

DOI: 10.12326/j.2096-9694.2026013

实时直接分析质谱结合机器学习的 栎属木材树种识别方法研究

刘守佳, 詹伟辉, 李天笑, 焦立超, 殷亚方

(中国林业科学研究院木材工业研究所; 中国林业科学研究院木材标本馆; 国家林业和草原局木材标本资源库;

CITES全球野生动植物鉴定实验室(木材与木制品), 北京 100091)



摘要: 高效提取木材特征信息从而建立科学准确的木材树种识别方法, 对推动生物多样性保护、林木资源可持续利用等具有重要现实意义。然而现有遗传学、计算机视觉等方法仍难以满足快速精准应用场景下的树种鉴定需求。以构造特征高度相似的3种栎属木材为研究对象: 美国白栎 (*Quercus alba*)、蒙古栎 (*Q. mongolica*) 和夏栎 (*Q. robur*), 对其开展树种识别方法研究。采用实时直接分析-飞行时间质谱技术 (Direct Analysis in Real Time-Time-of-Flight Mass Spectrometry, DART-TOF-MS) 获取木材化学指纹图谱, 通过构建逻辑回归、多层感知机、随机森林和K-近邻算法等4种机器学习方法分析木材质谱数据, 评估不同机器学习方法在树种识别中的可行性, 同时比较质谱数据预处理参数对识别模型性能的影响。研究表明: 逻辑回归算法具有最高的树种识别准确率 (达89.5%), 性能优于其他3种模型; 不同机器学习方法在处理复杂质谱特征时表现出较明显差异, 合理的分箱参数设置对模型性能具有关键影响, 在相对峰强度阈值为3%、分箱容差为20 mmu时模型整体表现最优; 3种栎属木材在木质素单体及其衍生物组成上具有高度相似性, 但在黄酮类化合物、木质素二聚体及鞣花酸的空间分布方面存在差异。DART-MS结合机器学习方法对3种栎属木材呈现出较强的识别能力, 可为木材树种的精准识别提供科学依据和技术路径。

关键词: 栎木; 木材识别; 机器学习; 实时直接分析质谱技术 (DART-MS); 逻辑回归算法

中图分类号: TS67; S781.1; TP181 文献标识码: A 文章编号: 2096-9694 (2026) 03-0020-10

Machine-Learning-Based Identification Method for *Quercus* Wood Species Using Direct Analysis in Real Time Mass Spectrometry

LIU Shoujia, ZHAN Weihui, LI Tianxiao, JIAO Lichao, YIN Yafang

(Research Institute of Wood Industry, Chinese Academy of Forestry; Wood Collection of Chinese Academy of Forestry;

Wood Specimen Resource Center of National Forestry and Grassland Administration;

CITES Wildlife Forensic Testing Laboratory (Timber and Timber Products), Beijing 100091, China)

Abstract: Establishing scientific and accurate wood species identification methods through the efficient extraction of characteristic information from different wood species holds significant practical importance for promoting biodiversity conservation and the sustainable utilization of forest resources.

收稿日期: 2026-01-21; 修改日期: 2026-03-07

基金项目: 中国林业科学研究院国际合作项目“木材标本资源信息深度挖掘与创新利用”(CAFYBB2022GA001)。

作者简介: 刘守佳 (1998—), 女, 博士研究生。Email: liushoujia1998@163.com

通讯作者: 殷亚方, 男, 研究员。Email: yafang@caf.ac.cn

However, existing techniques such as computer vision and DNA barcoding are still insufficient to meet the demand for rapid and accurate species discrimination, underscoring the need for methodological innovation in wood identification. In this study, three morphologically similar *Quercus* species common in trade: *Quercus alba*, *Quercus mongolic*, and *Quercus robur*, which are difficult to differentiate using conventional anatomical methods, were selected as investigated species to acquire spectral fingerprint data using Direct Analysis in Real Time-Time-of-Flight Mass Spectrometry (DART-TOF-MS). Four machine learning algorithms, including logistic regression, multilayer perceptron (MLP), random forest (RF), and k-nearest neighbors (KNN), were then applied to DART-TOF-MS data to assess their feasibility for species classification. We further investigated the influence of different spectral preprocessing parameters on model performance. The results showed that the logistic regression model achieved the highest classification accuracy (89.5%), outperforming the other three algorithms. Distinct differences were observed among the four machine learning approaches in their ability to handle complex spectral features, and appropriate binning strategies-particularly threshold control and mass tolerance selection-were found to be critical for optimal model performance, with the best results obtained at a parameter setting of 3% relative threshold and 20 mmu binning tolerance. Although the three *Quercus* species exhibited similarities in lignin monomers and their derivatives, significant differences were detected in the distribution of flavonoids, lignin dimers, and ellagic acid. The present results demonstrated that DART-TOF-MS combined with machine learning exhibited a strong capability for identifying the three *Quercus* wood species examined in this study, providing a scientific basis and technical pathway for the accurate identification of wood species.

Key words: *Quercus*; wood identification; machine learning; DART-MS; logistic regression algorithm

木材识别是维护森林生态安全、应对木材非法采伐贸易实践中的核心手段之一，对推动生物多样性保护、林木资源可持续利用具有重要意义，同时也是保障木制品质量、维护市场公平秩序、支撑文物精准修复等经济社会活动中不可或缺的技术保障。在木材工业与贸易领域，实现对“同属相近树种”的高效精准鉴定，一直是行业内长期面临且亟需解决的关键技术难题。

木材精准识别方法的实现，主要依托木材固有特征高效提取、关键判别标记筛选以及高质量科学数据库构建这三个核心环节。当前，依托具有准确植物分类学信息的木材标本资源，对不同树种间木材多源异构特征变异规律的认知不断加深，遗传学、计算机视觉、化学指纹等识别方法体系快速发展，突破传统木材解剖学的识别精度局限，为实现木材“种”水平的识别提供了可能^[1-2]。然而，通过遗传学信息识别方面，木材中脱氧核

糖核酸（deoxyribonucleic acid, DNA）常因树木木质部细胞程序化死亡过程而严重降解，加工处理或长期存放进一步增加了木材DNA提取与分析难度^[2-3]；计算机视觉方法主要依靠木材横切面宏观特征，而缺乏径切面和弦切面特征信息，导致在木材“种”水平的判别上仍存在一定程度的可靠性不稳定等问题。化学指纹方面，基于质谱的木材特征分析手段不断得到关注，特别是快速发展的实时直接分析质谱（Direct Analysis in Real-Time Mass Spectrometry, DART-MS）技术，因其无需经过复杂前处理即可快速获取木材化学指纹的优点，在高通量分析需求的木材识别领域具有较好的应用前景^[4-6]。

机器学习是数据挖掘的重要数据分析技术，能够从海量数据中学习数据的内在规律，进而挖掘出潜在的信息。近年来，机器学习的引入使得木材树种识别研究得到进一步发展和应用^[1,7]。从

高维、复杂的木材质谱数据中有效提取适用于树种精准判别的特征，是实现木材识别的重要基础。现有研究多采用的传统分析方法，如主成分分析（principal component analysis, PCA）、正交偏最小二乘判别分析（orthogonal partial least squares-discriminant analysis, OPLS-DA）大多基于线性假设，难以有效捕捉质谱数据中普遍存在的非线性关系，例如代谢物间的交互作用或复杂的噪声干扰^[5-6, 8]。相较于传统方法，随机森林、支持向量机、神经网络等机器学习方法能够自动学习变量间复杂的非线性模式，从而精确捕捉不同树种或产地木材特征的细微差别，在木材识别中展现出明显优势^[9-12]。同时，通过交叉验证与集成学习等机制，能有效避免此类方法产生过拟合问题，进而获得泛化能力更强、更稳健的预测模型，最终能够在多树种分类任务中实现更高的准确率^[13]。

为验证 DART-MS 与机器学习融合技术路线的可行性与有效性，研究选取构造特征高度相似的 3 种栎属木材：美国白栎（*Quercus alba*）、蒙古栎（*Quercus mongolica*）和夏栎（*Quercus robur*），其中蒙古栎为《濒危野生动植物种国际贸易公约》（CITES）附录 III 所列树种。采用 DART-MS 技术采集木材化学指纹图谱，并通过不同机器学习方法综合分析木材质谱数据，评估各方法在树种识别中的可行性和可靠度。研究旨在通过提升质谱数据的预处理流程与优化机器学习模型参数，构建一种快速、精准的木材化学指纹识别方法，为木材“种”水平识别提供可靠的理论依据和方法范式。

1 材料与方法

1.1 实验材料

选取壳斗科栎属的美国白栎、蒙古栎和夏栎 3 种木材，共收集 63 份来源于不同植株的样品。所有样品均来自于经正确定名的木材标本馆藏标本，保存条件为气干状态。蒙古栎样品由中国林业科学研究院木材标本馆（Wood Collection of Chinese Academy of Forestry, CAFw）提供，取样时间为 2024 年；美国白栎与夏栎样品由英国皇家植物园（Royal Botanic Gardens, Kew）提供，取样时间为 2021—2024 年间。

为提高与国内外质谱数据库的可比性，减少心/边材化学指纹图谱对树种识别模型的影响，研究均选择心材部位样品开展试验。样品产地与数量信息列于表 1。

表 1 三种栎属木材样品信息
Tab.1 Information on three *Quercus* wood samples

木材树种	产地	心材样本数量/份
美国白栎(<i>Quercus alba</i>)	美国	18
蒙古栎(<i>Quercus mongolica</i>)	中国黑龙江、北京、吉林	23
夏栎(<i>Quercus robur</i>)	西班牙、英国、克罗地亚、波兰	22

1.2 主要试剂

聚乙二醇 600、无水乙醇、丙三醇、番红染液、二甲苯、中性树胶等，均外购。

1.3 主要设备与仪器

滑走切片机（Leica SM2010R）、光学显微镜（Olympus BX50）、木材宏观图像采集仪器 iWood 3.0^[14]、DART-SVP 实时直接分析质谱离子源（IonSense）、高解析飞行时间质谱仪（JEOL USA）等。

1.4 试验方法

1.4.1 木材宏观构造图像采集 分别采用 80、180、240、400、600 目砂纸，依次对 3 种栎木样品横切面与纵切面进行砂光，至其表面平滑、无明显划痕。采用图像采集装置 iWood 3.0 对砂光后的样品截面进行图像采集，图像分辨率为 2 048 像素×2 048 像素。木材宏观构造特征参照国际木材解剖学家协会（International Association of Wood Anatomists, IAWA）阔叶树木材特征标准进行记载^[15]。

1.4.2 木材微观构造图像采集 制备木块尺寸为 10 mm×10 mm×10 mm，经过热水法煮软化、切片、染色、脱水、透明及封片等步骤，完成永久切片制备。将切片置于光学显微镜载物台上，采集三切面切片图像，图像分辨率为 4 096 像素×2 160 像素。木材微观构造特征参照 IAWA 阔叶树木材特征要求进行记录^[15]。

1.4.3 DART-TOF-MS 测试 采用 DART-SVP 实时直接分析质谱离子源结合高解析飞行时间质谱仪（TOF-MS）对 63 份样品进行质谱采集。从每份样

品上制备3个尺寸为15 mm×5 mm×1 mm的小木片,对每个小木片各进行1次质谱测试。测试时,用镊子夹取小木片置于离子源气出口与质谱仪进样口之间,在每完成一个样本的测试后,进行一次聚乙二醇600测试,用于质量校准。

DART-SVP实时直接分析质谱离子源参数设置:气体源为氦气,载气温度350℃。高解析飞行时间质谱仪参数设置:正离子模式,采谱范围质荷比(m/z)90~800。以聚乙二醇质谱数据为基准,逐一对所有样品的质谱数据进行校准,导出TXT格式数据用于分析与机器学习模型构建。

1.5 质谱数据机器学习模型构建

1.5.1 数据预处理 运用质谱分析软件 Mass Mountaineer 7.1.28.0 (Cody, USA),通过设置分箱容差与相对峰强度阈值两项参数,对原始质谱数据进行预处理,从而提高后续化学计量学模型的准确性和效率。分箱容差是指在对质谱数据做“分箱或对齐”预处理时,为了把代表同一离子的多个质荷比(m/z)测量值归入同一个变量而允许的误差窗口。相对峰强度阈值则是指把多个近似的质荷比测量值视为同一化学特征时所允许的质量偏差。分箱容差与相对峰强度阈值是一对关键参数,共同决定最终变量数量与分类准确率^[11]。共设置18种预处理参数,详细信息列于表2。

表2 DART-TOF-MS质谱数据预处理参数组合

Tab.2 Preprocessing parameter combinations of

DART-TOF-MS data

序号	相对峰强度 阈值/%	分箱容差/ mmu	序号	相对峰强度 阈值/%	分箱容差/ mmu
1	1	1	10	3	20
2	1	5	11	3	30
3	1	10	12	3	50
4	1	20	13	5	1
5	1	30	14	5	5
6	1	50	15	5	10
7	3	1	16	5	20
8	3	5	17	5	30
9	3	10	18	5	50

1.5.2 机器学习模型构建 首先划分质谱数据集。预处理后的质谱数据集按6:4的样本比例,划分为训练集与独立测试集,其中40%的样本作为独立测试集不参与模型训练与超参数调优。数据集

划分均在样本水平执行,避免同一样本的数据被拆分至两个数据集中,确保测试集完全由模型未接触过的独立样本构成。将剩余60%样本训练集输入机器学习模型,采用5倍交叉验证完成最优超参数的搜索与确定。最终利用独立测试集对训练后的模型进行评估,完成独立数据泛化验证。

选取构建逻辑回归(logistic regression, LR)、多层感知机(multilayer perceptron, MLP)、随机森林(random forest, RF)和K-近邻算法(K-nearest neighbor, KNN)等4种常用的机器学习算法。逻辑回归作为一种广义线性模型,通过对输入变量与分类结果之间的线性关系建模,能够有效处理二分类及多分类问题,具有较好的可解释性^[16]。多层感知机属于前馈神经网络,通过多层非线性映射学习输入与输出间的复杂关系,适用于处理非线性特征较强的数据^[17]。随机森林是一种集成学习方法,通过构建并集成多棵决策树以提升模型的鲁棒性和泛化能力,在处理高维数据及特征选择方面表现突出^[18]。K-近邻算法则基于样本间的距离度量,通过判定待分类样本在特征空间中最邻近的类别进行预测,具有实现简单、对数据分布无强假设的特点^[19]。

模型构建的软硬件配置如下:Windows 11 Pro 64位操作系统,Python 3.7.13编程语言,PyTorch 1.12.1深度学习框架,NVIDIA CUDA 10.2并行计算平台,Intel® Arc™ Graphics显卡,32 GB LPDDR5X内存,Intel®Core™ Ultra 7 258V中央处理器。

1.5.3 模型性能评估 采用准确率(Accuracy)、宏平均F1值(Macro-F1)与加权F1值(Weighted-F1)3类指标评估模型性能,各指标计算公式分别见式(1)~式(3)。

$$\text{Accuracy} = \frac{\sum_{i=1}^C (\text{TP}_i + \text{TN}_i)}{\sum_{i=1}^C (\text{TP}_i + \text{TN}_i + \text{FP}_i + \text{FN}_i)} \quad (1)$$

$$\text{Macro-F1} = \frac{1}{C} \sum_{i=1}^C \frac{2 \text{TP}_i}{2 \text{TP}_i + \text{FP}_i + \text{FN}_i} \quad (2)$$

$$\text{Weighted-F1} = \frac{1}{N} \sum_{i=1}^C n_i \cdot \frac{2 \text{TP}_i}{2 \text{TP}_i + \text{FP}_i + \text{FN}_i} \quad (3)$$

式中: TP_i 为第*i*类真正例; TN_i 为第*i*类真负例; FP_i 为第*i*类假正例; FN_i 为第*i*类假负例; N 为全数据集总样本量; C 为分类任务的总类别数量; n_i 为第*i*类样本总数量。

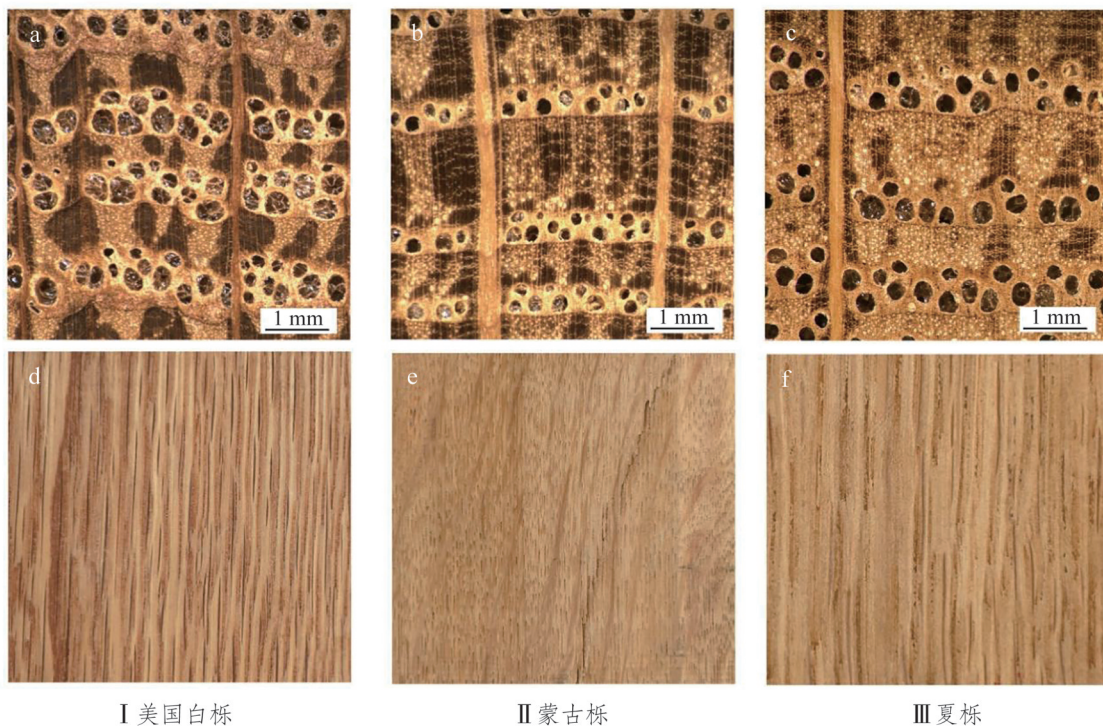
2 结果与分析

2.1 栎属木材宏、微观构造特征

3种栎属木材具有相似的宏观构造特征(图1)。心材颜色呈浅黄色至暗褐色;生长轮明显,为环孔材;在导管排列上,早材导管多为2~3列,三者晚材导管均呈火焰状排列;轴向薄壁组织呈环管状及带状;具宽木射线;木材纹理直,结构粗。

图2显示3种栎属木材横、径、弦三切面的微观构造特征。3个树种木材均以单管孔为主,侵填

体常见,具有单穿孔结构,导管间纹孔互列,存在环管管胞;木纤维细胞壁薄至厚;轴向薄壁组织主为星散聚合状、环管状及带状(宽1~3个细胞);木射线非叠生,以单列木射线为主,具有单列木射线及宽木射线,射线组织类型为同形,由方形和长方形薄壁细胞组成,射线细胞内常可见棱晶类结晶;导管射线间纹孔以刻痕状为主,部分存在类似管间纹孔式形态。总的看来,3种栎属木材宏观与微观构造特征具有高度相似性,难以通过单一解剖构造有效区分。



注:上行(a~c)为横切面;下行(d~f)为纵切面。

图1 三种栎属木材的宏观构造特征

Fig.1 Macroscopic features of three *Quercus* species (I *Q. alba*; II *Q. mongolica*; III *Q. robur*)

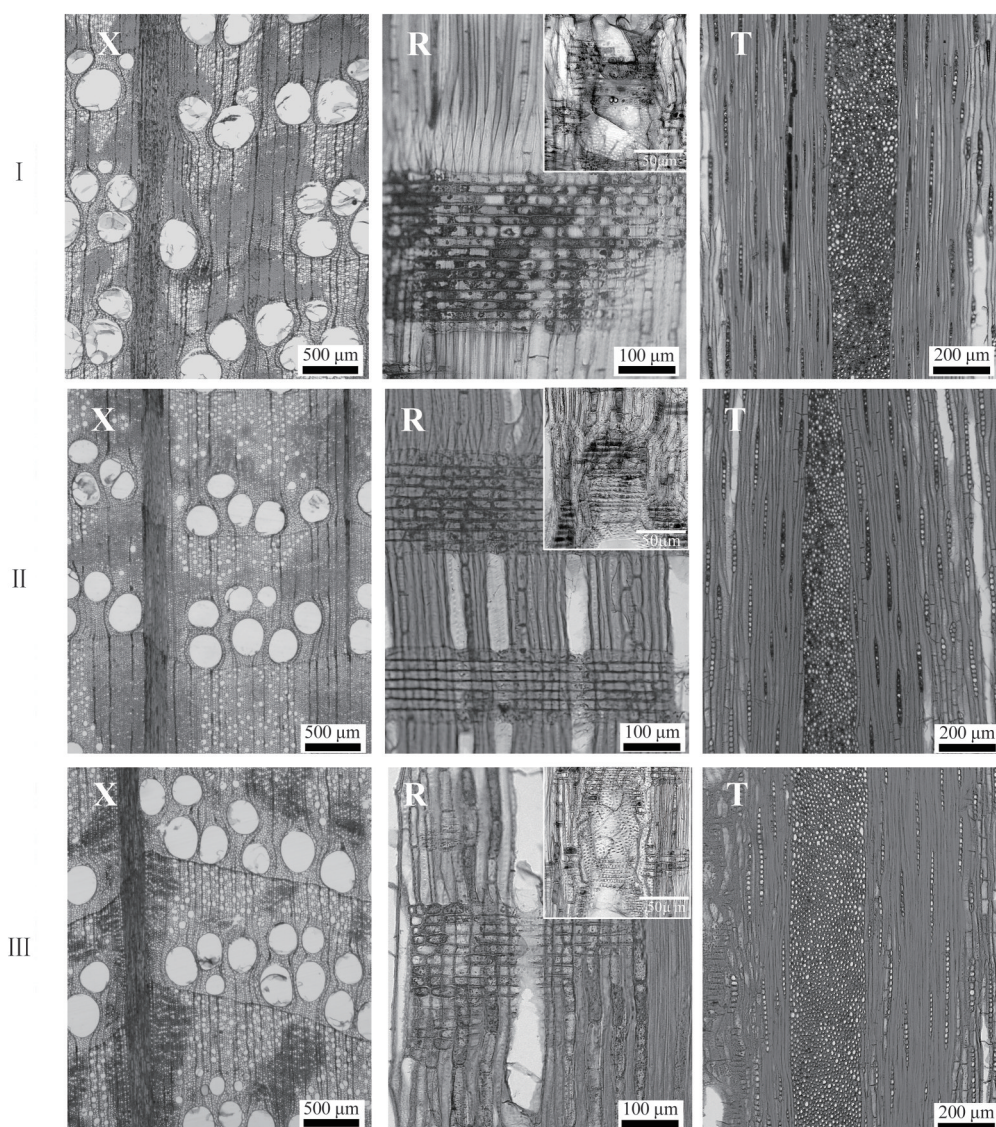
2.2 栎属木材DART-TOF-MS特征

3种栎属木材的质谱热图(图3a)显示,从整体分布上看,信号均主要富集于低质荷比区间(90~230),呈现密集且高强度特征;而在中高质荷比区间(230~800),信号则趋于稀疏,强度减弱。同时,不同树种在代谢物质的质谱分布上呈现一定差异。具体而言,美国白栎和蒙古栎的信号峰主要集中于低质荷比区间,分布相对集中;而夏栎在中等质荷比区间保持一定的信号强度。

23份蒙古栎木材样品在 m/z 249~252区间内虽

然均出现峰信号,但强度差异较大,其中第29号样品在该区间的相对峰强度低于1.5%。这种个体间变异主要源于基因型与环境因子的共同作用。即使是同一树种,不同产地的样本在生长过程中所经历的温度、降水等环境条件的差异,也会导致其化学表型产生波动,从而成为木材难以实现“种”水平精准识别的主要原因^[20]。

3种栎属木材在DART-TOF-MS正离子模式下的代表性质谱图如图3b所示。分析结果显示,三者均含有典型木质素降解产物,包括愈创木醇



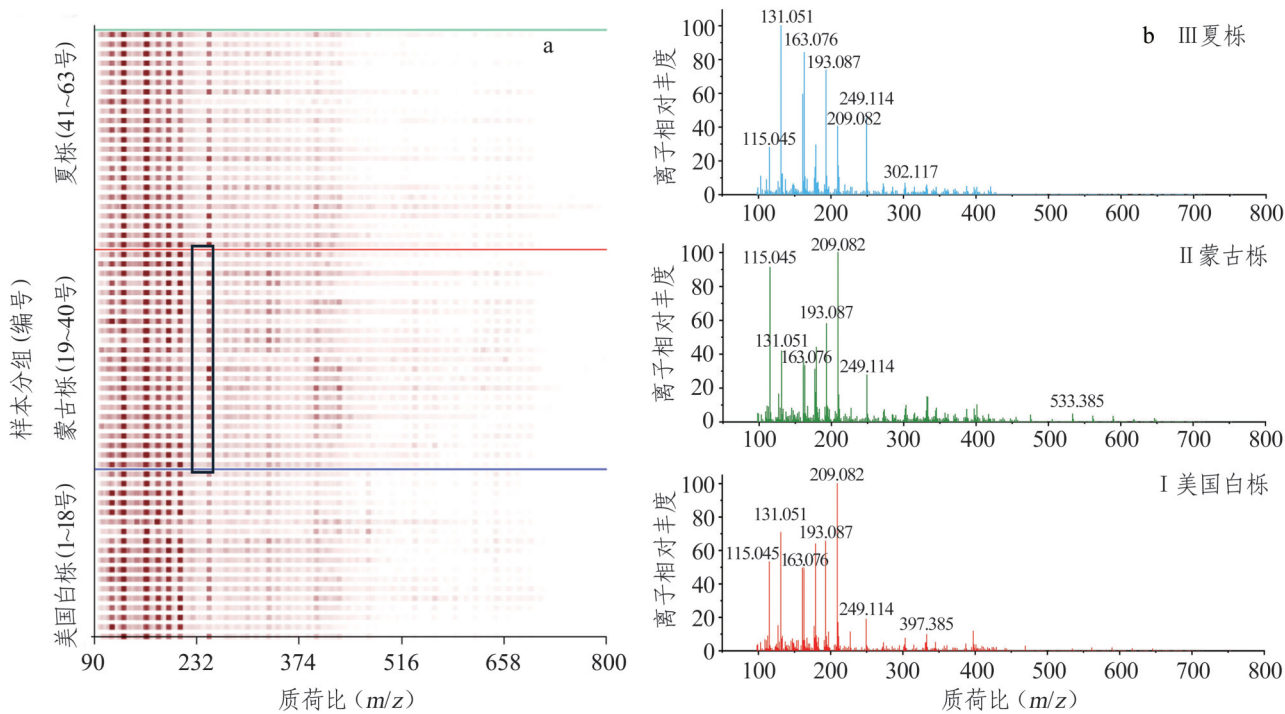
注: I 美国白栎, II 蒙古栎, III 夏栎; X 横切面, R 径切面, T 弦切面。

图2 三种栎属木材三切面微观构造特征

Fig.2 Microscopic features of three *Quercus* species (I *Q. alba*; II *Q. mongolica*; III *Q. robur*)

(Coniferyl alcohol)、愈创木醛 (Coniferaldehyde)、丁香醛 (Syringaldehyde) 及 辛阿醛 (Sinapaldehyde), 表明其木质素结构均由愈创基 (G型) 和紫丁香基 (S型) 单元构成, 符合典型阔叶材木质素的化学特征^[21-22]。值得注意的是, 辛阿醛在3种栎属木材中均呈现为强峰, 提示S型结构单元在栎属木材木质素结构中占重要地位。除共有特征外, 不同树种亦表现出化学成分的差异性。通过对3种栎属木材信号峰的检索, 部分信号峰的初步归属信息列于表3。通过对质谱信号峰的初步解析可明确, 3种栎属木材的信号峰差异主要

体现在黄酮类化合物、木质素二聚体及鞣花酸等物质上。在夏栎样本中检测到鞣花酸信号峰 (质荷比 m/z 302.117), 该类化合物通常来源于鞣质类物质的降解, 说明夏栎多酚类次生代谢产物丰富度明显高于另外2种栎木。该结果与 PUECH 等^[23]、CHATONNET 与 BOIDRON^[26] 关于夏栎在酿酒陈酿过程中释放复杂风味物质的研究相吻合。蒙古栎则在 m/z 533.385 出现特征离子峰, 经初步质谱归属推测为高分子多酚或鞣质片段, 显示出与美国白栎、夏栎的次生代谢物组成差异^[27]。尽管存在上述差异, 3种栎属木材在多个主峰信号 (如 m/z



注:a 每份样本的代表性质谱热图(颜色深浅代表离子的相对信号强度,深红色表示较高离子峰丰度);b 代表性质谱图。

图3 三种栎属木材质谱图
Fig.3 Mass spectrometry spectra of three *Quercus* species

表3 DART-MS部分信号峰归属的基本信息

Tab.3 Preliminary assignment of selected DART-MS spectra peaks

观测质量	理论质量	分子式	可能化合物(分子式)	关联树种	参考文献
273.1134	273.0758	$C_{15}H_{12}O_5$	Naringenin	美国白栎、夏栎	[24]
303.1241	303.1233	$C_{14}H_{14}O_3$	4,4'-Dihydroxy-3,3',5'-trimethoxystilbene(G,S 二聚体)	美国白栎、夏栎	[25]
333.1344	333.1338	$C_{18}H_{20}O_6$	4,4'-Dihydroxy-3,3',5,5'-tetramethoxystilbene(S,S 二聚体)	美国白栎、夏栎	[25]
387.1448	387.1444	$C_{21}H_{22}O_7$	7-Dehydromedioresinol(G,S 二聚体)	美国白栎	[25]
401.1618	401.1600	$C_{22}H_{24}O_8$	5,5'-Dimethoxy-8,8'-didehydro-3,4-divanillyltetrahydrofuran(S,S 二聚体)	美国白栎、夏栎	[25]
302.1973	302.1920	$C_{14}H_6O_8$	Ellagic acid(鞣花酸)	夏栎	[23, 24]

115.045、131.051、163.076、193.087、249.114等)上仍具有高度一致性,说明其基本化学组成较为相似。为进一步消除主观误差,提高鉴定精度,引入机器学习方法对获得的有效质谱峰进行综合比对与模式识别,以实现快速、客观的树种判别。

2.3 机器学习结合DART-TOF-MS的木材识别方法

2.3.1 质谱预处理参数优化 为评估质谱数据预处理对木材树种识别准确率的影响,比较4种机器学习模型在不同相对峰强度阈值与分箱容差组合下的树种识别准确率(图4)。总体而言,不同模型的性能随参数设置呈现差异化趋势,分箱参数(相对峰强度阈值与分箱容差)的设置对木材树种识别准确率产生较大影响。其中,逻辑回归模型

对分箱参数最为敏感,在较低相对峰强度阈值(1%)结合中等分箱容差(10、20 mmu)时准确率高于0.800;然而随着分箱容差进一步增大,准确率下降。这说明参数设置并非越高越优,过高参数在抑制噪音的同时,也可能削弱具有判断力的特征信息。相比之下,多层感知机模型的整体准确率多处于0.600~0.700之间,随参数变化的波动幅度较小,说明其对分箱设置较为鲁棒,但树种识别准确率较低。K-近邻算法与随机森林模型在多数参数组合下表现相对稳定,准确率主要集中在0.650~0.750区间,对分箱容差变化不敏感。综上,合理的分箱策略(尤其是相对峰强度阈值与分箱容差的协同选择)对提升模型识别性能具

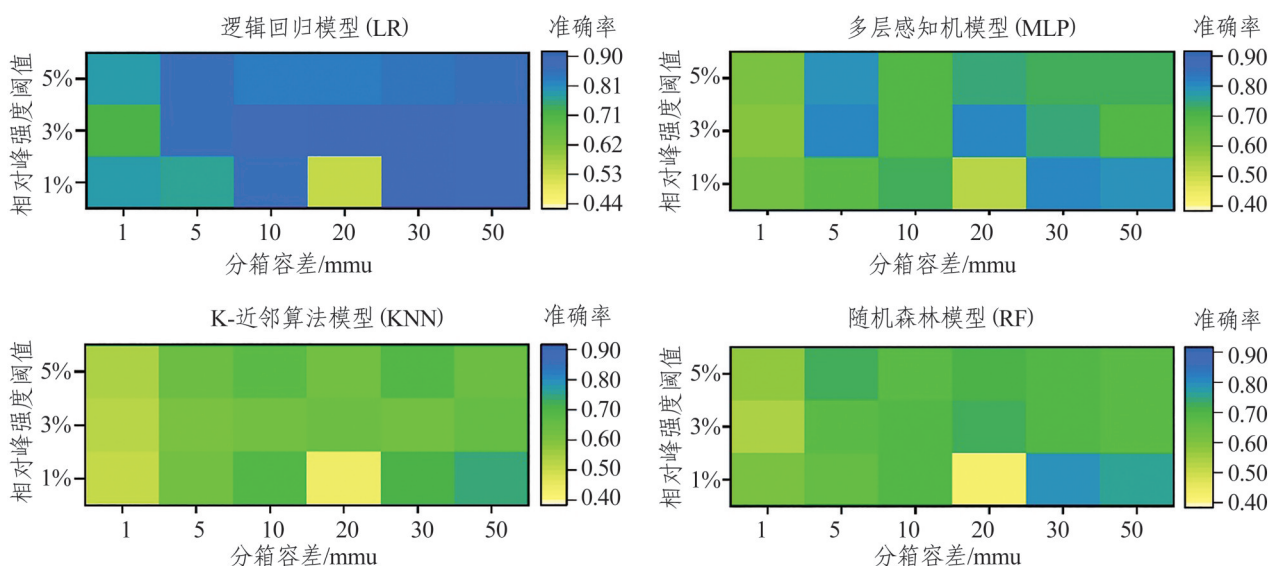


图4 四种机器学习模型在不同分箱参数条件下分类准确率

Fig.4 Classification accuracy of four machine learning models under different binning parameters

有关键作用，4种模型中逻辑回归在参数3%_20 mmu 预处理条件下获得最佳识别准确率0.895。同时对比4种模型的表现可知，不同机器学习方法在处理复杂质谱特征时存在自身结构的内在差异与适用边界。

尽管3种栎属木材在化学成分上具有一定相似性，但其心材中化学组分的相对含量及DART-TOF-MS所捕获的化学特征仍存在差异，从而为木材分类识别提供了化学依据。DEKLERCK等^[11]探讨了通过优化质谱预处理参数（分箱容差、相对峰强度阈值及变量数量）提高物种分类准确率的可能性，在楝科10个树种中采用无监督随机森林算法，分类准确率为0.780~0.822。本研究结果进一步支持了已有研究关于DART-MS可用于木材识别的结论^[8, 12]。

2.3.2 4种机器学习模型的性能 采用4种建模原理的机器学习算法，对DART-TOF-MS质谱数据构建分类模型，为木材树种的快速、准确识别提供方法学依据。基于质谱预处理参数对树种识别准确率的对比分析，表4展示各模型在最佳预处理参数（3%_20 mmu）设置下的性能差异。4种机器学习模型中，逻辑回归模型实现最优分类性能，其对3种栎属木材识别准确率达0.895。该模型不仅树种识别准确率最高，而且训练耗时仅31.59 s，远低于多层感知机与随机森林模型。

表4 四种机器学习模型的性能比较

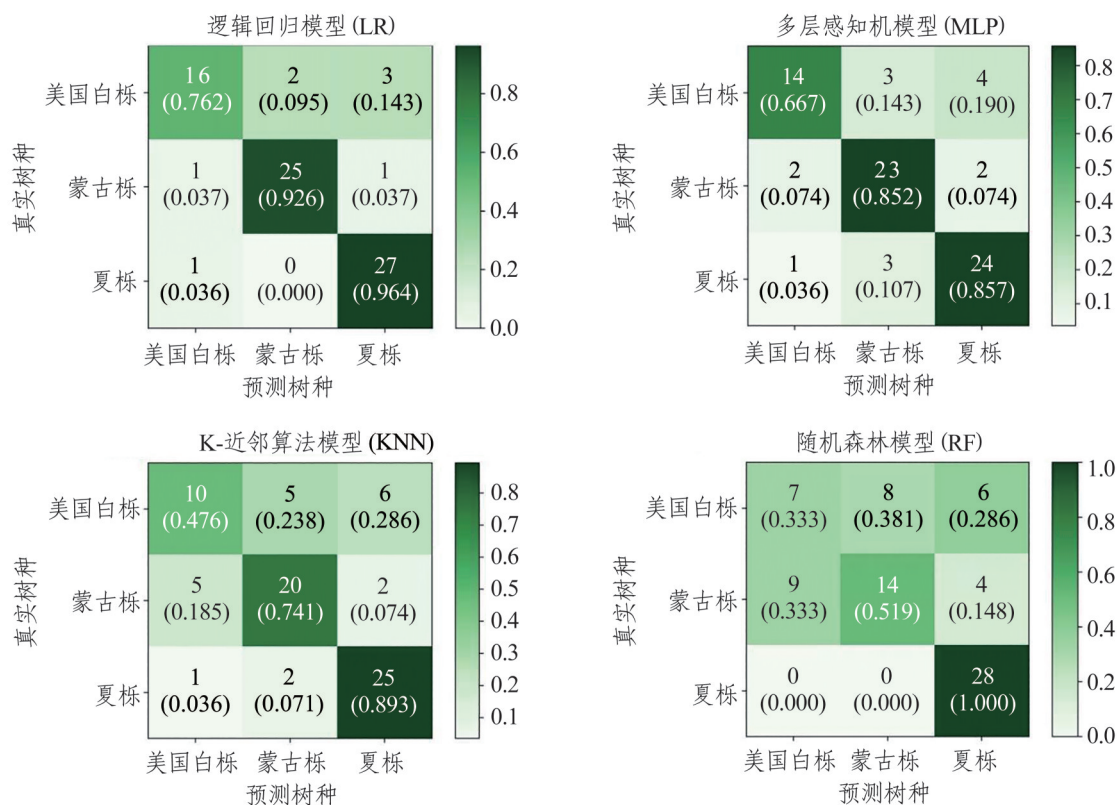
Tab.4 Performance comparison of four machine learning models

模型	准确率	训练时间/s	Macro-F1 值	Weighted-F1 值
逻辑回归	0.895	31.59	0.803	0.817
多层感知机	0.803	180.61	0.795	0.800
K-近邻算法	0.724	108.72	0.700	0.714
随机森林	0.645	15.60	0.599	0.620

注：预处理条件（相对峰强度阈值_分箱容差）为3%_20 mmu。

图5展示各模型的混淆矩阵结果。由图5可知，4种模型对夏栎的识别准确率均超过0.850，其中随机森林和逻辑回归两种模型的识别准确率分别达到1.000与0.964。此外，值得注意的是4种模型对美国白栎的识别表现较差，逻辑回归模型虽表现最佳，识别准确率也仅为0.762。

逻辑回归的线性判别能力与DART-TOF-MS质谱数据的特征结构高度匹配，同时其凸优化过程可快速收敛至全局最优解，显著优于多层感知机等非凸优化模型。此外，逻辑回归模型引入的正则化项有效抑制了高维冗余特征的干扰，提升了模型的泛化能力与稳定性。因此，在特征区分度较高的质谱数据场景中，逻辑回归等线性模型不仅具备较强的预测能力，也能兼顾计算效率与解释性，为木材树种的快速识别提供高效解决方案。研究结果进一步验证质谱预处理参数设置对树种识别准确率的影响，所构建的逻辑回归模型对3种



注：单元格内主数值为识别样本数，括弧内为该样本数占对应真实类别总样本数的比例。

图5 3种栎属木材的分类混淆矩阵

Fig.5 Confusion matrix of the classification for three *Quercus* species

高度相似的栎属木材识别准确率达0.895，表明该方法在木材树种识别中具有良好的可行性。同时，已有研究表明该方法对新鲜及干燥心材样本均适用，不易受取样位置或木材样本保存时间的影响，为木材标本的科学利用提供新途径^[12, 28-29]。由混淆矩阵结果可知，蒙古栎与夏栎的识别准确率分别为0.926、0.964，而美国白栎的识别准确率相对较低。主要原因可能是在正离子模式下，美国白栎与蒙古栎、夏栎的化学指纹特征重叠度较高，在特征空间中的边界更为接近；其次，美国白栎的样本量偏少，可能导致其种内变异覆盖不足，削弱了模型对美国白栎的稳定识别能力。

3 结论

1) 研究基于逻辑回归、多层感知机、随机森林和K-近邻算法等4种机器学习方法，对构造特征高度相似的3种栎属木材美国白栎、蒙古栎和夏栎的DART-TOF-MS质谱指纹图谱进行分析，评估其在树种识别中的可靠性和可行性。其中逻辑回归

算法构建模型的树种识别准确率最高（达0.895），优于其他3种模型。

2) 不同机器学习方法在处理复杂质谱特征时表现出较明显差异，合理的分箱策略（尤其是相对峰强度阈值控制与分箱容差选择）对模型性能具有关键影响，当参数设置为3%_20 mmu时模型整体表现最优。

3) 3种栎属木材在黄酮类化合物、木质素二聚体及鞣花酸的空间分布方面存在差异，这为基于质谱指纹信息的树种识别提供潜在的化学标志物依据。

本研究仍存在一定局限性，当前试验树种数量有限，后续应拓展更多树种验证方法的泛化能力。相较于木材解剖学分析、DNA条形码等分子标记技术，基于化学质谱的木材树种识别方法具有操作简便、检测分析周期短等优势。未来通过扩容完善质谱数据库，该技术在木材“种”水平的快速精准鉴定实践中，将展现突出的应用潜力。

参考文献:

- [1] LIU S J, HE T, WANG J J, et al. Can quantitative wood anatomy data coupled with machine learning analysis discriminate CITES species from their look-alikes[J]. Wood Science and Technology, 2022, 56(5): 1567-1583.
- [2] JIAO L C, YU M, WIEDENHOEF A C, et al. DNA barcode authentication and library development for the wood of six commercial *Pterocarpus* Species: the critical role of xylarium specimens[J/OL]. Scientific Reports, 2018, 8: 1945. <https://doi.org/10.1038/s41598-018-20381-6>.
- [3] JIAO L C, LU Y, HE T, et al. DNA barcoding for wood identification: global review of the last decade and future perspective[J]. IAWA Journal, 2020, 41(4): 620-643.
- [4] CODY R B, LARAMEE J A, DURST H D. Direct analysis in real time (DART) mass spectrometry of solid materials, surfaces, and liquids [J]. Journal of the American Society for Mass Spectrometry, 2005, 16(4): 681-685.
- [5] 张毛毛, 蒋劲东, 刘波, 等. 实时直接分析质谱技术在木材识别研究中的应用 [J]. 木材工业, 2019, 33(1): 29-33.
- ZHANG M M, JIANG J D, LIU B, et al. Application of direct analysis in real time mass spectrometry (DART-MS) for wood identification [J]. Wood Industry, 2019, 33(1): 29-33.
- [6] ZHANG M M, ZHAO G J, LIU B, et al. Wood discrimination analyses of *Pterocarpus tinctorius* and endangered *Pterocarpus santalinus* using DART-FTICR-MS coupled with multivariate statistics [J]. IAWA Journal, 2019, 40(1): 58-74.
- [7] RAVINDRAN P, WIEDENHOEF A C. Comparison of two forensic wood identification technologies for ten Meliaceae woods: computer vision versus mass spectrometry[J]. Wood Science and Technology, 2020, 54(5): 1139-1150.
- [8] ESPINOZA E O, LANCASTER C A, KREITALS N M, et al. Distinguishing wild from cultivated agarwood (*Aquilaria* spp.) using direct analysis in real time and time of-flight mass spectrometry[J]. Rapid Communications in Mass Spectrometry, 2014, 28(3): 281-289.
- [9] PARK G, LEE Y G, YOON Y S, et al. Machine learning-based species classification methods using DART-TOF-MS data for five coniferous wood species[J/OL]. Forests, 2022, 13(10): 1688. <https://doi.org/10.3390/f13101688>.
- [10] DEKLERCK V, FINCH K, GASSON P, et al. Comparison of species classification models of mass spectrometry data: Kernel discriminant analysis vs random forest; a case study of *Afrormosia* (*Pericopsis elata* (Harms) Meeuwen)[J]. Rapid Communications in Mass Spectrometry, 2017, 31(19): 1582-1588.
- [11] DEKLERCK V, MORTIER T, GOEDERS N, et al. A protocol for automated timber species identification using metabolome profiling [J]. Wood Science and Technology, 2019, 53(4): 953-965.
- [12] ZHANG M M, ZHAO G J, GUO J, et al. Timber species identification from chemical fingerprints using direct analysis in real time (DART) coupled to Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS): comparison of wood samples subjected to different treatments[J]. Holzforschung, 2019, 73(11): 975-985.
- [13] 路佳佳. 基于交叉验证的集成学习误差分析[J]. 计算机系统应用, 2023, 32(1): 302-309.
- LU J J. Error analysis of ensemble learning based on cross validation [J]. Computer Systems & Applications, 2023, 32(1): 302-309.
- [14] LIU S J, ZHENG C, WANG J J, et al. How to discriminate wood of CITES-listed tree species from their look-alikes: using an attention mechanism with the ResNet model on an enhanced macroscopic image dataset[J/OL]. Frontiers in Plant Science, 2024, 15: 1368885. <https://doi.org/10.3389/fpls.2024.1368885>.
- [15] IAWA Committee. IAWA list of microscopic features for hardwood identification[J]. IAWA Bulletin, 1989, 10(3): 219-332.
- [16] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 56-62.
- [17] 邱锡鹏. 神经网络与深度学习[M]. 北京: 机械工业出版社, 2020: 89-105.
- [18] 冯晓荣, 瞿国庆. 基于深度学习与随机森林的高维数据特征选择[J]. 计算机工程与设计, 2019, 40(9): 2494-2501.
- FENG X R, ZHAI G Q. Feature selection for high-dimensional data based on deep learning and random forest[J]. Computer Engineering and Design, 2019, 40(9): 2494-2501.
- [19] 李驰, 段雨梅. K近邻算法优化设计策略[J]. 电脑知识与技术, 2019, 15(31): 200-202, 211.
- LI C, DUAN Y M. Optimal design strategy of K-nearest neighbor algorithm[J]. Computer Knowledge and Technology, 2019, 15(31): 200-202, 211.
- [20] TRAORÉ M, MARTÍNEZ CORTIZAS A. Color and chemical composition of timber woods (*Daniellia oliveri*, *Isobberlinia doka*, *Khaya senegalensis*, and *Pterocarpus erinaceus*) from different locations in southern Mali[J/OL]. Forests, 2023, 14(4): 767. <https://doi.org/10.3390/f14040767>.
- [21] SJOSTROM E. Wood chemistry: fundamentals and applications[M]. Gulf Professional Publishing, 1993.
- [22] OSAKABE K, TSAO C C, LI L G, et al. Coniferyl aldehyde 5-hydroxylation and methylation direct syringyl lignin biosynthesis in angiosperms[J]. Proceedings of the National Academy of Sciences of the United States of America, 1999, 96(16): 8955-8960.
- [23] PUECH J L. Extraction of phenolic compounds from oak wood in model solutions and evolution of aromatic aldehydes in wines aged in oak barrels[J]. American Journal of Enology and Viticulture, 1987, 38 (3): 236-238.
- [24] WEI L M, MA R K, FU Y L. Differences in chemical constituents between *Dalbergia oliveri* heartwood and sapwood and their effect on wood color[J/OL]. Molecules, 2022, 27(22): 7978. <https://doi.org/10.3390/molecules27227978>.
- [25] CODY R B, ESPINOZA E O, PRICE E R, et al. Wood from hardwood angiosperms and coniferous gymnosperms shows distinctive lignin peaks in direct analysis in real time (DART) mass spectra[J]. Journal of the American Society for Mass Spectrometry, 2023, 34(4): 784-789.
- [26] CHATONNET P, BOIDRON J N. Identification and determination of volatile compounds released by oak wood used for aging of wines[J]. American Journal of Enology and Viticulture, 1989, 40(1): 76-83.
- [27] QUIDEAU S, DEFFIEUX D, DOUAT-CASASSUS C, et al. Plant polyphenols: chemical properties, biological activities, and synthesis [J]. Angewandte Chemie International Edition, 2011, 50(3): 586-621.
- [28] DEKLERCK V, LANCASTER C A, VAN ACKER J, et al. Chemical fingerprinting of wood sampled along a pith-to-bark gradient for individual comparison and provenance identification[J]. Forests, 2020, 11(1): 107.
- [29] ZHANG M M, GUO J, LU Y, et al. Similarity network fusion for aggregating headspace GC-MS and direct analysis in real time-mass spectrometry data from solid samples to enhance species identification efficiency of high-temperature heated wood[J/OL]. Journal of Wood Science, 2022, 68: 1-13. <https://doi.org/10.1186/s10086-022-02044-3>.

(本文编校 向琴)