Title

Extracting and Analyzing Biochemical Features from Nano Bioparticles for Disease Diagnosis using Surface-enhanced Raman Spectroscopy and Artificial Intelligence

Author

Li, Tieyi

Publication Date

2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Extracting and Analyzing Biochemical Features from Nano Bioparticles for Disease Diagnosis using Surface-enhanced Raman Spectroscopy and Artificial Intelligence

A dissertation submitted in partial satisfaction of

the requirements for the degree Doctor of Philosophy

in Materials Science and Engineering

by

Tieyi Li

2023

ABSTRACTION OF THE DISSERTATION

Extracting and Analyzing Biochemical Features

from Nano Bioparticles for Disease Diagnosis using Surface-enhanced

Raman Spectroscopy and Artificial Intelligence

by

Tieyi Li

Doctor of Philosophy in Materials Science and Engineering

University of California, Los Angeles, 2023

Professor Ya-Hong Xie, Chair

Disease diagnosis has long been a basis of modern medicine, enabling early intervention and effective treatment strategies. Recent advancements in nanotechnology have ushered in a new era of diagnostic techniques, with nanoscale bioparticles emerging as powerful tools in this endeavor. Nanoscale bioparticles, including extracellular vesicles, viruses, and other bioactive entities, have gained prominence due to their unique properties that make them ideal candidates for biomarker detection. These tiny structures, often measuring around 100 nanometers, carry a

wealth of molecular information reflective of the physiological and pathological states of the body. Their presence, composition, and abundance in biological fluids such as blood, saliva, and urine hold invaluable clues for diagnosing a wide range of diseases This dissertation presents a cutting-edge approach to disease diagnosis by integrating the analysis of nano bioparticles, Surface-Enhanced Raman Spectroscopy (SERS), and machine learning techniques targeting disease diagnosis. SERS, with its unparalleled sensitivity and specificity, serves as a powerful tool for the characterization of biomolecules. We investigate the feasibility of SERS in capturing the intricate spectral signatures of nano bioparticles, revealing valuable insights into their molecular composition. Moreover, machine learning models are harnessed to decipher this wealth of spectral data, enabling the identification of disease-specific biomarkers with unprecedented accuracy. The article encompasses a detailed exploration of exosome biology, the principles of SERS, the intricacies of machine learning based data analysis methodologies applied to spectral data, preliminary achievements in non-small cell lung cancer diagnostic study, and feasibility of identify SARS-CoV-2 biomarkers for COVID detection. We present a particular subgroup of exosomes derived from human bronchial epithelial cells possessing distinct spectral signatures that can be a potential indicator of non-small cell lung cancer early metastasis, and a rapid and accurate SERS based platform for COVID detection using salivary specimen, superior in some cases to RT-PCR and antigen test. The integration of these multidisciplinary approaches represents a significant step toward revolutionizing disease diagnosis through the convergence of nano bioparticle analysis, spectroscopy, and machine learning, offering a promising avenue for early and accurate disease detection in clinical settings.

The dissertation of Tieyi Li is approved.

Ximin He

Xiaochun Li

Aaswath P. Raman

Ya-Hong Xie, Committee Chair

University of California, Los Angeles

2023

Table of Contents

List of Figures

List of Tables

VITA

2018	B.S. in Astronomy Peking University, Beijing, China
2023	Ph.D. Candidate in Materials Science and Engineering University of California, Los Angeles

Chapter 1 Introduction

1.1 Disease diagnosis and the requirements

Disease diagnosis is always a critical aspect of healthcare. The importance of disease diagnosis has been emphasized since it is the very first step in managing patients' conditions. Advances and improvements in disease diagnosis facilitate better medical treatment by doctors and healthcare professionals. Thanks to the rapid development of modern technologies and tools, diagnosis methods including laboratory test (Chernecky & Berger, 2012), imaging scans (Al-Sharify et al., 2020; Pantanowitz et al., 2011), genetic testing (Burke, 2002; McPherson, 2006), endoscopy (Friedt & Welsch, 2013; Spiceland & Lodhia, 2018), biopsy (Elston et al., 2016; J. Liu et al., 2021), and point-of-care testing (Gubala et al., 2012; C. Wang et al., 2021) are invented and applied on medical cares. There are a series of metrics for evaluating a diagnostic technique. Rapid and accurate diagnosis is always the top priority (Knottnerus et al., 2002). In addition, complexity and cost of the operation, sample acquisition procedure (e.g., invasive or non-invasive), requirements on devices and operators are also practical considerations while trying to apply a novel technique into clinical application (Knottnerus et al., 2002).

1.2 Current development of applying nano-bioparticles analyses for disease diagnosis

Recently, diagnoses based on characterization and analyses of nano bioparticles (NBPs) have been showing significant potential. NBPs are a family of biological entities inside a human body that are much smaller than the typical cells and bacteria, which are usually in the dimensions of 30nm-500nm (Zhu et al., 2021). NBPs include extracellular vehicles (EVs, exosomes, microvesicles, apoptotic bodies etc.) (Van der Pol et al., 2012), viral-like particles (VLPs) (Zeltins, 2013), proteins and so on. EVs have been reported to possess informative biomarkers for multiple

diseases such as lung cancer (Fujita et al., 2015; Ren et al., 2019; Roman-Canal et al., 2019), breast cancer (Fathi et al., 2023; You et al., 2019), gastric cancer (K. Y. Chung et al., 2020; G. Li et al., 2021; Xue et al., 2023), Alzheimer's disease (Eren et al., 2022; Gallart-Palau et al., 2020), and Parkinson disease (C.-C. Chung et al., 2020; Vacchi et al., 2020). Immunofluorescence staining (Zitvogel et al., 1999), proteomics analysis (Welton et al., 2010), flow cytometry (Pospichalova et al., 2015), genomic analyses (Kalluri & LeBleu, 2016; Luo et al., 2019) are the commonly used characterization techniques to assess the contents of exosomes and study their roles in disease diagnosis. Viruses are invasive NBPs which exist in organs, tissue or circulating system (Zeltins, 2013). Virus detection for disease diagnosis includes polymerase chain reaction (PCR) (Watzinger et al., 2006a), enzyme-linked immunosorbent assay (ELISA) (Boonham et al., 2014), sequencing (Boonham et al., 2014; S. Liu et al., 2011), immunofluorescence assay (IFA) (Gardner & McQuillin, 2014; Madeley & Peiris, 2002) and viral culture (Leland & Ginocchio, 2007). As a summary, conventional characterization or detection techniques are based on a portion of the analytes and typically require a certain level volume concentration, in other words, they are ‘bulk analyses of NBPs. For the purposes of improving characterization sensitivity and accuracy, single particle characterization techniques have been recently investigated (Kibria et al., 2016; Raghu et al., 2018), such as Tunable resistive pulse sensing (TRPS) (Anderson et al., 2015) and microfluidic resistive pulse sensing (MRPS) (J. S. Kim et al., 2023). Single particle analyses advance bulk analyses mostly in terms of the capability of distinguishing informative versus non-informative analytes. At the same time, however, much more information and variation of the signal brings in more complexity and usually requires powerful and robust analyzing methods. In this study, we applied a rather young single particle characterization technique-SERS on extracting and analyzing

the biochemical signatures of NBPs, specifically exosomes and viruses then show its potential in pathological studies and clinical diagnoses.

1.3 SERS characterization of nano-bioparticles

SERS is an "upgraded version" of Raman spectroscopy by incorporating metallic surface plasmon and Raman scattering. It is a powerful analytical technique for detection and characterization at low analyte concentrations (Sharma et al., 2012). When laser reaches the surface of a metallic surface, the interactions between the electromagnetic (EM) field of laser and the free electrons of metal forms localized surface plasmon resonance (LSPR) that significantly enhanced the electric field. Much stronger Raman scattering occurs on the analytes' molecules that gives clearer representative "fingerprints" which exclusively indicates the unique biochemical contents. On the other hand, instead of targeting a portion of the nano particle, SERS generates spectroscopic signatures based on all the active molecular bonds included by the particle, therefore a more comprehensive picture could be drawn from the analyte, which contains the biomarker information on both membrane surface and inner plasm. Label-free property of SERS also greatly simplifies the sample preparation procedure. Considering the circulating activity of nano-bioparticle, proper non-invasive specimen acquisition could also be realized and promotes SERS as a more accurate and practical technology.

Despite all the advantages of SERS characterizing single nano particles, amplified variations have been observed within SERS due to multiple different modes of surface plasmon induced and fluctuations in Raman scattering. Moreover, tremendous spectroscopic data for every single nano particle is generated that lays more pressure on the data analyses. Efficient and robust data analysis techniques are required to extract useful biomarker information as well as eliminating redundant and irrelevant information. Single particle characterization typically boosts the amount

of work in measuring specimen since every single particle needs to be characterized. Therefore, a more efficient data collection procedure is needed to increase the throughput and ensure the integrity of the specimen characterization. We customized an automatic plus iterative particle scanning program and successfully boosted the spectroscopic data throughput by approximately ten times. Future improvements such as using multiple laser emission sources, increasing NBP volume concentration, improving spectrometer hardware could help overcome the data throughput bottleneck.

SERS characterization of NBPs requires well-designed platforms that allow reasonably high electromagnetic field enhancement and compatibility to NBPs in terms of dimensions and safety. Platforms with various nano structures (nano particles, nano pillars, nano spheres, nano bowls etc.) and metallic materials (silver, gold, copper, graphene etc.) have been reported in literature on tons of biological studies (tissues, cells, bacteria, virus, EVs, proteins, DNA/RNA etc.) and demonstrate different outcomes in enhancements. Based on the typical dimensions of our targets (exosomes and virus), 30-150nm, we used lithography to fabricate gold substrate featured by quasi-periodic nano pyramidal structures with dimension of approximately 200 nm to allow sufficient overlapping between the "hot spots" (region located near the metallic surface with extremely high EM field enhancement factor) and analytes. Microscopic observations prove sparse distribution of NBPs on the substrate and comparison with regular Raman spectroscopy shows much more representative spectral features and significant enhanced signal by a factor around 10^9. Sparse distribution of NBPs ensures single nano particle characterization during scanning. Enhanced signal-to-noise ratio (SNR) facilitates the following spectral data analyses by presenting more featured peaks. Laser is another important factor in measuring NBPs. Factors including wavelength, power, exposure time, spot size need optimization to prevent damaging biological

specimens, changing analytes' signature due to overheating. 488nm, 633nm, 785nm and 1064 nm are typically available for Raman spectroscopy, among the choices, 633 nm and 785 nm are most used in our research. Longer-wavelength laser gives less overheating and fluorescence due to lower energy photons. Overheating may burn the specimen especially when the analytes are sensitive to temperature and fluorescence always appears on the Raman spectra by smooth baselines that might mask the Raman scattering signal when the energy of the photon due to fluorescent emission is higher than that of the scattered photons. However, red laser excites weaker LSPR according to the gold absorption spectra. 785 nm laser turns out to be the best option which could suppress the fluorescence, maintain a good condition of the specimen, generate high SNR spectra.

1.4 Artificial intelligence

A powerful data analysis platform is essential in dealing with the tremendous amount of complex SERS spectral databases. During the past ten years, artificial intelligence (AI) has shown brilliant capabilities in analyzing data such as pattern recognition, image/video analysis, nature language processing (NLPs), signal processing, predictive modeling (Zhai et al., 2021). Numerous traditional models and neural networks (NNs) have been introduced to healthcare (Reddy et al., 2020), finance (Cao, 2020), entertainment (Hallur et al., 2021), scientific research (Xu et al., 2021) etc. Machine learning (ML), as a member of AIs, has played an important role in disease diagnosis and biological research (Ahsan et al., 2022). People have widely applied deep learning models in recognizing cancers and other diseases with medical imaging (CT scanning, MRI scanning) and monitoring healthy conditions with electronic devices (Barragán-Montero et al., 2021; Currie et al., 2019). Genomics and proteomics analyses are gradually assisted by ML to allow scientists to generate novel discoveries (Libbrecht & Noble, 2015; Swan et al., 2013).

Inspired by ML's usage in the field of healthcare, we applied multiple ML algorithms in analyzing SERS spectroscopic data of biomarker identification for disease diagnosis, such as classification, clustering, feature selection, spectral preprocessing, data collection and so on. As reported, regular Raman spectroscopy has large signal fluctuations due to multiple physical reasons, including photon statistics, shot noise, sample heterogeneity, instrumental noise etc. For SERS, in addition to the above reasons, the factors related to plasmon resonance also amplify the signal fluctuations, such as ununiform LSPR, distribution of hotspots, thermal effect (Sharma et al., 2012). All the above lead to highly complicated peak patterns of SERS spectrum and variations even for the same NBP. The target biomarker spectral features are usually overwhelmed by tremendous irrelevant features. A generalized, robust and representative pattern needs to be established as the standard "fingerprint" of the target NBP. Multiple algorithms persistent to variations in our research have been developed and utilized. Traditional statistical data analyses and processing algorithms are designed for preprocessing spectroscopic data including quality control, baseline subtraction, noise reduction, normalization. Hierarchical clustering analysis (HCA), nearest neighbors clustering etc. are applied to group and differentiate the bioparticle signatures. Support vector machine, AdaBoost of decision tree and NNs are put into identify and classify biomarker signatures for the purpose of detecting abnormal entities. Our studies indicate that our characterization and analytical setup could successfully identify cancers and COVID biomarkers (either exosomal or viral biomarkers), validated by cross-validations and blind tests. Correlation studies further show linkage between the properties of certain bioparticles and stages of diseases, which demonstrates potential capabilities in tracking patients' conditions and early diagnoses.

1.5 Achievements

As mentioned previously, we utilized our platform and method in several disease diagnoses studies including cancers, virus caused diseases, dementia etc. The main focuses of this thesis are non-small cell lung cancer (NSCLC) and corona virus related diseases.

1.5.1 NSCLC

Early detection of cancer is always a challenging task, which is also one of our goals regarding NSCLC. In our investigation on NSCLC, we found a special subpopulation of exosomes isolated from Human Bronchial Epithelial Cells (HBECs) that possess biomarkers indicating the metastasis and development of patients' NSCLC. Collected originally from patients' MPE, specific subpopulation of HBECs isolated by "constricted migration" technique has higher migration tendency (HBEC-HM) and HBEC-HM derived exosomes were characterized by SERS. Spectral signatures and the corresponding analyses prove unique signatures versus other exosome subpopulations (named unselected HBECs or HBEC-UN derived exosomes). To validate the spectral features of SERS depicting the proteomic and genomic information of NSCLC, feature selection based on ant colony optimization (ACOFS) was conducted to select the biomarker-informative peaks, which were then compared with exosome mass spectrometry characterization results. SERS features related to premalignancy were discovered to serve as the biomarker. Validations by exosomes derived from NSCLC patients' malignant pleural effusion (MPE) showed the existence of similar exosomal biomarkers found in HBEC-HM. We have found promising feasibility in NSCLC monitoring and early diagnosis, more studies including a large samples scale and improving robustness of the analytical methods are being conducted to push it towards clinical application.

1.5.2 Corona viruses

Since late 2019, the COVID-19 pandemic has been a hot topic for over three years. We have been working on improving saliva-based SARS-CoV-2 detection by incorporating our platform. Our SERS platform has shown great potential in achieving accurate, simple, and rapid testing when compared to the "gold standard" PCR test and other novel methods such as combining screening and tumor recognition. We established unique SERS spectral signatures of SARS-CoV-2 against other similar nano bioparticles such as SARS-CoV-1 and EVs, based on size and contents. Furthermore, our platform was able to distinguish between different SARS-CoV-2 variants, facilitating research and the discovery of new variants. Starting with "artificial" clinical samples prepared by spiking SARS-CoV-2 into human saliva, our platform achieved satisfactory detectability with sensitivity and specificity of 90% and 80%, respectively. Subsequently, we conducted blind tests on clinical samples from 5 COVID-19 patients and 5 healthy controls then achieved 80% sensitivity and 100% specificity. Based on these preliminary results, we are introducing more clinical samples for further validation and practical application.

1.6 Chapter overview

Chapter II explores the present and future advancements in utilizing NBP (Nanobioparticle) analysis for disease detection and diagnosis, encompassing exosomes, viruses, and more. Chapter III delves into the operational principles and the breadth of applying NBP analysis in diagnosis using SERS, including aspects such as specimen preparation, processing, data collection, and subsequent data analysis. Chapter IV furnishes an extensive overview of our materials and research methodologies. Lastly, Chapter V and chapter VI provide a specific showcase of our accomplishments in disease detection, highlighting NSCLC diagnosis and the detection of SARS-CoV-2.

1.7 References

Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022). Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. Healthcare, 10(3), 541.

Al-Sharify, Z. T., Al-Sharify, T. A., & Al-Sharify, N. T. (2020). A critical review on medical imaging techniques (CT and PET scans) in the medical field. 870(1), 012043.

Anderson, W., Lane, R., Korbie, D., & Trau, M. (2015). Observations of Tunable Resistive Pulse Sensing for Exosome Analysis: Improving System Sensitivity and Stability. Langmuir, 31(23), 6577–6587. https://doi.org/10.1021/acs.langmuir.5b01402

Barragán-Montero, A., Javaid, U., Valdés, G., Nguyen, D., Desbordes, P., Macq, B., Willems, S., Vandewinckele, L., Holmström, M., Löfman, F., Michiels, S., Souris, K., Sterpin, E., & Lee, J. A. (2021). Artificial intelligence and machine learning for medical imaging: A technology review. Physica Medica, 83, 242–256.

Boonham, N., Kreuze, J., Winter, S., van der Vlugt, R., Bergervoet, J., Tomlinson, J., & Mumford, R. (2014). Methods in virus diagnostics: From ELISA to next generation sequencing. Virus Research, 186, 20–31.

Burke, W. (2002). Genetic testing. New England Journal of Medicine, 347(23), 1867–1875.

Cao, L. (2020). AI in Finance: A Review. SSRN Electronic Journal.

Chernecky, C. C., & Berger, B. J. (2012). Laboratory tests and diagnostic procedures. Elsevier Health Sciences.

Chung, C.-C., Chan, L., Chen, J.-H., Bamodu, O. A., & Hong, C.-T. (2020). Neurofilament light chain level in plasma extracellular vesicles and Parkinson's disease. Therapeutic Advances in Neurological Disorders, 13, 1756286420975917.

Chung, K. Y., Quek, J. M., Neo, S. H., & Too, H. P. (2020). Polymer-based precipitation of extracellular vesicular miRNAs from serum improve gastric cancer miRNA biomarker performance. The Journal of Molecular Diagnostics, 22(5), 610–618.

Currie, G., Hawk, K. E., Rohren, E., Vial, A., & Klein, R. (2019). Machine Learning and Deep Learning in Medical Imaging: Intelligent Imaging. Journal of Medical Imaging and Radiation Sciences, 50(4), 477–487.

Elston, D. M., Stratman, E. J., & Miller, S. J. (2016). Skin biopsy: Biopsy issues in specific diseases. Journal of the American Academy of Dermatology, 74(1), 1–16.

Eren, E., Leoutsakos, J.-M., Troncoso, J., Lyketsos, C. G., Oh, E. S., & Kapogiannis, D. (2022). Neuronal-derived EV biomarkers track cognitive decline in Alzheimer's disease. Cells, 11(3), 436.

Fathi, M., Martinez-Paniagua, M., Rezvan, A., Montalvo, M. J., Mohanty, V., Chen, K., Mani, S. A., & Varadarajan, N. (2023). Identifying signatures of EV secretion in metastatic breast cancer through functional single-cell profiling. Iscience, 26(4), 106482.

Friedt, M., & Welsch, S. (2013). An update on pediatric endoscopy. European Journal of Medical Research, 18(1), 1–7.

Fujita, Y., Kosaka, N., Araya, J., Kuwano, K., & Ochiya, T. (2015). Extracellular vesicles in lung microenvironment and pathogenesis. Trends in Molecular Medicine, 21(9), 533–542.

Gallart-Palau, X., Guo, X., Serra, A., & Sze, S. K. (2020). Alzheimer's disease progression characterized by alterations in the molecular profiles and biogenesis of brain extracellular vesicles. Alzheimer's Research & Therapy, 12(1), 1–15.

Gardner, P. S., & McQuillin, J. (2014). Rapid virus diagnosis: Application of immunofluorescence. Butterworth-Heinemann.

Gubala, V., Harris, L. F., Ricco, A. J., Tan, M. X., & Williams, D. E. (2012). Point of care diagnostics: Status and future. Analytical Chemistry, 84(2), 487–515.

Hallur, G. G., Prabhu, S., & Aslekar, A. (2021). Entertainment in Era of AI, Big Data & IoT. In S. Das & S. Gochhait (Eds.), Digital Entertainment (pp. 87–109). Springer Nature Singapore.

Kalluri, R., & LeBleu, V. S. (2016). Discovery of double-stranded genomic DNA in circulating exosomes. 81, 275–280.

Kibria, G., Ramos, E. K., Lee, K. E., Bedoyan, S., Huang, S., Samaeekia, R., Athman, J. J., Harding, C. V., Lötvall, J., Harris, L., Thompson, C. L., & Liu, H. (2016). A rapid, automated surface protein profiling of single circulating exosomes in human blood. Scientific Reports, 6(1), Article 1.

Kim, J. S., Kwon, S. Y., Lee, J. Y., Kim, S. D., Kim, D. Y., Kim, H., Jang, N., Wang, J., Han, M., & Kong, S. H. (2023). High-throughput multi-gate microfluidic resistive pulse sensing for biological nanoparticle detection. Lab on a Chip, 23(7), 1945–1953.

Knottnerus, J. A., van Weel, C., & Muris, J. W. (2002). Evaluation of diagnostic procedures. Bmj, 324(7335), 477–480.

Leland, D. S., & Ginocchio, C. C. (2007). Role of cell culture for virus detection in the age of technology. Clinical Microbiology Reviews, 20(1), 49–78.

Li, G., Wang, G., Chi, F., Jia, Y., Wang, X., Mu, Q., Qin, K., Zhu, X., Pang, J., & Xu, B. (2021). Higher postoperative plasma EV PD-L1 predicts poor survival in patients with gastric cancer. Journal for Immunotherapy of Cancer, 9(3).

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), 321–332.

Liu, J., Chen, Y., Pei, F., Zeng, C., Yao, Y., Liao, W., & Zhao, Z. (2021). Extracellular vesicles in liquid biopsies: Potential for disease diagnosis. BioMed Research International, 2021.

Liu, S., Vijayendran, D., & Bonning, B. C. (2011). Next generation sequencing technologies for insect virus discovery. Viruses, 3(10), 1849–1869.

Luo, Y., Huang, L., Luo, W., Ye, S., & Hu, Q. (2019). Genomic analysis of lncRNA and mRNA profiles in circulating exosomes of patients with rheumatic heart disease. Biology Open, 8(12), bio045633.

Madeley, C., & Peiris, J. (2002). Methods in virus diagnosis: Immunofluorescence revisited. Journal of Clinical Virology, 25(2), 121–134.

McPherson, E. (2006). Genetic diagnosis and testing in clinical practice. Clinical Medicine & Research, 4(2), 123–129.

Pantanowitz, L., Valenstein, P. N., Evans, A. J., Kaplan, K. J., Pfeifer, J. D., Wilbur, D. C., Collins, L. C., & Colgan, T. J. (2011). Review of the current state of whole slide imaging in pathology. Journal of Pathology Informatics, 2(1), 36.

Pospichalova, V., Svoboda, J., Dave, Z., Kotrbova, A., Kaiser, K., Klemova, D., Ilkovics, L., Hampl, A., Crha, I., & Jandakova, E. (2015). Simplified protocol for flow cytometry analysis of

fluorescently labeled exosomes and microvesicles using dedicated flow cytometer. Journal of Extracellular Vesicles, 4(1), 25530.

Raghu, D., Christodoulides, J. A., Christophersen, M., Liu, J. L., Anderson, G. P., Robitaille, M., Byers, J. M., & Raphael, M. P. (2018). Nanoplasmonic pillars engineered for single exosome detection. PloS One, 13(8), e0202773.

Reddy, S., Allan, S., Coghlan, S., & Cooper, P. (2020). A governance model for the application of AI in health care. Journal of the American Medical Informatics Association, 27(3), 491–497.

Ren, W., Hou, J., Yang, C., Wang, H., Wu, S., Wu, Y., Zhao, X., & Lu, C. (2019). Extracellular vesicles secreted by hypoxia pre-challenged mesenchymal stem cells promote non-small cell lung cancer cell growth and mobility as well as macrophage M2 polarization via miR-21-5p delivery. Journal of Experimental & Clinical Cancer Research, 38, 1–14.

Roman-Canal, B., Moiola, C. P., Gatius, S., Bonnin, S., Ruiz-Miró, M., González, E., Ojanguren, A., Recuero, J. L., Gil-Moreno, A., & Falcón-Pérez, J. M. (2019). EV-associated miRNAs from pleural lavage as potential diagnostic biomarkers in lung cancer. Scientific Reports, 9(1), 15057.

Sharma, B., Frontiera, R. R., Henry, A.-I., Ringe, E., & Van Duyne, R. P. (2012). SERS: Materials, applications, and the future. Materials Today, 15(1–2), 16–25.

Spiceland, C. M., & Lodhia, N. (2018). Endoscopy in inflammatory bowel disease: Role in diagnosis, management, and treatment. World Journal of Gastroenterology, 24(35), 4014.

Swan, A. L., Mobasheri, A., Allaway, D., Liddell, S., & Bacardit, J. (2013). Application of Machine Learning to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology. OMICS: A Journal of Integrative Biology, 17(12), 595–610.

Vacchi, E., Burrello, J., Di Silvestre, D., Burrello, A., Bolis, S., Mauri, P., Vassalli, G., Cereda, C. W., Farina, C., & Barile, L. (2020). Immune profiling of plasma-derived extracellular vesicles identifies Parkinson disease. Neurology-Neuroimmunology Neuroinflammation, 7(6).

Van der Pol, E., Böing, A. N., Harrison, P., Sturk, A., & Nieuwland, R. (2012). Classification, functions, and clinical relevance of extracellular vesicles. Pharmacological Reviews, 64(3), 676–705.

Wang, C., Liu, M., Wang, Z., Li, S., Deng, Y., & He, N. (2021). Point-of-care diagnostics for infectious diseases: From methods to devices. Nano Today, 37, 101092.

Watzinger, F., Ebner, K., & Lion, T. (2006). Detection and monitoring of virus infections by real-time PCR. Molecular Aspects of Medicine, 27(2–3), 254–298.

Welton, J. L., Khanna, S., Giles, P. J., Brennan, P., Brewis, I. A., Staffurth, J., Mason, M. D., & Clayton, A. (2010). Proteomics analysis of bladder cancer exosomes. Molecular & Cellular Proteomics, 9(6), 1324–1338.

Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., Liu, X., Wu, Y., Dong, F., Qiu, C.-W., Qiu, J., Hua, K., Su, W., Wu, J., Xu, H., Han, Y., Fu, C., Yin, Z., Liu, M., … Zhang, J. (2021). Artificial intelligence: A powerful paradigm for scientific research. The Innovation, 2(4), 100179.

Xue, J., Qin, S., Ren, N., Guo, B., Shi, X., & Jia, E. (2023). Extracellular vesicle biomarkers in circulation for the diagnosis of gastric cancer: A systematic review and meta-analysis. Oncology Letters, 26(4), 1–16.

You, S., Barkalifa, R., Chaney, E. J., Tu, H., Park, J., Sorrells, J. E., Sun, Y., Liu, Y.-Z., Yang, L., & Chen, D. Z. (2019). Label-free visualization and characterization of extracellular vesicles in breast cancer. Proceedings of the National Academy of Sciences, 116(48), 24012–24018.

Zeltins, A. (2013). Construction and characterization of virus-like particles: A review. Molecular Biotechnology, 53(1), 92–107.

Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Istenic, A., Spector, M., Liu, J.-B., Yuan, J., & Li, Y. (2021). A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. Complexity, 2021, 1–18.

Zhu, X., Cicek, A., Li, Y., & Yanik, A. A. (2021). Plasmonic Nanopores: Optofluidic Separation of Nano-Bioparticles via Negative Depletion. In Nanopores. IntechOpen.

Zitvogel, L., Fernandez, N., Lozier, A., Wolfers, J., Regnault, A., Raposo, G., & Amigorena, S. (1999). Dendritic cells or their exosomes are effective biotherapies of cancer. European Journal of Cancer, 35, S36–S38.

Chapter 2 Standard operating procedure of SERS platform for characterizing and analyzing nano-bioparticles

We have been building a Standard Operating Procedure (SOP) for maintaining, upgrading, and applying SERS platform. As shown in Figure 2.1, the SOP is composed of four general sections, sample preparation, SERS characterization, data analysis, clinical usage sequentially. Each part has been undergoing optimization to maximize the overall performance.

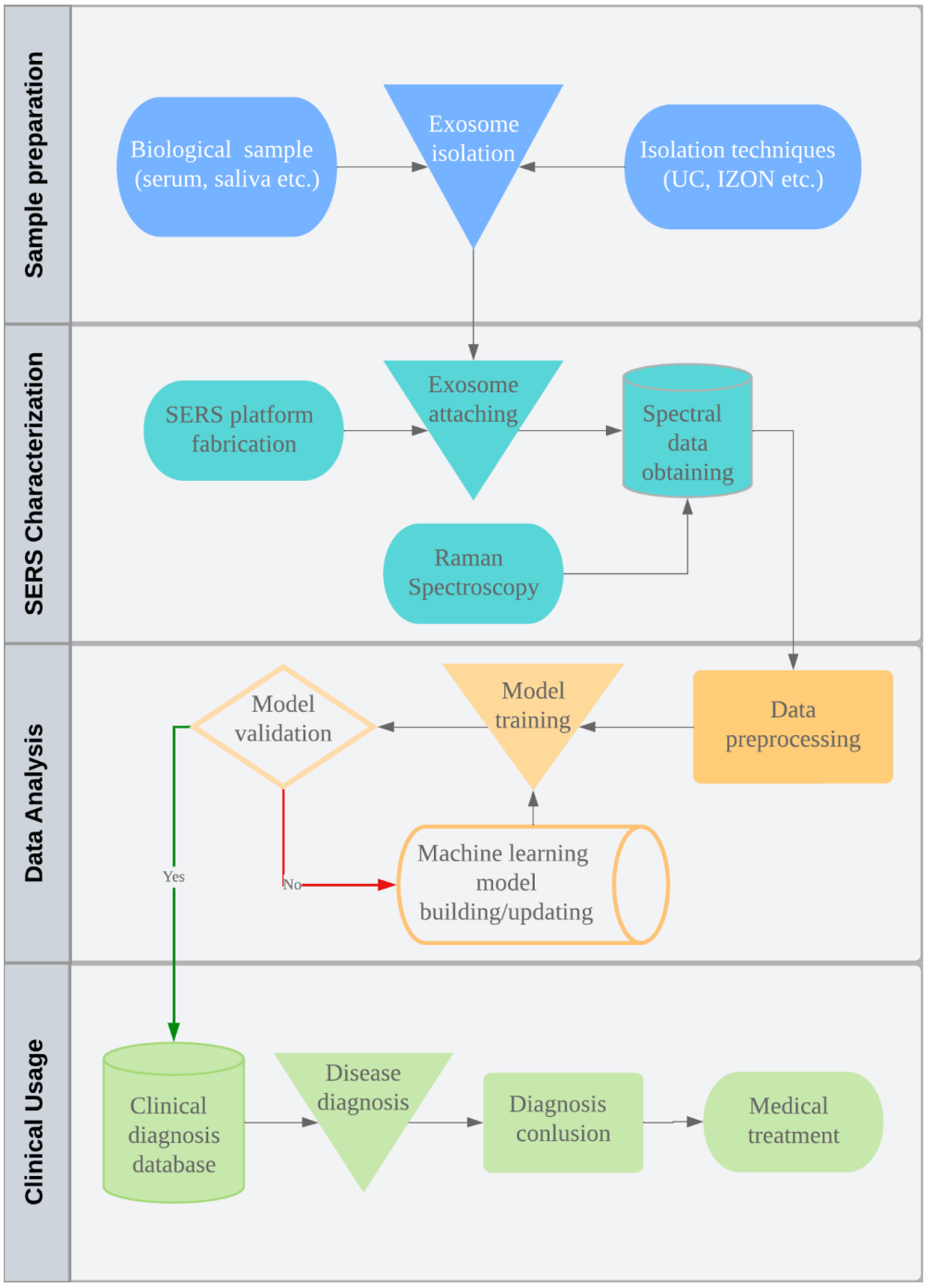


Figure 2.1 SOP of SERS-based platform for disease diagnosis.

2.1 Nano-bioparticle specimen preparation

Sample sources are typically human body fluid, which is mainly chosen by the abundance of the target biomarkers. There are also other factors when determining the sample sources, including invasiveness, complexity, time etc. For instance, we have collected bronchoalveolar lavage (BAL) and MPE for detecting lung cancer, cerebrospinal fluid (CSF) for dementia, serum or blood for breast cancer, saliva for COVID. Each of the body fluids has unique components and different physical properties (viscosity, density, volume etc.), according to which specially designed NBP isolation methods were utilized (Martins et al., 2023).

The most commonly and well-established technique is sequential ultracentrifugation (SUC) combined with ultrafiltration (M. Zhou et al., 2020). Typically, UC-based NBP isolation contains six to seven rounds of different centrifugal forces and time, as well as several ultrafiltration (UF, e.g., 0.22 μm column filter). UC can produce highly concentrated and pure nano bioparticles with optimized operation parameters, however, it usually needs complicated operations and professional ultracentrifuge that are rather costly. Recently, chromatography has been used in isolating NBPs from multiple biological fluids such as blood and urine (P. Li et al., 2017). In this technique, NBPs are separated by their physical properties including size, charge, and hydrophobicity. Typical chromatography includes size exclusion chromatography (SEC), ion exchange chromatography (IEC), affinity chromatography (AC). In our study, SEC (IZON ExoQuick columns) are used to isolate EVs from MPE, CSF, serum, and UC to isolate corona viruses from cell culture media, saliva. Isolated NBPs samples are resuspended in PBS with other special reagents.

2.2 SERS characterization

Well prepared specimens are loaded onto the SERS substrate for Raman characterization. The design and fabrication of SERS substrate for appropriate surface plasmon are one of the critical steps. Biological affinity and safety, compatibility to NBPs' sizes, enhancement factors (EFs) of surface plasmon resonance, feasibility of fabrication, productivity, compatibility to single particle detection are the key factors that impact the ultimate performance. There are numbers of material choices and nano structures reported with acceptable performance. As mentioned, gold, silver and copper are commonly used (Sharma et al., 2012). Nano pillars, nano bowls, nano particles, nano spheres etc. are frequently reported (Mo et al., 2016; Shen et al., 2009; Yue et al., 2020). Considering the size range of NBPs (30-150 nm), we utilized quasi periodic nano gold pyramidal structure made by lithography protocol. Each unit has dimensions of 200 nm × 200 nm × 250 nm (width × length × height). Pyramids are arranged in a hexagonal manner. The adjacent pyramids are spaced at 400 nm. FDTD simulation results show the "hotspots" are generally located at the lateral facet of each pyramid, which spread in a range of 100 nm with acceptable EFs (P. Wang et al., 2013). This "bottom-up gradual open" space allows NBPs fully overlapping with "hotspots", generating the spectral fingerprints of the entire intact NBP. More information from a single NBP increases the chance of discovering the unique features for identifying meaningful biomarkers. More details of fabrication are given in Section 3.3.4.

Loading specimens onto the substrate is rather straightforward. Typically, a tiny droplet (3-10 μL) of specimen is pipetted onto a 5 mm by 5 mm area of the substrate. For common biological buffer like PBS, the specimen droplet forms a "coffee ring" after drying since PBS solution typically doesn't wet the surface. However, other kinds of solutions (such as ethanol, DMSO) or specific solute added to PBS could cause changes to the contacting behavior between specimen

and substrate surface. In this case, grid barriers need to be placed on the surface to prevent different specimen mixing and control the NBPs distributions.

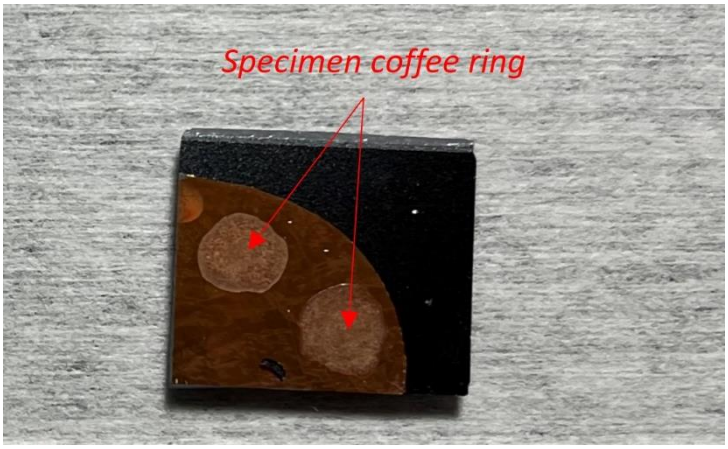


Figure 2.2 NBPs in PBS buffer forms coffee ring.

Subsequently, substrate with loaded specimens is placed in Raman spectrometer for characterization. To improve the spectroscopic data throughput, we run arial map acquisition on Raman spectrometer, which is basically a “scanning-characterizing” procedure. Searching maps with shorter exposure time and larger scanning area serve for capturing potential NBPs positions. Characterizing maps with longer exposure time run targeting the recorded NBPs positions and generate qualified spectral data. Since this “scanning-characterizing” procedure is a rather repetitive task, command script on Raman spectrometer support PC enabling auto-focus, maps management, signal detection, auto-characterization is installed to increase the data throughput. Compared to manual data collection, automatic maps acquisitions could boost data throughput by a factor of 5-10. Collected spectral data are ensembled into large datasets for the followed by analyses.

2.3 Data analysis

SERS spectral data analyses generally contain quality control, data formatting, preprocessing, fingerprints analyses. Efficient and accurate biomarker establishment requires spectroscopic data above a threshold of quality. A qualified spectrum is supposed to have reasonable numbers of well-shaped peaks and is free of fluorescence background, random noises and spikes due to cosmic rays (Sharma et al., 2012). Therefore, a filtering program is used to remove those spectra with insufficient qualities. All the spectra passing the filter undergo preprocessing including background subtraction, noise reduction, normalization. They are designed to remove the fluorescence interference and signal random fluctuations respectively on the data aspect. Fingerprints analyses include multiple distinct methods for different purposes. For instance, dimensionality reduction algorithms such as principal component analysis (PCA), linear discriminant analysis (LDA) etc. are used for visualizing the general data distribution and their simple linear relation among Raman shifts. SVMs, decision trees, NNs and other machine learning models are used to investigate and learn unique features of certain types of NBPs and output a predictive model for later tests. HCA, K-Nearest Neighbors Clustering (KNN) are for clustering subpopulation in NBPs from one specimen (e.g., SARS-CoV-2) and picking out the most representative biomarker for a disease. We also applied generic feature selection algorithms (ant colony optimization, particle swarm optimization-based feature selection) to build the linkage between Raman peak features and bio-chemical molecules' properties, such as up/down regulation of proteins or DNA/RNA mutations. With trustworthy biomarkers' fingerprints established, we then validate them and try to put into disease diagnosis.

2.4 Clinical usage

Systematic validations are strictly required. We conducted cross validations and blind tests for the purposes of optimizing protocol, tuning parameters, and evaluating our methodology. Clinical samples of investigated diseases are usually involved, typically, diseased samples versus healthy control samples are used to obtain sensitivity/specificity. In terms of selecting samples, there are questions related to biostatistics that need comprehensive investigation. For example, what is an appropriate size of clinical sample dataset? How many NBPs are needed to draw an enough clear picture of specimens? The scale of our clinical sample set is in the range of 30-100. We have achieved good grades in lung cancer diagnosis by exosomes and saliva based COVID detection. More clinical samples are planned according to biostatistical theories to further improve the solidity.

2.5 References

Li, P., Kaslan, M., Lee, S. H., Yao, J., & Gao, Z. (2017). Progress in Exosome Isolation Techniques. Theranostics, 7(3), 789–804.

Martins, T. S., Vaz, M., & Henriques, A. G. (2023). A review on comparative studies addressing exosome isolation methods from body fluids. Analytical and Bioanalytical Chemistry, 415(7), 1239–1263.

Mo, A. H., Landon, P. B., Gomez, K. S., Kang, H., Lee, J., Zhang, C., Janetanakit, W., Sant, V., Lu, T., Colburn, D. A., Akkiraju, S., Dossou, S., Cao, Y., Lee, K.-F., Varghese, S., Glinsky, G., & Lal, R. (2016). Magnetically-responsive silica–gold nanobowls for targeted delivery and SERS-based sensing. Nanoscale, 8(23), 11840–11850.

Sharma, B., Frontiera, R. R., Henry, A.-I., Ringe, E., & Van Duyne, R. P. (2012). SERS: Materials, applications, and the future. Materials Today, 15(1–2), 16–25.

Shen, X. S., Wang, G. Z., Hong, X., & Zhu, W. (2009). Nanospheres of silver nanoparticles: Agglomeration, surface morphology control and application as SERS substrates. Physical Chemistry Chemical Physics, 11(34), 7450.

Wang, P., Liang, O., Zhang, W., Schroeder, T., & Xie, Y. (2013). Ultra-Sensitive Graphene-Plasmonic Hybrid Platform for Label-Free Detection. Advanced Materials, 25(35), 4918–4924.

Yue, W., Gong, T., Long, X., Kravets, V., Gao, P., Pu, M., & Wang, C. (2020). Sensitive and reproducible surface-enhanced raman spectroscopy (SERS) with arrays of dimer-nanopillars. Sensors and Actuators B: Chemical, 322, 128563.

Zhou, M., Weber, S. R., Zhao, Y., Chen, H., & Sundstrom, J. M. (2020). Methods for exosome isolation and characterization. In Exosomes (pp. 23–38). Elsevier.

Chapter 3 Backgrounds, materials, methods

3.1 Nano-bioparticles

As stated in previous sections, NBPs refer to any nano-sized, discrete particles with biological functions which are suspended in a liquid medium. They play important roles in numerous physiological and pathological processes, such as substance transportation, immune response, neural transmission, disease development and so on. As such, NBPs often contain biomarkers for various biological processes, for example, cancer metastasis, neurodegenerative disorders, respiratory diseases. Our focuses are EVs, especially exosomes, and SARS-CoV-2 including its variant.

3.1.1. Extracellular vesicle (exosomes)

EVs are a heterogeneous family of structures/bodies enclosed by lipid bilayer that are secreted by cells into extracellular environments (Raposo & Stoorvogel, 2013). EVs were not believed to play significant roles in biological processes until the findings by Raposo et al in 1996 (Raposo et al., 1996), which proved the effect of EVs on immune responses. EVs are generally grouped into three classes based on size, origin and biogenesis pathways, which are exosomes, microvesicles (MVs), and apoptotic bodies, ranging 30-150 nm, 100-1000 nm, 1-5 μm respectively (Raposo & Stoorvogel, 2013). As demonstrated by Figure 3.1, exosomes are produced from the endolysosomal pathway and secreted from cells through the fusion of multivesicular bodies with the plasma membrane. In contrast, MVs are formed by budding from plasma membrane directly. Both types of EVs have the features of containing cytoplasmic proteins, lipid bilayer and nucleic acids.

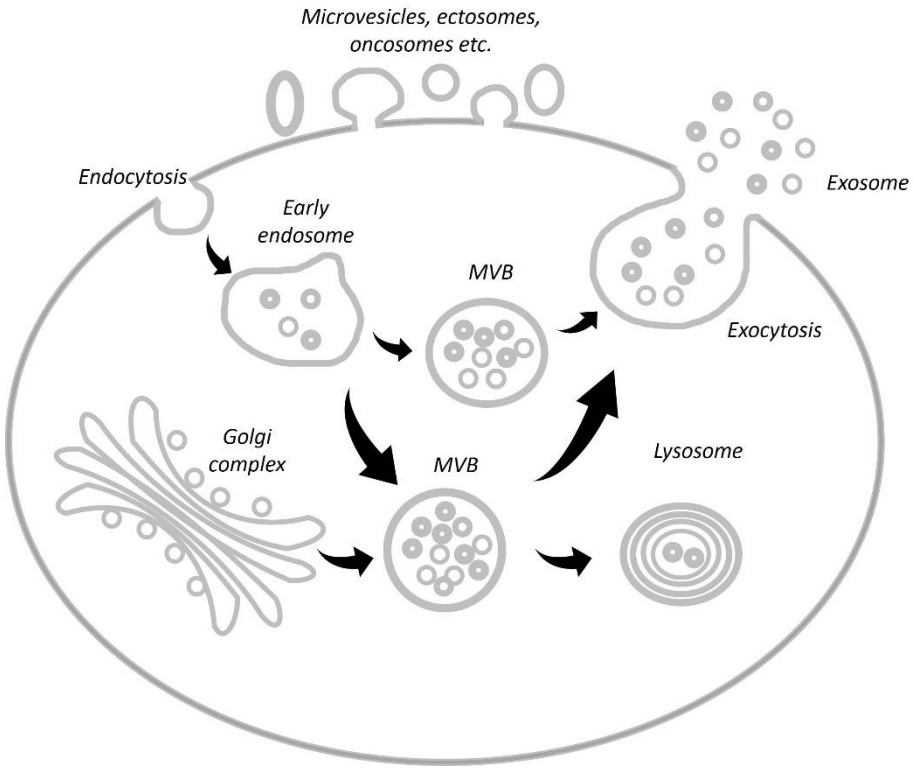


Figure 3.1 The biogenesis of EVs. EVs include exosomes, lysosomes, microvesicles, ectosomes, oncosomes etc.

Exosomes are believed to be a uniform population of vesicles of endocytic origin. They are formed by the inward budding of the multivesicular body (MVB) membrane, and cargo sorting into exosomes is facilitated by the endosomal sorting complex required for transport (ESCRT) and associated proteins such as ALIX and TSG101 (Colombo et al., 2013; Jiang et al., 2020; Koritzinsky et al., 2019). Moreover, in some cells, exosome production requires ceramide and neutral sphingomyelinase (Guo et al., 2015; Trajkovic et al., 2008), while in others, small GTPases such as RAB27A, RAB11, and RAB31 are involved in the fusion of MVBs with the cell membrane, leading to the secretion of exosomes (C. Hsu et al., 2010). Exosomes have the same membrane orientation as the cell of origin, like MVs. MVs represent a more heterogeneous population of vesicles formed by outward budding (Camussi et al., 2010). Extracellular vesicles, including

exosomes, serve as signalosomes for various biological processes. They are involved in antigen presentation (Lindenbergh et al., 2020), immune regulation (X. Zhou et al., 2020), and can directly activate cell surface receptors via protein and bioactive lipid ligands, transfer cell surface receptors (Christianson et al., 2013), and deliver effectors such as transcription factors, oncogenes, and infectious particles into recipient cells (Femminò et al., 2020). In addition, they contain various RNA species, including mRNAs, microRNAs, and non-coding RNAs, which can be functionally delivered to recipient cells.

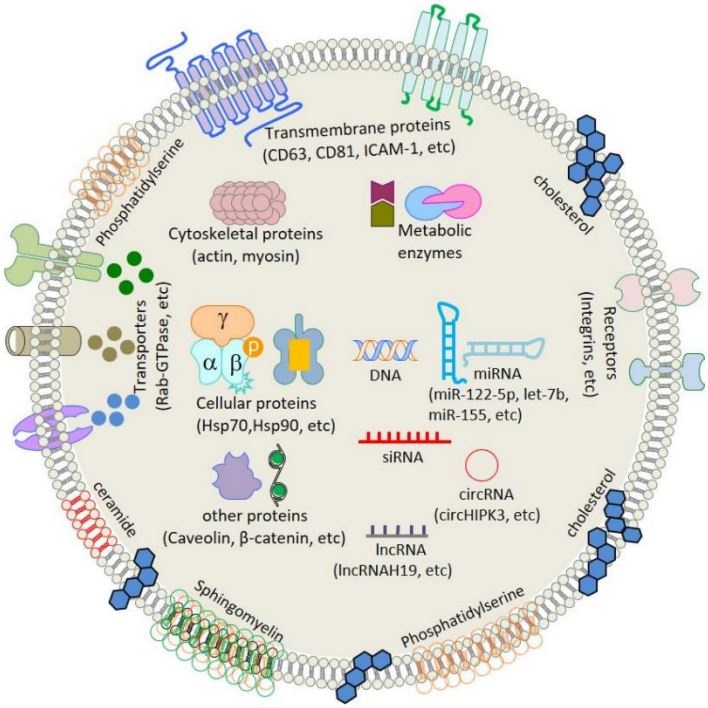


Figure 3.2 Diagram of exosome composition.

Exosomes are assumed initially as ‘trash bag’ for cells to exclude unwanted constituents (Van der Pol et al., 2012), however, studies have demonstrated that exosomes play significant roles in intercellular communication and have notably effect on both physiological and pathological processes in terms of the specially selected molecular components in exosomes. The biofunction

of exosomes is mainly determined by the host cell type and the composition of exosomes in terms of proteins, nucleic acids, carbohydrates. Recent studies showed a significant role of exosomes in alternative exclusion of proteins such as release of receptors and unwanted proteins, cell-to-cell signaling including antigen presentation and immune activation and suppression (Camussi et al., 2010; Femminò et al., 2020; Van der Pol et al., 2012). Besides, exosomes are also found to be enriched in mRNA and miRNA, thus serving as a shuttle for RNA to confer new functions to target cells (Das et al., 2019).

Proteomic and genomic studies reported that exosomes contain a specific subset of cellular proteins, nucleic acids and cytosol, some depend on the host cell type whereas others are found in most exosomes regardless of cell type (Kalluri & LeBleu, 2020; Pegtel & Gould, 2019), as shown in Figure 3.2 (Jan et al., 2021). Proteins from endosomes, plasma membrane and cytosol in the function of membrane transporters such as CD9, CD63 and CD81 are commonly found. Whereas those from nucleus, mitochondria and the Golgi complex mostly varied. miRNAs, as a newly found generic material that are exported outside cells and can serve as a communicator between cells, are found undergoing a specific selection of sequence during the formation of exosomes (Das et al., 2019). These observations reinforce the specificity of formation of exosomes that represent a specific subcellular portion and not a fully random process. Therefore, the molecular content of exosomes potentially reflects the origin and pathophysiological conditions of the releasing cells. Researchers claimed that exosomes contain cargo implicated in cancer, neurodegenerative, infectious diseases etc. (Howitt & Hill, 2016b; J.-H. Kim et al., 2018; Rangel-Ramírez et al., 2023).

According to the public exosome content database Exocarta, over 40,000 proteins, 1000 lipids, 7000 RNAs have been found in exosome from multiple organisms, which are indicative of pathophysiological conditions of their host cell (Keerthikumar et al., 2016). The enrichment of

diagnostic biomarkers that enables the detection of relative diseases during its early stage. Many protein biomarkers in circulating exosomes have been found to be potentially useful in disease diagnosis such as cancer and neurodegenerative diseases. Nilsson demonstrated the two known prostate cancer biomarkers, PCA-3 and TMPRSS2: ERG in exosomes isolated from urine (Nilsson et al., 2009). Exosomal amyloid peptides have also been demonstrated to accumulate in brain plaques of Alzheimer`s disease (AD) patients and the proven biomarker for AD, tau phosphorylated at Thr-181, is present at an upregulated level in exosomes from CSF of AD patients with mild symptoms (Xiao et al., 2017). In addition to protein biomarkers, exosomal nucleic acids such as mRNA and miRNA could also be diagnostic biomarkers. Fu et al found that the expression level of TRIM3 mRNA is notably decreased in exosomes isolated from gastric cancer patients` serum (H. Fu et al., 2018). In 2013, Tanaka et al. claimed that the exosomal miR-21 expression level is elevated in exosome isolated from patients suffering from esophageal squamous cell cancer (ESCC) (Tanaka et al., 2013). These findings strongly support the arguments of exosomes being biomarker carriers compared with conventional specimens such as serum or urine. More importantly, exosomes biomarkers from early obtainable biofluids such as saliva could be extremely suitable for clinical application (Y. Han et al., 2018). Generally, exosome biomarkers identification is still in the progress of investigation and their clinical value would be fully explored.

3.1.2 Viruses

SERS based single NBP analyses can play an important role in other types of NBP studies in addition to exosome. The family of viruses also contains tremendous biological information that can be analyzed in a single-particle manner. Detection of viruses' biomarkers can be more straightforward than exosome analysis, because as alien invader, the structure and function of

viruses are already exhaustively investigated, the role of SERS is more on built the standard signatures on known viral biomarker for the detection.

Viruses are usually incredibly small (typically 20-200nm) and consist of two main components (Johnson, 2000). (1) Genetic Material: The core of a virus contains genetic material, which can be either DNA or RNA. This genetic material carries the instructions for viral replication and is typically a single or double strand; Capsid: The genetic material is encased within a protective protein coat called a capsid. The capsid is composed of protein subunits called capsomers, which self-assemble around the genetic material. Some viruses also have an outer lipid envelope derived from the host cell's membrane. The identification of viral biomarkers is focusing on those two components. The classification of viruses depends on several criteria including genetic material, capsid structure, host specificity, mode of replication and so on.

SARS-CoV-2, investigated in our research, is composed of structural spike protein (S protein), lipid bilayer, membrane protein (M protein), envelop protein (E protein), and nucleocapsid protein (N protein) (Hasöksüz et al., 2020), demonstrated by Figure 3.3. Its viral genome length is over 30,000 bases, which is relatively longer than the typical RNA viruses. The SARS-CoV-2 genome also encodes 16-17 non-structural proteins (ns1 to ns17), including 3-chymotrypsin-like protease (3CLpro), papain-like protease (PLpro), helicase etc. S protein is a 1.273-amino acid trimetric glycoprotein composed of S1 attachment domain (1-686 residues) and S2 fusion domain (687-1,273 residues) (Bai et al., 2022). Part of S1, named receptor-binding domain (RBD, 306-545 residues), provides the main function of binding host cells by recognizing angiotensin-converting enzyme 2 (ACE2) and cellular protease TMPRSS2 on cells, followed by fusion between viral and host cells' membranes (L. Zhou et al., 2020). The E protein is involved in virus assembly and release since it plays a role in maintaining the structure of the viral envelop,

the M protein is also located on E protein for shaping the viral structure and interacting with host cells. The N protein is thus responsible for binding viral RNA genome and forming the ribonucleoprotein complex (RNP), it is also involved in virus replication and transcription by interacting with other viral proteins and host cells' proteins. The S protein and N protein are the main target of the immune system and are usually used for COVID diagnosis (H. Wang et al., 2020).

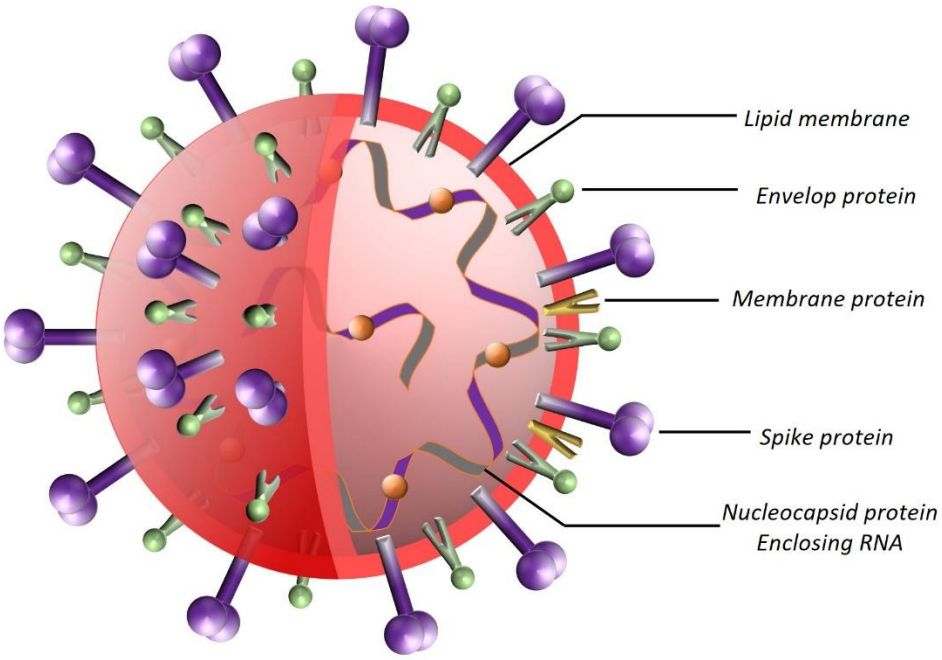


Figure 3.3 Diagram of SARS-CoV-2 structure and composition.

Current SARS-CoV-2 mutations occur mostly in the spike genes, enhancing the infectivity by increasing its ability to enter human cells. The mutations in genome include D614G, the RBD mutation N501Y, the RBD mutation E484K, N-terminal domain mutations, and non-spike mutations (Cosar et al., 2022). These mutations change the structure and amount of SARS-CoV-2 structural proteins, causing different variants possessing different transmissibility, disease severity and ability to evade human immune systems. Nowadays, SARS-CoV-2 has evolved into more than

50 different variants (Cosar et al., 2022; Magazine et al., 2022). Since SARS-CoV-2 belongs to the family of single strand RNA virus, RT-PCR has always been the most prevalent and reliable detection technique due to its accuracy and LOD (Kevadiya et al., 2021). Antigen detection has also been commonly used in the communities as a rapid self-check approach (Dinnes et al., 2022). We assume that mutations especially in S protein might provide SARS-CoV-2 unique discoverable biomarkers in terms of SERS based single NBP analysis. More investigation on its biological properties as well as detection approaches have been conducted to control the pandemic.

3.2 NBPs for disease diagnosis in literature

3.2.1 Studies on exosome-based disease diagnosis

Exosome-based disease diagnosis has been rapidly developed since early 2000s (Y. Zhang et al., 2020). As reported, exosomes play critical roles in intercellular communications mainly by transmitting molecules (proteins, lipids, nucleic acids etc.) (Alenquer & Amorim, 2015; Y. Zhang et al., 2020). Recent studies have shown that exosomes contain biomarkers for various diseases including cancer (W. Li et al., 2017; Soung et al., 2017), neurodegenerative diseases (Howitt & Hill, 2016a), infectious diseases (Fleming et al., 2014), making them a promising agent for diagnosis.

Exosomes have been identified as multifaceted regulators in cancer development by harboring molecules from cancer cells (Alenquer & Amorim, 2015). They have the ability to alter the tumor microenvironment, which affects adjacent cells (Carretero-González et al., 2018). Specifically, exosomes play a role in tumor growth and metastasis. Studies have demonstrated that exosomes can transfer genes associated with cancer growth promotion, which leads to the proliferation of metastatic cancer cells (Carretero-González et al., 2018; K. Li et al., 2019). Moreover, exosomal RNAs (exRNAs), including microRNAs and non-coding RNAs, have been

implicated as stimulators of cancer progression, such as in breast cancer and gastric cancer (M. Fu et al., 2019; Lakshmi et al., 2021; Rabinowits et al., 2009). It has been found that exosomal miRNAs can mediate silencing of downstream genes, thus promoting tumorigenesis in non-tumorigenic epithelial cells (Blackwell et al., 2017; Y. Liu et al., 2016). Research by J. Zhang and C. Chen et al revealed that exosomes derived from human umbilical cord blood derived EPCs have robust pro-angiogenic effects and can be incorporated into endothelial cells, significantly increasing their proliferation (Y. Hu et al., 2018). G. Sagar and R. Sah also found that exosomal adrenomedullin (AM) interacts with receptors on adipocytes, activating p38 and extracellular signal-regulated (ERK1/2) mitogen-activated protein kinases, which promote lipolysis in adipose tissue (Sagar et al., 2016). S.A. Melo et al reported that breast cancer-derived exosomes contain microRNAs associated with the RISC loading complex (RLC), which can initiate the formation of tumors in a Dicer-dependent manner in non-tumorigenic epithelial cells (Melo et al., 2014). Moreover, exosomes derived from highly metastatic lung cancer cells have been shown to induce vimentin expression and epithelial-to-mesenchymal transition (EMT) in HBECs, leading to migration, invasion, and proliferation in non-cancerous recipient cells (Rahman et al., 2016). Additionally, exosomes from different types of tumors have unique properties and are taken up by distinct resident cells due to integrin expression patterns (Soung et al., 2017). Therefore, exosomes may possess distinguishable biomarkers for various types of cancers and could serve as a bridge from normal cells to cancer cells, providing a possibility for cancer diagnosis.

Exosome specimens undergoing characterization for cancer diagnosis can be acquired from multiple sources. Body fluids and tissues are typical targets of collecting exosomes. Researchers fetch exosome specimens from blood or serum for prostate cancer (Malla et al., 2018), colorectal cancer (Matsumura et al., 2015), lung cancer (Taverna et al., 2016; Wu et al., 2020), from milk for

breast cancer (Xie et al., 2022), from malignant pleural effusion and bronchiolar lavage for non-small cell lung cancer etc. Unlike the exosomes in circulating system, tissue derived exosomes are assumed to possess more abundant biomarkers. Vella et al found that exosomes derived from brain tissue maintain the same traits of brain homogenate (L. Vella et al., 2016; L. J. Vella et al., 2017). It indicates that tissue derived exosomes could be another effective resource for disease diagnosis.

Depending on the type of exosomal biomarkers (proteins, RNAs etc.), different characterization technologies are applied, including immunofluorescence labeling, microscopy imaging, proteomics, genomics, surface plasmon resonance etc. Over the past years, proteomic analyses have been the most used methods to analyze exosomal protein biomarkers (Olver & Vidal, 2007). Mass spectrometry (MS) has been widely used in exosome related cancer detection. H. R. Larsen and K. Lund et al developed a capillary liquid chromatography MS platform in analyzing exosomes from breast cancer cell lines and found 27-Hydroxycholesterol associated with proliferation and metastasis in estrogen receptor breast cancer (Roberg-Larsen et al., 2017). K. Iha et al applied an ultrasensitive Enzyme-linked immunosorbent assay (ELISA) combined with thio-NAD cycling on detecting proteins in human cervical carcinoma derived exosomes' lumen and membrane fractions (Iha et al., 2022). After identifying that 221 proteins from urinary exosomes are differentially expressed in prostate cancer patients by MS, L. Wang et al applied antibody-based methods, including Western blot and ELISA, and investigated deeply on the expression of urinary exosomal biomarkers (flotillin 2, TMEM256, Rab3B etc.) (L. Wang et al., 2017). More generally, researchers always utilize western blotting to validate the standard exosomal biomarkers (CD63, CD9, CD81 etc.) (Y. Zhang et al., 2020). There are other aptamer-based methodologies emerging for overcoming the issues of MS and improving quantification. J. Webber et al used a novel affinity-based platform with specific protein binding reagents named SOMAmers to analyze

prostate cancer cells and succeeded in identifying over 300 proteins (Webber et al., 2014). Similarly, J. L. Welton, P. Brennan et al utilized SOMAscan assay, a multiplex aptamer-based protein array, to eliminate the effect of plasma proteins in order to discover unique proteins in prostate cancer cell derived exosomes (Welton et al., 2016).

Immunofluorescence-based techniques have been widely applied in analyzing proteins of various types of biological entities, such as cells, bacteria, viruses, EVs etc., due to its capabilities of specific targeting, multiplex sensing and feasibility to be integrated with microfluidic assays (Francisco-Cruz et al., 2020; Sood et al., 2016). It is essentially based on the antigen-antibody interaction between exosomal proteins and fluorescent reagents. Work carried out by Z. Zhao et al demonstrated tumor-derived circulating exosomes contain antigens as promising biomarker source for ovarian cancer diagnosis by employing ExoSearch and immunomagnetic beads (Zhao et al., 2016). In addition, S. Fang and H. Tian developed an immunocapture and quantification platform using IFKine Green Donkey anti-goat IgG and Dylight 549 goat anti-rabbit IgG for examining the clinical application of breast cancer derived exosomes (Fang et al., 2017). Despite the advantages of immunofluorescence-based protein detection, it also suffers from issues such as non-specific binding which leads to false-positive signals, low sensitivity with low-abundance proteins, requirements for specialized equipment like fluorescence microscope and expensive reagents could also be limitations (Francisco-Cruz et al., 2020; Shakes et al., 2012).

Compared with the diverse approaches for proteomics, genomic analyses of exosomes are typically based on sequencing. RNA sequencing or transcriptome sequencing (RNA seq) is a method using next-generation sequencing (NGS) to examine the sequences of exosomal RNA. As a supplemental approaches to proteomic analysis, genomics sometimes performs more sensitively and accurately in identifying biomarkers, especially for disease early diagnosis. Work done by D.

Wang et al revealed a coordinated increase in the levels of miR-146a-5p and miR-155-5p in colorectal cancer cells and exosomes, which promote the activation of cancer-associated fibroblasts through JAK2-STAT3/NR-κB signaling (D. Wang et al., 2022). J. Cai, L. Gong reported exosomal miR-6780b-5p correlated with EMT of ovarian cancer cells by sequencing exosomal RNAs (J. Cai et al., 2021). There are a lot more efforts focusing on the biological functions of exRNAs for more diseases, such as gastric cancer (F. Li et al., 2018), breast cancer (Rykova et al., 2008), lung cancer (Ni et al., 2023), neurodegenerative diseases (Saugstad et al., 2017). However, most sequencing methods are not cost effective and time consuming. It also encounters issues because of lacking optimized SOPs when the sample quantities are limited. Additionally, the biological role of exRNAs needs more sophisticated investigations (Kukurba & Montgomery, 2015).

3.2.2 Studies on virus and specifically COVID detection

Viruses are foreign invasive biological entities infecting human cells. While infecting a host cell, viruses hijack the host cellular machinery to produce more copies of themselves, leading to abnormalities. Similar to EVs, most viruses contain distinguishable biomarkers enabling early detection and development tracking (Alenquer & Amorim, 2015; M.-H. Zhang et al., 2023). Different technologies have been developed for virus detection, such as the most prevalent PCR, ELISA, NGS, viral culture, immunofluorescence etc. Targeting various components of viruses, those techniques have their own specific strength as well as drawbacks, appropriate approaches need to be selected for early, efficient, and accurate diagnosis.

The employment of PCR in investigation and detection of viruses has been regarded as the "gold standard", due to its capability of rendering high sensitivity and reproducibility, as well as a wide dynamic range (Bustin et al., 2005; Watzinger et al., 2006b). Targeting on the viral nucleic

acids, PCR undergoes exponential amplification of the target sequence, which leads to a limit of detection (LOD) down to the range of 1-10^2 copies/mL (Parker et al., 2015). Quantification of target sequence by real-time quantitative PCR (RQ-PCR) is based on the continuous measurement of accumulation or reduction of fluorescence during the amplification reaction (Bustin et al., 2005). Real-time PCR (RT-PCR) allows quantification based on the detection of the number of amplicons generated during each amplification cycle in a real-time mode. RT-PCR has been the most important technique in detection viral infection including influenza (Chu et al., 2015), HIV (Rutsaert et al., 2018), hepatitis viruses (Abe et al., 1999), corona viruses (Teymouri et al., 2021), respiratory syncytial virus (A. Hu et al., 2003) etc. During the COVID pandemic since 2019, RT-PCR has been playing the most important role in diagnosis and controlling infection situations (Teymouri et al., 2021). The tests involve collecting respiratory specimens (nasopharyngeal or oropharyngeal swab, sputum, saliva etc.) from suspected COVID patients then target viral RNA detection using specific primers and amplification. The cycle threshold value (Ct value) is then generated, which are inversely proportional to the viral concentration, determining the patients' infection status. The validate sensitivity and specificity of RT-PCR in COVID detection can reach as high as 90% and 95%, together with a LOD of 10-100 copies/mL (X. Wang et al., 2020). Owing to those features, RT-PCR for COVID detection is currently the most reliable and widely used technique. Nevertheless, requirements for professional equipment and operators raise the cost of RT-PCR based test, the preparation of primers for different SARS-CoV-2 variant also increases the complexity. Additionally, RT-PCR usually takes a long time to generate the outcome, which impedes its application as a rapid and affordable choice.

As a well-known characterization approach for EVs, ELISA also performs well in COVID detection. ELISA works based on the interaction between viral antigen and fluorophore-labeled

antibody (Van Elslande et al., 2020). The enzyme catalyzes a reaction producing a detectable signal, such as a color change or fluorescence signal. Commercial ELISA based COVID detection kits have been prevalently used as daily checking methods. It has the advantages of rapidity and cheapness, however, the sensitivity and LOD are much lower than RT-PCR (Kasetsirikul et al., 2020), especially when the viral load of the specimen is extremely low, leading to false negative for most of the early-staged patients. Therefore, ELISA is often used in conjunction with other diagnosis tests such as RT-PCR or serology tests, to confirm COVID cases.

Optical technologies have made significant progress in COVID detection (Lukose et al., 2021), providing another possibility for rapid and early diagnosis. Optical technologies generally focus on the spectroscopic responses that produce distinct spectral signatures. In contrast to PCR and antibody-based detection techniques, optical methods are usually label-free and non-invasive. D. L. Kitane and S. Loukman et al achieved 97% sensitivity and 98% specificity in quantitative COVID detection, validated by 280 clinical patient samples, based on multivariate analysis of Fourier Transform Infrared Spectroscopy (FTIR) of RNA extracts (Kitane et al., 2021). SARS-CoV-2 possesses unique spectral features located at 600-1350 cm^{-1}, 1500-1700 cm^{-1} and 2300-3900 cm^{-1}, they are attributed mainly to the viral RNA nucleobases. Raman spectroscopy is a similar spectroscopic characterization method as FTIR, which extracts the molecular vibrational modes by Raman scattering. C. Carlomagno et al extracted SARS-CoV-2 virus Raman fingerprints and applied a deep learning model for pattern identification that achieves above 95% accuracy (Carlomagno et al., 2021). Moreover, their choice of salivary specimen makes it more rapid and non-invasive. Similarly, S. A. Jadhav conducted works on integrating microfluidic platform with SERS for COVID detection, which potentially improves the LOD by immobilizing SARS-CoV-2 by antigen-antibody binding (Jadhav et al., 2021). Surface plasmon resonance (SPR) sensing chip

is also one of the approaches that will be used to diagnose SARS-CoV-2 in the near future. Work done by T. Akib and S. Mou fabricated a graphene-based multi-layer (Bk_7/Au/$PtSe_2$/Graphene) coated SPR biosensor for rapid detection of COVID. The performance of biosensors was evaluated numerically with different ligand-analytes and the optimized device improved sensitivity by adding layers of graphene (Akib et al., 2021).

There are a lot of novel technologies giving high sensitivity and additional advantages. G. Soufi and S. Iravani applies molecularly imprinted polymers on detection/recognition of a wide variety of viruses, they proved enough specificity, convenience, validity and reusability features on SARS-CoV-2, Human rhinovirus, Hepatitis A and B viruses, influenza A viruses etc. (Jamalipour Soufi et al., 2021). D. Haritha conducted chest X-ray imaging on COVID patients and designed an infection recognition model based on deep learning, CheXNet, for COVID detection. They successfully trained the model that detects 14 pathologies in the chestXray 14 dataset with 99.9% accuracy (Haritha et al., 2020). NGS, viral culture and mass spectrometry are additional powerful methods to extract information either from viral nucleic acids or proteins. According to the diagnostic requirement, appropriate technologies are supposed to be chosen to achieve optimal performance.

3.3 Acquisition of nano-bioparticles

3.3.1 Exosome isolation

Exosomes are released by most of the cells and are widely distributed in body fluids or cell culture media. The protocol of isolating exosomes from different sources is determined by factors including exosome concentration, density, viscosity, volume etc (P. Li et al., 2017). Most common exosome sources include serum, plasma, cell culture media, malignant pleural effusion (MPE), saliva, urine etc. Sequential ultracentrifugation (SUC) and size exclusion chromatography (SEC)

were utilized during the process, both methods rendered highly concentrated NBPs that were validated by transmission electron microscopy (TEM) and nanoparticle tracking analysis (NTA). We provided a common exosome isolation procedure from cell culture media as an example based on SUC below (Théry et al., 2006), as shown in Figure 3.4.

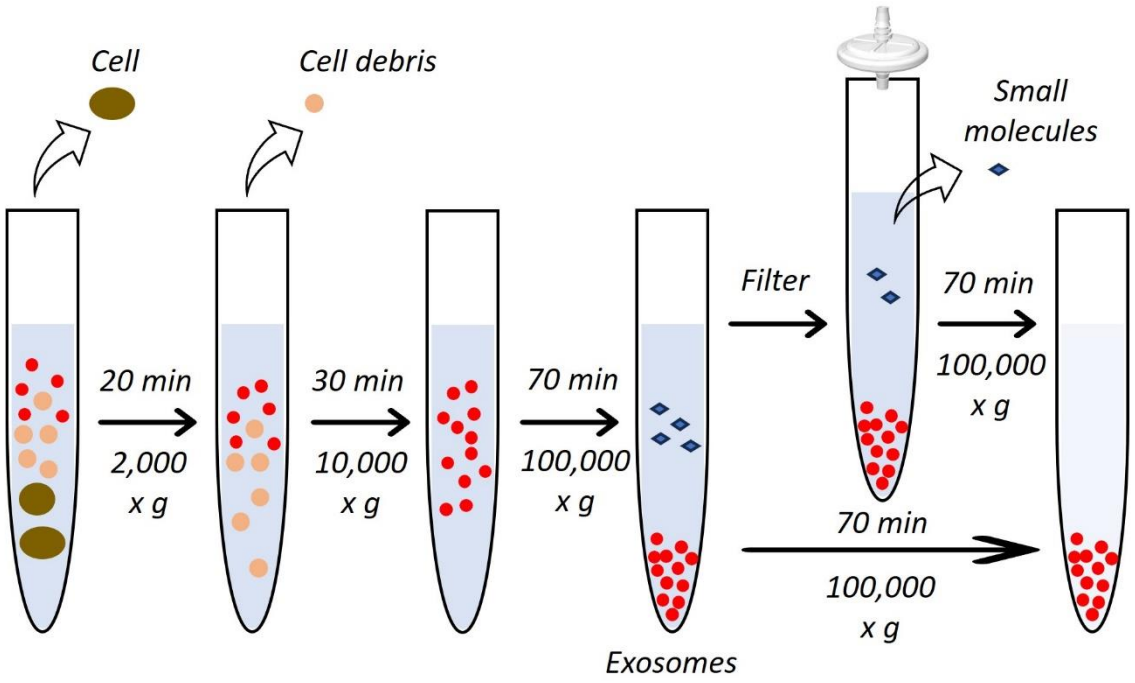


Figure 3.4 Procedure of isolating exosomes from body fluids based on ultracentrifugation.

Materials:

Cleared, conditioned medium,

Phosphate-buffered saline (PBS),

Beckman Optima TLX ultracentrifuge and TLA-100.3 fixed-angle rotor,

Polyallomer ultracentrifuge tubes,

Micropipette.

Steps:

1) Remove cells, dead cells and cell debris

a. Transfer the cleared, conditioned medium to centrifuge tubes,

b. Centrifuge 20 min at 2,000 × g and 4 ℃,

c. Pipet off the supernatant and transfer to ultracentrifuge polyallomer tubes,

d. Centrifuge 30 min at 10,000 × g and 4 ℃,

2) Collect exosome fraction

e. Transfer the supernatant to fresh tubes as step d,

f. Centrifuge at least 70 min at 100,000 × g and 4 ℃,

g. Remove the supernatant completely and exosomes are supposed to attach on tube wall,

3) Wash exosomes

h. Resuspend the pellet in PBS using micropipette,

i. Centrifuge 1 hour at 100,000 × g and 4 ℃,

j. Remove the supernatant completely as possible,

k. Repeat step j to concentrate exosomes (optional),

l. Resuspend the pellet in 20 – 50 μL PBS and store up to 1 year at -80 ℃.

Isolation of exosomes from viscous body fluid (such as plasma, saliva etc.) is slightly different from other sources (cell culture media, urine etc.) due to different viscosity and chemical contents. 0.22 μm filter devices are often needed for ultrafiltration.

Steps:

1) Dilute fluid with PBS and centrifuge 30 min at 2,000 × g and 4 ℃,

2) Transfer supernatant to ultracentrifuge tubes without pellet contamination,

3) Centrifuge 45 min at 12,000 × g and 4 ℃,

4) Transfer supernatant to fresh ultracentrifuge tubes and remove pellet,

5) Centrifuge 2 hours at 110,000 × g and 4 ℃,

6) Remove the supernatant and resuspend pellet in PBS,

7) Filter the suspension with 0.22 μm filter and collect in fresh ultracentrifuge tubes,

8) Centrifuge 70 min at 110,000 × g and 4 ℃, pour off the supernatant,

9) Resuspend the pellet and centrifuge 70 min at 110,000 × g and 4 ℃,

10) Resuspend the pellet in 20 – 50 μL PBS and store up to 1 year at -80 ℃.

In addition to SUC, IZON is one of the SEC commercial products that renders high purity and concentration (Patel et al., 2019). Automatic fraction collector (AFC) and qEV original 500 μL columns are used. qEV SEC columns separate particles based on their size as they pass through a column packed with a porous polysaccharide resin. As the sample passes through the column under gravity, smaller particles enter the resin pores on their way down then their exit from the column is delayed. Sequential volumes are collected after they exit the column and particles will be distributed across the volumes based on their size.

Materials:

Exosome specimen,

Phosphate-buffered saline (PBS),

IZON columns,

Polyallomer ultracentrifuge tubes,

Micropipette.

Steps:

1) Equilibrate the column and the samples buffer to be with in the operational temperature range of 18 – 24 ℃,

2) Remove the top cap and attach the column to the AFC,

3) Remove the bottom cap and allow the buffer to start running through the column,

4) Flush the column with at least two column volumes of PBS buffer to minimize the effects of sodium azide,

5) Load the prepared centrifuged sample onto the loading frit,

6) Immediately start collecting the buffer volume,

7) Allow the sample to run into the column, the column will stop flowing when all of the sample has entered the loading frit,

8) Top up the column with buffer and continue to collect the buffer volume,

9) Once the buffer volume is collected, continue to collect the Purified Collection Volume (PCV),

10) After NBPs fractions have been collected, clean and sanitize the column with 0.5 M NaOH to remove residual proteins,

11) Flush the column with PBS buffer and store for future usage.

3.3.2 SARS-CoV-2 isolation

SUC and SEC based protocol can also be used for SARS-CoV-2 isolation due to its similar size and other attributes. SARS-CoV-2 pelleting protocol is given below (Plavec et al., 2022). Figure 3.5 demonstrates the basic procedures.

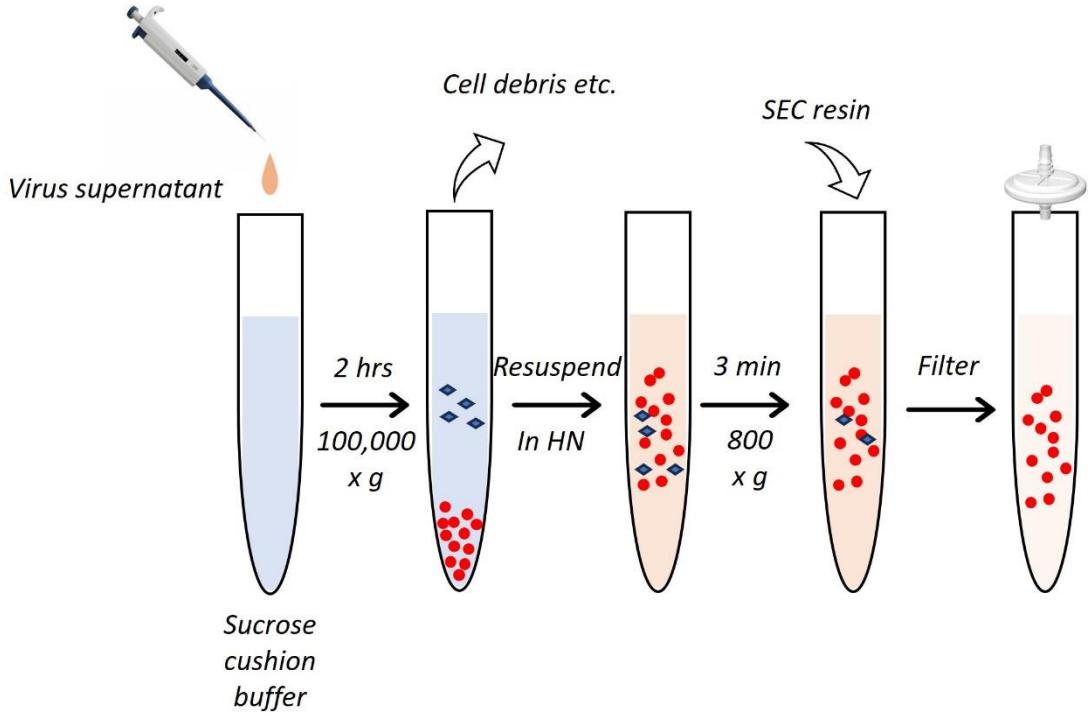


Figure 3.5 Procedure of isolating viruses from cell culture media based on ultracentrifugation.

Materials:

Virus infection cell culture media,

Phosphate-buffered saline (PBS),

SEC resin,

HN buffer,

100 kD filter,

Micropipette.

Steps:

1) Layer pre-cleared virus-containing supernatant on top of sucrose cushion in buffer including HEPES and NaCl.

2) Centrifuge 2 hours at 100,000 × g and 4 ℃,

3) Discard the supernatant and rinse the pellet with HN to remove leftover sucrose,

4) Resuspend the visible pellet in HN and store at -80 ℃.

The SEC protocol works in conjunction with centrifugation.

1) Add SEC resin to pre-cleared supernatant and rotate 20 min at 4 ℃ for mixing homogeneously,

2) Centrifuge the resin for 3 min at 800 × g and 4 ℃,

3) Collect the virus-containing supernatant,

4) Repeat step 1 to 3 to purify the virus-containing supernatant,

5) Filter the supernatant using 100 kD ultrafiltration to concentrate the virus specimen.

3.4 Surface-enhanced Raman Spectroscopy

3.4.1 Raman scattering

Raman spectroscopy is an efficient spectroscopic technique applied to observe vibrational, rotational, and other low-frequency modes in a system (Mulvaney & Keating, 2000). The basic physical mechanism in Raman spectroscopy is inelastic scattering between photons and phonons induced by monochromatic light, usually laser. The interaction between photons from laser and

molecular vibrational modes results in energy (frequency) of the laser photon shifting, according to the energy of scattered photons, the inelastic scattering is categorized into Rayleigh scattering, Stokes scattering and anti-Stokes scattering. There is no frequency shift for Rayleigh scattering, only the direction of the scattered photon changes, which is elastic scattering.

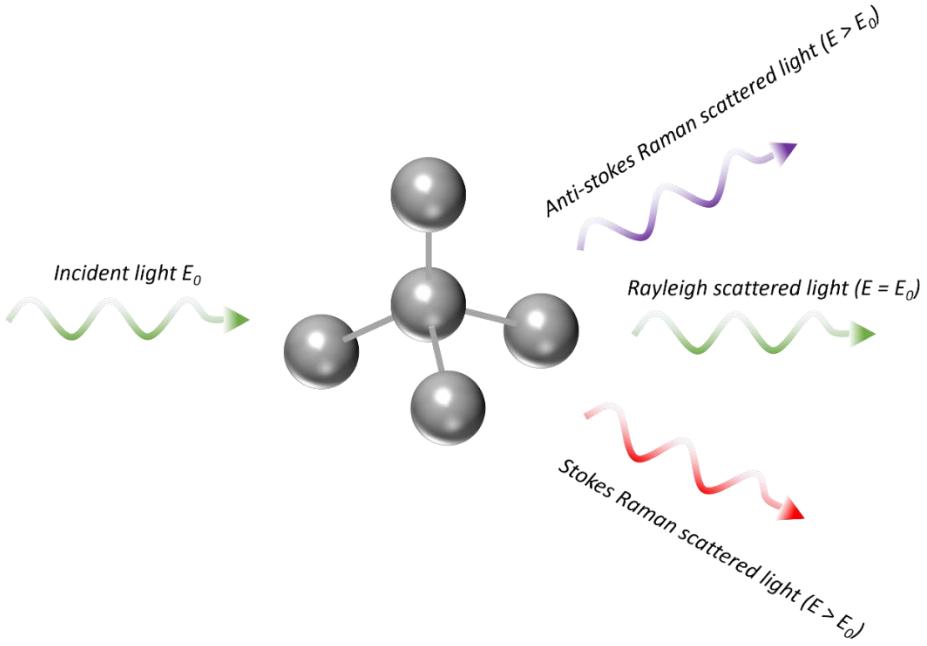


Figure 3.6 Diagram of physics process of Raman scattering.

Stokes scattering refers to the outcome that scattered photon has lower energy than the incident photon, and vice versa for anti-Stokes scattering. Different from fluorescence, which involves molecular energy transition between excited state and ground state, Raman scattering is due to phonon excitation and emission by interacting with photons, as shown by Figure 3.7.

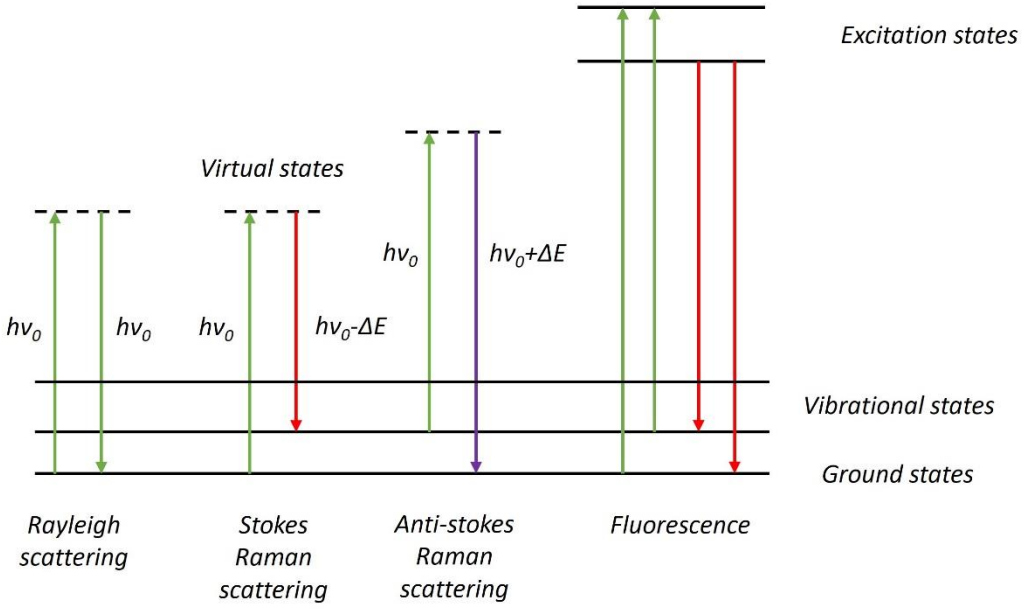


Figure 3.7 Energy-level diagram of that states in Raman spectra.

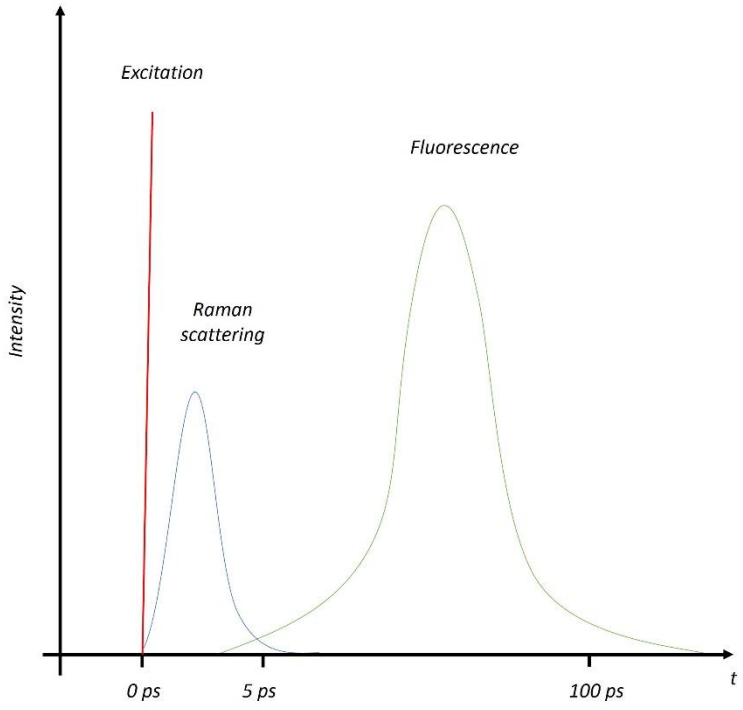


Figure 3.8 Temporal variation of excitation, fluorescence emission, Raman scattering.

In Raman spectroscopy, Raman intensity is plotted against Raman shift, which refers to the frequency change. Frequency shift of photons are determined by the molecular structural configuration, therefore specific Raman spectrum peak patterns provide "fingerprint" for molecules. For example, Figure 3.9 shows graphene possessing fingerprint peaks at 1325 cm^{-1} (D peak), 1589 cm^{-1} (G peak), and 2644 cm^{-1} (2D peak). The specadfstral fingerprint range of organic molecules is 500-1500 cm^{-1} (Fesenko et al., 2015).

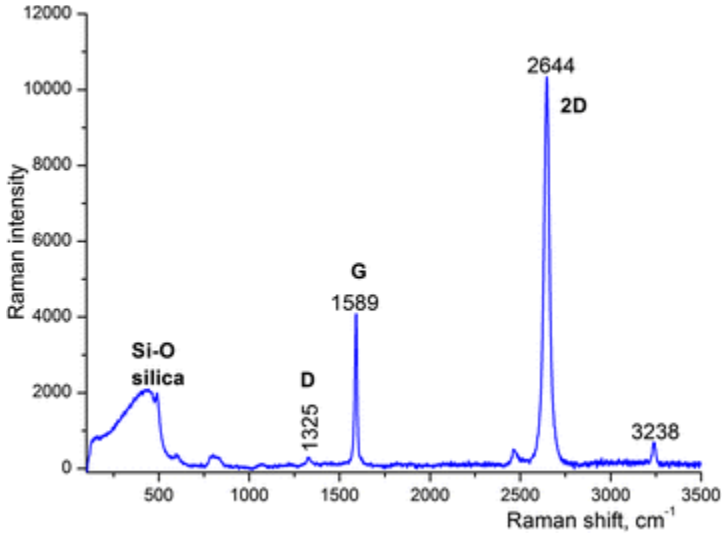


Figure 3.9 Raman spectra of single-layer graphene at 633 nm.

3.4.2 Surface enhancing mechanism

Since the surface enhancing phenomenon was first observed by M. Fleischmann and his colleagues in 1973, the theory and application have been rapidly developing in the past few decades (Cialla et al., 2012; Fleischmann et al., 1973). EM theory and chemical theory are proposed to explain the exact mechanism (Etchegoin & Le Ru, 2010). SERS has been utilized in many fields, such as chemical analysis and biosensing, due to its extremely high sensitivity and specificity (Cialla et al., 2012).

SERS effect is essentially due to the oscillation of metallic electron in the background of ionic metal cores induced by the time-varying electric field of incident light (Etchegoin & Le Ru, 2010). A small, isolated, illuminated metallic surface will form LSPR in response to the oscillating external electromagnetic field. Dipolar plasmon plays the most significant role in surface plasmon when the metallic particles are much smaller than the wavelength of the incident light, which can apply to most of the materials with free or nearly-free electron. When the dipoles are oscillating resonantly against the incident light, dipolar radiation occurs, as demonstrated by Figure 3.10. Certain positions near the metallic surface have greatly enhance EM field while others are deleted due to the coupling effect between the excited EM field and the incident light, this EM field reaches equilibrium in a few femtoseconds after the light induces. The Raman scattering is thus enhanced by the dipolar radiation together with the incident light. The SERS intensity is enhanced by a factor approximately equal to fourth power versus the local incident near field.

$$I_{SERS} \cong |E_{incident}|^4 \cdot I_{incident} \qquad 3\text{-}1$$

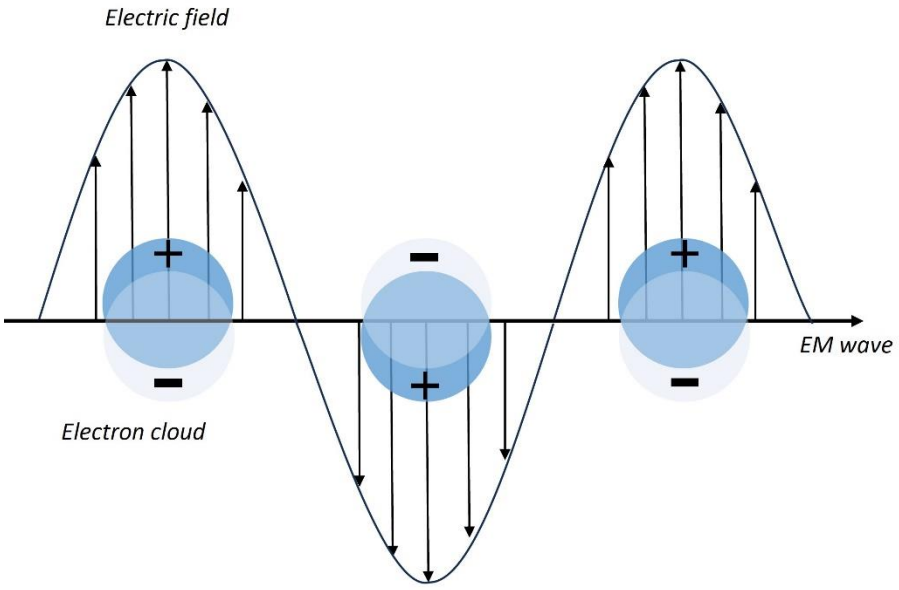


Figure 3.10 Diagram of plasmon resonance showing oscillation of electrons excited by the electromagnetic field of incident light.

SERS enhancing factor includes both linear and nonlinear optical effects, determined by the power of incident light $I_{incident}$. Typical enhancement factors range from 10^6 to 10^{15}, depending on the structure of rough surface, materials, and incident light (Le Ru et al., 2008). Additionally, SERS excitation is a near-field effect, which exists especially near the metallic surface, usually in the magnitude of nanometers. The electric field decays exponentially spatially off the metallic surface. Periodic nanostructures of metallic surface are currently the most widely used design for SERS substrate.

According to the mechanism of SERS, the enhancement of the electric field involves two parts, the plasmon resonance excitation, and the enhancement in polarizability due to chemical effects including charge-transfer excited states (Etchegoin & Le Ru, 2010). The first part gives an enhancement factor approximately equal to $|E(\omega)|^4$, $E(\omega)$ is the combination of local electric field enhancement factor at the incident light frequency and Stokes-shift frequency, which also depends

on the polarization of the dipole. The polarizability α of a small metallic surface is related directly to the dielectric function $\varepsilon(\lambda)$ and radius R,

$$\alpha = R^3 \frac{\varepsilon - 1}{\varepsilon + 2} \tag{3-2}$$

Combined with the dielectric function of Drude model, we have

$$\varepsilon = \varepsilon_b + 1 - \frac{\omega_p^2}{\omega^2 + i\omega\gamma} \tag{3-3}$$

Where the firm term ε_b is due to inter-band transition and is usually wavelength-dependent, ω_p is the metal's plasmon resonance whose square is proportional to the electron density, γ is the electronic-scattering rate which is inversely proportional to the electronic mean free path. Therefore, the polarizability equals to

$$\alpha = \frac{R^3\left(\varepsilon_b\omega^2 - \omega_p^2\right) + i\omega\gamma\varepsilon_b}{\left[(\varepsilon_b + 3)\omega^2 - \omega_p^2\right] + i\omega\gamma(\varepsilon_b + 3)} \tag{3-4}$$

So, the dipolar surface resonance occurs for a single metal particle in the case of

$$\omega = \frac{\omega_p}{\sqrt{\varepsilon_p + 3}} \tag{3-5}$$

The effect of interparticle coupling can positively increase the enhancing effect by specially design rough surface. Basically, when two dipoles are close enough, the mutual interaction of the

two nanoparticles leads to an increase in the magnitude of the electric field. The field of the incident light as well as its partner's intense field amplify the polarization for each nano-object. Well-engineered interacting nanostructure system can fulfill the goals of high field enhancement and reproducible SERS platform. Work done by Hao and Schatz shows the truncated tetrahedron structure with the dimension of 200 nm rendered a 4×10^{12} enhancement factor for single-molecule SERS (Hao et al., 2004). P. Wang and M. Xia demonstrated a SERS platform with hexagonally arranged nanopyramid resulting in enhancement factor around 10^{10} (P. Wang et al., 2013). J. Li and A. Wuethrich fabricated nanopillar array using lithographic approach made of gold-silver alloy which had enhancement factor approximately equal to 10^7 (J. Li et al., 2021).

The chemical mechanism first occurs by charge transfer due to the interaction between the enhanced electric field of localized surface plasmon and the target molecules (Morton & Jensen, 2009). It is molecule-specific and relies heavily on the local environment of the metal surface, as it arises from the overlap between the molecule's wave functions and the metal nanoparticle. This overlap leads to the renormalization of molecular orbitals and the emergence of new mixed charge-transfer states, both of which contribute to the chemical enhancement of the Raman signal. The chemical mechanism can be further classified into two types: non-resonant chemical mechanism (CHEM) and resonant charge-transfer chemical mechanism (CT).

CHEM arises from the interaction between the molecule and the metal surface, leading to the enhancement of the Raman signal (X. X. Han et al., 2022). CHEM is generally independent of the excitation wavelength and does not involve any electronic transition of the molecule. The CHEM mechanism is believed to arise from two main effects: the EM effect and the chemical effect. The EM effect arises from the strong local electric field produced by the metal surface, which enhances the excitation and Raman scattering cross-section of the molecule. The CE effect

arises from the chemical interactions between the molecule and the metal surface, which modify the molecular vibrations and their Raman scattering cross-section. CT occurs when the incident photon energy matches a molecular or charge-transfer excitation of the system. In this case, the molecule can undergo a charge transfer with the metal surface, resulting in the formation of a new charge-transfer state. This new state has a different Raman scattering cross-section than the ground state, leading to an enhancement of the Raman signal. The CT mechanism is generally more specific to the molecule and the metal surface than the CHEM mechanism, as it depends on the resonant excitation energy and the specific electronic structure of the molecule and the metal surface. Overall, both the non-resonant CHEM and the resonant CT mechanisms contribute to the enhancement of the Raman signal in SERS, with the CHEM mechanism generally dominating when the excitation wavelength is far from any molecular or charge-transfer excitation.

3.4.3 Advantages of SERS

SERS, as an NBPs characterization technique, provides many advantages over the other technologies. It serves as a complementary method to the typical proteomics and genomics to investigate the NBPs comprehensively. Due to the EM effect and CM effect, SERS can provide a million-fold enhancement to the sensitivity, making it feasible to detect trace amounts of molecules. Researchers have proven the capability of single-molecule detection using SERS, which has been a hot topic in physics, chemistry, and biology (J. Kneipp et al., 2008). K. Kneipp and Y. Wang exploited the scattering cross section (10^{-17}-10^{-16} cm^2/molecule) from SERS by studying single crystal violet molecule in aqueous colloidal sliver solution and observed the phenomenon of single molecule Raman scattering (K. Kneipp et al., 1997). S. M. Stranahan conducted more complicated studies on SERS hot spots using super-resolution optical imaging methods and observed the coupling between single molecule dipolar scattering and near-field hot spots (Stranahan & Willets,

2010). The investigation of single molecule can give information of its microscopic properties and structural transformation that help better understand the nature of molecular processes, which is not possible with ensemble of molecules because of averaging outcome. Enhancement of signal in SERS compensates the drawback of Raman scattering in characterizing single molecule-extremely small cross section (typically around 10^{-30}-10^{-25} cm^2), making it a good candidate to provide high degree of structural conformation of molecule. The spatial resolution of SERS can reach down to the scale of nanometers with appropriate substrate, it can be higher by combining with scanning probe microscopy techniques, such as atomic force microscopy (AFM), scanning tunneling microscopy (STM). Nowadays SERS opens novel perspectives in monitoring NBPs at the single particle level and brings exciting opportunities in biochemistry and biophysics. Given that tremendous amounts of particles/vesicles from different sources produced by human body, SERS based single vesicle technology will play a significant role in detecting biomarkers at low concentrations.

Raman scattering provides vibrational information through the interaction between phonons and photons, which depends on the molecular bonding configuration (Mulvaney & Keating, 2000). This yields valuable structural information, including bond length, angles, and vibrational modes, allowing for the identification of unknown compounds. SERS benefits from this fingerprinting property, increasing its selectivity in identifying target molecules in the presence of irrelevant ones. Similar in NBPs characterization, SERS benefits identifying target NBPs containing target biomarkers according to the unique spectroscopic signature incorporated by single vesicle characterization. Moreover, LSPR resolves the small cross-section problem, making SERS highly versatile for characterizing a wide range of samples, including solids, liquids, gases, and biological specimens. Typically, there is usually no label specifically required during Raman

test, whose function is basically facilitating the target molecules to give out recognizable signals, such as immunofluorescence staining, molecular beacons, mass spectrometry etc. The label-free nature of SERS simplifies specimen preparation, eliminating the need for labeling procedures that are typically a key step in other experiments. Instead, research attention shifts to substrate design and fabrication, peak pattern classification, and establishing the link between SERS signals and research objectives.

3.4.4 Substrate design and fabrication

The basic structure of our SERS substrate is quasi-periodic gold nano-pyramidal structure that is manufactured based on polystyrene sphere lithography approach. This design is determined through considering several factors including biological specimen compatibility, hot spot spatial location and enhancement factor, fabrication complexity and product quality, and repeatability.

The fabrication process refers to the SOP published previously by our group (P. Wang et al., 2015), shown in Figure 3.11. SiO_2/Si was washed by Piranha solution (H_2SO_4:H_2O_2 = 3:1, volume ratio) for 1 h at 70 °C, followed by rinsing with deionized water 3 times. Polystyrene spheres (Thermo Fisher Scientific, USA) were then applied to construct a monolayer on SiO_2/Si wafer (MSE Supplies, USA) surface via self-assembly to create hexagonal patterns. Subsequently, the substrate was dry etched by O_2 plasma under 200W for 50 s to shrink the polystyrene sphere size. The reduced polystyrene spheres act as the mask in the plasma etching process to remove the SiO_2 layer under exposure. Subsequently, the substrate was etched in 60% KOH solution (Sigma Aldrich, USA) for 2 mins to form periodic pyramidal reciprocal structures on the Si layer with patterned SiO_2 as a mask. A 200nm Au film was deposited on the mode and finally, epoxy was used to peel off the Au film which was attached to a new Si wafer. On the fabricated platform,

quasi-periodically hexagonally arranged Au nano-pyramids with base length of 200 nm, height of 200 nm was obtained.

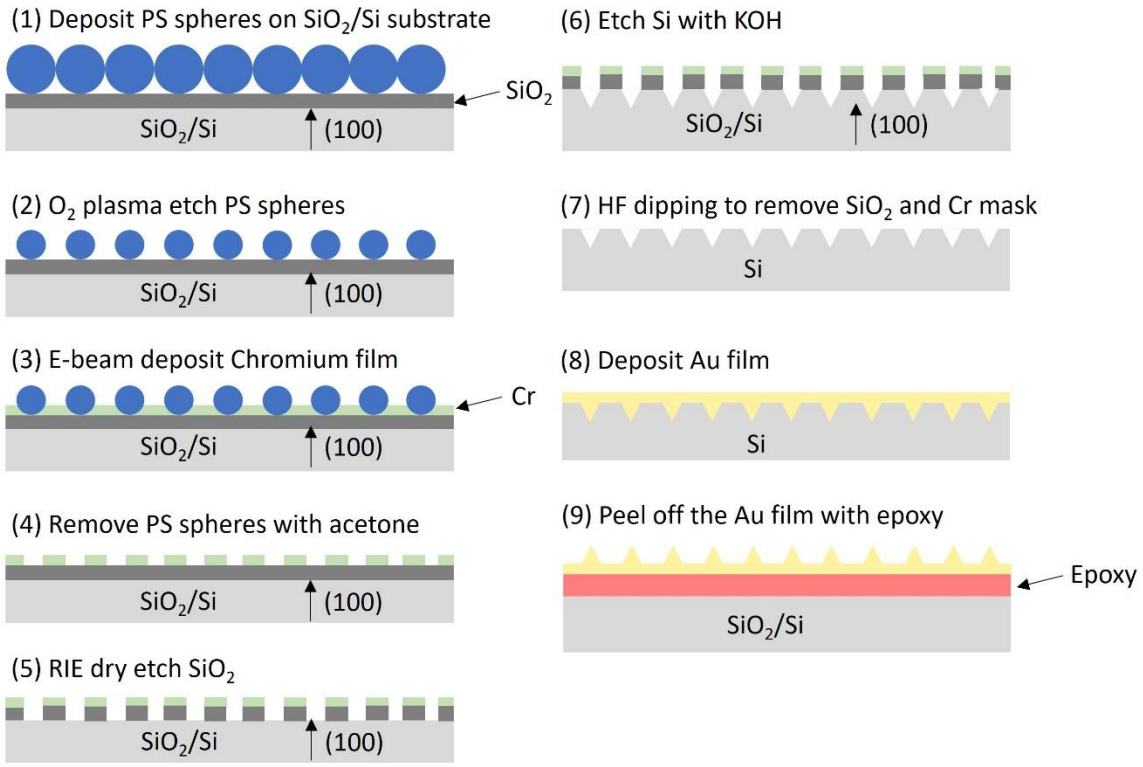


Figure 3.11 Procedure of fabricating periodic Au pyramidal SERS substrate based on lithography.

Scanning Electron Microscope (SEM, FEI Nova NanoSEM 230) was applied to evaluate the pyramidal shape of a single unit as well as the substrate surface pattern. Rhodamine 6G (R6G) was used as a Raman reporter to quantify the overall enhancement factor compared with plain gold substrate. Figure 3.12 shows the structure and SEM imaging of substrate, it proves well-formed pyramidal units with hexagonal arranging pattern.

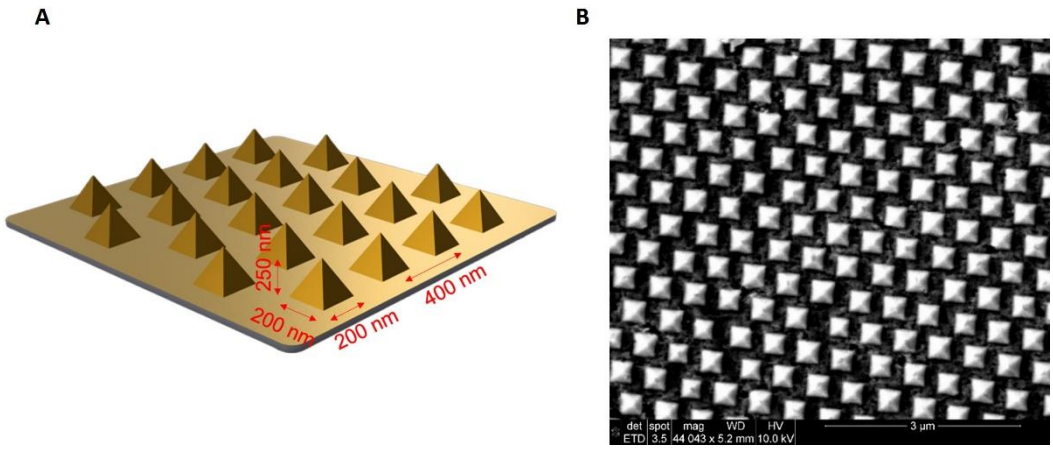


Figure 3.12 Au pyramidal structure. (A) Diagram of Au pyramidal SERS substrate surface structure. (B) SEM image of Au pyramidal SERS substrate characterization at 44043 ×.

We implemented FDTD (Finite difference time domain) simulations on calculating the spatial distribution of hot spots on the pyramidal surface. FDTD is a powerful tool to solve Maxwell's equations numerically to probe the EM field above the substrate surface (Zeng et al., 2016). We conducted EM field simulation with an infinite 2-dimensional (set by periodic boundary conditions) gold pyramidal pattern located on a thin layer of gold. Total-field scattered-field was set up as the laser source that is perpendicular to the substrate, with the electric field along the direction of x axis. FDTD was placed covering a 1200 nm (length) × 800 nm (width) × 800 nm (height) space, monitors were placed perpendicular to the z axis off the substrate by 100 nm, and perpendicular to y axis through the apexes of pyramids. The absolute intensity of electric field is shown in Figure 3.13B and Figure 3.13C from two different angles. As stated previously, the hotspots are located at the lateral facets of every single pyramid. Lateral facets facing the direction of x axis or -x axis have enhanced electric field, while the ones along the y axis have no enhancing outcome.

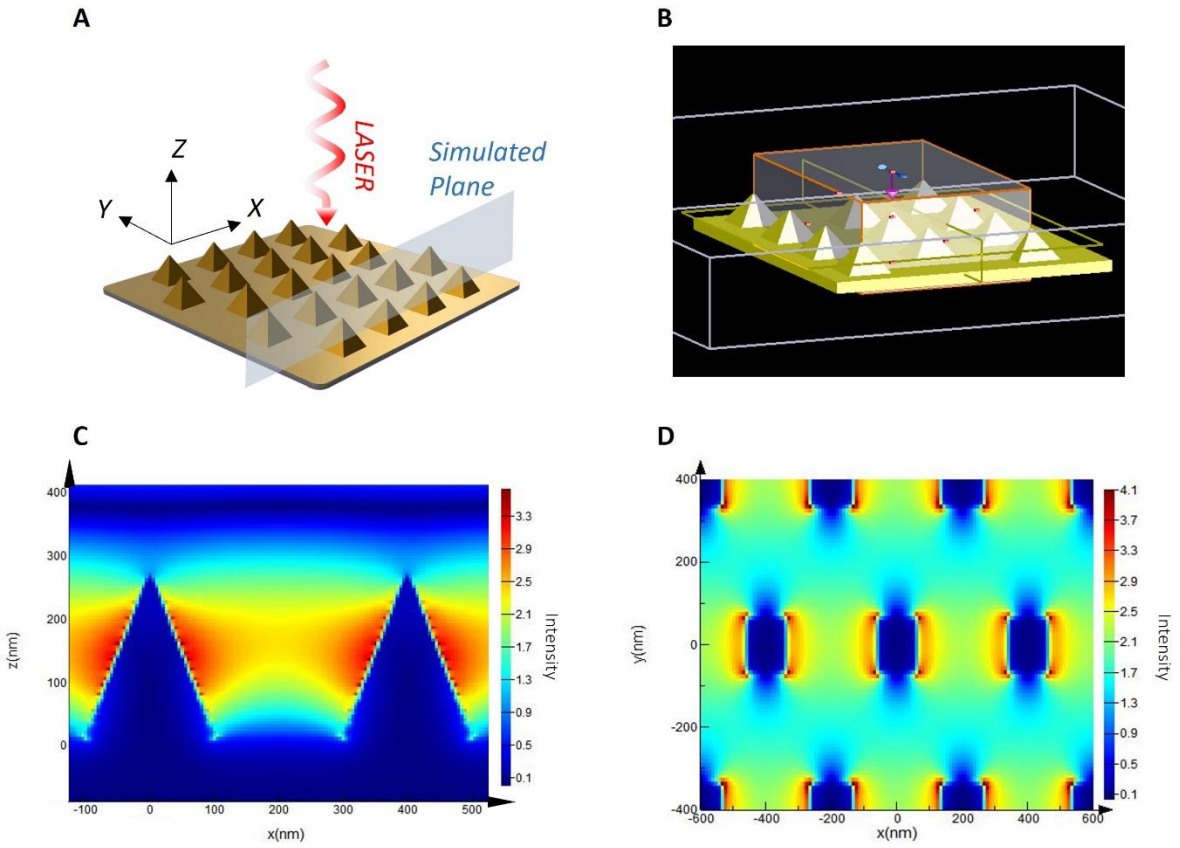


Figure 3.13 Au pyramidal substrate FDTD simulation. (A) Diagram of FDTD simulation of electromagnetic field of SERS platform. (B) Demonstration of FDTD simulation setup. (C) Electric field intensity distribution on x-z plane according to FDTD simulation results. (D) Electric field intensity distribution on x-y plane 100 nm above Au surface according to FDTD simulation results.

R6G molecule' signals in Figure 3.14 reveal significantly enhanced peaks' intensity as well as the SNR of spectrum, defined by the average peaks' intensity divided by the average random noise intensity.

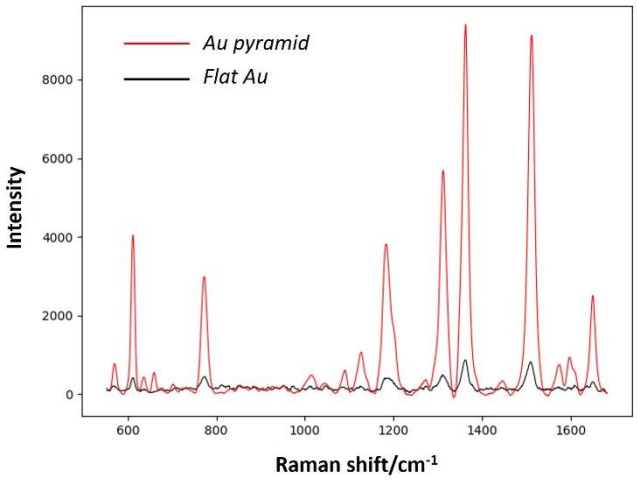


Figure 3.14 Raman spectra of R6G molecules on Au pyramidal substrate versus flat Au substrate.

3.5 Spectroscopic data collection

3.5.1 Raman map acquisition

Raman spectral data were immediately recorded using Raman spectrometer (Renishaw inVia Confocal Raman spectrometer, UK) under ambient conditions (20 °C, 1 atm), which is manually controlled by WiRE4.4 PC software. The map image acquisition function incorporated in the software was primarily used to collect numerical spectral data. A scanning then generating map acquisition approach is implemented in the NBP characterization to allow single particle detection. The setup of the maps is given in Table 3.1. The common parameters are 50× objective lens, laser wavelength 785 nm, Raman shift range 564-1680 cm^{-1}. Basically, a large square map (scanning map) covering an area of 300 μm by 300 μm was used to search for positions with NBPs' signals, which are supposed to show certain featured peaks. Those positions were recorded then characterized by a small square map (obtaining map), which would produce potential qualified spectra for followed-by data analysis. Having tested with different characterization approaches, we found our current method gives the most high-quality spectra in the same time duration. Figure

4.3(C) shows the spectra from a single NBP and Figure 4.3(D) shows the intensity map drawn around a recorded position at 1461 cm^{-1}.

The parameter combination of laser power and exposure time were determined to prevent sample burning, ensure spectra quality (SNR and fluorescence background), and data collection efficiency. Due to the uneven surface of the specimen, the maximal scanning map length and width were both 300 nm to avoid out-of-focus issue. 10 nm step size was chosen to avoid NBP repetitive scanning by the neighboring scanning points. For obtaining map, 5 nm × 5 nm map size is large enough to collect all qualified spectra belonging to a single NBP. As shown in Table 3.1, the total time for one cycle (scanning map plus obtaining maps) ranges from 12 minutes to 30 minutes.

Table 3.1 Parameters of scanning map and obtaining map.

Map type	**Laser Power**	**Exposure time**	**Map length**	**Map Width**	**Step size**	**Total time**
Scanning	50 mW	0.1s	300 nm	300 nm	10 nm	12 min
Obtaining	10 mW	0.5s	5 nm	5 nm	1 nm	25 sec

3.5.2 Automation of Raman measurement

We noticed that the repetitive style of Raman measurement could be accomplished by computer program engineering. To save the manpower and increase the scale of database, we made a python program enabling automated running scanning and obtaining maps. The basic procedure is demonstrated in Figure 3.15, and the cores of the program are the automated camera focusing section and the potential signal position selection algorithm. As the manual work, the automated measurement begins by several preparations including origin initialization, map acquisition map

initialization. Then the scanning map starts by controlling the software-hardware interaction. Regarding the potential signal position selection step, an algorithm calculating the numerical SNR was implemented to determine the keep-or-drop, the details are given in Section 3.6.1. A threshold of 8 was used to draw the boundary of keep-or-drop. Overall, by implementing the signal position selection algorithm, we could retain around 80% of the NBPs' positions compared to the manual work. As shown in Figure 3.15, obtaining maps runs by moving the sample staged according to the pre-calibrated coordinates. Recorded positions are characterized by the obtaining map template. Coming to moving the scanning map, a lens focusing algorithm based on computing the image sharpness is applied to adjust the stage height. The sharpness calculation is given by,

$$\boldsymbol{G}_x = \nabla_x \mathcal{L} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{x}}, \boldsymbol{G}_y = \nabla_y \mathcal{L} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{y}} \tag{3-6}$$

$$Sharpness = \frac{\sqrt{\sum_{i,j}^{N_L \times N_H} (\boldsymbol{G}_{xi}^2 + \boldsymbol{G}_{xj}^2 + \boldsymbol{G}_{yi}^2 + \boldsymbol{G}_{yi}^2)}}{N_L \times N_H} \tag{3-7}$$

Where $\boldsymbol{G}$ stands for gradient, L is the gray scale two-dimensional matrix of the image, N_L and N_H are the row and column length of the matrix. Through the automation, we can improve our data collection efficiency by a factor of 5 to 10 after summarizing 200 measurements, shown in Figure 3.16.

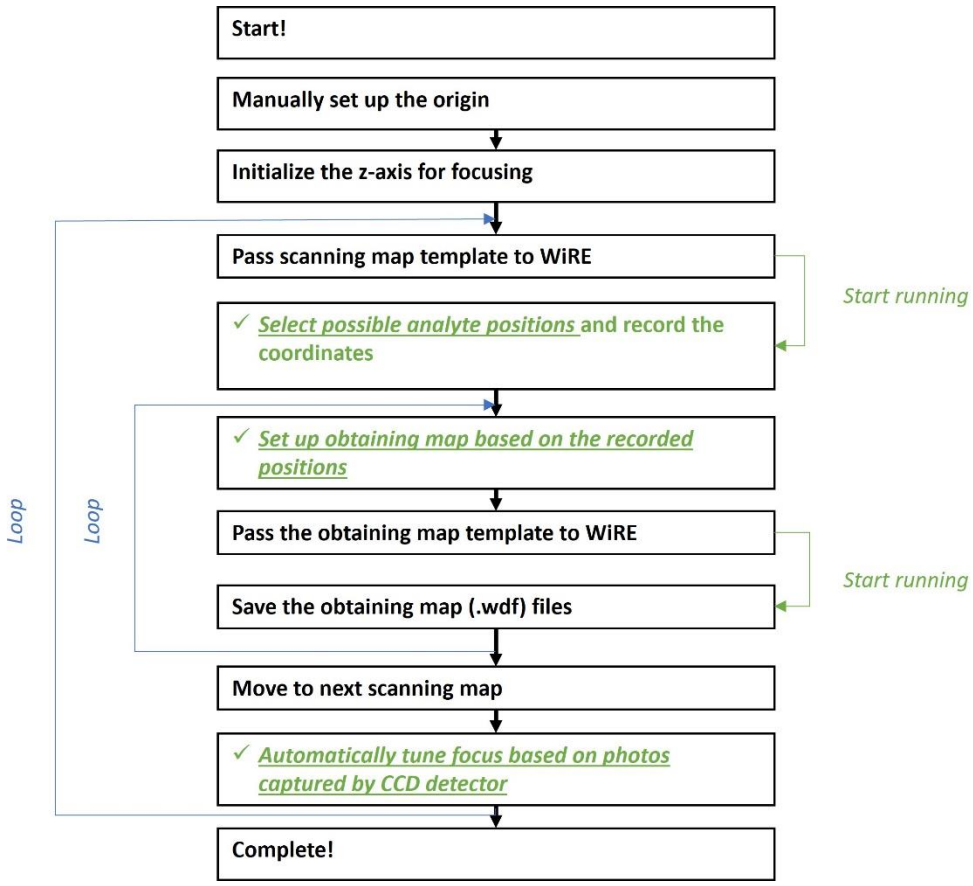


Figure 3.15 Workflow of automated SERS measurements.

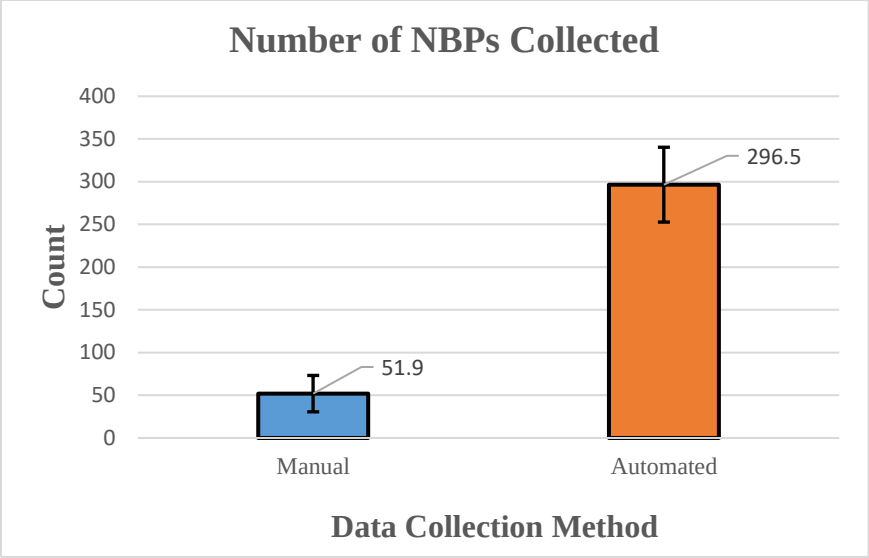


Figure 3.16 Number of NBPs collected. Comparison shows NBPs characterized between manual and automated measurements, the data throughput is increased by a factor 6.

The most significant part of this algorithm is currently selecting the potential analyte positions from the coarse map spectra, for which we chose to use a universal SNR calculation and cutoff by a predefined threshold. The performance of it directly determines our efficiency of collecting data, thus the subsequent data analyses. The current approach may not work for the specimens that have remarkable baselines (or peaks) at certain Raman bands. Those baseline or peaks contribute a lot to the spectral SNR even though these features are not informative to the NBPs. The non-informative baselines or peaks can be attributed to the factoring including solvent/solute, sample preparation, substrate preparation etc. Therefore, the selecting algorithm needs to specifically customize to deal with those special cases and ensure that informative NBPs' spectroscopic features are obtained to facilitate the single particle identification.

3.6 Data processing and analyses

Data processing is the final and most important step within the whole process. Given the enormous and complicated spectral data, well-designed and established analyzing algorithms are required to extract the most valuable and representative features to fulfill our purpose. Abundant features including relevant, irrelevant, redundant ones are usually present meanwhile. Infinite combinations of peak Raman band positions and intensity often requires a robust analytical approach that withstands unintentional variations. Conventional statistical tools (e.g., correlations, regressions etc.) based on bioinformatics and biostatistics are powerful techniques in data mining. Meanwhile, AI and machine learning demonstrate excellent capabilities in many fields, for example, image and voice recognition (Yuan et al., 2022), nature language processing (Mathews, 2019) and so on. Classification, clustering, regression, feature extraction and selection are widely used in spectroscopic data processing and analyses (Muto & Shiga, 2020). As such, we introduced several machine learning tools in our studies for different tasks and customized algorithms to optimize the performance. At the same time, a database of NBPs SERS spectral patterns is being established for efficient management and query. Currently our database is in dictionary format at the sample level, we are working on building a structured query language (SQL) configured database which incorporates all the available and informative NBPs' spectra. In addition, a succinct SERS spectral data processing graphical user interface (GUI) is under development.

3.6.1 Spectroscopic data quality control and preprocessing

A spectrum quality evaluation algorithm based on the SNR is implemented for spectral data quality control, by filtering noisy spectra without informative features. The SNR of a spectrum is calculated by the maximal peak intensity divided by the average noise level after baseline

subtraction. For more details, asymmetric least square (ALS) algorithm based baseline subtraction, Savitzky-Golay filtering based smoothing are involved in calculating the SNR.

3.6.1.1 Baseline (background) subtraction algorithm

The ALS method was used to subtract the fluorescence background of original Raman spectral data (Newey & Powell, 1987). A given spectrum is denoted by a vector $x = \{x_1, x_2, \ldots x_n\}$, the observed frequency (or wavelength) domain spectral intensities, typically a thousand components in our case. The smoothing series $z = \{z_1, z_2, \ldots z_n\}$ is faithful to $\boldsymbol{x}$, then the penalized least square's function is defined as the loss function in order to find the optimized solution to the original problem, which is composed of a fitness part and smoothness part:

$$\underset{\boldsymbol{W}}{\operatorname{argmin}} F = \underset{\boldsymbol{W}}{\operatorname{argmin}}[\sum_i w_i(x_i - z_i)^2 + \lambda \sum_i (\nabla^2 z_i)^2] \quad \text{3-8}$$

$$\Delta^2 z_i = (z_i - z_{i-1}) - (z_{i-1} - z_{i-2}), i = 1, 2, \ldots, n \quad \text{3-9}$$

Δ is a differential operator. The parameters w_i and λ are weight-fitness vector and penalty-control factor to tune the balance between smoothness and fitness. The definition of w_i is based on a parameter p typically in the range of 10^{-3} to 10^{-1},

$$w_i = \begin{cases} p, & (x_i - z_i) > 0 \\ 1 - p, & (x_i - z_i) \leq 0 \end{cases} \quad \text{3-10}$$

The optimized solution to equation 3-8 leads to a linear transformation:

$$(\boldsymbol{W} + \lambda \boldsymbol{D}^T \boldsymbol{D})\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x} \qquad 3\text{-}11$$

With $\boldsymbol{W}$ is the diagonal matrix for vector w, thus $\boldsymbol{W} = diag(w)$. $\boldsymbol{D}$ is the differential matrix, $\boldsymbol{Dz} = \Delta^2\boldsymbol{z}$. Basically, this method can estimate the true background but at the same time might eliminate the signal, therefore the balance parameters (λ is usually from 10^{-6} to 10^{-1}) should be finely tuned to minimize signal distortion. A baseline subtraction result is shown in Figure 3.17.

As stated, The baseline wandering problem occurs when the ALS baseline subtraction algorithm is confused by fluorescence background and Raman peak with large width. Residual baseline is observed to mask the featured peaks which could hinder the biomarker identification. In most of our preprocessing, fixed p and λ were used for the baseline subtraction tasks, which were chosen in an empirical manner. This method works for around 90% of cases, but for those specimens having a special baseline pattern, we need to modify the original algorithm to fit the rare cases. We are planning to introduce ML to learn the baseline patterns given a fluorescence database, then it will enable real-time adjustment of the baseline subtraction algorithm parameters. This adaptive capability will not only enhance the accuracy of our spectral analysis but also ensure that we can effectively address the outliers and special cases that were previously challenging to manage. Through this approach, we aim to make our spectral analysis more robust and versatile, ultimately advancing our capabilities in disease diagnosis and biomarker identification.

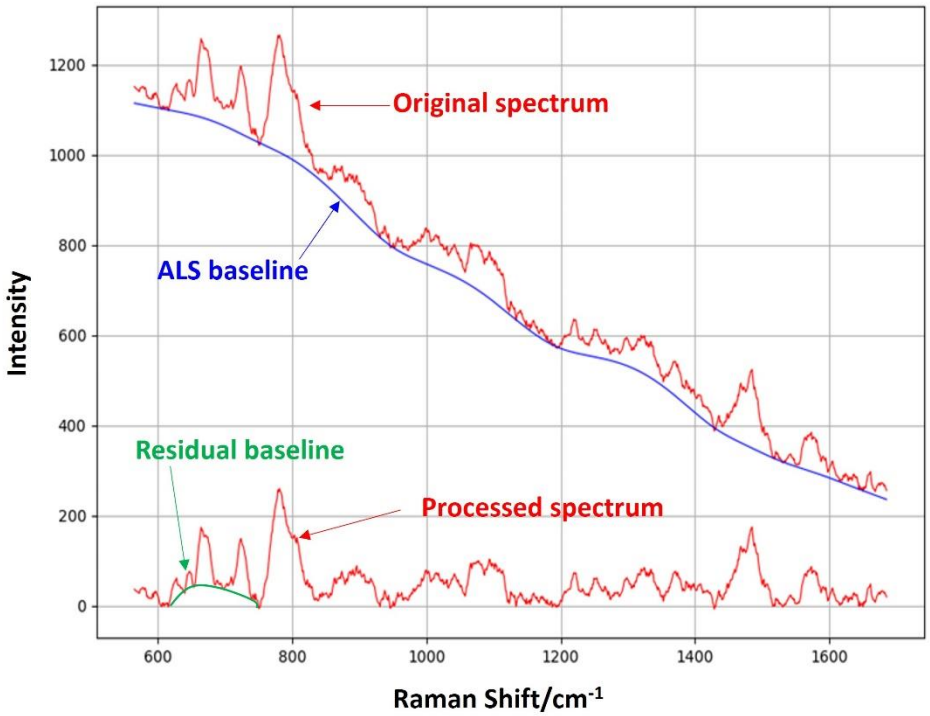


Figure 3.17 Demonstration of ALS based baseline subtraction. Original spectrum, baseline-subtracted spectrum, fitted baseline, and residual baseline are shown.

3.6.1.2 Smoothing (denoising) algorithm

Savitzky-Golay filtering method was applied to reducing the uncorrelated noise in Raman spectral data (Press & Teukolsky, 1990). This filtering algorithm was invented by Savitzky and Golay and attributed to least-squares smoothing that reduces noise while maintaining the shape and height of waveform peaks (e.g., Gaussian shaped spectral peaks).

The basic idea of Savitzky-Golay filtering is to conduct a local polynomial fitting in a preset shortened window of the whole-observed-wavelength-domain spectral intensities, then use the fitted value at the center of the window as the faithful transformation of the original corresponding value. A given spectrum $\boldsymbol{x}[n] = \{x_1, x_2, \ldots x_n\}$ is transformed to a faithful smoothing

series $\boldsymbol{p}[n] = \{p_1, p_2, \ldots\ p_n\}$ in the way that, for each component x_i in $\boldsymbol{x}$, considering its neighboring $2M$ samples centered at x_i, a polynomial function is applied to fit the $2M + 1$ samples:

$$\boldsymbol{p}[i] = \sum_{k=0}^{N} a_k i^k \qquad 3\text{-}12$$

With N is the order of the polynomial function. The optimized polynomial fitting can be found by minimizing the mean-squared approximation error for the group of input samples centered at x_i:

$$\underset{\boldsymbol{P}}{\operatorname{argmin}}\, \varepsilon = \underset{\boldsymbol{P}}{\operatorname{argmin}} \sum_{i=-M}^{M} (p[i] - x[i])^2 = \underset{a_k}{\operatorname{argmin}} \sum_{i=-M}^{M} (\sum_{k=0}^{N} a_k i^k - x[i])^2 \qquad 3\text{-}13$$

Where M is defined as the half-width of the approximation interval. Eventually, the smoothed output is obtained by evaluating $p[i]$ at the central point, that is $p[i]$, that is the 0^{th} polynomial coefficient. An example of smoothing result shown in Figure 3.18, which effectively filters out the messy noise.

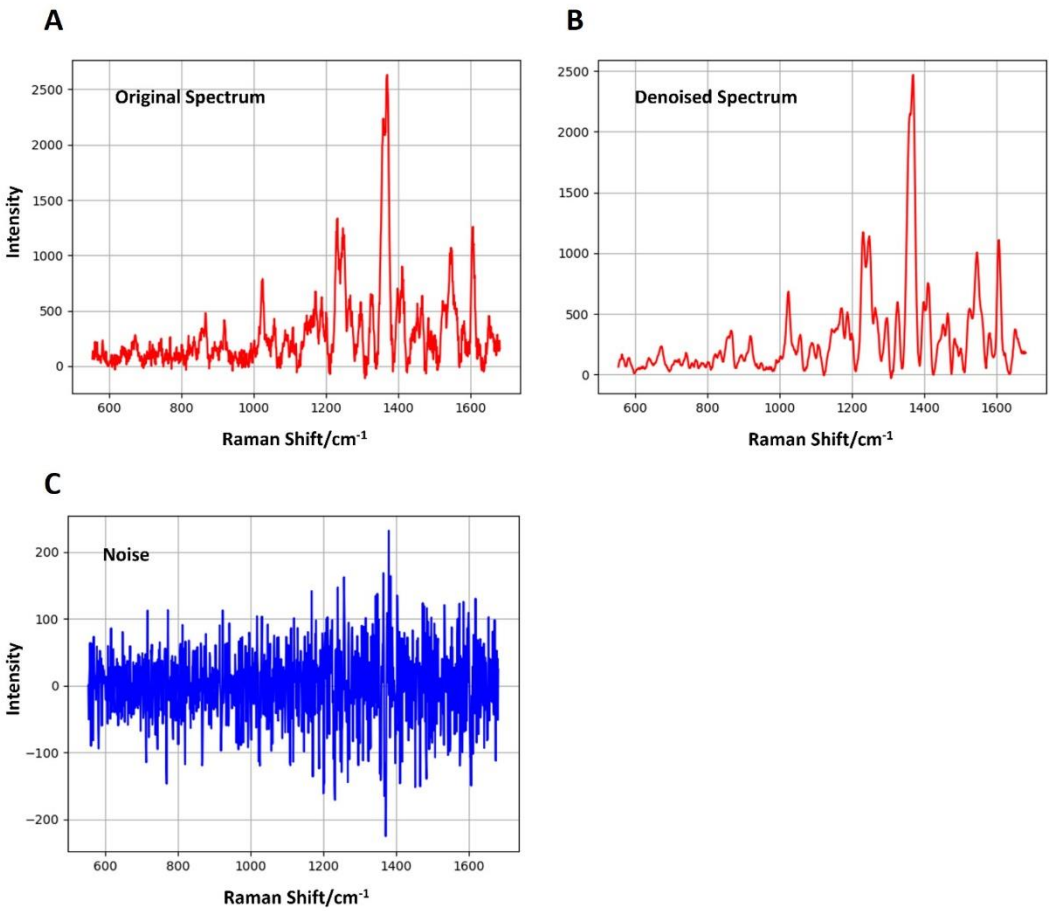


Figure 3.18 Demonstration of spectrum smoothing. (A) Original spectrum. (B) Smoothed spectrum. (C) Original spectrum subtracted by smoothed spectrum, i.e., noise spectrum.

3.6.1.3 Quality control

The spectrum quality evaluation algorithm is built on the ALS and filtering algorithms. The SNR of the baseline-subtracted spectrum serves as the metrics for evaluating the quality. As stated previously, the SNR is defined as the maximal peak intensity divided by the average intensity of the noise. By Savitzky-Golay filtering, a baseline-subtracted spectrum is split into the pure signal as well as the noise, as shown in Figure 3.18. Obviously, the SNR is given by

$$SNR = \frac{\max(Denoise\ spetrum)}{Avg(Noise)} \qquad 3\text{-}14$$

Practically, a spectrum with SNR higher than 25 has reasonable quality for presenting representative features, thus for data analysis. Figure 3.19 shows the SNR dropping from 120.8 to 6.9 as the spectrum becomes nosier.

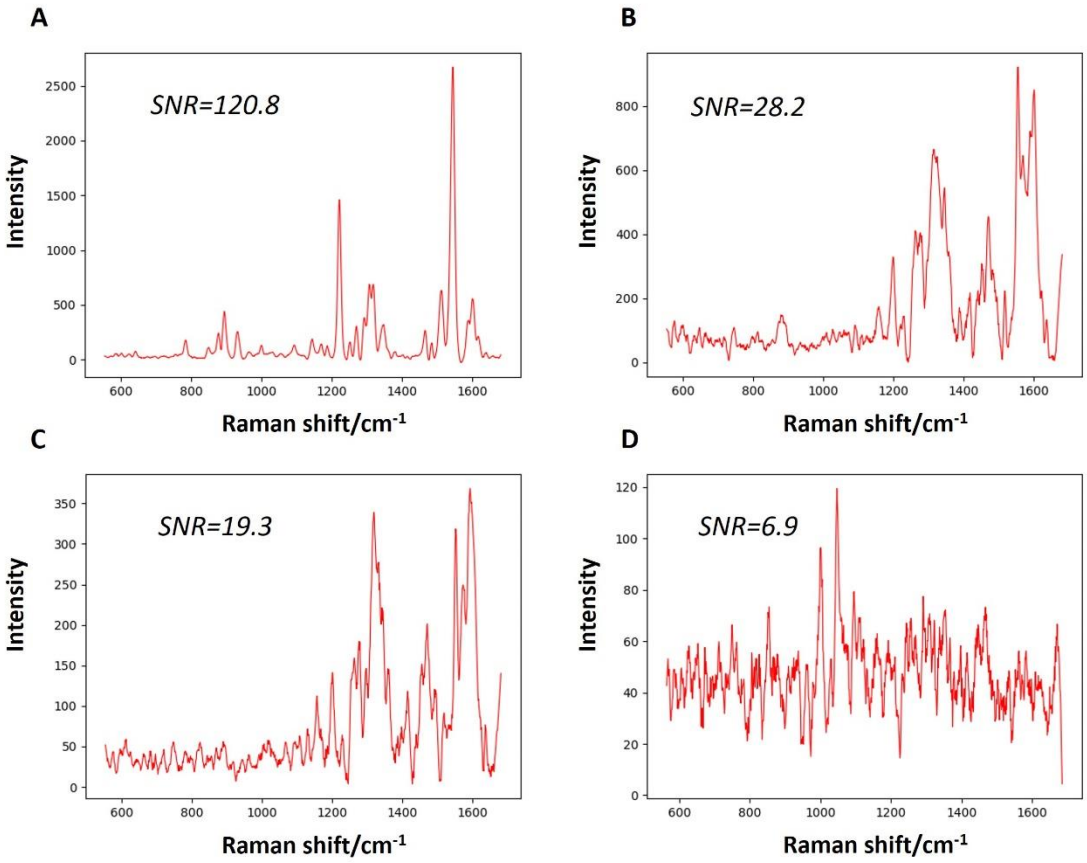


Figure 3.19 Spectra SNR. Spectra SNR decreases from 120.8 to 6.9, the quality is dropping accordingly.

3.6.2 Database

Considering the scale and attributes of the spectral database, we create dictionary-wise database to store our spectroscopic data. Each dictionary is created based on sample information

and type of study, which usually contains 10^2 to 10^5 spectra from NBPs. Keys and contents storing the sample information are given to every dictionary, including data matrix (matrix containing all the spectra, which are represented by one dimensional arrays), sample name, map index, label (numerical expression for each sample), group (high-level sample types), Raman shift (Raman shift range) and other auxiliary keys. Since we select a fixed Raman shift range to characterize the specimen, therefore the Raman shift information is the same across the samples and the spectral intensities are stored under a separate key-data matrix, which is intended to storage space as well as simplifying data analyses.

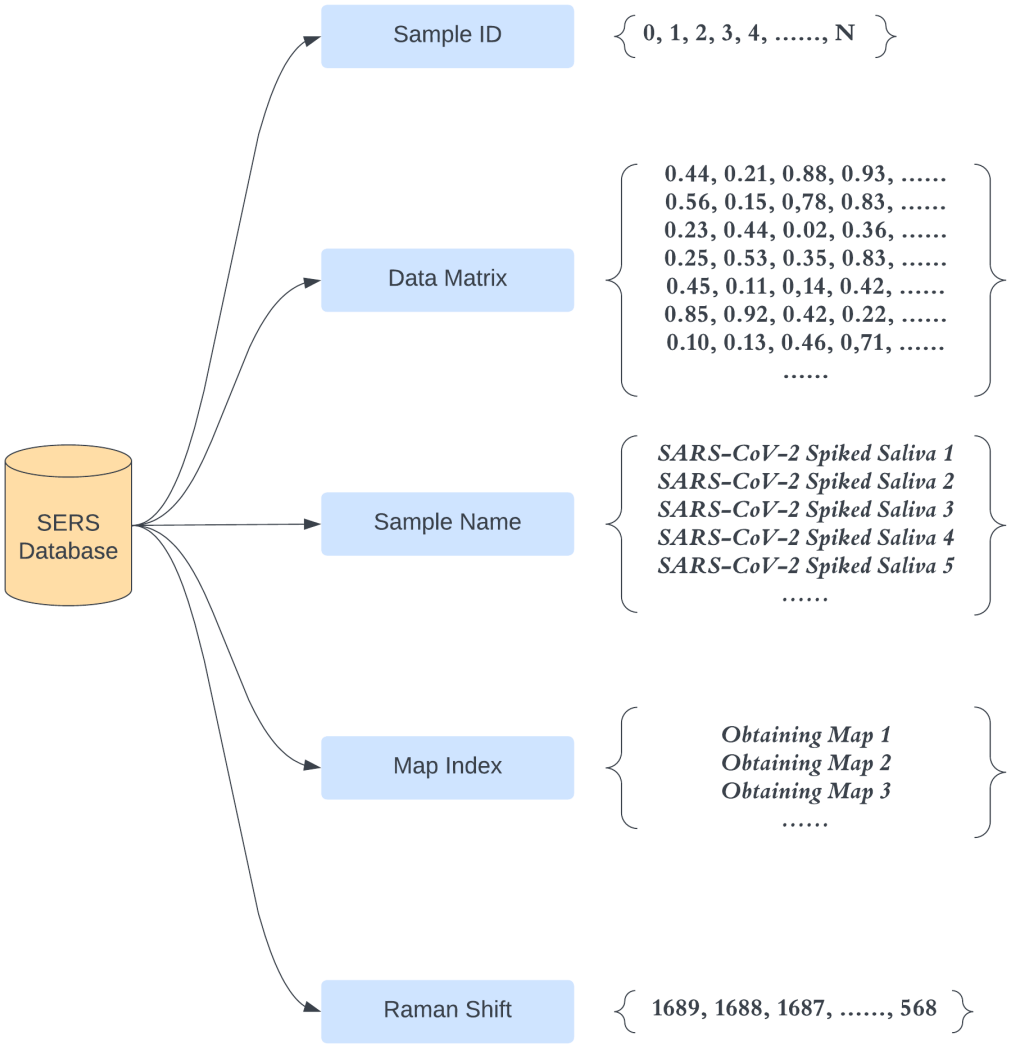


Figure 3.20 Diagram of SERS database structure.

With the demand for querying data in the future, SQL based data templates are being developed including all the samples as well as the test results. SQL is a well-known standard language for interacting with databases and conducting multiple operations, such as querying, inserting, updating, modifying database schemas, and managing user access. It will provide strong support for users to have access to the SERS fingerprints of NBPs and build data analysis models once established.

3.6.3 Artificial intelligence and machine learning

Artificial intelligence, especially machine learning, is believed to perform excellently on analyzing tremendous amounts of data and assist researchers for data mining and data science. With the rapid development of computing capabilities, such as the upgrading of AI graphic processing unit (GPU) by Apple Inc., Nvidia Inc and the invention of numerous analytical algorithms, AI based analytical methods are able to complete incredible complex and tremendous tasks efficiently. SERS spectra of NBPs are typically highly intricate due to a lot of different molecules contained within a single NBP. Unlike the small molecules such as amino acids, nucleic acid, NBPs' spectrum is usually composed of various combinations of peak positions, peaks widths, and peaks intensities. In addition, the uncertain fluorescence baselines and random noise make it extremely arduous to analyze and compile. Surface plasmon also adds a lot more variations to the regular Raman spectroscopy, therefore, a robust, powerful, and efficient analytical technique is required to carry on the analytical tasks. As introduced in the previous sections, we implemented multiple machine learning based analytical methods and algorithms, including dimensionality reduction analyses, supervised learning and unsupervised learning, feature extraction and selection, and performance evaluation metrics for a complete model designing, building, and testing platform.

Studies have shown that the introduction of machine learning could enhance the scale of scientific research. For example, machine learning models can classify spectroscopic data for biomarker discovery, disease diagnosis by identifying specific types of cells or tissues, and drug discovery and development. N. Banaei et al integrated microfluidic chip and two machine learning algorithms (K-nearest neighbor and classification tree) to build a SERS-based protein biomarker detection platform, which help them identify five protein biomarkers (CA19-9, HE4, MUC4, MMP7, and mesothelin) for recognizing disorder in cancer patients (Banaei et al., 2019). P.

Nguyen and B. Hong applied PCA and NN model to learn and predict the composition of 200-base long single-stranded DNA as biomarkers characterized by SERS spectral patterns (Nguyen et al., 2020). For evaluating the roles of machine learning models in biomarkers discovery, J. Li et al compared five general models including spectral decomposition, support vector regression, random forest regression, partial squares regression, and convolutional neural networks (CNN) to recognize the mixture components from a multiplexed mixture of seven SERS-active "nano-rattles" loaded with different dyes for mRNA biomarker detection (J. Q. Li et al., 2022). It turns out that CNN could successfully analyze SERS spectra from a singleplex, point-of-care assay that detects an mRNA biomarker for head and neck cancers. In addition to the application on spectral analyses, deep learning models greatly improve the interpretation of medical imaging, DNN have been widely used in CT-scan (Afshar et al., 2021; Al-Karawi et al., 2020; Kadry et al., 2020), X-rays imaging (Abed Mohammed et al., 2021; Chandra & Verma, 2020; Fusco et al., 2021; Rasheed et al., 2021), MRIs (Castillo T. et al., 2020; Eshaghi et al., 2021; Moradi et al., 2015) for disease diagnosis, treatment planning, and monitoring of diseases. As for the biological research field, spectra analyses or analyses of array-like data, such as mass spectrometry and sequencing, also provide significant information of biological properties. The involvement of machine learning analytical tools has been an important step forward (Hilario et al., 2003; Liebal et al., 2020; Petegrosso et al., 2020; Yang et al., 2020). Due to the nearly unlimited freedom of customizing models, machine learning can provide appropriate solutions to biological and medical problems.

3.6.3.1 Dimensionality reduction

Dimensionality reduction is a family of algorithms that are used in both machine learning and statistical analysis to reduce the original data high dimensionality space to a much smaller space while preserving the most informative features (Huang et al., 2019). It is employed to

overcome the “curse of dimensionality”, improve computational efficiency, remove redundant or irrelevant features, and visualize the original high-dimensional data (Poggio et al., 2017). Well-known algorithms include PCA, LDA, t-distributed stochastic neighbor embedding (TSNE), autoencoders, uniform manifold approximation and projection (UMAP) etc. Many algorithms are based on dimension linear transformations followed by projection or generating low-dimensional data distributions preserving the original high-dimensional distributions, which give succinct interpretations to the data from different perspectives. Dimensionality reduction can also be divided into supervised and unsupervised, depending on whether the data labels are involved. Supervised algorithms focus more on the classification by seeking transformed dimensions that provide the best separation of the data, while unsupervised dimensionality reductions try to group the data by either calculating the similarity, statistical variance etc. Research purposes are supposed to be clarified before choosing certain algorithms.

LDA is a supervised data analysis method first proposed by R. Fisher in differentiating flower types (Xanthopoulos et al., 2013). It determines a lower dimension space based on the original space that provides better separability of the data, which is computed based on the mean value and variance. As shown in Figure 3.21, a transformed dimension is computed to maximize the “distance” between two data groups. Massive data can be efficiently processed by LDA by simply solving a generalized eigenvalue problem and it works for both binary class and multi-class problems. Non-linear features can also be added by applying corresponding non-linear kernels.

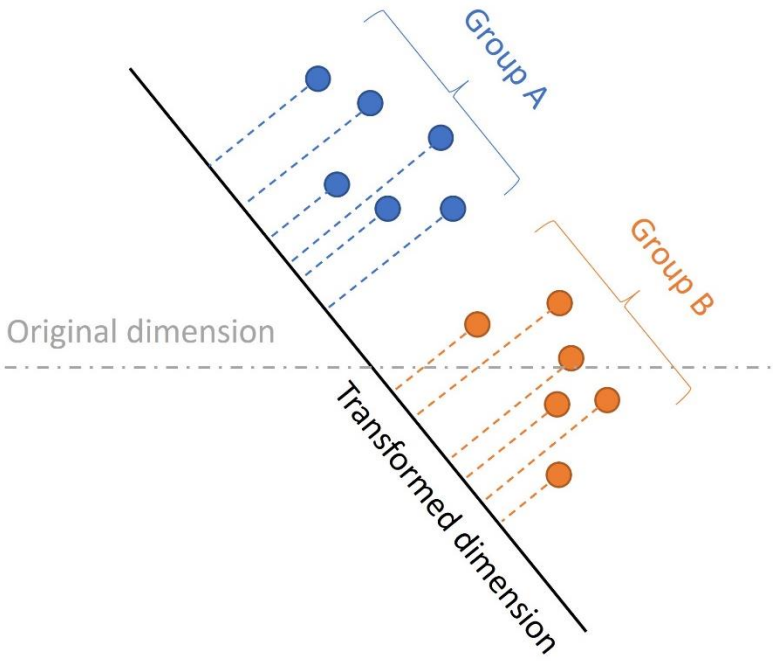


Figure 3.21 Simplified working principle of LDA by linear transformation.

For binary class problems, let $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, …, $\boldsymbol{x}_N$ be N data points belonging to two different classes A and B, the mean value for each class can be given by,

$$\overline{\boldsymbol{x}}_A = \frac{1}{N_A}\sum_{\boldsymbol{x}\in A}\boldsymbol{x}, \overline{\boldsymbol{x}}_B = \frac{1}{N_B}\sum_{\boldsymbol{x}\in B}\boldsymbol{x} \qquad 3\text{-}15$$

Where $\boldsymbol{N_A}$ and $\boldsymbol{N_B}$ are the data point number of group A and B, respectively. Similarly, the positive semidefinite variance matrix for each class can be written by

$$var(A) = \sum_{\boldsymbol{x}\in A}(\boldsymbol{x}-\overline{\boldsymbol{x}}_A)(\boldsymbol{x}-\overline{\boldsymbol{x}}_A)^T, var(B) = \sum_{\boldsymbol{x}\in B}(\boldsymbol{x}-\overline{\boldsymbol{x}}_B)(\boldsymbol{x}-\overline{\boldsymbol{x}}_B)^T \qquad 3\text{-}16$$

Which represents the variance within each class. On the other hand, the scatter matrix between the two groups is

$$\boldsymbol{S}_{AB} = (\overline{\boldsymbol{x}}_A-\overline{\boldsymbol{x}}_B)(\overline{\boldsymbol{x}}_A-\overline{\boldsymbol{x}}_B)^T \qquad 3\text{-}17$$

Standing for the distance of the means between the two groups. We are trying to find a transform dimension (or hyperplane) which maximizes the distance of the means between group A and B, at the same time minimizes the intra-class variance, therefore the loss function can be defined as

$$\max_{\Phi} \mathcal{L} = \max_{\Phi} \frac{\Phi^T \boldsymbol{S}_{AB} \Phi}{\Phi^T [\mathrm{var(A)} + \mathrm{var(B)}] \Phi} \tag{3-18}$$

Subject to

$$\Phi^T [\mathrm{var(A)} + \mathrm{var(B)}] \Phi = 1 \tag{3-19}$$

According to the Lagrangian multiplier method, the optimal Φ can be obtained by solving the eigenvalue and eigenvector problem.

$$\boldsymbol{S}_{AB} \Phi = \lambda [\mathrm{var(A)} + \mathrm{var(B)}] \Phi \tag{3-20}$$

Multi-class problem is an extension version of the binary class problem, the inter-class scatter matrix and the intra-class variance matrix.

$$\boldsymbol{S}_{intra} = \sum_{M} var(C_i) , \boldsymbol{S}_{inter} = \sum_{i=1}^{M} m_i (\overline{\boldsymbol{x}}_i - \overline{\boldsymbol{x}})(\overline{\boldsymbol{x}}_i - \overline{\boldsymbol{x}})^T \tag{3-21}$$

$$\overline{\boldsymbol{x}}_i = \frac{1}{m_i} \sum\nolimits_{\boldsymbol{x} \in i} \boldsymbol{x} , \overline{\boldsymbol{x}} = \frac{1}{N} \sum \boldsymbol{x} \tag{3-22}$$

Where *M* is the total number of classes, m_i denotes the data point number for each class. Then the optimal Φ is given by

$$\boldsymbol{S}_{inter} \Phi = \lambda \boldsymbol{S}_{intra} \Phi \tag{3-23}$$

In addition to dimensionality reduction, LDA can also be employed for classification tasks since it takes data labels into consideration for grouping the data points. Unknown samples can be assigned to preexist classes by projecting them onto the discriminant dimensions followed by

comparing their positions to the class boundaries. One important assumption underlying LDA is the multivariate Gaussian distribution of the data and equal class covariances. Violations of this assumption may require alternative methods like Quadratic Discriminant Analysis (QDA) or Regularized Discriminant Analysis (RDA). LDA also provides interpretability, as it assigns discriminant coefficients to each feature, indicating their relevance in the linear combination. This facilitates feature selection and understanding of the important factors contributing to the unique spectral features.

TSNE is another unsupervised dimensionality reduction algorithm frequently used in machine learning and data visualization. Unlike typical linear algorithms, TSNE tries to preserve the local structure of the data, rather than the global structure. It maps the original high-dimensional data into a much lower dimensional space (typically two dimensions) by maintaining the pairwise similarities between data points (Laurens & Hinton, 2008). TSNE constructs a similarity measurement between data points and optimizes a cost function to find an embedding that minimizes the Kullback-Leibler divergence (KL divergence) between the high-dimensional and low-dimensional representations (Hershey & Olsen, 2007).

Using the same notations as LDA, the similarity measurement between two data points x_i and x_j is given by the conditional probability p_{ij},

$$p_{ij} = \frac{\exp\left(-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2 / 2\sigma^2\right)}{\sum_{k \neq l} \exp\left(-\|\boldsymbol{x}_l - \boldsymbol{x}_k\|^2 / 2\sigma^2\right)} \qquad 3\text{-}24$$

In which p_{ij} will be relatively large for the data points near data points x_i, while for far data point p_{ij} will be negligible. σ is variance determined by optimizing the Shannon entropy [ref]. This is the core reason that TSNE focuses more on the local structure of the data instead of global. It then

creates low-dimensional data points to "represent" the original data points x_i. The counterparts of x_i and x_j, y_i and y_j, are also given a similarity measurement,

$$q_{ij} = \frac{\exp\left(-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2\right)}{\sum_{k \neq l} \exp\left(-\|\boldsymbol{x}_l - \boldsymbol{x}_k\|^2\right)} \quad \text{3-25}$$

In which the variance is set to $1/\sqrt{2}$ for simple calculation. It is worth noting that different variance for data points y_i only results in rescaling after dimensionality reduction. To minimize the mismatch between p_{ij} and q_{ij}, TSNE introduces KL divergences to measure the difference between the two distributions,

$$\mathcal{L} = \sum_i KL(P_i \parallel Q_i) = \sum_i \sum_j p_{i|j} log \frac{p_{i|j}}{q_{i|j}} \quad \text{3-26}$$

Where P_i and Q_i denote the sets of p_i and q_i for data point x_i in the original and mapped data sets, respectively. Optimizing the above loss function results in the low-dimensional form of original data set by minimizing the gradient of L,

$$\frac{\partial \mathcal{L}}{\partial y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j) \quad \text{3-27}$$

One potential problem of TSNE is the non-convex loss function, which leads to different results with different initialization states, several initializations are sometimes required to achieve the lowest KL divergence. TSNE is currently computationally expensive and is limited to fewer dimensional embeddings compared with other algorithms, moreover, it does not perverse the global structure well. Therefore, it is usually incorporated with other algorithms such as PCA to mitigate the pitfalls.

3.6.3.2 Classification or supervised learning

Classification is one of the fundamental tasks in machine learning which classifies the labelled data into separate predefined groups based on their features or characteristics. The goal of classification is to build a predictive model based on the training data (the input data) and generate accurate predictions to the unseen data. According to the number of classes or categories, classification can be divided into binary classification and multi-class classification, and the algorithms aim at learning decision boundaries that separate different classes within the feature space, the decision boundaries will be used later for unseen data predictions. Well-known classification algorithms include logistic regression, decision tree and its extensions (adaptive boosting algorithm, random forest algorithm etc.), support vector machine, naïve Bayes, and neural networks. Those models use different assumptions and strategies to learn from the data and have specific pros and cons. We have been using SVMs in our research based on the scale of dataset and universal performance (computational efficiency, accuracy etc.) compared with other models. The fundamental mathematics of SVMs is given below.

SVMs relies on the concepts of linear algebra and optimization then utilizes kernel tricks to increase its application on non-linear cases (Suthaharan, 2016). The goal of SVMs is to determine a hyperplane (plane in high-dimensional space) that separates the labelled data of different classes. It introduces the concept of margins to measure how well the hyperplane separates different classes, which are defined as the distance between the hyperplane and the nearest data points of different classes. For maximizing the margins, SVMs formulate the problem into a convex optimization problem solved by incorporating Lagrangian multipliers. SVMs classification is not limited to linear problems, for those data possessing non-linear structure, kernel tricks play significant roles for SVMs to learn their unique features.

Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ be N data points and $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N \in \{+1, -1\}$ are the corresponding labels. The hyperplane can be expressed by

$$\boldsymbol{\omega} \cdot \boldsymbol{x} + b = 0, \omega \in \mathcal{R}^N, b \in R \tag{3-28}$$

Corresponding to the decision function

$$f(\boldsymbol{x}) = sgn(\boldsymbol{\omega} \cdot \boldsymbol{x} + b) \tag{3-29}$$

Among all hyperplanes, there exists a specific one yielding the maximum margin that separating the classes, which is represented by

$$\max_{\boldsymbol{\omega},b}\{\min_{\boldsymbol{\omega},b}\|\boldsymbol{x} - \boldsymbol{x}_i\|\}, \boldsymbol{\omega} \cdot \boldsymbol{x} + b = 0, i = 1, 2, \ldots, N \tag{3-30}$$

As shown in Figure 3.22, the data points nearest to the hyperplane on different sides are the keys for determining the optimal hyperplane, which are named "support vectors". We can make the following adjustment to simplify the calculation,

$$\boldsymbol{\omega} \cdot \boldsymbol{x}_+ + b = +1, \boldsymbol{\omega} \cdot \boldsymbol{x}_- + b = -1 \tag{3-31}$$

Which only leads to rescaled coordinates. Therefore, the distance between the support vectors can be expressed by

$$d = \frac{2}{\|\boldsymbol{\omega}\|} \tag{3-32}$$

The original problem then becomes

$$\max_{\boldsymbol{\omega},b} \frac{2}{\|\boldsymbol{\omega}\|}, subject\ to\ y_i(\boldsymbol{\omega} \cdot \boldsymbol{x_i} + b) \geq 1, i = 1, 2, 3, \ldots, N \tag{3-33}$$

Or

$$\min_{\boldsymbol{\omega},b} \frac{1}{2}\|\boldsymbol{\omega}\|^2, subject\ to\ y_i(\boldsymbol{\omega} \cdot \boldsymbol{x_i} + b) \geq 1, i = 1, 2, 3, \ldots, N \tag{3-34}$$

Which becomes a convex optimization problem by the above modifications. A common way the solve this problem is through its Lagrangian dual

$$\max_{\boldsymbol{\lambda}} \min_{\boldsymbol{\omega},b} \mathcal{L}(\boldsymbol{\omega}, b, \boldsymbol{\lambda}) \tag{3-35}$$

Where

$$\mathcal{L}(\boldsymbol{\omega}, b, \boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{\omega}\|^2 - \sum_{i=1}^{N} \lambda_i [y_i(\boldsymbol{\omega} \cdot \boldsymbol{x_i} + b) - 1)] \tag{3-36}$$

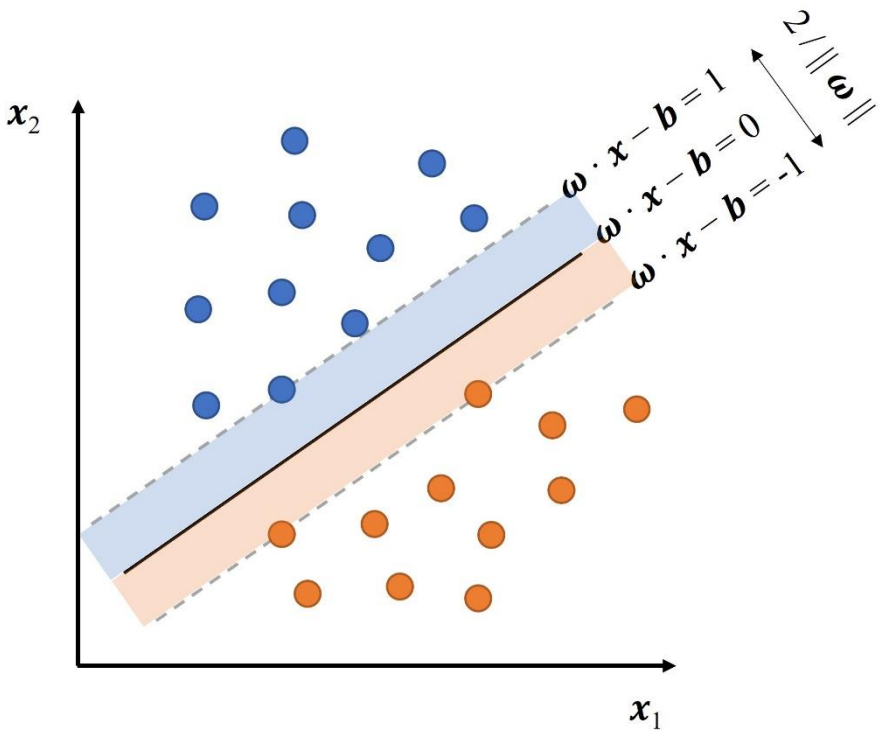


Figure 3.22 Demonstration of SVMs for classifying data points.

The primal problem and dual problem are closely related, and they usually have the same optimal solutions under certain conditions. Therefore, solving the dual problem can yield the optimal hyperplane for most cases.

To simplify the dual problem for easier computation, since $\boldsymbol{\mathcal{L}}(\boldsymbol{\omega}, b, \boldsymbol{\lambda})$ is convex, for any given $\boldsymbol{\lambda}$,

$$\frac{\partial}{\partial b}\mathcal{L}(\boldsymbol{\omega}, b, \boldsymbol{\lambda}) = 0, \frac{\partial}{\partial \boldsymbol{\omega}}\mathcal{L}(\boldsymbol{\omega}, b, \boldsymbol{\lambda}) = 0 \tag{3-37}$$

Render

$$\sum_{i=1}^{N} \lambda_i y_i = 0, \boldsymbol{\omega} = \sum_{i=1}^{N} \lambda_i y_i \boldsymbol{x_i} \tag{3-38}$$

By substituting Equation 3-38 to Equation 3-35 and 3-36, the dual proble1m can be written as

$$\max_{\lambda_i} \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{N} \lambda_i \lambda_j y_i y_j (\boldsymbol{x_i} \cdot \boldsymbol{x_j})$$
$$subject\ to\ \lambda_i \geq 0\ and \sum_{i=1}^{N} \lambda_i y_i = 0, i = 1, 2, 3, \ldots, N \tag{3-39}$$

And the hyperplane decision function is

$$f(\boldsymbol{x}) = sgn[\sum_{i=1}^{N} y_i \lambda_i \cdot (\boldsymbol{x} \cdot \boldsymbol{x_i}) + b] \tag{3-40}$$

Which is the typical problem in machine learning.

The involvement of kernel trick enables SVMs to effectively handle non-linearly separable data by implicitly mapping it to a higher-dimensional space where linear separation becomes possible. The kernel function calculates the similarity or distance between pairs of data points in

the original input space or the transformed feature space. It replaces the dot product between data points with a non-linear mapping, allowing SVMs to learn complex decision boundaries. Without kernel (or linear SVMs), we have

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x_i} \cdot \boldsymbol{x_j} \tag{3-41}$$

Multiple kernels can be used to map the original data into higher dimensional space, including linear kernel (original), polynomial kernel, Gaussian kernel (or Radial Basis Function, RBF), sigmoid kernel etc. Then we can rewrite the dual problem into a more general format,

$$\max_{\lambda_i} \sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{i,j=1}^{N} \lambda_i \lambda_j y_i y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j)$$
$$subject\ to\ \lambda_i \geq 0\ and \sum_{i=1}^{N} \lambda_i y_i = 0, i = 1, 2, 3, \ldots, N \tag{3-42}$$

And the corresponding decision function is

$$f(\boldsymbol{x}) = sgn[\sum_{i=1}^{N} y_i \lambda_i \cdot k(\boldsymbol{x}_i, \boldsymbol{x}_j) + b] \tag{3-43}$$

However, more real cases are that the hyperplane separating the classes does not exist due to the highly noisy data. Hereby, soft margin SVMs replaces hard margin SVMs by allowing a few instances violating $y_i(\omega \cdot x_i + b) \geq 1$, by introducing an extra term $\boldsymbol{\zeta}$, the constraints are relaxed to

$$y_i(\boldsymbol{\omega} \cdot \boldsymbol{x_i} + b) \geq 1 - \zeta_i, \zeta_i \geq 0, i = 1, 2, 3, \ldots, N \tag{3-44}$$

Therefore, the optimal classifier is established by balancing both $\boldsymbol{\omega}$ and the soft margin factor $\boldsymbol{\zeta}$, thus the objective function becomes

$$\frac{1}{2}\|\boldsymbol{\omega}\| + C\sum_{i=1}^{N}\zeta_i \qquad 3\text{-}45$$

In which C is the penalty hyperparameter controlling the margin and misclassification errors. Similarly, soft margin SVMs problem can be converted to dual problem and solved by quadratic optimization strategy (P. Chen et al., 2005).

In summary, SVMs offer the advantage of being effective in high-dimensional spaces, robust to overfitting, and versatile in handling both linear and non-linear classification tasks. SVMs perform over other algorithms on small and medium data set with high dimensionality. However, SVMs are sensitive to noise due to the concept of support vectors, computationally expensive for large datasets, and difficult to interpret (Suthaharan, 2016). For large and complicated datasets with non-linear features, even though multiple kernels are available to accommodate non-linear analyses, they are still based on certain assumptions on the data points distribution. Currently, neural networks and deep learning are believed to be better algorithms for large and complex data, however, their performance on SERS spectral data analyses hasn't been fully validated. Understanding these advantages and disadvantages helps determine the appropriate use and trade-offs of SVMs and other classification algorithms in machine learning applications.

3.6.3.3. Clustering or unsupervised learning

Unsupervised learning or clustering is another fundamental approach in machine learning, which groups the data points into clusters by calculating the inherent similarities between among data without prior knowledge of the data labels. The goal is essentially to discover the inherent structures, relationships, or natural groups within the data. Clustering algorithms can be grouped into four types, density-based, distribution-based, centroid-based, and hierarchical-based clustering. Each type uses unique metrics for similarity measurements and cluster determination.

Commonly used algorithms include K-means clustering (centroid-based), DBSCAN (Density-Based Spatial Clustering of Applications with Noise, density-based), Gaussian mixture models (GMM, distribution-based), HCA, hierarchical-based), BIRCH (balance iterative reducing and clustering using hierarchical, hierarchical-based), affinity propagation clustering, mean-shift clustering (hierarchical-based), agglomerative hierarchy clustering (hierarchical-based) etc. Specifically for NBPs' SERS data analysis, we need to consider the unique features of SERS spectra as well as the requirements. Some prerequisites or requirements are, (1) the total number of clusters is unknown; (2) customizable similarity metrics need to be available; (3) no knowledge of statistical distribution is available. Based on those concerns, we mainly implemented HCA in our study with customized "shifting Euclidean distance" as similarity metrics.

HCA utilizes an agglomerative or divisive method for clustering, shown in Figure 3.23. The former means that each instance starts from its own cluster and pairs of clusters are merged if they are "similar", while the latter means all instances start from one single cluster and splits recursively to form a hierarchy (Hubert, 2014; Köhn & Hubert, 2015). Upon the pairwise similarities obtained according to the predefined metrics, linkage criterion performs a critical role in determining pairs of clusters to be merged, or one cluster to be split. In our analysis of SERS spectra, "shifting Euclidean distance" is used as the similarity metrics, which is defined as,

$$dist(\boldsymbol{x}, \boldsymbol{y}) = \min_k \{\sum (x_{i\pm k} - y_{\mp k})^2\} \qquad 3\text{-}46$$

Where $\boldsymbol{x}$ and $\boldsymbol{y}$ represent two spectra, k is the shifting steps. The essential reason of applying shifting is due to the horizontal fluctuation on SERS spectrum due to the systematic error in calibration and random noise in photon scattering. The underlying idea is searching for the "optimal match" by shifting the pairwise spectra horizontally. As for the linkage criteria, there are

many options, such as complete linkage, single linkage, weighted average linkage, Ward linkage etc. Based on the assumptions that all the spectra included in a single cluster are supposed to share remarkable similarities, we used complete linkage (or maximum linkage), which is defined as,

$$d(X,Y) = \max_{\boldsymbol{x}\in X,\boldsymbol{y}\in Y} dist(\boldsymbol{x},\boldsymbol{y}) \qquad 3\text{-}47$$

In which $\boldsymbol{x}$ and $\boldsymbol{y}$ are two spectra, while X and Y are two clusters being compared. *dist* stands for the similarity (or distance) function. In this case, two clusters are joined into one only if all pairs of spectra have similarities below the predefined threshold, which guarantees that a single cluster exclusively corresponds to one spectral signature.

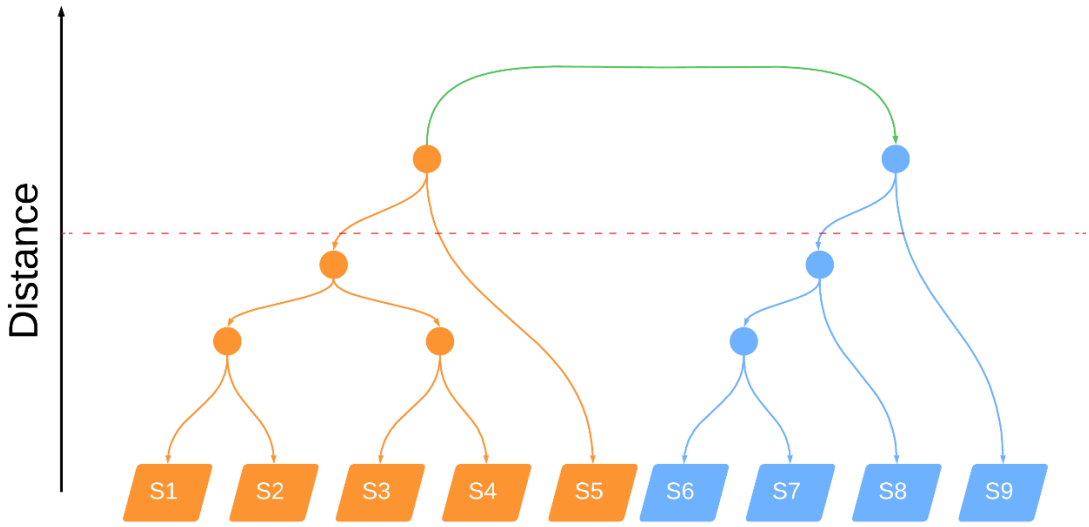


Figure 3.23 Demonstration of HCA algorithm for clustering analysis.

Overall, clustering analysis provides informative views of the inherent spectral features of NBPs, free from the knowledge of the labels. This method is very useful when multiple types of NBPs are present in one specimen, which helps us identify unique groups or discover new entities. In our study, clustering analysis serves as an auxiliary technique for classification by correcting

the training data labels and improving the prediction accuracy. However, it is still a challenge to determine the similarity metrics and the linkage, which can greatly affect the clustering results. HCA also assumes a hierarchical structure, which may not apply to all NBP dataset, we need to further increase the robustness of our SOP and give an objective interpretation of the dendrogram showing the hierarchical relations.

3.6.3.4 Feature selection and feature extraction

SERS spectral data typically have more than 10^3 features represented by the Raman shift range. There are informative features, irrelevant features, redundant features in the original feature space. Large feature (or dimensionality) space requires increasing computational resources for analysis and sometimes causes unintentional issues due to "curse of dimensionality". Those issues lead to lower efficiency in learning the spectral signatures of target NBPs as well as loss of the general and representative spectral features due to overfitting. To overcome these issues, researchers have put much effort into investigating powerful feature selection and feature extraction algorithms to avoid the above pitfalls. Both approaches aim to construct a reduced feature space that captures the essential characteristics of the data. Feature selection involves selecting a subset of the original features, while feature extraction transforms the original feature space into a more informative and compact representation, such as PCA, LDA, or autoencoders. Feature selection offers the advantage of interpretability and transparency and can be combined with other analytical tools for targeted and controlled analysis. Popular feature selection algorithms include recursive feature elimination (RFE), Chi-square, information gain, genetic algorithms, and more. Most algorithms select features based on their statistical properties such as correlation, variance, information gain etc., irrespective of the analysis goal. In contrast, genetic algorithms (GAs) search for the best feature subsets through generating many searching agents towards the

optimization of the analysis goal, for example, classification accuracy (Dorigo et al., 2006). GAs include many sub-algorithms different from the way of allocating searching agents and evaluating the “fitness”, which is defined as the performance of the solution domain. They offer an effective approach to tackle feature selection problems by mimicking the evolutionary process. The algorithm starts by generating an initial population of potential feature subsets, each represented as a chromosome. The fitness of each chromosome is evaluated based on a predefined evaluation metric, often classification accuracy or regression error. During each generation, a portion of the existing population is selected for the next generation. Individual solutions are selected through a fitness-based process, in which solutions with higher fitness values are typically more likely to be selected. Through repeated generations, GAs iteratively improve the fitness of the population, gradually converging towards an optimal or near-optimal feature subset. This process allows GAs to efficiently explore a large solution space and discover feature combinations that maximize the performance of the selected model. Typical algorithms such as ant colony optimization feature selection (ACOFS) and particle swarm optimization feature selection (PSOFS) are currently widely used for optimization problems, data mining, image and signal processing, neural networks architecture search and more (H. Peng et al., 2018; Sakri et al., 2018).

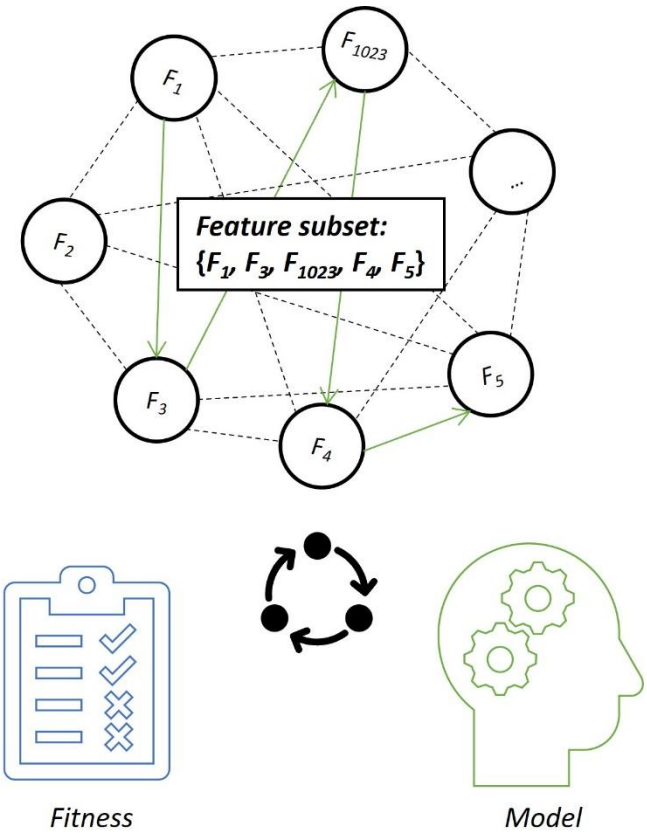


Figure 3.24 Diagram of ACOFS principle. Feature subsets are selected followed by evaluating fitness with incorporating predictive model, the optimization process keeps running until preset requirements are fulfilled.

ACOFS is a one of the GAs inspired by the foraging behavior of ants (H. Peng et al., 2018). While ACO is commonly used for solving optimization problems, it has also been applied to feature selection tasks. ACOFS starts with a population of artificial ants that traverse a search space, where each ant represents a potential feature subset. The goal is to find a feature subset that optimizes a specific objective function, such as classification accuracy or model complexity. The ants construct solutions by selecting features based on probabilistic decision rules and pheromone trails. Pheromone trails mimic the communication between ants by depositing and updating information about the quality of selected features. During the construction phase, ants evaluate the

quality of feature subsets and adjust the pheromone levels accordingly. The pheromone levels guide next generation ants to preferentially select features that have been previously regarded beneficial. As generations iteratively continue, it allows ants to explore the search space and converge towards the optimal feature subsets. The final feature subset is typically determined by selecting the best solution encountered during the iterations.

The detailed working principle of ACOFS is given in a series of steps below, and the pseudo-code is given in Figure 3.25.

Input: *X: M × N matrix, N-dimensional training data with M spectra.*
m (≤ N): number of features for the reduced feature space.
n_C_{max}: maximum of number of generations.
n_Ants: number of agents (ants) for each generation.
n_F: number of feature selected by each ants.
ρ: decay rate.
@hi: heuristic information function.
Output: *reduced feature space containing m features.*

Begin algorithm
Initialization
$\eta_i(t = 0) = @hi,\ i = 1, 2, \ldots, N.$ /* *pheromone* */
$\tau_i(t = 0) = c,\ i = 1, 2, \ldots, N.$ /* *heuristic information, c is a constant parameter* */
$S^l(t = 0) = \{f_1, f_2, \ldots, f_N\}.$ /* *local optimal feature subset*/
$S^g(t = 0) = \{f_1, f_2, \ldots, f_N\}.$ /* *global optimal feature subset* */
for $t = 1$ *to* n_C_{max} **do**
Place n_Ants on random features.
for $k = 1$ *to n_Ants* **do**
for $i = 1$ *to n_F* **do**
Choose the next unvisited feature according to the probability.
Move the k-th ant to the new selected feature f.
end for
$F^k(t) = \varphi \cdot Acc(S^k(t)) + (1 - \varphi) \cdot (1 - L(S^k(t)) / N).$ /* *Calculate the fitness for each ant* */
$S^l(t) = S^k(t;\ k = max\ F^k(t)).$ /* *Update local optimal feature subset*/
end for
$S^g(t) = S^l(t)$ **if** $F^l(t) \geq F^g(t)$ **else** $S^g(t) = S^g(t - 1)$. /* *Update global optimal feature subset*/
end for
Obtain the final reduced feature subspace S^g.
end algorithm

Figure 3.25 Pseudo-code of ACOFS.

Step 1: Initialization; assume N is the original feature space of dataset M including C classes. Initialize the pheromone τ with equal values and the heuristic information η with Fisher score for all N features. Preset the number of generations of feature selection t and the number of ants k for each generation.

$$\eta_i = \frac{\sum_{C_j}(\Omega_{i,C_j} - \Omega_{G_i})}{\sum_{C_j}\sigma_{i,C_j}^2} \tag{3-48}$$

In which $\Omega_{i,Cj}$ is the centroid for class C_j on feature i, and Ω_{Gj} is the global centroid on feature i, $\sigma_{i,Cj}$ is variance for class C_j on feature i

Step 2: Determine the feature subset size r according to the theory of sample-to-feature ratio.

Step 3: Generate k artificial ants for building feature subset.

Step 3: Using probabilistic transition rule to calculate the chance for each feature to be selected by each ant.

$$P_i^k(t) = \begin{cases} \dfrac{|\tau_i(t)|^\alpha |\eta_i(t)|^\beta}{\sum_{j \in f(i;k)} |\tau_j(t)|^\alpha |\eta_j(t)|^\beta}, & i \in f(i;k) \\ 0, & else \end{cases} \tag{3-49}$$

Which gives the probability for feature i to be selected during tth generation by the kth ant. $f(i;k)$ is the features that haven't been selected.

Step 4: Evaluate the fitness of subset $S^k(t)$ within each generation according to the fitness equation.

$$F = \varphi \cdot Acc[S^k(t)] + (1-\varphi)\left(1 - \frac{L(S^k(t))}{N}\right) \tag{3-50}$$

Where *Acc* denotes the performance metrics (classification accuracy, clustering separation etc.), L is the length of subset, φ is a weighing factor controlling the balance.

Step 5: Select the local best feature subset $S^l(t)$ from $S^k(t)$.

Step 6: Check if the termination conditions (predefined performance, number of generations etc.) are satisfied. If so, select the global optimal feature subset $S^g(t)$ from $S^l(t)$, otherwise continue the generations.

Step 7: Update the pheromone τ and the heuristic information η according to the rules of updating. The pheromone decays by ρ upon entering a new generation to increase the exploitability. Each feature is updated by the mean of fitness, given by $\Delta\tau^k$, the global optimal features are further rewarded by the last term $\Delta\tau^g$. e is the factor measuring rewarding degree for global optimal features.

$$\tau_i(t+1) = (1-\rho)\tau_i(t) + \sum_{k=1}^{N} \Delta\tau_i^k(t) + e\Delta\tau_i^g(t)$$

$$\Delta\tau_i^k(t) = \begin{cases} F_i^k/N, & i \in S^k(t) \\ 0, & else \end{cases} \qquad 3\text{-}51$$

$$\Delta\tau_i^g(t) = \begin{cases} F_i^g, & i \in S^g(t) \\ 0, & else \end{cases}$$

Step 8: Generate a new generation of ants and continue from step 3.

In conclusion, the basic idea behind ACOFS is selecting the optimal feature subset through guiding the ants to convergence with pheromone level. As a heuristic model, it has both the efficiency and exploitation to extract the features providing the most valuable information. Instead of conducting an extremely computationally expensive search on the original feature space, it is a hybrid search engine that combines the wrapper and filter approaches. ACOFS has been reported

to be able to figure out the optimal or near-optimal solutions and outperform other non-heuristic algorithms. S. Tabakhi et al compared the unsupervised ACOFS versus other eleven well-unknown univariate and multivariate feature selection methods (information gain, relevance-redundancy feature selection) using multiple classifiers (SVMs, decision tree, and Naïve Bayes) on different datasets (Tabakhi et al., 2014). Unsupervised ACOFS turned out to significantly outperform the other methods in terms of error rates and feature subset sizes and could be compatible with many classifiers (Kabir et al., 2012; Nayar et al., 2021). The parameter tuning for ACOFS is usually challenging and requires quite a bit of effort to optimize, and the computational complexity increases rapidly with the size of feature space growing. Researchers have combined multiple feature selection methods sequentially to save computational resources. Difference choices of heuristic information initialization affect the results as well, in addition to Fisher score, well-known metrics include information gain, gain ratio, symmetrical uncertainty, Gini index, and other filtering-based algorithms. ACOFS provides a rather flexible approach to investigating the SERS spectral features in depth, which helps a lot with interpreting biological properties of NBPs and discovering biomarkers. In the future, feature selection will be an inevitable step for obtaining bioinformation from SERS characterizations, more similar algorithms such as particle swarm algorithms, Tabu search, can be introduced for various data analysis tasks.

3.6.3.5 Implementation of machine learning methods

Machine learning methods play a pivotal role in our research, which focuses on the application of SERS for disease diagnosis. Specifically, in our study of early diagnosis for NSCLC, LDA dimensionality reduction is introduced to elucidate the spectral characteristics of exosome subgroups derived from HBEC. ACOFS is introduced to uncover the molecular information that contributes to distinguishing cancer-related exosome subgroups. Nearest Neighbors based spectral

matching algorithm serves as identifiers for extracting malignant signatures within clinical samples. Detailed explanations of these methods are provided in the subsequent chapters. Furthermore, in the investigation into COVID detection, SVMs serve as the primary predictive classifier for distinguishing SARS-CoV-2 viruses from irrelevant particles. HCA methods are employed to address labeling issues arising from the presence of diverse NBPs in virus specimens.

3.7 References

Abe, A., Inoue, K., Tanaka, T., Kato, J., Kajiyama, N., Kawaguchi, R., Tanaka, S., Yoshiba, M., & Kohara, M. (1999). Quantitation of Hepatitis B Virus Genomic DNA by Real-Time Detection PCR. Journal of Clinical Microbiology, 37(9), 2899–2903.

Abed Mohammed, M., Hameed Abdulkareem, K., Garcia-Zapirain, B., A. Mostafa, S., S. Maashi, M., S. Al-Waisy, A., Ahmed Subhi, M., Awad Mutlag, A., & Le, D.-N. (2021). A Comprehensive Investigation of Machine Learning Feature Extraction and Classification Methods for Automated Diagnosis of COVID-19 Based on X-Ray Images. Computers, Materials & Continua, 66(3), 3289–3310.

Afshar, P., Heidarian, S., Enshaei, N., Naderkhani, F., Rafiee, M. J., Oikonomou, A., Fard, F. B., Samimi, K., Plataniotis, K. N., & Mohammadi, A. (2021). COVID-CT-MD, COVID-19 computed tomography scan dataset applicable in machine learning and deep learning. Scientific Data, 8(1), 121.

Akib, T. B. A., Mou, S. F., Rahman, Md. M., Rana, Md. M., Islam, Md. R., Mehedi, I. M., Mahmud, M. A. P., & Kouzani, A. Z. (2021). Design and Numerical Analysis of a Graphene-Coated SPR Biosensor for Rapid Detection of the Novel Coronavirus. Sensors, 21(10), 3491.

Alenquer, M., & Amorim, M. (2015). Exosome Biogenesis, Regulation, and Function in Viral Infection. Viruses, 7(9), 5066–5083.

Al-Karawi, D., Al-Zaidi, S., Polus, N., & Jassim, S. (2020). Machine Learning Analysis of Chest CT Scan Images as a Complementary Digital Test of Coronavirus (COVID-19) Patients [Preprint]. Radiology and Imaging.

Bai, C., Zhong, Q., & Gao, G. F. (2022). Overview of SARS-CoV-2 genome-encoded proteins. Science China Life Sciences, 65(2), 280–294.

Banaei, N., Moshfegh, J., Mohseni-Kabir, A., Houghton, J. M., Sun, Y., & Kim, B. (2019). Machine learning algorithms enhance the specificity of cancer biomarker detection using SERS-based immunoassays in microfluidic chips. RSC Advances, 9(4), 1859–1868.

Blackwell, R., Foreman, K., & Gupta, G. (2017). The Role of Cancer-Derived Exosomes in Tumorigenicity & Epithelial-to-Mesenchymal Transition. Cancers, 9(12), 105.

Bustin, S. A., Benes, V., Nolan, T., & Pfaffl, M. W. (2005). Quantitative real-time RT-PCR – a perspective. Journal of Molecular Endocrinology, 34(3), 597–601.

Cai, J., Gong, L., Li, G., Guo, J., Yi, X., & Wang, Z. (2021). Exosomes in ovarian cancer ascites promote epithelial–mesenchymal transition of ovarian cancer cells by delivery of miR-6780b-5p. Cell Death & Disease, 12(2), 210.

Camussi, G., Deregibus, M. C., Bruno, S., Cantaluppi, V., & Biancone, L. (2010). Exosomes/microvesicles as a mechanism of cell-to-cell communication. Kidney International, 78(9), 838–848.

Carlomagno, C., Bertazioli, D., Gualerzi, A., Picciolini, S., Banfi, P. I., Lax, A., Messina, E., Navarro, J., Bianchi, L., Caronni, A., Marenco, F., Monteleone, S., Arienti, C., & Bedoni, M. (2021). COVID-19 salivary Raman fingerprint: Innovative approach for the detection of current and past SARS-CoV-2 infections. Scientific Reports, 11(1), 4943.

Carretero-González, A., Otero, I., Carril-Ajuria, L., De Velasco, G., & Manso, L. (2018). Exosomes: Definition, Role in Tumor Development and Clinical Implications. Cancer Microenvironment, 11(1), 13–21.

Castillo T., J. M., Arif, M., Niessen, W. J., Schoots, I. G., & Veenland, J. F. (2020). Automated Classification of Significant Prostate Cancer on MRI: A Systematic Review on the Performance of Machine Learning Applications. Cancers, 12(6), 1606.

Chandra, T. B., & Verma, K. (2020). Pneumonia Detection on Chest X-Ray Using Machine Learning Paradigm. In B. B. Chaudhuri, M. Nakagawa, P. Khanna, & S. Kumar (Eds.), Proceedings of 3rd International Conference on Computer Vision and Image Processing (Vol. 1022, pp. 21–33). Springer Singapore.

Chen, P., Lin, C., & Schölkopf, B. (2005). A tutorial on ν-support vector machines. Applied Stochastic Models in Business and Industry, 21(2), 111–136.

Christianson, H. C., Svensson, K. J., Van Kuppevelt, T. H., Li, J.-P., & Belting, M. (2013). Cancer cell exosomes depend on cell-surface heparan sulfate proteoglycans for their internalization and functional activity. Proceedings of the National Academy of Sciences, 110(43), 17380–17385.

Chu, H. Y., Englund, J. A., Huang, D., Scott, E., Chan, J. D., Jain, R., Pottinger, P. S., Lynch, J. B., Dellit, T. H., Jerome, K. R., & Kuypers, J. (2015). Impact of rapid influenza PCR testing on

hospitalization and antiviral use: A retrospective cohort study: Impact of Rapid Flu PCR Testing on Clinical Outcomes. Journal of Medical Virology, 87(12), 2021–2026.

Cialla, D., März, A., Böhme, R., Theil, F., Weber, K., Schmitt, M., & Popp, J. (2012). Surface-enhanced Raman spectroscopy (SERS): Progress and trends. Analytical and Bioanalytical Chemistry, 403(1), 27–54.

Colombo, M., Moita, C., Van Niel, G., Kowal, J., Vigneron, J., Benaroch, P., Manel, N., Moita, L. F., Théry, C., & Raposo, G. (2013). Analysis of ESCRT functions in exosome biogenesis, composition and secretion highlights the heterogeneity of extracellular vesicles. Journal of Cell Science, jcs.128868.

Cosar, B., Karagulleoglu, Z. Y., Unal, S., Ince, A. T., Uncuoglu, D. B., Tuncer, G., Kilinc, B. R., Ozkan, Y. E., Ozkoc, H. C., Demir, I. N., Eker, A., Karagoz, F., Simsek, S. Y., Yasar, B., Pala, M., Demir, A., Atak, I. N., Mendi, A. H., Bengi, V. U., … Demir-Dora, D. (2022). SARS-CoV-2 Mutations and their Viral Variants. Cytokine & Growth Factor Reviews, 63, 10–22.

Das, S., Ansel, K. M., Bitzer, M., Breakefield, X. O., Charest, A., Galas, D. J., Gerstein, M. B., Gupta, M., Milosavljevic, A., McManus, M. T., Patel, T., Raffai, R. L., Rozowsky, J., Roth, M. E., Saugstad, J. A., Van Keuren-Jensen, K., Weaver, A. M., Laurent, L. C., Abdel-Mageed, A. B., … Zhang, H.-G. (2019). The Extracellular RNA Communication Consortium: Establishing Foundational Knowledge and Technologies for Extracellular RNA Research. Cell, 177(2), 231–242.

Dinnes, J., Sharma, P., Berhane, S., Van Wyk, S. S., Nyaaba, N., Domen, J., Taylor, M., Cunningham, J., Davenport, C., Dittrich, S., Emperador, D., Hooft, L., Leeflang, M. M., McInnes, M. D., Spijker, R., Verbakel, J. Y., Takwoingi, Y., Taylor-Phillips, S., Van Den Bruel, A., …

Cochrane COVID-19 Diagnostic Test Accuracy Group. (2022). Rapid, point-of-care antigen tests for diagnosis of SARS-CoV-2 infection. Cochrane Database of Systematic Reviews, 2022(7).

Dorigo, M., Birattari, M., & Stutzle, T. (2006). Ant colony optimization. IEEE Computational Intelligence Magazine, 1(4), 28–39.

Eshaghi, A., Young, A. L., Wijeratne, P. A., Prados, F., Arnold, D. L., Narayanan, S., Guttmann, C. R. G., Barkhof, F., Alexander, D. C., Thompson, A. J., Chard, D., & Ciccarelli, O. (2021). Identifying multiple sclerosis subtypes using unsupervised machine learning and MRI data. Nature Communications, 12(1), 2078.

Etchegoin, P. G., & Le Ru, E. C. (2010). Basic Electromagnetic Theory of SERS. In S. Schlücker (Ed.), Surface Enhanced Raman Spectroscopy (1st ed., pp. 1–37). Wiley.

Fang, S., Tian, H., Li, X., Jin, D., Li, X., Kong, J., Yang, C., Yang, X., Lu, Y., Luo, Y., Lin, B., Niu, W., & Liu, T. (2017). Clinical application of a microfluidic chip for immunocapture and quantification of circulating exosomes to assist breast cancer diagnosis and molecular classification. PLOS ONE, 12(4), e0175050.

Femminò, S., Penna, C., Margarita, S., Comità, S., Brizzi, M. F., & Pagliaro, P. (2020). Extracellular vesicles and cardiovascular system: Biomarkers and Cardioprotective Effectors. Vascular Pharmacology, 135, 106790.

Fesenko, O., Dovbeshko, G., Dementjev, A., Karpicz, R., Kaplas, T., & Svirko, Y. (2015). Graphene-enhanced Raman spectroscopy of thymine adsorbed on single-layer graphene. Nanoscale Research Letters, 10(1), 163.

Fleischmann, M., Hendra, P. J., & McQuillan, A. J. (1973). Raman spectra from electrode surfaces. Journal of the Chemical Society, Chemical Communications, 3, 80.

Fleming, A., Sampey, G., Chung, M.-C., Bailey, C., Van Hoek, M. L., Kashanchi, F., & Hakami, R. M. (2014). The carrying pigeons of the cell: Exosomes and their role in infectious diseases caused by human pathogens. Pathogens and Disease, 71(2), 109–120.

Francisco-Cruz, A., Parra, E. R., Tetzlaff, M. T., & Wistuba, I. I. (2020). Multiplex Immunofluorescence Assays. In M. Thurin, A. Cesano, & F. M. Marincola (Eds.), Biomarkers for Immunotherapy of Cancer (Vol. 2055, pp. 467–495). Springer New York.

Fu, H., Yang, H., Zhang, X., Wang, B., Mao, J., Li, X., Wang, M., Zhang, B., Sun, Z., Qian, H., & Xu, W. (2018). Exosomal TRIM3 is a novel marker and therapy target for gastric cancer. Journal of Experimental & Clinical Cancer Research, 37(1), 162.

Fu, M., Gu, J., Jiang, P., Qian, H., Xu, W., & Zhang, X. (2019). Exosomes in gastric cancer: Roles, mechanisms, and applications. Molecular Cancer, 18(1), 41.

Fusco, R., Grassi, R., Granata, V., Setola, S. V., Grassi, F., Cozzi, D., Pecori, B., Izzo, F., & Petrillo, A. (2021). Artificial Intelligence and COVID-19 Using Chest CT Scan and Chest X-ray Images: Machine Learning and Deep Learning Approaches for Diagnosis and Treatment. Journal of Personalized Medicine, 11(10), 993.

Guo, B. B., Bellingham, S. A., & Hill, A. F. (2015). The Neutral Sphingomyelinase Pathway Regulates Packaging of the Prion Protein into Exosomes. Journal of Biological Chemistry, 290(6), 3455–3467.

Han, X. X., Rodriguez, R. S., Haynes, C. L., Ozaki, Y., & Zhao, B. (2022). Surface-enhanced Raman spectroscopy. Nature Reviews Methods Primers, 1(1), 87.

Han, Y., Jia, L., Zheng, Y., & Li, W. (2018). Salivary Exosomes: Emerging Roles in Systemic Disease. International Journal of Biological Sciences, 14(6), 633–643.

Hao, E., Schatz, G. C., & Hupp, J. T. (2004). Synthesis and Optical Properties of Anisotropic Metal Nanoparticles. Journal of Fluorescence, 14(4), 331–341.

Haritha, D., Pranathi, M. K., & Reethika, M. (2020). COVID Detection from Chest X-rays with DeepLearning: CheXNet. 2020 5th International Conference on Computing, Communication and Security (ICCCS), 1–5.

Hasöksüz, M., Kiliç, S., & Saraç, F. (2020). Coronaviruses and SARS-COV-2. TURKISH JOURNAL OF MEDICAL SCIENCES, 50(SI-1), 549–556.

Hershey, J. R., & Olsen, P. A. (2007). Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP ’07, IV-317-IV–320.

Hilario, M., Kalousis, A., Müller, M., & Pellegrini, C. (2003). Machine learning approaches to lung cancer prediction from mass spectra. PROTEOMICS, 3(9), 1716–1719.

Howitt, J., & Hill, A. F. (2016a). Exosomes in the Pathology of Neurodegenerative Diseases. Journal of Biological Chemistry, 291(52), 26589–26597.

Howitt, J., & Hill, A. F. (2016b). Exosomes in the Pathology of Neurodegenerative Diseases. Journal of Biological Chemistry, 291(52), 26589–26597.

Hsu, C., Morohashi, Y., Yoshimura, S., Manrique-Hoyos, N., Jung, S., Lauterbach, M. A., Bakhti, M., Grønborg, M., Möbius, W., Rhee, J., Barr, F. A., & Simons, M. (2010). Regulation of exosome secretion by Rab35 and its GTPase-activating proteins TBC1D10A–C. Journal of Cell Biology, 189(2), 223–232.

Hu, A., Colella, M., Tam, J. S., Rappaport, R., & Cheng, S.-M. (2003). Simultaneous Detection, Subgrouping, and Quantitation of Respiratory Syncytial Virus A and B by Real-Time PCR. Journal of Clinical Microbiology, 41(1), 149–154.

Hu, Y., Rao, S.-S., Wang, Z.-X., Cao, J., Tan, Y.-J., Luo, J., Li, H.-M., Zhang, W.-S., Chen, C.-Y., & Xie, H. (2018). Exosomes from human umbilical cord blood accelerate cutaneous wound healing through miR-21-3p-mediated promotion of angiogenesis and fibroblast function. Theranostics, 8(1), 169–184.

Huang, X., Wu, L., & Ye, Y. (2019). A Review on Dimensionality Reduction Techniques. International Journal of Pattern Recognition and Artificial Intelligence, 33(10), 1950017.

Hubert, L. J. (2014). Hierarchical Cluster Analysis. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, & J. L. Teugels (Eds.), Wiley StatsRef: Statistics Reference Online (1st ed.). Wiley.

Iha, K., Tsurusawa, N., Tsai, H.-Y., Lin, M.-W., Sonoda, H., Watabe, S., Yoshimura, T., & Ito, E. (2022). Ultrasensitive ELISA detection of proteins in separated lumen and membrane fractions of cancer cell exosomes. Analytical Biochemistry, 654, 114831.

Jadhav, S. A., Biji, P., Panthalingal, M. K., Murali Krishna, C., Rajkumar, S., Joshi, D. S., & Sundaram, N. (2021). Development of integrated microfluidic platform coupled with Surface-enhanced Raman Spectroscopy for diagnosis of COVID-19. Medical Hypotheses, 146, 110356.

Jamalipour Soufi, G., Iravani, S., & Varma, R. S. (2021). Molecularly imprinted polymers for the detection of viruses: Challenges and opportunities. The Analyst, 146(10), 3087–3100.

Jan, A. T., Rahman, S., Badierah, R., Lee, E. J., Mattar, E. H., Redwan, E. M., & Choi, I. (2021). Expedition into Exosome Biology: A Perspective of Progress from Discovery to Therapeutic Development. Cancers, 13(5), 1157.

Jiang, W., Ma, P., Deng, L., Liu, Z., Wang, X., Liu, X., & Long, G. (2020). Hepatitis A virus structural protein pX interacts with ALIX and promotes the secretion of virions and foreign proteins through exosome-like vesicles. Journal of Extracellular Vesicles, 9(1), 1716513.

Johnson, J. (2000). Structures of virus and virus-like particles. Current Opinion in Structural Biology, 10(2), 229–235.

Kabir, Md. M., Shahjahan, Md., & Murase, K. (2012). A new hybrid ant colony optimization algorithm for feature selection. Expert Systems with Applications, 39(3), 3747–3763.

Kadry, S., Rajinikanth, V., Rho, S., Raja, N. S. M., Rao, V. S., & Thanaraj, K. P. (2020). Development of a Machine-Learning System to Classify Lung CT Scan Images into Normal/COVID-19 Class.

Kalluri, R., & LeBleu, V. S. (2020). The biology, function, and biomedical applications of exosomes. Science, 367(6478), eaau6977.

Kasetsirikul, S., Umer, M., Soda, N., Sreejith, K. R., Shiddiky, M. J. A., & Nguyen, N.-T. (2020). Detection of the SARS-CoV-2 humanized antibody with paper-based ELISA. The Analyst, 145(23), 7680–7686.

Keerthikumar, S., Chisanga, D., Ariyaratne, D., Al Saffar, H., Anand, S., Zhao, K., Samuel, M., Pathan, M., Jois, M., Chilamkurti, N., Gangoda, L., & Mathivanan, S. (2016). ExoCarta: A Web-Based Compendium of Exosomal Cargo. Journal of Molecular Biology, 428(4), 688–692.

Kevadiya, B. D., Machhi, J., Herskovitz, J., Oleynikov, M. D., Blomberg, W. R., Bajwa, N., Soni, D., Das, S., Hasan, M., Patel, M., Senan, A. M., Gorantla, S., McMillan, J., Edagwa, B., Eisenberg, R., Gurumurthy, C. B., Reid, S. P. M., Punyadeera, C., Chang, L., & Gendelman, H. E. (2021). Diagnostics for SARS-CoV-2 infections. Nature Materials, 20(5), 593–605.

Kim, J.-H., Kim, E., & Lee, M. Y. (2018). Exosomes as diagnostic biomarkers in cancer. Molecular & Cellular Toxicology, 14(2), 113–122.

Kitane, D. L., Loukman, S., Marchoudi, N., Fernandez-Galiana, A., El Ansari, F. Z., Jouali, F., Badir, J., Gala, J.-L., Bertsimas, D., Azami, N., Lakbita, O., Moudam, O., Benhida, R., & Fekkak, J. (2021). A simple and fast spectroscopy-based technique for Covid-19 diagnosis. Scientific Reports, 11(1), 16740.

Kneipp, J., Kneipp, H., & Kneipp, K. (2008). SERS—a single-molecule and nanoscale tool for bioanalytics. Chemical Society Reviews, 37(5), 1052.

Kneipp, K., Wang, Y., Kneipp, H., Perelman, L. T., Itzkan, I., Dasari, R. R., & Feld, M. S. (1997). Single Molecule Detection Using Surface-Enhanced Raman Scattering (SERS). Physical Review Letters, 78(9), 1667–1670.

Köhn, H., & Hubert, L. J. (2015). Hierarchical Cluster Analysis. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, & J. L. Teugels (Eds.), Wiley StatsRef: Statistics Reference Online (1st ed., pp. 1–13). Wiley.

Koritzinsky, E. H., Street, J. M., Chari, R. R., Glispie, D. M., Bellomo, T. R., Aponte, A. M., Star, R. A., & Yuen, P. S. T. (2019). Circadian variation in the release of small extracellular vesicles can be normalized by vesicle number or TSG101. American Journal of Physiology-Renal Physiology, 317(5), F1098–F1110.

Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. Cold Spring Harbor Protocols, 2015(11), pdb.top084970.

Lakshmi, S., Hughes, T. A., & Priya, S. (2021). Exosomes and exosomal RNAs in breast cancer: A status update. European Journal of Cancer, 144, 252–268.

Laurens, V. der M., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(11).

Le Ru, E. C., Meyer, M., Blackie, E., & Etchegoin, P. G. (2008). Advanced aspects of electromagnetic SERS enhancement factors at a hot spot. Journal of Raman Spectroscopy, 39(9), 1127–1134.

Li, F., Yoshizawa, J. M., Kim, K.-M., Kanjanapangka, J., Grogan, T. R., Wang, X., Elashoff, D. E., Ishikawa, S., Chia, D., Liao, W., Akin, D., Yan, X., Lee, M.-S., Choi, R., Kim, S.-M., Kang, S.-Y., Bae, J.-M., Sohn, T.-S., Lee, J.-H., … Wong, D. T. W. (2018). Discovery and Validation of Salivary Extracellular RNA Biomarkers for Noninvasive Detection of Gastric Cancer. Clinical Chemistry, 64(10), 1513–1521.

Li, J. Q., Dukes, P. V., Lee, W., Sarkis, M., & Vo-Dinh, T. (2022). Machine learning using convolutional neural networks for SERS analysis of biomarkers in medical diagnostics. Journal of Raman Spectroscopy, 53(12), 2044–2057.

Li, K., Chen, Y., Li, A., Tan, C., & Liu, X. (2019). Exosomes play roles in sequential processes of tumor metastasis. International Journal of Cancer, 144(7), 1486–1495.

Li, P., Kaslan, M., Lee, S. H., Yao, J., & Gao, Z. (2017). Progress in Exosome Isolation Techniques. Theranostics, 7(3), 789–804.

Li, W., Li, C., Zhou, T., Liu, X., Liu, X., Li, X., & Chen, D. (2017). Role of exosomal proteins in cancer diagnosis. Molecular Cancer, 16(1), 145.

Liebal, U. W., Phan, A. N. T., Sudhakar, M., Raman, K., & Blank, L. M. (2020). Machine Learning Applications for Mass Spectrometry-Based Metabolomics. Metabolites, 10(6), 243.

Lindenbergh, M. F. S., Wubbolts, R., Borg, E. G. F., Van 'T Veld, E. M., Boes, M., & Stoorvogel, W. (2020). Dendritic cells release exosomes together with phagocytosed pathogen; potential implications for the role of exosomes in antigen presentation. Journal of Extracellular Vesicles, 9(1), 1798606.

Liu, Y., Gu, Y., Han, Y., Zhang, Q., Jiang, Z., Zhang, X., Huang, B., Xu, X., Zheng, J., & Cao, X. (2016). Tumor Exosomal RNAs Promote Lung Pre-metastatic Niche Formation by Activating Alveolar Epithelial TLR3 to Recruit Neutrophils. Cancer Cell, 30(2), 243–256.

Lukose, J., Chidangil, S., & George, S. D. (2021). Optical technologies for the detection of viruses like COVID-19: Progress and prospects. Biosensors and Bioelectronics, 178, 113004.

Magazine, N., Zhang, T., Wu, Y., McGee, M. C., Veggiani, G., & Huang, W. (2022). Mutations and Evolution of the SARS-CoV-2 Spike Protein. Viruses, 14(3), 640.

Malla, B., Aebersold, D. M., & Dal Pra, A. (2018). Protocol for serum exosomal miRNAs analysis in prostate cancer patients treated with radiotherapy. Journal of Translational Medicine, 16(1), 223.

Mathews, S. M. (2019). Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review. In K. Arai, R. Bhatia, & S. Kapoor (Eds.), Intelligent Computing (Vol. 998, pp. 1269–1292). Springer International Publishing.

Matsumura, T., Sugimachi, K., Iinuma, H., Takahashi, Y., Kurashige, J., Sawada, G., Ueda, M., Uchi, R., Ueo, H., Takano, Y., Shinden, Y., Eguchi, H., Yamamoto, H., Doki, Y., Mori, M., Ochiya, T., & Mimori, K. (2015). Exosomal microRNA in serum is a novel biomarker of recurrence in human colorectal cancer. British Journal of Cancer, 113(2), 275–281.

Melo, S. A., Sugimoto, H., O'Connell, J. T., Kato, N., Villanueva, A., Vidal, A., Qiu, L., Vitkin, E., Perelman, L. T., Melo, C. A., Lucci, A., Ivan, C., Calin, G. A., & Kalluri, R. (2014). Cancer Exosomes Perform Cell-Independent MicroRNA Biogenesis and Promote Tumorigenesis. Cancer Cell, 26(5), 707–721.

Moradi, E., Pepe, A., Gaser, C., Huttunen, H., & Tohka, J. (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. NeuroImage, 104, 398–412.

Morton, S. M., & Jensen, L. (2009). Understanding the Molecule−Surface Chemical Coupling in SERS. Journal of the American Chemical Society, 131(11), 4090–4098.

Mulvaney, S. P., & Keating, C. D. (2000). Raman Spectroscopy. Analytical Chemistry, 72(12), 145–158.

Muto, S., & Shiga, M. (2020). Application of machine learning techniques to electron microscopic/spectroscopic image data analysis. Microscopy, 69(2), 110–122.

Nayar, N., Gautam, S., Singh, P., & Mehta, G. (2021). Ant Colony Optimization: A Review of Literature and Application in Feature Selection. In S. Smys, V. E. Balas, K. A. Kamel, & P. Lafata (Eds.), Inventive Computation and Information Technologies (Vol. 173, pp. 285–297). Springer Singapore.

Newey, W. K., & Powell, J. L. (1987). Asymmetric Least Squares Estimation and Testing. Econometrica, 55(4), 819.

Nguyen, P. H. L., Hong, B., Rubin, S., & Fainman, Y. (2020). Machine learning for composition analysis of ssDNA using chemical enhancement in SERS. Biomedical Optics Express, 11(9), 5092.

Ni, Y., Zhang, W., Mu, G., Gu, Y., Wang, H., Wei, K., Xia, Y., Xie, X., Ge, Q., Tan, T., & Wang, J. (2023). Extracellular RNA profiles in non-small cell lung cancer plasma. Journal of Thoracic Disease, 15(5), 2742–2753.

Nilsson, J., Skog, J., Nordstrand, A., Baranov, V., Mincheva-Nilsson, L., Breakefield, X. O., & Widmark, A. (2009). Prostate cancer-derived urine exosomes: A novel approach to biomarkers for prostate cancer. British Journal of Cancer, 100(10), 1603–1607.

Olver, C., & Vidal, M. (2007). Proteomic Analysis of Secreted Exosomes. In E. Bertrand & M. Faupel (Eds.), Subcellular Proteomics (Vol. 43, pp. 99–131). Springer Netherlands.

Parker, J., Fowler, N., Walmsley, M. L., Schmidt, T., Scharrer, J., Kowaleski, J., Grimes, T., Hoyos, S., & Chen, J. (2015). Analytical Sensitivity Comparison between Singleplex Real-Time PCR and a Multiplex PCR Platform for Detecting Respiratory Viruses. PLOS ONE, 10(11), e0143164.

Patel, G. K., Khan, M. A., Zubair, H., Srivastava, S. K., Khushman, M., Singh, S., & Singh, A. P. (2019). Comparative analysis of exosome isolation methods using culture supernatant for optimum yield, purity and downstream applications. Scientific Reports, 9(1), 5335.

Pegtel, D. M., & Gould, S. J. (2019). Exosomes. Annual Review of Biochemistry, 88(1), 487–514.

Peng, H., Ying, C., Tan, S., Hu, B., & Sun, Z. (2018). An Improved Feature Selection Algorithm Based on Ant Colony Optimization. IEEE Access, 6, 69203–69209.

Petegrosso, R., Li, Z., & Kuang, R. (2020). Machine learning and statistical methods for clustering single-cell RNA-sequencing data. Briefings in Bioinformatics, 21(4), 1209–1223.

Plavec, Z., Domanska, A., Liu, X., Laine, P., Paulin, L., Varjosalo, M., Auvinen, P., Wolf, S. G., Anastasina, M., & Butcher, S. J. (2022). SARS-CoV-2 Production, Purification Methods and UV Inactivation for Proteomics and Structural Studies. Viruses, 14(9), 1989.

Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B., & Liao, Q. (2017). Why and when can deep-but not shallow-networks avoid the curse of dimensionality: A review. International Journal of Automation and Computing, 14(5), 503–519.

Press, W. H., & Teukolsky, S. A. (1990). Savitzky-Golay Smoothing Filters. Computers in Physics, 4(6), 669–672.

Rabinowits, G., Gerçel-Taylor, C., Day, J. M., Taylor, D. D., & Kloecker, G. H. (2009). Exosomal MicroRNA: A Diagnostic Marker for Lung Cancer. Clinical Lung Cancer, 10(1), 42–46.

Rahman, M. A., Barger, J. F., Lovat, F., Gao, M., Otterson, G. A., & Nana-Sinkam, P. (2016). Lung cancer exosomes as drivers of epithelial mesenchymal transition. Oncotarget, 7(34), 54852–54866.

Rangel-Ramírez, V. V., González-Sánchez, H. M., & Lucio-García, C. (2023). Exosomes: From biology to immunotherapy in infectious diseases. Infectious Diseases, 55(2), 79–107.

Raposo, G., Nijman, H. W., Stoorvogel, W., Liejendekker, R., Harding, C. V., Melief, C. J., & Geuze, H. J. (1996). B lymphocytes secrete antigen-presenting vesicles. The Journal of Experimental Medicine, 183(3), 1161–1172.

Raposo, G., & Stoorvogel, W. (2013). Extracellular vesicles: Exosomes, microvesicles, and friends. Journal of Cell Biology, 200(4), 373–383.

Rasheed, J., Hameed, A. A., Djeddi, C., Jamil, A., & Al-Turjman, F. (2021). A machine learning-based framework for diagnosis of COVID-19 from chest X-ray images. Interdisciplinary Sciences: Computational Life Sciences, 13(1), 103–117.

Roberg-Larsen, H., Lund, K., Seterdal, K. E., Solheim, S., Vehus, T., Solberg, N., Krauss, S., Lundanes, E., & Wilson, S. R. (2017). Mass spectrometric detection of 27-hydroxycholesterol in breast cancer exosomes. The Journal of Steroid Biochemistry and Molecular Biology, 169, 22–28.

Rutsaert, S., Bosman, K., Trypsteen, W., Nijhuis, M., & Vandekerckhove, L. (2018). Digital PCR as a tool to measure HIV persistence. Retrovirology, 15(1), 16.

Rykova, E. Y., Skvortsova, T. E., Hoffmann, A. L., Tamkovich, S. N., Starikov, A. V., Bryzgunova, O. E., Permjakova, V. I., Warnecke, J. M., Sczakiel, G., Vlassov, V. V., & Laktionov, P. P. (2008). Breast cancer diagnostics based on extracellular DNA and RNA circulating in blood. Biochemistry (Moscow) Supplement Series B: Biomedical Chemistry, 2(2), 208–213.

Sagar, G., Sah, R. P., Javeed, N., Dutta, S. K., Smyrk, T. C., Lau, J. S., Giorgadze, N., Tchkonia, T., Kirkland, J. L., Chari, S. T., & Mukhopadhyay, D. (2016). Pathogenesis of pancreatic cancer exosome-induced lipolysis in adipose tissue. Gut, 65(7), 1165–1174.

Sakri, S. B., Abdul Rashid, N. B., & Muhammad Zain, Z. (2018). Particle Swarm Optimization Feature Selection for Breast Cancer Recurrence Prediction. IEEE Access, 6, 29637–29647.

Saugstad, J. A., Lusardi, T. A., Van Keuren-Jensen, K. R., Phillips, J. I., Lind, B., Harrington, C. A., McFarland, T. J., Courtright, A. L., Reiman, R. A., Yeri, A. S., Kalani, M. Y. S., Adelson, P. D., Arango, J., Nolan, J. P., Duggan, E., Messer, K., Akers, J. C., Galasko, D. R., Quinn, J. F., … Hochberg, F. H. (2017). Analysis of extracellular RNA in cerebrospinal fluid. Journal of Extracellular Vesicles, 6(1), 1317577.

Shakes, D. C., Miller, D. M., & Nonet, M. L. (2012). Immunofluorescence Microscopy. In Methods in Cell Biology (Vol. 107, pp. 35–66). Elsevier.

Sood, A., Miller, A. M., Brogi, E., Sui, Y., Armenia, J., McDonough, E., Santamaria-Pang, A., Carlin, S., Stamper, A., Campos, C., Pang, Z., Li, Q., Port, E., Graeber, T. G., Schultz, N., Ginty, F., Larson, S. M., & Mellinghoff, I. K. (2016). Multiplexed immunofluorescence delineates proteomic cancer cell states associated with metabolism. JCI Insight, 1(6).

Soung, Y., Ford, S., Zhang, V., & Chung, J. (2017). Exosomes in Cancer Diagnostics. Cancers, 9(12), 8.

Stranahan, S. M., & Willets, K. A. (2010). Super-resolution Optical Imaging of Single-Molecule SERS Hot Spots. Nano Letters, 10(9), 3777–3784.

Suthaharan, S. (2016). Support Vector Machine. In S. Suthaharan, Machine Learning Models and Algorithms for Big Data Classification (Vol. 36, pp. 207–235). Springer US.

Tabakhi, S., Moradi, P., & Akhlaghian, F. (2014). An unsupervised feature selection algorithm based on ant colony optimization. Engineering Applications of Artificial Intelligence, 32, 112–123.

Tanaka, Y., Kamohara, H., Kinoshita, K., Kurashige, J., Ishimoto, T., Iwatsuki, M., Watanabe, M., & Baba, H. (2013). Clinical impact of serum exosomal microRNA-21 as a clinical biomarker in human esophageal squamous cell carcinoma: Exosomal MicroRNA-21 Expression in ESCC. Cancer, 119(6), 1159–1167.

Taverna, S., Giallombardo, M., Gil-Bazo, I., Carreca, A. P., Castiglia, M., Chacártegui, J., Araujo, A., Alessandro, R., Pauwels, P., Peeters, M., & Rolfo, C. (2016). Exosomes isolation and characterization in serum is feasible in non-small cell lung cancer patients: Critical analysis of evidence and potential role in clinical practice. Oncotarget, 7(19), 28748–28760.

Teymouri, M., Mollazadeh, S., Mortazavi, H., Naderi Ghale-noie, Z., Keyvani, V., Aghababaei, F., Hamblin, M. R., Abbaszadeh-Goudarzi, G., Pourghadamyari, H., Hashemian, S. M. R., & Mirzaei, H. (2021). Recent advances and challenges of RT-PCR tests for the diagnosis of COVID-19. Pathology - Research and Practice, 221, 153443.

Théry, C., Amigorena, S., Raposo, G., & Clayton, A. (2006). Isolation and Characterization of Exosomes from Cell Culture Supernatants and Biological Fluids. Current Protocols in Cell Biology, 30(1).

Trajkovic, K., Hsu, C., Chiantia, S., Rajendran, L., Wenzel, D., Wieland, F., Schwille, P., Brügger, B., & Simons, M. (2008). Ceramide Triggers Budding of Exosome Vesicles into Multivesicular Endosomes. Science, 319(5867), 1244–1247.

Van der Pol, E., Böing, A. N., Harrison, P., Sturk, A., & Nieuwland, R. (2012). Classification, functions, and clinical relevance of extracellular vesicles. Pharmacological Reviews, 64(3), 676–705.

Van Elslande, J., Houben, E., Depypere, M., Brackenier, A., Desmet, S., André, E., Van Ranst, M., Lagrou, K., & Vermeersch, P. (2020). Diagnostic performance of seven rapid IgG/IgM antibody tests and the Euroimmun IgA/IgG ELISA in COVID-19 patients. Clinical Microbiology and Infection, 26(8), 1082–1087.

Vella, L., Hill, A., & Cheng, L. (2016). Focus on Extracellular Vesicles: Exosomes and Their Role in Protein Trafficking and Biomarker Potential in Alzheimer's and Parkinson's Disease. International Journal of Molecular Sciences, 17(2), 173.

Vella, L. J., Scicluna, B. J., Cheng, L., Bawden, E. G., Masters, C. L., Ang, C., Willamson, N., McLean, C., Barnham, K. J., & Hill, A. F. (2017). A rigorous method to enrich for exosomes from brain tissue. Journal of Extracellular Vesicles, 6(1), 1348885.

Wang, D., Wang, X., Song, Y., Si, M., Sun, Y., Liu, X., Cui, S., Qu, X., & Yu, X. (2022). Exosomal miR-146a-5p and miR-155-5p promote CXCL12/CXCR7-induced metastasis of colorectal cancer by crosstalk with cancer-associated fibroblasts. Cell Death & Disease, 13(4), 380.

Wang, H., Li, X., Li, T., Zhang, S., Wang, L., Wu, X., & Liu, J. (2020). The genetic sequence, origin, and diagnosis of SARS-CoV-2. European Journal of Clinical Microbiology & Infectious Diseases, 39(9), 1629–1635.

Wang, J., Chen, J., & Sen, S. (2016). MicroRNA as Biomarkers and Diagnostics. Journal of Cellular Physiology, 231(1), 25–30.

Wang, L., Skotland, T., Berge, V., Sandvig, K., & Llorente, A. (2017). Exosomal proteins as prostate cancer biomarkers in urine: From mass spectrometry discovery to immunoassay-based validation. European Journal of Pharmaceutical Sciences, 98, 80–85.

Wang, P., Liang, O., Zhang, W., Schroeder, T., & Xie, Y. (2013). Ultra‐Sensitive Graphene‐Plasmonic Hybrid Platform for Label‐Free Detection. Advanced Materials, 25(35), 4918–4924.

Wang, P., Xia, M., Liang, O., Sun, K., Cipriano, A. F., Schroeder, T., Liu, H., & Xie, Y.-H. (2015). Label-Free SERS Selective Detection of Dopamine and Serotonin Using Graphene-Au Nanopyramid Heterostructure. Analytical Chemistry, 87(20), 10255–10261.

Wang, X., Yao, H., Xu, X., Zhang, P., Zhang, M., Shao, J., Xiao, Y., & Wang, H. (2020). Limits of Detection of 6 Approved RT–PCR Kits for the Novel SARS-Coronavirus-2 (SARS-CoV-2). Clinical Chemistry, 66(7), 977–979.

Watzinger, F., Ebner, K., & Lion, T. (2006). Detection and monitoring of virus infections by real-time PCR. Molecular Aspects of Medicine, 27(2–3), 254–298.

Webber, J., Stone, T. C., Katilius, E., Smith, B. C., Gordon, B., Mason, M. D., Tabi, Z., Brewis, I. A., & Clayton, A. (2014). Proteomics Analysis of Cancer Exosomes Using a Novel Modified

Aptamer-based Array (SOMAscanTM) Platform. Molecular & Cellular Proteomics, 13(4), 1050–1064.

Welton, J. L., Brennan, P., Gurney, M., Webber, J. P., Spary, L. K., Carton, D. G., Falcón-Pérez, J. M., Walton, S. P., Mason, M. D., Tabi, Z., & Clayton, A. (2016). Proteomics analysis of vesicles isolated from plasma and urine of prostate cancer patients using a multiplex, aptamer-based protein array. Journal of Extracellular Vesicles, 5(1), 31209.

Wu, Q., Yu, L., Lin, X., Zheng, Q., Zhang, S., Chen, D., Pan, X., & Huang, Y. (2020). Combination of Serum miRNAs with Serum Exosomal miRNAs in Early Diagnosis for Non-Small-Cell Lung Cancer. Cancer Management and Research, Volume 12, 485–495.

Xanthopoulos, P., Pardalos, P. M., & Trafalis, T. B. (2013). Linear Discriminant Analysis. In P. Xanthopoulos, P. M. Pardalos, & T. B. Trafalis, Robust Data Mining (pp. 27–33). Springer New York.

Xiao, T., Zhang, W., Jiao, B., Pan, C.-Z., Liu, X., & Shen, L. (2017). The role of exosomes in the pathogenesis of Alzheimer' disease. Translational Neurodegeneration, 6(1), 3.

Xie, Y., Su, X., Wen, Y., Zheng, C., & Li, M. (2022). Artificial Intelligent Label-Free SERS Profiling of Serum Exosomes for Breast Cancer Diagnosis and Postoperative Assessment. Nano Letters, 22(19), 7910–7918.

Yang, A., Zhang, W., Wang, J., Yang, K., Han, Y., & Zhang, L. (2020). Review on the Application of Machine Learning Algorithms in the Sequence Data Mining of DNA. Frontiers in Bioengineering and Biotechnology, 8, 1032.

Yuan, Y., Chen, L., Wu, H., & Li, L. (2022). Advanced agricultural disease image recognition technologies: A review. Information Processing in Agriculture, 9(1), 48–59.

Zeng, Z., Liu, Y., & Wei, J. (2016). Recent advances in surface-enhanced raman spectroscopy (SERS): Finite-difference time-domain (FDTD) method for SERS and sensing applications. TrAC Trends in Analytical Chemistry, 75, 162–173.

Zhang, M.-H., Yuan, Y.-F., Liu, L.-J., Wei, Y.-X., Yin, W.-Y., Zheng, L.-Z.-Y., Tang, Y.-Y., Lv, Z., & Zhu, F. (2023). Dysregulated microRNAs as a biomarker for diagnosis and prognosis of hepatitis B virus-associated hepatocellular carcinoma. World Journal of Gastroenterology, 29(31), 4706–4735.

Zhang, Y., Bi, J., Huang, J., Tang, Y., Du, S., & Li, P. (2020). Exosome: A Review of Its Classification, Isolation Techniques, Storage, Diagnostic and Targeted Therapy Applications. International Journal of Nanomedicine, Volume 15, 6917–6934.

Zhao, Z., Yang, Y., Zeng, Y., & He, M. (2016). A microfluidic ExoSearch chip for multiplexed exosome detection towards blood-based ovarian cancer diagnosis. Lab on a Chip, 16(3), 489–496.

Zhou, L., Xu, Z., Castiglione, G. M., Soiberman, U. S., Eberhart, C. G., & Duh, E. J. (2020). ACE2 and TMPRSS2 are expressed on the human ocular surface, suggesting susceptibility to SARS-CoV-2 infection. The Ocular Surface, 18(4), 537–544.

Zhou, X., Xie, F., Wang, L., Zhang, L., Zhang, S., Fang, M., & Zhou, F. (2020). The function and clinical application of extracellular vesicles in innate immune regulation. Cellular & Molecular Immunology, 17(4), 323–334.

Chapter 4 Prospect of detecting early metastasis of Non-small cell lung cancer by SERS plus machine learning

4.1 Non-small cell lung cancer and role of exosomes in metastasis

NSCLC is a major cause of cancer-related death globally and recurs in 30-55% of patients following surgery, most commonly as metastatic disease. Metastatic behavior, a hallmark of cancer, is often considered a late event, but recent findings suggest that the metastatic process may initiate during early-stage disease in some patients. Early-stage micrometastasis may often be present during surgery but below the level of clinical detection (Fontebasso & Dubinett, 2015; Salehi-Rad et al., 2020). Interestingly, clinical findings are consistent with laboratory-based studies indicating that metastatic dissemination may occur during early tumor development, particularly in the context of EMT in many cancers, including NSCLC (Eyles et al., 2010; Hüsemann et al., 2008; Podsypanina et al., 2008; Rhim et al., 2012). Exposure to carcinogens, inflammation, hypoxia, and aberrant genetic modulation leads to EMT activation and selection of cells within the premalignant lesions with metastatic potential, resulting in physiologically different and clonally linked airway lesions. Early migratory premalignant lesions may harbor decipherable targets that can be used to detect and intercept the malignant progression. Recent studies indicate a pattern of metastatic lung cancer disease progression (Hüyük et al., 2023; Kawano et al., 2002; Tang et al., 2021), suggesting that early detection and targeting of lung cancer at an early stage would dramatically enhance patient survival. Low-dose CT (LDCT) scan is the current standard of care for lung cancer screening, although fewer than 6% of the 15 million eligible high-risk US population take advantage of screening. Liquid biopsy techniques, which rely on cell-free DNA (cfDNA), circulating tumor cells (CTCs), or EVs to diagnose cancer, are being evaluated for early detection

and tracking tumor progression (Mukherjee et al., 2022). Thus far, cfDNA and CTC-based liquid biopsies appear to have their greatest utility in monitoring advanced disease. Thus, EV-based ultrasensitive diagnostic and screening tools have the potential to revolutionize the detection and treatment of lung cancer.

Reports suggest that cancer-derived exosomes can act via inter-cellular communication to induce EMT and metastasis in many cancers, including lung cancer (Y.-L. Hsu et al., 2017; Shimada & Minna, 2017). Examining tumor-derived exosomes in the blood and other bodily fluids may provide potential clues, serve as promising cancer biomarkers, and have potential for cancer diagnosis and prognosis (S. Chauhan et al., 2022). Exosomes are relatively physiologically stable and have unique early migratory fingerprints. Thus, they may be potentially examined using a single exosome-based detection method in developing an exosome-based liquid biopsy for identifying early lung cancer metastasis. Single exosome-based analysis enhances the likelihood of discovering distinct cancer-derived exosomes, particularly when the cancer-related exosome subpopulation is small. In contrast, bulk analysis is hindered by a high false negative rate because the cancer-related exosomal signal can be readily masked by the dominant normal exosomal fraction. We conducted single exosome-based characterization and analysis by combining SERS nano-pyramidal substrate and Raman map scanning methods.

Studying early migration/metastasis in lung cancer is a major challenge due to the paucity of relevant model systems. We have recently discovered a unique HM subpopulation of premalignant (expressing mutant KRAS-G12D and p53 knockdown), high-risk HBECs, using a novel "constricted migration" selection strategy with enhanced metastatic potential *in vivo* (Pagano et al., 2017). Comparative RNA-seq datasets illustrate the increased expression of key EMT genes in HBEC-HM compared to HBEC unselected cells. This unique subpopulation of HBEC-HM

offers a unique model to investigate premalignant cell migration and early metastasis (Paul et al., 2018). Just a small percentage of cells (<1%) in the UN population are highly migratory; hence most HBEC-UN cells can be considered as low migratory and show poor migratory and invasive properties and are distinctly different from the HBEC-HM population. To undertake molecular detection and analysis of early lung cancer metastatic phenotype, we isolated and characterized exosomes from these HBEC-UN and HBEC-HM cells and then studied their spectroscopic fingerprints. To further evaluate our data in the context of a malignant landscape, we have used MPE. MPE frequently results from tumor growth and metastatic progression and occurs in 30-35% of lung cancer patients. We have utilized a lung cancer patient MPE-biobank for exosome isolation and characterization from individual patient's MPE. Conventional methods (ELISA, PCR, SPR) used to identify exosomes are biomarker-driven and require labeling, making lung cancer exosome detection challenging, especially at the single-exosome level. Spectroscopic study of exosomes depends on spectral fingerprinting of molecular patterns, a new type of potential biomarker evaluated utilizing SERS for disease prediction.

SERS-based single-exosomal characterization can be highly sensitive, but achieving this goal is challenging and requires advanced methods (Kruglik et al., 2019). As shown in Figure 4.1, this technique of single-exosome fingerprinting is achieved by using quasi-periodic gold pyramids for enabling LSPR and spatial Raman map scanning to confirm the spectral source. The results have established the strength of our technology in differentiating exosomes from different subpopulations of premalignant cancer cells and is a potential strategy for single-exosome fingerprinting. This work is based on technical groundwork, and the results suggest that early lung metastatic cell-derived exosomes possess distinct molecular signatures and exhibit unique SERS scattering profiles. We utilized patient-derived MPE to confirm that these early metastatic

exosomal SERS signatures are also common in metastatic lung cancer-derived exosomes. ML-assisted SERS profiling of MPE-derived exosomes revealed the existence of HBEC-HM and HBEC-UN derived exosomal fingerprints, and the abundance of such HBEC-HM derived exosomal fingerprints correlated with the metastatic potential of NSCLC. We applied to preliminarily visualize the spectral signature differences among exosomes from different sources. In view of the high dimensionality and peak complexity of the SERS spectrum, we implemented feature selection and nearest neighbors-based algorithm to identify exosomal biomarkers. Our findings suggest that the ML-based SERS analysis of single exosomes could become a label-free, ultrasensitive, accurate detection technology for NSCLC early metastasis. Our approach has promise for future clinical applications that might facilitate the detection of micrometastasis and alter the treatment outcome of lung cancer.

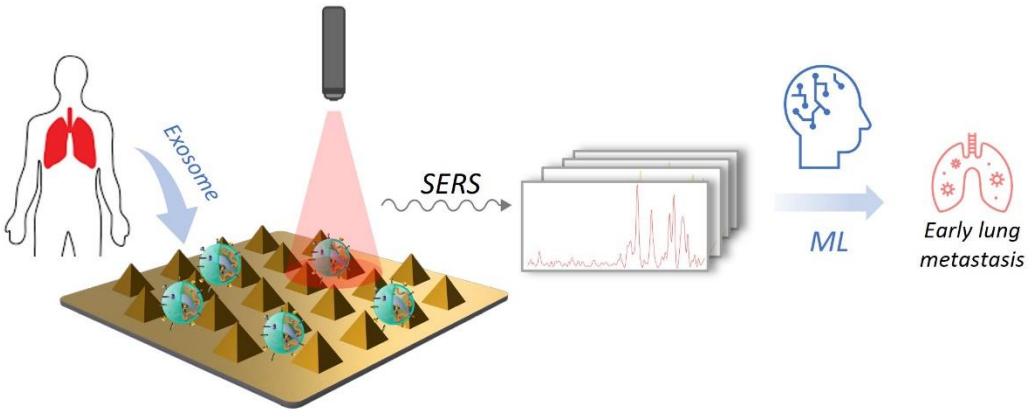


Figure 4.1 Procedure of single-vesicle SERS characterization and detection of early metastasis.

4.2 Materials and methods

4.2.1 Exosome isolation

Exosomes were purified by SUC as previously described. HBEC KRAS-Mut (G12D P53) Snail/Vector expressing HM and UN cells were cultured as described earlier (Pagano et al., 2017).

Cells were cultured in an FBS-free conditioned medium, pre-cleared of exosomes and protein aggregates prior to use for cell culture by ultracentrifugation. In brief, cell culture media were collected at 72 h after changing the medium for exosome isolation. Cell culture supernatants were first centrifuged at 300 g at 4 °C for 10 min and then at 2,000 g at 4°C for 15 min to remove contaminating cells and apoptotic bodies, respectively. The supernatants were then further centrifuged at 12,000 g at 4 °C for 45 min to remove sub-cellular debris. The clear supernatant was then filtered using 0.22 μm pore filters, followed by ultracentrifugation (Model, L8-M70, Beckman Coulter, USA) at 110,000 g at 4 °C for 70 mins. The resulting pellets were resuspended in pre-chilled PBS and again ultra-centrifuged at 110,000 g and 4 °C for 70 min. The final suspension is passed through 0.1 μm pore filters and subjected to characterization. The final exosome pellet was resuspended in 50-100 μL PBS for qNano measurement, and in a 2% paraformaldehyde (PFA) solution in Milli-Q water for TEM experiments.

4.2.2 Exosome characterization

4.2.2.1 Transmission electron microscopy

Formvar carbon-coated grids (FCF400-CU, Electron Microscopy Sciences) were glow-discharged on a Pelco easiGlow instrument (Ted Pella Inc., USA) for 2 min. Drops of PFA-fixed exosomes were then placed on the grids and incubated for 20 min in a dry environment. The grids were washed by floating them upside down on drops of Milli-Q water. The exosomes were further post-fixed in 1% glutaraldehyde for 5 min and stained successively in freshly prepared 2% uranyl acetate and 2% methylcellulose/0.4% uranyl acetate. Grids were imaged using a 100CX JEOL electron microscope at UCLA. Images were taken using a cooled slow-scan CCD camera at a magnification of 80,000 × following the published protocol (Yan et al., 2019).

4.2.2.2 Western immunoblotting

Exosome markers, CD63, CD81, Alix, CD9, and TSG101 antibodies have been purchased for Abcam (Cambridge, United Kingdom), and flow antibodies were purchased from BioLegend. Western blots were performed according to standard procedures (Paul et al., 2018; Yan et al., 2019). Cells grew to 80% confluence in T25 flasks for 72 hrs, followed by media collection. Cell/ Exosomes were isolated from the collected media and lysed with RIPA buffer using standard methods. 10 μg of each exosomal/ cell lysate was loaded per lane, and proteins were resolved by SDS-PAGE and transferred to an Immobilon-P Transfer Membrane (Millipore, Billerica, MA). The membranes were blocked with 5% nonfat milk and then incubated with primary antibodies diluted in a blocking solution according to the manufacturer's recommendations. Horseradish peroxidase-conjugated secondary antibodies (Bio-Rad, Hercules, CA) and enhanced chemiluminescence (ECL) reagent (Amersham Biosciences, Piscataway, NJ) were used for protein detection. Densitometry was performed in ImageJ using the "analyze gels" function.

4.2.2.3 Exosome size determination

The exosome pellets were resuspended in chilled PBS (Thermo Fisher Scientific, USA), pooled, and ultra-centrifuged at 110,000 ×g for 70 min at 4ºC. The final pellet of exosomes was resuspended in 50-100 μL PBS and stored temporarily at 4°C until use. As previously described, the exosome size and particle number were analyzed using the TRPS technology, qNano IZON system (Izon, Cambridge, MA, USA) (Greenberg et al., 2021; Maas et al., 2014). Standard beads were used to calibrate the system for voltage, stretch, pressure, and baseline current. The diluted exosome sample was passed through the NP100 nanopore (for a 50-200 nm size range), and the qNano IZON Control Suite software was used for data processing (Maas et al., 2014). The final exosome pellet was resuspended in PBS, and protein concentration was measured by BCA (Pierce,

Thermo Fisher Scientific). The integrity and time-dependent cellular uptake of exosomes were analyzed using fluorescently labeled with Dil dye (1,1'-Dioctadecyl-3,3,3',3'-Tetramethylindocarbocyanine Perchlorate) (Thermo Fisher Scientific, USA) and LSM880 confocal microscopy.

4.2.3 Cell lines and clinical samples

The HBEC cell line was generated from a patient's large airway and immortalized in the absence of viral oncogenes, as previously reported (Pagano et al., 2017). Typically, the HBEC-parental cells are non-tumorigenic (Grant et al., 2014; Ramirez et al., 2004); with the overexpression of KRAS G12D and P53 downregulation (HBEC-P53/KRAS or HBEC mut), cells begin to resemble "at-risk" epithelia. P53 silencing and KRAS activating mutations are strongly associated with NSCLC, and with the progressive accumulation of mutations like Snail, the at-risk cells gain anchorage-independent growth (AIG) in vitro and develop tumors and metastases in vivo (Grant et al., 2014; M. Sato et al., 2013). HBEC-based human carcinogenesis model is characterized by its potential for malignant conversion into cancer cells with metastatic capacity. We recently described the isolation of a highly migratory HBEC subpopulation and characterized them (Pagano et al., 2017; Paul et al., 2018). This unique bronchial subpopulation is isolated using our "constricted migration" based migration model and selected for their deformability by using cutting-edge physomic techniques, including deformability cytometry and Atomic Force Microscope. HBEC lines were cultures in keratinocyte serum-free media (Life Technologies) with 30 μg/mL bovine pituitary extract and 0.2 ng/mL recombinant EGF (Life Technologies). Cell culture maintenance and creation of HBEC-parental, -vector, and -Snail lines are described elsewhere (Pagano et al., 2017; Paul et al., 2018). Clinical samples and IRB information for use of MPE (patients' MPE) were provided by UC Davis, Department of Internal Medicine. Healthy

human serum specimens were purchased from STEMCELL Technologies Inc. Exosomes were isolated following the protocol in Section 3.3.1.

4.3 Results

4.3.2 SERS substrate and single-exosome fingerprinting

Scanning Electron Microscopy (SEM) images are collected to demonstrate the aerial arrangement of the Au-nano-pyramid and the exosome specimens on the top layer, as shown in Figure 4.3A and Figure 4.3B. Exosomes are located mainly on the lateral facet of the pyramids, highlighted by red circles, at which the SERS 'hot-spots' are located. Moreover, the impact of buffer crystallization is also shown by the 'cloudy' areas (white arrows) that blur the SEM imaging and decrease SERS spectral yield. Single vesicle identification is one of our platform's key advantages. The laser spot size of around 1 μm x 1 μm during the obtaining map step ensured single exosome characterization. Compared with bulk EV/exosomal analysis, the individual exosome-derived signature is measured by Raman map scanning on our platform, therefore, the cancer-derived exosomes can be studied explicitly for molecular signature biomarker extraction instead of being 'buried' by other exosome subpopulations. To demonstrate the spectral response of the exosomes on Raman map measurements, the Raman band of lipids (around 1450 cm^{-1}) (H. Sato et al., 2019) is selected for visualizing the intensity distribution around an exosome, as shown in Figure 4.3C. Figure 4.3D shows the Raman intensity profile of a single EV, and the circular outline agrees with the typical shape of EVs. The spatial spread of individual EV shown on intensity map (Figure 4.3D) is larger than the actual size due to the convolution of the laser beam of 1 um with and EV of 150 nm in size. According to Figure 4.3B, the average spacing between EVs is approximately several micrometers, which makes it possible for obtaining SERS signals from one single EV despite the laser beam convolution.

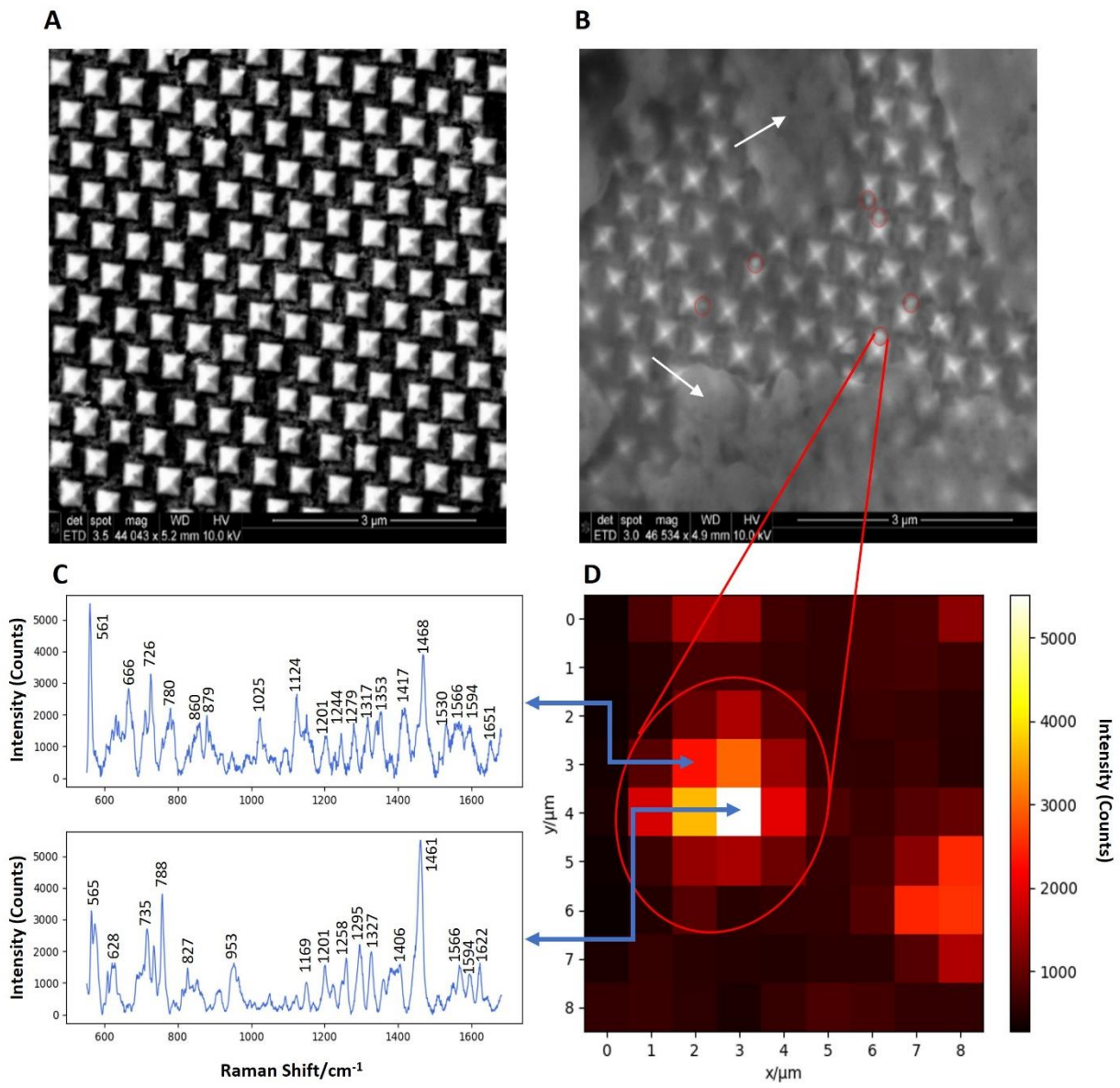


Figure 4.2 Microscopy characterization of SERS Au periodic pyramidal substrate. (A) SEM image of blank SERS substrate (magnification of 44043x). (B) SEM image of SERS substrate with EVs (magnification of 46534x, EVs are marked by red circles, crystals are marked by white arrows), while arrows mark the buffer crystallization. (C) SERS spectra from the center and margin of a single EV according to the intensity map (peak assignments are given in Table S1). (D) Raman intensity map of a single EV.

4.3.3 Exosome subgroups differentiation

We have established a "constricted migration" based selection strategy to isolate HBEC-HM and HBEC-UN cells (Figure 4.3A), which are premalignant cells (expressing KRAS-G12D and p53 knockdown: mutant (M)) (Pagano et al., 2017). Previous studies have shown that the overexpression of SNAIL is associated with increased motility and migration in NSCLC. We utilized genetically engineered HBECs to over-express SNAIL and compared Snail-modified vs. Snail-unmodified (Vector-modified) HBECs to assess their differences in migration. We found that the overexpression of SNAIL in HBEC-HM cells resulted in enhanced migration (Pagano et al., 2017). HBEC-M-S-HM cells show a significant difference in their actin cytoskeletal structure (Figure 4.3B) and exhibit enhanced EMT features and migration (Pagano et al., 2017). This subpopulation of HM-HBECs offers a unique model to investigate premalignant cell migration and early metastasis. Exosome samples derived from four different gene-modified cell lines were evaluated, including HBEC-M-S-HM (M stands for mutant, S stands for Snail-modified), HBEC-M-V-HM (V stands for vector, i.e., Snail-unmodified), HBEC-M-S-UN and HBEC-M-V-UN.

Our previous paper compared the ExoQuick kit vs. the SUC for exosome isolation. Though ExoQuick-mediated exosome isolation yielded a higher concentration of exosomes, the exosomes were more heterogeneous (Yan et al., 2019). Because we isolated exosomes from HBEC culture media and MPE, and the sample volume was not limited, the SUC method was utilized (Figure 4.3C) (Paul et al., 2018). We confirmed the presence, integrity, and size distribution of our isolated exosomes using TEM (Figure 4.3D). We verified that the preparation contained exosomes using western immunoblotting for EV/exosome markers, including CD63, CD81, Flotillin, and Hsp70. Calnexin was utilized as a negative marker. The western immunoblotting confirmed that the isolated exosomes were enriched with exosomal markers. Exosome size was characterized by

using a qNano Gold, which measures particles following the principle of Tunable Resistive Pulse Sensing (TRPS). The isolated exosome population showed a size distribution ranging from around 50-150 nm with a mean diameter of ~ 80-90 nm (Figure 4.3F). No significant differences were identified in the size of the exosomes isolated from HBEC-HM and the HBEC-UN exosomes (Figure 4.3F). The process was repeated with the MPE-derived exosomes. The exosomal protein concentration showed that the HBEC-M-S-HM had higher protein concentrations compared to the HBEC-M-V-UN cells, as measured by the BCA method (Figure 4.3G). In contrast, the HBEC-M-V-HM and HBEC-M-S-UN cells showed higher protein concentrations but were not statistically significant compared to HBEC3-UN-derived exosomes (Figure 4.3G). LC-MS/MS analyses and unsupervised hierarchical clustering revealed a clear distinction between HBEC-HM and HBEC-UN exosome proteomic profiles (data not shown), suggesting the HBEC-HM-derived exosomes are distinct from the HBEC-UN-derived exosomes, reminiscent of the parent cell types.

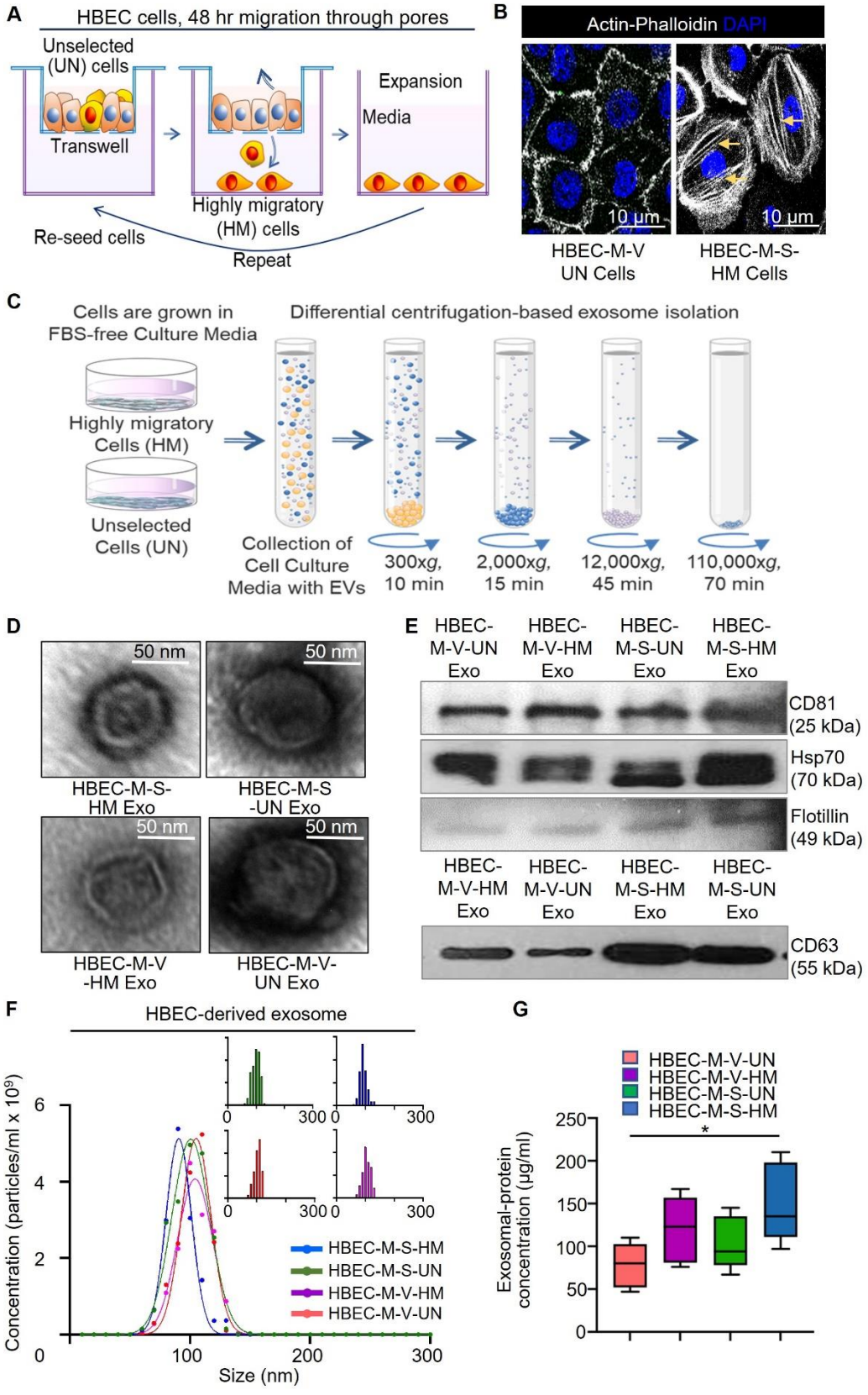

A
HBEC cells, 48 hr migration through pores
Unselected (UN) cells
Transwell
Highly migratory (HM) cells
Expansion
Media
Re-seed cells
Repeat
B
Actin-Phalloidin DAPI
10 μm
10 μm
HBEC-M-V UN Cells
HBEC-M-S-HM Cells
C
Cells are grown in FBS-free Culture Media
Differential centrifugation-based exosome isolation
Highly migratory Cells (HM)
Unselected Cells (UN)
Collection of Cell Culture Media with EVs
300xg, 10 min
2,000xg, 15 min
12,000xg, 45 min
110,000xg, 70 min
D
50 nm
HBEC-M-S-HM Exo
HBEC-M-S-UN Exo
HBEC-M-V-HM Exo
HBEC-M-V-UN Exo
E
HBEC-M-V-UN Exo
HBEC-M-V-HM Exo
HBEC-M-S-UN Exo
HBEC-M-S-HM Exo
CD81 (25 kDa)
Hsp70 (70 kDa)
Flotillin (49 kDa)
CD63 (55 kDa)
F
HBEC-derived exosome
Concentration (particles/ml x 10^9)
Size (nm)
HBEC-M-S-HM
HBEC-M-S-UN
HBEC-M-V-HM
HBEC-M-V-UN
G
Exosomal-protein concentration (μg/ml)
*

Figure 4.3 Isolation and characterization of HBEC-derived exosome. (A) Schematics showing the ex-vivo micropore selection technique-assisted isolation of the HBEC-HM. Cells were grown on the transwell, and cells that migrated through to the bottom well were propagated, and the process repeated. The UN contain a very tiny fraction (<1%) of cells that are HM, so the other UN cells do not pass through the membrane. (B) HBEC-HM display enhanced EMT positive phenotype with the profound actin cytoskeleton (stress fibers shown with yellow arrow) as studied using confocal images (Dronpa-Actin, DAPI-nucleus). (C) Schematics show the process of exosome isolation using sequential SUC. (D) Transmission electron microscopic characterization of exosomes derived from HBEC-M-S/V-HM/UN cells. (E) Western immunoblotting was used to probe for exosome markers on exosomal and respective cell lysates. (F) The exosome size was characterized using a qNano Gold instrument which measures particle size based on the principle of Tunable Resistive Pulse Sensing (TRPS). (G) Exosome protein concentration was measured and compared.

Raman map scanning SOP was used to characterize the samples and collect spectral data. Following the single NBP scanning protocol (the protocol details are given in the Methods section), we measured approximately 50 exosomes for each sample. Upon obtaining exosome spectra, we found heterogeneous signatures existing within and across the samples, indicating that the exosomes are from different sources. After obtaining qualified spectral data, we used statistical analysis and machine learning to differentiate the HBEC-HM-derived exosomes from the HBEC-UN-derived exosomes.

To visualize the spectral signatures of the exosomes attributed to the four cell lines, we plotted the representative spectra shown in Figure 4.4A. We then applied dimensionality reduction operation using LDA to visualize the differences of fingerprints. As a supervised learning algorithm, LDA searches for transformed dimensions that maximize the inter-group distance, representing the differences in spectral signatures between groups (Xanthopoulos et al., 2013). The

spectral features were further illustrated by projecting the data points to the first two dominant dimensions, i.e., LD1 and LD2. LDA analyses were performed using the Linear Discriminant Analysis package of the Python Scikit-learn module.

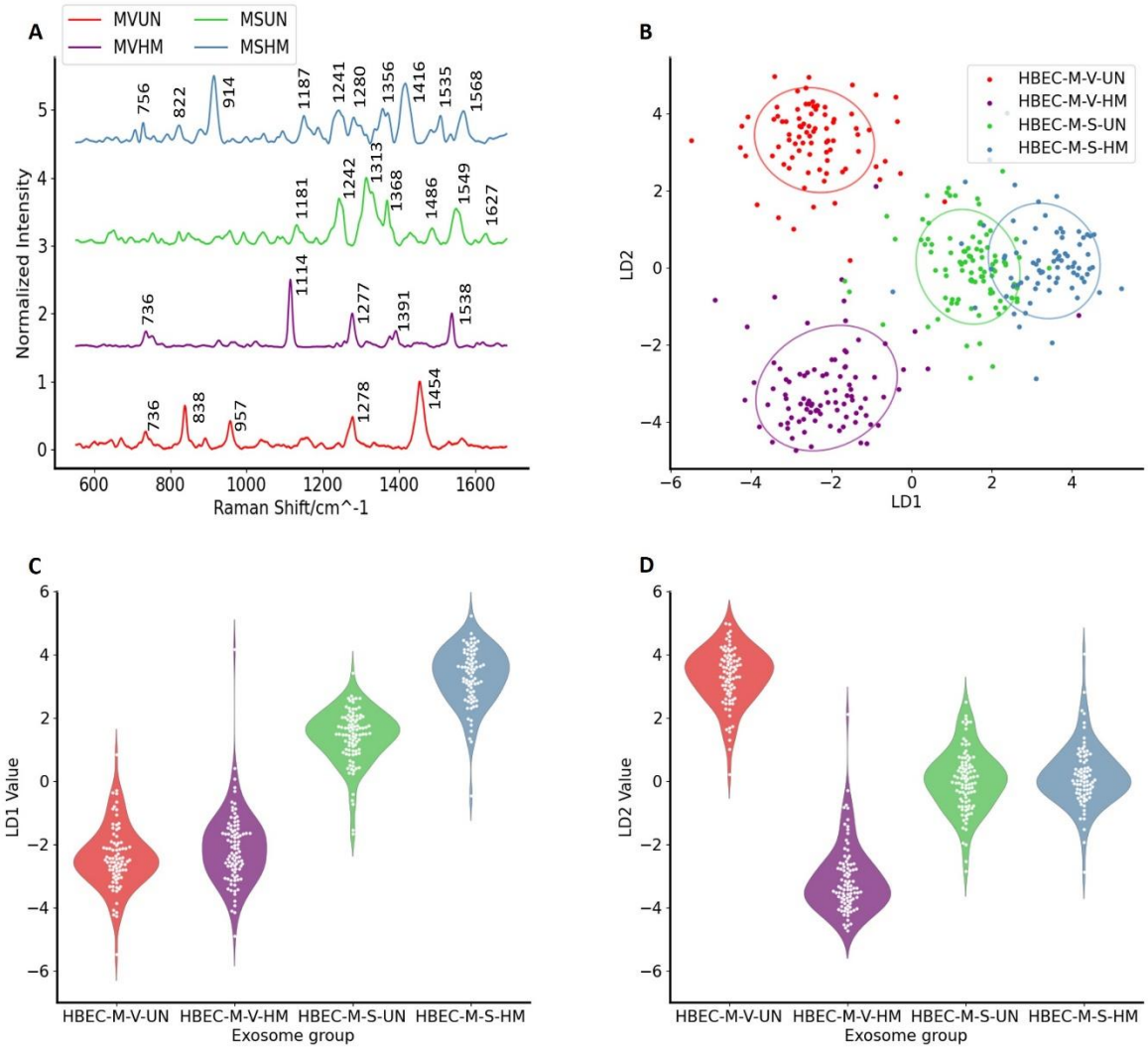


Figure 4.2 Statistical analyses of HBECs derived exosomal SERS fingerprints. (A) Averaged spectrum of exosomal fingerprints from each cell line (peak assignments are given in Table S1). (B) Clustering of exosomal fingerprints from four cell lines with LDA dimensionality reduction. (C) LD1 values of four cell lines' clusters. (D) LD2 values of four cell lines' clusters.

Interestingly, the overexpression of Snail adds special spectral features to exosomes, which may be due to their migratory properties as evident after data analyses. Additionally, Snail-overexpression decreases the dissimilarities between HBEC-M-S-UN and HBEC-M-S-HM exosomal signatures as indicated by their closer datapoints. The gene-modified group gives higher scores along the LD1 axis with positive mean values, while the control group is negative. Additionally, the HBEC-M-S highly migratory subgroup (HBEC-M-S-HM) oversteps the lower migratory subgroup (HBEC-M-S-UN) slightly, shown by the mean score of 3.7 and 1.7, however, they share an overlapping score range from 2.2 to 3.0. On the LD2 dimension, the higher migratory HBEC-M-V-HM group indicates a clearly lower value than the lower migratory HBEC-M-V-UN group by 3.8 versus -3.7, while the snail overexpressed exosome groups show nearly no difference. We can find that LD1 may extract the Snail modification features and slightly indicate the migratory difference among HBEC-M-S. LD2 shows on the different migratory properties of the Snail-unmodified exosomes. Therefore, the spectroscopic signature of exosomes derived from four cell types could be distinguished by statistical analyses, based on which we could identify exosomal fingerprints and the source cells. This raised the possibility that the exosomal fingers could be utilized in clinical biospecimen.

4.3.5 Illustration of SERS spectral signatures by feature selection

SERS signatures are reported to contain bio-molecular bonding information. In the previous section, we noted a small overlap between HBEC-HM and HBEC-UN-derived exosomes, given the results of LDA. To further investigate the signature differences, we applied feature selection analysis to investigate the principal spectral differences between the HM and UN groups. The process of feature selection is shown in Figure 4.5. Specifically, we used ACOFS and supervised learning (classification) for evaluating the differentiation. Within the original

spectrum's 1011 Raman shifts (features), there are redundant and irrelevant features that have no (or negative) contribution to the differentiation. Therefore, we implemented ACOFS to extract the top important features that provide the best differentiation between HBEC-HM and HBEC-UN derived exosomes. These features were evaluated for classification accuracy. As a heuristic algorithm, ACOFS gradually approaches the optimal feature subsets through generations of searching. Figure 4.6A demonstrates that the classification accuracy steadily increases as the searching generation grows, indicating that this feature selection procedure gradually removes the useless features while keeping only those that contribute to the differentiation, as shown in Figure 4.6B.

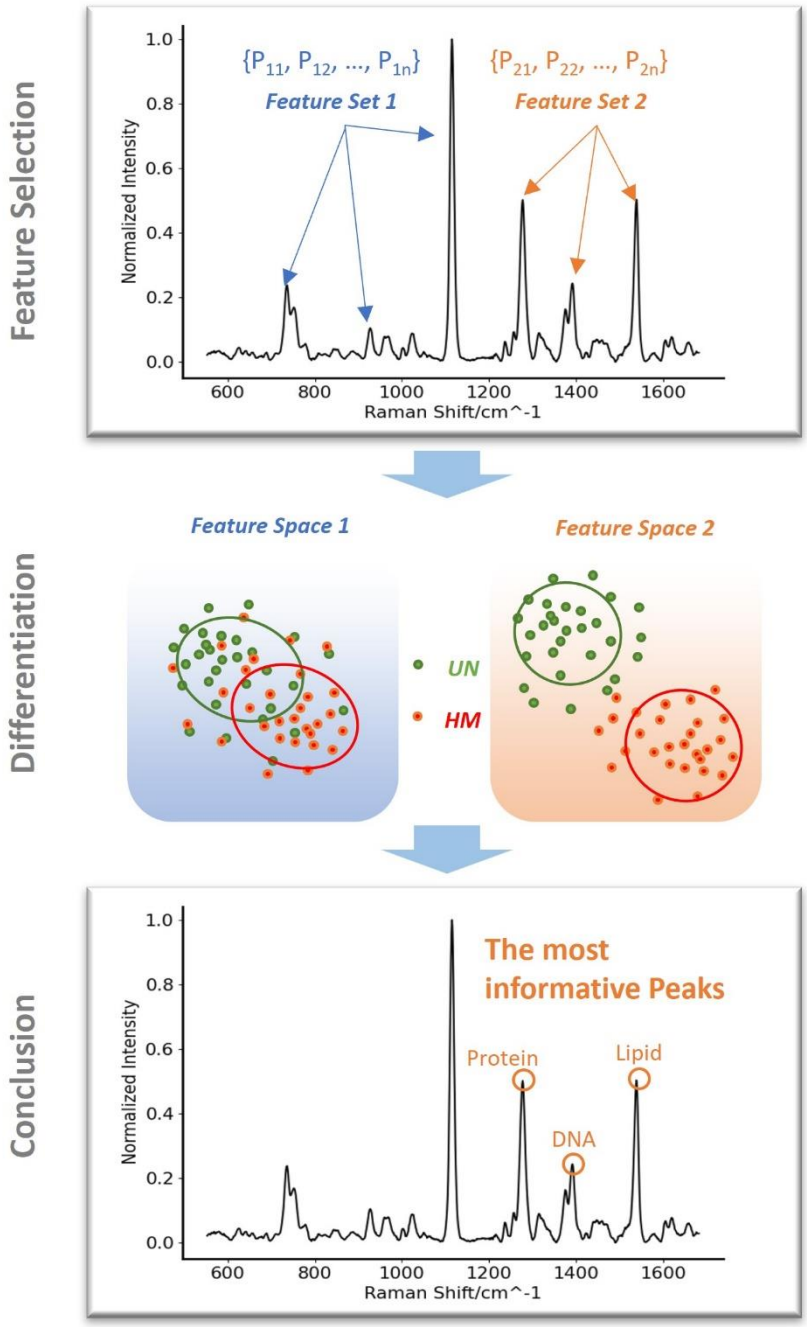


Figure 4.3 Procedure of feature selection. Two example feature subsets are compared in terms of the differentiation between HBEC-HM and HBEC-UN, the second subset gives a better outcome composed of peaks of protein, DNA, and lipid.

We conducted three rounds of feature selection to average the statistical fluctuations. After each round, every feature was given an importance score by running permutation importance (Altmann et al., 2010), which denoted the contribution to the differentiation. Figure 4.6C presents the importance of the top 20 features. According to the result, the most dominant Raman band differences between different migratory exosome subtypes lay approximately from 1200 cm^{-1} to 1600 cm^{-1}, similar feature Raman bands for NSCLC have been reported (Leng et al., 2023; Shin et al., 2018; H. Wang et al., 2018). The molecular information is listed in Table 4.1. The analysis results indicate genomic differences such as Guanine (1335 cm^{-1}) and Cytosine (1605 cm^{-1}, 1610 cm^{-1}). The Raman band at 1241 cm^{-1} is reported to be a nucleic acid biomarker for malignant tissues (Cheng et al., 2005). Besides, the results also reveal proteomic differences, namely glycine (1335 cm^{-1}), tyrosine (1605 cm^{-1}), Tryptophan (1624 cm^{-1}), proline (1547 cm^{-1}), phenylalanine (1581 cm^{-1}), β-carotene (1395 cm^{-1}) as well as amide I and III (1598 cm^{-1}, 1600 cm^{-1}, 1231 cm^{-1}). The molecules included in this Raman shift range are responsible for a variety of active biological processes, the differences on which may suggest different metabolism during cancer development (Leng et al., 2023). We anticipated that the HBEC derived exosomes may contain NSCLC metastatic molecular signatures, therefore studies involving patient specimens were conducted.

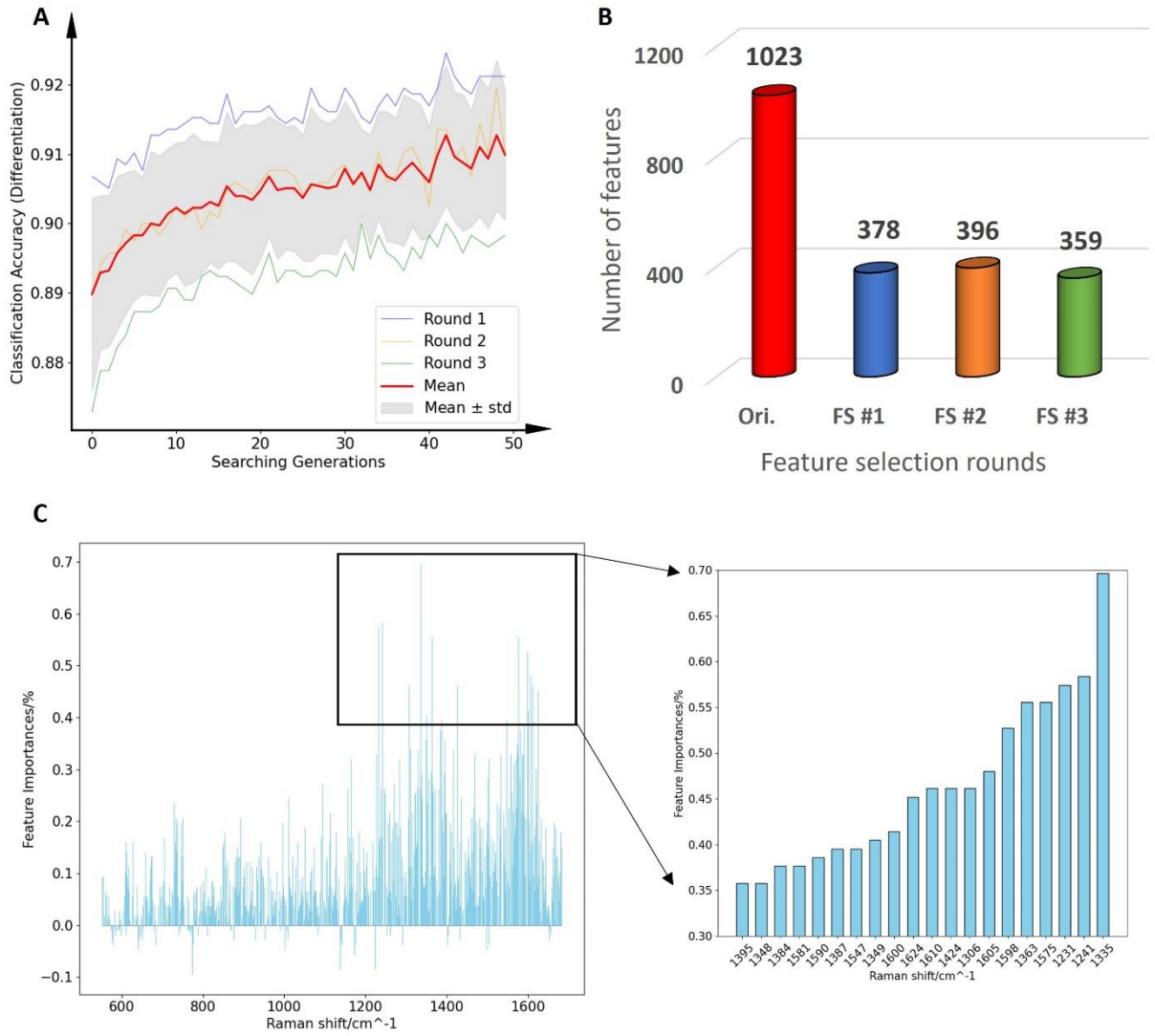


Figure 4.4 Classification accuracy changes during feature selection. (A) Differentiation between HBEC-HM and HBEC-UN (classification accuracy) gradually increases, which indicates that the optimal feature subsets are being extracted. (B) Comparison of the number of features among three rounds of feature selection and the original state. (C) Average importance scores of each feature (Raman shift) after feature selection. The top twenty important features are plotted and analyzed.

Table 4.1 Molecular information of the top twenty important features.

Top features (cm^-1)	**Peak assignments**
1335	Guanine, CH3CH2 wagging in nucleic acid and glycine backbone
1241	PO2- (asym.) belonging to nucleic acids (suggest an increase in nucleic acid in the malignant tissues)
1231	Amide III and CH2 wagging vibrations from glycine backbone and proline side chains
1575, 1573	Ring breathing modes in DNA; G, A
1363	Guanine (N7, B, Z-marker)
1598, 1600	C=O in Amide I
1306	CH3/CH2 twisting or bending mode of lipid/collagen
1424	Deoxyribose (B, Z-marker)
1610	Cytosine (NH2)
1624	Tryptophan
1348, 1349	Carbon particle
1547	Proline
1384, 1387	CH3 band
1590	Carbon particle
1581	δ(C=C), phenylalanine
1395	β-carotene

4.3.6 Elucidating High Migratory and Unselected exosomal features in patient samples

Upon establishing premalignant HBEC cell line derived exosomal signatures, we investigated whether the HBECs derived exosomal signatures could indicate the presence of metastasis in patients. Premalignant exosomal signatures were sought within exosome specimens derived from patients' MPE, subsequently correlation between the amount of premalignant exosomal signatures versus the status of NSCLC metastasis was studied. Patient information is given in Table S2. Characterization of these exosomes using TEM and western immunoblotting are presented in Figure 4.7. We used exosome specimens isolated from NSCLC patients' MPE as the test group to investigate the premalignant cell derived exosomes.

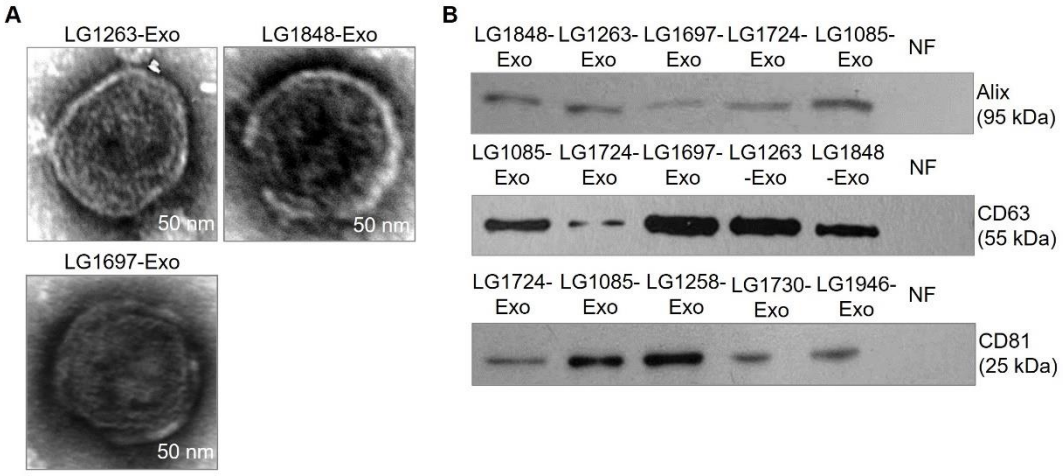


Figure 4.5 Characterization of patients' MPE exosomes. (A) TEM characterization of three example patients' exosome samples. (B) Western immunoblotting was used to probe for exosome markers.

It is well known that advanced stages of cancer in general and NSCLC (the majority of the MPE samples used in the current study) in specific are associated with significantly elevated presence of lymphocytes. Consequently, the possibility of lymphocyte derived exosomes dominating the exosome population in MPE must be considered. To this end, the single-exosome

characterization capability of our SERS-based biomolecular fingerprinting technique possesses unique advantages. It allows us to characterize exosomes individually, allowing for the premalignant exosomal biomarkers to not be masked by other exosome subpopulations. Extracting the unique SERS spectral signatures from HBEC-derived exosomes and combining these with our spectral data analysis algorithms enables the exclusive identification of premalignant exosomal SERS spectral features.

Our experimental studies shown in the next section indicate the power of such combination of biomolecular fingerprinting combined with machine learning. We examined exosomes from MPE of other metastatic cancer patients, i.e., individuals diagnosed with cancers of types other than NCSLC including small cell lung cancer (SCLC), tongue cancer, mesothelioma, as well as healthy human serum to assess the specificity of our assay. These experimental studies aim to determine whether our SERS-based platform could discriminate NSCLC-related exosomes in the presence of exosomes released by other types of cells.

SERS spectral datasets were subsequently collected for further analyses. A spectral signature matching program was used to query the spectral dataset for the HM and UN exosomal spectral signatures, shown in Figure 4.8. The nearest neighbors-based algorithm (Taunk et al., 2019) was used to identify the MPE/serum exosomes that have the same signature as the HBECs-derived exosomes. Specifically, the class of spectrum *i* in the patient-derived exosome dataset is determined by

$$C_i = \begin{cases} C_{argmin_{j;j\in HBECs}\, Sim(i;j)}, & Sim(i;j) < T \\ null, & else \end{cases} \qquad 4\text{-}1$$

Where the similarity score *Sim(i;j)* is given by modified Euclidean distance

$$Sim(i;j) = \sum_{offset=-t}^{t} \sqrt{\sum_{i=0}^{Raman\ shifts} (X_{i+t} - Y_i)^2} \qquad 4\text{-}2$$

X and *Y* are two different spectra, *t* is the spectral shifting *offset*. *T* is the threshold factor. The introduction of *t* is to neutralize the random errors during measurement.

The similarity between HBEC cell-derived exosomal spectrum and MPE-derived exosome spectrum was quantified. Successful matching was determined if the similarity score was within the predefined threshold. Modified Euclidean distance was used for calculating the similarity score, details are given in the Methods section. Figure 4.8B and 4.8C demonstrate the successful spectral matching of both the HBEC-HM-derived and HBEC-UN-derived exosomal signatures with MPE-derived signatures. Most spectral peaks belonging to the HBEC cell-derived exosomes and exosomes isolated from the patient MPE are colocalized, including the peak location, intensity, and width.

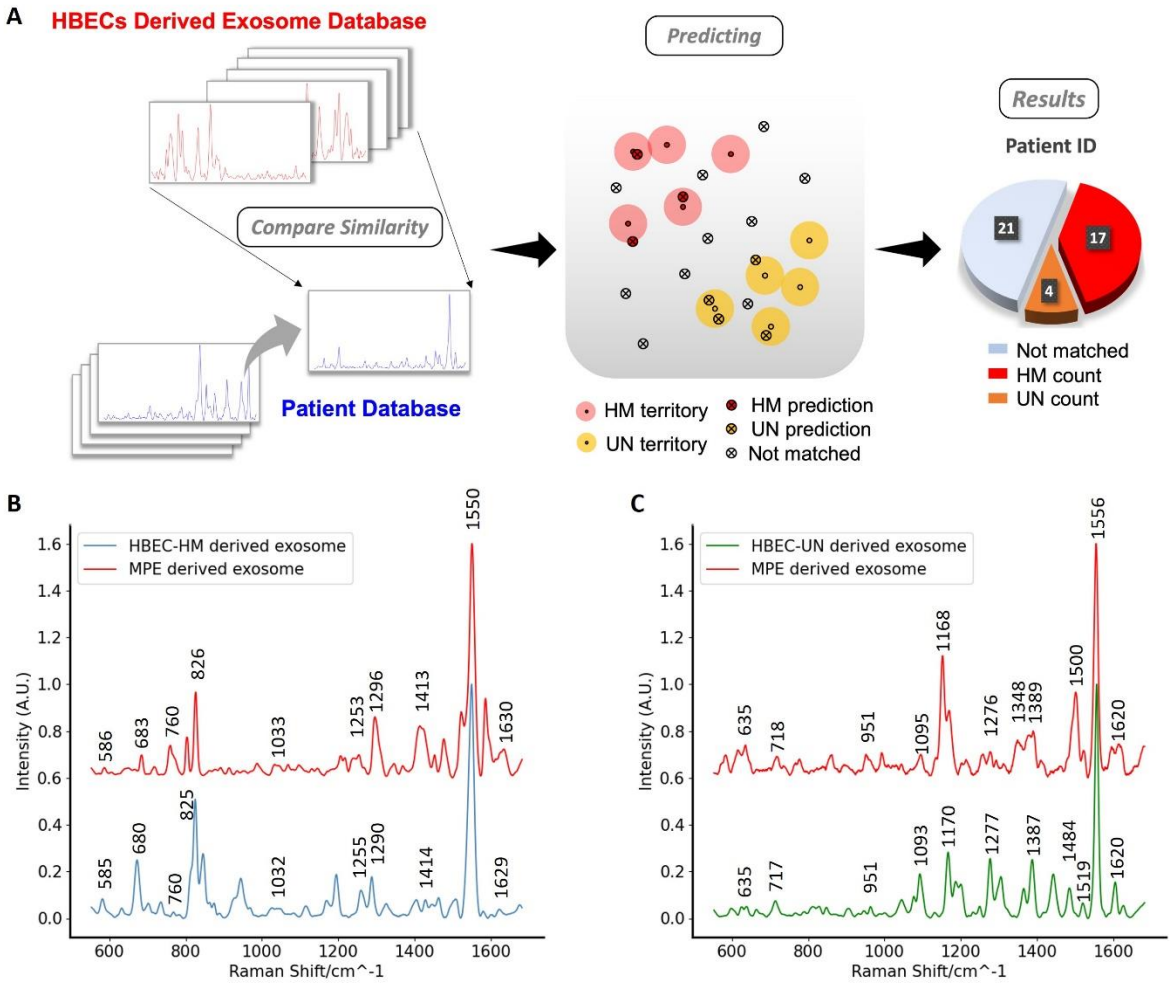


Figure 4.6 Spectral matching procedure. (A) Schematic procedure of Nearest Neighbors-based algorithm for identifying exosomal spectroscopic signatures. (B) Example of matched spectroscopic signatures between HBEC-HM exosomes and patient's MPE-derived exosomes (peak assignments are given in Table S1). (C) Example of matched spectroscopic signatures between HBEC-UN exosomes and patient's MPE-derived exosomes.

Table 4.2 and Figure 4.9 show the characteristic exosome counts in patients' samples. NSCLC patients show enrichment with highly migratory exosomal signatures, with an average of 3.4 exosomes per individual; Nearly no highly migratory exosomal signatures were found within patients' MPE with other metastasis, except only one HBEC-M-S-HM cell derived exosome is

found in patient LG1863, which is recognized as an outlier according to the box plot in figure 4.10B; Highly migratory exosomal signature does not exist in healthy human serum, indicating the high specificity of our exosomal biomarker based on single-exosome-fingerprinting technique. Compared to HBEC-UN derived exosomes, HBEC-HM-derived exosomes demonstrated more informative correlations among the three specimen groups-more explicit separations of three groups in terms of HBEC-HM-derived exosome count, which the counts of HBEC-UN derived exosomes haven't show meaningful correlations. We then evaluated the results by investigating the distribution counts for each group and the diagnostic ability by plotting receiver operating characteristics (ROC) curves, as shown in Figure 4.10. Figure 4.10A and Figure 4.10B show a clear distinguishment of NSCLC versus the other two groups. Focusing on diagnostic capability evaluation, areas under the curve (AUC) of 0.98, 1.00 were obtained for distinguishing NSCLC versus other metastasis, NSCLC versus healthy human serum respectively. Those results indicate that patients diagnosed with NSCLC possess clearly upregulated exosomes having highly similar signatures with HBEC-HM derived exosomes. One-group-versus-rest ROC curves (Figure 4.10E) were plotted as well for analyzing the prominence of a single group. ROC curve with 0.99 AUC was achieved after combining other metastasis and healthy human serum to a single group, and similar conclusions can be derived.

The limited number of characterized exosomes introduces statistical uncertainty that cannot be disregarded, hindering the ability to draw definitive conclusions. A larger sample dataset is required to further investigate the correlation between the number of HBEC-HM-derived exosomes and the severity of NSCLC. We are developing an automatic SERS characterization of single vesicle by modifying the Raman spectrometer controlling software, a higher throughput of spectral data (approximately increased by 10 times) is expected to be realized in the near future.

Consequently, our preliminary results indicate the presence of lung premalignant high migratory exosomal biomarkers (micrometastatic signatures) in patients' body fluid (metastatic disease) and could help indicate early NSCLC development.

Table 4.2 Counts of matched exosomal signatures.

Patient label	**Type of Cancer**	**Unselected**			**High Migratory**		
		M-V	M-S	Total	M-V	M-S	Total
LG1946	NSCLC	1	0	1	1	1	2
LG1730	NSCLC	0	0	0	0	1	1
LG1848	NSCLC	2	0	2	1	4	5
LG1263	NSCLC	0	0	0	2	2	4
LG1724	NSCLC	1	0	1	2	3	5
LG1697	NSCLC	1	1	2	0	3	3
LG1014	NSCLC	0	0	0	1	1	2
LG1014_A5	NSCLC	1	1	2	1	2	3
LC3	NSCLC	0	1	0	1	1	2
LG0782	NSCLC	1	0	1	3	0	3
LG1028	NSCLC	1	0	1	3	4	7
LC17	NSCLC	0	1	1	3	2	5
LG1085	NSCLC	1	1	0	0	1	1
LG1863	SCLC	0	1	1	1	0	1
LG1384	SCLC	0	0	0	0	0	0

LG1258	Mesothelioma	0	1	1	0	0	0
LC14	SCLC	0	1	1	0	0	0
LC5	Tongue cancer	0	0	0	0	0	0
CTRL01	Healthy human serum	0	0	0	0	0	0
CTRL02	Healthy human serum	0	0	0	0	0	0
CTRL03	Healthy human serum	1	0	1	0	0	0
CTRL04	Healthy human serum	0	2	2	0	0	0
CTRL05	Healthy human serum	0	0	0	0	0	0

Note: Patients suffering from different types of cancer are labeled with different colors (NSCLC Adenocarcinoma Stage IV with red, NSCLC with green, and other types with blue). NSCLC, Stage IV stands for NSCLC, Adenocarcinoma Stage IV; SCLC stands for Small-cell Lung Cancer; Total number of HBEC-UN and HBEC-HM derived exosomes are marked with yellow and red if detected.

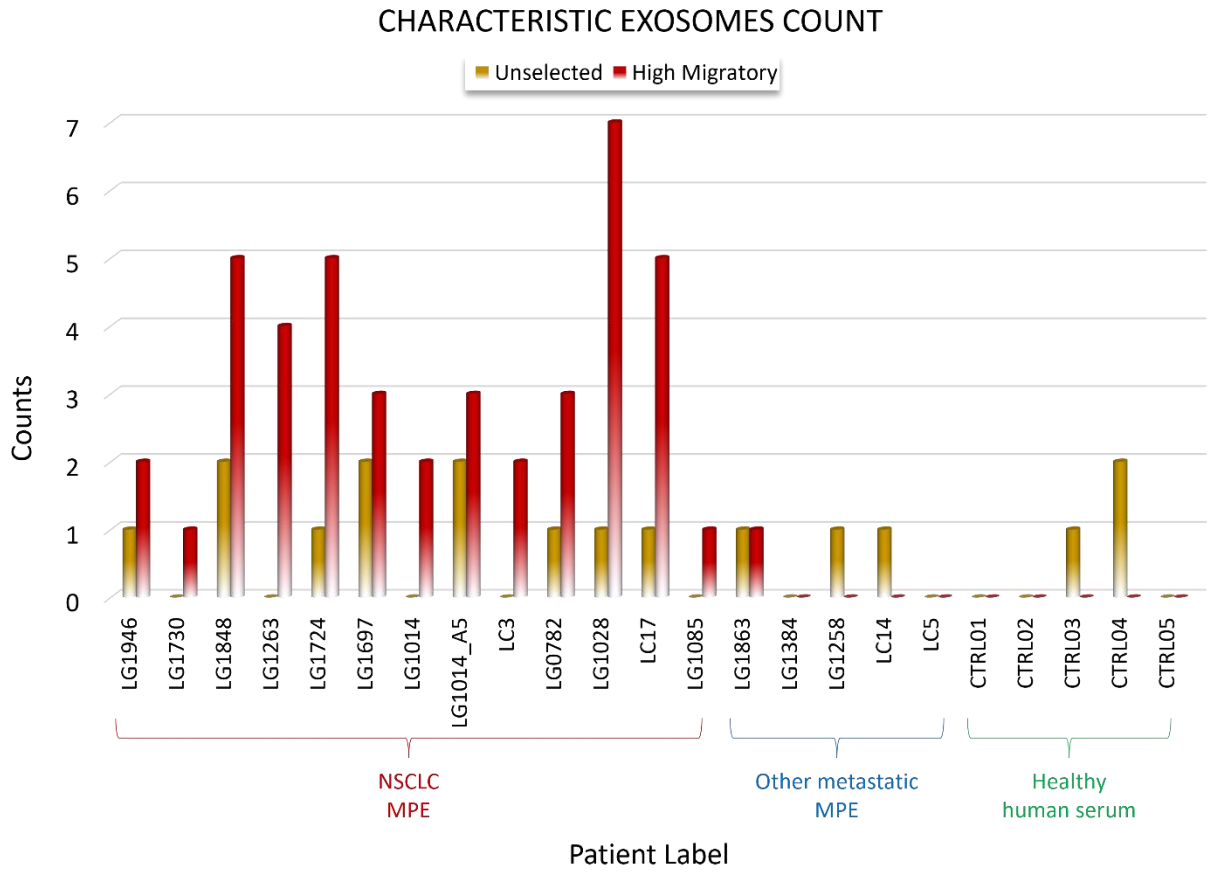
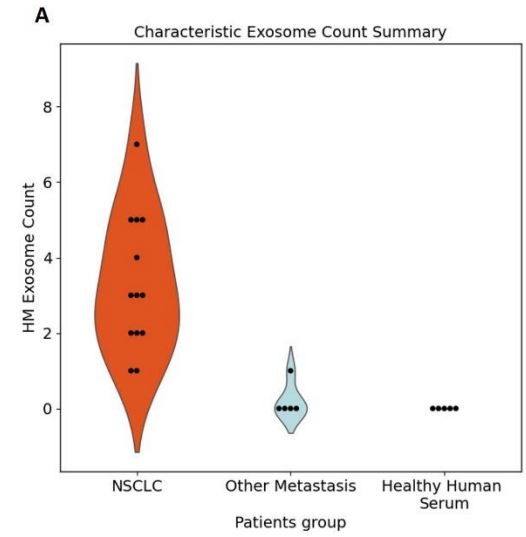
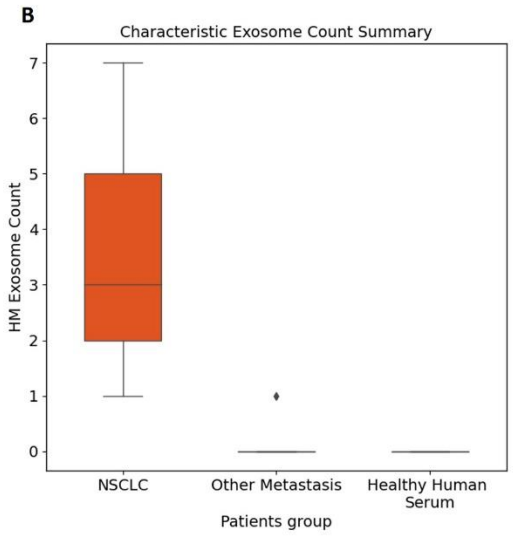
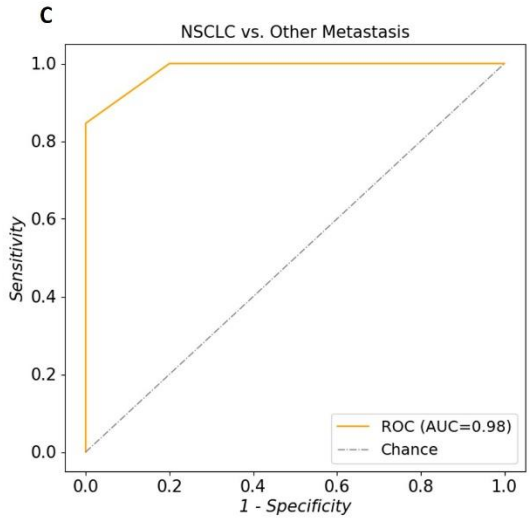
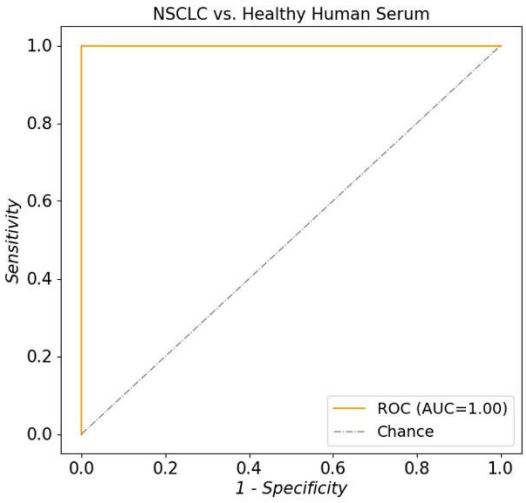


Figure 4.7 Summary of characteristic exosome counts.


A
Characteristic Exosome Count Summary
HM Exosome Count
0
2
4
6
8
NSCLC
Other Metastasis
Healthy Human Serum
Patients group


B
Characteristic Exosome Count Summary
HM Exosome Count
0
1
2
3
4
5
6
7
NSCLC
Other Metastasis
Healthy Human Serum
Patients group


C
NSCLC vs. Other Metastasis
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
ROC (AUC=0.98)
Chance
1 - Specificity


NSCLC vs. Healthy Human Serum
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
ROC (AUC=1.00)
Chance
1 - Specificity

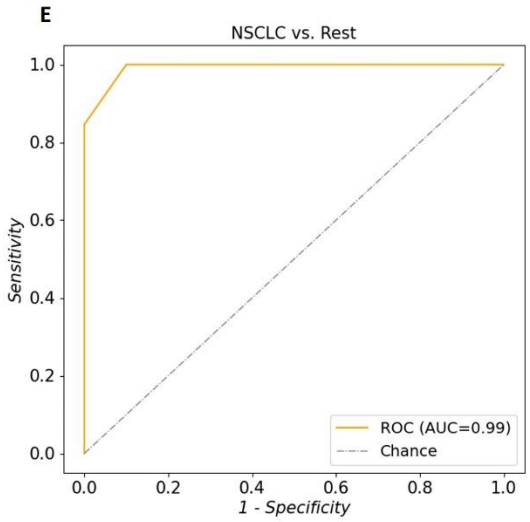

E
NSCLC vs. Rest
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
ROC (AUC=0.99)
Chance
1 - Specificity

Figure 4.8 Summary and evaluation of NSCLC characteristic exosome identification. (A) Violin plot of characteristic exosome count shown in chart 1. (B) Box plot of characteristic exosome count shown in chart 1. (C) ROC curve of binary classification between groups NSCLC and the rest. (D) ROC curve of binary classification between groups NSCLC and other metastasis. (E) ROC curve of binary classification between groups NSCLC and healthy human serum.

4.4 Conclusion

Our results highlight the promising feasibility of HBEC-HM (early metastatic) cell line-derived exosomes possessing potential biomarker signatures corresponding to metastatic disease in NSCLC. The differences in SERS signatures amongst the HBEC-M-HM and other group-derived exosomes and the existence of HBEC-M-HM cell-derived exosomal SERS signature in multiple advanced-stage NSCLC patients MPE supports our assumptions. Though highly promising, to further establish the importance of label-free single exosomal SERS-based biomarkers for clinical application in the early detection and interception of metastatic lung cancer, additional controlled studies designed to exclude irrelevant biological variations and uncertain statistical fluctuations need to be performed. Exosomes inherit biological features reminiscent of their parent cells, and therefore confounding factors, including sex, age, personal habits, other diseases, etc., might add uniqueness to the exosomal spectroscopic features. Besides, a limited sample size could also lead to ambiguous conclusions. Therefore, after establishing the preliminary feasibility, we plan to carry out repetitive experiments to evaluate our assumptions. In the future, more patient and healthy control samples will be characterized to further investigate and alleviate the effects of confounding factors.

LDA was implemented to visualize the spectroscopic signature differences among four exosome groups. We did find unique features of HBEC-HM cells derived exosomes versus HBEC-

UN cells derived exosomes. Even within the HBEC-HM cells, Snail-modified cell-derived exosomes show more uniqueness than Snail-unmodified. However, Snail overexpression did not change much on HBEC-UN cells derived exosomes. This observation implies the presence of non-metastatic premalignant cell subtypes and their inherent limited ability to metastasize, even after Snail overexpression. Based on our assumption that cells migrating through the transwell in the "constricted migration" process have unique features and are relevant to cancer propagation, HBEC-M-V-HM and HBEC-M-S-HM derived exosomal spectral signatures were combined to form the HBEC-HM group and compared to the HBEC-UN group. ACOFS ran a heuristic process to select the features on which the HBEC-HM group differs from HBEC-UN, which means the results are not decisive due to the "probabilistic moving" procedure in ACOFS. Therefore, a large number of feature selection rounds and a reasonable performance metric are critical components for increasing the reliability of the selected features. We ran three rounds of feature selection followed by averaging based on our dataset size and found negligible changes with increased rounds. Original features (1011) were reduced to approximately 380 according to the general optimal ratio between the number of features and the number of samples (spectra) (Hua et al., 2005). The top 20 features ranging from 1231 to 1610 were presented and given the molecular assignments for more explicit results demonstration. Current molecular information of Raman peaks are mostly the bonding within amino acids, nucleic acids, amide, lipids, etc. (Talari et al., 2015); no direct link to the proteomic/genomic compositions such as structural and functional proteins is established. Proteomics and genomics characterizations are essential to elucidate the exosomal biomarkers further and validate the molecular information agreements with SERS signatures. Accordingly, we have been planning Mass Spectrometry characterization on the exosome samples to extract the proteomic compositions and compare with spectral data.

We subsequently designed a nearest neighbors-based spectral matching algorithm to identify similar exosomal spectroscopic signatures, in which the number of neighbors is set to one and similarity metrics is modified Euclidean distance based on selected feature subset by ACOFS. A threshold of 0.07 was chosen empirically to determine the boundary between similar and dissimilar patterns, mainly based on the peaks' positions and intensities. We found that the spectra below 0.07 show reasonable overlapping peaks (shown in Figure 4.8B and 4.8C), and the ones over 0.08 start demonstrating visual differences. Therefore, we chose a stringent threshold to make our spectral matching more reliable. Nevertheless, the optimal algorithm and the threshold are supposed to be ultimately determined by well-designed cross-validations or even blind tests. Given our preliminary findings, we have planned studies with a large sample set and blind tests, as stated before. A standard pattern identification system is expected to be established. Preliminary informative correlations were found in the HBEC-HM-derived exosome signature counts in clinical samples, at the same time, larger datasets are being collected to render more evident conclusions.

This study demonstrates how single-exosome label-free spectroscopic signatures based on SERS and subsequent analyses supported by machine learning may predict early migratory phenotypes and aid in the early-stage detection, diagnosis, and interception of metastatic lung cancer. The technological platform comprises of an Au-graphene-plasmonic hybrid substrate for the increased collection of Raman spectra from individual exosomes. Single exosome SERS signatures were collected, characterized, and information-rich SERS spectra were examined using Nearest Neighbors algorithm to identify unique molecular "fingerprints." These fingerprints include abundant biological data and serve as the basis for this novel liquid biopsy approach. By refining the SERS platform fabrication protocol to increase the homogeneity of the pyramidal gold

patterns and introducing more powerful machine learning algorithms (neural networks etc.) persistent to statistical fluctuations, we will be able to make our biomarker fingerprints more robust and reliable.

Numerous peer-reviewed studies have shown that our proposed platform has an exceptionally high label-free specificity, regardless of the biological variability inherent in the patient data. As a result, we have developed a minimally invasive liquid biopsy approach with a single exosome detection capability, which is a significant advancement in the field of liquid biopsy for early cancer detection. This platform can become more efficient by introducing high-speed SERS single-particle data acquisition. Furthermore, by identifying human EVs/exosomes in MPE, we illustrate the clinical potential of our method. Using this research as a foundation, we anticipate that this method will provide precise measurement of unique EV subpopulations for extensive biological applications. To extend the span of our protocol, a less invasive patient specimen acquisition method can be introduced instead of extracting patients' MPE for detection. EVs are reported to circulate around human body fluid, which theoretically provides the possibility of using common medical test specimens, such as blood, plasma, serum, or sputum. This study raises a promising strategy to clinically detect NSCLC early metastasis and may be extended to other types of cancers (gastric cancer, breast cancer) with considerable enhancements implemented in the future.

4.5 References

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. Bioinformatics, 26(10), 1340–1347.

Carmicheal, J., Hayashi, C., Huang, X., Liu, L., Lu, Y., Krasnoslobodtsev, A., Lushnikov, A., Kshirsagar, P. G., Patel, A., Jain, M., Lyubchenko, Y. L., Lu, Y., Batra, S. K., & Kaur, S. (2019). Label-free characterization of exosome via surface enhanced Raman spectroscopy for the early detection of pancreatic cancer. Nanomedicine: Nanotechnology, Biology and Medicine, 16, 88–96.

Cheng, W.-T., Liu, M.-T., Liu, H.-N., & Lin, S.-Y. (2005). Micro-Raman spectroscopy used to identify and grade human skin pilomatrixoma. Microscopy Research and Technique, 68(2), 75–79.

Člupek, M., Matějka, P., & Volka, K. (2007). Noise reduction in Raman spectra: Finite impulse response filtration versus Savitzky-Golay smoothing. Journal of Raman Spectroscopy, 38(9), 1174–1179.

Dong, S., Wang, Y., Liu, Z., Zhang, W., Yi, K., Zhang, X., Zhang, X., Jiang, C., Yang, S., Wang, F., & Xiao, X. (2020). Beehive-Inspired Macroporous SERS Probe for Cancer Detection through Capturing and Analyzing Exosomes in Plasma. ACS Applied Materials & Interfaces, 12(4), 5136–5146.

Dorigo, M., Birattari, M., & Stutzle, T. (2006). Ant colony optimization. IEEE Computational Intelligence Magazine, 1(4), 28–39.

Eyles, J., Puaux, A.-L., Wang, X., Toh, B., Prakash, C., Hong, M., Tan, T. G., Zheng, L., Ong, L. C., Jin, Y., Kato, M., Prévost-Blondel, A., Chow, P., Yang, H., & Abastado, J.-P. (2010). Tumor cells disseminate early, but immunosurveillance limits metastatic outgrowth, in a mouse model of melanoma. Journal of Clinical Investigation, 120(6), 2030–2039.

Fontebasso, Y., & Dubinett, S. M. (2015). Drug Development for Metastasis Prevention. Critical Reviews in Oncogenesis, 20(5–6), 449–473.

Fraire, J. C., Stremersch, S., Bouckaert, D., Monteyne, T., De Beer, T., Wuytens, P., De Rycke, R., Skirtach, A. G., Raemdonck, K., De Smedt, S., & Braeckmans, K. (2019). Improved Label-Free Identification of Individual Exosome-like Vesicles with Au@Ag Nanoparticles as SERS Substrate. ACS Applied Materials & Interfaces, 11(43), 39424–39435.

Grant, J. L., Fishbein, M. C., Hong, L.-S., Krysan, K., Minna, J. D., Shay, J. W., Walser, T. C., & Dubinett, S. M. (2014). A Novel Molecular Pathway for Snail-Dependent, SPARC-Mediated Invasion in Non–Small Cell Lung Cancer Pathogenesis. Cancer Prevention Research, 7(1), 150–160.

Greenberg, J. W., Kim, H., Moustafa, A. A., Datta, A., Barata, P. C., Boulares, A. H., Abdel-Mageed, A. B., & Krane, L. S. (2021). Repurposing ketoconazole as an exosome directed adjunct to sunitinib in treating renal cell carcinoma. Scientific Reports, 11(1), 10200.

Guerrini, L., Garcia-Rico, E., O'Loghlen, A., Giannini, V., & Alvarez-Puebla, R. A. (2021). Surface-Enhanced Raman Scattering (SERS) Spectroscopy for Sensing and Characterization of Exosomes in Cancer Diagnosis. Cancers, 13(9), 2179.

Hsu, Y.-L., Hung, J.-Y., Chang, W.-A., Lin, Y.-S., Pan, Y.-C., Tsai, P.-H., Wu, C.-Y., & Kuo, P.-L. (2017). Hypoxic lung cancer-secreted exosomal miR-23a increased angiogenesis and vascular permeability by targeting prolyl hydroxylase and tight junction protein ZO-1. Oncogene, 36(34), 4929–4942.

Hua, J., Xiong, Z., Lowey, J., Suh, E., & Dougherty, E. R. (2005). Optimal number of features as a function of sample size for various classification rules. Bioinformatics, 21(8), 1509–1515.

Hüsemann, Y., Geigl, J. B., Schubert, F., Musiani, P., Meyer, M., Burghart, E., Forni, G., Eils, R., Fehm, T., Riethmüller, G., & Klein, C. A. (2008). Systemic Spread Is an Early Step in Breast Cancer. Cancer Cell, 13(1), 58–68.

Hüyük, M., Fiocco, M., Postmus, P. E., Cohen, D., & Von Der Thüsen, J. H. (2023). Systematic review and meta-analysis of the prognostic impact of lymph node micrometastasis and isolated tumour cells in patients with stage I–IIIA non-small cell lung cancer. Histopathology, 82(5), 650–663.

K. Paul, M. (Ed.). (2022). Extracellular Vesicles—Role in Diseases, Pathogenesis and Therapy (Vol. 13). IntechOpen.

Kawano, R., Hata, E., Ikeda, S., & Sakaguchi, H. (2002). Micrometastasis to lymph nodes in stage I left lung cancer patients. The Annals of Thoracic Surgery, 73(5), 1558–1562.

Kruglik, S. G., Royo, F., Guigner, J.-M., Palomo, L., Seksek, O., Turpin, P.-Y., Tatischeff, I., & Falcón-Pérez, J. M. (2019). Raman tweezers microspectroscopy of circa 100 nm extracellular vesicles. Nanoscale, 11(4), 1661–1679.

Lee, W., Nanou, A., Rikkert, L., Coumans, F. A. W., Otto, C., Terstappen, L. W. M. M., & Offerhaus, H. L. (2018). Label-Free Prostate Cancer Detection by Characterization of Extracellular Vesicles Using Raman Spectroscopy. Analytical Chemistry, 90(19), 11290–11296.

Leng, H., Chen, C., Chen, C., Chen, F., Du, Z., Chen, J., Yang, B., Zuo, E., Xiao, M., Lv, X., & Liu, P. (2023). Raman spectroscopy and FTIR spectroscopy fusion technology combined with deep learning: A novel cancer prediction method. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 285, 121839.

Li, J., Li, Y., Li, P., Zhang, Y., Du, L., Wang, Y., Zhang, C., & Wang, C. (2022). Exosome detection via surface-enhanced Raman spectroscopy for cancer diagnosis. Acta Biomaterialia, 144, 1–14.

Liu, Z., Li, T., Wang, Z., Liu, J., Huang, S., Min, B. H., An, J. Y., Kim, K. M., Kim, S., Chen, Y., Liu, H., Kim, Y., Wong, D. T. W., Huang, T. J., & Xie, Y.-H. (2022). Gold Nanopyramid Arrays for Non-Invasive Surface-Enhanced Raman Spectroscopy-Based Gastric Cancer Detection via sEVs. ACS Applied Nano Materials, 5(9), 12506–12517.

Maas, S. L. N., De Vrij, J., & Broekman, M. L. D. (2014). Quantification and Size-profiling of Extracellular Vesicles Using Tunable Resistive Pulse Sensing. Journal of Visualized Experiments, 92, 51623.

Mukherjee, A., Bisht, B., Dutta, S., & Paul, M. K. (2022). Current advances in the use of exosomes, liposomes, and bioengineered hybrid nanovesicles in cancer detection and therapy. Acta Pharmacologica Sinica, 43(11), 2759–2776.

Nalepa, J., & Kawulok, M. (2019). Selecting training sets for support vector machines: A review. Artificial Intelligence Review, 52(2), 857–900.

Pagano, P. C., Tran, L. M., Bendris, N., O'Byrne, S., Tse, H. T., Sharma, S., Hoech, J. W., Park, S. J., Liclican, E. L., Jing, Z., Li, R., Krysan, K., Paul, M. K., Fontebasso, Y., Larsen, J. E., Hakimi, S., Seki, A., Fishbein, M. C., Gimzewski, J. K., … Dubinett, S. M. (2017). Identification of a Human Airway Epithelial Cell Subpopulation with Altered Biophysical, Molecular, and Metastatic Properties. Cancer Prevention Research, 10(9), 514–524.

Palermo, G., Rippa, M., Conti, Y., Vestri, A., Castagna, R., Fusco, G., Suffredini, E., Zhou, J., Zyss, J., De Luca, A., & Petti, L. (2021). Plasmonic Metasurfaces Based on Pyramidal Nanoholes for High-Efficiency SERS Biosensing. ACS Applied Materials & Interfaces, 13(36), 43715–43725.

Pan, R., Liu, J., Wang, P., Wu, D., Chen, J., Wu, Y., & Li, G. (2022). Ultrasensitive CRISPR/Cas12a-Driven SERS Biosensor for On-Site Nucleic Acid Detection and Its Application to Milk Authenticity Testing. Journal of Agricultural and Food Chemistry, 70(14), 4484–4491.

Pang, Y., Wan, N., Shi, L., Wang, C., Sun, Z., Xiao, R., & Wang, S. (2019). Dual-recognition surface-enhanced Raman scattering(SERS)biosensor for pathogenic bacteria detection by using vancomycin-SERS tags and aptamer-Fe3O4@Au. Analytica Chimica Acta, 1077, 288–296.

Paul, M. K., Bisht, B., Darmawan, D. O., Chiou, R., Ha, V. L., Wallace, W. D., Chon, A. T., Hegab, A. E., Grogan, T., Elashoff, D. A., Alva-Ornelas, J. A., & Gomperts, B. N. (2014). Dynamic Changes in Intracellular ROS Levels Regulate Airway Basal Stem Cell Homeostasis through Nrf2-Dependent Notch Signaling. Cell Stem Cell, 15(2), 199–214.

Paul, M. K., Dutta, S., Bisht, B., Ramin, S.-R., Pagano, P., Bitan, G., Minna, J. D., & Dubinett, S. M. (2018). Abstract 2015: Exosomes secreted by highly migratory premalignant lung epithelial cells promote epithelial mesenchymal transition and migration. Cancer Research, 78(13_Supplement), 2015–2015.

Peng, H., Ying, C., Tan, S., Hu, B., & Sun, Z. (2018). An Improved Feature Selection Algorithm Based on Ant Colony Optimization. IEEE Access, 6, 69203–69209.

Podsypanina, K., Du, Y.-C. N., Jechlinger, M., Beverly, L. J., Hambardzumyan, D., & Varmus, H. (2008). Seeding and Propagation of Untransformed Mouse Mammary Cells in the Lung. Science, 321(5897), 1841–1844.

Ramirez, R. D., Sheridan, S., Girard, L., Sato, M., Kim, Y., Pollack, J., Peyton, M., Zou, Y., Kurie, J. M., DiMaio, J. M., Milchgrub, S., Smith, A. L., Souza, R. F., Gilbey, L., Zhang, X., Gandia, K., Vaughan, M. B., Wright, W. E., Gazdar, A. F., … Minna, J. D. (2004). Immortalization of Human Bronchial Epithelial Cells in the Absence of Viral Oncoproteins. Cancer Research, 64(24), 9027–9034.

Rhim, A. D., Mirek, E. T., Aiello, N. M., Maitra, A., Bailey, J. M., McAllister, F., Reichert, M., Beatty, G. L., Rustgi, A. K., Vonderheide, R. H., Leach, S. D., & Stanger, B. Z. (2012). EMT and Dissemination Precede Pancreatic Tumor Formation. Cell, 148(1–2), 349–361.

S. Chauhan, D., Mudaliar, P., Basu, S., Aich, J., & K. Paul, M. (2022). Tumor-Derived Exosome and Immune Modulation. In M. K. Paul (Ed.), Physiology (Vol. 13). IntechOpen.

Salehi-Rad, R., Li, R., Paul, M. K., Dubinett, S. M., & Liu, B. (2020). The Biology of Lung Cancer. Clinics in Chest Medicine, 41(1), 25–38.

Sato, H., Ishigaki, M., Taketani, A., & Andriana, B. B. (2019). Raman spectroscopy and its use for live cell and tissue analysis. Biomedical Spectroscopy and Imaging, 7(3–4), 97–104.

Sato, M., Larsen, J. E., Lee, W., Sun, H., Shames, D. S., Dalvi, M. P., Ramirez, R. D., Tang, H., DiMaio, J. M., Gao, B., Xie, Y., Wistuba, I. I., Gazdar, A. F., Shay, J. W., & Minna, J. D. (2013). Human Lung Epithelial Cells Progressed to Malignancy through Specific Oncogenic Manipulations. Molecular Cancer Research, 11(6), 638–650.

Shimada, Y., & Minna, J. D. (2017). Exosome mediated phenotypic changes in lung cancer pathophysiology. Translational Cancer Research, 6(S6), S1040–S1042.

Shin, H., Jeong, H., Park, J., Hong, S., & Choi, Y. (2018). Correlation between Cancerous Exosomes and Protein Markers Based on Surface-Enhanced Raman Spectroscopy (SERS) and Principal Component Analysis (PCA). ACS Sensors, 3(12), 2637–2643.

Talari, A. C. S., Movasaghi, Z., Rehman, S., & Rehman, I. U. (2015). Raman Spectroscopy of Biological Tissues. Applied Spectroscopy Reviews, 50(1), 46–111.

Tang, W.-F., Wu, M., Bao, H., Xu, Y., Lin, J.-S., Liang, Y., Zhang, Y., Chu, X.-P., Qiu, Z.-B., Su, J., Zhang, J.-T., Zhang, C., Xu, F.-P., Chen, J.-H., Fu, R., Chen, Y., Yang, T., Chen, Q.-K., Wu, T.-T., … Zhong, W.-Z. (2021). Timing and Origins of Local and Distant Metastases in Lung Cancer. Journal of Thoracic Oncology, 16(7), 1136–1148.

Taunk, K., De, S., Verma, S., & Swetapadma, A. (2019). A Brief Review of Nearest Neighbor Algorithm for Learning and Classification. 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 1255–1260.

Wang, H., Zhang, S., Wan, L., Sun, H., Tan, J., & Su, Q. (2018). Screening and staging for non-small cell lung cancer by serum laser Raman spectroscopy. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 201, 34–38.

Wang, P., Liang, O., Zhang, W., Schroeder, T., & Xie, Y. (2013). Ultra‐Sensitive Graphene‐Plasmonic Hybrid Platform for Label‐Free Detection. Advanced Materials, 25(35), 4918–4924.

Wang, P., Xia, M., Liang, O., Sun, K., Cipriano, A. F., Schroeder, T., Liu, H., & Xie, Y.-H. (2015). Label-Free SERS Selective Detection of Dopamine and Serotonin Using Graphene-Au Nanopyramid Heterostructure. Analytical Chemistry, 87(20), 10255–10261.

Xanthopoulos, P., Pardalos, P. M., & Trafalis, T. B. (2013). Linear Discriminant Analysis. In P. Xanthopoulos, P. M. Pardalos, & T. B. Trafalis, Robust Data Mining (pp. 27–33). Springer New York.

Yan, Z., Dutta, S., Liu, Z., Yu, X., Mesgarzadeh, N., Ji, F., Bitan, G., & Xie, Y.-H. (2019). A Label-Free Platform for Identification of Exosomes from Different Sources. ACS Sensors, 4(2), 488–497.

Yáñez-Mó, M., Siljander, P. R. -M., Andreu, Z., Bedina Zavec, A., Borràs, F. E., Buzas, E. I., Buzas, K., Casal, E., Cappello, F., Carvalho, J., Colás, E., Cordeiro-da Silva, A., Fais, S., Falcon-Perez, J. M., Ghobrial, I. M., Giebel, B., Gimona, M., Graner, M., Gursel, I., … De Wever, O. (2015). Biological properties of extracellular vesicles and their physiological functions. Journal of Extracellular Vesicles, 4(1), 27066.

Yu, X., Hayden, E. Y., Wang, P., Xia, M., Liang, O., Bai, Y., Teplow, D. B., & Xie, Y. (2020). Ultrasensitive amyloid β-protein quantification with high dynamic range using a hybrid graphene–gold surface-enhanced Raman spectroscopy platform. Journal of Raman Spectroscopy, 51(3), 432–441.

Chapter 5 Saliva-based COVID-2019 Detection by SERS and SVMs

5.1 Introduction to SERS detection of COVID

Since the emergence of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) in December 2019, more than 620 million cases and 6 million deaths have been reported till November 2022, as declared by World Health Organization (WHO) (Allan et al., 2022). The typical symptoms include fever, fatigue, severe respiratory illness, pneumonia as well as dyspnea. Recently, long-term damage to brain and heart have also been reported (Adjei et al., 2022). More SARS-CoV-2 variants have been emerging globally, such as the ones in the United Kingdom (B.1.1.7), the United States (B.1.429, Washington, B.1.1.529 or Omicron and Omicron BA.2) and India (B.1.617.2 or Delta) causing more rapid and wider spread of the pandemic around the world (Vasireddy et al., 2021). Currently, the SARS-CoV-2 strain Omicron BA.5 makes up around 62% of the COVID cases (Grewal et al., 2022). Though the mortality of the more recent variants has been much lower than the original strains (Adjei et al., 2022), the transmissibility has significantly increased (Araf et al., 2022; Challen et al., 2021).

SARS-CoV-2 belongs to the family of coronavirus of 60-140nm in vesicle size. It is composed of single-strand RNA, lipid bilayer membrane and structural proteins (spike protein, envelop protein, membrane protein and nucleocapsid protein) (Vasireddy et al., 2021). Currently the prevalent diagnostic technologies are RT-PCR and antigen test, which detect the viral RNA and the protein biomarkers (e.g., spike protein) (Chau et al., 2020). As SARS-CoV-2 belongs to the family of the single-stranded RNA viruses, RT-PCR is the most widely used detection tool due to its high accuracy, sensitivity, and Limit of Detection (LoD). The LoD of around 100 particles/mL,

sensitivity above 80% and specificity above 95% have been reported (Chau et al., 2020; Y.-S. Chung et al., 2021). It is worth noting that there are drawbacks of RT-PCR preventing it from becoming the optimal diagnostic technology for targeting highly mutable and contagious viruses. For most of the nucleic acid-based tests, highly specific primers are required in the reverse transcription step, therefore specific new primers are needed to deal with the mutated variants (Freeman et al., 1999). RT-PCR is also extremely sensitive to the viral load of the samples thus the viral concentration fluctuation of Nasopharyngeal swab specimens or salivary specimens could result in false positive/negative cases (Tahamtan & Ardebili, 2020). Moreover, sophisticated equipment, costly regents as well as professional operators are required for collection and analysis, which inevitably increases the time and consumption cost. In contrast, the faster test tool, antigen test, could generate results in 15-30 minutes. However, it is less reliable due to worse sensitivity and specificity (around 50% and 90%, respectively) (Yamayoshi et al., 2020). Fast, accurate and non-invasive detection tools are still needed to monitor the pandemic and potentially identify other highly infectious viruses in the future. In this report, we present the feasibility of applying SERS for rapid identification of viruses. A schematic procedure is provided in Figure 5.1.

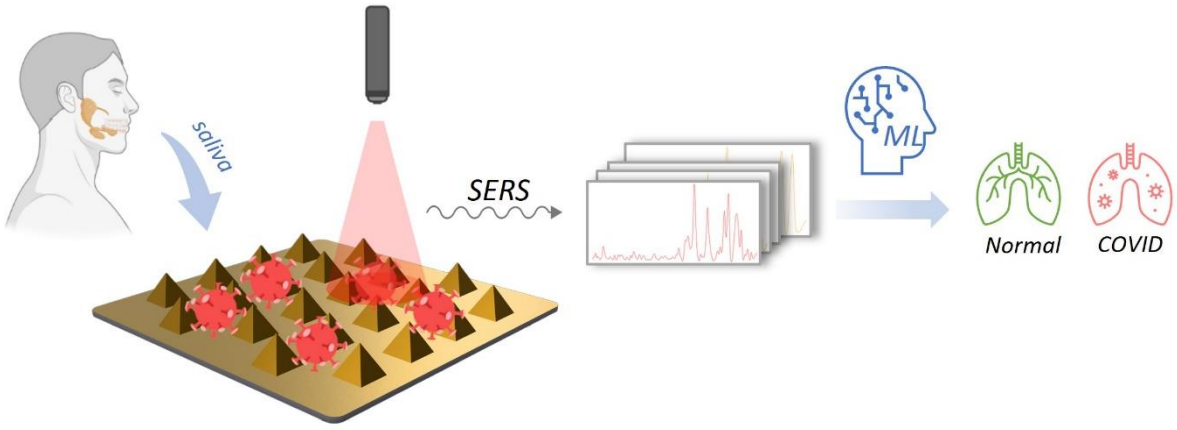


Figure 5.1 Schematic of SERS-based biosensing platform for virus detection.

Compared to antigen tests, SERS extracts SARS-CoV-2 biomarkers from multiple components, including structural protein, lipid bilayer and RNA strand (Sharma et al., 2012). Hereby, SERS has the advantages of drawing a more thorough picture over antigen test. Unlike nucleic acid based detecting technologies, SERS does not require complicated primers and regents nor special specimen treatment, therefore the estimated cost per test would be lower. Besides, SERS specimens can be isolated from different biofluids such as saliva, serum, urine and bronchoalveolar fluid, allowing for simple and non-invasive sample harvesting. Furthermore, SERS characterization for each sample requires a maximum of 1 to 6 hours, which makes it a more feasible "rapid-testing" method for SARS-CoV-2 compared to RT-PCR (Y.-S. Chung et al., 2021; Sharma et al., 2012). The Label-free feature of SERS-based test also makes it more amenable to scale up and adapt to more SARS-CoV-2 variants study.

SERS-based detection has been implemented for COVID detection. Improved detecting efficiency and limit of detection have been reported with uniquely designed biosensor setup (H. Chen et al., 2021). To prepare highly concentrated virus samples for SERS characterization, Sequential centrifugation and filtration are typically applied to isolate viruses from cell culture media (Stelzer-Braid et al., 2020). It has been reported that exosomes have similar size and density as viruses (30-150nm, 1.08–1.19 g/ml) (Bar-On et al., 2020; P. Zhang et al., 2019), therefore It is inevitable to exclude exosomes during virus isolation, which could lead to confusion in fingerprinting viruses. To establish the genuine fingerprint, exosomes' signatures need to be subtracted during either sample preparation or data processing.

This section demonstrates the feasibility of our SERS and machine learning- based fingerprinting and signature identification platform as being a potentially accurate and rapid saliva-based SARS-CoV-2 detection technique that could replace the current antigen test as a pandemic

monitoring tool. Figure 5.2 demonstrates the basic workflow. Briefly, SARS-CoV-2 virus samples were compared with SARS-CoV-1 virus and Vero-TMPRSS2 cell line- derived exosome samples and were successfully identified with 80% accuracy. We subsequently evaluated the diagnostic capabilities by comparing SARS-CoV-2 spiked human salivary samples versus healthy control. 10 SARS-CoV-2 spiked human salivary samples and 10 healthy controls salivary samples were applied to build the identifier. 90% sensitivity and 80% specificity were achieved afterward in blind test with the 20 samples. Using the above identification model, 5 COVID patients versus 5 healthy controls saliva samples were tested, 9 out of the 10 individuals are identified correctly. Detailed estimation of the advances and theoretical analysis of the feasibility of our platform is also provided.

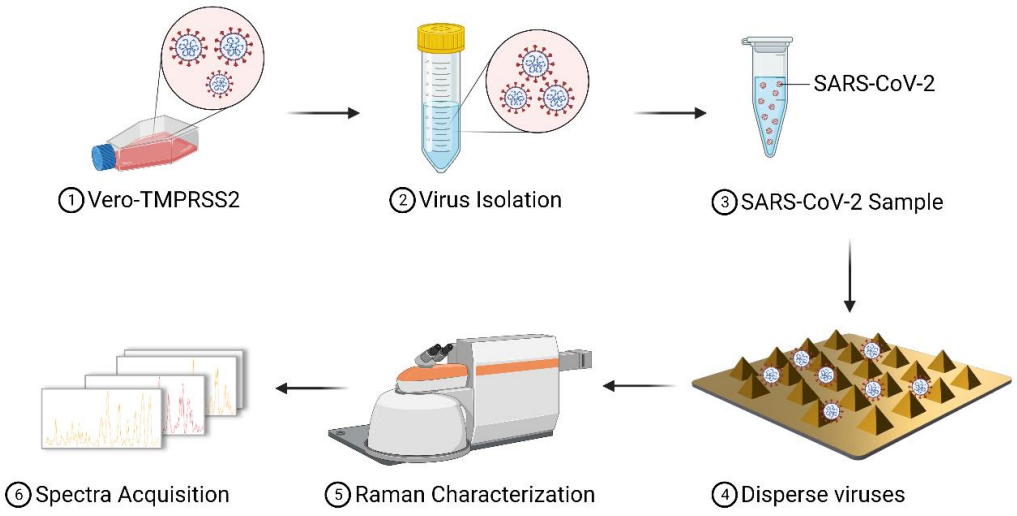


Figure 5.2 Schematic working flow of SERS characterization of SARS-CoV-2 specimens.

5.2 Methods and materials

5.2.1 Virus Sample preparation

The virus samples were produced, inactivated, and validated by the Institutional Biosafety Committee (IBC) for the University of California, San Diego following SARS-CoV-2 specimen

preparation(Carlin et al., 2023). All work with SARS-CoV-2 was conducted in biosafety level-3 conditions at the UCSD following the guidelines approved by the Institutional Biosafety Committee. Vero-TMPRSS2 cells are infected with viruses (either SARS-CoV-2 or SARS-CoV-2). Sequential centrifuge and filtration were used to isolate and purify the virus from cell culture media then the viruses were diluted in cell culture media (DMEM + 1% FBS + 10mM HEPES + 50 units/ml Penicillin and 50 µg/ml Streptomycin). Virus samples were then inactivated by heat (65°C for 30 minutes) (Pastorino et al., 2020) or UV (400 mJ/cm^2 delivered at UV 254nm) (Biasin et al., 2021). After inactivation, 10^8 to 10^{10} viruses per ml were estimated by ddPCR (RNA). Figure 5.3 shows a typical TEM (FEI TF20 High-resolution EM, USA) image of the specimen. Individual virus particles of about 50 nm diameter with the characteristic corona are clearly visible.

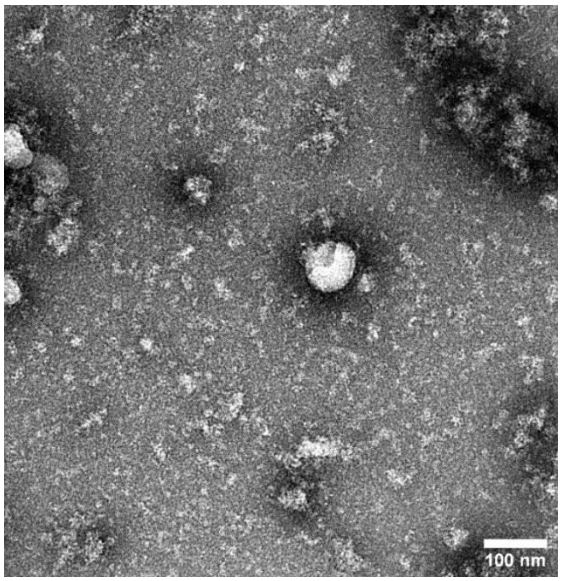


Figure 5.3 TEM image of SARS-CoV-2 specimen.

5.2.2 SARS-CoV-2 spiked human salivary samples preparation

The isolated and purified virus samples were used for preparing the SARS-CoV-2 spiked human salivary samples. The virus samples and salivary samples of healthy control were mixed

with the volume ratio that keeps the viral concentration around 10^8 particles/mL. Then the spiked salivary samples were aliquoted for multiple SERS testing.

5.2.3 SARS-CoV-2 clinical samples preparation

Archived saliva samples were obtained from an observational cohort study of hospitalized patients with COVID-19 from April 2020 until February 2021. The study was approved by the UCLA Institutional Review Board (#20-000473). Informed consent was obtained from all study participants. Patients with confirmed positive SARS-CoV-2 RT-PCR nasopharyngeal swabs were enrolled in an observational cohort study within 72 hours of admission. Exclusion criteria included pregnancy, hemoglobin < 8g/dL, or inability to provide informed consent. Blood specimens, nasopharyngeal swabs, and saliva were collected throughout hospitalization for up to 6 weeks. Demographic and clinical data, including laboratory results and therapeutics, were collected from the electronic medical records. Clinical severity was scored using the NIAID 8-point ordinal scale. A total of 10 samples were included in this study. Whole saliva was collected by passive drool into a cryovial. Samples were transported to the laboratory and immediately placed in -80 °C freezer for storage.

5.2.4 SERS characterization

SERS substrate fabrication follows the SOP described in Section 3.4.4. Spectra collection is similar to the single-particle-scanning methods stated in Section 3.5.1. For more details, a droplet of about 5 μL of the liquid sample was pipetted onto the surface of the SERS platform and dried under room ambient or in a vacuum desiccator typically within 15 minutes. The obtaining map yielded candidate spectra through which a spectra-selecting program traverses for establishing the spectral database. The rate of characterizing analytes is around 10-40 analytes/hour. According to our current spectral dataset size, approximately 1-6 hours are needed. As demonstrated by Figure

5.4, the spectra obtained have explicit Raman ranges with high signal-to-noise ratios. Peak assignments are given in Table S1.

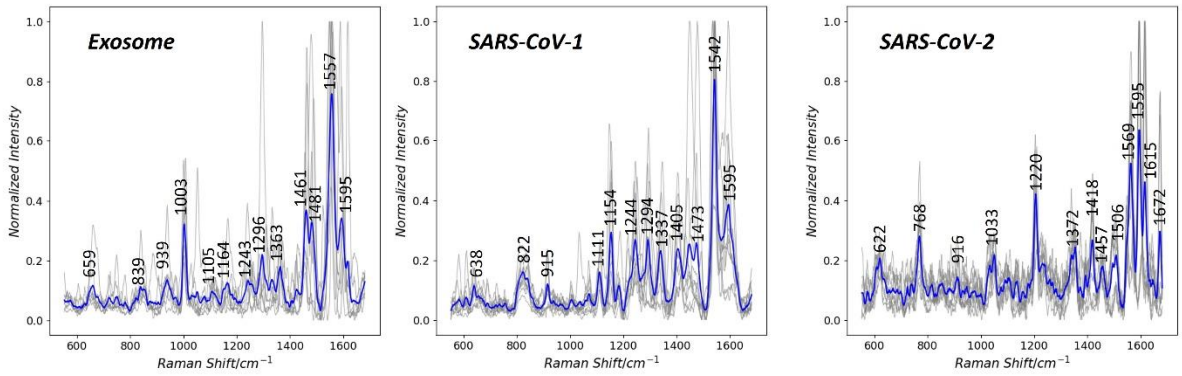


Figure 5.4 Spectra of Vero-TMPRSS2 exosome, SARS-CoV-1, SARS-CoV-2. Highly uniform spectra from the particles (gray lines) and averaged spectra (blue lines) demonstrate different patterns of different particles.

5.2.5 Method of spectral processing and data analysis

According to the preprocessing approaches stated in Section 3.6.1, Approximate 50 to 300 signal spots (depending on the particle concentration) were obtained for each sample to produce spectra that have 1023 Raman shifts in the range from 553 to 1581 cm^{-1}. Preprocessing steps are applied to alleviate the spectral signature fluctuations caused by sample variations, SERS platform heterogeneity, and instrument fluctuation. To elaborate, Fluorescence background subtraction and noise reduction are performed by batch processing based on asymmetric least square fitting (J. Peng et al., 2010) and Savitzky-Golay filtering (John et al., 2021), followed by min-max normalization that proportionally compresses the original intensity range to [0, 1]. A predictive model established by supervised learning or classification is the core of the proposed technology. It requires appropriate complexity of the classifier to prevent both underfitting and overfitting for the purpose of generalizing the characteristic signature effectively. We used SVM for the

classification tasks. Unsupervised learning or clustering analysis by HCA was also used as an auxiliary tool. Cross-validations are then applied to pre-evaluate our methodology given the labels and optimize the model settings, followed by tests for evaluating diagnostic capability. All the analyses are realized with Python using NumPy, SciPy and Scikit-learn modules.

5.3 Results

5.3.1 Single-vesicle techniques for viral detection

The single-vesicle detectability of SIM brings advantages in COVID detection. There are also several challenges originating from the working principle of single-vesicle detection. Most importantly, the feasibility of single-vesicle detection is determined by the standard signature of the target analyte (e.g., SARS-CoV-2) that we can refer to. The presence of EVs could potentially impact the procedure of obtaining the standard SERS spectral signature of SARS-CoV-2, as shown in Figure 5.5. The sample preparation step, the sample loading step, and the characterization step are all supposed to be conducted rigorously to prevent any possibility of contamination. The subsequent data processing step is also needed to get rid of irrelevant target analyte signatures. Secondly, though SERS dramatically increases the signal intensity of the analyte which facilitates much more sensitive detection, the inherent biological variabilities are also amplified. The signatures of SARS-CoV-2 from different SERS characterizations instances might fluctuate to some extent. Therefore, the intra-class (such as SARS-CoV-2) fluctuations versus the inter-class (such as SARS-CoV-2/EVs) differences must be validated to support the decision boundary. In addition, Single-vesicle characterization is usually performed in the manner of individual scanning, which greatly limits the data throughput. Much effort needs to be made to boost the data harvest rate and determine the characterization data size to make a sufficiently reliable diagnosis

conclusion. Due to the above concerns, we have performed the following experiments to establish the capability of SIM for SARS-CoV-2 detection.

5.3.2 Differentiation of SARS-CoV-2 versus SARS-CoV-1 virion in mixture of cell lysate

As a prerequisite step for establishing SIM identification of SARS-CoV-2 signature, we first evaluated the proposed platform in differentiating SARS-CoV-2 from other closely related virus types, including other types of virions and extracellular vesicles, of which the dimensions are close to the SARS-CoV-2 virus. SARS-CoV-1 is reported to share more than 70% genetic similarity with SARS-CoV-2 (Z. Cai et al., 2021), leading to highly similar structural components such as single-stranded RNA and spike protein, while the mutations make the latter less deadly but much more transmissible. With SARS-CoV-1 as a candidate, 10 SARS-CoV-1 specimens and 10 SARS-CoV-2 specimens were prepared and then characterized by SERS following our SERS map protocol. 50 to 70 spots rendering spectral signatures with high signal-to-noise ratio were collected for each sample, multiple spectra were saved per spot to account for the information of spectral intensity fluctuations, which allows for comprehensive training of the model by making it less sensitive to the slight changes.

In total, 1929 spectra from SARS-CoV-1 samples and 1559 from SARS-CoV-2 samples were recorded. Figure 5.4 are three examples of spectra set belonging to a single particle of Vero-TMPRSS2, SARS-CoV-1, SARS-CoV-2, respectively, in which multiple Raman ‘snapshots’ on different positions of a single particle and the average spectrum are presented. The peak assignment information is given in the supplementary material. Peaks in the spectra typically originate from the molecular bonds within amino acids, nucleic acid, Amide, C-C stretching or CH_n deformation etc. Multiple spectral patterns were discovered within each type of specimen (e.g., SARS-CoV-1) though the spectral signatures from a single particle are uniform, therefore a standard

representative signature is lacking. A possible reason is that SERS platform renders a superior sensitivity in detecting particles with extremely low concentration, the spectral signature is also prone to fluctuate due to the minor structural change of the molecule and the analyte-hotspot interaction. Hereby, we implemented the supervised and unsupervised learning model for building viral fingerprints, which would be used as a standard for virus identification.

The virus samples were purified from Vero-TMPRSS2 cells by sequential centrifugation, other biological particles with a similar dimension as the virus might be retained, leading to the non-ideal purity which could confuse the identifying model. Therefore, we implemented a control sample of Vero-TMPRSS2 cells under the same preparation manner expecting infection. The spectral signatures from the control act as background signals of the SARS-CoV-1 and SARS-CoV-2 spectral datasets. LDA was implemented to reduce the dimension of the spectra for clearer visualization of the datapoints distribution, in which the original spectra dataset was transformed into points with two-dimensional coordinates. LDA tries to group the spectra by maximizing the distance between the centroid of each group to the global centroid meanwhile minimizing intra-group variance. The inter-group distance conceptually represents the similarity between the corresponding spectra, as shown in Figure 5.5. It can be concluded that SARS-CoV-1 and SARS-CoV-2 clouds overlap with the Vero-TMPRSS2 in small portions, which are believed to be the non-virus particles examined in virus samples. Subsequently, HCA was used to cluster similar particles in Vero-TMPRSS2 and virus samples. Based on the groups clustered, we label the particles originally belonging to virus samples but clustered into Vero-TMPRSS2 as negative (i.e., non-SARS-CoV-2). We call this “label-correction process”, as shown in Figure 6.6A, 6.6B. Figure 5.6C, 5.6D, 5.6F present three similar SERS spectral signatures from different particles belonging

to the same cluster. The spectrum in Figure 5.6C was originally mislabeled by SARS-CoV-2 which would be corrected. Peak assignments are given in Table S1.

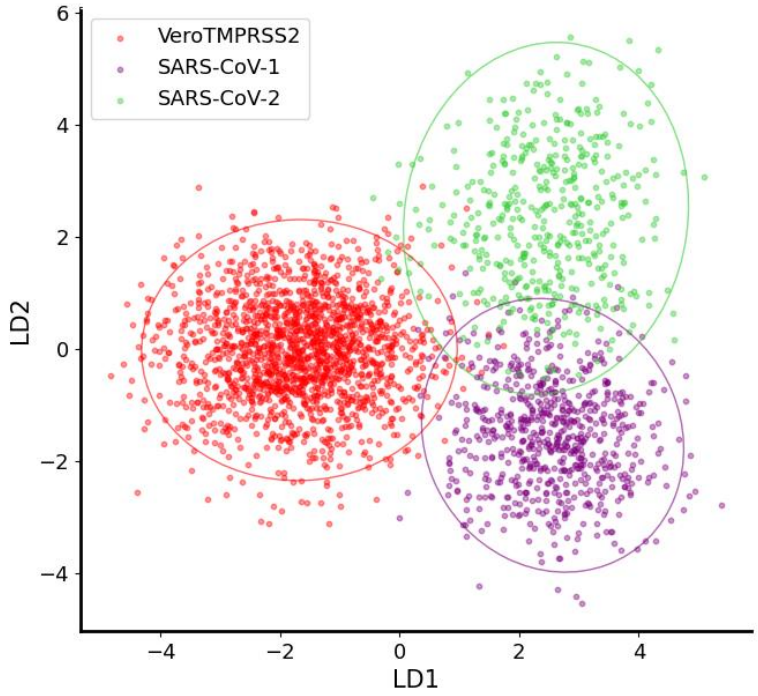


Figure 5.5 Linear Discriminant Analysis for dimensionality reduction. Spectral signatures of SARS-CoV-1, SARS-CoV-2, exosomes are processed by dimensionality reduction and visualized in the 2-dimensional plot.

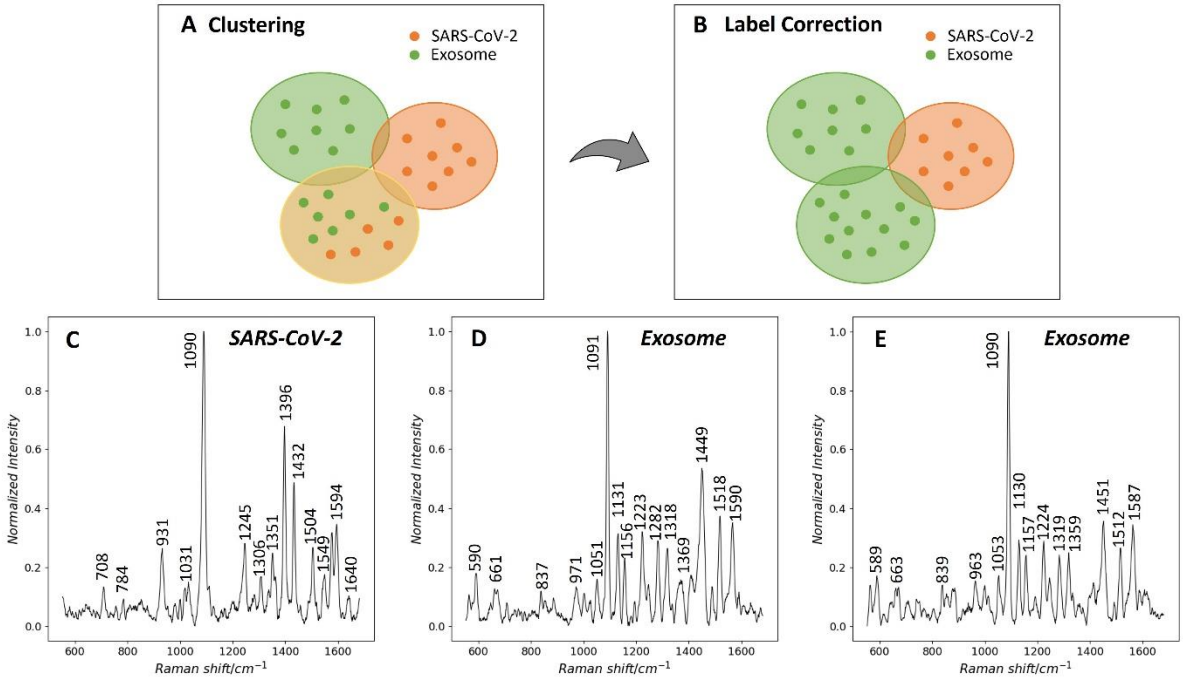


Figure 5.6 HCA for correcting the mislabeled exosomes. (A) Colored ovals are the clusters generated by HCA. Those clusters mixed by SARS-CoV-2 and exosome denote the existence of exosomes in SARS-CoV-2 specimen. (B) Exosomes' labels in the mixed clusters are corrected. (C), (D), (E) are three spectra attributed to different particles from the same cluster, where similar patterns are shown.

A binary classification model using support vector machines (SVM, RBF kernel, soft margin applied) was used in learning the characteristic fingerprints of SARS-CoV-1 and SARS-CoV-2. Due to the binary learning and predicting manner, the testing or validation spectra were either recognized as SARS-CoV-1 or SARS-CoV-2, based on the relative population ratio of SARS-CoV-1 and SARS-CoV-2 for each sample. Without loss of generality, we chose SARS-CoV-2 percentages (e.g., 50 found among 200 thus, 40.0%) as the score. Considering the various viral concentrations and non-virus particles in the specimens, we assigned the binary labels to non-SARS-CoV-2 (or negative) and SARS-CoV-2 (or positive) to avoid confusion and applied a threshold to draw the boundary between the score of two types of virions. It is important to mention

that the threshold was determined practically to maximize the cross-validation performance, also the sample threshold will be further applied or updated whenever more learning and predicting duties come.

During the training process, as more training instances are input, the model gradually learns the distinguishable features between the positive and the negative. Figure 5.7 shows the training error starts from 35% when 10% of the training process is done, and finally ends up with less than 5% after the training process is finished. Additionally, Figure 5.8 demonstrates a gradual separation between the scores of negative instances and positive instances.

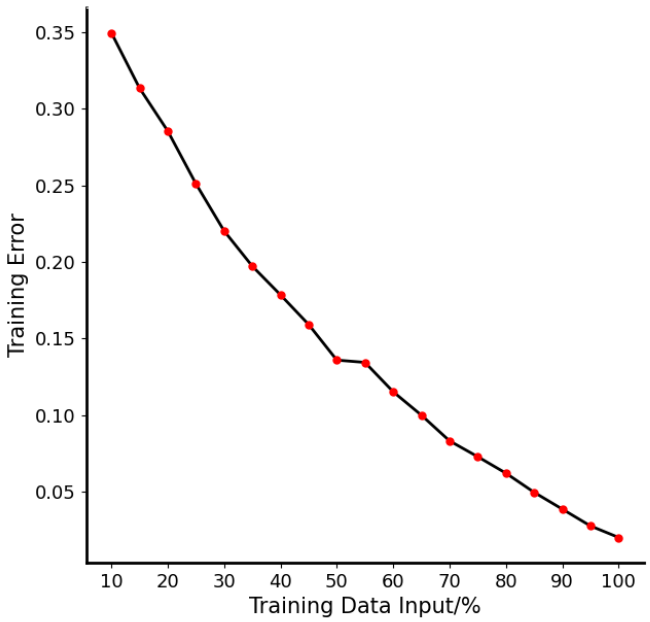


Figure 5.7 Model training process of training error. Training error gradually decreases as training instances being input.

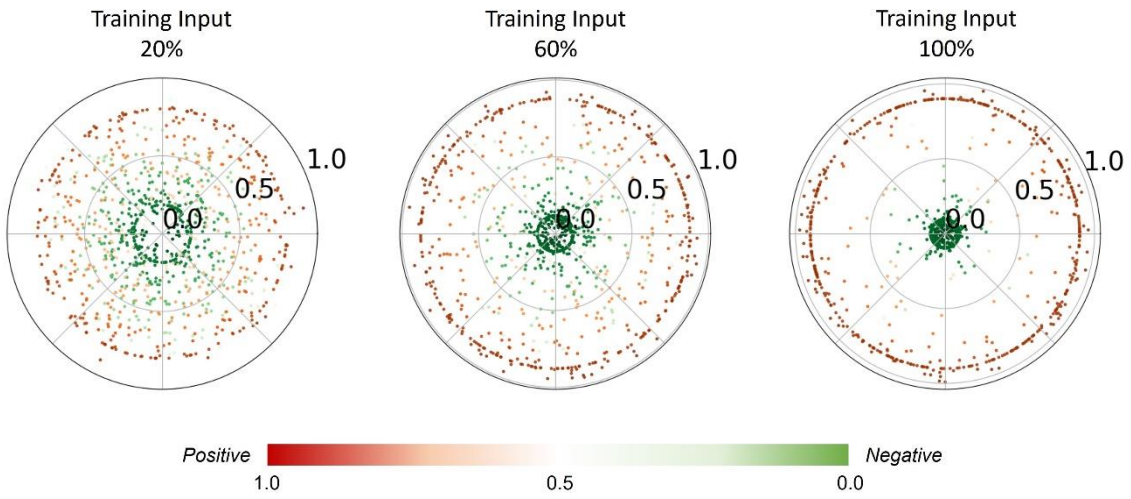


Figure 5.8 Model training process of datapoints separation. Scores of negative and positive instances gradually segregate.

We incorporated cross-validation for optimizing the classifier hyperparameters as well as choosing an appropriate threshold that generates the best predictive capability. Furthermore, to genuinely evaluate the predictive capability by alleviating the overfitting problem during validation, we applied 'leave pair of samples out' (LPSO) cross-validation. Demonstrated by Figure 5.9, In each round of validation, a pair of samples, one each from positive and negative groups respectively, are left out as the validation set while the remaining are the training set. The 'pair' manner is to ensure the sample balance in both training and validation. This process continues until every sample is traversed once as the validation set. A sample score (positive vesicle rate of a sample) list for all the samples is built once the cross-validation is completed, then the ROC curve is plotted together with the information of the true labels by adjusting the threshold.

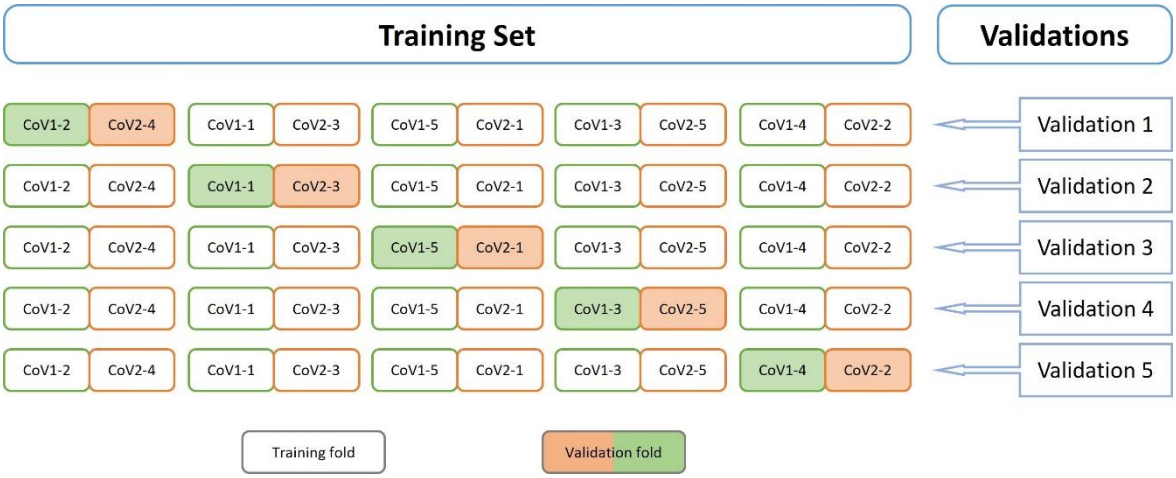


Figure 5.9 LPSO cross validation. Five rounds of cross-validation are conducted; In each round, training folds (unfilled blocks) and validation folds (filled blocks) are assigned for training and validating respectively.

Following the above protocol, the ROC curve is calculated and shown in Figure 5.10, which demonstrates an overall good pattern recognizing capability across all types of viruses. Accordingly, the scores of the samples were shown in the box plot of chart 1, based on the statistical properties of each cross-validation round, we applied the mean of positive sample quantile Q1 and negative sample quantile Q3 as the threshold to maximize the 'margin'. Figure 5.11 shows the fluctuations of the threshold (mean of Q1 and Q3) in cross validations. As indicated in Figure 5.11 and Table 5.1, a threshold of 0.300 was finalized which maximizes the average margin in cross-validations.

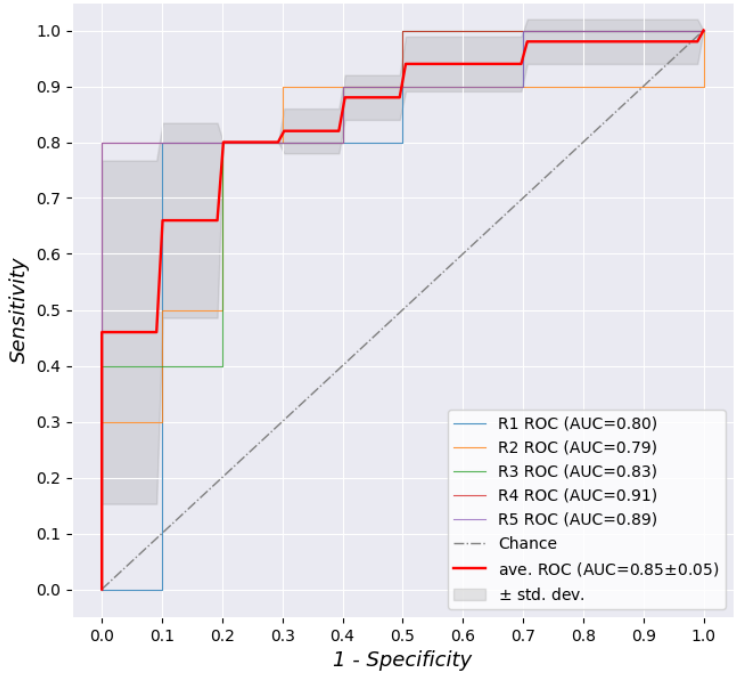


Figure 5.10 Individual and mean ROC curves of cross validations.

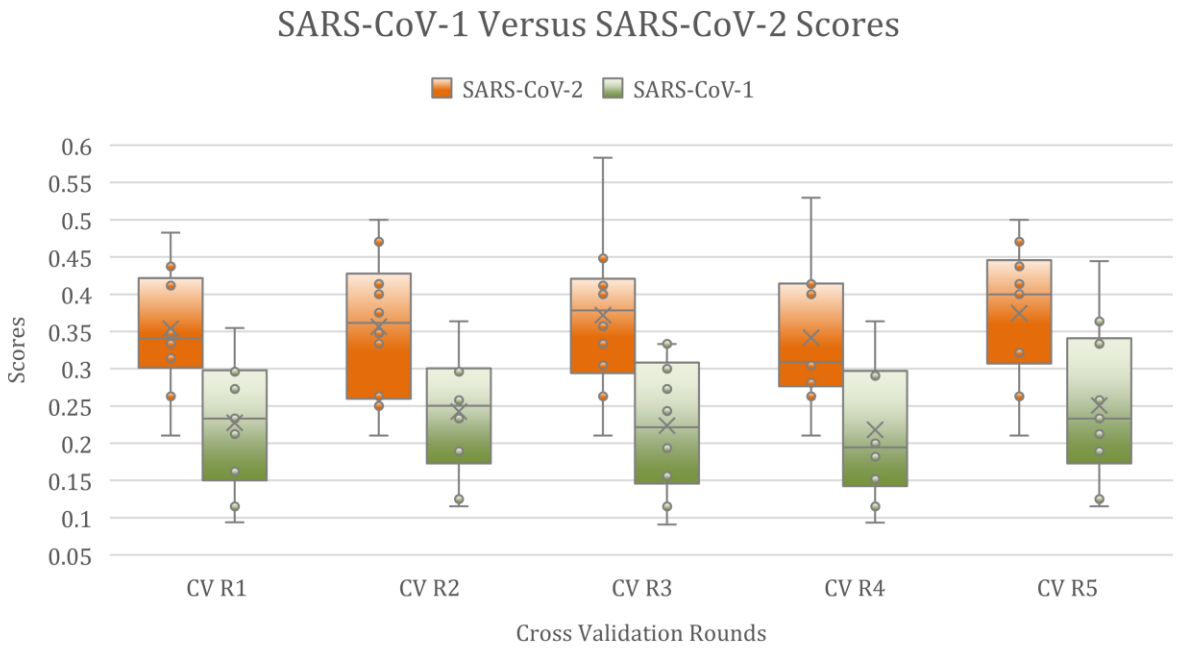


Figure 5.11 Sample scores distribution in the validation folds of cross validation rounds.

Table 5.1 Q1 and Q3 values of cross validations.

Cross-validation	Non-SARS-CoV-2(Q3)	SARS-CoV-2(Q1)	Q1&Q3 Mean
R1	0.290	0.316	0.303
R2	0.299	0.281	0.290
R3	0.293	0.312	0.302
R4	0.295	0.282	0.289
R5	0.314	0.322	0.319
AVE.	-	-	0.300

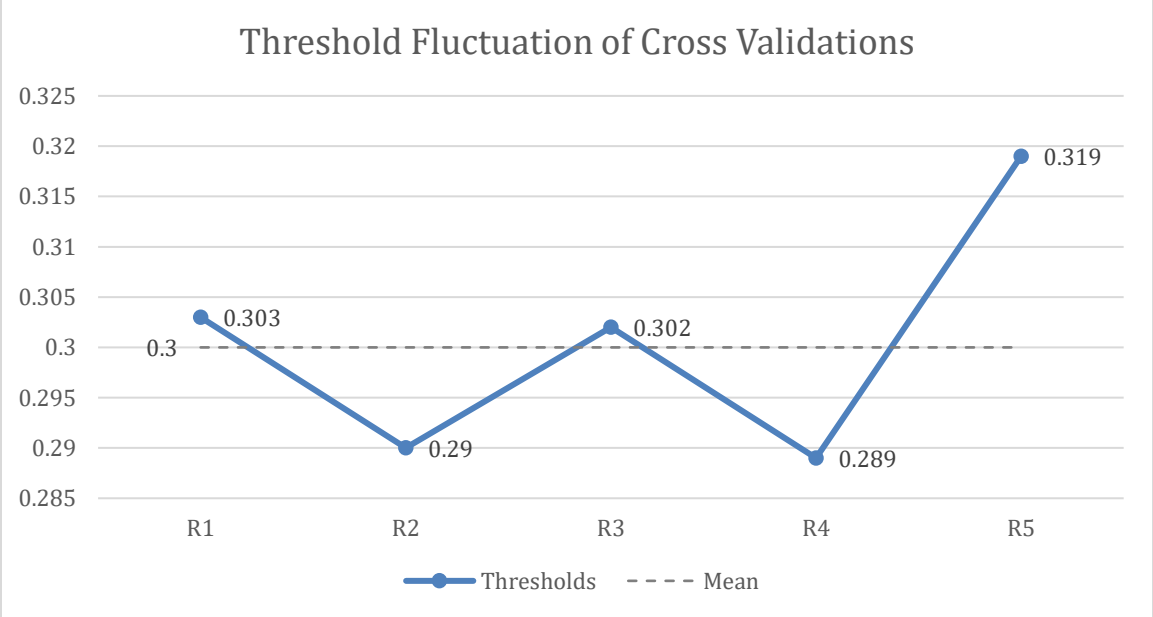


Figure 5.12 Fluctuations of threshold versus cross validation rounds.

Blind tests were subsequently performed after the classification model is optimized. 5 SARS-CoV-2 virus specimens versus 5 SARS-CoV-1 virus specimens were blinded to be given

predictions. Promising performance was given by the threshold equal to 30.0% and the sensitivity/specificity turned out to be 80%/80%. Table 5.2 shows the test results and Figure 5.13 shows the positive ratio generated by the classifier. This result combined with the LDA grouping demonstrates the feasibility of utilizing machine learning classifier and SERS to build a SARS-CoV-2 identifier, given that the specimen has a low diversity of the content (i.e., viruses and extracellular vesicles from Vero-TMPRSS2) and high viral load (10^8 – 10^{10} particles/mL).

Table 5.2 Blind test results of SARS-CoV-1 versus SARS-CoV-2.

Sample ID	**Negative**	**Positive**	**P.R.**	**Predictions**	**Labels**
1	50	10	16.7	Non-CoV-2	Non-CoV-2
2	54	12	18.2	Non-CoV-2	Non-CoV-2
3	38	14	26.9	Non-CoV-2	Non-CoV-2
4	39	12	23.5	Non-CoV-2	Non-CoV-2
5	39	24	38.1	Cov-2	Non-CoV-2
6	43	19	31.1	Cov-2	Cov-2
7	48	8	14.3	Non-CoV-2	Cov-2
8	33	15	31.2	Cov-2	Cov-2
9	40	24	37.5	Cov-2	Cov-2
10	38	18	32.1	Cov-2	Cov-2

Note: Negative, predicted Non-SARS-CoV-2 particles; Positive: predicted SARS-CoV-2 particles; P.R., Positive ratio (%)

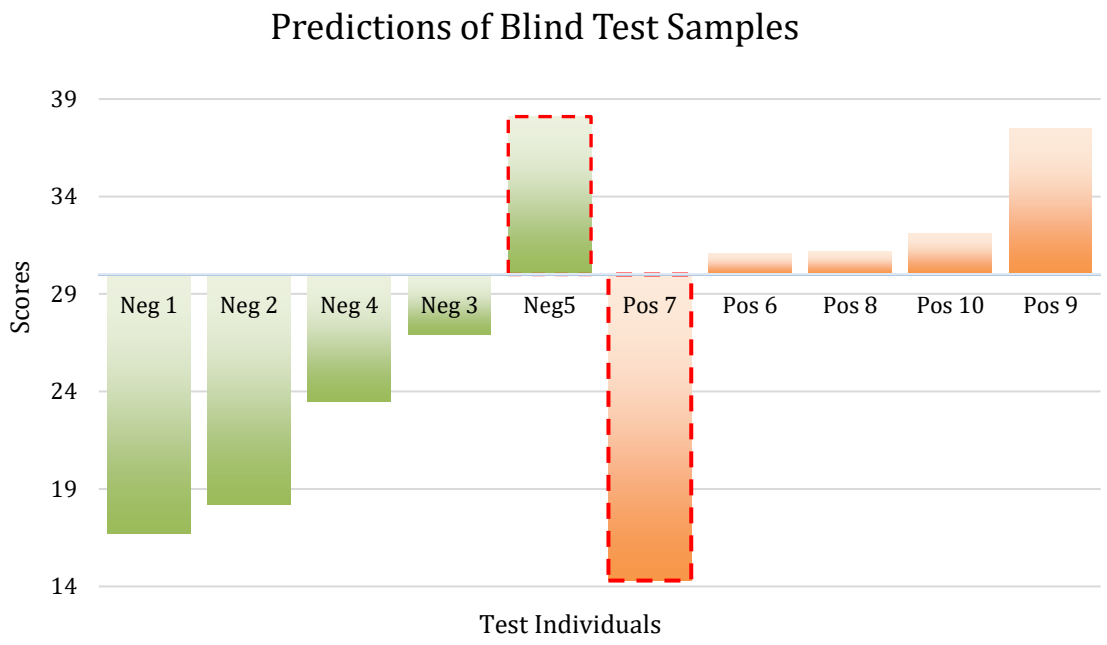


Figure 5.13 Sample scores of the blind test in distinguishing SARS-CoV-1 versus SARS-CoV-2.

5.3.3 Detection of SARS-CoV-2 in virus spiked human saliva

Given the capability of identifying SARS-CoV-2, we further evaluated our SERS fingerprinting plus SVMs protocol on the specimens with higher biological content complexity and closer to the clinical specimens, i.e., virus spiked saliva samples. Specifically, we introduced SARS-CoV-2 virus spiked saliva samples and healthy controls saliva samples as negative control. The preparation protocol of virus spiked saliva samples is given in the Materials and Methods section. A new SVMs classifier was trained using 10 SARS-CoV-2 virus spiked saliva samples versus 10 healthy control saliva samples. Around 50 analytes are collected for each sample, therefore the training dataset is composed of 999 analytes with 9689 spectra.

Like the data cleaning step in SARS-CoV-1 and SARS-CoV-2 study, the non-SARS-CoV-2 particles were subtracted from the SARS-CoV-2 spiked saliva training set by finding the spectral signatures overlapping between healthy control and SARS-CoV-2 spiked saliva. HCA was again

implemented in this background removal process. To ensure the objectivity of the classification and avoid information leakage, background removal is only done to the training set, excluding both the validation set and blind test set. The training set compositions before and after background removal were compared and shown in Figure 5.14.

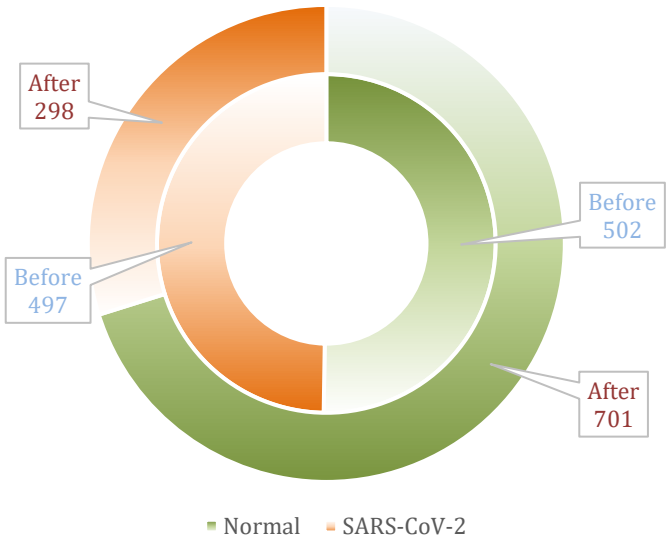


Figure 5.14 Number of training instances before and after label correction by clustering analysis.

Before launching into the blind test, LPSO cross-validation was done with SARS-CoV-2 spiked saliva (or positive) and healthy control (or negative) as the binary groups. As indicated by the ROC curve in Figure 5.15, 0.83 AUC was achieved in cross-validation, which showed reasonable performance. As the previous cross validations, the statistical analyses of the sample scores of cross-validations were presented in Figure 5.16 and Table 5.3, and the mean of positive quantile Q1 and negative quantile Q3 was chosen as the threshold that maximizes the margin between the two types. Figure 5.17 shows the threshold fluctuation. The trained model by ten virus

spiked saliva and ten healthy control individuals were used as classifier, together with a 0.259 as the score threshold.

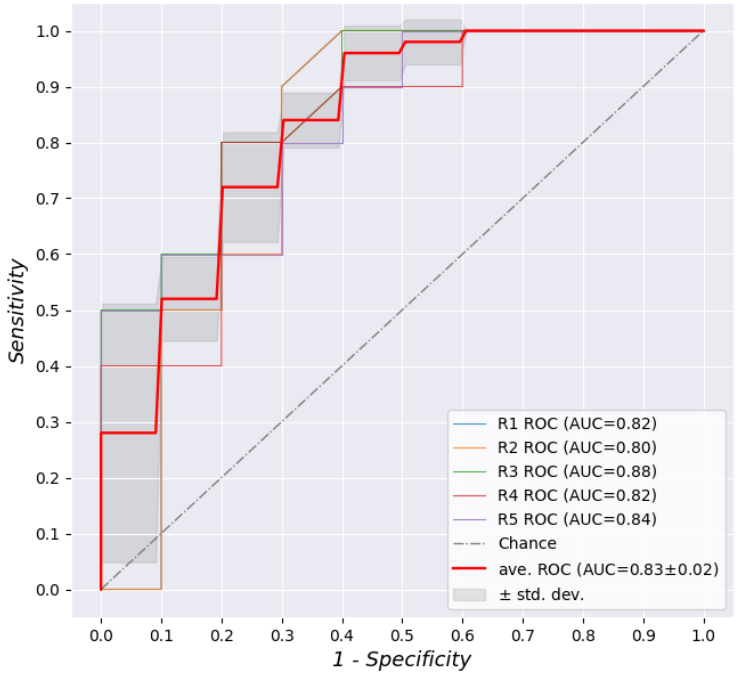


Figure 5.15 Individual and mean ROC curves of cross validations.

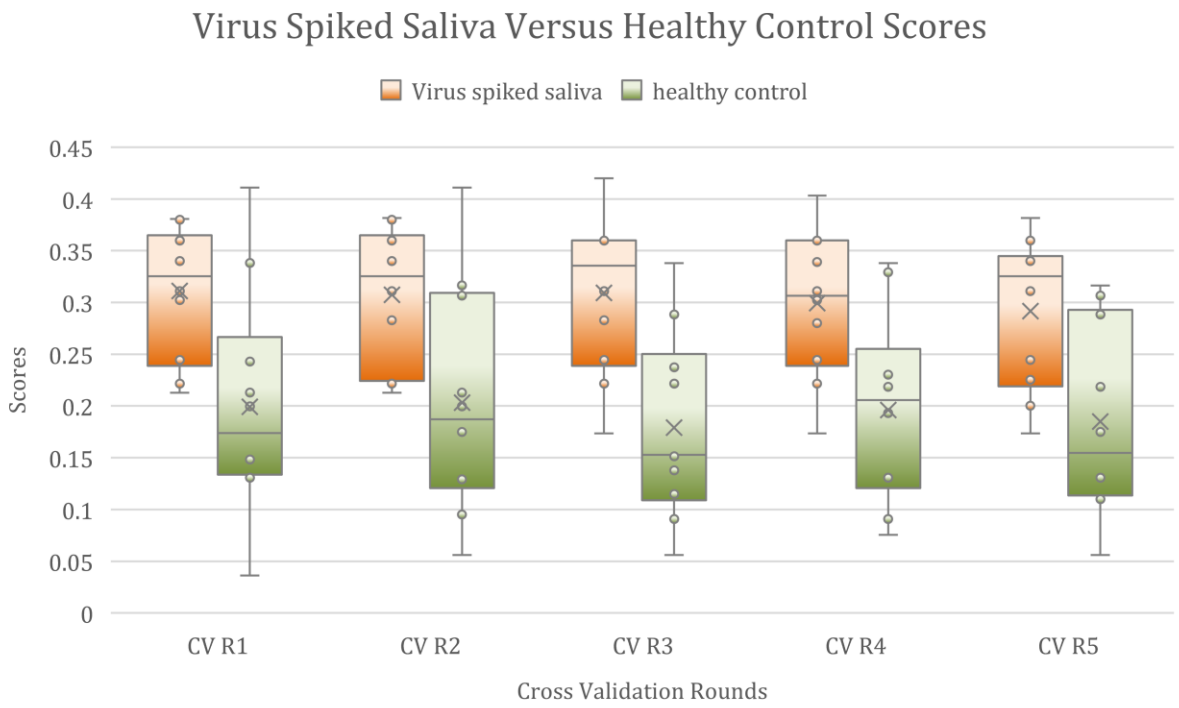


Figure 5.16 Sample scores distribution in the validation folds of cross validation rounds.

Table 5.3 Q1 and Q3 values of cross validations.

Cross-validation	**Virus Spiked Saliva**	**Healthy Control**	**Q1&Q3 Mean**
R1	0.259	0.235	0.247
R2	0.240	0.283	0.262
R3	0.254	0.233	0.244
R4	0.360	0.228	0.294
R5	0.230	0.271	0.251
AVE.	-	-	0.259

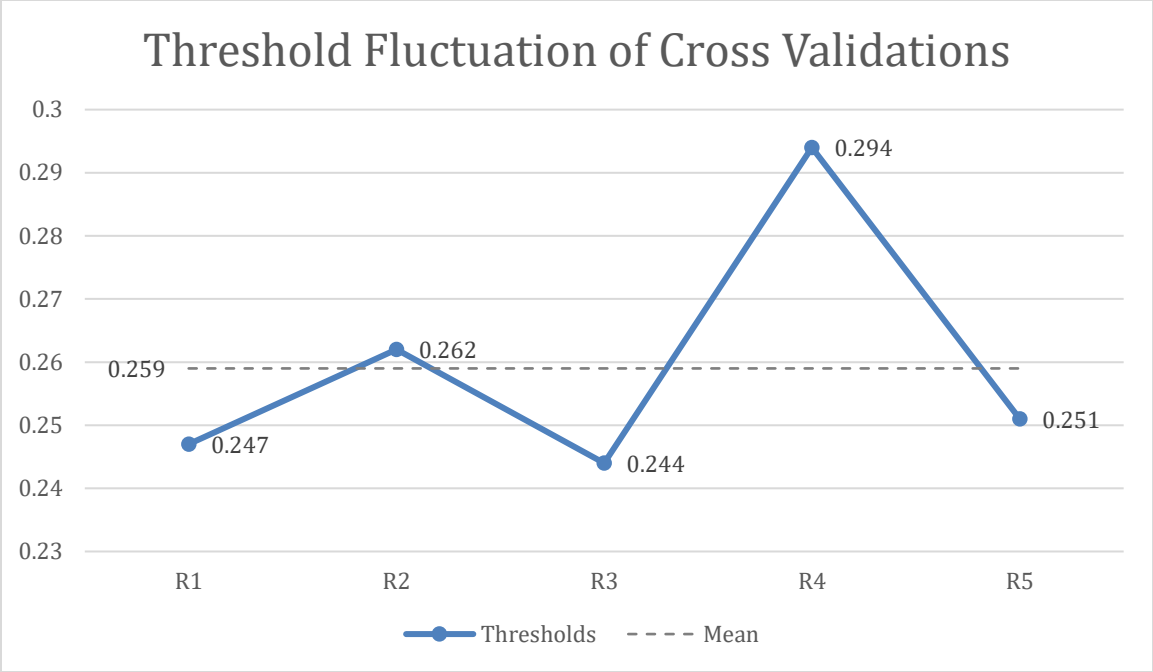


Figure 5.17 Fluctuations of threshold versus cross validation rounds.

Having trained the classifier, a blind test round with ten virus spiked saliva samples and ten healthy control saliva samples was then conducted. The virus spiked saliva samples were prepared following the same protocol as the cross-validation round, but with different healthy saliva backgrounds for mixing. This process is to simulate the various non-virus contents in human salivary specimens. The predictions and unblinding results are shown in Table 5.4 and Figure 5.18, and the corresponding decision matrix is presented in Table 5.5. 90% sensitivity and 80% specificity were achieved with one virus spiked individual and two healthy control individuals predicted incorrectly. The blind test outcome indicates a reasonable performance while trying to apply our platform in diagnosis.

We do also notice some potential pitfalls. First, samples 16, 17, 20 are right at the threshold decision line as shown in Figure 5.18, which decreases the robustness of the platform since the tolerance for statistical fluctuations is limited. Second, a blurrier decision boundary between the

positive/negative groups is present in the spiked saliva study compared to the virus in cell lysate study. This is demonstrated by the more positive/negative group scores overlapping, making it harder to draw an unambiguous decision boundary. The above potential pitfalls are due to the higher bioparticle complexity after spiking virus in the human salivary specimens. Therefore, decisive SARS-CoV-2 signatures are indispensable in improving the accuracy and robustness of our platform.

Table 5.4 Blind test results of SARS-CoV-2 spiked saliva versus healthy control saliva samples.

Sample ID	**Negative**	**Positive**	**P.R.**	**Predictions**	**Labels**
1	41	12	22.6	Control	Control
2	38	16	29.1	Virus	Virus
3	34	16	30.2	Virus	Virus
4	53	14	20.6	Control	Control
5	42	16	26.7	Virus	Virus
6	42	7	13.7	Control	Control
7	38	12	23.1	Control	Control
8	41	7	13.7	Control	Virus
9	25	11	29.7	Virus	Control
10	32	13	27.7	Virus	Virus
11	36	16	30.8	Virus	Virus
12	28	18	39.1	Virus	Control
13	35	10	22.2	Control	Control

14	35	15	30.0	Virus	Virus
15	38	11	22.4	Control	Control
16	37	13	26.0	Virus	Virus
17	37	13	26.0	Virus	Virus
18	36	13	26.5	Virus	Virus
19	40	11	21.6	Control	Control
20	35	12	25.5	Control	Control

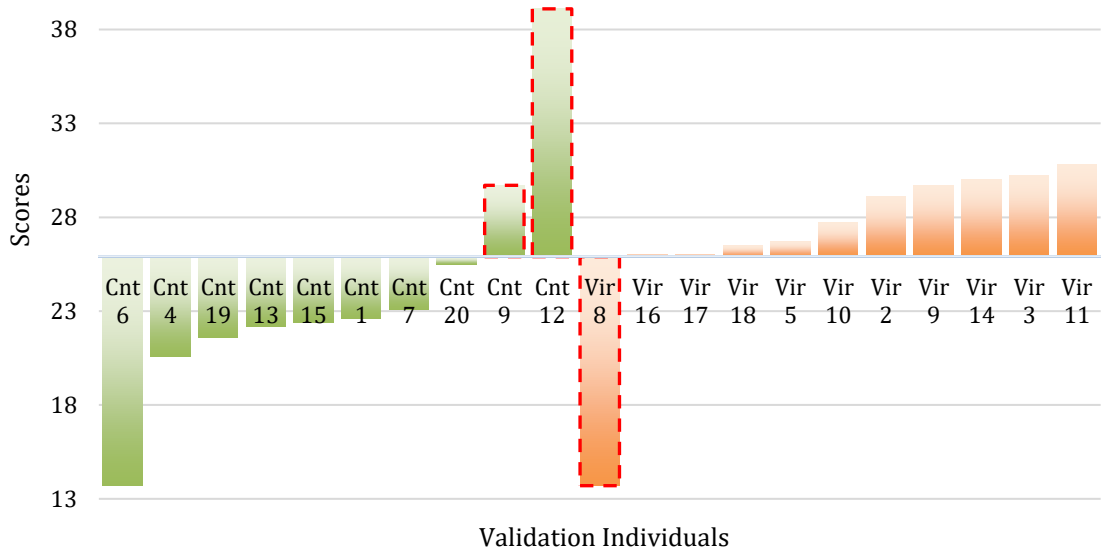


Figure 5.18 Sample scores of the blind test in distinguishing SARS-CoV-2 spiked saliva versus healthy control saliva.

Table 5.5 Confusion matrix of blind test with SARS-CoV-2 spiked saliva samples.

	Predicted Virus	***Predicted Healthy Control***
True Virus	9	1
True Healthy Control	2	8

5.3.4 Detection of SARS-CoV-2 in human saliva

All the studies are the prerequisites for successfully utilizing our platform in clinical diagnosis. Besides, the SARS-CoV-2 purified from Vero-TMPRSS2 cell media or SARS-CoV-2 spiked salivary specimens are simpler laboratory cases compared to the COVID patients' salivary specimens. Therefore, an additional test with clinical samples is necessary to evaluate the practical diagnostic capability.

Since SARS-CoV-2 spiked saliva samples can serve as a 'standard' repository for building the training set due to the presence of both SARS-CoV-2 virions and non-SARS-CoV-2 bioparticles (e.g., proteins, EVs), we applied the same trained classifier in the virus spiked saliva study based on the already proven predicting performance. The same threshold of 0.259 is used as well.

The detailed sample scores are shown in Table 5.6 and Figure 5.19. The final sensitivity and specificity turn out to be 100% and 80%, with only one healthy control predicted incorrectly. Among the correctly predicted samples, SN36's score is right at the decision boundary which will be sensitive to the whole training-predicting system, the remaining are clearly far from the decision boundary, as shown in Figure 5.19. Even though the small test set might be prone to statistical fluctuations, the preliminary success presents a promising application of the SERS platform in

SARS-CoV-2 diagnosis. Table 5.7 is the confusion matrix of the clinical test and Figure 5.20 shows the corresponding ROC curve.

Table 5.6 Results of blind test with clinical samples.

Sample ID	Negative	Positive	P.R.	Predictions	Labels	Ct Value
CLE92	177	77	30.3	Patient	Control	ND
CLE103	241	77	24.2	Control	Control	ND
HOS192	190	75	28.3	Patient	Patient	33.43
SN36	107	37	25.6	Control	Control	ND
HOS167	306	137	30.9	Patient	Patient	ND
HOS182	285	118	29.3	Patient	Patient	31.84
SN33	137	46	25.1	Control	Control	ND
HOS161	159	80	33.5	Patient	Patient	36.42
HOS189	118	47	28.5	Patient	Patient	29.36
SN34	244	67	21.5	Control	Control	ND

ND: Not detected

Table 5.7 Confusion matrix of blind test with clinical samples.

	Predicted Virus	*Predicted Healthy Control*
True Virus	5	0
True Healthy Control	1	4

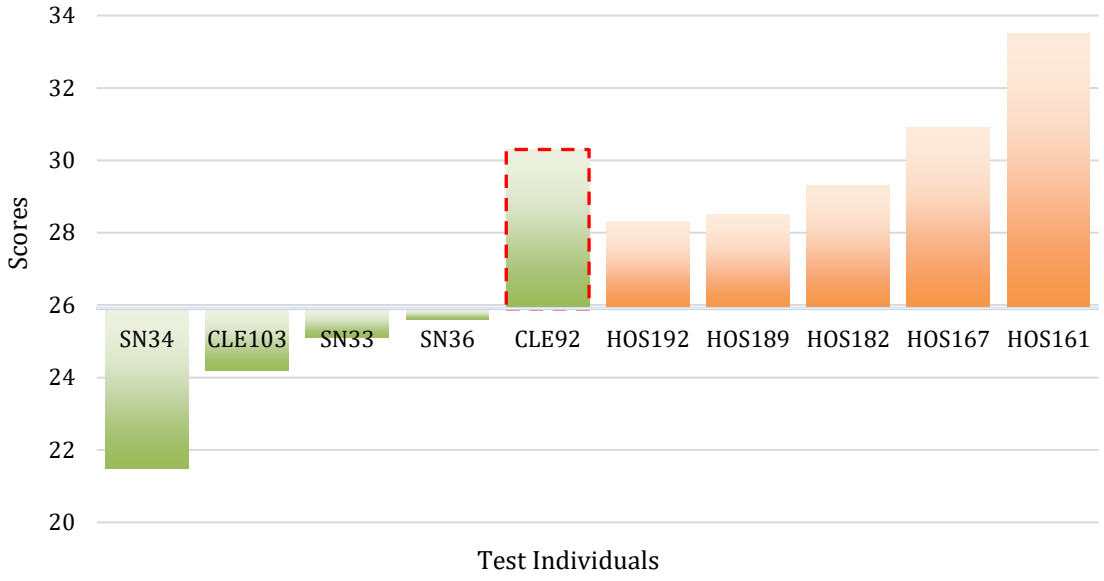


Figure 5.19 Sample scores of the clinical test in distinguishing COVID patients versus healthy controls.

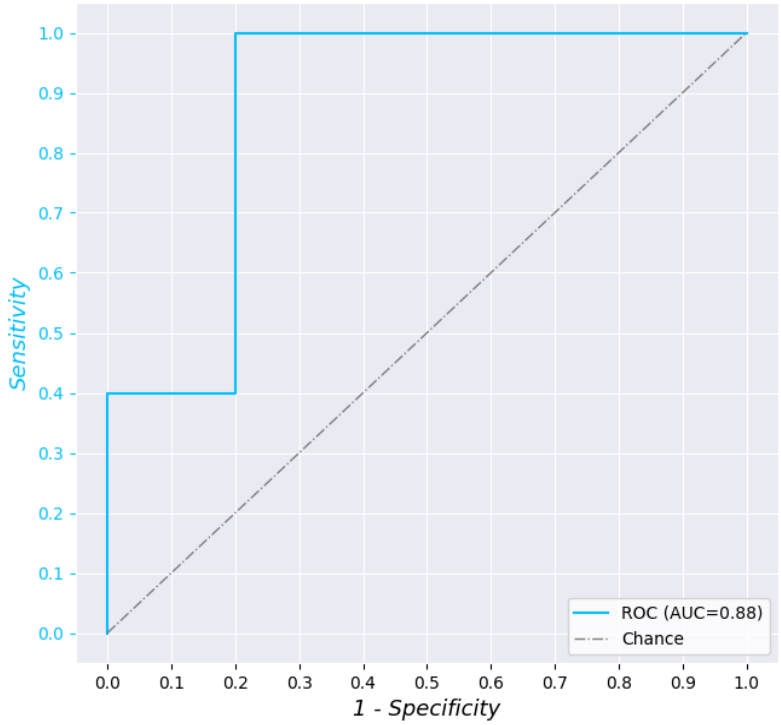


Figure 5.20 ROC curve of clinical sample blind test.

5.4 Conclusion

In this study, we utilized support vector machine incorporated with Radial Basis Function (RBF) kernel and soft margin regularization. To illustrate the fundamental working principle in identifying SARS-CoV-2 SERS spectral signatures, we consider the mathematical definition of the RBF and the training process under the hood. Within the RBF expression given in equation 5-1,

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\gamma \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2\right), \boldsymbol{x}_i \; and \boldsymbol{x}_j \; represent \; spectrum \qquad 5\text{-}1$$

Where γ is a constant. The SERS spectrum term $\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2$ is recognized as the square of Euclidean distance. The exponential term allows for attenuation to assign a higher weight to closely separated training samples, and to normalize the original squared Euclidean distance to zero and one. Therefore, the SVM algorithm essentially searches for an optimal decision boundary that minimizes the intra-group distance score (given by the kernel function), and at the same time maximizes the inter-group distance score. Consequently, the fundamental principle is essentially to analyze the similarity represented by the spectral peak property, which is determined by the biochemical content of the analyte. The final classifier is trained to build a distinguishing criterion to identify SARS-CoV-2 presence versus other non-SARS-CoV-2 content such as SARS-CoV-1 or extracellular vesicles.

In addition to the working principle of support vector machine classifier, one more prerequisite for successful classification is the need for intra-SARS-CoV-2 group spectral differences to be less prominent than the ones between SARS-CoV-2 group and non-SARS-CoV-2 group. SARS-CoV-2 is believed to have developed many variants with slightly different components. Among our studies, Washington strain was used to prepare virus spiked saliva samples while clinical samples were introduced without considering the mutant variant. The preliminary test performance provides indirect proof of our assumption.

Additionally, we translated the spectrum-level predictions given by the support vector machine classifier to a sample-level prediction by summarizing the instances belonging to each group. Then we chose a rather practical way to set up the decision boundary, which is based on cross-validation performance. The implicit reason is that we have quite limited knowledge about the viral load as well as the ratio of SARS-CoV-2 versus other particles. Fortunately, we could make the initial assumption that the genuine target (i.e., SARS-CoV-2) is present and only present in the virus spiked saliva specimens and patient specimens. Therefore, the positive group is bound to give a higher score than the negative group if enough analytes are characterized, due to the presence of the extra distinct SARS-CoV-2 group compared with the control group. This initial conclusion ensures that we can find the approximate position of the decision threshold via 'big data strategy', which is the one that optimizes the validation performance including 20 specimens in our study. Correspondingly, the threshold contains the information on the implicit ratio of the target particles versus non-target particles. It is believed that a larger sample set is more advantageous to diagnostic accuracy.

In conclusion, we demonstrated the feasibility of applying SERS and machine learning pattern recognition on SARS-CoV-2 detection by harvesting and analyzing SARS-CoV-2 isolated from cell culture media and virus spiked saliva samples. Clinical testing with 5 patients versus 5 healthy controls was completed with only one false positive, rendering 100% sensitivity and 80% specificity.

In terms of the advantages of our platform, firstly the label-free manner in fingerprinting and identifying SARS-CoV-2 greatly simplifies the reagent, equipment, and specialist requirement. Our well-established SERS platform fabrication protocol and automatic Raman characterization allow for less human involvement. Therefore, a simpler COVID test procedure and lower cost test

could be expected compared with RT-PCR. Additionally, like rapid antigen tests, the saliva-based specimen harvest protocol is fast and non-invasive. Virus isolation and purification are also not needed, which makes the preparation procedure for characterization simpler. The whole test duration using our platform is between 1-6 hours, mainly due to Raman scanning. Consequently, our platform offers a more accurate test performance than antigen test and a more rapid result yield than RT-PCR, those features could enable it to be a better pandemic monitoring technique.

Having demonstrated the feasibility in identifying SARS-CoV-2 Washington strain, SERS shows potential in contributing to distinguishing different variants. Multiclass classification will be conducted in place of binary classification. We have prepared multiple SARS-CoV-2 variants samples including B.1.351, B.1.1.7, BA.1, BA.5.1 etc. and are working on designing a supervised learning model appropriate to the multiclass classification task. Many algorithms have been reported to be efficient and accurate, such as Random Forest (Chaudhary et al., 2016), K-nearest Neighbors (Haixiang et al., 2016), Neural Networks (Minlong Lin et al., 2013). Foreseeing the challenges in differentiating SARS-CoV-2 variants with high similarity and the uniqueness of SERS spectrum, the collection of representative spectral data, the choice of classifier, model's parameters and even feature selections are supposed to be carefully organized.

As we mentioned, the clinical test sample size is small, which could only provide a preliminary indication of the potential of our platform's application for COVID tests. More COVID patient samples are particularly required, and appropriate rounds of double-blind tests are needed to validate the feasibility. More importantly, due to training data consideration, the classifier is built mainly on simulated samples - SARS-CoV-2 spiked saliva samples. Model parameters might vary while we are using clinical sample data for the training. Another key metrics to evaluate a detection technology is the Limit of Detection, repetitive studies of samples with

different viral loads have been planned. As a single particle characterization technique, a reliable throughput of data collection is needed to ensure the rate of capturing the target analyte. We are working on customizing the Raman spectrometer hardware and designing computer controlling software to enable automatic single particle characterization. All the above factors present challenges along the path of implementing SERS's advantages in COVID tests.

The current diagnostic methods for SARS-CoV-2 and its variants primarily rely on RT-PCR, with antigen tests as a convenient and rapid alternative testing. With the development of more diagnostic platforms, the limitations of the standard methods are expected to be addressed, for example, flexibility and point-of-care detectability. We introduced a SERS platform that could potentially be used for efficient diagnosis of SARS-CoV-2 and its future variants. It is expected that further advancements in high-throughput manufacturing techniques and modular design will enable the large-scale production of SERS-active substrates for virus detection. This scalability will play a pivotal role in meeting the increasing demand during outbreaks and ensuring timely and accurate diagnostics.

5.5 References

Adjei, S., Hong, K., Molinari, N.-A. M., Bull-Otterson, L., Ajani, U. A., Gundlapalli, A. V., Harris, A. M., Hsu, J., Kadri, S. S., Starnes, J., Yeoman, K., & Boehmer, T. K. (2022). Mortality Risk Among Patients Hospitalized Primarily for COVID-19 During the Omicron and Delta Variant Pandemic Periods—United States, April 2020–June 2022. MMWR. Morbidity and Mortality Weekly Report, 71(37), 1182–1189.

Allan, M., Lièvre, M., Laurenson-Schafer, H., De Barros, S., Jinnai, Y., Andrews, S., Stricker, T., Formigo, J. P., Schultz, C., Perrocheau, A., & Fitzner, J. (2022). The World Health Organization COVID-19 surveillance database. International Journal for Equity in Health, 21(S3), 167.

Araf, Y., Akter, F., Tang, Y., Fatemi, R., Parvez, Md. S. A., Zheng, C., & Hossain, Md. G. (2022). Omicron variant of SARS-CoV-2: Genomics, transmissibility, and responses to current COVID-19 vaccines. Journal of Medical Virology, 94(5), 1825–1832.

Bar-On, Y. M., Flamholz, A., Phillips, R., & Milo, R. (2020). SARS-CoV-2 (COVID-19) by the numbers. eLife, 9, e57309.

Bell, S. E. J., & Sirimuthu, N. M. S. (2006). Surface-Enhanced Raman Spectroscopy (SERS) for Sub-Micromolar Detection of DNA/RNA Mononucleotides. Journal of the American Chemical Society, 128(49), 15580–15581.

Biasin, M., Bianco, A., Pareschi, G., Cavalleri, A., Cavatorta, C., Fenizia, C., Galli, P., Lessio, L., Lualdi, M., Tombetti, E., Ambrosi, A., Redaelli, E. M. A., Saulle, I., Trabattoni, D., Zanutta, A., & Clerici, M. (2021). UV-C irradiation is highly effective in inactivating SARS-CoV-2 replication. Scientific Reports, 11(1), 6260.

Bruzas, I., Lum, W., Gorunmez, Z., & Sagle, L. (2018). Advances in surface-enhanced Raman spectroscopy (SERS) substrates for lipid and protein characterization: Sensing and beyond. The Analyst, 143(17), 3990–4008.

Cai, Z., Lu, C., He, J., Liu, L., Zou, Y., Zhang, Z., Zhu, Z., Ge, X., Wu, A., Jiang, T., Zheng, H., & Peng, Y. (2021). Identification and characterization of circRNAs encoded by MERS-CoV, SARS-CoV-1 and SARS-CoV-2. Briefings in Bioinformatics, 22(2), 1297–1308.

Carlin, A. F., Clark, A. E., Garretson, A. F., Bray, W., Porrachia, M., Santos, A. T., Rana, T. M., Chaillon, A., & Smith, D. M. (2023). Neutralizing Antibody Responses After Severe Acute Respiratory Syndrome Coronavirus 2 BA.2 and BA.2.12.1 Infection Do Not Neutralize BA.4 and

BA.5 and Can Be Blunted by Nirmatrelvir/Ritonavir Treatment. Open Forum Infectious Diseases, 10(4), ofad154.

Challen, R., Brooks-Pollock, E., Read, J. M., Dyson, L., Tsaneva-Atanasova, K., & Danon, L. (2021). Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: Matched cohort study. BMJ, n579.

Chau, C. H., Strope, J. D., & Figg, W. D. (2020). COVID-19 Clinical Diagnostics and Testing Technology. Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy, 40(8), 857–868.

Chaudhary, A., Kolhe, S., & Kamal, R. (2016). An improved random forest classifier for multi-class classification. Information Processing in Agriculture, 3(4), 215–222.

Chen, H., Park, S.-G., Choi, N., Kwon, H.-J., Kang, T., Lee, M.-K., & Choo, J. (2021). Sensitive Detection of SARS-CoV-2 Using a SERS-Based Aptasensor. ACS Sensors, 6(6), 2378–2385.

Chen, H., Park, S.-G., Choi, N., Moon, J.-I., Dang, H., Das, A., Lee, S., Kim, D.-G., Chen, L., & Choo, J. (2020). SERS imaging-based aptasensor for ultrasensitive and reproducible detection of influenza virus A. Biosensors and Bioelectronics, 167, 112496.

Chung, Y.-S., Lee, N.-J., Woo, S. H., Kim, J.-M., Kim, H. M., Jo, H. J., Park, Y. E., & Han, M.-G. (2021). Validation of real-time RT-PCR for detection of SARS-CoV-2 in the early stages of the COVID-19 outbreak in the Republic of Korea. Scientific Reports, 11(1), 14817.

El Amri, C., Baron, M.-H., & Maurel, M.-C. (2003). Adenine and RNA in mineral samples. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 59(11), 2645–2654.

Freeman, W. M., Walker, S. J., & Vrana, K. E. (1999). Quantitative RT-PCR: Pitfalls and Potential. BioTechniques, 26(1), 112–125.

Grewal, R., Kitchen, S. A., Nguyen, L., Buchan, S. A., Wilson, S. E., Costa, A. P., & Kwong, J. C. (2022). Effectiveness of a fourth dose of covid-19 mRNA vaccine against the omicron variant among long term care residents in Ontario, Canada: Test negative design study. BMJ, e071502.

Haixiang, G., Yijing, L., Yanan, L., Xiao, L., & Jinling, L. (2016). BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification. Engineering Applications of Artificial Intelligence, 49, 176–193.

John, A., Sadasivan, J., & Seelamantula, C. S. (2021). Adaptive Savitzky-Golay Filtering in Non-Gaussian Noise. IEEE Transactions on Signal Processing, 69, 5021–5036.

Kamińska, A., Witkowska, E., Winkler, K., Dzięcielewski, I., Weyher, J. L., & Waluk, J. (2015). Detection of Hepatitis B virus antigen from human blood: SERS immunoassay in a microfluidic system. Biosensors and Bioelectronics, 66, 461–467.

Kneipp, K. (2016). Chemical Contribution to SERS Enhancement: An Experimental Study on a Series of Polymethine Dyes on Silver Nanoaggregates. The Journal of Physical Chemistry C, 120(37), 21076–21081.

Kneipp, K., Kneipp, H., Kartha, V. B., Manoharan, R., Deinum, G., Itzkan, I., Dasari, R. R., & Feld, M. S. (1998). Detection and identification of a single DNA base molecule using surface-enhanced Raman scattering (SERS). Physical Review E, 57(6), R6281–R6284.

Lin, C., Liang, S., Peng, Y., Long, L., Li, Y., Huang, Z., Long, N. V., Luo, X., Liu, J., Li, Z., & Yang, Y. (2022). Visualized SERS Imaging of Single Molecule by Ag/Black Phosphorus Nanosheets. Nano-Micro Letters, 14(1), 75.

Luo, S.-C., Sivashanmugan, K., Liao, J.-D., Yao, C.-K., & Peng, H.-C. (2014). Nanofabricated SERS-active substrates for single-molecule to virus detection in vitro: A review. Biosensors and Bioelectronics, 61, 232–240. https://doi.org/10.1016/j.bios.2014.05.013

Minlong Lin, Ke Tang, & Xin Yao. (2013). Dynamic Sampling Approach to Training Neural Networks for Multiclass Imbalance Classification. IEEE Transactions on Neural Networks and Learning Systems, 24(4), 647–660.

Mosier-Boss, P. (2017). Review on SERS of Bacteria. Biosensors, 7(4), 51.

Palonpon, A. F., Ando, J., Yamakoshi, H., Dodo, K., Sodeoka, M., Kawata, S., & Fujita, K. (2013). Raman and SERS microscopy for molecular imaging of live cells. Nature Protocols, 8(4), 677–692.

Pastorino, B., Touret, F., Gilles, M., De Lamballerie, X., & Charrel, R. N. (2020). Heat Inactivation of Different Types of SARS-CoV-2 Samples: What Protocols for Biosafety, Molecular Detection and Serological Diagnostics? Viruses, 12(7), 735.

Peng, J., Peng, S., Jiang, A., Wei, J., Li, C., & Tan, J. (2010). Asymmetric least squares for multiple spectra baseline correction. Analytica Chimica Acta, 683(1), 63–68.

Shanmukh, S., Jones, L., Driskell, J., Zhao, Y., Dluhy, R., & Tripp, R. A. (2006). Rapid and Sensitive Detection of Respiratory Virus Molecular Signatures Using a Silver Nanorod Array SERS Substrate. Nano Letters, 6(11), 2630–2636.

Sharma, B., Frontiera, R. R., Henry, A.-I., Ringe, E., & Van Duyne, R. P. (2012). SERS: Materials, applications, and the future. Materials Today, 15(1–2), 16–25.

Stelzer-Braid, S., Walker, G. J., Aggarwal, A., Isaacs, S. R., Yeang, M., Naing, Z., Ospina Stella, A., Turville, S. G., & Rawlinson, W. D. (2020). Virus isolation of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) for diagnostic and research purposes. Pathology, 52(7), 760–763.

Stiles, P. L., Dieringer, J. A., Shah, N. C., & Van Duyne, R. P. (2008). Surface-Enhanced Raman Spectroscopy. Annual Review of Analytical Chemistry, 1(1), 601–626.

Tahamtan, A., & Ardebili, A. (2020). Real-time RT-PCR in COVID-19 detection: Issues affecting the results. Expert Review of Molecular Diagnostics, 20(5), 453–454.

Tavakkoli Yaraki, M., Tukova, A., & Wang, Y. (2022). Emerging SERS biosensors for the analysis of cells and extracellular vesicles. Nanoscale, 14(41), 15242–15268.

Vasireddy, D., Vanaparthy, R., Mohan, G., Malayala, S. V., & Atluri, P. (2021). Review of COVID-19 Variants and COVID-19 Vaccine Efficacy: What the Clinician Should Know? Journal of Clinical Medicine Research, 13(6), 317–325.

Wang, P., Liang, O., Zhang, W., Schroeder, T., & Xie, Y. (2013a). Ultra-Sensitive Graphene-Plasmonic Hybrid Platform for Label-Free Detection. Advanced Materials, 25(35), 4918–4924.

Wang, P., Liang, O., Zhang, W., Schroeder, T., & Xie, Y. (2013b). Ultra-Sensitive Graphene-Plasmonic Hybrid Platform for Label-Free Detection. Advanced Materials, 25(35), 4918–4924.

Yamayoshi, S., Sakai-Tagawa, Y., Koga, M., Akasaka, O., Nakachi, I., Koh, H., Maeda, K., Adachi, E., Saito, M., Nagai, H., Ikeuchi, K., Ogura, T., Baba, R., Fujita, K., Fukui, T., Ito, F., Hattori, S.,

Yamamoto, K., Nakamoto, T., … Kawaoka, Y. (2020). Comparison of Rapid Antigen Tests for COVID-19. Viruses, 12(12), 1420.

Zhang, P., Yeo, J. C., & Lim, C. T. (2019). Advances in Technologies for Purification and Enrichment of Extracellular Vesicles. SLAS Technology, 24(5), 477–488.

Chapter 6 Summary and prospects

In this dissertation, we delve deeply into the frontier of extracting and analyzing bioinformation from NBPs for disease diagnosis. NBPs exhibit unique formation processes and properties that allow certain types to serve as early indicators even before the onset of diseases. NBPs rich in bio-information hold great promise for future disease diagnosis and prognosis. Leveraging SERS technology's capability to characterize single NBPs, we achieve enhanced accuracy, selectivity, and specificity in detecting informative NBP biomarkers. SERS-based single NBP analysis captures spectral signatures stemming from the collective Raman scattering of all enclosed biomolecules, providing a comprehensive view compared to labeled techniques focusing on specific NBP biomarkers. The tremendous signal enhancement by surface plasmon also contributes to the amount of bioinformation from NBPs, as well as the level of complexity of information. However, this inherent complexity presents challenges in extracting disease-relevant information. We introduce a systematic SOP that combines SERS-based single NBP characterization with ML-based analyses for disease diagnosis. We lay the groundwork by providing fundamental insights into NBPs, with a special focus on exosomes and SARS-CoV-2, the NBPs studied in this dissertation. We also delve into the mathematical principles of data analysis methods, encompassing signal processing algorithms, generic algorithms, and machine learning techniques. This thesis highlights the application of our technique in diagnosing NSCLC and detecting SARS-CoV-2, demonstrating its feasibility in the medical field. This paves the way for investigating a wide array of NBPs. It's important to note that our achievements in this thesis represent preliminary feasibility demonstrations. Future research will encompass diverse NBPs, various diseases, robust diagnostic assessments, and the establishment of sophisticated platforms, ultimately propelling our technology towards clinical applications.

6.1 data throughput and labeled SERS methods

As stated previously, the development of single NBP characterization for disease significantly relies on accurate establishment of disease related standard spectral patterns and successful capture of informative NBP, which pave the path for the following steps such as data analyses, database building, clinical predictions and so on. These two prerequisites will become extremely challenging especially when the population of characteristic NBP is relatively low, such as in our study regarding NSCLC, the presence of lymphocytes derived exosomes will become a serious interference for the detection of cancer related exosomes. Failing to maintain the preconditions will lead to false positives due to wrong standard fingerprints building and false negatives due to missing informative NBPs in SERS scanning. In other words, data throughput issues will occur due to the limited NBP characterization rate and low concentration of informative NBPs, therefore it is pivotal to escalate the throughput of target disease specific NBPs within the allowance of time and resources.

Labeled SERS methods hereby show advantages. These techniques involve the use of engineered SERS-generating nanoparticles or substrates that are labeled with specific biomarker identifier or targeting agents. When these labeled entities interact with biological samples, such as blood or tissue, they provide a means to detect and identify disease-related molecules with remarkable precision and sensitivity (Y. Chen et al., 2023; Davis et al., 2018). In labeled SERS technologies, the labeled SERS platforms act as signal amplifiers. By attaching to specific disease markers, such as proteins or nucleic acids, it generates a highly specific and recognizable signal when analyzed using Raman spectroscopy. This allows for the detection of even trace amounts of disease biomarkers within a complex biological matrix. Another key advantage of labeled SERS in disease diagnosis is its ability to provide multiplexed detection (Gellner et al., 2009; Y. Wang et

al., 2016). Multiple identifiers labeled platforms can be used simultaneously, enabling the detection of multiple disease markers within a single sample. This makes it a valuable tool for diagnosing complex diseases, including cancer, infectious diseases, and neurological disorders. However, labeled SERS methods work only if the disease biomarkers are known and the corresponding "counterpart" (e.g., antibodies, receptors etc.) of the biomarker is available, which is still in development for the majority of diseases.

The advantages of labeled methods can effectively address potential data throughput challenges. As depicted in Figure 6.1, specific molecules can be immobilized onto the SERS substrate, facilitating the selective capture of NBPs from specific sources. In this instance, human ACE2 is affixed to the substrate, enabling the interaction and immobilization of viruses carrying the S protein, including SARS-CoV-2 and its variants. This functionalization of the SERS platform's surface essentially acts as a filter, preselecting informative NBPs based on their sources or attributes. This preselection process sets the stage for subsequent SERS scanning, substantially increasing the likelihood of capturing genuine disease biomarkers. Traditional virus isolation or purification techniques, like SEC, may inadvertently retain NBPs of similar dimensions as the virus within biopsy specimens (e.g., saliva, serum). By employing preselection, extraneous elements like exosomes, protein molecules, RNA/DNA fragments, and cell debris can be eliminated, reducing the risk of false biomarker identification.

This preselection step becomes even more crucial in the context of EV-based diagnostics since the specimen may contain EVs derived from a diverse array of cell types, with the target biomarker representing only a fraction of this population. By narrowing down the search space through preselection, the chances of successfully identifying disease-related biomarkers are significantly enhanced. The critical steps in this process involve the selection of functionalization

molecules and the method of immobilizing these molecules. It is imperative that the chosen functionalization molecules exhibit a strong affinity for the target NBPs and maintain reasonable stability when exposed to laser light. Additionally, immobilization techniques must secure the attachment of functionalization molecules without interfering with the biomarker signatures. Current methodologies often employ the use of cross-linkers, acting as a bridge between the metallic surface and protein molecules. Examples of such cross-linkers include DSP (Dithiobis[succinimidyl propionate], or Lomant's Reagent) (Xiang, 2004), Glutaraldehyde (Webster et al., 2007), Ethylene glycol bis (succinimidylsuccinate) (EGS) (Ding et al., 2016), N-Hydroxysuccinimide (NHS) esters (Mädler et al., 2009) and so on. Given that the surface plasmon hotspots are concentrated within 50 nm above the metallic surface, careful consideration must be given to factors such as the length of the cross-linker, the types of molecular bonds involved, and the Raman reactivity. These considerations are vital to ensure the successful capture and generation of NBP biomarkers.

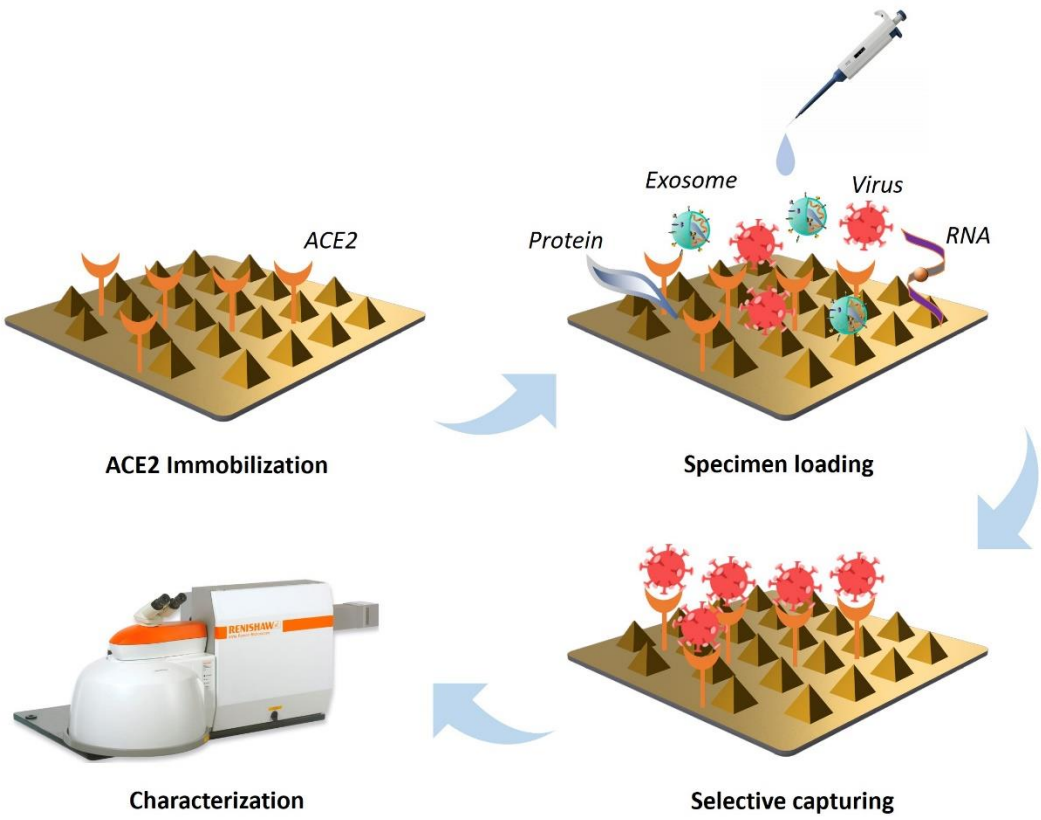


Figure 6.1 SOP of SERS surface functionalization by cross-linking.

6.2 reconstruction of molecular information from SERS spectral features

Our efforts have been directed towards promoting SERS as an alternative technique for proteomic characterization, because in many medical conditions, disorders are often indicated by abnormal levels of specific biomarkers, such as upregulated or downregulated proteins or RNAs. It becomes imperative to establish a clear connection between the spectral features observed in SERS and the composition and status of these biomolecules. This linkage is essential for advancing the application of SERS technology in pathology, medical diagnostics, medical treatment, and therapeutic interventions.

Raman or SERS spectroscopy is known for its ability to reveal molecular vibrational modes, with typical spectral peaks often associated with characteristic chemical bonds. However, when it comes to complex biomolecules like proteins, DNA, RNA, and lipids, these molecules share many of these characteristic chemical bonds. Consequently, it becomes highly challenging to deduce the specific status of biomolecules that are responsible for crucial biological activities (Sitjar et al., 2021). In other words, the level of information provided by SERS is lower than the information required by pathology. NBPs are composed of a collection of biomolecules and will produce increasingly complicated spectral features, making it harder to extrapolate the pathology.

Fortunately, primitive biomolecules including amino acids, nucleotides, lipids, phosphate are reported to produce highly uniform spectral features. Figure 6.2 shows the signatures of several types of amino acids. It provides us a chance to apply AI to learn the standard signatures and predict the primitive molecular composition given a random spectrum. A database consisting of the standard signatures is required and AI-driven models are expected to be trained to predict the molecular composition. Sophisticated algorithms and well-established evaluations metrics are also crucial. The successful execution of this research will promote SERS from molecule fingerprinting technology to another proteomic technique that investigates the structures, functions, and interactions of proteins within a biological system.

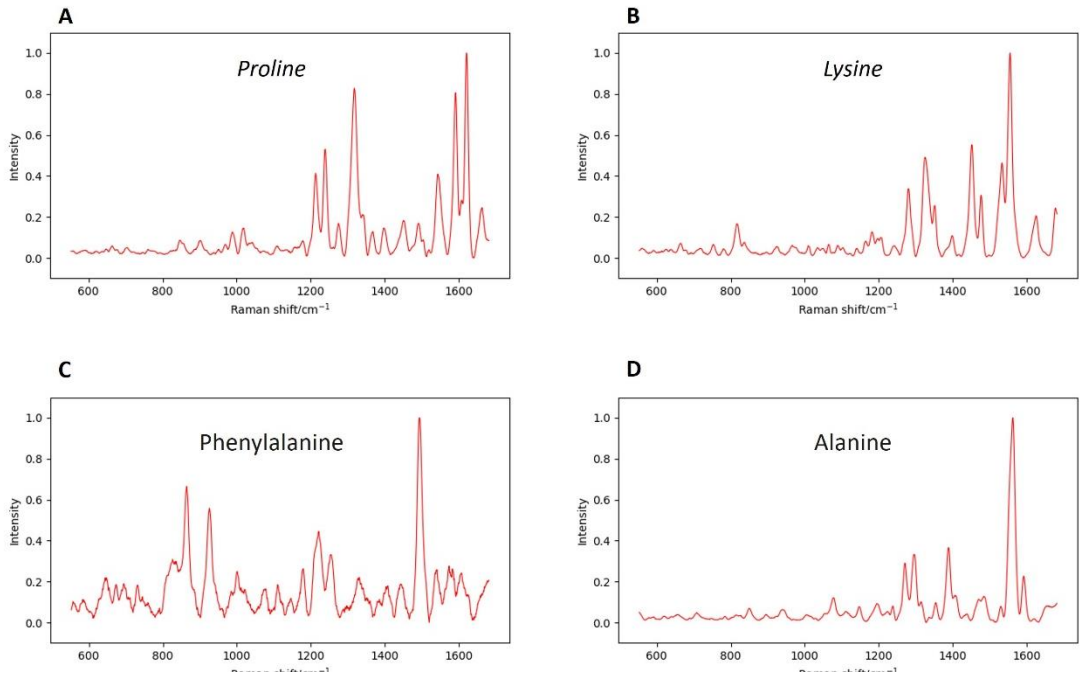


Figure 6.2 Standard spectra of primitive molecules.

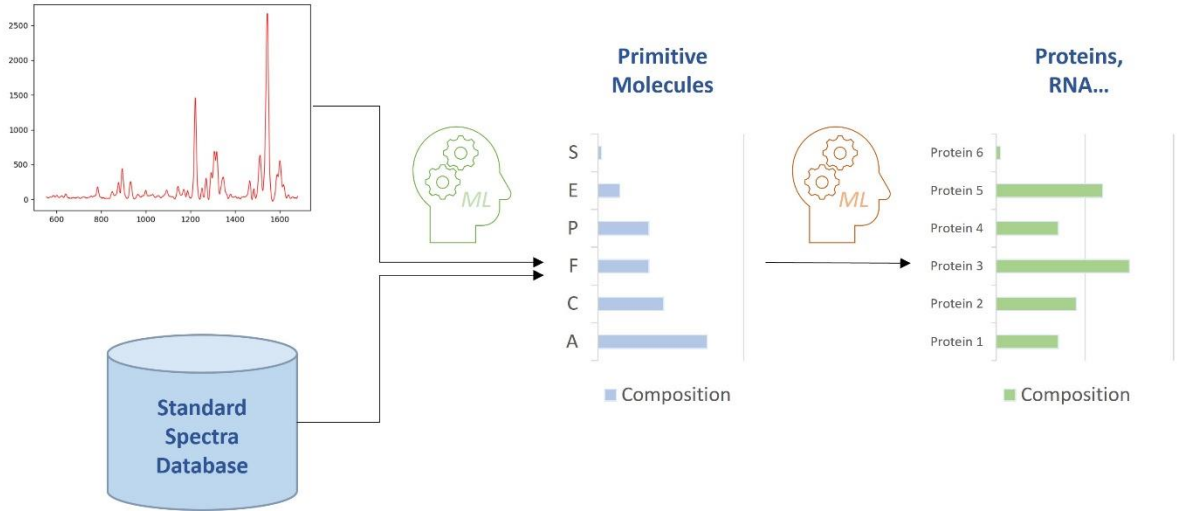


Figure 6.3 Anticipated workflow of molecular information reconstruction from SERS spectra. Primitive molecule composition is firstly predicted, followed by large biomolecules such as proteins, nucleotides composition is predicted.

6.3 References

Chen, Y., Yang, S., Shi, X., He, Z., Peng, H., Gu, G., Pang, X., Chen, H., Wang, Y., & Guo, L. (2023). Au@4-MBA@Ag NPs labeled SERS lateral flow immunoassay for ultrasensitive and quantitative detection of Salmonella enteritidis. Microchemical Journal, 193, 109134.

Davis, R., Campbell, J., Burkitt, S., Qiu, Z., Kang, S., Mehraein, M., Miyasato, D., Salinas, H., Liu, J., & Zavaleta, C. (2018). A Raman Imaging Approach Using CD47 Antibody-Labeled SERS Nanoparticles for Identifying Breast Cancer and Its Potential to Guide Surgical Resection. Nanomaterials, 8(11), 953.

Ding, Y.-H., Fan, S.-B., Li, S., Feng, B.-Y., Gao, N., Ye, K., He, S.-M., & Dong, M.-Q. (2016). Increasing the Depth of Mass-Spectrometry-Based Structural Analysis of Protein Complexes through the Use of Multiple Cross-Linkers. Analytical Chemistry, 88(8), 4461–4469.

Gellner, M., Kömpe, K., & Schlücker, S. (2009). Multiplexing with SERS labels using mixed SAMs of Raman reporter molecules. Analytical and Bioanalytical Chemistry, 394(7), 1839–1844.

Mädler, S., Bich, C., Touboul, D., & Zenobi, R. (2009). Chemical cross-linking with NHS esters: A systematic study on amino acid reactivities. Journal of Mass Spectrometry, 44(5), 694–706.

Sitjar, J., Liao, J.-D., Lee, H., Tsai, H.-P., Wang, J.-R., & Liu, P.-Y. (2021). Challenges of SERS technology as a non-nucleic acid or -antigen detection method for SARS-CoV-2 virus and its variants. Biosensors and Bioelectronics, 181, 113153.

Wang, Y., Kang, S., Doerksen, J. D., Glaser, A. K., & Liu, J. T. C. (2016). Surgical Guidance via Multiplexed Molecular Imaging of Fresh Tissues Labeled With SERS-Coded Nanoparticles. IEEE Journal of Selected Topics in Quantum Electronics, 22(4), 154–164.

Webster, A., Halling, M. D., & Grant, D. M. (2007). Metal complexation of chitosan and its glutaraldehyde cross-linked derivative. Carbohydrate Research, 342(9), 1189–1201.

Xiang, C. C. (2004). Using DSP, a reversible cross-linker, to fix tissue sections for immunostaining, microdissection and expression profiling. Nucleic Acids Research, 32(22), e185–e185.