第42卷，第9期
2022年9月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 42, No. 9, pp2740-2745
September, 2022

基于密度泛函理论与自举软缩减法的酒石酸太赫兹光谱特征谱区分析指认

唐 鑫，周胜灵*，祝诗平*，马羚凯，郑 权，普 京

西南大学工程技术学院，重庆 402160

摘 要 太赫兹时域光谱不但包含了样品的化学信息和物理信息，还承载了设备噪声、样品状态、环境参数等多方面的背景信息，其光谱的多元性可能影响模型的性能，降低预测精度。能否在复杂、重叠、变动背景下从光谱数据中提取目标组分的特征信息，去除冗余变量，筛选特征谱区，对太赫兹光谱定量、定性分析至关重要。以L-酒石酸为研究对象，在室温下采集6个浓度：10%，20%，40%，50%，60%和80%，共计342个样本的L-酒石酸太赫兹吸收光谱。利用密度泛函理论(DFT)中的B3LYP方法，基于6-31G*(d,p)基组对L-酒石酸单分子模型进行优化并对其太赫兹频谱特性进行理论模拟计算，分析对应特征波峰的分子振动模式，得到0.2～1.6 THz频段吸收谱。与实测吸收谱进行对比，实验所测结果与理论计算结果对应的吸收峰位置基本吻合。采用自举软缩减法(BOSS)对L-酒石酸的太赫兹吸收谱进行特征谱区筛选，并与竞争性自适应加权采样(CARS)、蒙特卡洛无信息变量消除法(MC-UVE)和间隔区间偏最小二乘法(iPLS)3种经典特征谱区筛选法进行对比，分析结果显示BOSS算法选取的有效谱区与DFT理论计算特征谱区重合度最优。分别使用全谱PLS，CARS-PLS，MC-UVE-PLS，iPLS及BOSS五种算法对L-酒石酸光谱进行建模回归分析，实验结果表明，四种谱区筛选方法相较于全谱PLS模型，预测精度均有所提高，其中BOSS算法预测能力提高最为显著，其交互验证均方根误差(RMSECV)、预测均方根误差(RMSEP)、训练集决定系数(R^2_{train})和测试集决定系数(R^2_{test})分别为0.026 0，0.026 0，0.988 1和0.987 5，相较其他模型有更高的预测精度和模型稳定性，为实现基于太赫兹光谱技术的快速定量检测提供了一种有效的方法。

关键词 太赫兹光谱；L-酒石酸；密度泛函；谱区筛选；自举软缩减法

中图分类号：O433.4 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2022)09-2740-06

引 言

太赫兹波(Terahertz，THz)频率介于0.1～10 THz之间，作为一门衔接了微观量子理论和经典电磁波理论的新兴科学，展现了许多独特的性质，如安全性、透视性和较强的分辨能力。THz光谱对固态分子的排列、振动和转动非常敏感[1]，可以通过振幅和相位信息，对材料的成分及其结构的细微变化进行有效的分析。

目前，国内外许多学者针对THz光谱信息筛选特征谱区，简化模型展开了研究。Hu等[2]利用竞争性自适应加权采样(CARS-PLS)对苯甲酸在1.0～3.0 THz频段建立光谱模型，得到相关系数为0.9953，较传统方法，检测精度大幅度提高。Li等[3]利用蒙特卡洛无信息变量消除法(MC-UVE-PLS)提取近红外光谱信息，用于测定棉籽中棉酚的含量，结果表明，MC-UVE-PLS算法可以通过提取有效信息和剔除无关变量达到特征谱区筛选的目的。Jiang等[4]通过间隔区间偏最小二乘法(iPLS)实现对0.2～1.6 THz的山梨酸和山梨酸钾样品THz光谱的特征谱区筛选，得到了很好的预测精度。自举软缩减法(bootstrapping soft shrinkage，BOSS)是由Deng于2016年提出[5]，该算法基于随机抽样统计技术(bootstrap sampling，BSS)和加权引导采样技术(weight bootstrap sampling，WBS)，用于生成变量的随机组合并构建子模型，对不同的变量赋予不同的权重。基于生成的权重模型，采用模型集群分析(MPA)结合PLS提取信息并建模分析。对上述步骤进行反复的变量迭代提取，简化变量空

收稿日期：2021-07-26，**修订日期**：2021-10-25

基金项目：国家自然科学基金面上项目(31771670，62005227)，重庆市自然科学基金面上项目(cstc2020jcyj-msxmX0300)资助

作者简介：唐 鑫，1993年生，西南大学工程技术学院硕士研究生 e-mail：tangxinyp@email.swu.edu.cn

*通讯作者 e-mail：zspswu@126.com；swuzhousl@163.com

间，权重较大的变量有更大概率得以保留。通过迭代计算中最小的交叉验证均方根误差(RMSECV)，确定最优变量集。对变量中存在共线性信息导致的模型不准确问题有良好的处理效果。

上述方法，多是从提升预测结果出发，通过迭代优化参数进行特征谱区筛选，但对筛选的特征谱区是否合理缺乏理论支撑，因此本文将通过量子化学方法对特征峰在分子结构振动的归属予以说明，对筛选的特征谱区正确性加以验证。随着计算机性能的显著提高，利用量子化学方法对分子振动模式进行仿真，分析光谱指纹特征的产生机理，得到了广泛应用。密度泛函理论(density functional theory，DFT)通过粒子密度函数来描述原子和分子等基态的特征，常用于计算和预测分子的物理结构和化学性质，基于其显著的“鲁棒性”，以简便有效的运算为分子特征提供合理准确的预测[6-7]。Zhou 等[8]利用 DFT 对卡马西平与烟酰胺、糖精和富马酸的单个组分的晶体振动进行了理论研究并得到 THz 光谱。实验结果显示，仿真的 THz 光谱成功再现了实验光谱中所有的晶体特征，证实了 DFT 理论在光谱指纹特性仿真计算中具有很高的计算效率和精度。Zhang 等[9]通过 THz-TDS 对尿囊素晶体的 THz 光谱吸收和色散特性(弱相互作用分析)进行解码，结果表明了 DFT 理论可以仿真晶体的 THz 光谱指纹特征。

综上所述，以 L-酒石酸(L-TA)为研究对象，建立并分析了基于 DFT 理论的分子模型及 THz 特征频谱，从理论计算角度验证了实验吸收峰来源判断的正确性。在此基础上，分别采用 CARS-PLS，MC-UVE-PLS，iPLS 和 BOSS 共 4 种方法对 L-TA 在 0.2～1.6 THz 频段进行特征谱区的筛选，建立了定量分析模型。

1　实验部分

1.1　装置

采用 EKSPLA 公司的 T-SPEC 太赫兹时域光谱设备。装置采用低温砷化镓(LT-GaAs)作为光导天线，FF50 飞秒激光器作为超短脉冲激光光源，中心波长为 1 064 nm，输出脉宽为 150 fs，发射器和探测器之间的光程约为 62.5 cm。光路通过光学透镜控制和校准，系统原理如图 1 所示。飞秒脉冲通过半波片，由分光镜将其分为强脉冲和弱脉冲。强脉冲经过斩波器后，由反射镜引导经可变延迟线入射到 LT-GaAs 光导天线上，产生 THz 电磁辐射脉冲，然后将该 THz 脉冲聚焦在待测试的样品上。弱脉冲作为探测光，待测样品发射的太赫兹脉冲与微弱脉冲进行合并，合并后的信号送入锁相放大器放大处理，得到含有待测样品信息的 THz 时域频谱。本实验在室温下进行，整个光谱处于充入氮气的封闭箱，相对湿度控制在 5%以下，以减少空气中水蒸气对 THz 波的吸收影响。

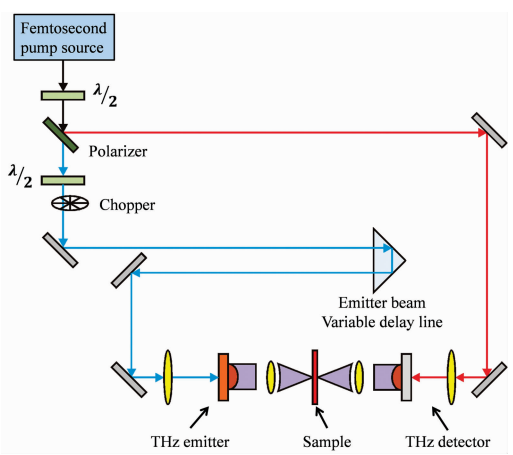

图 1　T-SPEC 太赫兹时域光谱设备原理图

Fig. 1　Schematic diagram of the T-SPEC terahertz time-domain spectroscopy equipment

1.2　样品制备与数据获取

实验样品 L-TA 和聚乙烯固体粉末均购于 Signa-aldrich 公司，样品纯度大于 98%，使用时未进一步提纯处理。将 L-TA 和聚乙烯混合，得到 L-TA 浓度分别为 10%，20%，40%，50%，60%和 80%的混合粉末，将其放入研钵中研磨以减少颗粒引起的散射效应，保证两者混合均匀。分别取各浓度适量粉末，置于 15 MPa 压力下持续压制 2 min，压制成约直径为 13 mm，厚度 1.0 mm 表面光滑无裂痕、上下面平行的圆形片状样本。将样片置于 T-SPEC 太赫兹时域光谱设备中，通过改变固定样片的二位平移台和 THz 波汇聚点的相对位置，每隔 0.2 mm 采集一个样本点的光谱数据。

1.3　光谱数据处理

1.3.1　光学参数

THz 波的原始信号是时域信号，经快速傅里叶变换(FFT)后，将参考信号 $r(t)$、样品信号 $s(t)$、样品厚度 d 和光速 c 作为输入，得到透射函数 $H(\omega)$和吸收系数 $\alpha_s(\omega)$[10]表达式如式(1)和式(2)

$$H(\omega)=\frac{FFT[s(t)]}{FFT[r(t)]} \tag{1}$$

$$\alpha_s(\omega)=\frac{2\omega K_s(\omega)}{c}=\frac{2}{d}\left\{\ln\left[\frac{4n_s(\omega)n_0}{|H(\omega)|[n_s(\omega)+n_0]^2}\right]\right\} \tag{2}$$

式(1)和式(2)中，$K_s(\omega)$为消光系数，ω 为角频率，n_0 和 n_s 分别为真空和被测样品复折射率。

1.3.2　原始光谱及光谱预处理

剔除明显异常数据，每个浓度获得 57 组，共计 342 组有效数据。由 FFT 得到吸收谱，如图 2(a)所示。实验光谱探测区间为 0～2.3 THz，根据信噪比选择 0.2～1.6 THz 区间进行分析。经多种预处理方法的比较，采用 Savtizky-Golay 平滑(SG)及均值中心化(mean centering)对数据预处理。处理后吸收谱如图 2(b)所示，1.1 THz 附近出现明显的特征波峰，这与论文资料一致[11]。各浓度光谱曲线有轻微的偏移，且相互叠加重合，掩盖了光谱的指纹特性。实验所得 THz 光谱特征波峰出现的偏移可能是因为实验过程中环境温度的波动所致。

1.3.3　样本选择方法

为了有效地覆盖多维向量空间，增加样本间的差异性和代表性，提高模型稳定性，采用 SPXY 方法[12]将样品分为训

练集和测试集，如表 1 所示。

表 1 L-TA 样本划分数据统计表

Table 1 Statistical table of L-TA sample division data

划分样本	样本数	最小值	最大值	均值	标准差
训练集	228	0	0.899 5	0.128 7	0.164 0
测试集	114	0	0.899 0	0.141 9	0.172 5

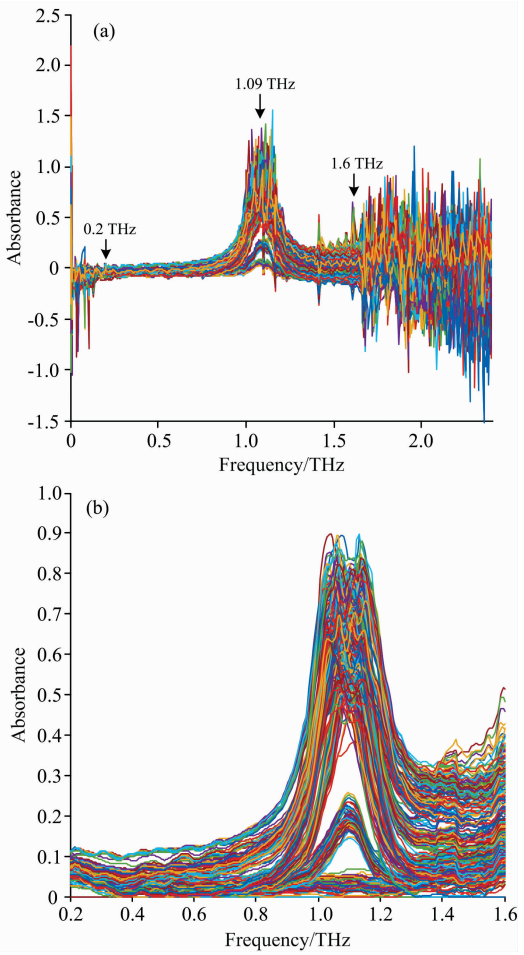

图 2 L-TA 样品的原始吸收谱(a)与经预处理后的吸收谱(b)

Fig. 2 The original absorption spectrum of the L-TA sample (a) and the pretreated absorption spectrum (b)

2 理论计算分析

2.1 DFT 仿真计算

使用 GAUSSIAN 16W 在 Windows 10 系统环境中进行仿真计算。由剑桥晶体数据中心获得单分子构型[13]如图 3 所示，采用 DFT 理论 B3LYP 方法，对氧原子和氢原子分别添加 d 和 p 轨道函数，以 6-31G*(d，p)为基组对单分子结构进行优化迭代计算[14-16]。相关研究证明，理论计算的频率是谐振频率，由于忽视了非谐振效应，基于此计算基频等可能会有一定误差。因此添加矫正因子(0.960)进行频率矫正。经优化结构的迭代计算与频率的计算，最终得到的 L-TA 计算结果没有虚频，说明得到了分子能量最稳定的结构。

在 0.2～1.6 THz 对 L-TA 单分子进行仿真计算。随机选取一组实验数据进行拟合。由图 3 可以看出，图中黑色曲线表示高斯仿真计算光谱，分子中 C1 相连羟基的转动带动碳链 C1—C2—C3—C4 进行集体振动，同时 C2—H4，C3—H5 和 O13—H14 形成的羟基有轻微摆动，造成了 L-TA 在 1.1 THz 处特征峰的出现。图 3 中红色曲线表示 L-TA 实验光谱拟合曲线，与仿真计算所得光谱的比较发现，实测光谱吸收峰较仿真计算光谱吸收峰向前偏移 0.02 THz，其轻微的谱峰偏移可能是由于实验测量是在室温下进行而理论仿真是基于绝对零度的计算下得出的。

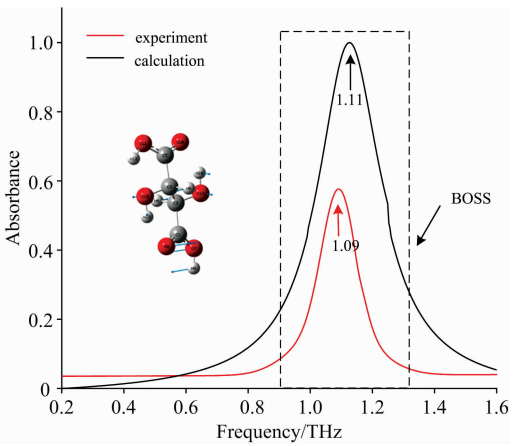

图 3 实验及仿真计算在 1.11 THz 分子振动示意图

Fig. 3 Schematic diagram of experimental and simulated molecular vibrations at 1.11 THz

2.2 基于 BOSS 算法的 L-TA 光谱的特征谱区筛选

2.2.1 变量的选取

在变量空间中利用 BBS 生成 K 个子集，对每个子集提取 BBS 选择的变量，并剔除重复变量以保持唯一性，对剔除后的变量赋予相同的权重(w)。

对提取的 K 个子集建立 PLS 模型，以交叉验证均方根误差(RMSECV)为模型判断标准，选取 RMSECV 最小值的模型为最佳模型，采用决定系数(R^2)对模型进行评价。

$$\mathrm{RMSECV}=\sqrt{\frac{\sum_{i=1}^{n}(y_{i,\ \mathrm{actual}}-y_{i,\ \mathrm{predicted}})^2}{n-1}} \tag{3}$$

$$R^2=1-\frac{\sum_{i=1}^{n}(y_{i,\ \mathrm{actual}}-y_{i,\ \mathrm{predicted}})^2}{\sum_{i=1}^{n}(y_{i,\ \mathrm{actual}}-\bar{y}_{i,\ \mathrm{actual}})^2} \tag{4}$$

式(3)和式(4)中，$y_{i,\ \mathrm{actual}}$ 为第 i 个样品参考方法的测定值，$y_{i,\ \mathrm{predicted}}$ 为第 i 个样品的预测值，$\bar{y}_{i,\ \mathrm{actual}}$ 为所有样品参考方法测定的平均值，n 为样品数。

将回归向量中的元素变为绝对值并归一化处理，再通过回归向量累加值结果赋予新的权重。

$$\omega_i=\sum_{k=1}^{K}b_{i,\ k} \tag{5}$$

式(5)中，K 是子模型的数量，$b_{i,\ k}$ 是第 k 个子模型中变量 i

的归一化回归系数绝对值。

根据变量新的权重选取新的子集，保证最佳子模型中回归系数较大的变量有更大的概率进入最优模型。重复以上步骤，直至新子集变量为 1。

3　结果与讨论

3.1　BOSS 算法对 L-TA 分析

采用 BOSS 算法对 L-TA 的 THz 光谱数据进行分析，由图 4(a)可以看出，RMSECV 值在 1～7 次迭代中稳定降低，由 0.030 0 减小为 0.026 7，在第 7～12 次迭代有持续上升趋势，由 0.0267 升为 0.028 7，随后增长趋势明显，在第 7 次迭代时有唯一最小的 RMSECV 值 0.026 7。因此，选取迭代数为 7 时的变量集进行分析。

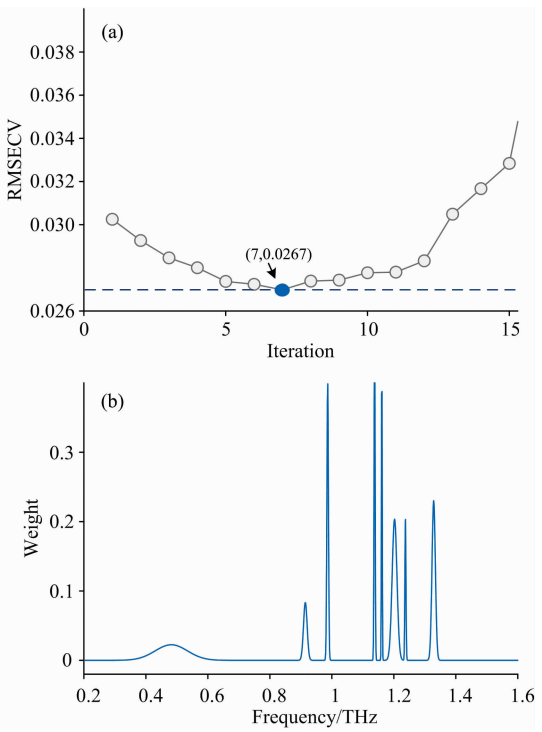

图 4　RMSECV 值随迭代次数变化图(a)和迭代次数为 7 时的权重图(b)

Fig. 4　RMSECV values with the number of iterations (a) and the weights at iteration number 7 (b)

提取第 7 次迭代中变量集赋予权重的值，如图 4(b)所示，在 0.4～0.6 THz 有一个权重波峰，但赋予权重较小。在 0.98～1.36 THz 频段，有非常明显的权重波峰，远大于 0.4～0.6 THz 的权重值。因此，根据权重值，将 0.98～1.36 THz 选为特征谱区。

3.2　不同建模方法比较

为了进一步验证 BOSS 算法对 L-TA 光谱数据特征谱区的筛选效果，将该算法与 iPLS，CARS-PLS 和 MC-UVE-PLS 三种经典的特征谱区选择方法及全谱 PLS 模型进行比较。BOSS，CARS-PLS 和 MC-UVE-PLS 模型参数均设置为 5 折交叉验证，采样次数为 500 次，最大潜在变量的数量为 10 个，迭代次数为 50 次，建模前均中心化预处理。对 iPLS 模型参数设置为 5 折交叉验证，建模前中心化预处理，主成分数量(principal components，PCs)经 PCA 回归确定为 5 个，子区间数范围为 2～60。

通过不同选择方法对 L-TA 的 THz 光谱进行特征谱段筛选。iPLS 特征谱区筛选效果受子区间数量设置的影响较大，子区间数为 4 时有最小的 RMSECV 和最大的 R^2。由图 5 可以看出，iPLS 选择谱区较为集中，但误差选择范围也大于其他算法，结果在表 2 得到验证，其 RMSECV 和 R^2 仅高于全谱 PLS 模型，低于其他三种算法。BOSS，CARS-PLS 和 MC-UVE 都选择了 0.45，0.93 和 1.1～1.24 THz 区域，这对应了 DFT 模型计算中 L-TA 分子 O—H 键转动的影响。CARS-PLS 和 MC-UVE-PLS 倾向于选择更多的变量，CARS-PLS 在 0.33 和 1.5 THz 附近选择了谱区，MC-UVE-PLS 在 0.2～0.4 和 0.92 THz 也有谱区的选择，但这些谱区未体现指纹特征，为无效信息谱区。相较于 CARS-PLS 和 MC-UVE-PLS，BOSS 的特征谱区选择更为集中，除 0.5 THz 附近有筛选的谱区，其他筛选谱区均在 0.93～1.34 THz 频段，与 DFT 仿真计算光谱特征谱区有良好的重合度，较完整地提取了特征波峰所在谱区，对特征谱区体现更加充分。

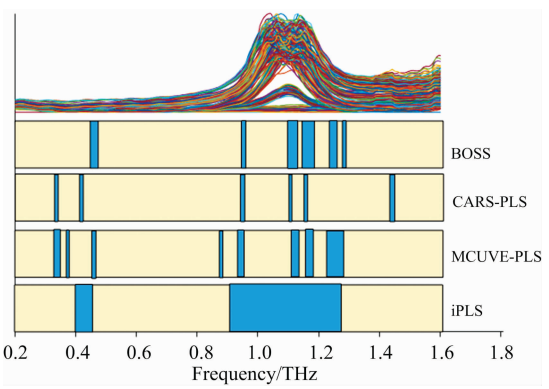

图 5　多方法对 L-TA THz 光谱的谱区筛选

Fig. 5　Multi-method filtering of L-TA THz spectra for frequency bands

表 2 总结了上述谱区筛选方法对 L-TA 的 THz 光谱计算效果。可以看出，四种谱区筛选方法相较于采用 PLS 法对全谱数据分析在测试集和验证预测精度均有所提高，BOSS 算法对预测能力的增强更为显著。与全谱 PLS 相比，BOSS 算法测试集决定系数 R^2_{train} 由 0.969 8 提高到 0.987 5，$RMSECV_{train}$ 由 0.039 降至 0.026，测试集 R^2_{test} 由 0.968 5 提高到 0.988 1，$RMSECV_{test}$ 由 0.042 降至 0.026，建模效果有显著提高。对比其他三种算法预测均方根误差(RMSEP)分布于 0.033～0.050 及全谱 PLS 的 RMSEP 为 0.066，BOSS 算法 RMSEP 为 0.026，有较大幅度减小，表现了更强的预测能力。对每种方法进行 50 次重复计算，BOSS 算法计算结果基本一致，其他计算方法均有一定波动，体现了 BOSS 算法较强的稳定性。

表 2 L-TA 光谱分析结果

Table 2 The spectral analysis of L-TA

Model	PCs_{train}	PCs_{test}	R^2_{train}	R^2_{test}	$RMSECV_{train}$	$RMSECV_{test}$	RMSEP
PLS	5	5	0.969 8	0.968 5	0.039	0.042	0.066
BOSS	9	9	0.987 5	0.988 1	0.026	0.026	0.026
iPLS	5	5	0.974 4	0.975 1	0.036	0.038	0.050
CARS-PLS	8	8	0.984 8	0.980 7	0.028	0.032	0.032
MC-UVE-PLS	9	10	0.984 3	0.986 3	0.031	0.028	0.033

结果表明，BOSS 算法相较于其他三种经典谱区算法，与 DFT 仿真光谱特征谱区有良好的一致性，谱区筛选效果更好，从而使模型具有更好的预测性能和稳定性能，这一研究也为其他复杂成分实验样品的定量分析提供了参考。

4 结 论

利用 THz-TDS 系统研究了 6 种不同浓度 L-TA 的 THz 光谱，基于 DFT 理论对 L-TA 单分子进行了仿真计算，分析了 1.11 THz 波峰处的振动模式，为筛选特征谱区的合理性提供了理论支撑。采用 BOSS 算法对 0.2～1.6 THz 频段吸收光谱进行特征谱区筛选，得到 R^2_{train} 为 0.987 5，R^2_{test} 为 0.988 1，RMSECV 为 0.026，RMSEP 为 0.026，相较 iPLS，CARS-PLS，MC-UVE-PLS 和全谱 PLS 建模，具有更佳的预测性能和稳定性能。结果表明，BOSS 算法选择的特征变量更为集中，选择 0.93～1.34 THz 频段的实验光谱与 DFT 理论仿真结果有良好的一致性。BOSS 算法结合 DFT 理论仿真计算可以简化模型，有效指认 THz 特征谱区，不仅有利于 L-TA 的分析，也为后续复杂样品如农产品的定量检测分析提供了参考。

References

[1] DI Zhi-gang, YAO Jian-quan, JIA Chun-rong, et al(邸志刚，姚建铨，贾春荣，等). Laser-infrared(激光与红外), 2011, 41(10): 1163.
[2] HU J, Chen R, Xu Z, et al. Sensors (Basel, Switzerland), 2021, 21(9): 3238.
[3] Li C, Zhao T, Li C, et al. Food Chemistry, 2017, 221: 990.
[4] Jiang Yuying, Li Guangming, Lü Ming, et al. Chin. Phys. B, 2020, 29(9): 145.
[5] Deng B, Yun Y, Cao D, et al. Analytica Chimica Acta, 2016, 908: 63.
[6] Grimme S, Antony J, Schwabe T, et al. Organic & Biomolecular Chemistry, 2007, 5(5): 741.
[7] Grimme S, Antony J, Ehrlich S, et al. Journal of Chemical Physics, 2010, 132(15): 154104.
[8] Zhou Q, Shen Y, Li Y, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2020, 236: 118346.
[9] Zhang Q, Chen T, Ma L, et al. Chemical Physics Letters, 2021, 767: 138350.
[10] Buzady A, Unferdorben M, Toth G, et al. Journal of Infrared Millimeter and Terahertz Waves, 2017, 38(8): 963.
[11] Soltani A, Gebauer D, Duschek L, et al. Chemistry-A European Journal, 2017, 23(57): 14128.
[12] Galvao R K H, Araujo M, José G, et al. Talanta, 2005, 67(4): 736.
[13] The Cambridge Crystallographic Data Centre: https://www.ccdc.cam.ac.uk/structures/.
[14] Lee C T, Yang W T, Parr R G. Phys. Rev. B, 1988, 37(2): 785.
[15] Becke A. Phys. Rev. A, 1988, 38(6): 3098.
[16] Maringolo M P, Tello A C M, Guimaraes A R, et al. J. Mol. Model., 2020, 26(10): 293.

Analysis and Identification of Terahertz Tartaric Acid Spectral Characteristic Region Based on Density Functional Theory and Bootstrapping Soft Shrinkage Method

TANG Xin, ZHOU Sheng-ling*, ZHU Shi-ping*, MA Ling-kai, ZHENG Quan, PU Jing
College of Engineering and Technology, Southwest University, Chongqing 402160, China

Abstract Terahertz time-domain spectroscopy contains the chemical and physical information of samples and indicates the background information related to equipment noise, sample status and environmental parameters. Its diversified spectrum may affect the model's performance and reduce the prediction accuracy. Therefore, extracting the characteristic information of target components, eliminatingredundant variables and screen the characteristic spectrum regions from the spectral data in a complex, overlapping and changing environment is of great significance for the quantitative and qualitative analysis of the terahertz spectrum. This paper collected the THz absorption spectra of 342 L-tartaric acid samples with concentrations of 10%, 20%, 40%, 50%, 60% and 80%. The B3LYP method in density functional theory (DFT) was used to optimize the monolecular model of L-tartaric acid based on 6-31G* (d, p) basis set, and the terahertz spectrum characteristics of the monolecular model were theoretically simulated. The molecular vibration modes corresponding to the characteristic wave peaks were analyzed, and the absorption spectra in the band of 0.2～1.6 THz were obtained. Compared with the measured absorption spectrum, the measured results agree well with the theoretical calculation results. The terahertz absorption spectrum of L-tartaric acid was screened using Bootstrapping soft shrinkage (BOSS). The competitive adaptive weighted sampling (CARS-PLS), Monte Carlo non-informational variable elimination (MC-UVE-PLS) and interval partial least square method (iPLS) were then compared and analyzed to obtain a better feature spectral region identification model. The analysis results indicate that the effective spectrum area obtained by the BOSS algorithm agrees better with the characteristic spectral region calculated by DFT theory. The L-tartaric acid spectrum modeling and regression analysis were conducted using full-spectrum PLS, CARS-PLS, MC-UVE-PLS, iPLS and BOSS algorithms. The experimental results imply that the prediction accuracy of the four spectral region screening methods is improved compared with the full spectrum PLS model. In addition, the prediction ability of the BOSS algorithm is improved most significantly by whose cross-validation root-mean-square error (RMSECV), prediction root-mean-square error (RMSEP), validation set determination coefficient (R^2_{test}) and test set determination coefficient (R^2_{train}) are 0.026 0, 0.026 0, 0.988 1 and 0.987 5 respectively, with higher prediction accuracy and model stability than other models. Therefore, it is foreseeable that, this study may provide an effective method for rapid and quantitative detection based on terahertz spectroscopy.

Keywords Terahertz spectrum; L-tartaric acid; Density functional theory; Spectral region; Bootstrapping Soft Shrinkage

(Received Jul. 26, 2021; accepted Oct. 25, 2021)

* Corresponding authors