# *INVERSE AND FORWARD MODELING TOOLS FOR BIOPHOTONIC DATA*

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

**Dissertation**

**(kumulativ)**

**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. Rola Houhou**

**geboren am 03 April 1985 in Barja, der Libanon**

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

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

**Tag der Verteidigung:** **13 Juli 2022**

# CONTENTS

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

**IR** Infrared spectroscopy

**SHG** Second-harmonic generator

**TPEF** Two-photon excitation microscopy

**CARS** Coherent anti-Stokes Raman scattering

**AI** Artificial intelligence

**ML** Machine learning

**DL** Deep learning

**HR** High-resolution

**LR** Low-resolution

**HQ** High-quality

**LQ** Low-quality

**PCA** Principal component analysis

**LDA** Linear discriminant analysis

**AUC** Area under the curve

**MSE** Mean squared error

**RMSE** Root mean squared error

**GS** Gerchberg-Saxton

**HIO** Hybrid Input-Output

**MEM** Maximum entropy method

**KK** Kramers-Kronig relation

**PCR** Principal component regression

**PLS** Partial least square regression

**SVM** Support vector machine

**CNN** Convolution neural network

**RNN** Recurrent neural network

**GAN** Generative adversarial network

**GRU** Gated recurrent unit

**LSTM** Long short-term memory

**NRB** Non-resonant background

**MM** Multimodal

**MAE** Mean absolute error

**PSNR** Peak signal-to-noise ratio

**ROI** Region of interest

**SSIM** Structural similarity index measure

**FDA** Functional data analysis

**SNR** Signal-to-noise ratio

**FPCA** Functional principal component analysis

**LOBOCV** Leave one batch out cross-validation

# LIST OF ACRONYMS

| | |
|---|---|
| $K$ | Linear or nonlinear operator |
| $x$ | An observation or measurement |
| $z$ | Parameters of the physical system or the real measurement |
| $I(\omega)$ | Intensity spectrum defined on frequency $\omega$ |
| $f(\omega)$ | Complex response function |
| $\nu$ | Normalized frequency |
| $\phi(\nu)$ | The phase of the complex response function |
| $N$ | Number of measurements |
| $L$ | Length of the spectrum |
| $I_{CARS}$ | The intensity of the CARS spectrum |
| $\chi^{(3)}$ | Nonlinear susceptibility |
| ${\chi_{nr}}^{(3)}$ | Non-resonant part |
| ${\chi_{nr}}^{(3)}$ | Raman resonant part |
| $\omega_{pu}$ | Frequency of the pump |
| $\omega_S$ | Frequency of the Stokes |
| $A_r$ | The amplitude of the $r^{\text{th}}$ Raman mode |
| $\Omega_r$ | The vibrational frequency of the $r^{\text{th}}$ Raman mode |
| $\gamma_r$ | The bandwidth of the $r^{\text{th}}$ Raman mode |
| $X_i(t)$ | A function defined on $t$, $\in L^2(I)$ |
| $L^2(I)$ | Hilbert space defined on $I$ |
| $I$ | A compact interval |
| $X$ | Matrix of observed data $\in \mathbb{R}^{N\times L}$ |
| $\psi_j$ | Set of basis functions defined on $I$ with an order $O$ |
| $c_j$ | Coefficients of the basis functions |

| | |
|---|---|
| $J$ | Number of basis functions |
| $\xi$ | Error threshold |
| $\varphi_k$ | Phase in the object plane |
| $\phi_k$ | Phase in the target plane |
| $I_{obj}$ | Object intensity |
| $I_{target}$ | Target intensity |
| $FFT$ | Fast Fourier transform |
| $S(\nu)$ | Power spectrum |
| $M$ | Number of poles |
| $b_0, b_k$ | Coefficients of the approximation |
| $R_l$ | Autocorrelation function |
| $\wp$ | Cauchy principal value |
| $\beta(t)$ | Weight function / eigenfunction |
| $f_i$ | Principal component scores |
| $\nu(t,s)$ | Covariance function |
| $\rho$ | eigenvalue |
| $V$ | Covariance operator |

# 1 INTRODUCTION

The outbreak of the recent COVID-19 pandemic has highlighted the challenges the healthcare systems of all countries of the world are facing. To address these challenges, new tools for clinical medicine to prevent such occurrence need to be developed. These developments should cover the three stages of clinical medicine, including the screening and the diagnosis of diseases, the treatment, and the surveillance of the patient response to treatment and the disease progression [1]. In healthcare systems, biomedical technologies are increasingly playing a role in all processes, from patient registration to data monitoring, from lab tests to self-care tools. From the design of X-ray machines to innovations in surgical practices, biomedical technologies improved our health and extended life expectancies [2], [3]. Of all biomedical technologies, biophotonics plays a crucial role in providing the most effective, lowest-cost approaches for diagnosing, treating, and preventing diseases [4]. In biophotonics, the molecular processes are analyzed by their interactions with light. The main idea is to utilize light to understand biological samples like biological tissue. To do so, optical processes such as reflection, absorption, elastic and inelastic scattering, and fluorescence are used to extract sample information. These processes can provide insights into the metabolic and pathological state of the tissue [5]. In this regard, both in vivo and ex vivo analyses can be implemented to study the structure and functions of molecules, cells, and organisms [6]. For instance, in vivo diagnosis of diseases using optical spectroscopy enables rapid clinical decisions without invasive biopsies. The incident light for in vivo scenarios can be delivered in a highly localized manner to tissue via optical fiber probes, which are placed within the working channels of minimally-invasive clinical tools, such as endoscopes [5].

The study of molecular processes using light has many benefits: it enables earlier and more accurate detection of diseases, customizes more effective treatments, causes fewer side effects, and is more practical due to the specific patient risk characterization and prediction of response to therapy [7]. However, several optical spectroscopic techniques exist to diagnose and monitor diseases and analyze molecular processes when biological samples interact with light. These techniques are considered part of molecular spectroscopy that explores in a non-invasive fashion the stereo-electronic, dynamic, and environmental effects of molecular systems, covering many chemistry areas [8]. In addition, novel techniques have enabled the study of many compounds, ranging from small molecules to metal complexes, organometallic compounds, and bio-related ions. For instance, fluorescence spectroscopy supplies information about the electronic excitation energy and the nature of the excited states. The key component in fluorescence spectroscopy is the fluorophore. In this technique, a molecule absorbs the energy of a specific wavelength and then emits energy at a different but equally specific wavelength. The amount and wavelength of the emitted energy depend on the fluorophore and its chemical environment [9], [10]. Moreover, infrared spectroscopy (IR) studies how light in the infrared region of the electromagnetic spectrum interacts with matter. It can measure molecular vibrations by recording the absorbed wavelengths of IR radiation. IR spectroscopy helps discover information about the structure of a compound or identify the functional groups, and it provides many absorption bands that represent the molecule fingerprint [11], [12].

Furthermore, dispersion and scattering are other optical processes that occur when light interacts with molecules in a gas, liquid, or solid. Most photons are scattered with the same energy as the incident photons, known as the elastic scattering or the Rayleigh scattering [13], [14]. However, a rare process, namely inelastic scattering or the Raman effect, occurs where a small number of these photons scatter at a different frequency than the incident photon [15], [16]. Similar to the IR spectroscopy, the Raman spectrum represents a vibrational fingerprint of the molecule(s). Both IR and Raman spectroscopy examine changes in vibrational and rotational states at the molecular level. However, infrared measures the amount of IR light absorbed while Raman measures the scattered light. Therefore, these technologies are considered complementary [16].

The optical methods described previously follow the linear superposition principle, which consists of adding the two separate inputs when these latter produce independent responses [17]. Besides these optical technologies, nonlinear optical spectroscopy is developed to obtain more information from samples than with linear spectroscopy. However, the superposition

principle no longer holds in nonlinear optics [18]. Therefore, many nonlinear optical techniques have been developed, varying between coherent and incoherent methods [19]. For instance, the second harmonic generation (SHG) is a coherent nonlinear process. In SHG, two incident photons at their fundamental frequency interacting with a medium are directly converted into a single photon at double frequency, without absorption or reemission of photons [20], [21]. SHG is an ideal tool for imaging collagen and other biological molecules breaking the centrosymmetry using a label-free approach. Furthermore, two-photon excitation microscopy (TPEF) is an incoherent technology motivated by the limitation of applying fluorescence spectroscopy to living systems due to photobleaching and phototoxicity. This limitation is controlled by maximizing the probability of detecting a signal photon per excitation event. Therefore, TPEF improves the detection of signal photons per excitation event, especially when imaging deep in highly scattering environments. In TPEF, two low-energy photons cooperate to cause a higher-energy electronic transition in a fluorescent molecule. It is a nonlinear process that depends on the second power of the light intensity [22]–[24]. Lastly, coherent anti-Stokes Raman scattering (CARS) microscopy involves two beams of light that simultaneously excite the sample. One is the pump beam, and the other of lower energy is the Stokes beam. CARS provides many advantages, including high resolution, high speed, high sensitivity, and non-invasive imaging of specific biomolecules without labeling [25]–[27]. Furthermore, it overcomes the limitation of the imaging speed of spontaneous Raman micro-imaging. Hence, these optical spectroscopy techniques provide detailed molecular structure and properties. They are considered important tools to examine the composition and nature of materials and chemicals in a non-destructive and non-intrusive manner. Nevertheless, unlike imaging, spectroscopic techniques have not achieved the same depth level due to different factors. Such factors may include a lack of sensitive and low-cost systems and dependence on expertise to interpret complex spectral signals [28].

Accordingly, the generated data from these spectroscopic techniques are enormous and often not directly utilized. Therefore, these data needs a specific treatment to extract the underlying information. The analysis of these data is related to chemometrics, the science of extracting information from measurements made on chemical systems using mathematical and statistical functions [29]. These measurements are either spectroscopic like near-infrared, fluorescence, Raman, chromatographic like gas chromatography, or physical like temperature, pressure, and concentrations. However, this thesis is limited to analyzing spectroscopic data acquired either as spectra or images. Typically, the importance of chemometric methods increases with the

increasing size of spectroscopic datasets [30], [31]. In addition, these methods are needed in various tasks, e.g., to correct artifacts and shortcomings of specific spectroscopic techniques.

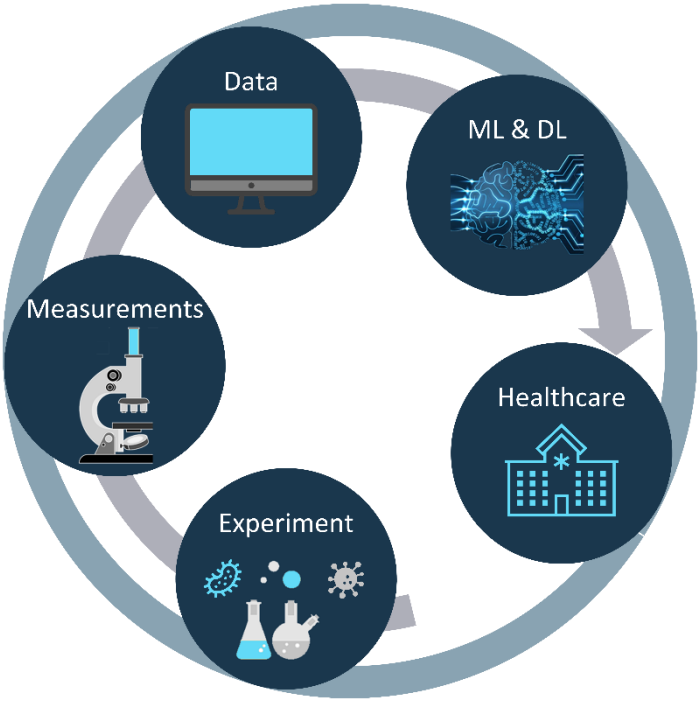


**Figure 1. The cycle of biophotonic analysis: from experimental design to artificial intelligence. The procedure of biophotonic applications started in the lab by designing the experiment. Then, linear or nonlinear optical spectroscopy is implemented, and data is subsequently collected. Afterward, artificial intelligence tools, including machine learning and deep learning, are applied to extract relevant information. Finally, setting effective and fast systems to improve the healthcare system.**

In summary, biphotonic analysis is indispensable for improving the healthcare system. Its cycle of biophotonic analysis, illustrated in Figure 1, started with the experimental design, followed by the measurements of the sample with a particular spectroscopic technique. Afterward, the corresponding data are collected, and finally, artificial intelligence (AI) techniques are applied, including machine learning (ML) and deep learning (DL) methods. These AI techniques can automate repetitive and expensive healthcare operations or assist physicians with real-time, data-driven insights.

In the last decade, relevant advancements in chemometric methods and the growth of new ones have been witnessed. These achievements have established new means for deeper investigations and characterizations for systems of increasing complexity. However, the analysis and interpretation of spectroscopic data are not always straightforward and pose

significant challenges. For instance, like any other type of measurement, optical measurements depend on detectors that convert the parameter to be measured into an electrical signal. Consequently, the measured signals can be very weak, and analysis of the noise characteristics of the signal source, the detectors, and the following electronic devices, become a part of the measuring procedure. In addition, some of the light parameters, such as the width of a femtosecond pulse, cannot be measured directly and require special optical instruments to be evaluated [32]. Therefore, mathematical and statistical methods are essential in predicting spectroscopic properties that can be directly compared with the experimental information or sometimes challenge the experiment itself [8]. These methods vary from chemometrics to deep learning techniques. However, chemometrics, machine learning, and deep learning methods can be grouped into three major sections: inverse modeling, pre-processing, and data modeling.

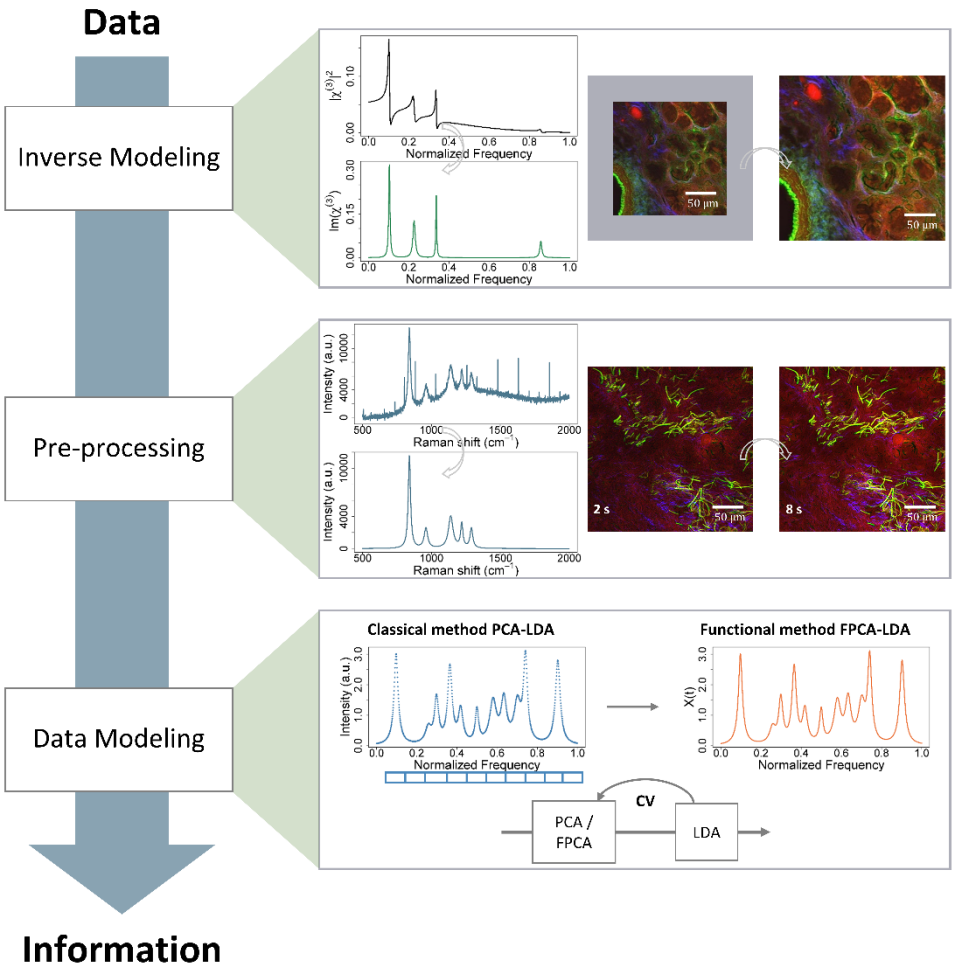


**Figure 2. Workflow of spectroscopic data analysis from inverse modeling to data modeling. In inverse modeling (the top row), the parameters are extracted to reconstruct what the measurements should be in reality. For instance, a Raman-like shape is extracted from the CARS spectrum, and spatially high-resolution images are constructed from low-resolution ones. Then, the pre-processing step (the middle row) constructs an improved version of the measurements since the measurement data is usually corrupted by artifacts. This step is applied for both spectral and image data. And finally, the data modeling (the bottom row) aims to extract the underlying information from data using a dimension reduction method and a classification method.**

First, inverse modeling is the procedure of recovering the parameters of the physical system from measurements that will enable the reconstruction of what these measurements should be in reality [33]. Next is the pre-processing step, which consists of a single or a combination of methods. The pre-processing methods construct a clean version of the data since the

measurements are usually distorted by unwanted artifacts, e.g., noises or background. And finally, the data modeling, also known as the approach of learning from data, in which the whole or a section of the data is used, and subsequently essential and relevant insights are extracted via feature extraction or selection [34], [35], classification, and regression. The three sections of the spectroscopic data analysis are shown in Figure 2. First, inverse modeling is applied either for spectral data or on images. In this step, for example, the Raman-like spectra are extracted from CARS spectra or super-resolution methods are implemented to extract spatially high-resolution (HR) images from low-resolution (LR) ones. The pre-processing step is then applied, where an improved version of the data is obtained. Since artifacts and noises distort the measurements due to the optical setup and other sources, various pre-processing methods can be implemented for spectra and images. Finally, the data modeling step is executed. This section aims to extract valuable information from the data, for example, marker values to diagnose diseases. Here, the relevant information is deduced by combining the dimension reduction method, the principal component analysis (PCA), and the classification method, the linear discriminant analysis (LDA).

Various mathematical and statistical techniques have been developed to extract meaningful and relevant insights from data. For instance, univariate or multivariate analysis can be implemented. In the univariate case, one variable is measured, or one is predicted, typically one wavelength is selected, and the absorbance change over time, for instance, is monitored [36]. This wavelength must not have contributions or overlap from other peaks. In contrast, multiple variables or predictions are used in the multivariate case, so the entire analysis typically utilizes the whole spectrum [37]. In this case, investigating the relationship between variables is allowed and reveals latent variation within a set of spectra. Therefore, multivariate analysis is performed on multiple sets of measurements, wavelengths, samples, and datasets where the analysis of variance and dependence between variables is crucial. The chemometric process begins by collecting data and then applying mathematical and statistical methods to extract relevant information from them. Chemometric methods remove redundant data, reduce variations not related to the analytical signal or image, and build models. The obtained information is related to the chemical process that gives knowledge about the system, which in turn facilitates decision-making.

Additionally, qualitative and quantitative analyses are employed in the data modeling section. For example, in classification models, the aim is often to predict a specific group with high accuracy. Hence, both data and labels are utilized to create a (classification) model, which is

subsequently evaluated using a cross-validation step. However, various cross-validation methods can be utilized. They vary from simple train/test splitting, k-fold cross-validation, bootstrapping, or Jackknife to batch-out cross-validation [38]. The cross-validation methods assess the model's generalization capability on an independent testing set. Accordingly, a confusion matrix is built, and the model performance is subsequently determined by evaluating the confusion matrix through the calculation of different metrics, e.g., sensitivity, specificity, accuracy. In addition, other techniques are utilized to test the model performance depending on the task studied, e.g., the area under the curve (AUC) [39], the mean, or the root mean squared error (MSE or RMSE) [40].

In chapter 2, the state of the art of existing mathematical methods is reviewed. Then, a detailed explanation of the researched questions and our proposed approaches and findings are provided in chapter 3.

# 2 STATE-OF-THE-ART

Mathematical and statistical methods can be used to extract as much information as possible from spectroscopic data. These methods can generally be categorized into data enhancement and data modeling. The data enhancement group belongs to the inverse problem category, which consists of recovering the parameters of the physical system from connected measurements. This recovery includes methods that either extract specific properties of the spectroscopic data known as inverse modeling or remove noise or artifacts from the measurements. In this context, the inverse modeling methods recover information about the system or its parameters from the measurements, which are directly inaccessible or expensive to acquire. However, data pre-processing methods are responsible for estimating the corrected version of the measurement data since it is disrupted by some artifacts usually resulting from the spectroscopic setup. Inverse problems arise in several signal processing applications, e.g., deblurring, denoising, super-resolution, reconstruction, segmentation, compressive sensing, inpainting. Moreover, the data modeling group is a forward problem that uses the measurement data to build a model that best represents the data and can extract relationships or predict specific characteristics of the sample or in the data set. A variety of unsolved problems exist in the inverse modeling, data pre-processing, and data modeling sections. Finding solutions for these problems is a step towards automated and fast diagnosis systems. However, the thesis is limited to answering some of the issues that are faced in spectroscopy, including phase retrieval, image denoising, super-resolution, and classification in the case of low-quality spectra.

## 2.1 Inverse Modeling

The inverse modeling is a category of the inverse problem class, which is generally formulated as follows

$$\boldsymbol{Kz = x,} \qquad \textbf{(1)}$$

where $x$ represents the measurements, $z$ describes the parameters of the physical system or the real measurements, and $K$ is the operator that connects $z$ and $x$ [41], [42]. The inverse problem aims to recover the parameters of the physical system from the measurements, e.g., the data.

In other words, it consists of using the measurements to extract hidden parameters or relevant properties of the physical system that are usually not available directly or complex to acquire. It has many applications, ranging from medical imaging and material characterization to parameter identification in systems biology.

For example, in signal and image processing, we often encounter the problem of finding an accurate estimation of the phase, which is challenging and sometimes impossible [43]. Generally, in optical measurements, the phase often contains a lot of information; however, only the square modulus of a signal is measured. Therefore, phase retrieval methods are implemented to recover the phase from the given magnitude measurement and subsequently reconstruct the full signal. Generally, the phase is more important than the magnitude in reconstructing a signal. Therefore, ignoring the phase and performing an inverse Fourier transform does not achieve adequate recovery [44]. Sophisticated measurement setups, e.g., holography, directly measure the phase by requiring interference with another known field. Since these methods are expensive and time-consuming, algorithmic methods can be used as an alternative to recover the phase from the given magnitude measurements. The reconstruction problem is well-known and arises in many engineering and applied physics areas, including optics, astronomical imaging, speech processing, computational biology, and blind deconvolution.

In image analysis, the phase encodes a lot of the structural content of the image, and important information is lost if the phase is not used in the reconstruction. Earlier approaches to phase retrieval were based on alternating projections, pioneered by the work of Gerchberg and Saxton [45], [46]. The Gerchberg-Saxton (GS) algorithm iteratively acquires the phase by starting with random initialization and using two intensity measurements obtained in time and Fourier domains (refer to the appendix for a detailed explanation of the algorithm). However, since the projections in the GS algorithm are between a convex set (for a time-domain) and a non-convex set (for the Fourier magnitude), the optimization often leads to a local minimum; consequently, the algorithm has limited recovery abilities even in a noiseless setting [47]. Modified versions of the GS algorithm are also popular in optical applications, e.g., the Hybrid Input-Output (HIO) algorithm [48], [49]. However, the convergence is not guaranteed to converge, and when it does, it might be to a local minimum.

In spectral analysis, the maximum entropy method (MEM) and the Kramers – Kronig relation (KK) are popular. These methods were mainly implemented in optical measurements, e.g., reflection spectroscopy [50], [51], and coherent anti-Stokes Raman scattering [52], [53]. First,

MEM states that any inference made from incomplete information should be based on the probability distribution with the maximum entropy permitted by the available data. In optical spectroscopy, only the intensity spectrum $I(\omega) = |f(\omega)|^2$ is measured, while the entire complex response function $f(\omega)$ is needed to obtain the desired material properties. Assuming that some additional information about $I(\nu)$, where $\nu$ is the normalized frequency, is available, the phase $\phi(\nu)$ is then retrieved and subsequently the imaginary and the real components of the function $f$ (refer to the appendix for mathematical explanation) [54]. Another popular method in optical spectroscopy is the Kramers – Kronig relation (KK), a mathematical relationship between the real and the imaginary components of a complex function developed according to the Cauchy residue theorem. As a result, the real component can be calculated if the imaginary component is available and vice versa. Since only the intensity spectrum is measured in optical spectroscopy, the KK relation can be modified to connect the phase with the intensity spectrum [55]. And subsequently, the phase is extracted, and both the imaginary and the real components are calculated (refer to the appendix for a detailed explanation of the method). Mainly, the MEM and KK methods require an error phase spectrum that involves a priori knowledge, and the problem of extrapolation beyond the range measured exists for the KK method.

Besides phase retrieval, another problem is encountered in many practical applications, including signal and image processing, which involves reconstructing a signal from a low number of measurements. This problem is called compressed sensing, formulated as shown in equation 1. However, the operator $K$ is a matrix $K \in \mathbb{C}^{N \times L}$ that models the measurement process where $x \in \mathbb{C}^L$ represents the observed data and $z \in \mathbb{C}^L$ is the signal of interest. Generally, it is impossible to recover $z$ from $x$ in the case $N < L$ without additional information. That is why it is suggested that the number of measurements $N$ must be as large as the signal length $L$ [56]. This fact also relates to the Shannon sampling theorem, which states that the sampling rate of a continuous-time signal must be twice its highest frequency to ensure successful reconstruction. Surprisingly, it has been shown that under certain assumptions, it is possible to reconstruct signals when the number $N$ of available measurements is smaller than the signal length $L$ [57].

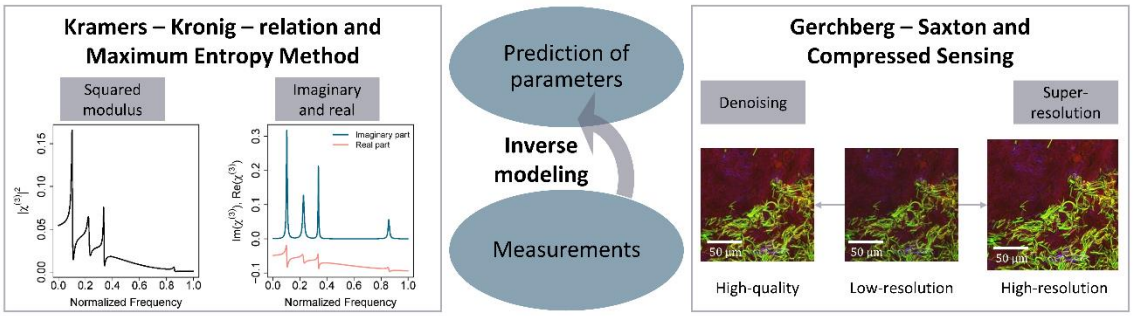


**Figure 3. Overview of inverse modeling methods used in signal and image processing. For instance, on the left panel, two existing methods KK and MEM, are shown, which can be used to extract the Raman-like spectrum from the CARS spectrum. The GS algorithm reconstructs a high-quality multimodal image on the right panel. However, the compressed sensing method reconstructs a spatially high-resolution multimodal image.**

The underlying assumption which makes all this possible is sparsity. The research area associated with this phenomenon has become known as compressive sensing, compressed sensing, compressive sampling, or sparse recovery. A signal is called sparse if most of its components are zero. As empirically observed, many real-world signals are compressible since they are well approximated by sparse signals often after an appropriate change of basis. The traditional approach of taking as many measurements as the signal is consuming since most of the coefficients are discarded in the compressed version of the signal [58]. Instead, one would want to acquire the compressed version of a signal "directly" via significantly fewer measured data than the signal length. In compressed sensing, the aim is to find the sparsest vector consistent with the measured data. A very popular and well-understood method is basis pursuit or $l_1$-minimization [59], which consists of finding the minimizer of the problem

$$\boldsymbol{min}\|\boldsymbol{z}\|_1 \; \boldsymbol{s.t.}\, \boldsymbol{Kz} = \boldsymbol{x}. \qquad \textbf{(2)}$$

The optimization problem can be solved with efficient methods from convex optimization since $l_1$-norm is a convex function. The algorithms used in compressive sensing can be divided into three categories: optimization methods, greedy methods, and thresholding-based methods [56]. In addition, various methods exist, e.g., basis pursuit, quadratically constrained basis pursuit, orthogonal matching pursuit, compressive sampling matching pursuit, basis thresholding, iterative hard thresholding, and hard thresholding pursuit. In Figure 3, inverse modeling methods, including phase retrieval and compressed sensing, are illustrated. The phase retrieval methods for signal processing through MEM and KK are displayed on the left panel. While on

the right panel, an additional phase retrieval method via GS and a compressed sensing method for image analysis are shown.

## 2.2 Pre-processing

Besides the process of interest, other processes often lead to different noise types, which can distort the measurement data. For instance, observations, i.e., spectra or images extracted from an optical setup, contain various distortions, e.g., Gaussian or Poisson noise, cosmic spikes, and background contribution. This distortion originated either from the sample or the process of the optical setup or fluctuation of the light source. Therefore, the problem occurs to estimate what the data is in reality, using these inadequate and noisy observations. We refer to this procedure as pre-processing, which is also part of the inverse problem category and it involves the correction of an image or a signal from disrupting contributions. Consequently, the pre-processing techniques aim to suppress unwanted distortions and improve the data. However, efforts for a common and best combination of pre-processing methods are widely investigated and attracted researchers in all spectroscopic fields. Although some rule of thumb exists, the workflow of pre-processing methods differs between spectroscopic methods and whether a spectrum or an image is analyzed. For instance, for spectral analysis, particularly Raman spectroscopy, a typical workflow suggested by Bocklitz et al. [60] starts by removing the spikes from the spectrum. A wavenumber calibration followed, and a baseline correction method headed a normalization technique afterward. Each of these steps uses a mathematical tool to get rid of a specific type of distortion.

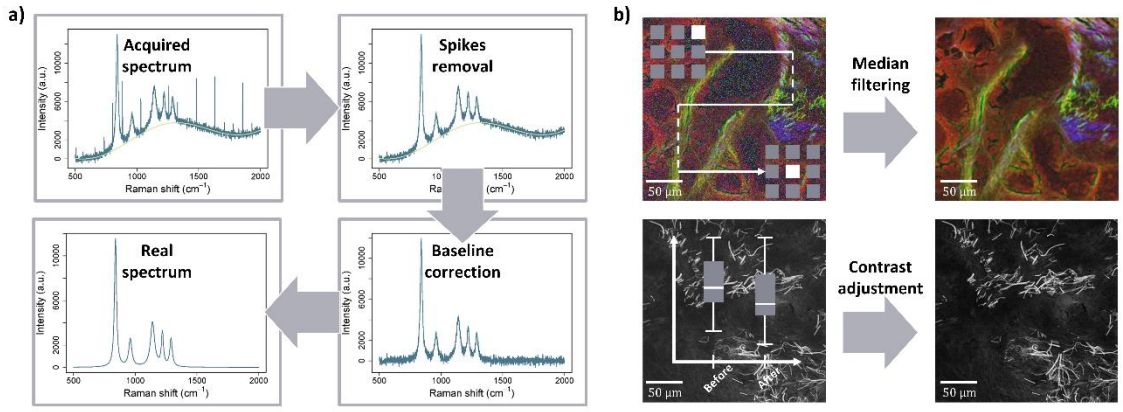


**Figure 4. Examples of some pre-processing methods used for spectral and image data. A typical workflow of pre-processing techniques for Raman spectra is shown in a). In contrast, the median filtering and the contrast adjustment applied on multimodal images are presented in b), which are standard pre-processing tools in image analysis.**

For image analysis, various categories exist, e.g., image filtering, contrast adjustment, and deblurring. These methods operate on images using specific mathematical relations and extract particular characteristics. Image filtering, for example, aims to emphasize certain features or remove other features [61]. However, contrast adjustment intends to enhance the brightness difference in the image between objects and their backgrounds [62]. Moreover, image deblurring aims to enhance image quality by removing distortion from blurry images [63]. Under the pre-processing group, various methods can be implemented to reach an improved version of the distorted data in hand. This group aims to prepare the data for further analysis in which an improved version of the data is essential for efficient analysis and decision-making. The pre-processing methods can either remove noise contributions, replace missing values, interpret or remove baselines, or even combine these targets. Some pre-processing methods for signal and image analysis are shown in Figure 4.

## 2.3 Data Modeling

Contrary to the inverse problem class, data modeling calculates what should be extracted from a particular data type. Typically, it consists of constructing and optimizing the parameters of a function that maps between the observations and some additional information about the data. For instance, one would like to classify specific data types or predict values from these observations or simply extract features from the data. Generally, data modeling methods are

grouped into either supervised and non-supervised methods, univariate and multivariate methods, or qualitative and quantitative methods. These task-oriented techniques have different objectives, e.g., dimension reduction, regression, and classification. An overview of unsupervised and supervised learning methods is shown in Figure 5.

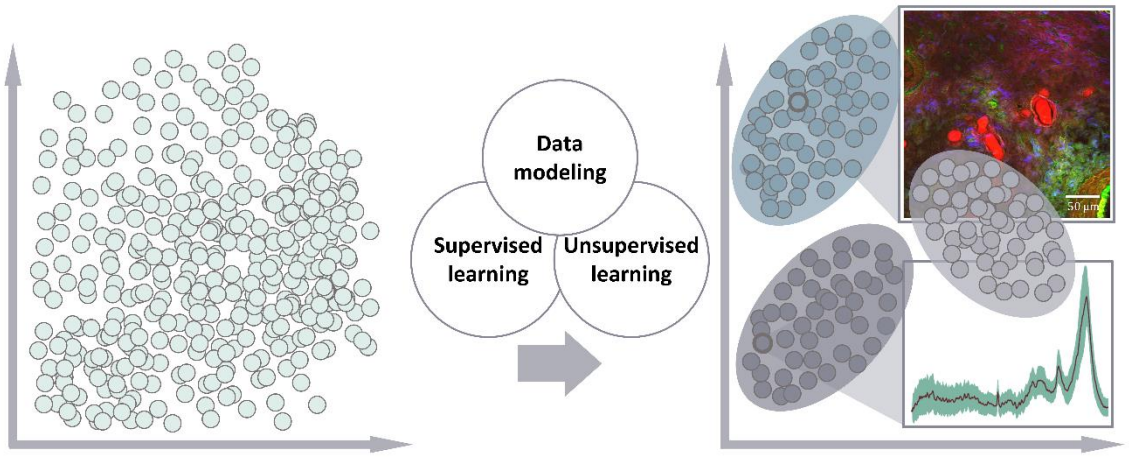


**Figure 5. Illustration of the data modeling. Data modeling can be divided into supervised and unsupervised methods. These methods are implemented in signal and image analysis.**

### 2.3.1 Unsupervised Learning

Due to the large amount of data acquired by optical measurements, it is often difficult to extract information, especially if unlabeled samples are involved. However, some techniques are developed to analyze unlabeled data, e.g., clustering and dimension reduction. The clustering methods consist of grouping data together that have some common (mathematical/statistical) characteristics. Such grouping can be performed by specifying the number of clusters beforehand and generating a k-means cluster analysis [64]. Using hierarchical clustering [65], a partition of the data can be generated without specifying the number of desired clusters. Besides clustering, dimension reduction tools are developed since the data size is enormous and using the whole data dimensionality sometimes leads to low performance due to redundancy and multicollinearity. These dimension reduction tools are primary steps implemented before further analysis. The further analysis can be grouped into feature extraction [66] or feature selection [35], [67]. The feature selection methods choose variables from the data that are highly important. These methods can be grouped into filter, wrapper, and embedded methods. In contrast, the feature extraction methods build a new subspace of variables that are easier to deal with since they represent a low-dimensional dataset. However, the feature extraction group consists of linear and nonlinear methods. For example, the

principal component analysis (PCA) is a linear method used frequently in feature extraction for spectroscopic analysis [68]. PCA reduces the number of features in a dataset while preserving as much information as possible by finding functions that approximate the data in the least-square sense.

### 2.3.2 Supervised Learning

In contrast, supervised learning methods depend on labeled samples and typically create a function that maps an input to a desired output, e.g., the labels. The supervised learning methods aim to approximate this mapping function that is used afterward to predict a new dataset. It includes regression and classification tasks. First, the regression consists of finding a combination (linear or nonlinear) between the features in the data to predict specific output variables, e.g., the concentration of drugs or other substances in the samples in Raman spectroscopy or the lifetime of fluorescence lifetime imaging microscopy. Various methods were developed for the regression task, including principal component regression (PCR) and partial least square regression (PLS). PCR apply first PCA to reduce dimensionality in the data and then train a regressor on the transformed data [69]. However, PLS differs from PCR by using labeled samples in the dimension reduction phase [70], [71]. Thus, PLS outperforms PCR in most cases. An important task that arises in analyzing spectroscopic data is classification. Similar to regression, it learns a mapping function between an input and a class label. The class label involves binary or multi-class labels, e.g., different kinds of species. Numerous methods were developed to differentiate between class labels and were frequently applied for spectroscopic data, e.g., linear discriminant analysis (LDA) and support vector machine (SVM). For instance, LDA involves finding linear discriminant functions that maximize the variances between the label groups and minimizing the variance within the label groups [72]–[74]. While in SVM, the aim is to search for an optimal hyperplane that optimally classifies the data [75], [76].

An illustration of the essential categories in the data modeling is shown in Figure 5, which involves supervised and unsupervised learning. Both learning categories are frequently applied in spectroscopy and represent useful, fast, and efficient tools to analyze samples of interest.

## 2.4 Deep Learning

Deep learning (DL) rose to prominence in the early 2010s, and it has achieved a revolution in the data science field with remarkable results. Different definitions for DL are present in

literature; however, DL is defined here in the context of inverse and forward problems. Previously the difference between the inverse and the forward problem was explained. In summary, through inverse modeling and pre-processing, the measurements are used to either extract specific parameters of the physical system or restore a better version of these measurements. In contrast, the forward problem transforms the measurements into meaningful outputs through data modeling. However, deep learning is a data-driven technique that can solve equally inverse and forward problems. An illustration of the inverse problem, e.g., data enhancement and pre-processing, as well as the forward problem, e.g., data modeling, in the context of deep learning, are displayed in Figure 6. DL is a subset of machine learning that consists of learning a function between an input and an output. Remarkably, DL does not require to choose the input and output carefully, like in the cases of inverse and forward problems. Instead, it works in both directions; the input and output can be chosen freely depending on the task. For instance, the input and output represent the measurements and the labels in a classification task, respectively. While in an inverse problem task, e.g., denoising, the input and the output represent the measurements and their clean version, respectively. Therefore, DL has increased its applications to include inverse and forward problems. In DL, feedforward deep network models are often meant and constructed. A feedforward network defines a mapping and learns parameters that result in the best prediction of this function. Deep learning is a vast field of research with multiple methods. These methods can be grouped into the convolution neural network (CNN) and the recurrent neural network (RNN).

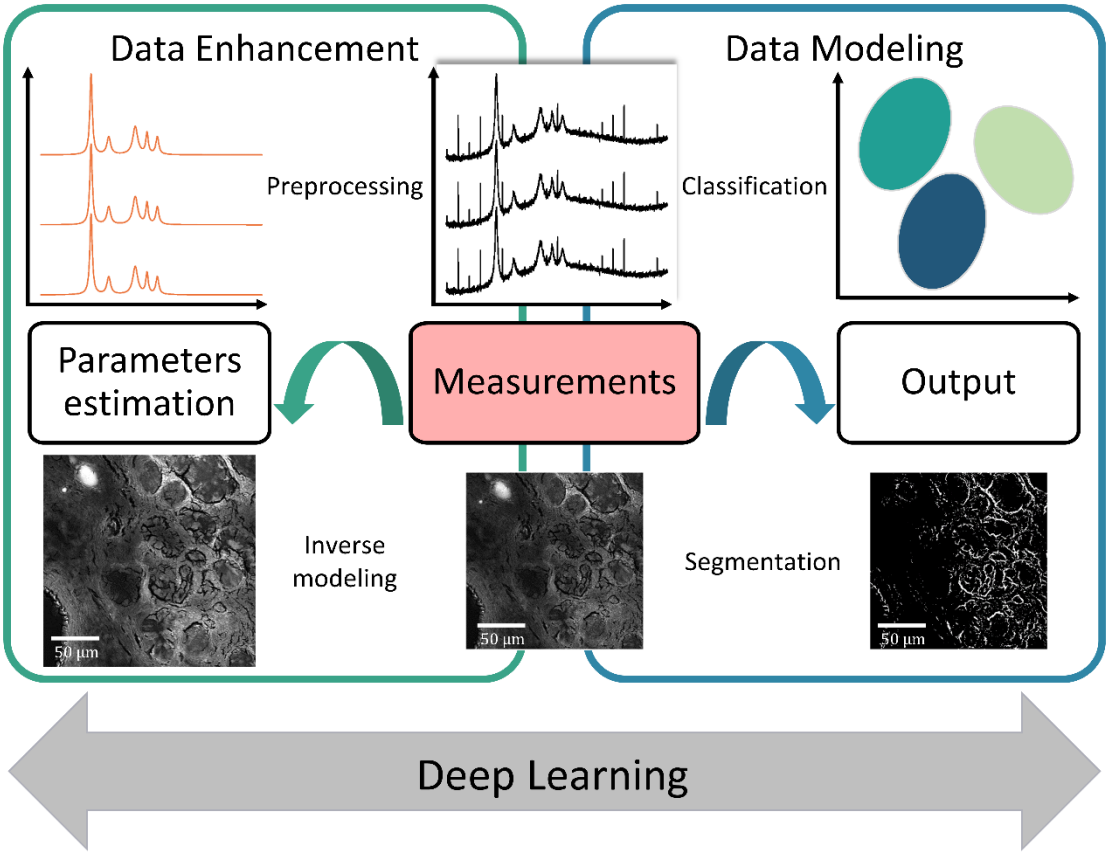


**Figure 6. Deep learning in the forward and inverse problem context. The inverse problem, e.g., data enhancement and pre-processing, uses the measurements to extract specific parameters of the physical system. In contrast, the forward problem transforms the measurements into meaningful outputs through data modeling. However, DL works in both directions; the input and output can be chosen freely depending on the task.**

One of the most popular deep neural networks is the CNN. CNN is a specialized neural network that applies convolutions between matrixes [77]. It has been tremendously successful in practical applications, e.g., object detection, image classification and segmentation, speech recognition, and video processing. A typical CNN has multiple layers; convolutional layers, non-linear activation layers, pooling layers, and fully connected layers. First, convolutional layers are performed, and it applies filters with specific kernel sizes to the original image to extract a feature map. The non-linearity layer is then applied, which is used to adjust or cut off the generated output. The next layer is the pooling layer, which consists of a down-sampling procedure to reduce the complexity of the input of the next layers. And finally, the fully connected layer is used similarly to neurons used in a traditional neural network. Each node in a fully connected layer is directly connected to every node in the previous and following layers. Furthermore, various improvements in CNN learning methodology and architecture were

performed to make CNN scalable to large, complex, and multi-class problems [78]–[80]. However, two specific CNN-based networks were developed, the generative adversarial networks [81] (GANs) and autoencoders [82]. GANs method is a particular CNN-based network that uses generative and adversarial models. In GANs, the two models, the generator and discriminator, play against each other. The generative (generator) model tries to produce fake outputs, while the discriminative model tries to discriminate artificially generated outputs from real ones. For example, the generator uses random numbers as input and returns an image. The generated image is then fed into the discriminator alongside ground-truth images. Finally, the discriminator returns probabilities, a number between 0 and 1, with 1 representing a prediction of an authentic image and 0 representing a fake image. Both networks are trying to optimize different and opposing objective functions or loss functions [83], [84]. The second important CNN-based network is the autoencoder. It represents another specific deep learning architecture that includes three components; encoder, bottleneck, and decoder. The encoder compresses the input and produces the bottleneck features, while the decoder reconstructs the input only using these bottleneck features. For instance, SegNet, an autoencoder, was developed for semantic pixel-wise segmentation. SegNet is a segmentation mechanism consisting of an encoder network, a corresponding decoder network, and a pixel-wise classification layer. In SegNet, the decoder network aims to map the low-resolution encoder feature maps to full input resolution feature maps for pixel-wise classification [85], [86].

On the other side, RNN is a different class of neural networks that allow previous outputs to be used as inputs while having hidden states. Typically, it deals with sequential data, e.g., time-series data. RNN is a network that can remember its previous input. It consists of storing information of previous inputs to generate the subsequent output of the sequence. It can process an input of any length, characterized by sharing weights across time. There are two significant obstacles in RNN: the exploding and the vanishing gradients. The exploding gradients appear when the algorithm assigns high importance to the weights. This problem can be solved by truncating or squashing the gradients. However, vanishing gradients occur when the values of a gradient are too small, and the model stops learning or needs too much time for optimization [87], [88]. Therefore, the gated recurrent unit [89] (GRU) and the long short-term memory [90] (LSTM) network have been developed to solve this. GRU was designed to handle the vanishing gradient problem by including a reset and update gate. However, LSTM was also intended for the same purpose. It uses three gates called input, output, and forget gate. The gates added in GRU and LSTM determine which information is to be retained for future predictions [91].

There are still not many deep learning applications in spectroscopy, particularly for biomedical applications. Furthermore, they could significantly complement the theoretical and experimental spectroscopy to accelerate the spectroscopic analysis, make predictions, and discover new characteristics.

# 3 CONTRIBUTION AND RESULTS

Briefly, the thesis was motivated by applying mathematical and statistical methods on spectral or image data that are either time-consuming to obtain or suffer from either low quality, or the existence of artifacts. We developed approaches and methods which work with such data in either forward or inverse problem sense. Finally, the application of the developed methods should help in clinical medicine, e.g., helping physicians in their decision making.

Therefore, the following questions are investigated: First, what is the main trend in chemometrics, machine learning, and deep learning? Second, can we acquire the Raman-like spectra from CARS spectra without a priori knowledge? Third, can we find a good denoising technique for reconstructing multimodal images? Finally, which method can be applied to better classify the low-quality Raman measurements?

Our contributions to address these questions are summarized in the following:

3.1. Chemometrics, machine learning, and deep learning methods are booming in spectroscopy. Therefore, the recent analysis techniques applied to chemical and spectroscopic measurements are reviewed. These methods are grouped into two main steps; data enhancement and modeling.

3.2. The coherent anti-Stokes Raman scattering (CARS) is implemented for faster measurements and higher signal strength. However, CARS spectra suffer from the non-resonant background (NRB) contribution. Existing methods cannot remove it entirely, and they are sensitive to the NRB strength. Therefore, we suggested the use of a deep learning network that predicts the Raman-like spectra directly from CARS spectra.

3.3. Multimodal images provide different information regarding the chemical composition of tissue samples. However, the measurement of high-quality images is time-consuming, which is unpracticable when a transfer to clinical application is needed. Therefore, we tested the phase retrieval method via GS, transfer learning method via DnCNN, and deep learning method via our developed network incSRCNN to acquire high-quality images from low-quality images, saving measurement time.

3.4. Although Raman data provide a fingerprint of the molecular structure of a sample, the measurement process takes a long time. Therefore, this study answers the following

research question: how can we improve the accuracy of a classification model when we have low quality of Raman data or Raman data with varying quality? In this regard, a novel method based on functional data analysis was implemented, which converts the Raman data into functions and then applies functional dimension reduction followed by a classification method.

The descriptions and discussions in the next sections will be based on the following publications and manuscripts (in the order of their appearance in the text, reprints are provided)

P1. R. Houhou and T. Bocklitz
Trends in artificial intelligence, machine learning, and chemometrics applied to chemical data
Analytical Science Advances, 2021, 2, 128-141

P2. R. Houhou, P. Barman, M. Schmitt, T. Meyer, J. Popp, and T. Bocklitz
Deep learning as a phase retrieval tool for CARS spectra
Optics Express, 2020, 28, 21002-21024

P3. R. Houhou, E. Quansa, T. Meyer-Zedler, M. Schmitt, Franziska Hoffmann, O. Guntinas-Lichius, J. Popp, and T. Bocklitz
Comparison of denoising tools for reconstruction of nonlinear multimodal images
Biomedical Optics Express, 2022, submitted

P4. R. Houhou, P. Rösch, J. Popp, and T. Bocklitz
Comparison of functional and discrete data analysis regimes for Raman spectra
Analytical and Bioanalytical Chemistry, 2021, 413, 5633-5644

## 3.1 Trends in chemometrics and machine learning methods for chemical data

As mentioned previously, optical devices used to measure biological samples produce massive amounts of data that cannot be manually analyzed. Therefore, chemometric, machine learning, and deep learning methods extract hidden information from these data by applying mathematical and statistical techniques. These methods are sample- and task-dependent. A review of recent chemometric, machine learning, and artificial intelligence methods utilized on chemical data is presented in the contribution [P1]. The investigation of these methods is limited to specific spectroscopic measurements and imaging approaches, including nuclear magnetic resonance, mass spectroscopy, vibrational spectroscopy, X-ray, atomic force microscopy, electron microscopy, and two-dimensional chromatography. Chemometric, machine learning, and deep learning methods were classified into two main groups: data enhancement and modeling.

The data enhancement group includes either reconstructing missing information or removing artifacts from observed measurements. In this context, different algorithms have been recently developed to either recover the structure and composition of materials, predict the material design, find the best pre-processing method, or combine techniques in which the order of use is essential. In contrast to these methods, deep learning is extensively applied to invert relevant characteristics or remove noise from the measured data via CNN, autoencoder, and LSTM.

On the other hand, the data modeling group applies chemometric and machine learning techniques to extract information from the data, either spectra or images. These methods were applied to identify features, discover biomarkers, and provide comprehensive information on chemical changes during a particular experiment. Most trends include the application of deep learning methods. The deep learning methods focused mainly on classification tasks via the CNN network, which proved to be more rapid and accurate than standard classification methods. Some chemometric, machine learning, and deep learning methods applied to spectroscopic measurements are presented in Figure 7.

In conclusion, attempts to improve predictive quality, robustness, and automation using chemometric, machine learning, and deep learning methods were performed in many application fields. Also, investigations to develop new AI-based techniques are increasing in chemistry as well.

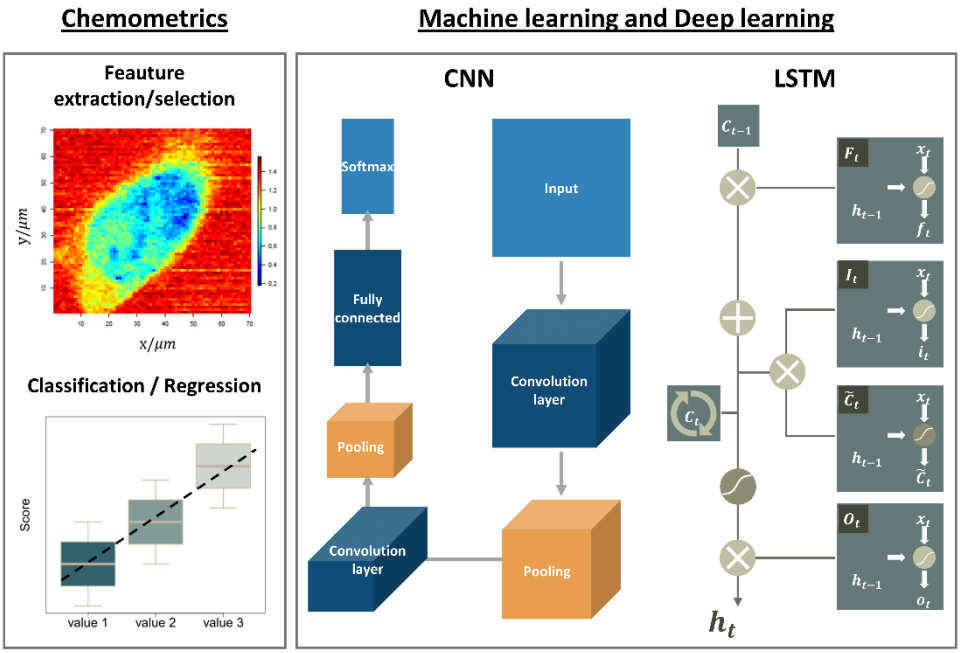


**Figure 7. An overview of chemometric, machine learning, and deep learning methods. Chemometric methods involve feature extraction or selection, regression, and classification tasks. Machine learning and deep learning, including CNN and LSTM networks, are implemented in spectroscopy.**

## 3.2 A novel approach to extract the non-resonant background from CARS spectra

Although spontaneous Raman scattering is a label-free, molecular fingerprint technique, it shows some drawbacks, including its weak spectroscopic signals, the fluorescence contributions of biological samples, and the long acquisition times [92]. Instead, CARS can be used since it produces much stronger vibrational sensitive signals than spontaneous Raman scattering for scatterers with a high concentration [93]. However, CARS spectra include the non-resonant background (NRB) contribution, which distorts the vibrational signals and leads to constructive and destructive interference effects [55]. Therefore, algorithmic phase retrieval methods are implemented to extract the phase without physically removing the non-resonant background, e.g., MEM and KK. However, MEM and KK require an accurate measurement of the non-resonant background. Therefore, we proposed an alternative solution for phase retrieval through deep learning using the LSTM network.

Our approach started by training a deep learning network via LSTM on an artificial dataset, then transferring it to an experimental dataset, and finally comparing the network's performance with the results of MEM and KK. The synthetic dataset was simulated using equation 17 (refer to the appendix for detailed explanation), and the total sample size was 4000. First, the simulated data was split into $^2/_3$ training and $^1/_3$ testing sets. The LSTM network approximates a function that maps the squared modulus to the imaginary component of the simulated training data by tuning its hyperparameters. Only one layer of LSTM is used, and after only 2 hours, an optimal network is trained. The trained network is used further to predict the artificial testing set and the experimental data. In Figure 8 a), the workflow of the deep learning network is illustrated. In b) of the same figure, a test CARS spectrum where the Raman resonances were created in both the strong and weak NRB regions is predicted by the network. As shown, the LSTM perfectly constructed the peak-like shape for the four resonances. Additionally, the peaks position and width perfectly reflect the theoretical ones in both NRB regions. Moreover, the LSTM network was able to remove the NRB completely.

Afterward, the real and imaginary components of the same spectrum were constructed by MEM and KK methods. In a) of Figure 9, the workflow of the algorithmic methods is illustrated. For MEM, a trial and error process was used to determine the optimal number of poles $M = 150$. In addition, a pre-processing step was necessary to remove the oscillations at the edges, which was done by replicating two small sub-regions on both edges and then removing them after the reconstruction of the spectrum.

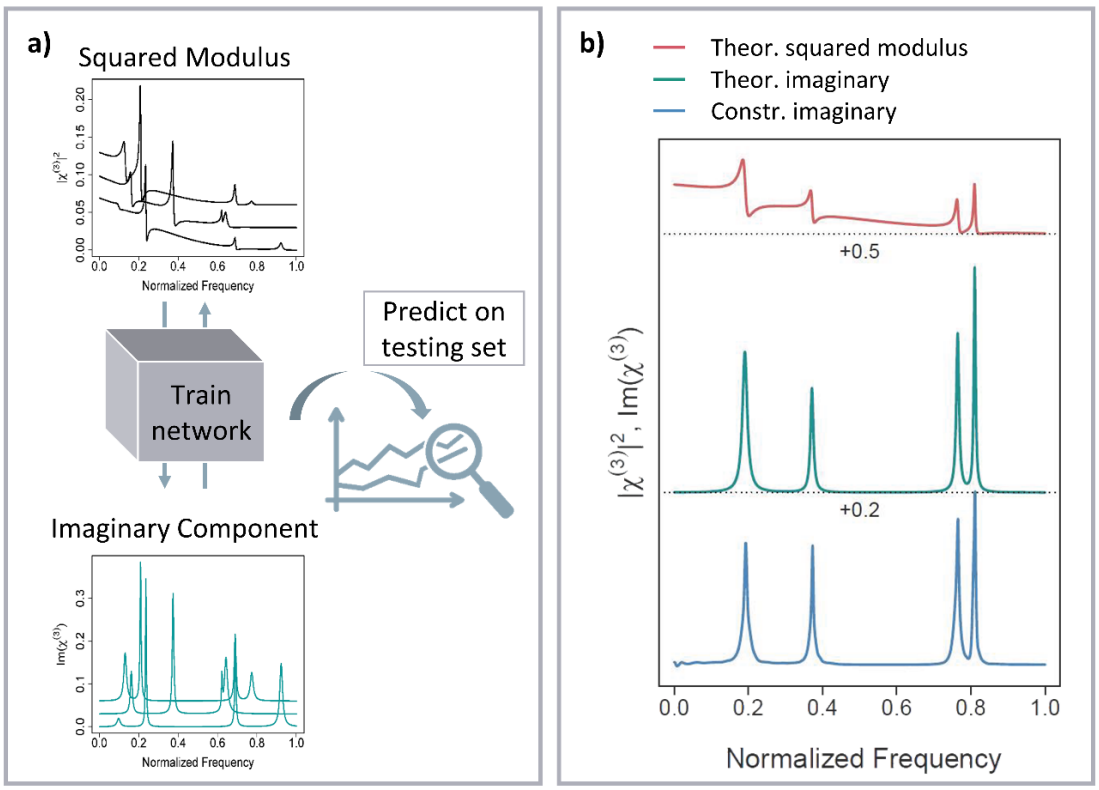


**Figure 8. The workflow and the performance of LSTM on one simulated CARS spectrum. In a), the LSTM network is first trained using a training dataset, then used to predict the testing set. The LSTM prediction on one CARS spectrum is shown in b) and compared with the theoretical component.**

As illustrated in b) of Figure 9, the MEM method extracted the imaginary and real components well, and the constructed squared modulus fits the theoretical one. In addition, the peaks shape, position, and width were estimated correctly in the imaginary part. Additionally, the constructed real component has a dispersive line shape and resembles the real theoretical component. Although the squared modulus was well reconstructed with the KK method, the constructed imaginary component shows a peak-like shape in the first two resonances and a dispersive line shape in the weak NRB region. The KK reconstruction is displayed in c) of Figure 9. Moreover, the constructed imaginary and real components are still distorted by a background contribution.

In conclusion, MEM and KK show an increased sensitivity regarding the NRB strength, which was reflected by the peaks shape flip in the imaginary and the real components or the background presence after the application of MEM and KK. However, the LSTM network predicted well the imaginary component and removed the background completely.

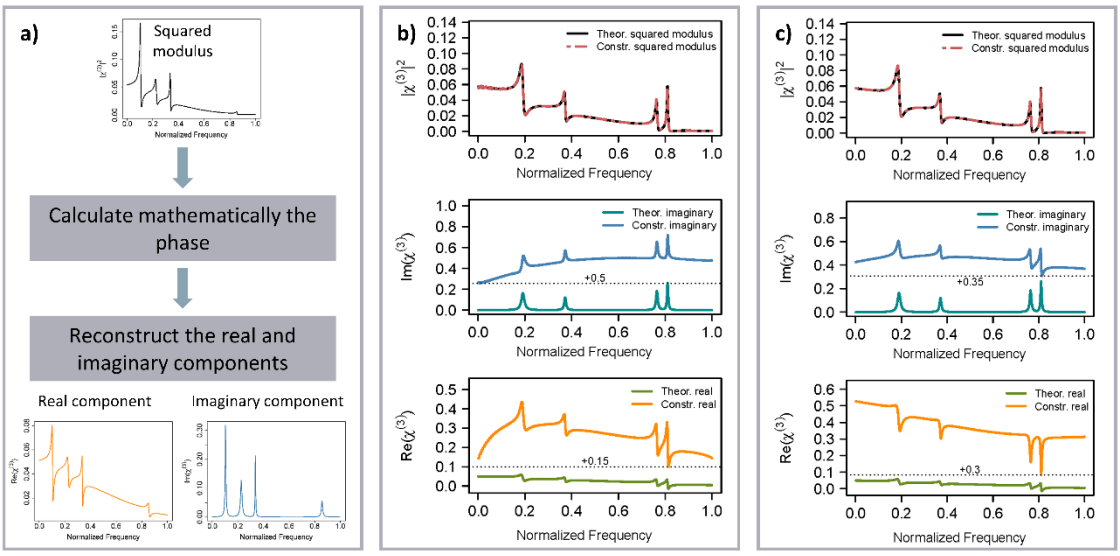


**Figure 9. The workflow and the performance of MEM and KK on a simulated CARS spectrum. The workflow of MEM and KK methods are illustrated in a). These methods are implemented on each spectrum where the phase is calculated, and then the imaginary as well as the real components are reconstructed. The performance of MEM and KK on one spectrum is visualized in b) and c), respectively. The background was not completely removed, and KK reconstruction showed a dispersive shape in the constructed imaginary components.**

The trained LSTM network was then validated on an experimentally measured broadband CARS spectrum of Acetonitrile and was compared to the MEM and KK reconstructions. The outputs of MEM and KK are illustrated in a) and b) of Figure 10, respectively. The reconstructed Acetonitrile spectrum fits the BCARS spectrum perfectly for both methods. However, the imaginary and real components were distorted, particularly in the region where the NRB is weak, visible by a dispersive line shape in the imaginary reconstruction. On the other hand, the LSTM prediction showed a good performance regarding the peaks position, amplitude, and width, illustrated in c) of Figure 10. In addition, the LSTM reconstruction did not need additional pre-processing.

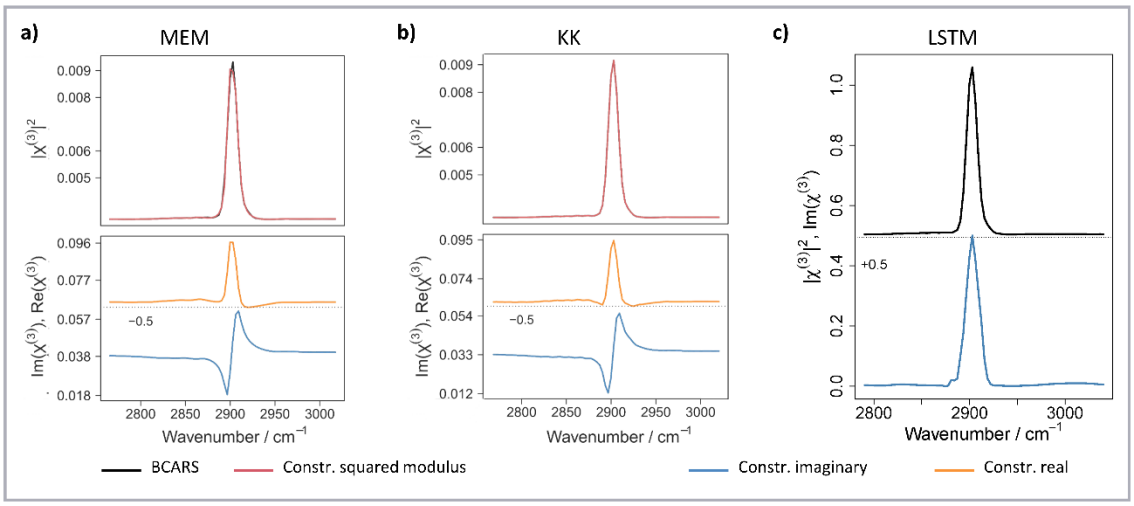


**Figure 10. The MEM, KK, and LSTM reconstructions of an experimental BCARS spectrum of Acetonitrile. The MEM and KK reconstructions of the real and imaginary components of the Acetonitrile spectrum are displayed in a) and b). Both methods showed a dispersive shape in the constructed imaginary component. However, the constructed imaginary components using the trained LSTM network on the same spectrum, displayed in c), showed a Lorentzian shape.**

In summary, the MEM and KK methods indicated an increased sensitivity regarding the strength of the NRB. Moreover, their reconstructions showed the flip of the peaks shape in the imaginary and the real components, and the background was still present after the retrieval. However, the LSTM network showed its potential since its results were not dependent on the strength of the non-resonant background. In conclusion, we used the deep learning technique for phase retrieval for the first time. As a result, the deep learning technique overcame MEM and KK regarding the peak shape and the removal of additional background contributions.

## 3.3 Comparison of denoising methods for multimodal images

Multimodal (MM) imaging is a relatively novel approach that combines two or more imaging methods and aims to acquire as much information as possible from the measured sample, e.g., tissue samples. For instance, the combination of the three nonlinear optical techniques, CARS, TPEF, and SHG, can acquire different information of molecules in the sample by different contrast mechanisms. In addition, the MM imaging approach provides high-quality (HQ) images, but the measurement of HQ images requires a significantly longer acquisition time compared to images of lower quality. Such faster MM imaging compromises the quality of the images due to the increase in the noise level, which will affect the identification of diseases or tissue abnormalities. Therefore, image denoising, a fundamental pre-processing technique, is crucial for medical applications since removing noise from images might also remove relevant information from the images. Therefore, choosing a suitable denoising method is necessary for an efficient and effective diagnosis. In P3, a classical phase retrieval method via GS was compared to two deep learning methods. These deep learning methods include a pre-trained network, namely DnCNN, and an own developed network with simple architecture derived from the super-resolution convolution neural network (SRCNN), referred to as incSRCNN.

The analysis was performed on MM images consisting of CARS, TPEF, and SHG modalities. It was split into two sections; first, artificial low-quality (LQ) images are created from HQ images by adding Poisson noise. The three methods, the GS algorithm, the pre-trained DnCNN, and the trained incSRCNN networks, were evaluated on these artificial LQ images. Then, the performance of these three methods was examined on experimental LQ images. In a) of Figure 11, an example of multimodal images including CARS, TPEF, and SHG modalities is shown. In the same figure, the workflow of the deep learning methods, where incSRCNN is trained with the CARS modality of the MM images (input, output) as (artificial LQ, experimental HQ), and the DnCNN is used as a transfer learning tool is illustrated in b). The parameters of the three methods can be found in the appendix.

The reconstruction of the artificial LQ image using the GS algorithm, the DnCNN, and the incSRCNN network is displayed in Figure 12. As shown, the GS reconstruction and three regions of interest (ROIs) preserve the structure of the image but include dark regions resulting from the source beam estimation. Furthermore, the overall similarity between the HQ and reconstructed images is significantly low. However, an improvement in the peak signal-to-noise ratio (PSNR) from 16.422 to 19.330 is deduced.

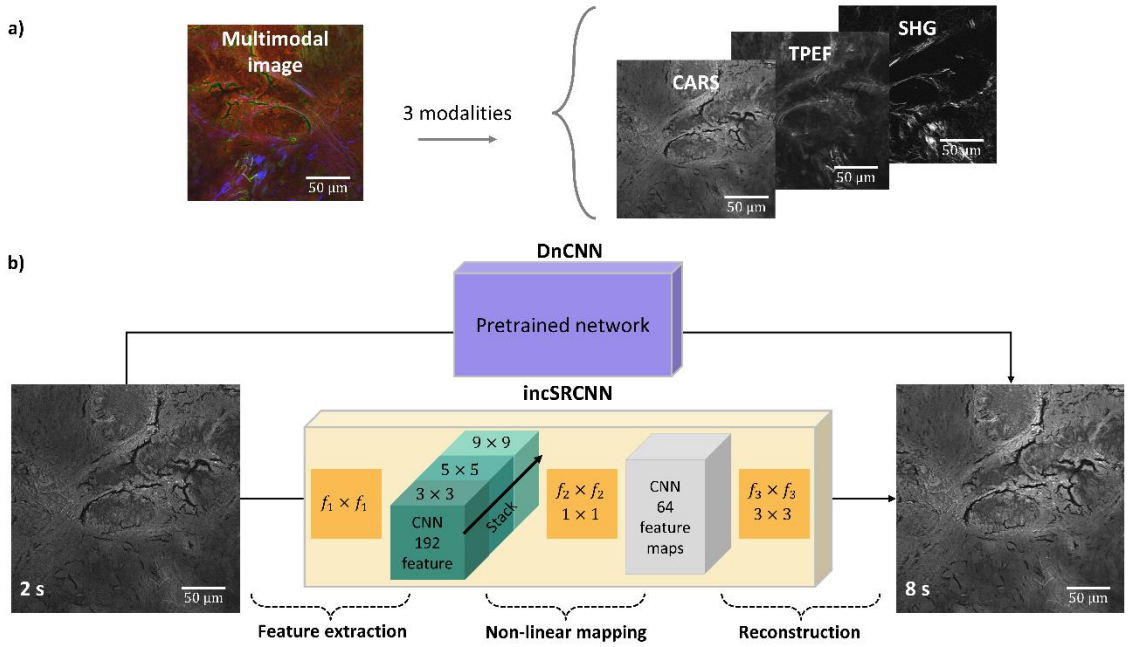


**Figure 11. Deep learning in image denoising for MM images. In a), an example of a multimodal image consisting of CARS, TPEF, and SHG modalities is shown. In b), two networks are tested separately on each modality. The incSRCNN was built using only the CARS modality and then predicting the other two modalities. However, a pre-trained network via DnCNN was used directly to predict each modality.**

In DnCNN reconstruction, the spatial structures and the color in the image are preserved, and the noise level is reduced. Consequently, the overall PSNR and SSIM were improved from 16.4 and 0.24 to 20.2 and 0.63, respectively. However, smoothed structures were shown in the three ROIs, which is critical for biomedical applications since some important information might be compromised and lost. Finally, the reconstructed image's spatial structures and color are preserved via the incSRCNN network, reducing the noise level. Consequently, the overall PSNR and SSIM were improved from 16.4 and 0.24 to 22 and 0.53, respectively. Although the incSRCNN reconstruction did not show smoothed regions, some dark areas were produced, affecting the PSNR and SSIM values.

The next step is to evaluate the three methods on experimental LQ images. Figure 13 showed that the GS reconstructed image preserves the spatial structures; however, it includes dark regions similar to the artificial case. As a result, the overall PSNR and SSIM decrease from 19.7 and 0.55 to 18 and 0.49, respectively.

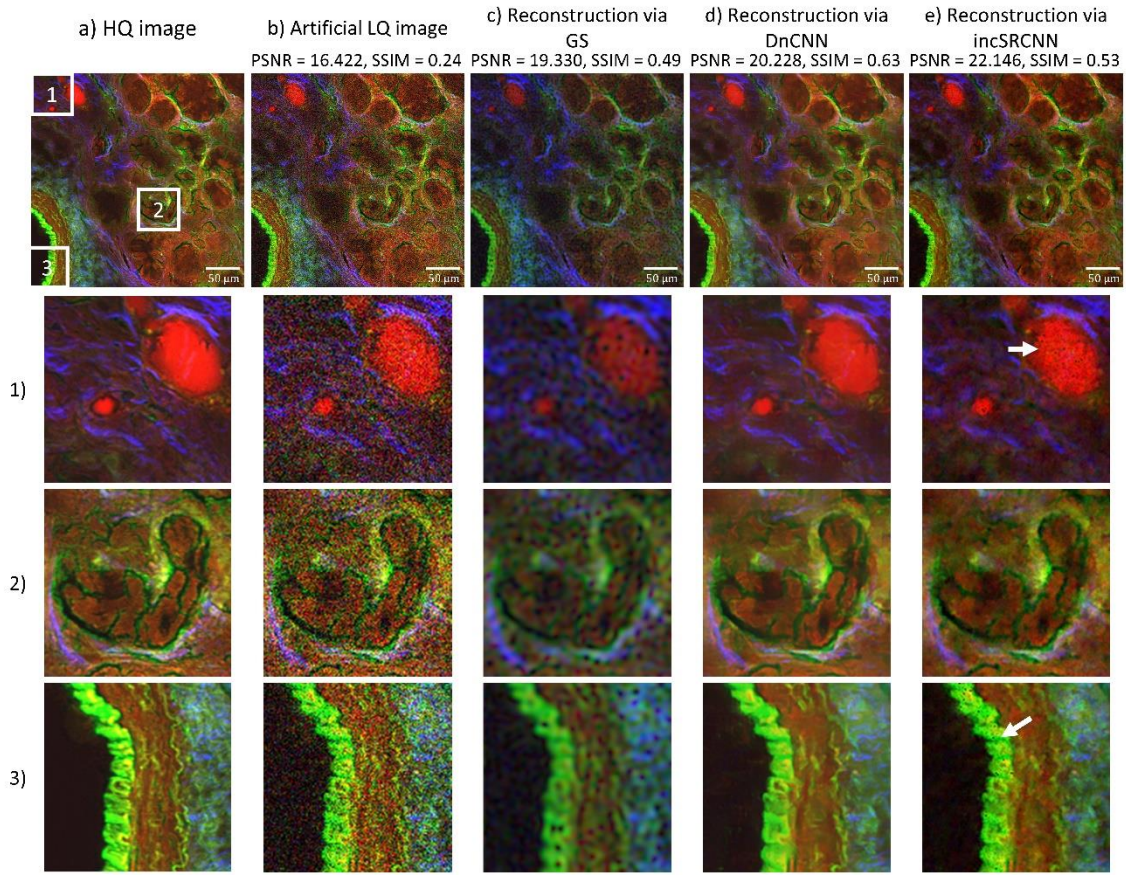


**Figure 12. Comparison of image denoising methods applied on artificial LQ images. The GS reconstruction shows dark regions resulting from the source beam estimation. Although the DnCNN and incSRCNN perform better than the GS algorithm, the DnCNN reconstruction shows a smoothed region absent in the incSRCNN case. However, the arrows in the incSRCNN reconstruction represent the inability of the network to estimate the correct values.**

In d) of the same figure, the spatial structure in the DnCNN reconstruction and three ROIs were preserved, and the noise level was slightly reduced. Moreover, the overall PSNR and SSIM were improved from 19.7 and 0.55 to 20.2 and 0.64, respectively. Similar to the artificial case, smoothed regions were produced, which might cause the removal of essential features that were highly sensitive in diagnosing diseases or abnormalities. Finally, the image's spatial structures and color were preserved via the incSRCNN network. Compared to the experiment LQ image, the overall PSNR and SSIM values slightly decreased. However, the incSRCNN network failed to estimate some regions, refer to arrows in the figure, which might cause the decrease in the PSRN and SSIM values.

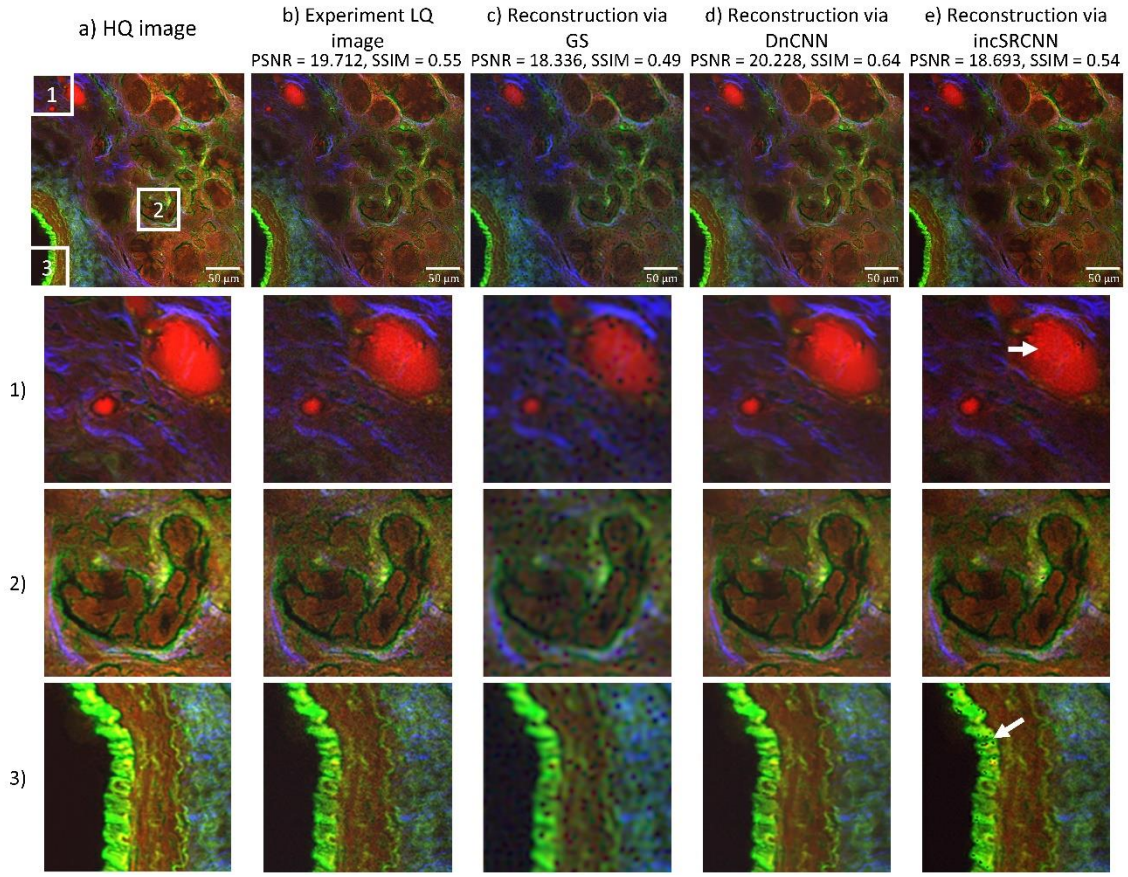


**Figure 13. Comparison of image denoising methods applied on experimental LQ images. The GS reconstruction shows dark regions resulting from the source beam estimation. Although the DnCNN and incSRCNN perform better than the GS algorithm, the DnCNN reconstruction shows a smoothed region absent in the incSRCNN case. However, the arrows in the incSRCNN reconstruction represent the inability of the network to estimate the correct values.**

To summarize, the GS reconstruction showed similar but poor performance for both artificial and experimental LQ images. It included dark regions, and the algorithm showed limited abilities even in noiseless settings. In addition, it seems that the algorithm converges to a local minimum that causes poor reconstructions. However, the DnCNN and the incSRCNN reconstructions preserved the colors and detailed structures for artificial and experimental LQ images. Both networks performed well in the artificial LQ case, but the DnCNN produced smoothed regions critical for medical applications. In the experimental case, the DnCNN network also produced smoothed region, which is a drawback compared to our proposed network that showed a slight reduction in the PSNR value due to the lack of data that the network could not train some regions.

## 3.4 A novel approach for analyzing Raman spectra

Chemometric methods can be split into discrete and functional methods. The discrete group, also called multivariate data analysis, considers the data as a set of independent points acquired on a specific interval. In contrast, data is modeled as a function or a curve in the functional group. However, in almost all of the literature, Raman data is always discretely analyzed. Nevertheless, this approach has a lot of room for improvement. On the one hand, the acquisition of Raman data requires a significant amount of time, creating some limitations in the spread of its applications. On the other hand, for a faster measurement of Raman data, a compromise in the spectral quality is inevitable. Therefore, we analyzed the Raman spectral data in the functional framework for the first time to allow the analysis of low-quality Raman spectra.

The classical PCA followed by LDA (PCA-LDA) in the discrete case was compared to the functional PCA followed by LDA (FPCA-LDA) to evaluate the performance of the functional data analysis on Raman spectra. This comparison was made on simulated and experimental Raman data. A detailed explanation of the simulated and experimental Raman data, the functional data analysis approximation, and the functional PCA (FPCA) can be found in the appendix.

First, the results of the functional approximation for all simulated cases are illustrated in Figure 14. In this figure, one spectrum per class for each SNR and a specific peak shift $\Delta\tilde{\nu} = 0.01$ are shown. In a) of the same figure, the original normal and abnormal spectra are plotted for all SNR values in each row. In b), the functional approximations for each class are illustrated for the nine cases of SNR. The number of basis functions is calculated based on the elbow method. It first calculates the root-mean-square error, and the optimal point is chosen, which refers to the largest distance to the line that joins the first and last values. Therefore, 190 B-spline basis functions in the functional approximation in all the simulation cases are used. In Figure 14, the functional approximation perfectly represents the original spectra with almost noiseless reconstruction for an SNR larger than 5. Although the functional approximation also fits noise for the lower SNR values due to a high number of the basis used, noise is significantly reduced in the approximation. In addition, it potentially maintains the peak shape in the case of lower SNR values. Then, the classification methods via PCA-LDA and FPCA-LDA were compared on the simulated Raman spectra.

For both methods, 10-fold cross-validation was used, and the number of components chosen was 50. The mean sensitivity was used as an evaluation metric, illustrated in a heat map in c)

Figure 14. In region c of Figure 14, PCA-LDA and FPCA-LDA perform perfectly with 100% mean sensitivity due to the clear distinction between the normal and abnormal classes. While in regions a and b, the FPCA-LDA performs better in most cases since the obtained functions contain less noise, and an improvement of the shape of the peaks in these functions was also detected. However, in a few cases, the PCA-LDA provides better mean sensitivity. The performance difference between FPCA-LDA and PCA-LDA is only significant in region b. The Kruskal-Wallis test calculated the significance of the performance difference and concluded that only region b showed a significant difference.

The next step is to check the performance of the functional data analysis on both raw and pre-processed experimental Raman spectra (details of experiment Raman can be found in the appendix).

In Figure 15, the PCA-LDA and FPCA-LDA methods were applied on pre-processed experimental data. The total number of spectra was 1131 spectra for the pre-processed datasets. Two cross-validation methods were tested, the leave one batch out cross-validation (LOBOCV) and the 10-fold cross-validation (10-fold CV). In the FPCA-LDA, the first step was to approximate the experimental Raman data into functions. Therefore, the elbow method was implemented, and the optimal number of basis functions was suggested equal to 80. However, we increased this number to 200 since we wanted the functional approximation to include the C-D/C-H region with high quality.

In the first row of Figure 15, the mean spectra per label are plotted on the left, and its functional version is shown on the right of the same figure. A slight reduction of noise in the functional version is deduced since the SNR increased from 302 to 307.14. Additionally, the Raman spectral features were conserved. The SNR is approximated by calculating the peak amplitude ratio at the C-H band (2930 $cm^{-1}$) to the standard deviation of the region between 2408 and 2578 $cm^{-1}$.

Moreover, the comparison between the PCA-LDA and FPCA-LDA using the LOBOCV and 10-fold CV are illustrated in the second and third row of Figure 15, respectively. The mean sensitivities are shown in panels a and b. Similar performance for both methods can be shown with a slight improvement in the FPCA-LDA method values. In addition, a reduction of the standard deviation is visible in the case of the FPCA-LDA method.

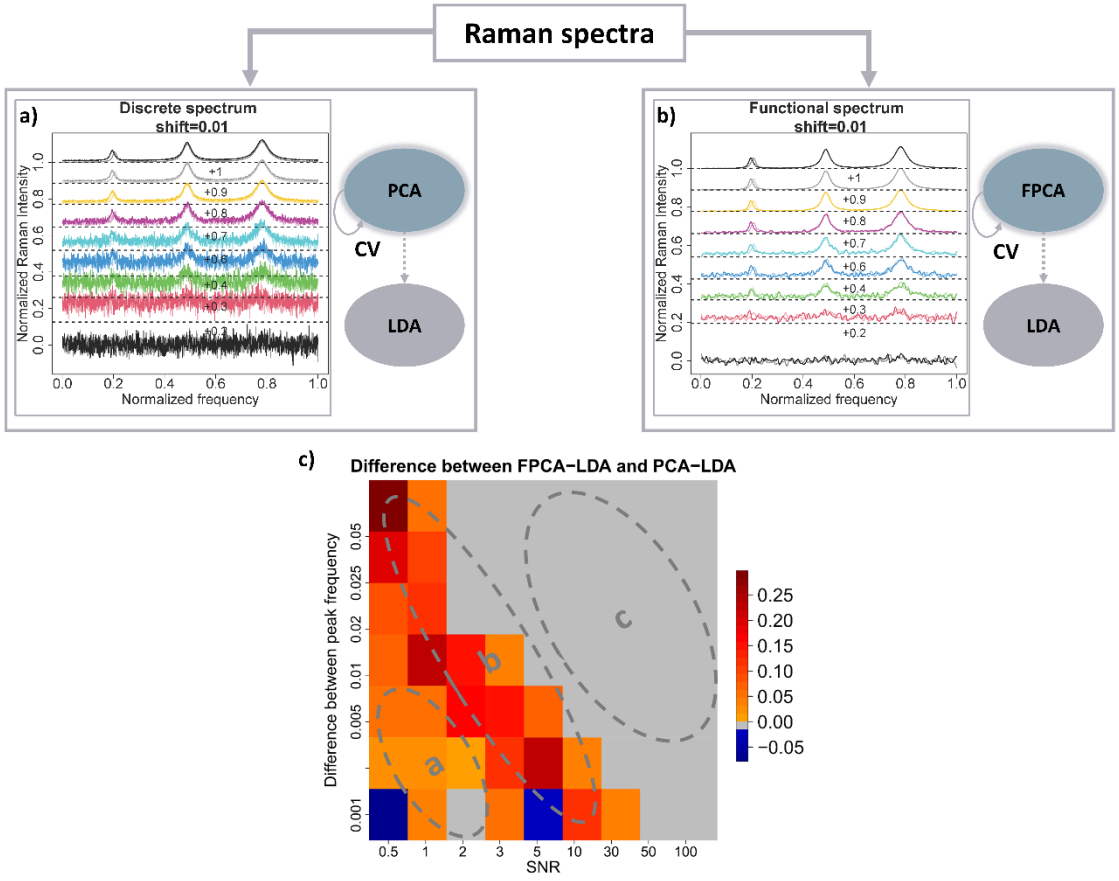


**Figure 14. The discrete and functional analysis of simulated Raman spectra. The discrete framework where simulated Raman spectra per SNR are shown in a). In b), the functional version for the same simulated Raman spectra is illustrated. Finally, the difference between the mean sensitivities of FPCA-LDA and PCA-LDA is represented in c).**

The maximum mean sensitivity for the PCA-LDA method refers to the model with 45 components with a mean sensitivity of $0.77 \pm 0.16$ in the LOBOCV case and 45 components with a value of $0.91 \pm 0.07$ in the 10-fold CV case. However, the maximum sensitivity for the FPCA-LDA method refers to a model with 50 components with a value of $0.79 \pm 0.14$ in the LOBOCV case and 47 components with a value of $0.91 \pm 0.06$ in the 10-fold CV case. The corresponding confusion matrices for the model with the highest mean sensitivities are illustrated in black and red for PCA-LDA and FPCA-LDA, respectively, in panels c and d for LOBOCV and 10-fold CV, respectively.

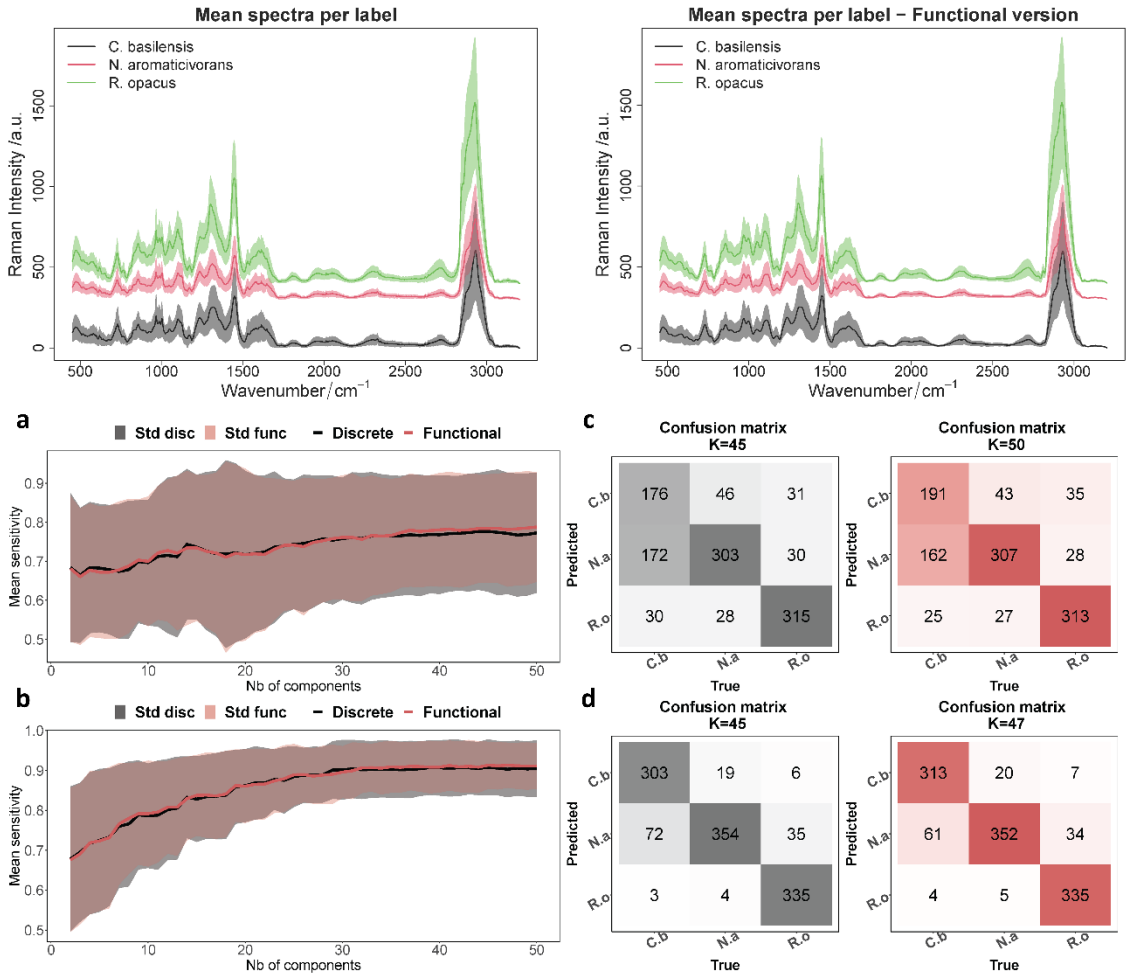


**Figure 15. The pre-processed experimental Raman data and its functional version, PCA-LDA and FPCA-LDA comparison. The mean spectra per label and its functional version are shown on the top row. The bottom row compares the mean sensitivities and confusion matrices of PCA-LDA and FPCA-LDA for LOBOCV, and a 10-Fold CV is illustrated.**

Even though functional data analysis resulted in smoothed function, this did not affect the classification output using both LOBOCV and 10-fold CV. Therefore, the functional approach preserved the important features needed for the classification. PCA-LDA and FPCA-LDA were also applied to the raw data, and similar performance was deduced.

In conclusion, functional data analysis can be considered a promising tool for analyzing Raman spectra, especially when the quality of the data is low. This property makes functional data analysis a great tool to analyze spectra acquired fast or in vivo, which yield low-quality Raman spectra. Furthermore, functional data analysis can be used with spectra with a different spectral resolution.

# 4 SUMMARY

The innovation in biophotonic technologies showed great potential and promise for the healthcare system's challenges in the 21st century. These advancements in light-based techniques allow studying biological samples at the cellular, tissue, and organ levels. To allow the translation of biophotonic technologies in the healthcare system, automated systems to support physicians in their decision-making are needed. That results from the fact that the datasets generated by these biophotonic techniques are massive, and the contained information cannot easily or directly be extracted. To solve this task, analytical methods such as chemometrics, machine learning, and deep learning are the cornerstone for fast information extraction. The first part of this thesis focuses on reviewing the recent trends in chemometrics and machine learning applied to chemical data. In this work, the main part is the division of the chemometric, machine learning, and deep learning methods into inverse modeling, pre-processing, and data modeling techniques. The inverse modeling, pre-processing, and data modeling groups aim to extract the physical system's parameters, acquire a clean version of distorted data, and train a model that identifies abnormalities. The review showed a significant focus in the literature on generating automated systems for faster processing and analysis in addition to the booming of deep learning applications for different spectroscopic data types. Afterward, I explored three research questions and argued that my contribution could be considered a step towards improving detection methods for the healthcare systems.

In the inverse modeling group, the first research question was motivated by the fact that Raman measurements can be performed non-invasively and provide a fingerprint signature of the molecule composition. A drawback is that these Raman measurements are slow. Hence, the CARS technique can be considered an alternative since faster measurements are possible, and an improved signal is obtained, which is crucial for biomedical applications. However, CARS measurements are distorted by the presence of NRB. Therefore, algorithmic methods have been applied to extract the Raman-like spectra from the CARS spectra. Although these methods can extract the Raman-like shape from the CARS spectra, they require either a priori knowledge about the system or a post-processing technique since they cannot completely remove the background. Therefore, different versions of algorithms were developed to tackle this issue,

which needs a priori knowledge. Can data-driven methods like deep learning remove the NRB without a priori knowledge and further data processing?

To answer our question, simulation and experimental data were both needed. Therefore, simulated data was first constructed where the squared modulus and the imaginary components are the CARS and Raman-like spectra, respectively. The synthetic dataset was split into training and testing sets. The deep learning method via the LSTM network was then trained using simulated training data by tuning its hyperparameters. Finally, two popular methods, MEM and KK, were applied to compare with the LSTM predictions. MEM and KK reconstructions showed some drawbacks; a background contribution still existed after the reconstruction, and a dispersive shape was created in the reconstructed imaginary component. These drawbacks confirmed that MEM and KK methods are very sensitive to the strength of the NRB and that both methods require an estimation of an error phase spectrum that involves a priori knowledge. In contrast, the deep learning approach successfully reconstructed the Raman-like shape where the background was removed entirely. The next step was to compare these findings in the case of experimental measured broadband CARS spectra. Therefore, the three methods were applied to the Acetonitrile BCARS spectrum. Although, the reconstructed spectrum perfectly fits the BCARS Acetonitrile spectrum for the MEM and KK methods, a clear distortion of the peak shape in MEM and KK imaginary reconstructions is present. On the other hand, LSTM prediction showed a good performance regarding the peaks amplitude and width, and no additional processing is needed compared to both reconstructions via MEM and KK methods. In conclusion, the LSTM overcame both standard techniques used for phase retrieval in terms of the non-dependency on the NRB strength and that no post-processing to remove the background completely is needed. Lastly, the LSTM prediction time is shorter than the MEM and KK methods.

In the pre-processing section, the image denoising problem is a fundamental task that is crucial for medical applications since removing noise from images might remove relevant information. To address this, we used multimodal images, which offer a lot of information about the complex structures in tissues and the chemical composition. The MM imaging provides spatially high-quality images but requires a significantly long acquisition process. However, faster MM imaging causes an increase in the noise level, which will reduce the image quality and compromise the identification of diseases or abnormalities. Therefore, a trade-off between fast imaging and a suitable denoising method is necessary for an efficient and effective diagnosis. Since deep learning showed success stories in computer vision, two DL-based

methods via DnCNN, a pre-trained network, and incSRCNN, a trained network with architecture inspired from SRCNN, were used and compared with a traditional phase retrieval method via the GS algorithm. The analysis was held on two types of images: simulated LQ images and experimental LQ images, and the aim is to reconstruct high-quality images from these noisy ones. These MM images involve CARS, TPEF, and SHG modalities. First, the three methods were applied on simulated LQ images constructed from the HQ images by adding Poisson noise and then on experimental LQ images. The GS algorithm was used independently on each modality for simulated and experimental LQ images. The reconstructed GS images showed bad performance since dark regions were produced. It seems that the algorithm was stuck in some local minimum and could not learn more. However, both the deep learning approaches perform better. First, the pre-trained network DnCNN was implemented as a transfer learning tool that was originally trained on natural images to fix corrupted images with noise and artifacts. Similarly, the DnCNN was employed separately on each modality. The DnCNN reconstructions on both the simulated and experimental LQ images at first glance were good and represented well the HQ images, particularly when comparing the PSNR and the SSIM, which generated the higher values. However, smoothed regions are produced, which means that some complex structures in the MM images were lost, which is crucial for detecting diseases or abnormalities. Finally, a simple CNN-based architecture, namely incSRCNN, was trained using the available small-sized MM images. The incSRCNN network was trained only using the CARS modality and then predicted the other TPEF and SHG modalities. Surprisingly, this network performs well, particularly for the artificial LQ images, since the complex structures were preserved with no smoothing regions. However, in the case of the experimental LQ images, the incSRCNN reconstruction did not improve, and the lack of data might be the reason. In this regard, worse experimental LQ images were constructed, and high-quality images were predicted using the incSRCNN reconstruction. The incSRCNN reconstruction showed better performance than the synthetic experiment LQ images. In summary, the deep learning networks showed promising results. Particularly, our proposed network consisted of simple architecture and was able to reconstruct the complex structures of the MM images. However, some artifacts were produced, which might result from the small data size used to train the network.

Finally, a different question was researched in the data modeling section focusing on the low-quality Raman spectra that are produced when the acquision time is reduced. However, this compromises the quality of the data. Therefore, the last part of the thesis investigates the search

for a method that performs better than standard classification tasks if low-quality Raman data is acquired. Generally, mathematical functions best describe Raman spectral data; however, only discrete points are measured due to the spectroscopic measurement setup. Therefore, we investigated the Raman spectral data in the functional framework for the first time. The first step is to approximate the Raman spectra using B-spline basis functions. Afterward, the functional PCA followed by LDA was applied, and results were compared to the standard PCA-LDA method. Both classification methods were performed on simulated and experimental Raman data. The simulated Raman spectra were divided into two classes (normal and abnormal group) with three peaks. The abnormal class was generated by slightly shifting one of the peaks from the peak position used in the normal group. 63 combination of different SNR values and peak shift were studied. Although the functional approximation also fits noise for the lower SNR values due to a high number of basis functions used, a significant reduction of noise is noticed in the approximation. It also potentially maintains the peak shape in the case of low SNR values. The functional PCA-LDA and PCA-LDA applied on the simulated Raman data perform statistically similarly in the following two cases. The first case refers to when the quality of the spectra is low (low SNR) in combination with a slight peak shift that both models do not work, and the second case is when the spectral quality is so good that both methods perform perfectly. However, when the SNR and the shift in the peak position values are inversely proportional, the functional PCA-LDA performs better. Then, we evaluated both approaches on experimental Raman spectra. A slight improvement in the classification performance reflects the finding on the simulated data since the experimental data is in accordance with the simulated data in terms of having high SNR. In conclusion, functional data analysis can be considered a promising tool for analyzing Raman spectra, especially when the quality of the data is low. This property makes the FDA a great tool to analyze spectra acquired fast or in vivo, which both yield low-quality Raman spectra.

# 5 ZUSAMMENFASSUNG

Die Innovationen im Bereich der biophotonischen Technologien haben deren großes Potenzial zum Adressieren von Herausforderungen des Gesundheitssystems im 21. Jahrhundert gezeigt. Durch die Fortschritte bei lichtbasierten Techniken können biologische Proben auf Zell-, Gewebe- und Organebene untersucht werden, was wiederum zur Verbesserung des Gesundheitssystems eingesetzt werden kann. Um dies zu erreichen muss die Biophotonik vermehrt in klinischen Studien eingesetzt werden, es müssen automatisierte Systeme aufgebaut werden und es müssen robusten Lösungen zur Unterstützung der Entscheidungsfindung des Arztes gefunden werden. Die mit diesen biophotonischen Techniken erzeugten Datensätze sind jedoch sehr umfangreich, und die in diesen Datensätzen enthaltenen Informationen lassen sich nicht einfach oder direkt extrahieren. Daher sind Analysemethoden wie Chemometrie, maschinelles Lernen und Deep Learning der Grundstein für eine schnelle Anwendung biophotonischer Messmethoden in klinischen Studien und die Entwicklung automatisierter Detektionssysteme. Darüber hinaus spielen diese Methoden eine entscheidende Rolle bei der Extraktion von Informationen aus komplexen Signalen und der Erkennung unerwarteter Ereignisse. Daher wurde in dieser Arbeit zunächst ein Überblick über die jüngsten Trends in der Chemometrie und im maschinellen Lernen bei der Anwendung auf chemische Daten gegeben. Im Hauptteil dieser Arbeit werden die chemometrischen, maschinellen Lern- und Deep-Learning-Methoden in die Gruppen inverse Modellierung, Vorverarbeitung und Datenmodellierung aufgeteilt. Die Gruppen "Inverse Modellierung", "Vorverarbeitung" und "Datenmodellierung" zielen darauf ab, die Parameter des physikalischen Systems zu invertieren, eine gesäuberte Version der Messdaten zu erhalten und ein Modell zu trainieren, das Änderungen in der Probe bzw. Gruppen identifiziert. Die Übersicht hat gezeigt, dass in der Literatur ein deutlicher Schwerpunkt auf der Entwicklung automatisierter Systeme für eine schnellere Verarbeitung und Analyse liegt, zusätzlich zum Boom der Deep-Learning-Anwendungen für verschiedene spektroskopische Methoden. Anschließend wurden drei Forschungsfragen in den oben genannten Gruppen untersucht und argumentiert, dass die Lösungen jeweils einen Beitrag zur Verbesserung der Gesundheitssysteme leisten können.

In der Gruppe „inverse Modellierung“ wurde die erste Forschungsfrage durch die Tatsache motiviert, dass Raman-Messungen eine nicht-invasive Messung darstellt, die einen Fingerabdruck der Probenzusammensetzung liefert. Jedoch ist diese Messung langsam, da der Raman-Effekt ein schwacher Prozess ist. In dieser Hinsicht kann die kohärente anti-Stokes Raman-Streuung (CARS)-Technik als nützliche Alternative betrachtet werden, da sie schneller Raman-Informationen und ein besseres Signal liefert, was für biomedizinische Anwendungen und klinische Versuche von entscheidender Bedeutung ist. Allerdings werden CARS-Messungen durch das Vorhandensein des nicht-resonanten Untergrunds (NRB) verzerrt. Aus diesem Grund wurden algorithmische Methoden angewandt, mittels derer die Raman-ähnlichen Spektren aus den CARS-Spektren extrahiert werden. Obwohl mit diesen Methoden die Raman-ähnliche Form aus den CARS-Spektren extrahiert werden kann, erfordern sie entweder A-priori-Wissen über das System oder eine Nachbearbeitungstechnik, da mit ihnen der Hintergrund nicht vollständig entfernt werden kann. Um dieses Problem zu lösen wurden verschiedene NRB-Korrekturen, wie zum Beispiel verschiedene Versionen von MEM und KK entwickelt. Doch ist es möglich, andere Methoden zu finden, die kein Vorwissen erfordern und den NRB vollständig entfernen?

Möglicherweise sind datengesteuerte Methoden die Lösung. Deshalb haben wir uns die Entwicklungen im Bereich des Deep Learning zunutze gemacht und diese Technik zur Beantwortung dieser Forschungsfrage eingesetzt. Zur Beantwortung eben dieser Frage wurden sowohl Simulations- als auch experimentelle Daten Genutzt. Zunächst wurden Simulationsdaten erstellt, bei denen das Betragsquadrat und die imaginäre Komponente die CARS- bzw. Raman-ähnlichen Spektren darstellen. Das synthetische Datensatz wurde in einen Trainings- und einen Testsatz aufgeteilt. Das Deep-Learning-Netz über das LSTM-Netz wurde dann anhand von Trainings- und Validierungssätzen trainiert, und ein Testsatz wurde für die Modell-Bewertung verwendet. Schließlich wurden zwei Standard-Methoden zur NRB-Korrektur, MEM und KK, angewendet, um sie mit der LSTM-Vorhersage zu vergleichen. Die MEM- und KK-Rekonstruktionen wiesen einige Nachteile auf: nach der Rekonstruktion ist noch ein Hintergrundbeitrag vorhanden. Darüber hinaus wurde bei Anwendung des KK-Algorithmus eine dispersive Form erzeugt. Diese Nachteile bestätigten, dass die MEM- und KK-Methoden sehr empfindlich auf die Stärke der NRB reagieren und dass beide Methoden eine Schätzung des Fehlerphasenspektrums erfordern, welches A-priori-Wissen voraussetzt. Im Gegensatz dazu rekonstruierte der Deep-Learning-Ansatz erfolgreich die Raman-ähnliche Form der Peaks, während der Hintergrund vollständig entfernt wurde. In einem nächsten

Schritt wurden diese Ergebnisse mit experimentell gemessenen breitbandigen CARS-Spektren verglichen. Daher wurden die drei Methoden auf das Acetonitril-BCARS-Spektrum angewandet. Für die MEM- und KK-Methode passte das rekonstruierete Spektrum perfekt zum gemessenen BCARS-Acetonitril-Spektrum. Bei den imaginären MEM- und KK-Rekonstruktionen ist jedoch eine deutliche Verzerrung der Peakform zu beobachten. Die LSTM-Vorhersage zeigte jedoch eine gute Leistung in Bezug auf die Peakamplituden und -breiten, und im Vergleich zu den beiden Rekonstruktionen mittels MEM- und KK-Methoden ist keine zusätzliche Datenverarbeitung erforderlich. Zusammenfassend lässt sich sagen, dass die LSTM-Methode den beiden für die NRB-Korrektur verwendeten Standardtechniken überlegen ist, da sie weniger von der NRB-Stärke abhängt und keine Nachbearbeitung zur vollständigen Entfernung des Untergrunds erfordert. Zusätzlich ist die LSTM-Vorhersagezeit kürzer als die Vorhersagezeit der MEM- und KK-Methode.

Im Abschnitt über die Vorverarbeitung ist die Bildentrauschung ein grundlegendes Vorverarbeitungsverfahren, das für medizinische Anwendungen von entscheidender Bedeutung ist, da durch die Entfernung von Rauschen aus Bildern wichtige Informationen verloren gehen können. Um dieses Problem zu lösen, haben wir multimodale Bilder verwendet, ein aussichtsreicher Ansatz, der mehr Informationen über die komplexen Strukturen im Gewebe und die chemische Zersetzung bietet. Diese MM-Bilder umfassen CARS-, TPEF- und SHG-Kanäle. Die MM-Bildgebung liefert qualitativ hochwertige Bilder, was aber einen wesentlich längeren Aufnahmeprozess benötigt. Eine schnellere MM-Bildgebung führt jedoch zu einem Anstieg des Rauschpegels, was die Bildqualität mindert und die Erkennung von Krankheiten oder Anomalien beeinträchtigt. Daher ist für eine effiziente und effektive Diagnose ein Kompromiss zwischen schneller Bildgebung und einer geeigneten Entrauschungsmethode erforderlich. Da Deep Learning im Bereich des Computer-Vision erfolgreich Anwendung findet, wurden zwei DL-basierte Methoden, das DnCNN, ein vortrainiertes Netzwerk, und incSRCNN, ein trainiertes Netzwerk mit einer von SRCNN inspirierten Architektur, verwendet und mit einer traditionellen Glättungsmethode über den GS-Algorithmus verglichen. Die Analyse wurde an zwei Arten von Bildern durchgeführt: an simulierten Bildern niedriger Qualität (LQ-Bildern) und an experimentellen LQ-Bildern, wobei das Ziel darin besteht, aus diesen verrauschten LQ-Bildern qualitativ hochwertige Bilder zu rekonstruieren. Zunächst wurden die drei Methoden auf simulierte LQ-Bilder, die aus den HQ-Bildern durch Hinzufügen von Poisson-Rauschen konstruiert wurden, und dann auf experimentelle LQ-Bilder angewendet. Der GS-Algorithmus wurde unabhängig für jeden

Kanal für simulierte und experimentelle LQ-Bilder verwendet. Die rekonstruierten GS-Bilder zeigten eine schlechte Qualität, da dunkle Bereiche erzeugt wurden. Es scheint, dass der Algorithmus in einem lokalen Minimum stecken geblieben ist und nicht weiter lernen konnte. Die beiden Deep-Learning-Ansätze schneiden jedoch besser ab. Zunächst wurde das vortrainierte Netzwerk DnCNN als Transfer-Learning-Tool implementiert, das ursprünglich auf natürlichen Bildern trainiert wurde, um beschädigte Bilder mit Rauschen und Artefakten zu korrigieren. In ähnlicher Weise wurde das DnCNN für jeden Kanal separat eingesetzt. Die DnCNN-Rekonstruktionen sowohl der simulierten als auch der experimentellen LQ-Bilder waren auf den ersten Blick gut und stellten die HQ-Bilder gut dar, insbesondere beim Vergleich des PSNR und des SSIM, welche hohe Werte ergaben. Es werden jedoch geglättete Regionen erzeugt, was bedeutet, dass einige komplexe Strukturen in den MM-Bildern verloren gingen, die für die Erkennung von Krankheiten oder Anomalien (möglicherweise) entscheidend sind. Schließlich wurde eine einfache CNN-basierte Architektur, nämlich incSRCNN, anhand der verfügbaren kleinen MM-Bilddatensatz trainiert. Das incSRCNN-Netzwerk wurde nur mit dem CARS-Kanal trainiert und sagte dann die anderen TPEF- und SHG-Kanäle voraus. Überraschenderweise schneidet dieses Netz gut ab, insbesondere bei den künstlichen LQ-Bildern, da die komplexen Strukturen ohne Glättungsbereiche erhalten blieben. Bei den experimentellen LQ-Bildern hat sich die incSRCNN-Rekonstruktion jedoch nicht verbessert, was auf den Mangel an Daten zurückzuführen sein könnte. In diesem Zusammenhang wurden schlechtere experimentelle LQ-Bilder konstruiert, und mit der incSRCNN-Rekonstruktion wurden qualitativ hochwertige Bilder vorhergesagt. Die incSRCNN-Rekonstruktion zeigte eine bessere Leistung als die synthetischen LQ-Bilder des Experiments.

Zum Schluss wurde im Abschnitt Datenmodellierung eine weitere Frage untersucht, die sich auf Raman-Spektren geringer Qualität konzentriert, die entshehen, wenn die Erfassungszeit reduziert wird. Dadurch wird jedoch die Qualität der Daten beeinträchtigt. Daher wird im letzten Teil der Arbeit nach einer Methode gesucht, die bei der Analyse von Raman-Daten geringer Qualität eine bessere Klassifikationsleistung erbringt als Standard-Klassifizierungsmethoden. Im Allgemeinen beschreiben mathematische Funktionen die Raman-Spektraldaten am besten, allerdings werden aufgrund des spektroskopischen Messaufbaus nur diskrete Punkte gemessen. Aus diesem Grund haben wir die Raman-Spektraldaten zum ersten Mal im Rahmen von Funktionen untersucht. In einem ersten Schritt werden die Raman-Spektren mit B-Spline-Basisfunktionen approximiert. Anschließend wurde die funktionale PCA, gefolgt von LDA, angewandt und die Ergebnisse mit der Standard-PCA-

LDA-Methode verglichen. Beide Klassifizierungsmethoden wurden an simulierten und experimentellen Raman-Daten durchgeführt. Die simulierten Raman-Spektren wurden in zwei Klassen (normale und abnormale Gruppe) mit drei Peaks unterteilt. Die abnormale Klasse wurde durch eine leichte Verschiebung eines der Peaks gegenüber der Peakposition in der normalen Gruppe erzeugt. Außerdem wurden zwei Szenarien betrachtet, jeweils mit und ohne zusätzlichen Untergrund. Es wurden 63 Kombinationen verschiedener SNR-Werte und Peakverschiebungen untersucht. Obwohl die funktionale Approximation in beiden Szenarien aufgrund der hohen Anzahl der verwendeten Basisfunktionen auch für die niedrigeren SNR-Werte mit Rauschen einhergeht, ist eine deutliche Verringerung des Rauschens bei der Approximation zu verzeichnen. Auch bei niedrigen SNR-Werten wird die Bandenform beibehalten. Die funktionale PCA-LDA und PCA-LDA, die auf die simulierten Raman-Daten angewandt werden, zeigen in den beiden folgenden Fällen statistisch ähnliche Ergebnisse. Der erste Fall bezieht sich auf eine niedrige Qualität der Spektren (niedriges SNR) in Kombination mit einer kleinen Peakverschiebung, so dass beide Modelle nicht gut funktionieren: Der zweite Fall ist durch eine hohe spektrale Qualität gekennzeichnet, sodass beide Methoden perfekt funktionieren. Wenn jedoch das SNR und die Verschiebung der Peak-Positionswerte umgekehrt proportional sind, schneidet die funktionale PCA-LDA besser ab. Anschließend haben wir beide Ansätze an experimentellen Raman-Spektren bewertet. Eine leichte Verbesserung der Klassifizierungsleistung spiegelt die Ergebnisse der simulierten Daten wider, da die experimentellen Daten den simulierten Daten mit hohem SNR ähnelten. Zusammenfassend lässt sich sagen, dass die funktionale Datenanalyse ein vielversprechendes Werkzeug für die Analyse von Raman-Spektren ist, insbesondere wenn die Qualität der Daten niedrig ist. Diese Eigenschaft macht die FDA zu einem hervorragenden Werkzeug für die Analyse von schnell oder in vivo aufgenommenen Spektren, die beide eine geringe Qualität aufweisen.

# 6 BIBLIOGRAPHY


[1] B. C. Wilson, M. Jermyn, and F. Leblond, 'Challenges and opportunities in clinical translation of biomedical optical spectroscopy and imaging', *J. Biomed. Opt.*, vol. 23, no. 3, p. 030901, Mar. 2018, doi: 10.1117/1.JBO.23.3.030901.

[2] '5 Ways Technology is Improving Health | UIC Health Informatics', *UIC Online Health Informatics*, Aug. 09, 2016. https://healthinformatics.uic.edu/blog/5-ways-technology-is-improving-health/ (accessed Aug. 04, 2021).

[3] '10 Biggest Technological Advancements for Healthcare in the Last Decade'. https://www.beckershospitalreview.com/healthcare-information-technology/10-biggest-technological-advancements-for-healthcare-in-the-last-decade.html (accessed Aug. 04, 2021).

[4] U. Optics, 'Biophotonic Technology And HIV - Custom Lens Design', *Universe Optics*, Oct. 31, 2017. https://www.universeoptics.com/biophotonic/ (accessed Dec. 29, 2021).

[5] J. A. Kim, D. J. Wales, and G.-Z. Yang, 'Optical spectroscopy for in vivo medical diagnosis—a review of the state of the art and future perspectives', *Prog. Biomed. Eng.*, vol. 2, no. 4, p. 042001, Aug. 2020, doi: 10.1088/2516-1091/abaaa3.

[6] G. Tucker *et al.*, 'Current Challenges and Potential Opportunities for the Pharmaceutical Sciences to Make Global Impact: An FIP Perspective', *J. Pharm. Sci.*, vol. 105, no. 9, pp. 2489–2497, Sep. 2016, doi: 10.1016/j.xphs.2015.12.001.

[7] M. Boenink, 'Molecular medicine and concepts of disease: the ethical value of a conceptual analysis of emerging biomedical technologies', *Med. Health Care Philos.*, vol. 13, no. 1, pp. 11–23, Feb. 2010, doi: 10.1007/s11019-009-9223-x.

[8] C. Puzzarini, M. P. de Lara-Castells, and M. J. Ramos, 'Challenges in spectroscopy: accuracy versus interpretation from isolated molecules to condensed phases', *Phys. Chem. Chem. Phys.*, vol. 21, no. 7, pp. 3395–3396, 2019, doi: 10.1039/C9CP90025J.

[9] M. Sauer, J. Hofkens, and J. Enderlein, Handbook of fluorescence spectroscopy and imaging: from single molecules to ensembles. Weinheim: Wiley-VCH, 2011.

[10] C. Albrecht, 'Joseph R. Lakowicz: Principles of fluorescence spectroscopy, 3rd Edition', *Anal. Bioanal. Chem.*, vol. 390, no. 5, pp. 1223–1224, Mar. 2008, doi: 10.1007/s00216-007-1822-x.

[11] G. F. Monnier, 'A review of infrared spectroscopy in microarchaeology: Methods, applications, and recent trends', *J. Archaeol. Sci. Rep.*, vol. 18, pp. 806–823, Apr. 2018, doi: 10.1016/j.jasrep.2017.12.029.

[12] H.-U. Gremlich and B. Yan, Infrared and Raman spectroscopy of biological materials. CRC press, 2000.

[13] M. Schmitt and J. Popp, 'Raman spectroscopy at the beginning of the twenty-first century', *J. Raman Spectrosc.*, vol. 37, no. 1-3, pp. 20–28, Jan. 2006, doi: 10.1002/jrs.1486.

[14] R. R. Jones, D. C. Hooper, L. Zhang, D. Wolverson, and V. K. Valev, 'Raman techniques: fundamentals and frontiers', *Nanoscale Res. Lett.*, vol. 14, no. 1, pp. 1–34, 2019.

[15] R. Petry, M. Schmitt, and J. Popp, 'Raman Spectroscopy—A Prospective Tool in the Life Sciences', *ChemPhysChem*, vol. 4, no. 1, pp. 14–30, 2003, doi: 10.1002/cphc.200390004.

[16] 'Handbook of Biophotonics, Volume 1: Basics and Techniques | Wiley', *Wiley.com*. https://www.wiley.com/en-us/Handbook+of+Biophotonics%2C+Volume+1%3A+Basics+and+Techniques-p-9783527410477 (accessed Dec. 08, 2021).

[17] 'The principle of superposition – x-engineer.org'. https://x-engineer.org/principle-superposition/ (accessed Dec. 29, 2021).

[18] A. Zheltikov, A. L'Huillier, and F. Krausz, 'Nonlinear Optics', in *Springer Handbook of Lasers and Optics*, F. Träger, Ed. Berlin, Heidelberg: Springer, 2012, pp. 161–251. doi: 10.1007/978-3-642-19409-2_4.

[19] J. M. Bueno, F. J. Ávila, and P. Artal, *Second Harmonic Generation Microscopy: A Tool for Quantitative Analysis of Tissues*. IntechOpen, 2016. doi: 10.5772/63493.

[20] R. M. Williams, W. R. Zipfel, and W. W. Webb, 'Interpreting Second-Harmonic Generation Images of Collagen I Fibrils', *Biophys. J.*, vol. 88, no. 2, pp. 1377–1386, Feb. 2005, doi: 10.1529/biophysj.104.047308.

[21] D. A. Kleinman, 'Theory of second harmonic generation of light', *Phys. Rev.*, vol. 128, no. 4, p. 1761, 1962.

[22] K. Svoboda and R. Yasuda, 'Principles of Two-Photon Excitation Microscopy and Its Applications to Neuroscience', *Neuron*, vol. 50, no. 6, pp. 823–839, Jun. 2006, doi: 10.1016/j.neuron.2006.05.019.

[23] M. Oheim, D. J. Michael, M. Geisbauer, D. Madsen, and R. H. Chow, 'Principles of two-photon excitation fluorescence microscopy and other nonlinear imaging approaches', *Adv. Drug Deliv. Rev.*, vol. 58, no. 7, pp. 788–808, 2006.

[24] P. T. So, C. Y. Dong, B. R. Masters, and K. M. Berland, 'Two-photon excitation fluorescence microscopy', *Annu. Rev. Biomed. Eng.*, vol. 2, no. 1, pp. 399–429, 2000.

[25] S. Li, Y. Li, R. Yi, L. Liu, and J. Qu, 'Coherent Anti-Stokes Raman Scattering Microscopy and Its Applications', *Front. Phys.*, vol. 0, 2020, doi: 10.3389/fphy.2020.598420.

[26] J.-X. Cheng and X. S. Xie, 'Coherent anti-Stokes Raman scattering microscopy: instrumentation, theory, and applications', *J. Phys. Chem. B*, vol. 108, no. 3, pp. 827–840, 2004.

[27] C. L. Evans and X. S. Xie, 'Coherent anti-Stokes Raman scattering microscopy: chemical imaging for biology and medicine', *Annu Rev Anal Chem*, vol. 1, pp. 883–909, 2008.

[28] N. Han, 'Computational and statistical approaches to optical spectroscopy', Thesis, Massachusetts Institute of Technology, 2018. Accessed: Dec. 29, 2021. [Online]. Available: https://dspace.mit.edu/handle/1721.1/120432

[29] T. Dearing, 'Fundamentals of Chemometrics and Modeling', 2010.

[30] L. A. Currie, ‘The importance of chemometrics in biomedical measurements’, 1991.

[31] R. G. Brereton *et al.*, ‘Chemometrics in analytical chemistry—part II: modeling, validation, and applications’, *Anal. Bioanal. Chem.*, vol. 410, no. 26, pp. 6691–6704, 2018.

[32] N. V. Tkachenko, Optical Spectroscopy: Methods and Instrumentations. Elsevier, 2006.

[33] C. R. Vogel, Computational methods for inverse problems. SIAM, 2002.

[34] J. Franklin, ‘The elements of statistical learning: data mining, inference and prediction’, *Math. Intell.*, vol. 27, no. 2, pp. 83–85, 2005.

[35] I. Guyon and A. Elisseeff, ‘An introduction to variable and feature selection’, *J. Mach. Learn. Res.*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[36] C. Camerlingo, I. Delfino, G. Perna, V. Capozzi, and M. Lepore, ‘Micro-Raman spectroscopy and univariate analysis for monitoring disease follow-up’, *Sensors*, vol. 11, no. 9, pp. 8309–8322, 2011.

[37] N. Ali, S. Girnus, P. Rösch, J. Popp, and T. Bocklitz, ‘Sample-Size Planning for Multivariate Data: A Raman-Spectroscopy-Based Example’, *Anal. Chem.*, vol. 90, no. 21, pp. 12485–12492, 2018, doi: 10.1021/acs.analchem.8b02167.

[38] P. Refaeilzadeh, L. Tang, and H. Liu, ‘Cross-validation.’, *Encycl. Database Syst.*, vol. 5, pp. 532–538, 2009.

[39] A. P. Bradley, ‘The use of the area under the ROC curve in the evaluation of machine learning algorithms’, *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, 1997.

[40] J. Huang and C. X. Ling, ‘Constructing New and Better Evaluation Measures for Machine Learning.’, in *IJCAI*, 2007, pp. 859–864.

[41] G. Bal, ‘Introduction to inverse problems’, Lect. Notes-Dep. Appl. Phys. Appl. Math. Columbia Univ. N. Y., 2012.

[42] F. D. M. Neto and A. J. da Silva Neto, *An introduction to inverse problems with applications*. Springer Science & Business Media, 2012.

[43] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, ‘Phase Retrieval with Application to Optical Imaging: A contemporary overview’, *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 87–109, May 2015, doi: 10.1109/MSP.2014.2352673.

[44] K. Jaganathan, Y. C. Eldar, and B. Hassibi, ‘Phase retrieval: An overview of recent developments’, *Opt. Compressive Imaging*, pp. 279–312, 2016.

[45] R. W. Gerchberg, ‘A practical algorithm for the determination of phase from image and diffraction plane pictures’, *Optik*, vol. 35, pp. 237–246, 1972.

[46] G. Whyte and J. Courtial, ‘Experimental demonstration of holographic three-dimensional light shaping using a Gerchberg–Saxton algorithm’, *New J. Phys.*, vol. 7, pp. 117–117, May 2005, doi: 10.1088/1367-2630/7/1/117.

[47] G. Yang, B. Dong, B. Gu, J. Zhuang, and O. K. Ersoy, ‘Gerchberg–Saxton and Yang–Gu algorithms for phase retrieval in a nonunitary transform system: a comparison’, *Appl. Opt.*, vol. 33, no. 2, pp. 209–218, Jan. 1994, doi: 10.1364/AO.33.000209.

[48] H. H. Bauschke, P. L. Combettes, and D. R. Luke, ‘Hybrid projection–reflection method for phase retrieval’, *JOSA A*, vol. 20, no. 6, pp. 1025–1034, 2003.

[49] J. G. Jo, S. J. Cho, M.-C. Park, Y. M. Jhon, and B.-K. Ju, 'Modified hybrid input-output algorithm for phase retrieval', in *25th International Vacuum Nanoelectronics Conference*, 2012, pp. 1–2.

[50] E. M. Vartiainen, K.-E. Peiponen, and T. Asakura, 'Maximum entropy model in reflection spectra analysis', *Opt. Commun.*, vol. 89, no. 1, pp. 37–40, 1992.

[51] E. Gornov, E. M. Vartiainen, and K.-E. Peiponen, 'Comparison of subtractive Kramers-Kronig analysis and maximum entropy model in resolving phase from finite spectral range reflectance data', *Appl. Opt.*, vol. 45, no. 25, pp. 6519–6524, 2006.

[52] M. T. Cicerone, K. A. Aamer, Y. J. Lee, and E. Vartiainen, 'Maximum entropy and time-domain Kramers–Kronig phase retrieval approaches are functionally equivalent for CARS microspectroscopy', *J. Raman Spectrosc.*, vol. 43, no. 5, pp. 637–643, 2012.

[53] E. M. Vartiainen, K.-E. Peiponen, H. Kishida, and T. Koda, 'Phase retrieval in nonlinear optical spectroscopy by the maximum-entropy method: an application to the $|\chi^{(3)}|$ spectra of polysilane', *JOSA B*, vol. 13, no. 10, pp. 2106–2114, Oct. 1996, doi: 10.1364/JOSAB.13.002106.

[54] E. M. Vartiainen, 'Phase retrieval approach for coherent anti-Stokes Raman scattering spectrum analysis', *JOSA B*, vol. 9, no. 8, pp. 1209–1214, Aug. 1992, doi: 10.1364/JOSAB.9.001209.

[55] C. H. Camp Jr, Y. J. Lee, and M. T. Cicerone, 'Quantitative, comparable coherent anti-Stokes Raman scattering (CARS) spectroscopy: correcting errors in phase retrieval', *J. Raman Spectrosc.*, vol. 47, no. 4, pp. 408–415, 2016.

[56] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. New York, NY: Springer New York, 2013. doi: 10.1007/978-0-8176-4948-7.

[57] E. J. Candès, 'Compressive sampling', in *Proceedings of the international congress of mathematicians*, 2006, vol. 3, pp. 1433–1452.

[58] M. F. Duarte and Y. C. Eldar, 'Structured compressed sensing: From theory to applications', *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4053–4085, 2011.

[59] H. Ohlsson, A. Y. Yang, R. Dong, and S. S. Sastry, 'Nonlinear basis pursuit', in *2013 asilomar conference on signals, systems and computers*, 2013, pp. 115–119.

[60] T. W. Bocklitz, S. Guo, O. Ryabchykov, and N. Vogler, 'Raman Based Molecular Imaging and Analytics: A Magic Bullet for Biomedical Applications!?', *Anal Chem*, p. 19, 2016.

[61] R. De Mets, A. Delon, M. Balland, O. Destaing, and I. Wang, 'Dynamic range and background filtering in raster image correlation spectroscopy', *J. Microsc.*, vol. 279, no. 2, pp. 123–138, 2020.

[62] C. Srisang, P. Asanithi, K. Siangchaew, A. Pokaipisit, and P. Limsuwan, 'Characterization of SiC in DLC/a-Si films prepared by pulsed filtered cathodic arc using Raman spectroscopy and XPS', *Appl. Surf. Sci.*, vol. 258, no. 15, pp. 5605–5609, 2012.

[63] T. H. Nguyen *et al.*, 'Denoising and deblurring of Fourier transform infrared spectroscopic imaging data', in *Computational Imaging X*, 2012, vol. 8296, p. 82960M.

[64] J. Dorney, F. Bonnier, A. Garcia, A. Casey, G. Chambers, and H. J. Byrne, 'Identifying and localizing intracellular nanoparticles using Raman spectroscopy', *Analyst*, vol. 137, no. 5, pp. 1111–1119, 2012.

[65] B. W. De Jong *et al.*, 'Discrimination between nontumor bladder tissue and tumor by Raman spectroscopy', *Anal. Chem.*, vol. 78, no. 22, pp. 7761–7769, 2006.

[66] P. Huang *et al.*, 'Study on glycoprotein terahertz time-domain spectroscopy based on composite multiscale entropy feature extraction method', *Spectrochim. Acta. A. Mol. Biomol. Spectrosc.*, vol. 229, p. 117948, 2020.

[67] M. Habibzadeh, A. Krzyżak, and T. Fevens, 'Comparative study of feature selection for white blood cell differential counts in low resolution images', in *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, 2014, pp. 216–227.

[68] S. Guo, P. Rösch, J. Popp, and T. Bocklitz, 'Modified PCA and PLS: Towards a better classification in Raman spectroscopy–based biological applications', *J. Chemom.*, vol. 34, no. 4, p. e3202, 2020, doi: https://doi.org/10.1002/cem.3202.

[69] R. B. Horton, E. Duranty, M. McConico, and F. Vogt, 'Fourier transform infrared (FT-IR) spectroscopy and improved principal component regression (PCR) for quantification of solid analytes in microalgae and bacteria', *Appl. Spectrosc.*, vol. 65, no. 4, pp. 442–453, 2011.

[70] A. I. Radu *et al.*, 'Toward food analytics: fast estimation of lycopene and β-carotene content in tomatoes based on surface enhanced Raman spectroscopy (SERS)', *Analyst*, vol. 141, no. 14, pp. 4447–4455, 2016.

[71] T. W. Bocklitz *et al.*, 'Pseudo-HE images derived from CARS/TPEF/SHG multimodal imaging in combination with Raman-spectroscopy as a pathological screening tool', *BMC Cancer*, vol. 16, no. 1, p. 534, Jul. 2016, doi: 10.1186/s12885-016-2520-x.

[72] S. Dochow *et al.*, 'Tumour cell identification by means of Raman spectroscopy in combination with optical traps and microfluidic environments', *Lab. Chip*, vol. 11, no. 8, pp. 1484–1490, 2011.

[73] A. Walter *et al.*, 'Raman spectroscopic detection of physiology changes in plasmid-bearing Escherichia coli with and without antibiotic treatment', *Anal. Bioanal. Chem.*, vol. 400, no. 9, pp. 2763–2773, 2011.

[74] C. Krafft, K. Thümmler, S. B. Sobottka, G. Schackert, and R. Salzer, 'Classification of malignant gliomas by infrared spectroscopy and linear discriminant analysis', *Biopolymers*, vol. 82, no. 4, pp. 301–305, 2006, doi: https://doi.org/10.1002/bip.20492.

[75] S. Li *et al.*, 'Noninvasive prostate cancer screening based on serum surface-enhanced Raman spectroscopy and support vector machine', *Appl. Phys. Lett.*, vol. 105, no. 9, p. 091104, 2014.

[76] N. Bergner *et al.*, 'Identification of primary tumors of brain metastases by Raman imaging and support vector machines', *Chemom. Intell. Lab. Syst.*, vol. 117, pp. 224–232, 2012.

[77] S. Albawi, T. A. Mohammed, and S. Al-Zawi, 'Understanding of a convolutional neural network', in *2017 International Conference on Engineering and Technology (ICET)*, 2017, pp. 1–6.

[78] J. Acquarelli, T. van Laarhoven, J. Gerretzen, T. N. Tran, L. M. C. Buydens, and E. Marchiori, 'Convolutional neural networks for vibrational spectroscopic data analysis', *Anal. Chim. Acta*, vol. 954, pp. 22–31, Feb. 2017, doi: 10.1016/j.aca.2016.12.010.

[79] S. Malek, F. Melgani, and Y. Bazi, 'One-dimensional convolutional neural networks for spectroscopic signal regression', *J. Chemom.*, vol. 32, no. 5, p. e2977, 2018.

[80] S. Guo, T. Mayerhöfer, S. Pahlow, U. Hübner, J. Popp, and T. Bocklitz, 'Deep learning for 'artefact'removal in infrared spectroscopy', *Analyst*, vol. 145, no. 15, pp. 5213–5220, 2020.

[81] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, 'Generative adversarial networks: An overview', *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, 2018.

[82] F. Q. Lauzon, 'An introduction to deep learning', in 2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA), 2012, pp. 1438–1439.

[83] S. Yu, H. Li, X. Li, Y. V. Fu, and F. Liu, 'Classification of pathogens by Raman spectroscopy combined with generative adversarial networks', *Sci. Total Environ.*, vol. 726, p. 138477, 2020.

[84] P. Pradhan, S. Guo, O. Ryabchykov, J. Popp, and T. W. Bocklitz, 'Deep learning a boon for biophotonics?', *J. Biophotonics*, p. e201960186, doi: 10.1002/jbio.201960186.

[85] L. Huang, X. Wu, Q. Peng, and X. Yu, 'Depth Semantic Segmentation of Tobacco Planting Areas from Unmanned Aerial Vehicle Remote Sensing Images in Plateau Mountains', *J. Spectrosc.*, vol. 2021, 2021.

[86] P. Pradhan et al., Semantic Segmentation of Non-linear Multimodal Images for Disease Grading of Inflammatory Bowel Disease: A SegNet-based Application. 2019. doi: 10.5220/0007314003960405.

[87] Z. C. Lipton, J. Berkowitz, and C. Elkan, 'A critical review of recurrent neural networks for sequence learning', *ArXiv Prepr. ArXiv150600019*, 2015.

[88] Y. Bengio, P. Simard, and P. Frasconi, 'Learning long-term dependencies with gradient descent is difficult', *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994, doi: 10.1109/72.279181.

[89] 'Identification of hepatitis B using Raman spectroscopy combined with gated recurrent unit and multiscale fusion convolutional neural network: Spectroscopy Letters: Vol 53, No 4'. https://www.tandfonline.com/doi/abs/10.1080/00387010.2020.1737944 (accessed Jan. 03, 2022).

[90] Z. Li *et al.*, 'Quantification of blood flow index in diffuse correlation spectroscopy using long short-term memory architecture', *Biomed. Opt. Express*, vol. 12, no. 7, pp. 4131–4146, Jul. 2021, doi: 10.1364/BOE.423777.

[91] Y. Yu, X. Si, C. Hu, and J. Zhang, 'A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures', *Neural Comput.*, vol. 31, no. 7, pp. 1235–1270, Jul. 2019, doi: 10.1162/neco_a_01199.

[92] C. Krafft, B. Dietzek, J. Popp, and M. Schmitt, 'Raman and coherent anti-Stokes Raman scattering microspectroscopy for biomedical applications', *J. Biomed. Opt.*, vol. 17, no. 4, p. 040801, 2012.

[93] M. Cui, B. R. Bachler, and J. P. Ogilvie, 'Comparing coherent and spontaneous Raman scattering under biological imaging conditions', *Opt. Lett.*, vol. 34, no. 6, pp. 773–775, Mar. 2009, doi: 10.1364/OL.34.000773.

[94] A. Zumbusch, G. R. Holtom, and X. S. Xie, 'Three-Dimensional Vibrational Imaging by Coherent Anti-Stokes Raman Scattering', *Phys. Rev. Lett.*, vol. 82, no. 20, pp. 4142–4145, May 1999, doi: 10.1103/PhysRevLett.82.4142.

[95] J. O. Ramsay, 'Functional Data Analysis', in *Encyclopedia of Statistical Sciences*, American Cancer Society, 2006. doi: 10.1002/0471667196.ess3138.

[96] J. O. Ramsay and C. J. Dalzell, 'Some Tools for Functional Data Analysis', *J. R. Stat. Soc. Ser. B Methodol.*, vol. 53, no. 3, pp. 539–561, 1991, doi: 10.1111/j.2517-6161.1991.tb01844.x.

[97] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, 'Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising', *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017, doi: 10.1109/TIP.2017.2662206.

# 7 PUBLICATIONS

# P1. TRENDS IN ARTIFICIAL INTELLIGENCE, MACHINE LEARNING AND CHEMOMETRICS APPLIED TO CHEMICAL DATA

Reprinted, with permission, from [Houhou, R, Bocklitz, T. Trends in artificial intelligence, machine learning and chemometrics applied to chemical data. Anal Sci Adv. 2021; 2: 128–141]. Copyright 2021 Analytical Science Advances.

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

| **Houhou, R[1], Bocklitz, T[2]. Trends in artificial intelligence, machine learning and chemometrics applied to chemical data. Anal Sci Adv. 2021; 2: 128-141** | | |
|---|---|---|
| Involved in | | |
| | 1 | 2 |
| Conceptual research design | × | × |
| Planning of research activities | × | × |
| Data collection | | |
| Data analyses and interpretation | | |
| Manuscript writing | × | × |
| Suggested publication equivalence value | 1.0 | |

128 Analytical Science Advances | Review doi.org/10.1002/ansa.202000162 Chemistry Europe European Chemical Societies Publishing

Received: 30 November 2020
Revised: 18 December 2020
Accepted: 21 December 2020

# Trends in artificial intelligence, machine learning, and chemometrics applied to chemical data

**Rola Houhou**[1,2] | **Thomas Bocklitz**[1,2]

[1] Institute of Physical Chemistry, Friedrich-Schiller-University Jena, Jena, Germany

[2] Department of Photonic Data Science, Member of Leibniz Research Alliance "Leibniz-Health Technologies", Leibniz Institute of Photonic Technologies, Jena, Germany

**Correspondence**
Thomas Bocklitz, Leibniz Institute of Photonic Technologies, Member of Leibniz Research Alliance "Leibniz-Health Technologies", 07745 Jena, Germany.
Email: thomas.bocklitz@uni-jena.de

**Funding information**
Deutsche Forschungsgemeinschaft, Grant/Award Number: CRC1076AquaDiva

## Abstract

Artificial intelligence-based methods such as chemometrics, machine learning, and deep learning are promising tools that lead to a clearer and better understanding of data. Only with these tools, data can be used to its full extent, and the gained knowledge on processes, interactions, and characteristics of the sample is maximized. Therefore, scientists are developing data science tools mentioned above to automatically and accurately extract information from data and increase the application possibilities of the respective data in various fields. Accordingly, AI-based techniques were utilized for chemical data since the 1970s and this review paper focuses on the recent trends of chemometrics, machine learning, and deep learning for chemical and spectroscopic data in 2020. In this regard, inverse modeling, preprocessing methods, and data modeling applied to spectra and image data for various measurement techniques are discussed.

KEYWORDS
2D chromatography, atomic force microscope, chemometrics, deep learning, electron microscope, inverse problem, machine learning, mass spectroscopy, nuclear magnetic resonance, preprocessing, vibrational spectroscopy, X-ray spectroscopy

## 1 | INTRODUCTION

A variety of measurement techniques to explore hidden aspects of a sample and to measure specific characteristics of a sample exists. Each of these measurement techniques exhibits its properties and is employed to measure a particular attribute of the sample, for example, the molecule structure at the atomic levels, the mass of particles or molecules, the isotopic signature of a sample, the absorbance, or the vibrational modes of molecules. However, the generated data from these measurements is often not directly utilized. Instead, chemometrics, machine learning, and deep learning tools are employed to extract underlying information from the data and link the data to specific applications. This review summarizes recent trends in the development of chemometrics, machine learning, and deep learning methods for a set of chemical important measurement methods like nuclear magnetic resonance (NMR), mass spectroscopy (MS), vibrational spectroscopy, X-ray spectroscopy, atomic force microscope (AFM), electron microscope (EM), and two-dimensional (2D) chromatography. Generally, the analysis of the respective data can be divided into two steps: data enhancement, for example, inverse modeling and preprocessing, and data modeling. A workflow showing the different steps involved in the analysis of chemical data is illustrated in Figure 1 and these steps are further discussed below.

The first step, data enhancement, is necessary because artifacts distort the generated data from all discussed measurement techniques. However, the enhancement can be achieved through either an inverse problem or a forward problem. A variety of methods to

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.
© 2021 The Authors. *Analytical Science Advances* published by Wiley-VCH GmbH

129 Analytical Science Advances | Review doi.org/10.1002/ansa.202000162 Chemistry Europe European Chemical Societies Publishing

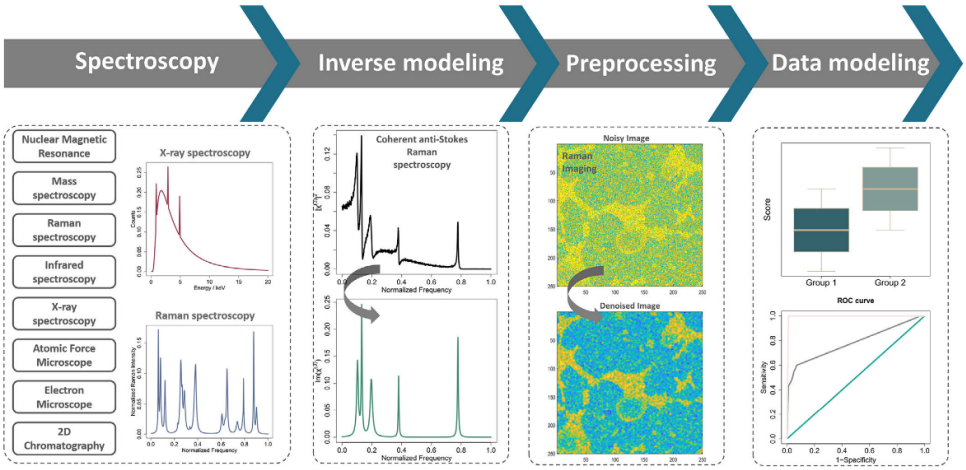


**FIGURE 1** General workflow of spectroscopic data analysis from data enhancement to data modeling. Inverse modeling, preprocessing, and data modeling represent key steps in this workflow.

solve or use both problems have been investigated. The preprocessing methods belong to the forward model type methods and are applied to remove different artifacts, for example, denoising,[1,2] baseline correction.[3–5] Moreover, the inverse problem is implemented using the observed data to estimate the original parameters of the chemical or physical system.[6,7] Learning from data via data modeling is the next step after the improvement of data is completed. In this regard, a wide range of mathematical and statistical techniques have been developed to extract important and relevant insights. Feature extraction[8,9] or selection,[10,11] classification,[12,13] and regression[14,15] are considered the essential categories of data modeling. Beside these classical techniques, deep learning gained popularity to solve spectroscopic and chemical applications.[16,17] Furthermore, the importance of chemometric methods increases to deal with the increasing size of spectroscopic datasets. Additionally, chemometric methods are needed to correct artifacts and shortcomings of specific spectroscopic technique. Therefore, the trends in 2020 and the limitations of applying chemometrics, machine learning, and deep learning methods on the aforementioned spectroscopic measurements are reviewed.

This review paper is divided into three main sections. The recent trends in applying chemometrics, machine learning, and deep learning methods to enhance spectroscopic data, including preprocessing methods and inverse modeling, are mentioned in section 1. Data modeling techniques for spectral data, including NMR, MS, vibrational spectroscopy, and X-ray, are covered in section 2. Finally, the advancements of chemometrics and artificial intelligence-based methods applied to imaging data involving AFM, EM, and 2D chromatography are discussed in section 3.

## 2 | CHEMOMETRICS, MACHINE LEARNING, AND DEEP LEARNING METHODS FOR ENHANCEMENT OF CHEMICAL DATA

In the last years, a growing interest in data analysis methods such as artificial intelligence-based methods can be recognized. For chemical applications, data generation is more beneficial if it is coupled with optimal techniques for the analysis of the generated data. The data analysis methods for the acquired chemical data are sample and task dependent. However, pre-analysis techniques are vital in evaluating chemical data regardless of the task to be solved. While chemical data are retrieved using various measurement techniques, the distortion of these data is very common. These distortions are called artifacts and result from the measurement devices, from the measurement or corrupting processes, and from the nature of the samples itself. The artifact removal or suppression leads to a data enhancement and is crucial for the data analysis to produce meaningful results. This enhancement of the chemical data can be categorized into two main types: the forward and the inverse problems. In the forward problem, the objective is to remove measurement artifacts and errors to determine the composition of the samples and the underlying structures of the chemical information. Concerning this problem, preprocessing techniques are being developed to remove the unwanted variations that limit the extraction of the underlying chemical-relevant structures. Instead, the inverse problem aims to reconstruct the missing information of the chemical/physical system, which was introduced through the measurement process. Recent trends in the application of the forward problems are referred to as preprocessing problems and the inverse problems are shown in the two following subsections.

130 Analytical Science Advances | Review doi.org/10.1002/ansa.202000162 Chemistry Europe European Chemical Societies Publishing

### 2.1 | Preprocessing

A detailed examination of preprocessing methods for a given data set is critical as these methods can also remove relevant chemical information. Therefore, the search for the best preprocessing method is vital, considering its impact on the subsequently performed data analysis and its outcome. These preprocessing methods can be employed to either remove noise contributions, replace missing values, interpret or remove baselines, or even a combination of these targets.[18] Depending on the studied problem, either a single preprocessing method or a combination of methods is applied to remove the underlying measurement errors and artifacts.

Nevertheless, retrieving the best preprocessing method or the best combination of techniques to remove all artifacts in the data at hand is challenging. Moreover, the preprocessing choice is the result of an exhaustive and long trial and error process. Consequently, a scientific effort is made to select appropriate preprocessing techniques. Recent papers concentrated on solving the trial and error problem by implementing automated software solutions that compare different preprocessing methods. An open-source python module "nippy" was developed by Torniainen et al[19] for semi-automatic comparisons of preprocessing techniques for near-infrared spectroscopy (NIRS). The presence of noise in the NIRS measurements significantly affects the multivariate methods needed for further analysis. Therefore, preprocessing categories are presented, including clipping, scatter correction, smoothing, derivatives, trimming, and resampling. This python-module was tested using two examples, and it resulted in a fast selection of different preprocessing combinations. The authors recommended following a specific order to apply these methods, which is crucial in NIR spectral data analysis. Further work to search for the best preprocessing strategies was presented in Martyna et al.[20] The authors proposed a novel concept to assess the preprocessing strategy using the ratio of between-groups to within-groups variance. This ratio was calculated on the first latent variable derived from the regularized multivariate analysis of variance (MANOVA). It is used to select the best preprocessing strategy that optimally highlights the differences between groups for highly multidimensional data. In addition, the search for the best preprocessing strategy was carried out using a genetic algorithm. Furthermore, the performance of this novel concept was verified by utilizing two forensic Raman spectral data sets. By assessing both problems in terms of discrimination, the authors successfully point out the sequence in which the preprocessing steps should be performed and extracted their most appropriate parameters. Since applying a preprocessing method might not entirely remove all artifacts, combining multiple methods is promising, and some of the recent publications deal with this combination. Roger et al.[21] developed a new approach to combine several pre-treatment techniques using sequential and orthogonalized partial least squares (SPORT). In their method, the authors applied different pre-treatment techniques to the same NIR transmission spectra including raw data, first and second derivatives, standard normal variate, and variable sorting for normalization. The sequential and orthogonalized partial least squares (SO-PLS) approach is used afterward to combine the resulting blocks of data. The application of SPORT showed good calibration performance in comparison with the existing stacking approach. With the booming of artificial intelligence, researchers utilized deep learning networks for various preprocessing tasks, for example, denoising. Zhang et al[22] tested two deep learning networks via the denoising autoencoder (DAE-1) and the stacked autoencoder (DAE-2) on NIR spectra. The results were compared to other denoising methods, including moving average smoothing, Savitzky-Golay smoothing (SGS), wavelet transform (WT), and empirical mode decomposition (EMD). In this regard, artificial and real NIR spectra were used for the model evaluation. The DAE-1 and DAE-2 applied on both simulated and real NIR spectra showed a better performance than the other methods. An additional study by Raulf et al[23] investigated the removal of disturbing scattering components from an infrared spectrum. The authors proposed deep learning via a convolutional neural network (CNN) as a preprocessing tool to remove the (resonant) Mie scattering. Their results showed that the deep learning approach is faster and can be used for a strong generalization across different tissue types. Their approach also overcame the trade-off between computation time and the bias of the corrected spectrum towards a reference spectrum. Moreover, Wahl et al[24] used a CNN as one single-step preprocessing for Raman spectra. In this paper, the CNN was trained using simulated data to handle three preprocessing steps, for example, cosmic ray removal, smoothing, and baseline subtraction. The preprocessing results were generally of higher quality than what was achieved using reference methods including second-difference, asymmetric least squares, and cross-validation. Additionally, the authors showed reliable results on measured Raman spectra from polyethylene, paraffin, and ethanol with background contamination from polystyrene. In conclusion, the authors proved deep learning as a promising tool for the automated preprocessing of Raman spectra. A drawback of the study was that the tested data basis was limited.

### 2.2 | Inverse modeling

Inverse modeling is the process of reconstructing missing information from observed measurements to identify its source or the corresponding model parameters. Inverse modeling tries to infer knowledge from given measurement data in the observation space $Y$ to the underlying unknown state $X$ of the sample or to a parameter function in the state space of $X$. General solutions for this problem do not exist or depends in an unstable way on the measurements, which is related to the ill-posed problem characteristics of inverse modeling.[25,26] A diversity of algorithms exists, and recent developments on this subject are listed below.

Yuan et al.[27] developed a new inverse modeling algorithm of a series of X-ray intensity measurements. Their objective was to recover the structure and composition of two-dimensional (2D) heterogeneous materials measured using X-ray spectroscopy by varying the beam's energy and position. Their method involves an iterative process of forward modeling, based on Monte-Carlo simulation, to determine the optimal structure to minimize the relative differences between the simulated and experimental characteristic X-ray intensities. In

131 Analytical Science Advances | Review doi.org/10.1002/ansa.202000162 Chemistry Europe European Chemical Societies Publishing

conclusion, the authors proved the feasibility of their approach in analyzing 2D heterogeneous materials for quantitative electron-induced X-ray. However, the input parameters such as beam positions, beam energies, and voxel size must be chosen appropriately. Another study by Hong et al[28] suggested using inverse modeling by combining X-ray computed tomography (XCT) testing and the finite element method to acquire the rust volume reduction coefficient. Conventional numerical models for corrosion expansion ignore the penetration of rust into the concrete matrix along the longitudinal direction. Their approach showed that the obtained reduction coefficient in the rust volume is linked to the rust expansion coefficient. The corrosion level obtained by XCT testing was significantly higher than what they found using the conventional corrosion model. Takeda et al[29] introduced a fundamental methodology of an automated system for material design. The authors built at the modeling stage a regression and a classification model to predict material properties and attributes. At the design stage, the trained predictive model is inverted to change and tune material structures. As a result, the two stages can achieve material design with user-demanded requirements. Also, the authors were able to inverse-design new molecular structures that satisfy the targeted LUMO energies. An inverse model was applied via long short-term memory (LSTM), a recurrent neural network-based, to retrieve the Raman-like shape from broadband coherent anti-Stokes Raman scattering (CARS)[29] by Houhou et al. The authors compared deep learning to other phase retrieval methods (maximum entropy method and Kramers – Kronig relation). The LSTM network outperformed these methods using artificial and experimental broadband CARS data. Additionally, the authors proved the stability of the deep learning method regarding the non-resonant background within CARS spectra. Guo et al[31] implemented a deep learning method in an inverse problem manner to remove artifacts from infrared spectra, which are caused by optical effects. Subsequently, the model could extract the pure absorption of the sample from the infrared measurements. The authors proposed an artifact removal approach based on a 1-dimensional U-Net shaped CNN using Poly (methyl methacrylate) as materials. The pure absorbance was successfully retrieved even when the absorbance is entirely overwhelmed by extensive artifacts. For the same objective, a different deep learning network was implemented by Magnussen et al[32] to recover the pure absorbance from infrared spectra. Initially, the Mie extinction extended multiplicative signal correction (ME-EMSC) algorithm extracts the pure absorbance from highly distorted spectra. Thereafter, the authors trained a deep learning network via the deep convolutional descattering autoencoder (DSAE) on a set of corrected infrared spectra. These corrected spectra were obtained using the ME-EMSC algorithm. Additionally, different reference spectra were used in this study to reflect the large variability in chemical features. In conclusion, the DSAE approach reduced the highly demanding computational time needed in the ME-EMSC algorithm for scatter correction. The DSAE outperformed the ME-EMSC correction in speed, robustness, and noise levels, while preserving the same chemical information in the corrected spectra.

After enhancing the data, the next step is to extract relevant information, which is achieved by applying chemometrics and machine learning techniques to the data, either spectra or images. An overview of chemometrics, machine learning, and deep learning methods is shown in Figure 2. Recent trends in applying chemometrics, machine learning, and deep learning techniques on spectral and image data are shown separately below.

## 3 | CHEMOMETRICS, MACHINE LEARNING, AND DEEP LEARNING METHODS FOR THE ANALYSIS OF SPECTRAL DATA

Spectroscopic measurements produce high-dimensional profiles containing a high amount of information, which can be optimally exploited using chemometrics, machine learning, and deep learning methods. These methods aim to discover the underlying chemical properties of the sample more precisely and accurately. Recent advances in applying chemometrics, machine learning, and deep learning methods used on spectral data generated by nuclear magnetic resonance (NMR), mass spectroscopy (MS), vibrational spectroscopy, and X-ray spectroscopy are discussed in the following subsections.

### 3.1 | Nuclear magnetic resonance

Nuclear magnetic resonance (NMR) describes a measurement principle where the nuclei of specific atoms are irradiated by a static magnetic field and then exposed to a second oscillating magnetic field.[35] The analysis of NMR spectra is challenging, and it is not straightforward to draw a conclusion directly from the spectra or even interpret the spectra without the use of chemometric methods. Therefore, combining NMR spectra with chemometrics, machine learning, and deep learning methods is beneficial, and some of the recent applications are discussed below.

The analysis of the 1D $^{1}$H NMR spectra of metabolomics samples is challenging since resonances overlap in specific chemical shift regions. However, Pérez et al[36] suggested using a chemometric approach through multivariate curve resolution-alternating least squares (MCR-ALS) to facilitate the steps of metabolites profiling and resonance integration. The authors proved the ability of this method to extract the concentrations and resonances from untargeted metabolites. Their approach was validated using 1D $^{1}$H NMR spectra from metabolomic profiling of zebrafish upon acrylamide exposure. Consequently, the authors recommended using their approach to identify spectral features and as biomarker discovery. Another issue for metabolomics NMR-based was presented by Miros et al.[37] in which the growth of *Hypericum perforatum*, or St. John's wort, plants are monitored under different light conditions. The authors developed a toolkit combining 1D $^{1}$H NMR spectra with multivariate analysis to extract differences in chemical profiles. As a result, specific metabolites were identified as markers for the difference between the plant growths under different light conditions and glutamine, sucrose, and fructose were found to be chemical markers of light conditions. Another area of interest was related to herbal medicines in which Zhao et al[38]

Review
doi.org/10.1002/ansa.202000162
Chemistry Europe
European Chemical Societies Publishing

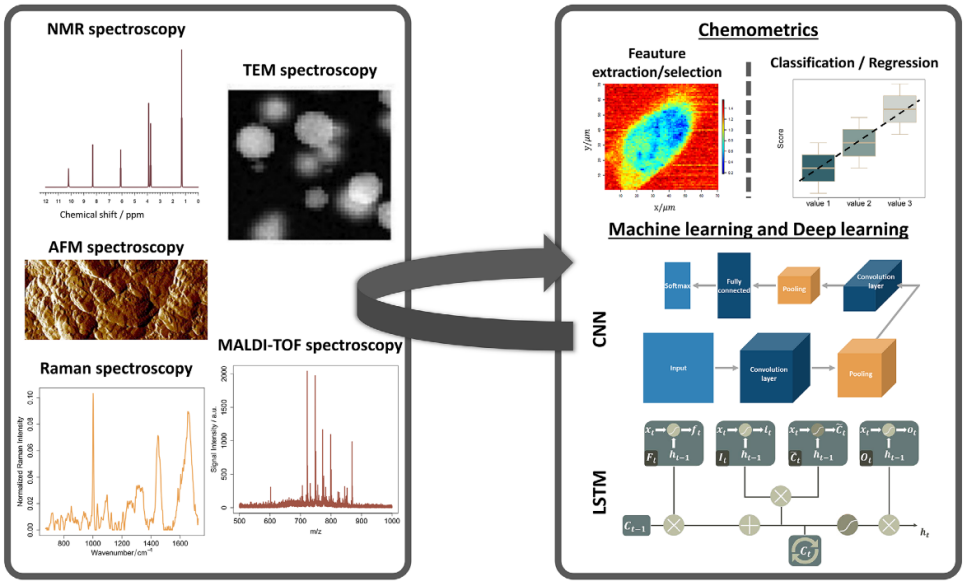


**FIGURE 2** An overview of chemometrics, machine learning, and deep learning methods applied to spectroscopic measurements. The AI-based methods like chemometrics, machine learning, and deep learning visualized on the right are utilized to the chemical/spectroscopic data. The data on the left is not representative for all data sources, but we focused in this review on these data types. [LSTM network, AFM, and TEM images are retrieved from references,[30,33,34] respectively]

analyzed the effects of the multistep processing on the chemical changes. The authors applied a chemometric method on $^1$H NMR spectra to provide comprehensive information on the chemical changes during the processing steps of the Danshen extract. For instance, the hierarchical classification analysis (HCA) clustered the samples according to the processing steps, which indicates that $^1$H NMR enables the identification of the critical control points based on information of the organic compounds present in the sample. Additionally, a principal component analysis (PCA) and an orthogonal partial least squares discriminant analysis (OPLS-DA) were applied to distinguish the major metabolite differences between the intermediates before and after the critical control point. The combination of $^1$H NMR and chemometrics proved to be an effective process quality control tool. Therefore, the authors proposed to apply their approach to other herbal medicines to identify critical control points and potential chemical markers. Marion et al[39] developed a new method, adaptive clustering around latent variables (AdaCLV), for simultaneous dimensionality reduction and variable clustering. This method was applied to NMR spectra and can be used for the identification of potential biomarkers. Briefly, AdaCLV filters out variables that do not vary significantly between samples and approximates cluster membership degrees on the remaining variables. The potential overlapping clusters are identified, and the ranking of variable importance within a cluster is achieved. Compared with other clustering methods, AdaCLV estimated latent variables and cluster membership with higher or equivalent precision, and it showed less sensitivity regarding the used hyperparameters. Further analysis was performed by Coimbra et al[40] through merging a time-domain nuclear magnetic resonance (TD-NMR) spectra with chemometric methods to determine the presence of formaldehyde in raw milk samples. PCA, partial least squares (PLS), and soft independent modeling of class analogy (SIMCA) were used to discriminate the samples regarding the level of milk adulteration. SIMCA overcame PCA and PLS with a good discrimination and a high predictive index. Consequently, TD-NMR combined with chemometrics proved its effectiveness for the dairy industry to check the formaldehyde levels in raw milk. Zhang et al[41] combined $^1$H NMR with chemometric methods to classify the monofloral Chinese honey based on botanical and geographical origins. $^1$H NMR spectra on samples of 8 classes were collected across China. A PCA could be used to successfully classify their botanic origins while the classification at different geographical levels was effectively distinguished using orthogonal partial least square discriminant analysis (OPLS-DA). This study reported several benefits, including a small sample amount, a simple preparation, a short testing time, and a non-targeted multi-species detection. Recent trends for chemical analysis also include research on the application of deep learning techniques. For instance, Kong et al.[42] proposed a combination of deep learning via a convolutional neural network (CNN) and the sparse matrix completion method to speed up 2D nanoscale NMR spectroscopy, which is vital for molecular structure determination. The use of a CNN successfully suppressed the

133 Analytical Science Advances | Review doi.org/10.1002/ansa.202000162 Chemistry Europe European Chemical Societies Publishing

observation noise and improved the sensitivity. An additional challenge was presented by Hou et al,[43] where the authors aim to assess the use of deep learning as a tool for rapid and accurate identification of edible oils. Therefore, two-dimensional CNN (2D-CNN) and one-dimensional CNN (1D-CNN) were used to classify different types of edible oils using different low-field nuclear magnetic resonance (LFNMR) spectra. The results showed that transverse relaxation decay signals analyzed by the 1D-CNN presented the best classification ability.

### 3.2 | Mass spectroscopy

Mass spectroscopy (MS) measures the mass of charge ratio of molecules. It is used to quantify known materials, identify unknown compounds within a sample, and elucidate the structure and chemical properties of different molecules. The fundamental principle involves the fragmentation of a compound or molecule into charged species, which are accelerated, deflected, and finally focused on a detector according to their mass and charge ratio. Ion deflection is based on charge, mass, and velocity, ions separation is based on mass to charge ($m/z$) ratio, and detection is proportional to the abundance of these ions.[44]

With rapidly growing chemometrics, machine learning, and deep learning methods, MS took its share of the pie. For instance, Duan et al[45] developed a new software, QPMASS, to analyze large-scale gas chromatography-mass spectrometry (GC-MS) data. GC-MS analysis generates many fragment ions for each analytic compound, making the tasks of sample deconvolution and peak alignment very challenging. To deal with this issue, the authors implemented parallel computing with an advanced dynamic programming approach. This approach aligns peaks from multiple samples based on the similarity of each pair calculated using retention time and mass spectra. The diagram of using dynamic programming and the parallel peak alignment in QPMASS is illustrated in Figure 3. As a result, QPMASS enabled fast processing of large-scale datasets and reduced false positive and false negative errors to be less than 5%.

A further challenge for GC-MS was tackled by Alkhalifah et al,[46] in which the search for automated algorithmic clustering methods was discussed. The authors developed VOCCluster, a python-based algorithm that quickly and efficiently analyzes features of deconvolved GC-MS breath data. Compared to a manual volatile organic compound (VOC) panel, the results showed a superior and faster performance with an accurate clustering of 96% of VOCs. Additional trends were introduced by Papagiannopoulou et al.[47] to identify pathogenic bacteria cells in urine samples. First, the authors implemented matrix-assisted laser desorption/ionization-time of flight mass spectrometry (MALDI-TOF MS) on individual bacterial cells and identified species rapidly and with acceptable accuracy. In this regard, a deep learning technique via CNN was applied. The results showed similar performance compared to traditional supervised machine learning algorithms, including logistic regression, random forests, and k-nearest neighbor classification. Further issues were tackled by Li et al,[48] who worked to decrease the false positive rate and to improve the low sensitivity arising from a database search engine. This engine is used to identify significant histocompatibility complex (MHC)-binding peptides in mass spectroscopy. The authors developed DeepRescore, a post-processing tool, to improve the sensitivity and reliability in peptide identification. Their approach combines peptide features derived from deep learning predictions with previously used features to rescore peptide-spectrum matches. The results showed that rescoring by DeepRescore on two public immunopeptidomics datasets increases both the sensitivity and reliability of the prediction of MHC-binding peptides. As well, it showed that the deep learning-derived features improved the performance.

### 3.3 | Vibrational spectroscopy

Vibrational spectroscopy is a non-destructive identification method that measures the vibrational energies of molecular vibrations in the sample. Each chemical bond has a unique vibrational energy, which will be different from one compound to another. This unique energy provides each compound with a unique fingerprint, which is vital in determining compound structures, identifying, and characterizing compounds, and identifying impurities. There are two types of vibrational spectroscopy discussed herein: infrared absorption and Raman spectroscopy. The main difference between these measurement techniques is that in infrared spectroscopy, the absolute frequencies at which a sample absorbs radiation are measured, while Raman spectroscopy measures inelastic scattering in a relative manner. These both are complementary as the vibrations feature different selection rules and both methods are essential to extract the full picture of the vibrational modes in a molecule.[49] Similar to other spectroscopic techniques, vibrational spectroscopy requires advanced data processing to extract meaningful information from spectra. Recent trends about the use of chemometrics, machine learning, and deep learning methods in Raman and infrared spectra are presented below.

Akpolat et al[50] discussed the use of a handheld Raman spectroscopic device with pattern recognition techniques for classification and quantification of different types of tomato carotenoids, which is of great interest for health issues. In this regard, samples with varying carotenoids profiles were non-destructively measured via a handheld Raman spectrometer. The derived spectra were analyzed for classification and quantification purposes using soft independent modeling of class analogy (SIMCA), artificial neural network (ANN), and partial least squares regression (PLSR). A good classification of tomatoes based on their carotenoid profile of 93% and 100% is shown using SIMCA and ANN, respectively. Besides this result, PLSR and ANN were able to achieve a good quantification of all-*trans*-lycopene. Consequently, the authors suggested using their approach as a tool for breeders to provide real-time information on carotenoid profiles. Another study aimed at the quantification of complex mixtures is discussed in Han et al[51] The authors developed a two-stage algorithm based on Bayesian modeling and implemented it on Raman spectra. First, a hierarchical Bayesian model was constructed to learn the peak representation for a target analyte spectrum. A reversible-jump

134 Analytical Science Advances | Review doi.org/10.1002/ansa.202000162 Chemistry Europe European Chemical Societies Publishing

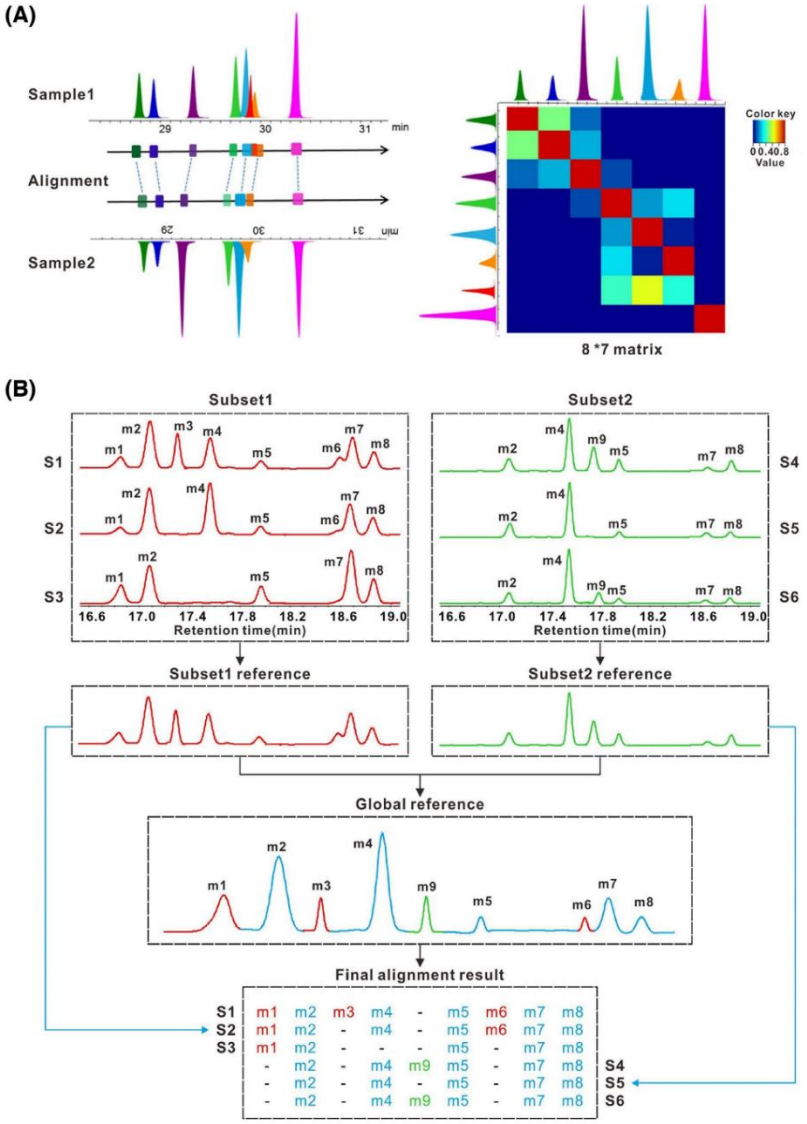


**FIGURE 3** Dynamic programming and parallel peak alignment in QPMASS. The diagram in panel (A) describes the dynamic programming for peak alignment. The left panel shows the matched peaks after alignment, and the right panel shows the score matrix for the peak alignment of two samples. In panel (B), the diagram illustrates the parallel peak alignment. The subset references are the consensus sample derived from the average mass spectra and retention time. The final alignment result describes the origin of the detected peaks. [Retrieved and adapted from reference[15]]

135 Analytical Science Advances | Review doi.org/10.1002/ansa.202000162 Chemistry Europe European Chemical Societies Publishing

Markov chain Monte Carlo (RJMCMC) is then used to estimate the target analyte concentration in a mixture using the peak variables learned in the first step. Their approach was implemented on both simulated and experimental spontaneous Raman spectral datasets and showed good quantification of glucose concentration. As a result, the authors suggested using this algorithm as a complementary tool for Raman spectroscopy-based mixture quantification studies. Since *Arcobacter* is an emerging foodborne pathogen that has become more important in recent years, Wang et al[52] were interested in fast identification of Arcobacter. The authors combined Raman spectroscopy with deep learning via a CNN to identify various species of *Arcobacter*. Their method achieved a high identification accuracy (97.2%) at the species level. Furthermore, a fully connected artificial neural network (ANN) was constructed on Raman spectra to determine the actual ratio of a specific *Arcobacter* species in a bacterial mixture. In their approach, the accuracy of Raman spectroscopy for bacterial species determination was improved and enabled rapid identification of *Arcobacter*. A further challenge in assessing endoscopic disease severity in Ulcerative colitis (UC) patients using Raman spectroscopy was explored by Kirchberger-Tolstik et al.[53] In this study, the endoscopic disease severity evaluation was performed according to the four Mayo subscores. The authors coupled Raman spectra with a one-dimensional CNN (1D-CNN) to identify the level of colonic inflammation and then applied first-order Taylor expansion to extract the important Raman bands for this classification. Their approach indicated a good classification performance and can be used further as a complementary method for UC characterization and diagnosis. Zafar et al[54] introduced a novel method that monitors the oxyhemoglobin changes produced by neuronal activations using functional near-infrared spectroscopy (fNIRS). The authors suggested using a kernel-based recursive least squares (KRLS) algorithm to reduce the detection time in fNIRS signals from the neuronal activation. In this manner, the KRLS algorithm with a Gaussian kernel was used. It showed the best performance for estimating both changes in oxyhemoglobin and deoxyhemoglobin. Therefore, a neuronal activation can be determined in about 0.1 s with fNIRS using KRLS prediction, enabling almost real-time detection if combined with electroencephalography. A different matter that involves the detection of petroleum presence in soil mixtures was covered by J. Galán-Freyle et al[55] A remote-sensed tool that combines artificial intelligence and a portable mid-infrared quantum cascade laser spectroscopy (QCL) system was developed. First, remote sensing combined with support vector machine (SVM) was used to detect the presence or absence of traces of petroleum in soil. Then, PCA, PLS-DA, and SVM were implemented to discriminate between the different soil types. Additionally, a statistical analysis method was developed to calculate limits of detection (LOD) and limits of decision (LD) from fits of the detection probability. As a result, a SVM provided better identification probabilities of soils that contains traces of petroleum. Le[56] proposed the use of deep learning with NIR for rapid analysis of cereal characteristics. First, the author applied the deep learning-stacked sparse autoencoder (SSAE) method on the corn and the rice datasets. This deep learning tool reduces the NIR data dimension, eliminates the interference information, and obtains advanced data features. A combination of the affine transformation (AT) and the extreme learning machine (ELM) was then established to predict the different types of cereals. Their approach provides a fast, efficient, and cost-effective method for cereal characteristics analysis. Xu et al[57] used functional near-infrared spectroscopy (fNIRS) to investigate hemodynamic fluctuations in the bilateral temporal cortices for typically developing (TD) children and children with autism spectrum disorder (ASD). The authors proposed the use of fNIRS time series to estimate the global time-varying behavior of brain activity and then combined two deep learning networks, LSTM and CNN, to explore the potential patterns of temporal variation for ASD identification. This global time-varying behavior is measured through the Augmented Dickey-Fuller (ADF) test. The ADF test showed that ASD children performed weaker stationarity in hemodynamic fluctuation variation than controls, as illustrated in Figure 4. Also, the proposed deep learning approach was able to differentiate between ASD and TD children accurately.

Accordingly, the characterization of the time-varying behavior of brain activity holds promising potential for a better understanding of the underlying causes of ASD. Furthermore, the deep learning framework has the potential for diagnosing children with the risk of ASD.

### 3.4 | X-ray

X-ray spectroscopy and X-ray intensity measurements enable the characterization of materials. This is done by X-ray excitation, for example, high energetic electromagnetic radiation, which results in the emission of characteristic wavelengths for the elements of the specimen/sample. These specific wavelengths can be used to generate insights in the elemental composition of the sample. X-ray spectroscopy can be used to address a range of scientific questions, from interactions of simples molecules to the structure of the human brain.[58] Chemometrics, machine learning, and deep learning methods proved to be great tools to solve practical problems, especially in chemistry and spectroscopy. However, their application in the X-ray field is not as broad as in other spectroscopic areas. Therefore, recent trends concerning the application of chemometrics, machine learning, and deep learning methods on X-ray spectra are mentioned below.

Otsuka et al.[59] investigated the effect of humidity-controlled storage of amorphous rebamipide (RB), RB form I, and RB solid dispersion with different surfactant and polymers. Their method is based on applying PCA on the dataset generated from power X-ray diffractograms (PXRD) and NIR spectra. The authors showed that the fusion of data from different sources resulted in correlations between NIR spectra and diffraction patterns in both neat RB and solid dispersion samples. As a result, the presented methods can be a useful model for evaluating amorphous active pharmaceutical ingredients without a standard sample. An additional study was motivated by the recent improvements of portable X-ray devices to detect meteorites from hot and cold deserts. In this study, chemometrics was used for the analysis of the X-ray data, which Allegreta et al.[60] measured using a portable energy dispersive X-ray fluorescence spectroscopy (pED-XPF) instrument. Moreover, the meteorite classification was achieved

136 Analytical Science Advances | Review doi.org/10.1002/ansa.202000162 Chemistry Europe European Chemical Societies Publishing

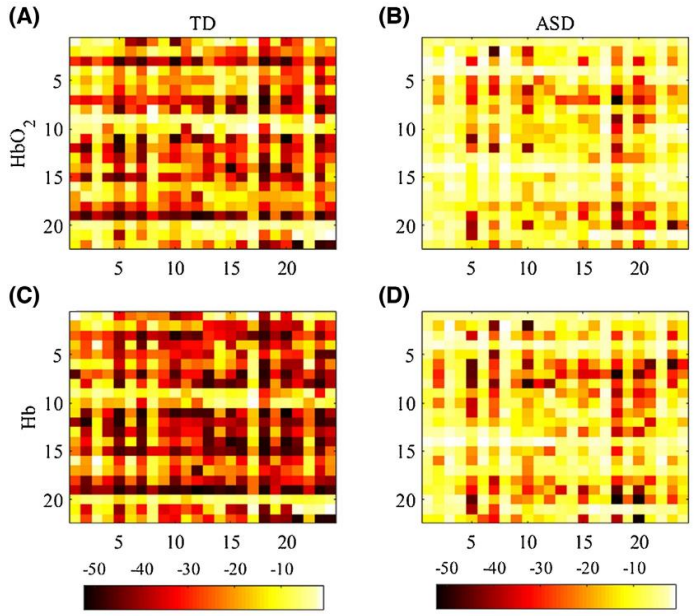


**FIGURE 4** ADF color map values for ASD and TD children. The *x*-axis indicates the number of optical channels, while the *y*-axis shows the numbering of subjects. A clear distinction in the response between ASD and TD children is concluded. [Retrieved and adapted from reference[57]].

by applying various chemometrics methods on standardized X-ray fluorescence (XRF) spectra. In this regard, PCA, cubic support vector machine (CSVM), fine kernel nearest neighbor (FKNN), subspace discriminant-ensemble classifiers (SD-EC), and subspace discriminant KNN-EC (SKNN-EC) methods were tested. Their approach allowed the rapid and trustable classification and discrimination of meteorites in macro-groups. 100% accuracy in sample classification was obtained using each of these machine learning methods. Consequently, their approach proved effective and promising for the differentiation and classification of real or supposed meteorites. Carbone et al.[61] proposed a graph-based deep learning architecture to predict the X-ray absorption near-edge structure (XANES) spectra of molecules. Briefly, XANES encodes vital information about the local chemical environment, but significant challenges arise from the material's complexity associated with chemical composition and structure. The author proved with their approach that the predicted spectra reproduce all prominent peaks, with 90% of the predicted peak locations within 1 eV of the ground truth. This method can also be used to provide a general-purpose, high-throughput capability for predicting spectral information of a broad range of materials, including molecules, crystals, and interfaces. Other issues were researched by Mullaliu et al,[62] who investigated the electrochemical activity in manganese hexacyanoferrate (MnHCF) by varying the interstitial ion content. The authors combined X-ray absorption spectroscopy and a MCR-ALS. Their approach intent to assess the structural and electronic modifications during Na release and Li insertion. As a result, MCR-ALS showed that water absorption affects the reaction dynamics only at the Fe site. Besides, the Mn local environment encountered a substantial yet reversible Jahn-Teller effect upon interstitial ion removal due to the formation of trivalent Mn. Furthermore, this is associated with decreased of equatorial Mn—N bond lengths by 10%.

## 4 | CHEMOMETRICS, MACHINE LEARNING, AND DEEP LEARNING METHODS FOR THE ANALYSIS OF IMAGE DATA

AI-based techniques like chemometrics, machine learning, and deep learning are used to analyze chemical image data for decades. These tools were adapted to different structures and types of images, from the simplest grayscale image to hyperspectral images, and provided new insights on their spatial and spectroscopic information. Recent trends in applying chemometrics, machine learning, and deep learning methods on image data from atomic force microscopy (AFM), electron microscopy (EM), and 2D chromatography are presented below.

137 Analytical Science Advances | Review doi.org/10.1002/ansa.202000162 Chemistry Europe European Chemical Societies Publishing

### 4.1 | Atomic force microscopy (AFM) and electron microscopy (EM)

Initially, atomic force microscopy (AFM) is a high-resolution imaging technique where a small probe with a sharp tip is scanned back and forth in a controlled manner across a sample to measure its surface. This procedure results in a topography of the sample at atomic resolution. AFM microscopy techniques can be performed in various scanning modes that enable nanoscale characterization of different material properties such as electrical, magnetic, and mechanical properties.[63]

On the other hand, Electron microscopy (EM) uses an electron beam to create an image of a sample. Because of the higher energy of the electrons compared with visible light, an electron microscope has a much higher resolution than a light microscope. EM can be used to investigate the microstructure of a wide range of biological and inorganic specimens. Moreover, it provides morphologic and crystallographic information.[64]

The incorporation of chemometrics, machine learning, and deep learning methods for AFM imaging techniques is limited, and recent applications are mentioned below. For instance, Yablon et al[65] investigated three different applications of machine learning in AFM imaging. In their first application, the authors applied two AI-based methods like neural networks and CNNs. These networks are trained to differentiate two different multicomponent polymer blends based on their AFM phase images. The results showed that CNN performs perfectly with 100% accuracy on the test data. A feature extraction approach was investigated in their second application to detect particles in an image with a complex background and many aggregates. In this manner, an initial logistic regression model was trained. The corresponding output was fed to a Hessian blob detection algorithm to isolate particles with a circular shape. Their proposed method significantly improves the particle identification compared with a commercial particle analysis package. Finally, the authors discussed the current status of autonomous instrumentation in AFM and its limitation, which needs a large amount of optimization. The additional issue concerned with the time required by the oscillating tip to reach the steady-state motion in AFM imaging was discussed in Javazm et al.[66] Due to AFM restricted scanning speed, the authors proposed an innovative imaging technique based on an artificial intelligence-based algorithm. Thereby, multiple artificial intelligence-based methods were investigated, including multilayer perceptron, radial basis function neural networks, and adaptive neural fuzzy inference system networks. Their approach aims to show the capability of artificial intelligence methods to estimate the surface topography directly. Therefore, the results showed that the multilayer perceptron overcame the other techniques in terms of surface characteristics estimation. In conclusion, the authors suggested using their approach to reach an accurate and fast estimation of the surface topography in AFM imaging. Further advantages of their method are that no closed-loop controller is needed, and the capability of estimating simultaneous the topography, the Hamaker parameter, and the tip-sample interaction force. Regarding the AFM limitation mentioned above, a further investigation was performed by Payam et al[67] The authors intended to explore the probe-sample interactions in dynamic AFM. Therefore, a novel approach for dynamic AFM data acquisition and imaging based on wavelet transform was applied in the photodetector data stream. By use of simulations, their approach was able to produce data including information about the transient response of amplitude and phase with the variation of material and sample topography properties. The method reliability was tested by comparing it with a standard lock-in amplifier (LIA) analysis. It showed the ability to reconstruct amplitude and phase images of standard samples, starting from time-domain data of actual measurements. The authors indicated that using their method would improve the measurement speed, reduce the loss of information and give access to a wealth of information about the transient response, which leads to the possibility of analyzing material properties in dynamic AFM.

Some of the challenges in electron microscopy (EM) imaging that are resolved using chemometrics, machine learning, and deep learning methods are presented below. For instance, Yu et al[68] addressed the limitations of traditional image recognition methods, such as the inability to obtain the complete pore space characteristics in scanning electron microscopy (SEM) images. Additionally, traditional image recognition methods for SEM images lead to poor segmentation results and a low accuracy. The authors implemented a semantic image segmentation technique based on artificial intelligence to analyze the pore characteristics and explore the relationship between the microscopic pore characteristics and the macroscopic permeability parameters of the sandstone in the SEM images. The results showed that the application of deep learning via CNNs accurately recognize images and allowed the automatic processing of microscopic images. Furthermore, this significantly improved the accuracy of the pore identification in rock samples. Li et al.[34] investigated a reaction-convection-diffusion model to track spatial-temporal patterns in scanning transmission electron microscopy (STEM) videos of Pt nanoparticle formation and graphene contamination. The authors developed a data-driven approach utilizing pixel-level information to infer the underlying partial differential equation (PDE) that governs the spatial-temporal patterns in STEM videos. The PDE model resulted in a redundant basis matrix, leading to non-unique numerical solutions. Therefore, the least angle regression algorithm (LARS) was utilized to reduce the ambiguity and to improve its interpretation. Additionally, the optimal parameter $\lambda$, used to balance model parsimony and descriptive capability, is determined by Mallow's Cp criteria (Cp). The Pearson correlation coefficient (PCC) is used to track discrepancies between experimental and estimated frames. The analysis applied to STEM multiple Pt particles video is illustrated in Figure 5.

Both the simulated and experimental datasets proved that the use of the PDE models has the potential to capture the characteristic behavior of spatial-temporal patterns at a mesoscopic scale in STEM videos and can be of great help for the investigations of complex time-evolving processes.

### 4.2 | Two-dimensional chromatography

Two-dimensional chromatography is a chromatographic technique that yields information on the chemical composition of a sample, by

Review
doi.org/10.1002/ansa.202000162
Chemistry Europe
European Chemical Societies Publishing

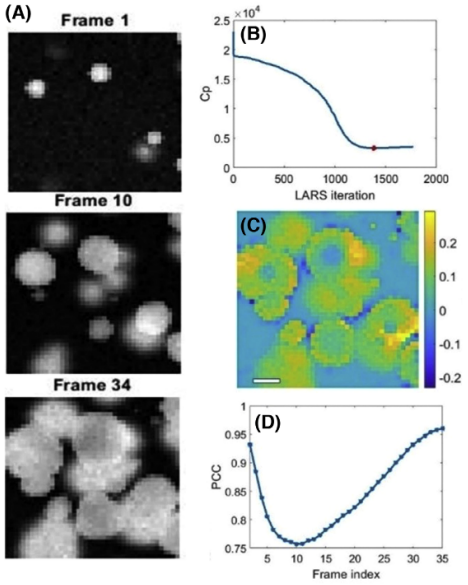


**FIGURE 5** Estimation based on multiple Pt particles STEM video. Snapshots of experimental data are illustrated in (A). The optimal parameter calculated using Cp criteria is shown in red in (B). The estimated frame is shown in (C). PCC in (D) proved the convergence to the experimental data between frame 11 and 35. [Retrieved from reference [34]]

combining two separation systems. Typically, two different chromatographic columns are connected in sequence, and an aliquot from the first column is injected in the second column. As the separation systems are often working differently in the columns, peaks that could not be separated using the first column can be separated using the second column in 2D chromatography.[69] A large amount of information is contained in high-resolution chromatography, and it is complicated to extract all relevant information and deduce correct and straightforward solutions. However, recent research for efficient chemometric data-processing strategies is presented below.

Huygens et al.[70] reported three evolutionary algorithms to enhance searches in the method development spaces of 1D- and 2D-chromatography, including genetic algorithms (GA), evolution strategies (ES), and covariance matrix adaptation-evolution strategy (CMA-ES). The authors compared these algorithms to a plain grid search. The results showed significant outperformance, especially in terms of the number of search runs needed to achieve a given separation quality. Additionally, the ES and GA performance followed a hyperbolic law in the large search run number limit. Subsequently, the convergence rate in the hyperbolic function can quantify the difference in the required number of search runs in these algorithms. A further problem on two-dimensional liquid chromatography was tackled by Pérez-Cova et al.[71] The authors employed a two-dimensional liquid chromatography method hyphenated simultaneously to two different detectors on a mixture of 31 pharmaceutical compounds. MCR-ALS was then used to evaluate the obtained two-dimensional chromatograms. The authors perform two evaluations. First they assessed the multilinear behavior of the high-dimensional data for each of the two detection modes. Second, they check the model performance for the multiset data obtained by fusion of the data coming from both detectors. Additionally, their approach proved that data fusion from the two detectors increased the ability for compound identification. Nagai et al.[72] identified a novel biomarker of hepatocellular carcinoma (HCC) in the human liver using multivariate analysis methods such as PCA and OPLS-DA. The data was extracted using an ultrahigh-performance liquid chromatography/quadrupole time-of-flight mass spectroscopy (UHPLC/QTOFMS) instrument equipped with a mixed-mode column. The results showed that novel biomarkers for HCC were identified with the global metabolomics/metabolic profiling (G-Met) method. Besides this, the difference in fatty acid species of triglyceride in tumor regions was demonstrated by high definition mass spectroscopy (HDMS) combined with UHPLC/QTOFMS. It showed localization in cryosections using desorption electrospray ionization-mass spectroscopy imaging (DESI-MSI). In conclusion, G-Met combined with UHPLC/QTOFMS and HDMS and distribution analysis by DESI-MSI is useful for characterizing tumor cell progression and discovering prospective biomarkers.

## 5 | SUMMARY AND OUTLOOK

Developments of artificial intelligence-based techniques like chemometrics, machine learning and deep learning have occupied the interest of researchers for decades. These data analysis techniques combined with spectroscopic measurements in chemistry and chemical data have gained popularity and yielded promising application possibilities in various fields from the food industry to biomedical applications. This review paper discussed the recent investigations on AI-based techniques for specific spectroscopic measurements and imaging approaches including NMR, MS, vibrational spectroscopy, X-ray, AFM, EM, and 2D chromatography. Each of these measurement techniques and the application tasks requires specific properties of the analysis methods. In that sense, the analysis methods are task and data-dependent and are discussed separately. However, the enhancement of the data quality is a common procedure in most of the reviewed studies. This enhancement is achieved by either applying inverse problems or preprocessing techniques. In this regard, deep learning techniques via CNN and LSTM became popular as solutions for the inverse problem for spectroscopic data. On the other hand, the modification of existing preprocessing techniques or their applications in new areas is a very common trend. Following the enhancement of data quality, data modeling for a variety of tasks was reviewed. For instance, the discrimination between different groups and the

quantification of a specific variable require either classification or regression methods. In the reviewed studies, PCA, PLS, SVM, SIMCA, ANN, PLSR, and deep learning were implemented for the spectroscopic measurements and the chemical data discussed herein (NMR, MS, vibrational spectroscopy, X-ray, AFM, EM, and 2D chromatography).

However, attempts to improve the predictive quality and robustness of artificial intelligence-based methods such as chemometrics, machine learning, and deep learning are performed in many application fields. Also, investigations to develop new AI-based techniques are increasing as well. Furthermore, strategies to overcome the lack of available data, especially in biomedical applications, and the advancement of data fusion methods are still subjects in further research.

## ACKNOWLEDGMENTS

The financial support from the Deutsche Forschungsgemeinschaft (DFG) via the CRC 1076 AquaDiva is highly appreciated.

## CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

## REFERENCES


1. Berry RJ, Ozaki Y. Comparison of wavelets and smoothing for denoising spectra for two-dimensional correlation spectroscopy. *Appl Spectrosc.* 2002;56(11):1462-1469. https://doi.org/10.1366/00037020260377779
2. Chen H, Xu W, Broderick N, Han J. An adaptive denoising method for Raman spectroscopy based on lifting wavelet transform. *J Raman Spectrosc.* 2018;49(9):1529-1539. https://doi.org/10.1002/jrs.5399
3. Guo S, Bocklitz T, Popp J. Optimization of Raman-spectrum baseline correction in biological application. *Analyst.* 2016;141(8):2396-2404. https://doi.org/10.1039/C6AN00041J
4. Shao L, Griffiths PR. Automatic baseline correction by wavelet transform for quantitative open-path fourier transform infrared spectroscopy. *Environ Sci Technol.* 2007;41(20):7054-7059. https://doi.org/10.1021/es062188d
5. Xi Y, Rocke DM. Baseline correction for NMR spectroscopic metabolomics data analysis. *BMC Bioinformatics.* 2008;9(1):324. https://doi.org/10.1186/1471-2105-9-324
6. Efitorov A, Burikov S, Dolenko T, Laptinskiy K, Dolenko S. Significant feature selection in neural network solution of an inverse problem in spectroscopy*. *Procedia Computer Science.* 2015;66:93-102. https://doi.org/10.1016/j.procs.2015.11.012
7. Dolenko SA, Burikov SA, Dolenko TA, Persiantsev IG. Adaptive methods for solving inverse problems in laser raman spectroscopy of multicomponent solutions. *Pattern Recognit Image Anal.* 2012;22(4):550-557. https://doi.org/10.1134/S1054661812040049
8. Sankaran S, Ehsani R. Visible-near infrared spectroscopy based citrus greening detection: Evaluation of spectral feature extraction techniques. *Crop Prot.* 2011;30(11):1508-1513. https://doi.org/10.1016/j.cropro.2011.07.005
9. Zhu C, Palmer GM, Breslin TM, Harter J, Ramanujam N. Diagnosis of breast cancer using diffuse reflectance spectroscopy: Comparison of a Monte Carlo versus partial least squares analysis based feature extraction technique. *Lasers Surg Med.* 2006;38(7):714-724. https://doi.org/10.1002/lsm.20356
10. Balabin RM, Smirnov SV. Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data. *Anal Chim Acta.* 2011;692(1):63-72. https://doi.org/10.1016/j.aca.2011.03.006
11. Li S, Chen G, Zhang Y, et al. Identification and characterization of colorectal cancer using Raman spectroscopy and feature selection techniques. *Opt Express, OE.* 2014;22(21):25895-25908. https://doi.org/10.1364/OE.22.025895
12. Faix O. Classification of Lignins from Different Botanical Origins by FT-IR Spectroscopy. *Holzforschung.* 1991;45(s1):21-28. https://doi.org/10.1515/hfsg.1991.45.s1.21
13. Geballe TR, Knapp GR, Leggett SK, et al. Toward Spectral Classification of L and T Dwarfs: Infrared and Optical Spectroscopy and Analysis. *Astrophys J.* 2002;564(1):466-481. https://doi.org/10.1086/324078
14. Nørgaard L, Saudland A, Wagner J, Nielsen JP, Munck L, Engelsen SB. Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. *Appl Spectrosc.* 2000;54(3):413-419. https://doi.org/10.1366/0003702001949500
15. Minasny B, McBratney AB. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. *Chemom Intell Lab Syst.* 2008;94(1):72-79. https://doi.org/10.1016/j.chemolab.2008.06.003
16. Chatzidakis M, Botton GA. Towards calibration-invariant spectroscopy using deep learning. *Sci Rep.* 2019;9(1):2126. https://doi.org/10.1038/s41598-019-38482-1
17. Kyathanahally SP, Döring A, Kreis R. Deep learning approaches for detection and removal of ghosting artifacts in MR spectroscopy. *Magn Reson Med.* 2018;80(3):851-863. https://doi.org/10.1002/mrm.27096
18. Mishra P, Biancolillo A, Roger JM, Marini F, Rutledge DN. New data preprocessing trends based on ensemble of multiple preprocessing techniques. *TrAC, Trends Anal Chem.* 2020;132(o):116045. https://doi.org/10.1016/j.trac.2020.116045
19. Torniainen J, Afara IO, Prakash M, Sarin JK, Stenroth L, Töyräs J. Open-source python module for automated preprocessing of near infrared spectroscopic data. *Anal Chim Acta.* 2020;1108:1-9. https://doi.org/10.1016/j.aca.2020.02.030
20. Martyna A, Menżyk A, Damin A, et al. Improving discrimination of Raman spectra by optimising preprocessing strategies on the basis of the ability to refine the relationship between variance components. *Chemom Intell Lab Syst.* Published online 2020:104029. https://doi.org/10.1016/j.chemolab.2020.104029
21. Roger J-M, Biancolillo A, Marini F. Sequential preprocessing through ORThogonalization (SPORT) and its application to near infrared spectroscopy. *Chemom Intell Lab Syst.* 2020;199:103975. https://doi.org/10.1016/j.chemolab.2020.103975
22. Zhang C, Zhou L, Zhao Y, Zhu S, Liu F, He Y. Noise reduction in the spectral domain of hyperspectral images using denoising autoencoder methods. *Chemom Intell Lab Syst.* Published online 2020:104063.
23. Raulf AP, Butke J, Menzen L, et al. Deep Neural Networks for the Correction of Mie Scattering in Fourier-Transformed Infrared Spectra of Biological Samples. arXiv preprint arXiv:200207681. Published online 2020.
24. Wahl J, Sjödahl M, Ramser K. Single-Step Preprocessing of Raman Spectra Using Convolutional Neural Networks. *Appl Spectrosc.* 2020;74(4):427-438.
25. Argoul P. Overview of Inverse Problems. 17.
26. Nakamura G. *Inverse Modeling An Introduction to the Theory and Methods of Inverse Problems and Data Assimilation.* IOP Publishing; 2015. https://doi.org/10.1088/978-0-7503-1218-9
27. Yuan Y, Demers H, Brodusch N, Wang X, Gauvin R. Inverse modeling for quantitative X-ray microanalysis applied to 2D heterogeneous materials. *Ultramicroscopy.* 2020;219:113117.
28. Hong S, Qin S, Dong P, et al. Quantification of rust penetration profile in reinforced concrete deduced by inverse modeling. *Cem Concr Compos.* Published online 2020:103622.
29. Takeda S, Hama T, Hsu H-H, et al. AI-driven Inverse Design System for Organic Molecules. *arXiv preprint arXiv:*200109038. Published online 2020.

30. Houhou R, Barman P, Schmitt M, Meyer T, Popp J, Bocklitz T. Deep learning as phase retrieval tool for CARS spectra. *Opt Express*. 2020;28(14):21002-21024.
31. Guo S, Mayerhöfer T, Pahlow S, Hübner U, Popp J, Bocklitz T. Deep learning for 'artefact'removal in infrared spectroscopy. *Analyst*. 2020;145(15):5213-5220.
32. Magnussen EA, Solheim JH, Blazhko U, et al. Deep convolutional neural network recovers pure absorbance spectra from highly scatter-distorted spectra of cells. *J Biophotonics*. n/a(n/a):e202000204. https://doi.org/10.1002/jbio.202000204
33. Hassenkam T, Fantner GE, Cutroni JA, Weaver JC, Morse DE, Hansma PK. High-resolution AFM imaging of intact and fractured trabecular bone. *Bone*. 2004;35(1):4-10. https://doi.org/10.1016/j.bone.2004.02.024
34. Li X, Dyck O, Unocic RR, Ievlev AV, Jesse S, Kalinin SV. Statistical learning of governing equations of dynamics from in-situ electron microscopy imaging data. *Materials & Design*. 2020;195:108973. https://doi.org/10.1016/j.matdes.2020.108973
35. The Basics of NMR. Accessed November 17, 2020. https://www.cis.rit.edu/htbooks/nmr/inside.htm
36. Pérez Y, Casado M, Raldúa D, et al. MCR-ALS analysis of 1 H NMR spectra by segments to study the zebrafish exposure to acrylamide. *Anal Bioanal Chem*. 2020;412(23):5695-5706.
37. Miros FN, Murch SJ, Shipley PR. Exploring feature selection of St John's wort grown under different light spectra using 1H-NMR spectroscopy. *Phytochem Anal*. Published online 2020.
38. Zhao F, Li W, Pan J, Chen Z, Qu H. A novel critical control point and chemical marker identification method for the multi-step process control of herbal medicines via NMR spectroscopy and chemometrics. *RSC Advances*. 2020;10(40):23801-23812.
39. Marion R, Govaerts B, von Sachs R. AdaCLV for Interpretable Variable Clustering and Dimensionality Reduction of Spectroscopic Data.; 2020.
40. Coimbra PT, Bathazar CF, Guimarães JT, et al. Detection of formaldehyde in raw milk by time domain nuclear magnetic resonance and chemometrics. *Food Control*. 2020;110:107006.
41. Zhang J, Chen H, Fan C, Gao S, Zhang Z, Bo L. Classification of the botanical and geographical origins of Chinese honey based on 1H NMR profile with chemometrics. *Food Res Int*. 2020;137:109714.
42. Kong X, Zhou L, Li Z, et al. Artificial intelligence enhanced two-dimensional nanoscale nuclear magnetic resonance spectroscopy. *npj Quantum Information*. 2020;6(1):1-10.
43. Hou X, Wang G, Wang X, Ge X, Fan Y, Nie S. Convolutional neural network based approach for classification of edible oils using low-field nuclear magnetic resonance. *J Food Compos Anal*. Published online 2020:103566.
44. Rajawat J, Jhingan G. Chapter 1 - Mass spectroscopy. In: Misra G, ed. *Data Processing Handbook for Complex Biological Data Sources*. Academic Press; 2019:1-20. https://doi.org/10.1016/B978-0-12-816548-5.00001-0
45. Duan L, Ma A, Meng X, Shen G, Qi X. QPMASS: A parallel peak alignment and quantification software for the analysis of large-scale gas chromatography-mass spectrometry (GC-MS)-based metabolomics datasets. *J Chromatogr A*. Published online 2020:460999.
46. Alkhalifah Y, Phillips I, Soltoggio A, et al. VOCCluster: Untargeted Metabolomics Feature Clustering Approach for Clinical Breath Gas Chromatography/Mass Spectrometry Data. *Anal Chem*. 2019;92(4):2937-2945.
47. Papagiannopoulou C, Parchen R, Rubbens P, Waegeman W. Fast Pathogen Identification Using Single-Cell Matrix-Assisted Laser Desorption/Ionization-Aerosol Time-of-Flight Mass Spectrometry Data and Deep Learning Methods. *Anal Chem*. 2020;92(11):7523-7531.
48. Li K, Jain A, Malovannaya A, Wen B, Zhang B. DeepRescore: Leveraging Deep Learning to Improve Peptide Identification in Immunopeptidomics. *Proteomics*. Published online 2020:1900334.
49. Vibrational Spectroscopy: Definition & Types - Video & Lesson Transcript. Study.com. Accessed November 17, 2020. https://study.com/academy/lesson/vibrational-spectroscopy-definition-types.html
50. Akpolat H, Barineau M, Jackson KA, et al. High-Throughput Phenotyping Approach for Screening Major Carotenoids of Tomato by Handheld Raman Spectroscopy Using Chemometric Methods. *Sensors*. 2020;20(13):3723.
51. Han N, Ram RJ. Bayesian modeling and computation for analyte quantification in complex mixtures using Raman spectroscopy. *Computational Statistics & Data Analysis*. 2020;143:106846.
52. Wang K, Chen L, Ma X, et al. Arcobacter Identification and Species Determination Using Raman Spectroscopy Combined with Neural Networks. *Appl Environ Microbiol*. 2020;86(20).
53. Kirchberger-Tolstik T, Pradhan P, Vieth M, et al. Towards an interpretable classifier for characterization of endoscopic Mayo scores in ulcerative colitis using Raman Spectroscopy. *Anal Chem*. Published online 2020.
54. Zafar A, Hong K-S. Reduction of onset delay in functional near-infrared spectroscopy: prediction of HbO/HbR signals. *Frontiers in Neurorobotics*. 2020;14:10.
55. Galán-Freyle NJ, Ospina-Castro ML, Medina-González AR, Villarreal-González R, Hernández-Rivera SP, Pacheco-Londoño LC. Artificial intelligence assisted Mid-infrared laser spectroscopy in situ detection of petroleum in soils. *Applied Sciences*. 2020;10(4):1319.
56. Le BT. Application of deep learning and near infrared spectroscopy in cereal analysis. *Vib Spectrosc*. 2020;106:103009. https://doi.org/10.1016/j.vibspec.2019.103009
57. Xu L, Liu Y, Yu J, et al. Characterizing autism spectrum disorder by deep learning spontaneous brain activity from functional near-infrared spectroscopy. *J Neurosci Methods*. 2020;331:108538.
58. X-ray - Fundamental characteristics. Encyclopedia Britannica. Accessed November 17, 2020. https://www.britannica.com/science/X-ray
59. Otsuka Y, Utsunomiya Y, Umeda D, Yonemochi E, Kawano Y, Hanawa T. Effect of polymers and storage relative humidity on amorphous rebamipide and its solid dispersion transformation: Multiple spectra chemometrics of powder X-Ray diffraction and near-infrared spectroscopy. *Pharmaceuticals*. 2020;13(7):147.
60. Allegretta I, Marangoni B, Manzari P, et al. Macro-classification of meteorites by portable energy dispersive X-ray fluorescence spectroscopy (pED-XRF), principal component analysis (PCA) and machine learning algorithms. *Talanta*. 2020;212:120785.
61. Carbone MR, Topsakal M, Lu D, Yoo S. Machine-Learning X-Ray Absorption Spectra to Quantitative Accuracy. *Phys Rev Lett*. 2020;124(15):156401. https://doi.org/10.1103/PhysRevLett.124.156401
62. Mullaliu A, Aquilanti G, Conti P, Giorgetti M, Passerini S. Effect of Water and Alkali-Ion Content on the Structure of Manganese (II) Hexacyanoferrate (II) by a Joint Operando X-ray Absorption Spectroscopy and Chemometric Approach. *ChemSusChem*. 2020;13(3):608-615.
63. AFM Microscopes - Introduction to Atomic Force Microscopy | Bruker. Bruker.com. Accessed November 17, 2020. https://www.bruker.com/products/surface-and-dimensional-analysis/atomic-force-microscopes/campaigns/afm-microscopes.html
64. *Electron Microscopy introduction*. WUR. Published March 21, 2014. Accessed November 17, 2020. https://www.wur.nl/en/Value-Creation-Cooperation/Facilities/Wageningen-Electron-Microscopy-Centre/Electron-Microscopy-intro.htm
65. *Machine learning to enhance atomic force microscopy analysis and operation*. Wiley Analytical Science. Accessed November 10, 2020. https://analyticalscience.wiley.com/do/10.1002/was.00010012

66. Javazm MR, Pishkenari HN. Observer Design for Topography Estimation in Atomic Force Microscopy Using Neural and Fuzzy Networks. *Ultramicroscopy*. Published online 2020:113008.
67. Payam AF, Biglarbeigi P, Morelli A, Lemoine P, McLaughlin J, Finlay D. Data acquisition and imaging using wavelet transform: a new path for high speed transient force microscopy. *Nanoscale Advances*. Published online 2020.
68. Yu Q, Xiong Z, Du C, et al. Identification of rock pore structures and permeabilities using electron microscopy experiments and deep learning interpretations. *Fuel*. 2020;268:117416.
69. Kilz P. Two-Dimensional Chromatography as an Essential Means for Understanding Macromolecular Structure. *Chromatographia*. 2004;59(1):3-14. https://doi.org/10.1365/s10337-003-0106-7
70. Huygens B, Efthymiadis K, Nowé A, Desmet G. Application of evolutionary algorithms to optimise one-and two-dimensional gradient chromatographic separations. *J Chromatogr A*. 2020;1628:461435.
71. Pérez-Cova M, Tauler R, Jaumot J. Chemometrics in comprehensive two-dimensional liquid chromatography: A study of the data structure and its multilinear behavior. *Chemom Intell Lab Syst*. Published online 2020:104009.
72. Nagai K, Uranbileg B, Chen Z, et al. Identification of novel biomarkers of hepatocellular carcinoma by high-definition mass spectrometry: Ultrahigh-performance liquid chromatography quadrupole time-of-flight mass spectrometry and desorption electrospray ionization mass spectrometry imaging. *Rapid Commun Mass Spectrom*. 2020;34(S1):e8551. https://doi.org/10.1002/rcm.8551

**How to cite this article:** Houhou R, Bocklitz T. Trends in artificial intelligence, machine learning and chemometrics applied to chemical data. *Anal Sci Adv*. 2021;2:128–141. https://doi.org/10.1002/ansa.202000162

# P2. DEEP LEARNING AS PHASE RETRIEVAL TOOL FOR CARS SPECTRA

Reprinted, with permission, from [R. Houhou, P. Barman, M. Schmitt, T. Meyer, J. Popp, and T. Bocklitz, Deep learning as phase retrieval tool for CARS spectra, Opt. Express, 2020, 28, 21002-21024]. Copyright 2020 Optics Express.

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

| **R. Houhou[1], P. Barman[2], M. Schmitt[3], T. Meyer[4], J. Popp[5], and T. Bocklitz[6], deep learning as phase retrieval tool for CARS spectra, Opt. Express, 2020, 28, 21002-21024** | | | | | | |
|---|---|---|---|---|---|---|
| Involved in | 1 | 2 | 3 | 4 | 5 | 6 |
| | | | | | | |
| Conceptual research design | × | | × | | × | × |
| Planning of research activities | × | | | | | × |
| Data collection | | × | | × | | |
| Data analyses and interpretation | × | × | × | × | × | × |
| Manuscript writing | × | × | × | × | × | × |
| Suggested publication equivalence value | 1.0 | | | | | |

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21002
Optics EXPRESS

# Deep learning as phase retrieval tool for CARS spectra

ROLA HOUHOU,[1,2] PARIJAT BARMAN,[2] MICHEAL SCHMITT,[1] TOBIAS MEYER,[1,2] JUERGEN POPP,[1,2] AND THOMAS BOCKLITZ[1,2,*]

[1]Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich-Schiller-University, Helmholtzweg 4, 07743 Jena, Germany
[2]Leibniz Institue of Photonic Technology, Albert-Einstein-straße 9, 07745 Jena, Germany
*thomas.bocklitz@uni-jena.de

**Abstract:** Finding efficient and reliable methods for the extraction of the phase in optical measurements is challenging and has been widely investigated. Although sophisticated optical settings, e.g. holography, measure directly the phase, the use of algorithmic methods has gained attention due to its efficiency, fast calculation and easy setup requirements. We investigated three phase retrieval methods: the maximum entropy technique (MEM), the Kramers-Kronig relation (KK), and for the first time deep learning using the Long Short-Term Memory network (LSTM). LSTM shows superior results for the phase retrieval problem of coherent anti-Stokes Raman spectra in comparison to MEM and KK.

Published by The Optical Society under the terms of the Creative Commons Attribution 4.0 License. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

## 1. Introduction

Spontaneous Raman scattering is an inelastic scattering based technique that describes information about the different vibrational, rotational and other transitions in the molecules [1]. Spontaneous Raman spectra are measured under the assumption of a weak electric field, where the molecular polarization $P$ is directly proportional to the electric field $E$ as follows $P = \epsilon_0 \chi E$. The symbols $\epsilon_0$ and $\chi$ represent the electric permittivity in vacuum and the molecular susceptibility, respectively [2]. Although spontaneous Raman scattering is a label-free, molecular fingerprint technique; the weak Raman spectroscopic signals in combination with fluorescence contributions of biological samples and the resulting long acquisition times are reducing the spread of its applications [3].

To overcome the aforementioned drawbacks, increasing the electric field strength can be considered as a solution. Accordingly, the induced polarization is expressed as Taylor series $P = \epsilon_0(\chi^{(1)}E^{(1)} + \chi^{(2)}E^{(2)} + \chi^{(3)}E^{(3)} + \cdots)$, where $\chi^{(n)}$ is the $n^{th}$ order susceptibility [2]. Our study focuses on the coherent Raman scattering (CRS) which results from third order polarization $P^{(3)} = \chi^{(3)}E^{(3)}$, in particular, coherent anti-Stokes Raman scattering (CARS). CARS produces much stronger vibrational sensitive signals than spontaneous Raman scattering for high concentration samples [4]. Therefore, it is useful as a diagnostic tool for determining the presence of chemical species using their Raman vibrational modes. Additionally, it can be applied to probe molecular structures and their environment, because the resonance frequency and relaxation rate often depend upon the molecular environment [2]. Due to these reasons, CARS is considered as an ideal tool for enhancing the Raman signals and reducing the acquisition time. These properties lead to first the focus on the development of CARS microscopy e.g. multiplex [5], broadband [6] and vibrational phase contrast CARS [7]. In addition to its advancement into the imaging of biomedical samples where a large number of publications deal with [8–12].

However, CARS faces a main challenge; the non-resonant background (NRB). This contribution distorts the vibrational signals as it is coherent with the CARS signal and leads to constructive and destructive interference effects [13]. By experimentally reducing the NRB generation, the CARS

#390413 https://doi.org/10.1364/OE.390413
Journal © 2020 Received 11 Feb 2020; revised 23 Apr 2020; accepted 24 Apr 2020; published 1 Jul 2020

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21003
Optics EXPRESS

spectral intensities will also dramatically reduce thus cancelling any advantage of the CARS measurements [14]. Therefore, the search for methods that extract the phase without physically removing the non-resonant background is essential. In this context, algorithmic phase retrieval methods have been widely explored. Most of these methods are based on iterative algorithms like the Gerchberg-Saxton algorithms derived from an error-reduction algorithm. This methods was first suggested for phase determination in image and diffraction plane images [15]. Another method which was frequently used for the phase retrieval problems is the maximum entropy method (MEM). MEM is based on the maximum entropy of the probability distribution from the available information. Accordingly, Vartiainen et al. investigated MEM in various papers; using the squared modulus spectra of substances like Polysilane [16], reflection data [17] and an improved version of MEM for reflection data [18]. Additionally, the Kramers-Kronig relation (KK) was popular as well and it is based on the causality of the physical process. This condition results in a certain relation between real and imaginary part of many physical quantities as they are represented by analytical functions. Grosse et al. [19] presented an analysis of a reflectance data using KK. While, Kuzmelo et al. [20] focuses on the KK constrained for optical spectra. Comparisons between KK and MEM for resolving the phase were also investigated by Gornov et al. [21] and Cicerone et al. [22]. Additionally, other algorithms such as basic input-output and hybrid input-output convex projection-based methods were tested by Bauschke et al. [23].

MEM and KK require an accurate measurement of the non-resonant background [13] and as far as we know, little research has been conducted on the behavior of these models when the strength of the NRB varies. In this manner, we want to evaluate the deep learning technique for the CARS spectra by varying the NRB conditions with comparison to MEM and KK. However, the extraction of the phase using deep learning was widely investigated in other fields. For digital holography Zhang et al. [24] used a U-type convolutional neural network (U-net) to recover the original phase of the microscopic images. Sinha et al. [25] built and tested a lensless imaging system where a deep neural network was trained and utilized to recover the phase of the objects. Goy et al. [26] demonstrated a superior performance of deep neural networks to recover objects comparing to the Gerchberg-Saxton phase retrieval algorithm.

We explore in this paper MEM and KK as two popular phase retrieval methods. We discuss their mathematical background and applications problems in section 2. Then, we present an alternative solution for the phase retrieval by means of deep learning using the LSTM network and corresponding details are described in section 3. Then, the impact of the strengths of the non-resonant background in the application of the three methods is tested using simulated data, which is shown in subsection 4.1. Then, we explore these three methods on two experimental BCARS spectra and the results are summarized in subsection 4.2. Finally, a conclusion is drawn in section 5.

## 2. MEM and KK: mathematical background

Theoretically, the CARS intensity $I_{CARS}$ is directly proportional to the squared modulus of the nonlinear susceptibility $|\chi^{(3)}|^2$ [27] and can be described as follows:

$$I_{CARS} \propto |\chi^{(3)}|^2 I_{pu}^2 I_s \,, \tag{1}$$

where $I_{pu}$ and $I_s$ are the pump and the Stokes intensity, respectively. The nonlinear susceptibility $\chi^{(3)}$ is the sum of a non-resonant part (NRB) $\chi_{nr}^{(3)}$ that appears due to the electronic contributions and a Raman resonant part $\chi_r^{(3)}$, which can be written as follows:

$$\chi^{(3)} = \chi_{nr}^{(3)} + \chi_r^{(3)} \,. \tag{2}$$

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21004
Optics EXPRESS

The non-resonant part is purely real and shows only a weak frequency dependency, while the resonant part is a complex function that can be described as Lorentz function:

$$\chi_r^{(3)} = \sum_r \frac{A_r}{\Omega_r - (\omega_{pu} - \omega_s) - i\gamma_r} , \tag{3}$$

where $(\omega_{pu} - \omega_s)$ is the difference of pump and Stokes frequency and $A_r$, $\Omega_r$ and $\gamma_r$ are the amplitude, the vibrational frequency and the bandwidth of the $r^{\text{th}}$ Raman mode, respectively. Moreover, the imaginary component of the resonant part is of interest as it holds the same information as the spontaneous Raman signal. However, this term is not directly measured, but determined in combination in the CARS spectrum [28]. More precisely, in CARS measurements, both the real and the imaginary components are unknown and only the squared modulus $S(\nu) = |\chi^{(3)}(\nu)|^2$ is available. Therefore, the use of mathematical methods that retrieve the phase $\varphi(\nu)$ from the modulus of the nonlinear susceptibility $|\chi^{(3)}(\nu)|$, related via

$$\chi^{(3)}(\nu) = |\chi^{(3)}(\nu)| \exp(i\varphi(\nu)) , \tag{4}$$

is crucial and was tested in this paper. As mentioned previously, a variety of methods are available for phase retrieval, but we focused on the application of the Maximum Entropy Method (MEM) and the Kramers-Kronig relation (KK). The application of these methods is different and a brief explanation of the methods is given in the following.

### 2.1. Maximum Entropy Method (MEM)

MEM is a probability-based technique, where the power spectrum $S(\nu)$ can be approximated on a defined normalized frequency $\nu$ range, described as

$$S(\nu) \approx \frac{|a_0|^2}{|1 + \sum_{k=1}^{M} a_k \exp(-2i\pi\nu k)|^2} , \tag{5}$$

where $a_0$, $a_k$ for $1 \le k \le M$ and $M$ represent the coefficients and the number of poles of the approximation, respectively. Then, by using Wiener-Khinchin theorem, which states that the power spectrum is equal to the Fourier transform of the autocorrelation function $R_j$ , we can write the following approximation:

$$\frac{|a_0|^2}{|1 + \sum_{k=1}^{M} a_k \exp(-2i\pi\nu k)|^2} \approx \sum_{j=-M}^{M} R_j \exp(-2i\pi\nu j) , \tag{6}$$

where the autocorrelation functions $R_j$ for $1 \le k \le M$ are obtained in Eq. (7) by applying the Fourier transform on the power spectrum:

$$R_j = \int_0^1 S(\nu) \exp(2i\pi\nu j) \mathrm{d}\nu; |j| \le M. \tag{7}$$

The coefficients $a_0$ and $a_k$ are estimated by solving the following Toeplitz matrix:

$$\begin{pmatrix} R_0 & R_{-1} & \dots & R_{-M} \\ R_1 & R_0 & \dots & R_{1-M} \\ \vdots & \vdots & \ddots & \vdots \\ R_M & R_{M-1} & \dots & R_0 \end{pmatrix} \begin{pmatrix} 1 \\ a_1 \\ \vdots \\ a_M \end{pmatrix} = \begin{pmatrix} |a_0|^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix} . \tag{8}$$

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21005
Optics EXPRESS

Afterwards, the estimated coefficients $|a_0|^2$ and $a_k$ for $1 \leq k \leq M$ are used to calculate $\chi^{(3)}$ as follows:

$$\chi^{(3)} = \frac{a_0}{1 + \sum_{k=1}^{M} a_k \exp(-2i\pi\nu k)}. \tag{9}$$

Since we calculated the squared modulus of $|a_0|^2$, the problem of finding the real and the imaginary parts of $\chi^{(3)}$ is reduced to finding the real and the imaginary part of $a_0$. As $a_0$ is constant and independent of $\nu$, the problem can be solved at a specific frequency position $\nu_0$. Two physically valid restrictions are available. If $\nu_0$ is far from any resonances;

$$\mathrm{Im}[\chi^{(3)}(\nu_0)] = 0, \tag{10}$$

then the constant phase $\varphi_0$ is defined as follows:

$$\begin{aligned} \tan\varphi_0 =& \mathrm{Im}(a_0)/\mathrm{Re}(a_0) \\ =& \mathrm{Im}(1 + \sum_{k=1}^{M} a_k \exp(-2i\pi\nu_0 k))/\mathrm{Re}(1 + \sum_{k=1}^{M} a_k \exp(-2i\pi\nu_0 k)) \,. \end{aligned} \tag{11}$$

Another restriction can be defined as the imaginary part $\mathrm{Im}(\chi^{(3)})$ has a local maximum at a resonance:

$$\frac{\partial}{\partial\nu}\mathrm{Im}[\chi^{(3)}(\nu)]_{\nu=\nu_r} = 0. \tag{12}$$

Then $\varphi_r$ is defined as follows:

$$\tan\varphi_r = \frac{\partial_\nu[\mathrm{Im}(1 + \sum_{k=1}^{M} a_k \exp(-2i\pi\nu k))|1 + \sum_{k=1}^{M} a_k \exp(-2i\pi\nu k)|^{-2}]_{\nu=\nu_r}}{\partial_\nu[\mathrm{Re}(1 + \sum_{k=1}^{M} a_k \exp(-2i\pi\nu k))|1 + \sum_{k=1}^{M} a_k \exp(-2i\pi\nu k)|^{-2}]_{\nu=\nu_r}} \,. \tag{13}$$

Subsequently, we can approximate the real and the imaginary part of $\chi^{(3)}$ using one of the restrictions mentioned either in Eq. (10) or in Eq. (12) [29]. The major drawback of the MEM method is that it requires an estimation of an error phase spectrum, which should be further removed from the retrieved MEM phase. However, this estimation involves a priori knowledge and in case of its absence, the method does not work properly. Additionally, a prerequisite for the method is the choice of the number of poles $M$ which is fundamental and selected using a trial and error process. This choice effects dramatically the approximation quality. Therefore, $M$ should be large enough to accurately reproduce the real and the imaginary components and adequately small to abandon any noises. In this context, the application of some preprocessing methods e.g. de-noising and baseline correction are crucial.

### 2.2. Kramers-Kronig relation (KK)

KK is a mathematical relationship between the real and the imaginary part of a complex function based on the Cauchy residue theorem. Therefore, we can calculate the imaginary part, if we know the real part, and vice versa as shown in Eq. (14) [13]:

$$\begin{aligned} \mathrm{Re}(\chi(i\nu)) &= \tfrac{-1}{\pi}\wp\textstyle\int_{-\infty}^{\infty} \frac{\mathrm{Im}(\chi(ix))}{x-\nu}\mathrm{d}x \\ \mathrm{Im}(\chi(i\nu)) &= \tfrac{1}{\pi}\wp\textstyle\int_{-\infty}^{\infty} \frac{\mathrm{Re}(\chi(ix))}{x-\nu}\mathrm{d}x \,, \end{aligned} \tag{14}$$

where $\wp$ is the Cauchy principle value. By taking the logarithm on both sides of Eq. (4) and then using the KK relation displayed in Eq. (13) we can deduce the phase from the squared modulus as follows:

$$\varphi(\nu) = -\frac{1}{\pi}\wp\int \frac{\ln(\sqrt{S(\nu')})}{\nu' - \nu}\mathrm{d}\nu' \,. \tag{15}$$

Hence, the phase was extracted by applying a discrete Hilbert transform on $\ln(\sqrt{S(\nu)})$. One of the drawbacks of KK method is the infinite behavior of $\ln(\sqrt{S(\nu)})$ which rises the problem of

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21006
Optics EXPRESS

extrapolation beyond the range of measurement. Similar to the MEM method, the KK method also requires an estimation of an error phase spectrum which involves a priori knowledge and resulted in phase distortion in case of its absence. Accordingly, various modifications to the KK methods were investigated. Furthermore, some preprocessing steps, e.g. de-noising, phase de-trending and baseline correction, are essential in order to reach an accurate reconstructed spectrum.

## 3. Deep learning: an alternative solution

Deep learning is a particular field of machine learning that is often implemented using neural network geometries. It works by easy processing units called layers, which are stacked. Deep learning extract useful information through these layers by optimizing their internal parameters via the backpropagation algorithm [30]. As in machine learning methods, an input is fed and an output is produced and both are controlled by using an optimizer and a loss function, respectively. These optimizer and loss function represent the core of the deep learning technique and work successively in such a way that the scores derived from the loss function are feed back to the network in order to adjust the weights by means of the optimizer [31].

Deep learning was able to overcome other machine learning methods in terms of performance in spite of being a black box tool which lacks knowledge of internal network working. However, it has a major drawback; its computation time which is significantly large. Despite this, the scope of its application expanded into diverse fields. In particular, it was widely applied in spectroscopy especially for classification and pattern recognition problems [32–34].

Due to the excellent performances achieved by deep learning in other fields [35–38], we want to test how the deep learning performs in the phase retrieval task. So, we used it as an indirect process to extract the imaginary component from the CARS squared modulus. The procedure is different than MEM and KK; we build first a model using the Long Short-Term Memory network. This network maps between the squared modulus and its imaginary component based on artificial data. Then, we use this model to predict the imaginary part or Raman like spectrum from the squared modulus of the testing set.

A large number of deep learning networks are available, but we implemented a network that deals with sequential data, and which has the same architecture as recurrent neural networks (RNNs). RNNs are gradient-based learning algorithm, which are used to predict sequential data. However, RNNs face a major challenge; it is not able to connect information when the input sequence range increases [39,40]. Thus, for long term dependencies, the gradient in RNNs will either vanish or explode. To solve this problem, the Long Short-Term Memory network (LSTM) was introduced [40,41]. In LSTM, four neural network layer; the input gate, the forget gate, the output gate and the cell state interacts in a particular way, which is described below. However, the core of this network is the cell state $C_t$. The cell state is linearly connected to the other layers and runs through the whole sequence. The process of the deep learning network using LSTM is shown in Fig. 1.

First, in LSTM, the output from a previous state $h_{t-1}$ and the current input vector $x_t$ are fed into the forget gate $F_t$. Its output i.e. the forget value $f_t$ is calculated as follows; 0 if no information is taken into consideration and 1 if all information is considered. And this is translated using the following equation:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)\,, \tag{16}$$

where $\sigma$ is a sigmoid function. $W_l$ and $b_l$ are the weight matrices and the bias vector parameters for each of the four layers, respectively. The subscript $l \in \{f, i, o, C\}$ characterized the layer. Then, we need to decide what information we want to store in the cell state. In this manner, two steps should be implemented. First, via the input gate layer $I_t$, a sigmoid function is used to decide on the values that we need to update. This layer $I_t$ is fed by two values; the current input

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21007
Optics EXPRESS

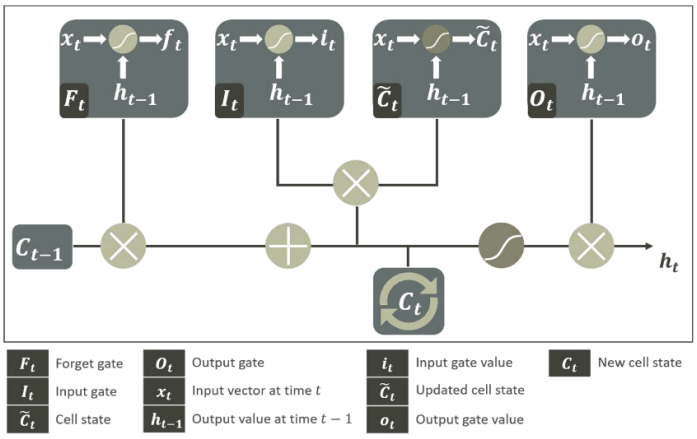


**Fig. 1.** The diagram of the Long Short-Term Memory network. LSTM is formed using four main gates; the input gate, the forget gate, the output gate and the cell state. These gates are connected in a particular way to learn the long term dependencies. See text for details.

vector $x_t$ and the output from previous state $h_{t-1}$ and it results with an input gate values $i_t$ as shown in Eq. (17):

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)\,. \tag{17}$$

Further, a tanh layer creates a vector of new candidate values $\widetilde{C}_t$ that could be added to the cell state using Eq. (18):

$$\widetilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C)\,. \tag{18}$$

Additionally, we need to use the old cell state $C_{t-1}$ and the updated cell state $\widetilde{C}_t$ to calculate the new cell state $C_t$. This is accomplished by linearly connecting both with the forget gate value $f_t$ and the input gate value $i_t$ using Eq. (19):

$$C_t = C_{t-1}f_t + i_t\widetilde{C}_t\,. \tag{19}$$

Finally, we need to decide what we want to output. Therefore, first we run a sigmoid function that decides which parts of the cell state we need to output and this is the output value $o_t$ from the output gate $O_t$. Then, a hyperbolic tangent tanh was implemented on the cell state and multiply it with the output from $O_t$ in order to only output the parts that we decided to as follows:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{20}$$

and

$$h_t = o_t \tanh C_t. \tag{21}$$

The aforementioned network is the simple LSTM nevertheless different versions are available [42]. The differences between these networks are minor but some of them are popular e.g. including "peephole connections" that let the gate layers look at the cell state which was introduced by Gers and Schmidhuber [43]. A comparison has been made between popular variants of LSTM by Greff et al. where the author finds that the compared LSTMs are similar [44]. However, in this paper we used the simple version of LSTM.

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21008
Optics EXPRESS

## 4. Results

### 4.1. Artificial data

In this paper, we focused on comparing the deep learning network LSTM as phase retrieval tool to MEM and KK models (displayed in Fig. 2) with respect to the strength of the non-resonant background. Therefore, we considered two scenarios. In scenario 1, the Raman resonances were constructed only in the region where the NRB is strong. In scenario 2, the Raman resonances were built in both the strong and the weak NRB regions. These both scenarios were chosen to determine the dependency of the three phase retrieval techniques to the NRB strength.

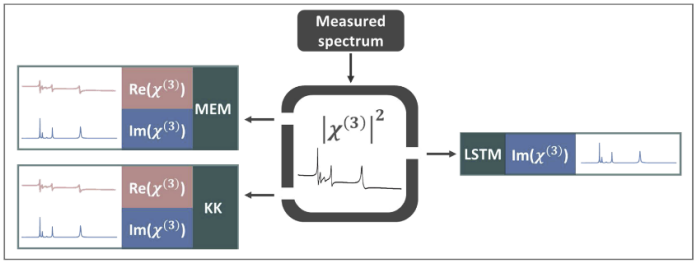


**Fig. 2.** The workflow of the three phase retrieval methods. Both MEM and KK extract both the real $\mathrm{Re}(\chi^{(3)})$ and the imaginary $\mathrm{Im}(\chi^{(3)})$ component from the squared modulus of the susceptibility $|\chi^{(3)}|^2$. In contrast, LSTM network predicts only the imaginary component $\mathrm{Im}(\chi^{(3)})$ using the squared modulus of the susceptibility $|\chi^{(3)}|^2$.

Initially, we construct simulated CARS spectra using Eq. (3) and this simulated data is used to assess the performance of the three phase retrieval techniques. The parameters used for the construction of the simulated spectra were chosen randomly (see Table 3). As shown in Eq. (3), the number of resonances needed to be defined a priori and the number was selected between 4 and 7. A range of amplitudes and bandwidths for the different resonances was defined between 0.001 and 0.5 and between 0.0001 and 0.01, respectively. The non-resonant background is defined as a constant value $\chi_{nr} = 0.5$ multiplied by a Gaussian function. In addition, we added noise in order to reflect the experimental conditions. As a result, the total sample size of the simulated CARS spectra was 4000.

Then, we used these simulated CARS spectra and applied all three phase retrieval methods. For MEM, we used an in-house written R algorithm. A trial and error process was implemented to determine the optimal number of poles which was equal to $M = 150$. Finally, Eq. (9) was used to extract the phase, where $v_0 = 0.6$ (refer to Table 3). A preprocessing step was necessary to remove the oscillations at the edges and this was done by replicating two small sub-regions on both edges and then removing them after the reconstruction of the spectra.

Afterwards, we applied the KK algorithm on the simulated CARS spectra, which was implemented in Python by Camp et al. [13].

Finally, we tested the deep learning technique using the LSTM network. In contrast to MEM and KK, the procedure of extracting the phase using the LSTM network is indirect. We used the network to map between the squared modulus and the imaginary components of the 4000 simulated CARS spectra. We split the squared modulus spectra into $\frac{2}{5}$ training set and $\frac{1}{5}$ testing set, where 20% of the training set were used as a validation set. First, we trained our network by tuning its hyperparametres (refer to Table 3). The Adam optimizer was used to update the parameters of the network and the maximum number of training epochs were set to 50. The

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21009
Optics EXPRESS

found optimal learning rate, optimal training batch size, and the optimal number of hidden units in the layer were found to be 0.005, 10 and 30, respectively. Then, we used the trained network in the prediction step on the testing set.

In scenario 1, four resonances were built in the strong NRB region. Using the MEM method, the constructed squared modulus fits well the theoretical one as shown in the first row of Fig. 3. On the other hand, the MEM algorithm was able to predict the peak like shape in the constructed imaginary component displayed in the second row of Fig. 3. The MEM based phase retrieval algorithm estimates correctly the position and the width of the peaks. However, it shows distortion in the amplitude and the shape in particular in the third peak which might be resulted from the presence of the background. In addition, the MEM algorithm constructs the real component shown in the third row of Fig. 3. The constructed real component has a dispersive line shape in the four positions and resembles to the theoretical real component. However, similarly to the constructed imaginary component, they were distorted by the presence of the background.

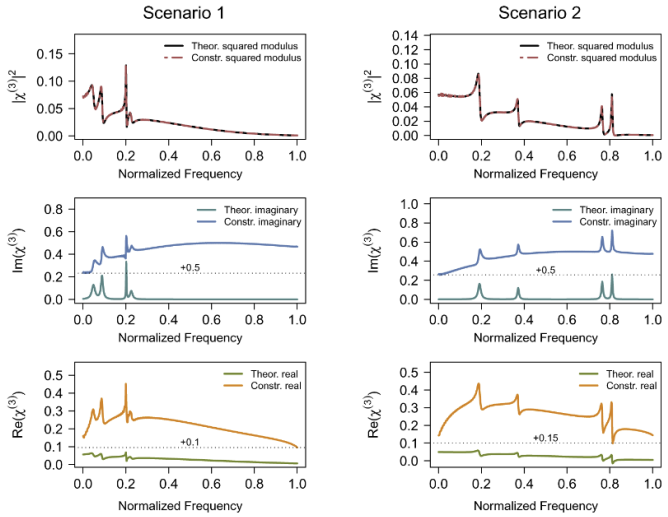


**Fig. 3.** Output of MEM applied on simulated CARS spectra in scenario 1 and 2 on the left and the right, respectively. In the second row, the constructed and the theoretical imaginary components for both scenarios are displayed in the upper and bottom part, respectively. In the third row the constructed and the theoretical real components for both scenarios are shown in the upper and bottom section, respectively. The dotted line in black represents the offset value.

For the same scenario (scenario 1), the results of the KK algorithm are shown in Fig. 4. In the first row, the constructed squared modulus is shown, which fits perfectly the theoretical one. If the constructed imaginary component and the theoretical one are compared, the KK algorithm was able to estimate the peak like shape for the four resonances. In addition, the position, width and amplitude of the peaks represent well the theoretical one. Additionally, the constructed real component of the KK algorithm showed a dispersive line shape in the four

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21010
Optics EXPRESS

positions. However, the background presence persist in both the imaginary and real components, which can be extracted afterwards using a baseline correction technique. [45]

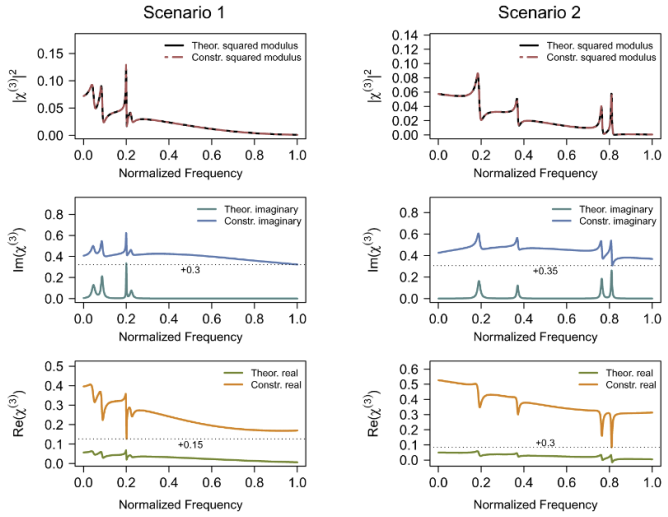


**Fig. 4.** Output of KK applied on simulated CARS spectra in scenario 1 and 2 on the left and the right, respectively. In the second row, the constructed and the theoretical imaginary components for both scenarios are displayed in the upper and bottom lines, respectively. In the third row, the constructed and the theoretical real components for both scenarios are shown in the upper and bottom lines, respectively. The dotted line in black represents the offset value.

Using our trained LSTM network, we predicted the imaginary component of the same spectra from scenario 1 as shown in the left part of Fig. 5. The LSTM network was able to correctly construct the peak like shape for the four resonances. The peak positions and widths were perfectly reflecting the theoretical ones. But, the amplitude for the first and fourth peak were smaller than those in the theoretical imaginary component. This might be a result of the parameter used to train the network. However, the LSTM method was able to remove completely the background.

In scenario 2, two resonances were built in the strong NRB region and two resonances in the weak NRB region. Using the MEM method, the constructed squared modulus fits well the theoretical one as shown in the first row of Fig. 3. Moreover, the MEM algorithm predicted the peak like shape in the constructed imaginary component in all four positions as shown in the second row of Fig. 3. It estimates correctly the peak positions, widths and amplitudes. Regarding the constructed real component, the MEM technique accurately estimated the dispersive line shape in the four positions displayed in the third row of Fig. 3. However, the constructed imaginary and real components were both distorted by the presence of the background.

The output of the KK algorithm on the spectrum from scenario 2 is shown in Fig. 4. The constructed squared modulus fits perfectly the theoretical one as shown in the first row of Fig. 4. In spite of this, the constructed imaginary component shows a peak like shape in the first two

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21011
Optics EXPRESS

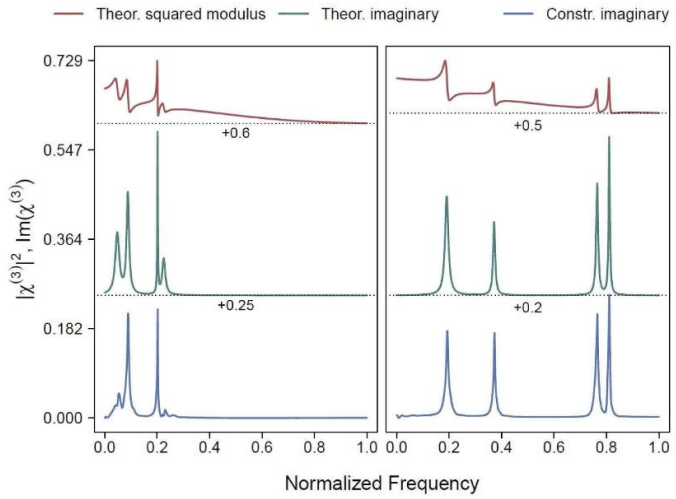


**Fig. 5.** Prediction of the imaginary components of the simulated spectra using the trained LSTM model in both scenarios. The theoretical squared modulus are shown in the upper line. The constructed and the theoretical imaginary components for both scenarios are displayed in the bottom and middle part, respectively. The dotted line in black represents the offset value.

resonances but a dispersive line shape in the weak NRB region as shown in the second row of Fig. 4. Similarly, the constructed real component shows a dispersive line shape in the first two resonances but a negative peak like shape in the weak NRB region which is displayed in the third row of Fig. 4. Additionally, the background remains in both the imaginary and the real components, which can be extracted afterwards using a baseline correction technique. [45]

Using the trained LSTM network, we predicted the imaginary component of the same spectra from scenario 2 as shown in the right part of Fig. 5. In this case, the LSTM was able to perfectly construct the peak like shape for the four resonances. The peak positions and widths were perfectly reflecting the theoretical ones in both NRB regions. Moreover, comparing with MEM and KK, LSTM was able to remove completely the background.

To evaluate the quality of the constructed components, we used the Root Mean Squared Error (RMSE) of all spectra in both scenarios. First, we calculated the RMSE per spectrum then the mean error and standard deviation of all spectra in each scenario were determined separately. As shown in Table 1, the constructed squared modulus of the MEM method fits the theoretical one with a mean error of $\approx 0.0264 \pm 0.005$ and $\approx 0.026 \pm 0.006$ in scenario 1 and 2, respectively. This represents a relative error of $3.97\% \pm 0.7\%$ and $3.7\% \pm 0.8\%$ in scenario 1 and 2, respectively. The mean errors of the constructed imaginary component were $\approx 0.1095 \pm 0.009$ and $\approx 0.1053 \pm 0.007$ in scenario 1 and 2, respectively. These values represent a relative error of $16.11\% \pm 1.3\%$ and $14.91\% \pm 0.9\%$. The constructed real component fits the theoretical one with a mean error of $\approx 0.0987 \pm 0.007$ and $\approx 0.0996 \pm 0.007$, which represents a relative error of $14.73\% \pm 1\%$ and $14.12\% \pm 0.9\%$, in scenario 1 and 2, respectively.

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21012
Optics EXPRESS

**Table 1. The mean and standard deviation of the Root Mean Squared Error (RMSE), which was calculated on single spectrum, for MEM, KK and LSTM methods. Comparing the constructed imaginary component of all three methods, the LSTM network overcame MEM and KK since the method shows a significantly lower mean error in both regions.**

| Scenario | MEM | | | KK | | | LSTM |
|---|---|---|---|---|---|---|---|
| | $\lvert\chi^{(3)}\rvert^2$ | $\mathrm{Im}(\chi^{(3)})$ | $\mathrm{Re}(\chi^{(3)})$ | $\lvert\chi^{(3)}\rvert^2$ | $\mathrm{Im}(\chi^{(3)})$ | $\mathrm{Re}(\chi^{(3)})$ | $\mathrm{Im}(\chi^{(3)})$ |
| 1 | 0.0264 | **0.1095** | 0.0987 | $4.69e^{-18}$ | **0.1018** | 0.0669 | **0.0115** |
| | ±0.005 | **±0.009** | ±0.007 | $\pm 4.73e^{-19}$ | **±0.011** | ±0.008 | **±0.003** |
| 2 | 0.026 | **0.1053** | 0.0996 | $4.63e^{-18}$ | **0.1039** | 0.068 | **0.0116** |
| | ±0.006 | **±0.009** | ±0.007 | $\pm 4.93e^{-19}$ | **±0.01** | ±0.01 | **±0.003** |

For the KK method the constructed squared modulus fits the theoretical one better then for the MEM technique showing a mean error of ≈ 0 with a very small variation in both scenarios. However, the constructed imaginary component represents the theoretical one with a mean error of ≈ 0.1018 ± 0.01 and ≈ 0.1039 ± 0.01, which represents a relative error of 15.33% ± 1.5% and 14.8% ± 1.4% in scenario 1 and 2 respectively. The constructed real component represents the theoretical one with a mean error of ≈ 0.0669 ± 0.008 and ≈ 0.068 ± 0.01, which represents a relative error of 10.07% ± 1.2% and 9.68% ± 1.4% in scenario 1 and 2 respectively. Furthermore, the LSTM network overcome both MEM and KK as shown in Table 1. The imaginary components of the testing set were constructed with an error of ≈ 0.01 ± 0.003, which represents a relative error of 1.67% ± 0.4% in both scenarios, respectively.

In general, the time for the execution of the MEM algorithm using the whole 4000 spectra is relatively short ≈ 10.584 minutes. Furthermore, the time for training the LSTM network is also relatively small ≈ 1.9 hours as a limited CPU was used for the execution. Additionally, the time for predicting a unique spectrum using MEM, KK and LSTM network is ≈ 0.18, ≈ 0.67 and ≈ 0.16 seconds, respectively, as shown in Table 2.

**Table 2. Comparison of the model training and the prediction times for the three methods. The time for training the LSTM network is ≈ 1.9 hours. And, the time for building the MEM model using the 4000 spectra is relatively small ≈ 10.58 min. However, the time to predict the imaginary component of one single spectrum using LSTM is smaller than the time for the MEM and KK based prediction.**

| Method | Training time | Prediction time |
|---|---|---|
| MEM | 10.58m | 0.18s |
| KK | 0 | 0.67s |
| LSTM | 1.91h | 0.16s |

After the evaluation of the two scenarios, the MEM technique and the KK algorithm were not able to accurately construct the real and the imaginary components without an accurate measurement of the NRB. It was shown that the output of both methods was strongly dependent on the NRB strength. In contrast the output of the deep learning network LSTM was almost independent of the NRB strength. Additionally, the LSTM could prove to be a powerful method for extracting the Raman like spectrum without the need for an accurate estimation of the non-resonant background.

### *4.2. Experimental data*

The validation of the above-mentioned methods is achieved by testing the three methods on experimental measured broadband CARS (BCARS) spectra. These spectra was generated using an optical source (FemtoFiber pro UCP, Toptica Photonics, Germany) with the ps pump pulse centered at 777 nm and a sub 25 fs broadband Stokes pulse covering ≈ 840 − 1100 nm. See Appendix A for further details on the setup. The comparison of the three phase retrieval methods

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21013
Optics EXPRESS

was performed using two experimental BCARS spectra, which reflect the two aforementioned scenarios. In scenario 1, the Toluene broadband CARS spectrum was tested, where the Raman resonances are in the strong non-resonant background region. In scenario 2, the Acetonitrile broadband CARS spectrum was used, where its Raman resonances exist in both the strong and the weak non-resonant background regions. Moreover, after a trial and error process, for the MEM method applied on both spectra, the optimal number of poles is $M = 150$ and Eq. (9) is used to extract the phase for $\nu_0 = 0.6$.

In scenario 1, the output of the MEM, KK and LSTM methods applied on the Toluene BCARS spectrum can be found in Appendix B (Fig. 8, Fig. 9 and Fig. 10, respectively). For the MEM and the KK methods, the reconstructed spectrum fits perfectly the BCARS Toluene spectrum and the imaginary and the real components were successfully retrieved. A comparison between the constructed imaginary components of the three methods for the Toluene BCARS spectrum is displayed in the second row of Fig. 6 in upper, middle and bottom lines, respectively. The three methods were able to successfully extract a Raman like spectrum. The outputs of MEM and KK, illustrated in the upper and middle lines of the second row in Fig. 6, show a good estimation of the peaks positions, widths and amplitudes. A clear distortion of the peaks shape is shown particularly in the region on the left. This was resolved by using the LSTM network as shown in the bottom line of Fig. 6. Therefore, the LSTM network shows a very good performance with regards to the peak amplitudes and widths. Nevertheless, the LSTM was not able to accurately predict the small peak at the band 1200 $cm^{-1}$ and this might result from the properties of the simulation data. In addition, the LSTM reconstruction does not need an additional preprocessing compared with the imaginary components of the MEM and the KK method.

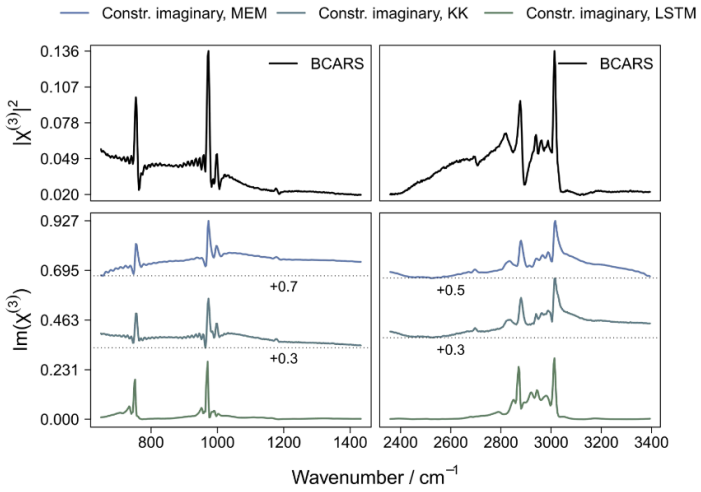


**Fig. 6.** The constructed imaginary component of BCARS Toluene spectrum using MEM, KK and LSTM. The BCARS Toluene spectrum is shown in the first row. In the second row, the constructed imaginary components by MEM, KK and LSTM are shown in upper, middle and bottom lines, respectively. The dotted line in black represents the offset value.

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21014
Optics EXPRESS

In scenario 2, the output of the MEM, KK and LSTM method applied on an Acetonitrile BCARS spectrum can be found in Appendix B (Fig. 11, Fig. 12 and Fig. 13, respectively ). For MEM and KK, the reconstructed spectrum fits perfectly the Acetonitrile BCARS spectrum but the imaginary and the real components were distorted, in particular in the region where the NRB is weak. A comparison between the constructed imaginary components using MEM, KK and LSTM for the Acetonitrile BCARS spectrum is displayed in the second row of Fig. 7 in upper, middle and bottom lines, respectively. For the MEM and the KK method, the imaginary components are dramatically distorted in the region where the NRB is weak. This is visible by the presence of a dispersive line shape as displayed in the upper and the middle lines of the second row in Fig. 7. In the region on the left, MEM and KK constructed imaginary components show a background contribution and small peak distortions. On the other hand, the LSTM network predicted correctly the peak positions and widths, but a small distortion is seen in the amplitudes in the strong NRB region. However, in the weak NRB region, the LSTM network could successfully extract the imaginary component which is seen by the accurate estimation of the positions, widths and amplitudes of the peaks. Nevertheless, LSTM was not able to accurately predict the small peak at the band 1337 $cm^{-1}$ and this might result from the properties of the simulation data. In addition, in LSTM, there is no need for additional preprocessing steps in contrast to the imaginary components from the MEM and the KK method.

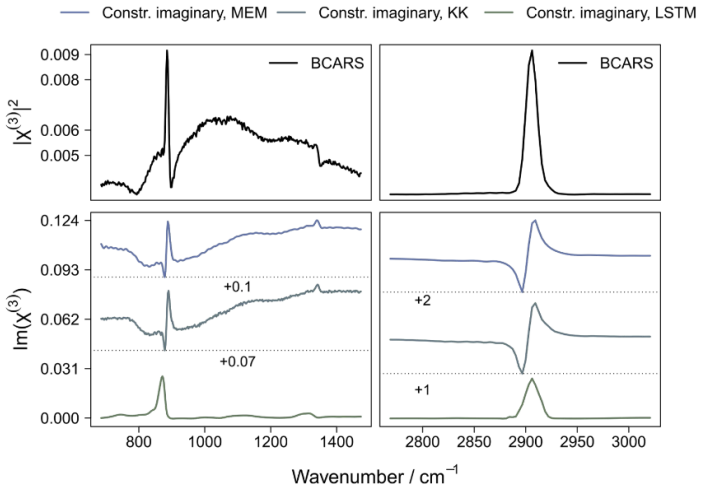


**Fig. 7.** The constructed imaginary components of a Acetonitrile BCARS spectrum using MEM, KK and LSTM. The Acetonitrile BCARS spectrum is shown in the first row. In the second row, the constructed imaginary components by MEM, KK and LSTM are shown in upper, middle and bottom lines, respectively. The dotted line in black represents the offset value.

In summary, the MEM and the KK method show an increased sensitivity regarding the strength of the NRB. This was presented either with the flip of the peaks shape in the imaginary and the real components or by the fact that a background is present after the retrieval. On the other hand,

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21015
Optics EXPRESS

the LSTM network showed its potential since its results are not dependent on the strength of the non-resonant background.

## 5. Conclusion

Coherent anti-Stokes Raman scattering offers many advantages over spontaneous Raman scattering among which the increase of the Raman signal, the elimination of fluorescence and the rapid time acquisition. Even though this optical technique improves the Raman vibrational modes, the CARS signal is distorted due to the contribution of the non-resonant background that has destructive and constructive role at the same time. Therefore, to extract the phase without experimentally removing the non-resonant background, algorithmic methods can be implemented. In this paper, we examined two popular methods for the phase retrieval; the maximum entropy method (MEM) and the Kramers-Kronig relation (KK). Furthermore, we applied deep learning using the long short term memory network as a phase retrieval tool. The imaginary components extracted from these three methods were compared by varying the non-resonant background into strong and weak regions.

Initially, the three methods were tested on a CARS simulated spectra for two different scenarios where the non-resonant background is strong and weak. In the scenario where the Raman resonances are built only in the strong NRB region, the imaginary components relative to MEM, KK and LSTM methods were correctly constructed. Their corresponding peaks shape reflect the theoretical ones. While in the scenario where some Raman resonances are built in the weak NRB region, the constructed imaginary components of the three methods differ. In MEM and LSTM, the peaks shape reflect the theoretical ones. While in KK, in particular in the region where the non-resonant background is weak, the constructed components were dramatically affected as the imaginary component shows a dispersive shape and the real component shows a peak like shape. However, for both MEM and KK methods, an additional preprocessing phase to remove the background is essential which is not needed when using the LSTM network. Additionally, the time for training the model in LSTM is high relatively to the other methods but the prediction time of one spectrum is significantly short in both scenarios $\approx 0.16$ second. Also, in both scenario the RMSE for the construction of the Raman like spectrum shows a mean error of $\approx 0.01$ for the LSTM network, while it shows a mean error of $\approx 0.11$ and $\approx 0.10$ for the MEM and KK methods, respectively.

Afterwards, we tested these three methods on two BCARS Acetonitrile and Toluene spectra. For BCARS Toluene spectrum, the constructed imaginary components via MEM and KK methods have a peak like shape with a minor distortion. Similarly, LSTM predicts well the peaks shape for the imaginary component plus the distortion mentioned previously was removed. On the other hand, for the BCARS Acetonitrile spectrum, the LSTM network predicts the imaginary component with a peak like shape contrarily to MEM and KK which show a dispersive shape in their constructed imaginary components. Additionally, in LSTM there is no need for further preprocessing step to remove the background in contrast to MEM and KK.

Therefore, the results of MEM, KK and LSTM for the simulated data and the experimental data can be interpreted similarly. MEM and KK are dependent on the strength of the NRB for both data types. In contrast the LSTM network is not dependent an the NRB strength and shows good phase retrieval performances for BCARS spectra.

As a conclusion, we used for the very first time the deep learning technique as a tool for the phase retrieval. To assess its potential, we varied the strength of the non-resonant background and we compared the deep learning output to those of MEM and KK methods. The deep learning technique overcame MEM and KK regarding the peak shape and the removal of additional background contribution. However, for future work, the development of the LSTM network by increasing the sample size, as it was relatively small, can be considered as a step forward towards a perfect phase extraction method.

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21016
Optics EXPRESS

### Appendix A. BCARS setup

In this section a description of the experimental BCARS setup is provided. As excitation source two co-seeded fiber lasers (Toptica FemtoFiber pro UCP) were used generating a narrow band (pump/probe) beam centered at 777 nm with a pulse duration of ~ 2.6 ps, 40 MHz repetition rate at an average power of ~ 10 mW (on the sample) and a super continuum (Stokes beam) generating pulses with a pulse duration of 18 fs within a range of 840-1100 nm with a repetition rate of 40 MHz, and an average power of 8 mW. The narrow band source was guided by means of mechanical delay stage to achieve control over temporal overlap of the two laser beams. The Stokes beam was passed through a beam expander (refractive telescope) to enlarge the beam diameter to match the back aperture of the objective. The two beams were then spatially combined by a dichroic beam combiner and focused on the sample by a 20x objective, 0.5 NA (Olympus, UPLFLN20X/0.5). The sample was placed on a three-axis translation stage and the beams were then collected and collimated in a confocal configuration by another 20X, 0.5 NA objective. The excitation beams and generated anti-Stokes beam were then passed through two shortpass filters (Semrock, Brightline multiphoton 770SP) to spectrally filter out the excitation beams and the generated anti-Stokes beam was then guided and focused by an achromatic lens to the entrance slit aperture of the spectrometer (Andor, Kymera 193i) attached with a CCD camera (Newton 920, DU920P-BEX2-DD). For the measurement of the solvents: Toluene (*) and Acetonitrile (*); were contained in a quartz cuvette with 1 mm of layer thickness (Hellma Macro-cuvette 110-QS).

### Appendix B. Results of the three methods applied on experimental BCARS data

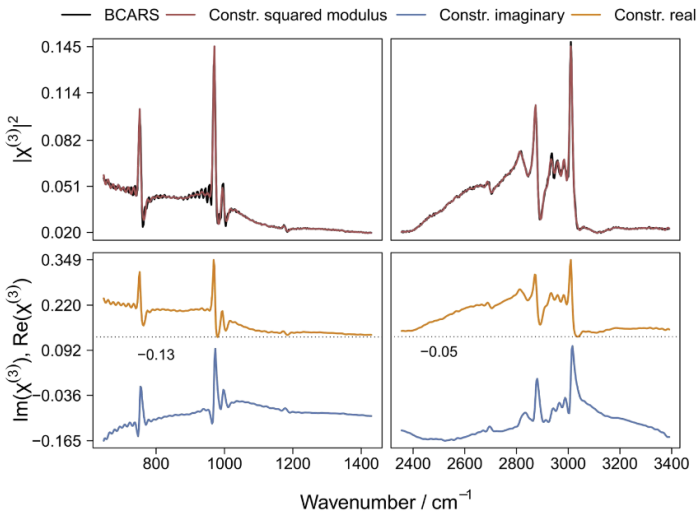


**Fig. 8.** Output of MEM for a Toluene BCARS spectrum. In the first row, the reconstructed spectrum and the BCARS spectrum are displayed. In the second row, the real and the imaginary components are shown in the upper and bottom lines, respectively. The dotted line in black represents the offset value.

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21017
Optics EXPRESS

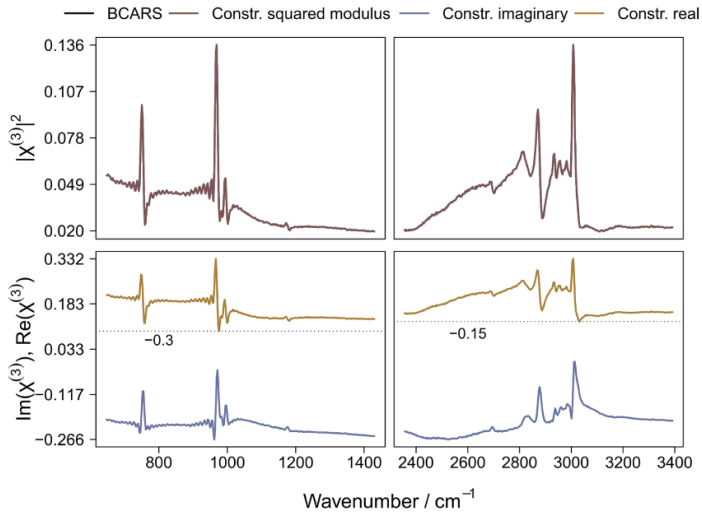


**Fig. 9.** Output of KK for a Toluene BCARS spectrum. In the first row, the reconstructed spectrum and the BCARS spectrum are displayed. In the second row, the real and the imaginary components are shown in the upper and bottom lines, respectively. The dotted line in black represents the offset value.

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21018
Optics EXPRESS

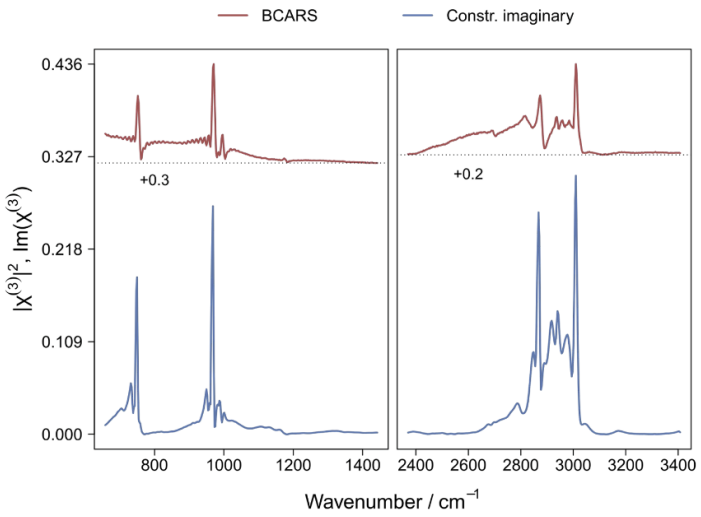


**Fig. 10.** Prediction of the imaginary component of the BCARS Toluene spectrum using the trained LSTM model. The BCARS Toluene spectrum and its predicted imaginary component are shown in the upper and bottom lines, respectively. The dotted line in black represents the offset value.

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21019
Optics EXPRESS

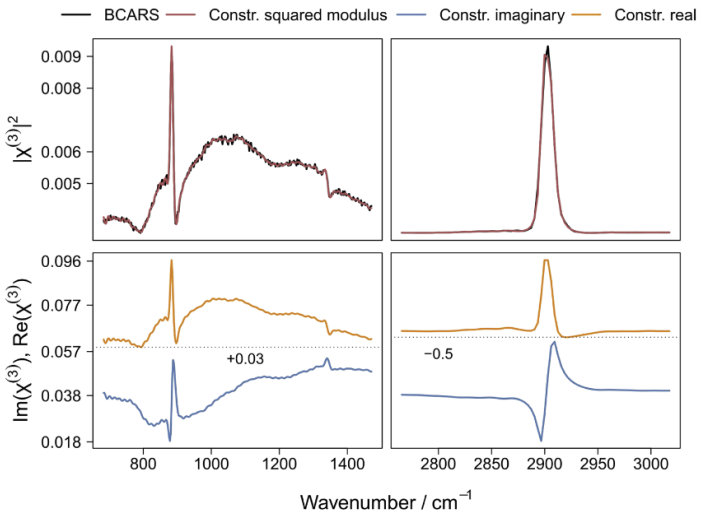


**Fig. 11.** Output of MEM for an Acetonitrile BCARS spectrum. In the first row, the reconstructed spectrum and the BCARS spectrum are displayed. In the second row, the real and the imaginary components are shown in the upper and bottom lines, respectively. The dotted line in black represents the offset value.

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21020
Optics EXPRESS

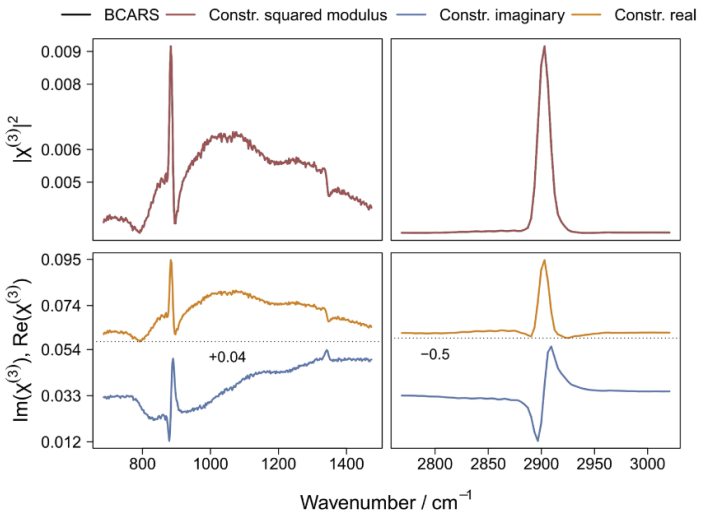


**Fig. 12.** Output of KK for a Acetonitrile BCARS spectrum. In the first row, the reconstructed spectrum and the BCARS spectrum are displayed. In the second row, the real and the imaginary components are shown in the upper and bottom lines, respectively. The dotted line in black represents the offset value.

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21021
Optics EXPRESS

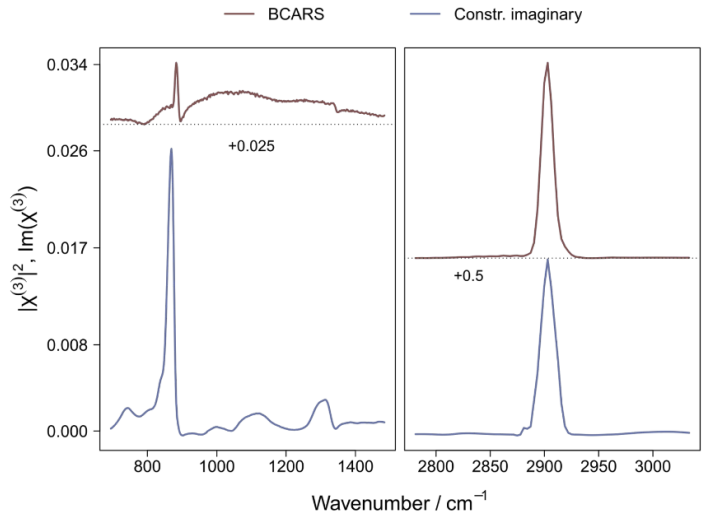


**Fig. 13.** Prediction of the imaginary components of the Acetonitrile BCARS spectrum using the trained LSTM model. The Acetonitrile BCARS spectrum and its predicted imaginary component are shown in the upper and bottom lines, respectively. The dotted line in black represents the offset value.

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21022
Optics EXPRESS

## Appendix C. Parameters for MEM and LSTM applied on simulation data

**Table 3. Parameters for the MEM method and the LSTM network applied to the simulated CARS spectra in both scenarios.**

| Parameters | | | | | |
|---|---|---|---|---|---|
| **Simulation data** | | **MEM** | | **LSTM** | |
| No. of resonances | [4, 7] | M | 150 | No. of Epochs | 50 |
| Amplitudes | [0.001, 0.5] | $\nu_0$ | 0.06 | Optimizer | Adam |
| Bandwidths | [0.0001, 0.01] | | | Learning rate | 0.005 |
| $\chi_{nr}$ | 0.5 | | | Batch size | 10 |
| No. of samples | 4000 | | | Validation size | 20% |
| | | | | No. of units in layer | 30 |

## Funding

Deutsche Forschungsgemeinschaft (CRC 1076 AquaDiva, CRC 1375 NOA); Leibniz-Gemeinschaft (SAS-2015-HKI-LWC (FastDrop)).

## Acknowledgments

The financial support from the Deutsche Forschungsgemeinschaft (DFG) via the CRC 1076 AquaDiva and the CRC 1375 NOA is highly appreciated. This work was also supported with funds from the funding line strategic networking of the Leibniz competition within the framework of the Leibniz Science Campus InfectoOptics SAS-2015-HKI-LWC (FastDrop).

## Disclosures

The authors declare no conflicts of interest.

## References

1. Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev, "Phase Retrieval with Application to Optical Imaging: A contemporary overview," IEEE Signal Process. Mag. **32**(3), 87–109 (2015).
2. J.-X. Cheng and X. S. Xie, Coherent Raman scattering microscopy (Taylor & Francis, 2016).
3. C. Krafft, B. Dietzek, J. Popp, and M. Schmitt, "Raman and coherent anti-Stokes Raman scattering microspectroscopy for biomedical applications," J. Biomed. Opt. **17**(4), 040801 (2012).
4. M. Cui, B. R. Bachler, and J. P. Ogilvie, "Comparing coherent and spontaneous Raman scattering under biological imaging conditions," Opt. Lett. **34**(6), 773–775 (2009).
5. J.-x. Cheng, A. Volkmer, L. D. Book, and X. S. Xie, "Multiplex coherent anti-Stokes Raman scattering microspectroscopy and study of lipid vesicles," J. Phys. Chem. B **106**(34), 8493–8498 (2002).
6. T. W. Kee and M. T. Cicerone, "Simple approach to one-laser, broadband coherent anti-Stokes Raman scattering microscopy," Opt. Lett. **29**(23), 2701–2703 (2004).
7. M. Jurna, J. P. Korterik, C. Otto, J. L. Herek, and H. L. Offerhaus, "Vibrational phase contrast microscopy by use of coherent anti-Stokes Raman scattering," Phys. Rev. Lett. **103**(4), 043905 (2009).
8. T. W. Bocklitz, F. S. Salah, N. Vogler, S. Heuke, O. Chernavskaia, C. Schmidt, M. J. Waldner, F. R. Greten, R. Bräuer, M. Schmitt, A. Stallmach, I. Petersen, and J. Popp, "Pseudo-HE images derived from CARS/TPEF/SHG multimodal imaging in combination with Raman-spectroscopy as a pathological screening tool," BMC Cancer **16**(1), 534 (2016).
9. O. Chernavskaia, S. Heuke, M. Vieth, O. Friedrich, S. Schürmann, R. Atreya, A. Stallmach, M. F. Neurath, M. Waldner, I. Petersen, M. Schmitt, T. Bocklitz, and J. Popp, "Beyond endoscopic assessment in inflammatory bowel disease: real-time histology of disease activity by non-linear multimodal imaging," Sci. Rep. **6**(1), 29239 (2016).
10. H. Hajjar, H. Boukhaddaoui, A. Rizgui, C. Sar, J. Berthelot, C. Perrin-Tricaud, H. Rigneault, and N. Tricaud, "Label-free non-linear microscopy to measure myelin outcome in a rodent model of Charcot-Marie-Tooth diseases," J. Biophotonics **11**(12), e201800186 (2018).
11. F. Masia, A. Glen, P. Stephens, W. Langbein, and P. Borri, "Label-free quantitative chemical imaging and classification analysis of adipogenesis using mouse embryonic stem cells," J. Biophotonics **11**(7), e201700219 (2018).

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21023
Optics EXPRESS

12. R. Galli, O. Uckermann, T. Sehm, E. Leipnitz, C. Hartmann, F. Sahm, E. Koch, G. Schackert, G. Steiner, and M. Kirsch, "Identification of distinctive features in human intracranial tumors by label-free nonlinear multimodal microscopy," J. Biophotonics **11**, e201800465 (2019).
13. C. H. Camp Jr, Y. J. Lee, and M. T. Cicerone, "Quantitative, comparable coherent anti-Stokes Raman scattering (CARS) spectroscopy: correcting errors in phase retrieval," J. Raman Spectrosc. **47**(4), 408–415 (2016).
14. M. R. Hamblin, P. Avci, and G. K. Gupta, Imaging in dermatology (Elsevier, 2016).
15. R. W. Gerchberg, "A practical algorithm for the determination of phase from image and diffraction plane pictures," Optik **35**, 237–246 (1972).
16. E. M. Vartiainen, K.-E. Peiponen, H. Kishida, and T. Koda, "Phase retrieval in nonlinear optical spectroscopy by the maximum-entropy method: an application to the $|\chi(3)|$ spectra of polysilane," J. Opt. Soc. Am. B **13**(10), 2106–2114 (1996).
17. E. Vartiainen, K.-E. Peiponen, and T. Asakura, "Maximum entropy model in reflection spectra analysis," Opt. Commun. **89**(1), 37–40 (1992).
18. E. Vartiainen, T. Asakura, and K.-E. Peiponen, "Generalized noniterative maximum entropy procedure for phase retrieval problems in optical spectroscopy," Opt. Commun. **104**(1-3), 149–156 (1993).
19. P. Grosse and V. Offermann, "Analysis of reflectance data using the Kramers-Kronig relations," Appl. Phys. A **52**(2), 138–144 (1991).
20. A. Kuzmenko, "Kramers–Kronig constrained variational analysis of optical spectra," Rev. Sci. Instrum. **76**(8), 083108 (2005).
21. E. Gornov, E. M. Vartiainen, and K.-E. Peiponen, "Comparison of subtractive Kramers-Kronig analysis and maximum entropy model in resolving phase from finite spectral range reflectance data," Appl. Opt. **45**(25), 6519–6524 (2006).
22. M. T. Cicerone, K. A. Aamer, Y. J. Lee, and E. Vartiainen, "Maximum entropy and time-domain Kramers–Kronig phase retrieval approaches are functionally equivalent for CARS microspectroscopy," J. Raman Spectrosc. **43**(5), 637–643 (2012).
23. H. H. Bauschke, P. L. Combettes, and D. R. Luke, "Hybrid projection–reflection method for phase retrieval," J. Opt. Soc. Am. A **20**(6), 1025–1034 (2003).
24. G. Zhang, T. Guan, Z. Shen, X. Wang, T. Hu, D. Wang, Y. He, and N. Xie, "Fast phase retrieval in off-axis digital holographic microscopy through deep learning," Opt. Express **26**(15), 19388–19405 (2018).
25. A. Sinha, J. Lee, S. Li, and G. Barbastathis, "Lensless computational imaging through deep learning," Optica **4**(9), 1117–1125 (2017).
26. A. Goy, K. Arthur, S. Li, and G. Barbastathis, "Low photon count phase retrieval using deep learning," Phys. Rev. Lett. **121**(24), 243902 (2018).
27. A. Zumbusch, G. R. Holtom, and X. S. Xie, "Three-dimensional vibrational imaging by coherent anti-Stokes Raman scattering," Phys. Rev. Lett. **82**(20), 4142–4145 (1999).
28. M. Müller and A. Zumbusch, "Coherent anti-stokes raman scattering microscopy," ChemPhysChem **8**(15), 2156–2170 (2007).
29. E. M. Vartiainen, "Phase retrieval approach for coherent anti-Stokes Raman scattering spectrum analysis," J. Opt. Soc. Am. B **9**(8), 1209–1214 (1992).
30. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**(7553), 436–444 (2015).
31. F. Chollet and J. J. Allaire, "Deep Learning with R, Ch. 5.4," (2018).
32. J. Houston, F. Glavin, and M. Madden, "Robust Classification of High-Dimensional Spectroscopy Data Using Deep Learning and Data Synthesis," J. Chem. Inf. Model. **60**(4), 1936–1954 (2020).
33. M. Chatzidakis and G. Botton, "Towards calibration-invariant spectroscopy using deep learning," Sci. Rep. **9**(1), 2126 (2019).
34. K. Ghosh, A. Stuke, M. Todorović, P. B. Jørgensen, M. N. Schmidt, A. Vehtari, and P. Rinke, "Deep learning spectroscopy: neural networks for molecular excitation spectra," Adv. Sci. **6**(9), 1801367 (2019).
35. M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, X. Zhang, J. Zhao, and K. Zieba, "End to end learning for self-driving cars," arXiv preprint arXiv:1604.07316 (2016).
36. R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," Hum. brain mapping **38**(11), 5391–5420 (2017).
37. A. Vieira, "Predicting online user behaviour using deep learning algorithms," arXiv preprint arXiv:1511.06247 (2015).
38. F. Jia, Y. Lei, J. Lin, X. Zhou, and N. Lu, "Deep neural networks: A promising tool for fault characteristic mining and intelligent diagnosis of rotating machinery with massive data," Mech. Syst. Signal Process. **72-73**, 303–315 (2016).
39. Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," arXiv preprint arXiv:1506.00019 (2015).
40. Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," IEEE Trans. Neural Netw. **5**(2), 157–166 (1994).
41. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation **9**(8), 1735–1780 (1997).
42. C. Olah, "Understanding LSTM networks, 2015," URL http://colah.github.io/posts/2015-08-Understanding-LSTMs (2015).

Research Article Vol. 28, No. 14 / 6 July 2020 / Optics Express 21024
Optics EXPRESS

43. F. A. Gers and J. Schmidhuber, "Recurrent nets that time and count," in *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, vol. 3, (2000), pp. 189–194.
44. K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A search space odyssey," IEEE Trans. Neural Netw. Learning Syst. **28**(10), 2222–2232 (2017).
45. S. Guo, T. Bocklitz, and J. Popp, "Optimization of Raman-spectrum baseline correction in biological application," Analyst **141**(8), 2396–2404 (2016).

# P3. COMPARISON OF DENOISING TOOLS FOR RECONSTRUCTION OF NONLINEAR MULTIMODAL IMAGES

Reprinted, with permission, from [R. Houhou, E. Quansah, T. Meyer-Zedler, M. Schmitt, Franziska Hoffmann, O. Guntinas-Lichius, J. Popp, and T. Bocklitz, Comparison of denoising tools for reconstruction of nonlinear multimodal images, Biomedical Optics Express, submitted on 28 March 2022].

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

| **R. Houhou[1], E. Quansah[2], T. Meyer-Zedler[3], M. Schmitt[4], Franziska Hoffmann[5], O. Guntinas-Lichius[6], J. Popp[7], and T. Bocklitz[8], Comparison of denoising tools for reconstruction of nonlinear multimodal images, Biomedical Optics Express, submitted on 28/03/2022** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Involved in | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| | | | | | | | | |
| Conceptual research design | × | | | × | | | × | × |
| Planning of research activities | × | | | | | | | × |
| Data collection | | × | × | | × | × | | |
| Data analyses and interpretation | × | | × | | | | | × |
| Manuscript writing | × | × | × | × | × | × | × | × |
| Suggested publication equivalence value | 1.0 | | | | | | | |

# Comparison of denoising tools for reconstruction of nonlinear multimodal images

Rola Houhou,[1,2] Elsie Quansah,[1,2] Tobias Meyer-Zedler,[1,2] Michael Schmitt,[1] Franziska Hoffmann,[3] Orlando Guntinas-Lichius,[3] Jürgen Popp,[1,2] and Thomas Bocklitz[1,2,*]

[1]*Institute of Physical Chemistry and Abbe Center of Photonics, Friedrich Schiller University, Helmholtzweg 4, 07743 Jena, Germany*
[2]*Leibniz Institute of Photonic Technology (Member of Leibniz Health Technologies), Albert-Einstein-Straße 9, 07745 Jena, Germany*
[3]*Department of Otorhinolaryngology, Institute of Phoniatry/Pedaudiology, Jena University Hospital, Jena, Germany*
*Thomas.bocklitz@uni-jena.de

**Abstract:** Biophotonic multimodal imaging techniques provide deep insights into biological samples such as cells or tissues. However, the measurement time increases dramatically when high-resolution multimodal images are required. To address this challenge, mathematical methods can be used to shorten the acquisition time for such high-quality images. In this research, we tested a standard phase retrieval method via the Gerchberg-Saxton algorithm and two artificial intelligence (AI) methods on head and neck tissues. The first AI method is a transfer learning-based technique that uses the pre-trained network DnCNN. The second is a trained network consisting of a simple architecture (incSRCNN) derived from the super-resolution convolutional neural network using augmented head and neck images. These methods reconstruct improved images using measured low-quality (LQ) images, which were measured in approximately 2 seconds. The evaluation was performed on artificial LQ images built from high-quality (HQ) images measured in 8 seconds by adding Poisson noise and then on experimental LQ images measured in 2 seconds. The results show the potential of using deep learning on these multimodal images to improve the data quality and reduce the acquisition time. Our proposed network has the advantage of having a simple architecture.

© 2022 Optica Publishing Group under the terms of the Optica Open Access Publishing Agreement

## 1. Introduction

Medical imaging is an important and active area of research with the potential to significantly improve the disease diagnosis and patient treatment. For decades, medical imaging modalities, e.g., X-ray, ultrasound imaging, and computerized tomography (CT) scan, have served as important tools to assist physicians in making their diagnostic decisions. Although several new especially optical imaging technologies have been developed in the last decades, their adoption in healthcare systems is still minimal. Nonlinear optical techniques, e.g., coherent anti-Stokes Raman scattering (CARS), two-photon excited fluorescence (TPEF), second-harmonic generation (SHG), and fluorescence lifetime imaging (FLIM) [1] are capable of measuring detailed information about the chemical composition and morphology of tissue sections with high spatial resolution. In particular, the simultaneous combination of two or more of these optical spectroscopic methods, called multimodal imaging (MM), allows maximizing the obtained chemical and morphological information of the measured tissues [2–6]. For instance, Vogler et al. [1] presented a microscopic experiment that combines three nonlinear optical techniques; CARS, TPEF, and SHG, and shows how different kinds of molecules and different contrast mechanisms can be obtained in one image measurement. The multimodal imaging approach provides high-quality (HQ) images, but the acquisition of such high-quality images

requires a relatively long acquisition process in comparison with low-quality images, because photon shot noise is the prominent noise source in nonlinear imaging techniques. Hence, the faster MM imaging required for real-time monitoring leads to an increase in the noise level of the images, which degrades their quality and affects the identification of tissues or their associated diseases, or abnormalities.

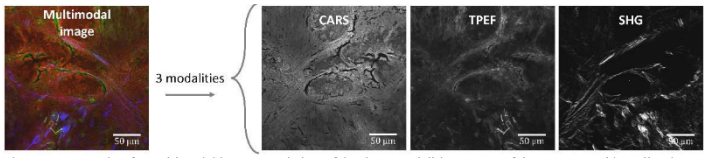


Fig. 1. An example of a multimodal image consisting of the three modalities CARS of the CH2-stretching vibration at 2850 cm-1, TPEF, and SHG is given.

In addition to increasing the acquisition time, image denoising is a fundamental preprocessing technique that can remove noise from images but may result in the loss of relevant information [7–9]. Consequently, the trade-off between fast imaging and a suitable denoising method needs to be balanced and optimized for an effective diagnostic imaging tool. The denoising algorithms vary from basic digital image filters to iterative reconstruction techniques. Therefore, choosing a suitable denoising method is not simple, and the restored images should maintain the following properties [8]. First, the details and edges that are critical in malignant tissue detection should be preserved. This means that the denoising algorithms should not produce artifacts and the recovered images should be similar to the original image. In addition, the algorithm should be computationally efficient and have low complexity, which is a prerequisite in medical applications that require immediate results. Finally, the denoising algorithms should not depend on vast amounts of data, which is not practical or readily accessible in medical imaging. Apart from the standard image denoising methods, deep learning featured a high potential for denoising and showed outstanding performance, especially in the processing of natural images and various medical imaging techniques, e.g., ultrasound imaging [10–12] and CT scan [13,14]. Therefore, we evaluated two deep learning methods on the multimodal images that comprise CARS, TPEF, SHG modalities and compared them with the phase retrieval method via Gerchberg-Saxton (GS) [15–19]. An example of a MM image is visualized in Fig. 1. We used two deep learning techniques; a pre-trained network, namely DnCNN [20], and our deep learning network with simple architecture derived from the super-resolution convolution neural network (SRCNN) [21,22] that we referred to as incSRCNN. In this manuscript, we briefly explain the three methods at the beginning and then describe the data and workflow. We then discuss the reconstruction of synthetic and experimental low-quality images using the GS algorithm, the DnCNN, and the incSRCNN networks is discussed. Finally, we summarize our results in the conclusion section.

## 2. Method

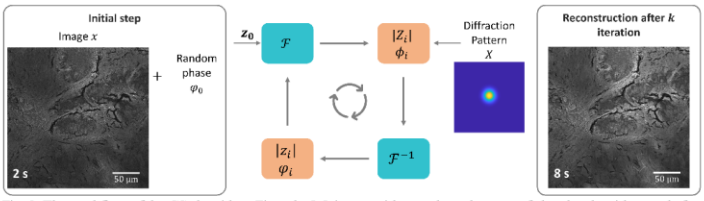


Fig. 2. The workflow of the GS algorithm. First, the LQ image with a random phase was fed to the algorithm, and after k iteration, the high-quality image was constructed. The GS algorithm depends on an estimation of the source, and Gaussian estimation was used.

This section briefly explains the implemented methods, grouped into a description of the classical phase retrieval problem and deep learning methods. First, the phase retrieval problem is implemented since it is applied to many problems. Several well-known phase retrieval algorithms exist, e.g., hybrid input-output (HIO) and Gerchberg-Saxton (GS). We focused on applying GS [15,17] to the MM images. Briefly, GS is the recovery of the phase using the measured image and the source object. It is considered an error-reduction algorithm that iteratively calculates the error until it converges. The GS algorithm is shown in Fig. 2, and it is applied independently on each channel where the phase and the modified amplitudes are determined iteratively, enabling the image reconstruction. Its input represents both the amplitudes of the sampled image and an estimation of the diffraction plane intensity. First, an initial phase $\boldsymbol{\varphi}_0$ is used by generating randomly uniform numbers between -$\pi$ and $\pi$. At iteration $k$, the initial field in the object plane is calculated using Eq. (1).

$$z_{\mathbf{k}} = \sqrt{x}\exp\left(\mathrm{i}\boldsymbol{\varphi}_{\mathbf{k}-1}\right) \tag{1}$$

The phase distribution in the target plane $\boldsymbol{\varphi}_{\mathbf{k}}$ is then calculated via the fast Fourier transform (*FFT*), as shown in Eq. (2).

$$\boldsymbol{\phi}_{\mathbf{k}} = \arg\left(FFT\left(z_{\mathbf{k}}\right)\right) \tag{2}$$

Eq. (3) combines the phase distribution in the target plane with the target intensity $\sqrt{\boldsymbol{X}}$ and finally the phase in the object plane $\boldsymbol{\varphi}_{\mathbf{k}}$ is recovered by using Eq. (4).

$$\boldsymbol{A}_{\mathbf{k}} = \sqrt{\boldsymbol{X}}\exp\left(\mathrm{i}\boldsymbol{\phi}_{\mathbf{k}}\right) \tag{3}$$

$$\boldsymbol{\varphi}_{\mathbf{k}} = \arg\left(FFT\left(\boldsymbol{A}_{\mathbf{k}}\right)\right). \tag{4}$$

Apart from the classical methods used in image denoising, artificial intelligence (AI) based methods have widely been used for restoring images, especially in computer vision and medical imaging, e.g., X-ray, CT imaging, and ultrasound scans. In artificial intelligence, high-quality images, which represent improved images in terms of signal-to-noise ratio (SNR), can be acquired by transfer learning or directly constructing deep learning techniques. Transfer learning consists of using knowledge obtained from one problem and transferring it to another related one. The deep learning method, on the other hand, trains a neural network with specific architecture using the available data, optimizing the parameter during training. In this manuscript, we evaluated both methods on the MM images.

First, a pre-trained neural network, the denoising convolutional neural networks (DnCNN), was used as a transfer learning tool. DnCNN was trained on natural images to correct noise and artifacts in corrupted images [20]. Briefly, DnCNN is a pre-trained network that outputs the residual image, i.e., the difference between the noisy observation and the latent clean image,

instead of predicting the denoised image. The architecture of this network is an adapted version of the VGG network [23] that is suitable for the image denoising task. Formally, the averaged mean squared error calculated in Eq. (*5*) between the desired residual images and estimated ones from noisy input

$$l(\theta) = \frac{1}{2N}\sum_{i=1}^{N} \left|\left| \mathfrak{R}\left(y_i;\theta\right) - \left(y_i - x_i\right) \right|\right|_F^2, \tag{5}$$

can be adopted as the loss function to learn the trainable parameters in DnCNN. $\mathfrak{R}(y)$ represents the residual mapping and $\left\{\left(y_i, x_i\right)\right\}_{i=1}^{N}$ represents $N$ noisy-clean training image patch pairs. In a nutshell, the DnCNN model has two main features: the residual learning formulation is adopted to learn $\mathfrak{R}(y)$, and batch normalization is incorporated to speed up training and boost the denoising performance.

Then, we constructed and trained a simple network which is a modified version of the super-resolution convolutional neural network (SRCNN) [21,22], namely incSRCNN. The architecture of this network is shown in Fig. 3. Like the SRCNN, the proposed network consists of three layers; however, it is implemented as a denoising task that outputs the same input size. The input image is convolved in the first layer with three different kernel sizes 3, 5, and 9 into 192 feature maps. The second layer then applies a 1×1 kernel to condense to 64 feature maps. Finally, the third layer uses a 3×3 kernel to construct the output image.

## 3. Data acquisition, description and workflow

### *3.1 Data acquisition and description*

The data used for developing the denoising method has been acquired using a laser scanning microscope (LSM510, Zeiss, Germany) equipped with a ps-laser system for coherent anti-Stokes Raman scattering (CARS), second harmonic generation (SHG), and two-photon excited fluorescence (TPEF) microscopy as described in detail previously [24]. Briefly, the sample is illuminated with two spatially and temporally synchronized laser pulse trains of ps pulse duration. The difference frequency of both lasers matches the symmetric CH2-stretching vibration at 2850 $cm^{-1}$. The pump laser is operating at 672.5 nm, the Stokes laser is at 832 nm. The specimen is illuminated through a 20x planapochromatic objective (Zeiss, Germany, NA=0.8) using a 50 mW pump and 70 mW of Stokes power. CARS and SHG signals are collected and detected in forward direction by PMT detectors. The signals are split by a 514 nm dichroic longpass mirror. The CARS signal is detected using a 550 nm bandpass filter, the SHG signal using a 415 nm bandpass filter. The TPEF signal is collected in epi-direction through the illumination objective and reflected by a 600 nm longpass dichroic mirror to the PMT detector. In front of the PMT the TPEF signal is filtered using a 650 nm shortpass filter and a 458/64nm bandpass filter (both Semrock, USA). All analyzed images have been acquired using 1.6 µs pixel dwell time, a field of view of 450 µm and 512 pixels length. For HQ images 16 frames have been averaged, for LQ images four frames averaging was applied.

The data represent head and neck tissue, with ten positions measured using the nonlinear multimodal imaging technique. The nonlinear multimodal imaging combines three modalities that are simultaneously excited using 672.5 nm pump and 832 nm Stokes and detected at 550 nm (CARS) at 458 nm (TPEF) and 415 nm (SHG). In this manuscript, we utilized high-quality (HQ) and (experimental) low-quality (LQ) images acquired within 8s and 2s, respectively. The HQ and LQ images were obtained by averaging 16 and 4 frames, respectively, and each has a spatial resolution of 512×512 pixels for a 450×450 $\mu m^2$ tile scan.

### *3.2 Workflow*

As mentioned before, we compared three denoising methods; phase retrieval via GS, the pre-trained deep network, DnCNN, and our proposed method, the incSRCNN. In the GS algorithm, each modality of the nonlinear multimodal imaging is processed independently. Since this algorithm depends greatly on knowing the source object, a Gaussian estimation is instead incorporated in the algorithm. For the DnCNN network, the pre-trained network was loaded and employed separately on each of the modalities of the nonlinear multimodal images to predict high-quality images. In the case of our proposed network, data augmentation is applied. Before data augmentation, one image was left aside for testing, and nine were split into 7 for the training and 2 for the validation. Various techniques can be considered for data augmentation; however, we used rotation, blurring, and Poisson noise for our medical images. In the analysis, we first created artificial LQ images from the HQ images by adding Poisson noise to these HQ images. The experiment and the artificial LQ images are rotated by 90°, 180°, and 270°, and the experiment LQ images were blurred using a Gaussian filter. The total number of images equals 63 for the training part and 18 for the validation. We simultaneously applied data augmentation for both HQ and LQ images. In addition, each image was split into 16 patches. Consequently, the total patch images in the training and validation sets are 1008 and 288 patch images for each channel, respectively. Since each modality of the nonlinear multimodal imaging techniques measures specific molecular contributions, we considered these channels as independent images. Accordingly, the total number of patch images equals 3024 and 864 for the training and validation sets, respectively.

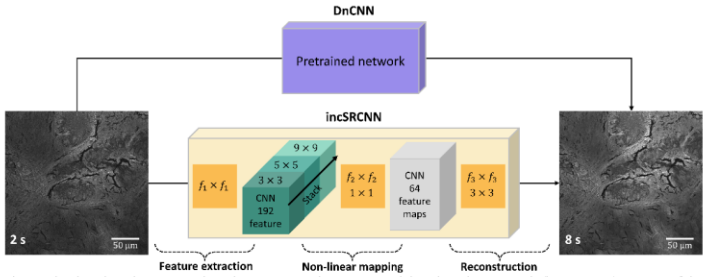


Fig. 3. The deep learning approaches via DnCNN and our proposed deep learning network (incSRCNN). On top of the figure, we used a pre-trained network, DnCNN, to predict MM images with higher quality. The architecture of our proposed network, incSRCNN, is shown on the bottom. This network represents a modified version of the SRCNN. Initially, the first layer convolves the input image with different kernel sizes into 192 feature maps. The second layer then applies a 1×1 kernel to condense to 64 feature maps. Finally, the third layer uses a 3×3 kernel to construct the output image.

## 4. Results

Our analysis was split into two sections; first, we created artificial low-quality (LQ) images from high-quality images (HQ) by adding Poisson noise. These artificial LQ images are created intentionally of lower quality as our experimental low-quality images to be used subsequently in the training of the incSRCNN network. Therefore, the incSRCNN can generalize and cover other measurements with a different setup and lower quality. We then evaluated the GS algorithm, the pre-trained DnCNN, and the trained incSRCNN networks on these artificial LQ images. We added a Poison distributed noise contribution with $\lambda$=2, which results in an average PSNR decrease from 19.7 to 16.4. Finally, we tested these three methods on the experimental LQ image and compared their performances. However, the image reconstructions evaluation is a tricky task, particularly for medical images, and as far as the authors know, no particular image metric is (always) recommended. Therefore, we used a panel of image metrics: the peak

signal-to-noise ratio (PSNR), the structural similarity index measure (SSIM), the image correlation coefficient (ICC) [25], and the mean absolute error (MAE). In addition, we visualized the residual images and the histogram of the residual images. Moreover, the time for reconstructing one channel by these methods was calculated in Table S1 in Supplement 1.

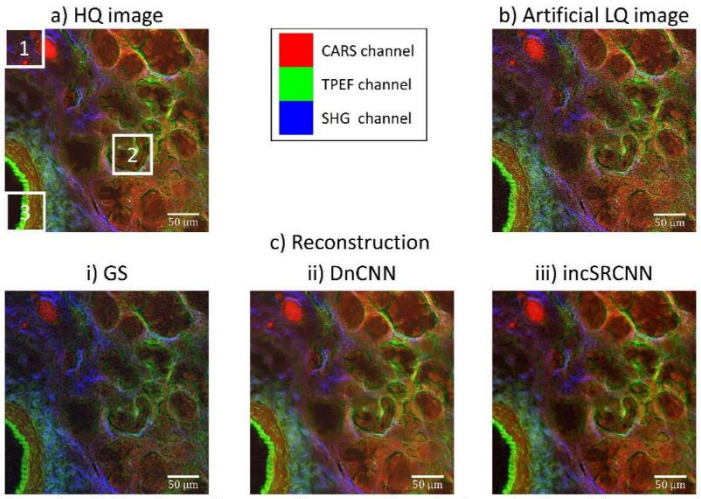


Fig. 4. The artificial LQ image with corresponding reconstructions using the GS algorithm and the DnCNN and incSRCNN networks. The experimental HQ and artificial LQ images are displayed in a) and b). The reconstructions of the artificial LQ image using the GS algorithm, the DnCNN network, and the incSRCNN network are shown in c)-(i, ii, iii), respectively. At first glance, the DnCNN network better represents the HQ image. However, the proposed incSRCNN network preserves detailed structures compared to the smooth region produced by the DnCNN network. Still, some artifacts were produced, resulting from the small data size used to train the network.

**Table 1. The PSNR, the SSIM, the ICC, and the MAE between the HQ image and the artificial LQ image and between the HQ image and the reconstructed images using the GS algorithm, the DnCNN, and incSRCNN networks**

| Image | Metric | Artificial LQ | GS | DnCNN | incSRCNN |
|---|---|---|---|---|---|
| CARS channel | PSNR | 14.767 | 14.932 | 19.228 | **21.190** |
| | SSIM | 0.27 | 0.39 | **0.60** | 0.53 |
| | ICC | 0.65 | 0.83 | **0.9** | 0.88 |
| | MAE | 0.14 | 0.15 | **0.05** | 0.07 |
| TPEF channel | PSNR | 16.099 | 21.344 | 20.787 | **22.997** |
| | SSIM | 0.17 | 0.53 | **0.62** | 0.52 |
| | ICC | 0.65 | 0.89 | **0.94** | 0.92 |
| | MAE | 0.12 | 0.06 | **0.03** | 0.06 |
| SHG channel | PSNR | 18.399 | 21.715 | 20.669 | **22.252** |
| | SSIM | 0.28 | 0.56 | **0.68** | 0.55 |
| | ICC | 0.75 | 0.85 | **0.93** | 0.9 |
| | MAE | 0.09 | 0.06 | **0.03** | 0.06 |
| MM image | PSNR | 16.422 | 19.330 | 20.228 | **22.146** |
| | SSIM | 0.24 | 0.49 | **0.63** | 0.53 |
| | ICC | 0.68 | 0.86 | **0.92** | 0.9 |
| | MAE | 0.12 | 0.09 | **0.04** | 0.06 |

First, the GS algorithm is implemented independently on the three channels that form the MM LQ images. The GS algorithm requires an approximation of the source beam and the LQ image as input. Therefore, the source beam is represented by Gaussian approximation (its illustration is shown as X in Fig. 2). A detailed explanation of the GS algorithm is discussed in the method section. The number of iterations that the algorithm carries on is 50000, and the code was built using Matlab 2020b (The MathWorks, Natick, MA).

The GS reconstruction of the artificial LQ image, displayed in Fig. 4 (c-i), generally preserves the structure but includes dark regions resulting from the Gaussian estimation. Furthermore, the overall similarity between the HQ and reconstructed images was significantly low. Since the CARS channel has a more complex structure than the TPEF and SHG channels, its reconstruction compared to the other two differs dramatically. In addition, although the noise level for the three channels decreases, only a slight improvement in the PSNR from 14.8 to 14.9 is shown for the CARS channel. However, the increase in the PSNR reached 21.3 and 21.7 from 16 and 18.4 for TPEF and SHG channels, respectively (refer to Table 1 for more details).

Moreover, the SSIM is improved for the three channels. Similar to the PSNR, an increase in the SSIM value is deduced in the three channels' reconstruction. For instance, the SSIM for the CARS, TPEF, and SHG reconstructions increases from 0.27, 0.17, and 0.28 to 0.39, 0.53, and 0.56, respectively. Furthermore, the average ICC and MAE of the reconstructed image are equal to 0.86 and 0.09, respectively, while their values for the artificial image were 0.68 and 0.12. For more insights about the reconstructions, the residual images of the HQ and the reconstructions images are visualized in Fig. S1 in Supplement 1 and compared to the residual images of the HQ and the artificial LQ images. Although the GS reconstruction was able to reconstruct some parts of the image compared with the artificial LQ case, various regions are still not well reconstructed. In addition, we illustrated in Supplement 1 Fig. S2 the histogram of the residual of the reconstructed image and compared it with the residual of the artificial LQ image. Although the PSNR, SSIM, ICC, and MAE showed improved values and the histogram of the residual images showed fewer variations, the reconstructions were poor and revealed darker regions. Afterward, we evaluated the GS algorithm on the experimental LQ image with the exact source estimation.

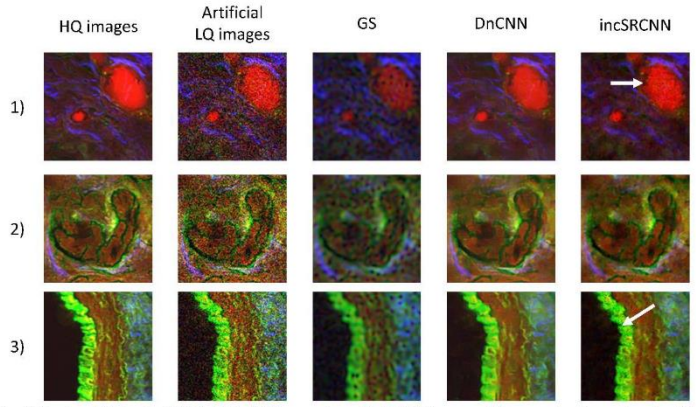


Fig. 5. Region of interests (ROIs) of the HQ image, the artificial LQ image, and its reconstructions using the GS algorithm, the DnCNN, and the incSRCNN networks. The GS algorithm produces blurry images with dark spots/regions. In the DnCNN reconstruction, some fine structures were lost while these structures were preserved using

the incSRCNN network. However, some black dots were produced as artifacts, which resulted from the small data size used to train the network.

In the experimental LQ image reconstruction, the PSNR for the TPEF and SHG channels increase from 20 to 21 and 22, respectively. However, the PSNR decreases from 19 to 14.9 in the CARS channel. In addition, the SSIM improved for only the TPEF channel but worsened in the CARS and the SHG channels. All characteristics are given in Table 2. Furthermore, the average ICC showed an improved correlation value from 0.78 to 0.86, but the average MAE showed an increased value. It is worth noticing that the worsened values are mainly related to the CARS channel reconstruction since the other channels presented acceptable results. Moreover, the GS reconstruction of the experimental LQ image does not differ from the artificial LQ reconstruction, which can be deduced from the residual images and the residual histogram in Supplement 1 Fig. S3 and Fig. S4, respectively. The reason is that the algorithm converged to a local minimum and could not improve more.

Next, we applied a pre-trained network, the DnCNN, to predict the reconstruction of the artificial and the experimental MM images. The DnCNN was implemented in Matlab 2020b (The MathWorks, Natick, MA). Similar to the GS algorithm, the DnCNN network was used independently on each of the three modalities. The reconstruction of the artificial LQ image is shown in Fig. 3 (c-ii). The spatial structures in the image are preserved, and the noise level is reduced. Furthermore, the PSNR has increased to 19.2, 20.8, and 20.7 for the CARS, TPEF, and SHG channels from 14.8, 16, and 18.4, respectively. In addition, the SSIM has significantly increased from 0.27, 0.17, 0.28 to 0.6, 0.62, 0.68 for the CARS, TPEF, SHG channels, respectively. Consequently, the overall PSNR and SSIM were improved from 16.4 and 0.24 to 20.2 and 0.63, respectively. Furthermore, the average ICC and MAE of the reconstructed image are equal to 0.92 and 0.04, while their values for the artificial image were 0.68 and 0.12. Fig. 4 shows three regions of interest (ROIs) for all reconstruction algorithms. The colors in the DnCNN reconstruction are well conserved, and the noise level in the reconstruction was reduced significantly. However, smoothed structures are displayed in these ROIs, which is critical for biomedical applications because some important information may be compromised and lost, affecting the diagnosis of tissue abnormalities and diseases. In addition, the residual images per channel between the HQ image and the reconstructed one are shown in Supplement 1 Fig. S1. In this figure, most of the values across the image are zero, which means that the DnCNN was able to reconstruct the exact value of the high-quality image.

Moreover, the histogram of the residual images between the HQ image and the DnCNN reconstruction in Supplement 1 Fig. S2 significantly reduces the values compared with the artificial LQ case. Therefore, the DnCNN reconstructed a good representation of the HQ image successfully. Afterward, we tested the performance of the DnCNN in the experimental LQ image. In Fig. 6 (c-ii), we showed the reconstruction using the DnCNN network, where the spatial structure in the image is preserved, and the noise level is slightly reduced. In Table 2, we compared the PSNR, SSIM, ICC, and MAE between the DnCNN reconstructions and the HQ image with the PSNR, SSIM, ICC, and MAE between the experiment LQ image and the HQ image. Compared to the experiment LQ image, we deduced a slight improvement in the PSNR, SSIM, ICC, and MAE values per channel and overall when using the DnCNN network. In Fig. 7, three ROIs showed a reduction in the noise level. Similar to the artificial case, smoothed regions were produced, which might cause the removal of important features that are highly sensitive in the diagnosis of diseases and abnormalities. In addition, we compared the residual images per channel of the DnCNN reconstructions with the residual images of the experiment LQ image in Supplement 1 Fig. S3. In this figure, almost similar residual values to the experimental LQ case can be detected. Furthermore, we visualized the histogram of the residual images of the DnCNN reconstruction and the experiment LQ images in Supplement 1 Fig. S4.

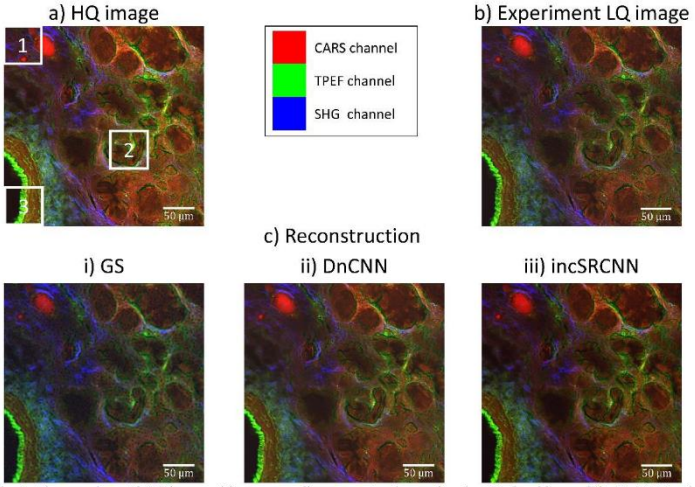


Fig. 6. The experimental LQ image with corresponding reconstructions using the GS algorithm and the DnCNN and incSRCNN networks. The experimental HQ and LQ images are displayed in a) and b), respectively. The reconstruction of the experiment LQ image using the GS algorithm, the DnCNN network, and the incSRCNN network is shown in c) – (i, ii, iii), respectively. At first glance, the DnCNN network better represents the HQ image. However, the incSRCNN reconstruction preserves detailed structures while the DnCNN reconstruction displays smoothed structures.

Finally, we evaluated our proposed network (incSRCNN) on the same artificial and experimental LQ images. The detailed architecture was described in the method section. The training of the network was performed by minimizing the mean absolute error (MAE)-based loss between the HQ images and the output of the incSRCNN network. The Adam algorithm was used for the optimization with a learning rate of $3e^{(-4)}$. A total of 1008 and 288 coupled HQ and LQ images were used for the training and the validation, respectively; refer to Supplement 1 Table S2 for more details. All computations were done using Google Colab. The total number of parameters to be trained is 20,481. We assessed different cases to train the network; three independent incSRCNN on each modality, one incSRCNN comprising all channels as separate data, and one incSRCNN that includes only the CARS channel. We found out that training with only the CARS channel produces better results. The training time of this network is approximately 10 minutes compared to 1 hour in the first and second cases.

The incSRCNN reconstruction of the artificial LQ image is shown in Fig. 3 (c-iii). The spatial structures and the color in the image are preserved, and the noise level is reduced. Furthermore, the PSNR has increased to 21, 23, and 22 for the CARS, TPEF, and SHG channels from 14.8, 16, and 18.4, respectively. In addition, the SSIM has significantly increased from 0.27, 0.17, 0.28 to 0.53, 0.52, 0.55 for the CARS, TPEF, SHG channels, respectively. Consequently, the overall PSNR and SSIM were improved from 16.2 and 0.24 to 22 and 0.53, respectively. Furthermore, the average ICC and MAE of the reconstructed image are equal to 0.9 and 0.06, while their values for the artificial image were 0.68 and 0.12. Fig. *4* shows three regions of interest (ROIs) of all reconstructions. The colors in the incSRCNN reconstruction are well conserved, and the noise level in the reconstruction was reduced significantly. In addition, the residual images per channel between the HQ image and the reconstructed one are shown in

Supplement 1 Fig. S1. In this figure, a significant reduction of the values is deduced compared to the artificial LQ case.

Moreover, the histogram of the residual images between the HQ image and the incSRCNN reconstruction in Supplement 1 Fig. S2 significantly reduces the values compared with the artificial LQ case. Therefore, the incSRCNN reconstructed a good representation of the HQ image successfully. Afterward, we tested the performance of the incSRCNN in the experimental LQ image. In Fig. 6 (c-iii), we showed the reconstruction using our proposed network, where the spatial structures and the color in the image are preserved; however, the noise level is slightly reduced. In addition, we compared in Table 2 the PSNR, SSIM, ICC, and MAE between the incSRCNN reconstructions and the HQ image with the PSNR, SSIM, ICC, and MAE between the experiment LQ image and the HQ image. Compared to the experiment LQ image, although the average ICC improved, the PSNR and SSIM values per channel overall slightly decreased. In Fig. 7, our proposed network preserves the color and spatial structures. However, the decrease mentioned above might result from the small data size that the network failed to estimate the values in some areas indicated by the arrow in the figure. However, we continued to assess this matter by checking the intensity values across an arbitrary region and evaluating the incSRCNN network on other noisy experimental LQ data. These noisy LQ data were derived from the experimental LQ images by adding Poisson noise. The results are illustrated in Supplement 1 Fig. S5 and Fig. S6. Figure S5 shows the intensity values for the HQ, LQ, and reconstructed images across the specified region in the image on the top left of the figure. The intensity values in GS differ totally from those in the LQ image, while the deep learning methods maintained a similar trend. However, the smoothed nature of the DnCNN reconstruction can be reflected by showing fewer details than in the incSRCNN reconstruction. Besides, the incSRCNN reconstructions for both noisy experimental LQ images illustrated in Figure S6 showed improvements in terms of PSNR, SSIM, ICC, and MAE values. In addition, we compared the residual images per channel of the incSRCNN reconstructions with the residual images of the experiment LQ image in Supplement 1 Fig. S3. In this figure, almost similar residual values to the experimental LQ case can be detected. Furthermore, we visualized the histogram of the residual images of the DnCNN reconstruction and the experiment LQ images in Supplement 1 Fig. S4.

**Table 2. The PSNR, SSIM, ICC, and the MAE between the HQ and the experimental LQ images and between the HQ and the reconstructed images using the GS algorithm, the DnCNN, and incSRCNN networks**

| Image | Metric | LQ | GS | DnCNN | incSRCNN |
|---|---|---|---|---|---|
| CARS channel | PSNR | 19.028 | 14.888 | **19.228** | 17.778 |
| | SSIM | 0.56 | 0.38 | **0.60** | 0.54 |
| | ICC | 0.86 | 0.82 | **0.88** | 0.86 |
| | MAE | 0.09 | 0.15 | **0.08** | 0.11 |
| TPEF channel | PSNR | 20.027 | 20.046 | **20.787** | 18.758 |
| | SSIM | 0.46 | 0.54 | **0.62** | 0.48 |
| | ICC | 0.76 | 0.79 | **0.82** | 0.8 |
| | MAE | 0.06 | 0.06 | **0.05** | 0.09 |
| SHG channel | PSNR | 20.081 | 20.075 | **20.669** | 19.556 |
| | SSIM | 0.63 | 0.54 | **0.68** | 0.59 |
| | ICC | 0.73 | **0.77** | **0.77** | 0.76 |
| | MAE | **0.05** | 0.07 | **0.05** | 0.06 |
| MM image | PSNR | 19.712 | 18.336 | **20.228** | 18.693 |
| | SSIM | 0.55 | 0.49 | **0.64** | 0.54 |
| | ICC | 0.78 | 0.79 | **0.82** | 0.81 |
| | MAE | 0.07 | 0.09 | **0.06** | 0.08 |

We previously discussed the performance of each method compared to the artificial and experimental LQ images. The GS reconstruction shows similar but poor performance for both artificial and experimental LQ images. The GS reconstructions include dark regions, and the algorithm showed limited abilities even in noiseless settings. In addition, it seems that the

algorithm converges to a local minimum that causes poor reconstructions. However, the DnCNN and the incSRCNN reconstructions preserved the colors and detailed structures. Both networks performed well in the artificial LQ case, but the DnCNN produced smoothed regions critical for medical applications. Our proposed network consists of a simple architecture that only uses the CARS channels and predicts the other two channels. As a result, the PSNR per channel reaches the highest incSRCNN reconstructions. Similar to the artificial case, the DnCNN and incSRCNN networks performed better. These two networks preserve the color and the spatial structures of the image. However, the DnCNN network produced smoothed region, which is a drawback compared to our proposed network that shows a slight reduction in the noise due to the lack of data that the network could not train some regions.

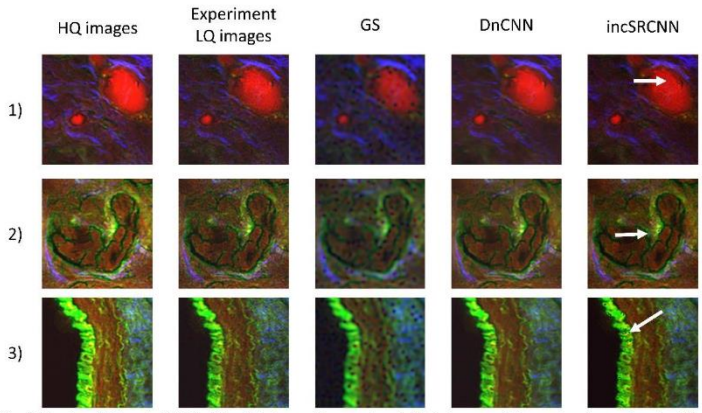


Fig. 7. Region of interest (ROIs) of the HQ image, the experimental LQ image, and its constructions using the GS algorithm, the DnCNN, and the incSRCNN networks. The GS algorithm produces a dark region due to the Gaussian estimation. In the DnCNN reconstruction, some fine structures are lost while were preserved using the incSRCNN network. However, some dark dots were produced, resulting from the lack of data used to train the network.

## 5. Conclusion

The multimodal imaging approach (MM), that combines the CARS, the TPEF, and the SHG modalities provide information on the structure of the measured tissue and its components. However, the MM approach often offers high-quality images, which take more time in the acquisition procedure than a faster MM image measurement, which results in MM images being distorted with noise and other artifacts. Therefore, image denoising techniques are helpful when fast measurements are needed or carried out. However, image denoising techniques feature the drawback that a suitable method needs to be chosen for different settings, which varies between application scenarios. In this context, we compared a classical phase retrieval method via Gerchberg-Saxton (GS) with two deep neural networks; the DnCNN and the incSRCNN. The DnCNN is a pre-trained network that we use as a transfer learning tool, while the incSRCNN is a trained network with our MM images. The data consists of MM images for neck and head tissue. First, we evaluated the GS algorithm, the DnCNN, and the incSRCNN networks on artificial LQ images. Afterward, we tested these three methods on an experimental LQ image.

The artificial LQ image was constructed from the HQ image by adding Poisson noise. The GS algorithm of the artificial LQ image showed poor reconstruction, where dark regions are produced caused by the Gaussian estimation used to describe the input beam. However, the

DnCNN and the incSRCNN reconstructions preserve the color and the spatial structures in the image and improve the PSNR, SSIM, ICC, and MAE compared to the artificial LQ image. However, the DnCNN produced smoothed region that might cause a compromise in the diagnosis of diseases and abnormalities. In addition, the incSRCNN showed the highest PSNR per channel.

Afterward, we compared the performance of the three methods on the experimental LQ image. Like the artificial case, the GS algorithm showed poor performance, and the DnCNN network preserved the color and spatial structures in the images, but smoothed regions were produced. However, the incSRCNN networks maintained the color and the spatial structures in the image and did not produce smoothed areas. However, our proposed network showed a slight decrease in the PSNR, which resulted from the lack of data. In conclusion, a priori knowledge of the beam source is vital for the GS reconstruction, and the algorithm has limited recovery abilities even in a noiseless setting.

In summary, the deep learning networks produced very promising results. However, the DnCNN network preserved the color and spatial structures of the image but produced smoothed regions, resulting in the loss of relevant information. However, our proposed network, the incSRCNN, consists of simple architecture, and it reconstructs the complex structures of the testing image and shows better PSNR than the other methods. Nevertheless, the incSRCNN network produced some artifacts represented by arrows in the zoomed figures, resulting from the lack of data used to train the network. On the other hand, the shorter time to reconstruct an HQ image, run on a limited CPU computer, is through the incSRCNN network. Additionally, only 0.22 seconds is needed for predicting a patch of the image.

**Funding.** This work is supported by the BMBF, funding program Photonics Research Germany (FKZ: 13N15706 (LPI-BT2-FSU) and 13N15710 (LPI-BT3-FSU)) and is integrated into the Leibniz Center for Photonics in Infection Research (LPI). The LPI initiated by Leibniz-IPHT, Leibniz-HKI, UKJ and FSU Jena is part of the BMBF national roadmap for research infrastructures. It is also supported by the Free State of Thuringia under the number 2019 FGR 0083 and co-financed by European Union funds within the European Social Fund (ESF) framework via the TAB-FG MorphoTox. In addition, the research was supported by the Digleben project (5575/10-9).

TMZ, MS, OGL, and JP gratefully acknowledge funding by the BMBF (TheraOptik, 13GW0370E). This project has also received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 101016923 (CRIMSON).

**Disclosures.** The authors declare no conflicts of interest.

**Data availability.** Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

**Supplemental document.** See Supplement 1 for supporting content.

## References

1. N. Vogler, A. Medyukhina, I. Latka, S. Kemper, M. Böhm, B. Dietzek, and J. Popp, "Towards multimodal nonlinear optical tomography – experimental methodology," Laser Phys. Lett. **8**, 617 (2011).
2. C. A. Patil, N. Bosschaart, M. D. Keller, T. G. van Leeuwen, and A. Mahadevan-Jansen, "Combined Raman spectroscopy and optical coherence tomography device for tissue characterization," Opt. Lett. **33**, 1135–1137 (2008).
3. P. C. Ashok, B. B. Praveen, N. Bellini, A. Riches, K. Dholakia, and C. S. Herrington, "Multi-modal approach using Raman spectroscopy and optical coherence tomography for the discrimination of colonic adenocarcinoma from normal colon," Biomed. Opt. Express **4**, 2179–2186 (2013).
4. A. T. Yeh, B. Kao, W. G. Jung, Z. Chen, J. S. N. M.d, and B. J. Tromberg, "Imaging wound healing using optical coherence tomography and multiphoton microscopy in an in vitro skin-equivalent tissue model," J. Biomed. Opt. **9**, 248–253 (2004).
5. N. Iftimia, R. D. Ferguson, M. Mujat, A. H. Patel, E. Z. Zhang, W. Fox, and M. Rajadhyaksha, "Combined reflectance confocal microscopy/optical coherence tomography imaging for skin burn assessment," Biomed. Opt. Express **4**, 680–695 (2013).
6. K. Kong, C. J. Rowlands, S. Varma, W. Perkins, I. H. Leach, A. A. Koloydenko, H. C. Williams, and I. Notingher, "Diagnosis of tumors during tissue-conserving surgery with integrated autofluorescence and Raman scattering microscopy," Proc. Natl. Acad. Sci. **110**, 15189 (2013).

7. N. Goel, A. Yadav, and B. M. Singh, "Medical image processing: A review," in *2016 Second International Innovative Applications of Computational Intelligence on Power, Energy and Controls with Their Impact on Humanity (CIPECH)* (2016), pp. 57–62.
8. S. Sagheer and S. George, "A review on medical image denoising algorithms - ScienceDirect," https://www.sciencedirect.com/science/article/pii/S1746809420301920.
9. JNTUH University, Telangana, India and also Dept of ECE, S R Engineering College (Autonomous), Warangal, India, S. Kollem, K. R. L. Reddy, and D. S. Rao, "A Review of Image Denoising and Segmentation Methods Based on Medical Images," Int. J. Mach. Learn. Comput. **9**, 288–295 (2019).
10. R. J. G. van Sloun, R. Cohen, and Y. C. Eldar, "Deep Learning in Ultrasound Imaging," Proc. IEEE **108**, 11–29 (2020).
11. "Deep learning for real-time semantic segmentation: Application in ultrasound imaging | Elsevier Enhanced Reader," https://reader.elsevier.com/reader/sd/pii/S0167865521000234?token=1F17833B8DB5CF25E3C38ADEFA3AE2D4975E6C1BFAC138281DF90C5AB29B50C52D715EE106463CD86F0980846FA8B4B9&originRegion=eu-west-1&originCreation=20211130102003.
12. S. Vedula, O. Senouf, A. M. Bronstein, O. V. Michailovich, and M. Zibulevsky, "Towards CT-quality Ultrasound Imaging using Deep Learning," ArXiv171006304 Phys. (2017).
13. M. Grewal, M. M. Srivastava, P. Kumar, and S. Varadarajan, "RADnet: Radiologist level accuracy using deep learning for hemorrhage detection in CT scans," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (2018), pp. 281–284.
14. A. Bhandary, G. A. Prabhu, V. Rajinikanth, K. P. Thanaraj, S. C. Satapathy, D. E. Robbins, C. Shasky, Y.-D. Zhang, J. M. R. S. Tavares, and N. S. M. Raja, "Deep-learning framework to detect lung abnormality – A study with chest X-Ray and lung CT scan images," Pattern Recognit. Lett. **129**, 271–278 (2020).
15. R. W. Gerchberg and W. O. Saxton, "Comment on `A method for the solution of the phase problem in electron microscopy'," J. Phys. Appl. Phys. **6**, L31–L32 (1973).
16. G. Yang, B. Dong, B. Gu, J. Zhuang, and O. K. Ersoy, "Gerchberg–Saxton and Yang–Gu algorithms for phase retrieval in a nonunitary transform system: a comparison," Appl. Opt. **33**, 209–218 (1994).
17. J. R. Fienup, "Phase retrieval algorithms: a comparison," Appl. Opt. **21**, 2758–2769 (1982).
18. F. Fogel, I. Waldspurger, and A. d'Aspremont, "Phase retrieval for imaging problems," Math. Program. Comput. **8**, 311–335 (2016).
19. G. Whyte and J. Courtial, "Experimental demonstration of holographic three-dimensional light shaping using a Gerchberg–Saxton algorithm," New J. Phys. **7**, 117–117 (2005).
20. K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising," IEEE Trans. Image Process. **26**, 3142–3155 (2017).
21. C. Dong, C. C. Loy, K. He, and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," ArXiv150100092 Cs (2015).
22. C. Dong, C. C. Loy, K. He, and X. Tang, "Image Super-Resolution Using Deep Convolutional Networks," IEEE Trans. Pattern Anal. Mach. Intell. **38**, 295–307 (2016).
23. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," ArXiv14091556 Cs (2015).
24. S. Heuke, N. Vogler, T. Meyer, D. Akimov, F. Kluschke, H. J. Rowert-Huber, J. Lademann, B. Dietzek, and J. Popp, "Detection and Discrimination of Non-Melanoma Skin Cancer by Multimodal Imaging," Healthc. Basel **1**, 64–83 (2013).
25. J. Xiao, Z. Liu, P. Zhao, Y. Li, and J. Huo, "Deep Learning Image Reconstruction Simulation for Electromagnetic Tomography," IEEE Sens. J. **18**, 3290–3298 (2018).

# Comparison of denoising tools for reconstruction of nonlinear multimodal images: supplement 1

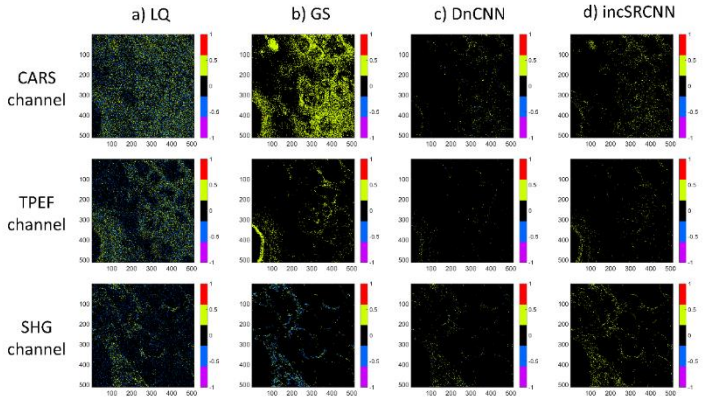


Fig. S1. The residual images between each channel of the experiment HQ image and the artificial LQ image (a) and its reconstruction using the GS algorithm (b), the DnCNN (c), and the incSRCNN (d) networks. The standard deviation of the residual of the artificial LQ image shown in a) is 0.2, 0.2, and 0.1 for CARS, TPEF, and SHG, respectively. In b), the standard deviation of the residual in CARS, TPEF, and SHG channels decreased to 0.1, 0.07, and 0.08, respectively; however, the spatial structure is still present, reflecting the poor performance of this algorithm. In c) since most image values are zero, the performance is good, and the standard deviation was reduced to 0.07, 0.05, 0.05 for CARS, TPEF, and SHG channels, respectively. Similar to the DnCNN, in d) the performance in incSRCNN is good, and the standard deviation was also reduced to 0.08, 0.06, and 0.06 for CARS, TPEF, and SHG channels, respectively.

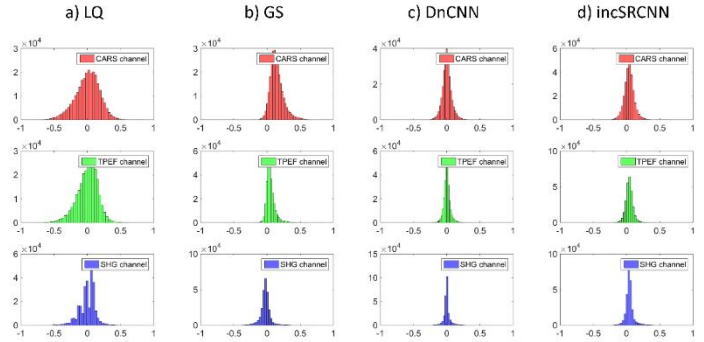


Fig. S2. The histogram of the residual images of the artificial LQ images (a) and its reconstruction using the GS algorithm (b), the DnCNN (c), and the incSRCNN (d) networks. In b), the GS reconstruction reduces values compared to the artificial LQ image. In c), a significant noise reduction is particularly shown for the CARS and TPEF channels.

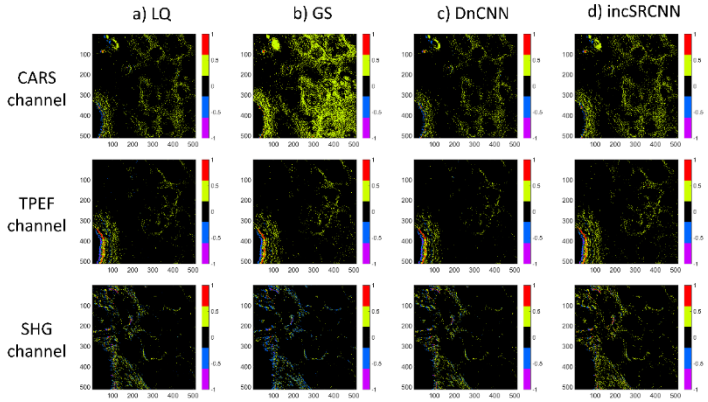


Fig. S3. The residual images between each channel of the experiment HQ image and the experimental LQ image (a) and its reconstruction using the GS algorithm (b), the DnCNN (c), and the incSRCNN (d) networks. The standard deviation of the residual of the experimental LQ image shown in a) is 0.09, 0.09, and 0.1 for CARS, TPEF, and SHG, respectively. In b), the standard deviation of the residual in the CARS channel increased to 0.1 and decreased to 0.09 and 0.09 for TPEF and SHG channels, respectively; however, the spatial structure is still present, reflecting the poor performance of this algorithm. In c), some of the structures are present, which suggests that the DnCNN network cannot perform as well as in the artificial images; however, the standard deviation is reduced to 0.08, 0.08, 0.09 for CARS, TPEF, and SHG channels, respectively. Similar to the DnCNN, in d) the structures are still present in the residual of incSRCNN; however, the standard deviation was also reduced to 0.08, 0.08, and 0.09 for CARS, TPEF, and SHG channels, respectively.

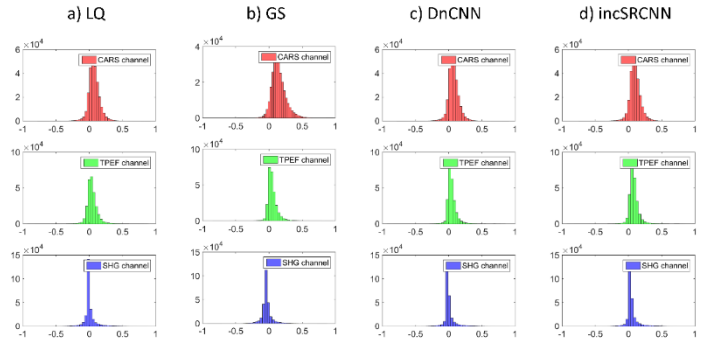


Fig. S4. The histogram of the residual images of the experimental LQ images (a) and its reconstruction using the GS algorithm (b), the DnCNN (c), and the incSRCNN (d) networks. In b), the GS reconstruction reduces values compared to the artificial LQ image. In c), a significant noise reduction is particularly shown for the CARS and TPEF channels.

**Table S1. The time in seconds for reconstructing one channel using the GS algorithm, DnCNN, and incSRCNN networks run on a limited CPU. The results showed the outperformance of the deep learning methods, particularly the incSRCNN, which refers to the simple architecture**

| GS | DnCNN | incSRCNN |
|---|---|---|
| 2871.2 | 40.5 | 7.6 (0.22 per one patch) |

**Table S2. The total number of images used in incSRCNN training before and after data augmentation**

| Set | The original number of images | Number of images after data augmentation per channel | Number of patches after data augmentation per channel | Total number of data |
|---|---|---|---|---|
| Training | 7 | 63 | 1008 | 3024 |
| Validation | 2 | 18 | 288 | 864 |
| Testing | 1 | - | - | - |

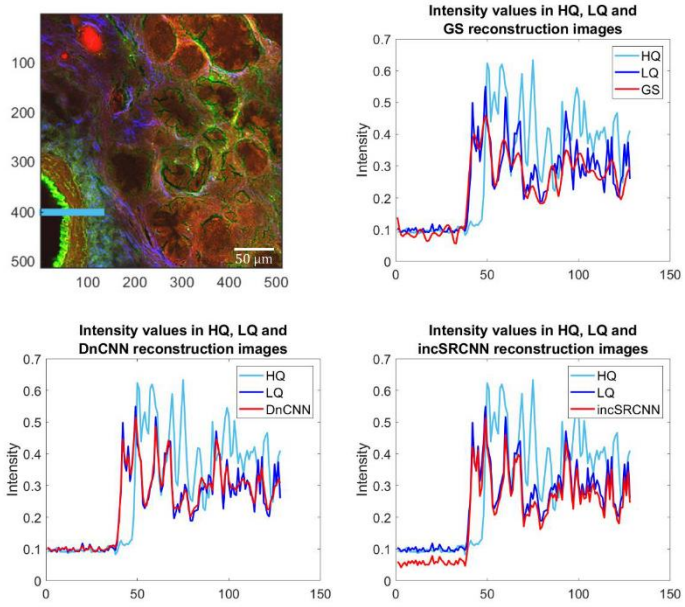


Fig. S5. The intensity values in HQ, experimental LQ, and reconstructions grayscale images on an arbitrary region. The intensity values in the GS reconstruction differ from those in the LQ image, and the trend changed totally. Although DnCNN retained the trend, it showed fewer details since it provides smoother results. While incSRCNN conserved details and retained trend in intensity values.

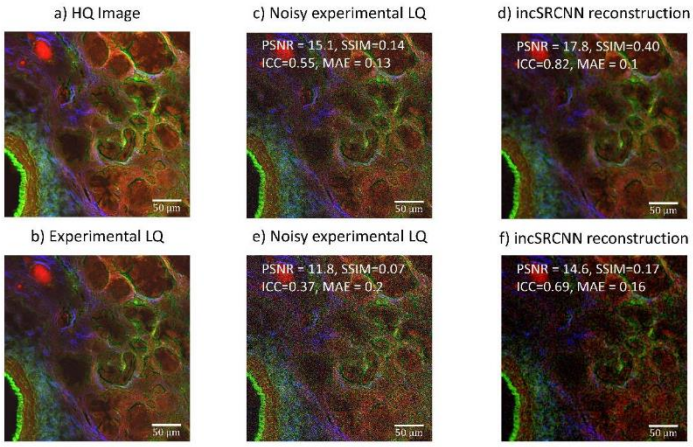


Fig. S6. The incSRCNN reconstructions of two noisy experimental LQ images. The two noisy images were derived from the experiment LQ by adding Poisson noise. The PSNR of the two noisy LQ images illustrated in c) and e) calculated with the experiment LQ image as reference was 16.67 and 11.89, respectively. The evaluation of the reconstructions is achieved through the PSNR, SSIM, ICC, and MAE metrics calculated with the HQ image as reference. In d) and f), the incSRCNN reconstructions preserved the structure, and all metrics showed improved values even though a high noise level was used.

# P4. COMPARISON OF FUNCTIONAL AND DISCRETE DATA ANALYSIS REGIMES FOR RAMAN SPECTRA

Reproduced, with permission, from [R. Houhou, P. Rösch, J. Popp, and T. Bocklitz, Comparison of functional and discrete data analysis regimes for Raman spectra, Analytical and Bioanalytical Chemistry, 413, 5633-5644, 2021] Copyright 2021 Springer Nature.

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

| **R. Houhou[1], P. Rösch[2], J. Popp[3], and T. Bocklitz[4], Comparison of functional and discrete data analysis regimes for Raman spectra, Analytical and Bioanalytical Chemistry, 413, 5633-5644, 2021** | | | | |
|---|---|---|---|---|
| Involved in | 1 | 2 | 3 | 4 |
| | | | | |
| Conceptual research design | × | | × | × |
| Planning of research activities | × | | | × |
| Data collection | | × | | |
| Data analyses and interpretation | × | × | × | × |
| Manuscript writing | × | × | × | × |
| Suggested publication equivalence value | 1.0 | | | |

Analytical and Bioanalytical Chemistry
https://doi.org/10.1007/s00216-021-03360-1

RESEARCH PAPER

# Comparison of functional and discrete data analysis regimes for Raman spectra

Rola Houhou[1,2] · Petra Rösch[1] · Jürgen Popp[1,2] · Thomas Bocklitz[1,2]

Received: 5 March 2021 / Revised: 13 April 2021 / Accepted: 16 April 2021
© The Author(s) 2021

**Abstract**
Raman spectral data are best described by mathematical functions; however, due to the spectroscopic measurement setup, only discrete points of these functions are measured. Therefore, we investigated the Raman spectral data for the first time in the functional framework. First, we approximated the Raman spectra by using B-spline basis functions. Afterwards, we applied the functional principal component analysis followed by the linear discriminant analysis (FPCA-LDA) and compared the results with those of the classical principal component analysis followed by the linear discriminant analysis (PCA-LDA). In this context, simulation and experimental Raman spectra were used. In the simulated Raman spectra, normal and abnormal spectra were used for a classification model, where the abnormal spectra were built by shifting one peak position. We showed that the mean sensitivities of the FPCA-LDA method were higher than the mean sensitivities of the PCA-LDA method, especially when the signal-to-noise ratio is low and the shift of the peak position is small. However, for a higher signal-to-noise ratio, both methods performed equally. Additionally, a slight improvement of the mean sensitivity could be shown if the FPCA-LDA method was applied to experimental Raman data.

**Keywords** Raman spectroscopy · Principal component analysis · Functional data analysis · B-splines · Functional principal component analysis

## Introduction

When light interacts with molecules within a sample volume, the result of the interaction is depending on both properties of the sample and the light. The study of such light-matter interactions is done by spectroscopic techniques. In this manuscript, we focus on Raman spectroscopy, which is an inelastic scattering process. The scattered light contains information about energy levels within the molecules, which result from the vibrational and rotational modes. In Raman spectroscopy, the obtained spectral data contain a vast amount of information concerning the molecules within the sample. However, this information cannot directly be used and the extraction of the information needs chemometric methods. Hence, the combination of these chemometric methods with Raman spectroscopy enables the extraction of the relevant information and increases the knowledge regarding the composition of a sample. The combination of Raman spectroscopy and chemometrics has gained popularity and can be used to address a number of tasks, like disease diagnostics and bacteria identification [1–6].

Typically, chemometric methods are split into univariate and multivariate methods, supervised and unsupervised methods, or qualitative and quantitative methods. These methods are typically understood in the discrete case, where the data is discrete. However, a different data analysis approach is needed if functions should be analyzed. Therefore, we can further separate chemometric methods into two additional groups, the discrete group and the functional data analysis group. The chemometric methods in the discrete group, which is also called multivariate data analysis, are most often implemented in Raman spectroscopy. In this group, the key concept is to consider each spectrum as a set of independent points acquired on a specific

Published in the topical collection *Recent Trends in (Bio)Analytical Chemistry* with guest editors Antje J. Baeumner and Günter Gauglitz.

✉ Thomas Bocklitz
thomas.bocklitz@uni-jena.de

[1] Institute of Physical Chemistry, Friedrich Schiller University Jena, Helmholtzweg 4, 07743 Jena, Germany

[2] Department of Photonic Data Science, Leibniz Institute of Photonic Technologies, Member of Leibniz Research Alliance "Leibniz-Health Technologies", Albert-Einstein-Str. 9, 07745 Jena, Germany

Published online: 15 May 2021

interval, e.g., wavenumber or frequency. Moreover, various models and tasks are contained in this group, e.g., dimension reduction, clustering, regression, and classification [7, 8]. For instance, the principal component analysis (PCA) is a well-known method to reduce the dimensionality in spectral data analysis [9–11]. For classification tasks, many methods are developed and implemented for spectroscopic measurements, e.g., linear discriminant analysis (LDA) [12–14], support vector machine [15–17], and neural networks [18]. In contrast to the methods in the discrete data analysis group, a spectrum is modeled as a function or a curve in the functional data analysis group. These functions are latent and infinite-dimensional, which cannot be calculated analytically and they need to be approximated. The study of these functions falls under the name functional data analysis (FDA), which originally was introduced by Ramsay et al. [19, 20]. It involves smoothing technique, data reduction, adjustment for clustering, functional linear modeling, and forecasting methods [21–24]. Initially, Ramsay et al. developed the FDA to analyze, model, and predict time series and then expanded the FDA to cover other data types. The advantages of using chemometric methods in the functional group are that FDA overcomes the curse of dimensionality. Moreover, the assumption that adjacent observations should be independent is not needed in the functional data analysis group, while in the discrete group, this assumption is needed but often violated.

The chemometric methods used in the functional data analysis group are adapted from the discrete ones. For instance, the functional principal component analysis (FPCA) is an adapted version of a discrete PCA. The FPCA deals with data in the form of functions and was developed by Dauxois et al. [25]. James et al. [26] introduced the functional linear discriminant analysis (FLDA) extended from the classical LDA method, where the predictor variables are curves or functions. Mas et al. [27] introduced the functional version of linear regression for random functions by considering the first-order derivative. Furthermore, the application of these functional methods was spread to various fields, including medicine [28–30], economics [31, 32], agriculture [33], linguistics [34–36], and behavior sciences [37]. However, as far as the authors know, they were never applied to Raman spectral data, and their application was limited to mass spectrometry [38] and near-infrared spectroscopy [39]. Although the Raman spectral data are in nature functions of wavenumber or frequency, they are acquired discretely on finite points due to the used spectroscopic setup. In this manuscript, we evaluated the analysis of Raman spectra in the functional framework for the first time. This evaluation was achieved by comparing the performance of the functional principal component analysis followed by the linear discriminant analysis (FPCA-LDA) to the classical principal component analysis followed by the linear discriminant analysis (PCA-LDA) on simulation and experimental Raman spectral data.

The manuscript is divided into four sections. First, we presented the theoretical background of the functional data analysis and the functional principal component analysis. The workflow of the two methods, the classical PCA-LDA and the FPCA-LDA, is explained in the "Material and method" section. In the "Results" section, the comparison between these two methods applied on simulation and experimental Raman data is shown. Finally, we summarized the main findings in the "Conclusion" section.

## Theoretical background

Functional data analysis (FDA) refers to the analysis of data in the form of functions. For instance, if our studied data is collected in a matrix $\boldsymbol{x} \in \mathbb{R}^{N \times p}$, then each row of this data is considered a function. The underlying idea of FDA is that we assume the existence of some functions $\boldsymbol{x}_i(t)$, $i = 1, \ldots, N$, giving rise to the observed data. Therefore, this function is treated as one entity instead of a sequence of individual measured variables.

### Functional approximation

The analysis of functions in the functional framework is done in some functional space, which is often assumed to be a Hilbert space, such as $L^2(I)$ defined on a compact interval $I$. However, we cannot analytically calculate these functions from the data, and we need to apply basis functions to approximate them. These basis functions represent a set of available functions $\varphi_k$, $k = 1, 2, \ldots, K$ that are mathematically independent of each other. With a linear combination of adequately large number of these basis functions, we can approximate our observed data. This approximation can be formulated by a linear expansion of $K$ known basis functions $\varphi_k$ as follows:

$$\boldsymbol{x}_i(t) = \sum_{k-1}^{K} c_k \varphi_k(t). \tag{1}$$

Likewise Eq. (1) can be expressed in matrix notation, as follows [40, 41]:

$$\boldsymbol{x}_i = \boldsymbol{C}' \boldsymbol{\phi} = \boldsymbol{\phi}' \boldsymbol{C}, \tag{2}$$

where $\boldsymbol{C}$ is the vector of length $K$ of the coefficients $c_k$ and $\boldsymbol{\phi}$ is the functional vector whose elements are the basis functions $\varphi_k$. The choice of the basis functions depends on the studied problem, where they should have features similar to those known to belong to the estimated functional data. However, most functional data analysis involves either a Fourier basis for periodic data or a B-spline basis for non-periodic data. The B-spline basis functions are the most common choice for approximating spectral data. These basis

functions are defined on a specific interval $I$ with an order $O$ and a knot vector. First, the interval $I$ is divided into subintervals where the corresponding endpoints are represented by the knot vector $\{t_1, \ldots, t_{K+O}\}$. After imposing continuity and smoothness conditions, the $K$ B-spline basis functions of order $O$ can be defined using Eq. (3):

$$\varphi_{k,1}(t) = \begin{cases} 1, & t_k \le t < t_{k+1} \\ 0, & else \end{cases}$$
$$\varphi_{k,O}(t) = \frac{t - t_k}{t_{k+O} - t_k}\varphi_{k,O-1}(t) + \frac{t_{k+O+1} - t}{t_{k+O+1} - t_{k+1}}\varphi_{k+1,O-1}(t). \quad (3)$$

Consequently, the approximation is achieved by calculating the coefficients $c_k$. Various methods can be implemented for this purpose; however, the simplest one is to minimize the sum of squared errors or the so-called least squares estimation using Eq. (4):

$$\mathrm{SMSSE}(\boldsymbol{x_i}) = \sum_{j-1}^{p} \left| x_{ij} - \sum_{k-1}^{K} c_k \varphi_k(t_j) \right|^2 = \|\boldsymbol{x_i} - \boldsymbol{\phi C}\|^2. \quad (4)$$

The choice of the number of basis functions $K$ has an influence on the approximation and it should be chosen carefully. With a high number of basis function, we risk fitting noise, and with a small $K$, we might remove some important aspect of the functional data. Therefore, methods exist, which can be used to obtain an optimal number of basis functions, e.g., the elbow method, the bias/variance trade-off method, and the stepwise variable selection method [42].

In contrast to the multivariate data analysis, the functional data in the FDA procedure are inherently infinite-dimensional, which makes the computation at any value $t$ possible. Besides, the underlying functions in FDA are smooth, but the observed data are often not due to the presence of noise in the measurements. Therefore, a higher level of variation in the observed data can make the extraction of a stable estimate of the functional data challenging. Furthermore, sparse and irregular observed longitudinal data can be analyzed in FDA.

## Functional principal components analysis (FPCA)

Functional principal component analysis (FPCA) is a key dimension reduction tool for functional data, and it is considered the most popular method in the functional analysis. Similarly to the classical principal component analysis (PCA), we need to examine the variance-covariance matrix/function in order to calculate the principal components. The theoretical background of FPCA [43] is shown below in detail; however, since it represents a functional version of the classical PCA, a short introduction to the theory of PCA is first presented.

The key idea of PCA is to construct the $i^{th}$ principal component $\boldsymbol{f_i}$ as a linear combination of the variable $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})'$ as shown in Eq. (5):

$$\boldsymbol{f_i} = \sum_{j-1}^{p} \beta_j x_{ij}, i = 1, \ldots, N \quad (5)$$

where $\beta_j$ is the weight coefficient of the observed values $x_{ij}$ of the $j^{th}$ variable. In PCA, we start by finding the weight vector $\boldsymbol{\beta_1} = (\beta_{11}, \ldots, \beta_{p1})'$ such as $f_{i1} = \sum_j \beta_{j1}\, x_{ij} = \beta_1' x_i$ have the largest possible mean square $N^{-1}\sum_i f_{i1}^2$ subject to the following constraint $\sum_j \beta_{j1}^2 = \|\boldsymbol{\beta_1}\|^2 = 1$. Then, we proceed to the second step and subsequent steps until reaching a desired number, which should be less or equal to the number of variables $p$. On the $m^{th}$ step, similar to the first step, we compute a new weight vector $\boldsymbol{\beta_m}$ and new values $f_{im}$; thus, the values $f_{im}$ have maximum mean square, subject to the constraint $\|\boldsymbol{\beta_m}\|^2 = 1$ and the $m-1$ additional constraints $\sum_j \beta_{jq}\, \beta_{jm} = \boldsymbol{\beta_q'\beta_m} = 0, q < m$. These $f_{im}$ are called the principal component scores.

In contrast to the classical PCA, the variable values in FPCA are function $\boldsymbol{x_i}(t)$, and the equivalent notation of $\boldsymbol{\beta}$ and $\boldsymbol{x}$ in FPCA are the functions $\beta(t)$ and $\boldsymbol{x}(t)$. Therefore, the principal component scores corresponding to the weight function are illustrated in Eq. (6):

$$\boldsymbol{f_i} = \int \boldsymbol{\beta x_i} = \int \boldsymbol{\beta}(t)\boldsymbol{x}(t)dt. \quad (6)$$

The first step in FPCA is to find the weight function $\boldsymbol{\beta_1}(s)$ in such a way that it maximizes $N^{-1}\sum_i f_i = N^{-1}\sum_i \left(\int \beta_1\, x_i\right)^2$ and subject to the unit sum of squares constraint $\int \boldsymbol{\beta_1}(s)^2 = 1$. Then, we proceed to the $m$ step, where we find the weight function $\beta_m$ that satisfies the orthogonality constraints $\int \boldsymbol{\beta_q\beta_m} = 0,\ q < m$. However, in most principal component analysis applications, finding the principal components is equivalent to finding the eigenvalues and eigenfunctions of the covariance function. Therefore, the covariance function $\boldsymbol{v}(\boldsymbol{s}, \boldsymbol{t})$ is defined as follows:

$$\boldsymbol{v}(\boldsymbol{t}, \boldsymbol{s}) = N^{-1} \sum_{i-1}^{N} \boldsymbol{x_i}(t)\boldsymbol{x_i}(s) \quad (7)$$

And each eigenfunction $\boldsymbol{\beta_j}(t)$ for an appropriate eigenvalue $\rho$ satisfies

$$\int \boldsymbol{v}(\boldsymbol{t}, \boldsymbol{s})\boldsymbol{\beta}(s)ds = \rho\boldsymbol{\beta}(t) \quad (8)$$

The left side of this equation is an integral transform $\boldsymbol{V}$ of the weight function $\boldsymbol{\beta}$ that can be defined by Eq. (9), and it is called the covariance operator $\boldsymbol{V}$. Therefore, we may also express the eigenequation directly as Eq. (10), where $\boldsymbol{\beta}$ is an eigenfunction rather than an eigenvector.

$$\boldsymbol{V\beta} = \int \boldsymbol{v}(., \boldsymbol{t})\boldsymbol{\beta}(t)dt \quad (9)$$
$$\boldsymbol{V\beta} = \rho\boldsymbol{\beta} \quad (10)$$

In classical PCA, the number of variables is equal to $p$. In contrast, in the case of functional PCA (FPCA), the number of variables refers to the number of function values which is

infinity. However, given that the functions $x_i(t)$ are not linearly dependent, the operator will have rank $N-1$, and there will be $N-1$ nonzero eigenvalues.

## Material and method

Our study is divided into two parts. First, we applied the classical principal component analysis as described previously on our data stored in a $N \times p$ matrix $\boldsymbol{x}$, and then we extracted the principal component scores matrix $\boldsymbol{\beta}$. Instead of applying the linear discriminant analysis directly on the data, we used these scores as input to the LDA model. This latter model calculates $L-1$ linear discriminant functions to separate $L$ groups, as shown in Eq. (11):

$$\boldsymbol{LD} = \boldsymbol{\Omega\beta} + \boldsymbol{b_0} \tag{11}$$

which represents a linear combination of the principal component scores, where $\boldsymbol{\Omega}$ represents the weight matrix and $\boldsymbol{b_0}$ is the bias. Briefly, these functions are calculated by maximizing the between-group variance illustrated in Eq. (12):

$$\boldsymbol{B} = \sum_{l=1}^{L} n_l \left(\overline{\beta}_l - \overline{\beta}\right)\left(\overline{\beta}_l - \overline{\beta}\right)' \tag{12}$$

and minimizing the within-group variation calculated in Eq. (13):

$$\boldsymbol{W} = \sum_{l=1}^{L} \left(\beta_{.l} - \overline{\beta}_l\right)\left(\beta_{.l} - \overline{\beta}_l\right)'. \tag{13}$$

In contrast to the discrete framework, we implemented the functional framework by the following steps. We transformed our discrete data into functions using cubic B-spline basis functions where the order $O$ is equal to 4 and we approximated the coefficients $c_k$ using Eq. (1) and the least square method. Then, the functional principal component analysis was applied to these functional versions of our data, and the functional scores were extracted. We then apply LDA on these functional scores. The testing sets in the classical PCA-LDA and FPCA-LDA were projected into the corresponding principal component space. Moreover, both methods were used inside a cross-validation loop and the results are obtained using group scripts developed in Gnu R software. The motivation and the workflow of both PCA-LDA and FPCA-LDA are illustrated in Fig. 1.

## Results

We tested the performance of the functional data analysis by comparing the functional principal component analysis followed by linear discriminant analysis (FPCA-LDA) and the classical principal component analysis followed by linear discriminant analysis (PCA-LDA) on both simulation and experimental Raman data.

### Simulation data

First, we simulated Raman spectra of two classes (a normal group and an abnormal group) with three peaks. The abnormal group was generated by slightly shifting one of the peaks from the peak position used in the normal group. This situation is often occurring in biomedical Raman spectroscopy, when for example the protein's secondary structure changes between two groups or if an isotope labeling was applied. We did that in two scenarios without adding a background and with adding a background contribution. The two classes in the simulated Raman spectra are referred to as normal and abnormal

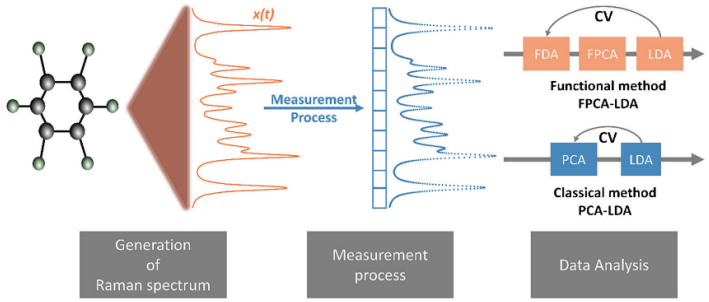


**Fig. 1** The motivation and the workflow of the PCA-LDA and the FPCA-LDA methods. On the left, the generation of a Raman spectrum is visualized, which yields a Raman spectrum in functional form. However, due to the measurement process (in the middle), we acquire the spectrum in a discrete manner. Due to the used multichannel detector, a measured Raman spectrum is characterized by a vector of intensities. On the right, the data analysis workflow that aims to compare the classical principal component analysis followed by linear discriminant analysis (PCA-LDA) and the functional principal component analysis followed by linear discriminant analysis (FPCA-LDA) is shown. Both methods include a cross-validation (CV) loop and the functional data is constructed by using B-splines

spectra. The number of pixels in each spectrum is equal to 1024. The abnormal spectra were constructed similarly to the normal spectra with three peaks, but only one of these peaks was shifted by one of the following values ($\Delta\tilde{\upsilon}$): 0.001, 0.003, 0.005, 0.01, 0.02, 0.025, 0.05. The total number of spectra in each class was set to 100 spectra. In addition, the simulation data was constructed for different signal-to-noise ratio (SNR) cases: 0.5, 1, 2, 3, 5, 10, 30, 50, 100. We end up with 63 cases of simulation data where each of the datasets includes 200 spectra. The same cases were constructed including a random background. The parameters of the simulation are summarized in Supplementary Information (ESM) Table S1. An example of the mean spectra for the simulated Raman where the shift in the peak position is equal to 0.05 and the SNR is 30 is shown in Fig. S1 in the ESM.

For the simulated Raman spectra without background, 190 B-splines are used in the FDA approximation. An illustration of the first 10 B-splines basis functions can be found in the ESM in Fig. S2. The choice of the number of basis functions is calculated based on the elbow method. We first calculated the root-mean-square error, and the optimal point is chosen, which refers to the largest distance to the line that joins the first and last values. After choosing the number of basis functions, we transform our discrete spectra into functional version using Eq. (1). The mean spectra for both discrete and functional versions in the case where $\Delta\tilde{\upsilon} = 0.05$ and SNR = 0.5 are illustrated in Fig. 2. On the left, the original mean spectra for the normal and the abnormal classes are shown in black and red, respectively. In comparison, the functional mean spectra for normal and abnormal classes are shown on the right in black and red, respectively. The functional spectra represent well the original spectra, where the shape of the peaks is preserved and a significant reduction of the noise was observed.

The results of the functional approximation for all the constructed cases are illustrated in Figs. S3 and S4 in the ESM. In ESM Fig. S3, we showed the mean spectra for the simulated Raman without background per shift when the SNR is equal to 0.5 for the original discrete spectra and its functional approximation. On the left, the original mean spectra for the simulation without background are shown. The black plot represents the mean spectra of the normal class, while the colored plots refer to the mean spectra of the abnormal class corresponding to the specific shift values $\Delta\tilde{\upsilon}$. On the right, the mean spectra of the functional approximation of the corresponding discrete spectra are illustrated. When comparing the functional mean spectra with the discrete mean spectra, we can deduce a significant reduction of noise and an improvement in the peak shape estimation.

In ESM Fig. S4, we showed one spectrum per class for each SNR and for a specific peak shift $\Delta\tilde{\upsilon} = 0.01$. On the left, the original spectra are illustrated. In each row, the normal and abnormal spectra are plotted for all SNR values. On the right, the functional approximations for each class are illustrated for the nine cases of SNR. In all these cases, we used 190 B-spline basis functions in the functional approximation. For a SNR which is larger than 5, the functional approximations perfectly represent the original spectra with almost noiseless reconstruction. Although the functional approximation also fit noise for the lower SNR values due to a high number of basis used, a significant reduction of noise is noticed in the approximation, and it maintains potentially the peak shape in the case of lower SNR values.

The two approaches (PCA-LDA and FPCA-LDA) explained previously were applied to the simulated Raman spectra that contain no background. For both methods, we used 10-fold cross-validation, and the number of components chosen

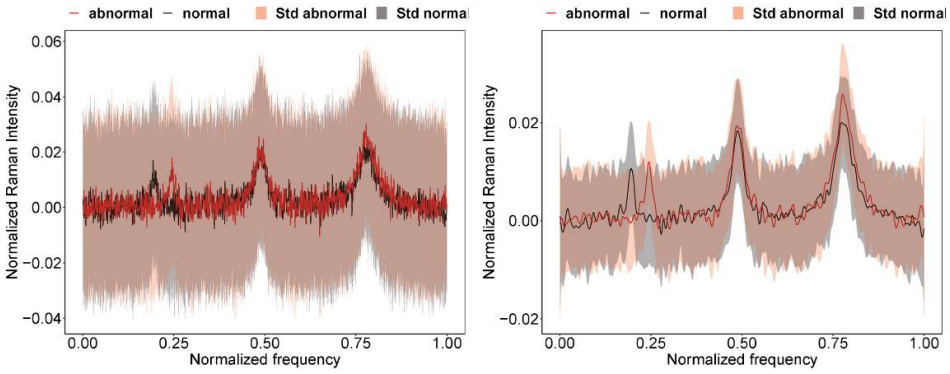


**Fig. 2** The mean spectra per class for the simulation without a background in the case where $\Delta\tilde{\upsilon} = 0.05$ and SNR = 0.5. The left plot represents the discrete simulated Raman mean spectra per each class, while the right plot represents the functional version of these simulated Raman spectra using 190 basis functions

was 50 components. The mean sensitivity was used as an evaluation metric, and it is illustrated in a heat map in Fig. 3. In region c of Fig. 3, PCA-LDA and FPCA-LDA perform perfectly with 100% mean sensitivity, due to the clear distinction between the normal and the abnormal classes. While in regions a and b of the same figure, the FPCA-LDA performs better in most of the cases, due to the fact that in the functional approximation, the obtained functions contain less noise and an improvement of the shape of the peaks in these functions was also detected. However, in few cases, the PCA-LDA provides better mean sensitivity. This result might refer to the choice of the number of basis functions that should not be too large that the approximation includes more noise or too small that the approximation excludes some relevant features in the approximation.

In the simulated Raman spectra that include a random background, 180 B-spline basis functions are used in the FDA approximation. The choice of the number of basis functions is calculated similarly to the simulation without background based on the elbow method. After choosing the number of basis functions, we transform our discrete spectra into functional versions using Eq. (1). The mean spectra of both discrete and functional versions in the case $\Delta\tilde{\upsilon} = 0.05$ and SNR = 0.5 are illustrated in the ESM in ESM Fig. S5. On the left, the original mean spectra for the normal and the abnormal classes are shown in black and red, respectively. The functional mean spectra for normal and abnormal classes are shown on the right in black and red, respectively. The functional spectra represent well the original spectra, where the shape of peaks is preserved, and a significant reduction of noise was observed.

The results of the functional approximation are illustrated in the ESM in Figs. S6 and S7. In ESM Fig. S6, we showed the mean spectra for the simulated Raman data with background per each shift and for SNR 0.5 for the original discrete spectra and its functional approximation. On the left, the original mean spectra for the simulation with a random background are illustrated. The black plot represents the mean spectrum of the normal class, while the colored plots refer to the mean spectra of the abnormal class corresponding to the specific shift values $\Delta\tilde{\upsilon}$. On the right of ESM Fig. S6, the mean spectra of the functional approximation of the corresponding discrete spectra are shown. When comparing the functional mean spectra with the discrete mean spectra, we can deduce a significant reduction of the noise and an improvement in the peak shape estimation.

In ESM Fig. S7, we showed for a specific peak shift position $\Delta\tilde{\upsilon} = 0.01$, one spectrum per class for each SNR case. On the left, the original spectra are illustrated. In each row, the normal and abnormal spectra are plotted for all SNR values. On the right, the functional approximations for each class are illustrated for the nine cases of SNR. In all these cases, we used 180 B-spline basis functions in the functional approximation. For a SNR which is larger than 5, the functional approximations perfectly represent the original spectra with almost noiseless construction. Although the functional approximation also fit noise for the lower SNR values due to a high number of basis used, a significant reduction of noise is observed in the approximation, and these functions preserved the peak shape in the case of lower SNR values.

The two approaches (PCA-LDA and FPCA-LDA) explained previously were applied on this simulated Raman spectra that contain a random background. For both methods, we used 10-fold cross-validation, and the number of components chosen was 50 components. The mean sensitivity was used as an evaluation metric, and the results are illustrated in a heat map in Fig. 4. In region c of Fig. 4, PCA-LDA and FPCA-LDA perform perfectly with 100% mean sensitivity, due to the clear distinction between the normal and the abnormal class. While in regions a and b of the same figure, in most cases, the FPCA-LDA performs better, due to the fact that in the functional approximation, less noise and an improvement of the shape of the peaks exist. However, in some cases, the

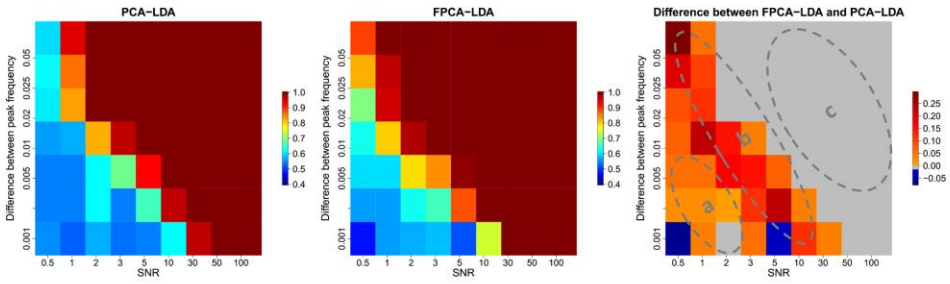


**Fig. 3** The mean sensitivities of applying PCA-LDA and FPCA-LDA on the simulated Raman data without background and the difference between both methods. The mean sensitivity of PCA-LDA, the mean sensitivity of FPCA-LDA, and the difference between the mean sensitivities of FPCA-LDA and PCA-LDA is illustrated on the left, middle, and right, respectively. In region c, both FPCA-LDA and PCA-LDA perform equally. While in regions a and b, the FPCA-LDA method performs better compared to PCA-LDA

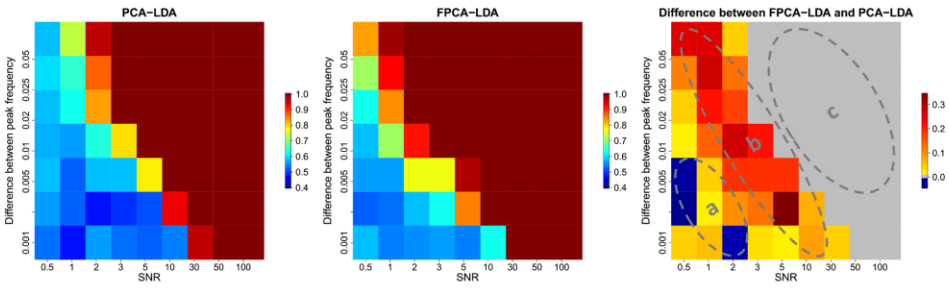


**Fig. 4** The mean sensitivity of applying PCA-LDA and FPCA-LDA on the simulated Raman data with background and the difference between both methods. The mean sensitivity of PCA-LDA, the mean sensitivity of FPCA-LDA, and the difference of the mean sensitivities are illustrated on the left, middle, and right, respectively. In region c, both FPCA-LDA and PCA-LDA perform equally. While in regions a and b, the FPCA-LDA method performs better compared to PCA-LDA

PCA-LDA provides a better mean sensitivity. This result might refer to the choice of the number of basis functions that should not be too large and too small. In order to test the significance of these findings, the non-parametric Kruskal-Wallis test was implemented. We tested the significance of the difference of the classification performance of both methods (FPCA-LDA and PCA-LDA). We performed this test for the simulation without and with background. All corresponding *p*-values are shown in ESM Fig. S10 where values less than 0.05 and values larger than 0.05 are highlighted in blue and in red, respectively.

In conclusion, we showed that FPCA-LDA and PCA-LDA perform equally in region c for both simulations (with and without background). However, FPCA-LDA performs better in both regions a and b of Fig. 3 and Fig. 4 for the simulation with and without background, respectively. The difference of performance is only significant in region b, which was determined by a Kruskal-Wallis test (ESM Fig. S10). This test showed that the performance difference of FPCA-LDA and PCA-LDA is only significant in region b for both simulations with and without background.

## Experimental data

The experimental data was published elsewhere [44]. Shortly, a Raman microscope with an excitation wavelength of 532 nm was used for the acquisition of the experimental spectra. The naphthalene-degrading soil bacteria *Rhodococcus opacus* DSM 8531 (*R. opacus*), *Novosphingobium aromaticivorans* DSM 12444 (*N. aromaticivorans*), and *Cupriavidus basilensis* DSM 9750 (*C. basilensis*) were included in the study and purchased from the Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures. Each Raman spectrum is a spectrum acquired from a single cell of around 75 to 100 single cells. Throughout the experiments, three batches were cultivated and measured. The aforementioned microorganisms are cultivated separately in water and in heavy water ($D_2O$). Through this fact, hydrogen atoms are exchanged by deuterium atoms and the C-H bond is exchanged by a C-D bond. The corresponding stretching vibration band (C-D stretching) is shifted in the Raman silent region (for more details on the experimental setup and preprocessing, we refer to Kumar et al. [44]). In the following, the two methods, PCA-LDA and FPCA-LDA, were applied on both raw and pre-processed experimental data. The total number of spectra used is 2262 and 1131 spectra for the raw experimental and pre-processed dataset, respectively. The preprocessing consists of a combination of two spectra for spike correction so the preprocessed dataset features less spectra than the raw dataset. We applied two approaches for two cross-validation, namely the leave one batch out cross-validation (LOBOCV) and the 10-fold cross-validation (10-fold CV).

The first step in the FPCA-LDA is to approximate our experimental Raman data into functions. Therefore, the B-spline basis functions are used for both the raw and pre-processed experimental Raman data. The number of basis functions implemented is equal to 200. The mean spectra for both raw and pre-processed Raman data and their functional mean spectra are shown in Fig. 5. The functional mean spectra (SNR = 278.25) for the raw experimental data represent well the original data (SNR = 272.38) with a slight reduction of noise as shown in the left panel of Fig. 5. In the right side of this figure, the functional mean spectra (SNR = 307.14) of the pre-processed experimental Raman data (SNR = 302) are illustrated. Like the raw functional mean, the pre-processed functional mean represents well the original data and a slight reduction of noise is observed. The elbow method suggested that the optimal number of basis functions is equal to 80 basis functions. However, we increased this number to 200 basis functions because we wanted the functional approximation to include the C-D/C-H region with high quality.

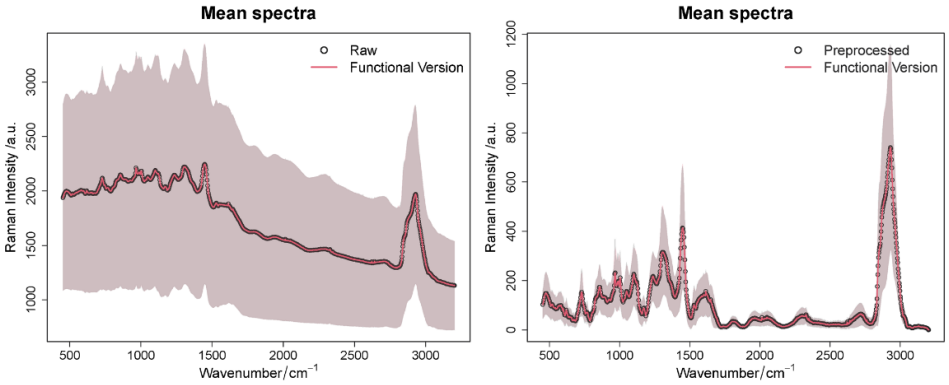


**Fig. 5** The discrete and the functional mean spectra for the raw and pre-processed Raman data. On the left, the raw Raman data is illustrated through their discrete mean spectra and their functional mean spectra in black and red, respectively. In the right panel, the pre-processed Raman data is also illustrated through their discrete mean spectrum and their functional mean spectrum in black and red, respectively

The functional data analysis was applied to the pre-processed experimental data for each of the classes mentioned above, and the discrete mean spectra and their functional version are shown in Fig. 6. A slight reduction of noise in the functional approximation is visible (see ESM Table S2), and this refers to the fact that a large number of basis functions are used and that the SNR of the pre-processed Raman data is high (approximately equal to SNR = 302). This SNR is approximated by calculating the ratio of the peak amplitude at the C-H band (2930 cm$^{-1}$) to the standard deviation of the region between 2408 and 2578 cm$^{-1}$. The functional approximation showed an improved SNR which is equal to 307.14.

We aim to compare both approaches that were explained in the workflow in Fig. 1. Therefore, we implemented the PCA-LDA method on the discrete pre-processed spectra and the FPCA-LDA method on the functional approximation of these spectra. In both methods, we tested the model using the two cross-validation schemes (LOBOCV and 10-fold CV), and the results are summarized in Fig. 7.

The comparison between the PCA-LDA and FPCA-LDA methods using the LOBOCV is illustrated in the first row of Fig. 7. The mean sensitivities are shown in panel a. Similar performance for both methods can be shown with a slight improvement in the values for the FPCA-LDA method, in

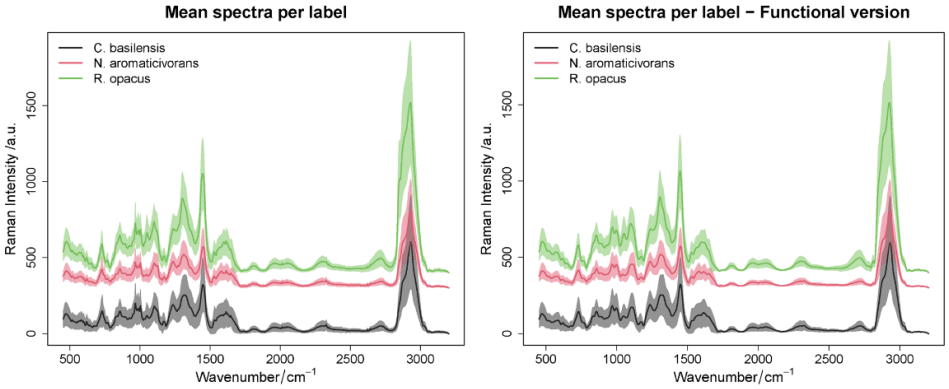


**Fig. 6** The mean spectra per class (*C. basilensis*, *N. aromaticivorans*, *R. opacus*) for the pre-processed experimental data (left) and their functional approximation (right). The discrete mean spectra per class are shown on the left. Their functional counter parts are illustrated on the right side. A reduction of noise in the functional version is deduced, where the mean intensity with the standard deviation for specific wavenumbers of both versions is illustrated in ESM Table S2. Beside this noise reduction, no Raman spectral features are removed

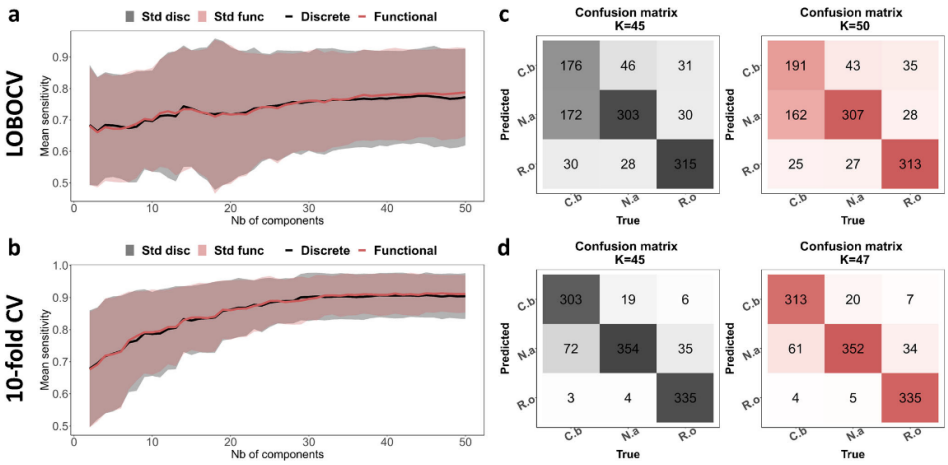


**Fig. 7** The mean sensitivities and the confusion matrices of PCA-LDA and FPCA-LDA methods using LOBOCV and 10-fold CV using the preprocessed Raman data. Panel **a** represents the mean sensitivities of PCA-LDA and FPCA-LDA methods using LOBOCV in black and red, respectively. The confusion matrices of the models with the highest mean sensitivity are illustrated in panel **c**. Panel **b** refers to the mean sensitivities of the PCA-LDA and the FPCA-LDA methods using 10-fold CV in black and red, respectively. The confusion matrices of the models with highest mean sensitivity are illustrated in panel **d**

particular in the region where the number of components is larger than 32. A reduction of the standard deviation is visible in the case of the FPCA-LDA method. Although the functional data analysis resulted in smoothed function, this did not affect the classification output. Therefore, the functional approach conserved the important features needed in the classification. The maximum mean sensitivity for the PCA-LDA method refers to the model with 45 components with a mean sensitivity of $0.77 \pm 0.16$. However, the maximum sensitivity for the FPCA-LDA method refers to a model with 50 components with a value of $0.79 \pm 0.14$. The corresponding confusion matrices for the model with the highest mean sensitivities are illustrated in panel c in black and red for PCA-LDA and FPCA-LDA, respectively. In the second row of Fig. 7, the mean sensitivities of PCA-LDA and FPCA-LDA using the 10-fold CV are shown. Similarly to panel a, both methods are performing analogously. A small reduction of the standard deviation is shown in the case of the FPCA-LDA. Even though functional data analysis resulted in smoothed function, this did not affect the classification output. Therefore, the functional approach conserved the important features needed for the classification. The maximum mean sensitivity for the PCA-LDA method refers to the model that includes 45 components with a value of $0.91 \pm 0.07$. However, the maximum sensitivity for the FPCA-LDA method refers to a model with 47 components with a value of $0.91 \pm 0.06$. The corresponding confusion matrices for the model with the highest mean sensitivities are illustrated in panel d for PCA-LDA and FPCA-LDA methods in black and red, respectively.

The functional data analysis was also applied to the raw experimental data, and the discrete mean spectra and their functional versions are shown in the ESM in Fig. S8. In these plots, a slight reduction of noise in the functional approximation is visible. This small noise reduction refers to the fact that a large number of basis functions are used and that our experimental Raman data contain less noise (SNR = 272.38). Also in the case of the raw data, we want to compare both approaches that were explained in the workflow in Fig. 1. Therefore, we implemented the PCA-LDA method on the discrete raw spectra and the FPCA-LDA method on the functional approximation of these spectra. In both methods, we tested the model with two cross-validation schemes (LOBOCV and 10-fold CV), and the results are summarized in the ESM in Fig. S9.

The comparison between the PCA-LDA and FPCA-LDA using the LOBOCV is illustrated in the first row of ESM Fig. S9. The mean sensitivities are shown in panel a. Similar performance can be shown for both methods with a slight improvement for the FPCA-LDA method. Particularly in the region where the number of components is larger than 40 and in the region where the number of components is around 20 components, the FPCA-LDA performs better.

Additionally, a reduction of the standard deviation is visible in the case of the FPCA-LDA. Although functional data analysis resulted in a smoothed function, this did not affect the classification output. Therefore, the functional approach conserved the important features needed in the classification. The maximum mean sensitivity for the PCA-LDA method was achieved by a model that includes 48 components and the mean sensitivity was $0.77 \pm 0.16$. However, the maximum sensitivity for the FPCA-LDA method was seen for a model with 44 components and the mean sensitivity was $0.78 \pm 0.15$. The corresponding confusion matrices for the model with the highest mean sensitivities are illustrated in panel c in black and red for PCA-LDA and FPCA-LDA methods, respectively. In the second row of ESM Fig. S9, the mean sensitivities of PCA-LDA and FPCA-LDA methods using the 10-fold CV are visible. Similar to panel a, both methods perform analogously with a slight improvement in the case where the FPCA-LDA method was applied, particularly in the region where the number of components is larger than 20 components. Furthermore, a reduction of the standard deviation is shown in the case of the FPCA-LDA method. Although functional data analysis resulted in smoothed function, this did not affect the classification output. Therefore, the functional approach conserved the important features needed for the classification. The maximum mean sensitivity for the PCA-LDA method refers to the model that includes 50 components with a value of $0.89 \pm 0.08$. However, the maximum sensitivity for the FPCA-LDA method could be seen for a model with 50 components ($0.9 \pm 0.06$). The corresponding confusion matrices for the models with highest mean sensitivities are illustrated in panel d in black and red for the PCA-LDA and the FPCA-LDA methods, respectively.

## Conclusion

In this manuscript, we tested the application of the functional data analysis on Raman spectral data. Therefore, we compared the functional approach, e.g., the functional principal component analysis followed by a linear discriminant analysis (FPCA-LDA) to the classical approach, the principal component analysis followed by linear discriminant analysis (PCA-LDA). Our study consists of testing both approaches on simulated and experimental Raman data. Within the simulated data, we investigated two scenarios: one with and one without background contribution. For both scenarios, we constructed 63 different simulations by changing the mean signal-to-noise ratio (SNR) and the shift in the peak position that was used to construct the abnormal class. A 10-fold cross-validation (10-fold CV) was used to evaluate the model performance in both scenarios. We could show that the functional approach (FPCA-LDA) performed better in region b of both Fig. 3 and Fig. 4, where the SNR and the shift in the peak position values are inversely proportional. However, both methods perform statistically similar in region a and region c of the same figures. In these both regions, either the quality of the spectra is low (low SNR) in combination with a small peak shift (region a) that both models do not work or the spectral quality is so good (region c) that both methods perform perfectly. These outcomes were similar for both scenarios (simulated Raman data with and without background). Then, we evaluated both approaches on experimental Raman spectra. Therefore, raw and pre-processed Raman data were used. Two cross-validation methods were implemented, the leave one batch out cross-validation (LOBOCV) and 10-fold CV. A slight improvement in the classification performance is shown when comparing the mean sensitivities between the PCA-LDA and the FPCA-LDA methods for the raw and the pre-processed experimental data. Moreover, these results were similar regarding the use of the cross-validation methods. This outcome of analyzing functional data on the experimental data is in accordance with the simulation data with or without background since the signal-to-noise ratio in the experimental data is high. However, the functional approach is sensitive to the choice of the number of basis functions, making this selection a very challenging task, which is an advantage of the classical approach. Furthermore, in terms of computation power, the functional version takes more time since an additional approximation step is needed before applying the functional PCA.

In conclusion, the functional data analysis can be considered a promising tool for analyzing Raman spectra, especially when the quality of the data is low. This property makes functional data analysis a great tool to analyze spectra which were acquired fast, or in vivo which both yield low-quality Raman spectra. It might be that the function data analysis is not so sensitive to a missing wavenumber calibration, because functions are utilized, and that functional data analysis can be used with spectra with a different spectral resolution. These points need further investigation and will be investigated in the future.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s00216-021-03360-1.

**Acknowledgements** This study is part of the Collaborative Research Centre AquaDiva of the Friedrich Schiller University Jena, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – SFB 1076 – Project Number 218627073. The funding of the DFG for the project BO 4700/4-1 and the Bundesministerium für Bildung und Forschung (BMBF) for the project LPI-BT1 (FKZ:13N15466) are highly acknowledged. The data was acquired by Vinay Kumar B.N. and data generation is acknowledged.

**Availability of data and material** Available

**Code availability** Available

**Funding** Open Access funding enabled and organized by Projekt DEAL. This study was funded by the Deutsche Forschungsgemeinschaft (CRC1076AquaDiva, BO 4700/4-1) and the BMBF (FKZ:13N15466, LPI-BT1).

## Declarations

**Conflict of interest** The authors declare no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## References

1. Ralbovsky NM, Lednev IK. Raman spectroscopy and chemometrics: a potential universal method for diagnosing cancer. Spectrochim Acta A Mol Biomol Spectrosc. 2019;219:463–87.
2. Bocklitz TW, Guo S, Ryabchykov O, Vogler N. Raman based molecular imaging and analytics: a magic bullet for biomedical applications!? Anal Chem. 2016;19.
3. Žukovskaja O, Ryabchykov O, Straßburger M, Heinekamp T, Brakhage AA, Hennings CJ, et al. Towards Raman spectroscopy of urine as screening tool. J Biophotonics. 2020;13:e201900143.
4. Matousek P, Stone N. Development of deep subsurface Raman spectroscopy for medical diagnosis and disease monitoring. Chem Soc Rev Royal Society of Chemistry. 2016;45:1794–802.
5. Lyng FM, Traynor D, Ramos IRM, Bonnier F, Byrne HJ. Raman spectroscopy for screening and diagnosis of cervical cancer. Anal Bioanal Chem. 2015;407:8279–89.
6. Gualerzi A, Niada S, Giannasi C, Picciolini S, Morasso C, Vanna R, et al. Raman spectroscopy uncovers biochemical tissue-related features of extracellular vesicles from mesenchymal stromal cells. Sci Rep. Nat Publ Group. 2017;7:9820.
7. Guo S, Ryabchykov O, Ali N, Houhou R, Bocklitz T. 4.13 - Comprehensive chemometrics. In: Brown S, Tauler R, Walczak B, editors. Compr Chemom Second Ed [Internet]. Oxford: Elsevier; 2020 [cited 2021 Mar 3]. p. 333–59. Available from: https://www.sciencedirect.com/science/article/pii/B9780124095472146001
8. Ryabchykov O, Guo S, Bocklitz T. 4. Analyzing Raman spectroscopic data [internet]. Micro-Raman Spectrosc. De Gruyter; 2020 [cited 2021 Mar 4]. p. 81–106. Available from: https://doi.org/10.1515/9783110515312-004/html.
9. Virtanen T, Reinikainen S-P, Kögler M, Mänttäri M, Viitala T, Kallioinen M. Real-time fouling monitoring with Raman spectroscopy. J Membr Sci. 2017;525:312–9.
10. Maslova OA, Guimbretière G, Ammar MR, Desgranges L, Jégou C, Canizarès A, et al. Raman imaging and principal component analysis-based data processing on uranium oxide ceramics. Mater Charact. 2017;129:260–9.
11. Guo S, Rösch P, Popp J, Bocklitz T. Modified PCA and PLS: towards a better classification in Raman spectroscopy–based biological applications. J Chemom 2020;34:e3202.
12. Vanna R, Morasso C, Marcinnò B, Piccotti F, Torti E, Altamura D, et al. Raman spectroscopy reveals that biochemical composition of breast microcalcifications correlates with Histopathologic features. Cancer Res American Association for Cancer Research. 2020;80:1762–72.
13. Ichimura T, Chiu L, Fujita K, Machiyama H, Yamaguchi T, Watanabe TM, et al. Non-label immune cell state prediction using Raman spectroscopy. Sci Rep Nature Publishing Group. 2016;6:37562.
14. Vogler N, Bocklitz T, Salah FS, Schmidt C, Bräuer R, Cui T, et al. Systematic evaluation of the biological variance within the Raman based colorectal tissue diagnostics. J Biophotonics. 2016;9:533–41.
15. Hunter R, Anis H. Genetic support vector machines as powerful tools for the analysis of biomedical Raman spectra. J Raman Spectrosc. 2018;49:1435–44.
16. Zheng X, Lv G, Zhang Y, Lv X, Gao Z, Tang J, et al. Rapid and non-invasive screening of high renin hypertension using Raman spectroscopy and different classification algorithms. Spectrochim Acta A Mol Biomol Spectrosc. 2019;215:244–8.
17. Kampe B, Kloß S, Bocklitz T, Rösch P, Popp J. Recursive feature elimination in Raman spectra with support vector machines. Front Optoelectron. 2017;10:273–9.
18. Jermyn M, Desroches J, Mercier J, Tremblay M-A, St-Arnaud K, Guiot M-C, et al. Neural networks improve brain cancer detection with Raman spectroscopy in the presence of operating room light artifacts. J Biomed Opt International Society for Optics and Photonics. 2016;21:094002.
19. Ramsay JO. Functional data analysis. Encycl Stat Sci [Internet]. American Cancer Society; 2006 [cited 2021 Feb 22]. Available from: https://doi.org/10.1002/0471667196.ess3138.
20. Ramsay JO, Dalzell CJ. Some tools for functional data analysis. J R Stat Soc Ser B Methodol. 1991;53:539–61.
21. Cuevas A. A partial overview of the theory of statistics with functional data. J Stat Plan Inference. 2014;147:1–23.
22. Ullah S, Finch CF. Applications of functional data analysis: a systematic review. BMC Med Res Methodol. 2013;13:43.
23. Wang J-L, Chiou J-M, Müller H-G. Functional data analysis. Annu Rev Stat Its Appl. 2016;3:257–95.
24. Muller H-G. International handbook (Encyclopedia) of statistical sciences. :5.
25. Dauxois J, Pousse A, Romain Y. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. J Multivar Anal. 1982;12:136–54.
26. James GM, Hastie TJ. Functional linear discriminant analysis for irregularly sampled curves. J R Stat Soc Ser B Stat Methodol. Wiley Online Library; 2001;63:533–50.
27. Mas A, Pumo B. Functional linear regression with derivatives. J Nonparametric Stat Taylor & Francis. 2009;21:19–40.
28. Ratcliffe SJ, Heller GZ, Leader LR. Functional data analysis with application to periodically stimulated foetal heart rate data. II: Functional logistic regression. Stat Med. 2002;21:1115–27.
29. Ratcliffe SJ, Leader LR, Heller GZ. Functional data analysis with application to periodically stimulated foetal heart rate data. I: Functional regression. Stat Med. 2002;21:1103–14.
30. Baladandayuthapani V, Mallick BK, Hong MY, Lupton JR, Turner ND, Carroll RJ. Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. Biometrics. 2008;64:64–73.
31. Ramsay JO, Ramsey JB. Functional data analysis of the dynamics of the monthly index of nondurable goods production. J Econ. 2002;107:327–44.

32. Kneip A, Utikal KJ. Inference for density families using functional principal component analysis. J Am Stat Assoc. Taylor & Francis. 2001;96:519–42.
33. Ogden RT, Miller CE, Takezawa K, Ninomiya S. Functional regression in crop lodging assessment with digital images. J Agric Biol Environ Stat. 2002;7:389.
34. Ramsay JO, Munhall KG, Gracco VL, Ostry DJ. Functional data analyses of lip motion. J Acoust Soc Am Acoustical Society of America. 1996;99:3718–27.
35. Lucero JC. Comparison of measures of variability of speech movement trajectories using synthetic records. J Speech Lang Hear Res American Speech-Language-Hearing Association. 2005;48:336–44.
36. Koenig LL, Lucero JC, Perlman E. Speech production variability in fricatives of children and adults: results of functional data analysis. J Acoust Soc Am Acoustical Society of America. 2008;124:3158–70.
37. Zhang Y, Müller H-G, Carey JR, Papadopoulos NT. Behavioral trajectories as predictors in event history analysis: male calling behavior forecasts medfly longevity. Mech Ageing Dev. 2006;127:680–6.
38. Harezlak J, Wu MC, Wang M, Schwartzman A, Christiani DC, Lin X. Biomarker discovery for arsenic exposure using functional data. Analysis and feature learning of mass spectrometry proteomic data. J Proteome Res Am Chem Soc. 2008;7:217–24.
39. Reiss PT, Ogden RT. Functional principal component regression and functional partial least squares. J Am Stat Assoc Taylor & Francis; 2007;102:984–996.
40. Tools for exploring functional data. Funct Data Anal [Internet]. New York, NY: Springer New York; 2005. p. 19–35. Available from: https://doi.org/10.1007/0-387-22751-2_2
41. From functional data to smooth functions. Funct Data Anal [Internet]. New York, NY: Springer New York; 2005. p. 37–58. Available from: https://doi.org/10.1007/0-387-22751-2_3
42. Smoothing functional data by least squares. Funct Data Anal [Internet]. New York, NY: Springer New York; 2005. p. 59–79. Available from: https://doi.org/10.1007/0-387-22751-2_4
43. Principal components analysis for functional data. Funct Data Anal [Internet]. New York, NY: Springer New York; 2005. p. 147–72. Available from: https://doi.org/10.1007/0-387-22751-2_8
44. Kumar BNV, Guo S, Bocklitz T, Rösch P, Popp J. Demonstration of carbon catabolite repression in naphthalene degrading soil bacteria via Raman spectroscopy based stable isotope probing. Anal Chem Am Chem Soc. 2016;88:7574–82.


**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Supplementary Information**

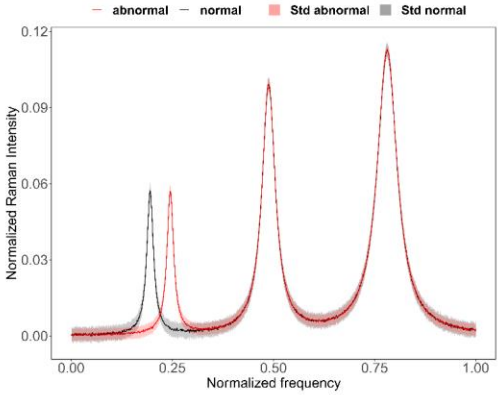


**Fig. S1** The mean spectra for the simulated Raman without background per class in the case of $\Delta\bar{\nu} = 0.05$ and $SNR = 30$. The mean spectra and its standard deviation for the normal simulated Raman are shown in black. While the mean spectra for the abnormal simulated Raman and its standard deviation are shown in red.

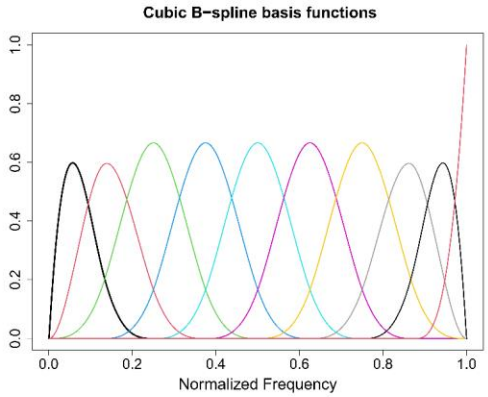


**Fig. S2** Cubic B-spline basis functions used in the FDA approximation. An illustration of the first ten basis functions is shown.

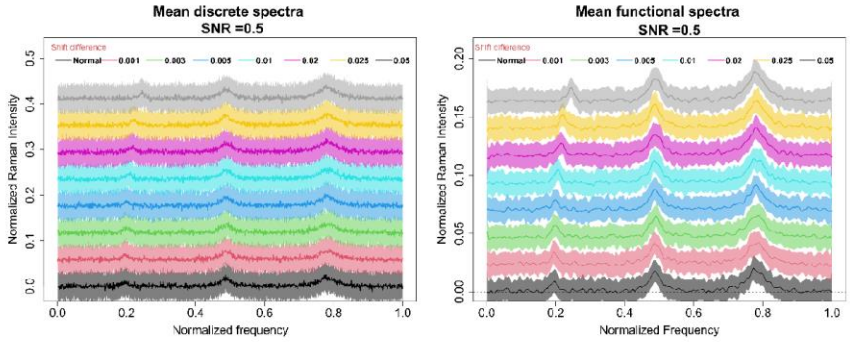


**Fig. S3** The discrete and the functional mean spectra per each shift for $SNR = 0.5$ for the simulated Raman without background. The left plot represents the discrete simulated Raman mean spectra, where the black plot represents the normal class, and the colored spectra represent the abnormal class for each $\Delta\tilde{v}$. On the right plot, the functional approximation using 190 basis functions is illustrated. Reduction of noise is shown in addition to the improvement in the shape of the peaks.

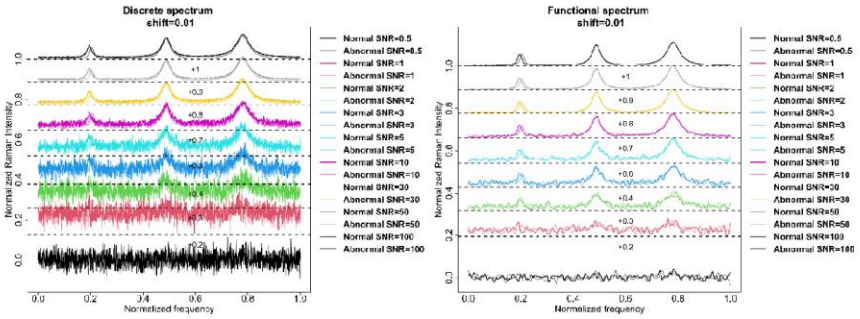


**Fig. S4** The discrete and the functional spectra per each SNR for $\Delta\tilde{v} = 0.01$ within the simulated Raman without background. The left plot represents on each row a simulated Raman spectrum for the normal and abnormal class. While the right plot represents the functional approximation of these spectra using 190 basis functions. Reduction of noise is shown in the right panel.

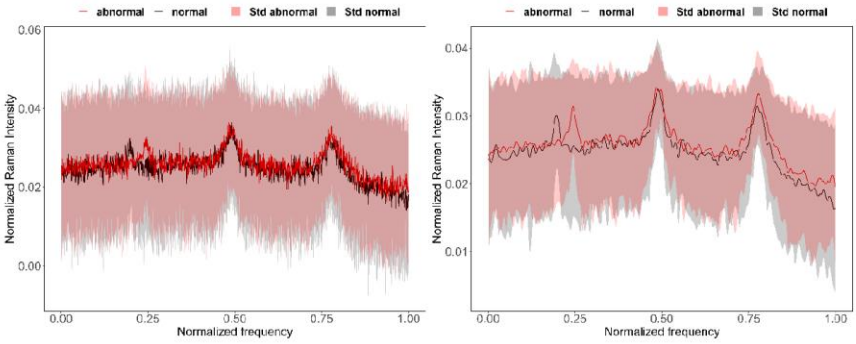


**Fig. S5** The mean spectra per class for the simulation with background ($\Delta\tilde{v} = 0.05$ and $SNR = 0.5$). The left plot represents the discrete simulated Raman mean spectra per class. While the right plot represents the functional version using 180 basis functions of these simulated Raman spectra.

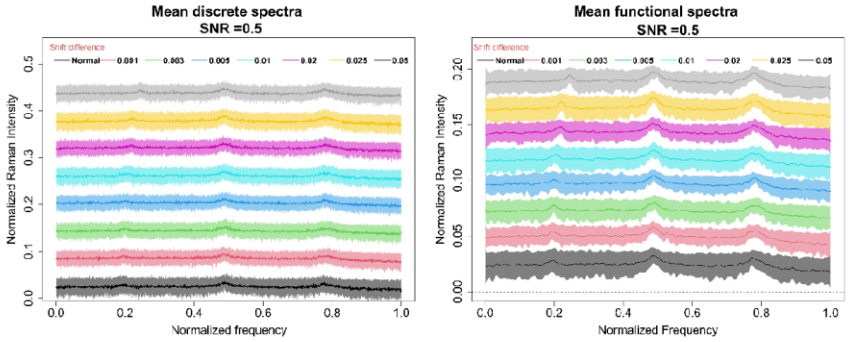


**Fig. S6** The discrete and the functional mean spectra per shift for $SNR = 0.5$ in the scenario with background. The left plot represents the discrete simulated Raman mean spectra. The black spectrum represents the normal class, while the colored spectra represent the abnormal class specified by $\Delta\tilde{\nu}$. While the right plot represents the functional approximation using 180 basis functions. Reduction of noise is shown in addition to the improvement in the peaks shape.

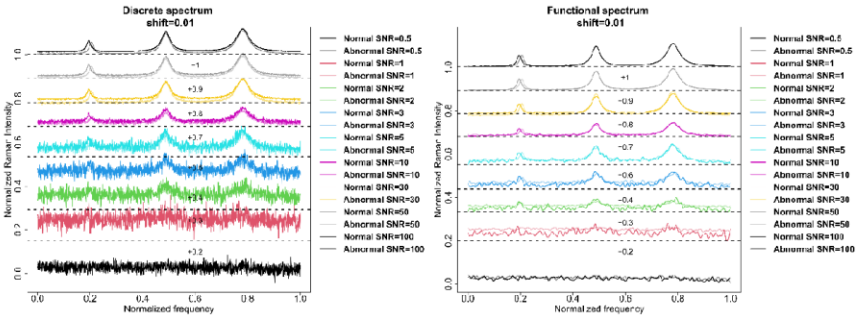


**Fig. S7** The discrete and the functional simulated Raman spectra per SNR for $\Delta\tilde{\nu} = 0.01$ with background. The left plot represents in each row a simulated Raman spectrum for the normal and abnormal class. While the right plot represents the functional approximation of these spectra using 180 basis functions. Reduction of noise is shown in the right panel.

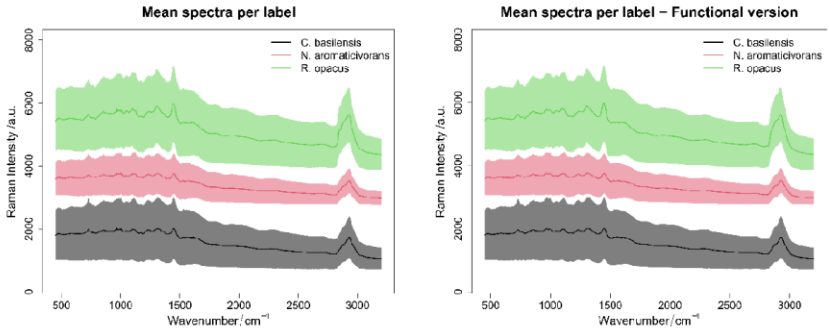


**Fig. S8** The mean spectra per label for the raw experimental data and their functional approximation. The mean discrete spectra per each label are shown on the left, while their functional mean spectra are illustrated on the right.

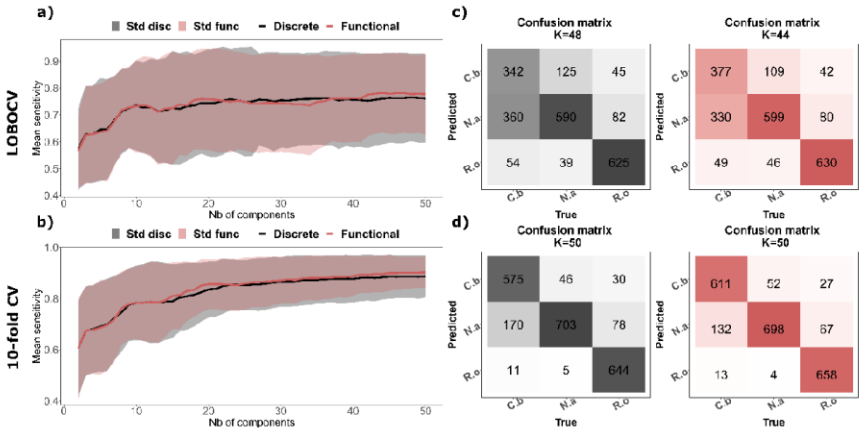


**Fig. S9** The mean sensitivity and the confusion matrices of PCA-LDA and FPCA-LDA methods using LOBOCV and 10-fold CV on the raw Raman data. Panel a) represents the mean sensitivities of the PCA-LDA and the FPCA-LDA methods using LOBOCV in black and red, respectively. The confusion matrices of these models for the corresponding highest mean sensitivity are illustrated in c). Panel b) refers to the mean sensitivities of PCA-LDA and FPCA-LDA methods using 10-fold CV in black and red, respectively. The confusion matrices of the models for the corresponding highest mean sensitivity are illustrated in panel d).

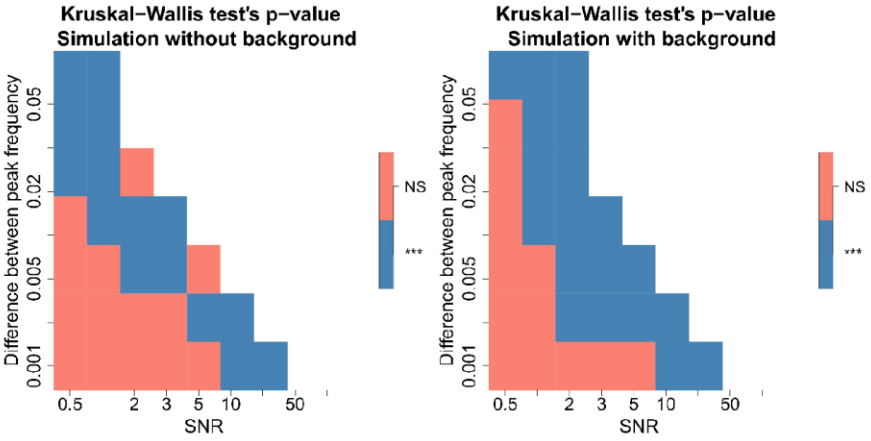


**Fig. S10** The p-values of the Kruskal-Wallis test for both simulation without and with background. The significance difference between PCA-LDA and FPCA-LDA classification is illustrated in blue while non significance difference is highlighted in red.

**Table S1** The parameter details of the simulated Raman spectra with and without background

| Parameters | Values |
|---|---|
| SNR | {0.5, 1, 2,3,5,10,30,50,100} |
| Labels | abnormal, normal |
| Shift peaks position $\Delta\tilde{\nu}$ used to construct the abnormal label | {0.001, 0.003, 0.005, 0.01, 0.02, 0.025, 0.05} |
| Number of pixels | 1024 |
| Peaks details: number, width, amplitude, position | 3, {0.01,0.02,0.03}, {0.005,0.0175,0.03}, {200,500,800} |
| Basis functions | Cubic B-spline basis of order 4 |
| $K$ | 190 for the simulated Raman spectra without background<br>180 for the simulated Raman spectra with background |

**Table S2** Mean intensity and standard deviation of specific wavenumber for the Pre-processed Raman spectra and its functional version displayed per microorganisms

| Wavenumber / $cm^{-1}$ | Mean intensity ± standard deviation | | | | | |
|---|---|---|---|---|---|---|
| | *R. opacus* | | *N. aromaticivorans* | | *C. basilensis* | |
| | Pre-processed Raman data | Functional version | Pre-processed Raman data | Functional version | Pre-processed Raman data | Functional version |
| 954 | 131.92 ± 81.84 | 131.44 ± 80.2 | 113.89 ± 62.84 | 114.28 ± 61.24 | 212.74 ± 83.42 | 216.17 ± 82.26 |
| 1402 | 122.58 ± 86.43 | 121.72 ± 85.33 | 99.24 ± 48.08 | 97.97 ± 45.84 | 204.92 ± 83.65 | 207.33 ± 82.02 |
| 2933 | 603.64 ± 307.98 | 597.1 ± 303.76 | 501.82 ± 214.03 | 495.06 ± 208.42 | 1104.44 ± 400.38 | 1080.6 ± 390.26 |

# 8 PEER-REVIEWED PUBLICATIONS

I. Martin Taubert, Will A. Overholt, Beatrix M. Heinze, Georgette Azemtsop Matanfack, Rola Houhou, Nico Jehmlich, Martin von Bergen, Petra Rösch, Jürgen Popp and Kirsten Küsel, Bolstering fitness via CO2 fixation and organic carbon uptake: mixotrophs in modern groundwater, ISME J, 1-10, 2021.

II. Martin Taubert, Will A. Overholt, Beatrix M. Heinze, Georgette Azemtsop Matanfack, Rola Houhou, Nico Jehmlich, Martin von Bergen, Petra Rösch, Jürgen Popp, Kirsten Küsel, Bolstering fitness via opportunistic CO2 fixation: mixotroph dominance in modern groundwater, bioRxiv, 2021.

III. Georgette Azemtsop Matanfack, Martin Taubert, Shuxia Guo, Rola Houhou, Thomas Bocklitz, Kirsten Küsel, Petra Rösch*, and Jürgen Popp, Influence of Carbon Sources on Quantification of Deuterium Incorporation in Heterotrophic Bacteria: A Raman-Stable Isotope Labeling Approach, Anal. Chem., 92, 16, 11429-11437, 2020.

IV. Shuxia Guo, Oleg Ryabchykov, Nairveen Ali, Rola Houhou, Thomas Bocklitz, Comprehensive Chemometrics, Elsevier, 333-59, 2020.

# 9 LIST OF CONFERENCES

- Rola Houhou, Jürgen Popp, and Thomas Bocklitz, Functional data analysis for Raman spectra, International Conference on Raman Spectroscopy, ICORS 2018, Jeju Island, South Korea, 26-31/08/2018, Poster

- Rola Houhou, Tobias Meyer, Parijat Barman, Michael Schmitt, Jürgen Popp, and Thomas Bocklitz, Phase retrieval methods for CARS spectra, Bunsentagung 2019, Jena, Germany, 30/05/-01/06/2019, Poster

- Rola Houhou, Jürgen Popp, and Thomas Bocklitz, PCA – LDA in functional and discrete framework applied to Raman spectra, 16th Scandinavian Symposium on Chemometrics, SSC 16, Oslo, Norway, 17-20/06/2019, Poster

# 10 List of Workshops

| **Speech and Vocal Training** | **Graduate Academy, Universität Jena** | **19-20/01/2018** |
|---|---|---|
| Good Scientific Practice and protecting Scientific Integrity | Graduate Academy, Universität Jena | 11-12/06/2018 |
| Data management with BExIS | iRTG AquaDiva | 18-19/06/2018 |
| Scientific Image Processing and Analysis | Graduate Academy, Universität Jena | 24-25/07/2018 |
| Effective Presentation with Science Slams | iRTG AquaDiva | 21-22/02/2019 |
| Fluid Flow and Reactive Transport in Natural Porous Media, Course I & II | iRTG AquaDiva | 03- 04/09/2019<br>&<br>22-23/10/2019 |
| Scientific Writing for Doctoral Researchers | iRTG AquaDiva | 18/09/2019 |
| Managing your PhD | iRTG AquaDiva | 19/09/2019 |

# 11 ACKNOWLEDGMENT

I would like to express my gratitude and warmest thanks to Prof. Dr. Jürgen Popp for allowing me to work in his group. My deep and sincere appreciation goes to my supervisor, PD Dr. Thomas Bocklitz, who made this work possible. His continuous support, advice, and guidance carried me through all the stages of my research time and writing of this thesis.

I want to dedicate a special thank you to the AquaDiva team, particularly to Dr. Anke Hädrich and Dr. Maria Fabisch. The team spirit and level of professionalism you characterized make the AquaDiva special. In addition, thanks to all AquaDiva Ph.D. students for our interdisciplinary discussions, the various courses, and the fun time we had together.

I would like to thank the following people for helping with the research: Prof. Dr. Michael Schmitt, Dr. Petra Rösch, Dr. Tobias Meyer-Zedler, Parijat Barman, Elsie Quansah, and Prof. Dr. med. Orlando Guntinas-Lichius.

I would like to also thank my colleagues in the AGBocklitz team (alumni and new) for all the enjoyable time. Additionally, thanks to the DFG and the BMFG for the financial support.

Special thanks go to my best friend and husband Mohamad; all our fruitful discussions, your tips, support, and understanding are highly appreciated. Thanks to my friends and family.

Last but not least, eternal gratitude for my dad.

# 12 APPENDIX

## 12.1 GS algorithm

The GS algorithm attempts to recover the phase using two intensity measurements. It is an error-reduction algorithm that iteratively calculates the error until it converges. The procedure is summarized as follows:

a. An initial phase $\varphi_0$ is used by generating randomly uniform numbers between $-\pi$ and $\pi$.
b. While the error is greater than $\xi$, the following steps are repeated:
   At iteration $k$,
   i. the initial field in the object plane is calculated as

$$\boldsymbol{a_k = \sqrt{I_{obj}}\exp(i\varphi_{k-1})} \quad \textbf{(3)}$$

   ii. The phase distribution in the target plane $\phi_k$ is then calculated via the fast Fourier transform ($FFT$)

$$\boldsymbol{\phi_k = arg\big(FFT(a_k)\big)} \quad \textbf{(4)}$$

   iii. Next, the phase distribution in the target plane with the target intensity $\sqrt{I_{target}}$ is combined as shown in equation 5

$$\boldsymbol{A_k = \sqrt{I_{target}}\exp(i\phi_k)} \quad \textbf{(5)}$$

   iv. And finally, the phase in the object plane $\varphi_k$is recovered as follows

$$\boldsymbol{\varphi_k = arg\big(FFT(A_k)\big)} \quad \textbf{(6)}$$

## 12.2 MEM method

MEM is a probability-based method, where the power spectrum $S(\nu)$ can be approximated on a defined normalized frequency $\nu$ range, described as

$$\boldsymbol{S(\nu) \approx \frac{b_0}{\left|1 + \sum_{k=1}^{M} b_k \exp\left(-i2\pi\nu\right)\right|^2}} \tag{7}$$

where $b_0, b_k$ for $1 \leq k \leq M$, and $M$ represent the coefficients and the number of poles of the approximation, respectively. Since the power spectrum is equal to the Fourier transform of the autocorrelation function $R_l$, we can write the following approximation

$$\boldsymbol{\frac{b_0}{\left|1 + \sum_{k=1}^{M} b_k \exp\left(-i2\pi\nu\right)\right|^2} \approx \sum_{l=-M}^{M} R_l \exp(-2i\pi\nu l)} \tag{8}$$

where $R_l = \int_0^{1/\Delta t} S(\nu) \exp(i2\pi\nu l)$. The coefficients $b_0$ and $b_k$ are estimated by solving the following Toeplitz matrix:

$$\begin{pmatrix} R_0 & R_{-1} & \cdots & R_{-M} \\ R_1 & R_0 & \cdots & R_{1-M} \\ \vdots & \vdots & \ddots & \vdots \\ M & R_{M-1} & \cdots & R_0 \end{pmatrix} \begin{pmatrix} 1 \\ b_1 \\ \vdots \\ b_M \end{pmatrix} = \begin{pmatrix} b_0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \tag{9}$$

Afterward, the problem can be solved at a specific frequency position $\nu_0$ where two physically valid restrictions are available. Consequently, the phase is retrieved, and the real and imaginary components are calculated.

## 12.3 KK method

Kramers – Kronig relation (KK) is a mathematical relationship between the real and the imaginary part of a complex function based on the Cauchy residues theorem. It is defined as follows:

$$\begin{aligned} \mathbf{Re}\left(\chi(\mathrm{i}\nu)\right) &= \frac{-1}{\pi} \wp \int_{-\infty}^{\infty} \frac{\mathbf{Im}\left(\chi(\mathrm{i}x)\right)}{x - \nu} dx \\ \mathbf{Im}\left(\chi(\mathrm{i}\nu)\right) &= \frac{1}{\pi} \wp \int_{-\infty}^{\infty} \frac{\mathbf{Re}\left(\chi(\mathrm{i}x)\right)}{x - \nu} \mathrm{d}x, \end{aligned} \tag{10}$$

where $\wp$ is the Cauchy principal value. By taking the logarithm on both sides and them using the KK relation, we can deduce the phase from the squared modulus as follows

$$\boldsymbol{\phi(\nu) = -\frac{1}{\pi} \wp \int \frac{\ln\left(\sqrt{S(\nu')}\right)}{\nu' - \nu} \mathrm{d}\nu'.} \tag{11}$$

Hence, the phase is then extracted by applying a discrete Hilbert transform on $\ln\left(\sqrt{S(\nu)}\right)$.

## 12.4 FPCA method

Functional principal component analysis (FPCA) is a dimension reduction tool for functional data, and it is considered the most popular method in FDA. Similar to the classical principal component analysis (PCA), the calculation of the principal components is done by examining the variance-covariance matrix/function.

However, in FPCA, the variable values are functions $X_i(t)$, and the equivalent notation of the weight vector $\beta$ and the variable $x_i$ in FPCA are the functions $\beta(t)$ and $X_i(t)$. Therefore, the principal component scores corresponding to the weight function is illustrated as follows

$$\boldsymbol{f_i = \int \beta X_i = \int \beta(t) X(t)\, dt} \qquad \textbf{(12)}$$

FPCA started by finding the weight function $\beta_1(s)$ that maximizes $N^{-1}\sum_i f_i = N^{-1}\sum_i(\int \beta_1 X_i)^2$ and subject to the unit sum of squares constraint $\int \beta_1(s)^2 = 1$. Then, at the $m$ step, the weight function $\beta_m$ is calculated in such a way that satisfies the orthogonality constraints $\int \beta_q \beta_m = 0, q < m$. Generally, in most PCA applications, finding the principal components is equivalent to finding the eigenvalues and eigenfunctions of the covariance function. Therefore, the covariance function $v(t,s)$ is defined as:

$$\boldsymbol{v(t,s) = N^{-1}\sum_{i=1}^{N} X_i(t) X_i(s)} \qquad \textbf{(13)}$$

And each eigenfunction $\beta_j(t)$ for an appropriate eigenvalue $\rho$ satisfies

$$\boldsymbol{\int v(t,s)\beta(s)ds = \rho\beta(t).} \qquad \textbf{(14)}$$

The left side of this equation is an integral transform $V$ of the weight function $\beta$ that can be defined by equation 15, and it is called the covariance operator $V$. Therefore, we may also express the eigenequation directly as equation 16, where $\beta$ is an eigenfunction rather than an eigenvector.

$$\boldsymbol{V\beta = \int v(.,t)\beta(t)dt} \qquad \textbf{(15)}$$

$$\boldsymbol{V\beta = \rho\beta.} \qquad (16)$$

In classical PCA, the number of variables is equal to $L$. In contrast, in the case of functional PCA (FPCA), the number of variables is infinity which refers to the number of function values.

## 12.5 CARS simulation

Theoretically, the CARS intensity $I_{CARS}$ is directly proportional to the squared modulus of the nonlinear susceptibility $\left|\chi^{(3)}\right|^2$ [94]. The nonlinear susceptibility $\chi^{(3)}$ is the sum of a non-resonant part (NRB) $\chi_{nr}^{(3)}$ that appears due to the electronic contributions and a Raman resonant part $\chi_r^{(3)}$ as follows $\chi^{(3)} = \chi_{nr}^{(3)} + \chi_r^{(3)}$. The non-resonant part is purely real and shows only a weak frequency dependency, while the resonant part is a complex function that can be described as Lorentz function:

$$\boldsymbol{\chi_r^{(3)}} = \sum_r \frac{\boldsymbol{A_r}}{\boldsymbol{\Omega_r - (\omega_{pu} - \omega_S) - i\gamma_r}}, \qquad \textbf{(17)}$$

where $(\omega_{pu} - \omega_S)$ is the difference of pump and Stokes frequency and $A_r, \Omega_r$, and $\gamma_r$ are the amplitude, the vibrational frequency, and the bandwidth of the $r^{\text{th}}$ Raman mode, respectively.

## 12.6 Raman simulation

Our study started primarily by simulating Raman data with different signal-to-noise ratio (SNR) cases: 0.5, 1, 2, 3, 5, 10, 30, 50, and 100. Afterward, the functional data analysis was tested for classification on these simulated Raman data and then on experimental Raman data. Therefore, the simulated Raman spectra are divided into two classes (normal and abnormal groups) with three peaks. The abnormal group was generated by slightly shifting one of the peaks from the peak position used in the normal group by one of the following values ($\Delta\tilde{\nu}$): 0.001, 0.003, 0.005, 0.01, 0.02, 0.025, and 0.05. This situation often occurs in biomedical Raman spectroscopy when, for example, the protein's secondary structure changes between two groups or if an isotope labeling is applied. The number of pixels in each spectrum is equal to 1024. The total number of spectra in each class was set to 100 spectra. Sixty-three cases of simulation data are built, where each dataset includes 200 spectra.

## 12.7 Experimental Raman data

Three microorganisms were included in the analysis; the naphthalene-degrading soil bacteria *R. opacus*, *N. aromaticivorans*, and *C.basilensis*. Throughout the experiments, three batches were cultivated and measured. The three microorganisms are cultivated separately in water and heavy water ($D_2O$). Through this fact, hydrogen atoms are exchanged by deuterium atoms, and a C-D bond exchanges the C-H bond.

## 12.8 FDA approximation

The functional data analysis (FDA) assumes the existence of some functions $X_i(t)$ in Hilbert space, e.g., $L^2(I)$ on a compact interval $I$, giving rise to the observed data $X \in \mathbb{R}^{N\times L}$. Therefore, each function is treated as one entity rather than a sequence of individual measured variables [95], [96]. Analytically, these functions cannot be calculated; instead, a set of basis functions $\psi_j, j = 1,2,\dots,J$ are used to approximate them by using the following equation:

$$\boldsymbol{X_i(t) = \sum_{j=1}^{J} c_j \psi_j(t),} \qquad \textbf{(18)}$$

where $c_j, j = 1,2,\dots,J$ represents the coefficients. The choice of the basis functions is various, but B-spline basis functions are the most common choice for spectral analysis. In brief, the B-spline basis function is a piecewise polynomial function defined on a specific interval $I$ with an order $O$ and a knot vector. Furthermore, several methods can be implemented to calculate the coefficients $c_j$; however, the least-squares estimation is utilized.

## 12.9 Parameters of GS algorithm, DnCNN, and incSRCNN networks

A Gaussian approximation is used in the GS algorithm since the source beam values are unavailable. The number of iterations that the algorithm carries on is 50000, and the code was built using Matlab 2020b (The MathWorks, Natick, MA).

The DnCNN [97] was also implemented in Matlab 2020b (The MathWorks, Natick, MA).

The proposed network (incSRCNN) includes a simple architecture consisting of three layers. The input image is convolved in the first layer with three different kernel sizes 3, 5, and 9 into 192 feature maps. The second layer then applies a $1 \times 1$ kernel to condense to 64 feature maps. Finally, the third layer uses a $3 \times 3$ kernel to construct the output image. The training of the network was performed by minimizing the mean absolute error (MAE)-based loss between the HQ images and the output of the incSRCNN network. The Adam algorithm was used for the optimization with a learning rate of $3e^{-4}$. A total of 1008 and 288 coupled HQ and LQ images were used for the training and the validation, respectively. All computations were done using Google Colab. The total number of parameters to be trained is 20,481. The training of incSRCNN includes only the CARS modality, and the training time of this network is approximately 10 minutes.

# 13 ERKLÄRUNGEN

**Selbständigkeitserklärung**

Ich erkläre, dass ich die vorliegende Arbeit selbständig und unter Verwendung der angegebenen Hilfsmittel, persönlichen Mitteilungen und Quellen angefertigt habe.

Rola Houhou

(Datum) (Ort) (Unterschrift)

**Erklärung zu den Eigenanteilen 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 der weiteren Doktoranden/ Doktorandinnen als Koautoren an den Publikationen und Zweitpublikationsrechten bei einer kumulativen Dissertation informiert und stimmen dem zu.

Die Anteile der Promovendin sowie der weiteren Doktoranden/Doktorandinnen als Koautoren an den Publikationen und Zweitpublikationsrechten bei einer kumulativen Dissertation sind in der Anlage aufgeführt.

Rola Houhou

(Datum) (Ort) (Unterschrift)

Ich bin mit der Abfassung der Dissertation als publikationsbasierte, d. h. kumulative, einverstanden und bestätige die vorstehenden Angaben.

PD Dr. Thomas Bockltiz

(Datum) (Ort) (Unterschrift)

Prof. Dr. Jürgen Popp

(Datum) (Ort) (Unterschrift)