Electronic Supplementary Information

Differentiation and classification of bacterial endotoxins based on surface enhanced Raman scattering and advanced machine learning

Yanjun Yang [a] *, Beibei Xu [b], James Haverstick [c], Nabil Ibtehaz [d], Artur Muszyński [e], Xianyan Chen [b], Muhammad E. H. Chowdhury [f], Susu M. Zughaier [g] *, Yiping Zhao [c] *

[a] School of Electrical and Computer Engineering, College of Engineering, The University of Georgia, Athens, GA 30602, USA

[b] Department of Statistics, The University of Georgia, Athens, GA 30602, USA

[c] Department of Physics and Astronomy, The University of Georgia, Athens, GA 30602, USA

[d] Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA

[e] Complex Carbohydrate Research Center, University of Georgia, Athens, GA 30602, USA

[f] Department of Electrical Engineering, College of Engineering, Qatar University, Doha, Qatar, PO. Box 2713

[g] Department of Basic Medical Sciences, College of Medicine, QU Health, Qatar University, Doha, Qatar, PO. Box 2713

S1. LPSs and controls used in this study

Table S1: Characterization of LPS structures used in this study.

LPS source	**Lipid A backbone and acylation pattern**[!]	**Polysaccharide (core-O-antigen type**[#]	**Biological activity***
E. coli-EH100 (Ra, rough LPS)	$(GlcN)_2$, $P_{2,}$ hexa-acyl (14:0(3-OH)$_3$, 14:0, 12:0)	Ra-type core oligosaccharide, no O-antigen	Active
E. coli-J5 (Rc, rough LPS)	$(GlcN)_2$, $P_{2,}$ hexa-acyl (14:0(3-OH)$_3$, 14:0, 12:0)	Rc-type core oligosaccharide, no O-antigen	Active
E. coli-O11:B4 Smooth LPS	$(GlcN)_2$, $P_{2,}$ hexa-acyl (14:0(3-OH)$_3$, 14:0, 12:0)	core oligosaccharide- O-antigen polymer smooth LPS	Active
E. coli-O128:B12 Smooth LPS	$(GlcN)_2$, $P_{2,}$ hexa-acyl (14:0(3-OH)$_3$, 14:0, 12:0)	core-O-antigen polymer;	Active
F. tularensis LVS Smooth LPS	$(GlcN)_2$, $P_{,}$ GalN, tetra-acyl (18:0(3-OH)$_3$, 16:0) or (16:0(3-OH)$_3$, 18:0)	core oligosaccharide -O-antigen polymer	Inactive (very weak agonist Does not induce TLR4)
H. pylori GU2 Smooth LPS	$(GlcN)_2$, PEtN, tetra-acyl; (18:0(3-OH)$_{3,}$ 16:0)	core oligosaccharide -O-antigen polymer	Inactive
M. catarrhalis LOS	$(GlcN)_2$, $P_{2,}$ PEtN, hepta-acyl (12:0(3-OH)$_4$, 12:0, 10:0)	core oligosaccharide, no O-antigen /lipooligosaccharide(LOS)	Active
P. aeruginosa Smooth LPS	$(GlcN)_2$, $P_{2,}$ hexa/hepta-acyl (12:0(3-OH)$_3$, 10:0 (3-OH)$_{1\ \text{or}\ 2}$, , 12:0, 16:0)	core oligosaccharide -O-antigen polymer	Active
S. minnesota Re595 Rough LPS	$(GlcN)_2$, $P_{2,}$ hexa-acyl (14:0(3-OH)$_3$, 14:0, 12:0)	Re- type truncated core oligosaccharide /]	Active
S. enerica serovar Typhimurium Rough LPS	$(GlcN)_2$, $P_{2,}$ hexa-acyl (14:0(3-OH)$_3$, 14:0, 12:0)	core oligosaccharide -O-antigen polymer /	Active
Sinorhizobium meliloti Rm1021 Smooth LPS	$(GlcN)_2$, $P_{2,}$ penta-acyl e.g (14:0(3-OH)$_2$, 18:0(3-OH)$_{2,}$ 28:0(27-OH)/ 28:0(27-βOme C4)	core oligosaccharide -O-antigen polymer	Inactive

Table Foot notes

!: The lipid A backbone and number and length of lipid A fatty acyl chains.

#: Complex structure of saccharide chain as poly or oligo or just truncated to core saccharides.

*: Immunostimulatory potency or ability to induce immune responses in macrophages

LPSs

In general, these 11 kinds of LPSs are representative of the most common bacteria that cause disease in humans. Bacterial pathogens such as various strains of *E. coli* and *Salmonella* can cause gastroenteritis or foodborne diseases as well as other infections such as urinary tract (UTI), sepsis, and nosocomial or healthcare acquired infections. *Pseudomonas aeruginosa* can cause wide range of infections in human such as lung infections, UTI, sepsis, skin, and other nosocomial infections. *H. pylori* is the common cause of stomach infection or gastritis which affect half of the global population. *Fransicella tularensis* is a serious bio-threat pathogen that causes tularemia which could be fatal in humans. All these LPSs represent gram negative bacterial pathogens that are often evolving into antibiotic resistant strains which render it difficult to treat.[1-4]

Non LPS controls

In this study, we tested cell wall biomarkers from Gram positive bacterial pathogens devoid of LPS such as *Staphylococcus aureus* that causes serious wide range of infections in humans including food poisoning, sepsis, soft tissue, joint and bone infections as well as nosocomial infections. Antibiotic resistant *S. aureus* known as MRSA is considered an urgent threat to human health as this pathogen spreads in community and healthcare setting increasing disease burden and treatment failure.[4, 5] *Bacillus subtilis* was also used in this study as is a common cause of food poisoning or foodborne infection.[6]

Staphylococcus aureus peptidoglycan (PGN) is a ~1000-fold less biologically active in inducing innate immune responses compared to LPS. *B. subtilis* lipotecioic acid (LTA) is similar to PGN in biological activity. Chitin is a linear aminopolyssaccharide polymer.

S2. SERS substrate and collected SERS spectra.

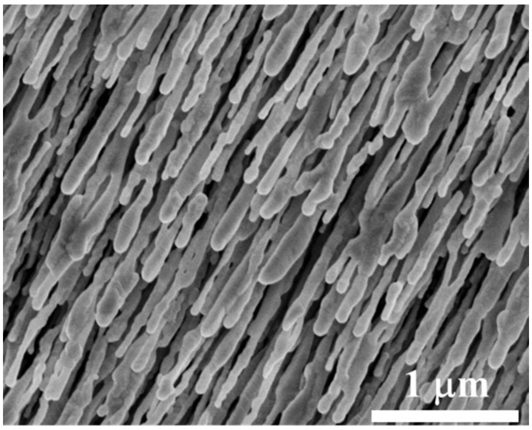

Fig. S1 SEM image of AgNR array SERS substrate.

Table S2. The number of SERS spectra taken and the average spectral correlation coefficient using different data pre-processing methods.

		Spectra number	Correlation coefficient		
	Baseline correction method		**Our method**	**Raw data**	**WiRE**
1	*E. coli*-EH100	390	0.986	0.978	0.899
2	*E. coli*-J5	409	0.984	0.966	0.941
3	*E. coli*-O11:B4	392	0.954	0.982	0.959
4	*E. coli*-O128:B12	394	0.994	0.980	0.956
5	*F. tularensis* LVS	415	0.993	0.990	0.986
6	*H. pylori* GU2	440	0.993	0.988	0.987
7	*M. catarrhalis*	367	0.987	0.985	0.833
8	*P. aeruginosa*	393	0.942	0.942	0.808
9	*S. meliloti* Rm1021	419	0.900	0.937	0.871
10	*S. minnesota* Re595	427	0.981	0.969	0.914
11	*S. enterica serovar Typhimurium*	396	0.984	0.976	0.968
12	*S. aureus* (LTA)	436	0.985	0.975	0.951
13	*B. subtilis* (PGN)	338	0.984	0.979	0.959
14	Chitin	408	0.990	0.984	0.966
	Average	**402**	**0.98 ± 0.03**	**0.97 ± 0.02**	**0.93 ± 0.06**

S3. Baseline correction method and parameter boundaries

Table S3. The boundaries for the parameters of baseline fitting function.

Definition	Parameter	Boundary
amplitude of the Gaussian function	A	$0.3 \le A \le 0.5$
center of the Gaussian peak	v_g	$75 \le v_g \le 300$
standard deviation of the Gaussian function	σ_g	$800 \le v_g \le 1800$
area of the Lorentzian function	L	$100 \le L \le 400$
center of the Lorentzian peak	v_l	$125 \le v_l \le 150$
width of the Lorentzian peak	σ_l	$400 \le \sigma_l \le 500$
“ground” level of the SERS spectrum	I_0	$-0.1 \le I_0 \le 0.1$

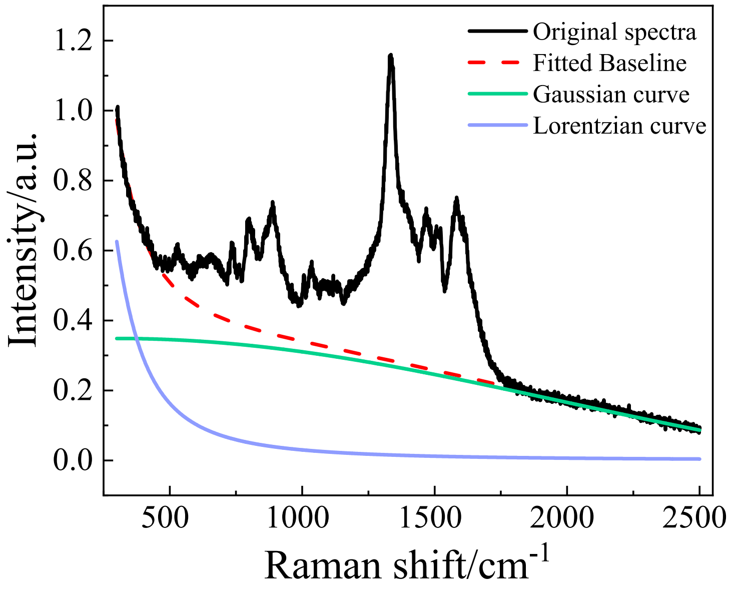

Fig. S2 Baseline correction: a typical raw SERS spectrum of *F. tularensis* LVS (solid black curve), a fitted baseline (dashed red curve), and the corresponding Gaussian curve (solid blue curve) and Lorentzian curve (solid green curve) in the baseline fitting.

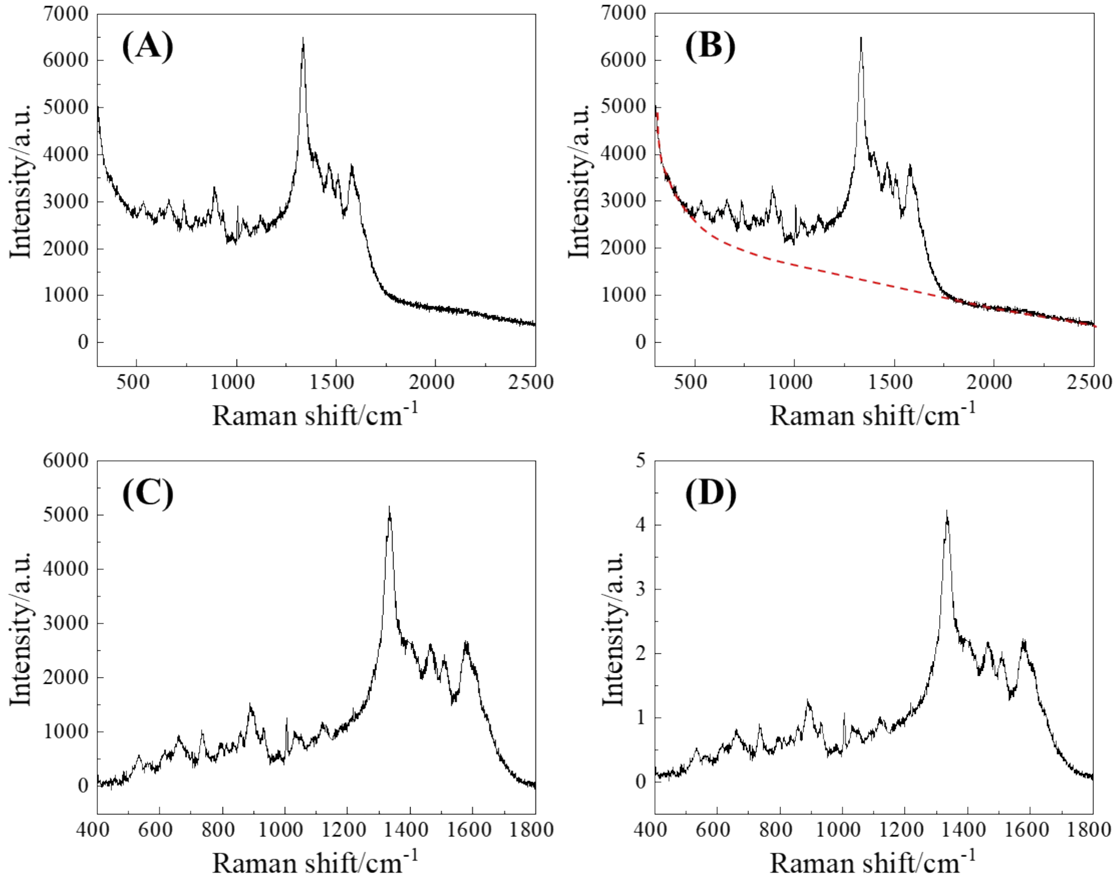

Fig. S3 (A) Typical raw SERS spectrum of *F. tularensis* LVS. (B) Baseline corrected by the Eq.1 (red dash line). (C) Baseline corrected spectrum. (D) Normalize the spectrum by the mean of the spectrum.

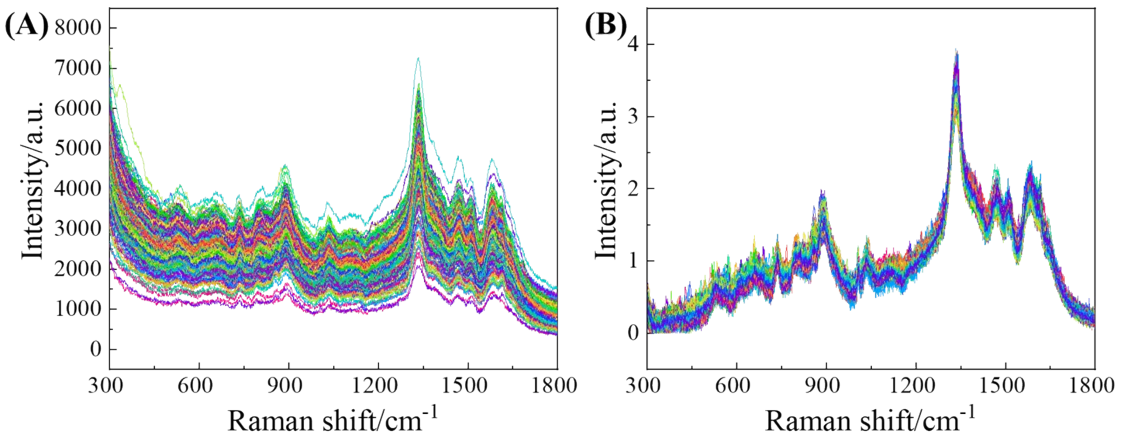

Fig. S4 (A) Original SERS spectra of *F. tularensis* LVS. (B) SERS spectra of *F. tularensis* LVS after pre-processing.

S4. The effect of different spectra pre-processing on MLAs

Table S4. Comparison of the results of the SVM trained models different with kernel functions.

	Accuracy
Linear	0.9998±0.0005
Polynomial	0.975±0.003
Sigmoid	0.931±0.009
RBF	0.999±0.001

Table S5. The accuracies obtained from different machine learning models for different data pre-processing methods.

Models	**Data pre-processing methods**			
	Our method	**Raw data**	**Raw data with normalization**	**WiRE correction**
RamanNet	1.0000 ± 0.0000	0.9980 ± 0.0003	0.9992 ± 0.0003	0.997 ± 0.002
SVM	0.9998 ± 0.0005	0.994 ± 0.003	0.995 ± 0.004	0.995 ± 0.002
RF	0.995 ± 0.001	0.973 ± 0.006	0.981 ± 0.002	0.975 ± 0.006
KNN	0.992 ± 0.003	0.979 ± 0.008	0.978 ± 0.006	0.953 ± 0.006
LDA	0.998 ± 0.001	0.988 ± 0.002	0.990 ± 0.002	0.965 ± 0.007
PLS-DA	0.98 ± 0.02	0.85 ± 0.02	0.924 ± 0.004	0.78 ± 0.02

Based on the overall spectral features of SERS spectra we obtained, a simple baseline correction method using Gaussian and Lorentzian function fitting was developed. **Table S2** summarized the average correlation coefficient between each individual spectrum and the average spectrum by using different baseline removal methods. A commercial WiRE baseline correction method based on polynomial fitting gave the average correlation coefficient ranging from 0.899 to 0.987, with an average coefficient of 0.93 ± 0.06, even worse than that (0.97 ± 0.02) of the original spectra set. Our baseline correction method shows the highest average correlation coefficient, 0.98 ± 0.03. In addition, the feature space from different data set is projected into a 2 dimensional-map using t-Distributed Stochastic Neighbor Embedding (tSNE).[7] Most clusters in the tSNE plot derived from PCA feature space from raw data (**Fig. S5A**) have curve shape and are overlapped due to the large variation of baselines. After the spectra were only normalized by area, the clusters are still overlapped (**Fig. S5C**). With WiRE baseline correction, only four clusters can be well separated, and other clusters are still overlapped. However, our baseline correction method shows that most clusters are free from overlaps and distributed properly with minimizing intraclass distance and maximizing interclass distance (**Fig. S5D**). This ensures better distinction among the classes, which increases classification performance as shown in **Table S5**. Compared to WiRE baseline correction, the accuracies of our baseline correction method increased significantly from 0.78 to 0.98 for PLS-DA model, and from 0.995 to 0.9998 for SVM model. Moreover, as shown in RamanNet embedding space, RamanNet model enables most clusters distribute properly without overlap and maximizing interclass distance, even for raw spectra (**Fig. S5B**), raw spectra with normalization (**Fig. S5D)**, and WiRE corrected spectra (**Fig. S5F)**, only several spectra were misclassified. Using our baseline correction method, all clusters are well separated with minimizing intraclass distance and maximizing interclass distance (**Fig. S5H)**, which can achieve a 100% accuracy. Although the accuracy increase is apparently insignificant for RamanNet model using different data pre-processing, we hope that our baseline correction method could facilitate the machine learning model to achieve higher accuracy for more complex data sets measured from real patient samples.

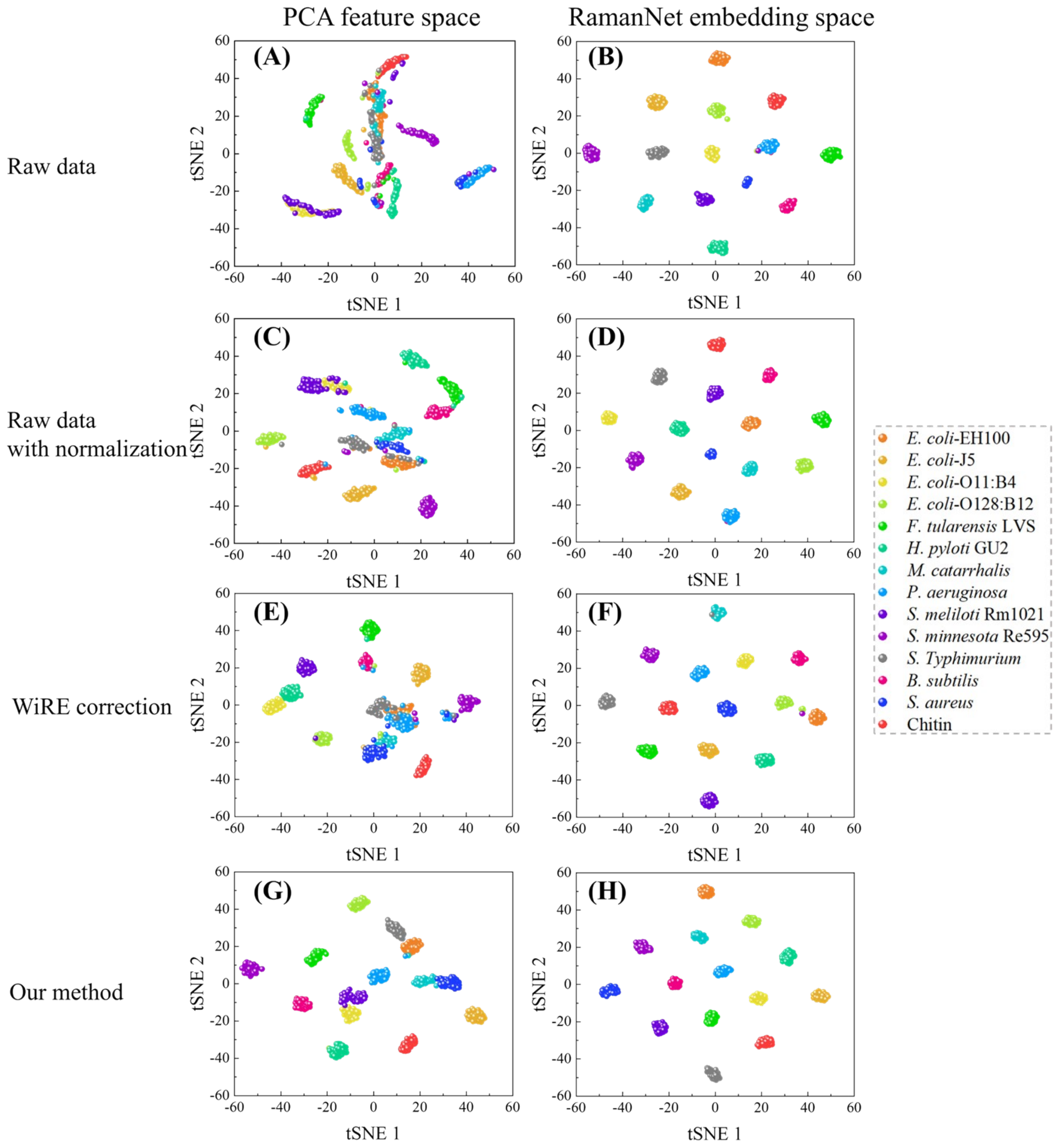

Fig. S5 The 256-dimensional feature space is projected into a 2 dimensional-map using tSNE based on SERS spectra from different data pre-processing methods. PCA feature space: (A) raw data, (C) raw data with normalization, (E) WiRE correction, (G) our method. RamanNet embedding space: (B) raw data, (D) raw data with normalization, (F) WiRE correction, (H) our method.

S5. SERS peak assignments

Table S6. SERS peak assignment for bacterial endotoxin structures[8-13].

Endotoxin	Peak (cm^{-1})	Peak assignment	Component
LPSs			
E. coli-EH100	668	δ(C-O-C) (glycosidic linkage), δ (C-C-C), δ(C-O-H)	O-Antigen
	809	δ(C-C-H), δ(C-O-H)	O-Antigen
	852	v (C-O-C)	saccharides (1,4 glycosidic link)
	920	δ(C-C-H), δ(C-O-H)	O-Antigen
	1003	v (C-O), v (C-C)	O-Antigen
	1043	v (C-O), v (C-C)	O-Antigen
	1142	δ(C-H)	Lipid A
	1340	δ(C-H)	Lipid A
	1381	δ(CH$_3$), v (C-N)	O-Antigen
	1445	Scissoring CH$_2$/CH$_3$, δ(C-O-C) (glycosidic linkage), δ(C-C-C), δ (C-O-H)	Lipid A
	1615	v (C-C) v (C-O)	Lipid A
E. coli-J5	672	δ(C-O-C)	fatty acid
	895	δ(C-C-H), δ(C-O-H)	O-Antigen
	1005	v (C-O), v (C-C)	O-Antigen
	1337	δ(C-H)	Lipid A
	1381	δ(CH$_3$), v (C-N)	O-Antigen
	1613	v (C-C) v (C-O)	Lipid A
E. coli-O11:B4	731	β(C-O-C)	carbohydrates
	792	v (C-O)	ring
	856	δ(C-C-H), δ(C-O-H)	O-Antigen
	1005	v (C-C) v (C-O)	O-Antigen
	1033	v (C-C) v (C-O)	O-Antigen
	1099	v (C-N)	O-Antigen
	1336	δ(C-H)	Lipid A
	1467	δ(C-O-C) (glycosidic linkage), δ (C-C-C), δ(C-O-H)	O-Antigen
	1508	v(C-N), δ(CH$_3$)	O-Antigen
	1583	v (C-O)	O-Antigen
	1618	v (C-C) v (C-O)	Lipid A
E. coli-O128:B12	672	β(C-O-C)	carbohydrates
	731	β(C-O-C)	carbohydrates
	790	v (C-O)	ring
	911	δ(C-C-H), δ(C-O-H)	O-Antigen
	1006	v (C-C) v (C-O)	O-Antigen
	1028	v (C-O)	carbohydrates
	1085	v (C-O)	Lipid A
	1331	β(C-H)	Lipid A
	1381	δ(CH$_3$), v (C-N)	O-Antigen
	1463	δ(C-O-C) (glycosidic linkage), δ (C-C-C), δ(C-O-H)	O-Antigen

	1580	ν (C-O)	O-Antigen
	1615	ν (C-O), ν (C-C)	Lipid A
F. tularensis LVS	531	δ(C-O-C) (glycosidic linkage)	O-Antigen
	662	δ(C-O-C) (glycosidic linkage), δ (C-C-C), δ(C-O-H)	O-Antigen
	736	Symmetric ν $(CH_3)_3$	Lipid A
	794	ν (C-O)	ring
	891	δ(C-C-H), δ(C-O-H)	O-Antigen
	1006	ν (C-O), ν (C-C)	O-Antigen
	1036	ν (C-O), ν (C-C)	O-Antigen
	1337	δ(C-H)	Lipid A
	1469	δ(C-O-C) (glycosidic linkage), δ (C-C-C), δ(C-O-H)	O-Antigen
	1508	ν(C-N), δ(CH_3)	O-Antigen
	1583	ν (C-O)	O-Antigen
H. pyloti GU2	548	δ(C-O-C) (glycosidic linkage)	O-Antigen
	656	δ(C-O-C) (glycosidic linkage), δ (C-C-C), δ(C-O-H)	O-Antigen
	735	Symmetric ν $(CH_3)_3$	Lipid A
	795	ν (C-O)	ring
	892	δ(C-C-H), δ(C-O-H)	O-Antigen
	1034	ν (C-O), ν (C-C)	O-Antigen
	1333	δ(C-H)	Lipid A
	1467	δ(C-O-C) (glycosidic linkage), δ (C-C-C), δ(C-O-H)	O-Antigen
	1508	ν(C-N), δ(CH_3)	O-Antigen
	1578	ν (C-O), N-H bending	O-Antigen
M. catarrhalis	560	β(CH_2)	ring
	665	δ(C-O-C) (glycosidic linkage), δ (C-C-C), δ(C-O-H)	O-Antigen
	759	Symmetric ν $(CH_3)_3$	Lipid A
	890	δ(C-C-H), δ(C-O-H)	O-Antigen
	1006	ν (C-O), ν (C-C)	O-Antigen
	1037	ν (C-O), ν (C-C)	O-Antigen
	1135	δ(C-H)	Lipid A
	1174	δ(C-O-C) (glycosidic linkage)	Lipid A
	1335	δ(C-H)	Lipid A
	1459	α (CH_2/CH_3), β(CH_2/CH_3)	Lipid A
	1616	ν (C-O), ν (C-C)	Lipid A
P. aeruginosa	542	δ(C-O-C) (glycosidic linkage)	O-Antigen
	648	δ(C-O-C) (glycosidic linkage), δ (C-C-C), δ(C-O-H)	O-Antigen
	734	β(C-O-C)	carbohydrates
	789	ν (C-O)	ring
	857	δ(C-C-H), δ(C-O-H)	O-Antigen
	894	δ(C-C-H), δ(C-O-H)	O-Antigen
	1004	ν (C-O), ν (C-C)	O-Antigen

	1047	ν (C-O), ν (C-C)	O-Antigen
	1079	ν (C-O), ν (C-C)	O-Antigen
	1141	δ(C-H)	Lipid A
	1326	δ(C-H), Bending CH_2	O-Antigen,Core
	1381	$\delta(CH_3)$, ν (C-N)	O-Antigen
	1458	Scissoring CH_2/CH_3, δ(C-O-C) (glycosidic linkage), δ(C-C-C), δ (C-O-H)	Lipid A
	1617	ν (C-O), ν (C-C)	Lipid A
S. meliloti	568	$\beta(CH_2)$	ring
	656	δ(C-O-C) (glycosidic linkage), δ (C-C-C), δ(C-O-H)	O-Antigen
	732	β(C-O-C)	carbohydrates
	791	ν (C-O)	ring
	863	δ(C-C-H), δ(C-O-H)	O-Antigen
	894	δ(C-C-H), δ(C-O-H)	O-Antigen
	1033	ν (C-O), ν (C-C)	O-Antigen
	1092	ν(C-N)	O-Antigen
	1270	dC-O-C (glycosidic linkage)	Lipid A
	1333	δ(C-H)	Lipid A
	1479	δ(C-O-C) (glycosidic linkage), δ (C-C-C), δ(C-O-H)	O-Antigen
	1586	ν (C-O)	O-Antigen
	1642	ν (C-O), ν (C-C)	Lipid A
S. minnesota	577	$\beta(CH_2)$	ring
	668	δ(C-O-C) (glycosidic linkage), δ (C-C-C), δ(C-O-H)	O-Antigen
	808	δ(C-C-H), δ(C-O-H)	O-Antigen
	916	δ(C-C-H), δ(C-O-H)	O-Antigen
	1042	ν (C-O), ν (C-C)	O-Antigen
	1139	δ(C-H)	Lipid A
	1280	δ(C-O-C) (glycosidic linkage)	Lipid A
	1381	$\delta(CH_3)$, ν (C-N)	O-Antigen
	1612	ν (C-O), ν (C-C)	Lipid A
S. typhimurium	669	δ(C-O-C) (glycosidic linkage), δ (C-C-C), δ(C-O-H)	O-Antigen
	808	δ(C-C-H), δ(C-O-H)	O-Antigen
	855	δ(C-O-C)	O-Antigen
	915	δ(C-C-H), δ(C-O-H)	O-Antigen
	1005	ν (C-O), ν (C-C)	O-Antigen
	1045	ν (C-O), ν (C-C)	O-Antigen
	1105	ν (C-O)	O-Antigen
	1141	δ(C-H)	Lipid A
	1337	δ(C-H)	Lipid A
	1381	$\delta(CH_3)$, ν (C-N)	O-Antigen
	1592	ν (C-O)	O-Antigen
	1614	ν (C-O), ν (C-C)	Lipid A
Control samples			

PGN from *S. aureus*[14]	541	δ(C-O-C) (glycosidic linkage)	N-Acetylglucosamine
	734	β(C-O-C)	carbohydrates
	808	δ(C-C-H), δ(C-O-H)	Glycerolphosphate repeating units
	1039	v (C-O), v (C-C)	Glycerolphosphate repeating units
	1094	v (C-O)	Glycerolphosphate repeating units
	1336	δ(C-H)	Glycerolphosphate repeating units
	1462	δ(C-O-C) (glycosidic linkage), δ (C-C-C), δ(C-O-H)	N-Acetylglucosamine, Glycerolphosphate repeating units
	1584	v (C-O)	carbohydrates
	1615	v (C-O), v (C-C)	Glycerolphosphate repeating units
LTA from *B. subtilis*[15]	669	δ(C-O-C) (glycosidic linkage), δ (C-C-C), δ(C-O-H)	N-Acetylglucosamine and N-Acetylmuramic acid linked to pentapeptides
	860	δ(C-C-H), δ(C-O-H)	pentapeptides
	891	δ(C-C-H), δ(C-O-H)	pentapeptides
	930	δ(C-C-H), δ(C-O-H)	pentapeptides
	1005	v (C-O), v (C-C)	pentaglycyl segment
	1136	δ(C-H)	pentapeptides bridge link
	1334	δ(C-H)	pentapeptides bridge link
	1453	Scissoring CH_2/CH_3, δ(C-O-C) (glycosidic linkage), δ(C-C-C), δ (C-O-H)	N-Acetylglucosamine and N-Acetylmuramic acid linked to pentapeptides bridge link
	1508	$\delta(CH_3)$	carbohydrates
	1584	v (C-O)	carbohydrates
	1614	v (C-O), v (C-C)	pentaglycyl segment
Chitin[16]	555	$\beta(CH_2)$	ring
	775	v (C-O)	ring
	892	δ(C-C-H), δ(C-O-H)	Polysaccharide linear polymer of β-1,4-linked N-Acetylglucosamine (GlcNAc) unit
	930	δ(C-C-H), δ(C-O-H)	Polysaccharide linear polymer of β-1,4-linked N-Acetylglucosamine (GlcNAc) unit
	1006	v (C-O), v (C-C)	Polysaccharide linear

			polymer of β-1,4-linked N-Acetylglucosamine (GlcNAc) unit
	1043	v (C-O), v (C-C)	Polysaccharide linear polymer l of β-1,4-linked N-Acetylglucosamine (GlcNAc) unit
	1134	δ(C-H)	Lipid anchor
	1304	δ(C-H), β(CH$_2$)	Polysaccharide linear polymer of β-1,4-linked N-Acetylglucosamine (GlcNAc) unit
	1449	Scissoring CH$_2$/CH$_3$, δ(C-O-C) (glycosidic linkage), δ(C-C-C), δ (C-O-H)	Polysaccharide linear polymer of β-1,4-linked N-Acetylglucosamine (GlcNAc) unit
	1614	v (C-O), v (C-C)	Lipid anchor

β, bending; δ, deformation; τ, twisting; v, stretching.

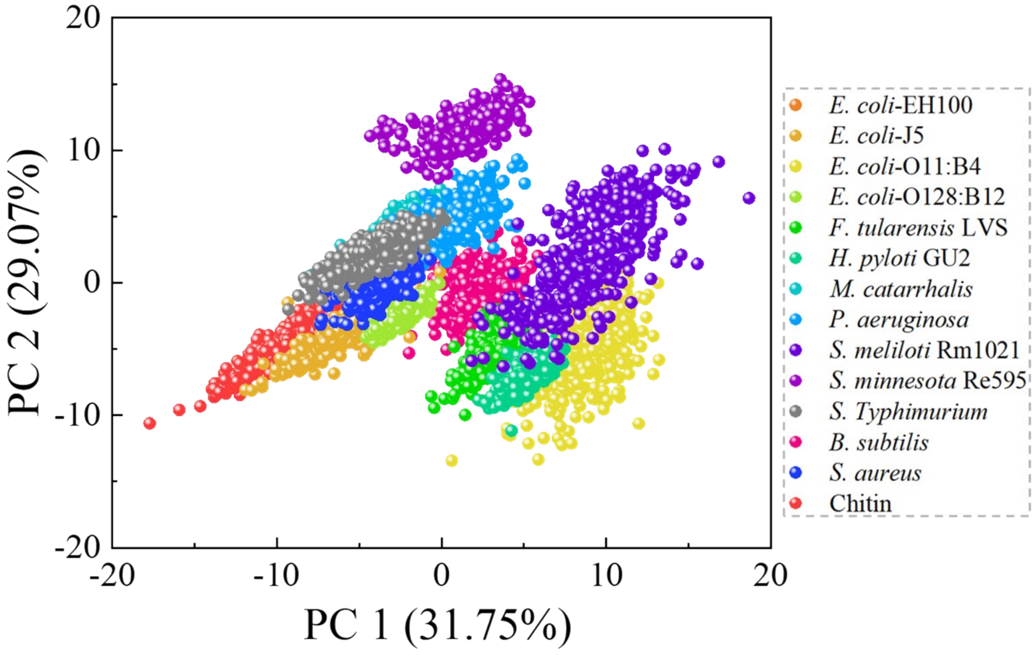

Fig. S6 The 2D PCA plot for SERS spectra of eleven bacterial endotoxin samples and three reference samples.

Table S7. SERS peak assignment for LPSs.

SERS peaks	Peak assignment	*E. coli*-EH100	*E. coli*-J5	*E. coli*-O11:B4	*E. coli*-O128:B12	*F. tularensis* LVS	*H. pyloti* GU2	*M. catarrhalis*	*P. aeruginosa*	*S. meliloti* Rm1021	*S. minnesota* Re595	*S. typhimurium*	PGN from *B. subtilis*	LTA from *S. aureus*	Chitin	Number of common peaks
531	δ(C-O-C) (glycosidic linkage)					🟦										1
542	δ(C-O-C) (glycosidic linkage)						🟦		🟦				🟨			3
560	$\beta(CH_2)$							🟦							🟨	2
568	$\beta(CH_2)$									🟦						1
577	$\beta(CH_2)$										🟦					1
648	δ(C-O-C) (glycosidic linkage), δ(C-C-C), δ(C-O-H)								🟦							1
656	δ(C-O-C) (glycosidic linkage), δ(C-C-C), δ(C-O-H)						🟦			🟦						2
669	δ(C-O-C) (glycosidic linkage), δ(C-C-C), δ(C-O-H), β(C-O-C)	🟦	🟦		🟦	🟦		🟦			🟦	🟦		🟨		8
735	β(C-O-C), Symmetric v $(CH_3)_3$			🟦	🟦	🟦	🟦		🟦	🟦			🟨			7
759	Symmetric v $(CH_3)_3$							🟦								1
775	v (C-O)														🟨	1
794	v (C-O)			🟦	🟦	🟦	🟦		🟦	🟦						6
808	δ(C-C-H), δ(C-O-H)	🟦									🟦	🟦	🟨			4
852	v (C-O-C)	🟦										🟦				2
857	δ(C-C-H), δ(C-O-H)			🟦					🟦	🟦				🟨		4
894	δ(C-C-H), δ(C-O-H)		🟦			🟦	🟦	🟦	🟦	🟦				🟨	🟨	8
917	δ(C-C-H), δ(C-O-H)	🟦			🟦						🟦	🟦				4

1003	v (C-C), v (C-O)	🟦	🟦	🟦	🟦	🟦		🟦	🟦			🟦		🟨	🟨	10
1033	v (C-C), v (C-O)			🟦	🟦	🟦	🟦	🟦		🟦			🟨			7
1047	v (C-C), v (C-O)	🟦							🟦		🟦	🟦			🟨	5
1079	v (C-C), v (C-O)				🟦				🟦							2
1095	v (C-N), v (C-O)			🟦						🟦		🟦	🟨			4
1136	δ(C-H)	🟦						🟦	🟦		🟦	🟦		🟨	🟨	7
1174	δ(C-O-C) (glycosidic linkage)							🟦								1
1270	δ(C-O-C) (glycosidic linkage)									🟦	🟦					2
1304	δ(C-H), Bending CH$_2$														🟨	1
1326	δ(C-H), Bending CH$_2$								🟦							1
1333	δ(C-H)	🟦	🟦	🟦	🟦	🟦	🟦	🟦		🟦		🟦	🟨	🟨		11
1387	δ(CH$_3$), v (C-N)	🟦	🟦		🟦				🟦		🟦	🟦				6
1453	Scissoring CH$_2$/CH$_3$, δ (C-O-C) (glycosidic linkage), δ(C-C-C), δ(C-O-H)	🟦						🟦	🟦					🟨	🟨	5
1469	δ(C-O-C) (glycosidic linkage), δ(C-C-C), δ(C-O-H)			🟦	🟦	🟦	🟦			🟦			🟨			6
1508	v(C-N), δ(CH$_3$)			🟦		🟦	🟦							🟨		4
1578	v (C-O), N-H bendinng						🟦									1
1592	v (C-O)			🟦	🟦	🟦				🟦		🟦	🟨	🟨		7
1614	v (C-C) v (C-O)	🟦	🟦	🟦	🟦			🟦	🟦		🟦	🟦	🟨	🟨	🟨	11
1642	v (C-C) v (C-O)									🟦						1

S6. Additional description of confusion matrix

Predicted Actual	Label 1	Label 2
Label 1	True positive (TP)	False Negative (FN)
Label 2	False Positive (FP)	True Negative (TN)

Fig. S7 Confusion matrix example.

Table S8 lists the seven performance measures used to evaluate the ML models. Accuracy calculates the proportion of correct predictions. Precision is the ratio of correct positive predictions to all positive predictions. Recall is the ratio of correct positive predictions to all positive observations. F1-score is the weighted average of Precision and Recall. Micro and Macro averages are two types of average of measures for multi-class setting. Micro average pools the performance over all samples, while macro calculates the measures for each class and average them. For example, Micro Precision pools all TP and FP over all classes and calculate the ratio of correct positive predictions to all positive predictions with pooled TP and FP, while Micro Precision calculates precision for all classes individually and then take the average.

Table S8. Equations for accuracy, precision, recall, F1 score.

	Equations
Accuracy	$Accuracy = \frac{TP + TN}{Total}$
Micro Precision	$Precision_{Micro} = \frac{\sum_{i=1}^{N} TP_i}{N \qquad N}$
Macro Precision	$Precision_{Macro} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i}$
Micro Recall	$Recall_{Micro} = \frac{\sum_{i=1}^{N} TP_i}{N \qquad N}$
Macro Recall	$Recall_{Macro} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FN_i}$
Micro F1-score	$F1_{Micro} = \frac{2 \times Precision_{Micro} \times Recall_{Micro}}{Precision_{Micro} + Recall_{Micro}}$
Macro F1-score	$F1_{Macro} = \frac{1}{N}\sum_{i=1}^{N}\frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i}$

Footnote: N = 14 labels; TP: True Positives; TN: True Negatives; FP: False Positives; FN: False Negatives.

S7. Additional results from MLAs

Table S9. Important features from SVM model.

Important features (SVM)	Peak assignment	*E. coli*-EH100	*E. coli*-J5	*E. coli*-O11:B4	*E. coli*-O128:B12	*F. tularensis* LVS	*H. pyloti* GU2	*M. catarrhalis*	*P. aeruginosa*	*S. meliloti* Rm1021	*S. minnesota* Re595	*S. typhimurium*	*B. subtilis* (PGN)	*S. aureus* (LTA)	Chitin	**Number of peaks**
670	δ (C-O-C) (glycosidic linkage), δ(C-C-C), δ(C-O-H)	🟦	🟦		🟦			🟦			🟦	🟦		🟨		7
734	β(C-O-C)			🟦	🟦	🟦	🟦		🟦	🟦			🟨			7
795	ν (C-O)			🟦	🟦	🟦	🟦		🟦	🟦						6
857	δ(C-C-H), δ(C-O-H)	🟦		🟦					🟦			🟦		🟨		5
930	δ(C-C-H), δ(C-O-H)													🟨	🟨	2
1003	ν (C-O), ν (C-C)	🟦	🟦	🟦	🟦	🟦		🟦	🟦			🟦		🟨	🟨	10
1048	ν (C-O), ν (C-C)	🟦									🟦	🟦			🟨	4
1333	δ(C-H)		🟦	🟦	🟦	🟦	🟦	🟦		🟦		🟦	🟨	🟨		10
1381	$\delta(CH_3)$, ν (CN)[17]	🟦			🟦			🟦	🟦		🟦	🟦			🟨	7
1593	ν (C-O)									🟦		🟦				2
1617	ν (C-O), ν (C-C)	🟦	🟦	🟦	🟦			🟦	🟦		🟦	🟦	🟨	🟨	🟨	11
Number of peaks		6	4	6	7	4	3	5	6	4	4	8	3	6	5	

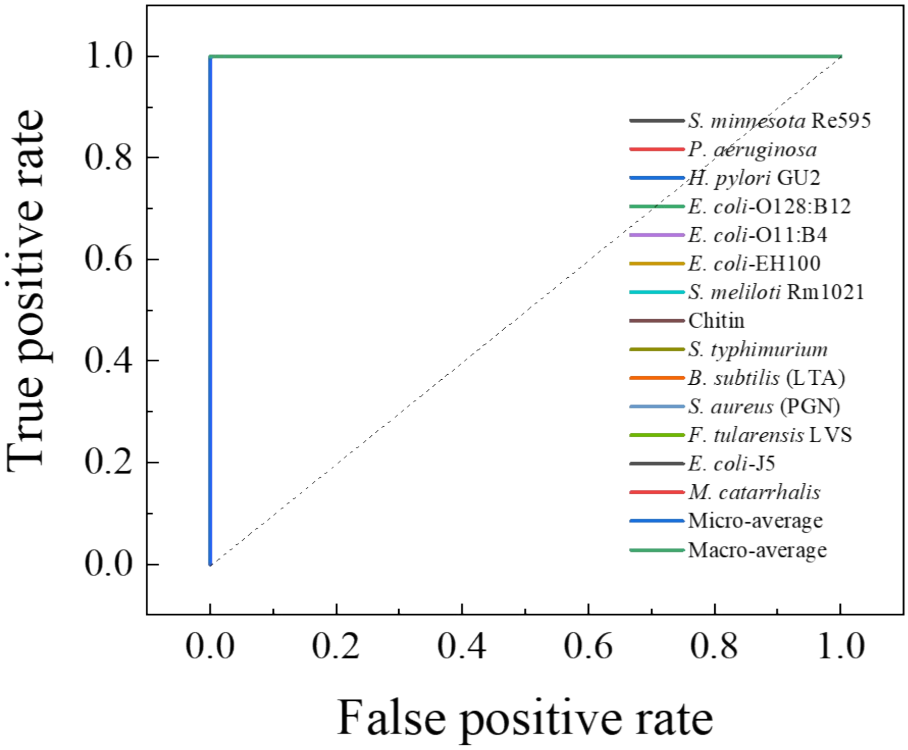

Fig. S8 The ROC curves using One-vs-Rest scheme for each LPS and Micro- and Macro-average ROC on all LPSs for the SVM classifier.

S8. Additional results from MLAs for LPS mixtures

Table S10. List of LPS mixtures and their biological significance.

	Mixture	Number of SERS spectra	Analyte 1	Analyte 2	Biological significance	Ref
1	ES	396	*E. coli*-O11:B4	*S. minnesota* Re595	Both pathogens are common cause of food borne infections and harbor biologically active LPS.	[18, 19]
2	FH	422	*F. tularensis* LVS	*H. pylori* GU2	*H. pylori* is common cause of gastritis affecting 50% of world population. In contrast, *F. tularensis* causes tularemia which is rare but serious infection. Both pathogens' LPS structures are biologically inactive and share similarities, yet they cause very different diseases in human.	[20, 21]
3	EH	409	*E. coli*-O11:B4	*H. pylori* GU2	Both pathogens are commonly causing disease in human, but their LPS structures vary greatly.	[18, 21]
4	SH	423	*S. minnesota* Re595	*H. pylori* GU2	Both pathogens are commonly causing disease in human, but their LPS structures vary greatly.	[21, 22]

Table S11. The accuracy obtained from different ML models using SERS spectra from single LPSs and mixtures.

Models	Accuracy
RamanNet	1.000 ± 0.000
SVM	0.997 ± 0.002
RF	0.995 ± 0.002
KNN	0.991 ± 0.003
LDA	0.9988 ± 0.0008
PLS-DA	0.928 ± 0.005

Fig. S9 The confusion matrix of the RamanNet model from SERS spectra of single LPS (11 LPS and 3 control samples) and 4 mixtures. Entries in the matrix represents the percentage of test spectra that are predicted by the SVM model as class (first row) given a ground truth of class (first column); entries along the diagonal represent the accuracies for each class.**(Below)**

Predicted **True**	*E. coli*-EH100	*E. coli*-J5	*E. coli*-O11:B4	*E. coli*-O128:B12	*F. tularensis* LVS	*H. pylori* GU2	*M. catarrhalis*	*P. aeruginosa*	*S. meliloti* Rm1021	*S. minnesota* Re595	*S. Typhimurium*	*S. aureus*	*B. subtilis*	Chitin	ES	FH	HE	HS
E. coli-EH100	100																	
E. coli-J5		100																
E. coli-O11:B4			100															
E. coli-O128:B12				100														
F. tularensis LVS					100													
H. pylori GU2						100												
M.catarrhalis							100											
P. aeruginosa								100										
S. meliloti Rm1021									100									
S. minnesota Re595										100								
S. Typhimurium											100							
S. aureus												100						
B. subtilis													100					
Chitin														100				
ES															100			
FH																100		
HE																	100	
HS																		100

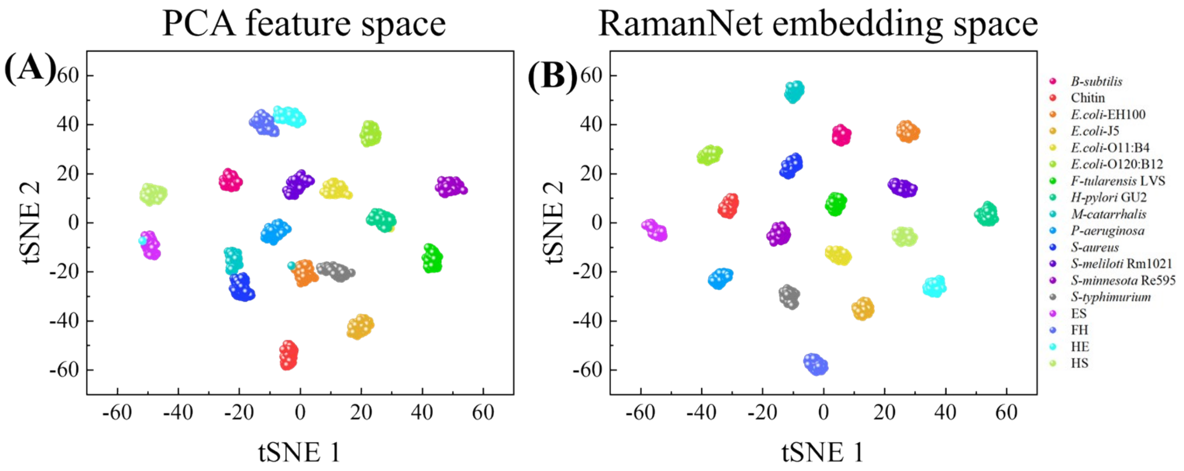

Fig. S10 The 256-dimensional feature space is projected into a 2 dimensional-map using tSNE based on SERS spectra from single LPS and mixture: (A) PCA feature space, (B) RamanNet embedding space.

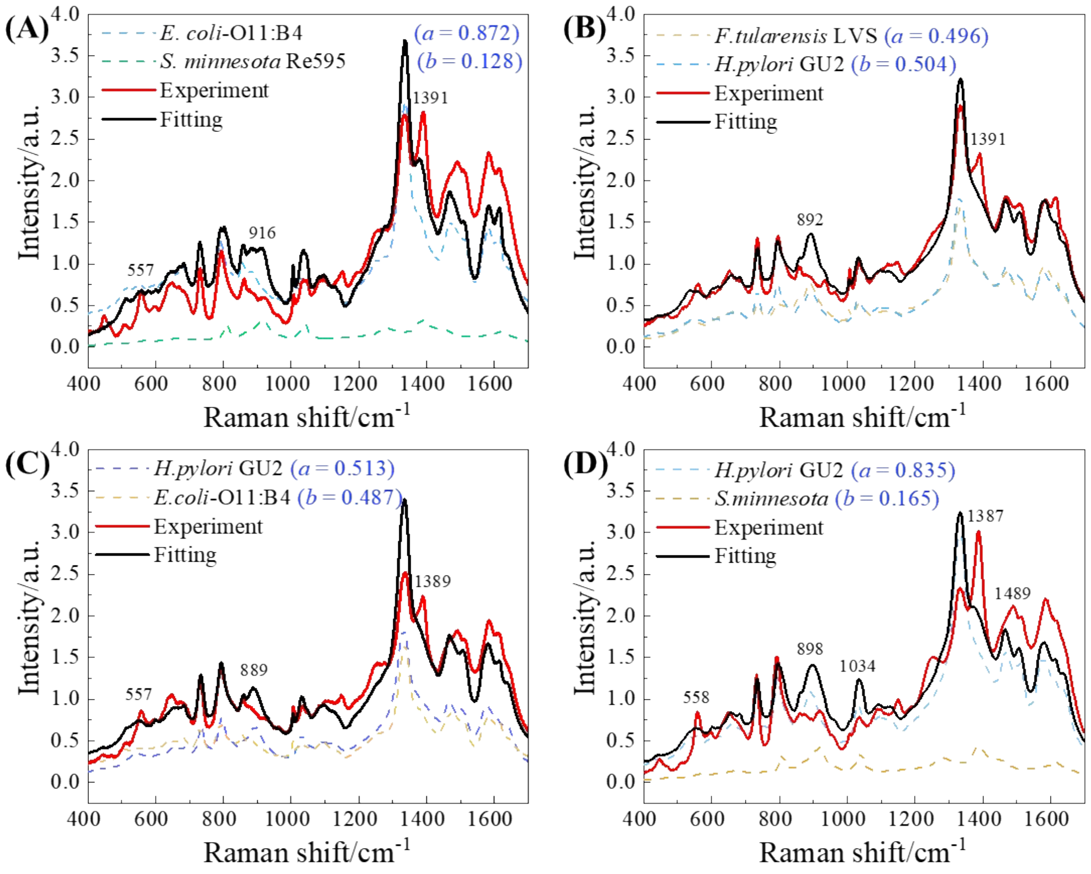

Fig. S11 Typical average SERS spectra (red curves) of four bacterial endotoxin mixtures and the linear combinations (black curves) of SERS spectra of the corresponding two LPSs: (A) *E. coli*-O11:B4 and *S. minnesota* Re595 (ES), (B) *F. tularensis* LVS and *H. pylori* GU2 (FH), (C) *E. coli*-O11:B4 and *H. pylori* GU2 (EH), and (D) *S. minnesota* Re595 and *H. pylori* GU2 (SH). Dash lines are the SERS contributions from each LPS.

References:

1. Fleckenstein, J. M.; Matthew Kuhlmann, F.; Sheikh, A., Acute Bacterial Gastroenteritis. *Gastroenterology Clinics of North America* **2021,** *50* (2), 283-304.

2. Cusack, R.; Garduno, A.; Elkholy, K.; Martín-Loeches, I., Novel investigational treatments for ventilator-associated pneumonia and critically ill patients in the intensive care unit. *Expert Opinion on Investigational Drugs* **2022,** *31* (2), 173-192.

3. https://ahpsr.who.int/publications/i/item/global-action-plan-on-antimicrobial-resistance.

4. https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance.

5. Lee, A. S.; de Lencastre, H.; Garau, J.; Kluytmans, J.; Malhotra-Kumar, S.; Peschel, A.; Harbarth, S., Methicillin-resistant Staphylococcus aureus. *Nature Reviews Disease Primers* **2018,** *4* (1), 18033.

6. Logan, N. A., Bacillus and relatives in foodborne illness. *Journal of Applied Microbiology* **2012,** *112* (3), 417-429.

7. Hinton, G.; Roweis, S., Stochastic neighbor embedding. In *Proceedings of the 15th International Conference on Neural Information Processing Systems*, MIT Press: 2002; pp 857–864.

8. Osorio-Román, I. O.; Aroca, R. F.; Astudillo, J.; Matsuhiro, B.; Vásquez, C.; Pérez, J. M., Characterization of bacteria using its O-antigen with surface-enhanced Raman scattering. *Analyst* **2010,** *135* (8), 1997-2001.

9. Stromberg, L. R.; Mendez, H. M.; Mukundan, H., Detection methods for lipopolysaccharides: past and present. *Escherichia coli-recent advances on physiology, pathogenesis biotechnological applications. InTech* **2017**, 141-168.

10. Wu, X.; Zhao, Y.; Zughaier, S. M., Highly Sensitive Detection and Differentiation of Endotoxins Derived from Bacterial Pathogens by Surface-Enhanced Raman Scattering. *Biosensors* **2021,** *11* (7), 234.

11. Mangini, M.; Verde, A.; Boraschi, D.; Puntes, V. F.; Italiani, P.; De Luca, A. C., Interaction of nanoparticles with endotoxin Importance in nanosafety testing and exploitation for endotoxin binding. *Nanotoxicology* **2021,** *15* (4), 558-576.

12. Töpfer, N.; Müller, M. M.; Dahms, M.; Ramoji, A.; Popp, J.; Slevogt, H.; Neugebauer, U., Raman spectroscopy reveals LPS-induced changes of biomolecular composition in monocytic THP-1 cells in a label-free manner. *Integrative Biology* **2019,** *11* (3), 87-98.

13. Alessandro, V.; Maria, M.; Stefano, M.; Diana, B.; Paola, I.; Anna Chiara De, L. In SERS-based nanotoxicology assessment of gold nanoparticles, *Proc. SPIE 11786, Optical Methods for Inspection, Characterization, and Imaging of Biomaterials V*, **2021,** 117861G.

14. Sharif, S.; Singh, M.; Kim, S. J.; Schaefer, J., Staphylococcus aureus Peptidoglycan Tertiary Structure from Carbon-13 Spin Diffusion. *Journal of the American Chemical Society* **2009,** *131* (20), 7023-7030.

15. Villéger, R.; Saad, N.; Grenier, K.; Falourd, X.; Foucat, L.; Urdaci, M. C.; Bressollier, P.; Ouk, T.-S., Characterization of lipoteichoic acid structures from three probiotic Bacillus strains: involvement of D-alanine in their biological activity. *Antonie van Leeuwenhoek* **2014,** *106* (4), 693-706.

16. Fernando, L. D.; Dickwella Widanage, M. C.; Penfield, J.; Lipton, A. S.; Washton, N.; Latgé, J.-P.; Wang, P.; Zhang, L.; Wang, T., Structural Polymorphism of Chitin and Chitosan in Fungal Cell Walls From Solid-State NMR and Principal Component Analysis. *Frontiers in*

Molecular Biosciences **2021,** *8*, https://doi.org/10.3389/fmolb.2021.727053.

17. Nowicka, A. B.; Czaplicka, M.; Kowalska, A. A.; Szymborski, T.; Kamińska, A., Flexible PET/ITO/Ag SERS Platform for Label-Free Detection of Pesticides. *Biosensors* **2019,** *9* (3), 111.

18. Caroff, M.; Novikov, A., Lipopolysaccharides: structure, function and bacterial identifications. *OCL* **2020,** *27*, 31.

19. Xiao, X.; Sankaranarayanan, K.; Khosla, C., Biosynthesis and structure–activity relationships of the lipid a family of glycolipids. *Current Opinion in Chemical Biology* **2017,** *40*, 127-137.

20. Okan, N. A.; Kasper, D. L., The atypical lipopolysaccharide of Francisella. *Carbohydrate Research* **2013,** *378*, 79-83.

21. Li, H.; Liao, T.; Debowski, A. W.; Tang, H.; Nilsson, H.-O.; Stubbs, K. A.; Marshall, B. J.; Benghezal, M., Lipopolysaccharide Structure and Biosynthesis in Helicobacter pylori. *Helicobacter* **2016,** *21* (6), 445-461.

22. Janusch, H.; Brecker, L.; Lindner, B.; Alexander, C.; Gronow, S.; Heine, H.; Ulmer, A. J.; Rietschel, E. T.; Zähringer, U., Structural and biological characterization of highly purified hepta-acyl lipid A present in the lipopolysaccharide of the Salmonella enterica sv. Minnesota Re deep rough mutant strain R595. *Journal of Endotoxin Research* **2002,** *8* (5), 343-356.