ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Domain-Level Classification of Archaea and Bacteria Using AI-Assisted Single-Cell Raman Spectroscopy

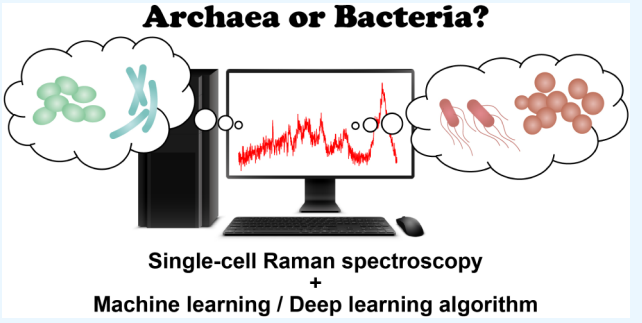
Nanako Kanno, Tatsuya Ohtani, Nodoka Oda, Shingo Kato, Moriya Ohkuma, and Shinsuke Shigeto*

Cite This: *ACS Omega* 2026, 11, 16913−16921

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Archaea and Bacteria are two fundamentally distinct domains of life that share prokaryotic traits, yet differ markedly in molecular and cellular architecture. While many archaeal species identified thus far have been found in extreme environments, recent metagenomic studies have revealed their widespread presence in moderate habitats, including soils, oceans, and even the human microbiome. However, archaea remain less well characterized than bacteria, largely due to the technical challenges associated with culturing and identifying these microorganisms. In this study, we present a culture-independent method for discriminating archaea from bacteria at the single-cell level using Raman spectroscopy combined with machine learning. We constructed a Raman spectral data set comprising 22 prokaryotic species (11 archaea and 11 bacteria) and developed a domain-level Archaea−Bacteria (AB) classifier using the LightGBM tree-based machine learning algorithm. Our AB classification model achieved an average classification accuracy of 89.1% and a sensitivity of 98.1% on eight representative species (including two independent held-out test species) with minimal data size and preprocessing. We also compared its performance to convolutional neural networks with transfer learning, a widely used deep learning approach. Our method provides a robust analytical framework for archaeal detection and represents a valuable addition to the microbiological toolkit, particularly for studying unculturable or low-abundance archaeal populations in complex microbial communities.



INTRODUCTION

As prokaryotes, archaea and bacteria share several features, such as small cell size (typically ∼0.5 to a few micrometers) and common shapes (spherical or rod-shaped), but they differ substantially at the molecular and genetic levels. One of the most notable differences in cellular architecture is the composition of their membranes: archaeal membranes consist of isoprenoid lipids that are ether-linked to *sn*-glycerol-1-phosphate, whereas bacterial membranes are composed of fatty acids linked to *sn*-glycerol-3-phosphate via ester bonds.[1,2] Many archaeal species inhabit extreme environments that are inhospitable to most other organisms, such as those with a high temperature, low pH, high salinity, or high pressure. In addition, archaea possess a variety of unique metabolic pathways that are uncommon in bacteria. For example, a group of archaea known as methanogens can produce methane from hydrogen and carbon dioxide (or other substrates such as acetate) under anaerobic conditions.[3]

Recent advances in genome sequencing have revealed that archaea are not confined to extreme environments but are widespread in more moderate habitats, including soils, oceans, and groundwater.[4,5] They have also been detected in the human body, such as in the gut, oral cavity, and even on the skin.[6,7] These findings highlight the ubiquity and diversity of archaea in Earth's ecosystems. Nevertheless, because of the considerable challenges associated with culturing archaea, much of the archaeal domain remains largely unexplored, representing a large body of the so-called "microbial dark matter".[8−10] Culturing archaea is often challenging due to their strict and diverse growth requirements, such as extremely high temperature, salinity, pH, or anaerobic conditions as well as their generally slow growth rates. As a result, the majority of archaeal species remain difficult or impossible to cultivate using conventional laboratory techniques (e.g., agar plate culture). Therefore, the development of a culture-independent and facile method for distinguishing archaea from bacteria at the single-cell level would significantly accelerate archaea research, which has fallen behind that of bacteria. Such a method would also advance our holistic understanding of archaeal biology.

Here, we present an Archaea−Bacteria (AB) classification model based on single-cell Raman spectroscopy and machine learning. There has been growing interest in applying artificial intelligence (AI)-assisted Raman spectroscopy to identify

Received: January 24, 2026
Revised: February 19, 2026
Accepted: February 26, 2026
Published: March 3, 2026

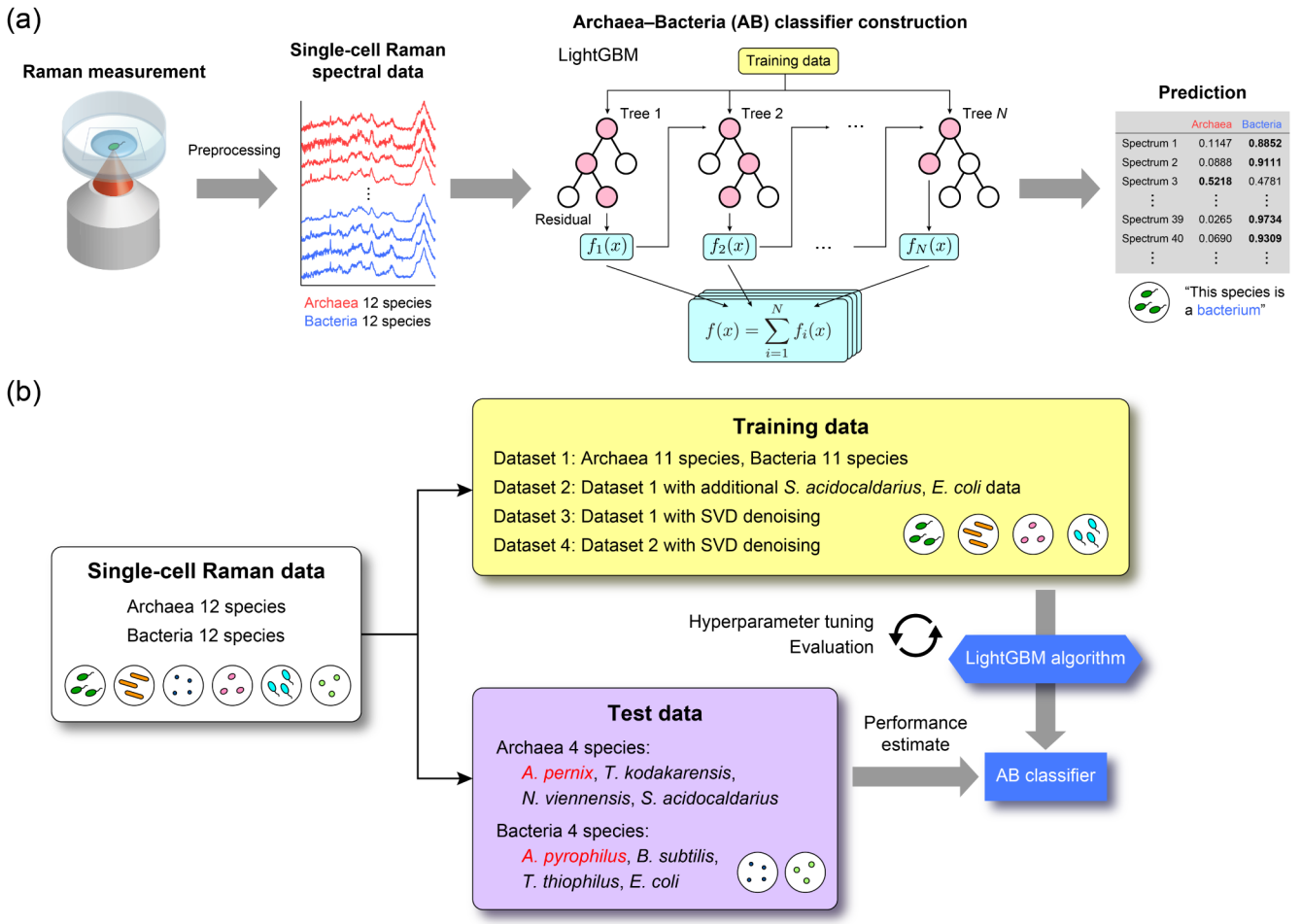


Figure 1. (a) Workflow of our AB classification using single-cell Raman spectral data combined with either machine learning (LightGBM) or deep learning (CNN). (b) Single-cell Raman data sets used for model training, validation, and test. We compared the performance of AB classifiers trained on four different data sets (1−4). Raman spectra of archaeal species *A. pernix* and bacterial species *A. pyrophilus* were used as blind test data. In addition, the spectra of three archaeal species (*T. kodakarensis*, *N. viennensis*, and *S. acidocaldarius*) and three bacterial species (*B. subtilis*, *T. thiophilus*, and *E. coli*), which are all different from those used for training, were used as test data. Although the workflows for LightGBM are illustrated in both parts (a) and (b) as examples, the same procedures were applied when using CNN (VGG16).

microorganisms.[11−13] Unlike sequencing-based methods, this approach is nondestructive, allowing for subsequent attempts of cell culturing and/or downstream analysis of identified cells. We previously reported high-accuracy (>98%) species-level classification of six prokaryotes (three archaeal and three bacterial species from different phyla) using a combination of random forest (RF) and single-cell Raman spectroscopy.[14] In the present work, we shift our focus from species-level classification to domain-level, binary (Archaea vs Bacteria) classification. Although species-level classification inevitably faces challenges due to the immense diversity of microorganisms, domain-level classification avoids such complexity and holds promise for real-world applications. These include the targeted detection of archaea in environmental samples, which are typically less abundant and more difficult to culture than bacteria.

We constructed a single-cell Raman spectral data set comprising 11 archaeal and 11 bacterial species from diverse prokaryotic lineages and then applied AB classification models using LightGBM[15] to classify eight species (four archaea and four bacteria, including one held-out test species for each group) into the two domains. LightGBM is a gradient boosting decision tree framework that utilizes a leafwise growth strategy and is known for its high accuracy and computational efficiency. It is one of the most widely used decision-tree-based methods for classification. Our LightGBM-based AB classifier achieved an average accuracy of 89.1% and a sensitivity of 98.1% for minimally preprocessed training data. By way of comparison, we also evaluated AB classification performance using a convolutional neural network (CNN), a deep learning method frequently adopted in Raman spectroscopic identification of microorganisms[16−19] and with a fundamentally different architecture from LightGBM. The approach presented herein offers a valuable addition to the analytical toolbox for the detection and identification of archaea at the single-cell level.

MATERIALS AND METHODS

Bacterial Strains and Sample Preparation

A total of 12 archaeal and 12 bacterial species were used in this study (see Table S1 for a complete list of the microorganisms used). All strains except for *Paracoccus denitrificans* Pd1222, which was a gift from Dr. Nobuhiko Nomura at the University of Tsukuba, were obtained from the Japan Collection of Microorganisms (JCM). *Bacillus subtilis*, *Escherichia coli*, *Herbaspirillum seropedicae*, *P. denitrificans*, and *Pseudomonas putida* were grown as in our previous studies; detailed protocols are given in ref 14 for the former two species and in ref 20 for the latter three. *Azotobacter chroococcum* and *Bradyrhizobium japonicum*

https://doi.org/10.1021/acsomega.6c00460
ACS Omega 2026, 11, 16913−16921

were cultured in BBL trypticase soy broth and in yeast mannitol broth containing 10.0 g/L mannitol, 1.0 g/L yeast extract, 0.5 g/L K_2HPO_4, 0.2 g/L $MgSO_4{\cdot}7H_2O$, and 0.1 g/L NaCl, respectively, at 200 rpm and 30 °C for 3 days. All of these strains were cultured prior to their use for Raman measurements. Cell cultures other than the seven strains listed above were used as received within 4 d after transfer from JCM.

Cells of each strain were harvested and washed three times with phosphate buffer solution (PBS) by centrifugation (8000−10000 × *g*, 30−60 s, room temperature, or 10000−12000 × *g*, 1−10 min, 4 °C for strains with a low cell density or a small cell size), followed by resuspension in PBS. PBS at pH 7.4 was used, but for marine (*Aeropyrum pernix*, *Pyrococcus furiosus*, *Nitrosopumilus piranensis*, and *Aquifex pyrophilus*) and acidophilic (*Thermoplasma acidophilum* and *Pyrobaculum islandicum*) strains, PBS with 3% NaCl and at pH 5.1 was used instead, respectively. The cell suspension was either diluted or concentrated for better microscopic observation of individual cells, and its 200 μL portion was transferred to a glass bottom dish for subsequent Raman measurements. Note that neither surface-enhanced Raman scattering substrate[19,21,22] nor sample drying was employed in this work.

Single-Cell Raman Spectroscopy

A laboratory-built confocal Raman microspectrometer, which has been described in detail previously,[23,24] was used to acquire Raman spectra from single cells of each strain (Figure 1a). Briefly, the 632.8 nm output of a He−Ne laser was used as the excitation light. The laser beam was magnified by a factor of ∼3 and subsequently introduced into an inverted Nikon microscope. The beam was focused onto the sample with an oil-immersion phase-contrast objective (100×, NA = 1.3), and backward Raman-scattered light was collected with the same objective. The laser power at the sample point was adjusted to 3 mW. The Raman excitation laser beam (spot size ∼1 μm) also served as laser tweezers, by which a single microbial cell in PBS was optically trapped and confined to the focal volume through photon pressure.[25] After confocal detection was achieved by means of a 100-μm pinhole, the Raman-scattered light was analyzed with a spectrograph equipped with a charge-coupled device (CCD) detector with 200 × 1600 pixels. The entire spectral window encompassing both fingerprint and CH-stretching regions with a spectral resolution of ∼5 cm^{-1} was recorded with a 30 s exposure time per cell. For each strain, 40 spectra were acquired from 40 different cells. For *Sulfolobus acidocaldarius* and *E. coli*, additional 80 spectra were measured from two biological replicates (40 spectra each) to enhance training data as well as to keep as independent test data (see below). Similarly, an additional 40 spectra were measured for *Thermococcus kodakarensis*, *Nitrososphaera viennensis*, *B. subtilis*, and *Thermodesulfovibrio thiophilus* from a biological replicate and used as independent test data. All Raman measurements were performed at room temperature.

Data Preprocessing

The recorded Raman spectra were subjected to the following spectral preprocessing that has been modified from what we developed in our previous work.[24] First, the removal of cosmic rays and subtraction of the PBS spectrum (average of 10 spectra) were performed. Because the wavenumber corresponding to each pixel varies slightly from day to day, the horizontal axes of Raman spectra measured on different days were aligned by reference to the emission lines of a standard neon lamp (see ref 24 for more details). After deleting the silent region (1802−2774 cm^{-1}), where no vibrational fundamentals are observed, the resulting fingerprint (671−1800 cm^{-1}) and CH-stretching (2775−3018 cm^{-1}) regions were individually smoothed using a 15-point Savitzky−Golay polynomial filter of degree 2. Subsequently, a linear baseline was determined for the smoothed spectrum and subtracted from both unsmoothed and smoothed spectra. The norm of the baseline-subtracted smoothed spectrum was calculated to vector-normalize the baseline-subtracted unsmoothed spectrum. The smoothing process was critical for an accurate estimation of the baseline and the norm of noisy single-cell Raman spectra (particularly those of some archaeal species). Finally, the normalized, unsmoothed spectra in the fingerprint and CH-stretching regions were concatenated, yielding a preprocessed single-cell Raman spectrum with 892 pixels (Figure 1a). Denoising using singular value decomposition (SVD) was further applied, where necessary, to a set of the 40 normalized spectra of each species acquired on the same day. In this SVD-based denoising, 10 SV components were retained. All of the above preprocessing was performed using Igor Pro 9.05 software (WaveMetrics).

Construction of AB Classification Models

Training and Test Data. The preprocessed single-cell Raman spectra acquired from 12 archaeal and 12 bacterial species were divided into training and test data sets. As shown in Figure 1b, we prepared four types of training data in the present study: a data set comprising 880 spectra without SVD denoising of 11 archaeal species and 11 bacterial species (data set 1); data set 1 with an additional 40 spectra for each of *S. acidocaldarius* and *E. coli* (data set 2; 960 spectra in total); data set 1 with SVD denoising (data set 3); and data set 2 with SVD denoising (data set 4). The spectra (40 spectra per species) of four archaeal (*A. pernix*, *T. kodakarensis*, *N. viennensis*, and *S. acidocaldarius*) and four bacterial (*A. pyrophilus*, *B. subtilis*, *T. thiophilus*, and *E. coli*) species were held out as test data. *A. pernix* and *A. pyrophilus* were kept aside from training, while the other six species were included in the training and validation of the AB classification model (Figure 1b). However, their test spectra were different from those used in training and validation. The SVD-denoised spectra were used as test data for data sets 3 and 4.

LightGBM Classification Models. The LightGBM algorithm was used to construct the AB binary classifiers. Hyperparameters, such as the maximum depth of each decision tree and the bagging fraction, were optimized using Optuna,[26] an automated hyperparameter optimization framework that relies on Bayesian optimization. To this end, the training data were split into training, validation, and Optuna evaluation sets in a 64:16:20 ratio. The validation fold (16%) was used for early stopping during the training of the AB classifier. Early stopping is a technique that halts training when the model's performance on the validation set ceases to improve. Log-loss values were calculated on the Optuna evaluation set (20%) to assess the model's performance under each hyperparameter configuration. Each Optuna study was set to run for 25 trials. The hyperparameters that were fixed or tuned in this study are summarized in Table S2. Using the optimized hyperparameters, an AB classification model was then trained on the full training set (880 or 960 spectra) and applied to classify 40 spectra of four archaeal and four bacterial species in the test set. It generated an output of the probabilities for the two classes (Archaea and Bacteria), with the class assigned based on the higher probability (Figure 1a). Model training was performed on a computer running Windows 11 with an Intel Core i5-14400F CPU and 16-GB RAM, using the LightGBM (ver. 4.3.0) and Optuna (ver. 3.6.0) packages in Python (ver. 3.11.7).

CNN (VGG16) Classification Models. To compare the performance of AB classification with that of the tree-based LightGBM, convolutional neural network (CNN) models were also adopted. CNN is one of the most commonly employed deep learning algorithms for Raman spectroscopic identification of biological samples.[16−19] In this study, we applied transfer learning using VGG16,[27] a representative CNN architecture pretrained on 1.2 million images from the ImageNet data set. For transfer learning, the pretrained VGG16 model without its fully connected layers was coupled to new fully connected layers trained on the single-cell Raman data. These new layers consisted of a flattened layer, a dense layer with 1032 units and ReLU activation, a dropout layer (rate = 0.3), and a final dense layer with 2 units and softmax activation. The model was trained using the Adam optimizer with a learning rate of 0.0001, for 10 epochs, and a batch size of 32. All single-cell Raman data were converted into 224 × 224 images for training, validation, and test. More specifically, each spectrum was displayed as a graph and saved in JPEG format on the Igor Pro software. The data set was split into training and validation sets in a 9:1 ratio. As with the LightGBM model, the CNN model was evaluated on independent test data of the four archaeal and four bacterial species. To implement VGG16, TensorFlow-GPU (ver. 2.6.0) and Keras (ver. 2.6.0) were used in Python (ver. 3.7.16). Model training was conducted on the same computer as described above with an NVIDIA GeForce RTX 4060 GPU (8 GB).

https://doi.org/10.1021/acsomega.6c00460
ACS Omega 2026, 11, 16913−16921

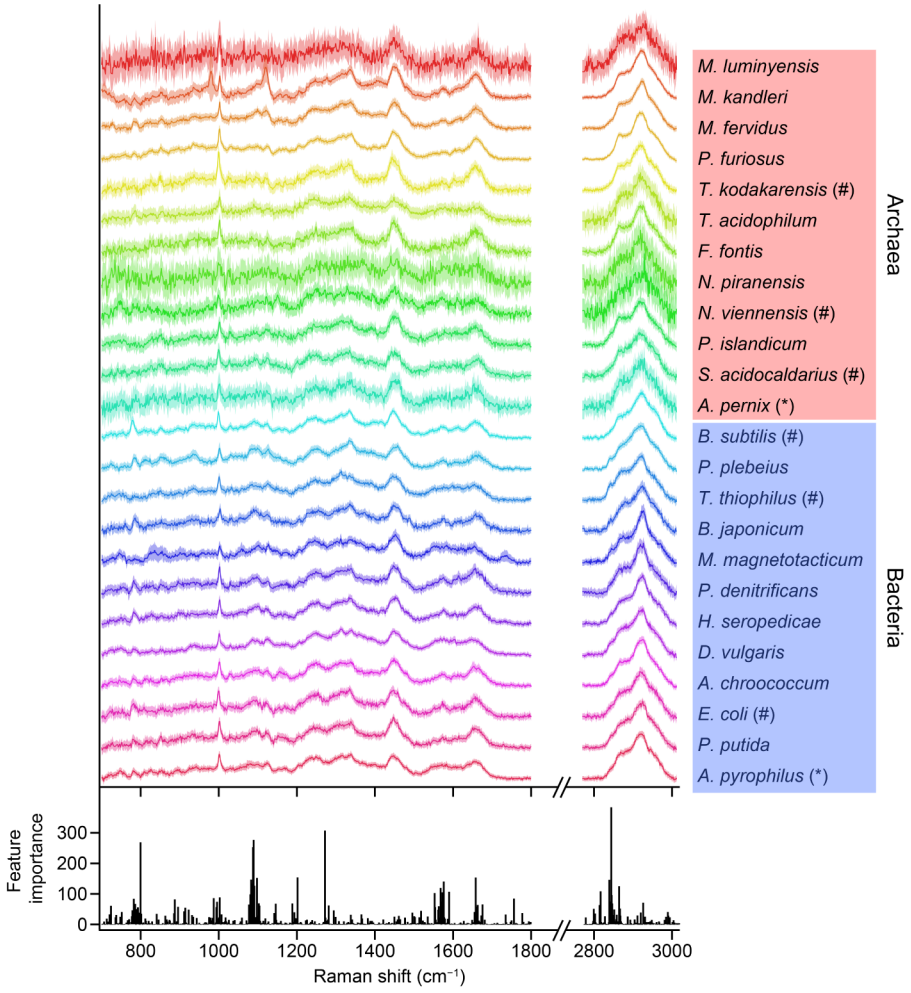


Figure 2. Average preprocessed Raman spectra (without SVD) for the 12 archaeal and 12 bacterial species (upper colored traces; $n = 40$ per species), along with a plot of feature importance scores (lower black trace). The shaded area represents the $\pm 1\sigma$ error envelope. Hash marks (#) and asterisks (*) indicate microbial species whose spectra were used for both training and test and held out as blind test data, respectively.

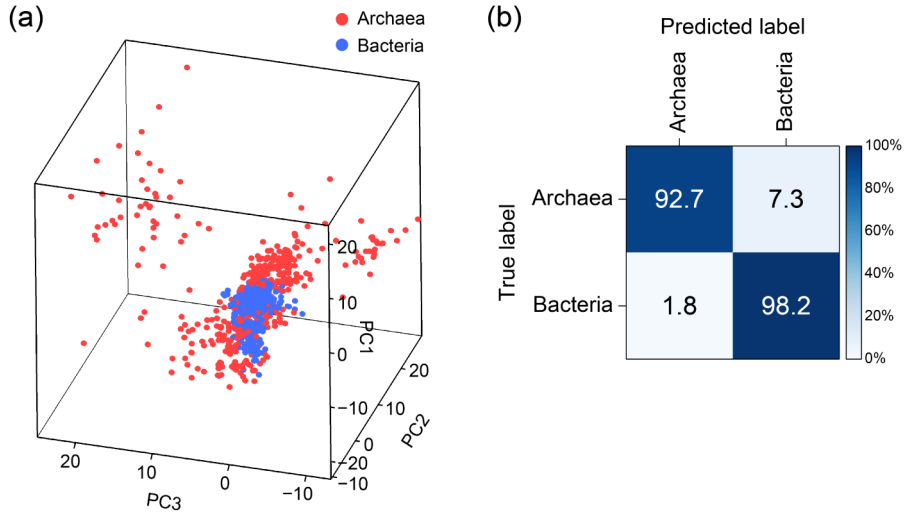


Figure 3. (a) Scatter plot of PC1, PC2, and PC3 scores obtained from the PCA of training data set 1. Red and blue dots represent individual spectra of archaeal and bacterial cells, respectively. (b) Confusion matrix showing the result of 10-fold cross-validation within training data set 1.

https://doi.org/10.1021/acsomega.6c00460
ACS Omega 2026, 11, 16913−16921

RESULTS

Single-Cell Raman Spectra of Archaeal and Bacterial Species

Figure 2 displays the average preprocessed Raman spectra (40 spectra per species) excited at 632.8 nm for the 12 archaeal and 12 bacterial species. The spectra of *T. kodakarensis*, *N. viennensis*, *S. acidocaldarius*, *B. subtilis*, *P. denitrificans*, *H. seropedicae*, *E. coli*, and *P. putida* were presented and discussed in our previous studies,[14,20] while the others are newly reported in this study. Overall, these spectra appear very similar and are dominated by Raman bands of DNA/RNA (~780 and ~1575 cm^{-1}), proteins (~1003, ~1250, ~1450, and ~1660 cm^{-1}), and lipids (~2870 cm^{-1}).[14,28,29] The assignments of these Raman bands are given in Table S3. Notably, *Methanopyrus kandleri*, a deep-sea-dwelling hyperthermophile, exhibits an intense Raman band at ~980 cm^{-1}, which is most likely attributable to the totally symmetric stretching mode of the SO_4^{2-} ion. This observation is consistent with the 532 nm-excited Raman spectrum of *M. kandleri* cells that we reported previously,[30] although the origin of the SO_4^{2-} ion remains unclear. Another feature evident from Figure 2 is the considerable variation in the noise levels and cell-to-cell spectral variability among species. This variation is particularly pronounced in archaeal species such as *Methanomassiliicoccus luminyensis*, *N. piranensis*, and *N. viennensis* because of their significantly smaller cell sizes compared to bacteria.

To visualize similarities among the single-cell Raman spectra, we performed principal component analysis (PCA) on training data set 1, which comprised 40 Raman spectra from each of the 11 archaeal and 11 bacterial species. As shown in the three-dimensional scatter plots of PC1, PC2, and PC3 (Figure 3a), archaeal and bacterial spectra did not form well-separated clusters, with archaeal spectra exhibiting a more dispersed distribution. This result suggests that, although the classification task of our interest is binary, it may become increasingly complex as the number of microbial species to be classified increases.

The performance of our LightGBM classification model, evaluated using 10-fold cross-validation on training data set 1, is presented as a confusion matrix in Figure 3b. Our AB classifier achieved a high accuracy of 95.5 ± 2.2% (± indicates the standard deviation across 10 train/test splits). As in our previous study using random forest,[14] this 10-fold cross-validation also yielded an importance score for each feature (i.e., Raman shift), which is plotted at the bottom of Figure 2. Most of the highly important features are localized in spectral regions near the prominent Raman bands described above, including those of DNA/RNA at ~780, ~1100, and ~1575 cm^{-1}, proteins at ~1660 cm^{-1}, and lipids at ~2850 cm^{-1}. This result ensures that our classification model predominantly relies on the Raman spectroscopic signatures of biomolecules present in archaeal and bacterial cells.

LightGBM Model Achieves Highly Accurate AB Classification

Next, we evaluated the performance of the LightGBM classification model on held-out test data consisting of the Raman spectra of *A. pernix*, *A. pyrophilus*, and six more species (three archaeal and three bacterial) selected from the training data sets. Note that these test data were completely independent of the training data. *A. pernix* and *A. pyrophilus* were chosen for testing due to their similar characteristics, as both are marine hyperthermophiles.[31,32] The model trained on training data set 1 achieved excellent overall performance, with an average classification accuracy of 89.1% across the eight species (dark green, Figure 4a). Among the archaeal species, the lowest accuracy was observed for *S. acidocaldarius* (95%), whereas among the bacterial species, *E. coli* exhibited the lowest accuracy at 27.5%. We attribute the exceptionally poor performance for *E. coli* to higher noise levels in its Raman spectra compared to other bacteria, possibly due to its smaller cell size in the stationary phase,[33−35] when the spectra were acquired.

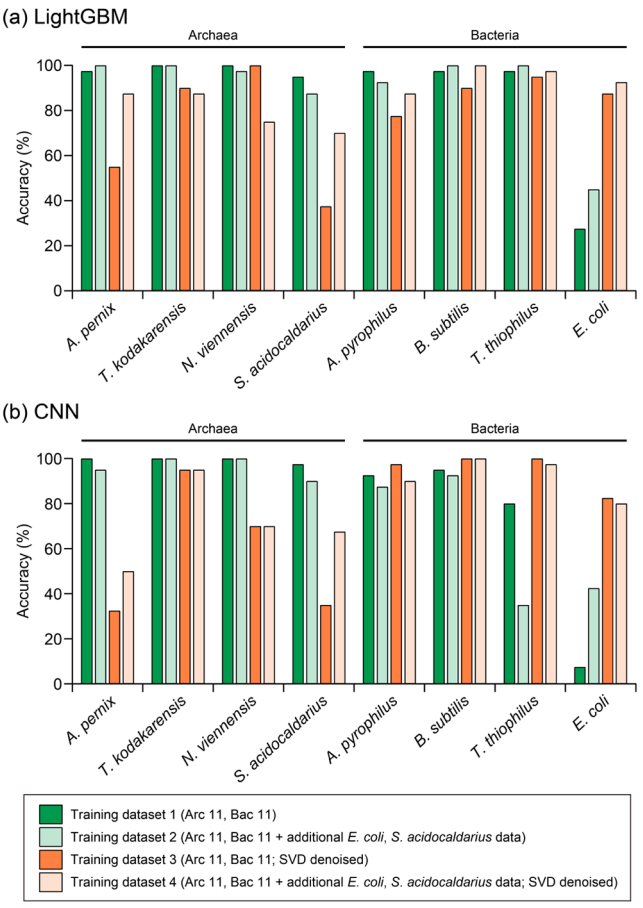


Figure 4. AB Classification accuracy obtained with LightGBM (a) and CNN (VGG16) (b) models for four archaeal species (*A. pernix*, *T. kodakarensis*, *N. viennensis*, and *S. acidocaldarius*) and four bacterial species (*A. pyrophilus*, *B. subtilis*, *T. thiophilus*, and *E. coli*). Of the eight species, *A. pernix* and *A. pyrophilus* are those that were not included in training data sets (indicated by asterisks in Figure 2). The other six species (designated by hash marks in Figure 2) were selected from those within training data sets. Results from four different training data sets 1−4 are compared: 1, dark green; 2, light green; 3, dark orange; 4, light orange.

To examine whether the classification accuracy for *E. coli* could be improved by increasing the training data, we added 40 additional real spectra each for *E. coli* and *S. acidocaldarius* to training data set 1, creating training data set 2. We emphasize that the added data are experimentally measured spectra and not simulated spectra as used in recent data augmentation strategies.[36−39] This data addition resulted in a modest improvement in the classification accuracy for *E. coli* (from 27.5% to 45%), with only a small increase in the overall average accuracy to 90.3% (light green, Figure 4a; see also Table 1).

More pronounced changes were observed when noise reduction using SVD (see Materials and Methods) was applied to both the training and test sets during the preprocessing of

https://doi.org/10.1021/acsomega.6c00460
ACS Omega 2026, 11, 16913−16921

Table 1. Comparison of the AB Classification Accuracy, Sensitivity, and Specificity for LightGBM and CNN Models Trained on Different Datasets

	LightGBM			CNN (VGG16)		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Data set 1	89.1	98.1	80.0	84.1	99.4	68.8
Data set 2	90.3	96.3	84.4	80.3	96.3	64.4
Data set 3	79.1	70.6	87.5	76.6	58.1	95.0
Data set 4	87.2	80.0	94.4	81.3	70.6	91.9

single-cell Raman spectra (training data set 3). The AB classification model trained on this data set achieved a significantly improved accuracy of 87.5% for *E. coli* (dark orange, Figure 4a). However, this improvement came at the cost of reduced accuracy for *A. pernix* (from 97.5% to 55%) and *S. acidocaldarius* (from 95% to 37.5%). This result suggests that for these two archaeal species the noise present in the original Raman spectra may have played a role in their classification.

The drop in classification accuracy for *A. pernix* and *S. acidocaldarius* observed with training data set 3 was largely mitigated by using the model trained on training data set 4, which corresponds to SVD-denoised data set 2. The accuracies for the two species recovered to 87.5% and 70%, respectively. However, because there were decent decreases in accuracy for other species, the overall average accuracy was 87.2%, slightly lower than that achieved with the model trained on data set 1 (Table 1).

The overall performance of the AB binary classifier trained on the four data sets is presented as confusion matrices in Figure 5a–d. For the model trained on data set 1, the sensitivity (i.e., recall for the Archaea class) and specificity (i.e., recall for the Bacteria class) were calculated to be 98.1% and 80.0%, respectively. The sensitivities and specificities of the other models trained on training data sets 2–4 are summarized in Table 1. High sensitivity is particularly desirable when applying this approach to environmental samples because misclassifying rare archaeal cells as abundant bacterial cells is more critical than the reverse case. The performance of the AB classifier is further demonstrated by the receiver operating characteristic (ROC) curves shown in Figure 5e. The area under the curve (AUC) exceeds 0.9 in all cases, indicating the high predictive capability of the LightGBM model.

Comparison with CNN

To check the robustness of our AI-assisted Raman spectroscopic classification of archaea and bacteria, we compared the AB classification performance of the deep learning approach CNN (VGG16) with that of the LightGBM model. CNN has a fundamentally different architecture from the relatively simple, tree-type LightGBM. Furthermore, while most previous studies have used one-dimensional Raman spectra as input for CNN model construction,[16−19] we employed two-dimensional images[40] (224 × 224 pixels; see Figure S1 for example). The classification accuracies for the same archaeal and bacterial species used in the LightGBM model are shown in Figure 4b for training data sets 1–4. The CNN model trained on training data set 1 achieved an average accuracy of 84.1% across the eight species, which is comparable to that of the LightGBM model. The average accuracies for models trained on training data sets 2–4 were 80.3%, 76.6%, and 81.3%, respectively (Table 1). Although there are some species-dependent differences in the variation of the classification accuracy across the four training data sets, the overall performance was found to be similar between the CNN and LightGBM models. The slightly lower accuracies of the CNN model may arise from the small data size used in this study for deep learning models. These results suggest that AB classification using single-cell Raman spectra is generally robust, regardless of whether it is based on decision trees (LightGBM) or neural networks (CNN) with 2D image data.

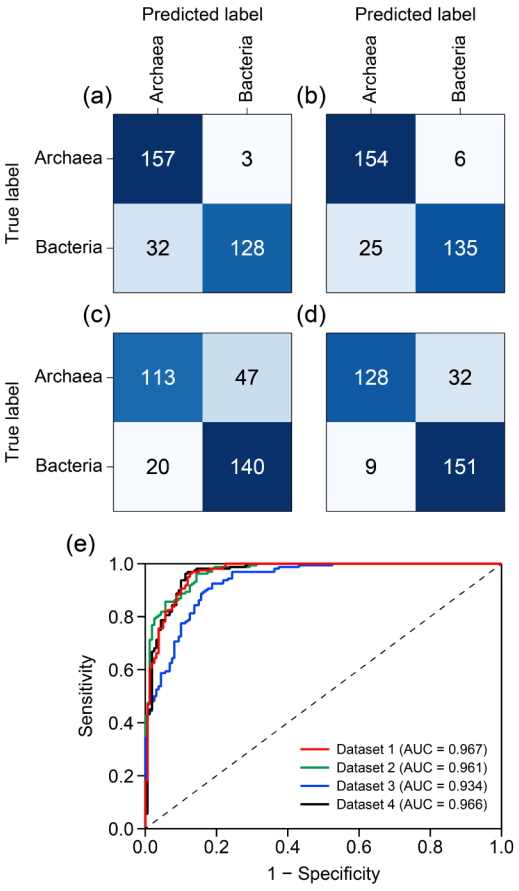


Figure 5. (a–d) Confusion matrices showing the AB classification results obtained using the LightGBM model trained on data set 1 (a), 2 (b), 3 (c), and 4 (d). The numbers in each matrix represent the number of spectra (40 spectra/species × 4 species = 160 spectra per class). (e) Corresponding ROC curves. The dashed line denotes the performance of random guessing (AUC = 0.5).

DISCUSSION

In this study, we developed and evaluated a single-cell AB classifier based on Raman spectroscopy and AI technologies. The classifier, which was trained on Raman spectra obtained from 11 archaeal and 11 bacterial species across diverse phylogenetic lineages, achieved an average classification accuracy of 89.1% and a sensitivity of 98.1% with LightGBM.

https://doi.org/10.1021/acsomega.6c00460
ACS Omega 2026, 11, 16913−16921

This performance was further benchmarked against a CNN model. These results underscore the feasibility of classifying single microbial cells at the domain level without labor-intensive cultivation and sequencing techniques.

The robust performance of the LightGBM classification model highlights the potential of tree-based machine learning models for microbial classification tasks. LightGBM required minimal preprocessing yet achieved comparable or superior performance to CNN, suggesting that classical machine learning algorithms remain competitive in such a binary classification. Furthermore, LightGBM's interpretability and faster training time make it particularly suitable for field-deployable or real-time applications.

It is important to note that our classifier focuses on domain-level identification rather than species-level identification. Some may view this as a step backward, given recent advances in AI-assisted Raman spectroscopy for microbial classification.[17−19] However, species-level classifiers trained on known strains may not generalize well to environmental microbial samples, which often contain rare, uncultured, or previously unknown species. This makes domain-level classification both practically important and technically challenging. While the binary classification simplifies the taxonomic complexity of prokaryotes, it also introduces difficulties due to substantial intradomain spectral variability. Nevertheless, domain-level classification is valuable in ecological and medical contexts, where distinguishing between Archaea and Bacteria is sufficient for subsequent analysis or intervention. For example, rapid detection of methanogenic archaea in anaerobic digesters[41] or human microbiomes[42] can inform diagnostic and therapeutic strategies.

One major advantage of Raman-based identification is that due to its nondestructive nature it allows for subsequent analysis of the same cell, such as cultivation attempts, genome sequencing, and metabolic assays. This enables the integration of phenotypic and genotypic information at the single-cell level, an important step toward a comprehensive understanding of microbial physiology and ecology. Raman-activated cell sorting systems[43,44] could be coupled with our classifier to physically isolate archaeal cells from mixed communities, facilitating the study of uncultured taxa.

Despite these advantages, this study has several limitations. First, although our training data set includes a taxonomically diverse panel of species, it represents only a small subset of the vast prokaryotic diversity found in nature. Expanding the data set to include more species, especially from poorly characterized archaeal lineages, is crucial for improving classification performance. Second, the Raman spectral profiles can be profoundly affected by growth conditions (as investigated in our previous study[20]), sample preparation methods, and instrumentation, which may affect reproducibility across laboratories. Thus, the standardization of protocols is required for future work.

Furthermore, we acknowledge that the present data set has a relatively low signal-to-noise ratio for several species, particularly small microbial cells such as *E. coli* and some of the archaeal strains analyzed. In Raman measurements of these cells, the spectral intensity is inherently limited by cell size. Increasing the signal to a level where noise becomes negligible would require prolonged acquisition times or higher laser power, which are incompatible with the practical objective of measuring cells in a near-native and minimally perturbed state. As a result, the measured spectra comprise varying proportions of biological signal and noise with the signal-to-noise ratio differing among species. The present study therefore does not assume uniform or high-quality spectra across all taxa but instead explicitly addresses how effectively archaea and bacteria can be distinguished under these experimentally constrained, real-world conditions. From this perspective, variations in classification performance observed after SVD denoising reflect the intrinsic difficulty of Raman-based microbial discrimination at low signal levels rather than indicating that the underlying biological signals are spurious.

In conclusion, our study demonstrates the feasibility and utility of machine-learning-assisted Raman spectroscopy for Archaea−Bacteria classification at the single-cell level. By bypassing the need for culturing or molecular probes, this approach offers a rapid, label-free, and minimally invasive tool for identifying archaea among microbial cells. As research continues to reveal the ubiquity and functional importance of archaea in diverse ecosystems, such tools will be essential for accelerating the exploration of archaeal biology and closing the knowledge gap that still exists relative to bacteria.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.6c00460.

List of the archaeal and bacterial strains used in this study (Table S1); hyperparameters tuned or fixed in the construction of an AB classifier using LightGBM (Table S2); assignments of major Raman bands (Table S3); and example images of single-cell Raman spectra used in the construction of an AB classifier using CNN (VGG16) (Figure S1) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Shinsuke Shigeto − *Department of Chemistry, Graduate School of Science and Technology, Kwansei Gakuin University, Sanda, Hyogo 669-1330, Japan;* orcid.org/0000-0002-2035-2068; Email: shigeto@kwansei.ac.jp

Authors

Nanako Kanno − *Department of Chemistry, Graduate School of Science and Technology, Kwansei Gakuin University, Sanda, Hyogo 669-1330, Japan*

Tatsuya Ohtani − *Department of Chemistry, Graduate School of Science and Technology, Kwansei Gakuin University, Sanda, Hyogo 669-1330, Japan*

Nodoka Oda − *Department of Chemistry, Graduate School of Science and Technology, Kwansei Gakuin University, Sanda, Hyogo 669-1330, Japan*

Shingo Kato − *Japan Collection of Microorganisms, RIKEN BioResource Research Center, Tsukuba, Ibaraki 305-0074, Japan*

Moriya Ohkuma − *Japan Collection of Microorganisms, RIKEN BioResource Research Center, Tsukuba, Ibaraki 305-0074, Japan*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.6c00460

Notes

The authors declare no competing financial interest.

https://doi.org/10.1021/acsomega.6c00460
ACS Omega 2026, 11, 16913−16921

■ ACKNOWLEDGMENTS

This work was supported by the Japan Society for the Promotion of Science KAKENHI Grant Numbers JP19H05681 (S.S.) and in part by the Morino Foundation for Molecular Science (S.S.).

■ REFERENCES

(1) Caforio, A.; Driessen, A. J. M. Archaeal phospholipids: Structural properties and biosynthesis. *Biochim. Biophys. Acta - Mol. Cell Biol. Lipids* **2017**, *1862*, 1325−1339.
(2) Harayama, T.; Riezman, H. Understanding the diversity of membrane lipid composition. *Nat. Rev. Mol. Cell Biol.* **2018**, *19*, 281−296.
(3) Methanogenesis: Ecology, Physiology, Biochemistry & Genetics. *Chapman & Hall Microbiology Series*. Ferry, J. G., Eds.; Springer: New York, 1993.
(4) De Long, E. F. Everything in moderation: Archaea as 'non-extremophiles'. *Curr. Opin. Genet. Dev.* **1998**, *8*, 649−654.
(5) Schleper, C.; Jurgens, G.; Jonuscheit, M. Genomic studies of uncultivated archaea. *Nat. Rev. Microbiol.* **2005**, *3*, 479−488.
(6) Probst, A. J.; Auerbach, A. K.; Moissl-Eichinger, C. Archaea on Human Skin. *PLoS One* **2013**, *8*, No. e65388.
(7) Koskinen, K.; Pausan Manuela, R.; Perras Alexandra, K.; Beck, M.; Bang, C.; Mora, M.; Schilhabel, A.; Schmitz, R.; Moissl-Eichinger, C. First Insights into the Diverse Human Archaeome: Specific Detection of Archaea in the Gastrointestinal Tract, Lung, and Nose and on Skin. *mBio* **2017**, *8*, No. e00824−00817.
(8) Solden, L.; Lloyd, K.; Wrighton, K. The bright side of microbial dark matter: Lessons learned from the uncultivated majority. *Curr. Opin. Microbiol.* **2016**, *31*, 217−226.
(9) Rinke, C.; Schwientek, P.; Sczyrba, A.; Ivanova, N. N.; Anderson, I. J.; Cheng, J.-F.; Darling, A.; Malfatti, S.; Swan, B. K.; Gies, E. A.; et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **2013**, *499*, 431−437.
(10) Lewis, W. H.; Tahon, G.; Geesink, P.; Sousa, D. Z.; Ettema, T. J. G. Innovations to culturing the uncultured microbial majority. *Nat. Rev. Microbiol.* **2021**, *19*, 225−240.
(11) Luo, R.; Popp, J.; Bocklitz, T. Deep Learning for Raman Spectroscopy: A Review. *Analytica* **2022**, *3*, 287−301.
(12) 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 Anal. Chem.* **2020**, *124*, 115796.
(13) Qi, Y.; Hu, D.; Jiang, Y.; Wu, Z.; Zheng, M.; Chen, E. X.; Liang, Y.; Sadi, M. A.; Zhang, K.; Chen, Y. P. Recent Progresses in Machine Learning Assisted Raman Spectroscopy. *Adv. Opt. Mater.* **2023**, *11*, 2203104.
(14) Kanno, N.; Kato, S.; Ohkuma, M.; Matsui, M.; Iwasaki, W.; Shigeto, S. Machine learning-assisted single-cell Raman fingerprinting for *in situ* and nondestructive classification of prokaryotes. *iScience* **2021**, *24*, 102975.
(15) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *31st Conference on Neural Information Processing Systems*; Advances in Neural Information Processing Systems, Long Beach, CA, USA, 2017.
(16) Lu, W.; Chen, X.; Wang, L.; Li, H.; Fu, Y. V. Combination of an Artificial Intelligence Approach and Laser Tweezers Raman Spectroscopy for Microbial Identification. *Anal. Chem.* **2020**, *92*, 6288−6296.
(17) Singh, S.; Kumbhar, D.; Reghu, D.; Venugopal, S. J.; Rekha, P. T.; Mohandas, S.; Rao, S.; Rangaiah, A.; Chunchanur, S. K.; Saini, D. K.; et al. Culture-Independent Raman Spectroscopic Identification of Bacterial Pathogens from Clinical Samples Using Deep Transfer Learning. *Anal. Chem.* **2022**, *94*, 14745−14754.
(18) Yu, S.; Li, X.; Lu, W.; Li, H.; Fu, Y. V.; Liu, F. Analysis of Raman Spectra by Using Deep Learning Methods in the Identification of Marine Pathogens. *Anal. Chem.* **2021**, *93*, 11089−11098.
(19) 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*, 4927.
(20) Oda, N.; Kanno, N.; Kato, S.; Ohkuma, M.; Shigeto, S. Improving Machine Learning−Based Bacterial Discrimination by Learning Single-Cell Raman Data From Multiple Growth Phases. *J. Raman Spectrosc* **2025**, *56*, 481−490.
(21) Tseng, Y.-M.; Chen, K.-L.; Chao, P.-H.; Han, Y.-Y.; Huang, N.-T. Deep Learning−Assisted Surface-Enhanced Raman Scattering for Rapid Bacterial Identification. *ACS Appl. Mater. Interfaces* **2023**, *15*, 26398−26406.
(22) Wang, L.; Tang, J.-W.; Li, F.; Usman, M.; Wu, C.-Y.; Liu, Q.-H.; Kang, H.-Q.; Liu, W.; Gu, B. Identification of Bacterial Pathogens at Genus and Species Levels through Combination of Raman Spectrometry and Deep-Learning Algorithms. *Microbiol. Spectr.* **2022**, *10*, No. e02580−02522.
(23) Matsuda, A.; Sakaguchi, N.; Shigeto, S. Can cells maintain their bioactivity in ionic liquids? A novel single-cell assessment by Raman microspectroscopy. *J. Raman Spectrosc.* **2019**, *50*, 768−777.
(24) Kanno, N.; Kato, S.; Ohkuma, M.; Matsui, M.; Iwasaki, W.; Shigeto, S. Nondestructive microbial discrimination using single-cell Raman spectra and random forest machine learning algorithm. *STAR Protoc.* **2022**, *3*, 101812.
(25) Ashkin, A. Acceleration and Trapping of Particles by Radiation Pressure. *Phys. Rev. Lett.* **1970**, *24*, 156−159.
(26) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. In *25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; ACM SIGKDD: Anchorage, AK, USA, 2019, pp. 2623−2631.
(27) Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv*, **2015**.
(28) Carey, P. *Biochemical Applications of Raman and Resonance Raman Spectroscopies*; Academic Press, 1982.
(29) Naumann, D. FT-Infrared and FT-Raman Spectroscopy in Biomedical Research. *Appl. Spectrosc. Rev.* **2001**, *36*, 239−298.
(30) Kanno, N.; Kato, S.; Itoh, T.; Ohkuma, M.; Shigeto, S. Resonance Raman analysis of intracellular vitamin B_{12} analogs in methanogenic archaea. *Anal Sci. Adv.* **2022**, *3*, 165−173.
(31) Sako, Y.; Nomura, N.; Uchida, A.; Ishida, Y.; Morii, H.; Koga, Y.; Hoaki, T.; Maruyama, T. *Aeropyrum pernix* gen. nov., sp. nov., a Novel Aerobic Hyperthermophilic Archaeon Growing at Temperatures up to 100°C. *Int. J. Syst. Evol. Microbiol.* **1996**, *46*, 1070−1077.
(32) Huber, R.; Wilharm, T.; Huber, D.; Trincone, A.; Burggraf, S.; König, H.; Reinhard, R.; Rockinger, I.; Fricke, H.; Stetter, K. O. *Aquifex pyrophilus* gen. nov. sp. nov., Represents a Novel Group of Marine Hyperthermophilic Hydrogen-Oxidizing Bacteria. *Syst. Appl. Microbiol.* **1992**, *15*, 340−351.
(33) Akerlund, T.; Nordström, K.; Bernander, R. Analysis of cell size and DNA content in exponentially growing and stationary-phase batch cultures of *Escherichia coli*. *J. Bacteriol.* **1995**, *177*, 6791−6797.
(34) Makinoshima, H.; Aizawa, S.-I.; Hayashi, H.; Miki, T.; Nishimura, A.; Ishihama, A. Growth Phase-Coupled Alterations in Cell Structure and Function of *Escherichia coli*. *J. Bacteriol.* **2003**, *185*, 1338−1345.
(35) Zhu, Y.; Mustafi, M.; Weisshaar, J. C. Biophysical Properties of *Escherichia coli* Cytoplasm in Stationary Phase by Superresolution Fluorescence Microscopy. *mBio* **2020**, *11*, No. e00143−00120.
(36) Kim, Y.; Lee, W. Distributed Raman Spectrum Data Augmentation System Using Federated Learning with Deep Generative Models. *Sensors* **2022**, *22*, 9900.
(37) Pavlou, E.; Kourkoumelis, N. Deep adversarial data augmentation for biomedical spectroscopy: Application to modelling Raman spectra of bone. *Chemom. Intell. Lab. Syst.* **2022**, *228*, 104634.
(38) Zhang, X.; Li, H.; Tian, X.; Chen, C.; Su, Y.; Li, M.; Lv, J.; Chen, C.; Lv, X. Application of spectral small-sample data combined with a method of spectral data augmentation fusion (SDA-Fusion) in cancer diagnosis. *Chemom. Intell. Lab. Syst.* **2022**, *231*, 104681.
(39) Wu, M.; Wang, S.; Pan, S.; Terentis, A. C.; Strasswimmer, J.; Zhu, X. Deep learning data augmentation for Raman spectroscopy cancer tissue classification. *Sci. Rep.* **2021**, *11*, 23842.

https://doi.org/10.1021/acsomega.6c00460
ACS Omega 2026, 11, 16913−16921

(40) Wattanavichean, N.; Boonchai, J.; Yodthong, S.; Preuksakarn, C.; Huang, C.-H.; Surasak, T. GFP Pattern Recognition in Raman Spectra by Modified VGG Networks for Localisation Tracking in Living Cells. *Eng. J.* **2021**, *25*, 151−160.
(41) Rabii, A.; Aldin, S.; Dahman, Y.; Elbeshbishy, E. A Review on Anaerobic Co-Digestion with a Focus on the Microbial Populations and the Effect of Multi-Stage Digester Configuration. *Energies* **2019**, *12*, 1106.
(42) Chaudhary, P. P.; Conway, P. L.; Schlundt, J. Methanogens in humans: Potentially beneficial or harmful for health. *Appl. Microbiol. Biotechnol.* **2018**, *102*, 3095−3104.
(43) Song, Y.; Yin, H.; Huang, W. E. Raman activated cell sorting. *Curr. Opin. Chem. Biol.* **2016**, *33*, 1−8.
(44) Tang, X.; Wu, Q.; Shang, L.; Liu, K.; Ge, Y.; Liang, P.; Li, B. Raman cell sorting for single-cell research. *Front. Bioeng. Biotechnol.* **2024**, *12*, 1389143.

https://doi.org/10.1021/acsomega.6c00460
ACS Omega 2026, 11, 16913−16921