Biosensors and Bioelectronics 247 (2024) 115946

Contents lists available at ScienceDirect

Biosensors and Bioelectronics

journal homepage: www.elsevier.com/locate/bios

Machine learning-driven SERS fingerprinting of disintegrated viral components for rapid detection of SARS-CoV-2 in environmental dust

Aditya Garg [a], Seth Hawks [b], Jin Pan [c], Wei Wang [c], Nisha Duggal [b], Linsey C. Marr [c], Peter Vikesland [c,**], Wei Zhou [a,*]

[a] Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, 24061, United States
[b] Department of Biomedical Sciences and Pathobiology, Virginia Tech, Blacksburg, VA, 24061, United States
[c] Department of Civil and Environmental Engineering, Virginia Tech, Blacksburg, VA, 24061, United States

ARTICLE INFO

Keywords:
Surface-enhanced Raman spectroscopy
SARS-CoV-2
Rapid environmental virus monitoring
Machine learning

ABSTRACT

Surveillance of airborne viruses in crowded indoor spaces is crucial for managing outbreaks, as highlighted by the SARS-CoV-2 pandemic. However, the rapid and on-site detection of fast-mutating viruses, such as SARS-CoV-2, in complex environmental backgrounds remains challenging. Our study introduces a machine learning (ML)-driven surface-enhanced Raman spectroscopy (SERS) approach for detecting viruses within environmental dust matrices. By decomposing intact virions into individual structural components via a Raman-background-free lysis protocol and concentrating them into nanogap SERS hotspots, we significantly enhance the SERS signal intensity and fingerprint information density from viral structural components. Utilizing Principal Component Analysis (PCA), we establish a robust connection between the SERS data of these structural components and their biological sequences, laying a solid foundation for virus detection through SERS. Furthermore, we demonstrate reliable quantitative detection of SARS-CoV-2 using identified SARS-CoV-2 peaks at concentrations down to 10^2 pfu/ml through Gaussian Process Regression (GPR) and a digital SERS methodology. Finally, applying a Principal Component Analysis-Linear Discriminant Analysis (PCA-LDA) algorithm, we identify SARS-CoV-2, influenza A virus, and Zika virus within an environmental dust background with over 86% accuracy. Therefore, our ML-driven SERS approach holds promise for rapid environmental virus monitoring to manage future outbreaks.

1. Introduction

The recent SARS-CoV-2 pandemic has highlighted the challenge of controlling the airborne spread of viruses in poorly ventilated indoor spaces (Bazant and Bush, 2021; Morawska et al., 2020). Numerous indoor "super-spreading events" have led to large SARS-CoV-2 outbreaks (Miller et al., 2021; Shen et al., 2020). Consequently, there is an urgent need for a rapid, low-cost, and field-deployable analytical method to detect pathogenic viruses in aerosols or on surfaces within congested indoor environments to help prevent and control nascent viral epidemics as soon as possible (Rahmani et al., 2020; Wang et al., 2023; Yao et al., 2021). Furthermore, a flexible detection technique capable of identifying various mutant virus strains is essential, given the rapidly mutating nature of viruses such as SARS-CoV-2 (Su et al., 2016).

Methods for directly detecting viruses can be categorized as targeted and non-targeted approaches. Targeted methods are typically based on the detection of amplified viral nucleic acids or viral antigens (Abdelhamid and Badr, 2021; Peeling et al., 2022; Yüce et al., 2021). Nucleic acid amplification strategies such as polymerase chain reaction (PCR) have facilitated the highly sensitive detection of SARS-CoV-2 in environmental samples (Liu et al., 2020; Rahmani et al., 2020; Santarpia et al., 2020). However, they are unsuitable for rapid on-site monitoring due to complex handling, specialized equipment, and expensive reagents. Several antigen detection methods have been developed for SARS-CoV-2 detection, typically relying on receptors (e.g., nanobodies, antibodies) to capture the target antigens, followed by their detection via various transduction methods (e.g., electrical (Fathi-Hafshejani et al., 2021; Seo et al., 2020), electrochemical (Eissa and Zourob, 2020), optical (Pinals et al., 2021), plasmonic (Ahmadivand et al., 2020, 2021; Park et al., 2022)). Antigen tests have demonstrated rapid, low-cost, and on-site SARS-CoV-2 detection (Kevadiya et al., 2021), even in environmental samples (Puthussery et al., 2023). However, these antigen tests,

* Corresponding author.
** Corresponding author.
E-mail addresses: pvikes@vt.edu (P. Vikesland), wzh@vt.edu (W. Zhou).

https://doi.org/10.1016/j.bios.2023.115946
Received 8 September 2023; Received in revised form 27 November 2023; Accepted 19 December 2023
Available online 20 December 2023

relying on predefined receptors, are inadequate for detecting mutant viruses and suffer from reliability issues in complex backgrounds, such as environmental matrices, due to non-specific interactions (Yao et al., 2021). In contrast, non-targeted methods measure holistic molecular fingerprints of viruses without receptors, enabling the surveillance of mutant strains. Traditional non-targeted methods like mass spectrometry (Liangou et al., 2021; Nachtigall et al., 2020) and nuclear magnetic resonance (Bizkarguenaga et al., 2022) are unsuitable for rapid on-site viral surveillance due to pretreatment steps and expensive equipment. Raman spectroscopy, compatible with handheld instrumentation, holds promise for rapid on-site detection of SARS-CoV-2 (Pezzotti et al., 2022). However, the spontaneous Raman scattering process, due to a meager quantum yield, lacks the sensitivity to detect low virus concentrations in environmental samples.

To address these limitations, surface-enhanced Raman spectroscopy (SERS) combines vibrational spectroscopy's molecular fingerprint specificity with plasmonic nanostructures' hotspot sensitivity, offering an ultrasensitive fingerprinting-based detection method (Garg et al., 2022; Langer et al., 2019). Therefore, SERS has enabled the ultrasensitive, non-targeted detection of various biomolecules, from small metabolites to large proteins (Zong et al., 2018). Recently, numerous studies have demonstrated the successful detection of SARS-CoV-2 in human fluids (e.g., saliva) using non-targeted SERS assays (Paria et al., 2022; Yang et al., 2022; Zhang et al., 2022). However, the application of non-targeted SERS for identifying SARS-CoV-2 in environmental samples remains unexplored mainly because applying SERS to virus surveillance in complex environmental matrices faces challenges. First, the size disparity between viruses (50–150 nm diameter) and sub-10 nm plasmonic nanogap hotspots restricts viral access to SERS-active zones, restricting attainable molecular information and hampering SERS detection sensitivity (Zhang et al., 2019). Second, identifying viruses in complex environmental matrices is challenging due to molecular signal interference from other biological components. Therefore, unsupervised (Garg et al., 2023; Ringnér, 2008) or supervised (Morais et al., 2020) machine learning (ML) methods are essential for conducting multivariate analysis of high-dimensional SERS datasets. Last, a limited understanding of the contributions of viral constituents like spike proteins, nucleocapsid proteins, and RNA to Raman spectra affects the reliability of SERS data interpretation in complex settings.

In this study, we demonstrate the ultrasensitive SERS detection of SARS-CoV-2 by decomposing viruses into their structural components and condensing them into nanogap SERS hotspots within a compact detection area comprising gold nanoparticle (NP) aggregates. We meticulously analyze the contributions of specific protein and nucleic acid constituents of SARS-CoV-2 to the SERS spectra of the decomposed virus. Leveraging Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), we demonstrate the capacity to pinpoint and characterize viruses based on their unique protein and nucleic acid profiles, thus forging a connection between label-free SERS data and amino acid and nucleotide sequences. Furthermore, we devise a multivariate Gaussian Process Regression (GPR) model that accurately quantifies SARS-CoV-2 concentrations from 10^3 to 10^6 pfu/mL. Lastly, by employing PCA-LDA, we identify three enveloped pathogenic viruses, SARS-CoV-2, influenza A virus, and Zika virus, amidst an environmental dust matrix, achieving high classification accuracies. These findings accentuate the capabilities of our approach as a rapid and precise tool for environmental surveillance of viruses.

2. Materials and methods

2.1. Synthesis of colloidal gold nanoparticles (AuNPs)

AuNPs were prepared through a seed-mediated growth approach. First, the AuNP seeds were synthesized by adding 3.88 mM Na_3Citrate to 100 mL of boiling 1 mM $HAuCl_4 \bullet 3H_2O$ with vigorous stirring and refluxing. The suspension was boiled for 15 min after the solution turned to wine red and then cooled at room temperature. To obtain the final AuNPs, 820 μL of the above-prepared AuNP seeds and 440 μL of 38.8 mM Na_3Citrate were successively added to 100 mL of boiling 0.254 mM $HAuCl_4 \bullet 3H_2O$ with vigorous stirring and refluxing for 30 min. After cooling down to room temperature, AuNPs were obtained and stored at 4 °C for future use.

2.2. Viral lysis protocol

Text S7 contains the protocols for the propagation of SARS-CoV-2, Zika virus, Influenza A virus, and Phi6. 100 μL of virus solutions were mixed with 1 μL of sodium dodecyl sulfate (SDS) in 1.5 mL microcentrifuge tubes. The tubes were placed in an ultrasonic bath sonicator (VEVOR ultrasonic cleaner, 40 kHz frequency) at 50 °C for 30 min.

2.3. SERS detection assay

Aluminum foil was surface silanized by vapor coating with Tridecafluoro-1,1,2,2-tetrahydrooctyl-1-trichlorosilane (TFOCS, Gelest Inc) in a vacuum chamber (Sidorova et al., 2009). 3 μL of analyte samples (e.g., lysed virus, structural proteins, RNA) and 6 μL of AuNPs were pipetted onto the silanized aluminum foil and evaporated at room temperature. For the SERS experiments, we used a confocal Raman microscope (alpha 300 RSA+, WItec, Germany) under 785 nm laser excitation (Xtra II, Toptica, Germany) using a 20x objective lens (5 mW laser power and 2s integration time). A spectrometer (UHTS300, Witec, Germany) containing a CCD camera (DU401A, Oxford Instruments, UK) was used to detect the backscattered photons. Each scan was performed over a 20 um * 20 um area consisting of 100 pixels.

2.4. SERS detection of SARS-CoV-2 structural proteins and RNA

SARS-CoV-2 S (SARS-CoV-2 Spike RBD (N487D) Protein), N (SARS-CoV-2 Nucleocapsid (R203M, D377Y) Protein), and E (SARS-CoV-2 (2019-nCoV) envelope(CoV-E) protein) proteins were purchased from Sino Biological. Aqueous solutions of the S, N, and E proteins were prepared (10 μg/mL), and SERS measurements were performed as described above. Viral RNA was extracted using the Qiagen QIAamp Viral RNA Mini kit (Qiagen, Hilden, Germany) according to the manufacturer's protocol. The elution volume was 60 μL. SERS detection was performed as described above. A negative control was prepared following the same procedure without the virus. The concentrations quantified on Qubit of extracted RNA from SARS-CoV-2, influenza A virus, and Zika virus were 0.8 ng/μL, 1 ng/μL, and 1.6 ng/μL, respectively. The three samples were diluted to a final concentration of 0.8 ng/μL in the buffer from the negative control sample. SERS detection was performed as described above.

2.5. SERS detection of SARS-CoV-2, Zika virus, and influenza A virus

To eliminate the effects of the cell supernatant background, the following sample groups were created. A) Control: SARS-CoV-2 background + influenza A virus background + Zika virus background; B) SARS-CoV-2 group: SARS-CoV-2 (10^5 pfu/mL) + influenza A virus background + Zika virus background; C) Influenza A virus group: influenza A virus(10^5 pfu/mL) + SARS-CoV-2 background + Zika virus background, and D) Zika virus group: Zika virus (10^5 pfu/mL) + SARS-CoV-2 background + influenza A virus background. Viral lysis and SERS detection were performed as described above.

2.6. SERS detection of SARS-CoV-2, Zika virus, and influenza A virus in environmental dust background

Dust was collected from a classroom HVAC filter by vacuuming and was suspended in ultrapure water at a concentration of 5 mg/ml. This stock was lysed using the protocol described above and passed through a

0.22 um filter. It was diluted to a concentration of 1 mg/ml for the experiments. The following sample groups were created again. A) Control: SARS-CoV-2 background + influenza A virus background + Zika virus background; B) SARS-CoV-2 group: SARS-CoV-2 (10^5 pfu/mL) + influenza A virus background + Zika virus background; C) Influenza A virus group: influenza A virus (10^5 pfu/mL) + SARS-CoV-2 background + Zika virus background, and D) Zika virus group: Zika virus (10^5 pfu/mL) + SARS-CoV-2 background + influenza A virus background. 50 μL of the virus sample groups were lysed using the protocol described above and were mixed with 50 μL of the lysed dust solution. SERS detection was performed as described above.

2.7. Multivariate analysis

Baseline correction and cosmic ray removal were performed using Project v4.1 software. The spectra whose maximum peak values were smaller than three times the noise level were discarded. The background noise intensity was determined using recorded signals in the spectral region at 2000 cm^{-1} without molecular Raman peaks (Nam et al., 2022). MATLAB was used for performing ERS calibration and data truncation. Lastly, R was used for performing PCA and LDA (Garg et al., 2023). The PCA-LDA classification results were obtained using the leave-one-out-cross-validation method. The sensitivity and specificity were calculated using the following expressions. Sensitivity = True positives/(True positives + False negatives); Specificity = True negatives/(True negatives + False positives). The multivariate GPR model was trained using the regression learner application in Mathworks MATLAB/SIMULINK (ver. R2022a). To validate the model, we conducted 5-fold cross-validation. Feature selection was performed using the Maximum Relevance Minimum Redundancy (MRMR) algorithm.

2.8. Digital SERS

The SERS maps at different SARS-CoV-2 concentrations were converted into a binary format (0 or 1) based on whether the pixel's intensity exceeded a predefined threshold. The threshold was set as average intensity plus three times the standard deviation (average+3SD) collected from a negative control sample, following the reported protocol (Godoy et al., 2020). Each pixel of the digitized map was multiplied by each corresponding pixel of the original SERS map, thus generating a digital SERS map. The sum of the pixel intensities for each map at various concentrations was used for quantification.

3. Results and discussion

3.1. SERS detection and multivariate analysis of lysed SARS-CoV-2

Fig. 1A outlines our streamlined approach for viral lysis and label-free SERS measurements. Initially, we developed a Raman-background-free viral lysis protocol to decompose viruses, ranging from 50 to 150 nm in size, into structural components that can fit into sub-10 nm plasmonic nanogap hotspots in Au NP aggregates. Next, the lysed virus solution was mixed with Au NPs and deposited on a silane-treated aluminum foil. The hydrophobic silane-treated aluminum foil promotes analyte enrichment within a compact detection region of ~0.8 mm^2 area, yielding ultrasensitive SERS detection. Fig. S1 shows a transmission electron microscopy (TEM) image of sub-5 nm hotspots generated by Au NP aggregates capable of generating high SERS enhancement factors (Ding et al., 2016). Finally, we captured the label-free SERS spectra of the lysed virus components. We observed that commercially available viral lysis buffers exhibit intense SERS signatures due to molecular components with high Raman cross-section that can cause spectral interference with the label-free SERS spectra of target viral analytes. For example, Fig. 1B presents the SERS spectra of lysis buffers used in a commercially available Flowflex antigen test and a nucleic acid extraction kit (QIAamp RNA kit), displaying notable SERS signatures. We observed that sodium dodecyl sulfate (SDS), an anionic detergent capable of disrupting viral envelopes (Miura et al., 2011; Thom et al., 2021), exhibits no discernible SERS peaks (Fig. 1B). Consequently, we developed a Raman-background-free viral lysis protocol combining SDS-based chemical disruption of the viral envelope with sonication and heat treatment. Sonication uses high-frequency sound waves (40 kHz) to agitate and lyse the viruses, while elevated temperatures (e.g., 50 °C) provide kinetic energy to accelerate the physical and chemical lysis of viruses.

We optimized the viral lysis protocol using Phi6, a well-established surrogate for enveloped pathogenic viruses (Aquino de Carvalho et al., 2017; Fedorenko et al., 2020) (**Text S1** and Figs. S3–5). We then applied the developed lysis protocol to obtain the label-free SERS spectra of lysed SARS-CoV-2 (Fig. 1C). Raman signal intensities were calibrated using the electronic Raman scattering (ERS) internal standard across all SERS measurements (Nam et al., 2020). Table S1 details the assigned

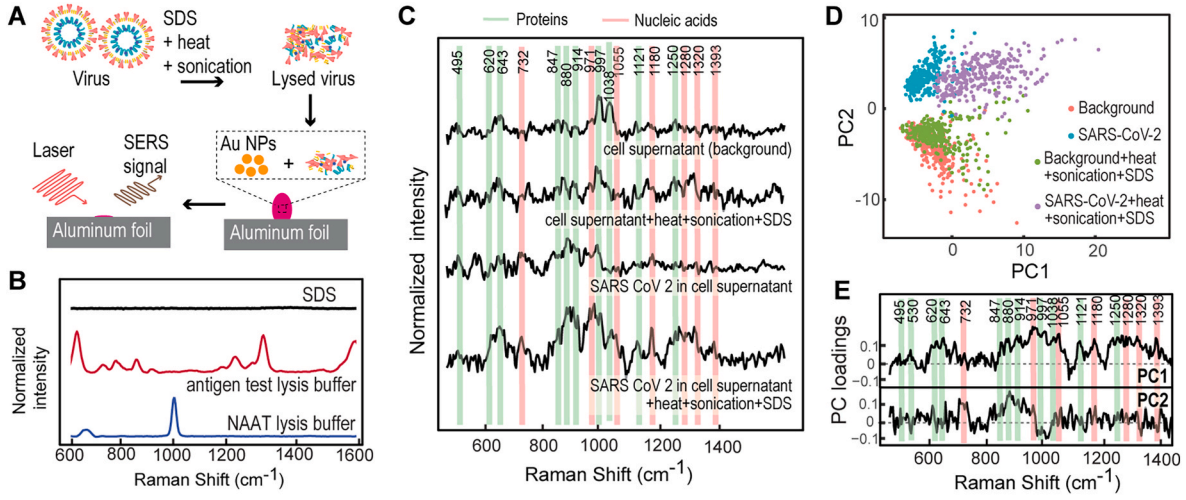

Fig. 1. SERS detection and multivariate analysis of lysed SARS-CoV-2. (A) Schematic illustration of the key steps for SERS detection of lysed viruses. (B) Average ERS-calibrated SERS spectra from several viral lysis buffer solutions. (C) Average ERS-calibrated SERS spectra of uninfected cell supernatant and SARS-CoV-2 without treatment and after treatment with heat, sonication, and SDS. Note: green lines and red lines mark the known positions of the protein peaks and nucleic acid peaks, respectively. (D) PC score scatter plot and (E) PC loadings from the PCA analysis of the SERS spectra from uninfected cell supernatant and SARS-CoV-2 (treated and untreated).

molecular origins of the observed SERS peaks. The control SERS spectrum from uninfected cell supernatant reveals protein-related Raman peaks at 620, 643, 997, and 1038 cm^{-1}. No significant changes in the spectrum were observed after treating the cell supernatant with SDS, sonication, and heat. In contrast to the control SERS spectra from the cell supernatant, the SERS spectrum of 10^6 pfu/mL SARS-CoV-2 exhibits an additional protein-related Raman peak at 880 cm^{-1}, attributed to tryptophan. Following treatment with SDS, heat, and sonication, we noted a significant increase in the intensity of several Raman peaks related to proteins (643, 880, 914, 997, 1038, 1121, and 1250 cm^{-1}) and nucleic acids (732, 971, 1055 and 1180, 1280, 1320, and 1393 cm^{-1}).

SARS-CoV-2 is comprised of spike (S), envelope (E), and membrane (M) proteins that form the virion and an RNA genome bound by nucleocapsid (N) proteins within the virion (Astuti, 2020; Wang et al., 2020). During label-free SERS detection of intact SARS-CoV-2, only a small portion of the S proteins, with a vertical size of approximately 5 nm, can access the sub-10 nm plasmonic nanogap hotspots. Consequently, the dominant protein-related peak at 880 cm^{-1} in the SERS spectrum of unlysed SARS-COV-2 likely originates from the S proteins on the SARS-CoV-2 virion surface. When lysed, the viral structure is dismantled, significantly increasing the accessibility for all virus structural components, both on the surface and inside the virion, to the plasmonic nanogap hotspots in the Au NP aggregates (Fig. S2).

We implemented PCA, an unsupervised multivariate ML analysis tool, to statistically identify the SERS peaks responsible for the differences among the various samples. The principal component (PC) score scatter plot shows a significant overlap between the data points from lysed and unlysed cell supernatant samples, while the SARS-CoV-2 samples are well separated (Fig. 1D). Compared to the cell supernatant samples that show negative PC1 and PC2 values, the SARS-CoV-2 samples are separately clustered with the unlysed SARS-CoV-2 sample exhibiting positive PC2 values and the lysed SARS-CoV-2 sample exhibiting positive PC1 and PC2 values. Fig. 1E shows the loading spectra for PC1 and PC2, revealing the contributions of different vibrational modes to the differences between samples. The protein-related Raman peak at 880 cm^{-1} positively contributes to PC2, signifying that this peak primarily accounts for the differences between the uninfected cell supernatant samples and unlysed SARS-CoV-2 samples. Several Raman peaks from proteins (495, 530, 620, 643, 880, 914, 997, 1038,1121 and 1250 cm^{-1}) and nucleic acids (971, 1055, 1180, 1280, 1320, and 1393 cm-1) positively contribute to PC1, statistically verifying that these peaks increase significantly in the SARS-CoV-2 samples after lysis.

3.2. SERS detection and multivariate analysis of SARS-CoV-2 structural components

To fully comprehend the origins of the observed peaks, a SERS database for SARS-CoV-2 components is vital. First, we obtained the SERS spectra of S, N, and E proteins in SARS-CoV-2 (Fig. 2A, Text S3). A strong peak at 880 cm^{-1} in S proteins aligns with our hypothesis that the untreated SARS-CoV-2 SERS peak at 880 cm^{-1} originates from S proteins on the SARS-CoV-2 virion surface. Due to their different amino acid compositions, the peak intensities corresponding to different vibrational modes varied in S, N, and E proteins (**Text S2**). We implemented PCA to extract the contributions of different Raman modes to the SERS spectra of S, N, and E proteins. The PC score scatter plot demonstrates that the SERS spectra data points from S, N, and E proteins are separated along PC3 (Fig. 2C). The data points from S proteins exhibit positive PC3 values, while most data points from N proteins show negative PC3 values. The PC3 loading spectrum shows that the peaks from tyrosine (643 cm^{-1}) and tryptophan (880 and 1353 cm^{-1}) positively contribute

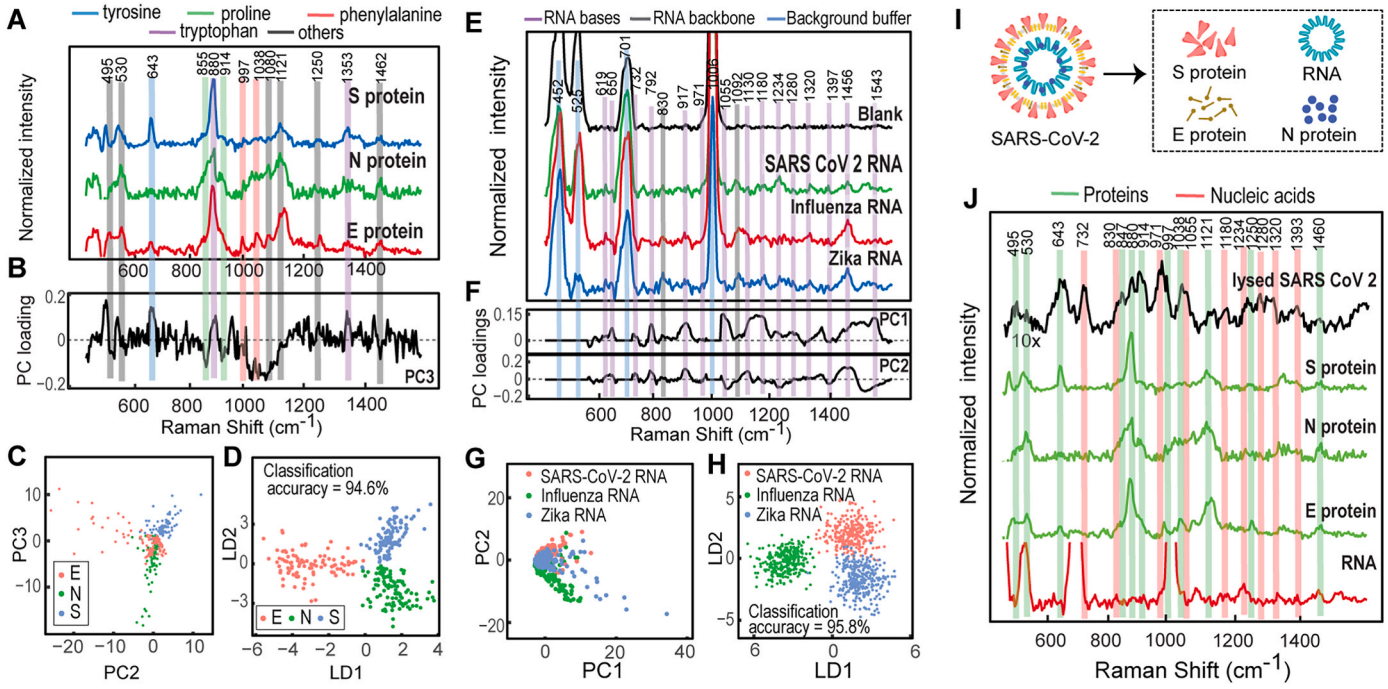

Fig. 2. SERS detection and multivariate analysis of SARS-CoV-2 structural components. (A) Average ERS-calibrated SERS spectra of the spike (S), nucleocapsid (N), and envelope (E) proteins in SARS-CoV-2. Note: blue, green, red, purple, and black lines indicate the known Raman peak positions for tyrosine, proline, phenylalanine, tryptophan, and other proteins, respectively. (B) PC loadings derived from the PCA analysis of the SERS spectra of S, N, and E proteins in SARS-CoV-2. (C) PCA and (D) PCA-LDA score scatter plots from S, N, and E proteins in SARS-CoV-2. (E) Average ERS-calibrated SERS spectra of extracted RNA from Zika virus, influenza A virus, and SARS-CoV-2. Note: purple, black, and blue lines indicate the known Raman peak positions for RNA bases, RNA backbone, and RNA extraction buffer, respectively. (F) PC loadings derived from the PCA analysis of the SERS spectra of extracted RNA from the Zika virus, influenza A virus, and SARS-CoV-2. (G) PCA and (H) PCA-LDA score scatter plots for SERS spectra from extracted RNA of Zika virus, influenza A virus, and SARS-CoV-2. (I) Schematic illustration depicting the structural components of SARS-CoV-2. (J) Average ERS-calibrated SERS spectra of lysed SARS-CoV-2 and those of SARS-CoV-2 structural components, including S proteins, N proteins, E proteins, and extracted RNA. Note: green and red lines mark the known Raman peak positions for the proteins and nucleic acids, respectively.

to PC3, indicating a significant contribution to S proteins, while the peaks from proline (855 and 914 cm^{-1}) negatively contribute to PC3, revealing considerable contributions to N proteins (Fig. 2B). Amino acid sequences confirmed higher tyrosine and lower proline percentages in S protein compared to N protein (Fig. S6, Text S2). However, the PC score scatter plot shows a significant overlap between the S, N, and E protein data points (Fig. 2C). We employed supervised PCA-LDA methods to classify the measured SERS spectra acquired from S, N, and E proteins (Fig. 2D). The PCA-LDA model can segregate the S, N, and E groups with an overall classification accuracy of 94.6% (Table S2). Thus, viral protein components can be identified based on their amino acid composition via multivariate analysis of label-free SERS datasets.

To investigate the feasibility of using SERS to identify different viruses, we obtained the SERS spectra of extracted RNA from SARS-CoV-2, influenza A virus, and Zika virus (Fig. 2E). Compared to the blank samples with peaks from the RNA extraction kit buffers, the extracted RNA samples exhibit several additional peaks from the RNA backbone and the four nucleotide bases: adenine, guanine, cytosine, and uracil **(Table S3)**. To extract the subtle differences between the SERS dataset of extracted RNA from SARS-CoV-2, influenza A virus, and Zika virus, we employed PCA. The PC1 and PC2 loadings display various features corresponding to the RNA bases (619, 650, 732, 792, 917, 1055, 1130, 1180, 1234, 1280, 1397, 1456, and 1543 cm^{-1}), indicating that the different nucleotide base compositions significantly contribute to the differences between the viral extracted RNA (Fig. 2F). While the PCA score scatter plot reveals distinct clusters with substantial overlap between the three samples (Fig. 2G), the PCA-LDA model segregates the three samples with a classification accuracy of 95.8% (Fig. 2H, Table S4). Therefore, viruses can be identified based on their nucleotide base composition differences via multivariate analysis of label-free SERS datasets. Thus, our in-depth analysis of viral structural components bridges the label-free SERS data to the amino acid and nucleotide sequences, establishing a robust foundation for SERS-based virus detection in complex matrices.

To understand the origins of the SERS spectra of lysed SARS-CoV-2, we compared the SERS spectrum of lysed SARS-CoV-2 to the SERS spectra of SARS-CoV-2 structural components (Fig. 2I). Indeed, the SERS peaks in lysed SARS-CoV-2 are a combination of the peaks from SARS-CoV-2 structural proteins and RNA (Fig. 2J), further validating that the SERS peaks of lysed SARS-CoV-2 originate from SARS-CoV-2 structural components.

3.3. Quantification of SARS-CoV-2 using multivariate regression

We obtained the SERS spectra of lysed SARS-CoV-2 at concentrations varying from 10^2 to 10^6 pfu/mL (Fig. 3A). The PC loading spectra revealed substantial contributions from the SARS-CoV-2 protein and RNA components to the separation (Fig. 3B), while PCA score scatter plots demonstrated the distinction between the SERS spectra of different SARS-CoV-2 concentrations from 10^2 to 10^6 pfu/mL (Fig. 3C). Although PCA effectively discriminates between various SARS-CoV-2 concentrations, it cannot provide quantitative predictions.

Numerous publications have demonstrated the dependable quantitative detection of SARS-CoV-2 using targeted SERS probes with signals from Raman reporter molecules (Cha et al., 2022; Park et al., 2022). While some reports have also explored quantitative analysis through non-targeted SERS methods, achieving reliable quantification has proven to be a challenge due to the interference posed by background components, including viral lysis buffers, viral inactivation solutions, and viral transport media. Existing literature often reports the limit of detection (LoD) by measuring the lowest detectable signals from the most substantial observable peaks ((Huang et al., 2023; Zhang et al., 2022)) or via multivariate regression using the entire spectrum ((Paria et al., 2022; Yang et al., 2022)). However, these analyses may include contributions from background components, leading to potentially inaccurate results. Our methodology addresses these challenges by combining Raman-background-free lysis, UV inactivation, and PCA-enabled identification of SARS-CoV-2 peaks. Furthermore, peak validation is achieved using a spectral library of SARS-CoV-2 structural components, ensuring reliable quantitative analysis **(Table S5)**.

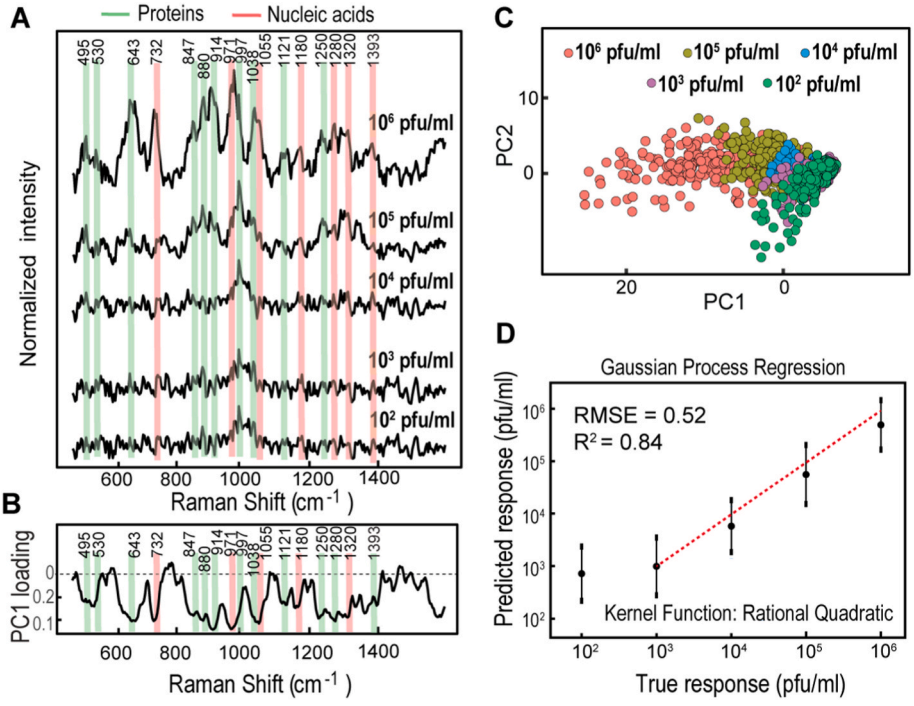

Fig. 3. Quantification of SARS-CoV-2 via multivariate regression. (A) Average ERS-calibrated SERS spectra of lysed SARS-CoV-2, with concentrations spanning 10^2-10^6 pfu/mL. Note: the green and red lines represent established Raman peak positions for proteins and nucleic acids, respectively. (B) PC loadings. (C) Scatter plot of PC scores, based on PCA analysis of the SERS spectra from lysed SARS-CoV-2, with concentrations ranging from 10^2 and 10^6 pfu/mL. (D) Prediction of SARS-CoV-2 concentration between 10^3 and 10^6 pfu/mL, using a GPR model featuring a rational quadratic kernel function (error bars represent standard deviation).

For reliable quantitive analysis, we developed a multivariate nonparametric regression model to map SERS measurements into predicted SARS-CoV-2 concentrations between 10^3 and 10^6 pfu/mL. As labeled in Figs. 3A and 17 SERS fingerprint peaks originating from SARS-CoV-2 structural components (Figs. 1E and 2J) were utilized to construct this regression model. Specifically, we employed GPR to create the model using the combined SERS dataset from lysed SARS-CoV-2 at concentrations between 10^2 and 10^6 pfu/mL. GPR is a nonparametric regression method that uses kernel functions to measure pattern similarity between the training and test datasets. Fig. 3D displays SARS-CoV-2 concentration predictions based on the GPR model with a rational quadratic kernel function. The GPR model exhibited strong agreement between the predicted and actual concentration values between 10^3 and 10^6 pfu/mL, with an *RMSE* of 0.52 and an R^2 of 0.84. To determine the LoD, we first employed the MRMR approach for wavenumber selection. This method identifies the wavenumbers from the regression model with the most correlation to SARS-CoV-2 concentrations and the least correlation between themselves. We selected the top three wavenumbers (997, 914, and 1320 cm^{-1}) and performed a digital SERS analysis to quantify the LoD. This digital SERS method can enable single molecule SERS analysis by filtering out data points below a predefined threshold and summing the remaining data points within the SERS maps at various concentrations, facilitating quantitive analysis (Godoy et al., 2020; Nam et al., 2021). Figs. S7A–C demonstrate the digital SERS maps at the three chosen wavenumbers across various SARS-CoV-2 concentrations between 10^2 and 10^6 pfu/mL. These maps reveal pixel values exceeding the threshold down to 10^2 pfu/mL. Figs. S7D–F illustrate the summed pixel intensities across the maps at various concentrations. Our results demonstrate a strong fit to the calibration curve, which was determined using a four-parameter sigmoidal fitting equation ($R^2 = 0.99$, 0.99, and 0.98 for 917,997, and 1320 cm^{-1}, respectively). The LoD was established at 10^2 pfu/ml highlighting the sensitivity of our approach **(Table S5)**.

3.4. *Detection of pathogenic enveloped viruses in environmental dust matrix*

We conducted SERS measurements for three distinct pathogenic enveloped viruses, including SARS-CoV-2, influenza A virus, and Zika virus. Fig. 4A shows the SERS spectra for the lysed virus samples and the control sample of uninfected cell supernatant **(Text S5)**. Since identifying virus types directly from the measured SERS spectra is challenging, we explored whether different viruses could be identified using ML multivariate analysis of their SERS spectral features. The PCA score plot reveals distinct clusters from different groups, with a slight overlap between control and virus samples likely due to their shared base media (Fig. 4C). Specifically, the control cluster is separated from the three virus clusters along PC1. The three virus clusters are separated along PC2, with significant overlap between the two respiratory viruses, SARS-CoV-2 and influenza A viruses. The PC1 and PC2 loadings show multiple peaks from proteins (643, 828, 847, 880, and 997 cm^{-1}) and nucleic acids (732, 971, 1055, and 1285 cm^{-1}) that contribute to the variance in the datasets, enabling clustering of the SERS datasets among different viruses (Fig. 4B).

We employed supervised PCA-LDA for data classification. The PCA-LDA model effectively segregates the different groups with classification sensitivities of 93.5%, 98.5%, and 96.3% for SARS-CoV-2, Zika virus, and influenza A virus, respectively (Fig. 4D, Table S6). The subtle

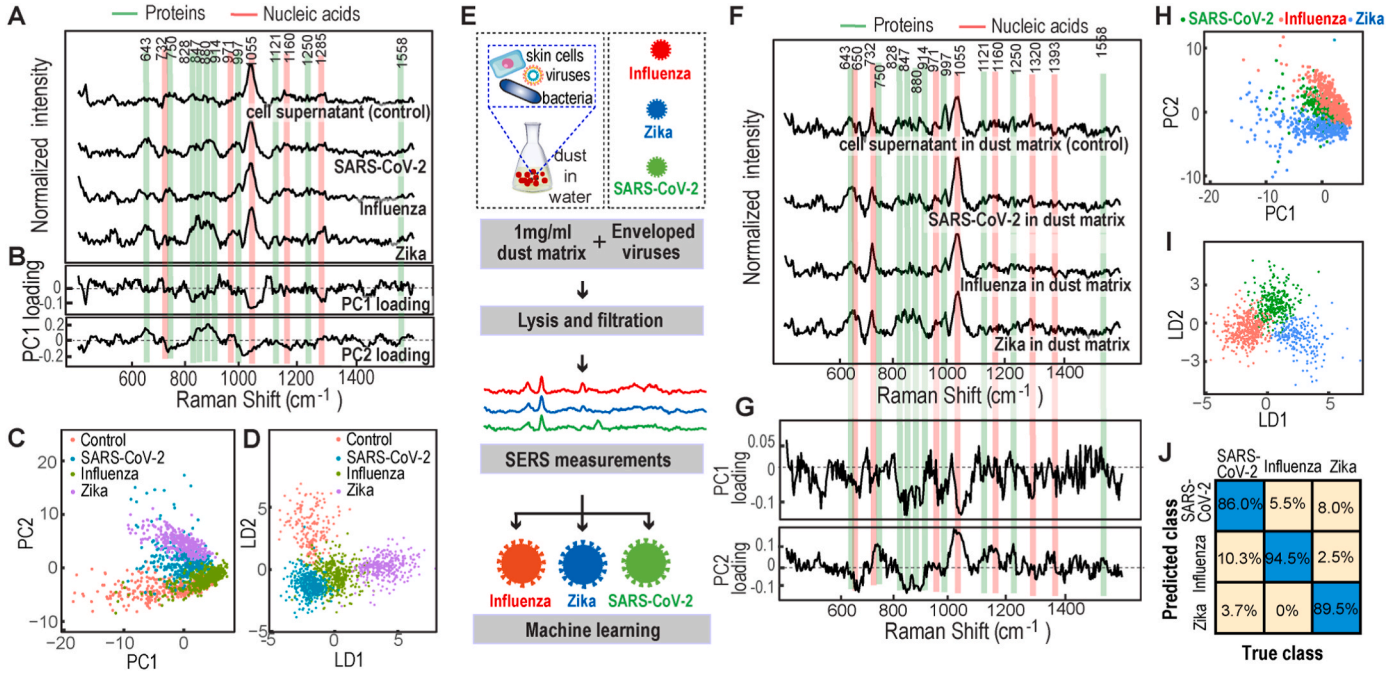

Fig. 4. Detection of pathogenic enveloped viruses in environmental dust matrix. (A) Average ERS-calibrated SERS spectra of background cell supernatant and enveloped pathogenic viruses, including SARS-CoV-2, influenza A virus, and Zika virus. The green and red lines indicate the established Raman peak positions for proteins and nucleic acids, respectively. (B) PC loadings. (C) Scatter plot of PC scores. (D) Scatter plot of PCA-LDA scores, derived from the multivariate analysis of the SERS spectra associated with background cell supernatant and enveloped pathogenic viruses such as SARS-CoV-2, influenza A virus, and Zika virus. (E) A schematic representation outlines the critical steps in identifying enveloped pathogenic viruses within an environmental dust matrix through label-free SERS coupled with machine learning. (F) Average ERS-calibrated SERS spectra for the environmental dust matrix amalgamated with background cell supernatant and enveloped pathogenic viruses, which include SARS-CoV-2, influenza A virus, and Zika virus. (G) PC loadings derived from the unsupervised PCA analysis of the SERS spectra for the environmental dust matrix combined with background cell supernatant and enveloped pathogenic viruses, such as SARS-CoV-2, influenza A virus, and Zika virus. (H) Scatter plot of PC scores from the same unsupervised PCA analysis. (I) Scatter plot of PCA-LDA scores and (J) Confusion matrix from the supervised PCA-LDA analysis of the SERS spectra concerning the environmental dust matrix amalgamated with background cell supernatant and enveloped pathogenic viruses, including SARS-CoV-2, influenza A virus, and Zika virus.

differences in the molecular fingerprints that enable accurate classification of different viruses can originate from (i) the varying genomic and proteomic compositions of the structural components of the different viruses and (ii) the distinct immune responses of cells to infection from different viruses. Furthermore, since environmental samples can contain multiple airborne viruses simultaneously, this study has also demonstrated the feasibility of using ML-empowered SERS to identify samples containing both SARS-CoV-2 and influenza A virus with high accuracy (Fig. S8 and Text S4).

To proactively forecast viral outbreaks, rapid and low-cost monitoring of airborne viruses is crucial in crowded indoor environments. Air in these environments contains high concentrations of dust particles (~0.0001 mg/L) with various biological components (e.g., dead skin, hair, bacteria, viruses) (Oomen et al., 2008). Thus, identifying airborne viruses in a dust matrix is essential to simulate real-life scenarios. We collected dust samples from classroom air filters, diluted them to 1 mg/mL, and spiked them with pathogenic enveloped viruses (SARS-CoV-2, influenza A virus, and Zika virus) at specific concentrations (Fig. 4E). This dust matrix simulates components from ~10,000 L of air sampled from crowded indoor spaces, eluted in 1 mL of solvent. The spiked dust samples underwent lysis to decompose the viruses and filtration to remove large dust particles. Label-free SERS measurements and multivariate analysis were performed to identify the pathogenic viruses in the dust matrix.

Fig. 4F presents the SERS spectra of different viruses in the environmental dust matrix. The SERS spectrum of environmental dust exhibits strong Raman peaks corresponding to nucleic acids, proteins, and lipids from various biological constituents in the sample (Fig. S9 and Text S6). The SERS signatures of the cell supernatant and viruses spiked in the environmental dust resemble the spectrum of the environmental dust due to the high load of nucleic acid, protein, and lipid components in the dust sample (Fig. 4F). The PCA score scatter plot in Fig. 4H reveals distinct clusters from different viruses, primarily separated along PC2, with substantial overlap between the SARS-CoV-2 and influenza A virus samples. Again, the PC1 and PC2 loadings display various features corresponding to virus-related proteins (643, 828, 847, 880, 997, 1121, 1250, and 1558 cm^{-1}) and nucleic acids (732, 971, 1055, 1160, and 1320 cm^{-1}) that significantly contribute to the differences between the samples in the dust matrix (Fig. 4G).

For virus classification and identification, we implemented supervised PCA-LDA. The LDA score scatter plots reveal a clear separation between the sample groups (Fig. 4I). From the confusion matrix in Fig. 4J, the sensitivities for the classification of SARS-CoV-2, influenza A virus, and Zika virus are 86.0%, 94.5%, and 89.5%, respectively. Furthermore, the specificities for the classification of SARS-CoV-2, influenza A virus, and Zika virus are 93.6%, 92.8%, and 98.4%, respectively, indicating the selective identification of viruses amongst complex environmental backgrounds.

4. Conclusions

We have demonstrated that by decomposing viruses and then concentrating them into nanogap plasmonic hotspots, SERS-based virus detection can be improved by increasing Raman signal intensity and offering rich Raman fingerprint information. The label-free SERS fingerprint information, which corresponds to virus-related proteomic and genomic information, when combined with ML data analytics, enables the identification of pathogenic viruses in complex environmental backgrounds. We successfully identified three pathogenic viruses with over 86% accuracy in an environmental dust background within 45 min. Previous research has documented the presence of a SARS-CoV-2 load of 48,000 gene copies/m^3 (equivalent to approximately 0.048–4.8 pfu/m^3 (Klimstra et al., 2020; Lin et al., 2022)) in air samples collected using personalized samplers (50L/minute) near COVID-19 patients in negative-pressure hospital rooms (Santarpia et al., 2020). Another study detected >0.013 pfu/m^3 of SARS-CoV-2 using high-flow rate samplers (150L/minute) in well-ventilated hospital areas with COVID-19 patients (Ang et al., 2022). Consequently, our method, with a detection limit of 0.3 pfu/test (100 pfu/ml), holds the potential to identify airborne viruses in indoor spaces with sampled air volumes of approximately 10–20 m^3. Such volumes can be effectively sampled within 1–2 h using high-flow rate samplers (e.g., 150L/minute), underscoring the practicality of our device for on-site environmental viral surveillance in indoor settings. Nevertheless, further studies are required with spiked dust samples and real sampled air to validate the translational utility of our system.

Although our ML-boosted SERS fingerprinting approach can be a valuable tool for the label-free screening of airborne viruses in indoor spaces, targeted bioanalysis methods still play an essential role in the absolute quantitative detection of positively screened viruses for evaluating their amount in the indoor environment. We envision that, when combined with on-site air sampling, an integrated microfluidics system, and a portable Raman spectrometer, the ML-boosted SERS fingerprinting approach can potentially enable rapid, on-site environmental virus surveillance, ultimately improving the management of future viral outbreaks.

CRediT authorship contribution statement

Aditya Garg: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing, Project administration, Software, Validation, Visualization. **Seth Hawks:** Data curation, Formal analysis, Writing – original draft. **Jin Pan:** Data curation, Investigation, Methodology. **Wei Wang:** Investigation, Methodology. **Nisha Duggal:** Methodology, Supervision, Investigation, Resources. **Linsey C. Marr:** Funding acquisition, Methodology, Project administration, Resources, Supervision. **Peter Vikesland:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing – original draft. **Wei Zhou:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by Wellcome Leap Inc (A5M3XP5X), by the US National Science Foundation grants OISE-1545756, CBET-2029911, and CBET-2231807. Laboratory and instrumentation support were provided by NanoEarth—a node of the NSF-supported NNCI (NSF award number #1542100). Additional support was provided by the Sustainable Nanotechnology Interdisciplinary Graduate Program (VTSuN IGEP), funded by Virginia Tech.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bios.2023.115946.

References

Abdelhamid, H.N., Badr, G., 2021. Nanotechnol.Environ. Eng. 6, 1–26.

Ahmadivand, A., Gerislioglu, B., Ahuja, R., Mishra, Y.K., 2020. Mater. Today 32, 108–130.
Ahmadivand, A., Gerislioglu, B., Ramezani, Z., Kaushik, A., Manickam, P., Ghoreishi, S. A., 2021. Biosens. Bioelectron. 177, 112971.
Ang, A.X., Luhung, I., Ahidjo, B.A., Drautz-Moses, D.I., Tambyah, P.A., Mok, C.K., Lau, K. J., Tham, S.M., Chu, J.J.H., Allen, D.M., 2022. Indoor Air 32 (1), e12930.
Aquino de Carvalho, N., Stachler, E.N., Cimabue, N., Bibby, K., 2017. Environ. Sci. Technol. 51 (15), 8692–8700.
Astuti, I., 2020. Diabetes & metabolic syndrome. Clin. Res. Rev. 14 (4), 407–412.
Bazant, M.Z., Bush, J.W., 2021. Proc. Natl. Acad. Sci. USA 118 (17), e2018995118.
Bizkarguenaga, M., Bruzzone, C., Gil-Redondo, R., SanJuan, I., Martin-Ruiz, I., Barriales, D., Palacios, A., Pasco, S.T., González-Valle, B., Laín, A., 2022. NMR Biomed. 35 (2), e4637.
Cha, H., Kim, H., Joung, Y., Kang, H., Moon, J., Jang, H., Park, S., Kwon, H.-J., Lee, I.-C., Kim, S., 2022. Biosens. Bioelectron. 202, 114008.
Ding, S.-Y., Yi, J., Li, J.-F., Ren, B., Wu, D.-Y., Panneerselvam, R., Tian, Z.-Q., 2016. Nat. Rev. Mater. 1 (6), 1–16.
Eissa, S., Zourob, M., 2020. Anal. Chem. 93 (3), 1826–1833.
Fathi-Hafshejani, P., Azam, N., Wang, L., Kuroda, M.A., Hamilton, M.C., Hasim, S., Mahjouri-Samani, M., 2021. ACS Nano 15 (7), 11461–11469.
Fedorenko, A., Grinberg, M., Orevi, T., Kashtan, N., 2020. Sci. Rep. 10 (1), 1–10.
Garg, A., Mejia, E., Nam, W., Nie, M., Wang, W., Vikesland, P., Zhou, W., 2022. Small, 2106887.
Garg, A., Nam, W., Wang, W., Vikesland, P., Zhou, W., 2023. ACS Sensors.
Godoy, N., García-Lojo, D., Sigoli, F., Pérez-Juste, J., Pastoriza-Santos, I., Mazali, I., 2020. Sensor. Actuator. B Chem. 320, 128412.
Huang, J., Wang, C., Wang, P., Mo, W., Zhou, M., Le, W., Qi, D., Wei, L., Fan, Q., Yang, Y., 2023. ACS Applied Materials & Interfaces.
Kevadiya, B.D., Machhi, J., Herskovitz, J., Oleynikov, M.D., Blomberg, W.R., Bajwa, N., Soni, D., Das, S., Hasan, M., Patel, M., 2021. Nat. Mater. 20 (5), 593–605.
Klimstra, W.B., Tilston-Lunel, N.L., Nambulli, S., Boslett, J., McMillen, C.M., Gilliland, T., Dunn, M.D., Sun, C., Wheeler, S.E., Wells, A., 2020. J. Gen. Virol. 101 (11), 1156.
Langer, J., Jimenez de Aberasturi, D., Aizpurua, J., Alvarez-Puebla, R.A., Auguié, B., Baumberg, J.J., Bazan, G.C., Bell, S.E., Boisen, A., Brolo, A.G., 2019. ACS Nano 14 (1), 28–117.
Liangou, A., Tasoglou, A., Huber, H.J., Wistrom, C., Brody, K., Menon, P.G., Bebekoski, T., Menschel, K., Davidson-Fiedler, M., DeMarco, K., 2021. E Clin. Med. 42, 101207.
Lin, Y.-C., Malott, R.J., Ward, L., Kiplagat, L., Pabbaraju, K., Gill, K., Berenger, B.M., Hu, J., Fonseca, K., Noyce, R.S., 2022. Sci. Rep. 12 (1), 5418.
Liu, Y., Ning, Z., Chen, Y., Guo, M., Liu, Y., Gali, N.K., Sun, L., Duan, Y., Cai, J., Westerdahl, D., 2020. Nature 582 (7813), 557–560.
Miller, S.L., Nazaroff, W.W., Jimenez, J.L., Boerstra, A., Buonanno, G., Dancer, S.J., Kurnitski, J., Marr, L.C., Morawska, L., Noakes, C., 2021. Indoor Air 31 (2), 314–323.
Miura, T., Masago, Y., Sano, D., Omura, T., 2011. Appl. Environ. Microbiol. 77 (12), 3975–3981.
Morais, C.L., Lima, K.M., Singh, M., Martin, F.L., 2020. Nat. Protoc. 15 (7), 2143–2162.
Morawska, L., Tang, J.W., Bahnfleth, W., Bluyssen, P.M., Boerstra, A., Buonanno, G., Cao, J., Dancer, S., Floto, A., Franchimon, F., 2020. Environ. Int. 142, 105832.
Nachtigall, F.M., Pereira, A., Trofymchuk, O.S., Santos, L.S., 2020. Nat. Biotechnol. 38 (10), 1168–1173.
Nam, W., Chen, H., Ren, X., Agah, M., Kim, I., Zhou, W., 2022. ACS Appl. Nano Mater. 5 (8), 10358–10368.
Nam, W., Kim, W., Zhou, W., You, E.-A., 2021. Nanoscale 13 (41), 17340–17349.
Nam, W., Zhao, Y., Song, J., Ali Safiabadi Tali, S., Kang, S., Zhu, W., Lezec, H.J., Agrawal, A., Vikesland, P.J., Zhou, W., 2020. J. Phys. Chem. Lett. 11 (22), 9543–9551.
Oomen, A.G., Janssen, P., Dusseldorp, A., Noorlander, C.W., 2008. RIVM Report 609021064.
Paria, D., Kwok, K.S., Raj, P., Zheng, P., Gracias, D.H., Barman, I., 2022. Nano Lett. 22 (9), 3620–3627.
Park, S., Jeon, C.S., Choi, N., Moon, J.-I., Lee, K.M., Pyun, S.H., Kang, T., Choo, J., 2022. Chem. Eng. J. 446, 137085.
Peeling, R.W., Heymann, D.L., Teo, Y.-Y., Garcia, P.J., 2022. Lancet 399 (10326), 757–768.
Pezzotti, G., Boschetto, F., Ohgitani, E., Fujita, Y., Shin-Ya, M., Adachi, T., Yamamoto, T., Kanamura, N., Marin, E., Zhu, W., 2022. Adv. Sci. 9 (3), 2103287.
Pinals, R.L., Ledesma, F., Yang, D., Navarro, N., Jeong, S., Pak, J.E., Kuo, L., Chuang, Y.-C., Cheng, Y.-W., Sun, H.-Y., 2021. Nano Lett. 21 (5), 2272–2280.
Puthussery, J.V., Ghumra, D.P., McBrearty, K.R., Doherty, B.M., Sumlin, B.J., Sarabandi, A., Mandal, A.G., Shetty, N.J., Gardiner, W.D., Magrecki, J.P., 2023. Nat. Commun. 14 (1), 3692.
Rahmani, A.R., Leili, M., Azarian, G., Poormohammadi, A., 2020. Sci. Total Environ. 740, 140207.
Ringnér, M., 2008. Nat. Biotechnol. 26 (3), 303–304.
Santarpia, J.L., Rivera, D.N., Herrera, V.L., Morwitzer, M.J., Creager, H.M., Santarpia, G. W., Crown, K.K., Brett-Major, D.M., Schnaubelt, E.R., Broadhurst, M.J., 2020. Sci. Rep. 10 (1), 12732.
Seo, G., Lee, G., Kim, M.J., Baek, S.-H., Choi, M., Ku, K.B., Lee, C.-S., Jun, S., Park, D., Kim, H.G., 2020. ACS Nano 14 (4), 5135–5142.
Shen, Y., Li, C., Dong, H., Wang, Z., Martinez, L., Sun, Z., Handel, A., Chen, Z., Chen, E., Ebell, M.H., 2020. JAMA Intern. Med. 180 (12), 1665–1671.
Sidorova, J.M., Li, N., Schwartz, D.C., Folch, A., Monnat Jr., R.J., 2009. Nat. Protoc. 4 (6), 849–861.
Su, S., Wong, G., Shi, W., Liu, J., Lai, A.C., Zhou, J., Liu, W., Bi, Y., Gao, G.F., 2016. Trends Microbiol. 24 (6), 490–502.
Thom, R.E., Eastaugh, L.S., O'Brien, L.M., Ulaeto, D.O., Findlay, J.S., Smither, S.J., Phelps, A.L., Stapleton, H.L., Hamblin, K.A., Weller, S.A., 2021. Front. Cell. Infect. Microbiol. 11, 716436.
Wang, M.-Y., Zhao, R., Gao, L.-J., Gao, X.-F., Wang, D.-P., Cao, J.-M., 2020. Front. Cell. Infect. Microbiol. 10, 587269.
Wang, W., Kang, S., Zhou, W., Vikesland, P.J., 2023. Environmental Science. Nano.
Yang, Y., Xu, B., Murray, J., Haverstick, J., Chen, X., Tripp, R.A., Zhao, Y., 2022. Biosens. Bioelectron. 217, 114721.
Yao, L., Zhu, W., Shi, J., Xu, T., Qu, G., Zhou, W., Yu, X.-F., Zhang, X., Jiang, G., 2021. Chem. Soc. Rev. 50 (6), 3656–3676.
Yüce, M., Filiztekin, E., Özkaya, K.G., 2021. Biosens. Bioelectron. 172, 112752.
Zhang, X., Zhang, X., Luo, C., Liu, Z., Chen, Y., Dong, S., Jiang, C., Yang, S., Wang, F., Xiao, X., 2019. Small 15 (11), 1805516.
Zhang, Z., Jiang, S., Wang, X., Dong, T., Wang, Y., Li, D., Gao, X., Qu, Z., Li, Y., 2022. Sensor. Actuator. B Chem. 359, 131568.
Zong, C., Xu, M., Xu, L.-J., Wei, T., Ma, X., Zheng, X.-S., Hu, R., Ren, B., 2018. Chem. Rev. 118 (10), 4946–4980.