基于非靶向代谢组学与可解释机器学习的白芍与赤芍饮片差异代谢物研究

陈 琪[1,2], 范国旗[1,2], 郭伊瑞[1], 左莉华[2,3], 李寒冰[4], 邓丹萍[5], 朱承慧[5], 孙 志[2,3*], 刘改枝[1*]

(1. 河南中医药大学药学院, 河南 郑州 450046; 2. 郑州大学第一附属医院药学部, 河南 郑州 450052; 3. 河南省精准医学临床质谱工程研究中心, 河南 郑州 450052; 4. 河南中医药大学医学院, 河南 郑州 450046; 5. 郴州市中医医院, 湖南 郴州 423000)

摘要: 本研究采用超高效液相色谱−四极杆−静电场轨道阱高分辨质谱 (UHPLC-Q-Orbitrap-HRMS) 对市售白芍与赤芍饮片进行非靶向代谢组学分析, 通过多元统计分析筛选差异代谢物, 结合可解释性机器学习方法, 构建了随机森林 (RF)、自适应增强 (AdaBoost) 和极端梯度提升 (XGBoost) 模型, 利用沙普利加性解释 (Shapley additive explanations, SHAP) 方法解析特征代谢物贡献度。共鉴定出44种代谢物, 包括黄酮类9种、酚酸类8种等代谢物, 其中12种代谢物在两类饮片间显著差异; 在测试集中, RF模型表现最佳; SHAP分析结果显示, 葡萄糖酸、槲皮素、山柰酚、*L*-天冬氨酸、二氢槲皮素、氧化芍药苷在模型中贡献度最高, *L*-天冬氨酸在白芍中丰度相对较高, 其余在赤芍中的丰度相对较高。该研究所建立的代谢组学结合机器学习的策略, 不仅为探讨白芍与赤芍功效差异提供了潜在化学线索, 同时也为其他易混淆中药的基原鉴别、产地溯源及特征代谢物挖掘提供了新的思路。

关键词: 白芍; 赤芍; 非靶向代谢组学; 差异代谢物; 机器学习; SHAP分析

中图分类号: R917 **文献标识码**: A **文章编号**: 0513-4870(2026)03-0871-10

Integrative analysis of differential metabolites between Paeoniae Radix Alba and Paeoniae Radix Rubra with non-targeted metabolomics and interpretable machine learning

CHEN Qi[1,2], FAN Guo-qi[1,2], GUO Yi-rui[1], ZUO Li-hua[2,3], LI Han-bing[4], DENG Dan-ping[5], ZHU Cheng-hui[5], SUN Zhi[2,3*], LIU Gai-zhi[1*]

(*1. College of Pharmacy, Henan University of Traditional Chinese Medicine, Zhengzhou 450046, China; 2. Department of Pharmacy, the First Affiliated Hospital of Zhengzhou University, Zhengzhou 450052, China; 3. Henan Engineering Research Center of Clinical Mass Spectrometry for Precision Medicine, Zhengzhou 450052, China; 4. School of Medicine, Henan University of Traditional Chinese Medicine, Zhengzhou 450046, China; 5. Chenzhou Municipal Hospital of Traditional Chinese Medicine, Chenzhou 423000, China*)

Abstract: Ultra-high performance liquid chromatography coupled with quadrupole-Orbitrap high-resolution mass spectrometry (UHPLC-Q-Orbitrap-HRMS) was applied for the non-targeted metabolomic characterization of Paeoniae Radix Alba and Paeoniae Radix Rubra decoction pieces. Differential metabolites were identified through

收稿日期: 2025-09-22; 修回日期: 2025-12-09.

基金项目: 国家自然科学基金面上项目 (82374018); 河南省“3个100”临床医学科学家培养专项 (HNCMS202432, HNCMS202411); 河南省中医药科学研究专项 (2023ZY2197); 河南省高等学校重点科研项目基础研究计划 (24A360025).

*通讯作者 E-mail: liugaizhi@126.com; sunzhi2013@163.com

DOI: 10.16438/j.0513-4870.2025-1131

multivariate statistical analysis. Interpretable machine learning methods, including random forest (RF), adaptive boosting (AdaBoost), and extreme gradient boosting (XGBoost), were further employed to build classification models, and the Shapley additive explanations (SHAP) algorithm was used to evaluate the contributions of characteristic metabolites. A total of 44 metabolites were identified, including 9 flavonoids, 8 phenolic acids, and other types of metabolites, among which 12 showed significant differences between the two kinds of decoction pieces. In the test set, the RF model exhibited superior performance. SHAP analysis indicated that gluconic acid, quercetin, kaempferol, *L*-aspartic acid, taxifolin, and oxypaeoniflorin significantly contributed to model discrimination, with *L*-aspartic acid showing elevated levels in Paeoniae Radix Alba, while the other metabolites were relatively higher in Paeoniae Radix Rubra. The metabonomics-machine learning strategy established in this study provides potential chemical evidence for elucidating the efficacy differences between Paeoniae Radix Alba and Paeoniae Radix Rubra and offers new insights into the identification, origin traceability, and characteristic marker discovery of other easily confused traditional Chinese medicines.

Key words: Paeoniae Radix Alba; Paeoniae Radix Rubra; non-targeted metabolomics; differential metabolite; machine learning; SHAP analysis

白芍 (Paeoniae Radix Alba) 与赤芍 (Paeoniae Radix Rubra) 是我国常用中药材, 均来源于毛茛科植物, 其中白芍来源于芍药 (*Paeonia lactiflora* Pall.) 的干燥根, 赤芍则为芍药或川赤芍 (*Paeonia veitchii* Lynch) 的干燥根[1]。然而, 两者在临床应用中表现出显著差异[2]。白芍以“养血调经、敛阴止汗”为主要功效, 常用于血虚、月经失调等证候, 并为当归芍药散、八珍益母丸等经典方剂的重要药材[3-5]。赤芍则以“清热凉血、散瘀止痛”为主, 常见于血热瘀阻相关疾病的治疗, 是血必净注射液等活血清热制剂的关键组成药物[6,7]。现有研究表明, 两者均含芍药苷、没食子酰芍药苷等共有成分, 但在化学成分的丰度分布与类别组成方面存在差异[8,9]; 其中, 白芍以单萜苷类化合物 (如芍药内酯苷、苯甲酰芍药苷) 相对富集, 而赤芍则以芍药苷及多酚类物质的积累更为显著[10-12]。现有研究虽已利用核磁共振技术开展代谢组学分析, 但液质联用技术的应用仍局限于氨基酸等少数成分的定量检测, 尚未实现对白芍与赤芍整体次生代谢谱的全面解析, 制约了对其物质基础差异的深入认识[13-15]。为此, 亟需引入高覆盖、非靶向的代谢组学策略, 结合先进的数据挖掘方法, 实现对二者整体代谢特征的系统性解析。

非靶向代谢组学技术能够在不依赖先验假设的条件下全面解析复杂样品的代谢组成, 为探索中药饮片的整体化学特征提供了有效工具[16,17]。但由于代谢组学数据维度高且噪声复杂, 传统多元统计方法在解释性和判别效能方面存在一定不足[18-20]。近年来, 机器学习方法因其在高维数据处理和非线性建模中的优势而逐渐应用于代谢组学研究[21]。具有可解释性的沙普利加性解释方法 (Shapley additive explanations, SHAP), 能够量化单一变量对模型预测结果的贡献, 不仅有助于识别可能具有判别意义的关键代谢物, 也为探讨不同药材在功效特征上的差异提供了新的视角[22]。

鉴于白芍与赤芍在临床应用与市场流通中主要以饮片形态存在, 且饮片加工较为规范、来源广泛, 能够在一定程度上反映实际药用特征, 因此本研究选择市售饮片作为研究对象。在此基础上, 结合超高效液相色谱−四极杆−静电场轨道阱高分辨质谱 (UHPLC-Q-Orbitrap-HRMS) 非靶向代谢组学与机器学习方法, 对两者的化学成分谱进行系统解析, 以揭示白芍与赤芍在代谢物组成上的整体差异。进一步利用SHAP算法筛选出具有判别价值的代谢物, 为阐释二者功效差异提供潜在物质基础线索; 同时也为中药质量控制与现代化研究提供新的思路。

材料与方法

仪器与试剂 Ultimate 3000超高效液相色谱仪、Q-Exactive型三重四极杆质谱仪及VORTEX-GENIE涡旋混匀仪、甲醇 (美国Thermo Fisher Scientific公司); 电子分析天平 (瑞士Mettler Toledo上海有限公司); HPLC级甲酸、酮洛芬标准品 (美国Sigma-Aldrich公司); 2-氯-*L*-苯丙氨酸对照品 (北京J&K Scientific公司)。

样品及来源 实验所用饮片样品均购自郑州市各药店, 共计白芍饮片 (编号BS1～BS15) 和赤芍饮片 (编号CS1～CS15) 各15批。经河南中医药大学李寒冰教授鉴定为毛茛科芍药的干燥根, 鉴定标准为中国药典 (2020年版), 样品来源及编号信息见表1。

试品溶液制备 取各批样品20 g粉碎过3号筛后, 精密称取1.0 g粉末置于50 mL离心管中, 加入25 mL含有内标物的甲醇溶液 (酮洛芬浓度为500 μg·mL^{-1},

Table 1　Sample information of Paeoniae Radix Alba (BS) and Paeoniae Radix Rubra (CS) decoction pieces

No.	Batch No.	Place of origin	No.	Batch No.	Place of origin
BS1	240302	Anhui	CS1	25021311	Inner Mongolia
BS2	A240916	Anhui	CS2	2502021042	Inner Mongolia
BS3	GX240401	Anhui	CS3	2310098	Inner Mongolia
BS4	241101	Anhui	CS4	241001	Inner Mongolia
BS5	GX241103	Anhui	CS5	B24062701-1	Inner Mongolia
BS6	241201	Anhui	CS6	250301	Inner Mongolia
BS7	A250202	Anhui	CS7	240403851	Inner Mongolia
BS8	D2411043	Anhui	CS8	250100921	Inner Mongolia
BS9	GX240201	Anhui	CS9	241001-1	Heilongjiang
BS10	240409	Anhui	CS10	250101-1	Heilongjiang
BS11	2241201	Anhui	CS11	240901-1	Heilongjiang
BS12	241003	Anhui	CS12	240501-1	Heilongjiang
BS13	241004	Anhui	CS13	241011-1	Heilongjiang
BS14	2230302	Anhui	CS14	241202-1	Heilongjiang
BS15	240904	Anhui	CS15	240401-1	Heilongjiang

2-氯-*L*-苯丙氨酸浓度为50 μg·mL^{-1})。室温下静置浸泡4 h后, 采用功率250 W、频率40 kHz的超声波清洗器提取20 min。提取后补加甲醇溶液至原体积, 混匀离心, 上清液经0.22 μm微孔滤膜过滤, 获得供试品溶液。为评价分析系统的稳定性与重复性, 将各供试品溶液分别取200 μL等体积混合, 制备质控样品 (quality control, QC), 用于监测仪器运行状态及分析过程中数据质量, 保障实验结果的稳定性与可重复性。

色谱条件　色谱分离采用ACQUITY UPLC HSS T3 C_{18}色谱柱 (100 mm × 2.1 mm, 1.8 μm, 美国Waters公司); 流动相为含0.1%甲酸的水 (A相) 与乙腈 (B相); 梯度洗脱程序如下: 0～1.0 min, 10% B; 1.0～5.0 min, 10%～98% B; 5.0～6.5 min, 98% B; 6.5～6.6 min, 98%～10% B; 6.6～8.0 min, 10% B; 流速0.3 mL·min^{-1}; 柱温为40 ℃; 进样体积为2 μL。

质谱条件　在正负离子模式下进行全扫描 (*m*/*z* 80～1 200), 喷雾电压分别为3.5 kV (正) 与2.8 kV (负), 毛细管温度设为350 ℃; 鞘气流速为40 arb (正) 和38 arb (负), 辅助气流速均为10 arb, 梯度碰撞能量依次设定为20、40、60 eV。

数据分析与处理　原始质谱数据由Thermo Xcalibur 3.0软件采集, 并通过Compound Discoverer 3.3平台完成峰提取、校准、对齐及代谢物识别, 结合mzCloud、mzVault和ChemSpider数据库进行峰注释, 保留峰面积相对标准偏差 (RSD) 小于30%的化合物[23,24]。采用Simca 14.1软件进行主成分分析 (principal component analysis, PCA) 和正交偏最小二乘判别分析 (orthogonal partial least squares-discriminant analysis, OPLS-DA)。差异代谢物的筛选标准设定为同时满足以下条件: 变量重要性投影值 (variable importance in projection, VIP) > 1.0, *P* 值 < 0.05, 且差异倍数 (fold change, FC) > 2.0 或 < 0.5[25], 相关可视化分析通过微生信完成。

机器学习建模　基于所筛选的差异代谢物分别构建随机森林 (random forest, RF)、自适应增强 (adaptive boosting, AdaBoost) 及极端梯度提升 (extreme gradient boosting, XGBoost) 3种机器学习模型。将数据集按7∶3随机划分为训练集和测试集, 训练集采用5折交叉验证进行模型稳定性评估[26,27]。在测试集中, 利用受试者工作特征曲线 (receiver operating characteristic, ROC) 和混淆矩阵对最佳模型的泛化能力进行评估。最后, 通过Python中SHAP方法, 对最佳模型进行特征贡献度解析和可视化分析。

结果与讨论

1　白芍和赤芍代谢物组成分析

QC样本总离子流图 (TIC) 分析显示, 正负离子模式下各色谱峰保留时间基本一致且信号响应高度重合, 表明检测方法具有良好的重现性和仪器稳定性 (图1), 为后续代谢物分析提供了技术保障[28]。在白芍与赤芍中共鉴定出44种代谢物。其中黄酮类9种、酚酸类8种、氨基酸及其衍生物8种、萜类5种、挥发油类5种、有机酸及其衍生物3种、鞣质类2种、糖类2种及其他类2种, 详细成分信息见表2。综上所述, 黄酮类化合物在两类样品中分布较为广泛, 是植物中重要的代谢物之一。

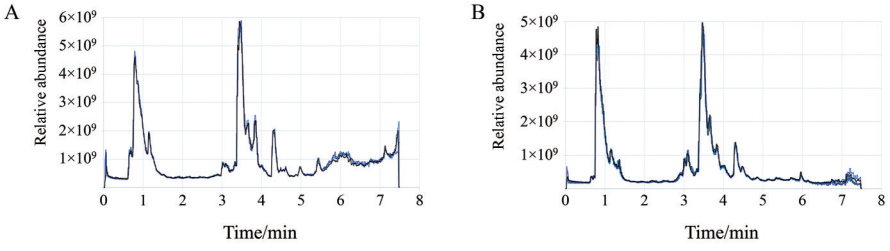


Figure 1　Total ion chromatograms of quality control samples of Paeoniae Radix Alba and Paeoniae Radix Rubra. A: Positive ionization mode; B: Negative ionization mode

Table 2 Identification of chemical constituents between Paeoniae Radix Alba and Paeoniae Radix Rubra

No.	Compound	t_R/min	Formula	Ion mode	*m/z*	Error/ppm	Fragment ion (*m/z*)
1	Taxifolin	3.81	$C_{15}H_{12}O_7$	$[M-H]^-$	304.058 4	0.37	285.040 4, 275.056 0, 241.050 6
2	Sucrose	1.22	$C_{12}H_{22}O_{11}$	$[M-H]^-$	342.116 4	0.46	179.055 4, 119.033 8, 71.012 5
3	Salicylic acid	4.26	$C_7H_6O_3$	$[M-H]^-$	138.030 6	−7.97	136.015 2, 93.033 3, 65.038 6
4	Quercetin	4.26	$C_{15}H_{10}O_7$	$[M-H]^-$	302.042 9	0.65	273.040 7, 257.045 5, 229.050 5
5	Pyrogallol	1.38	$C_6H_6O_3$	$[M-H]^-$	126.030 6	−9.10	81.033 5, 69.033 2
6	Protocatechuic acid	2.22	$C_7H_6O_4$	$[M-H]^-$	154.025 6	−6.34	110.031 8, 109.028 3, 108.987 4
7	Procyanidin B2	2.54	$C_{30}H_{26}O_{12}$	$[M+H]^+$	578.141 8	−1.12	299.055 1, 287.055 1, 271.060 0
8	Procyanidin B1	2.98	$C_{30}H_{26}O_{12}$	$[M-H]^-$	578.142 7	0.53	289.071 7, 245.081 7, 161.023 5
9	Palmitic acid	6.00	$C_{16}H_{32}O_2$	$[M-H]^-$	256.240 0	−1.08	211.206 1, 173.047 5, 119.945 9
10	Paeoniflorin	3.49	$C_{23}H_{28}O_{11}$	$[M-H]^-$	480.163 1	−0.14	357.119 2, 327.108 0, 195.065 4
11	Oxypaeoniflorin	3.03	$C_{23}H_{28}O_{12}$	$[M-H]^-$	496.158 2	0.30	345.119 6, 281.066 7, 177.054 8
12	Oleanolic acid	7.23	$C_{30}H_{48}O_3$	$[M-H]^-$	456.360 7	0.72	437.342 0, 409.347 0, 139.112 3
13	Methyl gallate	3.22	$C_8H_8O_5$	$[M-H]^-$	184.036 3	−4.61	168.005 6, 139.002 5, 111.007 6
14	Maltopentaose	0.79	$C_{30}H_{52}O_{26}$	$[M-H]^-$	828.275 5	0.98	707.225 5, 545.172 4, 425.130 1
15	Kaempferol	4.53	$C_{15}H_{10}O_6$	$[M-H]^-$	286.047 9	0.54	257.046 4, 213.055 2, 151.002 2
16	Gallic acid	2.09	$C_7H_6O_5$	$[M-H]^-$	170.020 7	−5.00	125.023 3, 124.015 7, 97.028 2
17	Ethyl gallate	3.72	$C_9H_{10}O_5$	$[M-H]^-$	198.052 2	−3.24	169.013 3, 140.010 0, 125.023 1
18	Ellagic acid	3.28	$C_{14}H_6O_8$	$[M+H]^+$	302.006 0	−0.85	285.002 6, 257.008 1, 219.016 8
19	Citric acid	1.18	$C_6H_8O_7$	$[M-H]^-$	192.026 2	−4.05	173.008 4, 154.997 5, 130.997 3
20	Catechin	2.34	$C_{15}H_{14}O_6$	$[M+H]^+$	290.079 0	−0.14	207.065 1, 179.070 6, 147.044 0
21	Benzoylpaeoniflorin	4.32	$C_{30}H_{32}O_{12}$	$[M+HCOO]^-$	584.189 8	0.68	177.054 8, 165.054 8, 135.044 0
22	Benzoic acid	4.65	$C_7H_6O_2$	$[M-H]^-$	122.035 6	−9.51	108.020 4, 93.033 2
23	Albiflorin	3.46	$C_{23}H_{28}O_{11}$	$[M+H]^+$	480.162 4	−1.66	179.070 1, 151.075 2, 109.028 5
24	Acetophenone	3.45	C_8H_8O	$[M+H]^+$	120.057 5	0.17	105.045 0, 93.070 3, 91.054 7
25	*γ*-Aminobutyric acid	0.78	$C_4H_9NO_2$	$[M+H]^+$	103.063 6	2.95	87.044 5, 86.060 5, 69.034 1, 60.081 5, 58.065 8
26	Valine	6.62	$C_5H_{11}NO_2$	$[M+H]^+$	117.079 2	1.52	73.029 0, 72.081 4, 59.073 7
27	Propyl gallate	3.17	$C_{10}H_{12}O_5$	$[M-H]^-$	212.067 9	−2.65	125.023 2, 123.080 4
28	Naringenin	4.50	$C_{15}H_{12}O_5$	$[M-H]^-$	272.068 5	0.08	253.050 6, 227.071 5, 177.018 5, 169.013 1, 151.002 6, 119.049 0
29	Threonine	0.81	$C_4H_9NO_3$	$[M+H]^+$	119.058 3	0.44	102.055 3, 74.060 6, 56.050 2
30	*L*-Phenylalanine	1.54	$C_9H_{11}NO_2$	$[M+H]^+$	165.078 9	−0.49	149.059 5, 131.049 1, 121.065 2, 120.080 8
31	Linoelaidic acid	7.15	$C_{18}H_{32}O_2$	$[M-H]^-$	280.240 3	0.06	261.222 3, 59.012 5
32	Glutamic acid	0.81	$C_5H_9NO_4$	$[M+H]^+$	147.053 0	−1.16	130.049 9, 102.055 3, 84.044 9
33	Leucine	0.90	$C_6H_{13}NO_2$	$[M+H]^+$	131.094 6	−0.57	86.096 9, 69.070 6
34	*L*-Aspartic acid	0.81	$C_4H_7NO_4$	$[M+H]^+$	133.037 4	−0.87	116.034 4, 88.039 8, 74.024 2
35	*L*-(-)-Carvone	3.21	$C_{10}H_{14}O$	$[M+H]^+$	150.104 4	−0.47	123.044 4, 95.086 0, 81.070 4
36	Gluconic acid	0.83	$C_6H_{12}O_7$	$[M-H]^-$	196.057 5	−4.09	177.039 7, 159.029 0, 141.017 9
37	Eugenol	3.76	$C_{10}H_{12}O_2$	$[M+H]^+$	164.083 6	−1.11	137.059 7, 133.064 8, 105.070 2
38	Eriodictyol	4.51	$C_{15}H_{12}O_6$	$[M-H]^-$	288.063 6	0.65	151.002 5, 135.043 9, 125.023 3
39	Epicatechin	3.13	$C_{15}H_{14}O_6$	$[M+H]^+$	290.078 5	−1.81	207.065 2, 179.070 5, 165.054 9, 147.044 0
40	Arginine	0.73	$C_6H_{14}N_4O_2$	$[M+H]^+$	174.111 5	−1.24	158.092 4, 130.097 6, 116.070 8, 112.087 1
41	*D*-(-)-Glutamine	0.78	$C_5H_{10}N_2O_3$	$[M-H]^-$	146.067 9	−8.31	127.050 1, 109.039 4
42	Camphor	0.31	$C_{10}H_{16}O$	$[M+H]^+$	152.120 1	−0.40	135.117 0, 109.101 4, 107.001 0, 97.065 2
43	1,2,3,6-Tetra-*O*-galloyl-*β*-*D*-glucose	3.48	$C_{34}H_{28}O_{22}$	$[M-2H]^{2-}$	393.045 8	0.82	635.089 2, 617.078 8, 579.089 8, 483.076 9, 465.067 7, 447.056 0
44	1,2,3,4,6-Pentagalloylglucose	3.58	$C_{41}H_{32}O_{26}$	$[M-2H]^{2-}$	469.051 3	0.07	295.045 6, 277.035 3, 211.024 3, 125.023 2

2 白芍和赤芍的多元统计分析

为分析白芍与赤芍化学轮廓的整体差异, 采用多元统计分析方法对正负离子模式合并数据进行PCA初步分析, 结果显示 (图2A), 白芍与赤芍样本呈现明显的分离趋势, 提示两组间可能存在化学成分差异, 而QC样本紧密聚集, 表明分析方法具有良好的技术重复性与数据稳定性[29]。后续将通过监督分析进一步筛选差异标志物。

采用OPLS-DA方法对白芍与赤芍样本进行有监督判别分析。结果显示 (图2B), 两组样本在得分图上

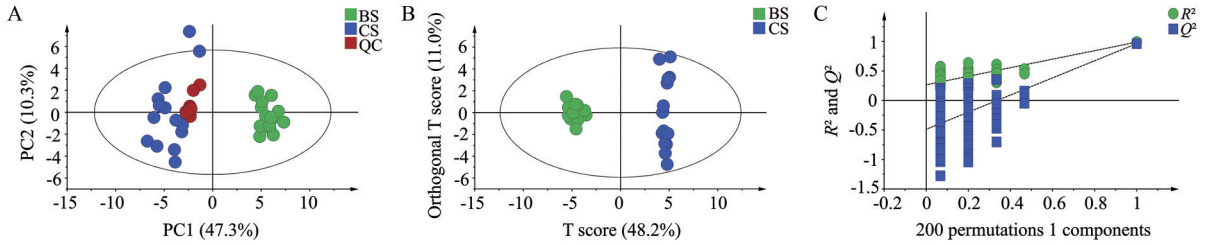


Figure 2　Multivariate statistical analysis of metabolomics of Paeoniae Radix Alba and Paeoniae Radix Rubra. A: Principal component analysis (PCA) score plot; B: Orthogonal partial least squares-discriminant analysis (OPLS-DA) score plot; C: Permutation test plot

显著分离, 且均分布于95%置信椭圆内, 表明模型参数具有较好的判别效果, 其中R^2Y为0.989, Q^2为0.963, 反映出模型具有优异的解释能力和预测能力[30]。白芍与赤芍样本分别聚集在置信区间的两侧, 反映出两组样本在化学成分上存在显著差异。为验证模型的可靠性, 进行200次置换检验, 结果显示 (图2C), 原始模型的R^2Y和Q^2值显著高于置换模型, 且Q^2回归线的截距为负值, 提示模型预测稳定且未过拟合, 为差异代谢物筛选提供了可靠依据[31]。

3　白芍与赤芍差异代谢物的筛选

基于OPLS-DA模型, 共筛选出白芍与赤芍之间的12种显著差异代谢物 (图3A), 其中, 3种代谢物在白芍组中相对含量显著升高, 包括*L*-天冬氨酸、谷氨酸及麦芽五糖; 其余9种代谢物则显著降低, 主要涉及槲皮素、没食子酸丙酯、氧化芍药苷等成分。对筛选的12种差异代谢物, 采用微生信进行聚类热图分析。热图中通过不同颜色表示代谢物在不同样本中的相对丰度, 橙色表示上调, 绿色表示下调。结果显示 (图3B), 白芍组样品中多数代谢物较赤芍组呈下调趋势, 提示两组在代谢特征上存在显著差异, 详细信息见表3。

4　基于机器学习模型构建与性能评估

基于所筛选的12种差异代谢物, 分别构建了RF、

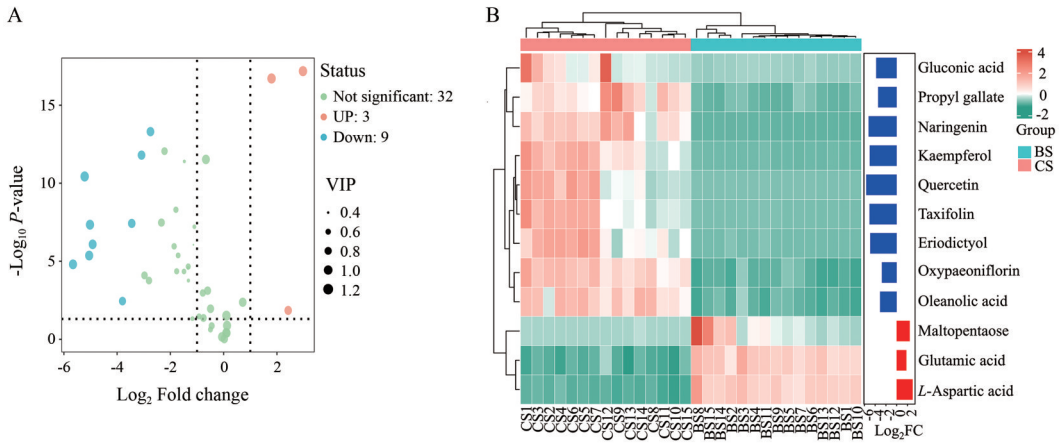


Figure 3　Analysis of differential metabolites between Paeoniae Radix Alba and Paeoniae Radix Rubra. A: Volcano plot of differential metabolites; B: Hierarchical clustering heatmap of differential metabolites. VIP: Variable importance in projection; FC: Fold change

Table 3　Characteristics of differential metabolites between Paeoniae Radix Alba and Paeoniae Radix Rubra

No.	Metabolite	Formula	Classify	BS *vs* CS			Type
				VIP	FC	*P* value	
1	Taxifolin	$C_{15}H_{12}O_7$	Flavonoids	1.27	0.03	4.30E-06	Down
2	Quercetin	$C_{15}H_{10}O_7$	Flavonoids	1.27	0.02	1.57E-05	Down
3	Oxypaeoniflorin	$C_{23}H_{28}O_{12}$	Terpenes	1.17	0.15	4.93E-14	Down
4	Oleanolic acid	$C_{30}H_{48}O_3$	Terpenes	1.16	0.12	1.58E-12	Down
5	Maltopentaose	$C_{30}H_{52}O_{26}$	Saccharides	1.13	5.36	1.41E-02	Up
6	Kaempferol	$C_{15}H_{10}O_6$	Flavonoids	1.29	0.03	4.54E-08	Down
7	Propyl gallate	$C_{10}H_{12}O_5$	Phenolic acids	1.15	0.09	3.75E-08	Down
8	Naringenin	$C_{15}H_{12}O_5$	Flavonoids	1.33	0.03	3.68E-11	Down
9	Glutamic acid	$C_5H_9NO_4$	Amino acids and their derivatives	1.34	3.48	1.94E-17	Up
10	*L*-Aspartic acid	$C_4H_7NO_4$	Amino acids and their derivatives	1.33	7.87	6.55E-18	Up
11	Gluconic acid	$C_6H_{12}O_7$	Organic acids and their derivatives	1.05	0.07	3.57E-03	Down
12	Eriodictyol	$C_{15}H_{12}O_6$	Flavonoids	1.24	0.03	8.23E-07	Down

AdaBoost、XGBoost 分类模型。将数据集按 7∶3 划分为训练集与测试集, 采用 5 折交叉验证评估模型性能。ROC 曲线分析显示, RF 模型的曲线下面积 (AUC) 高于 AdaBoost 和 XGBoost 模型 (图 4A), 表明其分类稳定性较好。学习曲线分析表明, RF 模型具有较好的拟合和泛化能力, 其训练得分及交叉验证得分均接近 1.0, 并随样本量增加迅速收敛且趋于一致 (图 4B～D)。综合比较各模型的性能指标 (表 4), RF 模型在预测准确性和稳定性方面均表现优异, 证明该模型在训练过程中具有较好的稳定性和泛化能力。

为进一步验证模型的泛化能力, 在完整训练集上重新训练该模型, 并在独立测试集上进行了验证。结果显示 (图 4E), RF 模型 ROC 曲线接近理论值。混淆矩阵分析进一步证实 (图 4F～H), 模型预测结果与实际分类之间高度一致, 而 AdaBoost 和 XGBoost 模型分别有一个白芍和赤芍样本的误判, 提示 RF 模型相比于其他模型, 具有良好的泛化能力, 可作为最终分类模型。

5 模型预测结果的解释性分析

基于博弈论中的 Shapley 值原理, SHAP 方法通过分配每个特征对模型输出的贡献, 能够精确评估各特征的边际贡献, 兼顾全局和局部解释的能力。这种方法在提升模型可解释性的同时, 确保了模型预测结果的透明性与可信度[32,33]。本研究采用 SHAP 算法对 RF 模型进行可解释性分析, 以评估各代谢物对分类模型的贡献。通过计算各特征的平均绝对 SHAP 值并进行贡献度排名 (图 5A), 发现葡萄糖酸 (gluconic acid)、槲皮素 (quercetin)、山柰酚 (kaempferol)、*L*-天冬氨酸 (*L*-aspartic acid)、二氢槲皮素 (taxifolin) 及氧化芍药苷 (oxypaeoniflorin) 6 种代谢物在模型中具有较高的平均贡献度, 提示这些代谢物在区分白芍与赤芍的预测模型中发挥关键作用。

为揭示白芍与赤芍的代谢差异, 本研究采用 SHAP 方法评估模型特征重要性。如图 5B 所示, SHAP 蜂群图中每个点代表样本对应特征的 SHAP 值, 点的颜色深浅反映了特征值的大小, 特征按其重要性降序排列。分析结果表明, 多数关键代谢物在赤芍组中呈现正向贡献, 即其含量增加倾向于判定为赤芍; 而 *L*-天冬氨酸与谷氨酸则表现出显著的负向贡献, 其高丰度特征与白芍的判别相关。

对筛选出的前 6 种代谢物峰面积进行组间比较

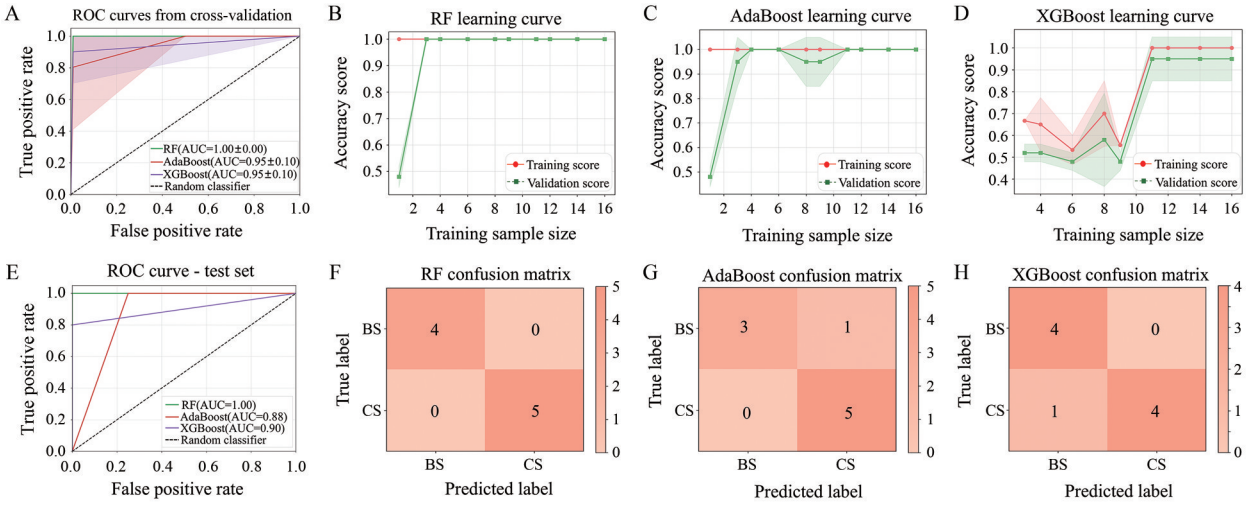


Figure 4 Comprehensive performance evaluation. A: ROC curves of the three machine learning classifiers (RF, AdaBoost, and XGBoost) obtained from 5-fold cross-validation; B−D: Learning curves of the RF, AdaBoost and XGBoost models; E: ROC curves of the RF, AdaBoost and XGBoost model on the test set; F−H: Confusion matrix of the RF, AdaBoost and XGBoost models on the test set. ROC: Receiver operating characteristic; RF: Random forest; AdaBoost: Adaptive boosting; XGBoost: Extreme gradient boosting

Table 4 Comparison of model performance using 5-fold cross-validation

Model	Dataset	Accuracy	Precision	Recall	F1	AUC
RF	Training set	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Validation set	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
AdaBoost	Training set	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Validation set	0.95 ± 0.10	0.97 ± 0.07	0.95 ± 0.10	0.95 ± 0.11	0.95 ± 0.10
XGBoost	Training set	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	Validation set	0.95 ± 0.10	0.97 ± 0.07	1.00 ± 0.10	0.95 ± 0.11	0.95 ± 0.10

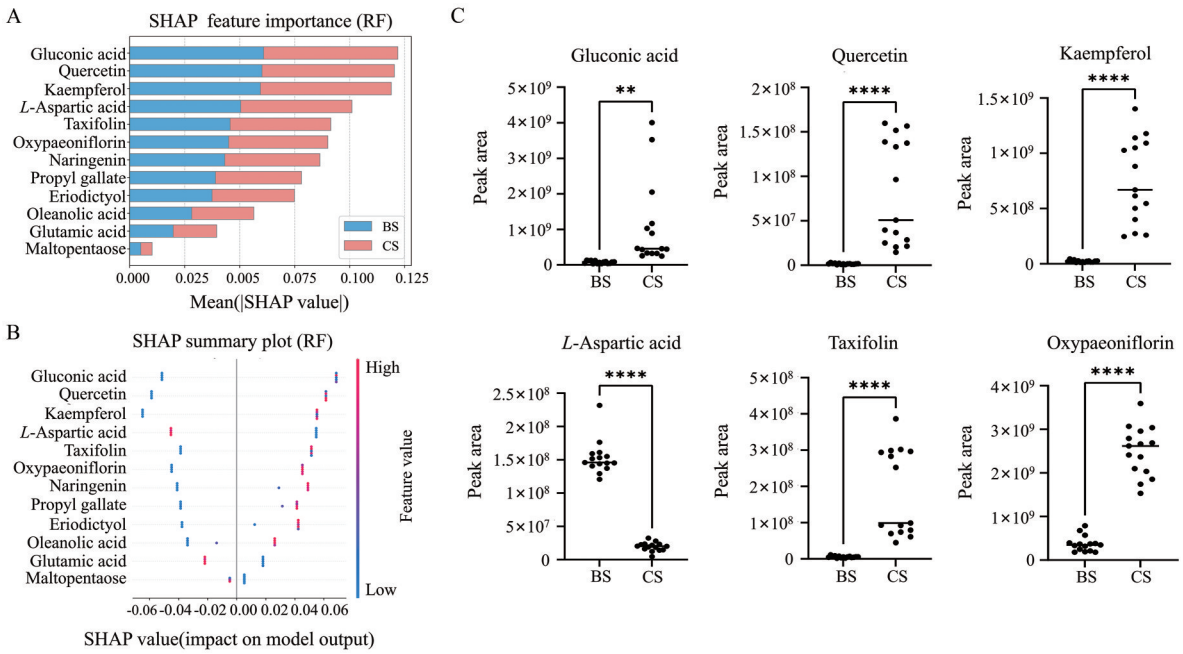


Figure 5　Importance assessment and differential comparison of characteristic metabolites based on SHAP value analysis. A: Importance ranking plot of characteristic metabolites based on SHAP value analysis; B: Beeswarm plot showing the contribution distribution of SHAP values for characteristic metabolites; C: Comparison of raw peak areas of the top six key characteristic metabolites between BS and CS. $n = 15$, $\bar{x} \pm s$. $^{**}P < 0.01$, $^{****}P < 0.000\ 1$. SHAP: Shapley additive explanation

(图 5C), 结果表明, 1 种代谢物在白芍中的含量高于赤芍, 其余 5 种代谢物在赤芍中的含量均显著高于白芍 ($P < 0.05$), 该结果与 SHAP 值分析相互验证, 证实了筛选获得的特征代谢物具有可靠的鉴别价值, 可作为区分白芍和赤芍的特征代谢物。该研究为两种药材的质量评价体系提供了新的实验数据和科学依据。

讨论

本研究基于非靶向代谢组学对白芍与赤芍饮片的化学成分进行了系统性分析, 共鉴定出 44 种代谢物, 涵盖黄酮类、酚酸类、氨基酸及其衍生物等多种化学类型, 较为全面地呈现了两种药材的代谢物组成特征。结果显示, 黄酮类化合物在两类饮片中分布最为丰富, 既往研究指出, 此类化合物可参与氧化应激及炎症反应调控, 并与血管功能和血液循环密切相关[34]。在鉴定的代谢物中, 共有 12 种代谢物在两类饮片间表现出显著差异, 提示白芍和赤芍在化学组成上具有较高区分度。其中白芍 *L*-天冬氨酸和谷氨酸含量较高, 与既往研究报道一致[14,15]。有研究表明, *L*-天冬氨酸和谷氨酸分别具有保肝和抗抑郁的作用, 这与白芍养血柔肝、疏肝解郁和柔肝止痛的功效高度契合[35,36]。此外, 赤芍中没食子酸丙酯、圣草酚及齐墩果酸含量显著升高, 已有文献提示其在清热、抗炎及血管保护等作用方面具有一定潜力, 这一发现与赤芍清热凉血、活血散瘀、泻火解毒的特点相符[37-39]。

进一步分别构建了 RF、AdaBoost 和 XGBoost 模型, 以比较不同算法在代谢组学判别任务中的适用性。其中, RF 表现最优, 测试集 AUC 达 1.00, 表明其对白芍与赤芍具有优异的判别能力, 且其对高维非线性数据的适应性与代谢组学特征高度契合[40]。SHAP 可解释性分析结果显示, 葡萄糖酸、槲皮素、山柰酚、二氢槲皮素及氧化芍药苷在赤芍中显著富集, 并在模型中具有较高贡献度。其中槲皮素作为一种黄酮类抗氧化剂, 可通过抑制核因子 κB (nuclear factor-κB, NF-κB) 通路显著减轻炎症反应[41]; 山柰酚不仅具有清除自由基的能力, 还能抑制血小板聚集, 从而促进血液循环[42]; 二氢槲皮素在稳定血管内皮功能和缓解氧化应激损伤方面表现突出[43]; 氧化芍药苷作为芍药苷的活性氧化衍生物, 兼具抗炎和镇痛作用[44]。上述 4 种成分与赤芍“清热凉血、散瘀止痛”的传统功效高度契合。葡萄糖酸虽无明确直接药理活性, 但其高 SHAP 值提示其可能作为关键代谢中间体, 参与活性成分的糖基化修饰或生物合成前体供应。综上, 这些关键代谢物不仅构成了两类饮片的化学判别基础, 也在功能层面为其传统功效差异提供了分子层面的解释。尽管其药效关联仍需实验验证, 但本研究为白芍与赤芍的物质基础解析及后续功能研究提供了新的线索。

该研究不仅揭示了白芍和赤芍在化学组成上的差

异, 也展示了代谢组学结合可解释性机器学习在中药研究中的优势。传统多元统计方法在高维数据分析中存在局限, 而机器学习模型可挖掘复杂变量关系, 并通过SHAP等工具实现结果可视化和特征重要性解析, 从而在提高分类准确性的同时明确关键差异代谢物[45,46]。该策略为中药饮片质量评价及潜在药效差异物质基础研究提供了方法学参考。

尽管本研究揭示了白芍与赤芍在代谢轮廓上的显著差异及其与传统功效的潜在关联, 但仍存在一定局限。首先, 样品来源地域分布不均, 白芍主要来自安徽, 赤芍多采自北方, 观测到的代谢差异可能受药材品种与产地因素共同影响。后续研究应收集来自相同产区的白芍和赤芍样品, 以更准确评估产地在代谢差异中的贡献, 同时, SHAP分析识别出的特征性代谢物虽具相关生物活性报道, 但其在白芍与赤芍药效差异中的具体贡献仍需通过体外或体内功能实验进一步验证。

综上, 本研究构建了基于非靶向代谢组学与可解释性机器学习的白芍和赤芍化学差异表征体系, 揭示了两类饮片在代谢物组成上的系统性差异, 并识别出多种具有潜在鉴别价值的关键差异代谢物。这些结果为后续探讨白芍与赤芍药效差异的物质基础研究提供了重要的指向性线索, 同时也为其他易混淆中药的基原鉴别、产地溯源及特征代谢物挖掘提供了新的思路。

作者贡献: 陈琪负责实验方案设计、实验开展、数据的整理及论文的撰写; 范国旗参与研究设计、实验实施及机器学习数据部分工作; 郭伊瑞参与代谢组学数据处理; 左莉华协助样品收集及实验方案优化; 李寒冰负责药材鉴定和质量控制; 邓丹萍、朱承慧协助论文修改和完善; 孙志、刘改枝负责研究总体设计、思路指导及指导文章修改。

利益冲突: 所有作者均声明不存在利益冲突。

References

[1] Chinese Pharmacopoeia Commission. Pharmacopoeia of the People's Republic of China (中华人民共和国药典) [S]. Beijing: China Medical Science Press, 2020: 108, 165.

[2] Tan YQ, Chen HW, Li J, et al. Efficacy, chemical constituents, and pharmacological actions of Radix Paeoniae Rubra and Radix Paeoniae Alba [J]. Front Pharmacol, 2020, 11: 1054.

[3] Leng JY, Ma YN, Zhang C, et al. Advances in chemical constituents, pharmacological effects and clinical application of Baishao (Paeoniae Radix Alba) [J]. J Liaoning Univ Tradit Chin Med (辽宁中医药大学学报), 2025, 27: 124-131.

[4] Zhang L. Clinical observation of Danggui-Shaoyao San with modified formula plus Tiao-Jing Zhi-Tong patches for treatment of primary dysmenorrhea [J]. J Pract Tradit Chin Med (实用中医药杂志), 2025, 41: 1139-1142.

[5] Ji QR, Zhang H, Wang XY, et al. Analysis of the dosing pattern of blood-activating preparations based on the 2020 edition of Pharmacopoeia of the People's Republic of China (Part I) [J]. J Liaoning Univ Tradit Chin Med (辽宁中医药大学学报), 2024, 26: 77-85.

[6] Yan LH, Jiang ZH, Zhen YC, et al. The substances and mechanism of Paeoniae Radix Rubra in the treatment of sepsis based on combination of UPLC-Q/TOF-MS with network pharmacology [J]. Asia-Pac Tradit Med (亚太传统医药), 2025, 21: 11-18.

[7] Li C, Wang P, Li M, et al. The current evidence for the treatment of sepsis with Xuebijing injection: Bioactive constituents, findings of clinical studies and potential mechanisms [J]. J Ethnopharmacol, 2021, 265: 113301.

[8] Parker S, May B, Zhang C, et al. A pharmacological review of bioactive constituents of *Paeonia lactiflora* Pallas and *Paeonia veitchii* Lynch [J]. Phytother Res, 2016, 30: 1445-1473.

[9] Xu SY, Liu WL, Liu XB, et al. DUS evaluation of nine intersubgeneric hybrids of *Paeonia lactiflora* and fingerprint analysis of the chemical components in the roots [J]. Front Chem, 2023, 11: 1158727.

[10] Zhu S, Yu XL, Wu YQ, et al. Genetic and chemical characterization of white and red peony root derived from *Paeonia lactiflora* [J]. J Nat Med, 2015, 69: 35-45.

[11] Liu J, Chen L, Fan CR, et al. Qualitative and quantitative analysis of major constituents of Paeoniae Radix Alba and Paeoniae Radix Rubra by HPLC-DAD-Q-TOF-MS/MS [J]. China J Chin Mater Med (中国中药杂志), 2015, 40: 1762-1770.

[12] Yu JJ, Wu JX, Liang YF, et al. Simultaneous quantitative determination of six active compounds from Paeoniae Rubra Radix and Paeoniae Alba Radix by HPLC [J]. Chin Tradit Herb Drugs (中草药), 2015, 46: 1673-1677.

[13] Shen ML, Zhang QY, Qin LP, et al. Single standard substance for the simultaneous determination of eleven components in the extract of Paeoniae Radix Alba (root of *Paeonia lactiflora* Pall.) [J]. J Anal Methods Chem, 2021, 2021: 1-12.

[14] Li SS, Zhang YP, Deng YH, et al. Evaluation of Baishao (Paeoniae Radix Alba) and Chishao (Paeoniae Radix Rubra) from different origins based on characteristic spectra of amino acids [J]. Digit Chin Med, 2021, 4: 120-129.

[15] Fan ML, Xing J, Li ZY, et al. Comparison on chemical constituents between Paeoniae Alba Radix and Paeoniae Rubra Radix using NMR based metabolomic approach [J]. Chin Tradit Herb Drugs (中草药), 2014, 45: 3230-3237.

[16] Ji PR, Yang XQ, Zhao XS. Application of metabolomics in quality control of traditional Chinese medicines: A review [J]. Front Plant Sci, 2024, 15: 1463666.

[17] Shen XT, Shen HF, Wang F, et al. Comparative non-targeted metabolomics reveals distinct metabolic profiles and functional traits in six mung bean (*Vigna radiata*) varieties [J]. PLoS One,

2025, 20: e0327962.

[18] Erdős B, Westerhuis JA, Adriaens ME, et al. Analysis of high-dimensional metabolomics data with complex temporal dynamics using RM-ASCA [J]. PLoS Comput Biol, 2023, 19: e1011221.

[19] Wang JJ. Biostatistical challenges in high-dimensional data analysis: Strategies and innovations [J]. Comput Mol Biol, 2024, 14: 163-172.

[20] Banerjee S, Nakrani P, Singh A, et al. DuReS: An R package for denoising experimental tandem mass spectra and metabolite annotation [J]. Anal Chem, 2025, 97: 11986-11992.

[21] Ma CH, Hao XL, Meng XF, et al. Survey on machine learning for multi-dimensional date query processing [J]. Chin J Comput (计算机学报), 2025, 48: 100-123.

[22] Bifarin OO. Interpretable machine learning with tree-based Shapley additive explanations: application to metabolomics datasets for binary classification [J]. PLoS One, 2023, 18: e0284315.

[23] Rivera-Pérez A, Garrido Frenich A. Comparison of data processing strategies using commercial *vs*. open-source software in GC-Orbitrap-HRMS untargeted metabolomics analysis for food authentication: Thyme geographical differentiation and marker identification as a case study [J]. Anal Bioanal Chem, 2024, 416: 4039-4055.

[24] Sheng H, Lu JL, Zhang RX, et al. Analysis of chemical constituents of Ludangshen oral liquid based on UPLC-Q-Exactive Orbitrap HRMS technology [J]. Food Drug (食品与药品), 2025, 27: 317-324.

[25] Guan YH, Liang ZW, Shi Y, et al. Differences in chemical constituents of *Polygonatum kingianum* before and after processing [J]. Chin Tradit Herb Drugs (中草药), 2024, 55: 3647-3658.

[26] Azis H. Assessing the performance of logistic regression in heart disease detection through 5-fold cross-validation [J]. Int J Artif Intell Med Issues, 2024, 2: 1-11.

[27] Ullah B, Kamran M, Rui Y. Predictive modeling of short-term rockburst for the stability of subsurface structures using machine learning approaches: t-SNE, K-means clustering and XGBoost [J]. Mathematics, 2022, 10: 449.

[28] Hu HR, Zhang RH, Du L, et al. Differential analysis of Yunnan Qiubei Chili metabolites based on non-targeted metabonomics [J]. Food Res Dev (食品研究与开发), 2024, 45: 155-163.

[29] Feng F, Huang C, Luosang DZ, et al. Serum metabolomic analysis of synchronous estrus in yaks based on UPLC-Q-TOF MS technology [J]. Animals, 2024, 14: 1399.

[30] Kim TJ, Park JG, Ahn SK, et al. Discrimination of adzuki bean (*Vigna angularis*) geographical origin by targeted and non-targeted metabolite profiling with gas chromatography time-of-flight mass spectrometry [J]. Metabolites, 2020, 10: 112.

[31] Murugesu S, Ibrahim Z, Ahmed QU, et al. Characterization of *α* -glucosidase inhibitors from *Clinacanthus nutans* Lindau leaves by gas chromatography-mass spectrometry-based metabolomics and molecular docking simulation [J]. Molecules, 2018, 23: 2402.

[32] Laref S, Harrou F, Sun Y, et al. Exploring antiviral drugs on monolayer black phosphorene: Atomistic theory and explainable machine learning-assisted platform [J]. Int J Mol Sci, 2024, 25: 4897.

[33] Wang HG, Hu MT, Liu XC. Reconstruction and interpretability analysis of China's food security level based on machine learning and SHAP algorithm [J]. J Chin Agric Univ (中国农业大学学报), 2025, 30: 264-274.

[34] Seelinger G, Merfort I, Schempp C. Anti-oxidant, anti-inflammatory and anti-allergic activities of luteolin [J]. Planta Med, 2008, 74: 1667-1677.

[35] Su R, Fu HL, Zhang QX, et al. Amplifying hepatic *L*-aspartate levels suppresses CCl_4-induced liver fibrosis by reversing glucocorticoid receptor *β* -mediated mitochondrial malfunction [J]. Pharmacol Res, 2024, 206: 107294.

[36] McIntyre RS, Jain R. Glutamatergic modulators for major depression from theory to clinical use [J]. CNS Drugs, 2024, 38: 869-890.

[37] Momeni Z, Danesh S, Ahmadpour M, et al. Protective roles and therapeutic effects of gallic acid in the treatment of cardiovascular diseases: Current trends and future directions [J]. Curr Med Chem, 2024, 31: 3733-3751.

[38] Lee S, Yang H, Son G, et al. Eriodictyol protects endothelial cells against oxidative stress-induced cell death through modulating ERK/Nrf2/ARE-dependent heme oxygenase-1 expression [J]. Int J Mol Sci, 2015, 16: 14526-14539.

[39] Wang DD, Zhang JP, Sun LL, et al. Protective effects and mechanisms of oleanolic acid on damaged human umbilical vein endothelial cells [J]. J Nanjing Med Univ Nat Sci (南京医科大学学报·自然科学版), 2012, 32: 1505-1510.

[40] Wang T, Li S, Wu Y, et al. Mechanistic investigation of Xuebijing for treatment of paraquat-induced pulmonary fibrosis by metabolomics and network pharmacology [J]. ACS Omega, 2021, 6: 19717-19730.

[41] Lan Y, Dong C, Wu M, et al. Quercetin ameliorates epithelial-mesenchymal transition and inflammation by targeting FSTL1 and modulating the NF-*κ*B pathway in pulmonary fibrosis [J]. Front Pharmacol, 2025, 16: 1594757.

[42] Fuentes E, Rodríguez L, Méndez D, et al. Inhibition of platelet aggregation by extracts and compounds from the leaves of chilean bean landraces (*Phaseolus vulgaris* L.) [J]. J Funct Foods, 2023, 100: 105388.

[43] Liskova S, Cacanyiova S, Cebova M, et al. Taxifolin reduces blood pressure *via* improvement of vascular function and mitigating the vascular inflammatory response in spontaneously hypertensive rats [J]. Int J Mol Sci, 2023, 24: 12616.

[44] Lv K, Jiang M. Ameliorative efficacy of oxypaeoniflorin, a traditional Chinese medicine monomer against adjuvant-induced

arthritic inflammation and pain [J]. Pharmacogn Mag, 2022, 18: 103.

[45] Cui T, Wang N, Shang L, et al. Application of metabolomics and MCDM approach in developing a novel strategy for disease diagnosis: A case study in primary Sjögren's syndrome [J]. J Pharm Biomed Anal, 2025, 266: 117080.

[46] Wang F, Meng QQ, Liu Q, et al. Integrated machine learning and network pharmacology approach for exploring the material basis of Buyang Huanwu Decoction for ischemic stroke [J]. Acta Pharm Sin (药学学报), 2025, 60: 627-636.