中国科学院大学
University of Chinese Academy of Sciences

中国科学院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences

BESIII

Spring 2026 IHEP ML Workshop

面向机器学习径迹重建的漂移室数据集

报告人：钱立宴

指导老师：张瑶、袁野

2026年4月15日

目录

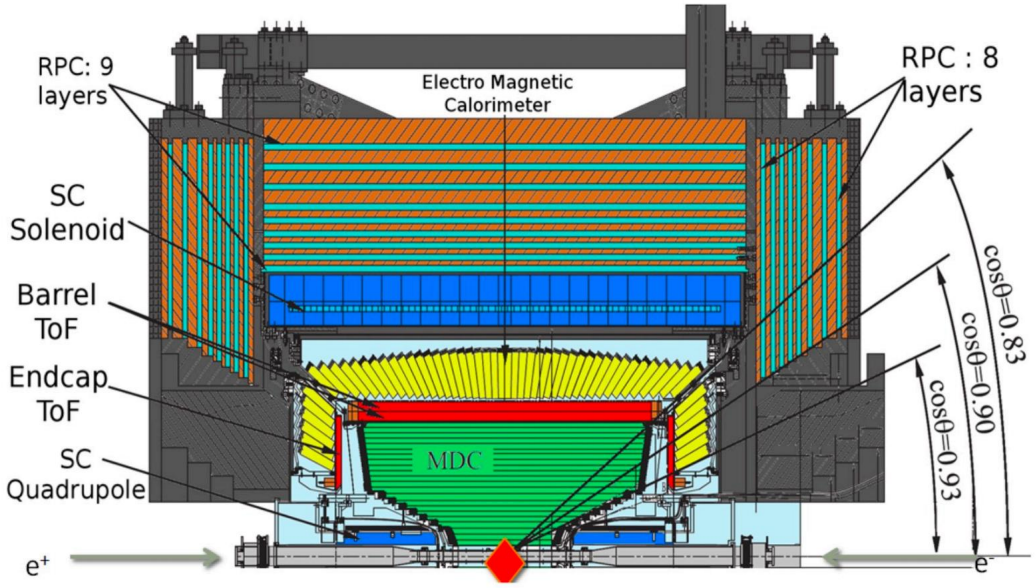
第一部分

工作背景及BESIII漂移室简介

Part1.工作背景及BESIII漂移室简介

宇宙万物由什么构成?

- 粒子物理学：标准模型
- 粒子物理实验：
  - 高能量前沿：CEPC、LHC…
  - **高亮度前沿：BEPCII**、STCF、Belle II…
  - 非加速器物理：LHAASO、JUNO…



BEPCII上的BESIII谱仪
（漂移室MDC为最内层径迹探测器）

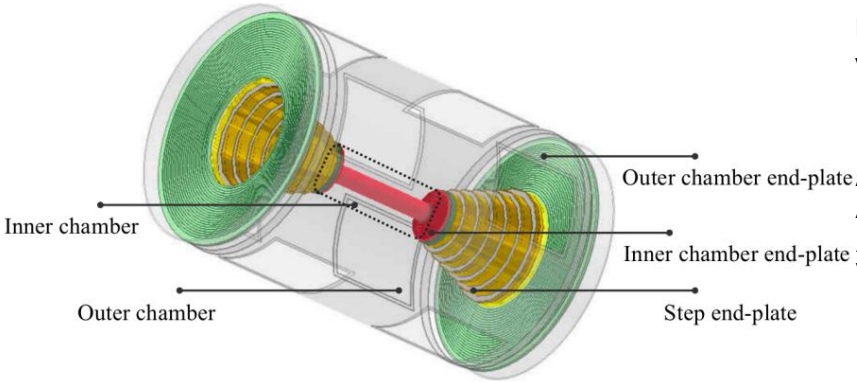
径迹重建任务

- 径迹探测器捕捉带电粒子对撞后留下的径迹，对径迹进行精准识别、关联与拟合，可推出粒子动量、顶点位置、粒子类型等关键物理量，重建精度直接影响粒子物理属性测量及新物理发现。
- 传统重建面临的挑战：性能高度依赖人工调参→难以适应复杂环境与探测器非理想效应；模块化的串行设计→难以实现端到端优化。
- 现有径迹重建机器学习数据集，多为针对硅探测器pileup事例，缺失针对漂移室的数据集：
  - TrackML 粒子跟踪挑战赛数据集→LHC
  - OpenDataDetector开放数据探测器生成模拟事例→（HL-）LHC

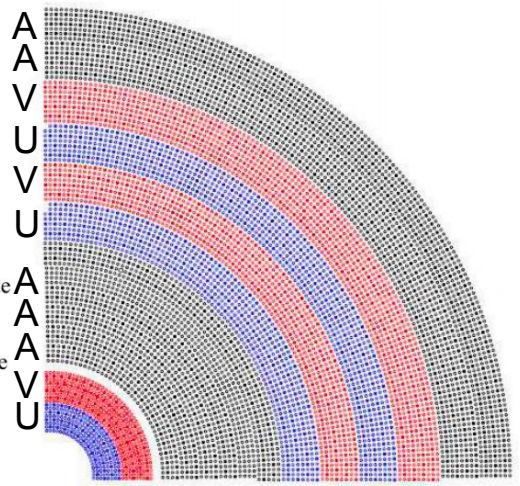
Part1.工作背景及BESIII漂移室简介

BESIII漂移室（MDC）

- BESIII上径迹探测器——漂移室（Multilayer Drift Chamber, MDC），任务是对带电径迹的精确测量，其性能直接影响物理分析的精度和能力
- 几何结构：
  - 圆柱体结构（分内室和外室两部分）——长度为 2582 mm，内半径 59 mm，外半径 810 mm，立体覆盖角为 $-0.93 < \cos\theta < 0.93$
  - 6796个类正方形漂移单元，中间一根信号
  - 丝，外围有 8-9 根场丝
  - 43 层信号丝层、11 个超层，超层结构与几何参数（见右表）
- 内室的升级→CGEM–IT（本工作基于MDC）



MDC整体结构示意图



MDC 1/4 丝层分布

超层结构与几何参数表

Superlayer	Type	N_{layer}	N_{wires}/layer	Radius (mm)	Length (mm)
1	U	4	40, 44, 48, 56	~79 – 115	780 – 816
2	V	4	64, 72, 80, 80	~127 – 162	828 – 864
3	A	4	76, 76, 88, 88	~197 – 246	1092 – 1272
4	A	4	100, 100, 112, 112	~262 – 311	1442 – 1612
5	A	4	128, 128, 140, 140	~327 – 375	1782 – 1952
6	U	4	160×4	~400 – 448	2174 – 2192
7	V	4	176×4	~464 – 514	2198 – 2216
8	U	4	208×4	~530 – 579	2222 – 2240
9	V	4	240×4	~595 – 642	2246 – 2264
10	A	4	256×4	~667 – 716	2276 – 2294
11	A	3	288×3	~732 – 763	2300 – 2306

• Notation: a×n denotes n layers each with number of wire a. A: axial superlayers, U: stereo superlayers with negative tilt angle, V: Stereo superlayers with positive tilt angle.

第二部分

BESIII漂移室径迹重建原理

Part2.BESIII漂移室径迹重建原理

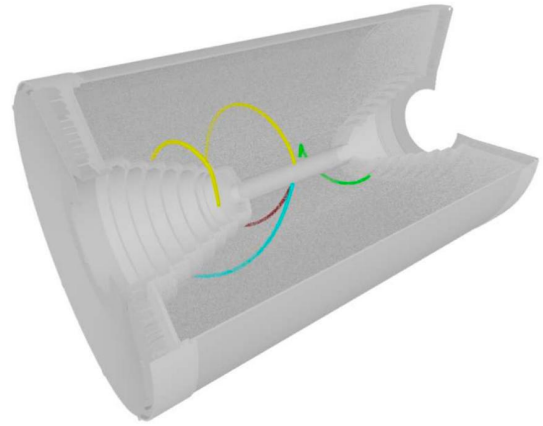
径迹重建原理

- 电离与漂移
  带电粒子使气体电离 → 电子在电场下向阳极丝漂
- 雪崩放大
  电子接近阳极丝时电场增强 → 引发气体二次电离 → 信号放大几个数量级
- 径迹寻迹：
  利用击中信息重建径迹 → 初步确定粒子空间位置与动量

算法：模板匹配（PAT）、径迹段寻找（TSF）、霍夫变换（HOUGH）

- 径迹拟合：Runge-Kutta、Kalman滤波、Genfit
  处理电离能损、多次库仑散射等复杂效应 → 由专门拟合算法完成

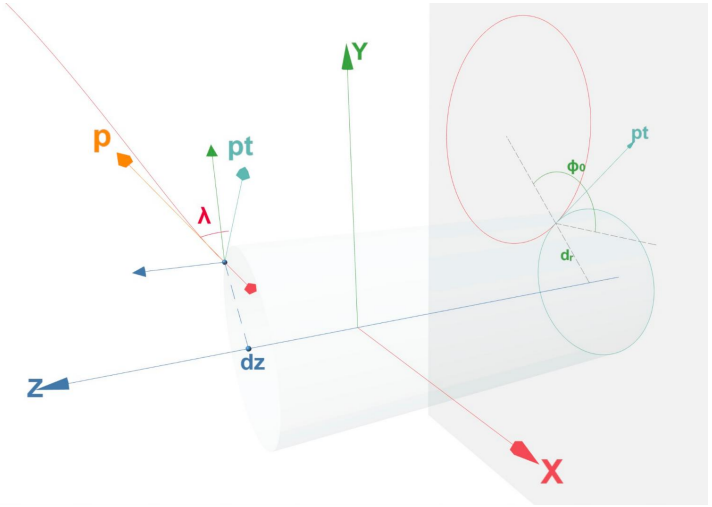
算法：Runge-Kutta、Kalman滤波、Genfit



带电粒子运动径迹 3D 示意图

带电粒子径迹模型

- 带电径迹空间螺旋线的参数
  - 径迹参数默认以原点 $O(0, 0, 0)$ 作为参考点
  - 使用螺旋线在 x-y 平面内距离参考点最近点处的 5 个参数 $(dr, \phi^0, \kappa, dz, \tan\lambda)^T$ 来定义一条螺旋线，在x-y平面内为一个圆



粒子径迹的螺旋线参数化示意图

第三部分

漂移室数据集介绍

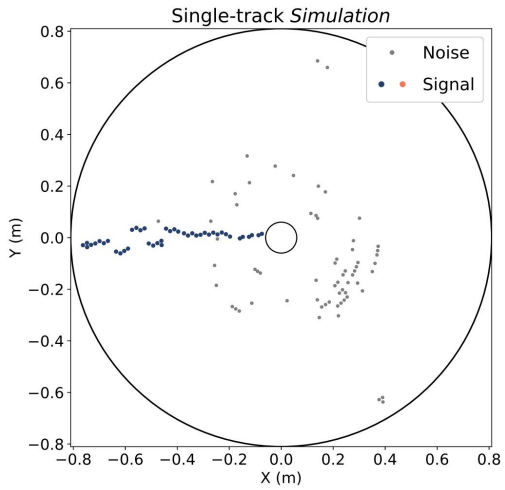
Part3.数据集介绍——数据集构成

样本种类

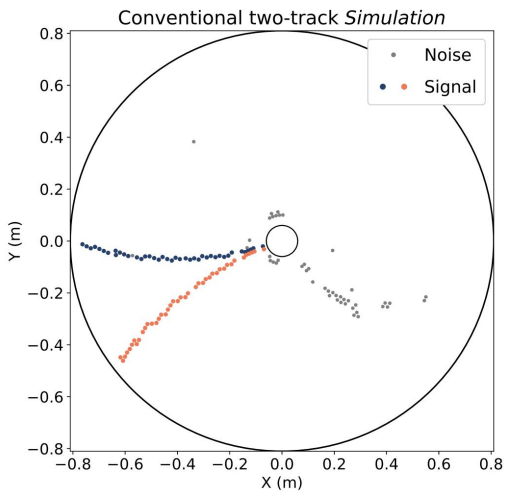
■ 为支撑径迹重建算法的基础研究并降低算法复杂度，样本种类如下：

事例类型	p_T [GeV/c]	$\cos\theta$	ϕ [rad]	粒子种类
单径迹 (*single-track*)	$0.15 \sim 1.5$	$-0.93 \sim 0.93$	$0 \sim 2\pi$	$e^{\pm}, \mu^{\pm}, \pi^{\pm}, K^{\pm}, p, \bar{p}$
常规双径迹 (*conventional two-track*)	$0.15 \sim 1.5$	$-0.93 \sim 0.93$	$0 \sim 2\pi$	$\pi^{+}\pi^{-}$
近邻双径迹 (*close-by two-track*)	$0.15 \sim 1.5$	$-0.93 \sim 0.93$	$\Delta\phi = 0.1 \sim 0.3$	$\pi^{+}\pi^{-}$

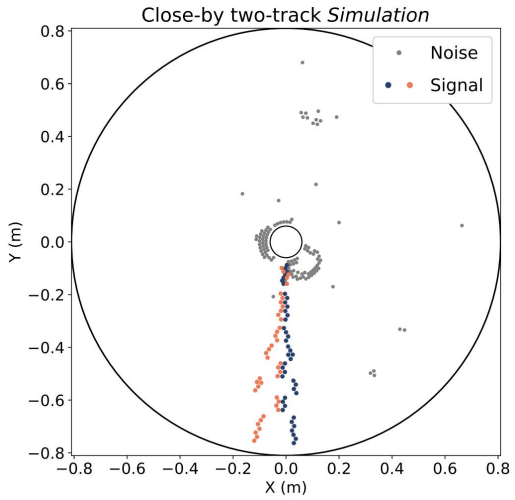
- 要求径迹的横动量 $p_T > 0.15$ GeV；
- 所有事例中的径迹运动学参数均在探测器可访问的相空间内进行均匀采样；
- 数据挑选：
  - ✓ 径迹击中必须经过6层
  - ✓ 径迹必须从目标粒子产生



单径迹事例显示



常规双径迹事例显示



近邻双径迹事例显示

Part3.数据集介绍——数据集构成

数据集规模与划分

- 单径迹事例数据集：涵盖电子、μ 子、π 介子、K 介子、质子共5类典型带电粒子，训练集与验证集合并包含每类粒子的正反粒子（如 π^+ 与 π^-）；
- 双径迹事例数据集：分为常规双径迹与近邻双径迹两个子类，各自独立构建训练集与验证集。

训练/验证集划分

10^5

每类事例规模

9 : 1

训练：验证

独立测试集构建

55 万

单径迹测试事例总量

覆盖5类粒子 × 2种电荷态，每类/态包含约 5.5 万个独立测试事例。

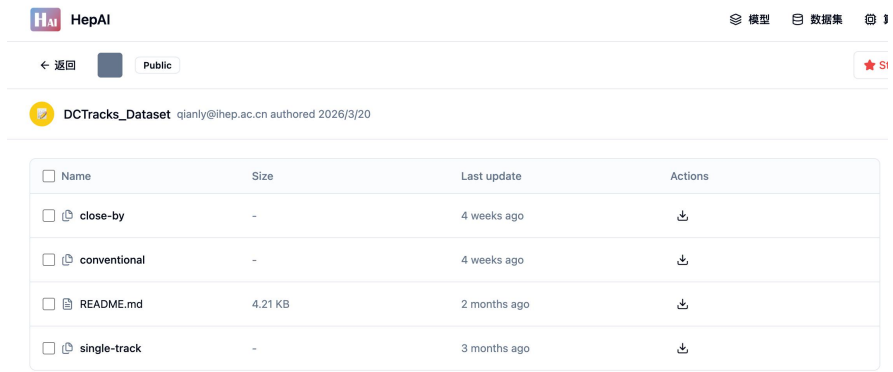
5 万

双径迹测试事例总量

覆盖每类双径迹约 2.5 万个独立测试事例。

数据集获取

- 所人工智能平台：ai.ihep.ac.cn



- 该数据集向外部研究人员开放共享，有使用意向者请发送邮件至 hepai@ihep.ac.cn 提交申请。

Part3.数据集介绍——特征与标签定义

数据集以逗号分隔值（CSV）格式存储，每个 CSV 文件包含固定数量的事例，文件中通过运行号run和事例号event两个字段唯一标识每一个事例；同一事例下的所有行对应同一事例中漂移室单元记录到的击中信息。每一行代表一个击中，包含两类数据：输入特征和监督学习所需的标签。

击中输入特征

特征名称	描述
空间特征	
middleX, middleY	信号丝两端中点的笛卡尔坐标（单位：cm），表征击中在 xy 平面上的投影位置。
wire,layer, slayer, locallayer	wire 为漂移单元号，信号丝的层级标识：layer 为全局层索引（0–42），slayer 为超层索引（0–10），locallayer 为超层内的局部层索引（0–3，最外层超层 0–2）。
测量特征	
rawDriftDist	由漂移时间经初始 T-X 关系刻度得到的漂移距离（单位：cm）。
rawDriftDistErr	对应漂移距离的估计不确定度（单位：cm）。

Part3.数据集介绍——特征与标签定义

击中级标签

■ 击中级标签为每个击中分配来自 MC 模拟的真值信息，用于监督分类、噪声抑制及击中-径迹关联等任务。

标签名称	描述
isSignal	二值标签：1 表示信号击中，0 表示噪声击中。
trackIndex	击中归属的模拟粒子唯一标识符。同一粒子产生的所有信号击中具有相同的正整数 trackIndex，噪声击中为 0。
posX, posY, posZ	击中的三维空间坐标（单位：GeV/c），噪声击中为 0。
momX, momY, momZ	击中的三维动量分量（单位：GeV/c），噪声击中为 0。
PID	击中类型标识符，采用 PDG 编号：e^-/e^+ 为 ±11，μ^-/μ^+ 为 ±13，π^+/π^- 为 ±211，K^+/K^- 为 ±321，$p/\bar{p}$ 为 ±2212，噪声击中为 0。
scaledFltLen	从粒子产生顶点到击中位置的飞行长度，经对应螺旋周长的归一化，可用于拟合辅助，噪声击中为 0。
lrAmbig	左右模糊性标志指示击中位于信号丝的左侧或右侧：-1 表示左侧，1 表示右侧，噪声击中为 0。
turnId	粒子螺旋运动的圈数与方向标识：绝对值为圈数（1= 一圈、≥2= 第 n 圈），正负号表示飞行方向（+ 向探测器外/- 向探测器内），0 为噪声击中。

Part3.数据集介绍——特征与标签定义

径迹级标签

径迹级标签关联至每条模拟粒子径迹，提供径迹全局参数的真值，用于监督回归任务。这些参数定义在粒子最靠近原点的 POCA 点处，包括该点的三维动量、空间坐标以及径迹的带电量。由于归属于同一条径迹的所有击中共享相同的径迹级标签，模型可在击中特征聚合后预测整条径迹的连续参数，实现从击中集合到径迹参数的直接映射。

标签名称	描述
initialMomX, initialMomY, initialMomZ	粒子在最靠近原点 $O(0,0,0)$ 的 POCA 点处的动量三分量（单位：GeV/c）。
initialPosX, initialPosY, initialPosZ	粒子在最靠近原点处的空间坐标（单位：cm）。
charge	径迹的带电量（+1 或 −1）。

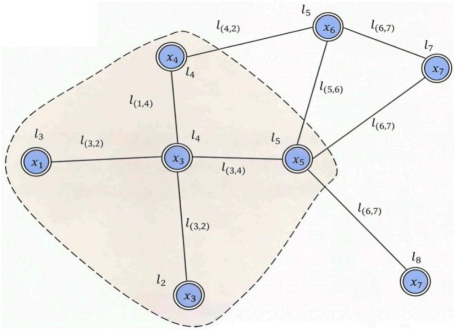
第四部分

基准实验测试

Part4.基准实验测试——GNNs模型

为什么选择GNNs?

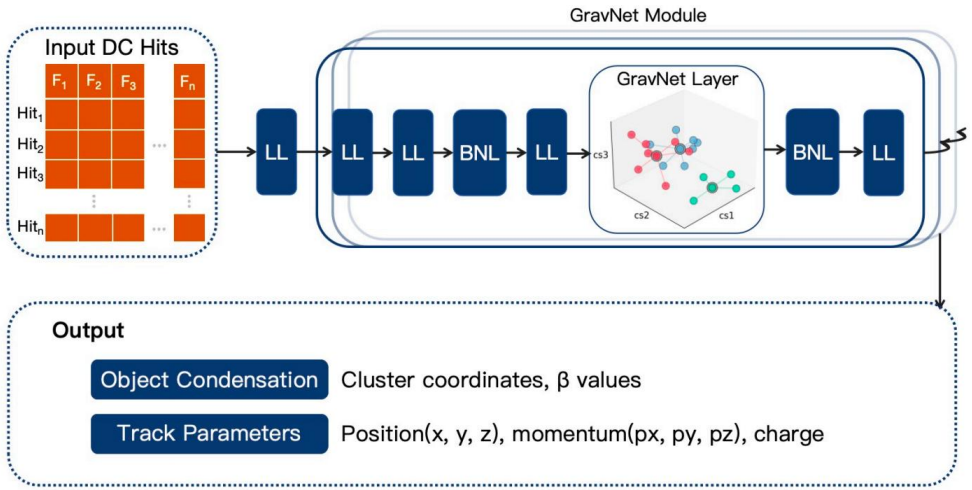
- 漂移室中径迹的数据天然具有图结构特性，且GNNs 能够有效适配不规则的探测器几何结构；每个击中视为图中的一个节点，节点特征即击中点的各个特征；而击中之间的空间关联则定义为边，从而将整个事例建模为一个图。
- GNNs可实现端到端重建（End-to-End），能直接输出径迹候选及其动量和顶点信息等。

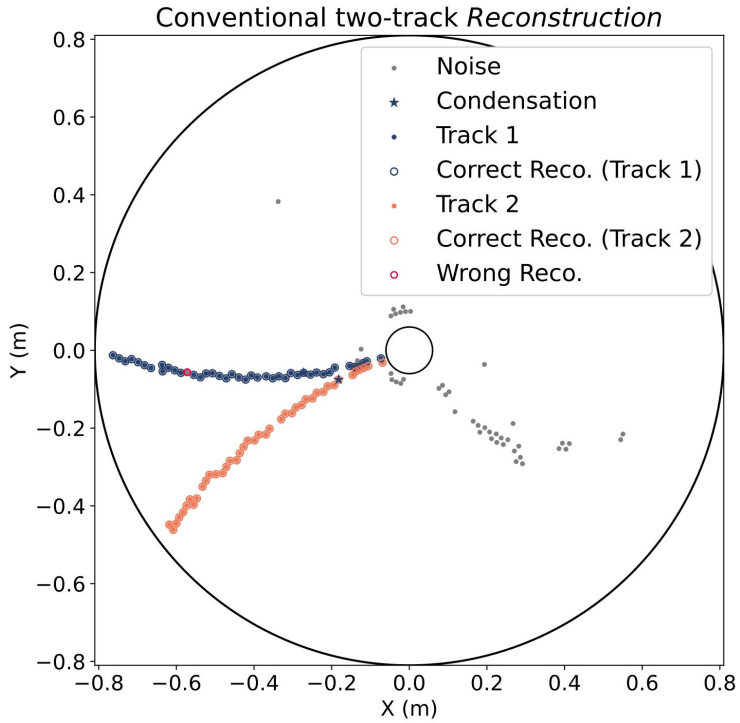


GNNs中的图结构

模型结构

- 以漂移室击中信号为直接输入，通过端到端的 GNNs 同步完成以下任务，为后续的径迹拟合环节提供高质量的初始值：
  - 预测每个事例中的径迹候选数量；
  - 推断每条径迹的运动学参数，包括三维动量、起点位置、电荷符号；
  - 推断每条径迹关联的击中点集合。





重建事例显示

Part4.基准实验测试——算法性能评估指标

该套多维度评估指标涵盖径迹寻找和径迹拟合两个阶段，利用 MC 真值信息，从击中效率与纯度、径迹效率、径迹电荷效率、电荷误判率、克隆径迹率与假径迹率以及径迹参数分辨等多个角度量化算法表现，为后续分析提供统一基准。

击中效率

径迹上与 truth 匹配的击中数与该径迹上可探测击中数之比：

$$\epsilon_{\text{hit}} = \frac{N_{\text{hit}}^{\text{matched}}}{N_{\text{hit}}^{\text{detectable}}}$$

$N_{\text{hit}}^{\text{matched}}$ 是径迹上与 truth 匹配的击中数，$N_{\text{hit}}^{\text{detectable}}$ 是该径迹上可探测击中数（即经过本底混合、数字化、阈值筛选和探测器效率损耗后的击中数）。

击中纯度

径迹上与 truth 匹配的击中数与重建径迹的击中数之比：

$$p_{\text{hit}} = \frac{N_{\text{hit}}^{\text{matched}}}{N_{\text{hit}}^{\text{assigned}}}$$

$N_{\text{hit}}^{\text{assigned}}$ 是重建径迹上的击中数。

匹配判据

- 在此，我们将至少拥有 6 个可探测击中的模拟粒子定义为一条可探测径迹。
- 若一条重建径迹满足以下径迹匹配判据，则将其判定为匹配径迹：$p_{\text{hit}} > 0.50$、$\epsilon_{\text{hit}} > 0.20$ 且 $N_{\text{hit}}^{\text{matched}} \geqslant 6$。
- 将不满足纯度或效率要求的重建径迹定义为假径迹（fake track）。若多条重建径迹同时满足同一条可探测径迹的匹配判据，则保留 ϵ_{hit} 最高的候选径迹作为匹配径迹，其余径迹称为克隆径迹（clone track）。

Part4.基准实验测试——算法性能评估指标

径迹效率

与 truth 匹配的匹配径迹数与可探测径迹数之比：

$$\epsilon_{\text{track}} = \frac{N_{\text{track}}^{\text{matched}}}{N_{\text{track}}^{\text{detectable}}}$$

$N_{\text{track}}^{\text{detectable}}$ 表示样本中可探测径迹数，
$N_{\text{track}}^{\text{matched}}$ 表示是与 truth 匹配的匹配径迹数。

径迹电荷效率，电荷误判率

被正确（错误）重建出电荷符号的可探测径迹占可探测径迹的比值：

$$\epsilon_{\text{track,q}} = \frac{N_{\text{track}}^{\text{matched,q-correct}}}{N_{\text{track}}^{\text{detectable}}} \qquad R_{\text{wrong,q}} = \frac{N_{\text{track}}^{\text{matched,q-incorrect}}}{N_{\text{track}}^{\text{detectable}}}$$

克隆率

克隆径迹总数与可探测径迹总数的比值：

$$R_{\text{clone}} = \frac{N_{\text{track}}^{\text{clone}}}{N_{\text{track}}^{\text{detectable}}}$$

其中，$N_{\text{track}}^{\text{clone}}$ 表示样本中克隆径迹的总数。

假径迹率

假径迹总数与可探测径迹总数的比值：

$$R_{\text{fake}} = \frac{N_{\text{track}}^{\text{fake}}}{N_{\text{track}}^{\text{detectable}}}$$

其中，$N_{\text{track}}^{\text{fake}}$ 表示样本中假径迹的总数。

Part4.基准实验测试——算法性能评估指标

径迹横动量分辨

重点关注电荷重建正确的径迹的 p_T。归一化残差定义为：

$$\eta_{p_T} = \frac{p_T^{reco} - p_T^{MC}}{p_T^{MC}}$$

对于无偏重建，η_{pT} 的分布通常符合高斯分布。横动量分辨率则量化为绝对残差分布在其中位数附近的 68% 覆盖区间：

$$r(p_T) = P_{68\%}\left(\left|\eta_{p_T} - P_{50\%}(\eta_{p_T})\right|\right)$$

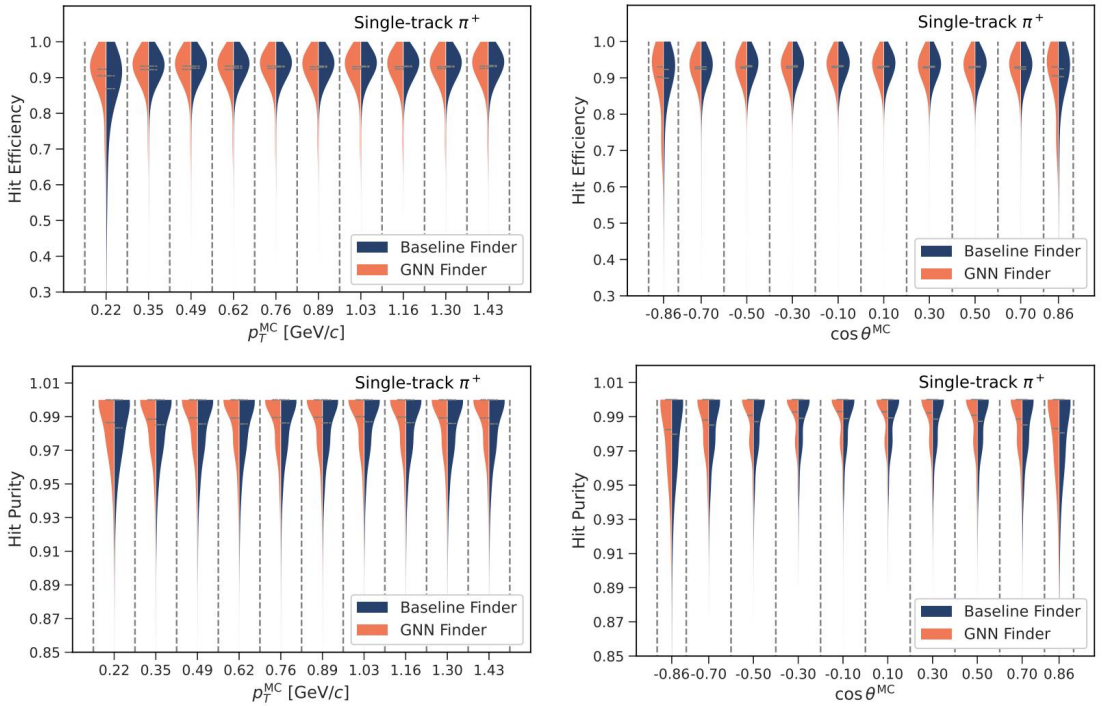
其中 P_q 表示分布的第 q 分位数，$P_{50\%}$ 为中位数。对于正态分布，该值对应于标准差。

Part4.基准实验测试——传统方法vs.GNNs方法性能对比

击中纯度与击中效率

- GNNs方法（GNN Finder） 与 传统方法（Baseline Finder ）共同找到的径迹，其击中效率与击中纯度统计结果汇总如下：

(%)	事例类型	ϵ_{hit}	p_{hit}
Baseline Finder	单径迹 (π^+)	$92.24^{+0.12}_{-0.12}$	$98.58^{+0.05}_{-0.05}$
GNN Finder		$92.20^{+0.12}_{-0.12}$	$98.91^{+0.05}_{-0.05}$
Baseline Finder	常规双径迹 ($\pi^+\pi^-$)	$90.87^{+0.14}_{-0.14}$	$97.93^{+0.07}_{-0.07}$
GNN Finder		$91.62^{+0.13}_{-0.13}$	$98.83^{+0.05}_{-0.05}$
Baseline Finder	近邻双径迹 ($\pi^+\pi^-$)	$91.26^{+0.16}_{-0.16}$	$97.95^{+0.08}_{-0.08}$
GNN Finder		$82.68^{+0.21}_{-0.21}$	$97.89^{+0.08}_{-0.08}$



击中效率与击中纯度随 p_T^{MC}和 cos_θ^{MC} 的变化

结论：在单径迹与常规双径迹上， *GNN Finder* 的击中效率与 *Baseline Finder* 基本相当，击中纯度则略高于传统算法，表明在无其它径迹干扰的情况 下，GNNs 方法不仅能达到与传统算法相当的基本性能水平，还在击中纯度上展现出小幅优势。

Part4.基准实验测试——传统vs.GNNs性能对比

径迹寻找与拟合效率——单径迹

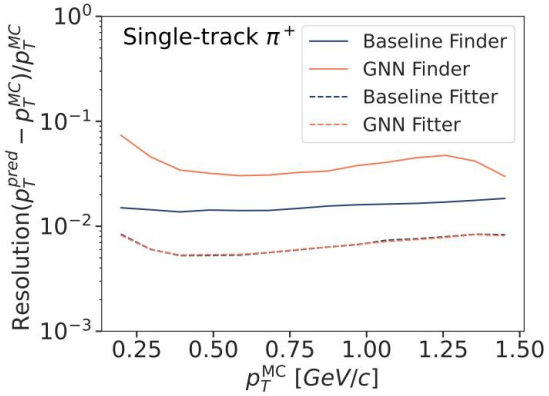
- GNNs方法（GNN Finder） 与 传统方法（Baseline Finder ）共同找到的径迹，其径迹寻迹与拟合效率统计结果汇总如下：

结论1：在径迹重建效率与拟合效率上，GNNs方法与传统方法相当；克隆抑制率GNNs方法更明显。

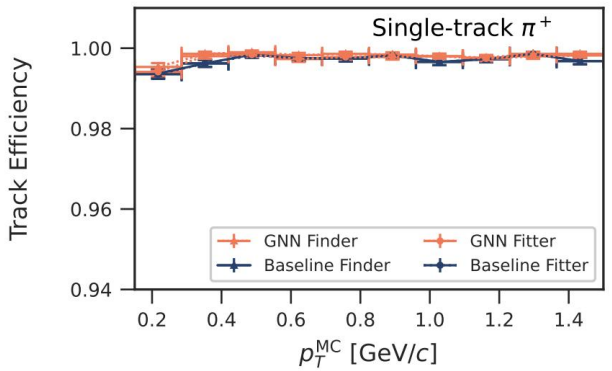
(%)	ϵ_{track}	$\epsilon_{\text{track,q}}$	R_{clone}	R_{fake}	$R_{\text{wrong,q}}$
单径迹 π^+ 事例					
Baseline Finder	$99.71^{+0.02}_{-0.02}$	$99.69^{+0.02}_{-0.02}$	$0.07^{+0.01}_{-0.01}$	0.01	$0.02^{+0.01}_{-0.01}$
GNN Finder	$99.81^{+0.02}_{-0.02}$	$99.55^{+0.03}_{-0.03}$	0.00	0.01	$0.27^{+0.02}_{-0.02}$
Baseline Fitter	$99.70^{+0.02}_{-0.02}$	$99.68^{+0.03}_{-0.03}$	$0.06^{+0.01}_{-0.01}$	0.01	$0.02^{+0.01}_{-0.01}$
GNN Fitter	$99.75^{+0.02}_{-0.02}$	$99.50^{+0.03}_{-0.03}$	0.00	0.01	$0.25^{+0.02}_{-0.02}$

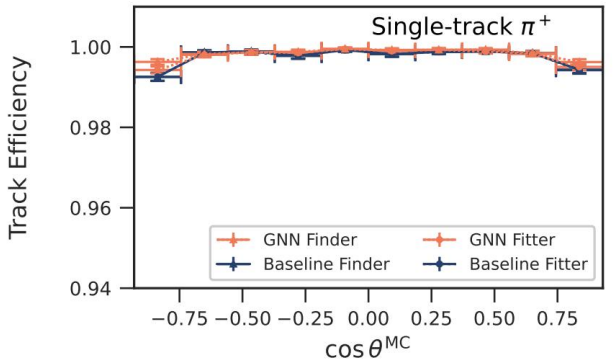
* 统计不确定度 < 0.01% 者，表中不予标注。

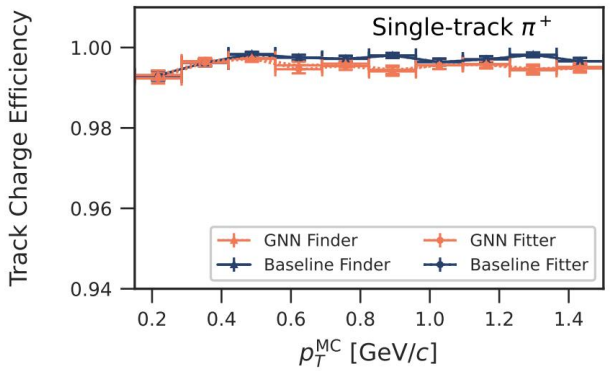
结论2：经过拟合后的径迹，GNNs方法与传统方法在横动量分辨上效果相当。

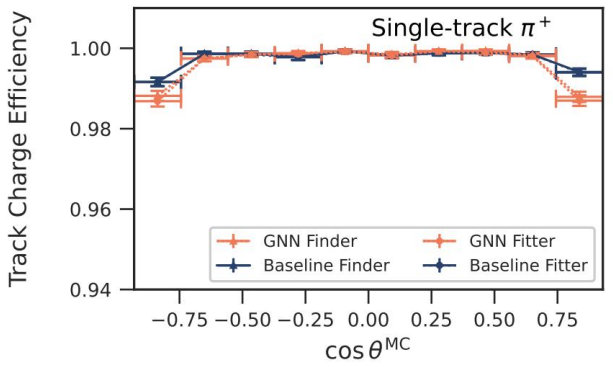


径迹横动量分辨随 p_T^{MC} 的变化









径迹效率和径迹电荷效率随 p_T^{MC} 和 cos_θ^{MC} 的变化

第五部分

工作总结

Part5.工作总结

工作总结

■ **核心成果：数据集与评估体系构建**

首次建立了第一个基于漂移室的机器学习径迹重建开放数据集，完整覆盖单径迹、常规双径迹及近邻双径迹三类典型物理事例，为不同机器学习算法对比提供了统一基准。

同时首次搭建了BESIII漂移室径迹重建的全流程算法框架（数据产生-寻迹-拟合-分析），可迁移应用于CGEM探测器以及COMET等其他粒子物理实验；

■ **初步结论：GNNs算法在BESIII漂移室上的可行性**

训练了BESIII上第一个端到端的GNN的径迹重建模型，发现GNNs方法在常规场景性能与传统方法相当，验证了可行性。

请各位老师批评指正！

BACK UP

- 单径迹 π^+，MC 真值以及由 Baseline Finder 和 GNN Finder 同时寻得并拟合的径迹的五参数（dr 、ϕ^0、κ、dz 和 $\tan\lambda$）。其他单径迹粒子种类的径迹参数分布与 π+ 类似，为简洁起见不再列出。