

基于 MALDI-TOF 的脂质 - 多肽指纹图谱 结合机器学习技术增强有机牛奶的真实性 认证

MALDI-064

摘要：本文介绍了一种利用基质辅助激光解吸电离飞行时间质谱（MALDI-TOF）对有机牛奶进行认证并确定其产地的方法。对牛奶中的脂质和多肽成分进行质谱采集，通过多元统计分析确定了区分有机和传统牛奶以及不同产地牛奶的关键脂质和多肽，8 种脂质和 6 种多肽在有机牛奶中含量较高，4 种脂质和 10 种多肽在不同产地的有机牛奶中含量差异显著。将脂质 - 多肽联合数据集与机器学习模型相结合，显著提高了分类准确率，优于基于单一数据集的模型。该方法展示了将 MALDI-TOF 与机器学习相结合用于牛奶真实性认证以防止乳制品欺诈的潜在应用价值。

关键词： MALDI-TOF 牛奶认证 养殖系统 产地 机器学习

技术特点：

- ❖ 提出了一种新颖的、简便、快速的 MALDI-TOF 质谱检测有机牛奶的方法。通过对脂质和多肽进行整合分析，并结合机器学习技术，实现了有机牛奶的认证以及其产地的识别。
- ❖ 采用了由 α - 氰基 -4- 羟基肉桂酸和 2, 5- 二羟基苯甲酸组成的二元基质，并结合真空干燥策略，提高了方法的重现性。

有机牛奶是一种纯天然、无污染的有机食品，它与普通牛奶的区别主要在于饲养方式、饲料和牛奶的处理过程。有机牛奶的生产过程严格遵守有机标准，高生产成本和低产量导致价格大幅上涨，增加了经济欺诈的风险，迫切需要强有力的认证方法来维护市场的完整性。

多种分析方法已被应用于牛奶的鉴定和溯源工作，包括近红外光谱（NIR）、气相色谱 - 质谱（GC-MS）、同位素比值质谱法（IRMS）、液相色谱串联质谱（LC-MS/MS）。然而，较复杂的样品制备和较高的时间成本限制了它们的适用性。

近年来，基质辅助激光解吸电离飞行时间质谱（MALDI-TOF）凭借其高通量、低成本、速度快的优势在食品分析领域初露头角，然而其在乳制品鉴定中的应用仍处于起步阶段，大多数研究仅限于分析单一类别的化合物，主要是蛋白质。然而，蛋白质、多肽和脂质都是牛奶中关键的营养成分，对多种化合物的综合分析仍较缺乏。本研究基于 MALDI-TOF 将脂质 - 多肽指纹图谱与机器学习技术相结合，引入了一种新的方法，使有机牛奶认证和产地识别成为可能。

■ 实验部分

1.1 仪器

岛津基质辅助激光解吸电离飞行时间质谱（MALDI-8030）

1.2 仪器参数

1.2.1 脂质分析参数

分析模式：	正离子模式	激光照射次数 (shots)：	200
激光器：	355 nm 固态激光器	脉冲引出质量 (Da)：	700
扫描频率 (Hz)：	100	扫描范围 (m/z)：	400-1000
激光强度：	60		

1.2.2 多肽分析参数

分析模式：	正离子模式	激光照射次数 (shots)：	1000
激光器：	355 nm 固态激光器	脉冲引出质量 (Da)：	3982
扫描频率 (Hz)：	200	扫描范围 (m/z)：	800-5000
激光强度：	70		

1.3 试剂与样品

基质：CHCA (α -氰基-4-羟基肉桂酸)；2, 5-二羟基苯甲酸 (DHB)；Super-DHB (S-DHB)。

样品：8 个品牌 (来自 8 个产地) 的市售盒装牛奶，每个品牌有机牛奶和传统牛奶各 3 份，共计 48 份样品。每份样品收集 6 个平行实验数据。

1.4 样品前处理

1.4.1 脂质分析方法：

样品制备：取 20 μ L 牛奶与 77 μ L 甲醇混合，旋涡 5 min。然后加入 258 μ L 甲基叔丁基醚 (MTBE)，再次旋涡。加入 65 μ L 超纯水进行双相分离，10000 rpm 离心 5 min，收集上相，蒸发干燥，用 100 μ L 甲醇重悬。

基质配制：称取 S-DHB 基质 10 mg 溶于 1 mL 70% 乙醇中。

点样方法：取等体积的样品制备溶液与基质溶液混合，取 1 μ L 滴在靶板上，室温晾干。

1.4.2 多肽分析方法：

样品制备：将 20 μ L 牛奶用 980 μ L 超纯水稀释至 1:50 (v/v)。

基质配制：首先将 20 mg DHB 溶解于 700 μ L 乙腈和 300 μ L 0.1% 三氟乙酸溶液中，并进行涡旋。将 20 mg CHCA 溶解于 700 μ L 乙腈和 300 μ L 5% 甲酸溶液中，旋涡后与 DHB 溶液等份混合。

点样方法：取等体积的样品制备溶液与基质溶液混合，取 1 μ L 滴在靶板上，真空干燥。

■ 结果与讨论

2.1 有机牛奶和传统牛奶的认证

将来自 8 个品牌的有机牛奶 (OM) 和传统牛奶 (CM) 共 48 份样品的脂质数据进行多元统计分析。对两组数据构建正交偏最小二乘判别分析 (OPLS-DA) 模型，如图 1a 所示，有机牛奶和传统牛奶样品得到明显的区分，模型拟合和预测能力较强 ($R^2Y=0.861$, $Q^2=0.804$, 越接近 1 表明性能越好)。为了验证模型的稳健性和准确性，进行了 200 次置换检验，结果显示 Q^2 截距相交于零以下，证实模型没有过拟合 (图 1b)。通过计算投影变量重要性 (VIP) 分数以确定关键的判别特征。其中，VIP>2 且 $p<0.05$ 的脂质有 8 个 (m/z 659.5、647.4、603.4、578.4、706.5、760.5、633.4、411.2)，它们被认为与区分有机牛奶和传统牛奶高度相关。热图 (图 1c) 和箱形图 (图 1d) 显示了这些差异脂质的相对丰度。

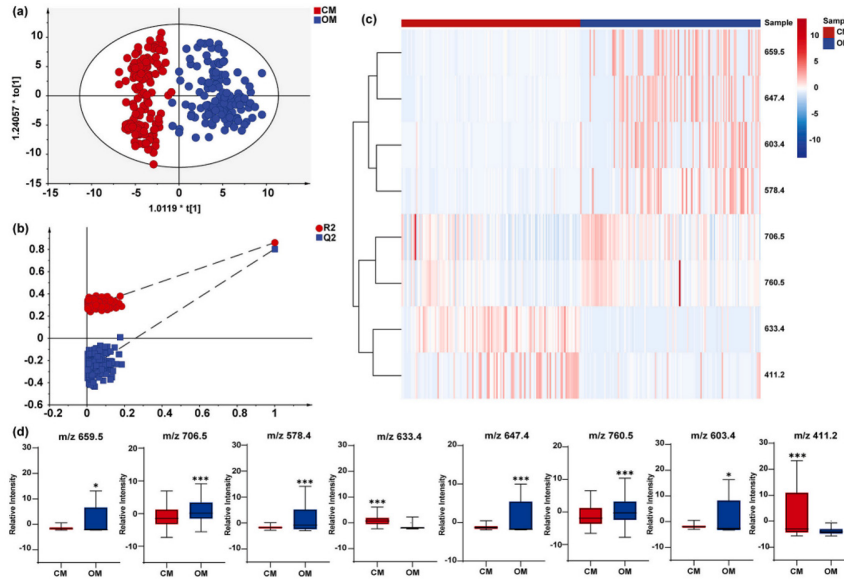


图 1 基于脂质的多元统计分析结果

(a) OPLS-DA 分析有机牛奶 (OM) 和传统牛奶 (CM) ; (b) OPLS-DA 置换检验结果; (c) 8 种脂质的热图; (d) 8 种脂质的箱形图。*, ** 和 *** 分别表示 OM 与 CM 之间的显著性水平 ($p < 0.05$, $p < 0.01$, $p < 0.001$)。

图片引自: Minrui Zhou, et al. Food Chemistry, Volume 475, 2025, 143346

基于多肽数据的 OPLS-DA 分析结果 (图 2a) 显示有机牛奶和传统牛奶之间存在明显差异, 模型拟合能力和预测能力强 ($R^2Y=0.854$, $Q^2=0.823$)。经过 200 次置换检验, 模型没有出现拟合 (图 2b)。热图 (图 2c) 和箱形图 (图 2d) 显示, 根据 $VIP > 2$ 和 $p < 0.05$ 选择的 6 种差异多肽 (m/z 1994.1、2107.2、2460.3、2618.0、3477.0 和 4142.9) 的相对丰度存在明显差异。

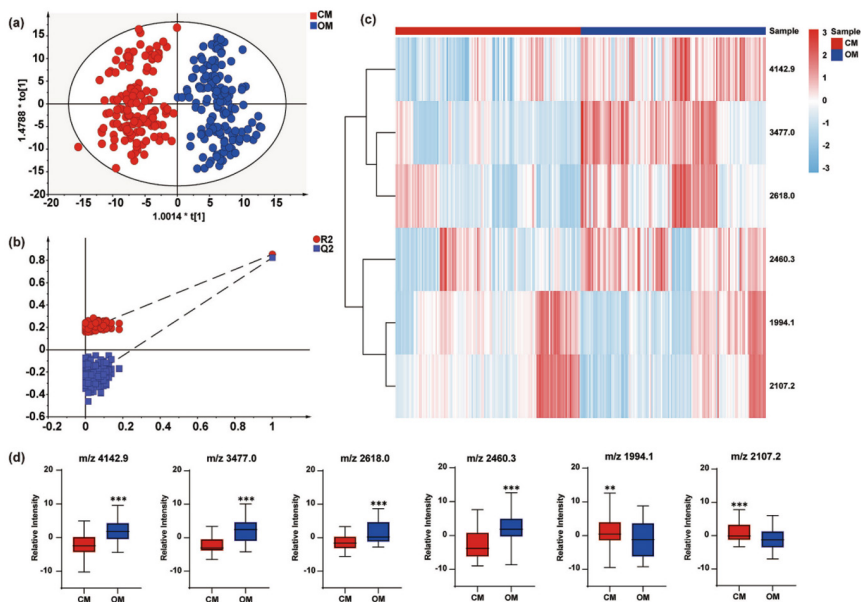


图 2 基于多肽的多元统计分析结果

(a) OPLS-DA 分析有机牛奶 (OM) 和传统牛奶 (CM) ; (b) OPLS-DA 置换检验结果; (c) 6 种多肽的热图; (d) 6 种多肽的箱形图。*, ** 和 *** 分别表示 OM 与 CM 之间的显著性水平 ($p < 0.05$, $p < 0.01$, $p < 0.001$)。

图片引自: Minrui Zhou, et al. Food Chemistry, Volume 475, 2025, 143346

2.2 不同产地的有机牛奶的认证

有机食品真实性验证不仅包括确认食品的有机状态，还包括验证其标注的产地。对八个不同产地的有机牛奶的脂质数据构建偏最小二乘判别分析 (PLS-DA) 模型，结果 (图 3a) 显示出良好的区域分类效果，模型拟合和预测能力较强 ($R^2Y = 0.881$, $Q^2=0.713$)。通过 200 次置换检验验证了模型的稳健性和准确性，证实不存在过拟合 (图 3b)。热图 (图 3c) 和箱形图 (图 3d) 显示了 4 种差异脂质 (m/z 758.5、760.5、663.4 和 664.4) 的相对丰度。其中，呼和浩特 (Hohhot)、巴彦淖尔 (Bayanaoer)、呼伦贝尔 (HulunBuir)、北京 (Beijing)、临沂 (Linyi) 和乌鲁木齐 (Urumqi) 都属于温带气候，使得这些地区的奶牛养殖方式具有相似性，因此牛奶样品品质谱特征具有重叠性。值得注意的是，尽管大理 (Dali) 和深圳 (Shenzhen) 都属于亚热带季风气候，大理牛奶样品却表现出独特的质谱特征，与其他 7 个地区有较大差异。这可能归因于大理独特的高海拔环境 (平均海拔 2243 米)，海拔在所有地区中最高，显著影响了其乳脂成分。

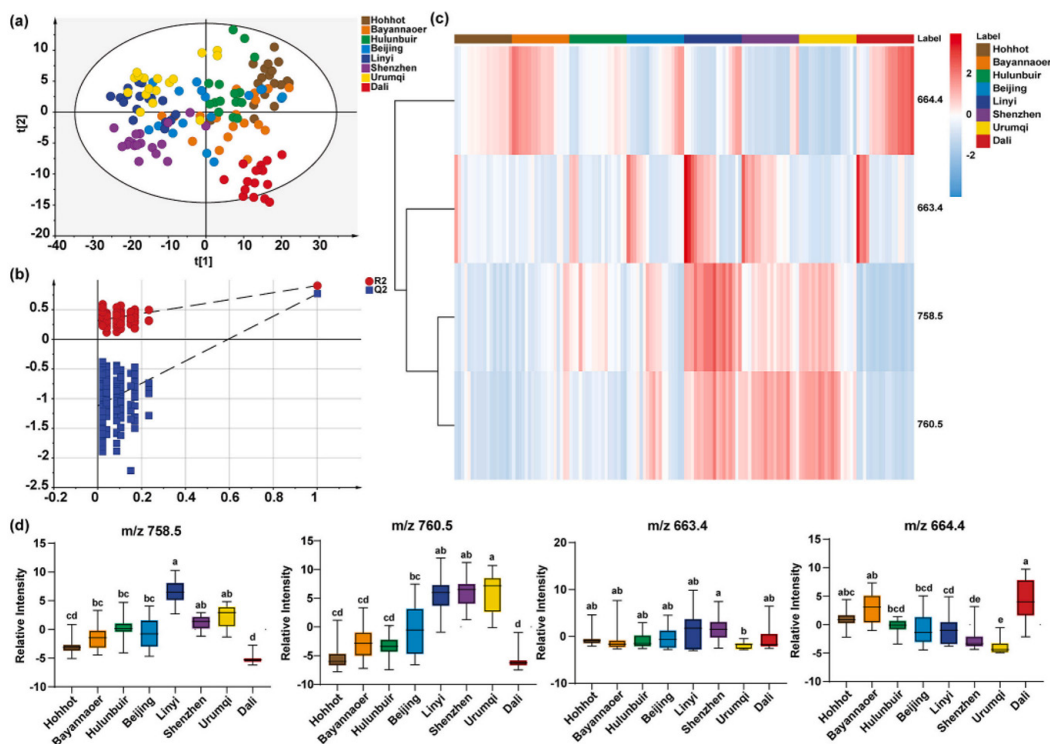


图3 基于脂质的多元统计分析结果 (不同产地)

(a) 八个不同产地的有机牛奶的 PLS-DA 结果; (b) PLS-DA 模型 200 次置换检验结果; (c) 4 种脂质的热图; (d) 4 种脂质的箱形图。采用 two-tailed Dunn 和 Kruskal-Wallis 检验进行统计学分析, $p < 0.05$ 认为有统计学意义。不同的小写字母表示不同地区的统计显著差异。图片引自: Minrui Zhou, et al. Food Chemistry, Volume 475, 2025, 143346

分析了来自八个不同产地的有机牛奶中的多肽, PLS-DA 模型 (图 4a) 实现了不同产地有机牛奶样品的完全分离 ($R^2Y = 0.919$, $Q^2 = 0.882$), 表明不同产地的有机牛奶中多肽水平差异显著。模型置换检验证实不存在过拟合 (图 4b)。热图 (图 4c) 和箱形图 (图 4d) 显示了 10 种差异肽 (m/z 1284.7、1991.1、2107.2、2331.1、2460.3、2618.0、2763.5、3477.0、3982.2 和 2528.6) 的相对丰度。这些多肽主要在大理地区有机牛奶样品中含量较高, 可能是由于高海拔环境对牛奶成分有较大影响。高海拔条件可能通过诱导低氧水平奶牛的代谢适应, 从而促进蛋白质合成并增加牛奶蛋白质含量, 从而显著影响牛奶成分。富含微量元素 (如硒和锌) 的高海拔土壤也可能促进蛋白质和肽的合成。

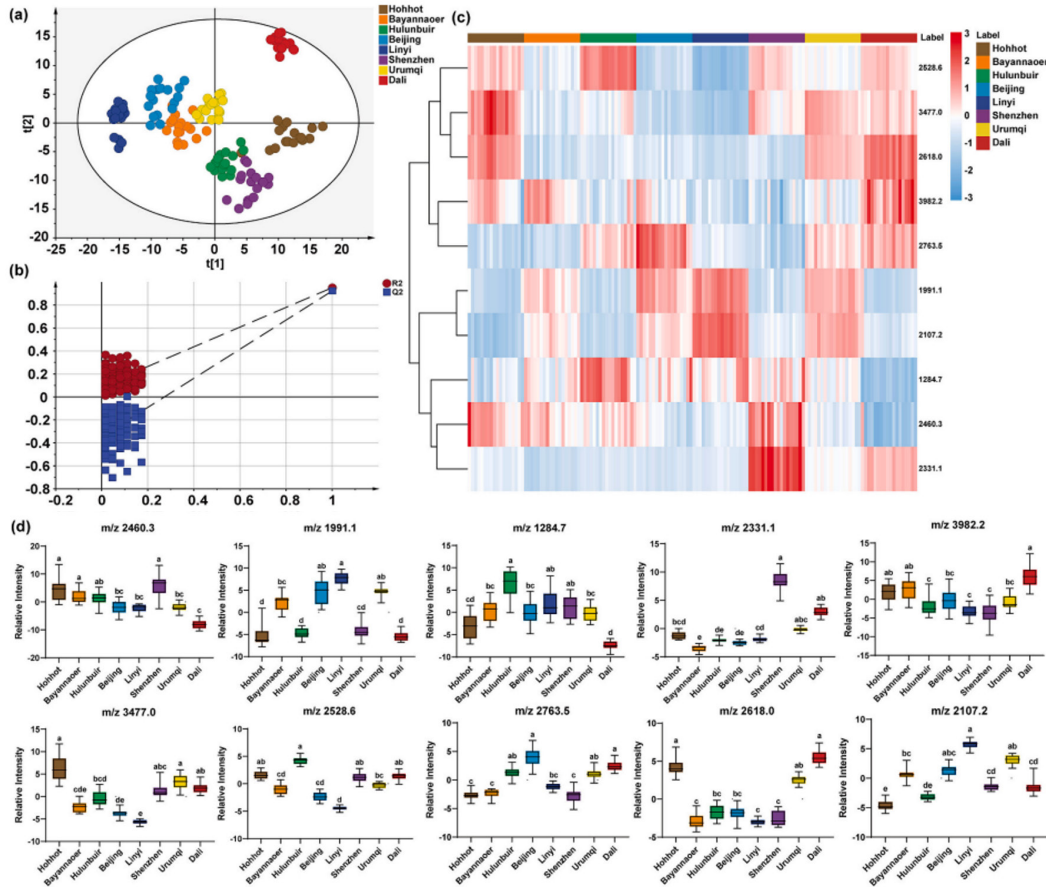


图 4 基于多肽的多元统计分析结果 (不同产地)

(a) 八个不同产地的有机牛奶的 PLS-DA 结果; (b) PLS-DA 模型 200 次置换检验结果; (c) 10 种多肽的热图; (d) 10 种多肽的箱形图。采用 two-tailed Dunn 和 Kruskal-Wallis 检验进行统计学分析, $p < 0.05$ 认为有统计学意义。不同的小写字母表示不同地区的统计显著差异。图片引自: Minrui Zhou, et al. Food Chemistry, Volume 475, 2025, 143346

2.3 机器学习分析

与单源数据分析相比, 整合多源信息能提供更全面和准确的评估。本研究首次将牛奶样品中脂质和多肽数据集结合起来构建机器学习模型, 以加强对乳制品的认证能力。使用单个数据集与联合数据集构建的模型的分​​类准确度如图 5a 所示。在使用联合数据集构建的 7 个模型中, ANN、SVM 和 KNN 模型在有机牛奶和传统牛奶的鉴别中达到 100% 的准确率, 其余模型也表现出较高的分类准确率 (91.6±3.1%)。相比之下, 仅使用单个脂质或多肽数据集构建的模型性能较低, 特别是 NB 模型, 其准确性为 52.3% (脂质数据集) 和 68.6% (多肽数据集)。联合数据集中 NB 模型的准确率显著提高到 87.2%。AUC 值的分布进一步证明了联合数据集的优势。如图 5b, 对于单个多肽数据集, 没有模型的 AUC 值能够达到 1, 模型的 AUC 值在 0.700-0.899 和 0.900-0.999 之间。同样, 对于单个脂质数据集, 没有一个模型的 AUC 值能够达到 1。然而, 当使用联合数据集, 三个模型的 AUC 值为 1, 而其余模型的 AUC 值在 0.900-0.999 之间。这表明联合数据集显著提高了模型性能。

此外, 联合数据集还提高了模型区分不同产地有机牛奶的准确性 (图 5c)。在 AUC 值方面, 使用联合数据集的所有模型的 AUC 值都在 0.900 以上, 其中四个模型的 AUC 达到 1 (图 5d)。

这些结果证实, 将脂质和多肽数据集与机器学习模型相结合大大提高了分类准确度, 为区分有机牛奶和传统牛奶以及识别其产地提供了强大的工具。

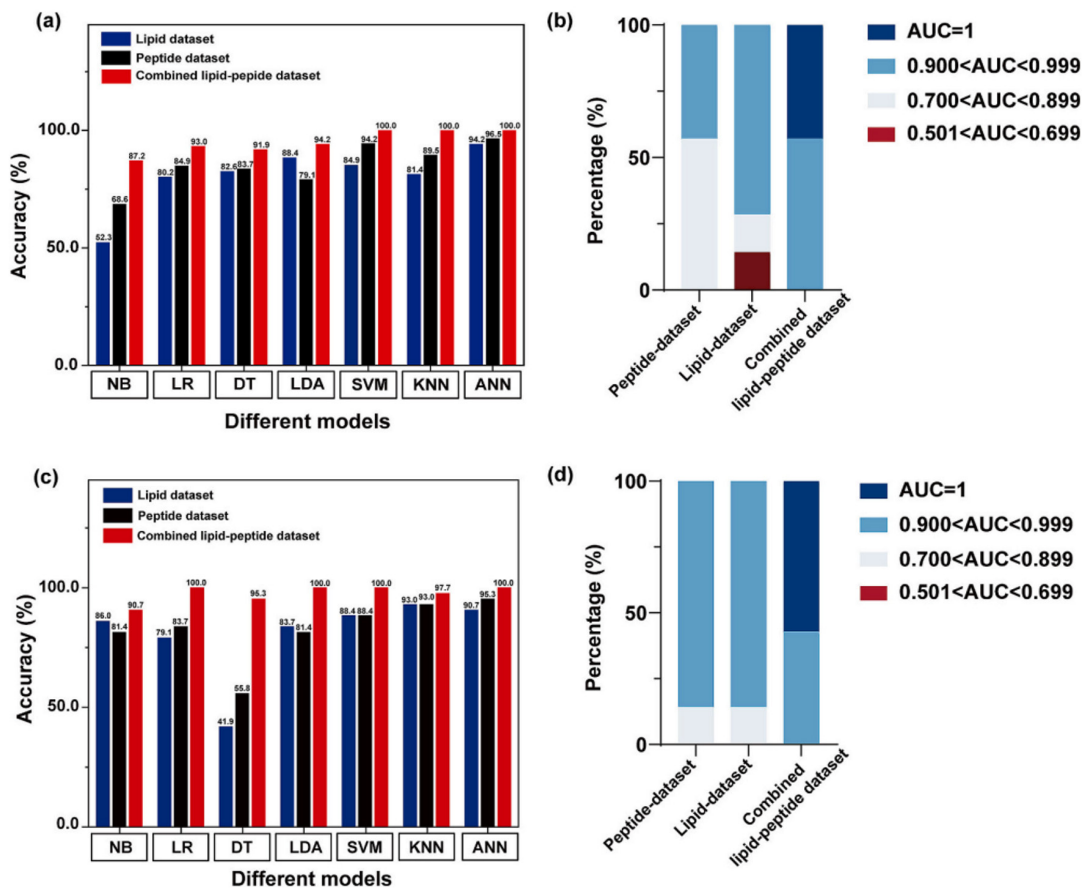


图 5 单个数据集和联合数据集的机器学习模型性能比较

有机牛奶和传统牛奶分类：(a) 每个模型在测试集上的准确性；(b) AUC 值在不同数据集上的分布。
不同产地的有机牛奶分类：(c) 每个模型在测试集上的准确性；(d) AUC 值在不同数据集上的分布。

图片引自：Minrui Zhou, et al. Food Chemistry, Volume 475, 2025, 143346

结论

本研究开发了一种新颖的方法，通过 MALDI-TOF 对牛奶中脂质和多肽进行分析，并结合机器学习技术，实现了有机牛奶的认证及其产地的识别。提出的多肽和脂质联合数据集与机器学习技术相结合，显著提高了机器学习模型在牛奶分类和产地识别方面的性能，克服了单一数据集模型准确率低的局限性。该方法为提高有机牛奶认证和追溯的准确性和可靠性提供了一个全面的框架，有望更广泛地应用于其他有机产品，以确保市场中有机产品的真实性以维护消费者的利益。

❖ 本研究由华南理工大学和岛津合作完成，研究成果已发表在《Food Chemistry》：

Minrui Zhou, Ting Li, et al. Enhanced authentication of organic milk using MALDI-TOF MS with combined lipid-peptide fingerprinting and machine learning integration[J]. Food Chemistry, Volume 475, 2025, 143346, <https://doi.org/10.1016/j.foodchem.2025.143346>

岛津应用云

