Characterizing viral samples using machine learning for Raman and absorption spectroscopy

Miad Boodaghidizaji[1], Shreya Milind Athalye[2], Sukirt Thakur[1], Ehsan Esmaili[1], Mohit S. Verma[2,3,4,*], Arezoo M. Ardekani[1,*]

[1] School of Mechanical Engineering, Purdue University, 585 Purdue Mall, West Lafayette, IN 47907, USA
[2] Department of Agricultural and Biological Engineering, Purdue University, West Lafayette, IN 47907, USA
[3] Weldon School of Biomedical Engineering, Purdue University, West Lafayette, IN 47907, USA
[4] Birck Nanotechnology Center, Purdue University, West Lafayette, IN 47907, USA

[*]Corresponding authors: msverma@purdue.edu, ardekani@purdue.edu

Abstract

Machine learning methods can be used as robust techniques to provide invaluable information for analyzing biological samples in pharmaceutical industries, such as predicting the concentration of viral particles of interest in biological samples. Here, we utilized both convolutional neural networks and random forests to predict the concentration of the samples containing measles, mumps, rubella, and varicella-zoster viruses (ProQuad®) based on Raman and absorption spectroscopy. We prepared Raman and absorption spectra datasets with known concentration values, then used the Raman and absorption signals individually and together to train RFs and CNNs. We demonstrated that both RFs and CNNs can make predictions with R^2 values as high as 95%. We proposed two different networks to jointly use the Raman and absorption spectra, where our results demonstrated that concatenating the Raman and absorption data increases the prediction accuracy compared to using either Raman or absorption spectrum alone. Additionally, we further verified the advantage of using joint Raman-absorption with principal component analysis (PCA). Furthermore, our method can be extended to characterize properties other than concentration, such as the type of viral particles.

Keywords: convolutional neural networks, random forest, Raman spectroscopy, absorption spectroscopy, principal component analysis

1. Introduction

The recent outbreak of COVID-19 proved the importance of robust anti-viral medications to stop the spread of pandemic viral infections. Anti-viral drugs and vaccines are the two major solutions to keep viral infections at bay. A recent study suggested that, for example, the COVID-19 vaccine saved approximately 20 million human lives in one year [1]. Measles, mumps, rubella, and varicella (MMRV) are common viral childhood diseases that can have serious complications. Developing efficient methods to mass produce MMRV paves the way for limiting the spread of the MMRV globally. The vaccine development flourished in the early 20th century, and Maurice Hilleman at Merck & Co., Inc., a pioneer in the development of vaccinations, developed Rubeovax™ in 1968, the first commercial live vaccine for measles [2]. Vaccine development and production have been continuously improving in upstream and downstream processing [3]. Vaccine production involves challenging processes such as viral vector development, effective purification, polishing steps, and formulation with stable storage conditions. These processes require comprehensive and continuous quality management to maintain the product's efficacy and ensure public safety. With the advancement in viral vector-driven gene therapies and vaccine production, there is a growing interest in improving the continuous production of virus-like particle (VLP)-based vaccines [4]. The development of continuous manufacturing processes in the vaccine industry demands rapid, robust, and continuous analytical methods (Process analytical technology (PAT) tools) to understand real-time manufacturing processes [5].

Non-invasive in-line sensors such as Raman probes (Raman spectroscopy) hold great potential due to their higher sensitivity to read the molecular fingerprints of chemical and biological molecules, species, or products [6, 7]. Raman spectra possess clear spectral features that can be easily assigned to different chemical compounds. Additionally, minimal sample preparation is sufficient for making accurate quantitative predictions using Raman spectra [8]. In other words, Raman spectroscopy provides invaluable information for various analyte molecules even in ultra-low concentrations [9]. Similarly, absorption spectroscopy is a robust technique that, owing to its high sensitivity and large signal-to-noise ratio [10], has the potential to be implemented as a great tool to make predictions. Generally, both Raman and absorption spectra have been widely used for particle detection and identification [11–14] and quantitative analysis [15–17].

Recently machine learning (ML) has become popular for making predictions based on spectroscopy data. Both supervised and unsupervised ML techniques have been applied to Raman signals to make predictions [18]. Particularly, Raman spectroscopy has been utilized for cancer predictions [18]. For instance, techniques, such as principal component analysis or artificial neural networks have been utilized for detecting cervical cancer [19]. Furthermore, Raman signals have been utilized for classification problems, such as classifying bacteria [5, 20–22] viral [23, 24], and fungal infections [25, 26]. Additionally, Raman spectroscopy has been applied for regression purposes, such as predicting the concentration of the markers of interest, such as sensing the pH and Lactate in body fluids [27]. Absorption spectroscopy also has been utilized for classification purposes, such as the characterization of proteins [28] classification of wines [29], and quantifying the concentration of organic acids [30]. Furthermore, the joint Raman and absorption spectra have been applied to predict the values of concentrations [31].

Previous studies, in particular, have confirmed the capability of ML techniques in making quantitative predictions based on Raman or absorption signals. However, a comparison of

these signals and their strength in making accurate ML-based predictions for viral samples, such as MMRV has not been studied before. Here, we aim to create methods based on Raman and absorption spectroscopy that enables monitoring of the concentration of the viral particles in well plates. Additionally, it is not known whether using Raman and absorption spectra simultaneously can boost the prediction accuracy compared to using only Raman or absorption spectra separately. In our previous study, we demonstrated that deep learning enables the efficient detection of bacteria, fungi, and mammalian cells in static dried-down conditions [22]. Following our previous study, we intend to build convolutional neural networks (CNNs) and random forests (RFs) models that accept the Raman or absorption spectra or their combination as the input and predict the concentration of samples containing MMRV.

2. Materials & Methods

2.1 Data acquisition

All these samples prepared in this study are based on the ProQuad®, which is a sterile, lyophilized, preservative-free, live virus vaccine that contains measles, mumps, rubella, and varicella-zoster viruses [32]. We procured ProQuad® (manufactured by Merck & Co., Inc., West Point, PA) from the Purdue College of pharmacy and stored it at -20°C. We prepared the linear dilutions of the ProQuad® vaccine with a step size of 4 % and an initial concentration of 7.20E+05 plaque forming units/ml (PFU/mL) (Lyophilized ProQuad® + 10 µL Diluent). Throughout this article, we refer to the number of infective particles within the sample (PFU) as particles. All the Raman spectra of the ProQuad® dilutions were collected with the Renishaw in Via™Qontorconfocal Raman microscope (Renishaw plc, Wotton-under-Edge, UK) [33]. We used a 785-nm excitation laser with 100 % (300mW) power and 10 s acquisition time (1 accumulation). The spectral resolution of the spectra was 1 cm^{-1}, and the spectrum ranged from 101 to 3200 cm^{-1}corresponding to 3194 Raman shifts. The samples were focused with a 5X objective of a microscope (LeicaDM2700M), and three replicate Raman spectra were collected for each dilution. The sample volume used for the measurement was 100 µL, and the substrate used for the measurements was a 96-well plate (Corning™3635 UV-Transparent Microplates). The experiment was repeated once. The raw Raman spectral data was collected using WiRE 5.5 software. Furthermore, we collected the absorption spectrum for ProQuad® dilutions using the BMG LABTECH, Inc microplate reader (CLARIOstar Plus, SN: 430-2173). The spectrum range was 220 nm to 1000 nm with a spectral resolution of 1 nm wavelength corresponding to 781 wavelengths. The sample volume used for the measurement was 100 µL, and the substrate used for the measurements was a 96-well plate (Corning™3635 UV-Transparent Microplates). We collected three spectral scans for each dilution. The experiment was repeated once. In total, the dataset includes Raman and absorption spectra for 25 different concentration values with three to six replicates for each value, making a total of 116 samples, where 20 % of this data is used for testing by 5-fold cross-validation as described in section 2.2.

2.2 Machine learning modeling

We adopt two widely used ML techniques to relate the Raman and absorption spectra to the concentration values: the random forest (RF) and the convolutional neural network (CNN) techniques. Before training, to ensure the reproducibility of the results, all the models are initialized by setting the seed number to zero. To assess the accuracy of predictions, we use the values of the coefficient of determination (R^2 scores). Further, to train the models, the 5-fold cross-validation technique is used both for the CNNs and RFs. In this method, the whole data is split into five sections, where the model is trained five times, and each time four sections are used as the training dataset and one section as the testing dataset. The 5-fold cross-validation model ensures that all the data points fall into the testing dataset at least once, preventing biased predictions. The Sklearn [34] and Pytorch [35] modules in Python are used for modeling the RFs and CNNs, respectively.

Convolutional neural network (CNN) is a supervised machine learning technique that, in our case, takes one-dimensional signals as the input and identifies the important parts of the signal, which paves the way for automatic learning of various features and hidden aspects in the signal that are important for the regression. In other words, CNN can capture the spatial and temporal dependencies in the Raman or absorption spectrum. The general architectures of the deep learning models used in this study are similar, i.e., a feed-forward single CNN consisting of four convolutional layers followed by four fully connected layers when either Raman or absorption spectrum is used as the input, as shown in Fig 1 a. However, when it comes to using both the Raman and absorption spectra as the input, we use two different designs. In one design, we concatenate the Raman and absorption signals and feed them into a single CNN, as shown in Fig 1 a. In another design, a double CNN is created for feeding the inputs, as demonstrated in Fig 1 b. In the double CNN, the Raman and absorption spectrum are first fed into two separate networks with four convolutional layers and then two fully connected layers. Eventually, the outputs of each network are concatenated and fed into a network with two fully connected layers. In all models, the architecture used for convolutional layers is based on residual mapping following the deep residual learning method [36]. The presence of residual blocks with shortcut connections between inputs and outputs boosts the training stability and paves the way for having deeper layers [36].

Furthermore, the kernel size used for all the convolutional layers is three with zero paddings and strides of one. Additionally, all the networks are trained for 6000 epochs (iterations), where a further increase in the epochs does not significantly boost the prediction accuracy. We use the mean squared loss function as the criterion for training with the back-propagation techniques, where we adopt the stochastic gradient descent with momentum and adaptive learning rate, Adam [37], where the weight decay and learning rate are set to 0.1 and 10^{-8}, respectively. Batch normalization and ReLU activation functions are applied consecutively at the end of each convolutional layer, and the ReLU function is applied at the end of each fully connected layer. After passing the last ReLU function, the data is mapped into one neuron as the output. The number of channels and neurons are hyperparameters that can be tuned for further accuracy. In the current study, we found that a maximum of 10 channels and 4000 neurons leads to sufficient accuracy while at the same time avoiding over-fitting.

Random Forest (RF) regression is a supervised machine-learning technique that utilizes the ensemble average of multiple decision trees to make final predictions [38]. Each one of the trees makes its prediction of the concentration. As shown in Fig 2, the Raman, absorption, or their concatenated spectrum is used as the input with the concentration as the output. RF is a powerful regression technique that runs efficiently on larger datasets. RFs are generally suitable for making predictions in the training range. Additionally, we use the bootstrapping technique, where we select multiple training samples from the original training sample, and

these different samples are used for training each one of these decision trees. Bootstrapping reduces over-fitting chances and stabilizes the network. The squared error criterion in scikit-learn [34] is used to measure the quality of splitting for 100 trees.

3. Results and discussions

We use CNN and RF as two powerful ML techniques, with different levels of preprocessing to identify the optimum predictions. Here, we discuss how the algorithms work with the test data generated using 5-fold cross-validation, where each fold can contain points both inside and outside of the training ranges. For CNN, we discuss whether a single or double CNN works better when both Raman and absorption spectra are used as the input.

In this study, CNN models are composed of multiple convolutional layers with a kernel size of 3, where, in each layer, by convolving around the signal, hidden features and patterns are learned. To expedite the learning process and improve the model performance, it is beneficial to preprocess the data before training the models. Thus, we apply baseline corrections and normalize the data using the standard normal variate method, i.e., subtracting each spectrum by its mean value and dividing by the standard deviation described by Romer et al. [39]. Additionally, normalizing the Raman spectrum makes intensities of the Raman and absorption spectrum to be approximately in the same order for further comparison. No baseline correction or normalization is required for the absorption spectrum since the difference between maximum and minimum values is relatively low. Additionally, normalizing the absorption data led to no significant boost in prediction accuracy. Finally, the Raman and absorption signals are smoothened using the Savitzky-Golay (SG) filter [39, 40]. Fig 3 demonstrates the Raman and absorption spectra before and after preprocessing for two different concentrations. In addition to normalization and applying filters, some studies trim the Raman spectrum to obtain the spectral range of interest [8]. In the current study, we did not observe any significant gain in the prediction accuracy when the Raman or absorption spectrum is trimmed, as we have shown, for example, for the RF method in the Appendix. Additionally, we analyzed how the predictions change with the subtraction of the control spectrum of solvent as described in the Appendix, where we noticed a reduction in the accuracy with the subtraction of the control spectrum. Therefore, we excluded the subtraction of the control spectrum step from preprocessing steps.

The R^2 coefficients for the values of the 5-fold predictions for both RF and CNN are listed in Table. 1. The average R^2 score for all the predictions is above 90%. However, the prediction accuracy is higher when the concatenated Raman-absorption spectrum is used for RF and CNN compared to the predictions based on either Raman or absorption spectrum. Furthermore, the prediction accuracy is slightly higher for RF compared to CNN in the hyperparameter space we studied. However, both RF and CNN lead to predictions with R^2 values as high as 98% when the joint Raman-absorption data is used. Additionally, we note that the single CNN demonstrates higher prediction accuracy compared to the double CNN, which might be attributed to the low predictability of the absorption spectrum compared to the Raman spectrum, as the prediction accuracy is higher when only Raman is used compared n to when only absorption data is used. Furthermore, we have made a comparison between RF and support vector machine (SVM) methods in Appendix, where we note that RF predictions are slightly more accurate than SVM. Additionally, we visually demonstrate how the predictions of CNN and RF vary for the testing dataset in one of the folds in the 5-fold dataset. As demonstrated in Fig 4, prediction values based on the Raman spectrum are more in line with the actual values as opposed to the absorption spectrum, where the average R^2 coefficient is lower. This difference can be attributed to the larger size of the Raman signal and, therefore,

larger regions of dissimilarity corresponding to different concentrations, which make Raman spectra more distinguishable from each other. Further, the use of joint Raman-absorption spectra boosts the prediction accuracy compared to the case when only Raman spectra are used.

The differences between the prediction accuracy of the Raman and the absorption spectra can further be understood through the principal component analysis (PCA). We use PCA to reduce the dimensionality of the Raman, absorption, and concatenated Raman-absorption spectra to 4, where the original size of the Raman and absorption spectra are 3194 and 781. Fig 5 demonstrates how principal coordinate (PC) values differ at different concentration values. The distinction between PCA points at different concentrations is more evident for the Raman-absorption spectrum as compared to the Raman or absorption spectrum. Additionally, we notice that for most cases, not only the prediction accuracy does not increase by conducting PCA, but also for the Raman and Raman-absorption data, the average R^2 values slightly decrease when we compare Fig 5 d, and Fig 4. Therefore, for the current dataset, dimensionality reduction does not improve the prediction accuracy.

4. Conclusion

In the current study, the possibility of using absorption, Raman, and joint Raman-absorption spectrum to determine the concentration of the samples containing viral particles was investigated. RF and CNN, as two different machine learning algorithms, were utilized for making predictions, and the prediction accuracy was monitored using 5-fold cross-validation. We demonstrated that with sufficient preprocessing, both the Raman and absorption spectra could be used to create a surrogate to predict the values of concentration. In most cases, the Raman spectrum leads to more accurate predictions compared to the absorption spectrum. Moreover, concatenating Raman and absorption spectra improves the prediction accuracy both for RF and CNN. Furthermore, PCA analysis sheds light on the advantage of joint spectra over single usage of Raman or absorption spectrum as the points corresponding to different concentrations are further separated. We have demonstrated that the joint utilization of the Raman and absorption spectra paves the way for the real-time measurements of the concentration of the viral particles in well plates, which can be extended to different static and dynamics settings, such as microfluidic devices with different flow conditions.

The key limitations of this study can be listed as follows. a) the predictions, in general, work well when the unknown concentration values lie in the range of training datasets. Given that here we focused on relatively large concentration values ($> 4 \times 105$/ml), the predictions for the low concentration values ($<< 4 \times 105$/ml) are not reliable. b) the predictions are valid only for ProQuad® samples. Further training data points corresponding to different types of viral particles are required to extend the applicability of the current method.

In future studies, we intend to extend the range of predictions and develop a graphical user interface, which accepts the raw Raman and absorption data as the input and predicts the values of concentrations for different ML methods. Indeed, the current study can serve as a basic block for developing completely automated software that can capture the values of concentration for different types of viral particles using different machine learning algorithms. Furthermore, we aim to extend the predictions to include Raman spectroscopy in microfluidics under different flow conditions.

Author contributions

Miad Boodaghidizaji: Conceptualization-Equal, Data curation-Lead, Formal analysis-Equal, Methodology-Lead, Software-Lead, Visualization-Lead, Writing – original draft-Lead, Writing – review & editing-Lead

Shreya Milind Athalye: Conceptualization-Equal, Data curation-Lead, Formal analysis-Equal, Visualization-Lead, Writing – original draft-Lead, Writing – review & editing-Supporting

Sukirt Thakur: Conceptualization-Equal, Formal analysis-Equal, Investigation-Equal, Methodology-Lead, Software-Supporting, Writing – original draft-Supporting

Ehsan Esmaili: Conceptualization-Equal, Data curation-Lead, Formal analysis-Equal, Investigation-Equal, Writing – original draft-Supporting

Mohit S. Verma: Conceptualization-Equal, Funding acquisition-Supporting, Project administration-Lead, Supervision-Lead, Writing – review & editing-Lead

Arezoo M. Ardekani: Conceptualization-Equal, Funding acquisition-Lead, Project administration-Lead, Supervision-Lead, Writing – review & editing-Lead

Acknowledgments

This work was performed under a Project Award Agreement from the National Institute for Innovation in Manufacturing Biopharmaceuticals (NIIMBL) and financial assistance award 70NANB21H085 from the U.S. Department of Commerce, National Institute of Standards and Technology. Miad Boodaghidizaji and Shreya Milind Athalye contributed equally to this work.

Conflict of interest

None declared

Ethics statement

None required

Data availability statement

The datasets generated and/or analyzed during the current study are available in the Mendeley data repository at http://dx.doi.org/10.17632/44sgp2jvj5.1

References

[1] O. J. Watson, G. Barnsley, J. Toor, A. B. Hogan, P. Winskill, and A. C. Ghani, "Global impact of the first year of covid-19 vaccination: a mathematical modelling study," The Lancet Infectious Diseases, 2022.

[2] T. H. Tulchinsky, "Maurice hilleman: Creator of vaccines that changed the world," Case Studies in Public Health, p. 443, 2018.

[3] J. T. Blue, J. R. Sinacola, and A. Bhambhani, "Process scale-up and optimization of lyophilized vaccine products," in Lyophilized Biologics and Vaccines, pp. 179–210, Springer, 2015.

[4] S. Guti´errez-Granados, F. G`odia, and L. Cervera, "Continuous manufacturing of viral particles," Current opinion in chemical engineering, vol. 22, pp. 107–114, 2018.

[5] M. K. Maruthamuthu, S. R. Rudge, A. M. Ardekani, M. R. Ladisch, and M. S. Verma, "Process analytical technologies and data analytics for the manufacture of monoclonal antibodies," Trends in biotechnology, vol. 38, no. 10, pp. 1169–1186, 2020.

[6] H. J. Butler, L. Ashton, B. Bird, G. Cinque, K. Curtis, J. Dorney, K. Esmonde-White, N. J. Fullwood, B. Gardner, P. L. Martin-Hirsch, et al., "Using raman spectroscopy to characterize biological materials," Nature protocols, vol. 11, no. 4, pp. 664–687, 2016.

[7] L. Rolinger, M. R¨udt, and J. Hubbuch, "A critical review of recent trends, and a future perspective of optical spectroscopy as pat in biopharmaceutical downstream processing," Analytical and bioanalytical chemistry, vol. 412, no. 9, pp. 2047–2064, 2020.

[8] F. Pian, Q. Wang, M. Wang, P. Shan, Z. Li, and Z. Ma, "A shallow convolutional neural network with elastic nets for blood glucose quantitative analysis using raman spectroscopy," Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, vol. 264, p. 120229, 2022.

[9] R. Panneerselvam, H. Sadat, E.-M. H¨ohn, A. Das, H. Noothalapati, and D. Belder, "Microfluidics and surface-enhanced raman spectroscopy, a win–win combination?," Lab on a Chip, vol. 22, no. 4, pp. 665– 682, 2022.

[10] S. B. Torrisi, M. R. Carbone, B. A. Rohr, J. H. Montoya, Y. Ha, J. Yano, S. K. Suram, and L. Hung, "Random forest machine learning models for interpretable x-ray absorption near-edge structure spectrum-property relationships," npj Computational Materials, vol. 6, no. 1, pp. 1–11, 2020.

[11] J. Probst, C. N. Borca, M. A. Newton, J. van Bokhoven, T. Huthwelker, S. Stavrakis, and A. deMello, "In situ x-ray absorption spectroscopy and droplet-based microfluidics: An analysis of calcium carbonate precipitation," ACS Measurement Science Au, vol. 1, no. 1, pp. 27–34, 2021.

[12] 5A. Nitkowski, L. Chen, and M. Lipson, "Cavity-enhanced on-chip absorption spectroscopy using microring resonators," Optics express, vol. 16, no. 16, pp. 11930–11936, 2008.

[13] A. Pallaoro, M. R. Hoonejani, G. B. Braun, C. D. Meinhart, and M. Moskovits, "Rapid identification by surface-enhanced raman spectroscopy of cancer cells at low

concentrations flowing in a microfluidic channel," Acs Nano, vol. 9, no. 4, pp. 4328–4336, 2015.

[14] S. E. Barnes, Z. T. Cygan, J. K. Yates, K. L. Beers, and E. J. Amis, "Raman spectroscopic monitoring of droplet polymerization in a microfluidic device," Analyst, vol. 131, no. 9, pp. 1027–1033, 2006.

[15] C. J. Strachan, T. Rades, K. C. Gordon, and J. Rantanen, "Raman spectroscopy for quantitative analysis of pharmaceutical solids," Journal of pharmacy and pharmacology, vol. 59, no. 2, pp. 179–192, 2007.

[16] W.-J. Bao, J. Li, J. Li, Q.-W. Zhang, Y. Liu, C.-F. Shi, and X.-H. Xia, "Au/znse-based surface enhanced infrared absorption spectroscopy as a universal platform for bioanalysis," Analytical chemistry, vol. 90, no. 6, pp. 3842–3848, 2018.

[17] E. E. Storey and A. S. Helmy, "Optimized preprocessing and machine learning for quantitative raman spectroscopy in biology," Journal of Raman Spectroscopy, vol. 50, no. 7, pp. 958–968, 2019.

[18] N. M. Ralbovsky and I. K. Lednev, "Towards development of a novel universal medical diagnostic method: Raman spectroscopy and machine learning," Chemical Society Reviews, vol. 49, no. 20, pp. 7428–7453, 2020.

[19] A. Daniel, A. Prakasarao, and S. Ganesan, "Near-infrared raman spectroscopy for estimating biochemical changes associated with different pathological conditions of cervix," Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, vol. 190, pp. 409–416, 2018.

[20] S. Khan, R. Ullah, S. Shahzad, N. Anbreen, M. Bilal, and A. Khan, "Analysis of tuberculosis disease through raman spectroscopy and machine learning," Photodiagnosis and photodynamic therapy, vol. 24, pp. 286–291, 2018.

[21] S. K. Koya, M. Brusatori, J. V. Martin, S. Yurgelevic, C. Huang, D. M. Liberati, G. W. Auner, and L. N. Diebel, "Rapid detection of clostridium difficile toxins in serum by raman spectroscopy," Journal of Surgical Research, vol. 232, pp. 195–201, 2018.

[22] M. K. Maruthamuthu, A. H. Raffiee, D. M. De Oliveira, A. M. Ardekani, and M. S. Verma, "Raman spectra-based deep learning: A tool to identify microbial contamination," MicrobiologyOpen, vol. 9, no. 11, p. e1122, 2020.

[23] A. Ditta, H. Nawaz, T. Mahmood, M. Majeed, M. Tahir, N. Rashid, M. Muddassar, A. Al-Saadi, and H. Byrne, "Principal components analysis of raman spectral data for screening of hepatitis c infection," Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, vol. 221, p. 117173, 2019.

[24] D. Tong, C. Chen, J. Zhang, G. Lv, X. Zheng, Z. Zhang, and X. Lv, "Application of raman spectroscopy in the detection of hepatitis b virus infection," Photodiagnosis and photodynamic therapy, vol. 28, pp. 248–252, 2019.

[25] Z. Guo, M. Wang, A. O. Barimah, Q. Chen, H. Li, J. Shi, H. R. El-Seedi, and X. Zou, "Label-free surface enhanced raman scattering spectroscopy for discrimination and detection of dominant apple spoilage fungus," International Journal of Food Microbiology, vol. 338, p. 108990, 2021.

[26]S. Dzurendov´a, V. Shapaval, V. Tafintseva, A. Kohler, D. Byrtusov´a, M. Szotkowski, I. M´arov´a, and B. Zim-mermann, "Assessment of biotechnologically important filamentous fungal biomass by fourier transform raman spectroscopy," International Journal of Molecular Sciences, vol. 22, no. 13, p. 6710, 2021.

[27]I. Olaetxea, A. Valero, E. Lopez, H. Lafuente, A. Izeta, I. Jaunarena, and A. Seifert, "Machine learning-assisted raman spectroscopy for ph and lactate sensing in body fluids," Analytical Chemistry, vol. 92, no. 20, pp. 13888–13895, 2020.

[28]J. Zhang, S. Ye, K. Zhong, Y. Zhang, Y. Chong, L. Zhao, H. Zhou, S. Guo, G. Zhang, B. Jiang, et al., "A machine-learning protocol for ultraviolet protein-backbone absorption spectroscopy under environmental fluctuations," The Journal of Physical Chemistry B, vol. 125, no. 23, pp. 6171–6178, 2021.

[29]A. Philippidis, E. Poulakis, R. Kontzedaki, E. Orfanakis, A. Symianaki, A. Zoumi, and M. Velegrakis, "Application of ultraviolet-visible absorption spectroscopy with machine learning techniques for the classi-fication of cretan wines," Foods, vol. 10, no. 1, p. 9, 2020.

[30]C. Wolf, D. Gaida, A. Stuhlsatz, T. Ludwig, S. McLoone, and M. Bongards, "Predicting organic acid con-centration from uv/vis spectrometry measurements–a comparison of machine learning techniques," Trans-actions of the Institute of Measurement and Control, vol. 35, no. 1, pp. 5–15, 2013.

[31]I. Isaev, N. Trifonov, O. Sarmanova, S. Burikov, T. Dolenko, K. Laptinskiy, and S. Dolenko, "Joint applica-tion of raman and optical absorption spectroscopy to determine concentrations of heavy metal ions in water using artificial neural networks," in Saratov Fall Meeting 2019: Laser Physics, Photonic Technologies, and Molecular Modeling, vol. 11458, p. 114580R, International Society for Optics and Photonics, 2020.

[32]B. J. Kuter, M. L. Hoffman Brown, J. Hartzel, W. R. Williams, K. A. Eves, S. Black, H. Shinefield, K. S.
Reisinger, C. D. Marchant, B. J. Sullivan, et al., "Safety and immunogenicity of a combination: Measles, mumps, rubella and varicella vaccine (proquad®)," Human Vaccines, vol. 2, no. 5, pp. 205–214, 2006.

[33]RENISHAW, "https://www.renishaw.com/en/raman-software–9450."

[34]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer,
R. Weiss, V. Dubourg, et al., "Scikit-learn: Machine learning in python," the Journal of machine Learning research, vol. 12, pp. 2825–2830, 2011.

[35]A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein,
L. Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," Advances in neural information processing systems, vol. 32, 2019.

[36]K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

[37]D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[38]U. Gr¨omping, “Variable importance assessment in regression: linear regression versus random forest,” The American Statistician, vol. 63, no. 4, pp. 308–319, 2009.

[39]S. Romero-Torres, J. D. P´erez-Ramos, K. R. Morris, and E. R. Grant, “Raman spectroscopy for tablet coat-ing thickness quantification and coating characterization in the presence of strong fluorescent interference,” Journal of pharmaceutical and biomedical analysis, vol. 41, no. 3, pp. 811–819, 2006.

[40]N. Gonz´alez-Viveros, P. G´omez-Gil, J. Castro-Ramos, and H. Cerecedo-N´u˜nez, “On the estimation of sugars concentrations using raman spectroscopy and artificial neural networks,” Food Chemistry, vol. 352, p. 129375, 2021.

[41]K. Ember, F. Daoust, M. Mahfoud, F. Dallaire, E. Z. Ahmad, T. Tran, A. Plante, M.-K. Diop, T. Nguyen,

A. St-Georges-Robillard, et al., “Saliva-based detection of covid-19 infection in a real-world setting us-ing reagent-free raman spectroscopy and machine learning,” Journal of biomedical optics, vol. 27, no. 2, p. 025002, 2022.

Table 1: The R^2 values of 5-fold cross-validation for the prediction of concentration for given Raman, absorption, and Raman-Absorption concatenated spectra

Fold	Absorption	Raman	Raman-Absorption	Concatenated R-A	Absorption	Raman	Raman-Absorption
1	0.975	0.976	0.981	0.992	0.977	0.974	0.988
2	0.948	0.973	0.980	0.974	0.974	0.983	0.995
3	0.942	0.980	0.977	0.992	0.964	0.979	0.991
4	0.835	0.964	0.931	0.946	0.881	0.953	0.981
5	0.952	0.984	0.980	0.991	0.989	0.982	0.990
Ave	0.930	0.975	0.969	0.979	0.955	0.974	0.989

Figure legends

Fig 1: Schematic view of the neural network structure used when a) Raman, absorption, or concatenated Raman-absorption spectrum is used as the input b) both Raman and absorption spectra are used as separate inputs. The number of layers shown is for illustration purposes and does not reflect the actual values.

Fig 2: Schematic view of the RF composed of multiple decision trees with either Raman, absorption, or concatenated Raman-absorption spectrum as the input. ``A" stands for the average. The number of nodes and trees shown are for illustration purposes and do not reflect the actual values.

Fig 3: Raw and preprocessed Raman and absorption plots at two different concentrations

Fig 4 a: Comparison of the predictions of the RF and CNN for one fold in the 5-fold cross-validation datasets when the Raman spectrum is used as the input

Fig 4 b: Comparison of the predictions of the RF and CNN for one fold in the 5-fold cross-validation datasets when absorption spectrum is used as the input

Fig 4 c: Comparison of the predictions of the RF and CNN for one fold in the 5-fold cross-validation datasets when the Raman-absorption spectrum is used as the input in the single network

Fig 4 d: Comparison of the predictions of the RF and CNN for one fold in the 5-fold cross-validation datasets when the Raman-absorption spectrum is used as the input in the double network

Fig 5 a: Comparison of the PCA plots at different concentration values for the Raman data.

Fig 5 b: Comparison of the PCA plots at different concentration values for the absorption data.

Fig 5 c: Comparison of the PCA plots at different concentration values for the concatenated Raman-absorption data.

Fig 5 d: Comparison of RF predictions with dimensionality reduction using PCA for different types of inputs. R-A stands for the concatenated Raman-absorption.

Appendix

Several ML algorithms can be used for making predictions using Raman and absorption spectra. Here, we present how the predictions might be different if one chooses a different ML algorithm, such as the support vector machine (SVM). Table A1 demonstrates the R^2 coefficients for the values of the 5-fold predictions for the SVM method. The results are very similar to the values listed for the RF method for the Raman and the concatenated Raman-absorption spectrum. However, for the absorption spectrum, we note that the RF predictions are more accurate than SVM.

Table A1: The R^2 values of 5-fold cross-validation for the prediction of concentration for the Raman, absorption, and concatenated Raman-absorption spectrum using the SVM method

	SVM		
Fold	Absorption	Raman	Raman-Absorption
1	0.664	0.980	0.991
2	0.698	0.984	0.990
3	0.777	0.981	0.989
4	0.601	0.968	0.980
5	0.621	0.984	0.993
Ave	0.672	0.979	0.988

Furthermore, we demonstrate how the prediction R^2 values change if the input dimension is reduced to 4 using PCA for the RF and SVM methods. As shown in Table A2, we note that dimensionality reduction, in this case, significantly decreases the R^2 values, particularly for the SVM method.

Table A2: The R^2 values of 5-fold cross-validation for the prediction of concentration for the Raman, absorption, and concatenated Raman-absorption spectrum using PCA

	RF			SVM		
Fold	Absorption	Raman	Raman-Absorption	Absorption	Raman	Raman-Absorption
1	0.958	0.926	0.933	0.664	0.731	0.701
2	0.972	0.935	0.960	0.698	0.746	0.790
3	0.949	0.951	0.940	0.777	0.623	0.578
4	0.873	0.914	0.973	0.601	0.758	0.793
5	0.973	0.957	0.973	0.621	0.695	0.538
Ave	0.945	0.937	0.956	0.672	0.710	0.680

The background noise can affect the Raman and absorption spectra, particularly at low Raman shifts and wavelengths. As a result, in this section, we remove the initial parts of the Raman (Raman shift < 300 cm^{-1}) and absorption spectrum (λ < 250 nm). As shown in Table A3, we note that the prediction accuracies do not change significantly with trimming. Therefore, we used the entire spectra for prediction. Indeed, one of the advantages of using machine learning techniques is that these techniques automatically detect which part of the signal is important. Fig A1 demonstrates the values of importance for the Raman and absorption spectra before and after trimming. The importance values are obtained automatically from the Sklearn importance attribute for the RF method [34]. As evident, we do not notice any significant shift in the important regions of the signals.

Table A3: The R^2 values of 5-fold cross-validation for the prediction of concentration for the trimmed Raman, absorption, and concatenated Raman-absorption spectrum using the RF method

	RF		
Fold	Absorption	Raman	Raman-Absorption
1	0.958	0.971	0.987
2	0.972	0.982	0.991
3	0.949	0.967	0.975
4	0.873	0.961	0.955
5	0.973	0.980	0.988
Ave	0.945	0.972	0.979

In this study, we did not subtract the control spectrum of the solvent (sterile water) from the Raman and absorption spectrum to minimize the amount of preprocessing. Here, we demonstrate how the prediction accuracies change if we subtract the control data from all the spectrums. Fig A2 demonstrates the comparison of the Raman and absorption spectrum for samples that contain viral particles. As evident, the presence of viral particles induces noticeable changes at most Raman shifts. Further, the absorption signal at all wavelengths is different when viral particles are introduced. Additionally, we presented the spectrum with the water data subtracted. Table A4 demonstrates R^2 values for predictions of the RF method using the spectrum with water data subtracted. We note that the R^2 values decrease with the subtraction of water data compared to the values presented in Table A3. Therefore, we excluded the subtraction of the water spectrum step in the preprocessing.

Table A4: The R^2 values of 5-fold cross-validation for the prediction of concentration for the trimmed Raman, absorption, and concatenated Raman-absorption spectrum using the RF method with control data being subtracted

	RF		
Fold	Absorption	Raman	Raman-Absorption
1	0.975	0.920	0.981
2	0.982	0.983	0.991
3	0.971	0.931	0.948
4	0.882	0.866	0.917
5	0.989	0.947	0.974
Ave	0.959	0.972	0.962

Appendix Fig A1 a: Importance of each point in the Raman shift in the regression based on the RF method

Appendix Fig A1 b: Importance of each point in the wavelength in the regression based on the RF method

Appendix Fig A2: Preprocessed Raman, absorption, and the control (sterile water) spectrum plots in addition to plots with background subtracted

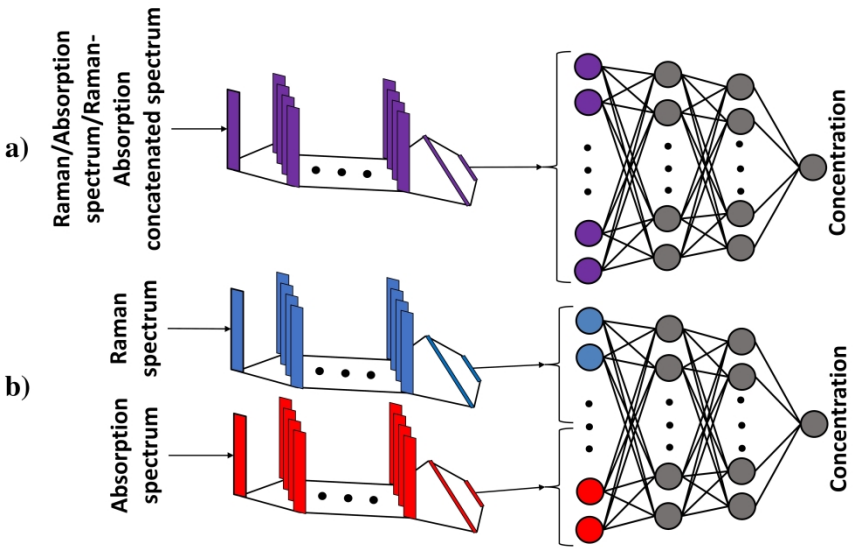


Fig 1

177x100mm (1200 x 1200 DPI)

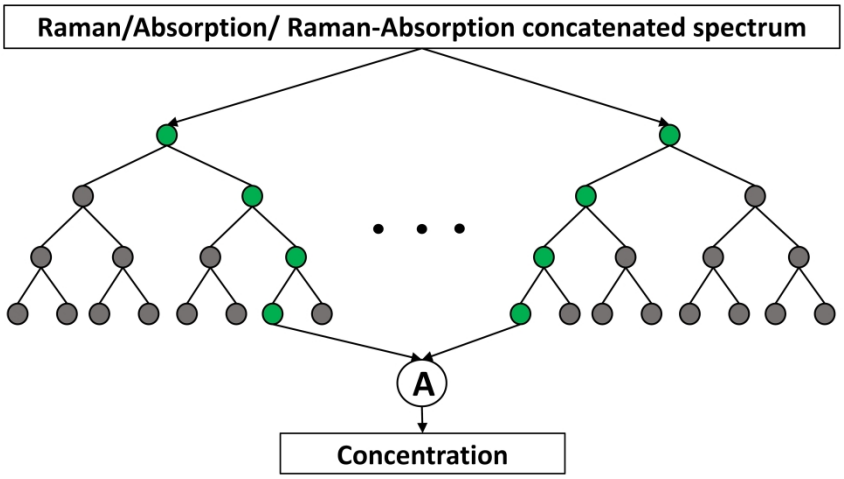


Fig 2

177x100mm (1200 x 1200 DPI)

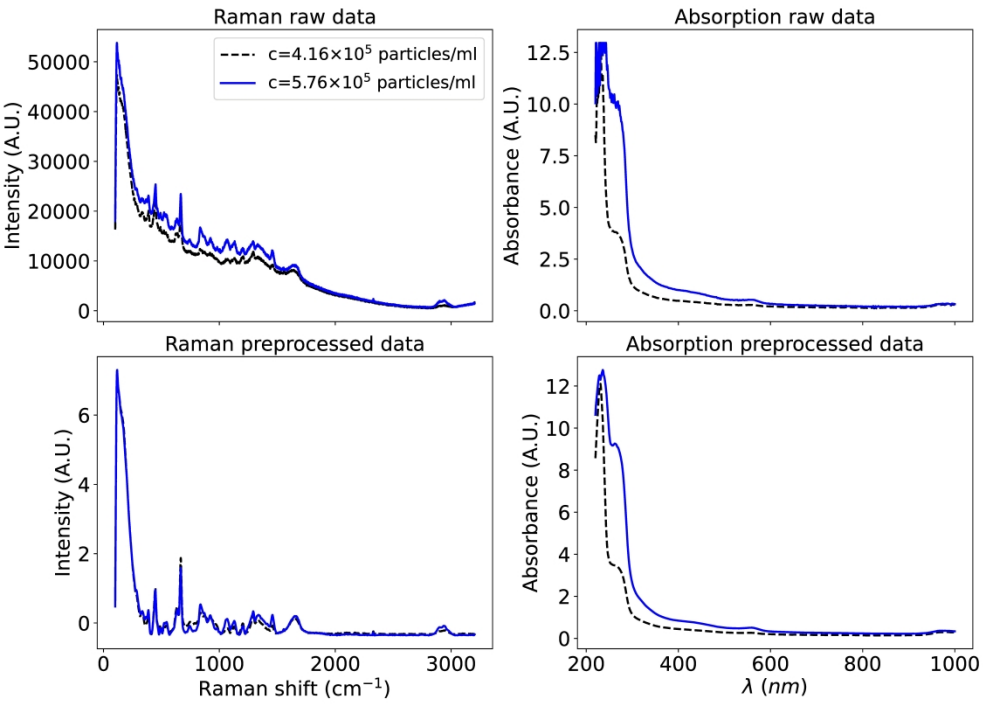


Fig 3

355x254mm (600 x 600 DPI)

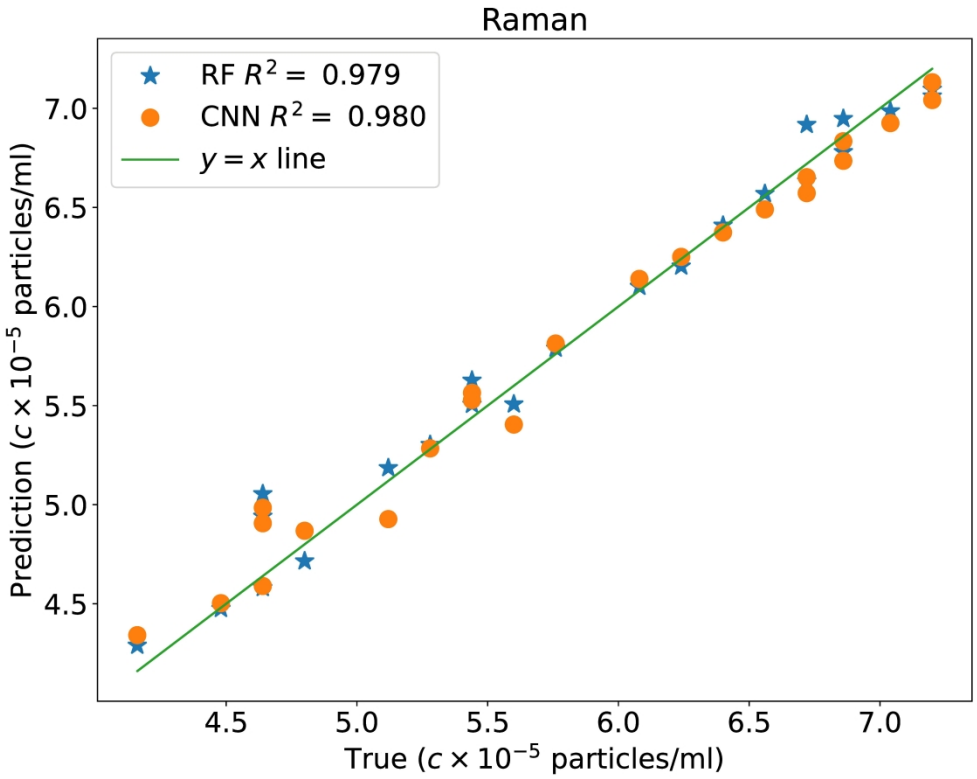


Fig 4a

254x203mm (600 x 600 DPI)

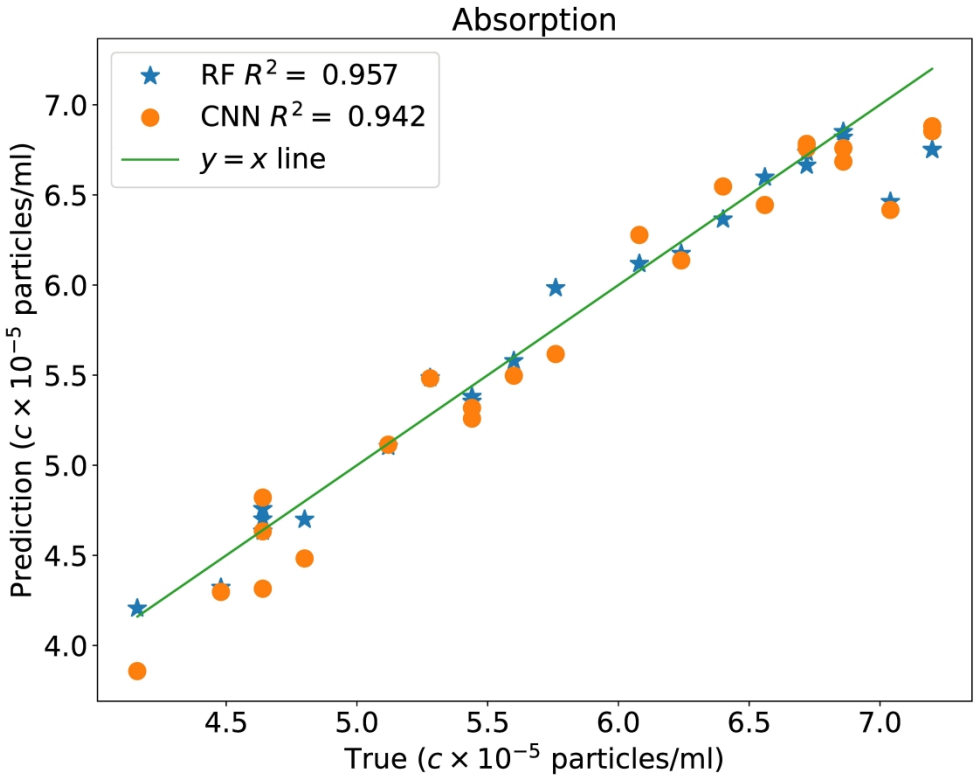


Fig 4b

254x203mm (600 x 600 DPI)

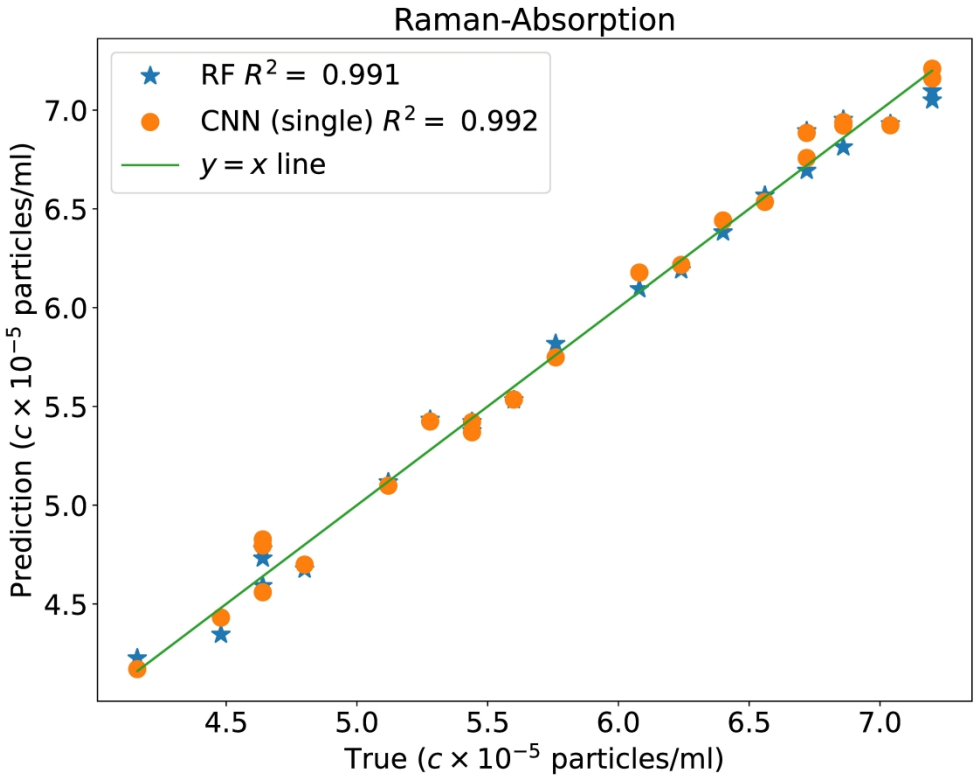


Fig 4c

254x203mm (600 x 600 DPI)

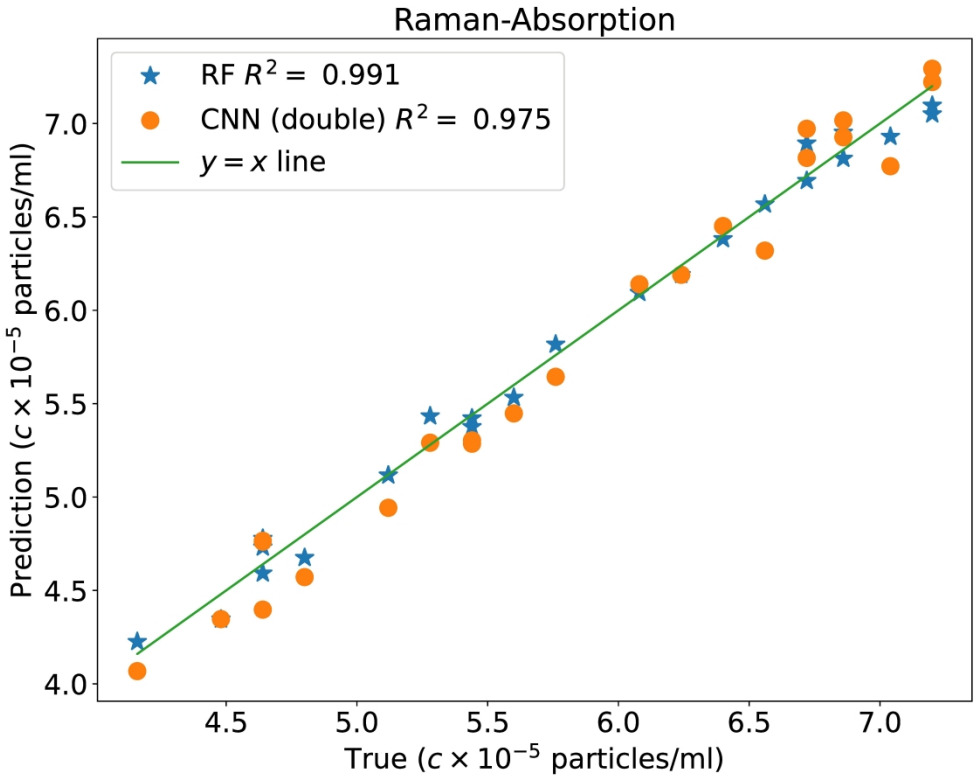


Fig 4d

254x203mm (600 x 600 DPI)

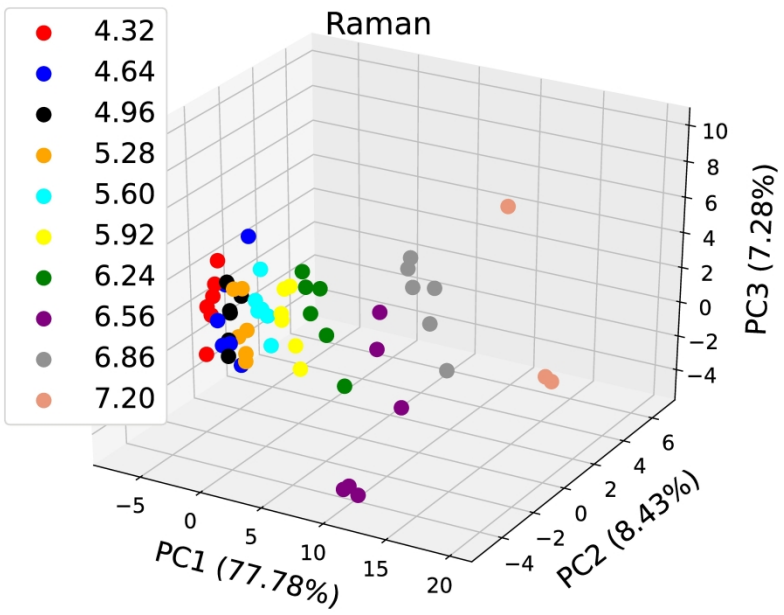


Fig 5a

254x203mm (600 x 600 DPI)

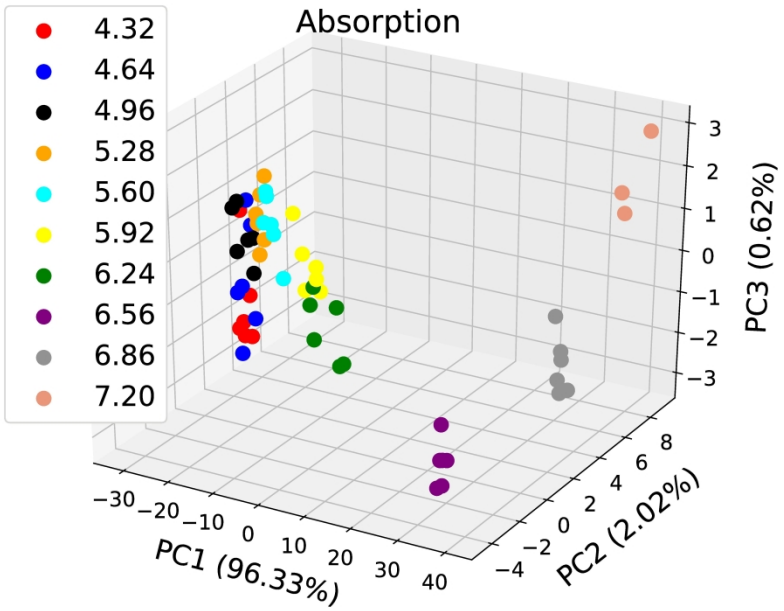


Fig 5b

254x203mm (600 x 600 DPI)

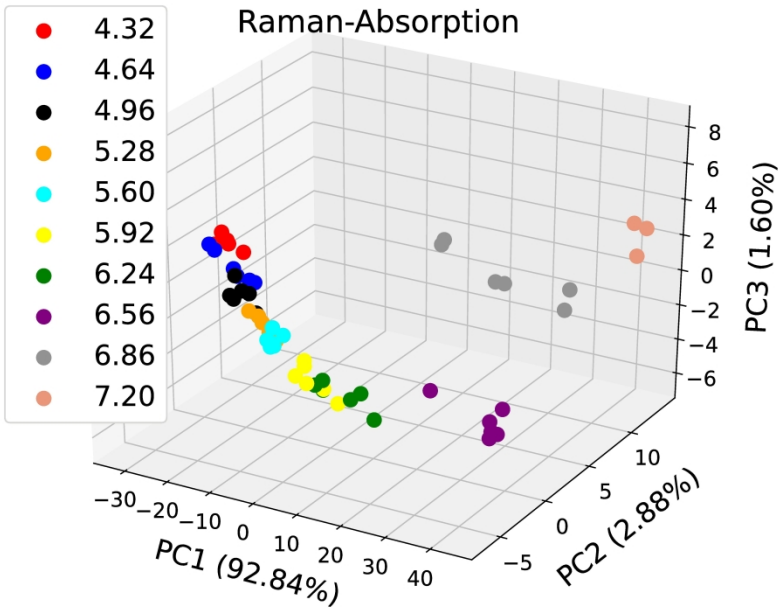


Fig 5c

254x203mm (600 x 600 DPI)

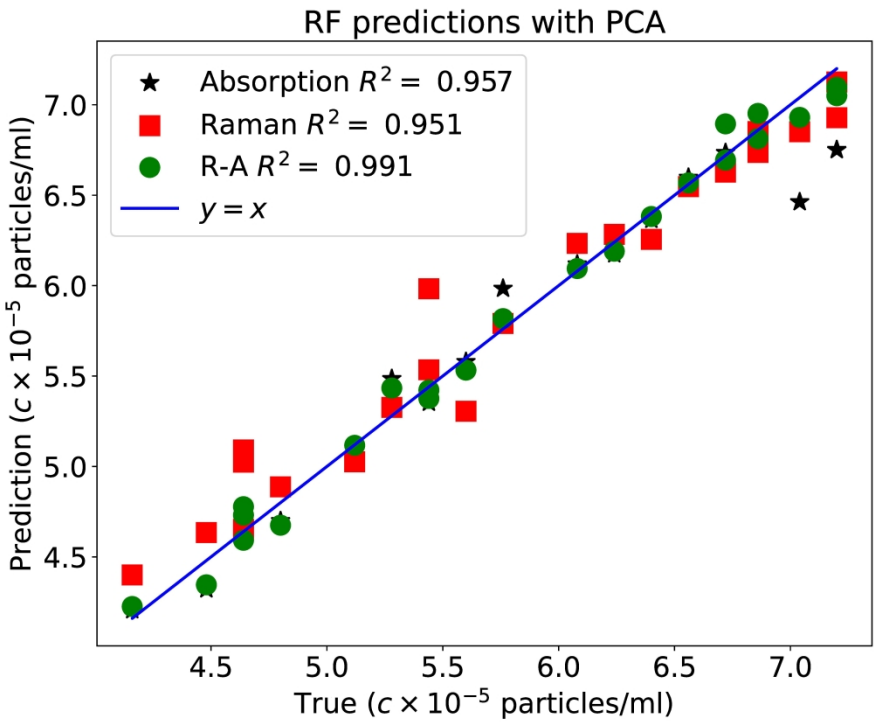
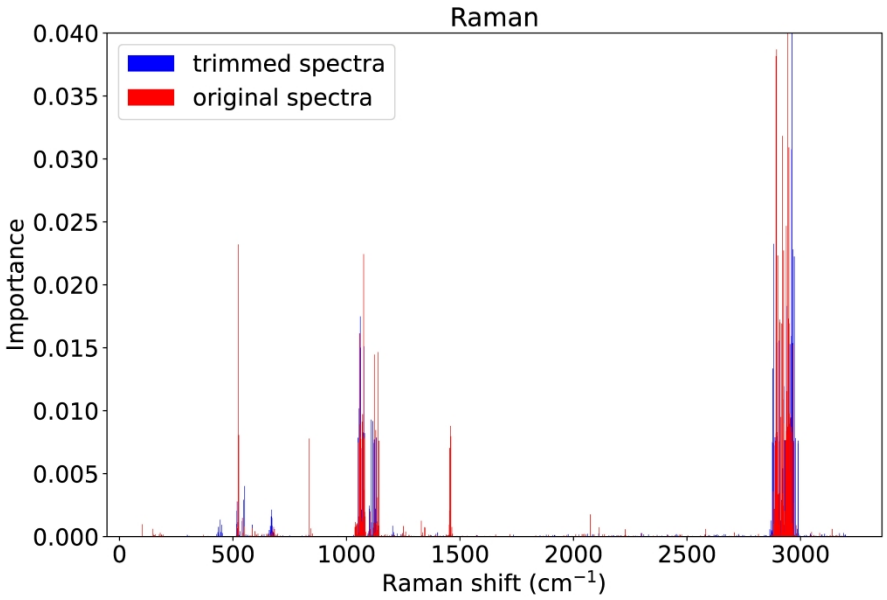


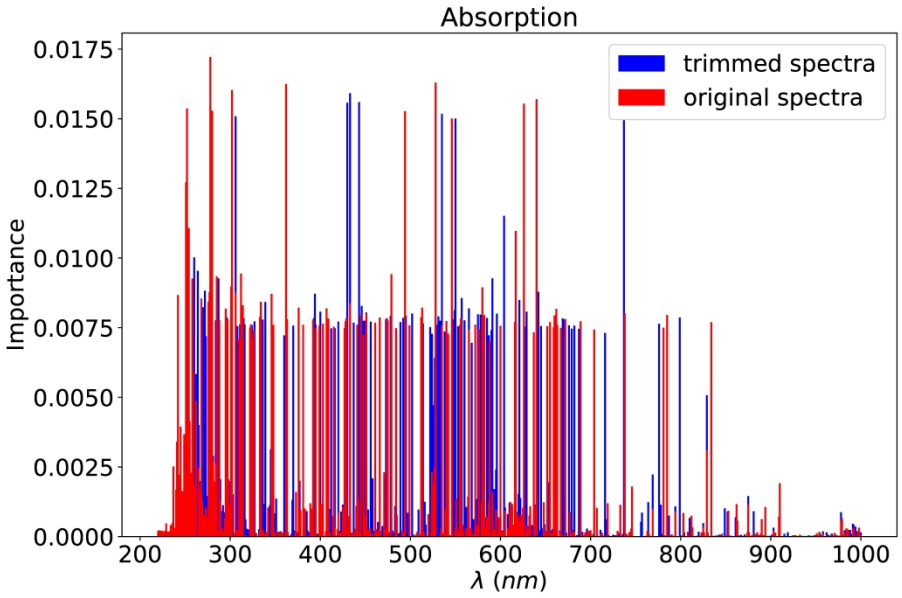
Fig 5 d

254x203mm (600 x 600 DPI)



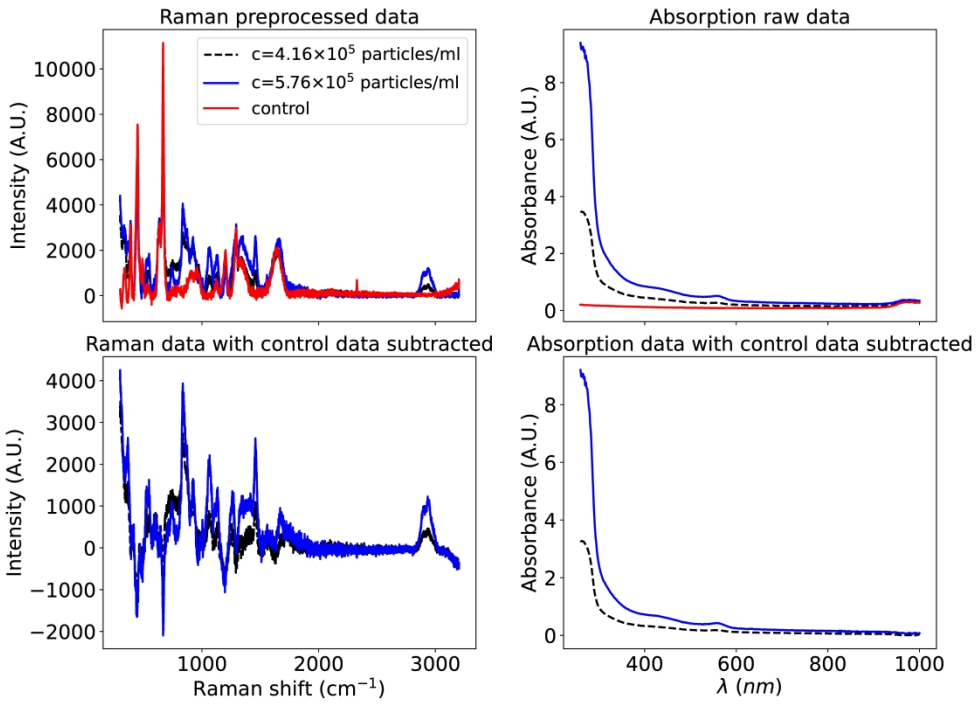
Appendix Fig A1 a

304x203mm (600 x 600 DPI)



Appendix Fig A1 b

304x203mm (600 x 600 DPI)



Appendix Fig A2

355x254mm (600 x 600 DPI)