Clinical Spectroscopy 5 (2023) 100025

Contents lists available at ScienceDirect

Clinical Spectroscopy

journal homepage: www.sciencedirect.com/journal/clinical-spectroscopy

Detection of several respiratory viruses with Surface-Enhanced Raman Spectroscopy coupled with Artificial Intelligence

Delphine Garsuault [a,*], Sanaa El Messaoudi [a], Mookkan Prabakaran [b], Ian Cheong [b], Anthony Boulanger [a], Marion Schmitt-Boulanger [a,*]

[a] GreenTropism, Paris 75008, France
[b] Temasek Life Sciences Laboratory, National University of Singapore, Singapore 119077, Singapore

ARTICLE INFO

Keywords:
Virus detection
SARS-CoV-2
SERS
Artificial intelligence

ABSTRACT

Diagnoses of viral infections are a challenge when facing a crisis like COVID-19, where their speed and reliability are critical to minimize diseases spread. The gold standard of diagnostics, quantitative Polymerase Chain Reaction, is time- and reagent-consuming and requires qualified personnel. Therefore, it is necessary to find new detection techniques to overcome these barriers. Surface Enhanced Raman Spectroscopy (SERS) is a detection method, based on light and metallic particles admixed with the samples, already used in different fields of research. In this study, we discriminate three respiratory viruses using a combination of SERS and Artificial Intelligence (AI). Our technique appears to be fast, reproducible, and reliable, achieving between 95 % and 100 % of accuracy, standing out as a powerful tool usable for viral diagnostics.

1. Introduction

Viruses are one of the major causes of diseases in the world [1]. They exhibit multitude of structural forms and have evolved many mechanisms to infect people and animals. Viruses of interest, as they can cause epidemics, include coronaviruses like the Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV), as well as Influenza viruses. Viral epidemics occur regularly and can reach pandemic status according to the infectivity and method of transmission, like COVID-19 outbreak, which started in Wuhan in 2019. To control these epidemics, it is crucial to quickly detect infected people and to treat or isolate them before the virus spreads to others [2,3]. Currently, the gold standard for detecting viruses is quantitative Polymerase Chain Reaction (qPCR). This is preceded by a Reverse Transcription (RT) step if the viral genetic material is RNA instead of DNA. Based on molecular biology principles, this technique is very sensitive and specific but requires qualified personnel, specialised equipment and expensive reagents, and is time consuming [4,5]. Thus, qPCR is not appropriate when fast detection and isolation is required as a strategy and unsuitable as point-of-care test. Other techniques are also used to detect viruses, such as antigenic or serologic tests. These are faster than qPCR but present other issues. Antigenic tests have too poor sensitivity and specificity [6,7], and serologic tests can only be used several weeks after the emergence of symptoms because it detects antibodies produced in reaction to infection [5]. Because these tests are not ideal tools for pandemic control, we need new innovative ways to detect and diagnose with fast, accurate, and reliable performances.

Last decades, Raman spectroscopy has emerged as a non-invasive, non-destructive and versatile technique to provide molecular information [8–12]. The signal obtained from the samples is defined by a difference in energy between the incident photon and the emitted one, that can be higher or lower. This difference creates a shift in the wavelength of the photon called Raman shift [13,14]. This shift depends on the chemical composition of the samples, which is why we obtain a unique fingerprint for each sample analysed with Raman spectroscopy or SERS [15]. Compared to Raman signal, which is very weak, SERS allows an enhancement by several orders of magnitude [10,16]. This enhancement is a combination of electromagnetic and chemical factors and is observed with different kind of nano- and microstructures, mainly metallic [10,14]. Surface Enhanced Raman Spectroscopy (SERS), thanks to its ability to provide valuable information about complex samples [16,17] is used in different fields such as detection of food contaminants [18], environmental field [4,14], or biomedical applications [4,19].

Raman spectroscopy and SERS are widely used in the biological field, and particularly for pathogens detection [9,20], and present several advantages compared to other spectroscopic techniques for this kind of studies. In Raman spectroscopy and SERS, there is no interference of

* Corresponding authors.
E-mail addresses: delphine.garsuault@gmail.com (D. Garsuault), marion.schmitt-boulanger@greentropism.com (M. Schmitt-Boulanger).

https://doi.org/10.1016/j.clispe.2023.100025
Received 3 February 2023; Received in revised form 5 July 2023; Accepted 10 July 2023
Available online 12 July 2023

water in the signal, a molecule very abundant in all biological samples, thus it is possible to analyse aqueous samples very easily. These techniques do not need any preparation of the sample, are very fast and are non-destructive. Moreover, Raman spectroscopy and SERS are very precise, as they result in a fingerprint of the whole sample, with information on specific chemical bond vibrations of the sample [10,14]. But, even with these advantages, their application in routine is limited [9,19,21–23]. This can be explained by reproducibility issues [9,19,21], difficulties to interpret the results [12,22], lower sensitivity than gold standards like qPCR or Enzyme Linked Immunosorbent Assay (ELISA) for example [20,23]. These problems may have different origins: a protocol that gives inconsistent results, a substrate or a laser wavelength not optimal to analyse the sample, difficulties on analysis of the spectra and various sources of noise [24–26]. To improve reproducibility and sensitivity, it is possible to functionalize the nanoparticles to target the analyte of interest [20]. A first treatment of spectra with statistical algorithms or the use of AI have also been proposed as solutions to improve reliability and reproducibility of SERS technique [9,24,27], but these solutions are not unique, as each lab has its own combination of algorithms. To improve SERS technique and be able to use it in routine in virus detection, we lack a common way of acquiring and treating data for analysis.

Different types of machine learning algorithms can be used to analyse SERS data. We chose Relevance Vector Machines (RVM) as they are a parsimonious probabilistic model used for regression and classification. This model is based on finding the optimal margin hyperplane which, where possible, properly classifies or separates data while being as far away as possible from all observations. The principle is to find a classifier, or a discrimination function, which quality of forecast is as great as possible. Furthermore, RVM is a technique based on a probabilistic formulation that not only provides a prediction, but also gives an informative predictive distribution [28,29].

In this work, we combined AI tools and SERS measurements to detect the presence of three respiratory viruses in cell cultures. We used gold particles, chosen for their biocompatibility, their efficacy, and because their signal in the SERS spectra does not overlap the signal of our analysed samples. These particles were not functionalized, and we used the same protocol for the whole study. This protocol is very simple as it consists solely of a direct mix of the sample and the particles, which is then deposited on an aluminium slide. We chose aluminium as it does not present any Raman activity with our choice of laser wavelength. SERS spectra acquired for each virus helped to build individual predictive models to distinguish positive and negative samples for each virus. Then, a global classification model was created to differentiate the viruses from one another with an accuracy of 95 %.

This work aims to improve the analysis of SERS spectra with an automation of the technique using the AI and to integrate this in a fast diagnostic solution. In SERS spectra, we can sometimes spot differences between samples with the naked eye, but it is difficult to determine to which sample these spectra correspond without an expert and tables of bands assignment. Our technique is very advantageous as it can recognize a sample automatically, by scanning a database in only a few seconds. Thus, it can become a helpful tool for diagnosis and recognition of biological samples. Our technique is also faster and need very few reagents compared to the current gold standard and, with the use of AI, it is easier and more reliable than other SERS detection techniques.

2. Materials and methods

SARS-CoV-2 viral strain (hCoV-19/Singapore/2/2020; GISASID Accession ID EPI_ISL_407987), human coronavirus 229E (ATCC; VR-740) and influenza A H1N1 were used for this study. Vero-E6 (CRL-1586), MDCK (CCL34) and MRC-5 (CCL171) cell lines were previously obtained from ATCC, USA. The SARS-CoV-2, hCoV-229E and Influenza A H1N1 viruses were grown in Vero-E6, MRC-5 and MDCK cells, respectively. Culture medium used for all these cell lines and viral cultures was Dulbeccos Modified Eagles Medium (DMEM, Gibco, Grand Island, NY, USA) complemented with 10 % Foetal Bovine Serum (FBS). The spectrometer STRam was purchased from Metrohm (USA) and operated with a 785 nm laser connected to a golden probe, furnished with the spectrometer. Spectral resolution of the spectrometer is under 6,0 cm^{-1} at 912 nm, according to the manufacturer. The spectrometer was calibrated with a polystyrene reference (furnished by the manufacturer) at the beginning of each day of experiment. Gold particles were purchased from Metrohm (USA). These particles were characterized by the provider: mean size of 100 nm, spherical shape and stabilized in Sodium citrate buffer. The initial concentration of these particles is 0,15 $g.L^{-1}$, corresponding to a molar concentration of $7{,}62.10^{-4}$ $mol.L^{-1}$. Aluminium slides were purchased from Jeulin (France).

Cells were infected with viruses at Multiplicity of Infection (MOI) of 0.5–1 and the supernatant was harvested two to three days after infection, depending on the cytopathic effect on the cells, observed under a microscope. The harvest consisted in pipetting 100 μL of supernatant and transferring it into a 1.5 mL tube. All culture samples were analysed just after this transfer, except for 12 samples (hCoV-229E) that were frozen and analysed the next day. For each virus, similar initial concentration was added to the cell culture and the virus grew in the same conditions, for the same number of days. Thus, the concentrations per virus were considered similar for all samples of this virus.

The gold particles were centrifuged for 45 min at 800 g at 4 °C and concentrated by discarding part of the supernatant (final concentration: 3,75 $g.L^{-1}$, corresponding to $1{,}90.10^{-2}$ $mol.L^{-1}$). These concentrated particles were then used for one to two weeks of experiments. The samples were prepared for SERS analysis by pipetting 10 μL of gold particles in a 1.5 mL tube and adding twice the volume of culture sample to the same tube. Then, the solution of particles and sample was mixed by pipetting several times to have uniformly dispersed gold particles inside the mix before being deposited on an aluminium slide. Three droplets of equal volume were deposited per sample. The slide was then placed under the laser for spectral acquisition. The laser probe was placed at 1 cm above the sample, as advised by the manufacturer, in order to have the best excitation of the whole sample. Nine spectra were acquired per sample, three for each droplet, with an acquisition time of 30 s at 100 % laser power (495 mW at the source, and 420 mW at the sample, according to the manufacturer). All experiments involving SARS-CoV-2 viral cultures were performed in a BSL-3. Other viruses were handled in BSL-2.

Spectra acquired from the samples were first analysed visually then using statistical algorithms. Different pretreatments were investigated to extract most of the discriminating information between samples, such as signal power normalization with Standard Normal Variate (SNV) [30], normalization by rescaling data between 0 and 1, or normalization by a simple maximum rescaling, resulting in a new maximum of 1 in all data. Other pretreatments can be used, like smoothing or derivative of the spectra using Savitzky-Golay (SG) algorithm [31], baseline reduction with Asymmetric Least Squares smoothing (ALS) [32,33] or dimensionality reduction with Principal Component Analysis (PCA) [34]. Signal power normalization pretreatments consist of a subtraction of the spectrum by its own mean followed by a division by its variance or its standard deviation. ALS smoothing consists of a calculation then correction of the baseline, allowing a better visualisation of the peaks. The baseline will be estimated using a polynomial fitted from the raw spectrum, which will then be subtracted from the spectrum. Smoothing consists of reducing signal noise. From a frequency point of view, this consists of attenuating, or even eliminating, the high parasitic frequencies that are considered not to be part of the RAMAN spectrum, while keeping the information useful. The derivative, on the other hand, reduces the drift of the baseline and highlights the spectral ranges that contain the discriminant information. In the case of a first derivative, the operation will emphasize the bandwidths, while the second derivative will emphasize the position of the peaks. These methods can be managed by applying the same algorithm, named Savitzky-Golay [31]. Finally,

Dimensionality reduction with PCA is characterised by a reduction of the number of variables describing the spectra, while conserving most of the information, that allows a simplification of their analysis. All these pretreatments were selected as they do not modify the shape of the spectra and can improve the signal-to-noise ratio then, consequently, the analysis of the information contained in the spectra by the model.

We used the Relevance Vector Machines (RVM) method as the classification algorithm [28,29]. This model was optimized using the AI developed by GreenTropism which selected the best pretreatments to apply as well as the best set of initialization parameters for the RVM. This selection of pretreatments is indicated for each virus analysis. The RVM algorithm was not modified.

To ensure robustness of the model and reliability of the predictions, we split our data into training and validation sets. The training set enables the algorithm to recognise and learn from existing data. The second step in the process of building the AI model is to optimize modelling parameters and hyperparameters. For this purpose, we used the K-Fold cross validation technique [35]. The last step of building a reliable model is to see how well it performs on unseen data, which is the aim of the validation phase. One can build a perfect model on the training data with no error, but it may fail to generalise for unseen data, phenomenon known as "overfitting". Thus, after training and validating the model using K-fold cross validation, we applied the chosen model on a new, completely independent dataset, the test set, to measure the models predictive power and ensure its robustness and generalisation capacity. The samples constituting the test sets of each virus were left out of training and validation and their spectra were taken several days after the ones of the training and validation sets to ensure independency.

To note, the spectra were analysed as a whole, without looking for specific bands. Raw data spectra are measured in "Relative intensity". This intensity is the one resulting from a subtraction of the background signal, automatically performed by the spectrometer as follow. The instrument collects the signal of the sample during the excitation by the laser, then collects the signal without any light (background signal) for the same amount of time. For example, here, we did acquisitions of 30 s. This means that the laser was on during 30 s, then the laser was off during 30 s and the instrument collected the signal during these 60 s. Thus, it is the relative intensity between the one collected from the sample under laser excitation and the one collected from the background. The baseline observed in the spectra is mainly caused by residual Rayleigh scattering at low Raman shift values or by the fluorescence of organic molecules intrinsic to the analysed sample or by contamination of the sample [36].

For visualisation purposes, we present here spectra pretreated with SNV and ALS. These two pretreatments were chosen as they allow a better comparison with naked eye between samples or classes (positive and negative) on the figures. SNV allows a reduction of differences in the global intensities of the signals and removes the multiplicative interferences of scatter and particle size without modifying the shape of the spectrum [30]. ALS smoothing allows a better observation of the peaks by correcting the baseline [32,33].

3. Results

Three different respiratory viruses are analysed in this study: two Coronaviruses and one Influenza virus. Our first choice of coronaviruses was Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), the virus at the origin of the current pandemic and causative agent of Coronavirus Disease 2019 (COVID-19). This virus is an enveloped betacoronavirus with its genetic information encoded in a positive-sense single-stranded RNA and is thought to have zoonotic origins [37,38]. The second coronavirus is an enveloped alphacoronavirus, the seasonal human coronavirus 229E (hCoV-229E). This virus is not zoonotic, with negative-sense single-stranded RNA [38]. These viruses cause mainly upper respiratory tract disorders but can also be associated with lower tract disorders. The third virus chosen for this study was the Influenza A swine virus H1N1. This virus infects mainly pigs but can spread to humans, causing upper and lower respiratory tract disorders that can be lethal. Like the two other viruses studied here, H1N1 virus is an enveloped single-stranded RNA virus [39].

3.1. Study of coronaviruses

3.1.1. Study of SERS reproducibility

Because reproducibility is a major issue in experiments using SERS, we firstly validated our protocol on this topic. Supplementary Fig. S1 shows the results of this first experiment with 6 SARS-CoV-2 positive samples, coming from 6 different viral cultures. We observed a very good reproducibility between the 54 spectra of these samples (see Supplementary Fig. S1a) and between the 9 spectra acquired from one of these samples (see Supplementary Fig. S1b). An analysis of variance was realized to validate the previous results (see Supplementary Fig. S2). The analysis was done between the 6 positive samples described above, averaging 54 spectra, (see Supplementary Fig. S2a) and between the 9 spectra acquired from Sample 1 (see Supplementary Fig. S2b). These experiments also asserted the reproducibility of the particles in themselves to validate the information obtained from the provider. We also measured three spectra of the aluminium slide as a negative control and almost no signal was obtained from it, as expected at this laser excitation wavelength. We can observe that the signal coming from aluminium slide is at the same level as the background signal, thus, the signal we observed in our experiments is considered coming solely from the samples (see Supplementary Fig. S3). The reproducibility obtained here is representative of the whole study, thus we show mean spectra with standard deviation for the following sections of this article.

3.1.2. Study of coronaviruses

After this first step of protocol validation, we began our work with the study of coronaviruses. We acquired the spectra of a total of 56 samples of SARS-CoV-2 (27 negative and 29 positive, including the six samples used for protocol validation), and 64 samples of hCoV-229E (32 positive and 32 negative). These samples came from different viral cultures with not-infected cell culture supernatant used as a negative control, and at different time points to ensure day by day and hour variability, to observe the representative variation that can exist between samples. The mean spectra with standard deviation acquired from the training and validation sets of these coronaviruses, and their corresponding negative samples, pretreated with a Standard Normal Variate (SNV) [30] and an Asymmetric Least Squares (ALS) [32,33] are shown on Fig. 1. We chose these pretreatments for visualization purposes only, as they allow a better comparison between two classes of spectra when observed with the naked eye. We can observe for SARS-CoV-2 that some regions show clear differences between the two populations of samples, as marked on Fig. 1a. Similarly, we can see differences for hCoV-229E marked on Fig. 1b. These differences are part of the signal analysed by the AI model but the fact they are visible to the trained eye does not necessarily mean they would be the most discriminant sections of the spectra for the AI to give its result.

We can also observe that some bands differ from positive and negative samples. For SARS-CoV-2, only visible on positive samples are four main bands: at 755 cm^{-1}, 1123 cm^{-1}, 1352 cm^{-1} and 1535 cm^{-1}. And for hCoV-229E, we can see the same four bands present in SARS-CoV-2 positive samples, suggesting they can represent common molecules of these two coronaviruses. We can also notice three more bands, at 860 cm^{-1}, 1430 cm^{-1} and 1564 cm^{-1}(see Supplementary Table S1 for tentative assignment of these bands).

These spectra constitute a database on SARS-CoV-2 and hCoV-229E in these conditions of culture. It was used to build and optimize the AI model before testing it. After collecting all the spectra for training and cross-validation sets [35], several types of pretreatments were checked by our AI to reduce the background signal, clean the data, and validate the models parameters before the analysis with the RVM algorithm. The

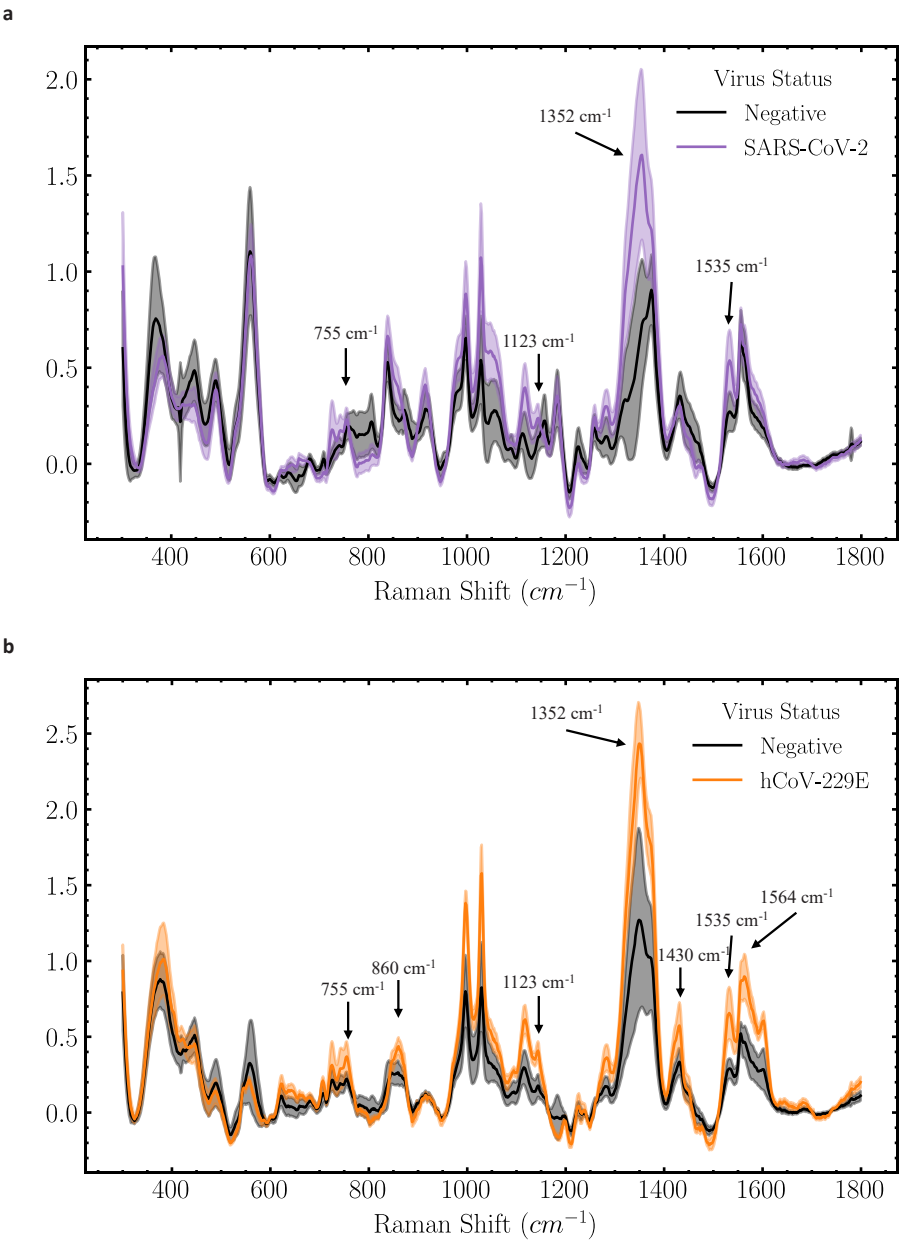

Fig. 1. SERS mean spectra (dark line) with standard deviation (light area) of train and validation sets of coronaviruses SARS-CoV-2 (purple) and hCoV-229E (orange) after preprocessing with SNV and ALS. A series of 9 spectra per sample have been collected to ensure a variability among the data. All spectra were acquired with a unique protocol. a) Spectra of SARS-CoV-2 positive (purple) and negative (black) samples. Coming from different viral cultures, 24 positive samples and 22 negative samples have been analysed. b) Spectra of hCoV-229E positive (orange) and negative (black) samples. 20 samples for each class were analysed.

combination of pretreatments with the RVM is called afterwards "AI model" or "model".

For SARS-CoV-2, the combination of pretreatments that gave the best results was a normalization followed by a Principal Component Analysis (PCA) [34] using 15 components with the following RVM. Next, this AI model was used on a test set containing 10 independent samples (5 positive and 5 negative) to evaluate a bad learning step: over or under fitted, and ensure its robustness and ability to give accurate predictions. The model produced perfect classification on this new dataset with 100 % accuracy (see Supplementary Fig. S4), predicting the class of all samples without any mistake. This means that the spectra of SARS-CoV-2 cultures, taken in their entirety, show sufficient differences between positive and negative samples to be perfectly classified by this model.

Concerning hCoV-299E, after data preprocessing with SNV followed by a dimensionality reduction by PCA using 15 components, the results obtained by the RVM algorithm in validation of the training set were excellent. However, we encountered some lack of generalization with the test set of 24 samples (12 positive and 12 negative), with a global accuracy of 64 %, a sensitivity of 18 %, and a specificity of 93 % (see Supplementary Fig. S5a). Additional analysis shows that the poorly recognized samples were from a frozen batch and can explain these results. Indeed, except for these 12 samples that were frozen, all samples of this study were analysed fresh, just after observation under the microscope for integrity and cytopathic effect when considered (positive samples only).

The frozen samples were removed from the database to fine-tune the model once again. With the exact same AI model algorithms as the one previously used for hCoV-229E, we improved the accuracy of the prediction to 93 %, with a sensitivity of 100 % and a specificity of 89 % (see Supplementary Fig. S5b). As expected, the frozen samples were the major cause of misclassification with the model. This highlights the fact that samples in test phase that are too different than the one from training and validation phase cannot be precisely predicted and may lead to errors.

These results validate our protocol combining SERS and AI as a technique capable of detecting positive and negative samples in viral cultures of coronaviruses. To ensure these results are reproducible, we

analysed another family of viruses.

3.2. Study of H1N1

After analysing these two coronaviruses, a respiratory virus from a different family was studied. We chose an Influenza A virus, H1N1. Because MDCK cells are very easy to cultivate and H1N1 infects them rapidly, it was possible to have more samples than for the other two viruses. We analysed a total of 166 samples (80 positive and 86 negative), including 69 samples (34 positive and 35 negative) in the test set. Spectra of the training and validation sets are shown on Fig. 2 after preprocessing with SNV and ALS.

We can observe marked differences between positive and negative samples on the mean spectra. Notably, between 600 cm^{-1} and 800 cm^{-1}, there are bands only visible on the positive spectra (see Supplementary Table S1 for bands assignment). Similarly to our study of the coronaviruses, we obtained excellent results in the training and validation phases, highlighting the fact that the positive samples could be almost perfectly distinguished from the negative samples by the model. During the optimisation process realised here by the AI, the best results of RVM were achieved using the following combination of pretreatments: a second order derivative with Savitzky-Golay (SG) algorithm [31], followed by a SNV and a PCA with 15 components. This model predicted the status of the test samples with an accuracy of 96 %, a sensitivity of 91 %, and a specificity of 99 % (Supplementary Fig. S6).

3.3. Comparison between the three viruses

After analysing the viruses one by one and comparing them to their corresponding negative samples, we examined the possibility to differentiate them from one another with our technique. To do so, we used the database previously created with the acquisitions from the analysis of each virus. We merged all the samples from the training and validation phases and mixed them together in a new dataset. We then separated them in two new sets: 80 % of the samples for training and 20 % for validation. We also merged the previous test sets to constitute a new test set for this experiment. Fig. 3 shows the mean spectra of the new training and validation sets with standard deviation of each virus after preprocessing with SNV and ALS.

When we compare the spectra of the three viruses we analysed, we can observe some bands present only in one virus as the one around 550 cm^{-1}, higher in intensity for SARS-CoV-2, the band between 650 cm^{-1} and 700 cm^{-1} only present in H1N1 virus signal, or the band at 1450 cm^{-1} higher in intensity and a little bit shifted compared to the other two viruses, for hCoV-229E (see Supplementary Table S1 for bands assignment). But we can also notice that the three spectral signatures are very similar. Because the culture medium is the same for all the samples, this similarity is consistent with the fact that the medium is the most present component and probably contributes the most to the whole signal.

Before analysing the new dataset with our AI model, we wanted to assess the natural separability of the groups. To do so, we used a Principal Component Analysis with 3 components to visualise clusters of data and eventual consistency with the viruses (Fig. 4).

The PCA analysis (Fig. 4a) showed general clustering and separation of the three classes with some degree of overlap between hCoV-229E and H1N1 in PC1. Class separation improved markedly with PC2, which represents 10 % of the data, and where we can see three separated groups representative of each virus. Because this analysis is an unsupervised classification, we can observe that the three viruses can linearly separate themselves naturally. The loadings of the PCA (Fig. 4b) show that several bands contribute to the Principal components. For PC1, representing most of the variance (89 %), the main contribution comes from the band at 1352 cm^{-1}, that we observed for the three viruses (see Supplementary Table S1). For PC2, representing 10 % of the variance of the PCA, in addition to the previous band observed, the bands that contribute the most seem to be the one at 1123 cm^{-1}, common to the three analysed viruses, and the band at 860 cm^{-1}, particularly observable in hCoV-229E. Finally, for PC3, we can observe a contribution of the 1352 cm^{-1} band once again, as well as the band at 734 cm^{-1}.

These findings were confirmed by building and testing an AI model, which gave excellent predictions in the new test set with an accuracy of 99 %, a sensitivity of 97 % and a specificity of 99 % (see Supplementary Fig. S7). The pretreatments used in this case were a second order derivative with SG algorithm, followed by a SNV and a PCA with 15 components. Thus, our technique is capable of classifying SARS-CoV2, hCoV-229E and H1N1 viruses with near-perfect accuracy and sensitivity.

After this comparison, we also added the negative samples to the previous datasets to build a four classes algorithm with the same

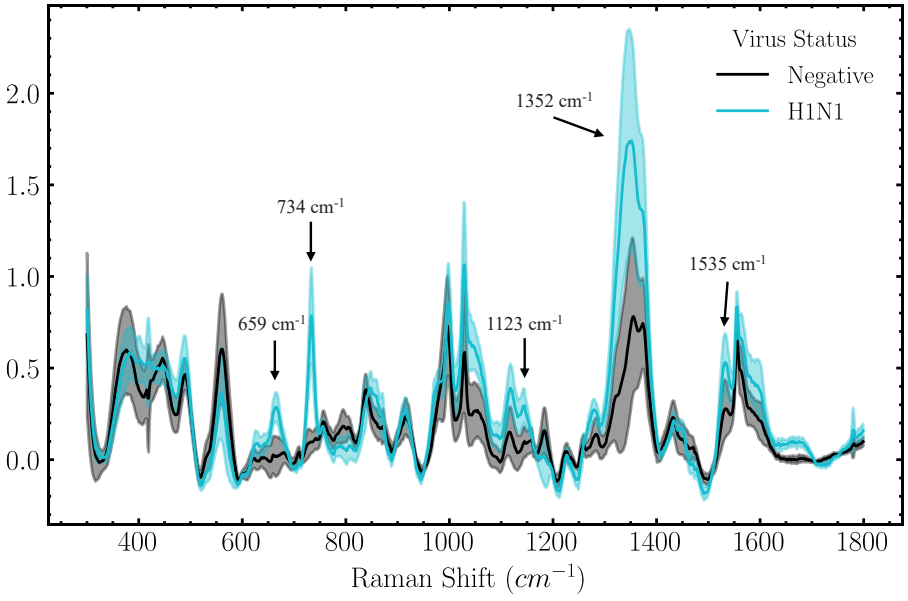

Fig. 2. SERS mean spectra (dark line) with standard deviation (light area) of train and validation sets of Influenza virus H1N1 (blue) after preprocessing with SNV and ALS. Coming from different viral cultures, 46 positive samples (light blue) and 51 negative samples (black) have been analysed using the same protocol. Once again, a series of 9 spectra per sample have been collected to ensure a certain level of variability among the data.

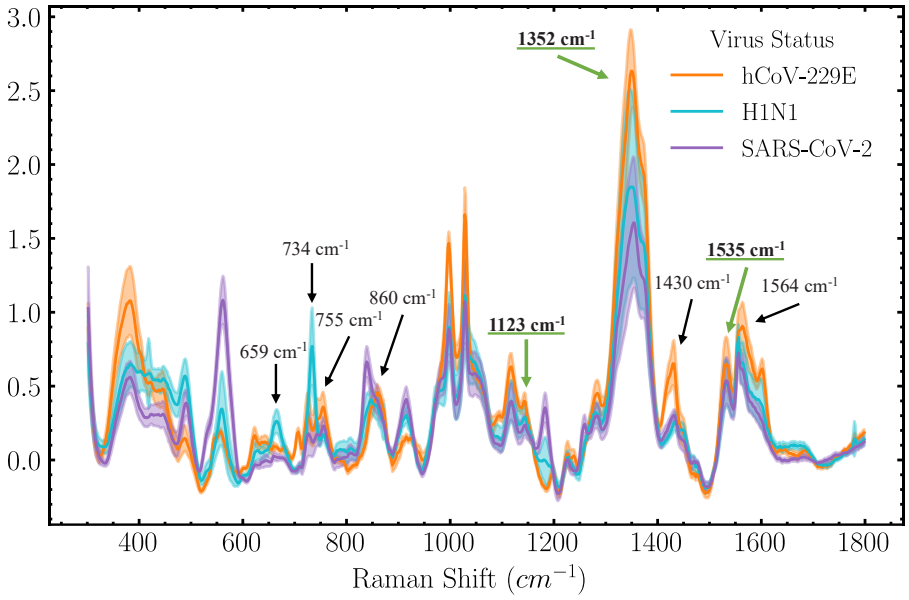

Fig. 3. SERS mean spectra (dark line) with standard deviation (light area) of three respiratory viruses (hCoV-229E in orange, H1N1 in blue and SARS-CoV-2 in purple) after preprocessing with SNV and ALS. Coming from different viral cultures, 20 hCoV-229E samples, 24 SARS-CoV-2 samples and 46 H1N1 samples have been analysed using different cells but same protocol. Bands underlined in green are common to the three viruses.

repartition as the previous ones: addition of the negative samples from training and validation sets of individual viruses to the new training and validation set, and addition of the negative samples from the test sets of individual viruses to the new test set. Once again, we can observe excellent results with a global accuracy of 95 %, a sensitivity for hCoV-229E, H1N1, SARS-CoV-2 and negative samples of 90 %, 97 %, 99 % and 96 %, respectively, as shown on Fig. 5.

We can see that 276 out of 288 negative spectra were correctly classified, 9 were misclassified as H1N1 positive spectra and 3 as SARS-CoV-2 positive spectra. 81 out of 90 hCoV-229E spectra were correctly classified and 9 were misclassified as negative spectra. 102 out of 105 H1N1 spectra were correctly classified and 3 were misclassified as hCoV-229E spectra. 80 out of 81 SARS-CoV-2 spectra were correctly classified and 1 was misclassified as a negative spectrum.

4. Discussion

In this study, we analysed three respiratory viruses: two coronaviruses, hCoV-229E and SARS-CoV-2, as well as an Influenza A virus, H1N1, with a combination of SERS and AI. We observed that these viruses are perfectly identifiable from their corresponding negative samples, and that it is possible to accurately differentiate them from one another, using gold particles not functionalized nor flagged. All the spectra and results obtained in this study constitute the beginning of a database on viruses analysed with our technique. We plan to broaden it to be able to detect and discriminate more viruses in the future. Once the database is created, the detection will be even faster and reliable as the best AI algorithms models will already be preselected.

For SERS detection of viruses, or biological samples in general, the use of metallic particles is the most described in previous studies. Gold and silver are the metals that are the most used in SERS experiments [1, 40]. Both kind of particles give good results in SERS experiments and react differently to the samples. In our work, we decided to use gold particles as they are thought to be more biocompatible and less prone to oxidisation. Moreover, silver has antimicrobial activity than can be problematic when studying pathogenic microorganisms [41]. These parameters seemed important to us to guaranty a good reproducibility of our technique, so we carefully analysed them to determine which would be better with our protocol. Particles can also be functionalised with antibodies, aptamers, or other structures capable of catching the analyte of interest and concentrate it. This functionalisation is widely used in SERS applications for biological samples [12,22,23,40]. Here, we wanted to use a simple protocol that can work with different types of analytes and decided to use bare nonmagnetic particles for this study.

We can observe with the preliminary experiments on reproducibility that there are some fluctuations in intensity of the signal of these samples. This is a particularity of SERS experiments, as this technique gives a fingerprint of the whole sample and not two samples are exactly the same. Our AI model, by analysing the spectra on their entirety and using pretreatments, can recognise patterns in the database that are not influenced by the slight differences in intensity. The AI model can recognise patterns that are statistically representative of the information we want to analyse.

The algorithms and models used to classify the spectra were not trained to recognize specific bands corresponding to chemical vibrations in the sample. On the contrary, they were trained to analyse the spectra as a whole and to highlight the differences and discriminate between positive and negative samples. To validate the results, we realized an analysis on the prominent bands for each virus [42–46]. Supplementary Table S1 presents these observations: these bands are only visible in positive samples and match with chemical bonds mostly present in proteins, which is coherent with the results given by the AI. Bands associated with Tyrosine and Tryptophan, particularly, are observed in the spectra we obtained. These two amino acids play a role in viral infection and replication, thus their presence is coherent with our analysis. One of the bands can also be representative of RNA, at 1123 cm^{-1}, as previously observed in literature [45], but, as we did not analysed RNA samples in our experiments, we cannot discriminate between a protein band and a RNA band.

In this study, we used viruses with different concentrations, between 10^5 TCID50.mL^{-1} and 10^{11} TCID50.mL^{-1}, to infect the cell cultures and were able to discriminate between positive and negative for all of them. These concentrations were the ones from the native viruses, with hCoV-229E at the concentration of 10^5 TCID50.mL^{-1}, SARS-CoV-2 at a concentration between 10^7 TCID50.mL^{-1} and 10^8 TCID50.mL^{-1}, and H1N1 at the concentration of 10^{11} TCID50.mL^{-1}.

When comparing the different spectra, we can observe differences between negative samples. This can be explained by the fact that three different cell lines were used in this study, Vero-E6, MDCK and MRC-5. However, the culture medium used to grow these cell lines and

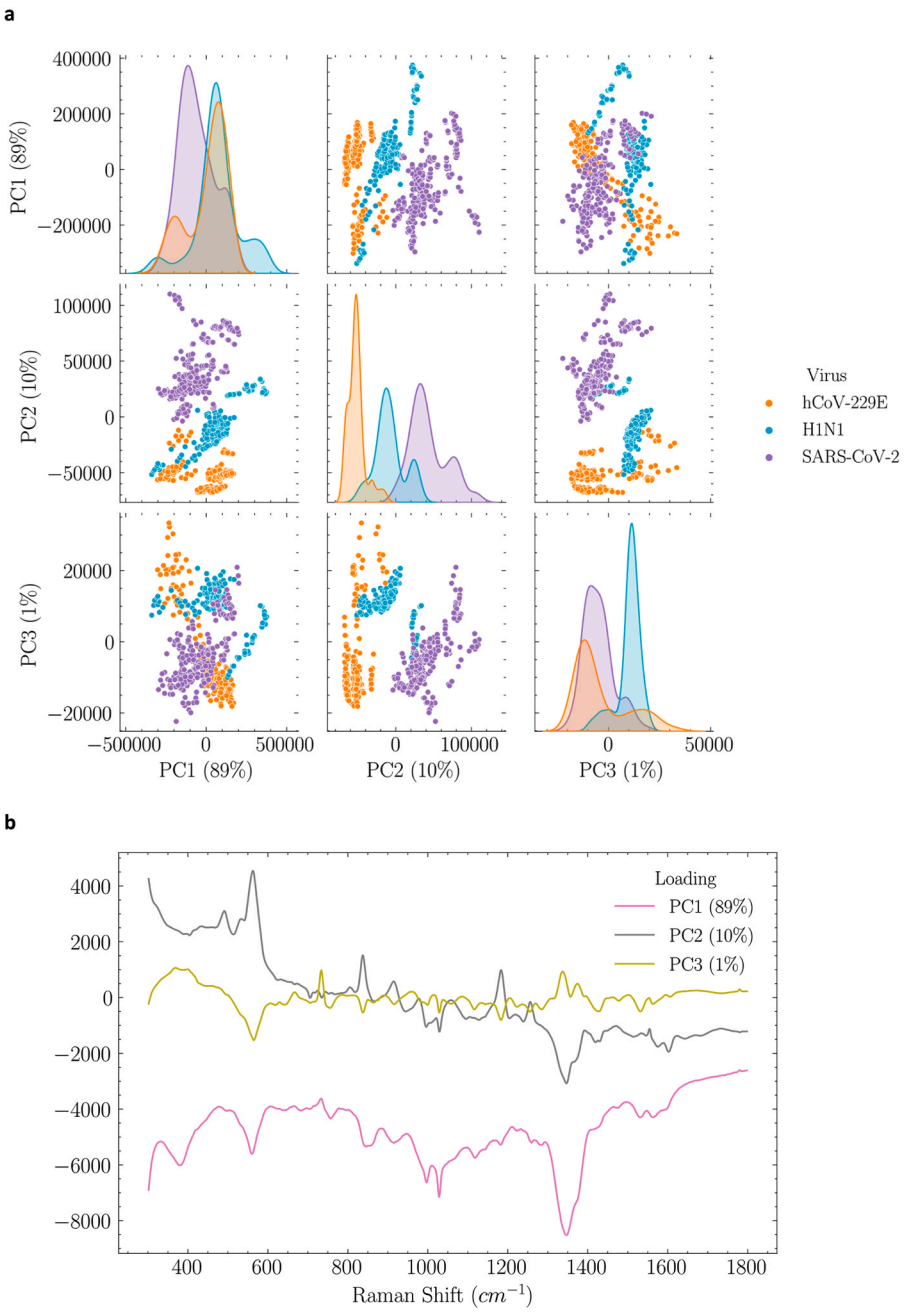

Fig. 4. Principal Component Analysis of the spectra from training and validation sets of the positive samples of the three viruses, hCoV-229E (Orange), H1N1 (Blue) and SARS-CoV-2 (Purple) and its loadings (PC1 in pink, PC2 in grey, PC3 in yellow). a) This PCA presents the same spectra than Fig. 3 with each point corresponding to the PCA score for one spectrum. The purpose is to study the separability of the three viruses on a three-dimensional space. The percentages represent the variance corresponding to each component (PC1/2/3). b) This figure represents the loadings for the three principal components of the presented PCA.

undertake the experiments was the same, DMEM. As we used the supernatant of the cell cultures to do the analysis, it can be slightly different because of the cell lines.

Studies on differentiation of bacteria with SERS already exist in literature [41], even if Raman spectroscopy is more used than SERS. For example, Ding et al. analysed three different strains of Salmonella with a combination of SERS and Convolutional Neural Network (CNN) [47]. Here, we validated our protocol of acquisition and analysis on viral culture samples, in a same culture medium. We demonstrated that SERS can give very reproducible and accurate results, as shown on the viruses analysed here. The combination of SERS with artificial intelligence gives very promising results in these conditions and could be used to detect rapidly viruses in more complex samples. A new study on samples in other media would be interesting, as well as one on patient samples, where there is more variability because of the microbiota of the patients.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Delphine Garsuault did the experiments with the help of Ian Cheong and Mookkan Prabakaran, who provided the biological material. Sanaa El Messaoudi, Delphine Garsuault and Ian Cheong analysed the data. Delphine Garsuault and Sanaa El Messaoudi wrote the manuscript. Marion Schmitt-Boulanger, Mookkan Prabakaran, Ian Cheong and Anthony Boulanger reviewed and corrected the manuscript. Ian Cheong, Mookkan Prabakaran and Marion Schmitt-Boulanger supervised and helped design the whole project.

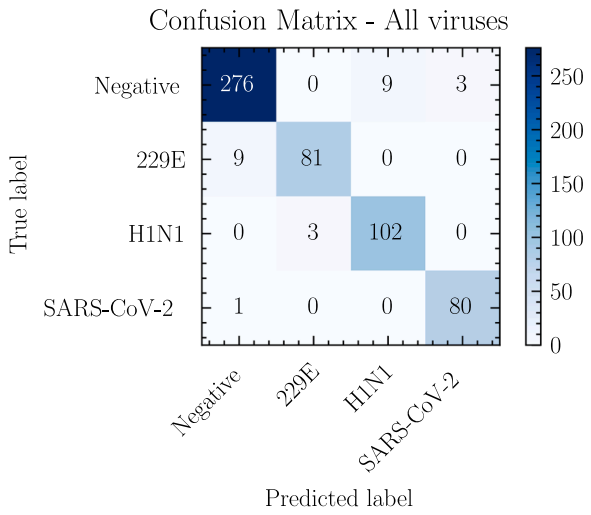

Fig. 5. Confusion matrix on all test data. This figure shows the predictions of the classification model comparing the true labels of the samples (y-axis) and the predicted labels by the model (x-axis) on all samples.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests. Delphine Garsuault, Sanaa El Messaoudi and Marion Schmitt-Boulanger are full-time employees of GreenTropism at the moment of the experiments. Anthony Boulanger. is full-time employee, Chief Technology Officer, and founder of GreenTropism. Other authors declare no conflict of interest.

Data Availability

Data will be made available on request.

Acknowledgments

We particularly thank Temasek Life Sciences Laboratory, Singapore for the use of its BSL-3 facilities and other ancillary support. We also thank Alexandre Banon, Data scientist, for his precious help.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.clispe.2023.100025.

References

[1] T.J. Moore, A.S. Moody, T.D. Payne, G.M. Sarabia, A.R. Daniel, B. Sharma, In vitro and in vivo SERS biosensing for disease diagnosis, Biosensors 8 (2) (2018), https://doi.org/10.3390/bios8020046.
[2] N. Younes, et al., Challenges in laboratory diagnosis of the novel coronavirus SARS-CoV-2, Viruses 12 (6) (2020), https://doi.org/10.3390/v12060582.
[3] A. Singh, V. Gupta, SARS-CoV-2 therapeutics: how far do we stand from a remedy? Pharmacol. Rep. 73 (3) (2021) 750–768, https://doi.org/10.1007/s43440-020-00204-0.
[4] M. Chisanga, H. Muhamadali, D.I. Ellis, R. Goodacre, Enhancing disease diagnosis: biomedical applications of surface-enhanced Raman scattering, Appl. Sci. 9 (6) (2019), https://doi.org/10.3390/app9061163.
[5] J. Sitjar, J.-D. Liao, H. Lee, H.-P. Tsai, J.-R. Wang, P.-Y. Liu, Challenges of SERS technology as a non-nucleic acid or -antigen detection method for SARS-CoV-2 virus and its variants, Biosens. Bioelectron. 181 (2021), 113153, https://doi.org/10.1016/j.bios.2021.113153.
[6] J. Dinnes, et al., Rapid, point-of-care antigen tests for diagnosis of SARS-CoV-2 infection, Cochrane Database Syst. Rev. 3 (2021), https://doi.org/10.1002/14651858.CD013705.pub2.
[7] E. Schuit, et al., Diagnostic accuracy of rapid antigen tests in asymptomatic and presymptomatic close contacts of individuals with confirmed SARS-CoV-2 infection: cross sectional study, BMJ 374 (2021) n1676, https://doi.org/10.1136/bmj.n1676.
[8] E. Cordero, I. Latka, C. Matthäus, I. Schie, J. Popp, In-vivo Raman spectroscopy: from basics to applications, J. Biomed. Opt. 23 (7) (2018) 1–23, https://doi.org/10.1117/1.JBO.23.7.071210.
[9] C.-C. Andrei, et al., SERS characterization of aggregated and isolated bacteria deposited on silver-based substrates, Anal. Bioanal. Chem. 413 (5) (2021) 1417–1428, https://doi.org/10.1007/s00216-020-03106-5.
[10] R. Pilot, R. Signorini, C. Durante, L. Orian, M. Bhamidipati, L. Fabris, A review on surface-enhanced Raman scattering, Biosensors 9 (2) (2019), https://doi.org/10.3390/bios9020057.
[11] M. Fleischmann, P.J. Hendra, A.J. McQuillan, Raman spectra of pyridine adsorbed at a silver electrode, Chem. Phys. Lett. 26 (2) (1974) 163–166, https://doi.org/10.1016/0009-2614(74)85388-1.
[12] S. Tanwar, S.K. Paidi, R. Prasad, R. Pandey, I. Barman, Advancing Raman spectroscopy from research to clinic: Translational potential and challenges, Spectrochim. Acta Part A: Mol. Biomol. Spectrosc. 260 (2021), 119957, https://doi.org/10.1016/j.saa.2021.119957.
[13] C.V. Raman, K.S. Krishnan, A new type of secondary radiation, Nature 121 (3048) (1928), https://doi.org/10.1038/121501c0.
[14] C.L. Haynes, A.D. McFarland, R.P. Van Duyne, Surface-enhanced Raman spectroscopy, Anal. Chem. 77 (17) (2005) 338 A–346 A, https://doi.org/10.1021/ac053456d.
[15] R. Wang, P. Yuan, M. Han, S. Xu, T. Wang, X. Wang, Asymmetry of Raman scattering by structure variation in space, Opt. Express OE 25 (15) (2017) 18378–18392, https://doi.org/10.1364/OE.25.018378.
[16] K.C. Bantz, et al., Recent progress in SERS biosensing, Phys. Chem. Chem. Phys. 13 (24) (2011) 11551–11567, https://doi.org/10.1039/C0CP01841D.
[17] S.-C. Luo, K. Sivashanmugan, J.-D. Liao, C.-K. Yao, H.-C. Peng, Nanofabricated SERS-active substrates for single-molecule to virus detection in vitro: a review, Biosens. Bioelectron. 61 (2014) 232–240, https://doi.org/10.1016/j.bios.2014.05.013.
[18] R. Pilot, SERS detection of food contaminants by means of portable Raman instruments, J. Raman Spectrosc. 49 (6) (2018) 954–981, https://doi.org/10.1002/jrs.5400.
[19] S. Abalde-Cela, P. Aldeanueva-Potel, C. Mateo-Mateo, L. Rodríguez-Lorenzo, R. A. Alvarez-Puebla, L.M. Liz-Marzán, Surface-enhanced Raman scattering biomedical applications of plasmonic colloidal particles, J. R. Soc. Interface 7 (2010) S435–S450, https://doi.org/10.1098/rsif.2010.0125.focus.
[20] F. Saviñon-Flores, et al., A review on SERS-based detection of human virus infections: influenza and coronavirus, Biosensors 11 (3) (2021), https://doi.org/10.3390/bios11030066.
[21] J.D. Driskell, Y. Zhu, C.D. Kirkwood, Y. Zhao, R.A. Dluhy, R.A. Tripp, Rapid and sensitive detection of rotavirus molecular signatures using surface enhanced Raman spectroscopy (avr), PLOS ONE 5 (4) (2010), e10222, https://doi.org/10.1371/journal.pone.0010222.
[22] O. Ambartsumyan, D. Gribanyov, V. Kukushkin, A. Kopylov, E. Zavyalova, SERS-based biosensors for virus determination with oligonucleotides as recognition elements, Int. J. Mol. Sci. 21 (9) (2020), https://doi.org/10.3390/ijms21093373.
[23] V.I. Kukushkin, et al., Highly sensitive detection of influenza virus with SERS aptasensor (avr), PLOS ONE 14 (4) (2019), e0216247, https://doi.org/10.1371/journal.pone.0216247.
[24] A.I. Pérez-Jiménez, D. Lyu, Z. Lu, G. Liu, B. Ren, Surface-enhanced Raman spectroscopy: benefits, trade-offs and future developments, Chem. Sci. 11 (18) (2020) 4563–4577, https://doi.org/10.1039/D0SC00809E.
[25] S.E.J. Bell, et al., Towards reliable and quantitative surface-enhanced Raman scattering (SERS): from key parameters to good analytical practice, Angew. Chem. Int. Ed. 59 (14) (2020) 5454–5462, https://doi.org/10.1002/anie.201908154.
[26] L.T. Kerr, H.J. Byrne, B.M. Hennelly, Optimal choice of sample substrate and laser wavelength for Raman spectroscopic analysis of biological specimen, Anal. Methods 7 (12) (2015) 5041–5052, https://doi.org/10.1039/C5AY00327J.
[27] C.-S. Ho, et al., Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning, Nat. Commun. 10 (1) (2019), https://doi.org/10.1038/s41467-019-12898-9.
[28] C.M. Bishop, Pattern Recognition and Machine Learning. Accessed: May 29, 2023. [Online]. Available: ⟨https://link.springer.com/book/9780387310732⟩.
[29] M.E. Tipping, Sparse bayesian learning and the relevance vector machine, J. Mach. Learn. Res. 1 (2001) 211–244, https://doi.org/10.1162/15324430152748236.
[30] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra, Appl. Spectrosc. 43 (5) (1989) 772–777, https://doi.org/10.1366/0003702894202201.
[31] Abraham Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (8) (1964) 1627–1639, https://doi.org/10.1021/ac60214a047.
[32] P. Eilers, H. Boelens, Baseline correction with asymmetric least squares smoothing, Unpubl. Manuscr. (2005).
[33] S. He, et al., Baseline correction for Raman spectra using an improved asymmetric least squares method, Anal. Methods 6 (12) (2014) 4402–4407, https://doi.org/10.1039/C4AY00068D.
[34] M.E. Tipping, C.M. Bishop, Mixtures of probabilistic principal component analyzers, Neural Comput. 11 (2) (1999) 443–482, https://doi.org/10.1162/089976699300016728.
[35] T. Fushiki, Estimation of prediction error by using K-fold cross-validation, Stat. Comput. 21 (2) (2011) 137–146, https://doi.org/10.1007/s11222-009-9153-8.

[36] J. Liu, J. Sun, X. Huang, G. Li, B. Liu, Goldindec: a novel algorithm for Raman spectrum baseline correction, Appl. Spectrosc. 69 (7) (2015) 834–842, https://doi.org/10.1366/14-07798.
[37] W.K. Baek, et al., A comprehensive review of severe acute respiratory syndrome coronavirus 2, Cureus 12 (5) (2020), https://doi.org/10.7759/cureus.7943.
[38] M. Hasöksüz, S. Kiliç, F. Saraç, Coronaviruses and SARS-COV-2, Turk. J. Med. Sci. 50 (9) (2020) 549–556, https://doi.org/10.3906/sag-2004-127.
[39] C. Peteranderl, S. Herold, C. Schmoldt, Human influenza virus infections, Semin Respir. Crit. Care Med. 37 (04) (2016) 487–500, https://doi.org/10.1055/s-0036-1584801.
[40] E. Mauriz, Recent progress in plasmonic biosensing schemes for virus detection, Sensors 20 (17) (2020), https://doi.org/10.3390/s20174745.
[41] P.A. Mosier-Boss, Review on SERS of bacteria, Biosensors 7 (4) (2017), https://doi.org/10.3390/bios7040051.
[42] J.C. Ramirez-Perez, D. Durigo, Surface-enhanced Raman spectroscopy (SERS) for characterization SARS-CoV-2, J. Saudi Chem. Soc. 26 (5) (2022), 101531, https://doi.org/10.1016/j.jscs.2022.101531.
[43] D. Němeček, G.J. Thomas, Chapter 16 - Raman Spectroscopy of Viruses and Viral Proteins, in: J. Laane (Ed.), Frontiers of Molecular Spectroscopy, Elsevier, Amsterdam, 2009, pp. 553–595, https://doi.org/10.1016/B978-0-444-53175-9.00016-7.
[44] G. Pezzotti, et al., Raman molecular fingerprints of SARS-CoV-2 British variant and the concept of Raman barcode, Adv. Sci. (Weinh.) 9 (3) (2022), e2103287, https://doi.org/10.1002/advs.202103287.
[45] B.R. Wood, et al., Infrared based saliva screening test for COVID-19, Angew. Chem. Int Ed. Engl. 60 (31) (2021) 17102–17107, https://doi.org/10.1002/anie.202104453.
[46] J. Lukose, et al., Raman spectroscopy for viral diagnostics, Biophys. Rev. 15 (2) (2023) 199–221, https://doi.org/10.1007/s12551-023-01059-4.
[47] J. Ding, et al., Rapid identification of pathogens by using surface-enhanced Raman spectroscopy and multi-scale convolutional neural network, Anal. Bioanal. Chem. 413 (14) (2021) 3801–3811, https://doi.org/10.1007/s00216-021-03332-5.