

doi: 10.12452/j.fxcxb.26061803

基于自适应动态窗口迭代的太赫兹光谱基线校正算法

申洁, 燕芳*, 李嘉辉, 向天明

(内蒙古科技大学 自动化与电气工程学院, 内蒙古 包头 014010)

摘要: 针对太赫兹光谱基线漂移掩盖本征吸收峰、降低混合物定量预测精度的问题, 提出一种基于自适应动态窗口迭代的基线校正算法。该方法首先采用 Savitzky-Golay 平滑抑制高频噪声, 随后利用梯度统计特征定位起峰点, 并结合二阶多项式多锚点拟合、迭代较小值替换和斜率变化率判定, 实现了复杂背景下基线的自适应估计。以谷氨酰胺、组氨酸、苏氨酸及二元混合物为对象, 利用透射式太赫兹时域光谱系统采集吸收谱, 建立偏最小二乘回归 (PLS)、卷积神经网络 (CNN)、反向传播神经网络 (BPNN) 及最小二乘支持向量机 (LSSVM) 模型, 基于校正后的光谱进行定量分析。结果表明, LSSVM 模型性能最优, R^2 为 0.9960, RMSEP 为 0.543 1%, RPD 为 15.72, MAE 为 0.427 1%。在该文实验体系和评价指标下, 所建方法优于 ModPoly、AsLS 和 airPLS, 说明该算法在保留吸收峰特征和提升太赫兹光谱定量分析性能方面具有一定应用潜力。

关键词: 太赫兹光谱; 基线校正; 动态窗口; 桥接拟合; 定量分析

中图分类号: O657.3; O436 **文献标识码:** A **文章编号:** 1004-4957(XXXX)XX-0001-13

Terahertz Spectroscopy Baseline Correction Algorithm Based on Adaptive Dynamic Window Iteration

SHEN Jie, YAN Fang*, LI Jia-hui, XIANG Tian-ming

(School of Automation and Electrical Engineering, Inner Mongolia University of Science and Technology, Baotou 014010, China)

Abstract: To address the problem that baseline drift in terahertz spectra may obscure intrinsic absorption peaks and reduce the quantitative prediction accuracy for mixtures, an adaptive dynamic-window iterative baseline correction algorithm was proposed. Savitzky-Golay smoothing was first used to suppress high-frequency noise. The peak onset was then located using gradient-based statistical features, and the baseline was adaptively estimated by combining multi-anchor second-order polynomial fitting, iterative minimum-value replacement, and a convergence criterion based on the rate of slope change. Glutamine, histidine, threonine, and two binary mixtures, Gln-His and His-Thr, were selected as representative samples. Their absorption spectra were collected using a transmission terahertz time-domain spectroscopy system, and partial least squares (PLS), convolutional neural network (CNN), back-propagation neural network (BPNN), and least-squares support vector machine (LSSVM) models were established for quantitative analysis of the corrected spectra. The results showed that the LSSVM model achieved the best performance, with an R^2 of 0.996 0, an RMSEP of 0.543 1%, an RPD of 15.72, and an MAE of 0.427 1%. Under the experimental system and evaluation metrics used in this study, the proposed method showed better quantitative prediction performance than ModPoly, AsLS, and airPLS, indicating its potential for baseline correction of terahertz spectra from amino acid pellet samples.

Key words: terahertz spectroscopy; baseline correction; dynamic window; bridging fitting; quantitative analysis

太赫兹波具有高穿透性、低电离能以及对生物大分子集体振动模式敏感等独特优势, 近年来, 太赫兹光谱技术作为一种新兴的非接触式检测手段, 在物质成分鉴定与定量分析领域展现出广阔的应用

收稿日期: 2026-05-18; 修回日期: 2026-06-19

基金项目: 内蒙古自治区重点研发和成果转化计划项目 (2025SYFHH0062)

* 通讯作者: 燕芳, 博士, 教授, 研究方向: 基于太赫兹技术的无损检测, E-mail: 0472yanfang@163.com

网络首发日期: XXXX-XX-XX

前景^[1-3]。然而,在太赫兹光谱采集过程中,受系统温漂、样品颗粒及环境散射、背景噪声干扰等因素影响,原始太赫兹光谱往往伴随着明显的非线性基线漂移现象^[4-5],这不仅可以掩盖物质的本征吸收特征,还会弱化吸收峰强度与化学组分浓度之间的线性相关性,进一步影响后续基于光谱的定量回归模型预测精度,因此,必须对采集的原始光谱进行预处理与基线校正,以提高后续定量分析的可靠性和准确性。

针对光谱预处理与非线性基线校正问题,国内外学者提出了多种方法^[6-9]。闫碧等^[10]设计了一种改进的自适应惩罚最小二乘法,借助平滑系数与惩罚权重的动态调节机制实现背景扣除。该方法能较好地实现光谱背景的自动剥离,但在处理太赫兹频段内连续分布的多个吸收峰时,其全局参数的分配容易受局部强吸收信号的干扰,从而在拟合过程中可能导致部分特征峰的强度受损。为了提高基线提取效率,胡颖惠等^[11]提出了一种基于降采样的光谱基线校正方法,通过削减采样率数量来筛选背景节点并进行插值拟合。这一方案显著缩短了运算耗时,然而在处理特征峰分布密集且宽度极不规则的太赫兹光谱时,固定步长的降采样机制很难对所有非线性的局部突变维持同等程度的响应灵敏度。此外,在基线拟合模型的选择上,如胡怡格^[12]在光谱分析算法研究中所指出的,常规的高阶多项式虽然具有较强的非线性拟合能力,但在逼近离散的光谱背景点时容易引发局部的龙格(Runge)震荡,造成基线畸变。综上所述,如何准确剥离背景基线同时防止特征信号受损,依然是太赫兹光谱分析领域面临的一项挑战。

本文提出了一种基于自适应动态窗口迭代的基线校正算法。该算法首先利用 Savitzky-Golay 平滑滤波抑制高频噪声;其次,提取光谱起始频段,并创新性地利用基于梯度统计特征的自适应判定准则构建动态参考窗口,定位特征峰的起始位置;然后,采用二阶多项式进行多锚点跨区域桥接拟合,并结合经典 ModPoly 算法的迭代较小值替换思想^[13]与斜率变化率判定进行收敛迭代。该算法从物理逻辑上有效避免了特征提取中的信号侵蚀现象,实现了对真实光谱背景的精确估计。

为验证该算法在太赫兹氨基酸压片光谱体系中的适用性,本文以谷氨酰胺、组氨酸和苏氨酸三种氨基酸单质,以及谷氨酰胺-组氨酸、组氨酸-苏氨酸两种二元混合物为研究对象,获取其在不同浓度梯度下的太赫兹光谱。将完成基线校正的光谱数据分别利用偏最小二乘回归(PLS)、卷积神经网络(CNN)、反向传播神经网络(BPNN)及最小二乘支持向量机(LSSVM)4种模型进行定量建模与对比验证;同时,选取 ModPoly、AsLS 和 airPLS 等常用基线校正方法作为对比算法,进一步从定量预测性能角度评估本文方法在该实验体系下的相对优势。实验结果表明,本文算法能有效消除基线干扰,使各定量模型的预测性能均有所提升,并在与传统基线校正方法的横向比较中表现出更优的预测精度,为后续在更多太赫兹光谱体系中的推广验证提供了参考。

1 实验部分

1.1 实验设备

本实验采用如图1所示的透射式太赫兹时域光谱系统进行光谱数据采集。

该系统由飞秒激光器、太赫兹发射和探测模块、光学延迟线及数据采集单元组成。飞秒激光器为锁模钛蓝宝石振荡器,输出激光脉冲的中心波长为 800 nm,脉冲宽度为 10 fs,重复频率为 75.42 MHz,平均输出功率为 300 mW。为减少空气中水分对太赫兹波的强烈吸收,整个光路系统置于密闭箱体中并在实验过程中持续充入高纯度氮气以保证湿度<5%,环境温度保持在 25℃。

1.2 样品制备

实验选取谷氨酰胺、组氨酸和苏氨酸3种氨基酸单质,以及谷氨酰胺与组氨酸的二元混合物、组氨酸与苏氨酸的二元混合物作为测试样品。3种氨基酸均购自日本 Wako Pure Chemical Industries 公司,纯度均≥99.0%。作为稀释剂的聚乙烯粉末购自美国 Sigma-Aldrich

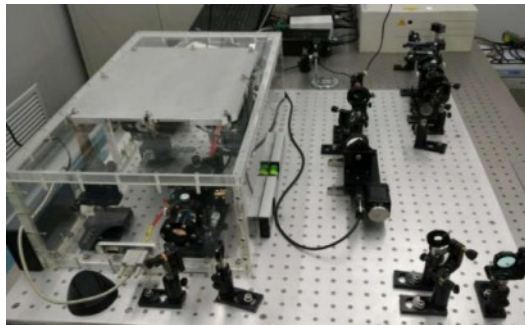


图1 透射式太赫兹时域光谱系统

Fig. 1 Transmission terahertz time-domain spectroscopy system

公司，微粒直径在 53~75 μm 之间。样品制备时使用精密电子天平严格称量待测样品与聚乙烯 (PE) 粉末，确保每个压片样品的总质量控制在 220 mg 左右。混合物置于玛瑙研钵中充分研磨约 5 min，并利用标准筛控制颗粒粒度，以最大程度减小样品颗粒对太赫兹波的米氏散射效应。混合均匀后的粉末装入模具，在 8 MPa 的压力下压制 2 min，制成直径 13 mm、厚度约 1.9 mm 的固体样片。为避免水分对太赫兹信号产生干扰，压制好的氨基酸样品均置于 60 $^{\circ}\text{C}$ 真空干燥箱中放置 2 h 以上，随后移入干燥器中冷却至室温保存。

本实验针对 3 种氨基酸单质及氨基酸二元混合物均设定了质量分数为 5%、10%、15%、20%、25% 和 30% 的浓度梯度，每种浓度的样品均独立制备 5 个平行样片。部分典型样片的具体质量配比如表 1 与表 2 所示。

表 1 组氨酸单质典型样片制备参数

Table 1 Preparation parameters of typical histidine pellets

Sample name	Concentration	PE mass (mg)	Analyte mass (mg)	Pellet thickness (mm)	Pellet mass (mg)
His	5%	209.1	11.2	1.89	216.8
His	10%	197.8	22.1	1.91	217.4
His	15%	187.3	32.8	1.90	215.9
His	20%	175.7	44.3	1.92	218.1
His	25%	165.2	54.8	1.90	216.3
His	30%	153.9	66.2	1.93	217.5

表 2 Gln-His 二元混合物样片制备参数

Table 2 Preparation parameters of Gln-His binary mixture pellets

Sample name	Concentration	PE mass (mg)	Gln mass (mg)	His mass (mg)	Pellet thickness (mm)	Pellet mass (mg)
Gln-His	5%	208.90	5.60	5.40	1.91	217.20
Gln-His	10%	198.20	10.90	11.10	1.90	216.50
Gln-His	15%	186.70	16.70	16.40	1.90	215.80
Gln-His	20%	176.30	21.80	22.10	1.93	218.00
Gln-His	25%	164.80	27.60	27.40	1.91	216.90
Gln-His	30%	154.10	32.90	33.10	1.92	217.70

1.3 原始光谱数据采集

本实验原始光谱采集在室温环境下进行。测试前，首先采集未放置样品时的时域信号，作为后续计算的参考信号 (Reference)。为减小压片不均匀带来的误差，每个样片均通过旋转角度测量 4 次，每次测试均进行多次脉冲累加平均以提高信噪比。基于上述采集方案，每个样片共获得 4 组原始光谱，即每种浓度梯度的 5 个平行样片共计获得 20 组原始光谱，每类样品设置 6 个浓度梯度，共获得 120 组原始光谱。本实验测试了 3 种氨基酸单质与 2 种氨基酸二元混合物，共获取 600 组有效的太赫兹时域光谱数据，为后续定量模型的建立提供了基础。

太赫兹光谱系统直接探测到的是携带样品信息的时域电场信号。设参考信号为 $E_{\text{ref}}(t)$ ，穿透样品后的信号为 $E_{\text{sam}}(t)$ 。通过快速傅里叶变换 (FFT)，可将时域信号转换为包含幅值和相位信息的频域信号，如公式 (1) 所示：

$$E(\omega) = \int E(t) e^{-i\omega t} dt \quad (1)$$

式中， ω 为角频率； t 为时间变量； i 为虚数单位。

样品的复透射系数 $H(\omega)$ 可由样品频域信号 $E_{\text{sam}}(\omega)$ 与参考频域信号 $E_{\text{ref}}(\omega)$ 的比值求得，如公式 (2) 所示：

$$H(\omega) = \frac{E_{\text{sam}}(\omega)}{E_{\text{ref}}(\omega)} = \rho(\omega) e^{-i\Delta\phi(\omega)} \quad (2)$$

$\rho(\omega)$ 为透射幅值比， $\Delta\phi(\omega)$ 为相位差。

考虑到压片样品前后存在空气/样品和样品/空气两个 Fresnel 透射界面，本文在忽略样品内部多次反射回波的条件下，对吸收系数进行界面透射校正^[14-15]。首先由相位信息估算样品折射率，如公式 (3) 所示：

$$n(\omega) = 1 + \frac{c\Delta\phi(\omega)}{\omega d} \quad (3)$$

式中, $n(\omega)$ 为样品折射率, c 为真空光速, d 为样品厚度。进一步结合 Fresnel 透射因子, 样品吸收系数可表示为公式 (4):

$$\alpha(\omega) = \frac{2}{d} \ln \left[\frac{4n(\omega)}{\rho(\omega)[n(\omega) + 1]^2} \right] \quad (4)$$

式中, $\alpha(\omega)$ 为样品吸收系数。式 (4) 考虑了样品前后两个界面的透射损耗, 忽略了样品内部多次反射回波。当内部反射回波在时域上可分离时, 可通过时间窗函数抑制其影响; 当反射回波与主脉冲部分重叠时, 该近似仍可能带来一定的系统误差。由于本文所有样品均采用相同的压片工艺、厚度控制和测试流程, 且不同基线校正方法均基于同一光谱提取流程进行比较, 因此该近似对不同预处理方法之间的相对性能评价影响较小。本实验中所有单质及混合物样片的吸收谱均基于上述物理模型提取。

为验证实验数据的可靠性与定量分析的可行性, 在提取吸收系数后, 进一步结合比尔-朗伯定律 (Beer-Lambert Law) 对样品的光谱特征进行了分析。依据该定律, 在一定浓度范围内, 物质对特定波长辐射的吸收程度与其浓度成正比。图 2 所示为经过 SG 平滑预处理后的太赫兹吸收光谱, 从图中可以观察到两类比较明显的光谱学特征: 首先, 受压片颗粒引起的米氏散射 (Mie scattering) 及系统低频响应波动影响, 原始光谱普遍存在随频率增加而逐步加剧的非线性基线漂移^[16]。这种宽频背景干扰与物质的本征吸收峰发生了重叠, 使得局部较弱吸收特征被遮盖; 其次, 随着浓度的增加, 在二元混合物体系中, 受氨基酸分子间氢键网络重组及偶极相互作用增强的影响, 不同浓度下的吸收峰不再是简单的等比例线性增长, 而是伴随出现了微小的非线性频移 (红移或蓝移) 现象^[17-18]。这种复杂的基线包裹与非线性的浓度偏移, 严重破坏了吸收系数与组分浓度之间的理想线性映射关系。为此, 本文提出了一种基于自适应动态窗口迭代的基线校正算法, 以有效校正背景干扰并还原光谱纯净特征, 从而为后续构建高精度的非线性定量预测模型奠定关键的数据基础。

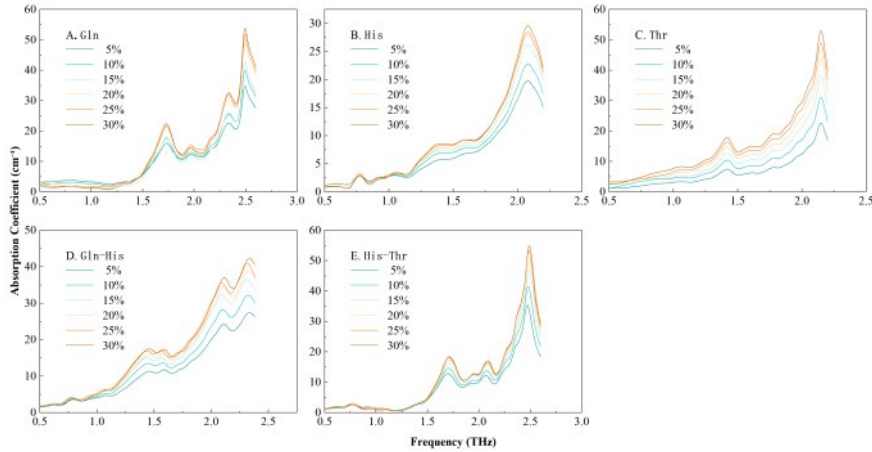


图 2 不同浓度下的氨基酸单质及二元混合物经 SG 平滑预处理后的太赫兹吸收光谱

Fig. 2 SG-smoothed THz absorption spectra of pure amino acids and binary mixtures at different concentrations

A. Gln; B. His; C. Thr; D. Gln-His; E. His-Thr (A. 谷氨酰胺; B. 组氨酸; C. 苏氨酸; D. 谷氨酰胺-组氨酸二元混合物; E. 组氨酸-苏氨酸二元混合物)

为进行后续的定量回归模型 (PLS、CNN、BPNN、LSSVM) 建立与评估, 本文对提取吸收系数后的 600 组太赫兹光谱数据进行训练集与预测集划分。考虑到同一样片重复测量所得光谱具有较高相似性, 若直接按单条光谱随机划分, 可能导致相似光谱同时进入训练集和预测集, 从而使模型性能评价偏高。因此, 本文以独立样片组作为数据划分的最小单位。每个独立样片经 4 次旋转测量获得的 4 组光谱始终整体划入训练集或预测集, 不进行拆分。每类样品包含 30 个独立样片组, 其中 21 个样片组作为训练集、9 个样片组作为预测集, 分别对应 84 组训练光谱和 36 组预测光谱; 5 类样品合计包括 105 个训练样片组和 45 个预测样片组, 对应 420 组训练光谱和 180 组预测光谱。划分过程按照样品类别和浓度

水平进行分层，并保证每次预测集均覆盖全部6个浓度水平。上述划分重复进行10次，最终以10次预测集评价指标的平均值表征模型性能。具体划分情况如表3所示。

表3 定量分析数据集划分表

Table 3 Dataset partitioning table for quantitative analysis

Sample name	Concentration levels	Total spectra	Training set	Prediction set	Repetitions
Gln	6	120	84	36	10
His	6	120	84	36	10
Thr	6	120	84	36	10
Gln-His	6	120	84	36	10
His-Thr	6	120	84	36	10
Total	-	600	420	180	-

1.4 定量分析模型与评价指标

为评价基线校正前后太赫兹光谱的定量分析性能，本文分别建立PLS、CNN、BPNN和LSSVM模型。模型输入为不同预处理条件下的太赫兹吸收光谱，输出为样品对应的浓度。为避免单一模型造成评价偏差，本文采用多种线性与非线性回归模型对校正效果进行对比分析。

为避免模型默认参数对基线校正效果评价造成影响，本文未直接采用默认参数。由于数据集采用10次独立样片组重复分组划分，因此在每次划分中，PLS、BPNN和LSSVM中需要选择的参数均仅利用该次训练集内部交叉验证重新确定。因此，对于不同重复划分以及不同光谱预处理条件，所得最优参数可能存在差异，预测集仅用于最终性能评价，不参与任何参数选择。

PLS模型采用训练集内部3折交叉验证，在1~15范围内选择潜变量数。BPNN采用单隐含层前馈神经网络，隐含层神经元数的候选值为10、20和30，并通过训练集内部3折交叉验证确定；训练函数采用缩放共轭梯度算法trainsecg，正则化参数设为0.001，最大训练次数为300，训练目标误差为 1×10^{-4} 。由于trainsecg通过梯度和尺度参数自适应调整更新步长，因此未设置固定学习率。

CNN采用轻量化一维卷积网络，依次包括输入层、卷积层、ReLU激活层、最大池化层、全连接层、ReLU激活层、Dropout层和回归输出层。卷积核数量为16，卷积核尺寸为3，池化窗口为2，全连接层神经元数为32，随机失活率为0.20，初始学习率为0.001，批量大小为32，最大训练轮数为50，优化器采用Adam。CNN网络结构和训练参数预先固定，并在原始光谱和校正光谱建模过程中保持一致。

LSSVM采用径向基核函数。正则化参数C的候选值为1000和10000；核函数参数 γ 以训练集标准化后样本间平方欧氏距离中位数的倒数为基准，并分别乘以0.3、1、3、10和30构成候选参数。每次划分均采用训练集内部5折交叉验证选择交叉验证误差较小的C和 γ 组合。所有模型的预测集均仅用于最终性能评价。

模型预测性能采用预测集决定系数（ R^2 ）、预测均方根误差（RMSEP）、相对预测偏差（RPD）和平均绝对误差（MAE）进行评价。一般而言， R^2 越接近1，RMSEP和MAE越小，RPD越大，表明模型的拟合效果和预测性能越好^[19]。各指标计算公式如下：

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (5)$$

$$\text{RMSEP} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (6)$$

$$\text{SD} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} \quad (7)$$

$$\text{RPD} = \frac{\text{SD}}{\text{RMSEP}} \quad (8)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (9)$$

式中, y_i 为第 i 个样本的真实浓度值; $\hat{y}_i$ 为第 i 个样本的预测浓度值; $\bar{y}$ 为预测集中真实浓度的平均值; n 为预测集样本数; SD 为预测集中真实浓度的标准差。

2 基于自适应动态窗口迭代的基线校正算法

本文提出的基于自适应动态窗口迭代的基线校正算法的整体执行流程如图 3 所示。首先, 利用 Savitzky-Golay (SG) 平滑滤波对实验获取的原始光谱进行预处理, 抑制系统高频随机噪声; 其次, 构建基于梯度统计特征的自适应动态窗口, 实现对弱吸收特征峰起点的精准锁定; 最后, 采用二阶多项式构建多锚点跨区域桥接拟合机制, 并结合迭代较小值替换逻辑逼近真实的底层物理散射包络。该算法的执行逻辑与原理如图 4 所示。

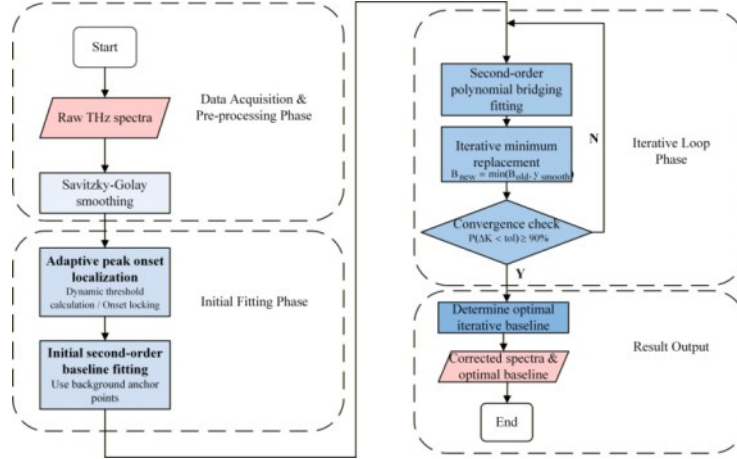


图 3 算法流程图

Fig. 3 Algorithm flowchart

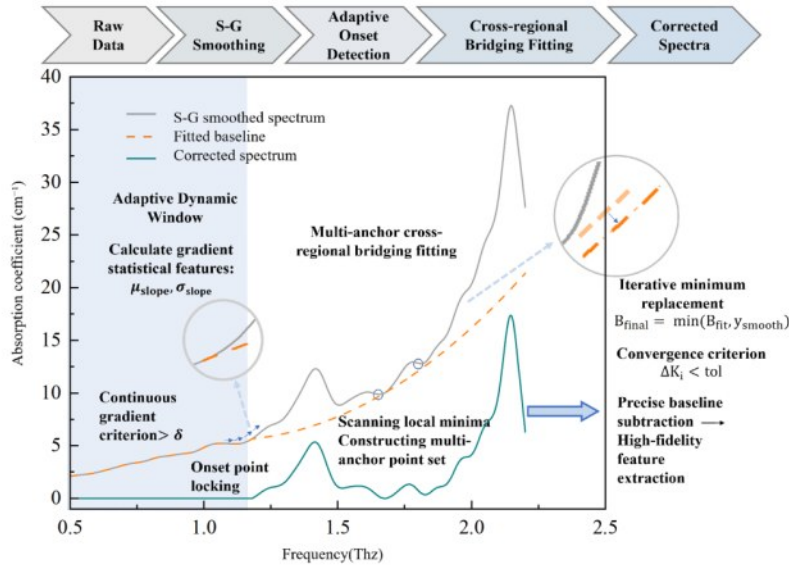


图 4 算法执行逻辑与原理

Fig. 4 Algorithm logic and principle

2.1 光谱预处理

SG 滤波是一种基于最小二乘准则的局部多项式拟合方法, 其核心逻辑是在特定长度的移动窗口内, 利用低阶多项式对采样点进行加权平滑^[20]。该算法能够在抑制太赫兹检测系统高频随机噪声的同时, 最大限度地保留特征吸收峰的本征形态与强度信息。经过平滑处理后, 窗口内中心采样点的预估值 $\hat{y}_i$ 满足公式 (10):

$$\hat{y}_i = \sum_{k=0}^p a_k f_i^k \quad (10)$$

式中, f_i 为频域采样点; a_k 为多项式拟合系数; p 为拟合阶次。

本实验中, SG 滤波窗口长度设置为 41 个频点, 多项式阶次取 $p=2$ 。该窗口长度与特征峰宽度较为匹配, 可在抑制噪声的同时保留峰形细节; 采用二阶多项式能够拟合局部的非线性趋势, 避免因过度平滑导致弱吸收峰消失。这一预处理步骤为后续基线提取提供了平滑的特征包络, 有效避免了噪声干扰引起的起峰定位误判。在此基础上, 光谱背景校正通常结合多项式拟合方法, 通过最小二乘法构建描述基线走势的连续背景曲线。

2.2 自适应起峰检测方法

在利用 SG 滤波完成光谱预处理之后, 针对传统固定窗口方法难以准确识别弱信号起峰点的问题, 本文提出一种基于梯度统计特征的自适应起峰检测方法。具体如图 4 左侧阴影区域及局部放大图所示, 其核心思想是利用动态窗口的梯度统计特性, 自适应构建起峰判别阈值, 并通过连续超阈值判断与回退机制实现对起峰位置的精确定位。

首先, 定义经过 SG 平滑预处理后的离散光谱序列为 $Y = \{y_1, y_2, \dots, y_N\}$, 其中 N 为采样点总数。为量化光谱序列的局部波动趋势, 计算其一阶向前差分序列作为梯度特征 $G = \{g_1, g_2, \dots, g_{N-1}\}$, 如公式 (11) 所示:

$$g_i = y_{i+1} - y_i \quad (11)$$

式中, g_i 为第 i 个采样点的一阶前向差分值; y_{i+1} 和 y_i 分别为第 $i+1$ 个和第 i 个采样点的光谱原始强度值。

随后, 提取光谱起始频段的 M 个连续数据点作为自适应动态窗口, 在该窗口内, 背景基线的梯度统计特性通过均值 μ_{slope} 与标准差 σ_{slope} 进行计算, 分别如公式 (12) 和 (13) 所示:

$$\mu_{\text{slope}} = \frac{1}{M} \sum_{i=1}^M g_i \quad (12)$$

$$\sigma_{\text{slope}} = \sqrt{\frac{1}{M-1} \sum_{i=1}^M (g_i - \mu_{\text{slope}})^2} \quad (13)$$

式中, M 为提取的动态参考窗口内的连续数据点总数。

基于上述统计参数, 构建起峰判定阈值 δ , 如公式 (14) 所示:

$$\delta = \mu_{\text{slope}} + \lambda \cdot \max(\sigma_{\text{slope}}, \varepsilon) \quad (14)$$

式中, λ 为检测敏感度的因子, 用于调节起峰判定阈值, ε 为极小正数, 本文取 $\varepsilon = 1 \times 10^{-4}$, 用于提供基础噪声容限, 避免在低噪声区域 σ_{slope} 过小导致阈值失去约束力。

根据公式 (14) 可知, λ 值越大, 起峰判定阈值越高, 弱吸收峰起点识别可能滞后; λ 值越小, 起峰判定阈值越低, 局部噪声波动可能被误判为吸收峰起点。为确定其取值, 本文在训练集内部对 λ 进行参数敏感性分析, 候选范围设为 3.0 ~ 5.0, 步长为 0.1。在固定 SG 平滑窗口长度为 41、SG 多项式阶次为 2、连续超阈值判定点数 L 为 10、回退步长 R 为 5 以及桥接拟合多项式阶次为 2 的条件下, 对于每个候选 λ , 首先利用该参数完成训练集光谱的基线校正, 随后建立 LSSVM 定量模型, 并采用训练集内部 3 折交叉验证的均方根误差作为评价指标。预测集不参与 λ 的选择。比较结果表明, $\lambda = 4.2$ 时训练集交叉验证均方根误差最小, 因此本文将检测敏感度因子设定为 4.2。

该方法的创新之处在于引入了连续性的统计判据, 以克服单点突变带来的偶然性误判。定义指示函数 $I(i)$ 用于表征单点梯度是否突破阈值, 如公式 (15) 所示:

$$I(i) = \begin{cases} 1, & g_i > \delta \\ 0, & g_i \leq \delta \end{cases} \quad (15)$$

式中, 当单点梯度 g_i 大于阈值 δ 时, 指示函数取 1, 反之取 0。

检测过程从参考窗口末端开始向后扫描, 设定评估窗口长度为 L (本文取 $L = 10$)。当且仅当局部窗口内的梯度序列连续满足超阈值条件时, 即满足公式 (16) 的严格约束:

$$\sum_{j=n}^{n+L-1} I(j) = L \quad (16)$$

式中, L 为设定的评估窗口长度; n 为首个满足该约束条件的初始突变点索引; j 为求和变量。

此时, 将首个满足该约束条件的索引确认为初始突变点。随后, 为了精确定位于纯背景的峰脚处, 需将该索引向低频方向回退 R 个步长 (本文设 $R = 5$), 即最终确定的有效特征峰起始锚点索引为 $n_{\text{start}} = n - R$ 。此方法能够自动区分由真实物质吸收引起的梯度突变与由系统基线缓慢漂移或高频噪声引起的微小波动, 从而显著提升了目标物质弱吸收边界识别的精度与鲁棒性。

2.3 多锚点跨区域桥接拟合

在精准锁定信号起始边界后, 为消除吸收峰区域数据对基线形态估计的干扰, 本文采用了一种多锚点跨区域桥接拟合方法。考虑到氨基酸吸收谱高频段仍分布有特征吸收峰, 本方法在提取起峰点前的全部纯背景数据段后, 通过利用一阶离散差分算子计算光谱的斜率, 自动全局扫描并定位高频区内特征吸收峰之间的极小值点。将起峰点前的纯背景数据与所有动态识别出的极小值数据联合, 共同构建去除吸收峰特征的非连续拟合锚点集。该锚点集包含 N 个离散频点, 构成平滑锚点序列 $\mathbf{y} = [y_1, y_2, \dots, y_N]^T$ 与频率坐标向量 $\mathbf{f} = [f_1, f_2, \dots, f_N]^T$ 。

针对上述提取的锚点集, 考虑到太赫兹光谱中的缓慢非线性基线漂移, 本文采用二阶多项式对上述锚点集进行多锚点跨区域桥接拟合。该多项式拟合过程可转换为标准的线性代数形式 $\mathbf{B} = \mathbf{X}\mathbf{c}$ 。其中, $\mathbf{c} = [c_0, c_1, c_2]^T$ 为拟合系数向量, $\mathbf{X}$ 为元素定义为 $X_{i,k} = f_i^{k-1}$ 的 $N \times 3$ 的范德蒙德 (Vandermonde) 矩阵。为求解最优系数, 该方法通过最小化观测向量与模型向量间的残差平方和构建目标函数 $J(\mathbf{c})$, 如公式 (17) 所示:

$$J(\mathbf{c}) = \|\mathbf{y} - \mathbf{X}\mathbf{c}\|_2^2 \quad (17)$$

式中, $\mathbf{y}$ 为观测向量 (即提取的锚点集); $\mathbf{X}$ 为范德蒙德 (Vandermonde) 矩阵; $\mathbf{c}$ 为待求的二阶多项式拟合系数向量; $\|\cdot\|_2$ 表示向量的二范数。

根据极值必要条件, 令导数 $\frac{\partial J}{\partial \mathbf{c}} = 0$, 推导出该无约束优化问题的闭式解析解, 即正规方程, 如公式 (18) 所示:

$$\mathbf{c} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (18)$$

式中, $\mathbf{X}^T$ 为矩阵 $\mathbf{X}$ 的转置; $(\mathbf{X}^T \mathbf{X})^{-1}$ 为矩阵 $\mathbf{X}^T \mathbf{X}$ 的逆矩阵。

进而求得当前多项式拟合的最小二乘解。上述 $N \times 3$ 阶设计矩阵实质上对应了二阶多项式的三项系数。在此选用二阶而非高阶多项式, 是基于对太赫兹光谱压片样品底层物理散射规律的考量。由于压片颗粒尺寸与太赫兹波长处于相近量级, 太赫兹波穿透该非均匀颗粒介质时, 颗粒散射会对透射谱产生显著影响, 并可能表现为米氏散射或符合 Rayleigh-Gans-Debye 近似的体散射特征^[16]。已有太赫兹粉末样品研究表明, 颗粒散射会引起随频率变化的透射能量损耗, 并在吸收谱中形成连续的宽频背景抬升^[4, 16]。因此, 本文采用二阶多项式描述这种缓慢变化的非线性背景趋势, 避免使用高阶多项式对局部吸收峰进行过度拟合。选用二阶多项式不仅是为了在数学上引入形态刚性, 以起到隐式结构正则化的作用, 从而规避高阶多项式在逼近离散节点时可能引发的龙格震荡现象; 更重要的是, 低阶多项式平滑递增的曲线形态在物理意义上契合由颗粒散射主导的光谱基线宏观漂移特征。这种由物理机制驱动的模式选择, 能够确保拟合基线真实反映连续的非线性散射背景, 有效防止高阶拟合曲线出现过拟合而错误地侵蚀局部的本征吸收峰。

为进一步确保拟合基线严格贴合真实光谱的底部包络, 防止基线侵入特征吸收区造成削峰, 鉴于常规最小二乘法极易受吸收峰突起干扰而导致基线整体抬升, 本算法创新性地引入了迭代较小值替换机制。该迭代逻辑与收敛判据如图 4 右侧流程图所示。在每一次迭代 j 中, 逐点比较当前拟合基线与平滑光谱的强度, 强制限制最终基线 B_{final} 在全频段内均不高于平滑光谱的强度。如公式 (19) 所示:

$$B_{\text{final}}^{(j)} = \min(B_{\text{fit}}^{(j)}(x_i), y_{\text{smooth}}(x_i)) \quad (19)$$

式中, x_i 为当前频率节点; $B_{\text{fit}}^{(j)}(x_i)$ 为拟合生成的初始基线; $y_{\text{smooth}}(x_i)$ 为经 SG 滤波后的平滑光谱值。

算法通过计算迭代前后斜率的变化率 ΔK_i 判断收敛状态, 如公式 (20) 所示:

$$\Delta K_i = \frac{|K_j(x_i) - K_{j-1}(x_i)|}{|K_{j-1}(x_i)| + \varepsilon} \quad (20)$$

式中, $K_j(x_i)$ 为第 j 次迭代后 x_i 处的斜率值, ε 为防止分母为零的极小常数。为平衡计算效率与拟合精度, 设定收敛阈值 $\text{tol} = 1 \times 10^{-3}$, 该值的选取参考了太赫兹吸收系数数据的量化精度 (通常为 10^{-3} 量级) 与斜率分布范围: 当斜率变化率 ΔK_i 降至此区间时, 基线形态的改变量已低于系统的随机噪声水平, 继续迭代将导致算法捕捉非物理的随机扰动。因此, 设定此阈值可有效避免过拟合, 确保基线在达到视觉与计算意义上的平稳状态时及时终止。

当全频段内超过 90% 的数据点满足斜率变化率 $\Delta K_i < \text{tol}$ 时停止迭代。为准确判定全局稳态并量化这一容错机制, 本文引入指示函数 $I(x_i)$, 当满足局部收敛条件时该函数取值为 1, 反之取值为 0。当全局满足公式 (21) 时即自动终止迭代:

$$\frac{1}{N} \sum_{i=1}^N I(x_i) \geq 90\% \quad (21)$$

式中, N 为全频段内的数据点总数; $I(x_i)$ 为局部收敛指示函数。

3 结果与讨论

为了全面验证本文提出的自适应动态窗口迭代基线校正算法的实际性能, 本节结合实验采集的 600 组太赫兹光谱数据, 从光谱形态特征还原、定量回归预测以及不同基线校正方法横向对比 3 个层面对处理结果展开验证。首先, 基于图 5 所示的氨基酸单质及其二元混合物校正前后的光谱形态对比, 直观展示算法在剥离复杂非线性基线及还原物质本征吸收峰方面的高保真效果; 其次, 将基线校正前后的光谱数据分别输入至 PLS、CNN、BPNN 和 LSSVM 四种模型中, 通过图 6 的回归散点图及核心评价指标, 量化分析本文算法对模型预测精度的提升作用; 最后, 固定采用 LSSVM 模型, 进一步与 ModPoly、AsLS 和 airPLS 等传统基线校正方法进行比较, 以验证本文方法在复杂光谱背景下的相对优势。

3.1 多光谱基线校正结果与特征提取

为直观验证本文提出的自适应动态窗口迭代算法在太赫兹光谱背景扣除中的有效性, 选取谷氨酰胺 (Gln)、组氨酸 (His)、苏氨酸 (Thr) 3 种单质及其二元混合物 (Gln-His、His-Thr) 的典型光谱, 对比基线校正前后的光谱形态变化。图 5A~E 分别展示了上述五种样品在基线校正前后的太赫兹吸收光谱。

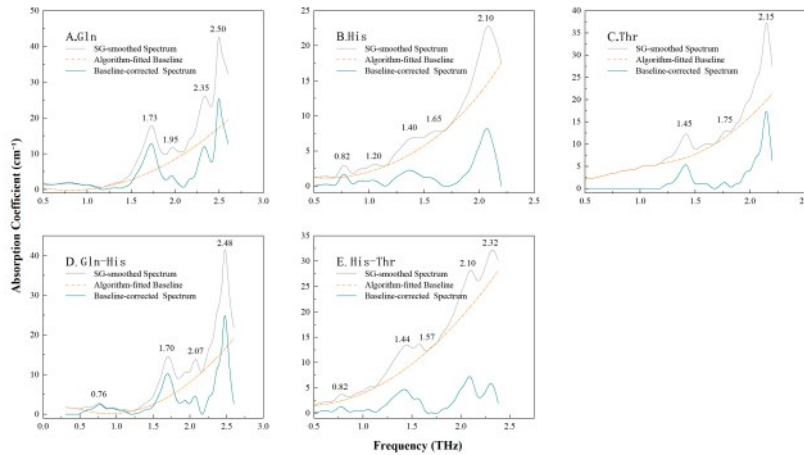


图 5 太赫兹光谱基线校正结果

Fig. 5 Baseline correction results of terahertz spectra

如图 5A~C 所示, 3 种氨基酸单质的原始光谱均呈现出明显的非线性基线漂移, 尤其在 1.5 THz 以上高频区域, 受颗粒散射及系统响应影响, 光谱基底随频率增加而显著抬升, 部分弱吸收特征被完全淹没。应用本文算法进行基线校正后, 光谱非特征吸收区域的基底被精确压至零点附近, 系统性吸收

系数偏移被有效消除。各物质的本征吸收峰得到了高保真还原：谷氨酰胺在 1.73 THz 处呈现明显次主峰，在 1.95 THz 和 2.35 THz 处有辅助吸收，2.50 THz 处为主峰，校正后所有峰位精准保留，峰形锐利无畸变；组氨酸在 0.82 THz 低频区存在特征弱吸收，1.20 THz 和 1.40 THz 处有次级峰，1.65 THz 处为过渡峰，2.10 THz 处为宽泛的主吸收带，校正后这些特征被完整呈现，尤其是 0.82 THz 弱峰在基线扣除后清晰可辨；苏氨酸的主峰位于 2.15 THz，次峰位于 1.45 THz，1.75 THz 处有辅助吸收，校正后主峰与次峰的峰位、峰高比及峰宽均与单质本征特性一致。上述结果得益于算法中多锚点跨区域桥接拟合与迭代较小值替换机制的协同作用，有效避免了传统方法中因基线拟合过度而导致的“信号切割”现象。

对于成分更为复杂、吸收频带密集且重叠严重的二元混合物体系，本文算法同样表现出优异的鲁棒性，不仅成功剥离了非线性背景，还精准捕捉了因分子间相互作用引起的峰位偏移、融合及双峰特征。如图 5D 所示，Gln-His 二元混合物的原始光谱在 0.7 ~ 2.6 THz 范围内呈现多个重叠吸收特征，基线干扰严重。校正后，两组分的特征峰被清晰分离并准确识别：混合物在 0.76 THz 处出现吸收峰，相较于组氨酸单质的 0.82 THz 特征峰发生了约 0.06 THz 的红移；在 1.70 THz 处形成宽峰，对应谷氨酰胺 1.73 THz 与组氨酸 1.65 THz 特征峰的交叠融合；主峰位于 2.48 THz，较谷氨酰胺单质的 2.50 THz 主峰发生轻微红移；此外在 2.07 THz 处也保留了组氨酸的宽吸收带特征。如图 5E 所示，His-Thr 二元混合物的校正光谱同样展现出优异的特征分离效果：混合物继承了组氨酸的 0.82 THz 和 2.10 THz 特征峰；苏氨酸的 1.45 THz 特征峰在混合体系中表现为 1.44 THz 与 1.57 THz 的宽频双峰特征；在 2.32 THz 处出现因分子间相互作用形成的重叠新突起。

上述峰位与文献基本一致^[17, 21-22]。峰位偏移、融合与双峰特征反映了不同氨基酸分子在混合态下的氢键网络重组及相互作用^[17-18]，校正后的光谱完整保留了这些物理化学信息。实验结果表明，本文算法不仅能够有效剥离单质光谱中的非线性散射背景，更能在混合物复杂重叠谱带中实现高保真的特征还原，为后续定量回归建模提供了高质量的数据基础。

3.2 基于校正光谱的定量回归分析

为进一步验证经自适应动态窗口迭代算法校正后的太赫兹光谱能够有效提升定量分析准确度与精度，将采集获取的氨基酸单质及二元混合物的原始吸收谱与完成基线校正后的校正光谱分别输入 PLS、CNN、BPNN 以及 LSSVM 四种模型进行浓度预测。

为直观比较不同模型在基线校正前后的浓度预测分布，图 6 展示了 10 次独立样片组分划分所得预测结果。图中每个散点表示相应样品类别在某一真实浓度水平下的平均预测值，模型的最终评价指标则由每次预测集中的全部样本分别计算，并取 10 次重复分组划分结果的平均值。

如图 6 所示，A、C、E、G 为基于原始光谱的定量预测结果，B、D、F、H 为基于校正光谱的定量预测结果，从左至右依次为 PLS、CNN、BPNN、LSSVM 模型的浓度预测回归散点图。纵向对比基线校正前后的回归散点图可知，原始光谱由于受严重背景干扰及非线性基线漂移的影响，各定量模型的预测点均呈现出较大离散度，大幅偏离理想对角线。而基于本文算法校正后的光谱，背景干扰被有效剥离，吸收特征与物质浓度之间的线性相关性得到修复，所有定量模型的预测点均更加紧密地聚集在理想回归线两侧。

横向对比不同模型的预测性能可知，非线性模型（CNN、BPNN、LSSVM）在处理混合物等复杂组分时均表现出较好的非线性映射能力。这种差异的根源在于：PLS 本质上是一种严格的线性回归方法，而太赫兹频段下混合体系的光谱并非单组分光谱的简单线性叠加，由于氨基酸分子间存在显著的氢键相互作用，导致吸收谱线出现了峰位偏移与非线性宽频包裹现象。

在 3 种非线性模型中，LSSVM 模型在所有测试中展现出最优的定标性能。CNN 和 BPNN 作为典型的人工神经网络模型，其网络权重的迭代优化高度依赖于大规模样本的支撑。由于太赫兹光谱数据通常呈现高维数、小样本的特点，神经网络在小样本条件下极易陷入局部最优解或发生过拟合现象^[23]，导致其预测泛化能力受限。相比之下，LSSVM 模型建立在统计学习理论和结构风险最小化原则之上^[24]。它通过将非线性问题映射到高维特征空间并转化为求解线性方程组，不仅有效规避了神经网络的局部极小值问题，而且在处理小样本、非线性及高维模式识别时具有先天的泛化优势。

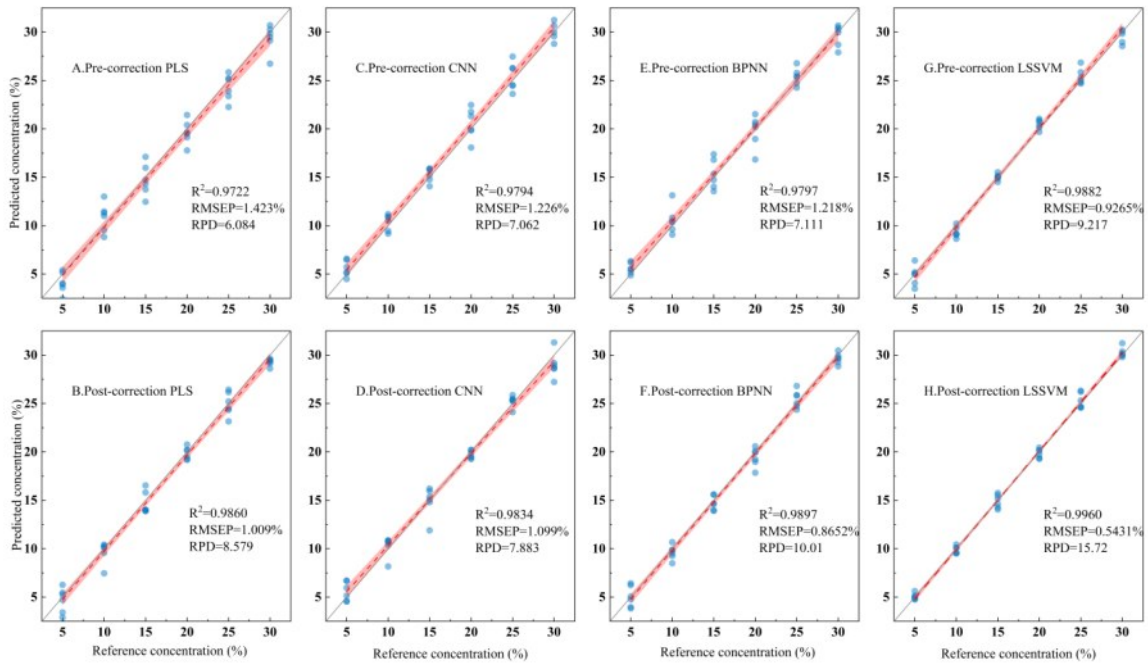


图 6 校正前后不同定量模型的浓度预测结果

Fig. 6 Concentration prediction results of different quantitative models before and after correction
before correction: A, C, E, G; after correction: B, D, F, H; the models are PLS, CNN, BPNN, and LSSVM, respectively

从物理机制上看, PLS作为一种线性回归方法, 其隐含的假设是吸收系数与浓度之间满足严格的线性关系。然而, 太赫兹光谱中的混合物体体系往往存在复杂的非线性效应: 首先, 随着浓度变化, 氨基酸分子间的氢键网络发生重组, 导致特征吸收峰位出现微小偏移 (如 Gln-His 混合物中 0.76 THz 和 2.48 THz 的红移), 这打破了单一频率下吸光度与浓度的线性比例关系; 其次, 不同分子间的偶极相互作用在混合态下可能产生新的振动耦合, 使得混合物的吸收谱不再是各组分光谱的简单线性叠加; 此外, 颗粒散射背景虽然被校正, 但残留的散射影响在高浓度下仍可能引入非线性贡献。PLS 无法拟合这些由分子间相互作用引起的非线性特征, 因此其预测精度相对受限。而非线性模型 (CNN、BPNN、LSSVM) 能够通过激活函数或核函数引入非线性映射, 自动学习浓度与吸收谱之间的复杂关联, 从而获得更优的拟合效果。

结合图 6 所示的预测散点分布及 10 次重复分组划分所得评价指标的平均结果可知, 采用本文算法校正后, LSSVM 模型的预测点分布最为集中, 整体预测偏差较小, 表现出最优的浓度预测能力。其预测集 R^2 由校正前的 0.9882 提升至 0.9960, RMSEP 由 0.9265% 降低至 0.5431%, RPD 由 9.217 提升至 15.72。由于图 6 中主要标注 R^2 、RMSEP 和 RPD 三项指标, 为进一步说明模型平均预测偏差, 本文补充比较了 LSSVM 模型的 MAE 结果, 其 MAE 由校正前的 0.6434% 降低至 0.4271%, 说明校正后模型的平均绝对预测误差进一步减小。上述结果表明, 本文算法能够有效削弱非线性基线漂移对光谱定量建模的影响, 使校正后的光谱更能反映样品浓度变化所对应的有效吸收信息。

3.3 不同基线校正方法的定量预测性能对比

为进一步验证本文所提自适应动态窗口迭代基线校正算法相较于传统基线校正方法在本实验数据集集中的表现, 选取 ModPoly^[13]、AsLS^[25] 和 airPLS^[26] 三种常用基线校正算法作为对比方法, 并将原始光谱及不同方法校正后的光谱分别输入 LSSVM 模型进行定量预测。为避免不同回归模型对基线校正效果评价造成干扰, 采用表现较优的 LSSVM 模型作为回归模型, 并采用 R^2 、RMSEP、RPD 和 MAE 四个测试集评价指标对不同基线校正方法的定量分析性能进行比较。同时, 为评价不同基线校正方法的计算效率, 在相同软硬件环境、相同输入数据规模和相同运行条件下, 统计各方法完成全部 600 组太赫兹光谱基线校正所需的时间。计时范围仅包括参数确定后各方法完成基线校正的计算过程, 不包括数据读取、参数寻优、定量模型训练、评价指标计算、绘图及文件保存等过程。各方法分别进行多次独

立运行,并以完成全部600组光谱校正所需总时间的平均值作为最终计算时间,结果如表4所示。

表4 不同基线校正方法结合 LSSVM 模型的预测性能对比

Table 4 Comparison of LSSVM prediction performance with different baseline correction methods

Data processing method	R ²	RMSEP (%)	RPD	MAE (%)	Time (s)
Raw spectra	0.988 2	0.926 5	9.217	0.643 4	—
ModPoly	0.990 6	0.827 9	10.31	0.627 7	0.721 3
AsLS	0.991 9	0.765 9	11.15	0.596 9	1.457
airPLS	0.988 6	0.911 8	9.365	0.598 4	0.623 9
Proposed	0.996 0	0.543 1	15.72	0.427 1	2.137

由表4可知,相比原始光谱,经过基线校正后的光谱整体上均能提升LSSVM模型的预测性能,说明基线漂移的有效剥离有助于增强太赫兹吸收特征与样品浓度之间的映射关系。其中,本文方法在4项测试集指标上均取得最优结果,R²达到0.996 0, RMSEP降至0.543 1%, RPD提升至15.72, MAE降低至0.427 1%。与原始光谱相比,本文方法使RMSEP降低约41.38%, RPD提升约70.55%, MAE降低约33.62%。

与传统基线校正方法相比,本文方法同样表现出更明显的优势。以对比方法中预测效果较好的AsLS为例,本文方法的RMSEP进一步降低约29.09%, RPD提升约41.99%, MAE降低约28.44%。这表明,本文算法通过自适应起峰检测、多锚点跨区域桥接拟合以及迭代较小值替换机制,能够准确地贴合真实光谱背景,减少传统方法在弱吸收峰和重叠吸收峰区域可能产生的特征削弱现象,从而为LSSVM模型提供了更高质量的光谱输入。

从计算时间来看,ModPoly、AsLS、airPLS和本文方法完成全部600组光谱基线校正的平均总耗时分别为0.721 3、1.457、0.623 9、2.137 s。本文方法由于增加了自适应起峰检测、多锚点跨区域桥接拟合、迭代较小值替换和收敛判定等步骤,计算时间高于3种对比方法。但其完成全部600组光谱校正仍仅需2.137 s,满足检测要求。结合其在R²、RMSEP、RPD和MAE等指标上的表现,说明本文方法能够在增加有限计算时间的情况下改善定量预测性能,并能够满足当前实验规模下批量太赫兹光谱预处理的计算需求。

4 结 论

针对太赫兹光谱中非线性基线漂移导致定量预测精度下降的问题,本文提出了一种基于自适应动态窗口迭代的基线校正算法。该方法结合SG平滑滤波、基于梯度统计特征的动态起峰检测、多锚点跨区域桥接拟合以及迭代较小值替换机制,实现了复杂背景下太赫兹光谱基线的精准估计与特征峰的高保真保留。以谷氨酰胺、组氨酸、苏氨酸及其二元混合物为研究对象的实验结果表明,校正后的光谱能够保留混合体系中的峰位偏移与重叠吸收特征,并显著提升定量模型的预测性能。其中,LSSVM模型表现最佳,R²达到0.996 0, RMSEP降至0.543 1%, RPD提升至15.72, MAE降至0.427 1%。综上,在本文氨基酸单质及二元混合物压片样品体系和相应评价指标下,所建方法较ModPoly、AsLS和airPLS等常用基线校正方法表现出更优的定量预测性能,说明该算法在当前实验条件下具有较好的应用潜力。但其在更多类型太赫兹光谱体系及复杂样品基质中的适用性仍有待进一步验证。

参考文献:

- [1] Ma H, Huang L L, Duan X, Li X T, Li B, Chen X. *Food Mach.* (马皓, 黄略略, 段续, 李潇婷, 李彬, 陈轩. 食品与机械), **2025**, 41 (2): 197–205.
- [2] Xu T, Lu X X, Shen Y, Zhang J B, Chang T Y, Cui H L, Zhang J. J. *Infrared Millim. Waves* (徐拓, 路星星, 沈雁, 张金波, 常天英, 崔洪亮, 张瑾. 红外与毫米波学报), **2023**, 42 (2): 201–207.
- [3] Zhang Z X, Zhang D L, Tian S J, Fang S Y, Zhao Y R, Zhao J, Hu J. *Spectrosc. Spectral Anal.* (张仲雄, 张东莉, 田世杰, 方世言, 赵艳茹, 赵娟, 胡瑾. 光谱学与光谱分析), **2021**, 41 (5): 1379–1386.
- [4] Qian H D, Li X X, Zhang Z W, Li L D, Guo R, Zhang Y Z, Deng H, Liu Q C. *Chin. J. Lasers* (钱浩东, 李晓霞, 张正文, 李领弟, 郭睿, 张宇, 邓琥, 刘泉澄. 中国激光), **2026**, 53 (2): 0211004.
- [5] Wu J, Wu X, Huang X H, Li J W, Zhang J J, Peng Y, Lu L. *Chin. J. Lasers* (吴静, 吴旭, 黄星皓, 李嘉伟, 张金晶, 彭滢, 陆林. 中国激光), **2024**, 51 (3): 0307206.
- [6] Tian C F, Li J J, Weng G J, Zhu J, Zhao J W. *Spectrosc. Spectral Anal.* (田超凡, 李剑君, 翁国军, 朱键, 赵军武. 光谱学与光谱分析), **2024**, 44 (4): 1073–1080.

- [7] Bai Y F, Luo H Y, Li Z W, Ding Y, Xiong W. *Acta Opt. Sin.* (白云飞, 罗海燕, 李志伟, 丁毅, 熊伟. 光学学报), **2024**, 44 (7): 0730001.
- [8] Liu L, Fan X G, Kang Z M, Wu Y, Wang X. *Spectrosc. Spectral Anal.* (刘龙, 范贤光, 康哲铭, 吴怡, 王昕. 光谱学与光谱分析), **2021**, 41 (1): 111–115.
- [9] Aradhye P D, Mandal S, Gray R D, Campbell C J. *Anal. Chem.*, **2025**, 97 (48): 26708–26719.
- [10] Yan B, Li Z B, Zhu X G, Liang H B, Ni P B. *J. Electron. Meas. Instrum.* (闫碧, 李忠兵, 朱迅果, 梁海波, 倪朋勃. 电子测量与仪器学报), **2025**, 39 (8): 230–240.
- [11] Hu Y H, Cao Z, Fu H J, Dai J S. *Spectrosc. Spectral Anal.* (胡颖惠, 曹政, 傅海军, 戴继生. 光谱学与光谱分析), **2025**, 45 (2): 351–357.
- [12] Hu Y G. *Research on Qualitative Analysis Algorithm Based on Fourier Transform Infrared Spectroscopy*. Yichang: China Three Gorges University (胡怡格. 基于傅里叶红外光谱的定性分析算法研究. 宜昌: 三峡大学), **2025**.
- [13] Lieber C A, Mahadevan-Jansen A. *Appl. Spectrosc.*, **2003**, 57 (11): 1363–1367.
- [14] Duvillaret L, Garet F, Coutaz J L. *Appl. Opt.*, **1999**, 38 (2): 409–415.
- [15] Dorney T D, Baraniuk R G, Mittleman D M. *J. Opt. Soc. Am. A*, **2001**, 18 (7): 1562–1571.
- [16] Yan F, Zhang Z H, Zhao X Y, Su H X, Li Z, Zhang H. *Spectrosc. Spectral Anal.* (燕芳, 张朝晖, 赵小燕, 苏海霞, 李智, 张寒. 光谱学与光谱分析), **2014**, 34 (6): 1473–1476.
- [17] Zeng R, Bian Y, Zhang X, Zhu Z, Yang B. *Spectroscopy*, **2022**, 37 (8): 36–44.
- [18] Li W W, Yan F, Liu Y S. *Spectrosc. Spectral Anal.* (李文文, 燕芳, 刘洋硕. 光谱学与光谱分析), **2025**, 45 (7): 1848–1856.
- [19] Bellon-Maurel V, Fernandez-Ahumada E, Palagos B, Roger J M, McBratney A B. *TrAC Trends Anal. Chem.*, **2010**, 29 (9): 1073–1081.
- [20] Savitzky A, Golay M J E. *Anal. Chem.*, **1964**, 36 (8): 1627–1639.
- [21] Yan F, Li W, Wang Z C. *Spectrosc. Spectral Anal.* (燕芳, 李伟, 王志春. 光谱学与光谱分析), **2020**, 40 (2): 397–402.
- [22] Tan L X, Chen M, Chen S W, Pan X X, Chao S S, Li J C, Liu R F, Yu H, Wang Y X, Wang P Y. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.*, **2025**, 343: 126491.
- [23] Liu Y S, Yan F, Li W W, Zhao M Y. *J. Instrum. Anal.* (刘洋硕, 燕芳, 李文文, 赵渺钰. 分析测试学报), **2024**, 43 (12): 1883–1892.
- [24] Suykens J A K, Vandewalle J. *Neural Process. Lett.*, **1999**, 9 (3): 293–300.
- [25] Eilers P H C, Boelens H F M. *Leiden University Medical Centre Report*, **2005**, 1 (1): 5.
- [26] Zhang Z M, Chen S, Liang Y Z. *Analyst*, **2010**, 135 (5): 1138–1146.

(责任编辑：盛文彦)