# AI-Ready科学数据集

## *AI-Ready Scientific Datasets*

**客座主编**



**陈刚**（1961– ），男，博士，中国科学院高能物理研究所研究员、博士生导师，国家高能物理科学数据中心主任。长期从事高能物理科学计算及大数据研究。推动高能物理分布式计算国际合作，为国际高能物理大规模数据处理和计算作出重要贡献。任国际高能物理科学计算协调委员会委员、国际高能物理数据全生命周期工作委员会委员等职。承担多项科学技术部、国家发展和改革委员会关于科学数据及人工智能的重大项目，推动高能物理科学数据共享及人工智能应用。



**邹自明**（1971– ），男，博士，中国科学院国家空间科学中心研究员、博士生导师，国家空间科学数据中心主任，国家重点研发计划首席科学家，中国空间科学学会常务理事，中国空间科学学会空间科学大数据专业委员会主任委员，全国科技平台标准化技术委员会（TC486）科学数据专家组组长。长期从事空间科学大数据与人工智能交叉领域研究，在科学数据治理理论、标准研制、空间信息组织与互操作、日地空间大数据系统工程、空间天气数据挖掘与知识发现等方面取得一系列成果，在国内外高水平学术期刊发表论文百余篇，获中国科学院杰出科技成就奖。



**崔辰州**（1976– ），男，博士，中国科学院国家天文台研究员，国家天文科学数据中心常务副主任。从事天文信息学研究二十余年，曾担任国际虚拟天文台联盟主席、国际天文学联合会（International Astronomical Union，IAU）数据与文献委员会主席，发起成立中国天文学会信息化工作委员会和IAU数据驱动天文科普教育工作组。发表科研论文120多篇，承担科研项目30余项，获得北京市、天津市科学技术进步奖二等奖。

# 导读

随着人工智能技术与科学数据深度融合，AI-Ready科学数据集作为“数据+AI”驱动科研范式变革的核心基础，在国家科学数据中心建设、跨领域AI科研应用与战略高技术突破中发挥着关键支撑作用。当前，尽管物理、天文、材料、海洋及冰冻圈等领域的科学数据规模激增，但普遍存在标准不一、质量参差、语义关联弱、流通风险大等突出问题，如何把原始科学数据转化为可直接支撑大模型训练与数据驱动发现的AI-Ready数据集，已成为亟待破解的基础性课题。

基于此，本刊以“AI-Ready科学数据集”为主题组织征文，特邀中国科学院陈刚、邹自明、崔辰州3位研究员担任客座主编。专题最终录用的8篇文章，围绕数据集建设实践、关键技术与质量治理、开放共享与安全溯源等主题展开系统探索。

在数据集建设实践方面，胡鹏等人针对我国散裂中子源数据语义表达与跨源融合的挑战，提出覆盖标准规范、数据管理、质量增强、数据融合与任务适配的5层技术框架，并在非弹性中子散射、页岩孔隙结构、中子衍射解析等场景构建了实验-模拟-文献融合的AI-Ready数据集。侯晓岚等人基于LAMOST与多源巡天数据构建了多模态数据治理流水线，深度融合Gaia、2MASS、DESI等异构信息，为天文大模型提供高质量数据燃料。高瑜蔚等人立足国家基础学科公共科学数据中心，提出培育权威数据库、遴选AI4S数据集、建设数据论文预印本平台等五大实践路径。翟军等人系统梳理美国国立卫生研究院的数据政策、Bridge2AI旗舰数据集与AI就绪度框架，为我国高质量数据集建设提供了国际借鉴。

在关键技术与质量治理方面，张智星等人面向冰冻圈野外观测数据，构建了数据接引阶段的质控流程，通过将规则质控与统计检测相结合，实现质量问题分层识别与可追溯标记。左维等人基于评价学构建了AI-Ready材料数据集质量评价框架，通过统一评价体系实现跨模态、跨维度的数据质量画像。席培奇等人面向对地观测数据提出知识图谱增强的多阶段语义检索方法，显著提升了检索精度与综合性能。

在开放共享与安全溯源方面，童心等人针对海洋AI-Ready数据集全流通过程中的新型安全威胁，构建了3级分级分类体系与五维量化评分模型，并融合数字水印与联盟链技术建立全链路可信溯源体系。

构建AI-Ready科学数据集，统筹科学数据资源的标准化高质量供给与安全可信流通，是释放数据要素价值、支撑AI for Science战略的重要基础。本次专题只是一个起点，期待这些探索凝聚学术与产业合力，推动AI-Ready科学数据集建设走向体系化、规模化，助力我国科研范式变革。