

面向中子源的高质量数据集构建技术研究

胡鹏^{1,2,3}, 张振宇^{1,3,4}, 胡皓^{1,3}, 李庆梦^{1,3}, 张正德^{1,3}, 熊东波^{1,3}, 齐法制^{1,2,3,4}

- 中国科学院高能物理研究所, 北京 100049;
- 散裂中子源科学中心, 广东东莞 523803;
- 国家高能物理科学数据中心, 北京 100049;
- 郑州大学计算机与人工智能学院, 河南郑州 450001

摘要

中国散裂中子源产生的中子科学数据在语义规范化表达、跨源融合对齐、AI任务的数据集构建等方面面临显著挑战, 难以直接支撑人工智能模型训练和数据驱动的科学发现。因此, 如何从复杂中子科学数据中构建面向AI4S的高质量数据集, 已成为推动中子科学研究向智能化转型的关键挑战。本文针对上述挑战, 在现有中子源数据治理体系的基础上, 提出了一种面向高质量数据集构建的五层技术框架, 具体包括: (1) 标准规范层, 统一元数据标准与数据格式, 实现多源异构数据的逻辑对齐; (2) 数据管理层, 构建多维数据存储与关联基座, 支撑大规模数据的快速检索与调用; (3) 质量增强层, 依托人工智能技术实现科学元数据的自动补齐与逻辑校验, 提升数据质量; (4) 数据融合层, 从数据流整合与大模型智能体两个维度推进, 探索实验数据与模拟、文本数据对齐; (5) 任务适配层, 利用大语言模型智能标注, 形成面向特定AI任务的端到端高质量数据集。基于该框架, 本文开展了在不同科研场景下的探索性验证: (1) 在非弹性中子散射研究中, 以实验数据为核心, 结合第一性原理计算和粒子输运模拟构建实验-模拟融合数据集, 有效缓解了实验数据稀缺等问题; (2) 在页岩孔隙结构研究中, 融合小角中子散射实验数据与文献数据, 构建多模态融合数据, 实现孔隙结构参数的高精度预测; (3) 在中子衍射结构解析中, 构建实验-模拟-文献协同的数据智能体, 实现精修过程的自动化执行与数据集生成, 并形成用于谱图解析和结构预测任务的AI-Ready数据集。最后, 本文针对面向中子源的高质量数据集构建提出相关思考: (1) 从源头和制度层面加强元数据质量的精细化把控; (2) 持续积极引入先进的数据与人工智能技术, 探索更高效、智能的构建方案; (3) 预研下一代具有自主进化能力的数据治理, 实现长期可持续发展、不断自我强化的数据基础设施。

关键词

中国散裂中子源; 高质量数据集; AI-Ready数据集; 数据融合; AI4S

中图分类号: TP311.13
issn.2096-0271.2021xxx

文献标志码: A

doi:10.11959/j.

*Research on high-quality dataset
construction technologies for*

neutron source facility

HU Peng^{1,2,3}, ZHANG Zhenyu^{1,3,4}, HU Hao^{1,3}, LI Qingmeng^{1,3}, ZHANG Zhengde^{1,3},
XIONG Dongbo^{1,3}, QI Fazhi^{1,2,3,4}

1. Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China;
2. Spallation Neutron Source Science Center, Dongguan 523803, China;
3. National High Energy Physics Science Data Center, Beijing 100049, China;
4. School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001, China

Abstract

Scientific data generated from the China Spallation Neutron Source face significant challenges in semantic standardization, cross-source alignment and integration, and dataset construction for AI tasks, making it difficult to directly support artificial intelligence model training and data-driven scientific discovery. Therefore, how to construct high-quality AI4S-oriented datasets from complex neutron science data has become a key challenge for advancing the intelligent transformation of neutron science research. To address these challenges, built upon the existing data governance system of neutron sources, this study proposes a five-layer technical framework for high-quality dataset construction, which specifically includes: (1) a standard specification layer, which unifies metadata standards and data formats to achieve logical alignment of multi-source heterogeneous data; (2) a data management layer, which establishes a multidimensional data storage and association infrastructure to support efficient retrieval and access of large-scale data; (3) a quality enhancement layer, which leverages artificial intelligence techniques to automatically complete and logically validate scientific metadata, thereby improving data quality; (4) a data fusion layer, which advances from both data flow integration and large-model-based intelligent agents to explore alignment between experimental data and simulation and textual data; and (5) a task adaptation layer, which utilizes large language models for intelligent annotation to construct end-to-end high-quality datasets tailored to specific AI tasks. Based on this framework, exploratory validations are conducted in multiple scientific scenarios: (1) in inelastic neutron scattering studies, an experiment-centered dataset integrating experimental data with first-principles calculations and particle transport simulations is constructed, effectively alleviating the issue of limited experimental data; (2) in shale pore structure research, multimodal fused data are constructed by integrating small-angle neutron scattering experimental data with literature data, enabling high-precision prediction of pore structure parameters; (3) in neutron diffraction structure analysis, an experiment-simulation-literature collaborative data agent is developed to automate refinement processes and dataset generation, forming AI-Ready datasets for spectral analysis and structure prediction tasks. Finally, this study presents perspectives on high-quality dataset construction for neutron sources. First, it is necessary to strengthen fine-grained control of metadata quality at the source and institutional levels. Second, continuous integration of advanced data and artificial intelligence technologies should be pursued to explore more efficient and intelligent construction approaches. Third, it is essential to proactively investigate next-generation data governance systems with self-evolving capabilities, enabling long-term sustainable development and continuously self-reinforcing data infrastructure.

Key words

China Spallation Neutron Source, High-quality Dataset, AI-Ready Dataset, Data Fusion, AI4S

1 引言

1.1 中子源 AI4S 的高质量数据集需求

中国散裂中子源是我国首台脉冲型散裂中子源，是面向材料、物理、化学、生物等基础科学前沿以及航空航天、高铁、化工等国家重大工程应用的中子散射综合研究平台。随着人工智能驱动的科学研究的快速发展，人工智能技术已广泛应用至中子源装置运行与领域科学研究的多个环节^[1-5]。不同于传统以存储和共享为目的的数据归档，面向 AI 任务的高质量数据集强调统一标准、完整科学语义、实验过程可追溯及数据的一致性与准确性，能够直接支撑模型训练、推理与科学发现^[6-8]。在中子实验方法学分析、新材料研发^[9-10]及重大工程应用等领域，高质量数据集已成为推动 AI4S 发展的关键基础。

1.2 中子源数据特征及高质量数据集构建挑战

中子源数据在 AI4S 应用中具有重要价值，但其复杂的数据结构特征也使高质量数据集构建面临显著挑战。从数据组织与应用特点来看，这些挑战主要体现在物理语义强耦合、多源多模态异构以及实验场景高度离散三个方面。

首先，中子源数据具有显著的物理语义强耦合特征。实验数据不仅包含探测信号本身，还高度依赖谱仪参数、束流条件、样品属性及原位环境等元数据。同时，数据处理过程涉及复杂物理模型与参数拟合，其结果对实验条件高度敏感。因此，实验数据、设备参数与数据处理流程之间形成

紧密关联，一旦元数据缺失或不规范，便可能导致实验结果不可复现或物理意义不可解析^[11]。这使得中子源数据难以脱离实验上下文进行统一描述与自动化建模。

其次，中子源数据具有多来源、多模态与语义异构并存的特点。除实验数据外，还包括理论计算模拟数据以及实验日志、学术文献等文本数据。这些数据在表达形式上涵盖谱图、图像、时间序列与文本等多种模态，在物理语义、时空尺度及数据精度方面存在显著差异^[12]。例如，实验数据反映真实物理过程，模拟数据基于理想模型，而文本数据则以自然语言描述知识信息。不同数据源之间缺乏统一的语义对齐机制，导致跨模态数据难以直接融合与协同利用。

此外，中子源数据还呈现实验场景多样与数据分布离散的特征。中子源服务于材料、物理、化学、生物等多个学科领域，不同实验方法学在测量原理、数据结构及分析流程上差异明显。与此同时，现有数据围绕实验过程组织，缺乏面向具体科学任务的输入输出结构定义与标准化标注方式，导致数据难以直接转化为可用于模型训练的任务数据集^[13]。

综上，中子源高质量数据集构建面临三类核心挑战：一是物理语义强耦合导致数据高度依赖实验上下文，增加了标准化表达与自动解析难度；二是多源多模态异构导致跨数据源缺乏统一语义对齐机制，制约数据融合与协同利用；三是实验场景高度离散导致数据与具体 AI 任务之间缺乏直接映射关系，难以形成标准化训练数据集。

2 中子源科学数据治理现状分析

2.1 基于DOMAS框架的数据治理

面向中国散裂中子源运行过程中产生的多模态、多来源科学数据，当前已逐步形成覆盖数据采集、传输、存储、处理、共享与服务的全生命周期数据治理体系。相较于传统以数据存储为核心的数据管理模式，现有体系已能够支持实验数据与元数据的关联组织，并初步具备模拟数据、文本数据等多源数据的统一管理能力和数据可发现性、可访问性及基础共享服务

方面形成了较完善的数据治理基础。

目前,中国散裂中子源的数据治理工作主要基于DOMAS框架开展^[14]。DOMAS是面向先进光源、中子源等大科学装置研发的通用数据管理框架,采用模块化架构,实现科学数据在全生命周期内的自动化组织、传输、存储与服务。该框架集成元数据管理、高效数据传输及数据服务等核心功能,并支持通过参数化配置适配不同实验装置需求。

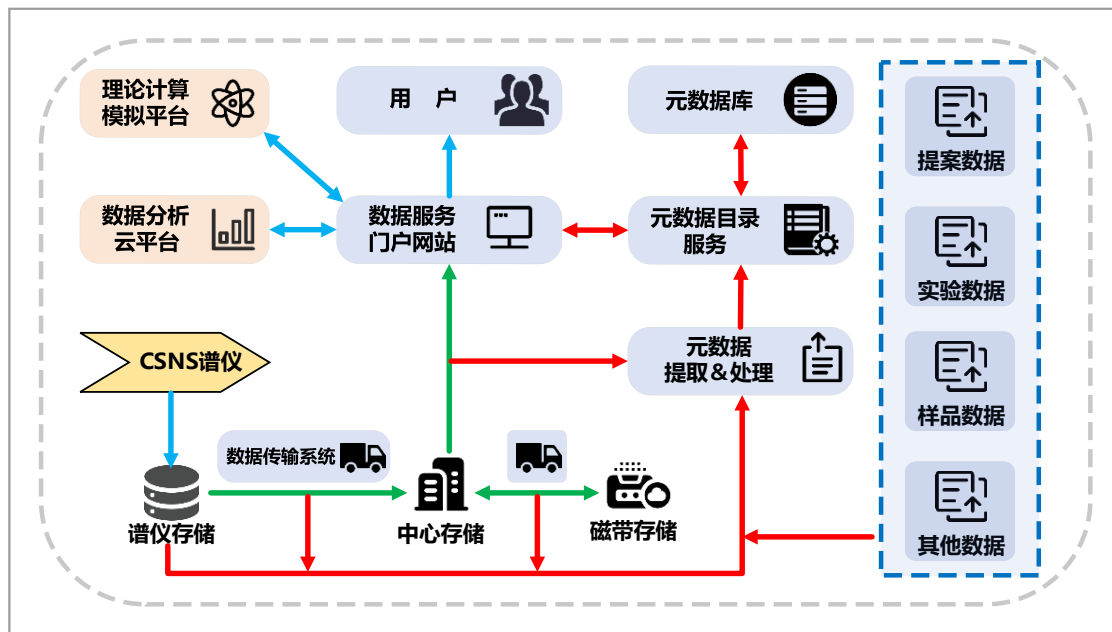


图1 中国散裂中子源数据管理架构设计图

2.2 现有数据治理方法的局限性

尽管当前基于 DOMAS 的中子源数据治理体系已具备较完善的数据管理与共享能力,但其核心目标仍在以下几个方面存在不足。

在物理语义表达方面, 现有体系已建立基于NeXus的元数据标准与结构化采集

机制,但部分原位环境参数、实验日志及数据处理过程信息仍依赖人工记录,数据规约与分析流程缺乏全流程自动追踪^[11]。同时,不同系统之间在命名规范与单位表达上仍存在不一致问题,导致数据语义完整性与可追溯性不足,数据难以脱离实验上下文实现独立理解与复用。

在多模态数据对齐与融合方面，不同数据源在组织粒度上存在结构性差异：实验数据以实验机时组织，模拟数据以计算任务为单元，文本数据以知识库为载体，粒度的不统一使数据关联与对齐融合的复杂度显著增加^[15]。更为关键的是，实验参数与模拟参数之间、结构化数据与文本知识之间尚未建立稳定的语义映射机制，多源数据融合在很大程度上仍依赖人工经验与定制化流程，缺乏通用的对齐融合范式。这一能力缺口直接制约了多来源、多模态数据集的自动化与规模化构建。

在面向 AI 任务的数据组织方面，当前数据体系仍主要围绕实验过程进行组织，缺乏针对具体科学任务的输入输出结构定义、统一标签体系与标准化标注机制^[16]。不同实验方法之间的数据结构与分析流程差异进一步增加了任务数据集构建难度，

使实验数据向 AI-Ready 数据集的转化过程仍高度依赖人工处理，难以满足 AI4S 对规模化、高质量数据供给的需求。

3 高质量数据集构建技术框架

3.1 总体架构设计

上述分析表明，面向 AI4S 对高质量数据集构建的新需求，现有中子源科学数据治理体系虽然能够实现数据规范化管理，但仍难以解决中子科学数据物理语义表达不足、多源异构数据难融合以及科学数据难以直接支撑 AI 任务等问题。为此，本文在现有中子源数据治理体系和 DOMAS 基础上，提出面向高质量数据集构建的五层技术框架，实现由数据管理向任务驱动数据供给的能力扩展。

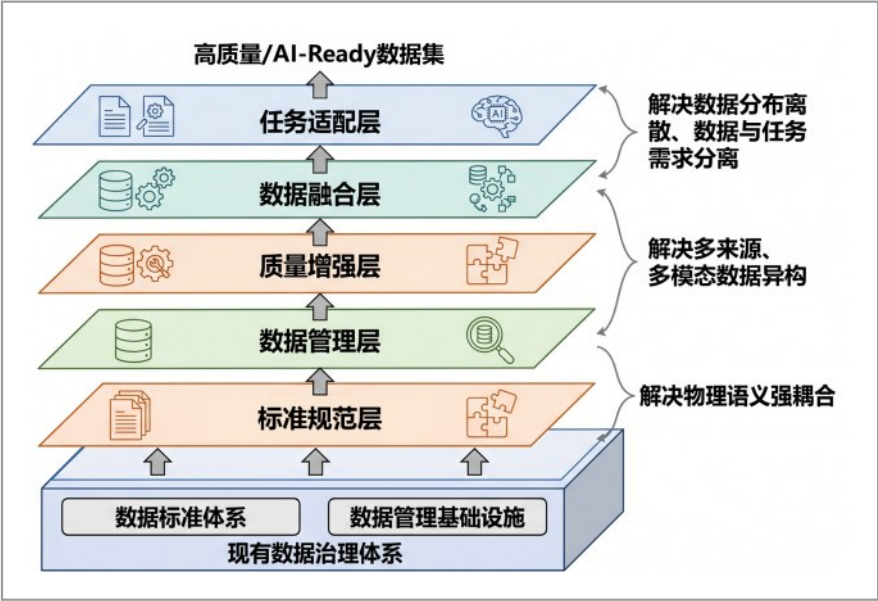


图2 高质量数据集构建总体框架

从功能定位上看，五层框架并非相互独立，而是形成逐级递进的数据处理流程，从基础构建、智能增强到任务化供给。其中，标准规范层负责统一数据格式、元数据模型与物理语义表达；数据管理层实现实验数据、模拟数据与文本数据的统一组织与关联管理；质量增强层通过规则校验与人工智能方法提升数据完整性与一致性；数据融合层实现跨来源、多模态数据的语义对齐与协同组织；任务适配层完成面向具体 AI 任务的数据组织，逐步实现高质量数据集的标准化构建与持续供给。

3.2 框架关键技术

3.2.1 标准规范层

针对中子源科学数据来源复杂、元数据表达不统一、物理语义缺失等问题，在 DOMAS 框架已有的数据标准体系之上构建标准规范层，以 NeXus 数据格式和现有元数据规范为基础，提出科学对象级语义建模、跨数据源语义统一以及 AI 导向的数据质量约束三项关键机制，为后续数据管理、跨模态融合和任务化组织提供统一的数据基础。

针对传统元数据以字段为组织单元、难以表达科学对象之间物理关联的问题，框架采用科学对象级语义建模方式，将样品、谱仪、实验模式、原位环境等抽象为统一科学对象，并建立对象之间的关联关系，实现元数据由字段级向对象级语义表达的转换，从而提高科学数据物理语义的完整性与可解释性。

为实现实验数据、模拟数据和文本数据之间的跨源融合，进一步建立跨数据源的统一语义体系。面向实验数据、模拟数据和文本知识数据，建立统一的数据对象描述规范和语义映射规则，将实验数据、

模拟数据和文本知识映射到统一语义空间，实现跨数据源的数据关联与语义对齐，为后续多源融合提供一致的数据表示。

考虑 AI 模型训练对完整上下文的依赖，提出最小必需元数据集。将样品属性、原位实验环境、数据处理过程等关键科学信息定义为强制记录项，从源头提升数据完整性和可追溯性。

3.2.2 数据管理层

尽管 DOMAS 已经实现实验数据管理，但实验数据、模拟数据和文本数据仍分散于不同系统，缺乏统一的数据组织方式，导致跨数据源关联困难，难以支撑后续的数据融合与 AI 任务构建。针对该问题，设计数据管理层，在 DOMAS 已有的实验数据管理能力基础上，扩展多源数据协同组织与关联管理机制，为质量增强、数据融合和任务适配提供统一的数据底座。

为解决不同来源数据分散管理的问题，数据管理层采用统一数据接入机制。在现有实验数据管理基础上，进一步接入样品管理、原位实验环境、模拟计算平台和文本知识平台等数据源，通过统一接口实现实验数据、模拟数据和文本数据的协同管理，补充实验上下文信息。

为解决不同系统之间缺乏统一关联标识的问题，进一步建立覆盖提案、实验、样品、谱仪、原位环境、模拟任务及文献等科学对象的统一标识体系。以实验过程为主线构建跨系统关联关系，实现实验数据、模拟数据和文本知识的统一组织与可追溯管理，为后续跨模态数据融合提供关联基础。

3.2.3 质量增强层

标准规范层和数据管理层实现了数据

规范化组织，但仍存在元数据缺失、跨系统数据不一致以及数据质量难以保证等问题。针对上述问题，构建了质量增强层，引入规则约束与大语言模型相结合的数据质量增强机制，实现元数据自动补全、一致性校验和质量闭环管理。

针对实验过程中存在的原位环境参数缺失、样品属性记录不完整以及数据处理参数未结构化等问题，构建多源元数据抽取与语义补全。面向实验日志、谱仪控制文件、样品信息表、模拟参数及文本文献等多源异构数据，通过LLM与结构化解析工具协同，实现关键科学元数据的自动抽取与补全，基于上下文语义进行字段补全与合理性推断，将标准规范层所定义的最小必需元数据集要求落实为可自动执行的补全操作，从而提升元数据覆盖率与结构完整性。

基于实验全周期流程的一致性约束与标准规范层定义的元数据规则体系，对中子实验数据的多层级结构关系进行一致性校验。可以有效识别跨系统数据之间的逻辑不一致问题，并通过规则引擎与LLM辅助分析进行异常标记与修正建议生成。例如，实验系统一致性校验：检查中子谱仪类型、实验模式、数据文件结构与对应实验参数之间是否匹配，确保数据生成来源与设备配置在结构上具有一致性。从工程层面保障数据的结构可信性与过程可追溯性。

为保证数据可靠性，该层建立自动校验和分级人工审核的混合质量控制机制。对于常规数据，采用自动化校验流程；对于高价值实验数据与异常检测结果，则引入专家复核与终审机制。经质量增强后的数据将以结构化形式回写至数据管理层。这一闭环回流机制使数据质量得以随实验积累持续演化迭代，而非停留于一次性清

洗，形成持续演化的数据质量闭环。

通过语义补全、一致性校验与分级审核的协同作用，质量增强层能够有效提升中子科学数据的语义完整性与物理一致性，为后续多模态数据融合与任务导向数据集构建提供数据质量保障。

3.2.4 数据融合层

由于不同数据来源在物理语义、组织粒度及表示方式上存在差异，仅依靠统一标识难以支撑跨模态联合分析。因此，在已有关联关系基础上，进一步设计数据融合层，通过统一语义表示、多源语义对齐及融合数据单元构建，实现不同来源科学数据的协同表达，为任务适配提供统一的数据输入。

在统一语义表示基础上，以科学对象为中心构建多源融合机制，将实验数据、模拟数据与文本知识进行联合组织。其中，实验数据提供真实观测信息，模拟数据补充理论计算结果，文本知识提供领域经验与背景知识。融合过程中，通过统一对象标识建立不同数据源之间的对应关系，并依据统一语义模型完成属性补充、语义互证和信息融合，使不同来源的数据形成完整的数据表达。

在语义关联基础上，该层将实验数据、模拟数据与文本知识进行统一组织，构建多源数据协同表达机制。三类数据在统一语义框架下形成互补关系，从而提升数据表达的完整性与解释能力。其中，实验数据提供真实观测结果，模拟数据提供理论计算支撑，文本数据提供领域先验知识与经验描述。三者的协同表达提升了数据在描述完整性与物理可解释性方面的能力，为后续任务适配层的标注与结构化输出提供更为丰富的语义基础。

为便于AI任务直接调用，提出多源融

合数据单元的组织形式。融合数据单元以样品、实验过程或特定科学对象进行构建，聚合其对应的实验数据、处理结果、模拟数据及相关文本知识，并保持各数据来源之间的关联关系，从而形成具有统一语义、多模态协同表达和完整上下文信息的数据实体。

最终，数据融合层输出面向科学对象的融合数据单元，为每个对象建立统一的多模态数据表示。该表示不仅保持实验、模拟与文本知识之间的语义关联，而且能够完整保留实验上下文及数据来源信息，为任务适配层的数据标注、样本构建和 AI-Ready 数据集生成提供统一的数据输入。至此，中子源科学数据实现了由多源并存向融合表达的转变。

3.2.5 任务适配层

尽管融合数据已经具备统一语义表达和完整的科学上下文，但其组织方式仍围绕科学对象和实验过程展开，尚不能直接满足 AI 模型训练的数据组织需求。不同 AI 任务对输入特征、输出标签及样本组织方式具有不同要求，因此需要进一步完成任务建模、样本生成和数据集封装。为此，本文设计任务适配层，实现融合数据向 AI-Ready 数据集的转换。

针对中子散射谱解析、晶体结构精修等典型 AI 场景，对原始物理过程进行任务化抽象，明确模型输入与输出的物理含义及训练样本组织方式。依据不同任务需求，从融合数据单元中提取实验条件、谱仪参数及样品属性作为输入特征，将散射谱特征、结构参数作为输出标签，构建符合机器学习样本之间的结构化映射关系，实现融合数据向训练样本的转换。

在样本映射基础上，结合大语言模型与规则引擎完成任务相关标签的自动化智

能生成。标注内容不仅包括分类或数值标签，还包括物理过程描述、特征区间识别及跨模态语义关联，利用大模型从实验上下文中自动推断物理语义并生成结构化标注。例如，在小角中子散射实验中，自动识别谱图特征变化并关联对应实验条件与样品材料，实现实验条件、谱图特征和结构精修的结构化表达。

在完成样本映射和智能标注后，对训练样本按照统一的数据集规范进行组织和封装，生成 AI-Ready 数据集。按照统一的数据集规范，对数据进行筛选、组织与封装，形成面向特定任务的数据集产品。数据集不仅包含输入输出数据，还包括任务定义、数据来源、处理流程及质量评估信息，支持模型训练、验证与复现实验。通过引入统一的数据集描述规范与版本管理机制，实现数据集的可复用、可迭代与可追溯。

通过任务建模、样本生成、智能标注及标准化封装，任务适配层实现了融合数据向 AI-Ready 数据集的自动转换，使科学数据能够直接服务于模型训练、模型评测及科学智能应用，为中子源 AI4S 提供统一的数据供给能力。

4 高质量数据集构建应用实例

基于上述五层技术框架，本文围绕中子科学 AI4S 场景开展多模态数据融合与高质量数据集构建实践，以实验数据为核心，结合模拟数据与文本知识数据，在非弹性中子散射、页岩孔隙结构研究及中子衍射智能分析等典型场景中进行了探索性验证。值得指出的是，当前相关实践主要面向高质量数据集构建流程验证和应用可行性探索，重点验证多源数据融合、质量增强和

任务适配流程的可实施性，整体仍处于初步探索阶段。相关智能化方法的可靠性、稳定性和泛化能力，以及对具体AI任务性能提升的实际贡献，仍需在后续工作中结合更大规模数据和更多应用场景开展系统研究。

4.1 非弹性中子散射实验模拟数据集构建

高能非弹中子散射是研究材料晶格动力学、声子输运及微观激发行为的重要实验方法，在热电材料、能源材料及凝聚态物理研究具有重要应用价值。然而，该类研究长期面临实验机时成本高、有效样本数量有限、理想真值谱难以直接获得以及多物理过程耦合复杂等问题，制约了人工智能方法在非弹散射数据解析中的应用^[17]。针对上述问题，本文依托五层框架开展实验-模拟协同的数据集构建实践。在实验数据主导的前提下，引入多尺度计算与粒子输运模拟结果，为实验数据提供物理解释、监督标签构造与样本增强支撑，从而提升实验数据的可学习性。

在标准规范层，围绕高能非弹性中子散射实验建立统一的数据描述体系。实验数据采用标准化格式组织，对中子飞行时间、动量转移 Q 、能量转移 E 、探测器几何参数、束流条件、样品环境参数及数据处理流程等核心元数据进行统一定义。同时，对第一性原理计算输出的力常数矩阵、声子色散结果以及理论动力学结构因子等模拟数据建立兼容的数据结构，实现实验数据与模拟结果在字段结构、单位体系及物理语义上的一致表达，为跨源数据协同利用提供基础支撑。

在数据管理层，以实验数据为主线开展多源数据统一管理。汇聚高能非弹谱仪实测TOF原始数据、谱仪配置参数、样品

组成信息、实验环境记录及数据处理结果，并关联管理基于VASP、Phonopy、BzScope及Neutron Cinema生成的大批量模拟数据。通过实验编号、样品编号、测量条件编号及计算任务编号建立统一标识体系，实现同一样品、同一实验条件下实验数据与模拟数据的精准关联，形成可追溯、可检索、可复用的实验主导型数据资源池。

在质量增强层，围绕实验数据可信度提升开展智能质量控制。针对实测谱图，自动识别声子谱连续性、散射强度合理性、布拉格峰位置一致性问题；针对元数据，自动补齐缺失的谱仪运行参数、样品环境条件及数据处理参数等。结合物理规则约束修正异常数值，辅以人工复核保障数据物理自洽与可追溯。

在数据融合层，以实验条件和样品对象为核心，通过统一标识将同一样品、同一实验条件下的实验谱图与理论计算结果进行对齐，将BzScope计算得到的理想动力学结构因子作为实验谱图解析的参考标签，用于构建监督学习样本。

在任务适配层，面向非弹散射数据智能解析任务，对融合数据进行结构化组织与标准化输出。最终形成以原始实测散射谱为输入，以动力学结构因子、峰位参数为输出的监督学习数据集，可直接用于谱图降噪、峰特征识别和参数反演等任务。数据集同时保留实验条件、样品属性及处理流程信息，支持模型训练、验证与可解释分析。

最终，根据上述流程构建形成了包含136条实验数据与5万余条模拟数据的高质量数据集，并已应用于非弹性中子散射谱图特征识别等任务，实现了非弹谱的实时解析。未来将进一步持续引入更多实验数据，不断提升模型解析效率、预测精度与

复杂场景适应能力。

4.2 实验-文献融合的页岩孔隙结构数据集构建

页岩储层孔隙结构表征是页岩油气开发与二氧化碳封存研究中的关键问题。小

角中子散射能够表征纳米尺度孔隙结构，但受实验成本高、样品获取困难及单次实验信息有限等因素影响，仅依赖实验数据难以支撑数据驱动建模。与此同时，大量相关实验结果分散于文献中，尚未形成可计算的数据资源。为此，本文开展实验数据与文献知识融合实践。

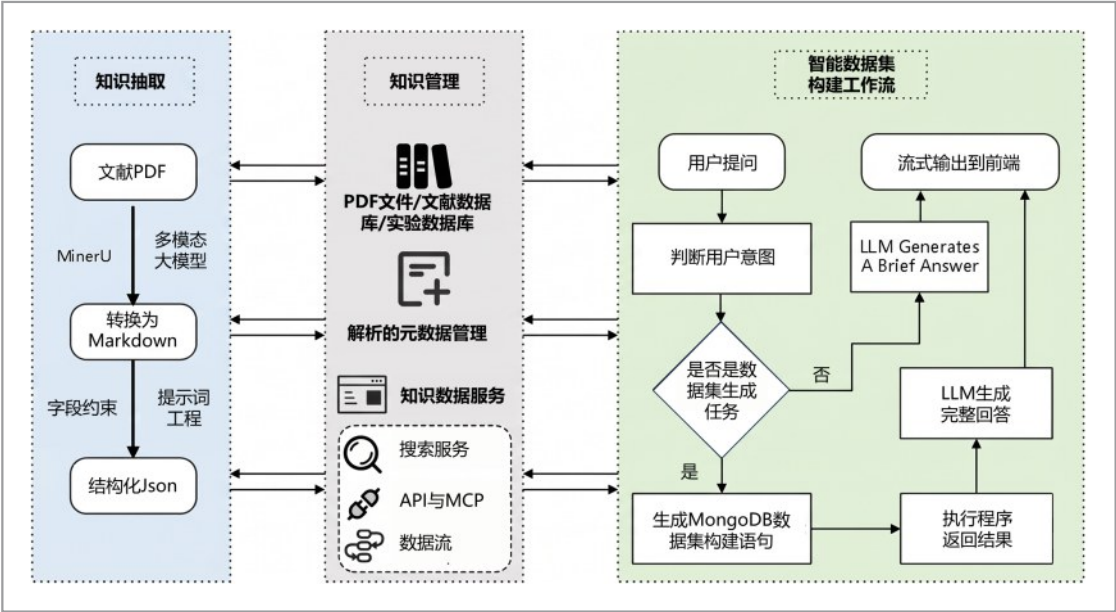


图3 AI-4-SPS系统整体构建框架

在标准规范层，针对实验数据与文献数据格式、单位、语义不一致问题，制定页岩样品、中子散射、压汞法、核磁共振法统一元数据标准，采用 NeXus 作为实验数据标准格式，文献抽取数据以 JSON/CSV 结构化存储。同时统一孔径范围、孔隙度、矿物组分、岩相类型等关键参数的单位体系，构建页岩储层领域术语本体，消除跨源数据语义歧义。

在数据管理层，以实验数据为主线实现数据统一汇聚与关联组织。汇聚中国散裂中子源页岩 SANS 原位实测数据，包括不同压力条件下获得的散射谱数据，以及

配套样品信息、实验日志、数据处理记录和谱仪运行参数。同时整合从 160 余篇文献中抽取的孔隙结构参数、矿物组成、及多方法学表征结果，构建样品-实验条件-散射谱-结构参数-文献数据多源数据资源池。通过统一标识、索引检索与版本管理机制，实现实验数据全流程可追溯管理。

在质量增强层，围绕数据质量提升与文献知识规范化开展联合增强。针对 SANS 实验数据，自动检查 Q 区间覆盖是否完整、背景扣除是否合理、散射曲线是否存在异常波动、重复实验结果是否一致，并校验实验条件记录与样品编号的一致性。

对于文献数据，利用 LLM 辅助完成从文献文本中实验参数、样品条件与物理指标的结构化抽取与补全；结合页岩孔隙结构理论，对孔隙度、连通率、孔径区间和测试方法之间的逻辑关系进行校验，识别并修正异常数值，确保数据准确、完整、物理自洽。

在数据融合层，开展实验、文献跨模态精准融合。对于具有明确来源标识的数据，通过唯一数据 ID 实现同盆地、同层位、同岩相及相近实验条件样本之间的直接配对。对于缺乏显示关联的数据，则利用语义嵌入模型对样品属性、测试条件与结构描述进行相似性匹配，实现基于语义相似性的辅助对齐。融合后，实验 SANS 数据提供真实观测约束，文献数据提供更广范围的孔隙结构统计信息及辅助表征结果，两者相互补充，有效扩充样本规模，形成覆盖中国和北美主要页岩盆地、包含多种测试方法的融合数据集。

在任务适配层，面向页岩孔隙率、孔径分布反演等任务，对融合数据进行结构化重组和智能标注。以实验散射谱特征、Q 区间相应参数、样品矿物组成及有机质含量作为输入特征，以孔隙结构参数与储层评价指标作为输出标签，构建标准化训练样品，为页岩油气勘探提供高质量数据支撑。

最终，基于上述流程构建形成包含 216 条高质量样本的多模态融合数据集。该数据集完成了实验数据与文献知识的结构化融合，为后续页岩孔隙结构参数预测等 AI 任务提供了标准化数据基础。相关模型训练与性能评估工作正在进一步开展，未来将结合预测精度、泛化能力等指标对数据集构建效果进行系统评价。

4.3 数据智能体在数据集构建中的

应用

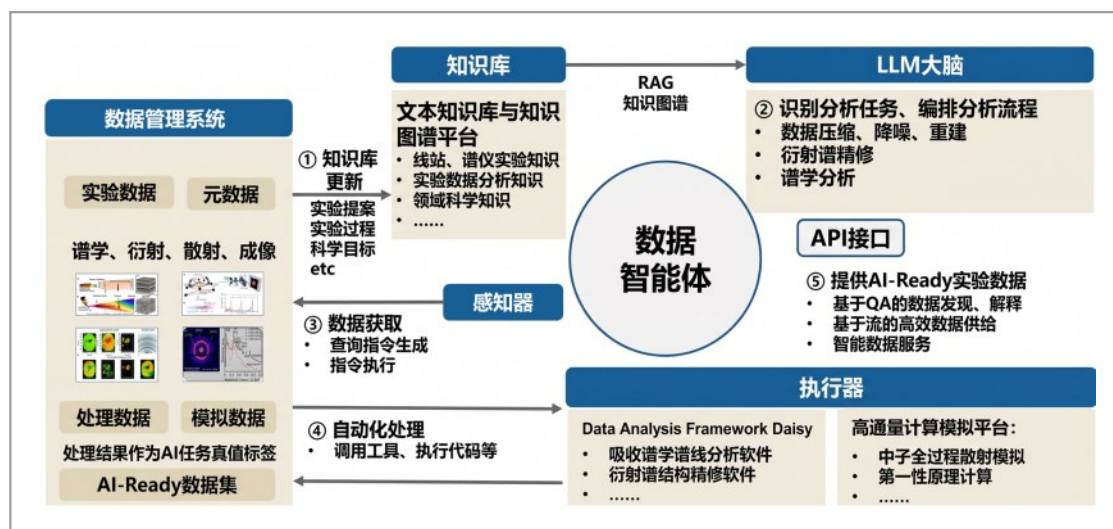
中子衍射结构精修是材料晶体结构解析中的关键环节，其典型核心流程通常包括初始结构选择、衍射峰拟合、参数优化与收敛判断等多个迭代步骤，具有流程长、依赖经验强、多源数据协同要求高等特点。传统方法中，实验与模拟数据往往由人工分别处理，缺乏统一的自动化协同机制。本节将数据智能体^[18]作为高质量数据集构建框架的任务执行与决策层，构建框架（见图 4）与智能体驱动的中子衍射自动精修数据生成流程，实现从数据治理到数据集构建的闭环贯通。智能体能够自动完成跨层数据调用、处理流程编排、工具协同与任务执行决策等流程，无需人工逐步干预。

在标准规范层，建立中子衍射领域统一数据标准体系。实验数据采用 NeXus 格式规范存储；模拟数据使用 CIF 晶体结构文件和仪器参数文件，并适配 GSAS-II 精修分析工具接口进行拟合，缩小理论与实验数据的误差值；为跨源数据融合奠定基础。

数据管理层实现多源数据统一汇聚与关联管理。依托 DOMAS 数据管理系统，汇聚中子衍射实验谱、中子衍射模拟谱及衍射谱精修理论知识。以材料与实验编号为唯一标识，将三类数据精准绑定，使智能体能够进行跨源数据检索与调用。

质量增强层依托大语言模型与物理规则引擎提升数据可靠性。校验衍射峰位、晶胞参数等物理逻辑合理性，剔除噪声数据与异常值；筛选物理自洽的高质量模拟结构数据，为高精度精修提供优质数据基底。

数据融合层通过智能体实现实验与模拟数据深度协同。用户提供初始晶体结构 CIF 文件作为精修起点，智能体以实验衍



射数据为真值约束，调用 GSAS-II 工具自动完成峰位拟合、晶胞参数优化、Rwp 收敛迭代等流程，驱动模拟数据向实验数据的逐步拟合。智能体通过迭代精修不断优化结构参数，最终输出低 Rwp 值的精修结果和更新后的 CIF 文件。这一过程本质上是一个“实验-模拟”双向校准的闭环，实验数据提供真实约束，模拟数据提供结构优化方向，二者协同驱动精修收敛。

任务适配层面向 AI 任务完成数据集标准化封装。基于上述流程，智能体已将实验谱、模拟谱与精修后结构数据进行有机整合，形成包含“实验衍射谱-理论衍射谱-精修后晶体结构”三元组的高质量理实融合数据集。目前，团队已基于该流程构建了示例小样本数据集，验证了智能体具备自动化构建理实融合数据集的能力，后续将持续扩大数据规模，为特定 AI 任务构建专用训练数据集，支撑中子科学领域的 AI4S 应用。

5 结论与展望

高质量元数据是科学数据治理的基础，也是数据共享复用、实验结果可追溯以及 AI4S 应用的重要前提。虽然本文建立了面向中子散射实验的元数据标准体系，并结合规则约束与智能补全提升了元数据质量，但复杂实验场景下仍存在语义一致性、自动采集能力及标准覆盖不足等问题。未来需要进一步加强元数据标准、实验设备和科研管理之间的协同，推动元数据采集、质量控制与共享规范一体化发展，为高质量数据集持续建设提供支撑^[8,15]。

人工智能技术正推动科学数据治理由人工驱动向智能协同演进。未来，可将大语言模型、知识图谱及智能体等技术嵌入数据治理全过程，实现实验记录解析、元数据抽取、数据关联、质量检查及数据集构建等环节的自动化与智能化，同时兼顾数据准备与处理过程中的成本与收益协同优化^[19]，提高高质量数据集的生产效率与一致性，促进 Data4AI 与 AI4Data 协同发展，为科研数据基础设施智能升级提供技术支撑。

随着实验数据、模拟数据与文献数据的深度融合，数据可信性将成为未来科学

数据治理的重要关注方向。模拟数据受模型假设、计算参数及适用范围等因素影响，可能存在一定的系统性偏差，因此未来将借鉴可信数据空间等先进治理架构^[20]，进一步加强融合数据的可信度评估与质量控制，建立实验验证、多源交叉验证及数据来源标注等机制，持续提升AI训练数据的可靠性，为AI4S应用提供更加可信的数据基础。

面向长期科学数据供给需求，未来数据治理应由静态管理进一步迈向动态演化，构建具备持续学习、自我优化能力的科学数据生态^[21]。通过融合持续积累的实验数据、模拟数据、领域知识及科研人员反馈，建立数据质量持续评估、语义修复、跨模态融合和任务数据集迭代更新机制，使数据体系能够不断完善和演进，为AI4S研究提供长期稳定、高质量的数据支撑。

参考文献：

[1] 李亚康,李健芳,胡鹏,等.人工智能在中子散射实验全生命周期中的应用综述[J]. 计算机工程, 2025,51(10):53-70.

Li Y K, Li J F, Hu P, et al. Review on application of artificial intelligence in full life-cycle of neutron scattering experiments[J]. Computer Engineering, 2025, 51(10): 53-70.

[2] BARAH O O, JOHN U K, EZE V H U, et al. Toward standardized, AI-driven multi-modal battery characterization: a review of neutron and synchrotron strategies across scales[J]. Ionics, 2025, 31(11): 11439-11461.

[3] WANG S, SHI H, WANG D, et al. Neutron-based characterization: A rising star in illuminating rechargeable lithium metal batteries[J]. Nano Energy, 2024, 122: 109337.

[4] SAMARAKOON A, TENNANT D A, YE F, et al. Integration of machine learning with neutron scattering for the Hamiltonian tuning of spin ice under pressure [J]. Communications Materials, 2022, 3(1).

[5] WANG H, DU R, LIU Z, et al. Machine learning in neutron scattering data analysis[J]. Journal of Radiation Research and Applied Sciences, 2024, 17 (2): 100870.

[6] GONG Y, LIU G, XUE Y, et al. A survey on dataset quality in machine learning [J]. Information and Software Technology, 2023, 162: 107268.

[7] MALERBA D, PASQUADIBISCEGLIE V. Data-Centric AI[J]. Journal of Intelligent Information Systems, 2024, 62(6): 1493-1502.

[8] GUILLEN-AGUINAGA M, AGUINAGA-ONTOSO E, GUILLEN-AGUINAGA L, et al. Data Quality in the Age of AI: A Review of Governance, Ethics, and the FAIR Principles[J]. Data, 2025, 10 (12): 201.

[9] TUR NYI R, MUKHOPADHYAY S. Neutron Spectroscopy and Computational Methods in investigation of Na ion Battery Materials: A Perspective[A/OL]. arXiv, 2023. <https://arxiv.org/abs/2304.08621>.

[10] CHENG M, FU C L, OKABE R, et al. Artificial intelligence-driven approaches for materials design and discovery[J]. Nature Materials, 2026, 25(2): 174-190.

[11] RAVI N, CHATURVEDI P, HUERTA E A, et al. FAIR principles for AI models with a practical application for accelerated high energy diffraction microscopy [J]. Scientific Data, 2022, 9(1). <http://dx.doi.org/10.1038/s41597-022-01712-9>. DOI:10.1038/s41597-022-01712-9.

[12] GHISLAT G, HERNANDEZ-

HERNANDEZ S, PIYAWAJANUSORN C, et al. Data-centric challenges with the application and adoption of artificial intelligence for drug discovery[A/OL]. arXiv, 2024. <https://arxiv.org/abs/2407.05150>. DOI: 10.48550/ARXIV.2407.05150.

[13] AHN H, LEE J. An Objective-Dependent and Reproducible Framework for AI Dataset Quality Assessment: In-Domain and Cross-Dataset Evidence From Building-Footprint Segmentation[J]. IEEE Access, 2026, 14: 103424-103434.

[14] Hu H, Lei L, Wang H, et al. DOMAS: a data management software framework for advanced light sources[J]. Synchrotron Radiation, 2024, 31(2): 312-321.

[15] BREWER W, WIDENER P, ANANTHARAJ V, et al. Data Readiness for Scientific AI at Scale[A/OL]. arXiv, 2025. <https://arxiv.org/abs/2507.23018>. DOI:10.48550/ARXIV.2507.23018.

[16] MITTAL S, THAKRAL K, SINGH R, et al. On responsible machine learning datasets emphasizing fairness, privacy and regulatory norms with examples in biometrics and healthcare[J]. Nature Machine Intelligence, 2024, 6(8): 936-949.

[17] FU C, CHENG M, HUNG N T, et al. AI-Driven Defect Engineering for Advanced Thermoelectric Materials[J]. Advanced Materials, 2025, 37(35). <http://dx.doi.org/10.1002/adma.202505642>. DOI: 10.1002/adma.202505642.

[18] Li Q, Wang H, Xiong D, et al., Rongzai agent: A Large Language Model-Based Autonomous Assistant for Rietveld Refinement of Neutron Diffraction Data[J]. arXiv preprint arXiv:2605.13911, 2026.

[19] 张群洪,李林,杨堃等.基于可信数据空间架构的高质量数据集建设[J].大数据, 2026. Zhang Q H, Li L, Yang K, et al. Construction of high-quality datasets based on trusted data space architecture[J]. Big Data Research, 2026.

[20] 崔文萱,丁小欧,程海峰,等.面向成本收益协同优化的高质量数据准备方法[J/OL]. 大数据, 2026. Cui W X, Ding X O, Cheng H F, et al. High-quality data preparation method for collaborative optimization of cost-benefit[J/OL]. Big Data Research, 2026.

[21] GHADERI D, SAHIMI A, ESSER B, et al. Why biology must prioritize data re-analysis in the era of artificial intelligence[J]. PLOS Biology, 2026, 24(7): e3003888.

录用日期: XXXX-XX-XX

通信作者:

基金项目: 中国科学院智能科学家平台、国家高能物理科学数据中心大湾区分中心/广东省物质科学数据中心(No. 2024B1212080001)。

Foundation Items: Robotic AI-Scientist platform of Chinese Academy of Sciences, and National High Energy Physics Science Data Center Guangdong-Hong Kong-Macao Greater Bay Area Branch / Guangdong Provincial Material Science Data Center (No. 2024B1212080001).