Supplementary Material for

Machine learning encodes urine and serum metabolic patterns for autoimmune diseases discrimination, classification and metabolic dysregulation analysis

Qiuyao Du[a,b], Xiao Wang[a,b], Junyu Chen[a,b], Yiran Wang[a,b], Wenlan Liu[c], Liping Wang[c], Huihui Liu[a,b,*], Lixia Jiang[d,*] & Zongxiu Nie[a,b,*]

[a]Beijing National Laboratory for Molecular Sciences, Key Laboratory of Analytical Chemistry for Living Biosystems, Institute of Chemistry, Chinese Academy of Sciences, Beijing 100190, China.
[b]University of Chinese Academy of Sciences, Beijing 100049, China.
[c]Shenzhen Second People's Hospital, The First Affiliated Hospital of Shenzhen University, Shenzhen 518000, China.
[d]Department of Laboratory Medicine, First Affiliated Hospital of Gannan Medical University, Ganzhou, Jiangxi Province 341000, China.

Correspondence should be addressed to Huihui Liu; hhliu@iccas.ac.cn, Lixia Jiang; jlx7310@gmu.edu.cn and Zongxiu Nie; znie@iccas.ac.cn

This part included:

Data processing and statistical analysis

Figures S1-S3

Tables S1-S23

Data processing and statistical analysis

Data extraction and alignment were carried out using MZmine-2.53, which was obtained at https://github.com/mzmine/mzmine2/releases/. The mass detection was performed on raw data and exact masses with mass level 1 by keeping the noise level at 3,000. Chromatograms were built using an ADAP module with a minimum height of 3,000, and *m/z* tolerance of 0.05 (or 20 ppm). For the chromatogram deconvolution, the local minimum search algorithm was used with the following settings: chromatographic threshold = 5%, minimum retention time range = 0.50 min, minimum relative height = 30%, minimum absolute height = 10,000, minimum ratio of the peak top/edge = 1.3, and peak duration range = 0.0-50.0 min[1]. After processing the HPLC-MS data, a total of 7326 RT-*m/z* ion peaks were found in urine samples and 6145 peaks in serum samples. The number of occurrences of each *m/z* in all spectra was calculated, and those less than 2/3 were eliminated. After selection, 572 *m/z* features of ion peaks for each spectrum of urine samples and 543 *m/z* features for serum samples were used as a basis for subsequent statistical analysis.

Machine learning was carried out with the Orange 3.31.1 module in Python 3.7. The build-in classifier neural network (NN), random forest (RF), logistic regression (LR), naive bayes (NB), support vector machine (SVM), adaboost (AB) and k-nearest neighbor (kNN) were applied. Model parameters were set as follows: NN: Hidden layers: 100, Activation: ReLu, Solver: Adam, Alpha: 0.0001, Max iterations: 200, Replicable training: True. RF: Number of trees: 10, Maximal number of considered features: unlimited, Replicable training: No, Maximal tree depth: unlimited, Stop splitting nodes with maximum instances: 5. LR: Regularization: Ridge (L2), C=1, class weights=False. NB: No additional parameter settings are performed. SVM: SVM type: SVM, C=1.0, ε=0.1, Kernel: RBF, exp(-auto|x-y|²), Numerical tolerance: 0.001, Iteration limt: 100. AB: Base estimator: tree, Number of estimators: 50, Algorithm (classification): Samme.r, Loss (regression): Linear. kNN: Number of neighbours: 5, Metric: Euclidean, Weight: Uniform.

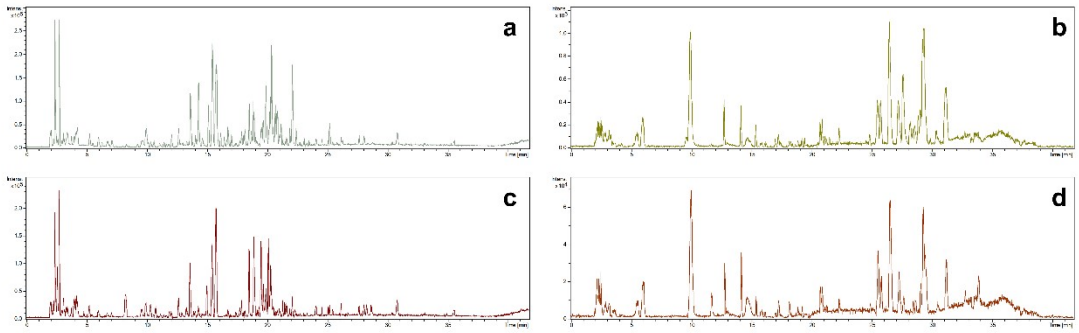

Figure S1. Total ion chromatograms (TIC) of healthy controls (HC) (a) and autoimmune diseases (ADs) patients (c) from urine samples; TIC of HC (b) and ADs patients (d) from serum samples.

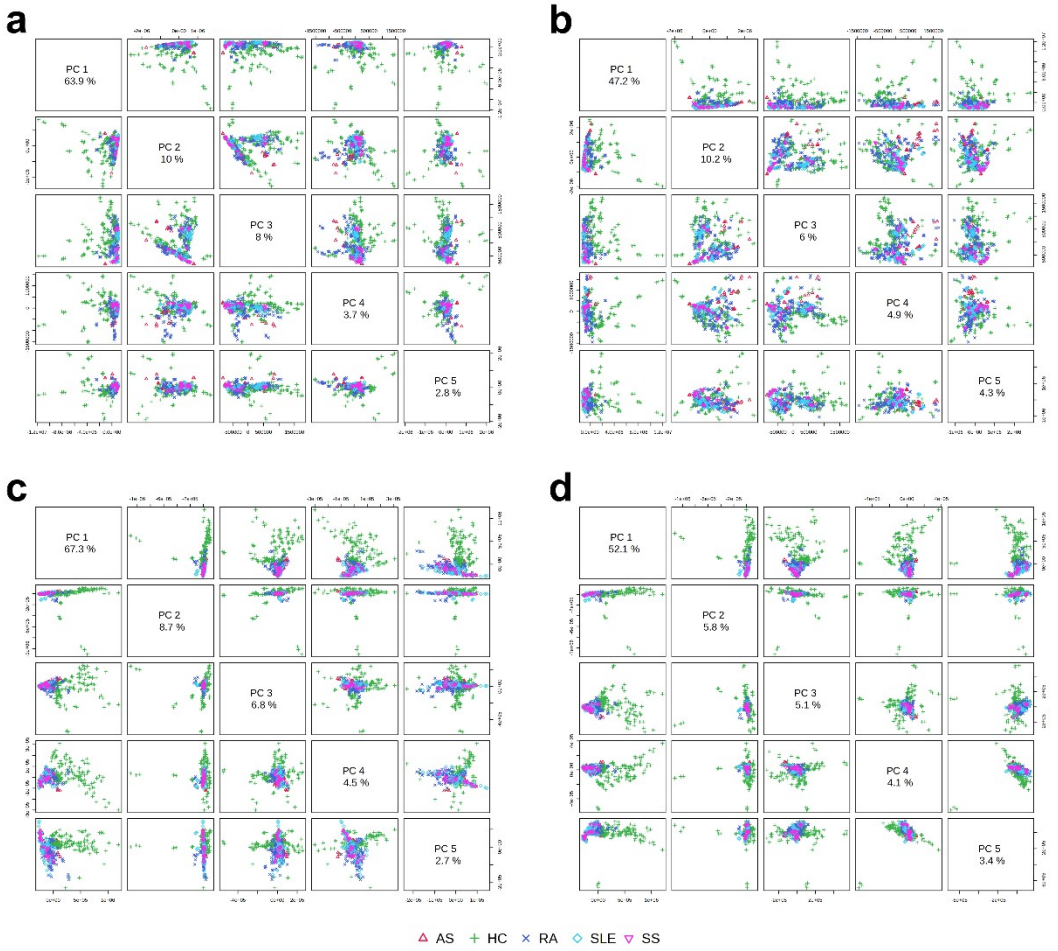

Figure S2. Overview of principal component analysis (PCA) results in the classification of autoimmune diseases (ADs) and healthy controls (HC). AS (ankylosing spondylitis), RA (rheumatoid arthritis), SLE (systemic lupus erythematosus), SS (sicca syndrome). PCA results with differential metabolites (a), and all *m/z* features (c) in urine samples; PCA results with differential metabolites (b), and all *m/z* features (d) in serum samples.

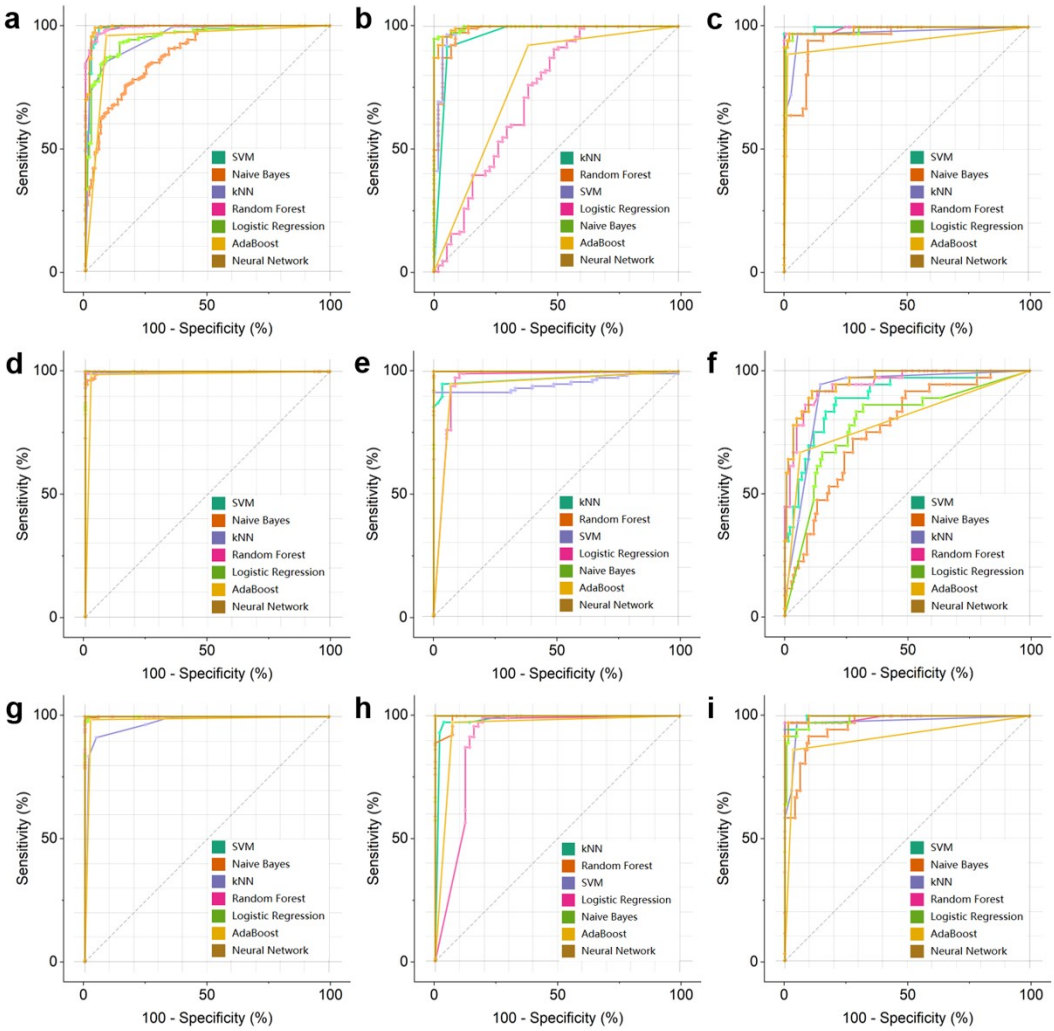

Figure S3. ROC curves of different classifiers with metabolites panels. ROC curves for autoimmune diseases (ADs) versus healthy controls (HC) in training cohort (a), testing cohort (b), and the classification of ADs and HC (AS vs SS vs SLE vs RA vs HC) (c) of metabolites panel with urine samples; ROC curves for ADs versus HC in training cohort (d), testing cohort (e), and the classification of ADs and HC (AS vs SS vs SLE vs RA vs HC) (f) of metabolites panel with serum samples; ROC curves for ADs versus HC in training cohort (g), testing cohort (h), and the classification of ADs and HC (AS vs SS vs SLE vs RA vs HC) (i) of metabolites panel with fusion model. AS (ankylosing spondylitis), SS (sicca syndrome), SLE (systemic lupus erythematosus), RA (rheumatoid arthritis).

Table S1 Age and gender distribution of study participants.

Groups	Number	Age						Gender	
		< 30	30-39	40-49	50-59	60-69	≥70	Male	Female
Healthy controls	62	8	5	14	19	8	8	28	34
Autoimmune diseases	121	11	12	21	40	19	18	39	82

Table S2 Metrics of classifiers for autoimmune diseases versus healthy controls (ADs vs HC) with

urine samples by LC-MS (training cohort).

Model	AUC	Accuracy	F1	Precision	Recall
Logistic Regression	1.000	1.000	1.000	1.000	1.000
Neural Network	1.000	0.997	0.997	0.997	0.997
Support Vector Machine	1.000	0.997	0.997	0.997	0.997
Random Forest	1.000	0.997	0.997	0.997	0.997
AdaBoost	0.996	0.997	0.997	0.997	0.997
Naive Bayes	0.995	0.957	0.957	0.958	0.957
k-Nearest Neighbor	0.968	0.861	0.857	0.862	0.861

Table S3 Metrics of classifiers for autoimmune diseases versus healthy controls (ADs vs HC) with urine samples by LC-MS (testing cohort).

Model	AUC	Accuracy	F1	Precision	Recall
Naive Bayes	1.000	0.989	0.988	0.989	0.989
Neural Network	1.000	0.960	0.959	0.962	0.960
Support Vector Machine	0.995	0.828	0.807	0.863	0.828
Random Forest	0.980	0.862	0.850	0.886	0.862
k-Nearest Neighbor	0.924	0.897	0.891	0.910	0.897
Logistic Regression	0.861	0.833	0.830	0.830	0.833
AdaBoost	0.614	0.747	0.688	0.816	0.747

Table S4 Metrics of classifiers for autoimmune diseases versus healthy controls (ADs vs HC) with serum samples by LC-MS (training cohort).

Model	AUC	Accuracy	F1	Precision	Recall
Random Forest	1.000	1.000	1.000	1.000	1.000
Logistic Regression	1.000	0.997	0.997	0.997	0.997
Support Vector Machine	1.000	0.995	0.995	0.995	0.995
k-Nearest Neighbor	1.000	0.984	0.984	0.985	0.984
Neural Network	0.999	0.995	0.995	0.995	0.995
Naive Bayes	0.979	0.957	0.958	0.960	0.957
AdaBoost	0.978	0.981	0.981	0.981	0.981

Table S5 Metrics of classifiers for autoimmune diseases versus healthy controls (ADs vs HC) with serum samples by LC-MS (testing cohort).

Model	AUC	Accuracy	F1	Precision	Recall
Neural Network	0.999	0.862	0.866	0.903	0.862
AdaBoost	0.987	0.983	0.983	0.984	0.983
Logistic Regression	0.982	0.943	0.943	0.945	0.943
k-Nearest Neighbor	0.972	0.885	0.888	0.910	0.885
Random Forest	0.972	0.845	0.849	0.878	0.845
Support Vector Machine	0.874	0.759	0.759	0.759	0.759
Naive Bayes	0.808	0.713	0.717	0.847	0.713

Table S6 Metrics of classifiers for autoimmune diseases versus healthy controls (ADs vs HC) with

fusion model by LC-MS (training cohort).

Model	AUC	Accuracy	F1	Precision	Recall
Random Forest	1.000	1.000	1.000	1.000	1.000
Support Vector Machine	1.000	1.000	1.000	1.000	1.000
Neural Network	1.000	0.997	0.997	0.997	0.997
Logistic Regression	1.000	0.997	0.997	0.997	0.997
Naive Bayes	0.990	0.987	0.987	0.987	0.987
AdaBoost	0.982	0.984	0.984	0.984	0.984
k-Nearest Neighbor	0.977	0.885	0.882	0.887	0.885

Table S7 Metrics of classifiers for autoimmune diseases versus healthy controls (ADs vs HC) with fusion model by LC-MS (testing cohort).

Model	AUC	Accuracy	F1	Precision	Recall
Random Forest	1.000	0.989	0.989	0.989	0.989
Neural Network	1.000	0.868	0.872	0.906	0.868
AdaBoost	0.987	0.983	0.983	0.984	0.983
Support Vector Machine	0.978	0.948	0.949	0.955	0.948
Logistic Regression	0.977	0.948	0.949	0.950	0.948
k-Nearest Neighbor	0.963	0.954	0.953	0.957	0.954
Naive Bayes	0.850	0.799	0.805	0.875	0.799

Table S8 Metrics of classifiers for the distinction of four autoimmune diseases (ADs) and healthy controls (HC) with urine samples by LC-MS.

Model	AUC	Accuracy	F1	Precision	Recall
Random Forest	1.000	0.989	0.989	0.989	0.989
Support Vector Machine	0.999	0.928	0.926	0.937	0.928
Logistic Regression	0.998	0.983	0.983	0.983	0.983
Neural Network	0.996	0.994	0.994	0.995	0.994
AdaBoost	0.983	0.972	0.972	0.973	0.972
Naive Bayes	0.965	0.822	0.822	0.829	0.822
k-Nearest Neighbor	0.959	0.678	0.679	0.686	0.678

AS vs SS vs SLE vs RA vs HC, AS: ankylosing spondylitis, SS: sicca syndrome, SLE: systemic lupus erythematosus, RA: rheumatoid arthritis.

Table S9 Metrics of classifiers for the distinction of four autoimmune diseases (ADs) and healthy

controls (HC) with serum samples by LC-MS.

Model	AUC	Accuracy	F1	Precision	Recall
Logistic Regression	0.993	0.967	0.967	0.967	0.967
Neural Network	0.993	0.956	0.955	0.957	0.956
Support Vector Machine	0.989	0.878	0.875	0.882	0.878
Random Forest	0.976	0.894	0.895	0.900	0.894
k-Nearest Neighbor	0.953	0.689	0.685	0.695	0.689
Naive Bayes	0.915	0.706	0.699	0.726	0.706
AdaBoost	0.865	0.783	0.782	0.785	0.783

AS vs SS vs SLE vs RA vs HC, AS: ankylosing spondylitis, SS: sicca syndrome, SLE: systemic lupus erythematosus, RA: rheumatoid arthritis.

Table S10 Metrics of classifiers for the distinction of four autoimmune diseases (ADs) and healthy controls (HC) with fusion model by LC-MS.

Model	AUC	Accuracy	F1	Precision	Recall
Naive Bayes	1.000	0.981	0.981	0.983	0.981
Random Forest	0.997	0.962	0.962	0.962	0.962
Logistic Regression	0.996	0.990	0.990	0.991	0.990
Support Vector Machine	0.990	0.876	0.880	0.924	0.876
k-Nearest Neighbor	0.951	0.657	0.655	0.690	0.657
AdaBoost	0.935	0.895	0.895	0.896	0.895
Neural Network	0.865	0.886	0.883	0.895	0.886

AS vs SS vs SLE vs RA vs HC, AS: ankylosing spondylitis, SS: sicca syndrome, SLE: systemic lupus erythematosus, RA: rheumatoid arthritis.

Table S11 Differential metabolites annotated in urine samples.

Experimental *m/z*	Theoretical *m/z*	Delta (ppm)	Formula	Adduct	Compound Name	Identification Database
205.0969	205.0972	1.4209	$C_{11}H_{12}N_2O_2$	$[M+H]^+$	L-Tryptophan	HMDB0000929
207.1103	207.1104	0.5286	$C_7H_{13}NO_2$	$[M+ACN+Na]^+$	Proline betaine	HMDB0004827
211.0574	211.0585	4.9780	$C_4H_{10}N_3O_5P$	$[M+NH_4-H_2O]^+$	Phosphocreatine	HMDB0001511
211.1694	211.1693	0.4881	$C_{13}H_{22}O_2$	$[M+H]^+$	Methyl (2E,6Z)-dodecadienoate	HMDB0031014
218.0121	218.0112	4.1292	$C_7H_6O_6S$	$[M+NH_4-H_2O]^+$	5-Sulfosalicylic acid	HMDB0011725
221.0919	221.0921	0.6809	$C_{11}H_{12}N_2O_3$	$[M+H]^+$	5-Hydroxy-L-tryptophan	HMDB0000472
225.1094	225.1090	1.8014	$C_8H_{17}NOS_2$	$[M+NH_4]^+$	Dihydrolipoamide	HMDB0000985
227.1251	227.1254	1.2909	$C_{10}H_{20}O_4$	$[M+Na]^+$	3,7-Dimethyl-3-octene-1,2,6,7-tetrol	HMDB0035133
229.1541	229.1547	2.7580	$C_{11}H_{20}N_2O_3$	$[M+H]^+$	Isoleucylproline	HMDB0011174
231.1588	231.1591	1.3055	$C_{12}H_{22}O_4$	$[M+H]^+$	Dodecanedioic acid	HMDB0000623
241.1544	241.1547	1.3553	$C_{12}H_{17}NO_3$	$[M+NH_4]^+$	Cerulenin	HMDB0015168
248.2111	248.2111	0.1014	$C_{30}H_{58}O_2$	$[M+2Na]^{2+}$	28-Methyl-27-nonacosenoic acid	HMDB0031951
249.1068	249.1070	0.7566	$C_4H_7N_3O$	$[2M+Na]^+$	Creatinine	HMDB0000562
249.2143	249.2154	4.3137	$C_{45}H_{86}O_6$	$[M+2H+Na]^{3+}$	TG(19:0/15:0/8:0)	HMDB0110080
267.1630	267.1634	1.4040	$C_{42}H_{71}O_{12}P$	$[M+3H]^{3+}$	PG(16:1(9Z)/PGJ2)	HMDB0268691
271.1648	271.1644	1.3439	$C_{14}H_{26}O_2$	$[M+2Na-H]^+$	Myristoleic acid	HMDB0002000
274.1035	274.1026	3.2291	$C_{11}H_{19}NO_4$	$[M+2Na-H]^+$	Butenylcarnitine	HMDB0013126
286.2010	286.2009	0.4168	$C_{31}H_{58}O_6$	$[M+2Na]^{2+}$	TG(8:0/8:0/12:0)	HMDB0072057
287.2037	287.2052	5.0997	$C_{48}H_{88}O_8$	$[M+3Na]^{3+}$	DG(a-25:0/PGF2alpha/0:0)	HMDB0298162
293.1464	293.1461	0.9989	$C_{40}H_{72}O_{15}P_2$	$[M+2H+Na]^{3+}$	PGP(i-14:0/20:4(6Z,8E,10E,14Z)-2OH(5S,12R))	HMDB0275029
300.2169	300.2165	1.4971	$C_{33}H_{62}O_6$	$[M+2Na]^{2+}$	TG(10:0/10:0/10:0)	HMDB0000548
310.2007	310.2009	0.5009	$C_{35}H_{58}O_6$	$[M+2Na]^{2+}$	DG(12:0/20:4(6E,8Z,11Z,14Z)+=O(5)/0:0)	HMDB0294705
355.1726	355.1727	0.2607	$C_{16}H_{28}O_7$	$[M+Na]^+$	6Z-8-Hydroxygeraniol 8-O-glucoside	HMDB0035025
358.2587	358.2596	2.5191	$C_{41}H_{72}O_8$	$[M+H+Na]^{2+}$	DG(18:0/PGE2/0:0)	HMDB0295863

359.2619	359.2625	1.5800	$C_{38}H_{70}NO_8P$	$[M+H+NH_4]^{2+}$	PE(15:0/18:3(6Z,9Z,12Z))	HMDB0008896

Table S12 Differential metabolites annotated in serum samples.

Experimental *m/z*	Theoretical *m/z*	Delta (ppm)	Formula	Adduct	Compound Name	Identification Database
119.0188	119.0190	1.8826	$C_{16}H_{12}O_8$	$[M+2H+Na]^{3+}$	Patuletin	HMDB0030802
169.0491	169.0495	2.1692	$C_8H_8O_4$	$[M+H]^+$	3,4-Dihydroxybenzeneacetic acid	HMDB0001336
188.0703	188.0706	1.6098	$C_{11}H_9NO_2$	$[M+H]^+$	Indoleacrylic acid	HMDB0000734
200.2370	200.2367	1.5720	$C_{13}H_{28}O$	$[M+NH_4-H_2O]^+$	Tridecanol	HMDB0013316
202.0557	202.0565	4.1371	$C_5H_6N_4O_5$	$[M+NH_4-H_2O]^+$	2-Oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline	HMDB0059663
204.1590	204.1594	2.0507	$C_{10}H_{18}O_3$	$[M+NH_4]^+$	3-Oxodecanoic acid	HMDB0010724
205.0968	205.0972	2.0905	$C_{11}H_{12}N_2O_2$	$[M+H]^+$	L-Tryptophan	HMDB0000929
212.0194	212.0186	3.9605	$C_{17}H_{14}O_{10}$	$[M+2Na]^{2+}$	3,4,5-trihydroxy-6-({7-oxo-7H-furo[3,2-g]chromen-4-yl}oxy)oxane-2-carboxylic acid	HMDB0129424
216.9223	216.9235	5.3105	C_6H_5BrO	$[M+2Na-H]^+$	Bromobenzene-2,3-oxide	HMDB0060446
222.1825	222.1818	3.3346	$C_9H_{21}N_2O_3$	$[M+NH_4]^+$	3-Hydroxy-N6,N6,N6-trimethyl-L-lysine	HMDB0001422
223.1690	223.1693	1.1791	$C_{28}H_{44}O_4$	$[M+2H]^{2+}$	MG(0:0/24:6(6Z,9Z,12Z,15Z,18Z,21Z)/0:0)	HMDB0011560
226.9509	226.9509	0.1222	$C_6H_4Cl_2O_5$	$[M+H]^+$	2,5-Dichloro-4-oxohex-2-enedioate	HMDB0060363
227.0827	227.0827	0.1228	$C_{13}H_{14}N_2O_4$	$[M+H-2H_2O]^+$	L-cis-Cyclo(aspartylphenylalanyl)	HMDB0031360
228.1954	228.1954	0.2184	$C_{27}H_{54}O_2$	$[M+2Na]^{2+}$	Heptacosanoic acid	HMDB0002063
249.1820	249.1825	1.9179	$C_{14}H_{26}O_2$	$[M+Na]^+$	5-Tetradecenoic acid	HMDB0000499
263.0892	263.0890	0.6552	$C_{12}H_{16}O_5$	$[M+Na]^+$	3-Carboxy-4-methyl-5-propyl-2-furanpropionic acid	HMDB0061112
274.2738	274.2741	0.9690	$C_{16}H_{32}O_2$	$[M+NH_4]^+$	Palmitic acid	HMDB0000220
277.0378	277.0377	0.5118	$C_{10}H_{12}O_7S$	$[M+H]^+$	3-[4-methoxy-3-(sulfooxy)phenyl]propanoic acid	HMDB0131144
282.2788	282.2791	0.9242	$C_{18}H_{35}NO$	$[M+H]^+$	Oleamide	HMDB0002117
288.2167	288.2169	0.5696	$C_{15}H_{29}NO_4$	$[M+H]^+$	Octanoylcarnitine	HMDB0000791

291.1929	291.1930	0.3379	$C_{15}H_{26}O$	$[M+H+HCOONa]^+$	Farnesol	HMDB0004305
299.1097	299.1101	1.3015	$C_{12}H_{20}O_7$	$[M+Na]^+$	Triethyl citrate	HMDB0034263
313.1544	313.1547	0.8975	$C_{18}H_{20}N_2O_3$	$[M+H]^+$	Phenylalanylphenylalanine	HMDB0013302
316.2473	316.2482	2.7037	$C_{17}H_{33}NO_4$	$[M+H]^+$	Decanoylcarnitine	HMDB0000651
317.1786	317.1782	1.3222	$C_{47}H_{82}O_{15}P_2$	$[M+3H]^{3+}$	PGP(i-19:0/22:6(4Z,7Z,11E,13Z,15E,19Z)-2OH(10S,17))	HMDB0275595
381.1300	381.1309	2.2685	$C_{20}H_{22}O_6$	$[M+Na]^+$	Sanshodiol	HMDB0030573
383.1152	383.1160	1.9984	$C_6H_{12}O_6$	$[2M+Na]^+$	D-Glucose	HMDB0000122
398.2559	398.2561	0.4509	$C_{41}H_{77}N_2O_8P$	$[M+H+K]^{2+}$	SM(d16:1/20:3(8Z,11Z,14Z)-2OH(5,6))	HMDB0290273
542.3216	542.3217	0.1950	$C_{26}H_{50}NO_7P$	$[M+Na]^+$	LysoPC(18:2(9Z,12Z)/0:0)	HMDB0010386
566.3207	566.3217	1.6956	$C_{28}H_{50}NO_7P$	$[M+Na]^+$	LysoPC(20:4(5Z,8Z,11Z,14Z)/0:0)	HMDB0010395

Table S13 Statistical information of annotated metabolites in urine samples.

Compound Name	Adduct	VIP score	f value	p value	FDR	Fisher's LSD
L-Tryptophan	$[M+H]^+$	1.53	6.44	4.63E-05	6.31E-05	AS - SLE; AS - SS; HC - SLE; HC - SS; RA - SLE; RA - SS
Proline betaine	$[M+ACN+Na]^+$	2.73	28.65	1.47E-21	7.55E-21	HC - AS; AS - SLE; AS - SS; HC - RA; HC - SLE; HC - SS; RA - SLE; RA - SS
Phosphocreatine	$[M+NH_4-H_2O]^+$	4.09	27.83	5.58E-21	2.72E-20	HC - AS; AS - SLE; AS - SS; HC - RA; HC - SLE; HC - SS; RA - SLE
Methyl (2E,6Z)-dodecadienoate	$[M+H]^+$	2.80	29.14	6.72E-22	3.49E-21	HC - AS; RA - AS; HC - RA; HC - SLE; HC - SS; RA - SLE; RA - SS
5-Sulfosalicylic acid	$[M+NH_4-H_2O]^+$	2.30	32.54	2.98E-24	1.79E-23	HC - AS; HC - RA; HC - SLE; HC - SS; RA - SLE; RA - SS
5-Hydroxy-L-tryptophan	$[M+H]^+$	2.97	21.51	1.89E-16	6.58E-16	AS - SLE; AS - SS; HC - RA; HC - SLE; HC - SS; RA - SLE; RA - SS
Dihydrolipoamide	$[M+NH_4]^+$	3.61	13.42	2.01E-10	4.62E-10	AS - HC; AS - RA; AS - SLE; AS - SS; HC - RA; HC - SLE; HC - SS; RA - SLE
3,7-Dimethyl-3-octene-1,2,6,7-tetrol	$[M+Na]^+$	4.21	63.50	8.68E-44	1.84E-42	HC - AS; AS - SLE; AS - SS; HC - RA; HC - SLE; HC - SS; RA - SLE; RA - SS
Isoleucylproline	$[M+H]^+$	5.79	25.50	2.50E-19	1.06E-18	AS - RA; AS - SLE; AS - SS; HC - RA; HC - SLE; HC - SS; RA - SLE;

						RA - SS
Dodecanedioic acid	$[M+H]^+$	2.11	56.21	1.79E-39	3.00E-38	HC - AS; AS - SLE; AS - SS; HC - RA; HC - SLE; HC - SS; RA - SLE
Cerulenin	$[M+NH_4]^+$	1.25	11.64	4.58E-09	9.35E-09	AS - SLE; AS - SS; HC - RA; HC - SLE; HC - SS; RA - SLE; RA - SS
28-Methyl-27-nonacosenoic acid	$[M+2Na]^{2+}$	2.40	50.62	4.73E-36	6.28E-35	HC - AS; AS - SLE; AS - SS; HC - RA; HC - SLE; HC - SS; RA - SLE
Creatinine	$[2M+Na]^+$	1.25	37.69	9.79E-28	7.36E-27	HC - AS; RA - AS; HC - RA; HC - SLE; HC - SS; RA - SLE; RA - SS
TG(19:0/15:0/8:0)	$[M+2H+Na]^{3+}$	1.89	130.94	1.57E-77	8.14E-76	HC - AS; RA - AS; HC - RA; HC - SLE; HC - SS; RA - SLE; RA - SS
PG(16:1(9Z)/PGJ2)	$[M+3H]^{3+}$	2.84	10.79	2.06E-08	3.86E-08	HC - RA; HC - SLE; HC - SS
Myristoleic acid	$[M+2Na-H]^+$	2.24	8.07	2.57E-06	3.98E-06	AS - SLE; AS - SS; HC - RA; HC - SLE; HC - SS; RA - SLE; RA - SS
Butenylcarnitine	$[M+2Na-H]^+$	1.52	32.55	2.91E-24	1.76E-23	HC - AS; AS - SLE; AS - SS; HC - RA; HC - SLE; HC - SS; RA - SLE; RA - SS
TG(8:0/8:0/12:0)	$[M+2Na]^{2+}$	14.27	30.99	3.47E-23	1.93E-22	HC - AS; HC - RA; HC - SLE; HC - SS
DG(a-25:0/PGF2alpha/0:0)	$[M+3Na]^{3+}$	2.34	23.50	6.81E-18	2.65E-17	HC - AS; HC - RA; HC - SLE; HC - SS
PGP(i-14:0/20:4(6Z,8E,10E,14Z)-2OH(5S,12R))	$[M+2H+Na]^{3+}$	1.21	14.27	4.60E-11	1.10E-10	AS - RA; AS - SLE; AS - SS; HC - RA; HC - SLE; HC - SS; RA - SLE; RA - SS
TG(10:0/10:0/10:0)	$[M+2Na]^{2+}$	5.07	41.70	2.22E-30	1.95E-29	HC - AS; AS - SLE; HC - RA; HC - SLE; HC - SS; RA - SLE
DG(12:0/20:4(6E,8Z,11Z,14Z)+=O(5)/0:0)	$[M+2Na]^{2+}$	6.31	25.57	2.24E-19	9.60E-19	HC - AS; HC - RA; HC - SLE; HC - SS
6Z-8-Hydroxygeraniol 8-O-glucoside	$[M+Na]^+$	3.43	55.43	5.29E-39	8.39E-38	HC - AS; RA - AS; HC - RA; HC - SLE; HC - SS; RA - SLE; RA - SS
DG(18:0/PGE2/0:0)	$[M+H+Na]^{2+}$	5.68	42.24	9.94E-31	9.15E-30	HC - AS; AS - RA; AS - SLE; AS - SS; HC - RA; HC - SLE; HC - SS
PE(15:0/18:3(6Z,9Z,12Z))	$[M+H+NH_4]^{2+}$	1.22	42.49	6.80E-31	6.37E-30	HC - AS; AS - RA; AS - SLE; AS - SS; HC - RA; HC - SLE; HC - SS

Table S14 Statistical information of annotated metabolites in serum samples.

Compound Name	Adduct	VIP score	f value	p value	FDR	Fisher's LSD
Patuletin	$[M+2H+Na]^{3+}$	1.39	89.90	3.36E-58	1.72E-57	HC - AS; RA - AS; HC - RA; HC - SLE; HC - SS; RA - SLE; RA - SS
3,4-Dihydroxybenzeneacetic acid	$[M+H]^+$	1.31	10.30	4.91E-08	6.28E-08	AS - RA; HC - RA; HC - SLE; HC - SS

Indoleacrylic acid	$[M+H]^+$	13.29	84.82	1.46E-55	6.72E-55	HC - AS; HC - RA; HC - SLE; HC - SS; RA - SLE
Tridecanol	$[M+NH_4-H_2O]^+$	1.88	178.75	1.38E-96	3.11E-95	HC - AS; HC - RA; HC - SLE; HC - SS
2-Oxo-4-hydroxy-4-carboxy-5-ureidoimidazoline	$[M+NH_4-H_2O]^+$	2.57	32.99	1.45E-24	2.55E-24	HC - AS; HC - RA; HC - SLE; HC - SS
3-Oxodecanoic acid	$[M+NH_4]^+$	1.92	81.94	4.89E-54	2.02E-53	HC - AS; RA - AS; HC - RA; HC - SLE; HC - SS; RA - SLE; RA - SS
L-Tryptophan	$[M+H]^+$	5.91	72.85	4.20E-49	1.41E-48	HC - AS; HC - RA; HC - SLE; HC - SS
3,4,5-trihydroxy-6-({7-oxo-7H-furo[3,2-g]chromen-4-yl}oxy)oxane-2-carboxylic acid	$[M+2Na]^{2+}$	1.55	3.78	4.87E-03	5.37E-03	HC - AS; HC - RA; HC - SLE; HC - SS
Bromobenzene-2,3-oxide	$[M+2Na-H]^+$	2.16	125.19	5.20E-75	4.54E-74	HC - AS; HC - RA; HC - SLE; HC - SS; RA - SLE
3-Hydroxy-N6,N6,N6-trimethyl-L-lysine	$[M+NH_4]^+$	1.57	162.01	2.87E-90	4.44E-89	HC - AS; HC - RA; HC - SLE; HC - SS
MG(0:0/24:6(6Z,9Z,12Z,15Z,18Z,21Z)/0:0)	$[M+2H]^{2+}$	1.85	82.26	3.30E-54	1.39E-53	HC - AS; HC - RA; HC - SLE; HC - SS
2,5-Dichloro-4-oxohex-2-enedioate	$[M+H]^+$	7.96	126.24	1.78E-75	1.67E-74	HC - AS; RA - AS; HC - RA; HC - SLE; HC - SS; RA - SLE; RA - SS
L-cis-Cyclo(aspartylphenylalanyl)	$[M+H-2H_2O]^+$	2.35	85.89	4.00E-56	1.90E-55	HC - AS; HC - RA; HC - SLE; HC - SS; RA - SLE
Heptacosanoic acid	$[M+2Na]^{2+}$	1.34	135.43	1.84E-79	1.99E-78	HC - AS; RA - AS; HC - RA; HC - SLE; HC - SS; RA - SLE
5-Tetradecenoic acid	$[M+Na]^+$	2.44	42.30	9.07E-31	1.83E-30	AS - HC; RA - AS; SLE - AS; RA - HC; SLE - HC; SS - HC; RA - SLE; RA - SS
3-Carboxy-4-methyl-5-propyl-2-furanpropionic acid	$[M+Na]^+$	6.48	66.52	1.57E-45	4.66E-45	HC - AS; HC - RA; HC - SLE; HC - SS; RA - SLE
Palmitic acid	$[M+NH_4]^+$	1.98	30.01	1.66E-22	2.79E-22	AS - HC; AS - RA; RA - HC; SLE - HC; SS - HC; SLE - RA; SS - RA
3-[4-methoxy-3-(sulfooxy)phenyl]propanoic acid	$[M+H]^+$	1.40	165.27	1.59E-91	2.61E-90	HC - AS; HC - RA; HC - SLE; HC - SS
Oleamide	$[M+H]^+$	2.46	75.77	1.04E-50	3.72E-50	HC - AS; RA - AS; HC - RA; HC - SLE; HC - SS; RA - SS
Octanoylcarnitine	$[M+H]^+$	1.32	92.93	9.70E-60	5.26E-59	HC - AS; HC - RA; HC - SLE; HC - SS
Farnesol	$[M+H+HCOONa]^+$	3.75	90.08	2.75E-58	1.42E-57	HC - AS; HC - RA; HC - SLE; HC - SS; RA - SLE; RA - SS

Triethyl citrate	$[M+Na]^+$	3.40	326.03	7.27E-141	1.97E-138	HC - AS; HC - RA; HC - SLE; HC - SS; RA - SLE
Phenylalanylphenylalanine	$[M+H]^+$	4.38	44.85	2.05E-32	4.44E-32	HC - AS; HC - RA; HC - SLE; HC - SS; RA - SLE
Decanoylcarnitine	$[M+H]^+$	1.74	96.32	1.92E-61	1.13E-60	HC - AS; HC - RA; HC - SLE; HC - SS
PGP(i-19:0/22:6(4Z,7Z,11E,13Z,15E,19Z)-2OH(10S,17))	$[M+3H]^{3+}$	1.57	170.09	2.31E-93	4.32E-92	HC - AS; HC - RA; HC - SLE; HC - SS; RA - SLE
Sanshodiol	$[M+Na]^+$	1.54	81.47	8.67E-54	3.51E-53	HC - AS; HC - RA; HC - SLE; HC - SS
D-Glucose	$[2M+Na]^+$	1.62	95.05	8.26E-61	4.76E-60	HC - AS; HC - RA; HC - SLE; HC - SS
SM(d16:1/20:3(8Z,11Z,14Z)-2OH(5,6))	$[M+H+K]^{2+}$	1.34	219.17	1.54E-110	7.01E-109	HC - AS; HC - RA; HC - SLE; HC - SS; RA - SLE; RA - SS
LysoPC(18:2(9Z,12Z)/0:0)	$[M+Na]^+$	1.59	42.75	4.64E-31	9.46E-31	HC - AS; HC - RA; HC - SLE; HC - SS; RA - SLE
LysoPC(20:4(5Z,8Z,11Z,14Z)/0:0)	$[M+Na]^+$	1.75	73.50	1.85E-49	6.32E-49	HC - AS; RA - AS; HC - RA; HC - SLE; HC - SS; RA - SLE

Table S15 Metrics of classifiers for autoimmune diseases versus healthy controls (ADs vs HC) with urine metabolites panel (training cohort).

Model	AUC	Accuracy	F1	Precision	Recall
Neural Network	0.994	0.979	0.979	0.979	0.979
Random Forest	0.993	0.955	0.955	0.955	0.955
Support Vector Machine	0.989	0.965	0.965	0.966	0.965
Logistic Regression	0.953	0.888	0.889	0.890	0.888
k-Nearest Neighbor	0.953	0.872	0.868	0.873	0.872
AdaBoost	0.937	0.944	0.944	0.944	0.944
Naive Bayes	0.889	0.805	0.806	0.806	0.805

Table S16 Metrics of classifiers for autoimmune diseases versus healthy controls (ADs vs HC) with urine metabolites panel (testing cohort).

Model	AUC	Accuracy	F1	Precision	Recall
Naive Bayes	0.996	0.966	0.966	0.967	0.966
Neural Network	0.994	0.954	0.953	0.957	0.954
Support Vector Machine	0.981	0.954	0.953	0.957	0.954
Random Forest	0.978	0.937	0.936	0.937	0.937
k-Nearest Neighbor	0.960	0.902	0.897	0.915	0.902
AdaBoost	0.769	0.822	0.815	0.819	0.822
Logistic Regression	0.724	0.776	0.758	0.774	0.776

Table S17 Metrics of classifiers for autoimmune diseases versus healthy controls (ADs vs HC) with serum metabolites panel (training cohort).

Model	AUC	Accuracy	F1	Precision	Recall
Logistic Regression	1.000	0.997	0.997	0.997	0.997
Neural Network	1.000	0.995	0.995	0.995	0.995
Random Forest	1.000	0.995	0.995	0.995	0.995
Support Vector Machine	1.000	0.989	0.989	0.990	0.989
k-Nearest Neighbor	0.998	0.981	0.981	0.982	0.981
Naive Bayes	0.998	0.963	0.963	0.966	0.963
AdaBoost	0.982	0.984	0.984	0.984	0.984

Table S18 Metrics of classifiers for autoimmune diseases versus healthy controls (ADs vs HC) with serum metabolites panel (testing cohort).

Model	AUC	Accuracy	F1	Precision	Recall
Random Forest	1.000	0.966	0.966	0.969	0.966
Neural Network	1.000	0.937	0.938	0.947	0.937
Naive Bayes	1.000	0.776	0.782	0.867	0.776
k-Nearest Neighbor	0.971	0.920	0.921	0.929	0.920
Logistic Regression	0.953	0.897	0.898	0.906	0.897
Support Vector Machine	0.949	0.920	0.921	0.935	0.920
AdaBoost	0.939	0.943	0.943	0.943	0.943

Table S19 Metrics of classifiers for autoimmune diseases versus healthy controls (ADs vs HC) with fusion model of metabolites panel (training cohort).

Model	AUC	Accuracy	F1	Precision	Recall
Support Vector Machine	1.000	1.000	1.000	1.000	1.000
Random Forest	1.000	0.995	0.995	0.995	0.995
Neural Network	1.000	1.000	1.000	1.000	1.000
Logistic Regression	1.000	0.992	0.992	0.992	0.992
Naive Bayes	1.000	0.984	0.984	0.985	0.984
AdaBoost	0.982	0.984	0.984	0.984	0.984
k-Nearest Neighbor	0.972	0.893	0.890	0.896	0.893

Table S20 Metrics of classifiers for autoimmune diseases versus healthy controls (ADs vs HC) with fusion model of metabolites panel (testing cohort).

Model	AUC	Accuracy	F1	Precision	Recall
Support Vector Machine	1.000	0.989	0.989	0.989	0.989
Naive Bayes	1.000	0.891	0.894	0.918	0.891
Neural Network	1.000	0.971	0.972	0.974	0.971
Random Forest	0.993	0.954	0.954	0.954	0.954
k-Nearest Neighbor	0.985	0.937	0.936	0.937	0.937
AdaBoost	0.952	0.960	0.960	0.960	0.960
Logistic Regression	0.883	0.897	0.897	0.897	0.897

Table S21 Metrics of classifiers for the distinction of four autoimmune diseases (ADs) and healthy controls (HC) with urine metabolites panel.

Model	AUC	Accuracy	F1	Precision	Recall
Random Forest	0.998	0.961	0.961	0.961	0.961
Neural Network	0.993	0.911	0.910	0.918	0.911
Logistic Regression	0.992	0.939	0.939	0.939	0.939
Support Vector Machine	0.980	0.844	0.844	0.856	0.844
k-Nearest Neighbor	0.946	0.706	0.705	0.707	0.706
Naive Bayes	0.941	0.700	0.706	0.723	0.700
AdaBoost	0.931	0.889	0.889	0.895	0.889

AS vs SS vs SLE vs RA vs HC, AS: ankylosing spondylitis, SS: sicca syndrome, SLE: systemic lupus erythematosus, RA: rheumatoid arthritis.

Table S22 Metrics of classifiers for the distinction of four autoimmune diseases (ADs) and healthy controls (HC) with serum metabolites panel.

Model	AUC	Accuracy	F1	Precision	Recall
Random Forest	0.974	0.833	0.832	0.834	0.833
Neural Network	0.962	0.800	0.801	0.806	0.800
k-Nearest Neighbor	0.920	0.589	0.576	0.574	0.589
Support Vector Machine	0.908	0.594	0.604	0.638	0.594
Naive Bayes	0.860	0.561	0.533	0.575	0.561
AdaBoost	0.858	0.772	0.772	0.773	0.772
Logistic Regression	0.820	0.644	0.641	0.642	0.644

AS vs SS vs SLE vs RA vs HC, AS: ankylosing spondylitis, SS: sicca syndrome, SLE: systemic lupus erythematosus, RA: rheumatoid arthritis.

Table S23 Metrics of classifiers for the distinction of four autoimmune diseases (ADs) and healthy controls (HC) with fusion model of metabolites panel.

Model	AUC	Accuracy	F1	Precision	Recall
Neural Network	0.999	0.972	0.972	0.973	0.972
Logistic Regression	0.995	0.956	0.956	0.956	0.956
Random Forest	0.994	0.967	0.967	0.971	0.967
Support Vector Machine	0.992	0.883	0.884	0.897	0.883
Naive Bayes	0.972	0.817	0.814	0.821	0.817
AdaBoost	0.948	0.917	0.917	0.918	0.917
k-Nearest Neighbor	0.941	0.694	0.697	0.703	0.694

AS vs SS vs SLE vs RA vs HC, AS: ankylosing spondylitis, SS: sicca syndrome, SLE: systemic lupus erythematosus, RA: rheumatoid arthritis.

1. G. Della Sala, A. Mangoni, V. Costantino and R. Teta, *Frontiers in Chemistry*, 2020, **8**.