so, every cluster in the false color image can be attributed to a certain medical related class. In the case of the study presented in reference 142, which is the basis of image Figure 7, these medical related groups were blood, epithelial tissue, connective tissue, mucus, and rest. Either these groups already incorporate a diagnostic group that, for example, a cancerous tissue or a defined tissue structure, is used to construct a diagnosis on a second layer. In the example of reference 142, epithelial tissue is employed to conduct a further classification into two disease types (ulcerative colitis and Crohns disease) and healthy controls. The SHP approach is employed for a large number of cancerous diseases[143,144] and for finding the primary tumor of brain metastases.[145,146] In that case, cancerous tissue on the first classification layer is further distinguished into groups, which reflect the location of the primary tumor. In reference 147, tissue of the oral cavity is investigated but not in an automated manner.

New Trends and Developments. Tissue imaging and diagnosis based on Raman spectroscopy features a unique potential, but the technique is rather slow and the signal strength is limited. The Raman signal is intrinsically low due to the small cross sections of the Raman process. Therefore, the Raman spectrum can be easily masked by fluorescence of the tissue. A possible solution is to apply high intense radiation or long-time integration to improve the signal quality, but the measurement process is enlarged. Additionally, this approach is limited by the maximum permissible exposure (MPE), and long measurements are not desired for studying biological tissues and diagnostics.[136] Therefore, a number of concepts have been introduced, which should speed up the measurement or increase the signal strength. Three of these concepts are surface enhanced Raman scattering (SERS), coherent anti-stokes Raman scattering (CARS), and stimulated Raman scattering (SRS).

Surface enhanced Raman scattering (SERS) uses metal surfaces and by choosing the excitation wavelength wisely together with the substance and manufacturing method of the metal surfaces, a field enhancement around the metal structures can be recognized. This is one enhancement mechanism and

DOI: 10.1021/acs.analchem.5b04665
Anal. Chem. 2016, 88, 133−151

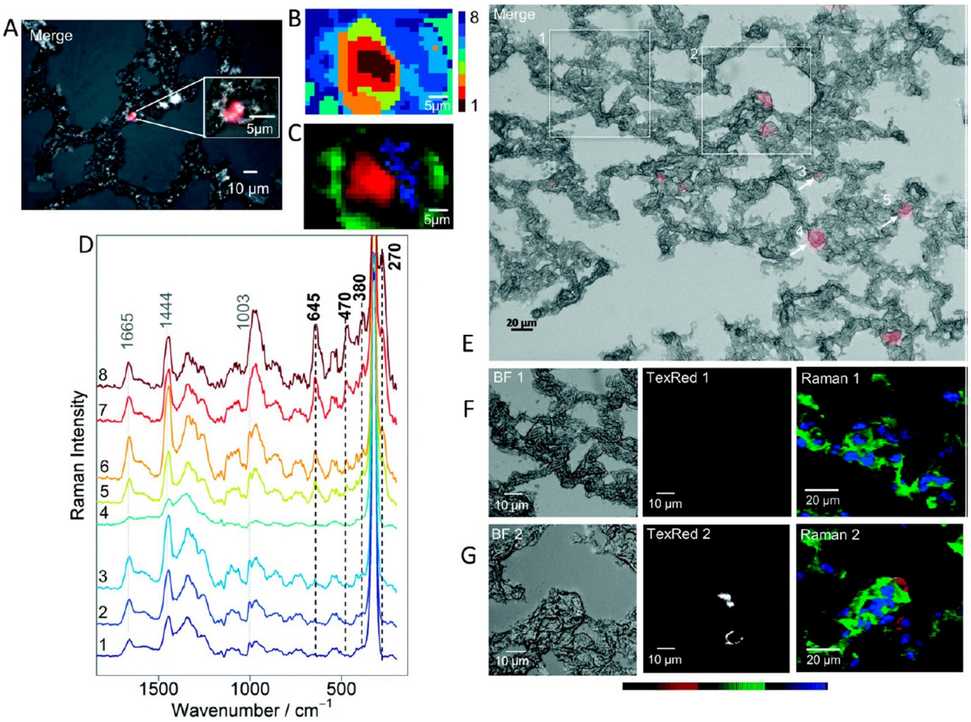


Figure 8. Raman and fluorescence images of rat lung sections incubated with a fluorescence dye (TxR) labeled ZrO_2 nanoparticles (NP) are set together.[175] (A) An overlay of a bright field image and the reddish fluorescence image of TexRed is shown together with a magnified inset. (B) A false-color image generated by a hierarchical cluster analysis of the Raman mapped region (inset region) is displayed. The colors represent the eight predefined clusters. (C) By expert knowledge, a marker band integration is carried out and these maker bands are characteristic bands for tissue, nucleic acids, and ZrO_2 colored in green, blue, and red, respectively. The color scale ranges from maximum and minimum value within the image. (D) The mean Raman spectra of the eight HCA clusters are given and prominent features are marked. Raman maps using calculated sum intensities of selected wavenumber regions are constructed. (E) Again, an overlay of the bright field image and red fluorescence image is shown. The white frames in part E are given in parts F and G in a zoomed version, and these regions were Raman mapped. For the inset, the Raman images show again tissue, nucleic acids, and ZrO_2 colored in green, blue, and red. By comparison of the bright field image, the fluorescence image and the corresponding Raman image are in good agreement and is visible. A. Silge, K. Brutigam, T. Bocklitz, P. Rsch, A. Vennemann, I. Schmitz, J. Popp and M. Wiemann, *Analyst*, **2015**, *140*, 5120 (ref 175). Reproduced by permission of The Royal Society of Chemistry.

the generated evanescent field is a way to obtain stronger Raman scattered light. In the last years, significant progress has been achieved in the application of SERS for tissue imaging. In particular, nanoparticles conjugated with specific antibodies are used as markers for a selectively enhancement of the Raman intensity of the corresponding antibodies.[148,149]

Other approaches for signal enhancement are coherent Raman techniques, like antistokes Raman scattering (CARS) or stimulated Raman scattering (SRS), which can provide orders of magnitude stronger signal, with higher temporal and spatial resolution compared to spontaneous Raman techniques.[150] Because of its high sensitivity in detecting lipids, proteins, and DNA, CARS has become a potential Raman imaging tool for biological samples.[150] For instance, it was employed in ex- and in vivo imaging of mouse ear, where a strong epi-CARS signal was obtained due to the backscattering of the intense forward-propagating CARS radiation in tissue.[151] Axonal myelin was investigated using forward and epi-detected CARS. This study was performed on spinal cord white matter which was isolated from guinea pigs.[152] Moreover, CARS was demonstrated feasible to diagnose lung carcinoma.[153] Most CARS based studies use single band approaches, but recently broad band techniques are applied as well. A high-speed broad band CARS system has been reported which can image the entire biologically relevant Raman window from 500 to 3500 cm^{-1}. With this system, healthy murine liver, pancreas tissue, as well as interfaces between xenograft brain tumors and the surrounding healthy brain matter were imaged.[154] Also other tissue types have been investigated using conventional CARS in combination with further nonlinear imaging modalities,[155,156] and those images can readily be used for diagnostic purposes. Similar to CARS but not disturbed by a nonresonant background is stimulated Raman scattering (SRS), which has also been used for studying biological tissues.[157] For further details of CARS and SRS, both technical and application issues, reviews 150, 158, and 159 are recommended.

Another challenge for Raman spectral imaging of tissues is the maximum penetration depth, which has been restricted to surface or near-surface applications. With spatially offset Raman spectroscopy (SORS), however, the measurement depth in tissues can be increased to 10−20 mm. This strategy has successfully been applied for investigation of bone disease, glucose level,[160] urology,[161] and soft tissues, for example, breast tissue.[162−164] The current research in the SORS field also includes investigations on the appropriate analysis procedures for the resulting SORS data.

Because of the development of fiber-optic Raman probes, a further miniaturization of Raman spectroscopy based instru-

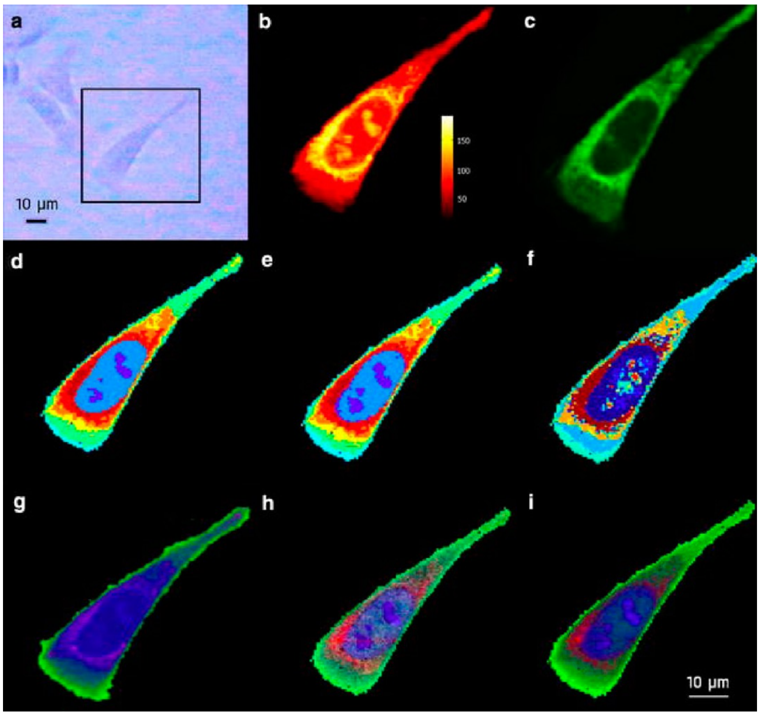


Figure 9. Different computational imaging methods are visualized on an example cell. A white light image (a) of the cell, an univariate intensity plot of the 2935 cm^{-1} Raman band (b), and a fluorescence image tracking mitochondria (c) are given. The images d and e correspond to hard clustering algorithms for image generation. While image d represents a HCA generated cell image, image e is generated by a KM using eight clusters. The image f is generated by a thresholded fuzzy cluster (FC) method with five clusters. The images g−i are constructed using a PCA, a VCA algorithm, and the N-FINDR using three components, which are colored by green, blue, and red. See reference 109 for a biospectroscopic interpretation of the clusters and the images. With kind permission from Springer Science+Business Media: *Theoretical Chemistry Accounts*, Spectral unmixing and clustering algorithms for assessment of single cells by Raman microscopic imaging, *130*, **2011**, 1249, Martin Hedegaard, Christian Matthäus, Soeren Hassing, Christoph Krafft, Max Diem, Jürgen Popp, Figure 1 (ref 109), Copyright Springer-Verlag 2011.

ments is possible.[165,166] Therefore, measurements of solid organs can be achieved, for example, lymph nodes, prostate, and breast.[167,168] As shown in reference 169 with a fiber-optic probe, Raman spectral diagnostics can support decision makers during cancer surgery.[169] Remarkably, several methods have been explored to obtain high-quality Raman spectra with very small or even no Raman background signal.[170,171] The recently developed exciting approach applies Raman spectral imaging for the resection margin localization during cancer surgery.[169] This is usually achieved by selective sampling to dramatically reduce the required acquisition time, where the spatial information on tissues is obtained by another optical technique or real-time Raman spectra.[120] For example, reference 169 shows a hand-held contact fiber optic probe which is developed for differentiation of normal and cancerous brain cells during surgery, which is related to prestudies.[146,172−174]

Recently, the combination of different analysis methods has been focused for Raman spectroscopic application in biological tissues.[176,177] Complementary methods, namely, Raman spectroscopy and matrix-assisted laser desorption/ionization (MALDI) mass spectrometric imaging (MSI), were combined to provide a deeper understanding of biological tissues.[178−181] Diffuse optical tomography methods were combined with Raman spectroscopy to obtain reliable and repeatable localized Raman signals from micro CT-imaged bones in vivo.[118] Another multimodal combination is the combined measurements of optical coherence tomography (OCT) and Raman spectra.[182−184] In this combination OCT acts as overview technique, while the Raman spectra of certain sites are used to predict tissue types and disease states. The so-called surface enhanced spatially offset Raman spectroscopy (SESORS) is able to measure tissues deep into 20−50 mm with up to four labeled nanoparticles,[185] where SORS and SERS are coupled. Another possibility to use two approaches for tissue diagnostics is the application of a dye or labeled structure, like a nanoparticle. In that respect the design of the labeled structure or dye in combination with the Raman excitation wavelength is crucial. In Silge et al.,[175] liver tissue was incubated with labeled nanoparticles and the fluorescence signal and Raman spectra are analyzed together. The image generation based on Raman spectroscopic imaging and fluorescence imaging is visualized in Figure 8. By employing a hierarchical cluster analysis of the Raman scan the nanoparticles could be localized and its surrounding tissue could be studied. A band integration model could be constructed with characteristic bands for tissue, nucleic acids, and ZrO_2, which were colored coded in green, blue, and red, respectively. Other approaches combining fluorescence based techniques and Raman spectroscopy are presented in the articles.[186,187] In reference 186, a combined fiber probe suitable for Raman spectroscopy and fluorescence lifetime imaging is presented. Another interesting combination approach is presented by Kong et al.,[187] where the autofluorescence image is used to select regions where a

Raman spectrum is measured and thus a faster acquisition of large areas can be carried out.

Cell Imaging and Diagnostics. Besides the analysis of tissue, Raman spectroscopy is also employed to study cells, which are not in a network like in tissue. The application of Raman spectroscopy to study single cells, their interaction with the surrounding, and the ongoing cellular processes is an emerging field, and there are a huge number of publications. We try to review these publications in a sorted manner. Therefore, we divided the cell studies in studies on eukaryotic cells, bacteria, and fungi. Bacteria are usually small, so they can be covered within the laser focus. Therefore, the inner structure of bacteria and spores are not easily accessible with Raman spectroscopy. Thus, most of the Raman based studies on bacteria are devoted to either classification or identification of groups of bacteria or to the investigation of the bacterial response on changes in the surrounding of the bacteria, like their growth medium. In the later studies, the bacteria are used as a biosensor. In contrast to that, eukaryotic cells and fungi are larger than the laser focus and thus the inner structure of cells can be investigated. Therefore, a huge application field for Raman spectroscopy is the imaging of cells, their inner structure, and changes introduced by a drug treatment or other influencing factors. In this regard often the response to a cancer drug on cancer cells is studied. Another approach is cyto-pathology based on Raman spectroscopy. Here, Raman spectroscopy is employed to differentiate the biochemical constitution of cancer cells and normal or benign cells. The biochemical composition, measured by a Raman spectrum, is then utilized with a pretrained classification model to conduct a diagnosis of a certain cell. In that regard a single cell based diagnosis would be possible, which would be highly beneficial for fine needle biopsies or brush biopsies. Within this section we will focus on eukaryotic cell studies; for bacterial cell studies, excellent reviews are available.[188,189]

Cell Imaging. First we like to focus on the Raman spectral imaging of eukaryotic cells.[191] If the Raman based imaging is carried out without a label, mathematical methods are used to generate an image. These statistical and mathematical methods are summarized in the Computational Methods section. The methods for image generation of cells typically belong to either hard clustering, soft clustering, or unmixing procedures. In order to get an idea of how these algorithms for Raman spectral imaging of cells work, Figure 9 provides a comparison and overview of the different methods. In Figure 9, a white light image of a cell is given. In comparison, panel b shows the application of a univariate image generation method using the intensity of the 2935 cm^{-1} band and panel c is a fluorescence image tracking mitochondria. The images in panels d and e are the result of hard clustering algorithms for image generation, a HCA and a KM using 8 clusters. The image f is generated utilizing a fuzzy cluster method (FC) with 5 clusters but thresholded before visualization. The images in panel g−i are constructed by multivariate unmixing and factor methods. These methods are a PCA (g), a VCA algorithm (h), and the N-FINDR (i) using three components, which are colored by green, blue, and red. In reference 109, these groups and the mean Raman spectra are interpreted. Further studies not involving a label for visualization purposes are in references 192−194. In reference 192, significant differences in Raman spectra of cytosol and nucleus in different cell-lines were demonstrated. The conclusion is that Raman spectra could quantify the cell state before and after the induction of differentiation in neuroblastoma and adipocytes. The article by Ghita et al.[194] demonstrates that Raman spectroscopy can be used to investigate stem cells. Raman spectroscopy detects the biochemical composition, which reflects the specialized function of differentiated cells. The authors claim that Raman spectroscopy can be utilized for the evaluation of culture conditions during differentiation, cell quality, and phenotype heterogeneity of stem cells. The article, reference 193 on the other hand, warns that a lack of consistency and transparency in the construction of false-color Raman images may lead to wrong interpretation and clues. The false-color images should reflect the biomedical question under investigation.

Beside this direct visualization, a common Raman spectroscopic based practice is to label the structure of interest either with an isotopic compound, like C13 or deuterium, or to use a SERS active nanoparticle. Such a methodology was applied in reference 195, where macrophages were studied. With the help of deuterium labeled fatty acids, the uptake and storage of these exogenously provided fatty acids could be investigated. Li et al.[196] employ gold nanoflowers as a SERS substrate, and they investigate the enhancement of gold nanoflowers and they demonstrate that these SERS substrates result in a strong Raman signal of living cells. Furthermore, they could show that nanoflowers feature excellent targeting properties and a high signal-to-noise ratio for SERS imaging. Panikkanvalappil et al.[197] demonstrate that SERS labeling can be utilized to distinguish cancer cell DNA from healthy cell DNA. The authors claim that the measured SERS spectra were highly reproducible and independent from the human cells and the highly complex composition of the cells. The same group demonstrated[198] that monitoring the mitosis of cells with functionalized gold nanocubes and SERS is possible. They studied the complex biological processes involved in mitosis within populations of healthy and cancerous cells. The authors could interpret their results in a way that a high number of proteins were converted from their helix to a sheet conformation. Another study involving SERS labels is presented by Xiao et al.[199] Within the study of the expression, spatial distribution as well as the endocytosis of EGFR, the Epidermal Growth Factor Receptor, in single breast cancer cells is investigated using SERS. Schie et al.[200] have developed a compound multiphoton microscopy and Raman spectroscopy method for label-free fatty acid chromatography of individual cellular lipid droplets. By comparing the relative amount of palmitic acid and oleic acid determined with Raman spectroscopy in individual lipid droplets with gas chromatography (GC) analysis of several millions of cell, the researchers could show that Raman spectroscopy performed equally to GC.

Classification of Cells. Another type of cell based Raman studies investigates the possibility of classifying cells based on their Raman spectra. The idea behind this approach is that every cell has its unique Raman spectral fingerprint. The application of such a Raman based prediction of cell type would be immense: by a minimal invasive brush biopsy or a fine needle biopsy cells are taken from a patient and by measuring a small number of Raman spectra a prediction of the disease state would be possible. In that way cytopathology might be revolutionized.

In order to test the feasibility of the approach, a number of recent studies are carried out. Pijanka et al.[201] applied Raman spectroscopy to differentiate between lung cancer cells and lung epithelial cells. The authors state that this separation is not sufficient as both cell types also have to be differentiated from

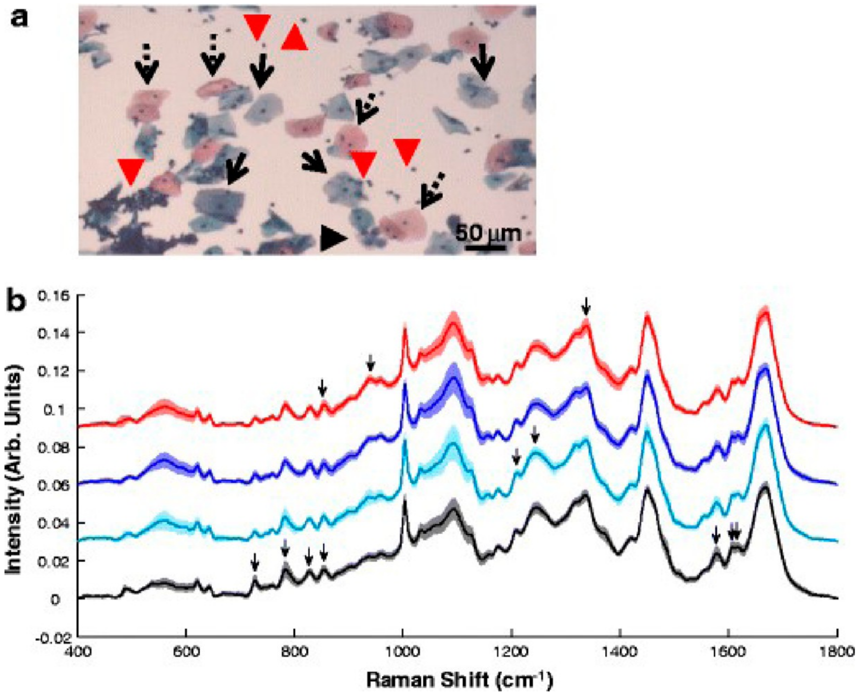


Figure 10. A typical application of Raman spectroscopy for cyto-pathological diagnostics is shown.[190] (a) A Pap-stained negative Thinprep slide shows different cell types or conditions. Here parabasal, intermediate, and superficial cells are marked by a black arrowhead, a solid arrow, and dashed arrows, respectively. White blood cells are indicated with a red arrowhead. Mean Raman spectra of these types and conditions are visualized below with a measure of their uncertainty. The color code is as follows: parabasal (light blue), intermediate (blue), superficial erythrocytes and white blood cells (black). With kind permission from Springer Science+Business Media: *Analytical and Bioanalytical Chemistry*, Raman spectroscopy for screening and diagnosis of cervical cancer, *407*, **2015**, 8279, Fiona M. Lyng, Damien Traynor, Iñes R. M. Ramos, Franck Bonnier, Hugh J. Byrne, Figure 4 (ref 190), Copyright Springer-Verlag Berlin Heidelberg 2015.

lung fibroblasts as well. Another study researches the possibility of a Raman based fine needle aspiration cytology.[202] Within this contribution, various classification models are constructed and evaluated by a cross-validation. One model was based on a SVM and designed to separate six breast cancer cell lines, T47-D, MT-3, MCF-7, JIMT-1, HCC-1143, and BT-20, from each other. The classification accuracy was around 99%. In contrast to the former study, Farhane et al.[203] investigate the influence of the location where spectra are measured on the classification performance. Normal and cancer cells from lung origin, adenocarcinoma cell lines, the Calu-1 cell line, and the BEAS2B normal immortalized bronchial epithelium cell line were studied. Raman spectra of different subcellular compartments (cytoplasm, nucleus, and nucleolus) were acquired. The result was that all regions can be utilized to differentiate normal and cancer cells, but for certain classification tasks only the nucleolar spectral profiles yielded a good classification performance. In the presented case, the two cancer cell lines could only be differentiated by the nucleolar Raman spectra. The publication by Lyng et al.[190] also deals with cell based cancer diagnosis. The authors demonstrate that Raman spectroscopy in combination with multivariate statistical analysis is sensitive to biochemical changes occurring due to cervical cancer. A good validation employing stains is demonstrated, which is related to Figure 10. The authors discussed recent advances and challenges for screening and diagnosis of cervical cancer. The principle, which is illustrated in Figure 10 applies to all Raman based cell studies. A number of cells are measured and compared with the gold standard or a reference method.

In this case a Pap-stained negative Thinprep slide is used to determine the cell types and conditions of the cells. After back tracing of the stained image to the corresponding Raman spectra, the spectra of the groups (parabasal, intermediate and superficial erythrocytes, and white blood cells) can be investigated.

Cellular Drug Response. Another application field for cell based Raman studies is the determination of the cell-drug response on the single cell level and the investigations of carrier systems for intracellular drug delivery.[204] Bi et al.[205] investigate different breast cancer cell lines and their behavior during lapatinib treatment. The authors could characterize the biochemical composition of different cancer cells (BT474, MCF-10A, HER2+ MCF-10A), which are either lapatinib resistant or sensitive. The results indicated a different lipogenesis of resistant cells compared to sensitive cells. In another study,[206] Raman spectroscopy was utilized to determine the spatial distribution of the drug erlotinib within cells. The study provided insights into the drug acting mechanism within the cells in a noninvasive manner. The authors demonstrate that the drug is colocalized with the EGFR protein at the membrane and erlotinib is metabolized within cells to its demethylated derivative. The latter fact was proven by the change of the Raman spectrum of erlotinib measured in cells compared to a reference erlotinib spectrum. A similar *in vitro* study was carried out by Bräutigam et al.[207] They investigated the monitoring of the effectiveness of anticancer drugs in living colon cancer cells (HT-29). By a number of Raman scans, morphological as well as biochemical changes could be observed while the cells were treated with the chemotherapeutic agent docetaxel. A quantification of the response time could be achieved, which may lead to a monitoring tool for the effectiveness of an ongoing chemotherapy. The authors of reference 208 investigate the molecular changes induced by drug treatments on cancer cell nuclei for an improved cancer therapeutic efficiency. In order to understand the mode of action, surface-enhanced Raman scattering (SERS) spectroscopy is applied to study Soma Gastric Cancer (SGC-7901) cells treated with two drugs.

By in situ SERS spectral analysis, the effects of two drugs (Hoechst33342 and doxorubicin) on biomolecules within the cell nuclei could be unraveled. In the study by Farhane et al.[209] the drug doxorubicin was studied within the cells of the lung cancer cell line A549. They employed multivariate statistics in combination with Raman spectroscopic imaging to study the doxorubicin interaction with cancer cells and induced spectral variations. The authors could show that Raman spectroscopy can localize the drug with subcellular resolution and determine the local biomolecular changes induced by doxorubicin. The apoptotic effect in the nuclear regions determined by Raman spectroscopy indicates that the method is capable to monitor the mechanisms of action and response of the cell on a molecular level. Schie et al. used line-scanning Raman microscopy to investigate doxorubicin-induced changes in leukemia T cells for drug exposure time up to 96 h.[210] It was shown that while spectra from individual cell locations were bad predictors of drug-induced changes; Raman spectra representing the total molecular changes in cells were highly reliable. In a follow-up publication, the researchers investigated doxorubicin-induced changes for drug-exposure times between 12 h and 24 h and showed that changes in the total molecular content already occur at 12 h post exposure.[211] Moreover, a reliable mean spectrum, which can be used to describe the total molecular content of a leukemia T cell, can be established based on 30 Raman spectra from randomly chosen cellular locations. Another study on breast cancer cells is carried out by Goel et al.[212] Within the study, the Pentoxifylline treatment induced spectral changes are elucidated. Beside the interpretation of these changes as a linear function of the drug dosage, a classification between a control group and PTX treated group was carried out indicating again the feasibility of Raman spectroscopy as a control tool. In the article, reference 213, the uptake and toxicity of nanoparticles in living cells are researched. Raman spectroscopy allowed for the localization of the nanoparticles inside NIH/3T3 fibroblasts and RAW 264.7 macrophages. From the spatial position of the nanoparticles within the cells and the intracellular concentration with respect to cellular constituents such as proteins and DNA, the uptake could be studied. Like the former also[214] focus on the potential and limitations of Raman spectroscopy to analyze living 3D samples. Raman spectroscopy allows for a labelfree monitoring of metabolic changes with high sensitivity and can be applied also in 3D cell cultures. They point out that application oriented and user-friendly systems are needed to use the unique potential of Raman spectroscopy for monitoring of cell—drug relationships.

Recent Developments for Cell Studies. Recent approaches and developments in the field of Raman based imaging and diagnostics of cells include combination of different methodologies, improvement of existing techniques, and methods for a higher throughput. Higher throughput can be achieved by a parallelization of the measurement or by methods which feature a higher quantum yield. A method which results in a stronger signal is the SERS method but also the mentioned CARS and SRS technique. A number of the SERS studies are already reviewed in the sections above. The stimulated Raman scattering (SRS) is a quite promising technique but is demanding in terms of lab equipment.[159] In the article, reference 215, SRS is used to monitor deuterated choline within living mammalian cell lines. They could show that the subcellular distributions of choline metabolites differed between cancer cells and benign cells. The isotope-based stimulated Raman scattering microscopy for studying choline may also be applied in vivo. As for the conventional Raman spectroscopy also, advanced image analysis and chemometric procedures are applied to analyze the SRS data. This is the subject of ongoing research. For example, in reference 216, a hyperspectral image analysis methodology for SRS data is presented. The authors could show that their spectral phasor analysis allows for a fast and reliable cell segmentation. Other approaches to speed up the measurement are parallelization methods, like line-scan Raman microscopy (LSRM),[217] taking measurements in parallel on a filter system[218] or perform the measurements in a microfluidic lab on a chip device.[219] A lot of ongoing research is devoted to the combination of Raman spectroscopy with other measurement approaches. Two example studies showed the combination of SERS and fluorescence based techniques[208] and Raman spectroscopy with atomic-force microscopy (AFM).[220] These methods try to use the complementary information within each kind of data in order to get a deeper insight into the cell morphology or biochemical changes ongoing via cancer transformation or treatment.

■ SUMMARY AND CONCLUSION

Within this contribution, we reviewed Raman spectroscopy for biomedical applications. We have put together the state of the art and new developments. The review was structured into instrumentation, computational methods, and applications. The time frame covered in these sections differ: In the instrumentation and computational methods sections, also older literature was reviewed, in order to give the reader the possibility to use the section as a tutorial and to see the progress made in the respective field. The application section was divided into cell- and tissue-based studies focusing on eukaryotic cells. The time frame covered within this section is the recent years from 2011, with accentuation on the last 2 years.

In every section an outlook was given indicating new trends and developments. Especially, the application of Raman spectroscopy to real world problems will become challenging and exciting. The development of faster imaging and measuring devices, fiber-based Raman probes, and hand-held Raman devices will get Raman spectroscopy ready for real world applications, for example, in clinics. Therefore, Raman spectroscopy research is moving from research laboratories to real application laboratories in biological, pharmaceutical, or clinical departments. The next step Raman spectroscopy has to take is a good performance in clinical trials to validate the findings and to translate this technique into clinical practice. To do so, a larger data basis and the development of SOPs will be an essential step to move beyond the proof of concept. For biological applications, the combination of Raman with other analytical or imaging techniques will be in the focus of further research. Therefore, Raman spectroscopy in combination with other techniques will answer the next important questions. To sum this contribution up, Raman based molecular imaging and analytics is a magic bullet for biomedical applications.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: Thomas.bocklitz@uni-jena.de.
*E-mail: juergen.popp@uni-jena.de.

Notes

The authors declare no competing financial interest.

Biographies

Thomas W. Bocklitz was born in 1982 and studied physics and mathematics at the Friedrich-Schiller-University. He received his diploma in theoretical physics in 2007 and a Ph.D. in chemometrics in 2011. Dr. Thomas Bocklitz is a junior research group leader for statistical data analysis and image analysis mostly for biophotonic applications. Dr. Bocklitz's research agenda is closely connected with the translation of physical information, measured by AFM, TERS, Raman spectroscopy, CARS, SHG, TPEF, into medical or biological relevant information. This research has led to 30 publications in peer-reviewed journals, one patent, and one book chapter.

Shuxia Guo, born in 1989, finished her bachelor's study in electrical engineering in 2011 at Xiamen University, China. In the same university, she started her research of chemometrics in 2012 and received her Master's degree in 2014. From then on Ms. Guo has been engaged in her Ph.D. research with a topic "chemometrical methods in biological diagnostics based on Raman spectroscopy" in the group of Prof. Jürgen Popp at the Institute for Physical Chemistry in Friedrich-Schiller-Universität Jena. Her main research interest is to establish robust approaches for Raman spectral processing in biological applications.

Oleg Ryabchykov received a bachelor's degree in applied physics at Taurida National V.I. Vernadsky University in 2011 and a master's degree in High Technologies at Taras Shevchenko National University of Kyiv in 2013. Since December 2013, he has been studying data and image analysis as a Ph.D. student at the Institute for Physical Chemistry at the Friedrich Schiller University Jena. His current research includes evaluation of statistical models and data processing for medical diagnosis and characterization by means of Raman spectroscopy and other optical methods.

Nadine Vogler studied bioinformatics at the University in Halle, Germany. After graduating in 2008, she received a scholarship from the German Science Foundation within the Jena School for Microbial Communication to perform her Ph.D. work. Her studies dealt with the application of CARS microscopy in life sciences and medicine. Since gaining her doctoral degree in September 2011, she joined the IPHT as a staff scientist within the research group "Molecular Imaging", where she continues her work on multimodal linear and nonlinear optical microscopy to classify cancer related structural changes in tissue.

Jürgen Popp studied chemistry at the universities of Erlangen and Würzburg. After his Ph.D. in Chemistry he joined Yale University for postdoctoral work. He subsequently returned to Würzburg University where he finished his habilitation in 2002. Since 2002 he holds a chair for Physical Chemistry at the Friedrich-Schiller University Jena. Since 2006 he is also the scientific director of the Leibniz Institute of Photonic Technology, Jena. His research interests are mainly concerned with biophotonics. In particular his expertise in the development and application of innovative Raman techniques for biomedical diagnosis should be emphasized.

■ ACKNOWLEDGMENTS

Financial support of the EU, the "Thüringer Ministerium für Wirtschaft, Wissenschaft und Digitale Gesellschaft", the "Thüringer Aufbaubank", the Federal Ministry of Education and Research, Germany (BMBF), the German Science Foundation, the Fonds der Chemischen Industrie, and the Carl-Zeiss Foundation are greatly acknowledged. The authors thank Sandro Heuke for the generation of image 1.

■ REFERENCES

(1) Smekal, A. *Naturwissenschaften* **1923**, *11*, 873−875.
(2) Raman, C. V. *Nature* **1928**, *121*, 619.
(3) Raman, C. V.; Krishnan, K. S. *Nature* **1928**, *121*, 501−502.
(4) McCreery, R. L. *Raman Spectroscopy for Chemical Analysis*; Chemical Analysis: A Series of Monographs of Analytical Chemistry and Its Applications, Vol. *157*; Wiley-Interscience: New York, 2000.
(5) Lewis, I. R., Edwards, H. G. M., Eds. *Handbook of Raman Spectroscopy: From the Research Laboratory to the Process Line*; Marcel Dekker: New York, 2001.
(6) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, 2008.
(7) Beebe, K.; Pell, R.; Seasholtz, M. *Chemometrics: A Practical Guide*; John Wiley and Sons: New York, 1998.
(8) Gould, R. The LASER, Light Amplification by Stimulated Emission of Radiation. In *The Ann Arbor Conference on Optical Pumping*, University of Michigan, Ann Arbor, MI, June 15−18, 1959.
(9) Boyle, W.; Smith, G. *Bell Syst. Tech. J.* **1970**, *49*, 587−593.
(10) Janesick, J. *Scientific Charge-Coupled Devices*; SPIE Press: Bellingham, WA, 2001.
(11) Lissberger, P. H. *J. Opt. Soc. Am.* **1959**, *49*, 121−122.
(12) Lissberger, P. H.; Wilcock, W. L. *J. Opt. Soc. Am.* **1959**, *49*, 126−128.
(13) Bradley, M. C., Ed. *Chemical and Biochemical Applications of Lasers*; Academic Press: New York, 1977.
(14) Parker, F. *Applications of Infrared, Raman, and Resonance Raman Spectroscopy in Biochemistry*; Plenum Press: New York, 1983.
(15) Mahadevan-Jansen, A.; Richards-Kortum, R. Raman Spectroscopy for Cancer Detection: A Review. Proceedings of the 19th International Conference-IEEE/EMBS, Chicago, IL, October 30− November 2, 1997; pp 2722 − 2728, Vol.6.
(16) Movasaghi, Z.; Rehman, S.; Rehman, I. U. *Appl. Spectrosc. Rev.* **2007**, *42*, 493−541.
(17) Zhu, X.; Xu, T.; Lin, Q.; Duan, Y. *Appl. Spectrosc. Rev.* **2014**, *49*, 64−82.
(18) McCreery, R. L. *Raman Spectroscopy for Chemical Analysis*; John Wiley & Sons, Inc.: New York, 2000; p 452.
(19) Gautam, R.; Samuel, A.; Sil, S.; Chaturvedi, D.; Dutta, A.; Ariese, F.; Umapathy, S. *Curr. Sci.* **2015**, *108*, 341−56.
(20) Turrell, G., Corset, J., Eds. *Raman Microscopy: Developments and Applications*; Elsevier Academic Press: San Diego, CA, 1996.
(21) Wei, D.; Chen, S.; Liu, Q. *Appl. Spectrosc. Rev.* **2015**, *50*, 387−406.
(22) Meng, Z.; Petrov, G. I.; Cheng, S.; Jo, J. A.; Lehmann, K. K.; Yakovlev, V. V.; Scully, M. O. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 12315−12320.
(23) Souza, N.; Zeiger, M.; Presser, V.; Mücklich, F. *RSC Adv.* **2015**, *5*, 62149−62159.
(24) Hecht, E. *Optics*, 4th ed.; Addison-Wesley: Reading, MA, 2002; p 680.
(25) Chase, B. *Anal. Chem.* **1987**, *59*, 881A−890A.
(26) Ellis, G.; Hendra, P. J.; Hodges, C. M.; Jawhari, T.; Jones, C. H.; Le Barazer, P.; Passingham, C.; Royaud, I. A. M.; Snchez-Blzquez, A.; Warnes, G. M. *Analyst* **1989**, *114*, 1061−6.
(27) Cutler, D. J. *Spectrochim. Acta* **1990**, *46*, 131−51.
(28) Zhao, J.; McCreery, R. L. *Appl. Spectrosc.* **1997**, *51*, 1687−97.

(29) Puppels, G. J.; de Mul, F. F. M.; Otto, C.; Greve, J.; Robert-Nicoud, M.; Arndt-Jovin, D. J.; Jovin, T. M. *Nature* **1990**, *347*, 301−303.

(30) Gerbig, Y. B.; Michaels, C. A.; Forster, A. M.; Hettenhouser, J. W.; Byrd, W. E.; Morris, D. J.; Cook, R. F. *Rev. Sci. Instrum.* **2012**, *83*, 125106.

(31) Gerbig, Y. B.; Michaels, C. A.; Cook, R. F. *J. Mater. Res.* **2015**, *30*, 390−406.

(32) Vallance, C.; Brouard, M.; Lauer, A.; Slater, C. S.; Halford, E.; Winter, B.; King, S. J.; Lee, J. W. L.; Pooley, D. E.; Sedgwick, I.; Turchetta, R.; Nomerotski, A.; John, J. J.; Hill, L. *Phys. Chem. Chem. Phys.* **2014**, *16*, 383−395.

(33) Bisson, P. J.; Whitten, J. E. *Rev. Sci. Instrum.* **2015**, *86*, 055107.

(34) Dzsaber, S.; Negyedi, M.; Bernáth, B.; Gyüre, B.; Fehér, T.; Kramberger, C.; Pichler, T.; Simon, F. *J. Raman Spectrosc.* **2015**, *46*, 327−332.

(35) Brückner, M.; Becker, K.; Popp, J.; Frosch, T. *Anal. Chim. Acta* **2015**, *894*, 76−84.

(36) Afseth, N. K.; Segtnan, V. H.; Wold, J. P. *Appl. Spectrosc.* **2006**, *60*, 1358−1367.

(37) Bocklitz, T.; Walter, A.; Hartmann, K.; Rösch, P.; Popp, J. *Anal. Chim. Acta* **2011**, *704*, 47−56.

(38) Gendrin, C.; Roggo, Y.; Collet, C. *J. Pharm. Biomed. Anal.* **2008**, *48*, 533−553.

(39) Vidal, M.; Amigo, J. M. *Chemom. Intell. Lab. Syst.* **2012**, *117*, 138−148.

(40) Mozharov, S.; Nordon, A.; Littlejohn, D.; Marquardt, B. *Appl. Spectrosc.* **2012**, *66*, 1326−1333.

(41) Phillips, G.; Harris, J. M. *Anal. Chem.* **1990**, *62*, 2351−2357.

(42) Cappel, U. B.; Bell, I. M.; Pickard, L. K. *Appl. Spectrosc.* **2010**, *64*, 195−200.

(43) Li, S.; Dai, L. *Appl. Spectrosc.* **2011**, *65*, 1300−1306.

(44) Ehrentreich, F.; Sümmchen, L. *Anal. Chem.* **2001**, *73*, 4364−4373.

(45) Schulze, H. G.; Turner, R. F. *Appl. Spectrosc.* **2014**, *68*, 185−191.

(46) McCreery, R. L. *Raman Spectroscopy for Chemical Analysis*; John Wiley & Sons: New York, 2000; Vol. *157*.

(47) Dörfer, T.; Bocklitz, T.; Tarcea, N.; Schmitt, M.; Popp, J. *Z. Phys. Chem.* **2011**, *225*, 753−764.

(48) Bocklitz, T.; Dörfer, T.; Heinke, R.; Schmitt, M.; Popp, J. *Spectrochim. Acta, Part A* **2015**, *149*, 544−549.

(49) Tseng, C.-H.; Ford, J. F.; Mann, C. K.; Vickers, T. J. *Appl. Spectrosc.* **1993**, *47*, 1808−1813.

(50) Zhang, Z.-M.; Chen, S.; Liang, Y.-Z. *Talanta* **2011**, *83*, 1108−1117.

(51) Fryling, M.; Frank, C. J.; McCreery, R. L. *Appl. Spectrosc.* **1993**, *47*, 1965−1974.

(52) Lin, D. H. M.; Manara, D.; Lindqvist-Reis, P.; Fanghnel, T.; Mayer, K. *Vib. Spectrosc.* **2014**, *73*, 102−110.

(53) Lieber, C. A.; Mahadevan-Jansen, A. *Appl. Spectrosc.* **2003**, *57*, 1363−1367.

(54) Zhang, Z.-M.; Chen, S.; Liang, Y.-Z.; Liu, Z.-X.; Zhang, Q.-M.; Ding, L.-X.; Ye, F.; Zhou, H. *J. Raman Spectrosc.* **2010**, *41*, 659−669.

(55) Ramos, P. M.; Ruisánchez, I. *J. Raman Spectrosc.* **2005**, *36*, 848−856.

(56) Perez-Pueyo, R.; Soneira, M. J.; Ruiz-Moreno, S. *Appl. Spectrosc.* **2010**, *64*, 595−600.

(57) Liu, H.; Zhang, Z.; Liu, S.; Yan, L.; Liu, T.; Zhang, T. *Appl. Spectrosc.* **2015**, *69*, 1013−1022.

(58) Baek, S.-J.; Park, A.; Kim, J.; Shen, A.; Hu, J. *Chemom. Intell. Lab. Syst.* **2009**, *98*, 24−30.

(59) Mosier-Boss, P.; Lieberman, S.; Newbery, R. *Appl. Spectrosc.* **1995**, *49*, 630−638.

(60) Zhao, J.; Lui, H.; McLean, D. I.; Zeng, H. *Appl. Spectrosc.* **2007**, *61*, 1225−1232.

(61) He, S.; Zhang, W.; Liu, L.; Huang, Y.; He, J.; Xie, W.; Wu, P.; Du, C. *Anal. Methods* **2014**, *6*, 4402−4407.

(62) Baek, S.-J.; Park, A.; Ahn, Y.-J.; Choo, J. *Analyst* **2015**, *140*, 250−257.

(63) Emry, J. R.; Olcott Marshall, A.; Marshall, C. P. *Geostand. Geoanal. Res.* **2015**, DOI: 10.1111/j.1751-908X.2015.00354.x.

(64) Bocklitz, T.; Schmitt, M.; Popp, J. In *Ex-Vivo and In-Vivo Optical Molecular Pathology*; Popp, J., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2014; Chapter 7 Image Processing−Chemometric Approaches to Analyze Optical Molecular Images, pp 215−248.

(65) Chen, S.; Lin, X.; Yuen, C.; Padmanabhan, S.; Beuerman, R. W.; Liu, Q. *Opt. Express* **2014**, *22*, 12102−12114.

(66) Beattie, J. R.; Glenn, J. V.; Boulton, M. E.; Stitt, A. W.; McGarvey, J. J. *J. Raman Spectrosc.* **2009**, *40*, 429−435.

(67) Engel, J.; Gerretzen, J.; Szymańska, E.; Jansen, J. J.; Downey, G.; Blanchet, L.; Buydens, L. M. *TrAC, Trends Anal. Chem.* **2013**, *50*, 96−106.

(68) Esquerre, C.; Gowen, A.; Burger, J.; Downey, G.; O'Donnell, C. *Chemom. Intell. Lab. Syst.* **2012**, *117*, 129−137.

(69) Schulze, H.; Turner, R. *Appl. Spectrosc.* **2015**, *69*, 643.

(70) Sacŕe, P.-Y.; De Bleye, C.; Chavez, P.-F.; Netchacovitch, L.; Hubert, P.; Ziemons, E. *J. Pharm. Biomed. Anal.* **2014**, *101*, 123−140 {JPBA} Reviews 2014..

(71) Kanungo, T.; Mount, D. M.; Netanyahu, N. S.; Piatko, C. D.; Silverman, R.; Wu, A. Y. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **2002**, *24*, 881−892.

(72) Likas, A.; Vlassis, N.; Verbeek, J. J. *Pattern Recognition* **2003**, *36*, 451−461.

(73) Ding, C.; He, X. K-means Clustering via Principal Component Analysis. In *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.

(74) Johnson, S. *Psychometrika* **1967**, *32*, 241−254.

(75) Navarro, J. F.; Frenk, C. S.; White, S. D. M. *Astrophys. J.* **1997**, *490*, 493.

(76) Muehlethaler, C.; Massonnet, G.; Esseiva, P. *Forensic Sci. Int.* **2011**, *209*, 173−182.

(77) Cooper, J. B. *Chemom. Intell. Lab. Syst.* **1999**, *46*, 231−247.

(78) Chiang, L. H.; Russell, E. L.; Braatz, R. D. *Chemom. Intell. Lab. Syst.* **2000**, *50*, 243−252.

(79) Guha, S.; Rastogi, R.; Shim, K. *SIGMOD Rec.* **1998**, *27*, 73−84.

(80) Zhang, T.; Ramakrishnan, R.; Livny, M. *SIGMOD Rec.* **1996**, *25*, 103−114.

(81) von Luxburg, U. *Statistics and Computing* **2007**, *17*, 395−416.

(82) Farkas, A.; Vajna, B.; Soti, P. L.; Nagy, Z. K.; Pataki, H.; Van der Gucht, F.; Marosi, G. *J. Raman Spectrosc.* **2015**, *46*, 566−576.

(83) Altman, E. I.; Marco, G.; Varetto, F. *Journal of Banking & Finance* **1994**, *18*, 505−529.

(84) Notingher, I.; Jell, G.; Notingher, P. L.; Bisson, I.; Tsigkou, O.; Polak, J. M.; Stevens, M. M.; Hench, L. L. *J. Mol. Struct.* **2005**, *744747*, 179−185.

(85) Li, M.; Yuan, B. *Pattern Recognition Letters* **2005**, *26*, 527−532.

(86) Howley, T.; Madden, M. *Artificial Intelligence Review* **2005**, *24*, 379−395.

(87) Thissen, U.; Pepers, M.; Ustun, B.; Melssen, W. J.; Buydens, L. M. C. *Chemom. Intell. Lab. Syst.* **2004**, *73*, 169−179.

(88) H., C.-W.; Lin, C.-J. *Neural Networks, IEEE Transactions on* **2002**, *13*, 415−425.

(89) Lee, J.; Chang, K.; Jun, C.-H.; Cho, R.-K.; Chung, H.; Lee, H. *Chemom. Intell. Lab. Syst.* **2015**, *147*, 139−146.

(90) Maji, S.; Berg, A. C.; Malik, J. Classification using intersection kernel support vector machines is efficient. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008, CVPR 2008*, Anchorage, AK, June 24−26, 2008; pp 1−8.

(91) Amari, S.; Wu, S. *Neural Networks* **1999**, *12*, 783−789.

(92) Boucher, T. F.; Ozanne, M. V.; Carmosino, M. L.; Dyar, M. D.; Mahadevan, S.; Breves, E. A.; Lepore, K. H.; Clegg, S. M. *Spectrochim. Acta, Part B* **2015**, *107*, 1−10.

(93) Smola, A.; Schölkopf, B. *Statistics and Computing* **2004**, *14*, 199−222.

(94) Chen, H.-Z.; Tang, G.-Q.; Ai, W.; Xu, L.-L.; Cai, K. *Biotechnol. Prog.* **2015**, DOI: 10.1002/btpr.2161.

(95) Chen, L.; Chu, C.; Huang, T.; Kong, X.; Cai, Y.-D. *Amino Acids* **2015**, *47*, 1485−1493.

(96) Tolstik, T.; Marquardt, C.; Beleites, C.; Matthäus, C.; Bielecki, C.; Bürger, M.; Krafft, C.; Dirsch, O.; Settmacher, U.; Popp, J.; Stallmach, A. *J. Cancer Res. Clin. Oncol.* **2015**, *141*, 407−418.

(97) Svensson, O.; Josefson, M.; Langkilde, F. W. *Chemom. Intell. Lab. Syst.* **1999**, *49*, 49−66.

(98) Ramirez Elias, M.; Guevara, E.; Gonzalez, F. J.; Zamora Pedraza, C.; Aguirre, R.; Juarez, B. Quality control of mezcal combining multivariate analysis techniques and Raman spectroscopy. In *25th International Conference on Electronics, Communications and Computers, CONIELECOMP 2015*, Cholula Puebla, Mexico, February 25−27, 2015; pp 121−123.

(99) Haaland, D. M.; Thomas, E. V. *Anal. Chem.* **1988**, *60*, 1193−1202.

(100) Ryder, A. G.; O'Connor, G. M.; Glynn, T. J. *J. Raman Spectrosc.* **2000**, *31*, 221−227.

(101) Norgaard, L.; Saudland, A.; Wagner, J.; Nielsen, J. P.; Munck, L.; Engelsen, S. B. *Appl. Spectrosc.* **2000**, *54*, 413−419.

(102) Armanfard, N.; Reilly, J.; Komeili, M. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **2015**, 1−1.

(103) Kim, H.; Choi, B. S.; Huh, M. Y. *Knowledge and Data Engineering, IEEE Transactions on* **2015**, *28*, 29−40.

(104) Sharma, M. J.; Yu, S. J. *Applied Mathematics and Computation* **2015**, *253*, 126−134.

(105) Lin, W.-C.; Tsai, C.-F.; Ke, S.-W.; You, M.-L. *Applied Soft Computing* **2015**, *37*, 296−302.

(106) Yuong Wong, S.; Siah Yap, K.; Jen Yap, H. *Neurocomputing* **2016**, *171*, 1431.

(107) Panda, M.; Abraham, A. *Neural Computing and Applications* **2015**, *26*, 507−523.

(108) Bezdek, J. C.; Ehrlich, R.; Full, W. *Comput. Geosci.* **1984**, *10*, 191−203.

(109) Hedegaard, M.; Matthäus, C.; Hassing, S.; Krafft, C.; Diem, M.; Popp, J. *Theor. Chem. Acc.* **2011**, *130*, 1249−1260.

(110) Zortea, M.; Plaza, A. *IEEE Geosci. Remote. S.* **2009**, *6*, 787−791.

(111) Nascimento, J. M. P.; Dias, J. M. B. *IEEE Transaction On Geosience And Remote Sensing* **2005**, *43*, 898−910.

(112) Felten, J.; Hall, H.; Jaumot, J.; Tauler, R.; de Juan, A.; Gorzsas, A. *Nat. Protoc.* **2015**, *10*, 217−240.

(113) Garrido, M.; Rius, F. X.; Larrechi, M. S. *Anal. Bioanal. Chem.* **2008**, *390*, 2059−2066.

(114) Jaumot, J.; de Juan, A.; Tauler, R. *Chemom. Intell. Lab. Syst.* **2015**, *140*, 1−12.

(115) Piqueras, S.; Krafft, C.; Beleites, C.; Egodage, K.; von Eggeling, F.; Guntinas-Lichius, O.; Popp, J.; Tauler, R.; de Juan, A. *Anal. Chim. Acta* **2015**, *881*, 24−36.

(116) Zhang, X.; Tauler, R. *Anal. Chim. Acta* **2013**, *762*, 25−38.

(117) Abramczyk, H.; Brozek-Pluska, B.; Surmacki, J.; Jablonska-Gajewicz, J.; Kordek, R. *Prog. Biophys. Mol. Biol.* **2012**, *108*, 74−81.

(118) Demers, J.-L. H.; Esmonde-White, F. W.; Esmonde-White, K. A.; Morris, M. D.; Pogue, B. W. *Biomed. Opt. Express* **2015**, *6*, 793−806.

(119) Diem, M.; Miljkovíc, M.; Bird, B.; Chernenko, T.; Schubert, J.; Marcsisin, E.; Mazur, A.; Kingston, E.; Zuser, E.; Papamarkakis, K.; Laver, N. *Spectroscopy* **2012**, *27*, 463−496.

(120) Kong, K.; Kendall, C.; Stone, N.; Notingher, I. *Adv. Drug Delivery Rev.* **2015**, *89*, 121.

(121) Morris, M. D.; Mandair, G. S. *Clin. Orthop. Relat. Res.* **2011**, *469*, 2160−2169.

(122) Smith, G. P.; McGoverin, C. M.; Fraser, S. J.; Gordon, K. C. *Adv. Drug Delivery Rev.* **2015**, *89*, 21.

(123) Talari, A. C. S.; Movasaghi, Z.; Rehman, S.; Rehman, I. u. *Appl. Spectrosc. Rev.* **2015**, *50*, 46−111.

(124) Bakker Schut, T.; Witjes, M.; Sterenborg, H.; Speelman, O.; Roodenburg, J.; Marple, E.; Bruining, H.; Puppels, G. *Anal. Chem.* **2000**, *72*, 6010−6018.

(125) Haka, A. S.; Volynskaya, Z.; Gardecki, J. A.; Nazemi, J.; Shenk, R.; Wang, N.; Dasari, R. R.; Fitzmaurice, M.; Feld, M. S. *J. Biomed. Opt.* **2009**, *14*, 054023−054023.

(126) Hashimoto, A.; Chiu, L.-d.; Sawada, K.; Ikeuchi, T.; Fujita, K.; Takedachi, M.; Yamaguchi, Y.; Kawata, S.; Murakami, S.; Tamiya, E. *J. Raman Spectrosc.* **2014**, *45*, 157−161.

(127) Huang, Z.; McWilliams, A.; Lui, H.; McLean, D. I.; Lam, S.; Zeng, H. *Int. J. Cancer* **2003**, *107*, 1047−1052.

(128) Kamemoto, L. E.; Misra, A. K.; Sharma, S. K.; Goodman, M. T.; Luk, H.; Dykes, A. C.; Acosta, T. *Appl. Spectrosc.* **2010**, *64*, 255−261.

(129) Kaminaka, S.; Yamazaki, H.; Ito, T.; Kohda, E.; Hamaguchi, H.-o. *J. Raman Spectrosc.* **2001**, *32*, 139−141.

(130) Kirsch, M.; Schackert, G.; Salzer, R.; Krafft, C. *Anal. Bioanal. Chem.* **2010**, *398*, 1707−1713.

(131) Krafft, C.; Belay, B.; Bergner, N.; Romeike, B. F.; Reichart, R.; Kalff, R.; Popp, J. *Analyst* **2012**, *137*, 5533−5537.

(132) Lui, H.; Zhao, J.; McLean, D.; Zeng, H. *Cancer Res.* **2012**, *72*, 2491−2500.

(133) Magee, N. D.; Beattie, J. R.; Carland, C.; Davis, R.; McManus, K.; Bradbury, I.; Fennell, D. A.; Hamilton, P. W.; Ennis, M.; McGarvey, J. J.; Elborn, J. S. *J. Biomed. Opt.* **2010**, *15*, 026015−026015.

(134) McManus, L. L.; Bonnier, F.; Burke, G. A.; Meenan, B. J.; Boyd, A. R.; Byrne, H. J. *Analyst* **2012**, *137*, 1559−1569.

(135) Nyman, J. S.; Makowski, A. J.; Patil, C. A.; Masui, T. P.; OQuinn, E. C.; Bi, X.; Guelcher, S. A.; Nicollela, D. P.; Mahadevan-Jansen, A. *Calcif. Tissue Int.* **2011**, *89*, 111−122.

(136) Papour, A.; Kwak, J. H.; Taylor, Z.; Wu, B.; Stafsudd, O.; Grundfest, W. *Biomed. Opt. Express* **2015**, *6*, 3892−3897.

(137) Saha, A.; Barman, I.; Dingari, N.; McGee, S.; Volynskaya, Z.; Galindo, L.; Liu, W.; Plecha, D.; Klein, N.; Dasari, R.; Fitzmaurice, M. *Biomed. Opt. Express* **2011**, *2*, 2792−2803.

(138) Shapiro, A.; Gofrit, O. N.; Pizov, G.; Cohen, J. K.; Maier, J. *Eur. Urol.* **2011**, *59*, 106−112.

(139) Surmacki, J.; Musial, J.; Kordek, R.; Abramczyk, H. *Mol. Cancer* **2013**, *12*, 48.

(140) Tay, L.-L.; Tremblay, R. G.; Hulse, J.; Zurakowski, B.; Thompson, M.; Bani-Yaghoub, M. *Analyst* **2011**, *136*, 1620−1626.

(141) Bi, X.; Walsh, A.; Mahadevan-Jansen, A.; Herline, A. *Dis. Colon Rectum* **2011**, *54*, 48−53.

(142) Bielecki, C.; Bocklitz, T.; Schmitt, M.; Krafft, C.; Marquardt, C.; Gharbi, A.; Knösel, T.; Stallmach, A.; Popp, J. *J. Biomed. Opt.* **2012**, *17*, 076030.

(143) Ergin, A.; Großerüschkamp, F.; Theisen, O.; Gerwert, K.; Remiszewski, S.; Thompson, C. M.; Diem, M. *Analyst* **2015**, *140*, 2465.

(144) Diem, M.; Mazur, A.; Lenau, K.; Schubert, J.; Bird, B.; Miljković, M.; Krafft, C.; Popp, J. *J. Biophoton.* **2013**, *6*, 855−886.

(145) Fullwood, L. M.; Clemens, G.; Griffiths, D.; Ashton, K.; Dawson, T. P.; Lea, R. W.; Davis, C.; Bonnier, F.; Byrne, H. J.; Baker, M. J. *Anal. Methods* **2014**, *6*, 3948−3961.

(146) Bergner, N.; Bocklitz, T.; Romeike, B. F.; Reichart, R.; Kalff, R.; Krafft, C.; Popp, J. *Chemom. Intell. Lab. Syst.* **2012**, *117*, 224−232.

(147) Behl, I.; Kukreja, L.; Deshmukh, A.; Singh, S. P.; Mamgain, H.; Hole, A. R.; Krishna, C. M. *J. Biomed. Opt.* **2014**, *19*, 126005.

(148) Schütz, M.; Steinigeweg, D.; Salehi, M.; Kömpe, K.; Schlücker, S. *Chem. Commun.* **2011**, *47*, 4216−4218.

(149) Vendrell, M.; Maiti, K. K.; Dhaliwal, K.; Chang, Y.-T. *Trends Biotechnol.* **2013**, *31*, 249−257.

(150) Cheng, J.-X.; Xie, X. S. *J. Phys. Chem. B* **2004**, *108*, 827−840.

(151) Evans, C. L.; Potma, E. O.; Puoris' haag, M.; Côte, D.; Lin, C. P.; Xie, X. S. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 16807−16812.

(152) Wang, H.; Fu, Y.; Zickmund, P.; Shi, R.; Cheng, J.-X. *Biophys. J.* **2005**, *89*, 581−591.

(153) Gao, L.; Wang, Z.; Li, F.; Hammoudi, A. A.; Thrall, M. J.; Cagle, P. T.; Wong, S. T. C. *Arch. Pathol. Lab. Med.* **2012**, *136*, 1502.

(154) Camp, C. H., Jr.; Lee, Y. J.; Heddleston, J. M.; Hartshorn, C. M.; Walker, A. R. H.; Rich, J. N.; Lathia, J. D.; Cicerone, M. T. *Nat. Photonics* **2014**, *8*, 627−634.

(155) Heuke, S.; Vogler, N.; Meyer, T.; Akimov, D.; Kluschke, F.; Röwert-Huber, H.-J.; Lademann, J.; Dietzek, B.; Popp, J. *Healthcare* **2013**, *1*, 64−83.

(156) Meyer, T.; Guntinas-Lichius, O.; von Eggeling, F.; Ernst, G.; Akimov, D.; Schmitt, M.; Dietzek, B.; Popp, J. *Head & Neckeck* **2013**, *35*, E280−E287.

(157) Saar, B. G.; Freudiger, C. W.; Reichman, J.; Stanley, C. M.; Holtom, G. R.; Xie, X. S. *Science* **2010**, *330*, 1368−1370.

(158) Evans, C. L.; Xie, X. S. *Annu. Rev. Anal. Chem.* **2008**, *1*, 883−909.

(159) Krafft, C.; Schie, I.; Meyer, T.; Schmitt, M.; Popp, J. *Chem. Soc. Rev.* **2016**, DOI: 10.1039/C5CS00564G.

(160) Matousek, P.; Stone, N. *Journal of biophotonics* **2013**, *6*, 7−19.

(161) Matousek, P.; Stone, N. *Analyst* **2009**, *134*, 1058−1066.

(162) Baker, R.; Matousek, P.; Ronayne, K. L.; Parker, A. W.; Rogers, K.; Stone, N. *Analyst* **2007**, *132*, 48−53.

(163) Keller, M. D.; Majumder, S. K.; Mahadevan-Jansen, A. *Opt. Lett.* **2009**, *34*, 926−928.

(164) Stone, N.; Baker, R.; Rogers, K.; Parker, A. W.; Matousek, P. *Analyst* **2007**, *132*, 899−905.

(165) Latka, I.; Dochow, S.; Krafft, C.; Dietzek, B.; Popp, J. *Laser Photonics Rev.* **2013**, *7*, 698.

(166) Krafft, C.; Dochow, S.; Latka, I.; Dietzek, B.; Popp, J. *Biomed. Spectrosc. Imaging* **2012**, *1*, 39−55.

(167) Kendall, C.; Day, J.; Hutchings, J.; Smith, B.; Shepherd, N.; Barr, H.; Stone, N. *Analyst* **2010**, *135*, 3038−3041.

(168) Shim, M. G.; Wilson, B. C.; Marple, E.; Wach, M. *Appl. Spectrosc.* **1999**, *53*, 619−627.

(169) Jermyn, M.; Mok, K.; Mercier, J.; Desroches, J.; Pichette, J.; Saint-Arnaud, K.; Bernstein, L.; Guiot, M.-C.; Petrecca, K.; Leblond, F. *Sci. Transl. Med.* **2015**, *7*, 274ra19−274ra19.

(170) Dochow, S.; Latka, I.; Becker, M.; Spittel, R.; Kobelke, J.; Schuster, K.; Graf, A.; Brückner, S.; Unger, S.; Rothhardt, M.; Dietzek, B.; Krafft, C.; Popp, J. *Opt. Express* **2012**, *20*, 20156−20169.

(171) Santos, L. F.; Wolthuis, R.; Koljenovic, S.; Almeida, R. M.; Puppels, G. J. *Anal. Chem.* **2005**, *77*, 6747−6752.

(172) Haka, A. S.; Volynskaya, Z.; Gardecki, J. A.; Nazemi, J.; Lyons, J.; Hicks, D.; Fitzmaurice, M.; Dasari, R. R.; Crowe, J. P.; Feld, M. S. *Cancer Res.* **2006**, *66*, 3317−3322.

(173) Kircher, M. F.; de la Zerda, A.; Jokerst, J. V.; Zavaleta, C. L.; Kempen, P. J.; Mittra, E.; Pitter, K.; Huang, R.; Campos, C.; Habte, F.; Sinclair, R.; Brennan, C. W.; Mellinghoff, I. K.; Holland, E. C.; Gambhir, S. S. *Nat. Med.* **2012**, *18*, 829−834.

(174) Bergner, N.; Krafft, C.; Geiger, K.; Kirsch, M.; Schackert, G.; Popp, J. *Anal. Bioanal. Chem.* **2012**, *403*, 719−725.

(175) Silge, A.; Brautigam, K.; Bocklitz, T.; Rosch, P.; Vennemann, A.; Schmitz, I.; Popp, J.; Wiemann, M. *Analyst* **2015**, *140*, 5120−8.

(176) Almendro, V.; Marusyk, A.; Polyak, K. *Annu. Rev. Pathol.: Mech. Dis.* **2013**, *8*, 277−302.

(177) Lochhead, P.; Chan, A. T.; Giovannucci, E.; Fuchs, C. S.; Wu, K.; Nishihara, R.; O'Brien, M.; Ogino, S. *Am. J. Gastroenterol.* **2014**, *109*, 1205−1214.

(178) Bocklitz, T.; Crecelius, A.; Matthäus, C.; Tarcea, N.; Von Eggeling, F.; Schmitt, M.; Schubert, U.; Popp, J. *Anal. Chem.* **2013**, *85*, 10829−10834.

(179) Bocklitz, T.; Bräutigam, K.; Urbanek, A.; Hoffmann, F.; von Eggeling, F.; Ernst, G.; Schmitt, M.; Schubert, U.; Guntinas-Lichius, O.; Popp, J. *Anal. Bioanal. Chem.* **2015**, *407*, 7865−7873.

(180) Ahlf, D. R.; Masyuko, R. N.; Hummon, A. B.; Bohn, P. W. *Analyst* **2014**, *139*, 4578−4585.

(181) Masyuko, R.; Lanni, E. J.; Sweedler, J. V.; Bohn, P. W. *Analyst* **2013**, *138*, 1924.

(182) Sudheendran, N.; Qi, J.; Young, E. D.; Lazar, A. J.; Lev, D. C.; Pollock, R. E.; Larin, K. V.; Shih, W.-C. *Laser Phys. Lett.* **2014**, *11*, 105602.

(183) Egodage, K.; Dochow, S.; Bocklitz, T.; Chernavskaia, O.; Matthaeus, C.; Schmitt, M.; Popp, J. *J. Biomed. Photonics Eng.* **2015**, *1*, 169−177.

(184) Ashok, P. C.; Praveen, B. B.; Bellini, N.; Riches, A.; Dholakia, K.; Herrington, C. S. *Biomed. Opt. Express* **2013**, *4*, 2179−2186.

(185) Stone, N.; Faulds, K.; Graham, D.; Matousek, P. *Anal. Chem.* **2010**, *82*, 3969−3973.

(186) Dochow, S.; Ma, D.; Latka, I.; Bocklitz, T.; Hartl, B.; Bec, J.; Fatakdawala, H.; Marple, E.; Urmey, K.; Wachsmann-Hogiu, S.; Schmitt, M.; Marcu, L.; Popp, J. *Anal. Bioanal. Chem.* **2015**, *407*, 8291−8301.

(187) Kong, K.; Rowlands, C. J.; Varma, S.; Perkins, W.; Leach, I. H.; Koloydenko, A. A.; Williams, H. C.; Notingher, I. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 15189−15194.

(188) Lu, X.; Al-Qadiri, H. M.; Lin, M.; Rasco, B. A. *Food Bioprocess Technol.* **2011**, *4*, 919−935.

(189) Pahlow, S.; Meisel, S.; Cialla-May, D.; Weber, K.; Rösch, P.; Popp, J. *Adv. Drug Delivery Rev.* **2015**, *89*, 105.

(190) Lyng, F. M.; Traynor, D.; Ramos, I. R.; Bonnier, F.; Byrne, H. J. *Anal. Bioanal. Chem.* **2015**, *407*, 8279−8289.

(191) Schie, I. W.; Huser, T. *Appl. Spectrosc.* **2013**, *67*, 813−828.

(192) Ichimura, T.; Chiu, L.-d.; Fujita, K.; Kawata, S.; Watanabe, T. M.; Yanagida, T.; Fujita, H. *PLoS One* **2014**, *9*, e84478.

(193) Ashton, L.; Hollywood, K. A.; Goodacre, R. *Analyst* **2015**, *140*, 1852−1858.

(194) Ghita, A.; Pascut, F.; Sottile, V.; Denning, C.; Notingher, I. *EPJ Techn. Instrum.* **2015**, *2*, 6.

(195) Stiebing, C.; Matthäus, C.; Krafft, C.; Keller, A.-A.; Weber, K.; Lorkowski, S.; Popp, J. *Anal. Bioanal. Chem.* **2014**, *406*, 7037−7046.

(196) Li, Q.; Jiang, Y.; Han, R.; Zhong, X.; Liu, S.; Li, Z.-Y.; Sha, Y.; Xu, D. *Small* **2013**, *9*, 927−932.

(197) Panikkanvalappil, S. R.; Mackey, M. A.; El-Sayed, M. A. *J. Am. Chem. Soc.* **2013**, *135*, 4815−4821.

(198) Panikkanvalappil, S. R.; Hira, S. M.; Mahmoud, M. A.; El-Sayed, M. A. *J. Am. Chem. Soc.* **2014**, *136*, 15961−15968.

(199) Xiao, L.; Harihar, S.; Welch, D. R.; Zhou, A. *Anal. Chim. Acta* **2014**, *843*, 73−82.

(200) Schie, I. W.; Nolte, L.; Pedersen, T. L.; Smith, Z.; Wu, J.; Yahiatene, I.; Newman, J. W.; Huser, T. *Analyst* **2013**, *138*, 6662−6670.

(201) Pijanka, J. K.; Stone, N.; Rutter, A. V.; Forsyth, N.; Sockalingum, G. D.; Yang, Y.; Sule-Suso, J. *Analyst* **2013**, *138*, 5052−5058.

(202) Becker-Putsche, M.; Bocklitz, T.; Clement, J. H.; Rösch, P.; Popp, J. *J. Biomed. Opt.* **2013**, *18*, 047001−047001.

(203) Farhane, Z.; Bonnier, F.; Casey, A.; Maguire, A.; O'Neill, L.; Byrne, H. J. *Analyst* **2015**, *140*, 5908−5919.

(204) Kann, B.; Offerhaus, H. L.; Windbergs, M.; Otto, C. *Adv. Drug Delivery Rev.* **2015**, *89*, 71−90.

(205) Bi, X.; Rexer, B.; Arteaga, C. L.; Guo, M.; Mahadevan-Jansen, A. *J. Biomed. Opt.* **2014**, *19*, 025001.

(206) El-Mashtoly, S. F.; Petersen, D.; Yosef, H. K.; Mosig, A.; Reinacher-Schick, A.; Kotting, C.; Gerwert, K. *Analyst* **2014**, *139*, 1155−1161.

(207) Bräutigam, K.; Bocklitz, T.; Schmitt, M.; Rösch, P.; Popp, J. *ChemPhysChem* **2013**, *14*, 550−553.

(208) Liang, L.; Huang, D.; Wang, H.; Li, H.; Xu, S.; Chang, Y.; Li, H.; Yang, Y.-W.; Liang, C.; Xu, W. *Anal. Chem.* **2015**, *87*, 2504−2510.

(209) Farhane, Z.; Bonnier, F.; Casey, A.; Byrne, H. J. *Analyst* **2015**, *140*, 4212−4223.

(210) Schie, I. W.; Alber, L.; Gryshuk, A. L.; Chan, J. W. *Analyst* **2014**, *139*, 2726−2733.

(211) Schie, I. W.; Chan, J. W. *J. Raman Spectrosc.* **2015**, DOI: 10.1002/jrs.4833.

(212) Goel, P. N.; Singh, S. P.; Murali Krishna, C.; Gude, R. P. *J. Innovative Opt. Health Sci.* **2015**, *08*, 1550004.

(213) Bräutigam, K.; Bocklitz, T.; Silge, A.; Dierker, C.; Ossig, R.; Schnekenburger, J.; Cialla, D.; Rösch, P.; Popp, J. *J. Mol. Struct.* **2014**, *1073*, 44−50.

(214) Charwat, V.; Schütze, K.; Holnthoner, W.; Lavrentieva, A.; Gangnus, R.; Hofbauer, P.; Hoffmann, C.; Angres, B.; Kasper, C. *J. Biotechnol.* **2015**, *205*, 70−81.

(215) Hu, F.; Wei, L.; Zheng, C.; Shen, Y.; Min, W. *Analyst* **2014**, *139*, 2312−2317.

(216) Fu, D.; Xie, X. S. *Anal. Chem.* **2014**, *86*, 4115−4119.

(217) Qi, J.; Shih, W.-C. *Appl. Opt.* **2014**, *53*, 2881−2885.

(218) Neugebauer, U.; Kurz, C.; Bocklitz, T.; Berger, T.; Velten, T.; Clement, J. H.; Krafft, C.; Popp, J. *Micromachines* **2014**, *5*, 204−215.

(219) Dochow, S.; Beleites, C.; Henkel, T.; Mayer, G.; Albert, J.; Clement, J.; Krafft, C.; Popp, J. *Anal. Bioanal. Chem.* **2013**, *405*, 2743−2746.

(220) Zhang, D.; Feng, Y.; Zhang, Q.; Su, X.; Lu, X.; Liu, S.; Zhong, L. *Spectrochim. Acta, Part A* **2015**, *141*, 216−222.

[P4] Toward food analytics: fast estimation of lycopene and β-carotene content in tomatoes based on surface enhanced Raman spectroscopy (SERS)

Reproduced from [Radu, A.I., Ryabchykov, O., Bocklitz, T.W., Huebner, U., Weber, K., Cialla-May, D. and Popp, J., 2016. Toward food analytics: fast estimation of lycopene and ß-carotene content in tomatoes based on surface enhanced Raman spectroscopy (SERS). *Analyst*, *141*(14), pp.4447-4455] with permission from the Royal Society of Chemistry.

Erklärungen zu den Eigenanteilen des Promovenden sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an der Publikation.

[1]Radu, A.I., [2]Ryabchykov, O., [3]Bocklitz, T.W., [4]Huebner, U., [5]Weber, K., [6]Cialla-May, D., [7]Popp, J., 2016. Toward food analytics: fast estimation of lycopene and ß-carotene content in tomatoes based on surface enhanced Raman spectroscopy (SERS). *Analyst*, *141*(14), pp.4447-4455							
Beteiligt an (Zutreffendes ankreuzen)							
	1	**2**	3	4	5	6	7
Konzeption des Forschungsansatzes	X	**X**			X	X	X
Planung der Untersuchungen	X		X	X	X	X	X
Datenerhebung	X						
Datenanalyse und Interpretation	X	**X**	X				
Schreiben des Manuskripts	X	**X**	X		X	X	X
Vorschlag Anrechnung Publikationsäquivalent	1.0	**0.5**					

Analyst

PAPER

View Article Online
View Journal | View Issue

Open Access Article. Published on 12 May 2016. Downloaded on 24/01/2018 11:55:43.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Cite this: *Analyst*, 2016, **141**, 4447

Toward food analytics: fast estimation of lycopene and β-carotene content in tomatoes based on surface enhanced Raman spectroscopy (SERS)†

Andreea Ioana Radu,[a,b] Oleg Ryabchykov,[a,b] Thomas Wilhelm Bocklitz,[a,b] Uwe Huebner,[b] Karina Weber,[a,b] Dana Cialla-May*[a,b] and Jürgen Popp[a,b]

Received 16th February 2016,
Accepted 12th May 2016

DOI: 10.1039/c6an00390g

www.rsc.org/analyst

Carotenoids are molecules that play important roles in both plant development and in the well-being of mammalian organisms. Therefore, various studies have been performed to characterize carotenoids' properties, distribution in nature and their health benefits upon ingestion. Nevertheless, there is a gap regarding a fast detection of them at the plant phase. Within this contribution we report the results obtained regarding the application of surface enhanced Raman spectroscopy (SERS) toward the differentiation of two carotenoid molecules (namely, lycopene and β-carotene) in tomato samples. To this end, an e-beam lithography (EBL) SERS-active substrate and a 488 nm excitation source were employed, and a relevant simulated matrix was prepared (by mixing the two carotenoids in defined percentages) and measured. Next, carotenoids were extracted from tomato plants and measured as well. Finally, a combination of principal component analysis and partial least squares regression (PCA-PLSR) was applied to process the data, and the obtained results were compared with HPLC measurements of the same extracts. A good agreement was obtained between the HPLC and the SERS results for most of the tomato samples.

Introduction

For the last several centuries, the scientific focus has been directed toward characterizing the functioning and necessities of the body. Moreover, currently, the role different molecules have, their pathway upon ingestion and their daily intake necessity is being documented. Among others, carotenoids have attracted much attention because of their large bioavailability[1–4] and their important roles in the mammalian organism.[5–10] Nevertheless, it has been found that only 50 out of more than 600 different known carotenoids[1–4] are actually present in the human diet, and only 5–6 of them are detectable in human plasma (α- and β-carotene, β-cryptoxanthin, lycopene, lutein and zeaxanthin).[5,11] The functions each of these carotenoids play in the body range from pro-vitamin A activity and antioxidant activity to radical scavenging. For instance, all-*trans*-β-carotene is the only carotenoid capable of oxidative cleavage into two all-*trans*-retinal molecules, and this process appears to have a feedback regulation property.[2,8,11,12] That is, β-carotene absorption and conversion to retinol partially depends on the individual's vitamin A availability. Moreover, according to different published statistics the necessary intake of vitamin A based on dietary sources of animal origin (*e.g.*, fatty fish, liver and eggs) is often not reached.[9–11,13] Further on, out of the 6 carotenoids detectable in the human plasma, lycopene was found to have the highest efficiency as an antioxidant capable of neutralizing reactive oxygen species (ROS)[5,14,15] and reducing both cell-division at the G0–G1 cell cycle phase and insulin-like growth of mitogens in various cancer cell lines.[5,16,17] There are, however, also negative effects of excessive carotenoid uptake in combination with smoking and alcohol drinking.[9,18] It is, accordingly, important from a health point of view to have a balanced dietary regime. Still, to achieve such a regime, information regarding the quality and composition of the food is needed.

The golden standard in carotenoid analysis is high performance liquid chromatography (HPLC).[19,20] The drawbacks of this method are high costs and limited specificity (because of co-elution).[21] Thus, there is potential for alternative analytical methods. Among others, Raman spectroscopy and surface enhanced Raman spectroscopy (SERS) were tested for analyzing mixtures of carotenoids in various matrices.[22–26] However, previous Raman studies have failed to obtain sufficient differentiation between the two carotenoids at lower concentrations, with HPLC coupled with UV-VIS or MS detection remaining

---

[a] *Friedrich Schiller University Jena, Institute of Physical Chemistry and Abbe Center of Photonics, Helmholtzweg 4, 07745 Jena, Germany*
[b] *Leibniz Institute of Photonic Technology Jena, Albert-Einstein-Str. 9, 07745 Jena, Germany. E-mail: dana.cialla-may@uni-jena.de*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c6an00390g

Open Access Article. Published on 12 May 2016. Downloaded on 24/01/2018 11:55:43.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

the better option for reliable food analytics.[27–29] Here, we present the first results of using SERS to differentiate between lycopene and β-carotene in tomatoes at different ripening stages. However, at this stage of the research, no advantage of the potential the method has towards being non-invasive was applied. Instead, a significant, but small amount of the full-food-batch production sample was taken apart, and an established extraction protocol was used to obtain analyte solutions that were measured by both SERS and HPLC. Regarding the SERS measurements, we designed a step-by-step experimental procedure that reaches towards developing a possible protocol for analyzing carotenoids from tomatoes by employing a relatively simple spectroscopic technique in combination with a statistical analytical tool. For this, the first step consisted of the characterization of the pure carotenoids. Next, different mixtures of β-carotene/lycopene solutions were prepared and measured to create a database and a statistical model that could be used for analysis of food extract samples. The available literature providing information about the amounts of the two carotenoids present in natural products was consulted to decide on the actual mixtures. Finally, a tomato extraction protocol was applied, and the tomato-extracts were measured and analyzed by applying the already existing statistical model. To verify the results of the proposed SERS approach, all tomato samples were also measured by HPLC.

Experimental

Chemicals and reagents

All reagents were of analytical or HPLC reagent grade. Lycopene (≥90% pure), β-carotene (≥95% pure) and 2,6-di-*tert*-butyl-4-methylphenol (BHT, ≥99% pure) were purchased from Sigma Aldrich (Steinheim, Germany). Methanol (≥99.5% pure) was purchased from Carl Roth (Karlsruhe, Germany). Tetrahydrofuran (THF, ≥99.9% pure) was purchased from Merck KGaA (Darmstadt, Germany).

Cherry tomatoes at different ripening stages were provided by local producers from the area of Jena, Germany. First, a series of tomatoes (series A) exhibiting different degrees of ripeness (yellow to red) were taken from the same tomato plant. The tomatoes were immediately frozen and stored at −20 °C until analysis. A second series of four tomatoes (all of them exhibiting the same degree of ripeness – all yellow) were gathered from one plant (series B). One tomato from this batch was frozen immediately. The remaining tomatoes were illuminated with an 11 W lamp for various periods of time leading to increasing degrees of ripeness. After illumination the tomatoes were frozen and stored at −20 °C.

SERS active substrates

For the development of the SERS active substrates e-beam lithography combined with ion-beam etching were used according to the protocol described by Huebner *et al.*[30,31] More exactly, a 4″ fused silica wafer was cleaned using a peroxymonosulfuric acid solution, and then a thin undercoating (hexamethyldisilazane – HMDS) and a 260 nm thick positive tone electron beam resist 'AR6200.09' (ALLRESIST GmbH) were spun on the wafer. Further on, the resist was baked for 3 min at 150 °C on a hotplate, and a 10 nm gold layer was evaporated on top of the resist. The electron beam exposure, which was performed by using the unique character projection-based electron beam technique,[31] of the shaped beam writer SB350OS (from Vistec Electron Beam GmbH) resulted in the formation of 48 chips per wafer (5 × 10 mm^2). Each of the obtained chips contains 4 gratings with a size of 1 × 1 mm^2 for the SERS investigations. The exposure and the removing of the gold layer were followed by the development of the resist in an AR 600-546 developer for 60 s and the IPA rinsing for 30 s. Next, the etching into the fused silica surface was performed with a CHF_3–SF_6–ICP etching process (Inductively Coupled Plasma – ICP) by using an ICP power of 300 W. The etch depth of the 2D gratings with a period of 436 nm is approximately 100 nm. Last, the residual resist was removed using an oxygen plasma, and the wafer was separated into single chips.

Silver films were deposited freshly (at the beginning of every measurement day) by means of thermal evaporation at an oil-free background pressure in the lower 10^{-7} mbar range. For this, the chips were mounted line of sight to the evaporation boat to let the vapor strike the substrate normal to the surface. High-purity 99.999% silver granules were used as raw material. The thickness as well as the deposition rate was controlled *in situ* using a quartz microbalance. The thickness of the silver layer was 40 nm. A scanning electron microscopy (SEM) image of the measuring fields used throughout the experiments is presented in Fig. 1. The image was obtained using a JEOL JSM-6700F system.

Sample preparation

For all of the experiments discussed in this study a mixed solvent of methanol and THF stabilized with 0.1% BHT (1 : 1, v/v) was used.

For the concentration dependent SERS measurements of the two analytes, stock solutions of 106 μM β-carotene and lycopene were prepared by dissolving the appropriate analyte quantities in MeOH/THF. Measuring solutions were prepared by dilution in MeOH/THF immediately prior to use. Final concentrations of 106, 90, 74, 58, 42, 26, 10, 9, 7.4, 5.8, 4.2, 2.6 and 1 μM were obtained for both analytes. All named solutions were used for the SERS measurements shortly after preparation. For each measurement, a new SERS substrate was used.

For the SERS measurements of the β-carotene/lycopene mixtures two stock solutions of 100 μM of each analyte were prepared, and the two analytes were mixed to obtain the analyte percentages shown in Table S1.† Subsequently, the resulting mixtures were measured. However, recording the full data set took a couple of hours, and the mixtures were stored at −20 °C for the needed time. For each mixture a different substrate was used.

The food samples were homogenized to obtain a puree. 2–5 g of each pure sample were mixed with 30 ml methanol/THF of a solution (1 : 1, v/v) and 200 mg magnesium

This journal is © The Royal Society of Chemistry 2016

Open Access Article. Published on 12 May 2016. Downloaded on 24/01/2018 11:55:43.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

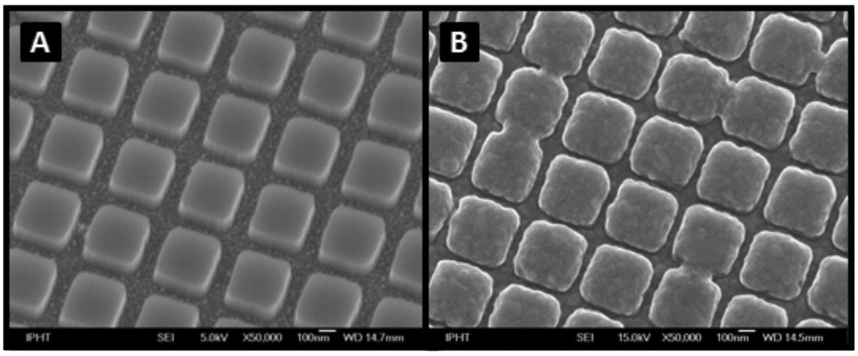

A
B
IPHT SEI 5.0kV X50,000 100nm WD 14.7mm
IPHT SEI 15.0kV X50,000 100nm WD 14.5mm

Fig. 1 SEM image of the measurement field of the SERS active substrate (grating pitch: 436 nm): a quartz-grating as the template without a silver film (A) and the quartz-grating covered with 40 nm silver as a SERS-substrate (B).

bicarbonate. The resulting mixture was stirred using an ultra turrax and filtered with a Buchner funnel. The procedure was repeated two times. The combined filtrates were evaporated to dryness and then dissolved in a defined volume of the extracting agent. The resulting sample was measured by both SERS and HPLC. For the SERS measurements a new substrate was used for each extract.

Spectroscopic measurements

The extinction spectra of the analytes were recorded using a Jasco V650 diode-array spectrophotometer.

For the SERS measurements, the substrates were incubated in the analyte solutions for 30 min and then dried in an Ar stream, which was chosen due to the instability of the carotenoid molecules under normal lab conditions. SERS spectra were recorded using a commercially available WITec confocal Raman system (WITec alpha 300 SR, WITec GmbH, Ulm, Germany) equipped with a 488 nm laser. The light was focused onto the sample *via* a 100× objective (NA 0.9), and the Raman scattered light was collected with the same microscope objective. An optical grating of 1800 g mm^{-1} was used resulting in a spectral resolution of ~2 cm^{-1}. Scans consisting of 100 point measurements were recorded with an integration time of 0.5 s per point. The power at the surface of the sample was adjusted to 20 μW. For each measured analyte (standard solution or sample extract) 13, scans were recorded.

Data analysis

All of the presented spectra were analyzed using R (version 3.0.2)[32] and plotted using Origin 8.5. For data analysis, spectra were first averaged over a 50-point measurement. The resulting spectra were wavenumber calibrated, cut to the relevant spectral range of 500–1700 cm^{-1}, background corrected using the sensitive nonlinear iterative peak (SNIP) algorithm,[33] spike corrected and, for the analysis of the mixtures and food extracts, normalized for the whole spectral range. For the statistical analysis, a principal component analysis (PCA)[34,35] (using a different number of components) was performed and followed by a partial least squares regression (PLSR)[34,35] (using a different number of components) analysis. Two types of cross-validation were performed. First, to build up the training data set, all values representing one concentration were removed, which was repeated for all of the applied concentrations (further referred to as M1). Second, 1% of the total number of measurements was randomly taken out for training (further referred to as M2). The optimal number of principal components for PCA and PLS was chosen, and a model was built using all of the measured data. The obtained model was applied to the test data to predict the food composition. A PCA was applied prior to the PLS regression because performing a regression of high-dimensional data within each repetition of the cross-validation loop would dramatically increase the processing time. On the other side, PCA is an unsupervised method that can be used for the dimensionality reduction of the data outside of the cross-validation loop. Consequently, the PLS regression was performed for low-dimensional data, so the time required for the model construction and evaluation was significantly decreased.

Limit of detection (LOD) values were defined according to the IUPAC norms and are equal to the signal of the blank plus three times the standard deviation of the blank.

HPLC measurements

The HPLC system consisted of a Shimadzu binary gradient system with a DGU-20A3R degassing unit, SIL-20AC autosampler, CTO-20AC column oven and SPD-20A UV/VIS detector. The injection volume was 50 μl, and the separation was performed on a 250 × 4.6 mm S-5 μm YMC 30 HPLC column. The mobile phase consisted of methanol (solvent A) and tertiary butyl methyl ether (tBME, solvent B). The total flow was 1.3 ml min^{-1}, and the column temperature was adjusted to 29 °C. The gradient started with 90% solvent A and 10% solvent B. A linear gradient was applied up to 55% solvent A and 45% solvent B (45 min) followed by another linear gradient up to 45% solvent A and 55% solvent B (5 min). This ratio was held constant for 5 min before returning to the starting conditions (90% solvent A) within 2 min.

Open Access Article. Published on 12 May 2016. Downloaded on 24/01/2018 11:55:43.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

The peaks were evaluated at 450 nm and 470 nm. The quantification was performed by an external calibration with standard solutions taking into account the internal standard. The limit of quantification (LOQ) determined by the signal to noise ratio was 0.03 μg ml^{-1}.

Results and discussion

Plants naturally produce carotenoids to color their flowers and fruits, to attract animals, to gather the light needed for photosynthesis and to protect chlorophyll from photo-damage.[1] According to the available literature[1,2,36,37] lycopene and β-carotene are two carotene molecules (out of 600 known) that can be found in the same plants, in different ratios, depending on the maturity/age of the plant. As depicted in Fig. S1† and largely presented in the literature,[1,2,23,36] in the plant biosynthesis, lycopene is first formed, and upon cyclization, it converts to either β-carotene or α-carotene, which further undergo conversions to other carotenoid molecules. Accordingly, at the different stages of fruit ripening the amount of lycopene and β-carotene also differs. Considering this and keeping in mind the different roles the two carotenoids have in the mammalian organism, it is important to be able to differentiate which plants contain high amounts of each of them. On the other hand, lycopene and β-carotene are very similar from a chemical and spectroscopic point of view (Fig. 2), making the differentiation rather difficult. Analyzing the extinction profile of the analytes (see Fig. 2A), a gain from the resonance contribution by using a 488 nm laser as an excitation source is expected. In a further step, SERRS measurements of the two analyte solutions having a concentration of 106 μM were performed, and the obtained data are shown in Fig. 2B. By analyzing these spectra, a number of differences in the two analytes concerning band intensities and band positions are identified. The ratio of the bands centered at 1526 and 1155 cm^{-1} and assigned to C=C in-phase stretching and C–C stretching vibrations of the polyene chain of the two molecules change when comparing the case of β-carotene with that of lycopene.[38,39] Further on, in the spectral range of 1230–1330 cm^{-1}, two different small bands can be observed. That is, the band centered at 1270 cm^{-1} and assigned to the C–H rocking vibration (also belonging to the polyene chain), and the one at 1287 cm^{-1} assigned to the ring methylene twist.[38] The ratio of these two also changes for the different molecules. Additionally, a 5 cm^{-1} shift of the band centered at approximately 1190 cm^{-1} (and assigned to the C–C stretching vibration) from one molecule to the other can be observed.[38]

As already mentioned, this study is directed towards the detection of β-carotene and lycopene out of a food matrix. To do so, different experimental steps were designed and performed. First, different concentrations of the independent pure analytes were measured before mixing them in different ratios (see Table S1† for the exact percentages). Upon performing these measurements and the analysis, a calibration curve was generated and used for estimating the presence of the two analytes in the studied tomatoes. The results were then compared with the current gold standard, HPLC.

SERRS spectra of different concentrations of the analytes ranging from 106 μM to 1 μM were measured to establish an understanding of the technique's sensitivity. As observed from the plots in Fig. S2,† detection down to a concentration of 10 μM and 26 μM were achieved for lycopene and β-carotene, respectively. Keeping this in mind and considering the already discussed plant-carotenoid transformation path (see Fig. S1†), different lycopene/β-carotene mixtures were prepared. Information regarding the individual percentages of these two analytes in each solution and the different individual concentrations are included in Table S1,† while Fig. 3 depicts the obtained SERRS spectra. As already mentioned, a change in the ratio of the bands centered at 1270 and 1287 cm^{-1} is observed in the case of the SERRS spectra of the pure analytes (Fig. 2B). The same observation is also valid for the spectra in Fig. 3, where a gradual change of the two bands' intensities

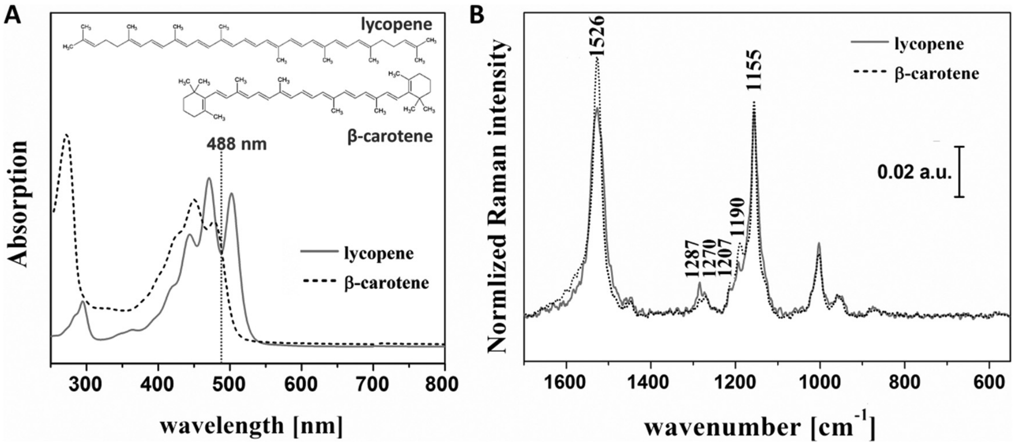


Fig. 2 Chemical structure and extinction spectra of lycopene, 18.7 nM and β-carotene, 6 μM (A) and SERRS spectra of the two analytes, 100 μM (B).

This journal is © The Royal Society of Chemistry 2016

Open Access Article. Published on 12 May 2016. Downloaded on 24/01/2018 11:55:43.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

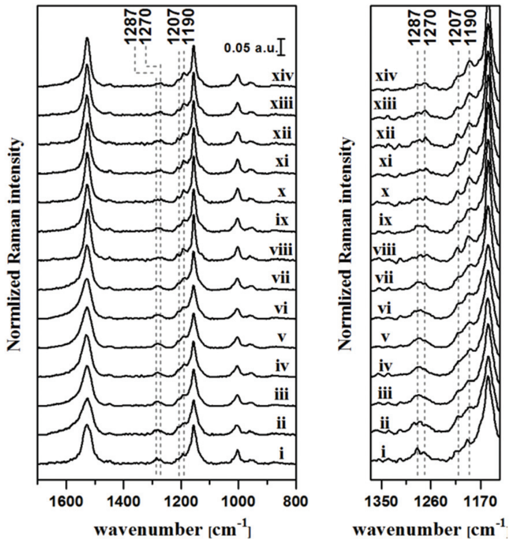


Fig. 3 SERRS spectra of the β-carotene/lycopene mixtures. The spectra are arranged based on the variation of the two analyte percentages: (i) 0% β-carotene and 100% lycopene (0% βc and 100% lyc), (ii) 8% βc and 92% lyc, (iii) 16% βc and 84% lyc, (iv) 24% βc and 76% lyc, (v) 32% βc and 68% lyc, (vi) 40% βc and 60% lyc, (vii) 48% βc and 52% lyc, (viii) 56% βc and 44% lyc, (ix) 64% βc and 36% lyc, (x) 72% βc and 28% lyc, (xi) 80% βc and 20% lyc, (xii) 88% βc and 12% lyc, (xiii) 96% βc and 4% lyc, and (xiv) 100% βc and 0% lyc. The graph on the right side depicts the spectral range between 1330 and 1160 cm^{-1} for better visualization.

occurs with a variation in the amount of the two analytes in the solution. Additionally, a change in the shapes of the bands centered at 1190 and 1207 cm^{-1} as the solution composition changes is observed. That is, the band centered at 1190 cm^{-1} becomes more defined with increasing in β-carotene fractions. To better comprehend and visualize these features and to show the potential of the SERRS technique in food analytics, the SERRS data were further analyzed by applying statistical methods. As mentioned, this consisted of a PCA-PLSR analysis considering the optimal number of components, selected from two different cross validation procedures (see Data analysis section of Experimental for further information). Nevertheless, before applying the PCA-PLSR analysis, each group of 50 spectra were averaged, resulting in a total number of 30 spectra per mixture that were further used for the statistical analysis. This step was performed to compensate for the widely discussed SERRS drawbacks regarding the chemical binding of the analyte to the substrate and the reproducibility of the larger scale SERRS measurements.[40,41] Additionally, when analyzing this result, one should consider that β-carotene percentages lower than 26% are lower than the lowest detectable analyte concentration achieved for measuring the pure analyte. The same is valid for percentages of lycopene lower than 10%. This was expected to negatively influence the root mean square error (RMSE)[34] value of the obtained regression results. The RMSE values obtained for the different considered PCA-PLS component numbers are depicted in Fig. 4. When analyzing this data, one realizes that by using more than 4 PCA components and 2 PLS components a saturation of the RMSE for both cross-validation approaches is achieved. Accordingly, to avoid overfitting, the chosen PCA-PLS combination for further analysis was limited to 4 PCA and 2 PLS components, having a RMSE value of 11.7%. The different cross-validated regression results are depicted in Fig. 5. The expected accuracy of the proposed SERRS method in predicting

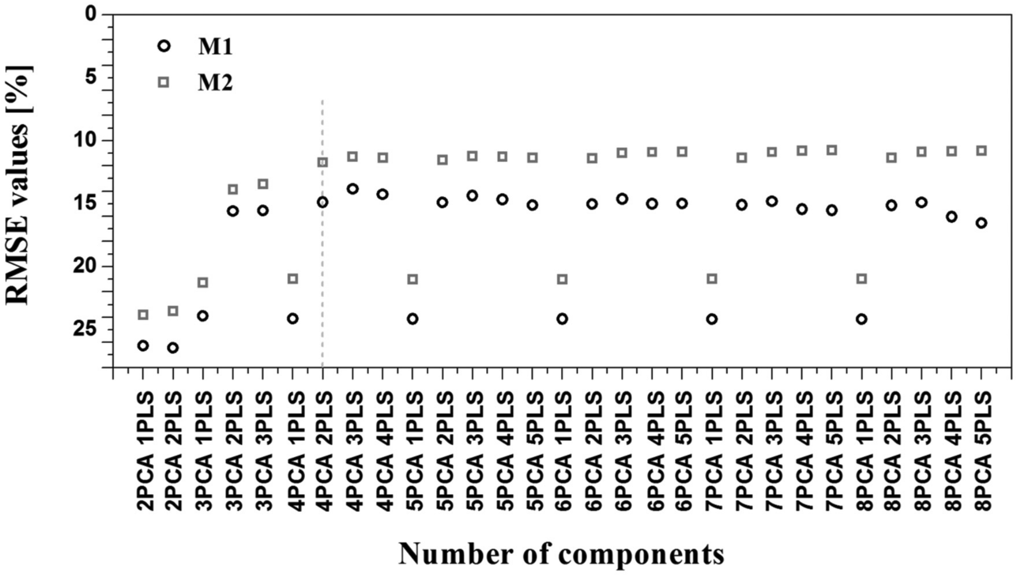


Fig. 4 RMSE values for various numbers of components used for PCA and for PLS.

Open Access Article. Published on 12 May 2016. Downloaded on 24/01/2018 11:55:43.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

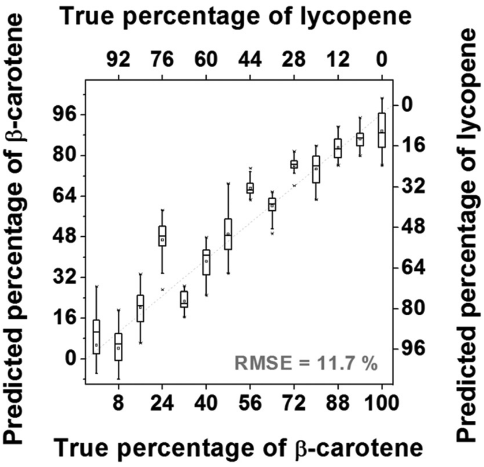


Fig. 5 Cross-validation analysis results obtained for the case of the 4-component PCA and 2-component PLS analysis.

the concentrations of the two investigated carotenoids in food samples is also around the same level. To assess the potential of our SERS approach as well as its limitations by employing real food samples, cherry tomatoes in different ripening stages were investigated.

To do so, the data already presented was used in an attempt to analyze two different series of cherry tomatoes differing in their degree of ripeness. Tomatoes in the A series were picked from the tomato plant upon reaching the different ripening stages, and tomatoes from the B series were picked from the tomato plant at the same early ripening stage (yellow). In the case of the B series, the different investigated ripening stages were achieved in simulated lab-ripening conditions consisting of illuminating the vegetable with an 11 W lamp for the needed period of time. Upon reaching a different ripening stage under the lab illumination conditions, one tomato out of the batch, corresponding to the new ripening stage, was frozen and stored at −20 °C until the extraction and measurement were performed. The analysis protocol depicted in Fig. 6 consists of 4 easy steps. First, an HPLC extraction protocol (described in the Sample preparation section) was performed individually for each of the tomatoes studied (i) to obtain the solutions (ii) used for incubating the SERRS substrates for 30 min (iii). Upon incubation, the substrates were dried using Ar (iv) and measured by applying a 488 nm excitation source. The obtained SERRS spectra are depicted in Fig. 7 together with pictures of the 4 different ripening stages considered throughout this study. As observed, the spectra are similar to the ones already discussed (see Fig. 2B). Nevertheless, there are a few differences in the spectral range of 1100–1300 cm^{-1}. First, a difference in the intensity of the band centered at 1190 cm^{-1} can be observed in both Fig. 7A and B as the color of the tomato changes from yellow (spectra A1 and B1) to red (spectra A4 and B4). The band intensity is higher in the case of the yellow tomato and comparably lower for the red tomatoes. By comparing this alone to the spectra in Fig. 2B one would expect that the red tomato has higher lycopene content than the yellow one. Further on, the shoulder that can be identified at approximately 1207 cm^{-1} in the spectra of the pure analytes (and that is assigned to the C–C stretching vibration)[39] develops into a well-defined band in the case of the tomato samples. This can also be observed in the case of the mixed carotenoid spectra (Fig. 3) and could be a result of the interaction of the different carotenoid molecules *via* the polyene chain. Additionally, the band centered at 1526 cm^{-1} in the case of the pure analytes (Fig. 2B) is blue shifted in the case of the tomato extracts by 6 cm^{-1} for the yellow tomatoes and 9 cm^{-1} for the red tomatoes (Fig. 7A and B). All of the named spectral changes can be caused by the interaction of the carotenoid molecules among themselves and with the SERRS active substrate. However, one should keep in mind that in the case of the tomato extracts the analyzed matrix has a higher degree of complexity than the one used for creating the analytical model. More exactly, other carotenoids, such as phytoene, phytofluene, ζ-carotene, γ-carotene and neurosporene, can also be found in the tomato fruit matrix and could have spectral contributions.[42] Nevertheless, according to the literature, the predominant molecule in the tomato fruit is lycopene followed by β-carotene.[26,42]

The preprocessing of the tomato SERRS spectra was performed by following the same steps performed in the case of the studied pure analyte. The value obtained for the RMSE is approximately 18.9%. The results (in the form of PLS scores obtained by applying the PCA-PLS regression analysis) are presented in Table 1 together with the HPLC measurements'

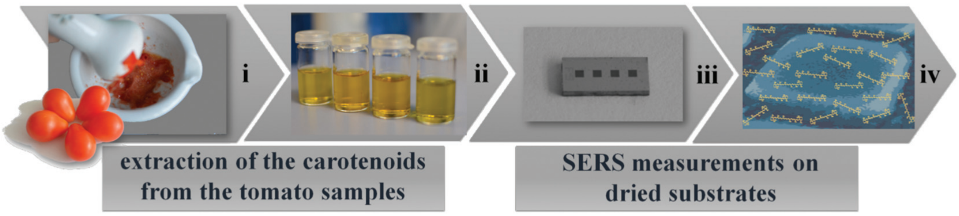

i
ii
iii
iv
extraction of the carotenoids from the tomato samples
SERS measurements on dried substrates

Fig. 6 Schematic representation of the analysis chain. As depicted, the first step consists of the preparation of the analytes to be measured. To this end, the described extraction protocol was applied (i) and the resulting solutions (ii) were used for incubating the SERRS active substrate (iii). Upon incubation the substrates were dyed with N_2 (iv) and measured by means of SERRS.

This journal is © The Royal Society of Chemistry 2016

Open Access Article. Published on 12 May 2016. Downloaded on 24/01/2018 11:55:43.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

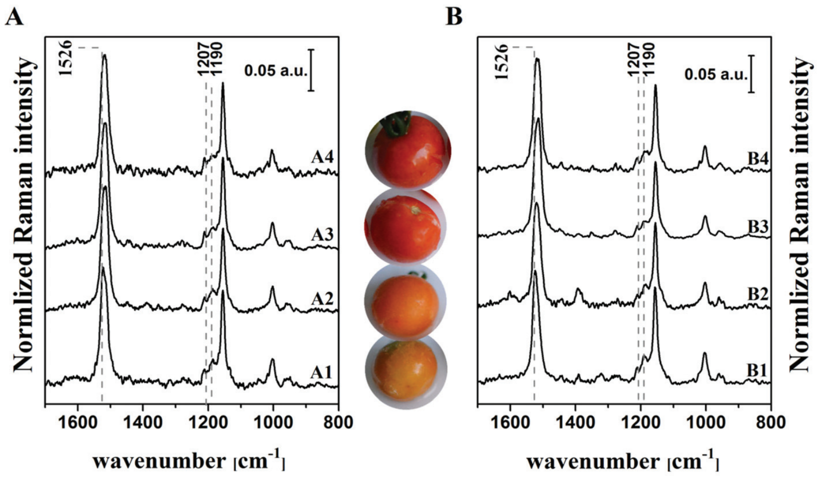


Fig. 7 SERRS spectra of the garden-ripening tomato batch (A) and the lab-ripening tomato batch (B), as well as sample pictures of the colors the tomatoes had when analyzed.

Table 1 Percentage of lycopene and β-carotene estimated to be present in the tomato extracts by means of HPLC and SERRS measurements

		SERRS results[a]		HPLC results			
		lyc	βc	lyc	βc	lyc[b]	βc[b]
		%	%	µM	µM	%	%
Plant-ripening	A1	41.5 ± 12.9	58.5 ± 12.9	10.7 ± 1.1	7.8 ± 0.8	57.9	42.1
	A2	70.1 ± 6.4	29.9 ± 6.4	30.0 ± 3.0	10.6 ± 1.1	74.0	26.0
	A3	67.2 ± 9.4	32.8 ± 9.4	51.5 ± 5.1	13.4 ± 1.3	79.3	20.7
	A4	59.6 ± 15.6	40.4 ± 15.6	20.8 ± 2.1	12.0 ± 1.2	63.4	36.6
Lab-ripening	B1	11.3 ± 10.4	88.7 ± 10.4	13.1 ± 1.3	10.7 ± 1.1	55.1	44.9
	B2	48.1 ± 9.0	51.9 ± 9.0	27.9 ± 2.8	13.2 ± 1.3	67.9	32.1
	B3	86.6 ± 5.7	13.4 ± 5.7	99.7 ± 9.9	14.6 ± 1.5	87.2	12.8
	B4	63.9 ± 4.8	36.1 ± 4.8	54.8 ± 5.5	18.0 ± 1.8	75.3	24.7

lyc – lycopene. βc – β-carotene. [a] PLS score value. [b] The % calculation was performed by considering that lycopene and β-carotene are the only two carotenoids present in the extract.

results. For the latter, the exact same extract was measured as in the case of the SERRS experiments. As observed from the table, a quite good agreement was obtained for the two analytical methods. In the case of the samples B1 and B2, however, the lycopene and β-carotene content predicted by SERRS and measured by HPLC presented different levels of the two carotenoids. This might be due to miss-assignments of other carotenoids present in the extract (*i.e.*, phytoene, phytofluene, ζ-carotene, γ-carotene and neurosporene) by SERRS, as at this ripening stage their presence in the plant is expected. An improved prediction of the tomato composition is expected when employing all of the mentioned carotenoids for building the model. This prediction is, however, beyond the aim of this study. A further observation that can be made by analyzing Table 1 is related to the variation in the lycopene composition in an adult tomato fruit. Upon reaching the red ripening stage, further storage of the tomato before consumption leads to a decrease of the lycopene content in favor of other carotenoids. This is important when deciding on a dietary regime building towards a health-improving result. A last, interesting observation, first noted during experimentation, confirms that a tomato's lycopene content still increases when the fruit ripening is achieved in lab/shop conditions. This is important regarding the transportation time needed from the actual plantations to the commercializing facilities.

Conclusions

The current paper presents the work performed toward analyzing cherry tomato fruits by means of SERRS. To this end, a rather simple but relevant simulated matrix was prepared. This matrix consisted of different mixtures of the two most prevalent carotenoids found in tomatoes, namely, β-carotene and lycopene. The percentages of the two carotenoids were varied to simulate possible compositions in vegetables, such as tomatoes. Upon statistical analysis, a regression curve was obtained and used to analyze the tomato samples. Further on,

Open Access Article. Published on 12 May 2016. Downloaded on 24/01/2018 11:55:43.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

two tomato series representative of garden-ripening and laboratory-ripening conditions were considered to test the developed analytical method and to comparatively access the lycopene/β-carotene abundance of market-available tomatoes. Accordingly, upon acquiring the needed tomatoes for the designed experiments, they were subjected to the same carotenoid-extraction protocol and measured by both SERRS and HPLC. The SERRS measurements were performed under the same conditions as the ones employed for the lycopene/β-carotene mixtures. Upon analyzing the data, we were able to estimate the abundance of the two carotenoids investigated in the tomato samples. Moreover, a good agreement was obtained between the HPLC and the SERRS results for most of the tomato samples. Additionally, both measurement methods registered a gradual increase of the lycopene content independent of the tomato ripening conditions investigated.

Acknowledgements

The funding of research projects “QuantiSERS” (03IPT513A) and “Jenaer Biochip Initiative 2.0” (03IPT513Y) within the framework of “InnoProfile Transfer – Unternehmen Region” of the Federal Ministry of Education and Research, Germany (BMBF) is gratefully acknowledged. Tobias Voigt, Dr Bianca Braha and Dr Bernd Giese from Food GmbH (Jena, Germany) are gratefully acknowledged for their expertise concerning carotenoid extraction from food and for providing the HPLC measurements. Sabine Schmidt and Torsten May are gratefully acknowledged for providing the different cherry tomato patches used throughout the experiments.

References

1 J. Hirschberg, *Curr. Opin. Plant Biol.*, 2001, **4**, 210–218.

2 E. H. Harrison, C. dela Sena, A. Eroglu and M. K. Fleshman, *Am. J. Clin. Nutr.*, 2012, **96**, 1189S–1192S.

3 R.-X. Yu, W. Köcher, M. E. Darvin, M. Büttner, S. Jung, B. N. Lee, C. Klotter, K. Hurrelmann, M. C. Meinke and J. Lademann, *J. Biophotonics*, 2014, **7**, 926–937.

4 M. E. Darvin, H. Richter, S. Ahlberg, S. F. Haag, M. C. Meinke, D. Le Quintrec, O. Doucet and J. Lademann, *J. Biophotonics*, 2014, **7**, 735–743.

5 A. Gajowik and M. M. Dobrzyńska, *Rocz. Panstw. Zakl. Hig.*, 2014, **65**, 263–271.

6 M. Dizdaroglu, *Free Radicals Biol. Med.*, 1991, **10**, 225–242.

7 J. E. Klaunig and L. M. Kamendulis, *Annu. Rev. Pharmacol. Toxicol.*, 2004, **44**, 239–267.

8 J. D. Ribaya-Mercado, F. S. Solon, M. A. Solon, M. A. Cabal-Barza, C. S. Perfecto, G. Tang, J. A. A. Solon, C. R. Fjeld and R. M. Russell, *Am. J. Clin. Nutr.*, 2000, **72**, 455–465.

9 M. Jenab, S. Salvini, C. H. van Gils, M. Brustad, S. Shakya-Shrestha, B. Buijsse, H. Verhagen, M. Touvier, C. Biessy, P. Wallstrom, K. Bouckaert, E. Lund, M. Waaseth, N. Roswall, A. M. Joensen, J. Linseisen, H. Boeing, E. Vasilopoulou, V. Dilis, S. Sieri, C. Sacerdote, P. Ferrari, J. Manjer, S. Nilsson, A. A. Welch, R. Travis, M. C. Boutron-Ruault, M. Niravong, H. B. Bueno-de-Mesquita, Y. T. van der Schouw, M. J. Tormo, A. Barricarte, E. Riboli, S. Bingham and N. Slimani, *Eur. J. Clin. Nutr.*, 0000, **63**, S150–S178.

10 A. Agudo, N. Slimani, M. C. Ocké, A. Naska, A. B. Miller, A. Kroke, C. Bamia, D. Karalis, P. Vineis, D. Palli, H. B. Bueno-de-Mesquita, P. H. M. Peeters, D. Engeset, A. Hjartåker, C. Navarro, C. M. Garcia, P. Wallström, J. X. Zhang, A. A. Welch, E. Spencer, C. Stripp, K. Overvad, F. Clavel-Chapelon, C. Casagrande and E. Riboli, *Public Health Nutr.*, 2002, **5**, 1179–1196.

11 T. Grune, G. Lietz, A. Palou, A. C. Ross, W. Stahl, G. Tang, D. Thurnham, S.-A. Yin and H. K. Biesalski, *J. Nutr.*, 2010, **140**, 2268S–2285S.

12 H. Bachmann, A. Desbarats, P. Pattison, M. Sedgewick, G. Riss, A. Wyss, N. Cardinault, C. Duszka, R. Goralczyk and P. Grolier, *J. Nutr.*, 2002, **132**, 3616–3622.

13 D. R. Tennant, J. Davidson and A. J. Day, *Br. J. Nutr.*, 2014, **112**, 1214–1225.

14 P. Di Mascio, S. Kaiser and H. Sies, *Arch. Biochem. Biophys.*, 1989, **274**, 532–538.

15 L. G. Rao, E. Guns and A. V. Rao, *Agric. Food Industry Hi-Tech*, 2003, **14**, 25–30.

16 J. Levy, E. Bosin, B. Feldman, Y. Giat, A. Miinster, M. Danilenko and Y. Sharoni, *Nutr. Cancer*, 1995, **24**, 257–266.

17 R. Matsushima-Nishiwaki, Y. Shidoji, S. Nishiwaki, T. Yamada, H. Moriwaki and Y. Muto, *Lipids*, 1995, **30**, 1029–1034.

18 T. Tanaka, M. Shnimizu and H. Moriwaki, *Molecules*, 2012, **17**, 3202.

19 S. F. P. R. H. C. P. P. S. D. K. Lloyd, in *Liquid Chromatography Applications*, Elsevier, Amsterdam, 2013, p. 667, DOI: 10.1016/B978-0-12-415806-1.01001-9.

20 Trends Food Sci. Technol., 2003, **14**(10), 438.

21 M. W. Dong, *LCGC North Am.*, 2013, **31**(6), 472–479.

22 A. M. Nikbakht, T. T. Hashjin, R. Malekfar and B. Gobadian, *J. Agric. Sci. Technol.*, 2011, **13**, 517–526.

23 R. Baranski, M. Baranska and H. Schulz, *Planta*, 2005, **222**, 448–457.

24 M. Køcks, S. O. Banke, B. Madsen, T. Vaz, M. Carvalheira, N. Pandega, I. Sousa and S. D. Nygaard, *Appl. Spectrosc.*, 2013, **67**, 681–687.

25 K. Hesterberg, S. Schanzer, A. Patzelt, W. Sterry, J. W. Fluhr, M. C. Meinke, J. Lademann and M. E. Darvin, *J. Biophotonics*, 2012, **5**, 33–39.

26 J. Qin, K. Chao and M. S. Kim, *Postharvest Biol. Technol.*, 2012, **71**, 21–31.

27 J. Trebolazabala, M. Maguregui, H. Morillas, A. de Diego and J. M. Madariaga, *Spectrochim. Acta, Part A*, 2013, **105**, 391–399.

28 M. Baranska, W. Schütze and H. Schulz, *Anal. Chem.*, 2006, **78**, 8456–8461.

This journal is © The Royal Society of Chemistry 2016

Open Access Article. Published on 12 May 2016. Downloaded on 24/01/2018 11:55:43.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

29 W. Liu, Z. Wang, Z. Zheng, L. Jiang, Y. Yang, L. Zhao and W. Su, *Chin. J. Chem.*, 2012, **30**, 2573–2580.

30 U. Huebner, K. Weber, D. Cialla, H. Schneidewind, M. Zeisberger, H. G. Meyer and J. Popp, *Microelectron. Eng.*, 2011, **88**, 1761–1763.

31 U. Huebner, M. Falkner, U. D. Zeitner, M. Banasch, K. Dietrich and E.-B. Kley, 30th European Mask and Lithography Conference, *Proc. SPIE*, 2014, **9231**, 92310E.

32 R. C. Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014.

33 M. Omer, H. Negm, R. Kinjo, Y.-W. Choi, K. Yoshida, T. Konstantin, M. Shibata, K. Shimahashi, H. Imon, H. Zen, T. Hori, T. Kii, K. Masuda and H. Ohgaki, in *Zero-Carbon Energy Kyoto 2012*, ed. T. Yao, Springer, Japan, 2013, ch. 27, pp. 245–252. DOI: 10.1007/978-4-431-54264-3_27.

34 R. Wehrens, *Chemometrics with R. Multivariate data analysis in the natural sciences and life Sciences*, 2011, DOI: 10.1007/978-3-642-17841-2.

35 T. Bocklitz, A. Walter, K. Hartmann, P. Rösch and J. Popp, *Anal. Chim. Acta*, 2011, **704**, 47–56.

36 L. Arab, S. Steck-Scott and P. Bowen, *Participation of lycopene and beta-carotene in carcinogenesis: defenders, aggressors, or passive bystanders?*, 2001.

37 S. Schlücker, A. Szeghalmi, M. Schmitt, J. Popp and W. Kiefer, *J. Raman Spectrosc.*, 2003, **34**, 413–419.

38 N. Tschirner, M. Schenderlein, K. Brose, E. Schlodder, M. A. Mroginski, C. Thomsen and P. Hildebrandt, *Phys. Chem. Chem. Phys.*, 2009, **11**, 11471–11478.

39 M. R. López-Ramírez, S. Sanchez-Cortes, M. Pérez-Méndez and G. Blanch, *J. Raman Spectrosc.*, 2010, **41**, 1170–1177.

40 C. Wang, C. J. Berg, C.-C. Hsu, B. A. Merrill and M. J. Tauber, *J. Phys. Chem. B*, 2012, **116**, 10617–10630.

41 V. R. Salares, N. M. Young, P. R. Carey and H. J. Bernstein, *J. Raman Spectrosc.*, 1977, **6**, 282–288.

42 F. Khachik, L. Carvalho, P. S. Bernstein, G. J. Muir, D.-Y. Zhao and N. B. Katz, *Exp. Biol. Med.*, 2002, **227**, 845–851.

[P5] Raman spectroscopic investigation of the human liver stem cell line HepaRG

Copyright 2018 Wiley. Reprinted with kind permission from [Ryabchykov, O., Bräutigam, K., Galler, K., Neugebauer, U., Mosig, A., Bocklitz, T. and Popp, J., 2018. Raman spectroscopic investigation of the human liver stem cell line HepaRG. *Journal of Raman Spectroscopy*, *49*(6), pp.935-942]

Erklärungen zu den Eigenanteilen des Promovenden sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an der Publikation.

[1]Ryabchykov, O., [2]Bräutigam, K., [3]Galler, K., [4]Neugebauer, U., [5]Mosig, A., [6]Bocklitz, T., [7]Popp, J., 2018. Raman spectroscopic investigation of the human liver stem cell line HepaRG. *Journal of Raman Spectroscopy*, *49*(6), pp.935-942							
Beteiligt an (Zutreffendes ankreuzen)							
	1	2	3	4	5	6	7
Konzeption des Forschungsansatzes	**X**	X					X
Planung der Untersuchungen		X		X	X	X	X
Datenerhebung		X	X				
Datenanalyse und Interpretation	**X**					X	
Schreiben des Manuskripts	**X**	X	X	X		X	X
Vorschlag Anrechnung Publikationsäquivalent	**1.0**	1.0					

Received: 1 November 2017 | Revised: 23 March 2018 | Accepted: 3 April 2018
DOI: 10.1002/jrs.5392

SPECIAL ISSUE - RESEARCH ARTICLE

WILEY Journal of RAMAN SPECTROSCOPY

Raman spectroscopic investigation of the human liver stem cell line HepaRG

Oleg Ryabchykov[1,2,*] | Katharina Bräutigam[1,2,*] | Kerstin Galler[2,3] |
Ute Neugebauer[1,2,3] | Alexander Mosig[3] | Thomas Bocklitz[1,2] | Jürgen Popp[1,2]

[1]Institute of Physical Chemistry and Abbe Center of Photonics, University of Jena, Helmholtzweg 4, Jena 07743, Germany

[2]Leibniz Institute of Photonic Technology, Albert-Einstein-Straße 9, Jena 07745, Germany

[3]Center for Sepsis Control and Care (CSCC), Jena University Hospital, Jena, Germany

Correspondence
Thomas Bocklitz, Institute of Physical Chemistry and Abbe Center of Photonics, University of Jena, Helmholtzweg 4, Jena 07743, Germany; or Leibniz Institute of Photonic Technology, Albert-Einstein-Straße 9, Jena 07745, Germany.
Email: thomas.bocklitz@uni-jena.de

Funding information
BMBF; Center for Sepsis Control and Care, Grant/Award Number: (FKZ 01EO1502); PhotoSkin, Grant/Award Number: (FKZ 13N13243); EU; HemoSpec, Grant/Award Number: (FP 7, CN 611682); PhotoSkin, Grant/Award Number: (FKZ 13N13243); Free State of Thuringia, Grant/Award Number: (FKZ 2015 FGI 0011 and 2016 FGI 0010); European Regional Development Fund (EFRE)

Abstract

In this work, Raman spectroscopic cell imaging approaches and a discrimination between HepG2, nondifferentiated hepatic stem cell line HepaRG, and differentiated hepatocyte-like HepaRG cells are presented. Raman spectroscopic imaging was used to visualize the cell nuclei by means of false color imaging using a marker band, and a cell segmentation was performed by means of clustering. Furthermore, a 3-class-classification model based on the mean Raman spectra of individual cells was established for a classification between different cell types. A high average sensitivity of 96% was achieved by the applied classification model. Based on the results of clustering and classification, the main spectral contributions to different cell types and cell segments were analyzed in detail. Thereby, HepG2, nondifferentiated hepatic stem cell line HepaRG, and differentiated hepatocyte-like HepaRG cells were Raman spectroscopically characterized and proven to be significantly different.

KEYWORDS
HepaRG, HepG2, Raman spectroscopic imaging

1 | INTRODUCTION

Understanding the interactions between eukaryotic cells and drugs is of utmost importance to (a) minimize the toxic impact of pharmaceuticals on the human body in an early stage of the drug design, (b) to personalize medication with a drug dosage tailored for an individual patient based on its own capability to metabolize pharmaceuticals, and (c) to react adequately to the course of a disease. Therefore, cell–drug interaction assays are used to study the interactions of drugs and cells. To implement such an assay towards an understanding of cell–drug interaction on a cellular and subcellular level, a model system is needed.

The most important organ of drug metabolism is the liver including their numerous enzyme systems, for example, the cytochrome P450 system, which is involved in a number of metabolic pathways. Due to the

---

[*] Oleg Ryabchykov and Katharina Bräutigam contributed equally to the presented work.

Copyright © 2018 John Wiley & Sons, Ltd.

hepatotoxicity, several drugs were withdrawn from the market during the last decades.[1] One reason for this situation is the failure of animal-based studies as a result of the weak correlation between the xenobiotic metabolism and hepatic toxicity between humans and animals.[2] Furthermore, primary human hepatocytes, which are known as the gold standard model for investigations concerning metabolism and toxicity, can only be obtained invasively by surgical operation of patients. Often, the span of life and activity of primary human hepatocytes are limited and phenotypical changes occur early in the cell cultivation.[3] In contrast, hepatocyte-like cell lines derived by tumors or oncogene immortalizations are easy to access and can be reproduced easily. One drawback of these hepatocyte-like cell lines is the loss of liver-specific functions, especially their enzyme activities.[4] One example of such a cell line is the often used HepG2 cell line.

The unique human liver HepaRG stem cells were derived from a differentiated human hepatoma at the Institut national de la santé et de la recherche médicale (Inserm) of France.[5] During cultivation, a differentiation process can be initiated. Thereafter, the HepaRG cells differentiate in two cell types: (a) hepatocyte-like and (b) biliary epithelial-like cells. However, within the frame of this work, we focus on the spectroscopic investigation of the cells differentiated into hepatocyte-like cells, as they are used for screening of treatments against hepatocellular carcinoma. In contrast to other cultivated liver cells, these hepatocyte-like cells are capable of exerting a huge amount of liver-specific functions, including cytochrome P450 expression.[5] In addition, the cells can be stored by cryopreservation and their functional activities are quite stable. Therefore, the HepaRG cell line is well suited for toxicity and metabolism studies and is a promising alternative to primary human hepatocytes with its described drawbacks.[2] To shed light on the unique character of the HepaRG cells and to reveal the changes occurring with the cell differentiation, Raman spectroscopy was applied as cell characterization tool. Prospectively, the spectroscopic monitoring of the differentiation process may allow ensuring that the drug screening is performed on well differentiated hepatocyte-like HepaRG cells. Moreover, it can be used for comparison of the differentiation efficacy depending on the substrates and laboratory setting.

Raman spectroscopy probes inherent molecular vibrations that provide a highly specific molecular fingerprint of the biochemical composition of biological samples, such as cells and tissues.[6] As Raman spectroscopy is a noninvasive vibrational spectroscopy approach and does not require any labels, it needs minimal sample preparation. Thus, vibrational spectroscopy was found suitable for the investigation of stem cells.[7] Particularly, Fourier-transform infrared spectroscopy has been applied to monitor the differentiation of rat bone marrow mesenchymal stem cells,[8] mouse stem cell-derived hepatocytes,[9] human mesenchymal stem cells,[10] and also embryonic stem cells of murine[11] and human origin.[12] On the other side, Raman spectroscopy has been applied for differentiating murine embryonic stem cells[13] and assessment of human osteoblast-like cells.[14] Moreover, Raman spectra of biological cells are useful to precisely describe the cellular content[15] and intracellular response on a treatment.[16] In combination with an optical microscope, a high spatial resolution can be achieved. The great potential of this so-called Raman spectroscopic imaging to retrieve biochemical information from single cells,[17] biological tissues,[18, 19] and body liquids[19] has been demonstrated. Besides these studies, other cell analytics[20, 21] and classification approaches[22] based on Raman spectroscopy have been published within the last years.

In this contribution, we report about the characterization of HepaRG cells by means of Raman spectroscopic imaging. Furthermore, a comparison of the spectral signatures of HepaRG and the often used liver carcinoma cell line HepG2 is shown. Moreover, an overview of the classification between two cell lines and a comparison between differentiated hepatocyte-like and nondifferentiated HepaRG cells are presented within this article. Overall, this study shows that a monitoring of the cell differentiation process by means of Raman spectroscopy is possible. Therefore, it shows an essential prerequisite to monitor the cancer treatment process of these cells by means of nondestructive, label-free Raman spectroscopy. This might improve the understanding of the working principle of the treatment.

2 | EXPERIMENTAL SECTION

2.1 | Cell cultivation

The cells were incubated at 5% CO_2, at 37 °C, in a humidified atmosphere. HepG2 cells (Cell Lines Service GmbH) were cultured in DMEM/F12 (Thermo Fisher Scientific Inc.) supplemented with 10% fetal bovine serum (Biochrom GmbH). The HepaRG cells (Biopredic International, Rennes, France; Figure 1) were treated according to Gripon *et al.* 2002.[23] Shortly, in order to achieve cell proliferation, HepaRG cells were cultured in Williams E without Glutamine (Thermo Fisher Scientific Inc.) supplemented with 10% fetal bovine serum (Biochrom GmbH), 5 μg/ml Insulin (Merck KGaA), 50 μg/ml Hydrocortisol-Hemisuccinat (Merck KGaA), and GlutaMax (Thermo Fisher Scientific Inc.). In order to achieve differentiation of the HepaRG into hepatocytes and bile duct epithelial cells, the culture medium was

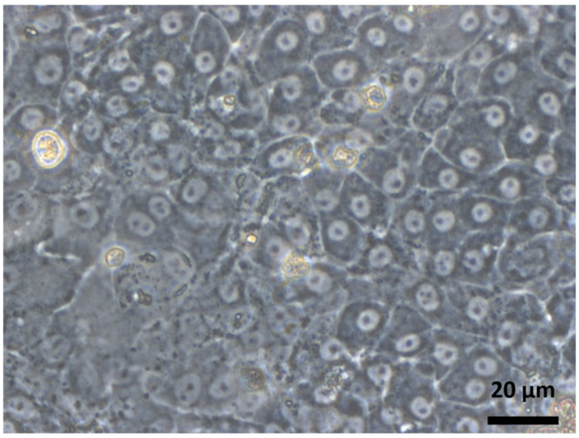


FIGURE 1 A bright-field microscopic image of HepaRG cells

prepared as just described and supplemented with 2% DMSO (Merck KGaA) in addition.

2.2 | Sample preparation

For the Raman spectroscopic experiments, cells were harvested by trypsinization and seeded on calcium fluoride slides in a 12-well plate. The same cultivation conditions (5% CO_2 at 37 °C) and the same medium as mentioned above were used. HepaRG cells, differentiated into hepatocyte-like and biliary epithelial cells, were transferred to the calcium fluoride slides, where the differentiated cells were cultivated for 5 days. HepG2 and undifferentiated HepaRG cells were cultivated on calcium fluoride slides up to 3 days under the microscopic control of the cell density on the slide. At least three slides for each batch were used. After the cultivation, the cells were fixed in 4% formaldehyde solution ("Rotifix", Carl Roth GmbH & Co. KG) and stored in phosphate buffered saline (PBS) solution (Biochrom AG) at 4 °C until the Raman measurement.

2.3 | Raman measurements

For Raman spectroscopic measurements, a commercially available confocal Raman microscope[24] CRM 300 (WITec GmbH, Germany) was used. The microscope was equipped with a diode laser operating at 785 nm with a laser power between 80 and 100 mW on the sample. Laser light was focused with a 60×/NA 1.0 water immersion objective on fixed cells in PBS solution. HepG2, HepaRG, and differentiated hepatocyte-like HepaRG cells were located on the slides by visual inspection of bright-field microscopic images. Then, single cells were Raman spectroscopic measured using the mapping mode by collecting Raman spectra at each point of a grid. A step size of 0.5 μm and an acquisition time of 2 s per single spectrum were used. In total, 37 HepG2 cells, 49 hepatocyte-like HepaRG, and 37 nondifferentiated HepaRG were measured. Although the used growth media were different, we do not expect them to be present while measuring, because the cells were fixed and stored in PBS solution prior to Raman measurements. Therefore, the differences related to the cell type should feature much larger impact on the Raman spectra than the difference in the growth medium.

2.4 | Data analysis

All the described data analysis was done in the statistical language R.[25] Prior to applying chemometric methods, the Raman spectra were pretreated. The development of this preprocessing workflow is important to achieve reliable results.[26] Thus, automatic calibration, background correction,[27] and noise elimination[28] have to be performed within the data preprocessing routine. In particular, the background was corrected using the sensitive nonlinear iterative peak (SNIP) clipping algorithm with 100 iterations, then the spectra were vector normalized, and a dimension reduction by means of principal component analysis (PCA) was applied. The average preprocessed Raman spectra and the standard deviations for each cell line are shown in Figure 4. Furthermore, the analysis was repeated with another baseline correction to verify that the discovered results are not dependent on the selected background correction method. In an alternative processing pipeline, the asymmetric least squares (ALS) baseline correction[29] was used instead of SNIP.

Two different approaches were used for the visualization of the cells according to the preprocessed Raman data. First, the unsupervised k-means clustering (Figure 3a) was utilized, which allows to segment the Raman spectral scan

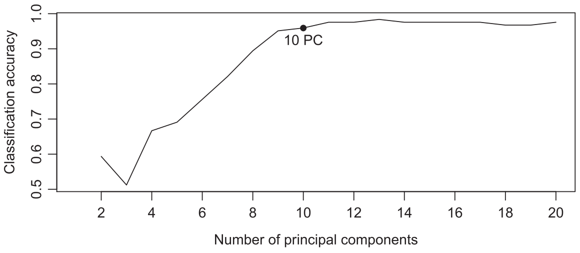


FIGURE 2 Leave-one-out-cross-validation of the 3-class LDA. The leave-one-cell-out-cross-validation accuracy of the LDA classification between the three cell lines (HepG2, nondifferentiated HepaRG, and differentiated hepatocyte-like HepaRG) for different quantity of variables is shown. The bend, or saturation, point (indicated with a dot) was estimated as an optimum for the number of PC. LDA = linear discriminant analysis; PC = principal component

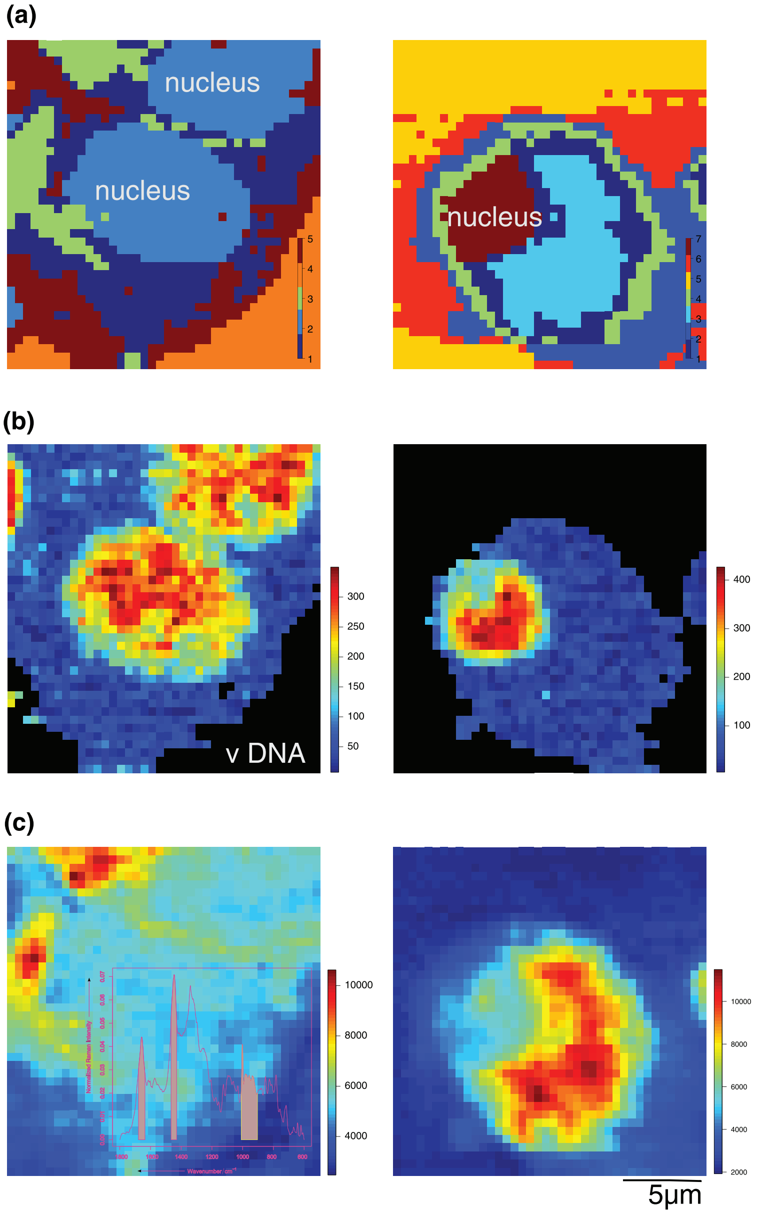


FIGURE 3 Raman spectroscopic imaging of hepatocyte-like HepaRG cells. The cell visualization with three different imaging approaches for two different cells is shown: (a) cell morphology by k-means cluster analysis; (b) nucleus visualization using integration of a DNA marker band at 785cm^{-1} (O-P-O stretching of DNA backbone); (c) identification of cell Raman spectra applying an integration over characteristic Raman peaks of cellular components. The scale bar is valid for all images

according to spectral features. Thus, different clusters visualize different cell components and the background. In contrast to the clustering method, which uses the whole available range of spectral information, a band integration approach (Figure 3b,c) uses the peak area to visualize a single molecular component.

In order to take the analysis from an unsupervised visualization level to the level of cell classification among different cell lines, a linear discriminant analysis (LDA)-based classification was carried out. The model was constructed based on the average Raman spectra of the cells. An advantage of such a linear method is the simplicity of the interpretation and its robustness. Thus, even if the LDA classification is built on the principal components (PCs) extracted by a PCA, the LDA loadings can be analyzed manually after back-projecting into spectral intensities using the inverse of PCA rotational matrix.

However, PCA is often followed by nontrivial issue of selecting the optimal number of PC components for model construction. Although the main reason for applying the dimension reduction is avoiding overfitting, it is also important to keep the key features of the spectra. Thus, the number of the variables should not be too small either. To select the optimal number, a leave-one-cell-out-cross-validation of the classification between the three cell lines was performed for different numbers of PCs from two to 20. The optimum has been selected as a saturation point of the LDA classification accuracy over the number of components (Figure 2, Figure S1). Moreover, the contribution to the data variance is less than 0.5% for each PC with index higher than 10. Thus, usage of more variables for the analysis was avoided.

3 | RESULTS AND DISCUSSION

3.1 | Investigation of differentiated hepatocyte-like HepaRG cells

For the characterization of the differentiated HepaRG cells, the Raman spectra of 49 hepatocyte-like cells were measured using the imaging mode. In detail, on every point of a predefined grid covering the cell area and its surrounding, a Raman spectrum was recorded. The step size was set to 0.5 μm. To handle this huge amount of spectra, different chemometric methods were applied. The result obtained from these analytical methods is visualized in Figure 3 for two examples of differentiated hepatocyte-like HepaRG cells. First of all, the multivariate k-means cluster analysis was carried out to visualize the different cell compartments (Figure 3a). Furthermore, the assignment of the nucleus was verified by the univariate method of band integration over the Raman peak at 785 cm^{-1} assigned to the vibration of the sugar-phosphate backbone of the DNA and RNA.[30] An integration over this band visualizes the position and the shape of the cell nucleus (Figure 3b). The non-cell area of the scan is indicated by black pixels. This background area was determined by setting a threshold and applying it to the false-color image shown in Figure 3c, which was obtained by an integration of three prominent Raman bands of biological substances.[22] The first band in the range from 905 to 1,010 cm^{-1} features contributions of the C-C stretching vibration[30] at 936 cm^{-1} and symmetric ring breathing mode[30] of phenylalanine at 1,002 cm^{-1}. The other two bands: 1,431 – 1,467 cm^{-1} and 1,637 – 1,679 cm^{-1} are affiliated to CH$_2$ deformation mode[30] at 1,149 cm^{-1} and Amid I[30] at 1,658 cm^{-1}. All the mentioned bands and a few other prominent peaks are depicted in Figure 4 and Figure S2 (for an alternative preprocessing). The additionally highlighted bands can be affiliated to CH$_2$ deformation mode[30] at 1,301 cm^{-1}, Amid III[31, 32] at 1,262 cm^{-1}, and C-C stretching mode at 1,128 cm^{-1}.

Furthermore, we performed a pairwise comparison of cytoplasm and cell nucleus differences between the cell types. To do so, we performed clustering on 12 selected cells (four cells per class), which featured clear distinction between nuclei and cytoplasm. The clustering results (Figure S4) and pairwise differences of spectra from subcellular regions, which are related to nuclei (Figure S5) and cytoplasm (Figure S6), are shown in the Supporting Information. In summary, it could be shown that the cell nucleus region shows the largest differences between the cell types.

The Raman spectroscopic imaging results allowed no further insights about the properties of the HepaRG cell line. Therefore, the comparison between differentiated hepatocyte-like and nondifferentiated HepaRG cells was shifted into the focus and was investigated subsequently. Additionally, a comparison of the spectral signature of HepaRG cells with the spectral signature of the often used liver carcinoma cell line HepG2 was carried out.

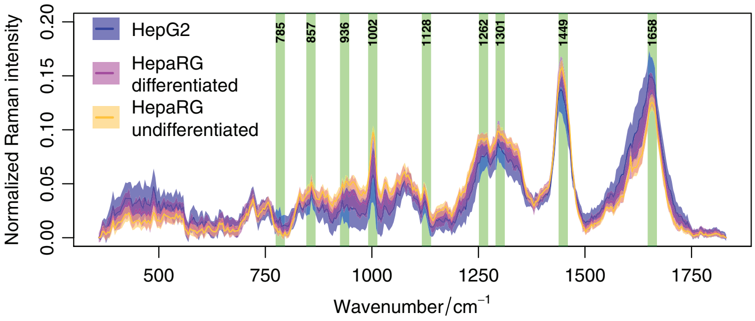


FIGURE 4 Mean Raman spectra. The solid lines depict mean Raman spectra for each of the investigated cell lines, and the semitransparent areas around the spectra show the standard deviation of the cell line groups. The detailed description of the labeled Raman peaks can be found in Section 3.1

3.2 | Discrimination between differentiated hepatocyte-like HepaRG cells, nondifferentiated HepaRG cells, and HepG2 cells

In addition to the visualization of the cell morphology of the HepaRG cells, a LDA was applied to discriminate between differentiated hepatocyte-like HepaRG cells, nondifferentiated HepaRG cells, and HepG2 cells, respectively. The result of the construction and evaluation of the LDA is presented in Figure 5, where the LD values of the mean Raman spectrum of each measured cell is visualized. Three clearly separated point clouds show that the three cell types are distinguishable. In detail, the point cloud of the standard HepG2 cell line and the point cloud of the nondifferentiated HepaRG cells are much closer to each other compared with the differentiated HepaRG cells. This finding is supported by visualizing the 95% confidence intervals of the LDA score values by ellipses (Figure 5). Thus, the differentiation of the HepaRG cells during cultivation has an important impact on the Raman spectroscopic signature of the liver cells and has a significant contribution to the discrimination of the cells. The result, that the nondifferentiated HepaRG cells and the HepG2 cells have a much more similar Raman spectroscopic signatures in comparison with the differentiated HepaRG cells correlates with the knowledge, that HepG2 cells and nondifferentiated HepaRG cells have less liver-specific functions and a lower enzyme activity compared with the differentiated hepatocyte-like HepaRG cells.[2]

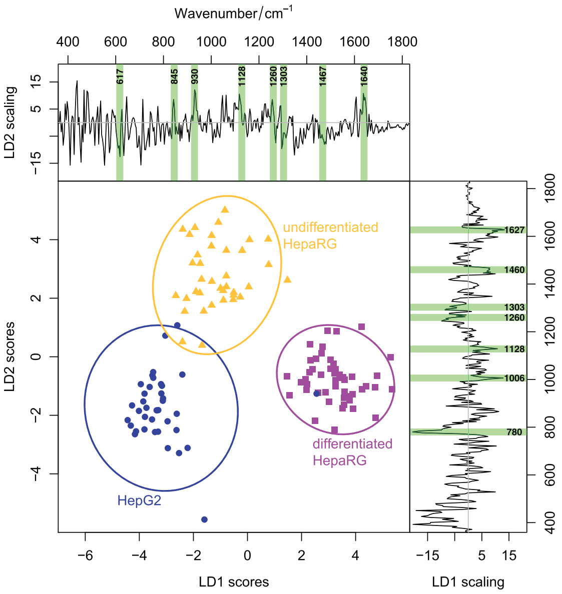


FIGURE 5 LDA classification scores and scaling vectors. Every point represents the LD scores of a mean Raman spectrum of a cell. The points of HepG2 cells (circles), nondifferentiated HepaRG cells (triangles), and differentiated hepatocyte-like HepaRG cells (squares) form distinct clouds. Additionally, the ellipses outline confidence ellipses on a level of 95% for the scores of each cell line. The LD scaling vectors related to the LD scores are shown on the right and the upper side of the image. These vectors represent the separation of the groups and characterize them according to their Raman spectroscopic signatures. See text for further details. LDA = linear discriminant analysis

Table 1 shows the corresponding confusion matrix for the classification with the above described LDA model: Each mean Raman spectrum of the differentiated hepatocyte-like HepaRG cells was correctly classified; 34 out of 37 mean Raman spectra of the standard HepG2 cells and 35 out of 37 Raman mean spectra of nondifferentiated HepaRG cells were correctly classified. As a result, the accuracy accounts to 96%. In addition, three of the false classified cells showed a stressed or apoptotic cell morphology in the visual analysis.[33] This could be a reason for the misclassification. As a result, the applied LDA model was well suited to distinguish between the HepG2, the nondifferentiated HepaRG, and the differentiated hepatocyte-like HepaRG cells. In particular, the differentiated hepatocyte-like HepaRG cells showed a higher dissimilarity from the other two investigated groups (HepG2 and nondifferentiated HepaRG cells). To ensure that the discovered findings do not originate from the choice of the baseline correction method, we repeated the analysis with ALS baseline correction and verified the similarity of the results (see Figures S1–S3). To avoid

TABLE 1 Confusion matrix of the LDA

	True labels		
Predicted labels	HepG2	HepaRG hepatocyte-like	HepaRG nondifferentiated
HepG2	**34**	0	2
HepaRG hepatocyte-like	1	**49**	0
HepaRG nondifferentiated	2	0	**35**

Note. The leave-one-cell-out-cross-validation shows high performance of classification between differentiated hepatocyte-like HepaRG cells, nondifferentiated HepaRG cells, and HepG2 cells. The diagonal elements of the confusion matrix refer to the correct predictions and are shown bold in the table. The accuracy was around 96%.

the introduction of a bias by manual parameter selection, we performed ALS baseline correction with the default parameters ($\lambda = 6$, $p = 0.05$, maxit = 20).[34] The resulting classification accuracy differed slightly from the preprocessing pipeline using the SNIP background correction. However, the general outcome of the analysis had the same trend. Thus, the alternative model (Figure S3) had similar loadings and scores (but reversed).

The analysis of the mean Raman spectra of the HepG2, the nondifferentiated HepaRG, and the differentiated hepatocyte-like HepaRG cells (Figure 4) identified spectral signatures of the cells lines. However, the investigation of the LD scaling vectors (Figure 5) did not identify marker bands, which are responsible for the good classification. Nevertheless, the sum of minor differences within the Raman spectroscopic signatures of the three cell types enabled the reliable classification. To validate the significance of peak contributions in the loadings, we compared pairwise the difference spectra of subcellular areas for different cell types for SNIP (Figure S8) and ALS (Figure S9) baseline correction. This validation revealed that the main differences occur in the nuclei regions, but there are no significant differences present below 600 cm^{-1}. Furthermore, both baseline estimations led to similar difference spectra.

The LD1 scaling vector in Figure 5 represents separation of differentiated HepaRG cells from HepG2 cells and undifferentiated HepaRG cells. This LD1 vector shows a change in the ratio between Amid I and Amid III[31, 32] (1,627 and 1,260 cm^{-1}) and a change of the deformation modes[31] of CH and CH_2 (1,303 and 1,460 cm^{-1}). Moreover, the scaling vector LD1 shows a decrease of O-P-O stretching vibrations of the sugar-phosphate backbone of DNA[31] (780 cm^{-1}). Furthermore, the LD1 shows an increase of amino acid related bands[31, 32] (1,006 cm^{-1}) and of the C-C stretching band[30] (1,128 cm^{-1}). On another side, the LD2 scaling vector, which separates HepG2 and undifferentiated HepaRG, shows a decrease of CH and CH_2 deformation modes and C-C ring twist of phenylalanine[30] (617 cm^{-1}). Additionally, an increase of the other mentioned peaks is visible in LD2 scaling vector. Among the peaks found in the LDs, the difference spectra clearly feature the sugar-phosphate backbone vibration of DNA at 780 cm^{-1} and the amino acid related vibration at 1,006 cm^{-1}. However, other peaks in the difference spectra are not strongly influencing the classification. We assume that these peaks reflect the Raman band variation in general, rather than being specific to the cell type.

4 | CONCLUSION

In the present study, the Raman spectroscopic signature of the differentiated hepatocyte-like liver cells HepaRG was characterized by comparing it with the Raman spectroscopic signature of undifferentiated HepaRG cells and the well-established liver carcinoma cell line HepG2. By utilizing an LDA, the three cell types were successfully classified with an accuracy of 96%. Thereby, the Raman spectroscopic signature of the promising differentiated hepatocyte-like HepaRG cells is distinct from the undifferentiated HepaRG and HepG2 cells, whose Raman spectroscopic signatures were closer to each other. In summary, the differences in chemical composition of the cell lines were shown by Raman spectroscopic characterization. Prospectively, the great potential of Raman imaging to retrieve biochemical information from single cells may be used to study the interaction between HepaRG cells and several drugs as well as enzyme activity assays. Further studies will check out multiple anticancer drug tests using the HepG2, nondifferentiated stem HepaRG, and differentiated hepatocyte-like HepaRG cells. Additionally, a study proving the reliability and stability of the Raman spectroscopic prediction of the cell differentiation based on larger sample sizes might be performed.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Nora Nowak and Julian Stobbe for measurements and analysis during their practical course as part of their study of chemistry. We thank R. Claus and J. Fischer (Dept. of Anesthesiology and Intensive Care Medicine, JUH) for support with HepG2 cells.

Financial support by the BMBF via the integrated research and treatment center *Center for Sepsis Control and Care* (FKZ 01EO1502) and the project *PhotoSkin* (FKZ 13N13243) are gratefully acknowledged. Funding from the EU via the project *HemoSpec* (FP 7, CN 611682) and co-founding of the project *PhotoSkin* via the 7th framework program of the EU are highly acknowledged. Financial support by the Carl-Zeiss Foundation and support by the Free State of Thuringia (FKZ 2015 FGI 0011 and 2016 FGI 0010) with co-financing from the European Union within the European Regional Development Fund (EFRE) are greatly acknowledged.

ORCID

Thomas Bocklitz http://orcid.org/0000-0003-2778-6624

REFERENCES

[1] K. E. Lasser, P. D. Allen, S. J. Woolhandler, D. U. Himmelstein, S. M. Wolfe, D. H. Bor. *JAMA* **2002**, *287*, 2215.

[2] A. Guillouzo, A. Corlu, C. Aninat, D. Glaise, F. Morel, C. Guguen-Guillouzo. *Chem.-Biol. Interact.* **2007**, *168*, 66.

[3] E. LeCluyse. *Eur. J. Pharm. Sci.* **2001**, *13*, 343.

[4] S. Wilkening, F. Stahl, A. Bader. *Drug Metab. Dispos.* **2003**, *31*, 1035.

[5] P. Gripon, S. Rumin, S. Urban, J. Le Seyec, D. Glaise, I. Cannie, C. Guyomard, J. Lucas, C. Trepo, C. Guguen-Guillouzo. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 15655.

[6] T. W. Bocklitz, S. Guo, O. Ryabchykov, N. Vogler, J. Popp. *Anal. Chem.* **2016**, *88*, 133.

[7] J. W. Chan, D. K. Lieu. *J. Biophotonics* **2009**, *2*, 656.

[8] D. Ye, W. Tanthanuch, K. Thumanu, A. Sangmalee, R. Parnpai, P. Heraud. *Analyst* **2012**, *137*, 4774.

[9] K. Thumanu, W. Tanthanuch, D. Ye, A. Sangmalee, C. Lorthongpanich, R. Parnpai, P. Heraud. *J. Biomed. Opt.* **2011**, *16*, 057005.

[10] C. Chonanant, N. Jearanaikoon, C. Leelayuwat, T. Limpaiboon, M. J. Tobin, P. Jearanaikoon, P. Heraud. *Analyst* **2011**, *136*, 2542.

[11] D. Ami, T. Neri, A. Natalello, P. Mereghetti, S. M. Doglia, M. Zanoni, M. Zuccotti, S. Garagna, C. A. Redi. *Biochim. Biophys. Acta Mol. Cell Res.* **2008**, *1783*, 98.

[12] P. Heraud, E. S. Ng, S. Caine, Q. C. Yu, C. Hirst, R. Mayberry, A. Bruce, B. R. Wood, D. McNaughton, E. G. Stanley, A. G. Elefanty. *Stem Cell Res.* **2010**, *4*, 140.

[13] I. Notingher, I. Bisson, J. M. Polak, L. L. Hench. *Vib. Spectrosc.* **2004**, *35*, 199.

[14] L. L. McManus, F. Bonnier, G. A. Burke, B. J. Meenan, A. R. Boyd, H. J. Byrne. *Analyst* **2012**, *137*, 1559.

[15] I. W. Schie, J. W. Chan. *J. Raman Spectrosc.* **2016**, *47*, 384.

[16] H. K. Yosef, L. Mavarani, A. Maghnouj, S. Hahn, S. F. El-Mashtoly, K. Gerwert. *Anal. Bioanal. Chem.* **2015**, *407*, 8321.

[17] F. Draux, P. Jeannesson, A. Beljebbar, A. Tfayli, N. Fourre, M. Manfait, J. Sulé-Suso, G. D. Sockalingum. *Analyst* **2009**, *134*, 542.

[18] F. M. Lyng, D. Traynor, I. R. M. Ramos, F. Bonnier, H. J. Byrne. *Anal. Bioanal. Chem.* **2015**, *407*, 8279.

[19] H. J. Butler, L. Ashton, B. Bird, G. Cinque, K. Curtis, J. Dorney, K. Esmonde-White, N. J. Fullwood, B. Gardner, P. L. Martin-Hirsch, M. J. Walsh, M. R. McAinsh, N. Stone, F. L. Martin. *Nat. Protoc.* **2016**, *11*, 664.

[20] K. Galler, K. Bräutigam, C. Grosse, J. Popp, U. Neugebauer. *Analyst* **2014**, *139*, 1237.

[21] K. Hartmann, M. Becker-Putsche, T. Bocklitz, K. Pachmann, A. Niendorf, P. Rösch, J. Popp. *Anal. Bioanal. Chem.* **2012**, *403*, 745.

[22] K. Bräutigam, T. Bocklitz, M. Schmitt, P. Rösch. *J. Popp, ChemPhysChem* **2013**, *14*, 550.

[23] P. Gripon, S. Rumin, S. Urban, J. L. Seyec, D. Glaise, I. Cannie, C. Guyomard, J. Lucas, C. Trepo, C. Guguen-Guillouzo. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 15655.

[24] L. Baia, K. Gigant, U. Posset, R. Petry, G. Schottner, W. Kiefer, J. Popp. *Vib. Spectrosc.* **2002**, *29*, 245.

[25] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria **2016**.

[26] T. Bocklitz, A. Walter, K. Hartmann, P. Rösch, J. Popp. *Anal. Chim. Acta* **2011**, *704*, 47.

[27] S. Guo, T. Bocklitz. *J. Popp, Analyst* **2016**, *141*, 2396.

[28] O. Ryabchykov, T. Bocklitz, A. Ramoji, U. Neugebauer, M. Foerster, C. Kroegel, M. Bauer, M. Kiehntopf, J. Popp. *Chemom. Intell. Lab. Syst.* **2016**, *155*, 1.

[29] P. H. Eilers, H. F. Boelens. *Leiden University Medical Centre Report* **2005**, *1*, 5.

[30] I. Notingher, L. Hench. *Expert Rev. Med. Devices* **2006**, *3*, 215.

[31] C. Krafft, T. Knetschke, R. H. W. Funk, R. Salzer. *Vib. Spectrosc.* **2005**, *38*, 85.

[32] I. Notingher, G. Jell, P. L. Notingher, I. Bisson, O. Tsigkou, J. M. Polak, M. M. Stevens, L. L. Hench. *J. Mol. Struct.* **2005**, *744*, 179.

[33] C. Krafft, T. Knetschke, R. H. W. Funk, R. Salzer. *Anal. Chem.* **2006**, *78*, 4424.

[34] K. H. Liland, B.-H Mevik, baseline: Baseline Correction of Spectra, **2015**.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Ryabchykov O, Bräutigam K, Galler K, et al. Raman spectroscopic investigation of the human liver stem cell line HepaRG. *J Raman Spectrosc*. 2018;49:935–942. https://doi.org/10.1002/jrs.5392

[P6] Fusion of MALDI spectrometric imaging and Raman spectroscopic data for the analysis of biological samples

Reprinted from [Ryabchykov, O., Popp, J. and Bocklitz, T., 2018. Fusion of MALDI spectrometric imaging and Raman spectroscopic data for the analysis of biological samples. *Frontiers in Chemistry*, *6*, p.257] in accordance with the Creative Commons license 4.0.

Erklärungen zu den Eigenanteilen des Promovenden sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an der Publikation.

[1]Ryabchykov, O., [2]Popp, J., [3]Bocklitz, T., 2018. Fusion of MALDI spectrometric imaging and Raman spectroscopic data for the analysis of biological samples. *Frontiers in Chemistry*, *6*, p.257			
Beteiligt an (Zutreffendes ankreuzen)			
	1	2	3
Konzeption des Forschungsansatzes	**X**	X	X
Planung der Untersuchungen	**X**		X
Datenerhebung			
Datenanalyse und Interpretation	**X**		X
Schreiben des Manuskripts	**X**	X	X
Vorschlag Anrechnung Publikationsäquivalent	**1.0**		

frontiers in Chemistry

ORIGINAL RESEARCH
published: 16 July 2018
doi: 10.3389/fchem.2018.00257

Fusion of MALDI Spectrometric Imaging and Raman Spectroscopic Data for the Analysis of Biological Samples

Oleg Ryabchykov[1,2], *Juergen Popp*[1,2] *and Thomas Bocklitz*[1,2]*

[1] *Spectroscopy and Imaging Research Department, Leibniz Institute of Photonic Technology, Member of Leibniz Health Technology, Jena, Germany,* [2] *Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich Schiller University Jena, Jena, Germany*

Despite of a large number of imaging techniques for the characterization of biological samples, no universal one has been reported yet. In this work, a data fusion approach was investigated for combining Raman spectroscopic data with matrix-assisted laser desorption/ionization (MALDI) mass spectrometric data. It betters the image analysis of biological samples because Raman and MALDI information can be complementary to each other. While MALDI spectrometry yields detailed information regarding the lipid content, Raman spectroscopy provides valuable information about the overall chemical composition of the sample. The combination of Raman spectroscopic and MALDI spectrometric imaging data helps distinguishing different regions within the sample with a higher precision than would be possible by using either technique. We demonstrate that a data weighting step within the data fusion is necessary to reveal additional spectral features. The selected weighting approach was evaluated by examining the proportions of variance within the data explained by the first principal components of a principal component analysis (PCA) and visualizing the PCA results for each data type and combined data. In summary, the presented data fusion approach provides a concrete guideline on how to combine Raman spectroscopic and MALDI spectrometric imaging data for biological analysis.

Keywords: MALDI-TOF, Raman imaging, data combination, data fusion, normalization, PCA

OPEN ACCESS

Edited by:
Hoang Vu Dang,
Hanoi University of Pharmacy, Vietnam

Reviewed by:
Xia Guan,
Louisiana State University,
United States
Ennio Carbone,
Università degli Studi Magna Græcia
di Catanzaro, Italy
Frédéric Jacques Cuisinier,
Université de Montpellier, France
Anna V. Sharikova,
University at Albany, United States

***Correspondence:**
Thomas Bocklitz
thomas.bocklitz@uni-jena.de

Specialty section:
This article was submitted to
Analytical Chemistry,
a section of the journal
Frontiers in Chemistry

Received: *17 December 2017*
Accepted: *08 June 2018*
Published: *16 July 2018*

Citation:
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. 6:257. doi: 10.3389/fchem.2018.00257

INTRODUCTION

Different analytical methods could be utilized for biomedical analysis (e.g., cells, and tissues, etc.) to highlight a certain aspect of the sample e.g., morphological microstructure, distribution of electronic chromophores, molecule classes, or special proteins. Among the label-free imaging approaches, matrix-assisted laser desorption/ionization (MALDI) spectrometry, and Raman microscopy are certainly among the most powerful imaging techniques for the investigation of biomedical samples. Raman spectroscopy is a non-destructive spectroscopic method, which provides complex molecular information about the general chemical composition of the sample with a rather high spatial resolution (Abbe limit) to highlight subcellular features (Kong et al., 2015). The drawback of Raman imaging lies in its weak scattering efficiency that makes sampling time rather long for large area imaging. Raman spectroscopic imaging has demonstrated its potential for biomedical diagnosis in numerous cancer-related studies (Tolstik et al., 2014), biological material analysis (Butler et al., 2016), cell characterization studies (Ramoji et al., 2012), and many other biomedical applications (Matousek and Stone, 2013; Ember et al., 2017).

On the other side, MALDI mass spectrometry provides information on specific substances, such as lipids or proteins (Fitzgerald et al., 1993). MALDI is a soft ionization technique utilized for mass-spectrometric imaging (Gessel et al., 2014) to determine large organic molecules and biomolecules undetected by conventional ionization techniques. This technique was employed in clinical parasitology (Singhal et al., 2016), microbial identification (Urwyler and Glaubitz, 2016), and cancer tissue investigation (Hinsch et al., 2017).

Raman spectroscopic and MALDI mass spectrometric imaging both offer a high molecular sensitivity. Moreover, Raman spectroscopy has been sequentially applied together with different mass spectrometric techniques to address a variety of biological tasks such as characterization of succinylated collagen (Kumar et al., 2011), investigation of microbial cells (Wagner, 2009), identification of fungal strains (Verwer et al., 2014) and characterization of lipid extracts from brain tissue (Köhler et al., 2009). In all the aforementioned studies, the Raman and mass spectrometric data are analyzed separately, and then summarized or compared to each other (Masyuko et al., 2014; Bocklitz et al., 2015; Muhamadali et al., 2016). To significantly increase the information content, Raman spectroscopic and MALDI mass spectrometric imaging data have to be co-registered (Bocklitz et al., 2013) followed by a high-level (distributed) data fusion. It means that each data type is analyzed separately to obtain the respective scores, which are then fused together. Alternatively, spectroscopic imaging can be used for mapping an area that is suitable for further investigation by means of MALDI spectrometric imaging (Fagerer et al., 2013) or a certain mass peak is used to define an area, from which the Raman spectra are analyzed (Bocklitz et al., 2013). Such a hierarchical pipeline corresponds to a decentralized data fusion approach.

In the present work, we introduced an analytical method to perform a low-level (centralized) fusion of Raman and MALDI imaging data. Because the experimental implementation of correlated imaging is challenging in many aspects (Masyuko et al., 2013), we utilized a computational approach to combine imaging data obtained by MALDI spectrometry and Raman spectroscopy. The correlation of Raman spectroscopy with mass spectrometric imaging techniques such as MALDI (Ahlf et al., 2014) or secondary ion mass spectrometry (SIMS) (Lanni et al., 2014) have proved its usefulness for biological applications. Moreover, a combination of MALDI imaging data with optical microscopy could attenuate instrumental effects (Van De Plas et al., 2015), and a joint analysis of vibrational and MALDI mass spectra could provide valuable information on brain tissue (Van De Plas et al., 2015; Lasch and Noda, 2017). Nevertheless, even if Raman and MALDI spectra are obtained by correlated imaging, each type of spectra shows its own specific features and should be preprocessed separately. Because the measurement techniques are based on different physical effects, the difference in data dimensionality and dynamic range can affect the contribution of each datatype in the analysis. Therefore, a weighting coefficient that balances the influence of Raman spectroscopic and MALDI spectrometric data in the data fusion center is required.

MATERIALS AND METHODS

Experimental Details

We demonstrated the data fusion on an example dataset of MALDI spectrometric and Raman spectroscopic scans obtained from the same mouse brain sample (*Mus musculus*) of 10 μm cryosection. The sample was cut on a cryostat, and then dried on a precooled conductive ITO-coated glass slide. Subsequently, Raman spectra were obtained using a confocal Raman microscope CRM-alpha300R (WITec, Ulm, Germany) and excited with a 633 nm HeNe laser (Melles Griot). The laser irradiation was adjusted in order to have about 10 mW power. The laser was coupled through an optical fiber into a Zeiss microscope. A spectral map was obtained by a raster scan with a 25 μm grid with a dwell time of 2 s and a pre-bleaching time of 1 s.

After the Raman scan, MALDI mass spectrometric imaging was performed with a common matrix alpha-cyano 4-hydroxy cinnamic acid (5 mg/mL) in 50% acetonitrile and 0.2% trifluoracetic acid. The ImagePrep station (Bruker Daltonics) was used to prepare and apply the matrix on the sample. The MALDI-time-of-flight (MALDI-TOF) spectrometric map was obtained on a Ultraflex III MALDI-TOF/TOF mass spectrometer (Bruker Daltonics, Bremen, Germany). A "smartbeam" laser ($\lambda = 355$ nm, repetition rate 200 Hz) was used. The spectrometer was calibrated with an external standard, a peptide calibration mixture (Bruker Daltonics). The measurements were performed in the positive reflectron mode with 500 shots per spectrum and spatial resolution of 75 μm.

Further experimental details for both data types and an example of a hierarchical data fusion implementation can be found in the report by Bocklitz et al. (2013). Nevertheless, in the context of a further discussion, it is important to highlight that in MALDI mass spectrometric imaging a matrix suitable for the analysis of the lipid content was applied.

Preprocessing of Raman Spectroscopic Data

The influence of corrupting effects (e.g., cosmic spikes, fluorescence) on Raman spectra cannot be avoided completely. Thus, the development of complex preprocessing routines (Bocklitz et al., 2011) is required. To allow further analysis of the Raman spectra obtained with different calibrations, all spectra need to be interpolated to the same wavenumber axis (Dörfer et al., 2011). Moreover, keeping all the spectra in a single data matrix simplifies a further processing routine, so it is advantageous to perform the calibration as one of the first steps of the preprocessing workflow (**Figure 1**). Besides the wavenumber calibration, intensity calibration should be performed for the comparison of the measurements obtained with different devices

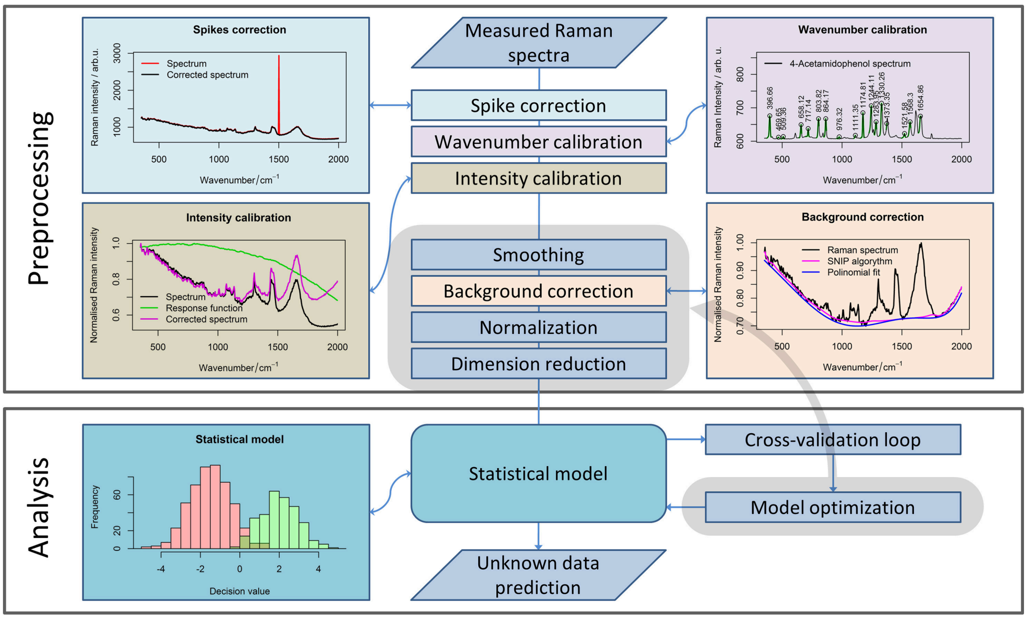


FIGURE 1 | A general pipeline for Raman data preprocessing. The workflow shows the main steps of the preprocessing routine necessary for robust Raman spectral analysis. Although some steps should be defined while planning the experiment, subsequent preprocessing methods (highlighted in gray area) and their parameters can be optimized for extracting the required information from the data.

or in the case where some changes in the measurement device have occurred (Dörfer et al., 2011).

The calibration is always needed for a reliable analysis, especially if the measurements were performed over a large time period, or settings of the device were changed between the measurements. In contrast, the following step within the preprocessing workflow (i.e., noise removal) is an optional step. However, among smoothing methods, only the running median with a relatively large window is applicable for cosmic ray noise removal. Unfortunately, filtering with a large window may corrupt the Raman bands themselves. Alternatively, 2–3 spectra per point can be acquired to eliminate the spikes that are not present in each spectrum. Nevertheless, this approach increases the measurement time dramatically. Therefore, this approach is not suitable for Raman imaging when a large number of spectra are recorded. Thus, specialized spike correction approaches like wavelet transform (Ehrentreich and Summchen, 2001), correlation methods (Cappel et al., 2010), calculation of the Laplacian of the spectral data matrix (Schulze and Turner, 2014; Ryabchykov et al., 2016), or a difference between the original and a smoothed spectrum (Zhang and Henson, 2007) must be used for spike removal.

The next step in the preprocessing workflow for Raman spectra is fluorescence background removal. In this work, the sensitive nonlinear iterative peak (SNIP) clipping algorithm (Ryan et al., 1988) was used for baseline estimation. The SNIP algorithm can be utilized for background estimation for a number of spectral measurements, like X-ray and mass spectra.

After baseline correction, the Raman spectra must be normalized (Afseth et al., 2006) to complete the basic preprocessing. There are several normalization approaches (e.g., vector normalization, normalization to integrated spectral intensity, or a single peak intensity value) that enhance the stability of the spectral data. In this work, we used vector normalization and l_1-normalization (Horn and Johnson, 1990) for Raman spectra. The difference between normalization to integrated spectral intensity and l_1-normalization is that the latter utilized absolute intensity values. As a result, the difference between both normalization approaches becomes more significant when negative values appear in the baseline corrected spectra due to noise or baseline correction artifacts.

Preprocessing of MALDI Spectrometric Data

Although the measurement techniques themselves differ dramatically for Raman and MALDI mass spectroscopic imaging data, the preprocessing of these data has a lot in common. The m/z values are set according to an internal calibration and may "float" slightly from one measurement to another. Therefore, a phase correction along the m/z axis must be performed within the preprocessing workflow (**Figure 2**) to ensure that the spectra obtained in different measurements are comparable. For this purpose, it is advisable to use the stable intense peaks within the phase correction routine (Gu et al., 2006).

From a theoretical point of view, MALDI spectra should not feature a spectral background. Nevertheless, in measured MALDI

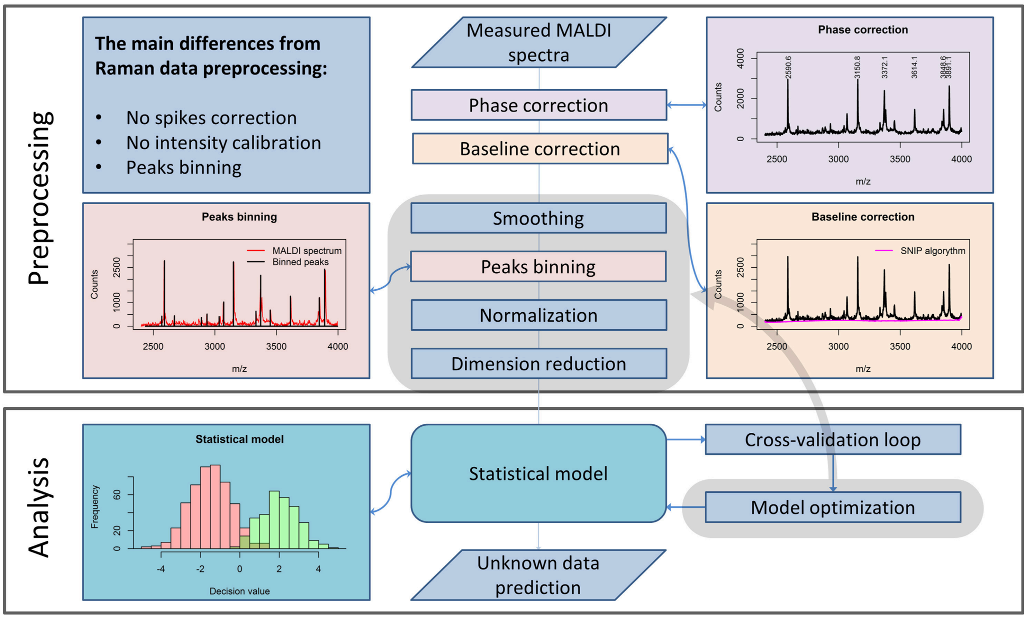


FIGURE 2 | A general pipeline for MALDI data preprocessing. The workflow shows the main steps of the preprocessing routine necessary for robust MALDI spectral data analysis and the main differences as compared to the Raman data preprocessing routine, described in **Figure 1**.

spectra a background is present. In literature, a background present in MALDI mass spectra is also known as "chemical noise background" (Krutchinsky and Chait, 2002). This type of noise results from matrix impurities and unstable ion clusters created during the sample scanning.

Similarly to Raman spectral preprocessing, the SNIP algorithm (Ryan et al., 1988) can be used to eliminate the background from mass spectra. Another complication in the analysis of MALDI spectra results from the fact that even after the phase correction, peak positions vary insignificantly among different spectra. An interpolation procedure, which is applied in Raman data preprocessing, would corrupt the sharp peaks found in MALDI spectra and is therefore not applied. To enable a direct comparison of the spectra, a binning procedure is applied. This procedure is based on the equalization of the m/z-values of peak positions within a certain range. Since the average peak width along the m/z axis increases with increased mass, the binning range is set with a so-called tolerance relative to the mass values. In contrast to Raman spectroscopy, intensity calibration for MALDI mass spectrometric imaging is not required. Nevertheless, normalization may be applied. Various types of normalization are used for MALDI mass spectroscopic imaging data: total ion count (TIC), vector norm (RMS), median, square root, logarithmic, and normalization to a noise level. In contrast to the Raman spectral data, MALDI mass spectra do not feature negative values. Thus, TIC normalization and normalization to l_1-norm, which is a sum of absolute values, are equal for MALDI spectra. If the significance level of the data is high, the normalization may be not necessary for the subsequent analysis.

Computational Details

For MALDI data acquisition and calibration, a flexImaging software version 3.0 (Bruker Daltonics) was used. The data processing was also performed in R (R Core Team, 2017) using packages akima (Gebhardt)[1], Peaks (Morhac)[2], readBrukerFlexData (Gibb)[3], rsvd (Erichson)[4], spatstat (Baddeley and Turner, 2005), and Spikes (Ryabchykov et al., 2016).

Prior to the data preprocessing and data fusion, the MALDI and Raman spectra were interpolated to the same (spatial) grid by utilizing a co-registration framework. Based on the false color images of Raman spectroscopic and MALDI spectrometric scans, 6 points clearly representing the same positions on every scan were manually selected. The coordinates of the Raman spectroscopic map were then transformed to the coordinate system of the MALDI mass spectrometric map. Subsequently, the Raman spectra were interpolated to the grid of the MALDI mass spectral map. To perform this interpolation, every point within the Raman grid was assigned to the nearest point within the MALDI grid. After that, the average of the Raman spectra, assigned to the same point within the MALDI grid, was

---

[1] Gebhardt, H. A. "akima: Interpolation of Irregularly and Regularly Spaced Data."
[2] Morhac, M. "Peaks: Peaks."
[3] Gibb, S. "readBrukerFlexData: Reads Mass Spectrometry Data in Bruker *flex Format."
[4] Erichson, N. B. "rsvd: Randomized Singular Value Decomposition."

calculated. Two spectral maps were thus obtained and aligned in a point-wise manner.

After the alignment, the Raman spectroscopic and MALDI mass spectrometric imaging data were preprocessed. During the preprocessing, the wavenumber calibration of the Raman spectra and the phase correction of MALDI spectra were performed. The MALDI mass spectrometric imaging data were subsequently subjected to noise removal, background correction, and TIC normalization. The Raman spectra were corrected for fluorescence background and vector normalized. The SNIP algorithm was used for background estimation in both cases.

After the preprocessing, Raman and MALDI mass spectral data differed in their dimensionality and in dynamic range. Data with different dynamic ranges would contribute unequally in a further analysis and consequently the spectral matrices have to be additionally weighed before performing the PCA. The weighting coefficient was selected as a ratio between the l_1-norms of the matrices, which are sums over the absolute values in the matrix. After the weighting, the data were combined in a single matrix and analyzed with a PCA. To illustrate the benefit of data fusion and weighting, we also analyzed the un-weighted data in a combined manner and each data type separately. We also investigated the case, where the same normalization approach was applied to both data types and no additional weighting is required. When the Raman spectra were normalized to the total spectral intensity, which is equivalent to TIC normalization of mass spectra, the data matrices had equal l_1-norms.

RESULTS AND DISCUSSION

Both Raman spectroscopic and MALDI mass spectrometric imaging data provide different insights into the chemical composition of the sample. Information on a broad range of molecules can be obtained from the Raman spectra. This information can be complemented by detailed information on lipid content, obtained from the MALDI data. To utilize both types of information together, a data fusion must be applied. This data fusion may be performed during different stages of the analysis workflow. Therefore, the architecture of the data processing workflow is dependent on the selected data fusion approach. These approaches can be divided into the following types (Castanedo, 2013):

- Centralized architecture (**Figure 3A**). The preprocessed data from different sources are combined in the data fusion center and are analyzed together.
- Decentralized architecture (**Figure 3B**). This scheme does not have a single data fusion center. The processing workflows are interacting at different processing stages. This architecture may provide multiple outputs or be represented as a hierarchical structure.
- Distributed architecture (**Figure 3C**). Each data type is preprocessed and analyzed separately. Subsequently, the output values are evaluated and combined to obtain a single result.

The decentralized and distributed architecture already showed their effectiveness for biomedical investigations (Bocklitz et al., 2013; Ahlf et al., 2014). The current work focuses on the centralized data fusion approach, also called low-level data fusion. In contrast to decentralized and distributed architectures, the centralized architecture shows a simpler workflow (**Figure 3A**). The data are combined in early steps of the analysis, directly after the preprocessing and even before the dimension reduction. At the data fusion center, where the different types of data are combined, an additional normalization or scaling of the data may be required to weight the influence of the different data types on the global model. The need for this weighting step arises from the differences in the data dimensionality, measurement units and dynamic ranges of the different measurement techniques. It is worth mentioning that the weighting is not a major issue in high-level data fusion approaches, which usually deal with standardized low-dimensional outputs of preliminary analysis in the data fusion center. However, a low-level data fusion (such as the applied centralized data fusion model) deals directly with preprocessed spectra of different types. Thus, the data scaling may dramatically influence extraction efficiency of the features.

To investigate the impact of data weighting, we searched for a marker that would allow an objective comparison of different data fusion and normalization approaches. This weighting scheme is designed for biological samples (i.e., a complex chemical composition), of which a large number of independent features have to be identified for appropriate description. By applying a PCA for dimension reduction, a large portion of the data variance is expected to be spread among multiple principal components (PCs) and the optimal approach should correspond to the slowest raise of the cumulative proportion of variance with a number of PCs.

The variances of the data explained by PCA are shown in the **Figure 4** where the normalization and fusion approaches (described in section Computational Details) are shown. Unfortunately, a direct comparison between cumulative proportions of variance obtained from Raman and MALDI mass spectral data, and their combined data is not suitable due to the different number of variables. However, different trends in the observed variance by the PCs in data with the same dimensionality can be interpreted. The left side of **Figure 4** shows that the variance of vector normalized Raman data is spread among a larger number of PCs than that of the total area normalized Raman data. This finding indicates that the vector normalization allows extracting a larger number of significant features from Raman data. Because the Raman spectra were vector normalized and the MALDI spectra were TIC normalized, the Raman data contribute more to the overall data variance than the MALDI data. Consequently, the PCA will focus on the variations in the Raman data and the variations in the MALDI data will have only a small influence. Alternatively, two datasets can be balanced by normalizing spectra of both types to their l_1-norms. By definition, this norm is a sum of absolute values. It takes dimensionality and scaling of the data into account, so no additional weighting is required. TIC normalization performed on MALDI data is already equal to l_1-normalization because

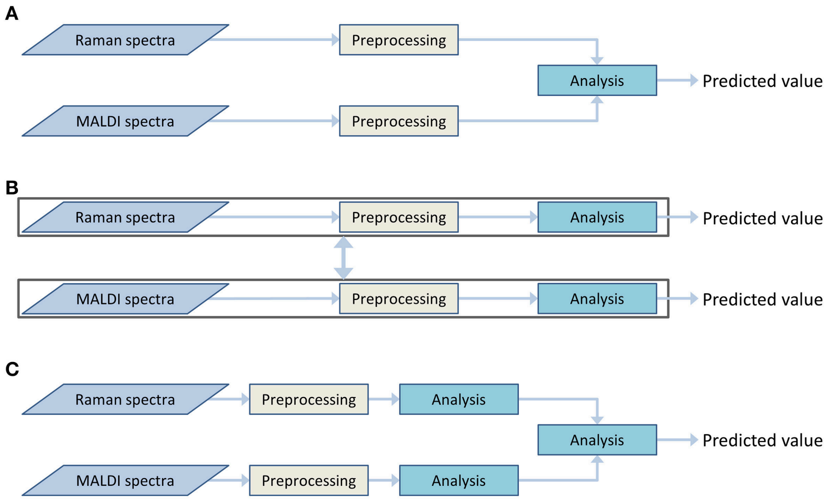


FIGURE 3 | Various data fusion architectures: centralized **(A)**, decentralized **(B)**, and distributed **(C)** architectures.

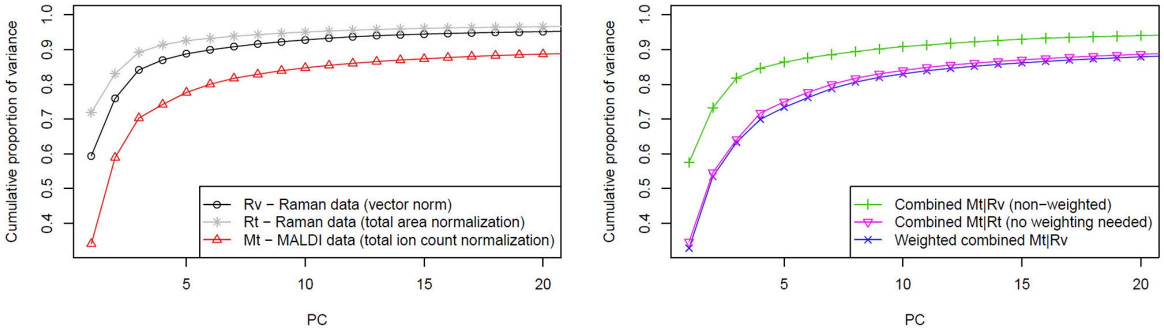


FIGURE 4 | Influence of weighting in the data fusion center on the PCA. The left side of the figure shows cumulative proportion of data variance explained by first 20 PCs for Raman data (normalized in two different ways) and for MALDI data. On the right side of the figure, a slower raise of cumulative proportion of weighted data variance in comparison to the non-weighted case is shown. This trend reflects that more independent features can be extracted from the data by applying weighting prior to the data fusion. As it is also shown in the plot on the right side, a similar effect can be reached by applying the same type of normalization for both data types.

there are no negative values present in the mass spectra. The right side of **Figure 4** clearly shows that there is a marked difference between the approach not taking the data scaling into account and the approaches based on weighting or identical normalization. However, no significant benefit was observed when comparing the weighting to identical normalization approach.

To further investigate the influence of weighting on data fusion, the weighting coefficient was varied in a range from 1 to 20 and a PCA utilized for every case. The extracted curves of the cumulative proportion of the variance were organized as a surface plot (**Figure 5**). To make the interpretation easier, the curves, which correspond to the data combination without weighting and with weighting based on the ratio of l_1-norms, are additionally highlighted in **Figure 5**. Although no single weighting coefficient is globally the best, the proposed weighting coefficient lies close to the area where the data variance is spread between multiple PCs. Thus, fusing data in this manner enables the PCA to extract a larger number of reliable features.

Although an optimal data fusion has been achieved as abovementioned, a direct comparison of cumulative proportions of variance explained by the PCA for data with different dimensionalities may be misleading. Hence, the results obtained from the combined approach and separated data analysis (**Figure 6**) were checked by means of inspecting the PCA loadings and scores. The first three PCs were visualized separately for the MALDI spectrometric imaging data (**Figures 6A,C**), Raman spectroscopic imaging data (**Figures 6B,D**), and their combination (**Figures 6E–G**).

The comparison of the PCA scores in **Figure 6** shows that the image of the MALDI-Raman combination (**Figure 6G**) depicts clearer spatial features of the sample (compared to **Figures 6C,D**). The corresponding false-color score composite (**Figure 6G**) is less noisy, and looks subjectively better than the images obtained separately from the MALDI mass spectrometric (**Figure 6C**) and Raman spectroscopic data (**Figure 6D**). Moreover, the loading vector of the third PC of the MALDI spectra (shown in blue color in **Figure 6A**) has positive and negative values related to isotopes of the same molecules. It means that it represents mostly noise and variations in the signal to noise ratio. On the other hand, the MALDI part of the loadings of the third PC in the combined analysis (shown in blue color in **Figure 6E**) reflects a joint behavior for the isotopes of the same ions. Moreover, the Raman part of this PC contains the peaks associated with lipids (Notingher and Hench, 2006), namely the C = C stretching region (1,655–1,680 cm^{-1}), and CH deformation band (1,420–1,480 cm^{-1}). Although these two peaks may also be associated with Amide I and CH deformations of proteins, there is a decrease in the protein-associated range (Notingher and Hench, 2006) in the wavenumber region 1,128–1,284 cm^{-1}. Furthermore, there are notable changes in the CH-stretching region (2,800–3,100 cm^{-1}). Thus, the third PC of the combined data represents the actual diversity in the lipid composition of the sample. The relationship of the CH stretching region of the Raman spectra to the changes in the lipid content can also be observed by a high correlation of the Raman spectral region with MALDI mass spectra (**Figure 7**).

Since both data types simultaneously reflect variations in lipid content, the specific changes in the correlation profiles (**Figure 7**) of the Raman and MALDI data are observed in the areas related to lipid bands in Raman spectra. Besides the contributions of lipids, which are found in the third PC, the fingerprint region of Raman spectra contains numerous peaks related to proteins and DNA. These Raman bands correlate with MALDI peaks both positively and negatively (**Figure 7**). The correlation of a certain MALDI peak with the Raman data shows a similar structure, but with an opposite sign. This sign change reflects changes in the contribution of specific lipids with respect to the overall increase of lipid content in the sample.

One of the non-lipid compounds, which feature strong Raman bands, is phenylalanine. Its symmetrical ring breathing mode and C-H in-plane mode are visible in the first two PCs at 1,004 and 1,030 cm^{-1}. Another peak related to phenylalanine can be found in the first two PCs at 1,104 cm^{-1} (Movasaghi et al., 2007). Aside of that, the first PC contains contributions of tryptophan at 760 cm^{-1} (Bonifacio et al., 2010). The protein backbone C-C_α stretching of collagen is present in the second PC at 936 cm^{-1} and the ν(C–C) protein backbone is located in the first two PCs at 816 cm^{-1} (Bonifacio et al., 2010). Also, prominent collagen-associated bands like Amide I and Amide III can be seen in the first PC at 1,655–1,680 and 1,220–1,284 cm^{-1}, respectively (Krafft et al., 2005; Notingher and Hench, 2006). Moreover, the peak at 1,647 cm^{-1} is associated with the random coil structure of proteins in general (Movasaghi et al., 2007). This peak is also present in the first two PCs.

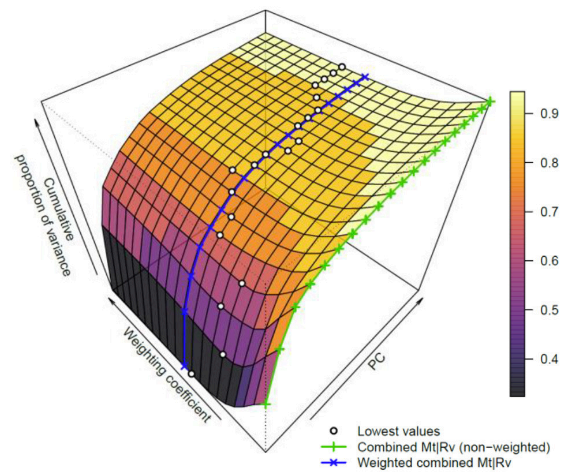


FIGURE 5 | Dependence of the variance explained by PCA using the weighting scheme. The surface plot covers the first 20 PCs and weighting coefficients between 1 and 20. The cumulative proportions of variance for the weighted and non-weighted cases are shown as blue and green lines, respectively (please refer to the online version for colors). Furthermore, the lowest variance is highlighted for each number of PCs with a dot. These dots represent an optimal unmixing for the related number of PCs. Although this optimum changes with respect to PC numbers, the used weighting coefficient based on l_1-norms clearly lies near the minimum of cumulative proportion of variance for a given number of PCs.

The main contribution to the first PC is the ratio between the fingerprint region of Raman spectra and C-H stretching region. On the other side, the fingerprint region of the second PC contains both positive and negative peaks, reflecting the changes in protein content. Along with the protein content, valuable information about DNA is obtained from the first two PCs of the Raman spectra. The peak at 1,180 cm^{-1} represents cytosine and guanine. Another DNA peak is located at 1,263 cm^{-1} and represents adenine and thymine (Movasaghi et al., 2007). All Raman spectral features provide a complex overview of the chemical composition of the mouse brain section. The MALDI data, on the other hand, extends the overview of the distribution of biomolecules based on Raman spectroscopy with detailed information about the lipid content composition.

CONCLUSION

In this paper, a data fusion scheme was investigated to analyze Raman spectroscopic and MALDI mass spectrometric imaging data together. We described the most significant corrupting effects influencing the analysis of Raman spectroscopic and MALDI mass spectrometric imaging data. The preprocessing workflows were shown for the suppression of these corrupting effects by means of calibration, noise reduction, background correction, and normalization for both data types. After the pretreatment steps, the importance of data weighting prior to data fusion is highlighted, especially when the data are

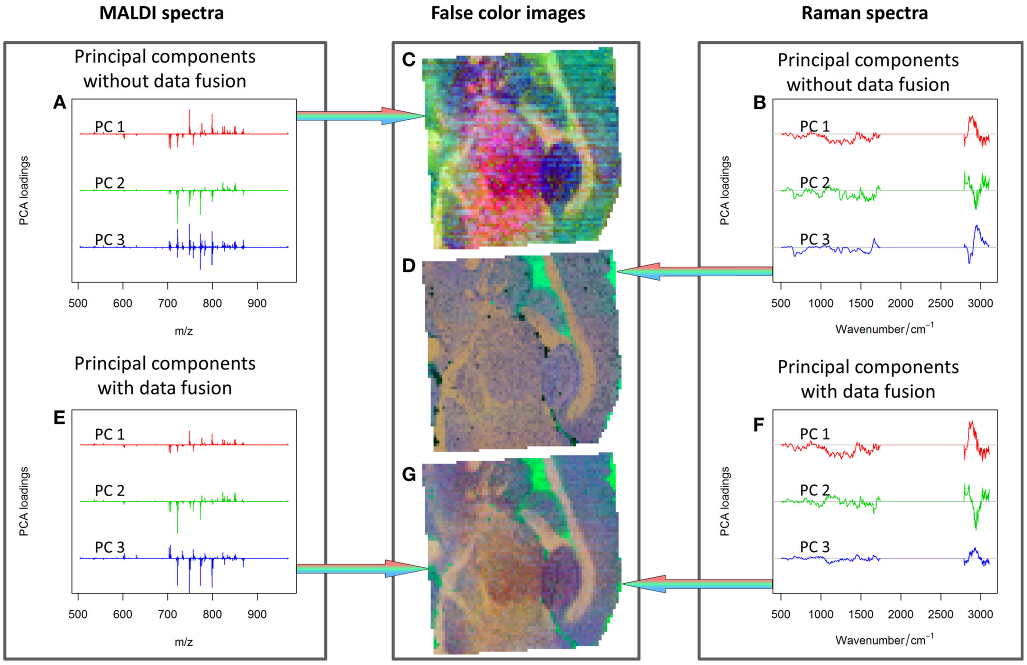


FIGURE 6 | PCA analysis: first three PCs calculated for MALDI spectra **(A)**, Raman spectra **(B)**, combined Raman-MALDI data **(E,F)** and their false-color score composites **(C,D,G)**. Red, green, and blue colors indicate the first, second and third PCs, respectively. Separate plots for the loadings and false color images can be found as Supplementary Material. The PCs composite image of the combined data **(G)** shows a smoother appearance, and the loadings after data fusion **(E,F)** are easier to interpret. See text for further details.

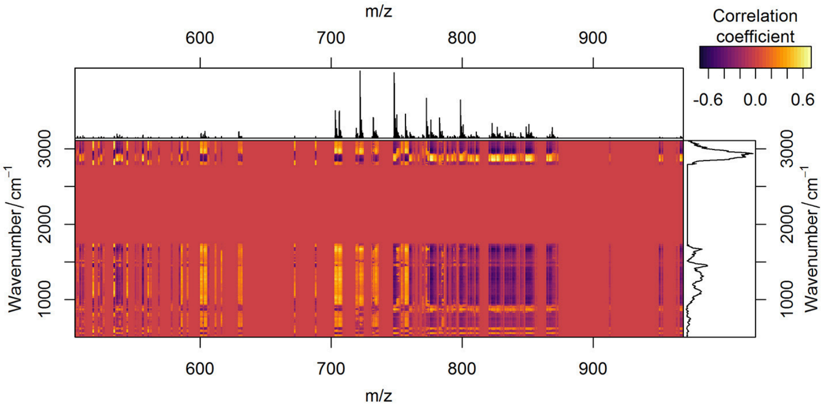


FIGURE 7 | Correlation between Raman spectroscopic and MALDI mass spectrometric data. Correlation of two data types after being preprocessed is depicted in yellow (positive values), red (zero), and violet (negative values) colors. Average preprocessed MALDI spectrum (on the top of the figure) and Raman spectrum (on the right side of the figure) are plotted for easier interpretation.

obtained from different sources and have different scales and dimensionalities. As there is no universal way of balancing the influence of data types on the analysis, optimization, and validation of weighting approaches should be done according to the specific data. In order to allow a judgment of the quality of a weighting, we proposed an approach that allows estimating the goodness of data weighting. This approach is based on analyzing proportions of data variance explained by PCs and we applied this approach by examining the cumulative variance. It was shown that the weighting, based on the ratio of l_1-norms of the data matrices, allows optimal unmixing of the example data set into features. Besides the comparison of different weighting schemes, the proposed method can be used for the comparison of normalization approaches. It was found that vector normalization allows better unmixing of the example Raman data as compared to the normalization to the integrated spectral intensity (l_1-norm). Besides the establishment of a weighting approach, we discovered that a nearly optimal result compared to the weighting is achieved if the spectra of both types are normalized to the same norm. We could demonstrate this by normalizing both types of spectra of an example dataset to the same norm. This was the l_1-norm in our example. However, it is important to keep in mind that this method of comparing the cumulative proportions of variance should be used only when a researcher is interested in maximizing the number of extracted independent features.

The revealing of additional meaningful features by means of optimal data fusion was demonstrated for the combination of Raman spectroscopic and MALDI mass spectrometric imaging data. We showed this by comparing the third PC extracted from each type of data separately and from the combined data. The MALDI-related part of the third combined component showed a clearer interpretation in comparison to the third loading obtained from the MALDI data alone. Moreover, the Raman-related part of the combined component reflected variations in lipid to protein ratio. This PC depicts a decrease in a protein-associated range that occurs along with an increase of bands related to the CH deformation and C=C stretching in lipids, which can be found in the regions 1,128–1,284, 1,420–1,480, and 1,655–1,680 cm^{-1}, respectively. Therefore, changes in the lipid to protein ratio and changes in lipid content itself can be observed simultaneously through the data fusion of Raman spectroscopic and MALDI mass spectrometric imaging data.

Finally, the advantage of the combined analysis was illustrated by a comparison of the PCA results visualized as false-color RGB images. These images were obtained separately for the preprocessed Raman and MALDI imaging data and for the combined data. Visual investigation of the images showed that the combined approach provides a sharper image with less noise contributions. This allows the conclusion that the data fusion increases reliability not only for the spectral but also for the spatial features present in the data.

ETHICS STATEMENT

This research is based on already published data provided to the authors by Bocklitz et al. (2013). For this reason, an ethics approval was not required as per institutional and national guidelines.

AUTHOR CONTRIBUTIONS

TB and JP initiated the study, supervised the study and discussed the results. OR performed the analysis including the development of the R scripts. TB performed the pre-study including the co-registration step. OR, JP, and TB wrote the manuscript.

ACKNOWLEDGMENTS

Financial support of the EU via the project HemoSpec (FP 7, CN 611682), co-funding of the EU for the project PhotoSkin (FKZ 13N13243) and support of the BMBF via the project PhotoSkin (FKZ 13N13243) and Uro-MDD (FKZ 03ZZ0444J) are highly acknowledged. The publication of this article was funded by the Open Access Fund of the Leibniz Association. Authors wish to thank Dr. Anna Crecelius for acquiring the data and to Prof. Dr. Ferdinand von Eggeling for helpful discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2018.00257/full#supplementary-material

Supplementary Image 1 | The plots from Figure 6 provided in vector format.

REFERENCES

Afseth, N. K., Segtnan, V. H., and Wold, J. P. (2006). Raman spectra of biological samples: a study of preprocessing methods. *Appl. Spectrosc.* 60, 1358–1367. doi: 10.1366/000370206779321454

Ahlf, D. R., Masyuko, R. N., Hummon, A. B., and Bohn, P. W. (2014). Correlated mass spectrometry imaging and confocal Raman microscopy for studies of three-dimensional cell culture sections. *Analyst* 139, 4578–4585. doi: 10.1039/C4AN00826J

Baddeley, A., and Turner, R. (2005). spatstat: an R package for analyzing spatial point patterns. *J. Stat. Softw.* 12:42. doi: 10.18637/jss.v012.i06

Bocklitz, T., Bräutigam, K., Urbanek, A., Hoffmann, F., Von Eggeling, F., Ernst, G., et al. (2015). Novel workflow for combining Raman spectroscopy and MALDI-MSI for tissue based studies. *Anal. Bioanal. Chem.* 407, 7865–7873. doi: 10.1007/s00216-015-8987-5

Bocklitz, T., Walter, A., Hartmann, K., Rösch, P., and Popp, J. (2011). How to pre-process Raman spectra for reliable and stable models? *Anal. Chim. Acta* 704, 47–56. doi: 10.1016/j.aca.2011.06.043

Bocklitz, T. W., Crecelius, A. C., Matthäus, C., Tarcea, N., Von Eggeling, F., Schmitt, M., et al. (2013). Deeper understanding of biological tissue: quantitative correlation of MALDI-TOF and Raman imaging. *Anal. Chem.* 85, 10829–10834. doi: 10.1021/ac402175c

Bonifacio, A., Beleites, C., Vittur, F., Marsich, E., Semeraro, S., Paoletti, S., et al. (2010). Chemical imaging of articular cartilage sections with Raman mapping, employing uni- and multi-variate methods for data analysis. *Analyst* 135, 3193–3204. doi: 10.1039/c0an00459f

Butler, H. J., Ashton, L., Bird, B., Cinque, G., Curtis, K., Dorney, J., et al. (2016). Using Raman spectroscopy to characterize biological materials. *Nat. Protocols* 11, 664–687. doi: 10.1038/nprot.2016.036

Cappel, U. B., Bell, I. M., and Pickard, L. K. (2010). Removing cosmic ray features from Raman map data by a refined nearest neighbor comparison method as a precursor for chemometric analysis. *Appl. Spectro.* 64, 195–200. doi: 10.1366/000370210790619528

Castanedo, F. (2013). A review of data fusion techniques. *Sci. World J.* 2013:19. doi: 10.1155/2013/704504

Dörfer, T., Bocklitz, T., Tarcea, N., Schmitt, M., and Popp, J. (2011). Checking and improving calibration of Raman spectra using chemometric approaches. *Zeitschrift Fur Phys. Chem.* 225, 753–764. doi: 10.1524/zpch.2011.0077

Ehrentreich, F., and Sümmchen, L. (2001). Spike removal and denoising of Raman spectra by wavelet transform methods. *Anal. Chem.* 73, 4364–4373. doi: 10.1021/ac0013756

Ember, K. J. I., Hoeve, M. A., McAughtrie, S. L., Bergholt, M. S., Dwyer, B. J., Stevens, M. M., et al. (2017). Raman spectroscopy and regenerative medicine: a review. *Regenerat. Med.* 2:12. doi: 10.1038/s41536-017-0014-3

Fagerer, S. R., Schmid, T., Ibáñez, A. J., Pabst, M., Steinhoff, R., Jefimovs, K., et al. (2013). Analysis of single algal cells by combining mass spectrometry with Raman and fluorescence mapping. *Analyst* 138, 6732–6736. doi: 10.1039/c3an01135f

Fitzgerald, M. C., Parr, G. R., and Smith, L. M. (1993). Basic matrixes for the matrix-assisted laser desorption/ionization mass spectrometry of proteins and oligonucleotides. *Anal. Chem.* 65, 3204–3211. doi: 10.1021/ac00070a007

Gessel, M. M., Norris, J. L., and Caprioli, R. M. (2014). MALDI imaging mass spectrometry: Spatial molecular analysis to enable a new age of discovery. *J. Prot.* 107, 71–82. doi: 10.1016/j.jprot.2014.03.021

Gu, M., Wang, Y., Zhao, X. G., and Gu, Z. M. (2006). Accurate mass filtering of ion chromatograms for metabolite identification using a unit mass resolution liquid chromatography/mass spectrometry system. *Rapid Commun. Mass Spectrosc.* 20, 764–770. doi: 10.1002/rcm.2377

Hinsch, A., Buchholz, M., Odinga, S., Borkowski, C., Koop, C., Izbicki, J. R., et al. (2017). MALDI imaging mass spectrometry reveals multiple clinically relevant masses in colorectal cancer using large-scale tissue microarrays. *J. Mass Spectro.* 52, 165–173. doi: 10.1002/jms.3916

Horn, R. A., and Johnson, C. R. (1990). *Matrix Analysis*. Cambridge University Press.

Köhler, M., Machill, S., Salzer, R., and Krafft, C. (2009). Characterization of lipid extracts from brain tissue and tumors using Raman spectroscopy and mass spectrometry. *Anal. Bioanal. Chem.* 393, 1513–1520. doi: 10.1007/s00216-008-2592-9

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. doi: 10.1016/j.addr.2015.03.009

Krafft, C., Knetschke, T., Funk, R. H. W., and Salzer, R. (2005). Identification of organelles and vesicles in single cells by Raman microspectroscopic mapping. *Vibrat. Spectrosc.* 38, 85–93. doi: 10.1016/j.vibspec.2005.02.008

Krutchinsky, A. N., and Chait, B. T. (2002). On the nature of the chemical noise in MALDI mass spectra. *J. Am. Soc. Mass Spectrosc.* 13, 129–134. doi: 10.1016/S1044-0305(01)00336-1

Kumar, R., Sripriya, R., Balaji, S., Senthil Kumar, M., and Sehgal, P. K. (2011). Physical characterization of succinylated type I collagen by Raman spectra and MALDI-TOF/MS and *in vitro* evaluation for biomedical applications. *J. Mol. Struct.* 994, 117–124. doi: 10.1016/j.molstruc.2011.03.005

Lanni, E. J., Masyuko, R. N., Driscoll, C. M., Dunham, S. J. B., Shrout, J. D., Bohn, P. W., et al. (2014). Correlated imaging with C60-SIMS and confocal raman microscopy: visualization of cell-scale molecular distributions in bacterial biofilms. *Anal. Chem.* 86, 10885–10891. doi: 10.1021/ac5030914

Lasch, P., and Noda, I. (2017). Two-dimensional correlation spectroscopy for multimodal analysis of FT-IR, Raman, and MALDI-TOF MS hyperspectral images with Hamster brain tissue. *Anal. Chem.* 89, 5008–5016. doi: 10.1021/acs.analchem.7b00332

Masyuko, R., Lanni, E. J., Sweedler, J. V., and Bohn, P. W. (2013). Correlated imaging - a grand challenge in chemical analysis. *Analyst* 138, 1924–1939. doi: 10.1039/c3an36416j

Masyuko, R. N., Lanni, E. J., Driscoll, C. M., Shrout, J. D., Sweedler, J. V., and Bohn, P. W. (2014). Spatial organization of *Pseudomonas aeruginosa* biofilms probed by combined matrix-assisted laser desorption ionization mass spectrometry and confocal Raman microscopy. *Analyst* 139, 5700–5708. doi: 10.1039/C4AN00435C

Matousek, P., and Stone, N. (2013). Recent advances in the development of Raman spectroscopy for deep non-invasive medical diagnosis. *J. Biophot.* 6, 7–19. doi: 10.1002/jbio.201200141

Movasaghi, Z., Rehman, S., and Rehman, I. U. (2007). Raman spectroscopy of biological tissues. *Appl. Spectro. Rev.* 42, 493–541. doi: 10.1080/05704920701551530

Muhamadali, H., Weaver, D., Subaihi, A., Almasoud, N., Trivedi, D. K., Ellis, D. I., et al. (2016). Chicken, beams, and Campylobacter: rapid differentiation of foodborne bacteria via vibrational spectroscopy and MALDI-mass spectrometry. *Analyst* 141, 111–122. doi: 10.1039/C5AN01945A

Notingher, I., and Hench, L. L. (2006). Raman microspectroscopy: a noninvasive tool for studies of individual living cells *in vitro*. *Exp. Rev. Med. Dev.* 3, 215–234. doi: 10.1586/17434440.3.2.215

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ramoji, A., Neugebauer, U., Bocklitz, T., Foerster, M., Kiehntopf, M., Bauer, M., et al. (2012). Toward a spectroscopic hemogram: Raman spectroscopic differentiation of the two most abundant leukocytes from peripheral blood. *Anal. Chem.* 84, 5335–5342. doi: 10.1021/ac3007363

Ryabchykov, O., Bocklitz, T., Ramoji, A., Neugebauer, U., Foerster, M., Kroegel, C., et al. (2016). Automatization of spike correction in Raman spectra of biological samples. *Chemometr. Intell. Lab. Syst.* 155, 1–6. doi: 10.1016/j.chemolab.2016.03.024

Ryan, C. G., Clayton, E., Griffin, W. L., Sie, S. H., and Cousens, D. R. (1988). SNIP, a statistics-sensitive background treatment for the quantitative analysis of PIXE spectra in geoscience applications. *Nuclear Instr. Methods Phys. Res. B* 34, 396–402. doi: 10.1016/0168-583X(88)90063-8

Schulze, H. G., and Turner, R. F. (2014). A two-dimensionally coincident second difference cosmic ray spike removal method for the fully automated processing of Raman spectra. *Appl. Spectro.* 68, 185–191. doi: 10.1366/13-07216

Singhal, N., Kumar, M., and Virdi, J. S. (2016). MALDI-TOF MS in clinical parasitology: applications, constraints and prospects. *Parasitology* 143, 1491–1500. doi: 10.1017/S0031182016001189

Tolstik, T., Marquardt, C., Matthäus, C., Bergner, N., Bielecki, C., Krafft, C., et al. (2014). Discrimination and classification of liver cancer cells and proliferation states by Raman spectroscopic imaging. *Analyst* 139, 6036–6043. doi: 10.1039/C4AN00211C

Urwyler, S. K., and Glaubitz, J. (2016). Advantage of MALDI-TOF-MS over biochemical-based phenotyping for microbial identification illustrated on industrial applications. *Lett. Appl. Microbiol.* 62, 130–137. doi: 10.1111/lam.12526

Van De Plas, R., Yang, J., Spraggins, J., and Caprioli, R. M. (2015). Image fusion of mass spectrometry and microscopy: a multimodality paradigm for molecular tissue mapping. *Nat. Methods* 12:366. doi: 10.1038/nmeth.3296

Verwer, P. E., Van Leeuwen, W. B., Girard, V., Monnin, V., Van Belkum, A., Staab, J. F., et al. (2014). Discrimination of Aspergillus lentulus from *Aspergillus fumigatus* by Raman spectroscopy and MALDI-TOF MS. *Eur. J. Clin. Microbiol. Infect. Dis.* 33, 245–251. doi: 10.1007/s10096-013-1951-4

Wagner, M. (2009). Single-cell ecophysiology of microbes as revealed by Raman microspectroscopy or secondary ion mass spectrometry imaging. *Ann. Rev. Microbiol.* 63, 411–429. doi: 10.1146/annurev.micro.091208.073233

Zhang, L., and Henson, M. J. (2007). A practical algorithm to remove cosmic spikes in Raman imaging data for pharmaceutical applications. *Appl. Spectro.* 61, 1015–1020. doi: 10.1366/000370207781745847

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Ryabchykov, Popp and Bocklitz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

[P7] UV-Raman spectroscopic identification of fungal spores important for respiratory diseases

Reprinted with permission from [Žukovskaja, O., Kloß, S., Blango, M.G., Ryabchykov, O., Kniemeyer, O., Brakhage, A.A., Bocklitz, T.W., Cialla-May, D., Weber, K. and Popp, J., 2018. UV-Raman Spectroscopic Identification of Fungal Spores Important for Respiratory Diseases. *Analytical chemistry*, *90*(15), pp.8912-8918]. Copyright 2018 American Chemical Society. Copyright 2018 Wiley.

Erklärungen zu den Eigenanteilen des Promovenden sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an der Publikation.

[1]Žukovskaja, O., [2]Kloss, S., [3]Blango, M. G., [4]Ryabchykov, O., [5]Kniemeyer, O., [6]Brakhage, A. A., [7]Bocklitz, T. W., [8]Cialla-May, D., [9]Weber, K., [10]Popp, J., 2018. UV-Raman Spectroscopic Identification of Fungal Spores Important for Respiratory Diseases. *Analytical chemistry*, *90*(15), pp.8912-8918										
Beteiligt an (Zutreffendes ankreuzen)										
	1	2	3	4	5	6	7	8	9	10
Konzeption des Forschungsansatzes	X	X	X		X	X	X	X	X	X
Planung der Untersuchungen	X	X	X				X	X	X	X
Datenerhebung	X	X								
Datenanalyse und Interpretation	X			**X**			X			
Schreiben des Manuskripts	X		X	**X**			X	X	X	X
Vorschlag Anrechnung Publikationsäquivalent	1.0			**0.5**						

analytical chemistry
Article
Cite This: Anal. Chem. XXXX, XXX, XXX−XXX
pubs.acs.org/ac

Downloaded via THURINGER UNIV LANDESBIBLIOTHEK on July 17, 2018 at 15:28:22 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

UV-Raman Spectroscopic Identification of Fungal Spores Important for Respiratory Diseases

Olga Žukovskaja,[†,‡,§] Sandra Kloß,[†] Matthew G. Blango,[∥] Oleg Ryabchykov,[†,§] Olaf Kniemeyer,[∥] Axel A. Brakhage,[∥,⊥] Thomas W. Bocklitz,[†,§] Dana Cialla-May,[†,‡,§] Karina Weber,[*,†,‡,§] and Jürgen Popp[†,‡,§]

[†]Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich Schiller University Jena, Helmholtzweg 4, 07745 Jena, Germany

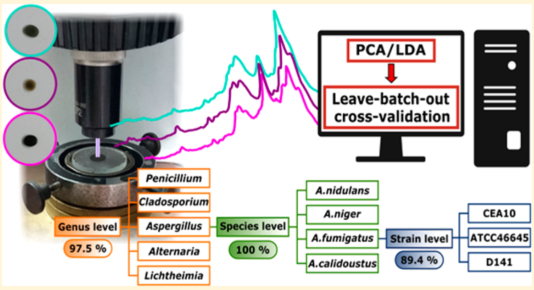
[‡]Research Campus Infectognostic, Philosophenweg 7, 07743 Jena, Germany

[§]Leibniz Institute of Photonic Technology Jena—Member of the Research Alliance "Leibniz Health Technologies", Albert-Einstein-Straße 9, 07745 Jena, Germany

[∥]Department of Molecular and Applied Microbiology, Leibniz Institute for Natural Product Research and Infection Biology (HKI), Adolf-Reichwein-Straße 23, 07745 Jena, Germany

[⊥]Department of Microbiology and Molecular Biology, Institute for Microbiology, Friedrich Schiller University Jena, Neugasse 25, 07743 Jena, Germany

Supporting Information



ABSTRACT: Fungal spores are one of several environmental factors responsible for causing respiratory diseases like asthma, chronic obstructive pulmonary disease (COPD), and aspergillosis. These spores also are able to trigger exacerbations during chronic forms of disease. Different fungal spores may contain different allergens and mycotoxins, therefore the health hazards are varying between the species. Thus, it is highly important quickly to identify the composition of fungal spores in the air. In this study, UV-Raman spectroscopy with an excitation wavelength of 244 nm was applied to investigate eight different fungal species implicated in respiratory diseases worldwide. Here, we demonstrate that darkly colored spores can be directly examined, and UV-Raman spectroscopy provides the information sufficient for classifying fungal spores. Classification models on the genus, species, and strain levels were built using a combination of principal component analysis and linear discriminant analysis followed by evaluation with leave-one-batch-out-cross-validation. At the genus level an accuracy of 97.5% was achieved, whereas on the species level four different *Aspergillus* species were classified with 100% accuracy. Finally, classifying three strains of *Aspergillus fumigatus* an accuracy of 89.4% was reached. These results demonstrate that UV-Raman spectroscopy in combination with innovative chemometrics allows for fast identification of fungal spores and can be a potential alternative to currently used time-consuming cultivation.

Hundreds of millions of people of all ages suffer every day from chronic respiratory diseases. According to the latest WHO estimates (2004), which are considered conservative, >300 million people are afflicted with asthma, 210 million people suffer from chronic obstructive pulmonary disease (COPD), while >400 million suffer from some other form of respiratory disease.[1] These diseases all have a negative impact on the quality of life of those affected and in many instances can be life threatening, especially in cases of acute exacerbation. Exacerbations of allergic respiratory disease occur frequently in COPD and asthma patients, likely triggered by exposure to environmental stimuli. Although debated, these exacerbations are generally characterized by increased airway inflammation, mucus production, and impaired lung function.

Outdoor fungal spores are one of several environmental factors responsible not only for causing respiratory diseases in humans,[2−4] but also for triggering exacerbations during chronic forms of lung diseases such as asthma and COPD. For example, it was reported that asthma hospitalization cases increase during thunderstorms due to the increased aerosolization of fungal spores.[5] Since different fungi may produce different allergens and mycotoxins, the severity of asthma exacerbation may vary between spores of different fungal taxa,[6] e.g., some patients have hypersensitivity only to *Aspergillus spec.*, which can lead to allergic bronchopulmonary aspergillosis

Received: March 7, 2018
Accepted: June 29, 2018
Published: June 29, 2018

ACS Publications © XXXX American Chemical Society

DOI: 10.1021/acs.analchem.8b01038
Anal. Chem. XXXX, XXX, XXX−XXX

(ABPA) or other complications in patients suffering from asthma, cystic fibrosis, or COPD.[7] Therefore, it would be beneficial to identify and eliminate fungal spores before they trigger respiratory disease or exacerbations in patients with chronic respiratory diseases.

Currently the "gold standard" for fungal identification is the cultivation of sampled organisms coupled with careful observation and measurement of macroscopic and microscopic morphological characteristics of the organism of interest, including reproductive structures like spores. However, this method is laborious and requires highly trained and experienced personnel. Additional difficulties in fungal identification include the presence of a wide range of nonculturable organisms and/or contamination of fungal cultures with fast-growing bacteria, as these factors make the analysis of morphological information complicated.[8] A fast, highly automated method, which can identify fungal spores via pattern recognition, would exclude the human factor from the identification process and would improve turn-around times for spore identification. As an alternative approach to the more classic culture methods, polymerase chain reaction (PCR)-based detection methods for fungal spore identification have been reported.[8−11] PCR methods require expensive reagents and design of target-specific primers and tend to produce false-positive results. Another possibility is offered by vibrational spectroscopy, a technique which utilizes molecular vibrations to provide information about the molecular composition, structure and behavior within a sample.

Among the vibrational spectroscopic methods, Raman spectroscopy is a very promising tool for the characterization of microorganisms.[12,13] This method offers specific molecular information about the chemical species in the samples in a noninvasive and label-free manner. The abilities of Raman spectroscopy for investigation and characterization of airborne allergens, such as individual pollen grains, has been demonstrated by various groups.[14−16] The characterization of fungal spores using Raman spectroscopy with visible excitation wavelengths has been previously also reported. However, in these studies, the excitation wavelengths utilized could only be applied to white or light-colored spores.[17,18] Several different microfungi spores relevant to indoor contamination were successfully characterized and identified with Raman spectroscopy, describing the biochemical composition of a single spore.[19] A study of C. Wang et al. reported measurements of Raman spectra in 1600−3400 cm^{-1} spectral range from individual pollen particles and Bermuda grass smut spores held in a photophoretic trap. Due to the small size of spores the spectrum had high background and only three Raman bands were visible.[20] Raman spectroscopy with visible excitation wavelengths was also used to discriminate *Aspergillus lentulus* from *Aspergillus fumigatus* with an accuracy of only 78%.[17] K. De Gussem et al. combined Raman spectroscopy with linear discriminant analysis (LDA) and reached around 90% accuracy in assigning the spectra of spores to the correct genus,[18] while identification on the species level was not possible. Instead of applying visible Raman excitation wavelengths various studies on mammalian cells,[21,22] bacteria,[23−25] and pollen[26] showed that utilizing electronically resonant excitation wavelengths in the UV region can be beneficial for the identification of microorganisms due to a selective resonant enhancement of the Raman signals of taxonomically important macromolecules, e.g., DNA/RNA bases and aromatic amino acids. Furthermore, the fluorescence background in UV resonance Raman spectra is negligible for excitation wavelengths below 260 nm, which lead to Raman spectra with high signal-to-noise ratio.[27−31]

In the present study, for the first time the UV-Raman spectroscopic identification of fungal spores is demonstrated. Eight different filamentous fungal species, implicated to a varied extent in respiratory diseases worldwide, were examined: *Aspergillus fumigatus*, *Aspergillus niger*, *Aspergillus nidulans*, *Aspergillus calidoustus*, *Cladosporium herbarum*, *Alternaria alternata*, *Penicillium rubens*, and *Lichtheimia corymbifera*. Except for the mucoralean fungus *Lichtheimia corymbifera*, a basal fungal lineage, all other species belong to the division Ascomycota. In contrast to previous studies,[17,18] highly pigmented spores were successfully investigated. The identification of spores was based on a combination of principal component analysis (PCA) and LDA of the UV resonance Raman spectral data and was applied at a genus, species, and strain level. The performance of the models was evaluated with leave-one batch-out cross-validation. The presented results highlight the possibility of UV-Raman spectroscopy as a promising method for the automated identification of fungal spores.

■ MATERIALS AND METHODS

Fungal Spores. For this study 11 fungal strains from 8 different species were utilized, namely *A. fumigatus*, *P. rubens*, *A. niger*, *C. herbarum*, *A. alternata*, *A. nidulans*, *A. calidoustus*, and *L. corymbifera* (Table S-1 of the Supporting Information, SI). The *A. fumigatus pksP* mutant has been described previously.[32,33] The fungi were grown for 7 days, in the dark, on 10 cm agar plates at room temperature (~22 °C). Growth was conducted on *Aspergillus* minimal media (AMM; 6.0 g/L $NaNO_3$, 0.52 g/L KCl, 1.52 g/L KH_2PO_4, 0.52 g/L $MgSO_4 \cdot 7H_2O$, 1% (wt/vol) glucose, and 1 mL of trace element solution [1 g/L $FeSO_4 \cdot 7H_2O$, 8.8 g/L $ZnSO_4 \cdot 7H_2O$, 0.4 g/L $CuSO_4 \cdot 5H_2O$, 0.15 g/L $MnSO_4 \cdot 4H_2O$, 0.1 g/L $Na_2B_4O_7 \cdot 10H_2O$, 0.05 g/L $(NH_4)_6Mo_7O_{24} \cdot 4H_2O$] per liter),[34] malt agar (MA) (Sigma-Aldrich), or modified SUP agar plates (4.1 g/L KH_2PO_4, 1.1 g/L NH_4Cl, 0.9 g/L K_2HPO_4, 0.1 g/L $MgSO_4$, 1% (wt/vol) glucose, and 0.5% (wt/vol) yeast extract).[35] AMM was supplemented with 0.06 mg/L biotin and 5 mM arginine for growth of *A. nidulans* A89. Conidia (asexually produced spores) from each strain were collected in 10 mL of water and separated from hyphae through a 40 μm pore filter. After centrifugation for 10 min at 1800 × *g*, conidia were resuspended in 3% (v/v) formaldehyde and incubated for 1 h for inactivation. Following formaldehyde-inactivation, each conidial-mixture was washed with water to remove any contaminating formaldehyde. The inactivation of the spores was confirmed by the lack of growth on appropriate media. Each growth condition was tested in triplicate, on three separate days. For the UV-Raman measurements 10 μL of spore suspensions was spread on a fused-silica surface and allowed to dry at room temperature.

Spectroscopic Instrumentation. Raman measurements were performed using a LabRam HR 800 spectrometer from Horiba Jobin-Yvon (Bensheim, Germany) with a 2400 lines/mm grating. As excitation wavelength, a frequency doubled argon-ion laser (Coherent Innova 300, MotoFReD, Coherent, Dieburg, Germany) operating at 244 nm was applied. The samples were illuminated via a microscope (Olympus BX 41) equipped with a 20× magnification antireflection coated UV objective (LMU UVB) with a numerical aperture of 0.4. The laser power at the laser head was ca. 10.5 mW and on the

DOI: 10.1021/acs.analchem.8b01038

sample ca. 0.8 mW at a spot diameter of ~1 μm, corresponding to ~10^5 W/cm^2 irradiance. The entrance slit was set to 300 μm. The Raman scattered light was detected by a nitrogen-cooled CCD camera with an integration time of 20 s with 2 accumulations. The samples were rotated with a speed 60 rotations/min and moved in the *x,y* direction after each rotation to obtain an average spectrum over a large sample area to minimize possible photodegradation by UV radiation. Ten spectra were collected per sample. Measurements were performed by two independent operators.

Data Analysis. Data processing was performed using an in-house developed script in the programming language R.[36] First, the spectra were wavenumber calibrated by using the Teflon Raman spectrum measured prior to each sample measurement as reference. Next, the Raman spectra were background corrected using the sensitive nonlinear iterative peak (SNIP) clipping algorithm[37] with a second-order clipping filter. Finally, all spectra were vector normalized and used as input for PCA, which was performed to reduce the dimensionality of the data while retaining the most significant information for classification. PCA was followed by LDA. The performance of the created LDA model for classification of fungal spores was estimated using the leave-one batch-out-cross-validation (LBOCV) approach.[38] In this method, one batch was held out from the data set, and the LDA model was redeveloped using the remaining spectra. The resultant model was then used to classify the removed batch. This process was repeated with every batch until all spectra were classified. In this manner LBOCV of the PCA-LDA model was utilized for all number of principle components (PCs). Then, an optimal number of PCs was chosen by finding a saturation point of the accuracy as a function of the number of PCs. For spectral comparison across the groups, mean Raman spectra for each fungal species were calculated using preprocessed, vector-normalized spectra of all batches.

■ RESULTS AND DISCUSSION

The primary goal of this study was to determine the feasibility of UV-Raman spectroscopy for the identification of melanised fungal spores. The selected fungal spores were all highly pigmented: black for *A. alternata* and *A. niger*, dark green for *A. fumigatus* strains and *P. rubens*, brown for *A. calidoustus* and *C. herbarum* and light gray for *L. corymbifera* and *A. nidulans* (Figure S-1). The difficulties of recording Raman spectra using visible excitation wavelengths from such dark conidia were previously reported; the dark pigments result in a strong interfering fluorescent signal, which masks the original spectrum.[17] Therefore, initially, the possibility of obtaining good quality spectra with UV-Raman from deeply colored, highly absorbing samples was investigated. For this purpose, wild-type *A. fumigatus* conidia were compared to non-pigmented conidia of a strain with a mutation in the *pksP* gene, essential for the formation of the gray-green spores containing dihydroxynaphthalene (DHN)-melanin.[32] The UV-Raman spectra of these samples are depicted in Figure 1.

During the measurements, the samples were continuously rotated; however, this rotation was insufficient to completely eliminate the laser-induced degradation of spore proteins into polymeric hydrogenated amorphous carbon. Thus, a carbon background manifesting itself by the two broad bands at approximately 1360 and 1610 cm^{-1} is observed.[39−41] For the nonpigmented *pksP* mutant strain, the graphitization level was lower, and therefore the bands originating from the spores in the region of 1500−1700 cm^{-1} are more prominent. However, the general shape of the spectra and the peak positions are the same for both samples. In the pigmented wild-type spores of *A. fumigatus*, the unique spore-related spectral features necessary for the spore identification process are still easily visible. Hence, it can be concluded that UV-Raman can be used for the measurement of highly pigmented spores.

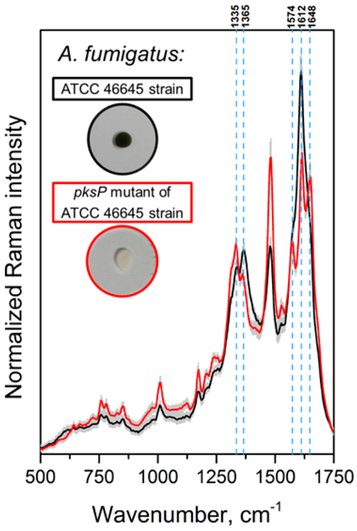


Figure 1. Mean UV-Raman spectra and their double standard deviation of the pigmented *A. fumigatus* ATCC 46645 strain and the nonpigmented *pksP* mutant strain.

To compare a diverse range of allergenic fungi, we had to cultivate these organisms on several different media to promote sporulation due to different nutrient requirements. To test the variability induced by different culture media, we cultivated *P. rubens* strain ATCC 28089 and *A. fumigatus* strain ATCC 46645 on both MA and AMM medium for direct comparison. The mean UV-Raman spectra of the collected spores are presented in Figure S-2. The influence of the culture medium on the spectra was visible and can be best seen in the 1250−1700 cm^{-1} wavenumber region. In the case of *A. fumigatus* grown on MA medium, the signal at 1365 cm^{-1} is more intense, whereas for *P. rubens* the peak at 1648 cm^{-1} is more prominent. Despite this observation, all strains grown on different agars were included into the classification model, proving that the differences from cultivation conditions are smaller than the differences due to species affiliation.

The representative preprocessed mean Raman spectra of the studied fungal spores are depicted in Figure 2.

For all species, primary bands were observed in the wavenumber region between 750 and 1700 cm^{-1}, which is typically associated with various nucleic acids and protein subunits in UV-Raman spectroscopy. The band positions are in good agreement with previously published UV-Raman spectroscopic studies of microorganisms.[24,23,42] The signal at 1648 cm^{-1} can be assigned to thymine and the one at 1612 cm^{-1} to the aromatic amino acids tyrosine and tryptophan. Guanine and adenine exhibit peaks at 1574 and 1479 cm^{-1}; they also contribute to the signal at 1335 cm^{-1}. The latter band (at 1333 cm^{-1}) also contains information from tyrosine. The signal at 1526 cm^{-1} can be assigned to cytosine. Thymine and

DOI: 10.1021/acs.analchem.8b01038

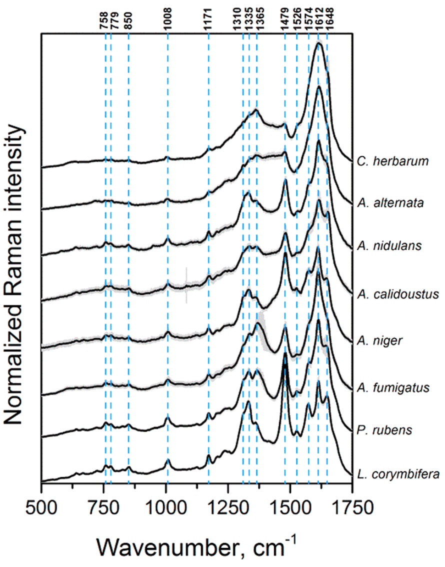


Figure 2. Mean UV-Raman spectra of the all batches from different fungal spores plotted together with double standard deviation (gray area). For the *A. fumigatus* the mean spectrum represents all investigated strains.

adenine exhibit a band at 1365 cm^{-1}. The peak at 1171 cm^{-1} can be assigned to tyrosine. Tryptophan exhibits characteristic bands at 758 and 1008 cm^{-1}. In this study, measurements were performed on the bulk sample under continuous rotation; so several spores contribute to one spectrum. Therefore, one spectrum already comprises metabolic and developmental diversity within the spore population. The reproducibility of these spectra was tested by measuring the Raman spectra from three independent batches.

It is evident from Figure 2 that the fungal spores belonging to different genera display distinct Raman spectral signatures and are easily distinguishable based on their spectra. The peak positions for all analyzed spores are the same, however the relative intensities for the bands in the region of 1250 and 1750 cm^{-1} differ. For the *C. herbarum* and *A. alternata* measurements, three Raman modes at 1574, 1612, and 1648 cm^{-1} are strongly convoluted due to a higher carbonization background. In addition to a visual comparison, a PCA/LDA model was applied and verified using the leave-one batch-out-cross-validation. Three different models on genus, species, and strain level were established (see Figure 3). Health hazards related to fungal spores may differ across genera and species due to the difference between allergenicity and types of mycotoxins produced. It is known for example that the severity of asthma exacerbation may vary between spores of different taxa.[6] To prevent exacerbations, it is important to define the composition of spores in the air on the genus level. Thus, the classification model was first trained to distinguish spores on the genus level. Prior to LDA, the data size was reduced by PCA. The optimal number of PCs was chosen by finding a saturation point of the accuracy as a function of the number of PCs; in this case, 6 PCs were selected (Figure S-3). The LDA model was trained with 2 batches of spores and then tested with one independent batch of the same strain. This allowed for three different batch permutations for validation and gave a reliable unbiased classification model. The sensitivity, specificity, and accuracy of the LDA model in each run were calculated and averaged. Table 1 shows the number of spectra that were classified correctly for each species of the fungi and also summarizes sensitivity and specificity.

Out of 1079 spectra, 27 were misclassified; this resulted in 97.5% accuracy. In Figure 4, the LDA score plot of the classification model is depicted, whereby each dot represents one spectrum and the ellipsoids correspond to confidence regions for the scores on a level of 95% of the spectral data for each class. The biggest amount of errors appeared to be due to the misclassification of *Aspergillus* species to the closely related genus *Penicillium*. The LDA score plot also nicely reflects the phylogenetic relationship between the different genera: *A. alternata* and *C. herbarium* belong to the same class of Dothideomycetes, while the closely related genera *Penicillium* and *Aspergillus* are grouped into the class Eurotiomycetes. The species *L. corymbifera* belongs to a lower group of fungi, the phylum Zygomycota that diverged early in the evolution of true fungi.[43]

In the second step, the model distinguishes between four different *Aspergillus* species. Several *Aspergillus* species are able to cause infections like invasive pulmonary aspergillosis or allergic reactions like ABPA, while others are nearly non-pathogenic.[44] In addition, antifungal drug susceptibility is wide-ranging among even quite similar organisms.[45] Thus, the

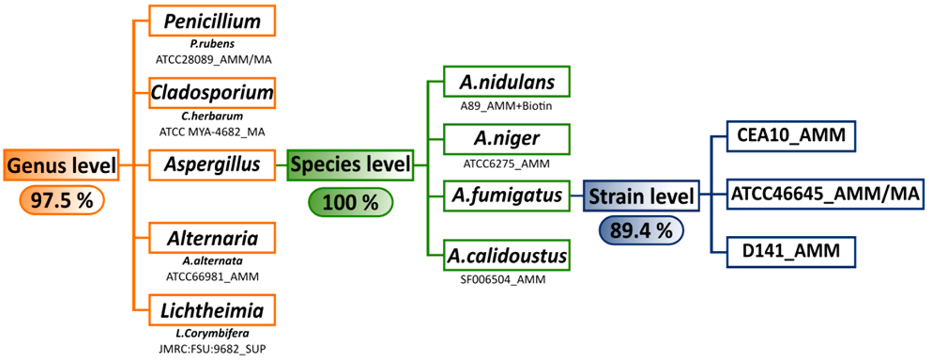


Figure 3. 3-Level PCA/LDA classification model of fungal spores cultured in three independent batches (the percentage represents the accuracies of the model achieved by LBOCV).

Table 1. Identification Results for Different Fungal Spores on the Genus Level

True Predicted	*Altenaria alternata*	*Aspergillus* species	*Cladosporium herbarum*	*Lichtheimia corymbifera*	*Penicillium rubens*	Sensitivity, %	Specificity, %
Altenaria alternata	90	0	0	0	0	**100**	**100**
***Aspergillus* species**	0	603	0	0	1	**95.87**	**99.78**
Cladosporium herbarum	0	4	90	0	0	**100**	**99.6**
Lichtheimia corymbifera	0	0	0	90	0	**100**	**100**
Penicillium rubens	0	22	0	0	179	**99.44**	**97.55**

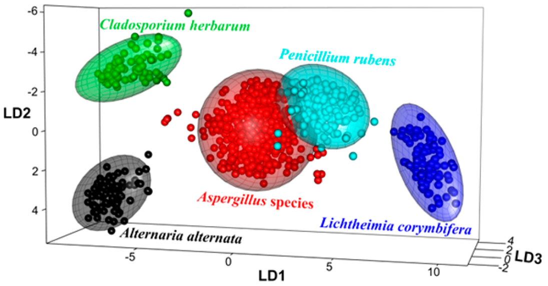


Figure 4. LDA score plots for the classification model on the genus level.

knowledge of the species identity may influence the choice of appropriate antifungal therapy. The visual differentiation of the spectra from *A. fumigatus*, *A. nidulans*, *A. calidoustus*, and *A. niger* is already not as obvious as it was for different genera (Figure 2). For building the identification model on the species level, the number of PCs was set to 7 (Figure S-4). All 629 spectra were assigned correctly resulting in 100% accuracy. The differentiation was based mostly on the different amount of DNA in the spores, mainly resulting in the intensity differences for the peaks at 1365, 1479, and 1648 cm^{-1} (Figure S-5). The achieved accuracy was higher than the one reported by P.E.B. Verwer et al., where the discrimination of *Aspergillus lentulus* from *Aspergillus fumigatus* with VIS-Raman spectroscopy was correct for 78% of spectra.[17]

Finally, the model was trained to classify three different strains of *A. fumigatus*, since diversity in virulence among different *A. fumigatus* strains is also well-documented.[46] The UV-Raman spectra of the three investigated *A. fumigatus* strains are presented in Figure 5. By comparing the spectra, it is clear that the main spectral characteristics of the investigated strains were the same, as a consequence of their close phylogenetic relationship. The LDA model was built with 10 PCs (Figure S-6), and the classification was correct for 89.4% of spectra. The results of this classification model are summarized in Table 2. The misclassification occurred mostly between the ATCC 46445 and D141 strains. The identification of fungal spores on the strain level using Raman spectroscopy was not previously reported.

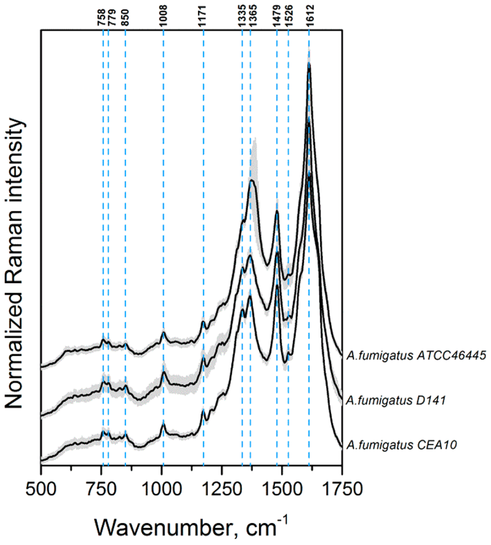


Figure 5. Mean UV-Raman spectra of three *A. fumigatus* strains.

Table 2. Results of the Identification of *A. fumigatus* Strains

True Predicted	ATCC 46645	CA10	D141	Sensitivity, %	Specificity, %
ATCC 46645	151	1	8	**83.89**	**94.97**
CEA10	6	89	0	**98.89**	**97.77**
D141	23	0	81	**91.01**	**91.48**

■ CONCLUSIONS

In the case of respiratory fungal infections, identification of the fungal agent is often incredibly challenging due to a diverse array of potential organisms and relatively poor diagnostic possibilities. Thus, the development of sensitive and automated methods for the identification of fungal spores is of particular medical interest. In this study, a combination of UV resonance

DOI: 10.1021/acs.analchem.8b01038

Raman spectroscopy and chemometrical methods allowed for a highly accurate identification of fungal spores on a genus, species, and even strain level. In all analyzed classification models, strains grown on different agars were included. The promising results presented here indicate that the differentiation of spectra is not determined by the cultivation conditions, but instead based on the differences in the amounts of DNA, RNA, and protein aromatic amino acid units in spores of each species. Moreover, it has been demonstrated that darkly colored spores can be directly examined with UV-Raman spectroscopy, and the contribution of amorphous carbon in the spectra does not inhibit the correct identification of the spores. The obtained classification ratios in the range between 89% and 100% demonstrate that UV resonance Raman spectroscopy represents a powerful tool for fungal spore identification and should be considered for further investigations.

■ ASSOCIATED CONTENT

S Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.8b01038.

> Table S-1: Strains used in the study; Figure S-1: Images of dried spore biomass from the studied fungal species on the fused silica substrate; Figure S-2: Mean UV-Raman spectra and their double standard deviation of the *A. fumigatus* (A) and *P. rubens* (B) strains cultured on different media; Figure S-3: Accuracy as a function of the number of PCs for the genus level; Figure S-4: Accuracy as a function of the number of PCs for the species level; Figure S-5: Loading vectors for the species level model; and Figure S-6: Accuracy as a function of the number of PCs for the strain level (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +49 (0)3641-206309. Fax: +49 (0)3641-206399. E-mail: karina.weber@leibniz-ipht.de (K.W.).

ORCID

Olaf Kniemeyer: 0000-0002-9493-6402
Axel A. Brakhage: 0000-0002-8814-4193
Thomas W. Bocklitz: 0000-0003-2778-6624
Karina Weber: 0000-0003-4907-8645
Jürgen Popp: 0000-0003-4257-593X

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We gratefully acknowledge the Federal Ministry of Education and Research, Germany (BMBF) for the continued support for the InfectoGnostics (13GW0096F) and EXASENS (13N13856 and 13N13861) grants.

■ REFERENCES

(1) WHO. *Global Surveillance, Prevention and Control of Chronic Respiratory Diseases. A Comprehensive Approach*; WHO: Switzerland, 2007.

(2) Kohler, J. R., Casadevall, A., Perfect, J. *Cold Spring Harbor Perspect. Med.* **2015**, *5*, a019273

(3) Sorenson, W. G. *Environ. Health Perspect* **1999**, *107*, 469−472.

(4) Wilson, L. S.; Reyes, C. M.; Stolpman, M.; Speckman, J.; Allen, K.; Beney, J. *Value Health* **2002**, *5*, 26−34.

(5) Dales, R. E.; Cakmak, S.; Judek, S.; Dann, T.; Coates, F.; Brook, J. R.; Burnett, R. T. *Chest* **2003**, *123*, 745−750.

(6) Atkinson, R. W.; Strachan, D. P.; Anderson, H. R.; Hajat, S.; Emberlin, J. *Occup. Environ. Med.* **2006**, *63*, 580−590.

(7) Agarwal, R.; Hazarika, B.; Gupta, D.; Aggarwal, A. N.; Chakrabarti, A.; Jindal, S. K. *Med. Mycol.* **2010**, *48*, 988−994.

(8) Dean, T. R.; Betancourt, D.; Menetrez, M. Y. *J. Microbiol. Methods* **2004**, *56*, 431−434.

(9) Hernandez, A.; Martinez, J. L.; Mellado, R. P. *World J. Microbiol. Biotechnol.* **1999**, *15*, 33−36.

(10) Ward, E. *Methods Mol. Biol.* **2009**, *508*, 147−159.

(11) Williams, R. H.; Ward, E.; McCartney, H. A. *Appl. Environ. Microbiol.* **2001**, *67*, 2453−2459.

(12) Stockel, S.; Kirchhoff, J.; Neugebauer, U.; Rosch, P.; Popp, J. *J. Raman Spectrosc.* **2016**, *47*, 89−109.

(13) Muhamadali, H.; Subaihi, A.; Mohammadtaheri, M.; Xu, Y.; Ellis, D. I.; Ramanathan, R.; Bansal, V.; Goodacre, R. *Analyst* **2016**, *141*, 5127−5136.

(14) Guedes, A.; Ribeiro, H.; Fernandez-Gonzalez, M.; Aira, M. J.; Abreu, I. *Talanta* **2014**, *119*, 473−478.

(15) Ivleva, N. P.; Niessner, R.; Panne, U. *Anal. Bioanal. Chem.* **2005**, *381*, 261−267.

(16) Schulte, F.; Lingott, J.; Panne, U.; Kneipp, J. *Anal. Chem.* **2008**, *80*, 9551−9556.

(17) Verwer, P. E. B.; Leeuwen, W. B.; Girard, V.; Monnin, V.; Belkum, A.; Staab, J. F.; Verbrugh, H. A.; Bakker-Woudenberg, I.; Sande, W. W. J. *Eur. J. Clin. Microbiol. Infect. Dis.* **2014**, *33*, 245−251.

(18) De Gussem, K.; Vandenabeele, P.; Verbeken, A.; Moens, L. *Anal. Bioanal. Chem.* **2007**, *387*, 2823−2832.

(19) Ghosal, S.; Macher, J. M.; Ahmed, K. *Environ. Sci. Technol.* **2012**, *46*, 6088−6095.

(20) Wang, C. J.; Pan, Y. L.; Hill, S. C.; Redding, B. *J. Quant. Spectrosc. Radiat. Transfer* **2015**, *153*, 4−12.

(21) Ashton, L.; Hogwood, C. E. M.; Tait, A. S.; Kuligowski, J.; Smales, C. M.; Bracewell, D. G.; Dickson, A. J.; Goodacre, R. *J. Chem. Technol. Biotechnol.* **2015**, *90*, 237−243.

(22) Yazdi, Y.; Ramanujam, N.; Lotan, R.; Mitchell, M. F.; Hittelman, W.; Richards-Kortum, R. *Appl. Spectrosc.* **1999**, *53*, 82−85.

(23) Gaus, K.; Rosch, P.; Petry, R.; Peschke, K. D.; Ronneberger, O.; Burkhardt, H.; Baumann, K.; Popp, J. *Biopolymers* **2006**, *82*, 286−290.

(24) Lopez-Diez, E. C.; Goodacre, R. *Anal. Chem.* **2004**, *76*, 585−591.

(25) Walter, A.; Schumacher, W.; Bocklitz, T.; Reinicke, M.; Rosch, P.; Kothe, E.; Popp, J. *Appl. Spectrosc.* **2011**, *65*, 1116−1125.

(26) Manoharan, R.; Ghiamati, E.; Britton, K. A.; Nelson, W. H.; Sperry, J. F. *Appl. Spectrosc.* **1991**, *45*, 307−311.

(27) Domes, C.; Domes, R.; Popp, J.; Pletz, M. W.; Frosch, T. *Anal. Chem.* **2017**, *89*, 9997−10003.

(28) Frosch, T.; Schmitt, M.; Noll, T.; Bringmann, G.; Schenzel, K.; Popp, J. *Anal. Chem.* **2007**, *79*, 986−993.

(29) Harz, M.; Krause, M.; Bartels, T.; Cramer, K.; Rosch, P.; Popp, J. *Anal. Chem.* **2008**, *80*, 1080−1086.

(30) Neugebauer, U.; Schmid, U.; Baumann, K.; Holzgrabe, U.; Ziebuhr, W.; Kozitskaya, S.; Kiefer, W.; Schmitt, M.; Popp, J. *Biopolymers* **2006**, *82*, 306−311.

(31) Petry, R.; Mastalerz, R.; Zahn, S.; Mayerhofer, T. G.; Volksch, G.; Viereck-Gotte, L.; Kreher-Hartmann, B.; Holz, L.; Lankers, M.; Popp, J. *ChemPhysChem* **2006**, *7*, 414−420.

(32) Jahn, B.; Koch, A.; Schmidt, A.; Wanner, G.; Gehringer, H.; Bhakdi, S.; Brakhage, A. A. *Infect. Immun.* **1997**, *65*, 5110−5117.

DOI: 10.1021/acs.analchem.8b01038
Anal. Chem. XXXX, XXX, XXX−XXX

(33) Langfelder, K.; Jahn, B.; Gehringer, H.; Schmidt, A.; Wanner, G.; Brakhage, A. A. *Med. Microbiol. Immunol.* **1998**, *187*, 79−89.
(34) Brakhage, A. A.; Vandenbrulle, J. *J. Bacteriol.* **1995**, *177*, 2781−2788.
(35) Schwartze, V. U.; Hoffmann, K.; Nyilasi, I.; Papp, T.; Vagvolgyi, C.; de Hoog, S.; Voigt, K.; Jacobsen, I. D. *PLoS One* **2012**, *7*, 11.
(36) R. D. C Team. *R*; R Foundation for Statistical Computing: Vienna, Austria, 2011.
(37) Ryan, C. G.; Clayton, E.; Griffin, W. L.; Sie, S. H.; Cousens, D. R. *Nucl. Instrum. Methods Phys. Res., Sect. B* **1988**, *34*, 396−402.
(38) Guo, S. X.; Bocklitz, T.; Neugebauer, U.; Popp, J. *Anal. Methods* **2017**, *9*, 4410−4417.
(39) Ferrari, A. C.; Robertson, J. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2001**, *64*, 13.
(40) Harz, M.; Claus, R. A.; Bockmeyer, C. L.; Baum, M.; Rosch, P.; Kentouche, K.; Deigner, H. P.; Popp, J. *Biopolymers* **2006**, *82*, 317−324.
(41) Kumamoto, Y.; Taguchi, A.; Smith, N. I.; Kawata, S. *Biomed. Opt. Express* **2011**, *2*, 927−936.
(42) Wu, Q.; Hamilton, T.; Nelson, W. H.; Elliott, S.; Sperry, J. F.; Wu, M. *Anal. Chem.* **2001**, *73*, 3432−3440.
(43) Blackwell, M.; Vilgalys, R.; James, T. Y.; Taylor, J. W. Fungi. Eumycota: mushrooms, sac fungi, yeast, molds, rusts, smuts, etc., 2012. http://tolweb.org/.
(44) Brakhage, A. A. *Curr. Drug Targets* **2005**, *6*, 875−886.
(45) Van Der Linden, J. W. M.; Warris, A.; Verweij, P. E. *Med. Mycol.* **2011**, *49*, S82−S89.
(46) Mondon, P.; De Champs, C.; Donadille, A.; Ambroise-Thomas, P.; Grillot, R. *J. Med. Microbiol.* **1996**, *45*, 186−191.

[P8] Surface enhanced Raman spectroscopy-detection of the uptake of mannose-modified nanoparticles by macrophages in vitro: A model for detection of vulnerable atherosclerotic plaques

Copyright 2018 Wiley. Reprinted with kind permission from [Dugandžić, V., Drikermann, D., Ryabchykov, O., Undisz, A., Vilotijević, I., Lorkowski, S., Bocklitz, T.W., Matthäus, C., Weber, K., Cialla-May, D. and Popp, J., 2018. Surface enhanced Raman spectroscopy-detection of the uptake of mannose-modified nanoparticles by macrophages in vitro: A model for detection of vulnerable atherosclerotic plaques. *Journal of Biophotonics*, p.e201800013. doi: 10.1002/jbio.201800013]

Erklärungen zu den Eigenanteilen des Promovenden sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an der Publikation.

^{1}Dugandžić, V., ^{2}Drikermann D., ^{3}Ryabchykov, O., ^{4}Undisz, A., ^{5}Vilotijević I., ^{6}Lorkowski, S., ^{7}Bocklitz, T.W., ^{8}Matthäus, C., ^{9}Weber, K., ^{10}Cialla-May, D., ^{11}Popp, J. 2018. Surface enhanced Raman spectroscopy-detection of the uptake of mannose-modified nanoparticles by macrophages in vitro: A model for detection of vulnerable atherosclerotic plaques. *Journal of Biophotonics,* e201800013, doi: 10.1002/jbio.201800013											
Beteiligt an (Zutreffendes ankreuzen)											
	1	2	**3**	4	5	6	7	8	9	10	11
Konzeption des Forschungsansatzes	X				X	X		X	X	X	X
Planung der Untersuchungen	X	X	**X**		X	X	X	X	X	X	X
Datenerhebung	X	X		X							
Datenanalyse und Interpretation	X	X	**X**		X		X				
Schreiben des Manuskripts	X		**X**		X	X	X		X	X	X
Vorschlag Anrechnung Publikationsäquivalent		0.5	**0.25**								

Received: 12 January 2018 | Accepted: 22 May 2018
DOI: 10.1002/jbio.201800013

Journal of BIOPHOTONICS

FULL ARTICLE

Surface enhanced Raman spectroscopy-detection of the uptake of mannose-modified nanoparticles by macrophages in vitro: A model for detection of vulnerable atherosclerotic plaques

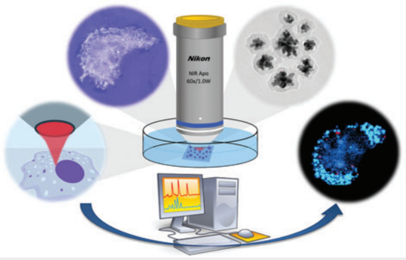
Vera Dugandžić[1,2,3] | Denis Drikermann[4] | Oleg Ryabchykov[1,2,3] | Andreas Undisz[5] | Ivan Vilotijević[4] | Stefan Lorkowski[3,6,7,8] | Thomas W. Bocklitz[1,2,3] | Christian Matthäus[1,2,3] | Karina Weber[1,2,8] | Dana Cialla-May[1,2,3]* | Jürgen Popp[1,2,3,8]

[1]Institute of Physical Chemistry, Friedrich-Schiller University Jena, Jena, Germany
[2]Leibniz Institute of Photonic Technology, Jena, Germany
[3]Abbe Center of Photonics, Friedrich Schiller University Jena, Jena, Germany
[4]Institute for Organic Chemistry and Macromolecular Chemistry, Friedrich Schiller University Jena, Jena, Germany
[5]Otto Schott Institute of Materials Research, Friedrich Schiller University Jena, Jena, Germany
[6]Institute of Nutrition, Friedrich Schiller University Jena, Jena, Germany
[7]Competence Cluster for Nutrition and Cardiovascular Health (nutriCARD), Halle-Jena-Leipzig, Germany
[8]Jena Centre for Soft Matter, (JCSM), Friedrich Schiller University Jena, Jena, Germany

***Correspondence**
Dana Cialla-May, Institute of Physical Chemistry, Friedrich-Schiller University Jena, Helmholtzweg 4, 07743 Jena, Germany.
Email: dana.cialla-may@leibniz-ipht.de

Funding information
Bundesministerium für Bildung und Forschung, Grant/Award Numbers: 03IPT513Y, 13N13856; Carl-Zeiss-Stiftung

Atherosclerosis is a process of thickening and stiffening of the arterial walls through the accumulation of lipids and fibrotic material, as a consequence of aging and unhealthy life style. However, not all arterial plaques lead to complications, which can lead to life-threatening events such as stroke and myocardial infarction. Diagnosis of the disease in early stages and identification of unstable atherosclerotic plaques are still challenging. It has been shown that the development of atherosclerotic plaques is an inflammatory process, where the accumulation of macrophages in the arterial walls is immanent in the early as well as late stages of the disease. We present a novel surface enhanced Raman spectroscopy (SERS)-based strategy for the detection of early stage atherosclerosis, based on the uptake of tagged gold nanoparticles by macrophages and subsequent detection by means of SERS. The results presented here provide a basis for future in vivo studies in animal models.



The workflow of tracing the SERS-active nanoparticle uptake by macrophages employing confocal Raman imaging.

KEYWORDS
atherosclerosis, gold nanoparticles, macrophages, mannose, SERS, silica coating

1 | INTRODUCTION

Cardiovascular diseases are the leading cause of death worldwide [1, 2]. Atherosclerosis, as a phenomenon of hardening and narrowing of arteries, is a major cause of cardiovascular diseases. Atherosclerosis is characterized by the formation of lesions in arterial walls, which contain inflammatory cells and smooth muscle cells, and are hallmarked by calcification, fibrosis and intracellular as well as extracellular accumulation of lipids [3]. The development of atherosclerotic plaques is a slow process which, to some extent, affects almost all people. Not all arterial plaques present danger and lead to sudden events such as stroke and myocardial infarction. Those that do not are denoted as stable plaques [4]. In the study conducted by van der Wal et al it was shown that an inflammatory process is involved in the rupture or erosion of atherosclerotic plaques, irrespective of the plaque morphology [5]. In simple terms, all

© 2018 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

vulnerable arterial plaques are characterized by the presence of macrophages [6].

The reliable diagnosis of vulnerable atherosclerotic lesions in the early stages of their development is a major challenge for modern medicine. Since there is no single indicator pointing toward the development of atherosclerotic lesions, health organizations have developed guidelines which help health practitioners to assess the risk of cardiovascular diseases including atherosclerosis [7]. The current state of the art in diagnosis of arterial plaques includes various blood tests, including LDL cholesterol levels [8] and inflammation biomarkers such as the high-sensitive C-reactive protein (hs-CRP) [9]. Various methods are available for the imaging of atherosclerotic plaques, such as magnetic resonance imaging (MRI) [10], computed tomography (CT) [11, 12], optical coherent tomography (OCT) [13] and intravascular ultrasound [14]. These methods require highly trained medical professionals for interpretation of results and while they shed light on morphological features of the atherosclerotic plaques, they fail to assess the vulnerability of the plaques. MRI and CT require employment of the suitable contrast agents, most often gadolinium based, which exhibit toxic effect and can lead to allergic reactions and kidney failure [15]. OCT is generally obstructed by blood and faces severe difficulties with imaging the aortic ostial lesions [16]. Therefore, there is a clear need for novel diagnostic methods which will allow the assessment of the vulnerability of plaques.

The last two decades have seen extensive research on the application of optical spectroscopic techniques for the detection and characterization of atherosclerotic plaques. Combined with OCT, fluorescence spectroscopy was applied to detect lipid-rich lesions stained by the lipid soluble dye indocyanine green [17]. Time-resolved laser-induced fluorescence spectroscopy is another powerful tool for the detection of prone-to-rupture atherosclerotic plaques in histopathological samples [18] and in vivo [19]. Infrared spectroscopy has been applied for the detection of atherosclerosis from dried blood samples in combination with an artificial neural network statistical approach [20]. Combined with vascular ultrasound, near-infrared spectroscopy has been employed to study the morphology of atherosclerotic plaques [21].

The high specificity of Raman spectroscopy and the possibility to directly measure biological samples have motivated the intense pursuit for the application of Raman spectroscopy in biomedical research. The development of a fiber probe-based Raman clinical setup, capable of intra-arterial probing, was first reported two decades ago [22], but these systems never made their way into routine application in clinics. Various groups have used Raman spectroscopy for the characterization of atherosclerotic lesions, contributing to the overall knowledge of the chemical composition and structural features of atherosclerotic plaques [23–28]. This includes reports on in vivo catheter-based Raman probe detection and characterization of plaques employing chemometric aproaches [29, 30]. A more fundamental study was performed by Stiebing et al, in which confocal Raman micro-spectroscopy was used to study the distribution and uptake dynamics of fatty acids and cholesterol by human macrophages [31–33].

To the best of our knowledge surface enhanced Raman spectroscopy (SERS) has not yet been employed for the detection of atherosclerosis. This fact is especially surprising in the light of the recent advances in nanoparticle synthesis. Plasmonic nanoparticles as SERS tags can be specifically designed to have excellent brightness allowing for simple signal readout, while specific targeting can be achieved through surface modification with specific molecules or antibodies. A similar approach has been employed for the detection of microscopic ovarian cancer using nano-probe based SERS [34]. The employment of SERS for the detection and risk assessment of atherosclerosis together with the use of specifically targeted SERS active nanoparticles would provide an easy and simple readout, independent of the interpretation by a trained medical professional. Furthermore, the direct readout relieves the SERS-based strategies from the need for applying statistical models. Due to the excellent sensitivity of SERS, the required amount of nanoparticles for the detection is minimal and the signal stability eliminates any signal bleaching issues. The only application of plasmonic nanoparticles for atherosclerosis detection is reported by Ankri et al [35], where gold nanorods were employed for the diffusion reflection measurement-based detection of atherosclerosis.

Since macrophages are abundant in prone-to-rupture atherosclerotic plaques, the ability to detect macrophages within the plaque would enable detection and early diagnosis of the vulnerable plaques. Herein, we report a novel strategy for the detection of dangerous atherosclerotic plaques based on SERS, which is applied for the detection in macrophage in vitro model. Our approach is based on the use of mannose-modified SERS-active gold nanoparticles tagged with a suitable reporter for targeting macrophages. The uptake of the nanoparticles by macrophages was traced by confocal Raman micro-spectroscopy in vitro and the results encourage further research into applications of SERS for both detection and characterization of vulnerable atherosclerotic plaques.

2 | EXPERIMENTAL

2.1 | Chemicals and instrumentation

Hydrogen tetrachloroaurate(III) hydrate (99.999%), (3-mercaptopropyl) trimethoxysilane (MTPMS) and succinic anhydride from Alfa Aesar and Trisodium citrate dihydrate, ammonium hydroxide (32%) and ethanol (99.8%) from Carl Roth were used. Polyvinylpyrrolidone with an average relative molecular weight of 55 000 (PVP K-55), tetraethylorthosilicate, 1,4-phenylene diisocyanide (PDI) and hydrazine hydrate were purchased from Sigma Aldrich. Extra dry dimethylformamide (DMF) and extra dry ethanol were supplied by ACROS organics and (3-aminopropyl) trimethoxysilane was supplied by abcr. All chemicals were used as received. Milli-Q grade water was used for nanoparticle preparation.

Glassware was washed with aqua regia and thoroughly rinsed with Milli-Q grade water. Nanoparticle synthesis and modifications were performed at room temperature. Centrifugation steps were performed at 5724*g* for 8 minutes if not indicated otherwise. Nanoparticle pellets were re-dispersed by ultrasonication. UV-Vis spectra were acquired applying a Jasco V-670 spectrophotometer. Malvern Nano-ZS Zetasizer was employed for DLS and Zeta potential measurements. The samples for DLS and Zeta potential measurements were prepared by diluting 10 µL of nanoparticle solution to 1 mL with Milly-Q water. TEM (Transmission electron microscopy) images were acquired using a JEOL JEM-3010 (300 keV). Raman measurements were performed on a WITec confocal Raman Microscope Alpha300 R coupled with a 785 nm CW diode laser. A Nikon water immersion objective with 60x magnification and numerical aperture (NA) of 1.00 was used. Raman spectra were detected with a cooled thermo-electric CCD. With the total laser power of 60 mW and a 1.00 NA water objective we estimate the peak intensity on the sample to be 240 $mW/\mu m^2$.

2.2 | Culture and differentiation of THP-1 cells

Human THP-1 macrophages were differentiated from THP-1 monocytes obtained from ATCC (Manassas, Virginia). Differentiation was induced by adding phorbol-12-myristate-13-acetate (0.1 mg/mL) and β-mercaptoethanol (50 µM) in L-glutamine supplemented RPMI 1640 medium for 96 hours under 5% CO2 atmosphere at 37 °C [36, 37]. For Raman micro-spectroscopy, the macrophages were grown on calcium fluoride slides with partial confluency.

2.3 | SERS-active gold nanoparticle synthesis

Branched gold nanoparticles were synthesized by reduction of tetrachloroauric acid with hydrazine hydrate in the presence of PVP K-55 and trisodium citrate as previously reported [38]. The solution containing 0.25 mM tetrachloroauric acid, 1.875 mM trisodium citrate and 0.3125 g/L PVP K-55 was prepared by adding up appropriate amount of stock solutions, respectively, with intense magnetic stirring. Three minutes after addition of PVP K-55, 0.25 mL of 60 mM hydrazine hydrate solution have been added into 9.75 mL of the above prepared solution under intensive magnetic stirring. Upon the addition of hydrazine hydrate, the color of the solution changed immediately to intensive blue. The prepared nanoparticles were characterized by UV-Vis spectrophotometry and TEM microscopy.

The SERS reporter PDI was added to the nanoparticles by mixing 50 µL of 0.01 mM solution of PDI in ethanol with 1 mL of the nanoparticle solution (Step a, Figure 1). This sample was denoted as AuNP@PDI.

2.4 | Silica encapsulation

Silica deposition was performed by a modified Stöber procedure by hydrolyzing tetraethylorthosilicate in a mixed water/ethanol solution in the presence of ammonia (Step b, Figure 1) [39]. Ten milliliter of the gold nanoparticle

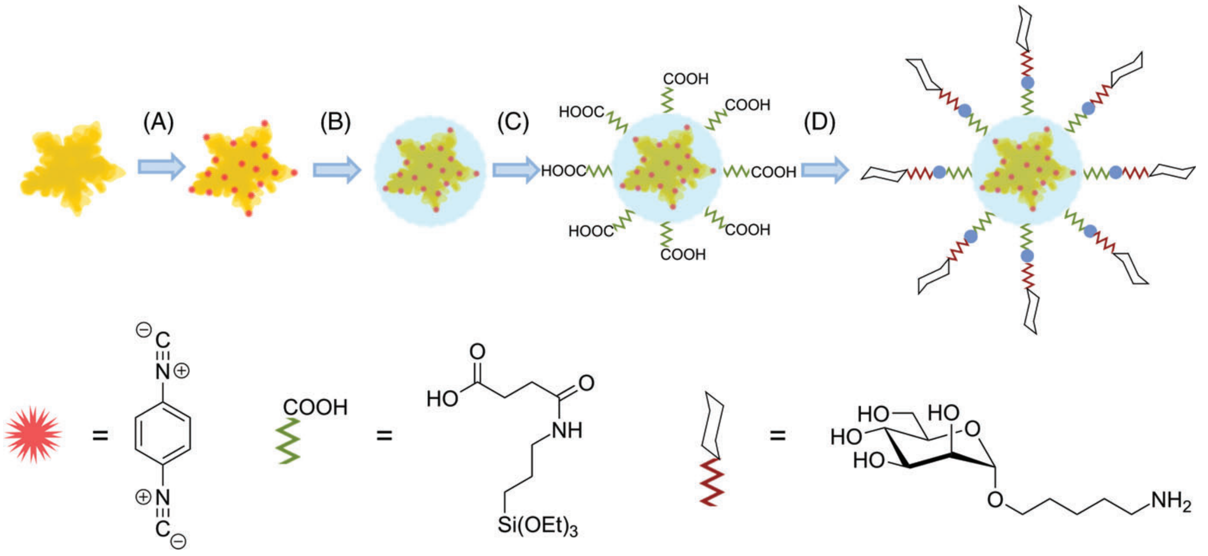


FIGURE 1 Schematic representation of the nanoparticle modification steps: (A) Addition of Raman reporter, PDI; (B) silica encapsulation; (C) functionalization of the silica surface with carboxyl groups; D) mannose binding to the nanoparticle surface via an EDC/NHS coupling reaction

suspension was concentrated to the volume of 1 mL by centrifugation and re-dispersion of the nanoparticles in 1 mL of the supernatant. The nanoparticles were then diluted with 3 mL of dry ethanol under magnetic stirring, followed by the addition of 450 μL of 0.01 mM (3-Mercaptopropyl) trimethoxysilane (MPTMS) in dry ethanol and 300 μL of 1 mM PDI in dry ethanol, respectively. The mixture was stirred for several minutes followed by addition of 3.33 mL of 10% (v/v) ammonium hydroxide solution in dry ethanol. The silica encapsulation was achieved by slow addition of 10% (v/v) tetraethoxysilane (TEOS) in dry ethanol over a total of 8 hours at a flowrate of 1.07 nL/s, having 1 hour break after every 30 minutes of the addition of TEOS. Addition of TEOS was precisely controlled by a neMESYS syringe pump system (Cetoni GmbH). After the reaction was completed, one third of the silanization mixture was removed from the reaction vessel, washed several times with water by centrifugation and re-dispersed in water. This sample was denoted as AuNP@PDI@silica. The remaining two thirds of the silanization mixture were used for further modification with mannose.

2.5 | Surface functionalization with mannose

Mannose functionalization of the silica surface was achieved via EDC (1-Ethyl-3-(3-dimethylaminopropyl)-carbodiimide)/NHS (N-Hydroxysuccinimide) promoted amide coupling starting from carboxyl functionalized silica coated gold nanoparticles and mannose bearing an amino-functionalized linker. The required mannose derivative was prepared in five steps via a modification of a previously reported procedure [40] from mannose and 5-aminopentanol. In order to bind amino functionalized mannose to the surface of silica, the nanoparticle surface was functionalized with carboxyl groups as described elsewhere (Step c, Figure 1) [41]. Briefly, 15.5 μL of (3-aminopropyl)triethoxysilane was mixed with an equimolar amount of succinic anhydride in 0.5 mL of dry DMF and left to react overnight with gentle stirring. The mixture was subsequently dissolved in 1 mL of dry ethanol and 22.1 μL of the resulting solution was added directly to the silanization mixture, followed by stirring for 90 minutes. The prepared carboxyl functionalized silica coated gold nanoparticles were centrifuged, washed twice with water and re-dispersed in 120 μL of water. Carboxyl modified nanoparticles were diluted with 130 μL of 30 mM MES (2-(N-morpholino)ethanesulfonic acid) buffer (pH 5.5). The obtained nanoparticle solution was mixed with 250 μL of freshly prepared solution containing 30 mg/mL EDC and 30 mg/mL NHS in 30 mM MES buffer (pH 5.5) and reacted for 10 minutes followed by addition of 250 μL of 5 mg/mL of amino modified mannose solution in water (Step d, Figure 1). The reaction could proceed overnight at room temperature with mixing. Nanoparticles were subsequently centrifuged at 3220 g for 6 minutes, washed with water and re-dispersed in 1 mL of water. This sample was denoted as AuNP@PDI@silica-man.

2.6 | Incubation of the nanoparticles with macrophages and Raman measurements

Mature macrophages were incubated with the prepared nanoparticle samples AuNP@PDI, AuNP@PDI@silica and AuNP@PDI@silica-man for 30 minutes as well as 2 hours. Prior to incubation, the samples were diluted to achieve equal optical density. After the incubation, the mature cells were washed with PBS buffer and fixed with 4% paraformaldehyde (PFA) solution for 20 minutes at room temperature.

The sample was located on a piezoelectrically driven microscope scanning stage and scanned through the laser focus in a raster pattern. Raman images were taken with a step size of 0.5 μm with an integration time of 0.1 second per step. The grating of 300 grooves/mm gave a spectral resolution of ~6 cm^{-1}. The cells were mapped with setting the laser focus in the middle of the cell respective to the vertical axis. Per sample, 10 to 14 cells were randomly chosen and imaged employing Raman microscopy.

2.7 | Data analysis

Data analysis was conducted using an in-house developed script in the programming language R [42] with use of the packages "imager" [43], "Peaks" [44] and "tiff" [45].

First, all spectra were background corrected using the SNIP (sensitive nonlinear iterative peak-clipping) algorithm [46], then the spikes were detected by the Laplacian operator and removed [47]. Subsequently, a Gaussian smoothing was applied. Unfortunately, within the spectra of high-fluorescence intensity, Raman signatures were not well-distinguishable, because large Poisson noise contributions decreased the signal-to-noise ratio of the Raman signal. Therefore the spectra that had the value of SD 20 times greater than the median value were excluded from further analysis. Furthermore, to obtain the visualization of the macrophages and estimate their area, the peak area in the range 1400 to 1500 cm^{-1}, originating from CH_2 scissoring of lipids, was used. In particular, the cell areas were estimated by k-means clustering with subsequent morphological closing and opening operations. However, in cases where the cell area was not estimated correctly, the fragments of other cells were present within the scan, or other artifacts appeared the cell area was selected manually. This manual procedure was done for 12 cells.

The presence of the nanoparticles in the cells was characterized by a peak in the range 2100 to 2200 cm^{-1}, which is attributed to the isocyano group of PDI. To detect the nanoparticles within the cells, the spectral range 2100 to 2200 cm^{-1} was integrated and an automated k-means-based threshold, which was adjusted for a high cut-off, was applied. In addition to the thresholding, an automated verification of the peak presence was performed by checking if the first derivative of a smoothed spectrum was changing the sign within the range 2120 to 2180 cm^{-1}. Only if the peak was found within the cell area and its integrated intensity was higher than the threshold, the pixel was marked and kept for further analysis.

To produce the false color images the peak around 1400 to 1500 cm^{-1} and the verified peaks in the range 2100 to 2200 cm^{-1} were clamped to the range from zero to the doubled value of 80th percentile and plotted in blue and red colors respectively. Technically, two separate images were created, one presenting the cell and the other one presenting the distribution of nanoparticles. These two images were further overlaid to give final images.

Subsequently, the total Raman intensity of the red channel as well as total number of the red pixels in each false color image of human THP-1 macrophages was calculated. These values were normalized by the size of the cell and used for the assessment of the abundance of the nanoparticles in each sample, raising the possibility to compare the uptake of each type of nanoparticles by human THP-1 macrophages. The Mann-Whitney *U* test was used to examine between-group differences. The Mann-Whitney *U* test is a nonparametric test used to determine whether the values from two samples have the same distribution which does not assume the distribution to be normal, what makes it more suitable than *t* test for non-normal or unknown distributions.

3 | RESULTS AND DISCUSSION

The concept for applying SERS and Raman spectroscopy for the detection and characterization of atherosclerotic plaques has been examined in a macrophage in vitro model, since macrophages are abundant in vulnerable atherosclerotic plaques. The branched gold nanoparticles were synthesized and modified with silica and mannose in a stepwise procedure. Branched nanoparticles were selected due to their well-documented efficient SERS activity without particle aggregation [38]. The prepared gold nanoparticles had an average size of 79 ± 18 nm, as estimated from TEM images. DLS measurements indicated a high polydispersity of the sample. However, the obtained hydrodynamic radii differ greatly from the values obtained from TEM. The measured Zeta potential was found to be −24.43 ± 0.25 mV. An appropriate reporter molecule was required for the nanoparticle uptake test traced by SERS and 1,4-phenylene diisocyanide appeared to be the best suited reporter molecule. PDI easily binds to gold surfaces and features isocyanide groups with strong signal in the Raman silent region of 2100 to 2200 cm^{-1}, allowing for the Raman detection within a complex matrix such as cells or tissue. The branched gold nanoparticles synthesized in the presence of PVP proved to be particularly suitable for functionalization with PDI since no agglomeration was observed. When round gold nanoparticles were used, extensive agglomeration was found due to the bidentate character of the PDI molecule. Silica encapsulation was carried out to prevent the gold nanoparticle from adsorbing matrix molecules and to allow for chemical modification of the particle surface [48]. The silica encapsulation resulted in a tremendous change of the zeta potential compared to the "naked" gold cores. The zeta potential of the silica coated nanoparticles was found to be 23.37 ± 0.17 mV. Subsequently, silanization with carboxy functionalized triethoxysilane (product of the reaction of (3-aminopropyl)triethoxysilane with succinic anhydride) was performed in order to further facilitating chemical modification of the surface of the particles by introducing carboxylic functional groups. Finally, mannose was bound to the SERS tag surface via an EDC/NHS promoted amide coupling reaction to produce the mannose-modified SERS tag particles. The binding of mannose was accompanied by a change of the zeta potential to the value of 27.2 ± 0.54 mV, indicating the successful binding.

A representative TEM image of the nanostructures modified with the reporter molecule PDI is shown in Figure 2A. The spiky structural features assured a high SERS intensity without agglomeration of these particles. To achieve better biocompatibility, the synthesized nanoparticles were coated with a silica shell (TEM image, Figure 2B). Silica encapsulation of gold nanoparticles is a widely used

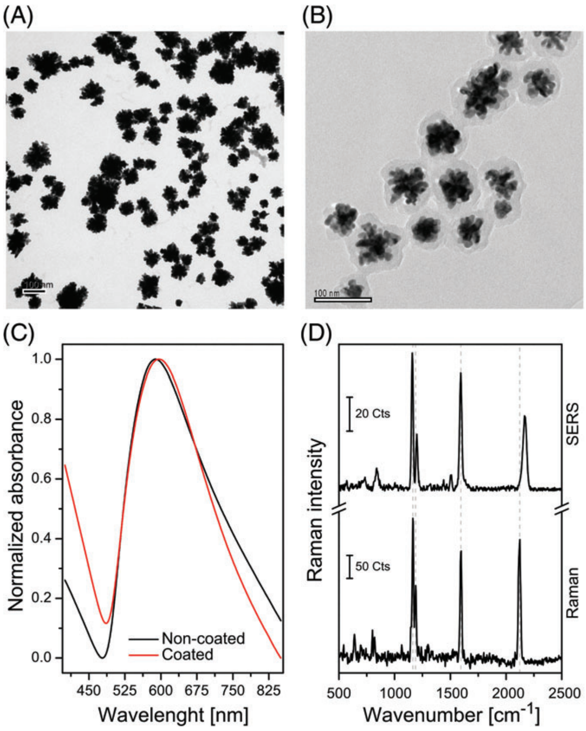


FIGURE 2 TEM image of (A) noncoated and (B) silica coated gold nanoparticles; (C) comparative UV-Vis spectrum of the gold nanoparticles before and after silica encapsulation; (D) Raman spectrum of PDI powder and SERS spectrum of a 1 × 10^{-5} M solution of PDI

method for achieving stability and biocompatibility of nanoparticles. Even though several publications report successful coating of round nanoparticles with silica [49, 50], coating of branched and star-shaped nanoparticles is rather demanding and hard to achieve. Slow addition of the siloxane precursor and priming of the gold surface appeared to be crucial for the successful silica encapsulation of the branched nanoparticles, avoiding the formation of empty silica nanoparticles. The addition of a small amount of (3-mercaptopropyl) trimethoxysilane before the silanization primed the surface and allowed for successful coating of the nanoparticles with silica; providing necessary hydroxyl groups on the gold surface. Further condensation of hydrolyzed TEOS was possible at this stage. The slow addition of the siloxane precursor assured a slower rate of polymerization of orthosilicic acid compared to the rate of the hydrolysis of the siloxane precursor, preventing nucleation of new silica nanoparticles. Consequently, the small shift of the localized surface plasmon resonance peak observed in the UV-Vis spectrum (Figure 2C) is due to the increase in the local refractive index around the particles [50]. The Raman spectrum of the reporter molecule PDI shows an intense peak at 2121 cm^{-1}, originating from the isocyanide (Figure 2D). In the SERS spectrum this peak is shifted to 2168 cm^{-1} indicating coordination of the isocyanide in PDI to the gold surface [51].

The appearance of the peak in a spectral window where no Raman features of cells are present is crucial for simple and rapid data processing and simplifies differentiation of the SERS signal of the nanoparticles from the Raman signals of the cells.

The abundance of macrophages in vulnerable atherosclerotic plaques makes them a perfect target for the uptake of our fabricated SERS nanoparticles, in particular due to their phagocytic character. Since mature macrophages express a mannose-specific C-type lectin receptor CD206 [52, 53], silica-coated gold nanoparticles whose surface is functionalized with mannose reduce the time needed for their uptake. A comprehensive investigation was conducted to assess the uptake of the prepared nanoparticles by macrophages. Mature macrophages were incubated with gold nanoparticles bearing only PDI reporter (AuNP@PDI), silica-coated gold nanoparticles (AuNP@PDI@silica) and mannose-modified silica-coated gold nanoparticles (AuNP@PSI@silica-man) for 30 minutes and for 2 hours, respectively. After incubation, macrophages were washed with PBS at room temperature to remove any physically deposited nanoparticles and fixed with 4% (w/v) PFA. Raman imaging was performed employing an integration time as short as 0.1 second, but still keeping the quality of the acquired spectra at a satisfactory level for subsequent data processing. All cell data was processed applying the same conditions using an in-house developed algorithm as described in the data analysis section. Since the cell figures are plotted using scaled intensities in desired regions, a peak recognition parameter was applied using a derivative approach to confirm that the reporter signal indeed originates from the PDI peak in the region of 2120 to 2180 cm^{-1}. The signal was considered positive only when the first derivative of the spectrum had a zero value at the spectral range where the reporter peak is expected. Applying this approach, false-positives arising from the background signal from the cells in which no reporter was found, were successfully eliminated. Corresponding reporter spectra from each red point in the Raman cell image are presented together with the Raman image for each investigated cell.

Figure 3 shows the Raman images of macrophages incubated for 30 minutes with gold nanoparticles bearing only the SERS reporter (AuNP@PDI). The corresponding Raman spectra of single pixels were plotted. The reporter signal was observed in only 3 out of 14 imaged cells (Figure 3A-C). In most cases, no specific signal of the Raman reporter was found; here, the spectra are dominated by the contributions of the cellular Raman signals (Figure 3D-N). The Raman measurement showed that gold nanoparticles without silica encapsulation were not taken up to a significant amount by macrophages after 30 minutes of incubation. Furthermore, no improvement of the particle uptake was observed even after 2 hours of incubation (Figure S1 in File S1, Supporting Information). An additional explanation of the observed low uptake of this type of nanoparticles could be derived taking into account the possibility that the SERS reporter—PDI—could have been desorbed from the nanoparticles' surface and replaced by competing molecules found in cells, resulting in an inability to detect the presence of the nanoparticles in cells by SERS.

In the next step, the silica-coated nanoparticles AuNP@PDI@silica were used for the incubation of macrophages. As shown in Figure 4, the Raman images as well as the corresponding spectra showing a specific signal assigned to the Raman reporter molecule are depicted for an incubation time of 30 minutes. As illustrated by these images, a specific signal of the SERS reporter was found in all randomly chosen and inspected cells, indicating that silica-coated nanoparticles were taken up by the macrophages to a higher extent than the non-coated AuNP@PDI particles.

It is evident that the uptake is not evenly distributed and varies from cell to cell. However, it is rational to expect that the uptake of nanoparticles will be proportional to the size of cells. The highest load of nanoparticles is observed in two cells depicted in Figure 4E and Figure 4J. It is, however, obvious that the size of these two cells is much bigger compared to other eight cells in the sample.

The improved uptake after silica encapsulation might be a result of an increased size of the nanoparticles upon silica encapsulation or/and different properties of the particle surface. However, no significant difference in the uptake was observed after 2 hours of incubation (Figure S2 in File S1) indicating that the maximum load was achieved already after 30 minutes of incubation.

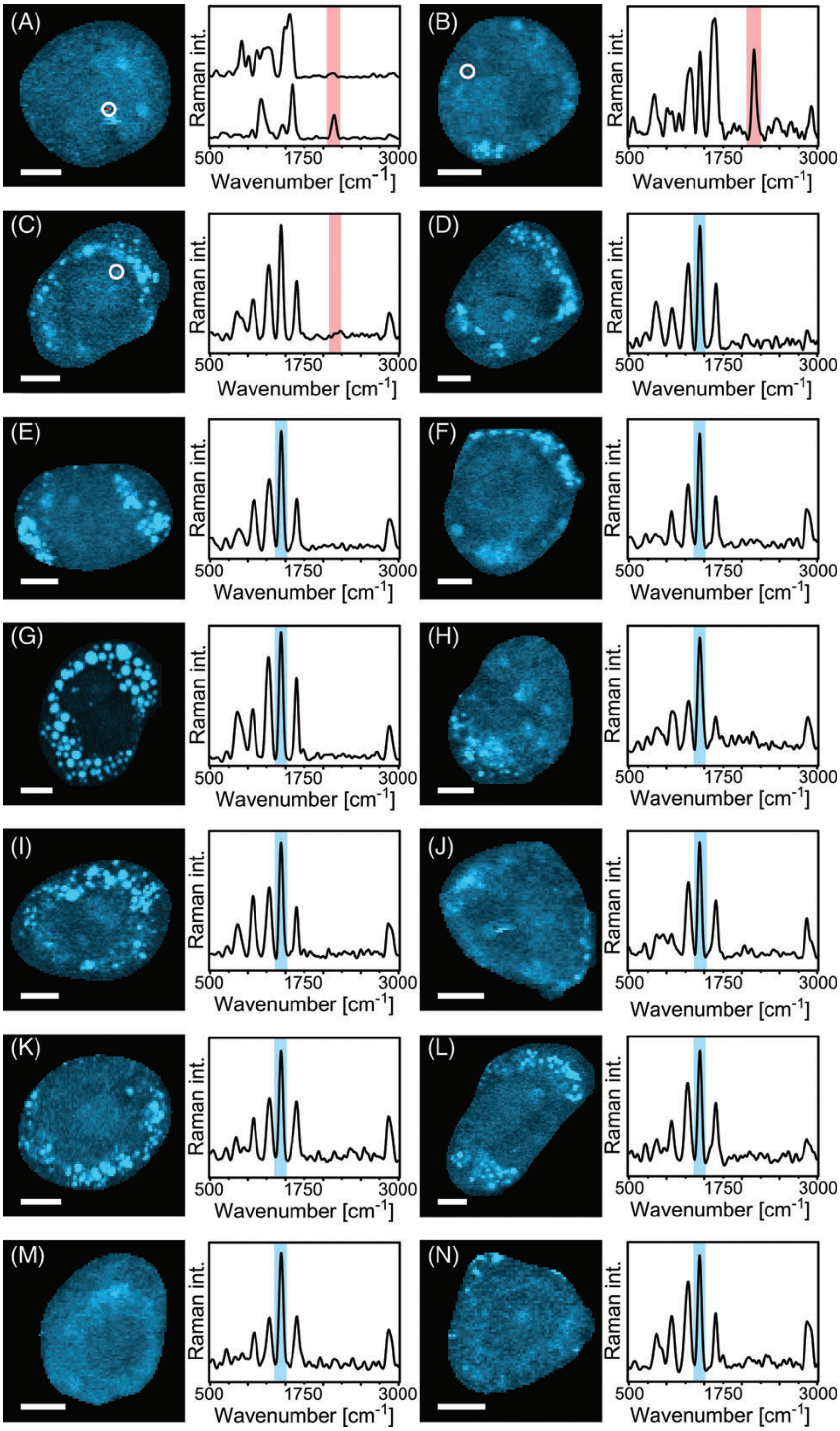


FIGURE 3 Raman images of human THP-1 macrophages incubated with AuNP@PDI for 30 minutes. (A-C) Raman images of macrophages together with corresponding point spectra, in which the signal of the Raman reporter was found (marker mode labeled with red). White circles indicate the location of red points in images for improved visibility; (D-N) Raman images of the macrophages, in which the signal of the Raman reporter was not found. The corresponding spectra present a Raman spectrum in a single pixel of a microphage image. The scale bar represents 10 μm

Finally, THP-1 macrophages were incubated for 30 minutes with the mannose-modified nanoparticles AuNP@PDI@silica-man. The Raman images as well as the corresponding pixel Raman spectra showing a contribution by the marker mode of the reporter molecule PDI are illustrated in Figure 5. The mannose modification further

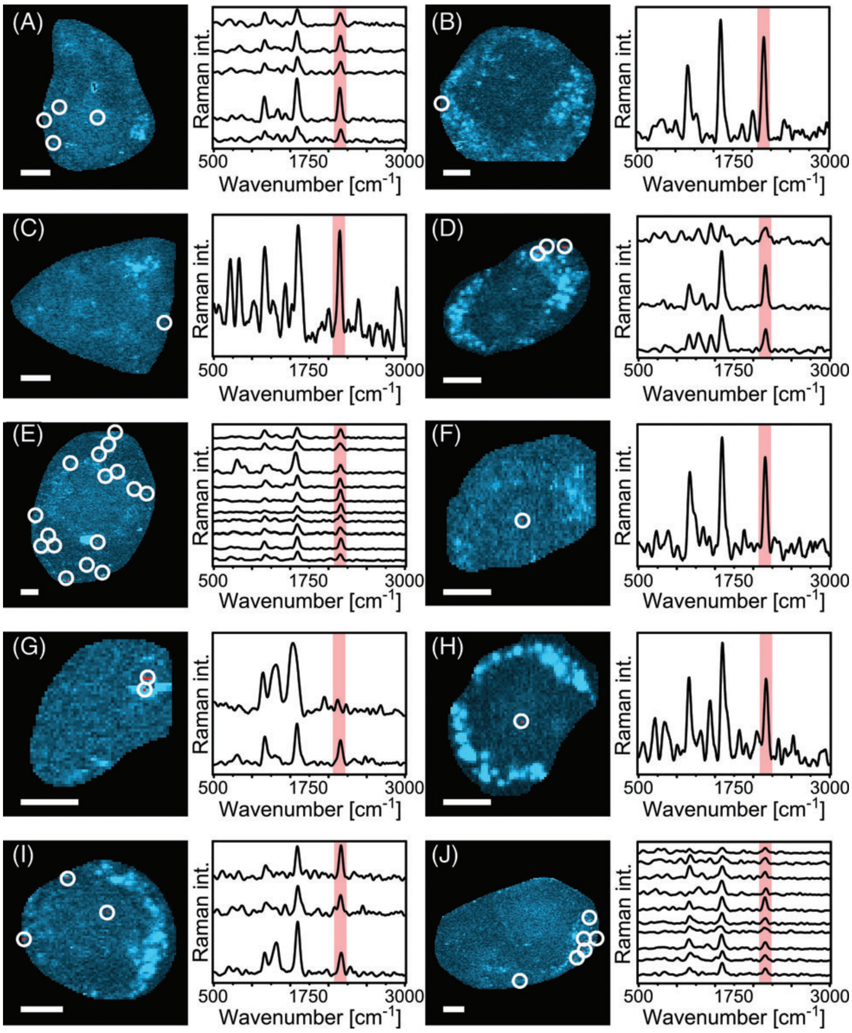


FIGURE 4 Raman images of human THP-1 macrophages incubated with AuNP@PDI@silica for 30 minutes together with corresponding point spectra extracted from each red point in the Raman image. The specific signal of the Raman reporter was found in all inspected cells (marker mode labeled with red). White circles indicate the location of red points in images for improved visibility. The scale bar represents 10 μm. *Cell E contained more than 10 red points. For clarity, only 10 randomly chosen spectra were plotted

improved the uptake of the nanoparticles leading to higher abundance of the reporter signal observed presumably due to the specific recognition of the mannose on nanoparticles by the mannose receptor on the macrophage surface. In the case of the mannose-modified nanoparticles, the uptake increased over time, leading to a significantly higher load of the nanoparticles in the macrophages after 2 hours of incubation (Figure S3 in File S1).

Evidently, the presence of mannose at the nanoparticles' surface facilitates their uptake, resulting in increased amounts of nanoparticle uptake over time as well as a significantly higher uptake of the nanoparticles after 30 minutes compared to the non-functionalized AuNP@PDI@silica particles. This is in accordance with the literature reports where mannose coating was utilized for the targeted delivery of nanoparticles to macrophages [52, 54].

In an attempt to compare the abundance of the nanoparticles in the cells incubated with different types of nanoparticles for 30 minutes and 2 hours, two semi-quantitative approaches have been employed. In a first approach, the total Raman intensity of the red channel (related to the number of nanoparticles taken up by the macrophages) in has been normalized by the total cell area in the image expressed in pixels and the average value has been calculated for all imaged cells in a sample. The obtained values are shown in the Figure 6A.

The lowest uptake was observed in the case of the nanoparticles bearing only SERS reporter (PDI), while the highest uptake was observed when the mannose-coated nanoparticles were used. Furthermore, no significant difference in uptake was observed for the samples AuNP@PDI and AuNP@PDI@silica with the increase of the incubation

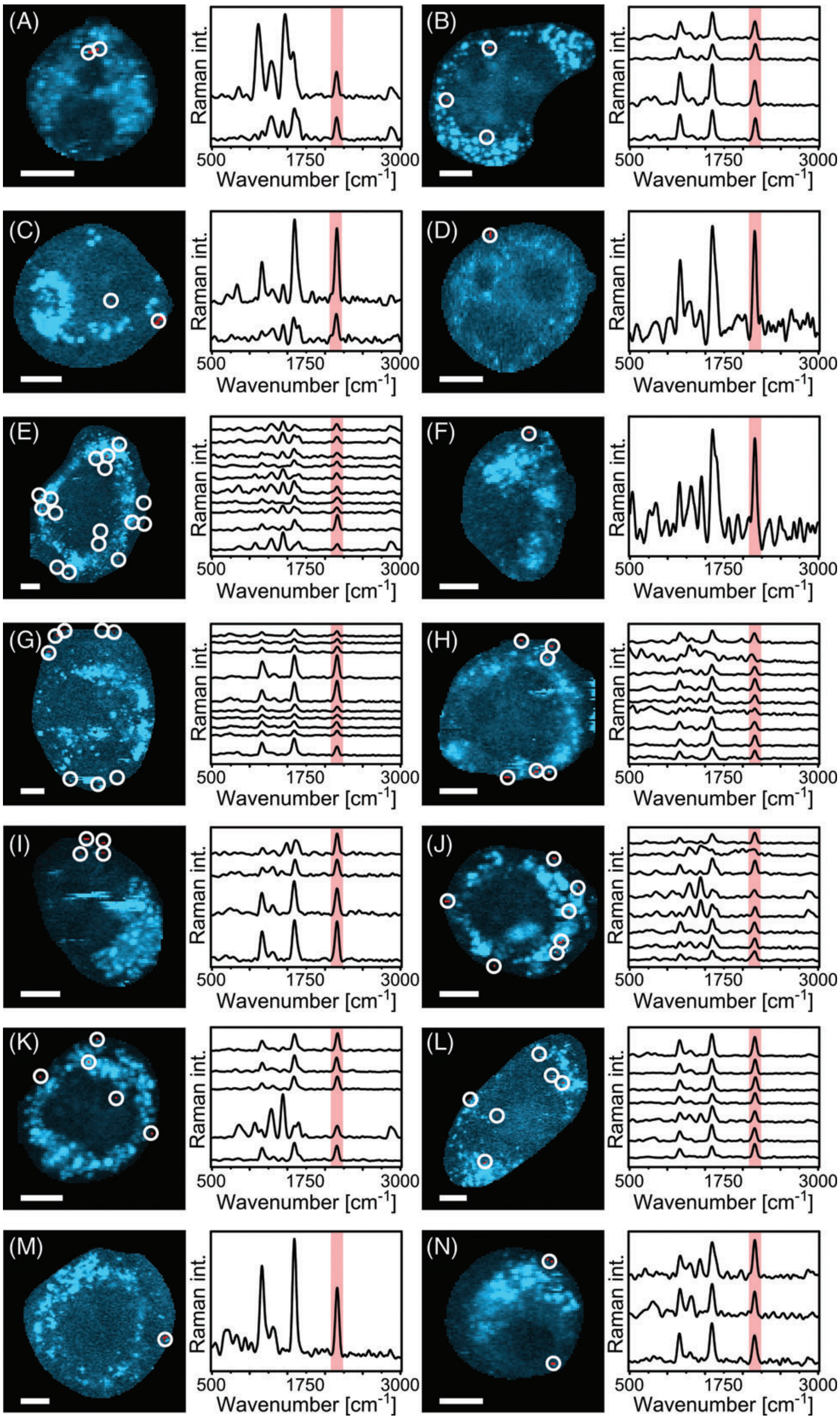


FIGURE 5 Raman images of human THP-1 macrophages incubated with AuNP@PDI@silica-man for 30 minutes together with the corresponding point spectra extracted from each red point in the Raman image. The specific signal of the Raman reporter was found in all inspected cells (marker mode labeled with red). The scale bar represents 10 μm. *Cell E contained more than 10 red points. For clarity, only 10 randomly chosen spectra were plotted. White circles indicate the location of red points in images for improved visibility

time, while in the case of the mannose-coated nanoparticles, the uptake rises dramatically with increased incubation time. The second approach was based on the total number of pixels in the image where the reporter signal was found. This number was normalized by the cell size expressed in pixels to obtain values which can serve as a semi-quantitative

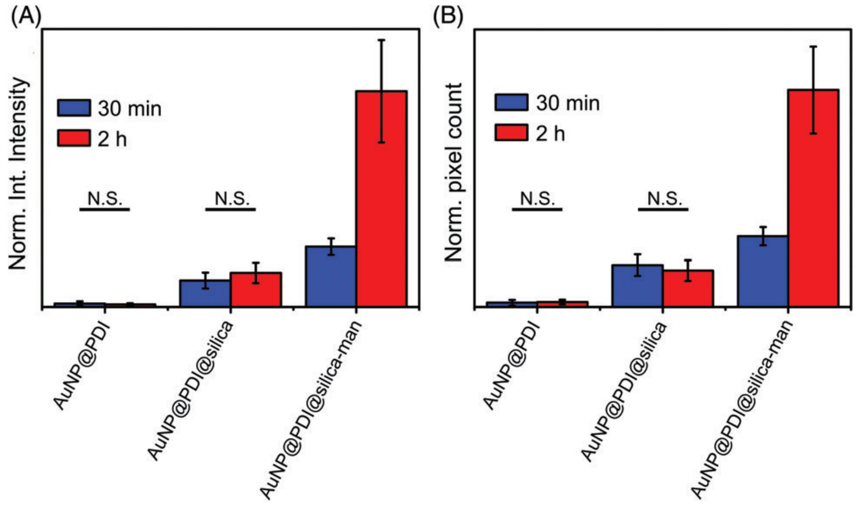


FIGURE 6 A comparison of the nanoparticle abundance in human THP-1 macrophages incubated with three types of nanoparticles (AuNP@PDI, AuNP@PDI@silica and AuNP@PDI@silica-man) for 30 minutes and 2 hours, respectively: (A) expressed as an average value of the total SERS intensity of the characteristic PDI band appearing in the Raman image normalized by cell size (Normalized integrated intensity); (B) expressed as an average value of the total number of pixels in which the PDI signal appears in a Raman image normalized by cell size (Normalized pixel count). Error bars represent the SE of the mean. Statistically significantly difference ($P < 0.05$) was found for all comparisons of mean values, except for those connected by a black bar and labeled with not significant (N.S.)

indication of the nanoparticle abundance in cells. Average values are calculated for all the samples and the obtained values are shown in Figure 6B. Again, the lowest uptake was found in the case of the nanoparticles bearing only the SERS reporter and the highest uptake was found for the mannose-coated nanoparticles with no significant change in the uptake for samples AuNP@PDI and AuNP@PDI@silica and a significant increase in the case of mannose-coated nanoparticles with the increase of the incubation time. Even though none of these parameters is fully quantitative, the same tendency present in both approaches serves as a strong indicator of the reliability of the applied strategy.

4 | CONCLUSION

In the present study we examined the possibility of applying SERS spectroscopy for the detection and characterization of vulnerable atherosclerotic plaques, by targeting macrophages as highly abundant cells in these lesions. Branched gold nanoparticles exhibiting SERS activity without particle agglomeration were synthesized and successfully surface-modified. The nanoparticles were first silica encapsulated to enhance their biocompatibility and to diminish potential toxicity. A successful silica encapsulation was achieved by slow addition of TEOS to the reaction mixture containing nanoparticles primed with MPTMS. Specific targeting of the macrophages was achieved by modification of the core shell Au-silica nanoparticles with mannose to allow specific binding to mannose receptors which are abundantly expressed on the surface of mature macrophages. Our results indicate that branched nanoparticles bearing only the SERS reporter are not suitable for the detection of macrophages due to the low uptake even after 2 hours of incubation. This low uptake of the branched nanoparticles could be a consequence of the spiky structure of the nanoparticles. Simple silica encapsulation already enhanced the uptake of the nanoparticles by macrophages, with a significant increase of uptake observed already after 30 minutes of incubation. Silica encapsulation is a widely used strategy to improve the biocompatibility of nanomaterials used for the application in biological systems, due to the high-chemical inertness of silica. The increase in the uptake of the silica-coated nanoparticles may also be a consequence of the increased size of the nanoparticles compared to the non-coated nanoparticles. As expected, the modification of the core-shell gold-silica nanoparticles with mannose led to the highest uptake of the nanoparticles within the short time frame of 30 minutes. Although a significant increase of the uptake of mannose-modified nanoparticles was observed after 2 hours, we deemed that incubation time of 30 minutes could be adequate for a clinical application. The described branched nanoparticles exhibit sufficient SERS activity without aggregation and allow for a better control of the size of the nanoparticles compared to methods based on SERS active nanoparticle agglomerates. A sensible choice of the SERS reporter, showing a signal in a spectral range where tissues and cells do not show Raman features (CN and CC triple bonds) is of utmost importance for in vivo applications. This allows for a simple and direct readout, avoiding the need for complex statistic posttreatment of the obtained datasets. With our results, in vivo studies in animal models appear promising. The uptake of mannose-modified nanoparticles by macrophages located in atherosclerotic plaques should be studied in more detail, paving the way for the future application of SERS for the detection and characterization of vulnerable atherosclerotic plaques in humans, for example using catheter Raman probes.

ACKNOWLEDGMENTS

Financial support from the Federal Ministry of Education and Research, Germany (BMBF) via the projects "JBCI2.0" (03IPT513Y) and "EXASENS" (13N13856) is gratefully acknowledged. Portion of this work was funded by the Carl-Zeiss Foundation endowed professorship (to I.V.). We thank Prof. Dr Andreas Henke (University Hospital Jena) for providing the human THP-1 monocytes and lab facilities for biological experiments and Dr Walter Müller (FSU Jena) for calculating the laser power density on the sample.

ORCID

Dana Cialla-May http://orcid.org/0000-0002-8577-1490

REFERENCES

[1] World Health Organization, www.who.int/mediacentre/factsheets/fs317/en/ (accessed: September, 2017).

[2] M. Naghavi, A. A. Abajobir, C. Abbafati, et al., *Lancet*, **2017**, *390*, 1151.

[3] M. C. Fishbein, *Cardiovasc. Pathol.* **2010**, *19*, 6.

[4] M. J. Davies, *Circulation*, **1996**, *94*, 2013.

[5] A. C. van der Wal, A. E. Becker, C. M. van der Loos, P. K. Das, *Circulation* **1994**, *89*, 36.

[6] K. J. Moore, F. J. Sheedy, E. A. Fisher, *Nat. Rev. Immunol.* **2013**, *13*, 709.

[7] P. Greenland, J. S. Alpert, G. A. Beller, E. J. Benjamin, M. J. Budoff, Z. A. Fayad, E. Foster, M. A. Hlatky, J. M. Hodgson, F. G. Kushner, M. S. Lauer, L. J. Shaw, S. C. Smith Jr., A. J. Taylor, W. S. Weintraub, N. K. Wenger, *J. Am. Coll. Cardiol.*, **2010**, *56*, e50.

[8] A. L. Catapano, A. Pirillo, G. D. Norata, *Br. J. Pharmacol.* **2017**, *174*(22), 3973.

[9] E. Corrado, M. Rizzo, G. Coppola, K. Fattouch, G. Novo, I. Marturana, F. Ferrara, S. Novo, *J. Atheroscler. Thromb.* **2010**, *17*, 1.

[10] F. Wiesmann, M. Szimtenings, A. Frydrychowicz, R. Illinger, A. Hunecke, E. Rommel, S. Neubauer, A. Haase, *Magn. Reson. Med.* **2003**, *50*, 69.

[11] J. A. Rumberger, D. B. Simons, L. A. Fitzpatrick, P. F. Sheedy, R. S. Schwartz, *Circulation*, **1995**, 92, 2157.

[12] S. Schroeder, A. F. Kopp, A. Baumbach, C. Meisner, A. Kuettner, C. Georg, B. Ohnesorge, C. Herdeg, C. D. Claussen, K. R. Karsch, *J. Am. Coll. Cardiol.* **2001**, *37*, 1430.

[13] X. Li, J. Li, J. Jing, T. Ma, S. Liang, J. Zhang, D. Mohar, A. Raney, S. Mahon, M. Brenner, P. Patel, K. K. Shung, Q. Zhou, Z. Chen, *IEEE J. Sel. Top. Q. Electron.* **2014**, *20*, 196.

[14] D. H. O'Leary, J. F. Polak, S. K. Wolfson, M. G. Bond, W. Bommer, S. Sheth, B. M. Psaty, A. R. Sharrett, T. A. Manolio, *Stroke* **1991**, *22*, 1155.

[15] M. Rogosnitzky, S. Branch, *Biometals* **2016**, *29*, 365.

[16] A. Alame, E. S. Brilakis, *Catheter. Cardiovasc. Interv.* **2016**, *87*, 241.

[17] S. Lee, M. W. Lee, H. S. Cho, J. W. Song, H. S. Nam, D. J. Oh, K. Park, W.-Y. Oh, H. Yoo, J. W. Kim, *Circ. Cardiovasc. Interv.* **2014**, *7*, 560.

[18] L. Marcu, J. A. Jo, Q. Fang, T. Papaioannou, T. Reil, J.-H. Qiao, J. D. Baker, J. A. Freischlag, M. C. Fishbein, *Atherosclerosis* **2009**, *204*, 156.

[19] J. Bec, D. M. Ma, D. R. Yankelevich, J. Liu, W. T. Ferrier, J. Southard, L. Marcu, *J. Biophotonics* **2014**, *7*, 281.

[20] A. S. Peters, J. Backhaus, A. Pfützner, M. Raster, G. Burgard, S. Demirel, D. Böckler, M. Hakimi, *Vib. Spectrosc.* **2017**, *92*, 20.

[21] S. K. Zacharias, R. D. Safian, R. D. Madder, I. D. Hanson, M. C. Pica, J. L. Smith, J. A. Goldstein, A. E. Abbas, *Vasc. Med.* **2016**, *21*, 337.

[22] I. James, F. Brennan, Y. Wang, R. R. Dasari, M. S. Feld, *Appl. Spectrosc.* **1997**, *51*, 201.

[23] A. S. Haka, J. R. Kramer, R. R. Dasari, M. Fitzmaurice, *J Biomed Opt.* **2011**, *16*, 011011.

[24] O. R. Šćepanović, M. Fitzmaurice, A. Miller, C.-R. Kong, Z. Volynskaya, R. R. Dasari, J. R. Kramer, M. S. Feld, *J Biomed Opt.* **2011**, *16*, 011009.

[25] S. W. E. van de Poll, D. J. M. Delsing, J. W. Jukema, H. M. G. Princen, L. M. Havekes, G. J. Puppels, A. van der Laarse, *Atherosclerosis* **2002**, *164*, 65.

[26] K. M. Marzec, T. P. Wrobel, A. Rygula, E. Maslak, A. Jasztal, A. Fedorowicz, S. Chlopicki, M. Baranska, *J. Biophotonics* **2014**, *7*, 744.

[27] A. Lattermann, C. Matthäus, N. Bergner, C. Beleites, B. F. Romeike, C. Krafft, B. R. Brehm, J. Popp, *J. Biophotonics* **2013**, *6*, 110.

[28] S. W. E. van de Poll, K. Kastelijn, T. C. Bakker Schut, C. Strijder, G. Pasterkamp, G. J. Puppels, A. van der Laarse, *Heart* **2003**, *89*, 1078.

[29] J. T. Motz, M. Fitzmaurice, A. Miller, S. J. Gandhi, A. S. Haka, L. H. Galindo, R. R. Dasari, J. R. Kramer, M. S. Feld, *J Biomed Opt.* **2006**, *11*, 021003.

[30] C. Matthäus, S. Dochow, G. Bergner, A. Lattermann, B. F. M. Romeike, E. T. Marple, C. Krafft, B. Dietzek, B. R. Brehm, J. Popp, *Anal. Chem.* **2012**, *84*, 7845.

[31] C. Stiebing, L. Schmölz, M. Wallert, C. Matthäus, S. Lorkowski, J. Popp, *J. Lipid Res.* **2017**, *58*, 876.

[32] C. Stiebing, T. Meyer, I. Rimke, C. Matthäus, M. Schmitt, S. Lorkowski, J. Popp, *J. Biophotonics* **2017**, *10*, 1217.

[33] C. Stiebing, C. Matthäus, C. Krafft, A.-A. Keller, K. Weber, S. Lorkowski, J. Popp, *Anal. Bioanal. Chem.* **2014**, *406*, 7037.

[34] A. Oseledchyk, C. Andreou, M. A. Wall, M. F. Kircher, *ACS Nano* **2017**, *11*, 1488.

[35] R. Ankri, D. Leshem-Lev, D. Fixler, R. Popovtzer, M. Motiei, R. Kornowski, E. Hochhauser, E. I. Lev, *Nano Lett.* **2014**, *14*, 2681.

[36] M. Schnoor, I. Buers, A. Sietmann, M. F. Brodde, O. Hofnagel, H. Robenek, S. Lorkowski, *J. Immunol. Methods* **2009**, *344*, 109.

[37] M. Schnoor, P. Cullen, J. Lorkowski, K. Stolle, H. Robenek, D. Troyer, J. Rauterberg, S. Lorkowski, *J. Immunol.* **2008**, *180*, 5707.

[38] G. H. Jeong, Y. W. Lee, M. Kim, S. W. Han, *J. Colloid Interface Sci.* **2009**, *329*, 97.

[39] W. Stöber, A. Fink, E. Bohn, *J. Colloid Interface Sci.* **1968**, *26*, 62.

[40] M. K. Patel, B. Vijayakrishnan, J. R. Koeppe, J. M. Chalker, K. J. Doores, B. G. Davis, *Chem. Commun.* **2010**, *46*, 9119.

[41] S. Kralj, M. Drofenik, D. Makovec, *J. Nanopart. Res.* **2011**, *13*, 2829.

[42] R Development Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria, the R Foundation for Statistical Computing. **2015**.

[43] S. Barthelme. **2017**, imager: Image Processing Library Based on 'CImg' R package, version 0.41.1

[44] M. Morhac **2012**, Peaks, R package, version 0.2

[45] S. Urbanek. **2013**, Package "tiff", R package, version 0.1-5.

[46] C. G. Ryan, E. Clayton, W. L. Griffin, S. H. Sie, D. R. Cousens, *Nucl. Instrum. Methods Phys. Res., Sect. B* **1988**, *34*, 396.

[47] O. Ryabchykov, T. Bocklitz, A. Ramoji, U. Neugebauer, M. Foerster, C. Kroegel, M. Bauer, M. Kiehntopf, J. Popp, *Chemom. Intell. Lab. Syst.* **2016**, *155*, 1.

[48] S. P. Mulvaney, M. D. Musick, C. D. Keating, M. J. Natan, *Langmuir* **2003**, *19*, 4784.

[49] S. M. Kang, B. S. Lee, S.-g. Lee, I. S. Choi, *Colloids Surf. A Physicochem. Eng. Asp.* **2008**, *313–314*, 150.

[50] L. M. Liz-Marzán, M. Giersig, P. Mulvaney, *Langmuir* **1996**, *12*, 4329.

[51] K. Nakamoto, *Applications in Coordination Chemistry*, John Wiley & Sons, Inc., Hoboken, New Jersey, **2008**, p. 1.

[52] S. Zhu, M. Niu, H. O'Mary, Z. Cui, *Mol. Pharm.* **2013**, *10*, 3525.

[53] J. D. Ernst, *Infect. Immun.* **1998**, *66*, 1277.

[54] Z. Cui, C.-H. Hsu, R. J. Mumper, *Drug Dev. Ind. Pharm.* **2003**, *29*, 689.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

File S1. Supporting Information

Figure S1 Raman images of human THP-1 macrophages incubated with AuNP@PDI for 2 hours. (A-D) Raman images of macrophages together with corresponding point spectra, in which the signal of the Raman reporter was found (marker mode labeled with red). White circles indicate the location of red points in images for improved visibility; (E-K) Raman images of human THP-1 macrophages, in which the signal of the Raman reporter was not found. The corresponding spectra present a Raman spectrum in a single pixel of a microphage image. The scale bar represents 10 μm

Figure S2 Raman images of human THP-1 macrophages incubated with AuNP@PDI@Silica for 2 hours. (A-I) Raman images of macrophages together with corresponding point spectra, in which the signal of the Raman reporter was found (marker mode labeled with red). White circles indicate the location of red points in images for improved visibility; (J) Raman image of a human THP-1 macrophage, in which the signal of the Raman reporter was not found. The corresponding spectra present a Raman spectrum in a single pixel of a microphage image. The scale bar represents 10 μm

Figure S3 Raman images of human THP-1 macrophages incubated with AuNP@PDI@Silica-man for 2 hours together with corresponding point spectra extracted from each red point in the Raman image. The specific signal of the Raman reporter was found in all inspected cells (marker mode labeled with red). White circles indicate the location of red points in images for improved visibility. The scale bar represents 10 μm. *Only 10 randomly chosen spectra were plotted in the cases of the cells where number of red points exceeded 10.

How to cite this article: Dugandžić V, Drikermann D, Ryabchykov O, et al. Surface enhanced Raman spectroscopy-detection of the uptake of mannose-modified nanoparticles by macrophages in vitro: A model for detection of vulnerable atherosclerotic plaques. *J. Biophotonics*. 2018;e201800013. https://doi.org/10.1002/jbio.201800013

List of publications and conference contributions

Refereed publications

1. Ryabchykov, O., Guo, S. and Bocklitz, T., 2018. Analyzing Raman spectroscopic data. *Physical Sciences Reviews*, doi:10.1515/psr-2017-0043.
2. Ryabchykov, O., Bocklitz, T., Ramoji, A., Neugebauer, U., Foerster, M., Kroegel, C., Bauer, M., Kiehntopf, M. and Popp, J., 2016. Automatization of spike correction in Raman spectra of biological samples. *Chemometrics and Intelligent Laboratory Systems*, *155*, pp.1-6.
3. Ryabchykov, O., Ramoji, A., Bocklitz, T., Foerster, M., Hagel, S., Kroegel, C., Bauer, M., Neugebauer, U. and Popp, J., 2016, September. Leukocyte subtypes classification by means of image processing. In *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on* (pp. 309-316). IEEE.
4. Bocklitz, T.W., Guo, S., Ryabchykov, O., Vogler, N. and Popp, J., 2015. Raman based molecular imaging and analytics: a magic bullet for biomedical applications!?. *Analytical chemistry*, *88*(1), pp.133-151.
5. Radu, A.I., Ryabchykov, O., Bocklitz, T.W., Huebner, U., Weber, K., Cialla-May, D. and Popp, J., 2016. Toward food analytics: fast estimation of lycopene and ß-carotene content in tomatoes based on surface enhanced Raman spectroscopy (SERS). *Analyst*, *141*(14), pp.4447-4455.
6. Ryabchykov, O., Bräutigam, K., Galler, K., Neugebauer, U., Mosig, A., Bocklitz, T. and Popp, J., 2018. Raman spectroscopic investigation of the human liver stem cell line HepaRG. *Journal of Raman Spectroscopy*, *49*(6), pp.935-942.
7. Ryabchykov, O., Popp, J. and Bocklitz, T., 2018. Fusion of MALDI Spectrometric imaging and Raman spectroscopic data for the analysis of biological samples. *Frontiers in Chemistry*, *6*, p.257.
8. Žukovskaja, O., Kloß, S., Blango, M.G., Ryabchykov, O., Kniemeyer, O., Brakhage, A.A., Bocklitz, T.W., Cialla-May, D., Weber, K. and Popp, J., 2018. UV-Raman spectroscopic identification of fungal spores important for respiratory diseases. *Analytical chemistry*, *90*(15), pp.8912-8918.

9. Dugandžić, V., Drikermann, D., Ryabchykov, O., Undisz, A., Vilotijević, I., Lorkowski, S., Bocklitz, T.W., Matthäus, C., Weber, K., Cialla-May, D. and Popp, J., 2018. Surface enhanced Raman spectroscopy-detection of the uptake of mannose-modified nanoparticles by macrophages in vitro: A model for detection of vulnerable atherosclerotic plaques. *Journal of Biophotonics*, p.e201800013. doi: 10.1002/jbio.201800013.
10. Lichodievskiy, V., Vysotskaya, N., Ryabchikov, O., Korsak, A., Chaikovsky, Y., Klimovskaya, A., Pedchenko, Y., Lutsyshyn, I. and Stadnyk, O., 2014. Application of oxidized silicon nanowires for nerve fibers regeneration. In *Advanced Materials Research* (Vol. 854, pp. 157-163). Trans Tech Publications.

Conferences

1. Ryabchykov, O., Bocklitz, T., Ramoji, A., Neugebauer, U., Bauer, M., Kiehntopf, M., Popp, J. (2014). Optimization and automatization of the analysis of Raman spectra for biomedical diagnosis. *24th International Conference on Raman Spectroscopy (ICORS 2014).* Poster presentation.
2. Ryabchykov, O., Ramoji, A., Bocklitz, T., Neugebauer, U., Bauer, M., Kiehntopf, M., Popp, J. (2015). Combination of image processing and Raman spectroscopy for automated white blood cell classification. *8th international conference on advanced vibrational spectroscopy 2015 (ICAVS8) in Vienna.* Poster presentation.
3. Ryabchykov, O., Bocklitz, T., Ramoji, A., Neugebauer, U., Bauer, M., Kiehntopf, M., Popp, J. (2015). Automation of data pre-processing for analysis by means of Raman spectroscopy. *DokDok2015.* Poster presentation.
4. Ryabchykov, O., Ramoji, A., Bocklitz, T., Foerster, M., Hagel, S., Kroegel, C., Bauer, M., Neugebauer, U., Popp, J. (2016). Leukocyte subtypes classification by means of image processing. *Federated conference on computer science and information systems (FedCSIS 2016), 6th International Workshop on Artificial Intelligence in Medical Applications (AIMA'16).* Talk presentation.
5. Ryabchykov, O., Bocklitz, T., Ramoji, A., Neugebauer, U., Popp, J. (2016). Microscopic cell images classification by means of pseudo-Zernike invariants. *DokDok2016.* Talk presentation.
6. Ryabchykov, O., Bocklitz, T., Neugebauer, U., Popp, J. (2017). Merging of vibrational and mass spectral data for investigation of biological samples. *15th Scandinavian Symposium on Chemometrics (SSC15).* Poster presentation.
7. Ryabchykov, O., Kirchberger-Tolstik, T., Popp, J., Bocklitz, T. (2018). Data fusion of Raman spectroscopic imaging and MALDI imaging for liver cancer diagnostic. *TOPIM TECH 2018.* Poster presentation.

Acknowledgments

I want to thank Prof. Dr. Jürgen Popp for giving me an opportunity to work on my doctoral thesis in his working group. Also, I would like to thank Prof. Dr. Ute Neugebauer for her helpful inputs and the opportunity to be a part of the research project's team.

I express my gratitude to the head of our junior working group Dr. Thomas Bocklitz, whose priceless intensive supervision and many hours of time spent on relevant scientific discussions made this work possible to get accomplished.

I want to thank Shuxia Guo, Olga Žukovskaja, Vera Dugandžić, Dr. Andreea Radu, Dr. Tatiana Kirchberger-Tolstik, Dr. Katharina Bräutigam and Dr. Anuradha Ramoji for our efficient and productive multidisciplinary collaborations.

And, of course, I am grateful for the professional, friendly, and positive atmosphere provided by colleagues from IPHT, UKJ, and IPC. It played a great role in my work.

In addition to current colleagues, I would like to express my gratitude to my former colleagues, and especially to my former supervisor Prof. Valerii Lozovski who always encouraged me to stay in research.

On a personal note, I am very grateful to my family, especially to my parents, Maryna and Anatoliy, and to my brother Oleksii for inspiring me throughout my whole life.

Finally, I want to thank my dear friends from Simferopol, Kyiv, and, of course, Jena for their comprehensive understanding and support in various branches of my life.

Erklärungen

Selbstständigkeitserklärung

Ich erkläre, dass ich die vorliegende Arbeit selbstständig und unter Verwendung der angegebenen Hilfsmittel, persönlichen Mitteilungen und Quellen angefertigt habe.

Oleg Ryabchykov

(Ort) (Datum) (Unterschrift)

Erklärung zu den Eigenanteilen des Promovenden/der Promovendin sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an den Publikationen und Zweitpublikationsrechten bei einer kumulativen Dissertation

Für alle in dieser kumulativen Dissertation verwendeten Manuskripte liegen die notwendigen Genehmigungen der Verlage („Reprint permissions“) für die Zweitpublikation vor.

Die Co-Autoren der in dieser kumulativen Dissertation verwendeten Manuskripte sind sowohl über die Nutzung, als auch über die oben angegebenen Eigenanteile informiert und stimmen dem zu (es wird empfohlen, diese grundsätzliche Zustimmung bereits mit Einreichung der Veröffentlichung einzuholen bzw. die Gewichtung der Anteile parallel zur Einreihung zu klären).Die Anteile der Co-Autoren an den Publikationen sind in diesem Kapitel aufgeführt.

Oleg Ryabchykov

(Ort) (Datum) (Unterschrift)

Einverständniserklärung des Betreuers

Ich bin mit der Abfassung der Dissertation als publikationsbasiert, d.h. kumulativ, einverstanden und bestätige die vorstehenden Angaben. Eine entsprechend begründete Befürwortung mit Angabe des wissenschaftlichen Anteils des Doktoranden/der Doktorandin an den verwendeten Publikationen werde ich parallel an den Rat der Fakultät der Chemisch Geowissenschaftlichen Fakultät richten.

Prof. Dr. Jürgen Popp ____
(Ort) (Datum) (Unterschrift)

PD Dr. Thomas Bocklitz ____
(Ort) (Datum) (Unterschrift)