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，秦小帅，李腾，黄性涛

zhaiyc@mail.sdu.end.cn

★2024年超级陶粲装置研讨会

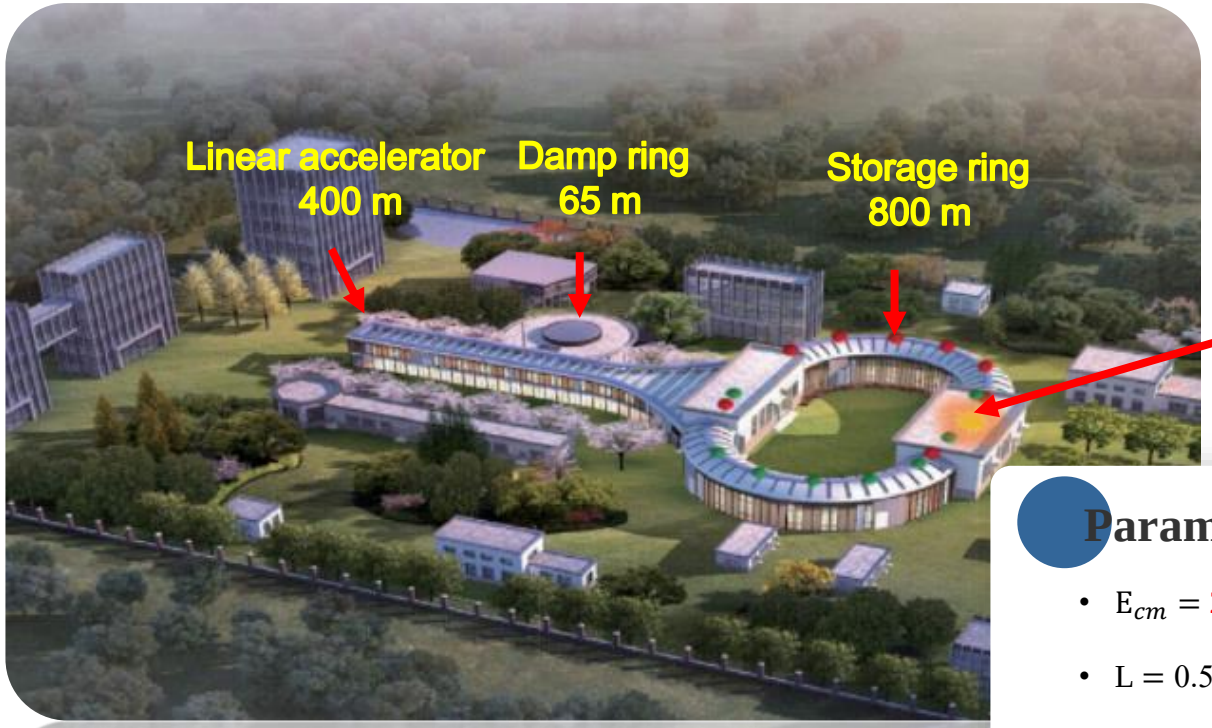
2024. 07. 09

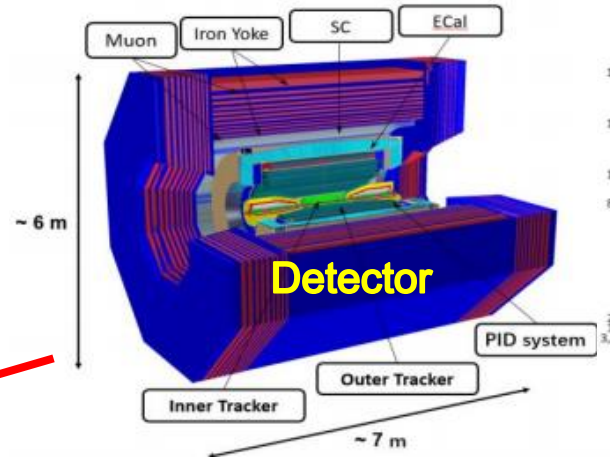
目录

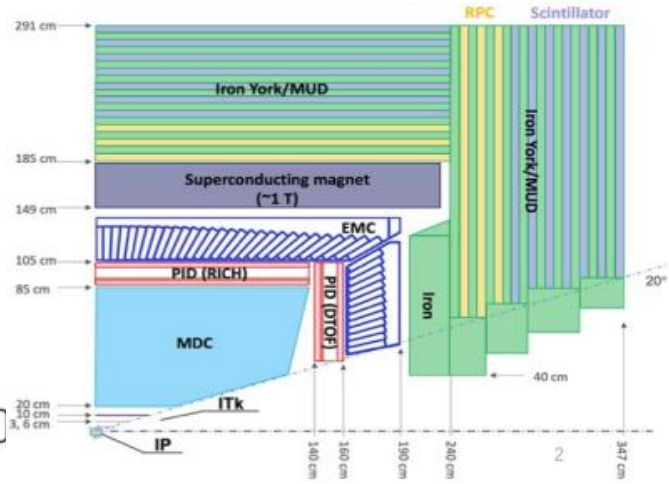
CONTENTS

简介

* 超级Tau-Charm工厂（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。







Schematic layout of the STCF detector concept

Parameters of STCF

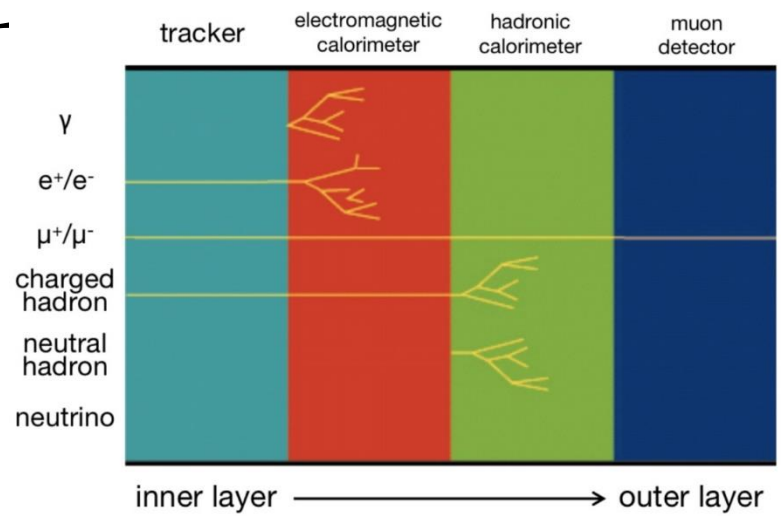
- E_{cm} = 2-7 GeV ,
- L = 0.5×10^{35} cm^{-2} s^{-1}
- Circumference: Double-ring, 600-800 m
- Crossing angle: 2×30mrad

Physics Objectives

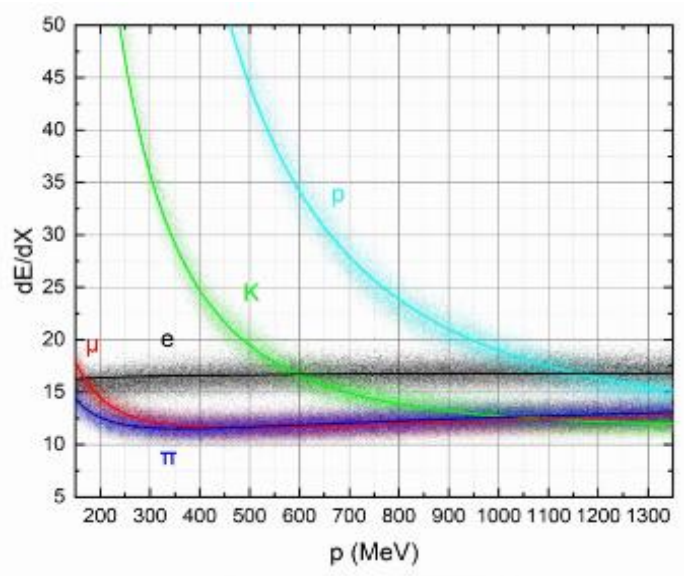
- Rich physics with c quark and τ leptons
- Non-perturbed strong interaction and new exotic hadronic states
- Studying flavor physics and CP violation physics
- Searching for new physics

简介

* **粒子识别(PID)是高能物理实验中最重要和常用的物理分析工具之一**
  - STCF作为中国下一代正负电子对撞机，对PID的准确性和效率提出了高要求，以满足其严格的标准
* **传统的PID方法已成功应用于各个对撞机物理实验中，但其效率较低，难以满足PID的要求**
  - 单个子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
  - 利用子检测器的测量结果来构造PID特征并进行加权处理较为复杂和困难
* **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检测器的信息，对径迹信息进行“智能联合”



探测器响应

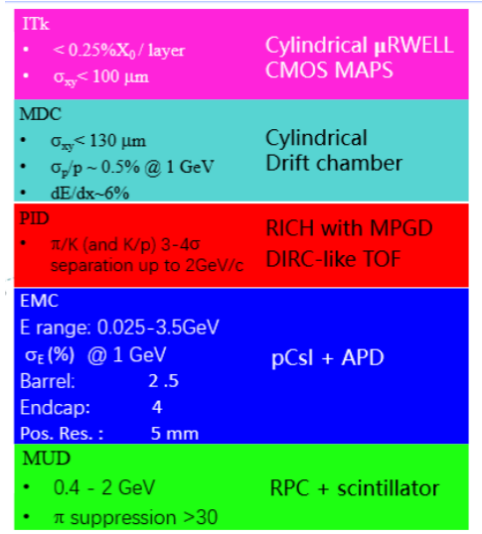


dE/dx 分辨性能.

简介

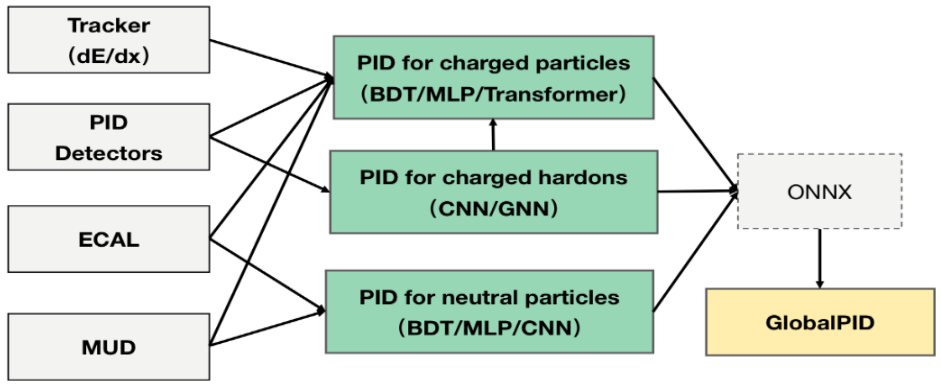
* **针对STCF实验，我们创新并开发基于机器学习的粒子识别算法**

- 带电粒子的GlobalPID算法(e/μ/π/K/p)
  - 结合所有子探测器的重建信息
  - 实现较佳PID性能
  - 尝试其他ML算法：Transformer…
- 中性粒子(γ/ 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



- *π/K (K/p) 3-4σ separation up to 2Gev/c*
- *μ/π up to 2Gev/c, π suppression ~ 3%*
- *Good discrimination power for γ/n/K_L^0*

PID要求



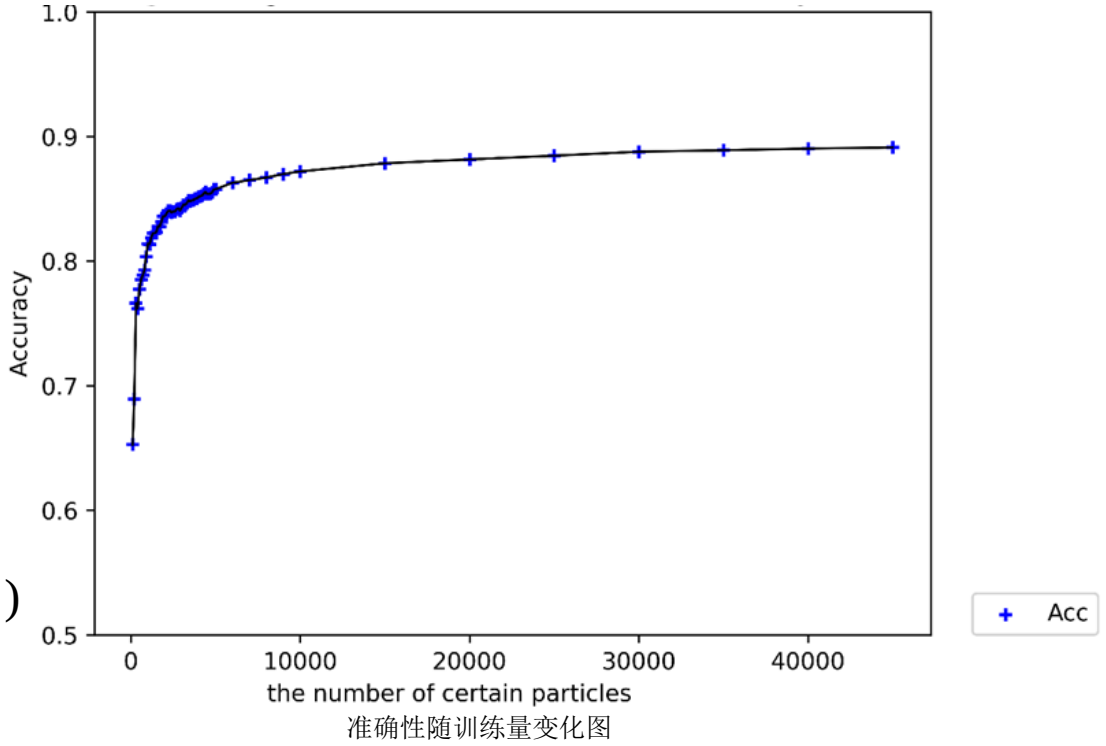
基于机器学习的STCF粒子识别算法

带电粒子的GlobalPID算法

I

数据样本

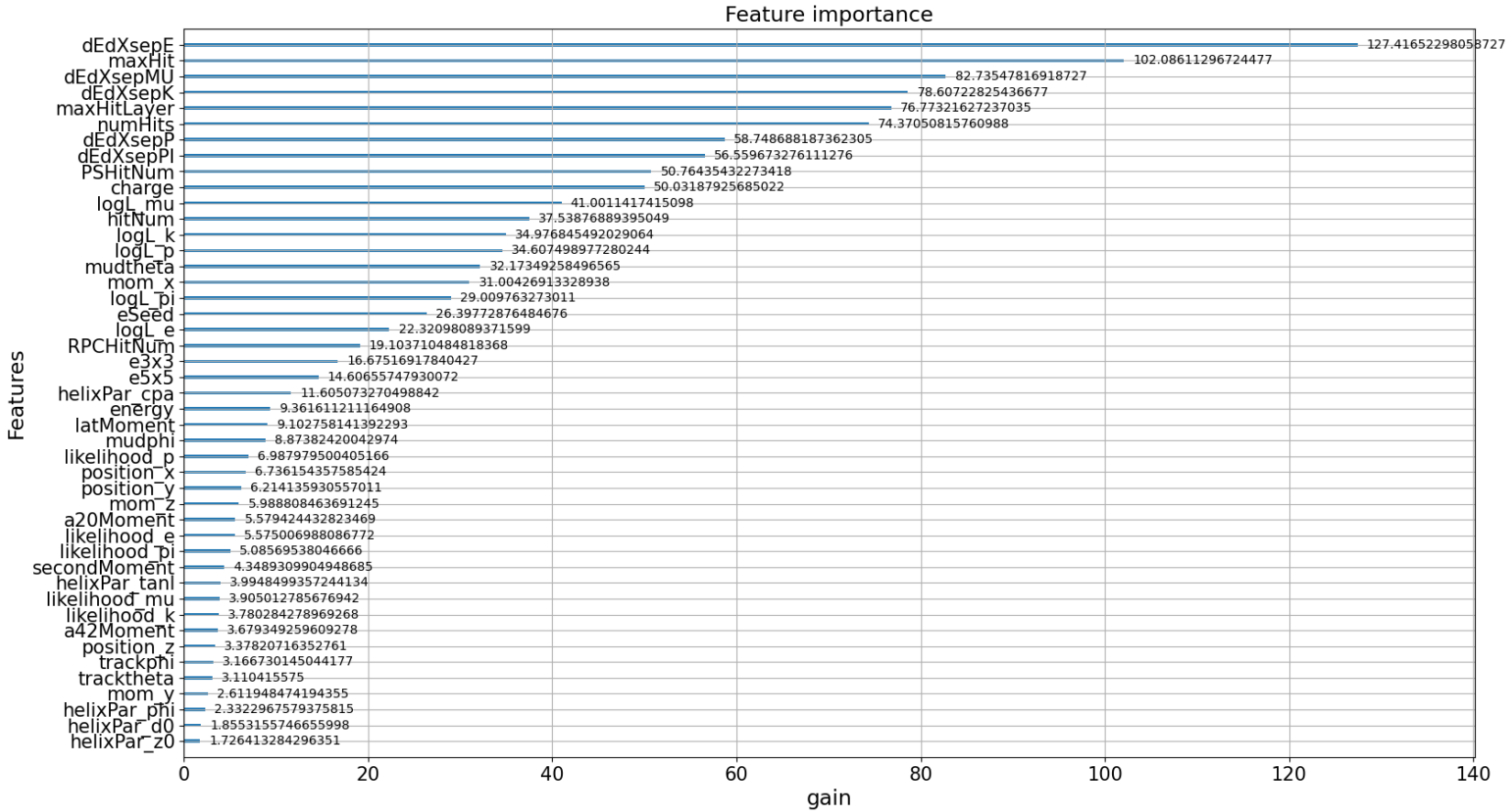
* **数据样本的质量**
  - 高统计量
  - 大动量展宽和角度覆盖范围
* **数据生成**
  - 基于OSCAR 2.5.0模拟和重建
  - 使用ParticleGun生成MC单粒子
  - 统计量：每种粒子类型50000 (e±, μ±, π±, K±, p±)
  - p∈(0.2, 2.4)Gev/c, θ∈(0°, 180°)， phi ∈(0°, 360°)
* **数据预处理**
  - 将动量和θ谱展平，以避免由p/θ分布引起的偏差
  - 训练:验证:测试 = 8:1:1



准确性随训练量变化图

特征选择

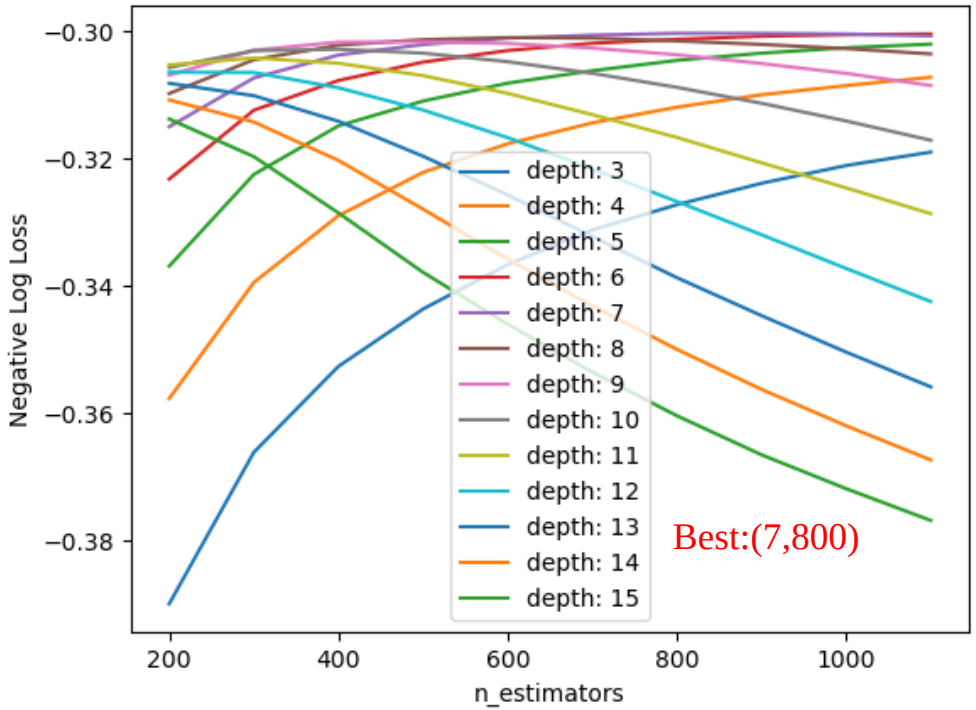
* 从大量相互关联的子探测器信息中选择最具信息量的特征子集可以帮助稳定模型训练过程
  - 已收集了Tracker/dEdx/RICH/DTOF/ECAL/MUD重建变量
  - 保留了45个特征，获得了特征的重要性分布（有关完整变量列表，请参见backup）。



超参数优化

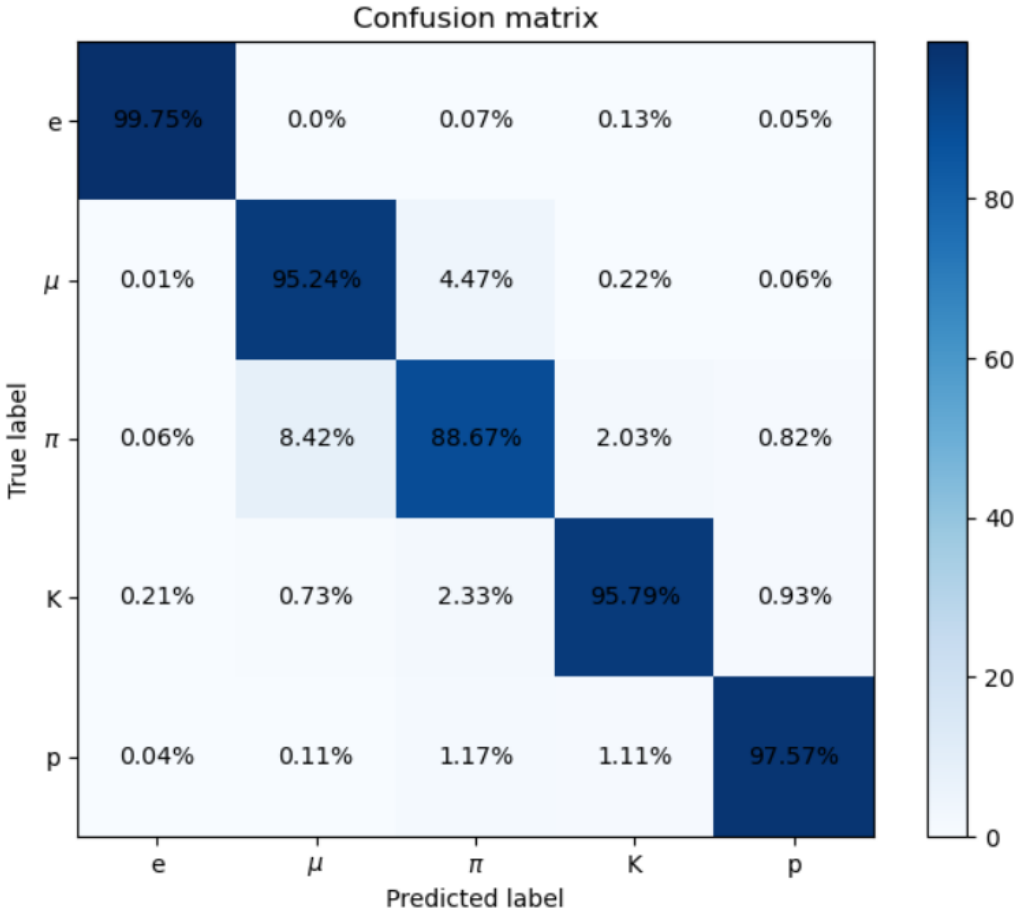
* **目标:BDT超参数的自动优化**
  - 少人工干预和时间成本
  - 提高模型效率和可靠性
* **基于GridSearchCV获取最优超参数**
  - 使用带电粒子间的分辨能力作为判据
  - max_depth(树深度)的搜索范围:[200,1200]
  - n_estimators(分类器个数)的搜索范围:[3,15]
* **最优超参数组合:**
  - max_depth: 7
  - n_estimators: 800

- 超参数调优

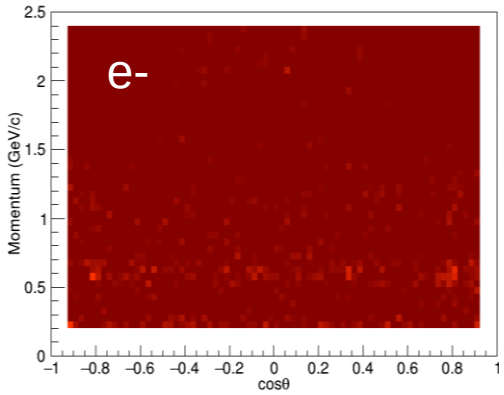


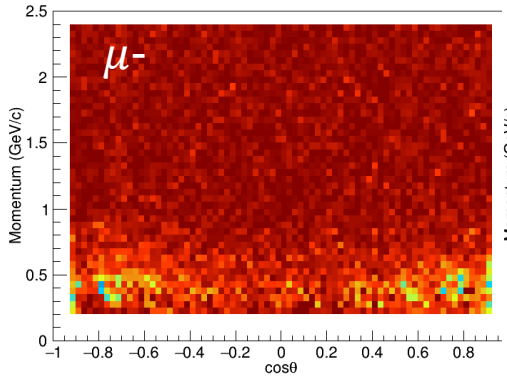
性能

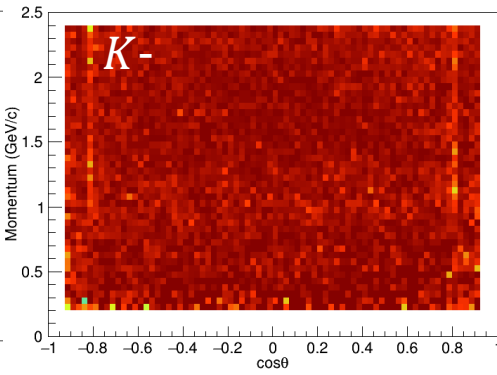
* **在GlobalPID中带电粒子识别性能（初步）**

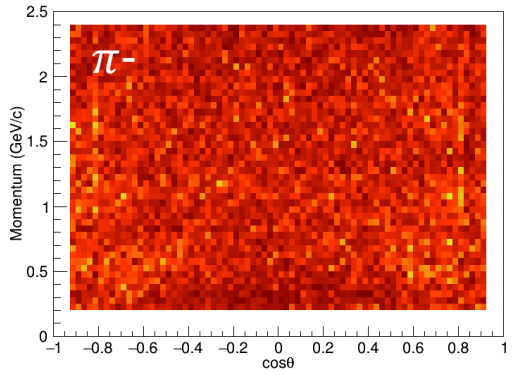


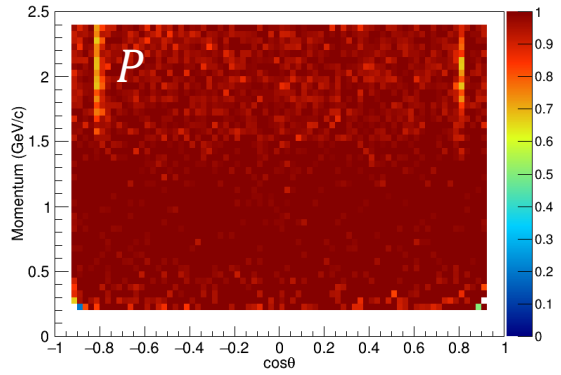
- PID 效率: $\frac{\text{选择正确的信号数}}{\text{信号总数}}$
- 对于轻子具有出色的区分性能, PID 效率 > 90%
- 目前对强子的鉴别性能次优









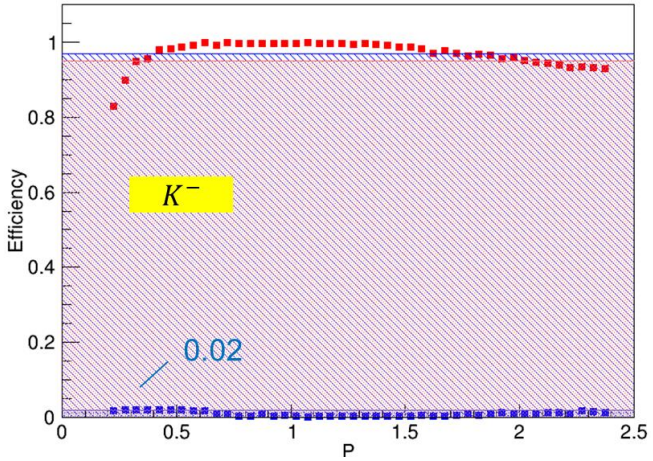


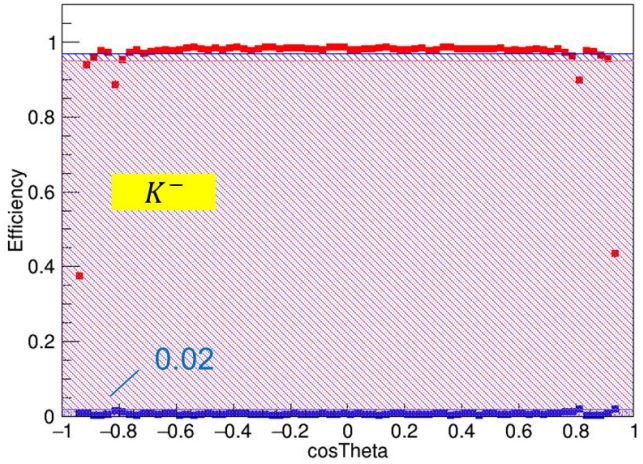
可以进一步优化，利用基于深度学习的PID算法在切伦科夫探测器上。

性能

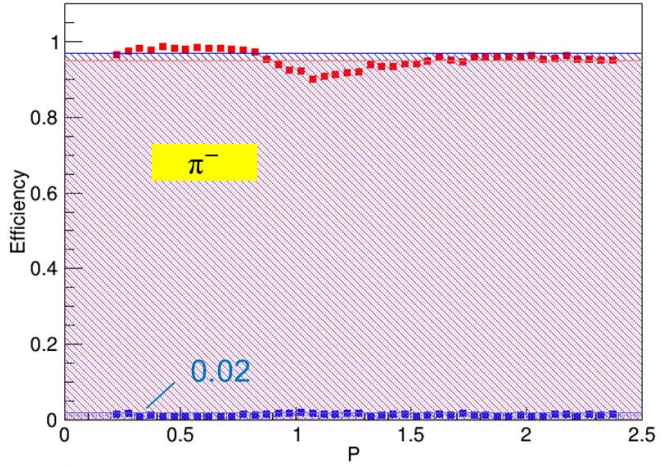
* **不同PID模式下的粒子鉴别性能：K/p,π/K,μ/π**

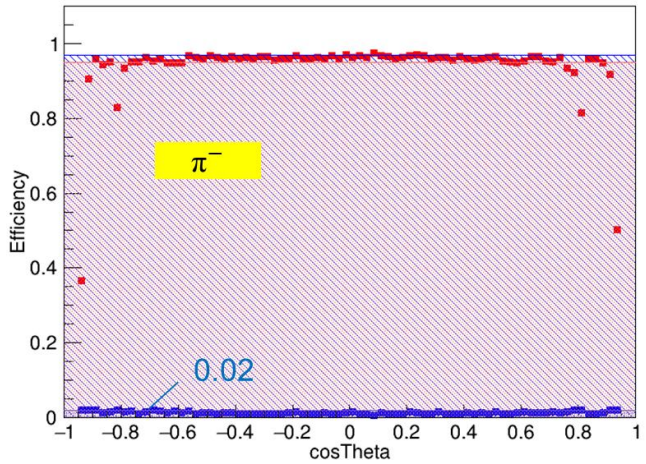
- **K/p (**误鉴别 <2% **):**
  - P < 0.4 Gev/c : PID 效率 > 80%
  - P 直到 2Gev/c： PID 效率 > 95%



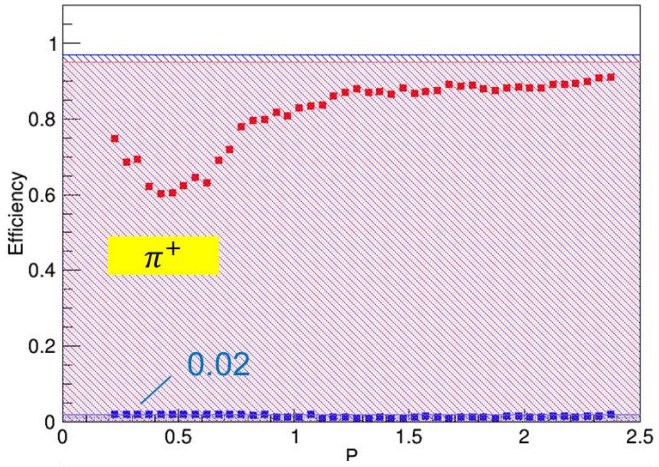


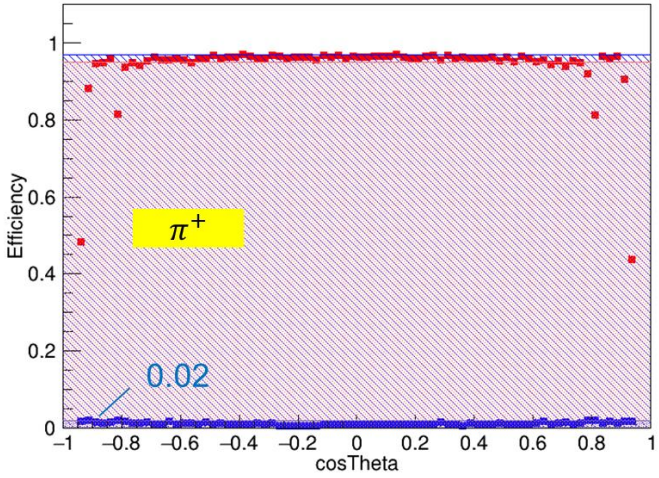
- **K/π (**误鉴别 <2% **):**
  - P < 0.8 Gev/c : PID 效率 ~ 97%
  - 动量大于 1.5Gev/c： PID 效率 ~ 95%





- **μ/π (**误鉴别 <2% **):**
  - PID 效率 : > 60%



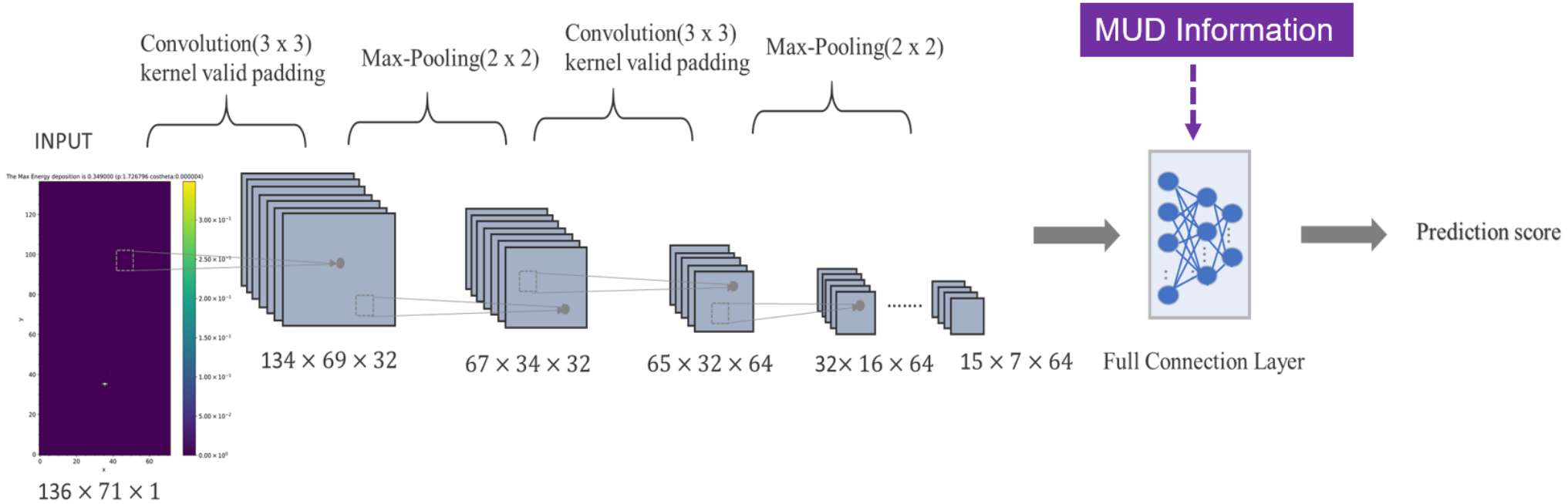


基于CNN的中性粒子鉴别

II

方法

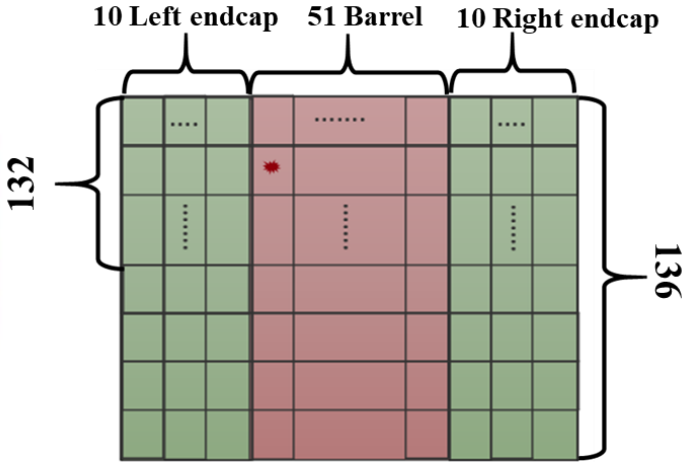
* **中性粒子鉴别通常需要综合考虑ECAL的能量沉积、时间响应以及MUD击中信息**
* **CNN在图像识别、目标检测和其他领域展现出了出色的性能和应用前景**
  - 分层结构：有效地从图像中提取隐藏特征
  - 局部连接性：有助于提取局部特征
  - 数据增强：提高模型的泛化能力。
  - 全连接层：将来添加MUD信息
* **基于CNN的中性粒子鉴别器初步实现**



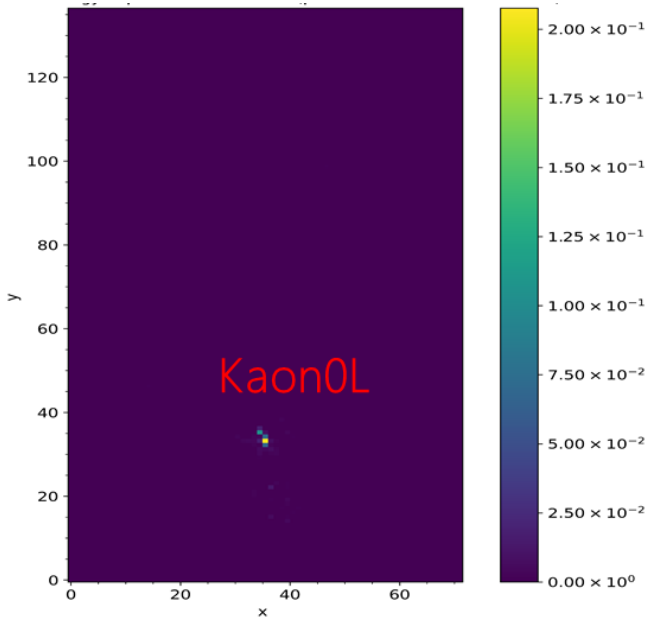
数据样本

* **ECAL设计（全吸收量能器）**
  - 桶部: 51 × 132 = 6732
  - 端盖: 3 × 85 + 102 + 136 × 2 = 1938
  - 晶体大小: 5 × 5 × 28(15 X_0) 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 ∈ （0，2.0）Gev/c,$\theta = 90°, \varphi = 0°$

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（初步）

•中子:

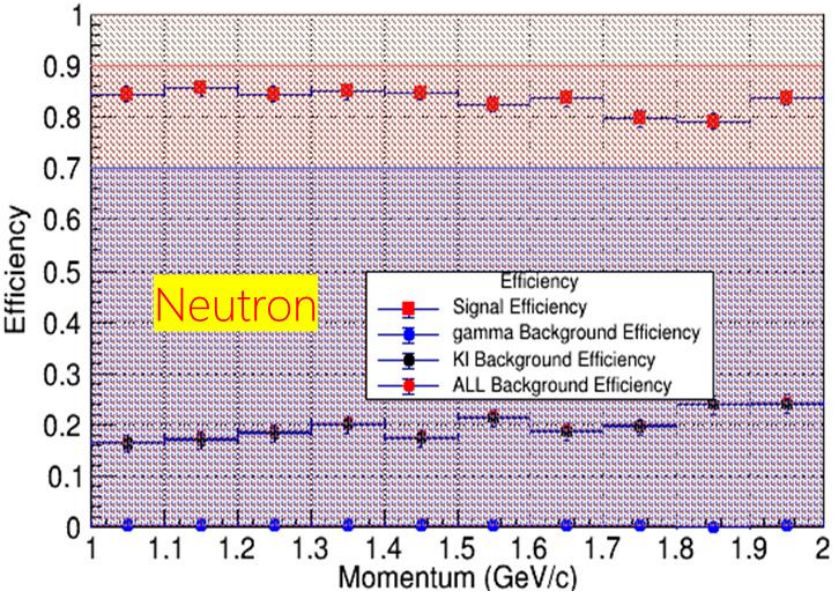
- 信号效率>80%
- 本底误鉴别约为20%，主要是KL

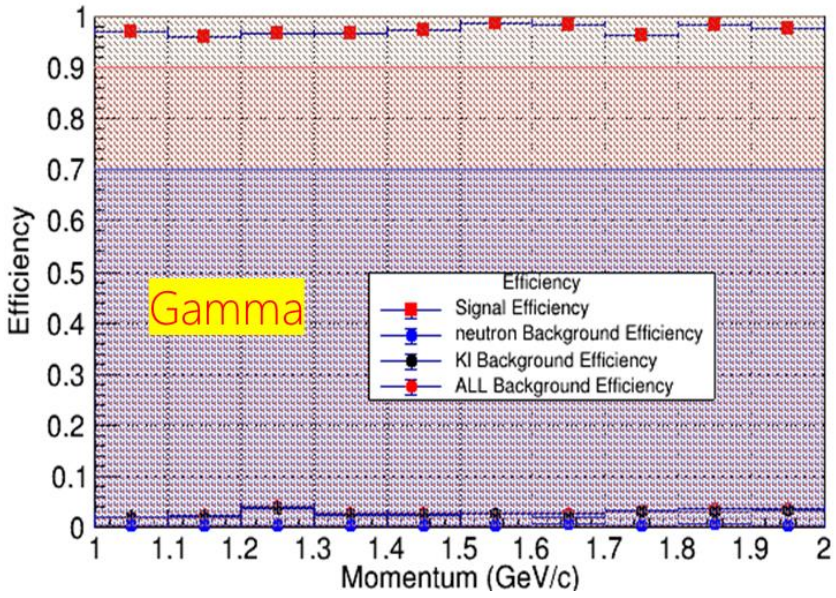
•光子:

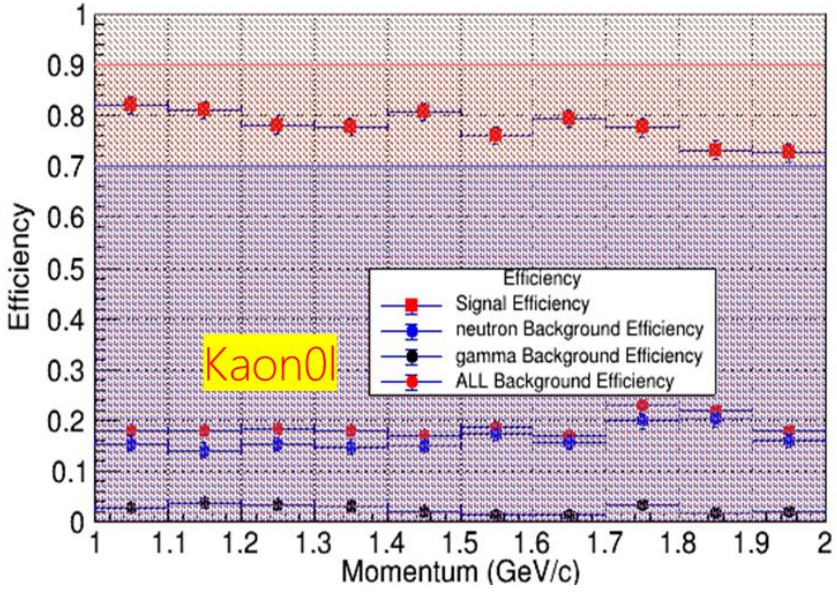
- 良好的光子鉴别性能
- 信号效率> 90%

•K_L:

- 信号效率> 70%
- 本底误鉴别约为20%，主要是中子



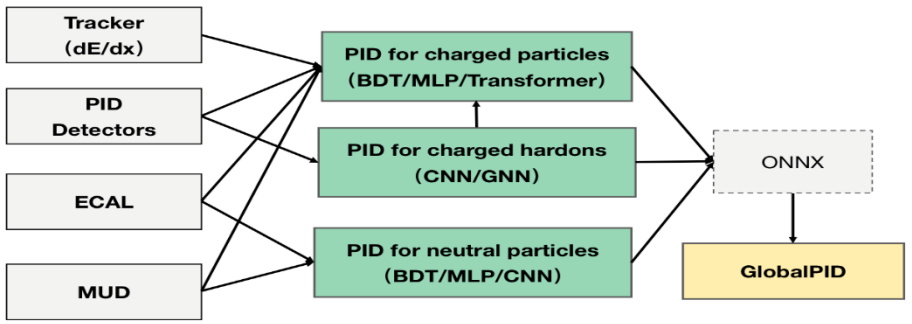




中子和KL鉴别能力仍需改进

GlobalPID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，参考“Get Started with Analysis in OSCAR”（author：周杭，王博，翟云聪，于明玉；周小蓉，艾小聪）
* **基于中子粒子鉴别的机器学习软件包有待开发。。**
* **GlobalPID包集成:所有软件包都将转移到ONNX框架中。。**



基于机器学习的STCF粒子识别算法

Therefore, in the STCF experiment, we have developed a new particle identification software package based on data-driven machine learning methods. The GlobalPIDSvc software package includes pre-trained BDT (based on XGBoost) model and algorithm, and is an important part of the OSCAR software package's Analysis branch.To help analysts become familiar with the software package and its functionalities, the user manual is as follows:

1 Users need to add the directive to include the GlobalPID header file in the source file of the instance selection program.

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

2 Users need to check and retrieve the GlobalPIDSvc instance in the initialize() function of the instance selection program.

```
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();
```

3 To obtain the information of a specific track which needs particle identification as well as the information of each subdetector.

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

4 Users can choose the PID mode, the currently supported PID modes are: All $(e/\mu/\pi/K/p)$, $\pi/K/p$, π/K, $e/\pi/K$, μ/π.

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

5 Users can obtain the predicted probabilities of the trajectory under five particle hypotheses.

```
float m_prob_e = m_pid->prob(Electron);
float m_prob_mu = m_pid->prob(Muon);
```

GlobalPID软件包

* MC 样本

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

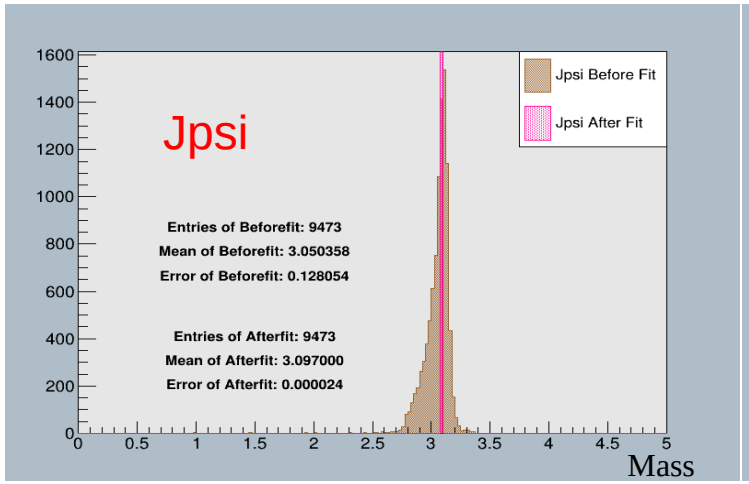
* 事例选择

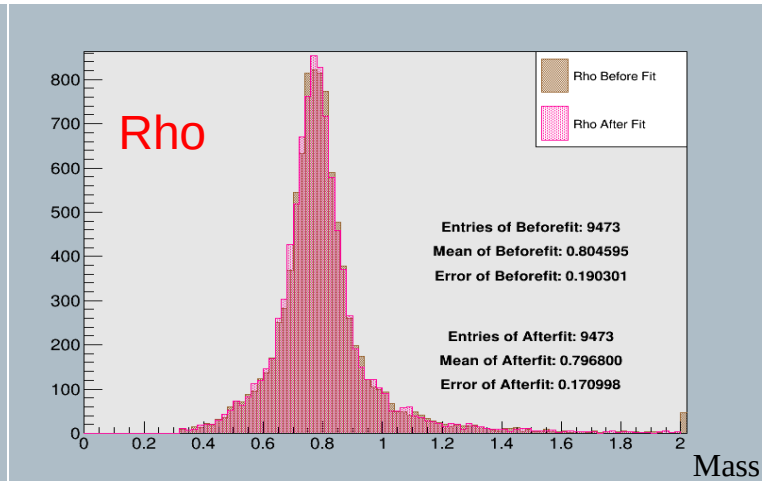
- 带电径迹的选择
  - $N_{\text{good charge track}} = 2$
  - Total charge=0
- 中性径迹的判断
  - 桶部光子Eγ>25MeV (|cosθ|<0.8)
  - 端盖光子Eγ>50MeV (0.86<|cosθ|<0.92)
  - $\Delta\theta$(重建径迹与truth之间theta差值) < 10
  - $N_{\text{good neutral track}} > 2$
- Global PID
  - 判选条件：Prob(π)>Prob(K)&Prob(π)>Prob(P)
- 顶点拟合&运动学拟合
  - $\chi^2 > 200$

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&中性径迹	14214	59.23%
带正电的径迹鉴别为pion	12608	88.70%
带负电的径迹鉴别为pion	12522	88.10%
两条径迹都鉴别为pion	11090	78.02%





总结

* **为了充分利用STCF检测器的性能，提出开发利用深度机器学习算法的PID软件系统**
* **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模式下带电粒子识别性能，驱动快速仿真工作
* **基于CNN的全局中性粒子识别器已初步实现**
* **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($e^+e^- \to \pi\pi$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

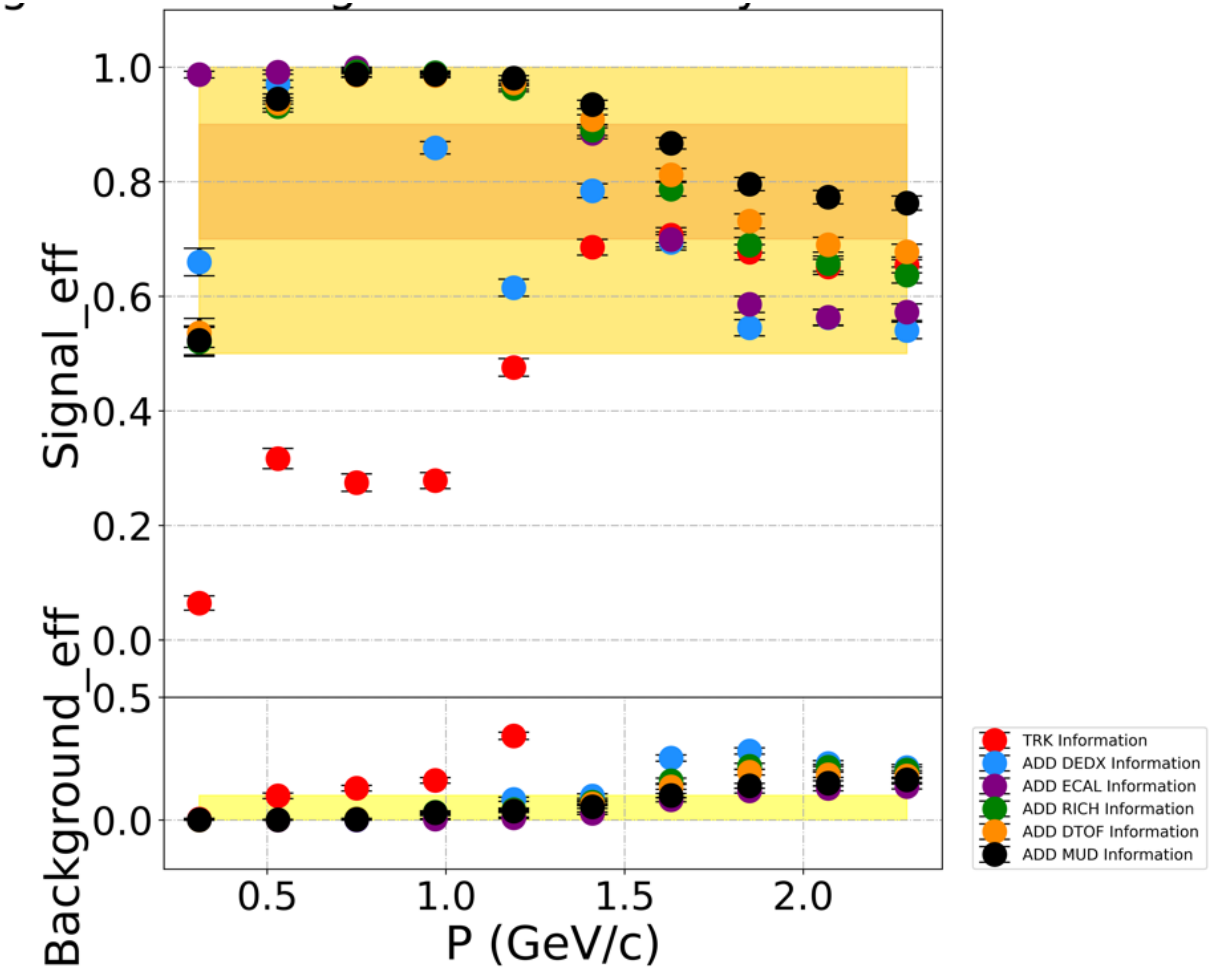
Backup 1

Features

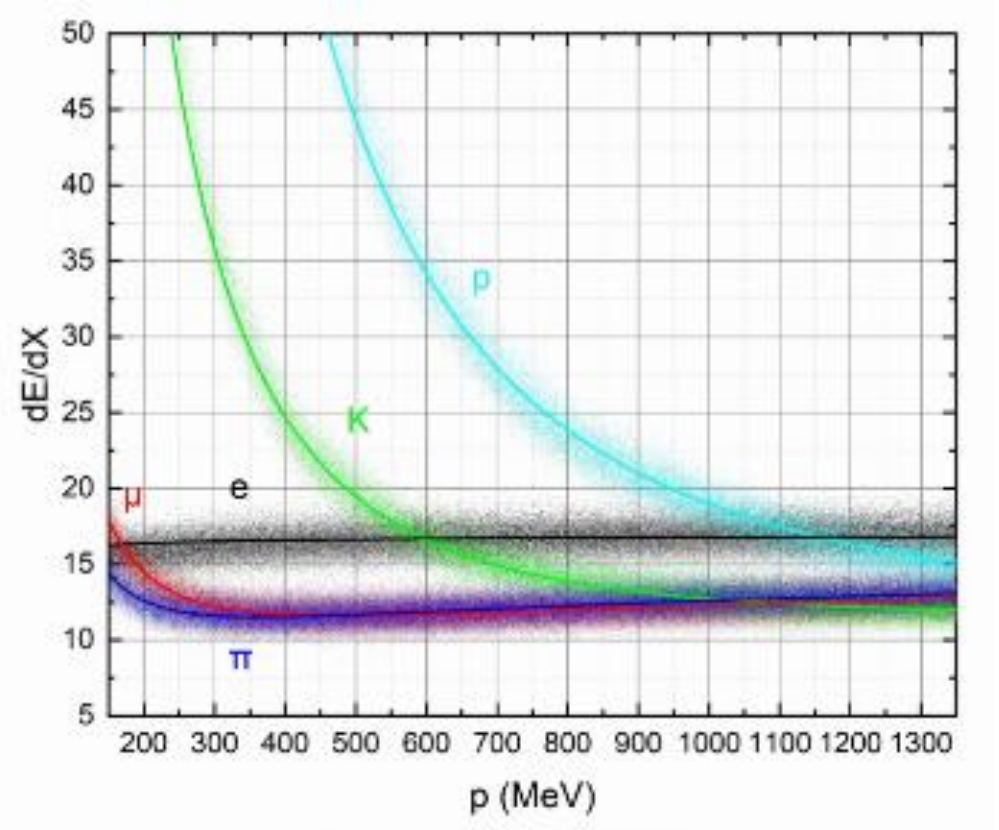
	特征量信息	说明
ReconstructinParticle	‘charge’	重建粒子的电荷
	‘momentum.x’	
	‘momentum.y’	粒子在xyz方向上的动量
	‘momentum.z’	
RecRICHLikelihood	‘likelihood_e’	该粒子假设为电子的可能性
	‘likelihood_mu’	该粒子假设为muon的可能性
	‘likelihood_k’	该粒子假设为kaon的可能性
	‘likelihood_pi’	该粒子假设为kaon的可能性
	‘likelihood_p’	该粒子假设为proton的可能性
DTOFPid	‘logL_e’	粒子分别在五种粒子假设下的可能性
	'logL_mu'	
	'logL_pi'	
	'logL_k'	
	'logL_p'	
TrackerRecTrack	‘helixPar_d0’	
	‘helixPar_phi’	
	‘helixPar_cpa’	
	‘helixPar_z0’	
	‘helixPar_tanl’	

	特征量信息	说明
DEDX	'dEdXsepE/MU/PI/K/P'	基于五种粒子假设下的chi2值
RecECALShower	‘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’	Shower的x坐标
	‘position.y’	Shower的y坐标
	‘position.z’	Shower的z坐标
	‘secondMoment ’	二阶矩阵
	‘LateralMoment ’	横向矩阵
	‘ZernikeMoment{2,0} ‘	Zernike2*0矩阵
	‘ZernikeMoment{4,2}’	Zernike4*2矩阵
MUDTrack	‘theta’	在极方向上的夹角
	‘phi’	在xy平面上的夹角
	‘hitNum’	在u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

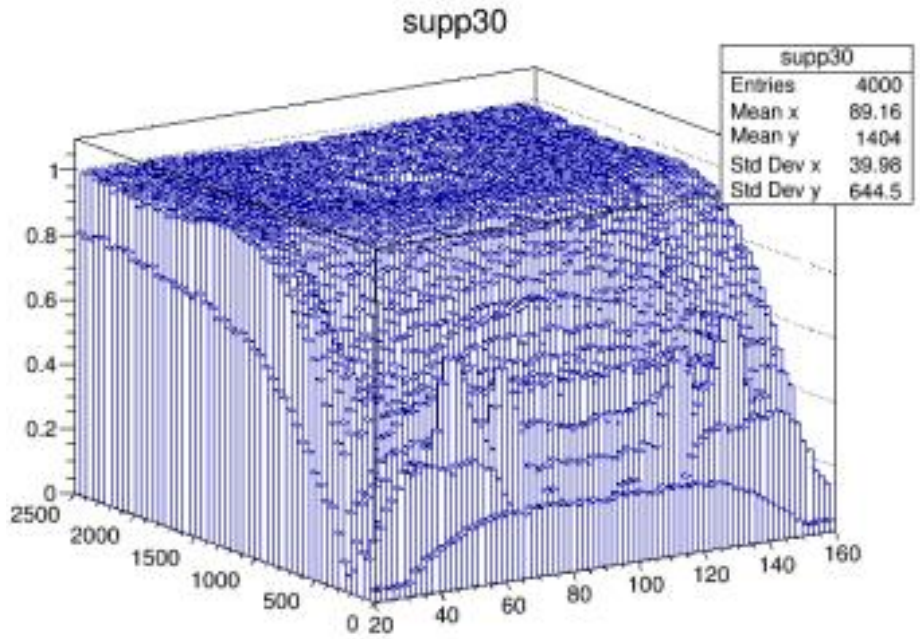
- *The signal efficiency of Proton*



- *DE/dx Sepa.*

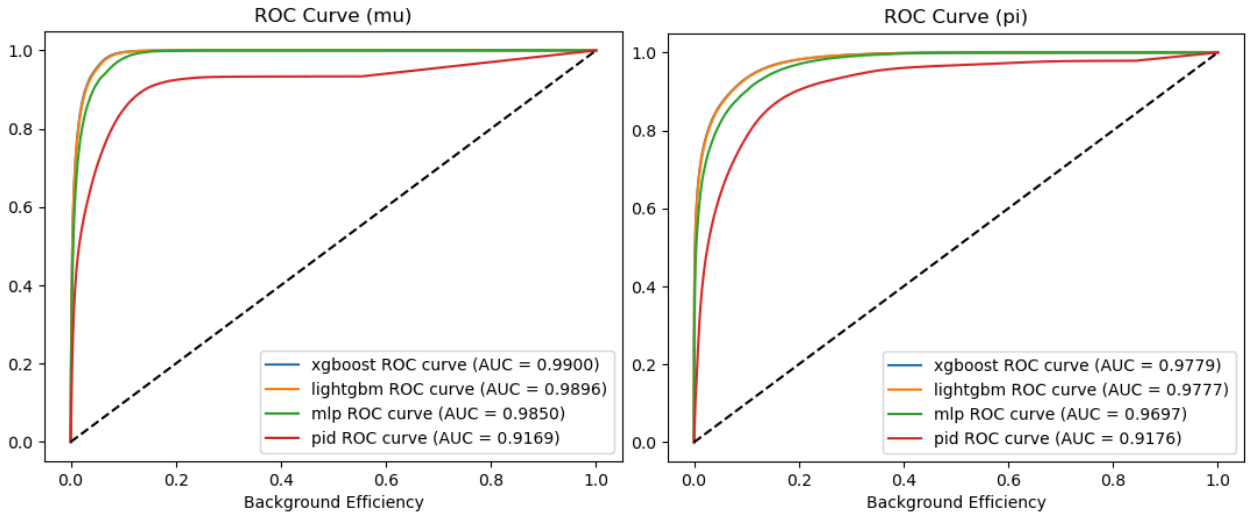


- *MUD Sepa.*



Backup 5

- 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