LHAASO第二次合作组会议

Data Express in Big Sciences期刊介绍

《大科学数据快报（英文）》

介绍人：程耀东

2026.10.21-26

PART 1 / DEFINITION

什么是数据论文?

定 义

数据论文是一种**以系统描述数据集为核心**的学术出版物。它不报告新的科学发现或分析结论，而是呈现数据集的背景、采集与处理方法、结构、质量与获取方式，使数据本身成为可被正式引用与追踪的学术产出。

与传统论文的核心区别

	传统研究论文	数据论文
主角	分析数据 → 得结论	**让数据被发现、复用、引用**
核心产出	新的科学发现	**高质量、可复用的数据集**
结论部分	必须有	**不需要**

数据论文的“主角”是数据，而非结论。

PART 2 / CORE STRUCTURE

数据论文的三要素

（一）数据集描述

数据**从哪里来**、

怎么产生的、

包含**什么内容**。

（二）分析示例或验证示例

可复现的演示：

一段分析流程或代码，展示如何提取物理量并得到预期结果

或用**已知标准样本**验证数据集正确性。

（三）数据集使用说明

获取方式：平台名称、访问链接、标识符（DOI / CSTR）、访问条件

使用环境：软件、库、工具及版本，软硬件依赖，代码获取方式

许可协议：数据授权使用方式

注意事项：已知局限、适用范围、推荐引用方式、后续更新计划

数据论文不需要结论与讨论部分。其价值在于数据的完整、准确与可用，而非得出新结论。

PART 3 / SAMPLE PAPER 01

范文一

SAMPLE 02

LHC 物理数据集

LHC physics dataset for unsupervised New Physics detection at 40 MHz 发表于Scientific Data

scientific data

Check for updates

OPEN

DATA DESCRIPTOR

LHC physics dataset for unsupervised New Physics detection at 40 MHz

Ekaterina Govorkova[1], Ema Puljak[1]✉, Thea Aarrestad[1], Maurizio Pierini[1], Kinga Anna Woźniak[1,3] & Jennifer Ngadiuba[2,4]

In the particle detectors at the Large Hadron Collider, hundreds of millions of proton-proton collisions are produced every second. If one could store the whole data stream produced in these collisions, tens of terabytes of data would be written to disk every second. The general-purpose experiments ATLAS and CMS reduce this overwhelming data volume to a sustainable level, by deciding in real-time whether each collision event should be kept for further analysis or be discarded. We introduce a dataset of proton collision events that emulates a typical data stream collected by such a real-time processing system, pre-filtered by requiring the presence of at least one electron or muon. This dataset could be used to develop novel event selection strategies and assess their sensitivity to new phenomena. In particular, we intend to stimulate a community-based effort towards the design of novel algorithms for performing unsupervised new physics detection, customized to fit the bandwidth, latency and computational resource constraints of the real-time event selection system of a typical particle detector.

Background & Summary

The proton bunches of the CERN Large Hadron Collider (LHC) cross paths $\mathcal{O}(10)$ million times per second in each of the experimental halls, possibly generating a *collision event* each time. In such an event multiple proton pairs may collide, possibly producing thousands of particles to be detected by the detectors located in each experimental hall. Detector sensors record the flow of emerging particles in the form of electronic signals. For the so-called general purpose detectors (ATLAS[1] and CMS[2]), this globally amounts to $\mathcal{O}(1\ \text{MB})$ of information. The resulting data throughput of $\mathcal{O}(10)$ TB/s is too large to be recorded. This is why these two detectors process data in real-time to select a small fraction of them (about 1000/s), compatible with downstream computing resources. This strategy was effective in providing the data needed to discover the Higgs boson[3,4].

This filtering system, usually referred to as the *trigger*, consists of a two-stage selection, as illustrated in Fig. 1. With the recent upgrade of their data acquisition system, the two other big detectors, LHCb and ALICE, have been equipped with a novel data processing capability that avoids the need to select events. Instead, their data acquisition follows a real-time data processing approach[5,6]: all the events enter a computer farm, where they are processed in real-time and stored in a reduced-size data format, consisting of high-level information sufficient for data analysis. For this reason, their case goes beyond the scope of this paper, and it is not considered further.

The first trigger stage, the Level-1 trigger (L1T), runs a set of algorithms deployed as logic circuits on custom electronic boards equipped with field-programmable gate arrays (FPGAs). This stage rejects more than 98% of the events, reducing the incoming data stream to 100k events/s. Due to the short time interval between two events (25 ns) and limited buffer capabilities, the entire pipeline of L1T algorithms has to be executed within $\mathcal{O}(1)\ \mu$s. The second stage, called the high-level trigger (HLT), consists of a computer farm processing events on commercial CPUs, running hundreds of complex selection algorithms within $\mathcal{O}(100)$ ms. The trigger selection algorithms are designed to guarantee a high acceptance rate for the physics processes under study.

When designing searches for unobserved physics phenomena, one typically considers specific theory-motivated scenarios. This *supervised* strategy has been proven successful when dealing with strongly theoretically motivated searches, e.g., the Higgs boson discovery[3,4]. However, this approach might become a limiting factor in the absence of strong theoretical guidance. The ATLAS and CMS trigger systems could be discarding interesting events, limiting the possibility of new physics discoveries.

[1]European Organization for Nuclear Research (CERN), CH-1211, Geneva 23, Switzerland. [2]Fermi National Accelerator Laboratory, Batavia, IL, 60510, USA. [3]Present address: University of Vienna, Vienna, Austria. [4]Present address: California Institute of Technology, Pasadena, CA, 91125, USA. ✉e-mail: ema.puljak@cern.ch

PART 3 / SAMPLE PAPER 02

范文二

SAMPLE 01

BESIII 实验数据集

Experimental dataset from BESIII detector at Beijing electron–positron collider

发表于《核技术（英文版）》

Nuclear Science and Techniques (2025) 36:218
https://doi.org/10.1007/s41365-025-01789-y

DATA ARTICLE

Experimental dataset from BESIII detector at Beijing electron–positron collider

Ming-Hua Liao[1] · Jian-Shu Liu[1] · Xin-Nan Wang[2] · Sheng-Sen Sun[2,3] · Zheng-Yun You[1]

Received: 1 April 2025 / Revised: 8 May 2025 / Accepted: 19 June 2025 / Published online: 12 September 2025
© The Author(s), under exclusive licence to China Science Publishing & Media Ltd. (Science Press), Shanghai Institute of Applied Physics, the Chinese Academy of Sciences, Chinese Nuclear Society 2025

Abstract
In the BESIII detector at Beijing electron–positron collider, billions of events from e^+e^- collisions were recorded. These events passing through the trigger system were saved in raw data format files. They play an important role in the study of physics in τ-charm energy region. Here, we published an e^+e^- collision dataset containing both Monte Carlo simulation samples and real data collected by the BESIII detector. The data pass through the detector trigger system, file format conversion, and physics information extraction and finally save the physics information and detector response in text format files. This dataset is publicly available and is intended to provide interested scientists and those outside of the BESIII collaboration with event information from BESIII, which can be used to understand physics research in e^+e^- collisions, developing visualization projects for physics education, public outreach, and science advocacy.

Keywords Electron–positron collision · BESIII · Data sharing · Education · Visualization

1 Introduction

The Beijing Electron–Positron Collider (BEPC) [1] has been in operation since 1989 and was upgraded to BEPCII in 2008. The BEPCII was designed to operate in τ-charm energy region. It is a double-ring multi-beam electron–positron collider with a circumference of 237.5 m, beam energy ranging from 1.0 to 2.45 GeV, and luminosity of 1×10^{33} $cm^{-2}s^{-1}$. BEPCII has 93 bunches arranged at 8 ns or 2.4 m intervals in each ring. The electron and positron beams collided at a horizontal crossing angle of ±11 mrad at the intersection.

Beijing Spectrometer III (BESIII) [2–4], a spectrometer running at BEPCII, can accurately record the information of the final state particles from e^+e^- collisions. This information was recorded using a two-level trigger system [5]. Level-1 (L1) is hardware-based, and level-3 (L3) is software-based, known as the event filter, which operates on an online computer farm. The L1 trigger aims to efficiently identify high-quality physics events and reduce cosmic rays and beam-related background noise. Data stored in the front-end buffers are retrieved once an L1 accept signal is received, approximately 6.4 ms after a collision. Subsequently, the data were sent to an online computer farm for event processing. The event stream undergoes filtering to further suppress background events through the L3 trigger before saving to the permanent storage device in the raw file format.

The BESIII Offline Software System (BOSS) [6, 7], which was developed using C++ language and object-oriented techniques, uses the event data service of Gaudi as a data manager to read raw data information. Reconstruction algorithms [8–10] can retrieve information from raw data and reconstruct physical information, which will be stored for the subsequent physics analysis.

This work was supported by the National Key Research and Development Program of China (No. 2023YFA1606000), National Natural Science Foundation of China (Nos. 12175321, 11975021, and U1932101), and National College Students Science and Technology Innovation Project of Sun Yat-sen University.

✉ Zheng-Yun You
youzhy5@mail.sysu.edu.cn

1 School of Physics, Sun Yat-Sen University, Guangzhou 510275, China

2 Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China

3 University of Chinese Academy of Sciences, Beijing 100049, China

范文

Experimental dataset from BESIII detector at Beijing electron–positron collider

3 Data records

The dataset can be accessed via the repository listed in the specification table. Table 1 lists the physical processes, selected number of events, and data types included in the dataset. The dataset consisted of two parts: one based on Monte Carlo (MC) simulation samples using the BESIII offline software and the other based on real data collected by the BESIII detector. The events in the dataset were saved in a text format with hit and track information from the detector response. The content of the dataset is introduced in the following sections.

Because some classical physics processes are submerged in huge background events in real data, it is not easy to select them directly. However, the behavior of MC events is close to real data from event displays and can help us determine whether rare physics events are signals. The MC samples include three major decay modes with large branching ratios for the ground state of vector charmonium J/ψ, such as the electromagnetic decay $J/\psi \to e^+e^-/\mu^+\mu^-$, and the hadronic decay $J/\psi \to \pi^+\pi^-\pi^0$ [24], in which $J/\psi \to e^+e^-$ is the decay channel where the charm quark was discovered in 1974 [25, 26]. Additionally, the decay modes of charmed meson D, such as $D^0 \to K^-\pi^+, \bar{D}^0 \to K^+\pi^-, D^+ \to K^-\pi^+\pi^+$, and $D^- \to K^+\pi^-\pi^-$ [27, 28], are provided, as they are the main single-tag reconstruction channels in D meson studies widely used in charm physics analysis. Correspondingly, the main decay channel of the lightest charmed baryon Λ_c^+, $\Lambda_c^+ \to pK^-\pi^+$ [29], is provided. The dataset also contains the semi-leptonic process $\Lambda_c^+ \to ne^+\nu_e$, and the $\bar{\Lambda}_c^-$ is reconstructed using some main decay channels [30]. The neutrons in the final states have a clear signature in the EMC with a distinct hit distribution. This physics process has also been applied in neutron identification studies at BESIII based on deep learning techniques [30].

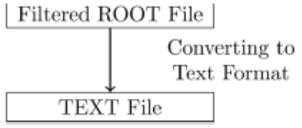


Fig. 2 The data conversion flow of BESIII event files

Table 1 The physics processes, number of events, and data types in the dataset

S.N	Physics processes	Events	Types
1	$J/\psi \to e^+e^-$	128	MC
2	$J/\psi \to \mu^+\mu^-$	126	MC
3	$J/\psi \to \pi^+\pi^-\pi^0$	155	MC
4	$D^0\bar{D}^0$, $D^0 \to K^-\pi^+, \bar{D}^0 \to K^+\pi^-$	122	MC
5	D^+D^-, $D^+ \to K^-\pi^+\pi^+$, $D^- \to K^+\pi^-\pi^-$	140	MC
6	$\Lambda_c^+\bar{\Lambda}_c^-$, $\Lambda_c^+ \to pK^-\pi^+$, $\bar{\Lambda}_c^- \to \bar{p}K^+\pi^-$	113	MC
7	$\Lambda_c^+ \to ne^+\nu_e$	104	MC
8	$e^+e^- \to \pi^+\pi^- J/\psi$ ($Z_c(3900)$)	10	Data
9	$e^+e^- \to e^+e^-$	141	Data
10	$e^+e^- \to \mu^+\mu^-$	121	Data
11	$e^+e^- \to \gamma\gamma$	150	Data
12	3097 MeV	150	Data
13	3686 MeV	150	Data
14	3773 MeV	144	Data
15	Cosmic ray	10	Data

The BESIII detector collected a large amount of real data in the τ-charm energy region, with the highest integrated luminosity at energy points of 3097 MeV, 3686 MeV, and 3773 MeV, which correspond to the production thresholds of $J/\psi, \psi(3686)$, and $D\bar{D}$, respectively. The dataset includes real data at these energy points, with approximately 150 events at each energy point, as well as cosmic ray backgrounds and quantum electrodynamics (QED) processes,

（一）数据集描述：共15类物理过程共计1000余个BESIII探测器事例，包括MC模拟数据和真实数据两部分。

EMC track, EMC hit, MUC track, and MUC hit. Tracks and hits contain a variety of digital information, such as momentum, energy, and detector identifiers. The definitions of digits in text files are stored in the specification document in the dataset named "Parameters_definition.txt". Users can refer to the documents for detailed data information.

4 Recommended repositories to store and find data

With approval from the BESIII collaboration, the dataset has been made public via the Science Data Bank with the link: https://doi.org/10.57760/sciencedb.21486. Users can obtain additional information from the BESIII official website (http://bes3.ihep.ac.cn/).

5 Technical validation

The dataset records a few important physical processes, as well as major background events. They have some remarkable features in terms of visualization. Herein, we present various visualization techniques to demonstrate the validity of the dataset.

The physical processes in the dataset can be visualized using BesVis, as shown in Fig. 3. It has an X-Y cross sectional view of the BESIII detector, where the tracks and hits can be intuitively displayed.

In addition to the two-dimensional (2D) display, a three-dimensional display was also provided using OpenGL, as shown in Fig. 4. This demonstrates a $\Lambda_c^+ \to ne^+\nu_e$ event, which contains four charged tracks and a clear cluster of EMC hits from neutron energy deposits.

Moreover, new visualization techniques have been developed to display these physical processes. For example, unity-based event displays have been used in BESIII [14, 34], JUNO [17], and other experiments [21]. It has the features of local running, multi-platform deployment, and better visualization effects. Figure 5 shows the 2D display of a $J/\psi \to \mu^+\mu^-$ event, where the muons penetrate the

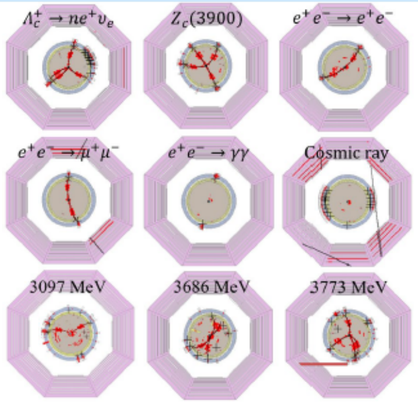


Fig. 3 (Color online) Various physics processes displayed with BesVis in X-Y view. From the inside out, the subdetectors are MDC, TOF, EMC, and MUC

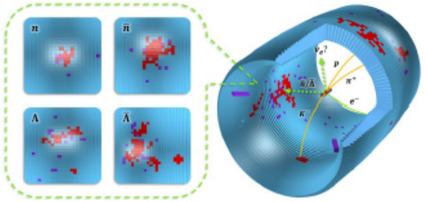


Fig. 4 (Color online) Visualization of a $\Lambda_c^+ \to ne^+\nu_e$ event using OpenGL [30]. The blue blocks represent the EMC crystals, and the red blocks indicate the hits of particles with energy deposited in the EMC

MUC, and a $Z_c(3900)$ event, which shows the approximate ψ-shaped tracks.

（二）数据验证：数据可视化，呈现了几个典型事例的径迹。

范文 *Experimental dataset from BESIII detector at Beijing electron–positron collider*

6 Usage notes

We propose three potential applications with this dataset.

1) The real data contain not only J/ψ, $\psi(3686)$, and $D\bar{D}$ production events, but also continuum backgrounds including QED processes, cosmic ray backgrounds, and other events. The dataset can be used to study the characteristics of different processes.
2) The dataset has full detector response information at the hit level and track level. Both MC simulation data and real data are provided. It can be used to study the detector response in simulation and track reconstruction algorithms for development of data processing techniques in offline software.
3) Visualization plays a crucial role in every aspect of high energy physics experiments. The dataset describes full event information in e^+e^- collisions. Combined with the BESIII detector geometry, it can be used to develop event display software with various new visualization techniques, not only for physics researches, but also for physics education and public outreach.

Author Contributions S.S. Sun and Z.Y. You conceived the experi

（三）数据集使用说明：该数据集可以用来干什么、授权何种使用方式。

PART 4 / IMPORTANCE

发表数据论文的重要性

数据论文已经成为科研出版生态的重要组成部分

01

专业期刊已成体系

国内：中国科学院于今年6月发布Data Express**英文数据刊群**——“1种综合旗舰刊+N种领域专业刊”的顶级数据期刊集群，旨在为全球科学数据交流提供中国枢纽；

02

科研范式正在转变

AI 在科学中的应用日益突出，未来的论文不仅为人类阅读，也将为AI解析。

结构化、机器可读的数据论文，是**AI for Science** 时代科研产出的关键载体。

03

对数据工作者贡献的认可

- ✓ 发布数据论文鼓励数据工作者专注于数据的管理与利用，无需承担额外的科研任务。
- ✓ 让装置用户实现一机时多成果，助力科研增量产出。

让数据本身成为可被引用、可被认证的**独立学术成果**。

国际同类数据期刊

以下三本期刊是当前国际数据论文生态中影响力显著、覆盖学科领域互补的代表性数据期刊

scientific data

开放获取期刊，发表来自自然科学、医学、工程和社会科学各领域的研究数据集描述以及有关研究数据共享的文章。

Scientific Data
IMPACT FACTOR · 2024 **6.9**



Data in Brief

SCOPE: 跨学科短篇数据描述
PUBLISHER: Elsevier
已发表数千篇数据论文



Earth System Science Data

SCOPE: 地球系统观测数据集的描述、发布与质量评估
PUBLISHER: Copernicus Publications

Data Express in Big Sciences

《大科学数据快报》

· 主办单位

中国科学院高能物理研究所

· 合办与出版单位

科学出版社

· 预计2026年底上线

· 期刊宗旨

着眼于大科学装置数据资源的构建与共享，规范数据描述论文的发表，重点聚焦具有科学价值、开放共享价值及实际应用潜力的数据集和研究成果。

期刊将探讨大科学装置数据获取与应用、人工智能的融合、技术创新、质量控制、管理模式及共享环境建设，促进学术交流与成果转化，推动开放科学的发展。



主编团队介绍

主编团队均来自**中国科学院高能物理研究所**，在大科学装置数据领域具有深厚积累。

主 编

EDITOR-IN-CHIEF



陈 刚

中国科学院高能物理研究所**研究员**

国家高能物理科学数据中心 · 主任

ASSOCIATE EDITORS · 副主编



齐法制

中国科学院高能物理研究所 研究员
计算中心 主任



程耀东

中国科学院高能物理研究所 研究员
计算中心 副主任



刘北江

中国科学院高能物理研究所
研究员



郝权

中国科学院高能物理研究所
研究员

创刊时间线

期刊将于2027年正式出版，诚邀来稿！

2026.09

期刊官网及投审稿系统上线

2026.12

累计上线5篇文章

2027.05

累计上线不少于15篇文章

现在投稿，就是最好的时机

- 共建新刊，成为前15篇文章的作者

PART 6 / NEW JOURNAL • 新刊DEBS重磅发布

期刊栏目设计

期刊围绕大科学装置数据全生命周期，共设 **5 个核心栏目**，覆盖从描述、采集到应用的完整链条。

01

Scientific Datasets and Databases
科学数据集及数据库

发表大科学领域数据论文（Data Paper/Data Descriptor），涵盖大科学装置产出的数据集、数据库、数据产品及其质量评估与验证。

02

Data Infrastructure and Computing for Big Science
大科学数据基础设施与计算

数据采集、高速传输、分布式存储、边缘/高性能计算

03

Big Data and AI in Big Science
大科学与人工智能

机器学习、深度学习、科学大模型在大科学数据分析中的应用

04

Data Mining, Management and Standards
数据挖掘、管理与标准

数据治理、元数据标准、FAIR原则、跨装置数据互操作

05

Cybersecurity and Data Governance
数据安全与治理

科学数据安全、隐私保护、数据共享政策、开放科学与伦理

诚邀投稿

Join us in advancing open science & scientific data sharing.

我们诚挚邀请数据工作者、科研人员及大科学装置团队投稿。
期刊将为作者提供以下优势与支持：

01

快速审稿

FAST-TRACK REVIEW

高效的同行评审流程，
缩短从投稿到发表的周期

02

开放获取

OPEN ACCESS

论文向全球读者开放，
提升成果传播力与影响力

03

不收取版面费

NO APC FEE

不收取任何发表费用

期待您的投稿，共建开放科学生态