Supporting Information

A new fusion strategy for rapid strain differentiation based on MALDI-TOF MS spectrometry and Raman spectra

Jian Song, [#a,c] Wenlong Liang, [#c,e] Hongtao Huang, [#d] Hongyan Jia, [a] Shouning Yang, [a] Chunlei Wang*[e] and Huayan Yang *[a,b]

[a] NMPA Key Laboratory for Research and Evaluation of Innovative Drug, Henan Key Laboratory of Organic Functional Molecule and Drug Innovation, Collaborative Innovation Center of Henan Province for Green Manufacturing of Fine Chemicals, School of Chemistry and Chemical Engineering, Henan Normal University, Xinxiang, Henan 453007, China

[b] Shanghai Applied Radiation Institute, School of Environmental and Chemical Engineering, Shanghai University, Shanghai 200444, China

[c] School of physics, Henan Normal University, Xinxiang, Henan 453007, China

[d] College of Educational Information Technology, Henan Normal University, Xinxiang, Henan 453007, China

[e] International Joint Laboratory of Catalytic Chemistry, College of Science, Shanghai University, Shanghai 20044, China.

* **Email:** Huayan Yang: yanghuayan@shu.edu.cn; Chunlei Wang: wangchunlei1982@shu.edu.cn.

[#] *These authors contributed equally.*

Feature Extractor (PCA-LDA)

Feature extraction is critical for improving the performance of most learning algorithms. Extracting different features not only reduces the complexity of the subsequent model training but also improves the accuracy of the subsequent classification and prediction as well as the efficiency of the machine learning models[1, 2]. Linear discriminant analysis (LDA) is a well-known method for dimension reduction [3]. The goal of LDA is to maximize the between-class scatter matrix measure (S_b) while minimizing the within-class scatter matrix measure (S_W). In other words, the LDA method not only magnifies the differences between different categories but also reduces the differences within the same category. Such an approach could make it easier for machine-learning models to identify category differences. However, the number of acquired spectra is often smaller than the vector dimension representing them. S_W is always singular, rendering the LDA model unusable.[4]. To improve the singularity of the S_W, the number of spectra is greater than the vector dimension, a dimension reduction method can be linked with LDA methods [5]. PCA is also a dimension reduction method for dimensionality reduction and yields projection directions that maximize the total scatter across all classes. The PCA method extracts the main features of the spectra and makes the within-class scatter matrix nonsingular. The PCA-processed data allows the LDA method to directly project from a high dimension to a low dimension[5]. If S_W is nonsingular, the optimal projection matrix (W_{opt}) of LDA method is denoted by:

$$W_{opt} = argmax_W \frac{|W^T S_b W|}{|W^T S_W W|}$$

The optimal projection matrix (W_{opt}) obtained by combining the PCA and LDA methods is denoted by:

$$W_{opt}^{\ T} = W_{lda}^{\ T} W_{pca}^{\ T} = argmax_w \frac{|W^T W_{pca}^{\ T} S_b W_{pca} W|}{|W^T W_{pca}^{\ T} S_W W_{pca} W|}{}^T W_{pca}^{\ T}$$

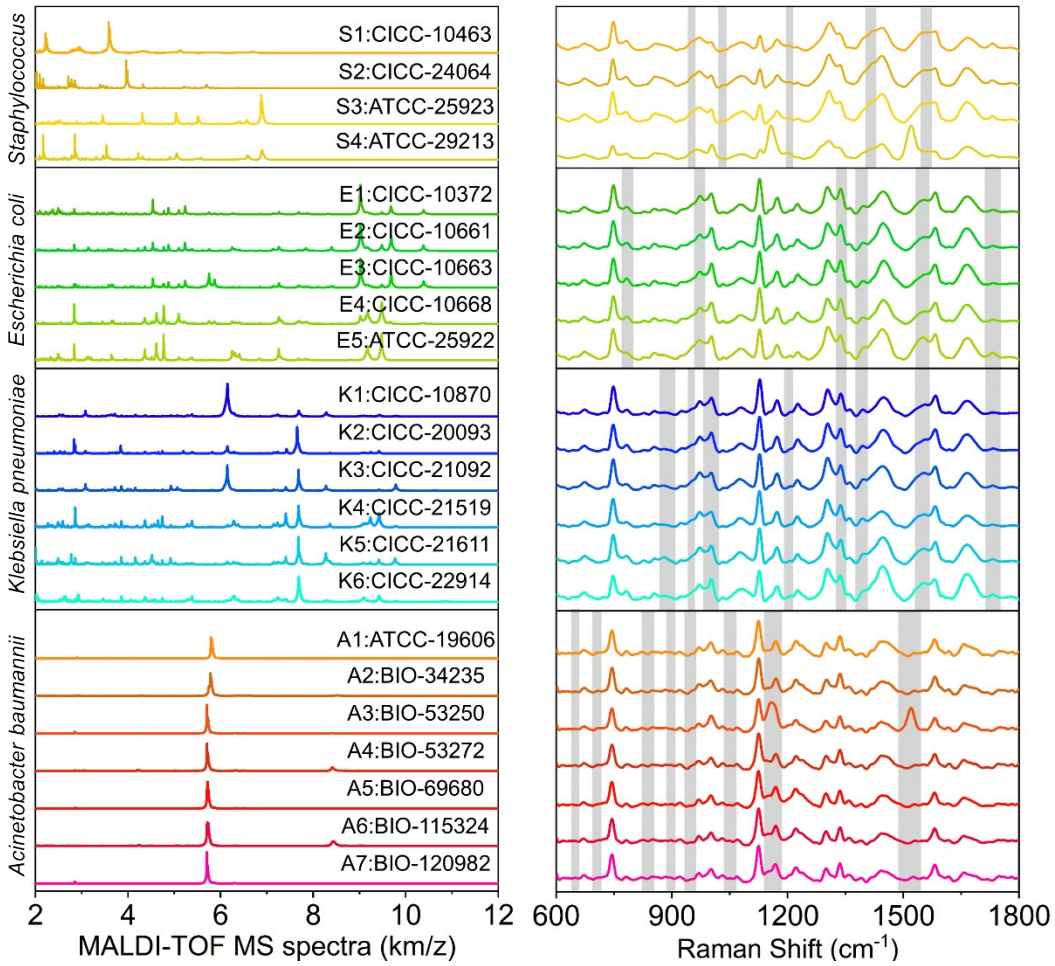

Fig. S1 MALDI-TOF MS spectra of 22 bacterial isolates from different genera (*Staphylococcus*, *Escherichia coli* (*E. coli*) and species (*Klebsiella pneumoniae* (*K. pneumoniae*), *Acinetobacter baumannii* (*A. baumannii*)) (LEFT). Average Raman spectra of 22 bacterial isolates (RIGHT).

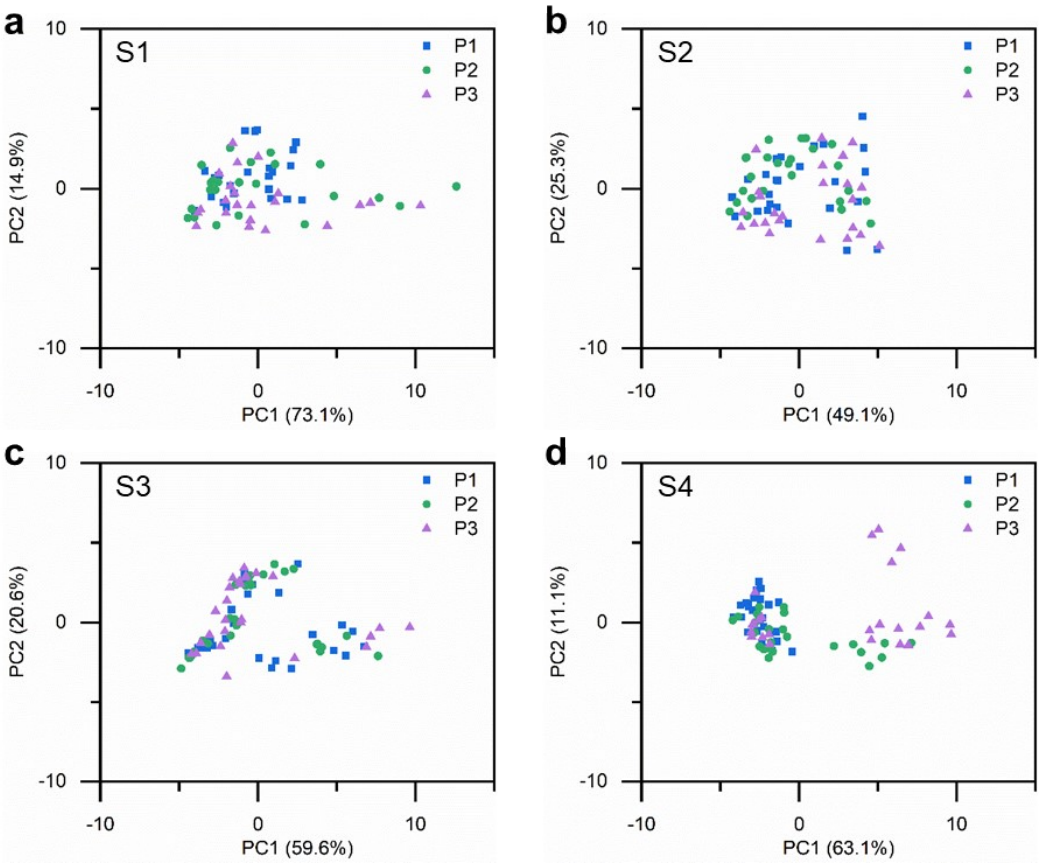

Fig S2. PCA of MALDI-TOF MS data of S1, S2, S3 and S4 form *Staphylococcus* visualized in a score plot.

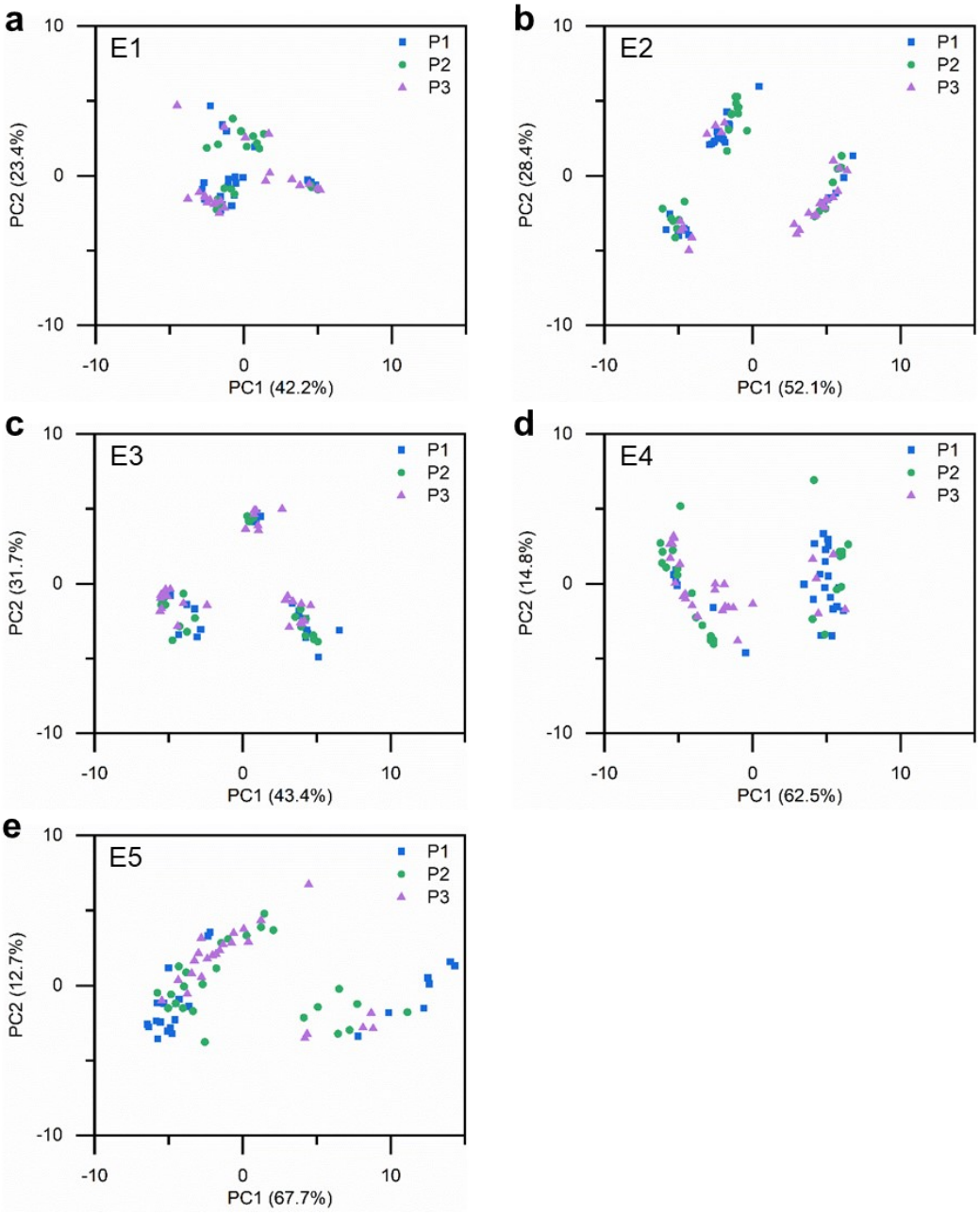

Fig S3. PCA of MALDI-TOF MS data of E1, E2, E3, E4 and E5 form *E. coli* visualized in a score plot.

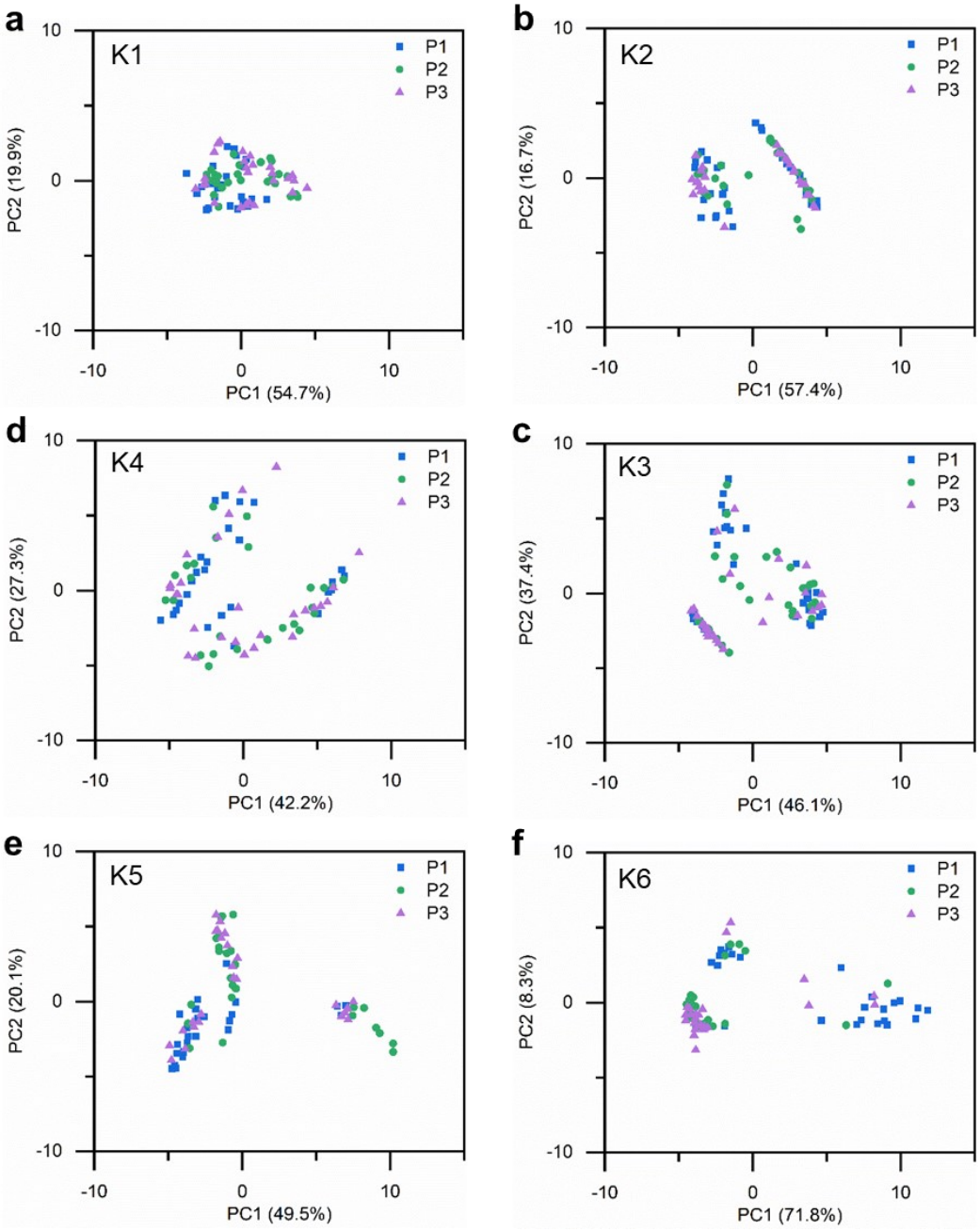

Fig S4. PCA of MALDI-TOF MS data of K1, K2, K3, K4, K5 and K6 form *K. pneumoniae* visualized in a score plot.

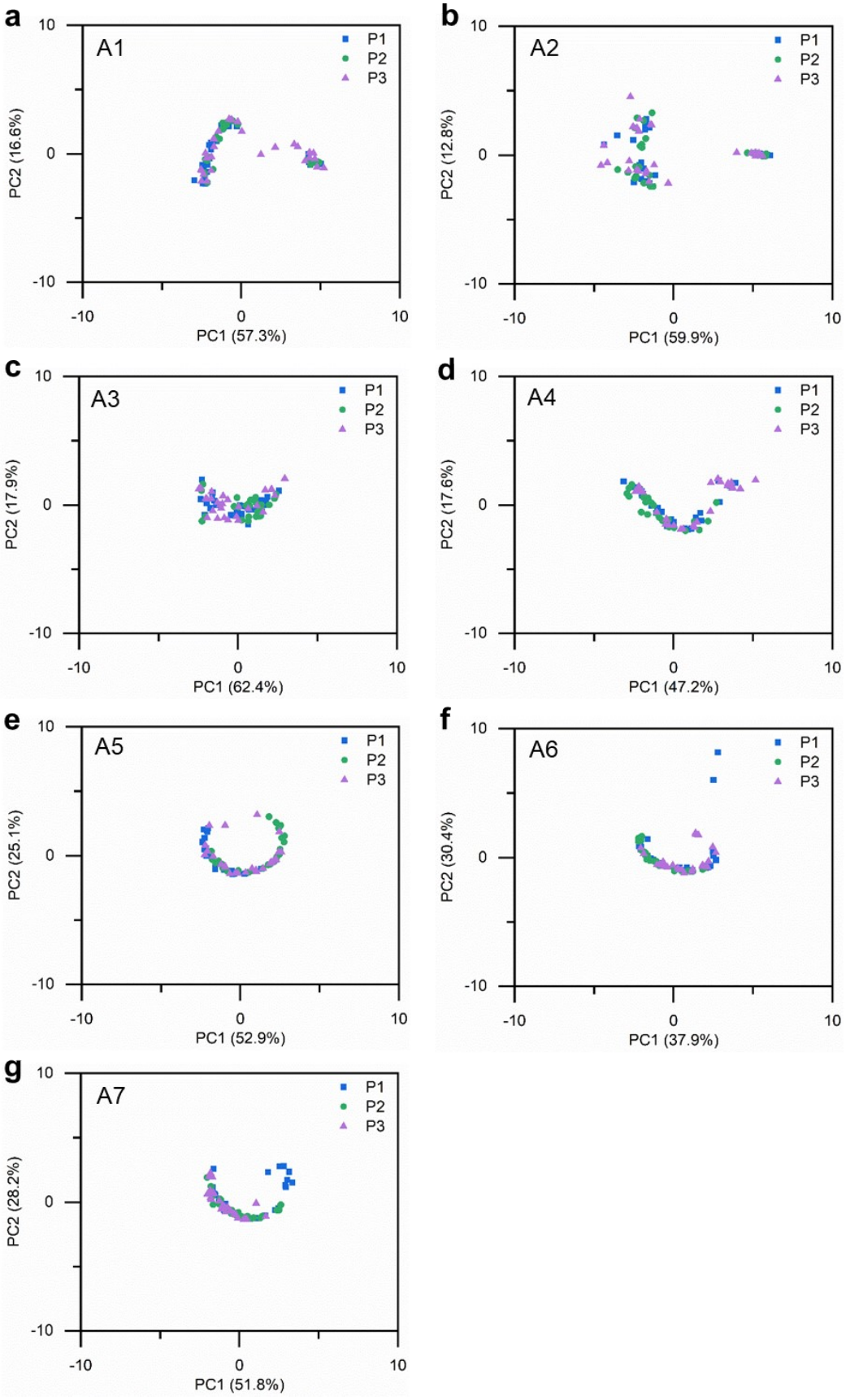

Fig S5. PCA of MALDI-TOF MS data of A1, A2, A3, A4, A5, A6 and A7 form *A. baumannii* visualized in a score plot.

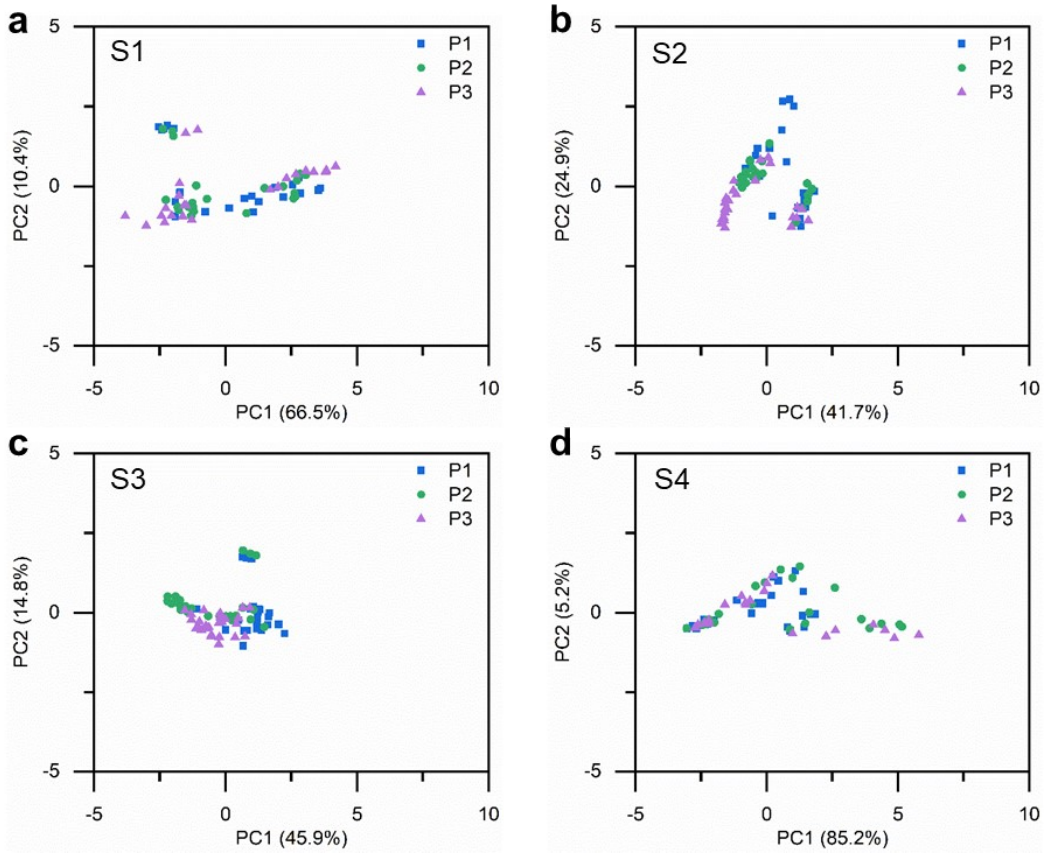

Fig S6. PCA of Raman spectra data of S1, S2 S3 and S4 form *Staphylococcus* visualized in a score plot.

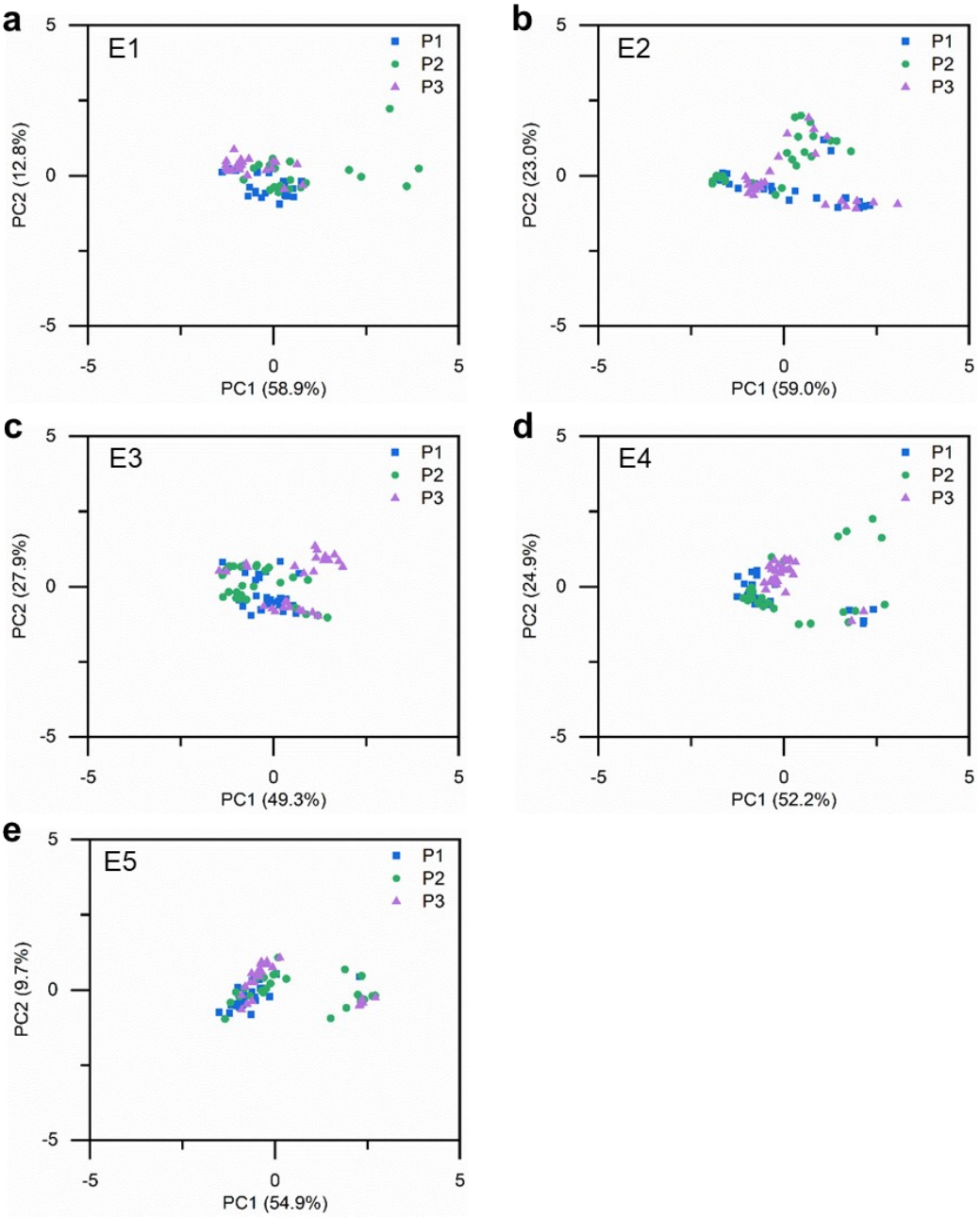

Fig S7. PCA of Raman spectra data of E1, E2, E3, E4 and E5 form *E. coli* visualized in a score plot.

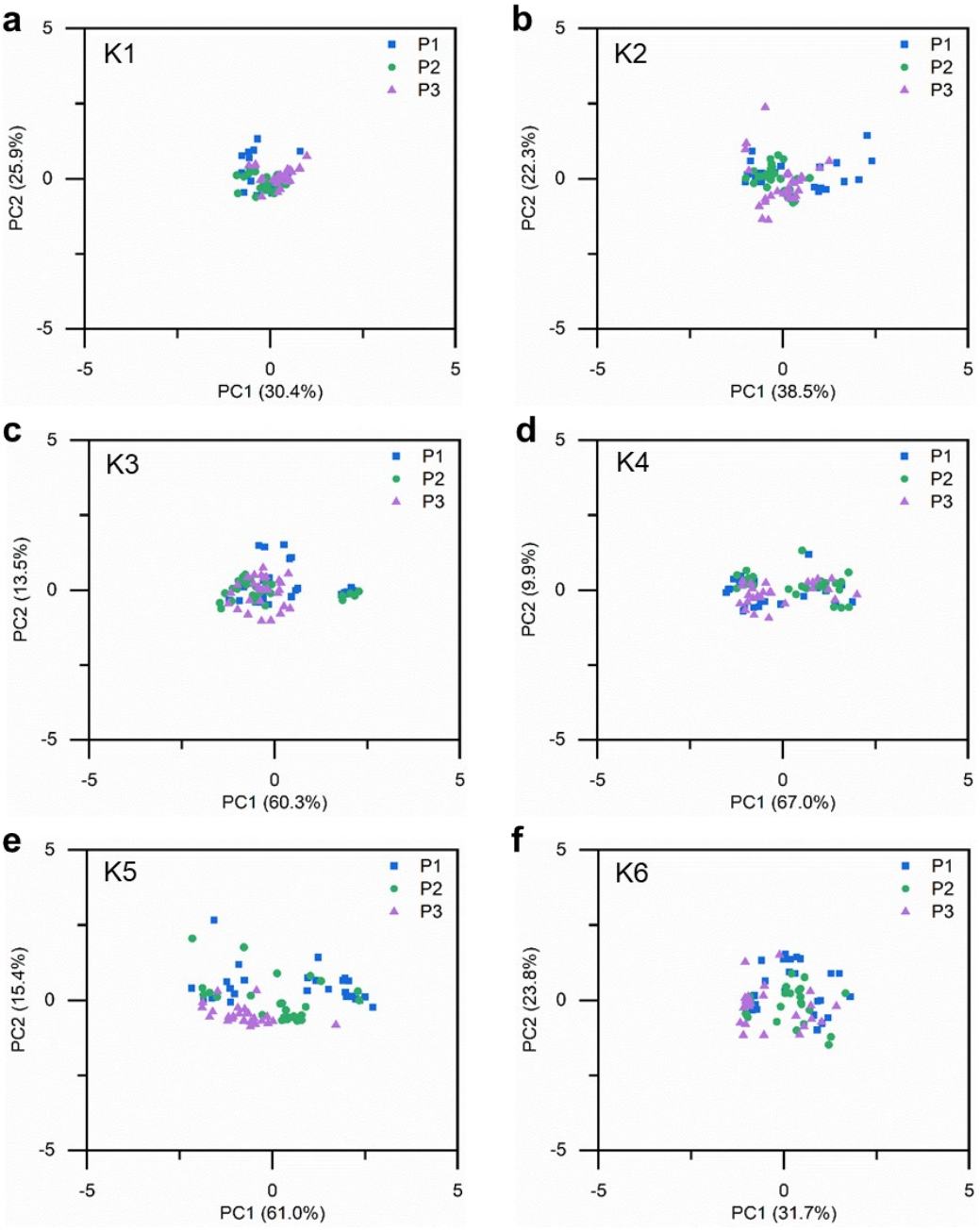

Fig S8. PCA of Raman spectra data of K1, K2, K3, K4, K5 and K6 form *K. pneumoniae* visualized in a score plot.

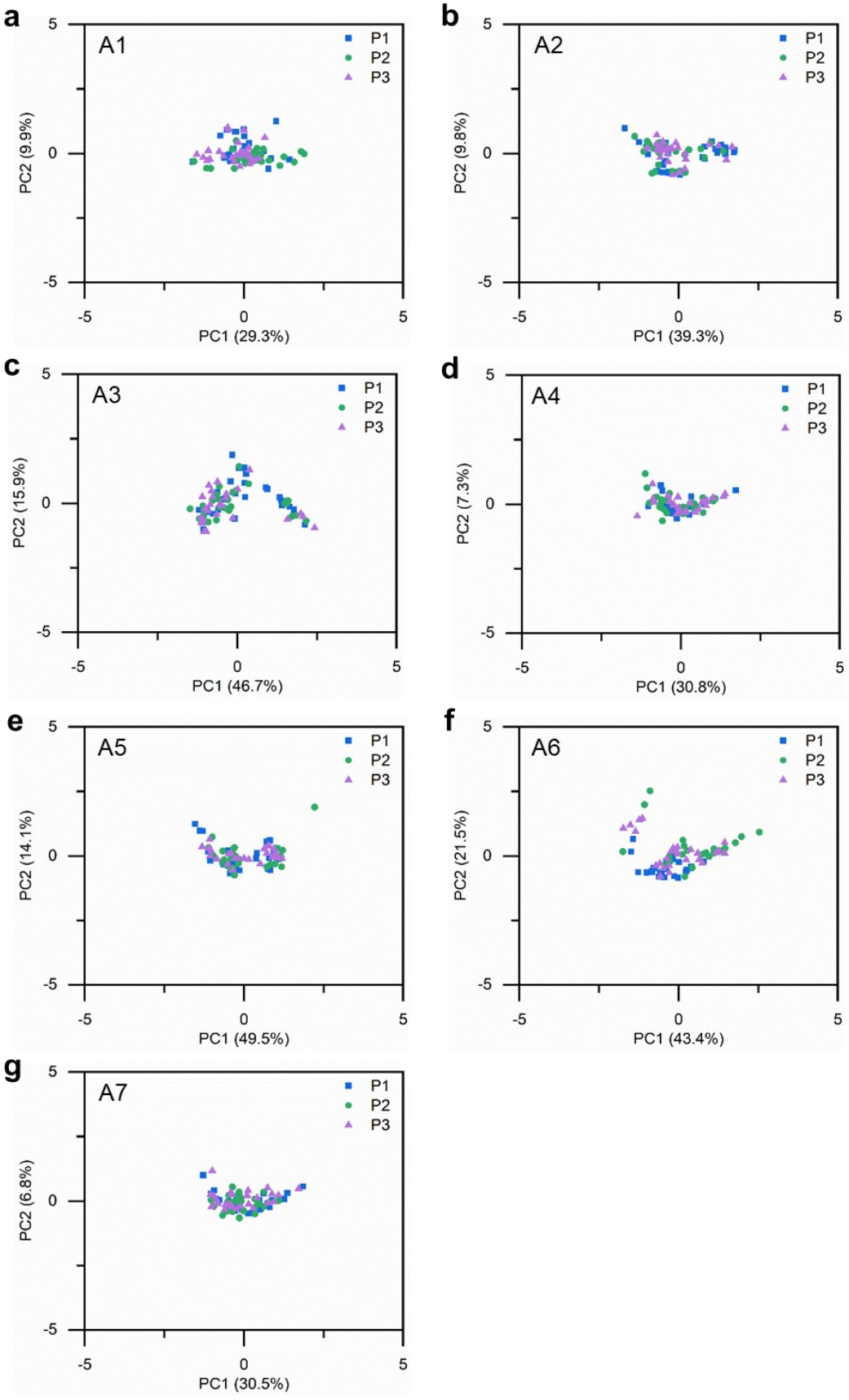

Fig S9. PCA of Raman spectra data of A1, A2, A3, A4, A5, A6 and A7 form *A. baumannii* visualized in a score plot.

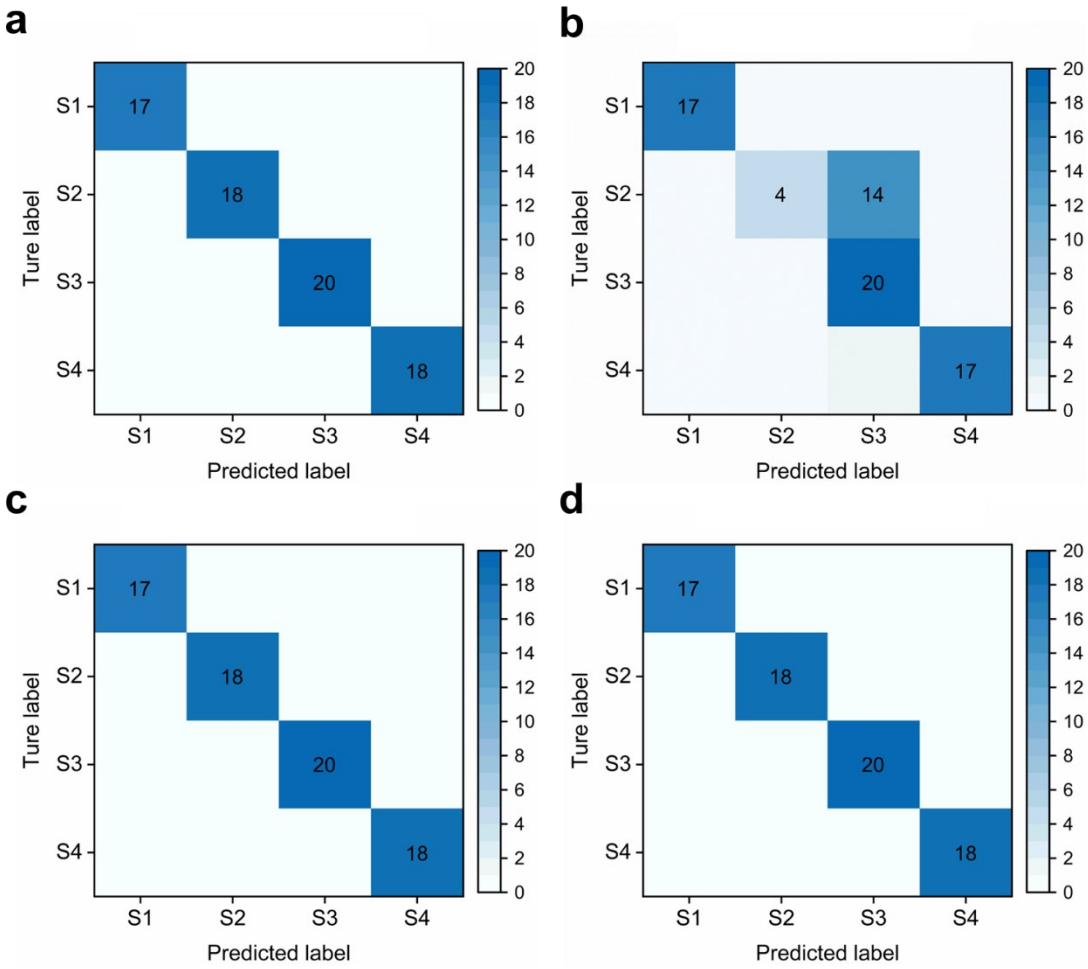

Fig. S10 The confusion matrix of the ANN model for *Staphylococcus* bacterial species. **a-d** The confusion matrix of *Staphylococcus* bacterial subspecies is based on MALDI-TOF MS spectral, Raman spectral, prototypical spectra fusion (PSF) and feature-extractor-based fusion (FEBF) dataset, respectively.

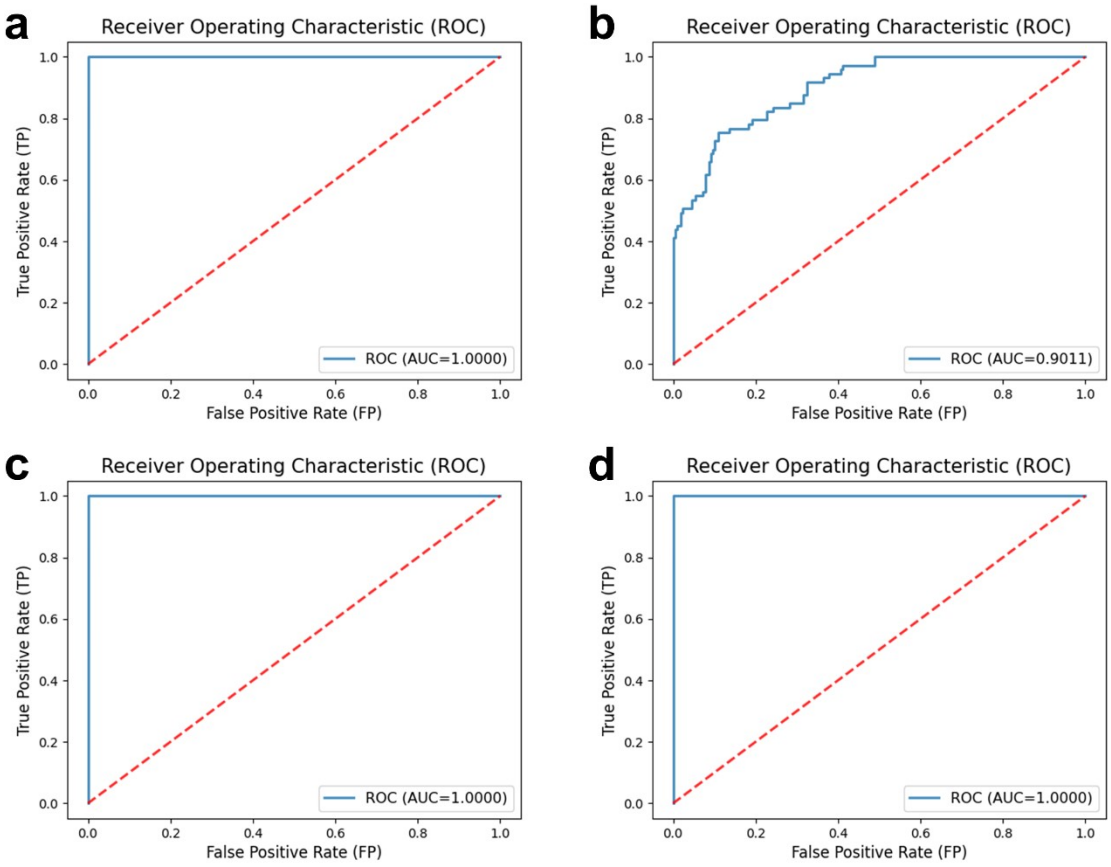

Fig. S11 The receiver operating characteristic curves (ROC) analysis of the ANN model for *Staphylococcus* bacterial species. **a-d** The ROC of *Staphylococcus* bacterial subspecies is based on MALDI-TOF MS spectral, Raman spectral, PSF and FEBF dataset, respectively.

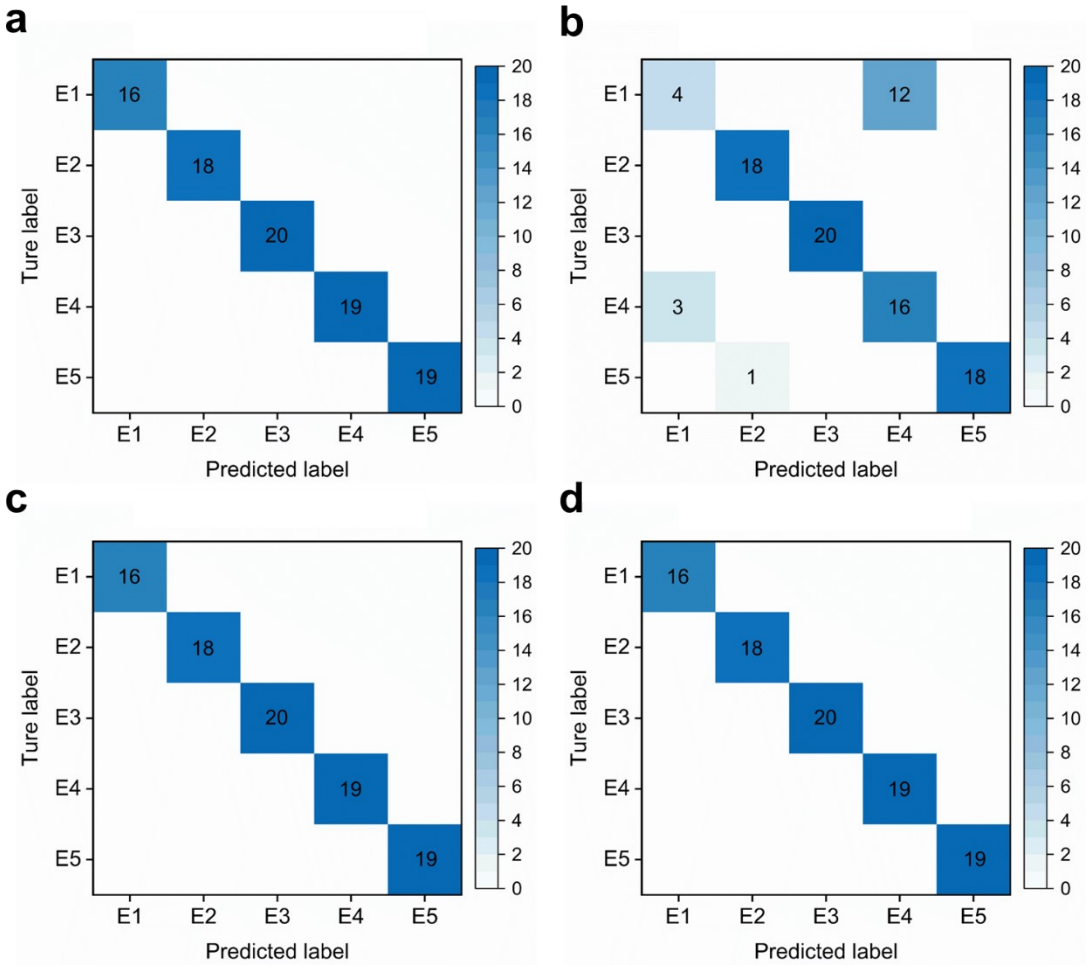

Fig. S12 The confusion matrix of the ANN model for *E. coli* bacterial species. **a-d** The confusion matrix of *E. coli* bacterial subspecies is based on MALDI-TOF MS spectral, Raman spectral, PSF and FEBF dataset, respectively.

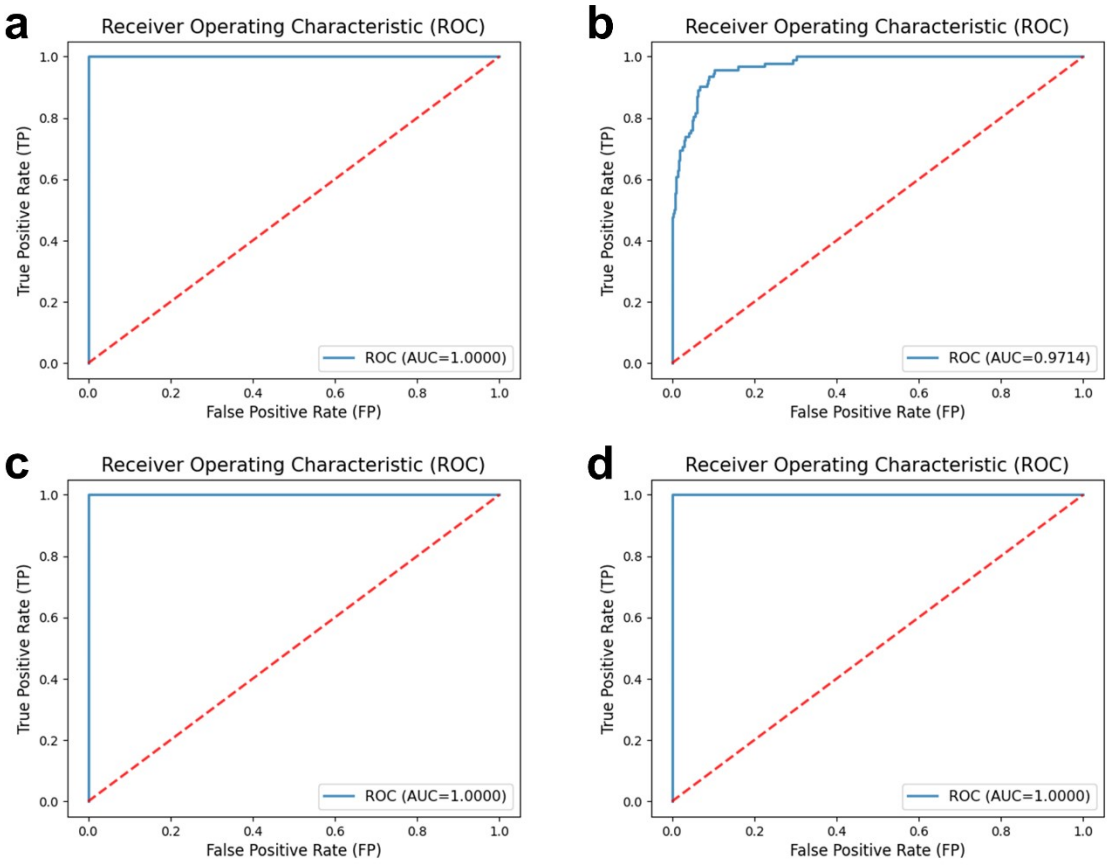

Fig. S13 The ROC of the ANN model for *E. coli* bacterial species. **a-d** The ROC of *E.coli* bacterial subspecies is based on MALDI-TOF MS spectral, Raman spectral, PSF and FEBF dataset, respectively.

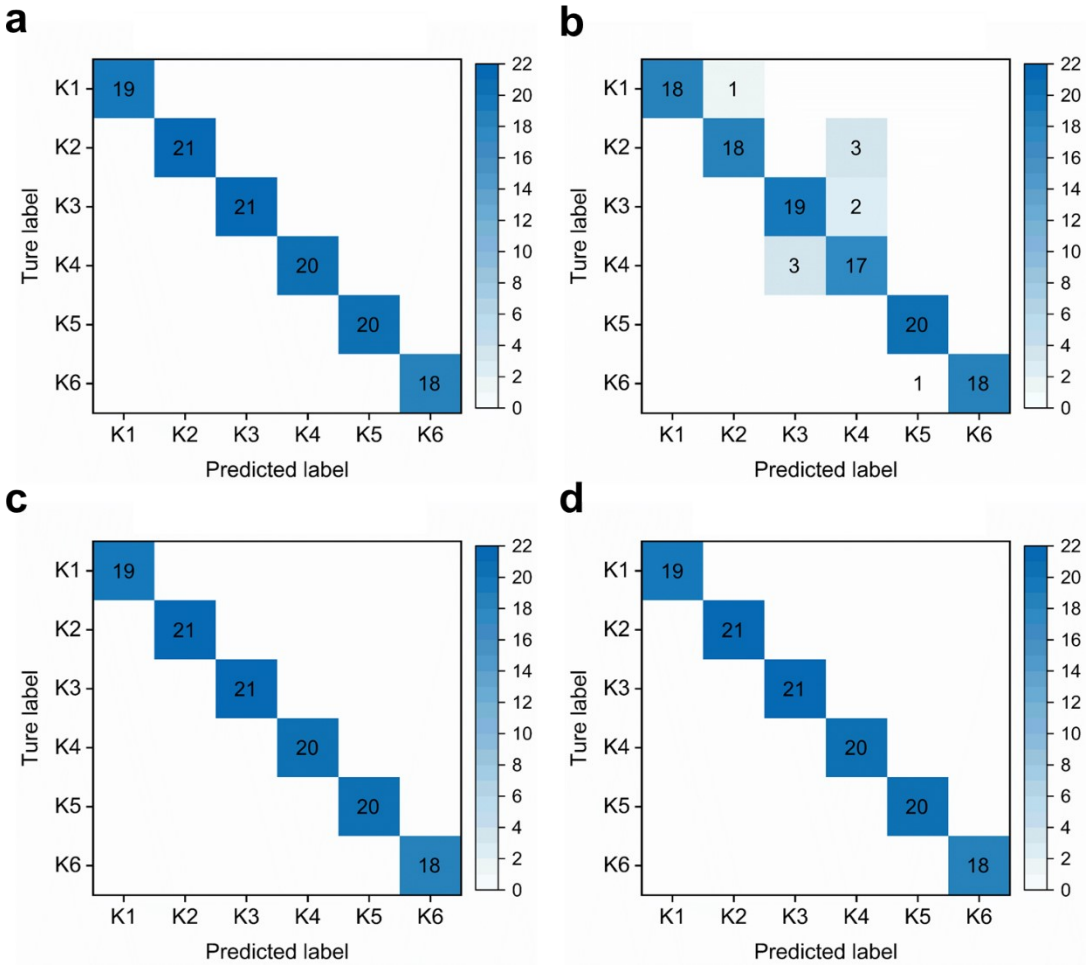

Fig. S14 The confusion matrix of the ANN model for *K. pneumoniae* bacterial subspecies. **a-d** The confusion matrix of *K. pneumoniae* bacterial subspecies is based on MALDI-TOF MS spectral, Raman spectral, PSF and FEBF dataset, respectively.

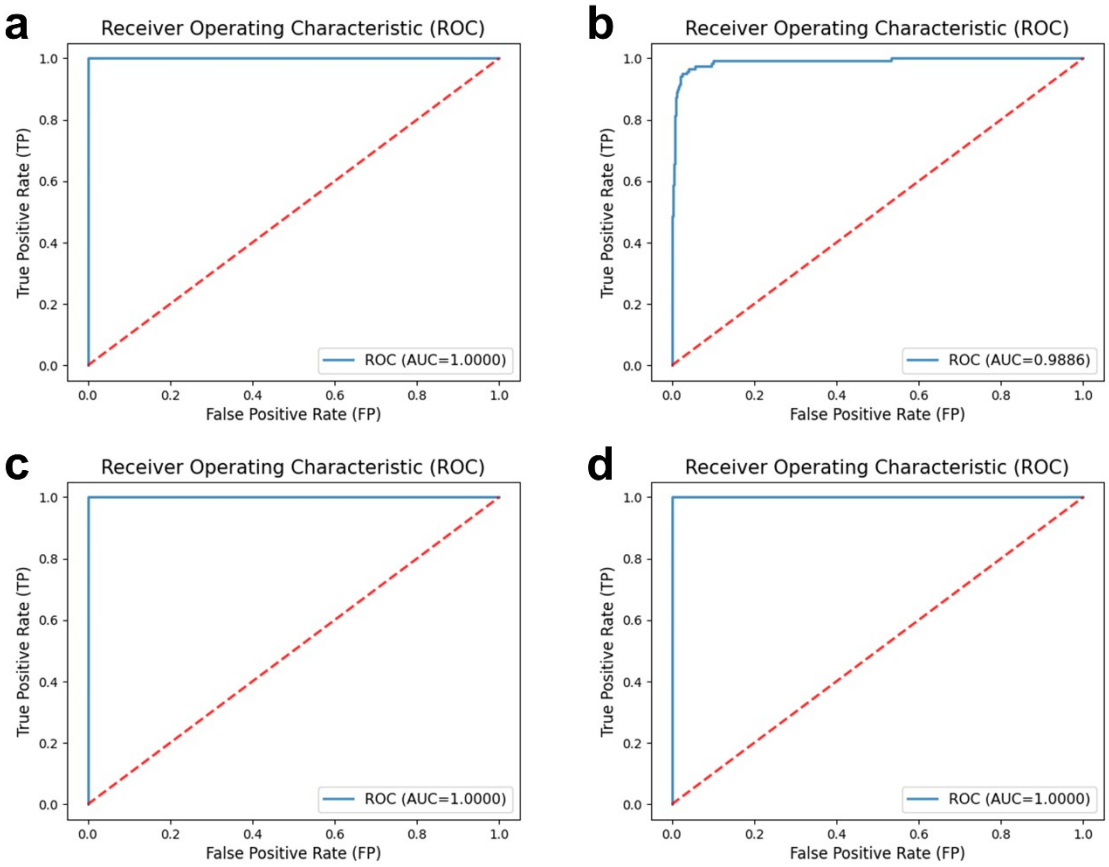

Fig. S15 The ROC of the ANN model for *K. pneumoniae* bacterial subspecies. **a-d** The ROC of *K. pneumoniae* bacterial subspecies is based on MALDI-TOF MS spectral, Raman spectral, PSF and FEBF dataset, respectively.

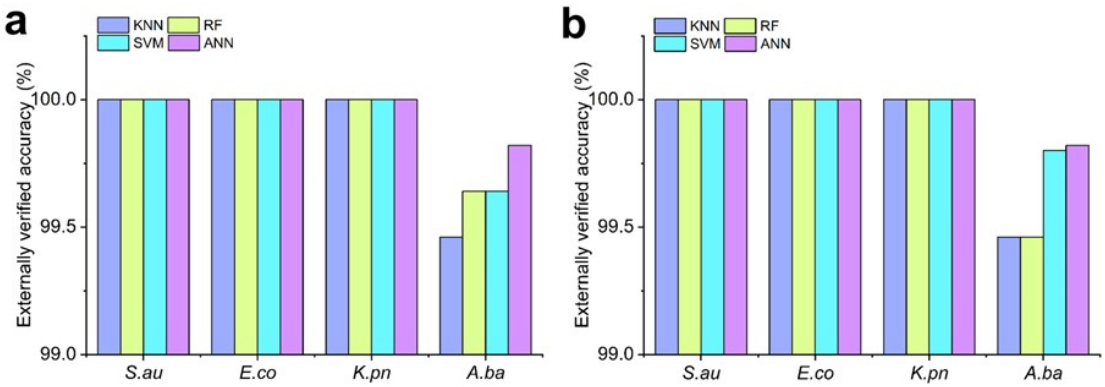

Fig. S16 a Ten-fold cross-validation accuracy of four commonly used machine methods applied to different bacteria isolates based on FEBF dataset. **b** Leave-one-out accuracy of four commonly used machine methods applied to different bacteria isolates based on FEBF dataset.

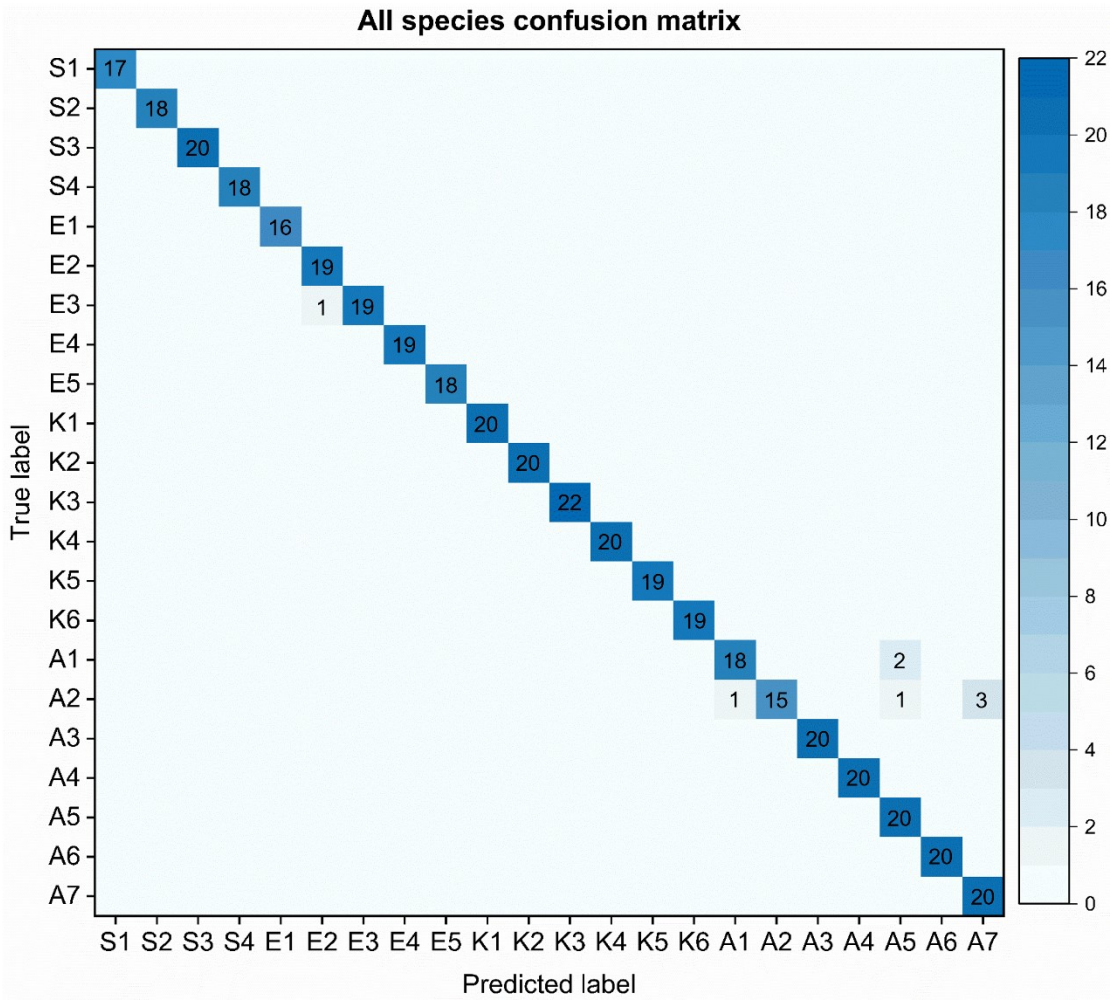

Fig. S17 The confusion matrix of all bacterial species based on corresponding prototypical spectra fusion dataset by ANN method.

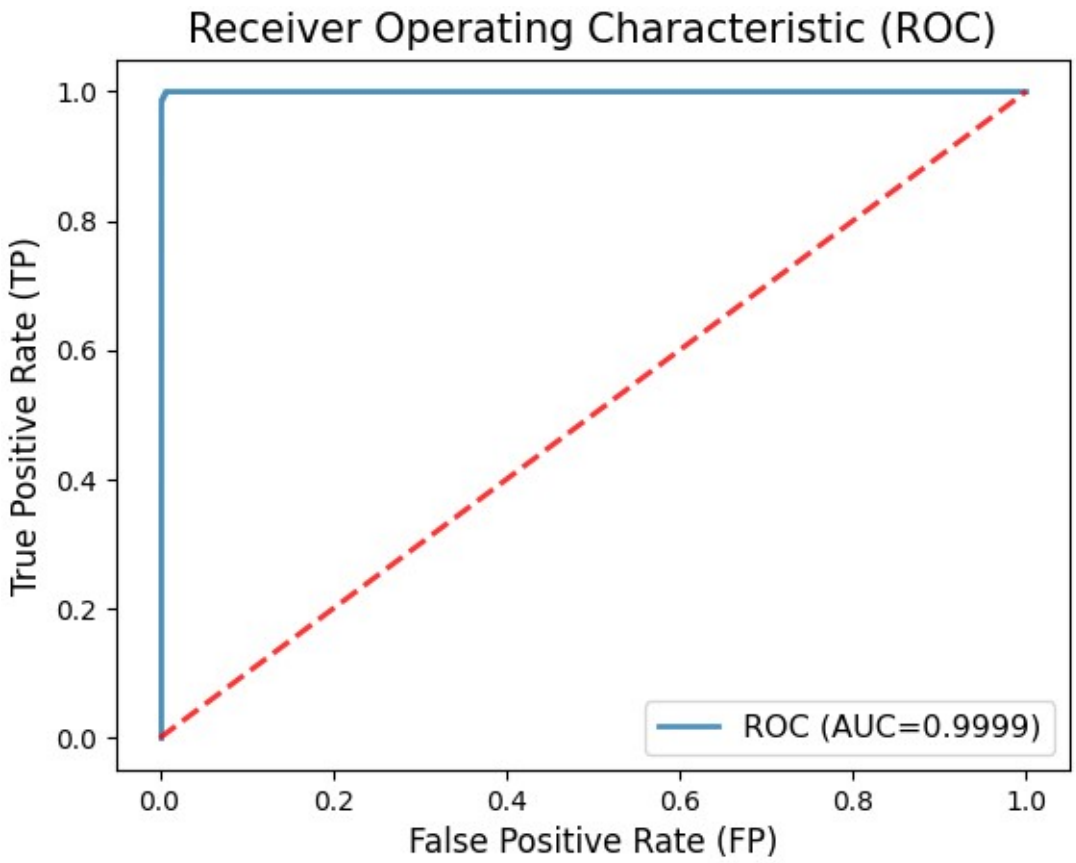

Fig. S18 The ROC of all bacterial species based on corresponding prototypical spectra fusion dataset by ANN method.

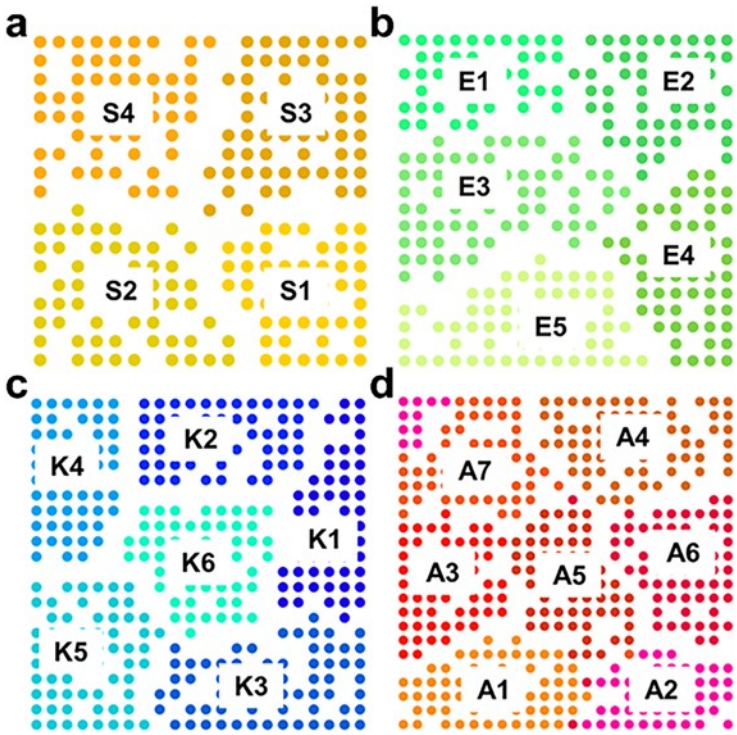

Fig. S19 SOM plot showing the clusters and clustering relationships of the differential feature data. **a, b** SOM plots of the *Staphylococcus* and *E. coli* species isolates. **c, d** SOM plots of the *K. pneumoniae* and *A. baumannii* subspecies isolates.

Table S1. The number of MALDI-TOF MS or Raman spectra (n) for each isolate.

Staphylococcus	**n**	***E. coli***	**n**	***K. pneumoniae***	**n**	***A. baumannii***	**n**
S1	69	E1	64	K1	78	A1	80
S2	74	E2	75	K2	82	A2	80
S3	78	E3	80	K3	87	A3	80
S4	72	E4	76	K4	79	A4	80
		E5	74	K5	78	A5	80
				K6	74	A6	80
						A7	80

Table S2. The number of FEBF spectra (n) for each isolate in the training set.

Staphylococcus	**n**	***E. coli***	**n**	***K. pneumoniae***	**n**	***A. baumannii***	**n**
S1	52	E1	48	K1	59	A1	60
S2	56	E2	57	K2	61	A2	60
S3	58	E3	60	K3	66	A3	60
S4	54	E4	57	K4	59	A4	60
		E5	55	K5	58	A5	60
				K6	56	A6	60
						A7	60

Table S3. The number of FEBF spectra (n) for each isolate in the testing set.

Staphylococcus	**n**	***E. coli***	**n**	***K. pneumoniae***	**n**	***A. baumannii***	**n**
S1	17	E1	16	K1	19	A1	20
S2	18	E2	18	K2	21	A2	20
S3	20	E3	20	K3	21	A3	20
S4	18	E4	19	K4	20	A4	20
		E5	19	K5	20	A5	20
				K6	18	A6	20
						A7	20

Table S4. Assignments of the major Raman peaks of four different species.

Staphylococcus **(cm^{-1})**	***E. coli*** **(cm^{-1})**	***K. pneumoniae*** **(cm^{-1})**	***A. banumannii*** **(cm^{-1})**	**Major peak assignment**	**Ref.**
620	620	621		Phenylalanine (skelet al)	6
643	640	644	648	Tyrosine	6
673	670	673	672	T, G(DNA/RNA)	7
724	723	725		adenine	6
746	748	746	744	Adenine (Nucleic acids)	8
778	783	783	781	Cytosine, uracil (ring, str)	6
	828	827	820	“exposed” Tyrosine	6
856	854	855	850	CC str, COC 1,4 glycosidic link	6
1001	1003	1000	1002	Phenylalanine	9
1080	1078	1077	1070	C–C or C–O str (lipid) C–C or PO2 str (nucleic acids)	10
1128	1126	1126	1125	Proteins: stretching C-N; Carbohydrates: str C-O	9
	1154	1154	1160	Lipids and nucleic acids (cytosine, guanine, adenine)	10
1229	1226	1225	1222	C-N and C-C str	6
	1245		1238	Amide III (β-Sheet)	11
1337	1339	1337	1336	Proteins: twisting (CH2, CH3)	11
	1364	1361	1360	Pyrimidineand imidazole rings (Nuleic acids)	8
1444	1449	1447	1445	C-H2 def	6
	1619	1620	1617	tyrosine	6
1666	1667	1666		Amide I	6
1732	1732	1732		>C=O ester str	6

[a] str=stretching; def=deformation; sym=symmetric; asym=antisymmetric.

Reference:

1. L. Feng, B. Wu, S. Zhu, J. Wang, Z. Su, F. Liu, Y. He and C. Zhang, *Front. Plant Sci.*, 2020, **11**, 577063.
2. L. Li, S. Zhang, Z. T. Zuo and Y. Z. Wang, *Journal of Chemometrics*, 2022, **36**, e3436.
3. F. Hallouche, A. E. Adams, O. R. Hinton, D. P. Surtees, V. Wadehra and G. V. Sherbet, *Analytical and quantitative cytology and histology*, 1993, **15**, 50-60.
4. J. Yang and J. Y. Yang, *Pattern Recognition*, 2003, **36**, 563-566.
5. P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, *IEEE Trans. Pattern Anal. Mach. Intell.*, 1997, **19**, 711-720.
6. K. Maquelin, C. Kirschner, L. P. Choo-Smith, N. van den Braak, H. P. Endtz, D. Naumann and G. J. Puppels, *Journal of microbiological methods*, 2002, **51**, 255-271.
7. J. C. Fraire, S. Stremersch, D. Bouckaert, T. Monteyne, T. De Beer, P. Wuytens, R. De Rycke, A. G. Skirtach, K. Raemdonck, S. De Smedt and K. Braeckmans, *ACS applied materials & interfaces*, 2019, **11**, 39424-39435.
8. S. G. Kruglik, F. Royo, J.-M. Guigner, L. Palomo, O. Seksek, P.-Y. Turpin, I. Tatischeff and J. M. Falcón-Pérez, *Nanoscale*, 2019, **11**, 1661-1679.
9. A. Rygula, K. Majzner, K. M. Marzec, A. Kaczor, M. Pilarczyk and M. Baranska, *J Raman Spectrosc*, 2013, **44**, 1061-1076.
10. M. Moreno, L. Raniero, E. Â. L. Arisawa, A. M. do Espírito Santo, E. A. P. dos Santos, R. A. Bitar and A. A. Martin, *Theor Chem Acc*, 2010, **125**, 329-334.
11. V. Shalabaeva, L. Lovato, R. La Rocca, G. C. Messina, M. Dipalo, E. Miele, M. Perrone, F. Gentile and F. De Angelis, *PLoS One*, 2017, **12**, e0175581.