Received: 7 December 2023 | Revised: 1 January 2024 | Accepted: 4 January 2024

DOI: 10.1002/jbio.202300510

JOURNAL OF BIOPHOTONICS

RESEARCH ARTICLE

Rapid and accurate identification of marine bacteria spores at a single-cell resolution by laser tweezers Raman spectroscopy and deep learning

Jianchang Hu[1,2] | **Lin He**[2] | **Guiwen Wang**[3] | **Liwei Liu**[1] | **Yiping Wang**[1] | **Jun Song**[1] | **Junle Qu**[1,4] | **Xiao Peng**[1] | **Yufeng Yuan**[2]

[1]State Key Laboratory of Radio Frequency Heterogeneous Integration (Shenzhen University), College of Physics and Optoelectronic Engineering, Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, Shenzhen University, Shenzhen, Guangdong, China

[2]School of Electronic Engineering and Intelligentization, Dongguan University of Technology, Dongguan, Guangdong, China

[3]Institute of Eco-Environmental Research, Guangxi Academy of Sciences, Nanning, Guangxi, China

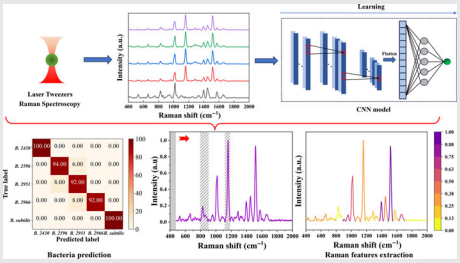
[4]Engineering Research Center of Optical Instrument and System, Ministry of Education, Shanghai Key Lab of Modern Optical System, School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, China

Correspondence
Xiao Peng, State Key Laboratory of Radio Frequency Heterogeneous Integration (Shenzhen University), College of Physics and Optoelectronic Engineering, Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, Shenzhen University, Shenzhen, Guangdong, 518060, China.
Email: pengxiao_px@szu.edu.cn

Yufeng Yuan, School of Electronic Engineering and Intelligentization, Dongguan University of Technology, Dongguan, Guangdong, 523808, China.
Email: yufengyuan@dgut.edu.cn

Funding information
National Key R&D Program of China, Grant/Award Number: 2021YFF0502900; National Natural Science Foundation of China, Grant/Award Numbers: 32060777, 12264005, 22327802, 62127819, 62175161, 62075137; Guangdong Basic and Applied Basic Research Foundation, Grant/Award Number: 2022A1515011845; Shenzhen Basic Research Program, Grant/Award Number: JCYJ20210324095810028; Shenzhen Science and Technology Program, Grant/Award Number: JCYJ20220818100202005; Shenzhen Key Laboratory of Photonics and Biophotonics, Grant/Award Number: ZDSYS20210623092006020; Dongguan

Abstract



Marine bacteria have been considered as important participants in revealing various carbon/sulfur/nitrogen cycles of marine ecosystem. Thus, how to accurately identify rare marine bacteria without a culture process is significant and valuable. In this work, we constructed a single-cell Raman spectra dataset from five living bacteria spores and utilized convolutional neural network to rapidly, accurately, nondestructively identify bacteria spores. The optimal CNN architecture can provide a prediction accuracy of five bacteria spore as high as 94.93% ± 1.78%. To evaluate the classification weight of extracted spectra features, we proposed a novel algorithm by occluding fingerprint Raman bands. Based on the relative classification weight arranged from large to small, four Raman bands located at 1518, 1397, 1666, and 1017 cm^{-1} mostly contribute to producing such high prediction accuracy. It can be foreseen that, LTRS combined with CNN approach have great potential for identifying marine bacteria, which cannot be cultured under normal condition.

KEYWORDS
bacteria identification, deep learning, extraction weight of spectra features, laser tweezers Raman spectroscopy (LTRS)

Science and Technology of Social Development Program, Grant/Award Number: 20231800936312; Dongguan University of Technology, Grant/Award Number: 221110080; Medical-Engineering Interdisciplinary Research Foundation of Shenzhen University, Grant/Award Numbers: 2023YG012, 2023YG033

1 | INTRODUCTION

Marine ecosystem is one of the most abundant biodiversity environments in world side, which is famous for microbial Garden of Eden [1]. Marine microorganisms including bacteria and archaea, occupy the vast majority of the total marine life [2]. Moreover, marine bacteria also play various important roles in maintaining ecosystem balance [3], degrading microplastics [4], and participating in carbon/sulfur/nitrogen cycles [5]. In addition, marine bacteria have been considered as important resources for producing clinically and industrially important secondary metabolites [6]. For example, carotenoids produced by marine bacteria have shown significant antioxidant properties, which are beneficial for preventing oxidative stress-related human diseases such as cardiovascular, neurodegenerative, and inflammatory diseases [7]. It is worth noting that, marine bacteria usually live in extreme conditions of deep ocean such as extreme low temperature, high pressure and high salinity, which are difficult for culturing marine bacteria using normal conditions [8]. Here, there is a challenge is that, how to accurately identify rare marine bacteria without a culture process in a single-cell resolution level is significant and valuable.

For non-culturable marine bacteria, classic phenotypic approaches including agar plate and broth culture cannot multiply bacteria rapidly. Nucleic acid amplification method is a classic tool to identify marine bacteria based on polymerase chain reaction (PCR) technology. For instance, the 16S ribosomal RNA gene sequence has proved to be powerful in studying the diversity and identifying marine microbial communities, especially for non-culturable microorganisms [9]. However, it is a destructive method, and it needs time-consuming cell process. In addition, the mass spectrometry represented by MALDI-TOF, has been introduced into bacteriology, showing high identification accuracy [10]. However, it requires a comprehensive spectral database, and the tested marine microorganisms should be presented in the spectral database. Then, it is not suitable to identify unknown marine bacteria collected from deep ocean region. Therefore, it is considerably significant to propose a novel recognition strategy of marine bacteria via a non-destructive approach with a resolution of single-cell level.

It is generally accepted that, Raman spectroscopy (RS) including surface enhanced Raman spectroscopy (SERS) has been confirmed to be a non-destructive analysis method employed for characterizing microorganism [11], monitoring plasmon-mediated chemical reactions [12], and controlling food quality [13], and so forth. For microorganism recognition, the measured Raman spectra of microorganism strains via confocal RS are averaged population features, which cannot accurately reveal the individual spectra information. It is generally accepted that, the heterogeneity of bacteria cells may exist at a single-cell level, highly depending on various colony strains, culture conditions, and other stressful factors [14]. These stated variables make it difficult for characterizing bacteria species in a single-cell resolution. Fortunately, the emergence of LTRS can efficiently solve this dilemma. LTRS was developed by successfully integrating laser optical tweezers with confocal RS technology. Laser optical tweezers can be employed to trap a single bacteria cell, and confocal RS can be used to collect the finger-print Raman scattering signal of trapped bacteria cell. Therefore, LTRS has great potential in studying the heterogeneity of Raman spectra measured from bacteria cells with a single-cell resolution [15].

In most microorganism cells, the main components named primary metabolites are quite similar, such as nucleic acids, proteins, and lipids, etc. Owing to various metabolic states, the levels of three main components are always changing. For some marine bacteria cells, they can produce high levels of secondary metabolites such as carotenoids, chlorophyll, and poly-beta-hydroxybutyrate (PHB). In the later stage of growth, the content of secondary metabolites in marine bacteria is even much higher than that of primary metabolites. Generally, these metabolites in a single cell are very few, and it is indeed difficult for performing the classification and identification of marine bacteria species only via LTRS technology.

To perform the classification of bacteria, there are two main classification model based on machine learning including unsupervised and supervised approaches. Both principal component analysis (PCA) [16] and hierarchical cluster analysis (HCA) [17] belong to unsupervised methods, and they can determine the identical classification by judging the spectra similarity from different

bacteria species. In addition, both soft independent modeling of class analogy (SIMCA) [18] and support vector machine (SVM) [19] are supervised learning models, having the functions of training and prediction spectra data. Unlike unsupervised model, there is one point needs to be clear is that, prior to prediction, it is required to define the labels for the training classification model. After the classification model is well trained, it can be employed to judge the ascription of new spectra without being trained. The SIMCA is developed on the basic of PCA classification results, and it determines the ascription of unknown spectra by calculating the distance between unknown spectra and PCA model. However, for PCA classification, it mainly depends on the number of artificially selected principal components. SVM is mainly employed to deal with linearly inseparable spectra data, having high prediction accuracy and specificity in binary classification. However, SVM cannot efficiently deal with multiple classification issues. Therefore, exploring more advanced classification algorithms employed for accurately identifying bacteria are highly desirable.

In artificial intelligence (AI) field, machine learning belongs to a subset of AI. Meanwhile, deep learning is an advanced evolution of machine learning. In addition, convolutional neural network (CNN) algorithm is one of the most popular architectures [20]. Originally, CNN was successfully employed to extract image features [21]. Afterward, CNN was introduced into spectra analysis field [22]. To date, CNN combined with Raman spectroscopy have shown great potential in characterizing microorganisms because of its powerful recognition ability [11a, 23]. It is worth noting that, CNN architecture act as a black box, and it cannot export the extracted spectra features directly. Although a host of studies have shown strong identification ability of CNN, few of them can reveal the contribution of spectra features for reported high identification accuracy.

Herein, we proposed a rapid, non-destructive, and accurate method for identifying marine bacteria with a single-cell resolution via LTRS combined with CNN model, as illustrated in Figure 1. The single-cell Raman spectra dataset was constructed by 750 Raman spectra of five bacteria, which was randomly divided into two parts: training (90%) dataset and testing (10%) dataset. The spectra feature of training (90%) dataset was deeply learned and memorized by CNN model, and the predicted dataset was employed to evaluate the classification performance of optimal CNN model. For the tested marine bacteria, the averaged prediction accuracy was approximately 94.93% ± 1.78%. More importantly, a

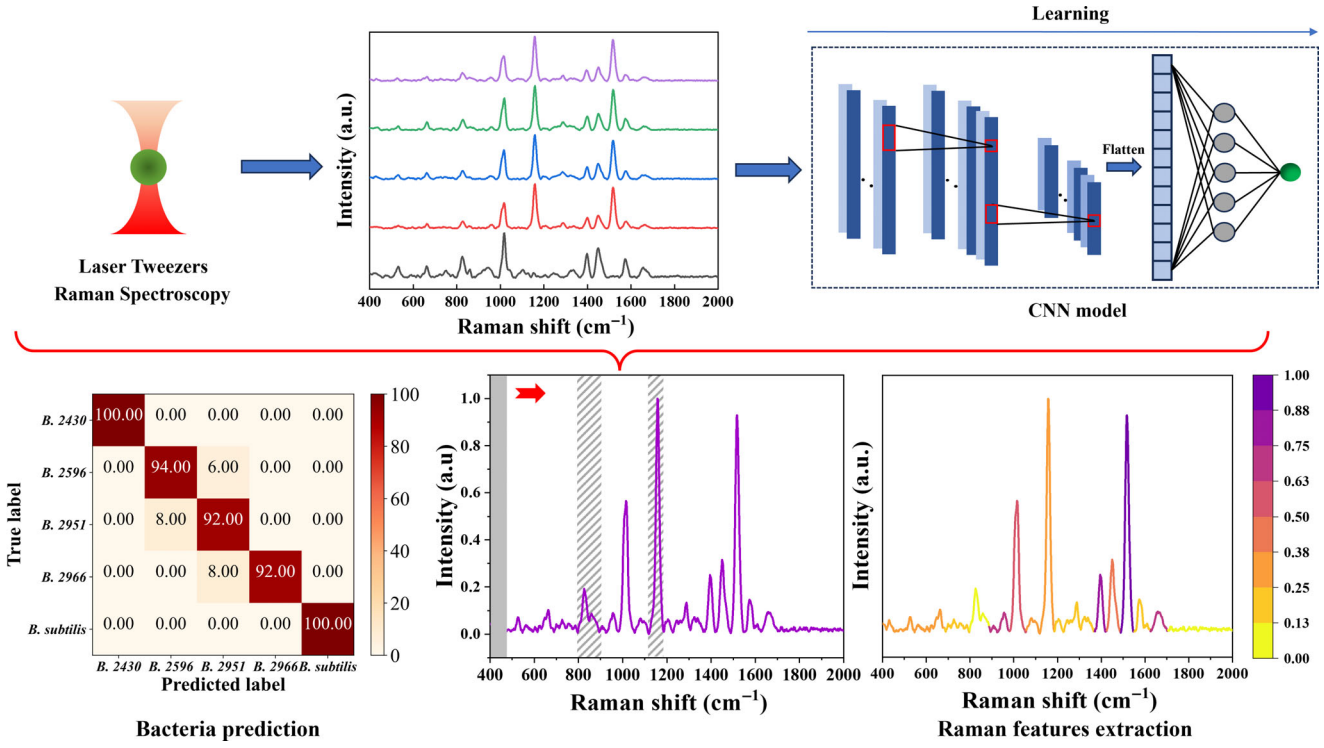


FIGURE 1 Illustration of rapid and accurate identification of marine bacteria spores using LTRS combined with CNN at a single-cell resolution. *Bacillus marisflavi* (MCCC1K02430) is short for *B. 2430*; *Bacillus aryabhata* (MCCC1K02966) is short for *B. 2966*; *Bacillus aerius* (MCCC1K02596) is short for *B. 2596*; *Bacillus nealsonii* (MCCC1K02951) is short for *B. 2951*; *Bacillus subtilis* (CICC63501) is short for *B. subtilis*.

novel approach to reveal the respective classification weight of Raman spectra features was proposed by introducing a sliding spectra interval. The visualization plot of classification weight clearly showed that, these four Raman bands located at 1518, 1397, 1666, and 1017 cm^{-1} mainly contribute to such high prediction accuracy of 94.93% ± 1.78%.

2 | MATERIALS AND METHODS

2.1 | Bacterial strains and cultivation of bacteria spores

In our study, two different culture mediums (Marine broth 2216, Nutrient broth) were employed to cultivate five bacillus strains including *Bacillus marisflavi* (MCCC1K02430), *Bacillus aryabhata* (MCCC1K02966), *Bacillus aerius* (MCCC1K02596), *Bacillus nealsonii* (MCCC1K02951), and *Bacillus subtilis* (CICC63501). Agar (1182GR500) was bought from BioFroxx (Einhausen, Germany). Marine broth 2216 (212185) purchased from Becton Dickinson (New Jersey, USA) was employed to cultivate marine bacteria strains. Nutrient broth (CM1168) purchased from Thermo Scientific (Basingstoke, UK) was employed to cultivate *Bacillus subtilis* (CICC63501), which was bought from the China Center of Industrial Culture Collection (CICC). It is worth noting that, the other four marine bacteria strains were collected from China Yellow Sea with a geographic coordinate (77° 31′ N, and 69° 19′ W in July 2010). In brief, four seawater water samples containing marine bacteria were collected into 10 liter Niskin bottles from various depths (MCCC1K02430, 5152 kilometers; MCCC1K02966, 2.7 kilometers; MCCC1K02596, 3.309 kilometers; MCCC1K02951, 2.3 kilometers). Afterward, these seawater water samples were rapidly transferred into 50 mL sterile centrifuge tubes and stored at 4°C fridge in dark condition. To obtain spore-forming marine bacteria, 50 mL seawater were filtered by a nylon filter. The obtained re-suspension was coated onto marine broth 2216 with 1.5% agar and cultivated at 30°C. Several days later, single colonies grown on agar medium were extracted and further cultivated in a new marine broth 2216 with 1.5% agar. To guarantee the purity of isolated marine bacteria strains, the process was repeated no less than three times. The chemical manganese sulfate ($MnSO_4$, Purity ≥99.5%) was bought from Tianjin Kemiou Chemical Reagent Co., Ltd. In all experimental process, the ultra-pure water (18.2 MΩ/cm) was utilized.

Agar plates composed of 1.87 g marine broth/nutrient broth, 0.75 g agar, and 0.25 mg $MnSO_4$ were selected to cultivate bacteria spores for 72 hours at a stable temperature of 36.5°C. Then, the bacteria spores could be obtained by centrifugation and washing process, respectively. Finally, the samples of bacteria spores were stored in a 4°C fridge.

2.2 | Laser tweezers Raman spectroscopy

To obtain the Raman scattering signal from a single cell, we built a setup of LTRS system excited by a near-infrared (NIR) wavelength at 780 nm, as illustrated in Figure 2. Unlike commercial equipment, only one laser beam was designed to form an optical well, and the same laser beam was employed to excite the Raman scattering of single bacteria spore in optical well synchronously. A NIR laser beam with a Gaussian profile was coupled into an inverted microscope setup (TE2000U, Nikon). With the help of an oil immersion objective (Nikon, 100×, N.A. = 1.30), the NIR laser beam was focused to form an invisible optical well. Under the physiological condition, single bacteria spore can be selected and captured by shifting the optical well. Meanwhile, the Raman scattering of trapped bacteria spore can be excited, which was collected by an oil immersion objective. Finally, the signal of Raman scattering was coupled into a LS785 spectrometer. In addition, phase contrast imaging of trapped bacteria spore was measured using a CCD (PIXIS 400BR, Princeton Instruments). The resolution of our home-built LTRS system was 6 cm^{-1}, which needs to be calibrated by the standard Raman scattering signal of 2.0 μm polystyrene spheres.

2.3 | Spectra measurement and data process

For each species of bacteria spore, no less than 150 single-cell Raman spectra were yielded. The measurement spectral interval was from 400 to 2000 cm^{-1}, responding to the finger-print region of spectral feature. The time for collecting Raman signal was fixed to be 10 s. Prior to each spectral measurement of bacteria spore, the excitation laser was shut off. In addition, three Raman spectra from background were also measured. Using self-compiled Matlab code, the obtained Raman spectra of bacteria spores were preprocessed by subtracting the background spectra, and correcting spectra baseline, respectively. Then, the preprocessed Raman spectra data were performed by a normalization method. Finally, a Raman spectra dataset containing no less than 750 single-cell Raman spectra of five species of bacteria spore was constructed. To compare with the Raman scattering signals obtained by LTRS, the population Raman spectra

of precipitation from five bacteria spores were measured by a confocal Raman spectrometer with an excitation wavelength of 532 nm.

2.4 | Atomic force microscopy (AFM) imaging of bacteria spore

To characterize the topological morphology of five bacteria spores, AFM images of five bacteria spores were measured by a Bioscience AFM system (NanoWizer 4, Bruker, USA). In brief, bacteria spore suspension was deposited onto a ploy-coated slide, and was dried for 1 h at home temperature. Afterward, the ploy-coated slide was fixed by sample holders. The AFM probe was made of a silicon tip having a resonance frequency of 320 kHz and an elasticity coefficient of 42 N/m. All the AFM images were scanned via a frequency of 1.0 Hz.

2.5 | Bacteria spore identification

The CNN model employed for bacteria spore classification was proposed, as shown in Figure 3. It contains seven principal components: an input layer, two convolutional layers, two max pooling layers, a fully-connected

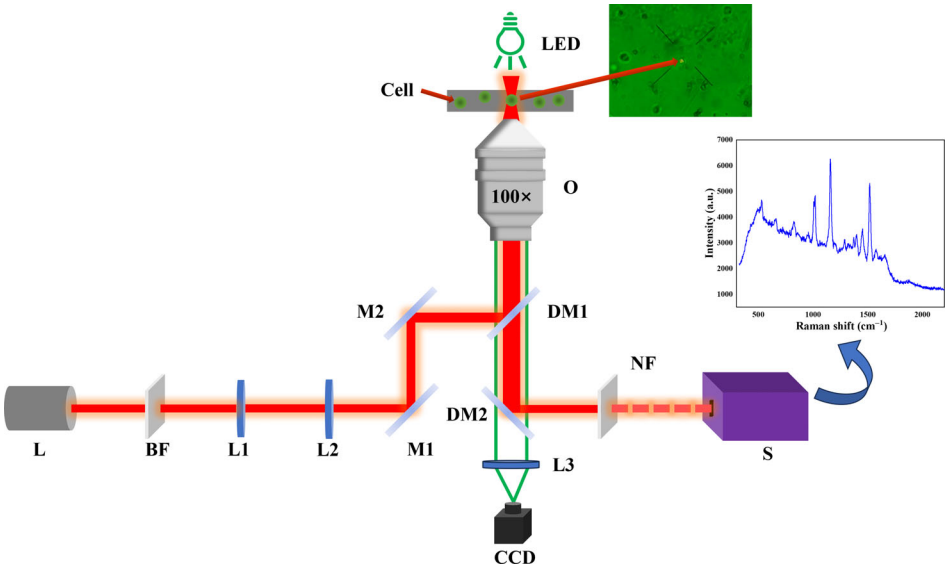


FIGURE 2 The built of LTRS with a single-cell resolution. BF, band filter; CCD, charge-coupled device; DM, dichroic mirror; L, lens; L, laser source; LED, Light-emitting diode; M, mirror; NF, notch filter; O, objective; S, spectrometer.

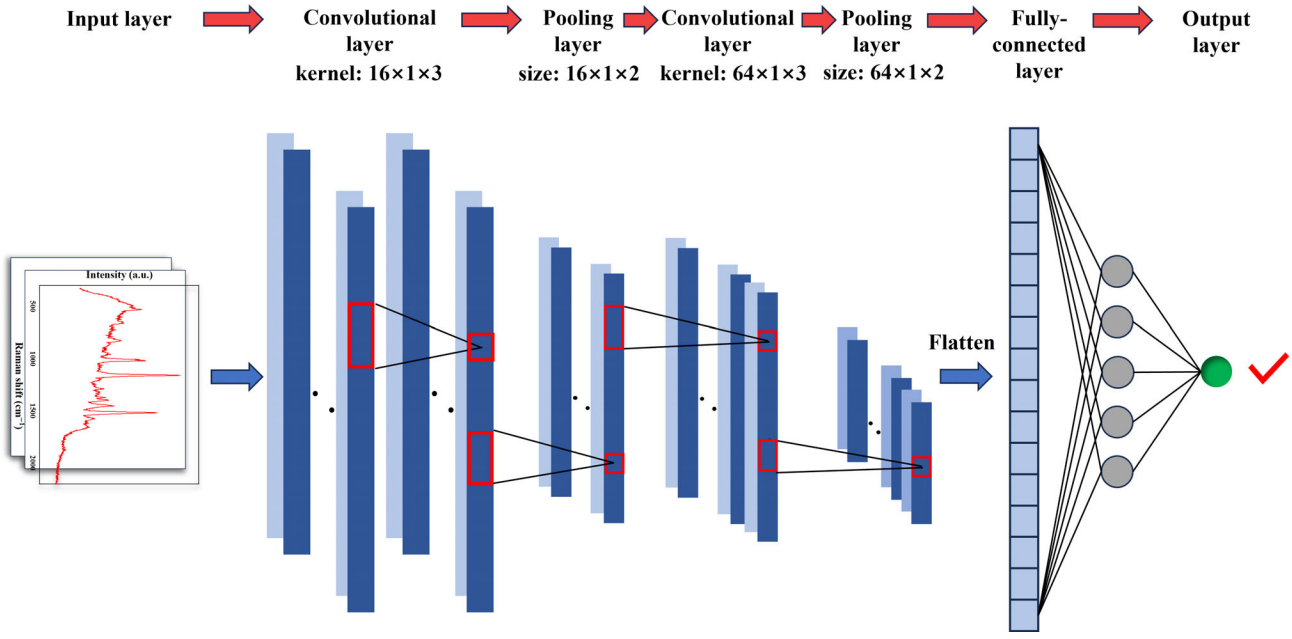


FIGURE 3 Illustration of CNN configuration.

layer and an output layer. It is worth noting that, input layer can randomly select 90% of the total Raman spectra dataset as training set, and the remain spectra data were automatically categorized as testing set. The purpose of convolutional layer was introduced to extract spectra features by controlling the kernel sizes. In this study, the kernel size was set to be 1×3. In addition, the number of filters in the first convolution layer was fixed to be 16, and the number of filters in the second convolution layer was fixed to be 64. Then, the spectra dataset was propagated to the max pooling layer with a filter size of 1×2. Here, the function of max pooling layer was to decrease the number of parameters and hold the feature maps. With help of both two convolutional layers and max pooling layers, a fully-connected layer was introduced to transform multi-dimensional data into one-dimensional data. Finally, the maximum probability of prediction dataset was output via a confusion matrix.

To show the superiority of our CNN model, conventional machine learning methods were also employed to process the same Raman spectra dataset. It is well-known that, conventional machine learning approaches including PCA, HCA, SIMCA, and SVM, are classical methods to achieve classification or pattern recognition. In PCA model, seven principal components were selected to extract the spectra features, and the scores of three principal components (PC-1, PC-2, PC-3) were depicted in a three-dimensional space. For HCA, the Euclidean distance was employed to classify five bacteria spores, and the classification tree was plotted. For both SIMCA and SVM, the Raman spectra dataset was divided into two parts: training set (90%), and testing set (10%). As SIMCA was performed based on various PCA models, the Coomans' plots defining as the distance between unlabeled Raman spectra data to well-trained PCA models could be obtained. For SVM, the confusion matrix was employed to evaluate the prediction classification performances.

3 | RESULTS AND DISCUSSION

3.1 | Average single-cell Raman spectra of bacteria spores

Figure 4 exhibits the averaged single-cell Raman spectra originating from five bacteria spores including *Bacillus marisflavi* (MCCC1K02430), *Bacillus aryabhata* (MCCC1K02966), *Bacillus aerius* (MCCC1K02596), *Bacillus nealsonii* (MCCC1K02951), and *Bacillus subtilis* (CICC63501), respectively. It can be found that, the single-cell Raman spectra from identical specie of bacteria spores have a tiny heterogeneity, due to the heterogeneity of individual spore, which is beneficial for developing the robustness of CNN model. More importantly, the obtained Raman spectra from single cell could reveal real-time signals on intracellular compounds. For example, Ca^{2+}-dipicolinic acid (Ca-DPA) is a typical biomarker in bacteria spore, whose Raman bands mainly locate at 660, 826, 1017, 1397, 1449, and 1576 cm^{-1}, as listed in Table 1. Unlike the *Bacillus subtilis* (CICC63501) specie, the four marine bacteria spores (MCCC1K02430, MCCC1K02966, MCCC1K02596, and MCCC1K02951) can synthesize the carotenoids. Moreover, there are strong Raman scattering signals of carotenoids (Table 1) located at 1158, and 1518 cm^{-1}, respectively. In addition, we also observed a relatively weak Raman band 1666 cm^{-1}, which belongs to vibration of Amide I (protein). To show the superiority of LTRS, the population Raman spectra of precipitation from five bacteria spores, was shown in Figure S1. Obviously, the Raman scattering

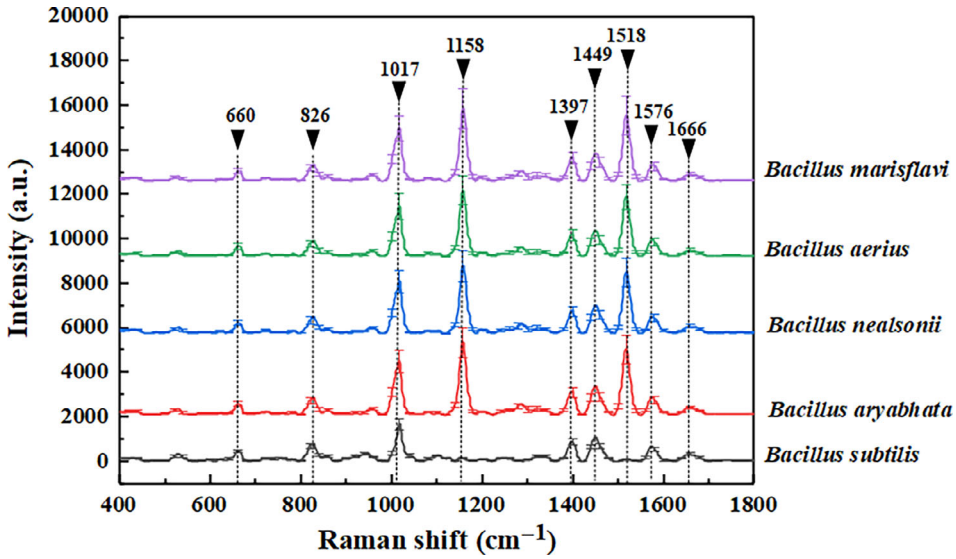


FIGURE 4 Typical single-cell Raman spectra of five bacteria spores. The excitation wavelength is 780 nm, and the measurement time is 10 s.

signals of Ca-DPA in four marine bacteria spores were significantly weakened. Moreover, compared with single-cell Raman spectra of bacteria spores, the signal-to-noise ratio (SNR) of population Raman spectra was poor. Therefore, single-cell Raman spectra approach can provide more rich information on intracellular compounds than confocal Raman spectroscopy. In addition, the morphology of five bacteria spores was characterized by AFM approach, as shown in Figure S2. It clearly showed that, it is also impossible to accurately recognize bacteria spores only via topological morphology method. For our obtained single-cell Raman spectra of bacteria spores, it is easy to distinguish the *Bacillus subtilis* from other four marine bacteria spores via the production of carotenoids. However, it is still difficult to recognize the four marine bacteria spores, because they have almost identical Raman features. Thus, it is significant to propose a novel strategy for classifying marine bacteria with high identification accuracy.

TABLE 1 Assignment of typical Raman bands in bacteria spores [24].

Raman band (cm^{-1})	Assignment	Vibration mode
660	Ca-DPA	Bending vibration of C—C in pyridine ring
826	Ca-DPA	Out-of-plane deformation of C—H
1017	Ca-DPA	Symmetric stretching of pyridine ring
1158	Carotenoids	Stretching vibrations of C—C and C═C
1397	Ca-DPA	Symmetric stretching of O—C—O
1449	Ca-DPA	Symmetric bending of C—H in pyridine ring
1518	Carotenoids	Stretching vibrations of C—C and C═C
1576	Ca-DPA	Asymmetric stretching of O—C—O
1666	Protein	Vibration of Amide I (beta-sheet)

3.2 | Photodamage effect on marine bacteria

It is generally accepted, NIR laser is considered to produce little photodamage for biological samples. However, the carotenoids in living marine bacteria spores have shown strong light sensitivity, and it is easy to rapidly degrade the carotenoids using the laser irradiation with high power. Therefore, it is still necessary to remove the spectra difference caused by photodamage. Under confocal conditions, the power on samples was approximately 8 mW. Here, we counted a batch of time-dependent Raman spectra of five bacteria spores in single-cell level, as depicted in Figure 5 and Figure S3. In a short measurement time of 20 s, a batch of single-cell Raman spectra were collected by every 1 s. Figure 5A showed that, the

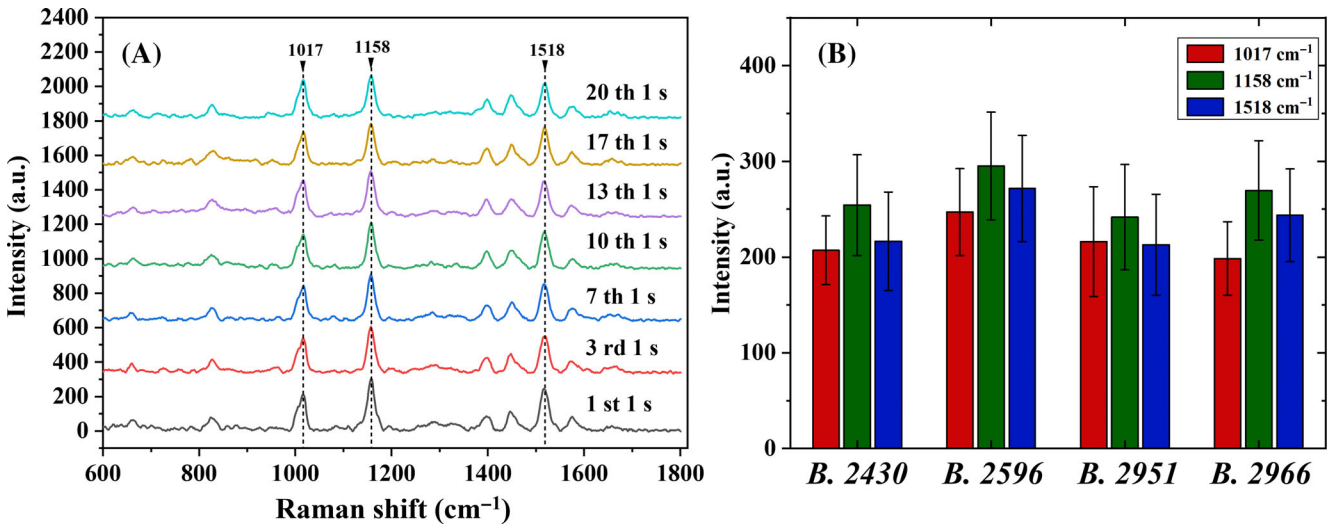


FIGURE 5 Photodamage effect on bacteria spores. (A) Time-dependent Raman spectra of *Bacillus marisflavi* (MCCC1K02430) cell measured by LTRS on various time points in single-cell analysis. The total collection time is 20 s, and the collection time of each Raman spectrum is 1 s. The laser power of 780 nm on sample is approximately 8 mW. (B) Statistic average intensity of three Raman bands at 1017, 1158, and 1518 cm^{-1}. For each bacteria species, no less than 20 cells were counted.

relative intensity of Raman band at 1017 cm^{-1} ascribed to Ca-DPA of a single cell was stable with increasing laser irradiation time. In addition, there was no significant change in two Raman bands (1158 and 1518 cm^{-1}) belong to carotenoids. Moreover, we calculated the averaged intensity of three Raman bands (1017, 1158, and 1518 cm^{-1}) based on 20 *Bacillus marisflavi* (MCCC1K02430), as shown in Figure 5B. It can be found that, the Raman intensity of single-cell analysis method is in accordance with the statistics result of 20 cells. Similar observations are also found for other four bacteria spores, as illustrated in Figures 5B and S3. In all, it indicates that the 780 nm laser with 8 mW showed little photodamage effect on bacteria spores, especially for marine bacteria spores.

3.3 | Bacteria spore identification with high sensitivity and specificity via CNN

To achieve high accuracy identification of five bacteria spores, the CNN architecture was employed to recognize the feature information on Raman spectra dataset from five bacteria spore species. Considering that the obtained Raman spectra data were one-dimensional data arrays, one-dimensional CNN model was selected. For a dataset containing Raman spectra with a number of 750, 90% of them was introduced into CNN architecture for training to obtain a best taxonomic model. The remained Raman spectra were automatically categorized into testing dataset. In the process of training, both loss function and gradient descent optimization algorithm play essential roles in obtaining an optimal taxonomic model. For example, if the loss function from validation dataset begin to show an upward trend, it may be overfitted. And if the selected learning rate is not appropriate, it is possible to be trapped in a local optimal solution. Therefore, it is vital to control the training epochs and determine an appropriate gradient descent optimizer. Generally, the loss function on both training and validation datasets is an important indicator for monitoring the overfitting during the CNN training process [21]. For example, if the loss function of training dataset is obviously lower than the validation dataset, it is likely to be overfitting. In addition, if the loss function of validation dataset tends to be a rising state, the overfitting is also likely to occur.

In our study, we selected an adaptive moment estimation (Adma) function with three parameters ($\beta1 = 0.9$, $\beta2 = 0.999$, and learning rate $= 0.00005$) to optimize the CNN model. And the validation dataset was randomly selected 20% of the training dataset through the calculation code. When the training epochs was fixed to be 100, the prediction accuracy was 78.67%, as shown in Figure S4A,B. The loss function showed that, the obtained CNN model was underfitting. However, the prediction accuracy was relatively poor. As the training epochs was increased to be 200, the prediction accuracy was developed to be 86.67% (Figure S4C,D). Meanwhile, the CNN model was still underfitting. Next, as the training epochs was further increased to be 300, both training and validation loss function exhibited a stable state, as depicted in Figure S4E. At the same time, the curve of training loss function was smaller than the validation loss. Moreover, the accuracy of both training and validation loss tended to be in a stable state (Figure S4F).

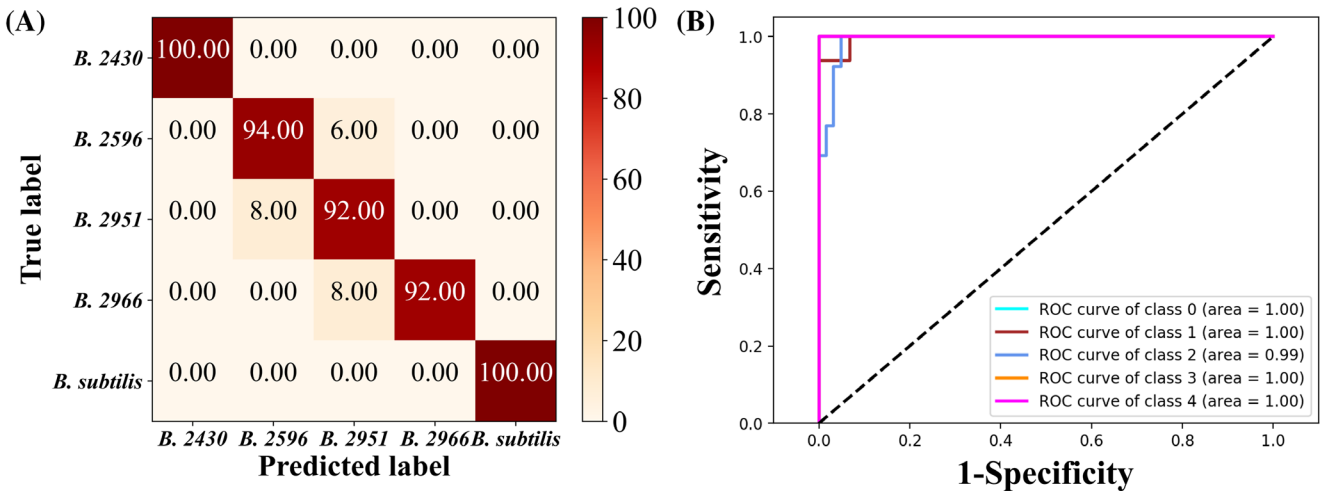


FIGURE 6 (A) Prediction accuracy of five bacteria spores obtained by running our optimal CNN model one time at 300 epochs. Noting that, the abbreviations such as *B. 2430*, *B. 2596*, *B. 2951*, *B. 2966*, and *B. subtilis* stand for *Bacillus marisflavi* (*MCCC1K02430*), *Bacillus aryabhata* (*MCCC1K02966*), *Bacillus aerius* (*MCCC1K02596*), *Bacillus nealsonii* (*MCCC1K02951*), *and Bacillus subtilis* (*CICC63501*), respectively. (B) High sensitivity and specificity provided by our proposed CNN architecture.

It suggested that, the optimal CNN model has shown no obvious indicator of both underfitting and overfitting. More importantly, the prediction accuracy was enhanced to be 96%, as shown in Figure 6A. In addition, more larger training epochs such as 500, 1000, 1500 were also studied to examine the overfitting of CNN model, as illustrated in Figure S5. It can be found that, the curves of validation loss function were obviously unstable, indicating that the CNN model began to be overfitting as the training epochs were more than 500. Finally, the optimal CNN model at 300 epochs was repeated by 100 times, an average classification accuracy of 94.93% ± 1.78% can be obtained.

To determine the specificity and sensitivity of our proposed CNN model, the receiver operating characteristic (ROC) curve was plotted in Figure 6B. For the five bacteria spores, the averaged area under ROC curve was larger than 0.99, suggesting that the optimal CNN architecture can identify five bacteria spores, due to both high specificity and high sensitivity.

To compare the superiority of CNN with high specificity and sensitivity, conventional machine learning approaches (PCA, HCA, SIMCA, SVM) were also employed to study the same Raman spectra datasets. For PCA, the total score from three maximum principal components (PC-1, PC-2, and PC-3) was determined to be 94%, as shown in Figure S6. However, the 3D plot of three maximum principal components clearly showed that, except for *Bacillus subtilis*, other four marine bacteria spores were closely gathered together. It indicated that, PCA approach cannot effectively classify five bacteria spores. Based on four PCA classification models, SIMCA (Figure S7) showed an average prediction accuracy of 40% and 41%, respectively. Using HCA model, a host of Raman spectra from five bacteria spores were unexpectedly categorized into one cluster, as shown in Figure S8. For the SVM method, the average classification and prediction accuracy was 50.67% and 56%, respectively (Figure S9). Here, we can conclude that, LTRS integrated with CNN method can realize accuracy identification of bacteria spores.

3.4 | Classification weight extraction based on occluded Raman bands

Although our optimal CNN architecture can efficiently identify five bacteria spores, we don't know which feature information should contribute to yielding such high specificity and sensitivity. Indeed, CNN has proven its strong recognition ability, but implicitly covers useful spectra features at the same time. To reveal and extract this feature information hidden in the black box of CNN, we proposed a novel algorithm named occluded Raman band approach to evaluate the weights of spectra features, as shown in Figure 7A. For each typical Raman band from 400 to 2000 cm^{-1}, we chose a sliding window to occlude it. Then, the prediction accuracy was calculated by running CNN model at least 10 times. It can be assumed that, a typical Raman band occluded by a sliding window produce a relatively large effect on the prediction accuracy of CNN model, and it is very likely to facilitate the precise identification of five bacteria spores. Here, the prediction accuracy of original spectra was defined as P_{Ori}, and the prediction accuracy of occluded Raman bands was called as P_{Occ}. Then, the relative weights of occluded Raman bands can be calculated by an equation ($C = \frac{P_{Ori} - P_{Occ}}{P_{Ori}} \times 100\%$). Based on the relative weights, the contribution of occluded Raman bands can be visually plotted, as shown in Figure 7B and Table 2. It clearly showed that, according to the relative classification weight arranged from large to small, these four Raman bands located at 1518, 1397, 1666, and 1017 cm^{-1} mostly contributed to such high specificity and sensitivity. According to the assignment of Raman bands, these

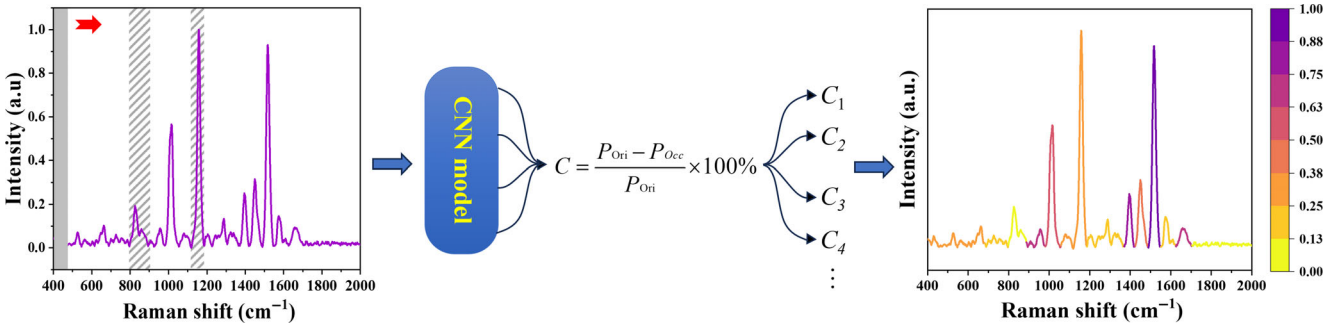


FIGURE 7 Illustration of classification weight extraction based on an occluded Raman band algorithm. C, contribution; P_{Occ}, prediction accuracy of occluded spectra; P_{Ori}, prediction accuracy of original spectra. The contribution value is the averaged contribution by running the CNN model more than 10 times. After obtaining the contribution of each occluded Raman band, the contribution value was normalized into an interval of [0, 1].

TABLE 2 The calculated prediction accuracy and classification weight from various occluded Raman bands.

Occluded Raman bands (cm^{-1})	Prediction accuracy	Classification weight
400–560	94.13%	0.09%
560–682	95.60%	0.06%
682–796	94.54%	0.05%
796–885	95.07%	0.00%
977–1055	93.60%	1.47%
1055–1184	94.13%	0.09%
1184–1368	95.34%	0.04%
1368–1423	93.07%	2.03%
1423–1487	93.87%	1.19%
1487–1549	92.67%	2.45%
1549–1623	94.53%	0.05%
1623–1704	93.47%	1.61%

differences in spectra features of five bacteria spores are mainly focused on intracellular compounds: carotenoids, Ca-DPA, and amide I.

4 | CONCLUSION

In this work, we proposed a novel approach to identify marine bacteria using LTRS combined with CNN model at a single-cell resolution. LTRS can measure real-time spectral information for intracellular compounds including Ca-DPA, carotenoids, and amide I. Compared to conventional machine learning methods, our proposed CNN model can realize a high prediction accuracy of 94.93% ± 1.78% for five bacteria spores. To figure out the contribution of spectra features, a novel method of classification weight extraction named occluded Raman band was proposed. Based on the relative classification weight arranged from large to small, the contribution from four Raman bands located at 1518, 1397, 1666, and 1017 cm^{-1} mainly facilitate identification of five bacteria spores. It can be expected that, our approach will provide a novel method for identifying unculturable marine bacteria with high accuracy, specificity, and sensitivity in future.

AUTHOR CONTRIBUTIONS

Yufeng Yuan conceived the study. Jianchang Hu performed bacteria cell cultivation, Raman spectra measurement, and construction of CNN. Lin He and Guiwen Wang helped the optimization of LTRS system. Guiwen Wang, Liwei Liu, Yiping Wang, Jun Song, Junle Qu, Xiao Peng, and Yufeng Yuan discussed and analyzed the experiment data. Jianchang Hu wrote the manuscript, and Yufeng Yuan revised the manuscript. Yufeng Yuan and Xiao Peng supervised the project.

ACKNOWLEDGMENTS

This work was partially supported by the National Key R&D Program of China (2021YFF0502900), the National Natural Science Foundation of China (62075137/62175161/62127819/22327802/12264005/32060-777), the Guangdong Basic and Applied Basic Research Foundation (2022A1515011845), Shenzhen Basic Research Program (JCYJ20210324095810028), Shenzhen Science and Technology Program (JCYJ20220818100202005), Shenzhen Key Laboratory of Photonics and Biophotonics (ZDSYS20210623092006020), Dongguan Science and Technology of Social Development Program (20231800936312), the high-level talent program of Dongguan University of Technology (No. 221110080), and Medical-Engineering Interdisciplinary Research Foundation of Shenzhen University (2023YG033/2023YG012).

CONFLICT OF INTEREST STATEMENT

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

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Liwei Liu https://orcid.org/0000-0002-4593-665X
Junle Qu https://orcid.org/0000-0001-7833-4711
Yufeng Yuan https://orcid.org/0000-0001-7472-4368

REFERENCES

[1] A.-S. Heiskanen, T. Berg, L. Uusitalo, H. Teixeira, A. Bruhn, D. Krause-Jensen, C. P. Lynam, A. G. Rossberg, S. Korpinen, M. C. Uyarra, A. Borja, *Front. Mar. Sci.* **2016**, *3*:00184. https://doi.org/10.3389/fmars.2016.00184
[2] H. C. Flemming, S. Wuertz, *Nat. Rev. Microbiol.* **2019**, *17*, 247.
[3] S. Das, N. Mangwani, *Oceanologia* **2015**, *57*, 349.
[4] (a) R. Gao, C. Sun, *J. Hazard Mater.* **2021**, *416*, 125928. (b) S. Oberbeckmann, M. Labrenz, *Ann. Rev. Mar. Sci.* **2020**, *12*, 209.
[5] (a) A. York, *Nat. Rev. Microbiol.* **2018**, *16*, 259. (b) Y. Han, M. Zhang, X. Chen, W. Zhai, E. Tan, K. Tang, *Environ. Int.* **2022**, *158*, 106889.
[6] (a) F. Pietra, *Nat. Prod. Rep.* **1997**, *14*, 453. (b) J. A. Moghaddam, T. Jautzus, M. Alanjary, C. Beemelmanns, *Org. Biomol. Chem.* **2021**, *19*, 123.
[7] M. A. Gammone, G. Riccioni, N. Orazio, *Mar. Drugs* **2015**, *13*, 6226.
[8] Z. Zhang, Y. Wu, X.-H. Zhang, *Deep Sea Research Part II: Topical Studies in Oceanography*, Vol. 155 **2018**, p. 34.
[9] F. Valenzuela-González, R. Casillas-Hernández, E. Villalpando, F. Vargas-Albores, *Cienc. Mar.* **2015**, *41*, 297.

[10] C. Lozano, M. Kielbasa, J.-C. Gaillard, G. Miotello, O. Pible, J. Armengaud, *Microorganisms* **2022**, *10*, 719.
[11] (a) 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, *Nat. Commun.* **2019**, *10*, 4927. (b) M. Tahir, M. I. Majeed, H. Nawaz, S. Ali, N. Rashid, M. Kashif, I. Ashfaq, W. Ahmad, K. Ghauri, F. Sattar, I. Jawad, M. A. Ghauri, M. A. Anwar, *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2020**, *237*, 118408.
[12] (a) B. B. Zhou, W. H. Ou, J. D. Shen, C. H. Zhao, J. Zhong, P. Du, H. D. Bian, P. Li, L. B. Yang, J. Lu, Y. Y. Li, *ACS Catal.* **2021**, *11*, 14898. (b) B. Zhou, J. Zhong, X. Tang, J.-H. Liu, J. Shen, C. Wang, W. Ou, H. Wang, L. Liu, J. Pan, J. Lu, Y. Y. Li, *J. Catal.* **2022**, *413*, 527.
[13] (a) B. Zhou, W. Ou, C. Zhao, J. Shen, G. Zhang, X. Tang, Z. Deng, G. Zhu, Y. Y. Li, J. Lu, *Chem. Eng. J.* **2021**, *426*, 130733. (b) L. Jiang, M. M. Hassan, S. Ali, H. Li, R. Sheng, Q. Chen, *Trends Food Sci. Technol.* **2021**, *112*, 225.
[14] B. Ryall, G. Eydallin, T. Ferenci, *Microbial. Mol. Biol. Rev.* **2012**, *76*, 597.
[15] M.-Y. Wu, W.-W. Li, G. Christie, P. Setlow, Y.-Q. Li, *Anal. Chem.* **2021**, *93*, 1443.
[16] A. Ditta, H. Nawaz, T. Mahmood, M. I. Majeed, M. Tahir, N. Rashid, M. Muddassar, A. A. Al-Saadi, H. J. Byrne, *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2019**, *221*, 117173.
[17] H. Ali, R. Ullah, S. Khan, M. Bilal, *Vib. Spectrosc.* **2019**, *102*, 112.
[18] T. Sanada, N. Yoshida, K. Kimura, H. Tsuboi, *Biol. Pharm. Bull.* **2021**, *44*, 691.
[19] X. Chen, X. Wu, C. Chen, C. Luo, Y. Shi, Z. Li, X. Lv, C. Chen, J. Su, L. Wu, *Sci. Rep.* **2023**, *13*, 5137.
[20] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, L. Farhan, *J. Big Data* **2021**, *8*, 53.
[21] R. Yamashita, M. Nishio, R. K. G. Do, K. Togashi, *Insights Imaging* **2018**, *9*, 611.
[22] X. Zhang, J. Xu, J. Yang, L. Chen, H. Zhou, X. Liu, H. Li, T. Lin, Y. Ying, *Anal. Chim. Acta* **2020**, *1119*, 41.
[23] (a) Y. Liu, J. Xu, Y. Tao, T. Fang, W. Du, A. Ye, *Analyst* **2020**, *145*, 3297. (b) F. Du, L. He, X. Lu, Y.-Q. Li, Y. Yuan, *Spectrochim. Acta, Part A* **2023**, *289*, 122216. (c) L. Huang, H. Sun, L. Sun, K. Shi, Y. Chen, X. Ren, Y. Ge, D. Jiang, X. Liu, W. Knoll, Q. Zhang, Y. Wang, *Nat Commun* **2023**, *14*, 48.
[24] (a) S.-S. Huang, D. Chen, P. L. Pelczar, V. R. Vepachedu, P. Setlow, Y.-Q. Li, *J. Bacteriol.* **2007**, *189*, 4681. (b) G. Wang, D. Paredes-Sabja, M. Sarker, C. Green, P. Setlow, Y. Q. Li, *J. Appl. Microbiol.* **2012**, *113*, 824. (c) V. S. Novikov, V. V. Kuzmin, S. M. Kuznetsov, M. E. Darvin, J. Lademann, E. A. Sagitova, L. Y. Ustynyuk, K. A. Prokhorov, G. Y. Nikolaeva, *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2021**, *255*, 119668. d Y. Briers, T. Staubli, M. C. Schmid, M. Wagner, M. Schuppler, M. J. Loessner, *PLoS One* **2012**, *7*, e38514.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: J. Hu, L. He, G. Wang, L. Liu, Y. Wang, J. Song, J. Qu, X. Peng, Y. Yuan, *J. Biophotonics* **2024**, e202300510. https://doi.org/10.1002/jbio.202300510