Supplementary Information for:

Functional Regression for SERS Spectrum Transformation Across Diverse Instruments

Tao Wang [1#], Yanjun Yang [2#], Haoran Lu [1], Jiaheng Cui [3], Xianyan Chen [4], Ping Ma [1*], Wenxuan Zhong [1*] and Yiping Zhao [2*]

[1] Department of Statistics, Franklin College of Arts and Sciences, University of Georgia, Athens, Georgia 30602, United States

[2] Department of Physics and Astronomy, Franklin College of Arts and Sciences, University of Georgia, Athens, Georgia 30602, United States

[3] School of Electrical and Computer Engineering, College of Engineering, The University of Georgia, Athens, GA 30602, United States

[4] Department of Epidemiology & Biostatistics, College of Public Health, The University of Georgia, Athens, GA 30602**,** United States

[#] These authors contribute equally to this work

* Corresponding Authors: E-mail: pingma@uga.edu; wenxuan@uga.edu; zhaoy@uga.edu

S1. Silver nanorod (AgNR) substrate fabrication.

AgNR arrays prepared by the oblique angle deposition (OAD) are excellent SERS substrates as reported previously.[1-4] Briefly, clean glass slides (0.5 inch × 0.5 inch) were loaded into a vacuum deposition chamber with the substrate normal antiparallel to the incident vapor direction. A layer of 20 nm-thick Ti film (Titanium pellets, Kurt J. Lesker, 99.995%) and a layer of 200 nm-thick Ag film Silver (Silver pellets, Kurt J. Lesker, 99.99%) were deposited in sequence at a rate of 0.2 nm/s and 0.3 nm/s, respectively. Then, the substrate normal was rotated to 86° relative to the incident vapor direction, and a thickness of 2000 nm Ag film was deposited at a rate of 0.3 nm/s to obtain the AgNR arrays. The entire evaporation process was conducted under a high vacuum condition with a pressure $< 3 \times 10^{-6}$ Torr. A typical SEM image of an AgNR substrate is shown in **Figure S1**, and a detailed deposition procedure and condition can be found in Song et al.[5] According to previous extensive studies, the AgNR substrates have been demonstrated to possess good SERS reproducibility with <10% relative standard deviation (RSD), high SERS enhancement factor up to 10^9, and large area uniformity.[1, 2, 6, 7]

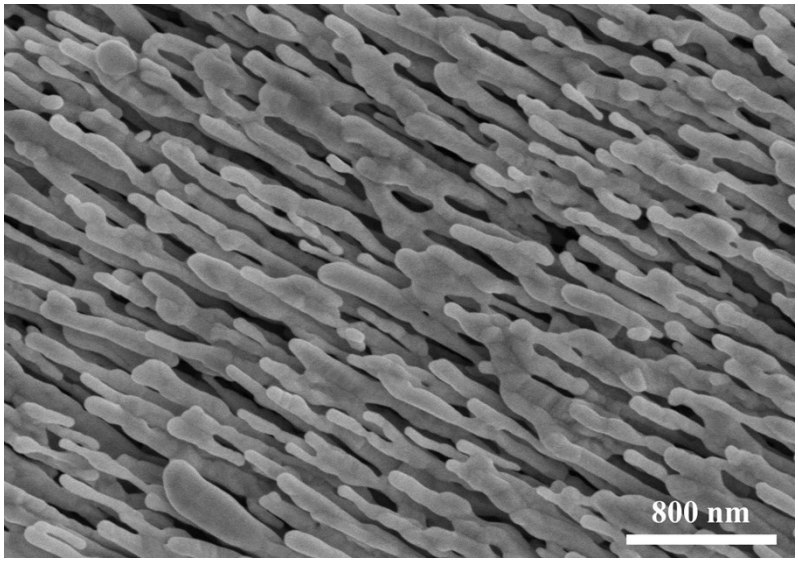

Figure S1. SEM image of AgNR array SERS substrate.

S2. The sources of bacterial samples.

Table S1 lists the bacterial biomarkers, endotoxins, and references for the SERS spectra collection. Seven bacteria biomarkers and two references were purchased from Sigma-Aldrich, including

lipoteichoic acid from *Staphylococcus aureus*, enterobactin from *Escherichia coli*, pyocyanin from *Pseudomonas aeruginosa*, β-Carotene, 2,5-Dihydroxybenzoic acid, 2,3-Dihydroxybenzoic acid, peptidoglycan (PGN) from *S. aureus*, vancomycin hydrochloride, and polymyxin B.

The lipopolysaccharides (LPSs) were extracted from the bacterial cells by hot phenol-water extraction procedure.[8] The water phases were dialyzed (12-14 kDa cutoff membrane), freeze-dried, washed in 90% EtOH to remove traces of phospholipids. The detailed purifications of specific LPS samples are in corresponding references: *Francisella tularensis* LVS;[9] LOS *Moraxella catarrhalis*;[10] *Pseudomonas aeruginosa*,[11] *S. meliloti*. The endotoxin of *Salmonella enterica* serovar, Typhimurium (S-type LPS), *Salmonella enterica* serovar, Minnesota Re595 (R-type LPS; Re), *E.coli*-EH100 (R-type LPS; Ra), LPS *E.coli*-O128:B12, *E.coli*-O11:B4, *E.coli*-J5 (R-type LPS; Rc) were obtained from Sigma Aldrich, and the LPS of *Helicobacter pylori* GU2 was obtained from Wako Pure Chemicals (Tokyo, Japan).

Table S1. Biomarkers and their bacterial information.

No.	Name	Bacteria	Short name or notes
1	lipoteichoic acid	*S. aureus*	LTA
2	enterobactin	*E. coli*	ENT siderophore
3	pyocyanin	*P. aeruginosa*	PCN
4	β-Carotene	multiple bacteria including mycobacteria	carotenoid pigment
5	2,5-Dihydroxybenzoic acid	mammalian siderophore	2,5-DHBA
6	2,3-Dihydroxybenzoic acid	bacterial siderophore	2,3-DHBA
7	vancomycin hydrochloride		(reference)
8	polymyxin B		(reference)
9	lipopolysaccharide	*P. aeruginosa*	
10	lipopolysaccharide	*F. tularensis* LVS	
11	lipopolysaccharide	*H. pylori* GU2	
12	lipopolysaccharide	*E. coli*-EH100	
13	lipopolysaccharide	*E. coli*-O11:B4	

14	lipopolysaccharide	*E. coli*-J5	
15	lipopolysaccharide	*S. meliloti* Rm1021	
16	peptidoglycan	*B. subtilis* (PGN)	
17	lipopolysaccharide	*M. catarrhalis*	
18	lipopolysaccharide	*E. coli*-O128:B12	
19	lipopolysaccharide	*S. minnesota* Re595	
20	lipopolysaccharide	*S. enterica serovar Typhimurium*	

S3. Raman instrument information.

Table S2. Specifications of five Raman spectrometers.

Parameter	Renishaw	Tec5	First defender	Rapid ID
Company	Renishaw	Tec5	Thermo Fisher Scientific	Wasatch Photonics
Laser wavelength (nm)	785	785	785	785
Laser power (mW)	Adjustable	Adjustable, 30 – 500 mW	Adjustable, 75mW, 125mW, 250mW	Adjustable, 450 mW
Detector operating temperature	-70 °C	-60 °C		-15 °C
Acquisition time		3 ms - 10 mins		≥ 8 ms
Spectral range (cm^{-1})	100-3200	300-3200	250 - 2875	250 - 2226
Spectral step (cm^{-1})	Ununiform interval	1	Ununiform interval	Ununiform interval
Spectral resolution (cm^{-1})	0.3 cm^{-1} (FWHM), Highest typically necessary: 1 cm^{-1}	5 cm^{-1} (FWHM)	7 - 10.5 cm^{-1} (FWHM)	7 cm^{-1}
Process temperature range	0°C to 40°C	5°C to 180°C	-20°C to +40°C	0°C to 40°C, non-condensing
Dimension	115.2 × 96.5 × 63.5 cm	40.0 × 10.0 × 10.0 cm	19.3 × 10.7 × 4.4 cm	24.3 × 17.2 × 6.0 cm
Weight	90 kg	5kg	800 g	< 2.3 kg

S4. Spectral comparison for different instruments:

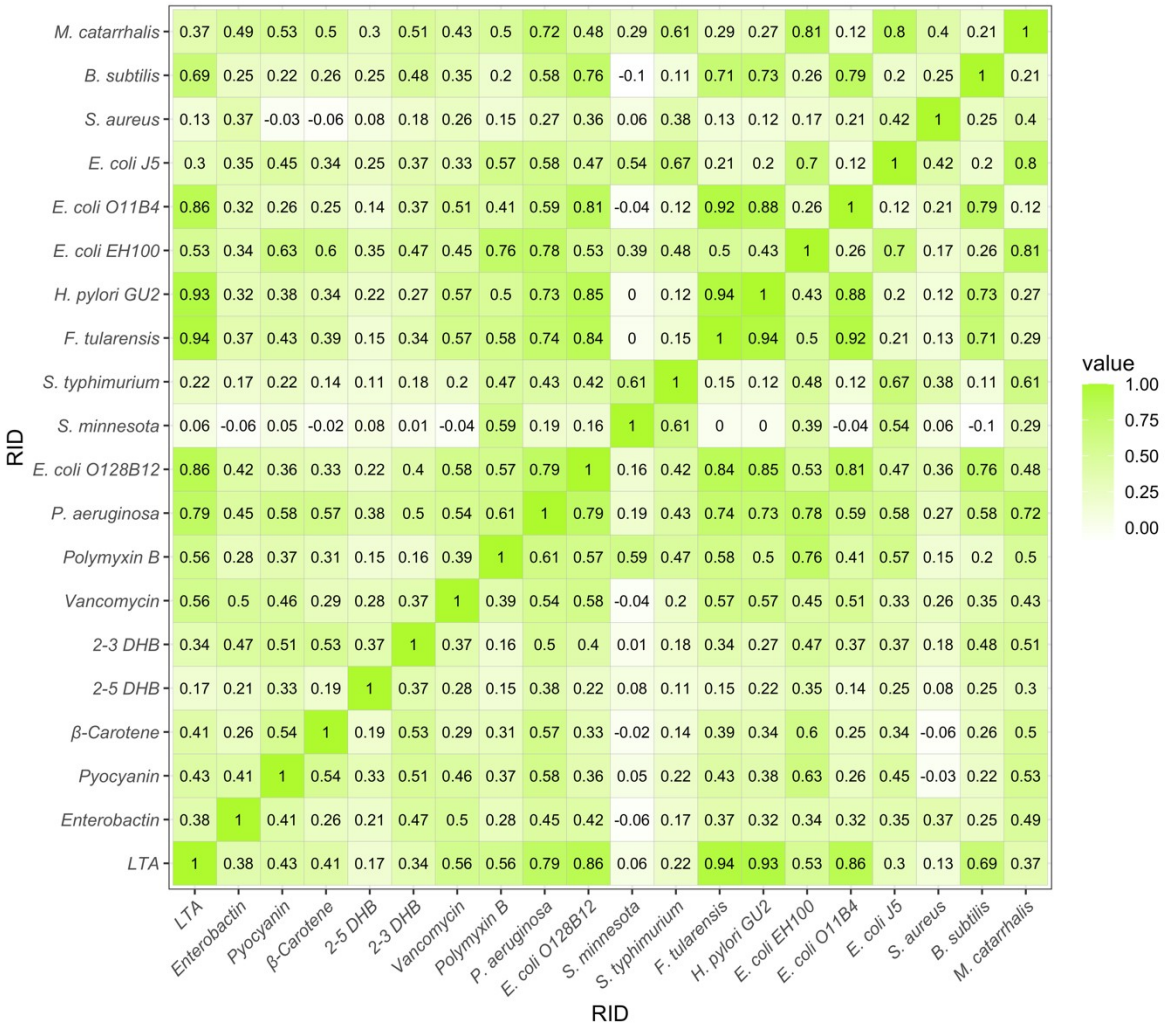

Figure S2. Pearson correlation coefficients between average RID spectra.

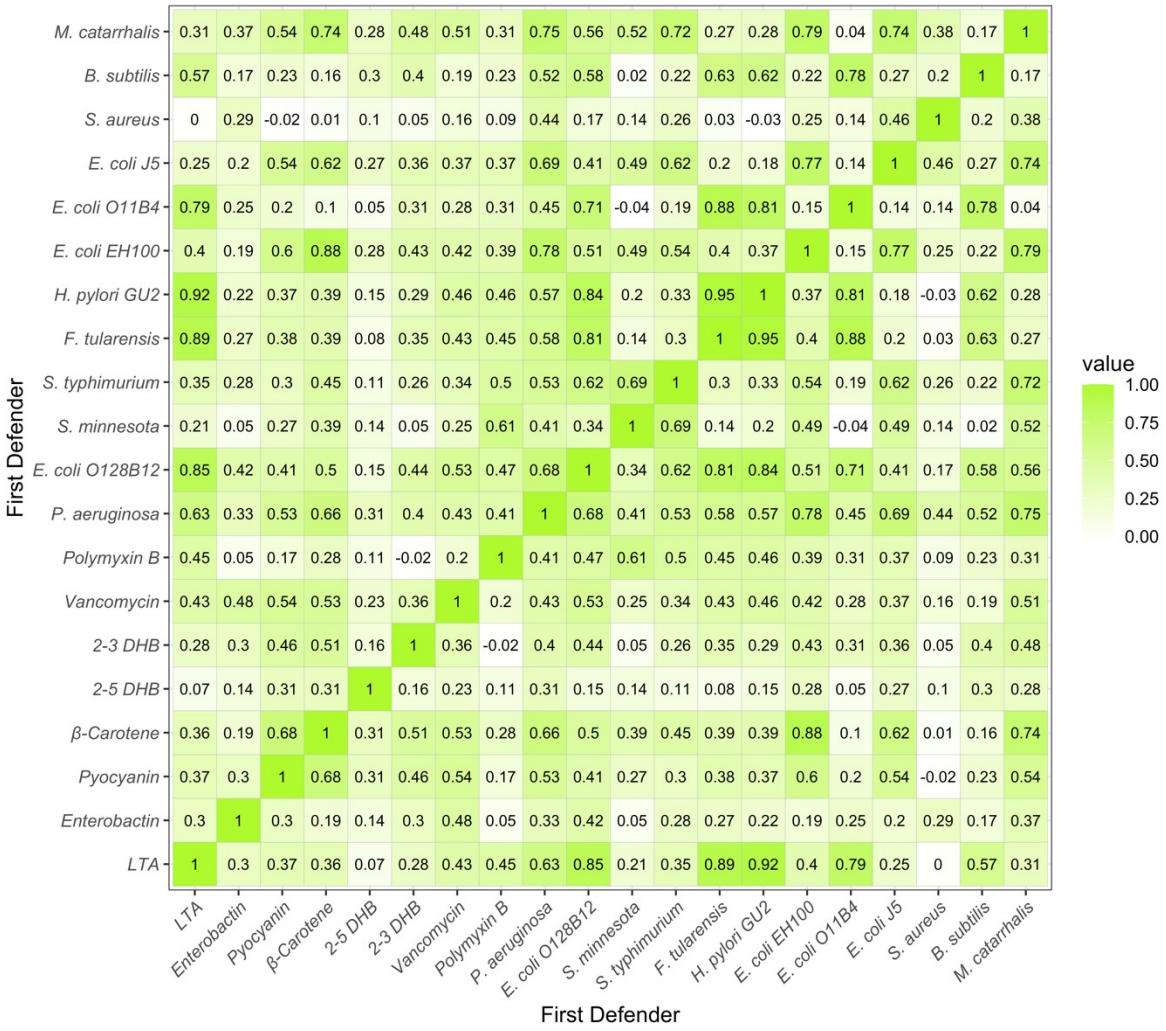

Figure S3. Pearson correlation coefficients between average FirstDefender spectra.

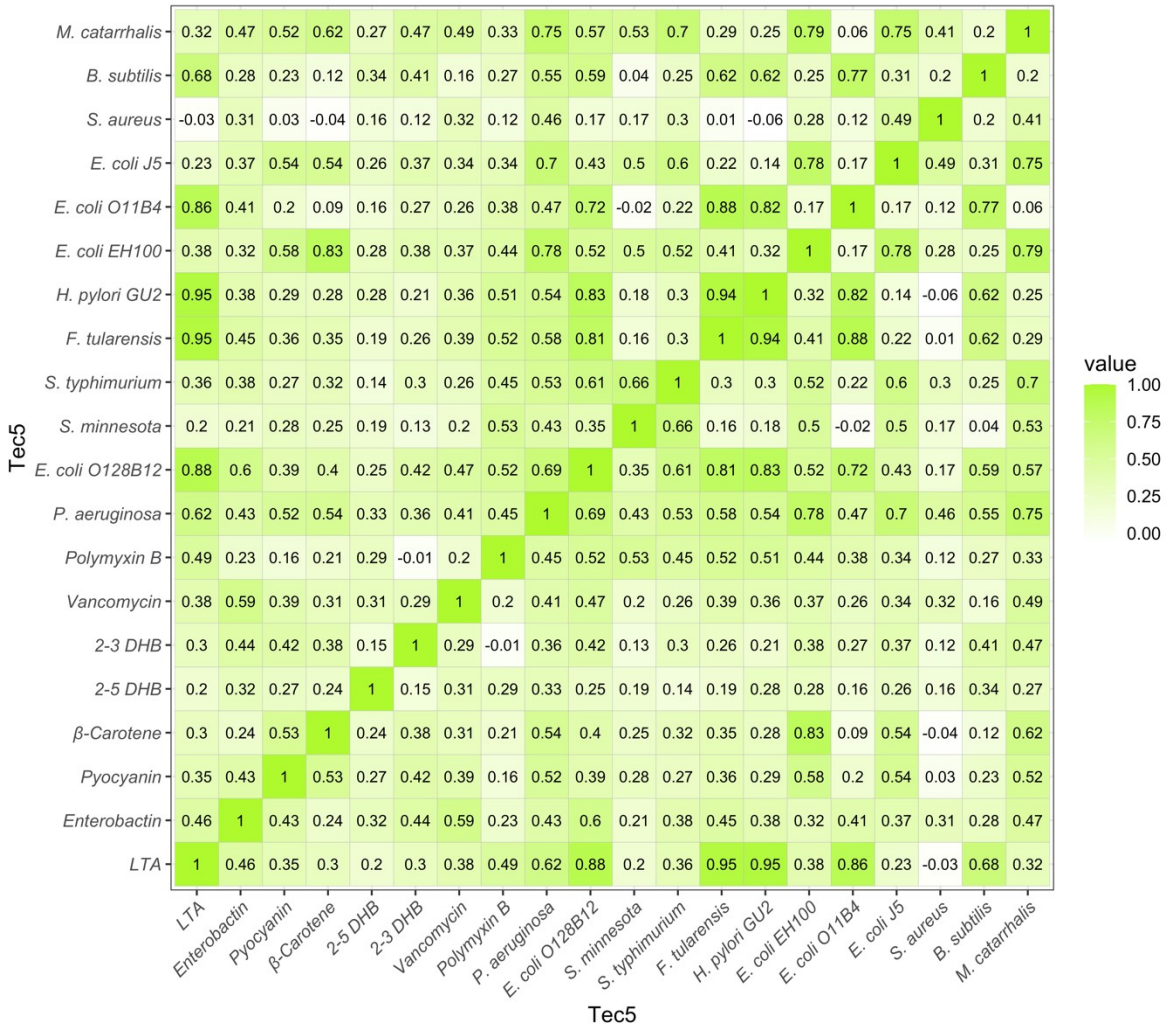

Figure S4. Pearson correlation coefficients between average Tec5 spectra.

S5. Average accuracies of different settings.

Table S3. Average accuracies of transformed individual spectrum (TIS), and original individual spectra (OIS) and original spectra with feature extraction (OISFE) of Tec5 at different training sizes.

Training Set Size	TIS		OIS (benchmark)		OISFE	
	CNN	SVM	CNN	SVM	CNN	SVM
14	0.991	0.991	0.981	0.986	0.989	0.996
15	0.991	0.993	0.980	0.988	0.987	0.997
16	0.996	0.997	0.981	0.992	0.984	0.996
19	0.994	0.988	0.957	0.976	0.966	0.987

For N = 14, 15 and 16, 30 repetitions were performed to get average accuracy. For N = 19, 20 iterations were performed to go over all the analytes.

Table S4. Average accuracies of transformed individual spectrum (TIS), and original individual spectra (OIS) and original spectra with feature extraction (OISFE) of RID at different training sizes.

Training Set Size	TIS		OIS (benchmark)		OISFE	
	CNN	SVM	CNN	SVM	CNN	SVM
14	0.964	0.893	0.887	0.917	0.960	0.960
15	0.958	0.911	0.869	0.916	0.960	0.956
16	0.966	0.935	0.896	0.94	0.955	0.959
19	0.979	0.92	0.872	0.888	0.924	0.904

For N = 14, 15 and 16, 30 repetitions were performed to get average accuracy. For N = 19, 20 iterations were performed to go over all the analytes.

Table S5. Average accuracies of transformed individual spectra (TIS), and original individual spectra (OIS) and original spectra with feature extraction (OISFE) of First Defender at different training sizes.

Training Set Size	TIS		OIS (benchmark)		OISFE	
	CNN	SVM	CNN	SVM	CNN	SVM
14	0.824	0.851	0.845	0.846	0.898	0.949
15	0.826	0.857	0.869	0.863	0.898	0.955
16	0.865	0.928	0.872	0.885	0.893	0.947
19	0.965	0.954	0.792	0.862	0.934	0.929

For $N = 14$, 15 and 16, 30 repetitions were performed to get average accuracy. For $N = 19$, 20 iterations were performed to go over all the analytes.

S6. Selected peaks on average Renishaw spectra.

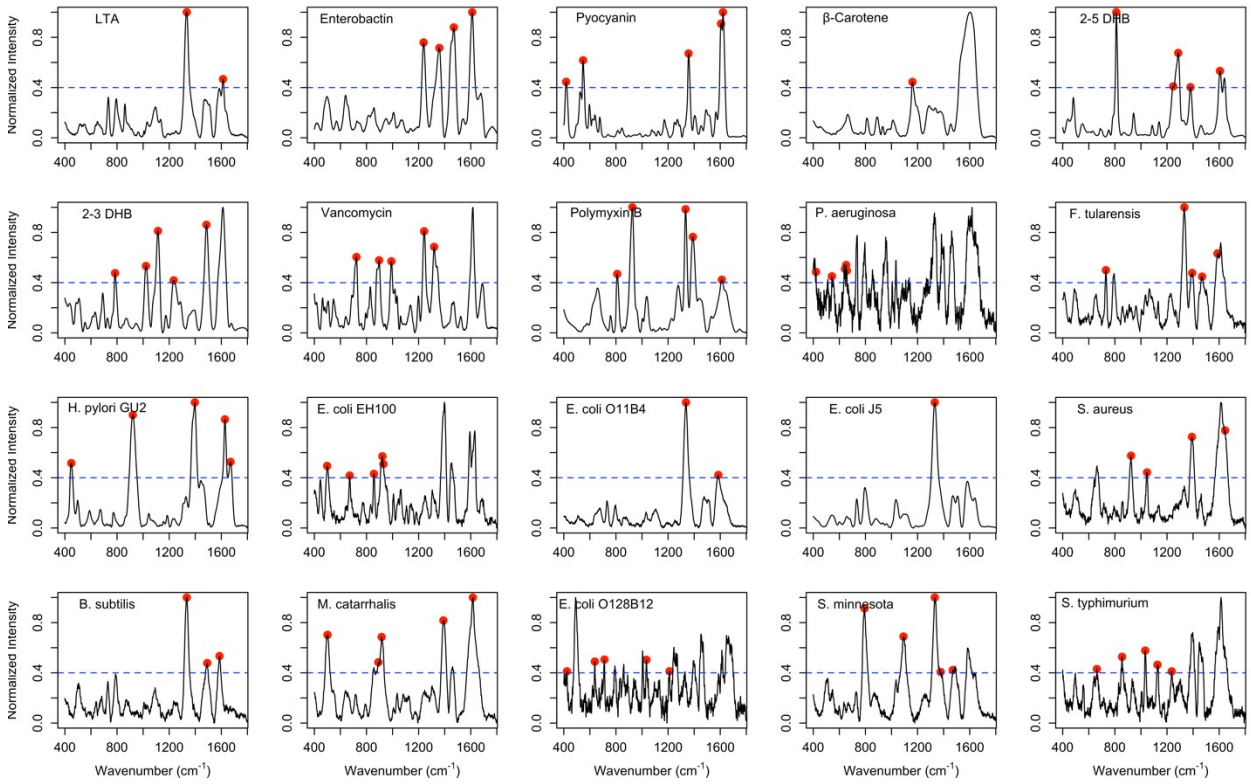

Figure S5. Selected peaks on average Renishaw spectra of 20 analytes.

S7. Classification results.

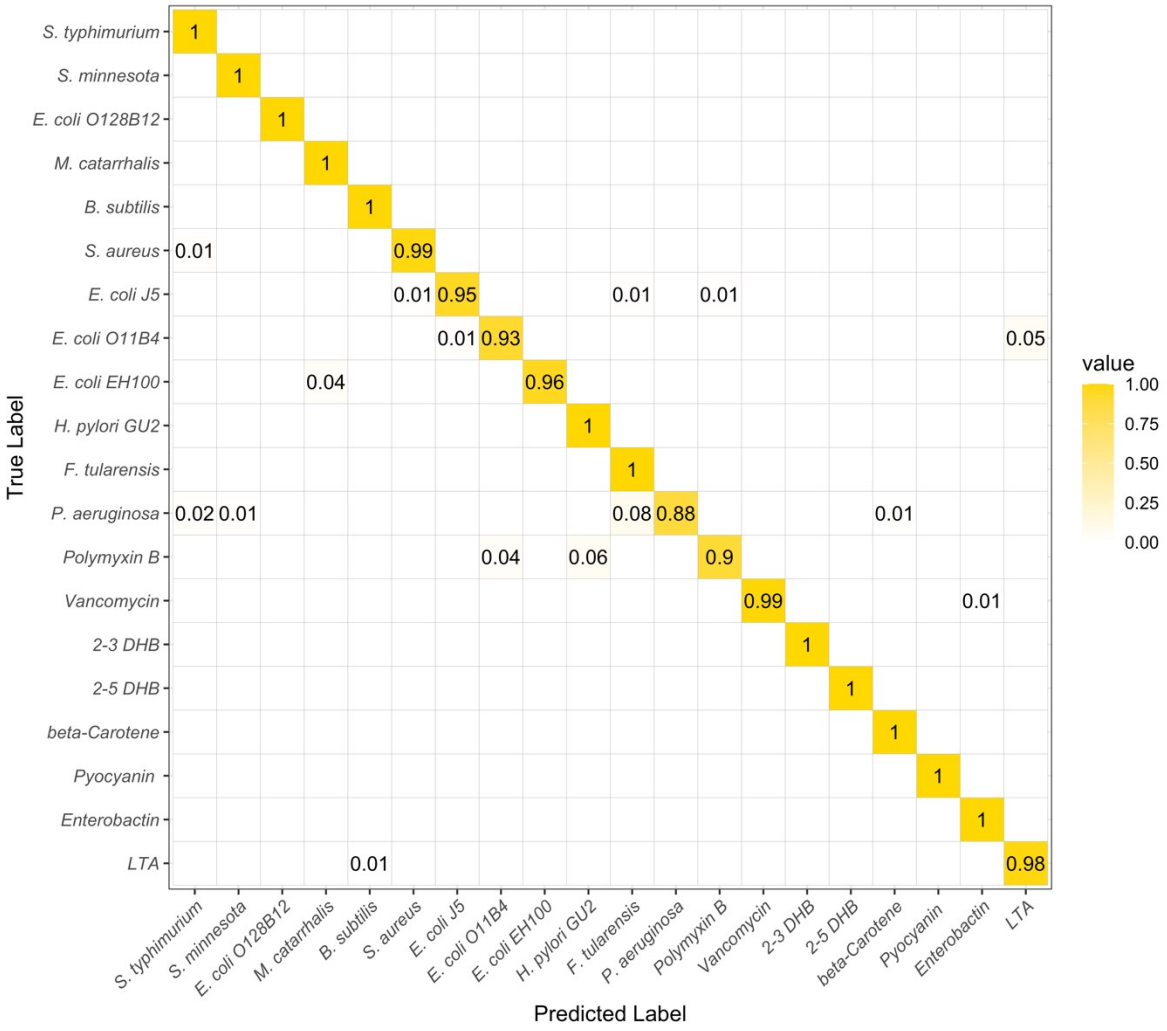

Figure S6. Confusion matrix of transformed RID individual spectra of CNN model over 20 repetitions of leave-one-out testing.

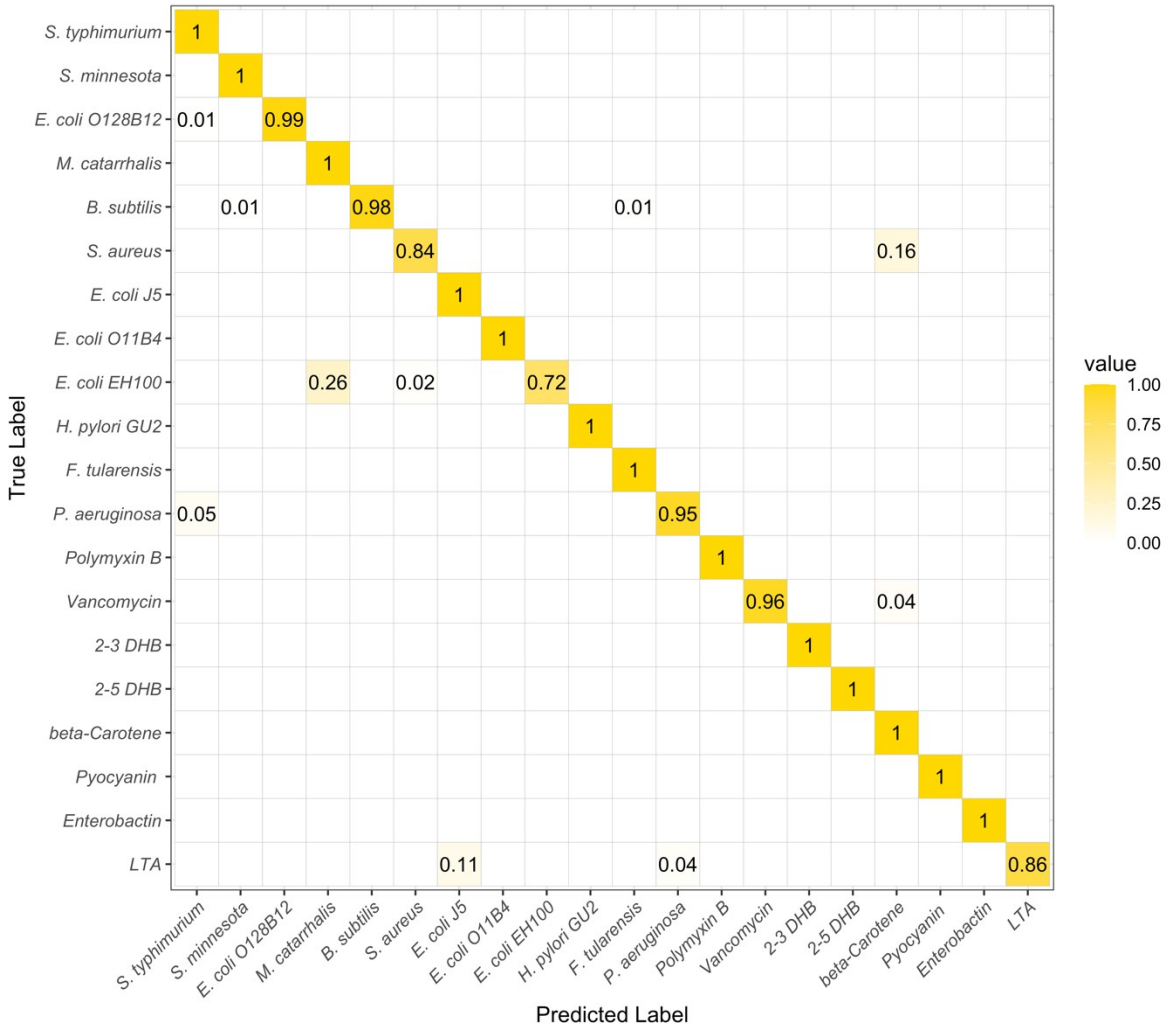

Figure S7. Confusion matrix of transformed FirstDefender individual spectra of CNN model over 20 repetitions of leave-one-out testing.

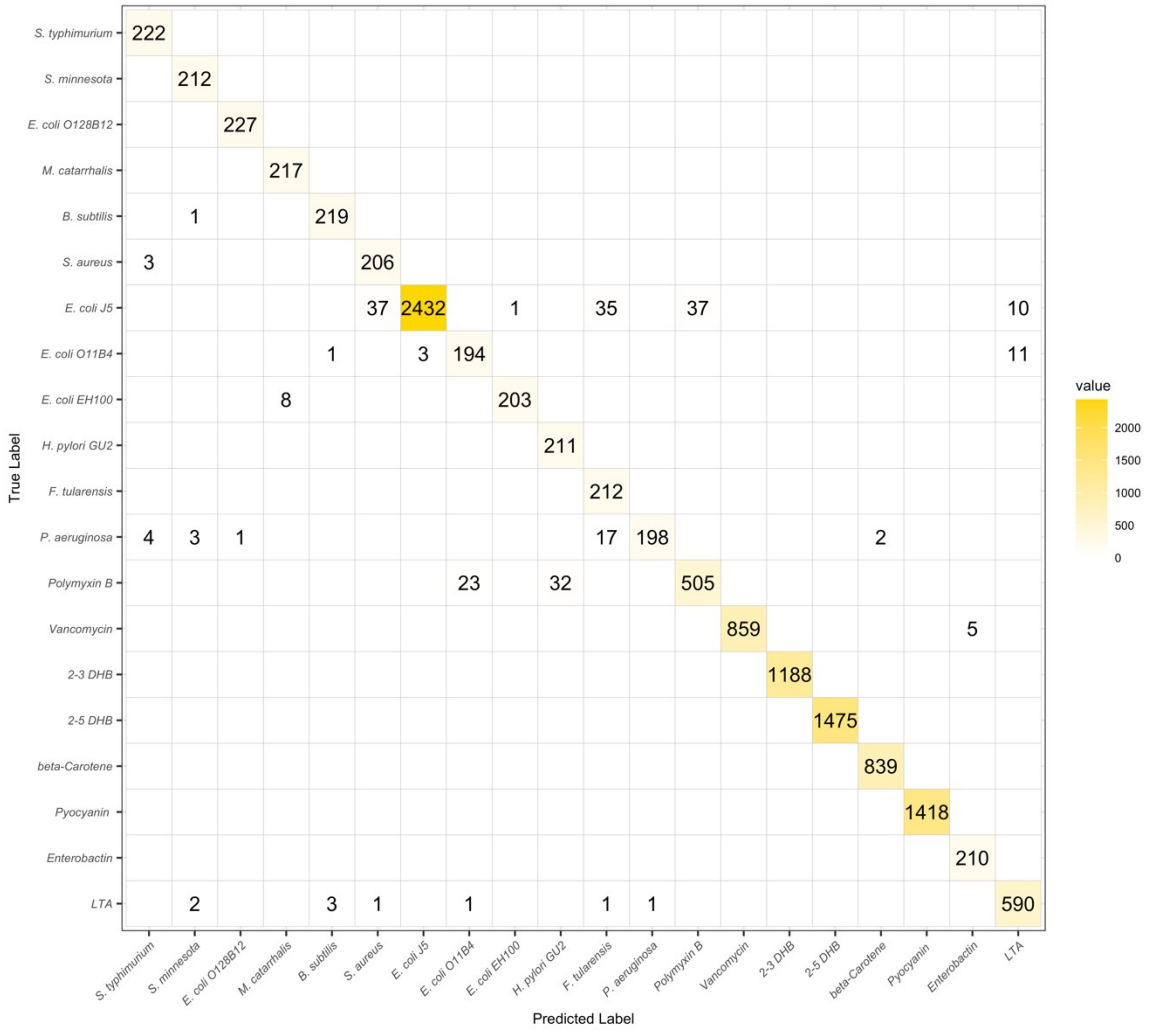

Figure S8. Confusion matrix of transformed RID individual spectra of CNN model over 20 repetitions of leave-one-out testing.

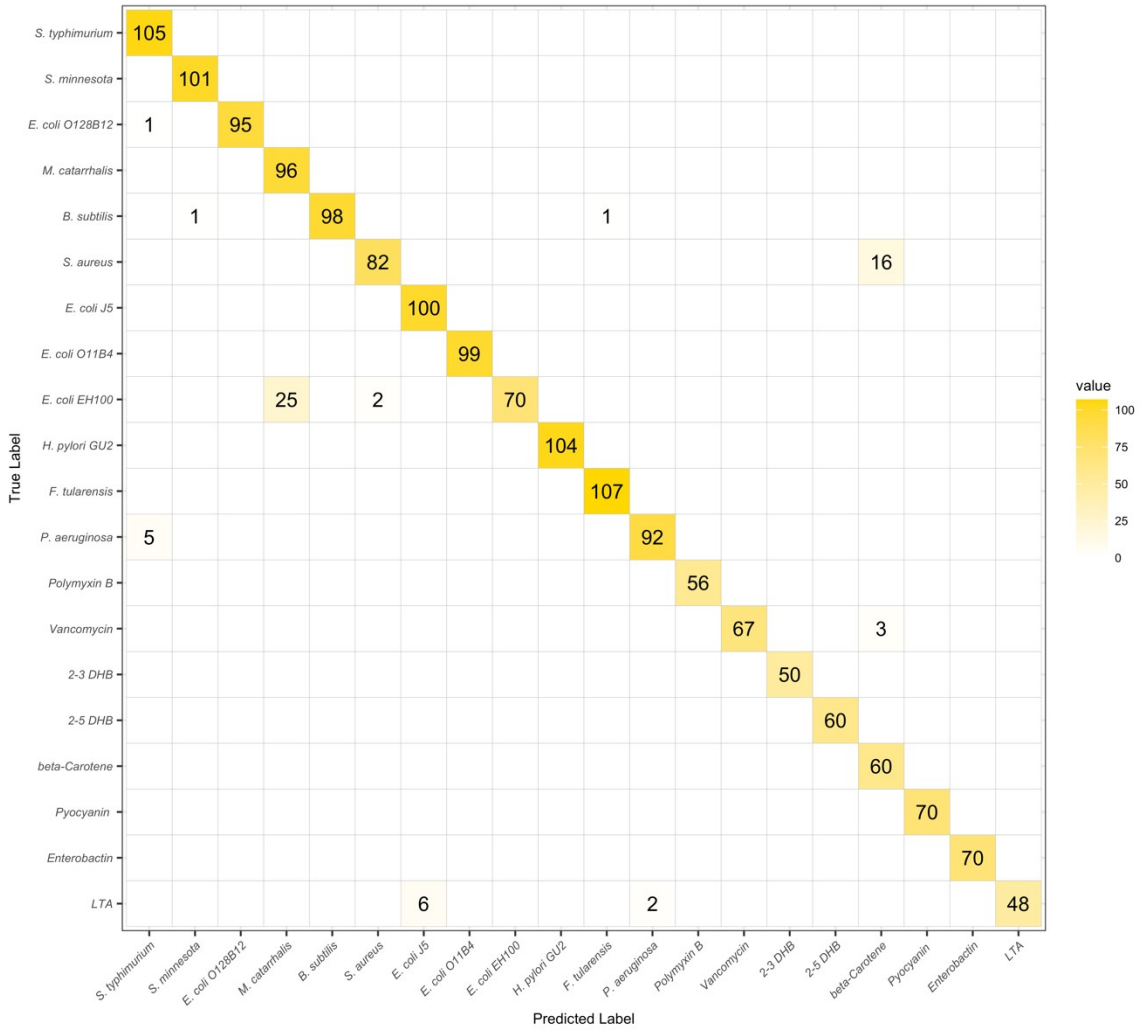

Figure S9. Confusion matrix of transformed FirstDefender individual spectra of CNN model over 20 repetitions of leave-one-out testing.

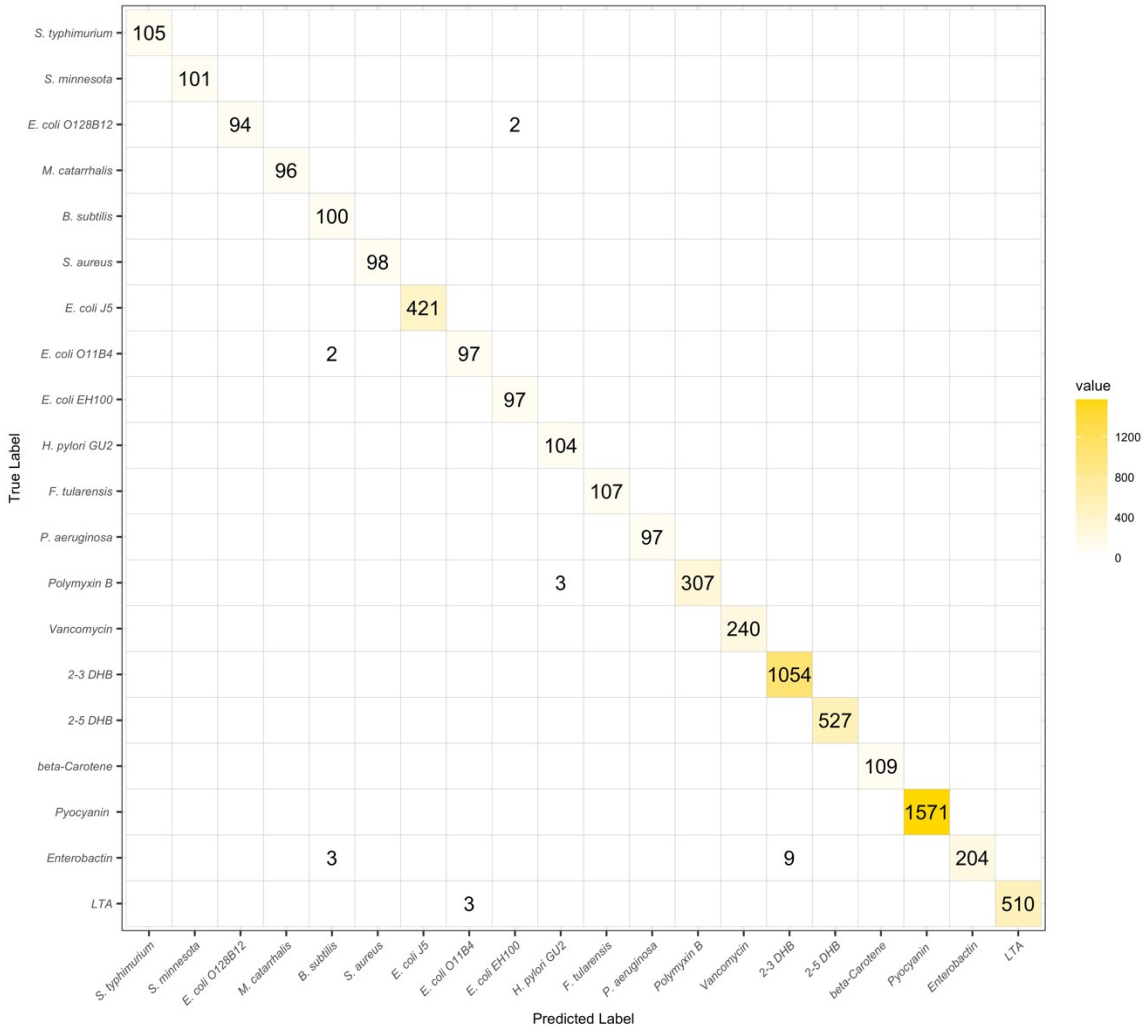

Figure S10. Confusion matrix of transformed Tec5 individual spectra of CNN model over 20 repetitions of leave-one-out testing.

Table S6. The number of spectra measured by RID, Tec5 and FirstDefender of 20 analytes.

No.	Name	Bacteria	RID	Tec5	First Defender
1	lipoteichoic acid	*S. aureus*	599	513	56
2	enterobactin	*E. coli*	210	216	70
3	pyocyanin	*P. aeruginosa*	1418	1571	70
4	β-Carotene	multiple bacteria including mycobacteria	839	109	60
5	2,5-Dihydroxybenzoic acid	mammalian siderophore	1475	527	60
6	2,3-Dihydroxybenzoic acid	bacterial siderophore	1188	1054	50
7	vancomycin hydrochloride		864	240	70
8	polymyxin B		560	310	56
9	lipopolysaccharide	*P. aeruginosa*	225	97	97
10	lipopolysaccharide	*F. tularensis* LVS	212	107	107
11	lipopolysaccharide	*H. pylori* GU2	211	104	104
12	lipopolysaccharide	*E. coli*-EH100	211	97	97
13	lipopolysaccharide	*E. coli*-O11:B4	209	99	99
14	lipopolysaccharide	*E. coli*-J5	2552	421	100
15	lipopolysaccharide	*S. meliloti* Rm1021	209	98	98
16	peptidoglycan	*B. subtilis* (PGN)	220	100	100
17	lipopolysaccharide	*M. catarrhalis*	217	96	96
18	lipopolysaccharide	*E. coli*-O128:B12	227	96	96
19	lipopolysaccharide	*S. minnesota* Re595	212	101	101
20	lipopolysaccharide	*S. enterica serovar Typhimurium*	222	105	105

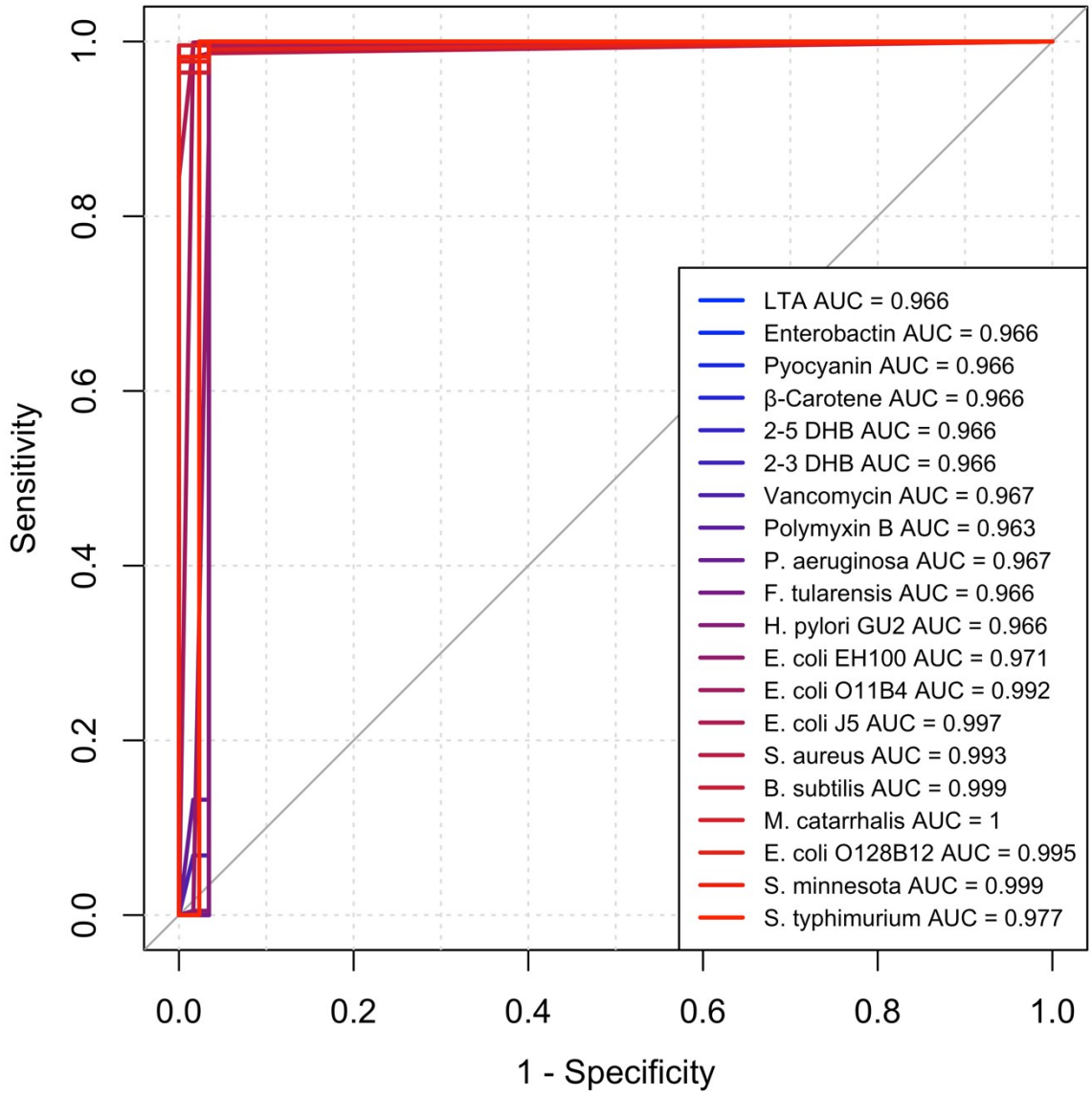

Figure S11. The obtained ROC curve of CNN using transformed individual RID spectra in the leave-one-out test.

Table S7. Precision, sensitivity, specificity and F1 score of CNN using transformed individual RID spectra in the leave-one-out test.

No.	Name	Bacteria	Precision	Sensitivity	Specificity	F1 Score
1	lipoteichoic acid	*S. aureus*	0.966	0.985	0.998	0.975
2	enterobactin	*E. coli*	0.977	1.000	1.000	0.988
3	pyocyanin	*P. aeruginosa*	1.000	1.000	1.000	1.000
4	β-Carotene	multiple bacteria including mycobacteria	0.998	1.000	1.000	0.999
5	2,5-Dihydroxybenzoic acid	mammalian siderophore	1.000	1.000	1.000	1.000
6	2,3-Dihydroxybenzoic acid	bacterial siderophore	1.000	1.000	1.000	1.000
7	vancomycin hydrochloride		1.000	0.994	1.000	0.997
8	polymyxin B		0.932	0.902	0.997	0.917
9	lipopolysaccharide	*P. aeruginosa*	0.995	0.880	1.000	0.934
10	lipopolysaccharide	*F. tularensis* LVS	0.800	1.000	0.996	0.889
11	lipopolysaccharide	*H. pylori* GU2	0.868	1.000	0.997	0.930
12	lipopolysaccharide	*E. coli*-EH100	0.995	0.962	1.000	0.978
13	lipopolysaccharide	*E. coli*-O11:B4	0.890	0.928	0.998	0.909
14	lipopolysaccharide	*E. coli*-J5	0.999	0.953	1.000	0.975
15	lipopolysaccharide	*S. meliloti* Rm1021	0.844	0.986	0.997	0.909
16	peptidoglycan	*B. subtilis* (PGN)	0.982	0.995	1.000	0.989
17	lipopolysaccharide	*M. catarrhalis*	0.964	1.000	0.999	0.982
18	lipopolysaccharide	*E. coli*-O128:B12	0.996	1.000	1.000	0.998
19	lipopolysaccharide	*S. minnesota* Re595	0.972	1.000	0.999	0.986
20	lipopolysaccharide	*S. enterica serovar Typhimurium*	0.969	1.000	0.999	0.984

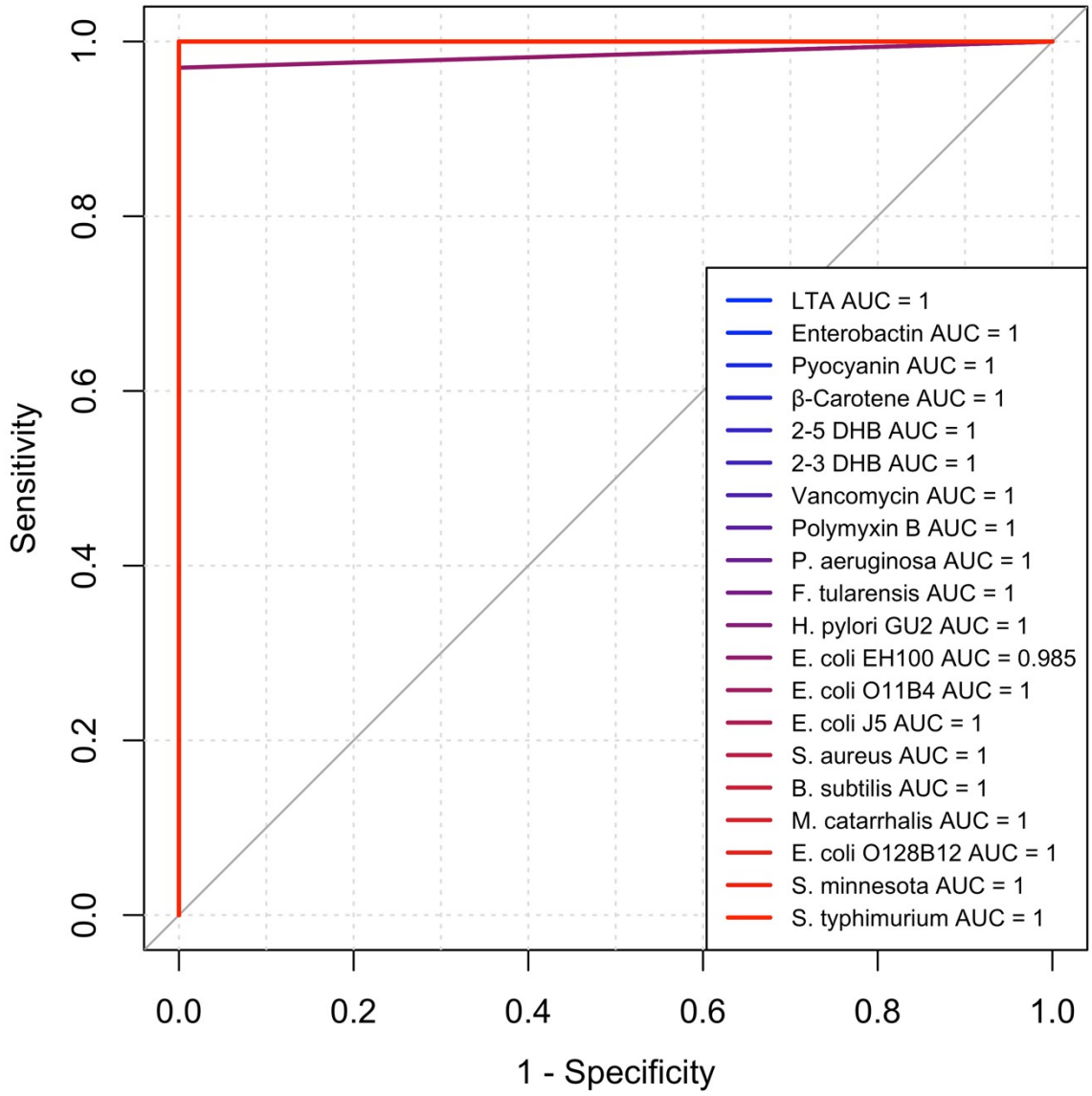

Figure S12. The obtained ROC curve of CNN using transformed individual Tec5 spectra in the leave-one-out test.

Table S8. Precision, sensitivity, specificity and F1 score of CNN using transformed individual Tec5 spectra in the leave-one-out test.

No.	Name	Bacteria	Precision	Sensitivity	Specificity	F1 Score
1	lipoteichoic acid	*S. aureus*	1.000	0.994	1.000	0.997
2	enterobactin	*E. coli*	1.000	0.944	1.000	0.971
3	pyocyanin	*P. aeruginosa*	1.000	1.000	1.000	1.000
4	β-Carotene	multiple bacteria including mycobacteria	1.000	1.000	1.000	1.000
5	2,5-Dihydroxybenzoic acid	mammalian siderophore	1.000	1.000	1.000	1.000
6	2,3-Dihydroxybenzoic acid	bacterial siderophore	0.992	1.000	0.998	0.996
7	vancomycin hydrochloride		1.000	1.000	1.000	1.000
8	polymyxin B		1.000	0.990	1.000	0.995
9	lipopolysaccharide	*P. aeruginosa*	1.000	1.000	1.000	1.000
10	lipopolysaccharide	*F. tularensis* LVS	1.000	1.000	1.000	1.000
11	lipopolysaccharide	*H. pylori* GU2	0.972	1.000	0.999	0.986
12	lipopolysaccharide	*E. coli*-EH100	0.980	1.000	1.000	0.990
13	lipopolysaccharide	*E. coli*-O11:B4	0.970	0.980	0.999	0.975
14	lipopolysaccharide	*E. coli*-J5	1.000	1.000	1.000	1.000
15	lipopolysaccharide	*S. meliloti* Rm1021	1.000	1.000	1.000	1.000
16	peptidoglycan	*B. subtilis* (PGN)	0.952	1.000	0.999	0.976
17	lipopolysaccharide	*M. catarrhalis*	1.000	1.000	1.000	1.000
18	lipopolysaccharide	*E. coli*-O128:B12	1.000	0.979	1.000	0.989
19	lipopolysaccharide	*S. minnesota* Re595	1.000	1.000	1.000	1.000
20	lipopolysaccharide	*S. enterica serovar Typhimurium*	1.000	1.000	1.000	1.000

Section S8 The B-spline Curves

A B-spline of order p+1 is a collection of piecewise polynomial functions $^{B_{i,p}(\Delta v)}$ of degree p. For a given sequence of wavenumbers, $^{\Delta v_0}$, $^{\Delta v_1,\ldots,\Delta v_m}$, there exists, up to a scaling factor, a unique spline $^{B_{i,0}(\Delta v)}$ that satisfies:

$$B_{i,0}(\Delta v) = \begin{cases} 1 \; if \; \Delta v_i \le \Delta v < \Delta v_{i+1} \\ 0 \; otherwise \end{cases}. \quad \text{(S1)}$$

The B-spline of high order p+1can be defined recursively as:

$$B_{i,p}(\Delta v) = \frac{\Delta v - \Delta v_i}{\Delta v_{i+p} - \Delta v_i} B_{i,p-1}(\Delta v) + \frac{\Delta v_{i+p+1} - \Delta v}{\Delta v_{i+p+1} - \Delta v_{i+1}} B_{i+1,p-1}(\Delta v). \quad \text{(S2)}$$

Then, the functional variables $\beta_1^j(\Delta v), \beta_2^j(\Delta v) \; and \; I_i^j(\Delta v), \; j = \{Renishaw, RID, Tec5, First \; Defender\}$ can be represented by n B-spline basis functions of order p+1 as:

$$\begin{cases} \beta_1^j(\Delta v) = \sum_{i=1}^{n} u_i^j B_{i,p}(\Delta v) \\ \beta_2^j(\Delta v) = \sum_{i=1}^{n} v_i^j B_{i,p}(\Delta v) \\ I_i^j(\Delta v) = \sum_{i=1}^{n} a_i^j B_{i,p}(\Delta v) \end{cases}. \quad \text{(S3)}$$

Hence, **Equation 2** in the **Functional regression method** section can be rewritten as:

$$\sum_{i=1}^{n} a_i^{RS} B_{i,p}(\Delta v) = \sum_{i=1}^{n} v_i^j B_{i,p}(\Delta v) + \sum_{i=1}^{n} u_i^j B_{i,p}(\Delta v) \sum_{i=1}^{n} a_i^j B_{i,p}(\Delta v) + e_i(\Delta v), \quad \text{(S4)}$$

where the coefficients $^{a_i^{RS}}$ and $^{a_i^j}$ $(i = 1,\ldots,n)$ can be estimated directly by fitting B-spline functions to the spectrum of Renishaw and instrument j while $^{v_i^j}$ and $^{u_i^j}$ $(i = 1,\ldots,n,)$ can be estimated via Equation 3 in the main text.

Notes and Reference

1. Y. J. Liu and Y. P. Zhao, *Physical Review B*, 2008, **78**, 075436.
2. J. D. Driskell, S. Shanmukh, Y. Liu, S. B. Chaney, X. J. Tang, Y. P. Zhao and R. A. Dluhy, *The Journal of Physical Chemistry C*, 2008, **112**, 895-901.
3. Y. J. Liu, H. Y. Chu and Y. P. Zhao, *The Journal of Physical Chemistry C*, 2010, **114**, 8176-8183.
4. Y. J. Liu, Z. Y. Zhang, Q. Zhao, R. A. Dluhy and Y. P. Zhao, *The Journal of Physical Chemistry C*, 2009, **113**, 9664-9669.
5. C. Song, Y. Yang, B. Yang, Y. Sun, Y. P. Zhao and L.-H. Wang, *Nanoscale*, 2016, **8**, 17365-17373.
6. Y. J. Liu, H. Y. Chu and Y. P. Zhao, *Journal of Physical Chemistry C*, 2010, **114**, 8176-8183.
7. Y.-J. Liu, Z.-Y. Zhang, Q. Zhao, R. Dluhy and Y.-P. Zhao, *The Journal of Physical Chemistry C*, 2009, **113**, 9664-9669.
8. O. Westphal, *Methods Carbohydrate Chem*, 1965, **5**, 83.
9. S. Soni, R. Ernst, A. Muszynski, N. Mohapatra, M. Perry, E. Vinogradov, R. Carlson and J. Gunn, *Frontiers in Microbiology*, 2010, **1**.
10. S. Gao, D. Peng, W. Zhang, A. Muszyński, R. W. Carlson and X.-X. Gu, *The FEBS Journal*, 2008, **275**, 5201-5214.
11. R. Davis Michael, A. Muszyński, V. Lollett Ivonne, L. Pritchett Christopher, W. Carlson Russell and B. Goldberg Joanna, *Journal of Bacteriology*, 2013, **195**, 1504-1514.