GreenTropism

Photonics and Artificial Intelligence
Combining technologies for pathogen diagnostics
Paris, Rouen, September 2022

This paper presents the current state of the art of the combination of photonic technologies, especially Raman and SERS spectroscopy, with Artificial Intelligence, in the application of pathogen detection (viruses, bacteria, fungi, etc.). After a brief review of the different technologies involved in this combination, we will present two use cases based on our most recent experiments, which highlight our advances.

Chapter 1 - Some background information

Pathogens

The word pathogen (from the Greek pathos, suffering, and -genes, producer of) describes in biology an organism capable of infecting a host and producing a disease. Many taxa therefore fall under this umbrella term, such as viruses, bacteria, fungi, etc. When a disease carried by a pathogen affects a large percentage of a population in a short period of time, we are dealing with an epidemic episode. If the same disease spreads worldwide, we are talking about a pandemic event, such as the Covid-19 disease caused by the SARS-CoV-2 virus.
Subjected in their own way to Darwinian evolution, pathogens are generally characterized by their infectivity (the ability to infect another target of the same species), their mutation rate and their lethality. These abilities to adapt strongly to new hosts (even new species), new environmental conditions, and to spread rapidly, to the point of becoming a threat to global health (as evidenced by variants of SARS-CoV-2), make pathogen control a major focus of research and development. This ongoing work has two main facets: treatment - drugs and vaccines - and diagnosis of this infectious disease. For some diseases and therefore some pathogens, late treatment (due to late diagnosis) may induce severe side effects or simply no treatment is currently available, making the diagnostic effort increasingly important and strategic.

Diagnosis of pathogens today relies on a large number of different biological tests for their identification: microscopic examination; culture of samples placed in conditions favorable to the growth of infectious agents; detection of molecules produced by the immune system (antibodies) in response to the infectious agent; detection of molecules originating from the infectious agent (antigens); and search for genetic material (DNA or RNA) of the infectious agent.
Often, it is not possible to identify the infectious agent with a single test, most of them require the intervention of specialized laboratories, a lot of time and can be very expensive.

Photonics

Just as the word «electronics» refers to the study and use of electrons, photonics refers to the science of photons: generation, detection, manipulation and processing.
For the clarity of this paragraph, we will develop here the branch of spectroscopy among the various applications available such as optics, photonic computing, plasmonics, etc.

Spectroscopy

In the 17th century, the decomposition of white light by triangular glass prisms became a subject of interest and recreational observation. Among his peers, Sir Isaac Newton was the first scientist to publish on the future discipline of spectroscopy - and in particular what would become spectroscopy in the visible range of the electromagnetic spectrum - in 1672, in a letter to the Royal Society containing «his new theory of light and color». It is also Newton who invented the word «spectrum» to describe the observed phenomenon.
In 1800, William Herschel noticed that the different colors of the spectrum to generate

different amounts of heat. He then tested his hypothesis by reproducing the prism experiment and adding thermometers under each color obtained. In order to have a «negative» control, he placed one beyond the red color and observed a higher temperature than that observed for the red light. He concludes that there is «an invisible form of light» after the red color, and thus highlights the infrared wavelengths.

Since Newton and Herschel, spectroscopy has encompassed more and more experiments, observations, technologies, and is known today as the field of study of the electromagnetic spectrum, and the interactions between photons and matter, allowing to understand the composition and structure of objects ranging from bacteria to stars, molecules, foods, beverages, cosmetics, chemicals, etc.

Raman spectroscopy is a technique for non-destructive analysis of the structure of a molecule. This technology is based on Raman scattering, a phenomenon predicted by Smekal in 1923 and observed by Râman in 1928 in organic liquids.
In general, when monochromatic radiation impinges on transparent material systems such as gases, liquids, and perfectly transparent solids, the vast majority of the light is transmitted without any variation, but a scattering phenomenon still occurs. Analysis of this scattered light reveals the existence of different frequencies. The scattering of light that does not involve a change in frequency, and therefore wavelength, corresponds to Rayleigh scattering and that with a variation of frequencies corresponds to Raman scattering.
During Raman scattering, an exchange of energy occurs between the sample and the monochromatic source (visible or near infrared). This energy exchange induces a variation of the frequency of the scattered photons compared to the excitation frequency. The variation then depends on the molecular structure of the sample that has been excited by the source. Raman signals are naturally weak. To increase the sensitivity of the measurement, exaltation by SERS effect is then often used. SERS signal exaltation uses metal surfaces with nanometer scale particles that are added to the sample to be analyzed. The most commonly used metals are gold and silver. When the molecule enters the vicinity of the metal structure, a significant enhancement of the Raman signal of a few orders of magnitude is observed. A very low concentration can thus be detected without having to functionalize the molecule of interest.

Photonic data and Artificial Intelligence

Spectroscopy (and more particularly vibrational spectroscopy which is our subject here) can be classified among learning technologies. In order to be able to interpret a spectrum, to determine if the analyzed sample belongs to class A or class B, or contains such quantity of such molecule, it is necessary to train algorithms to perform these tasks and calculations. The algorithms will rely on data pairs (spectrum; target parameter) to learn to recognize variations in the target parameter correlated to variations in the spectra (variations that can be direct and linear, or variations in ratios between peaks, etc.). As learning, algorithmic tests and increasing samples representative of the problem to be solved in the database, the performance and accuracy of the models in classification and quantification will increase, until a maximum threshold is reached. This working method is at the heart of Artificial Intelligence, which encompasses the often confused subfields of Machine Learning, Deep Learning and Chemometrics. The general working philosophy, whatever the typology of the algorithms, consists in designing experiments in such a way that a part of the samples (the database, i.e. spectra and metadata) is dedicated to the selection of the pre-processing algorithms and the models, then a second part for the optimization of their hyperparameters, and the last part to blind tests: samples that have never been used or seen by the models will be injected in the models in order to evaluate the accuracy and the robustness of the algorithms.

Since the early days of industrial applications of spectroscopy, easy-to-implement algorithms have been used to search for correlations between changes in reference data (e.g., a change in glucose concentration) and changes in spectral patterns (e.g., the change in specific peak amplitudes). This approach to multivariate statistics is commonly referred to as «chemometrics» (a term coined by Svante Wold in

1971) and is generalized as the science of extracting information from data sets generated by chemical systems.
The increase in computing power has since allowed the emergence of software and software components such as Artificial Intelligence developed by GreenTropism, which relies on both Artificial Intelligence and more computationally intensive algorithms and combinations of algorithms (such as artificial neural networks and genetic algorithms) to automate the chemometrics process, allowing researchers and industry to extract more information and generate more accurate and robust models faster. Thanks to Artificial Intelligence and the ability to train hundreds of models in a few seconds, it is also possible to manage this process with new metrics to evaluate the quality of input databases, the reliability of predictions and to better take into account the effects of hardware drifts, background interferences and weak but interesting signals in the spectra.

This combination of Artificial Intelligence and photonics technologies further enables new users and uses: new users, such as industrial operators, lab technicians and engineers, academics with no prior knowledge of chemometrics, democratizing the technology; and new uses that were difficult to achieve due to the difficulty of dealing with massive data and exploring thousands of algorithmic combinations, from preprocessing to modeling strategies.

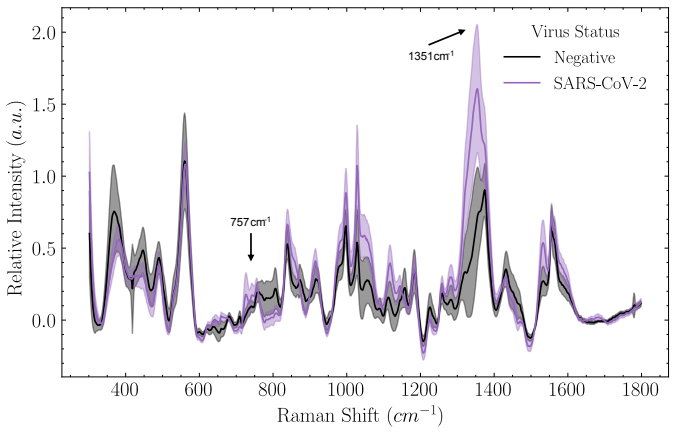
Chapter Two - Case Studies - Application to Pathogen Diagnostics

Pathogen diagnostics is one of these new uses and will now be developed with some use cases. The subject of pathogen diagnosis is indeed delicate because many additional factors, in addition to the mere absence or presence of pathogens, will interfere with the light and thus contribute to the signal: the presence/absence of additional molecules depending on the typology of the transport buffer, the lysis buffer, the sampling process and the origin of the sample: nasopharyngeal swab, saliva, cerebrospinal fluid, blood, etc.

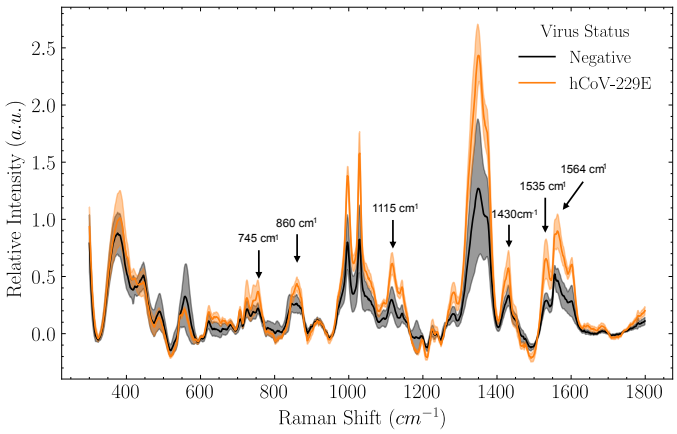
When considering the implementation of a sustainable population-based screening strategy, in terms of time, delay and use of consumer goods, photonics is a fairly obvious choice. It enables near-instantaneous collection of information, without chemicals or consumables. Combined with Artificial Intelligence, it is possible to train several models to identify several pathogens in the same sample, i.e. a multiplex analysis with a single spectrum.

Case Study 1: Detection of SARS-CoV-2, HCoV-229E and H1-N1 in SERS and Artificial Intelligence controlled samples.

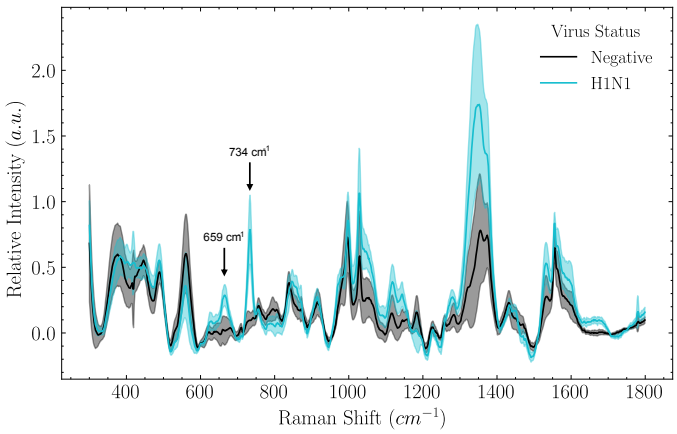
In this study, we analyzed three different respiratory viruses:
- SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2), the enveloped betacoronavirus responsible for the current pandemic and causative agent of coronavirus 2019 (COVID-19).
- hCoV-229E, the seasonal human alphacoronavirus 229E.
- H1N1, the swine influenza A virus.
The same protocol was applied to each virus: gold nanoparticles with a controlled range of diameters were gently mixed with the samples. Nine spectra were acquired during only 30 seconds, at a constant and controlled distance. In parallel, negative control spectra were acquired.
The spectra were mainly processed with SNV (Standard Normal Variate) normalization and ALS (Asymmetric Least Squares) smoothing.



Mean SERS spectra with standard deviation of SARS-CoV-2 coronavirus after pretreatments

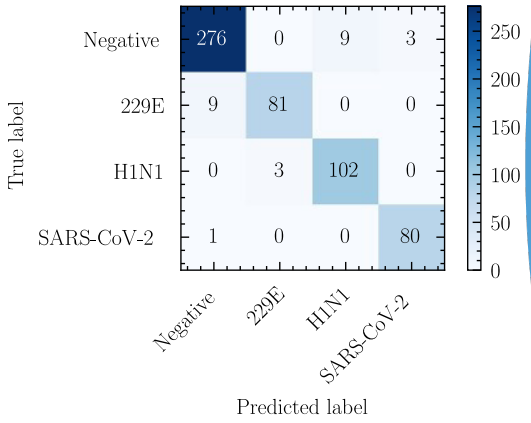


Mean SERS spectra with standard deviation of hCoV-229E coronavirus after pretreatments



Mean SERS spectra with standard deviation of H1N1 influenza virus after pretreatments

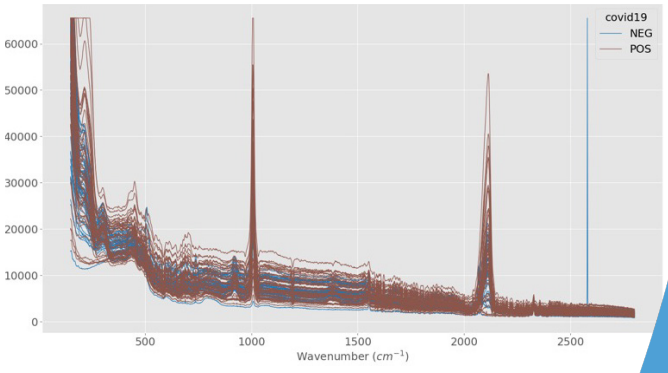
An AI model was built to perform the classification tasks on the entire dataset, and to handle all the issues related to measurements with a spectrometer, with a classical calibration/validation/testing approach (80% of the samples for the training set and 20% of the samples for the test set) to optimize the pre-processing and hyperparameters of the models. With a four-class classifier, we achieved excellent prediction performance in the test set with 99% accuracy, 97% sensitivity and 99% specificity. Therefore, our technique is able to classify SARS-CoV2, hCoV-229E and H1N1 viruses with near perfect accuracy and sensitivity.



Confusion matrix of the four classes

Case Study 2: Detection of SARS-CoV-2 in human samples (nasopharyngeal swabs) by SERS and AI

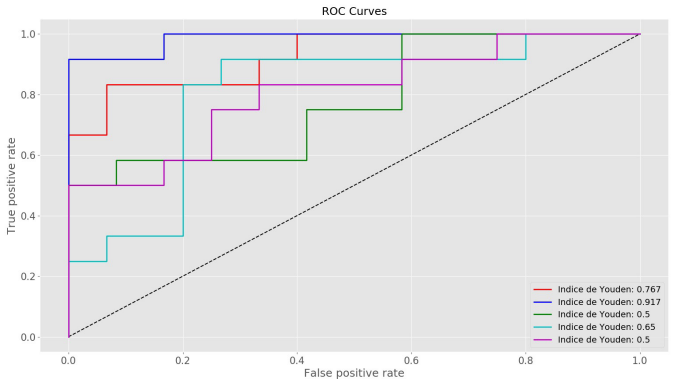
During one of the epidemic peaks, GreenTropism was able to access nasopharyngeal samples - with prior consent from patients - to test the validity of the combination of spectrometry and AI for field samples. The reference comparator for sample status was determined by RT-PCR.
The protocol remained the same as in case study 1: gold nanoparticles were added to a fraction of the samples, gently mixed, and analyzed by spectrometry after a drop was placed on the sample holder; the dataset was split for learning and blind testing.



Raw spectral database

After peak analysis, the first part of the spectra was linked to the presence of gold nanoparticles and discarded from the analysis. Different modeling strategies were tested and compared in order to take into account the increased variability of the samples: indeed, in addition to the positive or negative status of the samples, the samples were collected from patients without discrimination of sex or age, smokers and non-smokers, with different collection dates, etc.
As we are working on a binary classification, we have chosen to present the results using ROC curves, as is commonly done in the field of diagnosis and pathology, and to calculate the Youden index J, which is expressed as follows:
J = sensitivity + specificity - 1

In terms of interpretation of the results, a perfect classifier would be in the upper left corner, a random classifier would be on the diagonal, and the closer the Youden index is to 1, the better the model.



We finally arrived at a model characterized by a Youden index of 0.92 (dark blue curve), conducive to a test scaling step.
These experiments will need to be repeated in bioclinical trials on a larger number of samples and tested against strain mutations to ensure maximum reliability and performance of a photonics and AI based diagnostic.

Conclusion and perspectives

While the combination of photonics and chemometrics has been a powerful assemblage of technologies for several decades already, allowing in particular the conversion of spectra into numerical values or classes for industrial purposes (e.g. water content, protein content, classification for quality control), the amount of available algorithms, combinations of algorithms and hyperparameters for their fine tuning represented until now an obstacle to reach a higher level of accuracy and robustness. The introduction of GreenTropism's Artificial Intelligence into this ecosystem has enabled this and other breakthroughs, such as a drastic increase in efficiency in terms of data mining and model reliability, the ease with which it is possible to manage multiple spectrometers and datasets in parallel. In this particular work, we can also observe the capabilities of AI to handle the high molecular complexity of the samples (i.e., the noise of the spectra), whether nasopharyngeal or salivary, and a level of performance compatible with pathogen detection.

The author

Anthony Boulanger holds a PhD in microbiology and biostatistics, after which he joined the Veolia Group as a researcher in chemometrics and near infrared spectroscopy for environmental processes. In 2014, he founded GreenTropism to design and apply Artificial Intelligence algorithms for chemometrics automation and photonic data processing, for technologies such as vibrational spectroscopy, atomic spectroscopy and multi/hyperspectral imagers.

Acknowledgements

The results were obtained thanks to the work of the GreenTropism teams of the Laboratory and Data Science departments, in particular Dr. Delphine Garsuault, Tiffany Guédet, Florent Perez and Alexandre Banon, as well as Dr. Marion Schmitt-Boulanger for the coordination and management of the two case studies.

info@greentropism.com
www.greentropism.com
@greentropism
@greentropism