

高温超导 YBCO 薄膜制备与结构表征数据的标准化收集与治理方法

崔恩雨^{1)3)4) #} 杨立红^{2)†} 房俊^{3)4)†} 王宇彤²⁾ 张济全²⁾ 欧阳琛³⁾⁴⁾
邵嘉琪⁵⁾ 袁洁¹⁾²⁾

- 1) (中国科学院物理研究所, 凝聚态物质科学数据中心, 北京 100190)
- 2) (中国科学院物理研究所, 超导国家重点实验室, 北京 100190)
- 3) (北方工业大学, 人工智能与计算机学院, 北京 100144)
- 4) (北方工业大学, AI+领域应用关键技术北京市重点实验室, 北京 100144)
- 5) (北京建筑大学, 机电与车辆工程学院, 北京 102616)

摘要

高温超导钇钡铜氧 (YBa₂Cu₃O_{7-δ}, YBCO) 薄膜是高温超导电子学领域的战略级材料, 其微观结构和超导性能强烈依赖于制备工艺、氧含量及缺陷状态等因素。当前, YBCO 薄膜研发与规模化制备过程中仍普遍存在实验数据标准缺失、过程信息采集不完整、多源异构数据难以融合、样品与数据全链路追溯不足等问题, 制约了高质量工艺数据库构建及数据驱动优化方法的应用。针对上述瓶颈, 本文面向器件应用需求, 提出了一套覆盖“薄膜制备—结构表征—数据治理—电子实验记录归档”的 YBCO 薄膜全流程标准化数据收集与治理方法。该方法在脉冲激光沉积 (pulsed laser deposition, PLD) 生长环节构建多模态数据采集架构, 其中温度和真空度由两路独立相机进行连续采集, 并通过时间窗口匹配与线性插值重构实现时间对齐; 激光羽辉图像则利用激光器输出的 TTL 信号进行硬件触发采集, 实现羽辉图像与激光脉冲事件的同步记录; 在表征环节集成机器人送样、智能任务调度与多模式 XRD 自动测试流程, 实现样品流转、测试参数、衍射原始数据及样品标识的自动绑定。进一步地, 本文构建 Raw-Processed-Merged 三层数据存储与治理架构, 在完整保留原始实验证据的基础上, 通过光学字符识别 (optical character recognition, OCR) 后处理、物理规则校验、异常筛查、时间窗口匹配与线性插值重构等策略, 实现多源异步数据的清洗、标准化与统一时间轴融合。在此基础上, 本文构建了面向 YBCO 薄膜制备与表征的多模态实验数据库, 统一组织实验元数据、温度与真空度时序数据、羽辉图像、XRD 表征文件及 ELN 记录。依托结构化元数据模型和电子实验记录本 (electronic laboratory notebook, ELN) 接口, 将过程时序数据、表征文

件、多媒体附件及追溯标识封装为可检索、可审计的实验记录，形成从分析结果反向追踪至原始视频、OCR 识别结果和实验操作过程的闭环溯源体系。该体系将传统依赖人工经验和离散记录的实验流程转化为标准化、结构化、机器可读的数据资产，为开展 YBCO 薄膜工艺—结构关系研究、构建高保真实验数据库，并为后续进一步融合超导性能数据及开展数据驱动研究提供数据基础。本文数据集可在科学数据银行数据库 <https://doi.org/10.57760/sciencedb.j00213.00081> 中访问获取。（数据集私有访问链接：<https://www.scidb.cn/s/F7rMJb>）

关键词:YBCO 薄膜, XRD 自动化表征, 数据治理, 电子实验记录本(ELN)

PACS: 74.72.-h, 74.78.-w, 07.05.Kf, 81.70.-q

基金: 国家重点基础研究发展计划项目(批准号: 2022YFA1603903), 国家自然科学基金(批准号: U23A6015), 中国科学院基础研究青年团队(批准号: 2022YSBR-048)

同等贡献作者.

† 通信作者. E-mail: lihongyang@iphy.ac.cn; fangjun@ncut.edu.cn

第一作者. E-mail: 17812030907@163.com

1 引言

钇钡铜氧 ($\text{YBa}_2\text{Cu}_3\text{O}_{7-\delta}$, 简称 YBCO) 是典型的铜氧化物高温超导材料, 具有较高的超导转变温度 (T_c)、临界电流密度 (J_c)、较低的微波表面电阻 (R_s) 以及灵活的制备方式, 是高温超导薄膜物理研究和电子学器件应用的重要载体。随着薄膜制备技术的持续进步, YBCO 薄膜在超导转变温度、临界电流密度、表面平整度、大面积制备能力和均匀性等方面不断提升, 推动了其在微波滤波、探测、约瑟夫森结及高温超导电子学器件中的应用发展。然而, 面向器件制备需求, YBCO 薄膜仍需在结构质量、表面状态、氧掺杂均匀性、批次一致性和长期稳定性等方面实现更高水平的可控。尽管高质量 YBCO 薄膜制备技术已取得显著进展, 但在规模化研发和制备过程中稳定复现优异性能仍是制约其进一步应用的重要问题^[1-3]。

YBCO 薄膜的微观结构及最终超导性能受到制备条件、氧含量、缺陷分布和

界面状态等多因素共同影响^[4]。在脉冲激光沉积（pulsed laser deposition, PLD）制备过程中，靶材、沉积温度、背景氧压、激光能量密度、靶基距、退火气氛和降温曲线等参数会共同决定薄膜的晶体取向、结晶质量、氧掺杂水平、表面形貌和超导性能^[5-6]。这种对多参数强耦合效应的依赖，决定了传统依赖人工记录、试错法主导的研发模式已无法满足现代器件应用对高效率、低成本及高一致性的需求。

伴随材料基因组计划（Materials Genome Initiative）的推进，数据驱动的材料研发已成为国际共识^[7]。然而，在当前 YBCO 薄膜研发过程中，仍普遍存在实验数据标准不统一、过程信息采集不完整、多源异构数据难以融合，以及样品—工艺—表征—结果之间关联不足等问题，导致大量实验数据难以转化为可被机器学习算法直接利用的结构化数据，严重制约了数据检索与复用^[8-9]。具体而言，在 PLD 生长环节，温度、真空度、羽辉状态等动态过程信息往往缺乏连续化、同步化记录；在结构表征环节，XRD 峰位、峰宽、摇摆曲线等测试结果及其原始文件也常未与样品标识、测试参数自动绑定。上述问题造成了全生命周期数据链断裂，制约了高质量工艺数据库的构建，也限制了机器学习模型训练与逆向工艺设计方法在 YBCO 薄膜研发中的应用^[10]。

针对上述问题，本文面向器件应用场景，提出一套 YBCO 薄膜标准化数据收集与治理方法。该方法以 PLD 薄膜生长和 XRD 结构表征为核心环节，构建覆盖现场感知、原始数据留存、数据清洗、时间对齐、结构化封装和电子实验记录本入库的全流程数据治理体系。本文的主要工作包括：

第一，在薄膜制备过程中，建立多模态数据采集与时间关联架构，其中温度和真空度由两路独立相机连续采集，并通过软件进行时间对齐；激光羽辉图像利用激光器 TTL 信号进行硬件触发采集，实现羽辉图像与激光脉冲事件的同步记录；

第二，构建机器人辅助的 XRD 自动测试与样品流转机制，实现样品标识、测试任务、测量参数和原始数据文件的自动关联；

第三，提出 Raw-Processed-Merged 三层数据模型与多源异步数据时间对齐策略，并结合 ELN 结构化记录实现实验全生命周期追溯^[23]。在此基础上，构建面向 YBCO 薄膜制备与表征的多模态实验数据库，统一组织实验元数据、过程时序数据、羽辉图像、XRD 表征文件及 ELN 记录。该方法旨在将非结构化、离散化和经验化的实验信息转化为标准化、机器可读、可审计和可复用的数据资产，为开展 YBCO 薄膜工艺—结构关系研究，并为后续进一步融合超导性能数据和开展数据驱动分析提供基础。

第四，基于真实 YBCO 薄膜 PLD 生长实验开展端到端验证与系统性性能评估。以一次完整生长实验为案例，展示从多模态数据采集、OCR 识别与异常修正、时间对齐、XRD 自动化表征到 ELN 入库的全链路数据流转过程；在性能评估方面，定量验证 OCR 后处理准确率（61.8%→98.6%）、多源数据时间对齐精度（匹配率 99.99%，平均偏差 0.041 s）、机器人辅助 XRD 批量测试效率（人工作业时间减少 93%–97%）及 ELN 入库稳定性（重试后成功率 100%）；此外，以 15 批 YBCO 薄膜实验对工艺特征与 XRD(005)峰位进行初步相关性分析，验证所构建数据体系对工艺—结构探索性关联分析的支撑能力。

2 基于多模态感知的实验数据自动化采集体系

为实现对 YBCO 生长过程全维度的数字化映射，本研究设计了如图1所示的多模态数据采集架构。该架构围绕关键过程参数视觉化记录、羽辉瞬态图像同步采集、XRD自动化测试数据采集以及结构化元数据辅助录入四类通道展开，实现了生长过程参数、瞬态现象、结构表征结果与实验背景信息的关联融合。



图 1 YBCO 薄膜生长多模态数据采集架构

Fig. 1. Schematic architecture of the multimodal data acquisition system for YBCO thin-film growth.

2.1 关键过程参数的可视化时序采集

PLD生长过程中的腔体温度与真空度是决定薄膜结晶质量的核心连续变量。

不同于离散的手动记录，本系统采用高精度外置相机阵列（4K分辨率，6-12mm 变焦镜头）对设备的状态进行实时监控。通过固定支架锁定视场，并利用感兴趣区域（Region Of Interest, ROI）技术聚焦显示屏数字区域，系统可实时捕捉温度与真空度的动态变化^[11]。



图 2 温度/真空度视觉采集装置与 ROI 区域示意图

Fig. 2. Visual acquisition setup and regions of interest for temperature and chamber-pressure readings.

2.2 激光羽辉瞬态图像同步触发采集

激光轰击靶材产生的瞬态羽辉演化是脉冲激光沉积过程中物质输运与薄膜成核生长的直观表征。该过程由纳秒级激光脉冲触发，随后羽辉在微秒量级内发生快速膨胀、传播与衰减，羽辉发光与激光脉冲具有严格的时间关联。随着沉积时间延长，靶材表面烧蚀坑逐渐加深，羽辉的亮度、形态、高度及空间分布可能发生变化，进而影响到达基片表面的粒子通量和能量分布，对薄膜初始成核及后续生长阶段的生长速率、均匀性和结晶质量产生影响。然而，普通连续拍摄难以准确捕捉羽辉的有效演化时刻，因此本系统引入硬件触发机制：利用激光器输出的TTL 信号作为触发源，精确控制工业相机的曝光时刻，实现羽辉图像与激光脉冲事件的一一对应^[12]。该机制将肉眼难以分辨的瞬态羽辉过程转化为可记录、可追溯和可量化分析的图像序列，为后续分析激光能量密度、靶材烧蚀状态与薄膜生长行为之间的关联提供原始数据支撑。

2.3 机器人辅助的薄膜XRD自动化测试与数据采集

薄膜生长完成后，XRD表征是连接制备工艺与结构质量评价的关键环节。本研究主要采用日本Rigaku公司SmartLab型X射线衍射仪（9 kW）对YBCO薄膜进行结构表征，该设备配备Ge(220×2)双晶单色器、微区扫描部件，可实现薄膜的高分辨测试和局部区域的精细化扫描。

传统测试模式中，样品转移、参数配置、测试执行和数据文件整理等环节高度依赖人工操作，不仅测试时间成本较高，而且测试条件、样品状态、测试顺序及原始数据文件等信息难以实现全过程记录与追溯，且由于人工参与会引入较多人为误差。随着器件级应用样品量的快速增长，传统方式在效率、准确性和可追溯性方面的不足日益凸显。特别是在超导体材料研究中，YBCO 薄膜的晶体结构

测试依赖高分辨多轴测角仪实现晶面精确定位，对测试精度要求极高。然而，手动操作往往存在样品定位偏差，且样品调整及衍射曲线判读过程高度依赖人工，容易引入人为误差。为构建YBCO薄膜从生长过程到结构表征的完整数据链，本研究设计了一套机器人辅助的薄膜XRD自动化测试流程，将样品流转、测试任务调度及衍射数据采集进行统一关联与管理。

系统整体测试流程如图3所示。实验人员仅需将待测薄膜样品存放至智能样品柜，并录入样品编号、测试类型与定制化测试参数，常规测试场景可直接调用系统预设标准参数模板。智能调度系统根据样品存放位置与测试队列优先级自动生成、排序测试任务。测试启动后，机器人自动完成样品抓取与转运，将待测样品精准输送至XRD测试样品台。样品就位后，自动化测试系统接管全套表征流程，独立完成测试作业。



图 3 机器人辅助的薄膜 XRD 自动化测试与数据关联流程图

Fig. 3. Workflow for robot-assisted automated XRD characterization and data association of thin-films

由于商用XRD设备原厂控制系统通常未开放外部二次开发与远程调用接口，本研究未对原厂软件进行底层修改，而是将人工测试的标准化操作拆解为可复用、可自动化执行的基础测量单元，包括样品调整（Sample alignment）、位移调控（Move）、测试（Measurement）等核心模块。根据不同测试需求组合适配为标

准化测试方案，可覆盖薄膜多维度精细结构表征，包含用于物相鉴定的 2θ - θ 扫描、用于表征晶体缺陷与结晶度的摇摆曲线（Rocking curve）、用于测试薄膜厚度与密度的反射率测试（Reflectivity）以及用于解析薄膜应力状态与精细物相结构的二维 Reciprocal space mapping（Wide-range RSM）测试。

自动化任务执行过程中，系统可模拟标准化人工操作逻辑，自动完成测试流程调用、扫描角度区间、步长、扫描速率及数据命名等参数的批量配置；同时自动完成样品高度、倾角与中心位置的精准校正，保障衍射信号采集处于最优测试状态。参数配置与样品对位完成后，系统自动启动扫描采集，并在测试过程中实时监测设备界面状态与异常提示，实现全过程闭环监控。正式扫描前，系统可自动完成基线校验与状态自检，规避因设备状态异常或样品对位偏差导致的无效测试。

测试正常完成后，系统自动将原始XRD衍射图谱、全套测试参数与对应样品编号进行绑定归档；若测试过程出现设备异常、信号异常或对位失效等问题，系统可自动捕捉异常信息、生成日志记录并终止当前任务，杜绝无效数据录入与数据丢失。测试结束后，机器人自动将样品取回并归位至智能样品柜，实现样品全流程无人化流转。实验人员可后续统一取回样品，所有测试原始数据、流程参数与样品信息已完成自动关联归档，全程可追溯、可复盘。

此外，该自动化系统支持通过MES/SECS-GEM工业接口与生产线管控系统对接，可适配万片/年级检测需求，能够有效支撑规模化薄膜材料的制备、表征与工业化研发生产。

2.4 结构化元数据辅助录入

除过程数据外，样品信息（衬底类型、基底尺寸）、靶材参数及退火条件等

静态元数据同样需要标准化记录。为避免重复录入导致的格式不统一，设计了“历史参数记忆库”功能，支持实验人员调用历史参数模板进行快速调用与修正。这一设计有效解决了不同批次实验参数命名规范不一致的问题，确保了元数据的标准化与可追溯性。

3 多源异构数据的清洗与时间对齐治理

原始采集数据往往存在噪声大、格式杂、时间不同步等问题，直接的聚合无法支撑分析需求^[13]。

本研究提出了“三层存储 + 双轨对齐”的数据治理策略，整体处理流程见图4，三层存储数据模型见图5，确保数据在保留原始证据的同时具备高质量的分析属性。



图 4 温度与真空度数据清洗与对齐流程图

Fig. 4. Workflow for cleaning and aligning temperature and vacuum data

3.1 基于OCR的智能识别、后处理与异常筛查

针对相机捕获的温度与真空度图像，如图4所示，系统采用PaddleOCR深度学习模型进行字符识别^[14-15]。考虑到实际采集场景中可能存在反光、倾斜、运动模糊及字体变形等情况，OCR识别结果并不直接用于后续分析，而是依次经过字符

纠错、物理范围校验和异常筛查去噪。首先，通过规则引擎对OCR原始字符串进行字符纠错和格式规范化，包括小数点修正、单位字符处理及科学计数法重构等，并将规范化后的字符串转换为数值。对于能够成功解析的数据，进一步依据仪器可显示的数值范围及实验物理约束进行范围校验，包括有限数值判断、温度和真空度有效测量范围检查以及科学计数法指数完整性检查等。无法解析或明显违反约束的数据被标记为无效值。

为降低OCR瞬时误识别等异常噪声对后续数据分析的影响，对于通过物理范围校验但仍可能存在异常识别的数据，进一步进行局部时序一致性的异常筛查。温度数据直接在原始数值尺度下进行比较；考虑到真空度可能跨越多个数量级，首先进行对数变换：

$$p_i' = \log_{10}(p_i) , \quad (1)$$

其中， p_i 为第*i*个真空度读数。以当前采样点及其前后各5个有效采样点组成长度为11的局部窗口，计算窗口中位数 M_i 及中位绝对偏差（median absolute deviation, MAD）：

$$\text{MAD}_i = \text{median}(|x_j - M_i|) , \quad (2)$$

其中， x_i 为当前数据点， x_j 表示局部窗口内的温度值或经对数变换后的真空度值。为量化当前数据点相对于局部趋势的偏离程度，定义稳健标准化偏离指标：

$$S_i = \frac{|x_i - M_i|}{\max(1.4826\text{MAD}_i, \delta_x)} , \quad (3)$$

其中，系数1.4826用于在近似正态分布条件下将MAD转换为与标准差相当的稳健尺度估计； δ_x 为最小尺度参数，用于避免稳定平台阶段MAD接近零时造成异常度量失稳。本文温度数据取 $\delta_T = 0.1^\circ\text{C}$ ，真空度对数数据取 $\delta_p = 0.05$ 。采用中位数和MAD而非均值和标准差，可降低单个OCR误识别值对局部统计量的影响。

为避免将正常的小幅波动误判为异常，本文同时考虑稳健偏离程度和实际变

化幅度。对于温度数据，当 $S_i > 5$ 且 $|T_i - M_i| \geq 1.0^\circ\text{C}$ 时，将该点标记为候选异常点。

对于真空度数据，当 $S_i > 5$ 且 $|\log_{10}(p_i) - M_i| \geq 0.5$ 时，将该点标记为候选异常点。

候选异常点进一步结合变化持续性和OCR识别置信度进行判定。对于每个候选点，检查其后3个有效采样点：若其中至少2个数据点延续相同变化方向或保持新的数值水平，则认为该变化具有持续性，并作为真实工艺变化保留；若较大偏离仅出现在单个采样点，随后迅速恢复至原有水平，则优先判定为OCR瞬时误识别。对于升温、充氧和退火等动态工艺阶段，只要连续多个采样点支持相同变化趋势，即不作为异常剔除。OCR识别置信度仅作为辅助判据，不单独用于删除数据。

因此，本文的异常筛查主要针对OCR产生的孤立突跳误识别。最终综合物理范围约束、局部统计偏离程度、实际变化幅度、变化持续性和OCR置信度，将数据划分为有效值、候选异常值和无效值。有效值进入Processed层；候选异常值保留异常标记及处理日志以供人工复核；无法解析或明显违反物理约束的数据在Processed层中置为空值。所有原始识别结果均保留在Raw层，以保证异常处理过程可追溯和可审计。

3.2 “Raw-Processed-Merged” 三层数据分层策略

为保障数据的可追溯性与可用性，如图5所示本研究建立了三层数据模型。兼顾了原始数据留存、数据质量提升与多源数据融合等需求^[16-17]。

- **Raw 层**：原始数据留存层，用于完整保存 OCR 系统输出的原始识别结果、图像来源信息及毫秒级时间戳。该层不对识别内容进行人为修正，即使其中包含识别错误、缺失值或异常字符，也作为实验过程的原始“数字证据”永久保留，用于后续异常复核、数据追溯和结果审计。

- **Processed 层**：清洗与标准化数据层，在 Raw 层基础上对原始识别结果进

行数值提取、单位统一、异常值筛查、去噪处理和格式规范化，形成可靠的温度、真空度等有效过程数据集。该层数据具有更高的可读性和稳定性，可用于实验过程的实时监控、趋势分析以及不同实验批次之间的初步比较。

• **Merged 层：**多源融合与建模数据层，进一步针对温度、真空度等过程变量采样频率不一致的问题，采用时间窗口合并、线性插值重构等方式，将不同来源的数据融合到统一时间轴上，形成结构化的过程特征表。该层数据可直接服务于后续统计分析、机器学习建模以及工艺参数与薄膜结构和性能之间的关联研究。



图 5 Raw-Processed-Merged 三层数据关联图谱

Fig. 5. Raw-Processed-Merged three-layer data association graph

3.3 异源数据的时间对齐算法

由于双路相机独立工作，原始数据的采样时刻并不完全一致。为解决这一多源异步采样导致的时序不一致问题^[13]，本研究在数据处理阶段建立统一时间基准，并开发了时间窗口合并与线性插值重构两种时间对齐策略：

1. **时间窗口合并：**该方法以某一通道的时间戳 $t_i^{(A)}$ 为基准，在预设容差窗口 Δt 内搜索另一通道中时间最接近的读数。若满足时间差小于容差阈值，则将两组

读数合并为同一时刻的联合记录。若最近邻采样点的时间差超过预设容差窗口 Δt ，则认为该数据点在另一通道中不存在有效匹配对象，不生成联合记录。其匹配关系可表示为：

$$j^* = \arg \min_j |t_i^{(A)} - t_j^{(B)}|, \quad |t_i^{(A)} - t_{j^*}^{(B)}| \leq \Delta t, \quad (4)$$

其中， $t_i^{(A)}$ 表示基准通道第 i 个数据点的采样时刻， $t_j^{(B)}$ 表示待匹配通道第 j 个数据点的采样时刻， Δt 为允许的最大时间偏差。该方法不改变原始数据值，能够最大程度保留真实采样结果，适用于温度、真空度等变化相对平缓的工艺阶段。

2. 线性插值重构：对于需要形成连续高密度数据集的应用场景，本研究进一步构建统一时间轴 T_k ，并根据相邻两个真实采样点对目标时刻的数据进行线性估计。对于任一过程参数 x ，当 $t_i \leq T_k \leq t_{i+1}$ 时，其插值结果为：

$$x(T_k) = x(t_i) + \frac{x(t_{i+1}) - x(t_i)}{t_{i+1} - t_i} (T_k - t_i), \quad t_i \leq T_k \leq t_{i+1}, \quad (5)$$

其中， T_k 为统一时间轴上的目标时刻， $x(t_i)$ 和 $x(t_{i+1})$ 分别为该时刻前后相邻采样点的真实读数。该方法可在不显著改变整体变化趋势的前提下补齐缺失时刻的数据，从而生成采样间隔一致的连续数据序列，便于后续趋势分析、过程监控、特征提取和机器学习模型训练。

通过上述处理，系统能够将原本采样频率不同、时间点不完全一致的温度、真空度等数据统一到同一时间轴下，降低多源数据关联分析中的时间偏差。同时，根据不同应用需求，系统可生成多粒度数据集，例如 0.5 s 高分辨率数据用于过程细节分析，10 s 摘要数据用于电子实验记录本展示、长期趋势回顾和数据归档。

4 面向ELN的结构化数据封装与入库

仅有数值数据不足以支撑器件应用的深度分析，必须将过程数据与实验元数据、附件数据进行有机融合，形成完整的“实验故事链”。本研究设计了一套符

合JSON Schema标准的结构化记录模型，并通过ELN接口实现自动化归档^[18-20]。

4.1 多维数据结构化定义

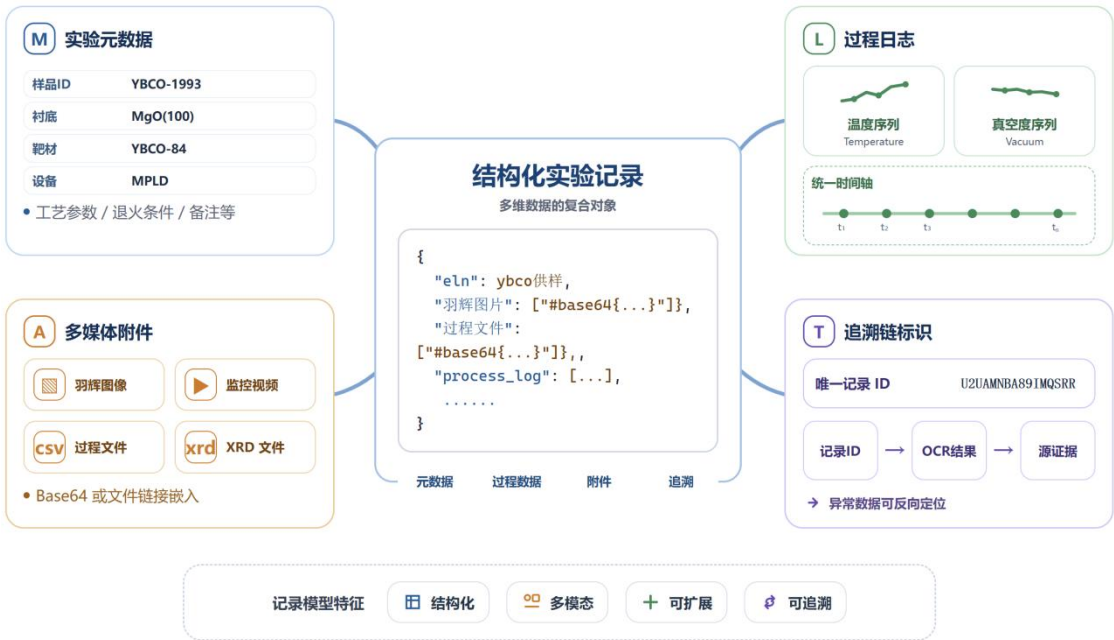


图 6 结构化实验记录模型示意图

Fig. 6. Schematic of the structured experimental record model

每一条目实验记录被定义为包含以下核心模块的复合对象：

- **实验元数据**：固定字段（衬底、靶材、设备编号）与可变字段（靶机距，退火温度等）。
- **过程日志**：经时间对齐后的温度/真空度序列数据。
- **多媒体附件**：羽辉原始图像序列、全过程监控视频、Raw/Processed CSV 文件、XRD 测试文件等，均通过 Base64 编码或直接链接形式嵌入记录。
- **追溯链标识**：每条记录均生成唯一 ID，关联至原始视频片段与 OCR 置信度评分。

4.2 ELN入库与备份机制

为应对海量数据上传可能导致的系统卡顿，系统采用分片上传策略：基础信息先行上传，随后按批次写入过程日志。同时，建立本地-云端双重备份机制，无论上传成功与否，均在本地保留完整数据包与上传状态日志（Success/Fail），确保数据资产的安全性与可追溯性^[21]。

5 实验数据库内容、典型数据与全生命周期追溯

基于上述多模态实验数据采集、清洗、时间对齐和结构化封装方法，本文构建了面向YBCO薄膜制备与结构表征的实验数据库。数据库以单次薄膜制备实验为基本组织单元，以样品编号作为主要关联标识，统一保存实验元数据、温度与真空度过程数据、激光羽辉图像、XRD结构表征文件及ELN实验记录。

本节首先介绍数据库的数据组成、组织结构和数据规模，随后以编号为2002的YBCO薄膜PLD生长实验作为端到端案例，展示从温度与真空度视频采集、OCR识别、异常清洗、时间对齐、羽辉图像关联、XRD自动化表征到ELN入库的完整数据流转过程。该案例的主要目的在于验证数据完整性、处理可复核性和全生命周期追溯能力。

5.1 数据库总体内容与组织方式

YBCO薄膜实验数据库覆盖薄膜制备、过程监测、结构表征、数据治理和实验记录归档等环节。数据库以单次PLD生长实验为基本数据单元，每次实验对应一条结构化实验记录，并通过样品编号、采集时间戳和实验编号关联不同来源的数据。

数据库主要包括六类数据。第一类为实验元数据，包括样品编号、实验编号、衬底类型、样品尺寸、靶材信息、设备编号、靶基距、沉积时间及退火条件等；

第二类为过程时序数据，包括温度和真空度的原始监控视频、OCR识别字符串、识别置信度、毫秒级时间戳及清洗后的标准化数值；第三类为羽辉图像数据，包括与激光触发事件关联的羽辉图像序列、采集时间等；第四类为XRD表征数据，包括 2θ - θ 扫描、摇摆曲线等原始测试文件和测试参数；第五类为数据治理结果，包括Raw、Processed和Merged三层数据文件和时间对齐日志；第六类为ELN归档数据，包括结构化实验记录、多媒体附件、数据文件链接等。

表 1 YBCO 薄膜实验数据库的主要数据内容

Table 1. Overview of multimodal data contents in the YBCO thin film experimental database.

数据类别	主要数据内容	数据形式
实验元数据	样品、衬底、靶材、设备及工艺条件	JSON、结构化表格
过程时序数据	原始视频、OCR 结果、清洗数值、时序曲线	MP4、CSV、图像
羽辉数据	羽辉瞬态图像、采集时间	JPG、JSON
XRD 数据	2θ - θ 、摇摆曲线等数据文件、测试参数	原始文件(RAS、RASX、TXT)、JSON
数据治理结果	Raw、Processed、Merged 文件及处理日志	CSV、JSON、日志
ELN 记录	元数据、过程曲线、附件及日志	JSON、文件夹、附件链接

截至本文数据统计时,数据库共收录YBCO薄膜实验样品56个,包含温度OCR记录147万条、真空度OCR记录约153万条、羽辉图像约59610帧、XRD测试文件81个,原始监控视频总时长约165小时,数据总容量约72GB。



删除[RainM97]: 本文所有数据集（包括实验元数据、过程时序数据、羽辉数据、XRD数据、数据治理结果及ELN记录等）均以JSON、CSV、MP4、JPG、RAS/RASX/TXT等格式在补充材料中给出。

5.2 实验案例与数据构成

本研究选取编号为2002的YBCO薄膜实验作为端到端验证案例。该薄膜采用PLD方法制备，实验过程中实时采集温度显示视频、真空度显示视频和激光羽辉图像。薄膜生长完成后，由机器人辅助XRD自动测试系统执行 2θ - θ 扫描和摇摆曲线测试，并将测试参数、原始数据文件和样品编号自动绑定。该实验的数据组成如表2所示。

表 2 YBCO-2002 实验的数据组成

Table 2. Data composition of the YBCO-2002 experimental dataset.	
数据项目	数据内容
样品编号	2002
实验编号	PU314NMENNUMK3C3
温度/真空度视频时长	43 分 22 秒
温度 OCR 记录数	9156
真空度 OCR 记录数	9545
羽辉图像数量	1038
XRD 测试模式	2θ - θ 扫描、摇摆曲线
ELN 入库状态	Success

5.3 OCR数据处理与异常修正

温度和真空度通过OCR实时提取显示屏读数。Raw层完整保存OCR原始字符串、识别置信度、图像帧编号和毫秒级时间戳。Processed层进一步完成字符纠错、

数值解析、科学计数法重构、物理范围校验和异常筛查。

图7给出了该实验中温度和真空度读数识别的典型示例。图7(a–d)展示了温度的识别结果。对于温度单位符号，Raw层允许出现“C”“°C”或单位缺失等形式，Processed层会在修正时自动补齐。图7(a)修正单位后为592.2 °C，是稳定显示状态下的正确识别结果。图7(b)和图7(c)展示了两类可被自动剔除的错误：在读数快速刷新或显示不清时，OCR分别输出68.9和38.3C，二者均显著偏离实验的合理温度范围，因此不会进入Processed层。相比之下，图7(d)中的587.2C虽然满足温度范围约束并被保留，但原始图像表明该帧处于显示屏快速变化过程的瞬时误读。

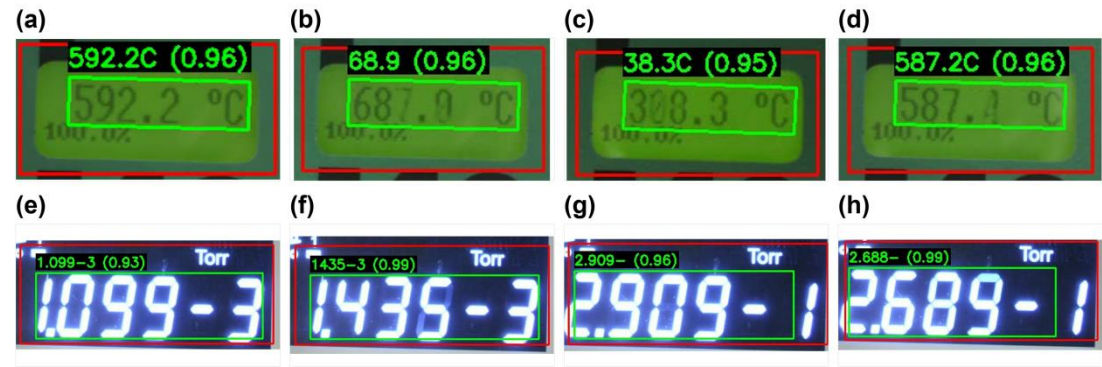


图 7 OCR 读数示例。(a) 592.2C；(b) 68.9；(c) 38.3C；(d) 587.2C；
(e) 1.099-3；(f) 1435-3；(g) 2.909-；(h) 2.688-

Fig. 7. OCR readout examples. (a) 592.2C; (b) 68.9; (c) 38.3C; (d) 587.2C;
(e) 1.099-3; (f) 1435-3; (g) 2.909-; (h) 2.688-.

图7(e–h)展示了真空度的识别结果。Processed层会自动进行科学计数法重构。图7(e)为稳定显示下的正确识别结果，OCR输出1.099-3，对应 1.099×10^{-3} Torr。图7(f)中小数点因数码管高亮和局部饱和未被识别，Raw层得到1435-3；Processed层依据真空度显示格式，将其修正为 1.435×10^{-3} Torr。图7(g)中指数位缺失，OCR仅输出2.909-，结合相邻时刻数值连续性和当前实验阶段的压力范围，自动补齐为 2.909×10^{-1} Torr。图7(h)展示了读数快速变化时的残留误差，OCR输出2.688-并通过局部格式校验，但原始图像显示该读数为瞬时误读。

上述结果表明，Processed层能够显著降低OCR识别格式和噪声对实验日志的影响，同时Raw层图像和字符串保留了必要的人工追溯依据。

5.4 温度与真空度时间对齐

为展示温度与真空度两路异步采样数据的对齐效果，选取同一实验中的局部片段进行可视化对比。图8(a)和图8(b)采用5 s局部窗口，图8(c)采用60 s局部窗口，主要用于放大展示原始采样点在短时间尺度上的错位现象，以及时间窗口合并和插值轴重构后两路数据在同一时刻上的对应关系。

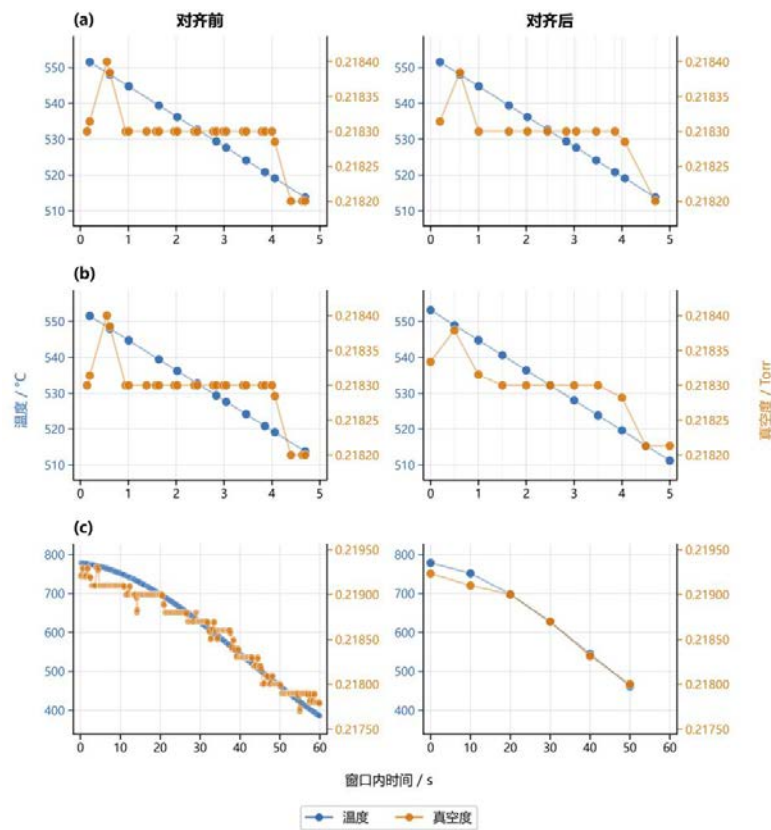


图 8 多源实验数据时间对齐与不同时间粒度重构效果对比。左列为对齐前温度与真空度的异步采样散点，右列为对齐后的联合记录结果。(a) 时间窗口合并结果；(b) 0.5 s 统一时间轴重构结果；(c) 10 s 统一时间轴摘要结果。

Fig. 8. Comparison of temporal alignment and multi-resolution reconstruction of experimental data. The left column shows asynchronous temperature and vacuum sampling points before alignment, while the right column shows the aligned joint records.(a) Time-window merging; (b) reconstruction on a unified 0.5 s time axis;(c) summarization on a unified 10 s time axis.

结果表明，时间窗口合并能够建立温度与真空度之间的近邻对应关系；0.5s 适合保留局部变化细节；10s则在降低数据密度的同时保留主要趋势，可用于实验记录归档和长期过程回顾。

5.5 羽辉采集结果

羽辉序列共包含1038帧JPG图像，分辨率均为 2448×2048 ，采集时间持续约1037 s。相邻帧平均间隔约为1.000 s。如图9所示，系统能够连续记录羽辉区域的发光形貌和亮度分布。典型帧中可见羽辉中心高亮区及外侧红黄色扩散边界，表明采集图像能够反映沉积过程中的羽辉状态。

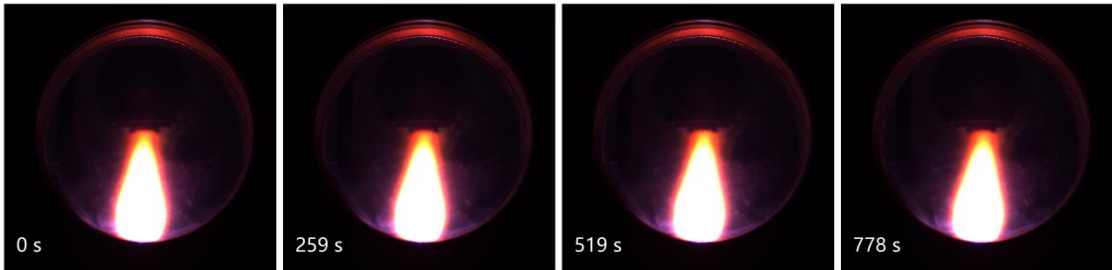


图 9 羽辉图像序列的实际采集结果。四帧图像按采集时间顺序排列，时间标注为相对于首帧的采集时间，分别为 0 s、259 s、519 s 和 778 s。

Fig. 9. Representative plume image sequence acquired during the experiment. The four frames are arranged in chronological order, with time labels given relative to the first frame: 0 s, 259 s, 519 s, and 778 s.

5.6 自动化XRD表征结果

薄膜生长完成后，机器人根据预设任务将样品从智能样品柜转运至XRD测试工位，自动完成样品装载、位置对准、 2θ - θ 扫描和摇摆曲线测试。任务创建时，操作人员仅需在统一界面中填写样品与衬底信息，并确认相应的扫描程序，主要配置内容见表3。其中， 2θ - θ 扫描用于获取较宽角度范围内的衍射信息，以识别薄膜及衬底的主要衍射峰；摇摆曲线测试则选取YBCO（005）反射作为目标峰，用于表征薄膜的结晶质量。在正式摇摆曲线测试前，进行005衍射峰的预扫描，该范围对应YBCO（005）反射附近的摇摆扫描角度区域，能够覆盖其预期峰位，

扫描结束后精确寻峰，并将 $2\theta/\theta$ 移动到衍射峰位置。系统根据预扫描结果自动确定YBCO（005）反射的实际峰中心，随后围绕该峰中心执行精细摇摆曲线扫描。

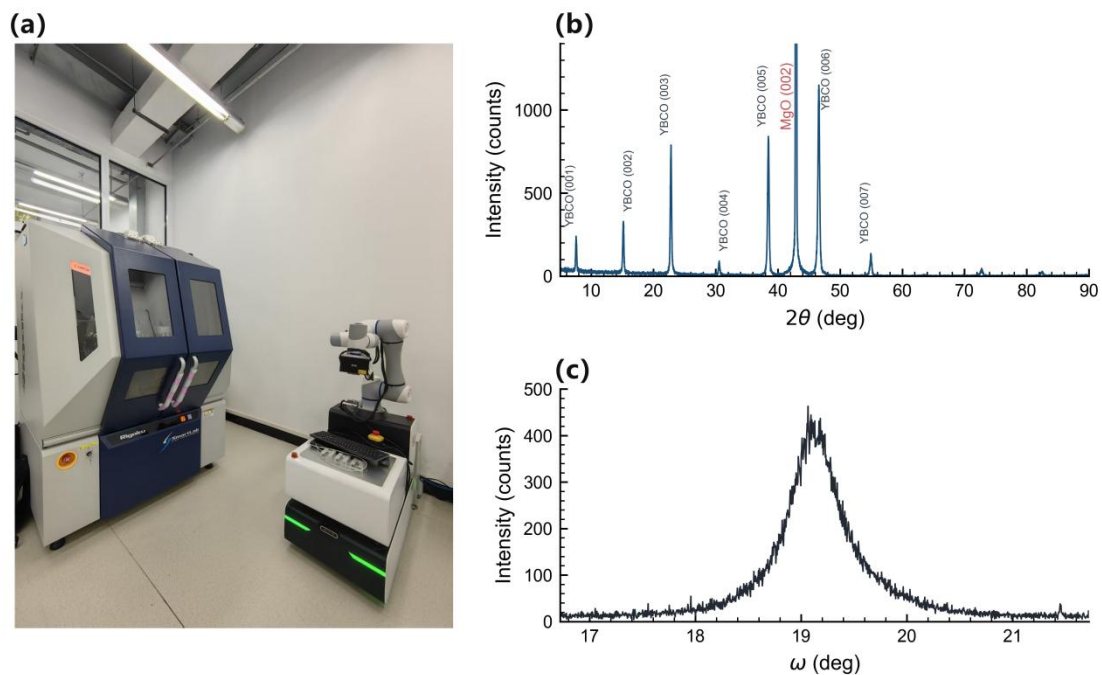


图 10 YBCO#2002 薄膜的自动化 XRD 表征：（a）机器人辅助 XRD 自动测试实景；（b） 2θ - θ 扫描图谱；（c）摇摆曲线结果

Fig. 10. Automated XRD characterization of the YBCO#2002 thin film: (a) robot-assisted automated XRD measurement setup; (b) 2θ - θ diffraction pattern; and (c) rocking curve.

表 3 自动化 XRD 任务所需的人工配置参数

Table 3. Manual configuration parameters required for automated XRD measurements.

配置模块	配置内容
样品信息	实验流程：XRD + Rocking Curve；样品尺寸：10 mm × 10 mm × 0.5 mm
衬底信息	衬底：MgO；表面取向：(001)；
2θ - θ 扫描	范围：5°-90°；步长：0.01°；速度：10°/min
005 衍射峰的	
预扫描	2θ 扫描范围：36°-41°；步长：0.01°；速度：2°/min
摇摆曲线扫描	以预扫描获得的峰中心 ω_0 为中心，扫描范围： ω_0-3° 至 ω_0+3° ； 步长：0.01°；速度：1°/min
数据管理	文件标识：2002；文件名称、保存路径和样品编号由系统自动关联

任务提交后，上述参数由系统自动下发至XRD设备。测试结束后，原始衍射文件、任务参数与样品编号自动绑定，避免了重复参数录入和人工文件命名，实现了机器人转运、自动测试与数据归档的一体化衔接。图10进一步展示了机器人辅助XRD测试实景图及其获得的 2θ - θ 扫描和摇摆曲线结果。

5.7 ELN入库与完整追溯案例



图 11 YBCO 薄膜实验数据的 ELN 结构化归档与完整追溯案例。(a) 实验元数据、羽辉图像、过程曲线、监控视频及 XRD 文件的统一入库；(b) 从 Merged 层经 Processed 层、Raw 层回溯至原始视频画面的数据追溯链。

Fig. 11. Structured ELN archiving and end-to-end traceability of YBCO thin-film experimental data. (a) Unified archiving of experimental metadata, plume-image sequences, process curves, monitoring videos, and XRD files; (b) traceability from the Merged layer through the Processed and Raw layers to the corresponding frame in the original monitoring video.

完成多源数据清洗、时间对齐与结构化封装后，统一写入ELN。如图11(a)所示，样品信息、生长条件、Raw-Processed-Merged各层数据文件及表征结果被集中归档，实现了不同来源实验数据的统一管理。

当过程曲线中出现异常数据点时，研究人员可依据时间戳和唯一标识，从Merged层记录逐级回溯至Processed层清洗结果、Raw层原始OCR数据以及对应的原始视频画面，如图11(b)所示。追溯过程中可以观察原始数值、处理后数值和时间信息，从而支持异常数据复核、处理规则检查及实验过程复现。从而能够实现实验数据从采集、处理、融合到归档的全生命周期追溯。

6 性能评估与实验验证

6.1 OCR识别准确率与后处理

为评估在实际实验采集场景下的OCR识别与后处理效果，从历史采集数据中随机抽取500个数据作为验证样本，其中温度字段250个、真空度字段250个。人工根据原始视频读取对应仪表读数，并将其作为对照真值。随后分别比较Raw原始识别结果和Processed后处理结果与人工真值的一致性。

考虑到实验现场可能存在反光、轻微倾斜、运动模糊和字体变形等情况，采用数值误差评价识别准确性。温度字段以人工读数为真值，误差不超过 $\pm 0.5^{\circ}\text{C}$ 判定为正确；真空度字段由于通常以科学计数法形式显示，要求指数位一致，且相对误差不超过1%判定为正确。若字段无法解析、被错误修正，或虽为合法数值但超过误差阈值，均判定为识别错误。

从表4可以看出，Raw层直接OCR结果在真实实验场景下存在较高的不稳定性。温度字段的Raw准确率为48.4%，如图7所示，主要错误来源包括显示屏刷新过程中的瞬时读数、反光或模糊造成的数字缺失，以及ROI区域选择偏大导致单

位字符“C”被识别。经过后处理后，温度字段正确数由121个提高到247个，准确率由48.4%提高到98.8%。

表 4 OCR 识别与后处理准确率统计

Table 4. Accuracy statistics of OCR recognition before and after post-processing.					
字段	抽样数	Raw 正确数	Raw 准确率	Processed	Processed
				正确数	准确率
温度	250	121	48.4%	247	98.80%
真空度	250	188	75.2%	246	98.40%
合计	500	309	61.8%	493	98.60%

对于真空度字段，仪表通常采用“1.099-3”表示“ 1.099×10^{-3} ”这类简写科学计数法。若将两种表示方式视为数值含义相同，仅比较有效数字和指数是否一致，则Raw层共有188个样本识别正确，准确率为75.2%。Raw层中的主要错误包括小数点缺失、指数位漏识别、数字字符误识别以及读数快速刷新引起的瞬时误读。经过科学计数法重构、格式修正、数值解析和物理规则校验后，真空度字段的正确数提高至246个，准确率达到98.4%。

总体来看，Raw层在500个验证样本中共有309个样本识别正确，整体准确率为61.8%；经过后处理后，Processed层的正确样本数提高至493个，整体准确率达到98.6%。结果表明，仅依赖OCR模型的原始字符输出，容易受到显示格式、图像质量和读数刷新状态等因素的影响；结合字段类型、科学计数法格式、物理范围约束和时间连续性等信息进行后处理，能够显著提高自动记录数据的准确性和可用性。剩余错误主要来自读数快速变化过程中产生的瞬时帧误读，此类结果在数值上可能仍处于合理范围内，因此难以仅依靠单点物理阈值进行识别和剔除。为保证数据的可追溯性，系统完整保留Raw层图像、原始识别字符串和时间戳，

以支持后续对异常数据及边界案例的定位、复核和处理规则审查。

6.2 时间对齐误差

为验证温度与真空度两路独立采集数据的时间对齐效果，并考察时间窗口参数对匹配结果的影响，本文选取5组独立实验数据进行对比验证，分别记为G1~G5。5组实验共包含温度参考点29454个、真空度记录点30607个，两路记录量相差1153个。以两路记录数均值为基准计算，记录量差异比例约为3.84%。所选实验覆盖过程参数相对稳定和动态变化的工艺阶段，以降低单一实验状态对评价结果的影响。

时间窗口匹配以温度记录作为参考，对每个温度采样点独立搜索真空度通道中时间戳最接近的数据点。当两者的绝对时间差不超过预设时间窗口 Δt 时，认为匹配成功并建立联合记录。匹配率定义为成功获得有效近邻的温度参考点数占全部温度参考点数的比例。匹配后的时间误差采用两路记录时间戳的绝对差值进行评价，并统计平均绝对时间差、95%分位绝对时间差和最大绝对时间差。

表 5 不同时间窗口参数下的时间匹配结果

Table 5. Temporal matching results under different time-window settings.

时间窗口/s	成功匹配点数	匹配率/%	平均绝对 时间差/s	95%分位绝对 时间差/s	最大绝对 时间差/s
0.05	22138	75.16	0.033	0.049	0.050
0.10	29325	99.56	0.041	0.072	0.100
0.20	29420	99.88	0.041	0.073	0.199
0.50	29436	99.94	0.041	0.073	0.491
1.00	29451	99.99	0.041	0.073	0.993

为确定合适的时间窗口参数，分别设置 Δt 为0.05、0.10、0.20、0.50和1.00 s，对5组实验数据进行统一测试，结果见表5。

由表5可知，当时间窗口由0.05 s增大至0.10 s时，匹配率由75.16%提高至99.56%；当时间窗口增大至0.20 s时，匹配率进一步达到99.88%。继续将时间窗口增大至0.50 s和1.00 s，匹配率相较于0.20 s仅分别提高0.06和0.11个百分点，但最大绝对时间差分别增大至0.491 s和0.993 s。与此同时，当时间窗口不小于0.10 s时，匹配成功数据的平均绝对时间差和95%分位绝对时间差基本保持稳定。因此，综合考虑匹配完整性和时间对应误差，本文最终选择 $\Delta t=0.20$ s作为时间窗口匹配参数。该参数能够覆盖绝大多数正常采样时刻之间的偏差，同时避免因过度放宽匹配条件而引入时间间隔较大的数据配对。采用 $\Delta t=0.20$ s后，进一步分别统计G1~G5五组实验的时间匹配结果，以验证该参数在不同实验记录中的适用性和稳定性，结果见表6。

表 6 $\Delta t=0.20$ s 时五组独立实验的时间匹配结果

Table 6. Temporal matching results for five independent experiments at $\Delta t=0.20$ s.							
实验	温度参	真空度记	成功匹配的	匹配	平均绝对	95%分位绝	最大绝对
组	考点数	录点数	温度点数	率/%	时间差/s	对时间差/s	时间差/s
G1	8709	8708	8709	100.00	0.030	0.035	0.161
G2	5972	5971	5972	100.00	0.031	0.036	0.173
G3	3491	3490	3491	100.00	0.030	0.035	0.143
G4	4070	4052	4038	99.21	0.048	0.068	0.177
G5	7212	8386	7210	99.97	0.063	0.085	0.199
合计	29454	30607	29420	99.88	0.041	0.073	0.199

5组实验的匹配率均不低于99.21%，其中G1~G3的匹配率达到100%，G4和

G5的匹配率分别为99.21%和99.97%。按实验组计算，匹配率均值为99.84%±0.35%。各组平均绝对时间差为0.030~0.063 s，95%分位绝对时间差为0.035~0.085 s，最大绝对时间差均未超过设定的0.20 s时间窗口。

在时间窗口匹配基础上，为获得采样间隔一致的连续过程数据，进一步将温度和真空度数据线性插值重构至统一时间轴，并分别生成0.5 s和10 s等间隔数据。其中，0.5 s数据主要用于过程细节分析和动态变化监测，10 s数据主要用于实验记录展示、长期趋势回顾和电子实验记录归档。验证过程中，将重构后的曲线重新映射至相应的原始单通道采样时刻，并与相同时刻的原始测量值逐点比较。采用平均绝对误差（mean absolute error, MAE）评价重构前后的数值差异。四组实验的重构误差见表7。

表 7 五组独立实验的统一时间轴重构误差

Table 7. Reconstruction errors on unified time axes for five independent experiments.				
实验组	0.5 s 温度	0.5 s 真空度	10 s 温度	10 s 真空度
	MAE/℃	MAE/Torr	MAE/℃	MAE/Torr
G1	0.06	2.40×10 ⁻⁴	0.29	6.50×10 ⁻⁴
G2	0.07	2.90×10 ⁻⁴	0.34	8.10×10 ⁻⁴
G3	0.05	2.10×10 ⁻⁴	0.27	5.90×10 ⁻⁴
G4	0.12	5.80×10 ⁻⁴	0.49	1.62×10 ⁻³
G5	0.16	7.60×10 ⁻⁴	0.67	2.15×10 ⁻³

五组实验的温度重构结果表现出较好的一致性。0.5 s重构条件下，温度MAE为0.05~0.16 ℃；10 s重构条件下，温度MAE为0.27~0.67 ℃。其中，G1~G3的过程参数变化相对平缓，重构误差较小；G4和G5包含较明显的动态变化，因此误差有所增加。总体而言，0.5 s重构的温度MAE均不超过0.2 ℃，10 s重构的

温度MAE均不超过0.7 °C。

真空度重构结果呈现相同规律。0.5 s重构的真空度MAE为 $2.10 \times 10^{-4} \sim 7.60 \times 10^{-4}$ Torr，10 s重构的真空度MAE为 $5.90 \times 10^{-4} \sim 2.15 \times 10^{-3}$ Torr。五组实验中，0.5 s重构误差均低于相应的10 s重构误差，说明较小的重采样间隔能够更充分地保留局部动态变化。10 s重构虽然对短时波动产生了一定平滑作用，但总体误差仍处于较低水平，能够保持温度和真空度数据的主要变化趋势。作为逐组误差评价的补充，将G1~G5五组有效数据合并后，进一步计算重构曲线与原始单通道曲线之间的Pearson相关系数和决定系数 R^2 ，结果显示，0.5 s重构下温度和真空度的Pearson相关系数分别为0.9998和0.9975；10 s重构下温度和真空度分别为0.9991和0.9922。说明无论是0.5s还是10 s，重构曲线均能较好保持原始单通道数据的变化趋势。

综合来看，时间窗口合并能够保证数据的直接对应关系，线性插值重构则进一步提高了时间轴的一致性和数据连续性，可满足后续过程趋势分析、实验记录展示与电子实验记录归档需求。

6.3 自动化XRD测试评估

以标准XRD 2θ - θ 测试流程(sample alignment, move 2θ - θ , Ry/Rx measurement, 2θ - θ measurement)进行评估。其中Ry/Rx measurement的起始角度为 -1° ，终止角度 1° ，步长为 0.01° ，速度为 $2^\circ/\text{min}$ ， 2θ - θ measurement的起始角度为 5° ，终止角度 80° ，步长为 0.01° ，速度为 $10^\circ/\text{min}$ 。

通过12 h连续测试窗口评估，若采用人工方式，单个样品测试时间约20 min，则理论上可完成约35个样品。然而该模式要求操作人员全程值守，连续完成样品取放、参数设置、样品对位、测试启动、文件命名和数据整理等操作。随着测试

时间延长，操作者疲劳、参数输入错误、光路配置错误、样品顺序混乱及文件归档不一致等问题会显著增加，从而影响数据可靠性和可追溯性。

在机器人辅助自动化模式下，单个样品的XRD自动化测试时间约为20.4 min，机器人调度与取放样平均约7-8 min，因此单样品完整周期约为27.4-28.4 min。按12 h连续运行计算，可完成约25-26个样品。虽然由于机器人转运和样品交接引入了额外时间，使单台仪器的理论样品数低于理想人工连续操作，但自动化模式下人员仅需完成测试参数录入和样品柜上样，约1-2 min/样品，12 h批量测试的人工作业时间仅约25-52 min，人员占用降低约93%-97%，如图12所示。

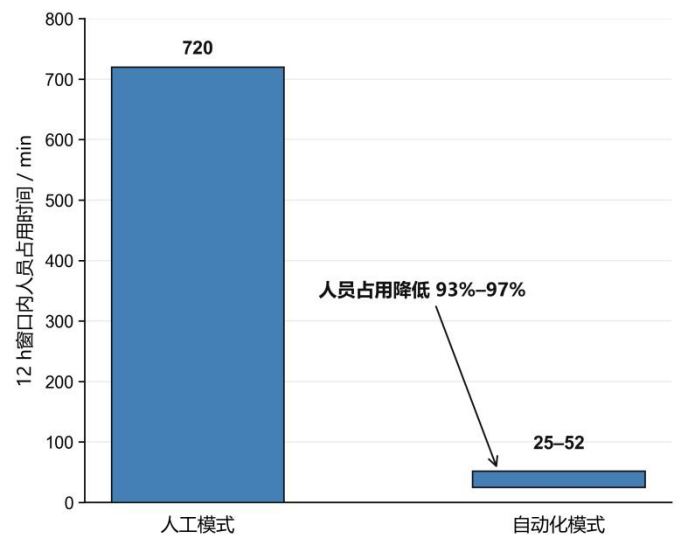


图 12 人工模式与自动化模式下 XRD 批量测试人员占用时间对比

Fig. 12. Comparison of personnel occupancy time between manual and automated XRD batch testing modes.

因此，该系统的主要优势是将高频、依赖经验的人工操作转化为标准化、可重复、可追溯的自动流程。在长时间批量测试场景中，自动化测试能够保持一致的样品对位逻辑、扫描参数、文件命名和数据绑定规则，避免人工连续值守导致的测试条件波动和数据管理错误。

6.4 ELN入库成功率

为验证在实际实验数据管理场景中的ELN数据入库稳定性，本文对运行过程中生成的上传日志进行统计分析。并将ELN上传过程中可能出现的异常归纳为四类：网络与连接异常、接口服务异常、权限与身份认证问题以及批量写入压力问题，如表8所示。

表 8 ELN 上传过程中的主要异常类型

Table 8. Classification of major exceptions occurring during the ELN data upload process.		
异常类型	可能原因	系统处理方式
网络与连接异常	网络波动、请求超时、连接中断	保留上传任务，重新发起请求
	ELN 接口临时无响应、服务器返回异常状态	保留数据，在 ELN 接口恢复后重试
接口服务异常	登录状态失效、访问令牌过期、权限校验失败	重新认证后继续上传
权限与身份认证问题	单次写入数据量较大、请求频率较高、接口处理负载增加	分批记录日志，并通过重试机制完成写入
批量写入压力问题		

表 9 ELN 上传过程日志统计（重试前单次请求）

Table 9. Statistical analysis of ELN upload logs for single requests before retry.				
上传类型	成功日志数	失败日志数	总请求数	单次请求成功率
新建 ELN 实验记录	101	11	112	90.18%
ELN 日志更新请求	2540	167	2707	93.83%
整体上传请求	2641	178	2819	93.69%

为了稳定的上传数据，将 ELN 上传操作分为两类：一类为“新建 ELN 实验

记录”，即首次将完整实验数据写入 ELN 平台并生成实验记录；另一类为“已有 ELN 记录的日志追加/更新”，即在实验记录创建完成后，继续向该记录补充实验日志、过程数据或附件信息。根据上传日志统计结果，重试前的单次请求成功率如表 9 所示。

在不考虑重试机制的情况下，首次 ELN 记录入库的单次请求成功率为 90.18%，ELN 日志更新请求的单次成功率为 93.83%，整体上传请求的单次成功率为 93.69%。这些失败日志主要反映上传过程中受到网络连接、接口响应、身份认证状态以及批量写入压力等因素影响时产生的瞬时异常。最终通过重试所有 ELN 记录均成功完成入库，最终入库成功率达到 100%。

6.5 工艺特征与YBCO（005）峰位的初步相关性分析

为初步检验所构建数据体系对工艺—结构关系分析的支撑能力，本节选取15批YBCO薄膜实验进行相关性分析。对于 c 轴取向的YBCO薄膜，（005）衍射峰的位置直接反映相应晶面的面间距变化，可用于表征薄膜 c 轴晶格状态。YBCO的 c 轴晶格参数对氧含量、晶格应变及缺陷状态较为敏感，而这些结构因素又与超导转变温度和临界电流密度等物理性能密切相关。因此，本文选取（005）峰位作为连接生长过程与薄膜结构状态的敏感表征指标。XRD数据首先以MgO衬底衍射峰作为内部参照进行角度校准，随后采用pseudo-Voigt函数拟合YBCO（005）衍射峰并提取峰位。分析变量包括生长时间、温度时序斜率、真空度时序斜率和羽辉平均灰度，分别对应实验元数据、温度过程、真空过程和原位羽辉图像信息。

温度和真空度时序斜率通过对相应生长时段的时序曲线进行线性拟合获得，用于表征实验过程中温度或真空度的总体变化方向及变化速率。羽辉平均灰度由生长过程中有效羽辉图像帧的灰度统计结果汇总得到，用于描述羽辉整体发光强度。上述指标均为对原始时序数据和图像数据的简单统计概括，本节主要用于验

证跨模态数据关联及基础统计分析流程，尚不以建立完整的工艺—结构预测模型为目标。

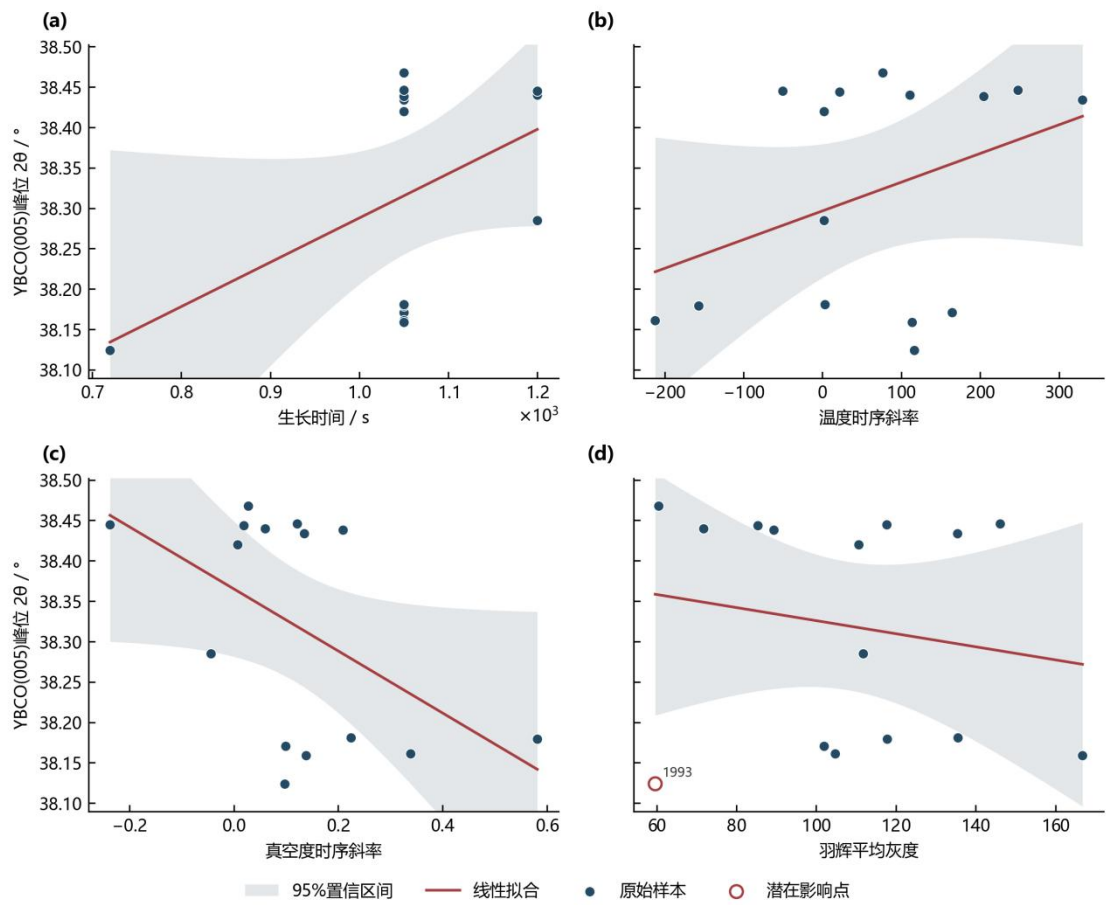


图 13 工艺及羽辉特征与 YBCO（005）峰位的相关性。（a）生长时间；（b）温度时序斜率；（c）真空度时序斜率；（d）羽辉平均灰度。

Fig. 13. Correlations of process and plume features with the YBCO (005) peak position. The panels show (a) growth time, (b) temperature time-series slope, (c) vacuum time-series slope, and (d) mean plume grayscale intensity.

采用Pearson相关检验评价各变量与YBCO（005）峰位之间的线性关联。

Pearson相关系数 r 用于表征线性相关的方向和程度，取值范围为 $-1\sim 1$ ，并通过显著性检验获得对应的 p 值。图13展示了15个样品的原始数据点、最小二乘线性拟合结果及拟合均值的95%置信区间，Pearson相关系数 r 的95%置信区间采用Fisher z 变换计算。考虑到本节同时进行了4项相关性检验，为降低多重比较导致的假阳性风险，进一步采用Benjamini - Hochberg方法对 p 值进行多重比较校正，并报告

错误发现率（false discovery rate, FDR）校正后的 q 值^[24]。多重比较校正后的统计学显著性以 $q<0.05$ 为判据。

由图13和表10可见,生长时间与YBCO(005)峰位呈中等程度正相关($r=0.440$, $p=0.100$), 温度时序斜率呈弱至中等程度正相关($r=0.369$, $p=0.176$)。真空度时序斜率的相关系数绝对值最大,表现为中等程度负相关($r=-0.505$, $p=0.055$)。羽辉平均灰度与峰位之间的线性相关性较弱($r=-0.180$, $p=0.521$)。

表 10 工艺及羽辉特征与 YBCO（005）峰位的 Pearson 相关性影响诊断

Table 10. Pearson correlations and influence diagnostics for process and plume features versus the YBCO (005) peak position.

分析变量	Pearson r	r 的 95% CI	p	BH-FDR q	潜在影响点
生长时间	0.440	[-0.093, 0.777]	0.100	0.201	无
温度时序斜率	0.369	[-0.177, 0.741]	0.176	0.235	无
真空度时序斜率	-0.505	[-0.808, 0.010]	0.055	0.201	无
羽辉平均灰度	-0.180	[-0.634, 0.366]	0.521	0.521	1993

经BH-FDR校正后,生长时间、温度时序斜率、真空度时序斜率和羽辉平均灰度对应的 q 值分别为0.201、0.235、0.201和0.521,均未达到0.05显著性水平。此外,四项Pearson相关系数的95%置信区间均包含0。因此,当前结果尚不能证明上述变量与YBCO（005）峰位之间存在统计学显著的总体线性关系。其中,真空度时序斜率的原始 p 值最接近0.05,可视为值得进一步检验的负相关趋势。

为评价单个样品对线性拟合结果的影响,进一步采用Cook距离和外部学生化残差进行回归诊断^[25]。本研究将Cook距离($D_i > 4/n$)或外部学生化残差绝对值大于2作为潜在影响点的经验筛查标准。由于 $n = 15$,Cook距离的筛查阈值为($4/15 = 0.267$)。诊断结果表明,生长时间、温度时序斜率和真空度时序斜率三项

分析中均未发现超过预设阈值的影响点。羽辉平均灰度分析中的1993号样品被识别为潜在影响点，其Cook距离为0.557，外部学生化残差为-2.129。保留全部15个样品时，羽辉平均灰度与YBCO（005）峰位的相关系数为-0.180；仅在敏感性分析中排除1993号样品后，相关系数变为-0.416，但仍未达到统计学显著水平（ $p=0.139$ ）。该结果说明，在当前样本量下，羽辉平均灰度的相关结果对个别样品具有一定敏感性。

基于实验样本与全局统计特征开展相关性检验，本分析并未得到工艺—结构间具备统计学显著性的线性关联结果。该现象来源于多方面的客观约束：一方面，本次分析样品规模有限，仅包含15个样品相关系数的置信区间较宽，统计估计仍具有较大不确定性；另一方面，温度和真空度目前仅采用整体时序斜率进行表征，羽辉图像也仅采用平均灰度这一全局亮度指标，原始数据中包含的阶段变化、局部波动和空间形态信息尚未得到充分利用。例如，具有相近总体斜率的两条温度或真空度曲线，可能在过程过冲、稳定时间、局部波动幅度和后期恢复速率等方面存在明显差异。后续可根据实际沉积过程划分不同工艺阶段，分别提取阶段均值、变化速率、稳定时间和波动幅度等特征。对于羽辉图像，可结合羽辉形成和传播过程开展具有物理意义的特征提取，包括羽辉面积、轴向长度、横向宽度、长宽比、质心位置、传播方向、轮廓波动等。对连续图像还可进一步计算这些特征随时间的漂移和波动，从而更完整地描述羽辉形态与动态稳定性。

该结果并不否定这套多模态数据库的分析价值，恰恰反映出材料工艺—结构关联挖掘对样本数量、精细化物理特征提取的双重要求。本节已实现实验元数据、温度与真空度时序数据、原位羽辉图像特征和XRD结构结果在样品层面的统一关联，并完成从特征提取、相关检验到异常影响诊断的分析流程。这表明所构建的数据体系能够为后续开展物理引导的特征提取和工艺—结构关系建模提供结构

化、可追溯的数据基础。综上，在进一步扩大样本量并补充超导性能数据后，可继续扩展至工艺—结构—性能关系研究。

7 结论

本文针对高温超导YBCO薄膜研发过程中实验数据采集不完整、多源数据难以关联以及全流程追溯不足等问题，探索了一套覆盖薄膜制备、结构表征、数据治理与ELN归档的全流程数据治理方法。在PLD生长过程中，温度和真空度由两路独立相机连续采集，并通过软件时间对齐实现联合记录；激光羽辉图像则利用激光器TTL信号进行硬件触发采集，实现羽辉图像与激光脉冲事件的同步记录。通过构建Raw-Processed-Merged三层数据模型与时间对齐方法，实现了多源异构实验数据的清洗、融合和追溯；通过机器人辅助XRD自动化测试与ELN结构化归档，实现了样品、工艺过程、结构表征文件和实验记录的自动关联。在此基础上，构建了面向YBCO薄膜制备与结构表征的多模态实验数据库，并通过典型实验验证了OCR后处理、软件时间对齐、自动化XRD测试和ELN入库等环节的有效性。

当前研究仍存在一定局限性：实验验证主要集中于制备过程数据与XRD结构表征结果，尚未系统纳入超导转变温度、临界电流密度和微波表面电阻等性能指标。因此，本文所构建的数据体系主要为YBCO薄膜工艺—结构关系研究提供结构化、可追溯的数据基础，并可在后续融合超导性能数据后进一步扩展至更完整的工艺—结构—性能研究。

未来将进一步扩展实验样本并纳入超导性能测试数据，完善工艺—结构—性能一体化数据集；将该采集治理方案拓展至更多薄膜制备与表征设备，同时探索结合大语言模型实现实验记录智能解析，进一步赋能高温超导薄膜的数据驱动研发。

8 致谢

本研究工作得到中国科学院超导专项的大力支持（项目编号：[SCZX-0102]）。

录用稿件，非最终出版稿

参考文献

- [1] Zhao Z X, Chen L Q, Yang Q S, Huang Y Z, Chen G H, Tang R M, Liu G R, Cui C G, Chen L, Wang L Z, Guo S Q, Li S L, Bi J Q 1987 *Chin. Sci. Bull.* **32** 412 (in Chinese) [赵忠贤, 陈立泉, 杨乾声, 黄玉珍, 陈赓华, 唐汝明, 刘贵荣, 崔长庚, 陈烈, 王连忠, 郭树权, 李山林, 毕建清 1987 科学通报 **32** 412]
- [2] Zhang J Q, Chen F C, Ma X P, Qiao D, Wang Y T, Yang J T, Feng Z P, Bai Z H, Xiong P Y, Wang J X, Li C G, Yang H X, Li Q, Tian S B, Gao X, Wang Y, Wang X, Yuan J, Zhu B Y, Jin K 2026 *Chin. Sci. Bull.* **71** 2387 (in Chinese) [张济全, 陈赋聪, 马小平, 乔丹, 王宇彤, 杨景婷, 冯中沛, 柏中华, 熊沛雨, 王家晓, 李春光, 杨槐馨, 李潜, 田士兵, 高翔, 王越, 王旭, 袁洁, 朱北沂, 金魁 2026 科学通报 **71** 2387]
- [3] Tian Q W, Yuan P S, Yu H Q, Wang S N, Liu X Y, Li L Y, You L X 2025 *Acta Phys. Sin.* **74** 057401 (in Chinese) [田清文, 原蒲升, 余慧勤, 汪书娜, 刘晓宇, 李凌云, 尤立星 2025 物理学报 **74** 057401]
- [4] Antončík F, Jankovský O, Hlášek T, Bartůněk V 2020 *Nanomaterials* **10** 1429
- [5] Dijkkamp D, Venkatesan T, Wu X D, Shaheen S A, Jisrawi N, Min-Lee Y H, McLean W L, Croft M 1987 *Appl. Phys. Lett.* **51** 619
- [6] Zhang D M, Li Z H, Zhong Z C, Guan L, Li X G 2011 *Pulsed Laser Deposition Dynamics* (Beijing: Science Press) pp45—64 (in Chinese) [张端明, 李智华, 钟志成, 关丽, 李小刚 2011 脉冲激光沉积动力学原理 (北京: 科学出版社) 第45—64页]
- [7] Price C C, Li Y S, Zhou G Y, Younas R, Zeng S S, Scanlon T H, Munro J M, Hinkle C L 2024 *Nano Lett.* **24** 14862
- [8] Ehrlinger L, Wöß W 2022 *Front. Big Data* **5** 850611
- [9] Bacco M, Kocian A, Chessa S, Crivello A, Barsocchi P 2024 *Data Brief* **57** 110969
- [10] Mitchell S N, Lahiff A, Cummings N, Hollocombe J, Boskamp B, Field R, Reddyhoff D, Zarebski K, Wilson A, Viola B, Burke M, Archibald B, Bessell P, Blackwell R, Boden L A, Brett A, Brett S, Dundas R, Enright J, Gonzalez-Beltran A N, Harris C, Hinder I, Hughes C D, Knight M, Mano V, McMonagle C, Mellor D, Mohr S, Marion G, Matthews L, McKendrick I J, Pooley C M, Porphyre T, Reeves A, Townsend E, Turner R, Walton J, Reeve R 2022 *Phil. Trans. R. Soc. A* **380** 20210300

- [11] Peng J Q, Zhou W, Han Y, Li M T, Liu W Q 2023 *Measurement* **222** 113615
- [12] IEEE 2020 IEEE Std 1588-2019: IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems (New York: IEEE)
- [13] Wang H, Li J, McDonald B E, Farrell T R, Huang X, Clancy E A 2023 *Sensors* **23** 2465
- [14] Li C X, Liu W W, Guo R Y, Yin X T, Jiang K T, Du Y K, Du Y N, Zhu L F, Lai B H, Hu X G, Yu D H, Ma Y J 2022 arXiv:2206.03001 [cs.CV]
- [15] Cui C, Sun T, Lin M H, Gao T Q, Zhang Y B, Liu J X, Wang X Q, Zhang Z L, Zhou C D, Liu H E, Zhang Y, Lv W Y, Huang K, Zhang Y C, Zhang J, Zhang J, Liu Y, Yu D H, Ma Y J 2025 arXiv:2507.05595 [cs.CV]
- [16] Azzabi S, Alfughi Z, Ouda A 2024 *Computers* **13** 183
- [17] Armbrust M, Ghodsi A, Xin R, Zaharia M 2021 *Proceedings of the 11th Annual Conference on Innovative Data Systems Research (CIDR 2021)* Virtual Event, January 11—15, 2021
- [18] JSON Schema: A Media Type for Describing JSON Documents, Wright A, Andrews H, Hutton B, Dennis G <https://json-schema.org/draft/2020-12/draft-bhutton-json-schema-00> [2026-09-03]
- [19] Siffa I C, Schäfer J, Becker M M 2022 *F1000Research* **11** 475
- [20] Buonsanti R 2023 *Chem. Mater.* **35** 805
- [21] Schubotz S, Schubotz M, Auernhammer G K 2025 *J. Open Res. Softw.* **13** 11
- [22] Buneman P, Khanna S, Tan W C 2001 *Database Theory — ICDT 2001: 8th International Conference* London, UK, January 4—6, 2001 p316
- [23] Wang D, Zhou M B, Huang D C, Li Y L, Lin Z F, Liu J D, Zhu T N, Zhu Y, Li M X, Xiao R J, Yuan J, Weng H M 2024 *Chin. Sci. Bull.* **69** 1164 (in Chinese) [王丹, 周明波, 黄东宸, 李云龙, 林泽丰, 刘俊德, 朱天念, 竺云, 李明星, 肖睿娟, 袁洁, 翁红明 2024 科学通报 **69** 1164]
- [24] Benjamini Y, Hochberg Y 1995 *J. R. Stat. Soc. Ser. B* **57** 289
- [25] Cook R D 2000 *Technometrics* **42** 65

Standardized Data Collection and Governance for Fabrication and Structural Characterization of High-Temperature Superconducting YBCO Thin-Films

1) (Condensed Matter Science Data Center, Institute of Physics, Chinese Academy of Sciences,
Beijing 100190, China)

2) (National Laboratory for Superconductivity, Institute of Physics, Chinese Academy of Sciences,
Beijing 100190, China)

3) (School of Artificial Intelligence and Computer Science, North China University of Technology,
Beijing 100144, China)

4) (Beijing Key Laboratory of Key Technologies for AI+ Domain Applications, North China University
of Technology, Beijing 100144, China)

5) (School of Mechanical-Electrical and Vehicle Engineering, Beijing University of Civil Engineering
and Architecture, Beijing 102616, China)

High-temperature superconducting yttrium barium copper oxide (YBa₂Cu₃O_{7-δ}, YBCO) thin-films are important materials for high-temperature superconducting electronics. Their microstructure and superconducting properties are highly sensitive to fabrication conditions, oxygen content, and defect states. However, the lack of standardized experimental data, incomplete process information, difficulties in integrating heterogeneous data, and insufficient sample-to-data traceability hinder the development of high-quality process databases and data-driven optimization methods. To address these challenges, this work

proposes a standardized data collection and governance method covering the fabrication and structural characterization of YBCO thin-films. During pulsed laser deposition (PLD), temperature and vacuum-pressure readings are continuously recorded using two independent cameras and temporally aligned through time-window matching and linear interpolation. Laser plume images are acquired through hardware triggering by the laser's TTL signal, enabling synchronization with laser pulse events. For structural characterization, robotic sample delivery, intelligent task scheduling, and automated multi-mode X-ray diffraction (XRD) testing are integrated to associate sample identifiers, testing parameters, and raw diffraction data. A three-layer Raw – Processed – Merged data storage and governance architecture is also developed. Original experimental evidence is preserved, while multi-source asynchronous data are cleaned, standardized, and integrated through optical character recognition (OCR) post-processing, physical-rule validation, anomaly screening, time-window matching, and linear interpolation. The processed data are aligned on a unified time axis and organized into structured datasets for subsequent analysis. Based on this framework, a multimodal experimental database is established to integrate experimental metadata, temperature and vacuum time-series data, plume images, XRD characterization files, and electronic laboratory notebook (ELN) records. A structured metadata model and ELN interface support searchable experimental records and end-to-end traceability from analytical results to original videos, OCR

outputs, and experimental operations. The proposed framework transforms fragmented experimental records into standardized, structured, and machine-readable data assets, providing a foundation for studying process–structure relationships in YBCO thin-films and supporting future data-driven research. The datasets presented in this paper, including temperature and vacuum time-series data, laser plume images, XRD characterization data, and ELN records, are available at <https://doi.org/10.57760/sciencedb.j00213.00081> (private access link: <https://www.scidb.cn/s/F7rMJb>).

Keywords: YBCO thin-films, automated XRD characterization, data governance, electronic laboratory notebook (ELN)

* Project supported by the National Key R&D Program of China (Grant No. 2022YFA1603903), the National Natural Science Foundation of China (Grant No.U23A6015), and CAS Project for Young Scientists in Basic Research(Grant No. 2022YSBR-048) .

These authors contributed equally.

† Corresponding author.E-mail:lihongyang@iphy.ac.cn; fangjun@ncut.edu.cn
The first author.E-mail:17812030907@163.com