arXiv:2512.15230v1 [hep-ex] 17 Dec 2025

ColliderML：OpenDataDetector 高亮度物理基准数据集首个版本

Doˇga Elitez欧洲核子研究中心与美因茨约翰内斯·古腾堡大学

Paul Gessinger欧洲核子研究中心

Daniel Murnane*
尼尔斯·玻尔研究所
与劳伦斯伯克利国家实验室

Marcus Selchou Raaholt
哥本哈根大学

Andreas Salzburger
CERN

Stine Kofoed Skov
哥本哈根大学

Andreas Stefl欧洲核子研究中心

Anna Zaborowska
CERN

摘要

我们介绍**ColliderML**[1] ——一个大型、开放且与实验无关的数据集，包含在高亮度大型强子对撞机条件下（$\sqrt{s} = 14$太电子伏，平均堆积 $\mu = 200$）完全模拟并数字化的质子-质子碰撞。COLLIDERML 提供了涵盖十种标准模型和超越标准模型过程的 100 万个事件，以及丰富的单粒子样本。所有数据均采用现代的次领先阶矩阵元计算和簇射、逼真的逐事件堆积叠加、经过验证的 OpenDataDetector 几何结构以及标准重建流程生成。该数据集通过适用于机器学习的 Hugging Face 平台发布，填补了机器学习（ML）研究在探测器级数据方面的一项重大空白。我们介绍其物理覆盖范围，以及生成、模拟、数字化和重建流水线；说明数据格式和获取方式，并展示初步的对撞机物理基准测试。

1 引言

粒子物理是一个亟需高质量模拟的领域，以匹配大型强子对撞机（LHC）等对撞机实验中数据采集的精度。比较观测到的行为与预测结果，或许能揭示新现象，但差异也可能源于模拟或数据采集过程中的各种误差。因此，尽可能真实地模拟预期物理过程和探测器本身（包括可能的错误校准），通常是发现或排除超出标准模型（BSM）物理的关键一步。总体而言，对撞机物理模拟分为两类：所谓的“全”模拟和“快速”模拟。两种模拟都从预测所有即时粒子产生和衰变的概率开始，其中包括强子化等效应。随后，在全模拟中，物理引擎会逐步模拟每个碰撞产物与周围探测器材料的相互作用，并计算散射、测量以及进一步衰变或产生粒子的影响。接下来，为了获得最高保真度的数据，必须将这种“真值”模拟数字化，以生成更真实的探测器响应，其中还包括电子噪声、错位等可能性。

*通讯作者：daniel.murnane@nbi.ku.dk
[1]文档：https://opendatadetector.github.io/ColliderML/，数据集：https://huggingface.co/datasets/CERN/ColliderML-Release-1

预印本。正在审稿。

设备、给定时间窗口内可测量能量范围的限制、电荷漂移等。

全模拟的计算开销很大，因此在许多情况下，可以容忍一定程度的保真度损失，转而采用快速模拟。加速粒子模拟有许多颇具前景的方法，例如使用基于全模拟训练的生成式机器学习模型。一种常见做法是将全模拟中的复杂相互作用参数化为一组可应用于碰撞产物的函数。例如，可以用一条展宽螺旋轨迹，替代带电粒子在磁场中的物质内传播时所经历的随机且嘈杂的路径。随后，可将螺旋线叠加到探测器上，并快速生成各种测量结果或下游相互作用。通过这种方式，可以提供海量快速模拟数据集，从而支持使用数千万乃至数十亿个样本训练大型机器学习模型。然而，使用快速模拟可能需要付出代价：机器学习模型训练所依据的是参数化结果，而非真实的模拟。对于这类模型检测新物理微弱迹象的能力，这种权衡会产生何种影响，至今尚不明确。为此，我们提出一个新的全模拟数据集 COLLIDERML，旨在发挥以下三种作用：

1. 一个涵盖各个尺度物理对象的资源库，从生成的部分子到物质中的能量沉积，再到重建径迹和喷注。这些数据经过逼真的数字化处理，可用于研究对撞机工作流各阶段重建所采用的机器学习辅助算法。我们特别指出，多阶段或多任务模型、异常检测模型和基础模型尤其适用于 COLLIDERML 数据集。2. 一个探索下一代探测器可能性的试验平台。COLLIDERML 事件规模庞大且噪声较多，探测器组件取自多种拟议中的未来技术（例如高磁场、高颗粒度量热、像素和条纹内径迹器）。为此，我们提供多种标准模型和超出标准模型道，包含完全模拟的（且不重复的）质子碰撞叠加，以及多种错误校准情况。3. 一种研究在快速模拟上训练模型所产生影响的方式。特别是，记忆、泛化性和对称性保持等某些行为，究竟是这些简化数据集所特有的，还是所有物理模拟的一般属性?

我们公开发布了四个子数据集，合称 COLLIDERML：涵盖十种产生道的逾一百万个经完整模拟的高光度碰撞事件；相同产生道的千万个硬散射事件；千万个单粒子模拟；以及三十万个经历了各种错误校准和其他严苛条件的事件。最终，所有事件都将在模拟与重建链中包含多个层级的数据，从初态粒子、击中到分析层级对象。本文聚焦于全叠加产生道的较低层级对象，这些对象将在 COLLIDERML 第1版中提供，目前大部分已可在 Hugging Face [1] 上获取。第2版和第3版将包含其余重建对象以及错误校准条件。发布计划见图 1。

本文详细介绍 COLLIDERML 的制作过程，以及为实现上述目标所作的选择。我们还提出了若干指标，可用于规范跟踪任务的报告方式。第 3 节介绍部分子生成以及粒子轨迹和衰变模拟所用的技术。第 4 节介绍内层探测器和量能器的数字化方法。第 5 节介绍径迹重建。最后，第 6 节概述面向用户的 Parquet 文件结构，以及数据的访问方式。

2 相关工作

对撞机模拟通常分为完整 模拟（基于 Geant4 的 [2, 3] 输运，可能包含真实数字化）或快速 模拟（参数化响应或学习型替代模型）。快速方法可扩展至10^7–10^8 个样本，但会用简化函数（例如参数化量能器响应）替代随机输运。这种速度与保真度之间的权衡，使大规模现代机器学习预训练和基准测试成为可能，但简化处理在多大程度上会造成学习和下游推理偏差，仍有待探究。具体而言，参数化越简化，细粒度效应和对象就越不真实。Delphes [4]是一种广受欢迎的参数化快速模拟方法。尽管近期发展已包括对对象进行更精细的处理，直至径迹级（例如协方差），但这种方法仍无法生成真实的径迹探测器响应和量能器能量沉积 [5]。

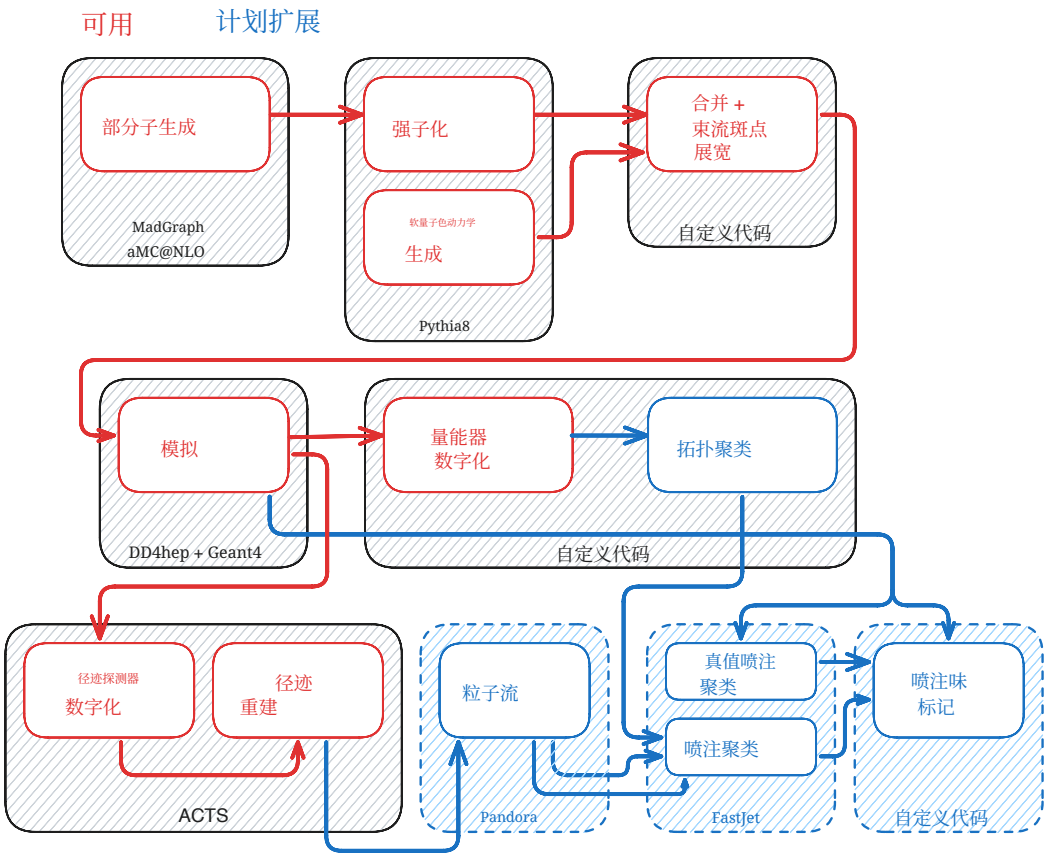


图 1：用于生成 COLLIDERML 及各版本中对象的技术

除了人工调优的参数化方法，学习型替代模型还利用生成模型加速子探测器模拟（例如用于电磁/强子簇射的*CaloGAN* [6]；ATLAS/CMS 快速模拟工具链中的快速量能器变体 [7, 8]）。这些方法以精确输运为代价，换取分布保真度和速度；它们通常能匹配低阶矩，但难以处理长尾、罕见拓扑或高粒度几何效应——而这些恰恰是下游机器学习敏感的情形。与此同时，学习型推理 框架利用分类器校正模拟与数据之间的差异，或展开探测器效应（例如*OmniFold*的迭代式仿真感知重加权 [9], 及相关的基于分类器的似然比估计 [10, 11]）。这些进展共同表明，机器学习既能替代模拟，也能修正模拟。

对撞机物理数据集拥有丰富的生态系统，涵盖分析（事例级统计）、高（粒子流对象）、中（径迹和量能器簇）以及低（探测器击中/体素）等抽象层级。我们在图 2 中总结了具有代表性的数据集及其规模。在高层级上，JetNet 和 JetClass[12] 提供用于监督学习和生成式建模的 10^6–10^8 喷注样本；ATLAS 顶夸克标记开放数据扩展了监督学习基准；SUSY（UCI）则提供经典的表格型高层特征。低层级资源包括 TrackML [13, 14]（硅径迹追踪）以及 CaloChallenge 的电磁/强子簇射 [15]。

为何选择COLLIDERML 大多数公开数据集要么只提供高层／中层抽象，要么只在受控条件下展示单个子探测器；很少有数据集能在同一处提供全仿真、全探测器事件，并包含真实且不重复的堆积、错误校准以及所有中间对象。快速模拟数据集的规模也促使人们训练超大型模型，而当训练样本少于 $\mathcal{O}(10^6)$ 时，可能会出现过拟合，令泛化能力的主张变得模糊。COLLIDERML 以类似高亮度大型强子对撞机的多尺度环境和可调节的真实性参数，针对这些不足，旨在研究全模拟条件下的记忆、归纳偏置和对称性保持。

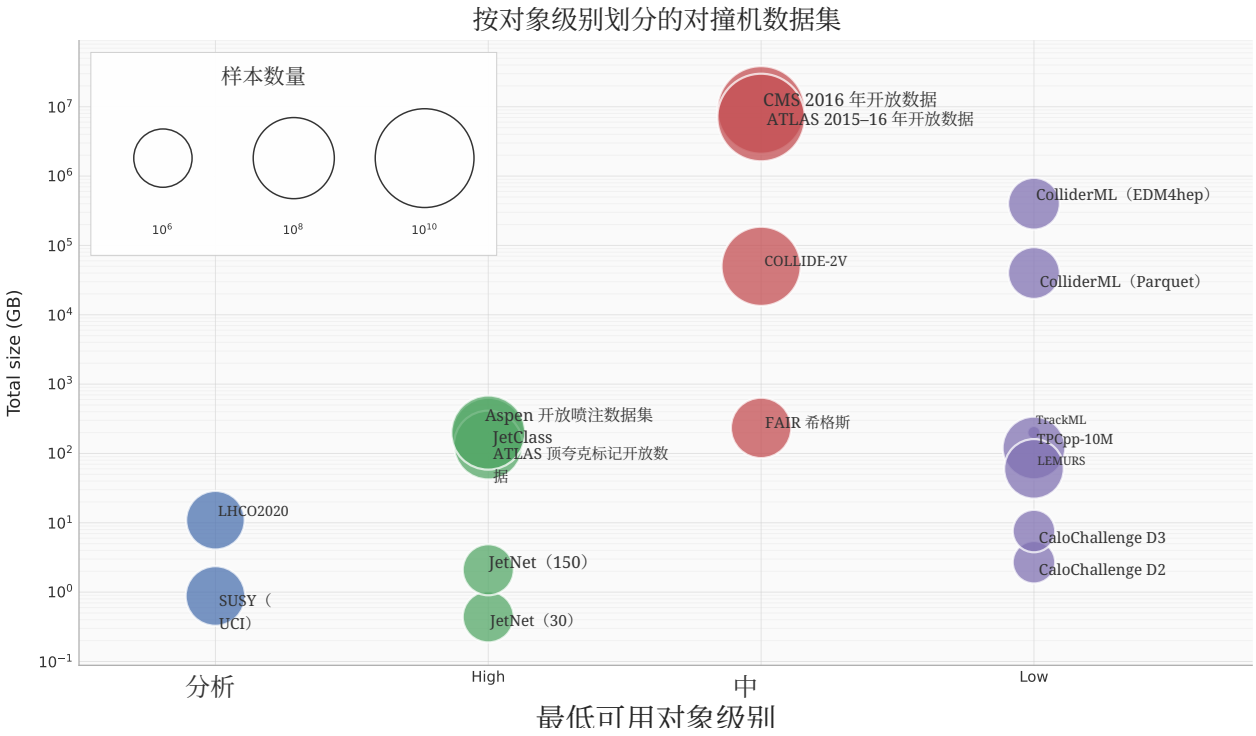


图2：COLLIDERML 与其他对撞机数据集的比较。该可视化图表背后的数值见 B 节。

具体而言，我们预计 ColliderML 非常适合测试基础模型在细粒度对象与高层对象之间的泛化能力，而以往模型只关注低层对象[16] 或高层对象 [17–19]。它应当足够真实，可用于研究机器学习算法在径迹重建和聚类 [20–28], 、快速模拟 [29–31] 以及粒子流 [32, 33] 方面的扩展性，且条件与大型强子对撞机探测器第二阶段升级的预期条件非常相似。最后，它还可用于考察物理学驱动模型；这类模型历来侧重于在快速模拟 [34–39]中进行测试。

与以往工作相比，COLLIDERML 旨在提供：(i) 每个事件的所有对象层级数据均可联合使用，以支持跨阶段研究；(ii) 逼真的堆积和探测器效应（未来版本还将包含错误校准），适用于对重建和机器学习进行压力测试；以及 (iii) 一项发布计划，先发布低层级对象，再发布高层级重建结果，以支持社区基线和基础模型预训练。

3 模拟

3.1 包含的物理过程

在 COLLIDERML（高堆积）的首个版本中，我们开放了表 1 所列的十种产生道，并简要说明了选择各产生道的原因。具体而言，我们纳入了有助于精确研究电弱扇区、希格斯扇区和顶夸克扇区中耦合与质量的标准模型产生道。我们还纳入了一种喷注占主导的堆积道，用于研究重建算法的性能。对于超出标准模型物理，我们选取了三种不同的信号特征，分别对应机器学习在重建和异常检测方面的关键挑战。首先，我们纳入了一种规范介导超对称破缺模型（GMSB）[40] ，其中长寿命中性ino次轻超对称粒子（NLSP）衰变为引力微子暗物质候选粒子（寿命为 $c\tau \approx 100$ 毫米）。该模型产生位移顶点和缺失能量信号，可作为新物理搜索及非标准径迹重建的目标，例如 ATLAS探测器 [41]在 Run 3 中开展的位移轻子搜索。

其次，我们纳入了隐藏谷（暗量子色动力学）产生道 [42], ，将其 Z' 建模为一个衰变至暗扇区的中介粒子，其中包含稳定和不稳定暗强子。由此产生半可见喷注——这类物体具有局域缺失能量和复杂子结构，可用作异常检测的基准示例 [43]。最后，我们纳入了理论依据充分的 Z′重共振模型 [44] ，用于对标准监督式峰搜索和分辨率研究进行基准测试。这些奇异信号是电子离子对撞机和未来环形对撞机等拟议未来设施物理计划的核心内容 [45–47]。

ID	过程标签	过程类型	备注
1	顶夸克对产生	$pp \to t\bar{t}$	顶夸克产生；对标准模型检验和复杂末态 ·
2	zee	$pp \to Z \to e^+e^-$	Drell-Yan过程；探测器校准和电弱探针。
3	Z玻色子衰变为μ子对	$pp \to Z \to \mu^+\mu^-$	Drell-Yan过程；探测器校准和径迹探针。
4	双希格斯	$gg \to HH$	胶子融合双希格斯；高光度基准研究 ·
5	双光子	$pp \to \gamma\gamma$	双光子产生；圈诱导QCD/电弱过程。
6	多喷注	$pp \to$ 喷注	QCD产生；强子信号的主导本底-性质 ·
7	ggf	$pp \to H$	胶子-胶子聚变；主要的希格斯玻色子产生道。
8	susy	$pp \to \tilde{g}\tilde{g}$	GMSB 胶微子产生； $\tilde{\chi}_1^0$ NLSP 提供位移顶点基准。
9	Z′	$pp \to Z'_{SSM}$	重共振；用于隆起搜寻的标准烛光。
10	隐藏谷	$pp \to Z' \to$ 暗	暗量子色动力学扇区；用于异常检测的半可见喷注。

表 1：COLLIDERML 中产生道列表（高堆积）。超出标准模型道（8–10）旨在代表不同的重建挑战。

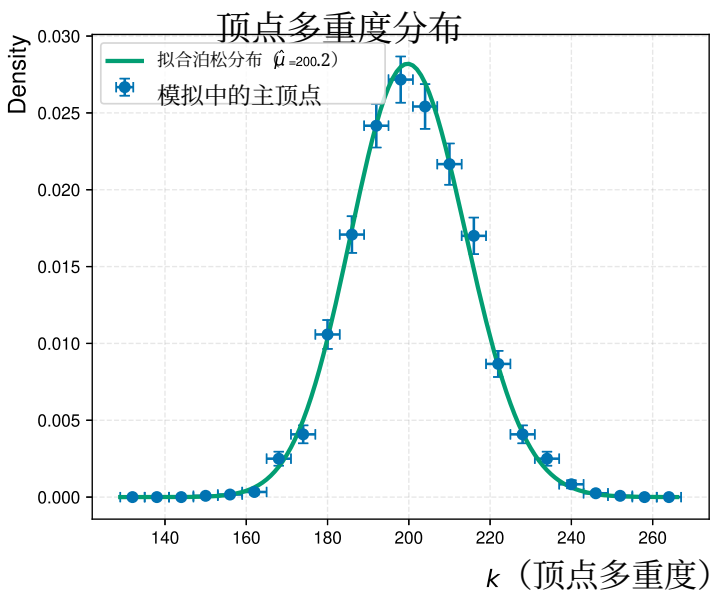
还包括仅硬散射样本（即无堆积）和粒子枪道，其中单粒子以均匀采样的初始方向、在设定的能级上发射。在这些单粒子道中，我们提供单个 $\mu^{\pm}$、单个 $e^{\pm}$、单个 $\pi^{\pm}$、单个 $K^{\pm}$ 和单个 γ，涵盖多个离散和连续能量区间。

硬散射矩阵元计算采用当前最先进的方法：使用 Madgraph aMC@NLO v3.5.8 [48] 对主要过程及前两个喷注进行次领头阶（NLO）计算。在 Madgraph 中，圈诱导过程（胶子融合和双希格斯）按领头阶建模。簇射由 Pythia8.313 [49], 提供；NLO 过程采用 FxFx匹配，领头阶过程采用 CKKW 匹配。超出标准模型道在 Pythia8 内原生建模，并按领头阶进行簇射。

我们特别谨慎地指出，矢量玻色子聚变与散射在上文使用的默认 aMC@NLO→ Pythia8 流水线设置中建模并不完善 [50, 51]。因此，我们避开了涉及这些现象的产生道，例如排除 $VV \to HH$，仅允许 $gg \to HH$ 作为双希格斯产生道[2]。

堆积生成。 在 ATLAS [52, 53] 和 CMS [54] 实验的第二阶段升级中，每次束团交叉预计将发生 200 次质子-质子碰撞。其中大多数是软量子色动力学散射，对大多数物理分析并不重要。不过，为了从本底中区分出感兴趣的硬散射碰撞，必须对这些散射进行建模和模拟。合作组采用的典型工作流程是：生成一个大型的堆积模拟池 P，抽取 $N \ll P$ 所谓的堆积“子事件”，并将其与一个硬散射子事件结合，生成

[2]我们注意到，VBS 和 VBF 产生道值得纳入，未来将发布这些数据；届时会使用 POWHEG-BOX 进行部分子生成，并在簇射时采用 Pythia 的偶极反冲方案。研究表明，这种做法与测量结果的符合程度要高得多 [51]。



Figu图 3：在 20 次事件中，主顶点的数量（单次硬散射和 N 个堆积子事件）等。拟合计 60 个事件数的泊松曲线验证了事件符合预期分布。

一个完整的堆积事件。这种方法在计算上效率很高，但可能导致事件内容大量重叠。对于只考虑分析层级对象的数据集，这或许不是重要影响；但在低层级上，这会导致同一批径迹探测器击中和量能器贡献在一个数据集中重复出现数十次甚至数百次。虽然尚未通过详尽研究证明这种重复内容一定会被记忆，我们仍谨慎行事，以避免 COLLIDERML 中出现这种可能性。

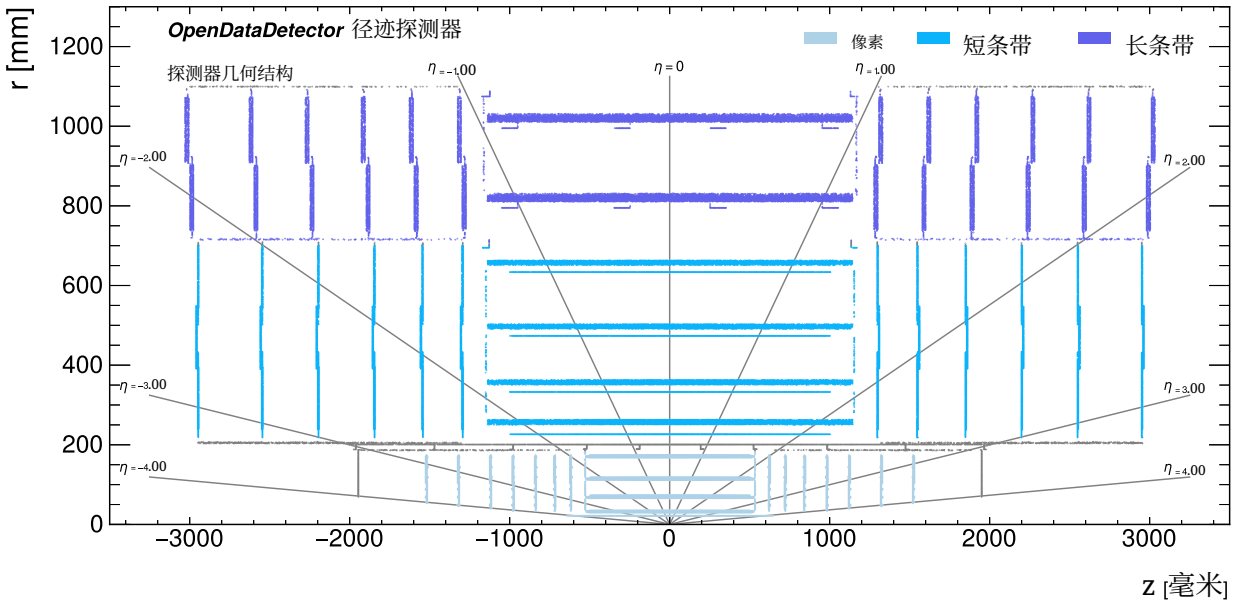
因此，每个堆积子事件都是独立生成的（形成约 2 亿个部分子子事件的池），并在模拟之前 无重复地合并。由于堆积不太可能成为详细分析的重点，我们使用 Pythia8 原生的领先阶矩阵元计算来生成该过程，随后在 Pythia8 中进行簇射。对于每个硬散射过程，从泊松分布中抽取 N 个堆积事件，其中 $\lambda = 200$，并将其合并到一个 HepMC 文件中。图 3 显示了这一过程符合预期。

顶点涂抹 在合并簇射后的稳定粒子子事件（即一个硬散射事件和 N 个堆积事件）时，我们模拟有限束斑的影响。每个子事件都根据以原点为中心的高斯分布随机偏移。空间和时间宽度设为$\sigma_{x,\ y} = 12.5$ μm、 $\sigma_z = 55.5$ 毫米和 $\sigma_t = 185$ ps。

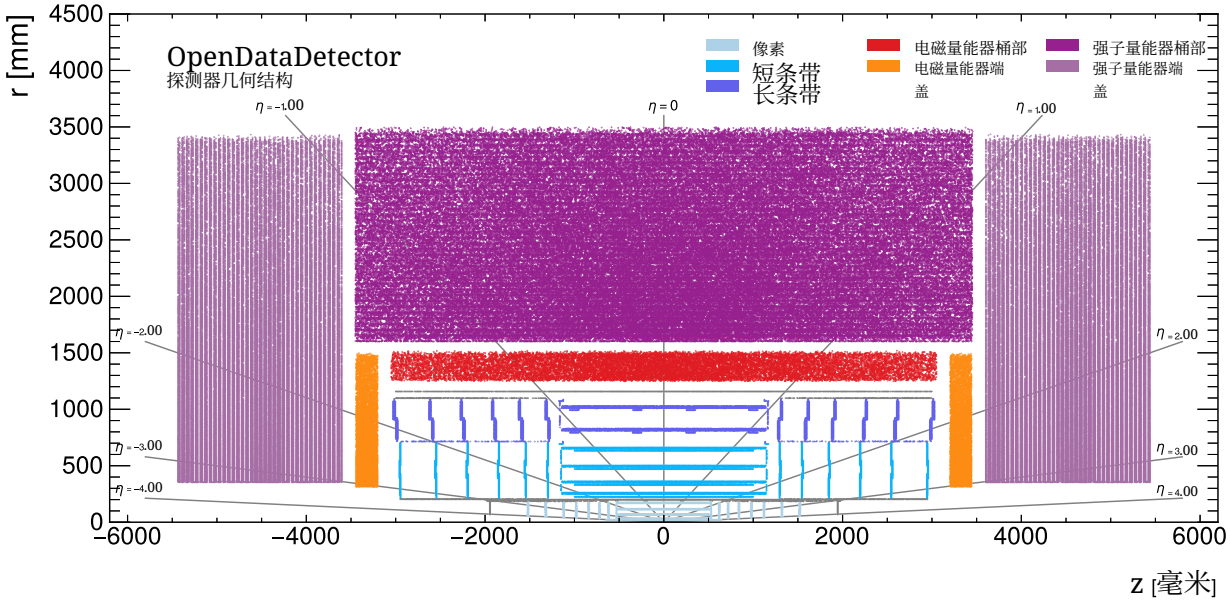
3.2 探测器描述

所采用的探测器几何结构属于 OpenDataDetector (ODD) [55], ，这是一个由 ACTS [56] 合作组织推动的项目，旨在以与实验无关的方式，为高光度和未来对撞机环境提供逼真的物理模拟。ODD 融合了现有及未来探测器的设计选择：一个受 ATLAS 第二阶段 ITk [57, 58], 启发、具有广泛覆盖范围的硅内径迹探测器 (IT)；一个类似于若干 FCCee 探测器方案的电磁量能器 (ECal)；以及一个类似于 CMS HGCal 子探测器和未来探测器 CLD、AHCAL 与 SiD 的强子量能器 (HCal) [59–63]。未来计划加入缪子探测系统。内径迹探测器由两个子系统组成：最内层采用高精度像素硅探测器，最外层采用精度较低的硅条探测器。径迹探测器各层合计在赝快度 $\eta \in [-3, 3]$ 范围内至少提供 12 层灵敏材料，并在 $\eta \in [-3.5, 3.5]$ 范围内提供 8 层。

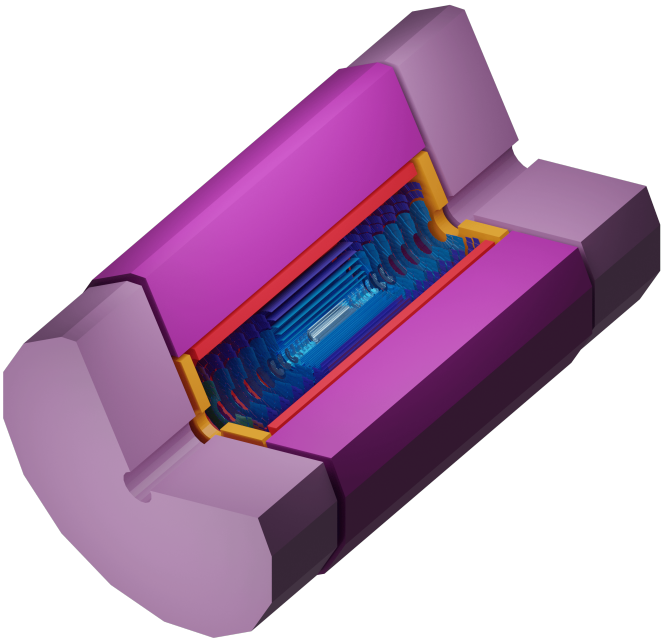
电磁量能器由 48 层厚度为 0.5 毫米的硅传感器组成，每层分为尺寸为 5.1 毫米 $\times$ 5.1 毫米的单元，并与厚度为 1.9 毫米的钨吸收层交错排列。读出，



(a) ODD径迹系统的纵向视图



(b) 包含量能器的 ODD 纵向视图



(a) ODD 径迹重建系统和量能器系统的等轴三维视图

图 5：ODD 内层径迹探测器由最内层的高颗粒度硅像素、中间层的短硅“条像素”（strixel）条带以及最外层的较长二维分辨率条带组成。预计该径迹探测器能够重建出良好的径迹，直至 $\eta \leq 3.5$ [64]。量能器具有高颗粒度，电磁量能器（ECal）的深度至少为 $27\ X_0$，桶部和端盖则为 $10\ \lambda_0$。

间隔件以及额外材料使每层的厚度增加了 2.65 毫米。强子量能器由 36 层厚 3 毫米的聚苯乙烯闪烁体组成，分割成尺寸为 30 毫米 $\times$ 30 毫米的单元。吸收体为 30 毫米厚的钢，剩余的 16 毫米用于读出。

本研究采用 OpenDataDetector 的默认螺线管磁场设置，该设置较探测器的先前版本略有更新（见 [64]）。现在，磁场由螺线管产生，在探测器中心达到 3 T；螺线管外部则配置为 0.5 T 的环境磁场。选择 3 T 是为了在径迹曲率方面设定一个新颖的工作点，介于 ATLAS探测器的 2 T 磁场 [65] 与 CMS 探测器的 4 T [66]之间。灵敏跟踪材料的外边界设为 $r \leq 1080$ 毫米和$z \in [-3030$ 毫米，3030 毫米]。这用于定义入射到量能器的粒子的真值处理，详见第 3.3 节。

3.3 粒子传播与相互作用

为了模拟 Pythia8 产生的每个初级粒子所经过的路径及其衰变产生的次级粒子，我们使用 Geant4 [67] 。具体而言，我们使用作为 Key4hep 框架一部分提供的 DD4hep 库 [68], 其中包含用于读取探测器几何描述的便捷函数，以及用于指定真值处理方式的函数，并最终提供标准的 EDM4hep 数据模型 [69]，以记录有关事例模拟的全部信息。

为更好地满足 COLLIDERML 数据集[3]的需求，我们对标准 DD4hep 配置做了一些调整。多线程以实验模式实现，这

[3]我们预计，这些更改将在不久的将来出现在 DD4hep 的主分支和大型强子对撞机计算网格（LCG）软件发行版中。

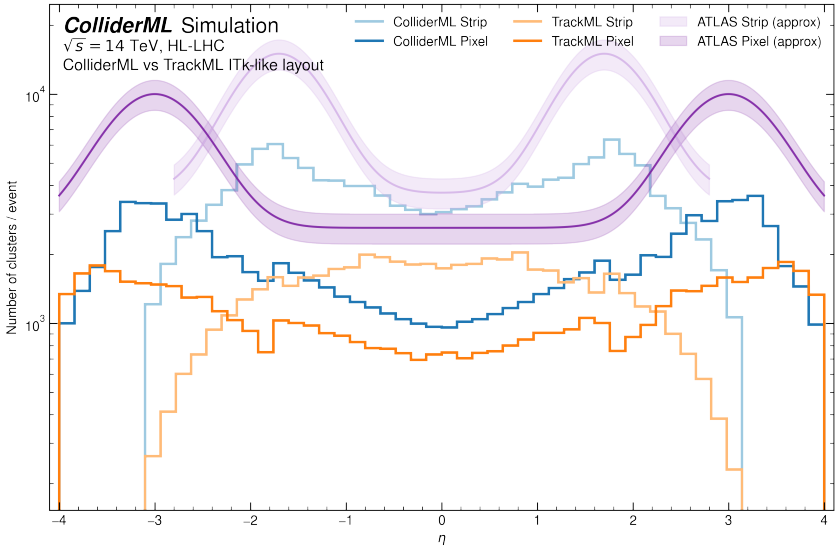


图 6：模拟 TrackML 数据集（橙色）、ATLAS Phase 2 内径迹探测器（紫色）和 COLLIDERML（蓝色）的内径迹探测器击中数随赝快度的分布。ATLAS 数据取自[70]

使每个 CPU节点的模拟吞吐量有望提高约 4 倍。这是因为多进程模式下的模拟受内存限制。模拟是 COLLIDERML流水线中成本最高的阶段，这项改进使可用事件数量在现有计算预算下从约 300,000 个增加到超过 1,000,000 个。附录 A.1 节介绍了对扩展行为的考察。

第二项定制旨在保留更多粒子衰变真值。在默认的 DD4hep 模拟中，绝大多数粒子都会被丢弃，导致之后所有子粒子命中都归因于少数几个母粒子。对于 COLLIDERML，我们引入了两个额外层级的真值处理，采用更为保守的策略：只要存在任何可能的用途，或丢弃信息可能引起疑虑或混淆，就倾向于保留所有粒子细节。默认处理和定制处理的完整说明见 A.2 节。简而言之，模拟层级保留的粒子遵循图 15b 所示的流程图。在这种情况下，只有簇射和模拟过程中产生的“记账”粒子，或在量能器中产生并衰变的极低能粒子会被丢弃。任何在灵敏材料中留下能量沉积的粒子都会被完整保留。

4 数字化

COLLIDERML 数据集提供了高度逼真的数字化，以模拟内径迹探测器和两个量能器中的探测器响应。

4.1 径迹探测器数字化

ODD 径迹探测器中的击中采用基于 ACTS 的几何通道划分方法进行数字化。在此过程中，从模拟期间产生的击中中提取每个粒子的位置和方向。随后将粒子径迹投影到平坦的传感器表面上。

使用分割来模拟实体粒子传感器中的分段读出。对于 COLLIDERML，矩形桶部和端盖中的梯形条带传感器均采用矩形分割。图9展示了所采用的这种矩形分割示意图。所采用分割的数值参数见表2。随后计算粒子穿过的每个分段中的粒子径迹长度，并将传感器厚度纳入考虑。根据粒子的径迹长度以及电荷、动量等粒子性质计算电荷沉积。计算得到的电荷

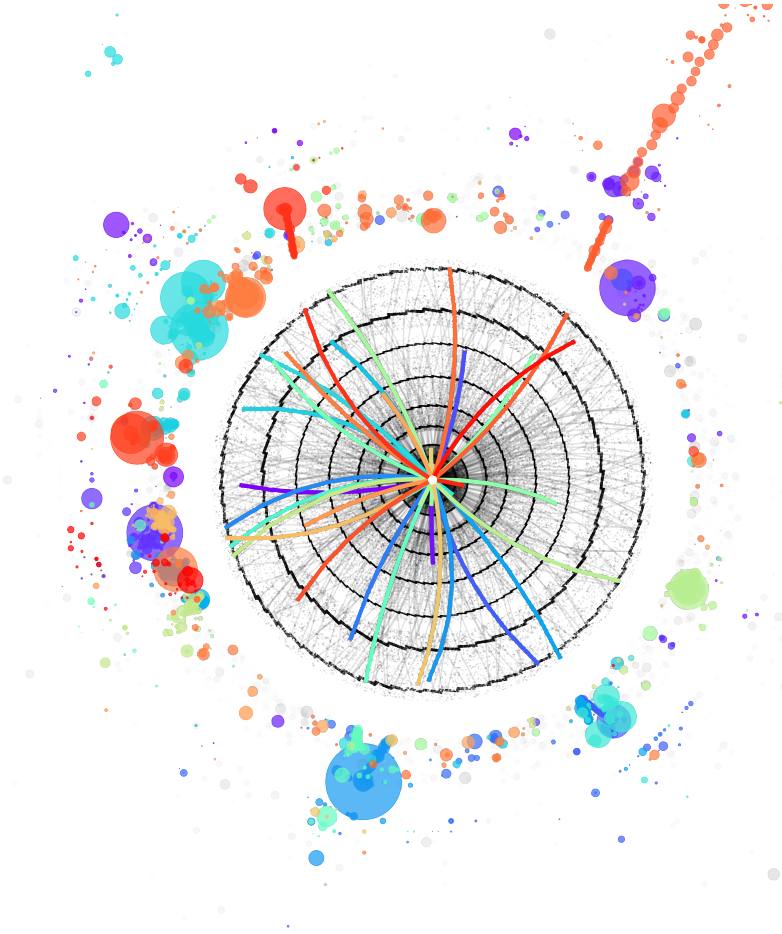


图7：Geant4 模拟中粒子穿过 ODD 的 tt 事件显示。灰色表示 ODD 中所有能量沉积的完整集合。选取部分粒子，并以不同颜色标示它们在 IT 和量能器中的能量沉积；每项量能器能量贡献的大小以圆圈大小表示。

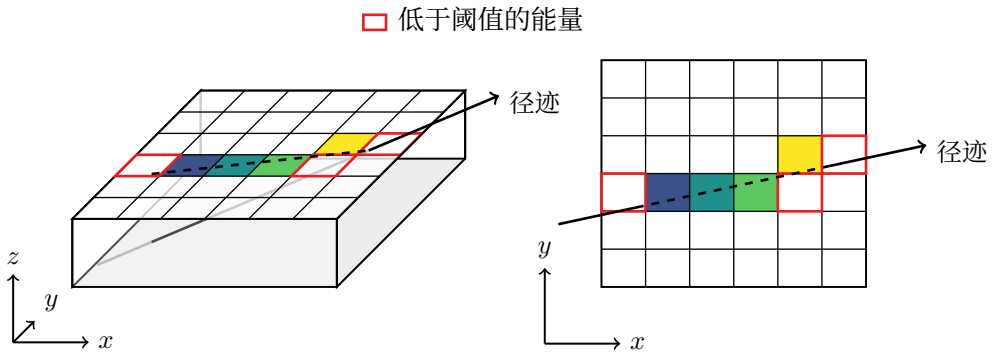
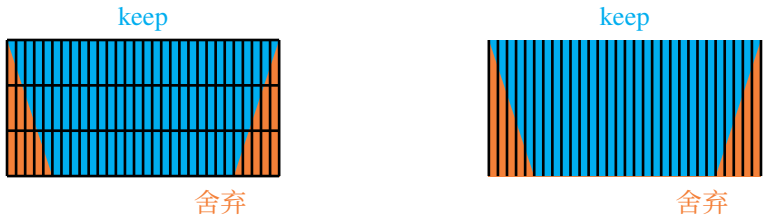


图8：示意图展示了如何通过将沉积能量离散化为传感器上的固定分割，对模拟带电粒子的交点进行数字化。



（a）短条带数字化　　（b）长条带数字化

图 9：使用矩形分割对条带探测器中的端盖传感器进行数字化的方法示意图。

随机进行展宽以模拟测量噪声效应，最后应用逐分段电荷阈值。

簇位置的计算方式有两种：直接取几何平均值（数字、条带），或取电荷加权平均值（模拟、像素）。在数据处理过程中，簇会保留已激活分段的完整列表，但输出数据集仅包含簇位置和方差。这些量为一维或二维，并受限于其所属的传感器。若要转换为三维空间中的一个点，只需取缺失维度上各分段的中心。更精确的三维位置重建（即“空间点”）将在未来版本中加入。

图10展示了探测器中的真值位置和重建位置。真值位置在所有传感器上均匀分布，而重建位置则根据相应子系统的分割呈离散分布。簇会保留与其所在击中的逻辑关联，因此可以从由簇重建的径迹回溯至原始粒子。

簇方差采用统计方法确定：对于每种传感器类别（像素、短/长条带以及桶部、端盖），计算簇位置与相关真值击中位置之间的残差 $\Delta l = l_{\text{rec}} - l_{\text{true}}$。这衡量了簇位置估计的经验不确定度。针对每种簇大小，计算各测量维度上残差分布的均方根（像素和短条带为两个维度，长条带为一个维度）。所得矩阵存储为查找表，并将均方根值作为方差赋给簇，以供下游处理使用。

4.2 量能器数字化

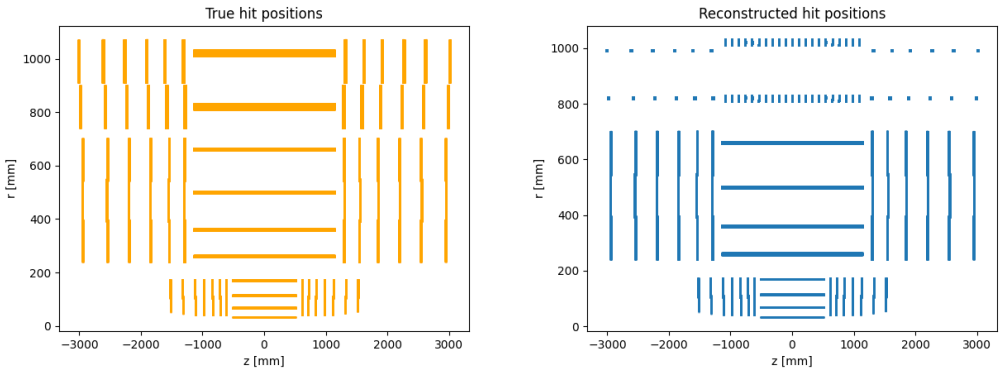
本次发布的量能器单元能量读数包含一小部分阈值效应，用于模拟探测器校准中将采用的初步筛选。首先，对每个贡献粒子的重建时间进行近似计算：从贡献沉积时间中减去飞行时间修正 dt （单位为纳秒）。其定义如下：

$$dt = \sqrt{x^2 + y^2 + z^2}/c - \delta \tag{1}$$

其中， x, y, z 是单元的全局位置，单位为毫米； $c = 300$ 毫米/纳秒约等于光速，而添加 $\delta = 0.1$ 作为小缓冲量，以处理时间窗口噪声。各项贡献的粒子产生时间必须落在事件起始时间（取为 0 纳秒）前后绝对值为 $[-1, 10]$ 纳秒的时间窗口内。

子系统	桶部间距 [μ米]	桶部编号	端盖间距 [μ米]	端盖编号	厚度 [μm]	位置类型
像素	50×50	17	50×50	16/18	125	模拟
短条带	80×500	24	80×500	23/25	200	数字
长条带	100	29	125	28/30	250	数字

表2：ODD 中不同子系统的传感器设计参数



(a) (b)

图10：探测器中的原始击中位置（a）和重建击中位置（b），展示了像素、短条带和长条带数字化的影响

保留下来的贡献按单元求和，并对电磁量能器单元应用极低的能量阈值：50 keV（强子量能器单元为 250 keV）。这两个阈值使粒子贡献数量减少约 90%，有击中的单元数量减少约 60%。

OpenDataDetector 中的量能器是采样量能器——其设计原理决定了它们会根据采样率记录能量沉积。通常，会根据重建对象的性能对各个单元进行扩展性校准，从而校准能量读数。例如，在单粒子模拟中，我们预期探测器内沉积能量的总和应等于粒子的总能量。对于 OpenDataDetector 中的电磁簇射，这种扩展性校准很容易得出；但要将其用于强子簇射，则需要正确实现粒子流重建。数据集的后续版本将为电磁量能器和强子量能器实现这一功能，并加入更逼真的量能器数字化（例如纳入饱和和噪声效应）及重建。对于希望进行校准的用户，我们先提供 A.3 节中的操作方法，以及校准后量能器的分辨率。

5 重建

COLLIDERML 数据集的一个重要组成部分，是能够同时使用探测器的低层直接读数，以及通过标准算法重建的高层对象。这样，研究人员既可以方便地将这些对象用于下游研究，也可以将其作为改进重建的基准。数据集的首个版本包含通过将内径迹探测器击中连接并拟合为粒子轨迹而生成的对象（径迹重建）。未来版本将包含量能器能量沉积的聚类（簇或拓扑聚类）、表征单个粒子的径迹与簇匹配（粒子流），以及将对象聚类为被认为源自单个初级粒子的稠密区域（喷注聚类）。

5.1 径迹重建

重建的径迹对象由 ACTS Examples 框架生成。ACTS（通用径迹软件）是一种与实验无关的径迹重建工具包，多个实验都在生产中使用它（例如 ATLAS、Faser 和 sPHENIX）。

ACTS 为径迹重建的不同阶段提供了一系列成熟的算法。使用连通分量标记（CCL）对硅像素和条形传感器上的激活进行聚类。随后，可通过空间点构建将得到的簇转换到探测器全局坐标系中。这些空间点可以组合成径迹种子，用于获得径迹参数的初步估计。基于组合卡尔曼滤波器（CKF）的通用径迹寻找算法，可以将潜在的径迹种子补全为完整径迹，并通过要求测量数达到最低限值来剔除伪径迹。图 11 展示了一组径迹和粒子示例。

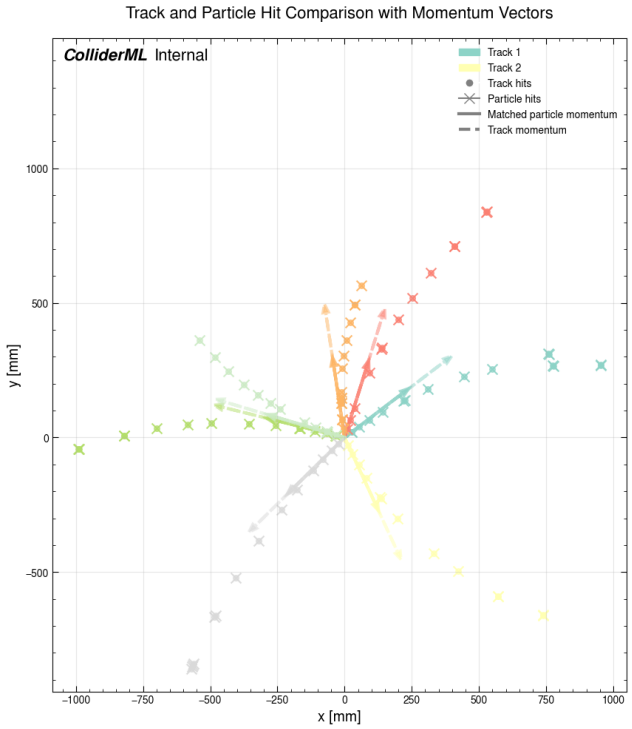


图 11：部分径迹及其匹配粒子和各自击中的可视化结果。可以看出，径迹的完整性较高（包含匹配粒子的大部分击中），纯度也较高（径迹主要由匹配粒子的击中构成）。径迹参数拟合良好，例如此处展示的初始动量角（向量的比例尺不代表实际大小）。

尽管所提供的基于 ACTS 的重建链尚未经过全面调优，但径迹重建的整体表现令人满意。ACTS Examples 框架定义了多种关键性能指标（KPI），包括效率、纯度、分辨率和拉量，定义如下：

效率 重建且唯一匹配到真值的径迹数 / 粒子数。
径迹内容纯度 径迹上正确分配的真值击中数 / 径迹上的击中数。
分辨率 径迹参数残差高斯核心的标准差。
拉偏量 径迹参数残差高斯核心的标准差 / 估计的径迹参数不确定度。

图 12 和图 13 报告了这类指标中具有代表性的一组。

6 数据结构与访问

图1所述流水线的输出由两部分组成：包含模拟的内径迹探测器和量能器能量沉积以及真值粒子的 EDM4hep ROOT 文件，以及包含数字化径迹探测器击中和径迹候选的自定义 ROOT 文件。为了让这套数据最大程度地适用于机器学习实验，我们将其处理为一组包含四个 Parquet 文件的数据集：粒子（表5）、径迹探测器_击中（表6）、量能器_击中（表 7）和径迹（表8）。表3简要列出了所有对象的特征。
数据可通过欧洲核子研究中心组织下的 Hugging Face Datasets 门户获取。该位置与 datasets Python 库兼容，因此可通过该接口使用多种下载和流式传输选项。例如，要获取双希格斯道中仅硬散射事例的粒子列表，可以使用 Python 调用：

```
from datasets import load_dataset ds =load_dataset("CERN/ColliderML-Release-1", "双希格斯_堆积0_粒子")
```

运行以下命令即可使用用于数据集访问和操作的便捷库：

```
pip install colliderml
```

文档见 https://opendatadetector.github.io/ColliderML/。该库包含用于可视化数据、为初态粒子指定子粒子、定义探测器生成区域，以及对事件进行下采样以达到特定堆积工作点的函数。最后一项对于基准测试尤为重要。我们建议，默认情况下，在 COLLIDERML 上获得的结果应针对每次束团交叉发生 200 次相互作用的堆积条件进行测试。不过，在开发研究中，报告仅硬散射、Run 1 条件（堆积 20）、Run 2 条件（堆积 40）或 Run 3 条件（堆积 60）下的性能也可能很有用。这样便可考察随着光度增加而产生的扩展行为。未来的工作将探索在该数据集上进行严格基准测试的指标和技术。

表 3：Release 1 中物理对象及其特征概览

对象	**主要特性**
粒子	**运动学：** 四动量 (p_x， p_y， p_z， E)、质量、电荷。 **顶点：** 产生位置 ($\vec{v}$， t) 和真实近地点 (d_0， z_0)。 **身份：** PDG 编号、唯一标识符、母粒子标识符。 **来源：** 初级粒子布尔值、初级顶点索引。
径迹探测器击中	**位置：** 测量值（x， y， z）与真值（x， y， z），时间。 **层级结构：** 体积、层、表面和探测器 ID。 **真值：** 指向生成该粒子的粒子_标识符。
量能器击中	**能量：** 总能量和单元位置 (x, y, z)。 **贡献：** 对该单元有贡献的真值粒子的嵌套列表（ID、能量分数、时间信息）。
径迹	**拟合螺旋线：** $d_0, z_0, \phi, \theta, q/p$. **关联：** 径迹探测器击中列表 `_ids`. **真值：** 多数_粒子_标识符（真值匹配）。

7 总结与展望

本文介绍了 COLLIDERML，这是一个大规模、开放且与实验无关的数据集，旨在弥合快速模拟基准与高亮度大型强子对撞机的真实高堆积条件之间的差距。COLLIDERML 提供了涵盖十种物理道的 100 万个经过真实数字化的完整模拟事件，为开发和压力测试新一代经典算法与机器学习算法提供了独特的试验平台。

首个版本聚焦于最低层级的对象——能量沉积、径迹和真值粒子——支持立即开展径迹重建和量能器校准工作。未来版本计划大幅扩展这一生态系统。第2版将引入更高层级的重建对象，包括粒子流候选体和喷注，以促进跨阶段研究。未来版本还将引入真实的探测器校准偏差和“挑战”数据集，专门用于检验现代机器学习架构的稳健性和领域适应能力。

我们在 Hugging Face 上公开这些数据，以降低机器学习社区的入门门槛。我们邀请从业者将 COLLIDERML 不仅用作静态基准，还用作动态环境，研究高亮度对撞机环境中重建与生成的基本极限；同时也欢迎大家为指标定义和重建方法贡献力量。

8 致谢

DM 得到了丹麦数据科学学院的资助，该学院由诺和诺德基金会（NNF21SA0069429）资助。PG 通过 CERN 下一代触发器项目获得埃里克与温迪·施密特战略创新基金的资助，资助协议编号为 SIF- 2023-004。本研究使用了国家能源研究科学计算中心的资源。该中心是美国能源部科学办公室用户设施，由美国能源部科学办公室资助，合同编号为 DE-AC02-05CH11231；本研究使用的 NERSC 资助编号为 HEP-ERCAP0034031。

参考文献

[1]Daniel Murnane, Paul Gessinger, Doga Elitez, Andreas Salzburger, Andreas Stefl, Anna Zaborowska, Marcus Selchou Raaholt 和 Stine Kofoed Skov。ColliderML-Release-1（修订版 9d79dd6），2025年。URL https://huggingface.co/datasets/CERN/ ColliderML-Release-1。[2] S. Agostinelli 等。GEANT4——一种模拟工具包。《核仪器与方法 *A*》，506：250–303，2003年。doi：10.1016/S0168-9002(03)01368-8。[3] J Allison。Geant4 的进展与应用。《*IEEE* 核科学汇刊》，53：270，2006年。doi：10.1109/TNS.2006.869826。URL https://cds.cern.ch/record/1035669。[4] Michele Selvaggi。DELPHES 3：用于通用对撞机实验快速模拟的模块化框架。《物理学会议系列杂志》，523：012033，2014年。doi：10.1088/1742-6596/523/1/012033。[5]Michele Selvaggi。Delphes 现状与计划。在 EUCaif 开放数据与模拟器日（LHC）上的报告，2025年11月。URL https://indico.cern.ch/event/1588582/ #11-delphes-status-and-plan。访问日期：2025-12-08。[6] Michela Paganini、Luke de Oliveira 和 Benjamin Nachman。CaloGAN：使用生成对抗网络在多层电磁量能器中模拟三维高能粒子簇射。《物理评论*D*》，97(1)，2018年1月。ISSN 2470-0029。doi：10.1103/physrevd. 97.014021。URL http://dx.doi.org/10.1103/PhysRevD.97.014021。[7] G. et al Aad。Atlfast3：ATLAS 新一代快速模拟。《大科学计算与软件》，6(1)，2022年3月。ISSN 2510-2044。doi：10.1007/s41781-021-00079-7。URL http://dx.doi.org/10.1007/s41781-021-00079-7。[8] Samuel Bein、Patrick Connor、Kevin Pedro、Peter Schleper 和 Moritz Wolf。使用机器学习改进快速模拟。《*EPJ*会议网络》，295：09032，2024年。ISSN 2100- 014X。doi：10.1051/epjconf/202429509032。URL http://dx.doi.org/10.1051/epjconf/ 202429509032。[9] Anders Andreassen、Patrick T. Komiske、Eric M. Metodiev、Benjamin Nachman 和 Jesse Thaler。OmniFold：一种同时展开所有可观测量的方法。《物理评论快报》，124(18)，2020年5月。ISSN 1079-7114。doi：10.1103/physrevlett.124.182001。URL http://dx.doi.org/10.1103/PhysRevLett.124.182001。[10] Kyle Cranmer、Juan Pavez 和 Gilles Louppe。使用经过校准的判别式分类器近似似然比，2016年。URL https://arxiv.org/abs/1506.02169。[11] Johann Brehmer、Kyle Cranmer、Gilles Louppe 和 Juan Pavez。使用机器学习约束有效场论。《物理评论快报》，121(11)，2018年9月。ISSN 1079-7114。doi：10.1103/physrevlett.121.111801。URL http://dx.doi.org/10.1103/PhysRevLett.121.111801。[12] Huilin Qu、Congqiao Li 和 Sitian Qian。JetClass：用于喷注物理深度学习的大规模数据集，2022年6月。URL https://doi.org/10.5281/zenodo.6619768。[13] Sabrina 等，Amrouche。跟踪机器学习挑战赛：精度阶段，第 231–264 页。施普林格国际出版公司，2019年11月。ISBN 9783030291358。doi：10. 1007/978-3-030-29135-8_9。URL http://dx.doi.org/10.1007/978-3-030-29135-8_9。

[14] Sabrina 等，Amrouche。跟踪机器学习挑战：吞吐量阶段。《大科学计算与软件》，7(1)，2023年2月。ISSN 2510-2044。doi：10.1007/ s41781-023-00094-w。URL http://dx.doi.org/10.1007/s41781-023-00094-w。[15] Claudius Krause 等。CaloChallenge 2022：面向快速量能器模拟的社区挑战，2025年。URL http://iopscience.iop.org/article/10.1088/1361-6633/ ae1304。[16] Andris Huang、Yash Melkani、Paolo Calafiura、Alina Lazar、Daniel Thomas Murnane、Minh-Tuan Pham 和 Xiangyang Ju。用于粒子径迹跟踪的语言模型，2024年。URL https: // arxiv.org/abs/2402.10239。[17] Wahid Bhimji、Chris Harris、Vinicius Mikuni 和 Benjamin Nachman。Omnilearned：用于喷注物理所有任务的基础模型框架，2025年。URL https://arxiv. org/abs/2510.24066。[18] Joschka Birk、Anna Hallin 和 Gregor Kasieczka。Omnijet-alpha：首个用于粒子物理的跨任务基础模型。《机器学习：科学与技术》，5(3)：035031，2024年8月。ISSN 2632-2153。doi：10.1088/2632-2153/ad66ad。URL http://dx.doi.org/10. 1088/2632-2153/ad66ad。[19] Tobias Golling、Lukas Heinrich、Michael Kagan、Samuel Klein、Matthew Leigh、Margarita Osadchy 和 John Andrew Raine。集合上的掩码粒子建模：迈向高能物理自监督基础模型，2024年。URL https://arxiv.org/abs/2401.13537。[20] Javier Duarte 和 Jean-Roch Vlimant。用于粒子径迹跟踪与重建的图神经网络，第 387–436 页。WORLD SCIENTIFIC，2022年2月。ISBN 9789811234033。doi：10.1142/9789811234033_0012。URL http://dx.doi.org/10.1142/9789811234033_0012。[21] Gage DeZoort、Savannah Thais、Javier Duarte、Vesal Razavimaleki、Markus Atkinson、Isobel Ojalvo、Mark Neubauer 和 Peter Elmer。通过边分类交互网络进行带电粒子径迹跟踪。《大科学计算与软件》，5(1)，2021年11月。ISSN 2510-2044。doi：10.1007/s41781-021-00073-z。URL http://dx.doi.org/10.1007/s41781-021-00073-z。[22] Caillou、Sylvain、Calafiura、Paolo、Ju、Xiangyang、Murnane、Daniel、Pham、Tuan、Rougier、Charline、Stark、Jan 和 Vallier、Alexis。ATLAS GNN4ITk 径迹重建链的物理性能。*EPJ Web ofConf.*，295：03030，2024年。doi：10.1051/epjconf/202429503030。URL https://doi.org/10.1051/epjconf/202429503030。[23] Murnane，Daniel。影响者损失：用于径迹重建的端到端几何表征学习。*EPJWeb of Conf.*，295：09016，2024年。doi：10.1051/epjconf/202429509016。URL https://doi.org/10.1051/epjconf/202429509016。[24] Ryan Liu、Paolo Calafiura、Steven Farrell、Xiangyang Ju、Daniel Thomas Murnane 和 Tuan Minh Pham。用于粒子径迹重建的分层图神经网络，2023年。URL https://arxiv.org/abs/2303.01640。[25] Xiangyang Ju、Daniel Murnane、Paolo Calafiura、Nicholas Choma、Sean Conlon、Steven Farrell、Yaoyuan Xu、Maria Spiropulu、Jean-Roch Vlimant、Adam Aurisano、Jeremy Hewes、Giuseppe Cerati、Lindsey Gray、Thomas Klijnsma、Jim Kowalkowski、Markus Atkinson、Mark Neubauer、Gage DeZoort、Savannah Thais、Aditi Chauhan、Alex Schuy、Shih-Chieh Hsu、Alex Ballow 和 Alina Lazar。用于高亮度大型强子对撞机粒子径迹跟踪的几何深度学习流水线性能。《欧洲物理学杂志 *C*》，81(10)，2021年10月。ISSN 1434-6052。doi：10.1140/epjc/ s10052-021-09675-8。URL http://dx.doi.org/10.1140/epjc/s10052-021-09675-8。[26] Jared Dynes Burleson、Sylvain Caillou、Paolo Calafiura、Jay Chan、Christophe Collard、Xiangyang Ju、Daniel Thomas Murnane、Mark Neubauer、Minh Tuan Pham、Charline Rougier、Jan Stark、Heberth Torres 和 Alexis Vallier。ATLAS GNN4ITk 径迹重建链的物理性能。技术报告，欧洲核子研究中心，日内瓦，2023年。URL https://cds.cern.ch/record/2882507。

[27] Samuel Van Stroud、Philippa Duckett、Max Hart、Nikita Pond、Sébastien Rettie、Gabriel Facini 和 Tim Scanlon。《用于高能物理中带电粒子径迹重建的变换器》，2024年。URL https://arxiv.org/abs/2411.07149。[28] Jan Kieseler。《目标凝聚：物理探测器、图结构和图像数据中的单阶段无网格多目标重建》。《欧洲物理学杂志 C》，80(9)，2020年9月。ISSN 1434-6052。doi：10.1140/epjc/s10052-020-08461-2。URL http://dx.doi.org/10.1140/epjc/s10052-020-08461-2。[29] Raghav Kansal、Javier Duarte、Hao Su、Breno Orzari、Thiago Tomei、Maurizio Pierini、Mary Touranakou、Jean-Roch Vlimant 和 Dimitrios Gunopulos。《利用消息传递生成对抗网络进行粒子云生成》，2022年。URL https://arxiv.org/abs/ 2106.11535。[30] Claudius Krause 和 David Shih。《Caloflow II：利用归一化流更快且仍准确地生成量能器簇射》，2023年。URL https://arxiv.org/abs/2110.11377。[31] Erik Buhmann、Gregor Kasieczka 和 Jesse Thaler。《EPiC-GAN：用于粒子喷注的等变点云生成》。《SciPost Physics》，15(4)，2023年10月。ISSN 2542-4653。doi：10.21468/ scipostphys.15.4.130。URL http://dx.doi.org/10.21468/SciPostPhys.15.4.130。[32] Nilotpal Kakati、Etienne Dreyer、Anna Ivina、Francesco Armando Di Bello、Lukas Heinrich、Marumi Kado 和 Eilam Gross。《HGPflow：将超图粒子流扩展至对撞机事件重建》，2025年。URL https://arxiv.org/abs/2410.23236。[33] Farouk Mokhtar。《CMS中基于机器学习的粒子流算法》，2025年。URL https: //arxiv.org/abs/2508.20541。[34] Shiqi Gong、Qi Meng、Jue Zhang、Huilin Qu、Congqiao Li、Sitian Qian、Weitao Du、Zhi-Ming Ma 和 Tie-Yan Liu。《用于喷注标记的高效洛伦兹等变图神经网络》。《高能物理学杂志》，2022(7)，2022年7月。ISSN 1029-8479。doi：10.1007/ jhep07(2022)030。URL http://dx.doi.org/10.1007/JHEP07(2022)030。[35] Alexander Bogatskiy、Timothy Hoffman、David W. Miller、Jan T. Offermann 和 Xiaoyang Liu。《粒子物理的可解释等变神经网络：Pelican》。《高能物理学杂志》，2024(3)，2024年3月。ISSN 1029-8479。doi：10.1007/jhep03(2024)113。URL http://dx.doi.org/10.1007/JHEP03(2024)113。[36] Jonas Spinner、Victor Bresó、Pim de Haan、Tilman Plehn、Jesse Thaler 和 Johann Brehmer。《用于高能物理的洛伦兹等变几何代数变换器》，2024年。URL https://arxiv.org/abs/2405.14806。[37] Daniel Murnane、Savannah Thais 和 Jason Wong。《用于喷注标记的半等变 GNN 架构》。《物理学杂志：会议系列》，2438(1):012121，2023年2月。ISSN 1742-6596。doi：10.1088/1742-6596/2438/1/012121。URL http://dx.doi.org/10.1088/ 1742-6596/2438/1/012121。[38] Daniel Murnane、Savannah Thais 和 Ameya Thete。《用于带电粒子径迹跟踪的等变图神经网络》，2023年。URL https://arxiv.org/abs/2304.05293。[39] Savannah Thais 和 Daniel Murnane。《等变性并非你所需的一切：刻画等变图神经网络在粒子物理任务中的效用》，2023年。URL https: //arxiv.org/abs/2311.03094。[40] Savas Dimopoulos、Scott Thomas 和 James D. Wells。《规范媒介超对称性破缺下的超对称粒子能谱学与电弱对称性破缺》。《核物理B》，488 (1–2):39–91，1997年3月。ISSN 0550-3213。doi：10.1016/s0550-3213(97)00030-8。URL http://dx.doi.org/10.1016/S0550-3213(97)00030-8。[41] G. Aad、E. Aakvaag、B. Abbott、S. Abdelhameed 和 et al Abeling, K. 在 $\sqrt{s}$ = 13 太电子伏和 13.6 太电子伏 pp 碰撞中利用ATLAS探测器搜寻位移轻子。《物理评论D》，112(1)，2025年7月。ISSN 2470-0029。doi：10.1103/w8hh-xf24。URL http://dx.doi.org/ 10.1103/w8hh-xf24。

–

[42] Matthew J. Strassler 和 Kathryn M. Zurek。《强子对撞机中隐藏谷的回响》。《物理学快报B》，651(5–6):374–379，2007年8月。ISSN 0370-2693。doi：10.1016/j.physletb. 2007.06.055。URL http://dx.doi.org/10.1016/j.physletb.2007.06.055。[43] A. Tumasyan、W. Adam、J. W. Andrejkovic、T. Bergauer 和 S. Chatterjee。《在 13 太电子伏质子-质子碰撞中搜寻强耦合暗物质的共振产生》。《高能物理学杂志》，2022(6)，2022年6月。ISSN 1029-8479。doi：10.1007/jhep06(2022)156。URL http://dx.doi.org/10.1007/JHEP06(2022)156。[44] Paul Langacker。《重 z' 规范玻色子的物理学》。《现代物理评论》，81 (3):1199–1228，2009年8月。ISSN 1539-0756。doi：10.1103/revmodphys.81.1199。URL http://dx.doi.org/10.1103/RevModPhys.81.1199。[45] R. Abdul Khalek、A. Accardi、J. Adam、D. Adamiak 和 W. Akers。《电子-离子对撞机的科学需求与探测器概念》。《核物理A》，1026:122447，2022年10月。ISSN 0375-9474。doi：10.1016/j.nuclphysa.2022.122447。URL http://dx.doi.org/10.1016/j.nuclphysa.2022.122447。[46] Michelangelo et al Mangano。《未来环形对撞机物理机遇》。技术报告 6，欧洲核子研究中心，日内瓦，2019年。URL https://cds.cern.ch/record/2651294。[47] Jorge et al de Blas。《物理简报：为2026年欧洲粒子物理学战略更新提供的意见》，2025年。URL http://cds.cern.ch/record/2944678。[48] J. Alwall 等。《树图级和次领头阶微分截面的自动化计算及其与部分子簇射模拟的匹配》。《JHEP》，07:079，2014年。doi：10.1007/JHEP07(2014)079。[49] T. Sjöstrand 等。《PYTHIA 8.2简介》。《Comput. Phys. Commun.》，191:159–177，2015年。doi：10.1016/j.cpc.2015.01.024。[50] Barbara Jäger、Alexander Karlberg、Simon Plätzer、Johannes Scheller 和 Marco Zaro。《通过矢量玻色子融合产生希格斯玻色子的部分子簇射效应》。《欧洲物理学杂志C》，80(8)，2020年8月。ISSN 1434-6052。doi：10.1140/epjc/s10052-020-8326-7。URL http://dx.doi.org/10.1140/epjc/s10052-020-8326-7。[51] Stefan Höche、Stephen Mrenna、Shay Payne、Christian T Preuss 和 Peter Skands。《使用 VINCIA 和 PYTHIA 研究矢量玻色子融合希格斯产生中的量子色动力学辐射》，2021年。URL https://arxiv. org/abs/2106.10987。[52] 《ATLAS内层径迹仪像素探测器技术设计报告》。技术报告 CERN-LHCC-2017-021；ATLAS-TDR-030，欧洲核子研究中心，日内瓦，2017年9月。URL http://cds.cern.ch/ record/2285585。[53] 《ATLAS内层径迹仪条纹探测器技术设计报告》。技术报告 CERN-LHCC-2017-005；ATLAS-TDR-025，欧洲核子研究中心，日内瓦，2017年4月。URL https://cds.cern.ch/ record/2257755。[54] 《CMS径迹仪第二阶段升级》。技术报告，欧洲核子研究中心，日内瓦，2017年。URL https://cds.cern.ch/record/2272264。[55] P. Gessinger-Befurt 等。《开放数据探测器径迹系统》。仪器学日（IN2P3）2023，演示文稿，2023年。URL https://indico.in2p3.fr/event/28938/contributions/ 123916/。[56] Xiaocong et al Ai。《通用径迹软件项目》。《大科学计算与软件》，6(1):8，2022年。doi：10.1007/s41781-021-00078-8。URL https://doi.org/10.1007/ s41781-021-00078-8。[57] ATLAS 合作组。《ATLAS内层径迹仪像素探测器技术设计报告》。技术报告 CERN-LHCC-2017-021，ATLAS-TDR-030，欧洲核子研究中心，2017年。URL https://cds.cern.ch/record/2285585。

[58] ATLAS 合作组。《ATLAS 内部径迹仪条带探测器技术设计报告》。技术报告 CERN-LHCC-2017-005，ATLAS-TDR-025，欧洲核子研究中心，2017年。网址：https://inspirehep.net/literature/1614102。[59] CLICdp/FCC-ee 探测器工作组。CLD——未来环形对撞机探测器概念。https://edms.cern.ch/document/2390058/1，2019年。CERN-LCD-2019-001。[60] CMS 合作组。《CMS 端盖量能器第二阶段升级》。技术报告 CERN-LHCC-2017-023，CMS-TDR-019，欧洲核子研究中心，2017年。网址：https://cds.cern.ch/record/2293646。[61] M. Bacchetta 等。CLD——未来环形对撞机探测器概念。欧洲核子研究中心报告，2019年。arXiv:1911.12230。[62] C. Adloff 等。CALICE 模拟强子量能器原型的建造与调试。*JINST*，5:P05004，2010年。doi: 10.1088/1748-0221/5/05/P05004。[63] H. Aihara、P. Burrows、M. Oreglia 等。SiD 意向书。*arXiv*，2009年。[64] Paul Gessinger-Befurt、Andreas Salzburger 和 Joana Niermann。开放数据探测器径迹系统。载于《物理学杂志：会议系列》，第2438卷，页码012110。IOP Publishing，2023年。[65] *ATLAS* 中央螺线管。技术设计报告。ATLAS。欧洲核子研究中心，日内瓦，1997年。doi: 10.17181/CERN.ZZVJ.2JYE。网址：https://cds.cern.ch/record/331067。电子版不可用。[66] *CMS* 磁体项目。技术设计报告。CMS。欧洲核子研究中心，日内瓦，1997年。doi: 10.17181/CERN.6ZU0.V4T9。网址：https://cds.cern.ch/record/331056。[67] S. Agostinelli 等。Geant4——一种模拟工具包。《核仪器与方法物理研究*A*辑：加速器、谱仪、探测器及相关设备》，506(3): 250–303，2003年。ISSN 0168-9002。doi: https://doi.org/10.1016/S0168-9002(03)01368-8。网址：https://www.sciencedirect.com/science/article/pii/S0168900203013688。[68] M. Frank、F. Gaede、M. Petric 和 A. Sailer。DD4hep：探测器描述工具包。载于*J.Phys. Conf. Ser.*，第513卷，页码022010，2014年。doi: 10.1088/1742-6596/513/2/022010。[69] Frank Gaede、Thomas Madlener、Placido Declara Fernandez、Gerardo Ganis、Benedikt Hegner、Clement Helsens、Andre Sailer、Graeme A Stewart 和 Valentin Voelkl。EDM4hep——高能物理实验通用事件数据模型。*PoS*，ICHEP2022:1237，2022年。doi: 10.22323/1.414.1237。网址：https://cds.cern.ch/record/2869508。[70] G. 等 Aad。高亮度大型强子对撞机运行条件下 ATLAS 内部径迹仪的预期径迹重建性能。《仪器学报》，20(02):P02018，2025年2月。ISSN 1748-0221。doi: 10.1088/1748-0221/20/02/p02018。网址：http://dx.doi.org/10.1088/1748-0221/20/02/P02018。[71] G. Bayatian 等。CMS 前向量能器楔形模块的设计、性能与校准。《欧洲物理学杂志*C*辑》，53:139–166，2008年。doi: 10.1140/epjc/s10052-007-0459-4。[72] Salim Cerci。利用 CMS 实验测量 13 太电子伏下的前向能流。技术报告，欧洲核子研究中心，日内瓦，2016年。网址：https://cds.cern.ch/record/2241148。[73] Gregor Kasieczka、Benjamin Nachman 和 David Shih。大型强子对撞机奥林匹克 2020 异常检测挑战赛官方数据集，2019年11月。网址：https://doi.org/10.5281/zenodo.3547721。[74] Thea 等 Aarrestad。暗机器异常评分挑战赛：大型强子对撞机的基准数据与模型无关的事件分类。*SciPost Physics*，12(1)，2022年1月。ISSN 2542-4653。doi: 10.21468/scipostphys.12.1.043。网址：http://dx.doi.org/10.21468/SciPostPhys.12.1.043。

[75] Eric A. Moreno、Philip Ploner、Ranit Das、Jennifer Ngadiuba、Abhijith Gandrakota、Benedikt Maier、Thea Aarrestad、Maciej Glowacki、Shiqi Kuang、Philip Harris、Javier Duarte、David Shih、Alex Tapper 和 Mia Liu。COLLIDE-2V：包含 7.5 亿个双视图大型强子对撞机事件、面向低延迟机器学习的数据集。HuggingFace Datasets，2025年。网址：https://huggingface.co/datasets/fastmachinelearning/collide-1m。报告：https://indico.cern.ch/event/1496673/contributions/6637964/.[76] Huilin Qu、Congqiao Li 和 Sitian Qian。用于喷注标记的粒子变换器，2024年。网址：https://arxiv.org/abs/2202.03772。[77] ATLAS 合作组。ATLAS 顶夸克标记开放数据集，2022年。网址：https://doi.org/10.7483/OPENDATA.ATLAS.FG5F.96GA。[78] Raghav Kansal、Javier Duarte、Hao Su、Breno Orzari、Thiago Tomei、Maurizio Pierini、Mary Touranakou、Jean-Roch Vlimant 和 Dimitrios Gunopulos。JetNet，2022年8月。网址：https://doi.org/10.5281/zenodo.6975118。[79] Raghav Kansal、Javier Duarte、Hao Su、Breno Orzari、Thiago Tomei、Maurizio Pierini、Mary Touranakou、Jean-Roch Vlimant 和 Dimitrios Gunopulos。JetNet150，2022年8月。网址：https://doi.org/10.5281/zenodo.6975117。[80] Michele Faucci Giannelli、Gregor Kasieczka、Claudius Krause、Ben Nachman、Dalila Salamani、David Shih 和 Anna Zaborowska。2022年快速量能器模拟挑战——数据集 2，2022年3月。网址：https://doi.org/10.5281/zenodo.6366271。[81] Michele Faucci Giannelli、Gregor Kasieczka、Claudius Krause、Ben Nachman、Dalila Salamani、David Shih 和 Anna Zaborowska。2022年快速量能器模拟挑战——数据集 3，2022年3月。网址：https://doi.org/10.5281/zenodo.6366324。[82] Shuhang Li、Yi Huang、David Park、Xihaier Luo、Haiwang Yu、Yeonju Go、Christopher Pinkenburg、Yuewei Lin、Shinjae Yoo、Joseph Osborn、Christof Roland、Jin Huang 和 Yihui Ren。TPCpp-10M：时间投影室中的模拟质子-质子碰撞数据，供人工智能基础模型使用，2025年8月。网址：https://doi.org/10.5281/zenodo.16970029。[83] Javier Duarte。含有 Hbb 标记机器学习研究所需喷注、径迹和次级顶点属性的样本，2019年。网址：https://doi.org/10.7483/OPENDATA.CMS.JGJX.MS7Q。[84] CMS 合作组。CMS 2016 年开放数据，2016年。网址：https://opendata.cern.ch/search?q=&f=type%3ADataset%2Bsubtype%3ASimulated&f=experiment%3ACMS&l=list&order=desc&p=1&s=10&sort=mostrecent。[85] ATLAS 合作组。ATLAS 实验 2015–2016 年开放数据研究用 DAOD_PHYSLITE 格式，2024年。网址：http://doi.org/10.7483/OPENDATA.ATLAS.9HK7.P5SI。[86]Daniel Whiteson。超对称。UCI 机器学习库，2014年。DOI：https://doi.org/10.24432/C54606。

区域	缩放因子
电磁量能器桶部	37.5
电磁量能器端盖	38.7
强子量能器桶部	45.0
强子量能器端盖	46.9

表4：开放数据探测器当前估算的缩放因子。完整的重建流程实现且簇射完成校准后，数值可能会发生变化（尤其是强子量能器的数值）。

A 附录／补充材料

A.1 DD4hep 多线程扩展行为

该数据集一项具有重要技术意义的进展，是首次在 DD4hep 模拟框架中采用大规模多线程。这对于提供目前可用的事件数量至关重要。要理解其中的原因，请注意，以往的并行化方案要求在每个核心上调用一个独立的 ddsim 实例，这会迅速耗尽计算节点的可用内存。因此，内存限制将模拟的多进程处理限制在每个计算节点 32 个实例。采用多线程后，我们发现扩展至 8 个并行线程时表现良好；如图14所示，每个节点最多可使用 128 个线程。

A.2 DD4hep 模拟真值处理

为了支持对粒子衰变重建或模拟开展深入研究，我们采用非常保守的真值处理方式。DD4hep 算法的流程如下：在粒子衰变／相互作用中，我们是否保留子粒子信息？如果保留，则将子粒子登记为一个粒子，其所有详细信息都写入粒子记录，所有能量沉积也都归属于该粒子。如果不保留，则将该子粒子标记为待丢弃，其能量沉积归属于母粒子。这种归属关系在模拟结束后确定，因此可能出现这样的情况：一个待丢弃的子粒子自身产生了一个应予保留的孙粒子。于是，子粒子的击中归属于其母粒子，而孙粒子的击中归属于其自身，依此类推。在这一框架下，DD4hep 真值处理的选择归结为：在什么条件下丢弃或保留一个粒子？

图15a展示了默认的真值处理。虽然这种处理方式能够生成紧凑的粒子真值记录，且可能适用于大多数标准分析用例，但也会导致多种不理想的行为。例如，量能器向径迹探测器中的反散射未得到正确处理。此外，在这种环境中，径迹重建可能会变得混乱，因为所有低能粒子都被归入其母粒子，因此它们的击中会在径迹探测器中呈现为树状径迹。最后，所有量能器衰变信息都会丢失。

A.3 量能器校准

这里介绍一种简单的OpenDataDetector量能器校准方法。COLLIDERML版本中每个单元的沉积能量已经过阈值处理，因此需要根据单元的位置应用表4中所述的修正因子。

通过检查单粒子样本的总能量，可以评估量能器校准的性能。我们使用单电子、光子和π介子来研究数字化的分辨率。粒子的生成能量在 300 兆电子伏至 1 太电子伏之间呈对数均匀分布。这些样本集已作为 COLLIDERML 的一部分提供。将量能器中沉积的数字化总能量与生成器粒子的初始能量进行比较，从而得到残差 $|E_{gen} - E_{calo}|$、相对残差（$|E_{gen} - E_{calo}|$）$/E_{gen}$ 或拉偏量（$E_{gen} - E_{calo}$）$/\sqrt{E_{calo}}$ 分布。图 16、图 17 和图 18 展示了这些分布。数字化对电子和光子的能量分辨效果非常好，

在高 p_{T} 且赝快度介于 -2 和 2 之间时，残差小于 10%。这与电磁量能器的分辨率上限一致，该上限由其 3% 的采样分数决定。

正如在真实的探测器描述中常见的那样，π介子的能量更难以分辨。即使横动量超过 10 吉电子伏特，分辨率也会停滞在 30%。在该数据集的后续迭代版本中，可通过专门调优或基于机器学习的校准算法来改善这一点。

需要注意，并非所有初态粒子（即该数据集版本中没有母粒子的粒子）都会进入探测器的覆盖范围。其中大多数类似于赝快度非常高的束流残余粒子。在每次 pp 相互作用中涉及的 14 太电子伏动能中，应用表4所述的校准后，约有 O(300) 吉电子伏特会沉积在量能器中。这与大型强子对撞机探测器中的观测结果相当 [71, 72]。

A.4 ObjectData 数据类型

下表详细列出了数据集中使用的确切数据类型和字段名称。

字段	Type	描述
事件_标识符	int32	唯一事件标识符
粒子_标识符	列表<uint64>	事件内唯一粒子标识符
pdg_标识符	列表<整数64>	PDG 粒子编码
mass	列表<浮点数32>	粒子静止质量（吉电子伏特/c2）
能量	列表<浮点数32>	粒子总能量（吉电子伏特）
电荷	list<float32>	电荷（以 e 为单位）
px、py、pz	list<float32>	动量分量（吉电子伏特/c）
vx、vy、vz	列表<浮点数32>	顶点位置（毫米）
time	列表<浮点数32>	产生时间（纳秒）
顶点_初级	列表<无符号整数16>	主顶点 ID（1=硬散射， [2,.....， N]=堆积—up)
近地点_横向冲击参数	列表<浮点数32>	粒子拟合后的径向冲击参数近地点
近地点_纵向冲击参数	列表<浮点数32>	粒子的纵向冲击参数拟合近地点
母粒子_标识符	列表<整数64>	母粒子的标识符
初级	列表<布尔值>	该粒子是否为初级粒子

表5：粒子数据结构。请注意，如果粒子没有母粒子，则 parent_id 的值为 -1 。

字段	Type	描述
事件_标识符	32位无符号整数	唯一事件标识符
x、y、z	列表<浮点数32>	测量的击中位置（毫米）
真实_x、真实_y、真实_z	列表<浮点数32>	数字化前的真实击中位置（毫米）
time	列表<浮点数32>	击中时间（纳秒）
粒子_标识符	列表<无符号整数64>	创建此击中的真值粒子
体积 _id	列表<无符号整数8>	探测器体积标识符
层_标识符	列表<无符号整数16>	探测器层编号
表面 _id	列表<无符号整数32>	传感器表面标识符
探测器	list<uint8>	探测器子系统代码

表 6：径迹探测器命中数据结构

字段	Type	描述
事件_标识符	32位无符号整数	唯一事件标识符
探测器	列表<无符号整数8>	量能器子系统名称

字段	Type	描述
总_能量	列表<浮点数32>	单元中沉积的总能量（吉电子伏特）
x, y, z	列表<浮点数32>	单元中心位置（毫米）
贡献_粒子_ID	列表<列表<无符号64位整数»	对该单元有贡献的粒子 ID
贡献_能量	列表<列表<32位浮点数»	每个粒子的能量贡献（吉电子伏特）
贡献_时间	列表<列表<32位浮点数»	每项贡献的时间（纳秒）

表 7：量能器击中数据。注意贡献使用了嵌套列表。

字段	Type	描述
事件_标识符	32位无符号整数	唯一事件标识符
径迹_标识符	列表<无符号整数16>	事件内的唯一轨迹标识符
多数_粒子_标识符	列表<无符号整数64>	在此径迹上击中数最多的真值粒子
d0	列表<浮点数32>	横向冲击参数（毫米）
z0	列表<浮点数32>	纵向冲击参数（毫米）
phi	列表<浮点数32>	方位角（弧度）
θ	列表<浮点数32>	极角（弧度）
qop	列表<浮点数32>	电荷与动量之比（e/吉电子伏特）
击中_ID	列表<列表<32位无符号整数»	分配给此径迹的径迹探测器命中编号列表

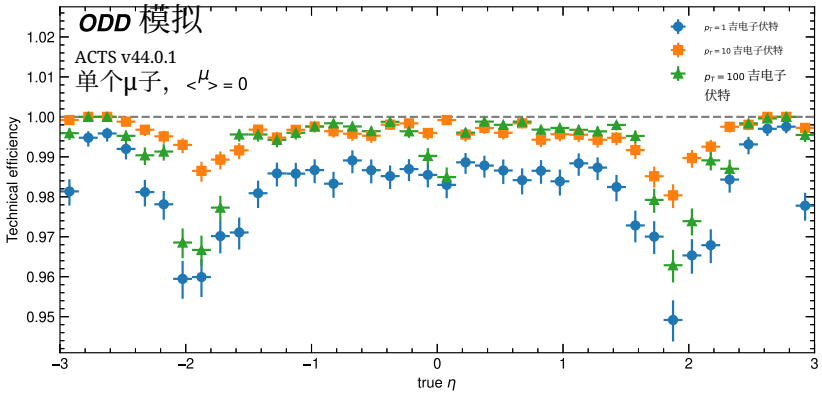
表8：径迹数据结构。径迹参数是由 ACTS 拟合螺旋线所得

B 公开对撞机数据集

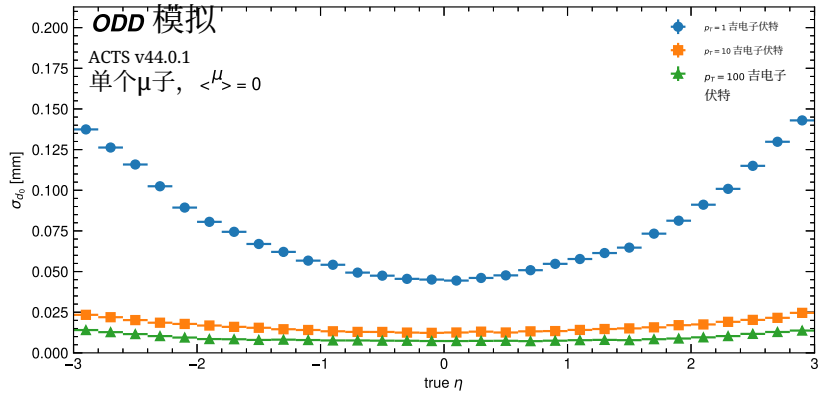
图 2 中包含以下数据集：

表9：涵盖不同对象粒度的公开对撞机物理数据集：低（探测器击中）、中（径迹/量能器簇）、高（粒子流组分）、分析（完全重建的粒子/表格数据）。大小为数据源报告的近似值。

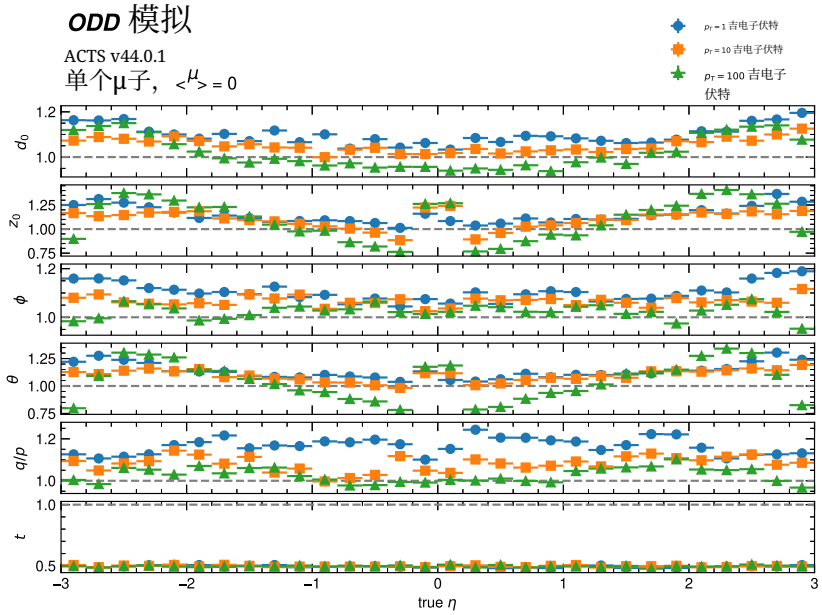
数据集	参考样本（约）级别		大小（约）
LHC Olympics 2020（LHCO2020）	[73]	4,000,000 高	11 吉字节
Dark Machines	[74]	约 10 亿次分析	34 吉字节
COLLIDE-2V	[75]	750,000,000 高	50 太字节
JetClass	[12, 76]	125,000,000 喷注 High	190 吉字节
ATLAS 顶夸克标记开放数据	[77]	44,500,000 个高能喷注	129 吉字节
JetNet30	[78]	880,000 个高能喷注	440 兆字节
JetNet150	[79]		2.1 吉字节
TrackML	[13, 14]	10,000 低	200 吉字节
CaloChallenge 2022	[80] (D2) [81] (D3)	200,000 个簇射 低	2.7 吉字节 7.6 吉字节
TPCpp–10M	[82]	10,000,000 低	120 吉字节
FAIR 希格斯衰变数据集	[83]	6,000,000 喷注 中	228 吉字节
CMS 开放数据（2016年）	[84]	约 100 亿 中	约 100 太字节
ATLAS 开放数据（2015–2016年）	[85]	约 100 亿 中	约 100 太字节
SUSY（UCI）	[86]	5,000,000 分析	880 兆字节
ColliderML	本研究	1,000,000 低	400 太字节（EDM4hep） 40 太字节（Parquet）



(a) 单个粒子的技术径迹寻找效率随真实 η μ p_T = 1 10 100 的变化，粒子能量为,,吉电子伏特。

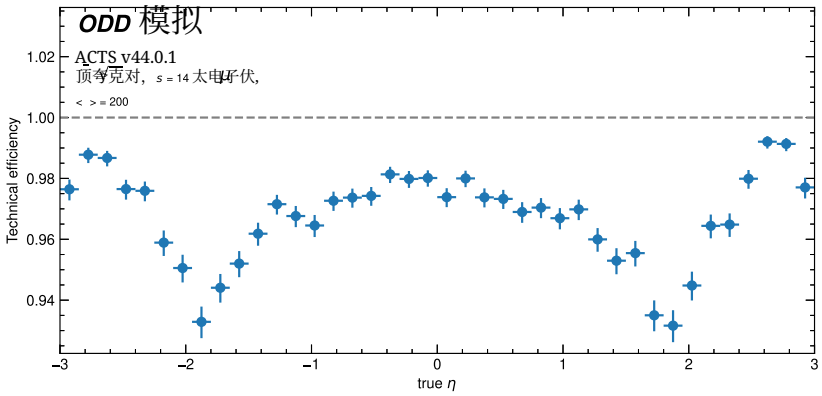


(b) 单个 μ 的径迹分辨率随真实 η 的变化，其中 $p_T = 1$ 为, 10,100 吉电子伏特。

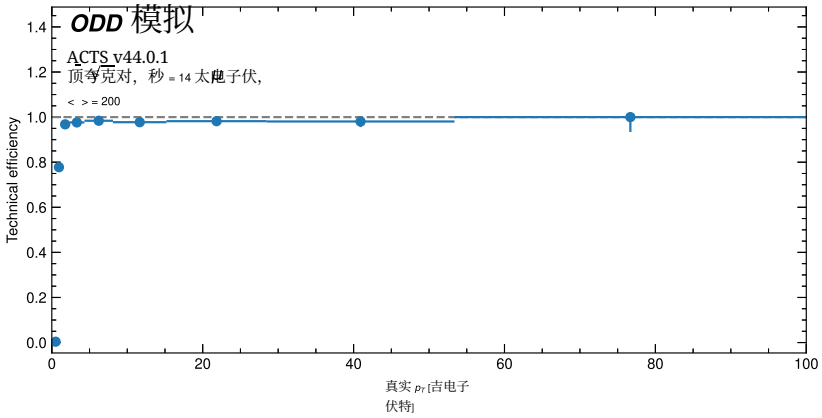


(c) 相互作用点处所有拟合径迹参数的拉偏宽度随真实 η 的变化，其中单个 μ 的能量为 $p_T = 1, 10, 100$ 吉电子伏特。

图 12



(a) 对于堆积为 $\langle\mu\rangle = 200$ 的 $t\bar{t}$ 个事件，技术径迹寻找效率随真实 η 的变化。



(b) 对于堆积为 的事件，技术径迹寻找效率随真实 p_T $t\bar{t}$ $\langle\mu\rangle =$ 的变化。

200

图 13

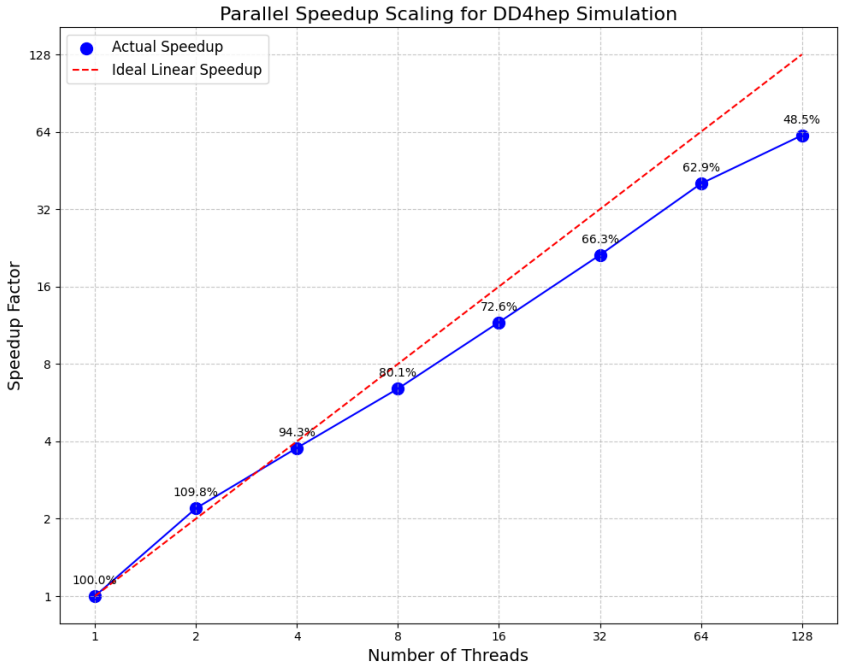
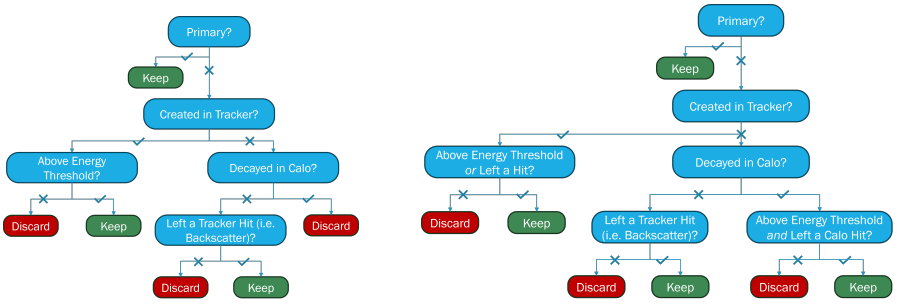
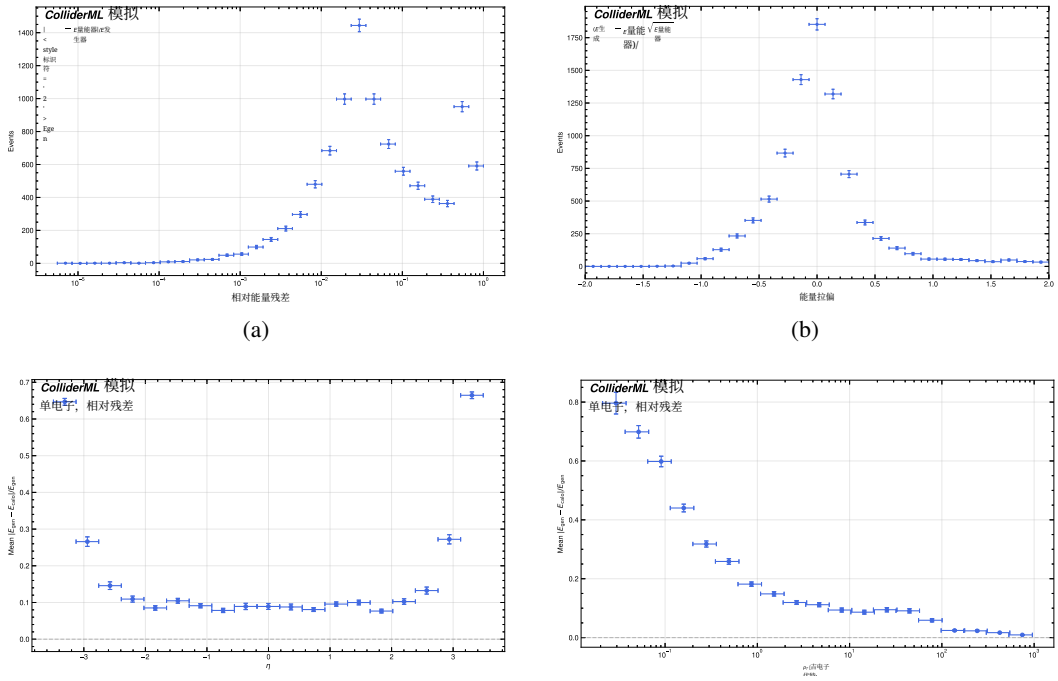


图14：多线程 DD4hep 的扩展性能。每个线程在 $\langle\mu\rangle = 10$ 下模拟 tt 个事件。加速因子定义为线程数乘以单线程运行时间与该线程数下多线程运行时间的比值。图中还给出了实测加速比相对于理想强扩展加速比的百分比。



(a) (b)

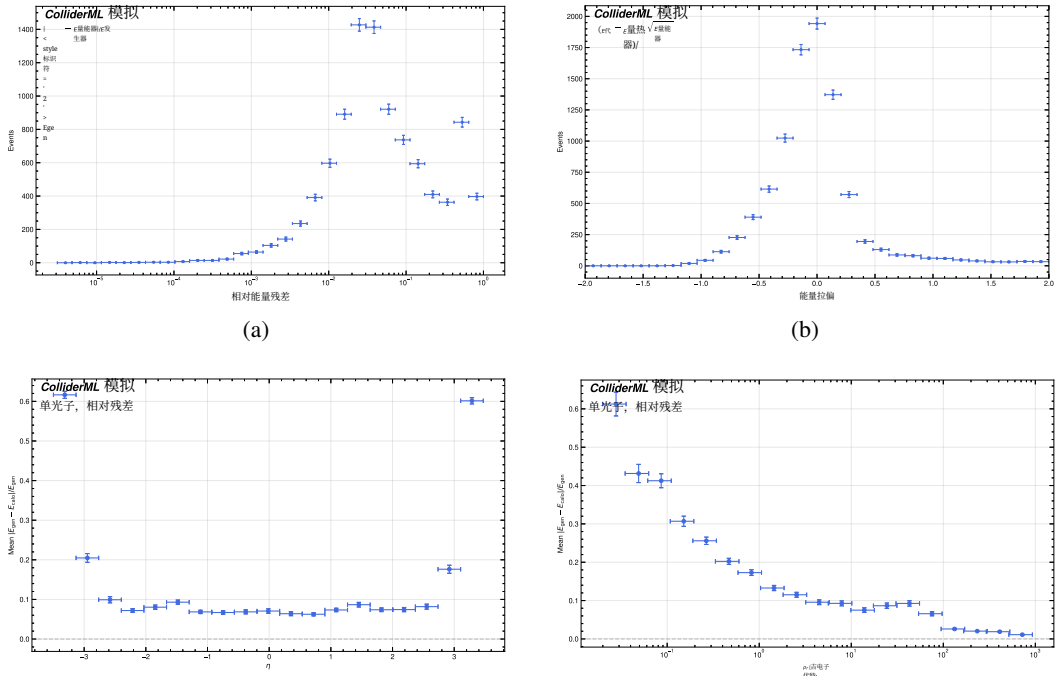
图15：真值处理逻辑流程。a）DD4hep 默认的真值处理。b）自定义的 COLLIDERML 处理，可更详细地跟踪探测器中的粒子信息



(a) (b)

(c) (d)

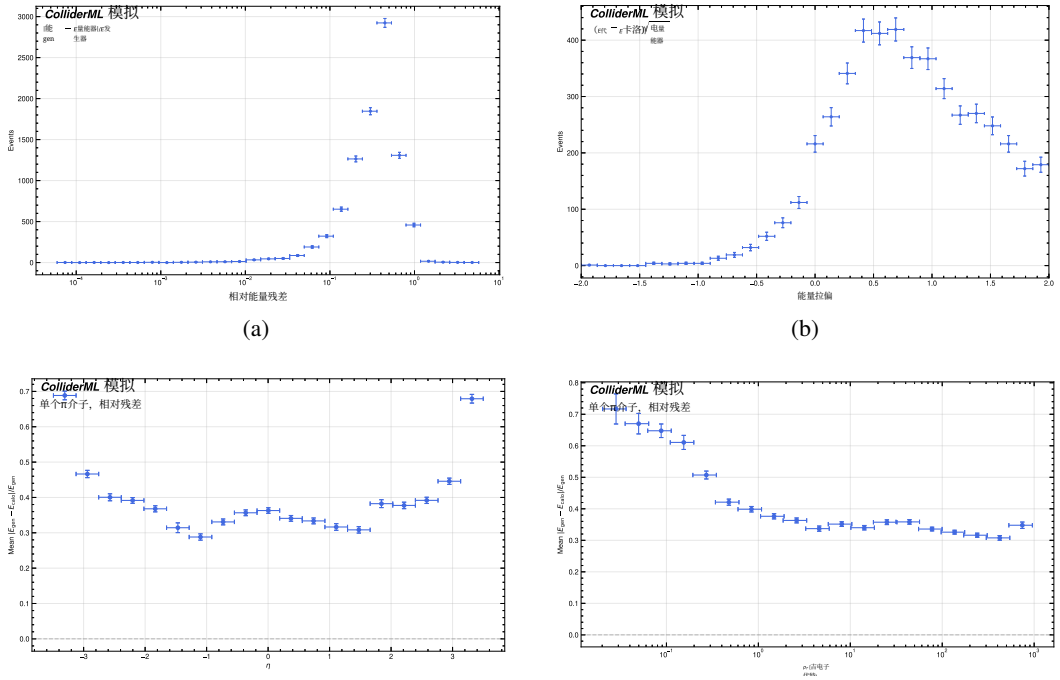
图16：单电子样本的量能器数字化验证结果：a）相对残差分布，b）拉偏分布，c）不同赝快度下的相对残差，以及 d）生成器粒子在不同横动量下的相对残差



(a) (b)

(c) (d)

图17：单光子样本的量能器数字化验证结果：a）相对残差分布，b）拉偏分布，c）不同赝快度下的相对残差，以及 d）生成器粒子在不同横动量下的相对残差



(a) (b)

(c) (d)

图18：单π介子样本的量能器数字化验证结果：a) 相对残差分布，b) 拉偏分布，c) 相对残差随赝快度的变化，以及 d) 相对残差随生成器粒子横动量的变化