iScience, Volume *24*

Supplemental information

Machine learning-assisted single-cell Raman fingerprinting for *in situ* and nondestructive classification of prokaryotes

Nanako Kanno, Shingo Kato, Moriya Ohkuma, Motomu Matsui, Wataru Iwasaki, and Shinsuke Shigeto

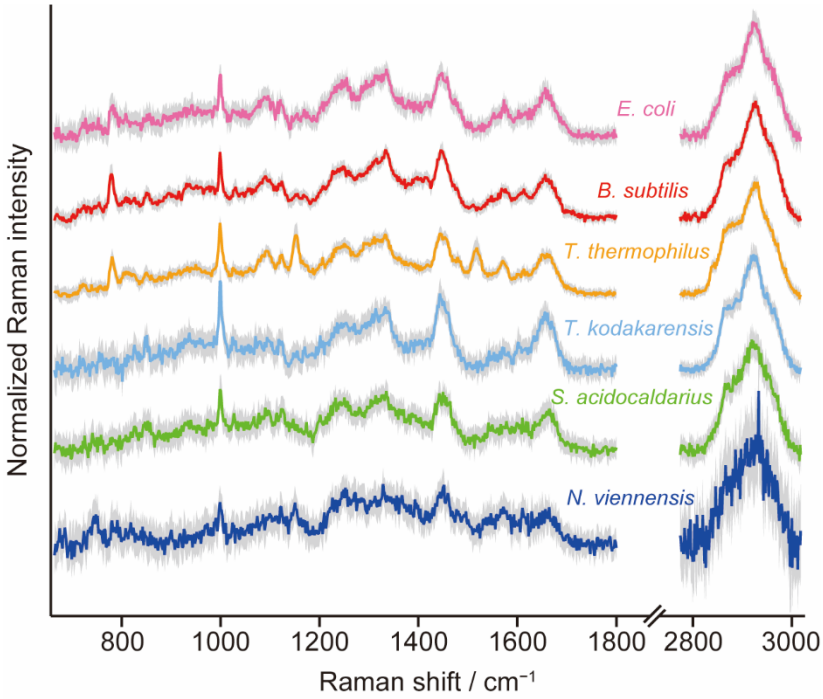


Figure S1. Average of the preprocessed Raman spectra for each of the six prokaryotic species, related to Figure 2.

The shaded area represents the 1σ error envelop.

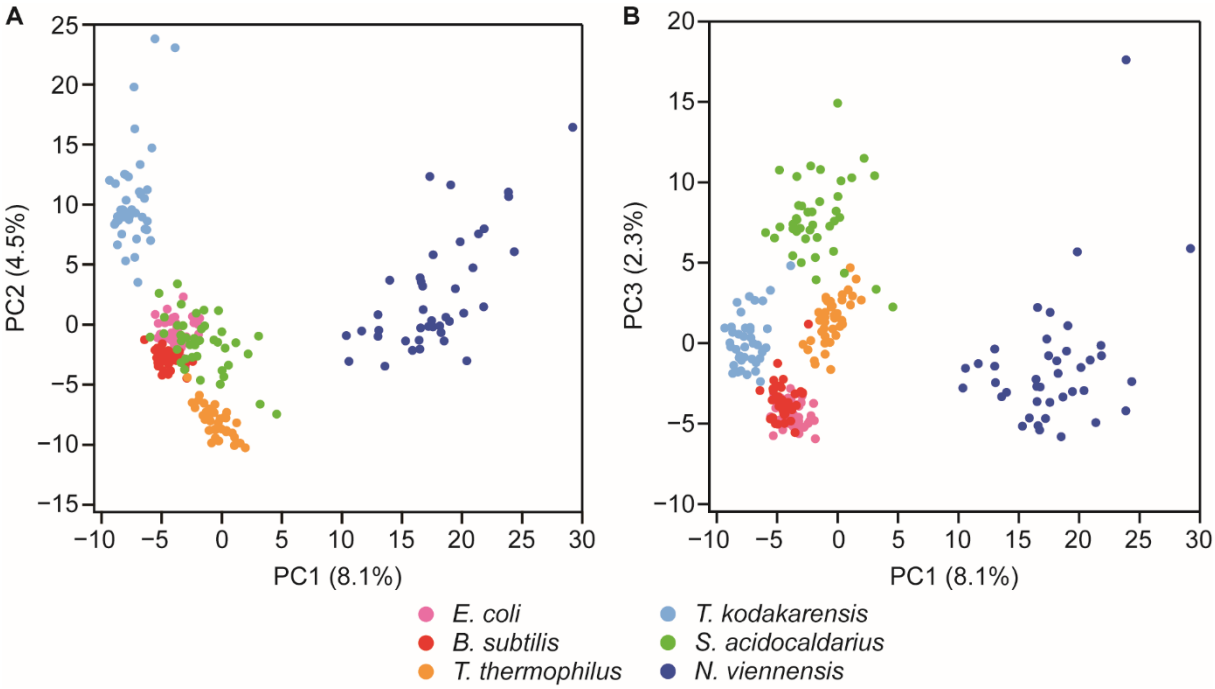


Figure S2. PCA score plots for the six species, related to Figure 2.

(A) PC1 vs. PC2.

(B) PC1 vs. PC3.

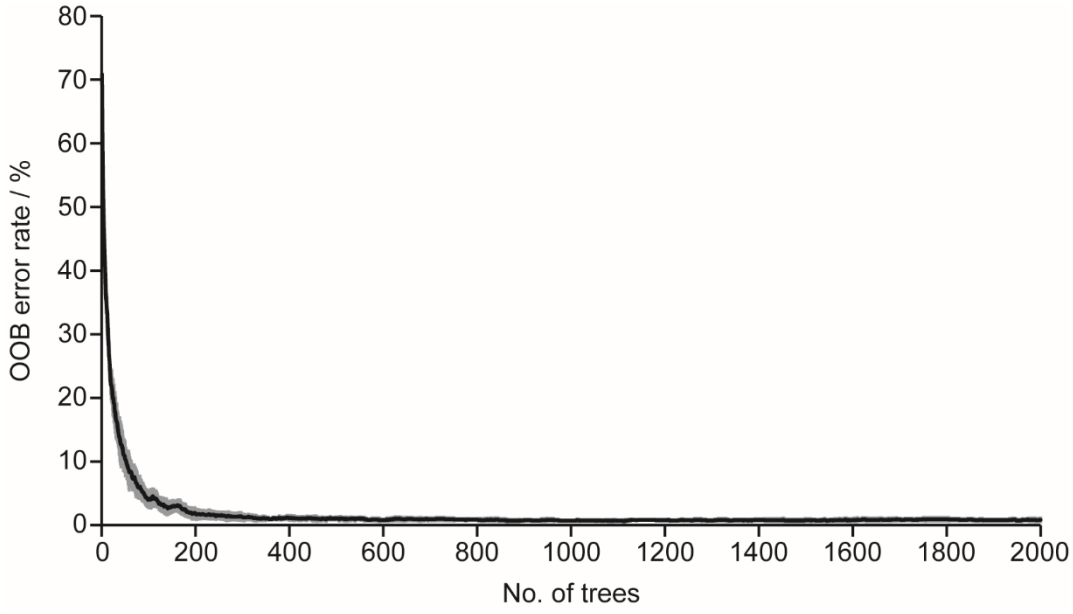


Figure S3. Out-of-bag (OOB) error rate of the random forest machine learning model for classifying the six prokaryotic species, related to Figure 2.

The black line shows the mean error rate and the gray part shows standard deviation across 10 training and validation splits.

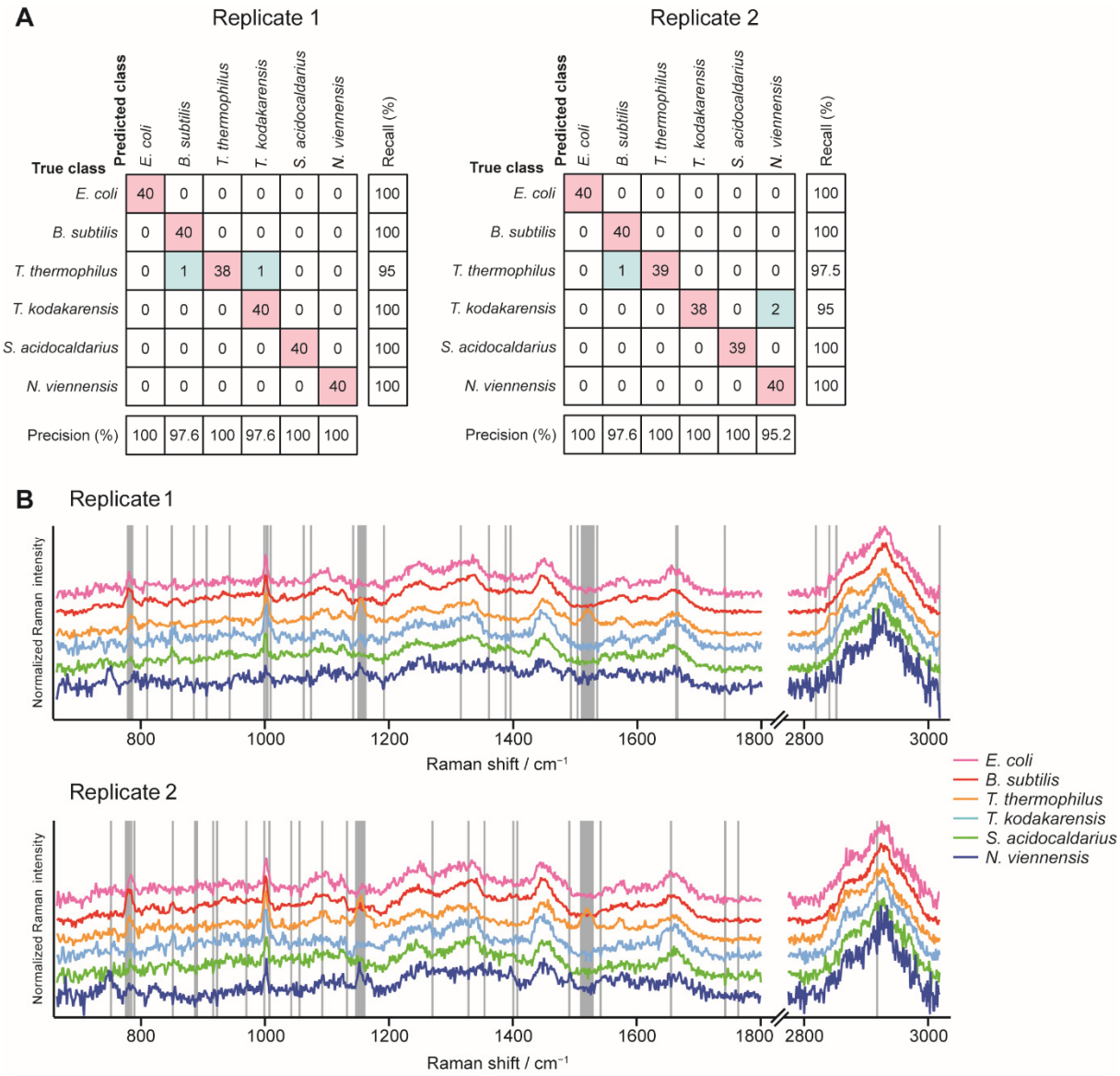


Figure S4. RF classification of the Raman datasets obtained from the other two independent batches of the six prokaryotic species (denoted replicates 1 and 2 here) than those shown in the main text, related to Figure 2.

(A) Confusion matrices. The overall validation accuracy was 99.2±1.7% for replicate 1 and 98.8±1.9% for replicate 2.

(B) Averaged Raman spectra of the six species, together with top 50 most important features (vertical lines).

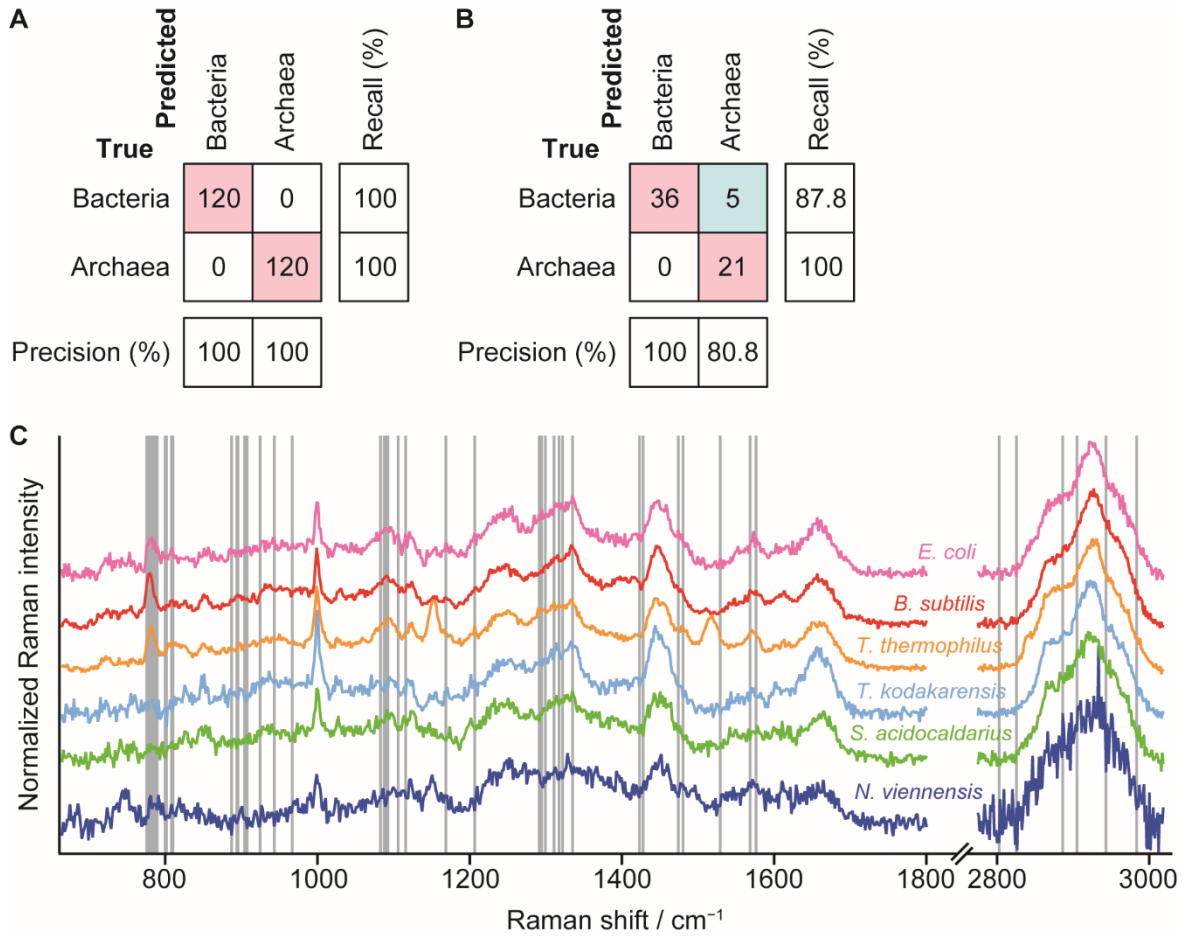


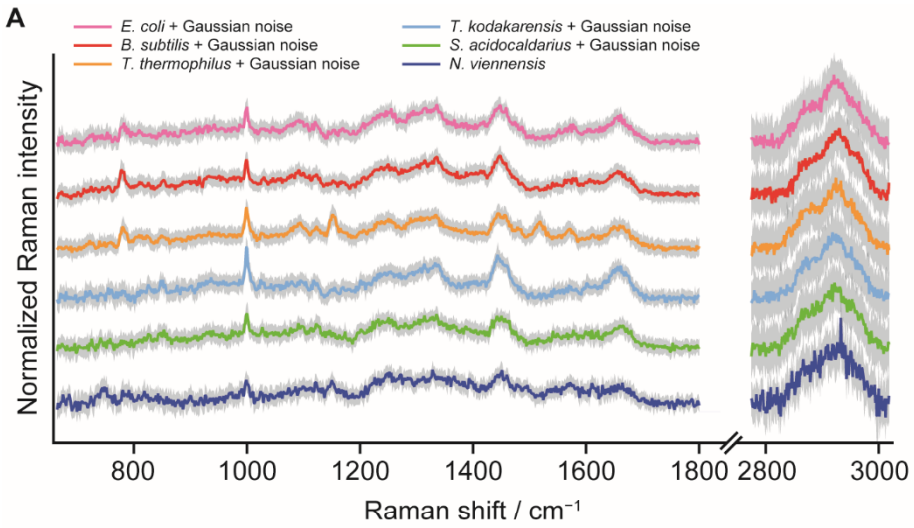
Figure S5. Bacteria–archaea binary classification, related to Figure 2.

The three bacterial species (*E. coli*, *B. subtilis*, and *T. thermophilus*) are labeled "Bacteria" and the three archaeal species (*T. kodakarensis*, *S. acidocaldarius*, and *N. viennesis*) "Archaea". The optimized numbers of trees (n_estimators) and features (max_features) were 200 and 29, respectively.

(A) Confusion matrix, **C**, for two classes in the model construction. Each entry of the confusion matrix, C_{ij}, represents the total number of spectra known to be in class *i* and predicted by the RF model to be in class *j* in 10-fold cross-validation. Correct classification results are shown in red boxes on the diagonal. The validation accuracy was 100%. The precision and recall rates are also shown in percentage.

(B) Confusion matrix, **D**, for two classes in the mixed population. Each entry of the confusion matrix, D_{ij}, represents the number of spectra known to be in class *i* and predicted by the RF model to be in class *j*. Correct classification results are shown in red boxes on the diagonal, and misclassification results in a blue box. The overall accuracy was 91.9%. The precision and recall rates are also shown in percentage.

(C) Top 50 most important features extracted from the two-class classification result are shown as vertical lines on the averaged spectra.



B

True \ Predicted	*E. coli*	*B. subtilis*	*T. thermophilus*	*T. kodakaraensis*	*S. acidocaldarius*	*N. viennensis*	Recall (%)
E. coli	25	13	0	0	2	0	62.5
B. subtilis	13	25	0	0	2	0	62.5
T. thermophilus	0	1	39	0	0	0	97.5
T. kodakaraensis	0	0	0	40	0	0	100
S. acidocaldarius	1	0	0	1	38	0	95.0
N. viennensis	0	0	0	0	0	40	100
Precision (%)	64.1	64.1	100	97.6	90.5	100	

C

True \ Predicted	*E. coli*	*B. subtilis*	*T. thermophilus*	*T. kodakaraensis*	*S. acidocaldarius*	*N. viennensis*	Recall (%)
B. subtilis	2	16	0	0	3	0	76.2
T. thermophilus	0	0	20	0	0	0	100
N. viennensis	0	0	0	0	3	18	85.7
Precision (%)	-	100	100	-	-	100	

Figure S6. Effect of artificial addition of Gaussian noise on the random forest classification results, related to Figures 2 and 3.

The optimized numbers of trees (n_estimators) and features was (max_features) were 700 and 50, respectively.

(A) Average of the noise-added Raman spectra for each of the five prokaryotic species other than *N. viennensis*. No noise was added to the *N. viennensis* spectra. The shaded area represents the 1σ error envelope.

(B) Confusion matrix, **C**, for six classes in the model construction. Each entry of the confusion matrix, C_{ij}, represents the total number of spectra known to be in class *i* and predicted by the RF model to be in class *j* in 10-fold cross-validation. Correct classification results are shown in red boxes on the diagonal, and misclassification results in blue boxes. The validation accuracy was 86.3±7.0%. The precision and recall rates are also shown in percentage.

(C) Performance breakdown for the mixed population. Correct classification results are shown in red boxes, and misclassification results in blue boxes. The values in the table represent the number of spectra (cells). The precision and recall rates are shown in percentage. The overall accuracy was 87.1%.

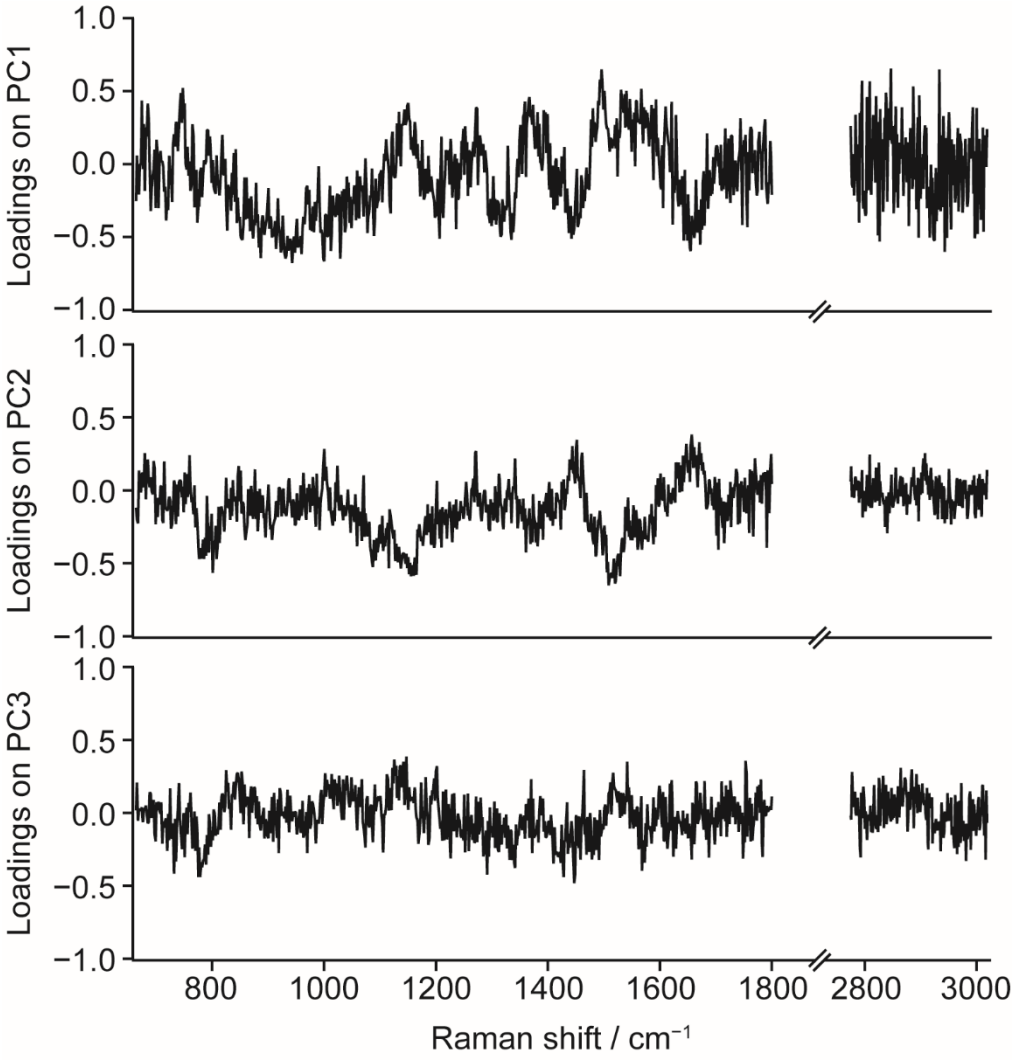


Figure S7. Loading spectra of the first three principal components (PC1, PC2, and PC3) derived from the PCA of the single-cell Raman dataset of the six prokaryotic species, related to Figure 2.

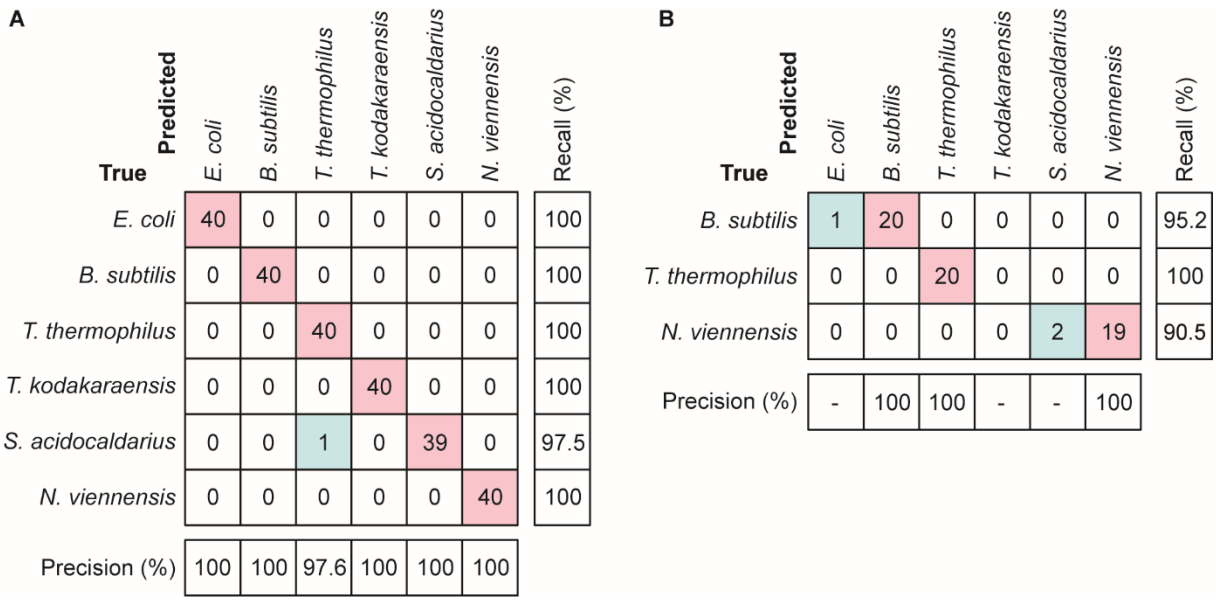


Figure S8. Random forest species classification using Raman spectra on which noise reduction with singular value decomposition (SVD) was performed, related to Figures 2 and 3.

(A,B) Performance breakdown. SVD components with negligibly small singular values were discarded as noise, and only the remaining components were used to reconstruct denoised data. Excessive denoising may lead to an increase in the similarity of noise patterns among the spectra, which may affect machine learning modeling. To avoid this, 30% of total SVD components were retained. The optimized numbers of trees (n_estimators) and features (max_features) were 300 and 29, respectively. The validation accuracy in (A) was 99.6±1.3%, and the overall accuracy in (B) was 95.2%.

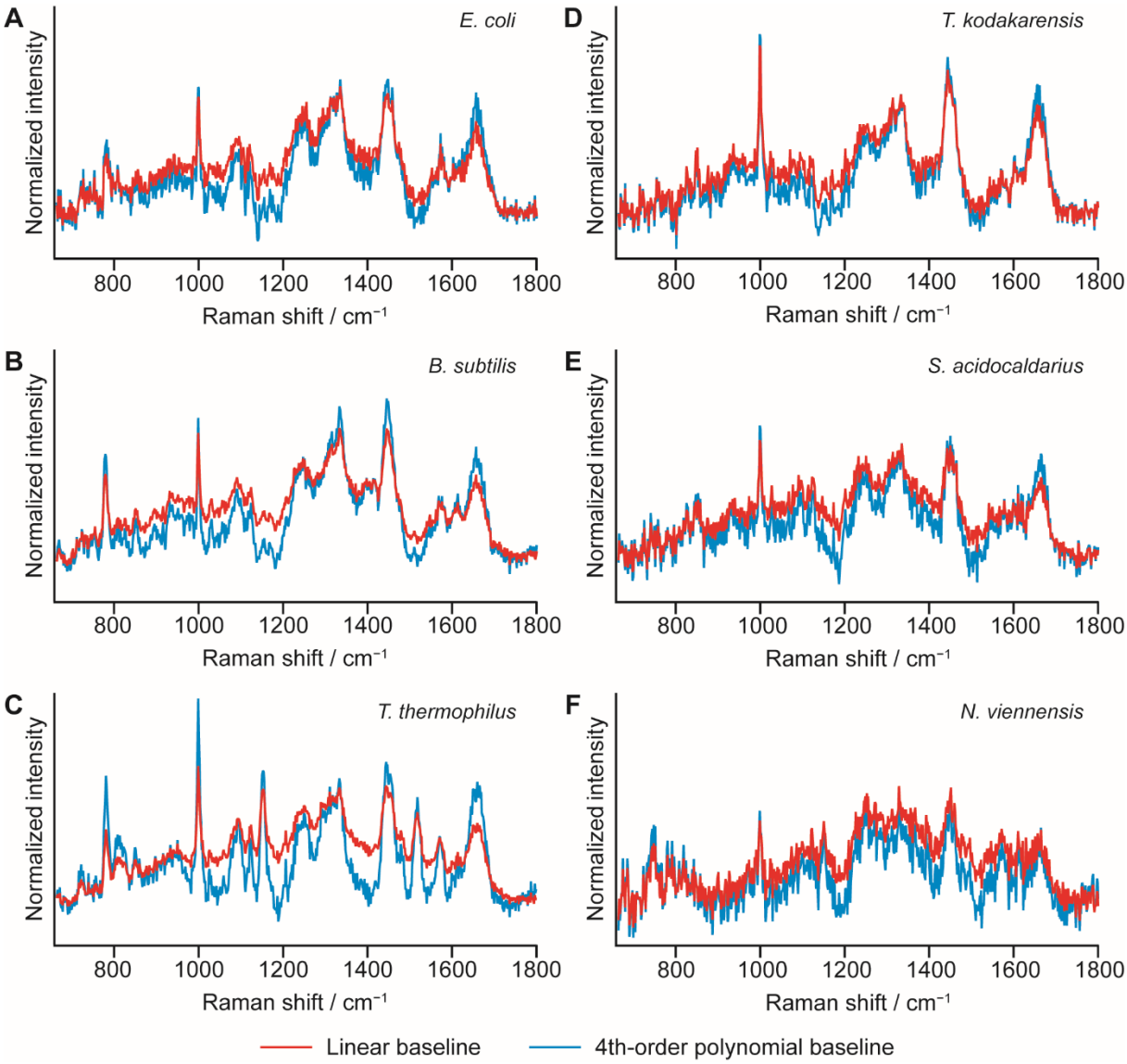


Figure S9. Averaged Raman spectra with linear (red line) and fourth-order polynomial (blue line) baseline subtracted, of the six prokaryotic species, related to Figure 1.

(A) *E. coli*

(B) *B. subtilis*

(C) *T. thermophilus*

(D) *T. kodakarensis*

(E) *S. acidocaldarius*

(F) *N. viennensis*

Table S1. Full list of the top 50 most important features, related to Figure 2 and Table 2.

Ranking	Importance[a]	Wavenumber (cm^{-1})	Assignment	Major molecular components
32	35.9	775.7	C, T, U, backbone (P-O-P)	DNA/RNA
11	65.9	777.4	C, T, U, backbone (P-O-P)	DNA/RNA
9	69.4	779.1	C, T, U, backbone (P-O-P)	DNA/RNA
35	34.1	780.9	C, T, U, backbone (P-O-P)	DNA/RNA
8	71.2	782.6	C, T, U, backbone (P-O-P)	DNA/RNA
48	29.8	786.0	C, T, U, backbone (P-O-P)	DNA/RNA
25	41.0	801.6		
29	39.1	931.5		
6	76.0	999.0	Phenylalanine ring-breathing	Proteins
1	100.0	1000.7	Phenylalanine ring-breathing	Proteins
34	35.0	1111.0		
36	34.1	1137.5	C–C str.	Carotenoids
46	30.0	1140.8	C–C str.	Carotenoids
42	31.2	1142.4	C–C str.	Carotenoids
49	29.5	1145.7	C–C str.	Carotenoids
30	38.6	1147.4	C–C str.	Carotenoids
22	42.9	1149.1	C–C str.	Carotenoids
5	77.2	1150.7	C–C str.	Carotenoids
13	56.2	1152.4	C–C str.	Carotenoids
14	54.4	1154.0	C–C str.	Carotenoids
10	69.1	1155.7	C–C str.	Carotenoids
24	41.1	1157.3	C–C str.	Carotenoids
37	33.8	1360.9		
26	40.9	1370.5		
44	30.4	1381.7		
50	29.2	1391.3		
33	35.2	1405.7		
39	33.4	1413.7		
27	40.7	1447.1	CH bend.	Proteins/lipids
45	30.1	1496.2		
23	42.0	1508.8	C=C str.	Carotenoids
16	51.7	1510.4	C=C str.	Carotenoids
28	39.9	1512.0	C=C str.	Carotenoids

4	83.7	1513.6	C=C str.	Carotenoids
7	71.4	1515.1	C=C str.	Carotenoids
2	90.9	1516.7	C=C str.	Carotenoids
3	85.2	1518.3	C=C str.	Carotenoids
12	65.3	1519.8	C=C str.	Carotenoids
15	51.7	1521.4	C=C str.	Carotenoids
41	32.3	1523.0	C=C str.	Carotenoids
19	45.7	1524.6	C=C str.	Carotenoids
31	37.0	1526.1	C=C str.	Carotenoids
20	45.4	1527.7	C=C str.	Carotenoids
38	33.4	1568.4	G, A	DNA/RNA
43	30.7	1655.4	Amide I	Proteins
40	33.2	1661.5	Amide I	Proteins
18	46.6	2844.3	CH_2 sym. str.	Lipids
17	47.3	2846.8	CH_2 sym. str.	Lipids
47	30.0	2943.3	CH str.	Proteins/lipids
21	44.8	2946.9	CH str.	Proteins/lipids

[a]The highest importance is scaled to 100.

Abbreviations: A, adenine; G, guanine; C, cytosine; T, thymine; U, uracil; str., stretching; bend., bending; sym., symmetric.