

# **Refinement of Raman spectra from extreme background and noise interferences – Cancer diagnostics using Raman spectroscopy**

Von der Fakultät für Maschinenbau, Verfahrens- und Energietechnik  
der Technischen Universität Bergakademie Freiberg

genehmigte

## **DISSERTATION**

zur Erlangung des akademischen Grades

Doktor-Ingenieur

(Dr.-Ing.)

vorgelegt

von Medhanie Tesfay Gebrekidan, M.Sc.

geboren am 03.08.1986 in Adigrat, Äthiopien

Gutachter:

Prof. Dr.-Ing. habil Andreas Bräuer

Prof. Dr.-Ing. Hans-Jürgen Koß

Tag der Verleihung: 17.01.2022

## **Versicherung**

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskriptes habe ich Unterstützungsleistungen von folgenden Personen erhalten:

Prof. Dr.-Ing. habil. Andreas Siegfried Bräuer

Weitere Personen waren an der Abfassung der vorliegenden Arbeit nicht beteiligt. Die Hilfe eines Promotionsberaters habe ich nicht in Anspruch genommen. Weitere Personen haben von mir keine geldwerten Leistungen für Arbeiten erhalten, die nicht als solche kenntlich gemacht worden sind. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

## Abstract

Raman spectroscopy is an optical measurement technique able to provide spectroscopic information that is molecule-specific and unique to the nature of the specimen under investigation. It is an invaluable analytical tool that finds application in several fields such as medicine and *in situ* chemical processing. Due to its high specificity and label-free features, Raman spectroscopy greatly impacted cancer diagnostics. However, retrieving and interpreting the Raman spectrum that contains the molecular information is challenging because of extreme background interference.

I have developed various spectra-processing approaches required to purify Raman spectra from noisy and heavily background interfered raw Raman spectra. In detail, these are a new noise reduction method based on vector casting and new deep neural networks for the efficient removal of noise and background. Several neural network models were trained on simulated spectra and then tested with experimental spectra. The here proposed approaches were compared with the state-of-the-art techniques via different signal-to-noise ratios, standard deviation, and the structural similarity index metric. The methods presented here perform well and are superior in comparison to what has been reported before, especially at small signal-to-noise ratios, and for extreme fluorescence interfered raw Raman spectra. Furthermore, the deep neural network-based methods do not rely on any human intervention.

The motivation behind this study is to make Raman spectroscopy, especially the shifted-excitation Raman difference spectroscopy (SERDS), an even better tool for process analytics and cancer diagnostics. The integration of the above-mentioned spectra-processing approaches into SERDS in combination with machine learning tools enabled the differentiation between physiological mucosa, non-malignant lesions, and oral squamous cell carcinomas with high accuracy, above the state of the art. The

distinguishable features obtained in the purified Raman spectra are assignable to different chemical compositions of the respective tissues. The feasibility of a similar approach for breast tumors was also investigated. The purified Raman spectra of normal breast tissue, fibroadenoma, and invasive carcinoma were discriminable with respect to the spectral features of proteins, lipids, and nucleic acid. These findings suggest the potential of SERDS combined with machine learning techniques as a universal tool for cancer diagnostics.

## Zusammenfassung der Ergebnisse der Dissertation

Die Raman-Spektroskopie ist eine optische Messtechnik, die in der Lage ist spektroskopische Information zu liefern, welche molekülspezifisch und einzigartig in Bezug auf die Eigenschaften der untersuchten Spezies sind. Sie ist ein unverzichtbares analytisches Instrument, das Anwendung in verschiedenen Bereichen findet, wie etwa der Medizin oder der *in situ* Beobachtung von chemischen Prozessen. Wegen ihren Eigenschaften, wie der hohen Spezifität und der Möglichkeit von Tracer-freien Messung, hat die Raman-Spektroskopie die Tumordiagnostik stark beeinflusst. Aufgrund einer äußerst starken Beeinflussung der Raman-Spektren durch Hintergrundsignale, ist das Isolieren und Interpretieren von Raman-Spektren eine große Herausforderung.

Im Rahmen dieser Arbeit wurden verschieden Ansätze der Spektrenbearbeitung entwickelt, die benötigt werden um Raman-Spektren aus verrauschten und stark mit Hintergrundsignalen behafteten Rohspektren zu extrahieren. Diese Ansätze beinhalten im Speziellen eine auf dem *vector-casting* basierende Methode zur Rauschminimierung und eine auf dem deep neural networks basierende Methoden zur Entfernung von Rauschen und Hintergrundsignalen. Verschiedene neuronale Netze wurden mittels simulierter Spektren trainiert und an experimentell gemessenen Spektren evaluiert. Die im Rahmen dieser Arbeit vorgeschlagenen Ansätze wurden mit alternativen Methoden auf dem aktuellen Stand der Entwicklung unter Zuhilfenahme von verschiedenen Signal-Rausch-Verhältnissen, Standardabweichungen und dem *structural similarity index* verglichen. Die hier entwickelten Ansätze zeigen gute Ergebnisse und sind bisher bekannten Methoden überlegen, vor allem für Raman-Spektren mit einem niedrigem Signal-Rausch-Verhältnis und extrem starken Fluoreszenz-Hintergrund. Zusätzlich erfordern die auf *deep neural networks* basierten Methoden keinerlei menschliches Eingreifen.

Die Motivation hinter dieser Arbeit ist die Verbesserung der Raman-Spektroskopie, vor allem der *shifted-excitation Raman difference spectroscopy* (SERDS) hin zu einem noch besseren Instrument in der Prozessanalytik und Tumordiagnostik. Die Integration der oben genannten Ansätze zur Spektrenbearbeitung von SERDS in Kombination mit Methoden des maschinellen Lernens ermöglichen es, physiologische Schleimhaut, nicht-maligne Läsionen und oralen Plattenepithelkarzinomen mit einer Genauigkeit zu unterscheiden, die bisherigen Methoden überlegen ist.

Die spezifischen Merkmale in den bearbeiteten Raman-Spektren können verschiedenen chemischen Zusammensetzungen in den jeweiligen Geweben zugeordnet werden. Die Übertragbarkeit auf einen ähnlichen Ansatz zu Erkennung von Brusttumoren wurde überprüft.

Die bereinigten Raman-Spektren von normalem Brustgewebe, Fibroadenoma und invasiven Mammakarzinom konnten mithilfe der spektralen Eigenschaften von Proteinen, Lipiden und Nukleinsäuren unterschieden werden. Diese Erkenntnisse lassen das Potential von SERDS in Kombination mit Ansätzen des maschinellen Lernens als universelles Werkzeug zur Tumordiagnose erkennen.

# Table of Contents

Versicherung .....	i
Abstract .....	ii
Zusammenfassung der Ergebnisse der Dissertation .....	iv
Table of Contents .....	vi
Abbreviations and symbols .....	vii
1 Introduction.....	1
2 State of the art of the purification of Raman spectra .....	4
2.1 Experimental methods for the enhancement of the signal-to-background ratio and the signal-to-noise ratio.....	6
2.2 Mathematical methods for the extraction of pure Raman spectra from raw spectra .....	9
2.3 Raman based cancer diagnostics .....	14
2.4 Neural networks for the evaluation of Raman spectra .....	17
2.5 Objective .....	19
3 Application relevant fundamentals .....	27
3.1 Basics of Raman spectroscopy .....	27
3.2 Simulation of raw Raman spectra.....	34
3.3 Shifted-excitation Raman difference Spectroscopy .....	37
3.4 Raman experimental setup .....	45
3.5 Mathematical method for Raman spectra refinement.....	48
3.6 Deep neural networks .....	64
4 Summary of the published results .....	74
4.1 A shifted-excitation Raman difference spectroscopy evaluation strategy for the efficient isolation of Raman spectra from extreme fluorescence interference .....	75
4.2 Vector casting for noise reduction .....	78
4.3 Refinement of spectra using a deep neural network; fully automated removal of noise and background.....	81
4.4 Breast Tumor Analysis using Shifted-Excitation Raman difference Spectroscopy .....	83
4.5 Optical diagnosis of clinically apparent lesions of oral cavity by label-free Raman spectroscopy .....	86
Conclusion.....	90

## Abbreviations and symbols

SNR	Signal-to-noise ratio	
SBR	Signal-to-background ratio	
CARS	Coherent anti-stokes Raman spectroscopy	
SRS	Stimulated Raman spectroscopy	
RRS	Resonance Raman spectroscopy	
SERS	Surface-enhanced Raman spectroscopy	
SPR	Surface plasmon resonance	
SERDS	Shifted-excitation Raman spectroscopy	
airPLS	Adaptive iteratively reweighted penalized least squares	
ALS	Asymmetric least squares	
arPLS	Asymmetrically reweighted penalized least squares	
CT	Computer tomography	
MRI	Magnetic resonance imaging	
SVM	Support vector machines	
CNN	Convolutional neural network	
HWHM	Half-width at half maximum	
FWHM	Full-width at half maximum	
ModPoly	Modified polynomial fit	
ML	Machine learning	
OSCC	Oral squamous cell carcinoma	
PCA	Principal component analysis	
LDA	Linear discriminant analysis	
IR	Infrared	
$E$	Photon energy	$J$
$h$	Plank's constant	$J \cdot s$
$\nu$	Frequency of light	$Hz$
$\lambda_i$	Incident photon wavelength	$nm$
$\lambda_s$	Scattered photon wavelength	$nm$
$r$	Raw Raman spectrum	-
$R$	Raman spectrum	-



$R_{recov}$	Recovered Raman spectrum	-
$n$	Noise spectrum	-
$B$	Background spectrum	-
$\bar{\nu}$	Raman shift	$cm^{-1}$
$\sigma$	HWHM of Gaussian peak	-
$\gamma$	HWHM of Lorentzian peak	-
$n_{ph}$	Shot noise	-
$n_{rd}$	Read-out noise	-
$n_{th}$	Thermal noise	-
$\Delta\lambda$	Excitation wavelength shift	$nm$
$w_i$	Weighting factor	-
$e$	Error signal	-
$\lambda$	Smoothing factor	-
$p$	Fit parameter	-
$\theta$	Machine learning trainable parameters	-

# 1 Introduction

Raman spectroscopy is an optical measurement technique for the remote and non-invasive analysis of the chemical composition of samples. It has been discovered in 1928 by C.V. Raman [1]. The main advantages of Raman spectroscopy are that often no sample preparation is required and that - from the detected signal spectrum one can quantify the chemical composition [2-4], the temperature [5-8], the state of phase [9, 10] or the heterogeneity [11, 12] of the sample. Therefore, Raman spectroscopy qualifies as analytical tool especially, when no physical samples can be or should be extracted from the measurement object, for example for the analysis of paintings or human tissue, or when the extraction of physical samples would drastically change the sample conditions, for example in reacting flows or high-pressure and/or high-temperature environments. With respect to cancer diagnostics, Raman spectroscopy enables objective tissue characterization *in vivo* and thus can avoid the resection of suspicious abnormal tissue, which eventually is only inflamed.

Within the scientific environment of the author at the Technische Universität Bergakademie Freiberg, Germany, Raman spectroscopy is a frequently used measurement technique. In order to explore new grounds in chemical engineering and medical diagnostics, the working group is obliged to advance Raman spectroscopy beyond the state of the art.

Medical Raman diagnostics often suffers from low signal-to-background ratios (SBR) and low signal-to-noise ratios (SNR). Both, low SBR and low SNR, complicate the reliable extraction of quantitative information from Raman spectra.

Therefore, the task of this thesis is the development of advanced methods for the evaluation of raw Raman spectra, which suffer from low SNRs and low SBRs. The methods to be developed are supposed to advance the state-of-the-art techniques with respect to

1. the degree of human intervention required to process the spectra
2. data manipulation caused by the evaluation method itself.

Therefore, this dissertation is structured as follows: The “Introduction” is followed by section two, which provides the state of the art with respect to the “Purification of Raman Spectra”. Section two is a compilation of published approaches and methods for the improvement of the SNR and SBR as well as the mathematical treatment of raw Raman spectra. All these approaches and methods have in common that they aim at obtaining quasi-interference free pure Raman spectra. Section three describes the “Application relevant fundamentals”. The fundamentals cover Raman spectra, experimental Raman setups, state of the art mathematical post-processing methods, and finally, machine learning for the processing and evaluation of spectra. Section four provides a frame around the five papers which constitute the results part of this thesis. Section five summarizes the main achievements and presents the conclusions.

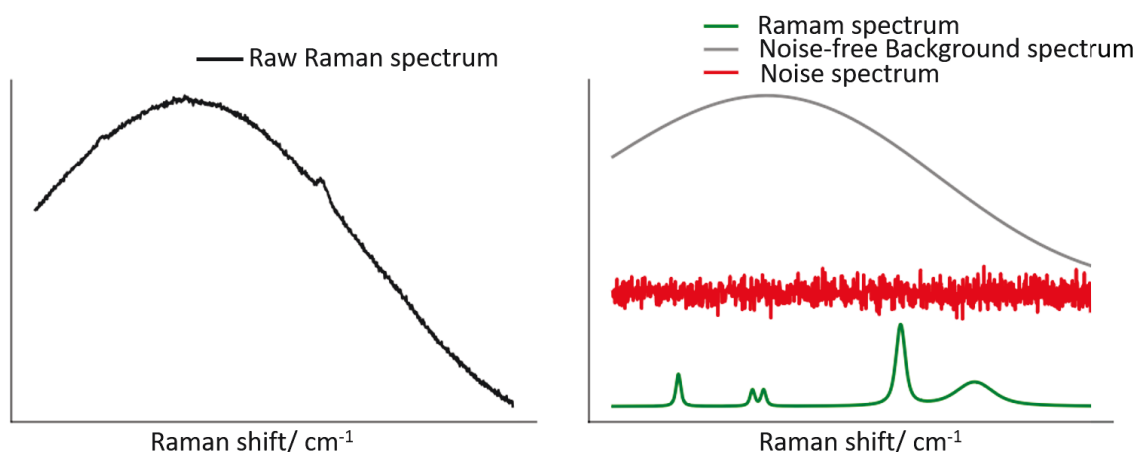
## References within section 1

1. C.V. Raman and K.S. Krishnan, *A new type of secondary radiation*. Nature, 1928. **121**(3048): p. 501-502.
2. J.J. Schuster, et al., *Online monitoring of the supercritical CO<sub>2</sub> extraction of hop*. The Journal of Supercritical Fluids, 2018. **133**: p. 139-145.
3. F.L.J. Cals, et al., *Raman spectroscopic analysis of the molecular composition of oral cavity squamous cell carcinoma and healthy tongue tissue*. Analyst, 2018. **143**(17): p. 4090-4102.
4. B. Volodin, et al., *Application of the shifted excitation Raman difference spectroscopy (SERDS) to the analysis of trace amounts of methanol in red wines*. Proceeding of SPIE 8939, 2014: p. 89390Y-89390Y-10.

5. L.A. Bahr, et al., *Temperature determination of superheated water vapor by rotational-vibrational Raman spectroscopy*. Optics Letters, 2018. **43**(18): p. 4477-4480.
6. R. Hickman and L. Liang, *Rotational temperature measurement in nitrogen using Raman spectroscopy*. Review of Scientific Instruments, 1972. **43**(5): p. 796-799.
7. T.C. Klima, et al., *Quantification of mixture composition, liquid-phase fraction and-temperature in transcritical sprays*. The Journal of Supercritical Fluids, 2020. **159**: p. 104777.
8. T.C. Klima and A.S. Braeuer, *Raman Thermometry in Water, Ethanol, and Ethanol/Nitrogen Mixtures from Ambient to Critical Conditions*. Analytical Chemistry, 2018. **91**(1): p. 1043-1048.
9. T.C. Klima and A.S. Braeuer, *Vapor-liquid-equilibria of fuel-nitrogen systems at engine-like conditions measured with Raman spectroscopy in micro capillaries*. Fuel, 2019. **238**: p. 312-319.
10. A. Hédoux, Y. Guinet, and M. Descamps, *The contribution of Raman spectroscopy to the analysis of phase transformations in pharmaceutical compounds*. International Journal of Pharmaceutics, 2011. **417**(1-2): p. 17-31.
11. L.A. Bahr, et al., *In situ analysis of aerosols by Raman spectroscopy—Crystalline particle polymorphism and gas-phase temperature*. Journal of Aerosol Science, 2018. **126**: p. 143-151.
12. D. Bassing and A.S. Braeuer, *The lag between micro-and macro-mixing in compressed fluid flows*. Chemical Engineering Science, 2017. **163**: p. 105-113.

## 2 State of the art of the purification of Raman spectra

Figure 2.1 shows a synthesized raw Raman spectrum on the left side. It is shown on the right side that the synthesized raw Raman spectrum is the summation of a noise-free pure Raman spectrum (green) of a substance, a synthesized noise-free background spectrum (grey) and a synthesized noise spectrum (red). The synthesized noise spectrum regards the noise that in reality would be contained in the pure Raman spectrum and the background spectrum. Furthermore, it regards read-out noise and thermal noise, inherent to the detectors usually used in Raman spectroscopy. The application relevant fundamentals on the synthesis of raw Raman spectra and the simulation of pure Raman spectra are provided in section three in detail and shall here be taken as granted.



*Figure 2.1: Illustration of interferences to Raman measurements. Synthesized raw Raman spectrum (left black curve) which is a summation of synthesized noise free background spectrum (grey curve), synthesized noise spectrum (red curve) and a pure Raman spectrum (green curve).*

The pure Raman spectrum contains the desired information about the sample, which can be the chemical composition [1-3], the temperature [4-7], the polymorphism [8, 9] and others [10-13]. Unfortunately, in a real measurement scenario, one would measure the raw Raman spectrum. The extraction of the pure Raman spectrum from the raw Raman spectrum is a rather challenging task, especially if – unlike shown in the synthetization scenario in Figure 2.1 – one doesn't know neither the pure Raman spectrum, the background spectrum nor the noise spectrum in advance.

One can say that the extraction of the pure Raman spectrum from a raw Raman spectrum is the simpler, the larger the signal-to-noise ratio (SNR) is and the larger the signal-to-background ratio (SBR) is. In Figure 2.1 the SNR and the SBR are 5 dB and -40 dB, respectively. The definitions of SNR and SBR are provided later in section three. It is thus not surprising that different experimental methods have been developed for the enhancement of the pure Raman signal relative to the background and relative to the noise. One of the following subsections aims at providing the state of the art of these methods.

Nevertheless, under certain circumstances, these experimental methods are not applicable and/or – though these methods are applied – the resulting SBRs and SNRs are still low. Then, mathematical evaluation methods can help extract the pure Raman spectrum from the raw Raman spectrum. The second following subsection aims at presenting the state of the art of these mathematical evaluation methods.

It must be mentioned that this section “State of the art” is rather a compilation of the respective publications than a detailed description of the methods. If the mentioned methods are relevant for this thesis, they are described in detail in the section “Application relevant fundamentals”.

Other following subsections provide the state of the art with respect to Raman-based cancer diagnostics and the application of neural networks for the evaluation of Raman spectra.

## **2.1 Experimental methods for the enhancement of the signal-to-background ratio and the signal-to-noise ratio**

There are various possibilities to increase the SBR or SNR. This can be achieved by either increasing the Raman signal or by decreasing the background and the noise. Increasing the excitation laser power is a simple method for increasing the Raman signal, but often not applicable, simply because of laser radiation can harm the sample under investigation, due to photodegradation, thermal decomposition or burning [14]. Therefore, methods are reported here, which are not simply based on the linear relation between laser excitation power and signal intensity.

Non-linear variants of Raman spectroscopy, such as coherent anti-stokes Raman spectroscopy (CARS) [15] or stimulated Raman spectroscopy (SRS) [16, 17] exploit the non-linear relation between the excitation intensity and the Raman signal intensity. This non-linear interdependency exists above a certain threshold of minimum excitation intensity. Above this threshold an increase in the excitation intensity results in a disproportionate increase in the signal intensity. Short-pulse lasers are required for the experimental realization of such non-linear Raman variants, because of what the entire experimental setup is complex compared to the linear variant of Raman spectroscopy. Additionally, the acquisition of entire spectra and the interpretation of the obtained spectra are challenging [18].

Resonance Raman spectroscopy (RRS) [19-22] uses excitation wavelengths close or identical to an electronic transition of a molecule contained in the sample. Due to the near-resonant excitation of the transition, the Raman signal assignable specifically to a related transition is much stronger than compared to not-near-resonant excitation. Therefore, only

specific Raman peaks, belonging to the near-resonantly excited transitions, but not the entire Raman spectrum, can be enhanced. Consequently, RRS is mainly applied for the qualitative detection of various species or for the quantification of the concentration of only one species rather than for quantitative mixture analysis. Wu et al. [21] applied this technique for human breast cancer diagnosis. They reported enhancement of specific Raman signatures of normal human breast tissue when excited with 532 nm excitation wavelength due to resonance effect as compared to Raman spectra of the cancerous human breast tissue. Zhou et al. [22] also carried out a resonance Raman spectroscopic investigation of the discrimination between human brain normal, benign and cancer meningeal tissues using 532 nm excitation wavelength. Their findings indicated a resonance enhancement of amide II protein molecule in the spectra of malignant meningioma tissue.

Surface-enhanced Raman spectroscopy (SERS) [23, 24] makes use of metallic nanoparticles such as gold or silver at very close proximity of the sample under investigation. Laser excitation of these roughened metallic nanoparticles causes collective oscillation of their surface electrons, termed surface plasmon resonance (SPR). The resonance of the frequency of the incident light with this plasmonic surface leads to a great enhancement of the electromagnetic field around the nanoparticles ("hot spot region"). A molecule placed in the vicinity of this hot spot then experiences enhanced electric field strength. This increase in the strength of the electric field enhances the intensity of the Raman signal significantly. Eryilmaz et al. [25] applied SERS for the rapid analysis of total protein content. They reported the ability of SERS to quantify total protein fraction in milk samples. Zhai et al. [26] carried out SERS analysis of residual pesticides in apple samples. They concluded that SERS has a great potential to serve as a useful means for detecting pesticide residues. Qian et al. [27] also reported the application of SERS for lung cancer detection using saliva samples. The metallic nanoparticles are the key element of a SERS experiment. Fabrication of these metallic particles/substrates for



high Raman signal enhancement and reproducibility is still a challenge [28]. This makes label-free SERS spectra very complex and a full band assignment is often not possible. Current developments of this method also applied SERS-tags [29] for indirect detection of samples under investigation. These tags are created by attaching intrinsically strong Raman scattering molecules (Raman reporters) to the nanoparticles. Therefore, an enhanced Raman spectrum of the Raman reporter molecule is collected instead of the Raman spectrum of the molecule under investigation. Generally, the reliability of the SERS method depends on the appropriate sample preparation and adequate data preprocessing to improve SNR and SBR [30].

The time-gated [31-34] approach also belongs to the experimental methods. It takes advantage of different life-times of Raman scattering and fluorescence emissions. If the excitation laser is shortly pulsed or if the excitation laser power is temporally modulated, the pure Raman spectrum can be separated from the fluorescence background based on analyzing the time lag between the excitation and the signal modulation [31, 35]. If entire spectra should be recorded, it requires pulsed lasers with rapidly time gated spectral detectors. However, if the Raman signal should be measured only for a certain Raman-shift, the usage of a temporally modulated excitation laser power with a temporally highly resolving zero-dimensional detector, such as a photo-diode, is sufficient. Summarizing, the realization of this technique either depends on the availability of expensive equipment or provides Raman signal at specific Raman shifts only [36]. Lipiaainen et al. [34] applied this method for the quantitative assessment of ternary mixtures of solid-state forms of a model drug called piroxicam (PRX). They demonstrated that time-gated Raman spectroscopy has the potential for quantifying ternary mixtures of fluorescent pharmaceuticals during drug development and manufacturing.

Shifted-excitation Raman difference spectroscopy (SERDS) [37] is a method for pure Raman spectrum extraction [38]. As SERDS plays an important role within this thesis, a detailed description of its working principle will follow in the “Application relevant

fundamentals”. It follows Kasha’s rule [39], which states that the fluorescence emission is unaltered for a small change in the excitation photon energy, but the Raman spectrum shifts according to the excitation photon energy change [37]. Elimination of fluorescence succeeds by the subtraction of two raw Raman spectra acquired with slightly different excitation wavelengths (photon energies). After subtraction and under ideal/perfect circumstances, the difference spectrum is free of background contributions. Besides removing fluorescence, this method can suppress other undesired interferences, such as ambient light and detector etaloning effects. This shows that the technique has the potential to be applied under daylight conditions, where spectroscopy measurements are usually not performed. Theurer et al. [40] applied SERDS for soil analysis and reported a 10-fold improvement of the signal-to-background ratio. Korinth et al. [12] also applied this method for pollen classification and they reported advantages of SERDS in scenarios where Raman spectra are affected by variations due to detector etaloning, ambient light and intensive background interference. However, in practice when the conditions are non-ideal, photobleaching and optical filter characteristics affect varying background contributions in the two raw Raman spectra. This imposes difficulties to obtain background free difference spectra. Consequently, mathematical post-processing methods often help further enhance the SNR and SBR, which will be reported in the following subsection.

## **2.2 Mathematical methods for the extraction of pure Raman spectra from raw spectra**

If the experimental methods for the enhancement of the SBR and the SNR did not provide the anticipated success or were not applicable or available, mathematical methods can help extract the pure Raman spectrum from the raw Raman spectrum. The acquired raw spectra are processed, once they have been recorded. Therefore, these mathematical methods are referred to as post-measurement processing methods. Again, they aim at enhancing the SNR and the SBR. All the methods mentioned below are described in detail in the section three “Application relevant fundamentals”.

Examples of commonly used methods for de-noising of raw Raman spectra include the Savitzky-Golay (SG)-smoother [41], perfect smoother [42], and smoothing based on wavelet transform [43, 44]. The Savitzky-Golay smoother applies least-squares fitting of polynomials of specified order to connected data points contained in a moving window of specified size. The value of the data point which is situated in the center of the window is then updated from the value of the polynomial fit function at this data point. This whole process of polynomial fitting and updating new value within a shifted new spectral window is repeated for all data points of the raw Raman spectra. Appropriate selection of polynomial order and window size determines the performance of this method. The Savitzky-Golay smoother is considered as a standard method for denoising of raw Raman spectra [45-47]. The Perfect smoother has been shown to provide an improvement over the Savitzky-Golay smoother [48]. It is based on penalized least squares fitting. In order to improve the SNR, the denoised Raman spectrum has to be smoother but the smoother it is, the more it will deviate from the original spectrum. The main idea of perfect smoother is therefore to find a balance between the fidelity of the denoised raw spectra in comparison to the original spectra and the roughness of the denoised spectra. This involves tuning a parameter to find an optimum between these two conflicting goals. Recently, Zeng et al. [49] used this method when applying Raman spectroscopy for mixture analysis.

Smoothing based on wavelets involves transformation of the noisy raw Raman spectra into a wavelet domain by decomposing it into a set of specified orthonormal wavelet basis functions. The noise is suppressed in the wavelet domain by thresholding the wavelet coefficients which belong to the noise. Then the not-suppressed coefficients are reverse transformed to obtain the noise-reduced spectra. In this method, the selection of the wavelet basis functions, decomposition level and the thresholding rule have a great impact on its performance. This method was also proposed to denoise raw Raman spectra and reported better performance compared to the other methods [50-52].

Mathematical techniques based on polynomial fitting [53-55] and penalized least squares [56-59], are most common approaches to improve SBR in Raman spectroscopy. These mathematical techniques assume that the background is a broadband curve. Polynomial fitting techniques use a least-squares based polynomial fitting of specified order to approximate the broadband curve. Lieber et al. [55] modified the single polynomial fit of the background by rather involving a series of polynomial fits and they termed their method as modified polynomial fit (ModPoly). During the first iteration they fit a polynomial function with a specified order to the original raw Raman spectrum. Then they compared the polynomial-fit spectrum with the original raw Raman spectra along the whole spectral range. Afterwards, at each wavenumber the raw Raman spectrum intensity value is replaced by the value of the polynomial-fit spectrum whenever it exceeds the intensity value of the polynomial-fit. This creates a new spectrum with trimmed peaks. In a second iteration, they fit a new polynomial function to the new spectrum with the trimmed peaks. Again, at each wavenumber and whenever the new spectrum is above the new polynomial-fit function, it is replaced by the intensity value of the new polynomial-fit. This procedure repeats for further iterations and after every iteration the polynomial-fit drifts towards the broadband background. This method systematically excludes the Raman peaks in every iteration while the polynomial-fit function approximates the broadband background. Andersen et al. [60] applied this method recently while using Raman spectroscopy for the prediction of water holding capacity and pH in *longissimus lumborum* of pork. Methods based on penalized least squares use the principle of the perfect smoother [42] to balance the fidelity of the fit function in comparison to the raw Raman spectra and smoothness of the fit function. They include a weighting vector to treat peak regions and peak-free regions of the raw Raman spectra differently, thus systematically only fitting to the background signal of the raw Raman spectra. Eilers [61] was the first to propose penalized least squares to improve SBR. His method, asymmetric least squares (ALS) [61], updates the weighting vector iteratively such that a small weight is assigned to the data points where the original raw Raman spectra is above the fit function (vaguely

peak regions). Otherwise a large weight is assigned. This fitting and weight updating continues until no difference is detected between consecutive fit functions. This method employs two adjustable parameters, the smoothness parameter  $\lambda$  for balancing the fidelity and roughness of the fit function and the weighting parameter  $p$  for updating the weighting vector. An adaptive iteratively reweighted penalized least squares (airPLS) [59] was proposed as an improvement of the ALS method. Unlike the ALS method it assigns different weights automatically according to the difference between the raw Raman spectra and the fit functions in every iteration. Therefore, only one parameter – smoothness - needs to be adjusted. In order to improve the applicability of penalized least squares methods in noisy raw Raman spectra, an asymmetrically reweighted penalized least squares smoothing (arPLS) was also proposed by Baek et al. [56]. Methods based on penalized least squares are widely used for background correction of raw Raman spectra because they are fast and avoid the need for peak detection [45, 62, 63]. Background approximation methods that are based on approaches, other than polynomial fitting and penalized least squares fitting, have been reported [36, 64, 65]. A detailed review of background correction methods can be found elsewhere [66].

Due to their low cost and high flexibility, post-measurement processing techniques are widely applied for the refinement of pure Raman spectra. However, the choice of a suitable method or a combination of these methods dramatically impacts the Raman spectrum interpretation for qualitative and quantitative evaluation. These techniques involve tunable parameters that must be explicitly tuned for different raw Raman spectra acquired with different experimental settings or sample matrices. Thus, the post-processing techniques, especially when the adjustable parameters are not chosen properly, can distort Raman spectral signatures and impact the reliability of further spectral analysis. This imposes an obstacle for automation and real-time monitoring using Raman spectroscopy, especially when both the SBR and the SNR are small. The effect of tunable parameters is illustrated for the cases of the Savitzky-Golay smoother and asymmetrically reweighted penalized

least squares smoothing (arPLS; see above) in Figure 2.2 considering a raw Raman spectrum having an SNR and SBR of 15 dB and -14 dB, respectively.

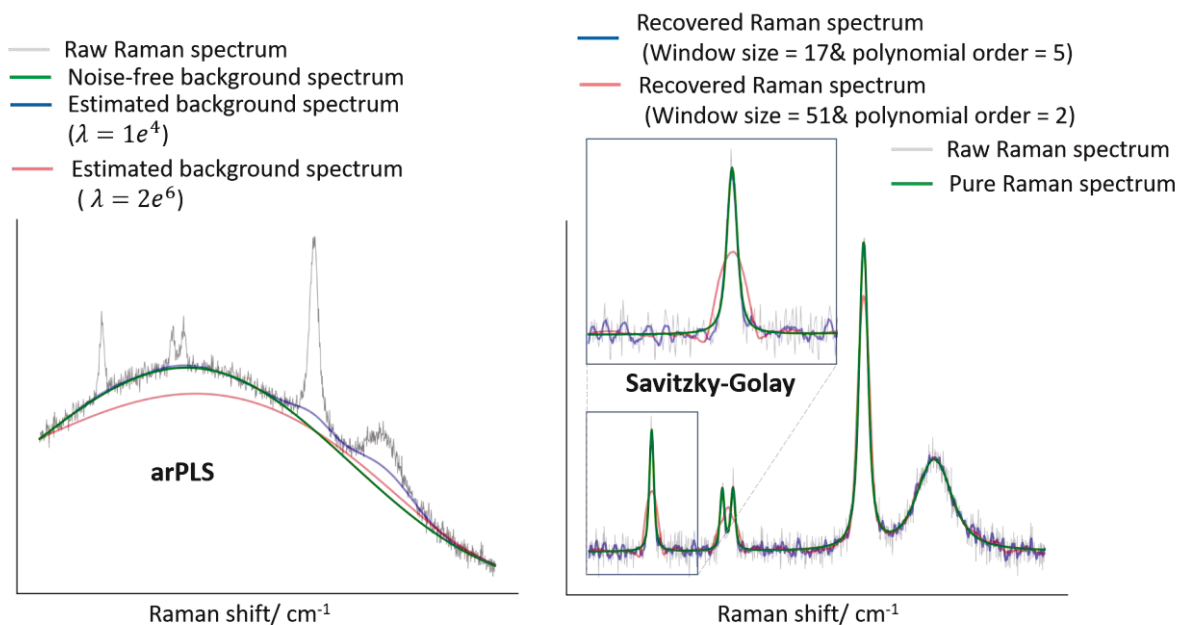


Figure 2.2: Illustration of the effect of tunable parameters: (left) arPLS method for background estimation applied to a raw Raman spectrum (grey curve) having SBR of -14 dB, setting the smoothness parameter to  $\lambda = 1e^4$  (blue curve) and to  $\lambda = 2e^6$  (red curve). (right) Savitzky-Golay smoother applied to denoise a raw Raman spectrum (grey curve) having a SNR of 15 dB, setting the window size to 17 and the polynomial order to 5 (blue curve), and the window size to 51 and the polynomial to 2 (red curve).

For example, in the case of the Savitzky-Golay smoother (see Figure 2.2 right side), selecting a larger moving window size or a smaller polynomial order means smoothing more the noisy spectrum. But the more the noise is smoothed out, the more the sharp Raman peaks are distorted. This becomes clear when comparing in Figure 2.2 (right) the pure Raman spectrum (green) with the two “Savitzky-Golay recovered” Raman spectra (blue and red). They were derived from the raw Raman spectrum (grey) by smoothing it with the Savitzky-Golay smoother with a window size of 17 and a polynomial order of 5,

respectively with a window size of 51 and a polynomial order of 2. Considering the larger window size and the lower polynomial order, the SNR is improved compared to the smaller window size and larger polynomial degree, but vice versa the original signature of the Raman peak is dramatically altered.

On the left side of Figure 2.2 the effect of parameter tuning on background correction is shown considering arPLS as an example. Setting the smoothness of the arPLS to  $\lambda = 1e^4$  achieves a better estimation of the background signal under the narrow Raman peaks. However, the estimated background signal interacts with the Raman spectrum, especially in the region of broad Raman peaks/bands. Increasing the smoothness parameter to  $\lambda = 2e^6$  improves the estimation of the background under the broad peaks/bands but the estimated background deviates in the region of the narrow Raman peaks.

### **2.3 Raman based cancer diagnostics**

Cancer is a major public health problem worldwide and the leading cause of premature death in most countries [67, 68]. According to the world cancer report 2020 [69], cancer is the first or second leading cause of premature death (i.e. at ages 30 to 69 years) in 134 of 183 countries. Accurate and early diagnosis combined with adequate surgical treatment can reduce this high mortality rate [70]. Most cancer types also develop through a pre-malignant stage. Hence, early detection of pre-malignant lesions can prevent further development of cancer. Presently, cancer diagnosis requires different approaches depending on the type of cancer. These approaches involve visual inspection of suspicious lesions, lab tests for blood, urine and other body fluid samples, and medical imaging methods [71]. New developments in screening technologies such as ultrasonography, computer tomography (CT), mammography, and magnetic resonance imaging (MRI) have improved cancer diagnosis regarding quality, selectivity, and diagnostic time. However, the final decision has still to be made using histopathological

analysis of suspicious tissue biopsies or diagnostic excisions. Histopathological examination, which involves specialized knowledge of a pathologist, is currently the gold standard for discriminating normal against diseased tissues. It involves fixation, micro-sectioning, and routinely staining with hematoxylin and eosin (H&E) of the biopsied or resected tissues. The pathologist is then responsible for making a final decision based on a microscopic examination of the H&E stained section. This approach is invasive and demands considerable time and cost. Further, the distinction between malignant, premalignant and benign tumors often involves repeated biopsies/excisions, which increases the diagnosis time and the risk of patient harm and anxiety. Another issue is sampling error, because the biopsy may not be representative of the whole lesion.

Surgery is a primary treatment of most cancer patients where surgeons decide on accurate localization of tumors and identification of their margins [70]. During the surgical treatment, the complete resection of cancer, while preserving healthy structures, increases the 5-year survival significantly and improves the patient's quality of life [71]. However, delineating the entire tumor without affecting surrounding healthy tissue is challenging because surgeons commonly rely on visual and physical inspection to define the borders between healthy and tumor regions. This may lead to tumor recurrence due to the residual tumor after surgery, which is associated with a poor prognosis and may need additional surgery. A histopathological analysis is also not applicable for tumor margin assessment during surgery due to a lack of the capability for providing immediate feedback. Currently, to support assessment of intraoperative resection margins, the surgeons need direct pathologist involvement for an intra-operative consultation using a small piece of frozen sections of suspicious tissue sampled from the wound bed [72]. This process takes about 5 -10 minutes, excluding the pathologist's time to study it under the microscope. This examination also has to be done while the patient is under anesthesia on the operating table. The shortcomings of available techniques for *in vivo* and real-time cancer diagnostics have fueled research to develop a portable label-free, non-invasive



and inexpensive method for accurate and real-time cancer diagnosis and intraoperative surgical guidance.

Carcinogenesis changes tissue biochemistry that occurs even before the appearance of clinical symptoms, not visible to the naked eye, neither under the optical microscope [73]. Raman spectroscopy provides an intrinsic molecular fingerprint of chemical composition of tissue, an advantage that is lacking in other diagnostic techniques [74]. This provides the capability to detect biological tissue changes during the onset and progression of carcinogenesis, which is reflected in their Raman spectra. Thus, Raman spectroscopy is a potential candidate to address these unmet clinical needs that can complement established cancer screening tools as well as accurate guidance during the surgical treatment.

Raman spectroscopy has been explored in numerous studies employing various types of measurement systems, targeting different types of cancers, both in *ex vivo* and *in vivo* settings. Since the calendar year of 2018, more than 60 thousand scientific papers on the application of Raman spectroscopy for cancer diagnosis were published according to google scholar when searched for keyword combination 'Raman and cancer diagnostics'. Perhaps one of the most striking recent demonstration of the application of Raman spectroscopy for cancer detection is by Desroches et al. [75]. They engineered a tiny Raman probe into a commercially available biopsy system allowing brain tumor analysis before tissue harvesting. By analyzing the high wavenumber Raman spectra ( $>2600\text{ cm}^{-1}$ ), they reported a brain tumor detection accuracy with a sensitivity and specificity of 80% and 90%, respectively. They preferred the high wavenumber Raman region, because the laser-induced fluorescence contribution (background interference) in this region is low. However, for biological tissues, approximately 90% of the molecular Raman signatures are found in the 'fingerprint' spectral region at low wavenumbers covering the range from about  $500\text{ cm}^{-1}$  to  $1800\text{ cm}^{-1}$ . Thus, analyzing the fingerprint spectral region would - most probably - have been given an even better detection

accuracy. However, the separation of the pure Raman spectra from the intensive laser-induced fluorescence (background) in this region is challenging. Moreover, designing a tiny Raman probe using a single fiber for both, excitation and collection in the fingerprint spectral region is tricky due to background interference originating from the non-linear interaction inside the fiber.

Surface-enhanced Raman spectroscopy (SERS) is also regarded as a powerful emerging technique for intraoperative tumor detection and can provide orders of magnitude enhancement of the Raman signal [23]. This technique involves an injection of SERS tags [29] into tumor-bearing samples for indirect tumor detection. Due to the enhanced permeation and retention effect, the SERS tags accumulate in the tumor; thus, significantly higher intensity of Raman signal from the Raman reporters of the nanoparticles can be collected from the tumor as compared to normal tissue. This intensity difference of the Raman signal from tumor and normal tissue is then considered for tumor detection. Though interesting and much progress has been achieved in recent years, many questions remain regarding biocompatibility of used SERS tags and effects of substrate variation in spectral reproducibility. Furthermore, the less expensive and less complex spontaneous Raman spectroscopy is sufficient for cancer diagnostics under many circumstances [70]. Tissue laser-induced fluorescence and detector noise are the main challenges of Raman spectroscopy for its application in cancer diagnosis.

## **2.4 Neural networks for the evaluation of Raman spectra**

Machine learning techniques have frequently been utilized for the evaluation/interpretation of Raman spectra. They learn meaningful representations, or characteristics or patterns contained in the Raman spectra that help to answer a variety of scientific questions. Partial least squares (PLS) [76] based regression models were applied for the quantification and composition assessments of mixtures. Machine learning algorithms such as linear discriminant analysis (LDA) [77], principal component analysis (PCA) [78], support vector

machines (SVM) [79], gradient boosting [80] and random forest [81] have been also reported by other researchers [82-85] for an appropriate transformation of the Raman spectra into meaningful categories, such as for instance that obtained from biological samples of healthy tissues and tissues with a disease. Once these models, whether regression or classification, are built based on known information (training data), they can then be used for the interpretation of test Raman spectra, if the test Raman spectra are contained within the “room” of training data. Extrapolations are not possible. If for example a model is built on experimental raw spectra from pig-tissue, it is not applicable for the interpretation of human-tissue. In this way machine learning algorithms can be used to interpret the information content of the Raman spectra.

Deep learning, a sub-field of machine learning, has risen to the forefront of the artificial intelligence community. It shows excellent performances, especially in the fields of computer vision and pattern recognition. Motivated by the recent and widespread success of deep learning, deep neural networks - especially convolutional neural networks (CNN) - were applied to tackle challenges of Raman spectroscopy. Liu et al. [86] trained a CNN for mineral species classification using raw Raman spectra without spectral preprocessing, such as background correction, noise reduction, and feature extraction. Fan et al. [87] developed a CNN model to identify components in mixtures using raw Raman spectra. They evaluated their model using simulated and experimentally acquired raw Raman spectra of fluid and powder mixtures. They reported that CNN could learn spectral features and identify substances in both, artificial and real spectral datasets of mixtures. Xia et al. [88] also used a CNN model to extract features from raw Raman spectra. The extracted features were fed into a support vector machine (SVM) classifier to detect oral tongue squamous cell carcinoma. These findings showed the potential of CNN for the interpretation of raw Raman spectra with respect to mineral species recognition, mixture composition analysis or human tissue characterization. However, the usage of the trained CNN models is very much restricted to the specific applications. Data

quality also plays an important role on the performance of deep learning models [89]. Thus, a better performance could have been achieved if pure Raman spectra were evaluated instead of the raw Raman spectra.

## 2.5 Objective

The previous sections have shown that extensive research has been undertaken to develop methods which seek to advance the practical limitations of Raman spectroscopy. It has been illustrated that these limitations find their origin in the low Raman signal levels in relation to large background and noise interferences. Many of the undertaken efforts focused on the development of algorithms, which intend to separate the information-rich pure Raman spectrum from the undesired background and noise interferences. However, the mathematical extraction of pure Raman spectra from heavily interfered raw Raman spectra – according to the state of the art – relies on human-intervention-dependent algorithms. “Human-intervention-dependent” means that humans have to choose at least few input parameters. Taking for example a typical polynomial smoothing algorithm, the human has to choose a window size and a polynomial degree. As a consequence, the mathematically derived pure Raman spectrum depends on the human intervention. Furthermore, there is always the risk of altering the pure Raman spectrum when the raw Raman spectrum is mathematically processed.

Figure 2.3 shows an example of a raw Raman spectrum featuring a low SNR and a low SBR acquired from healthy human tissue of the oral cavity. The background and noise spectra overwhelmingly interfere with the pure Raman spectrum and impose complications to interpret the information which is important for cancer diagnostics. This thesis aims to develop and implement spectral processing methods for the successful refinement of pure Raman spectra from such high noise level and intensive background interferences with less spectral manipulation and no or minimal reliance on input parameters that have to be chosen by humans. According to Figure 2.3 the refined Raman

spectra can then be further given to machine learning algorithms for spectra interpretation. With respect to human tissue characterization, spectra interpretation can mean the classification of tissue as healthy, premalignant or malignant. The potential of the developed methods is demonstrated mainly by their application to cancer/tissue diagnostics, as herein extreme levels of background interferences are typical.

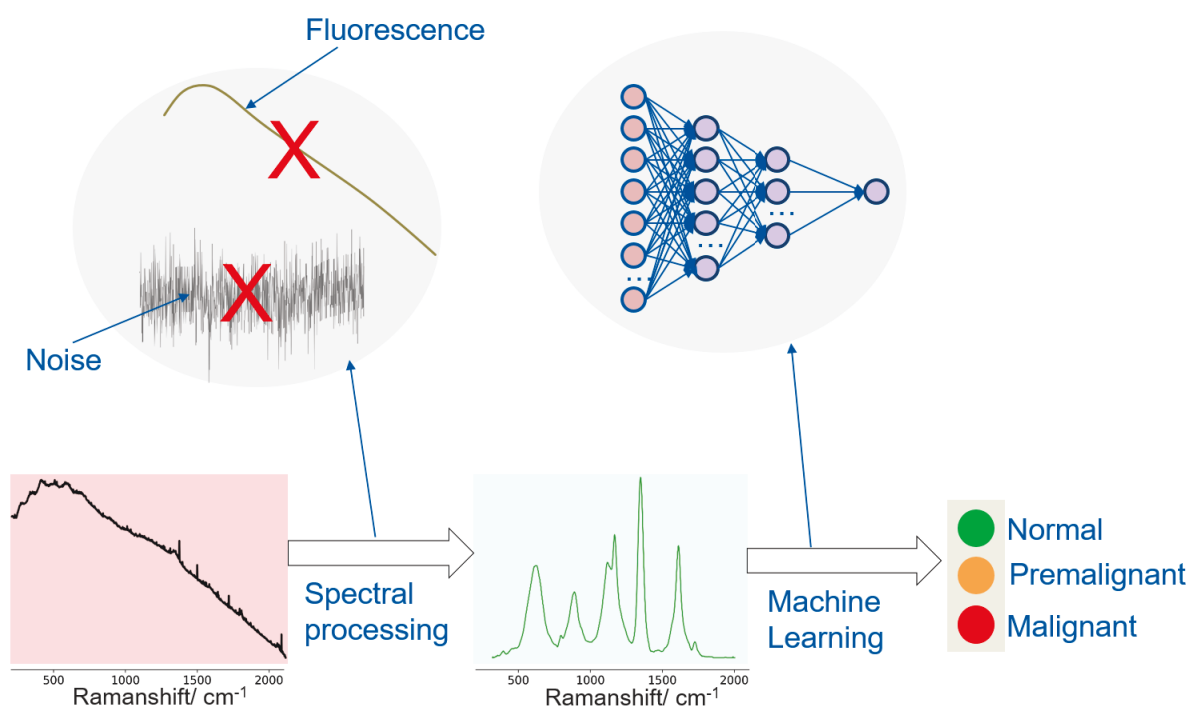


Figure 2.3: Overview of Raman spectroscopy combined with spectral processing and machine learning for cancer diagnosis.

## References within section two

1. J.J. Schuster, et al., *Online monitoring of the supercritical CO<sub>2</sub> extraction of hop*. The Journal of Supercritical Fluids, 2018. **133**: p. 139-145.
2. F.L.J. Cals, et al., *Raman spectroscopic analysis of the molecular composition of oral cavity squamous cell carcinoma and healthy tongue tissue*. Analyst, 2018. **143**(17): p. 4090-4102.

3. B. Volodin, et al., *Application of the shifted excitation Raman difference spectroscopy (SERDS) to the analysis of trace amounts of methanol in red wines*. Proceeding of SPIE 8939, 2014: p. 89390Y-89390Y-10.
4. L.A. Bahr, et al., *Temperature determination of superheated water vapor by rotational-vibrational Raman spectroscopy*. Optics Letters, 2018. **43**(18): p. 4477-4480.
5. R. Hickman and L. Liang, *Rotational temperature measurement in nitrogen using Raman spectroscopy*. Review of Scientific Instruments, 1972. **43**(5): p. 796-799.
6. T.C. Klima, et al., *Quantification of mixture composition, liquid-phase fraction and-temperature in transcritical sprays*. The Journal of Supercritical Fluids, 2020. **159**: p. 104777.
7. T.C. Klima and A.S. Braeuer, *Raman Thermometry in Water, Ethanol, and Ethanol/Nitrogen Mixtures from Ambient to Critical Conditions*. Analytical Chemistry, 2018. **91**(1): p. 1043-1048.
8. L.A. Bahr, et al., *In situ analysis of aerosols by Raman spectroscopy–Crystalline particle polymorphism and gas-phase temperature*. Journal of Aerosol Science, 2018. **126**: p. 143-151.
9. T.C. Klima and A.S. Braeuer, *Vapor-liquid-equilibria of fuel-nitrogen systems at engine-like conditions measured with Raman spectroscopy in micro capillaries*. Fuel, 2019. **238**: p. 312-319.
10. G. Fevotte, *In situ Raman spectroscopy for in-line control of pharmaceutical crystallization and solids elaboration processes: A review*. Chemical Engineering Research and Design, 2007. **85**(7): p. 906-920.
11. J. Zhu, et al., *Surface-enhanced Raman spectroscopy investigation on human breast cancer cells*. Chemistry Central Journal, 2013. **7**(1): p. 37.
12. F. Korinith, et al., *New methodology to process shifted excitation Raman difference spectroscopy data: A case study of pollen classification*. Scientific Reports, 2020. **10**(1): p. 1-12.
13. P. Vandenabeele, *Raman spectroscopy in art and archaeology*. Journal of Raman Spectroscopy, 2004. **35**(8 - 9): p. 607-609.
14. P. Vandenabeele, *Practical Raman spectroscopy: an introduction*. 2013 ed. 2013, Chichester, West Sussex: Wiley Online Library.
15. W.M. Tolles, et al., *A review of the theory and application of coherent anti-Stokes Raman spectroscopy (CARS)*. Applied Spectroscopy, 1977. **31**(4): p. 253-271.
16. R.W. Hellwarth, *Theory of stimulated Raman scattering*. Physical Review, 1963. **130**(5): p. 1850.
17. H. Lin, et al. *Deep learning spectroscopic stimulated Raman scattering microscopy*. in *Multiphoton Microscopy in the Biomedical Sciences XIX*. 2019. International Society for Optics and Photonics.
18. R.S. Das and Y. Agrawal, *Raman spectroscopy: recent advancements, techniques and applications*. Vibrational Spectroscopy, 2011. **57**(2): p. 163-176.

19. C.-H. Liu, et al., *Resonance Raman and Raman spectroscopy for breast cancer detection*. Technology in Cancer Research & Treatment, 2013. **12**(4): p. 371-382.
20. B. Robert, *Resonance Raman spectroscopy*. Photosynthesis Research, 2009. **101**(2): p. 147-155.
21. B. Wu, et al. *Characterization and discrimination of human breast cancer and normal breast tissues using resonance Raman spectroscopy*. in *Optical Biopsy XVI: Toward Real-Time Spectroscopic Imaging and Diagnosis*. 2018. International Society for Optics and Photonics.
22. Y. Zhou, et al., *Visible resonance Raman spectroscopy in human brain tissues*, in *Neurophotonics and Biomedical Spectroscopy*. 2019, Elsevier. p. 65-106.
23. J. Register, et al., *Shifted - excitation Raman difference spectroscopy for the detection of SERS - encoded gold nanostar probes*. Journal of Raman Spectroscopy, 2018. **49**(12): p. 1961-1967.
24. C.L. Haynes, A.D. McFarland, and R.P. Van Duyne, *Surface-enhanced Raman spectroscopy*. 2005, ACS Publications.
25. M. Eryilmaz, et al., *Rapid quantification of total protein with surface - enhanced Raman spectroscopy using o - phthalaldehyde*. Journal of Raman Spectroscopy, 2017. **48**(5): p. 653-658.
26. C. Zhai, et al., *Extraction and identification of mixed pesticides' Raman signal and establishment of their prediction models*. Journal of Raman Spectroscopy, 2017. **48**(3): p. 494-500.
27. K. Qian, et al., *New method of lung cancer detection by saliva test using surface - enhanced Raman spectroscopy*. Thoracic Cancer, 2018. **9**(11): p. 1556-1561.
28. R. Panneerselvam, et al., *Surface-enhanced Raman spectroscopy: bottlenecks and future directions*. Chemical Communications, 2018. **54**(1): p. 10-25.
29. Y. Wang, B. Yan, and L. Chen, *SERS tags: novel optical nanoprobe for bioanalysis*. Chemical Reviews, 2013. **113**(3): p. 1391-1428.
30. C. Zong, et al., *Surface-enhanced Raman spectroscopy for bioanalysis: reliability and challenges*. Chemical Reviews, 2018. **118**(10): p. 4946-4980.
31. J. Kostamovaara, et al., *Fluorescence suppression in Raman spectroscopy using a time-gated CMOS SPAD*. Optics Express, 2013. **21**(25): p. 31632-31645.
32. M. Kögler and B. Heilala, *Time-gated Raman spectroscopy—a review*. Measurement Science and Technology, 2020. **32**(1): p. 012002.
33. F. Gyakwaa, et al., *Characterisation of Binary Phase Mixtures of Magnesium-Aluminate Spinel and Calcium-Aluminates Using Time-Gated Raman Spectroscopy*. ISIJ International, 2020. **60**(5): p. 988-997.
34. T. Lipiäinen, et al., *Time-gated Raman spectroscopy for quantitative determination of solid-state forms of fluorescent pharmaceuticals*. Analytical Chemistry, 2018. **90**(7): p. 4832-4839.

35. J. Harris, et al., *Sub-nanosecond time-resolved rejection of fluorescence from Raman spectra*. Analytical Chemistry, 1976. **48**(13): p. 1937-1943.
36. M. Koch, et al., *Iterative morphological and mollifier - based baseline correction for Raman spectra*. Journal of Raman Spectroscopy, 2017. **48**(2): p. 336-342.
37. A.P. Shreve, N.J. Cherepy, and R.A. Mathies, *Effective rejection of fluorescence interference in Raman spectroscopy using a shifted excitation difference technique*. Applied Spectroscopy, 1992. **46**(4): p. 707-711.
38. F. Korinth, et al., *Wide Field Spectral Imaging with Shifted Excitation Raman Difference Spectroscopy Using the Nod and Shuffle Technique*. Sensors, 2020. **20**(23): p. 6723.
39. M. Kasha, *Characterization of electronic transitions in complex molecules*. Discuss. Faraday Soc., 1950. **9**: p. 14-19.
40. L.S. Theurer, M. Maiwald, and B. Sumpf, *Shifted excitation Raman difference spectroscopy: A promising tool for the investigation of soil*. European Journal of Soil Science, 2021. **72**(1): p. 120-124.
41. A. Savitzky and M.J. Golay, *Smoothing and differentiation of data by simplified least squares procedures*. Analytical Chemistry, 1964. **36**(8): p. 1627-1639.
42. P.H. Eilers, *A perfect smoother*. Analytical Chemistry, 2003. **75**(14): p. 3631-3636.
43. F. Ehrentreich and L. Sümmchen, *Spike removal and denoising of Raman spectra by wavelet transform methods*. Analytical Chemistry, 2001. **73**(17): p. 4364-4373.
44. P.M. Ramos and I. Ruisánchez, *Noise and background removal in Raman spectra of ancient pigments using wavelet transform*. Journal of Raman Spectroscopy, 2005. **36**(9): p. 848-856.
45. A. Ergin, et al. *Detection and analysis of glucose at metabolic concentration using Raman spectroscopy*. in 2003 IEEE 29th Annual Proceedings of Bioengineering Conference. 2003. IEEE.
46. X. Wang, et al., *Quantitative analysis of BPO additive in flour via Raman hyperspectral imaging technology*. European Food Research and Technology, 2017. **243**(12): p. 2265-2273.
47. M. Zhao, et al., *Performances of full cross-validation partial least squares regression models developed using Raman spectral data for the prediction of bull beef sensory attributes*. Data in Brief, 2018. **19**: p. 1355-1360.
48. S.J. Barton, T.E. Ward, and B.M. Hennelly, *Algorithm for optimal denoising of Raman spectra*. Analytical Methods, 2018. **10**(30): p. 3759-3769.
49. H.T. Zeng, et al., *Mixture analysis using non - negative elastic net for Raman spectroscopy*. Journal of Chemometrics, 2020. **34**(10): p. e3293.
50. Y. Xi, et al., *A novel pre-processing algorithm based on the wavelet transform for Raman spectrum*. Applied Spectroscopy, 2018. **72**(12): p. 1752-1763.



51. P. Prakash, et al., *Wavelet-Based Noise Removal from Raman Signal to Study PLD Coated Forsterite–Hydroxyapatite Thin Film on Stainless Steel 316L Substrate*. Journal of Applied Spectroscopy, 2020. **87**(3): p. 545-552.
52. L. Pan, et al., *Noise Reduction Technique for Raman Spectrum using Deep Learning Network*. arXiv preprint arXiv:2009.04067, 2020.
53. F. Gan, G. Ruan, and J. Mo, *Baseline correction by improved iterative polynomial fitting with automatic threshold*. Chemometrics and Intelligent Laboratory Systems, 2006. **82**(1-2): p. 59-65.
54. A. Cao, et al., *A robust method for automated background subtraction of tissue fluorescence*. Journal of Raman Spectroscopy, 2007. **38**(9): p. 1199-1205.
55. C.A. Lieber and A. Mahadevan-Jansen, *Automated method for subtraction of fluorescence from biological Raman spectra*. Applied Spectroscopy, 2003. **57**(11): p. 1363-1367.
56. S.-J. Baek, et al., *Baseline correction using asymmetrically reweighted penalized least squares smoothing*. Analyst, 2015. **140**(1): p. 250-257.
57. S. He, et al., *Baseline correction for Raman spectra using an improved asymmetric least squares method*. Analytical Methods, 2014. **6**(12): p. 4402-4407.
58. J. Peng, et al., *Asymmetric least squares for multiple spectra baseline correction*. Analytica Chimica Acta, 2010. **683**(1): p. 63-68.
59. Z.-M. Zhang, S. Chen, and Y.-Z. Liang, *Baseline correction using adaptive iteratively reweighted penalized least squares*. Analyst, 2010. **135**(5): p. 1138-1146.
60. P.V. Andersen, et al., *Prediction of water holding capacity and pH in porcine longissimus lumborum using Raman spectroscopy*. Meat Science, 2021. **172**: p. 108357.
61. P.H. Eilers and H.F. Boelens, *Baseline correction with asymmetric least squares smoothing*. Leiden University Medical Centre Report, 2005. **1**(1): p. 5.
62. H. Yuan, et al., *Classification and detection of testosterone propionate and nandrolone residues in duck meat using surface-enhanced Raman spectroscopy coupled with multivariate analysis*. Poultry Science, 2021. **100**(1): p. 296-301.
63. X. Zheng, et al., *Raman spectroscopy for rapid and inexpensive diagnosis of echinococcosis using the adaptive iteratively reweighted penalized least squares-Kennard–stone-back propagation neural network*. Laser Physics Letters, 2018. **15**(8): p. 085702.
64. Y. Liu, W. Cai, and X. Shao, *Intelligent background correction using an adaptive lifting wavelet*. Chemometrics and Intelligent Laboratory Systems, 2013. **125**: p. 11-17.
65. K.H. Liland, A. Kohler, and N.K. Afseth, *Model - based pre - processing in Raman spectroscopy of biological samples*. Journal of Raman Spectroscopy, 2016. **47**(6): p. 643-650.

66. G. Schulze, et al., *Investigation of selected baseline removal techniques as candidates for automated implementation*. Applied Spectroscopy, 2005. **59**(5): p. 545-574.
67. F. Bray, et al., *Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries*. CA Cancer J Clin, 2018. **68**(6): p. 394-424.
68. R.L. Siegel, K.D. Miller, and A. Jemal, *Cancer statistics, 2019*. CA: a Cancer Journal for Clinicians, 2019. **69**(1): p. 7-34.
69. C. Wild, E. Weiderpass, and B. Stewart, *World cancer report: cancer research for cancer prevention*. Lyon: International Agency for Research on Cancer, 2020.
70. N.M. Ralbovsky and I.K. Lednev, *Raman spectroscopy and chemometrics: A potential universal method for diagnosing cancer*. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2019. **219**: p. 463-487.
71. N.M. Ralbovsky and I.K. Lednev, *Towards development of a novel universal medical diagnostic method: Raman spectroscopy and machine learning*. Chemical Society Reviews, 2020.
72. J.M. Jorns, et al., *Intraoperative frozen section analysis of margins in breast conserving surgery significantly decreases reoperative rates: one-year experience at an ambulatory surgical center*. American Journal of Clinical Pathology, 2012. **138**(5): p. 657-669.
73. A. Mahadevan-Jansen and R.R. Richards-Kortum, *Raman spectroscopy for the detection of cancers and precancers*. Journal of Biomedical Optics, 1996. **1**(1): p. 31-71.
74. Q. Li, Q. Gao, and G. Zhang, *Classification for breast cancer diagnosis with Raman spectroscopy*. Biomedical Optics Express, 2014. **5**(7): p. 2435-2445.
75. J. Desroches, et al., *A new method using Raman spectroscopy for in vivo targeted brain cancer tissue biopsy*. Scientific Reports, 2018. **8**(1): p. 1-10.
76. C.A.F. de Oliveira Penido, et al., *Quantification of cocaine in ternary mixtures using partial least squares regression applied to Raman and Fourier transform infrared spectroscopy*. Journal of Raman Spectroscopy, 2017. **48**(12): p. 1732-1743.
77. H. Kim, B.L. Drake, and H. Park, *Multiclass classifiers based on dimension reduction with generalized LDA*. Pattern Recognition, 2007. **40**(11): p. 2939-2945.
78. I. Jolliffe, *Principal component analysis*. 2002: Wiley Online Library.
79. W.S. Noble, *What is a support vector machine?* Nature Biotechnology, 2006. **24**(12): p. 1565-1567.
80. A. Natekin and A. Knoll, *Gradient boosting machines, a tutorial*. Frontiers in Neurorobotics, 2013. **7**: p. 21.
81. G. Biau and E. Scornet, *A random forest guided tour*. Official Journal of the Spanish Society of Statistics and Operations Research (TEST), 2016. **25**(2): p. 197-227.

82. K. Noack, et al., *Combined shifted-excitation Raman difference spectroscopy and support vector regression for monitoring the algal production of complex polysaccharides*. Analyst, 2013. **138**(19): p. 5639-5646.
83. M.-J. Jeng, et al., *Raman spectroscopy analysis for optical diagnosis of oral cancer detection*. Journal of Clinical Medicine, 2019. **8**(9): p. 1313.
84. H. Sun, et al., *Application of KPCA combined with SVM in Raman spectral discrimination*. Optik, 2019. **184**: p. 214-219.
85. X. Zhang, et al., *Raman spectroscopy of follicular fluid and plasma with machine-learning algorithms for polycystic ovary syndrome screening*. Molecular and Cellular Endocrinology, 2021. **523**: p. 111139.
86. J. Liu, et al., *Deep convolutional neural networks for Raman spectrum recognition: a unified solution*. Analyst, 2017. **142**(21): p. 4067-4074.
87. X. Fan, et al., *Deep learning-based component identification for the Raman spectra of mixtures*. Analyst, 2019. **144**(5): p. 1789-1798.
88. J. Xia, et al., *Analysis and classification of oral tongue squamous cell carcinoma based on Raman spectroscopy and convolutional neural networks*. Journal of Modern Optics, 2020: p. 1-9.
89. V. Gudivada, A. Apon, and J. Ding, *Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations*. International Journal on Advances in Software, 2017. **10**(1): p. 1-20.

### 3 Application relevant fundamentals

This section briefly describes the fundamentals necessary to being able to follow the descriptions and interpretations given in this work. First, the basics of Raman spectroscopy are covered including its experimental challenges. Second, the simulation of raw spectra composed of noise, background and pure Raman spectra is described. Third, the basics of Shifted-Excitation Raman Difference Spectroscopy (SERDS) and its challenges are presented. Fourth, the Raman experimental setup used for measurements in this thesis is introduced. Fifth, details of mathematical methods for noise reduction and background correction are provided. The effect of adjustable parameters and their influence on mathematically derived Raman spectra is provided with the help of simulated raw Raman spectra. Finally, this section lays out the required knowledge about machine learning with a focus on deep neural networks.

#### 3.1 Basics of Raman spectroscopy

This subsection describes Raman spectroscopy briefly. The details of Raman spectroscopy are presented elsewhere [1]. Here the main focus will be that a Raman spectrum can be considered a fingerprint of a molecule and that it is very likely that the fluorescence emission (if existing) spectrally overlaps with it.

Figure 3.1 shows the term-energy of a diatomic molecule as a function of the displacement  $R$  of the separation between the cores of the two nuclei. For describing the principle of light-matter interaction a simple diatomic molecule is considered. But the concept is transferrable to more complex molecules because transitions in these complex molecules

can also be considered as transitions between two nuclei. The energy of a molecule is composed of the electronic energy, the vibrational energy and the rotational energy.

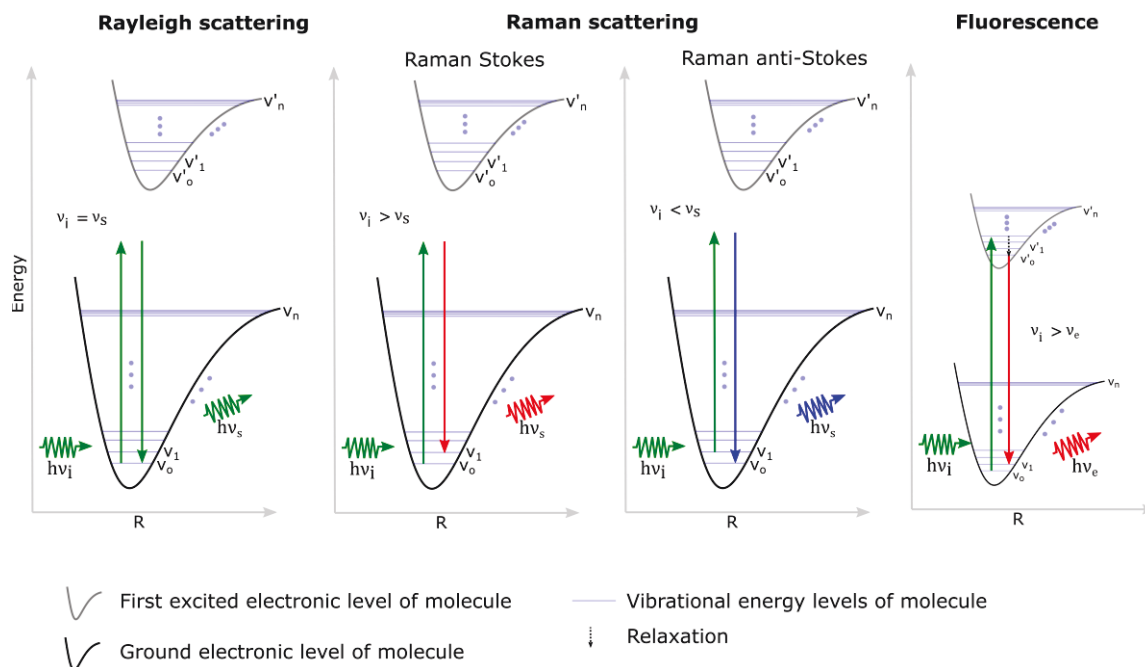


Figure 3.1: Illustration of the energy transfer between an incident photon and a molecule during different light scattering processes and laser-induced fluorescence emission considering a diatomic molecule for simplicity as a function of the distance between the cores of the two nuclei. Rayleigh scattering (left), Raman scattering (middle) and absorption-emission process (right)

The vibrational energy is a function of the mass of the nuclei, the displacement of the separation distance between the two nuclei (amplitude of vibration) and the “spring constant” (bond-properties) of the bond between the nuclei. The electronic energy depends only on the orbitals occupied by the electrons of the molecule. The rotational energy is a function of the mass of the nuclei, the distance between them and the rotational frequency. Therefore, the term-energy and also the energy difference between energy levels of any molecule depend on the specific properties of the molecule such as the mass of the nuclei, the spring constant and the distance between the nuclei. Only the potential curves of the ground and first excited electronic energy levels of the molecule, including its vibrational energy levels, are depicted in Figure 3.1. The superscript ‘

differentiates the vibrational energy levels of the excited electronic state ( $v'_0, v'_1, \dots, v'_n$ ) from the vibrational energy levels of the electronic ground state ( $v_0, v_1, \dots, v_n$ ). The rotational energy levels are excluded because they are not relevant for this study.

Light can be considered as composed of photons and the energy of one photon

$$E = h\nu \quad 3.1$$

is proportional to the frequency of light  $\nu$ , where  $h$  is the Plank's constant. Depending on the photon energy and energy difference between energy levels and properties of the molecules (such as Raman active or IR active), there are different possible ways that light can interact with a molecule. In the following only Raman scattering, Rayleigh scattering and laser-induced fluorescence will be discussed, as they are relevant for this thesis.

Figure 3.1 illustrates these three interactions between an incident photon and a molecule during a scattering process or a laser-induced fluorescence event. A scattering process occurs when the incident photon cannot be absorbed by the molecule but can be scattered, which means that the photon “collides” with the molecule and consequently is scattered away from its original path. This scattering process can be elastic or inelastic. In the first case, elastic scattering (Rayleigh scattering), there is no change of energy of neither the molecule nor the incident photon. Thus no energy exchange occurs during the collision. This is represented in Figure 3.1 (left) by the green arrows of identical length of the incident photon and scattered photon. In this context, the scattered photon has the same energy as the incident photon, but the propagation directions of the photon before and after the collision can be different. In the inelastic scattering process, which is termed Raman scattering, there is an energy transfer between the molecule and the incident photon. If the molecule receives energy from the incident photon, the scattering process is called Raman Stokes. Then the scattered radiation is red-shifted with respect to the incident radiation which means that the scattered photon has less energy than the incident

photon. This is represented, in Figure 3.1 second from left by the red arrow whose length is shorter than the green arrow length used to represent the incident photon. If energy is transferred from the molecule to the incident photon it is called Raman anti-Stokes scattering. The Raman anti-Stokes scattered radiation is blue-shifted with respect to the incident radiation which means that the scattered photon carries more energy after the scattering process (see Figure 3.1 second from right). Raman anti-Stokes is out of scope of this study. Only the Raman-Stokes effect is exploited.

Figure 3.1 (right) shows a laser-induced fluorescence process. As a first step the molecule absorbs an incident photon. Afterwards it undergoes a transition from the vibrational energy level  $v_0$  of its ground electronic state to the vibration energy level  $v'_1$  of its excited electronic state as shown in the figure with green arrow. The molecule then relaxes to the vibrational ground state  $v'_0$  releasing energy through non-radiative processes but it is still in its excited electronic state. Afterwards, it emits a new photon and undergoes a transition back to one of the vibration energy levels of its electronic ground state. In this specific example (see Figure 3.1 right) it undergoes a transition to the vibration energy level  $v_1$  of its electronic ground state. Since excited molecules can undergo transitions to several of their vibrational energy levels of the electronic ground state, the laser-induced fluorescence emission features a broadband spectrum. Radiation due to the laser-induced fluorescence process is also red-shifted, meaning the emitted photons have less energy than the incident photons. Thus, the fluorescence emission (if existing) spectrally overlaps with the Raman scattered signal, as both occur red-shifted with respect to the excitation photons (photons before the light-matter interaction). Biological tissue contains a large variety of molecules with many different functional groups and sizes. Consequently, when biological tissue is irradiated with monochromatic radiation, Raman scattered light as well as laser-induced fluorescence emissions emerge.

During Raman Stokes scattering the molecule undergoes a transition from its ground vibrational energy level ( $\nu_0$ ) to its first excited vibrational energy level ( $\nu_1$ ). Since every molecule has its own unique vibrational energy levels, the difference in energy between the incident photon and the scattered photon carries relevant information specific for that molecule that scattered the photon. This difference in energy is referred to as Raman shift  $\bar{\nu}$

$$\bar{\nu} = \left( \frac{1}{\lambda_i} - \frac{1}{\lambda_s} \right) \quad 3.2$$

in Raman spectroscopy, which is calculated as the difference in energy between the incident photon having a wavelength of  $\lambda_i$  and a scattered photon of wavelength  $\lambda_s$ . Therefore the Raman spectrum is a plot of the number of scattered photons as a function of the Raman shift.

What has been described for a diatomic molecule can be transferred to more complex molecules, such as in the presented example to ethanol. Ethanol is composed of carbon, hydrogen and oxygen. Existing intramolecular transitions are for example the C-H vibration between C and H nuclei in the methyl group, the O-H vibration between O and H nuclei in the hydroxyl group, the C-O vibration between the C and the O nuclei in the hydroxyl group and many more. Figure 3.2 shows that each peak in the Raman spectrum of ethanol can be assigned to an intramolecular transition, where the peak position (Raman shift) quantifies the energy of this transition and the peak height (or area under the peak) represents the probability that the respective transition scatters photons according to the Raman mechanism. In the case of ethanol, the C-H vibration is the most intense peak, as ethanol features in each molecule five of those bonds between C and H nuclei. It is mentioned in section 3.2 why the Raman peaks are not monochromatic lines but feature specific peak profile shapes.



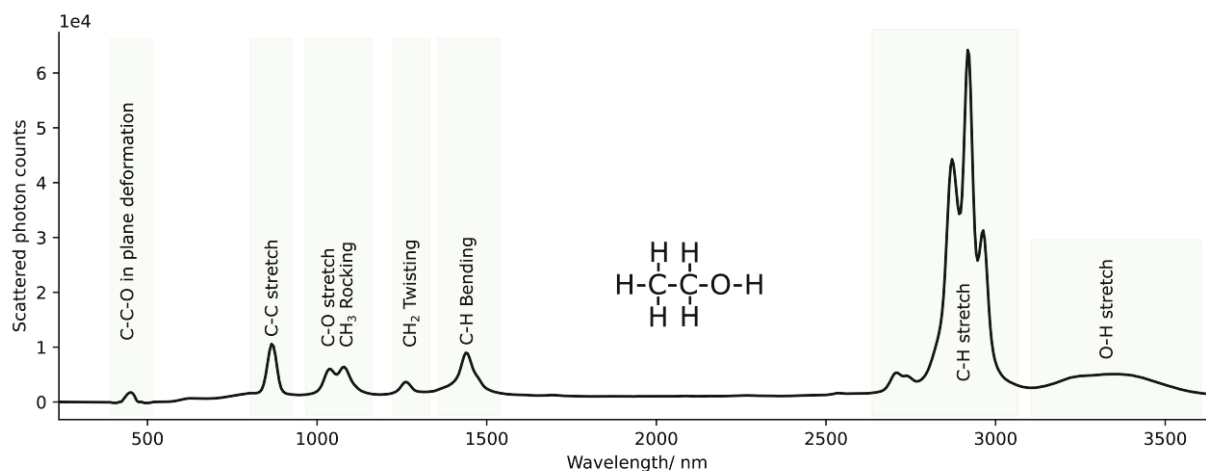


Figure 3.2: Raman spectrum of pure ethanol acquired with an excitation wavelength of 532 nm. Raman spectral signatures are assigned to molecular vibration modes of ethanol.

The experimentally acquired Raman spectrum of ethanol can be taken as pure Raman spectrum, as ethanol molecules cannot be excited to fluoresce under the here prevailing experimental conditions. In order to show the interference of the pure Raman spectrum with a broadband fluorescence emission a dye species was added to the ethanol. The dye cryptocyanine was chosen in order to generate a fluorescence emission once excited with the Raman excitation laser @ 785 nm.

Figure 3.3 shows the raw Raman spectra of two mixtures of the dye in pure ethanol with 0.015 and 0.123 weight fractions of the dye. The fluorescence is typically broad and featureless. It is also several orders of magnitude more intense than the Raman spectrum especially for the higher concentration of the dye (0.123 wt%), even though, the dye has a very low absorption coefficient at the wavelength of 785 nm. The green boxes in Figure 3.3 show the Raman signatures of pure ethanol. The Raman signatures are clearly visible in the raw Raman spectrum (black curve) of the mixture with less concentration of the dye but they are completely buried by fluorescence in the raw Raman spectrum of the mixture with higher concentration of the dye and are hardly visible. This imposes a barrier to

acquiring high-quality Raman signals especially for applications which involve fluorescent samples. The fluorescence presence also adds significant shot noise to the Raman spectrum, thus lowering the signal-to-noise ratio (SNR).

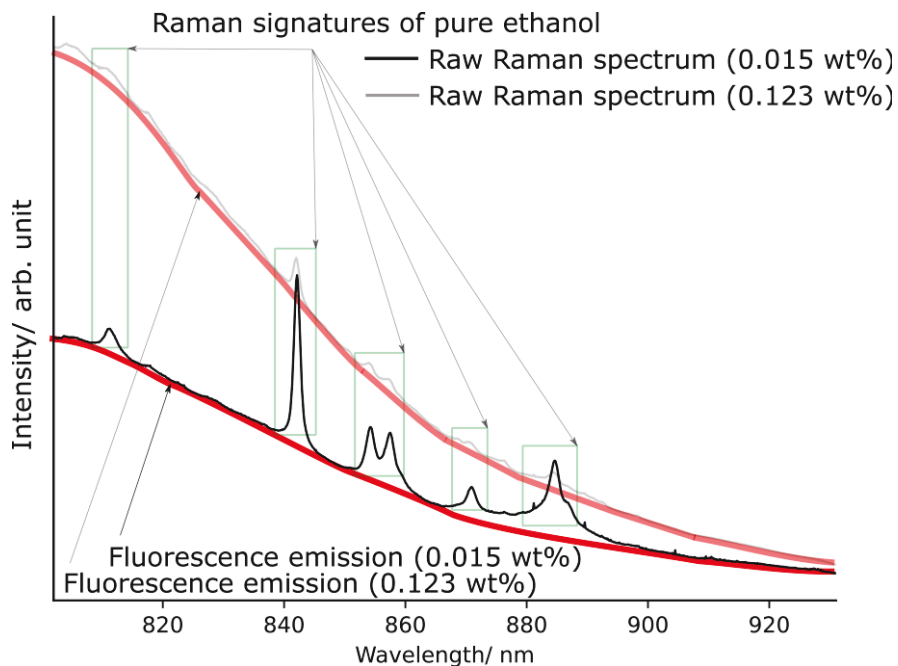


Figure 3.3: Raw Raman spectra of two mixtures of cryptocyanine dye 0.015wt% and 0.123wt% in pure ethanol shown in black and gray curves respectively. The estimation of the fluorescence emission is shown in red and light red. The Raman signatures of pure ethanol are indicated with green rectangular boxes.

Spectra like the experimental one shown in Figure 3.3 can also be synthesized by simulation. Simulation enables the rapid generation of raw Raman spectra, which especially in the case of artificial neural networks can serve as comprehensive training data.

### 3.2 Simulation of raw Raman spectra

The raw Raman spectrum  $r(\bar{\nu})$

$$r(\bar{\nu}) = R(\bar{\nu}) + B(\bar{\nu}) + n(\bar{\nu}) \quad 3.3$$

consists of a pure Raman spectrum  $R(\bar{\nu})$ , a background spectrum  $B(\bar{\nu})$  and a noise spectrum  $n(\bar{\nu})$ , where  $\bar{\nu}$  is Raman shift and computed as per equation 3.2.

Theoretically, the pure Raman spectrum consists of quasi-monochromatic spectral lines and bands as a function of the Raman shift reflecting the characteristics of the sample under investigation. However, several spectral line broadening physical mechanisms affect the Raman peak shape. Such broadening effects can be due to collisions which give rise to Lorentzian profile of the Raman peaks or thermal Doppler broadening which results in a Gaussian profile of the Raman peaks. The combination of both broadening mechanisms creates a Voigt profile in Raman spectral line shapes. Further broadening mechanisms can stem also from the Raman instrument such as for instance linewidth of the excitation laser sources. Raman peaks thus feature a distribution around their center frequency which can be fitted by Lorentzian, Gaussian or Voigt profiles. Considering this, a pure Raman spectrum  $R(\bar{\nu})$

$$R(\bar{\nu}) = \sum_{i=0}^M \begin{cases} P_h \frac{\Re[\text{wofz}(((\bar{\nu} - \bar{\nu}_0) + j\gamma)/\sigma\sqrt{2})]}{\sigma\sqrt{2\pi}} & \text{Voigt} \\ P_h \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(\bar{\nu} - \bar{\nu}_0)^2}{2\sigma^2}\right) & \text{Gaussian} \\ P_h \frac{\gamma/\pi}{(\bar{\nu} - \bar{\nu}_0)^2 + \gamma^2} & \text{Lorentzian} \end{cases} \quad 3.4$$

can be synthesized as a summation of peaks, where each of the peaks can feature a Gaussian-, a Lorentzian- or a Voigt profile. The different profiles respect Doppler broadening, collisional broadening, optical effects and a combination of them. Each of the

profiles is characterized by three parameters. The first is the central position of the Raman peak  $\bar{\nu}_0$ , between  $\bar{\nu} = 0 \text{ cm}^{-1}$  and  $\bar{\nu} = 4200 \text{ cm}^{-1}$ , which covers the entire Raman spectral range of physically existing Raman transitions. The second is the height ( $P_h$ ) of the Raman peaks. The intensity of the Raman peak depends on different factors such as the excitation wavelengths, the laser power, the number of molecules and others. Applying normalization by scaling between 0 and 1, the intensity of the Raman peak can be varied between 0 and 1. The third parameter is the width of the Raman peak. The half-width at half maximum (HWHM) of the Gaussian (which is  $\sigma$ ) and of the Lorentzian (which is  $\gamma$ ) peak profile can be varied to simulated Raman peaks having different widths.

Noise originates from two primary sources; device-associated noise and signal-associated noise. Read out noise and thermal noise are the two device-associated noises and shot noise is signal-associated noise. Read out noise (Gaussian noise) depends on the read-out frequency of the detector, while thermal noise (Poisson noise) occurs from thermally generated electrons in the silicon structure of the CCD detector and correlates to temperature. Shot noise (Poisson noise) arises due to the discrete nature and random arrival times of acquired photons. Thus, noise in raw Raman spectra consists of Poisson noise and Gaussian noise. But above certain noise levels, the Poisson noise can be approximated as Gaussian noise[2, 3]. With this assumption, the noise spectrum  $n(\bar{\nu})$

$$n(\bar{\nu}) = e(\bar{\nu}) \sqrt{n_{ph}^2(\bar{\nu}) + n_{th}^2 + n_{rd}^2} \quad 3.5$$

can be modeled as the summation of shot noise  $n_{ph}$  (also referred to as photon noise), thermal noise  $n_{th}$  and read-out noise  $n_{rd}$ .  $e(\bar{\nu})$  is Gaussian noise having a standard deviation of one and mean of zero. The shot noise

$$n_{ph}(\bar{\nu}) = \sqrt{R(\bar{\nu})} \quad 3.6$$

is the square root of the pure Raman spectrum, and with this a function of the variable  $\bar{\nu}$ . The shot noise is larger at spectral ranges with large signal level, while spectral regions without signal feature no shot noise. This noise level is also dependent on the background signal. Therefore, high background interference also bury the Raman signal in its large shot noise.

The thermal noise

$$n_{th} = \sqrt{B} \quad 3.7$$

is approximated by the square root of the thermal background  $B$ . The thermal background is independent of the signal. Therefore it is supposed to be a constant over the whole spectral range  $\bar{\nu}$ . The read-out noise does also not depend on the signal. Thus it is considered as a constant  $c$  over  $\bar{\nu}$ .

$$n_{rd} = c \quad 3.8$$

Following the above spectroscopic noise simulation approaches, a noise featuring different SNR

$$SNR = 10 \log_{10} \frac{\int_{\bar{\nu}=0cm^{-1}}^{\bar{\nu}=4200cm^{-1}} (R(\bar{\nu}))^2 d\bar{\nu}}{\int_{\bar{\nu}=0cm^{-1}}^{\bar{\nu}=4200cm^{-1}} (n(\bar{\nu}))^2 d\bar{\nu}} \quad 3.9$$

in decibel (dB) can be synthesized.

In this study the background spectrum  $B(\bar{\nu})$  is assumed to be broadband and featureless. Therefore it can be approximated as polynomial, exponential, sinusoidal, Gaussian or sigmoidal distribution or as combination of those functions. Of course, not all background interferences are simply broadband and featureless. For instance surface enhanced Raman spectroscopy (SERS) is known to feature rather complex backgrounds, which

under certain circumstances are composed of several narrowband peaks. However, complex background spectra with narrowband peaks are out of scope of this study. But this does not mean the techniques developed in this study are not applicable for non-varying complex backgrounds.

In order to consider different strength of the background level in the simulated raw Raman spectra, the Raman signal to background ratio

$$SBR = 10 \log_{10} \frac{\int_{\bar{\nu}=0cm^{-1}}^{\bar{\nu}=4200cm^{-1}} R(\bar{\nu}) d\bar{\nu}}{\int_{\bar{\nu}=0cm^{-1}}^{\bar{\nu}=4200cm^{-1}} B(\bar{\nu}) d\bar{\nu}} \quad 3.10$$

is determined. By varying this parameter several background spectra with different strength can be simulated.

### 3.3 Shifted-excitation Raman difference Spectroscopy

Shifted-excitation Raman difference spectroscopy is a powerful method for extracting recovered Raman spectra from heavily fluorescence interfered raw Raman spectra. This technique is based on Kasha's rule, which states that an excited molecule relaxes to the vibrational ground state in the excited electronic state, before the fluorescence emission takes place. This is when the collisional frequency between molecules exceeds the inverse lifetime of the molecule in the excited state, which under the here considered condition is granted. Figure 3.4 (left) shows a sketch of the term-energy of a molecule, which can absorb green excitation radiation and then emit a fluorescence photon. This can be for example a fluorophore compound contained in blood or biological tissue. As it can be seen from the figure, regardless of whether the molecule absorbs the dark green photon (dark green arrow) or the more photon energy containing light green photon (light green photon), according to Kasha's rule the fluorescence emission always occurs from the vibrational ground state  $\nu'_0$ . Incident photons with slightly different photon energies (slightly different wavelengths) can promote the molecule to different vibrational states in

the excited electronic state; however the emission occurs from the vibration ground state of its electronic excited state. This makes the fluorescence emission spectrum theoretically independent of slight changes of excitation wavelength. Figure 3.4 (right) shows a sketch of the term-energy levels of a molecule that undergoes first a Raman Stokes transition when excited with the dark green excitation wavelength and then a Raman Stokes transition when excited with the light green excitation wavelength. In this case, Raman scattered photons shift in energy in proportion to the excitation energy as shown in Figure 3.4 (right). The length of the arrows indicating in Figure 3.4 (right) the Raman scattered photons illustrates this. Of course, the Raman-shift is still the same and not effected by the excitation wavelength. In order to make the photon energy shift visible in a spectrum, one has to either consider the wavelength on the abscissa or, if one prefers to keep the Raman-shift on the abscissa, has to relate the Raman shift for both transition to the same excitation wavelength.

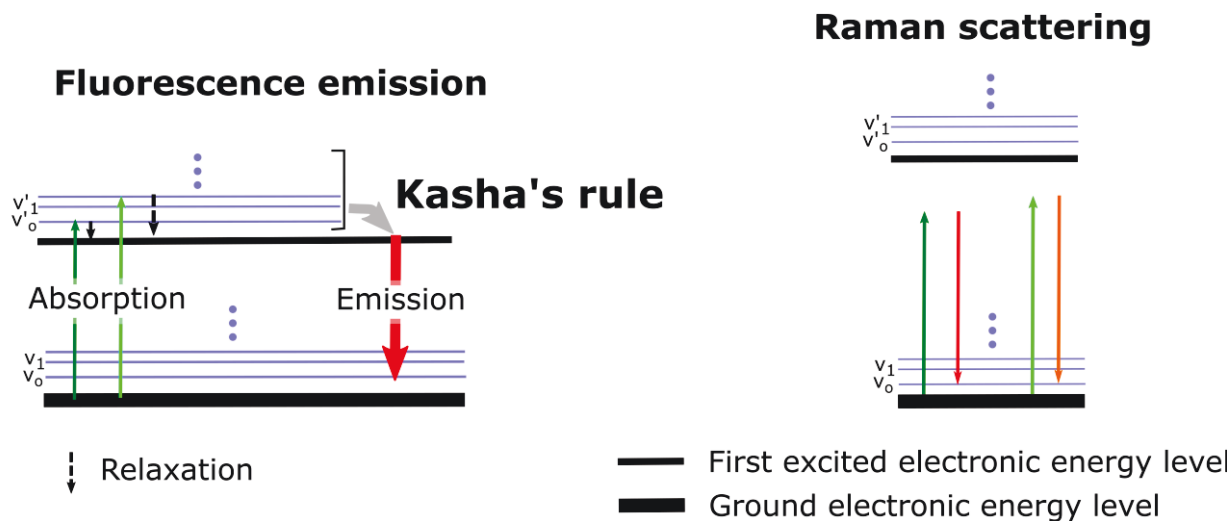


Figure 3.4: (left) shows a sketch of the term-energy levels of a molecule (for example fluorophore molecules in biological tissues) that can be excited to fluoresce with a Raman excitation laser. (right) shows a sketch of the term-energy levels of a molecule that cannot be excited to fluoresce with the same laser.

Figure 3.5 shows as thin solid lines one synthesized raw Raman spectrum  $r_{\lambda}(\bar{\nu})$  excited with the excitation wavelength  $\lambda$  and another raw Raman spectrum  $r_{\lambda+\Delta\lambda}(\bar{\nu})$  excited with slightly different excitation wavelength  $\lambda + \Delta\lambda$ . Due to Kasha's rule, the fluorescence background which is shown as thick gray solid line is identical in both raw Raman spectra whereas the Raman peaks, when related to the same excitation wavelength, are shifted according to the photon energy shift corresponding to  $\Delta\lambda$ . Thus the fluorescence is canceled out in the difference spectrum

$$\Delta r(\bar{\nu}) = r_{\lambda}(\bar{\nu}) - r_{\lambda+\Delta\lambda}(\bar{\nu}) \quad 3.11$$

and the difference spectrum only contains the Raman information. The difference spectrum is shown as dashed blue curve in Figure 3.5. The pure Raman spectrum can then be recovered from this difference spectrum after reconstruction. I will discuss later the reconstruction method and the effect of different excitation wavelength shift  $\Delta\lambda$  on the reconstruction performance.

The charm of SERDS is first that it eliminates fluorescence via an experimental method and based on a physical mechanism. Secondly, this technique removes not only fluorescence but also any other non-varying background interferences such as ambient light, etaloning and others. However, in reality, the background spectra for both excitation wavelengths are not always identical, because of which the SERDS spectrum has to be mathematically treated before reconstruction and also after reconstruction. This is shown in detail in the results section.



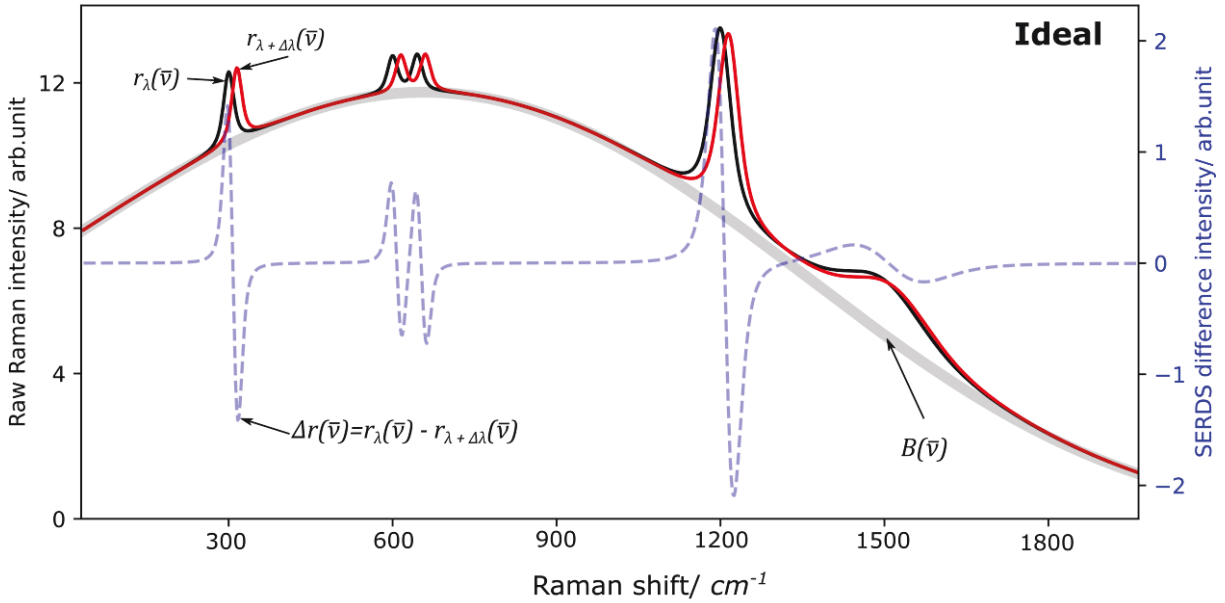


Figure 3.5: Illustration of the SERDS principle. Simulated raw Raman spectrum and its shifted version (only Raman spectrum is shifted) are shown in black and red curves respectively. The common background is shown as thick grey curve. The SERDS difference spectrum is shown as dashed blue curve.

Figure 3.6 shows the working principle of the integration method used to reconstruct the recovered Raman spectrum from the difference spectrum. The difference spectrum computed as per equation 3.11 is shown as dashed blue line. At first, the intensity values of the to-be-reconstructed Raman spectrum  $R_{recov}(\bar{\nu})$  are all initialized to zero, which is shown by the dashed black horizontal line in Figure 3.6. Afterwards, the intensity of  $R_{recov}(\bar{\nu}_i)$  at a specified Raman shift  $\bar{\nu}_i$  is computed by adding the intensity value of the difference spectrum  $\Delta r(\bar{\nu}_i)$  and the intensity value of the updated reconstructed Raman spectrum  $R_{recov}(\bar{\nu}_i - \Delta\bar{\nu})$  shifted by the reconstruction Raman shift  $\Delta\bar{\nu}$

$$\Delta\bar{\nu} = \bar{\nu}_i - \bar{\nu}_{i-1} \quad 3.12$$

to the left (lower Raman shift) from the specified Raman shift  $\bar{\nu}_i$ .

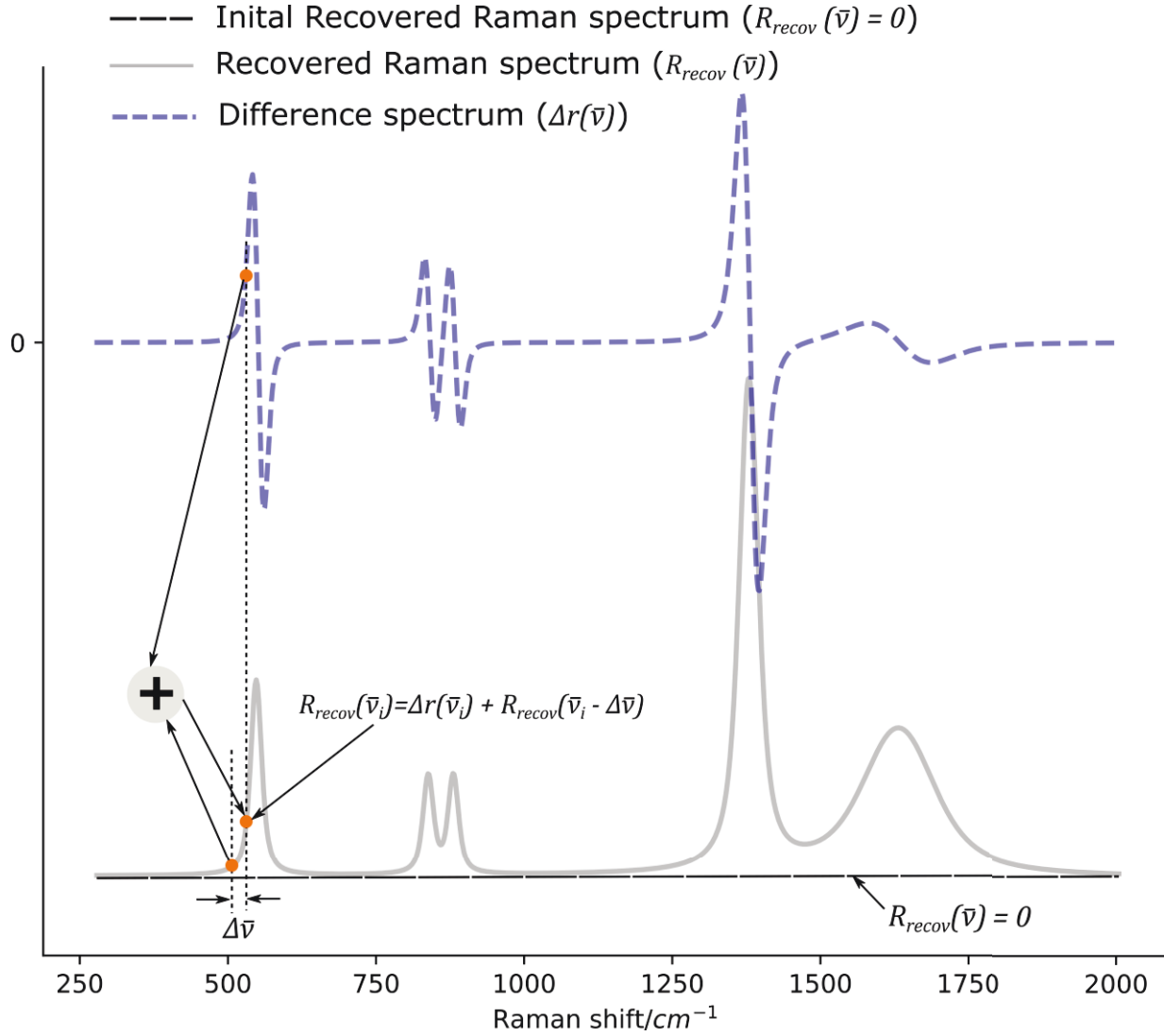


Figure 3.6: Illustration of the working principle of the integration method for the reconstruction of the recovered Raman spectrum (gray curve) from a difference spectrum (dashed blue curve).

This procedure is repeated for all Raman shift positions on the spectral axis from the left to the right side of the spectrum until the whole Raman spectrum is recovered as shown as gray curve in Figure 3.6. Depending on the excitation wavelength shift  $\Delta\lambda$ , this reconstruction by integration method introduces peak shifting and peak widening effects on the recovered Raman spectra.

The excitation wavelength shift  $\Delta\lambda$  is an important SERDS parameter. Care must be taken in selecting this parameter. In order to illustrate the effect of this parameter, a raw Raman spectrum  $r_\lambda(\bar{\nu})$  consisting of only one single Raman peak is considered (see Figure 3.7 black curve).

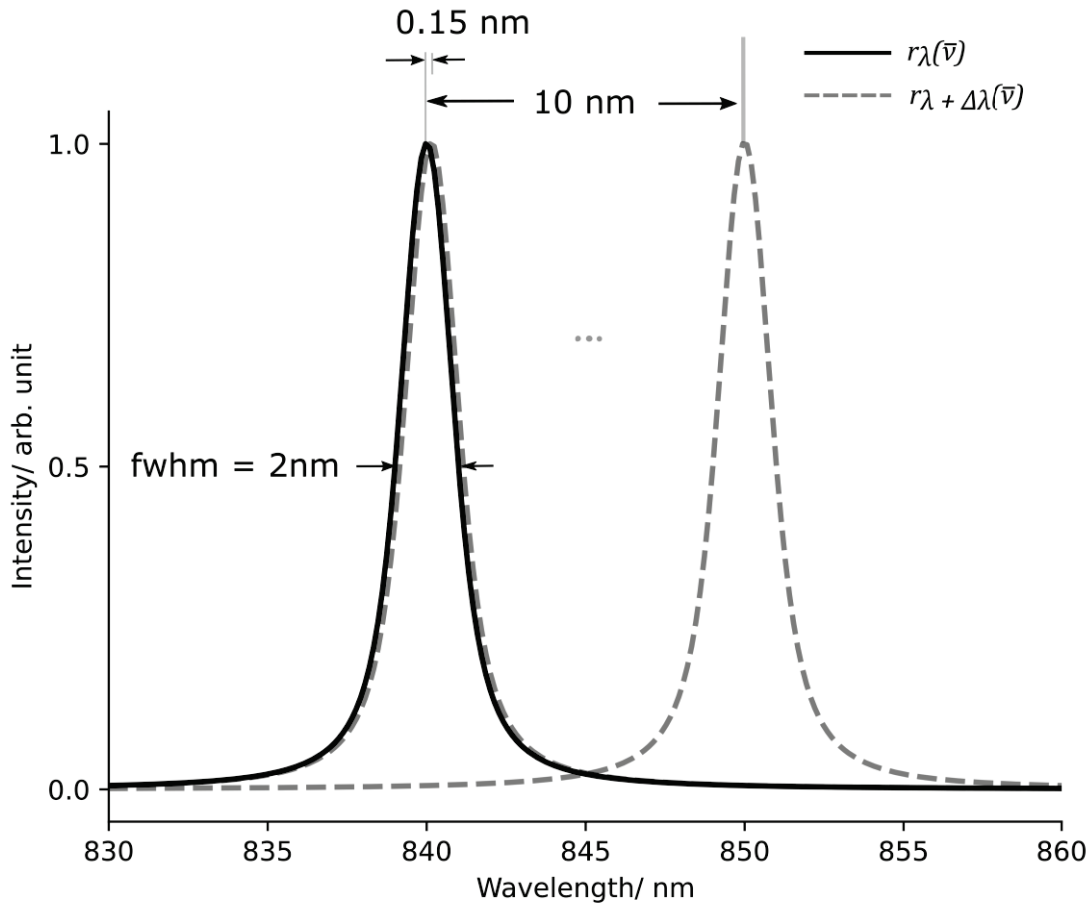


Figure 3.7: Simulated SERDS spectra consisting of only one Raman peak. (black curve) simulated Raman spectrum consists of single Raman peak with 2 nm linewidth. (dashed gray curves) simulated Raman Raman spectra shifted to the right with 0.15 nm (smallest shift) to 10 nm (largest shift)

The peak features a Voigt profile. The Raman peak has a linewidth of 2 nm at full width half maximum (FWHM). The SERDS principle is then imitated by simulating several raw Raman spectra  $r_{\lambda+\Delta\lambda}(\bar{\nu})$  consisting also of single Raman peaks, but excited with a slightly

shifted excitation wavelengths  $\Delta\lambda$ ; therefore the Raman peaks (see Figure 3.7 dashed gray curves) are synthesized to be shifted from the original Raman peak. Only the Raman spectra shifted with the smallest  $\Delta\lambda = 0.15$  nm and largest  $\Delta\lambda = 10$  nm excitation wavelength shift are shown as dashed gray curves in Figure 3.7. In order to determine the effect of the excitation wavelength shift to the signal-to-noise ratio of the difference spectrum, a noise signal was added to the Raman spectra. Then the SERDS difference spectra are computed based on equation 3.11 by subtracting the shifted raw Raman spectra from the original raw Raman spectrum. Figure 3.8 left side (blue dashed curve) shows the signal-to-noise ratio and the relative spectral resolution (black solid curve) of the reconstructed Raman peak as a function of the excitation wavelength shift  $\Delta\lambda$ . The SNR was computed based on equation 3.9 for each difference spectrum relative to the corresponding noise-free difference spectra (before reconstruction). The relative spectral resolution was computed after the reconstruction of the recovered Raman spectra and will be discussed later. As it can be seen from Figure 3.8 (left), the SNR curve reveals that selecting a small excitation wavelength shift  $\Delta\lambda$  corresponds to low SNR of the difference spectrum. On the contrary a small excitation wavelength shift achieves a high spectral resolution of the recovered Raman spectra as shown in Figure 3.8 (left) in black curve. Increasing the excitation wavelength shift, which decreases the peak area overlap in the SERDS raw Raman spectra, improves the SNR of the SERDS difference spectrum, initially, in a linear relationship. Then the slope decreases and approaches a constant value when the SERDS Raman peaks are separated completely at large wavelength shifts. Thus, high excitation wavelength shifts comforts to the improvement of the SNR.

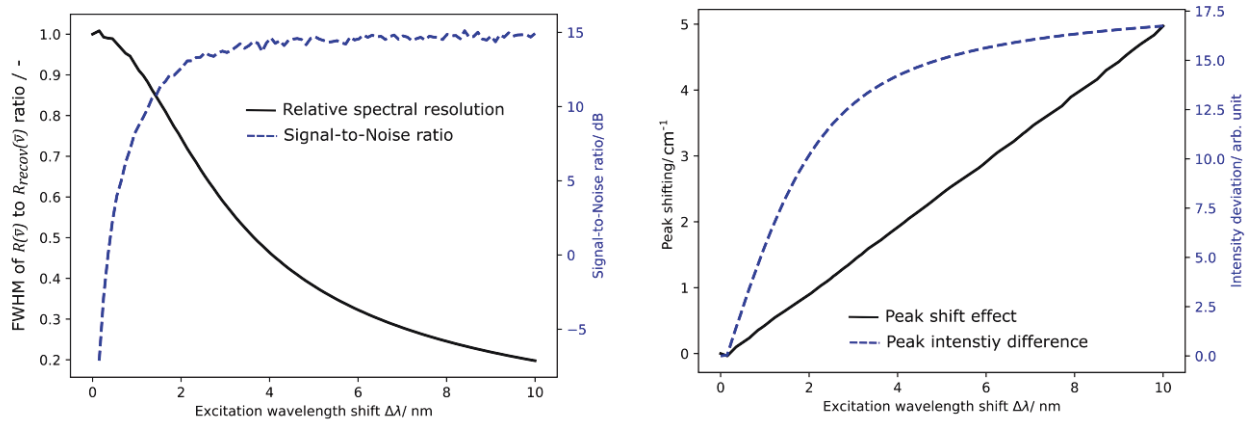


Figure 3.8: (left) plot of signal-to-noise ratios (dashed blue curve) of SERDS difference spectra as a function of wavelength shifts between SERDS raw spectra and plot of spectral resolution of reconstructed Raman spectra (black curve) at different wavelengths shifts between SERDS raw spectra. (right) maximum intensity of reconstructed Raman spectra (dashed blue curve) as a function of wavelength shifts between SERDS raw spectra and peak shift of reconstructed Raman spectra (black curve) with respect to the original Raman spectrum (black curve Figure 3.7) as a function of wavelength shifts between SERDS raw spectra.

The recovered Raman spectra are reconstructed using the integration method from the SERDS difference spectra. And the influence of the  $\Delta\lambda$  on the shape of the recovered Raman spectra is analyzed. For each recovered Raman spectrum the peak shift effect

$$\Delta P_{\text{position}}(\Delta\lambda) = \bar{\nu}_{\max(R_{\text{recov}}(\bar{\nu}))} - \bar{\nu}_{\max(R(\bar{\nu}))} \quad 3.13$$

is computed as the difference between the Raman shift position  $\bar{\nu}_{\max(R(\bar{\nu}))}$  of the peak of the recovered Raman spectra  $R_{\text{recov}}(\bar{\nu})$  and the Raman shift position  $\bar{\nu}_{\max(R(\bar{\nu}))}$  of the peak of the pure Raman spectrum  $R(\bar{\nu})$ . Figure 3.8 right side (black curve) shows the peak shift effect which is varying in a linear fashion with the excitation wavelength shift. Figure 3.8 right side (dashed blue curve) shows the peak intensity difference between the reconstructed Raman spectra and the pure Raman spectrum as a function of excitation wavelength shift. The peak intensity difference  $\Delta P_{\text{intensity}}(\Delta\lambda)$

$$\Delta P_{intensity}(\Delta\lambda) = \max(R_{recov}(\bar{\nu})) - \max(R(\bar{\nu})) \quad 3.14$$

is the maximum intensity of the recovered Raman spectra minus the maximum intensity of the pure Raman spectrum. Figure 3.8 (right) shows as dashed blue curve the effect of excitation wavelength shift  $\Delta\lambda$  on the peak intensity of recovered Raman peak. The excitation wavelength shift also imposes a peak widening effect on the recovered Raman spectrum. This effect is computed as relative spectral resolution. Figure 3.8 left side (black curve) shows the relative spectral resolution of the recovered Raman spectra with respect to the pure Raman spectrum as a function of excitation wavelength shift. The relative spectral resolution  $rP_{width}(\Delta\lambda)$

$$rP_{width}(\Delta\lambda) = \frac{FWHM_{R(\bar{\nu})}}{FWHM_{R_{recov}(\bar{\nu})}} \quad 3.15$$

is computed as the ratio of the linewidth of the pure Raman spectrum to the linewidth of the recovered Raman spectra. The result shows the loss of spectral resolution (widening of recovered Raman peaks) with an increase in excitation wavelength shift.

This shows there is always a compromise in choosing the excitation wavelength shift. A small shift achieves better spectral resolution at the cost of reduced signal strength of broadband Raman bands due to considerable peak area overlap. On the contrary, setting a large wavelength shift results in a reduced spectral resolution and distorted narrow spectral signatures to gain improved signal strength of broader peaks.

### 3.4 Raman experimental setup

In this subsection the general experimental setup of the Raman device, as it has been used with slight modifications throughout the thesis, is described. In order to being able

to cope with the special needs of the specific application, slight modifications had to be realized.

Figure 3.9 shows the experimental setup of the self-engineered Raman sensor.

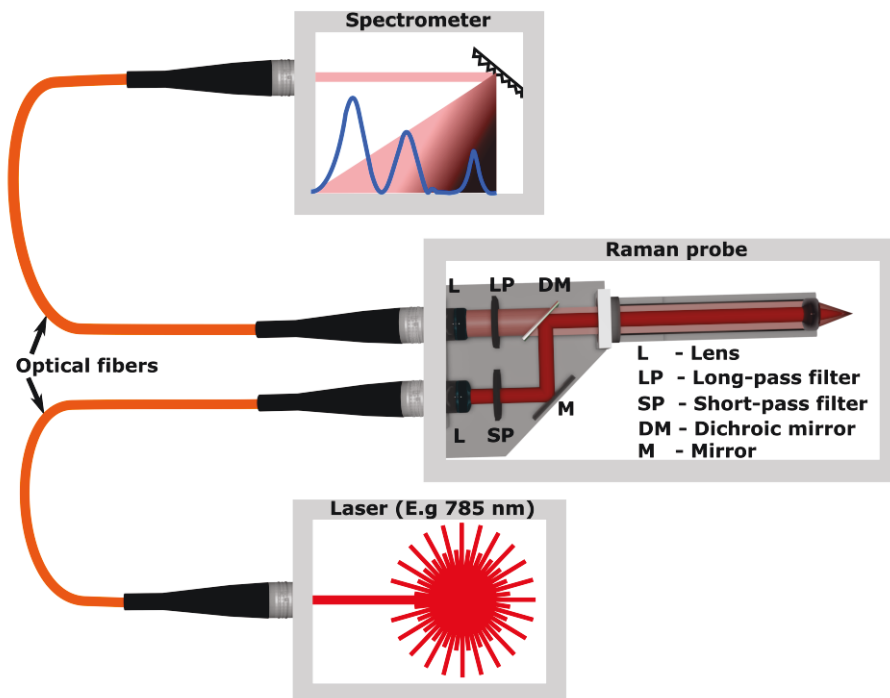


Figure 3.9: Sketch of self-engineered Raman sensor consisting of the Laser source, the spectrometer and the Raman probe containing optical lenses, optical filters and one dichroitic mirror.

A tunable diode laser with a variable emission wavelength between 770 and 810 nm was used as the excitation light source. The excitation beam is launched into a glass fiber, which guides the laser radiation to a Raman probe. A first lens collimates the laser radiation coming out of the excitation fiber. Inside the Raman probe, a short pass filter

(785 nm) suppresses wavelengths longer than 785 nm originating from fiber-light interactions when the excitation light passes through the glass fiber. The excitation laser beam is then reflected via a dichroic mirror, which is highly reflective for the excitation wavelength but transparent for wavelengths longer than 785 nm. It is then focused through a lens onto the sample. A portion of the excited signals (this is mainly elastic light scattering signals, fluorescence and the desired Raman signals) is detected in back-scattering direction through the same lens. The red-shifted fluorescence and Raman signals pass the dichroic mirror towards another lens focusing them onto a detection glass fiber bundle guiding the signals from the Raman probe to the spectrometer. The spectrometer analyzes the collected raw Raman spectra. The elastic light scattering signals are filtered out first by the dichroic mirror reflecting them towards the excitation glass fiber and second by a long pass filter mounted between the dichroic mirror and the signal focusing lens.

During my thesis I used two different lasers as excitation light source and two different spectrometers for signal detection. These main components of the Raman experiments are specified in the Table 3.1 below.

*Table 3.1: Specification of main components of the Raman experiments*

<b>Laser</b>	<b>Manufacturer</b>	<b>Excitation wavelength range</b>	<b>Laser power</b>
Toptica DLpro	Toptica	770 – 810 nm	1.5 W
Cobolt samba	Hübner Photonics	532 nm	500 mW - 1.5 W



<b>Spectrometer</b>	<b>Manufacturer</b>	<b>Detection wavelegnth range</b>	<b>Spectral resolution</b>
Ventana 785	Ocean optics	800 – 940 nm	10 cm <sup>-1</sup>
QEPro 532	Ocean optics	532 – 700 nm	11 cm <sup>-1</sup>

### **3.5 Mathematical method for Raman spectra refinement**

Raman spectra refinement refers to the enhancement of the signal-to-noise ratio (SNR) and signal-to-background ratio (SBR) as given by equations 3.9 and 3.10, respectively. These ratios are directly proportional to the pure Raman signal and inversely proportional to the interferences (noise or background). Therefore, the enhancement of these ratios can be achieved by either the enhancement of the pure Raman signal relative to the background and the noise or the reduction/removal of the interferences. Different experimental methods have been developed based on the enhancement of the pure Raman signal. Mathematical post-measurement processing methods follow the reduction or removal of the interferences to enhance the SNR and the SBR. These mathematical methods are usually applied when the experimental methods for the enhancement of the SBR and the SNR cannot provide the anticipated success or cannot be applied. So far, there is no single mathematical technique which can handle both simultaneously, the enhancement of SNR and SBR. Therefore, the mathematical post-measurement methods are broadly categorized as noise reduction methods or background correction methods. In my thesis I had to compare the results of the post-measurement processing methods that I have developed against state of the art post-measurement processing methods. Therefore, I here provide details on those post-measurement processing methods for noise reduction and background correction, which I have already mentioned in section two of this thesis.

### 3.5.1 Enhancement of the Signal-to-Noise Ratio (SNR) of Raman spectra

Figure 3.10 illustrates an overview of the principle of mathematical methods for noise reduction. A noise-interfered raw Raman spectrum (red in Figure 3.10) is given as an input to the noise reduction algorithm, which is shown in the figure as black rectangle. This algorithm is supposed to recover the Raman spectrum and to eliminate the noise. The fading of the red curve to the right of the noise reduction algorithm is supposed to illustrate that the noise is removed after passing through the noise reduction algorithm whereas the Raman spectrum (green in Figure 3.10) is recovered. It has been shown in Figure 2.2 that

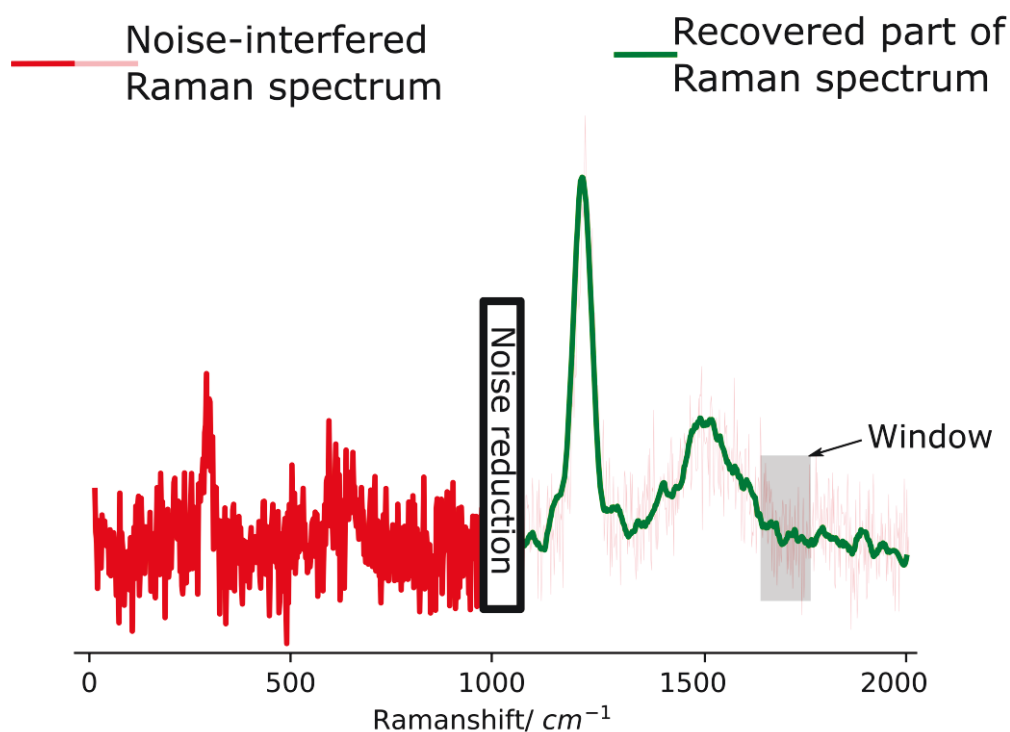


Figure 3.10: Overview of noise reduction. Noise-interfered Raman spectrum is shown partly in red and light red curves. Recovered part of Raman spectrum is shown in green curve.

mathematical post-measurement methods potentially can modify the pure Raman spectrum. Consequently, the recovered Raman spectrum should be taken only as an approximation of the pure (real) Raman spectrum.

Within the grey window shown in Figure 3.10 the recovered Raman spectrum (green curve) is varying slowly whereas the noise is fluctuating randomly (light red curve). The simplest way to smooth out the noise is to replace each data point of the raw Raman spectrum by the average of its neighboring data points within a given moving window (window size defines the number of data points that are averaged). The smoothness of the resulting spectrum is determined by the size of the moving window. But “moving-window-averaging” always reduces peak maxima and also introduces shifting of the peak maxima. The Savitzky-Golay smoother (SG-smoother) [4] does not average within the moving window but fits a polynomial function to the data points within the moving window. Then the center data point of the moving window is updated from the value of the polynomial function at the center point location. Figure 3.11 shows the SG-smoother as applied to de-noise a raw Raman spectrum. The noise-interfered raw Raman spectrum is shown as gray curve. The data points of the raw Raman spectrum within a moving window (light blue rectangle in Figure 3.11) are fitted with a polynomial function of a specified order (a polynomial of third order in Figure 3.11). The value of the to-be-recovered Raman spectrum, which is the center data point of the moving window (green circle in Figure 3.11) is computed from the fit-polynomial function at the Raman shift position of this center data point. Then the light blue window is shifted by one data point to the right and a new polynomial fit of the same order is applied to the data points within the shifted window. This step is not shown in Figure 3.11. The center data point of this shifted window is again updated from the value of the fit-polynomial function evaluated at the Raman shift position of this central data point. In this way all data points of the recovered Raman spectrum function (see red curve in Figure 3.11) are updated one after the other by shifting of the light blue window of data points to the right of the raw Raman spectrum followed by fitting

and evaluating of a polynomial function. The size of the moving window and the polynomial order are adjustable parameters of the SG-smoother to be provided through human intervention. Their effects will be discussed later.

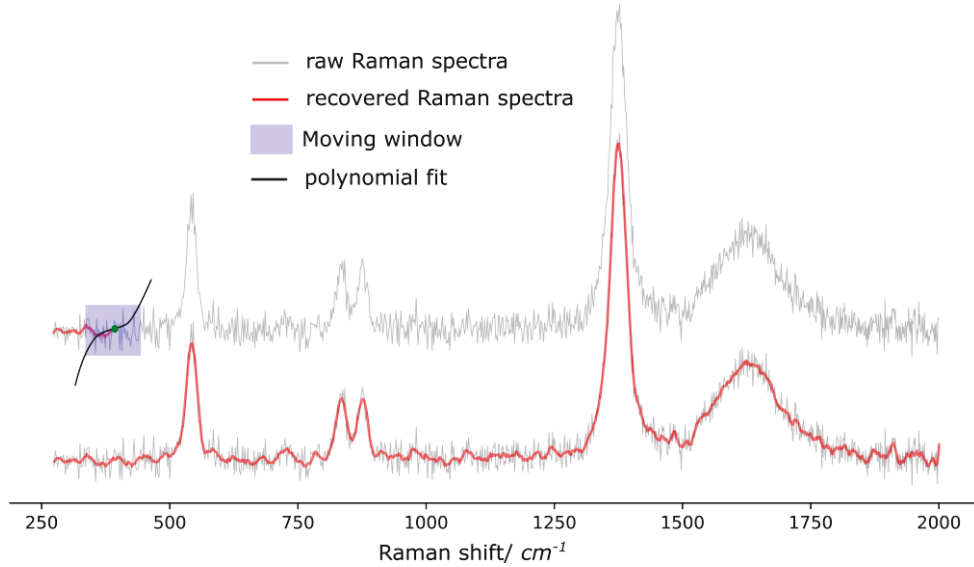


Figure 3.11: Illustration of the SG-smoother. The raw Raman spectra are shown as gray curves. The SG-smoother recovered Raman spectrum is shown as red curve. A third order polynomial function (black curve) is fitted to the data points within the moving window (light blue rectangle).

A perfect smoother (PS-smoother) [5] is an alternative to the SG-smoother, which can be programmed in less than 10 lines of code. Unlike the SG-smoother, the perfect smoother approximates the noise-interfered raw Raman spectrum by penalizing; (1) the fidelity of the recovered Raman spectrum to the noise-interfered raw Raman spectrum and (2) the roughness of the recovered Raman spectrum. The fidelity of the recovered Raman spectrum to the raw Raman spectrum

$$e(\vec{v}) = \sum_{i=1}^N |r(\vec{v}_i) - R_{recov}(\vec{v}_i)|^2 \quad 3.16$$

is the least square of the difference between the raw Raman spectrum  $r(\bar{\nu})$  and the to-be-recovered Raman spectrum  $R_{recov}(\bar{\nu})$ . The roughness of the recovered Raman spectrum

$$s(\bar{\nu}) = \sum_{i=1}^N \Delta^d R_{recov}(\bar{\nu}_i)$$

is obtained from the difference with a degree  $d$  between consecutive data points of the recovered Raman spectrum. For instance

$$\Delta^1 R_{recov}(\bar{\nu}_i) = R_{recov}(\bar{\nu}_i) - R_{recov}(\bar{\nu}_{i-1}) \quad 3.17a$$

$$\Delta^2 R_{recov}(\bar{\nu}_i) = R_{recov}(\bar{\nu}_i) - 2 \cdot R_{recov}(\bar{\nu}_{i-1}) + R_{recov}(\bar{\nu}_{i-2}) \quad 3.17b$$

$$\Delta^3 R_{recov}(\bar{\nu}_i) = R_{recov}(\bar{\nu}_i) - 3 \cdot R_{recov}(\bar{\nu}_{i-1}) + 3 \cdot R_{recov}(\bar{\nu}_{i-2}) - R_{recov}(\bar{\nu}_{i-3}) \quad 3.17c$$

are the computation of the roughness of the recovered Raman spectrum for  $d = 1, 2$ , and  $3$ , respectively.

Then the objective function

$$Q = \min_{R_{recov}(\bar{\nu})} (e(\bar{\nu}) + \lambda \cdot s(\bar{\nu})) \quad 3.18$$

is a balanced combination of the two goals, the fidelity and the roughness, where the smoothness parameter  $\lambda$  is to be adjusted through human intervention. The finally recovered Raman spectrum is the solution which minimizes equation 3.18. The perfect smoother involves two parameters  $d$  and  $\lambda$  to be adjusted through human intervention.

Noise reduction based on wavelets (wavelet-smoother) [6] involves transformation of the noisy raw Raman spectra into a wavelet domain by decomposing it into a set of specified orthonormal wavelet basis functions. The procedure is shown in Figure 3.12. In a first

step, one of the wavelet families (Symlets, daubechies and soon) and a decomposition level ( $1, 2, \dots, \log_2 N$  where  $N$  is the total number of data points of the raw Raman spectrum) have to be specified through human intervention. Then in the second and third steps, the noise is suppressed in the wavelet domain by thresholding the wavelet coefficients which belong to the noise. In these two steps two adjustable parameters, threshold selection rule (Bayes, minmax, universal and others) and the thresholding function (hard, soft, median and others) are required to be specified. Then, in the final step, the not-suppressed coefficients are reversely transformed to obtain the noise-reduced recovered Raman spectrum. The choice of the adjustable parameters plays an important role in the performance of wavelet transformation for noise reduction. The impact of some of these parameters will be discussed using synthesized raw Raman spectra later.

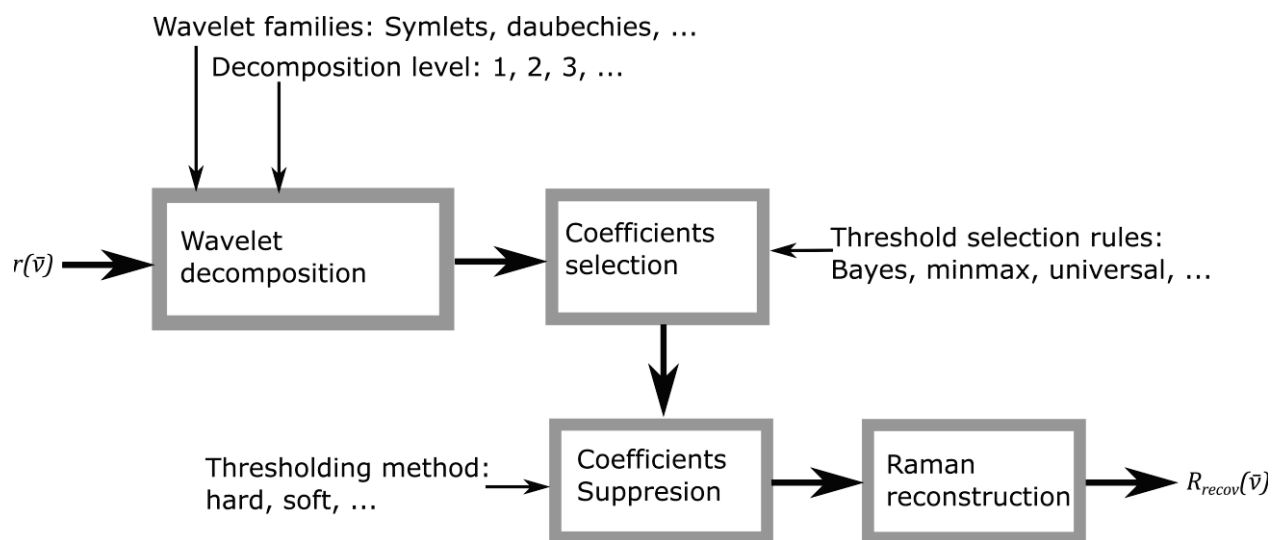


Figure 3.12: Graphical illustration of the procedures of wavelet transformation to denoise raw Raman spectrum.

Figure 3.13 shows the effect of adjustable parameters on the performance of the mathematical methods for noise reduction discussed above. A noise-interfered raw

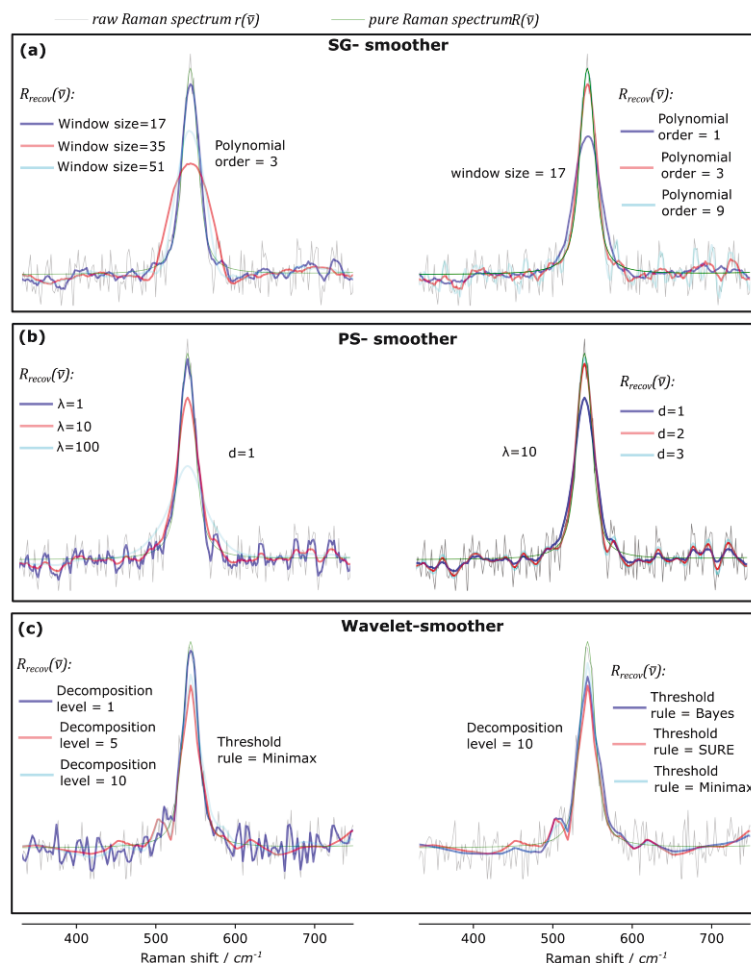


Figure 3.13: Illustration of effect of adjustable parameters on the performance of the mathematical methods for noise reduction. The raw Raman spectrum and the pure Raman spectrum are shown as gray and green curves, respectively. (a) The recovered Raman spectra are shown in blue, red and cyan using the SG-smoother with window sizes of 17, 35 and 51, respectively (left). The recovered Raman spectra are shown in blue, red and cyan using the SG-smoother with polynomial orders of 1, 3 and 9, respectively (right). (b) The recovered Raman spectra are shown in blue, red and cyan using the PS-smoother setting  $\lambda = 1$ ,  $\lambda = 10$  and  $\lambda = 100$ , respectively (left). The recovered Raman spectra are shown in blue, red and cyan using PS-smoother setting  $d = 1$ ,  $d = 2$  and  $d = 3$ , respectively (right). (c) The recovered Raman spectra are shown in blue, red and cyan using the wavelet-smoother with decomposition levels of 1, 5 and 10, respectively (left). The recovered Raman spectra are shown in blue, red and cyan using the wavelet-smoother setting the threshold rule to Bayes, SURE and Minimax respectively (right).

Raman spectrum composed of only one peak is shown as gray curve. The corresponding pure Raman spectrum is shown as green curve. The SG-smoother was applied with three

different moving window sizes and a fixed third order polynomial function to eliminate the noise. Here it is intended to illustrate the effect of varying window size on the performance of the SG-smoother.

Figure 3.13a (left) shows the SG-recovered Raman spectra as blue, red and cyan curves, respectively, when the moving window size was set to 17, 35 and 51 data points and keeping the polynomial order at 3. Increasing the window size achieves smoother recovered Raman spectra but the recovered Raman peak gets significantly distorted. The recovered Raman peak becomes broader with an increase in window size, which means loss of spectral resolution. Figure 3.13a (right) shows the effect of varying the polynomial order of the SG-smoother while keeping the window size fixed at 17. The recovered Raman spectrum obtained by setting the polynomial order to 1, 3, and 9 are shown in blue, red and cyan respectively. A low polynomial order results in distortion of the recovered Raman peak, but a high polynomial order shows an improved preservation of the recovered Raman peak. However, a high polynomial order also features a low SNR enhancement.

In order to analyze the influence of the adjustable parameters of PS-smoother, the raw Raman spectrum of Figure 3.13 was also recovered using three different values of the smoothness parameter  $\lambda$  of the PS-smoother. The result is shown in Figure 3.13b (left) as blue, red and cyan curves, respectively, for  $\lambda = 1$ ,  $\lambda = 10$  and  $\lambda = 100$ . For small values of  $\lambda$ , the recovered Raman spectrum is rough and becomes smoother at larger values of  $\lambda$ . But the PS-smoother with larger  $\lambda$ -values also interacts negatively with the recovered Raman peak. Figure 3.13b (right) shows the effect of the second parameter of PS-smoother. The recovered Raman spectrum obtained by setting  $d = 1$ ,  $d = 2$  and  $d = 3$  of PS-smoother are shown, respectively, as blue, red and cyan curves. Setting  $d = 1$  enhanced the SNR, but at the cost of a trimmed recovered Raman peak.



Figure 3.13c shows the effect of the threshold selection rules and decomposition level to the performance of wavelet transformation for noise reduction. Wavelet-smoother features many parameters but here only the effect of these two parameters is discussed. But the other parameters as well have similar effects on the performance of the wavelet-smoother. Figure 3.13c (left) shows the recovered Raman spectra as blue, red and cyan curves when the decomposition level was set to 1, 5 and 10, respectively. Choice of a small decomposition level achieves noisier recovered Raman spectra. Setting a large decomposition level achieves smoother recovered Raman spectra, but it negatively affects the recovered Raman peak. Figure 3.13c (right) shows the recovered Raman spectrum as blue, red and cyan curves obtained using wavelet-smoother with three threshold selection rules, such as Bayes, SURE and minimax, respectively. The three parameters achieve comparable smooth recovered Raman spectra. But setting the threshold selection rule to SURE provides the worst recovered Raman spectrum.

### **3.5.2 Enhancement of Signal-to-Background Ratio (SBR) of Raman spectra**

Figure 3.14 shows the general working principle of a background correction method for the enhancement of the SBR. The simulated raw Raman spectrum is shown partly as red and partly as light red curve. The noise-free background spectrum is shown as gray curve. The background correction algorithm, which is illustrated in Figure 3.14 as black rectangle, can be viewed as peeling-off the background from the raw Raman spectrum and extract the pure Raman spectrum. The part of the recovered Raman spectrum after removal of the background is shown to the right of the background correction method as green curve.

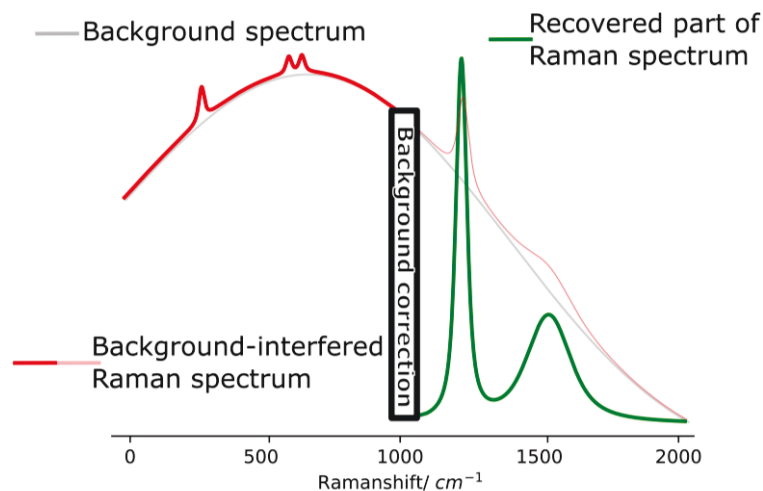


Figure 3.14: General working principle of background correction methods

Common background correction methods involve polynomial fitting and penalized least squares fitting to approximate the background spectrum with an assumption that the background spectrum is a broadband curve. The modified polynomial fit method (ModPoly) is one of these methods, which applies iterative polynomial fitting to approximate the background spectrum. Figure 3.15 shows graphically the working principle of ModPoly. Simulated raw Raman spectrum and noise-free background spectrum are shown in light blue and thick light red curves. During the first iteration a polynomial of specified order is applied to fit the raw Raman spectrum. This first approximation of the recovered background spectrum is shown in Figure 3.15 as dashed blue curve.

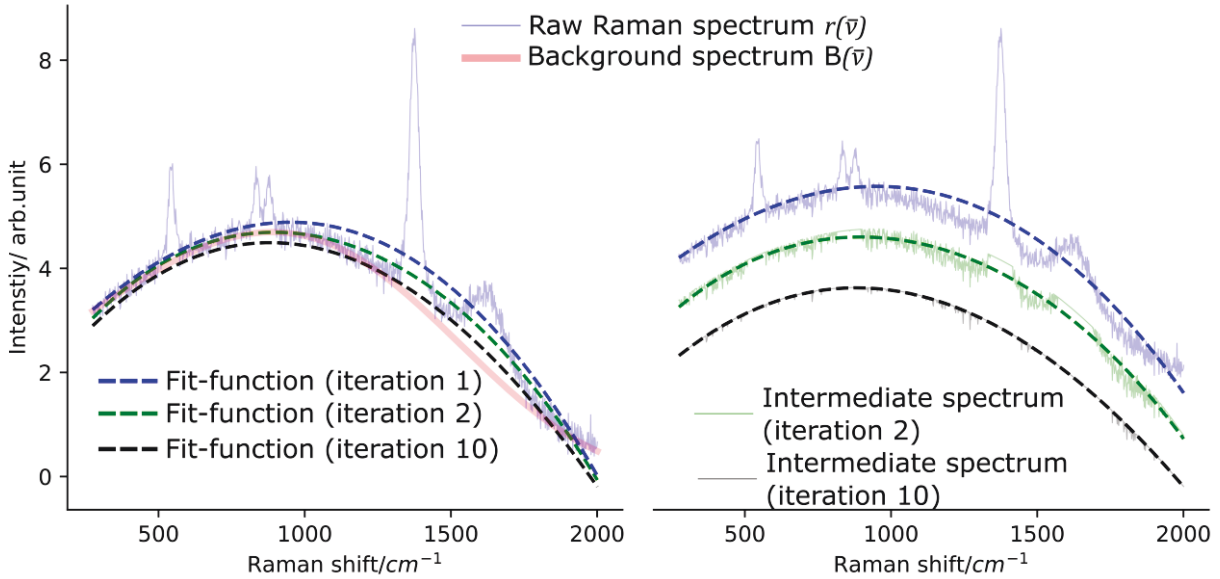


Figure 3.15: Illustration of the ModPoly method. The raw Raman spectrum and the real background spectrum are shown as light blue and as thick light red curves. The recovered background spectra at iteration 1, iteration 2 and iteration 10 are shown as dashed blue, dashed green and dashed black curves. Intermediate spectra with trimmed Raman peaks at iteration 2 and iteration 10 are shown as light green and light black curves

Then the recovered background spectrum  $B_{recov}(\bar{\nu})$  is compared with the raw Raman spectrum along the whole spectral range. Afterwards, at each Raman shift position the intensity value of raw Raman spectrum is replaced by the value of the recovered background spectrum whenever it exceeds the intensity value of the recovered background spectrum. This creates a new intermediate spectrum with trimmed peaks as shown in Figure 3.15 as light green curve on the right. In a second iteration, a new polynomial function is fitted to the newly created intermediate spectrum. This second approximation of the recovered background spectrum is shown in Figure 3.15 as dashed green curve on the right. Then, the light green intermediate spectrum is compared with this dashed green recovered background spectrum. Again, whenever the light green intermediate spectrum is above the dashed green recovered background spectrum, it is replaced by the intensity value of the dashed green recovered background spectrum. This

procedure repeats for further iterations. The recovered background spectrum after iteration 10 is shown as dashed black curve in Figure 3.15 on the left. With every iteration, the recovered background spectrum  $B_{recov}(\bar{\nu})$  drifts towards approximating the background spectrum  $B(\bar{\nu})$ . This method systematically excludes the Raman peaks in every iteration, while the recovered background spectrum approximates the background spectrum. In this method the polynomial order has to be adjusted through human intervention.

Methods based on penalized least squares modify the perfect smoother to be applied to approximate the background spectrum. Asymmetric least squares smoothing (ALS) is one of these methods. This method includes only a weighting factor to the fidelity (equation 3.16)

$$e(\bar{\nu}) = \sum_{i=1}^N w_i \cdot |r(\bar{\nu}_i) - B_{recov}(\bar{\nu}_i)|^2 \quad 3.19$$

of the fit-function (to-be-recovered background spectrum)  $B_{recov}(\bar{\nu})$  to the raw Raman spectrum  $r(\bar{\nu})$ . The weighting factor

$$w_i = \begin{cases} p & \text{if } r(\bar{\nu}_i) > B_{recov}(\bar{\nu}_i) \\ 1 - p & \text{if } r(\bar{\nu}_i) < B_{recov}(\bar{\nu}_i) \end{cases} \quad 3.20$$

is computed by introducing a new parameter  $p$  to be adjusted through human intervention. The main idea of introducing the weighting factor is to treat peak regions and peak-free regions of the raw Raman spectra differently, thus systematically only fitting to the background spectrum of the raw Raman spectrum. Figure 3.16 illustrates the principle of ALS.

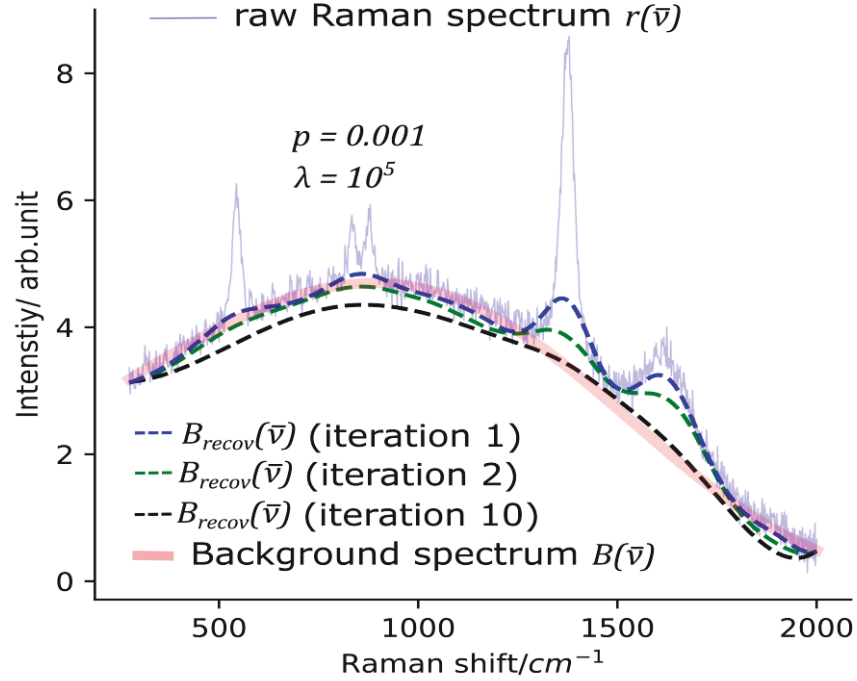


Figure 3.16: Illustration of the ALS method. The raw Raman spectrum and the background spectrum are shown as light blue and thick light red curves. Recovered background spectra at iteration 1, iteration 2 and iteration 10 are shown as dashed blue, dashed green and dashed black curves.

The raw Raman spectrum and the noise-free background spectrum are shown as light blue and thick light red curves, respectively. In the first iteration the recovered background spectrum  $B_{recov}(\bar{\nu})$  is computed by solving equation 3.18 which is now a function of also the weighting factor  $w_i$  since the fidelity equation (equation 3.19) is modified. This first approximation of the background spectrum is shown by the dashed blue curve in Figure 3.16. Then the weighting factor ( $p = 0.001$ ) is updated based on equation 3.20. Here the recovered background spectrum is compared with the raw Raman spectrum along the whole spectral range. Whenever the intensity value of the raw Raman spectrum exceeds the intensity value of the recovered background spectrum, a small weighting factor is assigned. And whenever the intensity value of recovered background spectrum exceeds the intensity value of the raw Raman spectrum, a larger weighting factor is assigned. In a

second iteration, a new recovered background spectrum is estimated again using equation 3.18 but with the updated weighting factor. The recovered background spectrum in iteration two is shown as dashed green curve in Figure 3.16. In the next iteration, again the weighting factor is updated by comparing the new recovered background spectrum and the raw Raman spectrum. This fitting and weight updating continues until no difference is detected between the consecutively recovered background spectra. The recovered background spectrum at iteration 10 is shown as dashed black curve in Figure 3.16. Thus, as it can be seen in Figure 3.16, after every iteration the recovered background spectrum drifts towards approximating the background spectrum. This method involves two adjustable parameters to be provided through human intervention.

An adaptive iteratively reweighted penalized least squares method (airPLS) [7] was also proposed as an improvement of the ALS method. Its difference from ALS is that airPLS assigns the weighting factors automatically according to the difference between the raw Raman spectrum and the recovered background spectrum in every iteration. Thus, it excludes the weighting parameter  $p$  of the ALS method. Therefore, only one parameter  $\lambda$  needs to be adjusted. The ALS and airPLS methods assign large weights whenever the recovered background spectrum is above the raw Raman spectrum and assign small weights whenever the recovered background spectrum is below the spectrum of the previous iteration. However, noise is distributed above the background spectrum as well as below it. The airPLS method therefore inherently underestimates the real background. Therefore the recovered background after iteration 10 in the figure above is mainly below the real background.

In order to take this into account, an asymmetrically reweighted penalized least squares smoothing (arPLS) was proposed by Baek et al. [8]. This method includes an estimation of the noise level in every iteration and adjusts the weights accordingly.

Figure 3.17 shows the effect of adjustable parameters on the performance of the ModPoly and ALS background correction methods. Two raw Raman spectra are simulated. One with a simple (left in Figure 3.17) and one with a complex background (right in Figure 3.17). Figure 3.17a shows the recovered background spectra in dashed black and dashed green curves which are obtained when using the ModPoly method by setting the polynomial order to 4 and 10, respectively. The choice of high polynomial order increases the flexibility of the recovered background spectrum. Figure 3.17a (right) shows that the ModPoly with a polynomial order of 10 approximates the complex background spectrum better than a polynomial order of 4. But this flexibility also implies that the recovered background spectrum also takes broad Raman bands as background. Figure 3.17b shows the recovered background spectra in dashed black and dashed green curves obtained using the ALS method by setting the smoothness parameter to  $\lambda = 10^4$  and  $\lambda = 10^5$  respectively. In both cases the weighting parameter  $p$  is kept constant ( $p = 0.001$ ). As it can be seen from Figure 3.17b (left)  $\lambda = 10^4$  achieves a better estimation of the background spectrum under the narrow Raman peaks. However, the estimated background signal interacts with the Raman spectrum, especially in the region of broad Raman peaks/bands. Increasing  $\lambda$  by one order of magnitude improves the estimation of the recovered background under the broad peaks/bands, but the estimated background deviates in the region of the narrow Raman peaks.

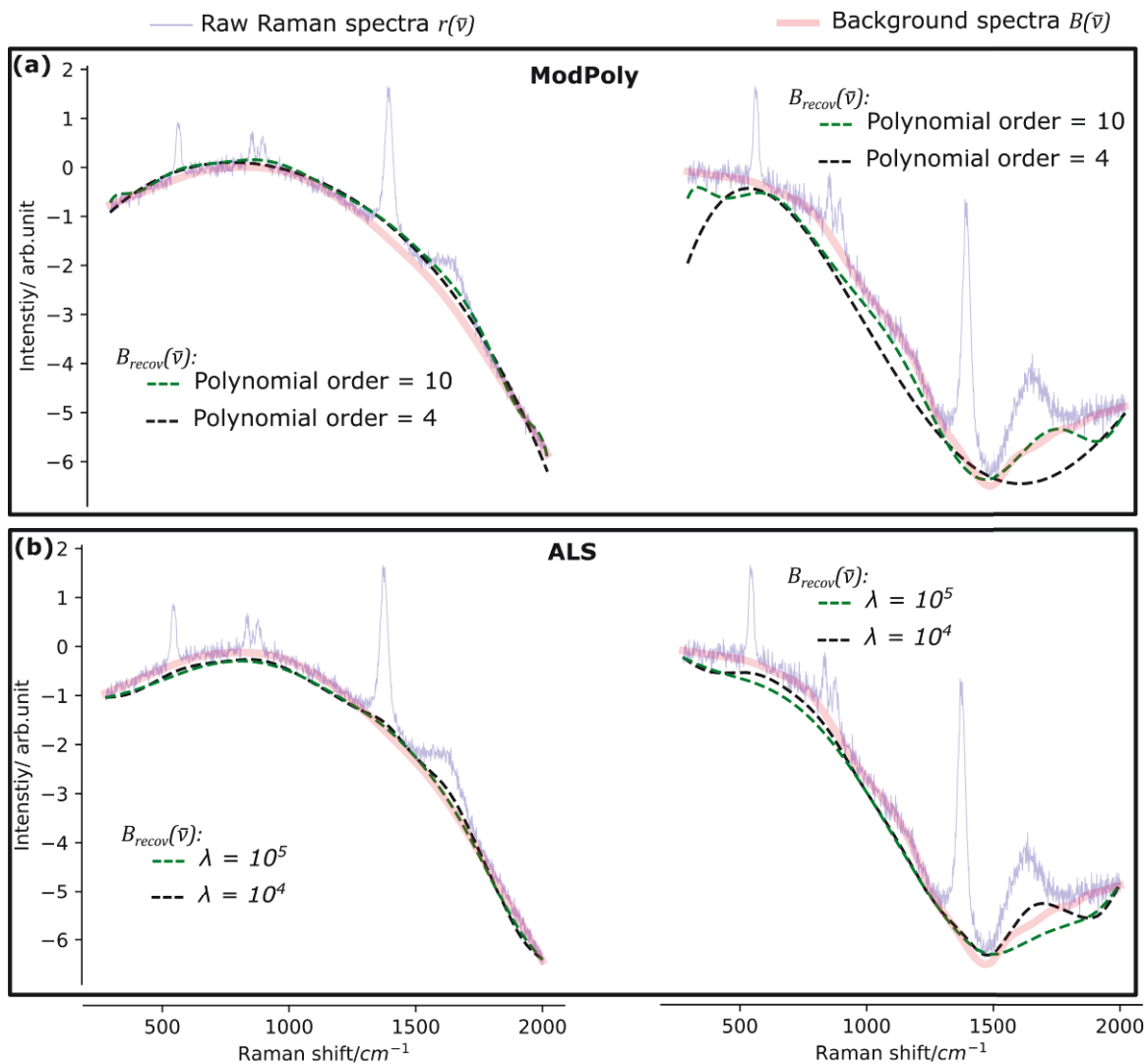


Figure 3.17: Illustration of the effect of adjustable parameters on the performance of the mathematical methods for background correction. Two raw Raman spectra are shown as light blue curves simulated with a simple background spectrum (left column) and complex background spectrum (right column). The background spectra are shown as thick light red curves. (a) Recovered background spectra  $B_{\text{recov}}(\bar{\nu})$  are shown as dashed black and green curves obtained using Modpoly by setting the polynomial order to 4 and 10, respectively. (b) Recovered background spectra  $B_{\text{recov}}(\bar{\nu})$  are shown as dashed black and green curves obtained using ALS by setting  $\lambda = 10^4$  and  $\lambda = 10^5$ , respectively.

For both background correction methods, a parameter, which works to approximate one background spectrum, may not work for a background spectrum of different shape. This



requires repetitive human interventions in case the raw Raman spectra feature different background spectra. In Figure 3.17, both methods work better to approximate the simple background spectrum but fail to approximate the complex background spectrum. In the results section it will be shown that dividing the raw Raman spectra into sub-regions and applying ALS to the sub-regions separately, improves the approximation of complex backgrounds.

### 3.6 Deep neural networks

This section aims at providing a brief introduction into deep neural networks. More details of the field can be found elsewhere [9, 10]. The main focus will be on convolutional neural networks (CNN) and U-Net deep neural network architectures, which are relevant for this thesis.

Machine learning (ML) is all about learning meaningful representations of the input data that help to make sense of the data and map to the corresponding target. ML learns the meaningful representations from exposure to known pairs of input data and corresponding targets. Figure 3.18 briefly illustrates this principle. Input data and corresponding answers are provided to the ML technique as an input. Based on this information, the technique trains a model which maps the input data to the answers. This trained model is then utilized to predict an answer for unknown new input data.

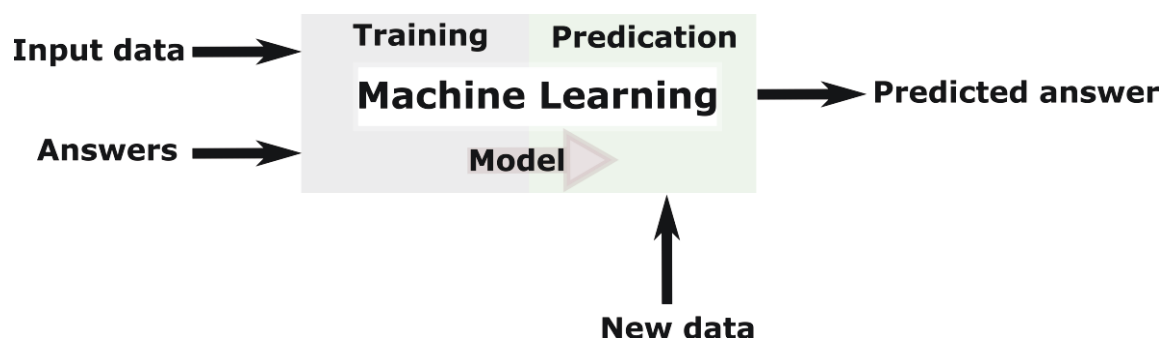


Figure 3.18: Illustration of the principle of machine learning (ML)

Deep learning is a branch of machine learning that involves multiple levels of data representations to map the input data to their corresponding final target [9]. It requires large amounts of data to imitate complex interrelationship, whose inputs and outputs are far apart. One example is highly fluorescence interfered raw Raman spectra as input data and as output purified Raman spectra. Another example are raw Raman spectra as input and as output the malignant or non-malignant categorization in cancer diagnostics. In order to accomplish this success, it uses mathematical entities named neural networks, which are capable of representing complicated interrelationships functions through a composition of linear relationships.

Figure 3.19 illustrates in a nutshell, how deep neural networks work in the case of the refinement of Raman spectra. Raw Raman spectra  $r(\bar{\nu})$  were considered as an input to the input layer of the network. In Figure 3.19 only one raw Raman spectrum is shown as input. Following the input layer, one hidden layer consists of  $m$  nodes  $a_m$ .

For simplicity of description, only two hidden layers are considered, but depending on the problem at hand, several hidden layers can be required to implement a deep neural network. The value of each node is computed in a two-step procedure. First, its value is computed through a linear combination of all its input data (data left of the node). For instance, the value of the first node ( $a_1$ ) of the hidden layer

$$a_1 = w_{11} \cdot r(\bar{\nu}_1) + w_{12} \cdot r(\bar{\nu}_2) + \dots + w_{1n} \cdot r(\bar{\nu}_n) + b_1 \quad 3.21$$

is equal to the weighted sum of the raw Raman spectrum plus a bias  $b$ .

Second the value of the node passes through non-linear function ( $f(\cdot)$ ) which is called activation function. Thus, the values of all the nodes  $a_m$  of one hidden layer in matrix form can be defined as

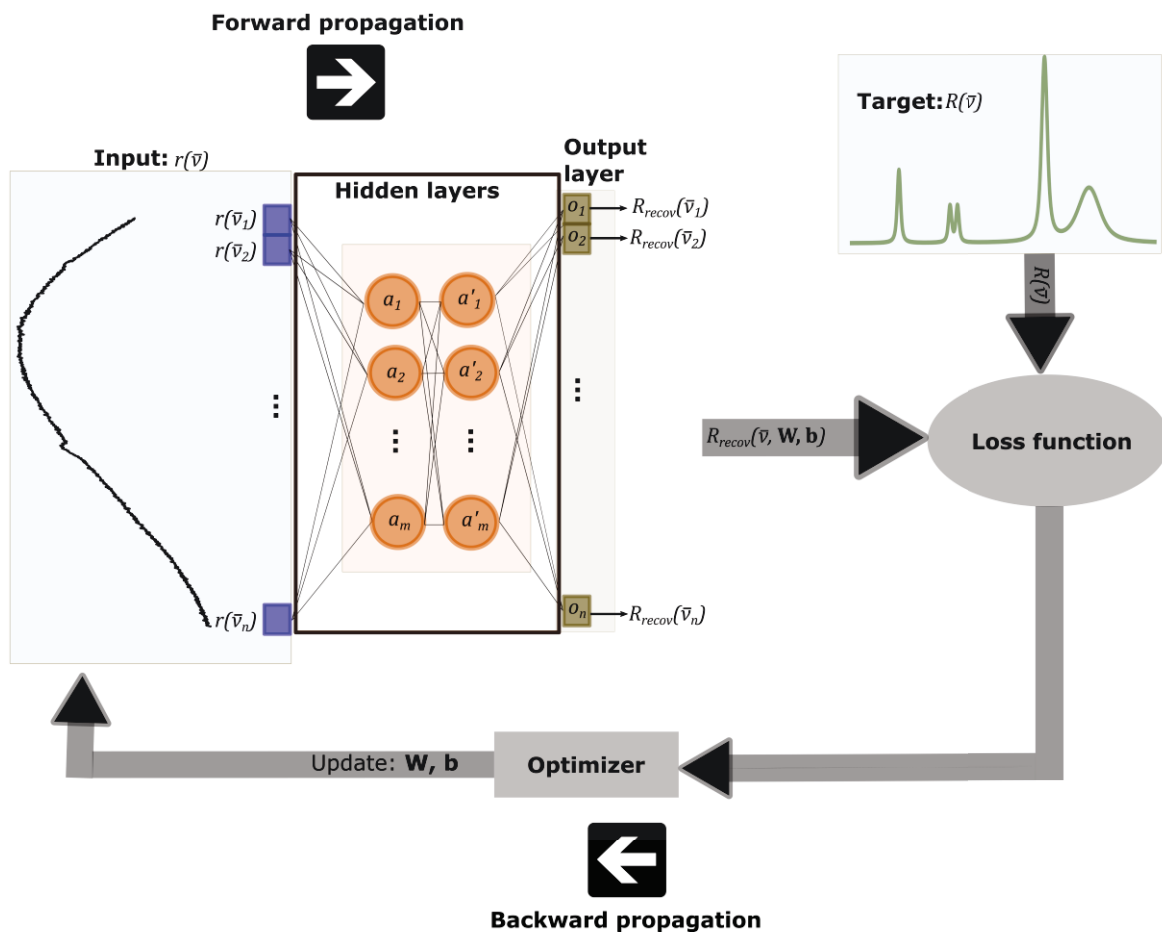


Figure 3.19: Illustration of the principle of neural networks in a nutshell

$$\mathbf{a} = \mathbf{f}(\mathbf{W}^T \cdot r(\bar{v}) + \mathbf{b}) \quad 3.22$$

with the weighting matrix  $\mathbf{W}$

$$\mathbf{W} = \begin{bmatrix} w_{11} & \cdots & w_{m1} \\ \vdots & \ddots & \vdots \\ w_{1n} & \cdots & w_{mn} \end{bmatrix} \quad 3.23$$

and the bias matrix  $\mathbf{b}$

$$\mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix} \quad 3.24$$

The parameters contained in the weighting matrix  $\mathbf{W}$  and contained in the bias matrix  $\mathbf{b}$  are trainable parameters, which have to be adjusted through optimization by exposing the neural network to known pairs of raw Raman spectra and pure Raman spectra.

The non-linear activation function enables neural networks to learn a complex pattern of input data [9]. At present, the rectified linear unit (ReLU) non-linear activation function

$$\mathbf{f}(\mathbf{W}^T \cdot r(\bar{\nu}) + \mathbf{b}) = \max(\mathbf{W}^T \cdot r(\bar{\nu}) + \mathbf{b}, \quad 0) \quad 3.25$$

is a good default for all hidden layers and it is simply equal to the half-wave rectifier.

Going from one layer to the next involves computation of a weighted sum of the inputs from the previous layer plus a bias and pass the result through a non-linear function. Therefore, the value of the nodes in the output layer (see Figure 3.19)

$$\mathbf{o} = f(\mathbf{W}'^T \cdot \mathbf{a}' + \mathbf{b}') \quad 3.26$$

is the weighted sum of the values of the nodes of the last hidden layer added to a bias and then activated through an activation function. Depending on the task, the activation function can be defined differently in the output layer. For instance, sigmoid functions are used for binary classification problems, whereas linear or none can be used for refinement of Raman spectra (regression task). Finally the recovered Raman spectra

$$R_{recov}(\bar{v}) = \mathbf{o}$$

is obtained from the output layer and it is a function of the weights and bias as shown in Figure 3.19. The computation of the output from the input in a cascade is called forward propagation. The details of backward propagation, loss function and optimization which are shown in Figure 3.19 are discussed in the following paragraphs.

Figure 3.20 illustrates a convolutional neural network (CNN). CNN consists of two layers; convolutional layers and pooling layers in addition to the fully-connected layers that are shown in Figure 3.19. Figure 3.20 (light green box) shows the convolutional layer operation considering a kernel with three elements. The convolutional layer focuses on using learnable kernel  $\mathbf{K}$  that stride through the input data, perform element-wise multiplication and averaging, and pass the result through a non-linear activation, commonly the ReLU. Its role is to reduce the input data into a more comfortable form and at the same time to not lose the input features that are critical for getting a good prediction. The kernel size is referred to as hyper-parameter, as it is not a learning parameter of the neural network, but its elements are trainable parameters to be adjusted through optimization like the weight and bias parameters. Figure 3.20 (light red box) shows pooling operation for a pool size of two; however, the pool size is a hyper-parameter that can be tuned. The pooling operation involves gliding through the input data (the input to the pooling layer is the output of the convolutional layer) by taking average- or maximum of a specified number of input data points. The pooling layer is utilized for the reduction of the dimensionality of the representations and thus reduces the computational complexity of the CNNs [9]. In the specific example presented in Figure 3.20 after the pooling operation the number of input data points is reduced by half. CNNs usually consist of fully-connected layers as it is shown in Figure 3.19. They are finally responsible for mapping the extracted features of the convolutional layers to the predicted output.

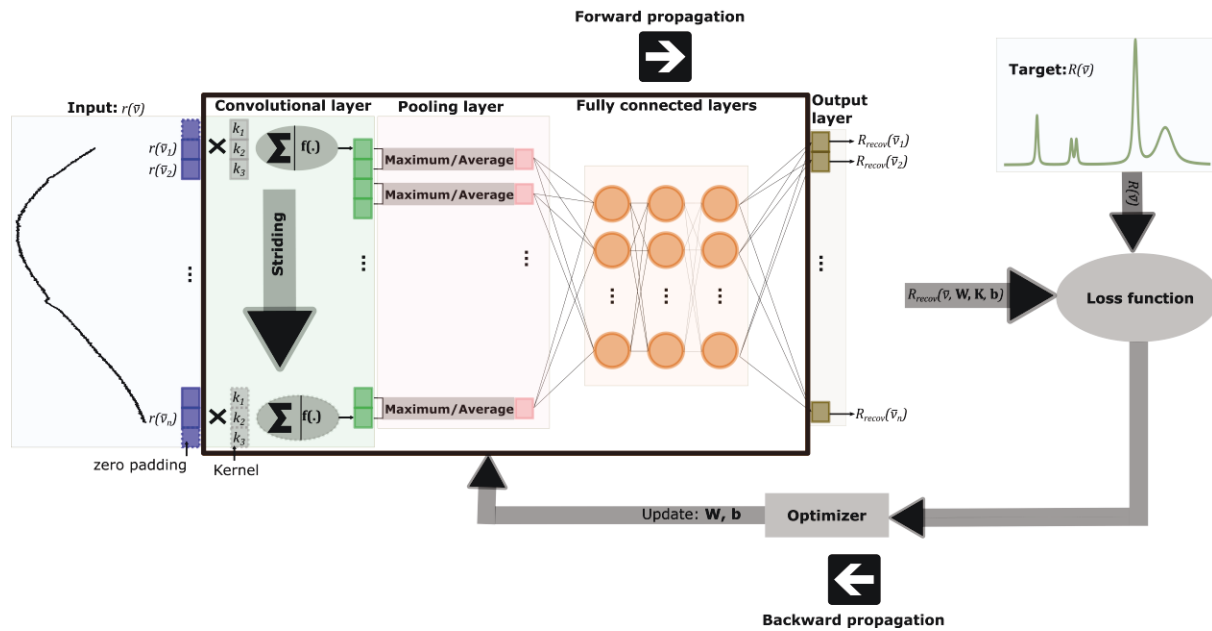


Figure 3.20: Graphical illustration of convolutional neural networks

The parameters

$$\theta = [\mathbf{W}, \mathbf{K}, \mathbf{b}] \quad 3.27$$

which are called weights, kernels and bias, respectively, are adjustable/learnable parameters of the convolutional neural network which can be envisioned as a steering wheel that drives, in the specific example shown in Figure 3.20, the raw Raman spectra towards the pure Raman spectra. A typical deep learning system can have hundreds of millions of these adjustable parameters. Proper adjustment of these parameters  $\theta$  requires computation of a loss function (objective function)  $L(\theta)$  that evaluates the error between the predicted output  $R_{recon}(\bar{\nu}, \mathbf{W}, \mathbf{K}, \mathbf{b})$  and true targets  $R(\bar{\nu})$ . A mean squared error (MSE) loss function

$$L(W, K, b) = \sum_{i=1}^N |R(\bar{v}_i) - R_{recov}(\bar{v}_i)|^2 \quad 3.28$$

is usually utilized for regression learning tasks such as the refinement of Raman spectra. The objective function can be envisioned as a hilly landscape in the high-dimensional parameter space of the neural network. The gradient of this objective function indicates the direction of steepest descent towards the minimum of the error between the predicted output (recovered Raman spectra) and target output (pure Raman spectrum). The computation of the gradient applies the chain rule for derivatives, where differentiation is repeatedly applied through all the layers starting from the output layer all the way to the input. This is called backward propagation. Once the gradients are computed, it is straightforward to update the parameters  $\theta^i$

$$\theta^i = \theta^{i-1} - \eta \frac{\partial L(\theta)}{\partial \theta} \quad 3.29$$

as a function of their current values  $\theta^{i-1}$ , following the steepest descent by a specified increment called learning rate  $\eta$  where  $i$  is the number of iterations also called number of epochs. This optimization approach is called gradient descent optimizer and it iteratively updates the parameters for a given number of iterations (epochs) which is the number of forward and backward propagation passes through the network to be set by the user. The learning rate is the most critical hyper-parameter to be tuned during the training of deep neural networks. Small learning rates will make the learning process converge slowly, whereas large learning rates could fail to converge. One usually starts with a learning rate value and then decreases it monotonically during learning [9]. Instead of reducing the learning rate during training, Smith et al [11] also suggested varying the learning rate cyclically as a function of epoch between reasonable boundary values. Besides the learning rate, there are also many hyper-parameters to tweak, such as the number of

hidden layers and the number of nodes per layer, number of convolutional layers and kernel size, and others.

With the large number of adjustable parameters, a CNN has an incredible amount of freedom to fit a wide variety of complex datasets. However, this great flexibility means it is also prone to overfitting a scenario, where the model performs only well on the training dataset but fails to generalize. Therefore, after each epoch, the performance of the system is always measured on a different dataset (validation dataset) not included in the training dataset. This helps monitor the learning process and interrupts the training when the model starts to overfit. *Dropout* is a popular regularization approach to tackle overfitting in deep neural networks. It is a fairly simple algorithm such that at every training epoch some of the nodes are entirely ignored based on a given dropout percentage.

Based on convolutional neural networks, there are several deep neural network architectures. U-Net is one of these architectures originally proposed by Ronneberger [12] for medical image segmentation. “U” is not an acronym here, it has to be understood like in the expression U-turn, where one reverses the direction. Figure 3.21 shows the architecture of the U-Net as applied to refinement of Raman spectra. It applies a network of convolutional layers to perform the prediction of the desired output from noisy and background interfered raw Raman spectra. The network architecture consists of an encoder network followed by a decoder. The encoder takes a raw Raman spectrum as an input. It extracts spatial features from the spectra, thus yielding a multilevel feature representation of the information to map the input to the desired output. The decoder then reconstructs the output, which has the same pixel space as the input, using the extracted information of the encoder.



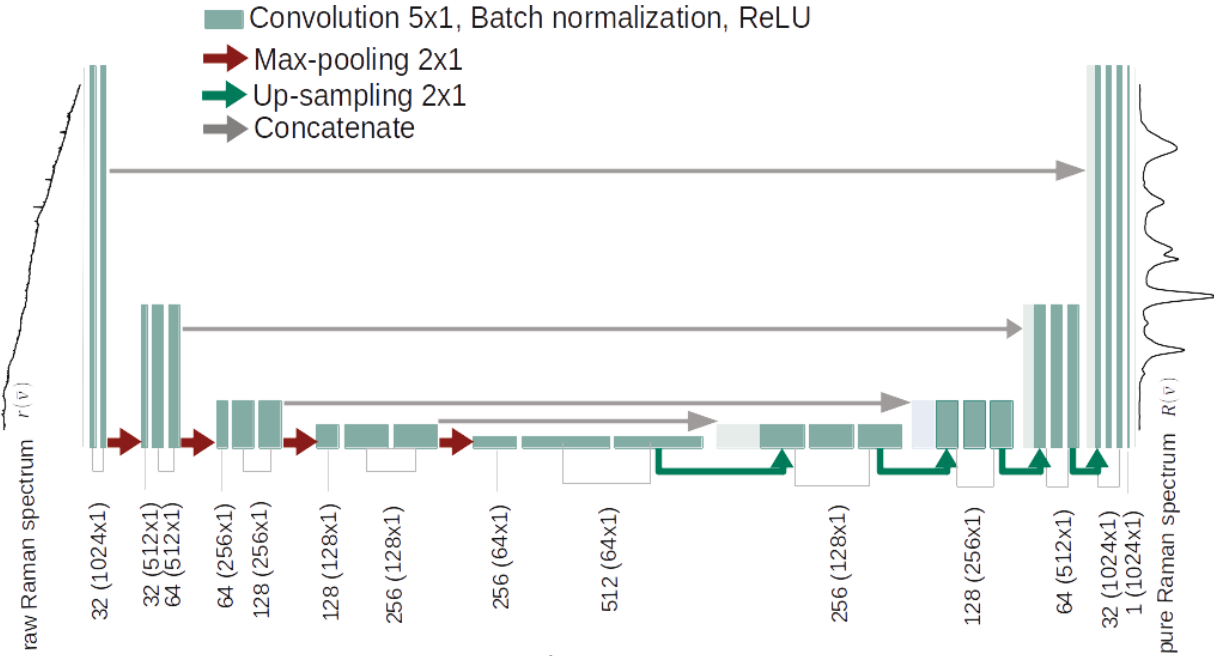


Figure 3.21: Architecture of the U-Net: light green boxes show feature maps after the application of convolutions followed by batch normalization and Rectified Linear Unit (Relu) activation. The dimensions of the feature maps are indicated under each box. Half-grey boxes illustrate feature maps for which half of their channels are copied from the encoder side. The raw spectra input to the network is shown in the left most and its equivalent output Raman spectrum is shown in the right most of the network. Brown, green and grey arrows show max-pooling, up-sampling and concatenate operations, respectively.

The encoder part was implemented by applying a typical formation of convolutional neural networks for classification tasks. It consists of two  $5 \times 1$  convolution operations followed by the max-pooling operation with a pooling size of  $2 \times 1$ . This exact sequence was repeated four times, and at the end of the encoder, two  $5 \times 1$  convolution operations were added to connect the encoder to the decoder. After each down-sampling step, the number of filters of the convolution operations was doubled, which doubled the number of feature channels.

The decoder part applies a repeated application of up-sampling and concatenation, followed by regular  $5 \times 1$  convolution operations. It first up-samples the feature channels using a  $2 \times 1$  transposed convolution operation. Then the up-sampled feature channels,

which are now reduced by half, get appended by the output of the corresponding convolutional layer of the encoder prior to the pooling operation. The concatenation step allows the network to retrieve the information lost by pooling operation during encoding. After the up-sampling and concatenation, the feature map propagated through two  $5 \times 1$  consecutive convolutional layers followed by batch normalization. Similar to the encoder, this sequential operation of up-sampling, concatenation and convolution operations repeated four times, halving the feature channels at each stage. Rectified Linear Unit (ReLU) was used to activate the output of all convolution layers of both of the encoder and decoder networks. At the end of the decoder, a  $1 \times 1$  convolutional layer with no activation function was used to retrieve the predicted output.

### References within section three

1. A. Braeuer, *In situ spectroscopic techniques at high pressure*. 2015: Elsevier.
2. F. Ehrentreich and L. Sümmchen, *Spike removal and denoising of Raman spectra by wavelet transform methods*. Analytical Chemistry, 2001. **73**(17): p. 4364-4373.
3. S.J. Barton, T.E. Ward, and B.M. Hennelly, *Algorithm for optimal denoising of Raman spectra*. Analytical Methods, 2018. **10**(30): p. 3759-3769.
4. A. Savitzky and M.J. Golay, *Smoothing and differentiation of data by simplified least squares procedures*. Analytical Chemistry, 1964. **36**(8): p. 1627-1639.
5. P.H. Eilers, *A perfect smoother*. Analytical Chemistry, 2003. **75**(14): p. 3631-3636.
6. G.W. Horgan, *Using wavelets for data smoothing: a simulation study*. Journal of Applied Statistics, 1999. **26**(8): p. 923-932.
7. Z.-M. Zhang, S. Chen, and Y.-Z. Liang, *Baseline correction using adaptive iteratively reweighted penalized least squares*. Analyst, 2010. **135**(5): p. 1138-1146.
8. S.-J. Baek, et al., *Baseline correction using asymmetrically reweighted penalized least squares smoothing*. Analyst, 2015. **140**(1): p. 250-257.
9. C. Francois, *Deep learning with Python*. 2017, Manning Publications Company.
10. N. Ketkar and E. Santana, *Deep Learning with Python*. Vol. 1. 2017: Springer.
11. L.N. Smith. *Cyclical learning rates for training neural networks*. in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2017. IEEE.
12. O. Ronneberger, P. Fischer, and T. Brox. *U-net: Convolutional networks for biomedical image segmentation*. in *International Conference on Medical image computing and computer-assisted intervention*. 2015. Springer.

## 4 Summary of the published results

The results relevant for my thesis have been published in 5 manuscripts in peer-reviewed international journals. Three manuscripts are related to the development of Raman and Raman spectral processing techniques beyond the state of the art. In detail these are:

1. M.T. Gebrekidan, C. Knipfer, F. Stelzle, J. Popp, S. Will, A. Braeuer  
A Shifted-Excitation Raman Difference Spectroscopy (SERDS) evaluation strategy  
for the efficient isolation of Raman spectra from extreme fluorescence interference  
Journal of Raman Spectroscopy 47, 198–209 (2016)
2. M.T. Gebrekidan, Ch. Knipfer, A.S. Braeuer  
Vector casting for noise reduction  
Journal of Raman Spectroscopy 51, 731–743 (2020)
3. M.T. Gebrekidan, Ch. Knipfer, A.S. Braeuer  
Refinement of spectra using a deep neural network; fully automated removal of  
noise and background  
Journal of Raman Spectroscopy 52, 723-736 (2021)

Two manuscripts are related to the application of the developed techniques for cancer diagnostics. In detail these are:

4. Gebrekidan T.M., Erber R., Hartmann A., Fasching P.A., Emons J., Beckmann  
M.W., Braeuer A.S.  
Breast tumor analysis using Shifted-Excitation Raman Difference Spectroscopy

(SERDS)

Technology in Cancer Research & Treatment 17, (2018)

5. L. Matthies, M.T. Gebrekidan, J.F. Tegtmeyer, N. Oetter, M. Rohde, T. Vollkommer, R. Smeets, W. Wilczak, F. Stelzle, M. Gosau, A.S. Braeuer, Ch. Knipfer

Optical Diagnosis of Oral Cavity Lesions by label-free Raman Spectroscopy  
Biomedical Optics Express 12, 836-851 (2021)

The subsections that follow summarize each of the respective manuscripts. The title of each subsection is equal to the title of the respective manuscript. The subsections do neither contain a motivation, the state of the art nor details on the experimental setup, as all this information has been provided in the previous sections of this thesis. Whenever appropriate, references are made to these previous sections.

#### **4.1 A shifted-excitation Raman difference spectroscopy evaluation strategy for the efficient isolation of Raman spectra from extreme fluorescence interference**

The working principle of the shifted-excitation Raman difference spectroscopy (SERDS) method is described in subsection 3.3. Practically, the SERDS method is subject to two challenges, which are that (i) the difference spectrum still contains left-over fluorescence background (due to variation in the intensity of excitation source and photobleaching) and that (ii) the difference spectrum features a low SNR even lower than that of the two raw Raman spectra (because the noise can be additive while the signal is subtractive). The methods I have developed for encountering these challenges are reported in [1] and are summarized below in the context of Figure 4.1. At the top of Figure 4.1 the two raw Raman spectra required for the SERDS method are shown. At the bottom one can see the recovered Raman spectrum as the results of the developed methods.

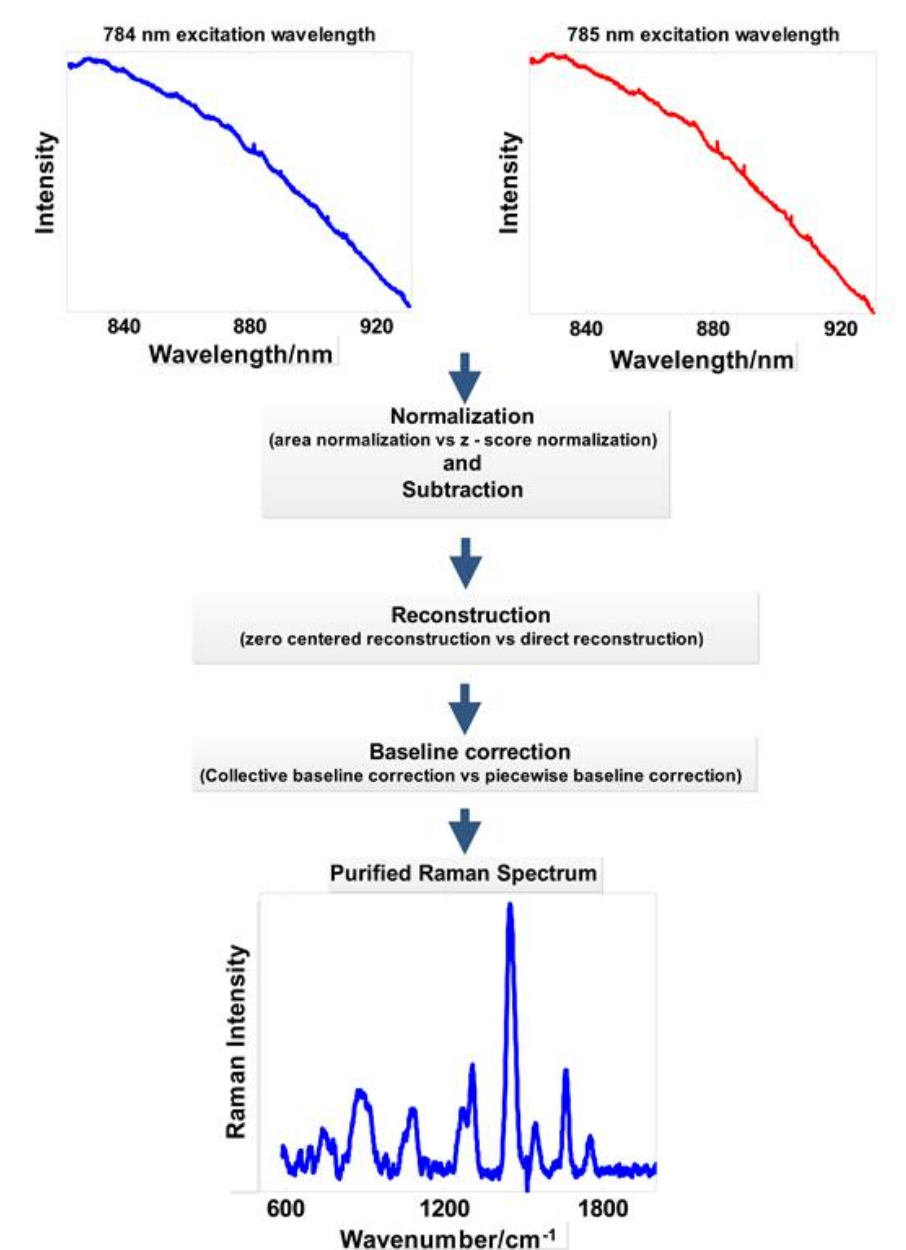


Figure 4.1 Overview of the processing steps required for the purification of the pure Raman spectrum from the interfering fluorescence. (source: Ref. [1]).

The contribution of intensity variation to the left-over fluorescence can be overcome following spectral normalization (first step in Figure 4.1). However, since the photobleaching effect changes the fluorescence contribution, the difference spectrum obtained after applying the normalization still features a left-over fluorescence. In this study zero-centering technique is recommended before spectral reconstruction for further reduction of the left-over fluorescence.

Spectral reconstruction based on the integration method was applied to the difference spectrum after zero-centering. Due to the low SNR of the difference spectrum, this method considers a Raman shift difference between consecutive spectral data points to reconstruct the Raman spectrum. This improves the SNR of the reconstructed Raman spectrum but at the cost of Raman peak shifting and Raman band broadening, which is discussed in detail in [1].

The reconstructed Raman spectrum still contains a fluorescence background though significantly reduced in contrast to the raw Raman spectra. Finally, a mathematical baseline correction algorithm was implemented based on piecewise asymmetric least squares (see subsection 3.5) to remove the left-over fluorescence from the reconstructed Raman spectrum.

The capacity of the developed methods is illustrated in Figure 4.2, which compares a pure Raman spectrum of ethanol with the one reconstructed from a heavily fluorescence interfered ethanol/dye-solution (top in Figure 4.2). The reconstruction was even able to detect the Raman signature of the dye cryptocyanine.

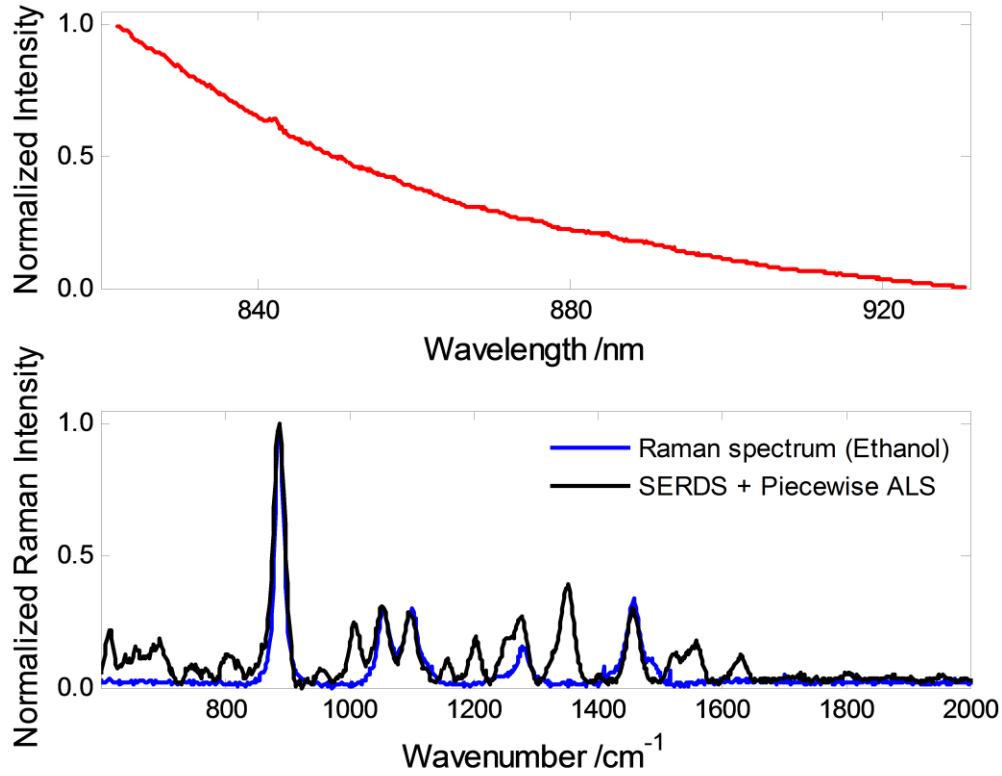


Figure 4.2: (top) Raw spectrum taken from the cryptocyanine-ethanol-solution and (bottom) the comparison of a pure Raman spectrum acquired from pure ethanol with a Raman spectrum reconstructed according to the here proposed method from the raw spectrum shown at the top. (source: Ref. [1]).

## 4.2 Vector casting for noise reduction

Subsection 3.5.1 showed that the removal of noise is a trade-off between smoothing the spectrum but also modifying the peaks (shape, central position, peak maximum).

In [2] a completely novel noise reduction method based on casting vectors is reported. The vector casting method initially searches for the top and bottom envelopes of the noisy spectra. On this account, in a first-level, all peaks and valleys, irrespectively of whether the peak or valley is due to noise or due to a real signal were identified. These first-level peaks and valleys are a good first estimate of the top and bottom envelopes of the raw

Raman spectrum. In a second-level, each of these first-level envelopes are decomposed into peak/valley pairs. This process is iterated and at the third-level, peak/valley pairs of the second-level are decomposed into further peak/valley pairs. The consecutive data points of third-level valleys of the second-level valleys of the first-level peaks were then considered as possible left and right border of a signal peak. Then, the presence of a real signal peaks is regarded by considering two criteria; (i) whether the height of the potential signal peak is greater than the maximum value of the difference between two consecutive data points of the noisy spectra and (ii) whether the slopes of the linear fits of the first-level peaks and valleys left of the maximum of the potential peak are both positive and right of the potential peak are both negative. Following this, a categorization of the first-level peak/ valley pairs of data points into part of peak region and peak free region were carried out. Afterwards, the top and bottom envelopes, which are the first-level peaks and valleys, are smoothed considering an asymmetric moving average windows.

In order to retrieve the noise-reduced spectrum, vectors are casted within the margin of the above computed smooth envelopes from a starting already noise-reduced data point to subsequent not yet noise-reduced data points. Vectors that cross either of the envelopes are deleted. Then the not yet noise-reduced data point that is situated one increment right of the already noise-reduced data point is computed as the mean of the intercept of the remaining vectors. The performance of the vector casting method was compared with the state-of-the-art noise reduction approaches using simulated Raman spectra of various SNRs and experimentally acquired Raman spectra. The method performs well, especially at small SNRs, when quantified by different signal-to-noise ratio metrics. And it comparably relies on a minimum of human input. The drawbacks of this method are that it involves peak detection and takes longer execution time, which depends on the number of vectors to be casted.

Figure 4.3 compares the vector casting method for noise reduction with the Savitzky-Golay and the wavelet methods. It can be seen that the vector casting method results in



excellent noise reduction, least modifies the peak characteristics and thus is superior (under the tested circumstances) to the other methods.

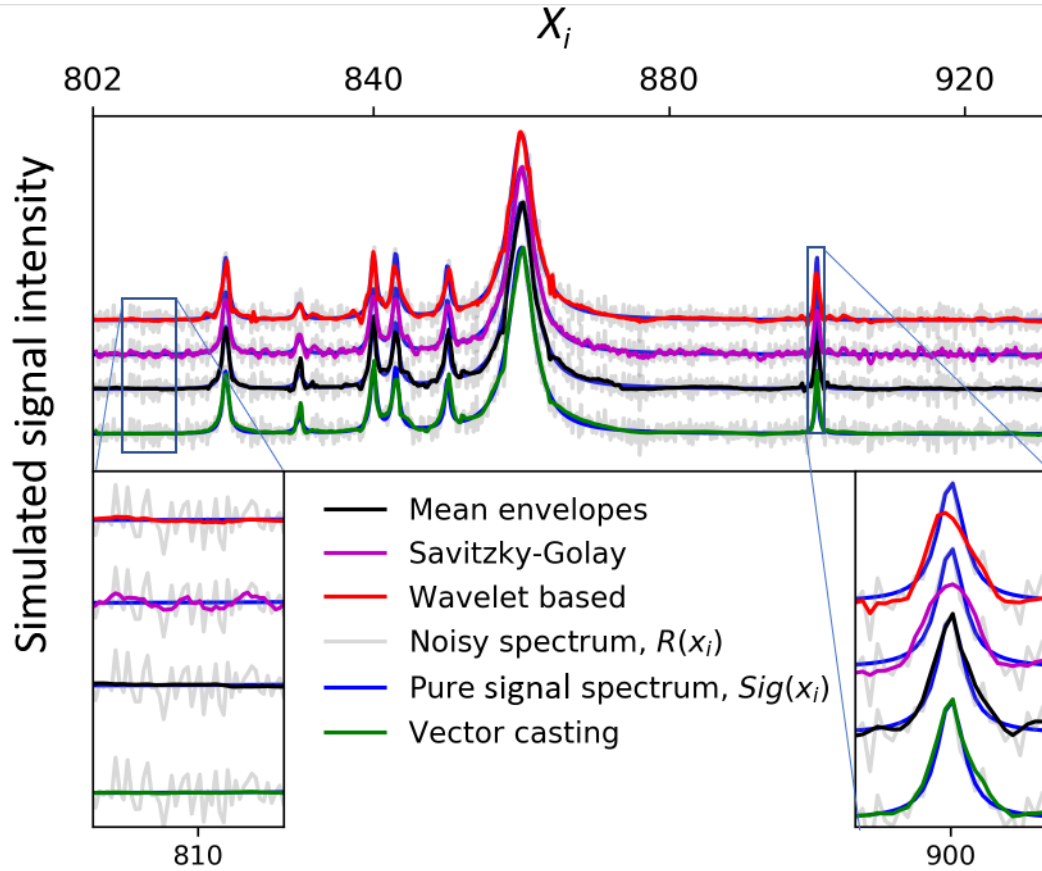


Figure 4.3: Comparison of noise-reduced spectra  $r(x_i)$  (black line) using (a) vector casting method, (b) envelope-finder algorithm, (c) wavelet based smoothing and (d) Savitzky-Golay filter with respect to the pure signal spectrum  $R(x_i)$  (blue line). The original noisy spectrum is shown as gray line. (source: Ref. [2]).

### **4.3 Refinement of spectra using a deep neural network; fully automated removal of noise and background**

Compared to the vector casting method (previous subsection), the deep neural network approach preserves the peak shapes and at the same time it efficiently removes the noise. Moreover, the neural network model does not at all rely on any human intervention.

The architecture of the U-Net network has already been introduced in Figure 3.21 in subsection 3.6. The U-Net has been trained with hundreds of thousands of synthetic raw Raman spectra that were simulated according to the description provided in subsection 3.2. The respective reference [3] contains different application scenarios where the trained U-Net provided noise-reduction and background-elimination capabilities which were superior to the state of the art methods. Here its capability is demonstrated in Figure 4.4 with respect to a highly fluorescence interfered cryptocyanine/ethanol-solution. The higher the cryptocyanine weight fraction in the solution, the more the raw Raman spectra was interfered with fluorescence background. The results revealed the very positive performance of the U-Net method in the preservation of Raman spectral features and efficient removal of noise. The performance of the different background correction methods was also quantified based on different SNRs metrics and the structure similarity index. The metrics confirmed the superiority of U-Net with respect to the other background correction methods. The structure similarity index obtained, when having treated the raw Raman spectra of cryptocyanine/ethanol-solutions with three different spectra processing methods are reported in Figure 4.5. This figure clearly demonstrates the superiority of the U-Net method relative to the state of the art methods arPLS and ModPoly (compare subsection 3.5.2).

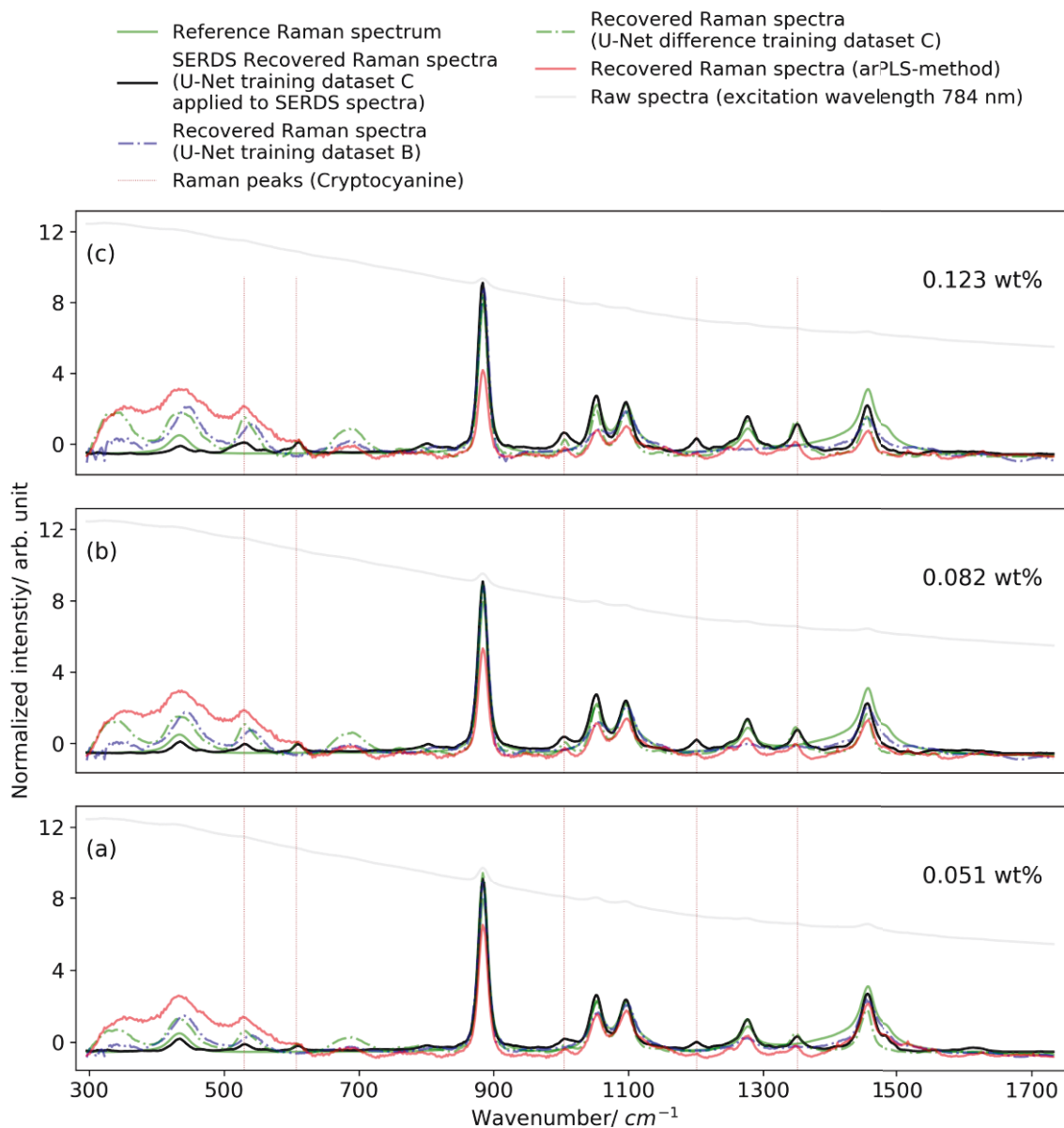


Figure 4.4: Raw spectra (grey lines) of ethanol/cryptocyanine-mixture excited at 784 nm wavelength. Recovered Raman spectra obtained using arPLS, the U-Net method trained by training dataset B, the U-Net difference method trained by training dataset C and the U-Net difference model when applied to SERDS spectra (The SERDS spectrum is not shown). The reference Raman spectrum of ethanol is shown in green. (a) Raw Raman spectrum from ethanol/cryptocyanine-mixture of 0.051 wt% weight fraction of cryptocyanine. (b) Raw Raman spectra obtained from ethanol/cryptocyanine-mixture of 0.082 wt% weight fraction of cryptocyanine and (c) Raw Raman spectra obtained from ethanol/cryptocyanine-mixture of 0.123 wt% weight fraction of cryptocyanine. (source: Ref. [3]).

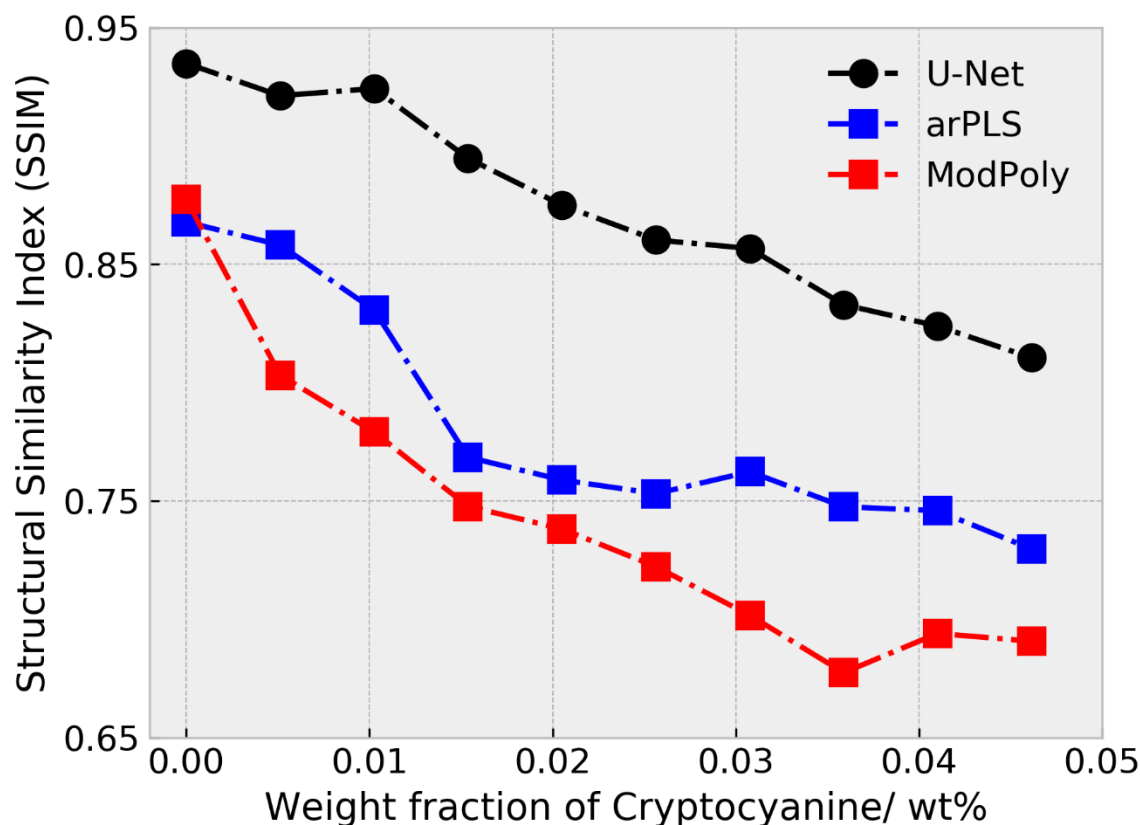


Figure 4.5: Comparison of Raman extraction performance of U-Net (black circles), arPLS (blue squares) and ModPoly (red squares) using structural similarity index metric. (source: Ref. [3])

#### 4.4 Breast Tumor Analysis using Shifted-Excitation Raman difference Spectroscopy

SERDS, combined with machine learning tools, is investigated in this study for the analysis of breast tumor and the details are reported in [4]. The Raman spectra refinement approaches proposed in this study are applied to separate the Raman spectrum from extreme fluorescence interference. The purified Raman spectra are then fed to machine learning tools for spectra interpretation. On this account *ex vivo* classification of resected and formalin-fixed breast tissue samples as normal tissue, fibroadenoma or invasive

carcinoma was investigated using the respective purified Raman spectra. The normal breast tissue is mainly composed of fat with only small contributions of collagen; as such one can clearly see the Raman signatures in the raw Raman spectra (see Figure 4.6a dashed black curve) which is governed by less fluorescence background. On the contrary, the raw Raman spectra of fibroadenoma and invasive carcinoma (see Figure 4.6 black and gray curves respectively) are dominated by extremely strong fluorescence, which hindered the visibility of Raman signatures. Thus, the raw Raman spectra have to be refined for further spectra interpretation. Thus SERDS with further mathematical methods based on curve fitting and simple data processing regimes as reported in [1] was applied to recover the Raman spectra. Figure 4.6 shows the recovered Raman spectra. The different signatures of Raman spectra

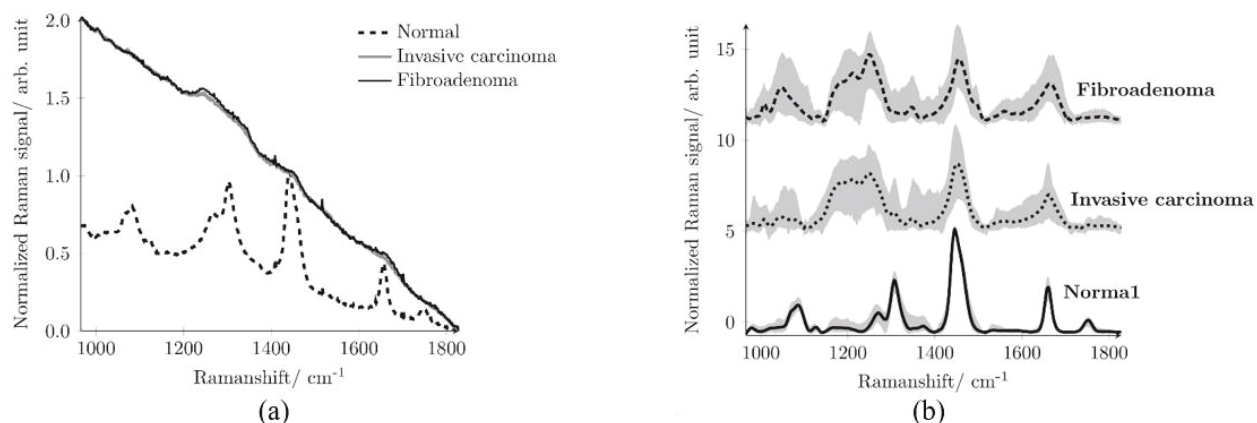


Figure 4.6: (a) Mean raw spectra acquired from the three different breast tissue sites, normal (broken black line), fibroadenoma (solid black line), and invasive breast carcinoma (solid gray line) using 785 nm excitation wavelength. (b) Reconstructed mean Raman spectrum of fibroadenoma (dashed black line), invasive carcinoma (dotted black line) and normal tissue (solid black line) together with standard deviation (shaded gray area). (source: Ref. [4])

are assignable to respective molecular tissue composition. The majority of the Raman signatures of the healthy tissue originate from lipids. On the other hand, many of the Raman spectral features of the tumor tissues are assigned to protein. Several clear Raman signatures differentiate between normal breast tissue and tumor tissue,

irrespective of whether the tumor is fibroadenoma or invasive carcinoma. On the contrary the chemical composition of the tumor tissues are very similar to each other. Thus their difference in the Raman spectra are less pronounced.

A principal component analysis and linear discriminant analysis (PCA/LDA) was then applied to the purified Raman spectra to objectively classify between the different tissue types. The result shows excellent classification of normal breast tissue, fibroadenoma and invasive breast carcinoma. Figure 4.7 summarizes the performance of the classification

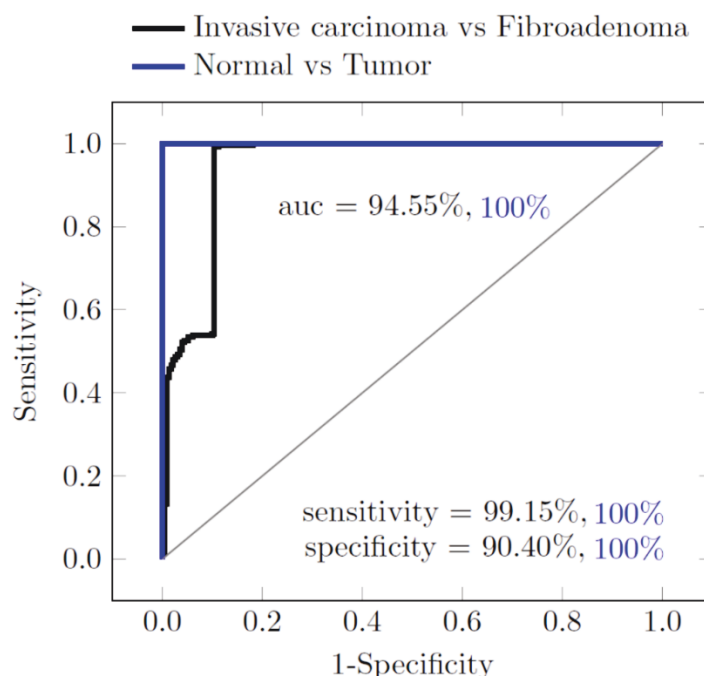


Figure 4.7: Receiver-operating characteristic (ROC) curve explaining the performance of the PCA/LDA classifier, between normal and tumor tissue (blue curve, sensitivity of 100%, specificity of 100 and area under the curve (auc) of 1), and between invasive breast carcinoma and fibroadenoma (black curve, sensitivity of 99.15%, specificity of 90.40% and auc equal to 0.9455). (source: Ref. [4])

per Receiver operating characteristic (ROC) curve. Tumor tissues (fibroadenoma and invasive carcinoma) were identified with 100 % sensitivity, and healthy tissues were identified with 100% specificity. Additionally the results indicate a high sensitivity of 99.15% for invasive breast carcinoma and a high specificity of 90.40% for fibroadenoma.

#### **4.5 Optical diagnosis of clinically apparent lesions of oral cavity by label-free Raman spectroscopy**

Oral squamous cell carcinoma (OSCC) is one of the most prevalent cancers and frequently preceded by non-malignant lesions. In [5] SERDS combined with classical machine learning (PCA/LDA) was applied for the diagnosis of clinically apparent lesions of the oral cavity. Unlike the breast, the oral cavity is complex, it contains different types of tissues, and depending on the anatomical location, content and amount of tissue can be different. Thus, the Raman spectrum of the healthy oral tissue contains diversified features, and the raw Raman spectra are also governed by less to extreme intensive fluorescence background. Figure 4.8 shows the Raman spectra of physiological tissue (green curve), OSCC (red curve) and non-malignant lesions (black curve) that were averaged from 42, 95 and 43 reconstructed Raman spectra, respectively. The non-malignant lesions involved distinct histopathological entities such as inflammation, leukoplakia, dysplasia, hyperkeratosis and irritant fibroma. Annotations assign the spectral Raman signatures to their known molecular origin. The differences between physiological and pathological tissues (non-malignant and OSCC) are noticeable in the whole spectral region analyzed, although more pronounced in the fingerprint region ( $1200\text{--}1800\text{ cm}^{-1}$ ). The differences between malignant and non-malignant lesions were less pronounced and occurred mainly in the region between  $800$  to  $1400\text{ cm}^{-1}$ . The spectral features of physiological tissue of the oral cavity reflected dominant contribution from lipid molecules, a strong  $\text{CH}_2$  band around  $1448\text{ cm}^{-1}$ , two sharp peaks/bands around the amide III region, a sharper peak around the amide I region and altered nucleic acid spectral features. In contrast, spectra from non-malignant and OSCC tissues feature broader amide I and III regions, a shifted and weaker  $\text{CH}_2$  band and a strong phenylalanine peak.

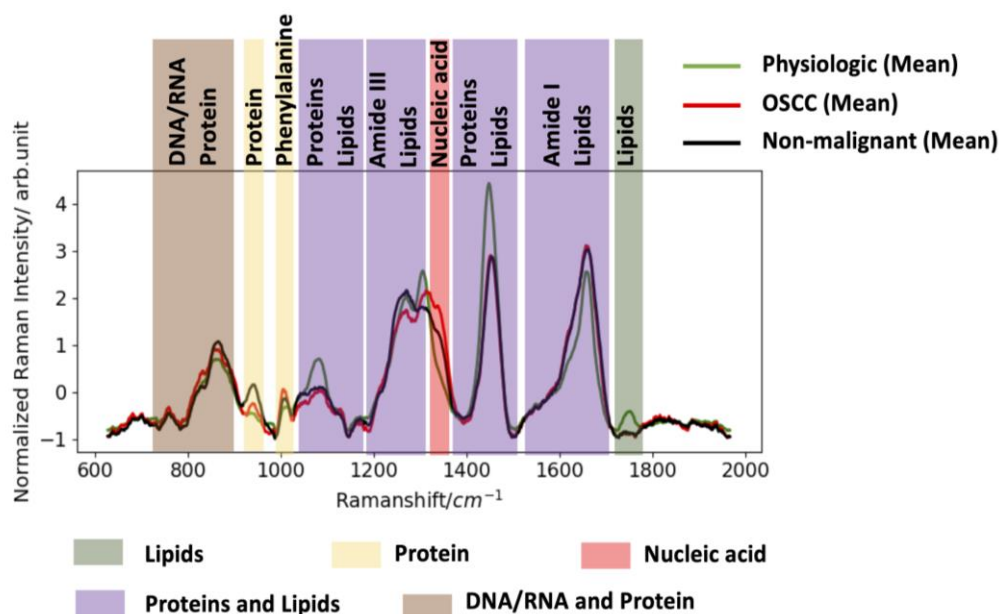


Figure 4.8: Reconstructed mean Raman spectra of physiological oral tissue (green line), malignant (red) and non-malignant lesions (black) and with peak position assignment to their respective molecular origin. Lipids (light green), protein (light yellow), both proteins and lipids (light magenta), nucleic acid (light red), and proteins, nucleic acid and carbohydrates (light brown). (source: Ref. [5])

The feasibility of the classification of physiological, non-malignant and malignant lesions based on their reconstructed Raman spectra was explored using PCA/LDA. Figure 4.9 illustrates the receiver-operating characteristic curve (ROC) that visualizes the performance of the classification. In the physiological mucosa vs. non-malignant lesions classification, it was found a 5-fold cross-validation accuracy of 95.3 % at a sensitivity of 95.4 % and specificity of 95.2 %. The area under the curve was found to be 0.99, with only four out of 85 tissue spectra being misclassified. Spectral classification of OSCC against physiological mucosa gave an overall accuracy of 89.8 % at a sensitivity of 93.7 % and specificity of 81.0 %. Here, a total of 123 out of 137 were correctly classified while eight physiological tissue samples were falsely classified as OSCC and six OSCC were misclassified as physiological tissue. The respective AUC was determined to be 0.90.



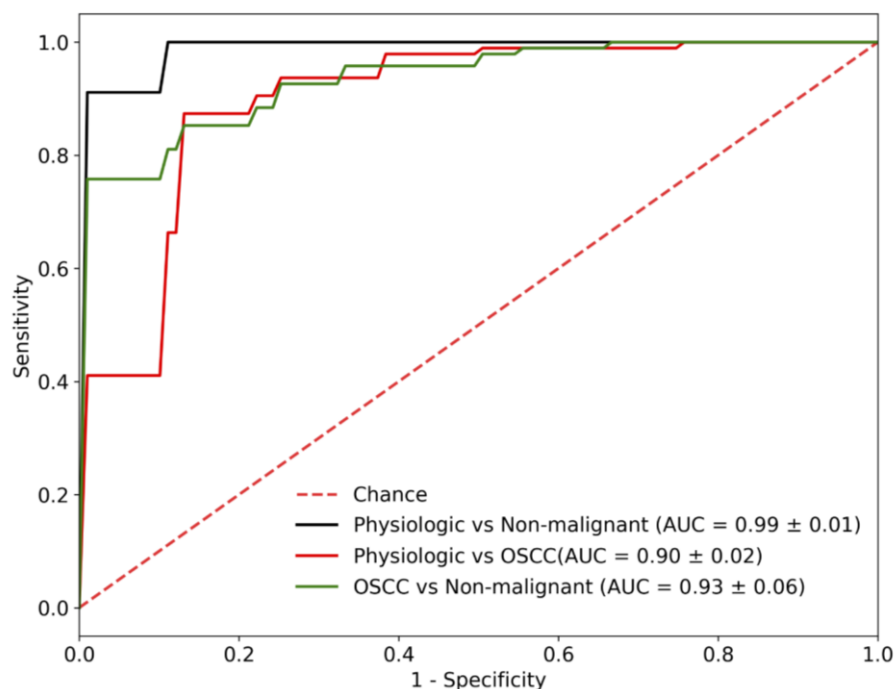


Figure 4.9: Receiver-operating characteristic curve. Performance of the PCA-LDA classifiers, between physiological mucosa and non-malignant lesions (black curve), physiologic vs. OSCC (red curve), and OSCC against non-malignant lesions (green curve). (source: Ref. [5])

Classification of OSCC against non-malignant tissues resulted in a sensitivity of 93.7 % and specificity of 76.7 %. In total, 122 out of 138 pathological tissues were correctly classified while 16 were falsely classified giving an overall accuracy of 88.4 % and an AUC of 0.93. The respective reference [5] provides additional details and discussions.

#### Reference within section four

1. M.T. Gebrekidan, et al., *A shifted - excitation Raman difference spectroscopy (SERDS) evaluation strategy for the efficient isolation of Raman spectra from extreme fluorescence interference*. Journal of Raman Spectroscopy, 2016. **47**(2): p. 198-209.
2. M.T. Gebrekidan, C. Knipfer, and A.S. Braeuer, *Vector casting for noise reduction*. Journal of Raman Spectroscopy, 2020. **51**(4): p. 731-743.

3. M.T. Gebrekidan, C. Knipfer, and A.S. Braeuer, *Refinement of spectra using a deep neural network: Fully automated removal of noise and background*. Journal of Raman Spectroscopy, 2021. **52**(3): p. 723-736.
4. M.T. Gebrekidan, et al., *Breast Tumor Analysis Using Shifted-Excitation Raman Difference Spectroscopy (SERDS)*. Technology in cancer research & treatment, 2018. **17**: p. 1533033818782532.
5. L. Matthies, et al., *Optical diagnosis of oral cavity lesions by label-free Raman spectroscopy*. Biomedical Optics Express, 2021. **12**(2): p. 836-851.

## Conclusion

Raman spectroscopy gains potency for a large variety of practical applications in different fields. Fiber optics facilitate the minimally invasive analysis in hazardous environments and *in vivo* investigations of hardly accessible samples such as inside the human body. On the contrary, intensive fluorescence interferences and low SNR pose considerable challenges to the interpretability of Raman spectra. Especially these two challenges are encountered by this thesis. This thesis reports the development of methods for the purification of Raman spectra from heavily interfered raw Raman spectra. The capabilities of the methods are demonstrated by their application to chemical engineering tasks or to cancer diagnostics in a medical context.

Different methods for the processing of spectra that are relevant for the purification of the Raman spectra are developed and tested in the context of noise reduction and fluorescence rejection. The results show that it is possible to extract pure Raman spectra from heavily fluorescence-interfered raw Raman spectra, which feature low SNR. Based on the comparison with the state of the art techniques, the approaches proposed in this study significantly improve the Raman spectra post-processing. Even though explored and tested for Raman spectroscopy, these techniques can also be applied for other techniques such as diffractometry, emission spectroscopy, absorbance spectroscopy or chromatography. The exploration of deep neural networks for Raman spectra refinement showed promising results. In addition to the significant improvement of purification of Raman spectra, such an approach does not rely on any input parameters that humans

have to choose. Therefore, a pre-trained neural network can play a significant role in the automation of Raman spectral processing.

In the scope of this study I demonstrated the potential of the proposed techniques for broadband backgrounds. However, the proposed techniques are also promising to efficiently recover Raman spectra from other complex backgrounds which might be composed of several non-variant narrowband peaks. Throughout all post-processing steps, care has been taken not to eliminate Raman signals from the processed spectra.

According to the world cancer report, cancer is a major public health problem worldwide and the leading cause of premature death in most countries. In this study, I demonstrated the effectiveness of Raman spectra preprocessing approaches in combination with machine learning tools to identify breast and oral squamous cell carcinoma. I explored whether physiological mucosa, non-malignant lesions and oral cancer can be differentiated. To this end, pure Raman spectra of the respective tissues were isolated from heavily fluorescence interfered raw spectra using the proposed Raman purification techniques. The results demonstrate that malignant and non-malignant lesions can be differentiated from the physiological tissue of the oral cavity with a high accuracy, despite considerable heterogeneity of these lesions. The technique showed excellent results for the correct distinction of non-malignant lesions, oral squamous cell carcinoma and physiological mucosa. Invasive breast carcinoma can also be differentiated from fibroadenoma with a high accuracy by isolating the pure Raman spectrum of the respective tissue from its heavily fluorescence interfered raw spectrum. The differentiation between diseased and non-diseased was also found to be correlated to and determined by the spectral features of protein, lipid and nucleic acid.

Putting everything in a nutshell, I have developed a method which outperforms state of the art methods with respect to noise reduction and background elimination and which additionally does not rely on any human intervention.