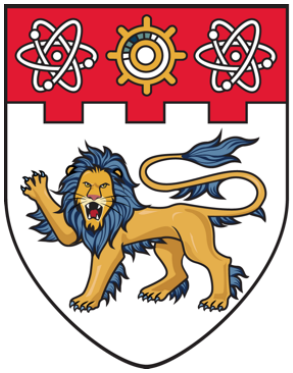


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

WHERE DATA SCIENCE MEETS SURFACE-ENHANCED RAMAN SCATTERING: HARNESSING FINGERPRINT VARIATIONS TO DRIVE PRACTICAL SENSING APPLICATIONS

LEONG YONG XIANG

SCHOOL OF CHEMISTRY, CHEMICAL ENGINEERING AND BIOTECHNOLOGY

2024

WHERE DATA SCIENCE MEETS SURFACE-ENHANCED RAMAN SCATTERING: HARNESSING FINGERPRINT VARIATIONS TO DRIVE PRACTICAL SENSING APPLICATIONS

LEONG YONG XIANG

SCHOOL OF CHEMISTRY, CHEMICAL ENGINEERING AND BIOTECHNOLOGY

A thesis submitted to the Nanyang Technological University in partial fulfilment of the requirement for the degree of Doctor of Philosophy

2024

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research done by me except where otherwise stated in this thesis. The thesis work has not been submitted for a degree or professional qualification to any other university or institution. I declare that this thesis is written by myself and is free of plagiarism and of sufficient grammatical clarity to be examined. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

15 Jan 2024

. . . . . .

Date

. . . . . .

Leong Yong Xiang

*please sign on the water mark area.

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it of sufficient grammatical clarity to be examined. To the best of my knowledge, the thesis is free of plagiarism and the research and writing are those of the candidate's except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

15 Jan 2024

. . . . . .

Date

. . . . . .

Prof. Ling Xing Yi

*please sign on the water mark area.

Authorship Attribution Statement

(B) This thesis contains material from 3 paper(s) published in the following peer-reviewed journal(s) in which I am listed as an author.

Chapter 1 and 6 are published as Yong Xiang Leong, Emily Xi Tan, Shi Xuan Leong, Charlynn Sher Lin Koh, Lam Bang Thanh Nguyen, Jaslyn Ru Ting Chen, Kelin Xia, and Xing Yi Ling. Where Nanosensors Meet Machine Learning: Prospects and Challenges in Detecting Disease X. *ACS Nano*, **16**, 9, 13279-13293 (2022). DOI: 10.1021/acsnano.2c05731.

The contributions of the co-authors are as follows:

- Ms. Tan Xi Emily, Dr. Leong Shi Xuan, Dr. Koh Sher Lin Charlynn, Assoc. Prof. Xia Kelin, Prof. Ling Xing Yi, and I conceptualized the topic.
- I conceptualized and organized the discussion points.
- I drafted the manuscript with inputs from Ms. Tan Xi Emily, Dr. Leong Shi Xuan, Dr. Koh Sher Lin Charlynn, Mr. Nguyen Bang Thanh Lam, Ms. Chen Ru Ting Jaslyn. The manuscript was revised with inputs from Assoc. Prof. Xia Kelin and Prof. Ling Xing Yi.
- Ms. Tan Xi Emily provided inputs on data visualization.
- All authors gave approval to the final version of the manuscript.

Chapter 2 is published as Yong Xiang Leong, Yih Hong Lee, Charlynn Sher Lin Koh, Gia Chuong Phan-Quang, Xuemei Han, In Yee Phang, and Xing Yi Ling. Surface-Enhanced Raman Scattering (SERS) Taster: A Machine-Learning-Driven Multireceptor Platform for

Multiplex Profiling of Wine Flavors. *Nano Lett.*, **21**, 6, 2642-2649 (2021). DOI: 10.1021/acs.nanolett.1c00416.

The contributions of the co-authors are as follows:

- Prof. Ling Xing Yi provided the initial project direction and supervised the research.
- Dr. Koh Sher Lin Charlynn, Dr. Phang In Yee, Prof. Ling Xing Yi and I designed the study.
- I performed laboratory experiments at the School of Chemistry, Chemical Engineering and Biotechnology and analyzed the results.
- I prepared the manuscript drafts. The manuscript was revised with inputs from Dr. Lee Yih Hong, Dr. Koh Sher Lin Charlynn, and Prof. Ling Xing Yi.
- Dr. Phan-Quang Gia Chuong and Dr. Han Xuemei provided inputs on data visualization.
- All authors gave approval to the final version of the manuscript.

Chapter 4 is published as Shi Xuan Leong[#], Yong Xiang Leong[#], Emily Xi Tan, Howard Yi Fan Sim, Charlynn Sher Lin Koh, Yih Hong Lee, Carice Chong, Li Shiuan Ng, Jaslyn Ru Ting Chen, Desmond Wei Cheng Pang, Lam Bang Thanh Nguyen, Siew Kheng Boong, Xuemei Han, Ya-Chuan Kao, Yi Heng Chua, Gia Chuong Phan-Quang, In Yee Phang, Hiang Kwee Lee, Mohammad Yazid Abdad, Nguan Soon Tan, and Xing Yi Ling. Noninvasive and Point-of-Care Surface-Enhanced Raman Scattering (SERS)-Based Breathalyzer for Mass Screening of Coronavirus Disease 2019 (COVID-19) under 5 min. *ACS Nano*, **16**, 2, 2629-2639 (2022). DOI: 10.1021/acsnano.1c09371.

[#] These authors contributed equally.

The contributions of the co-authors are as follows:

- Prof. Ling Xing Yi provided the initial project direction and supervised the research.
- Dr. Leong Shi Xuan, Dr. Sim Yi Fan Howard, Dr. Koh Sher Lin Charlynn, and Prof. Ling Xing Yi and I designed the study.
- Dr. Leong Shi Xuan, Ms. Tan Xi Emily, Dr. Sim Yi Fan Howard, Dr. Koh Sher Lin Charlynn, Dr. Lee Yih Hong, Ms. Chong Carice, Ms. Ng Li Shiuan, Ms. Chen Ru Ting Jaslyn, Mr. Pang Wei Cheng Desmond, Mr. Nguyen Bang Thanh Lam, Dr. Han Xuemei, Dr. Kao Ya-Chuan, Dr. Phan-Quang Gia Chuong, and Dr. Phang In Yee collected breath samples at the clinical trial sites.
- Dr. Leong Shi Xuan performed laboratory work at the School of Chemistry, Chemical Engineering and Biotechnology to establish the chemical basis of our study and analyzed the results.
- I performed data analysis for samples collected from the clinical trials, constructed, and fine-tuned the machine learning model.
- Dr. Leong Shi Xuan and I prepared the manuscript drafts. The manuscript was revised with inputs from Dr. Sim Yi Fan Howard, Dr. Koh Sher Lin Charlynn, Dr. Phan-Quang Gia Chuong, Dr. Han Xuemei, Asst. Prof. Lee Hiang Kwee, Dr. Mohammad Yazid Abdad, Assoc. Prof. Tan Nguan Soon, and Prof. Ling Xing Yi.
- All authors gave approval to the final version of the manuscript.

15 Jan 2024

. . . . . .

Date

. . . . . .

Leong Yong Xiang

*please sign on the water mark area.

Abstract

The advent of data science in many facets of science is a clear testament to its ability to revolutionize modern scientific discoveries. This thesis explores its application in surface-enhanced Raman scattering (SERS), a powerful spectroscopic technique that offers molecule-specific readout with high sensitivity. Despite having immense potential, practical SERS sensing applications remains hindered by poor surface affinities of the target analytes and complexity of the media they are present within. From a unique data science perspective, we design a strategy which leverages multiple molecular receptors that aim to induce receptor-analyte chemical interactions with different facets of the analyte at the plasmonic surface. The collective spectral output forms a holistic SERS 'super-profile' which accumulates all subtle variances embedded within and bolsters machine learning (ML) predictive models. Crucially, we demonstrate improved analyte specificities in detecting flavor compounds at the laboratory scale and breath volatile organic compounds in an actual clinical trial even in the presence of matrix interferences. To facilitate smart receptor selection, we introduce a ML-driven recommender system that maximizes SERS variance within the super-profile by selectively excluding excess uninformative features. Finally, we explore data augmentation techniques in overcoming class imbalance issues and construct robust predictive models that can be swiftly deployed for mass screenings during infectious disease outbreaks. Overall, these findings highlight the synergistic relationship between SERS and data science and are key in accelerating the practical translation of SERS sensors for diverse applications.

Acknowledgements

I would like to express my utmost gratitude to my supervisor Prof. Ling Xing Yi for the opportunity to pursue postgraduate studies in her research group. Under her astute guidance, I was able to solidify my research fundamentals and grow immensely as a budding researcher. I am grateful to the School of Chemistry, Chemical Engineering and Biotechnology, Nanyang Technological University for financially supporting my postgraduate studies and providing access to state-of-the-art laboratory equipment and facilities. I am also thankful for the support and timely feedback from my thesis advisory committee members – Prof. Xing Bengang and Dr. Alessandra Bonanni – throughout the course of my studies. In addition, I appreciate the insightful inputs from Dr. Phang In Yee as well as Assoc. Prof. Xia Kelin with regards to the application of machine learning in many aspects of my research.

It has been my privilege to work with many stellar individuals from the Ling group who have since left the group to pursue greater heights – Asst. Prof. Lee Hiang Kwee, Dr. Lee Yih Hong, Dr. Han Xuemei, Dr. Phan-Quang Gia Chuong, Dr. Koh Sher Lin Charlynn, Dr. Sim Yi Fan Howard, Dr. Kao Ya-Chuan, and Dr. Leong Shi Xuan. Our friendly banter and uninhibited exchange of ideas during lunch are fond memories which I will cherish for years to come. I would like to specially thank Dr. Koh Sher Lin Charlynn for being an incredibly cognizant and patient mentor, as well as Dr. Lee Yih Hong for being an articulate and attentive writing coach when guiding me in my very first publication. I am equally fortunate to be able to work with many exceptional juniors still in the Ling group – Ms. Tan Xi Emily, Mr. Nguyen Bang Thanh Lam, and Ms. Chen Ru Ting Jaslyn. I trust that each of you will continue to shine brightly and excel in whichever path you decide to take in future. On this note, I would also like to extend my appreciation to many talented undergraduates that I have worked with, and I hope I was able to spark your interest in scientific research with some of the exciting challenges we face.

This thesis is a milestone that I would never achieve without the wholehearted support of my loved ones. To my family – my parents, brother, sister, and grandmother, thank you for the unconditional love and unwavering trust in all my endeavors. I hope I made you guys proud. To my decade-long friends Kelvin, Wen Yu, and Jun Yan, thank you for making life outside of research so much fun during our gatherings. To my lab confidante Emily, thank you for listening to my occasional rants. To my love Felix, you are the biggest source of my strength throughout this journey and will continue to be as we head down the path to forever – I love you.

Finally, thank you for reading this thesis and I hope these findings will benefit you as much as it did for me.

Table of Contents

List of Abbreviations

2,4-DCA	2,4-dichloroanisole
2,6-DCA	2,6-dichloroanisole
3,5-DCA	3,5-dichloroanisole
2,4,6-TBA	2,4,6-tribromoanisole
2,4,6-TCA	2,4,6-trichloroanisole
2,3,5,6-TeCA	2,3,5,6-tetrachloroanisole
AEF	Analytical enhancement factor
AI	Artificial intelligence
airPLS	Adaptive iteratively reweighted penalized least-squares
ATP	4-aminothiophenol
AUC	Area under curve
BERT	Bidirectional encoder representations from transformers
$B(OH)_2$	4-mercaptophenylboronic acid
BPA	Bisphenol A
Br	4-bromothiophenol
BTP	4-bromothiophenol
BVOC	Breath volatile organic compound
CEM	Contrastive explanation method
CH_3	4-methylbenzenethiol
CHEM	Chemical
CHO	4-methylthiobenzaldehyde
CI	Confidence interval
CoD	Curse of dimensionality
COOH	4-mercaptobenzoic acid

COVID-19	Coronavirus disease 2019
CS	Chondroitin sulfate
Ct	Cycle threshold
DA	Discriminant analysis
DCGAN	Deep convolutional generative adversarial network
DFT	Density functional theory
DL	Deep learning
DMAB	4,4-dimercaptoazobenzene
DT	Decision tree
ECFP	Extended-connectivity fingerprints
EF	Enhancement factor
EM	Electromagnetic
ENN	Edited nearest neighbors
FWHM	Full width at half maximum
GAN	Generative adversarial network
GC-MS	Gas chromatography coupled mass spectrometry
ISOMAP	Isometric mapping
kNN	k-Nearest neighbors
LASSO	Least absolute shrinkage and selection operator
LC-MS	Liquid chromatography coupled mass spectrometry
LDI-MS	Laser desorption/ionization mass spectrometry
LIME	Local interpretable model-agnostic explanations
LSPR	Localized surface plasmon resonance
LV	Latent variable
MBA	4-mercaptobenzoic acid

MBT	4-methylbenzenethiol
MCC	Matthews' correlation coefficient
MH	3-mercaptohexanol
MHA	3-mecaptohexyl acetate
MIL	Material Institute Lavoisier
ML	Machine learning
MLP	Multi-layered perceptron
MOF	Metal-organic framework
MP	4-mercaptophenol
MPY	4-mercaptopyridine
MTBH	4-methylthiobenzaldehyde
NEG	COVID-negative
NH_2	4-aminothiophenol
NIR	Near-infrared
NLP	Natural language processing
NN	Neural network
NPV	Negative predictive value
NT	2-naphthalenethiol
OH	4-mercaptophenol
PAH	Polycyclic aromatic hydrocarbon
PC	Principal component
PCA	Principal component analysis
PCR	Polymerase chain reaction
PLS	Partial least-squares
PLSDA	Partial least-squares discriminant analysis

POS	COVID-positive
PPV	Positive predictive value
PY	4-mercaptopyridine
R6G	Rhodamine 6G
RF	Random forest
RMSE	Root mean-squared error
RS	Recommender system
SD	Standard deviation
SEM	Scanning electron microscopy
SERS	Surface-enhanced Raman scattering
SHAP	Shapley additive explanations
SMILES	Simplified molecular input line entry system
SMOTE	Synthetic minority oversampling technique
SNR	Signal-to-noise ratio
SVM	Support vector machine
SVMDA	Support vector machine discriminant analysis
SVMR	Support vector machine regression
SWCNT	Single-walled carbon nanotube
TS	Tanimoto similarity
t-SNE	t-distributed stochastic neighbor embedding
UMAP	Uniform manifold approximation and projection
VOC	Volatile organic compound
XGBoost	Extreme gradient boosting tree
ZIF	Zeolitic imidazolate framework

Chapter 1 Introduction: The Role of Data Science in SERS

Abstract. Surface-enhanced Raman scattering (SERS) is a powerful vibrational spectroscopic technique leveraging plasmonic nanoparticles to significantly enhance the Raman fingerprints of molecules. With high detection sensitivity, portable instruments, and the ability to provide near-instantaneous readout, it is unsurprising that SERS has been applied in the detection of numerous chemical and biological analytes across diverse applications. While promising, the poor analyte surface affinities and presence of interferences in the sample matrix severely limits the ability of SERS in applications beyond the laboratory scale. In this chapter, I first introduce the pivotal role of data science in addressing these problems through feature transformation, data augmentation and predictive modelling. Then, I summarize recent progress in harnessing data science approaches for rapid on-site predictions of infection diseases. To conclude this chapter, I outline the motivation and objectives of this thesis.

1.1 Surface-enhanced Raman scattering (SERS)

1.1.1 Principles of SERS

SERS is a surface-sensitive vibrational spectroscopic technique that employs plasmonic nanoparticles to significantly enhance the Raman fingerprints of molecules when they are close (< 10 nm) to the plasmonic surface (**Figure 1-1**).[1-2] These plasmonic nanoparticles can be metallic in nature (such as Ag, Au, Cu, Al) or semiconducting materials (for instance, metal oxides like ZnO and TiO_2 or single element semiconductors like graphene).[3-5] Fundamentally, Raman enhancement via plasmonic nanoparticles can be attributed to either or both the electromagnetic (EM) and chemical (CHEM) mechanisms.[6] For EM enhancement, an incident laser irradiation triggers the collective oscillation of the material's conduction band electrons that is locally confined on the surface of the nanoparticle – a phenomenon known as localized surface plasmon resonance (LSPR). Coupling of these LSPRs with the incident laser creates intense secondary electric fields that amplify both the excited and emitted radiation, accounting for 10^3 – 10^8-fold Raman signal enhancement. Notably, the strength of this EM enhancement is dictated by nanoparticle morphologies, its dielectric properties, and the availability of neighboring particles for inter-particle plasmonic coupling. For CHEM enhancement, the Raman signal improves up to 10^3-fold when the target molecules are chemically adsorbed on the plasmonic surface because their increased polarizability facilitates the charge transfer effect. Unlike the EM enhancement which may occur independently if analytes are merely physically confined on the plasmonic surface, the CHEM enhancement often occur together with the EM enhancement due to the need for chemical interactions. This concerted effect enables an overall Raman signal boost up to 10^{11}-fold, which is often key in allowing SERS to achieve detection of analytes down to the single particle level.[7-8] To date, SERS has demonstrated immense potential across chemical[9-10], biological[11-12], and environmental/food[13-15] applications.

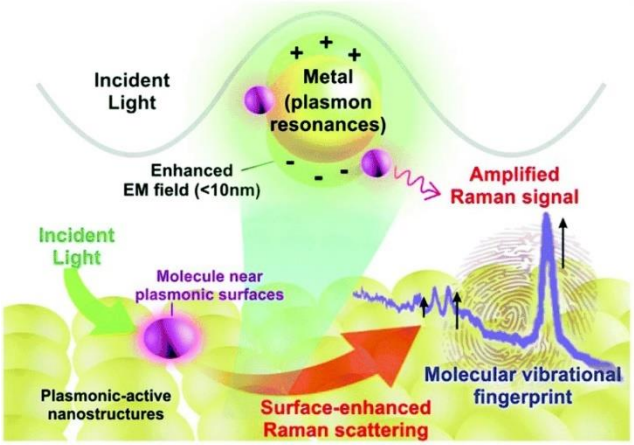


Figure 1-1. EM enhancement mechanism to boost Raman signals. Adapted with permission from ref. 6. Copyright 2019 Royal Society of Chemistry.

The overall Raman signal enhancement can be measured using either the enhancement factor (EF) or the analytical enhancement factor (AEF). Although both EF and AEF compare the SERS signal intensity against the Raman signal intensity, EF accounts for the number of molecules affected by the field enhancement whereas AEF relates to the analyte concentrations. In general, EF serves to benchmark SERS performances across different platforms, but it can be difficult to accurately estimate the number of molecules present within the laser excitation volume. AEF is hence a useful alternative to gauge their analytical performance. These metrics are calculated by:

$$\mathrm{EF} = \frac{\mathrm{I_{SERS}}}{\mathrm{I_{Raman}}} \times \frac{\mathrm{N_{Raman}}}{\mathrm{N_{SERS}}}$$

$$\mathrm{AEF} = \frac{\mathrm{I_{SERS}}}{\mathrm{I_{Raman}}} \times \frac{\mathrm{C_{Raman}}}{\mathrm{C_{SERS}}}$$

where I refers to the signal intensity, N refers to the number of molecules present within the laser excitation volume and C refers to the concentration of molecules in the sample.

1.1.2 Analyte confinement strategies

Since signal enhancement relies heavily on molecules being at or close to the plasmonic surface, an important hurdle faced in SERS-based analyte detection is the poor analyte affinity to the nanoparticles. To address this challenge, many physical and chemical strategies have been devised to confine analytes near the plasmonic surface. In this section, I will discuss two physical and two chemical confinement strategies in detail.

Physical confinement strategies refer to methods that aim to physically restrict analyte molecules at the nanoparticle surface, without relying on chemical interactions between them. These strategies are attractive as they are universally applicable to a diverse range of analytes. Since target SERS analytes are commonly present within an aqueous or polar solvent, one strategy involves controlling the surface wettability of substrates – such as superhydrophobic surfaces – so that analytes are physically concentrated within a small area when dried.[16] Such a surface is characterized by having an advancing contact angle of > 150 ° when a polar solvent droplet containing the sample is in contact with the SERS substrate.[17-18] This repelling effect can arise from physically roughened surfaces fabricated via lithographic means[19] or chemically induced by attaching ligands with hydrophobic functional groups onto the nanoparticles.[20] In fact, superhydrophobic surfaces also occur naturally, such as in lotus leaves or rose petals, due to the presence of unique hierarchical nanostructures. For example, one study utilized a Ag coated Taro leaf to harness its anti-water surface wetting properties, with a water contact angle of 154 ° as a result of ~20 µm micro-papillae on the leaf surface (**Figure 1-2**).[21] Using this SERS substrate, the authors demonstrated that a 4 µL water droplet containing Rhodamine 6G (R6G) dries in an area of 0.3 mm^2, which is 17-fold smaller than a regular hydrophobic substrate with a contact angle of 90 ° (~5 mm^2) and 43-fold smaller than a hydrophilic substrate with a contact angle of 30 ° (~13 mm^2). The Ag-coated Taro leaf SERS substrate exhibited an EF of ~10^6 and was able to attain a 10^{-8} M detection limit for R6G.

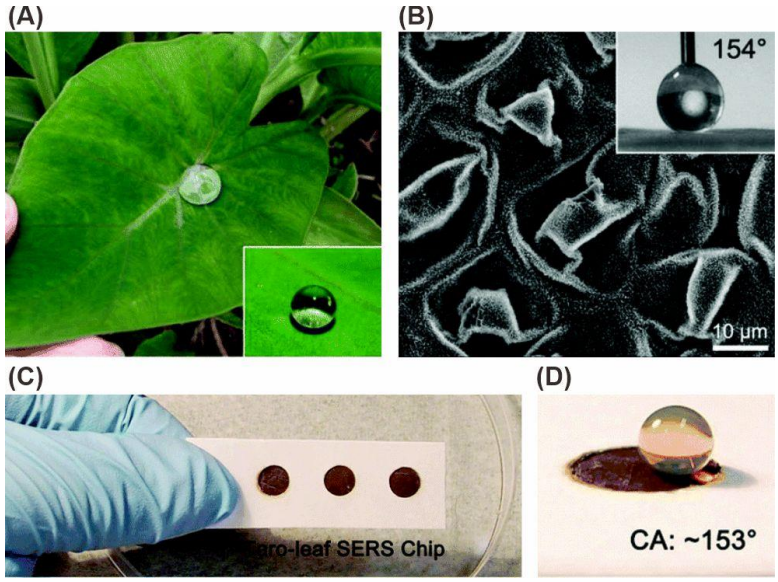


Figure 1-2. Superhydrophobic Ag coated Taro leaf SERS substrate. (A) Digital image of a Taro leaf. The inset shows a water droplet deposited on the leaf, showing its anti-water surface properties. (B) Scanning electron microscopy (SEM) image of the micro-papillae structures on the Taro leaf surface. (C, D) The Ag coated Taro leaf SERS substrate and its water contact angle. Adapted with permission from ref. 21. Copyright 2016 Royal Society of Chemistry.

Another physical strategy to confine analytes is to coat a layer of sorbent material atop the plasmonic nanoparticles that temporarily traps analyte molecules within them. Specifically, metal-organic frameworks (MOF) are excellent choices to fulfill this role given their highly porous crystalline structure formed by an extensive coordination network of metal ions and organic linker molecules.[22] Their pore sizes can be judiciously tuned by selecting different combinations of metal ions and organic linkers. For example, zeolitic imidazolate framework (ZIF)-8 with a Zn core and 2-methylimidazolate linkers and Material Institute Lavoisier (MIL)-101 with a Cr_3O trimer core and 1,4-benzenedicarboxylic acid linkers have pore apertures of 3.4 Å and 8.6 Å respectively.[23-24] As opposed to superhydrophobic SERS substrates which primarily target liquid or aqueous soluble analytes, MOF-SERS systems are proficient in improving gas-phase SERS sensing. This is important because the low analyte concentrations

in the gaseous phase coupled with the poor analyte affinity to the plasmonic surface is often a huge challenge to address. In one study, a ZIF-8 encapsulated Ag nanocube SERS platform was used to concentrate and detect toxic vapors such as toluene down to 200 ppm and polycyclic aromatic hydrocarbons (PAH) such as 2-naphthalenethiol (NT) down to 50 ppb, below their legal exposure limits (**Figure 1-3**).[25] Crucially, the authors highlighted the importance of the ZIF-8 coating in allowing a 5-fold amplification in SERS signal intensity in detecting other volatile organic compounds (VOC) like 4-methylbenzenethiol (MBT).

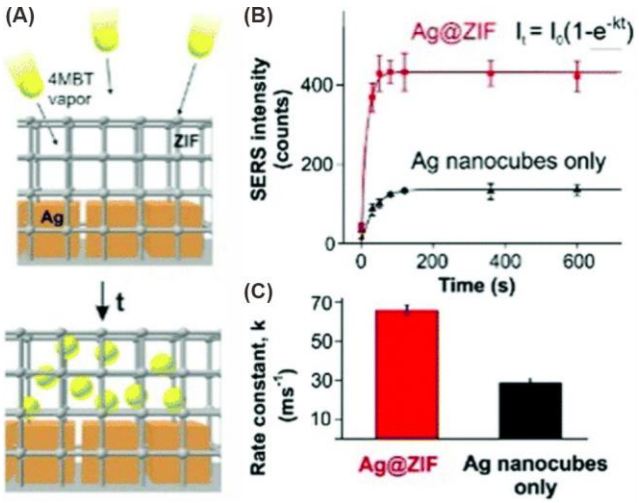


Figure 1-3. ZIF-8 encapsulated Ag nanocube SERS platform to detect toxic vapors, PAHs, and other VOCs. (A) The use of ZIF-8 to trap gaseous molecules close to the Ag nanocubes. (B) Time-dependent SERS analysis comparing platforms with (red) and without (black) the ZIF-8 film upon incubation of MBT vapor. (C) Rate constants of the respective platforms based on the MBT characteristic SERS peak at 1079 cm^{-1}. Adapted with permission from ref. 25. Copyright 2018 Royal Society of Chemistry.

While both strategies have shown promising potential across many applications, it is important to recognize their drawbacks. An obvious concern for superhydrophobic surfaces is their compatibility with non-polar liquids/solvent, but this can be mitigated by introducing

omniphobic properties onto the superhydrophobic surfaces.[26] Aside from this issue, two major problems remain with physical confinement strategies. First, if the analytes inherently have small Raman cross-sections, the SERS enhancement may still be lacking even if they are near the plasmonic surface. Next, the selected hydrophobic/omniphobic chemical moiety or MOF cannot contain conflicting Raman signals as they will also be amplified and potentially mask the desired analyte signals.

In these aspects, chemical confinement strategies are appealing alternatives with two primary modalities. In the first approach, a chemical coupling reaction is induced to enlarge the Raman cross-section of the target analyte. Common reactions include azo coupling between a diazonium and an aromatic compound or the click reaction between an azide and an alkyne. For instance, bisphenol A (BPA) is an environmental toxin with no affinity to Ag and hence no distinguishable SERS signal.[27] However, when azo coupled with an aryl-diazonium cation, a SERS peak at 1406 cm^{-1} attributed to the N=N stretching mode emerges, which allows monitoring of BPA down to 10^{-17} M (> 10^5-folds lower than the regulatory limit) (**Figure 1-4**). Such high sensitivity stems from the Raman cross-section enlargement as the final product shows extended resonance compared to BPA.

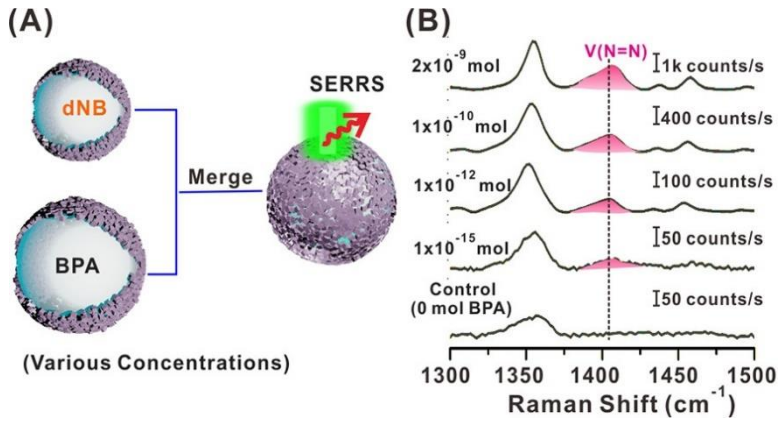


Figure 1-4. (A) Merging two plasmonic liquid marbles to induce the azo coupling reaction between diazonium nitrobenzene and BPA. (B) SERS spectra of the resultant compound

showing the emergence of the N=N stretching mode at 1406 cm^{-1}. Adapted with permission from ref. 27. Copyright 2017 American Chemical Society.

Although chemical coupling is highly sensitive and analyte-specific, it is often not a practical option as the scope of analytes available for detection is severely limited by the nature of the coupling reaction. This method also limits the ability to detect multiple analytes within the sample matrix concurrently. Hence, receptor-driven SERS is a promising alternative that has garnered increased interest in recent years. This strategy makes use of aromatic molecules with (1) an 'anchor' component that forms strong covalent bonds with plasmonic materials and (2) an 'active' component that promotes generic intermolecular interactions with functional groups present on the target analyte(s) (**Figure 1-5**).[6] The anchor often manifests as a free thiol or silane group that forms strong M-S (with metals such as Ag, Au) or M-O-Si bonds (with metals such as Cu, Al with naturally occurring oxide layers). The active functional group can then induce hydrogen bonding, electrostatic, aromatic donor-acceptor or van der Waals' interactions depending on the functional groups present on the analyte moieties.

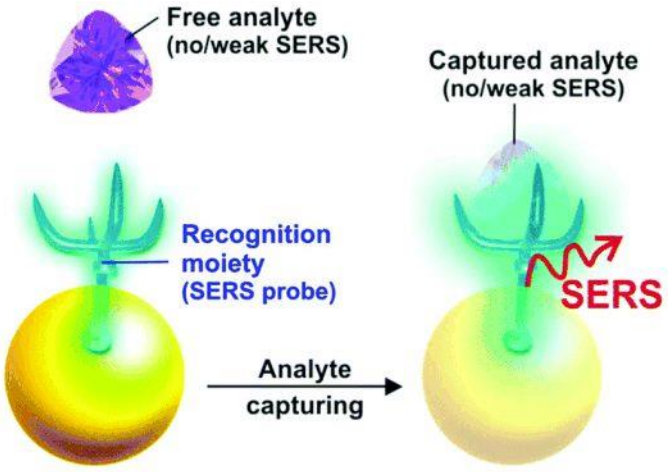


Figure 1-5. Mechanism of receptor-driven SERS sensing. Adapted with permission from ref. 6. Copyright 2019 Royal Society of Chemistry.

Importantly, the receptor SERS signals are inherently strong due to its aromaticity and hence can function as a transducer to indicate the presence of analytes through its signal variations. While this does not provide direct indication of the analyte's molecular structure, its unique structural and electronic properties will affect the receptor-analyte interaction orientation and strength. This results in slight differences in the acquired SERS signal variations between two analytes, thereby differentiating them. Since the focus is now on the receptor SERS signals, concerns over the analyte Raman cross-sections as well as conflicting Raman signals from the molecular receptors are no longer an issue. Moreover, because the intermolecular interactions are generic, a receptor can interact with more than one type of analyte. Recently, a receptor-driven SERS platform based on 4-mercaptopyridine (MPY) was used to differentiate four isomeric chondroitin sulfate (CS) disaccharides (**Figure 1-6**).[9] The authors leveraged a 'charge and shape complementarity' approach to induce site-specific multidentate electrostatic interactions between MPY and each CS disaccharide and was able to attain > 97% accuracy in classifying all four CSs along with five other interferences with close structural similarities. This study highlights the potential of receptor-driven SERS in both chemical analyte confinement as well as analyte-specific differentiation.

Among the four physical and chemical analyte confinement strategies discussed above, receptor-driven SERS sensing carries the most potential for practical translation. Aside from addressing problems associated with small analyte Raman cross-sections, potential interfering platform SERS signals and overly strict receptor-analyte pairing requirements, receptor-driven SERS platforms are compatible with many plasmonic materials, assembly techniques and are simple to fabricate. Many receptor-driven SERS platforms have also demonstrated the ability to conduct multiplex analyte sensing amidst sample matrices containing interfering molecules, which is crucial in meeting the demands of complex real-world applications.

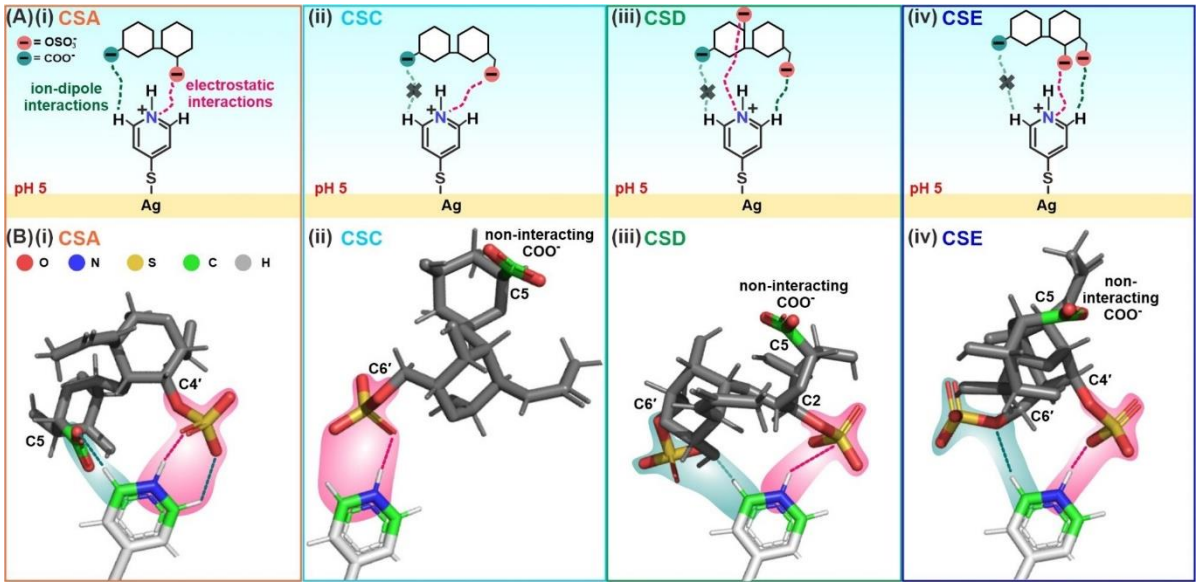


Figure 1-6. Energy-optimized MPY^{+}/CS^{-} interaction complexes of four CS disaccharides, showing site-specific multidentate interactions formed because of 'charge and shape complementarity' with the MPY receptor. Adapted with permission from ref. 9. Copyright 2023 Wiley-VCH GmbH.

1.2 Data science in SERS

1.2.1 Chemometrics and data science

Analyte classification and/or quantification are the main goal(s) of most SERS-based sensors when applied in practice. These sensors can be used to determine the amount of disease biomarkers in biofluids for biomedical diagnostics or toxic gases in the air for environmental surveillance.[11,13] To do so, the experimentally collected SERS spectra must undergo a series of data processing and analyses to identify key peak variations in the form of position shifts, intensity fluctuations or shape changes. Traditionally, this involves manually looking at the peaks of SERS spectra before and after analyte exposure to compare their differences. The inefficient and subjective nature of such manual methods naturally prompts the exploration of more objective and robust methods for SERS data analysis. In 1995, the term 'chemometrics' was first coined by Svante Wold, referring to the area in chemistry that deals with the extraction

of chemical information from raw experimental data.[30] Wold proposed two important points. First, he proposed that multivariate models – citing principal component analysis (PCA) and partial least-squares (PLS) as examples – can derive unexpected patterns by considering the joint effect of all input variables, as opposed to traditional analyses where only one or very few variables are studied at any time. This ensures the most amount of chemical information can be extracted from the raw data and resolves any potential conflicts amongst the input variables. Next, he proposed that the entire process of data analysis should closely conform with our knowledge of chemistry. This prevents conclusions that are scientifically unsound and advises against over-reliance on multivariate models.

Today, the incorporation of computer science, mathematics, and statistics to analyze data is now more formally termed as data science and can be applied to any domain within and beyond the natural sciences. Notably, this shift in terminology reflects the expansion in the range of techniques used – including advanced artificial intelligence (AI) and machine learning (ML) algorithms – as well as its generalizability across different domains. Nevertheless, the fundamental points proposed by Wold with regards to chemometrics remains highly relevant. In chemistry, data science manifests itself in three main areas – (1) autonomously navigating experimental parameter spaces to drive scientific discoveries, (2) designing optimized workflows using feedback from past experiments, and (3) constructing predictive models based on experimental data (**Figure 1-7**).[31] These areas are expansive and covers almost all aspects of chemistry. In the following sections, I will cover some of the data science techniques that are applicable in SERS data analytics and showcase some recent achievements in SERS-based sensing with emphasis on the pivotal role of data science.

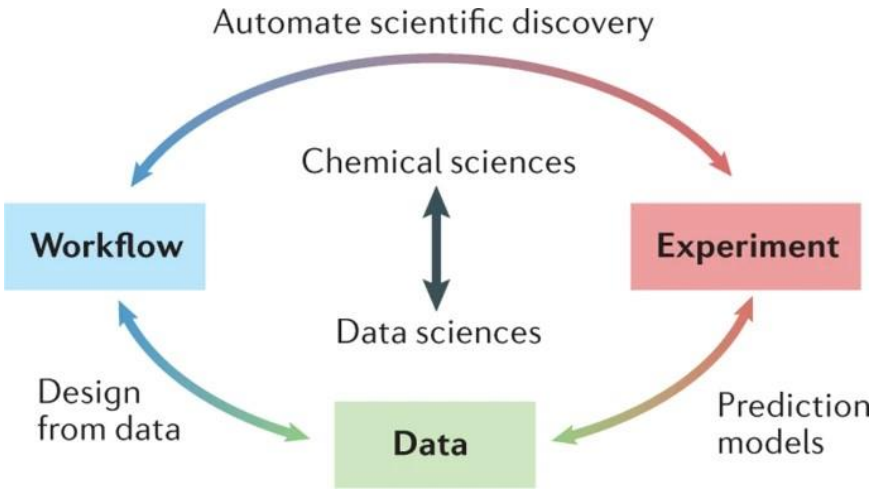


Figure 1-7. Relationship between data science and chemistry. Adapted with permission from ref. 31. Copyright 2022 Springer Nature.

1.2.2 Data science strategies

Data science involves the study of data to extract meaningful insights. Broadly, this includes all strategies related to the handling and processing of data, such as data storage, transformation, augmentation, modelling, and visualization (**Figure 1-8**). For this thesis, I will mainly focus on data transformation, augmentation and modelling as these areas are more relevant to the research and development of SERS-based sensors.

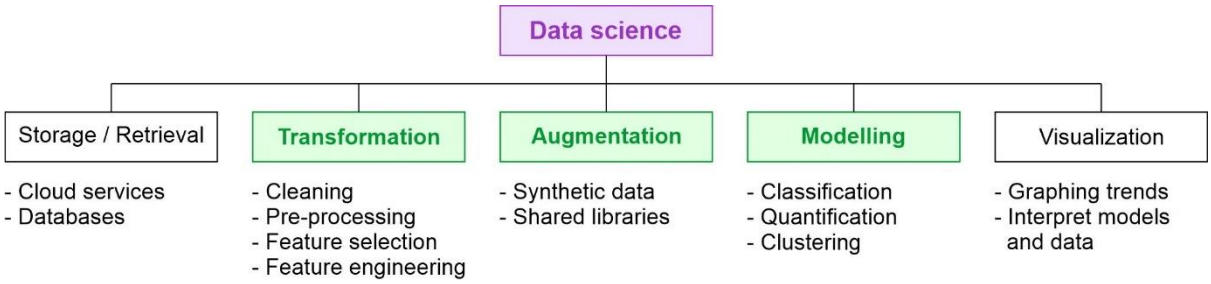


Figure 1-8. Data science strategies related to SERS, with discussion focus on transformation, augmentation, and modelling data.

Data transformation techniques aim to convert raw data into a structured format that is ready for analysis. It is almost always advantageous to apply data transformation techniques prior to predictive modelling as the raw data often contains information in an unstructured or sometimes misleading manner. This is akin to trying to find a specific book in a library, but all the books are scattered on the ground. What data transformation does is to first rearrange all the books back on shelves in an orderly fashion, so that we can find the book we want easily. In SERS, data transformation includes data cleaning, pre-processing, feature selection and feature engineering. Here, data cleaning is intuitive – we simply remove spectra that are erroneous. For example, if the SERS spectrum shows abnormally low signal-to-noise ratio (SNR) or does not contain signals we typically expect such as those from molecular receptors, it should be removed. It is crucial that these erroneous spectra do not create outliers that mislead and potentially skew the predictive model. However, removal is not always necessary if the error can be corrected. For example, cosmic ray signals are single wavenumber spikes that contribute large but meaningless differences between two SERS spectra as they originate from exposure of cosmic rays to the detector (**Figure 1-9A**).[32] This can significantly skew multivariate models such as PCA as all input Raman wavenumbers are of equal importance. Nevertheless, these cosmic ray spikes can be easily removed by outlier detection metrics such as the Hotelling's t-squared, Mahalanobis distance or *Q* residuals to preserve the otherwise useful sample data.[33] These metrics are widely available with most modern Raman measurement software as built-in processing tools.

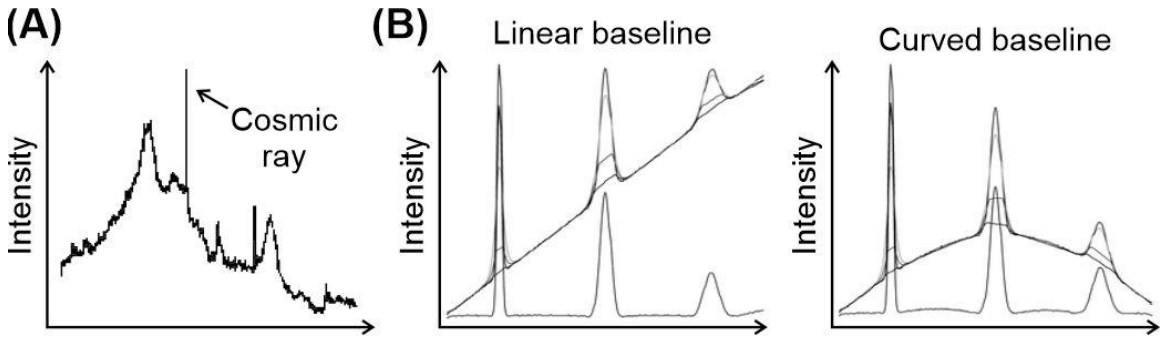


Figure 1-9. Data pre-processing methods. (A) Cosmic ray signal. (B) Linear and curved baseline correction using the adaptive iteratively reweighted penalized least-squares algorithm. Adapted with permission from ref. 32 and 34. Copyright 1990 American Chemical Society and 2010 Royal Society of Chemistry.

Data pre-processing refer to the use of mathematical and statistical tools to treat raw data and is often essential when dealing with SERS spectra. This is because an acquired SERS spectrum is heavily influenced by numerous external factors such as measurement conditions, instrument response, as well as the type of substrate and sample matrix. Often, variations may arise due to inevitable inconsistencies in the external environment instead of the actual relationship with the target analyte(s), which will confuse both humans and ML algorithms. While rigorously designed and executed experiments and well-calibrated Raman spectrometers will alleviate these external influences, it is impossible to completely remove them. It is crucial to ensure that these inconsistencies do not manipulate the decision-making models which rely heavily on the input data being representative of the underlying chemical meaning. Baseline correction, normalization, and smoothing are the most common techniques employed to correct these inconsistencies pertaining to SERS spectra (**Figure 1-9B**). Aside from rectification, data pre-processing is also important in structuring data in a manner that best reflect similarities or differences between samples. These methods aim to process raw SERS spectra such that variations become more pronounced. Examples include taking derivatives to highlight shifts in

peak centers and incorporating clutter removal to mask uninformative signals. A summary of common pre-processing techniques pertaining to SERS data analysis is shown in Table 1-1 below.

Table 1-1. Summary of different data pre-processing techniques for SERS spectral processing.

Category	Purpose	Methods
Baseline correction	Remove the floating spectral baseline	• Weighted least-squares • Polynomial fitting[34] • Adaptive iteratively reweighted penalized least-squares[35] • Asymmetric least-squares[36]
Normalization	Signal standardization across samples	• Simple normalization • Area under curve • Standard normal variate[37]
Smoothing	Removes spectral noise	• Savitzky-Golay filtering[38] • Maximum likelihood estimation[39]
Scaling and Centering[40]	Correcting offsets in data within sample (scaling) and across samples (centering)	• Autoscale • Centroid scaling • Mean/Median centering • Class centering
Derivative	Identifies critical turning points in the spectra	• First derivative[41] • Second derivative
Clutter removal	Apply filters to identify features that should be down-weighted or removed	• Generalized least-squares weighting[42] • External parameter orthogonalization[43]

Feature engineering refers to the addition, deletion, combination, or mutation of SERS data with the goal of improving ML model performance. Feature selection, which is essentially targeted deletion of data, is the most common technique adopted and is sometimes discussed as a standalone topic. Formally, feature selection is the process where a subset of relevant SERS features is selected for subsequent input in predictive models. While optional, feature selection provides several key advantages. First, assuming no weights are added to the input SERS features, feature selection will place increased emphasis on the selected features. Since SERS peaks do not occur at every point along a spectrum, it is meaningful to select regions containing peaks of interest and in doing so allow ML algorithms to focus on studying peak variations as opposed to noise. Next, feature selection effectively reduces the dimensionality of SERS data. This is especially useful for high dimensional spectral data, which are typically represented with thousands of individual features. If left unchecked, a phenomenon known as the curse of dimensionality (CoD) may cause an accuracy decrease in the resulting predictive model.[44-45] For SERS, this is because the excess number of uninformative features hinder the precise retrieval of key variances within the spectra. At present, there is no set rule when conducting feature selection. An intuitive method is to assess the importance of all input features and select features above a set threshold for predictive modelling. In one study, this method was applied to select SERS peaks important in the qualitative trace detection of PAHs, attaining an improvement in detection sensitivity from 5 ppb to 0.1 ppb.[46] Although promising, it is noteworthy that feature selection should be exercised with caution as the exclusion of features may result in a skewed model even if the results concur with one's expectations. It is hence crucial to corroborate both experimental and theoretical SERS variances with ML-identified important variables to ensure the ML model is built upon chemically relevant information.

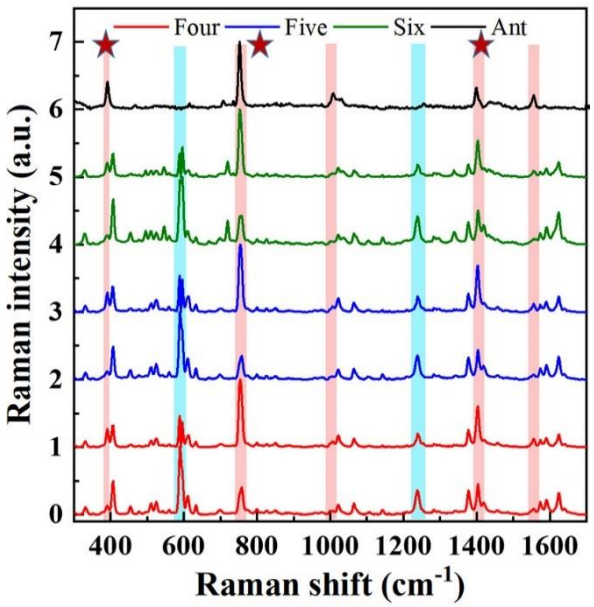


Figure 1-10. Identifying and selecting important SERS peaks related to anthracene detection. The red stars indicate the top three peaks at 398 cm^{-1} (importance = 1), 752 cm^{-1} (importance = 0.7), and 1398 cm^{-1} (importance = 0.3). Adapted with permission from ref. 46. Copyright 2022 American Chemical Society.

Aside from feature selection, combination of features is an emerging area pertaining to feature engineering in SERS data analytics. Specifically, this includes the combination of (1) several sources of SERS data – thereby constituting an array-like SERS platform (**Figure 1-11A**), and (2) SERS data with data from other analytical methods – thereby constituting a hybrid multimodal analytical method (**Figure 1-11B**).[47] In this thesis, I am interested in leveraging array-based SERS detection, where multiple receptors are used to detect the same analyte(s). The strategies will be further detailed in section 1.3 below.

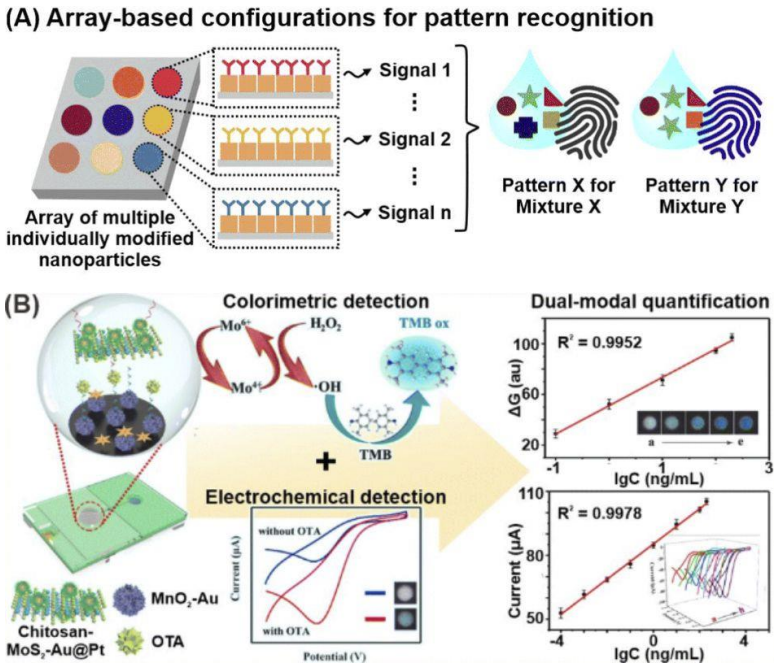


Figure 1-11. Combination of features arising from (A) several sources of SERS data – array-based SERS detection, and (B) SERS data with data from other analytical techniques – hybrid multimodal method. Adapted with permission under a Creative Commons CC BY License from ref. 47. Copyright 2022 Royal Society of Chemistry.

Unlike transformative techniques, data augmentation strategies primarily aim to boost ML training data by artificially creating modified copies of existing data. The reasons to do so are multifold. First, it is practically impossible to collect large amounts of samples from sources which are difficult to access, such as patients with cancers or rare disorders. Next, it is often resource-intensive to collect large amounts of data – in terms of cost, materials, manpower, and time. However, many complex ML algorithms such as neural networks, demand large training data sets to fully elucidate complex relationships between analytes and/or interferences but can offer unparalleled predictive prowess if successfully constructed. Augmenting small samples is a middle ground that takes advantage of these complex ML algorithms while reducing the cost of data collection. In one study, several augmentation techniques were compared for their ability to amplify the Raman spectra of serum samples obtained from 44 lung cancer patients

(**Figure 1-12**).[48] These techniques include adding random noise, creating baseline offsets, using the synthetic minority oversampling technique (SMOTE), and using generative adversarial networks (GAN).[49-51] Interestingly, they found that the fused dataset, which comprises data from all five augmentation strategies, showed the best performance across almost all of the classification models tested. This could be because the combination of augmentation methods effectively mimicked the high variability of actual serum samples, and this cannot be reflected by any single technique alone. More importantly, this shows that augmentation strategies can boost predictive accuracies of complex ML models in the event where real data is limited.

Dataset	Classification methods					
	SVM_{RBF}	SVM_{Linear}	SVM_{Poly}	CNN	ResNet	$\bar{u}$
Original dataset	74.67%	79.70%	75.56%	84.15%	86.37%	**80.09%**
Adding noise	75.70%	81.78%	79.85%	85.49%	88.15%	**82.19%**
Offset, slope, multiplication	76.44%	82.22%	78.96%	86.22%	87.70%	**82.31%**
SMOTE	77.19%	81.33%	81.19%	86.67%	88.89%	**83.05%**
GAN	80.44%	79.70%	77.04%	87.70%	89.04%	**82.78%**
DCGAN	78.67%	82.22%	78.37%	85.04%	87.41%	**82.34%**
Fused dataset	81.48%	84.59%	81.78%	87.11%	90.07%	**85.01%**

Figure 1-12. Classification accuracies obtained using the original and augmented Raman data to predict lung cancers using serum samples from patients. Adapted with permission from ref. 48. Copyright 2022 Elsevier.

1.2.3 Recent advances in disease detection

The most critical motivation of utilizing AI and data science in SERS data analytics is to harness their ability to provide predictions for future data which are not part of the existing data set – also known as predictive modelling. Here, we highlight recent advances in infectious disease detection using nanosensors with emphasis on the pivotal role of various ML predictive models.[52] Notably, an established ML model translates raw signal outputs of the nanosensors into comprehensible results reflecting the infection status of an individual at the point of care. This is pertinent when considering the nature of infectious diseases such as the recent Coronavirus Disease 2019 (COVID-19) outbreak. To achieve this, ML algorithms incorporate various methods to isolate and deconvolute complex signals acquired by nanosensors in response to each biomarker, while assessing the overall signal variation in response to the entire panel of biomarkers. In this context, the ML algorithms used are typically supervised in nature and can be broadly classified into several main concepts (1) regression-based algorithms, (2) support vector machines (SVM), (3) tree-based ensembles, and (4) neural networks (NN) (**Figure 1-13**).

Regression-based ML models aim to relate input features to dependent variables through linear or nonlinear functions, making them useful to provide quantitative analyses of biomarkers. For example, PLS regression was used to construct a calibration curve based on input SERS spectral data to quantify the amount of pregnane in urine samples.[11] The PLS model quantified pregnane in 40 patient urine samples with minimal error, between 0 – 3.1% against standard liquid chromatography-mass spectrometry (LC-MS) analyses. Extensions to the PLS algorithm allow it to conduct discriminant analysis (DA) for binary classification tasks by projecting features as latent variables (like principal components) separated by a linear discriminator which demarcates the two classes.[53-54] For instance, PLSDA was used to distinguish 501 breath samples of healthy individuals from individuals infected with COVID-

19 using high dimensional multi-receptor SERS spectral data.[12] The PLSDA model achieved an average classification sensitivity of 96.2% and specificity of 99.9%, even among asymptomatic individuals. Other notable regression-based ML algorithms include logistic, sparse, elastic-net and the least absolute shrinkage and selection operator (LASSO) regression algorithms.[55-58] Overall, regression-based models offer a simple and intuitive approach to classify and quantify biomarkers as the correlation is based on statistical principles which are well-understood. While most regression algorithms can be generalized or extended to accommodate nonlinear feature-to-variable relationships, ML algorithms which are inherently capable of handling such nonlinearity are often preferred.

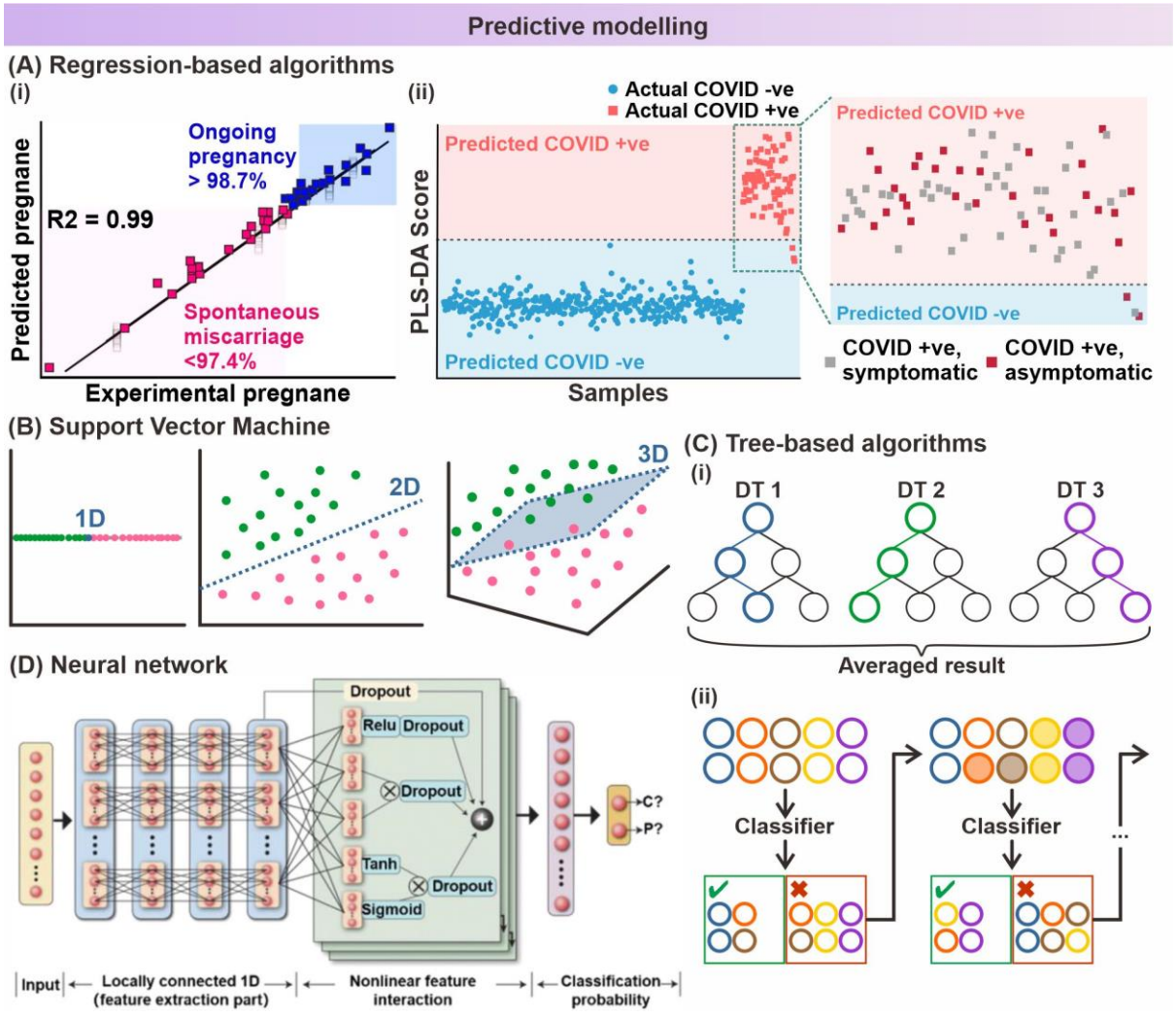


Figure 1-13. Various predictive models and their applications towards disease detection. (A) Regression-based algorithms – (i) PLS regression, (ii) PLSDA, (B) SVM, (C) Tree-based algorithms – (i) Decision trees, (ii) Gradient boosted trees. (D) NN. Adapted with permission from ref. 52. Copyright 2022 American Chemical Society.

SVM is one example of ML algorithms which can handle both linear and nonlinear relationships in complex biological matrices through the judicious selection of kernel functions (such as radial basis functions), which map features into a transformed high dimensional feature space.[59-60] Briefly, SVM identifies the separating hyperplane that maximizes the distance to the nearest data point of any class within a given feature space to achieve classification and regression.[61] For example, SVM enabled the simultaneous classification of 12 VOCs with 81.7% accuracy based on electronic nose signals recording the optical transmittance of liquid crystal-ionic liquid droplets embedded within a biopolymeric matrix.[62] Notably, nonlinear ML algorithms such as SVM can be highly effective in distinguishing structurally similar biomarkers, which exhibit competitive or interdependent interactions with nanosensors in conjunction with tailored surface functionalities. For example, structurally similar molecules like α- and γ-thrombin were classified with a positive predictive value of 97.2% and 99.0%, respectively, using aptamer functionalized Si nanopores and measuring the protein translocation events.[63] Overall, SVM presents an attractive option especially for nonlinear feature relationships, which is common when dealing with complex biological matrices or target biomarkers of very similar molecular structures. However, depending on the matrix complexity and nature of the biomarkers of interest, regression or SVM models, which rely on single estimators may be insufficient to elucidate the intricate feature relationships.

In contrast to single model-based algorithms (regression, SVM), tree-based ensemble algorithms provide an alternative approach because they construct numerous models to address

the classification or regression problem. The central idea behind such an ensemble approach is to create a single resulting model from the combination of all the individually constructed models, such that the resulting model outperforms any of the individual entities, with higher robustness and generalizability.[64] One way to combine all the constructed models is simply to average the predictions from each model – such as the creation of a random forest (RF) model using numerous independent decision tree (DT) models.[65] For example, a RF classifier predicted 91 proteins that adsorb to single-walled carbon nanotubes with 78% accuracy and 70% precision, thereby identifying proteins with high binding affinities.[66] Another way is to sequentially combine each constructed model, such that a new model is trained with the knowledge of the entire error that the whole ensemble has learnt so far – such as gradient boosting techniques.[67] For example, an extreme gradient boosted tree (XGBoost) algorithm trained on fluorescent microscopy tracking of hundreds to thousands of poly(ethylene glycol)-coated polystyrene nanoparticles effectively predicted the neurodevelopmental age of rats with 86.64% accuracy by studying changes in the brain extracellular matrix.[68] Overall, tree-based ensemble algorithms have immense potential in dealing with complicated feature relationships arising from concurrent monitoring of multiple biomarkers due to their inherent ability to form robust and generalizable models.

Alternatively, NNs can fulfill similar roles to tree-based ensemble algorithms in deciphering subtle yet complex biomarker relationships through mimicking biological neural networks with artificial neurons, connectors, and layered architecture. One common variant of NNs are feed-forward networks, where artificial neurons are arranged with unidirectional connections between them and each response to an input is independent of the previous outcome.[69] Perceptron-based algorithms are a classic example of feed-forward NNs, which isolate individual signals by employing an array of nonspecific receptors to capture certain features before analyzing the ensemble response, thus simulating human brain perception.[70]

For example, a perceptron-based algorithm was used to diagnose ovarian and endometrium cancers based on near-infrared (NIR) spectroscopic data of a photoluminescent sensor array comprising 132 DNA-wrapped single-walled carbon nanotubes (SWCNT) with varying chiralities.[71] Importantly, the perceptron-based model classified three key biomarkers (CA-125, HE4, and YKL-40) within uterine lavage samples with 91 – 100% accuracy even in sub-nanomolar ranges, signifying the strong potential for point-of-care applications. Another popular variant of NNs are deep learning (DL)-based networks, which incorporate more layers than the general NN structure to achieve progressive extraction of key features that improve model performance and illuminate potentially causal relationships between them.[72] For instance, a DL algorithm was used to study 20 serum metabolite fingerprints arising from 6 differential permutated metabolic pathways within native serum for swift and accurate diagnosis of stroke.[73] This is crucial in allowing the model to attain a classification sensitivity of 88.24%, specificity of 80%, and area under curve (AUC) of 0.845 via the analyses of nanoparticle-assisted laser desorption/ionization mass spectrometry (LDI-MS) signal data. Overall, neural networks offer a high degree of user customizability through tailored network design for specific applications, which can be advantageous to tree-based ensemble algorithms. However, they do require a much larger input data set (typically above thousands) to form robust models and are comparatively more computationally expensive. Collectively, all these examples explicate ML algorithms as a powerful and versatile toolkit for the thorough interrogation of a selected panel of biomarkers to enable precise detection of infectious diseases.

1.3 Thesis motivation and objectives

The proliferation of data science in many aspects of nanosensing is well-justified by the widespread success in many applications beyond disease detection. However, for SERS, most remain successful only at the laboratory scale. The central aim of this thesis is therefore

to explore a data science-centric approach in harnessing the full potential of SERS and drive SERS towards practical sensing applications.

In the first section of this chapter, I highlighted numerous advantages with receptor-driven SERS sensing. While promising, its practical translation is hindered by the occasional inability to distinguish analytes and interferences with highly similar molecular structures. The root cause of this problem is the indirect nature of such sensing methods, where the presence of the analyte is associated with fluctuations in the strong receptor SERS signals. With similar analytes that interact with receptors in similar orientations and/or strengths, it is unsurprising that the resultant receptor signal variations can be virtually indistinguishable in some cases. To address this concern, I propose an array-based sensor harnessing both direct and indirect sensing with multiple functionalized molecular receptors in Chapter 2. This concept entails accumulating variances across all molecular receptors by concatenating their resulting SERS spectra so that the collective SERS 'super-profile' has enhanced analyte specificity. Spectral concatenation is essentially a data transformation and feature engineering strategy that allows the ML model to pick up variances from all input SERS features. Crucially, we illustrate the distinct improvement in flavor classification and multiplex quantification using a traditional single receptor platform and our array-based multi-receptor platform.

With an array-based platform, selection of receptors becomes a key consideration as it will directly influence its performance. At present, selection is based on chemical intuition and prior practical experiences or references from literature. However, determining the number and type of receptors in this manner is subjective and cannot be optimized. In Chapter 3, I propose a data-driven recommender system approach to objectively and efficiently select the optimal number and type of molecular receptors tailored to specific classification problems. The system adopts a four-stage 'identify, filter, rank and recommend' approach to attain smart receptor selection by assessing their relative importance in resolving the problem at hand. In addition,

the optimized receptor-problem pairings can be logged within a recommender database that allows extensions to ‘unseen’ problems using a collaborative filtering approach.

In Chapter 4, I showcase the application of an array-based multi-receptor platform in the differentiation breath profiles of individuals infected with COVID-19. The platform detects fluctuations in the volatile organic compound composition within exhaled breath with the aid of a ML predictive model. In the context of an epidemic / pandemic outbreak, our SERS-based sensor is highly suited as a mass screening tool that provides on-site results in a rapid and non-invasive fashion. This is the first time an array-based multi-receptor SERS sensor is deployed in an actual infectious disease outbreak with promising clinical trial results.

The success achieved in the clinical trial, albeit small-scale, motivates further data collection with the aim to refine our ML model. For mass screening applications, the imbalance in positive and negative classes coupled with the relative inaccessibility of positive samples in practice presents a significant challenge in the process of ML model construction. Therefore, in Chapter 5, I explore data augmentation to synthetically amplify the proportion of COVID-positive sample data. In doing so, we also investigate the optimal data size and the class ratio when constructing a robust ML predictive model. This is crucial in addressing concerns on how swiftly a robust ML model can be constructed in the face of an infectious disease outbreak.

Finally, I conclude my thesis by summarizing key findings from chapters 2 to 5 and offer my perspective in the future of SERS with data science. I envision that data science holds the key in harnessing the decades of research efforts in SERS to realize practical applications as a standalone or complementary analytical technique across diverse fields.

References

1. Cardinal, M. F.; Vander Ende, E.; Hackler, R. A.; McAnally, M. O.; Stair, P. C.; Schatz, G. C.; Van Duyne, R. P., *Chem. Soc. Rev.*, 2017, 46, 3886-3903.
2. Ding, S.-Y.; You, E.-M.; Tian, Z.-Q.; Moskovits, M., *Chem. Soc. Rev.*, 2017, 46, 4042-4076.
3. Álvarez-Puebla, R. A., *J. Phys. Chem. Lett.*, 2012, 3, 857-866.
4. Schlücker, S., *Angew. Chem., Int. Ed.*, 2014, 53, 4756-4795.
5. Wang, X.; Zhang, E.; Shi, H.; Tao, Y.; Ren, X., *Analyst*, 2022, 147, 1257-1272.
6. Lee, H. K.; Lee, Y. H.; Koh, C. S. L.; Phan-Quang, G. C.; Han, X.; Lay, C. L.; Sim, H. Y. F.; Kao, Y.-C.; An, Q.; Ling, X. Y., *Chem. Soc. Rev.*, 2019, 48, 731-756.
7. Goh, M. S.; Lee, Y. H.; Pedireddy, S.; Phang, I. Y.; Tiiu, W. W.; Tan, J. M. R.; Ling, X. Y., *Langmuir*, 2012, 28, 14441-14449.
8. Liu, Y.; Pedireddy, S.; Lee, Y. H.; Hegde, R. S.; Tiiu, W. W.; Cui, Y.; Ling, X. Y., *Small*, 2014, 10, 4940-4950.
9. Leong, S. X.; Kao, Y.-C.; Han, X.; Poh, Z. W.; Chen, J. R. T.; Tan, E. X.; Leong, Y. X.; Lee, Y. H.; Teo, W. X.; Yip, G. W.; Lam, Y.; Ling, X. Y., *Angew. Chem. Int. Ed.*, 2023, e202309610.
10. Leong, S. X.; Koh, C. S. L.; Sim, H. Y. F.; Lee, Y. H.; Han, X.; Phan-Quang, G. C.; Ling, X. Y., *ACS Nano*, 2021, 15, 1817-1825.
11. Kao, Y.-C.; Han, X.; Lee, Y. H.; Lee, H. K.; Phan-Quang, G. C.; Lay, C. L.; Sim, H. Y. F.; Phua, V. J. X.; Ng, L. S.; Ku, C. W.; Tan, T. C.; Phang, I. Y.; Tan, N. S.; Ling, X. Y., *ACS Nano*, 2020, 14, 2542-2552.
12. Leong, S. X.; Leong, Y. X.; Tan, E. X.; Sim, H. Y. F.; Koh, C. S. L.; Lee, Y. H.; Chong, C.; Ng, L. S.; Chen, J. R. T.; Pang, D. W. C.; Nguyen, B. T. L.; Boong, S. K.; Han, X.; Kao,

Y.-C.; Chua, Y. H.; Phan-Quang, G. C.; Phang, I. Y.; Lee, H. K.; Mohammad, Y. A.; Tan, N. S.; Ling, X. Y., *ACS Nano*, 2022, 16, 2629-2639.

13. Nguyen, B. T. L.; Leong, Y. X.; Koh, C. S. L.; Leong, S. X.; Boong, S. K.; Sim, H. Y. F.; Phan-Quang, G. C.; Phang, I. Y.; Ling, X. Y., *Angew. Chem. Int. Ed.*, 2022, 134, e202207447.
14. Leong, Y. X.; Lee, Y. H.; Koh, C. S. L.; Phan-Quang, G. C.; Han, X.; Phang, I. Y.; Ling, X. Y., *Nano Lett.*, 2021, 21, 2642-2649.
15. Phan-Quang, G. C.; Yang, N.; Lee, H. K.; Sim, H. Y. F.; Koh, C. S. L.; Kao, Y.-C.; Wong, Z. C.; Tan, E. K. M.; Miao, Y.-E.; Fan, W.; Liu, T.; Phang, I. Y.; Ling, X. Y., *ACS Nano*, 2019, 13, 12090-12099.
16. De Angelis, F.; Gentile, F.; Mecarini, F.; Das, G.; Moretti, M.; Candeloro, P.; Coluccio, M. L.; Cojoc, G.; Accardo, A.; Liberale, C.; Zaccaria, R. P.; Perozziello, G.; Tirinato, L.; Toma, A.; Cuda, G.; Cingolani, R.; Di Fabrizio, E., *Nat. Photonics*, 2011, 5, 682.
17. Xue, Z. X.; Wang, S. T.; Lin, L.; Chen, L.; Liu, M. J.; Feng, L.; Jiang, L., *Adv. Mater.*, 2011, 23, 4270– 4273.
18. Ling, X. Y.; Phang, I. Y.; Vancso, G. J.; Huskens, J.; Reinhoudt, D. N., *Langmuir*, 2009, 25, 3260– 3263.
19. Tan, J. M. R.; Ruan, J. J.; Lee, H. K.; Phang, I. Y.; Ling, X. Y., *Phys. Chem. Chem. Phys.*, 2014, 16, 26983-26990.
20. Lee, H. K.; Lee, Y. H.; Zhang, Q.; Phang, I. Y.; Tan, J. M. R.; Cui, Y.; Ling, X. Y., *ACS Appl. Mater. Interfaces*, 2013, 5, 11409-11418.
21. Huang, J.-A.; Zhang, Y.-L.; Zhao, Y.; Zhang, X.-L.; Sun, M.-L.; Zhang, W.; *Nanoscale*, 2016, 8, 11487-11493.
22. Rösler, C.; Fischer, R. A., *CrystEngComm*, 2015, 17, 199-217.
23. Paul, A.; Kaur, I. B.; Muthukumar, S.; Prasad, S., *ACS Omega*, 2022, 7, 26993-27003.

24. Chen, Y. F.; Babarao, R.; Sandler, S. I.; Jiang, J. W., *Langmuir*, 2016, 26, 8743-8750.

25. Koh, C. S. L.; Lee, H. K.; Han, X.; Sim, H. Y. F.; Ling, X. Y., *Chem. Commun.*, 2018, 54, 2546-2549.

26. Li, X.; Lee, H. K.; Phang, I. Y.; Lee, C. K.; Ling, X. Y., *Anal. Chem.*, 2014, 86, 10437-10444.

27. Han, X.; Koh, C. S. L.; Lee, H. K.; Chew, W. S.; Ling, X. Y., *ACS Appl. Mater. Interf.*, 2017, 9, 39635-39640.

28. Lay, C. L.; Koh, C. S. L.; Wang, J.; Lee, Y. H.; Jiang, R.; Yang, Y.; Yang, Z.; Phang, I. Y.; Ling, X. Y., *Nanoscale*, 2018, 10, 575-581.

29. Kaja, S.; Mathews, A. V.; Venuganti, V. V. K.; Nag, A., *Langmuir*, 2023, 39, 5591-5601.

30. Wold, S., *Chemometr. Intell. Lab Syst.*, 1995, 30, 109-115.

31. Yano, J.; Gaffney, K. J.; Gregoire, J.; Hung, L.; Ourmazd, A.; Schrier, J.; Sethian, J. A.; Toma, F. M., *Nat. Rev. Chem.*, 2022, 6, 357-370.

32. Phillips, G. R.; Harris, J. M., *Anal. Chem.*, 1990, 62, 2351-2357.

33. Penny, K. I.; Jolliffe, I. T., *J. R. Stat. Soc. D.*, 2001, 50, 295-307.

34. Lieber, C. A.; Mahadevan-Jansen, A., *Appl. Spectrosc.*, 2003, 57, 1363-1367.

35. Zhang, Z.-M.; Chen, S.; Liang, Y.-Z., *Analyst*, 2010, 135, 1138-1146.

36. He, S.; Zhang, W.; Liu, L.; Huang, Y.; He, J.; Xie, W.; Wu, P.; Du, C., *Anal. Methods*, 2014, 6, 4402-4407.

37. Fatima, A.; Cyril, G.; Vincent, V.; Stéphane, J.; Olivier, P., *Analyst*, 2020, 145, 2945-2957.

38. Barton, S. J.; Ward, T. E.; Hennelly, B. M., *Anal. Methods*, 2018, 10, 3759-3769.

39. Levina, E.; Wagaman, A. S.; Callender, A. F.; Mandair, G. S.; Morris, M. D., *J. Chemom.*, 2007, 21, 24-34.

40. Bro, R.; Smilde, A. K., *J. Chemom.*, 2003, 17, 16-33.

41. Navas, N.; Romero-Pastor, J.; Manzano, E.; Cardell, C., *J. Raman Spectrosc.*, 2011, 41, 1486-1493.

42. Martens, H.; Høy, M.; Wise, B. M.; Bro, R.; Brockhoff, P. B., *J. Chemom.*, 2003, 17, 153-165.

43. Zorzetti, B. M.; Shaver, J. M.; Harynuk, J. J., *Anal. Chim. Acta*, 2011, 694, 31-37.

44. Altman, N.; Krzywinski, C., *Nat Methods*, 2018, 15, 399-400.

45. Berisha, V.; Krantsevich, C.; Hahn, P. R.; Hahn, S.; Dasarathy, G.; Turaga, P.; Liss, J., *Npj Digit. Med.*, 2021, 4, 153.

46. Luo, S. H.; Wang, W. L.; Zhou, Z. F.; Xie, Y.; Ren, B.; Liu, G. K.; Tian, Z. Q., *Anal. Chem.*, 2022, 94, 10151-10158.

47. Leong, S. X.; Leong, Y. X.; Koh, C. S. L.; Tan, E. X.; Nguyen, L. B. T.; Chen, J. R. T.; Chong, C.; Pang, D. W. C.; Sim, H. Y. F.; Liang, X.; Tan, N. S.; Ling, X. Y., *Chem. Sci.*, 2022, 13, 11009-11029.

48. Zhang, X.; Li, H.; Tian, X.; Chen, C.; Su, Y.; Li, M.; Lv, J.; Chen, C.; Lv, X., *Chemom. Intell. Lab. Syst.*, 2022, 231, 104681.

49. Sil, S.; Mukherjee, R.; Kumbhar, D.; Reghu, D.; Shrungar, D.; Kumar, N. S.; Singh, U. K.; Umapathy, S., *J. Raman Spectrosc.*, 2021, 52, 2648-2659.

50. Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, P. W., *J. Artif. Intell.*, 2002, 16, 321-357.

51. Frid-Adar, M.; Klang, E.; Amitai, J., *IEEE International Symposium on Biomedical Imaging*, 2018, 289-293.

52. Leong, Y. X.; Tan, E. X.; Leong, S. X.; Koh, C. S. L.; Nguyen, L. B. T.; Chen, J. R. T.; Xia, K.; Ling, X. Y., *ACS Nano*, 2022, 16, 13279-13293.

53. Brereton, R. G.; Lloyd, G. R., *J. Chemom.*, 2014, 28, 213-225.

54. Abdi, H., *WIREs Comp. Stat.*, 2010, 2, 97-106.

55. Huang, L.; Wang, L.; Hu, X.; Chen, S.; Tao, Y.; Su, H.; Yang, J.; Xu, W.; Vedarethinam, V.; Wu, S.; Liu, B.; Wan, X.; Lou, J.; Wang, Q.; Qian, K., *Nat. Commun.*, 2020, 11, 3556.
56. Lassau, N.; Ammari, S.; Chouzenoux, E.; Gortais, H.; Herent, P.; Devilder, M.; Soliman, S.; Meyrignac, O.; Talabard, M.-P.; Lamarque, J.-P.; Dubois, R.; Loiseau, N.; Trichelair, P.; Bendjebbar, E.; Garcia, G.; Balleyguier, C.; Merad, M.; Stoclin, A.; Jegou, S.; Griscelli, F.; Tetelboum, N.; Li, Y.; Verma, S.; Terris, M.; Dardouri, T.; Gupta, K.; Neacsu, A.; Chemouni, F.; Sefta, M.; Jehanno, P.; Bousaid, I.; Boursin, Y.; Planchet, E.; Azoulay, M.; Dachary, J.; Brulport, F.; Gonzalez, A.; Dehaene, O.; Schiratti, J.-B.; Schutte, K.; Pesquet, J.-C.; Talbot, H.; Pronier, E.; Wainrib, G.; Clozel, T.; Barlesi, F.; Bellin, M.-F.; Blum, M. G. B., *Nat. Commun.*, 2021, 12, 634.
57. Li, R. Y.; Di Felice, R.; Rohs, R.; Lidar, D. A., *npj Quantum Inf.*, 2018, 4, 14.
58. Culos, A.; Tsai, A. S.; Stanley, N.; Becker, M.; Ghaemi, M. S.; McIlwain, D. R.; Fallahzadeh, R.; Tanada, A.; Nassar, H.; Espinosa, C.; Xenochristou, M.; Ganio, E.; Peterson, L.; Han, X.; Stelzer, I. A.; Ando, K.; Gaudilliere, D.; Phongpreecha, T.; Marić, I.; Chang, A. L.; Shaw, G. M.; Stevenson, D. K.; Bendall, S.; Davis, K. L.; Fantl, W.; Nolan, G. P.; Hastie, T.; Tibshirani, R.; Angst, M. S.; Gaudilliere, B.; Aghaeepour, N., *Nat. Mach. Intell.*, 2020, 2, 619-628.
59. Luts, J.; Ojeda, F.; Van de Plas, R.; De Moor, B.; Van Huffel, S.; Suykens, J. A. K., *Anal. Chim. Acta*, 2010, 665, 129-145.
60. Kuo, B. C.; Ho, H. H.; Li, C. H.; Hung, C. C.; Taur, J. S., *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2014, 7, 317-326.
61. Noble, W. S., *Nat. Biotechnol.*, 2006, 24, 1565-1567.
62. Ramou, E.; Palma, S. I. C. J.; Roque, A. C. A., *ACS Appl. Mater. Interf.*, 2022, 14, 6261-6273.

63. Reynaud, L.; Bouchet-Spinelli, A.; Janot, J.-M.; Buhot, A.; Balme, S.; Raillon, C., *Anal. Chem.*, 2021, 93, 7889-7897.

64. Rokach, L., *Artif. Intell. Rev.*, 2010, 33, 1-39.

65. Breiman, L., *Mach. Learn.*, 2001, 45, 5-32.

66. Ouassil, N.; Pinals, R. L.; Bonis-O'Donnell, J. T. D.; Wang, J. W.; Landry, M. P., *Sci. Adv.*, 2022, 8, eabm0898.

67. Natekin, A.; Knoll, A., *Front. Neurorobot.*, 2013, 7, 1.

68. McKenna, M.; Shackelford, D.; Ferreira Pontes, H.; Ball, B.; Nance, E., *ACS Nano*, 2021, 15, 8559-8573.

69. Jain, A. K.; Jianchang, M.; Mohiuddin, K. M., *Computer*, 1996, 29, 31-44.

70. Gardner, M. W.; Dorling, S. R., *Atmos. Environ.*, 1998, 32, 2627-2636.

71. Yaari, Z.; Yang, Y.; Apfelbaum, E.; Cupo, C.; Settle, A. H.; Cullen, Q.; Cai, W.; Roche, K. L.; Levine, D. A.; Fleisher, M.; Ramanathan, L.; Zheng, M.; Jagota, A.; Heller, D. A., *Sci. Adv.*, 2021, 7, eabj0852.

72. Sze, V.; Chen, Y. H.; Yang, T. J.; Emer, J. S., *Proceedings of the IEEE*, 2017, 105, 2295-2329.

73. Xu, W.; Lin, J.; Gao, M.; Chen, Y.; Cao, J.; Pu, J.; Huang, L.; Zhao, J.; Qian, K., *Adv. Sci.*, 2020, 7, 2002021.

Chapter 2 SERS Taster: A Machine Learning-Driven Multireceptor Platform for Multiplex Profiling of Wine Flavors

Abstract. Integrating ML with SERS accelerates the development of practical sensing devices. Such integration, in combination with direct detection or indirect analyte capturing strategies, is key to achieve high predictive accuracies even in complex matrices. However, in-depth understanding of spectral variations arising from specific chemical interactions is essential to prevent model overfit. Herein, we design a ML-driven 'SERS Taster' to simultaneously harness useful vibrational information from multiple receptors for enhanced multiplex profiling of five wine flavor molecules at parts-per-million levels. Our receptors employ numerous non-covalent interactions to capture chemical functionalities within flavor molecules. By strategically combining all receptor-flavor SERS spectra, we construct comprehensive 'SERS super-profiles' for predictive analytics using chemometrics. We elucidate crucial molecular-level interactions in flavor identification, and further demonstrate the differentiation of primary, secondary, and tertiary alcohol functionalities. Our SERS taster also achieves perfect accuracies in multiplex flavor quantification in an artificial wine matrix.

2.1 Introduction

The integration of ML with SERS presents significant potential in translating research-based SERS platforms for diverse applications.[1-3] These applications demand SERS platforms to achieve ultra-trace detection of multiple molecules with weak Raman scattering cross-sections, via either direct detection or indirect analyte capturing with molecular receptors that promote receptor-analyte interactions.[4-5] Vibrational fingerprints resulting from direct detection of these molecules are often insignificant, while subtle changes in receptor fingerprints are difficult to pinpoint through manual visual inspection. ML algorithms perform automated analyses of large SERS spectral datasets across entire spectral windows, eliminating subjective judgements to attain unparalleled accuracies.[6-7] More importantly, they unveil intricate data patterns which enable predictive analytics crucial for Raman/SERS-based applications.[8-10] However, there is an inherent risk of over-relying on these powerful algorithms to achieve desired outcomes without thoroughly understanding chemical interactions which occur at the molecular level. The resulting overfit models crumble when introducing new data or attempting to predict properties of an unknown sample. Consequently, establishing a strong correlation between chemical knowledge and chemometric model outputs is critical to harness the inherent strengths of ML approaches in extracting and comprehending complex SERS fingerprints.

Herein, we design a ML-driven 'SERS taster', capable of achieving multiplex profiling of five wine flavor molecules with 100% accuracy at parts-per-million levels (**Figure 2-1**). Our strategy employs multiple carefully designed receptor-flavor chemical interactions to capture all active chemical functionalities within flavor molecules. By serially combining all receptor-flavor spectra as a 'SERS super-profile', we construct a more complete spectroscopic profile for each molecule. These compound SERS super-profiles are comprehensively analyzed using

ML-driven chemometric models, enabling unambiguous identification and multiplex quantification of wine flavors.

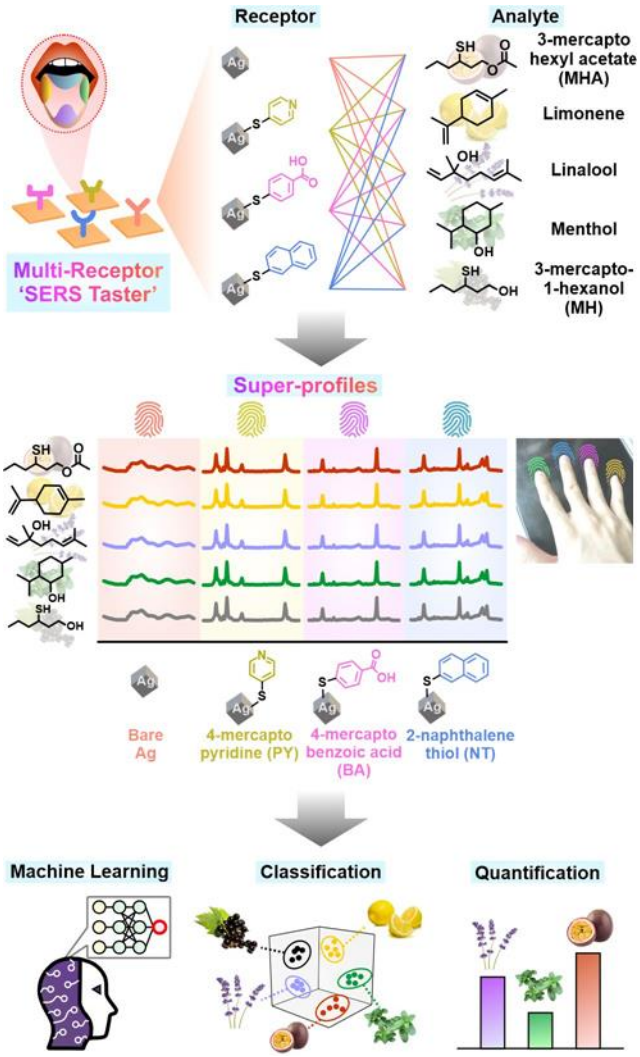


Figure 2-1. Our SERS Taster design comprises four independent SERS substrates with unique molecular receptors (Bare Ag, MPY, MBA, NT) to distinguish wine flavor compounds (Menthol, Linalool, Limonene, MHA, MH). Through specific receptor-flavor chemical interactions, we construct SERS super-profiles for each flavor molecule by strategically combining all receptor SERS spectra in series. Analysis of these super-profiles using ML-driven chemometric models enable enhanced identification and quantification of flavor molecules.

To achieve this, we first judiciously select four surface receptors to introduce a wide range of receptor-flavor chemical interactions. Next, we select five representative wine flavors, including higher aliphatic alcohols (menthol), terpenes (linalool, limonene) and sulfur-containing compounds (3-mercaptohexyl acetate (MHA), 3-mercapto-1-hexanol (MH)) (**Figure 2-2**).[11-12] These flavor molecules have weak Raman scattering cross-sections and are challenging to detect even with advanced chromatographic techniques. The interactions between individual receptor-flavor pairs produce SERS spectral variations that corroborate both experimentally and *in silico*.

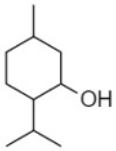
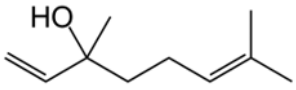
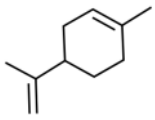
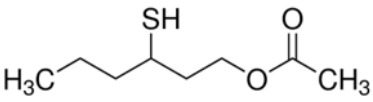
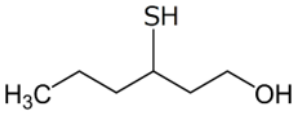
Wine flavor	Molecular structure	Types of wines present in	Possible intermolecular interactions
Menthol		Many red wine types such as Bordeaux, Cabernet Sauvignon, Shiraz	Hydrogen bonding; van der Waals' forces
Linalool		White wine types such as Muscat, Galicia	Hydrogen bonding; π-π interactions; van der Waals' forces
Limonene		White wine types such as Chardonnay, Viognier	π-π interactions; van der Waals' forces
3-mercaptohexyl acetate (MHA)		Red wine types, such as Cabernet Sauvignon, Merlot and white wine types such as Sauvignon Blanc and Semillon	Electrostatic interactions; van der Waals' forces; Ag-S bonding
3-mercapto-1-hexanol (MH)		Red wine types, such as Cabernet Sauvignon, white wint types such as Sauvignon Blanc and rosé wines	Hydrogen bonding; van der Waals' forces; Ag-S bonding

Figure 2-2. Molecular structures of representative molecules for various wine flavors, their occurrence in various wines, along with possible interactions with receptor molecules.[12-18]

We further amplify these variations by combining individual SERS spectra into a compound SERS super-profile for every flavor. PCA of these super-profiles achieve complete flavor identification, enabling even the discrimination of alcohols with varying degrees of substitution. We further employ support vector machine discriminant analysis (SVMDA) to

quantitatively classify all flavors with 100% accuracy. In contrast, flavor classification accuracy drops dramatically to 33% with single receptors. Finally, our SERS taster achieves perfect accuracy in the multiplex quantification of wine flavors in an artificial wine matrix despite potential interferences. Our SERS taster overcomes current limitations in wine flavor profiling as a highly sensitive SERS platform that requires minimal sample pre-treatment and provides ease of multiplex detection. Collectively, our findings pave the way in the development of innovative SERS sensors for flavor chemistry and a myriad of applications extending beyond.

2.2 Results and discussion

2.2.1 Overview of SERS Taster

Our SERS taster design incorporates multiple molecular receptors grafted onto Ag nanocube surfaces to capture and confine target flavor molecules close to the SERS platform for enhanced signals. To begin, we prepare densely packed Ag nanocube arrays using the Langmuir-Blodgett technique (edge length = 117 ± 6 nm; particle density = 32 nanocubes/μm^2; **Figure 2-3**, **2-4A**). The combination of strong EM enhancement from the sharp edges of Ag nanocubes and inter-particle plasmonic coupling give rise to a high SERS enhancement of 1.9 × 10^6 using 4-mercaptopyridine (MPY) as the receptor molecule.[2, 5, 19] The SERS EF is calculated as follows:

$$\text{SERS EF} = \frac{\mathrm{I_{SERS}}}{\mathrm{I_{Raman}}} \times \frac{\mathrm{N_{Raman}}}{\mathrm{N_{SERS}}}$$

In solution

$$\text{Confocal laser volume, } \mathrm{V_{laser}} = \pi \times \frac{x}{2} \times \frac{y}{2} \times z$$

$$= 9.55 \times 10^{9}\ \mathrm{nm^3}$$

$$= 9.55 \times 10^{-18}\ \mathrm{m^3}$$

where x = 910 nm, y = 680 nm, z = 4320 nm are measured confocal resolution in x, y and z dimensions in solution.

$$N_{Raman} = V_{laser} \times C_{Receptor} \times \text{Avogadro's no.}$$

$$= 9.55 \times 10^{-18}\ m^3 \times 10\ mol/m^3 \times 6.02 \times 10^{23}\ \text{molecules/mol}$$

$$= 1.26 \times 10^7\ \text{molecules}$$

On substrate

$$\text{Area of laser spot, } A_{laser} = \pi \times \frac{x}{2} \times \frac{y}{2}$$

$$= 1.55 \times 10^5\ nm^2\ \text{(in air)}$$

where x = 520 nm, y = 380 nm are measured confocal resolution in x and y dimensions in air.

$$\text{Estimated particle density, } P_{cubes} = 32\ \text{nanocubes}/\mu m^2$$

$$\text{No. of Ag nanocubes within laser spot, } N_{cubes} = P_{cubes} \times A_{laser}$$

$$= 4.97\ \text{nanocubes}$$

$$\text{Exposed surface area of Ag nanocubes, } S_{cubes} = N_{cubes} \times A_{cubes}$$

$$= 4.97 \times (117^2)$$

$$= 6.80 \times 10^4\ nm^2$$

where A_{cubes} is the surface area of the nanocubes exposed to receptor molecules (top facet).

$$N_{SERS} = S_{cubes} \times D_{receptors}$$

$$= 6.80 \times 10^4\ nm^2 \times 0.329\ \text{molecules}/nm^2$$

$$= 2.24 \times 10^4\ \text{molecules}$$

where $D_{receptor} = 3.29 \times 10^{13}$ molecules/cm^2, according to literature.[20]

For the 1098 cm^{-1} peak in MPY:

$$\text{SERS EF} = \frac{14988.1}{4.5} \times \frac{1.26 \times 10^7}{2.24 \times 10^4}$$

$$= \mathbf{1.9 \times 10^6}$$

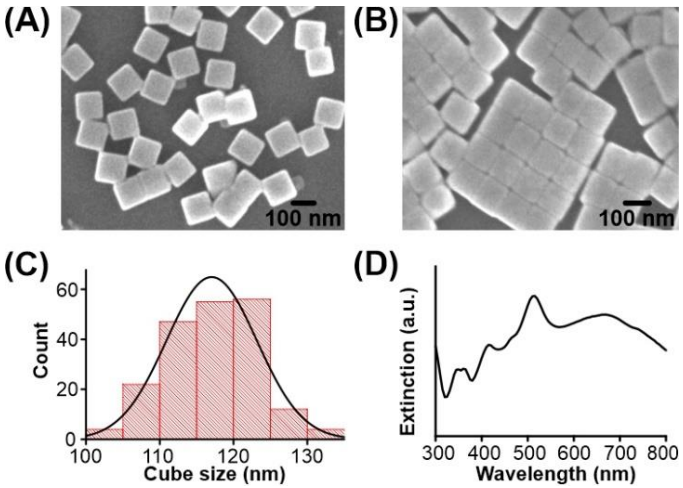


Figure 2-3. (A) SEM image of the synthesized Ag nanocubes. (B) Magnified SEM image showing packing density of Ag nanocubes after self-assembly using the Langmuir-Blodgett technique. (C) Edge length distribution of Ag nanocubes. The average edge length is 117 ± 6 nm. (D) UV-vis extinction spectrum of colloidal Ag in ethanol. The peaks at 355, 415, 510 and 670 nm can be assigned to octupole (355 nm), quadrupole (415, 510 nm) and dipole (670 nm) resonances.[21]

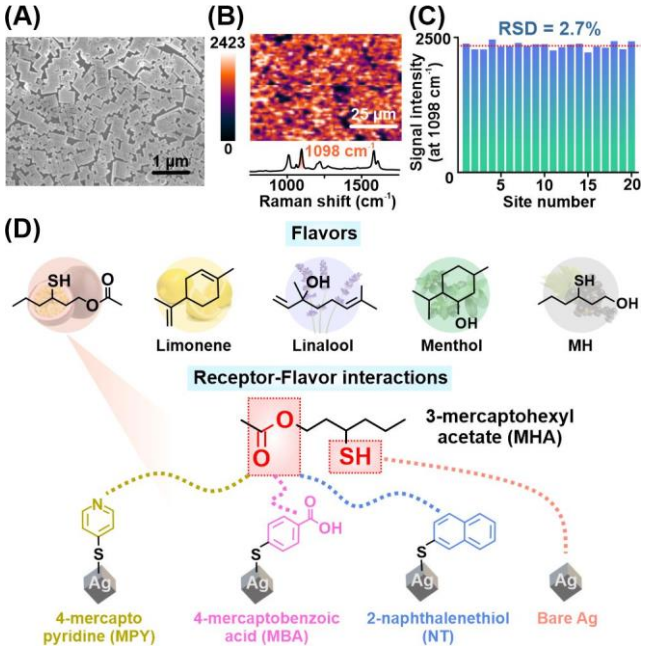


Figure 2-4. (A) SEM image of the substrate surface showing the packing density of Ag nanocubes. (B) Hyperspectral SERS map generated using the well-defined C-S stretching mode at 1098 cm^{-1}. (C) Relative standard deviation of the 1098 cm^{-1} peak across selected

spectra, showing high signal homogeneity of the platform. The red dotted line indicates the mean signal intensity. (D) Scheme depicting MHA exposure to each receptor platform, forming receptor-flavor interactions.

The hyperspectral SERS map exhibits highly consistent signal intensities across an approximate area of 5 mm^2, with a relative standard deviation of 2.7%, indicating homogeneous enhancement capabilities (**Figure 2-4B**, **2-4C**). The high SERS enhancement and low intensity variation indicate a robust and homogeneous SERS platform which is crucial to successfully detect flavor molecules with weak Raman scattering cross-sections.

The key strategy of our SERS taster involves constructing a holistic spectroscopic profile for each flavor molecule. We achieve this by introducing multiple receptors to promote interactions with various active chemical functionalities on flavor molecules (**Figure 2-4D**). In our SERS Taster, we select 4-mercaptopyridine (MPY), 4-mercaptobenzoic acid (MBA), 2-naphthalenethiol (NT) and a bare Ag surface as receptors. These receptors promote electrostatic interactions, hydrogen bonding, π-π interactions, van der Waals' forces and Ag-thiolate bonding with flavor chemical functionalities such as alcohols and esters. These chemical interactions confine flavor molecules near the Taster surface, producing characteristic receptor SERS spectral changes. Apart from bare Ag, these thiolated receptors allows the formation of self-assembled monolayers on Ag nanocubes. They also contain aromatic rings which exhibit larger Raman cross-sections, thereby amplifying spectral changes upon interaction.[22]

2.2.2 Profiling passionfruit flavor

Using MHA as a model wine flavor, we demonstrate that the characteristic spectral variations observed with our SERS taster corroborates with density functional theory (DFT) simulations. MHA is a passion fruit flavor commonly found in wines such as Cabernet Sauvignon and Merlot with dominant influence on the eventual wine flavor.[12]

We examine the experimental SERS spectra obtained using MPY, MBA, NT, and bare Ag before and after exposure to aqueous MHA (10^{-3} M), assigning key vibrational modes using DFT. In control experiments without MHA, MPY exhibits characteristic twin in-plane C-H deformations at 1201, 1220 cm^{-1} and twin C-C/C-N pyridine ring stretching ($\nu_{CC/CN}$) at 1583, 1611 cm^{-1} respectively (**Figure 2-5Ai**, **2-6A**).[23] The presence of MHA intensifies the 1220 cm^{-1} peak and weakens the 1611 cm^{-1} peak. Formation of electrostatic interactions between the MPY nitrogen and MHA oxygen polarizes the aromatic C-H bonds in MPY, lowering the energy requirement for C-H bond vibration and restricting $\nu_{CC/CN}$ within the aromatic ring (**Figure 2-5Bi**). In the DFT-simulated SERS spectrum without MHA, the twin peak at 1246 and 1278 cm^{-1} can be indexed to in-plane C-H deformation while the peak at 1613 cm^{-1} can be indexed to $\nu_{CC/CN}$ (**Figure 2-7A**). In the presence of MHA, the 1278 cm^{-1} peak intensifies relative to the 1246 cm^{-1} peak while the 1613 cm^{-1} peak weakens relative to both the 1246 and 1278 cm^{-1} peaks.

For MBA, a broad feature including peaks at 1358, 1382 and 1425 cm^{-1} are indexed to symmetric carboxylate stretching (ν_{OCO-}) (**Figure 2-5Aii**, **2-6B**).[24] Interactions with MHA intensifies the 1382 cm^{-1} peak with a concomitant blue-shift of the feature to 1390, 1418 and 1435 cm^{-1}, respectively. Electrostatic interactions between the carboxylate group and MHA's acetate carbon lowers electron density within the carboxylate moiety, reducing the energy required for ν_{OCO-} (**Figure 2-5Bii**). In the DFT-simulated SERS spectrum, the peaks at 1340, 1390 and 1438 cm^{-1} can be indexed to symmetric ν_{OCO-}. Upon addition of MHA, the 1390 cm^{-1}

^{1} peak intensifies significantly relative to the 1340 and 1438 cm^{-1} peaks (**Figure 2-7B**). The 1340 and 1438 cm^{-1} peaks do not show significant increase in intensity as the symmetric ν_{OCO}- contribution from these two peaks are lesser.

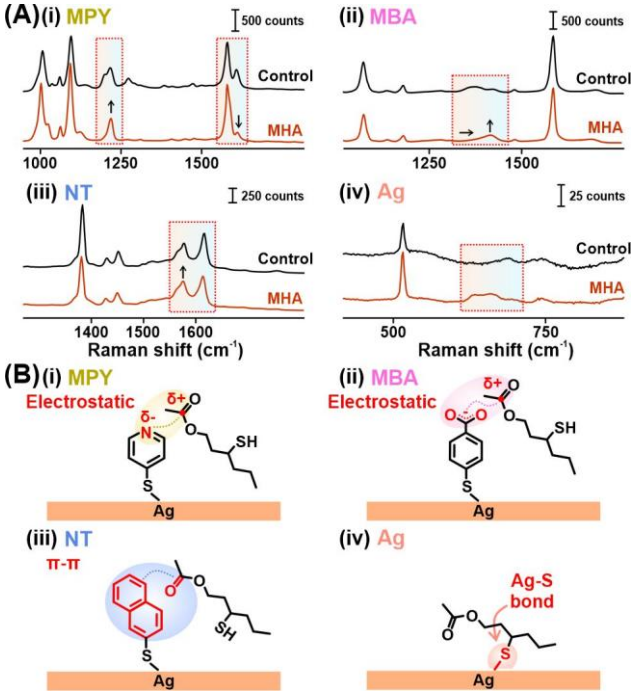


Figure 2-5. (A) Experimental SERS spectra obtained for each receptor in the presence of MHA (red) and absence of MHA with an aqueous background (grey): (i) MPY (ii) MBA (iii) NT (iv) Ag. (B) Scheme depicting receptor-flavor interactions at the molecular level, including the type of interaction, and interacting sites for (i) PY (ii) BA (iii) NT (iv) Ag.

For NT, the twin peak at 1571 and 1582 cm^{-1} and a peak at 1621 cm^{-1} are indexed to asymmetric and symmetric C-C ring stretching (ν_{CC}), respectively (**Figure 2-5Aiii**, **2-6C**).[25] Presence of MHA increases the twin peak intensity relative to the 1621 cm^{-1} peak (**Table 2-1**). π-π interactions between the naphthalene ring and MHA's carbonyl results in polarization of NT's aromatic C-C bonds, leading to a decrease in energy requirement for asymmetric ν_{CC} (**Figure 2-5Biii**). In the DFT-simulated SERS spectrum, the peaks at 1598, 1632 and 1665 cm^{-}

[1] can be indexed to ν_{CC} of the naphthalene ring (**Figure 2-7C**). In the presence of MHA, the intensity of peaks at 1598 and 1632 cm^{-1} increases relative to the 1665 cm^{-1} peak. We quantify this increase by comparing the peak ratios in both our experimental and simulated spectra, designating the peaks 1 to 3 by ascending wavenumbers. In both cases, peaks 1 and 2 show an increase in intensity relative to peak 3. While all three peaks account for different ν_{CC}, only peak 3 involve ν_{CC} that is symmetric about the C2 axes of the naphthalene ring.

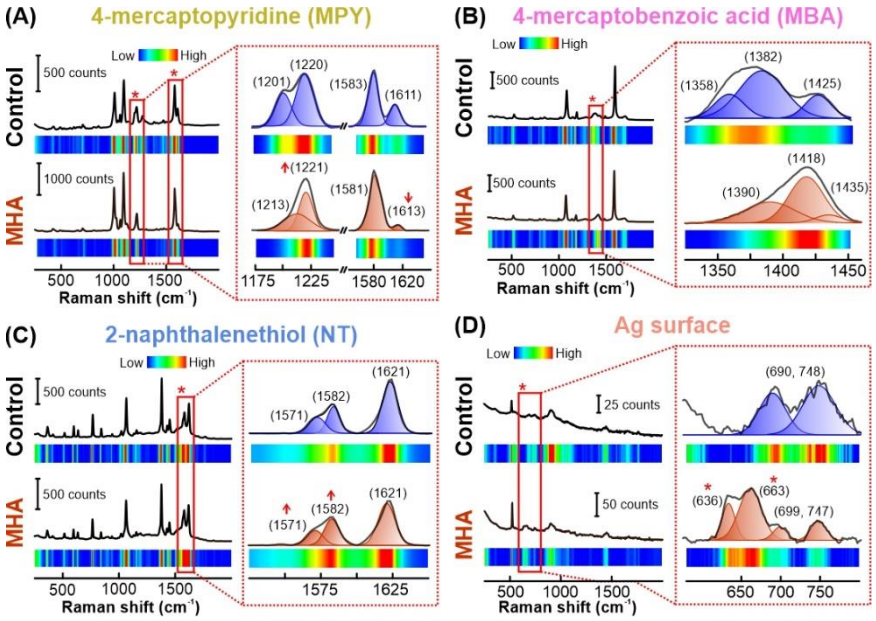


Figure 2-6. In-depth analysis of experimental spectral differences through peak deconvolution. (A) MPY – 1220 cm^{-1} peak intensifies and 1611 cm^{-1} peak weakens. (B) MBA – 1382 cm^{-1} peak intensifies and the feature blue-shifts to 1390, 1418 and 1435 cm^{-1} respectively. (C) NT – 1571 and 1582 cm^{-1} peaks intensify relative to the 1621 cm^{-1} peak. (D) Ag surface – Emergence of MHA peaks at 636, 663 cm^{-1}. The color bar represents the relative peak intensity.

Finally, for the bare Ag surface, addition of MHA results in an emergence of two additional peaks at 636 and 663 cm^{-1}, indexed to acetate wagging (π_{OCO}) and bending (δ_{HCO}) modes of MHA (**Figure 2-5Aiv**, **2-6D**). The formation of strong Ag-thiolate bond between Ag and the MHA thiol group brings MHA in proximity of the strong electromagnetic enhancement

provided by the Ag nanocubes, leading to the observation of these signals (**Figure 2-5Biv**). In the DFT-simulated spectrum, presence of MHA induces the emergence of SERS peaks at 670 and 760 cm^{-1} that are indexed to the acetate wagging (π_{OCO}) and bending modes (δ_{HCO}) of MHA (**Figure 2-7D**).

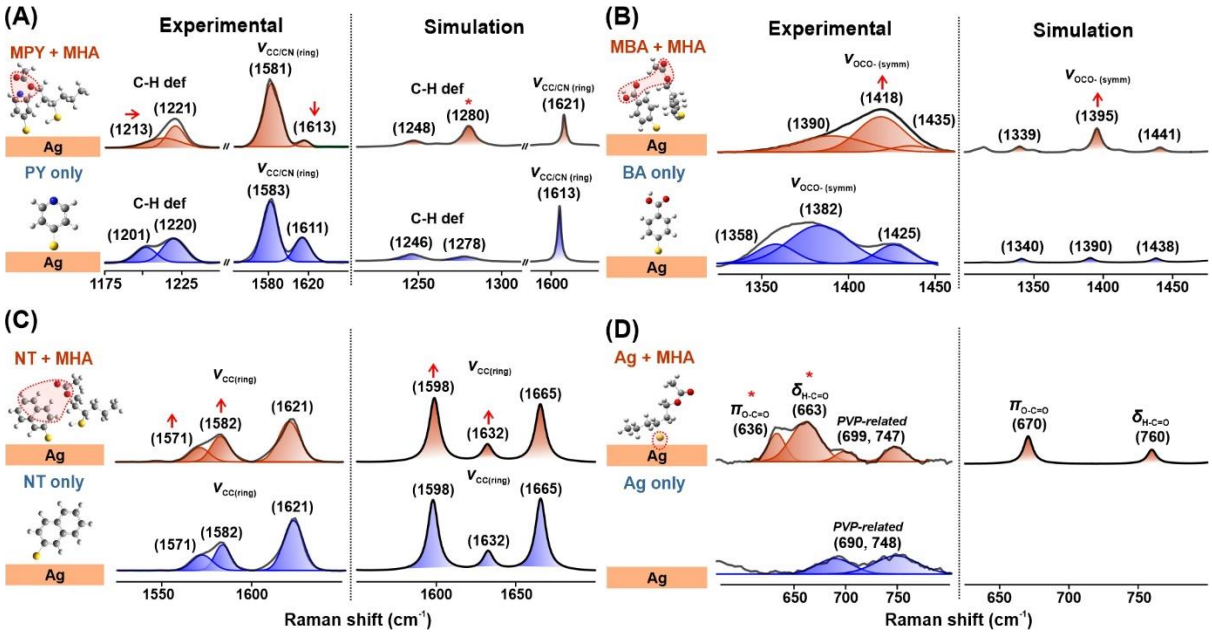


Figure 2-7. Experimental SERS spectra compared to density functional theory (DFT) simulations of the receptors in the presence (red) and absence of MHA (blue). (A) MPY – 1278 cm^{-1} peak (C-H def) intensifies and 1613 cm^{-1} peak ($\nu_{CC/CN}$) weakens. (B) MBA – 1390 cm^{-1} peak (ν_{OCO-}) intensifies, and blue-shift to 1395 cm^{-1}. (C) NT – 1598 and 1632 cm^{-1} peaks (ν_{CC}) intensify while 1665 cm^{-1} peak remains constant. (D) Emergence of MHA peaks at 670 and 760 cm^{-1} (π_{OCO} ; δ_{HCO}).

Collectively, the spectral changes observed experimentally corroborates strongly with the DFT-simulated spectral changes. These computationally optimized molecular structures provide critical insight to the receptor-flavor chemical interactions occurring at the molecular level. Crucially, we highlight the ability of individual receptors to interact with different

functional groups of a single flavor molecule which collectively contribute to the re-construction of its chemical profile.

Table 2-1. Comparison of experimental and DFT simulated peak intensities (a. u.) for the ring stretching modes of NT. Underlined values highlight the increase in peak intensity of peaks 1 and 2 relative to peak 3.

Experimental spectrum

	Peak 1: 1571 cm^{-1}	**Peak 2: 1582 cm^{-1}**	**Peak 3: 1621 cm^{-1}**
MHA	142.35 ± 18.39	252.04 ± 20.73	403.04 ± 7.61
Control	153.54 ± 12.67	254.98 ± 24.35	494.03 ± 7.96
	Peak 1 : Peak 2	**Peak 1 : Peak 3**	**Peak 2 : Peak 3**
MHA	0.56	**0.35**	**0.63**
Control	0.60	0.31	0.52
Percentage change	6.7%	**12.9%**	**21.2%**

DFT simulated spectrum

	Peak 1: 1598 cm^{-1}	**Peak 2: 1632 cm^{-1}**	**Peak 3: 1665 cm^{-1}**
MHA	684.21	197.25	618.82
Control	754.35	218.07	768.71
	Peak 1 : Peak 2	**Peak 1 : Peak 3**	**Peak 2 : Peak 3**
MHA	3.47	**1.11**	**0.32**
Control	3.46	0.98	0.28
Percentage change	0.3%	**13.3%**	**14.3%**

2.2.3 Constructing SERS super-profiles

Leveraging the useful vibrational information conferred by each receptor, we strategically construct a SERS super-profile for MHA through horizontal combination (**Figure 2-8**). The resulting super-profile comprises all spectral variations arising from receptor-MHA interactions with MPY, MBA, NT, and Ag. In contrast, using a single receptor limits interactions to a single flavor functional group. Hence, distinct SERS spectral changes with a single receptor merely describes a partial profile of the flavor chemical structure. We subsequently construct SERS super-profiles for four additional flavor molecules (menthol, linalool, limonene, and MH) present in a large variety of red, white and rosé wines where they contribute to specific sensory attributes that constitute the overall wine flavor.[26]

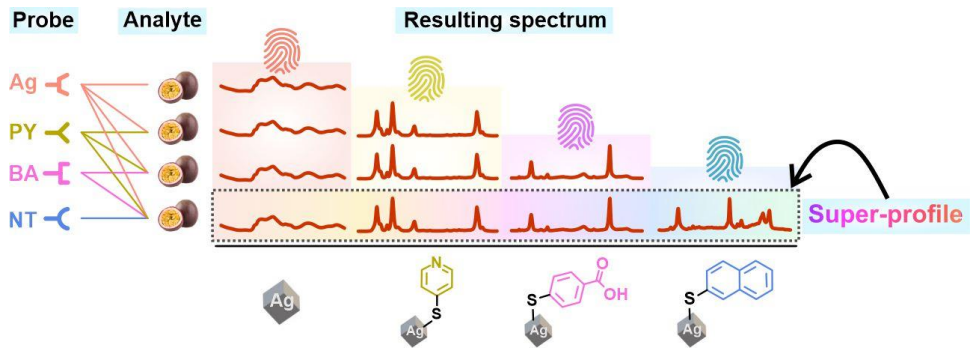


Figure 2-8. Scheme depicting the construction of a SERS super-profile by horizontally combining all four receptor spectra in series.

To illustrate the superiority of our SERS super-profiles in identifying and classifying wine flavors, we employ PCA to distinguish between super-profiles of different flavor molecules. PCA offers unparalleled accuracy in scrutinizing the full spectral window with minimal analytical errors, effectively identifying even subtle spectral variations. Based on the 2D PCA score plots (**Figure 2-9Ai**), our SERS taster exhibits distinct and compact clusters of data representing individual flavor molecules including the control. Each data cluster is

encapsulated within a 95% confidence ellipse. The clear separation between these confidence ellipses indicates that our SERS taster effectively differentiates all five flavor molecules.

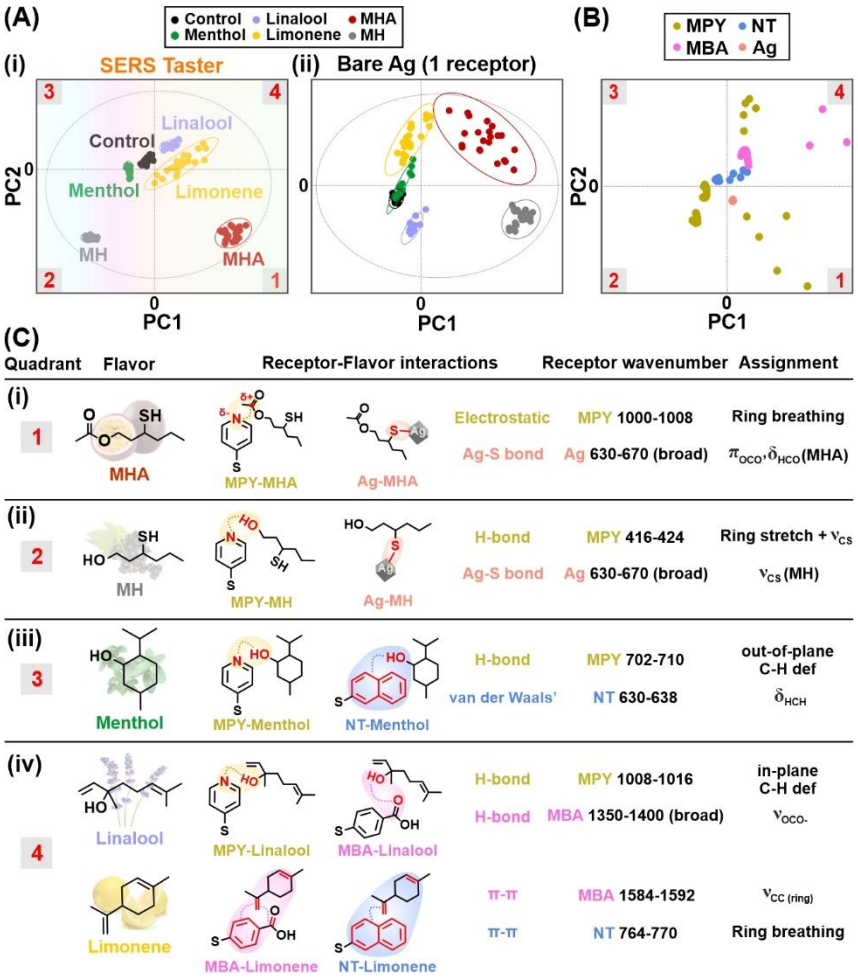


Figure 2-9. ML-driven classification of flavor molecules using 144 super-profiles (576 SERS spectra). (A) PCA score plot of the first two principal components showing the relative flavor data cluster separation using our SERS Taster. (B) PCA biplot of the first two principal components, which highlights key spectral regions where variations exert dominant influence over component scores. The specific spectral regions, their corresponding receptor-flavor interactions and vibrational assignment are summarized based on their influence in the flavor data cluster positions in each of the quadrants of the score plot (i) Quadrant 1 (ii) Quadrant 2 (iii) Quadrant 3 (iv) Quadrant 4. (C) Table summarizing key receptor-flavor interactions derived from the PCA biplot.

In contrast, as the number of receptors decreases, the relative ability to separate these data clusters diminishes (**Figure 2-10**). The use of a single receptor (Ag) results in overlaps between the confidence ellipses of menthol, limonene, and the control (**Figure 2-9Aii**). This overlap shows that the single receptor model cannot differentiate flavors due to insufficient SERS spectral variances. In addition, the larger confidence ellipses signify high intra-cluster variance. This comparison highlights the advantage of our SERS super-profiles over single receptor-flavor interactions for distinct flavor identification.

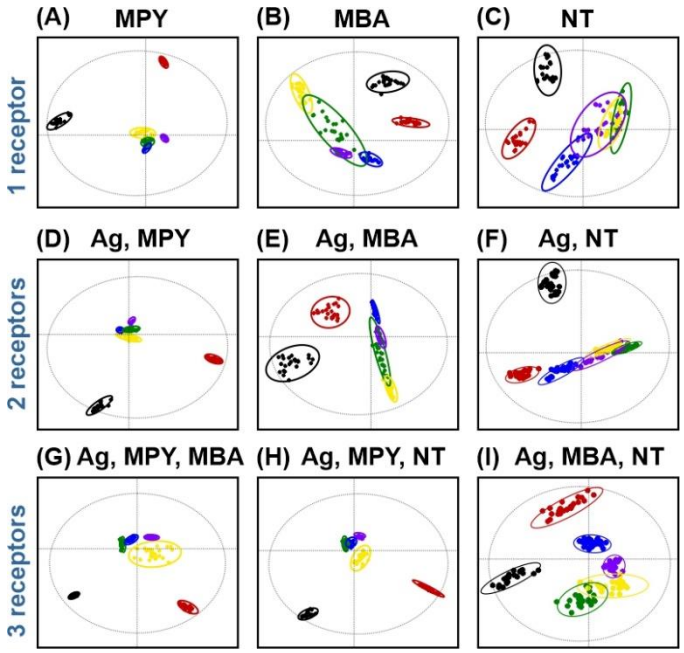


Figure 2-10. Principal component analysis (PCA) score plots for other receptor combinations with receptor number ≤ 3. (A) MPY only. (B) MBA only. (C) NT only. (D) Ag + MPY. (E) Ag + MBA (F) Ag + NT (G) Ag + MPY + MBA (H) Ag + MPY + NT (I) Ag + MBA + NT.

To further elucidate the underlying chemical meaning behind the PCA scores, we scrutinize the PCA biplot (**Figure 2-9B**, **2-9C**). The biplot features key receptor spectral regions which drive the separation of each flavor data cluster to different quadrants of the PCA

score plot.[27-28] We first deconstruct the biplot into four quadrants and relate these spectral regions to specific receptor-flavor interactions causing the variation.

Firstly, PCA classifies MHA in the first quadrant of the score plot (**Figure 2-9Ai**, lower right). In this quadrant, electrostatic interactions of MHA with MPY result in significant spectral variations in MPY's pyridine ring breathing mode at the 1000 – 1008 cm^{-1} region (**Figure 2-9Ci**, **2-11**).[23] The proximity of MHA with the bare Ag surface also amplifies MHA's π_{OCO} and δ_{HCO} at the 630 – 670 cm^{-1} region.

MH lies in the second quadrant of the score plot, at the opposing end of the PC1 axis compared to MHA (**Figure 2-9Ai**, lower left). Two reasons contribute to this classification. First, hydrogen bonding between MH and MPY influences MPY's concurrent pyridine ring stretch and C-S stretching (ν_{CS}) mode at the 416 – 424 cm^{-1} region (**Figure 2-9Cii**, **2-11**).[23] Next, Ag-thiolate interactions between MH and Ag magnifies MH's ν_{CS} at 630 – 670 cm^{-1} region. In this region, the Ag-MH spectrum is different from the Ag-MHA spectrum due to inherent differences in molecular structure and interactivity with the Ag surface. These unique peak shape changes promote the classification of MH in the opposing quadrant of MHA.

Menthol positions itself between the second and third quadrant of the score plot (**Figure 2-9Ai**, upper left). Crucially, hydrogen bonding between menthol and MPY causes peak ratio changes for MPY's out-of-plane C-H deformation mode at the 702 – 710 cm^{-1} region (**Figure 2-9Ciii**, **2-11**).[23] In addition, van der Waals' interactions between menthol and NT causes a subtle decrease in peak intensity of the C-H bending (δ_{HCH}) mode of NT at the 630 – 638 cm^{-1} region.[25] These variations differentiate menthol from the flavorless control.

Finally, linalool and limonene emerge as separate clusters within the fourth quadrant of the score plot (**Figure 2-9Ai**, upper right). For linalool, hydrogen bonding between linalool and MPY increases the peak intensity of MPY's in-plane C-H deformation mode at the 1008 – 1016 cm^{-1} region (**Figure 2-9Civ**, **2-11**).[23] Concurrently, hydrogen bonding between linalool

and MBA result in a distinct peak ratio change involving MBA's ν_{OCO-} at the 1350 – 1400 cm^{-1} region.[24] For limonene, π-π interactions between limonene and MBA result in a red shift of MBA's ν_{CC} at the 1584 – 1592 cm^{-1} region (**Figure 2-9Civ**, **2-11**).[24] π-π interactions between limonene and NT also result in a decrease in peak intensity of NT's ring breathing mode at the 764 – 770 cm^{-1} region.[25]

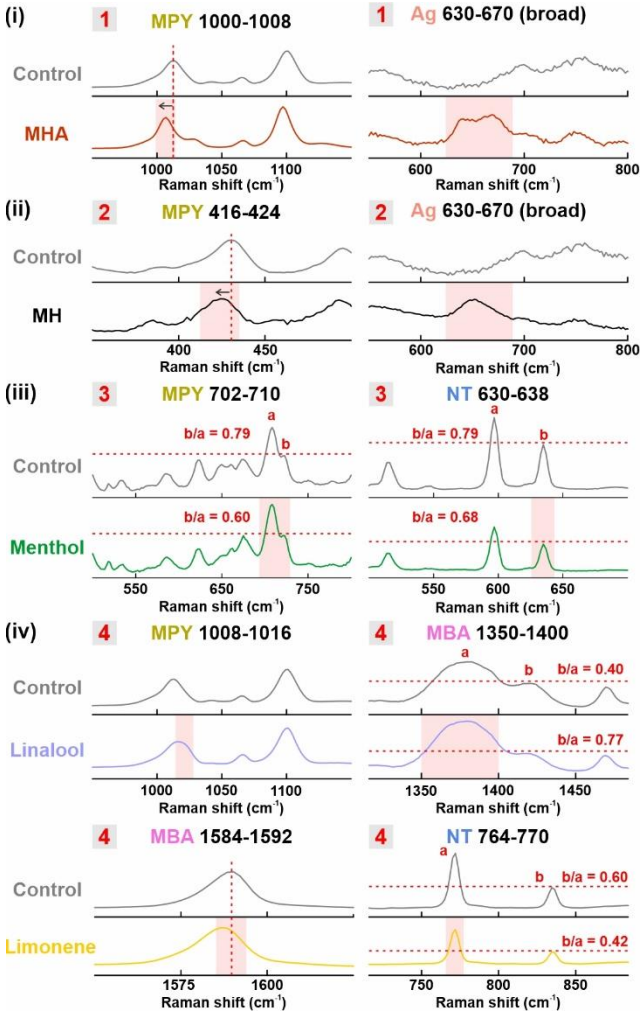


Figure 2-11. Experimental SERS spectra showing key spectral regions highlighting critical spectral variations in the PCA biplot. The spectra are arranged according to the flavor separated within the labelled quadrants – (i) quadrant 1 (ii) quadrant 2 (iii) quadrant 3 (iv) quadrant 4 – respectively.

Collectively, these differences drive the cluster separation of linalool and limonene from the flavorless control. Notably, we demonstrate that our SERS Taster distinguishes primary (MH), secondary (menthol) and tertiary (linalool) alcohols by classifying them in different quadrants with distinct SERS spectral changes. This successful classification is driven by our information-rich SERS super-profiles which amalgamate and magnify all spectral variance arising from individual receptor-flavor interactions.

By examining the PCA biplot, we use our knowledge of chemical interactions occurring at the molecular level to unravel how the chemometric model classifies different flavors as distinct clusters. This bridges the gap between SERS spectral inputs and chemometric model outputs, ensuring our model is built upon valid receptor-flavor spectral variation and not meaningless background variations.

To quantitatively evaluate the predictive capability of our SERS Taster, we construct confusion matrices using SVMDA (**Figure 2-12**). SVMDA is a supervised machine learning model that allows us to predict the identity of flavor molecules by examining their SERS super-profiles, with a high degree of flexibility and robustness.[29-31] In the first model, we introduce super-profiles used in the PCA earlier. In the second model, we introduce SERS spectra derived by exposing flavor molecules only to bare Ag.

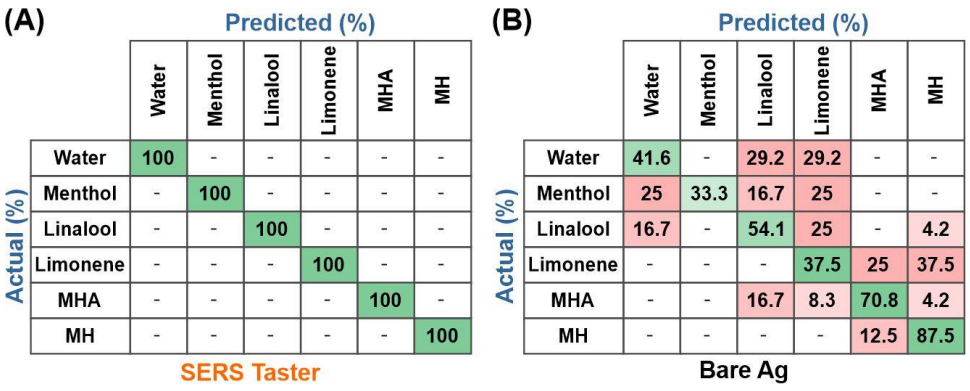


Figure 2-12. Confusion matrices obtained using SVM-DA for (A) our SERS Taster and (B) a bare Ag surface.

From the resulting confusion matrices, we affirm that our SERS Taster achieves 100% accuracy in the classification of all flavors, including the control (**Figure 2-12A**, **Table 2-2**). Conversely, bare Ag shows a wide accuracy range of 33.3 – 87.5% (**Figure 2-12B**, **Table 2-2**).

Table 2-2. Summary table for the overall classification accuracy, rate of false positives and false negatives. False negative refers to a data point that belongs to a flavor molecule but is incorrectly classified as a control data. False positive (for a single flavor molecule) refers to a data point belonging to another flavor molecule incorrectly classified as itself.

	Ag surface only	**SERS Taster**
Average classification accuracy	**54.2%** (78/144)	**100%**
False negative	**6.9%** (10/144)	**0%**
False positive (for a single flavor molecule):		
Menthol	**0%**	**0%**
Linalool	**10.4%** (15/144)	**0%**
Limonene	**14.6%** (21/144)	**0%**
MHA	**6.3%** (9/144)	**0%**
MH	**7.6%** (11/144)	**0%**

Notably, the Ag surface classifies thiolated flavor molecules (MHA, MH) with higher accuracy due to formation of the strong Ag-thiolate bond bringing them close to the plasmonic surface for SERS enhancement. In contrast, non-thiolated flavor molecules (menthol, linalool, limonene) do not interact well with the Ag surface and are prone to misclassification. These prediction outcomes demonstrate that our SERS taster effectively predicts the identity of an unknown flavor molecule with high accuracy through analysis of its SERS super-profile.

2.2.4 Multiplex flavor quantification

To enhance the applicability of our SERS taster in actual flavor analysis, we demonstrate the ability of our SERS taster to simultaneously quantify two flavor molecules in an artificial wine matrix as a proof-of-concept (**Figure 2-13**). We select MHA and MH as both are sulfur-containing molecules that exhibit intense fruity notes, concurrently found in many wine types, such as Sauvignon blanc and Cabernet Sauvignon.[12]

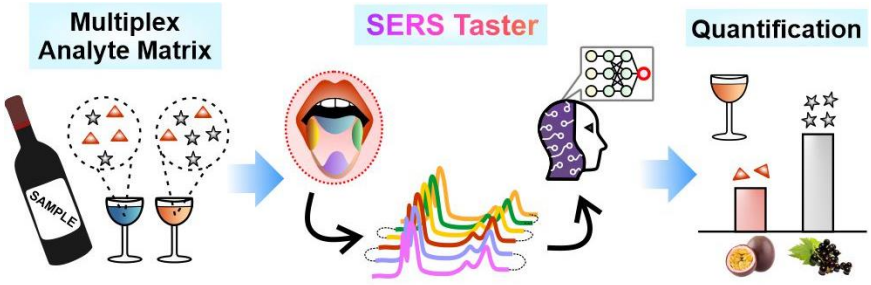


Figure 2-13. Scheme depicting multiplex quantification of flavors using our SERS Taster.

Using SVM regression (SVMR), we construct calibration curves to compare the quantification accuracy of our SERS taster and a single receptor platform (MBA). The flavor concentrations range from 2 – 10 μM (approximately 0.2 – 2 ppm, **Figure 2-14**). For our SERS taster, the calibration curves show near ideal linearity with high prediction coefficients of 0.998 for both MHA and MH (**Figure 2-15Ai**). In contrast, using only MBA yields prediction coefficients of 0.964 and 0.952 for MHA and MH respectively (**Figure 2-15Aii**).

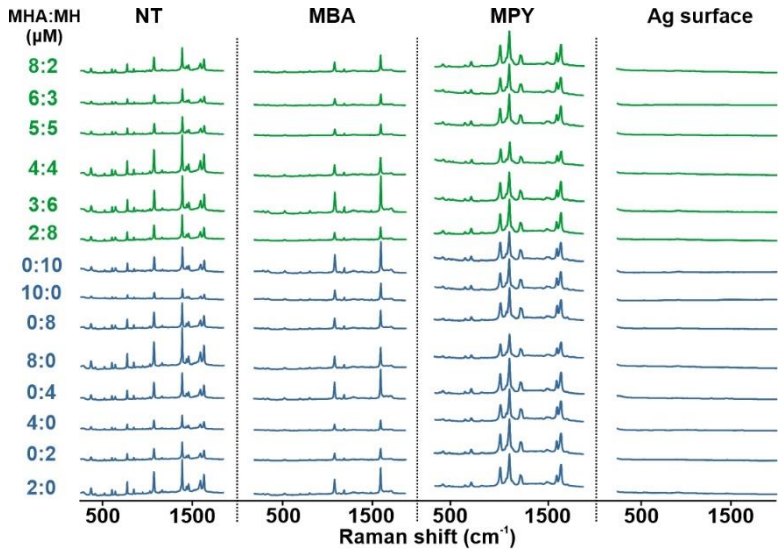
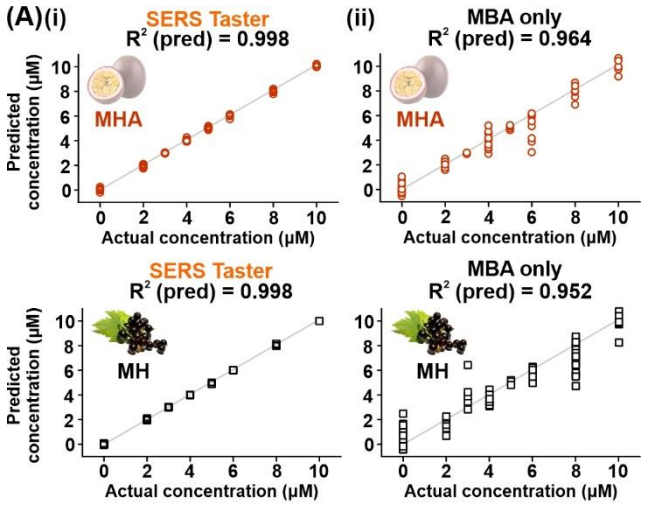


Figure 2-14. SERS spectra involving varying concentrations of MHA and MH within artificial wine used to construct calibration curves using SVMR.



(B)

Artificial wine #	Actual [MHA] (μM)	Predicted [MHA] (μM) (% difference) SERS Taster	MBA	Actual [MH] (μM)	Predicted [MH] (μM) (% difference) SERS Taster	MBA
1	2	2.1 (+5%)	2.4 (+20%)	8	7.7 (–4%)	6.9 (–14%)
2	3	2.9 (–3%)	3.3 (+10%)	6	5.8 (–3%)	5.2 (–13%)
3	4	3.9 (–3%)	3.3 (–18%)	4	4.1 (+3%)	3.4 (–15%)
4	5	5.1 (+2%)	4.5 (–10%)	5	5.0 (0%)	5.7 (+14%)
5	6	6.0 (0%)	6.6 (+10%)	3	3.1 (+3%)	3.4 (–13%)
6	8	7.8 (–3%)	7.1 (–11%)	2	2.0 (0%)	2.3 (+15%)

Figure 2-15. Multiplex quantification of flavors (MHA and MH) in artificial wine. (A) Calibration curves obtained using SVMR for MHA and MH using (i) our SERS Taster and (ii)

a single receptor platform (MBA). (B) Comparison of quantification accuracy between our SERS Taster and a single receptor (MBA) platform using six artificial wine samples with varying concentrations of MHA and MH. For each sample, the predicted flavor concentration, and its deviation from the actual concentration (% difference) is shown.

Next, we prepare six artificial wine samples with varying concentrations of both flavors and expose them to both platforms. Using the calibration curves, our SERS taster show excellent quantification accuracies for both flavors, ranging from 95 – 100% (**Figure 2-15B**). Notably, the difference between actual and predicted flavor concentrations in the artificial wine samples 0 – 0.3 μM. In contrast, we observe lower quantification accuracies of 80 – 90% using only BA. The large range also indicates higher inconsistencies arising from the model's inability to pick up minute spectral changes. The predicted concentrations also exhibit larger deviations of 0.3 – 1.1 μM. Crucially, we show that our SERS Taster exhibits enhanced sensitivity to fluctuations in flavor concentrations as opposed to conventional single receptor platforms.

2.3 Conclusion

In conclusion, we have designed a machine learning-driven multi-receptor SERS taster that enables multiplex profiling of five wine flavors with 100% accuracies at parts-per-million levels. Notably, our two-pronged approach utilizes multiple molecular receptors to generate rich SERS spectral variances and machine learning-driven chemometric models to extract these variances with unparalleled precision. First, the use of four targeted receptors effectively captures a more complete spectroscopic profile of each flavor molecule through unique receptor-flavor chemical interactions that induce distinct spectral variations. By strategically combining all receptor SERS spectra, we construct compound SERS super-profiles

encompassing these interactions that collectively aid in the re-construction of a flavor chemical profile. Next, using PCA and SVMDA, we exemplify the importance of our multi-receptor approach where only our SERS Taster achieves unambiguous identification of all five flavors. Crucially, we elucidate the complex PCA scores by examining the PCA biplot, establishing a robust correlation of the chemometric output with our knowledge of chemical interactions occurring at the molecular level. Importantly, we demonstrate the ability of our SERS Taster in distinguishing primary, secondary and tertiary alcohols. We further highlight the promising potential of our SERS Taster in achieving multiplex quantification of wine flavors within an artificial wine matrix with potential interferences, showcasing high quantification accuracies up to 100%. A comparison of these results with similar platforms using only a single receptor clearly illustrates the superiority of our SERS Taster in identifying and quantifying wine flavors. This combination of SERS with machine learning-driven chemometrics thus creates a rapid and highly sensitive analytical approach for multiplex detection of small molecules. Our SERS Taster tackles current limitations faced in chemical analysis of flavor compounds, providing a potential paradigm shift for food-related studies and a myriad of applications extending beyond.

2.4 Materials and methods

Synthesis and purification of Ag nanocubes. Ag nanocubes were synthesized in high yield using the polyol reduction method.[32] 20 mL of 1,5-pentanediol was added to a 100 mL round-bottom flask and heated to 190 °C for 10 min. Aliquots of 250 μL of poly(vinylpyrrolidone) and 500 μL of $AgNO_3$ precursor solutions were then added in alternation to the reaction flask until the reaction mixture turned reddish-brown. The reaction mixture was repeatedly washed by ethanol and centrifuged before being subjected to vacuum filtration using polyvinylidene

fluoride filter membranes (Durapore®) with pore sizes 5 μm, 0.65 μm, 0.45 μm and 0.22 μm to remove impurities.

Self-assembly of Ag nanocubes using the Langmuir-Blodgett technique. Oxygen plasma (FEMTO SCIENCE, CUTE-MP/R, 100W) was used to clean Si substrates (2 μm × 2.5 μm in size) for 5 min before immersing into the Langmuir-Blodgett trough (KSV NIMA, KN1002). The surface pressure was zeroed prior to the addition of Ag nanocubes. 700 μL of purified Ag nanocubes were dispersed in 1050 μL of chloroform and carefully added to the surface of the water. The mechanical barrier was then gradually adjusted at a fixed compression rate of 2 mm/s till the surface pressure reached 16 mN/m. The substrate was then lifted at a fixed rate of 2 mm/s while maintaining the surface pressure.

Surface functionalization of Ag nanocubes. Each receptor (NT, MBA, MPY) solution were prepared as separate 10 mM solutions in 1:1 ethanol/2-propanol. The substrates were then immersed in 5 mL of a single receptor solution for at least 12 hours. The substrates were removed and carefully washed with ethanol. To prepare the unfunctionalized Ag surface, the substrate was immersed in 10 mL of 0.5 M KI for 30 minutes. The substrate was then removed, washed, and used for SERS measurements immediately.

Platform characterization. Scanning electron microscopy (SEM) was performed using a JEOL-JSM-7600F microscope at an accelerating voltage of 5 kV. UV-vis spectroscopy was performed using a Cary 60 UV-vis spectrometer. SERS measurements were performed using x-y imaging mode of the Ramantouch microspectrometer (Nanophoton Inc., Osaka, Japan) with a 532 nm excitation laser (power = 0.4 mW). A 50× (N.A. = 0.55) objective lens was used

with 10 s acquisition time for data collection. All SERS spectra were obtained by averaging at least 120 individual SERS spectra within the SERS image.

SERS measurements of flavors. The functionalized substrates were individually immersed in 200 μL of aqueous analyte and measured separately. SERS measurements were performed using a hyperspectral x-y imaging mode with an acquisition time of 10 s per line and a laser power of 0.4 mW.

Density functional theory (DFT) simulations. The DFT simulations were carried out using the unrestricted B3LYP exchange-correlation functional in the Gaussian 09 computational chemistry package. The LANL2DZ basis set was used for Ag while the 6-31G(d,p) basis set was used for all other atoms. The Ag surface was modeled using a reported triangle comprising six Ag atoms. The triangular Ag cluster was first geometrically optimized before placing each receptor molecule (NT, MBA, MPY) at the vertex. The whole system was then relaxed with all Ag atoms fixed. Finally, the analyte molecule was placed near the receptor before allowing the whole system to relax with all Ag atoms fixed again.

Constructing SERS super-profiles. The spectral range selected for analysis ranged from 250 to 2000 cm^{-1} for a single SERS spectrum. Two SERS spectra were horizontally combined by arithmetically adding a constant value of 2000 to the wavenumber values of the second spectra. This is repeated up to four SERS spectra. The wavenumbers of the compound spectrum can be correlated back to the original wavenumber values by subtracting the constant value added. The gap between each SERS spectrum (0 – 250 cm^{-1}) is ignored in the analysis. 24 SERS spectra were collected for each flavor per receptor, totaling to 576 SERS spectra for 5 flavors

+ 1 flavorless control and 4 receptors (24 × 6 × 4). These spectra are then combined to form 144 SERS super-profiles (576 ÷ 4).

Machine learning analysis. Machine learning analyses (PCA, SVMDA, SVMR) were conducted using SOLO v8.8 (Stand Alone Chemometrics Software, Eigenvector Research, Inc.). For all models, we apply a standardized set of pre-processing methods which include baseline correction using the automatic weighted least squares method, extended multiplicative scatter correction, normalization, and median centering. All models were cross-validated using venetian blinds, with 10 splits and a blind thickness of 1.

Multiplex flavor quantification. The artificial wine matrix comprises 86% water, 12% ethanol, 1% glycerol (to represent sugars) and 1% tartaric acid (to represent acids).[33] A total of 14 different combinations of flavor concentrations were tested (shown in **Figure 2-14**). For each artificial wine sample, 16 SERS super-profiles were constructed, totaling 224 SERS super-profiles. This data set is then split into a calibration set comprising 142 SERS super-profiles (or SERS spectra for the single receptor) and a validation set comprising 82 SERS super-profiles. Finally, 8 SERS super-profiles were individually constructed for 6 'unknown' artificial wine samples as the test data set.

References

1. Kao, Y.-C.; Han, X.; Lee, Y. H.; Lee, H. K.; Phan-Quang, G. C.; Lay, C. L.; Sim, H. Y. F.; Phua, V. J. X.; Ng, L. S.; Ku, C. W.; Tan, T. C.; Phang, I. Y.; Tan, N. S.; Ling, X. Y., *ACS Nano*, 2020, 14, 2542-2552.
2. Phan-Quang, G. C.; Yang, N.; Lee, H. K.; Sim, H. Y. F.; Koh, C. S. L.; Kao, Y.-C.; Wong, Z. C.; Tan, E. K. M.; Miao, Y.-E.; Fan, W.; Liu, T.; Phang, I. Y.; Ling, X. Y., *ACS Nano*, 2019, 13, 12090-12099.
3. Xu, M.-L.; Gao, Y.; Han, X. X.; Zhao, B., *J. Agric. Food Chem.*, 2017, 65, 6719-6726.
4. Lee, H. K.; Lee, Y. H.; Koh, C. S. L.; Phan-Quang, G. C.; Han, X.; Lay, C. L.; Sim, H. Y. F.; Kao, Y.-C.; An, Q.; Ling, X. Y., *Chem. Soc. Rev.*, 2019, 48, 731-756.
5. Koh, C. S. L.; Lee, H. K.; Han, X.; Sim, H. Y. F.; Ling, X. Y., *Chem. Commun.*, 2018, 54, 2546-2549.
6. Leong, S. X.; Koh, L. K.; Koh, C. S. L.; Phan-Quang, G. C.; Lee, H. K.; Ling, X. Y., *ACS Appl. Mater. Interfaces*, 2020, 12, 33421-33427.
7. Dina, N. E.; Gherman, A. M. R.; Chiş, V.; Sârbu, C.; Wieser, A.; Bauer, D.; Haisch, C., *Anal. Chem.*, 2018, 90, 2484-2492.
8. Shin, H.; Oh, S.; Hong, S.; Kang, M.; Kang, D.; Ji, Y.-g.; Choi, B. H.; Kang, K.-W.; Jeong, H.; Park, Y.; Hong, S.; Kim, H. K.; Choi, Y., *ACS Nano*, 2020, 14, 5435-5444.
9. Ho, C.-S.; Jean, N.; Hogan, C. A.; Blackmon, L.; Jeffrey, S. S.; Holodniy, M.; Banaei, N.; Saleh, A. A. E.; Ermon, S.; Dionne, J., *Nat. Commun.*, 2019, 10, 4927.
10. Lussier, F.; Missirlis, D.; Spatz, J. P.; Masson, J.-F., *ACS Nano*, 2019, 13, 1403-1411.
11. Fariña, L.; Boido, E.; Carrau, F.; Versini, G.; Dellacassa, E., *J. Agric. Food Chem.*, 2005, 53, 1633-1636.
12. Tominaga, T.; Niclass, Y.; Frérot, E.; Dubourdieu, D., *J. Agric. Food Chem.*, 2006, 54, 7251-7255.

13. Rodríguez-Bencomo, J. J.; Schneider, R.; Lepoutre, J. P.; Rigou, P., *J. Chromat. A*, 2009, 1216, 5640-5646.

14. Picard, M.; Franc, C.; de Revel, G.; Marchand, S., *Anal. Chim. Acta*, 2018, 1001, 168-178.

15. Lattey, K. A.; Bramley, B. R.; Francis, I. L., *Aust. J. Grape Wine Res.*, 2010, 16, 189-202.

16. Williams, P. J.; Strauss, C. R.; Wilson, B., *J. Agric. Food Chem.*, 1980, 28, 766-771.

17. López-Vázquez, C.; Bollaín, M. H.; Moser, S.; Orriols, I., *J. Agric. Food Chem.*, 2010, 58, 9657-9665.

18. Siebert, T. E.; Barter, S. R.; de Barros Lopes, M. A.; Herderich, M. J.; Francis, I. L., *Food Chem.*, 2018, 256, 286-296.

19. Phan-Quang, G. C.; Lee, H. K.; Phang, I. Y.; Ling, X. Y., *Angew. Chem. Int. Ed.*, 2015, 54, 9691-9695.

20. Gui, J. Y.; Lu, F.; Stern, D. A.; Hubbard, A. T., *J. Electroanal. Chem. Interfacial Electrochem.*, 1990, 292, 245-262.

21. Lee, Y. H.; Chen, H.; Xu, Q.-H.; Wang, J., *J. Phys. Chem. C*, 2011, 115, 7997-8004.

22. Lee, H. K.; Lee, Y. H.; Morabito, J. V.; Liu, Y.; Koh, C. S. L.; Phang, I. Y.; Pedireddy, S.; Han, X.; Chou, L.-Y.; Tsung, C.-K.; Ling, X. Y., *J. Am. Chem. Soc.*, 2017, 139, 11513-11518.

23. Zhang, L.; Bai, Y.; Shang, Z.; Zhang, Y.; Mo, Y., *J. Raman Spectrosc.*, 2007, 38, 1106-1111.

24. Ho, C.-H.; Lee, S., *Colloids Surf. A Physicochem. Eng. Asp.*, 2015, 474, 29-35.

25. Agarwal, N. R.; Lucotti, A.; Tommasini, M.; Neri, F.; Trusso, S.; Ossi, P. M., *Sens. Actuators B Chem.*, 2016, 237, 545-555.

26. Polášková, P.; Herszage, J.; Ebeler, S. E., *Chem. Soc. Rev.*, 2008, 37, 2478-2489.

27. Elci, S. G.; Moyano, D. F.; Rana, S.; Tonga, G. Y.; Phillips, R. L.; Bunz, U. H. F.; Rotello, V. M., *Chem. Sci.*, 2013, 4, 2076-2080.

28. Sipos, L.; Kovács, Z.; Sági-Kiss, V.; Csiki, T.; Kókai, Z.; Fekete, A.; Héberger, K., *Food Chem.*, 2012, 135, 2947-2953.

29. Belousov, A. I.; Verzakov, S. A.; von Frese, J., *Chemometr. Intell. Lab Syst.*, 2002, 64, 15-25.

30. Alexandre Marcelo, M. C.; Martins, C. A.; Pozebon, D.; Ferrão, M. F., *Anal. Methods*, 2014, 6, 7621-7627.

31. Dixon, S. J.; Brereton, R. G., *Chemometr. Intell. Lab Syst.*, 2009, 95, 1-17

32. Tao, A.; Sinsermsuksakul, P.; Yang, P., *Angew. Chem. Int. Ed.*, 2006, 45, 4597-4601.

33. Sumby, K. M.; Grbin, P. R.; Jiranek, V., *Food Chem.*, 2010, 121, 1-16.

Chapter 3 Guiding Smart Receptor Selection using a Machine Learning-Driven SERS-based Recommender System for Tailored Structural Analog Differentiation

Abstract. Multiple molecular receptors gather and amplify SERS signal variances from non-covalent receptor-analyte interactions, providing enhanced analyte specificity. Currently, the number and type of receptors are manually determined based on chemical intuition and trial-and-error experimentation. Here, we introduce a ML-driven SERS receptor recommender system (RRS) with a four-stage 'identify, filter, rank and recommend' approach to recommend optimal receptor combinations, achieving 96.6% accuracy in the classification of five haloanisole analogs. Our RRS provides an objective framework to curate input features by assessing their classification importance, maximizing SERS variance while minimizing the curse of dimensionality effect. By constructing a recommender database with all combinations of two to five haloanisole classification problems, our RRS can harness collaborative filtering to extrapolate receptor recommendation for an 'unseen' six haloanisole problem. Our RRS strives to unlock the full potential of data-driven SERS sensing for increasingly challenging practical applications involving structurally similar analytes.

3.1 Introduction

SERS is an attractive spectroscopic technique often employed for trace analyte sensing due to its ability to provide fingerprint-specific readout. The inherent surface sensitive nature of SERS underscores the significance of using molecular receptors, particularly when dealing with analytes with poor affinity to plasmonic surfaces and/or have small Raman cross-sections.[1-3] These receptors play a critical role in recognizing, chemically interacting with, and confining analytes for effective SERS enhancement, as well as transducing receptor-analyte interactions as receptor signal changes. Traditionally, highly selective receptors that form complementary interactions with a single analyte, such as the antigen-antibody pair, are appealing for their signal specificity. However, generic molecular receptors that can concurrently form non-covalent interactions with a range of analytes have gained popularity in recent years because of their high flexibility and ability to probe different facets of the target analyte's functionality. Furthermore, by serially combining multiple receptor-analyte SERS spectra into a 'super-profile', we can gather and amplify SERS signal variations detected by each receptor. ML predictive models subsequently harness these variations to elucidate complex analyte interrelationships and enable molecular identification and quantification with enhanced specificity even in the case of isomeric compounds.[4-6]

At present, the number and type of molecular receptors are manually determined based on chemical intuition and trial-and-error experimentation, and this process is subjective and prone to human biases. It is critical to establish a smart framework to objectively select the types of receptors for SERS measurements that can best interact with target analytes and give rise to optimal and distinctive signal variations. Additionally, the number of receptors also play an important role. If the receptors are too few or too similar, there will not be sufficient SERS signal variation and information to differentiate the analytes. Conversely, if too many receptors are involved, the number of SERS features would greatly overwhelm the number of samples,

leading to a problem known as the 'curse of dimensionality' (CoD). CoD is known to reduce the predictive accuracy of ML models in classifying or quantifying analytes because the large proportion of uninformative background noise or unimportant features hinders the precise retrieval of key spectral variances arising from the analytes.[7-8] This raises a key question pertaining the research gap in molecular receptor design – which and how many receptors are sufficient?

Herein, we introduce a ML-driven SERS receptor recommender system (RRS) framework that effectively selects the optimal number and types of molecular receptors for SERS detection and achieve 96.6% accuracy in the classification of five haloanisole structural analogs (**Figure 3-1**). Inspired by recommender systems ubiquitous in search engines, social media, and e-commerce, our SERS RRS comprises a four-stage 'identify, filter, rank and recommend' algorithmic approach. Structurally, our target haloanisoles – 2,4-dichloroanisole (2,4-DCA), 2,6-dichloroanisole (2,6-DCA), 3,5-dichloroanisole (3,5-DCA), 2,4,6-trichloroanisole (2,4,6-TCA), and 2,3,5,6-tetrachloroanisole (2,3,5,6-TeCA) – each possess an anisole backbone but differ by their halogen substitution number and positions. These haloanisoles are commonly associated with musty odors in wine, negatively affecting their quality.[9] For receptors, we shortlist nine thiolated molecules with diverse functional groups, including amino (NH2), bromo (Br), carboxylic acid (COOH), methyl (CH3), hydroxyl (OH), boronic acid (B(OH)2), pyridine (PY), aldehyde (CHO) and naphthalene (NT). Our experimental and *in silico* SERS studies affirm that non-covalent interactions between the receptor-haloanisole pairs do induce significant variations in the resulting SERS spectra.

Our four-stage approach begins with an 'identify' stage where our RRS identifies and groups SERS peaks with a normalized intensity above 0.01 as feature groups. Each feature group represents a Raman vibrational mode and can be correlated to chemical interactions at the molecular level. Subsequently, in the 'filter' stage, our RRS removes the signal-less

background regions that remain unchanged before and after analyte exposure. In the 'rank' stage, our RRS ranks feature groups across all nine receptors using their eXtreme gradient boosting tree (XGBoost) importance scores. Finally, in the 'recommend' stage, our RRS recommends the receptor combination that attains the highest XGBoost classification accuracy.

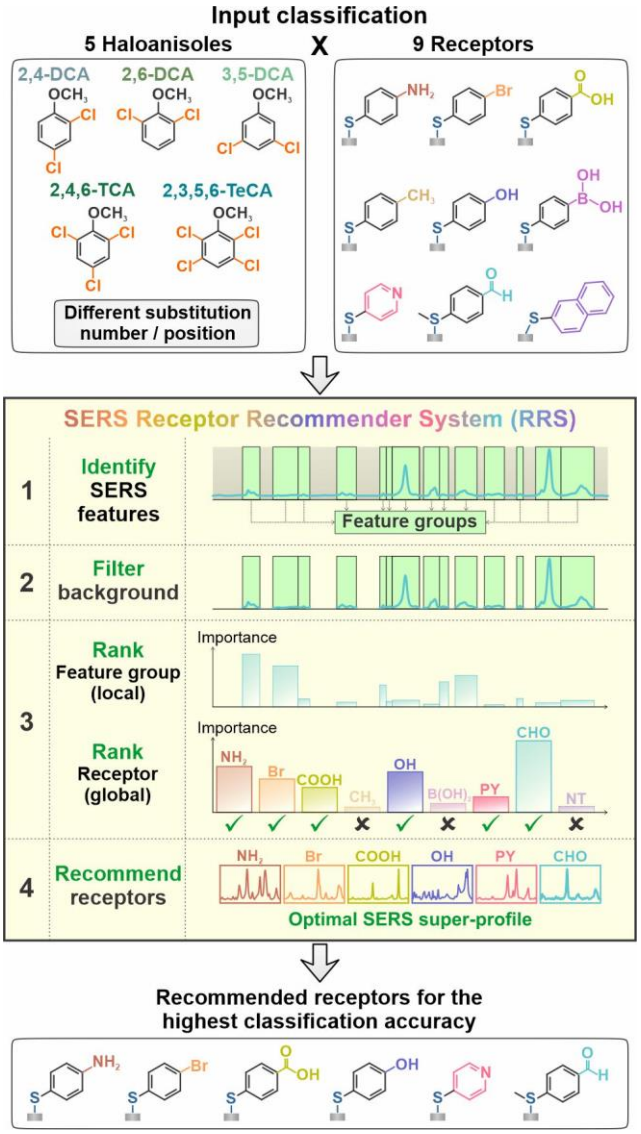


Figure 3-1. Our RRS is a ML-driven framework that adopts a four-stage 'identify, filter, rank and recommend' approach to analyze SERS data and select the optimal number and type of receptors for SERS detection. The framework identifies and filters SERS peaks that correspond to Raman vibrational modes, before ranking and selecting receptors based on their classification importance.

We further construct a recommender database using receptor importance scores from all combinations of two to five haloanisole classifications. This allows our RRS to leverage a k-nearest neighbor (kNN) collaborative filtering approach by retrieving similar classifications within the database and predicting receptor importance for a six-haloanisole classification involving an untrained haloanisole. We demonstrate the ability to derive a seven-receptor combination that achieves the best accuracy in classifying the six haloanisoles, effectively providing receptor recommendations even before collecting experimental data. Overall, our SERS RRS provides an objective framework for smart receptor selection to unlock the full potential of data-driven ML-SERS sensing, which is invaluable in tackling increasingly challenging practical applications involving analytes with a high degree of structural similarity.

3.2 Results and discussion

3.2.1 Overview of our SERS RRS framework

Our SERS RRS is a ML-driven framework that employs a four-stage 'identify, filter, rank and recommend' approach to analyze SERS spectra and select an optimal set of receptors that most effectively capture SERS signal variances for the classification of structurally similar haloanisole analogs. To begin, we prepare a series of receptor-functionalized SERS substrates by independently grafting each of the nine thiolated molecular receptors onto an array of drop-casted Ag nanocubes (edge length = 115 ± 5 nm, purity = 90%, **Figure 3-2**). These receptors contain a myriad of functional groups, namely amino (4-aminothiophenol, NH_2), bromo (4-bromothiophenol, Br), carboxylic acid (4-mercaptobenzoic acid, COOH), methyl (4-methylbenzenethiol, CH_3), hydroxyl (4-mercaptophenol, OH), boronic acid (4-mercaptophenylboronic acid, $B(OH)_2$), pyridine (4-mercaptopyridine, PY), aldehyde (4-(methylthio)benzaldehyde, CHO) and naphthalene (2-naphthalenethiol, NT). We observe highly homogeneous receptor SERS signals with a low signal standard deviation ranging from

4.03 – 6.49% based on their respective C-S stretching vibrational mode spanning over a large area of 5 mm^2 across 100 substrates (**Figure 3-3**).

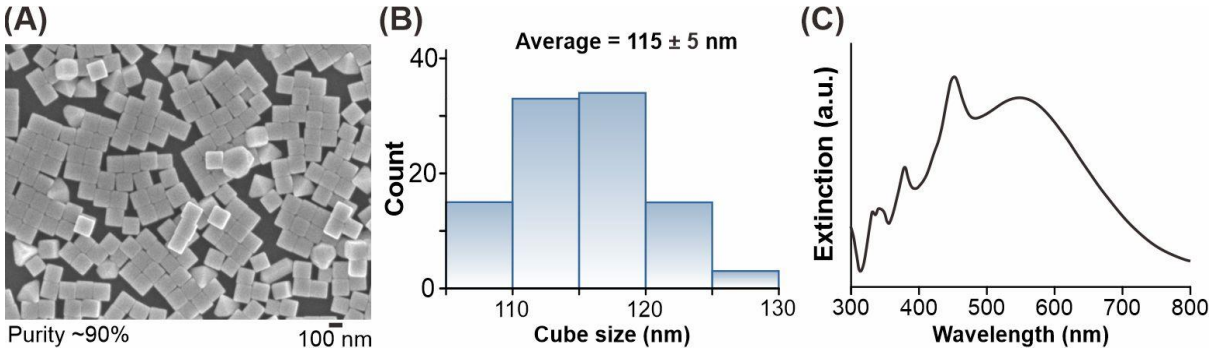


Figure 3-2. (A) SEM image of synthesized and purified Ag nanocubes (purity ~90%) using the polyol reduction method. (B) Size distribution of the Ag nanocubes. (C) UV-vis spectrum of the Ag nanocubes. The peaks at 358, 411, 512 and 648 nm can be assigned to octupole (358 nm), quadrupole (411, 512 nm) and dipole (648 nm) resonances.

We select haloanisoles as our analyte because of their high structural similarity since they each possess an anisole backbone and only differ by the number and position of their chlorine substituents (**Figure 3-4A**). For instance, 2,4-DCA, 2,6-DCA, and 3,5-DCA each have two chlorine substituents at different positions, 2,4,6-TCA has three, and 2,3,5,6-TeCA has four chlorine substituents. We postulate that the high electronegativity of these chlorine substituents will exert significant steric and electronic repulsion when adjacent to the methoxy group, which is expected to be the major interacting site with complementary receptor groups. This will in turn influence the energetically favorable orientation of the receptor-haloanisole interaction complex, as the chlorine substituents will maximize their distance from the aromatic ring of the receptor. In addition, the chlorine substituents exert stronger electron withdrawing effects with an increasing number of chlorines attached to the aromatic moiety of the haloanisoles. As a result, the combination of both the steric and electronic repulsion as well as

the electron withdrawing effects will weaken the non-covalent receptor-haloanisole interactions, including hydrogen bonding, halogen bonding, dipole interactions, aromatic donor-acceptor interactions and van der Waals' interaction. Hence, depending on the number and position of these chlorine substituents on the haloanisoles, the varying strengths of the non-covalent interactions will induce haloanisole-specific receptor peak shifts and intensity fluctuations that will allow us to differentiate the haloanisoles.

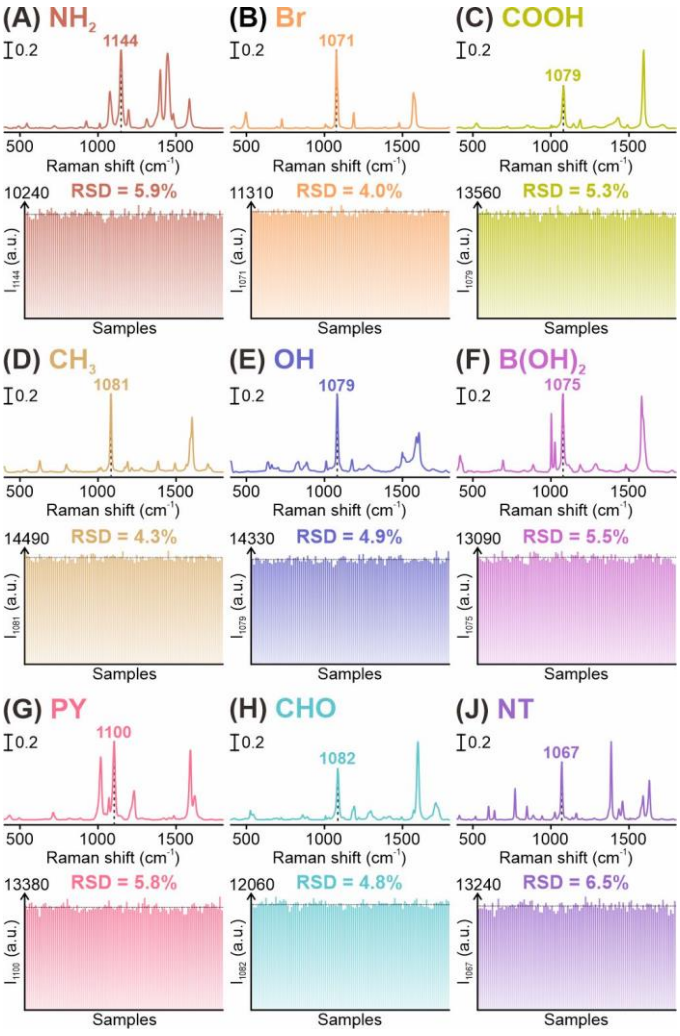


Figure 3-3. For each receptor, a representative SERS spectrum and a bar chart indicating the relative signal standard deviation across 100 substrates is shown – (A) NH_2, (B) Br, (C) COOH, (D) CH_3, (E) OH, (F) $B(OH)_2$, (G) PY, (H) CHO, (J) NT.

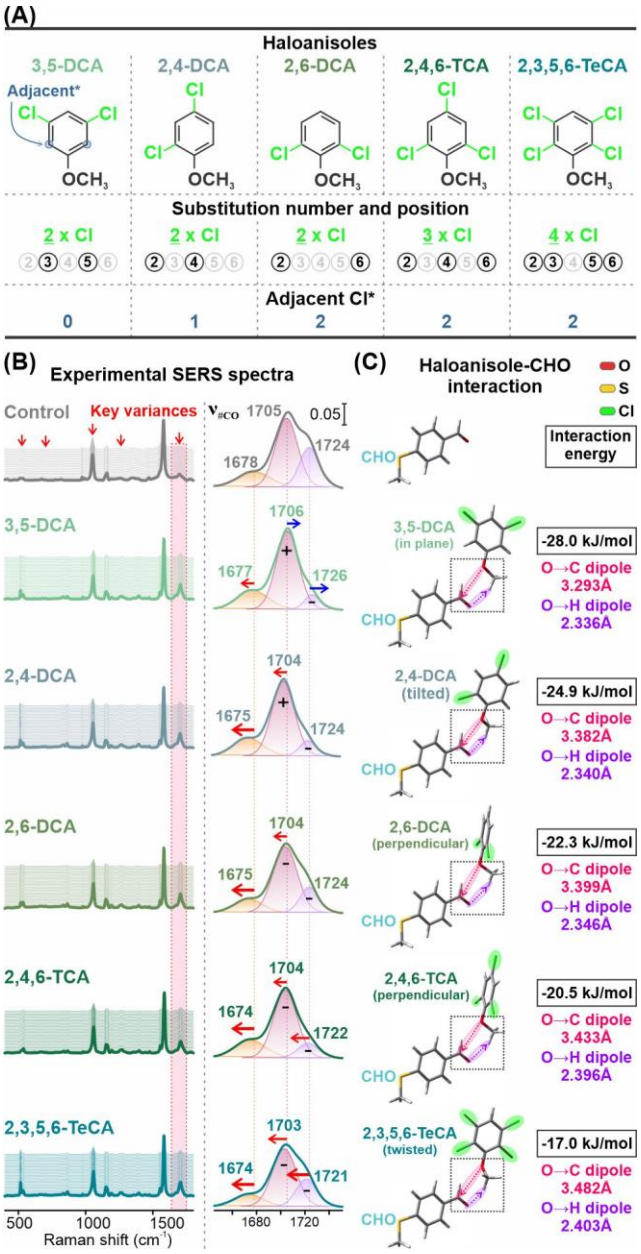


Figure 3-4. In-depth SERS spectral analysis using the CHO receptor. (A) Comparing the structural differences between the haloanisole analogs. (B) SERS spectra of the CHO receptor in the absence (control) and in the presence of each haloanisole. Detailed peak shifts and intensity changes in the 1644 – 1755 cm^{-1} region are shown, which contains peaks indexed to the aldehyde C=O stretching vibration. (C) Configurations of the interaction complex simulated using density functional theory (DFT). The descriptors in brackets describe the relative orientation of the haloanisole chlorine substituents to the aromatic ring of the CHO receptor.

3.2.2 Establishing the chemical meaning behind our RRS

To investigate changes in receptor SERS spectra after receptor-haloanisole interactions, we acquire 1620 SERS spectra by measuring receptor spectra in the absence of haloanisole (control) and in the presence of each of the five haloanisoles individually as the foundation of our RRS framework (**Figure 3-4B, Figure 3-5**). We use CHO as our model receptor for the discussion and focus on regions within the experimental CHO SERS spectrum with noticeable peaks, deconvoluting all overlapping peaks to identify their corresponding Raman vibrational modes based on density functional theory (DFT) simulations (**Figure 3-6, Table 3-1**).

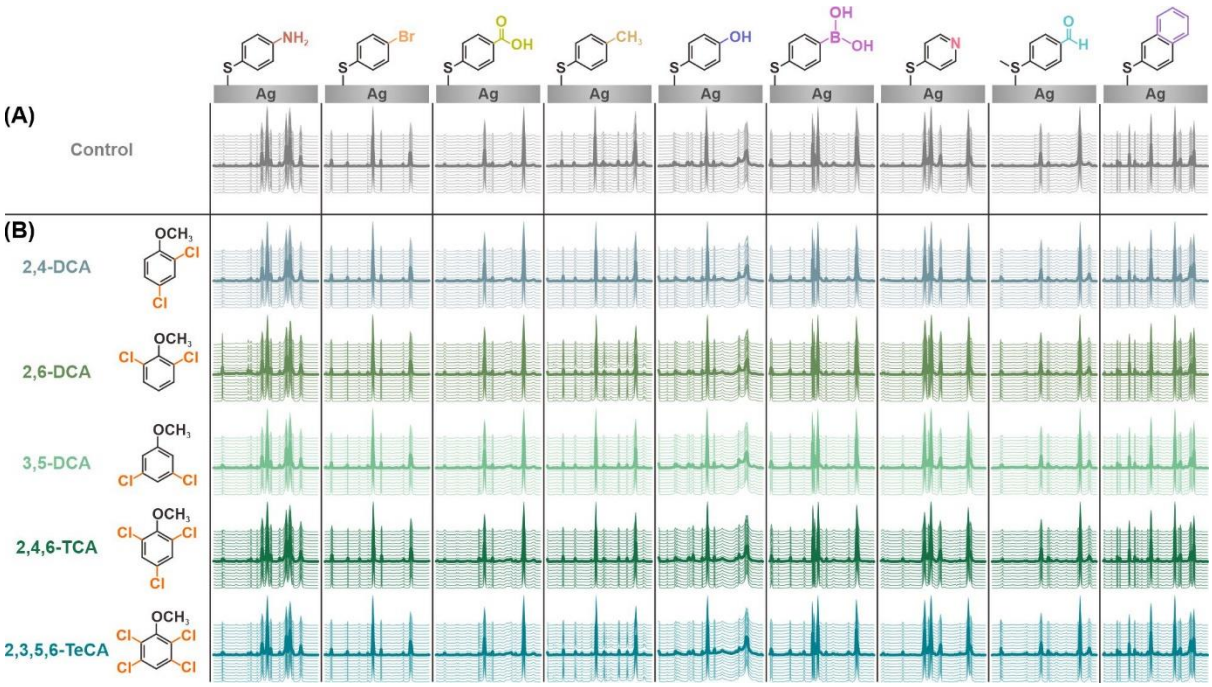


Figure 3-5. Experimental SERS spectra. (A) SERS spectra of receptors in the absence of haloanisoles (control). (B) SERS spectra of each receptor in the presence of each haloanisole.

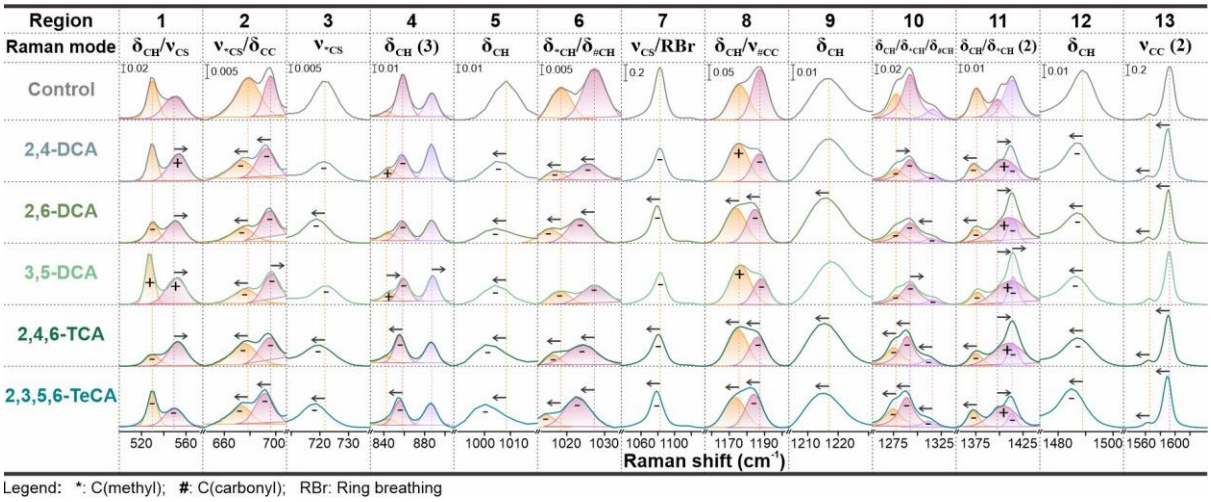


Figure 3-6. Key SERS variances with the CHO receptor. Regions in the experimental CHO SERS spectrum with observable peaks that are deconvoluted and assigned with their respective Raman vibrational modes based on DFT simulations. Specific peak variations in the presence of each haloanisole are shown and compared with the control.

Table 3-1. CHO receptor Raman vibrational modes. List of Raman vibrational modes for the CHO receptor in the absence of haloanisoles (control). The symbols – asterisk (*) and hash (#) – are used to denote the carbon of the methyl group and the carbon of the aldehyde group respectively.

Peak no.	Peak Region (experimental, cm^{-1})	Peak Position (experimental, cm^{-1})	Peak Position (DFT, cm^{-1})	Raman Vibrational Mode
1	500 – 562	526	477	Aromatic CH bending
2		542	510	CS stretching
3	644 – 708	679	671	*CS stretching
4		696	690	Aromatic CC bending
5	709 – 737	721	691	*CS stretching
6	826 – 908	852	793	Aromatic CH bending (wagging)
7		858	813	
8		886	815	
9	994 – 1010	1004	949	Aromatic CH bending (twisting)
10	1011 – 1034	1018	957	*CH bending (wagging)
11		1026	994	#CH bending (wagging)
12	1035 – 1136	1082	1056	CS stretching

13		1117	1097	Aromatic ring breathing
14	1156 – 1203	1176	1149	Aromatic CH bending (scissoring)
15		1186	1193	#CC stretching
16	1204 – 1231	1218	1276	Aromatic CH bending (rocking)
17	1252 – 1340	1280	1293	Aromatic CC stretching (asymmetric)
18		1292	1317	*CH bending (wagging)
19		1315	1370	#CH bending (scissoring)
20	1354 – 1442	1375	1395	Aromatic CH bending (scissoring)
21		1402	1420	*CH bending (twisting)
22		1414	1434	*CH bending (scissoring)
23	1473 – 1503	1489	1470	Aromatic CH bending
24	1530 – 1643	1566	1548	Aromatic CC stretching (asymmetric)
25		1593	1580	Aromatic CC stretching (symmetric)
26	1644 – 1755	1678	1726	#CO stretching (H-bonding)^
27		1705		#CO stretching (Self-associated)^
28		1724		#CO stretching (Isolated)^

^ The #C=O stretching modes are split into three peaks experimentally due to H-bonding with solvent moieties (1678 cm^{-1}) and self-association (1705 cm^{-1}).[10-11]

Overall, we observe distinct CHO spectral variations after haloanisole exposure that clearly distinguishes haloanisoles from the control, as well as multiple CHO peak shifts and intensity fluctuations that allows us to distinguish between different haloanisoles. Specifically, for the control CHO receptor spectra in the absence of haloanisoles, we observe an intrinsic CHO's C=O stretching mode at 1724 cm^{-1} as well as two other C=O stretching modes as a result of receptor-solvent hydrogen bonding at 1678 cm^{-1}, and receptor-receptor self-association at 1705 cm^{-1}.[10-11] Upon exposing the CHO receptor to the haloanisoles, a pair of dipole-dipole interactions form between (1) an electron rich methoxy's oxygen from the haloanisole and an electron poor carbonyl's carbon from CHO (O→C dipole) and (2) an

electron rich carbonyl's oxygen from CHO and an electron poor methoxy's hydrogen from the haloanisole (O→H dipole) (**Figure 3-4C**).

Notably, the pair of dipole interactions are strongest for 3,5-DCA at -28.0 kJ/mol with no adjacent chlorine substituents to the methoxy group, followed by 2,4-DCA at -24.9 kJ/mol with one adjacent chlorine. For 2,6-DCA, 2,4,6-TCA, and 2,3,5,6-TeCA with two adjacent chlorine substituents, their interaction energies are -22.3, -20.5, and -17.0 kJ/mol respectively, which indicates a decrease in dipole interaction strength with an increase in number of chlorine substituents. This trend concurs with the increase in inter-atomic distances between the anisole methoxy's C and CHO carbonyl's C (O→C dipole) as well as between the CHO carbonyl's O and anisole methoxy's H (O→H dipole) from 3,5-DCA with the strongest dipole interactions to 2,3,5,6-TeCA with the weakest. These findings are in close agreement with our postulation that more adjacent chlorine atoms to the methoxy group induces larger steric and electronic repulsion while more chlorine substituents increases the electron withdrawing effect, both of which will decrease the strength of the dipole interactions. As a result, in the experimental SERS spectra of the di-substituted 3,5-DCA with no adjacent chlorines, 2,4-DCA with one adjacent chlorine and 2,6-DCA with two adjacent chlorines, we can first isolate 3,5-DCA based on the blue-shift of the C=O stretching mode from 1724 to 1726 cm^{-1} (**Figure 3-4B**). This is because the CHO-3,5-DCA dipole interactions is the only pair that is strong enough to polarize the CHO bonds and reduce the energy required for the intrinsic CHO C=O stretching mode at 1724 cm^{-1}. When comparing 2,4-DCA and 2,6-DCA, the former shows an increase in the 1705 cm^{-1} peak intensity from 0.195 to 0.211 while the latter shows a decrease in intensity from 0.195 to 0.187. The decrease in peak intensity at 1705 cm^{-1} for 2,6-DCA stems from the perpendicular configuration of its CHO-2,6-DCA complex, aimed to minimize steric and electronic repulsion between the two adjacent chlorine substituents of the haloanisole and the aromatic ring of CHO, but inevitably hinders receptor-receptor interactions. Finally, for the tri-

substituted 2,4,6-TCA, the 1705 and 1724 cm^{-1} peaks red-shifted to 1704 and 1722 cm^{-1} respectively while for the tetra-substituted 2,3,5,6-TeCA the peaks red-shifted to 1703 and 1721 cm^{-1} respectively. As compared to 2,6-DCA, 2,4,6-TCA and 2,3,5,6-TeCA contain one and two more chlorine substituents respectively, and the larger number of chlorines will hinder receptor-solvent and receptor-receptor interactions to a larger extent, causing an increase in energy required for their C=O stretching modes. We establish strong corroboration between our experimental observations and DFT simulations pertaining the CHO-haloanisole interactions and SERS spectral variations, which is crucial in ensuring effective ML-driven classification of the haloanisoles.

Beyond CHO, we show that receptor-haloanisole interactions also give rise to distinct SERS spectral variations for all other receptors. We utilize 2,4,6-TCA as our model haloanisole to illustrate SERS spectral variations arising from the formation of non-covalent interactions across all nine receptors and corroborate our findings with DFT simulations (**Figure 3-7**). Across all nine receptors, the receptor-2,4,6-TCA binding energies of -0.64 – 4.90 kcal/mol is larger than the 2,4,6-TCA-2,4,6-TCA binding energy of -1.69 kcal/mol, which affirms the preferential formation of receptor-2,4,6-TCA interactions (**Table 3-2**). For NH_2, PY and CHO, dipole-dipole interactions occur between the electron rich receptor N/O atoms and the electron deficient methoxy C in 2,4,6-TCA. In NH_2, the presence of 2,4,6-TCA causes a blue-shift of the azobenzene CN bending peak from 923 to 924 cm^{-1} as the C-N bond weakens due to dipole-dipole interactions, confirmed by the increase in C-N bond distance from 1.373 to 1.377 Å.[12] For COOH, OH and $B(OH)_2$, hydrogen bonding occurs between the receptor OH group and the methoxy O in 2,4,6-TCA. In COOH, the presence of 2,4,6-TCA causes a blue-shift of the concurrent CS stretching and CO bending peak from 718 to 719 cm^{-1} as the C=O bond weakens due to hydrogen bond formation, confirmed by the increase in C=O bond distance from 1.216 to 1.220 Å. Their respective hydrogen bond distance of 2.824, 2.823 and 2.972 Å concurs with

their receptor-2,4,6-TCA binding energies of 2.19, 3.27 and 1.90 kcal/mol. For CH_3 and NT, aromatic donor-acceptor and van der Waals' interactions occur between the aromatic rings of the receptors and 2,4,6-TCA, with 2,4,6-TCA being the aromatic acceptor in both cases due to the strongly electron withdrawing chlorine substituents.[13] In NT, the presence of 2,4,6-TCA causes a blue-shift of the concomitant CC and CH bending peak from 884 to 888 cm^{-1} as the aromatic C-C bonds weaken due to aromatic donor-acceptor interactions, confirmed by the increase in the average C-C bond distance from 1.411 to 1.412 Å. The larger aromatic ring system in NT provides stronger aromatic donor-acceptor interactions as compared to MBT, as shown by their receptor-2,4,6-TCA binding energies of -0.64 and 1.91 kcal/mol respectively. For Br, the presence of 2,4,6-TCA induces halogen bonding between the Br of BTP and O of 2,4,6-TCA, causing the peak indexed to concurrent CS and CBr stretching to blue-shift from 1071 to 1072 cm^{-1}.[14] This is attributed to the weakening of the C-Br bond, confirmed by the increase in C-Br bond distance from 1.908Å to 1.910 Å.

Collectively, these experimental peak shifts are referenced using a consistent internal standard Si peak at 520 cm^{-1} that does not shift before and after 2,4,6-TCA exposure and are well-supported by DFT-simulated SERS peak changes (**Figure 3-8**). Importantly, the observation of distinct SERS spectral changes in all nine receptors when exposed to 2,4,6-TCA and the differences in their relative interaction type and strength form the basis of receptor selection using our SERS RRS. Using the knowledge gained from this analysis, we can design a ML framework that leverages the unique chemical information offered by each receptor in the form of SERS peak variations and put together a combination of receptors that best differentiate the structurally similar haloanisoles.

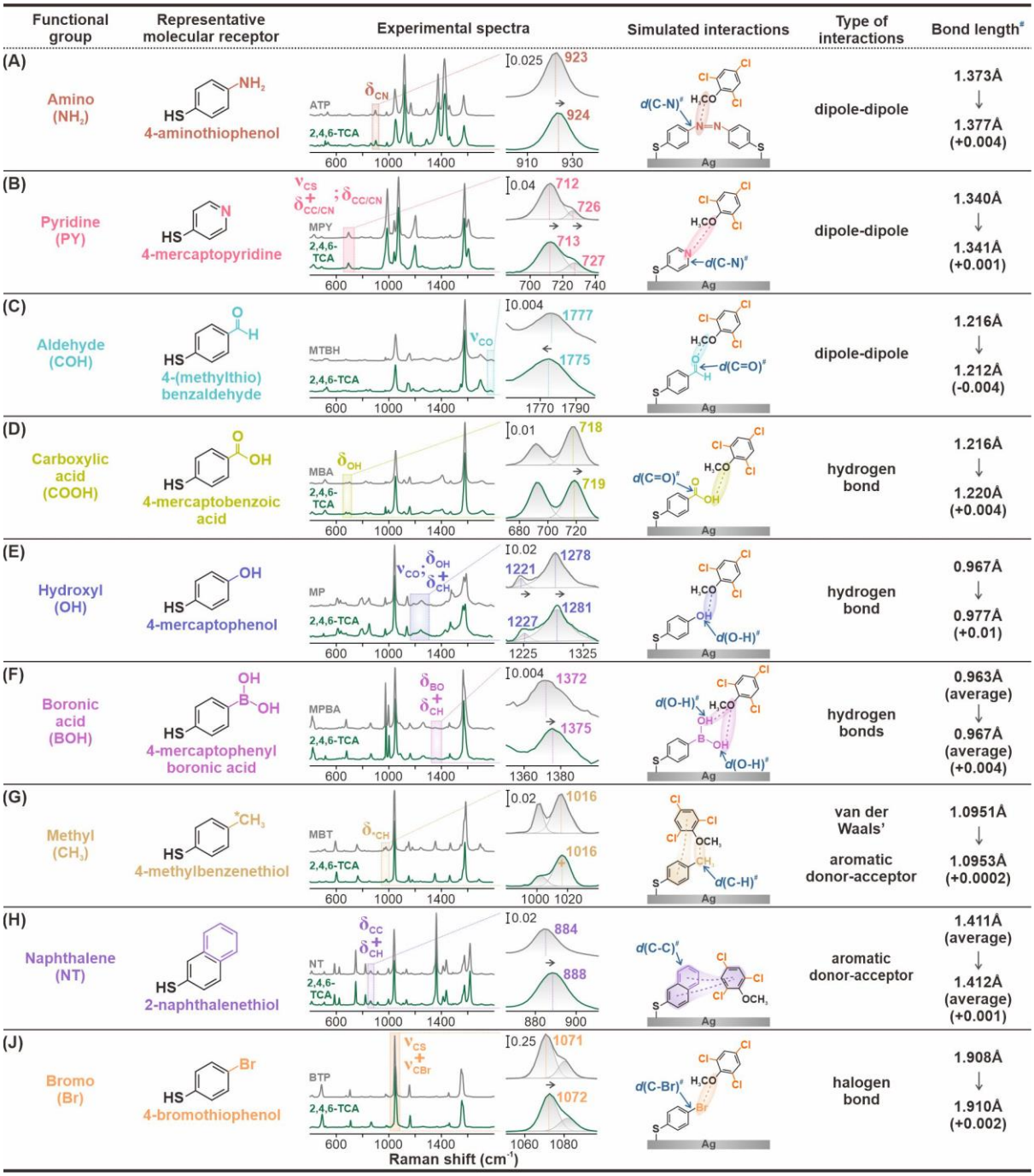


Figure 3-7. Key SERS variances across receptors with 2,4,6-TCA. Comparison of the receptor only and the receptor-2,4,6-TCA SERS spectra for each of the nine receptors – (A) NH_2, (B) PY, (C) CHO, (D) COOH, (E) OH, (F) $B(OH)_2$, (G) CH_3, (H) NT, (J) Br – reveal changes in characteristic receptor vibrational modes, arising from various non-covalent interactions at the molecular level.

Table 3-2. Binding energies of the receptor-2,4,6-TCA and 2,4,6-TCA-2,4,6-TCA complexes. Binding energies are calculated using the following equation:

$$\text{Binding energy (E)} = (E(B3LYP)_{probe} + E(B3LYP)_{analyte}) - E(B3LYP)_{probe\text{-}analyte}$$

A positive binding energy indicates a favorable, exothermic process upon probe-analyte interaction while a negative binding energy indicates an unfavorable process.[5]

Type of complex	Binding energy (kJ/mol)
NH_2-TCA	13.3
Br-TCA	6.2
COOH-TCA	17.5
CH_3-TCA	-2.7
OH-TCA	13.7
$B(OH)_2$-TCA	8.0
PY-TCA	15.2
CHO-TCA	20.5
NT-TCA	12.2
TCA-TCA	-7.1

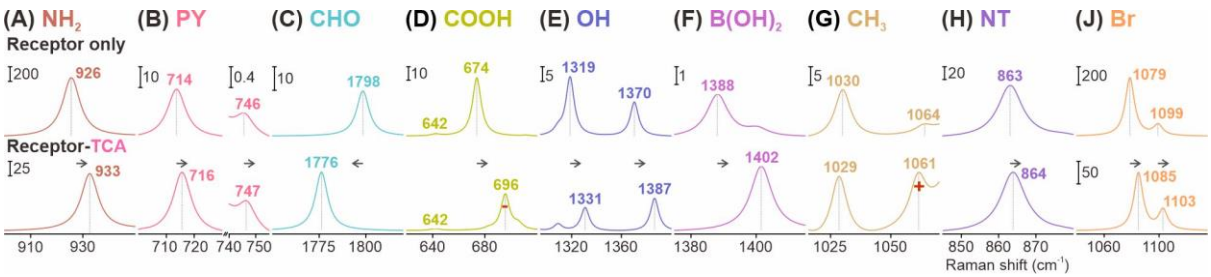


Figure 3-8. DFT-simulated SERS peak shifts. (A) NH_2, (B) Br, (C) COOH, (D) CH_3, (E) OH, (F) $B(OH)_2$, (G) PY, (H) CHO, (J) NT.

3.2.3 Our four-stage 'identify, filter, rank and recommend' approach

To objectively select receptors that provide unique spectral variations induced by receptor-haloanisole interactions, we introduce a strategic four-stage 'identify, filter, rank and recommend' algorithmic approach as part of our RRS framework. The idea is to acquire SERS spectra of receptors after exposure to each haloanisole and extract only key receptor SERS peaks with assigned Raman vibrational modes and analyze their peak shifts and intensity fluctuations. This enables ML algorithms to efficiently determine receptor combinations that produce the most effective cumulative spectral changes in the receptor-haloanisole SERS spectra and therefore achieve the highest accuracy in differentiating the target haloanisoles. To do so, we employ XGBoost as our ML classification model to classify haloanisoles as it is a versatile tree-based algorithm ideal for supervised classification tasks.[15] Additionally, a key advantage in utilizing XGBoost is its built-in function designed to evaluate feature importance by measuring how each input feature contributes to accuracy improvements. For our purpose, this function is particularly important because it enables us to examine and quantify the importance of individual spectral regions within the full spectrum in enhancing the accuracy in haloanisole classification.

In the 'identify' stage, we set out to identify a group of wavenumbers close to an identified spectra peak to form a 'feature group' (**Figure 3-9A**). Crucially, the formation of feature groups allows us to compute the importance of a receptor SERS peak in the haloanisole classification, which can be directly correlated to a specific receptor vibrational mode and hence provide key chemical information relating to the molecular structure of the haloanisoles. Feature groups are formed by setting a threshold of 0.01 normalized intensity to selectively identify SERS peaks after baseline correction and normalization, while avoiding misidentification of instrumental noise as a peak (**Figure 3-10**). Using the full CHO spectrum as input for the XGBoost classification of all five haloanisoles including a control, we attain

63.7 – 72.2% accuracy in classifying each haloanisole and 98.6% for the control (**Figure 3-9B**). Without forming feature groups, the XGBoost feature importance highlights random single wavenumbers that merely describe parts of the CHO peaks and have no chemical meaning.

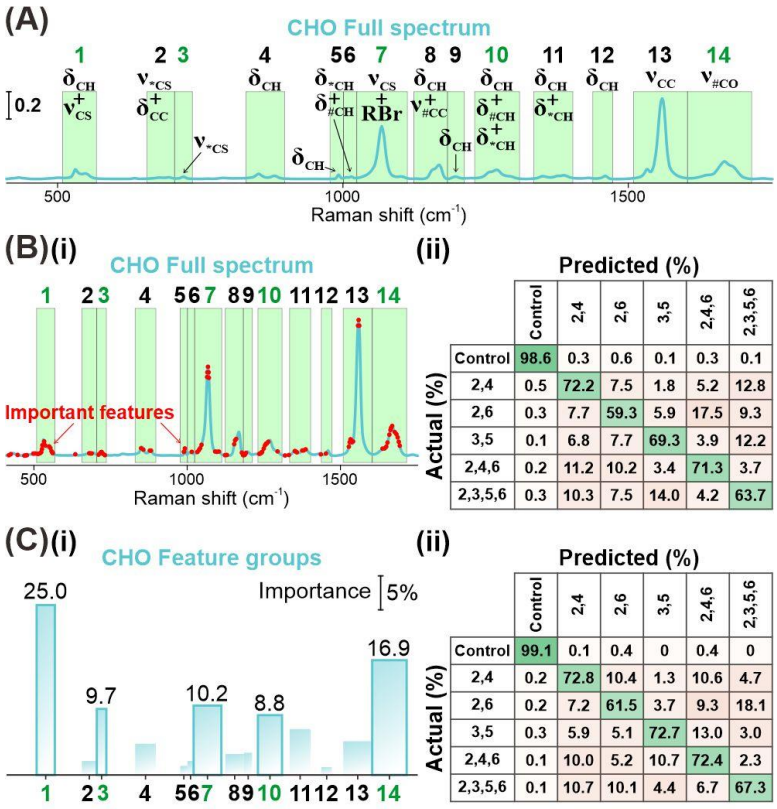


Figure 3-9. Formation and optimization of CHO feature groups. (A) Identifying CHO feature groups and assigning their Raman vibrational modes. (B) XGBoost classification using the full CHO spectrum showing (i) the important wavenumbers and (ii) the confusion matrix. (C) XGBoost classification using only the CHO feature groups showing (i) the important feature groups and (ii) the confusion matrix.

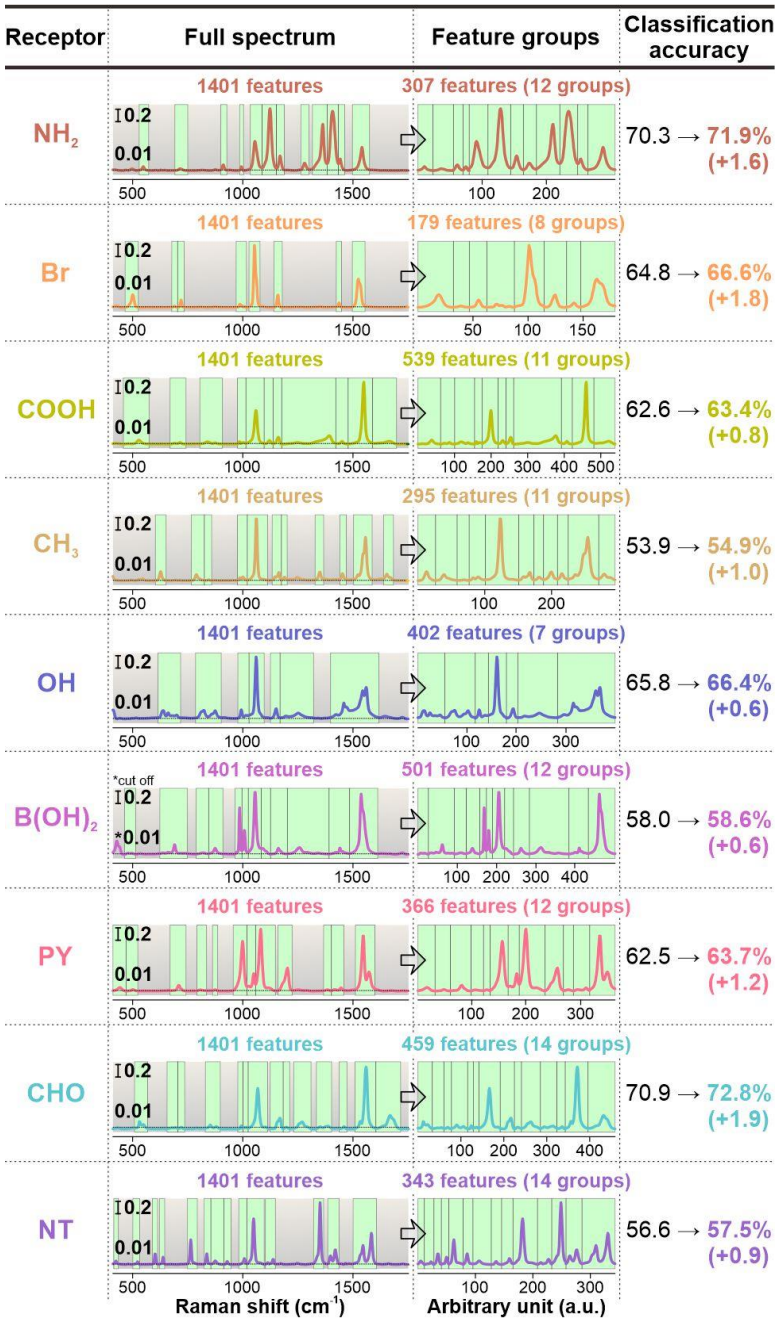


Figure 3-10. SERS spectrum of each receptor before and after removal of background regions. The identified SERS peaks are grouped into feature groups to improve classification accuracy and allow meaningful XGBoost feature importance analyses. The asterisk (*) indicates a twin peak of $B(OH)_2$ that was not considered as the left shoulder was cut off at 400 cm^{-1}.

In the 'filter' stage, we exclude all wavenumbers that are not identified as part of a receptor spectra peak. These excluded wavenumbers describe signal-less background regions

that are not chemically relevant and do not change before and after haloanisole exposure. This allows us to utilize only CHO feature groups for the XGBoost classification, where we attain slight improvements in classification accuracy to 67.3 – 72.8% for the haloanisoles and 99.1% for the control (**Figure 3-9C**). This improvement stems from the mitigation of the CoD effect at the single receptor level, where signal-less background regions negatively influence XGBoost classification due to instrument noise signals. More importantly, the XGBoost feature importance analyses highlights feature groups 1, 3, 7, 10 and 14 as the top five CHO peaks that are the most pertinent in distinguishing the haloanisoles and the control. In close agreement with our analyses earlier, these peaks are assigned to C-H bending, C-S stretching, ring breathing, and C=O stretching modes which are polarized due to CHO-haloanisole dipole interactions. Hence, as opposed to analyzing the feature importance of random single wavenumbers, the importance of feature groups is inherently more chemically meaningful and will significantly bolster the interpretability of the XGBoost models and better guide our RRS in recommending receptors. Similar improvements are also observed across all other receptors, where using feature groups alone will improve the XGBoost classification of the haloanisoles as opposed to using the full SERS spectrum (**Figure 3-11A**). It is noteworthy that while some receptors have lesser SERS peaks than others – and therefore will have more wavenumbers excluded as part of background regions – their ability to distinguish haloanisoles is not affected.

Moving further, the 'rank' stage is split into two levels – (1) the 'local' level which refers to the ranking of feature groups within a receptor, and (2) the 'global' level which refers to the ranking of individual receptors across all available receptors. At the local level of the 'rank' stage, we first sort all feature groups in order of their importance scores and then sequentially remove feature groups with the lowest importance until the highest classification accuracy is attained (**Figure 3-11B**). The motivation for doing so is because some receptor SERS peaks either produce insignificant variations after haloanisole exposure or produce

variations to the same extent which does not allow distinction of the haloanisoles. For example, the peak in CHO feature group 12 at 1489 cm^{-1} showed red-shift and peak intensity decrease to the same extent for all five haloanisoles and thus is ranked lowest based on its XGBoost importance score (**Figure 3-6**). Similar to the removal of uninformative background regions, we illustrate a slight improvement in the classification accuracy from 72.8 to 73.9% when the lowest ranked feature groups 5 and 12 are removed. This improvement is also seen across all nine receptors, which indicates that uninformative feature groups will contribute to CoD and should be removed.

(A)

Receptor	Full spectrum		Feature groups	
	Number of wavenumbers	Classification Accuracy	Number of wavenumbers (groups)	Classification Accuracy
NH_2	1401	70.3	307 (12)	71.9% (+1.6)
Br	1401	64.8	179 (8)	66.6% (+1.8)
COOH	1401	62.6	539 (11)	63.4% (+0.8)
CH_3	1401	53.9	295 (11)	54.9% (+1.0)
OH	1401	65.8	402 (7)	66.4% (+0.6)
$B(OH)_2$	1401	58.0	501 (12)	58.6% (+0.6)
PY	1401	62.5	366 (12)	63.7% (+1.2)
CHO	1401	70.9	459 (14)	72.8% (+1.9)
NT	1401	56.6	343 (14)	57.5% (+0.9)

(B)

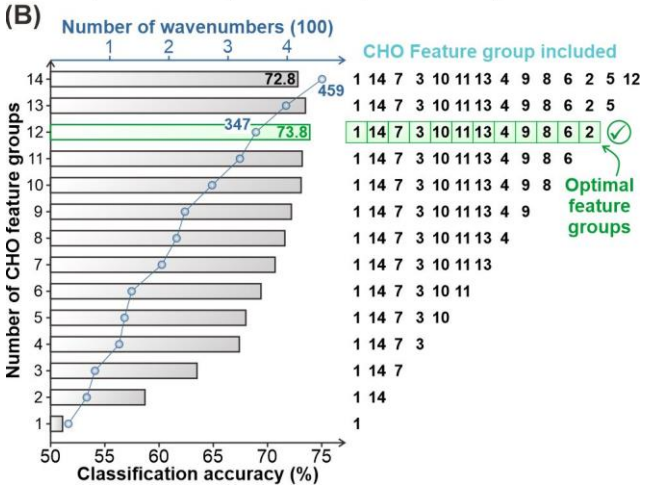
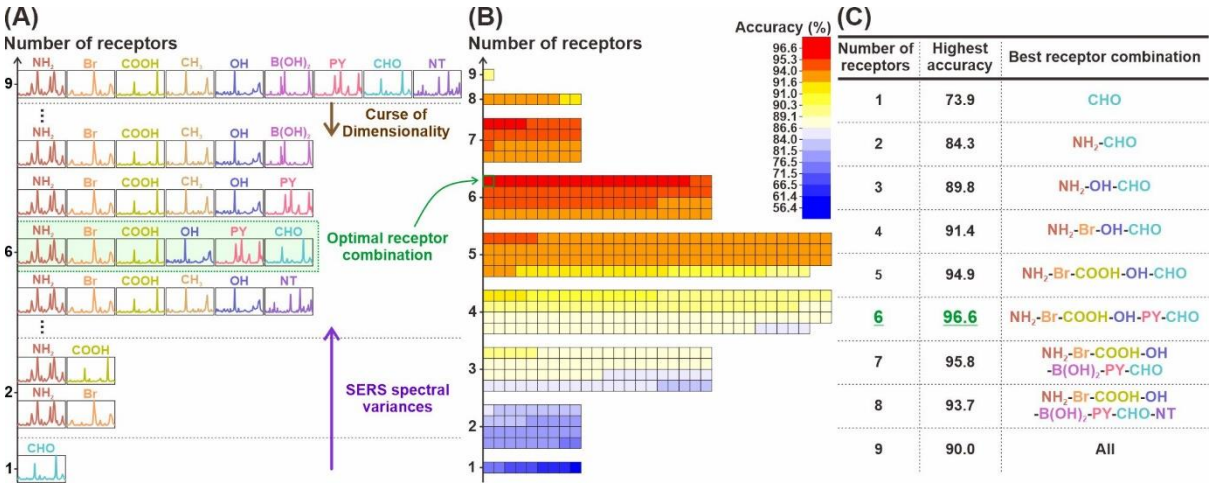


Figure 3-11. (A) Classification accuracy attained using a full SERS spectrum as input as opposed to feature groups as input across all receptors. (B) Optimizing included CHO feature

groups by sequentially excluding the least important feature groups and comparing the classification accuracy attained.

At the global level of the 'rank' stage, we construct SERS 'super-profiles' by serially combining all feature groups originating from different receptor combinations (**Figure 3-12A**). When combining the feature groups from two receptors, we effectively consider SERS variations arising from receptor-haloanisole interactions in both receptors, thereby increasing the total number of variations for enhanced ML analyses. From the chemistry perspective, this is akin to utilizing multiple receptors to either (1) probe different facets of a target analyte and derive a clearer picture of the analyte's structural characteristics or (2) provide compound spectral variations that accentuates the differences between haloanisoles. However, a key bottleneck in such multi-receptor SERS sensing strategies is the onset of CoD effects due to the wavenumber-intensive nature of SERS spectral data. Therefore, to comprehensively investigate the optimal number and combination of receptors, our RRS constructs all 502 combinations of SERS super-profiles involving two to nine receptors and rank their attained classification accuracies (**Figure 3-12B, Figure 3-12C**). When increasing the number of receptors from one to six, we observe a significant increase in the highest attained XGBoost classification accuracy from 73.9% (using a single receptor) to 96.6% (using a six-receptor combination). This is because each additional receptor brings a different set of SERS peak variances arising from non-covalent receptor-haloanisole interactions such as hydrogen bonding, halogen bonding, dipole-dipole interactions, aromatic donor-acceptor interactions as well as van der Waals' interactions. Therefore, the combined SERS super-profile will contain a larger number of variances that distinguish the haloanisoles better than any single receptor. However, when we increase the number of receptors from six to nine, we note a decrease in accuracy from 96.6% to 90.0%. This decrease stems from the onset of CoD as each additional

receptor do not provide sufficient new SERS variances to differentiate the haloanisoles than what is already present within the super-profile. As such, each receptor added beyond the optimal merely obstructs the retrieval of key spectral variances arising from each haloanisole. Finally, in the 'recommend' stage, this optimal six-receptor combination comprising NH_2-Br-COOH-OH-PY-CHO is recommended for the five haloanisole classification.



Number of receptors	Highest accuracy	Best receptor combination
1	73.9	CHO
2	84.3	NH_2-CHO
3	89.8	NH_2-OH-CHO
4	91.4	NH_2-Br-OH-CHO
5	94.9	NH_2-Br-COOH-OH-CHO
6	96.6	NH_2-Br-COOH-OH-PY-CHO
7	95.8	NH_2-Br-COOH-OH -$B(OH)_2$-PY-CHO
8	93.7	NH_2-Br-COOH-OH -$B(OH)_2$-PY-CHO-NT
9	90.0	All

Figure 3-12. Optimizing SERS super-profiles by objective selection of receptors. (A) Formation of an optimal SERS super-profile with maximum SERS variances and incurs minimal CoD effects. (B) XGBoost classification accuracies across all 511 combinations of one to nine-receptor SERS spectra/super-profiles. (C) The highest accuracy attained using a specific number of receptors along with its best receptor combination is stated.

Crucially, we highlight the flexibility of our four-stage approach in the RRS framework with an alternative workflow that begins with the formation of a SERS super-profile comprising all receptors that should be considered (**Figure 3-13**). The key advantage of this workflow is that it does not require the formation of all possible super-profile combinations. This is important because if 20 receptors were considered, the total number of combinations will be 1048575, and thus will require significantly longer processing time.

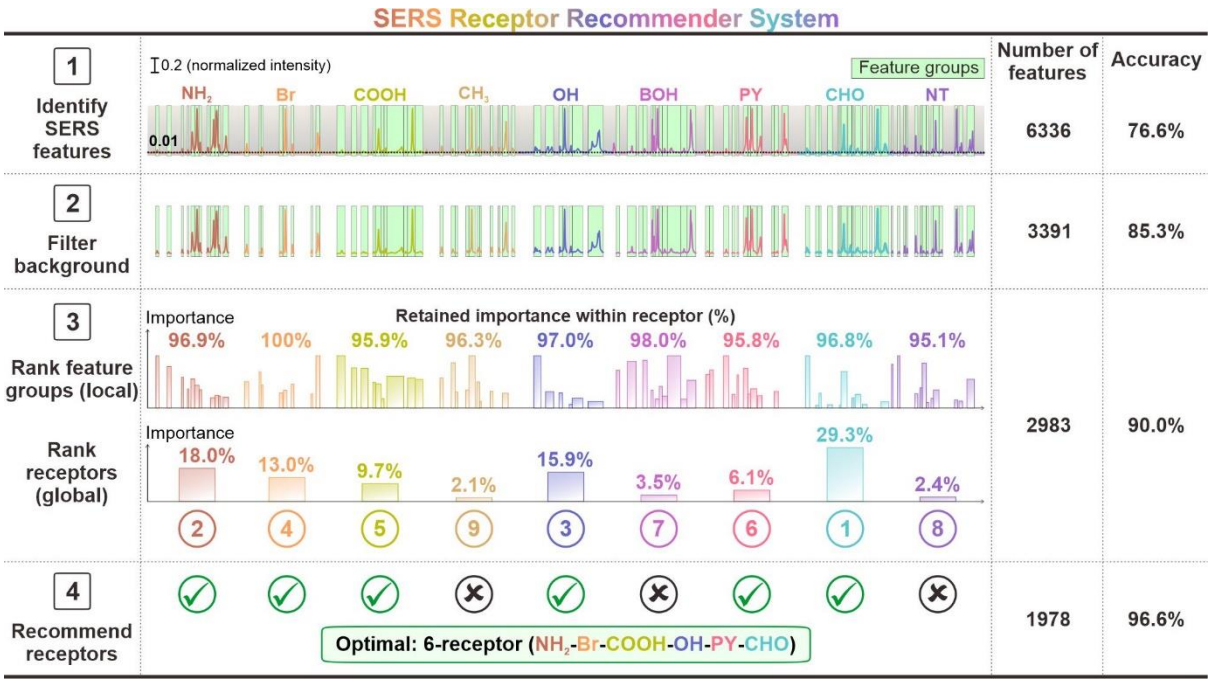


Figure 3-13. Stage-wise results from our four stage 'identify, filter, rank and recommend' approach. At each stage, the number of SERS features and the highest attained XGBoost classification accuracy is stated.

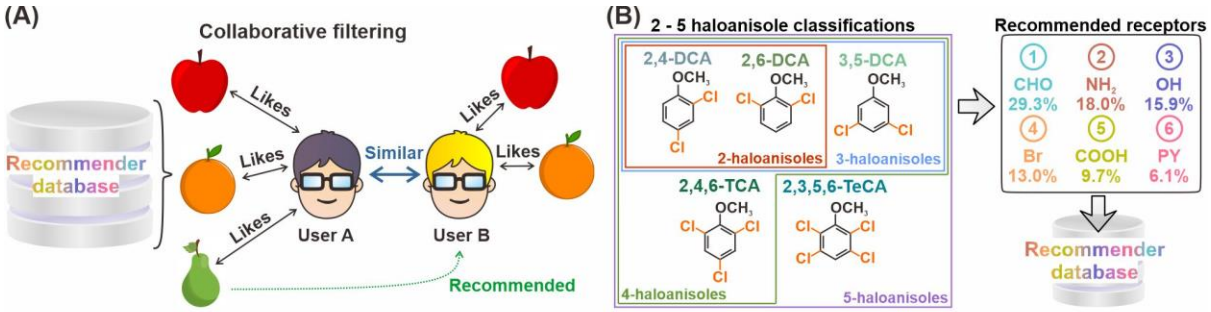
To overcome this, we first combine all receptor SERS spectra into a SERS super-profile. In this context, the nine receptors have a total of 6336 wavenumbers and attain a classification accuracy of 76.6%. In the 'identify' and 'filter' stages, all receptor SERS peaks are identified and grouped into feature groups while the signal-less background regions are excluded in the same fashion. At this point, the classification accuracy has already improved to 85.3%. In the local level of the 'rank' stage, unimportant feature groups within each receptor were removed, which further improves the classification accuracy to 90.0%. In the global level of the 'rank' stage, instead of ranking the classification accuracies attained by different super-profiles, we rank the XGBoost importance of each receptor within the nine-receptor super-profile. From the receptor importance, we then select the top-n receptors to test, where n denotes a specific number of receptors. For example, for a two-receptor combination, we only include the top-2 receptors, hence we test the CHO-NH_2 super-profile. This reduces the total number of receptor

combinations from 511 to 9, which greatly enhances the scalability of our RRS in dealing with large numbers of receptors. In doing so, we demonstrate that we can arrive at the same optimal six-receptor combination of NH_2-Br-COOH-OH-PY-CHO in the 'recommend' stage, which attains the highest accuracy of 96.6%.

3.2.4 Collaborative filtering using a recommender database

With a smart framework to objectively select receptors, we further demonstrate the ability of our RRS to leverage collaborative filtering and directly recommend a set of optimal receptors for a six haloanisole classification with an untrained haloanisole (2,4,6-TBA) even before collecting its experimental data. Collaborative filtering is a common technique used in many RSs to provide informed recommendations by building a database of item preferences based on users in the past (**Figure 3-14A**).[16] In general, when there is a new user, the RS searches the database for 'neighbors' who have historically had similar taste to the new user so that the probability of a recommendation being relevant is higher. Drawing parallel to this approach, we construct a recommender database by recording the global receptor importance across 26 different combinations of two to five haloanisole classifications (**Figure 3-14B, Figure 3-14C**). Here, the 'users' are the 26 classifications, the 'items' are the 9 receptors, and the user-item preference are indicated by the receptor importance scores. Importantly, we observe that with an increasing number of haloanisoles present in the classification, the number of receptors in the best receptor combination also increases, where a two, three, four, and five haloanisole classification would optimally require 2-3, 3-4, 4-5, and 6 receptors respectively. This is because with an increasing number of haloanisoles, the high degree of structural similarity between the haloanisoles will inevitably increase the difficulty in differentiating them, which in turn requires more receptors to contribute crucial SERS spectral variances. In addition, we note that CHO occurs most frequently as part of the best receptor combination,

with a total of 20 times, which concurs with its individual performance with the five haloanisole classification as the best performing receptor. Fundamentally, this stems from its ability to interact favorably with the haloanisoles as exemplified by its highest CHO-haloanisole complex binding energy (**Table 3-2**).



(C)

Receptor importance (%): 0 – 35

Recommender database

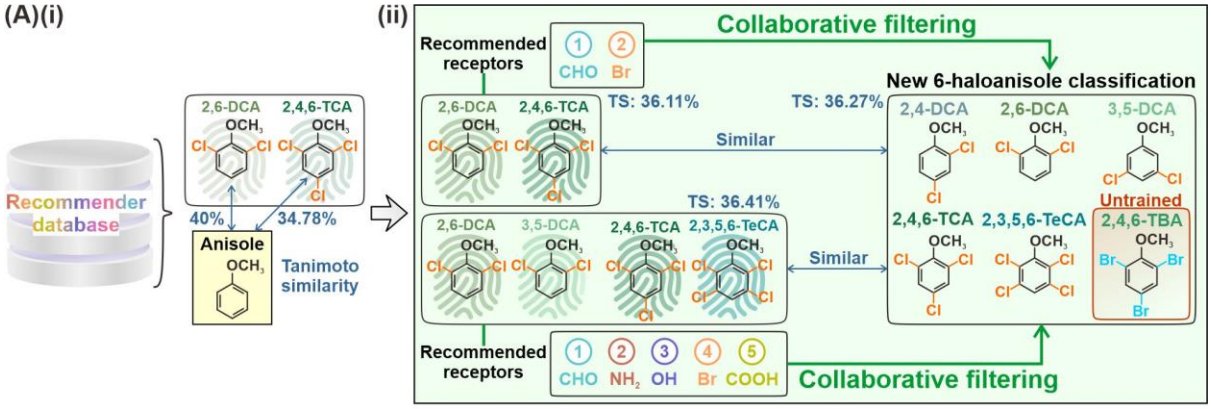
2 Haloanisole	Tanimoto similarity (%)	NH_2	Br	COOH	CH_3	OH	$B(OH)_2$	PY	CHO	NT	Best receptor combination
2,4/2,6	38.72	✓				✓			✓		NH_2-OH-CHO
2,4/3,5	40.64	✓				✓			✓		NH_2-OH-CHO
2,4/2,4,6	36.11		✓	✓							Br-COOH
2,4/2,3,5,6	32.22		✓	✓							Br-COOH
2,6/3,5	41.92	✓				✓			✓		NH_2-OH-CHO
2,6/2,4,6	37.39		✓						✓		Br-CHO
2,6/2,3,5,6	33.50	✓		✓							NH_2-COOH
3,5/2,4,6	39.31	✓				✓					NH_2-OH
3,5/2,3,5,6	35.42					✓			✓		OH-CHO
2,4,6/2,3,5,6	30.89		✓	✓					✓		Br-COOH-CHO
3 Haloanisole											
2,4/2,6/3,5	40.42	✓	✓			✓			✓		NH_2-Br-OH-CHO
2,4/2,6/2,4,6	37.40		✓	✓		✓			✓		Br-COOH-OH-CHO
2,4/2,6/2,3,5,6	34.81		✓	✓					✓		Br-COOH-CHO
2,4/3,5/2,4,6	38.68	✓	✓			✓			✓		NH_2-Br-OH-CHO
2,4/3,5/2,3,5,6	36.09	✓		✓		✓					NH_2-COOH-OH
2,4/2,4,6/2,3,5,6	33.07	✓				✓		✓			NH_2-OH-PY
2,6/3,5/2,4,6	39.54	✓	✓			✓			✓		NH_2-Br-OH-CHO
2,6/3,5/2,3,5,6	36.95	✓	✓	✓					✓		NH_2-Br-COOH-CHO
2,6/2,4,6/2,3,5,6	33.93	✓		✓					✓		NH_2-COOH-CHO
3,5/2,4,6/2,3,5,6	35.21			✓				✓	✓		COOH-PY-CHO
4 Haloanisole											
2,4/2,6/3,5/2,4,6	39.01	✓	✓			✓		✓	✓		NH_2-Br-OH-PY-CHO
2,4/2,6/3,5/2,3,5,6	37.07	✓	✓	✓				✓	✓		NH_2-Br-COOH-PY-CHO
2,4/2,6/2,4,6/2,3,5,6	34.80	✓	✓	✓		✓			✓		NH_2-Br-COOH-OH-CHO
2,4/3,5/2,4,6/2,3,5,6	35.76	✓		✓		✓			✓		NH_2-COOH-OH-CHO
2,6/3,5/2,4,6/2,3,5,6	36.41	✓	✓	✓		✓			✓		NH_2-Br-COOH-OH-CHO
5 Haloanisole											
2,4/2,6/3,5/2,4,6/2,3,5,6	36.61	✓	✓	✓		✓		✓	✓		NH_2-Br-COOH-OH-PY-CHO

Figure 3-14. Constructing a SERS RRS recommender database. (A) Collaborative filtering in RSs, where the RS provides informed recommendations to a new user by studying users in the database with similar tastes. (B) Our recommender database comprises 26 different

classifications (users) involving two to five haloanisoles. For each classification, the global receptor importance (user-item preference) is obtained and recorded. (C) Global receptor importance scores, best receptor combinations, and their averaged pairwise TS for all 26 classifications.

The goal of our RRS is to elucidate the receptor importance scores (user-item preference) of the new six haloanisole classification (new user) from the recommender database and leverage them to recommend a set of receptors (items) that is optimal (**Figure 3-15Aii**). To do so, our RRS must (1) search within the database for classifications that are similar to the new six haloanisole classification (users that have similar taste) and (2) determine how many classifications should be considered when predicting the receptor importance scores (how many 'neighbors'). First, we utilize the averaged pairwise Tanimoto similarity (TS) as a metric to compare two classifications (**Figure 3-15Ai**).[17-18] The TS is a well-established measure of the relative structural similarity between different molecules ranging from 0 (completely dissimilar) to 1 (the same molecule) and can be calculated using the simplified molecular input line entry system (SMILES) notation of each haloanisole.[19-20] A pairwise TS is calculated by comparing one haloanisole (such as 2,4-DCA) with an anisole as the origin since all haloanisoles possess the same anisole backbone (**Table 3-3**). For a two haloanisole classification (such as 2,4-DCA and 2,6-DCA), an averaged pairwise TS is tabulated between 2,4-DCA and 2,6-DCA (**Table 3-4**). Hence, the averaged pairwise TS allows us to compare the relative similarity between a two haloanisole classification and a three haloanisole classification by measuring the degree of similarity among the haloanisoles present in each classification. In doing so, we hypothesize that if a new classification comprise a set of haloanisoles that is comparable to other classifications within the database, the optimal receptors for both classifications will likely be similar. The 26 classifications in our

recommender database have averaged pairwise TS values ranging from 30.89 to 41.92%, while the six haloanisole has a value of 36.27% which falls within this range (**Figure 3-14B, Figure 3-15Aii**). This is because 2,4,6-TBA has a pairwise TS of 34.59%, which is lower than all other haloanisoles but higher than 2,3,5,6-TeCA (27%) indicating that it is structurally more similar to an anisole than 2,3,5,6-TeCA (**Table 3-3**). Subsequently, our RRS sorts the classifications within our recommender database starting from the smallest to the largest difference in their averaged pairwise TS values as compared to the six haloanisole classification (**Table 3-4**).



(B)

(i) Collaborative filtering recommendation

Accuracy	Receptor importance (%)								
	NH_2	Br	COOH	CH_3	OH	$B(OH)_2$	PY	CHO	NT
95.7	18.7	11.2	8.4	0.7	22.2	5.1	7.2	22.3	4.2
RMSE = 3.2%	✓	✓	✓	✗	✓	✓	✓	✓	✗

(ii) Four-stage recommendation

95.7	14.2	15.4	5.9	3.0	19.6	8.9	5.1	25.6	2.3
	✓	✓	✓	✗	✓	✓	✓	✓	✗

Figure 3-15. Collaborative filtering using a recommender database. (A)(i) Calculating the pairwise TS. (ii) Based on the averaged pairwise TS, similar classifications are selected from the recommender database using a kNN approach. The global receptor importance scores from these classifications are arithmetically averaged to predict receptor importance for the new six haloanisole classification. (B) Receptor importance scores for the six haloanisole classification (i) predicted by collaborative filtering and (ii) calculated by our four-stage approach.

Table 3-3. Pairwise TS compared with an anisole.

No.	Haloanisole	Pairwise Tanimoto similarity (%)
1	2,4-DCA	37.43
2	2,6-DCA	40.00
3	3,5-DCA	43.84
4	2,4,6-TCA	34.78
5	2,3,5,6-TeCA	27.00
6	**2,4,6-TBA (untrained)**	**34.59**

Table 3-4. TS across all 26 classifications involving 2 – 5 haloanisoles. 'U' denotes the classification involving an untrained haloanisole (2,4,6-TBA).

No.	Haloanisoles	TS (%)
U	**2,4-DCA/2,6-DCA/3,5-DCA/2,4,6-TCA/2,3,5,6-TeCA/2,4,6-TBA**	**36.27**
1	2,6-DCA/3,5-DCA/2,4,6-TCA/2,3,5,6-TeCA	36.41
2	2,4-DCA/2,4,6-TCA	36.09
3	2,4-DCA/3,5-DCA/2,3,5,6-TeCA	36.61
4	2,4-DCA/2,6-DCA/3,5-DCA/2,4,6-TCA/2,3,5,6-TeCA	35.76
5	2,4-DCA/3,5-DCA/2,4,6-TCA/2,3,5,6-TeCA	36.95
6	2,6-DCA/3,5-DCA/2,3,5,6-TeCA	37.07
7	2,4-DCA/2,6-DCA/3,5-DCA/2,3,5,6-TeCA	35.42
8	3,5-DCA/2,3,5,6-TeCA	35.21
9	3,5-DCA/2,4,6-TCA/2,3,5,6-TeCA	37.39
10	2,6-DCA/2,4,6-TCA	37.40
11	2,4-DCA/2,6-DCA/2,4,6-TCA	34.81
12	2,4-DCA/2,6-DCA/2,3,5,6-TeCA	34.80
13	2,4-DCA/2,6-DCA/2,4,6-TCA/2,3,5,6-TeCA	33.93
14	2,6-DCA/2,4,6-TCA/2,3,5,6-TeCA	38.68
15	2,4-DCA/3,5-DCA/2,4,6-TCA	38.72
16	2,4-DCA/2,6-DCA	39.01
17	2,4-DCA/2,6-DCA/3,5-DCA/2,4,6-TCA	33.50

18	2,6-DCA/2,3,5,6-TeCA	39.31
19	3,5-DCA/2,4,6-TCA	33.07
20	2,4-DCA/2,4,6-TCA/2,3,5,6-TeCA	39.54
21	2,6-DCA/3,5-DCA/2,4,6-TCA	32.22
22	2,4-DCA/2,3,5,6-TeCA	40.42
23	2,4-DCA/2,6-DCA/3,5-DCA	40.64
24	2,4-DCA/3,5-DCA	30.89
25	2,4,6-TCA/2,3,5,6-TeCA	41.92
26	2,6-DCA/3,5-DCA	36.95

Finally, we design a kNN collaborative filtering algorithm for our RRS to elucidate the classifications within our recommender database that should be included when predicting receptor importance scores for the six haloanisole classification and thereby recommend the optimal receptors. Briefly, the algorithm takes the arithmetic average of the receptor importance scores from k neighboring classifications incrementally, starting from the one with the closest averaged pairwise TS value.[21] We obtain k as 15 by retrospectively predicting the receptor importance values for the five haloanisole classification using our collaborative filtering approach and comparing our results with the actual scores obtained with our four-stage approach earlier (**Table 3-5**). This is because the root mean-squared error (RMSE) of predicted importance scores is lowest at k = 15. Therefore, our RRS predicts the receptor importance scores for the six haloanisoles classification by taking the arithmetic average of the receptor importance scores from the top 15 classifications.

Using these scores, our RRS recommends a seven-receptor combination (NH_2-Br-COOH-OH-BOH-PY-CHO) where each receptor included has an importance contribution above 5% (**Figure 3-15Bi**). Our RRS recommends a larger number of receptors (7) for a six haloanisole classification as compared to a five haloanisole classification (6) because it learns

from the positive relationship between the number of haloanisoles in question and number of receptors required as shown by past classifications.

Table 3-5. Determining the optimal k value for our kNN-driven collaborative filtering by averaging the receptor importance scores from the top k classifications and comparing the result to our four-stage approach for the five haloanisole classification. The RMSE between the actual and predicted importance scores are used to guide the selection of k.

k	RMSE
1	3.26
2	5.41
3	5.88
4	4.06
5	4.96
6	4.99
7	5.03
8	3.54
9	5.12
10	4.98
11	4.54
12	5.05
13	3.60
14	4.83
15	**3.14**
16	4.93
17	5.03
18	6.71
19	6.12
20	6.82
21	4.91

22	6.22
23	4.35
24	3.62
25	4.68

We further collect receptor-2,4,6-TBA experimental SERS spectra and perform the four-stage approach as a benchmark to assess the quality of our collaborative filtering recommendation. Crucially, we show that this seven-receptor combination is also the receptor combination recommended by the four-stage approach, attaining the highest classification accuracy of 95.7% (**Figure 3-15Bii**). Additionally, based on the actual receptor importance scores with the four-stage approach, we illustrate the high accuracy in predicting receptor importance scores using our collaborative filtering approach with a low RMSE of 3.2%. The high relevance stems from the predictability of receptor-haloanisole interactions with similar groups of haloanisoles, which are firmly guided by chemical principles unlike complicated human behavior and preferences. In contrast to conventional methods for receptor selection based on chemical intuition or trial-and-error experimentation, our RRS adopts a data-driven approach that would remove the need for subjective decisions and inconsistencies in random testing. Since the RRS recommends based on past information recorded in the recommender database, we can obtain an optimal receptor combination even before collecting experimental data for the new classification as long as we can optimize k. More importantly, our proof-of-concept application here demonstrates both the immense potential of RSs in elucidating optimal receptors for SERS-based sensing applications as well as the ability to harness the full potential of multi-receptor SERS super-profiles without worrying about potential CoD influences.

3.3 Conclusion

In conclusion, we introduced a SERS RRS with a strategic four-stage 'identify, filter, rank and recommend' approach to objectively select molecular receptors to form SERS super-profiles with maximum variance and minimal CoD effects. Our in-depth SERS spectral analysis first affirmed the basis of receptor-driven SERS sensing, whereby haloanisole-specific receptor peak variations can be detected even with structurally similar compounds due to differences in their receptor-haloanisole interaction strengths. In the 'identify' stage, our RRS forms feature groups that comprise receptor SERS peaks to allow direct correlation between XGBoost model performance and specific receptor vibrational modes, which imbues enhanced model interpretability. In the 'filter' stage, our RRS excludes background regions that are signal-less and do not vary before and after haloanisole exposure, significantly reducing the total number of included SERS wavenumbers and mitigating the CoD effect. In the 'rank' stage, the local level further optimizes the included receptor feature groups while the global level comprehensively derives the receptor combination of NH_2-Br-COOH-OH-PY-CHO which attains the highest classification accuracy of 96.6% and is recommended in the 'recommend' stage. When a large number of receptors is concerned, our RRS adopts an alternative approach where a SERS super-profile comprising all receptors is constructed so that their contributions can be ranked at the global level of the 'rank' stage. This allows efficient derivation of the optimal receptor combination without the need to construct all super-profile combinations. For a new classification involving untrained 2,4,6-TBA, our RRS leverages the averaged pairwise TS as a metric to retrieve similar classifications recorded within a recommender database. Subsequently, our RRS employs a kNN-driven collaborative filtering approach by averaging the receptor importance scores of 15 similar classifications and recommends the optimal receptor combination without having to collect experimental data involving 2,4,6-TBA. Our RRS resolves current limitations in manual trial-and-error methods in receptor optimization for

multi-receptor SERS platforms, providing a potential paradigm shift for such array-based sensing strategies to achieve precise molecular differentiation with generic receptors and non-specific interactions.

3.4 Materials and methods

Chemicals. Silver nitrate ($AgNO_3$, ≥ 99%), 1,5-pentanediol (PD, ≥ 97%), poly(vinylpyrrolidone) (PVP, M_w ~55000), 4-aminothiophenol (NH_2, ≥ 97.0%), 4-bromothiophenol (Br, 95%), 4-mercaptobenzoic acid (COOH, 99%), 4-methylbenzenethiol (CH_3, 98%), 4-mercaptophenol (OH, 97%), 4-mercaptophenylboronic acid ($B(OH)_2$, 90%), 4-mercaptopyridine (PY, 95%), 4-(methylthio)benzaldehyde (CHO, 95%), 2-naphthalenethiol (NT, 99%), 2,4-dichloroanisole (2,4-DCA, 97%), 2,6-dichloroanisole (2,6-DCA, 97%), 3,5-dichloroanisole (3,5-DCA, 98%), 2,4,6-trichloroanisole (2,4,6-TCA, 99%), 2,4,6-tribromoanisole (2,4,6-TBA, 99%), 2,3,5,6-tetrachloroanisole (2,3,5,6-TeCA, ≥ 98%) were purchased from Sigma Aldrich. Copper (II) chloride was purchased from Alfa Aesar. Ethanol (EtOH, ACS, ISO, Reag. Ph Eur) was purchased from Merck. 2-propanol (IPA, ≥ 99.7%) was purchased from J. T. Baker., Avantor® inc. Milli-Q water (18.2 MΩ·cm) was purified using a Sartorius Arium® 611 UV ultrapure water system. All reagents were used without further purifications.

Synthesis and purification of Ag nanocubes. The polyol reduction method was used to synthesize Ag nanocubes in high yield.[22] In brief, 20 mL of PD was added to a 100 mL round-bottom flask and heated at 190 °C for 10 min. 250 µL of PVP and 500 µL of $AgNO_3$ precursor solution were added in alternation to the reaction mixture until the observation of reddish-brown coloration. The reaction mixture was left to cool to room temperature before several wash cycles involving the addition of ethanol and subsequent centrifugation, discarding the

supernatant. The resultant mixture was subjected to vacuum filtration with polyvinylidene fluoride filter membranes (Durapore®) with pore sizes 5 µm, 0.65 µm, 0.45 µm and 0.22 µm to remove impurities.

Receptor functionalization and sensor preparation. The purified nanocubes were individually functionalized with 10 mM of 9 different molecular receptors. Respective amounts of receptor solutions (ATP: 10 µL, BTP: 160 µL, MBA: 310 µL, MBT: 60 µL, MP: 240 µL, MPBA: 200 µL, MPY: 20 µL, MTBH: 240 µL, NT: 80 µL) were added dropwise to 2 mL of filtered Ag nanocube solution in 1:1 EtOH/IPA and stirred for 3 h. The resultant mixture was washed with 1:1 EtOH/IPA solution for several times before re-dispersing in 0.5 mL of 4:1 H_2O/EtOH. 1.5 µL of receptor functionalized Ag nanocubes were drop cast onto a Si wafer (1 µm × 1 µm in size) and dried under ambient conditions prior to use.

Characterizations. Scanning Electron Microscopy (SEM) images were acquired using the JEOL-JSM-7600F microscope at an accelerating voltage of 5 kV. Each haloansiole was separately exposed to individual receptor-functionalized SERS sensor to generate 30 SERS spectra (9 × 5 × 30 = 1350), along with 20 blank receptor SERS spectra (9 × 1 × 20 = 180), totaling 1530 spectra. Similarly, 30 SERS spectra were collected from each receptor for the 'unknown' haloanisole (246TBA), totaling 270 spectra (9 × 30 = 270). SERS measurements were performed using the x-y imaging mode of the Ramantouch microspectrometer (Nanophoton Inc., Osaka, Japan) with a 532 nm excitation laser (power = 0.4 mW). A 20× objective lens (N.A. = 0.4) was used with 30 s acquisition time for data collection. All SERS spectra were averaged across 100 individual SERS spectrum within the SERS map.

DFT simulations. DFT simulations were carried out using the Gaussian 16 computational chemistry package with the unrestricted B3LYP exchange-correlation functional. The LANL2DZ basis set was used for Ag while the 6-31G(d, p) basis set was used for all other atoms. The Ag surface was modelled using a triangular Ag cluster based on previous reports and was geometrically optimized before placing each receptor molecule (Br, COOH, CH_3, OH, $B(OH)_2$, PY, CHO, NT) at the vertex. For NH_2, the simulation was done using the 4,4-dimercaptoazobenzene (DMAB) dimer attached to two Ag clusters at the vertex. The whole system was then relaxed with all Ag atoms fixed. Finally, the individual haloanisole molecule was placed near the receptor before allowing the whole system to relax with all Ag atoms fixed.

Spectrum pre-processing and formation of SERS super-profiles. The acquired SERS spectra were broadly trimmed using a set of Python processing codes to retain only the 400 to 1800 cm^{-1} region where most of the Raman fingerprint peaks are present. The spectrum was then baselined using the adaptive iteratively reweighted penalized least squares (airPLS) algorithm and normalized by setting the maximum peak intensity to 1.[23] To form a SERS super-profile, two receptor spectra were concatenated along their Raman shift (cm^{-1}) axis.

Identify peaks, group features, and filter background. An arbitrary threshold of 0.01 normalized intensity was set to guide the automated identification of SERS peaks. Once a peak maximum is identified, the peak is selected by including all SERS features based on their Raman shift (cm^{-1}) until a second arbitrary threshold of 0.001 normalized intensity. The selected region is grouped as a single feature group and was allowed to comprise multiple blended peaks. Raman vibrational modes were assigned to each peak within the feature groups by corroborating with the respective DFT-simulated receptor SERS spectrum. All unselected SERS features were filtered as background signals.

Classification model. The open source XGBoost algorithm was used to construct the classification models used.[15] Hyperparameter optimizations were carried out using Optuna for each constructed XGBoost model over 10 trials to obtain an optimal set of model parameters that gives the lowest model logarithmic loss.[24] The optimized parameters include (1) the number of gradient boosted trees, (2) the maximum tree depth, (3) the booster's learning rate, (4) the type of booster used, (5) the minimum loss required to make further partition of a leaf node, (6) the subsample ratio of the training instance and (7) the subsample ratio of columns when constructing each tree. Subsequently, unique optimized models were constructed over 50 random states and their accuracies were averaged to give the final classification accuracy.

Grouped feature importance and relative receptor importance. Feature importance scores were obtained using the built-in XGBoost feature importance calculation method, using the gain type, which measures the improvement in accuracy brought by a feature to the branches it is on. Within each receptor, grouped feature importance scores were obtained by summing all contributions by discrete wavenumbers in the defined SERS peak regions. The relative receptor importance scores were then obtained by summing all grouped feature importance scores associated with the receptor in a nine-receptor SERS super-profile. Receptors were ranked in order of importance (from lowest to highest) and sequentially included until the peak classification accuracy is attained. The combination that provides the maximum classification accuracy recommended as the optimal receptor combination pertaining to a specific classification.

Collaborative filtering using a recommender database. A recommender database is constructed using all 26 combinations of 2 – 5 haloanisole classifications (10+10+5+1). For each haloanisole, we calculate their pairwise TS with an anisole as the origin since all

haloanisoles possess the same anisole backbone. To do so, we first derive the extended-connectivity fingerprints (ECFP) of individual haloanisoles using the open-source RDKit package with Morgan algorithm refinement.[25-26] Then, we utilize the SMILES notation to represent their respective molecular structures. The molecules are then compared to output a pairwise TS value. For each classification, the averaged pairwise TS is calculated by averaging all pairwise TS values of the haloanisoles within the classification. We subsequently rank the 26 classifications according to their relative similarity with the new six-haloanisole classification based on their respective averaged pairwise TS values. Then, a kNN approach was used to predict the optimal receptors.[21] The recommendation relevance was evaluated by calculating (1) the RMSE of the predicted importance scores and (2) the classification accuracy using the predicted set of receptors.

References

1. Lee, H. K.; Lee, Y. H.; Koh, C. S. L.; Phan-Quang, G. C.; Han, X.; Lay, C. L.; Sim, H. Y. F.; Kao, Y.-C.; An, Q.; Ling, X. Y., *Chem. Soc. Rev.*, 2019, 48, 731-756.
2. Leong, Y. X.; Tan, E. X.; Leong, S. X.; Koh, C. S. L.; Nguyen, L. B. T.; Chen, J. R. T.; Xia, K.; Ling, X. Y., *ACS Nano*, 2022, 16, 13279-13293.
3. Leong, S. X.; Leong, Y. X.; Koh, C. S. L.; Tan, E. X.; Nguyen, L. B. T.; Chen, J. R. T.; Chong, C.; Pang, D. W. C.; Sim, H. Y. F.; Liang, X.; Tan, N. S.; Ling, X. Y., *Chem. Sci.*, 2022, 13, 11009-11029.
4. Leong, Y. X.; Lee, Y. H.; Koh, C. S. L.; Phan-Quang, G. C.; Han, X.; Phang, I. Y.; Ling, X. Y., *Nano Lett.*, 2021, 21, 2642-2649.
5. Nguyen, L. B. T.; Leong, Y. X.; Koh, C. S. L.; Leong, S. X.; Boong, S. K.; Sim, H. Y. F.; Phan-Quang, G. C.; Phang, I. Y.; Ling, X. Y., *Angew. Chem. Int. Ed.*, 2022, 61, e202207447.
6. Leong, S. X.; Leong, Y. X.; Tan, E. X.; Sim, H. Y. F.; Koh, C. S. L.; Lee, Y. H.; Chong, C.; Ng, L. S.; Chen, J. R. T.; Pang, D. W. C.; Nguyen, L. B. T.; Boong, S. K.; Han, X.; Kao, Y.-C.; Chua, Y. H.; Phan-Quang, G. C.; Phang, I. Y.; Lee, H. K.; Mohammad, Y. A.; Tan, N. S.; Ling, X. Y., *ACS Nano*, 2022, 16, 2629-2639.
7. Altman, N.; Krzywinski, M., *Nat. Methods*, 2018, 15, 399-400.
8. Berisha, V.; Krantsevich, C.; Hahn, P. R.; Hahn, S.; Dasarathy, G.; Turaga, P.; Liss, J., *npj Digit. Med.*, 2021, 4, 153.
9. Campillo, N.; Viñas, P.; Cacho, J. I.; Peñalver, R.; Hernández-Córdoba, M., *J. Chromatogr. A*, 2010, 1217, 7323-7330.
10. Singh, D. K.; Srivastava, S. K.; Schlücker, S.; Singh, R. K.; Asthana, B. P., *J. Raman Spectrosc.*, 2011, 42, 851-858.

11. Chan, W. S.; Ma, C.; Kwok, W. M.; Phillips, D. L., *J. Phys. Chem. A*, 2005, 109, 3454-3469.
12. Fang, Y.; Li, Y.; Xu, H.; Sun, M., *Langmuir*, 2010, 26, 7737-7746.
13. Martinez, C. R.; Iverson, B. L., *Chem. Sci.*, 2012, 3, 2191-2201.
14. Cavallo, G.; Metrangolo, P.; Milani, R.; Pilati, T.; Priimagi, A.; Resnati, G.; Terraneo, G., *Chem. Rev.*, 2016, 116, 2478-2601.
15. Chen, T.; Guestrin, C., In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, 785-794.
16. Sarwar, B.; Karypis, G.; Konstan, J.; Riedl, J., *Proceedings of the 10th international conference on World Wide Web*, 2001, 285-295.
17. Wei, J.; He, J.; Chen, K.; Zhou, Y.; Tang, Z., *Expert Syst. Appl.*, 2017, 69, 29-39.
18. Ahn, H. J., *Inf. Sci.*, 2008, 178, 37-51.
19. Kuwahara, H.; Gao, X., *J. Cheminformatics*, 2021, 13, 27.
20. Bajusz, D.; Rácz, A.; Héberger, K., *J. Cheminformatics*, 2015, 7, 20.
21. Bobadilla, J.; Ortega, F.; Hernando, A.; Bernal, J., *Knowl.-Based Syst.*, 2012, 26, 225-238.
22. Tao, A.; Sinsermsuksakul, P.; Yang, P., *Angew. Chem. Int. Ed.*, 2006, 45, 4597-4601.
23. Zhang, Z.-M.; Chen, S.; Liang, Y.-Z., *Analyst*, 2010, 135, 1138-1146.
24. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M., In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019, 2623-2631.
25. Rogers, D.; Hahn, M., *J. Chem. Inf. Model.*, 2010, 50, 742-754.
26. Morgan, H. L., *J. Chem. Doc.*, 1965, 5, 107-113.

Chapter 4 Noninvasive and Point-of-Care SERS-based Breathalyzer for Mass Screening of COVID-19 under 5 min

Abstract. Mass screening for COVID-19 requires tests that are simple to administer yet able to provide accurate results on-site. Current analytical tools require costly and bulky instruments, limiting their deployment. Here, we design a hand-held SERS-based breathalyzer that detects fluctuations in breath volatile organic compounds to identify COVID-19 infected individuals in under five minutes. Our multi-receptor SERS sensor achieved > 95% sensitivity and specificity in a clinical trial comprising 501 participants, regardless of their displayed symptoms. Our robust machine learning model identifies key variations in SERS fingerprints arising from receptor-analyte interactions with different breath profile compositions for high throughput classification. The observed experimental SERS variations can be attributed to many reported potential COVID-19 breath biomarkers and are well-supported by theoretical evidence. Our design strives to spur the development of next-generation, noninvasive exhaled breath diagnostic toolkits tailored for mass screening purposes.

4.1 Introduction

Developing mass screening tools that are simple to administer yet can provide accurate results on-site to identify infectious, yet asymptomatic individuals is one of the key strategies to control the spread of COVID-19. These screening tools aim to complement diagnostic polymerase chain reaction (PCR) tests by filtering out most healthy individuals from the general population, thereby avoid potentially overloading the test facilities. Exhaled human breath comprises a myriad of volatile organic compounds (VOC) such as aldehydes, ketones, and alcohols that fluctuate in concentration upon COVID-19 infection due to coronavirus-induced immune responses and metabolic changes.[1-4] This presents an attractive noninvasive and simple method to rapidly distinguish infected and healthy individuals regardless of their displayed symptoms. Currently, gas chromatography coupled mass spectrometry (GC-MS) is the gold standard to separate and identify volatile compounds in exhaled breath.[5-7] However, these instruments are costly and bulky, making them inflexible to upscale and integrate as a mass screening tool for on-site deployment. In addition, breath sample collection and analyses cannot be done in parallel because of the need to exhale directly into the instrument. As such, there is an urgent need to develop a simple, portable, and inexpensive mass screening tool that can analyze COVID-19 related breath VOCs (BVOC).

Herein, we design a SERS-based breathalyzer to distinguish BVOC profiles of COVID-positive individuals, achieving >95% sensitivity and specificity across 501 participants from clinical case-control studies conducted in Singapore when benchmarked using PCR tests (**Figure 4-1**). Our breathalyzer comprises a SERS sensor that is enclosed within a custom hand-held, single-use breath chamber to facilitate the safe collection of breath samples. Breath samples are collected by exhaling continuously into the breathalyzer for 10 s and measured on-site using a portable Raman spectrometer, providing test results within 5 min through our ML predictive model (**Figure 4-2**). When exposed to exhaled breath, multiple molecular receptors

with diverse chemical functionalities on interact via ion-dipole interactions or hydrogen bonding with the BVOCs present. These interactions induce specific spectral variations which can be concatenated as a SERS 'super-profile' to accentuate minute differences in BVOC compositions between COVID-positive and COVID-negative individuals. We affirm through experimental and *in silico* simulations using pure VOCs as benchmark that these spectral variations arise due to fluctuations in specific VOCs which have been reported as potential COVID-19 biomarkers.

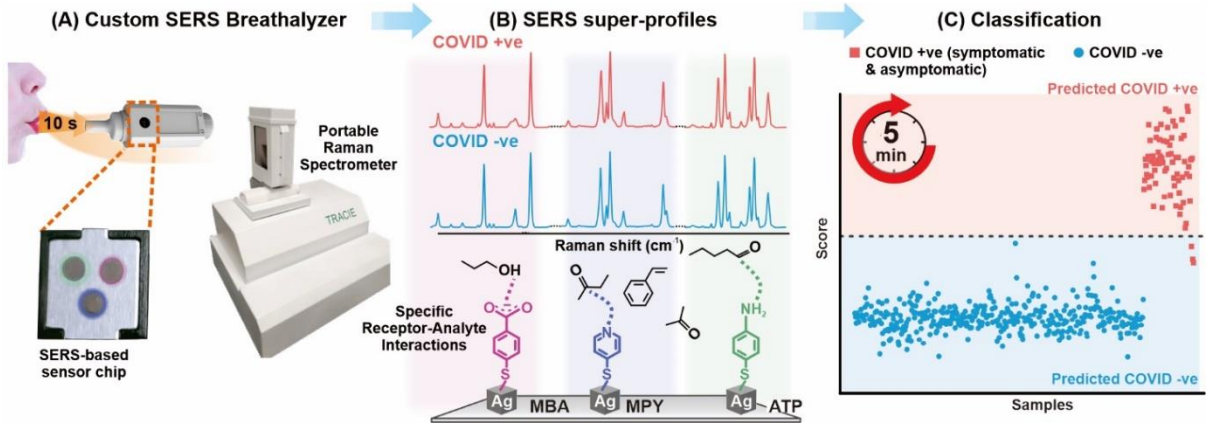


Figure 4-1. Our SERS-based sensor is enclosed within a custom hand-held, single-use breath chamber that allows safe collection of exhaled breath. Using a portable Raman spectrometer, the sample is measured on-site, providing screening results in under 5 min with our machine learning predictive model.

Crucially, we highlight that the observed spectral differences are not influenced by the displayed symptoms and other confounding factors such as the age, gender, smoking habits, and time since last consumed meal of the participants. Our partial least-squares discriminant analysis (PLSDA) machine learning model can be seamlessly integrated with most Raman measurement software to provide results directly after sample measurement. Additionally, our workflow strategically separates sample collection and measurement to provide high flexibility

when deployed in practice. Overall, our study demonstrates the potential of SERS-based sensors in analyzing and differentiating breath metabolites for rapid and noninvasive disease detection. It is a decisive step towards the practical translation of SERS-based sensors for next-generation point-of-care diagnostic toolkits, particularly for respiratory diseases beyond COVID-19.

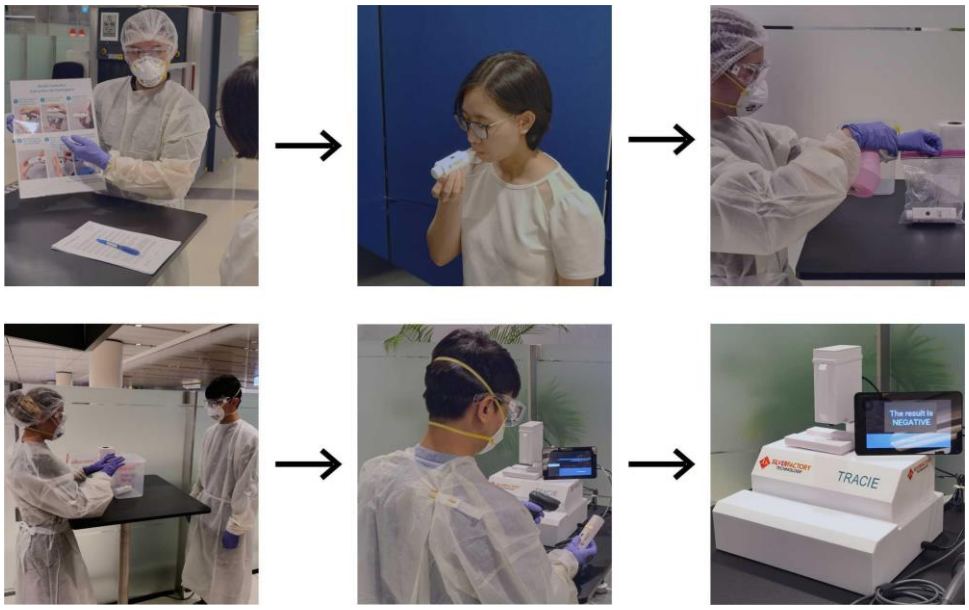


Figure 4-2. The recruitment workflow involves four main segments – (1) preliminary briefing by a clinical trial staff, (2) participant exhaling into the breathalyzer, (3) disinfecting the breathalyzer, (4) measuring SERS signals and providing results on the spot.

4.2 Results and discussion

4.2.1 Sensor fabrication and characterization

To distinguish between COVID-positive and COVID-negative exhaled breath profiles, we designed a SERS sensor with multiple molecular receptors that induces intermolecular interactions with the BVOCs present. Our SERS sensor comprises an array of Ag nanocubes (edge length = 120 ± 5 nm, **Figure 4-3**) with high AEF of 1.4×10^{10}, calculated using R6G as the reference compound (**Figure 4-4**). The calculation is as follows:

I_{SERS} = 1269 ± 44 counts

I_{Raman} = 1784 ± 19 counts

$$AEF = \frac{I_{SERS}}{I_{Raman}} \times \frac{C_{Raman}}{C_{SERS}}$$

$$= \frac{1269}{17.84} \times \frac{2 \times 10^{-2}}{10^{10}}$$

$$= 1.4 \times 10^{10}$$

where C_{SERS} and C_{Raman} are the concentrations of Rhodamine 6G measured using our SERS sensor (10^{-10} M) and normal Raman (2×10^{-2} M) respectively, while I_{SERS} and I_{Raman} are the signal intensities recorded using SERS and normal Raman at their respective concentrations per unit time.

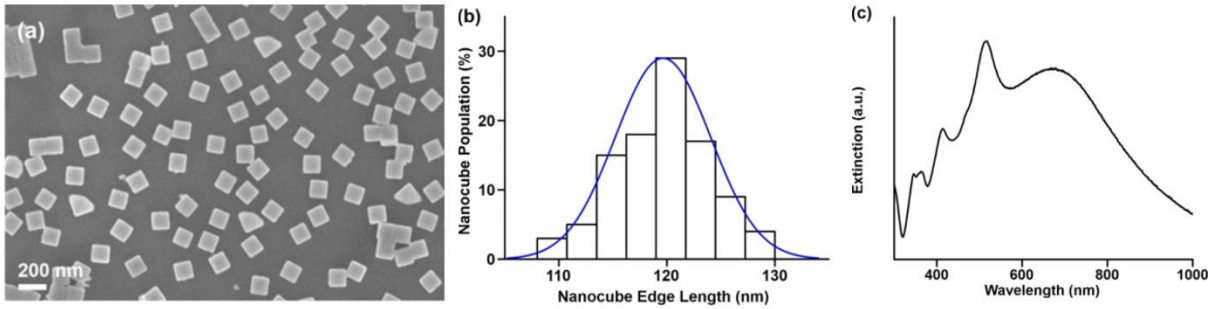


Figure 4-3. (A) SEM image of the Ag nanocubes. (B) Size distribution of the Ag nanocubes, with an average edge length of 120 ± 5 nm. (C) Extinction spectrum of the Ag nanocubes, exhibiting clear plasmonic resonances.

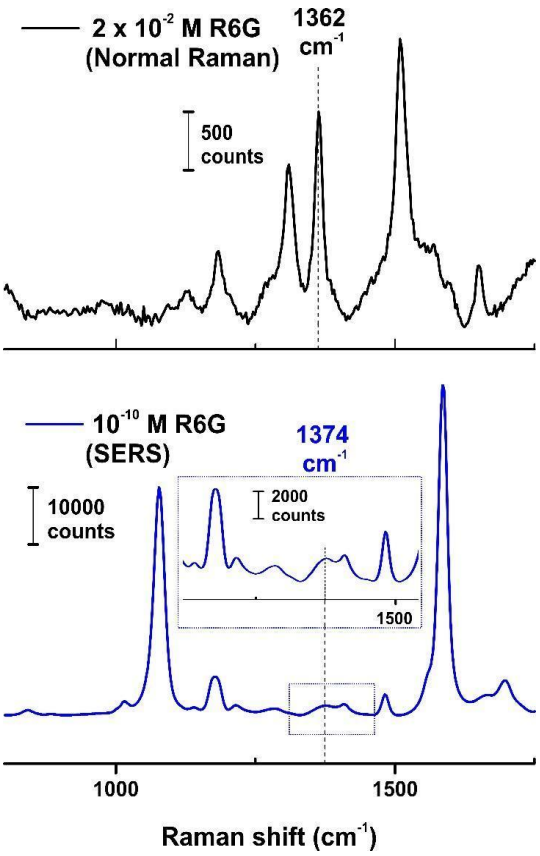


Figure 4-4. Normal Raman spectrum (2×10^{-2} M, black) on a clean aluminum surface and SERS spectrum (10^{-10} M, blue) of R6G on our SERS sensor.

The strong electromagnetic enhancement arises from the sharp Ag nanocube edges and intense inter-nanocube plasmonic coupling, which enables ultrasensitive analyte detection.[8] The Ag nanocubes were separately functionalized with 4-mercaptobenzoic acid (MBA), 4-mercaptopyridine (MPY), and 4-aminothiophenol (ATP), which possess the carboxylic acid, pyridine, and hydroxyl functional groups respectively. These functional groups allow the formation of hydrogen bonding, ion–dipole interactions and π–π interactions which confines the BVOCs close to the plasmonic surface for effective SERS enhancement.[9-10] In addition, our SERS sensor displays < 4% signal standard deviation across at least 50 substrates per receptor, which indicates excellent sensor reproducibility (**Figure 4-5**). Our SERS sensors were subsequently placed within a single-use breath chamber and sealed under vacuum for safety

and hygiene. Importantly, the Ag nanocubes remain stable after assembly and breath exposure, with no significant signs of oxidation after 5 days (**Figure 4-6**).

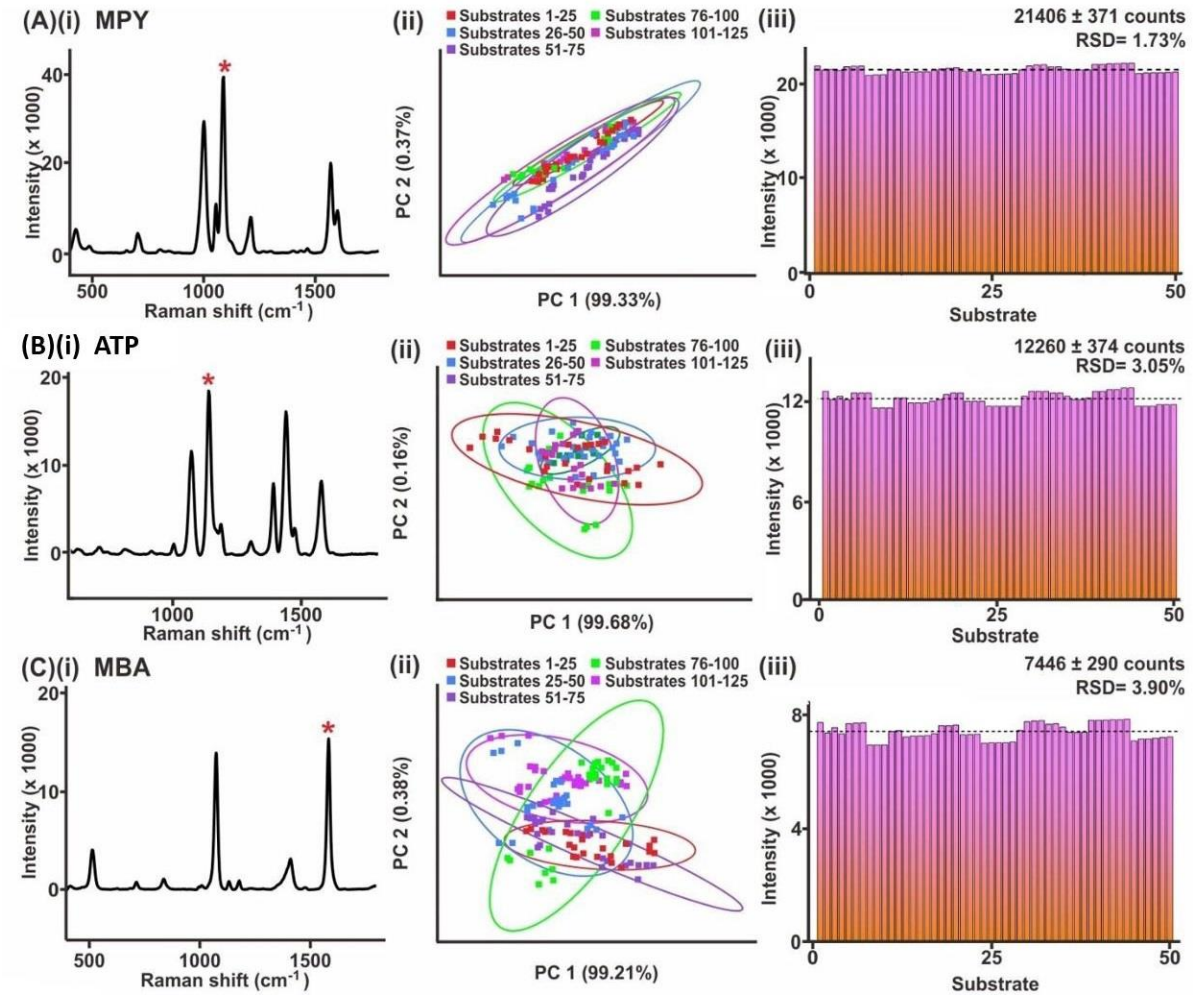


Figure 4-5. Sensor reproducibility test for (A) MPY, (B) ATP, and (C) MBA. For each receptor, the (i) representative SERS spectra, (ii) PCA score plot of 125 substrates divided into 5 subsets with substrates 1 to 25 collected on day 1, substrates 26 to 50 on day 2, substrates 51 to 75 on day 3, substrates 76 to 100 on day 4, and substrates 101 to 125 on day 5, and (iii) SERS signal homogeneities across 50 substrates are shown.

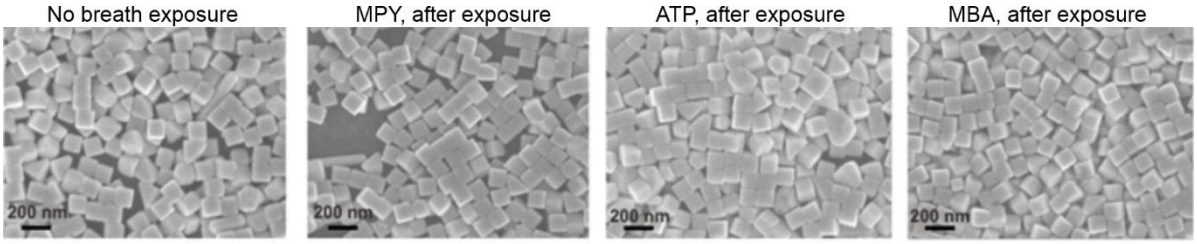


Figure 4-6. SEM images of Ag nanocubes after assembly (before breath exposure), and after breath exposure for each receptor.

4.2.2 Chemical analysis of breath profiles

In a comparative case-control clinical study conducted in Singapore involving 501 participants, we evaluated the efficacy of our SERS-based breathalyzer in identifying COVID-positive individuals. Each participant was provided with a breathalyzer where they had to take a deep breath before exhaling continuously for 10 s to collect alveolar air from deeper regions within the lung involved in the exchange of VOCs between the respiratory and circulatory systems (**Figure 4-2**).[7, 11] Subsequently, the breathalyzer was disinfected and placed within a transparent resealable bag to incubate for at least 2 min, allowing some time for the BVOCs to interact with the receptors on our SERS sensor prior to SERS measurement (**Figure 4-7**). The whole process takes less than 5 min, which is ideal for high throughput screening exercises during epidemics/pandemics. To benchmark our results, we also collected a nasopharyngeal swab sample from each participant for a PCR test within 48 h of breath collection. Across the 501 participants, 74 (14.8%) tested COVID-positive based on their PCR test results, with 31 showing no symptoms at the point of testing (**Figure 4-8A**).

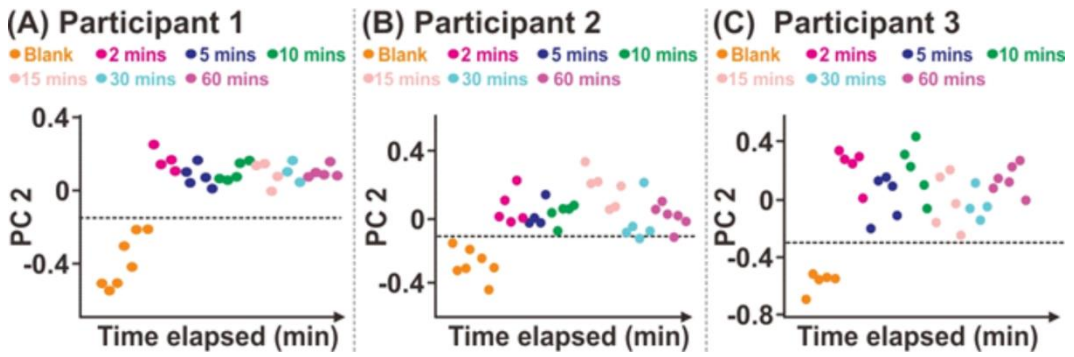


Figure 4-7. Determining the minimum incubation time to allow formation of receptor-VOC interactions. Across all 3 participants, significant spectral changes occurred after 2 min of breathalyzer incubation and remained consistent thereafter.

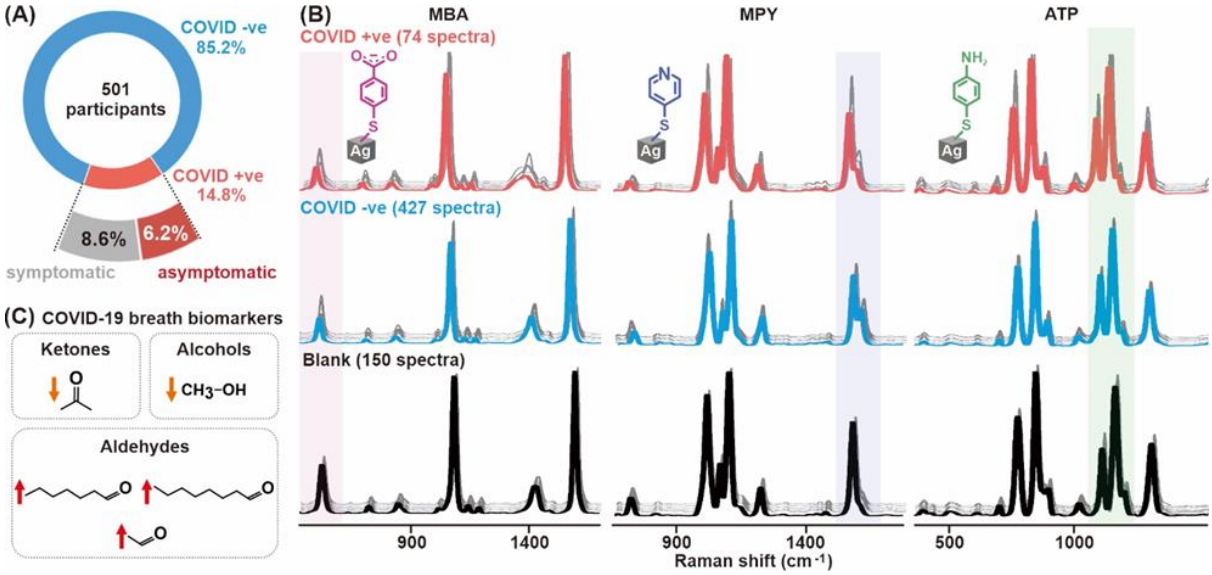


Figure 4-8. SERS spectra of exhaled breath samples collected in the clinical trial. (A) Proportion of COVID-positive and COVID-negative participants. (B) SERS spectra of each receptor in the absence ('blank') and presence of COVID-positive and COVID-negative breath samples. Highlighted regions indicate peaks with significant changes. (C) Reported potential COVID-19 breath biomarkers and their molecular structures. In COVID-positive breath samples, their relative concentration changes are indicated with arrows.

Upon analyzing the SERS spectra of each receptor in the absence of breath sample ('blank'), presence of COVID-positive breath (74 samples) and presence of COVID-negative breath (427 samples), we identified key spectral differences that distinguish each class from the others (**Figure 4-8B**). To investigate the origins of these spectral differences, we compare them to changes observed with specific VOCs that were reported to be potential COVID-19 breath biomarkers, including methanol, ethanal, heptanal, octanal and acetone (**Figure 4-8C**).[1-3] Water vapor is also included as the main interference to the SERS signals acquired. Similar to the breath sample collection, the SERS sensor was incubated with saturated neat standards of each VOC individually within an enclosed system (**Table 4-1**). The saturated vapor concentration is calculated based on the ideal gas equation as follows:

$$PV = nRT$$

where P is the saturated vapor pressure at 35 °C (Pa), V is volume of enclosed vial (cm^3), n is the number of moles of each VOC (mol), R is the universal gas constant (8.314×10^6 cm^3 Pa K^{-1} mol^{-1}) and T is the incubation temperature (K).

Rearranging the equation,

$$\text{Saturated vapor concentration (g cm}^{-3}) = \frac{V}{RT} \times MW$$

Where MW is the molecular weight of the target analyte (g mol^{-1}).

The saturated concentration can be converted from g cm^{-3} to ppm by the following relationship,

$$\text{Saturated concentration (ppm)} = \text{Saturated concentration (g cm}^{-3}) \times 10^6$$

Table 4-1. Saturated vapor concentrations of each potential COVID-positive breath biomarker.

Compound	MW (g mol^{-1})	Saturated vapor pressure (Pa)	Saturated vapor concentration (ppm)
ethanal	44.05	143,280	2465
heptanal	100.21	569	22
octanal	128.212	80	4
acetone	58.08	37,519	851
methanol	32.04	21,697	271
water	18.02	5,626	40

For MBA, the C-S stretching (νCS) peak at 521 cm^{-1} decreases in intensity from 0.29 ± 0.03 in blank to 0.19 ± 0.05 and 0.22 ± 0.09 in the presence of COVID-positive and COVID-negative breath, respectively. Notably, the decrease in νCS intensity is larger for COVID-positive breath than COVID-negative breath (**Figure 4-9A**).[12] The νCS mode is indicative of the relative polarizability of the C-S bond and is primarily influenced by ion-dipole interactions with carbonyl compounds such as ethanal and heptanal or hydrogen bonding with hydroxyl-containing compounds such as methanol and water.[13] In pure VOC experiments, a decrease in νCS is also observed when exposed to ethanal and heptanal while an increase is observed when exposed to methanol and water. It is noteworthy that the overall decrease in νCS intensity when exposed to water indicates that the relative humidity of exhaled breath is not a significant source of interference to the MBA νCS mode.

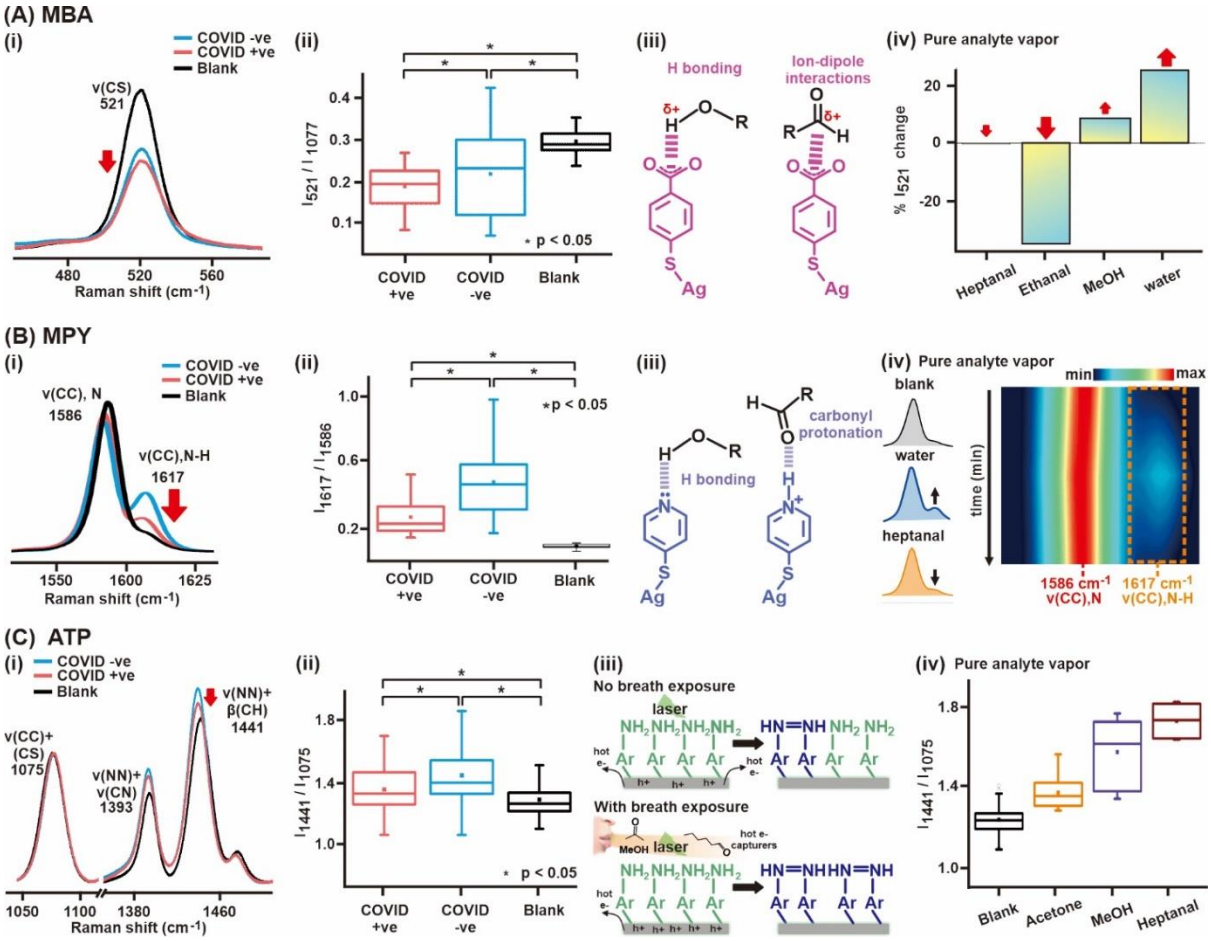


Figure 4-9. SERS analysis of COVID-positive, COVID-negative, and blank spectra. (A) MBA – (i) MBA SERS peak at 521 cm^{-1}. (ii) Box plot comparison of the 521 cm^{-1} peak intensity. (iii) Formation of ion-dipole interactions and hydrogen bonding between MBA and the VOCs. (iv) Percentage change in the 521 cm^{-1} peak intensity. (B) MPY – (i) MPY I_{1617}/I_{1586} SERS peak intensity ratio. (ii) Box plot comparison of the I_{1617}/I_{1586} SERS peak intensity ratio. (iii) Formation of hydrogen bonding between protonated and deprotonated MPY with the VOCs. (iv) Evolution of the 1550 – 1625 cm^{-1} region upon first exposure to water vapor, followed by heptanal vapor. The changes in peak intensity ratio are illustrated in the inset. (C) ATP – (i) ATP 1030 – 1600 cm^{-1} region. (ii) Box plot comparison of the 1441 cm^{-1} SERS peak intensity. (iii) Laser-induced ATP dimerization to DMAB in the presence of BVOCs acting as hot electron acceptors. (iv) Box plot comparison of the 1441 cm^{-1} peak after exposure to neat VOCs.

The differences in box plots are determined to be statistically significant at $p < 0.05$ using the Mann-Whitney rank sum test, as indicated by the asterisk (*).

For MPY, the ratio of the aromatic C=C stretching (νCC) twin peak at 1586 and 1617 cm^{-1} (I_{1617}/I_{1586}) increases from 0.091 ± 0.011 in blank to 0.265 ± 0.116 and 0.477 ± 0.194 in the presence of COVID-positive and COVID-negative breath, respectively. Notably, the decrease in I_{1617}/I_{1586} is smaller for COVID-positive breath than COVID-negative breath (**Figure 4-9B**). As the 1586 and 1617 cm^{-1} peaks are indexed to νCC when the nitrogen is deprotonated and protonated respectively, the I_{1617}/I_{1586} describes the amount of protonated pyridine species present.[14] Prior to breath exposure, I_{1617}/I_{1586} is low, indicating that most MPY is in the deprotonated state. After exposure to breath, pseudo-protonated MPY are formed via hydrogen bonding with hydroxyl-containing compounds such as methanol and water, resulting in an increase in I_{1617}/I_{1586}.[15-16] In pure VOC experiments, a similar increase in I_{1617}/I_{1586} is also observed when exposed to water but subsequently decreases when exposed to heptanal. This is because the carbonyl compounds compete for protons to form protonated carbonyl species, which slightly lowers the relative amount of pseudo-protonated pyridine species formed.

For ATP, the azobenzene N=N stretching coupled with C-H bending (νNN + βCH) peak at 1441 cm^{-1} increases in intensity from 1.272 ± 0.116 in blanks to 1.339 ± 0.179 and 1.430 ± 0.187 in the presence of COVID-positive and COVID-negative breath, respectively. Notably, the increase in νNN + βCH is smaller for COVID-positive breath than COVID-negative breath (**Figure 4-9C**). The νNN + βCH vibrational mode originates from 4,4-dimercaptoazobenzene (DMAB), which is formed by laser-induced dimerization of ATP.[17] The increase in νNN + βCH indicates a larger extent of ATP dimerization as facilitated by the BVOCs present.[18-20] In pure VOC experiments, a similar increase in νNN + βCH is also observed when exposed to acetone, methanol and heptanal.

The observation of a larger decrease in νCS peak intensity for MBA, smaller decrease in I_{1617}/I_{1586} for MPY and smaller increase in νNN + βCH for ATP is consistent with reduced acetone and methanol levels as well as increased ethanal, heptanal and octanal levels reported in literature.[1-3] The strong correlations between our experimental spectral findings and reported potential COVID-19 biomarkers affirm that our SERS sensor effectively captures the changes in BVOC concentrations within the COVID-positive breath profile. Due to the non-specific nature of our receptors, our SERS sensor records the cumulative response of each receptor to all BVOCs present. When combining the different SERS responses of individual receptors, these spectral changes can reinforce one another to form characteristic SERS 'breath-prints' that can be used as unique identifiers of an individual's COVID-19 infection status. Such an array recognition technique is highly advantageous because it eliminates the need to isolate and identify individual components for class differentiation, which is tedious and cumbersome.

4.2.3 Constructing the classification model

With an in-depth understanding of the spectral regions contributing to the differentiation of breath profiles based on their COVID-19 infection status, we construct a binary classification model using PLSDA to achieve rapid, high throughput analyses. PLSDA is an established technique that maximizes and combines the largest SERS spectral covariances between different datasets as latent variables (LV) to achieve maximum differentiation between COVID-19 positive and negative breath profiles.[21-23] The PLSDA algorithm requires minimal computational power and produces classification scores that are easily comprehensible, making it particularly suitable for our application as a mass screening tool. Before constructing the PLSDA model, SERS spectra derived from all three receptors are baseline corrected, normalized, and concatenated as a single SERS super-profile (**Figure 4-10**).

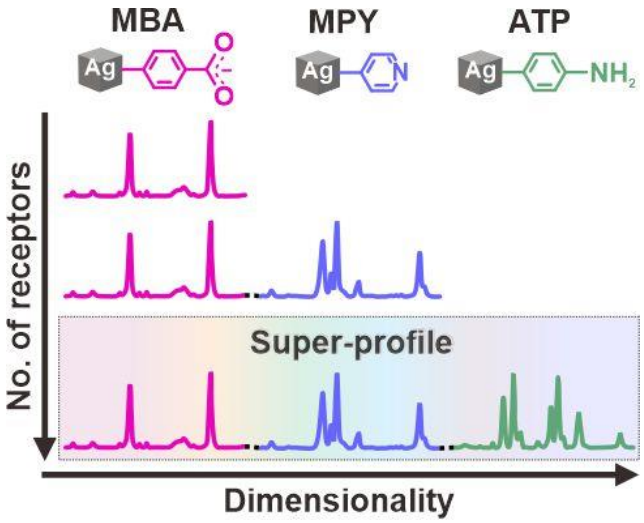


Figure 4-10. Formation of SERS super-profiles using spectral information from multiple receptors to increase the data dimensionality.

Each SERS super-profile effectively harnesses spectral variances arising from receptor-BVOC interactions, creating an additive effect that enhances the differentiation of COVID-positive and negative classes. Next, a random stratified sampling algorithm is used to split the dataset into a train set and prediction set comprising 80% and 20% of the original dataset respectively over 50 different iterations to generate 50 classification outcomes with each prediction set. Such iterations minimize any potential issues with selection bias, chance classification outcomes, and model overfit.[24-25]

Overall, the PLSDA model achieves an average classification sensitivity of 96.2% and specificity of 99.9% when distinguishing COVID-positive and negative breath profiles as benchmarked using PCR tests (**Figure 4-11A**). The sensitivity and specificity is calculated as follows:

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$= \frac{71}{71+3} \times 100\%$$

$$= 96\%$$

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}$$

$$= \frac{427}{427 + 0} \times 100\%$$

$$= 100\%$$

The low average false-negative rate of 3.8% is superior to commercially available antigen rapid tests with reported false-negative rates of 10 – 30% and is comparable to PCR tests given similar sample sizes.[26-28] Notably, asymptomatic COVID-positive individuals can be accurately classified, indicating that characteristic BVOC changes do occur and can be detected regardless of symptoms (**Figure 4-11A inset**). This is consistent with recent studies reporting that the lack of symptoms does not preclude internal physiological changes.[29-30] Timely detection of such asymptomatic individuals through active mass screening is especially critical to disrupt the silent viral spread into local communities that often remains undetected until a massive outbreak occurs.

We further use the PLSDA score plot and loadings plot to highlight how different receptor spectral regions influence the classification outcome, to establish a robust relationship between the classification results and previously identified regions which showed distinct differences. The first two LVs of the PLSDA loadings plot is important in describing regions contributing to the largest variances between the two classes.[31-32]

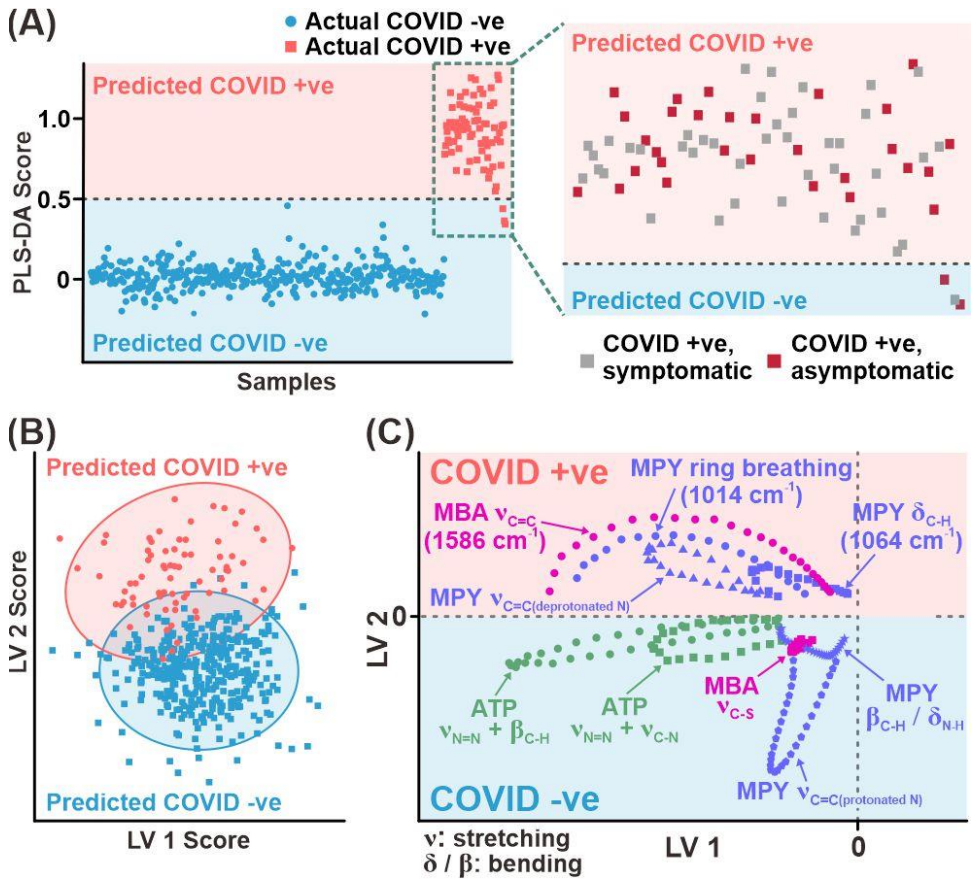


Figure 4-11. PLSDA for rapid, high throughput classification of breath profiles based on their COVID-19 infection status. (A) PLSDA score plot derived from the classification of individual SERS super-profiles showing clear distinction between the breath profiles of COVID-positive and COVID-negative individuals. Inset shows the zoomed-in segment of the PLSDA score plot for COVID-positive individuals, illustrating that symptoms do not affect their classification scores. (B) PLSDA score plot of the first two LVs, highlighting the influence of LV 2 in classifying COVID-positive and COVID-negative individuals. (C) PLSDA loadings plot for the first two LVs to illustrate specific receptor vibrational modes which influence the classification of COVID-positive and COVID-negative individuals.

From the score plot, we observe that COVID-positive breath samples typically show more positive LV 2 scores while COVID-negative breath samples show more negative LV 2 scores (**Figure 4-11B**). We note that the distribution of data points along LV 1 for both class

groups is due to intraclass variances, which can be attributed to variations in BVOC concentrations among different individuals.[33] Nonetheless, this does not affect the COVID-positive/negative clustering along LV 2. In combination with the loadings plot, we can then correlate spectral regions which are assigned positive LV 2 scores as regions contributing more significantly to a COVID-positive classification outcome, and *vice versa* (**Figure 4-11C**). For instance, MBA's C-S stretching (1077 cm^{-1}), MPY's ring breathing (1014 cm^{-1}) and C=C stretching (deprotonated N) (1586 cm^{-1}), and ATP's C-H bending + C-N stretching (1143 cm^{-1}) and C-H bending (1186 cm^{-1}) are assigned positive LV 2 scores. This signifies that the cumulative effect of peak intensity and/or peak position variances from these vibrational modes contribute to the classification of a breath profile as COVID-positive. On the other hand, MBA's C-S stretching (521 cm^{-1}), MPY's C-H + N-H bending (1224 cm^{-1}) and C=C stretching (protonated N) (1617 cm^{-1}), and ATP's N=N + C-N stretching (1393 cm^{-1}) and N=N stretching + C-H bending (1441 cm^{-1}) are assigned negative LV 2 scores and therefore are crucial in classifying COVID-negative breath profiles. This thus affirms that the amalgamation of multiple receptor spectral changes in our SERS super-profiles are important in assigning the COVID-positive or COVID-negative class. Furthermore, it proves that our model is built upon valid spectra variances arising from chemical interactions between receptor–BVOC and the change in BVOC concentrations and not spectral noise.

To emphasize the importance of the multi-receptor SERS super-profile, we demonstrate the distinct sensitivity improvement from 80 to 96.2% when comparing a single SERS receptor with our SERS super-profile sensor (**Figure 4-12**). An increase in the number of correctly classified COVID-positive breath profiles can be observed as the number of receptors increases from one to three. This increase exemplifies that each receptor imbues enhanced distinguishing capabilities to our SERS sensor by increasing the total number of distinct features between the breath profiles of COVID-positive and COVID-negative individuals. Such an approach is

critical for complex sample matrices to allow our SERS sensor to record a more complete description of the differences in breath profiles. Notably, a high specificity can be achieved even with a single receptor as it is comparatively easier to distinguish a sensor that is exposed to breath, than accurately identifying a COVID-positive breath profile.

No. of receptors	Sensitivity	Specificity
MBA	80.0%	99.7%
MPY	90.5%	99.7%
ATP	84.9%	99.7%
MBA+MPY	92.2%	99.8%
MBA+ATP	94.6%	99.9%
MPY+ATP	95.0%	99.8%
MBA+MPY+ATP	96.2%	99.9%

Figure 4-12. Classification sensitivity and specificity for an increasing number of receptors using averaged classification outcomes across 50 model iterations.

4.2.4 Model analysis in relation to clinical trial

Through rigorous analysis of our clinical trial results, we highlight the key strengths of our SERS sensor based on its performance given a specific use case. The overall sensitivity of 96.2% (95% CI: 91.8 – 100%) and specificity of 99.9% (95% CI: 99.7 – 100%) can be derived by constructing a confusion matrix using the averaged classification outcomes across 50 model iterations (**Figure 4-13**). Both the positive and negative predictive values (PPV and NPV) are > 99%, indicating high accuracy of our PLSDA model in predicting the presence of COVID-19 at the disease prevalence of our clinical studies.[34]

	Actual positive	Actual negative	
Predicted positive	71 (71.16)	0 (0.28)	PPV = 99.6% (98.2-100%)
Predicted negative	3 (2.84)	427 (426.72)	NPV = 99.3% (98.6-100%)
	Sensitivity = 96.2% (91.8-100%)	Specificity = 99.9% (99.7-100%)	

Figure 4-13. Confusion matrix of the averaged classification outcomes across 50 model iterations. Values in green and brown indicate correct and incorrect classification outcomes, respectively. Actual values before rounding off are given in gray brackets. The sensitivity, specificity, positive, and negative prediction values are in blue, with their corresponding 95% confidence intervals in gray brackets directly below.

When considering the model sensitivity in relation to displayed COVID-19 symptoms, we note that our model shows high sensitivities of 97.7% and 93.6% for both symptomatic and asymptomatic individuals (**Figure 4-14A**). The slightly lower sensitivity when identifying asymptomatic COVID-positive individuals could be due to the limited sample size of 31 participants. To enhance the ability of our sensor in picking up asymptomatic COVID-positive individuals, it is essential to collect more data to elucidate spectral features which are important for their classification. In addition, out of 70 participants with reported comorbidities including asthma and thyroid dysfunctions (4 COVID-positive, 66 COVID-negative), all 70 participants are accurately classified in their respective COVID-19 infection status (**Figure 4-15**). This indicates that the presence of pre-existing medical conditions does not affect the prediction outcome of our SERS sensor.

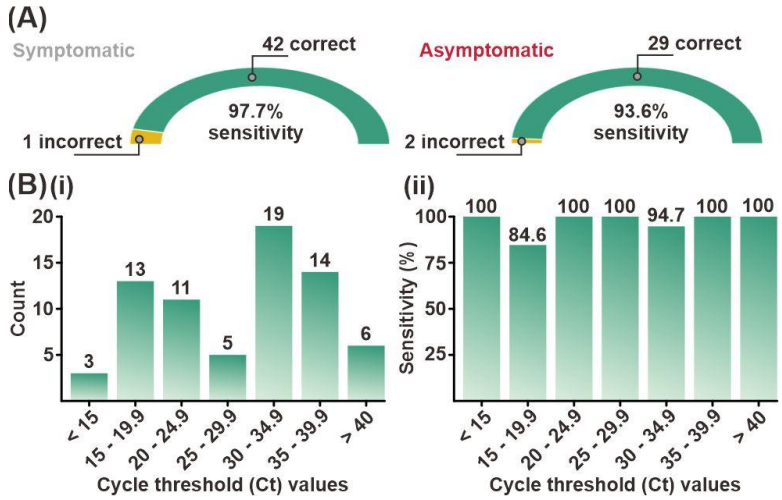


Figure 4-14. (A) Sensitivity of our SERS sensor in the classification of symptomatic and asymptomatic COVID-positive individuals. (B) The number of COVID-positive participants based on their respective Ct values determined by a PCR test and the model sensitivity at each Ct range.

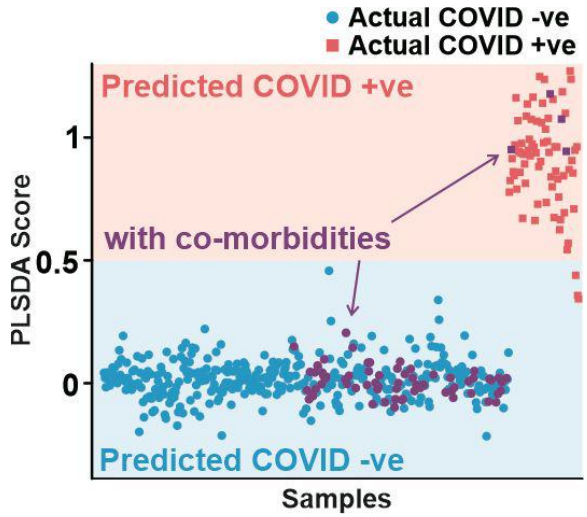


Figure 4-15. PLSDA score plot highlighting participants with reported comorbidities in purple.

Importantly, we demonstrate good representation of individuals at various stages of COVID-19 infection in our clinical trial with PCR cycle threshold (Ct) values ranging from <15 to >40 (**Figure 4-14B**). The PCR Ct value indicates the relative viral load in an infected individual, whereby a low Ct value is equivalent to a high viral load. Notably, the high sensitivity of our SERS sensor across a large range of Ct values indicates that there are distinct

BVOC differences for all COVID-positive individuals regardless of the viral load in their bodies. This is crucial in ensuring our breathalyzer's effectiveness in picking up infected individuals across all stages of infection, as these individuals may still be potentially infectious.[35-37]

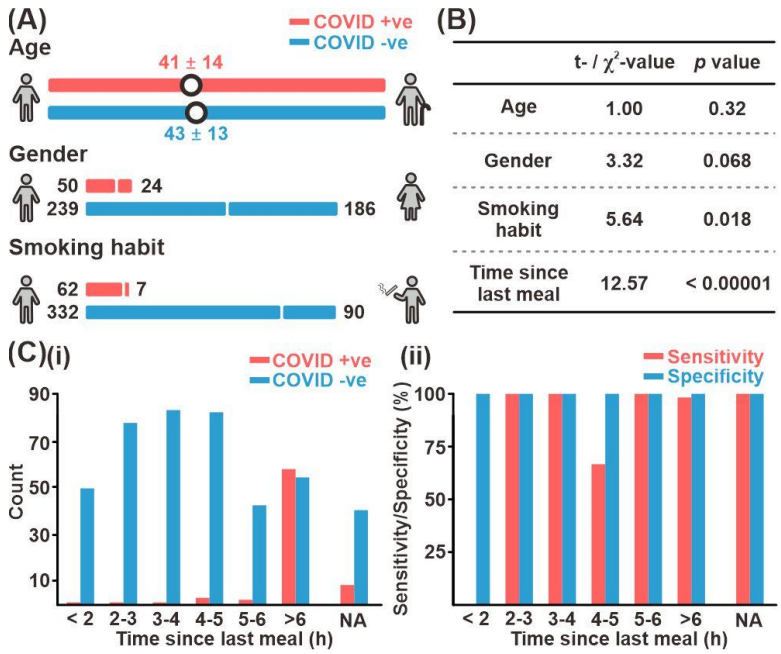


Figure 4-16. (A) Participant demographics such as their mean age, gender, and smoking habits. (B) Statistical test results of potential confounding factors such as participants' age, gender, smoking habits, and time since the last meal using either the t test or χ^2 test, with their corresponding p-value. (C) Analysis of time since the last meal as a potential confounding factor based on (i) distribution of time since last meal of all participants (a small number of participants were unable to recall this information (denoted as NA) and (ii) the model sensitivity and specificity at each time range.

In addition, we ascertain that other potential confounding factors such as age, gender, smoking habits, and time since the last meal do not significantly influence our classification, by employing the t test and χ^2 test (**Figure 4-16**). Given a mean age of 41 ± 14 years old for

COVID-positive participants and 43 ± 13 years old for COVID-negative participants, age did not significantly influence our classification at the 95% confidence level ($t = 1.00$, $p = 0.32$). Similarly, the gender distribution of 50 males and 24 females across COVID-positive participants and 239 males and 186 females across COVID-negative participants did not affect our classification ($\chi^2 = 3.32$, $p = 0.068$). When investigating the effect of smoking habits, since only 6 COVID-positive and 90 COVID-negative participants smoke, we set a higher critical value of $p = 0.01$. At a 99% confidence level, the participants' smoking habits ($\chi^2 = 5.64$, $p = 0.018$) did not impact our classifications. However, a larger sample size is likely necessary to arrive at a more statistically robust conclusion for this factor. In terms of the time elapsed since the participants' last meal, there is significant imbalance as most COVID-positive participants (58 out of 66) did not consume any food for > 6 h prior to the breathalyzer test. This resulted in a significant difference between the average time since the last meal ($t = 12.57$, $p < 0.00001$) for COVID-positive and COVID-negative participants. This difference stems from an inherent limitation in operational protocol, as breath samples were typically collected from COVID-positive individuals by nurses before their breakfast at a specified timing. In contrast, breath samples from COVID-negative individuals were collected after disembarking from a flight, which had meals provided, with no restrictions as to when they are allowed to consume any food. Nonetheless, we note that even with such an imbalance, the high classification specificity of 99.9% is a clear indicator that the differences between COVID-positive and COVID-negative breath samples were much more pronounced compared to any differences in breath composition resulting from food consumption.

4.3 Conclusion

In conclusion, we showcase our design of a SERS-based breathalyzer for rapid, noninvasive screening of individuals for COVID-19, achieving a sensitivity of 96.2% and specificity of 99.9%. Through the strategic use of multiple molecular receptors to capture and interact with various BVOCs in exhaled breath, we generate highly informative SERS super-profiles that harness each receptor's distinguishing power. Fundamentally, we establish good qualitative agreement between our observed SERS spectral variances with those induced by pure VOC vapors of several potential COVID-19 biomarkers. The in-depth understanding of these spectral differences allows us to construct a robust PLSDA model which attains a false negative rate superior to commercially available antigen rapid tests and comparable to that of PCR tests. In addition, the classification accuracy is independent of whether the individual displays COVID-19-related symptoms and other confounding factors such as age, gender, and smoking habits before breath collection. Most importantly, our test is simple, easy to administer, and requires only 5 min from sample collection to output of results for rapid turnover. As the world adjusts to a new normal, government strategies are shifting toward scaling up of COVID-19 testing, contact tracing, and vaccination. In this aspect, our breathalyzer can play a significant role in fulfilling this goal by supporting mass screening capabilities even at locations with high human traffic. Breath collection and measurements can be performed in parallel, which overcomes the current bottleneck in conventional GC–MS methods for breath analysis, making it suitable for testing in diverse settings and locations like schools, airports, and events like weddings, religious events, and conferences. Moreover, our findings from this work lay the foundation for next-generation breath-based detection of other respiratory and/or non-respiratory related diseases using SERS.

4.4 Materials and methods

Chemicals. Silver nitrate, 1,5-pentanediol (PD), poly(vinylpyrrolidone) (PVP; Mw ~55,000), 4-mercatopyridine (MPY), 4-mercaptobenzoic acid (MBA) and 4-aminothiophenol (ATP) were purchased from Sigma Aldrich. Copper (II) chloride was purchased from Alfa Aesar. Ethanol (ACS, ISO, Reag. Ph Eur) was obtained from Merck. Milli-Q water (> 18.0 MΩ. cm) was purified with a Sartorius Arium® 611 UV ultrapure water system. All reagents were used without further purification.

Synthesis of Ag nanocubes. Ag nanocubes were synthesized via the polyol method described in literature.[38] Briefly, 0.50 g of silver nitrate and 0.86 µg copper (II) chloride were dissolved in PD in a scintillation vial. Separately, 0.25 g of PVP was dissolved in PD. Using a temperature - controlled silicone oil bath, 20 mL of PD was heated for 10 min. The two precursor solutions were then injected into the hot reaction flask at different rates: 500 µL of silver nitrate solution every minute and 250 µL of PVP solution every 30s. This addition was stopped once the solution turned ochre. The Ag nanocubes were purified via several rounds of centrifugation and subsequently stored in ethanol. Scanning electron microscopic (SEM) imaging was carried out using JEOL-JSM-7600F electron microscope at an accelerating voltage of 5 kV.

Thiophenol functionalization of Ag nanocubes. Functionalization of Ag nanocubes surfaces was performed through individual ligand exchange reactions. 50 µL of 10 mM of a thiophenol solution (MPY, MBA, ATP) were separately added to 1 mL of Ag nanocubes each and allowed to stir overnight. The functionalized Ag nanocubes were then purified via centrifugation and dispersed in 1 mL ethanol.[39]

Sensor chip and breathalyzer fabrication. An automated liquid dispensing system (Y&D 7300N Smart Robot; Y&D Technology Co. Ltd.) was used to dispense the functionalized Ag nanocubes. The functionalized Ag nanocubes were first dispersed in aqueous solutions, carefully loaded into the dispensing system, then precisely dispensed onto an aluminum plate (5 μm × 5 μm in size). The dispensed Ag nanocubes were then allowed to dry under controlled conditions (24 °C with relative humidity of 40%). SERS signals of the dried droplets were measured to ensure sensor chip signal reproducibility and consistency before they were individually assembled into a breathalyzer. The assembled breathalyzer and an accompanying cap were vacuum-sealed prior to its usage during clinical trials.

Breath sample collection. Participants aged between 18 – 99 were recruited at multiple study sites for clinical trials, including the National Center for Infectious Diseases and Changi International Airport in Singapore. All recruitment protocols were covered under NTU's IRB-2020-12-012 and IRB-2021-03-046. Study participants were adequately briefed regarding the research goals and aims, and their consent were sought prior to sample collection. All breathalyzers were de-identified from the study participants with the use of unique subject identification numbers. During sample collection, a sealed vacuum package containing the breathalyzer was handed to the participant. The participant was directed to blow gently and continuously into the breathalyzer mouthpiece for 10 s before affixing the safety cap. The breathalyzer was then disinfected with 70% ethanol before SERS measurement. Each breathalyzer is fitted with medical grade HEPA filter at the outlet to isolate any pathogens present within the breath chamber and prevent escape into the external environment.

SERS measurement of breath samples. SERS measurements were conducted using the portable Metrohm Raman spectrometer (Mira DS) with an excitation wavelength of 785 nm,

laser power of 50 mW and an acquisition time of 0.05 s with an average of 5 raster scans. The spectral window of 400 to 1800 cm^{-1} was used for data analyses. Spectral pre-processing includes baseline correction using the adaptive iteratively reweighted penalized least squares (airPLS) algorithm and min-max normalization.[40] The processed SERS spectra from all three receptors were then concatenated into a SERS super-profile representing the breath profile of a participant.

Model building. The partial least-squares discriminant analysis (PLS-DA) models were constructed using the Python-based scikit-learn package.[41] In one iteration, data was first split into a 80% train and 20% test set using random state = 1. The train set was optimized and cross-validated using a k-fold cross-validation algorithm, with k = 10. Root-mean-squared errors resulting from the train set classification and averaged cross-validation classifications were derived and used to determine the number of latent variables selected for a PLS-DA model. The test set was then used to assess the outcome of our classification model through calculating its sensitivity and specificity. This process was then repeated for an additional 49 iterations using random states 2 to 50 to derive the averaged sensitivity and specificity of our SERS sensor.

SERS measurements of pure analyte vapor. The SERS sensor is incubated separately with 200 μL of a target analyte at 35 °C in an enclosed 20 mL vial. SERS detection was performed after 6 h of incubation using the same spectrometer system, measurement parameters and data pre-processing.

Participant Statistics. Participant statistics for categorical variables such as age and gender were presented as number (%). Continuous variables such as intensity ratios were presented as

mean ± standard deviation. The statistical significance of each variable between blanks and COVID-positive, blanks and COVID-negative, as well as COVID-positive and COVID-negative were assessed with Mann Whitney rank sum test. All tests were two-tailed with $p < 0.05$ as the significance threshold. Calculations were performed using the OriginPro 9.0 software. The statistical significance of each confounding factor on the classification was assessed using either a t-test (for continuous variable) or a chi-square test (categorical variable). The choice of statistical test depends on several parameters including the variable type (categorical/continuous) and distributions (normal/non-normal).

DFT simulations. The calculations on the interaction of Ag surface with various target analyte molecules were carried out using the unrestricted B3LYP exchange-correlation functional, as implemented in the Gaussian 09 computational chemistry package. The 6-31G(d, p) basis set was used for all atoms except Ag, for which the LANL2DZ basis set was employed. The Ag surface was modelled using a reported triangle consisting of 6 Ag atoms.[42] After geometry optimization of the triangular Ag cluster, each target analyte molecule was then placed near the Ag cluster (< 2 Å) and the entire system was re-optimized before obtaining the simulated spectra.

References

1. Chen, H.; Qi, X.; Ma, J.; Zhang, C.; Feng, H.;Yao, M., *medRxiv*, 2020, 2020.06.21.20136523.
2. Grassin-Delyle, S.; Roquencourt, C.; Moine, P.; Saffroy, G.; Carn, S.; Heming, N.; Fleuriet, J.; Salvator, H.; Naline, E.; Couderc, L.-J.; Devillier, P.; Thévenot, E. A.; Annane, D., *eBioMedicine*, 2021, 63.
3. Ruszkiewicz, D. M.; Sanders, D.; O'Brien, R.; Hempel, F.; Reed, M. J.; Riepe, A. C.; Bailie, K.; Brodrick, E.; Darnley, K.; Ellerkmann, R.; Mueller, O.; Skarysz, A.; Truss, M.; Wortelmann, T.; Yordanov, S.; Thomas, C. L. P.; Schaaf, B.; Eddleston, M., *EClinicalMedicine*, 2020, 29.
4. Gupta, A.; Madhavan, M. V.; Sehgal, K.; Nair, N.; Mahajan, S.; Sehrawat, T. S.; Bikdeli, B.; Ahluwalia, N.; Ausiello, J. C.; Wan, E. Y.; Freedberg, D. E.; Kirtane, A. J.; Parikh, S. A.; Maurer, M. S.; Nordvig, A. S.; Accili, D.; Bathon, J. M.; Mohan, S.; Bauer, K. A.; Leon, M. B.; Krumholz, H. M.; Uriel, N.; Mehra, M. R.; Elkind, M. S. V.; Stone, G. W.; Schwartz, A.; Ho, D. D.; Bilezikian, J. P.; Landry, D. W., *Nat. Med.*, 2020, 26, 1017-1032.
5. Cao, W.; Duan, Y., *Crit. Rev. Anal. Chem.*, 2007, 37, 3-13.
6. Kim, K. H.; Jahan, S. A.; Kabir, E., *Trends Anal. Chem.*, 2012, 33, 1-8.
7. Lawal, O.; Ahmed, W. M.; Nijsen, T. M. E.; Goodacre, R.; Fowler, S. J., *Metabolomics*, 2017, 13, 110.
8. Koh, C. S. L.; Lee, H. K.; Han, X.; Sim, H. Y. F.; Ling, X. Y., *Chem. Commun.*, 2018, 54, 2546-2549.
9. Lee, H. K.; Lee, Y. H.; Koh, C. S. L.; Phan-Quang, G. C.; Han, X.; Lay, C. L.; Sim, H. Y. F.; Kao, Y.-C.; An, Q.; Ling, X. Y., *Chem. Soc. Rev.*, 2019, 48, 731-756.
10. Leong, Y. X.; Lee, Y. H.; Koh, C. S. L.; Phan-Quang, G. C.; Han, X.; Phang, I. Y.; Ling, X. Y., *Nano Lett.*, 2021, 21, 2642-2649.

11. Sethi, S.; Nanda, R.; Chakraborty, T., *Clin. Microbiol. Rev.*, 2013, 26, 462-75.
12. Capocefalo, A.; Mammucari, D.; Brasili, F.; Fasolato, C.; Bordi, F.; Postorino, P.; Domenici, F., *Front. Chem.*, 2019, 7, 413.
13. Wang, Y.; Ji, W.; Sui, H.; Kitahama, Y.; Ruan, W.; Ozaki, Y.; Zhao, B., *J. Phys. Chem. C*, 2014, 118, 10191-10197.
14. Guerrini, L.; Rodriguez-Loureiro, I.; Correa-Duarte, M. A.; Lee, Y. H.; Ling, X. Y.; García de Abajo, F. J.; Alvarez-Puebla, R. A., *Nanoscale*, 2014, 6, 8368-8375.
15. Bi, L.; Wang, Y.; Yang, Y.; Li, Y.; Mo, S.; Zheng, Q.; Chen, L., *ACS Appl. Mater. Interf.*, 2018, 10, 15381-15387.
16. Wang, Y.; Yu, Z.; Ji, W.; Tanaka, Y.; Sui, H.; Zhao, B.; Ozaki, Y., *Angew. Chem. Int. Ed.*, 2014, 53, 13866-13870.
17. Xu, P.; Kang, L.; Mack, N. H.; Schanze, K. S.; Han, X.; Wang, H.-L., *Sci. Rep.*, 2013, 3, 2997.
18. Huang, Y.-F.; Zhu, H.-P.; Liu, G.-K.; Wu, D.-Y.; Ren, B.; Tian, Z.-Q., *J. Am. Chem. Soc.*, 2010, 132, 9244-9246.
19. Liu, Y.; Yang, D.; Zhao, Y.; Yang, Y.; Wu, S.; Wang, J.; Xia, L.; Song, P., *Heliyon*, 2019, 5, e01545.
20. Wu, D.-Y.; Zhao, L.-B.; Liu, X.-M.; Huang, R.; Huang, Y.-F.; Ren, B.; Tian, Z.-Q., *Chem. Commun.*, 2011, 47, 2520-2522.
21. Barker, M.; Rayens, W., *J. Chemom.*, 2003, 17, 166-173.
22. Brereton, R. G.; Lloyd, G. R., *J. Chemom.*, 2014, 28, 213-225.
23. Wold, S.; Sjöström, M.; Eriksson, L., *Chemometr. Intell. Lab. Syst.*, 2001, 58, 109-130.
24. Westerhuis, J. A.; Hoefsloot, H. C. J.; Smit, S.; Vis, D. J.; Smilde, A. K.; van Velzen, E. J. J.; van Duijnhoven, J. P. M.; van Dorsten, F. A., *Metabolomics*, 2008, 4, 81-89.
25. Xu, Q.-S.; Liang, Y.-Z., *Chemometr. Intell. Lab. Syst.*, 2001, 56, 1-11.

26. Fouzas, S., *Lancet Infect. Dis.*, 2021, 21, 1068-1069.
27. Wikramaratna, P.; Paton, R. S.; Ghafari, M.; Lourenço, J., *medRxiv*, 2020, 2020.04.05.20053355.
28. Xie, X.; Zhong, Z.; Zhao, W.; Zheng, C.; Wang, F.; Liu, J., *Radiology*, 2020, 296, E41-E45.
29. Chang, M. C.; Lee, W.; Hur, J.; Park, D., *Respiration*, 2020, 99, 748-754.
30. Meng, H.; Xiong, R.; He, R.; Lin, W.; Hao, B.; Zhang, L.; Lu, Z.; Shen, X.; Fan, T.; Jiang, W.; Yang, W.; Li, T.; Chen, J.; Geng, Q., *J Infect.*, 2020, 81, e33-e39.
31. Goodpaster, A. M.; Kennedy, M. A., *Chemometr. Intell. Lab. Syst.*, 2011, 109, 162-170.
32. Xia, J.; Sinelnikov, I. V.; Han, B.; Wishart, D. S., *Nucleic Acids Res.*, 2015, 43, W251-7.
33. Phillips, M.; Herrera, J.; Krishnan, S.; Zain, M.; Greenberg, J.; Cataneo, R. N., *J Chromatogr. B Biomed. Sci. Appl.*, 1999, 729, 75-88.
34. Wong, H. B.; Lim, G. H., *Proc. Singapore Healthc.*, 2011, 20, 316-318.
35. Gandhi, M.; Yokoe, D. S.; Havlir, D. V., *N Engl J Med.*, 2020, 382, 2158-2160.
36. Tian, D.; Lin, Z.; Kriner, E. M.; Esneault, D. J.; Tran, J.; DeVoto, J. C.; Okami, N.; Greenberg, R. M.; Yanofsky, S.; Ratnayaka, S.; Tran, N.; Livaccari, M.; Lampp, M. L.; Wang, N.; Tim, S.; Norton, P.; Scott, J.; Hu, T. Y.; Garry, R.; Hamm, L.; Delafontaine, P.; Yin, X. M., *J Mol. Diagn.*, 2021, 23, 1078-1084.
37. Gao, Z.; Xu, Y.; Sun, C.; Wang, X.; Guo, Y.; Qiu, S.; Ma, K., *J Microbiol. Immunol. Infect.*, 2021, 54, 12-16.
38. Tao, A.; Sinsermsuksakul, P.; Yang, P., *Angew. Chem. Int. Ed.*, 2006, 45, 4597-4601.
39. Sim, H. Y. F.; Lee, H. K.; Han, X.; Koh, C. S. L.; Phan-Quang, G. C.; Lay, C. L.; Kao, Y.-C.; Phang, I. Y.; Yeow, E. K. L.; Ling, X. Y., *Angew. Chem. Int. Ed.*, 2018, 57, 17058-17062.
40. Zhang, Z.-M.; Chen, S.; Liang, Y.-Z., *Analyst*, 2010, 135, 1138-1146.

41. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R., *J Mach. Learn. Res.*, 2011, 12, 2825-2830.
42. You, T.-T.; Yin, P.-G.; Jiang, L.; Lang, X.-F.; Guo, L.; Yang, S.-H., *Phys. Chem. Chem. Phys.*, 2012, 14, 6817-6825.

Chapter 5 Augmentation-boosted Machine Learning to Promote Breath-based SERS Toolkits for Mass Screening: A Large-Scale COVID-19 Study

Abstract. Hand-held SERS breathalyzers are attractive toolkits for mass screening applications when considering non-pharmaceutical interventions to contain infectious diseases as they are non-invasive, simple to administer and offer point-of-care results. However, constructing a robust ML model is challenging due to the huge class imbalance, difficulty in obtaining positive samples from diseased individuals in discomfort, and inaccessibility of healthcare institutions during outbreaks. Here, we propose an ensemble augmentation strategy to boost COVID-positive sample data and construct a balanced ML model that achieves 99.8% sensitivity in classifying POS individuals across > 73000 exhaled breath samples. Importantly, we extract key information from SERS peaks by converting them into peak parameters using a pseudo-Voigt fitting function. Our ML model demonstrates significant improvements in sensitivity when the positive samples were amplified by 20-times and negative samples were reduced by 9%. Overall, our findings bolster the applicability of hand-held SERS breathalyzers for mass screenings during future outbreaks.

5.1 Introduction

Conducting mass screening exercises at airports and large-scale gatherings is crucial during an infectious disease outbreak to enable swift identification and isolation of infected persons.[1-2] Unlike diagnostic tests, mass screening tools primarily aim to separate the large proportion of negative cases from the positive or suspicious ones, thereby reducing the total cost of materials, manpower, and time required for confirmatory tests.[3] In the recent COVID-19 pandemic, hand-held breathalyzers embedded with surface-enhanced Raman scattering (SERS)-based nanosensors showed promising potential as simple, non-invasive, and point-of-care test kits tailored for mass screenings.[4] The nanosensor contains Ag nanocubes functionalized with multiple molecular receptors that reflect fluctuations in the amount of BVOC such as alcohols, aldehydes, and ketones through key SERS peak variations.[5-7] Participants simply had to exhale into the breathalyzer for ten seconds and a ML model will identify and analyze these SERS variances to classify individuals infected by the SARS-CoV-2 virus within five minutes. With continued development, these hand-held SERS breathalyzers could assume a pivotal role in mass screening when considering non-pharmaceutical interventions to contain infectious diseases in future outbreaks.[8]

However, constructing a robust ML model applicable for mass screening is inherently challenging as the overwhelming majority (> 95%) of healthy individuals creates a huge imbalance between the positive and negative classes. This imbalance will significantly impact the training of many ML classifiers, including powerful multi-layered perceptron (MLP) and deep learning models, skewing them heavily in favor of identifying healthy rather than diseased persons.[9-11] The inability to accurately identify diseased individuals is especially detrimental at places where people gather in large groups as they will potentially cause an uncontrolled spread of the disease. The problem is exacerbated by the difficulty in accessing healthcare institutions during outbreaks and collecting positive samples from people who are in discomfort due to the

disease symptoms. Hence, there is an urgent need for alternative strategies to resolve the imbalance between positive and negative classes and accelerate the construction of an accurate ML model that qualifies SERS-based breathalyzers for deployment as mass screening toolkits.

Herein, we introduce an ensemble augmentation strategy to boost COVID-positive (POS) sample data and construct a balanced ML model that achieves 99.8% sensitivity and a 93.8% positive predictive value (PPV) in classifying COVID-positive individuals across > 73000 exhaled breath samples (**Figure 5-1**). Our strategy entails a combination of over-sampling POS data by creating synthetic SERS spectra using the synthetic minority over-sampling technique (SMOTE) and under-sampling COVID-negative (NEG) data using the edited nearest neighbors (ENN) method.[12-13]

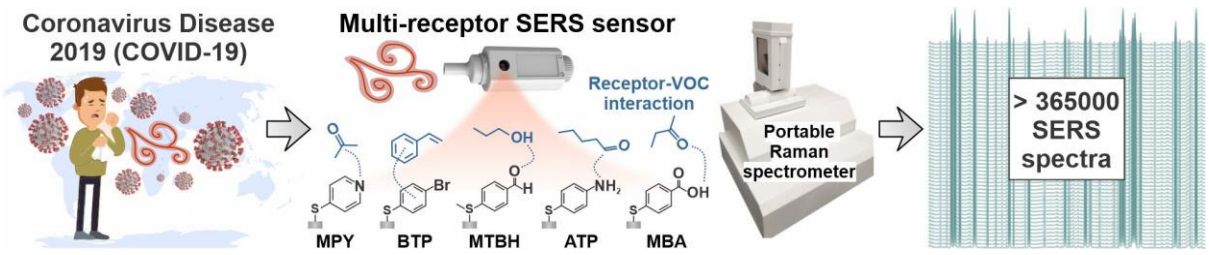


Figure 5-1. Collecting over 73000 exhaled breath samples using our multi-receptor SERS sensor embedded within a hand-held breathalyzer, culminating in > 365000 SERS spectra.

To comprehensively study the ability to establish a robust ML model for mass screening applications, we collected a total of 2460 samples from POS individuals (3.37%) and 70547 samples from NEG individuals over a period of three months in Singapore as an extension to our clinical case-control study. We built upon our previous findings by leveraging an array of five molecular receptors to elicit receptor-BVOC interactions with key biomarkers in the exhaled breath samples and combined their spectra as a SERS 'super-profile' with enhanced ability to distinguish subtle differences in their BVOC compositions. Crucially, we conducted feature engineering on the super-profiles to extract key information from the SERS peaks,

namely (1) peak position, (2) peak intensity, (3) its full width at half maximum (FWHM), (4) the proportion of Gaussian to Lorentzian character, and (5) its skewness. In doing so, we lower data dimensionality and accelerate ML computation by removing noise and redundant features, which is key when analyzing big data. Using the SMOTEENN algorithm, we highlight the need to overcome data imbalance as the MLP model demonstrated a 43.8% increase in sensitivity and a 31.4% increase in PPV when POS samples are augmented by 20-times their original amount and NEG samples are reduced by 9%. Overall, our findings further bolster the immense potential of hand-held SERS breathalyzers as toolkits for mass screening application at the point-of-care during infectious disease outbreaks.

5.2 Results and discussion

5.2.1 Data collection and feature engineering

To effectively simulate a mass screening setting and leverage the data collected for ML model construction, we collected and measured > 73000 exhaled breath samples over a period of three months in a continuation of our clinical study. We adopted a similar approach to fabricate our SERS sensor, which comprises Ag nanocubes functionalized with 4-mercaptopyridine (MPY), 4-mercaptobenzoic acid (MBA) and 4-aminothiophenol (ATP), with two additional receptors 4-methylthiobenzaldehyde (MTBH) and 4-bromothiophenol (BTP). Each receptor contains a thiol group that forms strong Ag-S covalent bond with the nanocubes and an additional functional group that promotes receptor-BVOC interactions with the key biomarkers such as ethanal, acetone, and methanol present in our exhaled breath.[5-7] These interactions result in variations in the receptor SERS peaks, which becomes more pronounced when concatenated into a single SERS super-profile. In our previous clinical study, we used a partial least-squares discriminant analysis (PLSDA) model to analyze these super-profiles and achieved 96.2% sensitivity in identifying POS individuals as opposed to 80% when using a

single receptor. Hence, we rationalize the inclusion of two additional receptors as they provide more SERS variations to further improve the ability to distinguish POS and NEG breath samples.

After data collection, we first judiciously clean and pre-process the raw data prior to ML model construction. When dealing with big data, it is crucial to remove outlier samples arising from human or instrumental errors that may confuse the subsequent ML training. We calculated the signal-to-noise (SNR) ratio for each receptor based on the regions highlighted in yellow (signal) and grey (noise) respectively (**Figure 5-2**).

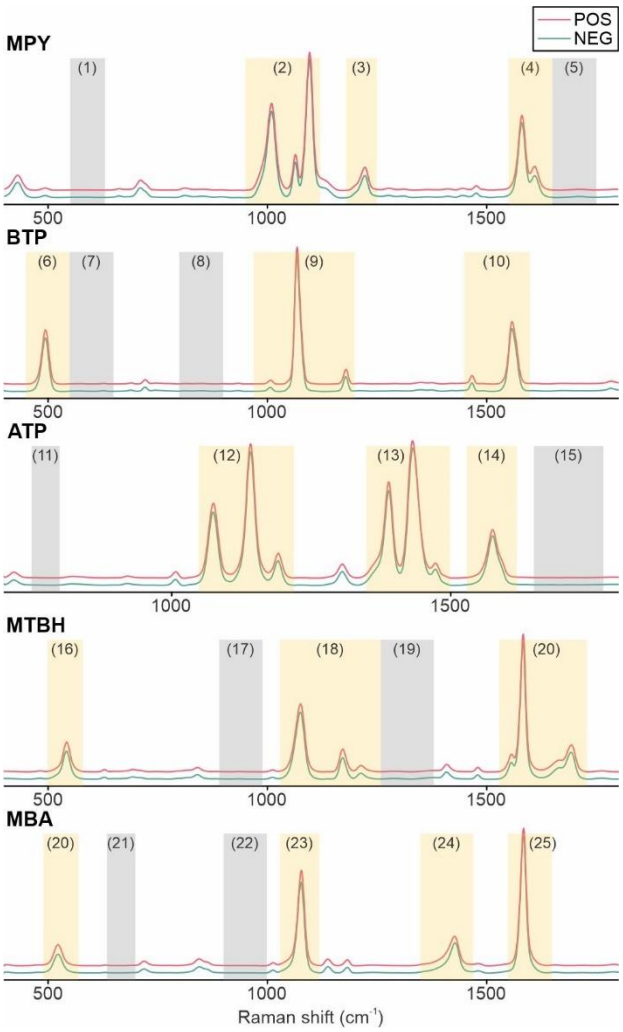


Figure 5-2. Regions designated as SERS signals (yellow) and spectral noise (grey).

The regions are determined such that each receptor has a SNR region ratio of 2.35, with three signal regions and two noise regions to ensure fair comparison across receptors. Samples that have at least one individual receptor SNR outside of two standard deviations (SD) from the mean (mean – 2 × SD) will be removed as the remaining SERS spectra will no longer form a complete super-profile. In doing so, we remove a total of 173 POS (6.57%) and 7165 NEG samples (9.21%) (**Table 5-1, 5-2**). After cleaning, the remaining SERS spectra were pre-processed using the adaptive iteratively reweighted penalized least-squares (airPLS) baseline correction and min-max normalization (**Figure 5-3**).

Table 5-1. Samples removed by data cleaning based on their SNR.

POS samples:

Receptor	SNR (mean)	SNR (SD)	Mean – 2SD	No. reject
MPY	184.1319	48.1411	87.8556	72 (2.73%)
BTP	85.3565	14.3529	56.6507	51 (1.94%)
ATP	9.3899	0.7487	7.8925	28 (1.06%)
MTBH	30.6526	3.9500	22.7526	55 (2.09%)
MBA	6.0962	1.5579	2.9803	9 (0.34%)

NEG samples:

MPY	207.5395	50.0069	107.5258	2728 (3.51%)
BTP	88.8007	14.3476	60.1056	2837 (3.65%)
ATP	9.0127	0.6321	7.7484	2001 (2.57%)
MTBH	31.7907	3.3072	25.1763	1600 (2.06%)
MBA	7.2565	1.9345	3.3876	772 (0.99%)

Table 5-2. Summary of removed samples from data cleaning.

POS samples:

Before	After	Total rejected	Overlap
2633	2460	173 (6.57%)	43 (1.63%)
NEG samples:			
77761	70547	7165 (9.21%)	2773 (3.57%)

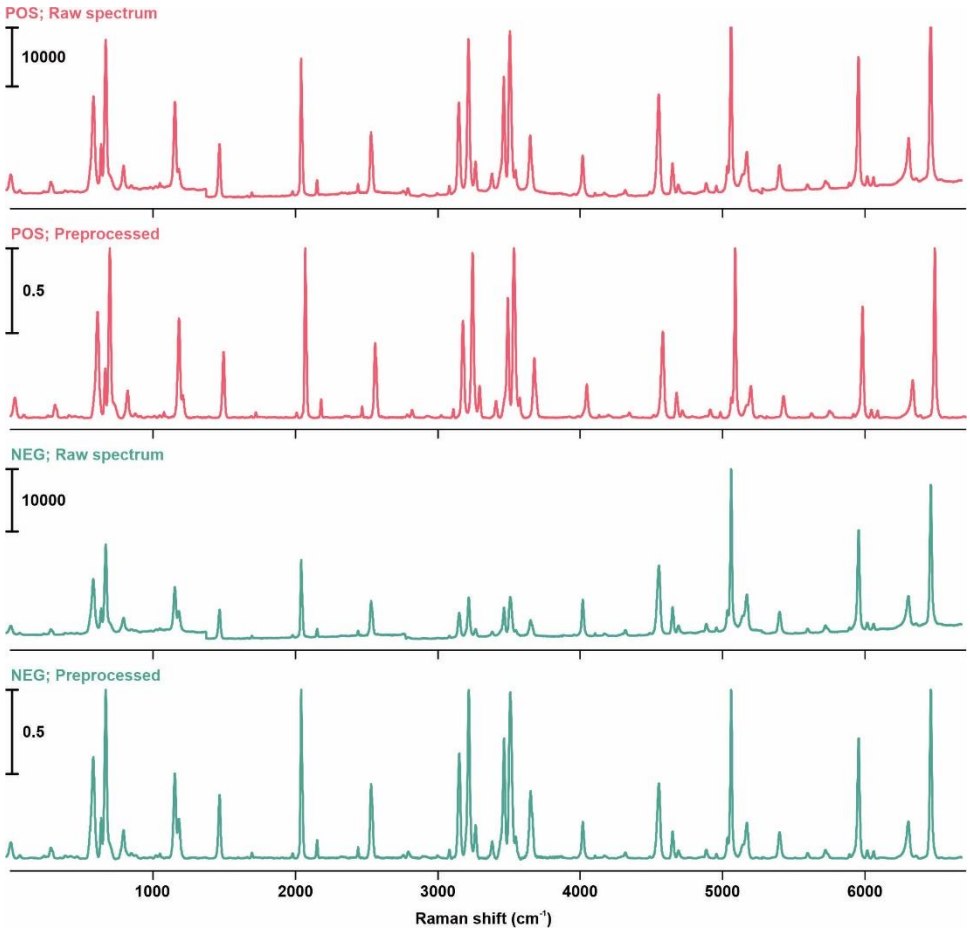


Figure 5-3. Representative POS and NEG super-profiles before and after baseline correction and normalization.

Subsequently, we adopt a feature engineering approach to parameterize SERS super-profiles based on their distinct receptor SERS peaks (**Figure 5-4**). We first assign Raman

vibrational modes to the receptor SERS peaks and select major peaks which have been identified to influence the classification COVID breath profiles based on in-depth chemical investigations in our previous clinical study. These peaks are fitted using a pseudo-Voigt fitting function, which is a convolution of the Gaussian and Lorentzian functions, that is suitable for most spectral peaks. After fitting, we compute five peak parameters – (1) position, (2) intensity, (3) FWHM, (4) fraction, and (5) skew – which collectively describe a SERS peak in terms of its position and shape (**Figure 5-5**).

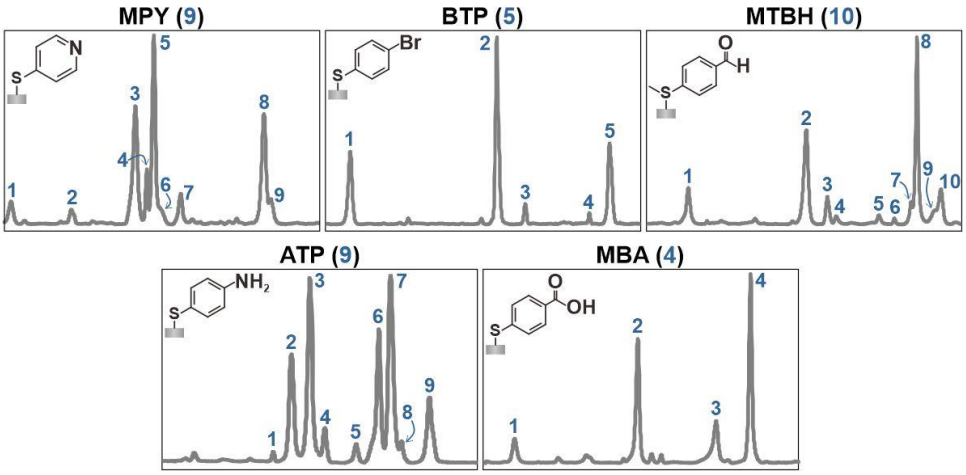


Figure 5-4. Selected peaks from each receptor.

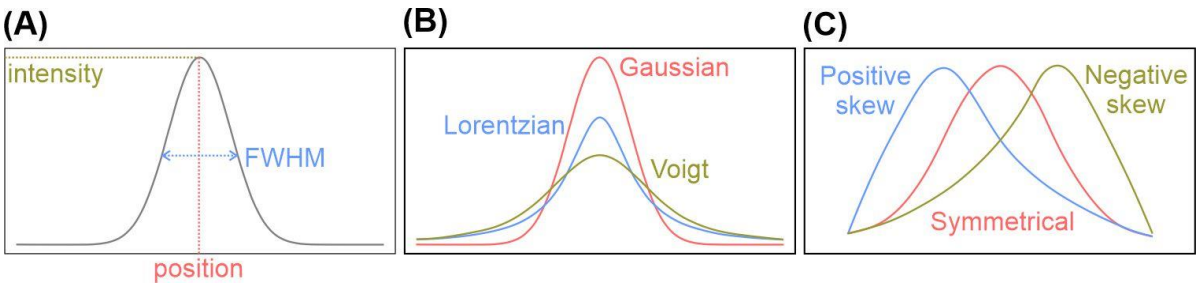


Figure 5-5. Graphical illustration of each peak parameter. (A) Peak position (x-value), intensity (y-value), FWHM. (B) Gaussian, Lorentzian, and Voigt distributions. (C) Symmetrical distributions and distributions with positive and negative skew.

The motivation for feature engineering is multifold. First, converting SERS super-profiles which comprise > 6700 features into 185 peak parameters (37 peaks × 5 parameters) will significantly lower the data dimensionality. For complex neural network ML models such as MLP, this will drastically shorten computational time, which is critical when processing big data. In addition, it also removes unwanted background noise and uninformative features that do not contribute to the classification of breath profiles. Finally, the peak parameters provide more chemically meaningful analysis as they directly correspond to the changes in Raman vibrational modes by reflecting peak shifts, intensity fluctuations, and/or shape differences when receptor-BVOC interactions occur at the molecular level.

5.2.2 Resolving the class imbalance problem

In an infectious disease outbreak, mass screening exercises are typically conducted at places where large groups gather to sieve out individuals who may be infected and isolate them. There are three key characteristics of mass screenings. First, screening tests are not designed to accurately diagnose a person of the disease infection. Rather, they are designed to identify both positive and suspicious/borderline cases, such as in very early or late stages of infection, so that the person can undergo an affirmative diagnostic test. In the case of COVID-19, hand-held SERS-based breathalyzers can be flexibly deployed at airports, large-scale events, and even at workplaces for routine mass screening. They are simple to administer, non-invasive, and can provide results on-site within five minutes, causing minimal disruptions to everyone involved. When a positive test result is received, the person can then undergo a nasopharyngeal swab to obtain a sample for the PCR test which determines the actual diagnosis. Meanwhile, this person should be isolated to prevent the potential spread of disease. Mass screening exercises have been proved to be an effected non-pharmaceutical intervention to contain infectious diseases and protect human lives.[1-3]

Next, the inherent nature of mass screening exercises involve an overwhelming majority of individuals who are most likely feeling well and do not show symptoms of the disease. This is unlike healthcare institutions, where diagnostic, rather than screening tests, are conducted as people who visit are generally unwell in the first place and hence require a confirmatory diagnosis. With no intervention, ML models trained with such significant class imbalances will place equal importance in correctly classifying POS and NEG classes which maximizes accuracy. In extreme cases, the ML model may optimize towards a local maxima where all samples are predicted as the majority class, losing the purpose of classification.[11] However, the cost of misclassifying POS cases is very high in the context of infectious disease outbreaks, as one person can easily transmit a contagious disease to more than one person.

Finally, data collection is a significant roadblock to the construction of robust ML models. During infectious disease outbreaks, it is difficult to obtain access to POS samples from healthcare institutions as stringent protocols must be established to contain the disease. Furthermore, admitted patients are generally feeling discomfort from disease symptoms and may not be medically fit to provide samples. A conundrum is hence formed – on one hand, we are provided with very few POS samples to train ML models, but on the other hand, we need to establish a balanced and robust ML model representative of the population for application in mass screenings.

Considering these concerns, we envisioned that a data-driven approach would resolve class imbalance without incurring too much time and resources to collect POS samples. Specifically, an ensemble approach which combines the under-sampling of NEG samples and over-sampling of POS samples will restore balance in the classes (**Figure 5-6**). Sampling techniques have been widely used in ML to resolve unequal distribution in classes to improve ML training on such datasets.[13-14]

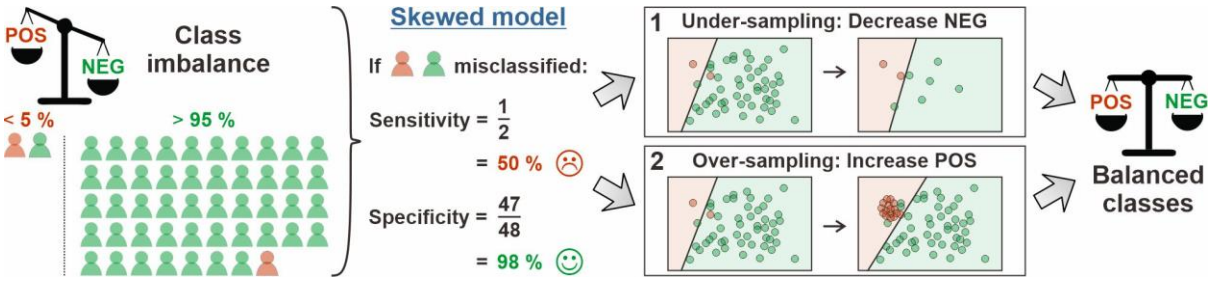


Figure 5-6. Class imbalance in the context of mass screenings. The cost of misclassifying POS samples is reflected as a drastic decrease in model sensitivity. An ensemble approach which combines under- and over-sampling is proposed to restore balance between the classes.

The concept of random under-sampling is simple – a subset of NEG samples is selected from the pool with no replacement. However, such randomness reduces the reproducibility of each iteration and creates significant unwanted variations. ENN is an alternative approach that systematically removes NEG samples which obscure the boundary separating POS and NEG, thereby creating a clearer distinction between the classes (**Figure 5-7**). The algorithm works by identifying k nearest neighbors from an observation and finding the majority class of the neighbors. If the observation and the majority class is different, then both the observation and the k neighbors are deleted from the dataset.

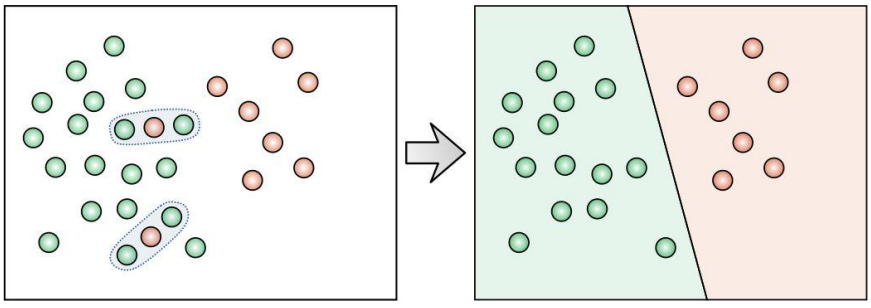


Figure 5-7. ENN approach to under-sample the majority class (green), by removing samples that obscure the boundary separating the two classes.

Similarly, random over-sampling simply involves selecting from currently available POS samples with replacement. While this is effective in inflating the absolute number of POS samples, it does not accurately reflect the class variance associated. On the other hand, SMOTE is an approach that generates synthetic samples by interpolation which preserves and reflects the class variance that would otherwise be obtained experimentally (**Figure 5-8**). SMOTE selects an observation from the minority class and creates a new observation by computing the difference in feature values with its nearest neighbor and multiplying the resultant by a random number between 0 to 1.[12] This creates a diverse set of synthetic observations. Notably, we affirm that these synthetic data have distributions which confirm to the original data, using the MBA receptor as example (**Figure 5-9**).

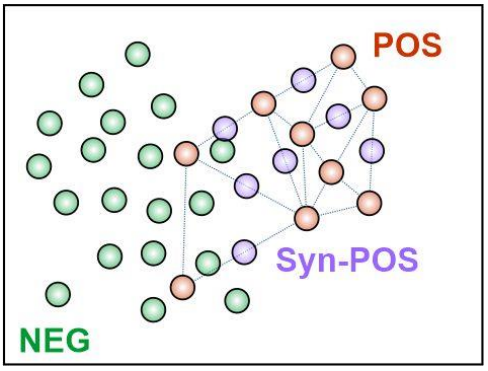


Figure 5-8. SMOTE creates synthetic POS samples by interpolation.

To systematically assess the impact of (1) under-sampling using ENN, (2) over-sampling using SMOTE, and (3) ensemble augmentation using SMOTEENN, we established multiple MLP models to classify POS and NEG samples. MLP is a feed-forward artificial neural network comprising fully connected layers with a non-linear activation function that is trained using the backpropagation method.[16-17] As opposed to traditional ML models, MLP effectively models the complexity involved in differentiating subtle class differences, as shown in various studies.[18-19]

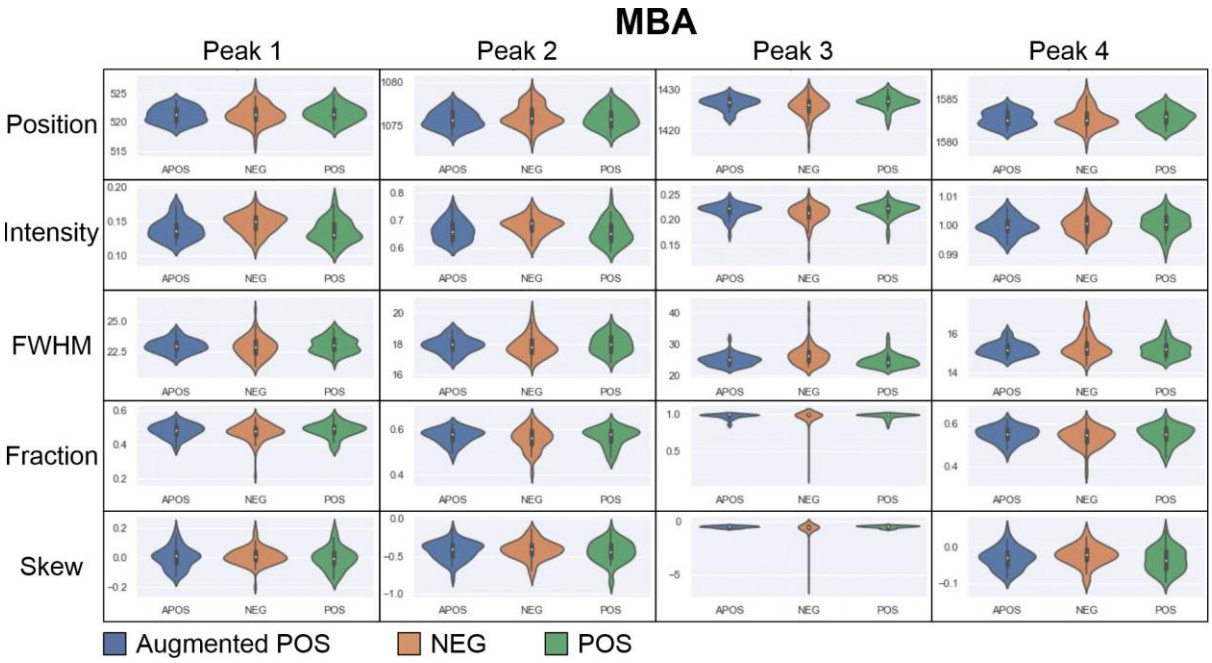


Figure 5-9. Violin plots of the peak parameter distributions for each of the 4 MBA receptor peaks. The augmented POS peak parameters follow the distribution of the POS samples as expected from the SMOTE algorithm and is contrasted with the NEG samples.

We assess the performance of the model using three metrics – sensitivity, PPV, and Matthews' correlation coefficient (MCC) as calculated by:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{P}}$$

$$\text{Positive predictive value} = \frac{\text{TP}}{\text{PP}}$$

$$\text{Matthews}' \text{ correlation coefficient} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

where P is the number of actual positive samples, PP is the number of samples predicted as positive, TP is the number of predicted positive samples which are indeed positive (hit), TN is the number of predicted negative samples which are indeed negative (correct rejection), FP is the number of predicted positive samples which are negative (false alarm), and FN is the number of predicted negative samples which are positive (miss).

Here, MCC is used as it is a better measure of the quality of binary classification than other metrics like accuracy and F1 score and performs well even when the classes are severely imbalanced.[20-22] We first fix the number of POS samples and investigate the inclusion of an increasing number of NEG samples (**Figure 5-10**). Without under-sampling, the MLP model attains a sensitivity of 56.0%, PPV of 62.4%, and a MCC of 53.5% (**Table 5-3**). With under-sampling alone, the MLP model attains a sensitivity of 77.5% (+ 21.5%), PPV of 89.1% (+26.7%), and a MCC of 75.8% (+ 22.3%). This occurs at a POS-to-NEG ratio of 0.492, and a NEG sample number of 5000. The drastic improvement in scoring metrics clearly indicate the detrimental effects of severe class imbalance in influencing the ML model performance. However, we note that under-sampling alone is inefficient, as we significantly under-utilize the NEG samples available and will still be trapped in the bottleneck that is collecting sufficient POS samples.

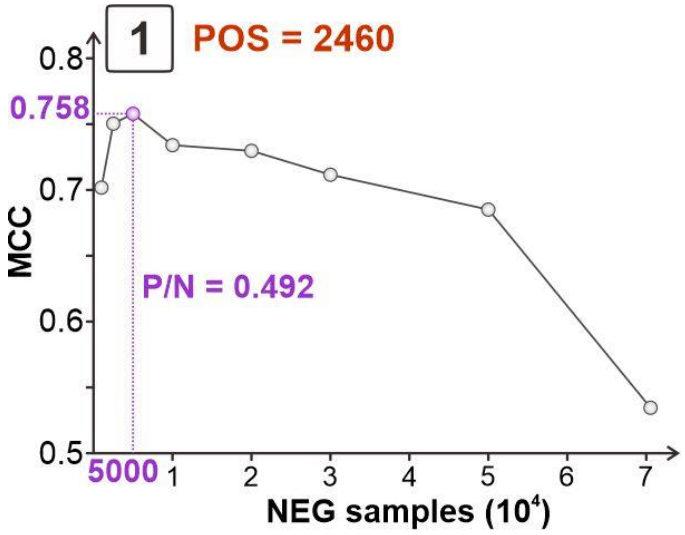


Figure 5-10. Investigating the effect of under-sampling on our MLP model by fixing the number of POS samples.

Table 5-3. Results of under-sampling using our MLP model.

POS	NEG	Ratio	Sensitivity	PPV	MCC
2460	70547	0.035	0.5602	0.6244	0.5345
2460	50000	0.049	0.5789	0.8418	0.6851
2460	30000	0.082	0.6285	0.8515	0.7116
2460	20000	0.123	0.6569	0.8690	0.7297
2460	10000	0.246	0.6951	0.8776	0.7340
2460	5000	0.492	0.7752	0.8906	0.7579
2460	2500	0.984	0.8228	0.9093	0.7504
2460	1000	2.460	0.8724	0.9291	0.7018

Next, we fix the number of NEG samples and investigate the inclusion of an increasing number of POS samples by synthetically generating POS data using SMOTE (**Figure 5-11**). With SMOTE, the MLP model attains a sensitivity of 98.0% (+ 42.0%), PPV of 87.9% (+25.5%), and a MCC of 89.0% (+ 35.5%) (**Table 5-4**). This occurs at a POS-to-NEG ratio of 0.389, and a POS sample number of 27423. Like under-sampling, over-sampling with SMOTE also caused drastic improvements in the MLP model as it alleviates the class imbalance. Crucially, we note that while SMOTE augments POS data effectively, it does not account for neighboring NEG data (**Figure 5-8**). Hence, we envision that an ensemble SMOTEENN method would harness the ability to better distinguish between the POS and NEG classes, while creating representative synthetic POS samples.

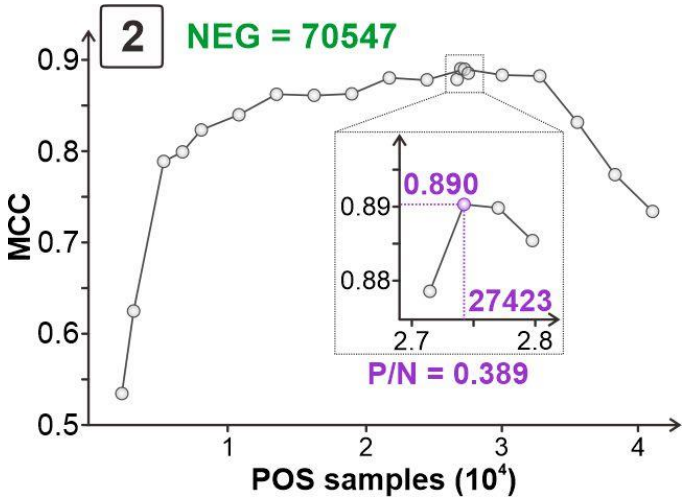


Figure 5-11. Investigating the effect of over-sampling on our MLP model by fixing the number of NEG samples.

Table 5-4. Results of over-sampling using our MLP model.

POS	NEG	Ratio	Sensitivity	PPV	MCC
2460	70547	0.0349	0.5602	0.6244	0.5345
3324	70547	0.0471	0.7114	0.6477	0.6249
5540	70547	0.0785	0.8964	0.7494	0.7887
6925	70547	0.0982	0.9155	0.7598	0.7992
8310	70547	0.1178	0.9406	0.7769	0.8233
11080	70547	0.1571	0.9441	0.8097	0.8398
13850	70547	0.1963	0.9635	0.8333	0.8623
16620	70547	0.2356	0.9653	0.8343	0.8612
19390	70547	0.2749	0.9516	0.8529	0.8628
22160	70547	0.3141	0.9639	0.8709	0.8802
24930	70547	0.3534	0.9632	0.8725	0.8782
27146	70547	0.3848	0.9601	0.8785	0.8786

27423	70547	0.3887	0.9797	0.8786	0.8903
27700	70547	0.3926	0.9617	0.8928	0.8898
27977	70547	0.3966	0.9693	0.8815	0.8854
30470	70547	0.4319	0.9618	0.8890	0.8834
33240	70547	0.4712	0.9578	0.8945	0.8823
36010	70547	0.5104	0.9165	0.8700	0.8316
38780	70547	0.5497	0.8634	0.8643	0.7742
41550	70547	0.5890	0.8607	0.8187	0.7340

Finally, with an ensemble SMOTEENN approach, the MLP model attains a sensitivity of 99.8% (+ 43.8%), PPV of 93.8% (+ 31.4%), and a MCC of 94.0% (+ 40.5%) (**Figure 5-12, Table 5-5**). This occurs at a POS-to-NEG ratio of 0.764, and a POS sample number of 49028. Interestingly, we note that with SMOTEENN, the number of POS samples after augmentation increased by 1.8-fold and in turn increased the optimal POS-to-NEG sample ratio. This means that the synthetic samples generated by SMOTE becomes more effective with the removal of NEG samples that is at the border separating the two classes. Intuitively, it is inherently challenging to pick up subtle differences in breath profiles of POS and NEG individuals even with an analytical technique like SERS that can provide molecular fingerprinting. If too many POS samples lie near the border separating the two classes, then SMOTE alone may blur the line between the two classes when interpolating. Hence, SMOTEENN is a superior approach.

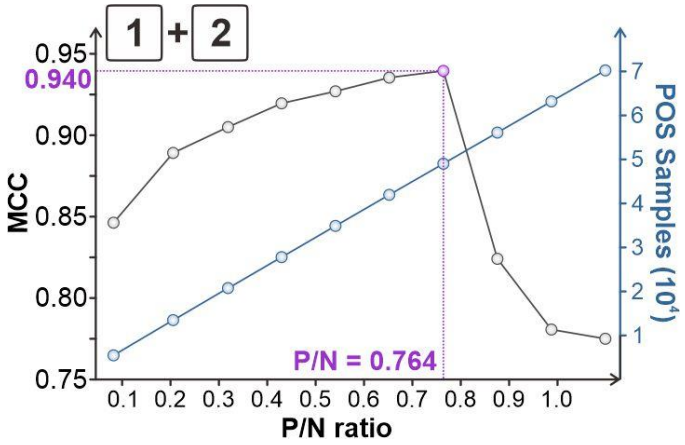


Figure 5-12. Investigating the effect of an ensemble under- and over-sampling method using SMOTEENN on our MLP model with varying POS-to-NEG sample ratio.

Table 5-5. Results of ensemble under- and over-sampling using our MLP model.

POS	NEG	Ratio	Sensitivity	PPV	MCC
5484	66752	0.082	0.9867	0.7783	0.8463
13493	65691	0.205	0.9934	0.8475	0.8891
20770	65156	0.319	0.9961	0.8759	0.9050
27798	64705	0.430	0.9979	0.9006	0.9196
34896	64506	0.541	0.9987	0.9152	0.9270
41958	64338	0.652	0.9986	0.9290	0.9354
49028	64167	0.764	0.9982	0.9382	0.9395
56103	64066	0.876	0.9260	0.8936	0.8241
63184	63990	0.987	0.8688	0.9069	0.7807
70236	63914	1.099	0.9052	0.8866	0.7750

5.3 Conclusion

In conclusion, we proposed an ensemble data augmentation strategy that harnesses the advantages of both over- and under-sampling to overcome the inherently large class imbalance problem associated with mass screening applications. Our SMOTEENN approach improved the MLP model significantly by amplifying POS samples to 20-times its original and reducing NEG samples slightly by 9%, achieving 99.8% sensitivity and 93.8% PPV in classifying COVID-positive individuals across > 73000 exhaled breath samples. Importantly, we conduct data cleaning, pre-processing, and feature engineering prior to ML model construction to reduce data dimensionality, simplify the ML model complexity, and shorten the subsequent computational time required. Our findings is crucial in bolstering the applicability of hand-held SERS breathalyzers in mass screening applications.

5.4 Materials and methods

Chemicals. Silver nitrate, 1,5-pentanediol (PD), poly(vinylpyrrolidone) (PVP; Mw ~55,000), 4-mercatopyridine (MPY), 4-mercaptobenzoic acid (MBA), 4-aminothiophenol (ATP), 4-bromothiophenol (BTP), and 4-methylthiobenzaldehyde (MTBH) were purchased from Sigma Aldrich. Copper (II) chloride was purchased from Alfa Aesar. Ethanol (ACS, ISO, Reag. Ph Eur) was obtained from Merck. Milli-Q water (> 18.0 MΩ. cm) was purified with a Sartorius Arium® 611 UV ultrapure water system. All reagents were used without further purification.

Data collection and SERS measurements. The synthesis of Ag nanocubes, thiophenol functionalization, and sensor chip fabrication is in accordance with our previous clinical study, except for two additional receptors (MTBH, BTP).[4] The breathalyzer fabrication and assembly conditions are kept identical. The study was conducted in the same fashion at Changi International Airport and Bright Vision Hospital in Singapore, following protocols covered

under NTU's IRB-2020-12-012 and IRB-2021-03-046. SERS measurements were done using the portable Zolix Raman spectrometer, with an excitation wavelength of 785 nm, laser power of 50 mW, and acquisition time of 0.05 s. The spectral window of 400 to 1800 cm^{-1} was processed using airPLS baseline and min-max normalization.

Data cleaning and peak parameterization. The raw SERS spectra were first cleaned using a custom Python script. We first established the spectral regions that constitute the signals and noise respectively. For MPY, the signal regions comprise 950 – 1120, 1180 – 1250, and 1540 – 1650 cm^{-1}, while the noise regions comprise 550 – 600, and 1650 – 1750 cm^{-1}. For BTP, the signal regions comprise 450 – 550, 970 – 1200, and 1450 – 1600 cm^{-1}, while the noise regions comprise 551 – 650, and 800 – 900 cm^{-1}. For MTBH, the signal regions comprise 500 – 580, 1030 – 1260, and 1530 – 1730 cm^{-1}, while the noise regions comprise 890 – 990, and 1260 – 1380 cm^{-1}. For ATP, the signal regions comprise 1050 – 1220, 1350 – 1500, and 1530 – 1620 cm^{-1}, while the noise regions comprise 750 – 800, and 1650 – 1775 cm^{-1}. For MBA, the signal regions comprise 490 – 570, 1030 – 1120, 1350 – 1470, and 1550 – 1650 cm^{-1}, while the noise regions comprise 635 – 700, and 900 – 1000 cm^{-1}. The identified SERS peaks were fitted using a pseudo-Voigt fitting function with non-linear least-squares minimization obtained from the LMfit Python library.

ML model construction. The MLP models were constructed using the Python-based scikit-learn package.[23] The SMOTE, ENN, and SMOTEENN sampling algorithms were adopted from the Imbalanced-learn toolbox.[24] In one iteration, the POS and NEG data were split into 10 folds using random state = 1, where 9 folds were used to train the MLP model and 1-fold was used to test it. This process was repeated so that each fold became the test set at least once. The results were averaged to obtain the sensitivity, PPV, and MCC scores for the model.

References

1. Johanna, N.; Citrawijaya, H.; Wangge, G., *J. Public Health Res.*, 2020, 9, 4.
2. Girum, T.; Lentiro, K.; Geremew, M.; Migora, B.; Shewamare, S., *Trop. Med. Health*, 2020, 48, 91.
3. Feng, Z.; Zhang, Y.; Pan, Y.; Zhang, D.; Zhang, L.; Wang, Q., *Med. Rev.*, 2022, 2, 197-212.
4. Leong, S. X.; Leong, Y. X.; Tan, E. X.; Sim, H. Y. F.; Koh, C. S. L.; Lee, Y. H.; Chong, C.; Ng, L. S.; Chen, J. R. T.; Pang, D. W. C.; Nguyen, L. B. T.; Boong, S. K.; Han, X.; Kao, Y.-C.; Chua, Y. H.; Phan-Quang, G. C.; Phang, I. Y.; Lee, H. K.; Mohammad, Y. A.; Tan, N. S.; Ling, X. Y., *ACS Nano*, 2022, 16, 2629-2639.
5. Chen, H. ; Qi, X. ; Ma, J. ; Zhang, C. ; Feng, H. ; Yao, M., *medRxiv*, 2020, 2020.06.21.20136523.
6. Grassin-Delyle, S.; Roquencourt, C.; Moine, P.; Saffroy, G.; Carn, S.; Heming, N.; Fleuriet, J.; Salvator, H.; Naline, E.; Couderc, L.-J.; Devillier, P.; Thévenot, E. A.; Annane, D., *EbioMedicine*, 2021, 63, 103154.
7. Ruszkiewicz, D. M.; Sanders, D.; O'Brien, R.; Hempel, F.; Reed, M. J.; Riepe, A. C.; Bailie, K.; Brodrick, E.; Darnley, K.; Ellerkmann, R.; Mueller, O.; Skarysz, A.; Truss, M.; Wortelmann, T.; Yordanov, S.; Thomas, C. L. P.; Schaaf, B.; Eddleston, M., *EClinicalMedicine*, 2020, 29-30.
8. Leong, Y. X.; Tan, E. X.; Leong, S. X.; Koh, C. S. L.; Nguyen, L. B. T.; Chen, J. R. T.; Xia, K.; Ling, X. Y., *ACS Nano*, 16, 13279-13293.
9. Japkowicz, N.; Stephen, S., *Intell. Data Anal.*, 2002, 6, 429-449.
10. Mazurowski, M. A.; Habas, P. A.; Zurada, J. M.; Lo, J. Y.; Baker, J. A.; Tourassi, G. D., *Adv. Neural Netw. Res.*, 2008, 21, 427-436.
11. Leevy, J. L.; Khoshgoftaar, T. M.; Bauder, R. A.; Seliya, N., *J. Big Data*, 2018, 5, 42.

12. Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P., *J. Artif. Intell. Res.*, 2002, 16, 321-357.
13. Batista, G. E. A. P. A.; Prati, R. C.; Monard, M. C., *SIGKDD Explor. Newsl.*, 2004, 6, 20-29.
14. Mohammed, R.; Rawashdeh, J.; Abdullah, M., *2020 11th International Conference on Information and Communication Systems (ICICS)*, 2020, 243-248.
15. Wilson, D. L., *IEEE Transactions on Systems, Man, and Cybernetics*, 1972, 3, 408-421.
16. Baum, E. B., *J. Complex.*, 1988, 4, 193-215.
17. Gardner, M. W.; Dorling, S. R., *Atmos. Envrion.*, 1998, 32, 2627-2636.
18. Han, J.-K.; Oh, J.; Yun, G.-J.; Yoo, D.; Kim, M.-S.; Yu, J.-M.; Choi, S.-Y.; Choi, Y.-K., *Sci. Adv.*, 2021, 7, eabg8836.
19. Park, K.; Yuk, H.; Yang, M.; Cho, J.; Lee, H.; Kim, J., *Sci. Robot.*, 2022, 7, eabm7187.
20. Matthews, B. W., *Biochim. Biophys. Acta, Prot.*, 1975, 405, 442-451.
21. Chicco, D.; Jurman, G., *BMC Genomics*, 2020, 21, 6.
22. Chicco, D.; Tötsch, N.; Jurman, G., *BioData Mining*, 2021, 14, 13.
23. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R., *J Mach. Learn. Res.*, 2011, 12, 2825-2830.
24. Lemaˆıtre, G.; Nogueira, F.; Aridas, C. K., *J. Mach. Learn. Res.*, 2017, 18, 1-5.

Chapter 6 Conclusion: The future of SERS with data science

6.1 Overall summary

Generic molecular receptors that promote non-specific receptor-analyte interaction allow sensing of analytes with inherently small Raman cross-sections and mitigate concerns with interfering platform SERS signals and strict receptor-analyte compatibility. However, relying on a single receptor is impractical as they are often unable to distinguish analytes with high structural similarity. Recently, data science strategies have emerged as a powerful tool to complement many analytical techniques in the nanoscale sensing of disease biomarkers. Considering its proliferation and success, this thesis primarily aims to explore the integration of data science strategies with SERS in overcoming longstanding bottlenecks with its translation towards practical applications.

In chapter 2, I first introduced a multi-receptor SERS Taster platform incorporating an array of four functionalized molecular receptors with both direct and indirect sensing capabilities. Adopting a data science approach, we concatenated all input features originating from the four receptor SERS spectra into a single SERS super-profile. The super-profile effectively amalgamates individual receptor SERS variances, which in turn forms a more complete spectroscopic profile of a target analyte, imbuing enhanced specificity. In the classification and multiplex quantification of five flavor compounds, the super-profile attained near-perfect accuracies in contrast to merely 33% with individual receptors. In addition, our ML model outcome can be closely correlated to our chemical domain knowledge of molecular-level interactions using in-depth analysis of the PCA biplot. In doing so, we also demonstrated the ability of our SERS Taster in distinguishing primary, secondary and tertiary alcohols. Our SERS Taster reinforces the potential of receptor-driven array-based SERS sensing in providing enhanced analyte specificities despite leveraging non-specific interaction modalities through a unique data science perspective.

In chapter 3, I addressed concerns with receptor selection for such multi-receptor SERS platforms. Currently, selection is conducted based on chemical intuition or manual trial-and-error experimentation. We proposed a ML-driven SERS-based receptor RS that objectively selects receptors to construct an optimized SERS super-profile using a four-stage 'identify, filter, rank and recommend' approach. Selecting the right receptor combination is important because (1) excess receptors will cause a decrease in ML performance due to the CoD effect, and (2) receptor-analyte interaction type and strength can vary significantly, hence should be tailored to the analytes concerned. Our SERS-based receptor RS harnesses chemical domain knowledge to selectively retain key receptor peak variances while removing background and uninformative features, thereby maximizing SERS variances, and minimizing the CoD effects. This results in a 25.7% improvement in classification accuracy from a single receptor to an optimized 6-receptor super-profile. Crucially, we highlight the ability to extrapolate receptor recommendations to unseen problems by leveraging a kNN-based collaborative filtering approach, even before analyzing spectra associated with the unknown. Our RS is key in paving the way for our multi-receptor SERS sensors to achieve precise molecular differentiation through optimizing receptor combinations.

In chapter 4, I showcased the integration of our multi-receptor SERS sensor within a hand-held breathalyzer for rapid, non-invasive screening of individuals for COVID-19 using their exhaled breath. We established good qualitative agreement between our observed SERS spectral changes and those caused by reported potential COVID-19 breath biomarkers, affirming the basis for BVOC profile differentiation. Crucially, our multi-receptor SERS sensor attained a 96.2% sensitivity and 99.9% specificity in a comparative case-control clinical trial involving 501 participants, which is superior to commercially available antigen tests and comparable to PCR tests. This outcome is independent of displayed COVID-19 symptoms and other confounding factors such as age, gender, and smoking habits. Our breathalyzer test is

simple and easy to administer, requiring only 5 min from sample collection to result output. Albeit small scale, these results underscore the potential of multi-receptor SERS sensors in mass screening applications for future epidemics/pandemics and other disease diagnostics involving breath BVOC as biomarkers.

In chapter 5, I proposed an ensemble data augmentation strategy that combines under- and over-sampling to overcome the huge class imbalance commonly associated with mass screening applications where healthy people form the overwhelming majority. With a balanced ML model, we achieved 99.8% sensitivity and 93.8% PPV in classifying COVID-positive individuals across > 73000 exhaled breath samples. We highlighted the importance of parameterizing SERS peaks to reduce data dimensionality and simplifying the ML model complexity, which in turn shortens the computation time significantly. Using our synthetic minority over-sampling technique coupled edited nearest neighbors' algorithm (SMOTEENN), we illustrate 43.8% increase in sensitivity and 31.4% increase in PPV when POS samples are amplified by 20-times their original and NEG samples are reduced by 9%. Our findings will be crucial in realizing the application of hand-held SERS breathalyzers for mass screenings during future infectious disease outbreaks.

6.2 Outlook

In this thesis, I conducted several investigations where I harnessed data science to drive the application of multi-receptor SERS sensors in practice. Moving forward, I postulate that data science strategies will continue to impact nanoscale SERS sensing of disease biomarkers and beyond. Specifically, this comprises five potential research directions – (1) revolutionizing nanoparticle discoveries, (2) strengthening result-to-knowledge relationships, (3) extrapolating potential disease correlations, (4) harnessing advanced ML algorithms, and (5) constructing massive open-source data repositories (**Figure 6-1**).

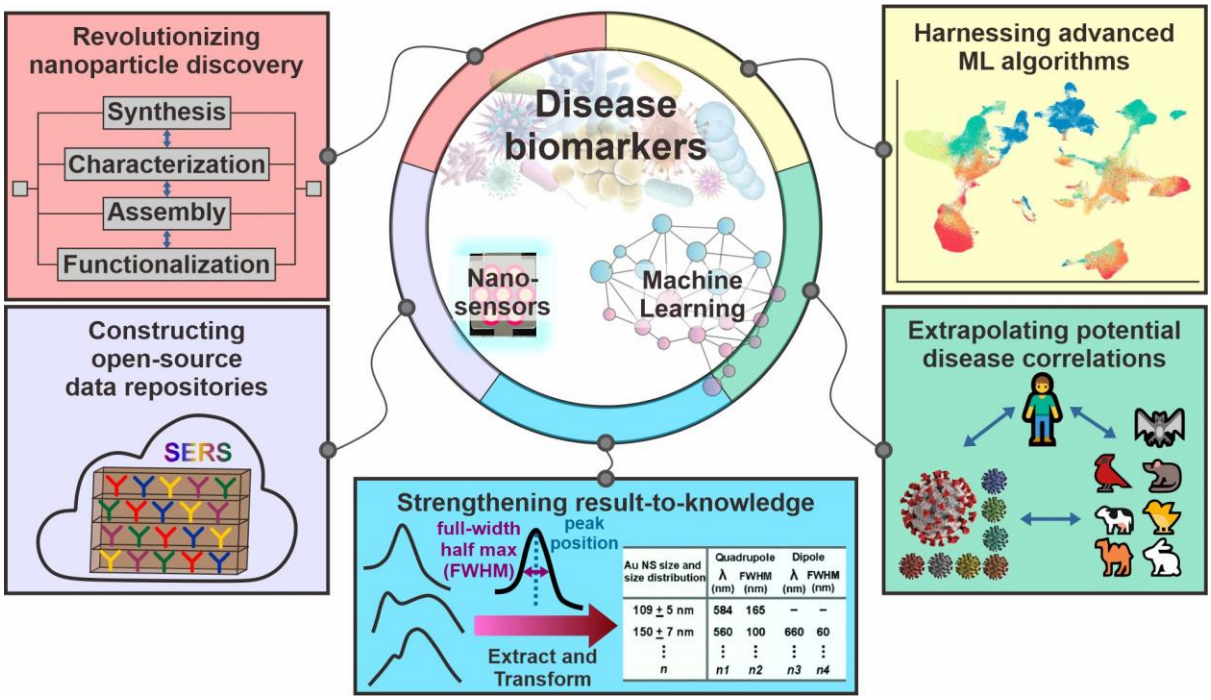


Figure 6-1. Outlook on how data science will further bolster efforts in developing SERS-based sensors for disease biomarker detection and beyond.

6.2.1 Revolutionizing nanoparticle discoveries

Plasmonic nanoparticles are the building blocks of every SERS sensor as they provide the medium to confine plasmons at the nanoscale, giving rise to SERS enhancements. Hence, it is always beneficial to continuously discover new plasmonic nanoparticles that can attain stronger SERS enhancements. However, this discovery process often incur huge costs in materials, labor, and time. There is also significant error involved when factoring in the high variance associated with manually performing synthesis experiments in practice. With state-of-the-art robotics and data science strategies, these costs can be drastically reduced. For example, a robotic arm can be programmed to conduct automatic reagent addition at specific time intervals according to experimental protocol.[1-6] If sufficient precursors are provided, the platform can run tirelessly throughout the day, minimizing labor costs, and maximizing time efficiency. Evolutionary ML algorithms are ideal for such a discovery process, as they offer

feedback after each experiment cycle which swiftly optimizes towards a desired target property (such as LSPR peak position or width).[7-9] In doing so, material costs are also reduced as the ML algorithms filter unfeasible precursor combinations. Additionally, parallelized ML algorithms can link segmented robotic modules for concurrent nanoparticle synthesis, characterization, assembly, and/or functionalization, thereby creating a one-stop autonomous workflow. There is significant potential for future research in this area towards the discovery of enhanced SERS sensors as well as their large-scale production with high reproducibility.

6.2.2 Strengthening result-to-knowledge relationships

With increasingly sophisticated sample matrices and the need to differentiate analytes with similar molecular structures, SERS variances are often subtle and embedded within a complex spectrum. While ML algorithms may be able to identify these variances, they can be inherently difficult to interpret. As such, these models commonly termed as 'black boxes', since the user cannot completely comprehend how the model arrives at these results. In chapters 2 and 3, we applied model-agnostic methods such as PCA loadings and feature importance analysis which can anchor ML outputs to molecular-level interactions based on established scientific reasoning. In chapter 5, we leveraged domain knowledge to guide feature engineering and transformed raw data into structured and meaningful inputs such as SERS peak attributes to construct generalizable and accurate ML models. The transformation of the raw spectroscopic data into key peak features simultaneously eliminates noise and increases data interpretability which enables the algorithms to efficiently detect patterns within. Notably, such parameterization also reduces computational cost and complexity.[10] Model interpretability is a crucial area in SERS-based sensing as it is the only avenue to determine if the ML predictive models should be relied upon. It is pertinent to further explore alternative methods to interpret ML models and/or devise inherently interpretable ML models.

6.2.3 Extrapolating potential disease correlations

In the context of infectious disease detection discussed in chapters 4 and 5, it is difficult to anticipate how an unknown disease would manifest in a future outbreak. Since current sensing strategies revolve around first acquiring clear biomarker target(s), this process can potentially be hastened if we can draw parallels to the knowledge we have about existing diseases. One such example is to utilize data science and ML to investigate the mechanism of inter-species antibody cross-reactivity between animal and human coronaviruses to anticipate any zoonotic spillovers of animal viruses to humans considering the COVID-19 pandemic. In one report, the authors screened about 12900 coronavirus-associated peptides using a gradient boosted tree model and found that single monoclonal antibodies were responsible for mediating the cross-reactivity extending from animal to human coronaviruses.[11] Such a ML model built to screen for multiple coronavirus antigens in parallel based on human serum antibody signatures is expected to be useful for diagnostic applications, particularly in the early stages of future epidemics/pandemics. In this aspect, ML algorithms can potentially be extended to narrow the scope of biomarker screening, thereby reducing the time taken to locate key biomarkers relevant to the detection of new diseases.

6.2.4 Harnessing advanced ML algorithms

At present, we have yet to harness the full suite of ML algorithms for SERS signal processing. For instance, other non-linear dimensionality reduction and visualization methods such as the uniform manifold approximation and projection (UMAP), t-distributed stochastic neighbor embedding (t-SNE) and isometric mapping (ISOMAP) are useful alternatives to conventional PCA.[12-14] To improve model interpretability, other techniques such as the local interpretable model-agnostic explanations (LIME), Shapley additive explanations (SHAP) and contrastive explanation method (CEM) can play crucial roles in avoiding potential pitfalls such

as model overfitting.[15-16] In addition, advanced natural language processing (NLP) tools such as bidirectional encoder representations from transformers (BERT) show significant promise in not only data mining tasks from available literature, but also 1D and 2D signal analysis problems.[17-18] Similarly, geometric deep learning methods such as graph neural networks have demonstrated superior performance in biomolecular data analysis, drug design and can be further used in molecular design and discovery. For example, a neural network-based model, AlphaFold, demonstrates highly accurate prediction of protein structures with atomic accuracies even when no similar structures is known.[19] Other diverse applications include computing complex RNA structures, and discovering hidden optical responses from specific electromagnetic nanostructures.[20-21] These advanced ML algorithms will spark a paradigm shift in the area of SERS signal processing, allowing SERS-based sensors to achieve unparalleled detection sensitivities and specificities even for elusive analytes.

6.2.5 Constructing massive open-source data repositories

To harness the full potential of advanced ML algorithms such as deep neural networks, there is a need to access large amounts of training data to ensure model robustness. In recent decades, the scientific community has collectively contributed to constructing massive, well-annotated data repositories with petabytes of data from various research disciplines, which have fueled many important discoveries.[22-25] The hallmark of such a data repository is the completion of the Human Genome Project, where the sequencing of the entire human genome allowed the creation of high resolution genetic maps and promoted genomic-scale technologies like high throughput oligonucleotide synthesis.[26] For SERS-based sensing, it is similarly plausible to set up these libraries to allow unrestricted access to experimental data, ranging from fundamental information about certain analytes to de-identified sample data from real applications. For example, a standardized SERS receptor library can be established to record a

standard spectrum under specific conditions along with all discovered vibrational mode changes associated with receptor-analyte interaction to streamline data aggregation and avoid duplication of effort. In the context of an epidemic/pandemic outbreak, the inclusion of data from real applications would significantly accelerate the development of a robust ML model by consolidating the efforts of multiple research teams. Data sharing via these open-source repositories would be key in accelerating the development of practical SERS-based sensors in diverse applications.

References

1. Li, J.; Ballmer, S. G.; Gillis, E. P.; Fujii, S.; Schmidt, M. J.; Palazzolo, A. M. E.; Lehmann, J. W.; Morehouse, G. F.; Burke, M. D., *Science*, 2015, 347, 1221-1226.
2. Dragone, V.; Sans, V.; Henson, A. B.; Granda, J. M.; Cronin, L., *Nat. Commun.*, 2017, 8, 15733.
3. Bédard, A. -C.; Adamo, A.; Aroh, K. C.; Russell, M. G.; Bedermann, A. A.; Torosian, J.; Yue, B.; Jensen, K. F.; Jamison, T. F., *Science*, 2018, 361, 1220-1225.
4. Granda, J. M.; Donina, L.; Dragone, V.; Long, D.-L; Cronin, L., *Nature*, 2018, 559, 377-381.
5. Steiner, S.; Wolf, J.; Glatzel, S.; Andreou, A.; Granda, J. M.; Keenan, G.; Hinkley, T.; Aragon-Camarasa, G.; Kitson, P. J.; Angelone, D.; Cronin, L., *Science*, 2018, 363, eaav2211.
6. Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai, Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; Rankin, N.; Harris, B.; Sprick, R. S.; Cooper, A. I., *Nature*, 2020, 583, 237-241.
7. Salley, D.; Keenan, G.; Grizou, J.; Sharma, A.; Martín, S.; Cronin, L., *Nat. Commun.*, 2020, 11, 2771.
8. Berardo, E.; Turcani, L.; Miklitz, M.; Jelfs, K. E., *Chem. Sci.*, 2018, 9, 8513-8527.
9. Jennings, P. C.; Lysgaard, S.; Hummelshøj, J. S.; Vegge, T.; Bligaard, T., *npj Comput. Mater.*, 2019, 5, 46.
10. Shiratori, K.; Bishop, L. D. C.; Ostovar, B.; Baiyasi, R.; Cai, Y.-Y.; Rossky, P. J.; Landes, C. F.; Link, S., *J. Phys. Chem. C*, 2021, 125, 19353-19361.
11. Klompus, S.; Leviatan, S.; Vogl, T.; Mazor, R. D.; Kalka, I. N.; Stoler-Barak, L.; Nathan, N.; Peres, A.; Moss, L.; Godneva, A.; Tikva, S. K. B.; Shinar, E.; Cohen-Dvashi, H.;

Gabizon, R.; London, N.; Diskin, R.; Yaari, G.; Weinberger, A.; Shulman, Z.; Segal, E., *Sci. Immun.*, 2021, 6, eabe9950.

12. Becht, E.; McInnes, L.; Healy, J.; Dutertre, C.-A.; Kwok, I. W. H.; Ng, L. G.; Ginhoux, F.; Newell, E. W., *Nat. Biotech.*, 2019, 37, 38-44.
13. Wu, N.; Zhang, X.-Y.; Xia, J.; Li, X.; Yang, T.; Wang, J.-H., *ACS Nano*, 2021, 15, 19522-19534.
14. Epps, R. W.; Bowen, M. S.; Volk, A. A.; Abdel-Latif, K.; Han, S.; Reyes, K. G.; Amassian, A.; Abolhasani, M., *Adv. Mater.*, 2020, 32, 2001626.
15. Chew, A. K.; Pedersen, J. A.; Van Lehn, R. C., *ACS Nano*, 2022, 16, 6282-6292.
16. Pilania, G., *Comput. Mater. Sci.*, 2021, 193, 110360.
17. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, 1, 4171-4186.
18. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V., *arXiv*, 2019, 1907.11692.
19. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D., *Nature*, 2021, 596, 583-589.
20. Townshend, R. J. L.; Eismann, S.; Watkins, A. M.; Rangan, R.; Karelina, M.; Das, R.; Dror, R. O., *Science*, 2021, 373, 1047-1051.
21. Kiarashinejad, Y.; Zandehshahvar, M.; Abdollahramezani, S.; Hemmatyar, O.; Pourabolghasem, R.; Adibi, A., *Adv. Intell. Syst.*, 2020, 2, 1900132.

22. Camacho, D. M.; Collins, K. M.; Powers, R. K.; Costello, J. C.; Collins, J. J., *Cell*, 2018, 173, 1581-1592.

23. Chang, K.; Creighton, C. J.; Davis, C.; Donehower, L.; Drummond, J.; Wheeler, D.; Ally, A.; Balasundaram, M.; Birol, I.; Butterfield, Y. S. N.; Chu, A.; Chuah, E.; Chun, H.-J. E.; Dhalla, N.; Guin, R.; Hirst, M.; Hirst, C.; Holt, R. A.; Jones, S. J. M.; Lee, D.; Li, H. I.; Marra, M. A.; Mayo, M.; Moore, R. A.; Mungall, A. J.; Robertson, A. G.; Schein, J. E.; Sipahimalani, P.; Tam, A.; Thiessen, N.; Varhol, R. J.; Beroukhim, R.; Bhatt, A. S.; Brooks, A. N.; Cherniack, A. D.; Freeman, S. S.; Gabriel, S. B.; Helman, E.; Jung, J.; Meyerson, M.; Ojesina, A. I.; Pedamallu, C. S.; Saksena, G.; Schumacher, S. E.; Tabak, B.; Zack, T.; Lander, E. S.; Bristow, C. A.; Hadjipanayis, A.; Haseley, P.; Kucherlapati, R.; Lee, S.; Lee, E.; Luquette, L. J.; Mahadeshwar, H. S.; Pantazi, A.; Parfenov, M.; Park, P. J.; Protopopov, A.; Ren, X.; Santoso, N.; Seidman, J.; Seth, S.; Song, X.; Tang, J.; Xi, R.; Xu, A. W.; Yang, L.; Zeng, D.; Auman, J. T.; Balu, S.; Buda, E.; Fan, C.; Hoadley, K. A.; Jones, C. D.; Meng, S.; Mieczkowski, P. A.; Parker, J. S.; Perou, C. M.; Roach, J.; Shi, Y.; Silva, G. O.; Tan, D.; Veluvolu, U.; Waring, S.; Wilkerson, M. D.; Wu, J.; Zhao, W.; Bodenheimer, T.; Hayes, D. N.; Hoyle, A. P.; Jeffreys, S. R.; Mose, L. E.; Simons, J. V.; Soloway, M. G.; Baylin, S. B.; Berman, B. P.; Bootwalla, M. S.; Danilova, L.; Herman, J. G.; Hinoue, T.; Laird, P. W.; Rhie, S. K.; Shen, H.; Triche, T.; Weisenberger, D. J.; Carter, S. L.; Cibulskis, K.; Chin, L.; Zhang, J.; Getz, G.; Sougnez, C.; Wang, M.; Saksena, G.; Carter, S. L.; Cibulskis, K.; Chin, L.; Zhang, J.; Getz, G.; Dinh, H.; Doddapaneni, H. V.; Gibbs, R.; Gunaratne, P.; Han, Y.; Kalra, D.; Kovar, C.; Lewis, L.; Morgan, M.; Morton, D.; Muzny, D.; Reid, J.; Xi, L.; Cho, J.; DiCara, D.; Frazer, S.; Gehlenborg, N.; Heiman, D. I.; Kim, J.; Lawrence, M. S.; Lin, P.; Liu, Y.; Noble, M. S.; Stojanov, P.; Voet, D.; Zhang, H.; Zou, L.; Stewart, C.; Bernard, B.; Bressler, R.; Eakin, A.; Iype, L.; Knijnenburg, T.; Kramer,

R.; Kreisberg, R.; Leinonen, K.; Lin, J.; Liu, Y.; Miller, M.; Reynolds, S. M.; Rovira, H.; Shmulevich, I.; Thorsson, V.; Yang, D.; Zhang, W.; Amin, S.; Wu, C.-J.; Wu, C.-C.; Akbani, R.; Aldape, K.; Baggerly, K. A.; Broom, B.; Casasent, T. D.; Cleland, J.; Creighton, C.; Dodda, D.; Edgerton, M.; Han, L.; Herbrich, S. M.; Ju, Z.; Kim, H.; Lerner, S.; Li, J.; Liang, H.; Liu, W.; Lorenzi, P. L.; Lu, Y.; Melott, J.; Mills, G. B.; Nguyen, L.; Su, X.; Verhaak, R.; Wang, W.; Weinstein, J. N.; Wong, A.; Yang, Y.; Yao, J.; Yao, R.; Yoshihara, K.; Yuan, Y.; Yung, A. K.; Zhang, N.; Zheng, S.; Ryan, M.; Kane, D. W.; Aksoy, B. A.; Ciriello, G.; Dresdner, G.; Gao, J.; Gross, B.; Jacobsen, A.; Kahles, A.; Ladanyi, M.; Lee, W.; Lehmann, K.-V.; Miller, M. L.; Ramirez, R.; Rätsch, G.; Reva, B.; Sander, C.; Schultz, N.; Senbabaoglu, Y.; Shen, R.; Sinha, R.; Sumer, S. O.; Sun, Y.; Taylor, B. S.; Weinhold, N.; Fei, S.; Spellman, P.; Benz, C.; Carlin, D.; Cline, M.; Craft, B.; Ellrott, K.; Goldman, M.; Haussler, D.; Ma, S.; Ng, S.; Paull, E.; Radenbaugh, A.; Salama, S.; Sokolov, A.; Stuart, J. M.; Swatloski, T.; Uzunangelov, V.; Waltman, P.; Yau, C.; Zhu, J.; Hamilton, S. R.; Getz, G.; Sougnez, C.; Abbott, S.; Abbott, R.; Dees, N. D.; Delehaunty, K.; Ding, L.; Dooling, D. J.; Eldred, J. M.; Fronick, C. C.; Fulton, R.; Fulton, L. L.; Kalicki-Veizer, J.; Kanchi, K.-L.; Kandoth, C.; Koboldt, D. C.; Larson, D. E.; Ley, T. J.; Lin, L.; Lu, C.; Magrini, V. J.; Mardis, E. R.; McLellan, M. D.; McMichael, J. F.; Miller, C. A.; O'Laughlin, M.; Pohl, C.; Schmidt, H.; Smith, S. M.; Walker, J.; Wallis, J. W.; Wendl, M. C.; Wilson, R. K.; Wylie, T.; Zhang, Q.; Burton, R.; Jensen, M. A.; Kahn, A.; Pihl, T.; Pot, D.; Wan, Y.; Levine, D. A.; Black, A. D.; Bowen, J.; The Cancer Genome Atlas Research, N.; Genome Characterization, C.; Genome Data Analysis, C.; Sequencing, C.; Data Coordinating, C.; Tissue Source, S.; Biospecimen Core Resource, C., *Nat. Genet.*, 2013, 45, 1113-1120.

24. Turnbaugh, P. J.; Ley, R. E.; Hamady, M.; Fraser-Liggett, C. M.; Knight, R.; Gordon, J. I., *Nature*, 2007, 449, 804-810.

25. Consortium, E. P., *Nature*, 2012, 489, 57-74.

26. Collins, F. S.; Morgan, M.; Patrinos, A., *Science*, 2003, 300, 286-290.