

新数据分析方法在 BES III 实验的应用*

李 科[†] 苑长征

(中国科学院高能物理研究所 北京 100049)

2025-02-25 收到

[†] email: like@ihep.ac.cn

DOI: 10.7693/wl20250405

CSTR: 32040.14.wl20250405

New data analysis methods at the Beijing Spectrometer III

LI Ke[†] YUAN Chang-Zheng

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

摘 要 自 1974 年发现 J/ψ 粒子至今半个世纪以来, 伴随着加速器技术和探测器技术的发展, 高能物理实验收集的数据在数量和复杂度上都有多个量级的提升。如今北京谱仪 III 实验收集的 J/ψ 粒子超过 10^{10} 个, 比当年丁肇中发现 J 粒子的实验统计量提高了 8 个数量级, 实验数据的分析方法在先进计算技术和算法的加持下也经历了重大变革。从传统的统计方法到决策树和深度学习, 研究人员不断探索更高效的方式从海量数据中快速、精确地提取物理信息。在北京谱仪 III 实验中, 多个多变量分析和机器学习模型被用于探测器模拟、径迹重建、粒子鉴别和事例挑选等, 显著提高了实验灵敏度和效率。近几年大语言模型展现出的强大的文本和代码生成能力, 为自动化、智能化数据分析提供了可能。基于这一理念, 研究人员开发了“赛博士”(Dr. SAI)智能体系统, 用于提升高能物理实验数据的分析效率和获取物理结果的速度, 这种变革将对高能物理的研究带来深刻的影响, 并有可能引起科研范式的改变。

关键词 统计分析, 多变量分析, 机器学习, 人工智能

Abstract Since the discovery of the J/ψ particle in 1974, the past half-century has witnessed remarkable advancements in accelerator and detector technology, leading to a many orders-of-magnitude increase in both the volume and complexity of data collected in high-energy physics experiments. Today, experiments at the Beijing Spectrometer III (BES III) have accumulated more than 10 billion J/ψ events, marking an 8 orders-of-magnitude improvement in statistics compared with the original J/ψ discovery by Samuel C.C. Ting. Concurrently, the methodologies for analyzing experimental data have undergone significant evolution, driven by breakthroughs in computational technologies and algorithms. From traditional statistical methods to boosted decision trees and deep learning, researchers have continuously explored more efficient ways to rapidly and accurately extract physical information from the large datasets. Within the BES III experiment, advanced multivariate analysis and machine learning models have been employed for detector simulation, charged track reconstruction, particle identification, and event selection, significantly enhancing experimental sensitivity and efficiency. In recent years, the emergence of large language models with their powerful text and code generation capabilities has opened new possibilities for automated physics analysis. Inspired by this, researchers have developed the AI agent “Dr. SAI” to accelerate the extraction of physical results in high-energy physics experiments. This approach is expected to make a substantial impact on high-energy physics research and may potentially usher in a transformative paradigm shift in scientific discovery.

Keywords statistical analysis, multivariable analysis, machine learning, artificial intelligence

1 引言

1933年6月10日爱因斯坦在英国牛津大学所作的斯宾塞(Herbert Spencer)讲座中^[1]说:“纯粹的逻辑思维不能给我们任何关于经验世界的知识;一切关于实在的知识,都是从经验开始,又终结于经验。用纯粹逻辑方法所得到的命题,对于实在来说是完全空洞的。”自伽利略开始的科学家即通过大量实验数据的积累与分析,从中提炼科学规律,以验证和拓展人类对自然的认知。作为所有科学基础的物理学,一直走在科学发现的前沿。在过去的一个世纪里,物理学的发展与计算技术的进步相辅相成,也推动了数据分析方法的革命性进步。

在所有物理实验中,高能物理实验是最为复杂的。其核心目标是研究构成万物的基本单元——粒子——本身的性质和它们之间的相互作用。然而,这些微观粒子无法被人类肉眼直接观测到,科学家必须借助一系列精密的实验装置来观察和测量它们。这些装置被称为探测器,它们的功能类似于一个显微镜和照相机的组合,能够“放大”粒子与探测器物质之间的相互作用,并精确记录相应信息。这些实验数据正是理解微观世界的关键。科学家通过数据分析,从中提取出粒子的产生和衰变性质,并进一步总结出支配这些行为的深层次物理规律。目前世界上最成功的基础理论——粒子物理的标准模型——正是通过这种方式建立起来的。

2 传统的数据分析方法

粒子物理探测器通常比较庞大且由多种不同的子探测器构成,以测量不同类型粒子的信息。北京谱仪Ⅲ(BESⅢ)探测器^[2]作为一个典型的运行在正负电子对撞机上的磁谱仪,长11 m、宽6 m、高9 m,总重700多吨,由主漂移室、飞行时间计数器、电磁量能器、缪子探测器等子探测器,以

及超导螺线管磁铁、读出电子学系统、触发判选系统、数据获取系统、控制监视系统、离线数据处理等部分组成。

粒子物理探测器读出的信号通常都是电信号,如电脉冲的幅度、谱形、时间等。BESⅢ探测器共有三万多道读出电子学,每秒钟记录最高达5000个事例,实时数据量每秒50 MB。数据处理的第一步是将这些电信号收集并校准成更具有物理意义的信号,如电离能损、沉积能量等。每一个物理事例本质上都是由这些信号以及相应子探测器的几何信息共同组成的。接下来需要进行的就重建过程,即将收集到的信号转化为实际的物理量,如带电粒子的运动轨迹、动量、能量、粒子种类等。

得到这些数据后,怎样提取出有用的物理信息是让科学家头疼的难题。由于粒子物理实验装置复杂,背后的物理过程繁多,收集的数据量巨大,从而很难从海量数据中找到人们感兴趣的事例。丁肇中先生曾将粒子物理实验数据分析难度比喻为在一个大如波士顿市的地方从无数的雨滴中找到一个特定的雨滴^[3],难度无异于大海捞针。在1974年发现 J/ψ 粒子时期,科学家已经开始使用计算机辅助分析海量的数据,但使用的方法相对比较简单,即依据以前的经验来定义信号事例的各种选择条件,从而挑选出最独特的那个雨滴。这种方法虽在不断进步,但核心思想一直延续至今,仍是所有高能物理实验普遍采用的模式,BESⅢ实验自2009年运行开始也大致采用这种数据分析方式^[2]。这种方式的典型效率是一个研究生经过一年的专业课程学习进入实验室工作,经过大约半年的时间学习和掌握相关计算机操作和软件分析环境以及基础数据分析训练,然后在两到三年内完成一个物理过程的分析,撰写论文并发表结果。这样一篇论文的内容和结果也就是通常一个博士学位论文的主要内容。

3 发展的数据分析方法

在早期的高能物理实验中,由于计算机技术尚未广泛应用,数据分析更多依赖于研究人员的

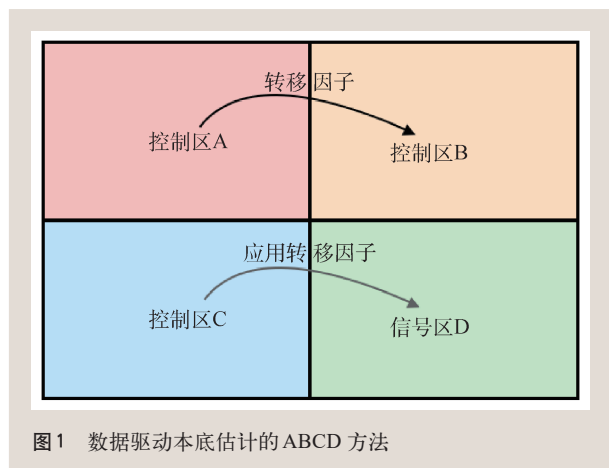
* 国家自然科学基金(批准号:12361141819)资助项目

经验来完成,包括事例挑选、本底分析、物理参数测量、探测效率估计等。随着计算机技术的发展,蒙特卡罗方法被引入数据分析中,用于模拟特定物理过程以更好地优化事例选择、估计本底和效率等。尽管这一方法在很多情况下有效,但由于该方法强烈依赖于人们对粒子与探测器物质相互作用和探测器几何结构的理解和描述,往往不能提供足够精确的预测,特别是会影响对本底的估算和效率的确定等。这就需要基于真实数据对蒙特卡罗方法进行修正,为此发展了数据驱动的本底估计方法以更好估计本底事例的贡献^[4]以及基于数据的效率测量。为了更有效地利用多个子探测器的信息,发展了基于多变量分析或深度学习的事例选择方法,以提高分析的灵敏度,从而可以更高效地利用数据得到物理结果。

3.1 数据驱动的本底估计方法

这是目前物理研究中广泛采用的一种方法^[4]。具体的做法是寻找一个与目标观测量无关的观测量,利用它来定义多个控制样本。由于这些控制样本与目标观测量无关,因此它们之间的关系应与在目标观测量上定义的样本相同。基于这一假设,人们可以从数据中估计出信号区间的本底事例贡献,一个经典的例子是ABCD方法,它通常用于通过控制区间来估计信号区的本底事例。

该方法的基本思路是选择两个不相关的变量 X 和 Y 用于定义4个区间,如图1所示,其中变量 X 分布在A和B区间,变量 Y 分布在C和D区间。



区间D代表信号区,其余三个为控制区间。通过适当的选择条件,可以合理假设几乎所有的信号事例都存在于区间D,而非区间A、B和C。由于 X 和 Y 两个变量的分布不相关,可以预期在真实的本底事例中,A区间与B区间的比值和C区间与D区间的比值一致,从而可以通过ABC三个控制区间估计D区间的本底事例贡献。如果 X 和 Y 之间存在相关性,则在用ABC估计D区间时需要修正这些相关性引起的偏差,这些修正通常通过蒙特卡罗模拟获得。

与传统的蒙特卡罗方法相比,数据驱动的方法使用了更少的先验知识和模拟信息,因此对本底的估计更为可靠。除了ABCD方法外,还有许多其他方法,但其基本思路都是通过构建几乎只包含本底事例的多个控制样本,并利用这些控制样本之间的关联来估计信号区的本底事例。

3.2 多变量分析方法

随着物理研究的深入,人们关注的物理信号在总事例样本中的比例越来越小,而且与本底事例的区分越来越不明显,通常很难仅通过单一观测量就从大量本底事例中挑选出合适的信号事例。对于每一个事例,人们通常会选择多个变量。首先,选取一个变量并设定一个阈值,根据这个阈值可以判断该事例是信号还是本底;随后,再选择另一个变量重复这一过程。阈值的设定以及变量的顺序是通过带有标记的样本进行训练确定的,目的是尽可能将信号和本底分开。由于这些变量之间往往存在关联,上面的在多个观测量上依次作选择挑选信号的方法会造成效率的快速下降。为解决这个问题,可以采用综合多个变量的方法,即多变量分析方法^[5]进行事例选择。在高能物理实验中,决策树(boosted decision tree, BDT)就是一种常用的多变量分析方法。

BESⅢ实验目前已经积累了世界上最大、最干净的粲介子样本,用以精确测量粲介子的性质。粲介子在正负电子湮灭过程中成对产生,主要通过弱相互作用衰变。在很多感兴趣的粲介子衰变中会产生中微子,而中微子几乎不与探测器发生

相互作用，因此无法被探测器直接收集。这使得挑选此类事例变得十分困难，尤其是在粲介子衰变末态中包含多个中微子的情况下，误组合的本底事例往往远多于信号事例。

在 $D_s^- \rightarrow \tau^- \nu$ 事例中，既包含直接衰变产生的中微子，又包含 τ^- 衰变产生的中微子，BESⅢ实验最近采用了BDT方法对其进行了选择和迄今为止最精确的测量^[6]。分析综合应用了标记侧、信号侧和完整事例的9个变量进行信号和本底区分。如图2所示，与传统方法相比，BDT方法能够更高效地挑选出信号事例，从而显著提高了测量的精度。

3.3 深度学习

决策树等多变量分析方法通过结合多个变量，能够提供更好的信号与本底事例的区分能力。然而，随着统计量的增加以及各种非线性效应的影响，BDT逐渐表现出局限性。为了应对这些挑战，研究人员转向了更为高效的深度学习方法如卷积神经网络(CNN)和图神经网络(GNN)，它们在分类和鉴别等问题上相比传统的BDT表现得更加优越。

在BESⅢ实验中，中子鉴别一直是一个巨大挑战。由于中子不带电荷，无法在内层探测器中留下痕迹，且在电磁量能器中只能沉积少量能量，因此与其他中性粒子(如 K_L^0 或由 Λ 粒子衰变产生的中子和 π^0)很难区分。最近，研究人员利用图神经网络的方法，基于从数据中获得的样本(如 $J/\psi \rightarrow \bar{p}\pi^+n$ 反应)训练中子鉴别器。具体来说，研究人员将电磁量能器中各个晶体的观测量(如位置、沉积能量等)及它们之间的关联构建为一个二维图。这类图包含了不同粒子与电磁量能器相互作用的信息，用于判断粒子的种类。

为了更好地处理真实数据与蒙特卡罗模拟之间的差异，研究团队选择了另一个样本，对蒙特卡罗模拟进行校准，使其与实际数据更加匹配。通过这一方法，图神经网络在中子鉴别上的性能

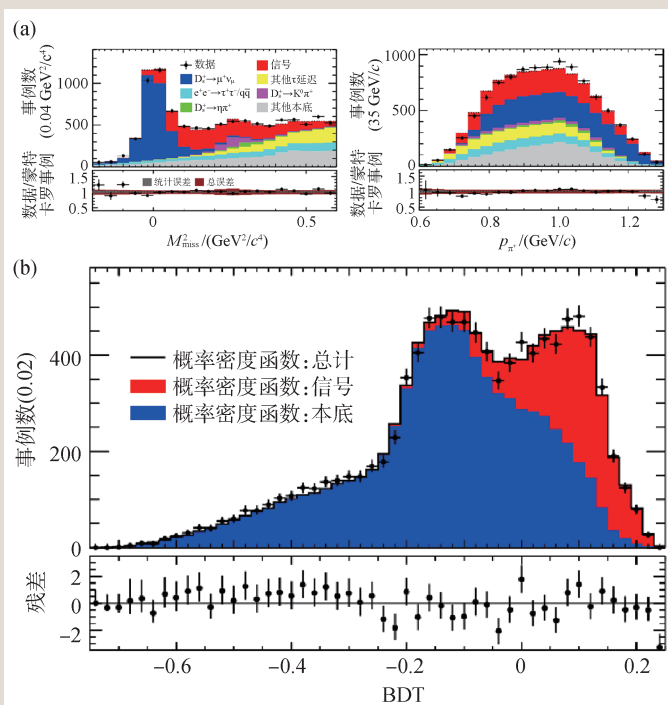


图2 BESⅢ实验测量 $D_s^- \rightarrow \tau^- \nu$ 过程的分支比 (a)丢失质量的平方和 π 粒子动量沿 x 轴的分布; (b)用BDT综合各个变量得出的结果,在此分布上信号和本底可以更好区分

得到了显著提升，得益于此，人们首次发现了 Λ_c^+ 到中子的 β 衰变^[7]。图3展示了BESⅢ探测器中观测到的中子、反中子、 Λ 超子及反 Λ 超子在量能器中的击中信息。

4 基于人工智能的未来数据分析模式

近年来，Transformer模型的诞生为机器学习领域带来了革命性的进展，尤其是在基于该模型构建的预训练通用大语言模型(如 ChatGPT、DeepSeek等)方面。这些模型通过训练海量的文本数据，具备了根据给定的文本预测后续的文本内容，在文本和代码生成方面表现出可与人类媲美

的能力。高能物理实验的数据处理极其复杂，不仅涉及庞大的数据量，且处理过程繁琐复杂，许多物理分析步骤存在重复性。大部分任务都涉及文本或代码生成，或可以转换为此类任务。因此，可以引入大语言模型来自动化完成高能物理实验的数据分析。基于此理念，BESⅢ实验开发了“赛

博士”(Dr.SAI)AI智能体^[8]系统(整体结构如图4所示),用于BESⅢ实验的数据处理。

“赛博士”的核心是通用大语言模型。研究团队在开源的大模型(如LLaMa3)基础上,结合高能物理领域的数据进行了二次预训练和微调,开发了一个专注于该领域的大模型——溪悟^[9],并基于溪悟,进一步开发了一系列工具,以提升AI结果的准确性。为了缓解AI生成内容时出现幻觉的问题,BESⅢ实验的内部数据,包括原始文档、专家交流记录以及用于数据处理的源代码等,都被存储在一个矢量库中,并应用增强检索技术进行管理。此外,还针对不同类型的任务开发了多个智能体,并引入了高能物理实验中的多级验证系统,对每一步的结果都实时地进行检验,以确保AI生成的结果在可预见的范围内是正确的。

这些智能体的基础是大语言模型,如溪悟或者国际上最流行的GPT、DeepSeek等,它们是“赛博士”的大脑;跟BESⅢ实验和高能物理相关的知识就存储在记忆器,也就是前文所提的矢量库中;在此基础上开发的多个智能体,用于分解大的任务到若干个子任务,用于代码生成,用于

在特定计算环境内执行程序,以及用于内部的对话和智能搜索等,多个智能体协同工作以解决实际数据分析中的复杂问题。

目前,已正式发布了“赛博士”的第一个版本^[8],其可在接收到人类清晰指令的前提下针对简单任务自动分解为更小的子任务、自动产生代码、自动运行并检查结果。研究人员正在利用最先进的AI技术对其不断升级,相信在不久的将来,随着其性能的不不断提升,研究人员将能够借助“赛博士”显著提高在BESⅢ物理分析中的科研效率。

以“赛博士”为基础进一步发展“赛教授”AI智能体系统将更广泛地应用于高能物理的数据分析以及物理测量结果的理论解析中。在这个系统中,AI作为替代模型将实现多维数据的映射,并应用于径迹重建、粒子鉴别、事例分类、异常检测、快速模拟等任务,显著增强科研人员的研究能力。BESⅢ实验将进一步开发事例解析算法,通过机器学习自动识别事例衰变链。该算法不再关注单一粒子或单一衰变模式的挑选,而是尝试解释整个事件的所有可能衰变链。这种算法的应用

将不仅提高数据分析的效率和对各种物理过程的研究能力,还将极大促进BESⅢ实验寻找新物理现象和精确测量物理参数等方面的科研产出。粒子物理研究将从“积累单个过程测量结果推测物理规律”跃升到“从大量过程的测量结果总结物理规律”的阶段,并可以快速实现对于理论预言过程的测量和检验,极大加速实验—分析—理论的反馈循环,高效促进基础物理的探索和发展。

5 小结

几十年来,高能物理实验发生了巨大的飞跃。加速器和探测

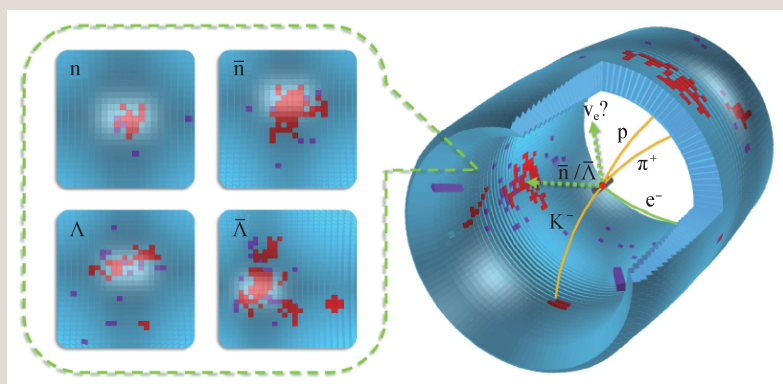


图3 BESⅢ实验电磁量能器中的中子、反中子、 Λ 超子及反 Λ 超子击中信息,颜色深浅代表沉积能量的大小

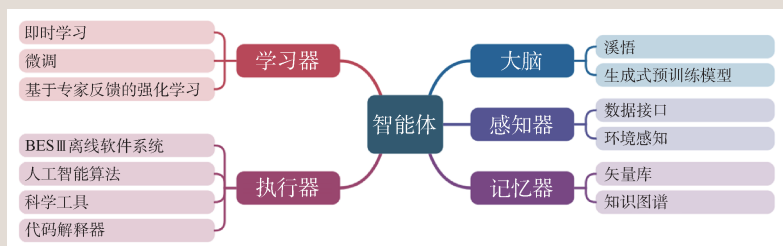


图4 “赛博士”AI智能体的整体结构图

器技术的显著提升使得实验数据迅猛增长。仅仅 50 年的时间就实现了从最初 BNL-E598 实验发现的 242 个发展到如今 BESⅢ实验收集的超过百亿个 J/ψ 粒子的跃升,数据量的增长跨越了 8 个量级,其他高能物理实验的数据也有类似的增长。

为了应对这种巨大的数据增长,许多新的数据分析方法应运而生,并在解决实验数据处理的问题上发挥了至关重要的作用。BESⅢ作为精确测量前沿的高能物理实验,在数据分析方法的改进上做出了多项尝试并取得了重要进展。“赛博士”AI 智能体的发布宣告基于人工智能的高能物理数据分析时代拉开序幕。展望未来,随着数据的不断深入挖掘和物理规律的总结,人们有望突破当前的科学瓶颈。或许有一天,在人工智能的帮助下,人们可以更高效、更智能地从海量数据中发现新的物理规律,为人类知识的拓展作出新的贡献。

致谢 感谢 BESⅢ合作组以及“赛博士”AI 智能体研究团队在本文内容研究中做出的重要贡献。

参考文献

- [1] 许良英等. 爱因斯坦文集(第一卷). 北京:商务印书馆,1976. p.312
- [2] Ablikim M *et al* (BESⅢ Collaboration). Nucl. Instrum. Meth. A, 2010, 614:345
- [3] 陈洁琦, 淳麟. 寻找带颜色的雨滴——丁肇中的科学风采. 上海:上海科技教育出版社,2002
- [4] Hatfield P W, Gaffney J A, Anderson G J *et al*. Nature, 2021, 593:351
- [5] Hoecker A, Speckmayer P, Stelzer J *et al*. 2007, arXiv: physics/0703039
- [6] Ablikim M *et al* (BESⅢ Collaboration). Phys. Rev. D, 2023, 108:092014
- [7] Ablikim M *et al* (BESⅢ Collaboration). Nature Communications, 2025, 16:681
- [8] Li K. AI agent for BESⅢ——Dr. SAI. 2024 量子计算和机器学习研讨会. 长春. 2024 年 8 月 2—8 日
- [9] Zhang Z, Zhang Y, Yao H *et al*. 2024, arXiv:2404.08001

**CIOE**
中国光博会

光学全产业链生态链展示及交流平台

2025年9月10-12日
深圳国际会展中心(宝安)

精密光学展&摄像头技术及应用展
PRECISION OPTICS EXPO & CAMERA EXPO

同期展会

信息通信展
INFORMATION COMMUNICATIONS EXPO

激光技术及智能制造展
LASER TECHNOLOGY & INTELLIGENT MANUFACTURING EXPO

红外技术及应用展
INFRARED APPLICATIONS EXPO

智能传感展
INTELLIGENT SENSING EXPO

光电子创新展
PHOTONICS INNOVATION EXPO

AR&VR展
AR&VR EXPO

新型显示技术展
DISPLAY TECHNOLOGY EXPO

第26届中国国际光电博览会

240,000M²展出面积
130,000+ 专业观众
3,800+ 参展企业



扫码登记 免费参观