APPLIED SCIENCES AND ENGINEERING

Open-set deep learning–enabled single-cell Raman spectroscopy for rapid identification of airborne pathogens in real-world environments

Longji Zhu[1], Yunan Yang[1,2], Fei Xu[1], Xinyu Lu[3], Mingrui Shuai[4,5], Zhulin An[4], Xiaomeng Chen[2], Hu Li[1], Francis L. Martin[6,7], Peter J. Vikesland[8], Bin Ren[3], Zhong-Qun Tian[3], Yong-Guan Zhu[1,9]*, Li Cui[1]*

Pathogenic bioaerosols are critical for outbreaks of airborne disease; however, rapidly and accurately identifying pathogens directly from complex air environments remains highly challenging. We present an advanced method that combines open-set deep learning (OSDL) with single-cell Raman spectroscopy to identify pathogens in real-world air containing diverse unknown indigenous bacteria that cannot be fully included in training sets. To test and further enhance identification, we constructed the Raman datasets of aerosolized bacteria. Through optimizing OSDL algorithms and training strategies, Raman-OSDL achieves 93% accuracy for five target airborne pathogens, 84% accuracy for untrained air bacteria, and 36% reduction in false positive rates compared to conventional close-set algorithms. It offers a high detection sensitivity down to 1:1000. When applied to real air containing >4600 bacterial species, our method accurately identifies single or multiple pathogens simultaneously within an hour. This single-cell tool advances rapidly surveilling pathogens in complex environments to prevent infection transmission.

INTRODUCTION

Bacterial infections are a leading cause of global mortality and are responsible for >7.7 million deaths annually (*1*, *2*). Bio-aerosols are critical vehicles for airborne pathogen transmission between the environment and humans (*3*–*5*). Respiratory activities such as coughing, sneezing, talking, and breathing by individuals suffering from respiratory infections produce pathogen-laden aerosols that can travel long distances and remain suspended in air for hours (*3*, *6*). If inhaled, then pathogen-laden aerosols potentially colonize the airways or lungs and may lead to respiratory disease (*7*). Aerosol transmission is a major source of nosocomial infections, particularly in operating rooms and intensive care units (ICUs) (*8*). Accordingly, timely and accurate detection of airborne pathogens is essential to contain infectious disease outbreaks at their source and to reduce nosocomial infection rates (*9*, *10*).

Despite the pervasive threats of airborne pathogens, tools to rapidly identify them in real environments are limited. Conventional culture-based approaches take several days to yield results (*11*, *12*). While matrix-assisted laser desorption ionization–time-of-flight mass spectrometry (MALDI-TOF MS) enables pathogen identification within a few minutes (*13*), the method nonetheless requires lengthy microbial cultivation to form colonies prior to identification (*14*). In addition, cultivation methods are futile for bacteria that adopt a viable but nonculturable (VBNC) state as a survival strategy (*15*). Culture-independent methods, such as polymerase chain reaction, isothermal amplification, enzyme-linked immunosorbent assay, or biosensors, can offer relatively rapid detection (*16*–*20*). However, they are heavily reliant on preselected recognition elements (e.g., specific primers, antibodies, and aptamers) and are limited by the type of pathogens that can be simultaneously identified, and their typically low airborne densities. DNA sequencing enables unbiased simultaneous detection of nearly all pathogens from a given environment (*21*, *22*), but requires substantial bacterial biomass for library construction and long timeframes (several to tens of days) for sequencing. Therefore, advanced techniques for rapid, accurate, and universal detection of airborne pathogens are urgently needed.

Single-cell Raman spectroscopy combined with deep learning (Raman-DL) provides a promising approach to identify pathogens in a rapid and culture-free manner (*23*–*25*). By training a model using Raman spectra of pathogenic bacteria and testing bacterial types within the training set (close-set identification), DL enables discrimination of subtle differences in Raman spectral fingerprints at the species and strain levels (*26*–*29*). Unfortunately, conventional close-set Raman-DL identification is limited to the bacterial spectra included in the training set. When applying Raman-DL to interrogate pathogens in real-world environments, high misclassification rates (i.e., false positives) are typical. The reason is that in real environments, a limited number of pathogenic bacteria coexist with large numbers of highly diverse nonpathogenic indigenous bacteria. In addition, because up to 99% of environmental bacteria are as yet unculturable (*30*, *31*) and bacterial diversity can be 100 to 1000× higher than that of pathogenic bacteria, it is impossible to encompass all classes of environmental bacteria within Raman-DL training sets. The presence of these unseen/unknown bacteria for the training model greatly impairs the performance of traditional close-set DL approaches that are necessarily forced to choose from the

[1]Key Lab of Urban Environment and Health, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen 361021, China. [2]College of Life Science, Northeast Agricultural University, Harbin 150030, China. [3]State Key Laboratory of Physical Chemistry of Solid Surfaces, College of Chemistry and Chemical Engineering, Xiamen University, Xiamen 361005, China. [4]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China. [5]Anhui University, Hefei 230601, China. [6]Biocel UK Ltd., Hull HU10 6TS, UK. [7]Department of Cellular Pathology, Blackpool Teaching Hospitals NHS Foundation Trust, Whinney Heys Road, Blackpool FY3 8NR, UK. [8]Department of Civil and Environmental Engineering, Virginia Tech, Blacksburg, VA 24061, USA. [9]State Key Lab of Urban and Regional Ecology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China.
*Corresponding author. Email: ygzhu@rcees.ac.cn (Y.-G.Z.); lcui@iue.ac.cn (L.C.)

known classes included in the training set, thus leading to false-positive results (*32*, *33*). To address the diversity in real-world situations, bacterial identification should be open-set, i.e., the model should not only accurately classify known classes but also appropriately exclude unseen/unknown classes.

Recently, open-set DL (OSDL) approaches have shown potential toward improving the accuracy of image recognition in real-world settings, such as facial and medical images (*33*, *34*). However, the utility for OSDL to enable single-cell Raman identification of pathogens in real environments remains underexplored. Compared with face images, Raman spectra of bacteria comprising around 225 features (from 600 to 1800 cm^{-1}) are more complex and vary with bacterial physiology. Therefore, to ensure the identification accuracy specific for airborne pathogens, Raman datasets of pathogens in an aerosol state should be established for model training purposes.

To address the above challenges and advance single-cell Raman spectroscopy as a tool to identify pathogens in real-world air environments, we designed the microbial aerosol generation and collection (MAGC) device as a means to collect pathogen-laden aerosols, and then established the first Raman spectral training datasets of pathogens collected from air environments. This device allows us to build a Raman spectral dataset that is closer to real-world airborne pathogens. Moreover, an advanced OSDL algorithm was developed by introducing two open-set strategies (softmax with threshold and openmax with threshold) to a DL attentional neural network (aNN) model. This is, to our knowledge, the first combination of DL and open-set strategy in the identification of bacteria based on Raman signal in real-world environments. By optimizing the algorithm threshold and testing it using a dataset containing both known pathogens and unknown air bacteria, the Raman-OSDL not only maintained a high 93% identification accuracy for target pathogens but also significantly reduced (~36%) the misclassification of unknown air bacteria as pathogens (i.e., false-positive rate) compared to traditional close-set DL. The optimized open-set aNN model was then applied to eight real-world air samples containing either single or multiple pathogens within a population of >4600 other bacterial species as determined by metagenomic sequencing. The positive presence of high- or low-abundance pathogens reported by Raman-OSDL was cross validated by metagenomic sequencing and culture methods. The entire process including aerosol collection, single-cell Raman acquisition, and OSDL output requires approximately 1 hour for high-abundance pathogens. These results demonstrate the capability of Raman-OSDL to rapidly, reliably, and simultaneously identify multiple airborne pathogens in real-world air environments that contain diverse microbes. Raman-OSDL obviously outperforms the close-set approach and provides a breakthrough that transitions Raman-based pathogen identification from pure culture to real-world environmental settings.

RESULTS

Establishing a RAPD platform for single-cell Raman identification of airborne pathogens

Raman spectra of bacteria are highly sensitive to bacterial physiology, which readily varies with living state and local environment. To identify airborne pathogens, an important premise is to collect a training dataset consisting of Raman spectra of bacteria under air conditions that can be used by algorithms for learning. Toward this objective, a rapid airborne pathogen detection (RAPD) platform comprising three essential components was established. RAPD includes a MAGC device (01), an airborne bacterial pretreatment and Raman acquisition system (02), and a DL model (03) (Fig. 1). The MAGC device was specially designed to enable aerosol generation, collection, ventilation, sterilization, and flexible switching of aerosol sources that contain specific bacteria (fig. S1). The microbial aerosol generation module releases pathogens into a sterile air chamber to form well-defined pathogenic aerosols. Then, a collection module is used to capture air samples containing specific pathogens on a gelatine membrane filter. Gelatine can be dissolved in water and is proven to exert no interference on bacterial Raman spectra (fig. S2). Device throughput is ~50 liter/min, allowing the collection of 1000 liter of air samples sufficient for pathogen detection within 20 min. Considering that air bacterial concentration is generally between 10^4 and 10^6 m^{-3}, and the abundance of pathogens inside is in the range of 10^{-3} to 10^{-2} (*35*–*37*), collecting 1000 liter of air is expected to have at least 10^4 cells, which are sufficient to detect even low-abundance pathogens inside. Ventilation and sterilization modules facilitate switching to other air samples those contain different pathogenic aerosols.

Bio-aerosols containing either target pathogens or nontarget bacteria were collected separately and successively to construct the training database. Five common airborne pathogenic bacteria were used as targets: *Staphylococcus aureus*, *Escherichia coli*, *Pseudomonas aeruginosa*, *Salmonella enterica*, and *Acinetobacter baylyi*. These pathogens have been found to be prevalent in both indoor air (hospital ICU) (*38*) and outdoor air (near sewage) (*39*) environments and cause millions of infection-related deaths each year (*2*). When selecting these pathogenic species, their distribution in the phylogenetic tree was not considered, as the main difficulties we need to overcome is how to avoid interference from highly diverse unknown bacteria in real-world environments on the identification of target pathogens. Nontarget bioaerosols included five nonpathogenic bacteria isolated from air (i.e., *Exiguobacterium acetylicum*, *Priestia megaterium*, *Bacillus velezensis*, *Bacillus cereus*, and *Staphylococcus lentus*) as well as indigenous airborne bacteria directly sampled from indoor and outdoor environments. Through sample collection, pretreatment, and Raman spectral analyses of these bio-aerosols from three separate batches, we constructed a Raman training dataset comprising 7552 spectra that included 6149 spectra from the five pathogenic bacteria and 1403 spectra from the nontarget air bacteria. This dataset was used to train and test the DL model to identify pathogens from real-world air environments (04 and 05 in Fig. 1) that contain not only target pathogens but also highly diverse bacterial classes not considered during training.

Figure 2A compares the single-cell Raman spectra of five pathogens obtained from laboratory liquid culture and from the MAGC device under air conditions. Both liquid and air bacteria were washed by water and air-dried on aluminum foil before Raman measurement. The only difference is that the air-suspended bacteria experienced aerosol states and adapted in the air environment. Unsupervised principal components analysis was used to classify and visualize these species using a two-dimensional score plot (Fig. 2B). The five pathogens in liquid culture or air conditions formed distinct clusters on either side of the line (Bray-Curtis distance, $P < 0.01$). The distances between the same species under liquid and air conditions were larger than those between different species under the same condition. Specifically, the relative intensities of peaks at approximately 1322 and 1578 cm^{-1} in the air environment were

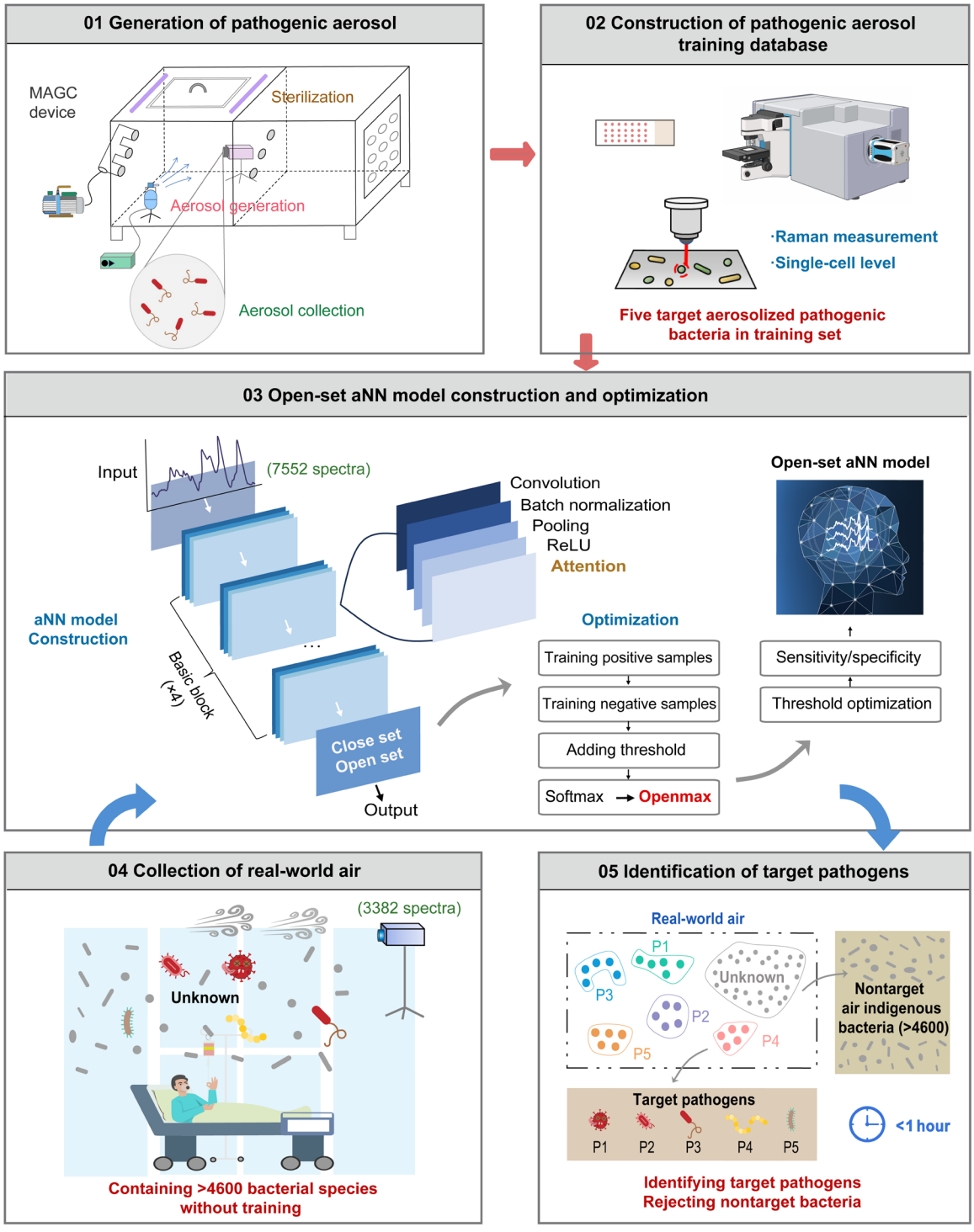

Fig. 1. Overview of the RAPD workflow. Airborne samples containing specific pathogens are generated and collected by the microbial aerosol generation and collection (MAGC) device (01). Bacteria in aerosol states were then subjected to single-cell Raman measurement with only a removing filter membrane pretreatment. The obtained spectra were used to construct the training database of bacteria in aerosol states (02). OSDLs strategies are developed, trained/tested with the constructed database, and optimized with artificially blended spectra containing both target and nontarget air bacteria for classifying airborne pathogens (03). Through the use of the pretrained open-set aNN model, single or multiple target pathogens in real-world air environments containing diverse, unknown indigenous air bacteria (04) that cannot be fully included in the training dataset were successfully identified (05).

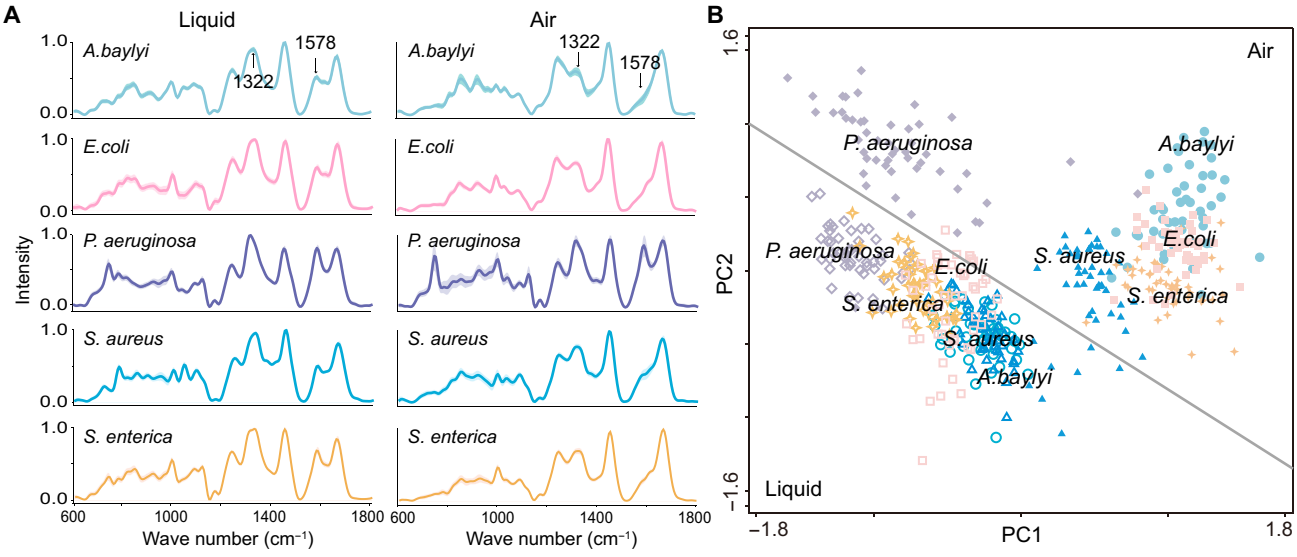

Fig. 2. Comparison of Raman spectra for pathogens in liquid and air conditions. Average Raman spectra (**A**) and principal components analysis (**B**) of the spectra of each pathogen ($n = 50$) under both liquid and air conditions.

significantly lower than those in the liquid culture ($P < 0.05$) except for *P. aeruginosa*. These findings demonstrate the impact of changes in bacterial physiology in liquid and air on Raman spectral features, thus confirming the necessity of establishing Raman training models using pathogens in an aerosol state as opposed to conventional laboratory liquid cultures. Previous studies about the effect of growth and storage conditions (e.g., culture media, storage temperature/media, and time) on Raman spectra of bacteria also ended up suspending the bacteria in water, but the spectral differences were still captured, indicating that bacterial physiological state can be maintained after washing (*40*–*42*).

Developing Raman-OSDL algorithm to identify pathogens and unknown bacterial classes in air

To identify pathogens in real air environments, it is infeasible to collect spectra of all bacterial species present; there are always unknown bacterial classes that the training set has never seen before. Here, to simulate complex air environments, a database of Raman spectra containing both known target pathogens and unknown air bacterial classes was constructed. The performance of four close-set ML/DL algorithms was initially tested in this scenario, including linear discriminant analysis (LDA), random forest (RF), support vector machine (SVM), and aNN (fig. S3). When both the training and testing sets were exposed to the same type of five air pathogens generated by the MAGC device, all four close-set ML/DL models displayed satisfactory accuracy with the aNN algorithm achieving the highest accuracy of 97%. Despite the good performance of the models, this was not realistic of complex air samples. Accordingly, the airborne bacterial classes that were not trained by aNN were newly introduced at the testing stage (labeled "unknown" class, Fig. 3C). This unknown class contains different indigenous airborne bacteria collected directly from outdoor air (test set 1, Fig. 3H). Unfortunately, when testing the "well-trained" aNN model with the unknown classes, the false-positive rate was as high as 100% (Fig. 3C). This result occurs because the traditional close-set algorithms inevitably classify the unknown air classes into one of the known pathogen classes included in the training set (Fig. 3B). To reduce this false positive rate, an "others" class (negative sample) containing five nonpathogenic bacteria isolated from air and bacteria collected directly from indoor air (i.e., indoor air 1) without cultivation were introduced in the training set to allow the model to see and learn (training set 1, Fig. 3H). We emphasize that the bacteria in the others class only represent a limited proportion of airborne bacteria because acquiring spectra of all airborne bacteria is infeasible. However, when we tested the aNN model with the unknown class containing spectra of airborne bacteria collected from outdoor air (i.e., not included in training), only 48% of the unknown dataset were correctly identified as the "others" class and up to 52% was misclassified as one of the five pathogens (i.e., false positive, Fig. 3D). Notably, none of the samples were classified as the "unknown" class. By comparison, for all the classes that were initially trained on aNN including five pathogens and "others" class, >96% accuracy was obtained. The above results indicate the obvious limitation of close-set approaches in testing a new class that was not included in the training model. Although it is possible to train some additional classes outside of the target pathogens of interest, it is not possible to train all real-world classes. In this case, close-set algorithms are not applicable for identification of bacteria in real air environments.

We introduced an open-set algorithm to the aNN to optimize its real-world application. An ideal open-set algorithm should be able to reduce the misclassification rate of unknown classes while retaining a high accuracy for target classes. Here, two open-set strategies were investigated. Compared with the close-set DL, the first strategy used was to introduce a threshold to the softmax activation function in the deep networks (Fig. 3A). Softmax has the capacity to generate a probability distribution for known class labels and to identify samples to the class with the highest probability (*43*). After setting a threshold for the probability of the softmax function, only when a Raman spectrum in the test set exhibited sufficient confidence (e.g., exceeding a threshold of 0.98) to a class in the training set, it was classified as that class; otherwise, it was categorized as

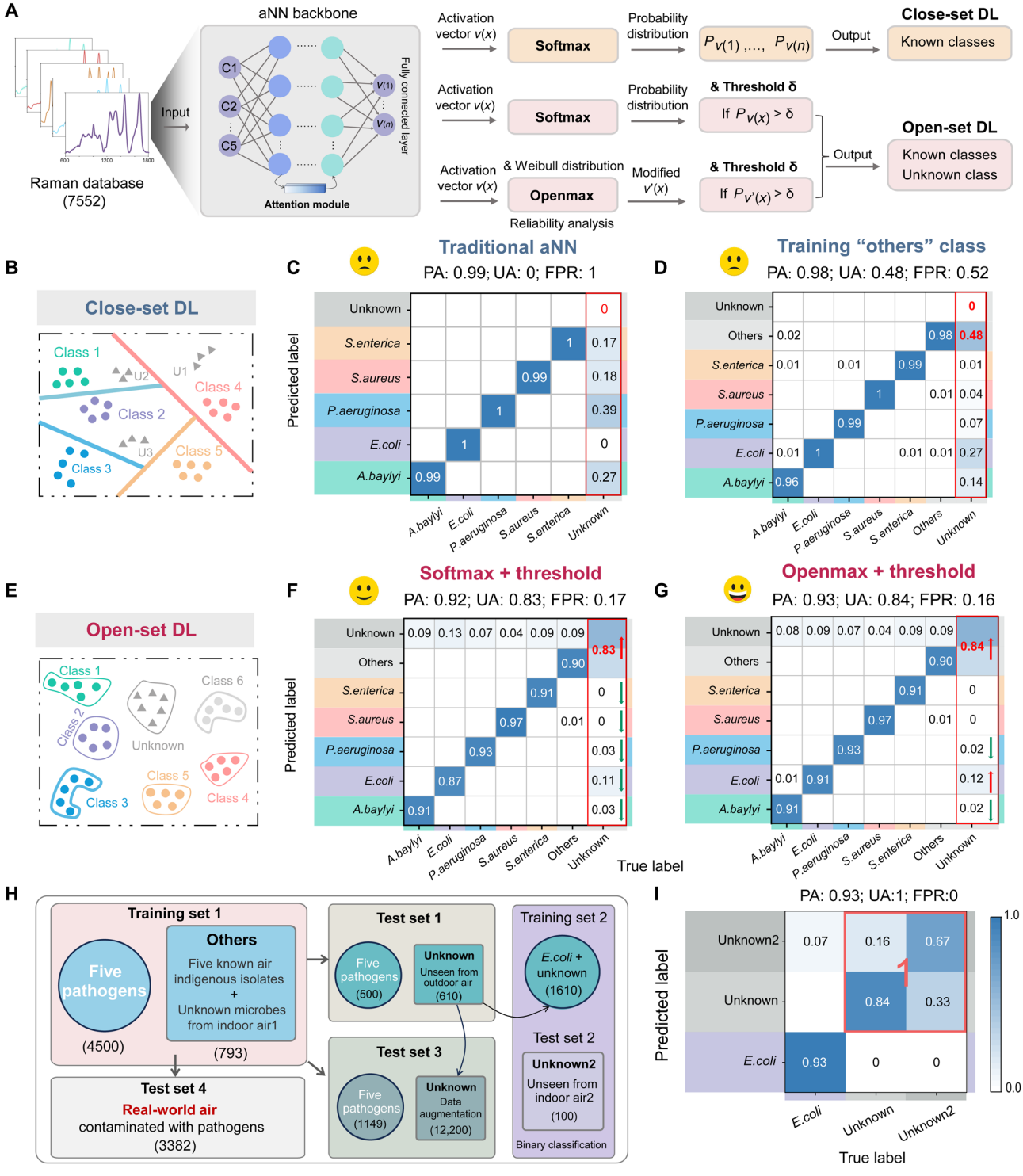

Fig. 3. Development of Raman-open set DL algorithm to identify pathogens and unknown classes in air. (**A**) Strategies for optimizing aNN algorithm from close set to open set. (**B**) Schematic diagrams of close-set DL. U1, U2, and U3 represent unknown classes in air environments. (**C** and **D**) Confusion matrix for five target airborne pathogens and unknown bacterial classes based on the close-set aNN before (C) and after (D) introducing new others class (H) to the training set. The other class represents a negative sample class including five nonpathogenic bacteria isolated from air and bacteria collected directly from indoor air without cultivation. The unknown class means air microbiota without containing the five target pathogens. (**E**) Schematic diagrams of OSDL. (**F** and **G**) Confusion matrix for five target airborne pathogens, others, and unknown classes based on two open-set aNN strategies of softmax + threshold (F) and openmax + threshold (G). (**H**) Composition of the dataset and the number of spectra used for model training and testing. Curved arrows indicate the source of the data set. All spectra in the test set were not used in the training set. (**I**) Retrained binary classification model to further distinguish *E. coli* from unknown bacteria classes. Unknown and Unknown2 were air bacteria obtained from two different locations in independent experiments. PA, pathogen accuracy; UA, unknown accuracy; FPR, false-positive rate.

unknown. When training and testing this open-set algorithm (i.e., softmax + threshold) using the same dataset as that in Fig. 3D, the accuracy of the unknown class significantly increased from 48 to 83% (Fig. 3F). Unfortunately, this improvement was accompanied by a decrease in the average accuracy for the five pathogens from 98 to 92%. Further, even at this high threshold of 0.98, some false-positive results had high-probability scores, and the false-positive rate was still 17%. These results indicate that merely adding a threshold for probabilities may not be sufficient to solve the open-set identification problem (*44*).

The second open-set strategy was to introduce an openmax function as an alternative to the softmax activation function into the final fully connected layer of the aNN (Fig. 3A) (*33*). Different from softmax, in openmax function, the activation vector of the penultimate layer was recalculated on the basis of Meta-Recognition and the Weibull distribution to identify system failure. Openmax can produce not only the probabilities of samples belonging to each known class but also the probabilities of samples not belonging to any known class, thus enabling prediction of unknown classes (Fig. 3E) (*45*). When performing this open-set aNN model using the "openmax + threshold" strategy (threshold: 0.98) with the same datasets, identification accuracies of 93 and 84% were achieved for the target pathogens and the unknown classes, respectively (Fig. 3G). Compared to the "softmax + threshold" strategy, the openmax + threshold strategy not only maintains high identification accuracy for most pathogens (90 to 97%) but also results in an additional 4% increase in the accuracy to identify *E. coli*, a 1% increase for identifying unknown, and a 1% decrease in the total false-positive rate. Ten-fold cross validation results confirmed the robustness of the open-set aNN model (fig. S4). Although the accuracy of 93% for the target pathogens is not as high as the close-set results (98 to 99%), this is understandable because the open-set model has to balance the identification of known classes with the rejection of unknown classes.

We further noticed that the false-positive rate of misidentifying nonpathogenic air bacteria to *E. coli* (12%) is much higher than the other four pathogens (0 or 2%) (Fig. 3G). The reason could be due to the likelihood that the Raman spectral signatures of some nonpathogenic air bacteria are highly similar to those of *E. coli*. To effectively distinguish *E. coli* from diverse unknown airborne bacteria, the *E. coli* and unknown class were trained a second time using a higher threshold of 0.996 (training set 2, Fig. 3H). Subsequently, we tested the model using a new unknown sample named "unknown2" collected from another indoor air setting (i.e., indoor air 2) (test set 2). This sample encompassed diverse spectra of airborne microbes but lacked the five target pathogens in the training set, which was verified by pure culture methods. As expected, the identification accuracies of *E. coli* and unknown samples were improved to 93 and 100%, respectively, and the false-positive rate declined to zero (Fig. 3I).

Together, these results indicate that the open-set aNN model can not only accurately identify target pathogens included in the training dataset (output as specific classes) but also reliably reject classes not present in the training dataset (output as unknown class). The Raman-OSDL method obviously outperforms the close-set approach and thus provides a novel means to detect pathogens in real-world air environments containing a large diversity of microbes.

Exploring the optimal parameter and detection limit of single-cell Raman-OSDL

It is notable that the thresholds used in the above softmax and openmax of the open-set aNN model were the optimized values we obtained. Hereby, we show how the model performance is affected by varying thresholds (Fig. 4A). The identification accuracy of the pathogen and unknown classes exhibited inverse trends with an increase in threshold. A larger threshold resulted in a rapid increase in the identification accuracy of unknown classes but a slow decrease in that of known pathogens. To simultaneously ensure a high accuracy (~90%) of both target pathogens and unknown classes, the threshold at 0.98 was selected as the optimal value to enable balanced performance of the open-set aNN model. However, the threshold is adjustable according to the risks of pathogens in practical applications. For example, for high-risk pathogens, to ensure a high true-positive rate of pathogens, the threshold needs to be set lower. In contrast, for daily surveillance of low-risk pathogens, to keep a low false-positive rate, the threshold needs to be set higher. Compared with softmax, openmax shows similar accuracy for pathogens but higher accuracy and lower error for the unknown class, especially under a lower threshold range (0.90 to 0.98). In addition, the accuracy and loss curves gradually converge with an increase in epoch number (Fig. 4C), thus suggesting that the 100 training epochs used here were sufficient to achieve a robust model.

Using the above-optimized parameters, we further explored the detection limit of the open-set aNN model. This study was conducted under the consideration that the abundance of pathogens in air may fluctuate wildly as an epidemic swells and recedes with time and location. For example, the pathogen/background bacteria ratio may be relatively low during the early stage of an outbreak but can increase to higher levels as an epidemic progress. This ratio also could be high in highly polluted air near feedlots and wastewater treatment plants but relatively low in ICU environments and typical home settings. Here, a total 30 test sets (3 different ratios × 10 mixed datasets) were created by artificially blending the single-cell Raman spectra of the five target pathogens with spectra of air microbes collected directly from air at decreasing ratios of 1:10, 1:100, and 1:1000, respectively (Fig. 4B). Notably, all the microbes in the test (test set 3, Fig. 3H) were collected from air, and their Raman spectra were never involved in model training. A total of 1149 pathogen spectra and 12,200 unknown spectra (augmented from 610 unknown spectra in test set 1, Fig. 3H) were used as the test set for randomly selected pathogens and unknowns according to different ratios. To ensure the reliability of the results, the sensitivities and specificities of the model for each class were determined by 10× random selection tests of spectra from pathogens and unknowns at different ratios (Fig. 4D). The sensitivity (true positives) curves for all five pathogens and the unknowns exhibited a nearly consistent trend without decrease in pathogen abundance down to 1:1000 (two-tailed Student's t test, $P > 0.05$). For the specificity (true negativity), except for the slight decrease in *Acinetobacter baylyi* from 100% at 1:10 to 99.97% at 1:1000 (two-tailed Student's t test, $P < 0.01$), the other five classes were not impaired by the decrease of pathogen abundance (two-tailed Student's t test, $P > 0.05$). The stable sensitivity and specificity indicate that Raman-OSDL can detect over a broad concentration range of pathogens. It indicates that even when just one pathogen is present in a milieu of up to 1000 nontarget microorganisms, it can be reliably identified. This

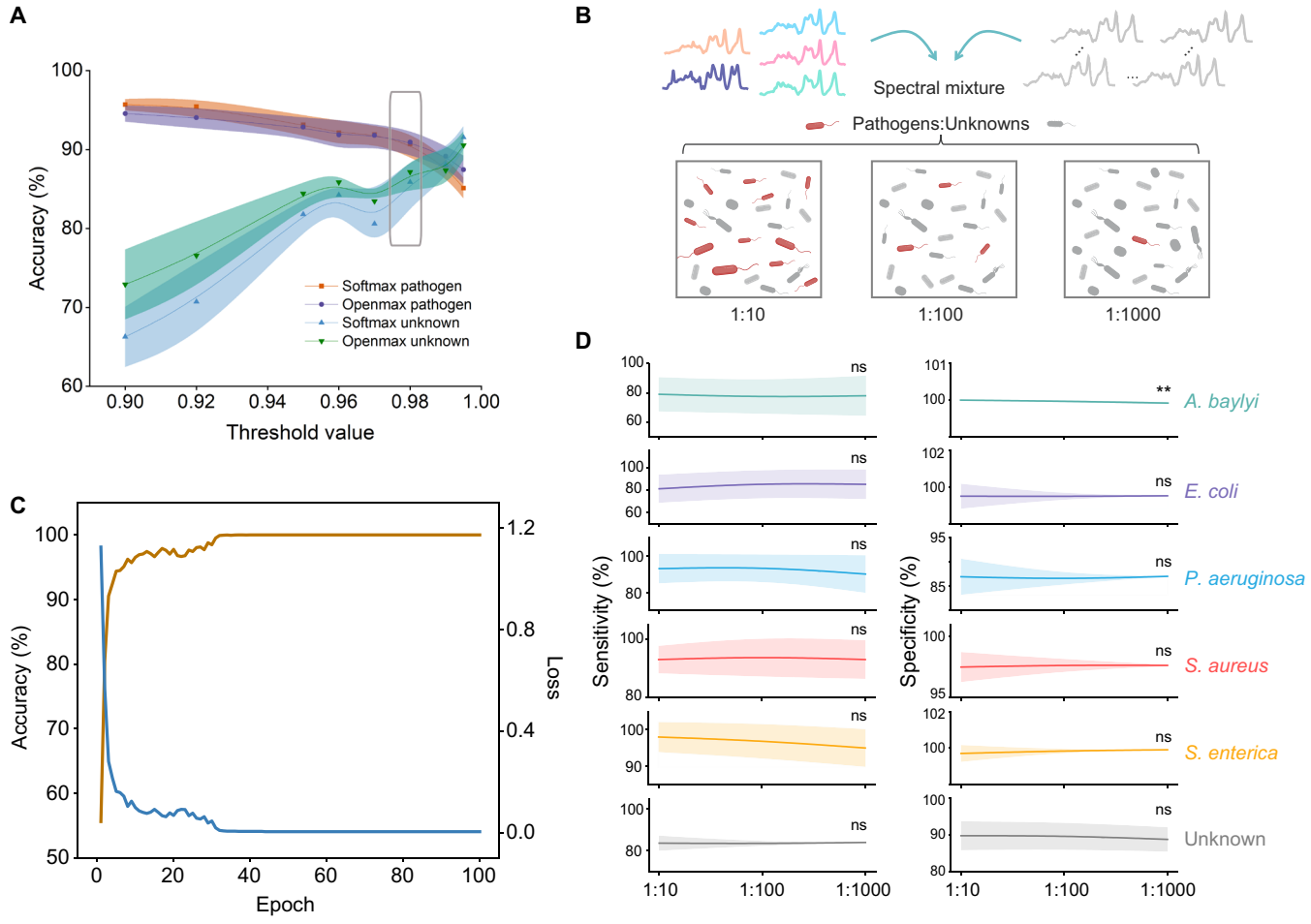

Fig. 4. Exploration of model parameters and detection limits. (**A**) Selection of threshold values for the probability used in softmax and openmax function of the open-set aNN model by considering the identification accuracy of both pathogens and unknown classes. (**B**) Schematic diagram illustrating the mixing of Raman spectra of pathogens and unknowns at ratios of 1:10 (10:100), 1:100 (10:1000), and 1:1000 (10:10000), respectively. The numbers in parentheses represent the actual number of spectra, which are randomly selected from the test set 3 in Fig. 3H. (**C**) Accuracy and loss curves of the open-set aNN model. (**D**) Exploration of the detection limit by calculating the sensitivity and specificity for the identification accuracy of target pathogens among background microbes at ratios shown in (B). **, significant difference observed between 1:10 and 1:1000 group (two-tailed Student's *t* test, $P < 0.01$). ns, no significant difference.

result is very important and is a prerequisite for the timely detection of pathogens.

Extending single-cell Raman-OSDL to identify pathogens in real-world air environments

To demonstrate that the Raman-OSDL approach can be extended to real-world air environments, we tested our model on two air samples spiked with either a single pathogen (*Salmonella enterica*) or five typical air pathogens (*S. aureus*, *E. coli*, *P. aeruginosa*, *S. enterica*, and *A. baylyi*). Notably, different from the above where we used artificially mixed spectra of pathogens and airborne microbes, here pathogenic aerosol was directly released into the real-world air with the diverse indigenous air microbes and then collected simultaneously. These indigenous microbes constituted >4600 different species as sequenced and annotated by metagenomic sequencing, indicating the high diversity of air microbial communities (fig. S5). Single-cell Raman spectra were then acquired without prior knowledge of the corresponding species identity. This situation well fits the expected real-world sampling scenarios.

A total of 2111 single-cell Raman spectra (test set 4 in Fig. 3H) were acquired from the air and then inputted into the pretrained Raman-OSDL model for testing. Because the approach was completely blinded and excluded any linkage between species identify and spectra, it was not possible to provide an accuracy value for identification. Instead, Raman-OSDL was used to classify target pathogens and unknown airborne bacteria based on their single-cell Raman spectra, and then this information was used to calculate the abundance of target pathogens within the airborne microbial community. Figure 5A shows the abundance results. For an air sample contaminated with *S. enterica*, the Raman-OSDL revealed that *S. enterica* was the dominant pathogen. A separate pathogen, *S. aureus*, was identified at a relatively low abundance in air. For reporting purposes that should avoid false positives and false alarms, we set a true-positive threshold for each pathogen based on the mean value +3 SD of the false-positive rate determined from the 10-fold cross validation of the open-set aNN model (Figs. 3G and 5A). The thresholds ranged from 0.5 to 4.2% for each pathogen are shown as the short red lines in Fig. 5 (A and B). In addition,

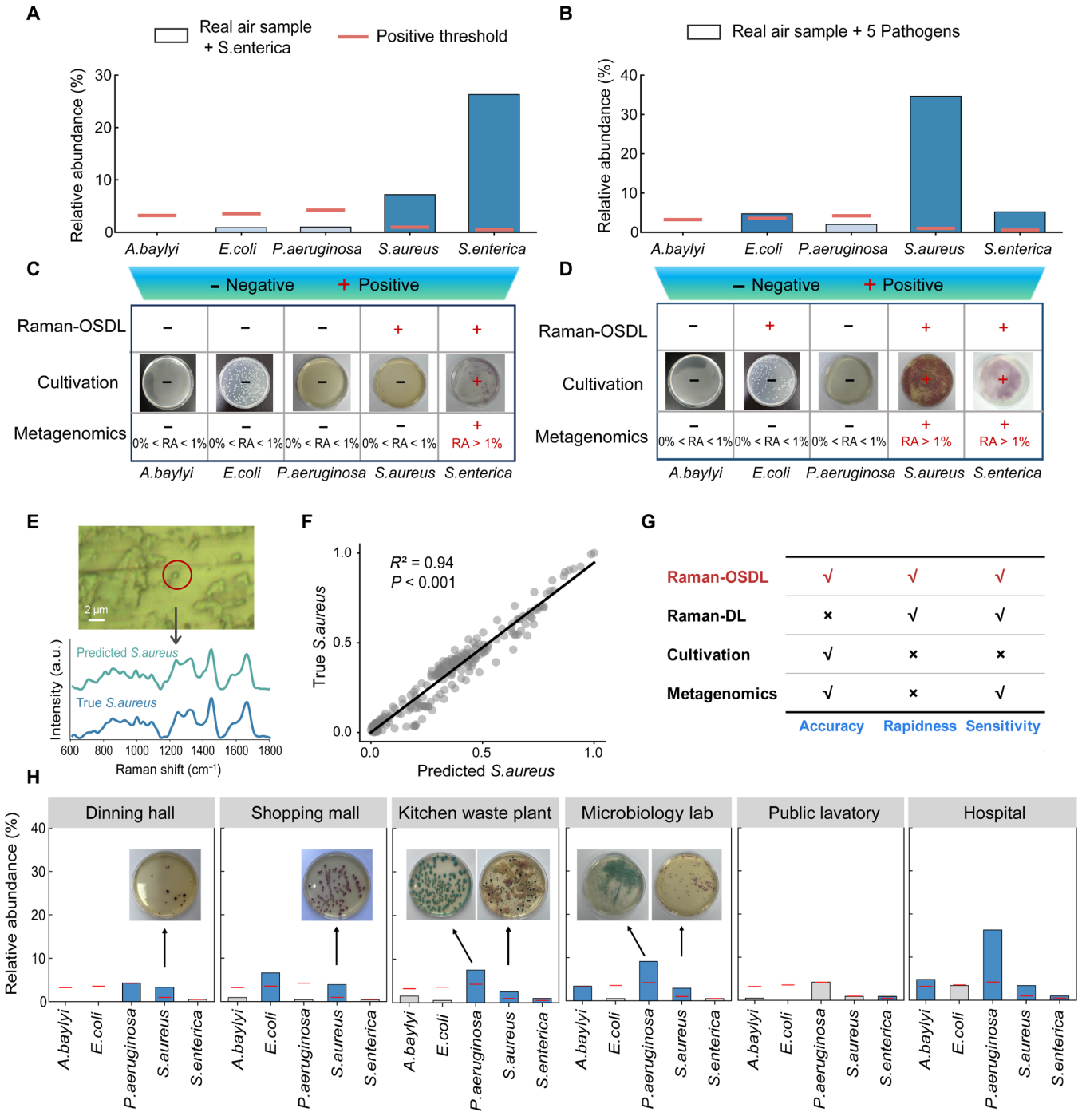

Fig. 5. Identification of single or multiple pathogens in real-world air samples. Relative abundance of the Raman-OSDL identified pathogen in real air samples contaminated with (**A**) single *S. enterica* and (**B**) five pathogens (i.e., *A. baylyi*, *E. coli*, *P. aeruginosa*, *S. aureus*, and *S. enterica*). Red lines represent the true-positive threshold for each pathogen, which is calculated as the mean value +3 SD from the false positive results in Fig. 3G. (**C** and **D**) Comparison of the identification results among Raman-OSDL, traditional cultivation and metagenomics methods in two air samples contaminated with *S. enterica* (C) and five pathogens (D). "– Negative": relative abundance (RA) < 1% or no specific colony; "+ Positive": relative abundance (RA) ≥ 1% or ≥ 1 specific colony. (**E**) Microscopic image of one predicted *S. aureus* cell (outlined by a red circle) on aluminum and comparison of its Raman spectra with that of a true *S. aureus* used in training. (**F**) Linear regression analysis of Raman spectra between the predicted and true *S. aureus* in (E). (**G**) Performance comparison between Raman-OSDL, Raman-close set DL, and two traditional methods used for real-world pathogen identification. (**H**) Identification results of six more real-world air environments using Raman-OSDL and traditional cultivation methods. Blue and gray columns represent the positive and negative Raman-OSDL results, respectively. The plate photographs only show positive results of the cultivation method.

considering that the outbreak of a pandemic usually requires pathogens to reach sufficient concentrations in airborne communities, thus we set 1% as the cutoff for a positive result based on the relative abundance of pathogens frequently observed in various air environments (*37*). Positives were reported only when the result for each pathogen exceeds both the true-positive threshold and the cutoff value simultaneously. On the basis of this criterium, both *S. enterica* and *S. aureus* were reported as positive in the air, while the other three pathogens were not.

To verify the results of our Raman-OSDL approach, the collected air sample was cultured on pathogen selective medium. This is the conventional way to identify pathogens based on the specific color of colonies (Fig. 5C and table S1). The colony of *S. enterica* with a specific color was detected, consistent with Raman-OSDL identification. However, none of the other four pathogens formed cultured colonies, including *S. aureus*. To explore the reasons for this discrepancy, we further compared the morphological and Raman spectral characteristics of the Raman identified *S. aureus* and a true *S. aureus* strain. The identified *S. aureus* in the air sample exhibits the typical spherical shape of *S. aureus* under the microscope (Fig. 5E). Moreover, its characteristic Raman peaks match well with those of the *S. aureus* used for training. The corresponding correlation coefficient between the two sets of Raman spectra was 0.94 ($P < 0.001$; Fig. 5F) and was much higher than that with the other four true pathogens (0.76, 0.79, 0.78, and 0.87) (fig. S6). To further confirm the results, metagenomic sequencing of this air sample was performed, and we found *S. aureus* in the metagenome with a relative abundance of <1%. The apparent absence of this species following culture may be due to their subsistence in a VBNC state or low abundance that results in the bacteria being challenging to isolate via plating on agar. These results demonstrate the accuracy of Raman-OSDL in identifying *S. aureus*, even when present in the VBNC state or at low abundance.

We further verified the performance of Raman-OSDL in testing an air sample contaminated with five pathogens. Using the same reporting criteria, the positive presence of *S. aureus* and *S. enterica* in air was consistently reported by Raman-OSDL, metagenomics, and pure culture (Fig. 5, B and D). For *E. coli*, positive results were reported by Raman-OSDL with a relative abundance >1%, but not via culture (no specific blue colony), or metagenomic sequencing with a low abundance (<1%) (fig. S5). For the other two airborne pathogens, *A. baylyi* and *P. aeruginosa*, all three methods produced negative results (relative abundance <1% or no specific colony), indicating that the abundances of these two pathogens in the air environment were too low to be reported as positive.

To test the generalizability of Raman-OSDL, we further extended the test to six more real-world air samples (1271 spectra, test set 4 in Fig. 3H) from a range of environments in which the presence of pathogens is completely unknown, including dining hall, shopping mall, kitchen waste plant, microbiology laboratory, public lavatory, and hospital. Both culture-dependent chromogenic method and culture-independent single-cell Raman method were used to detect pathogens from these air samples (Fig. 5H). One to four types of pathogens were identified in these samples via both methods. In addition, for all the positive pathogens identified by cultivation method, Raman method reported consistent pathogen-positive results, while one additional Raman-reported pathogen was not observed to grow into colonies in some air samples. The reason could be due to their too low abundance to be isolated via plating or in dead or VBNC state in response to stresses such as disinfection. These results demonstrated the generalizability of Raman combined with OSDL in identifying pathogens in diverse air environments.

Notably, using the RAPD platform and the pretrained Raman-OSDL model, the entire process from air sampling to obtaining test results requires just 1 hour for pathogens at abundance of >1%, including air sampling (<20 min, 1000 liters), pretreatment (<20 min, washing three times), Raman measurement (<20 min, ~400 spectra), and identification (<1 min). This end-to-end RAPD platform operates at a speed more than 10× faster than traditional culture methods (>12 hours to several days) and metagenomic sequencing (several days) (*21*, *22*). Compared with traditional methods, the Raman-OSDL shows better comprehensive performance in terms of accuracy, rapidness, and sensitivity (Fig. 5G).

DISCUSSION

Here, we advance single-cell Raman spectroscopy as a means for rapid, accurate, and culture-free identification of airborne pathogens in real-world air environments. These heterogeneous matrices contain not only pathogens but also diverse unknown indigenous bacteria that cannot be fully included in a training model. This effort represents a breakthrough in Raman-based pathogen identification as it transitions from pure cultures to real-world environmental settings. This transition is achieved by solving two challenges. The first is the development of an open-set aNN DL model that enables not only effective identification of known target pathogens but also rejection of diverse unknown airborne bacteria that previously generated misclassification and false positives. The second is the design of an MAGC device that enables construction of a Raman spectral training dataset of bacteria in aerosol states that improves identification accuracy. The established Raman-OSDL method performs airborne pathogen identification within 1 hour from air sampling and single-cell Raman acquisition to final classification output. The sensitivity is down to 1:1000, enabling identification of a broad concentration range of target pathogens, including those at low abundances in air environments. We envisage our approach as a rapid and sensitive tool for surveilling and identifying airborne pathogens in indoor and outdoor environments that will assist in prevention of airborne infection transmission.

Direct identification of pathogens from complex environmental microbial communities has long been a major challenge for close-set DL-based Raman spectral identification. Environmental microbiota are highly complex and include diverse and dynamically changing microorganisms. The air investigated in this study harbors almost all major bacteria phyla (e.g., Proteobacteria, Bacteroidetes, Actinobacteria, and Firmicutes) (fig. S5). Moreover, >99% of natural microorganisms have been reported as unculturable in a laboratory setting (*30*, *31*). Such high diversity and high unculturability make it impossible to get pure isolates of all airborne bacteria such that they can be included in Raman spectral training models. Although traditional close-set DL-based Raman spectroscopy exhibits good ability when all the testing classes are known during training, unseen/unknown classes not encountered during training emerge frequently in real-world environments, thus causing high false-positive rates that greatly limits the application of close-set DL.

Here, inspired by the advent of OSDL as a means to address real-world image recognition challenges (*33*), we proposed two OSDL

strategies of softmax + threshold and openmax + threshold. Their performances were tested and optimized for Raman identification of five typical pathogens collected from air in the presence of multiple unknown airborne species. For the openmax + threshold strategy, the average identification accuracy of the five airborne pathogens reached up to 93%, closely matching or even surpassing previous close-set ML accuracies for pure-cultured isolates (table S2) (*28*, *46*). This open-set method achieved 84% accuracy in identifying previously unseen airborne bacterial classes and reduced the false-positive rate by 36% in comparison with the conventional close-set DL. This is an obvious improvement and demonstrates the success of Raman-OSDL in avoiding interference from unseen environmental microorganisms and its applicability in real-world pathogen identification. Moreover, in cases wherein some nonpathogenic bacteria may exhibit very similar spectral characteristics to the target pathogen, such as *E. coli* shown here (Fig. 3G), we proposed a further strategy of employing a binary classification model with a higher threshold to distinguish *E. coli* from other unknown bacteria and decreased the false-positive rate from 12 to 0% (Fig. 3I). With the established training model, we successfully and simultaneously identified single and multiple pathogens directly collected from eight real-world air environments despite having completely blind information on the linkage between the collected spectra and microbial species. The identification results were cross validated by both metagenomic sequencing and culture-based methods, fully demonstrating its accuracy. Notably, the whole Raman dataset includes a total of ~23,000 spectra from target pathogens, nonpathogenic air bacteria, and eight real-world air microbiota. The associated bacterial diversity and dataset size are higher than most previous works (*28*, *46*–*48*).

The developed single-cell Raman-OSDL method displays advantages in its rapidness and high sensitivity. It takes approximately 1 hour for pathogens at abundance of >1% from air sampling to classification output, at least 10× faster than culture-based methods (*12*, *14*). Single-cell detection allows direct identification of indigenous airborne bacteria without lengthy cultivation. Compared to metagenomic sequencing, it requires much fewer cell numbers and less air volume, thereby reducing the overall air sampling time from several hours to minutes. Moreover, single-cell resolution enables identification of airborne bacteria in the VBNC state that often constitutes an important reservoir of pathogens in the environment (*49*). The detection sensitivity is down to a pathogen concentration of 1:1000, outperforming culture-based methods in effectively detecting low-abundance pathogens in practical settings.

The established RAPD platform relies upon an open-source OSDL algorithm and commercially available components that can be easily used by people lacking expertise in microbiology or computer science. In addition, the aNN DL algorithm used in this work has been previously demonstrated to be able to identify pathogen-derived extracellular vesicles down to species and even strain level (*47*). Thus, it should not be a big challenge to use the present RAPD platform to identify pathogens at a higher species or even strain levels. RAPD also allows for easy generalizability to other pathogenic bacteria and fungi in air environments. Users can establish training libraries for Raman spectra of any airborne pathogens of concern and can identify them in various air environments. Beyond the airborne aerosol, the idea of single-cell Raman combined with open-set identification may be extendable to the detection of microorganisms in other complex environmental matrices, such as soil, water, and the gut. Moreover, the open-set algorithm may have broad applications in other spectra-based identification, such as mass spectra (*50*), nuclear magnetic resonance spectra (*51*), and Fourier transform infrared spectra (*52*).

To achieve real-time surveillance of pathogens in practical applications, sample throughput and automatization require future improvements. High throughput potentially can reduce the time required to acquire large numbers of single cells from an environment and can accelerate the identification of low-abundance pathogens. Our throughput using the current micro-Raman spectrometer can be approximately 1000 cells/hour (3600/3 = 1200). Using a state-of-art Raman detector, the shortest time to acquire Raman spectra of single bacteria can be 0.01 s, thus potentially increasing the throughput to >300,000 cells/hour (*46*). For automation, recently developed Raman spectrometer that combines microfluidics, positive dielectrophoresis, and software has enabled automatic sample introduction and single-cell Raman acquisition (*53*, *54*). In a preliminary study, this system enabled automatic acquisition of high-quality spectra and classification of five airborne pathogens used in this study (fig. S7). Our RAPD hardware can potentially be integrated with it to automatically monitor airborne pathogens and improve detection speed and throughput. These improvements represent a promising future direction for deploying Raman-OSDL for rapid surveillance of environmental pathogens.

To the best of our knowledge, this study marks the first application of Raman-OSDL for rapid, accurate, and simultaneous identification of multiple pathogens in complex real-world environments. Such a technique is urgently needed for early warning and safeguarding against potential outbreaks and the spread of biosafety concerns in various environments. Raman-OSDL represents a major advancement for this purpose under One Health perspective.

MATERIALS AND METHODS

Pathogen-laden aerosol generation and collection

Five model bacterial pathogen isolates—including *A. baylyi*, *E. coli*, *P. aeruginosa*, *S. aureus*, and *S. enterica*—were cultured in Luria-Bertani (LB) media at 37°C overnight as preparation for microbial aerosol generation (see table S3 for full isolate information). These pathogens represent the airborne pathogens commonly detected in aerosols collected from high-risk environments such as hospitals and open wastewater canals (*39*, *55*, *56*). Inspired by previous efforts (*57*), we constructed the present MAGC device for pathogen-laden aerosol generation and collection. The MAGC device is composed of four modules that perform different functions, including the microbial aerosol generation module (g), air sampling module (f), sterilization module (j), and the ventilation system (i, e, and d) (Fig. 1 and fig. S1). By using the MAGC device, air samples containing specific pathogens can be obtained to construct a pathogenic bacterial Raman dataset. The device workflow is as follows:

1) The sterilization module (j) sterilizes the air in the device for 30 min using ultraviolet light.

2) The ventilation system (i, e, and d) works for 10 min to remove residual microbial cell material. Concurrently, clean outside air is brought into the device after removing outside microbes using a 0.22-μm polycarbonate filter.

3) The aerosol generation module (g) works to convert liquid pathogen cultures into aerosols and releases them into the air.

4) After suspended in the air for an hour, microbial aerosol was collected by the air sampling module (f, at a rate of 50 liter min^{-1}) with a 3-μm (pore size) gelatine filter membrane (12602-80-ALK, Sartorius stedim biotech). To avoid the effect of gelatine on bacterial Raman signal, the gelatine membrane was then dissolved in sterile water, in which the microbes was centrifuged and washed with sterile water to leave only bacteria for further Raman spectral and metagenomic analysis. We further compared the Raman spectra of (i) bacteria obtained via the above processes, (ii) gelatine dissolved in water, (iii) potential gelatine residue after washing, and (iv) aluminum foil substrate using the same acquisition time (fig. S2). Raman signal of bacteria is much stronger than the dissolved gelatine, and the Raman signal of potential gelatine residue after washing [air-dried on aluminum (Al) foil] almost has no difference with aluminum foil substrate. These results indicated that the gelatine after washing has no contribution to bacterial Raman signal. Before collecting the next pathogenic aerosol, the MAGC device was sterilized and ventilated for 30 min, respectively.

Single-cell Raman spectroscopy measurements and data pre-processing

An aliquot of 3 μl of air-suspended/liquid pathogen samples after washed by sterile water was immediately spotted on an Al foil substrate and air-dried at room temperature. Raman spectra of the dried samples were measured using a LabRAM Aramis (HORIBA Jonin-Yvon, Japan) confocal micro-Raman system with a 532-nm Nd:YAG (Yttrium Aluminium Garnet) excitation laser and with a grating of 300 g/mm. A 100× objective (Olympus, 0.90 numerical aperture) was used for Raman spectra acquisition. A 30 μm–by–30 μm *XY* map was taken using the mapping mode, with 3-μm spacing between spots to avoid overlap between different cells. Most spectra are acquired at the single-cell level due to the similar size between the laser spot and the bacterial cell (~1 μm). To maintain a consistent acquisition condition for all bacteria, the acquisition time for each spot was 3 s with three replicates (3 s × 3). Actually, an acquisition time of 3 s for once (3 s × 1) can already produce highly reproducible spectra (fig. S8). A small part of the spectra taken from multilayer or nonbacterial regions were excluded. The spectral range between 600 and 1800 cm^{-1} was used as the fingerprint for bacterial identification. The baseline of the spectra was automatically corrected, and the spectral intensity was normalized between 0 and 1 using Python scripts. In addition, only the spectra in fig. S7 were acquired using a Raman Flow Cytometry named FlowRACS (Qingdao Single-cell Biotech, China) as a preliminary study for automatic acquisition of spectra (*54*).

Raman dataset

The Raman dataset consists of spectra for the five airborne pathogens (i.e., *A. baylyi*, *E. coli*, *P. aeruginosa*, *S. aureus*, and *S. enterica*), five nonpathogenic isolates from air (i.e., *Exiguobacterium acetylicum*, *P. megaterium*, *Bacillus velezensis*, *B. cereus*, and *S. lentus*), and unknown microbes collected from real indoor and outdoor air environments. The Raman dataset contains 7552 spectra including 6149 spectra for the five airborne pathogens (>1000 for each class) and 1403 spectra for the unknown microbes collected at three measurement times. The above-unknown spectra contain various microbial species from air but do not contain the target five pathogens (verified by cultivation). In addition, another 3382 spectra were collected from two real-world air samples spiked with pathogens and six more real-world air samples from a range of environments in which the presence of pathogens is completely unknown, including dining hall, shopping mall, kitchen waste plant, microbiology laboratory, public lavatory, and hospital. All of the spectra were divided into six parts including two training sets and four test sets (Fig. 3H). Training set 1 and test set 1 are used to optimize and test the DL algorithm from close-set to open-set. For one of pathogens that show high false-positive rate even with openmax + threshold strategy, a further binary training and testing strategy (training set 2 and test set 2) with a higher threshold was developed to decrease misidentification of nonpathogenic air bacteria to *E. coli*. The purpose of test set 3 with many more spectra is to explore the detection limit of the above optimized DL model. Please note by now, training set and test set 1 to test set 3 are all from spectra with identity labels so as to calculate the identification accuracy. Last, with the established training model, test set 4 from bacteria in real air samples (without identity label) was collected to test its performance in complex real-world microbiota.

Traditional machine learning model

Three classical machine learning models, LDA (*58*), SVM (*59*), RF (*60*), and one DL model, aNN (*47*), were tested to compare their performance in airborne pathogen identification. As previously reported (*47*), the aNN model had better performance at bacterial identification than traditional algorithms and was optimized on the basis of traditional convolutional neural network (CNN) (*28*) by the addition of an attention module. According to the results of the confusion matrix and the receiver operating characteristic (ROC) curve, the best performing model was used for subsequent pathogen identification of real-world air samples.

Open-set aNN model and training details

The open-set aNN model used in this work is based on the innovative integration of a previously reported aNN architecture (*47*) and the OSDL (*33*) algorithms. The innovation of this study is not the algorithm itself, but the perfect combination of the two to effectively solve new practical environmental problems. The construction of aNN includes four convolution modules, four attention modules, and a fully connected layer (Fig. 1). The convolution block comprises four layers: convolution, batch normalization, pooling, and activation. The attention module was designed to enhance the significance of specific features within hidden layers through adaptive weighting, thereby amplifying the importance of crucial features. The aNN model with the attention module can better extract the difference between bacterial Raman spectra in two dimensions including channel attention and wave number attention. Detailed mechanisms of the attention module were described in our previous work (*47*).

Traditional DL methods, including aNN, fall under the close-set condition, whereby all testing classes are known at time of training. This is often not applicable to real-world environments containing complex unknown classes. This is mainly due to the close nature of the softmax layer of traditional DL algorithms. The softmax activation function can produce a probability distribution over known classes and classify the input into one of the known classes with the highest probability, even if the probability of each class is low (Fig. 3A). To adapt the aNN model into OSDL, a threshold δ was added to the maximum probability of the softmax output (softmax + threshold). Although the softmax + threshold strategy may improve OSDL performance somewhat, it is still not sufficient to identify unknown classes. A previous study has shown that a false-positive result may also produce high-probability scores (*44*). Considering this

situation, we further adapted the aNN model by introducing a new model layer, openmax (*33*), to replace the softmax layer. Compared with the softmax layer, openmax can evaluate the probability that the input belongs to each known class and the probability that it does not belong to any known class (i.e., the probability of unknown class) (*33*). The openmax layer modified the activation vector from the fully connected layer and then created a probability distribution of each class based on the Weibull distribution, which can check the reliability of the result (*60*). The input will be classified as the known class only if the maximum probability does not belong to the unknown class and is higher than the threshold value; otherwise, it will be classified as the unknown class (*45*). By removing the restriction that the sum of probabilities of known classes is 1 and rejecting inputs that are far from known classes, openmax can correctly handle unknown/unseen classes during identification. The detailed steps for openmax computation are shown below in Algorithm 1 (*33*)

Algorithm 1. Detailed steps for openmax computation.

Require: Activation vector (AV) for $\mathbf{v}(x) = v_1(x),\cdots, v_n(x)$

Require: means μ_j and libMR models $\rho_j = (\tau_i, \lambda_i, \kappa_i)$

Require: α, the number of "top" classes to revise

Let $s(i) = \text{argsort}(v_j(x))$; Let $\omega_j = 1$

for $i = 1,\ldots, \alpha$ do

$$\omega_{s(i)}(x) = 1 - \frac{\alpha - i}{\alpha} e^{-\left(\frac{\|x-\tau_{s(i)}\|}{\lambda_{s(i)}}\right)^{\kappa_{s(i)}}}$$

end for

Revise activation vector $v'(x) = v(x) \circ \omega(x)$

Define $v'_0(x) = \sum_i v_i(x)(1-\omega_i(x))$

$$P'(y=j\,|\,x) = \frac{e^{v'_j(x)}}{\sum_{i=0}^{n} e^{v'_i(x)}}$$

Let $y^* = \text{argmax}_j\, P'(y=j\,|\,x)$

Reject input if $y^* == 0$ or $P'(y=y^*\,|\,x) < \delta$

where mean μ_j is the mean activation vector over only the correctly classified training samples. For each class j, ρ_j represents an estimation value of the probability of an input being an outlier based on the extreme value theory and Weibull distribution. Parameters $\tau_i, \lambda_i, \kappa_i$ are associated with the data shifting, Weibull scale, and shape, respectively. The weight (ω) for the α largest activation classes was computed and used to scale the Weibull distribution probability. The v' is the AV after revised and P' is the final openmax probability. δ is the threshold of the model.

The aNN model was trained using the Stochastic Gradient Descent (SGD) optimizer with learning rate of 0.1, momentum of 0.9, weight decay of 5×10^{-4}, and batch size of 64. The predefined training epoch is 100 and 10-fold cross validation was conducted to verify the robustness of the model. To test the ability of the model to address real-world situations, all classes during training were known, while both known and unknown classes existed during testing.

Validation using real air samples contaminated with pathogens

To validate the performance of the Raman-OSDL technology in the real world, pathogenic aerosols were directly released to a real-world air environment containing original indigenous microbes to produce two air environments contaminated with either single- or multiple-pathogens. The first air sample dataset, consisting of ~900 spectra, was collected from the air environment contaminated with a single pathogen (i.e., *S. enterica*). The second air sample dataset, consisting of ~1200 spectra, was collected from an air environment contaminated with multiple pathogens (i.e., *A. baylyi*, *E. coli*, *P. aeruginosa*, *S. aureus*, and *S. enterica*). To ensure the reliability of the results, we used the pretrained model and performed the same procedure without any fine-tuning adjustment. The detection results of the real air samples were verified by traditional methods, including cultivation and metagenomic sequencing. Cultivation was conducted using CHROMagar chromogenic media (CHROMagar, Paris, France) for each target pathogen (table S1). Colonies with specific colors were picked and used for full-length sequencing of the *16S rRNA* gene using primers 27F (5′-AGAGTTTGATCCTGGCTCAG-3′) and 1492R (5′-TACGGYTACCTTGTTACGACTT-3′) to further species validation. Moreover, metagenomic sequencing was used to analyze the microbial composition of pathogen-contaminated air samples according to the previous procedure (*61*). Last, another six real-world air samples (~1200 spectra) from a range of environments in which the presence of pathogens was completely unknown were used to demonstrate the generalizability of the method in different air environments.

Statistical analysis

A two-tailed Student's *t* test was implemented in SPSS 25.0 to test the statistical significance between two groups with P values < 0.05 considered significant.

Supplementary Materials

This PDF file includes:

Figs. S1 to S8

Tables S1 to S3

REFERENCES AND NOTES

1. K. E. Jones, N. G. Patel, M. A. Levy, A. Storeygard, D. Balk, J. L. Gittleman, P. Daszak, Global trends in emerging infectious diseases. *Nature* **451**, 990–993 (2008).
2. GBD 2019 Antimicrobial Resistance Collaborators, Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet* **400**, 2221–2248 (2022).
3. C. C. Wang, K. A. Prather, J. Sznitman, J. L. Jimenez, S. S. Lakdawala, Z. Tufekci, L. C. Marr, Airborne transmission of respiratory viruses. *Science* **373**, eabd9149 (2021).
4. J. Wei, Y. Li, Airborne spread of infectious agents in the indoor environment. *Am. J. Infect. Control* **44**, S102–S108 (2016).
5. S. Herfst, M. Böhringer, B. Karo, P. Lawrence, N. S. Lewis, M. J. Mina, C. J. Russell, J. Steel, R. L. de Swart, C. Menge, Drivers of airborne human-to-human pathogen transmission. *Curr. Opin. Virol.* **22**, 22–29 (2017).
6. E. Huynh, A. Olinger, D. Woolley, R. K. Kohli, J. M. Choczynski, J. F. Davies, K. Lin, L. C. Marr, R. D. Davis, Evidence for a semisolid phase state of aerosols and droplets relevant to the airborne and surface survival of pathogens. *Proc. Natl. Acad. Sci. U.S.A* **119**, e2109750119 (2022).
7. S. Faridi, K. Naddafi, H. Kashani, R. Nabizadeh, M. Alimohammadi, F. Momeniha, S. Faridi, S. Niazi, A. Zare, A. Gholampour, M. Hoseini, Z. Pourpak, M. S. Hassanvand, M. Yunesian, Bioaerosol exposure and circulating biomarkers in a panel of elderly subjects and healthy young adults. *Sci. Total Environ.* **593-594**, 380–389 (2017).
8. R. E. Stockwell, E. L. Ballard, P. O'Rourke, L. D. Knibbs, L. Morawska, S. C. Bell, Indoor hospital air and the impact of ventilation on bioaerosols: a systematic review. *J. Hosp. Infect.* **103**, 175–184 (2019).
9. S. Chen, Y.-W. Su, J. Sun, T. Chen, Y. Zheng, L.-J. Sui, S. Yang, C. Liu, P. Wang, T. Li, Q. Chi, H. Sun, J. Chen, B.-Q. Xu, Z. Huang, Y. Fang, Label-free single-particle imaging approach

for ultra-rapid detection of pathogenic bacteria in clinical samples. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2206990119 (2022).
10. L. C. J. Alcantara, L. Amenga-Etego, R. Andersson, M. Bhaumik, Y. K. Choi, H. Decaluwe, J. Geoghegan, B. L. Haagmans, S. López, M. M. Mukhtar, E. Nelwan, E. A. Rahal, K. Sato, E. H. Sklan, Y. S. C. Fang, Methods for fighting emerging pathogens. *Nat. Methods* **19**, 395–397 (2022).
11. P. Rajapaksha, A. Elbourne, S. Gangadoo, R. Brown, D. Cozzolino, J. Chapman, A review of methods for the detection of pathogenic microorganisms. *Analyst* **144**, 396–411 (2019).
12. L. Váradi, J. L. Luo, D. E. Hibbs, J. D. Perry, R. J. Anderson, S. Orenga, P. W. Groundwater, Methods for the detection and identification of pathogenic bacteria: past, present, and future. *Chem. Soc. Rev.* **46**, 4818–4832 (2017).
13. A. Bizzini, G. Greub, Matrix-assisted laser desorption ionization time-of-flight mass spectrometry, a revolution in clinical microbial identification. *Clin. Microbiol. Infect.* **16**, 1614–1619 (2010).
14. R. Patel, MALDI-TOF MS for the Diagnosis of Infectious Diseases. *Clin. Chem.* **61**, 100–111 (2015).
15. J. D. Oliver, Recent findings on the viable but nonculturable state in pathogenic bacteria. *FEMS Microbiol. Rev.* **34**, 415–425 (2010).
16. C. Klumpp-Thomas, H. Kalish, M. Drew, S. Hunsberger, K. Snead, M. P. Fay, J. Mehalko, A. Shunmugavel, V. Wall, P. Frank, J.-P. Denson, M. Hong, G. Gulten, S. Messing, J. Hicks, S. Michael, W. Gillette, M. D. Hall, M. J. Memoli, D. Esposito, K. Sadtler, Standardization of ELISA protocols for serosurveys of the SARS-CoV-2 pandemic using clinical and at-home blood sampling. *Nat. Commun.* **12**, 113 (2021).
17. L. Cui, D. Zhang, K. Yang, X. Zhang, Y.-G. Zhu, Perspective on Surface-Enhanced Raman Spectroscopic Investigation of Microbial World. *Anal. Chem.* **91**, 15345–15354 (2019).
18. H. Zhang, X. Ma, Y. Liu, N. Duan, S. Wu, Z. Wang, B. Xu, Gold nanoparticles enhanced SERS aptasensor for the simultaneous detection of Salmonella typhimurium and Staphylococcus aureus. *Biosens. Bioelectron.* **74**, 872–877 (2015).
19. Z. Yang, Low-cost and rapid sensors for wastewater surveillance at low-resource settings. *Nat. Water* **1**, 405–407 (2023).
20. W. Witkowska McConnell, C. Davis, S. R. Sabir, A. Garrett, A. Bradley-Stewart, P. Jajesniak, J. Reboud, G. Xu, Z. Yang, R. Gunson, E. C. Thomson, J. M. Cooper, Paper microfluidic implementation of loop mediated isothermal amplification for early diagnosis of hepatitis C virus. *Nat. Commun.* **12**, 6994 (2021).
21. W. Gu, X. Deng, M. Lee, Y. D. Sucu, S. Arevalo, D. Stryke, S. Federman, A. Gopez, K. Reyes, K. Zorn, H. Sample, G. Yu, G. Ishpuniani, B. Briggs, E. D. Chow, A. Berger, M. R. Wilson, C. Wang, E. Hsu, S. Miller, J. L. DeRisi, C. Y. Chiu, Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids. *Nat. Med.* **27**, 115–124 (2021).
22. C. Y. Chiu, S. A. Miller, Clinical metagenomics. *Nat. Rev. Genet.* **20**, 341–355 (2019).
23. D. Wang, P. He, Z. Wang, G. Li, N. Majed, A. Z. Gu, Advances in single cell Raman spectroscopy technologies for biological and environmental applications. *Curr. Opin. Biotechnol.* **64**, 218–229 (2020).
24. N. M. Ralbovsky, I. K. Lednev, Towards development of a novel universal medical diagnostic method: Raman spectroscopy and machine learning. *Chem. Soc. Rev.* **49**, 7428–7453 (2020).
25. K. S. Lee, Z. Landry, F. C. Pereira, M. Wagner, D. Berry, W. E. Huang, G. T. Taylor, J. Kneipp, J. Popp, M. Zhang, J. X. Cheng, R. Stocker, Raman microspectroscopy for microbiology. *Nat. Rev. Method. Primers* **1**, 80 (2021).
26. W. Lu, X. Chen, L. Wang, H. Li, Y. V. Fu, Combination of an artificial intelligence approach and laser tweezers raman spectroscopy for microbial identification. *Anal. Chem.* **92**, 6288–6296 (2020).
27. S. Guo, J. Popp, T. Bocklitz, Chemometric analysis in Raman spectroscopy from experimental design to machine learning–based modeling. *Nat. Protoc.* **16**, 5426–5459 (2021).
28. C.-S. Ho, N. Jean, C. A. Hogan, L. Blackmon, S. S. Jeffrey, M. Holodniy, N. Banaei, A. A. E. Saleh, S. Ermon, J. Dionne, Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. *Nat. Commun.* **10**, 4927 (2019).
29. S. Kloß, P. Rösch, W. Pfister, M. Kiehntopf, J. Popp, Toward culture-free raman spectroscopic identification of pathogens in ascitic fluid. *Anal. Chem.* **87**, 937–943 (2015).
30. U. Hofer, The majority is uncultured. *Nat. Rev. Microbiol.* **16**, 716–717 (2018).
31. K. G. Lloyd, A. D. Steen, J. Ladau, J. Yin, L. Crosby, Phylogenetically novel uncultured microbial cells dominate earth microbiomes. *mSystems* **3**, e00055-18 (2018).
32. W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, T. E. Boult, Toward Open Set Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1757–1772 (2013).
33. A. Bendale, T. E. Boult, Towards open set deep networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), pp. 1563–1572.
34. M. Wang, T. Lin, L. Wang, A. Lin, K. Zou, X. Xu, Y. Zhou, Y. Peng, Q. Meng, Y. Qian, G. Deng, Z. Wu, J. Chen, J. Lin, M. Zhang, W. Zhu, C. Zhang, D. Zhang, R. S. M. Goh, Y. Liu, C. P. Pang, X. Chen, H. Chen, H. Fu, Uncertainty-inspired open set learning for retinal anomaly identification. *Nat. Commun.* **14**, 6757 (2023).
35. A. R. K. Gollakota, S. Gautam, M. Santosh, H. A. Sudan, R. Gandhi, V. Sam Jebadurai, C.-M. Shu, Bioaerosols: Characterization, pathways, sampling strategies, and challenges to geo-environment and health. *Gondwana Res.* **99**, 178–203 (2021).
36. H. Li, X.-Y. Zhou, X.-R. Yang, Y.-G. Zhu, Y.-W. Hong, J.-Q. Su, Spatial and seasonal variation of the airborne microbiome in a rapidly developing city of China. *Sci. Total Environ.* **665**, 61–68 (2019).
37. J. Zhao, L. Jin, D. Wu, J. Xie, J. Li, X. Fu, Z. Cong, P. Fu, Y. Zhang, X. Luo, X. Feng, G. Zhang, J. M. Tiedje, X. Li, Global airborne bacterial community—interactions with Earth's microbiomes and anthropogenic activities. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2204465119 (2022).
38. J.-L. Vincent, J. Rello, J. Marshall, E. Silva, A. Anzueto, C. D. Martin, R. Moreno, J. Lipman, C. Gomersall, Y. Sakr, K. Reinhart, International study of the prevalence and outcomes of infection in intensive care units. *JAMA* **302**, 2323–2329 (2009).
39. O. Ginn, L. Rocha-Melogno, A. Bivins, S. Lowry, M. Cardelino, D. Nichols, S. N. Tripathi, F. Soria, M. Andrade, M. Bergin, M. A. Deshusses, J. Brown, Detection and quantification of enteric pathogens in aerosols near open wastewater canals in cities with poor sanitation. *Environ. Sci. Technol.* **55**, 14758–14771 (2021).
40. C. Hanson, M. M. Bishop, J. T. Barney, E. Vargis, Effect of growth media and phase on Raman spectra and discrimination of mycobacteria. *J. Biophotonics* **12**, e201900150 (2019).
41. C. Wichmann, M. Chhallani, T. Bocklitz, P. Rösch, J. Popp, Simulation of Transportation and Storage and Their Influence on Raman Spectra of Bacteria. *Anal. Chem.* **91**, 13688–13694 (2019).
42. R. Mukherjee, T. Verma, D. Nandi, S. Umapathy, Understanding the effects of culture conditions in bacterial growth: A biochemical perspective using Raman microscopy. *J. Biophotonics* **13**, e201900233 (2020).
43. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2015), pp.1–9.
44. A. Nguyen, J. Yosinski, J. Clune, Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2015), pp. 427–436.
45. C. Qu, C. Liu, Y. Gu, S. Chai, C. Feng, B. Chen, Open-set gas recognition: A case-study based on an electronic nose dataset. *Sens. Actuators B Chem.* **360**, 131652 (2022).
46. J. Xu, X. Yi, G. Jin, D. Peng, G. Fan, X. Xu, X. Chen, H. Yin, J. M. Cooper, W. E. Huang, High-speed diagnosis of bacterial pathogens at the single cell level by raman microspectroscopy with machine learning filters and denoising autoencoders. *ACS Chem. Biol.* **17**, 376–385 (2022).
47. Y.-F. Qin, X.-Y. Lu, Z. Shi, Q.-S. Huang, X. Wang, B. Ren, L. Cui, Deep learning-enabled raman spectroscopic identification of pathogen-derived extracellular vesicles and the biogenesis process. *Anal. Chem.* **94**, 12416–12426 (2022).
48. S. Yan, S. Wang, J. Qiu, M. Li, D. Li, D. Xu, Raman spectroscopy combined with machine learning for rapid detection of food-borne pathogens at the single-cell level. *Talanta* **226**, 122195 (2021).
49. M. Del Mar Lleò, D. Benedetti, M. C. Tafi, C. Signoretto, P. Canepari, Inhibition of the resuscitation from the viable but non-culturable state in Enterococcus faecalis. *Environ. Microbiol.* **9**, 2313–2320 (2007).
50. J. Cox, Prediction of peptide mass spectral libraries with machine learning. *Nat. Biotechnol.* **41**, 33–43 (2023).
51. D.-W. Li, A. L. Hansen, C. Yuan, L. Bruschweiler-Li, R. Brüschweiler, DEEP picker is a deep neural network for accurate deconvolution of complex two-dimensional NMR spectra. *Nat. Commun.* **12**, 5229 (2021).
52. Y. Liu, W. Yao, F. Qin, L. Zhou, Y. Zheng, Spectral classification of large-scale blended (Micro)plastics using FT-IR raw spectra and image-based machine learning. *Environ. Sci. Technol.* **57**, 6656–6663 (2023).
53. X. Wang, L. Ren, Z. Diao, Y. He, J. Zhang, M. Liu, Y. Li, L. Sun, R. Chen, Y. Ji, J. Xu, B. Ma, Robust spontaneous raman flow cytometry for single-cell metabolic phenome profiling via pDEP-DLD-RFC. *Adv. Sci.* **10**, e2207497 (2023).
54. X. Wang, Y. Xin, L. Ren, Z. Sun, P. Zhu, Y. Ji, C. Li, J. Xu, B. Ma, Positive dielectrophoresis–based Raman-activated droplet sorting for culture-free and label-free screening of enzyme function in vivo. *Sci. Adv.* **6**, eabb3521 (2020).
55. X.-L. Gao, M.-F. Shao, Q. Wang, L.-T. Wang, W.-Y. Fang, F. Ouyang, J. Li, Airborne microbial communities in the atmospheric environment of urban hospitals in China. *J. Hazard. Mater.* **349**, 10–17 (2018).
56. F. B. Solomon, F. W. Wadilo, A. A. Arota, Y. L. Abraham, Antibiotic resistant airborne bacteria and their multidrug resistance pattern at University teaching referral Hospital in South Ethiopia. *Ann. Clin. Microbiol. Antimicrob.* **16**, 29 (2017).
57. S.-Y.-D. Zhou, H. Li, M. Giles, R. Neilson, X. Yang, J. Su, Microbial flow within an air-phyllosphere-soil continuum. *Front. Microbiol.* **11**, 615481 (2021).
58. F. L. Martin, J. G. Kelly, V. Llabjani, P. L. Martin-Hirsch, I. I. Patel, J. Trevisan, N. J. Fullwood, M. J. Walsh, Distinguishing cell types or populations based on the computational analysis of their infrared spectra. *Nat. Protoc.* **5**, 1748–1760 (2010).

59. Y. Yang, B. Xu, J. Murray, J. Haverstick, X. Chen, R. A. Tripp, Y. Zhao, Rapid and quantitative detection of respiratory viruses using surface-enhanced Raman spectroscopy and machine learning. *Biosens. Bioelectron.* **217**, 114721 (2022).
60. Y. Zhou, S. Shang, X. Song, S. Zhang, T. You, L. Zhang, Intelligent radar jamming recognition in open set environment based on deep learning networks. *Remote Sens.* **14**, 6220 (2022).
61. L. Zhu, R. Li, K. Yang, F. Xu, C. Lin, Q. Chen, D. Zhu, Q. Sun, Y.-G. Zhu, L. Cui, Quantifying health risks of plastisphere antibiotic resistome and deciphering driving mechanisms in an urbanizing watershed. *Water Res.* **245**, 120574 (2023).
62. Y.-M. Tseng, K.-L. Chen, P.-H. Chao, Y.-Y. Han, N.-T. Huang, Deep learning–assisted surface-enhanced raman scattering for rapid bacterial identification. *ACS Appl. Mater. Interfaces* **15**, 26398–26406 (2023).
63. Z. Liu, Y. Xue, C. Yang, B. Li, Y. Zhang, Rapid identification and drug resistance screening of respiratory pathogens based on single-cell Raman spectroscopy. *Front. Microbiol.* **14**, 1065173 (2023).

Acknowledgments
Funding: This work was supported by the National Natural Science Foundation of China (42021005 to Y.G.Z and 32100083 to L.Z.), the National Key R&D Program of China (2022YFF0713100 to L.C.), the National Natural Science Foundation of China (22176186, to L.C.), and the Chinese Academy of Sciences (ZDBS-LY-DQC027 to L.C.). **Author contributions:** Conceptualization: L.Z., F.X., B.R., Z.-Q.T., P.J.V., Y.-G.Z., and L.C. Methodology: L.Z., F.X., X.L., Y.-G.Z., and L.C. Investigation: L.Z., Y.Y., F.X., and H.L. Software: L.Z., X.L., M.S., and Z.A. Formal analysis: L.Z., H.L., and M.S. Resources: L.Z., F.X., B.R., Y.-G.Z., and L.C. Data curation: L.Z., Y.Y., and F.X. Validation: L.Z., Y.Y., F.X., M.S., and H.L. Visualization: L.Z., F.X., M.S., H.L., F.L.M., Z.-Q.T., and L.C. Supervision: B.R. and Y.-G.Z. Project administration: L.Z., Y.Y., Y.-G.Z., and L.C. Funding acquisition: L.Z., Y.-G.Z., and L.C. Writing—original draft: L.Z., X.C., B.R., and L.C. Writing—review and editing: L.Z., F.L.M., P.J.V., and Y.-G.Z. **Competing interests:** The authors declare that they have no competing interests. A Chinese patent application based on this work is pending (application number: 202211075462.0). **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Data used for training and test the DL models have been deposited in ScienceDB database at https://doi.org/10.57760/sciencedb.15628. The code used in this study is available in ScienceDB database at https://doi.org/10.57760/sciencedb.12074.

Submitted 12 April 2024
Accepted 3 December 2024
Published 8 January 2025
10.1126/sciadv.adp7991