High-speed diagnosis of bacterial pathogens at single cell level by Raman microspectroscopy with machine learning filters and denoising autoencoders

Jiabao Xu[1], Xiaofei Yi[2,3], Guilan Jin[2,3], Di Peng[2,3], Gaoya Fan[2,3], Xiaogang Xu[4,5], Xin Chen[4,5], Huabing Yin[6], Jon M. Cooper[6], Wei E. Huang[1]*

[1]Department of Engineering Science, University of Oxford; Oxford, Oxford, OX1 3PJ, U.K.

[2]Shanghai Hesen Biotechnology Co., Ltd; Shanghai, 201802, China.

[3]Shanghai D-band Medical Instrument Co., Ltd; Shanghai, 201802, China.

[4]Institute of Antibiotics, Huashan Hospital, Fudan University; Shanghai, 200040, China.

[5]National Clinical Research Center for Aging and Medicine, Huashan Hospital, Fudan University; Shanghai, 200040, China.

[6]James Watt School of Engineering, University of Glasgow, Glasgow, G12 8LT, U.K.

*Corresponding author. Wei E. Huang. Email: wei.huang@eng.ox.ac.uk

Keywords: Raman microspectroscopy, single cell, pathogen detection, machine learning, signal-to-noise ratio, diagnosis

Abstract: Accurate and rapid identification of infectious bacteria is important in medicine. Raman microspectroscopy holds great promise in performing label-free identification at the single-cell level. However, due to the naturally weak Raman signal, it is a challenge to build extensive databases and achieve both accurate and fast identification. Here, we used signal-to-noise ratio (SNR) as a standard indicator for Raman data quality and performed bacterial identification using 11,141 single-cell Raman spectra from 9 bacterial strains. Subsequently, using two machine learning methods, a simple filter and a neural network-based denoising autoencoder (DAE), we demonstrated 92% (simple filter using 1-second/cell spectra) and 84% (DAE using 0.1-second/cell spectra) identification accuracy. Our machine learning-aided Raman analysis paves the way for high-speed Raman micro-spectroscopic clinical diagnostics.

Introduction

Bacterial infections account for at least 700,000 deaths globally each year[1]. As antimicrobial resistance rapidly emerges, these infections are estimated to cause 10 million deaths and a $100 trillion economic burden by 2050[1]. Current diagnostic methods in clinical practices require lengthy cell culturing for one to a few days, before samples can be sent for pathogen identification and antibiotic susceptibility testing (AST). Broad-spectrum antibiotics are often prescribed prophylactically before diagnosis, increasing the prevalence of circulating and often inappropriate drug treatment[2]. New methods for rapid, accurate, and culture-free identification of bacterial infection are urgently required for early and accurate treatment as part of programmes of antimicrobial stewardship, aimed at reducing the emergence of resistant strains of micro-organisms.

Raman microspectroscopy is label-free vibrational spectroscopy, which is often implemented within a confocal optical configuration, able to interrogate biochemical profiles of bacteria at the single-cell level. A single-cell Raman spectrum (SCRS) represents a collection of vibrational frequencies characteristic of the cellular biomolecules present in a single cell, providing a unique fingerprint, namely, "Raman phenotype". Recent applications of Raman microspectroscopy in bacterial characterization have shown great potential to differentiate bacteria by genus, species and strain[3–10].

Single-cell resolution afforded by confocal Raman microspectroscopy makes it possible to detect clinical samples with low bacterial population, such as those with blood infections (typically <10 CFU/mL)[11]. Compared with other culture-free methods, including single-cell sequencing and molecular tagging[12–15], Raman microspectroscopy provides a unique advantage for phenotype identification, as it can be performed without specifically designed probes or labels. Several studies have demonstrated that Raman microspectroscopy can identify pathogens and perform AST directly from clinical samples[4,10,16,17]. A recent study containing an extensive dataset of 30 common bacterial pathogens illustrated 82% isolate-level accuracy and 97% antibiotic treatment level accuracy using Raman microspectroscopy and deep learning[3], demonstrating that the technique has the potential to become the next-generation diagnostic tool for rapid single-cell identification in clinical practice.

However, key challenges remain. Spontaneous Raman scattering is intrinsically weak, with around 1 in 10^7 to 10^8 scattering photons being detected, and consequently collection times of between 15–60 second per spectrum are needed to acquire data with a sufficient signal-to-noise ratio (SNR) for successful analysis and classification tasks[4,10,18–24]. This compromise between

the throughput of data acquisition and SNR significantly limits the collection of a large number of spectra to establish a database, underpinning the required, reliable training sets for machine learning to identify clinically relevant species and strains. Currently, the low speed of acquiring Raman spectra is a major obstacle for building such databases, with an additional problem being the high intra- and inter-sample variations found in Raman spectra, even from the same strain[25].

To overcome this, we developed standard Raman operation protocols that enable us to achieve optimal and accurate classification of bacterial phenotypes, ready to be translated into clinical practice. In doing so, we proposed to use the SNR as a standard criterion for quality control of SCRS, and then applied machine learning methods to speed up pathogen identification by analyzing low-SNR SCRS with short acquisition time. In this study, we studied 9 clinically relevant bacterial strains using a range of Raman acquisition times from 0.01 to 15 s, which yielded a total number of 11,141 SCRS. By using this large database with multiple parameter-performance correlations, we inspected the SNRs, unsupervised visualization, and classification performance under all acquisition conditions to provide a detailed generic guidance for the bacterial identification tasks. We developed two machine learning methods to optimize identification performance on low-SNR spectra. With an easy-to-use SNR filter, we improved the classification performance to 92% accuracy using spectra acquired with a throughput of 1 s/cell. With an advanced denoising autoencoder (DAE), we achieved 97% accuracy identifying bacterial species using spectra acquired with 1 s/cell speed or 84% accuracy with 0.1 s/cell speed. Accordingly we demonstrate that Raman microspectroscopy, in combination of machine learning for SCRS analysis, can make ultra-rapid and accurate bacterial identification possible, paving the way for building large databases and translating the methods into clinical practice.

Experimental Section

Bacterial strains and culture conditions

All 9 bacterial strains used in this study were provided by Huashan Hospital in Shanghai, China (**Table 1**), including *Staphylococcus epidermidis* ATCC 12228, *Acinetobacter baumanii* ATCC 19606, *Staphylococcus aureus* ATCC 25923, *Staphylococcus aureus* ATCC 29213, *Enterococcus faecalis* ATCC 29212, *Escherichia coli* ATCC 35218, *Escherichia coli* ATCC 25922, *Pseudomonas aeruginosa* ATCC 27853, and *Klebsiella pneumonia* ATCC 700603. All strains were grown on tryptone soya agar (TSA) plates at 37 °C for 24 h. One of the colonies was then suspended in 5 mL of tryptone soy broth (TSB) medium and incubated at 37 °C for

16 h with shaking at 180 rpm. After the cells reached the stationary phase, 1 mL of the sample was washed three times with sterile water. After resuspension in 1 mL sterile water, 2 μL of each sample was deposited onto an aluminum-coated slide and allowed to dry at room temperature. For each strain, two independent batches were prepared.

Single-cell Raman measurements, processing, and t-SNE visualization

Raman spectra of single cells were acquired by a Raman microscope (Alpha300R, WITec, Germany) equipped with a 532-nm laser. The laser beam was focused onto the sample with a 100× objective (100×/NA=0.9, ZEISS, Germany) with a power of approximately 7 mW applied on the sample. Cells were measured with a grating of 1200 mm/g, and the exposure time was set as 0.01, 0.1, 1, 10, or 15 s on the same cell. The spectral resolution was ~2 cm^{-1}, and a wavenumber range of 280–2186 cm^{-1} was chosen. For each sample, ~250 cells were measured for one bacterial strain at 5 acquisition times. A total of 11,141 single-cell Raman spectra were acquired for all conditions.

Pre-processing of the raw Raman spectra involved quality control by removing abnormally/burnt high-intensity spectra, comic-ray correction, and baseline fitting and subtraction to remove autofluorescence. Spectral normalization was done by vector normalization of the entire spectral region to correct general instrumentation fluctuation as well as sample and experimental variables without strongly interfering with the nature of the biological content.

A method of t-distributed stochastic neighbor embedding (t-SNE) was used to embed the high-dimensional Raman dataset in a two-dimensional space by minimizing the Kullback−Leibler divergence between the two probability distributions in respective dimensional spaces[26]. It was aimed to reduce the high dimensionality and collinearity of the dataset and further aid visualization.

LDA classification models

At each acquisition time, spectra were divided into a training set and a testing set with a 9:1 split ratio. Both the training and testing sets contained a balanced number of samples from seven bacterial species. While the training set was used to train a classification model, the test set was used to evaluate the model performance independently. Pre-processing of the Raman dataset involved scaling, centering, and dimension reduction by principal component analysis (PCA)

to the first hundred principal components. The pre-processed Raman spectra were used as the inputs into the LDA classifier, a linear classifier, to find a linear combination of features that separate different classes. Ten-fold cross-validation repeated by five times was used during model construction. Model performance was evaluated by the independent test set. Performance measures for all models were computed as accuracy for each class and an overall accuracy rate.

SNR calculation and filters

Calculation of SNR for each Raman spectrum was carried out using the formulae below:

$$SNR = \frac{\max(S) - \text{average}(N)}{sd(N)}$$

$$sd(N) = \sqrt{\frac{\sum_i^n (N_i - \bar{N})^2}{n-1}}$$

where S is the signal at the highest intensity, N is the noise region (1800–1900 cm^{-1} in biological samples), and $sd(N)$ is the standard deviation of the noise.

An SNR filter was constructed based on individual SNRs which removed the half of total Raman spectra for each group, namely, the first quantile (25% of the lowest individual SNR) and the fourth quantile (25% of the highest individual SNR) of the data. The selection of filter was based on the best performance improvement achieved in the classification task. Among removing 10% (5% with lowest SNR and 5% with highest SNR), 25% (12.5% with lowest SNR and 12.5% with highest SNR), and 50% (25% with lowest SNR and 25% with highest SNR) of the total data points, the SNR filters that removing half of the dataset have outperformed other filters.

DAE architecture and method

The DAE architecture was based on multilayer perception (MLP) feedforward artificial neural networks (NN) trained with stochastic gradient descent using back-propagation[27,28]. The change in each weight w of the jth node in the nth training sample is:

$$\Delta w_{ji}(n) = -\eta \frac{\partial \mathcal{E}(n)}{\partial v_j(n)} y_i(n)$$

where η is the learning rate, $\mathcal{E}$ is the error, v is the induced local field, and y_i is the output from the previous neuron.

The DAE has a bottleneck structure featuring an NN encoder that compresses the data dimension, a code that preserves the important attributes, and a NN decoder that reconstructs the code into spectral data with the Raman features but higher SNR. The architecture hyperparameters, including the numbers of hidden layers and units, activation function, dropout rate, and regularization, were examined with a grid search of multiple combinations. The best model was determined by the lowest mean squared error (MSE):

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

where y_i is the predicted outcome in terms of probability of each species and $\hat{y}_i$ is the expected outcome of each species which equals to one.

A low-SNR dataset (0.1- or 1-s acquisition time) and a high-SNR dataset (1- or 10-s acquisition time) measured at the same cell positions were selected for DAE training and evaluation, respectively. The training and validation sets account for 70% of the data acquired at a particular acquisition time and the training set account for 30%. An independent low-SNR dataset (0.1- or 1-s acquisition time) was selected as a training set. No labels were given during the training process (unsupervised). The autoencoder was trained by minimizing the reconstruction error between the reconstructed output and the high-SNR validation frame and an additional term of regularization, discouraging memorization/overfitting. The independent low-SNR testing set was then inputted into the trained DAE to evaluate performance and reconstructed into the DAE-transformed spectra.

Data availability

Data used in this study is available at https://doi.org/10.6084/m9.figshare.14377259.v1. All code is available at https://doi.org/10.6084/m9.figshare.14377286.v1.

Results & Discussion

Matching acquisition speed with SNR

We obtained the datasets by measuring Raman spectra on 9 bacterial strains from 7 species (**Table 1**), covering 66% of all clinical bacterial isolates from 52 Chinese hospitals in 2019[29]. Seven of the strains are quality control standards for clinical antimicrobial susceptibility tests according to Clinical and Laboratory Standards Institute (CLSI)[30]. For each strain, 250 single-

cell Raman spectra (SCRS) were measured, and acquisition times of 0.01, 0.1, 1, 5, 10, and 15 s were separately used on the same bacterial cell for determining the speed limit in a classification task. The power of the 532-nm continuous laser source, used in the Raman spectrometer, was 7 mW, which has previously been shown capable of acquiring high-quality spectra without damaging on the cell's molecular composition for short periods of time[31]. In total, 11,141 Raman spectra were collected and analyzed.

Table 1. **List of bacterial strains used in this study (bacterial Raman spectra n = 11,141).**

Species	Strain	Gram	Description
Staphylococcus aureus	ATCC 25923	Positive	Quality control strain for antibiotics susceptibility testing
	ATCC 29213	Positive	Quality control strain for antibiotics susceptibility testing
Escherichia coli	ATCC 25922	Negative	Quality control strain for antibiotics susceptibility testing
	ATCC 35218	Negative	Quality control strain for antibiotics susceptibility testing
Pseudomonas aeruginosa	ATCC 27853	Negative	Quality control strain for antibiotics susceptibility testing
Klebsiella pneumoniae	ATCC 700603	Negative	Quality control strain for antibiotics susceptibility testing
Staphylococcus epidermidis	ATCC 12228	Positive	Quality control strain
Acinetobacter baumannii	ATCC 19606	Negative	Quality control strain
Enterococcus faecalis	ATCC 29212	Positive	Quality control strain for antibiotics susceptibility testing

Acquisition speed is an essential parameter that determines the quality of a SCRS through the SNR. There are other constituting factors that contribute to the final SCRS output, including spectrometer configuration (sampling aperture, filters, objective lenses, detector, etc.) together with excitation wavelength, power as well as exposure time[25]. To unify the standard of spectral measurements, we used the SNR as a universal indicator to the quality of the spectral output, independent to the instrument and parameter settings. We propose that the SNR is the defining parameter, critical to subsequent data analysis and machine learning tasks. We calculated the averaged SNR for each bacterial species at each of the 5 acquisition speeds and showed the distribution of SNRs of each group (**Table S1, Figure S1** and **Figure S2**) using the formulae below:

$$SNR = \frac{\max(S) - \text{average}(N)}{sd(N)}$$

$$sd(N) = \sqrt{\frac{\sum_i^n (N_i - \bar{N})^2}{n-1}}$$

where S is the signal at the highest intensity, N is the noise region (1800–1900 cm^{-1} in biological samples), and $sd(N)$ is the standard deviation of the noise.

As stated, a key challenge of Raman spectroscopy is that the Raman signal is naturally weak, and as such there is always a compromise between measurement speed/throughput of SCRS and the output SNR. **Figure 1A** shows an example SCRS set of *Escherichia coli* 35219, from low SNR and high speed to high SNR and low speed. Building upon such exemplar data, we acquired a database of 11,141 SCRS and calculated SNR at each acquisition time for each of the 9 strains (**Figure 1B**, Figure S1, Figure S2 and Table S1). Large fluctuations of the SNR was observed within the same strain under the same measurement conditions due to the heterogeneity of individual cells (Figure S2). For example, with 1-s acquisition time, spectra of *Acinetobacter baumannii* 19606 yielded an averaged SNR of 8.1 with a range between 0 and 22 (Table S1). We also observed that the SNR value was not linearly correlated with acquisition time. Increasing acquisition time beyond 1 s gave less improvement on the output SNR, as a longer acquisition time did not necessarily mean a higher SNR, and in fact, could lead to photo-decomposition of local metabolites or proteins thereby introducing intra-sample variability. For example, in *Pseudomonas aeruginosa* 27853, a long acquisition time of 15 s yielded spectra with lower SNR than 10 s (Figure 1B and Figure S1). Finally, we noted that differences of SCRS quality across species were evident. An acquisition time of 1 s yielded an average SNR of 6.1 in *Staphylococcus epidermis*, 9.4 in *Klebsiella pneumoniae*, and 26.9 in *P. aeruginosa* (Table S1).

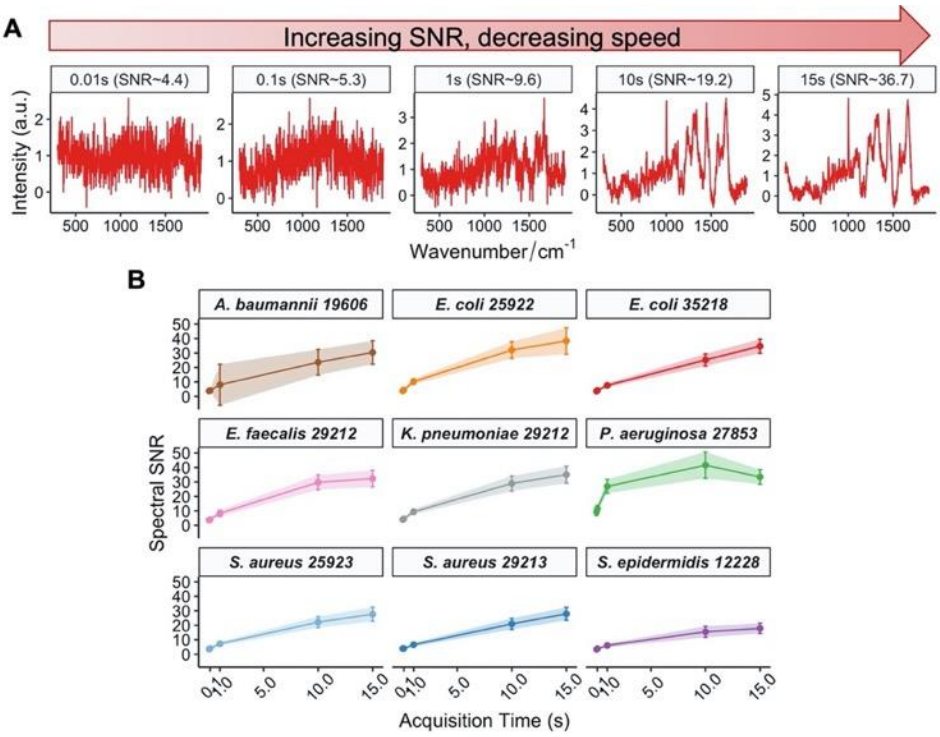


Figure 1. **Matching acquisition speed with SNR. (A)** Compromise made between speed and SNR showing by example spectra of *E. coli* 35219. **(B)** Spectral SNR of 9 bacterial strains with different acquisition time of 0.01, 0.1, 1, 10 or 15 s, each containing ~250 spectra from 2 replicates.

Given this degree of complexity and variability in the spectra collected, we demonstrated the significant SNR variations under identical measuring conditions, both within the same sample and across different species/strains. To make various SCRS comparable in different Raman spectroscopic studies, we propose that SNR can be used as a quality-control parameter because it represents the final spectral output.

Classification of bacteria at different speeds and SNRs

We then examined the underlying feature preservation and robustness in the dataset across different acquisition speeds and SNRs. We first tested the performance of the dataset in an unsupervised visualization task. **Figure 2A** shows the averaged Raman spectra for 7 species in the fingerprint region (300–1800 cm^{-1}), which entails most of the biomolecular vibrational modes within a single cell, providing a unique observable characteristics, so-called "Raman phenotype".

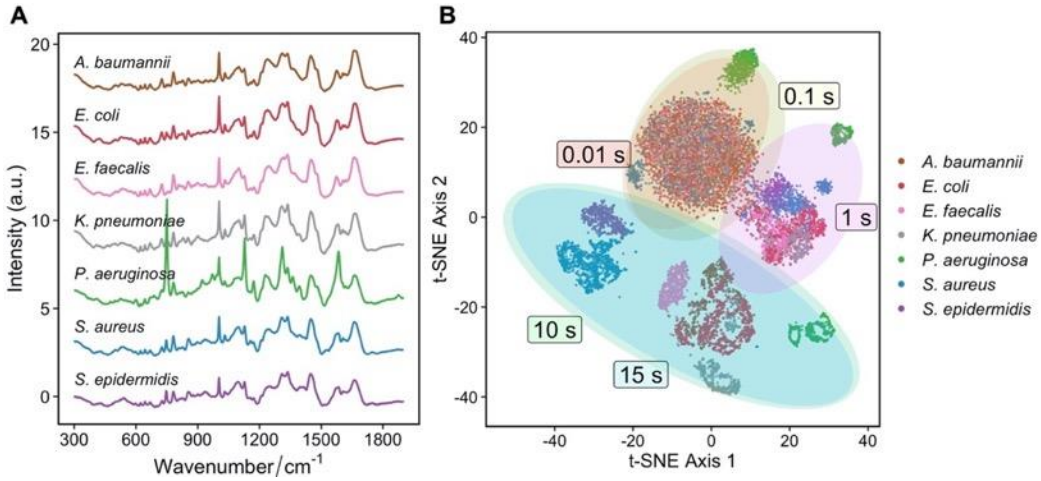


Figure 2. **Raman spectra and unsupervised t-SNE visualization. (A)** Averaged Raman spectra of 7 species from all acquisition time and **(B)** t-SNE visualisation of SCRS showing clusters of 7 species at different acquisition time.

Due to the high dimensionality of a Raman spectrum, multivariate dimension-reducing techniques, such as principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) algorithms, are commonly used for feature extraction and visualization[32]. The t-SNE plot of all data clearly separates spectra into clusters regarding different acquisition speeds (**Figure 2B**). SCRS acquired at 10 and 15 s/cell overlap as they had similar attributes. At 1 s/cell, SCRS showed some discrimination power at the species level and form clusters of different bacterial species. Clusters of SCRS acquired at 0.01 and 0.1 s/cell overlapped and showed no discrimination power over 6 of the 7 species. Raman spectra of *P. aeruginosa* cells were most distinct due to the resonance effect of cytochrome c molecules at 749, 1128, 1312 and 1584 cm^{-1}, resulting in high SNR (Figure 2A) and displaying distinct clusters in all acquisition settings (Figure 2B).

Next, we tested the performance of the dataset in supervised classification tasks. Data at each acquisition speed was split into a training set and a testing set with a 9:1 ratio. Each of the 5 training sets at either 0.01, 0.1, 1, 10, or 15 s/cell was used as an input to train a linear discriminant analysis (LDA) classification model, and each of the 5 test sets was used to evaluate the model performance.

Figure 3A summarizes the 25 model performances in overall classification accuracy. The model generally performed better when the training and testing sets had the same acquisition time or similar SNR. For example, the model trained with 10-s/cell SCRS illustrated 100% accuracy when classifying 10-s/cell SCRS but only 56% with 1-s/cell SCRS and 97% with 15-

s/cell SCRS. A testing set with a higher SNR could be beneficial in the case of an 8% accuracy increase using the 10-s/cell testing set for a model constructed from 1-s/cell SCRS. As expected, the model performed better with increasing acquisition time and SNR. The model performance increased from 45% for 0.01-s/cell, 52% for 0.1-s/cell, 85% for 1-s/cell, to 100% for 10-s/cell and 15-s/cell SCRS (Figure 3A). This result is consistent with a recent study using a classification model for 30 strains of pathogenic bacteria based on Raman spectra and deep learning, which reported an 82% accuracy rate on 1-s spectra[3].

Here, we have demonstrated that Raman spectral SNR is directly related to the performance of a classification model. Therefore, we propose that SNR is a better optimization parameter than the acquisition time, per se, and thus provides a standard criterion for quality evaluation of SCRS in spectroscopic practice, regardless of the instrument used and parameter settings.

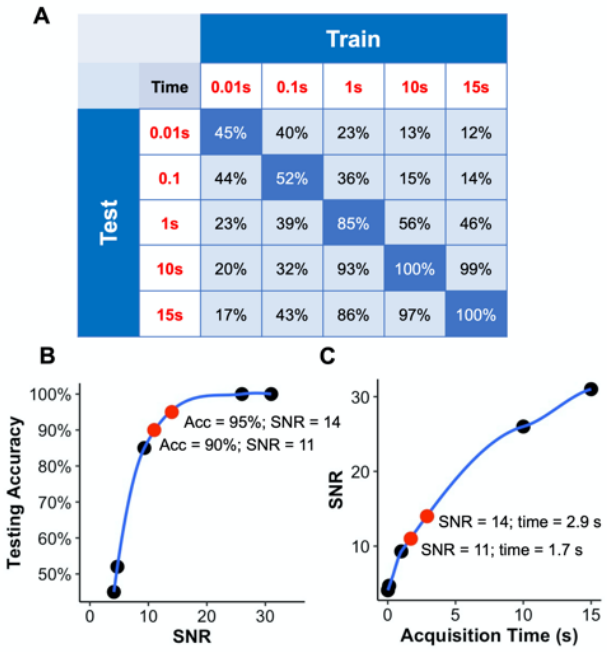


Figure 3. Bacterial classification based on Raman spectra and LDA models. (A) Classification performance on testing sets based on Raman spectra of 7 species at different acquisition times. Spectra at each acquisition parameter were split into a training set and a testing set where the training set was used for model training and the testing set was used for model evaluation. The diagonal was highlighted where the training and testing sets have matched acquisition parameters. **(B)** Average spectral SNR vs testing accuracy. An SNR of 11 is predicted to achieve desired accuracy of 90%; an SNR of 14 is predicted to achieve desired accuracy of 95%. **(C)** Spectral acquisition time vs average SNR. An SNR of 11 corresponds to

an acquisition time of 1.7 s; an SNR of 14 corresponds to an acquisition time of 2.9 s. The points were fitted by locally estimated scatterplot smoothing (LOESS) with a span = 0.8.

Figures 3B and **3C** show the testing accuracy versus the averaged SNR and the averaged SNR versus the acquisition time. The plots predict that for a desired test accuracy at 90%, an SNR of 11 is required, which corresponds to an acquisition time of 1.7 s/cell. An ideal accuracy at 95% will require an SNR of 14 and an acquisition time of 2.9 s/cell.

Notably, most current Raman spectroscopic studies ranging from the detection and identification of bacteria to the characterization of their metabolism and antibiotic responses use long acquisition times of 15–60 s aiming for better SNR output and downstream analysis[4,10,18–24]. The long measurement time impedes the collection of a large number of spectra into a dataset, which is crucial for machine learning identification tasks and their applications in real-world practices. Our results illustrate good classification performance on low-SNR spectra with a simple linear model, which required lower computational cost compared with complex algorithms such as deep learning. We show between correlations of speed, SNR and accuracy, thereby providing references for future applications with their desired levels of classification performance.

Optimizing high-speed classification by an SNR filter

After giving general references of speed and SNR for a classification task, we next sought to provide solutions to optimize performances of the low-SNR spectra. Based upon the observation of best classification performances achieved in spectra with the same acquisition time (Figure 3A), we hypothesized that a consistent SNR was a critical factor in data analysis and machine learning approaches, and confining SNR within certain limits will improve the classification performance.

Therefore, we explored high-speed low-SNR spectra acquired with 1-s/cell exposure time. We created an SNR filter to pass through the 1-s dataset. Based on the calculation of SNRs of individual SCRS, the filter removed the first quantile (25% of the lowest individual SNR) and the fourth quantile (25% of the highest individual SNR) of the data (**Figure 4**). Due to the unique characteristics of *P. aeruginosa* spectra (**Figure 4A**), LDA visualization before and after the filter application was shown without the spectra of *P. aeruginosa* (**Figure 4B**). Surprisingly, despite reducing the total spectra number by half, the LDA clustering illustrated significant improvement after passing through the SNR filter with more confined and distinct clusters of each bacteria species. A similar improvement was also observed at the strain level (**Figure S3**),

which could be particularly important for bacterial isolates, despite being identical species, with different antibiotic susceptibility profiles.

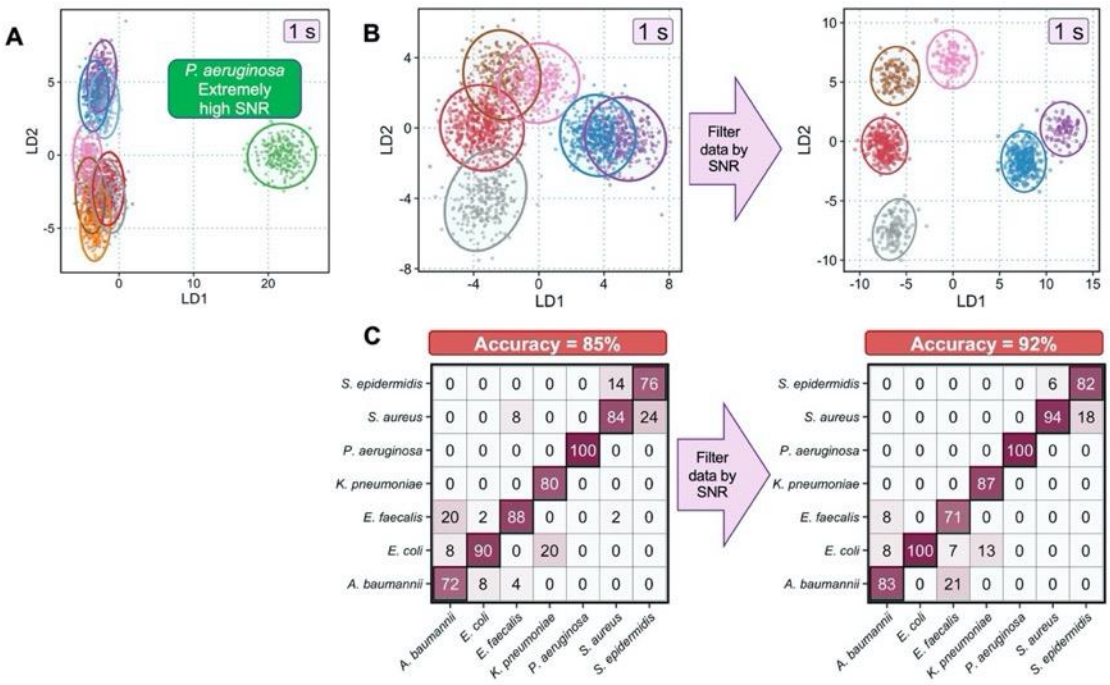


Figure 4. **Optimizing high-speed low-SNR spectra by a SNR filter which removes spectra with high (quantile > 0.75) or low SNR (quantile < 0.25) of each species. (A)** LDA plot before the SNR filter shows a distinct cluster of *P. aeruginosa* due to its unique spectral features and high SNR. **(B)** LDA clustering after a SNR filter shows much better predictive power at the species level. (**C)** Confusion matrix of an LDA classification model shown in percentages of the accurately classified species. Columns shows the actual percentages and rows as the predicted percentages. Performance of the model improves from an accuracy of 85% to 92% after a SNR filter.

We then evaluated the performance of the filtered dataset in a classification model (**Figure 4C**). Better species classification prediction was observed in five out of the seven bacterial species. In particular, classification within the same *Staphylococcus* genus (*Staphylococcus aureus* and *Staphylococcus epidermidis*) improved from 84% and 76% to 94% and 82%. The predictive power of the LDA model improved from an overall accuracy rate of 85% to 92% (Figure 4C and **Table 2**), consistent with the LDA visualization results (Figure 4B). After the SNR filter, *E. faecalis* was misclassified as either *A. baumannii* or *E. coli*. Correspondingly, accuracy rates for *A. baumannii* and *E. coli* were improved from 72% and 90% to 100% and 83%, respectively. Therefore, misclassification of *E. faecalis* was compensated by improvement in other species.

The selection criteria were likely chosen by the algorithm to minimize a global error and optimize the overall performance.

These results demonstrate that the supervised model performance has significantly improved performance by merely confining the SNR values within a certain range for spectra with low acquisition time. It verifies that SNR values should be considered as an essential parameter during spectroscopic measurement. Consistent SNR will help control data quality and reduce variations of intra- and inter-sampling that is independent on biological features, thereby contributing to improve classification performance.

Optimizing high-speed classification by a denoising autoencoder (DAE)

Neural networks have been applied with considerable success to a broad range of problems and, more recently, have been adapted in spectroscopic works[3,33–37]. Autoencoders are unsupervised learning technique in which neural networks are leveraged for the task of representation learning. Here, we present an advanced denoising autoencoder (DAE), based on multilayer perceptron (MLP) neural networks, for denoising the low-SNR Raman spectra. **Figure 5A** demonstrates the scheme of the DAE. The autoencoder has a neural network architecture with a bottleneck, which turns a high-dimensional spectroscopic input into a latent low-dimensional code (encoder) and then reconstructs an output that resembles the input with the latent code (the decoder). Due to the dimension reduction and feature learning during the process, the reconstructed output would have preserved the more essential features and therefore have a higher SNR than the noisy input.

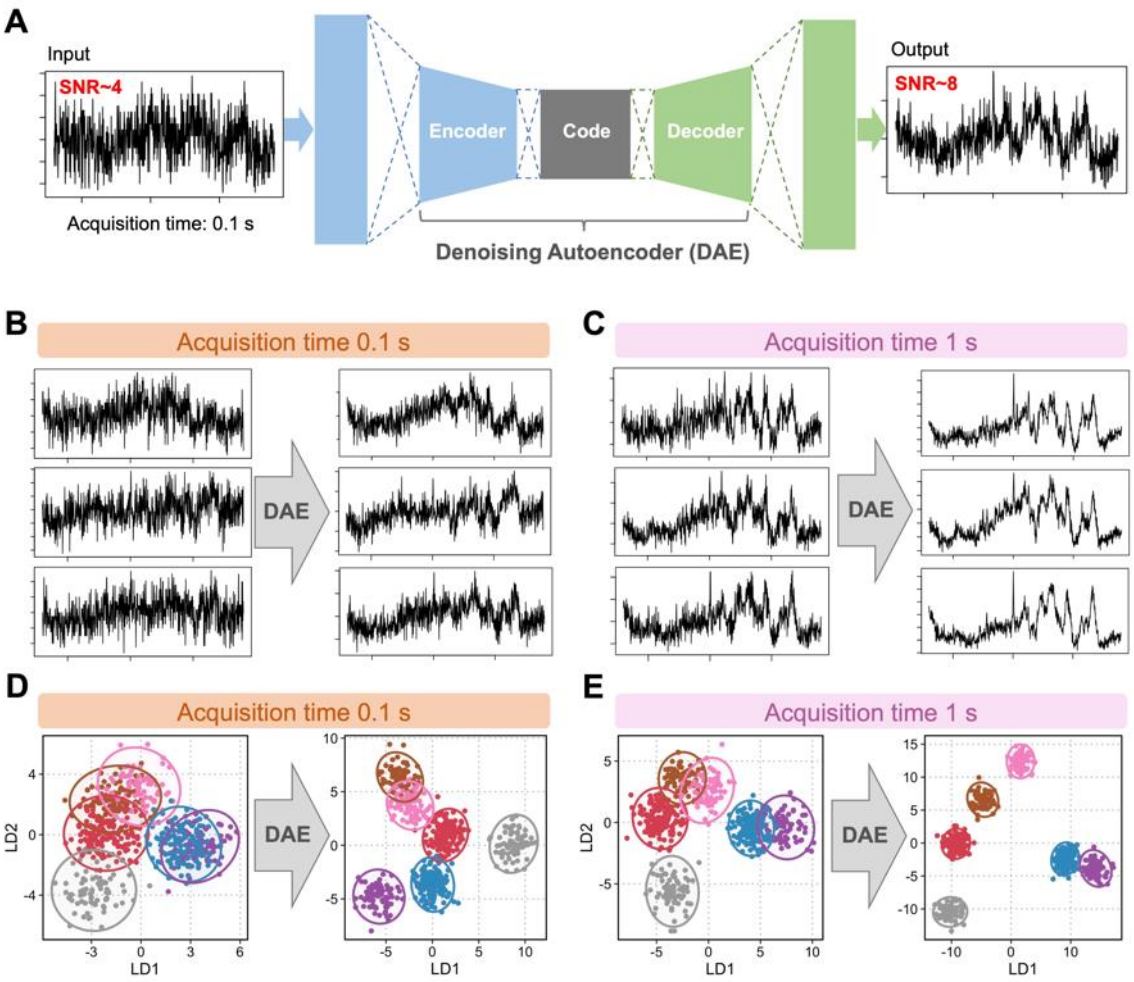


Figure 5. **A DAE for improving spectral SNR and optimizing classification performance.** **(A)** The DAE consists of an encoder that encodes the original Raman input into low-dimensional features, a code that stores the features, and a decoder that transform the code into an output that resembles the input. An example low-SNR spectrum illustrates an improvement in SNR by two times after passing through the DAE. **(B)** Examples of SNR improvement of 0.1-s spectra after DAE. **(C)** Examples of SNR improvement of 1-s spectra after DAE. **(D)** Improvement in LDA clustering of 0.1-s spectra after DAE. **(E)** Improvement in LDA clustering of 1-s spectra after DAE.

The DAE used three sets of data: a low-SNR training dataset, a high-SNR validation dataset, and a low-SNR testing dataset. The first two datasets are paired datasets with matching indexes of same cells, used to train the algorithm on how to transform low-SNR to high-SNR spectra. The third low-SNR testing dataset is an independent set of data used to evaluate the performance of the trained DAE model. The learning ability of the DAE is achieved by learning and preserving the intrinsic differences in the low-SNR spectra and validate using the high-

SNR spectra, thereby only "remembering" the most important features and "forgetting" the noise. After being trained, independent datasets can be passed through the DAE model to be transformed with higher SNRs.

An example testing spectrum acquired for 0.1 s/cell after passing through the DAE showed an improvement of SNR from 4 to 8, equivalent to an SNR of a 1-s/cell SCRS (Figure 4A). We further applied the trained DAE to an entire test set of either 0.1-s/cell SCRS (**Figure 5B**) or 1-s/cell SCRS (**Figure 5C**). All SCRS exhibited noise reductions after DAE transformation, as presented in the three examples of the 0.1-s or 1-s/cell dataset. The averaged SNR was increased from 4.1 to 7.2 for the 0.1-s/cell spectra and from 8.0 to 17.0 for the 1-s/cell spectra.

Table 2. **Summary of performance improvement by the SNR filter or DAE.**

	Raw spectra	**SNR-filtered spectra**	**DAE-processed spectra**
LDA classification accuracy at 0.1s	52%	65%	84%
LDA classification accuracy at 1s	85%	92%	97%
Advantages	/	Easy to apply; can be incorporated during routine measurements	More improvement on low-SNR data
Disadvantages	/	Less improvement on low-SNR data	More complex; need additional training data pairs

Next, we examined the clustering visualization and corresponding classification performance after DAE transformation. Both of the 0.1- and 1-s/cell SCRS show visually more distinct and tighter clusters of different bacterial species than the original spectra, exhibiting good feature preservation and extraction from the DAE (**Figure 5D** and **5E**). The DAE-transformed spectra were then split in the same manner as the original spectra and put into an LDA classifier for training and testing the model. The performance on the testing set has significantly improved after DAE reconstruction (**Table 2**). The model's overall accuracy increased from 52% to 84% for the 0.1-s SCRS, which is similar to the result achieved by the original 1-s spectra. The

transformed 1-s spectra improved from an accuracy rate of 85% to 97%, approaching the high performance achieved by the untransformed 10-s spectra.

By applying a trained DAE to low-SNR SCRS, we demonstrate significant improvements on both clustering and classification results, with the increases of SNR by two times and acquisition speed by ten times.

Conclusions

Raman microspectroscopy holds great promise as a technique for label-free and culture-independent identification of pathogens. In this work, we address the key challenges that have hindered its clinical application, involving the long data acquisition times (15–60 s per spectrum), lack of reference standard and quality control during measurements, and lack of optimization methods for high-speed and low-SNR SCRS. These challenges have led to difficulties in building large and robust Raman databases consisting of a comprehensive collection of bacterial pathogens, as the foundation of a diagnostic testing tool.

We performed classification on 9 bacterial strains with clinical relevance using a range of acquisition times from 0.01 to 15 s. By analyzing an extensive database with multiple parameter comparisons, we reported the output SNRs of spectra and found high inter- and intra-sample variations. Given that many variables determine the final spectra quality, we proposed that a common standard of SNR would be a good criterion for quality control. We also reported classification performance in all acquisition conditions and illustrated good classification results on low-SNR spectra (85% for 1-s spectra) with a simple linear classifier model. By providing general references on speed-SNR-accuracy correlations, we hope to guide future applications with their desired classification performance levels.

Next, we reported two machine learning methods for realizing rapid bacterial identification. The first method involved an easy-to-apply SNR filter which improved the classification accuracy from 85% to 92% using spectra with 1-s acquisition time. It further validated that SNR should be used as the quality control factor during Raman measurements and/or Raman spectra-based classification analysis. Although the filter was less effective in classifying the 0.1-s/cell spectra, it has the advantages of its easy applicability and potential to be incorporated into routine spectroscopic measurements as part of the quality control standards (Table 2). The second method involved a more advanced DAE based on MLP neural nets. After DAE transformation, the 0.1-s spectra showed an improved SNR, equivalent to the untransformed 1-s spectra and a classification accuracy of 84%. With the DAE being more complex and requiring

additional training/testing datasets, it shows the potential to achieve ultra-rapid Raman measurements with 0.1-s acquisition time (Table 2).

This approach also has the potential to be applied to spectroscopic data acquired by models of Raman spectrometers with much lower prices and volumes (e.g. portable and hand-held systems) on the market which could be ideal for clinical settings. These systems, compared with high-performance systems, usually have limitations in e.g. CCD detector, shorter optical pathway, and confocality. These disadvantages will usually result in problems of spectral resolution and weaker signals. By learning essential biomolecular features among noisy background, our approach provides a solution for decoding noisy and less resolved spectra, possibly coming from more affordable Raman systems. The same signal-to-noise standard should be maintained in such systems as it is a quality-control criterion, regardless of device model and acquisition parameters.

With methods proposed in this study, rapid Raman measurements can be realized for quickly generating large sets of data. Through the combination of standardized protocol and machine learning, we can overcome the inherently weak signal in Raman analysis, paving the way for the translation of Raman spectroscopy into clinical application for on-time identification of pathogens and AST.

Supporting Information

The Supporting Information is available free of charge via the internet at http://pubs.acs.org. Figures S1 to S3 and Tables S1 (PDF).

Acknowledgements

J.X., H.Y., J.M.C and W.E.H. thank Innovate UK (project reference number 104984). X.Y., G.J., D.P., G.F., X.X. and X.C are grateful to the grants received from National Key Research and Development Project (2018YFE0101800), Shanghai Municipal Science and Technology Commission (18411950601). J.X., J.W. and W.E.H. also thank the Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences for financial support.

Notes

The authors declare no competing financial interest.

References

(1) *No Time to Wait: Securing the Future from Drug-Resistant Infections*; Interagency Coordination Group on Antimicrobial Resistance, 2019.

(2) Fleming-Dutra, K. E.; Hersh, A. L.; Shapiro, D. J.; Bartoces, M.; Enns, E. A.; File, T. M.; Finkelstein, J. A.; Gerber, J. S.; Hyun, D. Y.; Linder, J. A. et al. Prevalence of Inappropriate Antibiotic Prescriptions Among US Ambulatory Care Visits, 2010-2011. *JAMA* **2016**, *315* (17), 1864–1873. https://doi.org/10.1001/jama.2016.4151.

(3) Ho, C.-S.; Jean, N.; Hogan, C. A.; Blackmon, L.; Jeffrey, S. S.; Holodniy, M.; Banaei, N.; Saleh, A. A. E.; Ermon, S.; Dionne, J. Rapid Identification of Pathogenic Bacteria Using Raman Spectroscopy and Deep Learning. *Nat Commun* **2019**, *10* (1), 1–8. https://doi.org/10.1038/s41467-019-12898-9.

(4) Harz, M.; Kiehntopf, M.; Stockel, S.; Rosch, P.; Straube, E.; Deufel, T.; Popp, J. Direct Analysis of Clinical Relevant Single Bacterial Cells from Cerebrospinal Fluid during Bacterial Meningitis by Means of Micro-Raman Spectroscopy. *J Biophotonics* **2009**, *2* (1–2), 70–80. https://doi.org/10.1002/jbio.200810068.

(5) Harz, M.; Rösch, P.; Popp, J. Vibrational Spectroscopy—A Powerful Tool for the Rapid Identification of Microbial Cells at the Single-Cell Level. *Cytometry Part A* **2009**, *75A* (2), 104–113. https://doi.org/10.1002/cyto.a.20682.

(6) Kastanos, E. K.; Kyriakides, A.; Hadjigeorgiou, K.; Pitris, C. A Novel Method for Urinary Tract Infection Diagnosis and Antibiogram Using Raman Spectroscopy. *Journal of Raman Spectroscopy* **2010**, *41* (9), 958–963. https://doi.org/10.1002/jrs.2540.

(7) Hamasha, K.; Mohaidat, Q. I.; Putnam, R. A.; Woodman, R. C.; Palchaudhuri, S.; Rehse, S. J. Sensitive and Specific Discrimination of Pathogenic and Nonpathogenic Escherichia Coli Using Raman Spectroscopy—a Comparison of Two Multivariate Analysis Techniques. *Biomed. Opt. Express, BOE* **2013**, *4* (4), 481–489. https://doi.org/10.1364/BOE.4.000481.

(8) Schie, I. W.; Huser, T. Methods and Applications of Raman Microspectroscopy to Single-Cell Analysis. *Appl Spectrosc* **2013**, *67* (8), 813–828. https://doi.org/10.1366/12-06971.

(9) Stöckel, S.; Kirchhoff, J.; Neugebauer, U.; Rösch, P.; Popp, J. The Application of Raman Spectroscopy for the Detection and Identification of Microorganisms. *Journal of Raman Spectroscopy* **2016**, *47* (1), 89–109. https://doi.org/10.1002/jrs.4844.

(10) Kloß, S.; Ro, P. Toward Culture-Free Raman Spectroscopic Identification of Pathogens in Ascitic Fluid. *Anal. Chem.* **2015**, 7.

(11) Reimer, L. G.; Wilson, M. L.; Weinstein, M. P. Update on Detection of Bacteremia and Fungemia. *Clin Microbiol Rev* **1997**, *10* (3), 444–465.

(12) Cronquist, A. B.; Mody, R. K.; Atkinson, R.; Besser, J.; D'Angelo, M. T.; Hurd, S.; Robinson, T.; Nicholson, C.; Mahon, B. E. Impacts of Culture-Independent Diagnostic Practices on Public Health Surveillance for Bacterial Enteric Pathogens. *Clinical Infectious Diseases* **2012**, *54* (suppl_5), S432–S439. https://doi.org/10.1093/cid/cis267.

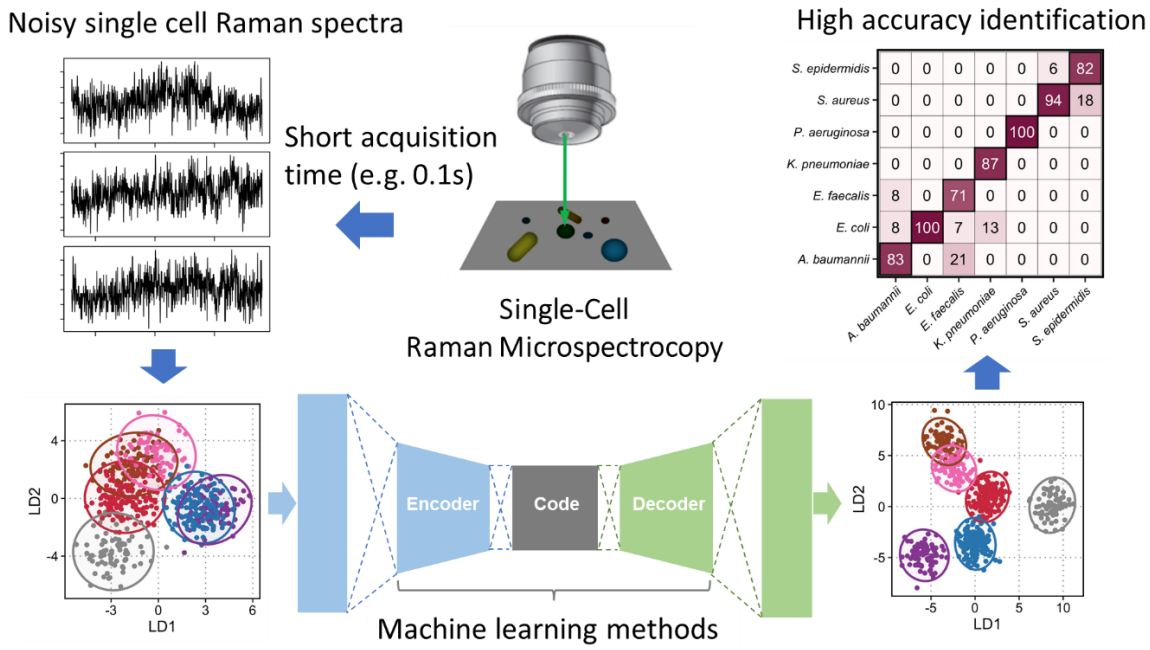
(13) Kang, D.-K.; Ali, M. M.; Zhang, K.; Huang, S. S.; Peterson, E.; Digman, M. A.; Gratton, E.; Zhao, W. Rapid Detection of Single Bacteria in Unprocessed Blood Using Integrated Comprehensive Droplet Digital Detection. *Nature Communications* **2014**, *5* (1), 5427. https://doi.org/10.1038/ncomms6427.

(14) Maurer, F. P.; Christner, M.; Hentschke, M.; Rohde, H. Advances in Rapid Identification and Susceptibility Testing of Bacteria in the Clinical Microbiology Laboratory: Implications for Patient Care and Antimicrobial Stewardship Programs. *Infect Dis Rep* **2017**, *9* (1). https://doi.org/10.4081/idr.2017.6839.

(15) Chung, J.; Kang, J. S.; Jurng, J. S.; Jung, J. H.; Kim, B. C. Fast and Continuous Microorganism Detection Using Aptamer-Conjugated Fluorescent Nanoparticles on an Optofluidic Platform. *Biosensors and Bioelectronics* **2015**, *67*, 303–308. https://doi.org/10.1016/j.bios.2014.08.039.
(16) Schroder, U. C.; Bokeloh, F.; O'Sullivan, M.; Glaser, U.; Wolf, K.; Pfister, W.; Popp, J.; Ducree, J.; Neugebauer, U. Rapid, Culture-Independent, Optical Diagnostics of Centrifugally Captured Bacteria from Urine Samples. *Biomicrofluidics* **2015**, *9* (4), 044118. https://doi.org/10.1063/1.4928070.
(17) Yi, X.; Song, Y.; Xu, X.; Peng, D.; Wang, J.; Qie, X.; Lin, K.; Yu, M.; Ge, M.; Wang, Y.; Zhang, D. et al. Development of a Fast Raman-Assisted Antibiotic Susceptibility Test (FRAST) for the Antibiotic Resistance Analysis of Clinical Urine and Blood Samples. *Anal Chem* **2021**, *93* (12), 5098–5106. https://doi.org/10.1021/acs.analchem.0c04709.
(18) Assmann, C.; Kirchhoff, J.; Beleites, C.; Hey, J.; Kostudis, S.; Pfister, W.; Schlattmann, P.; Popp, J.; Neugebauer, U. Identification of Vancomycin Interaction with Enterococcus Faecalis within 30 min of Interaction Time Using Raman Spectroscopy. *Anal Bioanal Chem* **2015**, *407* (27), 8343–8352. https://doi.org/10.1007/s00216-015-8912-y.
(19) Münchberg, U.; Rösch, P.; Bauer, M.; Popp, J. Raman Spectroscopic Identification of Single Bacterial Cells under Antibiotic Influence. *Anal Bioanal Chem* **2014**, *406* (13), 3041–3050. https://doi.org/10.1007/s00216-014-7747-2.
(20) Athamneh, A. I. M.; Alajlouni, R. A.; Wallace, R. S.; Seleem, M. N.; Senger, R. S. Phenotypic Profiling of Antibiotic Response Signatures in Escherichia Coli Using Raman Spectroscopy. *Antimicrobial Agents and Chemotherapy* **2014**, *58* (3), 1302–1314. https://doi.org/10.1128/AAC.02098-13.
(21) Moritz, T. J.; Taylor, D. S.; Polage, C. R.; Krol, D. M.; Lane, S. M.; Chan, J. W. Effect of Cefazolin Treatment on the Nonresonant Raman Signatures of the Metabolic State of Individual Escherichia Coli Cells. *Anal. Chem.* **2010**, *82* (7), 2703–2710. https://doi.org/10.1021/ac902351a.
(22) Palchaudhuri, S.; Rehse, S. J.; Hamasha, K.; Syed, T.; Kurtovic, E.; Kurtovic, E.; Stenger, J. Raman Spectroscopy of Xylitol Uptake and Metabolism in Gram-Positive and Gram-Negative Bacteria. *Appl Environ Microbiol* **2011**, *77* (1), 131–137. https://doi.org/10.1128/AEM.01458-10.
(23) Xu, J.; Webb, I.; Poole, P.; Huang, W. E. Label-Free Discrimination of Rhizobial Bacteroids and Mutants by Single-Cell Raman Microspectroscopy. *Anal. Chem.* **2017**, *89* (12), 6336–6340. https://doi.org/10.1021/acs.analchem.7b01160.
(24) Xu, J.; Preciado-Llanes, L.; Aulicino, A.; Decker, C. M.; Depke, M.; Gesell Salazar, M.; Schmidt, F.; Simmons, A.; Huang, W. E. Single-Cell and Time-Resolved Profiling of Intracellular Salmonella Metabolism in Primary Human Cells. *Anal. Chem.* **2019**, *91* (12), 7729–7737. https://doi.org/10.1021/acs.analchem.9b01010.
(25) Butler, H. J.; Ashton, L.; Bird, B.; Cinque, G.; Curtis, K.; Dorney, J.; Esmonde-White, K.; Fullwood, N. J.; Gardner, B.; Martin-Hirsch, P. L. et al. Using Raman Spectroscopy to Characterize Biological Materials. *Nat Protoc* **2016**, *11* (4), 664–687. https://doi.org/10.1038/nprot.2016.036.
(26) Maaten, L. van der; Hinton, G. Visualizing Data Using T-SNE. *Journal of Machine Learning Research* **2008**, *9* (Nov), 2579–2605.
(27) Kramer, M. A. Nonlinear Principal Component Analysis Using Autoassociative Neural Networks. *AIChE Journal* **1991**, *37* (2), 233–243. https://doi.org/10.1002/aic.690370209.

(28) Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.-A. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *J. Mach. Learn. Res.* **2010**, *11*, 3371–3408.
(29) http://www.chinets.com/Data/AntibioticDrugFast (accessed 2021 -01 -27).
(30) Weinstein, M. P. Performance Standards for Antimicrobial Susceptibility Testing, 30th Edition. Clinical and Laboratory Standards Institute January 21, 2020.
(31) Yuan, X.; Song, Y.; Song, Y.; Xu, J.; Wu, Y.; Glidle, A.; Cusack, M.; Ijaz, U. Z.; Cooper, J. M.; Huang, W. E.; Yin, H. Effect of Laser Irradiation on Cell Function and Its Implications in Raman Spectroscopy. *Appl. Environ. Microbiol.* **2018**, *84* (8), e02508-17. https://doi.org/10.1128/AEM.02508-17.
(32) Hsu, C.-C.; Xu, J.; Brinkhof, B.; Wang, H.; Cui, Z.; Huang, W. E.; Ye, H. A Single-Cell Raman-Based Platform to Identify Developmental Stages of Human Pluripotent Stem Cell-Derived Neurons. *PNAS* **2020**. https://doi.org/10.1073/pnas.2001906117.
(33) Fan, X.; Ming, W.; Zeng, H.; Zhang, Z.; Lu, H. Deep Learning-Based Component Identification for the Raman Spectra of Mixtures. *Analyst* **2019**, *144* (5), 1789–1798. https://doi.org/10.1039/C8AN02212G.
(34) Lussier, F.; Thibault, V.; Charron, B.; Wallace, G. Q.; Masson, J.-F. Deep Learning and Artificial Intelligence Methods for Raman and Surface-Enhanced Raman Scattering. *TrAC Trends in Analytical Chemistry* **2020**, *124*, 115796. https://doi.org/10.1016/j.trac.2019.115796.
(35) Le, B. T. Application of Deep Learning and near Infrared Spectroscopy in Cereal Analysis. *Vibrational Spectroscopy* **2020**, *106*, 103009. https://doi.org/10.1016/j.vibspec.2019.103009.
(36) Gordienko, Y.; Kochura, Y.; Alienin, O.; Rokovyi, O.; Stirenko, S.; Gang, P.; Hui, J.; Zeng, W. Dimensionality Reduction in Deep Learning for Chest X-Ray Analysis of Lung Cancer. *2018 Tenth International Conference on Advanced Computational Intelligence (ICACI)* **2018**, 878–883. https://doi.org/10.1109/ICACI.2018.8377579.
(37) Houston, J.; Glavin, F. G.; Madden, M. G. Robust Classification of High-Dimensional Spectroscopy Data Using Deep Learning and Data Synthesis. *J. Chem. Inf. Model.* **2020**, *60* (4), 1936–1954. https://doi.org/10.1021/acs.jcim.9b01037.

Table of Contents



Machine learning-aided Raman analysis renders high-speed identification of pathogens possible at 0.1 second per single-cell spectrum, paving the way for building extensive datasets and translating the technique to on-time clinical identification of pathogens and antibiotic guidance.