



STCF实验中基于机器学习的粒子鉴别算法

The particle identification algorithm based on machine learning in the STCF

*报告人: 翟云聪

zhaiyc@mail.sdu.edu.cn

★2025年超级陶粲装置研讨会

2025. 07. 04

為天下儲人材 為國家圖富強



目录

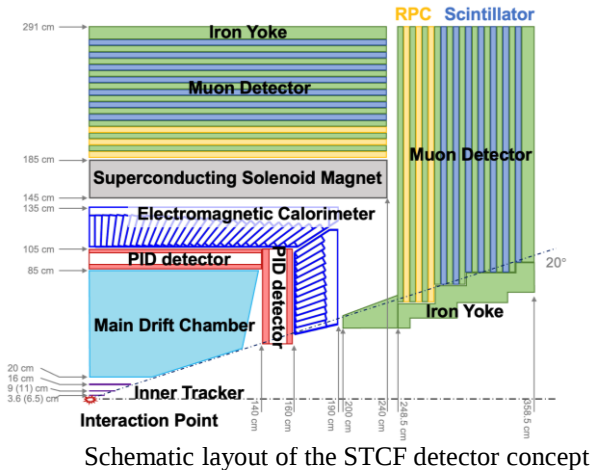
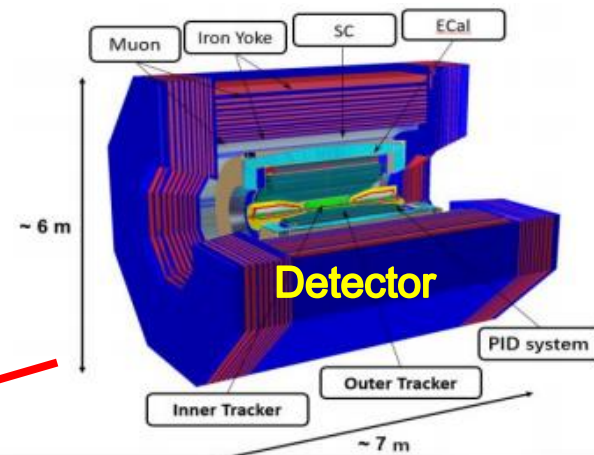
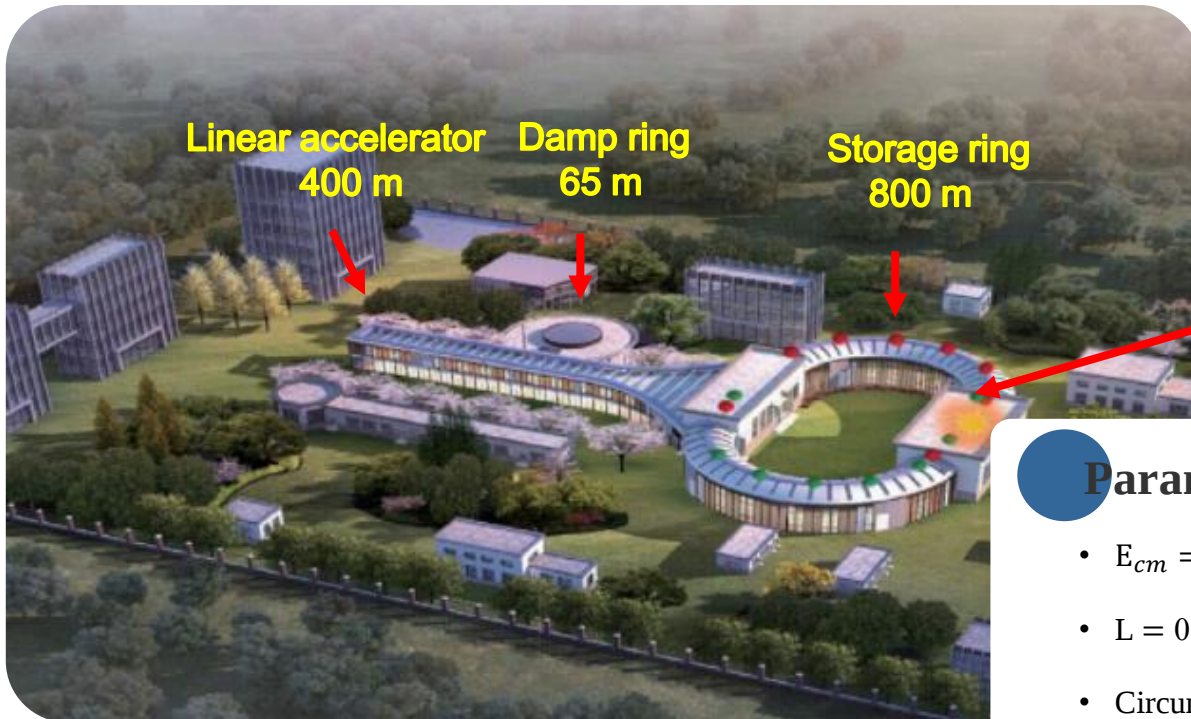
CONTENTS

- 简介
- 带电粒子的GlobalPID算法
- 基于CNN的中性粒子鉴别
- GlobalPID 软件包
- 总结

简介

* **超级陶粲装置 (STCF)** 是中国未来基于加速器粒子物理大科学设施的重要选择之一

- 丰富物理目标：精确测量CKM矩阵元、研究强子和轻子领域中的CP破缺、探索强子谱学以及寻找超越标准模型的新物理现象.....



Parameters of STCF

- $E_{cm} = 2\text{-}7 \text{ GeV}$,
- $L = 0.5 \times 10^{35} \text{ cm}^{-2} \text{ s}^{-1}$
- Circumference: Double-ring, 600-800 m
- Crossing angle: $2 \times 30 \text{ mrad}$

Frontiers of Physics		Contents
ISSN 2019-01-01	Volume 19-N	1
February 2019		2
Introduction		5
Physics		8
2.1 Motivation		8
2.1.1 Challenges in particle physics and the τ -charm energy region		8
2.1.2 Physics potential at the STCF		9
2.2 Charmonium and XYZ physics		12
2.2.1 The XYZ puzzles		12
2.2.2 Limitations of current experiments		13
2.2.3 Opportunities for solving the XYZ puzzles		14
2.2.4 Opportunities in higher charmonium states		15
2.3 Charmed hadron physics		16
2.3.1 Charmed mesons		16
2.3.2 Charmed baryons		20
2.4 Tau physics		23
2.4.1 Precision measurement of the τ properties		23
2.4.2 Determination of the SM parameters		24
2.4.3 CP symmetry tests		26
2.4.4 Flavor-violating τ decays		27
2.5 Topics in QCD studies and light hadron physics		27
2.5.1 QCD physics		28
2.5.2 Spectroscopy		29
2.5.3 Precision tests with light hadrons		30
2.5.4 Tests of the CPT invariance with J/ψ decays		34
2.6 New light particles beyond the SM		36
2.6.1 Particles in the dark sector		37
2.6.2 Millicharged particles		38
2.7 Summary		40

简介

* 作为探索上述粒子物理核心问题的关键支撑之一，STCF对**粒子鉴别(PID)**准确性和效率提出了较高要求

- 依赖于探测器硬件的先进性
- 依赖于粒子鉴别算法的创新，算法的优化决定了能否充分发挥探测器的高性能探测能力

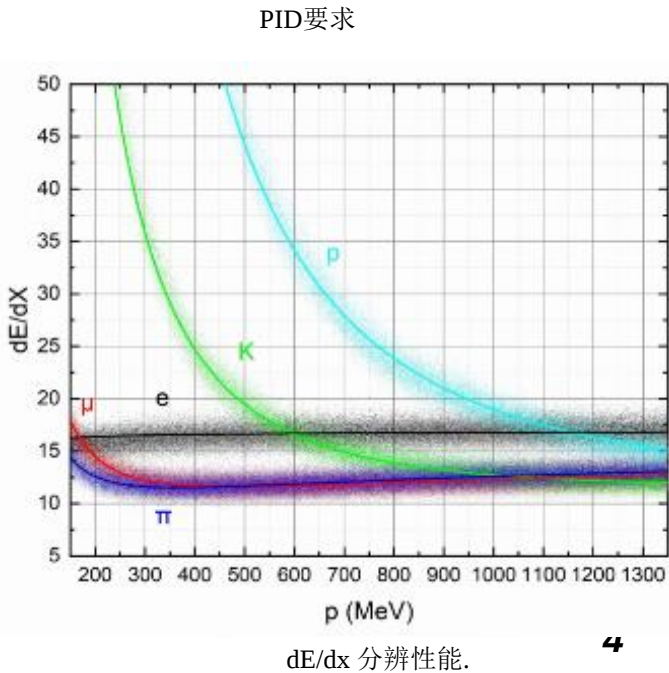
Process	Physics Interest	Optimized Subdetector	Requirements
$\tau \rightarrow K_s \pi \nu_\tau$, $J/\psi \rightarrow \Lambda \bar{\Lambda}$, $D_{(s)}$ tag	CPV in the τ sector, CPV in the hyperon sector, Charm physics	ITK+MDC	acceptance: 93% of 4π ; trk. eff.: > 99% at $p_T > 0.3$ GeV/c; > 90% at $p_T = 0.1$ GeV/c $\sigma_p/p = 0.5\%$, $\sigma_{\gamma\phi} = 130 \mu\text{m}$ at 1 GeV/c
$e^+e^- \rightarrow KK + X$, $D_{(s)}$ decays	Fragmentation function, CKM matrix, LQCD etc.	PID	π/K and K/π misidentification rate < 2% PID efficiency of hadrons > 97% at $p < 2$ GeV/c
$\tau \rightarrow \mu\mu\mu$, $\tau \rightarrow \gamma\mu$, $D_s \rightarrow \mu\nu$	cLFV decay of τ , CKM matrix, LQCD etc.	PID+MUD	μ/π suppression power over 30 at $p < 2$ GeV/c, μ efficiency over 95% at $p = 1$ GeV/c
$\tau \rightarrow \gamma\mu$, $\psi(3686) \rightarrow \gamma\eta(2S)$	cLFV decay of τ , Charmonium transition	EMC	$\sigma_E/E \approx 2.5\%$ at $E = 1$ GeV $\sigma_{\text{pos}} \approx 5$ mm at $E = 1$ GeV
$e^+e^- \rightarrow n\bar{n}$, $D_0 \rightarrow K_L \pi^+ \pi^-$	Nucleon structure Unity of CKM triangle	EMC+MUD	$\sigma_T = \frac{300}{\sqrt{p^3(\text{GeV}^3)}} \text{ ps}$

* 传统PID算法- -最大似然法为粒子鉴别提供了一种经典且直观的概率性判别工具

- 单个探测器的粒子鉴别能力往往存在不足，例如dE/dx方法，对高动量 π/K 和 μ/π 等粒子的鉴别存在较大困难
- 在多变量复杂关联条件下进行全局优化较为复杂和困难

* 基于数据驱动的机器学习方法为PID性能的提升开辟了新途径

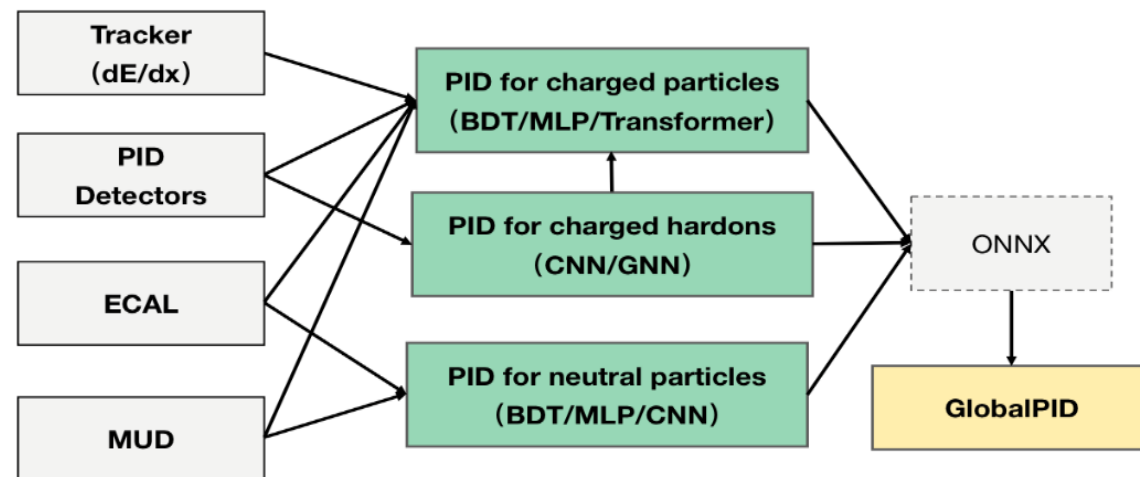
- 较强的建模和泛化能力
- 善于连接来自所有子探测器的信息，对径迹信息进行“智能联合”



简介

* 针对STCF实验，创新开发基于机器学习的粒子鉴别算法

- 带电粒子的GlobalPID算法($e/\mu/\pi/K/p$)
 - 联合所有子探测器的重建信息
 - 使用BDT(基于XGBoost)作为基线
 - 尝试其他ML算法: Transformer...



基于机器学习的STCF粒子鉴别算法

- 中性粒子($\gamma/K_L^0/n$)的鉴别
 - 有效地将光子与中子和 K_L^0 分离
 - ECAL原始信息: 在ECAL内的能量沉积、时间响应, MUD 的 击中响应
 - 基于经典卷积神经网络 (CNN) 在ECAL上开展中性粒子鉴别研究

带电粒子的 GlobalPID算法

I



数据样本

* 数据样本的质量

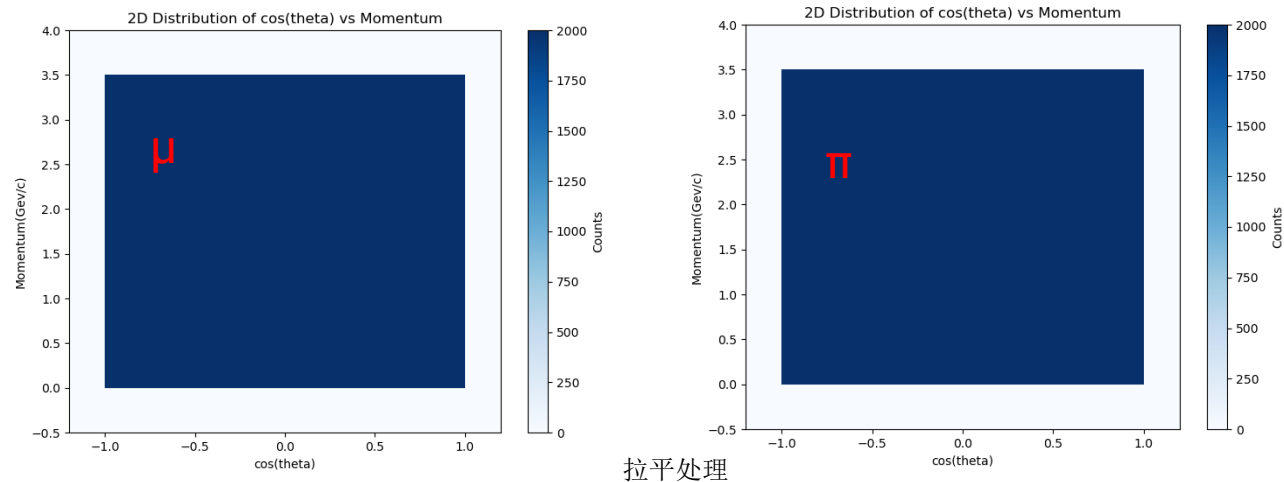
- 高统计量
- 大动量展宽和角度覆盖范围

* 数据生成

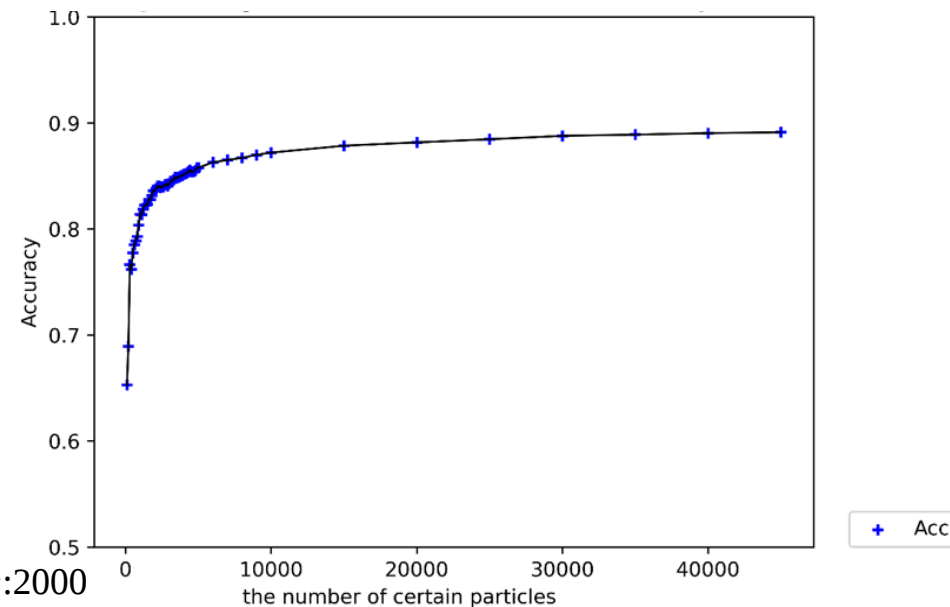
- 基于OSCAR 2.6.2模拟和重建(ITKM+BTOF)
- 使用ParticleGun生成MC单粒子
- 统计量($e^\pm, \mu^\pm, \pi^\pm, K^\pm, p^\pm$): 3×10^6
- $p \in (0.03, 3.5) \text{ GeV/c}$, $\theta \in (0^\circ, 180^\circ)$, $\varphi \in (0^\circ, 360^\circ)$

* 数据预处理

- 将动量和 θ 谱展平, 以避免由 p/θ 分布引起的偏差
 - 动量 (GeV/c): $0 \sim 3.5 \text{ GeV/c}$, 划分 14 个区间
 - 角度 (θ): $-1 \sim 1$, 划分 20 个区间
- 训练:验证:测试 = 8:1:1



拉平处理

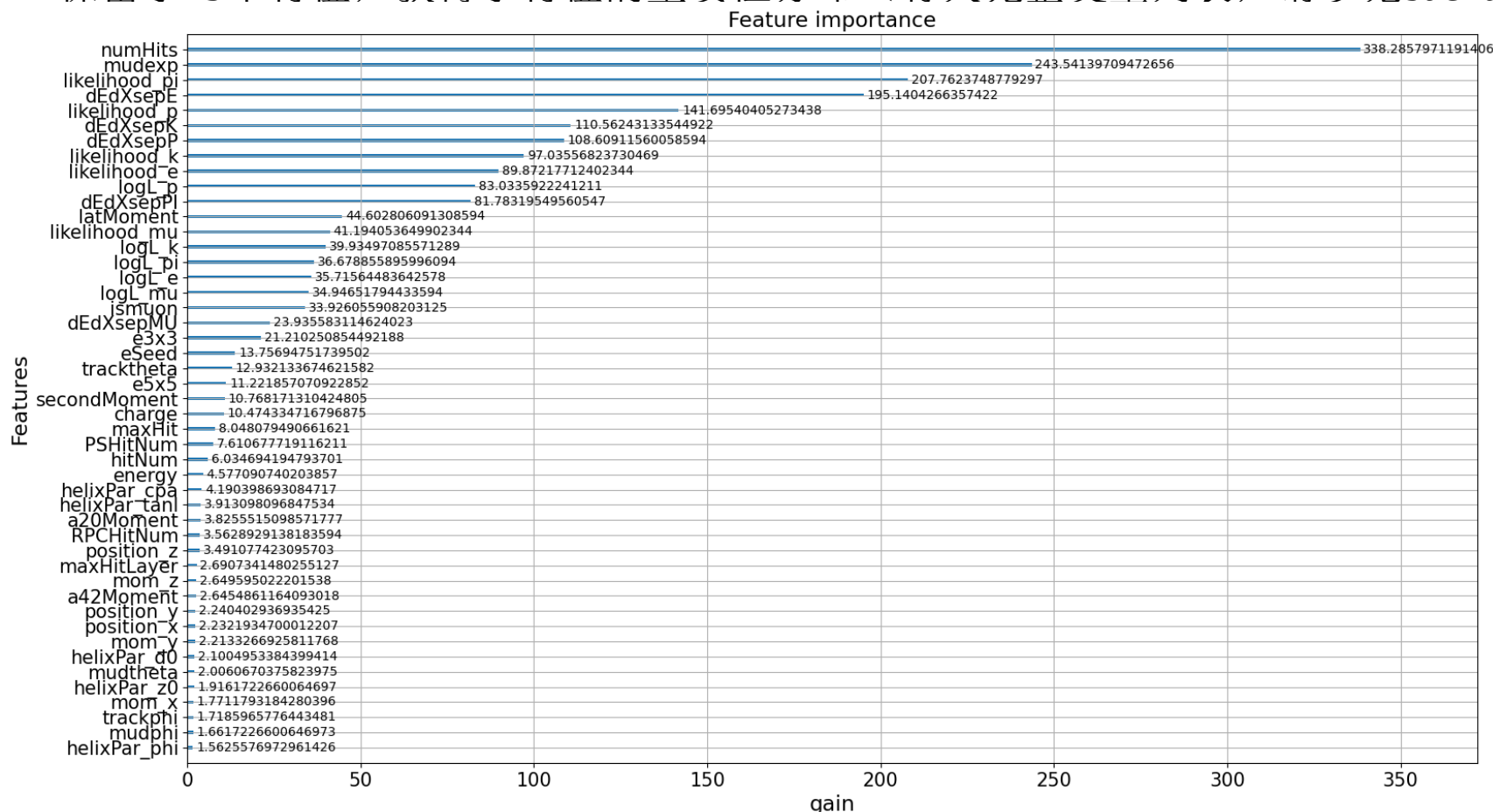


准确性随训练量变化图

特征选择

* 从大量相互关联的子探测器信息中选择最具信息量的特征子集可以帮助稳定模型训练过程

- 已收集了Tracker/dEdx/BTOF/DTOF/ECAL/MUD重建变量（增加MUD信息）
- 保留了49个特征，获得了特征的重要性分布（有关完整变量列表，请参见backup）。



超参数优化

* 目标:BDT超参数的自动优化

- 减少人工干预和时间成本
- 提高模型效率和可靠性

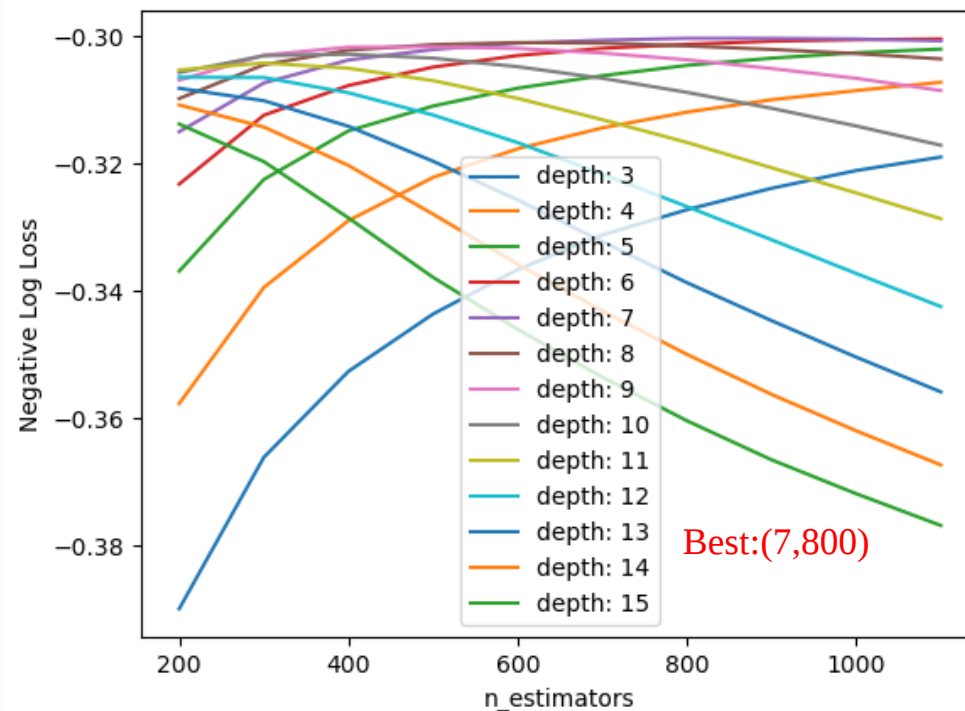
* 基于GridSearchCV获取最优超参数

- 使用带电粒子间的分辨能力作为判据
- max_depth(树深度)的搜索范围:[200,1200]
- n_estimators(分类器个数)的搜索范围:[3,15]

* 最优超参数组合:

- max_depth: 7
- n_estimators: 800

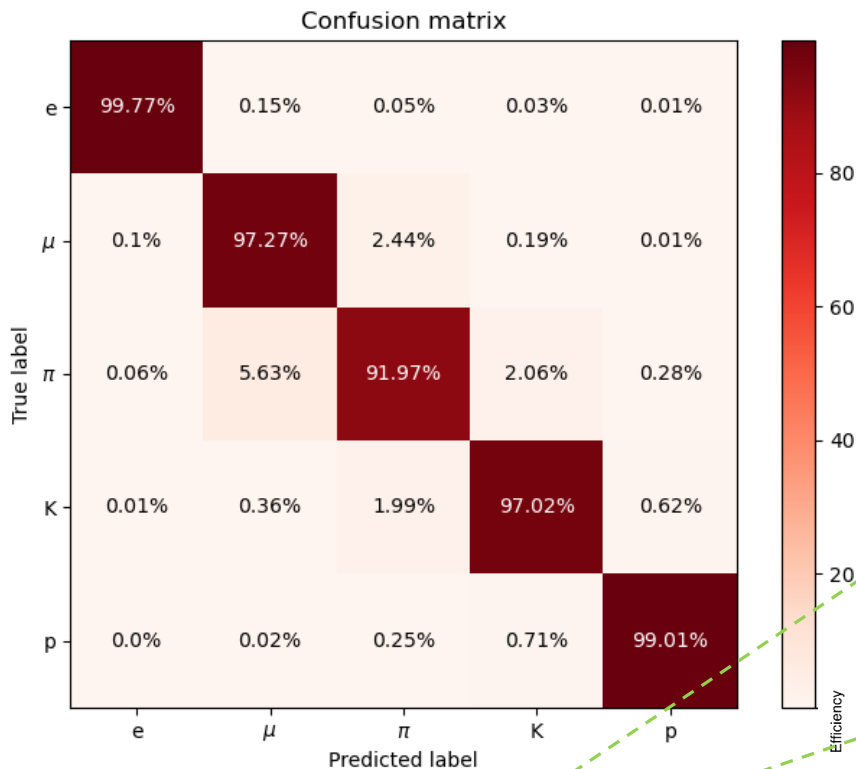
• 超参数调优



性能

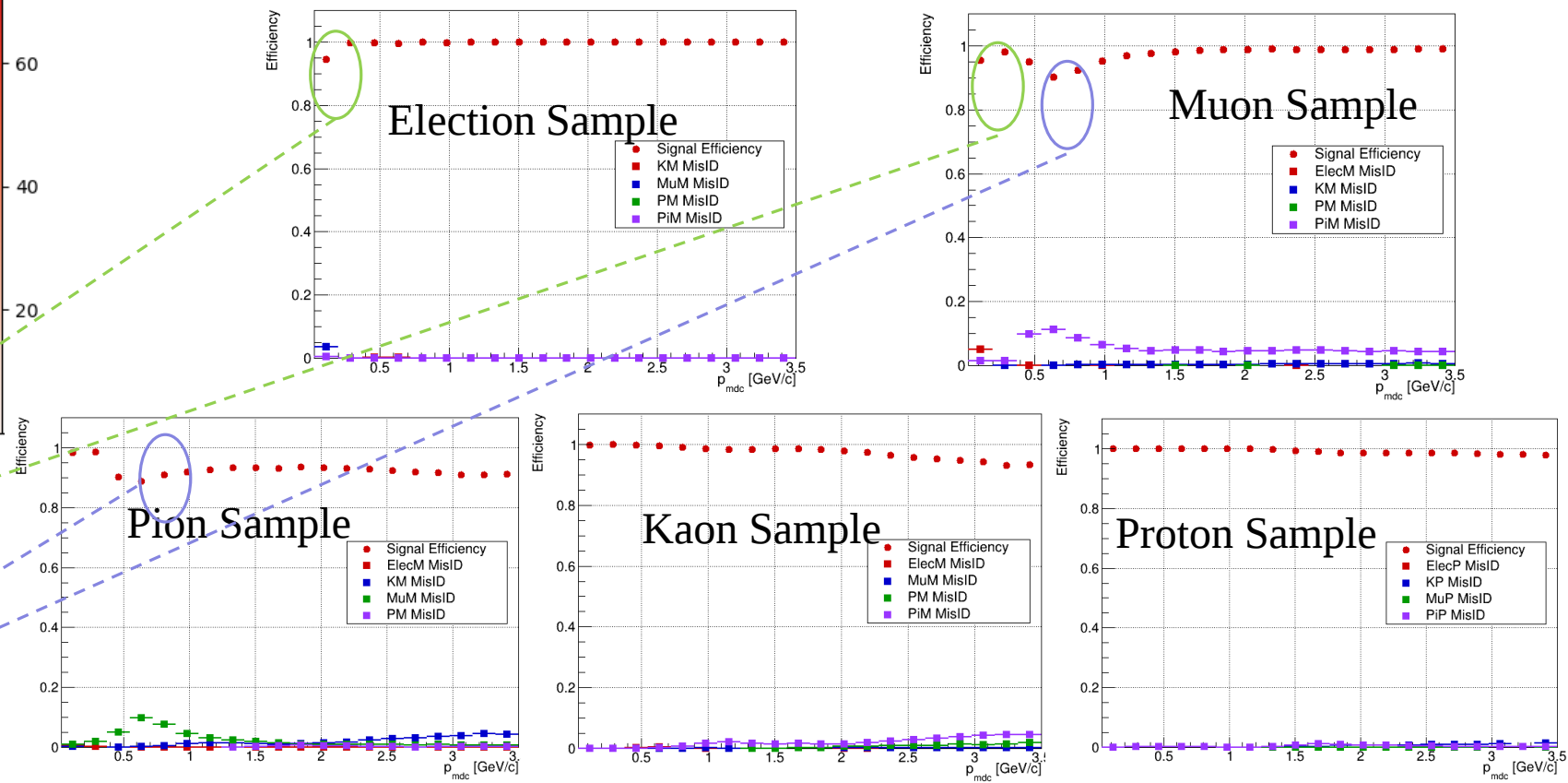
* 在GlobalPID中带电粒子鉴别性能

- 电子和质子的鉴别效率~ 99%
- K 介子的鉴别效率高于 95%
- μ 子的鉴别效率高于 90% (当动量大于 1.5 GeV 时, 效率超过 95%)
- 在大多数动量区间, π 介子的鉴别效率高于 90%



仅有dE/dx在动量低于100MeV时有效, 但是在区分e/ μ 方面性能较差

dE/dx和MUD在区分 π/μ 方面性能相对较差

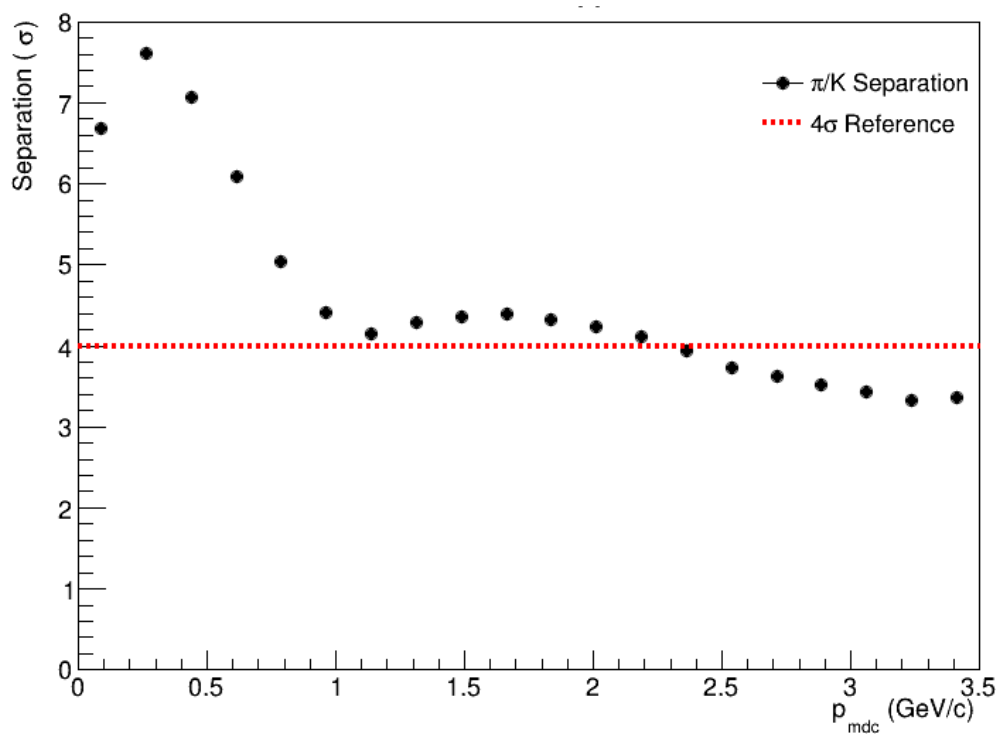


性能

* 不同PID模式下的分辨能力: $\pi/K, \mu/\pi$

- $h_m = \log(p_m(x_1)) - \log(p_m(x_2))$, $p_m(x_i)$ 为 m 粒子在 x_i 粒子假设下的概率
- 归一化概率密度函数: $\text{Scale}(h_m)$

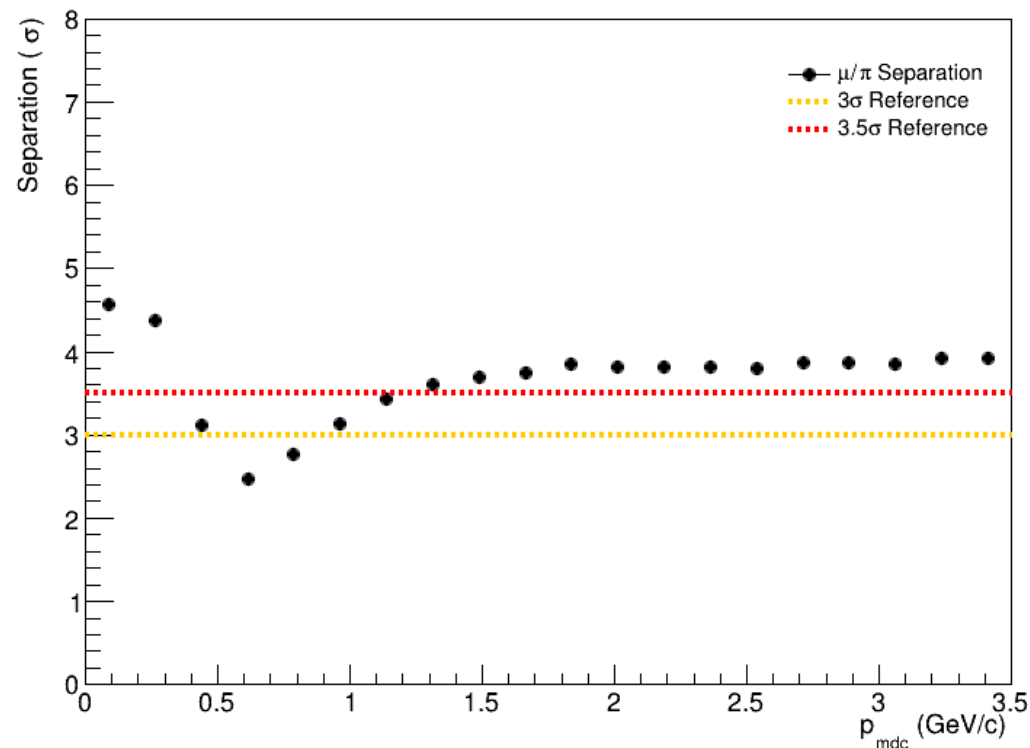
π/K Separation



- $P < 2 \text{ GeV/c}$: 分辨能力超过 4σ
- 全动量空间: 分辨能力超过 3σ

- 计算两分布重叠面积: $C = \int \min[\text{Scale}(h_1), \text{Scale}(h_2)] dx$
- Separation 计算: $\sqrt{2} \cdot \text{erf}^{-1}(1 - C)$, $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$

π/μ Separation

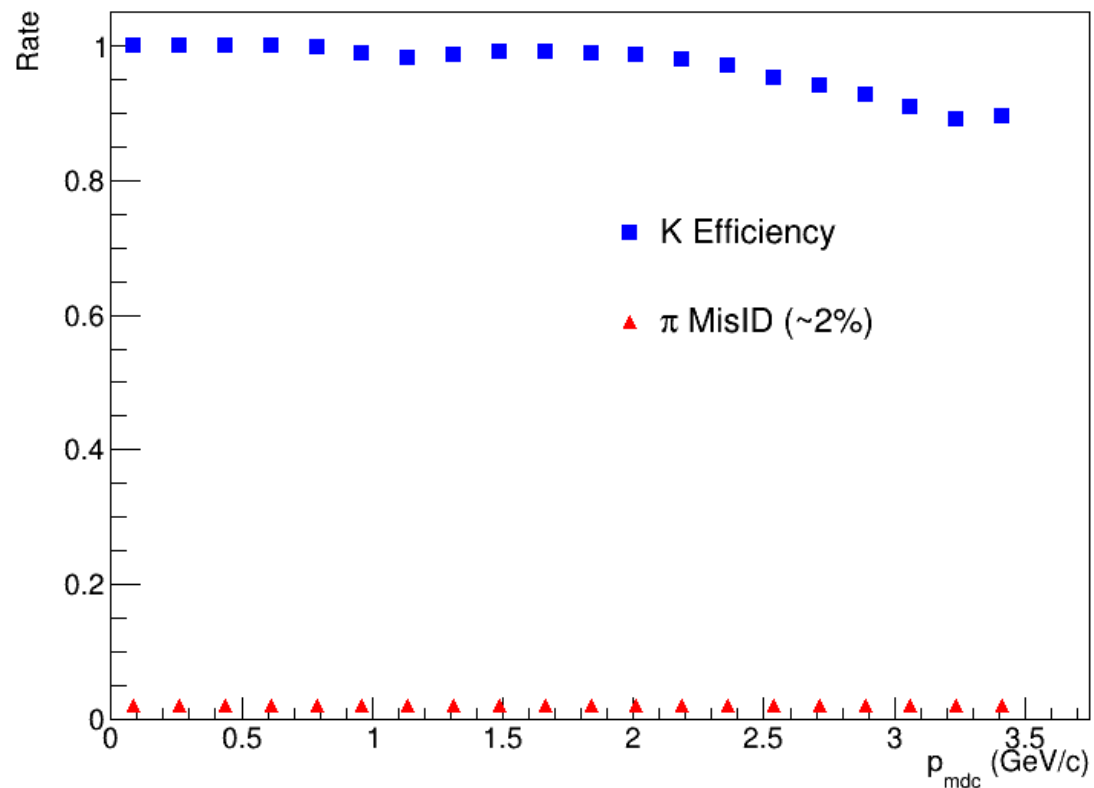


- $P > 1 \text{ GeV/c}$: 分辨能力超过 3σ
- $P > 1.5 \text{ GeV/c}$: 分辨能力超过 3.5σ

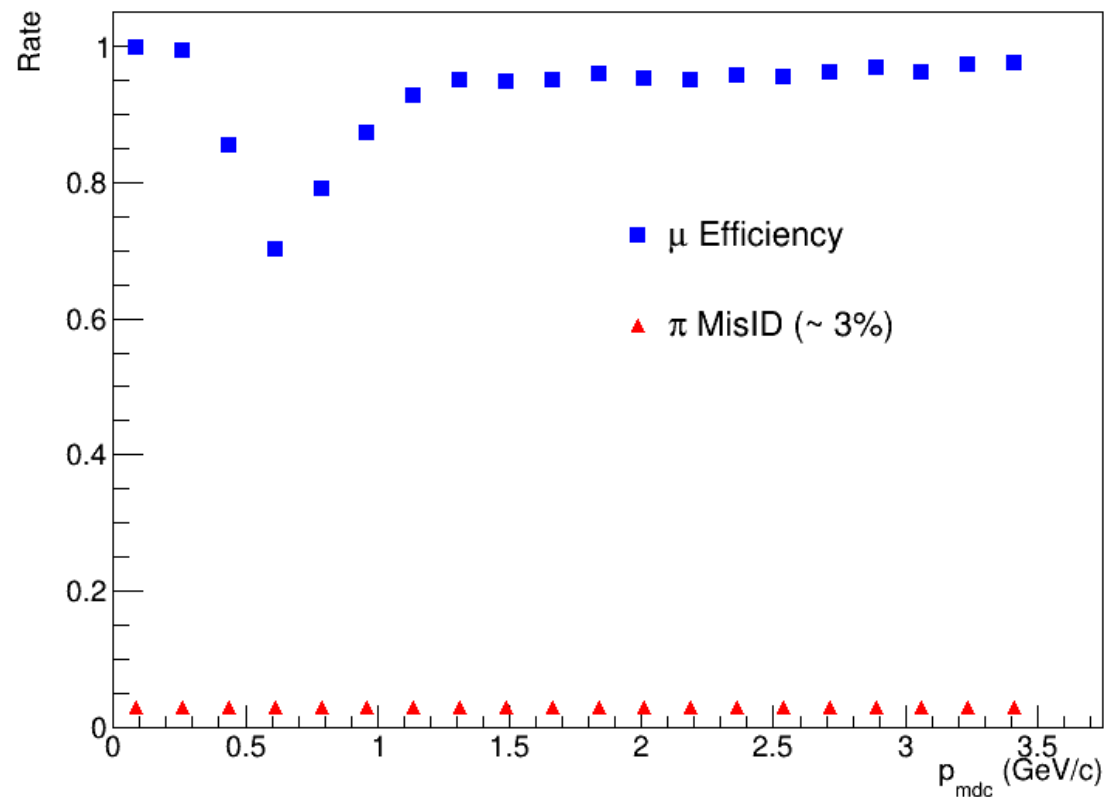
性能

* 不同PID模式下的信号效率/误判率: π/K (~2% π), μ/π (~3% π)

π/K (~2% π)



π/μ (~3% π)



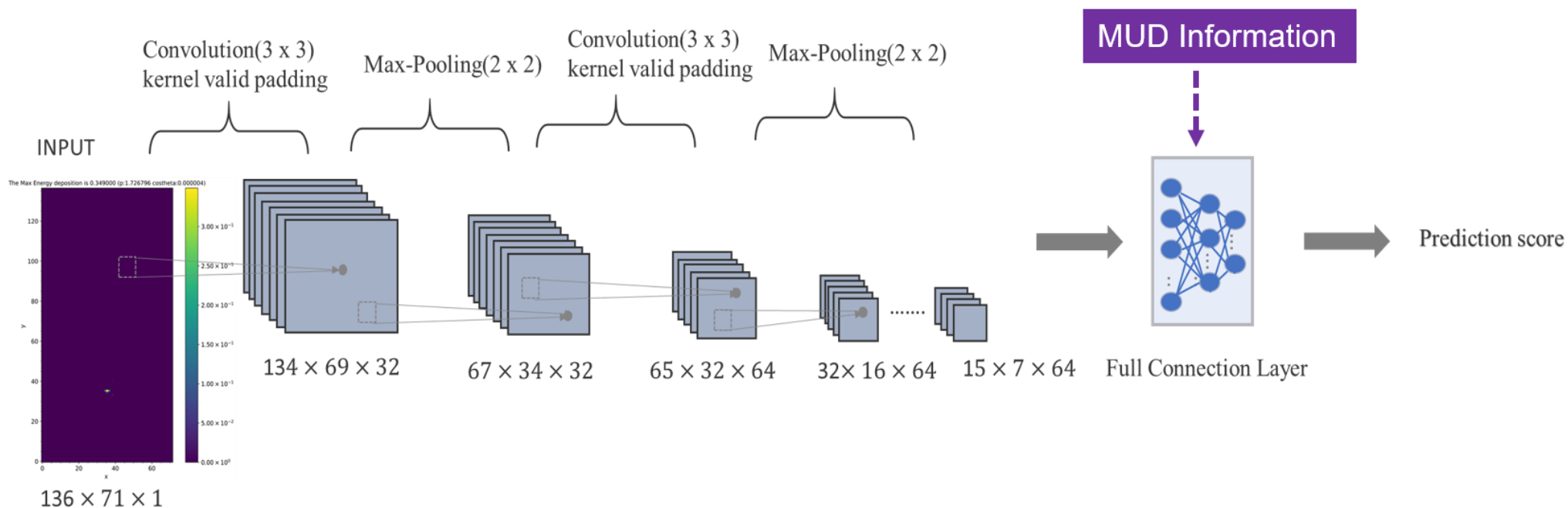
基于CNN的中性 粒子鉴别

II



方法

- * 中性粒子鉴别通常需要综合考虑ECAL的能量沉积、时间响应以及MUD击中信息
- * CNN在图像识别、目标检测和其他领域展现出了出色的性能和应用前景
 - 分层结构：有效地从图像中提取隐藏特征
 - 数据增强：提高模型的泛化能力
 - 局部连接性：有助于提取局部特征
 - 全连接层：将来添加MUD信息
- * 基于CNN的中性粒子鉴别器初步实现



数据样本

* ECAL设计（全吸收量能器）

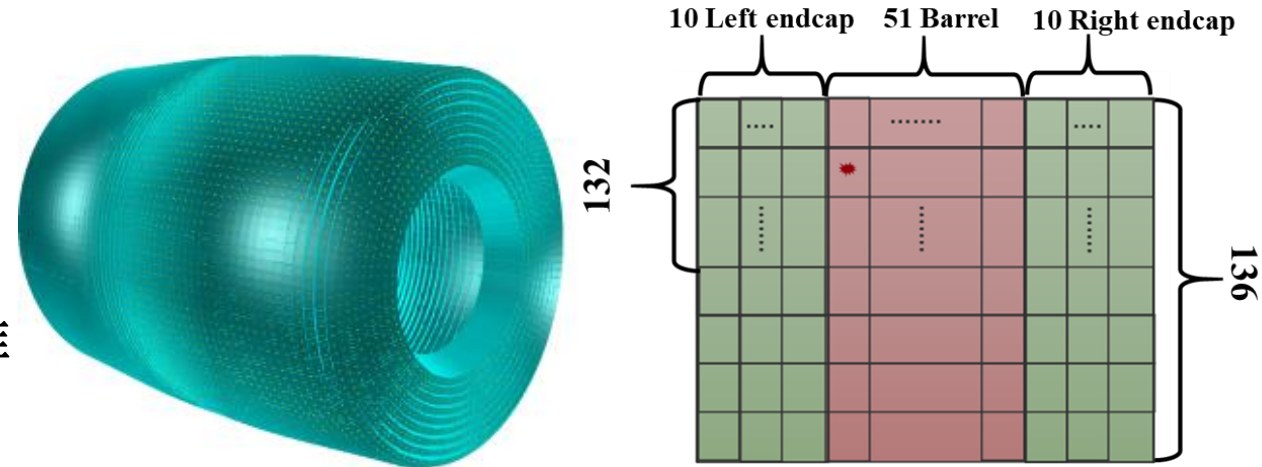
- 桶部: $51 \times 132 = 6732$
- 端盖: $3 \times 85 + 102 + 136 \times 2 = 1938$
- 晶体大小: $5 \times 5 \times 28(15 X_0) \text{ cm}^3$

* 将ECAL中的能量沉积转换为尺寸71 x 136的二维像素图

- X-坐标：**位置信息**
 - 左端盖/ 右端盖（0-9/61-70）
 - 桶部（10-60）
- Y-坐标：**CrystalID**
- 值：**晶体内的能量沉积**

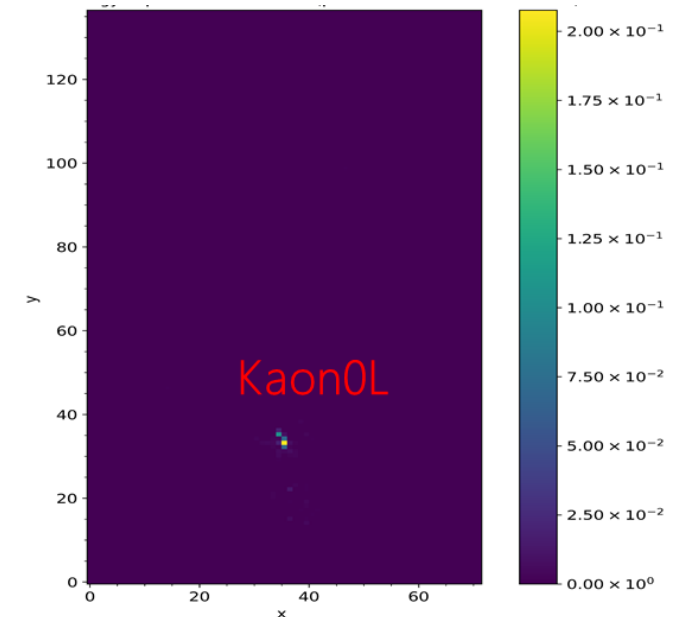
* 中性粒子数据样本

- $\gamma/K_L^0/n$
- ParticleGun产生
- 100,000(每种中性粒子)
- $P \in (0 \text{ y } 2.0 \text{ v GeV/c}, \theta = 90^\circ, \varphi = 0^\circ)$



ECAL模型

能量沉积像素图



K_L^0 的能量沉积

性能

* 分析ECAL中能量沉积分布，给出中性粒子的鉴别性能（初步）

•中子:

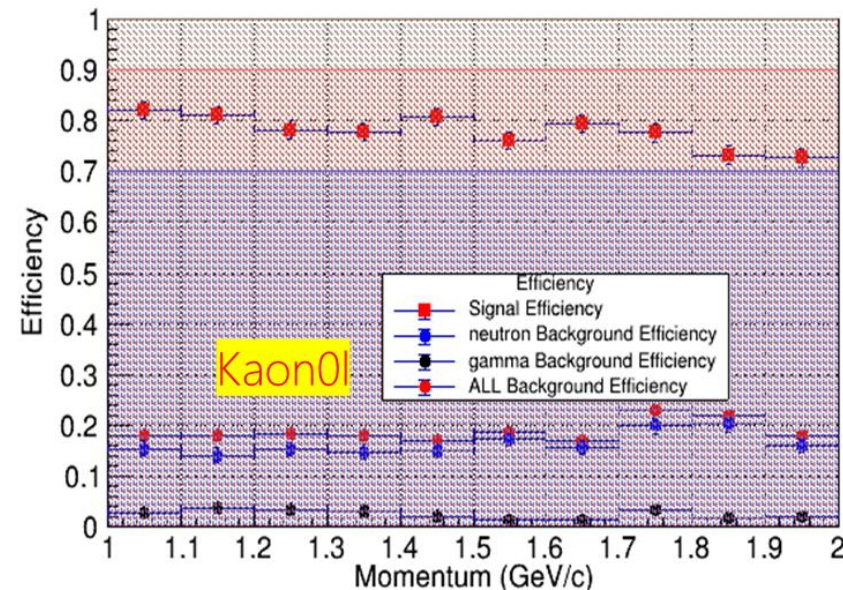
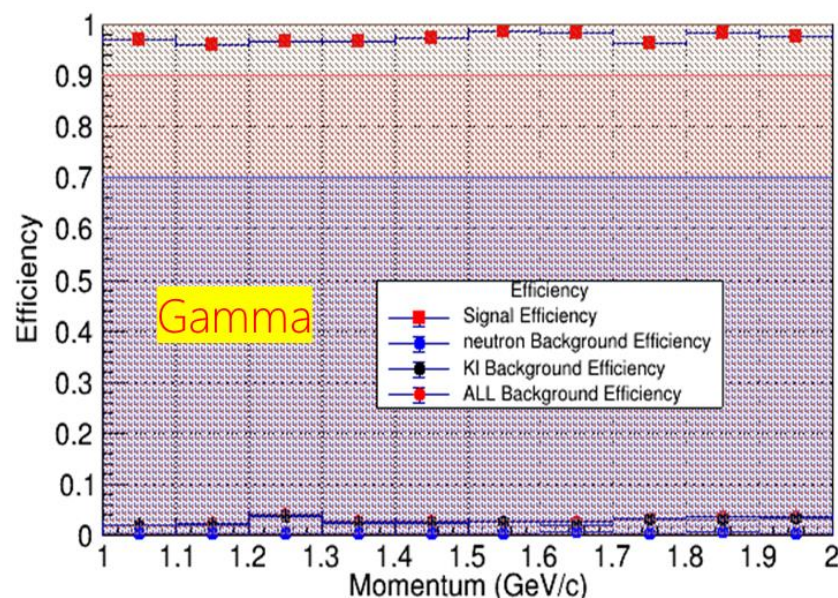
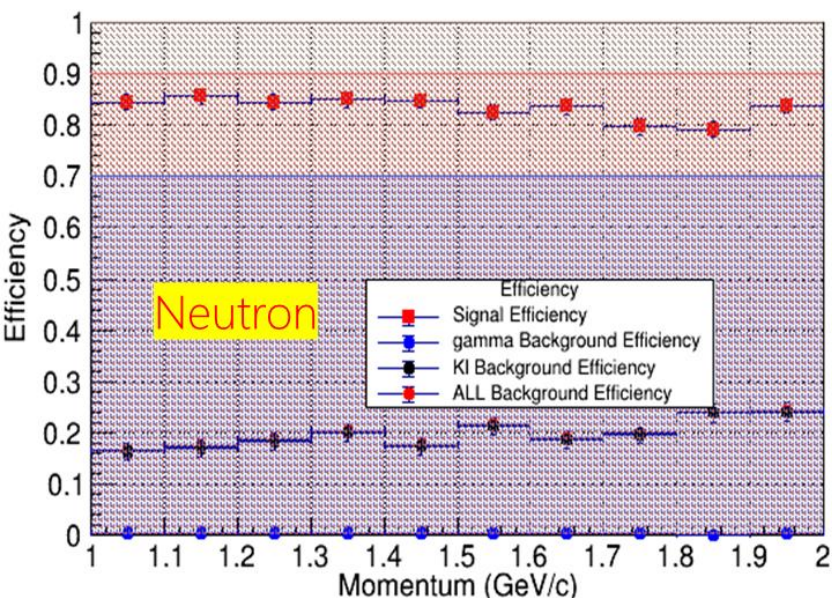
- 信号效率>80%
- 本底误鉴别约为20%，主要是 K_L^0

•光子:

- 良好的光子鉴别性能
- 信号效率> 90%

• K_L^0 :

- 信号效率> 70%
- 本底误鉴别约为20%，主要是中子



中子和 K_L^0 鉴别能力仍需改进

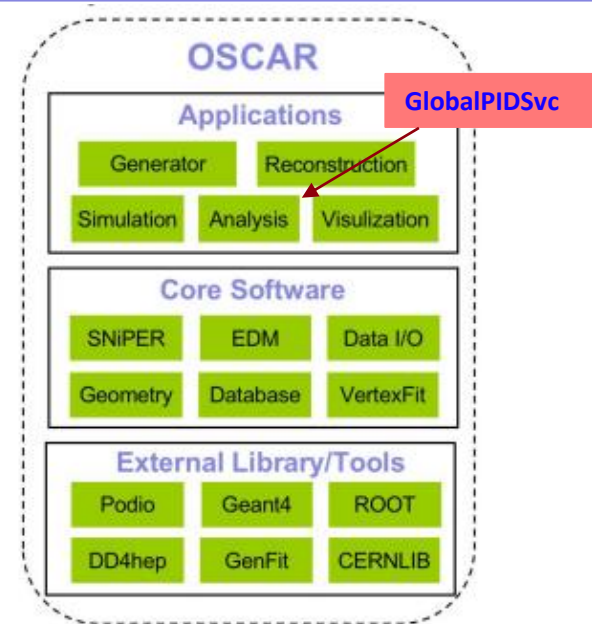
GlobalPIDSvc软件包

* BDT模型和GlobalPID算法已集成到OSCAR软件中，可供分析和研究

- 基于XGBoost的C-API并集成了预训练模型
- 对用户透明，为用户提供了简单的界面和用户手册

* 基于中子粒子鉴别的机器学习软件包有待开发.....

* GlobalPIDSvc包集成:所有软件包都将转移到ONNX框架中.....



Global PID软件包用户手册

GlobalPID软件包使用说明

一、在运行源文件中添加头文件

```
#include "GlobalPID/GlobalPIDSvc.h"
```

二、在运行源文件中关于GlobalPIDSvc调用

```
SniperPtr<GlobalPIDSvc> _globalpidsvc(getParent(), "GlobalPIDSvc");  
if (_globalpidsvc.valid()) {  
    LogInfo << "the GlobalPIDSvc instance is retrieved" << std::endl;  
}  
else {  
    LogError << "Failed to get the GlobalPIDSvc instance!" << std::endl;  
    return false;  
}  
  
m_pid = _globalpidsvc.data();
```

三、获取某一条径迹的径迹信息以及各个子探测器信息

```
m_pid->calculate(RecParticle);
```

四、选择PID模式

```
m_pid->setmode(m_pid->onlyKaon() | m_pid->onlyPion() | m_pid->onlyProton());  
m_pid->setmode(m_pid->onlyPionKaonProton());
```

五、得到该条径迹在五种粒子假设下的预测概率

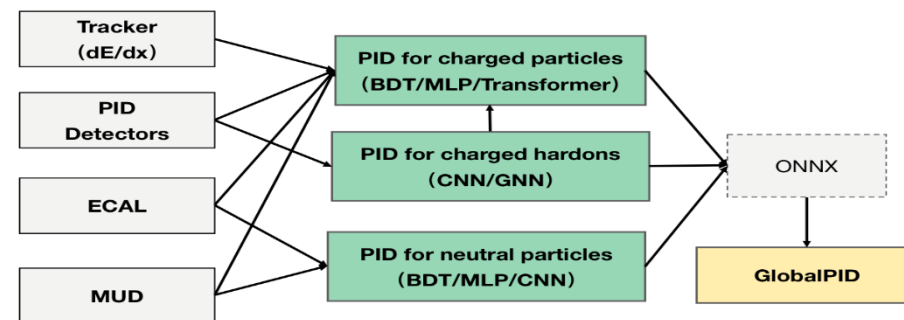
```
float m_prob_e = m_pid->prob(Electron); // (在Electron假设下的预测概率)  
float m_prob_mu = m_pid->prob(Muon); // (在Muon假设下的预测概率)  
float m_prob_pi = m_pid->prob(Pion); // (在Pion假设下的预测概率)  
float m_prob_k = m_pid->prob(Kaon); // (在Kaon假设下的预测概率)  
float m_prob_p = m_pid->prob(Proton); // (在Proton假设下的预测概率)
```

六、Python配置

```
import GlobalPID  
pidsvc = task.createSvc("GlobalPIDSvc")  
pidsvc.property("SetModelPath").set(topdir+"/Analysis/GlobalPID/src/xgb.mod  
el")  
pidsvc.property("SetMethod").set("XGBoost")
```

七、附录

- 目前可支持的PID模式有: All(e/mu/pi/K/p), pi/K/p, pi/K, e/pi/K, mu/pi

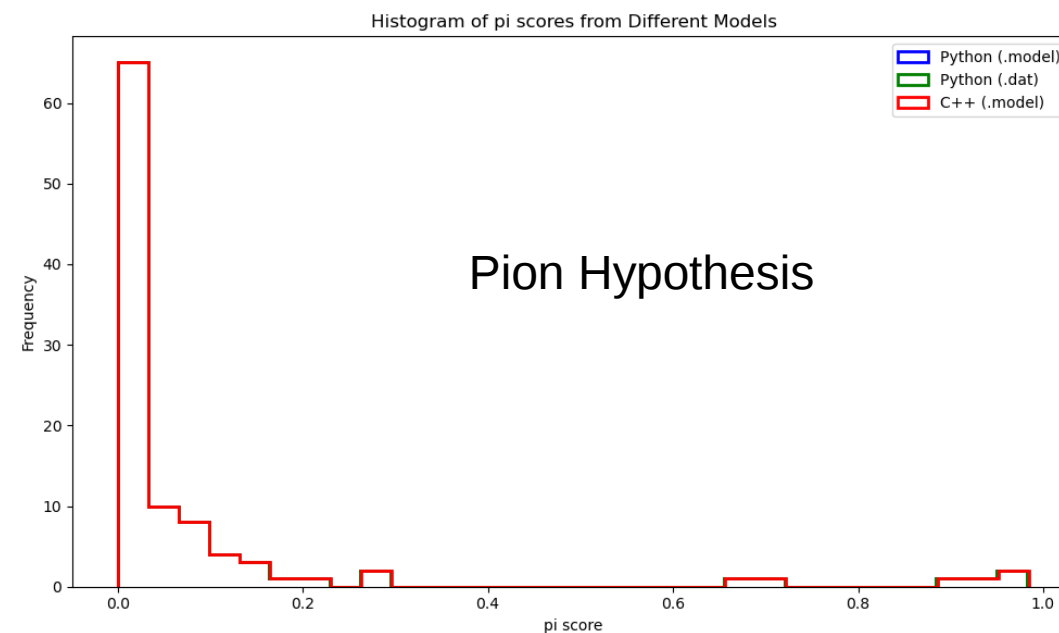
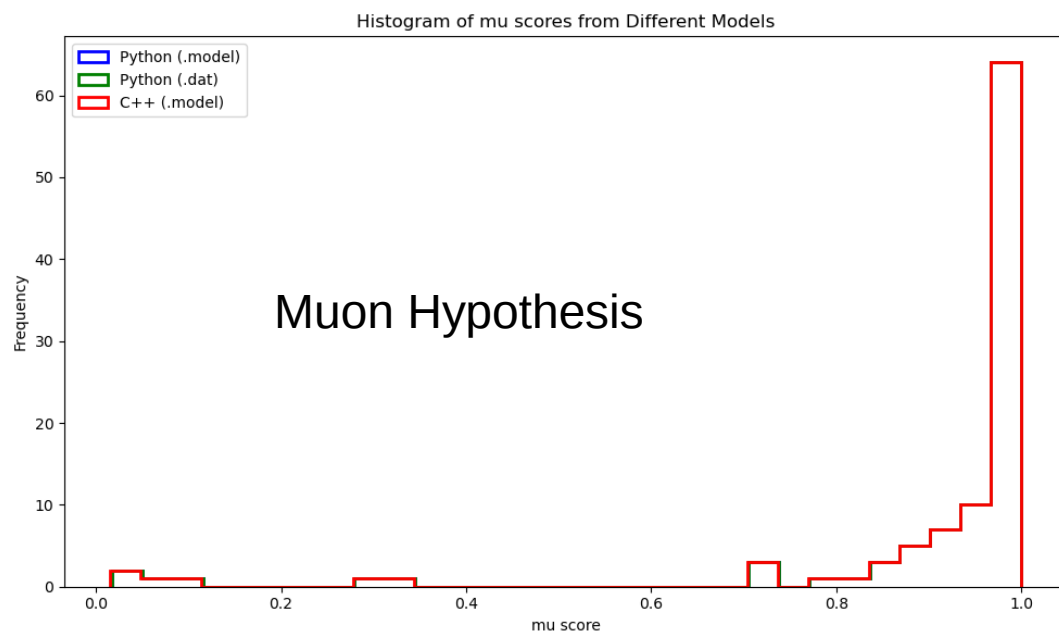


基于机器学习的STCF粒子鉴别算法

GlobalPID软件包

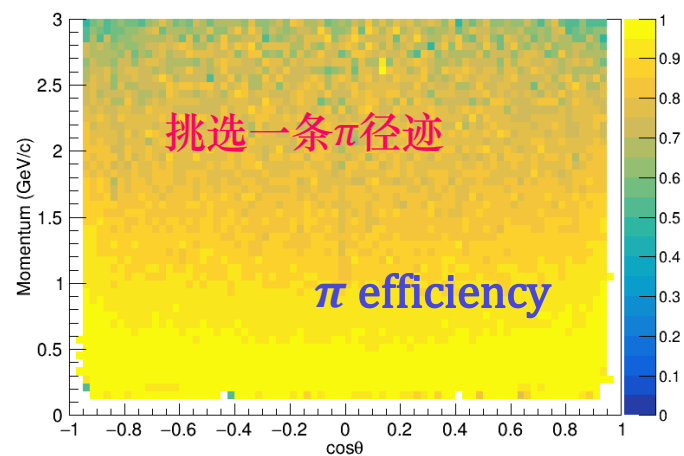
* GlobalPIDSvc软件包输出检查

- μ 单粒子单个事例检查
- Python和C++接口输出预测结果一致



总结

- * 为了充分利用STCF探测器的性能，提出开发利用深度机器学习算法的**PID**软件算法
- * 基于数据驱动方法，**BDT**(基于**XGBoost**) 作为基线被用作在STCF实验中鉴别带电粒子
 - 综合所有子探测器信息
 - 提供不同PID模式下带电粒子鉴别性能： π/K 分辨能力在 2 GeV/c 范围内保持在 4σ 以上，全动量空间范围内可达到 3σ 及以上；在动量大于 1.5 GeV/c 时，对 μ/π 鉴别性能达到 3.5σ 以上
- * 基于**CNN**的全局中性粒子鉴别器已初步实现
- * 初步性能测试已经显示出机器学习模型有能力提供良好的粒子鉴别**PID**性能，但需要进一步检查和验证
- * 搭建**GlobalPID**软件包能够鉴别带电粒子，可用于分析和研究
- * 然而，还需要进一步工作：
 - 使用更先进的模型（如**Transformer**）进一步提升PID性能
 - 在物理分析过程中检验软件包性能($e^+e^- \rightarrow \pi\pi X$ @ 7 GeV Collins效应)
 - 根据物理研究的具体要求优化机器学习模型
 - 所有软件包将转移到**ONNX**框架中



THANKS



Backup



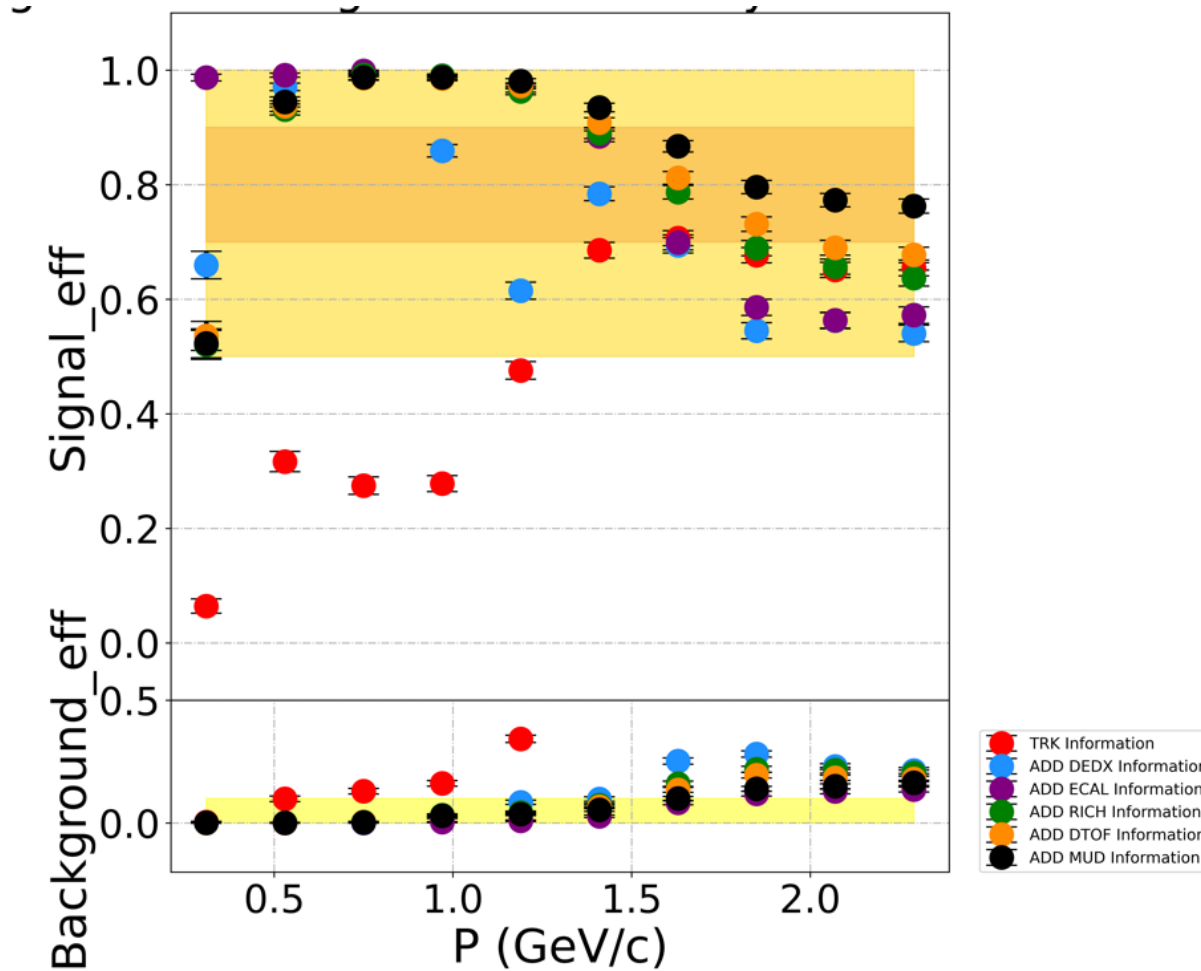
Features

探测器名称	特征量信息	说明
Track	'Charge'	重建粒子的电荷
	'Mom_x', 'Mom_y', 'Mom_z'	粒子在 x, y, z 方向上的动量
	'HelixPar_d0'	螺旋线五参数: 螺旋线上在 $x-y$ 平面内与参考点的距离最小的点 ($p0$) 与参考点的距离
	'HelixPar_φ'	$x-y$ 平面上圆心与参考点的连线方位角
	'HelixPar_cpa'	径迹横动量倒数, 符号与带电径迹的电荷符号相同
	'HelixPar_z0'	$x-y$ 平面上螺旋线上到参考点最近的点的 z 坐标 ($p0$ 的 z 坐标)
	'HelixPar_tan1'	螺旋线倾斜度 ($\frac{dz}{dx}$)
BTOF	'Likelihood_e'	假设该粒子为电子的似然值
	'Likelihood_μ'	假设该粒子为 μ 子的似然值
	'Likelihood_k'	假设该粒子为 K 介子的似然值
	'Likelihood_π'	假设该粒子为 π 介子的似然值
	'Likelihood_p'	假设该粒子为质子的似然值
dEdX	'dEdXsep'	基于五种粒子假设下的 χ^2 值
ECAL	'NumHits'	在 ECAL 里的击中数目
	'Energy'	重建粒子的能量
	'ESeed'	种子的能量
	'E3x3'	3×3 晶体内的能量沉积
	'E5x5'	5×5 晶体内的能量沉积
	'Position_x', 'Position_y', 'Position_z'	簇设的空间坐标
	'SecondMoment'	二阶矩阵
	'LateralMoment'	横向矩阵
	'ZernikeMoment{2,0}', 'ZernikeMoment{4,2}'	Zernike 矩阵 ($2 \times 0, 4 \times 2$)
MUD	'θ'	在极方向上的夹角
	'φ'	在 xy 平面上的夹角
	'HitNum'	在 μ 子探测器里的击中数
	'RPCHitNum'	在电阻板室 (RPC) 中的击中
	'PSHitNum'	在塑料闪烁体探测器上的击中
	'MaxHit'	有最大击中数所在层的击中数
	'MaxHitLayer'	有最多击中数目的层数
	'MudExp'	鉴别为 μ 子的可能性
	'IsMuon'	鉴别为 μ 子的结果 (是为 1, 否为 0)
DTOF	'logL_e'	粒子在电子假设下的似然值
	'Logl_μ'	粒子在 μ 子假设下的似然值
	'Logl_π'	粒子在 π 介子假设下的似然值
	'Logl_k'	粒子在 K 介子假设下的似然值
	'Logl_p'	粒子在质子假设下的似然值



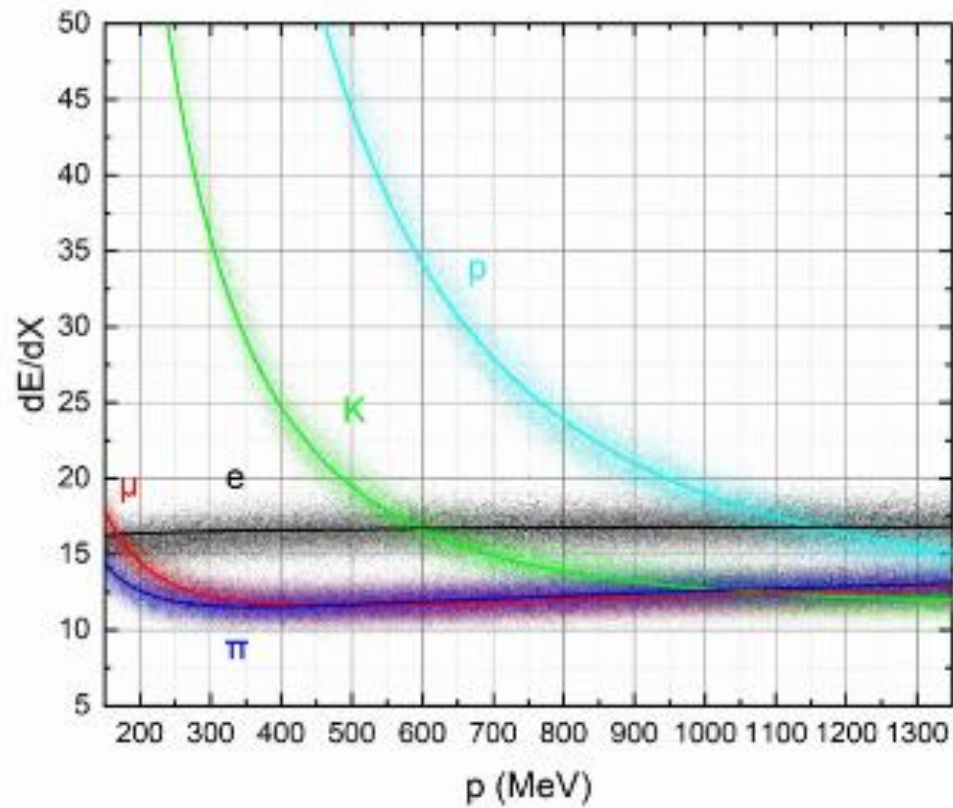


● *The signal efficiency of Proton*

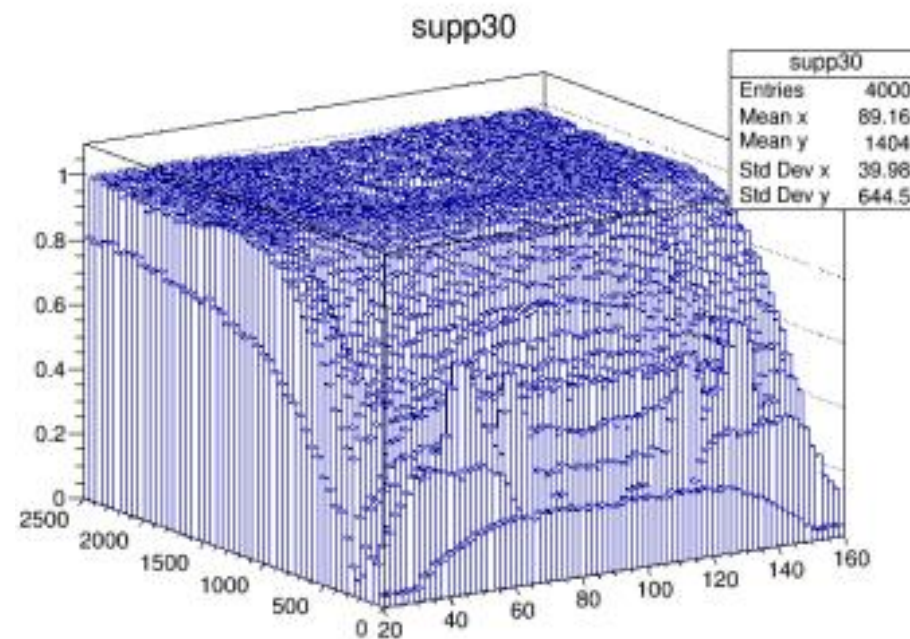




● *DE/dx Sepa.*

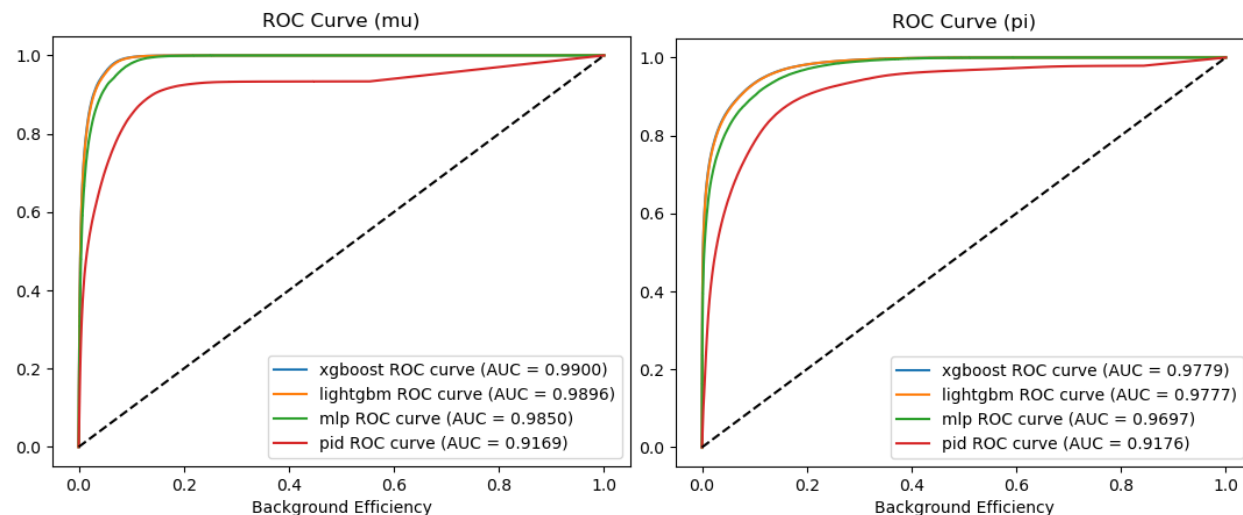


● *MUD Sepa.*





- ❖ Based on the selected features, various models are studied and tested:
 - Boosted decision tree based on XGBoost and LightGBM
 - Deep neural network
 - Support vector machine
- ❖ Model optimization is based on a combination of grid search and bayasian optimization



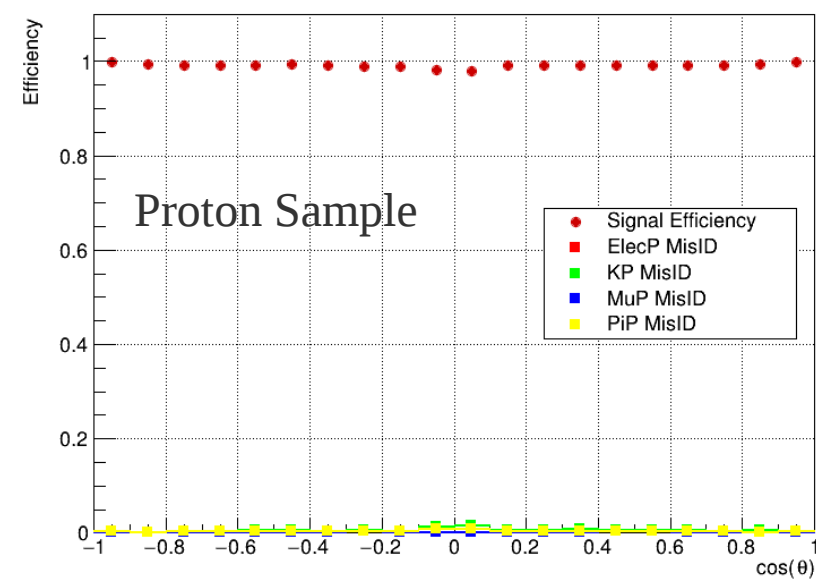
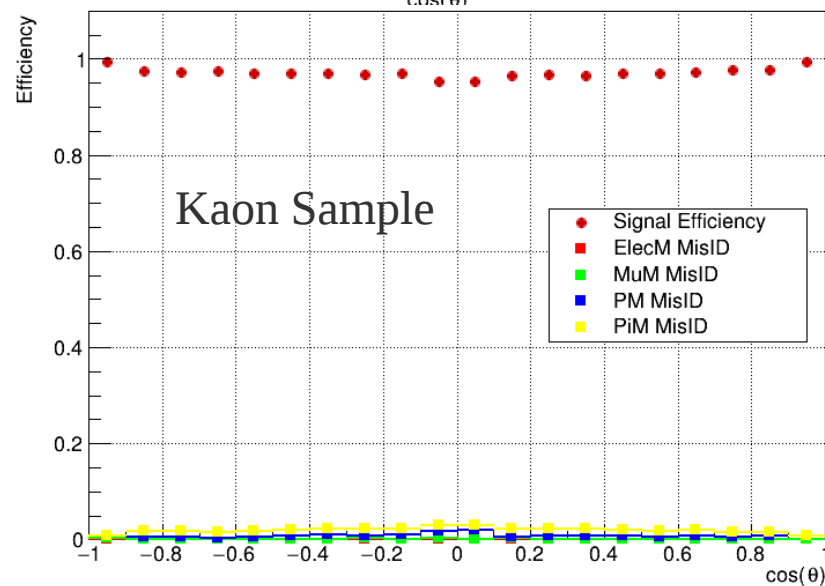
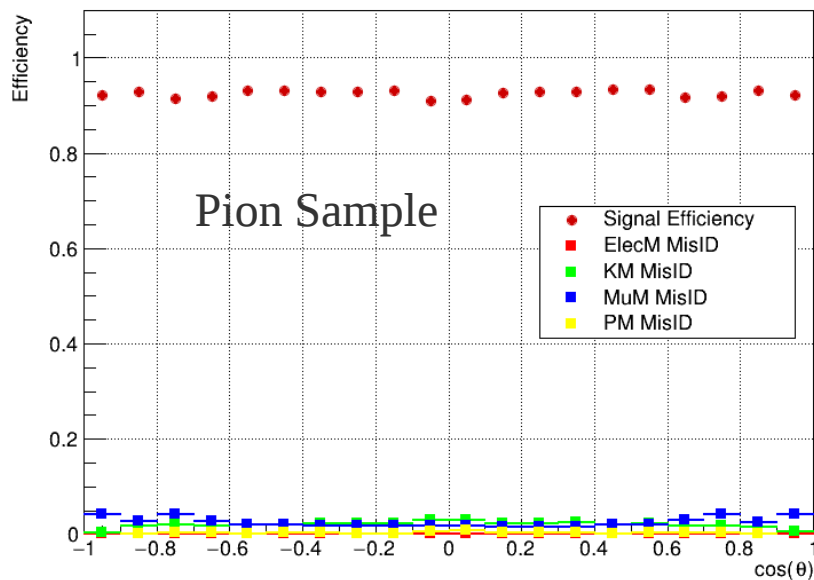
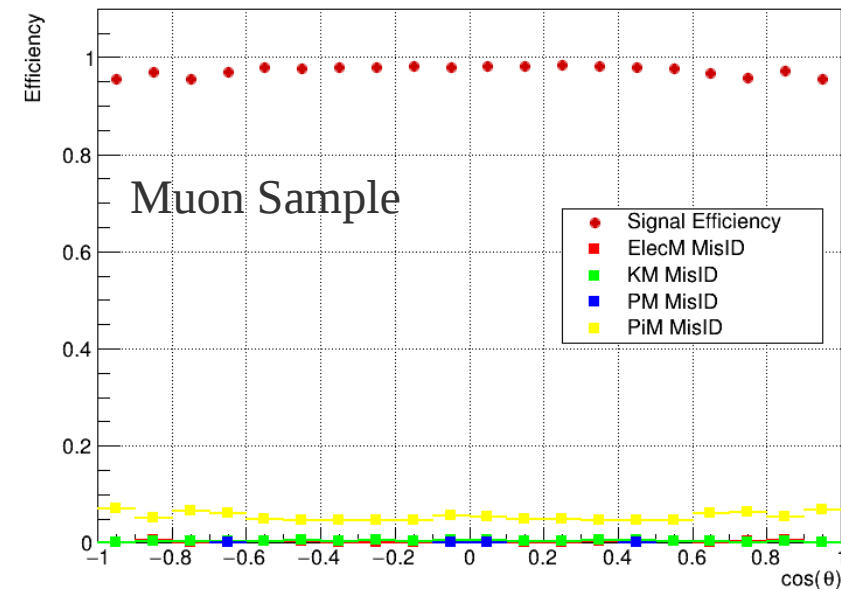
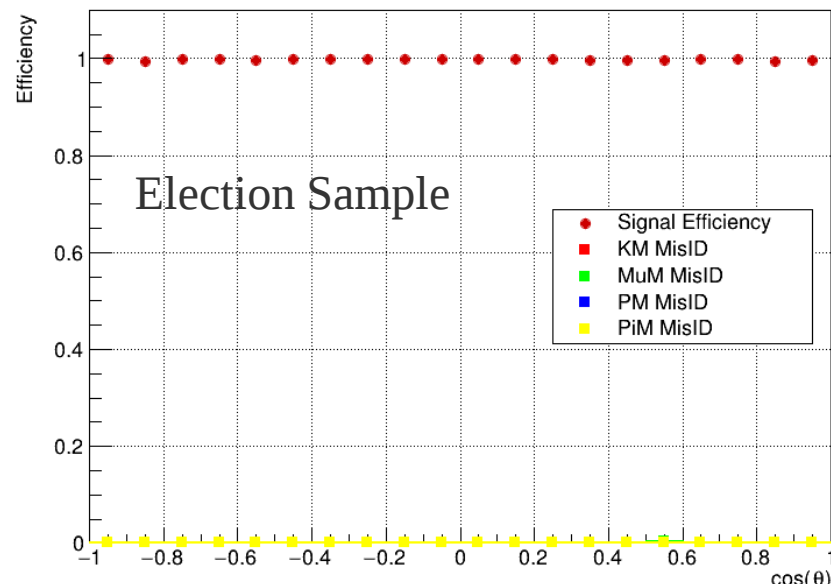
*BDT (XGBoost) is
chosen given its
performance and
tranparency
max depth: 7
n estimators: 400*





Backup 6

- GlobalPID :



GlobalPID软件包

*

MC 样本

- Oscar 版本: 2.5.0
- $\sqrt{s} = 3.097 \text{ GeV}$
- Exclusive MC : $J/\psi \rightarrow \rho \pi^0 \rightarrow \pi^+ \pi^- \gamma \gamma$

*

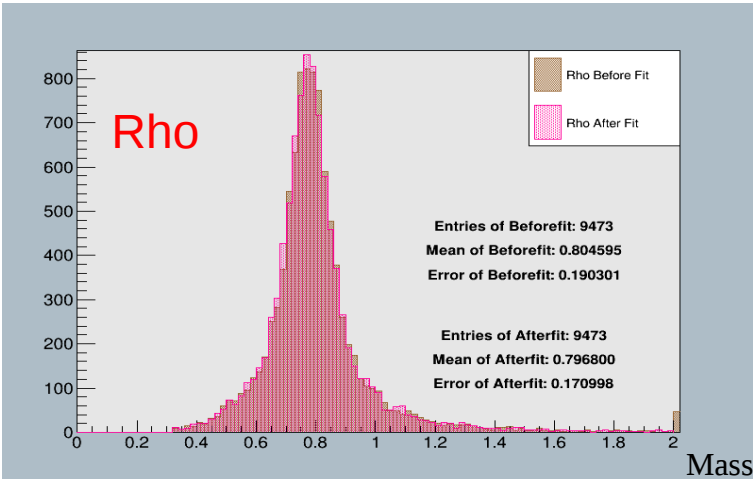
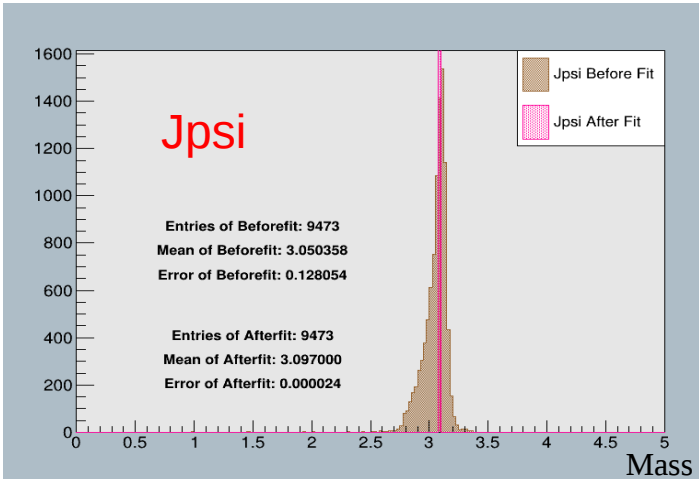
事例选择

- 带电径迹的选择
 - $N_{\text{good charge track}} = 2$
 - Total charge=0
- 中性径迹的判断
 - 桶部光子 $E_\gamma > 25 \text{ MeV}$ ($|\cos\theta| < 0.8$)
 - 端盖光子 $E_\gamma > 50 \text{ MeV}$ ($0.86 < |\cos\theta| < 0.92$)
 - $\Delta\theta(\text{桶部光子与端盖光子}) < 10$
 - $N_{\text{good neutral track}} > 2$
- Global PID
 - 判选条件 $2 \text{ Prob}(\pi) > \text{Prob}(e)$
& $\text{Prob}(\pi) > \text{Prob}(\mu)$ &
 $\text{Prob}(\pi) > \text{Prob}(K) \& \text{Prob}(\pi) > \text{Prob}(P)$
- 顶点拟合&运动学拟合
 - $\chi^2 > 200$

GlobalPID软件包可以
提供良好的鉴别能力

* 效率检查（初步结果）

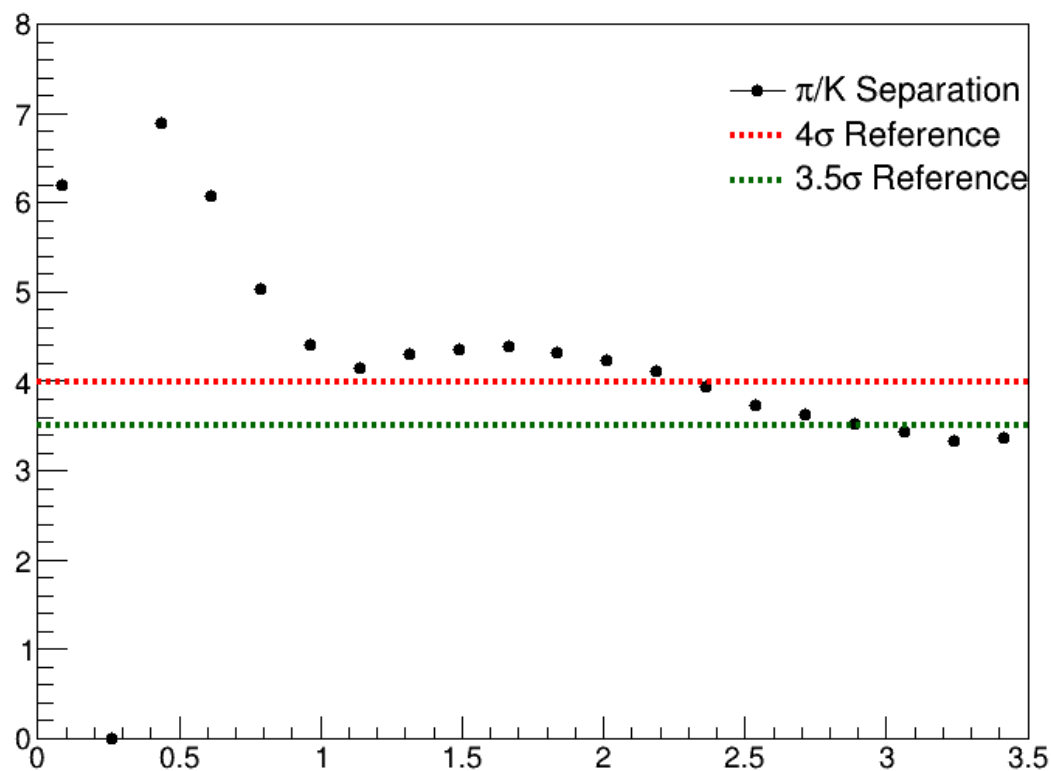
选择条件	事例数目	效率
事例数目	24000	
带电径迹	14217	59.24%
中性径迹	14214	59.23%
带正电的径迹鉴别为pion	12477	87.78%
带负电的径迹鉴别为pion	12411	87.32%
两条径迹都鉴别为pion	10869	76.47%
顶点拟合&运动学拟合	9473	66.65%



性能

* 不同PID模式下的分辨能力: π/K ($\sim 2\% \pi$), μ/π ($\sim 3\% \pi$)

π/K ($\sim 2\% \pi$)



π/μ ($\sim 3\% \pi$)

