Journal of Oceanology and Limnology
Vol. 44, No. 2, pp. 728-742, 2026
https://doi.org/10.1007/s00343-025-5087-z

Research Paper

Unsupervised classification of environmental marine microbes using a Raman spectra-based deep learning framework*

Fengyun LÜ[1, 2, #], Yanmei ZHANG[2, 3, #], Hongtao LIANG[1, **], Luyang SUN[2, 3, **]

[1] *School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266100, China*
[2] *Single-Cell Center, CAS Key Laboratory of Biofuels, Shandong Key Laboratory of Energy Genetics, Qingdao New Energy Shandong Laboratory, Shandong Energy Institute, Qingdao Institute of Bioenergy and Bioprocess Technology, Chinese Academy of Sciences, Qingdao 266101, China*
[3] *College of Life Science, University of Chinese Academy of Sciences, Beijing 101408, China*

Received Mar. 24, 2025; accepted in principle Apr. 22, 2025; accepted for publication May 28, 2025

Abstract The development of microfluidic-based Raman devices has enabled the acquisition of thousands of single-cell Raman spectra from complex microbial populations, offering new avenues for characterizing environmental bacteria. We proposed and validated the convolutional deep embedded clustering (CDEC) model, a clustering framework designed for the unsupervised classification of marine microbial Raman spectra. A four-stage analytical framework was employed to systematically assess the performance of the CDEC model, with increasing dataset complexity ranging from mock communities of pure bacterial cultures to natural microbial populations isolated from seawater. The CDEC algorithm consistently distinguished bacterial species across all stages, achieving an average accuracy of 97.27%, and surpassed the baseline methods in performance. In natural microbial communities, the model resolved eight distinct clusters with unique Raman spectral markers, revealing underlying metabolic heterogeneity. These results highlight the scalability and utility of the CDEC model for studying marine microbial diversity and characterizing microbial composition and metabolic traits at single-cell resolution.

Keyword: marine microbe; Raman spectroscopy; deep clustering; convolutional network

1 INTRODUCTION

Marine microorganisms play a fundamental role in oceanic ecosystems, influencing global biogeochemical cycles, symbiotic relationships, and even human health. These microbes drive critical processes such as carbon fixation, nitrogen cycling, and organic matter decomposition (Falkowski et al., 2008). Despite their ecological and biotechnological importance, their classification remains a significant challenge due to their vast diversity and the limitations of conventional taxonomy (Taş et al., 2021).

Classification of marine microbes has traditionally relied on 16S rRNA gene sequencing, a widely used molecular approach that provides high-resolution phylogenetic insights into microbial community composition (Langille et al., 2013). However, its effectiveness is limited by low taxonomic resolution and an inability to capture functional traits, restricting identification to the species or strain level. Moreover, horizontal gene transfer and genomic plasticity drive substantial functional divergence among closely related taxa, diminishing the predictive power of taxonomy-based approaches (Martiny et al., 2015). Recent metagenomic studies found that functional gene

* Supported by the National Key Research and Development Program of China (No. 2022YFF0713102), the Key Technology Research and Development Program of Shandong Province (No. 2024KJHZ033), and the National Natural Science Foundation of China (No. 32200448)
** Corresponding authors: lht@qust.edu.cn; sunly@qibebt.ac.cn
# Fengyun LÜ and Yanmei ZHANG contributed equally to this work and should be regarded as co-first authors.

composition offers a more robust framework for understanding microbial-environment interactions and that functional traits exhibit stronger and more stable associations with environmental variables than taxonomic profiles (Terzin et al., 2025). Yet, sequence-based methods remain constrained by complex sample preparation, computational demands, and the potential loss of intra-population heterogeneity in bulk analyses (van Rossum et al., 2020). These limitations underscore the need for a rapid, non-invasive, and high-throughput classification strategy in marine microbial research.

Raman spectroscopy overcomes key limitations of traditional microbial classification methods, offering a transformative and non-invasive approach. Its ability to analyze unculturable and complex microbiomes without extensive preprocessing makes it particularly valuable for marine microbial research (Hong et al., 2021). Despite its advantages, the analysis of microbial Raman spectra faces significant challenges. The data are high-dimensional, subject to noise, and exhibit minimal spectral differences between closely related species, making separation inherently difficult. Moreover, supervised "close-set" learning approaches can only predict classes included in their training data (Ho et al., 2019). When applied to real-world, uncategorized environmental samples, these models often fail, yielding high false-positive rates due to the presence of unknown and unculturable microorganisms (Parks et al., 2020). Although some studies have proposed "open-set" learning frameworks, such as OpenMax (Bendale and Boult, 2016) and OSDL (Zhu et al., 2025), which aim to reject out-of-distribution samples using threshold mechanisms, they still rely on pre-established spectral databases that are often affected by various batch effects introduced by the state of the Raman device, sample preparation procedures, and environmental conditions. These factors significantly impact spectral data quality, introducing batch effects that reduce the consistency and reliability of the database, thereby limiting their ability to generalize across complex, real-world environments (Guo et al., 2018a; Frempong et al., 2024).

Unsupervised learning methods, on the other hand, do not rely on labeled data and offer an alternative strategy for Raman spectral analysis. Techniques such as principal component analysis (PCA), K-means (K) clustering, and density-based spatial clustering of applications with noise (DBSCAN) have been used in spectra clustering based on intrinsic similarity, enabling microbial differentiation without predefined categories (Tang et al., 2021). More recently, deep learning has further advanced these unsupervised approaches by introducing diverse architectures capable of learning latent representations of Raman spectra. Techniques such as variational autoencoders (VAE), encoder-Transformer models, as well as other deep learning frameworks have demonstrated their effectiveness in tasks ranging from spectral clustering to microbial identification (Thrift et al., 2020; Guo et al., 2022; Sun et al., 2024). In addition, Yu et al. (2020, 2021) demonstrated the application of LSTMs and GANs in accurately identifying marine bacteria isolated from various sources, including the intestines of infected *Urechis unicinctus* larvae and marine environments such as seawater. Among deep learning methods, the DEC (deep embedded clustering) (Xie et al., 2016) uniquely integrates clustering objectives into feature learning, refining cluster assignments through Kullback-Leibler divergence minimization. This approach enables robust spectral representations without large labeled datasets, making it ideal for microbial community analysis. DEC has also proven effective in single-cell RNA-seq batch effect removal and clustering (Li et al., 2020).

In this study, we applied the DEC framework to unsupervised clustering of marine microbial Raman spectra, aiming to facilitate the classification of metabolic functions based on Raman spectroscopy in environmental microbiology. To effectively capture local patterns and intricate structural relationships within the data, we re-engineered the traditional DEC framework by introducing the CDEC (convolutional deep embedded clustering) model, a convolutional denoising autoencoder that enhances feature learning efficiency through expanded feature channels. This modification of convolutional layers enables the model to better capture local spectral features, while pooling layers directly reduce encoding dimensions and mitigating risks of gradient explosion and overfitting. These architectural improvements make CDEC particularly effective for processing complex, high-dimensional, noise-laden Raman spectral data. We evaluated the algorithm's effectiveness across four distinct datasets, ranging from pure cultures and real environmental samples, demonstrating its stability and clustering performance compared to other machine learning methods. Overall, our findings highlight the potential of this approach for the rapid

functional classification of marine microbial communities, facilitating its broader application in environmental microbiology.

2 MATERIAL AND METHOD

2.1 Raman dataset

2.1.1 Microalgae dataset

Three microalgal species (*Chlorella meneghiniana* FACHB-2217, *Synechocystis* PCC 6803, and *Chlamydomonas reinhardtii* GY-H49) were cultivated using TAP medium, BG11 medium, and CSI medium, respectively, until reaching the stationary phase. After centrifugation, cells were washed with deionized water (ddH_2O) for three times and loaded into a capillary tube. Single-cell Raman spectroscopy was performed using a CAST-R system (Qingdao Single-Cell Biotechnology Co., Ltd.) by a 532-nm laser with about 10-mW output power. Spectra were acquired for 0.5 s with a CCD grating of 300 lines/mm.

2.1.2 Marine bacteria-6 dataset

Six marine bacterial strains (*Metabacillus idriensis*, *Vibrio neocaledonicus*, *Exiguobacterium mexicanum*, *Pseudoalteromonas carrageenovora*, *Fictibacillus barbaricus*, and *Halopseudomonas oceani*) were purified from marine environments and taxonomically identified via 16S rRNA sequencing. The bacteria were cultured on marine LB medium for 10 h. Cells were then harvested by centrifugation, washed three times with ddH_2O, and resuspended in phosphate-buffered saline (PBS). High-throughput flow Raman spectroscopy was conducted using a Flow-RACS system (Qingdao Single-Cell Biotechnology Co., Ltd.) by a 532-nm laser with about 250-mW power for 0.8-s accumulation.

2.1.3 Marine environment dataset

Seawater samples were collected from Kiaochow Bay. A total of 2 L of seawater was sequentially filtered through 8-, 2-, and 0.22-μm sterile membrane filters. The environmental samples retained on the 0.22-μm filter were enriched by repeatedly washing the filter three times with 0.5 mL of PBS solution, then were centrifuged for 15 min at 3 000×*g* at 4 °C in a swinging-bucket rotor centrifuge, and the pellet was resuspended in 50 μL of PBS solution. An aliquot of 10 μL was taken for high-throughput Raman spectroscopy using the Flow-RACS system. The spectral acquisition time was set to 0.8 s with a final laser power of approximately 250 mW. This experimental procedure was repeated twice to ensure consistency and reliability.

2.2 Raman spectral preprocessing

All Raman spectra underwent a standardized preprocessing workflow to enhance signal quality and reduce artifacts. Raw spectral data were initially subjected to cosmic ray removal to eliminate sporadic high-intensity spikes that could interfere with subsequent analysis. Partitioned polynomial baseline removal was then performed to mitigate the effects of fluorescence interference and instrument dark currents, which typically manifest as broad, sloping backgrounds in Raman spectra. Spectral smoothing using a Savitzky-Golay algorithm was applied to reduce random noise while preserving peak shapes and intensities characteristic of molecular vibrations. Following noise reduction, spectra were normalized to account for variations in acquisition parameters and sample positioning, enabling reliable inter-sample comparison. This preprocessing pipeline was implemented using the RamEX package (version 0.99.0) in R (Zhang et al., 2025), with all parameters maintained consistently across all samples to ensure analytical reproducibility.

2.3 Convolutional deep embedded clustering (CDEC)

As illustrated in Fig.1, the CDEC model comprised two core modules: a representation learning module and a clustering module. The representation learning module reduced data dimensionality and learns feature representations, while the clustering module performed latent variable clustering.

2.3.1 Representation learning module

This module consisted of an encoder E and a decoder D. The input data x was mapped to the latent space via a convolutional denoising encoder, producing latent variable z. The variable z was then reconstructed by the decoder into $\hat{x}$. The reconstruction loss L_r was calculated using the mean squared error function (MSE), quantifying the discrepancy between the encoder input and decoder output:

$$L_r = \mathrm{MSE}(x,\hat{x}) = \frac{1}{N}\sum_{i=1}^{N}\| x_i - \hat{x}_i \|_2^2, \quad (1)$$

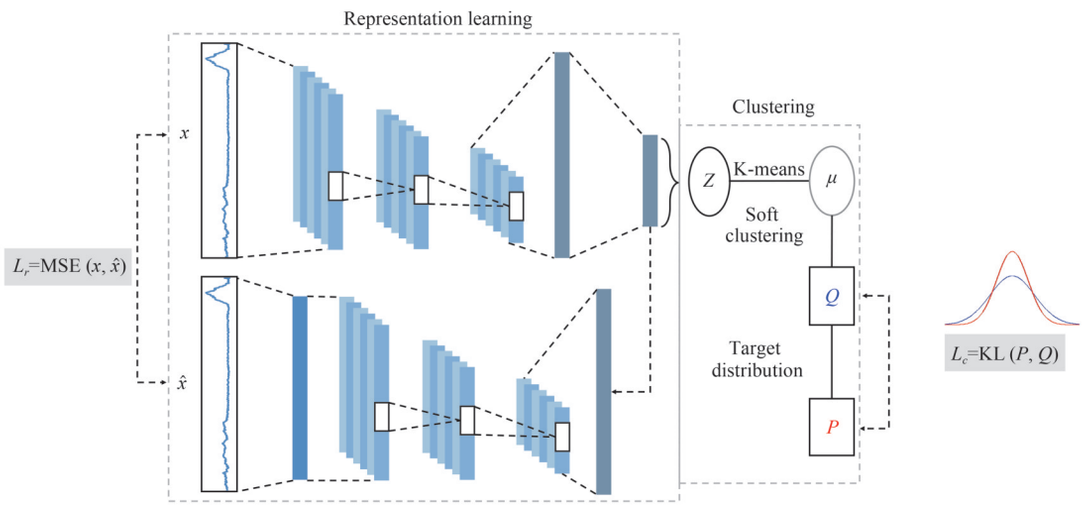

Fig.1 CDEC (convolutional deep embedded clustering) network structure

where N was the number of given spectra, and for the i^{th} spectra x_i denoted the original spectrum while $\hat{x}_i$ represented the reconstructed output produced by the decoder. This loss ensured the autoencoder preserved essential spectral features while discarding noise, serving as a foundation for subsequent clustering.

The convolutional denoising encoder ε processed noise-corrupted inputs x', generated as:

$$x' = x + N\left(\mu, \sigma^2\right). \tag{2}$$

Here, Gaussian noise with μ=0 and σ=0.2 was added to simulate real-world fluorescence interference, forcing the encoder to learn noise-invariant representations.

Then mapped them to low-dimensional latent variables z:

$$z = E\left(x'\right). \tag{3}$$

The encoder architecture included multiple convolutional and linear layers. Specifically, three 1D convolutional layers were used, with kernel sizes of 7, 5, 3, and output channels of 32, 64, and 128, respectively. These layers employed stride=2 and zero-padding, each followed by a ReLU activation and average pooling to enhance training stability. The final embedding was obtained via a linear layer projecting to a 16-dimensional latent space.

The convolutional layers employed 1D kernels to capture local spectral patterns (e.g., characteristic peaks in specific wavenumber ranges), while expanded feature channels enhanced the ability of modeling to represent complex spectral variations. Each convolutional operation was defined as:

$$h^{(l)} = \text{AvgPool}\left(f\left(W^{(l)} * h^{(l-1)} + b^{(l)}\right)\right), \tag{4}$$

where $h^{(l)}$ (the output of the l^{th} layer) was computed by convolving the previous layer's output $h^{(l-1)}$ with learnable kernels $W^{(l)}$ and biases $b^{(l)}$ of the encoder layers, * denoted the convolution operation, and $f(\cdot)$ was the ReLU activation function, and AvgPool($\cdot$) denotes the average pooling operation applied after ReLU. For the last convolutional layer without pooling, the output was directly obtained after ReLU activation. For a 1D convolution with input $h^{(l-1)} \in \mathbb{R}^{C_{\text{in}} \times L_{\text{in}}}$ and kernel $W^{(l)} \in \mathbb{R}^{C_{\text{out}} \times C_{\text{in}} \times K}$, the output at channel c_{out} and position i was calculated as:

$$h^{(l)}_{c_{\text{out}},i} = \text{AvgPool}\left(f\left(\sum_{c_{\text{in}}=1}^{C_{\text{in}}} \sum_{k=1}^{K} W^{(l)}_{c_{\text{out}},c_{\text{in}},k} \cdot h^{(l-1)}_{c_{\text{in}},i+k-1} + b^{(l)}_{c_{\text{out}}}\right)\right), \tag{5}$$

where C_{in} and C_{out} denoted the total numbers of input and output channels, respectively; c_{in} and c_{out} denoted the corresponding channel indices; L_{in} was the input length; K was the kernel size; and i indexed the spatial position along the input. This formulation enabled the model to integrate local neighborhood information across spectral channels, effectively extracting hierarchical representations for subsequent processing.

The final latent variable z was computed as:

$$z = w_l h_{\text{flatten}} + b_l, \tag{6}$$

where w_l and b_l were the weights and biases of the linear layer, and h_{flatten} was the flattened convolutional output.

The decoder D reconstructed the input data from z:

$$\hat{x} = D(z), \tag{7}$$

where the decoder architecture was designed to recover biologically relevant spectral features while suppressing high-frequency noise. It consisted of four transposed convolutional layers with kernel sizes of 3, 5, and 7, and output channels of 128, 64, and 32, respectively. Each transposed convolution used a stride of 1 and zero-padding to maintain the output length, and was followed by a ReLU activation function and an average pooling layer (except for the final layer). These transposed convolutional layers were followed by linear layers to further refine the reconstructed spectra.

Each transposed convolutional operation was defined as:

$$\hat{h}^{(l)} = \mathrm{AvgPool}\left(f\left(W^{(l)} * h^{(l-1)} + b^{(l)} \right) \right), \tag{8}$$

where $h^{(l)}$ was the output of the l^{th} transposed convolutional layer, $W^{(l)}$ and $b^{(l)}$ were the learnable kernel weights and biases of the decoder layers, * denoted the transposed convolution operation, and $f(\cdot)$ is the ReLU activation function. For a 1D transposed convolution with input $h^{(l-1)} \in \mathbb{R}^{C_{\mathrm{in}} \times L_{\mathrm{in}}}$ and kernel $W^{(l)} \in \mathbb{R}^{C_{\mathrm{in}} \times C_{\mathrm{out}} \times K}$, the output at channel c_{out} and position i was calculated as:

$$\hat{h}^{(l)}_{c_{\mathrm{out}},i} = \mathrm{AvgPool}\left(f\left(\sum_{c_{\mathrm{in}}=1}^{C_{\mathrm{in}}} \sum_{k=1}^{K} W^{(l)}_{c_{\mathrm{out}},c_{\mathrm{in}},k} \cdot \hat{h}^{(l-1)}_{c_{\mathrm{in}},i+k-1} + b^{(l)}_{c_{\mathrm{out}}} \right) \right), \tag{9}$$

where C_{in} and C_{out} denoted the total numbers of input and output channels, respectively; c_{in} and c_{out} denoted the corresponding channel indices; L_{in} was the input length; K was the kernel size; and i indexed the spatial position along the output. The transposed convolution operation increased the spatial dimension, enabling the decoder to reconstruct high-resolution spectral features from compressed latent representations.

The final reconstructed output $\hat{x}$ was:

$$\hat{x} = w_{\mathrm{f}} h_{\mathrm{flatten}} + b_{\mathrm{f}}, \tag{10}$$

where w_{f} and b_{f} were the final adjustment layer parameters, and h_{flatten} was the flattened transposed convolutional output.

Functionally, the decoder worked synergistically with the encoder: while the encoder distilled spectral data into low-dimensional metabolic fingerprints, the decoder ensured that these fingerprints remained interpretable by mapping them back to biologically meaningful spectral patterns.

2.3.2 Clustering module

Latent variables z underwent K-means clustering to initialize cluster centers μ_j, followed by soft clustering to generate probabilistic assignments q_{ij}. An auxiliary distribution p_{ij}, derived from q_{ij}, acted as a self-supervision mechanism to refine high-confidence cluster assignments. The clustering loss L_c, defined as the Kullback-Leibler (KL) divergence between p_{ij} and q_{ij}, was:

$$L_c = \mathrm{KL}\left(P \,/\!/\, Q \right) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}, \tag{11}$$

where p is the target distribution derived from q, By backpropagating L_c, the encoder dynamically refined feature extraction to align with cluster structures, enabling end-to-end optimization.

The soft assignment q_{ij} and auxiliary target distribution p_{ij} were computed as:

$$q_{ij} = \frac{\left(1 + \left\| z_i - \mu_j \right\|^2 / \alpha \right)^{-\frac{\alpha+1}{2}}}{\sum_{j'} \left(1 + \left\| z_i - \mu_{j'} \right\|^2 / \alpha \right)^{-\frac{\alpha+1}{2}}}, \tag{12}$$

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_{j'} q_{ij}^2 / \sum_i q_{ij'}}, \tag{13}$$

where α=1 (Student's t-distribution degrees of freedom).

2.3.3 Optimization

The CDEC model was optimized using the Adam optimizer, with learning rates and batch sizes dynamically adapted to the scale and complexity of each dataset. To ensure robust feature learning while avoiding overfitting, smaller datasets (e.g., the Marine Bacteria-3 dataset with 210 spectra and the Microalgae Dataset with 1 320 spectra) employed a conservative learning rate of 0.000 01 and a batch size of 32, whereas larger datasets (e.g., the marine bacteria-6 dataset with 9 969 spectra and the marine environment dataset with 11 267 spectra) utilized a higher learning rate of 0.001 and a batch size of 256 to expedite convergence.

The reconstruction loss L_r (Eq.1) ensured that the autoencoder effectively learned feature representations of the data during the parameter initialization phase. This phase allowed the model to learn noise-resistant representations of key metabolic features.

Following initialization, the model transitioned to the clustering optimization phase, where the

clustering loss L_c (Eq.11), designed to jointly optimize feature extraction and cluster assignments, was propagated through the optimizer to update encoder parameters and minimize the KL divergence loss.

These two loss functions were jointly optimized, enabling the encoder to refine its ability to extract clustering-specific features.

2.4 Evaluation metrics

Three metrics assess clustering performance: accuracy (ACC), normalized mutual information (NMI), and adjusted rand index (ARI). Higher values for these metrics generally indicate superior clustering performance.

Let the ground truth labels be defined as Y with $y_i \in [1, n]$, and predicted hard distribution as $\hat{Y}$ with $\hat{y}_i \in [1, n]$.

ACC measures the average correct classification rate of clustered samples:

$$\mathrm{ACC}\left(\hat{Y}, Y\right) = \max_{g} \frac{1}{N} \sum_{i=1}^{N} 1\left\{y_i = g\left(\hat{y}_i\right)\right\}, \tag{14}$$

where g represents the set of all possible one-to-one mappings between $\hat{Y}$ and Y, while the optimal mapping is efficiently determined using the Hungarian algorithm (Jonker and Volgenant, 1986).

NMI quantifies the mutual information between predicted and ground truth labels, with values ranging in [0,1]:

$$\mathrm{NMI}(\hat{Y}, Y) = \frac{I(\hat{Y}, Y)}{\frac{1}{2}\left[H\left(\hat{Y}\right) + H(Y)\right]}, \tag{15}$$

where $H(\cdot)$ denotes entropy and $I(\cdot)$ represents mutual information. ARI evaluates pairwise sample similarities between $\hat{Y}$ and Y, yielding scores in [-1, 1], where 1 indicates perfect agreement, 0 random assignment, and -1 complete disagreement:

$$\mathrm{ARI}\left(\hat{Y}, Y\right) = \frac{\mathrm{RI} - E(\mathrm{RI})}{\max(\mathrm{RI}) - E(\mathrm{RI})}, \tag{16}$$

with $\mathrm{RI} = \frac{a+b}{C_2^{n_{\text{samples}}}}$, where a and b represent concordant and discordant pairs, respectively.

To comprehensively evaluate clustering performance, two additional metrics—silhouettecoefficient (SC) and davies-bouldin index (DBI)—were used.

SC quantifies both intra-cluster cohesion and inter-cluster separation, ranging from [-1, 1], with higher values indicating better clustering. For each sample, the coefficient $s(i)$ is defined as:

$$s(i) = \frac{\left[b(i) - a(i)\right]}{\max\left\{a(i), b(i)\right\}}, \tag{17}$$

where $a(i)$ is the average distance from sample i to other samples in the same cluster, and $b(i)$ is the smallest average distance to samples in any other cluster. The global SC is the mean of $s(i)$ across all samples.

DBI evaluates cluster compactness and separation, with lower values indicating superior clustering. It is defined as:

$$\mathrm{DBI} = 1/K \sum_{i=1}^{K} \max_{j \neq 1} \left(\frac{\sigma_i + \sigma_j}{d\left(c_i, c_j\right)} \right), \tag{18}$$

where K is the number of clusters, σ_i is the average distance from samples in cluster i to its centroid, and $d(c_i, c_j)$ is the Euclidean distance between centroids of clusters i and j.

3 RESULT

To illustrate the superior performance of the CDEC algorithm in Raman spectral analysis of marine microorganisms, we conducted a four-stage evaluation in which the dataset complexity was progressively increased. First, three morphologically distinct microalgae species (Fig.2) were chosen to test basic discriminatory power under highly similar pigment-dominated spectra. We then analyzed three pure culture marine bacterial species acquired from online repositories (Fig.3), further challenging the model with moderate complexity. Next, we introduced six bacterial species from our own cultures (Fig.4), encompassing diverse sources such as seawater and sediment, to approximate real-world conditions. Finally, we assessed CDEC on authentic marine environmental samples, capturing the full heterogeneity and complexity of the native microbial community (Fig.5). At each stage, CDEC was compared against four well-established clustering algorithms—K-means (MacQueen, 1967), Louvain (Blondel et al., 2008), DBSCAN (Tang et al., 2021), and Deep-BID, a deep clustering method integrating batch correction, cell clustering, and denoising (Qin et al., 2024)—and consistently observed that CDEC outperformed these methods across different degrees of spectral complexity and data diversity.

3.1 Microalgae-3 dataset

We first evaluated the performance of CDEC on a dataset comprising three morphologically and

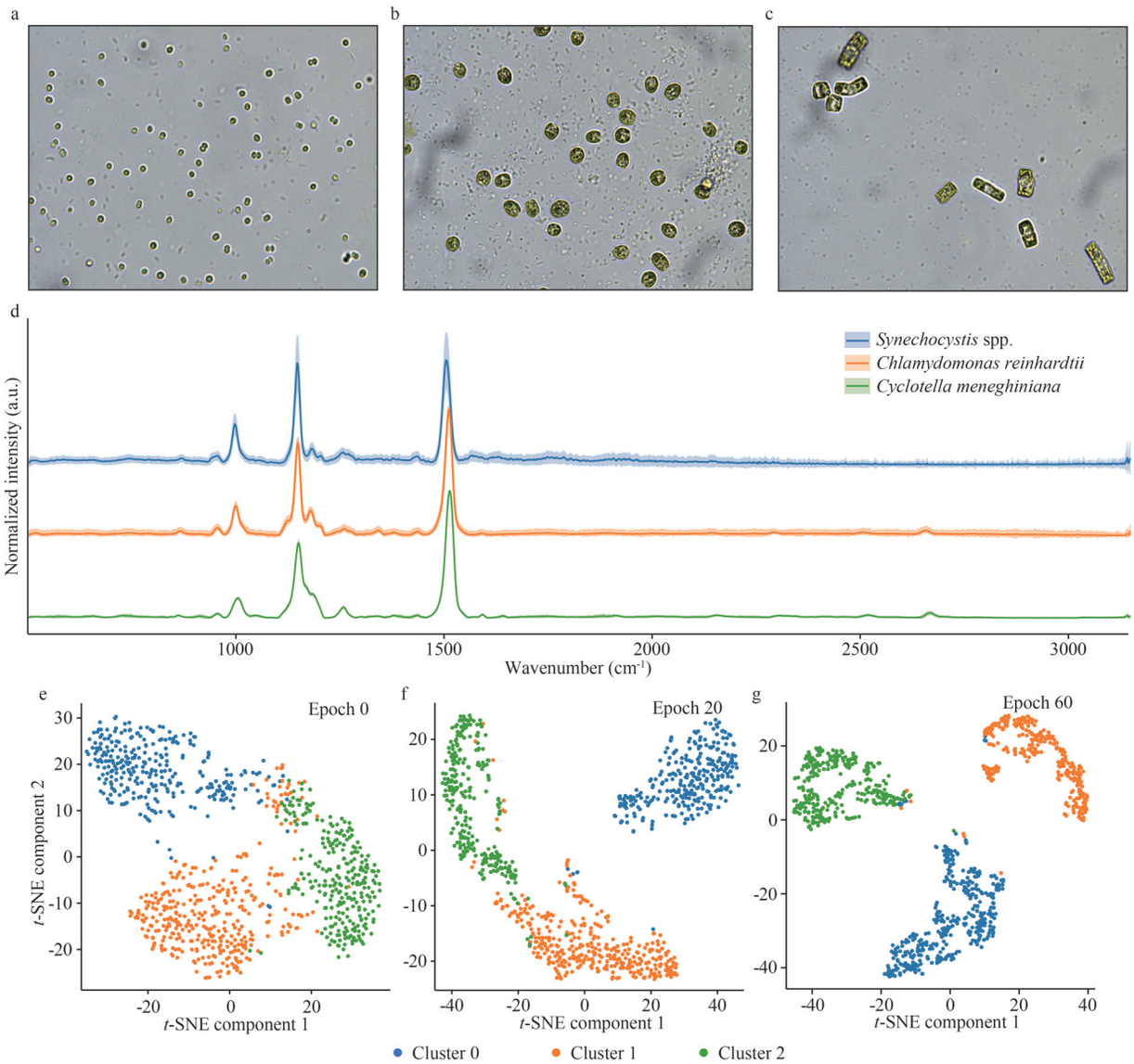

Fig.2 CDEC correctly identified three different species of microalgae

a–c. images of three types of microalgae: a. *Synechocystis* spp.; b. *Chlamydomonas reinhardtii*; c. *Cyclotella meneghiniana*; d. mean Raman spectra of these microalgae; e–g. *t*-SNE plots showing clustering distribution at training epochs 0, 20, and 60.

compositionally distinct microalgae species: *Synechocystis* spp., *Chlamydomonas reinhardtii*, and *Cyclotella meneghiniana* (Fig.2a–c). Despite their clear morphological differences under microscopic observation, the Raman spectra of these species were dominated by overlapping pigment-related peaks (e.g., 1 001, 1 150, and 1 510 cm^{-1}), which posed a challenge for differentiation. Initial inspection of the raw spectra revealed minimal inter-species differences, as distinctive features from lipids and proteins were masked by strong pigment signals (Fig.2d). A preliminary *t*-SNE visualization of the spectral data (Fig.2e) confirmed the difficulty of manual differentiation, with significant overlap among clusters. CDEC, however, progressively refined feature representations through iterative clustering, increasing inter-cluster separations and reducing intra-cluster variability (Fig.2f–g). To evaluate the performance of CDEC against other clustering methods, we employed three metrics: ACC, which measured overall classification accuracy, NMI, which measured the shared information between predicted and true labels, and ARI, which assessed the consistency of cluster assignments (see Section 2.4 for details).

For this dataset, baseline methods were tuned individually to ensure fair comparison. Specifically, K-means was run with k=3, DBSCAN used ε=0.2

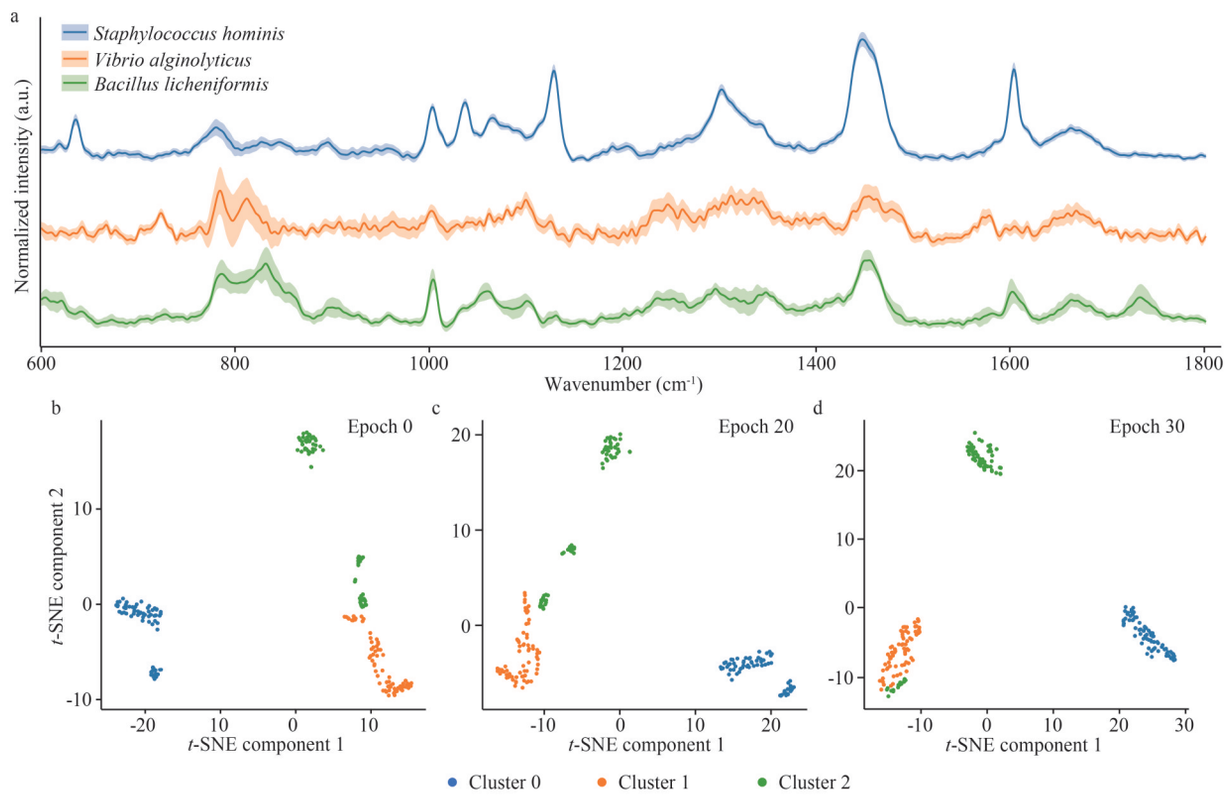

Fig.3 CDEC performance on limited dataset

a. Raman spectra of marine bacteria; b–d. *t*-SNE plots showing clustering distribution at training epochs 0, 20, and 30.

and minPts=5, Louvain applied a resolution parameter of 1.2, and Deep-BID was trained using the same hyperparameters as CDEC.

By the final iteration, CDEC achieved near-perfect cluster separation, with ACC NMI, and ARI reaching 99.46%, 97.13%, and 98.34%, respectively (Table 1). CDEC showed improved performance over baseline methods, including Louvain (NMI: 73.55%) and Deep-BID (ACC: 95.23%), exceeding the latter by over 4% in ACC and more than 10% in NMI and ARI. This improvement was likely due to its convolutional denoising autoencoder, which effectively reduced fluorescence noise and enhanced discriminative spectral features, allowing accurate clustering even in the presence of high spectral similarity.

3.2 Marine bacteria-3 dataset

We next employed a Marine bacteria dataset, which consists of three species (*Bacillus licheniformis*, *Vibrio alginolyticus*, and *Staphylococcus hominis*) to examine CDEC's performance under a slightly complexity with distinct Raman spectral features. The dataset was sourced from a publicly available spectral dataset published by Yu et al. (2020) These bacteria are commonly found in marine environments and have been widely used as test datasets in various species identification and classification studies, making them a suitable baseline for performance evaluation (Yu et al., 2020; Cui et al., 2022; Gracia Moisés et al., 2023).

To ensure fair comparison on this dataset, K-means was run with k=3, DBSCAN with ε=0.2 and minPts=10, and Louvain with resolution=1.0. Deep-BID was trained using the same hyperparameters as CDEC.

Initial examination of average Raman spectra showed that *Staphylococcus hominis* spectra exhibited uniform and distinct peak intensities, resulting in its early segregation into a distinct cluster (Fig.3a–b). In contrast, *Bacillus* and *Vibrio* exhibited spectral variability in peak positions, leading to partial overlaps in initial clustering. CDEC progressively resolved this ambiguity through expanding feature channels with 1D convolutional layers to amplify discriminative features, such as subtle intensity differences and peak shifts. Although the limited sample size (210

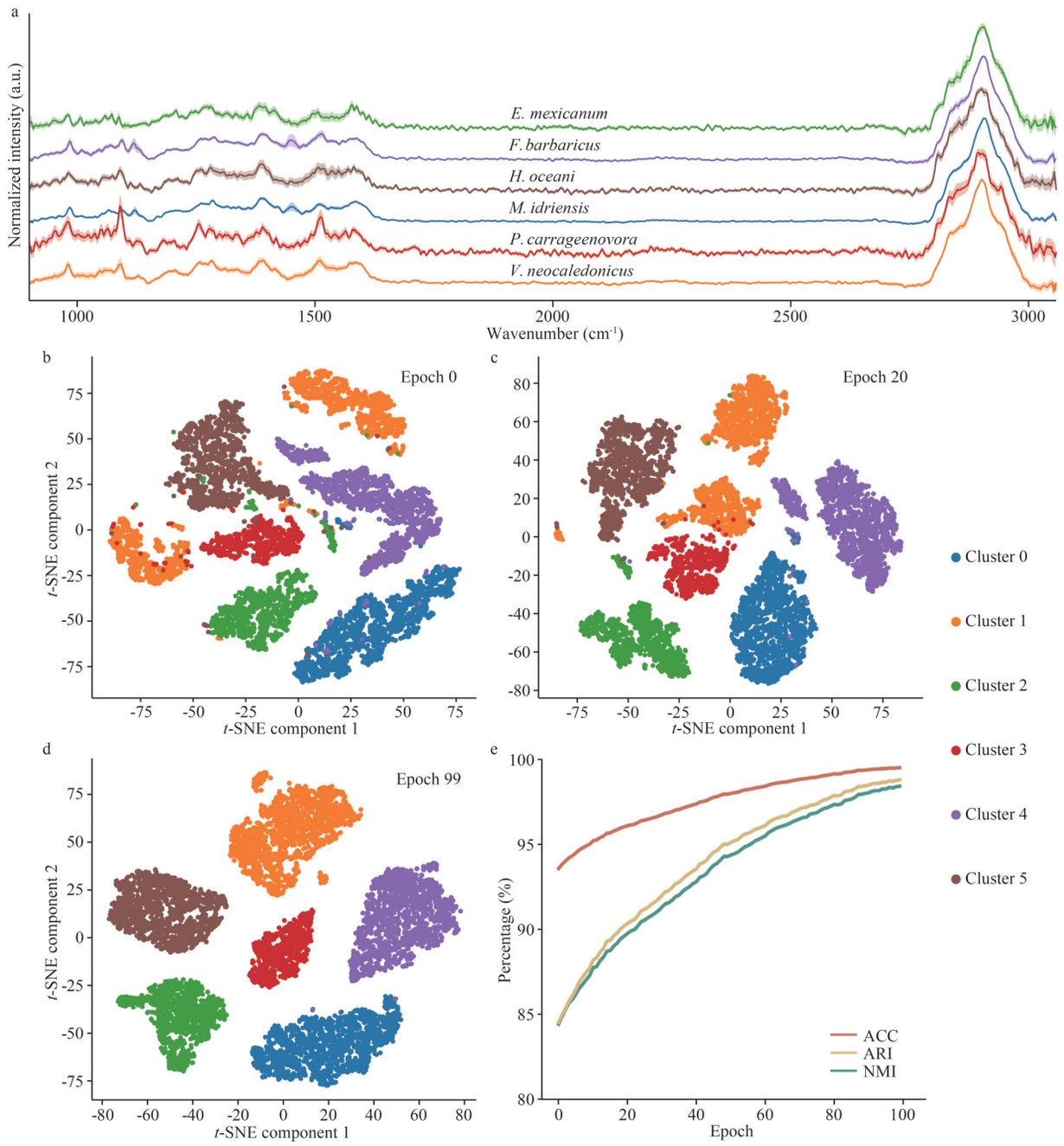

Fig.4 Evaluation of CDEC on a six-species marine bacterial dataset

a. mean Raman spectra; b–d. clustering evolution (*t*-SNE) across training epochs 0 (initial), 20 (mid-optimization), and 99 (final), demonstrating progressive subcluster merging; e. ACC/NMI/ARI for each epoch.

Table 1 Performance on microalgae-3

Method	ACC	NMI	ARI
K-means	63.73%	64.24%	48.37%
DBSCAN	66.73%	65.65%	51.61%
Louvain	66.84%	73.55%	62.52%
DEEP-BID	95.23%	84.52%	86.53%
CDEC	**99.46%**	**97.13%**	**98.34%**

The best performance in each metric was highlighted in bold.

spectra) might constrain best performance, CDEC aggregated fragmented *Vibrio* and *Bacillus* subgroups into coherent clusters. Comparative analysis showed that CDEC outperformed non-deep methods and achieved competitive results compared to Deep-BID (Table 2). On this dataset, Louvain method showed the lowest performance, likely due to a lack of network structure richness caused by the

Table 2 Performance on marine bacteria-3

Method	ACC	NMI	ARI
K-means	85.23%	72.52%	66.05%
DBSCAN	81.42%	78.05%	74.74%
Louvain	60.95%	70.75%	53.45%
DEEP-BID	91.91%	**82.11%**	78.67%
CDEC	**92.85%**	81.03%	**80.76%**

The best performance in each metric was highlighted in bold.

limited sample size. Among non-deep methods, DBSCAN achieved the best overall metrics, though its ACC was slightly lower than K-means. While Deep-BID achieved a marginally higher NMI (82.11% vs. 81.03%), CDEC showed slightly better ACC (92.85% vs. 91.91%) and ARI (80.76% vs. 78.67%). These results highlight the robustness of CDEC on small datasets, which was likely attributed to its noise-resistant encoding, iterative self-supervision, and joint optimization.

As training progressed, CDEC iteratively refined these ambiguous boundaries (Fig.4c–d), achieving a complete separation of the three species.

3.3 Marine bacteria-6 dataset

To further assess robustness in higher dimensions and imbalanced conditions, we tested CDEC on the Marine Bacteria-6 dataset, which consisted of six species with varying sample sizes (889–2 074) and diverse ecological backgrounds, isolated from a range of habitats, including marine sediment, deep-sea environments, coastal waters, sponge mud, and even low-salt or soil environments affected by water (Table 3). Each species played a unique ecological role, making this dataset ideal for simulating realistic marine bacterial communities in challenging conditions.

Baseline methods were configured with dataset-specific parameters selected to ensure fair comparison. Key settings include K-means (k=6), DBSCAN (ε=0.5, minPts=20), Louvain (resolution=1), and Deep-BID (with the same hyperparameters as CDEC).

Full-spectrum Raman profiles displayed clear species-specific characteristic peaks in both position and intensity, highlighting the rich discriminative information suitable for feature extraction (Fig.4a). Initially, clustering showed distinct but partially overlapping groups, with some species fragmented into subclusters—for instance, *Vibrio neocaledonicus* was distributed multiple smaller subgroups (Fig.4b). Through iterative optimization (Fig.4c–d), these fragmented subclusters progressively merged, resulting in tighter intra-cluster cohesion and clearer inter-cluster boundaries. Consequently, clustering performance metrics (ACC, NMI, ARI) showed steady improvement over training epochs (Fig.4e), ultimately achieving near-perfect species separation with ACC of 99.52%, NMI of 98.43%, and ARI of 98.81% (Table 4).

These metrics exceeded the second-best method (K-means ACC: 92.22%) by over 7%, highlighting CDEC's robust capacity to manage both high dimensionality and sample imbalance. Among the four baseline methods, K-means unexpectedly achieved the best performance, whereas DBSCAN performed markedly worse, with its highest scores falling to less than half of those achieved by

Table 3 Summary of marine bacterial species in marine bacteria-6

Species	Role	Common habitat	Ecological function	Gram	Reference
Metabacillus idriensis	Hydrolyze starch	Sponge mud	Promoting the carbon cycle	G+	Zhu et al. (2016); Avdiyuk et al. (2024)
Vibrio neocaledonicus	Corrosion inhibition	Marine sludge	Metal corrosion control	G-	Moradi et al. (2015)
Exiguobacterium mexicanum	Nitrogen removal	Marine sediment	Regulation of nitrogen cycle	G+	Cui et al. (2021)
Pseudoalteromonas carrageenovora	Degraded polysaccharide	Coastal and algae-rich waters	Promotes the recycling of organic matter and the reuse of nutrients such as carbon and sulfur	G-	Zhu et al. (2018); Genicot et al. (2014)
Fictibacillus barbaricus	Hydrolyze certain polysaccharides	Low-salt, freshwater or water-affected soil environments	Involved in organic degradation and nutrient cycling	G+	Glaeser et al. (2013)
Halopseudomonas oceani	Degradation, pressure resistance, temperature resistance	Deep sea	Potential of pollutant degradation and bioremediation	G-	Kruse et al. (2024)

Table 4 Performance on marine bacteria-6

Method	ACC	NMI	ARI
K-means	92.22%	82.93%	81.81%
DBSCAN	37.66%	36.44%	19.48%
Louvain	82.94%	86.87%	81.16%
DEEP-BID	85.07%	72.11%	72.63%
CDEC	**99.52%**	**98.43%**	**98.81%**

The best performance in each metric was highlighted in bold.

K-means. The distinct performance likely reflected DBSCAN's vulnerability to noise and uneven data density, challenges inherent to the imbalanced dataset. CDEC's denoising architecture effectively reduced fluorescence interference, while its clustering module adaptively prioritized rare species, ensuring that larger clusters did not dominate the results.

3.4 Marine environmental dataset

Finally, we evaluated the performance of CDEC on real environmental samples taken from uncultured seawater, which contained highly diverse microbial populations and complex background noise. A total of 11 267 single-cell Raman spectra were measured from the population using the microfluid-based Raman device (Wang et al., 2023), providing a detailed representation of the community's metabolic functionality and heterogeneity at the single-cell level. By applying the CDEC algorithm, the spectral data were effectively grouped into eight clusters, each exhibiting unique Raman peak intensities, reflecting the underlying compositional and functional variability within the microbial population (Fig.5a & c).

To enable a fair and representative evaluation, the parameters of all baseline methods were carefully adjusted for this dataset. K-means was run with a cluster number determined via the elbow method, DBSCAN was configured using ε=0.5 and minPts=10, Louvain's resolution was set to 0.9, and Deep-BID was trained with same hyperparameters as CDEC.

To validate the performance of CDEC on unlabeled environmental samples, we analyzed SC and DBI (Fig.5b) (see Section 2.4 for details). These unsupervised metrics confirmed CDEC's ability to capture intrinsic microbial diversity without relying on ground-truth labels. In contrast, the three non-deep-learning clustering methods yielded inferior results, characterized by low SC and high DBI. While Deep-BID achieved a comparable SC score to our method (0.85 vs. 0.89), its DBI value was significantly higher (1.91 vs. 0.15 for CDEC). This stark contrast in DBI highlighted CDEC's superior robustness against noise and heterogeneity inherent in real-world datasets.

Biomarker analysis of the clusters identified distinct biochemical signatures (Fig.5c), including protein-related peaks (e.g., Amide III band at 1 310–1 350 cm^{-1} and Amide I band at 1 647 cm^{-1}), nucleic acid-related peaks (e.g., guanine vibration at 1 310–1 325 cm^{-1} and nucleic acid backbone mode at 1 330–1 340 cm^{-1}), and lipid-related peaks (e.g., CH_2 bending vibration at 1 320–1 330 cm^{-1} and triglyceride C-C stretching at 890 cm^{-1}). These biomolecular markers highlighted the metabolic and functional roles of proteins, nucleic acids, and lipids in the microbial subpopulations. Among the eight clusters, clusters 0, 4, and 6 were grouped into the same clade (Fig.5c) and exhibited significantly higher intensities in key biomarker regions (peak values >0.2), indicating their association with metabolically active microbial populations. In contrast, the remaining five clusters showed relatively lower intensities (~0.1), which may reflect differences in metabolic states or compositional variations within the community (Fig.5c). Notably, all clusters displayed clear signals in the fingerprint region (1 000–1 800 cm^{-1}) and pronounced CH-related peaks, confirming their biological origin (Fig.5d).

Overall, these findings underscored the ability of the CDEC algorithm, coupled with single-cell Raman spectroscopy, to resolve biochemical and functional heterogeneity in complex microbial communities, shedding light on the rapid characterization of the metabolic and ecological composition of environmental microbes.

4 DISCUSSION

With the development of microfluidic-based Raman devices (Wang et al., 2023), one can easily obtain thousands of single-cell Raman spectra from a population, enabling new opportunities for characterizing environmental bacteria. In this study, we developed and validated the CDEC model to facilitate the unsupervised classification of marine microbial Raman spectra, providing a method for exploring marine microbes from a single-cell metabolic perspective. To systematically evaluate its performance, we conducted a four-stage analysis

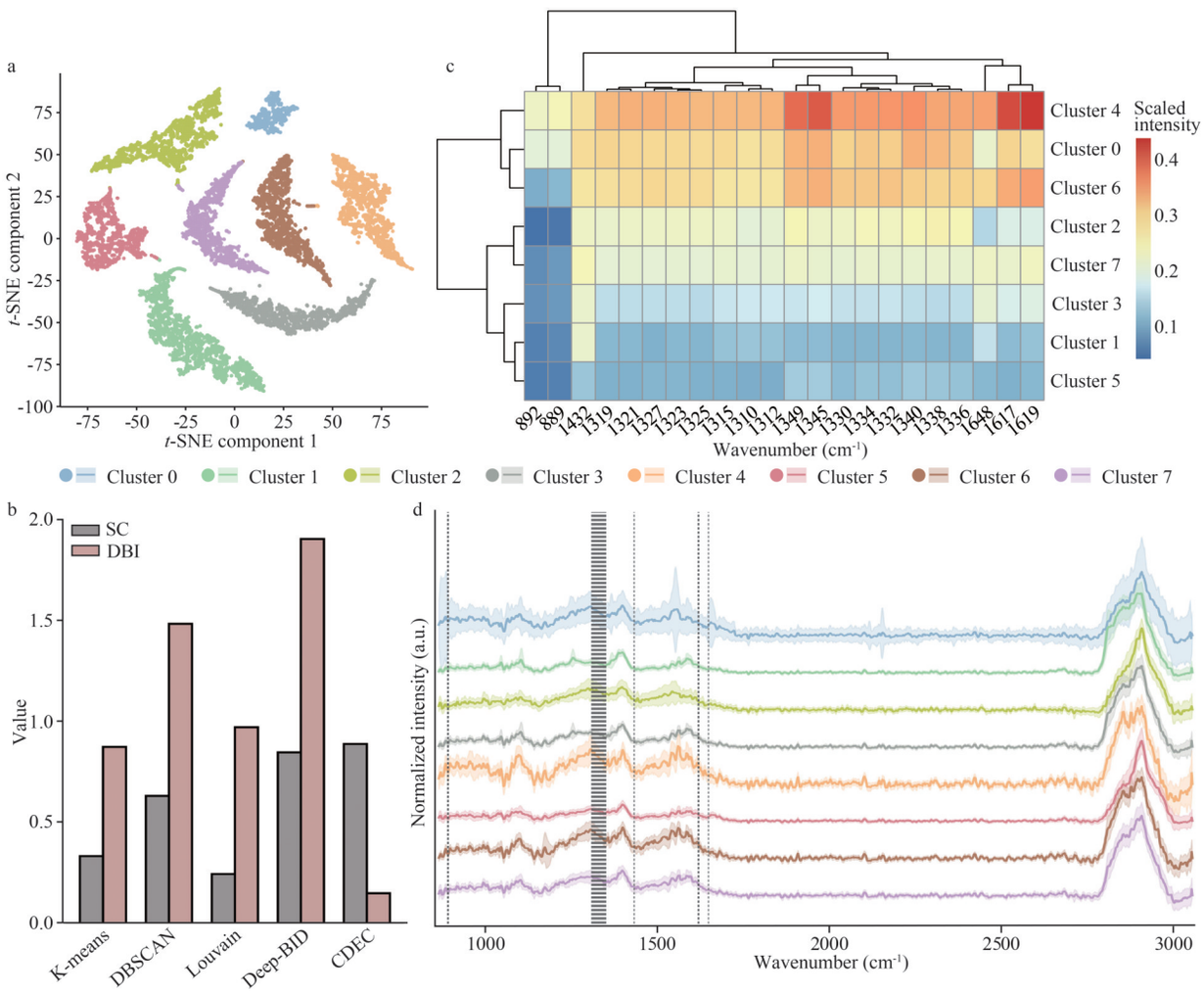

Fig.5 Clustering of marine environment data

a. *t*-SNE visualization showing clustering distribution of marine environmental data; b. SC and DBI of different clustering methods; c. Raman makers for clusters identified by CDEC; d. mean Raman spectrum for each cluster, with dashed lines representing Raman markers.

with increasing dataset complexity, starting with three morphologically distinct microalgae species with similar pigment-dominated spectra (Fig.2), followed by three pure culture marine bacterial species from online repositories (Fig.3), then six bacterial species from diverse marine environments such as seawater and sediment (Fig.4), and finally authentic marine environmental samples (Fig.5) representing the full complexity of native microbial communities. This stepwise approach allowed us to comprehensively assess the robustness and adaptability of the CDEC algorithm across diverse scenarios.

At each stage, CDEC was compared with four widely used clustering algorithms—K-means, Louvain, DBSCAN, and Deep-BID—to evaluate its performance across varying levels of spectral complexity and data diversity. CDEC consistently achieved clustering ACC exceeding 90% across all datasets. Additionally, improvements in NMI and ARI were achieved. For example, CDEC showed NMI increases of 25%, 19%, and 53% compared to K-means, Louvain, and DBSCAN, respectively, as well as ARI improvements of 41%, 41%, and 90% (Tables 1, 2, & 4). The results indicate that CDEC effectively overcomes challenges including high dimensionality, noise, and inter-class similarity in the Raman spectral analysis of marine microbes.

5 CONCLUSION

The success of CDEC lies in its end-to-end co-optimization framework, which integrates feature representation and clustering, supported by a

denoising convolutional autoencoder. This approach effectively captures both local spectral peaks and global spectral trends, which are essential for distinguishing closely related species. While aligning with recent advancements in unsupervised Raman spectral analysis, such as RamanCluster (Sun et al., 2024), which uses transformers for improved representation learning, CDEC uniquely incorporates clustering loss into the optimization process. This integration enables the encoder to extract discriminative features, thereby enhancing clustering accuracy.

For environmental microbes, functional profiling is often performed using metagenomic or metatranscriptomic sequencing (Burke et al., 2011). Here we demonstrate that single-cell Raman spectroscopy-based clustering provides a rapid approach to distinguish cell groups with distinct spectral profiles, offering a preliminary means to infer the functional potential of microbial communities. Critically, Raman spectroscopy captures real-time phenotypic-level insights such as metabolic states (e.g., lipid accumulation, protein synthesis) at single-cell resolution, while sequencing methods reveal genotype-level information including taxonomic identity and functional gene annotations. Recent advances have further highlighted the potential of integrating Raman spectroscopy with molecular profiling techniques. Research by Le Reste et al. (2021) demonstrated that combining Raman spectroscopy with transcriptomic data enabled successful prediction of patient prognosis and tumor subtype classification in glioblastoma multiforme. In the study by Kobayashi-Kirschvink et al. (2024), Raman spectroscopy was employed to non-destructively predict the gene expression profiles within live cells, thus offering a novel approach for dynamically monitoring changes in cell states. Despite its speed and efficiency, Raman spectroscopy is inherently limited in the depth and resolution of information it can provide, falling short of the accuracy achieved by sequencing-based approaches. Future efforts should focus on developing multimodal frameworks that integrate Raman-based clustering with sequencing validation, aiming to enhance the resolution and accuracy of microbial functional profiling and achieve a more comprehensive understanding of microbial community structure and function.

Furthermore, a key challenge for Raman spectroscopy lies in addressing batch effects, which can arise from both external factors, such as instrumentation variability, and intrinsic biological factors, including subtle metabolic shifts within samples (Guo et al., 2018b; Frempong et al., 2024). Standardized experimental protocols are essential to mitigate variability, but overcoming both external and biological sources of batch effects will be crucial for improving the reliability and robustness of Raman-based methods in microbial ecology.

6 DATA AVAILABILITY STATEMENT

Data used in this work are available upon request from the corresponding author.

References

Avdiyuk K V, Ivanytsia V O, Varbanets L D et al. 2024. Bacteria of deep-sea sediments of the black sea breaking down keratin. *Mikrobiolohichnyi Zhurnal*, **86**(3): 18-26, https://doi.org/10.15407/microbiolj86.03.018.

Bendale A, Boult T E. 2016. Towards open set deep networks. *In*: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Las Vegas, USA. p.1563-1572, https://doi.org/10.1109/CVPR.2016.173.

Blondel V D, Guillaume J L, Lambiotte R et al. 2008. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics*: *Theory and Experiment*, **2008**(1): P10008, https://doi.org/10.1088/1742-5468/2008/10/P10008.

Burke C, Steinberg P, Rusch D et al. 2011. Bacterial community assembly based on functional genes rather than species. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(34): 14288-14293, https://doi.org/10.1073/pnas.1101591108.

Cui D Y, Kong L C, Wang Y et al. 2022. *In situ* identification of environmental microorganisms with Raman spectroscopy. *Environmental Science and Ecotechnology*, **11**: 100187, https://doi.org/10.1016/j.ese.2022.100187.

Cui Y, Cui Y W, Huang J L. 2021. A novel halophilic *Exiguobacterium mexicanum* strain removes nitrogen from saline wastewater via heterotrophic nitrification and aerobic denitrification. *Bioresource Technology*, **333**: 125189, https://doi.org/10.1016/j.biortech.2021.125189.

Falkowski P G, Fenchel T, Delong E F. 2008. The microbial engines that drive Earth's biogeochemical cycles. *Science*, **320**(5879): 1034-1039, https://doi.org/10.1126/science.1153213.

Frempong S B, Salbreiter M, Mostafapour S et al. 2024. Illuminating the tiny world: a navigation guide for proper Raman studies on microorganisms. *Molecules*, **29**(5): 1077, https://doi.org/10.3390/molecules29051077.

Genicot S M, Groisillier A, Rogniaux H et al. 2014. Discovery of a novel iota carrageenan sulfatase isolated

from the marine bacterium *Pseudoalteromonas carrageenovora*. *Frontiers in Chemistry*, **2**: 67, https://doi.org/10.3389/fchem.2014.00067.

Glaeser S P, Dott W, Busse H J et al. 2013. *Fictibacillus phosphorivorans* gen. nov., sp. nov. and proposal to reclassify *Bacillus arsenicus*, *Bacillus barbaricus*, *Bacillus macauensis*, *Bacillus nanhaiensis*, *Bacillus rigui*, *Bacillus solisalsi* and *Bacillus gelatini* in the genus *Fictibacillus*. *International Journal of Systematic and Evolutionary Microbiology*, **63**(Pt 8): 2934-2944, https://doi.org/10.1099/ijs.0.049171-0.

Gracia Moisés A, Vitoria Pascual I, Imas González J J et al. 2023. Data augmentation techniques for machine learning applied to optical spectroscopy datasets in agrifood applications: a comprehensive review. *Sensors*, **23**(20): 8562, https://doi.org/10.3390/s23208562.

Guo S X, Heinke R, Stöckel S et al. 2018a. Model transfer for Raman-spectroscopy-based bacterial classification. *Journal of Raman Spectroscopy*, **49**(4): 627-637, https://doi.org/10.1002/jrs.5343.

Guo S X, Kohler A, Zimmermann B et al. 2018b. Extended multiplicative signal correction based model transfer for raman spectroscopy in biological applications. *Analytical Chemistry*, **90**(16): 9787-9795, https://doi.org/10.1021/acs.analchem.8b01536.

Guo Y X, Jin W Q, Wang W L et al. 2022. Unsupervised convolutional variational autoencoder deep embedding clustering for Raman spectra. *Analytical Methods*, **14**(39): 3898-3910, https://doi.org/10.1039/d2ay01184k.

Ho C S, Jean N, Hogan C A et al. 2019. Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. *Nature Communications*, **10**(1): 4927, https://doi.org/10.1038/s41467-019-12898-9.

Hong J K, Kim S B, Lyou E S et al. 2021. Microbial phenomics linking the phenotype to function: the potential of Raman spectroscopy. *Journal of Microbiology*, **59**(3): 249-258, https://doi.org/10.1007/s12275-021-0590-1.

Jonker R, Volgenant T. 1986. Improving the Hungarian assignment algorithm. *Operations Research Letters*, **5**(4): 171-175, https://doi.org/10.1016/0167-6377(86)90073-8.

Kobayashi-Kirschvink K J, Comiter C S, Gaddam S et al. 2024. Prediction of single-cell RNA expression profiles in live cells by Raman microscopy with Raman2RNA. *Nature Biotechnology*, **42**(11): 1726-1734, https://doi.org/10.1038/s41587-023-02082-2.

Kruse L, Loeschcke A, de Witt J et al. 2024. *Halopseudomonas* species: cultivation and molecular genetic tools. *Microbial Biotechnology*, **17**(1): e14369, https://doi.org/10.1111/1751-7915.14369.

Langille M G I, Zaneveld J, Caporaso J G et al. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, **31**(9): 814-821, https://doi.org/10.1038/nbt.2676.

Le Reste P J, Pilalis E, Aubry M et al. 2021. Integration of Raman spectra with transcriptome data in glioblastoma multiforme defines tumour subtypes and predicts patient outcome. *Journal of Cellular and Molecular Medicine*, **25**(23): 10846-10856, https://doi.org/10.1111/jcmm.16902.

Li X J, Wang K, Lyu Y et al. 2020. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nature Communications*, **11**(1): 2338, https://doi.org/10.1038/s41467-020-15851-3.

MacQueen J. 1967. Some methods for classification and analysis of multivariate observations. *In*: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Statistical Laboratory of the University, Berkeley, USA. p.281-297.

Martiny J B H, Jones S E, Lennon J T et al. 2015. Microbiomes in light of traits: a phylogenetic perspective. *Science*, **350**(6261): aac9323, https://doi.org/10.1126/science.aac9323.

Moradi M, Song Z L, Tao X. 2015. Introducing a novel bacterium, *Vibrio neocaledonicus* sp., with the highest corrosion inhibition efficiency. *Electrochemistry Communications*, **51**: 64-68, https://doi.org/10.1016/j.elecom.2014.12.007.

Parks D H, Chuvochina M, Chaumeil P A et al. 2020. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature Biotechnology*, **38**(9): 1079-1086, https://doi.org/10.1038/s41587-020-0501-8.

Qin L, Zhang G Y, Zhang S Q et al. 2024. Deep batch integration and Denoise of single-cell RNA-Seq data. *Advanced Science*, **11**(29): e2308934, https://doi.org/10.1002/advs.202308934.

Sun Z J, Wang Z, & Jiang M Q. 2024. RamanCluster: a deep clustering-based framework for unsupervised Raman spectral identification of pathogenic bacteria. *Talanta*, **275**: 126076, https://doi.org/10.1016/j.talanta.2024.126076.

Tang J W, Liu Q H, Yin X C et al. 2021. Comparative analysis of machine learning algorithms on surface enhanced Raman spectra of clinical *Staphylococcus* species. *Frontiers in Microbiology*, **12**: 696921, https://doi.org/10.3389/fmicb.2021.696921.

Taş N, de Jong A E E, Li Y M et al. 2021. Metagenomic tools in microbial ecology research. *Current Opinion in Biotechnology*, **67**: 184-191, https://doi.org/10.1016/j.copbio.2021.01.019.

Terzin M, Robbins S J, Bell S C et al. 2025. Gene content of seawater microbes is a strong predictor of water chemistry across the great barrier reef. *Microbiome*, **13**(1): 11, https://doi.org/10.1186/s40168-024-01972-0.

Thrift W J, Ronaghi S, Samad M et al. 2020. Deep learning analysis of vibrational spectra of bacterial lysate for rapid antimicrobial susceptibility testing. *ACS Nano*, **14**(11): 15336-15348, https://doi.org/10.1021/acsnano.0c05693.

Van Rossum T, Ferretti P, Maistrenko O M et al. 2020. Diversity within species: interpreting strains in microbiomes. *Nature Reviews Microbiology*, **18**(9): 491-506, https://doi.org/10.1038/s41579-020-0368-1.

Wang X X, Ren L H, Diao Z D et al. 2023. Robust spontaneous Raman flow Cytometry for single-cell

metabolic Phenome profiling via pDEP-DLD-RFC. *Advanced Science*, **10**(16): 2207497, https://doi.org/10.1002/advs.202207497.

Xie J Y, Girshick R B, Farhadi A. 2016. Unsupervised deep embedding for clustering analysis. *In*: Proceedings of the 33rd International Conference on Machine Learning. ICML, New York, USA. p.478-487.

Yu S X, Li H F, Li X et al. 2020. Classification of pathogens by Raman spectroscopy combined with generative adversarial networks. *Science of the Total Environment*, **726**: 138477, https://doi.org/10.1016/j.scitotenv.2020.138477.

Yu S X, Li X, Lu W L et al. 2021. Analysis of Raman spectra by using deep learning methods in the identification of marine pathogens. *Analytical Chemistry*, **93**(32): 11089-11098, https://doi.org/10.1021/acs.analchem.1c00431.

Zhang Y M, Jing G C, Chen R Z et al. 2025. RamEx: an R package for high-throughput microbial ramanome analyses with accurate quality assessment. *bioRxiv*, https://doi.org/10.1101/2025.03.10.642505.

Zhu D C, Zhang P P, Niu L L et al. 2016. *Bacillus ectoiniformans* sp. nov., a halotolerant bacterium isolated from deep-sea sediments. *International Journal of Systematic and Evolutionary Microbiology*, **66**(2): 616-622, https://doi.org/10.1099/ijsem.0.000763.

Zhu L J, Yang Y N, Xu F et al. 2025. Open-set deep learning–enabled single-cell Raman spectroscopy for rapid identification of airborne pathogens in real-world environments. *Science Advances*, **11**(2): eadp7991, https://doi.org/10.1126/sciadv.adp7991.

Zhu Y B, Qiao C C, Li H B et al. 2018. Improvement thermostability of *Pseudoalteromonas carrageenovora* arylsulfatase by rational design. *International Journal of Biological Macromolecules*, **108**: 953-959, https://doi.org/10.1016/j.ijbiomac.2017.11.014.