UiT The Arctic University of Norway

Department of Physics and Technology

Raman-spectroscopy of extracellular vesicles and self-supervised deep learning

Bridging optics, chemistry, and computer science

Mathias N. Jensen

A dissertation for the degree of philosophiae doctor – October 2023

Abstract

This thesis explores the prospect of a waveguide device for optical trapping and Raman spectroscopy of single biological nanoparticles. The aim of the work is to trap and measure particles in parallel, exploiting a multichannel spectrometer to collect measurements from the particles separately, drastically increasing the throughput of the method. The challenge of induced Raman background in the waveguide is investigated, it is found that UV-written SiO_2 waveguides produce a Raman background lower than -107.4 dB, which is 15 dB lower than in Si_3N_4. The Raman background in UV-SiO_2 waveguides is shown to be without prominent features in the biological fingerprint region (800-1700 cm^{-1}).
Furthermode, it is shown that a convolutional neural network can be developed for analysis of tomographic scans of silicon boules, with the goal of developing a general machine learning platform.The developed method successfully detected the quality of the crystalline structure of the silicon from the tomographic scans, achieving an accuracy of 98.7%. It is shown that the method remains accurate despite a reduction in signal-to-noise ratio in excess of 10 dB, demonstrating good robustness to noise.
An autoencoder architecture constructed from the previous convolutional neural network is shown to be able to recovering the Raman spectra of extracellular vesicles from spectra contaminated by the background generated by SiO_2 waveguides. The method is shown to be able to recover the spectra with very high fidelity, increasing the signal-to-noise ratio from -18$\pm$3 dB to 5.4 dB. Furthermore, the model is demonstrated to be capable of differentiating the spectra of particles with different biological origins well, using learned components corresponding to chemical elements in the particles.
Further developments of this autoencoder design demonstrates that it can also be made capable of adapting to variations in noise level, frequency dis-

tortions, and spectral resolution, making it able to accept spectra from different measurement systems. The features extracted by the autoencoder demonstrate good differentiation of biological nanoparticles by their Raman spectra. The differentiation is verified by an external classifier network, which is demonstrated to achieve a sensitivity of 92.2% and a specificity of 92.3% in detecting the biological origins of the spectra. The model is demonstrated to both de-noise the spectra and to be robust against noise and distortions in the spectra, demonstrated by the classification accuracy remaining over 80% for spectra with noise, frequency distortions, and with up to 80% of the spectra missing.

The combination of UV-written SiO_2 waveguides with integrated optics is promising for a high-throughput Raman-on-chip device capable of parallel trapping and measurement. Further augmentation of the method with machine learning is shown to solve the challenge of induced background in the waveguides. The machine learning methods developed for this purpose also demonstrate the ability to differentiate nanoparticles with high accuracy and significance in the presence of noise and distortion.

Acknowledgements

I would first and foremost express my gratitude to my supervisor, Olav Gaute Hellesø, who introduced me to the world of experimental physics. He has guided me well through a my journey through my masters degree, and now my PhD, during which he has supported my efforts and left me knowing and understanding more than I often realize.

I would also like to extend a profound thank you to my fellow researchers and peers in the UMO group. Through their helpfullness, resourcefullness, and their cheerfullness, they have filled my PhD journey with opportunities and possibilities. I would like to give special thanks to my co-supervisors, Jana Jagerská and Benjamin Ricaud, who have provided me with support and have always been quick to offer solutions.

I am also grateful to the Department of Physics and Technology as a whole, as more people from its various branches and offices have helped me thank I can list here.

I would also like to give thanks to the Medical cell BioPhysics group at the University of Twente, especially Cees Otto, for giving me the opportunity to stay and learn at their lab. I would also like to thank Sergei Kruglik at Sorbonne University for his contributions and hospitality. I am also grateful to Omri Snir and his associates at the Thrombosis Research Group for helping me bridge the gap between physics and biology .

Lastly, I would like to thank Jon-Richard, Abhishek and Rabiul for our jovial discussions and talks covering anything and everything, both within and far from our fields of work.

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

In fields from biology and medicine to chemistry and material science, the chemical composition of materials and substances defines what they are and how they interact with the world. Therefore, determining the chemistry and structure of something is paramount to understanding it and deciding what to do with it. In many ways, the endeavor to measure just this can be considered as one of the fundamental motivations behind the scientific process, leading from alchemy and the four elements to modern chemistry and quantum mechanics. Modern chemistry as we know it came into maturity as late as 1669 when the German chemist Hennig Brandt discovered phosphorus [75], filling in the first box in the periodic table.Since then, chemistry has come a long way, both in filling the periodic table, and understanding the nature of the elements on it. However, as the field of chemistry slowly evolved into what we know today, as did the field of optics. As the fields of chemistry and optics matured, they began crossing more and more into each other, starting with the idea presented by Irish physicist Sir George Stokes in 1864 [67] in using fluorescent chemical tags to enable chemistry to be imaged. While this idea spawned a plethora of powerful techniques that remain at the forefront of biological and chemical imaging even today, the concept of fluorescent imaging remains limited by the selection of fluorophores used to chemically label the sample, and this selection is again limited by the pre-existing knowledge of the chemistry of the sample. Thus, a method that would allow such imaging without this

burden of chemical labeling has been sought after since this limitation was uncovered. A mechanism that enabled such a method began to emerge in the 1920's, first as an idea by Austrian physicist Adolf Smekal in 1923[66] and then later observed in 1928 by Indian physicist C. V. Raman[61], with the latter giving it its name: Raman scattering. Through Raman scattering, the bonds between atoms can be interrogated using only light, thus making it possible to investigate the chemistry of a sample by analysing the spectrum of the scattered light. Because this effect works directly on the molecules of the sample, no chemical modification or labeling is required. The method of analysing the Raman scattering, or Raman spectroscopy, of a sample is therefore a completely label-free method, making it a powerful investigative method and a viable contender to more traditional chemical imaging like fluorescent imaging. In addition to not requiring pre-existing knowledge of the sample chemistry or specially designed fluorophores, the use of label-free Raman spectroscopy also preserves the chemistry of the sample. This solves another challenge observed in fluorescent imaging, namely toxicity in the sample, making the act of measuring the chemistry detrimental to the sample.

Because of its ability to perform label-free analysis, and to do so non-invasively and non-destructively, Raman spectroscopy has become a method of interest in systems of complex chemistry, notably biology[50, 69, 47]. While there are many methods of analysing the chemistry of biological materials, the fact that Raman spectroscopy uses only light affords it a unique advantage of being able to analyse the chemistry in very small volumes. Because the analysis is conducted using light, the size of the volume that can be interrogated is in principle only limited by the diffraction limit of the system, which through modern optics can readily be made as small as 200 nm in width. With the progress of modern instrumentation, with increasingly higher sensitivity and resolution, along with powerful and spectrally pure lasers, Raman spectroscopy is capable of measuring the chemistry of individual cells[33, 58] and even measuring the chemistry at different locations within a single cell. This enables chemical imaging down to the submicron level, opening the internal chemistry of cells for researchers to see. Similarly, Raman spectroscopy can be used to chemically analyse objects smaller than cells, such as biological nanoparticles. When used in combination with optical tweezers[38, 51] it becomes possible to isolate, confine, and measure the chemistry of nanoparticles smaller than 100 nm on a per-particle basis. One such type of nanoparticle that is of special interest is

the extracellular vesicle (EV), which are "messenger" particles emitted by cells. The communication between cells, facilitated by EVs, has been shown to play a significant role in the function of cells in groups[11], and it has been demonstrated that the chemical makeup of EVs can be used as a biomarker for a myriad of conditions[56, 12, 76, 3, 71]. Optical trapping and Raman spectroscopy therefore enables us to analyse EVs and detect these biomarkers, and provides a promising avenue for detecting[19] and diagnosing various conditions.

The ability to measure single nanoparticles enables Raman spectroscopy to be a powerful tool in research and diagnostics, but this selectivity comes with its own challenges. The principal challenge stems directly from one of its principal advantages, namely that it analyses few or single particles at a time. Because nanoparticles like EVs exist in the millions pr. millilitre, making a statistically significant inference on the state of cells they originate from require the analysis of a large amount of particles. With only a few of them being measured in each acquisition, this leads to very low throughput, making the method slow and cumbersome. So, to make Raman spectroscopy of EVs a viable method for diagnosis, the throughput must be increased. One way this can be achieved is to forego the single particle selectivity, and instead measure the particles in bulk, thus increasing the throughput. However, this gives limited information on the distribution of the particles by reducing it to a mean measurement, which reduces the available information and thus the quality of the inference made from it. Another approach to increase the throughput is through parallelization of the acquisition, collecting Raman scattering from multiple particles at the same time without mixing the generated spectra. To achieve this, the measurement system must be designed such that it has multiple sample sites, each capable of trapping one or few nanoparticles, and collecting the scattered light from each of the sites separately. A way of achieving such a design is through the use of integrated optics, by using a photonic chip with waveguides to replace the lenses in a traditional Raman system, it becomes possible to create an array of micron-sized structures. Through careful design, these structures can be made to serve as trapping sites for single nanoparticles, and by adapting the collection system to a multi-channel spectrometer, it becomes possible to project the light generated from each of the structures onto individual channels. Thus, we investigate if a waveguide chip can be fabricated into a Raman-on-chip device capable of facilitating Raman spectroscopy of multiple nanoparticles

in parallel, which can be further combined with microfluidics to provide high-throughput, single particle Raman spectra.
Once collected, the Raman spectra contains a wealth of information on the sample that generated it, but for this information to be usable, it must be analysed and interpreted. With the level of complexity of such information, and the amount of data generated by a Raman-on-chip device, this analysis becomes a significant challenge. To realize the goal of using Raman spectroscopy of EVs as a tool for research and diagnostics, the physical method must be complemented by a method of analysis that can translate the Raman spectra into usable information. With the available processing power of modern computers, machine learning is a strong contender for such an analysis method. The flexibility, adaptability, and ability to detect complex patterns in data has made machine learning methods powerful techniques for a wide range of applications. To apply machine learning to the Raman spectra of EVs, a suitable method must be developed. To do this, we first consider tomographic scans of Silicon as a test case. The goal is to recognize the quality of the crystalline structure from the transmission of infrared light through the structure. This data contains complex features similar to features in Raman spectra, and because the ground truth of the crystalline structure is readily available, this serves as a good test bench to develop a machine learning architecture through supervised learning. This architecture is then used as the basis for the Raman analysis method by making the model capable of self-supervised learning. The resulting architecture is capable of label-free learning from the label-free Raman spectra of EVs and other nanoparticles, and able to extract valuable information from a large number of spectra, and thus complement the high throughput of the waveguide Raman-on-chip device.

1.2 Scope of the thesis

In this thesis, the prospects of a waveguide-based Raman-on-chip device for extracellular vesicles are investigated, and a machine learning method is developed for this application. The focus of the thesis will be on three main areas: 1) Investigation of a waveguide device for trapping and exciting Raman scattering from nanoparticles approximately 100 nm in size. 2) Development of a machine learning architecture for determining the quality of presumed monocrystalline silicon by recognizing features in infrared

transmission tomography scans of silicon boules. 3) Expansion of the machine learning architecture into a model capable of extracting information from Raman spectra of biological nanoparticles through self-supervised learning.

1.3 Structure

The the thesis is divided into four main chapters, with the last three covering the themes outlined in the scope of the thesis.
Chapter 2 of the thesis will present an overarching view of the theory and state of the art of Raman spectroscopy and some select current techniques. This chapter also introduces optical trapping.
Chapter 3 of the thesis will focus on waveguide-based Raman spectroscopy. An overview of waveguide-based Raman techniques will be presented and its advantages and challenges introduced. The potential of Raman on waveguide will be illustrated here and the approach for the project will be presented.
Chapter 4 of the thesis will describe the development of the fundamental elements of the machine learning architecture. In this chapter, tomographic scanning by infrared transmission through monocrystalline silicon boules will be presented. A machine learning architecture is developed for determining the quality and intactness of the crystalline structure from the tomographic scans.
Chapter 5 of the thesis will describe the adaptation of the machine learning architecture to consider Raman spectra of biological nanoparticles. The machine learning architecture developed in chapter 4 will be expanded into a specialized autoencoder architecture that is capable of self-supervised learning from the Raman spectra of nanoparticles.

1.4 Publications

Paper I
M. N. Jensen, James C. Gates, Alex I. Flint, and Olav Gaute Hellesø, "Demonstrating low Raman background in UV-written SiO_2 waveguides", Optics Express, vol. 31, no. 19, pp. 31093-31107, Sep. 2023. **Author contributions:** Mathias N. Jensen performed all experimental work and data

analysis. James C. Gates and Alex I. Flint developed the UV-writing process and fabricated the waveguide chips. Olav Gaute Hellesø conceived the idea and oversaw the work. Mathias N. Jensen wrote the initial draft and Olav Gaute Hellesø finalized the manuscript for submission. All authors contributed to revision of the manuscript before publication.

Paper II

M. N. Jensen and Olav Gaute Hellesø, "Evaluation of crystalline structure quality of Czochralski-silicon using near-infrared tomography", Journal of Crystal Growth, vol. 581, no. 1, pp. 126527, Apr. 2022. **Author contributions:** Mathias N. Jensen conceived the original idea, performed all experimental work and data analysis. Olav Gaute Hellesø suggested and oversaw the work and oversaw the work. Mathias N. Jensen wrote the initial draft and Olav Gaute Hellesø finalized the manuscript for submission. Both authors contributed to revision of the manuscript before publication.

Paper III

M. N. Jensen, Eduarda M. Guerreiro, Agustin Enciso-Martinez, Sergei G. Kruglik, Cees Otto, Omri Snir, Benjamin Ricaud, and Olav Gaute Hellesø, "Identification of extracellular vesicles from their Raman spectra via self-supervised learning", submitted to Nature Scientific Reports. **Author contributions:** Mathias N. Jensen concieved the idea and implemented the method. Eduarda M. Guierreiro and Omri Snir prepared and provided samples for data generation. Agustin Enciso-Martinez and Sergei G. Kruglik conducted the experimental work. Cees Otto conducted and oversaw parts of the experimental work. Benjamin Ricaud oversaw the development of the architecture. Olav Gaute Hellesø oversaw the work. Mathias N. Jensen wrote the initial draft, Olav Gaute Hellesø and Benjamin Ricaud finalized the manuscript for submission.

Chapter 2

Raman spectroscopy

2.1 History of Raman spectroscopy

The concept of what we now know as Raman scattering has appeared multiple times through scientific history, receiving significant attention in the early 1900's. One of the initial discussions that inspired the discovery of Raman scattering was published by Sir Joseph Larmor [41] in 1919, who deliberated on the findings of Lord Rayleigh regarding the scattering of light by particles. He explored the implications of such scattering, how Rayleigh explained the blue hue of the daytime sky as an effect of scattering from molecules in air, and how or if this or similar effects could arise in liquids and solids. This deliberation served as a primer for the then still young C. V. Raman, who would incubate his own ideas on the phenomenon in the years following the publication of Larmor. Raman eventually began to express these ideas, among others in a brief monograph in 1922 [61], where he brought up the relationship between the quantum mechanics of light and those of molecules.
As Raman explored and refined his idea of quantum mechanics in scattering, he also supervised students working experimentally, some of which began reporting unexpected scattering from liquids. From 1923 to 1925, two of his students, K. R. Ramanathan and K. S. Krishnan, observed scattering in liquids where the scattered light and the incident light had different color. This provided one of the earliest concrete answers to the deliberations of Larmor that primed Raman's investigations. Until early 1928, Raman and his students continued their experimental work, eventually ex-

panding the investigation from liquids to solids, including glass and ice. Having realized that they had discovered a significant new phenomenon, Raman wasted no time communicating their findings which were published in May of 1928[62]. With the observation that scattering interactions could result in wavelength shifts between the incident and scattered radiation, and that the amount of shift and intensity distributions varied significantly between chemicals, a new method of chemical analysis was born.

2.2 Fundamental mechanism

Molecules as elastic bodies

Molecules were previously considered as point masses with no internal structure of mechanics due to their small size. But while they are indeed very small, they are in fact not as rigid as classical physics consider them to be. Quantum mechanics has revealed that molecules are actually very complex and can have a number of fluctuations in their internal structure. Raman scattering is one of the physical consequences of this. From the structure of a particular molecule, for instance ethanol as depicted in fig. 2.1, we can consider a certain level of motion for each of the atoms in the molecule relative to the rest of the molecule. These vibrations can be described as a collection of harmonic oscillators with the stiffness of each oscillator given by the atomic bond it corresponds to. The motion of an oscillator can be considered sinusoidal at resonance: $x\left(t\right) = Acos\left(\omega t + \phi\right)$, with A denoting the amplitude of the oscillation, ω denoting the frequency of the oscillation, and ϕ denoting the phase of the oscillation. Such an oscillation will have a resonance frequency at $\omega = \sqrt{k/m}$ and a total energy of $E = k \cdot A^2$. In the model of the molecule, the stiffness k is given by the mutual electric fields of the atoms involved. The stiffness will also be directional as the fields, and thereby the reciprocal forces on the atom are anisotropic. This gives rise to multiple modes of molecular vibrations depending on the number of atoms involved and their configuration. For instance, the -CH_2- in the middle of the ethanol molecule shown in fig. 2.1 exhibits a scissoring mode, one of six modes of vibrations such a group can assume. As each mode interacts differently with the overall electric field of the molecule, each mode has a unique stiffness and thus a unique resonance frequency.

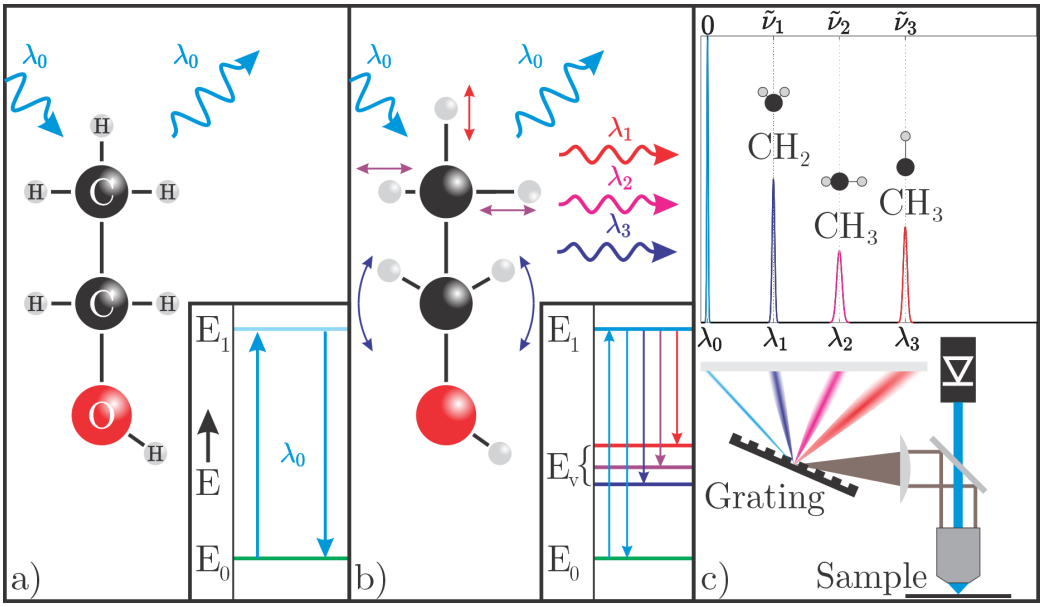


Figure 2.1: Conceptual sketch of the effects of molecular vibrations in a molecule of ethanol. a) Purely elastic interaction between the incoming photon and the electric field of the molecule, resulting in Rayleigh scattering. b) Mix of elastic and inelastic interaction between the photon and the molecule with three vibrational energy states (E_v), resulting in Raman scattering at three longer wavelengths. c) Conceptual sketch of a measurement system and the resulting spectrum showing the elastic (Rayleigh) scattering at wavelength λ_0 and the inelastic (Raman) scattering at longer wavelengths λ_1, λ_2 and λ_3.

Atomic groups as dipoles

While the mechanism behind Raman scattering can be conceptualized as a mechanical oscillator, the true nature of the mechanism is better described as a quantum harmonic oscillator formed by dipole fields in the molecule. In the Rayleigh regime, the entire molecule can be though of as a dipole that resonates with the exciting field. As illustrated in fig. 2.2, this interaction causes the electron cloud to "swing" around the molecule, inducing a global dipole in the molecule. This perturbation then resonates with the exciting field and produces a wave with the same frequency. In Raman scattering, the perturbation of the electron cloud upsets the equilibrium of the molecule, this results in the perturbation of the structure of the molecule as it responds to the shift in the electron field. The shift in the structure of the molecule then creates a local asymmetry in parts of the molecule. Due to differences in electronegativity, certain atoms (such as carbon) attract electrons to a higher degree than other atoms (such as hydrogen). Thus, even though the electrons are shared between them in a covalent bond, the electrons are more likely to be near the atom with the higher electronegativity. When a local asymmetry occurs, this effect makes one side of the pair/group more negative than the other, and thus a dipole is induced. Compared to the Rayleigh dipole, the Raman dipole typically has a much lower dipole moment and resonates at frequencies much lower than the exciting field.
To evaluate the effect of the dipoles on the electric field, and thus on the emitted radiation, we must first quantify the magnitude of the dipole moments. The dipole moment $\mathbf{d}$ induced by an external field $\mathbf{E}$ depends on the polarizability of the particle(s) α:

$$\mathbf{d} = \alpha \mathbf{E}. \tag{2.1}$$

In a true three dimensional system, the polarizability takes the form of a 3x3 tensor:

$$\alpha = \begin{bmatrix} \alpha_{\mathbf{xx}} & \alpha_{\mathbf{xy}} & \alpha_{\mathbf{xz}} \\ \alpha_{\mathbf{yx}} & \alpha_{\mathbf{yy}} & \alpha_{\mathbf{yz}} \\ \alpha_{\mathbf{zx}} & \alpha_{\mathbf{zy}} & \alpha_{\mathbf{zz}} \end{bmatrix}, \tag{2.2}$$

where $\alpha_{\mathbf{vu}}$ denotes the polarizability of the particle along the unit vector $\mathbf{v}$ when subjected to an electric field with polarization $\mathbf{u}$. In a quantum harmonic oscillator, like a Raman-active molecule, the polarizability is additionally dependent on the state that the molecule is in and will be

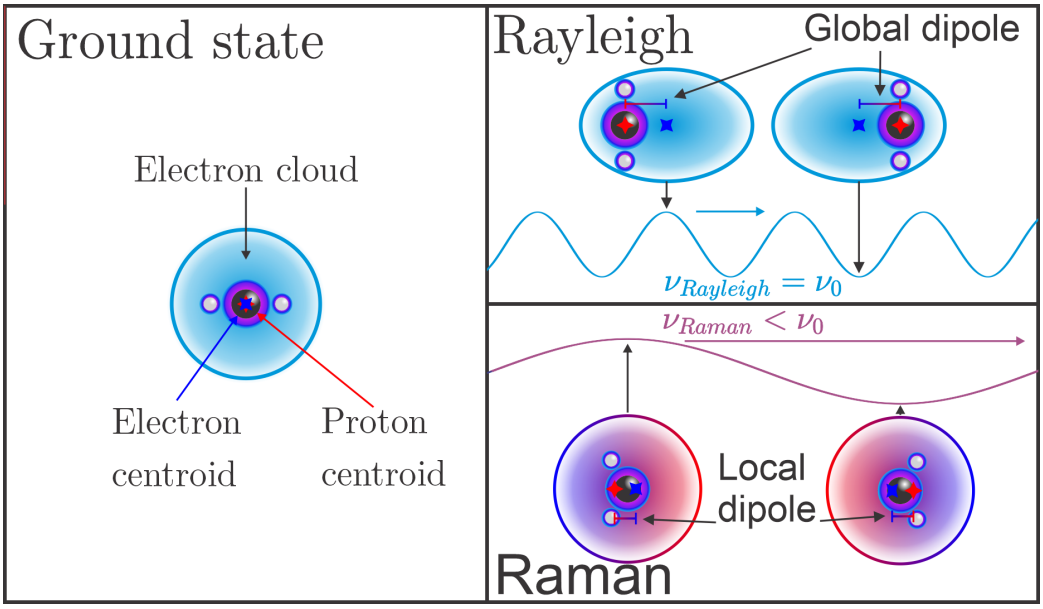


Figure 2.2: Conceptual sketch of the dipole effects in Rayleigh and Raman scattering. In Rayleigh scattering, the electron cloud of the molecule is perturbed by the exciting field, creating a dipole with a given moment. In Raman scattering, the perturbation of the electron cloud creates perturbations in the structure which makes the molecule locally asymmetrical. This also creates a local dipole which resonates at a lower frequency and has a smaller moment than the global dipole.

in after it has been polarized. In a simplified setting, we assume that the molecule is at rest in its ground state, denoted as $|0\rangle$, in the beginning of the process. As a photon of energy $\hbar\omega_I$ interacts with the molecule, the molecule transitions to a virtual vibronic state $|n\rangle$ of higher energy. The molecule will then transition back to a lower energy state $|f\rangle$ and emit a photon of energy $\hbar\omega_S$. This can be described using a rewrite of the Kramers-Heisenberg formula for a second order system[53]:

$$[\alpha_{\mathbf{vu}}]_{f0} = \frac{1}{\hbar} \sum_{|n\rangle} \left[\frac{\langle f|\mu_{\mathbf{v}}|n\rangle\langle n|\mu_{\mathbf{u}}|0\rangle}{\omega_{n0} - \omega_I - i\gamma_n} + \frac{\langle f|\mu_{\mathbf{u}}|n\rangle\langle n|\mu_{\mathbf{v}}|0\rangle}{\omega_{n0} + \omega_S + i\gamma_n} \right], \tag{2.3}$$

which describes the polarizability for a transition from the ground state $|0\rangle$ to a single vibrational state $|f\rangle$. The full expression of the polarizability given in eq. 2.3 normally then simplified through approximation. Under the assumption that the excited state lifetime γ_n is much longer than the period of either of the frequencies involved, the complex term $i\gamma_n$ can be neglected. Furthermore, we can assume that the photon energy of the incident field is larger than, or equal to, the band-gap energy of the molecule such that the intermediate state $|n\rangle$ is always an excited electronic state, thus we can omit summation over the ground states in eq. 2.3. Lastly we can assume that the transfer to and from the intermediate state $|n\rangle$ is equal to the electronic transfer to and from that state such that the transition frequencies $\omega_n 0$ and $\omega_f n$ can be said to equal to the electronic transition frequencies $\omega_n^e 0$ and $\omega_f^e n$. Using these three approximations, and by using the Born-Oppenheimer approximation to separate the motion of the nuclei from the electrons, the polarizability described in eq. 2.3 can be approximated as:

$$[\alpha_{\mathbf{vu}}]_{f0} = \langle f|\alpha^e_{\mathbf{vu}}(\omega_I; Q)|0\rangle, \tag{2.4}$$

where $\alpha^e_{\mathbf{vu}}(\omega_I; Q)$ denotes the frequency dependent electic polarizability operator for the molecule given the normal coordinates Q. In polyatomic molecules, α^e is commonly expressed by its Taylor expansion to the linear order:

$$[\alpha_{\mathbf{vu}}]_{f0} = \sqrt{\frac{\hbar}{2\omega_{|f\rangle}} \left(\frac{\partial \alpha^e_{\mathbf{vu}}}{\partial Q_{|f\rangle}} \right)}, \tag{2.5}$$

where $Q_{|f\rangle}$ is the normal mode coordinates of the vibrational mode corresponding to the state $|f\rangle$. In discrete terms, this can be transformed into

the sum operation:

$$\alpha_{\mathbf{vu}} = \sum_{i=1}^{3N-6} \left[\frac{\hbar}{\sqrt{2\omega_i}} \frac{\partial \alpha^e_{\mathbf{vu}}}{\partial Q_i} \right], \tag{2.6}$$

such that the Raman spectrum can be expressed where each Raman mode Q_i exists at an energy level given by ω_i with an intensity given the polarizability $\alpha^e_{\mathbf{vu}}$ depending on that state. The Raman spectrum $I(\tilde{\nu})$ can therefore be seen as expressed by:

$$I(\tilde{\nu}) = I_0 \left[\delta(\tilde{\nu}) + \sum_{i=1}^{3N-6} A_i \delta(E_i - hc_0\tilde{\nu}) \right], \tag{2.7}$$

where mode i occurs at at energy E_i, dependent on ω_i, with an amplitude of A_i depending on the polarizability as well as the density and Raman cross-section of the sample. Though, in the practical case, the observed spectrum will be the result of convolution with the spectral distribution of the incident field and convolution of the Raman mode states with diffuse states, resulting in a smooth, continuous expression.

2.3 Conventional methods

2.3.1 Raman imaging

One of the typical applications of Raman spectroscopy is as an imaging technique, analagous to fluorescence microscopy as shown in fig. 2.3. Fluorescence imaging normally acquires the fluorescent light using band-pass filters in the imaging system, only allowing the known emission wavelengths of one specific fluorophore through to the camera sensor at one time. While a similar approach can be used for Raman imaging, there are other considerations that must be made when attempting to image the Raman scattering, such as the signal strength and the number of wavelengths to be probed. Because of this, Raman imaging often employs a much more comprehensive hyperspectral imaging scheme to acquire the necessary signal to make chemical determinations. In Raman imaging, as shown in fig. 2.4, the images are instead collected as a more complete stack of images, where each image is an image of the sample plane for one specific wavelength. Due to the intrinsically weaker Raman scattering in comparison to fluorescence, the exciting field must be made significantly stronger. One common

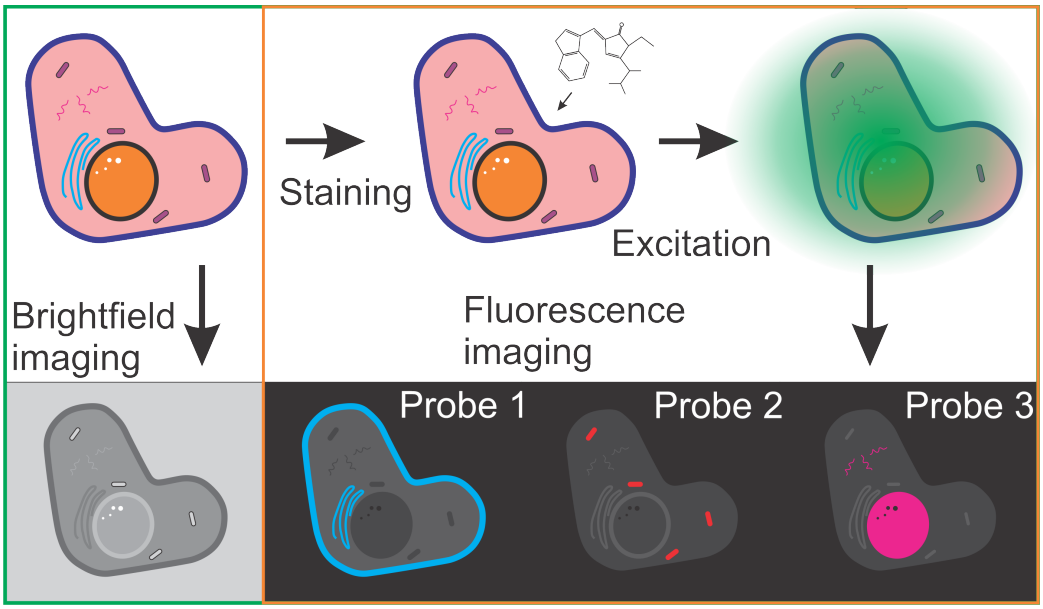


Figure 2.3: Conceptual sketch of fluorescence imaging of a cell. The enhanced contrast and specificity of fluorescence imaging compared to brightfield imaging is shown in the bottom row.

method of achieving this without bleaching or otherwise damaging the cell is to use a point scan method. In this approach, a confocal microscope configuration is often used. The use of a confocal system allwos for the excitation of and collection from volumes in the sample plane with sizes down to the diffraction limit. The high out-of-focus rejection performance of a confocal geometry also improves contrast in the images while also directing the exciting field such that the delivered dose in the cell is highly localized. Acquiring the images on a pixel-by-pixel basis also makes the configuration flexible in terms of resolution, allowing for the images to be acquired with any desired resolution down to the diffraction limited resolution of the system. This also makes it possible to use an arbitrary method of collecting the spectral information, enabling the use of high-performance spectrometers for this application such that the spectra can be collected over a wide range and with sub-nanometer spectral resolution. A significant drawback in the use of confocal Raman microscopy, similar to conventional confocal microscopy, is that the point scanning method is much slower than wide-field imaging and is therefore not well suited to live samples. This can be remedied to a degree by hybridizing the point scanning method of confocal Raman with the methods employed with fluorescence imaging, for instance by limiting the spectral range that is aquired such that more light can be collected per pixel per time, allowing for shorter acquisition times. Other, more advanced methods of exciting Raman scattering can also be employed to increase the signal strength to facilitate wide-field Raman imaging.

2.3.2 Plasmonic enhancement

One of the more advanced methods of exciting Raman scattering is through the use of plasmonic enhancements in the sample plane. In such a case, a surface is partially or completely clad in a material that can support surface plasmons, such as gold or silver, against which the sample is placed. When such a surface is exposed to an incident field with a frequency below the plasma frequency, the incident photons couple with the electron, producing oscillations in the cloud and forming surface plasmons. The oscillations of the electron cloud near the surface thus creates a very strong electric field in the near-field region around the plasmonic material, thus the Langmuir waves create a highly localized volume with significant field enhancement. This field can then couple to other molecules and atoms similar to how it would in a far-field region and thus also excite effects such

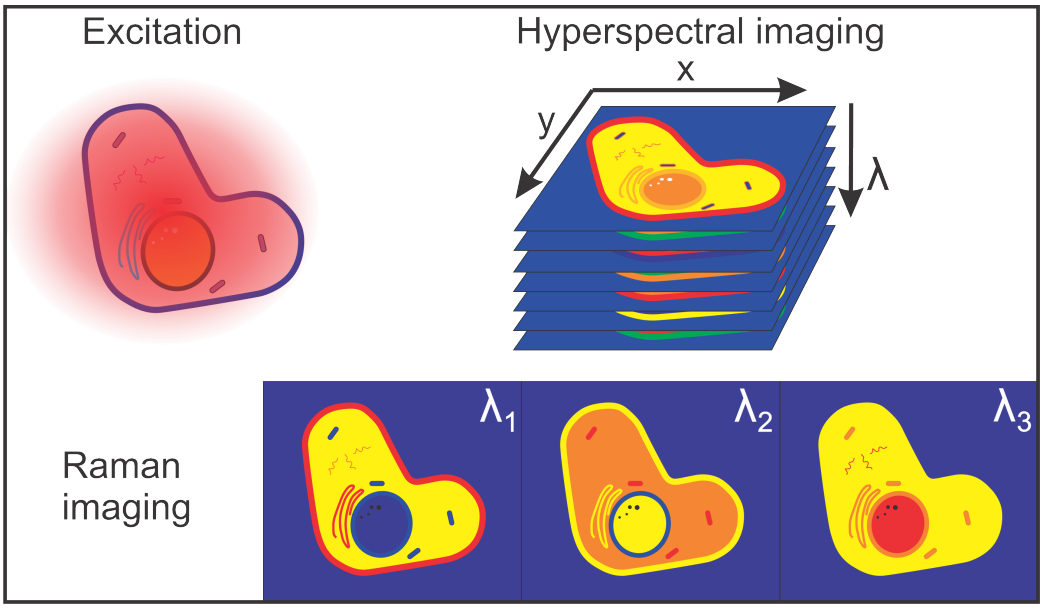


Figure 2.4: Conceptual sketch of Raman imaging of a cell. The unstained cell is exposed to a high intensity monochromatic light to excite Raman scattering. The images are then collected as a hyperspectral stack with each pixel of the stack forming a comprehensive spectrum corresponding to the Raman scattering in a small volume of the image plane.

as Raman-scattering. The principle of Surface Enhanced Raman Spectroscopy (SERS), as shown in fig. 2.5, is to exploit this property to produce enhanced Raman signals from a sample material, allowing for a lower detection threshold and purer signals from the material. Because of the increase in field strength surrounding the plasmonic material, the generated Raman scattering will also be increased up to a factor of 10^{12}[60], but due to the use of surface plasmons as an intermediary, the increase in signal will not be uniform across the spectrum. Since the field that interacts

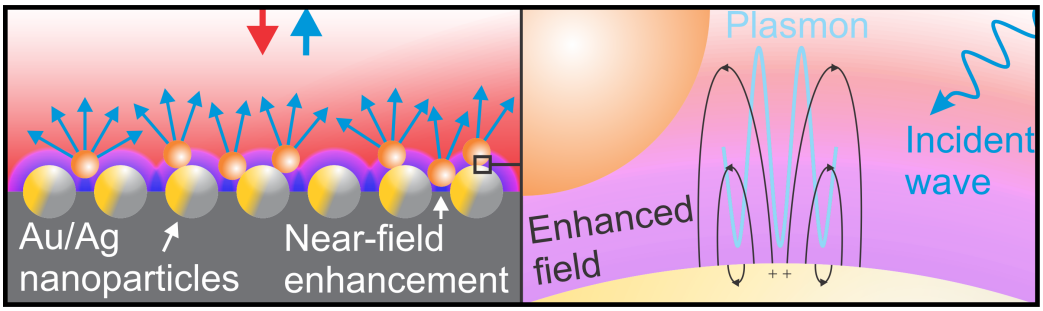


Figure 2.5: Conceptual sketch of plasmonic enhancement. The incident wave induces oscillations in the electron cloud of a metallic nanoparticle, producing a surface plasmon. The plasmon localizes the energy of the incident wave and acts as an enhanced field, coupling the incident wave to the sample particle in the near-field.

with the analyte is almost exclusively from the plasmons, the strength of the exciting field will depend on the enhancement of the incident field, which depends on the characteristics of the incident field and the characteristics of the surface plasmons. The greatest field enhancement occurs when the frequency of the incident field becomes equal to the natural plasmon frequency of the material, thus adding a consideration to which excitation wavelength to use and which plasmonic material to use. Another complication is the fact that the generated Raman scattering must also interact with the surface plasmons before the generated waves can couple back to the far-field region, thus making the enhancement of the field vary across the spectrum. This means that certain regions of the Raman spectra, close to the plasmon frequency, will be significantly more enhanced than other regions, resulting in the spectra becoming distorted. This can make it challenging to compare the results of SERS with results

from other, non-enhancement methods such as confocal Raman microspectroscopy and can give the illusion of a higher presence of certain chemical features while the higher amplitude is actually caused by fluctuations in the enhancement.

2.3.3 Stimulated scattering

In addition to enhancing the exciting field, as is done in SERS, the generation of Raman scattering can also be enhanced. As discussed in chapter 2.2, the Raman modes can be considered oscillators that are powered by the incident field, and like most oscillators, they can be driven either by random perturbations or deliberately by a resonating source. The latter is the principle of stimulated Raman scattering (SRS), where two waves are used to deliberately drive the oscillators, thereby generating more Raman scattering. One effect of the stimulated driving compared to the spontaneous driving of the mode(s) is that the modes oscillate in a coherent manner in the stimulated case, resulting in them having a uniform phase and polarization. These effects, and the increased directionality of the generated light, allows SRS to generate signals several orders of magnitude stronger than spontaneous Raman spectroscopy.
While the SRS effect can be invoked using a single wavelength pair, as shown in fig. 2.6, with great effect on the signal amplitude, doing so only reveals the amplitude of one specific vibrational mode with an energy equal to the difference between the photon energies. This is one drawback, as spontaneous excitation produces a response for all modes while stimulated excitation must use light sources that deliberately probe the spectrum. The probing of the frequency spectrum can be achieved using two main principles:

- Narrowband probing: Using at least one tunable laser source, either as stokes or pump source, such that the energy difference can be swept over the desired range.
- Broadband probing: Using one narrowband source, usually the pump, and one broadband source, usually as a stokes probe, to probe the spectrum in one shot.

In the narrowband setting, the fundamental method is to use a fixed continuous wave pump beam and a tunable continuous wave stokes beam that

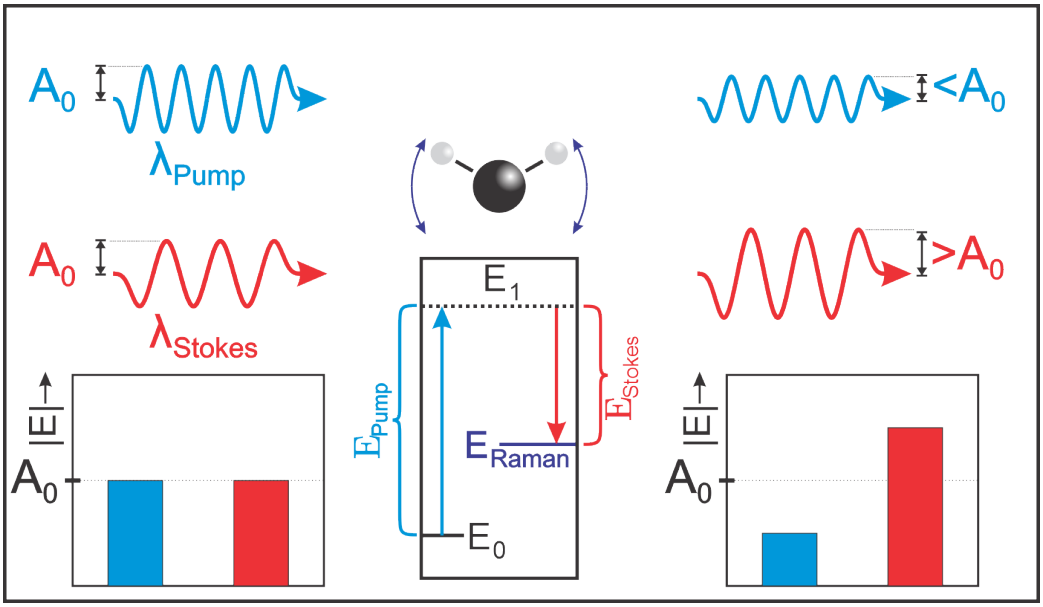


Figure 2.6: Conceptual sketch of SRS. As the photon energy difference between the incident pump field and stokes field becomes equal to the energy of a Raman mode, the transition between the vibrational ground state and the Raman active state is strongly driven. The driving of the transitions is observed by an energy transfer between the pump and stokes beams in SRS or from the stokes to the pump in CARS.

intersects in a common volume in the sample. In order to separate the two, at least one beam is modulated such that the output light intensity can be measured using a photodiode, or similar device, when both pump and stokes are active, and when only one of the beams, normally the pump, is active. The gain of the signal can be improved can be further improved by using a pulsed source, such as a picosecond laser source [68] modulated at megahertz frequencies, such that more excitation power is delivered to the sample volume within each modulation period and thus generating more signal. To maintain coherence and synchronize the pulses, both the pump and stokes beams are often generated from the same picosecond source using a non-linear optical device, such as an optical parameteric oscillator[78], which often provides the tuning of the stokes beam as well. Even though these methods can be used to accelerate the scanning of multiple Raman modes, narrowband SRS still suffers from slow acquisition

since it is still limited to a spectral point scan method. The requirement of high quality laser pulses with fixed field distributions also necessitates the use of high quality picosecond lasers and optical modulators, which are often specific use products and are relatively expensive in comparison to continuous wave lasers.

By comparison, in the broadband setting, the excitation does not require a tunable source, but requires that one of the sources, commonly the stokes, produces a consistent broadband spectrum. As with the narrowband setting, the spectrum is detected by distortions in the stokes beam while using modulation of the pump beam to separate the stokes beam spectrum and the stimulated Raman spectrum. This is often achieved by using a broadband femtosecond laser as the stokes probe and using a picosecond narrowband source as the pump source to achieve broadband femtosecond stimulated Raman scattering (FSRS)[39]. The excitation of the Raman modes is primed by the picosecond pump beam before the stimulated emission is driven by the broadband femtosecond stokes beam. Because of the relatively "long" picosecond duration of the pump beam in comparison to the "short" femtosecond stokes pulse, the modes excited by the pump are allowed to interact with the full spectrum of the pump before the vibrational modes dephase and degrade. The Raman spectrum of the sample is then observed by distortions in the probe spectrum, and can be quantified by observing the spectrum of the probe in the presence of the pump and in its absence such that the pure spectrum can be isolated from the stokes probe spectrum. Compared to the narrowband setting, this gives much more information in each acquisition, as the entire spectrum is acquired in one shot, while still exploiting the signal gain given by the stimulated scattering effect. One downside of this approach, in addition to the additional cost associated with having both a picosecond and a broadband femtosecond laser source, is that measuring the Raman spectrum requires a full fledged spectrometer, while the narrowband excited SRS can be measured using only a point source such as a photodiode with a lock-in amplifier.

These approaches can also be combined with other techniques such as surface enhanced FSRS [23] and can be made to exploit the temporal aspects of the vibrational mode transitions to extract additional information. Again, FSRS is an example of this, where the difference in pulse length between the pump and stokes beams makes the output spectrum vary in time (see figs. 3 and 5 in [39]). This allows the extraction of more data from

each measurement, making the combination more powerful in collecting information, particularly regarding the internal mechanics of the vibrational mode transitions and lifetimes [49].

2.4 Optical trapping

As mentioned earlier, one of the principal advantages of Raman spectroscopy in comparison to other chemical analysis methods is that Raman spectra can be acquired from significantly smaller volumes. The smallest volume is, in the general sense, only limited by the diffraction limit of the collection system. The diffraction limit can readily be made smaller than 300 nm in diameter with Rayleigh lengths shorter that 500 nm, resulting in volumes smaller than 0.05 μm^3. However, most results and works don't exploit the potential of Raman spectroscopy when it comes to single particles in the micron regime and below. To exploit this potential, both the exciting field and the collection spot must be diffraction limited in the same spot, which can be achieved with a confocal Raman microscope. By combining a confocal design with optical trapping, single particles of sub-micron size can be confined, isolated, and measured with accuracy.
In optical trapping, the optical field itself is used to affect a force onto a particle such that it can be isolated from the other particles in the analyte, and thus allowing it to be measured separately from the rest. The principle of optical trapping revolves around the interaction between four forces on the particle:

- Brownian motion
- Drag forces from flow
- Optical forces

Browninan motion is a well known phenomenon where a particle spontaneously moves as a result of its temperature. The direction of the spontaneous motion is random at each instance, and the distance of the motion Δx is a stochastic quantity described by a Gaussian probability distribution:

$$\Delta x = \mathcal{N}\left(0, \frac{k_B T}{3\pi r \eta}\Delta t\right), \tag{2.8}$$

where k_B denotes Boltzmann's constant, T denotes the temperature, η denotes the viscosity of the surrounding medium, and r denotes the particle size. Thus, the smaller the particle, and the higher the temperature, the more intense the Brownian motion will become. To achieve stable trapping of a particle subject to Brownian motion, the trapping force must be sufficient to overcome the Brownian motion. Thus, to attain a stable trap, the total potential of the trap must be at least ten times the kinetic energy of the Brownian motion [4]:

$$U(\mathbf{r}, r) \geq 10 k_B T. \tag{2.9}$$

In addition to Brownian motion, a particle suspended in a medium is also subject to the flow of the medium, which must also be considered. Flow in the suspension medium can either be deliberately induced, such as through microfluidics, or it can be accidental, such as thermally induced convective flow. Ideally, the flow in a trapping system is zero, but if there is flow then the drag force $\mathbf{F}$ must be accounted for:

$$\mathbf{F} = 6\pi\eta r\mathbf{v}, \tag{2.10}$$

where η denotes the dynamic viscosity of the medium, r denotes the particle size, and $\mathbf{v}$ denotes the local flow of the medium. Optical traps typically give a trapping force of less than 1 pN/mW[52]. Thus, 1 μm sphere suspended in water and trapped using a field delivering 10 mW of power can tolerate up to 260 $\mu m/s$ of flow.
The optical forces acting on the particle must thus be strong enough to overpower the force of flow in the suspension medium and provide a potential well significant enough to counteract the Brownian motion of the particle. In terms of the interaction between the field and the particle, the description of the interaction is determined by the refractive index and size of the particle. For particles whose size is comparable to the wavelength λ of the field, Mie theory is applicable to describe the interaction. However, in the case of EVs, with a typical radius r of 50 nm[59], the particles are significantly smaller than the wavelengths useful in Raman scattering (¿500 nm). Thus, with a relatively low refractive index n_p of 1.4[25], they fulfill the requirement of a Rayleigh particle:

$$r << \frac{\lambda}{4\pi|n_p - 1|}, |n_p - 1| << 1. \tag{2.11}$$

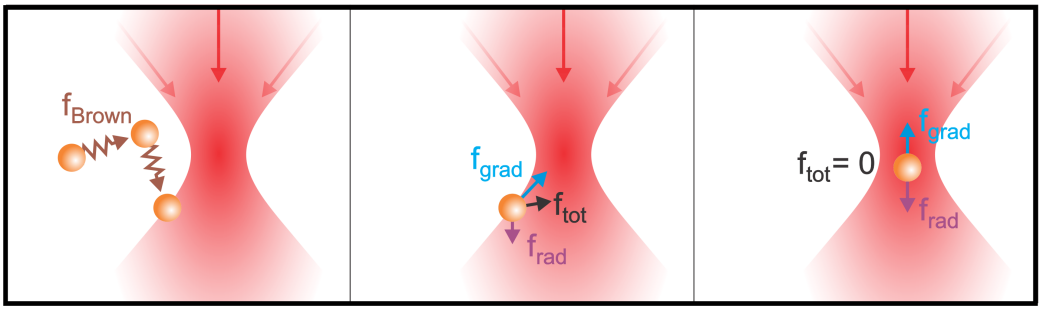


Figure 2.7: Conceptual sketch of optical trapping. A particle is suspended in a medium, where it randomly moves due to the force of Brownian motion $\mathbf{F}_{Brown}$. Once the particle approaches the beam, it will be affected by the radiation force $\mathbf{F}_{rad}$, pushing the particle along the optical axis, and the gradient force $\mathbf{F}_{grad}$, pushing the particle towards the focus of the beam. If the gradient force $\mathbf{F}_{grad}$ is sufficiently strong, the particle moves towards a a stable trapped position close to the beam waist.

In contrast to Mie theory, a particle subject to Rayleigh theory scatters homogeneously from its surface. The interaction between the field and the particle surface can thus be considered isotropic and uniform for a Rayleigh particle.
The optical forces acting on a Rayleigh particle can be divided into two forces: the scattering force and the gradient force. The scattering force occurs due to the incident field being scattered by the boundary of the particle, thus a force will be applied directly on the particle by the incident field along the fields propagation vector. The scattered radiation will then diverge based on the properties of the surface, resulting in further force being affected on the particle. The net force depends on the scattering geometry of the particle, and may not be symmetrical about the optical axis in cases where the scattering geometry on the particle is not symmetric.
For a Rayleigh particle of radius r with a refractive index n_p suspended in a medium with a refractive index n_m the scattering force $\mathbf{F_{scat}}$ can be described[30] as:

$$\mathbf{F}_{scat}(\mathbf{r}) = \frac{8\pi n_p k^4 r^6}{3c_0} \frac{(n_p/n_m)^2 - 1}{(n_p/n_m)^2 + 2} I(\mathbf{r})\,\widehat{\mathbf{z}}, \tag{2.12}$$

when subjected to a field $\mathbf{I}(\mathbf{r})$ propagating along a vector $\widehat{\mathbf{z}}$ with a wavevector $\mathbf{k}$. This requires that the refractive index of the particle n_p be greater

than the refractive index of the medium n_m. For a focused field, the plane wave field $\mathbf{I}(\mathbf{r})$ with a uniform energy flux along the propagation vector $\widehat{\mathbf{z}}$ is replaced by a field $I(\mathbf{r},\mathbf{v})$ with a directional energy flux along $\mathbf{v}$. The energy flux is given by the averaged Poynting vector $\langle \mathbf{S}(\mathbf{r},\mathbf{v})\rangle = 0.5Re\left[\mathbf{E}(\mathbf{r},\mathbf{v}) \times \mathbf{H}^{*}(\mathbf{r},\mathbf{v})\right]$. For a focused Gaussian beam with a beam waist w_0, the flux through the focal plane becomes $I_0 = P/\pi w_0^2$ for a given power P and the time-averaged Poynting vector in the focal plane becomes:

$$\langle \mathbf{S}(\mathbf{r},\mathbf{v})\rangle = \frac{P}{w_0^2} e^{-2|r|^2/w_0^2}\widehat{\mathbf{z}}, \tag{2.13}$$

such that the scattering force from a focused Gaussian beam can be expressed as:

$$F_{scat}(\mathbf{r},\mathbf{v}) = \frac{8\pi n_p k^4 r^6}{3c}\frac{m^2-1}{m^2+2}\frac{P}{w_0^2} e^{-2|r|^2/w_0^2}\widehat{\mathbf{z}}, \tag{2.14}$$

at the focal plane where the propagation vector $\mathbf{v}$ becomes equal to the optical axis vector $\widehat{\mathbf{z}}$.
For a non-uniform beam, the particle will also be subject to a force acted upon it by the gradient of the beam. The principle of the gradient force is that a dielectric particle, behaving as a dipole, will be attracted up the gradient of the optical field, pulling it towards the point of highest intensity which occurs at the focus of the beam. The gradient force acting on a Rayleigh particle can be described[30] by the expression:

$$\mathbf{F}_{grad}(\mathbf{r},\mathbf{v}) = \frac{2\pi n_p r^3}{c}\frac{(n_p/n_m)^2-1}{(n_p/n_m)^2+2}\nabla I(\mathbf{r},\mathbf{v}), \tag{2.15}$$

where $\nabla I(\mathbf{r},\mathbf{v})$ is the gradient of the field intensity at position r along direction v. When the refractive index n_p of the particle is greater than the refractive index n_m of the medium, $(n_p/n_m)^2 - 1$ becomes positive, thus the force $\mathbf{F}_{grad}$ will push the particle up the gradient towards the highest intensity. The total optical force on a Rayleigh particle will therefore be the sum contribution of the scattering and gradient forces:

$$\mathbf{F}_{\Sigma}(\mathbf{r},\mathbf{v}) = \mathbf{F}_{grad}(\mathbf{r},\mathbf{v}) + \mathbf{F}_{scat}(\mathbf{r},\mathbf{v}). \tag{2.16}$$

In the absence of other forces, the particle will thus settle into a stable position where $\mathbf{F}_{\Sigma} = 0$, as shown in fig. 2.7. meaning that the point where a particle settles into a stable position is where the two forces cancel each

other out. Since the scattering force always acts along the direction of the local wavefront, i.e. $\widehat{\mathbf{z}}$ near the beam waist, and the gradient force will always attract the particle to the focus of the beam, we see that the position where this occurs must be on the optical axis and slightly away from the beam waist such that the gradient force "pulls" on the particle as much as the scattering force "pushes" on the particle. Negating all other forces, a Rayleigh particle acted upon by a focused gaussian beam will therefore settle at a point in the field where the gradient force $F_{grad}(\mathbf{r},\mathbf{v})$, drawing the particle towards the focus, and the scattering force $F_{scat}(\mathbf{r},\mathbf{v})$, pushing the particle away from the focusing aperture, become equal:

$$F_{grad}(\mathbf{r},\mathbf{v}) = -F_{scat}(\mathbf{r},\mathbf{v}), \tag{2.17}$$

which allow us to substitute in Eqs. (2.15) and (2.12):

$$\frac{2\pi n_p r^3}{c}\frac{m^2-1}{m^2+2}\nabla I(\mathbf{r},\mathbf{v}) = \frac{8\pi n_p k^4 r^6}{3c}\frac{m^2-1}{m^2+2} I(\mathbf{r})\,\widehat{\mathbf{z}}, \tag{2.18}$$

which can be simplified into:

$$\nabla I(\mathbf{r},\mathbf{v}) = \frac{4k^4 r^3}{3} I(\mathbf{r})\,\widehat{\mathbf{z}}. \tag{2.19}$$

As the field is known to be Gaussian, the field can be expressed as:

$$I(\mathbf{r}) = \left(\frac{2P}{\pi w_0^2}\right)\frac{1}{1+(2\tilde{z})^2} e^{\frac{-2\left(\tilde{x}^2+\tilde{y}^2\right)}{1+(2*\tilde{z})^2}}, \tag{2.20}$$

using the normalized coordinates $\tilde{x} = x/w_0$, $\tilde{y} = y/w_0$, and $\tilde{z} = z/kw_0$ for a known beam waist w_0, wave vector k and power P. Under the assumption that the stable site will occur at $\tilde{x} = \tilde{y} = 0$, Eq. (2.19) can be written as:

$$-\frac{8\tilde{z}/(kw_0^2)}{1+(2\tilde{z})^2}\frac{1}{1+(2\tilde{z})^2}\left(\frac{2P}{\pi w_0^2}\right)\widehat{\mathbf{z}} = -\frac{4k^4 r^3}{3}\frac{1}{1+(2\tilde{z})^2}\left(\frac{2P}{\pi w_0^2}\right)\widehat{\mathbf{z}}, \tag{2.21}$$

which yields the trap position:

$$\tilde{z} = \frac{k^5 w_0^2 r^3}{3-4k^5 w_0^2 r^3}. \tag{2.22}$$

This imposes a limit $k^5 w_0^2 > 3/4r^3$ for trapping to occur in the absence of other forces. As mentioned earlier, the forces of the optical trap must also be stronger than the sum force of the flow and the Brownian motion for trapping to occur. As stated, the potential of the trap $U = \int_{\mathbf{r}} F_\Sigma(\mathbf{r},\mathbf{v})\,d\mathbf{r}$ is generally required to be larger than $10k_bT$ for the trapping to be stable.

Chapter 3

Waveguide devices for trapping and Raman spectroscopy

While what we now think of as waveguides are a relatively new phenomenon, the concept of them has been discovered multiple times through history, going as far back as the notion of an acoustic waveguide dating back to the 1700's. The first demonstration of optical guiding in a material in 1842 by Swiss physicist Jean-Daniel Colladon [13] and the first waveguides began slowly coming into existence through the 1930's [54]. From there, interest in optical guiding saw a resurgence in the 1960's with the invention of the laser and the following interest in using optical guiding for communication purposes. The modern notion of the dielectric optical waveguide was presented in 1966 by Charles Kao[36], where a cylindrical glass waveguide was developed and the guided modes were observed.
In modern waveguides, the scale, losses, and cost have all been significantly reduced, getting well into the nanometer scale and allowing for the construction of controllable optical devices only tens of micrometers across [2]. This has allowed a multitude of integrated optics devices, from comparatively simple devices, such as Mach-Zehnder interferometers [24], to complex devices, such as spectrometers [79], and even devices to facilitate super-resolution microscopy [31]. This development has also enabled the expansion of waveguide based sensors for spectroscopy, such as infrared spectroscopy [72] and its neighbour, Raman spectroscopy.
In this chapter, the concept of waveguide enhanced Raman spectroscopy (WERS) is explored along with the most common techniques used to achieve this. The prospect of using waveguides to facilitate optical trapping is

also be introduced and how this can be used in conjunction with WERS-techniques. The challenge of Raman-background in waveguides, being a significant limitation in WERS, is explored and experimentally evaluated to determine a viable waveguide material. Lastly, the design of a multi-trap Raman-on-chip device is considered and evaluated.

3.1 Waveguide enhanced Raman spectroscopy

Many methods of implementing waveguide enhanced Raman spectroscopy (WERS) have been investigated[20, 70, 73]. Two types form the bulk of the approaches to the technique:

- Evanescent field interaction
- Structure on chip

These two approaches differ in their goals, with evanescent field interaction focusing on relatively large sample volumes to achieve high sensitivity while using a structured chip design generally focuses on small volumes to achieve high selectivity.

Evanescent field interaction

The practice of using the evanescent field of a waveguide to interrogate an analyte is the most widely used method of WERS. The principle of the method, as shown in fig. 3.1, is that the waveguide projects its evanescent field into the analyte, where it interacts with the matter and its vibrational modes. The generated Raman scattering from this is then coupled back into the waveguide, causing the incident field and the fields containing the Raman signal to co-propagate in the waveguide. The principal advantage of using the evanescent field to interrogate the analyte is the significant increase in the interaction length, and thus the interaction volume. While a focused beam can achieve much higher intensities and can collect the generated scattering more efficiently by producing a tight envelope, the same factor also limits the interrogation volume and thus the signal. Thus, by replacing the conventional lens-based path, with a spot volume of a few cubic micrometers, with a waveguide, the length of the interaction path can be increased by several orders of magnitude. Using a high refractive index step in the waveguide design, the waveguide can be made to project a

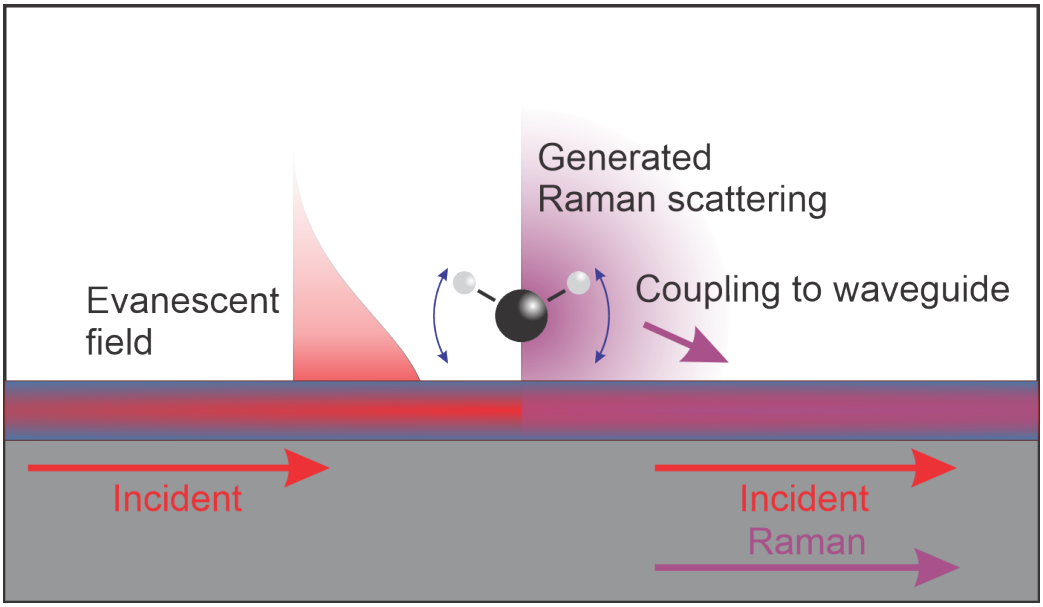


Figure 3.1: Conceptual sketch of evanescent field WERS. The evanescent field of the waveguide interacts with the analyte, generating Raman scattering. The scattered light is then partially coupled back into the waveguide.

strong evanescent field several hundred nanometers away, and by giving the analyte access to the space both above, on the sides, and sometimes below the waveguide, the interaction cross section can readily be made well over a square micrometer, allowing the excitation and collection from a large volume surrounding the waveguide. Thus, with only one millimeter of exposed waveguide, the interaction volume can easily be made to be as much as 1 000 cubic micrometers, increasing the volume by two to three orders of magnitude in comparison to a lens-based system. Since the Raman scattering of each interaction couples back into the waveguide where it propagates with the incident field, the total signal increases linearly with each interaction, thus with the length of its path. Furthermore, another advantage of using a high refractive index waveguide is that it allows for the path to be curved across the device. Curving the path, or even making it into a spiral pattern, makes it possible to further increase the path length, from a few millimeters to several centimeters[16].
Another benefit of exciting and collecting Raman scattering at the sur-

face of a waveguide is the Purcell effect which occurs due to coupling of the quantum oscillations of the analyte with those of the near field of the waveguide, producing a local enhancement of the spontaneous emission rate of the Raman scattering[15]. Thus, even with a reduction in field intensity due to the relatively weak evanescent field, the overall signal can be enhanced by as much as 500 times per centimeter of waveguide compared to a confocal system[16].
However, while the potential of using this setting of waveguide enhancement is very high, there are also significant challenges to it. One of the principal challenges is regarding the loss in the waveguide as both the incident and Raman fields propagate along it. As stated previously, the enhancement theoretically scales linearly with length, but it also scales with the intensity of the field, and when there is loss in the waveguide the field will decrease exponentially with length. This is also the case for the Raman fields that propagate in the waveguide, which will eventually also decay towards zero with length. Thus, the loss in the waveguide gives a fundamental limit to the length and achievable enhancement as the exponential decay eventually becomes larger than the linear enhancement. One solution to this is to operate the waveguide in back-propagation mode, by reading out the Raman signal from the facet where the incident light couples in. While this does not overcome the limit, it allows the waveguide to be longer than the loss limit without risking that the propagation loss compromises the signal.
Another significant issue with using WERS, both in this setting and in others, is that the waveguide itself will also produce Raman scattering, which will co-propagate with the incident and the sample Raman in the waveguide. This signal will thus produce a significant background in the measurements, and will need to be corrected for in post-processing. This challenge will be further addressed later in this thesis.

Structure on chip

Another approach to enhancing the Raman signal is to use more deliberate designs to invoke a stronger Raman response. Plasmonic structures can be fabricated on the chip in order to localize the field and produce nanometer scale near field enhancement[37]. This combination of plasmonic enhancement and WERS enables the best of the two to be exploited more fully. By using the long interaction length afforded by the waveguide, as well as

the high field enhancement given by the plasmonic effects, such devices can thus detect substances with an even lower detection threshold than either one alone[28, 57]. However, in this setting the weaknesses of both methods build on top of each other as well, resulting in the resonance distortion and the transmission/coupling distortion with wavelength compounding on each other, and may easily produce false impressions from the spectra. The increased interaction with the plasmonic structures in this setting also significantly increases the potential for heating along the waveguide surface. The plasmonic enhancement is also non-selective, the result being that any material in close proximity to the plasmonically active material will also experience an enhancement of its Raman scattering. Unless careful design is used, this can result in undesired elements, such as the waveguide material surrounding the plasmonic structure, contaminating the spectrum. Plasmonic enhancements are most commonly achieved by depositing nanoparticles of a material that forms surface plasmons, commonly gold or silver, onto a substrate, which in this case is a waveguide as shown in fig. 3.2. This can be enhanced further by creating more deliberate structures on, in or around the waveguide, such as nanoantennea[1] or metasurfaces[5]. The use of precision electron-beam lithography also enables the construction of such structures with much more predictable geometries, thus giving a more reliable enhancement effect. One challenge with this approach is the need for this precision, as both nanoantennae and metasurfaces are highly sensitive to deviations in their geometry. This makes fabrication much more challenging than the more common evanescent field WERS, and makes the device much more vulnerable to breakage and wear in the structure. This also presents an issue with biological samples, as the device would have to be cleaned and disinfected after each trial to preserve bio-safety. If the structure cannot withstand the cleaning process, then the chip essentially becomes a single-use product, which makes the device unpractical and gives rise to problems regarding repeatability.

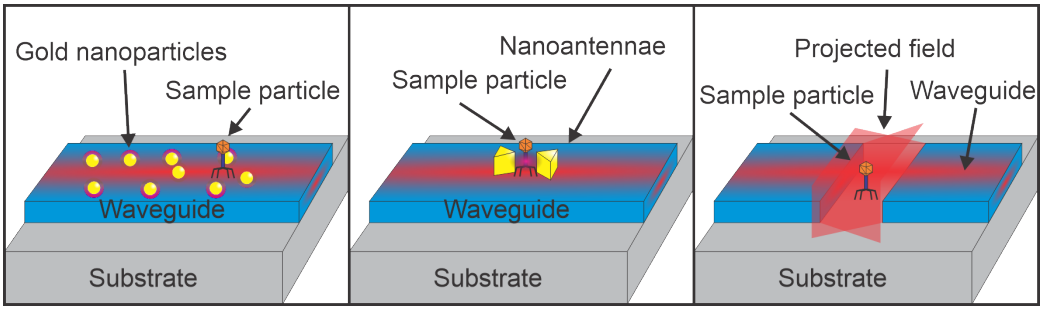


Figure 3.2: Conceptual sketch of three approaches in waveguide enhanced Raman spectroscopy. Left: simple plasmonic enhancement using stochastically deposited gold nanoparticles, creating strong local field enhancements around the beads. Center: plasmonic enhancement using gold nanoantennae, a strong field enhancement is created at the tips of the antennae enabling the generation of strong Raman scattering from the sample particle. Right: projecting the field of the waveguides into a cavity holding the sample, the size of the cavity can be designed to fit one or few particles, enabling trapping and Raman spectroscopy of single particles.

Another approach to WERS revolves around the use of the waveguides more as conduits instead of devices by using the micro/nanometer scale geometry of the waveguides to project the guided field into small volumes[8] as shown in fig. 3.2, exploiting the fact that integrated optics can be used to create microscopic devices with multiple functions. The principal advantage of this setting is the prospect of miniaturization of the desired function, allowing for a device to be created with the ability to measure a specific characteristic with smaller unit cells than what is achievable using a lens-based system[45]. While this setting does not give a direct enhancement similar to the other settings presented here, it does have the advantage of confining the volume to a micrometer size, while the other two methods generally considers a much larger volume. The trade off between enhancement through increased volume and the ability to interrogate small volumes pays off by making this setting capable of performing specific measurements rather than a bulk measurement of the analyte. Thus, if the analyte is heterogenous in nature, both evanescent field and plasmonically enhanced WERS approaches would not be intrinsically able to detect or characterize this heterogenity while the tip projection setting would. This give the tip projection setting access to more information regarding

the sample, and can reveal aspects of the sample that is of high value, especially in the field of biology[29].

3.2 Optical trapping on with waveguide devices

Optical trapping using a lens-based system is a well tested method, and has been in use for decades for analyzing particles down to less than 10 nanometers [51], and can thus be considered a reliable method of micro/nanomanipulation. However, these methods are fundamentally diffraction limited, meaning that even with a short wavelength excitation and a high NA immersion objective, the focused spot is still several hundred nanometers in width. This means that both the confining force and the expected volume of confinement for a particle in the order of 100 nm in size is quite poor, especially for biological nanoparticles with low refractive index. One significant advantage of using a waveguide design for creating this trapping field is that the waveguide can be made to have a relatively high refractive index, thus reducing the diffraction limit compared to free-space or water-based suspension fluids. It has been demonstrated that waveguide-based optical traps can be made to trap particles below 100 nm in size using slot waveguides, enabling trapping of objects as small as DNA strands[77]. Going even further, through the use of photonic crystal waveguides it is even possible to trap objects down to the atomic size[46]. Thus, using waveguide devices it is possible to trap particles much smaller than what is possible using free-space optical systems.

By combining the optical trapping with Raman spectroscopy on a waveguide it becomes possible to create devices that can readily trap, excite, and collect Raman scattering from nanoparticles. A distinct advantage of using a waveguide device in lieu of free-space optics is the microscopic size of the traps. As a waveguide trapping device generally has a trap whose size is in the order of a few microns, it becomes possible to create multiple sites on a single chip. By combining multiple sites with a deliberate design and waveguide circuitry, such as Y-junctions, the device can be made to have multiple traps active at the same time. With an appropriately powerful laser source coupled to the device, it is thus possible to create a Raman-on-chip device that can trap a particle in each of the traps at the same time,

and induce Raman scattering in the process.

3.3 Raman scattering in waveguides

As with conventional WERS, a Raman-on-chip device using single particle optical trapping is also susceptible to the induced Raman scattering in the waveguide itself. The magnitude of the Raman scattering in the waveguide is dependent on the path length of the waveguide circuitry, thus reducing the Raman scattering in the waveguide is crucial for the viability of a Raman-on-chip device. To evaluate the significance of the Raman scattering, a suitable theoretical framework must be introduced.

3.3.1 Theory of Raman scattering in guided modes

The theory explored earlier in chapter 2.2.2 illustrates the Raman scattering of single molecules, free from each other. For the Raman scattering in waveguides, this theory must be carried over into solid materials. In addition to the waveguide being a solid material, be it amorphous or crystalline, it is also carrying an electric field, therefore the interaction between the field and the material must also be considered. Therefore, the spectrum of the light created by a wave propagating in a waveguide can be divided into three elements: Fluorescence, Raman scattering by the molecules of the material, and distortions caused by fluctuations in the material properties.

The first, and strongest, of the observable effects produced in the waveguide by a propagating wave is fluorescence. Unlike Raman scattering, fluorescence is determined by the electron structure of the atoms in the material, thus the wavelengths of the emissions are fixed while in Raman scattering it is dependent on the excitation wavelength. The wavelengths that can excite fluorescence are also fixed, and are generally in the higher energy end of the visual spectrum, making it possible to avoid this effect by using lower energy excitation. Therefore, by using a laser source in the lower energy range of the visual spectrum, such as wavelengths longer than 600 nm, fluorescence can often be avoided.

In contrast to fluorescence, the emission energies of Raman scattering are not fixed, but are instead dependent on the excitation energy. While the magnitude of generated Raman scattering is dependent on the excitation wavelength, becoming stronger with higher photon energies, it is not possible to eliminate it in the same manner as fluorescence. However, by appropriately selecting the material of the waveguide, it is possible to isolate

the strongest Raman features of the waveguide from those in the sample to be analysed. Furthermore, as discussed in chapter 2.2.2, the number of Raman active modes in a material is directly dependent on the chemical complexity of the material, with molecules of N atoms supports $3N - 6$ vibrational modes. In a solid material, there will be additional modes due to the bonds between molecules that makes the material a solid, causing weaker and more diffuse Raman features to appear in the spectrum. Therefore, by selecting a waveguide material with few atoms in its molecules it is possible to make the Raman spectrum of the material have few, sparse features.

Using a chemically simple material to form the waveguide makes it possible to make the Raman scattering in the waveguide have sparse features, which through proper selection of the material can be made to exist outside the frequency range of interest for the samples Raman spectrum. However, the spectrum in the waveguide will also contain an underlying profile caused by variations in the material parameters resulting in spectral noise. In terms of optical waves, the most significant material property that affects the waves is the refractive index, both the real and imaginary. Thus, perturbations of the material, primarily through temperature fluctuations, will perturb optical waves propagating in the material through variations in the refractive index. The temperature fluctuations in the waveguide can be considered as weak distortions caused by spontaneous heat fluxes. By considering these heat fluxes, which can be considered to be stochastic, we form the basis for the stochastic noise in the propagating waves.

The Raman scattering in guided modes has been explored in several works [7, 22, 21], but has largely been limited to the lower frequency ranges due considering thermal diffusion as the principal mechanism of heat flux affecting the temperature field. A newer work[42] has instead considered the stochastic fluctuations in the high-frequency regime, where the spontaneous heat fluctuations in the material have a lifetime much shorter that the propagation time given by diffusion. In this context, the diffusion rate of the temperature fluctuation can be considered to be negligible, thus making the perturbations of the thermal noise entirely dependent on the spontaneous heat fluctuations. This allows the spectral noise induced by the thermal field to be modelled[42] as:

$$I(\Omega)/I(0) = \left(4\pi^2 \frac{L}{{\lambda_0}^2}\right)\left(\left\langle \delta n^2 \right\rangle \frac{\ell^3}{\ell^2 + 2W^2}\gamma\right) e^{-\gamma|\Omega|}, \tag{3.1}$$

where λ_0 denotes the exciting wavelength, L denotes the length of the guided mode, W denotes the mean width of the mode, $\langle \delta n^2 \rangle$ denotes the expected variance of the refractive index changes due to the thermal fluctuations, ℓ denotes the spatial correlation length of the thermal field, and γ denotes the temporal correlation time of the thermal field. The spatial correlation factor $\ell = \sqrt{\tau\kappa/\rho C_V}$ can be considered an extended material property, in the sense that it is largely defined by the thermal conductivity κ, density ρ, and heat capacity C_v of the material. The factor τ, is the relaxation time of the heat fluctuations, which depends on the motion of the elementary charges in the material and is related to the overall internal energy in the material. The temporal correlation time γ must then be non-zero and dependent on the spatial correlation length ℓ to account for the assumed zero rate of diffusion in the high-frequency regime. To link the two, the velocity of the heat fluctuations must be included, but as the velocity is itself a stochastic variable, it is modelled by a Gaussian distribution function with a determined standard deviation σ_v. The temporal correlation time can then be defined by the length of the spatial correlation and the width of the mode as: $\gamma = \sqrt{(\ell^2 + 2W^2)}/\sqrt{2}\sigma_v$.
From eq. 3.1, we can see that a material whose spatial correlation length ℓ is as short as possible and whose temporal correlation time γ is as long as possible is most beneficial, as this would reduce the overall noise produced in the waveguide. From the definition of ℓ we see its square root dependence on the thermal conductivity κ and the inverse square root dependence on the density ρ and the heat capacity C_v, meaning that a material with low thermal conductivity, high density and high heat capacity is poised to have a short correlation length ℓ. Given that the correlation time γ is positively dependent on both the correlation length ℓ and the mode width W, a material with a short correlation length should have a large mode to produce a long correlation time γ. Under the constraint of using a single-mode waveguide, this implies that the best waveguide material is one with low refractive index contrast.

3.4 UV-written silica waveguides

3.4.1 Motivation

As described in chapter 3.3, the potential of a waveguide based Raman device is strongly tied to the induced Raman scattering in the device itself. Thus, the investigation into the viability of such a device must begin with an investigation into the Raman scattering induced in the waveguide by a propagating wave. As illustrated in chapter 3.3.3, the principal contributing factor to the intensity of the Raman scattering, and the features in the scattered spectrum, is the material of the waveguide, and thus the selection of this material is of principal importance.
To meet the outlined goal of having a low spatial correlation length ℓ and high temporal correlation time γ in the model described in eq. 3.1, we investigate a candidate material previously used in Raman spectroscopy to consider it as the platform for a single particle Raman-on-chip device. It is known that Raman spectroscopy using SiO_2 optical fibers is viable, both for bulk measurements in-vivo[14] and for single microparticles[18, 17], thus we investigate if a waveguide built from the same material can achieve similar performance. Given the fact that SiO_2 has a thermal conductivity more than 13 times lower[40] and a thermal capacity more than 4 times higher[32] than Si_3N_4 despite having almost the same density[55], we expect the correlation length ℓ to be significantly shorter for SiO_2 than for Si_3N_4. Also, considering that SiO_2 has a low refractive index of 1.44[43] relative to 2.0 [6] for Si_3N_4, it is possible to create much larger single-mode waveguides, thus increasing the temporal correlation time γ. The resulting spectrum generated in a waveguide of SiO_2 should have a lower magnitude than that of Si_3N_4 according to eq. 3.1.
Based on this, we have chosen to investigate the use of SiO_2-based waveguides fabricated with low refractive index steps and large modes through the use of UV-writing, as shown in fig.3.3. These waveguides are fabricated by Dr. James Gates and his group at the University of Southampton and they are designed to emulate the properties of optical fibers, but on a chip. The waveguides are fabricated by first growing a layer of thermal oxide onto a silicon wafer, which forms the bottom cladding of the waveguide. A layer of silica is deposited onto the thermal oxide by means of flame hydrolysis deposition. This layer is doped with germanium and either boron or, less commonly, phosphorous, and will serve as the core of the waveg-

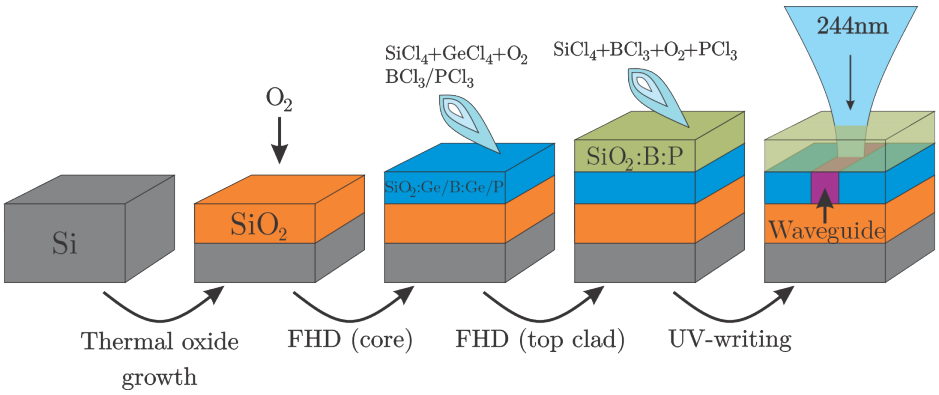


Figure 3.3: Sketch of the fabrication procedure for UV-written silica waveguides. The process is shown in four fabrication steps where the layers are first created onto a silicon substrate before a 244nm UV-laser is used to write the waveguide into the core layer.

uide. Lastly, a layer of silica doped with boron and phosphorous is deposited on top of the core layer, serving as the top cladding of the waveguide. The dopant concentration in the top layer is selected such that the top layer and the core have the same refractive index, which regularizes the vertical and horizontal mode field profiles. The waveguide itself is created in the last step, where a UV-laser is used to induce a refractive index change in the core, which occurs due to the photorefractive properties of the germanium and boron dopants. The result is an approximately rectangular waveguide core, surrounded on all sides by silica with approximately the same refractive index.

Due to choice of doping levels and the weakness of the photorefractive effect, the refractive index contrast between the written waveguide and the surrounding cladding is less than 1.7%[44, 26], which is similar to that of an optical fiber. This gives the waveguide a large single mode waveguide with a very weak evanescent field. The embedded nature of the waveguide makes the evanescent field unavailabe to any particles on the surface of the chip. The high transparency of silica in combination with the low refractive index contrast has the benefit of low propagation losses well below 1dB/cm[44].

The combination of low in-coupling and propagation losses as well as the known low background of silica allows for high power to be delivered along

the waveguide and for the Raman background to be low, making the UV-written silica waveguides promising for single particle Raman spectroscopy. To determine the significance of the background and to fine-tune the design of the waveguide, the precise background must be measured and its absolute intensity evaluated such that the UV-written silica waveguides can be compared to other platforms. This is the topic of the next section and Paper I. As the evanescent field is inaccessible to particles on the surface of the chip, structures must be etched into the waveguide to allow sample particles access to the field. This is the topic of chapter 3.3.5.

3.4.2 Experimental design and findings

Measuring the Raman spectrum of a given material, or in this case a waveguide, in arbitrary units is rather trivial. All that is needed is a pure enough laser, a spectrometer that is sensitive to the relevant wavelength range, and some coupling optics to connect them to the sample. However, to determine if silica is a better material than others, we must make the resulting spectra comparable. To achieve this, the intensity cannot be expressed in arbitrary units, but must be expressed in absolute units. To do this, the acquisition must be calibrated to express the power of the spectrum relative to the input power, which requires a more complicated measurement scheme.
In the work presented in Paper I, this is achieved by the setup illustrated in fig. 3.4 by using a second laser source with a wavelength in the passband of the longpass filter, which allows for the spectrometer to be calibrated to absolute units. The power of the second laser source, which can be measured using a powermeter at key points in the system (P1-3 in fig.3.4), also after the longpass filter (LP in fig.3.4). The acquired spectra can be calibrated to absolute units compatible with the expression in eq. 3.1. The calibration method and the resulting spectra are detailed in Paper I.
To summarize the results from Paper I, the UV-written waveguides were found to have a Raman spectrum with no pronounced features, populated purely by broad, weak peaks. The power level of the Raman spectrum generated by the UV-written SiO_2 was found to have a maximum of -107 dB at 425 cm^{-1} when excited by a 785 nm pump source and a maximum of -106 dB at 436 cm^{-1} when excited by a 660 nm pump source. By comparison, the power level of the Raman spectrum generated by a Si_3N_4 exhib-

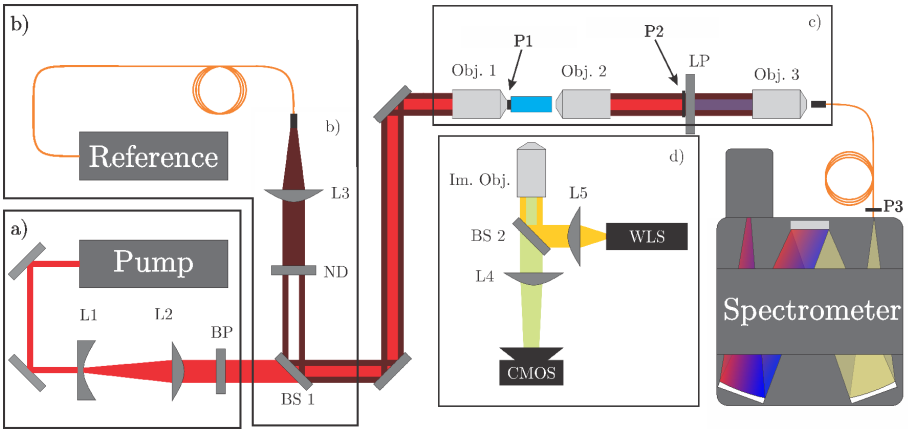


Figure 3.4: Sketch of the setup used to measure the Raman spectrum in absolute units. The Raman spectrum of the waveguide is collected using an in-line configuration where the outcoupled light is passed through a long-pass filter and coupled to a spectrometer. A second laser source, acting as a power reference, is coupled into a common path with the main source and measured both by power meter and by spectrometer to provide the calibration.

ited a sharp peak centered around 2323 cm^{-1} with a maximum power of -99 dB and an overall maximum power of -93 dB at 188 cm^{-1}. The results reveals that the UV-written SiO_2 waveguides produce Raman scattering with 8 dB lower power level overall and 15 dB lower power level in the biological fingerprint region (800-1700 cm^{-1}) compared to Si_3N_4, demonstrating that the UV-SiO_2 are well suited for a Raman-on-chip device. Comparison with a commercial SiO_2 fiber reveals that the UV-written waveguides have a power level 10 dB higher than the fiber, demonstrating that the waveguides are still not as suitable as fibers.

3.5 Status of on-chip trapping and Raman spectroscopy

3.5.1 Motivation

To trap particles, the waveguide chips are designed such that two counter-propagating beams are guided by waveguides leading to a trapping site, as shown in fig. 3.5. The trapping site is formed by a trench intersecting the waveguide where the two beams interfere, the standing waves produced by the interference create one or more potential wells where a particle is trapped. An orthogonal aperture above the trench collects the Raman scat-

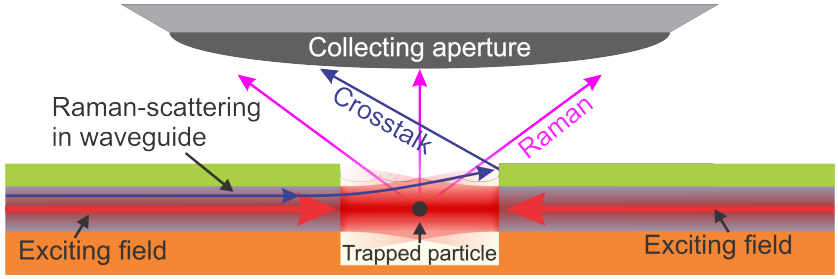


Figure 3.5: Design of a simple waveguide trap. The two counter-propagating modes are projected into a narrow open space intersecting the waveguides, where the particle is suspended in a solution. The standing waves of the counter-propagating beams produce the gradient for trapping while exciting Raman scattering from the particle. The generated Raman scattering is collected using an orthogonal collection aperture.

tering from the trapped particle. The orthogonal configuration reduces the amount of background from the waveguide collected by the aperture. However, some of the Raman scattering of the waveguide will inevitably couple into the aperture through crosstalk between it and the waveguide, as illustrated in fig. 3.5. Thus, given the Raman-background of the waveguides and the Raman spectrum of a PS-bead reported in Paper 1, we can infer the signal given a set crosstalk, as shown in fig. 3.6. In the case of -10 dB crosstalk between the waveguide and the collecting aperture, the signal a 500 nm particle is barely detectable, with only the ring breathing mode at 1001 cm^{-1} being slightly visible. With a crosstalk of -20 dB, the features of the polystyrene particle begin to manifest more clearly, with

both the breathing mode at 1001 cm^{-1} and the stretching modes at 1527 and 1604 cm^{-1} of the rings visible. With a crosstalk as low as -30 dB, the full spectrum of the polystyrene becomes apparent, demonstrating that a crosstalk between -20 and -30 dB produces spectra with a usable signal-to-noise ratio. The geometry of the trapping sites, and the overall design of

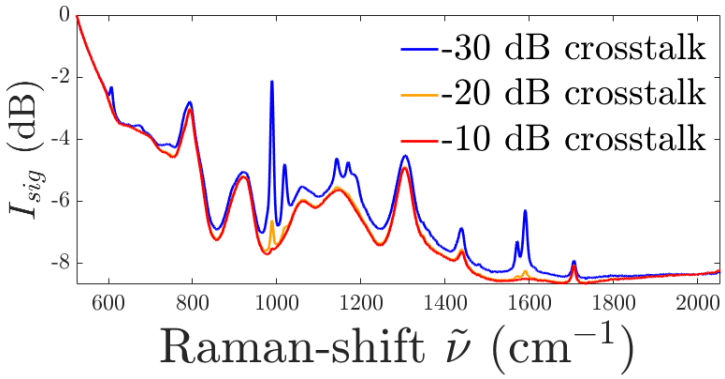


Figure 3.6: Inferred signal of a 500 nm diameter particle trapped in a SiO_2 waveguide trap. It is shown that a crosstalk of -30 dB between the waveguide and the aperture makes the signal of the particle clearly distinguishable from the background. The signal is still visible with a crosstalk of -20 dB, but is significantly obscured by the background. With a crosstalk of -10 dB, the signal of the particle is barely detectable and difficult to distinguish from the background without signal processing.

the waveguides as well as the measurement system, must therefore be designed such that the crosstalk between the waveguides and the collection system be kept as low as possible. In addition, the field projected by the waveguides must also be sufficiently powerful and free of distortions such that particles can be trapped by it.
While single trapping sites have been shown to be viable[8], the true purpose of creating a trapping chip is to create multiple traps on the same chip, thereby enabling the trapping of multiple particles as shown in fig. 3.7. Using the single trap design and a chip with multiple waveguides, a device with multiple trapping sites can be fabricated. As with the single-trap design, the multi-trap design uses a single orthogonal aperture to collect the Raman scattering from the particles. To differentiate this method from bulk methods, such as conventional WERS, the collection system is

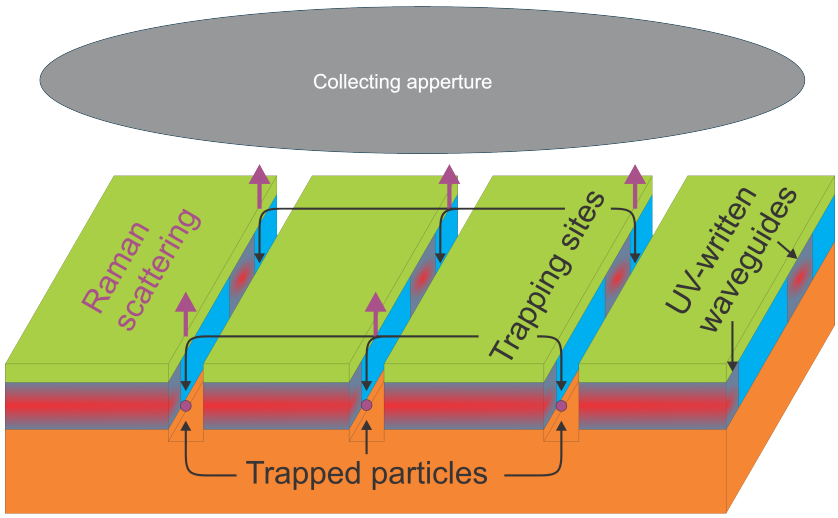


Figure 3.7: Sketch of waveguide chip with multiple trap sites. The chip design replicates the trap sites shown in fig. 3.5, expanding the single trap design to a trapping array. Through low losses, multiple trap sites can be chained on a single waveguide such that multiple particles can be trapped and excited by the same beam.

designed such that the Raman scattering from each trapping site is individually collected. This can be achieved by having the collecting aperture image the trapped particles on separate channels of an optical spectrum analyzer, as shown in fig. 3.8.

3.5.2 Prototype 1

The first prototype design uses a traps formed by etching trenches orthogonally into the waveguide chip, as shown in fig. 3.9. The chip was fabricated at the University of Southampton and characterized in the work described in chapter 3.4.2 and in Paper I. Once characterized, the trap designs were etched at the Norwegian University of Science and Technology (NTNU) by Dr. Marek Vlk also working on this project. The design uses the etched walls of the trenches to act as the projecting facets of the waveguides for trapping. The sidewalls are thus required to be especially smooth and steep. The design of the trenches require a width of 1.5 μm and a depth sufficient to penetrate the core layer of the waveguides. The combination

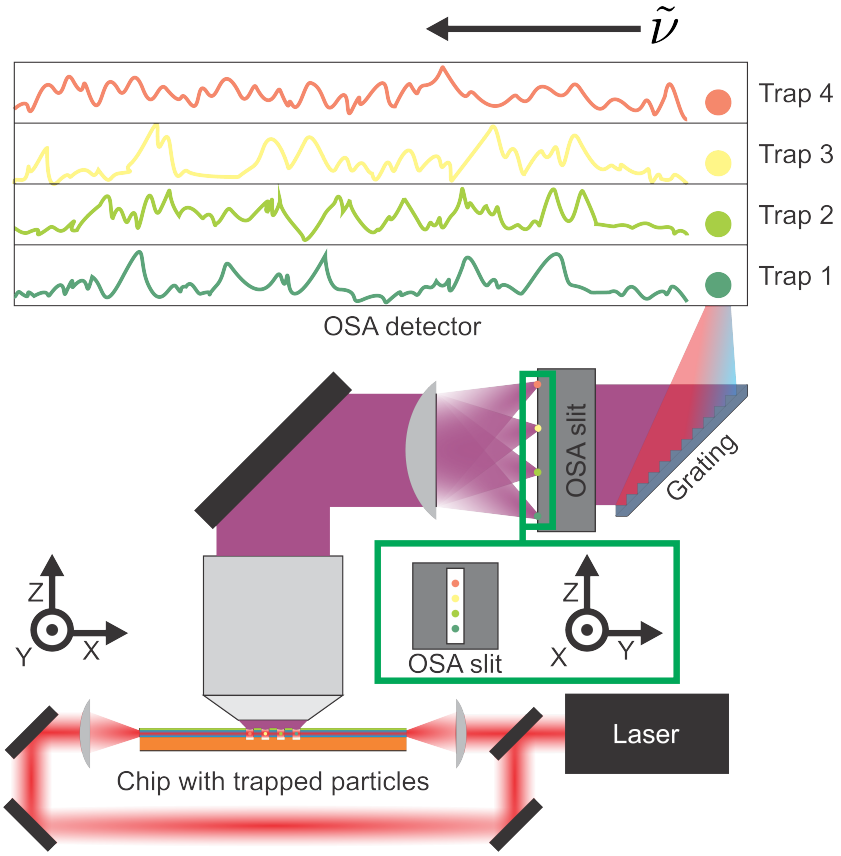


Figure 3.8: Sketch of a setup using the multi-trap waveguide chip to trap and excite multiple particles in parallel. By having the collecting aperture be part of an imaging system that projects an image of the trapped particle on the entrance slit of an OSA. With the individual particles being imaged at separate points along the entrance slit, the light from each of the particles will be projected on separate channels of the OSA, allowing them to be analyzed separately.

of these requirements made it challenging to achieve an etched structure of sufficient quality for the device to function. The intrinsic resistance of SiO_2 to etching agents further complicated the process, making it impossible to use a simple resist mask for etching. The process instead required the use of a deposited chromium mask that itself required a silicon mask, resulting in a process of three deposition steps, followed by lithography of the resist,

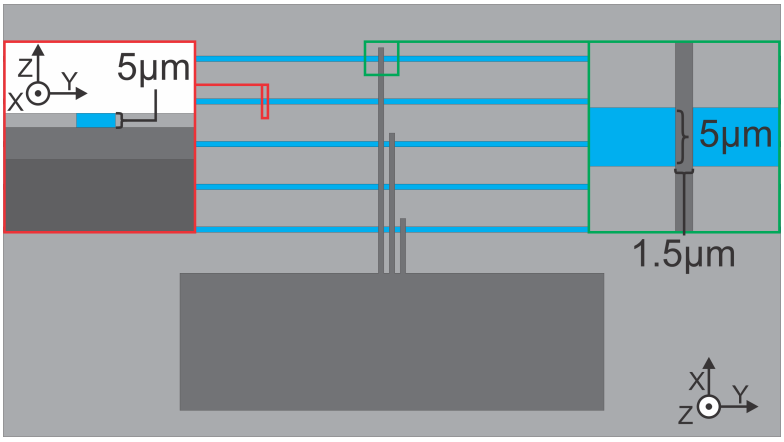
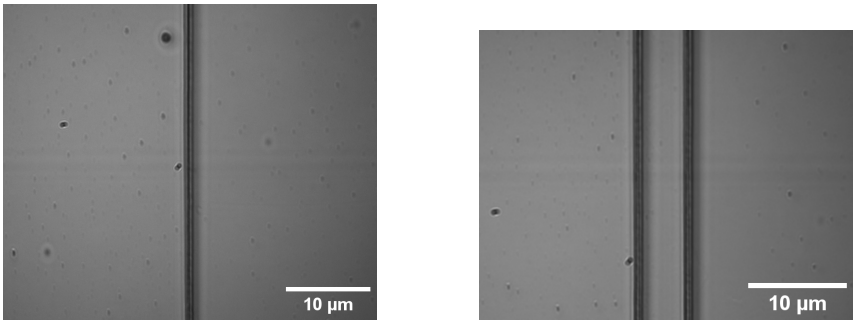


Figure 3.9: Sketch of the trap site design for the first prototype chip. The trenches and the reservoir are shown in dark grey with the waveguides shown in blue. The traps are formed by 1.5 μm wide trenches across the waveguides. The chip is designed such that some waveguides have single trenches, some double trenches and one with triple trenches.

and etching via three stages of reactive ion etching to produce the trenches shown in fig. 3.10 of the prototype chip. Testing of the prototype chip re-



(a) Brightfield image of a waveguide crossing a single trench.

(b) Brightfield image of a waveguide crossing a double trench.

Figure 3.10: Microscope images of the fabricated prototype 1 chip showing the trenches. The waveguides are weakly visible as horizontal lines in the images.

vealed an expected drop in transmission across the trenches ($\approx 5dB$), but otherwise with the same transmission characteristics as previously mea-

sured. During testing, it became clear that several of the waveguides exhibited a variable drop in transmission, despite being etched with the same trench design. Due to the absence of a top cladding on these waveguides, it was initially suspected that this variable change in transmission was due to chipping along the input facets of the waveguide. This was solved by thoroughly polishing the chip. Polishing reduced some of the losses, but the variance was still present and high loss persisted for some of the waveguides on the chip.
Several attempts were made at trapping 1 μm and 0.5 μm polystyrene beads using the prototype chip, but only one attempt yielded results. Further attempts revealed a progressively increasing propagation loss and a degradation in the lateral confinement of the guided mode, eventually rendering the waveguides inoperative. As this effect was also noticed in the un-etched waveguides without a top cladding, but not in the top clad waveguides, it is speculated that the absence of the top cladding may have lead to chemical degradation of the core layer.

3.5.3 Prototype 2

The second prototype design deviates from the trench-design for the trapping sites by using tapers at the end of the waveguide in lieu of the blunt endings of the waveguides, as shown in fig. 3.11. This allows for the medium surrounding the taper (water) to act as the side-cladding in the final micrometers of the waveguide leading up to the trapping site. The local increase in refractive index step allows for the mode to be narrowed down to approximately $1.5\mu m$ by the time it reaches the trap instead of the full $5\mu m$ width of the embedded waveguide. Together with a thinner core layer, this allows for better coupling to a particle smaller than the width of the waveguide. The second prototype chip was manufactured in the same process as the first, but due to a technical issue during fabrication the first chip with this design was damaged.

3.5.4 Prototype 3

The third prototype design expands on the taper design by leaving thin sections of the waveguide in the gap between them, these will hence be referred to as wires. By using the suspension of the particles as the side

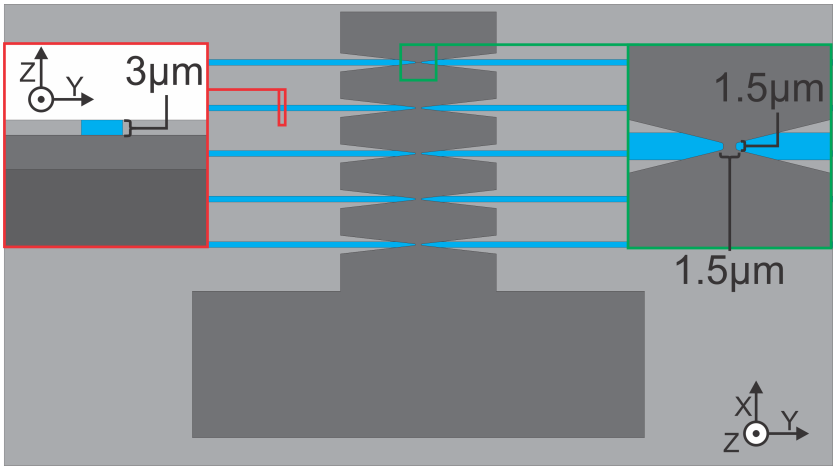
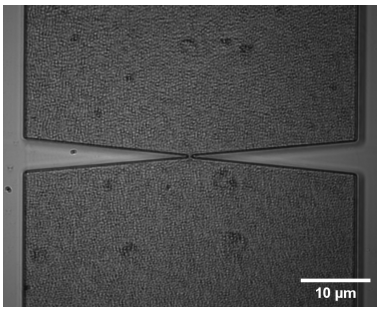
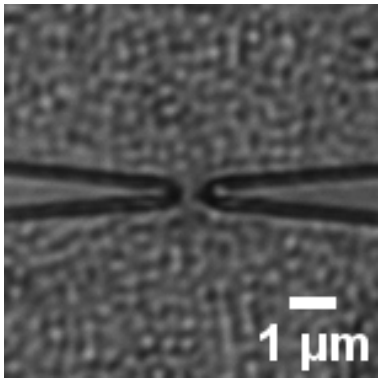


Figure 3.11: Sketch of the trap design for the second prototype chip.



(a) Brightfield image of a trapping site in the second prototype chip.



(b) Brightfield image of a trapping site near the tip of the tapers in the second prototype chip.

Figure 3.12: Microscope images of the fabricated prototype 2 taper design.

cladding of the waveguides, giving a higher refractive index step, the lateral mode size in the wires can be reduces, thus projecting a denser field onto the particles in the trap. These wires will then be chained together, separated by gaps that form the trap site, as shown in fig. 3.13. Due to the small size of the wire structures, fabrication is challenging and the dimensions of the resulting wires are thus not as designed. Furthermore, the same problem as in prototype 2, being the misalignment of the tapers, still

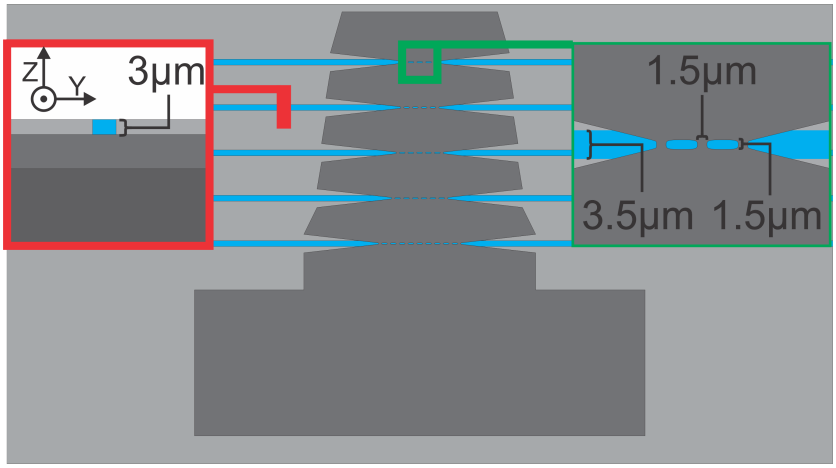
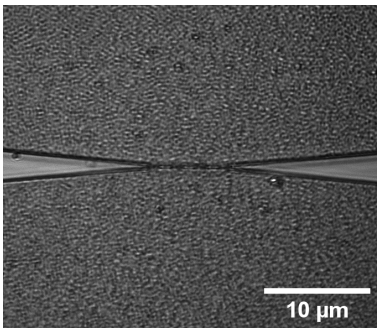
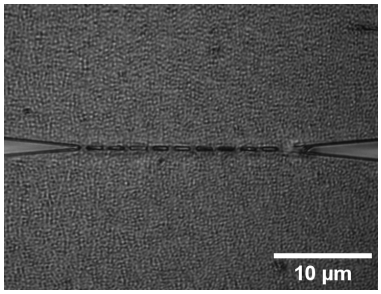


Figure 3.13: Sketch of the trap site design for the third prototype chip.

presents a significant challenge. Due to the poor visibility of the waveguides, the tapers were misaligned, rendering the first prototype 3 unviable. As the equipment required to fabricate more became unavailable shortly after this, no further attempts could be made.



(a) Brightfield image of a trapping site in the third prototype chip with three wire elements.



(b) Brightfield image of a trapping site in the third prototype chip with nine wire elements.

Figure 3.14: Microscope images of the fabricated prototype 3 wire design. The chip was fabricated with three and nine wire segments on different waveguides to evaluate the losses of multiple wires in series.

Chapter 4

Convolutional neural network for quality control of silicon ingots

4.1 Introduction

The challenges presented by the background in waveguide-based devices can be significant, as illustrated in chapter 3, but there are ways to mitigate its effects on the signal. This can be achieved by altering the design of the waveguide, both in terms of geometry, in the choice of material, and in the fabrication process. However, there are limits to what can be reliably achieved in a physical device, thus the background will always be present and is likely to pose a challenge regardless of the design. Another area where this can be mitigated is in the post-processing of the data, relying on powerful signal processing techniques to mitigate the presence of the noise in the signal. In simple terms, the acquired signal can be considered a superimposed signal onto a background:

$$I_{sig} = I_{background} + I_{sample},$$

where the background can ideally be removed by simple subtraction of the background. The background can often be approximated using previous measurements of the material, as shown in chapter 2, but the precise perturbations in the signal and variations in its intensity cannot be so easily measured. A subtraction using a background such as this will leave distortions in the signal and given the magnitude of the background relative to

the expected sample signal, these distortions are likely to be large in comparison to the sample signal.
We must therefore move to more advanced, adaptive filtering methods to remove the background without introducing such distortions. In the current state of the art, the epitome of such methods is machine learning, which can be designed to adapt to a very wide range of patterns and is capable of discerning signals from highly mixed data. Because such algorithms are highly capable at separating out data in this manner, they have significant potential for their use in separating the Raman spectra of a nanoparticle from the Raman background of the waveguide.
In order to implement a machine learning algorithm for Raman spectroscopy, a suitable algorithm must be developed. To achieve this, the attention of the project was temporarily shifted to another project: infrared tomography, which was an extension on the topic of my master-thesis[34]. The goal of this is to use the tomography setting as a testing ground for developing a machine learning algorithm that can later be transferred over for use in Raman spectroscopy.
The tomography project exploits the infrared transparency of silicon to observe the internal structure of a silicon ingot using laser scanning tomography. The goal of the project was to see distortions in the crystalline structure of the ingot using the transmission scan data. Because the intended use of such silicon ingots is in high-quality photovoltaic cells, there are strict requirements to the crystal quality, notably that it must be monocrystalline. In the case of significant distortions in the crystalline structure during production, the monocrystalline rapidly breaks down into multicrystalline structure. Because of the overlapping crystal structures in multicrystalline material, the electron mobility, and thus the overall electrical properties of the material becomes inhomogeneous. The result is that the efficiency, and predictability of the local electrical properties of the cell. Thus, the cell loses its practical and commercial value.
The distortions are challenging to detect during production and are therefore often not noticed until the product approaches the wafer stage near the end of the production line. The result of this is that faulty material passes through a significant length of the process before detection and thus costs time, resources and effort while not returning a usable product. The aim of the master-project was to detect defects in the crystalline structure during production[34, 35], while the aim for this chapter, and Paper II, is to detect defects in the silicon boules post-production. This chapter de-

scribes the production process for monocrystalline silicon boules, and the tomographic scanning method. The topic of tomographic scanning and silicon production is quite far from the topic of the thesis, but the methods developed for this topic, as presented in Paper II, contribute significantly to the main topic of this thesis.

4.2 Czochralski process

To produce monocrystalline ingots from materials like silicon, special fabrication processes are required and the most widely used process for this is the Czochralski process. This process is focused on using a template crystal, referred to as a seed, that the ingot is grown onto such that the ingot becomes an extension of the seed. This method allows for the growth of monocrystalline ingots up to several hundred kilograms each, while maintaining the original high-quality crystal structure of the seed throughout the material.
The core of the growth process is that a small rod of crystalline silicon, referred to as the seed, is used as a template for the silicon boule. This is achieved by keeping the seed crystal below the melting temperature of the material, which for silicon is 1 687 K, and introducing it to a body of molten silicon, which is kept just above the melting temperature. The boule is then produced by allowing the melt to freeze onto the seed crystal, thus using the structure of the seed to provide the nucleation points for the silicon as it freezes onto the seed, resulting in the newly frozen material assuming the same crystalline structure. As the melt begins freezing onto the seed, the seed is slowly retracted from the melt, as shown in fig. 4.1, such that the newly solidified melt takes its place as a nucleation source. By controlling the rotation of the seed, and its rate of retraction, relative to the temperature gradient at the contact surface, the crystal can be grown to produce a cylindrical ingot of desired diameter. This process is then continued until the melt is nearing depletion, when the growth process undergoes a controlled termination by gradually reducing the diameter of the interface to produce a conical section at the end of the ingot, often referred to as a "tail".
While this process is well tested and is generally quite reliable, it is also sensitive to a wide range of conditions during production, from contamination to temperature anomalies or vibrations. Because the potential of the

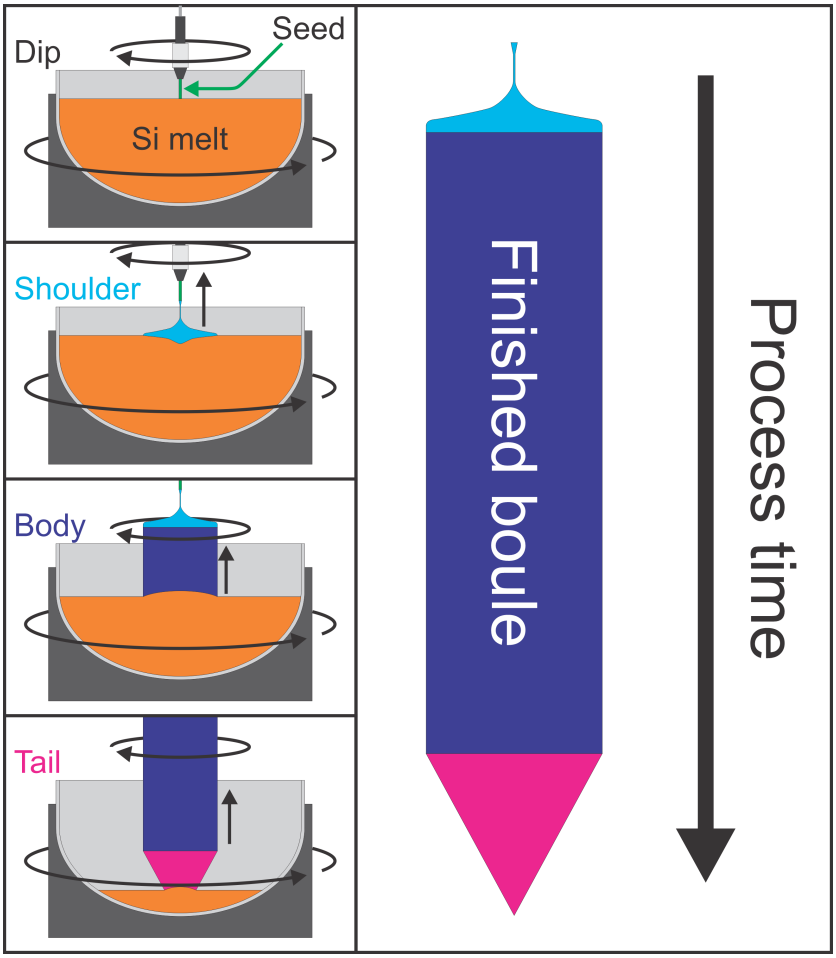


Figure 4.1: Sketch of the production a monocrystalline silicon boule through the Czochralski process. The boule is produced in four stages, starting with the dipping of the seed into the melt. The crystal is then widened during the shoulder section by reducing the pull rate while controlling the rotation of the crystal. Once the desired width has been achieved, a cylindrical section called the body is pulled. The body section contains the usable material of the boule and comprises the majority of the mass of the boule. When the melt is nearing depletion, the diameter of the crystal melt interface is reduced to produce a conical section, called the tail, which allows for controlled termination of the growth process.

nucleation into the crystal structure is low in comparison to nucleation to any point, small disruptions can cause crystal structures other than those of the seed to begin growing in the ingot during the process. Once this process begins, it cannot be reversed and the multicrystalline structure will begin to spread laterally until the entire cross section of the ingot is multicrystalline, as shown in fig. 4.2. Therefore it is important to determine the location where the multicrystalline structure begins to grow, such that the section of the boule which is multicrystalline can be removed with as little loss of monocrystalline material as possible.

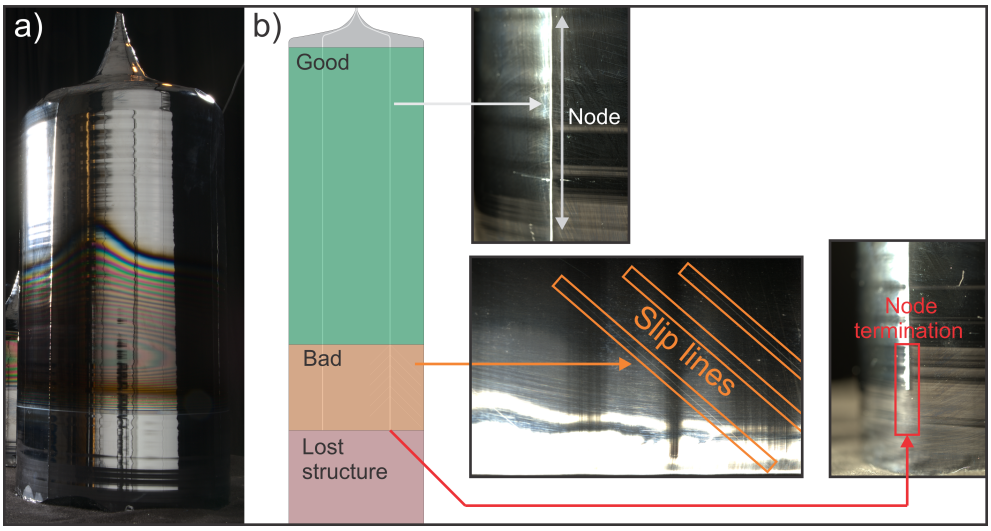


Figure 4.2: Illustration of the segments of a silicon boule with structure loss and the symptoms of structure loss. a) Photograph of a 36kg prematurely terminated silicon boule used as a sample in Paper II. b) Sketch of the sections in a boule with structure loss, illustrating the formation of slip lines at the onset of internal structure loss and eventually the disappearance of the nodes at the onset of full structure loss.

The detection of the interface between the mono- and multicrystalline material is a significant challenge in quality assurance of such ingots, largely because its occurrence produces subtle signs. The clearest of these signs is the disappearance of the "nodes" of the ingot, which are four equally spaced ridges following along the length of the ingot as a result of the cubic nature of silicon crystals. Thus, once the bulk material stops being one large crystalline structure, the nodes will also stop appearing. Using

the nodes as an indicator of structure loss is a reliable and widely used method, but it also yields limited information. As the multicrystalline growth begins near the center of the cross section before spreading outwards in a conical pattern, the onset of multicrystalline growth always precedes the disappearance of the nodes and other surface phenomena. The length between the onset of surface phenomena and true loss of structure is often estimated in quality assurance, but the accuracy of this estimate is relatively low.

4.3 Infrared transmission of silicon

The purpose of this work is to investigate a possible tool for measuring properties of the material below the surface of the ingot. The chosen tool for this is laser scanning of the ingot.
An initial work[34] explored infrared laser scanning as a method of measuring the deflection of the crystal-melt interface during production of monocrystalline silicon. Because the magnitude of the deflection of the interface is indicative of the temperature surrounding the crystal during growth, it is also tied to the induced thermal stress, which is a significant factor in the generation of crystal defects. Thus, by monitoring the deflection of the interface, it becomes possible to monitor the state of the process and estimate the likelihood that defects will be generated in the crystal.
The transparency of silicon in the infrared spectrum permits the transmission of an infrared laser beam through the material, which can then be exploited to probe the internal volume of the silicon boules. In the work presented in this chapter, and in Paper II, the concept of laser scanning is expanded from monitoring the interface deflection to monitoring the crystalline structure through the transmission of the infrared beam. A mid-infrared laser source is used to generate the scanning beam which is transmitted through the ingot and detected on the opposite side. By introducing rotation of the crystal and translation of the laser and detector along the axis of rotation, a full tomographic scan of the boule can be made.
Because silicon is a semiconductor material, there exists a region of "forbidden" electron energy states between the valence band and the conduction band. This means that for an electron to transition between bands, the electron must be provided with sufficient energy to cross this gap. If provided with a lower energy than this, the electron will be motivated to

go into the band gap, and since there are no actual states that correspond to the energies in the band gap, this transition cannot occur and thus the electron cannot accept the energy. When this energy is provided by a photon, the result is that a threshold energy, and thereby wavelength, exists where no band-to-band transitions can be induced by the photon. In indirect band gap materials, which also includes silicon, the lowest energy transition between the valence and conduction bands also involves a momentum change provided by a phonon, requiring a joint event between a photon and a phonon for a band-to-band transition to occur, which reduces the rate of such occurrences when the photon energy is below the required energy for a photon only transition. This results in three main spectral ranges for the absorption in silicon, as shown in fig. 4.3:

- A_d: Where the photon energy is above the threshold for a photon only transition, occurring in Silicon for a photon energy of 3.2 eV or higher[74].
- A_i: Where the photon energy is above the band gap energy but requires a phonon assist to make the electron transition, occurring in Silicon for photon energies between 1.11 eV and 3.2 eV[74].
- B: Where the photon energy is lower than the required energy for an electron band transition, occurring in Silicon for photon energies below 1.12 eV at room temperature.

Thus, for wavelengths longer than 1.11 μm, the electrons are not provided enough energy to transition between bands, thus the electrons cannot absorb the photons and the absorption of the photons is low. The absorption spectrum of Silicon measured by Schinke *et al.*[64] shown in fig. 4.3, demonstrating how the absorption is strongest above the direct band gap energy (A_d), weaker between the indirect and direct band gap energies (A_i), and drops dramatically below the intrinsic band gap energy (B). It has been shown that the absorption spectrum can be modeled[34] using the

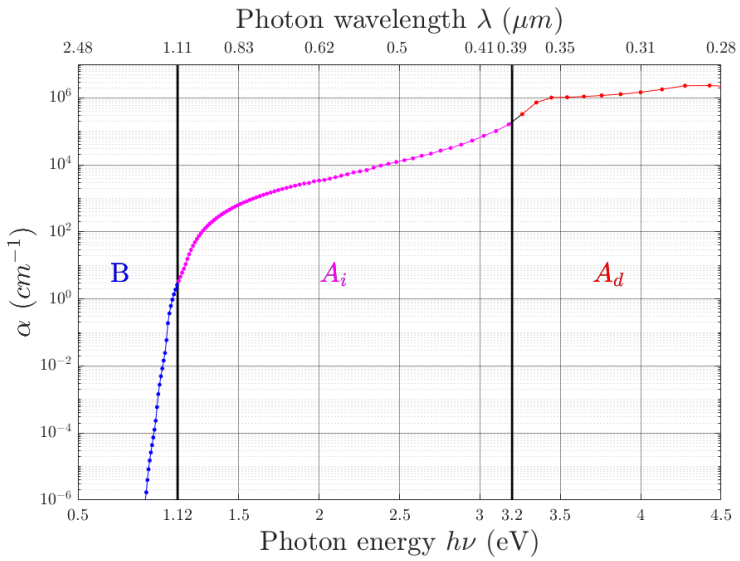


Figure 4.3: Absorption spectrum of silicon[64]. The spectrum is shown divided into the

expression:

$$\alpha_{Ex,B+A_i}(\nu,T) = \chi(h\nu)\left\{\frac{C_1(T)}{E_{g,D}(\mathbf{0},T)^2[h\nu - E_{g,D}(\mathbf{0},T)]^2} + \frac{C_2(T)}{E_{g,D}(\mathbf{p}_0,T)^2[h\nu - E_{g,D}(\mathbf{p}_0,T)]^2} + \frac{C_3(T)}{{E_a}^2(T)[h\nu - E_a(T)]^2}\right\}, \times D\eta(\mathbf{p}_0,T)(h\nu - E_g(T))^2 h\nu + C_{FCA}(\nu,T)\nu^{-2}N \quad (4.1)$$

where $C_n(T)$ denotes the coupling between the photon/phonon event and the transition of the electrons, $E(T)$ denotes the energies of the transitions, $\eta(T)$ denotes the phonon energies, and $C_{FCA}(\nu,T)$ denotes the absorption coefficient for photons by free carriers created by the dopants in the material.

4.4 Experimental design

The experiment is uses a mid-infrared laser source ($\lambda = 1.6\mu m$) mounted on a vertical translation stage to project a beam into the sample ingot.

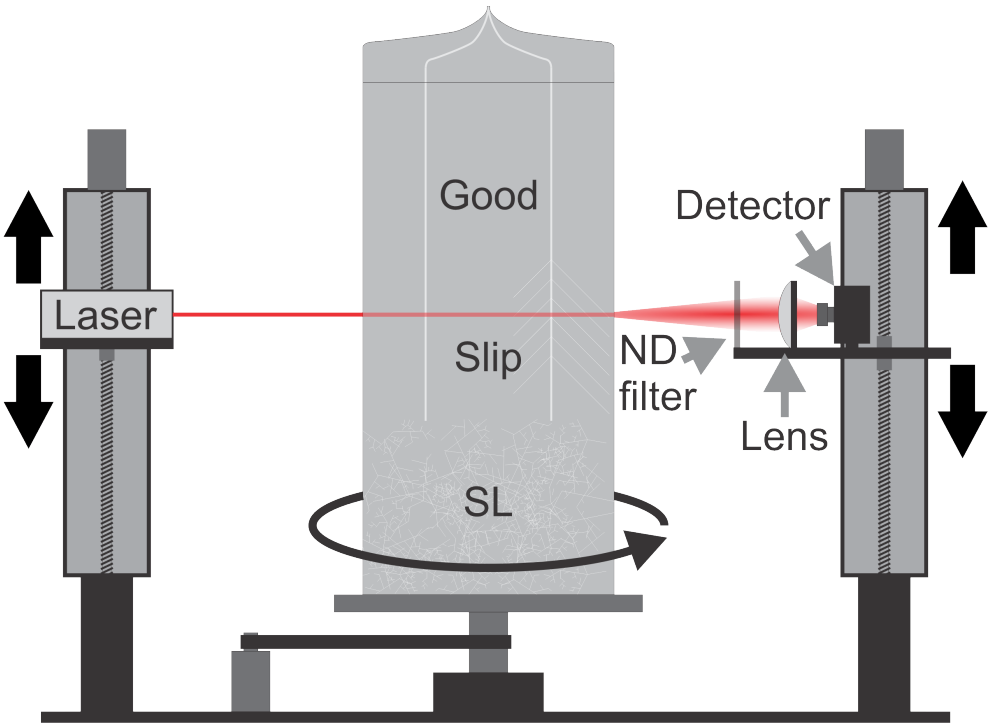


Figure 4.4: Sketch of the setup used in the infrared tomography of silicon boules. The laser and detector assemblies move independently along the vertical axis to maintain alignment with the beam transmitted through the boule. A lens is used to collect thetransmitted beam and a neutral density (ND) filter is used to prevent saturation of the detector.

The transmitted beam emerging from the other side of the ingot is collected by a lens and focused into a detector, also mounted on an identical vertical translation stage. The set-up is shown in fig. 4.4. The sample can be scanned vertically by moving the laser and detector independently, which together with rotation of the sample ingot enables tomographic scanning of the sample. The material of the sample is evaluated by the relative transmission of the beam, which is collected in vertical slices as the sample rotates to reconstruct two dimensional transmission maps of the cross section.
In the case of an ideal bulk material, the transmission map is expected to be homogeneous with the transmission being independent of the angle of the azimuth. However, due to the high refractive index of the silicon (3.44

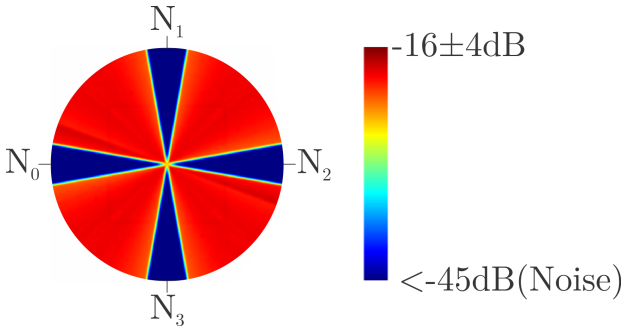


Figure 4.5: Expected transmission profile for a cross section of bulk monocrystalline silicon. The transmission is given by eq. 4.1 in the bulk material but is deflected by the nodes N, resulting in no transmission.

[9]), it is expected that the protrusions of the nodes shown in fig. 4.2 will deflect the incident beam and disperse the output beam, resulting in the transmission map having four equally spaced sections of low/no transmission resulting in a transmission map similar to what is shown in fig. 4.5. To investigate if the measured transmission maps can be used to evaluate the integrity and quality of the crystalline structure, transmission maps of the cross sections through the lengths of the samples are acquired, both in the known multicrystalline and presumed monocrystalline sections of the samples. The transmission maps are filtered with the goal of classifying the structure as intact or not intact, and to quantify the quality of the structure, by using machine learning methods.

4.5 Convolutional Neural Networks

Machine learning has been progressively becoming a more and more powerful alternative to traditional signal processing, and has proven itself efficient and capable at analysing complex patterns in data. The most commonly used machine learning method is the neural network, which uses simple computational units, called neurons. The neurons perform simple multiplication of the inputs with learned weights in order to determine their outputs. These neurons are generally combined in large numbers, with some acting in parallel to produce layers with a "width" given by the number of neurons in parallel, and some in subsequent layers that act in series to give the network "depth". By increasing the width of a network,

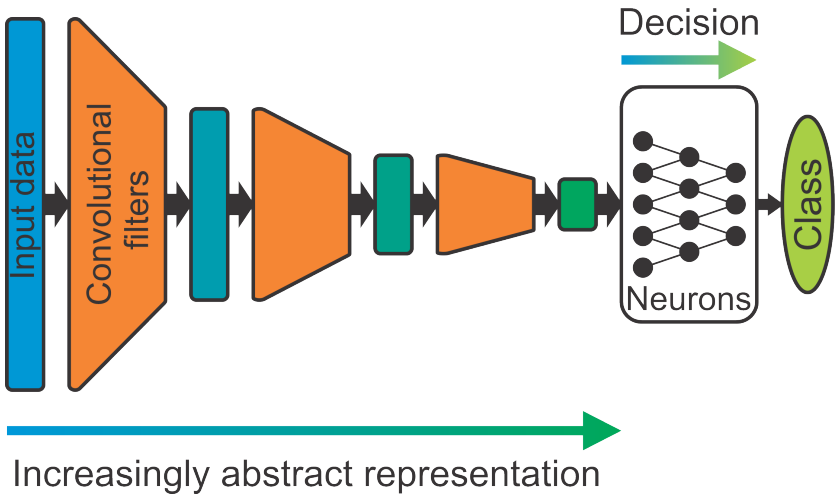


Figure 4.6: Sketch of a general CNN model. The input data is passed through layers of convolutional filters that re-represent the data in increasingly abstract ways. The size of the data is reduced by eliminating the weakest responses after each level of filtering such that the abstract information is concentrated.

it is possible for the network to react to more unique cases of its input, and by increasing the depth of the network, it becomes capable of reacting to more abstract combinations of cases. While such neural networks are can be adapted to a variety of tasks, they are also limited by the way they interact with the data.

One of the principal limitations of conventional neural networks is the required size of the network for a given size. As each neuron must learn a weight for each input data point, processing a large continuous vector, such as a spectrum, or matrix, such as an image, the number of weights quickly becomes large and the model becomes computationally expensive. An alternative to this is to replace the initial neurons with learned filters and convolving the input with these filters as shown in fig. 4.6. By using convolutional filters to process the data before passing it to the fully connected neurons, the model can concentrate the information in the data to make it more compatible with fully connected neural layers, thus reducing the required number of neurons. This also makes the model able to recognize specific patterns in the data and react to them regardless of where in the data these patterns occur. Neural networks that use this, re-

ferred to as Convolutional Neural Networks or CNN's, have been proven superior to conventional neural networks when fed data that is naturally continuous[27], such as those observed in the tomographic scans of the Silicon ingots measured in Paper II.
The variant of CNN used in this work is based on the well known VGG16 architecture[65] but it is modified to accept vector instead of matrix inputs. The chosen architecture consists of a convolutional pre-processing block followed by two parallel neural network heads that produce the outputs of the network. One of the heads is a regression head with linear activation, designed to yield a numeric quality number between zero and one. This should reflect the quality of the structure, with one indicating perfect structure and zero indicating total breakdown to multicrystalline structure. The other neural network head is a classification head, intended to classify the structure as either good, lost structure, or noise only, based on the transmission map. The convolutional layers in the model allows for the transmission profiles to be re-represented such that the features that are indicative of multicrystalline structure are enhanced, as shown in fig. 4.7. The concentrated features are then processed by the classifier and quality regressor heads to evaluate the structure.

4.5.1 Modular architecture and evolution

To make the architecture as flexible as possible, it is designed to be modular, meaning that the architecture is built from a set of arbitrary build parameters, as shown in fig. 4.8a. This allows the architecture to be altered at will, with the structure, such as the number of convolutional layers, size of the filters, number of neurons in each head etc., being decided by a total of 13 hyper-parameters. The optimal selection of these hyper-parameters presents an additional challenge, as the parameter space becomes large and the connection between the hyper-parameters and the performance are poorly known. To meet this challenge, evolution is implemented to converge the hyper-parameters towards optimal values. This treats the hyper-parameters as genes, allowing the networks to be created in groups to form a population in each generation, as shown in fig. 4.8b). As each generation is trained on a common data set, they are also tested on a common set to evaluate their performance. The way their hyper-parameters are passed to the next generation is, much like genes in nature, determined by the performance of the network possessing those parameters. By allowing the best

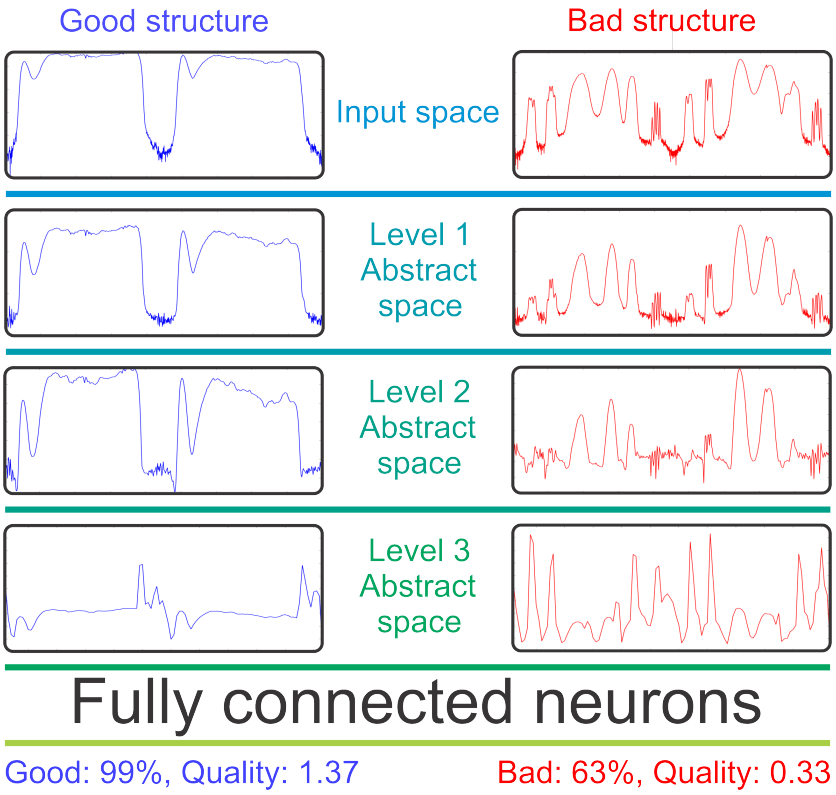


Figure 4.7: Representation of transmission profiles through convolutional layers in the model. The representations enhance the features that differentiate the good and bad structure, passing only the most valuable information to the neurons. This allows the neurons to estimate the likelihood that the structure is either good or bad, and to predict a numerical quality value that reflects how good the observed structure is.

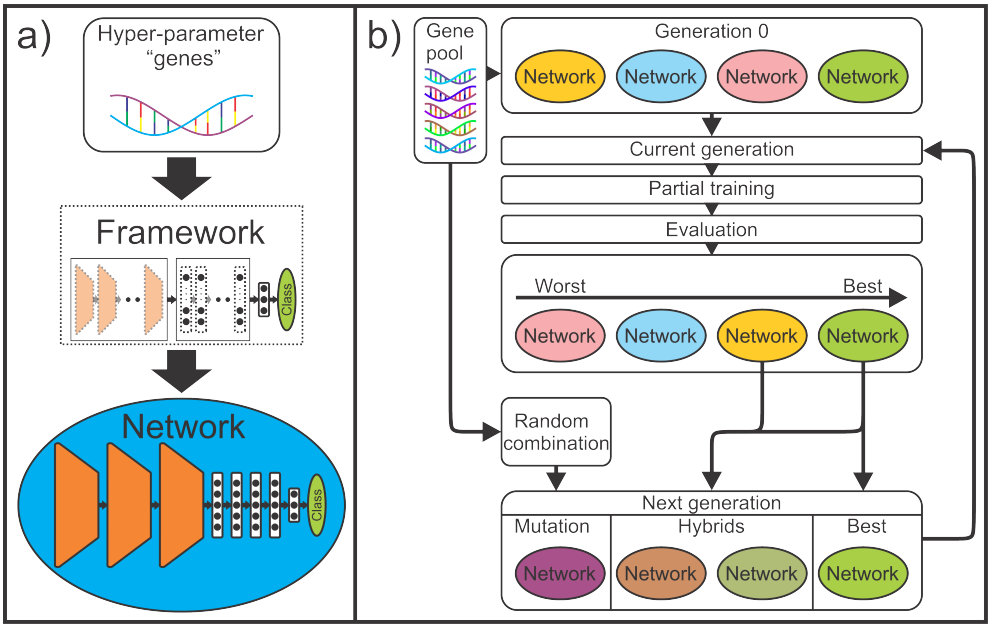


Figure 4.8: Schematic of the CNN and its use of hyper-parameter "genes". a) The model is designed as a framework without a predetermined architecture, instead using the hyper-parameters, or "genes", to construct the network with the architecture determined by the genes. b) The evolution process using a genetic algorithm. The initial generation is created from a gene pool which is then passed to the evolution loop. The generation is trained and evaluated such that the best performing networks are separated from the rest and used as a base for the next generation.

performing network to pass its "genes" to the next generation, the best in each subsequent generation is at least as good as in the previous generation. By allowing the "genes" of the best 60% networks to mix, convergence towards optimal set of hyper parameters is encouraged. And, by making 25% of the next generation from randomly selected genes, "mutations" and thus exploration of the hyper-parameter space are encouraged. The CNN is allowed to evolve over five generations using this data before converging to an optimal architecture with 20 convolutional layers, 5 neuron layers in the classifier, and 6 neuron layers in the quality regressor, giving a network with a total of 40.3 million parameters. The optimal network architecture is then trained on a subset of the data for 40 epochs

before testing in a separate subset. The classification of the cross sections demonstrated a 98.7% accuracy on the test set, showing a good ability to detect the loss of structure. Testing also revealed that the classification results are robust to noise, tolerating a reduction in signal-to-noise ratio from 23dB down to 5dB without a change in predicted class. The quality number given by the regressor conforms to expectations with high contrast and low noise. The regressor is shown to be robust towards noise by tolerating up to 9.9dB of noise with less than a 5% deviation in prediction. Further details and results are given in Paper II.

Chapter 5

Self-supervised processing by Raman autoencoder

5.1 Introduction

While machine learning methods like CNN's are indeed powerful and versatile, they are also fundamentally limited by the information available to them. In order for them to learn, they must be fed a large amount of example data for them learn from. The network attempts to predict a quality of the data, such as its class. The model learns by numerically quantifying how wrong the predictions are, producing what is referred to as the "loss" of the predictions. From the computed loss, the model estimates the gradient map of its parameters from such that following down the slope of the gradient minimizes the loss, and thus leads to the optimal parameters. This is referred to as supervised learning, where the data used for training comes with metadata, or labels, that describes the underlying truth about the aspects of the data we are interested in. As an example: If a CNN is tasked with recognizing a house or a car from a picture, it is trained on a set of thousands of images of houses and cars where, for each, it predicts the likelihood that the image is of a house or a car. This prediction, say 82% chance of a house and 18% chance of a car, is compared with the ground truth, that the image is of a house. If the prediction of "house" had been higher, then the loss for would be lower, and vice-versa. Such supervised learning is therefore only applicable when the features are well known and can be provided along with the data as reliable labels. In

other cases, where this information is unknown or poorly known, supervised learning cannot be applied as it has no basis for determining the loss of a prediction. Raman spectra can be considered an example of this, as conducting supervised learning on Raman spectra would require knowledge of the chemicals of interest in the samples that generated the spectra. To achieve this, a dataset would have to be constructed by measuring a variety of samples containing known concentrations and mixtures of the chemicals of interest. While this can be achieved, training an algorithm to be sufficiently robust for biological applications would require thousands of such measurements, which would take considerable time and effort to prepare, acquire, process, and verify, in order to create the necessary dataset to train a supervised method.
The ideal alternative is an algorithm that can learn from data without labels, instead learning from the data itself. This is widely referred to as unsupervised learning, as the algorithm is not provided any metainformation about the data. There are different ways this can be achieved, depending on the desired function of the algorithm and what types of predictions one wishes to have the algorithm make. One of these methods is referred to as an autoencoder, which consists of two neural networks joined together, as shown in fig. 5.1. The function of an autoencoder is that the first network, referred to as the encoder, accepts some training data which it processes into an output. This output forms an internal data space referred to as the "latent" space. The second network, referred to as the decoder, takes information in the latent space as an input and attempts to reconstruct the original data. Learning is achieved by comparing the reconstructed data from the decoder to the original input to the encoder. The goal is for the encoder to learn to recognize significant features or patterns in the input data such that the information in the input data can be described using only the comparatively small latent space. The latent space can therefore be considered analogous to the labels a CNN would use to train, except that the latent space representations are generated purely by the network. The overarching goal of using an autoencoder on Raman spectra of EV's is therefore to have the algorithm determine and detect important features of the spectra, with the ideal case being that it manages to determine latent representations where each dimension corresponds to a chemical of interest in the EV. This is where the CNN developed for the silicon-project comes into play. As an autoencoder is conceptually two neural networks, such as CNN's, connected by a latent space, a convolutional autoencoder[48] can

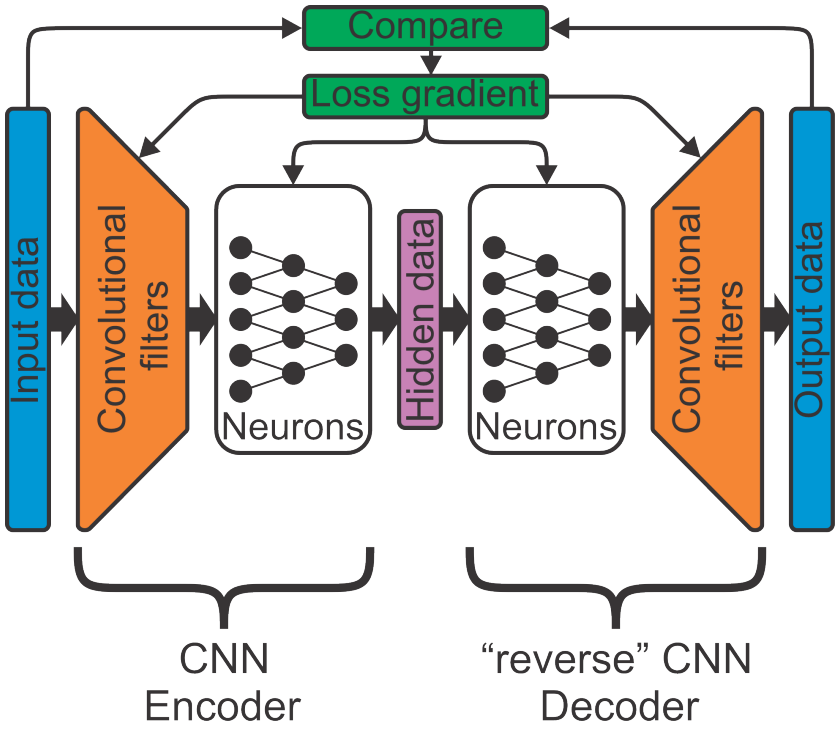


Figure 5.1: Schematic of a generic convolutional autoencoder. The encoder network functions in the same way as a conventional neural network, except instead of predicting classes, it predicts abstract latent information from the input. The integrity of the information is verified by the decoder, which attempts to recreate the input from the latent information. Learning is achieved by comparing the recreated data to the original data, thus enabling self-supervised learning that does not require labels.

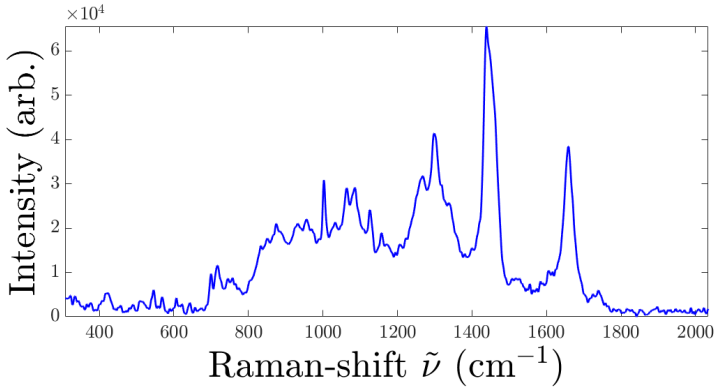


Figure 5.2: Raman spectrum of an EV derived from blood platelet in its natural state. The spectrum can be shown to consist of densely packed Raman modes from approximately 700 cm^{-1} to 1400 cm^{-1}.

be built using the same structure as was developed for the silicon-project in chapter 4. Thus, by relying on the modular nature of the already developed CNN and adding certain features, a powerful autoencoder algorithm can be built to handle unlabeled Raman spectra.
Through measurement systems like confocal Raman microscopes, the spectra of EV's can be determined and from this, information can be extracted. However, because of the complex chemical makeup of EV's, these spectra , as shown in fig. 5.2, are also relatively complex compared to the sparse peaks normally observed in materials consisting of a single molecule. Thus, using only a few select peaks to evaluate the spectra omits the majority of the information in the spectra, giving a poor description of the samples.
To fully exploit the richness of the information in the spectra, and to make the analysis method capable of detecting patterns in the spectra, a specialized autoencoder is developed to analyze the spectra. The goal of this work is to produce an autoencoder that is capable of decomposing the spectra of EV's such that EV's can be differentiated using biochemically significant factors. To motivate the algorithm to learn these factors, we wish to encourage the model to learn to recognize features belonging to the same chemical group, such as proteins, as a single factor. While accomplishing this would require reference information, such as spectra of pure proteins, we can motivate the model towards this goal by modifying its learning outcome. For the learned components to be true representations of the chem-

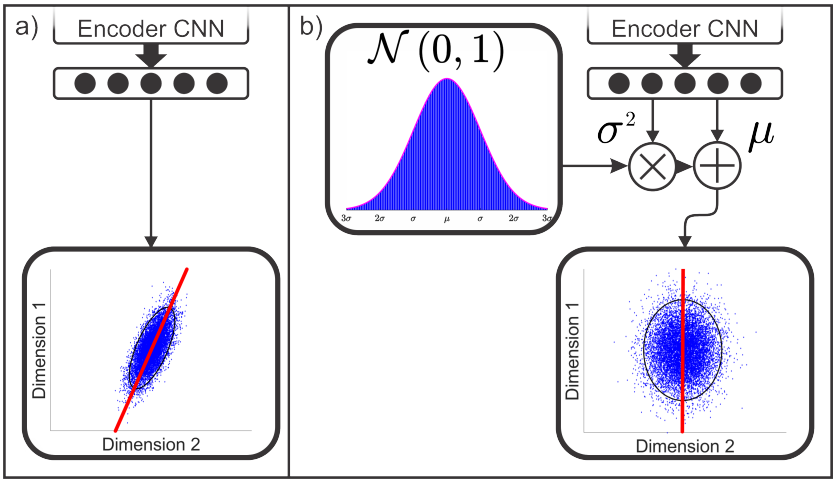


Figure 5.3: Illustration of the effect of using Gaussian re-sampling to generate the latent information in variational autoencoders. By motivating the network to represent the latent information as uncorrelated Gaussian variables, the network is motivated to learn the latent information as linearly independent.

ical elements of the spectrum, they must be independent, while this is difficult to accomplish, it can be approximated. By modeling the latent space representations using uncorrelated Gaussian distributions, the dimensions of the latent space can be motivated to be linearly independent[63]. This is achieved by using a a variational autoencoder[10] which uses Gaussian re-sampling to generate the latent space representations such that they are uncorrelated, as illustrated in fig. 5.3. This is motivated through the loss function, where the Kullback-Leibler divergence between the distribution of the latent space representations and uncorrelated Gaussian distributions is an element. The more the latent representations resemble uncorrelated, and thus linearly independent, variables, the lower the loss is. How strongly this affects the learning process is managed by a prescribed parameter β in the autoencoder to give the architecture flexibility to exploit this feature more fully. In terms of chemistry, this means that the model is encouraged to associate Raman features that correlate strongly, such as those belonging to a specific chemical, with one specific dimension of the latent space rather than representing this information over multiple dimensions. This give the autoencoder an advantage over simpler decomposition

methods, such as principal component analysis.

5.2 De-noising autoencoder

In addition to decomposing input data, effectively training the encoder as a feature extractor network, autoencoders can also be trained to make modifications to an input. This is readily achieved by modifying the data before passing it to the encoder, while comparing the reconstruction from the decoder with the unmodified data. By using the known spectrum of SiO_2 waveguides, as described in chapter 3.4.2 and Paper I, as a source of additive noise we can train the model to remove the waveguide background, as shown in fig. 5.4. The architecture is constructed from the CNN developed for the silicon-project in chapter 4 and Paper II, albeit with different hyper-parameters. The CNN is repurposed to serve as the encoder, and a reversed version, replacing the convolution and MaxPool elements with deconvolution and upsampling, is made to serve as the decoder. Thus, the same features that gave good performance for the CNN in the silicon-project are inherited by the Raman autoencoder. It also inherits the same modular construction and its use of hyper-parameters as genes. This also enables it to evolve as the CNN did, albeit with 23 hyper-parameters describing the deeper autoencoder.
The autoencoder is further augmented by the introduction of a Fourier element in the loss function. When the data passes through the autoencoder, especially the narrow latent space, some information is invariably lost, resulting in a difference between the reconstruction and the original. When the loss of the reconstruction is based purely on the reconstructed vector, for instance the root mean square of the difference between the reconstruction and original, most of the lost information is in the sharp features of the data, effectively making the process a low-pass filter. This degrades sharp features in the spectra, such as the sharp peak of the Phenylalanin, which occurs at 1006.3 cm^{-1} and is commonly used as a protein indicator. Potentially significant information is thus vulnerable to degradation. The Fourier element compares the Fourier transform of the reconstruction to the original. This motivates the autoencoder is to preserve information that presents with either a high amplitude in the data or in the Fourier transform of the data, or both. Sharp features, which with high amplitudes in Fourier space, will contribute stronger to the loss if degraded. Using

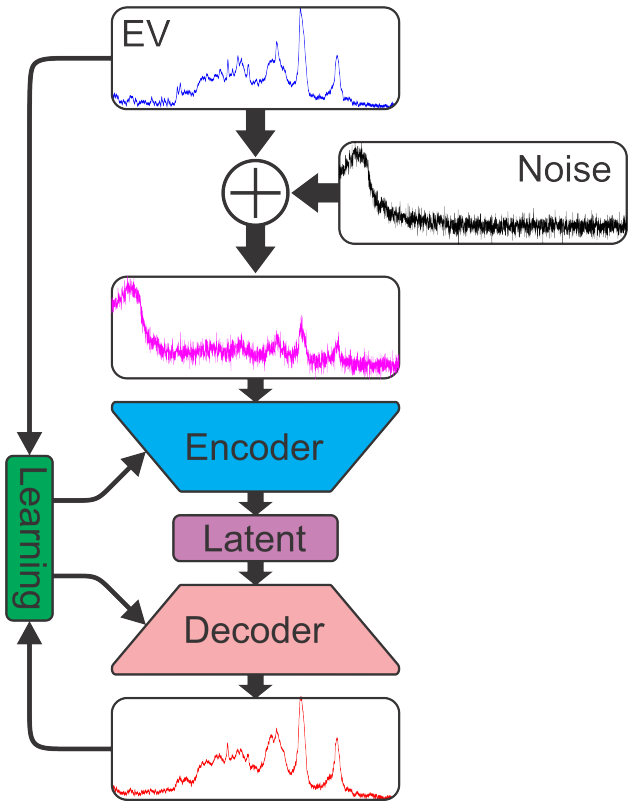


Figure 5.4: Schematic of the training of a de-noising autoencoder. The original data, being Raman spectra of EVs, is combined with noise corresponding to the background of SiO_2 waveguides before being fed to the encoder. The reconstructed spectra are compared with the original Raman spectra such that the model learns to remove the introduced noise.

both the normal and Fourier loss therefore encourages the reconstruction to preserve both sharp and blunt features better, allowing the autoencoder to learn a wider variety of patterns in the spectra.
The architecture shown in fig. 5.5 is created from the described framework through evolution. The model is trained on a dataset created from 279 Raman spectra of blood platelet derived EVs. The dataset is augmented with randomly generated Gaussian white noise, reducing the signal-to-noise ratio to from approximately 22 dB to 13 dB, or a factor of eight reduction. This emulates a condition where the particle size is halved, or if the power density of the field at the particle is reduced by a factor 8, equating to a mode width of 1 μm. A randomly selected background level is added to the noisy signal, bringing the mean signal-to-noise ratio down to -18±3 dB. This is done to emulate the expected background level in a waveguide trap as described in chapter 3.5 and shown in fig. 3.6.
The trained network is subsequently tested on 81 isolated spectra to verify the quality of the reconstructed spectra, as shown in fig. 5.6. The recreations are quite similar in comparison to the true noise-free original spectra and they clearly express the features in the spectra, both the low-frequency complexes and the high-frequency peaks. The increase in SNR is significant, increasing by an average of 20.5dB, or a factor of 105 from the noisy input spectra to the reconstructed spectra. With the exception of one peak due to numerical error in the data, the performance is reliable and excellent, increasing the signal-to-noise ratio from -18±3 dB to 5.4 dB.
The de-noising performance of the autoencoder shows that the background spectrum induced in a SiO_2 waveguide can be overcome. Demonstrating that the Raman spectra of EVs can be recovered with high fidelity, despite being heavily contaminated by a waveguide background with a random magnitude. The models ability to overcome Gaussian noise in addition to the background also demonstrates that Raman spectra of EVs can be collected with a significant reduction in the delivered power to the particle, showing that sufficient Raman scattering can be generated with a large mode size.
By passing artificially created latent representations to the decoder and observing the created spectra, we can determine the spectral components the autoencoder associates with each dimension. It was found that one of the components agree well with what is expected for a protein, and one of the components agree well with what is expected from a lipid, as shown in fig. 5.7. Using the latent dimensions considered to be associated with lipids

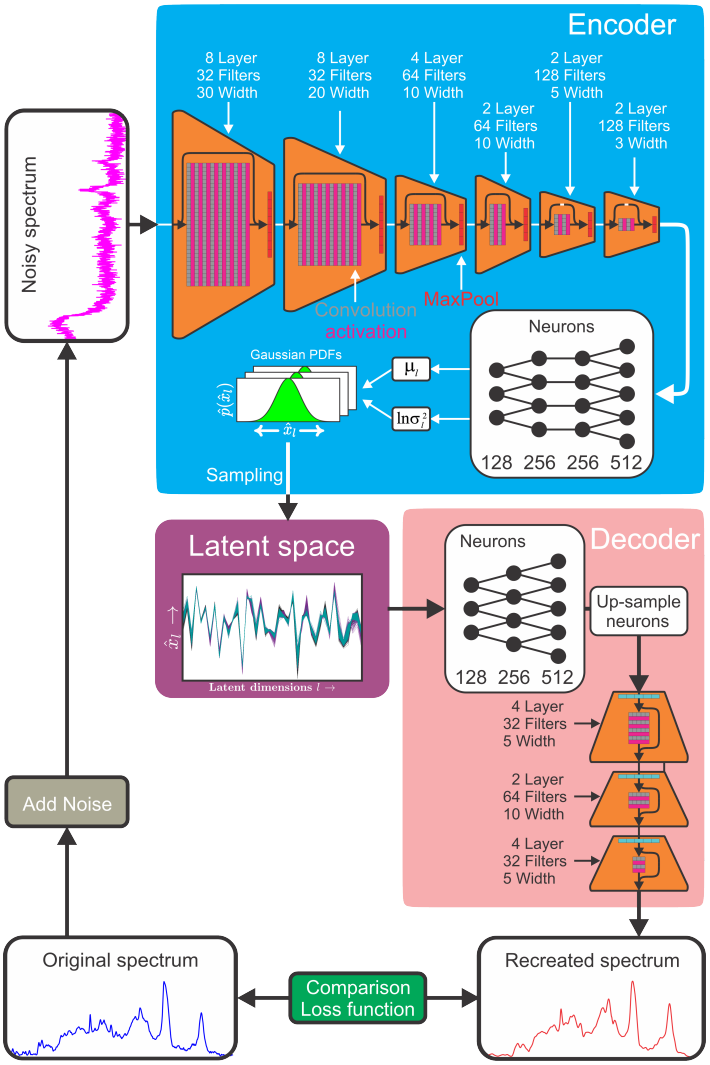
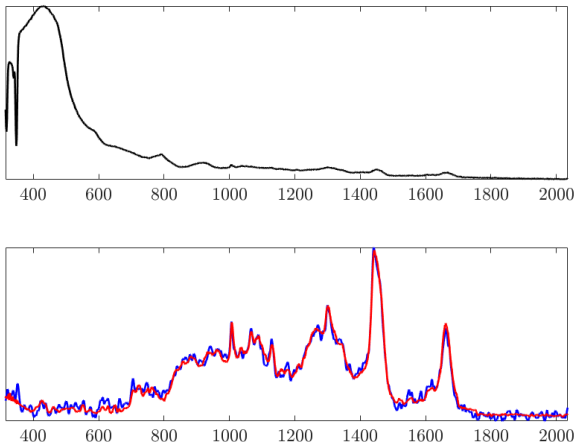
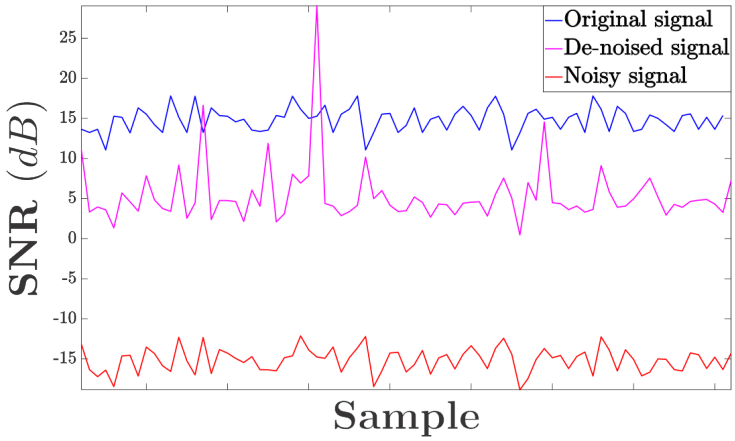


Figure 5.5: Schematic of the de-noising autoencoder. The encoder takes a noisy spectrum as its input and runs it through six blocks of convolutional filters with a total depth of 26 filter layers. The filtered spectrum is then passed through four layers of fully-connected neurons that predict the mean and variance of the latent Gaussian distributions. The latent space data is sampled from the described Gaussian PDFs. The decoder takes the latent data and runs it through three layers of fully-connected neurons followed by three blocks of up-sampling with convolutional filters. The reconstructed spectrum generated by the decoder is compared with the noise-free original spectrum to compute the loss.



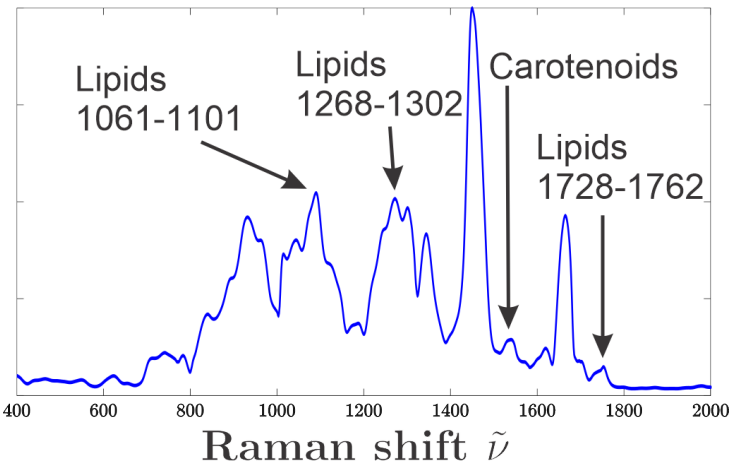
(a) Comparison between the noisy spectrum (black), the original spectrum (blue), and reconstructed spectrum (red).



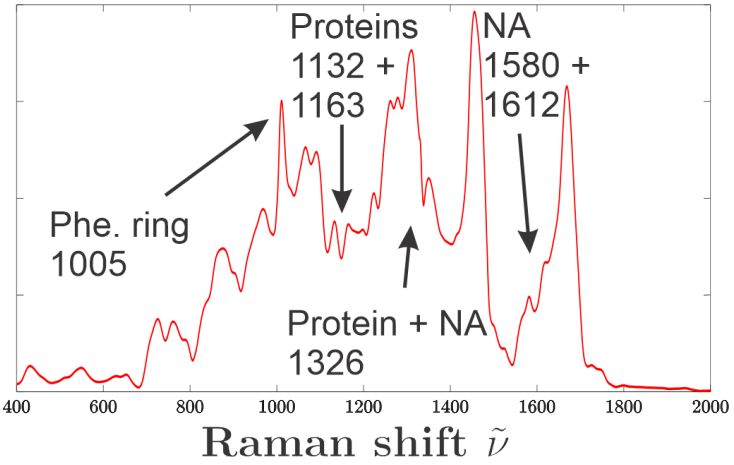
(b) Plot of the signal-to-noise ratio of the original, noisy, and recreated spectra.

Figure 5.6: Graphs showing the performance of the autoencoder. The spectra are reconstructed with a high level of accuracy in the presence of noise. The autoencoder is able to reconstruct the spectra with a mean SNR of 5.4dB, an increase of 20.5dB from the noisy input spectra.

and proteins, we investigate the distribution of the EV spectra in the la-



(a) Lipid associated component with assignments.



(b) Protein associated component with assignments.

Figure 5.7: Components created by probing the autoencoder latent space. The component shown in a) exhibits features consistent with both lipids and carotenoids, therefore it is considered to be associated with lipid content in the EVs. The component shown in b) exhibits features consistent with proteins and nucleic acids (NA), notably the clear Phenylalanin-peak at 1006.3 cm^{-1} (here 1005 cm^{-1}), and is therefore considered to be associated with the protein content in the EVs.

tent space. By passing the spectrum of an EV through to the encoder and intercepting the latent space representation, we form a datapoint in the la-

tent space, reflecting how the model views the spectrum. By specifically looking at the lipid and protein associated dimensions, we see that the model recognizes differences that correspond with the nature of the EVs, as shown in fig. 5.8. The platelet and THP-1 derived EVs form two distinct clusters that are well separated. This indicates that the model sees a clear difference in the EVs originating from the two sources, which agrees well with expectations as platelets and monocytes, like THP-1 cells, have very different biological functions. Perhaps most interesting is the results from the platelet derived EVs where we see that the unactivated platelet derived EVs in fig. 5.8a form a wide distribution with some registering strongly in the lipid associated dimension and some in the protein associated dimension. In contrast, we see that the activated platelets, both with thrombin and calcium, form a much tighter cluster in the high-lipid end of the control distribution. This indicates that the model sees them as both more homogeneous and that they consistently agree more with the lipid associated component. This is in agreement with the expectation, as we expect that the activated platelets will begin producing EVs sending a common signal, and thus their EVs will become more homogeneous. It also agrees with the expectation that the activated platelet derived EVs should contain little protein. Fig. 5.8b shows the latent representations of the same EVs, but with the same type and magnitude of noise as we expect to see in the waveguide. The distributions are smeared out compared to the distributions in fig. 5.8a, taking on an "L"-shape. However, the THP-1 and platelet derived EVs are still well differentiated, and the same trend is present in the platelet derived EVs as seen in fig. 5.8a.

5.3 Differentiation model for EVs

A new model was created to specialize in the differentiation of EVs using their Raman spectra, building on the de-noising autoencoder architecture. To achieve this, a larger sample set was required as the current set only included 281 spectra. It would be too time consuming to significantly expand the set using an existing confocal setup, both due to preparation of the sample and the measurement time. Instead, additional measurements from another lab (Medical Cell BioPhysics, University of Twente, The Netherlands) were included in the set, increasing the number of spectra to 2667. While this significantly increased the sample size, it also introduced chal-

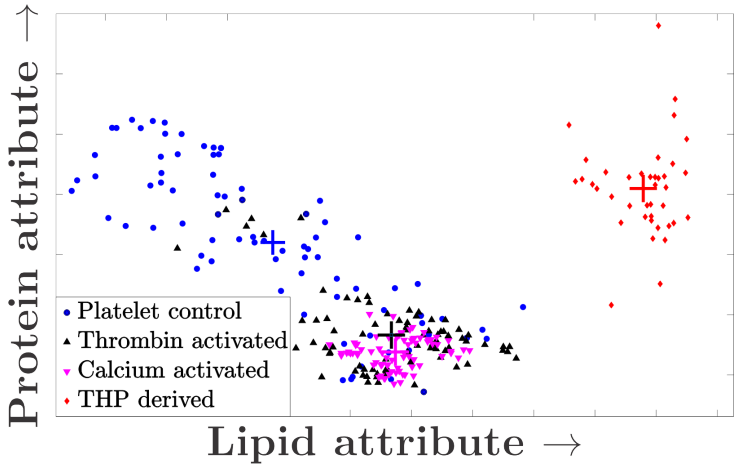
lenges due to the separate origins of the measurements. Uncertainty regarding the chemical status of the samples and the influence of the preparation on it also introduced challenges to the work.

5.3.1 Adaptive frequency architecture

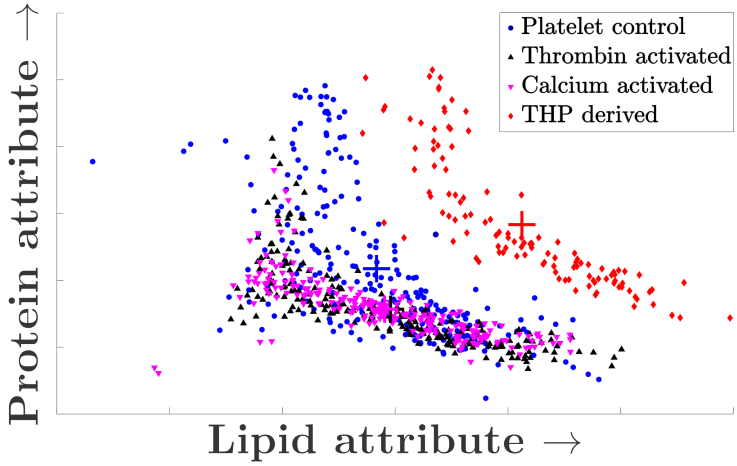
The most significant differences between the datasets from the two labs are the spectral range and resolution of the measurements. The measurements used for the de-noising autoencoder have a common spectral range of 1726 cm^{-1} (from 306 cm^{-1} to 2032 cm^{-1}) while the measurements from Twente have a range of up to 3371 cm^{-1} (from 304 cm^{-1} to 3675 cm^{-1}). The sample set is thus highly heterogeneous. There is also a significant difference in the noise level of the two measurements, which pose a further challenge for the model. Thus, to be able to use datasets from both sources in training the same autoencoder, the training scheme must be configured such that these system/lab-dependent factors do not influence the learning. One way of achieving this is to normalize the spectra, for instance by truncating and interpolating the spectra to the same spectral range and resolution. However, the strength of neural networks is their ability to adapt to different conditions. Applying efforts to clean the dataset fails to exploit this strength. Thus, instead of altering the dataset to make it palatable to the network, we instead investigate making the network able to adapt to differences in the datasets.

The primary way of achieving adaptability is by having the network consider the spectral range explicitly rather than implicitly, by passing information about the spectral range and resolution explicitly to the network, as is described in Paper III. By allowing it access to the spectral information, the network is provided with the information needed to adapt to changes in the spectral range such that information from different sources can be processed by the same algorithm. This is achieved by passing the spectra to the autoencoder on two parallel channels, one that contains the intensities of the Raman spectra and one that contains the wavenumber shifts corresponding to the intensities. As the wavenumber shifts follow a predictable form, a fourth degree polynomial, this can be parameterized instead of passing each the entire vector through the network. The parameters bypass the convolutional portion of the network and are merged with the intensity-related information in the neuron section of the network. The two forms of information can then intermingle in the neuron section before

being passed to the latent space by the encoder. The decoder processes the mixed information in order to reconstruct and separate the two information streams again, with the intensity-related information passing through deconvolution layers and the wavenumber-related information passing into a polynomial function to reconstruct the wavenumber vector originally passed to the encoder. Further details and results on application to EVs and other nanoparticles are presented in Paper III.



(a) Scatter plot of the latent space representations of the Raman spectra of EVs without the presence of artificial noise in the spectrum, as shown in fig.5.2.



(b) Scatter plot of the latent space representations of the Raman spectra of EVs in the presence of Gaussian noise and waveguide background, as shown in fig.5.4.

Figure 5.8: Scatter plots of EV spectra in the latent space of the de-noising autoencoder. The scatter plot in a) is generated using the source spectra, without any added noise, resulting in two well differentiated clusters corresponding to the THP-1 and platelet derived EVs. It is shown that the the EVs from activated platelts form a distinct sub-cluster in the larger cluster of the un-activated control platelets. In b) the same source data is used, only with added Gaussian noise and waveguide background, resulting in "L"-shaped clusters. The same behaviour is demonstrated in b) as in a) with the THP-1 and platelet derived EV distributions remaining differentiated, albeit with less separation. The same sub-clustering of the activated paltelet derived EVs is also present in b).

Chapter 6

Conclusions and future work

6.1 Conclusions

This thesis studies the prospect of using Raman spectroscopy as a tool for analysing extracellular vesicles and other biological nanoparticles, and presents a machine learning architecture for the analysis of the spectra. UV-written SiO_2 waveguides are evaluated as a candidate for a waveguide-based Raman-on-chip device. The UV-written SiO_2 waveguides are characterized, with focus on the background generated by the induced Raman scattering and how the background impacts measured spectra of nanoparticles, to evaluate the viability of a SiO_2-based Raman-on-chip device. A variational autoencoder model is developed to compensate for the generated noise and background in a waveguide device such that the spectrum of a measured nanoparticle can be recovered independent of the background level. The autoencoder is developed further to allow it to consider Raman spectra with variable resolution and noise, and is shown to be able to differentiate nanoparticles well.
In Paper I, waveguides fabricated by UV-writing in SiO_2 are characterized to evaluate their induced Raman scattering. The Raman scattering of several waveguides with different fabrication parameters is measured using two different excitation wavelengths. The generated Raman spectra are calibrated using a secondary laser source such that the magnitude of the spectra is expressed in absolute units. Measurements reveal a significantly lower Raman background in UV-written SiO_2 waveguides compared to Si_3N_4, but that the SiO_2 waveguides still produce a stronger background

than comparable fibres. Measurements of a particle analogue give an estimated achievable signal-to-noise ratio of -15 dB for a 100 nm diameter particle when trapped by a SiO_2 waveguide device.
In Paper II, a convolutional neural network model is developed for determining the quality of monocrystalline silicon from infrared transmission tomography. Three samples of silicon boules are subjected to a full tomographic scan using a near infrared laser source. The measured transmission profiles for each step along the longitudinal axis of the boule is processed by a convolutional neural network to determine the quality of the crystalline structure of the silicon. The network is created through a process of evolution using a genetic algorithm to determine the architecture of the network. The developed model achieved a 92.2% accuracy in predicting the intactness of the monocrystalline structure of the silicon.
In Paper III, the convolutional neural network of Paper II is expanded into a variational autoencoder designed for feature extraction from Raman spectra of biological nanoparticles. The developed autoencoder uses a convolutional encoder to compress the spectra into a 100 dimensional latent space containing the extracted features. The novelty of the autoencoder is the explicit consideration of the frequency aspect of the spectra, allowing the model access to both the intensity at each point in the spectra as well as the wavenumber at each point. This allowed the model to accept spectra with variable spectral resolution and range, enabling it to use data from two different laboratories with different measurement systems. The model is also shown to be robust against noise and distortions in the spectra, demonstrating significant de-noising capabilities. The extracted features of the Raman spectra are used to classify the spectra to the biological origin of the particles, achieving an accuracy of 92.2% and the ability to discern, among others, particles from prostate cancer patients from non-cancer controls with very high accuracy.

6.2 Future work

Due to technical challenges, the fabricated prototypes for the Raman-on-chip device were not viable, and due to limited access to fabrication facilities, further prototypes could not be fabricated. The future work on these devices would therefore include the fabrication of new chips with the prototype 3 trap design (wires). Once fabricated, the devices will be tested

using 1 μm and 0.5 μm polystyrene beads before continuing to biological nanoparticles, primarily extracellular vesicles. With successful trapping of nanoparticles in a single trap, the set-up will be modified to collect from multiple traps on the chip.
The developed machine learning method, using an adaptive frequency autoencoder, can be expanded by introducing more samples from additional laboratories. The goal of this work is to expand its capabilities beyond the samples it has been trained on, including larger particles such as staphylococcus aureus bacteria and cellular elements, such that its applicability can be broadened to cover more biological samples. Future work on this model also includes an expanded study into its potential to de-noise spectra with very low signal levels and its capabilities in reconstructing spectra from incomplete input spectra.

Bibliography

[1] Aftab Ahmed and Reuven Gordon. Directivity enhanced raman spectroscopy using nanoantennas. *Nano letters*, 11(4):1800–1803, 2011.

[2] Yuta Akihama, Yoshiaki Kanamori, and Kazuhiro Hane. Ultra-small silicon waveguide coupler switch using gap-variable mechanism. *Opt. Express*, 19(24):23658–23663, Nov 2011.

[3] Anush Arakelyan, Wendy Fitzgerald, Sonia Zicari, Christophe Vanpouille, and Leonid Margolis. Extracellular vesicles carry hiv env and facilitate hiv infection of human lymphoid tissue. *Scientific reports*, 7(1):1695, 2017.

[4] A. Ashkin, J. M. Dziedzic, J. E. Bjorkholm, and Steven Chu. Observation of a single-beam gradient force optical trap for dielectric particles. *Opt. Lett.*, 11(5):288–290, May 1986.

[5] Sencer Ayas, Hasan Guner, Burak Turker, Okan Oner Ekiz, Faruk Dirisaglik, Ali Kemal Okyay, and Aykutlu Dˆana. Raman enhancement on a broadband meta-surface. *ACS nano*, 6(8):6852–6861, 2012.

[6] Leonid Yu. Beliaev, Evgeniy Shkondin, Andrei V. Lavrinenko, and Osamu Takayama. Optical, structural and composition properties of silicon nitride films deposited by reactive radio-frequency sputtering, low pressure and plasma-enhanced chemical vapor deposition. *Thin Solid Films*, 763:139568, 2022.

[7] K. Blotekjaer. Thermal noise in optical fibers and its influence on long-distance coherent communication systems. *Journal of Lightwave Technology*, 10(1):36–41, 1992.

[8] Martijn Boerkamp, Thijs van Leest, Jeroen Heldens, Arne Leinse, Marcel Hoekman, Rene Heideman, and Jacob Caro. On-chip optical trapping and raman spectroscopy using a triplex dual-waveguide trap. *Opt. Express*, 22(25):30528–30537, Dec 2014.

[9] Deane Chandler-Horowitz and Paul M Amirtharaj. High-accuracy, midinfrared (450cm- 1 ω 4000cm- 1) refractive index values of silicon. *Journal of Applied physics*, 97(12), 2005.

[10] Tingting Chen, Xueping Liu, Bizhong Xia, Wei Wang, and Yongzhi Lai. Unsupervised anomaly detection of industrial robots using sliding-window convolutional variational autoencoder. *IEEE Access*, 8:47072–47081, 2020.

[11] Lesley Cheng and Andrew F Hill. Therapeutically harnessing extracellular vesicles. *Nature Reviews Drug Discovery*, 21(5):379–399, 2022.

[12] Maria Chiara Ciferri, Rodolfo Quarto, and Roberta Tasso. Extracellular vesicles as biomarkers and therapeutic tools: From pre-clinical to clinical applications. *Biology*, 10(5):359, 2021.

[13] Daniel Colladon. On the reflections of a ray of light inside a parabolic liquid stream. *Comptes Rendus*, 15(800-802):15, 1842.

[14] Eliana Cordero, Ines Latka, Christian Matthäus, Iwan W. Schie, and Jürgen Popp. In-vivo Raman spectroscopy: from basics to applications. *Journal of Biomedical Optics*, 23(7):071210, 2018.

[15] Ashim Dhakal, Ananth Subramanian, Roel Baets, and Nicolas Le Thomas. The role of index contrast in the efficiency of absoprtion and emission of a luminescent particle near a slab waveguide. In *16th European Conference on Integrated Optics (ECIO-2012)*, 2012.

[16] Ashim Dhakal, Ananth Z. Subramanian, Pieter Wuytens, Frédéric Peyskens, Nicolas Le Thomas, and Roel Baets. Evanescent excitation and collection of spontaneous raman spectra using silicon nitride nanophotonic waveguides. *Opt. Lett.*, 39(13):4025–4028, Jul 2014.

[17] Sebastian Dochow, Martin Becker, Ron Spittel, Claudia Beleites, Sarmiza Stanca, Ines Latka, Kay Schuster, Jens Kobelke, Sonja Unger, Thomas Henkel, Günter Mayer, Jens Albert, Manfred Rothhardt,

Christoph Krafft, and Jürgen Popp. Raman-on-chip device and detection fibres with fibre bragg grating for analysis of solutions and particles. *Lab Chip*, 13:1109–1113, 2013.

[18] Sebastian Dochow, Christoph Krafft, Ute Neugebauer, Thomas Bocklitz, Thomas Henkel, Günter Mayer, Jens Albert, and Jürgen Popp. Tumour cell identification by means of raman spectroscopy in combination with optical traps and microfluidic environments. *Lab Chip*, 11:1484–1490, 2011.

[19] Agustin Enciso-Martinez, Edwin Van Der Pol, Chi M. Hau, Rienk Nieuwland, Ton G. Van Leeuwen, Leon W.M.M. Terstappen, and Cees Otto. Label-free identification and chemical characterisation of single extracellular vesicles and lipoproteins by synchronous rayleigh and raman scattering. *Journal of Extracellular Vesicles*, 9(1):1730134, 2020.

[20] Mohamed A. Ettabib, Almudena Marti, Zhen Liu, Bethany M. Bowden, Michalis N. Zervas, Philip N. Bartlett, and James S. Wilkinson. Waveguide enhanced raman spectroscopy for biosensing: A review. *ACS Sensors*, 6(6):2025–2045, JUN 25 2021.

[21] Scott Foster. Low-frequency thermal noise in optical fiber cavities. *Physical Review A*, 86(4), OCT 1 2012.

[22] Scott Foster, Alexei Tikhomirov, and Mark Milnes. Fundamental thermal noise in distributed feedback fiber lasers. *IEEE Journal of Quantum Electronics*, 43(5):378–384, 2007.

[23] Renee R Frontiera, Anne-Isabelle Henry, Natalie L Gruenke, and Richard P Van Duyne. Surface-enhanced femtosecond stimulated raman spectroscopy. *The journal of physical chemistry letters*, 2(10):1199–1203, 2011.

[24] Yongkang Gao, Qiaoqiang Gan, Zheming Xin, Xuanhong Cheng, and Filbert J Bartoli. Plasmonic mach–zehnder interferometer for ultrasensitive on-chip biosensing. *ACS nano*, 5(12):9836–9844, 2011.

[25] Chris Gardiner, Michael Shaw, Patrick Hole, Jonathan Smith, Dionne Tannetta, Christopher W Redman, and Ian L Sargent. Measurement

of refractive index by nanoparticle tracking analysis reveals heterogeneity in extracellular vesicles. *Journal of extracellular vesicles*, 3(1):25361, 2014.

[26] Paul C. Gow, Rex H. S. Bannerman, Paolo L. Mennea, Christopher Holmes, James C. Gates, and Peter G. R. Smith. Direct uv written integrated planar waveguides using a 213 nm laser. *Opt. Express*, 27(20):29133–29138, Sep 2019.

[27] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, and Tsuhan Chen. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377, 2018.

[28] Yuejiao Gu, Shuping Xu, Haibo Li, Shaoyan Wang, Ming Cong, John R. Lombardi, and Weiqing Xu. Waveguide-enhanced surface plasmons for ultrasensitive sers detection. *The Journal of Physical Chemistry Letters*, 4(18):3153–3157, 2013.

[29] Alice Gualerzi, Sander Alexander Antonius Kooijmans, Stefania Niada, Silvia Picciolini, Anna Teresa Brini, Giovanni Camussi, and Marzia Bedoni. Raman spectroscopy as a quick tool to assess purity of extracellular vesicle preparations and predict their functionality. *Journal of Extracellular Vesicles*, 8(1):1568780, 2019. PMID: 30728924.

[30] Yasuhiro Harada and Toshimitsu Asakura. Radiation forces on a dielectric sphere in the rayleigh scattering regime. *Optics communications*, 124(5-6):529–541, 1996.

[31] Øystein Ivar Helle, Firehun Tsige Dullo, Marcel Lahrberg, Jean-Claude Tinguely, Olav Gaute Hellesø, and Balpreet Singh Ahluwalia. Structured illumination microscopy using a photonic chip. *Nature photonics*, 14(7):431–438, 2020.

[32] S Inaba, S Oda, and K Morinaga. Heat capacity of oxide glasses measured by ac calorimetry. *Journal of Non-Crystalline Solids*, 306(1):42–49, JUL 2 2002.

[33] Heera Jayan, Hongbin Pu, and Da-Wen Sun. Recent developments in raman spectral analysis of microbial single cells: Techniques and appli-

cations. *Critical Reviews in Food Science and Nutrition*, 62(16):4294–4308, 2022. PMID: 34251940.

[34] Mathias N Jensen. Mir-based in-situ measurement of silicon crystal-melt interface. Master's thesis, UiT Norges arktiske universitet, 2020.

[35] Mathias Novik Jensen and Olav Gaute Hellesø. Measuring the end-face of silicon boules using mid-infrared laser scanning. *CrystEngComm*, 23(26):4648–4657, 2021.

[36] K Charles Kao and George A Hockham. Dielectric-fibre surface waveguides for optical frequencies. In *Proceedings of the Institution of Electrical Engineers*, volume 113, pages 1151–1158. IET, 1966.

[37] Lingbo Kong, Changwon Lee, Christopher M. Earhart, Bernardo Cordovez, and James W. Chan. A nanotweezer system for evanescent wave excited surface enhanced raman spectroscopy (sers) of single nanoparticles. *Opt. Express*, 23(5):6793–6802, Mar 2015.

[38] Sergei G. Kruglik, Felix Royo, Jean-Michel Guigner, Laura Palomo, Olivier Seksek, Pierre-Yves Turpin, Irene Tatischeff, and Juan M. Falcon-Perez. Raman tweezers microspectroscopy of circa 100 nm extracellular vesicles. *Nanoscale*, 11(4):1661–1679, JAN 28 2019.

[39] Philipp Kukura, David W McCamant, and Richard A Mathies. Femtosecond stimulated raman spectroscopy. *Annu. Rev. Phys. Chem.*, 58:461–488, 2007.

[40] Jason M. Larkin and Alan J. H. McGaughey. Thermal conductivity accumulation in amorphous silica and amorphous silicon. *Phys. Rev. B*, 89:144303, Apr 2014.

[41] Joseph Larmor. Xv. the principle of molecular scattering of radiation. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 37(217):161–163, 1919.

[42] Nicolas Le Thomas, Ashim Dhakal, Ali Raza, Frédéric Peyskens, and Roel Baets. Impact of fundamental thermodynamic fluctuations on light propagating in photonic waveguides made of amorphous materials. *Optica*, 5(4):328–336, Apr 2018.

[43] H. J. Lee, F. Abdullah, S. D. Emami, and A. Ismail. Fiber modeling and simulation of effective refractive index for tapered fiber with finite element method. In *2016 IEEE 6th International Conference on Photonics (ICP)*, pages 1–3, 2016.

[44] G. Lepert, M. Trupke, E. A. Hinds, H. Rogers, J. C. Gates, and P. G. R. Smith. Demonstration of uv-written waveguides, bragg gratings and cavities at 780 nm, and an original experimental measurement of group delay. *Opt. Express*, 19(25):24933–24943, Dec 2011.

[45] Gyllion B. Loozen, Arnica Karuna, Mohammad M. R. Fanood, Erik Schreuder, and Jacob Caro. Integrated photonics multi-waveguide devices for optical trapping and raman spectroscopy: design, fabrication and performance demonstration. *Beilstein Journal of Nanotechnology*, 11:829–842, 2020.

[46] Xingsheng Luan, Jean-Baptiste Béguin, Alex P Burgers, Zhongzhong Qin, Su-Peng Yu, and Harry J Kimble. The integration of photonic crystal waveguides with atom arrays in optical tweezers. *Advanced Quantum Technologies*, 3(11):2000008, 2020.

[47] Ramasamy Manoharan, Yang Wang, and Michael S Feld. Histochemical analysis of biological tissues using raman spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 52(2):215–249, 1996.

[48] Jonathan Masci, Ueli Meier, Dan Ciresan, and Juergen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In T Honkela, W Duch, M Girolami, and S Kaski, editors, *ARTIFICIAL NEURAL NETWORKS AND MACHINE LEARNING - ICANN 2011, PT I*, volume 6791 of *Lecture Notes in Computer Science*, pages 52–59. Aalto Univ, Sch Sci, Dept Informat & Comp Sci, 2011. 21st International Conference on Artificial Neural Networks, ICANN 2011, Aalto Univ Sch Sci, Espoo, FINLAND, JUN 14-17, 2011.

[49] David W McCamant, Philipp Kukura, and Richard A Mathies. Femtosecond time-resolved stimulated raman spectroscopy: Application to the ultrafast internal conversion in β-carotene. *The Journal of Physical Chemistry A*, 107(40):8208–8214, 2003.

[50] Zanyar Movasaghi, Shazza Rehman, and Ihtesham U Rehman. Raman spectroscopy of biological tissues. *Applied Spectroscopy Reviews*, 42(5):493–541, 2007.

[51] Randa Mrad, Sergei G Kruglik, Nassim Ben Brahim, Rafik Ben Chaabane, and Michel Negrerie. Raman tweezers microspectroscopy of functionalized 4.2 nm diameter cdse nanocrystals in water reveals changed ligand vibrational modes by a metal cation. *The Journal of Physical Chemistry C*, 123(40):24912–24918, 2019.

[52] T.A. Nieminen, H. Rubinsztein-Dunlop, N.R. Heckenberg, and A.I. Bishop. Numerical modelling of optical trapping. *Computer Physics Communications*, 142(1):468–471, 2001. Conference on Computational Physics 2000: "New Challenges for the New Millenium".

[53] K. Ruud P. Norman and T. Saue. *Principles and Practices of Molecular Properties: Theory, Modeling, and Simulations*. Wiley, 2018.

[54] K.S. Packard. The origin of waveguides: A case of multiple rediscovery. *IEEE Transactions on Microwave Theory and Techniques*, 32(9):961–969, 1984.

[55] G. Pan, N. Yu, B. Meehan, T. W. Hawkins, J. Ballato, and P. D. Dragic. Thermo-optic coefficient of b2o3 and geo2 co-doped silica fibers. *Opt. Mater. Express*, 10(7):1509–1521, Jul 2020.

[56] Jelle Penders, Anika Nagelkerke, Eoghan M Cunnane, Simon V Pedersen, Isaac J Pence, R Charles Coombes, and Molly M Stevens. Single particle automated raman trapping analysis of breast cancer cell-derived extracellular vesicles as cancer biomarkers. *ACS nano*, 15(11):18192–18205, 2021.

[57] Frédéric Peyskens, Pieter Wuytens, Ali Raza, Pol Van Dorpe, and Roel Baets. Waveguide excitation and collection of surface-enhanced raman scattering from a single plasmonic antenna. *Nanophotonics*, 7(7):1299–1306, 2018.

[58] Giuseppe Pezzotti. Raman spectroscopy in cell biology and microbiology. *Journal of Raman Spectroscopy*, 52(12):2348–2443, 2021.

[59] Michelle L Pleet, Sean Cook, Vera A Tang, Emily Stack, Verity J Ford, Joanne Lannigan, Ngoc Do, Ellie Wenger, Jean-Luc Fraikin, Steven Jacobson, et al. Extracellular vesicle refractive index derivation utilizing orthogonal characterization. *Nano Letters*, 2023.

[60] Richard C Prince, Renee R Frontiera, and Eric O Potma. Stimulated raman scattering: from bulk to nano. *Chemical reviews*, 117(7):5070–5094, 2017.

[61] CV Raman. Diffraction by molecular clusters and the quantum structure of light. *Nature*, 109(2736):444–445, 1922.

[62] CV Raman and KS Krishnan. A new class of spectra due to secondary radiation. part i. 1928.

[63] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In EP Xing and T Jebara, editors, *International Conference On Machine Learning, Vol 32 (cycle 2)*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, 2014. International Conference on Machine Learning, Bejing, PEOPLES R CHINA, JUN 22-24, 2014.

[64] Carsten Schinke, P Christian Peest, Jan Schmidt, Rolf Brendel, Karsten Bothe, Malte R Vogt, Ingo Kröger, Stefan Winter, Alfred Schirmacher, Siew Lim, et al. Uncertainty analysis for the coefficient of band-to-band absorption of crystalline silicon. *Aip Advances*, 5(6), 2015.

[65] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[66] Adolf Smekal. Zur quantentheorie der dispersion. *Naturwissenschaften*, 11(43):873, 1923.

[67] G. G. Stokes. Xxxiv.—on the application of the optical properties of bodies to the detection and discrimination of organic substances. *J. Chem. Soc.*, 17:304–318, 1864.

[68] Jeffrey L Suhalim, Chao-Yu Chung, Magnus B Lilledahl, Ryan S Lim, Moshe Levi, Bruce J Tromberg, and Eric O Potma. Characterization

of cholesterol crystals in atherosclerotic plaques using stimulated raman scattering and second-harmonic generation microscopy. *Biophysical journal*, 102(8):1988–1995, 2012.

[69] Abdullah Chandra Sekhar Talari, Zanyar Movasaghi, Shazza Rehman, and Ihtesham Ur Rehman. Raman spectroscopy of biological tissues. *Applied spectroscopy reviews*, 50(1):46–111, 2015.

[70] Nathan F. Tyndall, Todd H. Stievater, Dmitry A. Kozak, Kee Koo, R. Andrew McGill, Marcel W. Pruessner, William S. Rabinovich, and Scott A. Holmstrom. Waveguide-enhanced raman spectroscopy of trace chemical warfare agent simulants. *Opt. Lett.*, 43(19):4803–4806, Oct 2018.

[71] Raghavendra Upadhya and Ashok K Shetty. Extracellular vesicles for the diagnosis and treatment of parkinson's disease. *Aging and disease*, 12(6):1438, 2021.

[72] Marek Vlk, Anurup Datta, Sebastián Alberti, Henock Demessie Yallew, Vinita Mittal, Ganapathy Senthil Murugan, and Jana Jágerská. Extraordinary evanescent field confinement waveguide sensor for mid-infrared trace gas spectroscopy. *Light: Science & Applications*, 10(1):26, 2021.

[73] Pengyi Wang and Benjamin L Miller. Waveguide-enhanced raman spectroscopy (wers): an emerging chip-based tool for chemical and biological sensing. *Sensors*, 22(23):9058, 2022.

[74] Qianqian Wang, Bo Xu, Jian Sun, Hanyu Liu, Zhisheng Zhao, Dongli Yu, Changzeng Fan, and Julong He. Direct band gap silicon allotropes. *Journal of the American Chemical Society*, 136(28):9826–9829, 2014.

[75] Mary Elvira Weeks. The discovery of the elements. xxi. supplementary note on the discovery of phosphorus. *Journal of Chemical Education*, 10(5):302, 1933.

[76] Yongwei Xiao, Lei Zheng, Xiaofeng Zou, Jigang Wang, Jianing Zhong, and Tianyu Zhong. Extracellular vesicles in type 2 diabetes mellitus: key roles in pathogenesis, complications, and therapy. *Journal of extracellular vesicles*, 8(1):1625677, 2019.

[77] Allen HJ Yang, Sean D Moore, Bradley S Schmidt, Matthew Klug, Michal Lipson, and David Erickson. Optical manipulation of nanoparticles and biomolecules in sub-wavelength slot waveguides. *Nature*, 457(7225):71–75, 2009.

[78] Delong Zhang, Ping Wang, Mikhail N Slipchenko, Dor Ben-Amotz, Andrew M Weiner, and Ji-Xin Cheng. Quantitative vibrational imaging by hyperspectral stimulated raman scattering microscopy and multivariate curve resolution analysis. *Analytical chemistry*, 85(1):98–106, 2013.

[79] Long Zhang, Ming Zhang, Tangnan Chen, Dajian Liu, Shihan Hong, and Daoxin Dai. Ultrahigh-resolution on-chip spectrometer with silicon photonic resonators. *Opto-Electronic Advances*, 5(7):210100–1, 2022.

Paper I: Demonstrating low Raman background in UV-written SiO_2 waveguides

Published in Optics Express, September 2023.
Authors: Mathias N. Jensen, James C. Gates, Alex I. Flint, and Olav Gaute Hellesø

Contribution notes: Mathias N. Jensen performed all experimental work and data analysis. James C. Gates and Alex I. Flint developed the UV-writing process and fabricated the waveguide chips. Olav Gaute Hellesø conceived the idea and oversaw the work. Mathias N. Jensen wrote the initial draft and Olav Gaute Hellesø finalized the manuscript for submission. All authors contributed to revision of the manuscript before publication.

Research Article | Vol. 31, No. 19 / 11 Sep 2023 / Optics Express 31092
Optics EXPRESS

Demonstrating low Raman background in UV-written SiO_2 waveguides

MATHIAS NOVIK JENSEN,[1] JAMES C. GATES,[2] ALEX I. FLINT,[2] AND OLAV GAUTE HELLESØ[1,*]

[1] *Department of Physics and Technology, UiT The Arctic University of Norway, Hansine Hansens veg 18, 9019 Tromsø, Norway*
[2] *Optoelectronics Research Centre, University of Southampton, University Road, SO17 1BJ Southampton, UK*
[*] *olav.gauge.helleso@uit.no*

Abstract: Raman spectroscopy can give a chemical 'fingerprint' from both inorganic and organic samples, and has become a viable method of measuring the chemical composition of single biological particles. In parallel, integration of waveguides and microfluidics allows for the creation of miniaturized optical sensors in lab-on-a-chip devices. The prospect of combining integrated optics and Raman spectroscopy for Raman-on-chip offers new opportunities for optical sensing. A major limitation for this is the Raman background of the waveguide. This background is very low for optical fibers but remains a challenge for planar waveguides. In this work, we demonstrate that UV-written SiO_2 waveguides, designed to mimic the performance of optical fibers, offer a significantly lower background than competing waveguide materials such as Si_3N_4. The Raman scattering in the waveguides is measured in absolute units and compared to that of optical fibers and Si_3N_4 waveguides. A limited study of the sensitivity of the Raman scattering to changes in pump wavelength and in waveguide design is also conducted. It is revealed that UV-written SiO_2 waveguides offer a Raman background lower than −107.4 dB relative to a 785 nm pump and −106.5 dB relative to a 660 nm pump. Furthermore, the UV-written SiO_2 waveguide demonstrates a 15 dB lower Raman background than a Si_3N_4 waveguide and is only 8.7 – 10.3 dB higher than optical fibers. Comparison with a polystyrene bead (in free space, diameter 7 µm) reveal an achievable peak SNR of 10.4 dB, showing the potential of UV-SiO_2 as a platform for a Raman-on-chip device capable of measuring single particles.

1. Introduction

Waveguide enhanced Raman spectroscopy (WERS) offers long interaction lengths and strong interaction with an analyte by using the evanescent field of a high-index contrast waveguide. Coated nanophotonic silicon nitride waveguides have been used for detecting traces of chemical warfare agent stimulants down to 5 ppb [1]. Micromolar levels of cyclohexane have been probed in aqueous solutions with slot waveguides, again coated and made of silicon nitride [2]. This demonstrates the applicability to both gasses and liquids. The waveguides can be made by standard fabrication methods, are robust and can easily be integrated with microfluidics to make a lab-on-a-chip. Optical components like directional couplers, wavelength filters and grating input couplers can be fully integrated, and further integration or hybrid assembly with lasers and detectors is possible. As a first step towards integration, a packaged, fiber-coupled sensor has been demonstrated, with an integrated directional coupler for splitting the forward-propagating pump and the backward-propagating signal [3]. A recent and excellent review of WERS gives more details about the technique [4]. For biological applications, Raman spectroscopy offers label-free detection and chemical analysis. Currently, confocal Raman microscopy is gaining importance in the fields of biochemistry and microbiology [5–7] due to its ability to selectively sample cells and smaller volumes. For biological particles, the combination of Raman microscopy with optical

#498795 https://doi.org/10.1364/OE.498795
Journal © 2023 Received 23 Jun 2023; revised 18 Aug 2023; accepted 18 Aug 2023; published 6 Sep 2023

Optics EXPRESS

trapping can analyse particles down to the nanoscale [8]. The extension of WERS to applications in microbiology and for the analysis of biological nanoparticles may significantly increase the capabilities of lab-on-a-chip systems for these field. As a first step in this direction, the use of TriPleX waveguides has been proposed [9]. However, there are some significant hurdles to overcome. WERS is an alternative to surface enhanced Raman spectroscopy (SERS), which gives significantly higher Raman enhancement. A comparison of the two methods is beyond the scope of this article and we refer to the many good reviews of SERS, e.g. [10–12].

Among the limitations and challenges for waveguide enhanced Raman spectroscopy are propagation losses that limit the interaction length and the Raman background of the material. The Raman background induced from the waveguide material acts as a noise-signal that overlays the signal from the analyte/sample, thus the detection performance of a WERS device is fundamentally limited by the characteristics of the waveguide material. The application of waveguides for Raman spectroscopy can be divided into three general concepts:

- Evanescent field interaction: Interaction between the evanescent field and a homogeneous analyte along the waveguide length gives a large interaction volume and high sensitivity.
- Direct field interaction: Employing a slot or porous waveguide allows the analyte to intersect with the centre of the guided mode, again potentially giving high sensitivity for a homogeneous analyte.
- Field projection: To analyse a nanoparticle, high intensity at a point is necessary. This can be obtained with a hole in the waveguide, a trench across it or a structure that focus the mode onto the particle. Light is thus not guided, but projected onto the analyte for interaction in a small volume.

While both evanescent and direct field interactions give high coupling efficiencies and sensitivity, they are both most efficient for homogeneous analytes. Our work is focused on nanoparticles, for which field projection into a micron-sized volume is suitable. Thus, an embedded waveguide with negligible evanescent field interaction and high transmission is optimal. Optical fibers have extremely low propagation losses and it has been demonstrated that they also have a very low Raman background [13]. In this work, we investigate silica-based waveguides that are designed and fabricated to mimic the performance of silica fibers. The waveguides are made by UV-writing in doped silica, as will be described later. The waveguides have low refractive index contrast, a relatively large core, low propagation losses and low losses when end-coupling to an optical fiber [14–16]. However, as the waveguides are buried, with silica on all sides, there is no evanescent field available to do Raman spectroscopy of an analyte. Our approach will be to etch structures into the waveguide core (e.g. holes or trenches) for access to the field, for analysis of biological nanoparticles. Using a waveguide to project the trapping/exciting rather than using the evanescent field to excite Raman scattering enables the creation of a compact chip device with one or more micron-scale trapping sites suitable for nanoparticles.

In this article, we investigate the intrinsic properties of the waveguides for Raman scattering, without an etched interaction volume and without an analyte. The waveguides are measured using an in-line measurement scheme with a high-power laser acting as pump and a secondary, low power laser acting as a reference to obtain measurements on an absolute scale. The measurements are repeated for two pump wavelengths (660 nm and 785 nm) to evaluate the wavelength sensitivity of the background. This information will decide the choice of pump laser for future applications. Raman scattering increases with $1/\lambda^4$, possibly making a shorter wavelength favourable, but the background from the waveguide also depends on the wavelength. Thus, an experimental study is necessary to find which wavelength gives the best signal-to-noise ratio for a given waveguide and sample. The results are compared with Raman scattering in an optical fiber and in

Optics EXPRESS

a polystyrene bead (diameter 7 μm), with the first serving as a 'gold' standard and the second to estimate the achievable signal-to-noise ratio for an easily reproducible case. The results are also compared with a model and with published values. A limited study of the influence of waveguide dimensions, composition, and cladding material is included for the UV-written silica waveguides.

The Raman background of four common waveguide materials, Al_2O_3, Si_3N_4, Ta_2O_5, and TiO_2, has been compared previously [17], and its wavelength dependence for Si_3N_4 and Ta_2O_5 was studied by D. Coucheron et al. [18]. Only N. Le Thomas et al. [13] has previously reported absolute values for the Raman background (of Si_3N_4). Here, we report the absolute values for the Raman background of UV-written silica waveguides in comparison to previous measurements of Si_3N_4. Given the similarity between the silica waveguides and silica fibers in terms of mode size and composition, the absolute values for the Raman background of select fibers are also reported here.

The UV-written silica waveguides studied here have a low index contrast relative to those considered in previous works. The relative similarity between the investigated silica waveguides and silica fibers, which have been successfully implemented as Raman-probes, promotes such waveguides as an integrated optics platform for on-chip Raman spectroscopy.

Membrane waveguides has recently emerged as a new waveguide geometry [19], with a thin core surrounded by an analyte. The results presented here are also relevant for silica membrane waveguides, where the index contrast is between (undoped) silica and water. Thus, although the present results are for burried silica waveguides, several approaches can be envisioned to exploit the low Raman background measured by modifying the waveguide geometry locally or along the entire length, by making a membrane waveguide.

2. Model and expectations for Raman background in waveguides

Waveguide enhanced Raman spectroscopy requires the pump laser to propagate through a waveguide core made of a dense material with small cross-section, and it is thus expected that this propagation will generate substantial Raman scattering in the device itself. As a consequence, the Raman scattering collected from an analyte will also contain the Raman scattering from the waveguide as a background signal. This background represents the fundamental noise limit for Raman spectroscopy of the analyte, as stated in the introduction. The Raman spectrum of SiO_2 is readily available [20,21], but its intensity relative to the pump laser is necessary, for the waveguide considered, to compare it with the spectrum of an analyte. Before preceding to this measurement, a model for the background is useful for interpreting the results, although several of the parameters must be obtained by fitting to the measurements. N. Le Thomas *et al.* have proposed a model for the Raman scattering in optical waveguides [13]. The main equation of the model will be described and used here. The model aims to express the fundamental level of the Raman scattering in a dielectric waveguide by considering the stochastic fluctuations of the induced thermal field and the subsequent noise induced in the guided wave. In contrast to previous models based on standard diffusion [22,23], this model is derived from the concept of "frozen" thermal diffusion, where the decay time of a diffusion-driven heat flux is considered to be significantly longer than the decay of spontaneous heat fluxes induced in the medium. The shift in perspective from the "slow" diffusion to the much faster stochastic heat fluxes allows the effect of the fluctuations at higher frequencies, such as those relevant to Raman spectroscopy, to be considered. The fundamental level of frequency noise induced in a wave propagating through a medium can then be predicted using knowledge of the behaviour of the stochastic heat fluxes through their temporal and spatial correlation in the medium. The model is given by Eq. (10) in [13]:

$$I(\Omega) = {A_0}^2 \left\{ \delta(\Omega) + 4\pi^2 \left\langle \delta n^2 \right\rangle \frac{L\ell}{{\lambda_0}^2} \frac{\ell^2}{\ell^2 + 2W^2} \gamma e^{-\gamma|\Omega|} \right\}, \tag{1}$$

Optics EXPRESS

where A_0 denotes the initial field amplitude, $\delta(\Omega)$ the spectrum of the initial field, $\langle \delta n^2 \rangle$ the variance of the refractive index due to thermal fluctuations, L the length of the mode, λ_0 the excitation wavelength, and W the mode width. The additional variables ℓ and γ describe the spatial and temporal correlations of fluctuations in the temperature field, respectively.

As shown in Eq. (1), the inverse square dependence of the Raman intensity on the radial frequency shift Ω as seen in Eq. (35) in [23] is replaced by an exponential dependence due to the explicit consideration of temporal correlations in the thermal field. In the case of low frequency shifts Ω, the thermal fluctuations in the mode are assumed to be governed primarily by diffusion in the guiding medium. This results in an approximation of the Raman spectrum that is proportional to the inverse square of the frequency shift Ω. However, when Ω becomes large, the period of the propagating wave becomes much shorter than the correlation time of the diffusion, i.e. the inducing field oscillates faster than diffusion can propagate the generated heat. Using this assumption, the diffusion can be considered as a steady state phenomenon rather than reactive to the propagating wave. In this setting, both the temporal and spatial correlations of the thermal field can be considered as strong influences. Pursuing this assumption leads to an approximation of the Raman spectrum that is exponentially dependent on Ω when Ω is large. The influence of the temporal correlation is accounted for through the introduction of the characteristic time γ of the correlations. Similarly, the spatial correlations are accounted for through a characteristic length ℓ which has a linear dependence on γ (see Eq. (9) in [13]). Using these variables, along with the length L and width W of the guided mode, the thermal field δT can be modeled and connected to the optical field through the expected variance of the refractive index change $\langle \delta n^2 \rangle$. As we intend to filter out the pump wavelength ($\Omega = 0$), we see that the dirac-delta term $\delta(\Omega)$ becomes zero for all relevant Ω, allowing us to remove it from the expression. By also allowing the exciting field intensity A_0^2 to be an input variable, the model in Eq. (1) can be rewritten as:

$$I(\Omega)/I(0) = \left(4\pi^2 \frac{L}{{\lambda_0}^2}\right)\left(\langle \delta n^2 \rangle \frac{\ell^3}{\ell^2 + 2W^2}\gamma\right) e^{-\gamma|\Omega|}. \tag{2}$$

As the waveguide length L and pump wavelength λ_0 are known *a priori*, these can be collapsed into a known quantity $\alpha = 4\pi^2 L/\lambda_0^2$ for convenience. Of the four remaining variables, all but the mode width W are strongly dependent on the characteristics of the temperature field T and its fluctuations δT. Given the known dependencies on the material parameters listed in Table 1, it is possible to estimate these variables given the findings of N. Le Thomas [13]. However, as the exact characteristics of the thermal field fluctuations δT are not known, the variables dependent on it are collapsed into a fit parameter β. For convenience, the unknown characteristic time γ is replaced by a fit parameter ϵ such that the model can be expressed as a function of the wavenumber shift $\tilde{\nu}$ in cm^{-1} rather than the radial frequency shift Ω in rad/s. The model in Eq. (2) is thus rewritten as:

$$I(\tilde{\nu})/I(0) = \alpha\beta e^{-\epsilon\tilde{\nu}}, \tag{3}$$

with

$$\beta = \langle \delta n^2 \rangle \frac{\ell^3}{\ell^2 + 2W^2}\gamma \quad \text{and} \quad \epsilon = 2 \cdot 10^2 \pi c \gamma.$$

The model can then be fitted to the experimental results using the known parameter α and the fit parameters β and ϵ. Together with the model for Raman scattering in waveguides, N. Le Thomas *et al.* presented measurements of the Raman background in Si_3N_4-waveguides. The results showed a peak of −91.9 dB (normalised to 1 cm length, see Fig. 4 in [13]) with a decay given by a the characteristic time $\gamma = 13$ fs. Considering the material parameters of Si_3N_4 [13], as shown in Table 1, we see a number of differences between the parameters of Si_3N_4 and those of SiO_2. In the following, these differences will be described to find expectation values for the Raman background of UV-written silica waveguides.

Optics EXPRESS

Table 1. Material parameters for SiO_2 [25–29] and Si_3N_4 [13]: thermal conductivity κ_0, density ρ, heat capacity C_V, thermal expansion coefficient α_L, thermo-optic coefficient $\partial n/\partial T$, and refractive index $n_\phi hi$. (*Mean values)

	$\kappa_0\left(\frac{W}{mK}\right)$	$\rho\left(\frac{g}{cm^3}\right)$	$C_V\left(\frac{J}{gK}\right)$	$\alpha_L\left(K^{-1}\right)$	$\frac{\partial n}{\partial T}\left(K^{-1}\right)$	n_ϕ
SiO_2	1.9	2.2*	0.73	$5.4{\cdot}10^{-7}$	$1.1{\cdot}10^{-5}$	1.5
Si_3N_4	25	2.5	0.17	$3.9{\cdot}10^{-6}$	$5{\cdot}10^{-5}$	1.8
Ratio	$9.3{\cdot}10^{-2}$	0.88	4.4	0.14	0.2	0.80

The thermal conductivity of SiO_2 is substantially lower than for Si_3N_4, which is expected to shorten the characteristic length ℓ of SiO_2 due to its square root dependence on thermal conductivity [13]. Factoring in the slightly lower density ρ and significantly higher heat capacity C_V we estimate the thermal diffusivity of SiO_2 to be 55 times weaker than for Si_3N_4, implying a correlation length ℓ that is 86.5% shorter for the same relaxation time τ. Considering that both the thermal expansion coefficient α and thermo-optic coefficient $\frac{\partial n}{\partial T}$ of SiO_2 is almost an order of magnitude smaller than for Si_3N_4 as well as the lower refractive index n_ϕ, it is implied that the variance of the refractive index $<\delta n^2>$ is approximately 96% smaller for SiO_2 than for Si_3N_4.

These material properties of SiO_2 compared to Si_3N_4 implies that the induced background of SiO_2 is significantly lower than for Si_3N_4. Factoring in the significantly larger mode field diameter of the UV-SiO_2 waveguides [15], it is implied that the Raman intensity of the UV-SiO_2 waveguide which is 18.3 dB weaker than for Si_3N_4, further implying a peak Raman intensity of -113.5 dB at $\tilde{\nu} = 530\text{cm}^{-1}$. This is in agreement with the measurements made by N. Le Thomas *et al.* on the PM-fiber (see Fig. S1 in [24]), as its peak corresponds to ≈ -114 dB when scaled to the same length.

As the materials of a PM-fiber and the silica waveguides are very similar, it is expected that the Raman spectrum of the waveguide will follow the same pattern as the PM-fiber as shown in Fig. S1 in the supplementary work of N. Le Thomas *et al.* [24]. However, given that the PM-fiber has a stated mode field diameter (MFD) of 5.3 μm while the UV-written waveguide has a MFD of approximately 5 μm, and since it is assumed that $\ell \ll W$, it is expected that the $\alpha\beta$-scalar for the waveguide is approximately 12% smaller than for the PM-fiber and its characteristic time $\approx$ 6% longer, assuming all other factors are equal. From this it is expected that the fiber and the waveguides will perform approximately the same, with the waveguide having a peak that is implied to be 1.1 dB lower than the fiber.

Regarding wavelength dependence, the equation derived by N. Le Thomas shows an inverse square dependence on the pump wavelength which implies a 1.5 dB stronger Raman background induced by a 660 nm pump compared to a 785 nm pump.

3. Experimental design

3.1. Waveguide fabrication

As depicted in Fig. 1, the fabrication begins by growing an oxide layer on a 150 mm diameter, 1 mm thick silicon wafer using a wet thermal process, common in microfabrication. On all the devices presented in this work, this layer is at least 15 μm thick and is high-purity silica. This layer will eventually serve as the bottom cladding of the waveguide, isolating the guided mode from the silicon underneath. Flame hydrolysis deposition (FHD) is then used to deposit a thick layer of doped silicate glass to serve as the core layer of the waveguide. The core is doped with germanium to provide UV photosensitivity, which is further enhanced by adding boron. Germanium and boron co-doping is conventionally used in UV-photosensitive optical fibers and planar waveguides. However, in this work, we have also investigated germanium and phosphorous co-doping. Lastly, an optional layer is added on top of the core to serve as a top

Optics EXPRESS

cladding to the waveguide. This layer is doped with phosphorus and boron to match the refractive index of the bottom silica-clad layer. The wafer is then diced into dies of approximately 10 x 20 mm and hydrogen loaded at 120 bar for several days to further enhance the UV photosensitivity [15,30]. The chip is then irradiated with two focused beams from a frequency-doubled argon-ion laser operating at a wavelength of 244 nm, see Fig. 1. UV exposure induces a localised increase in refractive index, thereby forming a channel waveguide.

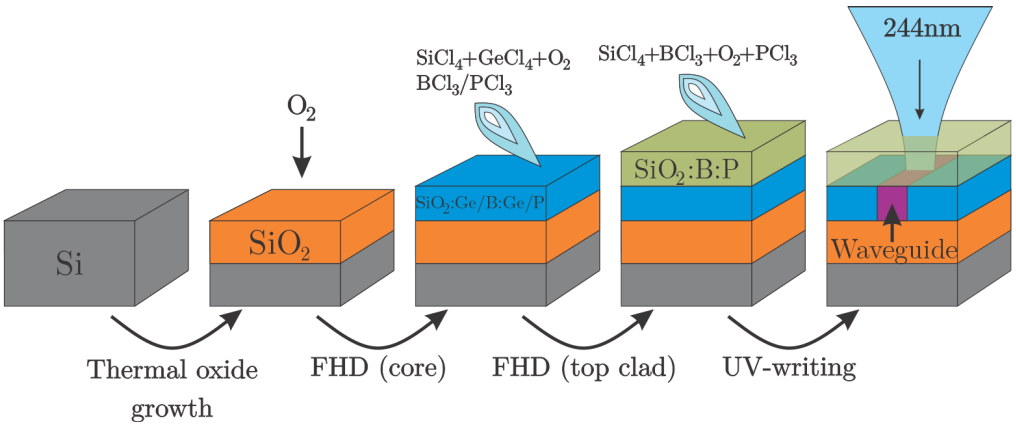


Fig. 1. Fabrication process for UV-SiO2 waveguides

Due to the weakness of the photorefractive effect, the Δn of the waveguide is low (typically $5 \cdot 10^{-3}$ [31]), resulting in a large mode and low NA, similar to standard optical fibers. The low Δn, along with the intrinsic smoothness of an etch-free waveguide, contribute to low propagation losses in the guided mode. The low loss, combined with the low NA ($\approx$ 0.1 [15]) allows these waveguides to project light with high power and low divergence into free space, but with lower intensity than a waveguide with high Δn. The low NA gives low loss across a gap in the waveguide, e.g. a hole or a trench, enabling several gaps in series for analysis of several samples along the waveguide. The large mode of these singlemode waveguides (MFD $\approx$ 5 µm) contributes to a low background in itself, as described in sec. 2.

3.2. Samples

Five waveguide chips are considered in this work to give a limited study of the sensitivity to waveguide design, all waveguides are of length $\approx$ 20 mm and are single mode for 660/785 nm wavelengths:

- Chip A: $\approx$ 5 µm **Ge+B** doped core layer with 17 µm B+P doped **top cladding**, MFD $\approx$ 5 µm @ 780 nm
- Chip B: $\approx$ 3 µm **Ge+B** doped core layer with $\approx$ 15 µm B+P doped **top cladding**, MFD $\approx$ 5 µm
- Chip C: $\approx$ 3 µm **Ge+P** doped core layer with **no top cladding**, MFD $\approx$ 5 µm
- Chip D_1 & D_2: $\approx$ 3 µm **Ge+B** doped core layer with **no top cladding**, MFD $\approx$ 5 µm

3.3. Experimental setup

The experimental setup, depicted in Fig. 2, consists of three main sections (a-c) and one auxiliary section (d).

Optics EXPRESS

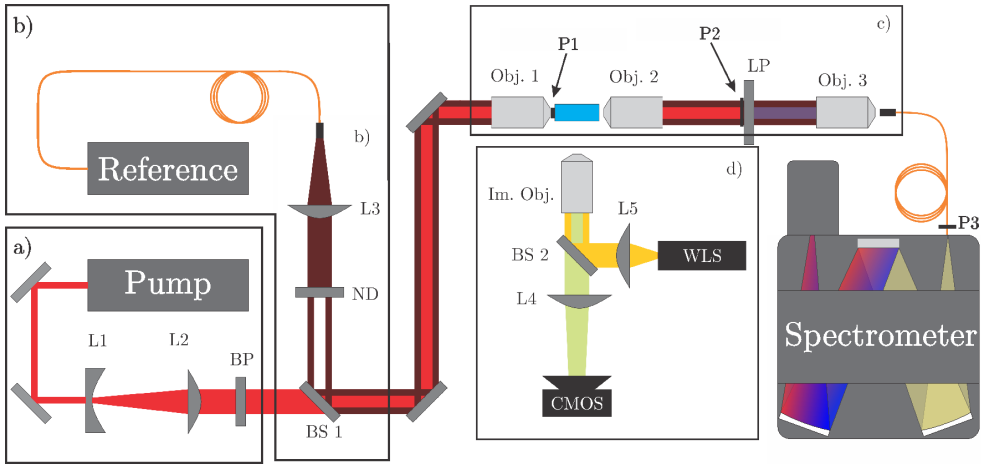


Fig. 2. Sketch of experimental setup for acquiring power Raman spectra of a sample waveguide. The path of the pump beam (red) moves through the entire setup untill removal by a long pass filter (LP) after out-coupling from the waveguide. A longer wavelength reference laser (brown) is joined to the pump path by a beam-splitte (BS 1) such that it bypasses the long pass filter and can be recorded by the spectrometer for calibration.

Section a

Section a is the origin of the pump beam, containing a high-power laser along with beam conditioning optics. To achieve best coupling to the waveguide, the output beam of the pump laser is expanded and collimated by a Galilean beam expander (L1 and L2) such that it yields a plane wave field with gaussian profile and width that is compatible with the back aperture of the in-coupling objective. The beam is also passed through a narrow band-pass filter (BP) such that the side-bands of the laser are suppressed, with special attention to the longer wavelengths.

Section b

Section b is the origin of the reference beam, accepting a fiber coupled laser source of wavelength longer than the pump beam and merging it to a common path with the pump beam. The fiber output is collimated by an appropriate lens (L3) such that it yields a gaussian plane wave that is coupled to a common path with the pump beam by a 90:10 beam splitter (BS 1).

Section c

Section c is the central part of the setup, containing the waveguide as well as the in-coupling (obj. 1) and out-coupling (obj.2) objectives. The output of the waveguide is, after collimation by the out-coupling objective, passed through a long-pass filter (LP) that removes the transmitted pump beam, leaving the Raman scattering and the reference beam. The filtered output is then coupled to a fiber by a final objective (obj. 3) and passed to the spectrometer.

Section d

Section d is a microscope for imaging the surface of the chip and for assisting in coupling to and from the waveguide, thus is an auxiliary section of the setup and does not contribute to the Raman measurements. This tower consists of a white light source (WLS) focused on the back focal plane of the imaging objective (Im. Obj.) by a lens (L5) relayed via a pellicle beam splitter (BS 2). Images are obtained by focusing the backscatter by a tube lens (L4) onto a camera (CMOS) for acquisition.

Optics EXPRESS

Components

- Pump: 660 nm DPSS (HüBNER, Cobolt 05-01) or 785 nm DPSS (CrystaLaser, DL785-100)
- Reference: 686 nm diode laser (ThorLabs, LP685-SF15) or 826 nm diode laser (ThorLabs, LPS-830-FC)
- Spectrometer: Multi-grating spectrometer (300/g, 600/g & 1200/g), focal length 320 mm (Teledyne Princeton, IsoPlane SCT320) with deep depletion CCD (Teledyne Princeton, BLAZE 400BR)
- Obj. 1: 10x objective (Olympus PLN10X, 10x 0.25NA) mounted on three axis stage with closed loop piezo (ThorLabs, MAX331D/M)
- Obj. 2: 10x objective (Olympus PLN10X, 10x 0.25NA) mounted on three axis stage (ThorLabs, MAX313D/M)
- Obj. 3: 10x objective (Olympus PLN10X, 10x 0.25NA) mounted on three axis stage (ThorLabs, 313D/M)
- BP: 660 ± 13 nm band-pass for 660 nm pump (Semrock, BrightLine FF01-660/13-25), 785 ± 3 nm for 785 nm pump (Edmund optics, 64-257)
- LP: 664 nm ultrasteep long-pass for 660 nm pump (Semrock, RazorEdge LP02-664RU-25) or 785 nm ultrasteep long-pass for 785 nm (Semrock, RazorEdge LP92-785RE-25)

3.4. Power calibration

Determining the spectrum of the Raman scattering induced in the waveguide is a relatively trivial task, requiring only subtraction of the background and correcting for the spectral sensitivity of the setup. However, in order to compare different waveguides and infer the intensity of the background in a Raman-on-chip device built from those waveguides requires a more thorough calibration such that the spectra can be expressed in absolute intensity rather than arbitrary units. This is why the reference laser (see Fig. 2) is necessary, providing a power reference that bypasses the long-pass filter such that it can be used as an intermediate bridge to compare the Raman spectra with the pump intensity. This requires a set of common measurement points (P1 and P2 in Fig. 2) where the pump and reference beams can be compared and a common point where the reference beam and the Raman scattering can be compared (P3 in Fig. 2). The goal of this is to obtain a calibration spectrum $C(\tilde{\nu})$ such that the intensity spectrum $I(\tilde{\nu})$ can be expressed from the measured spectrum $S(\tilde{\nu})$ as:

$$I(\tilde{\nu}) = C(\tilde{\nu})\, S(\tilde{\nu})\, (mW). \tag{4}$$

In order to express the measured spectrum as a calibrated spectrum, the measured spectrum must first be corrected for spectral background and sensitivity. As the background spectrum $S_{BG}(\tilde{\nu})$ introduces a constant bias to the measured spectrum $S'(\tilde{\nu})$, this must first be subtracted. The spectral sensitivity $A(\tilde{\nu})$ of the setup must also be accounted for such that the spectrum is not distorted. This is done by measuring a known source (Teledyne Princeton Halogen calibration lamp) and determining the relative response of the setup. The spectral sensitivity $A(\tilde{\nu})$ is then determined from the curve such that it scales to 1 at the reference laser wavelength

Optics EXPRESS

$(A\left(\lambda = \lambda_{Ref}\right) = 1)$. The true spectrum can then be approximated by the unbiased and distortion-free estimate spectrum:

$$\hat{S}\left(\tilde{\nu}\right) = A^{-1}\left(\tilde{\nu}\right)\left[S'\left(\tilde{\nu}\right) - S_{BG}\left(\tilde{\nu}\right)\right], \tag{5}$$

allowing Eq. (4) to be expressed as:

$$I\left(\tilde{\nu}\right) = c_1 A^{-1}\left(\tilde{\nu}\right) \cdot \left(S'\left(\tilde{\nu}\right) - S_{BG}\left(\tilde{\nu}\right)\right), \tag{6}$$

where $C\left(\tilde{\nu}\right)$ is replaced by $c_1 A^{-1}\left(\tilde{\nu}\right)$.

With the spectral elements of $C\left(\tilde{\nu}\right)$ being separated into the sensitivity curve $A^{-1}\left(\tilde{\nu}\right)$, the remaining calibration coefficient c_1 can be obtained by measuring the intensity of the pump and reference laser at in-coupling to the waveguide (P1 in Fig. 2) and at the fiber output (P3 in Fig. 2). This is done by first determining the scaling between mW and CCD counts through the coefficient c_0 using the measured intensity of the reference laser at the spectrometer (P3 in Fig. 2) $P3_{Ref}$ and a measured spectrum of the reference laser $S_{Ref}\left(\tilde{\nu}\right)$ along with the known attenuation factor A_{ND} of the neutral density filter (ND in Fig. 2). The coefficient c_0 can thus be expressed as:

$$c_0 = \frac{P3_{Ref}}{A_{ND}\sum_{n=0]}^{N}\hat{S}\left[\tilde{\nu}_n\right]},$$

where $\hat{S}\left[\tilde{\nu}_n\right]$ is the discrete spectrum of the reference laser source.

Finally, since the setup will have a difference in coupling and propagation losses for the pump and reference beams, this must also be taken into account. Using measurement points P1 and P3, the pump transmission can be expressed as:

$$T_{pump} = \frac{P3_{pump}}{P1_{pump}},$$

and the reference transmission as:

$$T_{Ref} = \frac{P3_{Ref}}{P1_{Ref}},$$

such that the calibration coefficient c_1 can be expressed as:

$$c_1 = c_0 \cdot \frac{T_{Ref}}{T_{Pump}}. \tag{7}$$

We can then substitute Eq. (7) into Eq. (6) to express the calibrated spectrum in Eq. (4) using measurable factors:

$$I\left(\tilde{\nu}\right) = \frac{T_{Ref}}{T_{Pump}}\frac{P3_{Ref}}{P1_{Pump}A_{ND}\sum_{n=0]}^{N}\hat{S}\left[\tilde{\nu}_n\right]} \cdot A^{-1}\left(\tilde{\nu}\right)\left(\hat{S}\left(\tilde{\nu}\right) - S_{BG}\left(\tilde{\nu}\right)\right). \tag{8}$$

3.5. Composite spectra

To fully take advantage of the spectrometer's capabilities and the fact that the fiber-coupled input (ThorLabs, SM-830) gives an effective slit width of approx. 5 µm, we choose to use the finest grating available (1200g/mm) to achieve a dispersion of 2.30 nm/mm at the focal plane. With a CCD pixel size of 20 µm, this yields a per pixel resolution of 0.05 nm but limits the spectral range of each acquisition to 52.3 nm. Therefore, the entire range of the Raman scattering (240 – 365 nm) cannot be captured in a single acquisition without compromising resolution. A one-shot acquisition also demands the dynamic range of the spectrum is within the dynamic range of the CCD (48.2 dB) and above the noise floor. Given the expected exponential decay of the signal with increasing wavenumber shift, as discussed in Sec. 2 a uniform spectral sensitivity risks

Optics EXPRESS

either over-saturating the CCD at low wavenumber shift, where the signal is strong, or losing the signal to noise at high wavenumber shift, where the signal is weak.

To solve this, we propose dividing the spectral range into segments that can be individually measured and later be merged in post-processing to obtain a composite spectrum covering an arbitrary spectral range. This allows us to acquire the entire spectral range without sacrificing the resolution afforded by the fine grating. This also allows us to dynamically select the exposure time and number of repeat exposures for each segment separately such that the SNR and dynamic range usage can be normalized for each segment separately.

This method is implemented through an automated script (Python 3.8.10) that partitions the desired range into a set of overlapping segments, each with their own exposure time t_e and number of averaged acquisitions n_{avg}. The acquisition parameters t_e and n_{avg} are then estimated using an initial guess for the exponential decay of the spectra such that the expected signal fills $\approx 10\%$ of the dynamic range of the CCD. A test acquisition of three spectra per segment is then made such that the parameters SNR and dynamic range usage can be estimated. This is achieved by a rough separation of the signal and CCD noise by low-pass filtering the measurement, using the low-frequency elements as a signal estimator and the high-frequency elements as a noise estimator. The signal estimator is then used to tune the dynamic range usage through the exposure time t_e while the noise estimator is used to tune the SNR to an acceptable level ($\geq$ 10dB) through increasing the number of averaged spectra n_{avg} in that segment. Using the determined parameters, the spectra of each segment is acquired and cleaned for background and cosmic rays. Using a least squares fit of their overlap, the segments are adjusted for their varying sensitivity and level of dark signal such that they are brought to a uniform scale with the first segment and can then be merged to form the composite spectrum.

4. Results

4.1. Pump wavelength: 600 nm vs. 785 nm

As previously mentioned in Sec. 1, one of the objectives of this work is to evaluate the sensitivity of the Raman scattering of the UV-SiO_2 waveguides to the wavelength of the pump lasers. To this end, the experiment is repeated with two sets of pump and reference lasers, first using a 660 nm pump complemented by a 686 nm reference and then using a 785 nm pump complemented by a 826 nm reference. These two pump wavelengths were chosen because of the availability of high-power lasers with high spectral purity and that both wavelengths are commonly used in Raman spectroscopy.

As described by N. Le Thomas *et al.* [13] and shown in the α component in Eq. (3), the intensity of the Raman scattering in the guided mode is expected to have an inverse square dependence on the pump wavelength λ_0. From this, it is expected that using $\lambda_0 = 660$ nm will induce Raman scattering approximately 1.5 dB stronger than using $\lambda_0 = 785$ nm, indicating that 785 nm may be favourable for a Raman-on-chip device. The use of 785 nm may also help reduce undesirable fluorescence in the waveguide and/or analyte compared to 660 nm, thus separating the Raman spectrum from the flourescence spectrum. However, the signal from a particle intersecting the beam path is expected to be proportional to λ^{-4}, implying a 3 dB increase in signal when using 660 nm instead of 785 nm as the pump wavelength, potentially compensating for the increased background in the waveguide. Another benefit of using a 660 nm pump is that it allows a longer range of wavenumber shift to be measured using a high-quality silicon-based CCD.

This enables us to acquire measurements up to 5 400cm^{-1} without exceeding the useful range of the spectrometer camera ($\lambda \leq 1025$ nm) while using 785 nm only allows for measurements up to 3 000 cm^{-1}. A pump of 660 nm thus allows for measurement of features in a wavenumber range where the background from the waveguide is expected to be negligible. One significant challenge with using shorter pump wavelength is the increased potential to induce fluorescence,

Optics EXPRESS

both from the analyte and the waveguide material, which would mask the Raman signal. The measurements are obtained from two waveguide chips, chip A for 785 nm and chip B for 660 nm. For reference, the same fiber as used by N. Le Thomas *et al.* [24] is measured with both pump wavelengths, as is a 7 μm polystyrene bead to demonstrate the achievable SNR compared to a particle. As shown in Fig. 3, both the waveguides and the PM-fiber exhibit stronger Raman scattering when excited with 660 nm, notably so with the appearance of a flattening of the spectra as the shift exceeds ≈ 1500 cm^{-1}. Given that this flattening forms a wide bulge and is only present when excited with 660 nm, this is more consistent with fluorescence than Raman. Aside from this, we can see that the levels and features of the two pairs remain almost identical prior to the flattening and that the features of the PS-bead are almost identical.

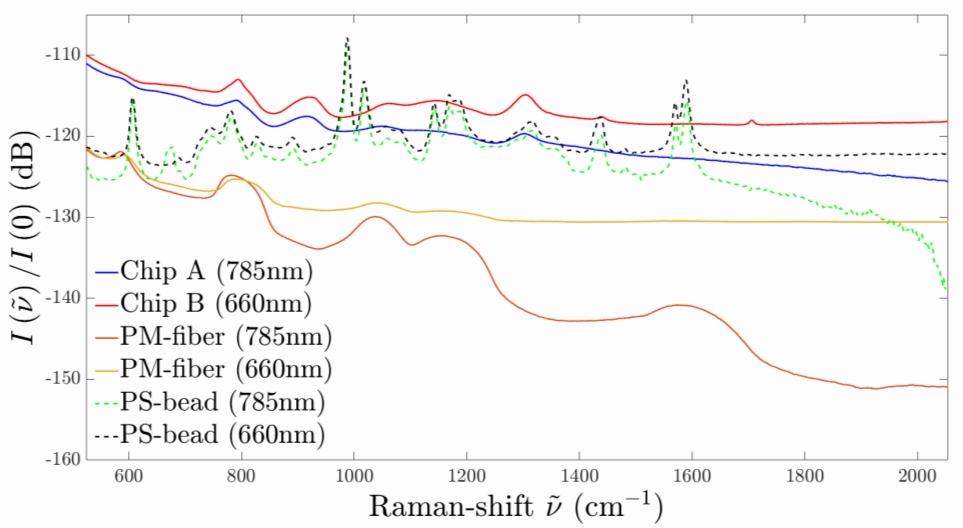


Fig. 3. Raman background spectra for waveguide chips A and B (with top cladding) compared to a PM-fiber with a single 7 μm polystyrene bead for reference, measured for two separate pump wavelengths (660 nm and 785 nm).

In terms of features, we see that both waveguides exhibit a peak at ≈ 920 cm^{-1} that is absent in the spectrum of the fiber, this can be readily assigned to the stretching mode of Ge-O-Si [32] due to the known high concentration of germanium in the waveguide cores. We can also observe that both waveguides produce peaks at ≈ 1310 cm^{-1} that corresponds well with B-O^- [33] as is also expected due to the boron-doping of the core. The remaining features at ≈ 1440 cm^{-1} and ≈ 1710 cm^{-1} may also be due to B-O-B and $[BO_3]^+$ modes, but due to the ambiguity of features near those shifts, we are hesitant to make the assignment. One other noteworthy observation in the Raman spectra is the significantly weaker feature at ≈ 580 cm^{-1} in the waveguides compared to the fiber. This is commonly assigned as a defect mode of Si-O-Si [34], which diminishes with increasing dopant concentration as observed here.

Because of the higher background produced by the waveguide when excited using 660 nm compared to 785 nm and the fact that the PS-bead, serving as an analogue for future particles, shows only a weak increase in feature intensity when excited using 660 nm, it is concluded that 785 nm is the preferred pump wavelength for this type of waveguide device.

4.2. *Dependence on doping and cladding*

Another objective is to determine the sensitivity of the Raman scattering on the doping of the core layer and the use of a top cladding. In this section, three additional chips (C,D_1 and D_2) are

Optics EXPRESS

measured using a 785 nm pump. For reference, both the fully clad NIR-waveguide (chip A) and the PM-fiber shown in Fig. 3 are included. As mentioned in Sec.3.2, both chips C and D are manufactured without top cladding while chips A and B are manufactured with a top cladding (>10 μm B+P doped SiO_2). Chip C is doped with phosphorous instead of boron while chips D are doped similarly to chips A and B. As shown in Fig. 4, the results from chip A and chips D_1 and D_2 match well, as expected, and it can be seen that the measurements obtained from two waveguides from D_1 and two from D_2 group well, supporting consistency of the power calibration. We also see that the phosphorous doped chip C performs significantly poorer than the other chips, producing a noticeably flatter Raman spectrum with a higher baseline than the other samples. The measured transmission through chip C was up to 5.6dB lower than either chip D and the mode at the output was poorly defined with significant slab guiding in the core layer relative to the guiding in the UV-written waveguide. The high degree of slab guiding indicates poor lateral confinement, likely due a low Δn being induced by the photorefractive effect without the presence of boron doping. Several of the chips, notably chip D_1, also displayed a significant variance in transmission between waveguides, varying as much as 5.4dB. The variance in transmission among the topless waveguides suggests chipping or defects at the facet because of a lack of the protective top cladding. The low and varying transmission for some of the waveguides may thus be due to poor in-coupling and surface defects, rather than absorption or scattering in the core itself.

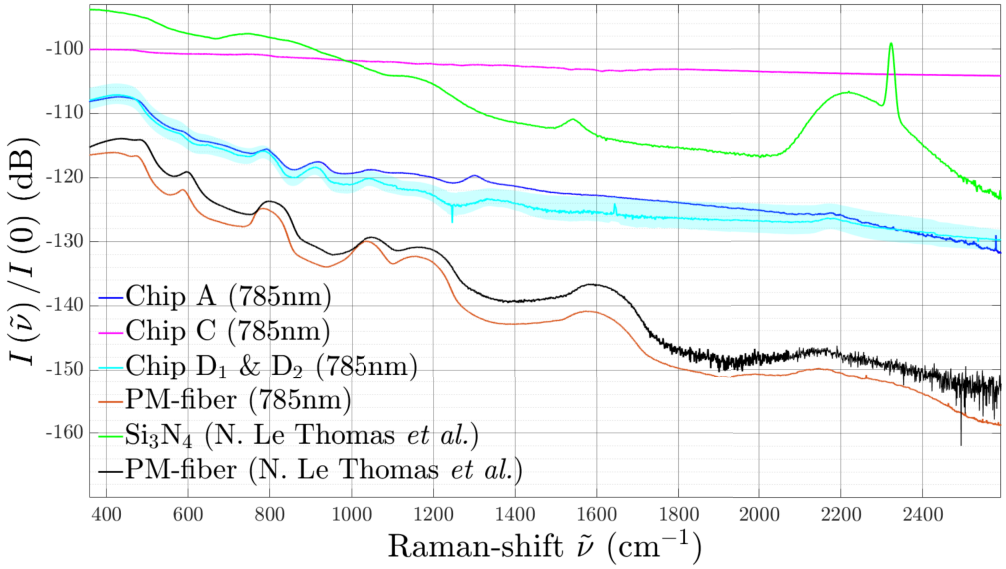


Fig. 4. Measurement of all chips with PM-fiber for comparison. Previously reported measurements of Si_3N_4 and similar PM-fiber are also included for reference.

4.3. *Fitting the spectra to the model*

In this section, the measured spectra are fitted to the model described in eq. (3) and the parameters of the fit are compared between the measured waveguides and previously reported measurements. The fits are shown in Fig. 5. The general level of the measurements follow the model in Eq. (3) well, with the majority of deviance being due to specific features in the Raman spectra. We also see from the parameters and peak intensity in Table 2 that our measurements of the PM-fiber agree with those made by N. Le Thomas *et al.* [24] in both profile and intensity. From the listed intensities we see that all of the Ge+B-doped SiO_2 (Chips A, B, D_1 and D_2) have a negligible

Optics EXPRESS

difference in peak amplitude, while the P+Ge-doped SiO_2 (Chip C) has a peak 6 dB higher than the rest, emphasizing the negative effect of phosphorous in the core. Lastly, we see that the Raman scattering in our UV-SiO_2 waveguides is more than 12 dB weaker than what is reported for Si_3N_4, reinforcing their advantage for Raman-applications.

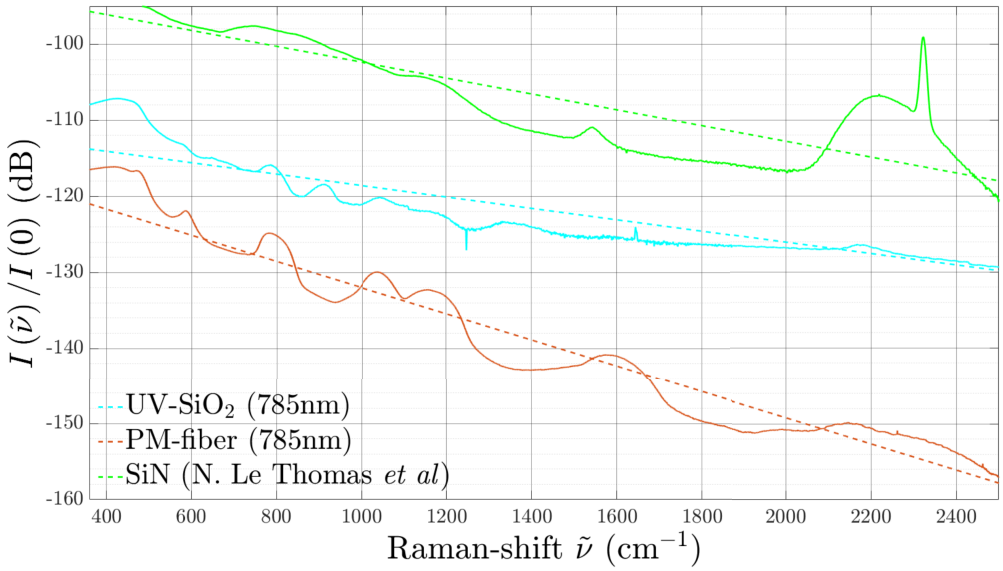


Fig. 5. Fit of the model in eq. (3) to measurements of chips A, B, C, D_1 and D_2 with a similar PM-fiber measured in this work and in the work of N. Le Thomas *et al.* as a reference. The fit of reported measurements of a Si_3N_4 waveguide is also added for comparison with the UV-written SiO_2 waveguides.

Table 2. Table of fit parameters for spectra in Fig. 5 to Eq. (3) as described in Sec. 2. *All measurements are scaled to emulate a length of 1cm

Sample\Parameter	α^* $\left(m^{-1}\right)$	β (ms)	ϵ (m)	I_{max} (dB)	RMSe (dB)
Chip A	6.4e">+27	2.3e-32	2.2e-5	−107	2.2
Chip B (660nm)	9.1e">+27	5.2e-33	5.7e-6	−106	3.8
Chip C	6.4e">+27	5.8e-32	3.7e-6	−100	0.3
Chip D	6.4e">+27	1.5e-32	1.9e-5	−107	2.4
PM780-fiber	6.4e">+27	5.2e-33	4.0e-5	−116	2.6
PM780-fiber [24]	6.4e">+27	1.0e-32	4.1e-5	−114	2.9
Si_3N_4 [13]	6.4e">+27	1.0e-30	2.4e-5	−94	4.3

5. Conclusion

The Raman background of UV-written silica waveguides has been measured and compared to other platforms, notably Si_3N_4 waveguides and (silica) optical fibers. To obtain results in absolute terms, the acquired spectra were calibrated to the input intensity. This was achieved using a separate laser source, acting as a reference and coupled into a common path with the pump laser, such that both the Raman spectra and the reference laser could be measured with the same configuration. Furthermore, by combining the spectrum of a reference source measured by

Optics EXPRESS

a spectrometer with the power of the pump and reference source measured by a photodiode power meter, the acquired Raman spectra were calibrated to absolute terms. To take full advantage of the setup, a high-resolution grating (1200 lines/mm) is used and the spectra are acquired in adjoining segments, allowing the full spectral and dynamic range of the spectrometer to be used. The acquisition time and number of repetitions was set independently for each section, to exploit the full range of the CCD spectrometer and to tailor the sensitivity to the expected intensity for the section. This allowed a high signal-to-noise ration for a very large dynamic range (from −100 to −160 dB relative to the pump laser).

The background of the UV-written SiO_2 waveguides was measured for two excitation wavelengths, 660 nm and 785 nm. Waveguides written into five chips were characterized, and for reference two optical fibers and a 7 μm polystyrene bead were also measured. This revealed a Raman intensity of less than −107.4 dB in the biochemical fingerprint region for a waveguide excited by 785 nm, and less than −106.5 dB when excited by 660 nm (normalised to 1 cm length). The difference increased for increasing wavenumber shifts, leading to the conclusion that 785 nm is better suited than 660 nm when using these waveguides. This conclusion depends on signal-to-noise ratio, which depends on the analyte, and it was shown that it holds for a polystyrene bead as it gave the same Raman signal for both wavelengths. The largest peak in the Raman spectrum of a 7 μm PS-bead was 10.4 dB higher than the waveguide background. This shows that a good signal-to-noise ratio can be obtained for microparticles with all the background from a 1 cm long waveguide collected as noise. The signal scales with the intensity and the interaction volume, with the first depending on the waveguide structure for illuminating the particle (e.g. hole, trench or taper) and the second on the diameter of the particle. The noise depends on how much of the waveguide background is collected (e.g. by a microscope objective or a collection waveguide). The achievable signal-to-noise ratio for nanoparticles thus depends on the interaction structure and the collection method. We will investigate this in future work.

The background induced in the best waveguide was 8.7 − 10.3 dB higher than for optical fibers and approximately 15 dB smaller than for Si_3N_4 waveguides [13]. UV-written waveguides thus present a very promising alternative for on-chip Raman spectroscopy, but there is still room for improvement when comparing with optical fibers.

A limited study of the impact of doping was made, with phosphorous doping giving significantly higher background than boron, with −100.0 dB and −106.5 dB, respectively, for a wavenumber shift of 410 cm^{-1} and 436 cm^{-1}. In addition, the phosphorous doping resulted in a significant flattening of the spectrum and poor waveguide confinement. This may have influenced the result. The presence or not of a top-cladding did not influence the background significantly (for boron doped samples). The background of four top-clad waveguides from two chips deviated by less than 5.7 dB over the entire fingerprint region, showing good repeatability considering the mean level of −121 dB.

This work has only considered the noise related to Raman background of the waveguides, and the logical next step will be to modify the waveguides to obtain signals from an analyte. Several procedures will be tested, notably etching trenches across the waveguides. Approaches for incorporating more complex structures, such as tapers and/or nanoantennas will also be explored.

Funding. The publication charges for this article have been funded by a grant from the publication fund of UiT The Arctic University of Norway. Norges Forskningsråd (302333).

Acknowledgments. The authors thank Dr. Cees Otto for useful discussions and suggestions. The authors would also like to thank Dr. Nicolas Le Thomas for providing the original data from his work [13,24] on Si_3N_4 waveguides and a SiO_2 PM-fiber.

Disclosures. The authors declare no conflic of interest.

Data availability. Acquired data is available in open source database (Dataverse UiT, DOI:10.18710/R6JHV7). The spectra of PM-fibers acquired by N. Le Thomas *et al.* and displayed in Figs. 4 and 5 are detailed in Ref. [13] and [24].

Optics EXPRESS

References

1. N. F. Tyndall, T. H. Stievater, D. A. Kozak, K. Koo, R. A. McGill, M. W. Pruessner, W. S. Rabinovich, and S. A. Holmstrom, "Waveguide-enhanced raman spectroscopy of trace chemical warfare agent simulants," Opt. Lett. **43**(19), 4803–4806 (2018).
2. N. L. Thomas, Z. Liu, C. Lin, H. Zhao, and R. Baets, "Raman on-chip: current status and future tracks," in *Integrated Optics: Devices, Materials, and Technologies XXV*, vol. 11689 S. M. García-Blanco and P. Cheben, eds., International Society for Optics and Photonics (SPIE, 2021), p. 1168908.
3. D. M. Kita, J. Michon, and J. Hu, "A packaged, fiber-coupled waveguide-enhanced raman spectroscopic sensor," Opt. Express **28**(10), 14963–14972 (2020).
4. M. A. Ettabib, A. Marti, Z. Liu, B. M. Bowden, M. N. Zervas, P. N. Bartlett, and J. S. Wilkinson, "Waveguide enhanced raman spectroscopy for biosensing: A review," ACS Sens. **6**(6), 2025–2045 (2021).
5. H. Jayan, H. Pu, and D.-W. Sun, "Recent developments in raman spectral analysis of microbial single cells: Techniques and applications," Crit. Rev. Food Sci. Nutr. **62**(16), 4294–4308 (2022). PMID: 34251940.
6. G. Pezzotti, "Raman spectroscopy in cell biology and microbiology," J. Raman Spectrosc. **52**(12), 2348–2443 (2021).
7. E. S. L. J. Hong, S. B. Kim, and T. K. Lee, "Microbial phenomics linking the phenotype to function: The potential of raman spectroscopy," J. Microbiol. **59**(3), 249–258 (2021).
8. W. Lee, A. Nanou, L. Rikkert, F. A. W. Coumans, C. Otto, L. W. M. M. Terstappen, and H. L. Offerhaus, "Label-free prostate cancer detection by characterization of extracellular vesicles using raman spectroscopy," Anal. Chem. **90**(19), 11290–11296 (2018). PMID: 30157378.
9. M. Boerkamp, T. van Leest, J. Heldens, A. Leinse, M. Hoekman, R. Heideman, and J. Caro, "On-chip optical trapping and raman spectroscopy using a triplex dual-waveguide trap," Opt. Express **22**(25), 30528–30537 (2014).
10. J. Langer, D. J. de Aberasturi, and J. Aizpurua, *et al.*, "Present and future of surface-enhanced raman scattering," ACS Nano **14**(1), 28–117 (2020).
11. M. Fan, G. F. S. Andrade, and A. G. Brolo, "A review on recent advances in the applications of surface-enhanced raman scattering in analytical chemistry," Anal. Chim. Acta **1097**, 1–29 (2020).
12. C. Zong, M. Xu, L.-J. Xu, T. Wei, X. Ma, X.-S. Zheng, R. Hu, and B. Ren, "Surface-enhanced raman spectroscopy for bioanalysis: Reliability and challenges," Chem. Rev. **118**(10), 4946–4980 (2018).
13. N. Le Thomas, A. Dhakal, A. Raza, F. Peyskens, and R. Baets, "Impact of fundamental thermodynamic fluctuations on light propagating in photonic waveguides made of amorphous materials," Optica **5**(4), 328–336 (2018).
14. Q. S. Ahmed, P. C. Gow, C. Holmes, P. L. Mennea, J. W. Field, R. H. Bannerman, D. H. Smith, C. B. Gawith, P. G. Smith, and J. C. Gates, "Direct uv written waveguides and bragg gratings in doped planar silica using a 213 nm laser," Electron. Lett. **57**(8), 331–333 (2021).
15. G. Lepert, M. Trupke, E. A. Hinds, H. Rogers, J. C. Gates, and P. G. R. Smith, "Demonstration of uv-written waveguides, bragg gratings and cavities at 780 nm, and an original experimental measurement of group delay," Opt. Express **19**(25), 24933–24943 (2011).
16. P. C. Gow, R. H. S. Bannerman, P. L. Mennea, C. Holmes, J. C. Gates, and P. G. R. Smith, "Direct uv written integrated planar waveguides using a 213 nm laser," Opt. Express **27**(20), 29133–29138 (2019).
17. A. Raza, S. Clemmen, P. Wuytens, M. de Goede, A. S. K. Tong, N. Le Thomas, C. Liu, J. Suntivich, A. G. Skirtach, S. M. Garcia-Blanco, D. J. Blumenthal, J. S. Wilkinson, and R. Baets, "High index contrast photonic platforms for on-chip raman spectroscopy," Opt. Express **27**(16), 23067–23079 (2019).
18. D. A. Coucheron, Ø. I. Helle, J. S. Wilkinson, G. S. Murugan, C. Domínguez, H. Angelskår, and B. S. Ahluwalia, "Study of waveguide background at visible wavelengths for on-chip nanoscopy," Opt. Express **29**(13), 20735–20746 (2021).
19. M. Vlk, A. Datta, S. Alberti, A. Aksnes, G. S. Murugan, and J. Jagerska, "High-aspect-ratio free-standing membrane waveguides for mid-infrared nanophotonics," in *2021 Conference on Lasers and Electr-optics (CLEO)*, (IEEE, 2021), Conference on Lasers and Electro-Optics. Conference on Lasers and Electro-Optics (CLEO), ELECTR NETWORK, MAY 09-14, 2021.
20. D. Giordano, D. González-García, J. K. Russell, S. Raneri, D. Bersani, L. Fornasini, D. Di Genova, S. Ferrando, M. Kaliwoda, P. P. Lottici, M. Smit, and D. B. Dingwell, "A calibrated database of raman spectra for natural silicate glasses: implications for modelling melt physical properties," J. Raman Spectrosc. **51**(9), 1822–1838 (2020).
21. T. Geisler, L. Dohmen, C. Lenting, and M. B. K. Fritzsche, "Real-time in situ observations of reaction and transport phenomena during silicate glass corrosion by fluid-cell raman spectroscopy," Nat. Mater. **18**(4), 342–348 (2019).
22. S. Foster, "Low-frequency thermal noise in optical fiber cavities," Phys. Rev. A **86**(4), 043801 (2012).
23. S. Foster, A. Tikhomirov, and M. Milnes, "Fundamental thermal noise in distributed feedback fiber lasers," IEEE J. Quantum Electron. **43**(5), 378–384 (2007).
24. N. L. Thomas, A. Dhakal, A. Raza, F. Peyskens, and R. G. Baets, "Supplement 1.pdf," (2018).
25. J. M. Larkin and A. J. H. McGaughey, "Thermal conductivity accumulation in amorphous silica and amorphous silicon," Phys. Rev. B **89**(14), 144303 (2014).
26. S. Inaba, S. Oda, and K. Morinaga, "Heat capacity of oxide glasses measured by ac calorimetry," J. Non-Cryst. Solids **306**(1), 42–49 (2002).
27. H. Gao, Y. Jiang, Y. Cui, L. Zhang, J. Jia, and L. Jiang, "Investigation on the thermo-optic coefficient of silica fiber within a wide temperature range," J. Lightwave Technol. **36**(24), 5881–5886 (2018).

Optics EXPRESS

28. G. Pan, N. Yu, B. Meehan, T. W. Hawkins, J. Ballato, and P. D. Dragic, "Thermo-optic coefficient of b_2o_3 and geo_2 co-doped silica fibers," Opt. Mater. Express **10**(7), 1509–1521 (2020).
29. H. J. Lee, F. Abdullah, S. D. Emami, and A. Ismail, "Fiber modeling and simulation of effective refractive index for tapered fiber with finite element method," in *2016 IEEE 6th International Conference on Photonics (ICP)*, (2016), pp. 1–3.
30. H. L. Rogers, C. Holmes, J. C. Gates, and P. G. R. Smith, "Analysis of dispersion characteristics of planar waveguides via multi-order interrogation of integrated bragg gratings," IEEE Photonics J. **4**(2), 310–316 (2012).
31. R. M. Parker, J. C. Gates, M. C. Grossel, and P. G. Smith, "A temperature-insensitive bragg grating sensor—using orthogonal polarisation modes for in situ temperature compensation," Sens. Actuators, B **145**(1), 428–432 (2010).
32. J. Zhao, Z. Yang, C. Yu, J. Qiu, and Z. Song, "Influence of glass composition on photoluminescence from ge^{2+} or ag nano-cluster in germanate glasses for white light-emitting diodes," J. Am. Ceram. Soc. **102**(3), 1169–1179 (2019).
33. A. K. Yadav and P. Singh, "A review of the structures of oxide glasses by raman spectroscopy," RSC Adv. **5**(83), 67583–67609 (2015).
34. K. Sasan, A. Lange, T. D. Yee, N. Dudukovic, D. T. Nguyen, M. A. Johnson, O. D. Herrera, J. H. Yoo, A. M. Sawvel, M. E. Ellis, C. M. Mah, R. Ryerson, L. L. Wong, T. Suratwala, J. F. Destino, and R. Dylla-Spears, "Additive manufacturing of optical quality germania-silica glasses," ACS Appl. Mater. Interfaces **12**(5), 6736–6741 (2020).

Paper II: Evaluation of crystalline structure quality of Czochralski-silicon using near-infrared tomography

Published in Journal of Crystal Growth, April 2022. **Authors:** Mathias N. Jensen and Olav Gaute Hellesø

Contribution notes: Mathias N. Jensen concieved the original idea, performed all experimental work and data analysis. Olav Gaute Hellesø suggested and oversaw the work and oversaw the work. Mathias N. Jensen wrote the initial draft and Olav Gaute Hellesø finalized the manuscript for submission. Both authors contributed to revision of the manuscript before publication.

Journal of Crystal Growth 583 (2022) 126527

Contents lists available at ScienceDirect

Journal of Crystal Growth

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

Evaluation of crystalline structure quality of Czochralski-silicon using near-infrared tomography

Mathias N. Jensen, Olav Gaute Hellesø *

Department of Physics and Technology, UiT The Arctic University of Norway, Hansine Hansens veg 18, Tromsø 9019, Norway

ARTICLE INFO

Communicated by Thierry Duffar

Keywords:
A2. Czochralski method
A1. Tomography
A1. Crystal structure
A1. Defects

ABSTRACT

In this work, three silicon samples are subject to tomographic scans using a 1.6μm laser. The samples were prematurely terminated due to anomalies during the Czhochralski-process. They are taken as analogues of the in situ crystal, where one sample has known aberrant structure in its lowermost 45 mm. The results of the tomographic scans show a distinct difference in transmission profile between the material of known poor mono-crystalline structure and assumed good structure. Three different analysis tools are constructed and applied to quantify the quality of the structure from the results of the tomographic scans. The first two analysis tools are applied as correlation filters constructed from patterns resembling the indicative transmission profiles of high-quality structure, one pattern being an ideal square wave and the other being experimentally determined from the measurements. Both correlation filters yield clear differentiation of low- vs. high-quality material. The final analysis tool is a deep convolutional neural network (deep CNN) evolved from a predetermined architecture configuration using a genetic algorithm. The trained CNN is shown to differentiate the usable high-quality material from the unusable material with a 98.7% accuracy on a testing set of 76 profiles and successfully assigns quality factors to the material that are in good agreement with the correlation filters and previous observations.

1. Introduction

In the production of silicon crystals, a mono-crystalline ingot, or boule, is ”pulled” from a crucible of molten material. This process relies on strict control of the thermal conditions surrounding the growth interface as even small deviations can cause the formation of anomalies, such as crystal dislocations or, in extreme cases, a complete loss of mono-crystalline structure[1]. This results in the material forming intersecting crystal lattices of various sizes and orientation, producing material with non-homogeneous macroscopic properties. As the goal of this type of growth process is to obtain high quality material with a homogeneous structure throughout, such deviations are detrimental for the outcome. For an intact mono-crystalline structure, four ridges parallel to the growth axis can be observed along the crystal sidewall. These ridges, or nodes, are caused by the cubic nature of the silicon crystal and can be used as indicators for an intact structure. Inversely, their disappearance can also be taken as the indicator for a loss of mono-crystalline structure[2]. However, this can only be observed after the structure has been lost throughout the cross section of the crystal, making it poorly suited both for determining the precise point where the usable mono-crystalline material ends and for predicting the occurrence of structure loss. This work aims to explore the use of tomographic scanning through the center of such crystals using a near-infrared laser to detect variations in transmission related to crystal abnormalities such as dislocations or aberrant structure. Three crystal samples that have been prematurely separated from the melt due to process anomalies are used. One of the samples, having a clear loss of structure, has been shown in a previous work[3] to have an anomalous transmission vs. angle profile when compared to the other two. The transmission vs. azimuth profiles through the aforementioned samples are obtained over a range of scan heights and analysed using three pattern recognition methods to provide a measure of the quality of the mono-crystalline structure (degree of homogeneity) and determine the starting point of the unusable material (lost macroscopic structure). The outcome is assessed for quality assurance applications and as an in situ monitoring tool during the production of mono-Si in the Czochralski-process.

* Corresponding author.
E-mail addresses: mathias.n.jensen@uit.no (M.N. Jensen), olav.gaute.helleso@uit.no (O.G. Hellesø).

https://doi.org/10.1016/j.jcrysgro.2022.126527
Received 10 November 2021; Received in revised form 5 January 2022; Accepted 9 January 2022
Available online 22 January 2022

2. Hypothesis and expectations

Given that pure silicon has an intrinsic band gap energy of $1.12eV$ at $300K$[4], no direct band-to-band electron transitions can occur for photon energies lower than $1.12eV$, and therefore the material is expected to be highly transparent for wavelengths exceeding $1.1\mu m$[5]. Because of this, a laser of wavelength $1.6\mu m$ ($0.775eV$) is used in this work. However, as the material of the samples is extrinsic in nature, absorption can still occur through electron transitions to and from free carrier states induced by the presence of the dopant (Boron). The magnitude of this absorption is given from the work of Schroder et al. [6]:

$$\alpha = \frac{q^3\lambda^2 p}{4\pi^2\varepsilon_0 c^3 n m^{*2}\mu} \approx 2.7{\cdot}10^{-18}\lambda^2 p \tag{1}$$

The dopant concentration specified by the manufacturer is $1.45{\cdot}10^{16} \geqslant p \geqslant 2.70{\cdot}10^{16}$ atoms pr. cm^3 for a main section positions $0 \geqslant z \geqslant 1500mm$. Assuming a constant diameter of $200mm$, this gives an expected absorption of $-13 \pm 4dB$ for a full length sample. Accounting for the reflective loss, assuming normal angle of incidence and a refractive index of 3.44[7], the expected transmission is $-16 \pm 4dB$ for the samples used in this work.

As explored in a previous work[3], the geometry of the protruding ridges of the nodes is expected to obscure the beam of the laser, resulting in no detected transmission for the azimuth of the nodes. In the case of high-quality mono-crystalline structure, it is expected that the transmission should remain mostly constant between the nodes, producing a transmission cross-section similar to what is shown in Fig. 1. In the findings of the previous work, it was also noted that while samples with "good" overall structure (highly ordered) yielded profiles that corresponded well with the expectation, a sample with known "bad" overall structure (high defect density) yielded a very different profile. In the regions where the mono-crystalline structure was compromised, the profile exhibited sparse, but intense peaks, in lieu of flat transmission in the "clear" sections between the node azimuths. The profile was shown to have eight dominant peaks at azimuths leading and lagging the four nodes by approximately 15° for the low-quality material, in contrast to the four wide plateaus observed for high-quality material[3]. From the work of Meyer et al.[8], it is known that the slip planes between homogeneously structured lattices exhibit highly anisotropic scattering, resulting in a "smearing" effect on the transmitted radiation image (see Fig. 7 of Meyer *et al.*[8]) with a scattering direction parallel to the slip plane. It is postulated that the eight distinct peaks in the profile are artefacts produced by a similar effect except that there are, in our case, multiple slip planes formed by the interfaces of the pseudo-randomly oriented lattices that form after a loss of mono-crystalline structure. From this, it is hypothesized that this abnormal transmission profile may be used as an indicator for the presence of such chaotic structure and may therefore be used as a measure of the quality of the mono-crystalline structure in cylindrical silicon ingots.

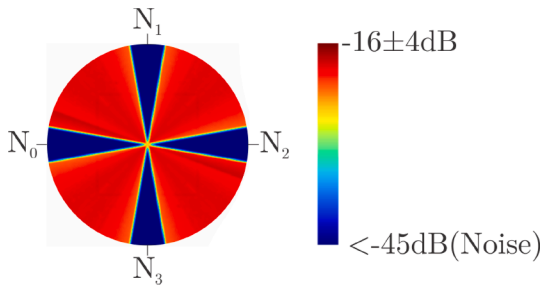


Fig. 1. Cross section of idealized transmission profile shown with nodes (N_{0-3}).

3. Experimental method

3.1. Samples

The samples shown in Fig. 2, conforming to the same specification and originating from the same foundry, are p-type (Boron doped) crystalline silicon boules manufactured to a specified diameter of 200 mm, with their dimensions, mass and bottom surface deflection (h) listed in Table 1. All three samples are the result of premature separation from the melt, resulting in the growth process being abruptly terminated such that they are effectively "snapshots" of the crystal during growth, and may therefore be used as analogues for the material in situ.

While all samples are expected to have some anomalies due to the separation shock, sample 1 shows clear indication (node termination) of unexpected loss of structure. These indications are not present in neither of two other samples. The loss of structure in sample 1 is evident by the absence of node lines in the lowermost 45 mm of the sample, see Fig. 3, and is supported by the presence of slip lines at varying positions surrounding the node lines higher up on the sample, see Fig. 3. Sample 1 is thus taken as an example of having "bad" material in its lower section, while samples 2 and 3 are taken as having mostly "good" material.

3.2. Experimental setup

The experimental setup is shown in Fig. 4, for a sample with lost structure similar to what is expected from sample 1. The central part of the setup consists of a rotating platform upon which the sample is placed. The platform is belt-driven and actuated by a DC motor providing a total gear ratio of 1:180. The laser (New Focus, Velocity set to $1600nm$) is mounted on a vertically oriented motorized translation stage (Standa, 8MT50-150BS) with a half-wave plate (Thorlabs, WPLH05M) on its output to control the beam polarization. A detector (Electro-Optical Systems, IGA-010-TE2-H) is mounted on an identical translation stage (Standa, 8MT50-150BS) such that it can travel parallel to the laser over a range of $140mm$ along the z-axis. Due to the small active area of the detector ($1mm$ diameter), a focusing lens (L) is added, increasing the pupil diameter of the detector to $25.4mm$ with an NA of 0.03. Due to the high and fixed sensitivity of the detector amplifier ($9{\cdot}10^7 V/W$), an ND-filter (effective OD 3.0 for $\lambda = 1.6\mu m$) is added to avoid saturation. This gives a high signal-to-noise ratio and matches the detectable intensity range with the observed transmission intensity, allowing better use of the detector range.

The signal from the detector is collected using a DAQ (National Instruments, USB-6009). An ad hoc timing system is implemented to serve as feedback for the DC motor, feeding a clock signal to the DAQ with a frequency of one pulse per rotation. The speed of the DC motor is manually set using a constant voltage source, while control of the

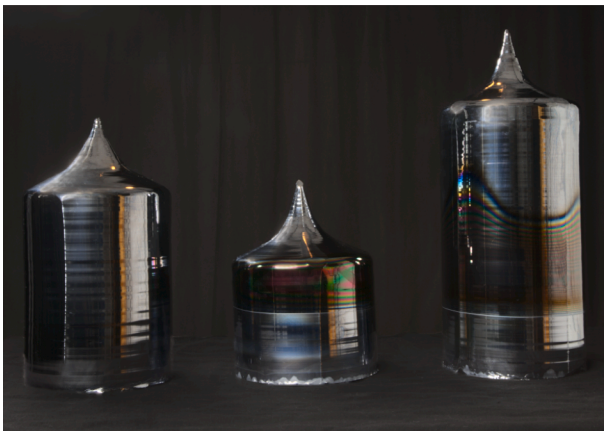


Fig. 2. Picture of silicon samples 1–3, from the left.

Table 1
Sample specifications.

	Length (*mm*)	Diameter (*mm*)	Mass (*kg*)	*h* (*mm*)
Sample 1	276 ± 1	213 ± 1	23.7	19.7 ± 0.1
Sample 2	162 ± 1	211 ± 1	10.1	11.0 ± 0.1
Sample 3	389 ± 1	212 ± 1	34.4	13.8 ± 0.1

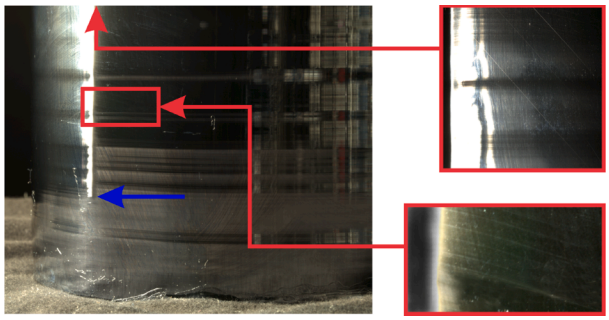


Fig. 3. Picture of sample 1 showing node termination (blue arrow) and slip lines (red frames).

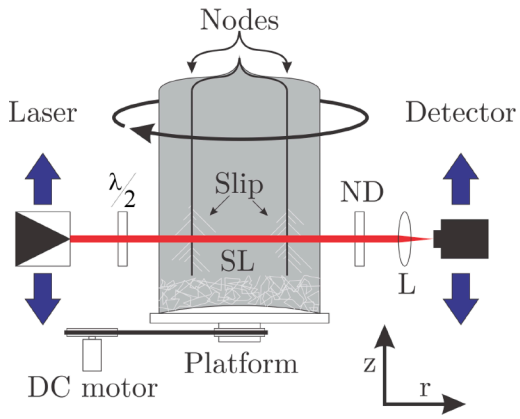


Fig. 4. Outline of setup for measuring a sample with macroscopic dislocations (slip lines) and complete collapse of mono-crystalline structure (SL).

translation stages, as well as data capture and processing, is automated (Python, V3.9.4 X64).

3.3. Scanning procedure

The scanning procedure is initiated by a mapping of optimal alignment for a predetermined set of scan positions (z-axis). This is conducted by setting the laser position to a predetermined scan position and moving the detector z-position in a range $\pm 5mm$ about the laser position.

For each detector position, data is acquired over a set number of rotations (1 by default), which is analyzed and assigned a score value. This score is determined as the mean of the upper 25th percentile of the data, encouraging the finding of a high transmission alignment, divided by the standard deviation in the percentile, discouraging the finding of an alignment that favours one or few peaks. The optimal alignment associated with each laser position is then set as the detector position with the highest score value.

Once a map of the optimal alignment positions has been obtained for a sample, the tomographic scan begins. The algorithm acquires a data set from each of the scan positions, henceforth referred to as a "slice". Each slice consists of data from a set number of revolutions (3 by default) acquired at maximum sampling frequency (24 kHz). The ad hoc timing system is used to ensure that each data-point in the slice is accurately mapped to the true azimuth at the time of acquisition. This is performed by using the clock pulse from the ad hoc timing system to confirm the rotation frequency of the crystal and setting the acquisition time accordingly. Using the clock pulse as a fixed reference for the acquisitions ensures that all slices begin at the same azimuth, and thus use a common coordinate system. To maintain accurate tracking of the rotation frequency and azimuth, the timing system periodically re-calibrates after a set number of revolutions (10 by default), in addition to the initial calibration.

4. Experimental results

The lowermost section of the samples include the concave bottom surfaces, with deflections listed in Table 1. No transmission is expected due to reflection by the bottom surface[3]. Therefore, a starting scan height of $10.5mm$ is selected. All three samples are scanned at 166 z-positions with a resolution of $0.5mm$ from $10.5mm$ to $35.5mm$ above the separation plane (bottom edge), and a resolution of $1mm$ from $35.5mm$ up to $150.5mm$.

As depicted in Fig. 5, the transmission profiles for samples 2 and 3 (Figs. 5b and c, respectively) remain largely constant throughout the samples, as expected. However, sample 1 exhibits a very different profile, that changes drastically from the sparse, sharp peaks for the lower slices to a wider plateau for the highest slice, approaching the profiles seen in samples 2 and 3. In terms of total transmission, the peaks of the profiles through samples 1 and 2 appear to remain in the vicinity of $-20dB$, which is at the lower end of the expected range. Transmission through sample 3 is somewhat lower than for the other two samples, with peaks around $-23dB$. While the absolute intensity is lower than expected, the signal-to-noise ratio is high, at $25 \pm 2dB$, and the profiles of the slices clearly distinguishes high-transmission azimuths from low-transmission azimuths.

The images shown in Fig. 6 are reconstructions of the transmission map through the crystals, obtained from the data presented in Fig. 5, using a simplified inverse Radon transform with one data-point per azimuth. The intensity of the images is given by the magnitude of the transmission for a given azimuth, but is subject to normalization and Gaussian smoothing on a per-slice basis. The images thus show the relative transmission throughout the cross section, separate from the absolute scale shown in Fig. 5.

As can be seen in the top row of Fig. 6, all three samples produce patters similar to the predicted pattern illustrated in Fig. 1, showing four clear sectors with relatively high and uniform transmission. In the next row down, this is no longer the case, as sample 1 exhibits sparse peaks with azimuths approximately symmetric around the nodes. Both samples 2 and, especially 3, continue to display a transmission profile similar to the predicted profile. In the lowermost row, the same dichotomy is present, albeit the signal-to-noise ratio for all samples is reduced. As previously stated in Section 3.1, sample 1 shows clear evidence of abnormal crystalline structure, both from the termination of the nodes and the presence of slip lines as shown in Fig. 3. From this knowledge and the observed patterns in transmission for the slices intersecting these areas, a correlation between quality of the mono-crystalline structure and the transmission profile is implied.

5. Data processing and analysis

5.1. Correlation filtering

The simplest method of differentiating two underlying patterns is through the use of correlation filtering by using a target pattern (i.e. transmission profile of high-quality material) and determining the conformity of a separate pattern to the target (i.e. quality factor) by the cross-correlation of the two. As we are interested in a scalar value for this, the peak cross-correlation between the target and measured profiles

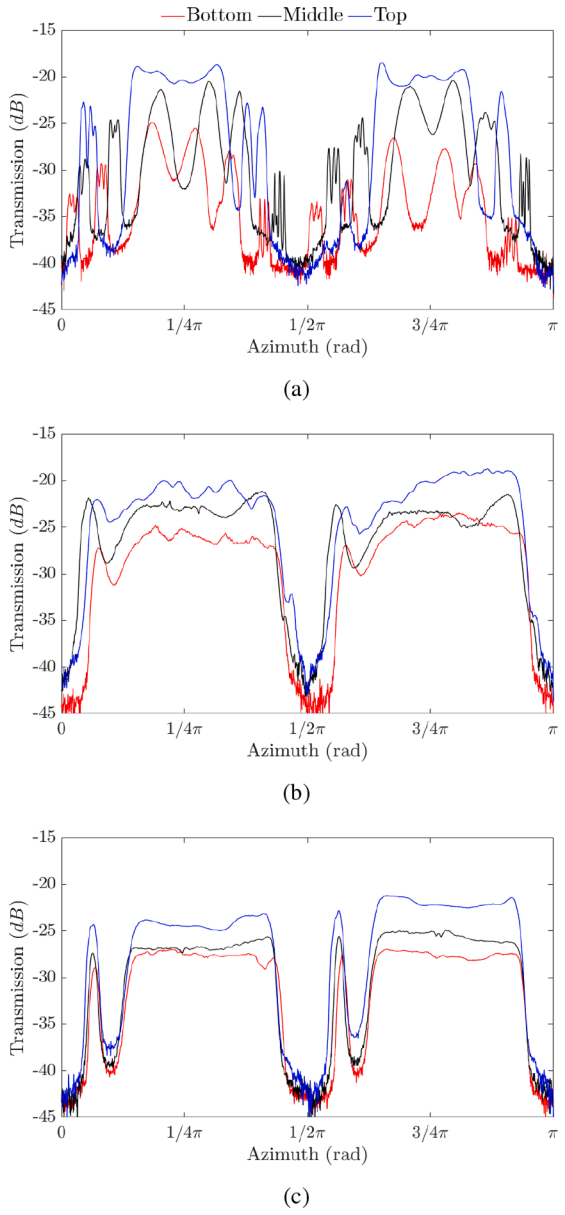


Fig. 5. Absolute transmission through the three samples at the lowermost (Bottom), intermediate (Middle,$z = 35.5mm$), and the uppermost common height (Top, $z = 140.5mm$). (a) Sample 1 (b) Sample 2 (c) Sample 3.

are taken as the raw output. The raw output of all slices considered are then normalized to a range $[0, 1]$ to make them easily comparable.

Correlation filtering is computationally inexpensive, especially in the case of a single dimension as here, which allows for easy integration as a real time feedback mechanism. A significant disadvantage, however, is the low order of complexity and thus limited ability to model complex patterns. Another shortcoming of such filters is that they are sensitive to background levels, scale and noise/distortions. These factors must therefore be attenuated by pre-processing the input passed to the filter, but the effectiveness of this is limited and often relies on assumptions regarding the properties of the factors.

5.1.1. *Ideal filter*

As previously stated, the transmission profile through a high quality crystal is expected to form two wide, pseudo-flat plateaus that are well distinguished from the underlying noise and background, as is the case throughout samples 2 and 3 (see Figs. 5 and 6).

The obvious candidate for modelling such a profile is a square wave

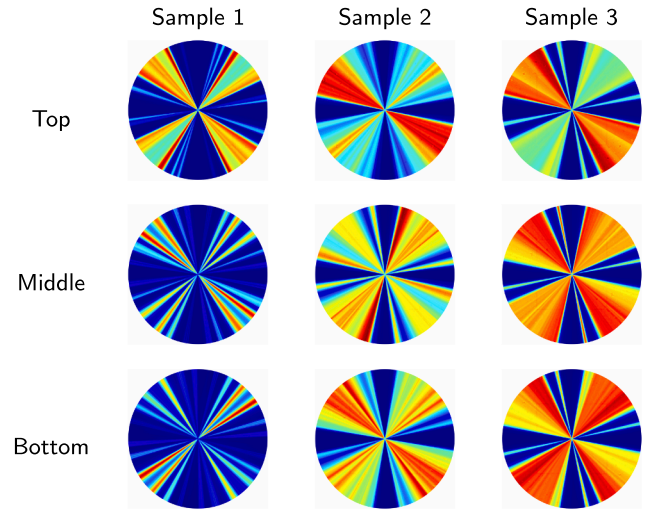


Fig. 6. Reconstruction of transmission cross-sections obtained from the measurements, with normalized linear scale indicating dark red as maximum transmission.

as shown in Fig. 7a, producing a slice cross-section as shown in Fig. 1. From this target profile, a filter is constructed and applied to the data as a correlation filter, producing the results shown in Fig. 7b. It is clear that the filter responds quite well to samples 2 and 3, returning a relatively

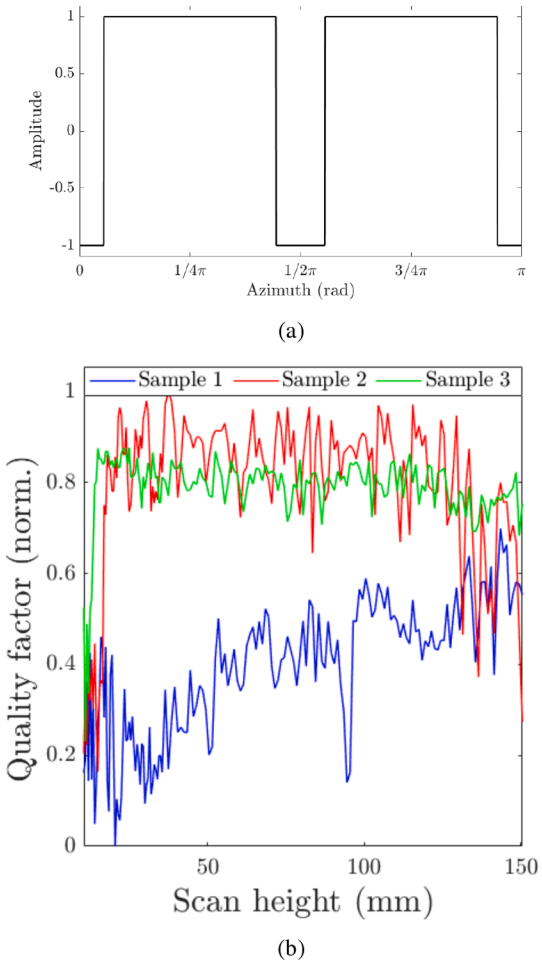


Fig. 7. Ideal correlation filter used to determine crystal quality. (a) Target profile (b) Filtering result.

high and constant value for all slices above the interface peak. Contrary to this, the filter responds relatively poorly to sample 1, returning a low response that increases towards that of samples 2 and 3 as the scan position increases. The response of this filter agrees well with the hypothesis and the observations made previously regarding the condition of the crystalline structure in sample 1 versus that of samples 2 and 3. The results of this filtering are encouraging and differentiates the known high-quality structure from the known low-quality structure with good separation of the filter response of sample 1 from that of samples 2 and 3. However, the filter response, and thus assigned quality factor, contains a high degree of noise, especially for samples 1 and 2, introducing uncertainty in determining whether the material is of usable quality or not.

5.1.2. Experimentally obtained filter

Another target profile is constructed based on the experimentally determined profiles of the known high-quality material of samples 2 and 3. The target profile is defined as the mean profile of all slices above the apex of the bottom surface (transmitted beam is detected) for samples 2 and 3. The target profile is shown in Fig. 8a normalized to $[-1, 1]$. This filter is applied in the same manner as the ideal filter and the result of the correlation filtering is shown in Fig. 8b. The filter response of the experimental filter to the data is highly similar to the response of the ideal filter in regards to its differentiation of sample 1 from samples 2 and 3. However, it appears to give a slight reduction in the noise of the response, but also gives slightly poorer separation of sample 1 from samples 2 and 3. In contrast to the ideal filter, the experimental filter better illustrates the change in transmission profiles in the slices of sample 1 as the scan height increases, showing a gradual increase until it merges with samples 2 and 3. However, in terms of differentiating the high-quality and low-quality structure, the experimental filter shows inferior performance compared to the ideal filter as the separation between sample 1 and samples 2 and 3 is small compared to the noise, giving an unfavourable overlap between them.

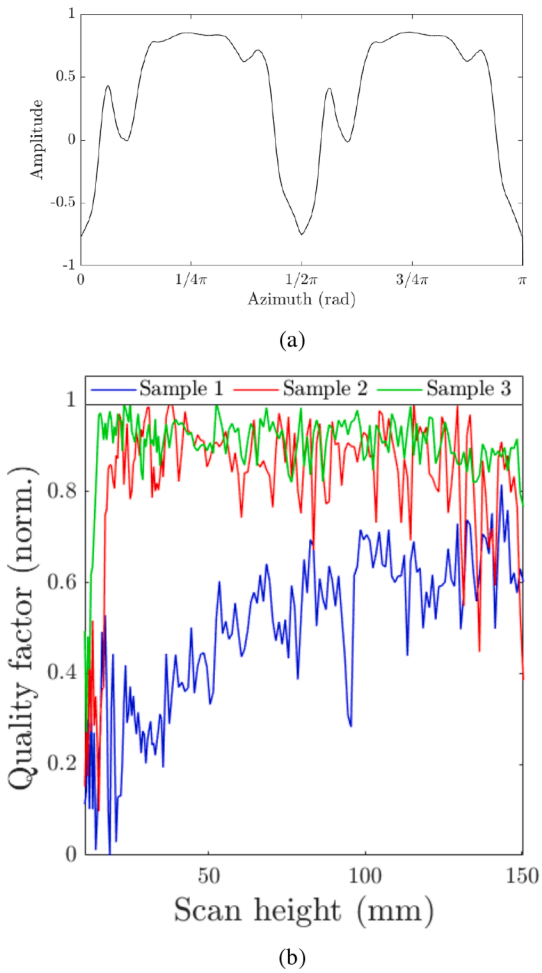


Fig. 8. Experimental correlation filter. (a) Target profile (b) Filtering result.

5.2. Neural network

As the underlying mechanisms causing the difference in the profiles of sample 1 from samples 2 and 3 are currently not well known, the features of the profiles may include complexities that are not immediately evident. While the results from both the ideal and experimental filter illustrate a clear contrast between the lower sections of sample 1 and the remaining slices, these filters may be overly simplistic and the results may therefore be, to a degree, circumstantial.

An alternative to the use of a predetermined target profile for correlation filtering is to design a filtering mechanism that adapts itself to the observed data, namely a machine learning model. The most logical type of machine learning model to use for this kind of analysis is a convolutional neural network (CNN).

The use of models such as CNNs, and other variants of neural networks[9], is becoming increasingly popular in the field of data analysis, especially for patterns that are not well known or subject to distortion/noise. The driving reason for this is that, contrary to the stiffness of correlation filters, CNNs are highly flexible and can learn to adapt to a wide range of patterns and features. Deep convolutional neural networks, with their increased depth (number of layers), are also able to learn highly abstract and complex patterns, making them able to model any pattern (given sufficient depth and width) and can learn to adapt to high degrees of distortions/noise in the data.

However, contrary to correlation filters, neural networks, and especially deep networks, rely on a longer sequence of operations that involve many more parameters, often several million, rather than the single operation of correlation filtering. Neural networks must also be taught how to determine its output, which requires a sizable pool of examples with known attributes (quality factor) and a time-consuming training phase, before it can be implemented. This makes then both computationally expensive and requires data with known properties to learn from, making them challenging to implement in real-time.

5.2.1. Architecture

A CNN consists of two main sections, the convolutional sub-network and a fully connected sub-network. The convolutional sub-network usually contains many, relatively small, learned filters and, as the name suggests, forms an output by discretely convolving the input with the filter. Such networks usually have many layers of such filters in series to form an abstract representation of the input and each layer commonly uses multiple independent filters in parallel to produce multiple output channels containing different representations of the input. A CNN normally uses this convolutional sub-network as a preprocessing mechanism that reduces the input to a concentrate which is then fed to the fully connected (FC) sub-network. The fully connected sub-network is a classical neural network consisting of a determined number of layers (depth), each with a determined number of neurons (width). After the input has passed through the fully connected sub-network, it reaches the head of the network, which produces the final output of the network based on the output of the last layer of the fully connected sub-network. The architecture of the chosen deep CNN is shown in Fig. 9.

As the input to the network is a one dimensional vector, all of the convolutional filters are also vectors. As shown in Fig. 9, this architecture also employs connections that directly bypass the filters, so-called "skip-connections" common to the ResNet architecture[10]. The use of skip-connections encourages the network to learn filters that modify the input to produce the output, rather than creating an entirely new representation. This added bypass helps the network learn faster by

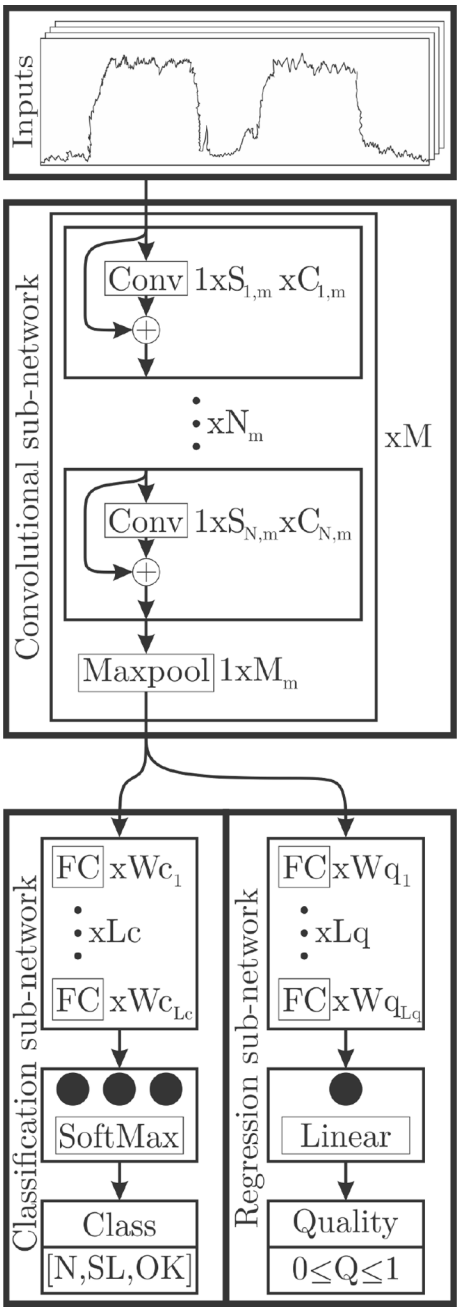


Fig. 9. Generalized architecture of deep CNN with modular convolutional blocks, taking the transmission profiles as its input and returning a quality factor and prediction of the usability of the material (OK vs. N or SL).

smoothing out the loss surface[11], allowing the gradient descent algorithm to more easily converge to the optimal network parameters.

The overall structure of the network is designed to be highly modular, by grouping the layers into blocks that can be configured as needed. The convolutional sub-network contains M convolutional blocks, each terminated by a maxpool layer to reduce the size of the block's output. For each block $m \in [1..M]$, there are N_m convolutional layers, and each layer $n \in [1...N_m]$ has a defined filter size $S_{n,m}$ and number of parallel filters $C_{n,m}$ that can be configured individually, as needed.

Once the input has been passed through all convolutional blocks, it is returned as the output of the convolutional sub-network and passed to two parallel, fully-connected sub-network. The two sub-network are independent and may be configured separately with any chosen depth (Lc and Lq) and width (Wc_k for $k \in [1..L_c]$ and Wq_l for $l \in [1..L_q]$). The classification sub-network aims to learn a way of predicting if the input originates from a noise-only signal (N), the profile of a slice with low-quality structure (SL), or the profile of a slice with high-quality structure (OK). This is achieved using three terminating neurons that take the final layer of the classification sub-network as input, weighting them and summing them before passing the result to a softmax activation function that redefines the raw output of the neurons into a probability distribution. The predicted class is then taken as the class with the highest probability given by the softmax output. The regression sub-network assigns a numerical value to the quality of the structure, based on the input profile, to create an output that is comparable with the filter responses shown in Figs. 7b and 8b. The output of this sub-network is determined as the weighted sum of the last layer in the sub-network and returned as a number (Q).

5.2.2. Configuration and evolution

Given that the presented data is in the form of single-dimension vectors, a small network should be sufficiently detailed, but the balance between the size and complexity of the convolutional, classification, and regression sub-networks must be chosen carefully. Due to the unknown complexity of the underlying features and their distribution in the data set, a larger network with many parallel convolution filters may provide greater performance than a small one. However, such a large network would require more example profiles to learn from and may be prone to over-fitting to the data without learning the underlying pattern. Conversely, a small network with more convolutional layers, but few parallel filters, would be superior in forming abstract patterns due to the extensive transformation of the input through many convolutional layers. This may help the network learn the underlying pattern better, especially on small training sets, but the lack of parllel representations of the input may limit the level of complexity the network can account for.

Because of this challenge, further inspiration is taken from nature by introducing the concept of evolution through a genetic algorithm[12], loosely based on the method employed by Dahou et al.[13]. This algorithm defines 13 hyper-parameters ("genes") that define the structure of the networks, these then define:

- Number of convolution blocks and layer configurations within each block
- Size of convolution filters in each block
- Number of parallel filters in each block
- Width and depth of classification and regression FC layers, separately
- Drop out rate for convolution, regression, and classification sub-nets separately
- Activation function for convolution, regression, and classification sub-nets separately

This algorithm then creates a population of five networks from a predetermined pool of configuration hyper-parameters using a random combination of the hyper-parameters. These five networks (generation zero) are then partially trained over 10 epochs on a common training set before evaluation on a separate common test set, defining the fitness score as the inverse of the loss (cross-entropy). To prevent devolution, the fittest network is always passed to the next generation such that the best of every subsequent generation is always as good or better than the previous one, driving the "survival of the fittest"-mechanism. The top two networks of the previous generation are then hybridized to produce two new networks. This process randomly combines the hyper-parameters ("genes") of the "parents" to produce "offspring" with a configuration containing only those "genes" that prove most advantageous, while providing a mechanism for mutation. The remaining two slots of the next generation are filled by producing two random networks from the original pool of configurations and hybridizing them with the two top performing networks from the previous generation. This is done to increase the genetic diversity of the population, increasing the

likelihood that advantageous "genes" outside those of generation zero are introduced into subsequent generations. After six generations of evolution, the process is terminated and the best performing network is taken as the "apex"-network, giving the configuration listen in Table 2.

In addition to the evolved configuration shown in Table 2, the convolution sub-network employs a rectified linear (ReLU) activation function with dropout rate 0.4, the classification sub-net employs a ReLU activation with dropout rate 0.5, and the classification sub-net employs a sigmoid activation with dropout 0.3.

5.2.3. Training and results

The original data set used to generate Figs. 7b and 8b contains 166 profiles for each of the three samples (498 in total). This is not sufficient to train the CNN, especially since the true nature and quality of the material in sample 1 is unknown. Therefore the data set is augmented by adding profiles obtained from earlier phases of the experiment, when polarization sensitivity was investigated. These profiles showed no discernible sensitivity to polarization but are taken at different scan heights from the final phase. Thus, they are not evaluated in parallel to the main set, but can be considered to represent the same patterns, bringing the total number of profiles to 996. Due to the true nature and quality associated with certain profiles from sample 1 being unknown, these are omitted, reducing the number of profiles viable for training to 675. To encourage the network to learn the underlying pattern instead of specific features of the profiles, the data set is augmented by creating synthetic profiles from the measured real profiles. These synthetic profiles are randomly selected from the expanded set of 675 profiles and a random selection of these are again modified with added Gaussian noise (mean of zero and standard deviation $-13dB$ of profile amplitude) to emulate measurements with a lower signal-to-noise ratio. Since the transmission for a given azimuth is assumed to be independent of preceding measurements, a change in direction of rotation is emulated by time-reversing (reversing the mapping of data-points to azimuth) a set of randomly selected profiles from the expanded set. The augmentation generated 508 synthetic profiles (75.3% of real profiles), giving 1183 profiles usable for training the CNN.

The profiles in the data set are then randomly shuffled to homogenize the set, increasing the likelihood that the subsets of it used by the CNN during training and testing are representative of the whole. The profiles in the set are also given a class label and a quality label for the CNN to use as reference during training. As the exact numerical value of the material quality is not well known, noise is added to the quality labels to encourage the model to learn a more abstract quantification rather than a discrete binary one.

The shuffled and labeled data set is then randomly separated into a testing and a validation set containing 10% of profiles each, while the remaining 80% is reserved for training.

As training during the evolution phase uses the same source data as here, there is a risk of overlap between the training set used during evolution and the testing or validation sets used here, which would bias the results. To prevent this, the "apex"-network is rebuilt such that its hyper-parameters ("genes") are kept, but all parameters learned from training are reset.

The "apex"-network is then trained on the generated training set over a total of 40 epochs using an Adam optimizer with a learning rate reducing from 10^{-2} to 10^{-5} over four steps (10 epochs per learning rate). The regression sub-network is subsequently trained on top of the main network, using the same training data and procedure.

Once trained, the performance of the network is evaluated using the testing set, giving an overall prediction accuracy of 92.2% and the confusion matrix shown in Table 3. The CNN predicted 98.7% of the known "good" material profiles (*OK*) as "good" ($\widehat{OK}$) while none of the known "bad" (*SL*) or pure noise (*N*) are misclassified as "good". However, there is some misclassification between the pure noise and the "bad" material as shown by 38.5% of the true "bad" being misclassified as noise and 13.3% of true noise being misclassified as "bad" material. The pure noise case is only present for prematurely separated samples (due to the concave bottom surfaces) and not in a complete boule. Thus, the important factor is the overlap between the "bad" and "good" predictions, which can be argued to be taken as less than or equal to the 1.3% overlap between *OK* and $\widehat{N}$.

While the performance of the network in differentiating the noise only (*N*) from the "bad" structure (*SL*) is not optimal, its capability in differentiating the unusable (*N* or *SL*) from the usable material (*OK*) is shown to be excellent with one misclassification among 75 true good profiles. Upon investigation, it is found that the misclassified profile is a weak signal with artificially added noise, making it appear as pure noise. It can also be observed that the CNN responds to the synthetic and real data in the same manner, predicting the same class and approximately equal quality for both real data and the derived synthetic data. This confirms that the synthetic data contains the same underlying pattern as is present in the real data and that the CNN is able to recognize the pattern in the presence of modifying factors such as noise or a change in direction of rotation.

Applying the network to the data set presented in Figs. 7b and 8b yields the classification shown in Fig. 10a and the quality factor shown in Fig. 10b. As seen in Fig. 10a, the network returns a high degree of certainty that there is no bad structure in samples 2 and 3, but that there is a high density of such occurrences in the lower section of sample 1. The quality factor shown in Fig. 10b, agrees with the results of correlation filtering shown in Figs. 7b and 8b, but shows better separation of the low-quality material of sample 1 and the high-quality of samples 2 and 3, while also exhibiting significantly less noise. Both the quality factor and the classification of "good" vs. "bad" structure agrees well with what is inferred in Section 2 regarding the quality and viability, of the material in sample 1 increasing with distance from the onset of structure loss.

5.2.4. Robustness of the trained model

The purpose of using a CNN instead of the simpler forms of filtering, such as described in Section 2, is its lower sensitivity to noise in the input signal. To illustrate this, the trained CNN is fed two previously unseen profiles, one with known low quality and one with known high quality, while observing the class (N, SL or OK) and quality factor predicted by the CNN.

The two profiles are replicated 100 times and contaminated with randomly generated noise before being passed to the CNN, the robustness of the network is then evaluated by increasing the severity of the added noise until the output of the CNN is affected by either a change in predicted class (ΔC) or a 5% change in assigned quality-factor(ΔQ). For fairness, three types of noise are used: Gaussian ($\mathcal{N}(0,\sigma)$), Poisson ($A{\cdot}\mathcal{P}(\lambda)$), and noisy sine ($B{\cdot}Sin(\propto 1/p) + \mathcal{N}(0, B/10)$). These noise types

Table 2
Configuration of CNN.

Block	Layers	Size	Channels	Parameters
Conv 1	5	5	32	21 k
Conv 2	5	5	64	92 k
Conv 3	5	5	128	369 k
Conv 4	5	5	256	1.5 M
FC Class	5	700	-	22.2 M
FC Regr.	6	512	-	16.1 M
			Total	40.3 M

Table 3
CNN confusion matrix for test data set.

	$\widehat{N}$	$\widehat{SL}$	$\widehat{OK}$
N	86.7%	13.3%	0.0%
SL	38.5%	61.5%	0.0%
OK	1.3%	0.0%	98.7%

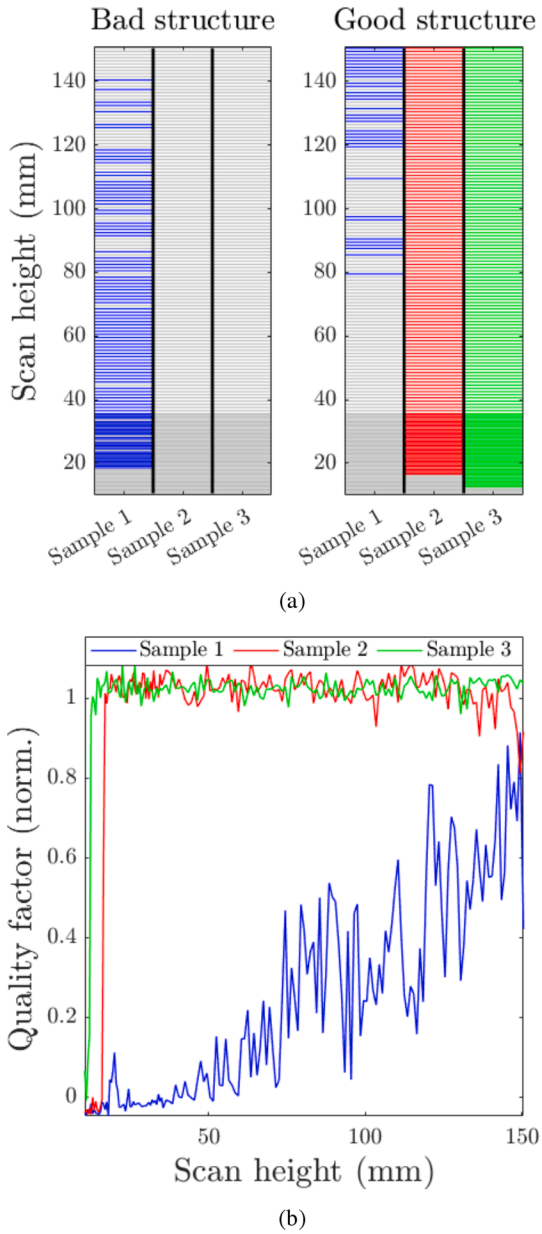


Fig. 10. Result of deep learning filter. (a) classification of low- vs. high-quality predictions (b) Assigned quality factor of material vs. slice height.

are then tried for all combinations of their parameters (1000 points pr. parameter) and are considered for the extremes of the noise types, returning the signal-to-noise thresholds listed in Table 4. As shown from the noise testing, the CNN is quite robust against all three types of noise, showing robust predictions for input profiles with an induced signal-to-noise ratio down to $5dB$, compared to the measured SNR of $> 23dB$ as mentioned in Section 4.

The regression sub-network is shown to be more robust than the classification sub-network, being able to withstand $1.4-9.9dB$ lower SNR without deviating more than 5% in predictions. Notably, it is also observed that all the noise-induced misclassifications are on the low-quality profile while the high-quality profile is not misclassified, implying a higher false positive rate than false negative rate. Albeit not shown, both classification and regression outputs of the CNN exhibit a slightly higher ($1.0-1.7dB$) tolerance to sinusoidal noise if the period of the noise is a whole fraction of a rotation ($360°, 180°, 120°...36°$).

Table 4
CNN noise threshold.

Noise type	Parameter	Threshold	ΔC	ΔQ
Gaussian	-	σ^{-1}	$2.6dB$	$-5.2dB$
Poisson	$\lambda = 0.2$	A^{-1}	$-7.0dB$	$-16.9dB$
	$\lambda = 20$	A^{-1}	$-6.3dB$	$-7.7dB$
Sine	$p \in [0, 10]$	B^{-1}	$5.0dB$	$-7.7dB$

6. Conclusion

Three samples of mono-crystalline silicon with diameters of $212 \pm 2mm$ are observed to have a transmission of $-21 \pm 2dB$ for a near-infrared laser of wavelength $1.6\mu m$. Tomographic scanning is performed on all three samples over a lateral range of $140mm$ from $10.5mm$ to $150.5mm$ above the separation plane. A total of 498 transmission profiles, measured as a function of transmission vs. crystal azimuth, are obtained for the three samples.

The transmission profiles are shown to produce a consistent pattern for a given vertical slice (z-position) of the crystal. All recorded transmission profiles from all three samples exhibit the same four blackout-zones due to beam refraction/obstruction by the node geometries. The profiles are also shown to exhibit unique features for slices intersecting material of known defective crystalline structure that differs notably from the profiles for slices intersecting intact mono-crystalline structure.

A quality-score can be determined from the transmission profile of a slice using targeted correlation filtering. An idealized square wave, with negative amplitude around the node azimuths, is used as the target profile to assign a quality factor to the measured profiles to produce a quality factor that is in agreement with observations and expectations regarding the samples. A second target filter is obtained from the measurement as the mean profile of assumed intact material from two of the samples. This filter also produced a viable quality factor in the presence of suspected poor/no structure, albeit with a higher degree of overlap between the known poor- and high-quality material compared to the idealized square wave filter.

A deep convolutional neural network (CNN) is also investigated as an analysis tool. The CNN is implemented as a modular architecture with a configuration determined by a set of 13 hyper-parameters that define the properties of each of its elements. The hyper-parameters of the CNN are determined by a process of evolution using a genetic algorithm to create a CNN whose configuration is best suited to learning the patterns observed in the profiles. The evolved CNN is configured with 20 convolutional layers, preceding a fully connected classification head of five layers and a fully connected regression head of six layers, giving a total depth of $20+5/6$ layers and 40.3 million parameters.

The CNN is then trained over a total of 40 epochs and tested on a separate testing set of 118 profiles, achieving an accuracy in differentiating the assumed intact structure from the known defective structure (or noise) of 98.7%. The predictions of the CNN results in a quality factor consistent with the results of both correlation filters, albeit with significantly reduced noise and vastly improved contrast between low- and high-quality material. The CNN also successfully classifies all slices of all three samples, yielding a map of lost vs. intact structure that agrees well with observations and expectations regarding the samples.

The observations and results of this work show that a consistent pattern in the transmission profiles coincide with the state of the crystalline structure, and that this can be used to quantify the quality of the structure. As these experiments are conducted at room temperature, it is feasible that these methods could also work for complete boules under the same conditions. The non-destructive nature of this method allows for quality testing to be conducted on every produced boule without the need for slicing, and subsequent loss of material. The method also make it possible to determine the precise boundary between usable and unusable material, enabling smaller margins to be used when removing the unusable material, thus improving material yield.

Because the temperature of the material during production is significantly higher than for the conducted experiments, it is unknown if these methods can be adapted as a real-time feedback system. Further investigations on near/mid-infrared transmission through silicon at high temperatures ($\geqslant 1400K$) are required.

Validation of the method proposed in this work would require additional material samples and added measurements of these, presumably through destructive methods such as carrier density imaging and/or lateral photovoltaic scanning of the sliced samples, to provide a more detailed reference point for the assessment of the method and analysis. Future work on this concept would also include investigations of other deep-learning architectures, both more comprehensive evolved networks and known established architectures such as ResNet, VGGXX, and GoogLeNet.

Author contributions

The second author suggested the problem and the authors conceived the measurement methods together. The first author conducted the experimental work and the data analysis. The authors contributed equally on the written work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors gratefully acknowledge financial support from the Research Council of Norway (No. 302333), and would like to thank Norwegian Crystals AS for supplying the samples and relevant reference information.

References

[1] A. Lanterne, G. Gaspar, Y. Hu, E. Øvrelid, M. Di Sabatino, Characterization of the loss of the dislocation-free growth during czochralski silicon pulling, J. Cryst. Growth 458 (2017) 120–128.
[2] Øyvind S. Sortland, E.J. Øvrelid, M. M'Hamdi, M. Di, Sabatino, Investigation of pinholes in czochralski silicon ingots in relation to structure loss, J. Cryst. Growth 510 (2019) 1–6.
[3] M.N. Jensen, O.G. Hellesø, Measuring the end-face of silicon boules using mid-infrared laser scanning, CrystEngComm 23 (2021) 4648–4657.
[4] V. Alex, S. Finkbeiner, J. Weber, Temperature dependence of the indirect energy gap in crystalline silicon, J. Appl. Phys. 79 (1996) 6943–6946.
[5] S.E. Aw, H.S. Tan, C.K. Ong, Optical absorption measurements of band-gap shrinkage in moderately and heavily doped silicon, J. Phys.: Condens. Matter 3 (1991) 8213–8223.
[6] D.K. Schroder, R.N. Thomas, J.C. Swartz, Free carrier absorption in silicon, IEEE J. Solid-State Circuits 13 (1978) 180–187.
[7] D. Chandler-Horowitz, P.M. Amirtharaj, High-accuracy, midinfrared (450cm-1–4000cm-1) refractive index values of silicon, J. Appl. Phys. 97 (2005) 123526.
[8] M. Meyer, M.H. Miles, T. Ninomiya, Some electrical and optical effects of dislocations in semiconductors, J. Appl. Phys. 38 (1967) 4481–4486.
[9] M. Avci, S. Yamacli, Neural network reinforced point defect concentration estimation model for czochralski-grown silicon crystals, Mathematical and Computer Modelling 51 (2010) 857–862. 2008 International Workshop on Scientific Computing in Electronics Engineering (WSCEE 2008).
[10] Z. Wu, C. Shen, A. van den Hengel, Wider or deeper: Revisiting the resnet model for visual recognition, Pattern Recogn. 90 (2019) 119–133.
[11] H. Li, Z. Xu, G. Taylor, C. Studer, T. Goldstein, Visualizing the loss landscape of neural nets, 2018. arXiv:1712.09913.
[12] J. Shapiro, Genetic Algorithms in Machine Learning, Springer Berlin Heidelberg, Berlin, Heidelberg, 2001, pp. 146–168. doi: 10.1007/3-540-44673-7_7.
[13] A. Dahou, M.A. Elaziz, J. Zhou, S. Xiong, Arabic sentiment classification using convolutional neural network and differential evolution algorithm, Computational Intelligence and Neuroscience 2019 (2019) 2537689.

Paper III: Identification of extracellular vesicles from their Raman spectra via self-supervised learning

Submitted to Nature Scientific Reports, October 2023. **Authors:** Mathias N. Jensen, Eduarda M. Guierreiro, Agustin Enciso-Martinez, Sergei G. Kruglik, Cees Otto, Omri Snir, Benjamin Ricaud, and Olav Gaute Hellesø
Contribution notes: Mathias N. Jensen concieved the idea and implemented the method. Eduarda M. Guierreiro and Omri Snir prepared and provided samples for data generation. Agustin Enciso-Martinez and Sergei G. Kruglik conducted the experimental work. Cees Otto conducted and oversaw parts of the experimental work. Benjamin Ricaud oversaw the development of the architecture. Olav Gaute Hellesø oversaw the work. Mathias N. Jensen wrote the initial draft, Olav Gaute Hellesø and Benjamin Ricaud finalized the manuscript for submission.

Identification of extracellular vesicles from their Raman spectra via self-supervised learning

Mathias N. Jensen[1], Eduarda M. Guerreiro[2], Agustin Enciso-Martinez[3,4,5], Sergei G. Kruglik[6], Cees Otto[7], Omri Snir[2,8], Benjamin Ricaud[1], and Olav Gaute Hellesø[1,*]

[1]Department of Physics and Technology, UiT The Arctic University of Norway, Tromsø, Norway.
[2]Thrombosis Research Group (TREC), Department of Clinical Medicine, UiT The Arctic University of Norway, Tromsø, Norway.
[3]Oncode Institute and Ten Dijke/Chemical Signaling Laboratory, Department of Cell and Chemical Biology, Leiden University Medical Center, Leiden, The Netherlands
[4]Amsterdam Vesicle Center, Department of Biomedical Engineering and Physics, Amsterdam University Medical Centers, Amsterdam, The Netherlands
[5]Laboratory of Experimental Clinical Chemistry, Department of Clinical Chemistry, Amsterdam University Medical Centers, Amsterdam, The Netherlands
[6]CNRS, Institut de Biologie Paris-Seine, Laboratoire Jean Perrin, Sorbonne University, Paris, France
[7]Department of Medical Cell BioPhysics, TechMed Centre, University of Twente, Enschede, The Netherlands
[8]Department of Medical Biology, UiT The Arctic University of Norway, Tromsø, Norway
[*]olav.gaute.helleso@uit.no

ABSTRACT

Extracellular vesicles (EVs) and nanoparticles released from cells attract interest for their possible role in health and diseases. The detection and characterization of EVs and other biological nanoparticles is challenging due to the lack of specialized methodologies. Raman spectroscopy, however, has been suggested as a novel approach for biochemical analysis of nanoparticles. To extract information from the spectra, a novel deep learning architecture is explored as a versatile variant of autoencoders. The proposed architecture considers the frequency range separately from the intensity of the spectra. This enables the model to adapt to the frequency range, rather than requiring that all spectra be pre-processed to the same frequency range as it was trained on. It is demonstrated that the proposed architecture accepts Raman spectra of nanoparticles from multiple biological origins and laboratories. High reconstruction accuracy is maintained despite large variances in frequency range and noise level. It is also shown that the architecture is able to cluster nanoparticles by their Raman spectra and differentiate them by their origin without pre-processing of the spectra or supervision during learning. The model performs label-free differentiation of nanoparticles from 13 biological sources with high fidelity, including separating extracellular vesicles from activated vs. unactivated blood platelets and nanoparticles from prostate cancer patients vs. non-cancer controls. The differentiation is evaluated by creating a neural network classifier that observes the features extracted by the model to classify the samples according to their origin. The classification reveals a test sensitivity of 92.2% and selectivity of 92.3% over 769 nanoparticles measured at two different labs with two different measurement configurations.

Introduction

Extracellular vesicles (EVs) are nanostructures confined by a lipid bilayer, produced by all types of cells, and released into the extracellular space. EVs play crucial roles in cell-to-cell communication and coagulation. Given their abundance in biological fluids and circulation, they have been identified as biomarkers for various inflammatory diseases and cancer[1–4]. This has increased the interest in EVs and the development in the field, as documented in recent reviews[5,6]. The biochemical composition of extracellular vesicles (EVs) is highly heterogeneous. The nanoscale size of EVs and other biological nanoparticles poses a significant challenge regarding analysis, and the field still lacks standardized analytical approaches. Currently, researchers employ several methods, including nanoparticle tracking analysis (NTA)[7], transmission electron microscopy (TEM), flow cytometry[8], and various chemical/biological techniques[9]. NTA measures the size distribution of EVs by tracking their Brownian motion. While it provides valuable size information, it does not offer insights into the biochemical properties of EVs, and size distribution alone is not sufficient to attribute EVs to their cellular origin. To use EVs as a biomarker for various diseases, suitable characterization and data analysis methods must be found.

Raman spectroscopy is a label-free methodology providing information about the chemical composition of Evs at the single, or near-to-single EV level. Thus, Raman spectroscopy is a viable alternative to established chemical analysis tools,

such as magnetic resonance and mass spectrometry[10,11]. By combining Raman spectroscopy with optical trapping, single particles down to nanoscopic sizes can be isolated and measured independently[12,13]. This combination is very promising for characterization of EVs. However, it is a challenge to decode the Raman spectra into the corresponding chemical mixture. For simple materials, such as polymers, decomposing the Raman spectra is often a relatively trivial task due to the inherent sparsity of vibrational modes in such materials[14,15]. For more complex samples, such as biological materials, their features are both numerous and often overlapping[16–18]. As an example, the true proportion of each biomolecule in EVs is rarely available, and there is thus no benchmark for decoding the Raman spectra. As Raman scattering is very weak, the spectra contain various amounts of noise (on y-axis) and the wavelenth shift (along x-axis) can be stretched or shifted, depending on the optical spectrometer and the calibration of it. To further complicate the picture, different labs have different set-ups, giving Raman spectra with different range for the wavelength shift.

Our aim is to have a flexible model that can extract high-quality, chemically significant information from the spectra and use this information to classify EVs and other biological nanoparticles in the presence of noise and variations in wavelength calibration using data from multiple sources. The complex challenge of using Raman spectra of biological nanoparticles to classify and, ultimately, use the spectra of the nanoparticles as a biomarker for diseases, requires a flexible analysis method that can handle large variations in the input, but still extract information that reflects the chemical composition of the EVs and other nanoparticles.

Common approaches to analyse Raman spectra are signal processing and analysis methods that decompose the spectra into their more fundamental components, which can, in some cases, be associated with known biochemicals. Principal component analysis[19] and k-means clustering[20] are two common methods applied for this purpose. They are often complemented by a classification method, such as linear discriminant analysis[21] or a support vector machine[22]. A method has also been proposed for relating Raman spectra to the biomolecular composition of the sample, called biomolecular component analysis[?]. Some of the data presented here (from Sorbonne University) has been analysed by biomolecular component analysis[23]. While these methods have demonstrated their usefulness, they are limited by their relatively simple function and thus their limited ability to consider complex patterns and dependencies in data. Neural networks and deep learning are very efficient methods for analysing data. The ability to learn underlying aspects in the data and make inferences based on highly non-linear relations has made neural networks, and derivative architectures such as convolutional neural networks, prevalent in the field of data analysis and they have been successfully applied on spectral information[1,24,25]. Another significant advantage of neural networks is their adaptability to noise in the data and variations in the signal background[1,26]. However, these methods often require several thousand examples to learn from and these examples have to be rigorously curated to avoid biasing the model. The last requirement often implies that the setting of the data must be made uniform, with the same frequency range and resolution for all spectra. This puts strict restrictions on the data that can be used, and calibration drift is a common problem which is not acceptable to such a model. For data from multiple sources with differences in range and resolution, the solution becomes pre-processing of the data by truncation, interpolation and normalization, in an attempt to emulate uniform settings.

We propose a neural network architecture specially taylored to handle Raman spectra from samples for which we only have a few labels. It is based on a self-supervised training approach. The architecture and training take care of the specificities of the spectra, of the noise properties from the measurements and on the variability of the recording from different places and devices. The general achitecture is based on a Variational AutoEncoder[27] (VAE). However, since the data from multiple sites can have different ranges, the autoencoder uses a novel formatting. It considers the signal (y-axis) and the frequency range (x-axis) separately in order to handle both noise (on y-axis) and changes in wavelength calibration (x-axis). This split is also useful for equalizing the input data size by applying a resampling step with interpolation. The architecture further includes a suitably sized latent space and a loss function adapted to picking-up the significant spectral information.

We adopt the standard approach in self-supervised learning[28]. First, we train the autoencoder in a self-supervised manner, i.e. without labels. Training data are generated by adding Gaussian noise, wavelength shift and clipping of the original spectra. In addition, the training data contains spectra from other types of particles (liposomes) and pure noise. The task of the network is to recover the original spectra from the artificially corrupted ones. While training on this data, the network will identify important information[29] in the input that enables it to recover the original spectra. The network will build an inner representation of the data, in what is called a latent space. Data encoded in this latent space should contain only the essential information to reconstruct the original spectra, without any noise, and this information will ideally approximate the chemical information in the particles.

In the second stage of the learning process, we take advantage of the latent space and use it as the input of a second deep neural network which will learn to associate the data to labels given by the origin of the EVs. The latent representations used as input should reflect the chemical information of the particles, contained inside the spectra and be free from noise. This is the advantage of using the latent representation of the first, self-supervised, network. The classification task is then made easier, allowing the use of a smaller network and a reduced need for labelled data. Further details are given in the next section.

Raman spectra from two laboratories, at University of Twente (Netherlands) and at Sorbonne University (France), are used

to train and test the model. For simplicity, the two datasets are referred to as 'Twente' and 'Paris'. The Twente dataset was larger than the Paris dataset, which gives an imbalance when training the model. Variations in measurement method, notably excitation wavelength and acquired wavelength range, and in sample preparation can give features and distortions in the data that the model should learn to disregard, thus recognizing the particles more by their chemical information rather than the condition of the data. For one case the two laboratories analysed EVs derived from the same, single cell type: blood platelets. For this case, it is important to see if the model recognizes the origin of the EVs or at which lab they were analysed.

The performance of the model regarding reconstruction of spectra, extraction of information-rich features, and classification is evaluated. The classification of EVs based on the information extracted by the model is used to verify the quality of the information and the viability of using Raman spectra of EVs as biomarkers for a range of conditions. The analysis methodology proposed in this article will be a valuable tool in that process, as it can handle data from several laboratories with variations in measurement settings. These problems and differences between datasets can be found for many applications, including other types of spectroscopy and in general for one-dimensional datasets from different sources.

Methods

Sample preparation

The samples considered in this work are EVs and nanoparticles from 13 biological origins, acquired with two different measurement systems in two different laboratories. Out of the 13 origins, three are comercially available cell line cultures, two are from bulk human blood, two are from extracted platelets, two are derived from sampled human cells, and the last four are lipoproteins. See Table S1 in Supplementary Information for an overview.

The three cell lines used are LNCaP, PC3 and THP-1, these are used to investigate if they are recognized as different from the human derived samples and how they are grouped relative to each other. The two samples from bulk blood are EVs taken from blood plasma and red blood cells (hence RBC) to investigate how the model perceives their similarity and difference. The two samples from blood platelets are isolated from the other elements of blood and are presumed to be from pure cultures. These are included to investigate whether Raman of their EVs can be used to determine if the platelets are activated (clotting) or not. The platelet derived EV data comes from three datasets:

- 1. Control: platelet EVs from untreated platelets from donors
- 2. Activated A23: EVs from artificially activated platelets using a calcium-activator (A23187) prepared by and measured at the Sorbonne Université
- 3. Activated Trap: EVs from artificially activated platelets using a thrombin-activator (TRAP6) prepared by and measured at the Sorbonne Université
- 3. Mixed: platelet derived nanoparticles from platelet concentrate prepared by and measured at the University of Twente

The two human cell derived samples originate from two groups, one from patients afflicted with prostate cancer and one from non-afflicted controls. This data is included to investigate if the model can determine if the biological nanoparticles originate from a healthy or afflicted person and to investigate how these two are clustered relative to each other. Lastly, the four types of lipoproteins (CM, HDL, LDL and VLDL) are included to investigate how the model reacts to something that is not an EV, but chemically similar. This also poses a special challenge for the model as the number of datapoints from lipoproteins is comparatively low and, due to their small size, also has a very low signal to noise ratio.

Data acquisition

The Raman spectra of the various EVs are collected with optical tweezers using two separate measurement systems[12,30]. In both systems, the trapping and excitation of the Raman scattering is performed by a single high power laser relayed by a high NA objective. The same objective also collects the Raman scattered light in backpropagation mode. In system 1, the laser source is an Ar-ion pumped Ti:Sapphire laser delivering approximately 100mW at 780nm to the sample volume via a water dipping objective (Olympus LUMFL, 60X, NA = 1.1). The backpropagated response is collected by the objective and passed to a 500mm focal length grating spectrograph with a liquid nitrogen cooled detector and a $50\mu m$ slit aperture, acquiring the spectra over a range of 309-2035cm^{-1}. In system 2, built by Ing. Aufried Lenferink at University of Twente, the laser source is a Krypton-ion laser (Coherent, INNOVA 90-K) delivering 70mW at 647nm to the sample volume via a non-immersion objective (Olympus, 40x NA = 0.95). The backpropagated response is collected by the objective and passed to a prism-based spectrometer built in-house with a Peltier-cooled detector, acquiring the spectra over a range of 301-3655cm^{-1}.

Datasets

The data used in this work consists of 2667 spectra from 13 origins as described previously in Sec. and listed in Table S1. One of the challenges is that the wavenumber range and center varies from spectrum to spectrum, especially since the 'Paris' dataset originates from a system with a much shorter wavenumber range than the 'Twente' dataset. The difference in the wavenumber range is shown in fig.1, and the figure also shows the difference in the signal-to-noise ratio between the two datasets. There is also a significant variation in the preparation of the samples, particularly the activation of the platelets, as the protocols and concentrations vary between the two labs.

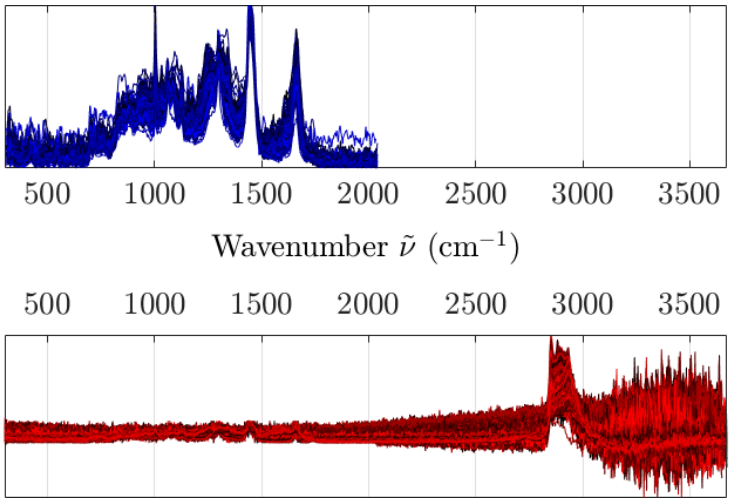


Figure 1. Spectra of the dataset divided into two main groups by the wavenumber range. The top graph shows the spectra from Sorbonne, with a wavenumber range of 307-2041cm^{-1} and the bottom graph shows the spectra from Twente, with a wavenumber range of 300-3674cm^{-1}

Machine learning methodology

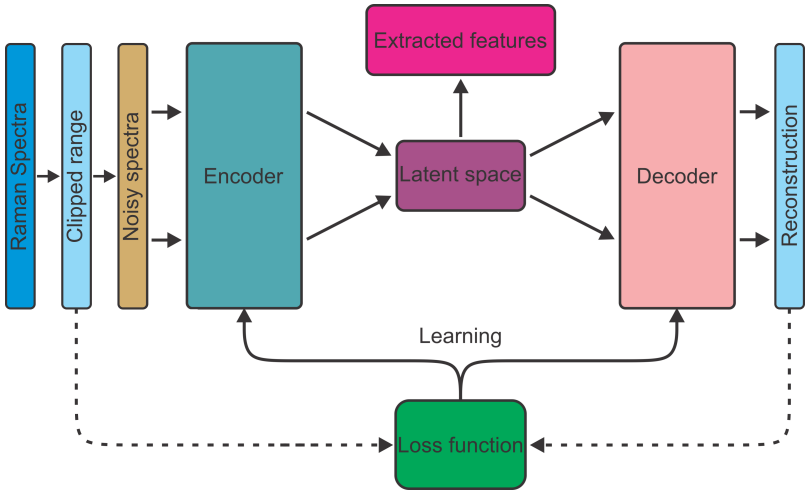


Figure 2. Schematic of the autoencoder. The encoder takes the vector of noisy and / or clipped Raman spectra as input which it processes into a compressed representation in the latent space containing the extracted features of the spectra. The decoder then takes the latent representations of the spectra as input and attempts to reconstruct the original Raman spectra using this information. Learning occurs by computing a loss between the original Raman spectra and the reconstructed spectra and passing the gradient to the encoder and decoder.

The desired outcome of the self-supervised deep neural network is to be able to extract chemically relevant information from the Raman spectra of sample particles, and for the quality of that information to be sufficient to reliably group particles

by their chemical nature. To achieve this, we apply an autoencoder[27,31] to the Raman spectra, as shown in fig. 2 and 3. The goal of the encoder is to learn to extract the most valuable information from the spectra and pass this on to the latent space, which would ideally be a representation of the chemical nature of the sample. The quality of this information is verified by the decoder, whose purpose is to reconstruct the original spectra using only the information available in the latent space. If the quality and completeness of the information passed to the latent space by the encoder is high, then the decoder should be able to reconstruct the spectra with high accuracy.

In order to focus on the important information, spectra are first clipped at random positions with the beginning and/or end removed. This forces the network to be robust to input with different spectral range. Secondly, noise of different amplitude is added to the clipped spectra. The loop for the self-supervised phase 1 learning is achieved by computing a loss, quantifying the amount of lost information, between the original (possibly clipped) spectra and the reconstructions produced by the decoder. Once this phase of learning is complete, the autoencoder should have learnt to reject noise and be robust to clipping; whatever spectra is passed to the encoder should produce an approximately equivalent representation in the latent space, containing high quality chemically related information. Phase 2 learning can then commence by interfacing the classifier with the latent space of the autoencoder and passing the training data through the encoder such that a training set is generated in the latent space. The classifier then relies on this training set and the known origins of the EVs as labels for its learning.

Variational autoencoder

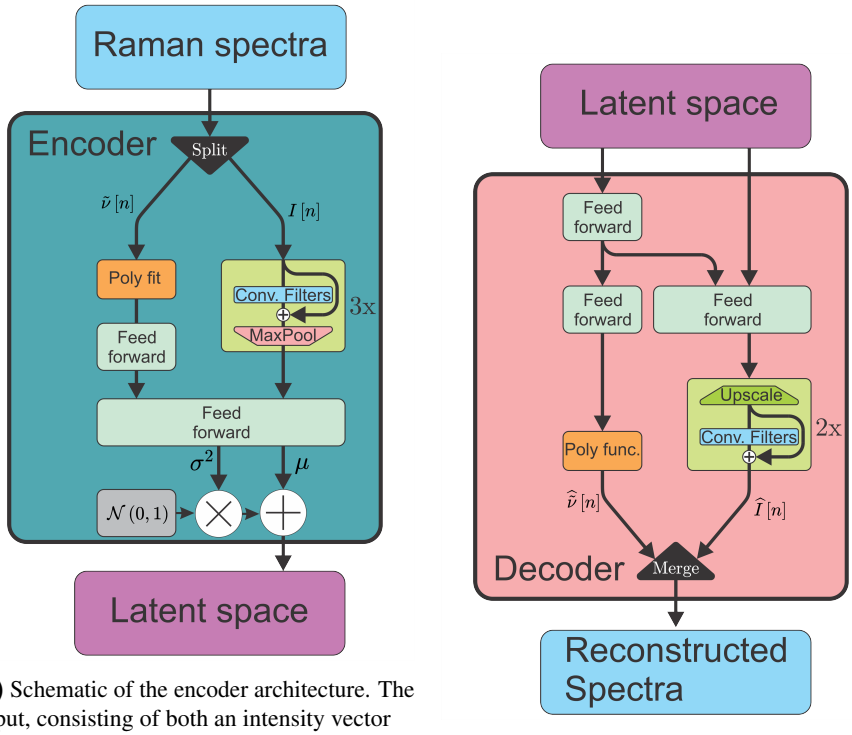
Since the Raman spectra are naturally continuous along the frequency axis, we base our autoencoder architecture on convolutional layers. To facilitate better learning of a deep model, we also implement skip connections across the convolutional layers, akin to those introduced in the ResNet architecture[32], which allows the gradient to bypass the layers during learning. We also implement Gaussian re-sampling as is used in variational autoencoders[33] to encourage the learning of uncorrelated features in the latent space.

As the goal of the autoencoder is to learn features in the Raman spectra that correspond to chemical aspects in the sample, we wish to make the information regarding one chemical group to be represented in a single dimension of the latent space. By using a variational autoencoder, which attempts to learn the latent features as uncorrelated gaussian variables, we motivate the model to express the observed chemical features such that they are as independent as possible. Our hypothesis is that it motivates the network to express patterns that correlate strongly, such as the CH2 peaks and lipid complexes, in one latent space dimension and other spectral signatures that correlate weakly, such as the amino and carotenoid complexes, into separate dimensions. The end goal of this is to make the learned representations of a single chemical be expressed in a single latent dimension instead of being spread out over multiple dimensions. That way, we hypothesise that the latent space will form a more realistic, and thus valuable, map of the chemical features in the sample, and where the different classes of EVs can be easily separated.

Adaptive frequency range

A variational autoencoder based on convolution is very capable of processing data such as Raman spectra when expressed as a numerical vector of intensities, but it will lack essential information about the frequency range. When interpreting the Raman spectra of a sample, be it by a human or by a machine, identifying both the peak intensity and its position in the spectrum are crucial. Since our data arises from multiple sites with different frequency ranges, if only the shape of the spectrum is made available to the model, then it is poorly equipped for inferring purposeful information from the spectrum. Conventionally, this information is passed to the model *implicitly* by pre-processing the dataset such that each index of the spectrum vector corresponds to the same wavenumber shift. This way, the model learns to associate a peak with an underlying component (i.e. a chemical) by where the peak appears in the vector rather than its wavenumber shift. In our context, this is not possible.

Instead, we propose making information regarding the frequency range available to the model such that it can be considered *explicitly*. This is implemented in the encoder as shown in fig. 3a by passing both the intensity and frequency vectors to the model on two separate channels, and allowing them to pass through two information pipelines through the model. The intensity vector passes through convolutional filters that extract the spectral features while the frequency vector is fitted by a fourth degree polynomial. The parameters of the frequency polynomial are then pre-processed by a feed forward block before being concatenated with the feature information and processed together in the main feed forward block of the encoder. Similarly as shown in fig. 3b, the decoder extracts the frequency related information from the last ten dimensions of the latent space, processes it through two feed forward blocks, and reconstructs the frequency range using a fourth degree polynomial whose parameters are given by the feed forward block. The intensity vector is reconstructed by passing the frequency and intensity related information from the latent space to a feed forward block whose output goes through a series of up-sampling convolution filters. The reconstructed frequency and intensity vectors are then concatenated to produce an output of the same format as the input. The final architecture has a total depth of 36 layers including the adaptive frequency neurons, making the model quite deep. Further details on the architecture are available in Table S2.



(a) Schematic of the encoder architecture. The input, consisting of both an intensity vector $I[n]$ and a wavenumber vector $\tilde{\nu}[n]$, is split to follow two paths. The two paths separate the frequency (x) and intensity (y) axes of the spectra. They only meet in the main feed forward networks of the encoder and the decoder, as well as the latent space.

(b) Schematic of the decoder architecture. The latent space data is again split to follow two paths. The frequency vector is reconstructed by a polynomial whose parameters are given by feed forward layers and the intensity vector is reconstructed by convolutional filters.

Figure 3. Schematic of the encoder and decoder of the autoencoder architecture.

Loss function

During training, the learning outcome of the autoencoder is to reconstruct the data as accurately as possible. Phase 1 learning must therefore use a loss function whose metrics reflect this and motivates the learning to extract chemical information from the spectra. To achieve this, Fourier loss is added to the more standard spatial and Kullback-Leibler loss. **The spatial loss function** describes the difference between the clipped original spectra and the reconstruction in the most direct manner and serves as the most fundamental loss function for the autoencoder during phase 1 learning. The metric of this loss is the root mean square error between the original and the reconstruction. **The Kullback-Leibler loss function** describe the difference between the discrete distributions in the latent space and a gaussian distribution. The need for this loss arises from the gaussian re-sampling in the last layer of the encoder and is described by the Kullback-Leibler divergence between the output of the last feed forward layer in the encoder, which produces the means μ and variances σ^2 for the re-sampling, and a a gaussian distribution with a mean of zero and a variance of 1. Thus, the more the latent representations approximate uncorrelated gaussian distributed variables, the lower the KL loss will become. **The Fourier loss function** describes the difference between the original and the reconstruction in Fourier space. The metric of this loss function is the sum of the mean square difference of the frequency and phase between the input and the reconstruction. The reason for including this loss is to counter unwanted effects of the simple spatial loss function, namely that the spatial loss is most sensitive to low-frequency features[34] such as large, smooth slopes. This effect encourage the autoencoder to consider the low-frequency elements more strongly, thus implicitly learning to low-pass filter the input[35]. By adding a Fourier loss, the autoencoder is forced to consider the higher frequency elements as well and, by adding a mask that attenuates frequencies outside the relevant bandwidth, encourage it to learn to preserve high-frequency elements of the spectra, such as sharp peaks.

The total loss function is a composite loss formed by these losses to form a total function :

$$L_\Sigma = (1-\alpha)\cdot(L_{RMS,I} + \gamma L_{RMS,\tilde{\nu}} + \beta L_{KL}) + \alpha L_{Fourier},$$

where $L_{RMS,I}$ is the RMS difference between the original and reconstructed intensity vectors, $L_{RMS,\tilde{\nu}}$ is the RMS difference

between the original and reconstructed frequency vectors, L_{KL} is the Kullback-Leibler loss, and $L_{Fourier}$ is the Fourier loss. The balancing parameters α, β and γ are used to moderate the relative strength of the losses. The parameter α balances the ratio of the Fourier loss relative to the other losses, this is set to 0.3 during training to make the spatial and KL losses slightly dominant over the Fourier loss. The parameter β determines the strength of the KL loss and is set to 5 to ensure the KL loss is significant during learning, but not strong enough to be clearly dominant. The parameter γ is used to balance the strength of the loss of the intensity and frequency reconstructions, to compensate for the lower order of magnitude in the frequency vector compared to the intensity vector.

Classifier head

The classifier head exists outside the autoencoder and only passively interacts with it through observing the latent space. This performs classification of the spectra through a conventional feed forward-approach where the latent space is passed to a series of feed forward networks with dropout, batch normalization, and skip connections. Due to the high level of pre-processing by the encoder, this network can be made relatively small, consisting of five layers of 128 neurons only. The output of this processing is passed to a classification head consisting of a feed forward layer with a softmax activation which produces a one-hot encoded class prediction.

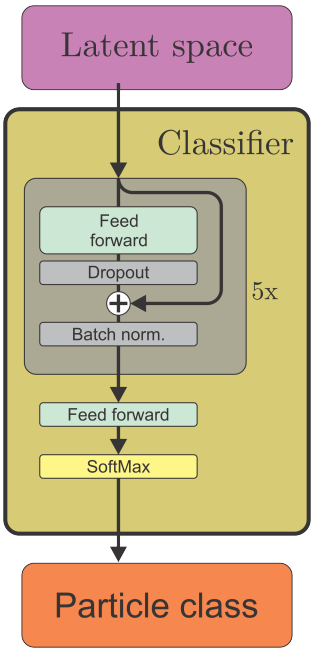


Figure 4. Schematic of the classifier head.

Training scheme

Artificial noise in training

As shown in fig. 1, the available spectra forms two distinct datasets, one from Paris and one from Twente. The most distinct differences between these two datasets are: the difference in signal-to-noise ratio, the difference in frequency range, and the difference in calibration drifts. For the autoencoder to function as intended, it must be trained to be robust against these differences, and thus the presence of these factors and their magnitudes must be more homogeneous in the training set. To achieve this, we augment the training set with three types of noise to emulate these conditions: noise in the intensity to emulate variations in the signal-to-noise ratio, clipping of the frequency range to emulate different acquisition ranges, and distortions in the frequency axis to emulate calibration drift.

The intensity noise is implemented as additative gaussian noise in the intensity of the spectra. The magnitude of this noise is calculated from the root mean square of the intensity before noise such that the variance of the gaussian noise is a prescribed amount of the RMS of the intensity. In training, the variance is decided to be equal to -5dB of the RMS of the intensity, calculated on a per-spectrum basis to make the significance of the noise more equal between the spectra. **The range clipping** is implemented by selecting a random start and stop for the range then removing the parts of the spectrum that is outside this range. The start of the new range is determined by random selection from the beginning of each spectrum to a prescribed

maximum, which is set to 800cm^{-1}. The stop of the new range is determined similarly as a random point between a prescribed minimum and the end of each spectrum, the minimum is set to 1500cm^{-1} in training. This results in up to 79 % of the spectrum being clipped. **The frequency noise** is implemented by first fitting a fourth degree polynomial to the frequency vector: $P_4[I,;\beta] = \beta_4 I^4 + \beta_3 I^3 + \beta_2 I^2 + \beta_1 I + \beta_0 \approx \tilde{\nu}[I]$. For each of the parameters β_n, a random amount of distortion $\hat{\beta}_n$ equivalent to 2% of β_n is selected to create a distorting polynomial $\hat{P}[I,\hat{\beta}]$. The distorting polynomial is then added to the frequency vector to create distortions in the vector that emulate poor calibration of the spectrometer: $\tilde{\nu}_N[I] = \tilde{\nu}[I] + \hat{P}[I,\hat{\beta}]$.

A last addendum to the training set is a set of randomly selected spectra whose intensities are purely gaussian noise. These are introduced to the training set to force the model to recognize a zero-signal condition, and thus discourage it from considering the spectra as variations on a "mean" spectrum but rather as intrinsically unique and express them as more unique in the latent space. Experiments have shown that this is beneficial (not reported here).

Training phases

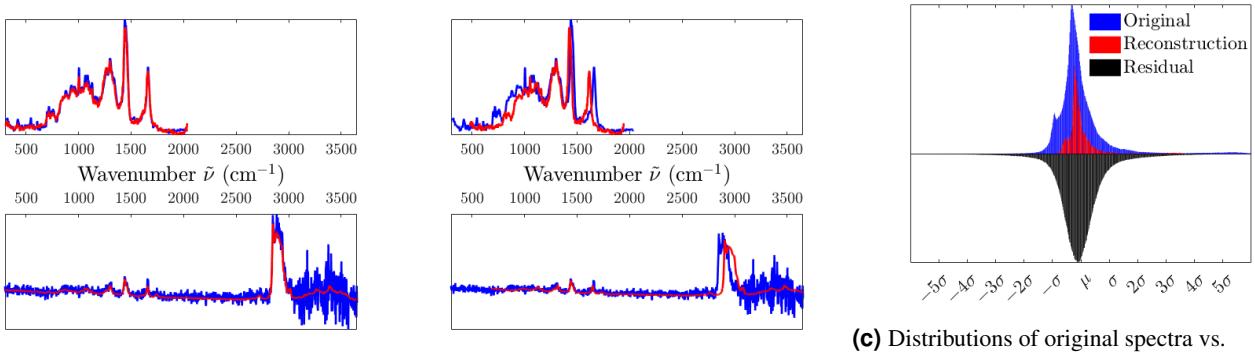
The raw spectra are first separated into the training set (70%, 1898 spectra) and the testing set (30%, 769 spectra) which are then isolated from each other in the datastructure. During phase 1 training, a new noisy set is generated from the raw training set for each epoch of training to encourage the model to be robust against noise. These training sets are generated by adding together one copy of the raw training set, three copies with the described noise types, and one copy of noise-only signals. The generated training set is then randomly shuffled before being passed to the model for training.

The model is then built with the determined achitecture and allowed to train on the generated training sets for 100 epochs with a learning rate of 10^{-4} before concluding phase 1 of the learning.

Once trained, the encoder of the model is used to generate latent representations of the training data which is then used to train the classifier. The classifier is fed the latent training set and tasked with classifying the data into 13 classes corresponding to the known particle origins which are used as labels.

Results

Reconstruction performance



(a) Reconstruction of intensity vector only by model.

(b) Reconstruction of both intensity and frequency vectors by model.

(c) Distributions of original spectra vs. reconstructed spectra, and the residual between them.

Figure 5. Reconstruction performance of the model, showing the input in blue and the reconstruction in red. The model is tasked with reconstructing the same data in both plots, but only the intensity vector in a) and both frequency and intensity vectors in b). The histograms in c) show the statistical distribution of the original spectra and the reconstructed spectra as well as the distribution of their residual difference.

The first measure of the quality of the model is the accuracy of the reconstructions it produces. The trained network is provided with unaltered spectra from the test set and asked to reconstruct them, producing the results shown in fig. 5. The models ability to reconstruct the spectra is excellent, despite the high levels of noise and their variability. Note the significant reduction in noise in both cases and the preservation of sharp features, such as the phenylalanin peak at 1003.6 cm^{-1}. Fig. 5b shows that the reconstruction of the frequency vector is not perfect, producing an elongation or compression of the spectra. However, the reconstruction is adapting to changes in the range, approximates it well, and, most importantly, the shapes of the features are preserved. By comparison, truncating and interpolating the data such that the adaptation is not required removes a significant part of the spectra, see fig. S1.

The most notable difference between the original spectra and the reconstructions is the significant noise reduction, especially for high wavenumbers (> 3000/cm). The autoencoder is trained on data with artificial noise and has learned to remove this

noise. Therefore, when presented with spectra that only have the natural noise, such as those in the test set, it will attempt to remove the natural noise in the same way it learned to remove the artificial noise. The performance of this process can be evaluated by investigating the difference between the original and the reconstructed spectra. The distribution of this residual is shown in fig.5c. In the case of perfect noise removal, the residual should consist of noise, which is presumed to have a Gaussian distribution. The Kullback-Leibler divergence between the distribution of the residual and a Gaussian distribution reveals a normalized divergence 3.3 times lower than for the original spectra, indicating that the residual is largely gaussian and thus largely noise.

Feature extraction capabilities

The high reconstruction accuracy of the model indicates that the information preserved in the latent space is of high quality and that noise is not preserved. The next step is to evaluate the relevance and significance of the information as a means to characterize the sample. This is done by passing the test data through the encoder of the network and extracting the predicted means of the latent space distributions such that the "perception" of the encoder can be investigated.

By using t-distributed stochastic neighbour embedding (t-SNE)[36], we project the 100 dimensional data in the latent space down to two dimensions and use the known origins of the data as labels to illustrate the latent space representations as shown in fig. 6. The t-SNE plots are used to reveal the structure of the latent space and show which samples are grouped together, both in latent space and in the t-SNE plots. The principle of t-SNE is to reduce the dimensionality while keeping points that are close in the high dimensional space close in the low dimensional one. Points in a t-SNE plot do not reflect where they are in the latent space, but their proximity to other points reflects their proximity in the latent space. Note that points in the t-SNE plot may appear more clustered than they are in reality[37]. Hence the clusters in the next figures mean that points are located close to each other in the latent space but they may not be well isolated from the others.

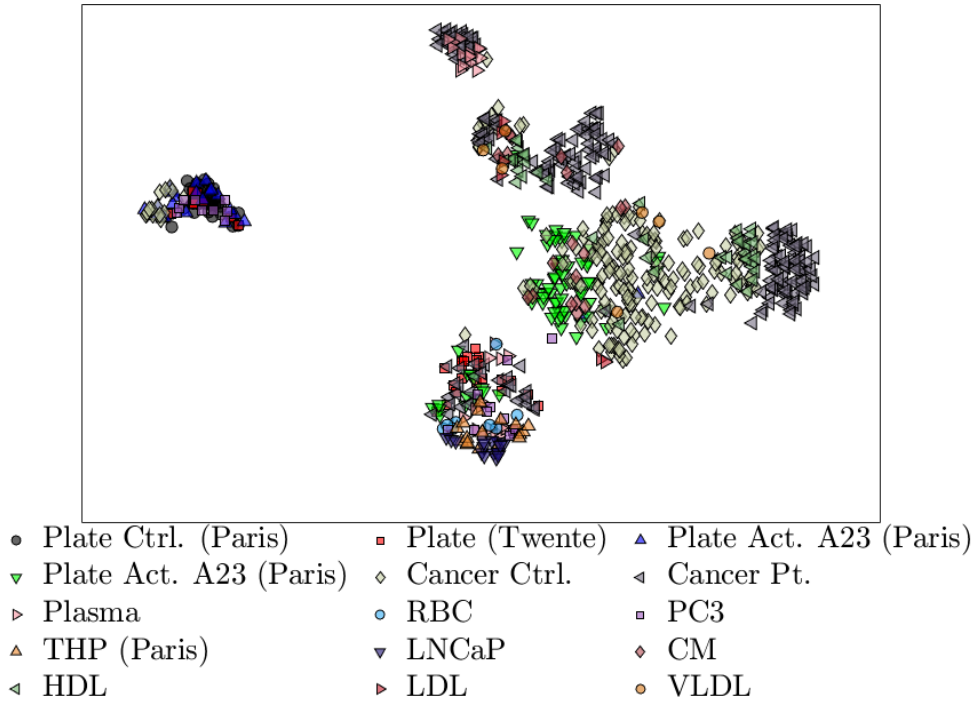
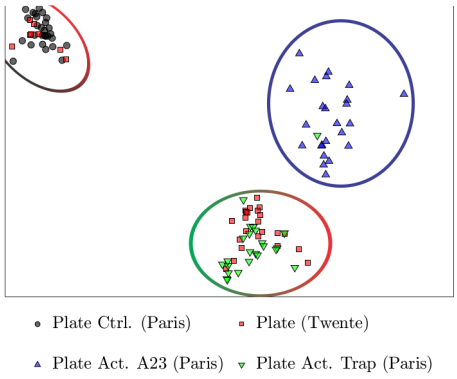
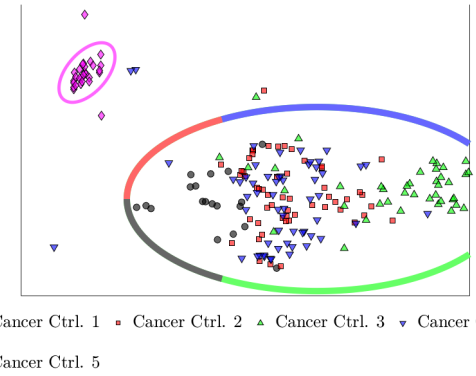


Figure 6. Two-dimensional t-SNE projection of the latent representations of spectra from the test set. This is a general overview of the latent space. Six clusters of samples can be seen, showing that the network is able to "see" similarities in the data. However, the main results are difficult to discern from this global view. In the next figure, subsets of the data are shown.

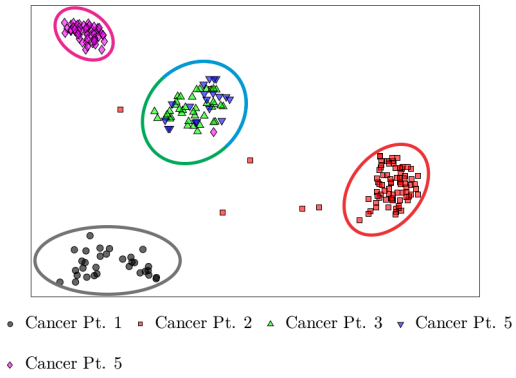
The plot shows six clusters, indicating that the model sees varying degrees of similarity and dissimilarity between the particles. Due to the large number of datapoints and labels, it is difficult to discern them in fig. 6. For better clarity, a few labels have been selected and plotted in fig. 7. Some more selections can be found in fig. S3. Note especially, in fig. 7a, the partial overlap between some of the platelet particles from Twente with the control and Trap-activated platelet particles from Paris. This indicates that the model recognizes them as similar in spite of the difference in the condition of their measurements. Furthermore for the Paris dataset, EVs from activated platelets are clearly separated from controls. Regarding the controls 1 to 4 in fig. 7b, they are well mixed in one cluster, while control 5 is separately clustered. In contrast, the five prostate cancer



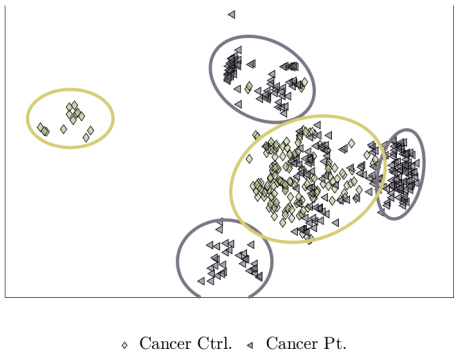
(a) t-SNE projection for platelet derived EVs. The TRAP6-activated platelets (green) form a common group with some platelet particles from Twente (red). Some of the latter also overlap with the platelet control EVs from Paris (black). The A23 (calcium, black)) activated platelets form a distinct group.



(b) t-SNE projection for nanoparticles from (cancer) controls. Controls 1-4 form a common wide group (multi-color) while the EVs from Control 5 forms a smaller, distinct group.



(c) t-SNE projection for nanoparticles from prostate cancer patients. Patients 3 and 4 form a common group (green/blue), while the remaining patients each form separate groups.



(d) t-SNE projection for nanoparticles from prostate cancer patients vs. controls. With the exception of Control 5, the controls form one group near the center. The prostate cancer patients form separate clusters surrounding the control cluster.

Figure 7. Projections of the latent space of the autoencoder with true origin labeling. The 100 dimensions of the latent space are squeezed down to 2 by a t-SNE model and plotted using the known origins of the particles. It is demonstrated that the self-supervised clustering of the autoencoder agrees well with the true particle types, and can recognize particles as the same regardless of the lab they were measured in.

patients in fig. 7c mostly form separate clusters, with the exception of patients 3 and 4 which share a cluster. Lastly, fig. 7d shows that the cancer patients and the controls are mostly well separated from each other. The underlying data show that the partial overlap of cancer patients is with controls 1-4, while control 5 is in a separate cluster, as in fig. 7b.

Classification accuracy

After the classifier has been trained on the latent representation of the training set produced by the encoder, the test set is passed through the encoder and to the classifier. The classifier is trained to recognize the same origins as shown by the labels in fig. 6 except for the merging of the platelet particles from Twente and platelet control particles from Paris into one label. The resulting confusion matrix of the classifications are shown in fig. 8, with supplementary matrices with percentages and for ICA in Fig. S4. As an example from the matrix, it shows that the cancer patients can be perfectly differentiated from non-cancer controls by the model. In cases of EVs from bulk blood, the model is also able to differentiate blood plasma and red blood cells with a high degree of accuracy, producing no misclassification between the two. Over a total of 769 spectra in the test set 709 are correctly classified, yielding a true positive rate (sensitivity) of 92.2% and true negative rate (selectivity) of 92.3%. Note that the highest rate of mutual misclassification is between the control platelets and plasma derived particles. A likely explanation for this is

that a large portion of the particles measured in the platelet test were determined to be lipoproteins[38], resulting in the spectra more closely resembling those of plasma, see fig. S3d). The noise robustness of the classifier is also tested by introducing the same clipping to the test data as for the training data, resulting in a sensitivity of 84.7% and a selectivity of 86.2% for clipped data. Further degradation of the data by noise or frequency distortions similar to the training case resulted in the classifier maintaining a sensitivity of 81.5% and a selectivity of 83.8% or higher. The resulting confusion matrices are shown in fig. S4.

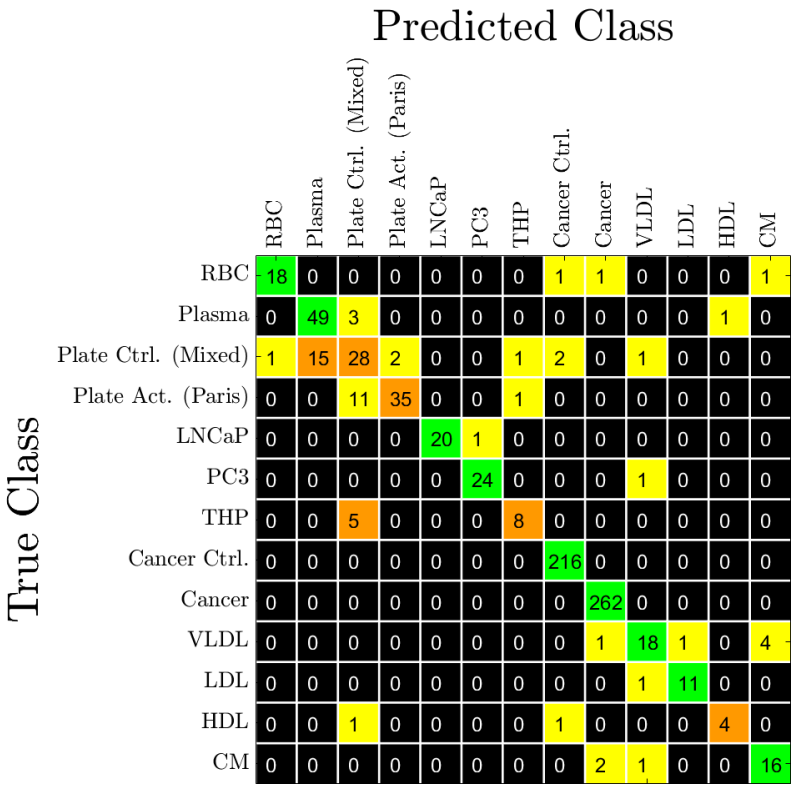


Figure 8. Confusion matrix for the classifier trained on the latent space representations of the spectra. The colors indicates the percentage correct predictions, with yellow for 1% to 25% correct, orange for 25% to 50% correct, and green for more than 75% correct prediction. The numbers show the number of analysed spectra for each case.

Discussion

In this work we have constructed a novel that is able to reconstruct the Raman spectra of single (or near-to-single) EV with a high degree of accuracy and with strong de-noising properties. The model is shown have to high level of adaptability as it can fit well data from different origins, with differences in frequency range and noise levels. The extracted features are of high quality, enabling to classify and cluster EVs and other biological nanoparticles based on their cellular origin and activation state, as well as an overall condition (health/disease). This is verified by the classifier, which attains more than 90 % accuracy in classifying the nanoparticles to their correct origins with high sensitivity and selectivity.

In the reconstructions shown in fig. 5, we see that the autoencoder perceives all of the significant features and reconstructs them well; both the low-frequency slopes and complexes, as well as the high-frequency peaks. From the frequency range reconstructions shown in fig. 5b we see that the reconstructed frequency vectors are reconstructed well, but with some degree of error. This error can readily be attributed to the weighting of the learning process, where the intensity vector, containing the actual Raman features, is given more importance in reconstruction. Nevertheless, the adaptive behavior of the reconstructed frequency vector clearly shows that the model actively considers it in its determination of the features in the spectrum and that the model can adapt to data with very different frequency ranges from different labs.

Figs. 6 and 7 show that the extracted features form natural clusters that correspond well with the samples that produced the spectra. The most notable of these results is shown in fig. 7a. For the dataset from Paris, EVs from un-activated control platelets are well differentiated from the EVs from activated platelets, and the model recognizes a difference between activation with Thrombin (Trap) or Calcium (A23). For the dataset from Twente, some platelet particles overlap with the controls from

Paris, while some overlap with those activated with Thrombin from Paris. The partial overlap for platelet particles from Twente and platelet control particles from Paris, despite their Raman spectra having very different noise levels and frequency ranges, illustrates the achieved strength and flexibility of the model for variations in the measurement method. The separate clusters for EVs from platelets activated with Thrombin (Trap) and with Calcium (A23), both Paris, further emphasizes the models ability to differentiate EVs. The proximity of the Trap-activated platelets and platelets from Twente to the plasma particles also indicates that the model sees them as similar. This may be due to a higher prevalence of protein in these particles, making them more similar to the plasma derived particles which are mainly lipoproteins. This can be explained by the fact that the platelets used to generate this data comes from stored platelet concentrate[38] while the Paris platelets come from fresh blood, which can influence the composition of the platelets and their derived nanoparticles. Regarding fig. 7b, four of the (non-cancer) controls are clustered together, indicating that the model can see them as similar despite originating from different controls. But one control forms a tight, separate cluster, indicating that the model sees a difference relative to controls 1-4. As the medical record and possible medication of the controls is not known, this difference cannot be further investigated.

Previous studies have shown that EVs from different cell lines and primary cells carry different biochemical features, as analyzed by Raman spectroscopy and PCA[19,39]. Such analysis, however, did not generate a clear distinction, as shown in fig. S2. It also could not blindly associate features extracted from Raman spectra of EVs with their cells of origin. The features extracted from the Raman spectra of EVs by the autoencoder enable classification according to the cellular origin, as well as between cancer patients and subjects who do not have cancer. Such classification opens new frontiers and possible uses of EVs for the diagnosis and prediction of disease. Of particular note is the specific identification of EVs from activated platelets, which were clearly separated from EVs isolated from resting platelets, monocytes or erythrocytes. The exclusivity of the features of nanoparticles from platelets was further demonstrated by grouping nanoparticles from platelets generated in two independent labs, each using different instruments and operators. Platelet activation and subsequent release of EVs is thought to play a role in various thrombotic conditions, such as venous thromboembolism (VTE)[40]. Therefore, it would be interesting to investigate whether our autoencoder and classifier can detect elevated levels of EVs from activated platelets in the plasma of VTE patients and whether this could facilitate future predictions or diagnoses of the disease. Our results also warrant further investigation in the field of cancer, where patients exhibit a distinct EV profile.

The quality of the information extracted from the spectra by the model is demonstrated by the achieved accuracy of the classifier. By taking the extracted features given by the latent space of the autoencoder, the classifier achieves both a sensitivity and selectivity of over 90% across the test set. Most notable is the fact that there is no overlap between the cancer and non-cancer derived particles, meaning that the model is capable of detecting prostate cancer with perfect sensitivity and selectivity. This indicates that the model perceives a significant difference between the EVs from cancer compared to non-cancer.The classifier also gives an 82.3% accuracy in classifying platelet derived EVs as activated or unactivated, with a sensitivity of 76.1% and selectivity of 93.3% for detection of activated platelets. The lower sensitivity of this test indicates that the model sees a relatively small difference between the activated and un-activated platelet derived EVs, which is not unexpected given their intrinsic similarity due to them having a common origin. This is also illustrated in fig. 7a by the partial overlap between the platelet particles from Twente and un-activated platelets from Paris, as discussed above. There is also a significant overlap between the platelets from Twente and the plasma derived particles, resulting in a sensitivity of only 65.1% for detecting control platelets from plasma.

Conclusion

We have demonstrated that an autoencoder, with a depth of 21 convolutional layers and eight fully connected layers, can learn to reconstruct a wide variety of Raman spectra with a highly variable signal-to-noise ratio and with a variable frequency range. The ability to consider spectra from separate measurement systems, while maintaining high reconstruction accuracy underscores the capabilities of the model. The de-noising performance of the model is also shown to be promising, leaving a residual difference between the spectra and reconstructions that follows a gaussian distribution, indicating that the residual is largely random noise that is filtered out by the model. The model is also shown to be capable of extracting valuable, chemically related information from the spectra, which allows it to perform label-free clustering of particles by their similarity despite the significant differences in the measurement methods.

It is shown that the model is capable of recognizing activated platelet derived EVs and recognize mixed activated platelet derived particles from both the lab in Paris and in Twente as similar, while also recognizing that EVs from un-activated platelets, and platelets activated by artificial calcium activators, are different from EVs activated using thrombin. This demonstrates both the viability of using Raman spectroscopy as a means of detecting platelet activation and the viability of using the demonstrated model to reliably extract valuable information from those spectra. For prostate cancer derived EVs, individual patients are mostly separated, while the EVs from controls are mostly clustered together. The reason for this cannot be further investigating without more data and knowing the medical conditions of the patients and the controls. The model readily differentiates the cancer patients from the non-cancer controls, with a perfect sensitivity and selectivity, demonstrating the viability of Raman

spectroscopy and the model. In cases of bulk blood, the model is also able to differentiate blood plasma and red blood cells with a high degree of accuracy, producing no misclassification between the two.

References

1. Lee, W., Lenferink, A. T. M., Otto, C. & Offerhaus, H. L. Classifying raman spectra of extracellular vesicles based on convolutional neural networks for prostate cancer detection. *J. Raman spectroscopy* **51**, 293–300, DOI: 10.1002/jrs.5770 (2020).
2. Xiao, Y. *et al.* Extracellular vesicles in type 2 diabetes mellitus: key roles in pathogenesis, complications, and therapy. *J. extracellular vesicles* **8**, 1625677 (2019).
3. Arakelyan, A., Fitzgerald, W., Zicari, S., Vanpouille, C. & Margolis, L. Extracellular vesicles carry hiv env and facilitate hiv infection of human lymphoid tissue. *Sci. reports* **7**, 1695 (2017).
4. Upadhya, R. & Shetty, A. K. Extracellular vesicles for the diagnosis and treatment of parkinson's disease. *Aging disease* **12**, 1438 (2021).
5. Cheng, L. & Hill, A. F. Therapeutically harnessing extracellular vesicles. *Nat. Rev. Drug Discov.* **21**, 379–399 (2022).
6. Ciferri, M. C., Quarto, R. & Tasso, R. Extracellular vesicles as biomarkers and therapeutic tools: From pre-clinical to clinical applications. *Biology* **10**, 359 (2021).
7. Dragovic, R. A. *et al.* Sizing and phenotyping of cellular vesicles using nanoparticle tracking analysis. *Nanomedicine: Nanotechnology, Biol. Medicine* **7**, 780–788 (2011).
8. Chandler, W. L. Measurement of microvesicle levels in human blood using flow cytometry. *Cytom. Part B: Clin. Cytom.* **90**, 326–336 (2016).
9. Edit I. Buzás, C. L., Chris Gardiner & Smith, Z. J. Single particle analysis: Methods for detection of platelet extracellular vesicles in suspension (excluding flow cytometry). *Platelets* **28**, 249–255, DOI: 10.1080/09537104.2016.1260704 (2017). PMID: 28033028, https://doi.org/10.1080/09537104.2016.1260704.
10. Xu, Y. *et al.* Raman spectroscopy coupled with chemometrics for food authentication: A review. *TrAC-trends Anal. Chem.* **131**, DOI: 10.1016/j.trac.2020.116017 (2020).
11. Hess, C. New advances in using raman spectroscopy for the characterization of catalysts and catalytic reactions. *Chem. Soc. Rev.* **50**, 3519–3564, DOI: 10.1039/d0cs01059f (2021).
12. Kruglik, S. G. *et al.* Raman tweezers microspectroscopy of circa 100 nm extracellular vesicles. *Nanoscale* **11**, 1661–1679, DOI: 10.1039/c8nr04677h (2019).
13. Mrad, R., Kruglik, S. G., Ben Brahim, N., Ben Chaabane, R. & Negrerie, M. Raman tweezers microspectroscopy of functionalized 4.2 nm diameter cdse nanocrystals in water reveals changed ligand vibrational modes by a metal cation. *J. Phys. Chem. C* **123**, 24912–24918, DOI: 10.1021/acs.jpcc.9b06756 (2019).
14. Zhou, X.-X., Liu, R., Hao, L.-T. & Liu, J.-F. Identification of polystyrene nanoplastics using surface enhanced raman spectroscopy. *Talanta* **221**, DOI: 10.1016/j.talanta.2020.121552 (2021).
15. Dong, M. *et al.* Raman spectra and surface changes of microplastics weathered under natural environments. *Sci. Of The Total. Environ.* **739**, DOI: 10.1016/j.scitotenv.2020.139990 (2020).
16. Dina, N. E. *et al.* Sers-based antibiotic susceptibility testing: Towards point-of-care clinical diagnosis. *Biosens. & Bioelectron.* **219**, DOI: 10.1016/j.bios.2022.114843 (2023).
17. Yaraki, M. T., Tukova, A. & Wang, Y. Emerging sers biosensors for the analysis of cells and extracellular vesicles. *Nanoscale* **14**, 15242–15268, DOI: 10.1039/d2nr03005e (2022).
18. Li, Q. *et al.* Design and synthesis of sers materials for in vivo molecular imaging and biosensing. *Adv. Sci.* DOI: 10.1002/advs.202202051 (2023).
19. Lee, W. *et al.* Label-free prostate cancer detection by characterization of extracellular vesicles using raman spectroscopy. *Anal. chemistry* **90**, 11290–11296 (2018).
20. Kothari, R. *et al.* Raman spectroscopy and artificial intelligence to predict the bayesian probability of breast cancer. *Sci. reports* **11**, 6482 (2021).
21. Gualerzi, A. *et al.* Raman spectroscopy as a quick tool to assess purity of extracellular vesicle preparations and predict their functionality. *J. Extracell. Vesicles* **8**, 1568780, DOI: 10.1080/20013078.2019.1568780 (2019). PMID: 30728924, https://doi.org/10.1080/20013078.2019.1568780.

22. Yin, P. *et al.* Facile peg-based isolation and classification of cancer extracellular vesicles and particles with label-free surface-enhanced raman scattering and pattern recognition algorithm. *Analyst* **146**, 1949–1955 (2021).

23. E.M. Guerreiro, S. S. N. L. B. J. G. F. S. S. B. A. N. K. P. N. P. J. H. O. G. H., S. G. Kruglik & Snir, O. Extracellular vesicles from activated platelets possess a phospholipid-rich biomolecular profile and enhance prothrombinase activity. *submitted* **TBA**, TBA (2023).

24. Ghosh, K. *et al.* Deep learning spectroscopy: Neural networks for molecular excitation spectra. *Adv. Sci.* **6**, DOI: 10.1002/advs.201801367 (2019).

25. Fan, X., Ming, W., Zeng, H., Zhang, Z. & Lu, H. Deep learning-based component identification for the raman spectra of mixtures. *Analyst* **144**, 1789–1798, DOI: 10.1039/c8an02212g (2019).

26. Akbarimajd, A. *et al.* Learning-to-augment incorporated noise-robust deep cnn for detection of covid-19 in noisy x-ray images. *J. Comput. Sci.* **63**, 101763 (2022).

27. Goodfellow, I., Bengio, Y. & Courville, A. *Deep learning* (MIT press, 2016).

28. Balestriero, R. *et al.* A cookbook of self-supervised learning. *arXiv preprint arXiv:2304.12210* (2023).

29. Lee, J. D., Lei, Q., Saunshi, N. & ZHUO, J. Predicting what you already know helps: Provable self-supervised learning. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. & Vaughan, J. W. (eds.) *Advances in Neural Information Processing Systems*, vol. 34, 309–323 (Curran Associates, Inc., 2021).

30. Enciso-Martinez, A. *et al.* Synchronized rayleigh and raman scattering for the characterization of single optically trapped extracellular vesicles. *Nanomedicine-Nanotechnology Biol. And Medicine* **24**, DOI: 10.1016/j.nano.2019.102109 (2020).

31. Kingma, D. P., Welling, M. *et al.* An introduction to variational autoencoders. *Foundations Trends Mach. Learn.* **12**, 307–392 (2019).

32. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016).

33. Rezende, D. J., Mohamed, S. & Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In Xing, E. & Jebara, T. (eds.) *International Conference On Machine Learning, Vol 32 (cycle 2)*, vol. 32 of *Proceedings of Machine Learning Research*, 1278–1286 (2014). International Conference on Machine Learning, Bejing, PEOPLES R CHINA, JUN 22-24, 2014.

34. Rahaman, N. *et al.* On the spectral bias of neural networks. In Chaudhuri, K. & Salakhutdinov, R. (eds.) *Proceedings of the 36th International Conference on Machine Learning*, vol. 97 of *Proceedings of Machine Learning Research*, 5301–5310 (PMLR, 2019).

35. Björk, S., Myhre, J. N. & Haugland Johansen, T. Simpler is better: Spectral regularization and up-sampling techniques for variational autoencoders. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3778–3782, DOI: 10.1109/ICASSP43922.2022.9746027 (2022).

36. Van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *J. machine learning research* **9** (2008).

37. Wattenberg, M., Viégas, F. & Johnson, I. How to use t-sne effectively. *Distill* DOI: 10.23915/distill.00002 (2016).

38. Enciso-Martinez, A. *et al.* Label-free identification and chemical characterisation of single extracellular vesicles and lipoproteins by synchronous rayleigh and raman scattering. *J. Extracell. Vesicles* **9**, 1730134, DOI: https://doi.org/10.1080/20013078.2020.1730134 (2020). https://onlinelibrary.wiley.com/doi/pdf/10.1080/20013078.2020.1730134.

39. Penders, J. *et al.* Single particle automated raman trapping analysis of breast cancer cell-derived extracellular vesicles as cancer biomarkers. *ACS nano* **15**, 18192–18205 (2021).

40. Snir, O. *et al.* Plasma levels of platelet-derived microvesicles are associated with risk of future venous thromboembolism. *J. Thromb. Haemostasis* **20**, 899–908 (2022).

Acknowledgements

The project was funded by the Research Council of Norway (grant no. 302333). B.R. acknowledges the support of Visual Intelligence, funded by the Research Council of Norway (grant no. 309439). The measurements at the University of Twente were carried out within the Perspectief Program Cancer ID [14193], which was in part financed by the Netherlands Organization for Scientific Research–Domain Applied and Engineering Sciences (NWO-TTW). Ing. Aufried Lenferink is acknowledged for building the measurement system at the University of Twente.

Author contributions statement

M.N.J. conceived of the architecture, implemented the model and wrote the manuscript draft. A.M. conducted experiments producing 2284 of the measurements. S.K. conducted the experiments resulting in 279 of the measurements. O.S. and E.G. prepared and provided the samples resulting in 279 of the measurements. C.O. conducted and oversaw the experiments resulting in 2284 of the measurements. B.R. oversaw the development of the architecture and writing of the manuscript. O.G.H initiated the work, managed the project and participated in writing the manuscript. All authors reviewed the manuscript.

Additional information

See supplementary information.

Competing interests The authors declare no competing interests.

SUPPLEMENTARY INFORMATION
Identification of extracellular vesicles from their Raman spectra via self-supervised learning

Mathias N. Jensen[1], Eduarda M. Guerreiro[2], Agustin Enciso-Martinez[3,4,5], Sergei G. Kruglik[6], Cees Otto[7], Omri Snir[2,8], Benjamin Ricaud[1], and Olav Gaute Hellesø[1,*]

[1]Department of Physics and Technology, UiT The Arctic University of Norway, Tromsø, Norway.
[2]Thrombosis Research Group (TREC), Department of Clinical Medicine, UiT The Arctic University of Norway, Tromsø, Norway.
[3]Oncode Institute and Ten Dijke/Chemical Signaling Laboratory, Department of Cell and Chemical Biology, Leiden University Medical Center, Leiden, The Netherlands
[4]Amsterdam Vesicle Center, Department of Biomedical Engineering and Physics, Amsterdam University Medical Centers, Amsterdam, The Netherlands
[5]Laboratory of Experimental Clinical Chemistry, Department of Clinical Chemistry, Amsterdam University Medical Centers, Amsterdam, The Netherlands
[6]CNRS, Institut de Biologie Paris-Seine, Laboratoire Jean Perrin, Sorbonne University, Paris, France
[7]Department of Medical Cell BioPhysics, TechMed Centre, University of Twente, Enschede, The Netherlands
[8]Department of Medical Biology, UiT The Arctic University of Norway, Tromsø, Norway
[*]olav.gaute.helleso@uit.no

ABSTRACT

Extracellular vesicles (EVs) and nanoparticles released from cells attract interest for their possible role in health and diseases. The detection and characterization of EVs and other biological nanoparticles is challenging due to the lack of specialized methodologies. Raman spectroscopy, however, has been suggested as a novel approach for biochemical analysis of nanoparticles. To extract information from the spectra, a novel deep learning architecture is explored as a versatile variant of autoencoders. The proposed architecture considers the frequency range separately from the intensity of the spectra. This enables the model to adapt to the frequency range, rather than requiring that all spectra be pre-processed to the same frequency range as it was trained on. It is demonstrated that the proposed architecture accepts Raman spectra of nanoparticles from multiple biological origins and laboratories. High reconstruction accuracy is maintained despite large variances in frequency range and noise level. It is also shown that the architecture is able to cluster nanoparticles by their Raman spectra and differentiate them by their origin without pre-processing of the spectra or supervision during learning. The model performs label-free differentiation of nanoparticles from 13 biological sources with high fidelity, including separating extracellular vesicles from activated vs. unactivated blood platelets and nanoparticles from prostate cancer patients vs. non-cancer controls. The differentiation is evaluated by creating a neural network classifier that observes the features extracted by the model to classify the samples according to their origin. The classification reveals a test sensitivity of 92.2% and selectivity of 92.4% over 783 nanoparticles measured at two different labs with two different measurement configurations.

Sample preparation

For 'Paris' dataset, sample prepared at UiT The Arctic University of Norway and analysed at Sorbonne University

Collection of clinical samples has been approved by the regional ethical committee for Medical and Health Research Ethics (REK200825). All participants were above the age of 18, did not suffer from illness, or use medication; all gave a written informed consent. Blood was drawn by venipuncture of the antecubital vein using a 21-gauge needle and minimal stasis. Blood was collected into Vacuette 6 ml Z tubes with no additives (Greiner Bio-One, Kremsmünster, Austria); the first tube was discarded. Acidic citrate dextrose buffer (ACD, 39 mM citric acid, 75 mM sodium citrate, 135 mM [D]-glucose, pH 4.5) and 2,8 mM Prostaglandin E1 (PGE1, MedChemExpress, Monmouth Junction, NJ, USA) were added rapidly to the blood to prevent blood coagulation and platelet activation, respectively. In addition, 3 ml of blood were drawn into K2EDTA Vacuette tubes (Greiner Bio-One, Kremsmünster, Austria) for cell count and analysis using an ABX MicrosES60 (ABX Diagnostics, Montpellier, France). Following addition of ACD and PGE1, blood was centrifuged at 140 xg for 15 min with no breaks (room

temperature) using a Megafuge 1.0 (Heraeus Sepatech) centrifuge equipped with a swing bucket rotor BS4402/A to generate platelet rich plasma (PRP).

Platelet pellets were recovered from PRP following centrifugation at 900 xg for 15 min at room temperature, washed twice with HEPES-NaCl buffer (10 mM HEPES, 0.85% NaCl, pH 7.4) and 2.8 mM PGE1, and resuspended in Tyrode-HEPES buffer (10 mM HEPES, 0.85% NaCl, 1 mM MgCl2, 2 mM CaCl2, 3 mM KCl, pH 7.4). Platelets (250x106 in 1 ml Tyrode-HEPES buffer) were stimulated with 100 µM thrombin receptor activator peptide 6 (TRAP-6, MedChemExpress, Monmouth Junction, NJ, USA) or 2 µM calcium ionophore A23187 (Sigma-Aldrich, USA) and incubated for 15 min at 37°C. Saline was added for time-matched unstimulated control platelets. Following 15 min, EDTA was added to platelet suspensions (activated and time-matched control) at a final concentration of 10 mM to stop platelet activation and platelets were sedimented at 2,500 xg for 10 min at room temperature. Supernatant was transferred to a new tube before proceeding with isolation of platelet-Evs, and the platelet pellets were resuspended in 1% paraformaldehyde (PFA) in PBS for assessment of platelet activation by flow cytometry. Platelet derived (micro)particles, hereafter referred to as platelet-Evs, were isolated from supernatant by centrifugation at 20,000 xg for 30 min at 4°C, using a 5810R Eppendorf centrifuge with a fixed angle rotor FA-45-30-11. EV pellets were resuspended in 1/10 of their initial volume (i.e., 10x concentrated) in a buffer suitable for the respective downstream analysis.

For EV characterization by Raman spectroscopy, Nanoparticle Tracking Analysis (NTA), and cryogenic transmission electron microscopy (Cryo-TEM), EV pellets were resuspended in PBS. For analysis of procoagulant activity of Evs by procoagulant phospholipid clotting assay and thrombin generation by calibrated automated hrombogram (CAT), Evs were re-suspended in pooled EV-depleted plasma (EVDP). EVDP was prepared as previously described by Ramberg *et al.*[1]. Briefly, citrated blood from 10 healthy individuals was centrifuged twice at 2,500 xg to produce platelet free plasma (PFP). PFP was subjected to ultracentrifugation (100,000 xg, 60 min) at 16°C (Beckman Optima LE-80 K Ultracentrifuge, rotor SW40TI, Beckman Coulter, USA). EVDP samples were pooled, aliquoted and stored at -80°C until further use.

For analysis of the prothrombinase complex activity Evs were resuspended in 20 mM HEPES, 150 mM NaCl buffer. All EV suspensions were stored at -80°C until use.

For 'Twente' dataset, sample prepared and analysed at University of Twente

Plasma from healthy donors and cancer patients

Blood was obtained from non-fasting healthy donors and mCRPC patients (both N=5) after written informed consent in accordance with the Helsinki Declaration and approved by the medical-ethical assessment committee of the Academic Medical Center, University of Amsterdam (NL 64623.018.18). Refer to Table S1 for donors clinical data. Whole blood was collected from each donor using a 21G needle, and the first vacutainer was discarded. Next, three citrate vacutainers of 2.7 mL (BD Biosciences, San Jose, CA) were collected and mixed gently by inversion. The vacutainers were centrifuged at 2500 g for 15 minutes at 20 °C without brake (Rotina 380R, Hettich, Tuttlingen, Germany). Plasma was collected up to 0.5 cm above the pellet, pooled and centrifuged in a conical base tube (10 mL; Sarstedt, Nimbrecht, Germany) at 2500 g for 15 minutes at 20 °C. The supernatant was deposited in aliquots of 75 µL (Sarstedt), which were snap frozen in liquid N2 and stored at 80 °C until use. Samples were thawed in a water bath at 37 °C immediately before use.

Lipoprotein particles (LPs) Human high density lipoprotein (HDL), low density lipoprotein (LDL), very low density lipoprotein (VLDL) and chylomicrons (CM) were acquired from Sigma-Aldrich Chemie N. V. (The Netherlands). HDL (Cat. No.: L8039), LDL (Cat. No.: 437644), VLDL (Cat. No.: 437647) and CM (Cat. No.: SRP6304) had a purity of $\geq$ 95% by electrophoresis, as specified by the provider.

LNCaP-derived EVs Cells from the prostate cancer cell line LNCaP (ATCC, CRL-1740, USA) were cultured at 37°C and 5% CO2 in RPMI-1640 with L-glutamine medium (Lonza, Cat. No.: 12-702F) supplemented with 10% (v/v) fetal bovine serum (FBS), 10 units/mL penicillin and 10 mg/mL streptomycin. Cells were seeded at a density of 10,000 cells/cm2 as recommended by ATCC and medium was refreshed every second day. At 80-90% confluence, cells were washed three times with phosphate buffer solution (PBS) and cultured in FBS-free RPMI-1640 with L-glutamine medium (Lonza, Cat. No.: 12-702F) supplemented with 1 unit/mL penicillin and 1µg/mL streptomycin. After 2-3 days of culture, cell supernatant was collected in a 15 mL tube (Cellstar® tubes, Grenier Bio-one BV, Alphen a/d Rijn, The Netherlands) and centrifuged at 500 g at room temperature for 10 minutes (centrifuge 5804, Eppendorf, Hamburg, Germany). Next, the supernatant containing LNCaP-derived EVs was collected and stored in aliquots (Greiner Bio-one) at -80 °C until use. Samples were thawed in a water bath at 37 °C immediately before use. LNCaP-derived EVs are referred to as LNCaP EVs throughout the text.

Red blood cell (RBC) - derived EVs RBC-derived EVs were obtained from RBC concentrate (150 mL, Sanquin Bloodbank, Amsterdam, The Netherlands) and diluted 1:1 with filtered PBS. Samples were centrifuged three times at 1560 g for 20 minutes at 20 °C (Rotina 46RS centrifuge, Hettich, Tuttlingen, Germany). The supernatant containing EVs was pooled and distributed in aliquots of 50 µL, which were snap frozen in liquid N2 for 15 minutes and stored in aliquots (Sarstedt) at -80 °C until use. Samples were thawed in a water bath at 37 °C before use. RBC-derived EVs are referred to as RBC EVs throughout the text.

For other particles used from Twente, see the methods-section of Martinez *et al.*[2]

Table S1. Clinical data on the cancer patients afflicted with Metastasized castration-resistant cancer (mCRPC) and control donors.

Subject	Date of diagnosis	Date of CRPC diagnosis	Treatment	Age	PSA*	LDH	ALP	tChol	HDL	LDL	TGL
Pt 1	05/2017	10/2018	Abiterone Prednisone	66	0.6	202.9	1.6	3.89	1.27	2.13	1.08
Pt 2	2006	04/2018	Enzalutamide	86	3.4	165.4	1.4	5.44	1.57	3.24	1.41
Pt 3	2000	04/2018	Abiterone Prednisone	75	1.2	175.6	5.0	3.14	1.37	1.39	0.86
Pt 4	NA	NA	NA	77	17.6	165.3	3.3	5.45	2.22	2.82	0.92
Pt 5	NA	NA	NA	82	10.4	240.3	6.4	5.75	1.36	3.59	1.78
HD min	NA	NA	NA	19	NA	125.8	1.2	3.45	0.93	1.81	0.42
HD max	NA	NA	NA	40	NA	270.3	3.2	4.38	1.45	2.47	1.93

Pt: patient, HD: healthy donor (N=5), PSA: prostate specific antigen *last determined PSA level before inclusion
LDH: Lactate dyhydrogenase (U/L 37C), ALP: Alkaline phospatase (U/L 37C), tChol: total cholesterol (mmol/L)
HDL: high density lipoproteins (mmol/L), LDL: low density lipoproteins cholesterol (mmol/L), TGL: Triglyceride (mmol/L).

Dataset

Table S2. Overview of samples in the data set with number of samples, distribution and ranges. The heterogenity in frequency range is clear, with some having a range of approximately 3300 cm^{-1} and some with a range of approximately 1700 cm^{-1}. There is also a large variability in the number of samples for each of the origins, with 898 cancer patient derived particles and only 19 high density lipoproteins.

Origin/Properties	Subspecies	$N_{samples}$	$\tilde{\nu}$ min (cm^{-1})	$\tilde{\nu}$ max (cm^{-1})
Plasma		153	301	3672
Cancer ctrl.		745	300	3652
Cancer		898	300	3652
RBC		56	300	3668
Platelet	Ctrl. (Paris)	82	307	2036
	Act. A23 (Paris)	77	308	2035
	Act. Trap (Paris)	81	307	2035
	Mixed (Twente)	184	300	3671
THP		41	315	2041
LNCaP		74	300	3673
PC3		94	300	3668
CM		64	300	3648
HDL		19	300	3652
LDL		47	302	3652
VLDL		69	300	3648

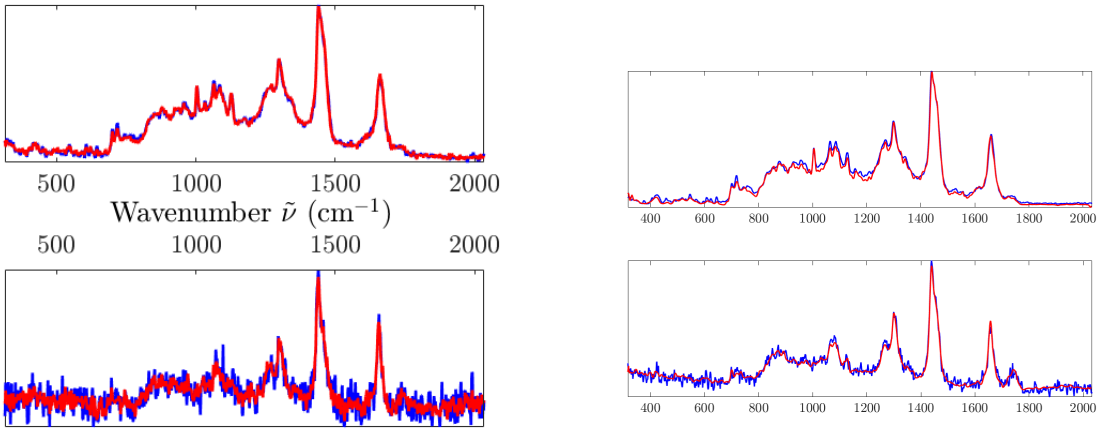
Architecture

Table S3. The final architecture of the encoder and decoder with the total number of parameters.

Element	Block	Layers	Filters	Width	Parameters
Encoder	Conv. Block E1	6	32	5	30 912
	Conv. Block E2	6	64	3	74 112
	Conv. Block E3	4	128	3	197 120
	Feed forward EFreq.	3	64	1	8 704
	Feed forward E	4	512	1	2 475 571
	Latent space	1	1	110	0
Decoder	Feed Forward DFreq.1	3	64	1	9 024
	Feed forward D	4	512	1	872 448
	Feed Forward DFreq.2	1	5	1	30
	Conv. Block D1	2	64	3	12 992
	Conv. Block D2	3	32	3	22 656
Classifier	Feed forward C	5	128	1	78 976

Supplementary results

Reconstruction



(a) Reconstruction by base model with explicit frequency range consideration.

(b) Reconstruction by a simple autoencoder model without skip or explicit frequency range consideration.

Figure S1. Reconstruction performance of comparative methods. The PCA reconstruction shown in a) is limited to 100 components, same as the autoencoder, and results in the reconstruction being highly accurate, but also preserves a significant amount of noise. For the autoencoder reconstruction shown in b) it is shown that the accuracy is high, and the process eliminates a significant amount of noise but the lack of adaptivity to frequency removes a significant portion of the spectra.

Clustering

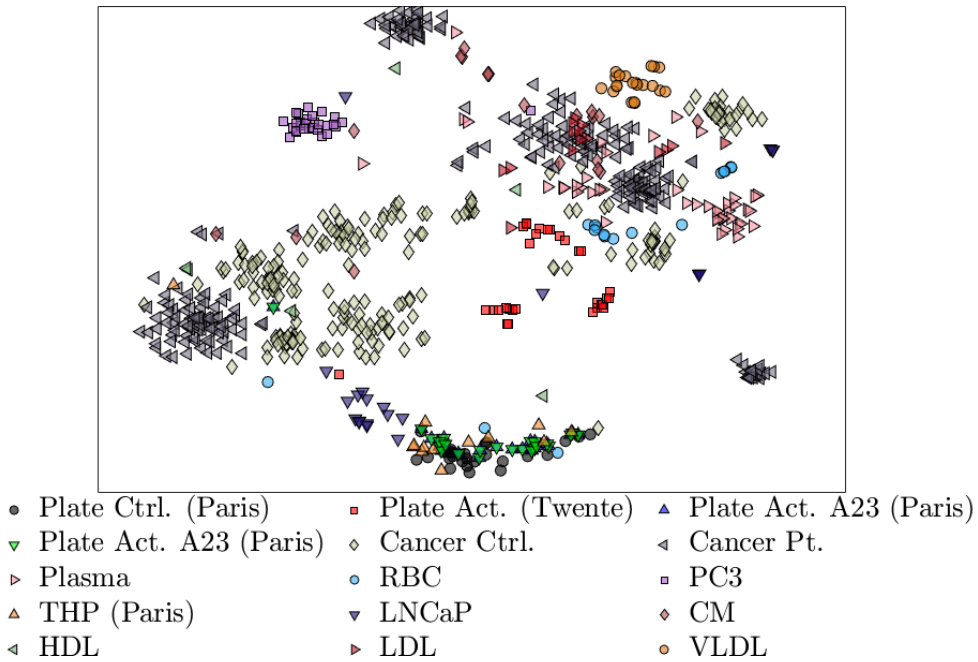
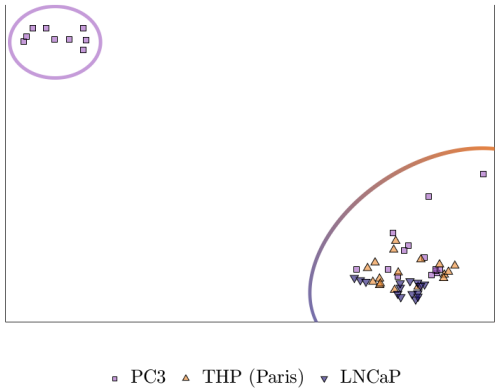
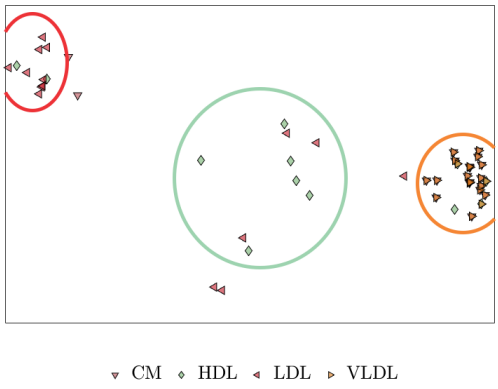


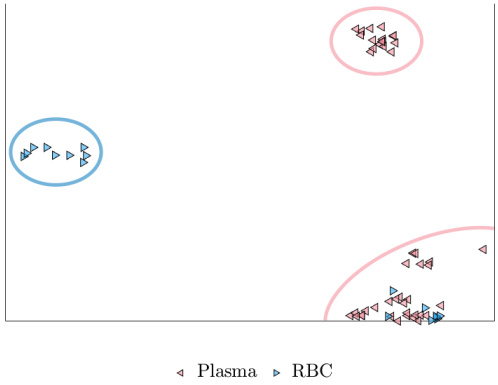
Figure S2. t-SNE projection of PCA components for test data. The particle origins cluster chaotically, forming several smaller clusters, notably the smeared cluster at the center bottom which reflect samples from Paris. The clusters are highly intermixed, similar to the autoencoder results, but the PCA results are more spread out. See especially the particles from PC3,LNCaP, and THP-1 which are differentiated by the PCA despite their intrinsic similarity and unlike the autoencoder, which recognizes them as similar.



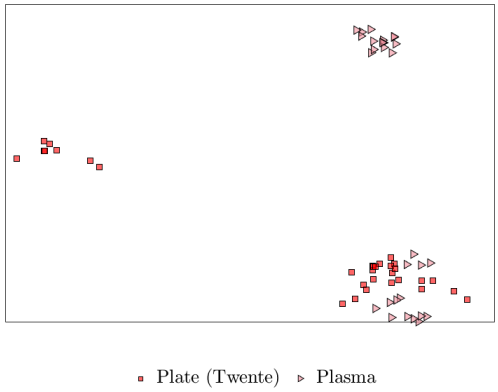
(a) t-SNE projected latent space for cell culture derived EVs. The PC3 and LNCaP derived nanoparticles cluster well and form tight, well distinguishable groups. The THP derived particles form a loose, poorly defined group that overlaps with the PC3 particles.



(b) t-SNE projected latent space for lipoproteins. The low density lipoproteints (LDL) form a distinct cluster that is well separated, while the high density lipoproteins (HDL) form a diffuse but differentiable cluster. The chylomicrons (CM) and the very low density lipoproteins (VLDL) form a common cluster. However, as spectra of both particle types had an extremely poor signal-to-noise ratio, it is possible that this is largely a noise response.



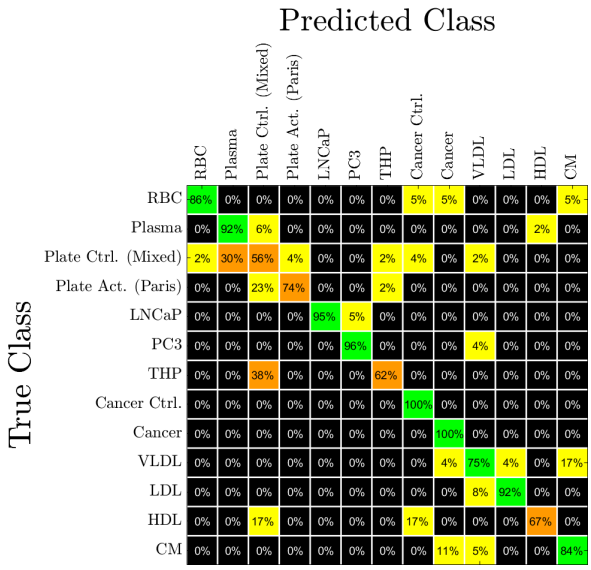
(c) t-SNE projected latent space for bulk blood elements. The red blood cells (RBC) form a distinct cluster that is well differentiated from the plasma, albeit with some outliers that overlap with the plasma. The plasma derived particles form two distinct clusters with one partially overlapping with the RBC particles.



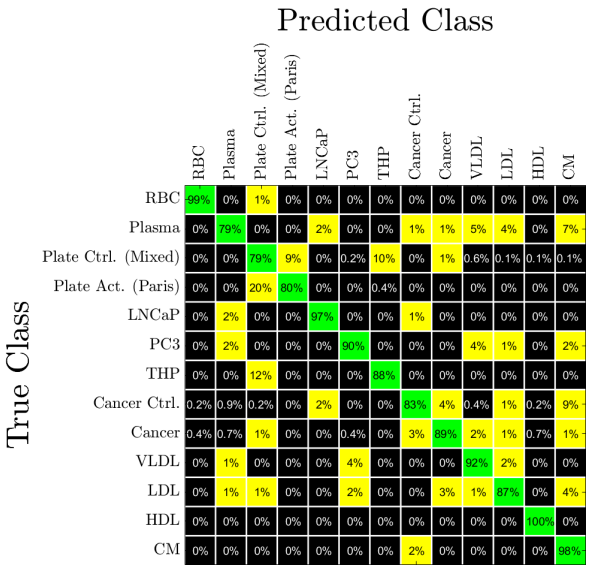
(d) t-SNE projected latent space for platelets and plasma. The platelets and the plasma derived nanoparticles form a common group in the lower right. This agrees with expectations as the platelets originate from stored concentrate, which is known to cause activation and generation of lipoproteins, making the platelet derived nanoparticles more similar to the general plasma derived particles.

Figure S3. t-SNE projection of the latent space for the autoencoder and for various selections of EVs and nanoparticles.

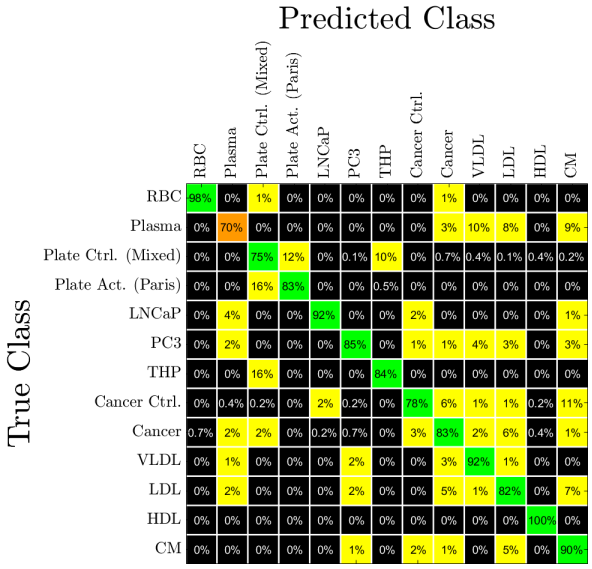
Classification



(a) Confusion matrix for unaltered test data.



(b) Confusion matrix for randomly clipped test data



(c) Confusion matrix for test set with clipped and noisy test data.

Figure S4. Confusion matrices for noise augmented test sets. The percentages are given by the number of true labels, e.g. 66% RBC indicates that 66% of the true RBC samples are classified as RBC by the model. The confusion matrix in a) shows the results of classification on the source data of the test set, without any artificial distortion or noise. The confusion matrix in b) shows the results of classification on the test set augmented with random clipping of the spectra, as was done to the training set during learning. The results show a general reduction in accuracy doe to increased misclassification, indicated by yellow, but remains generally good, indicated by the green on the diagonal. The confusion matrix in c) shows the result of classification on the test set augmented with random clipping, intensity noise, and frequency distortion in the same manner as the training set during learning. The accuracy is further reduced, indicated by the orange and yellow, due to increased misclassification. The colors indicates the percentage correct predictions, with yellow for 1% to 25% correct, orange for 25% to 50% correct, and green for more than 75% correct prediction.

References

1. Ramberg, C. The role of plasma extracellular vesicles and procoagulant phospholipid activity in venous thromboembolism. (2021).

2. Enciso-Martinez, A. *et al.* Label-free identification and chemical characterisation of single extracellular vesicles and lipoproteins by synchronous rayleigh and raman scattering. *J. Extracell. Vesicles* **9**, 1730134, DOI: https://doi.org/10.1080/20013078.2020.1730134 (2020). https://onlinelibrary.wiley.com/doi/pdf/10.1080/20013078.2020.1730134.