山东大学（青岛）
SHANDONG UNIVERSITY, QINGDAO

前沿交叉科学青岛研究院
INSTITUTE OF FRONTIER AND INTERDISCIPLINARY SCIENCE

FTCF2024
USTC Hefei
The 2024 International Workshop on Future Tau Charm Facilities
January 14-18, 2024

GlobalPID Algorithms Based on Machine Learning for STCF

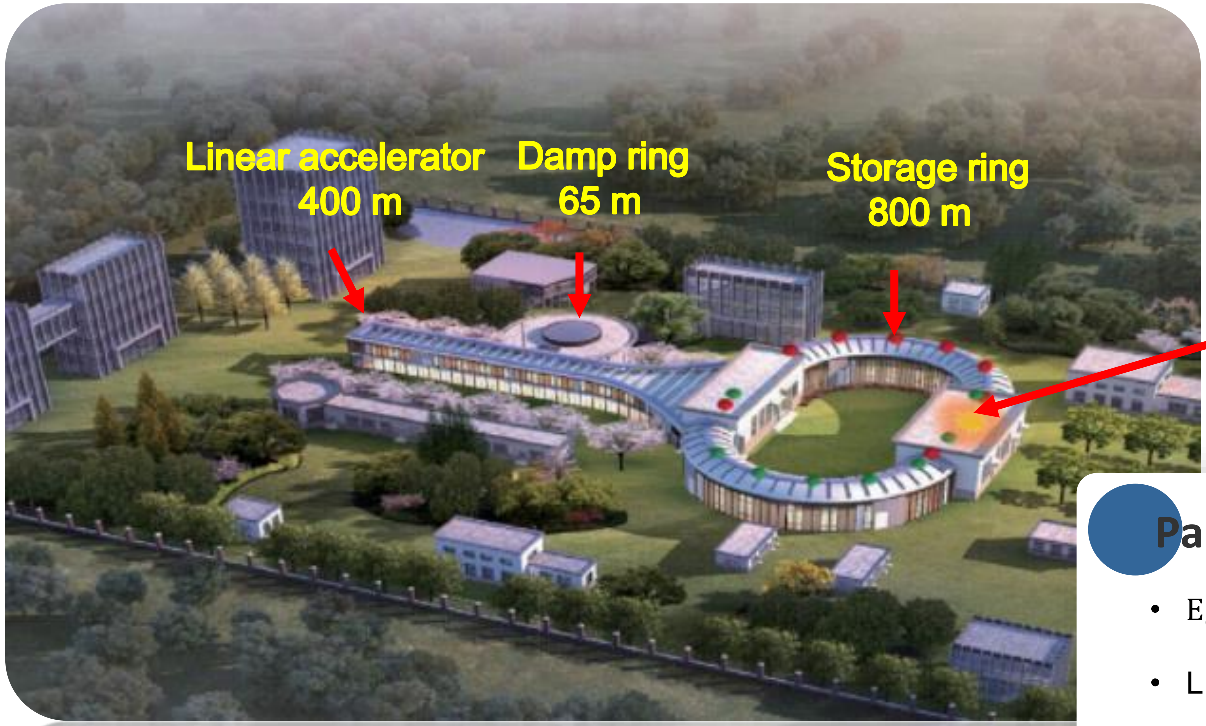
Yuncong Zhai, Zhipeng Yao,Teng li, Xingtao Huang
Shandong University
2024.01.17

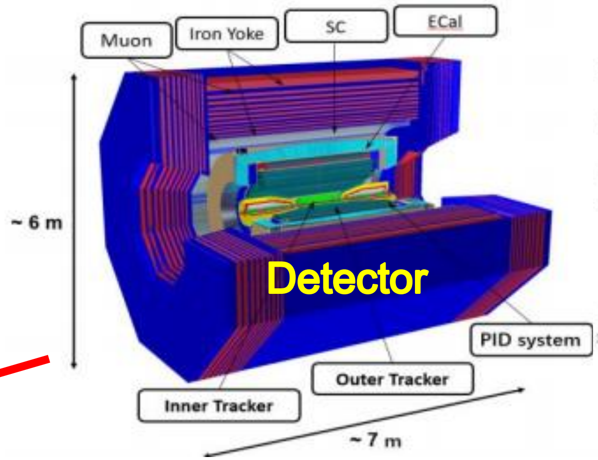
目录

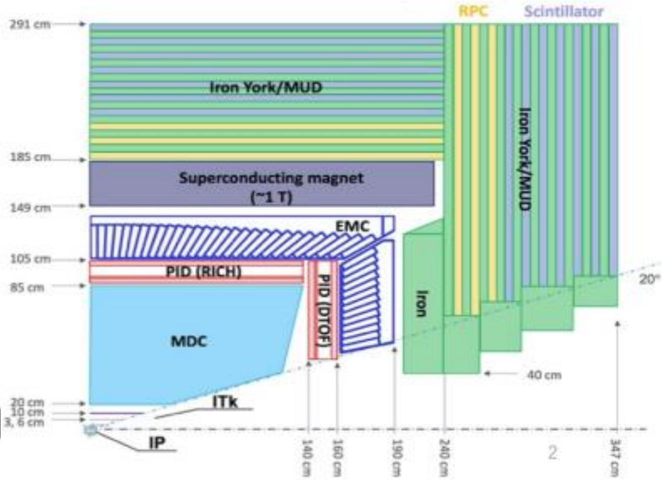
CONTENTS

Introduction

- The Super Tau Charm Facility (STCF) is an important option for China's future accelerator-based particle physics large-scale scientific facility.







Schematic layout of the STCF detector concept

Parameters of STCF

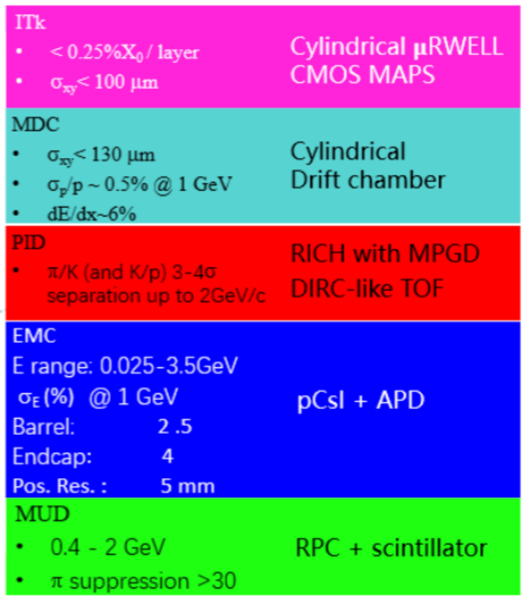
- E_{cm} = 2-7 GeV ,
- L = 0.5×10^{35} cm^{-2} s^{-1}
- Circumference: Double-ring, 600-800 m
- Crossing angle: 2×30mrad

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

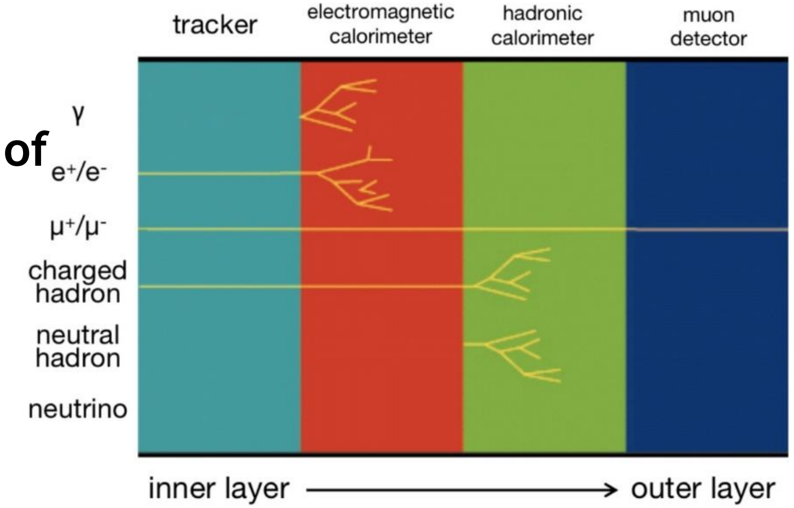
Introduction

- Particle identification (PID) is one of the most important and commonly used tools for the physics analysis in STCF.
- The PID algorithm performance is crutial for exploiting the potential of STCF detectors.



ITk • $< 0.25\% X_0$ / layer • $\sigma_{xy} < 100\ \mu m$	Cylindrical μRWELL CMOS MAPS
MDC • $\sigma_{xy} < 130\ \mu m$ • $\sigma_p/p \sim 0.5\%$ @ 1 GeV • dE/dx~6%	Cylindrical Drift chamber
PID • π/K (and K/p) 3-4σ separation up to 2GeV/c	RICH with MPGD DIRC-like TOF
EMC E range: 0.025-3.5GeV σ_E(%) @ 1 GeV Barrel: 2.5 Endcap: 4 Pos. Res.: 5 mm	pCsI + APD
MUD • 0.4 - 2 GeV • π suppression >30	RPC + scintillator

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



- **Better particle identification usually requires the combination of information from multiple sub-detectors.**
  - Single sub-detector is often sub-optimal
  - Usually difficult for traditional PID algorithms to combine all sub-detectors

Introduction

- **The data-driven machine learning (ML) has provided a powerful toolbox for PID.**
  - Advantage: Extracting effective information from large amounts of interrelate data
  - Widely applied and opening up new possibilities in high-energy physics experiments.
  - Achieved outstanding results in the field of PID.
    - LHCb、BelleII、CMS and ALICE …….
    - Main methods : Boosted Decision Tree (BDT) and Neural Networks (NN)

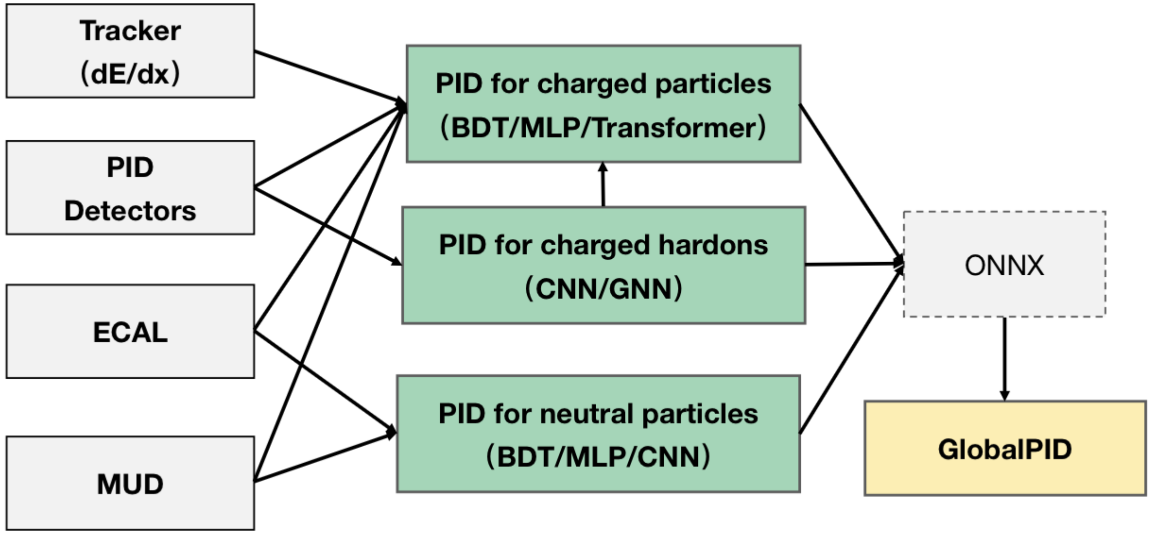
- **Innovated and developed a Global Particle Identification (GlobalPID) software algorithm based on the ML techniques.**
  - Targeting at particle identification problem at the STCF experiment
  - Exploration the physical potential
  - Achieve optimal PID performance
  - To boost the progress of physics analysis work

Introduction

– Identification of charged particles (e/μ/π/K/p)

* Combine all sub-detectors reconstruction information
* Taking BDT (based on XGBoost) as a baseline model
* Other ML algorithms tested as well :

  MLP, SVM,Transformer …..



– Charged hadrons discrimination

* e.g. DTOF raw information: The hit position and time of Cherenkov photons on the sensor
* Based on classical convolutional neural network (CNN) on PID detectors
  - Improve hadron discrimination power
  - As the input for charged particleID

– Neutral particle ($\gamma/K_L^0/n$) identification

* Fully utilize energy deposition, time response within the ECAL and MUD hit pattern
* A convolutional neural network is developed for neutral particle identification

Identification For Charged Particles

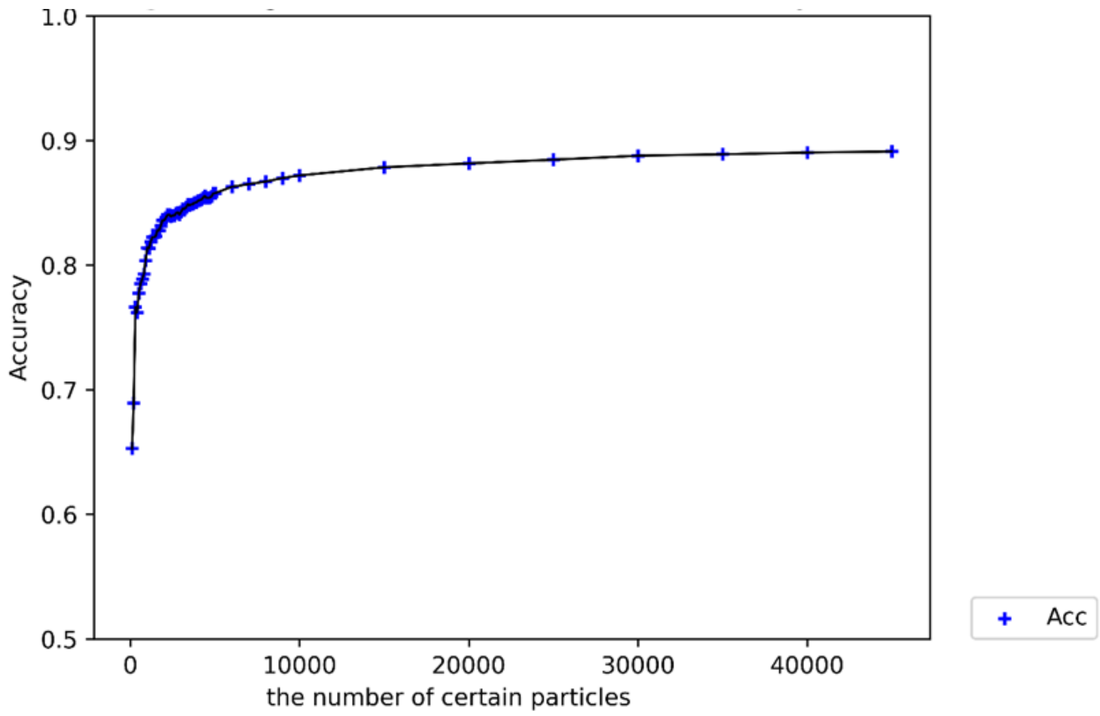
I

SHANDONG UNIVERSITY

Data Sample

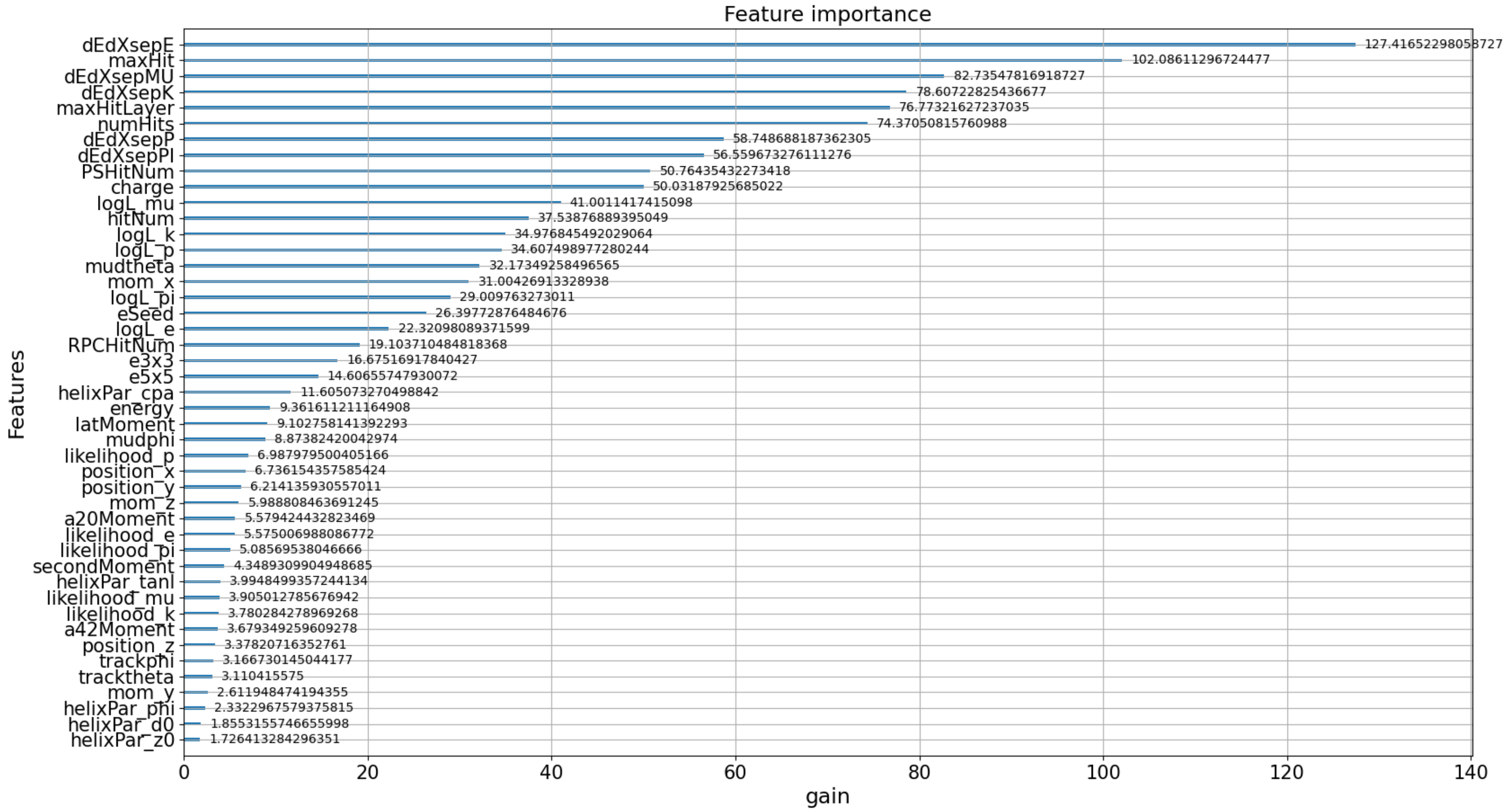
- **The quality of the data samples**
  - High statistics
  - large momentum and angle coverage
- **Data production**
  - Based on OSCAR simulation and reconstruction
  - MC single charged track using ParticleGun
  - 50000 tracks for each type (e±, μ±, π±, K±, p±)
  - p∈(0.2, 2.4)Gev/c, θ∈(20°, 160°)， phi = 0°
- **Pre-processing**
  - Flatten momentum and θ spectrum to avoid bias due to p/θ distribution
  - Train:Validation:Test = 8:1:1

- Accuracy varies with the number of traing tracks



Training and Tuning : Feature Selection

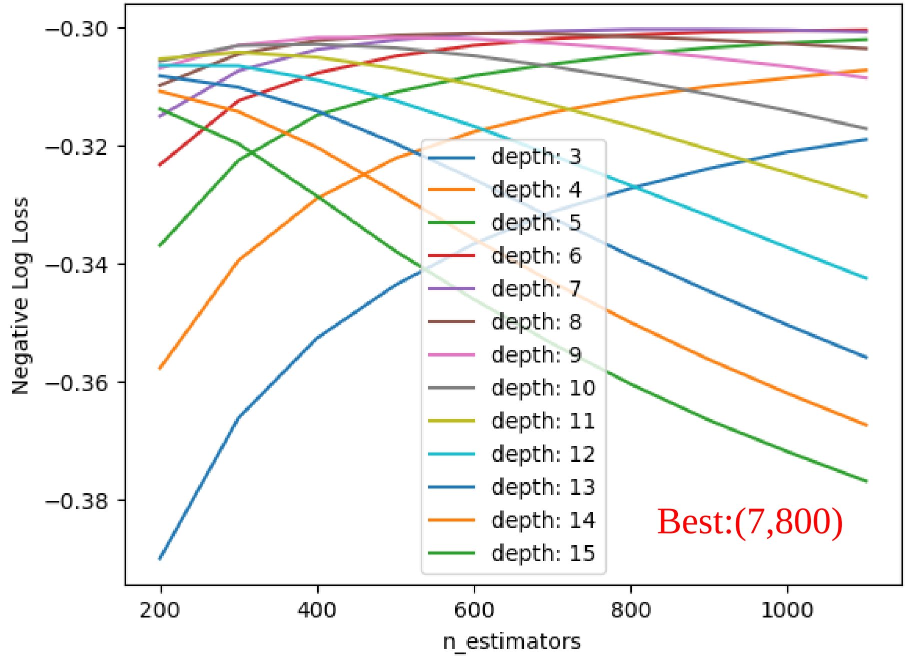
- Selecting a subset of the most informative features from large amount of interrelated sub-detectors information can help stabilize the model training process
- Tracker/dEdx/RICH/DTOF/ECAL/MUD reconstructed variables have been collected
- 45 features are kept, feature importance distribution of the features is obtained (Full list of variables please see backup slides)



Training and Tuning : Optimal Hyperparameters

- **Target: automated optimization of BDT hyperparameters**
  - Reduce manual intervention and time costs
  - Improve model efficiency and reliability
- **Optimal hyperparameters are obtained based on GridSearchCV**
  - Discrimination power between charged particles are used as criteria
  - Search range of max_depth: [200,1200]
  - Search range of n_estimators: [3,15]
- **Selected hyperparameters**
  - max_depth: 7
  - n_estimators: 800

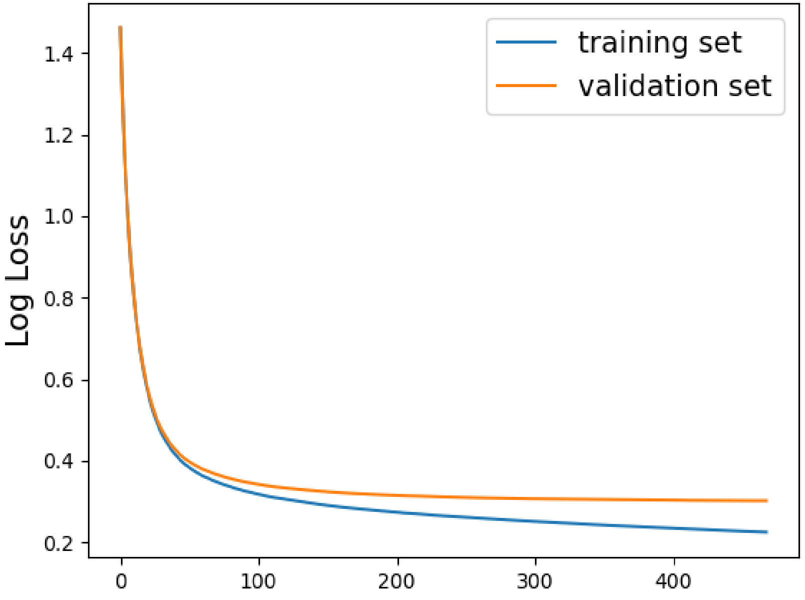
- Tunning of hyper-parameters



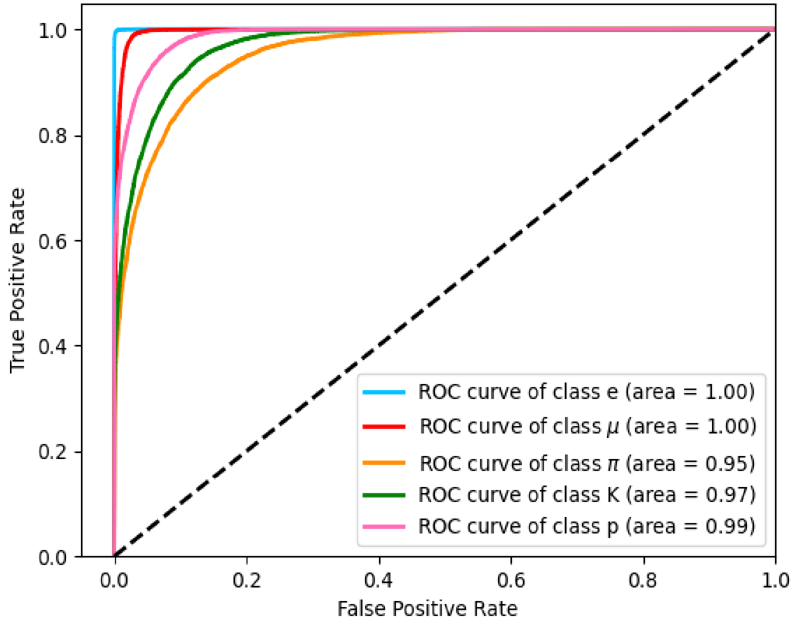
Performance

- BDT model(based on XGBoost) is trained and optimized to discriminate (e, μ, π, k, P)
- Preliminary results have been obtained
  - Good performance for leptons
  - Hadron performance is sub-optimal at the moment. Expecting better performance with updated PID reconstruction algorithms

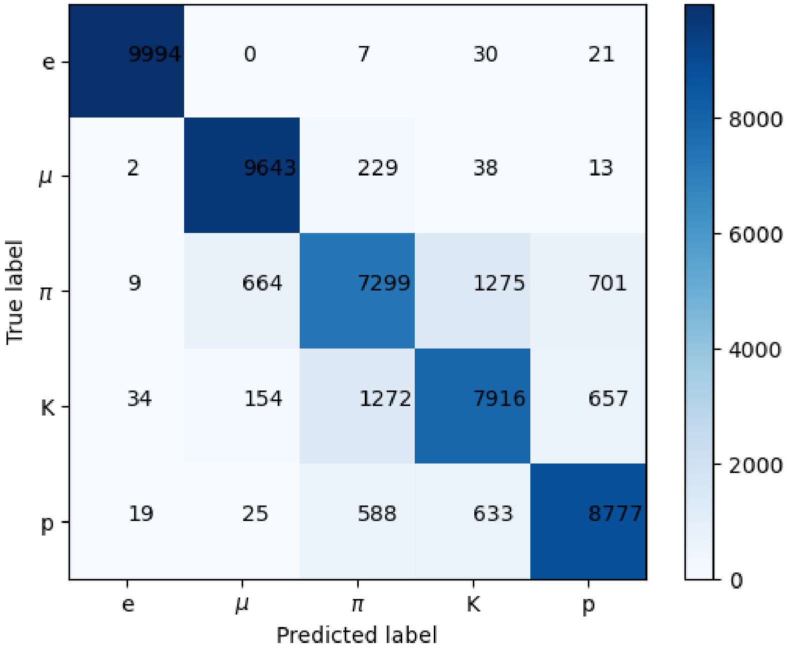
- **Learning curve**



- **The receiver operating characteristic curve (ROC curve)**

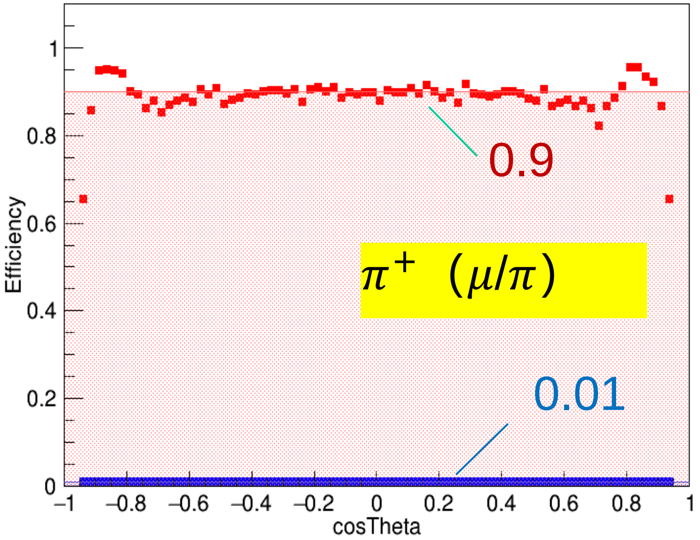


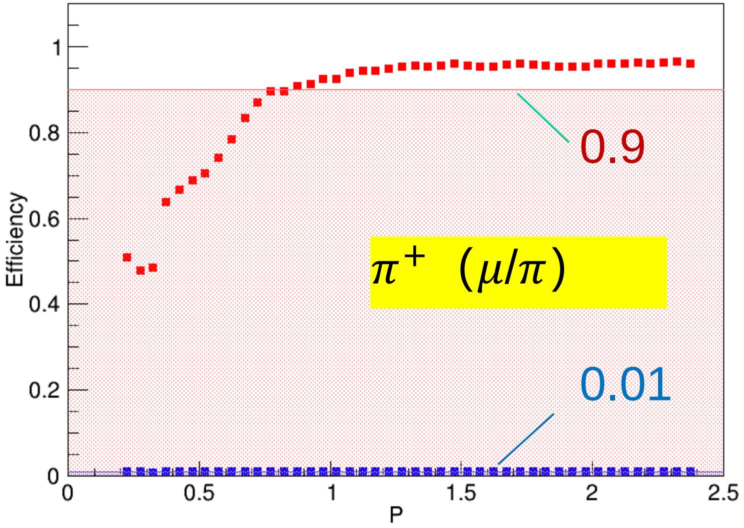
- **Confusion matrix**



Performance

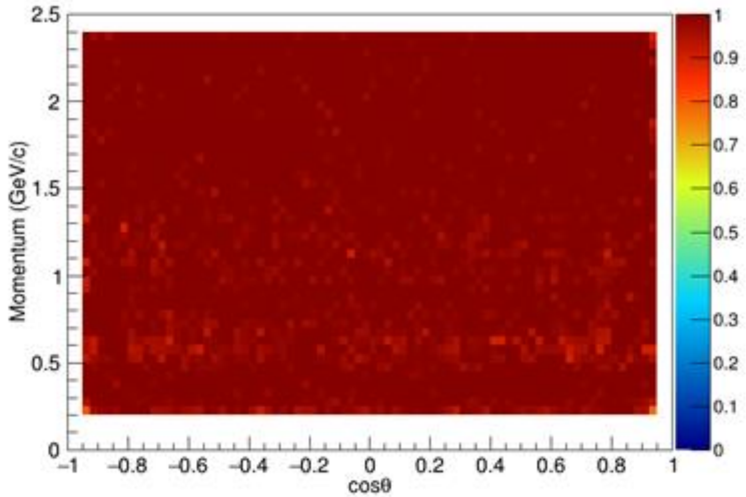
- **Charged Particle Identification Performance In GlobalPID (preliminary)**
  - Signal efficiency : $\frac{The\ number\ of\ \text{signal}\ selected\ correctly}{The\ total\ number\ of\ \text{signal}}$
  - Particle discrimination performance in different PID modes : All(e/μ/π/K/p),π/K/p,π/K,e/π/K,μ/π



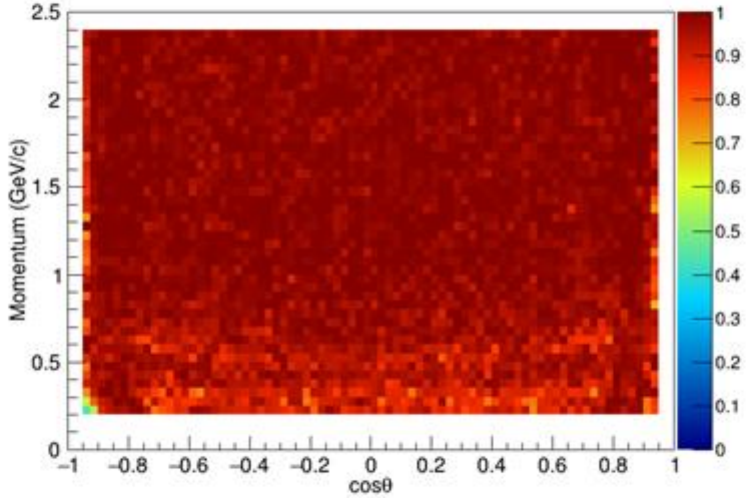


- The signal efficiency and background misidentification rate(no more than 1%) for π at different momentum and angles.

e(Five Particles Identification)



μ(Five Particles Identification)



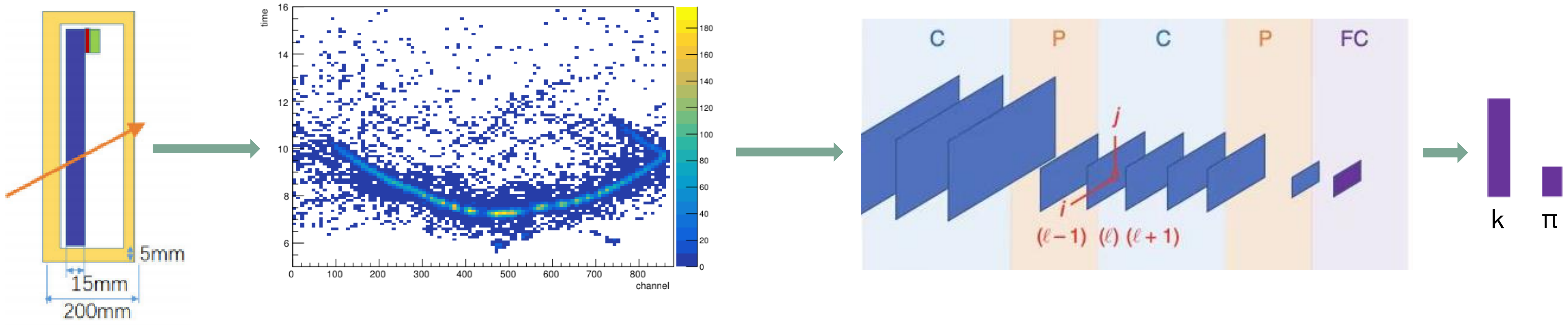
- The signal efficiency for e and μ

STCF DTOF based on classical convolutional neural network π/k discrimination

Zhipeng Yao

The DTOF is located on the end cap of the STCF PID system and is based on an total internal reflection Cherenkov time-of-flight detector.

* Using the hit position and time of Cherenkov photons on the photomultiplier tube, a two-dimensional pixel map is constructed and a convolutional neural network is developed for π/k discrimination, further enhancing the PID performance of the DTOF.



The darker areas in the image indicate a higher probability of Cherenkov photons being detected at the corresponding channel at the given time. The overall image represents the topological structure of Cherenkov photons produced by different particles.

STCF DTOF based on classical convolutional neural network π/k discrimination

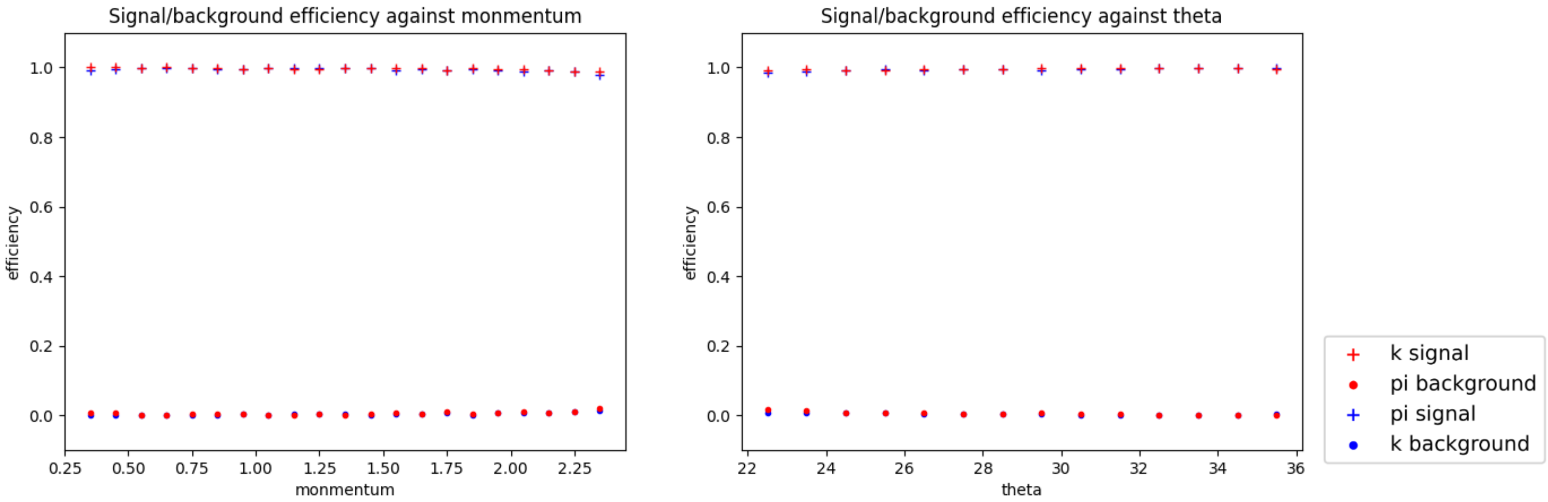
Zhipeng Yao

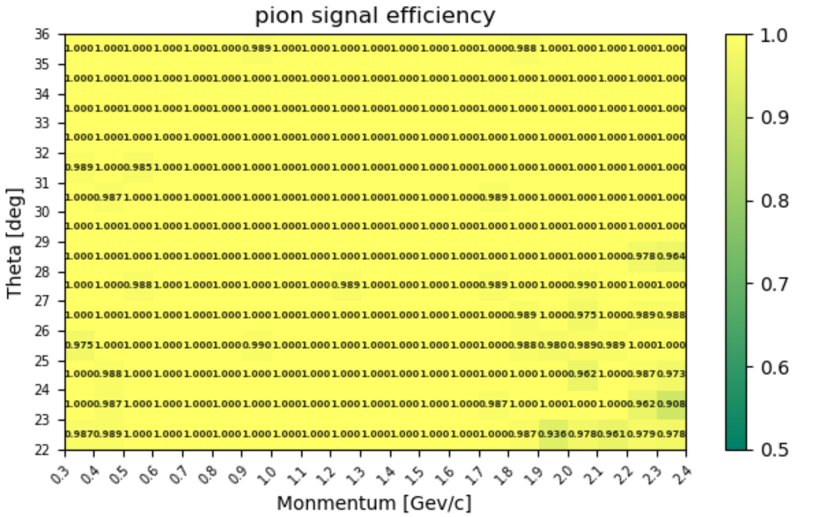
- model： **EfficientNetV2-S** Accuracy = 99.46%

Stage	Operator	Stride	#Channels	#Layers
0	Conv3x3	2	24	1
1	Fused-MBConv1, k3x3	1	24	2
2	Fused-MBConv4, k3x3	2	48	4
3	Fused-MBConv4, k3x3	2	64	4
4	MBConv4, k3x3, SE0.25	2	128	6
5	MBConv6, k3x3, SE0.25	1	160	9
6	MBConv6, k3x3, SE0.25	2	256	15
7	Conv1x1 & Pooling & FC	-	1280	1

- Using EfficientNetV2-S as the baseline model and optimizing
- Training:Adding momentum and position information of particle hits in the DTOF outside of the fully connected layers

- The signal efficiency and background misidentification rate for pions/kaons at different momentum and angles.





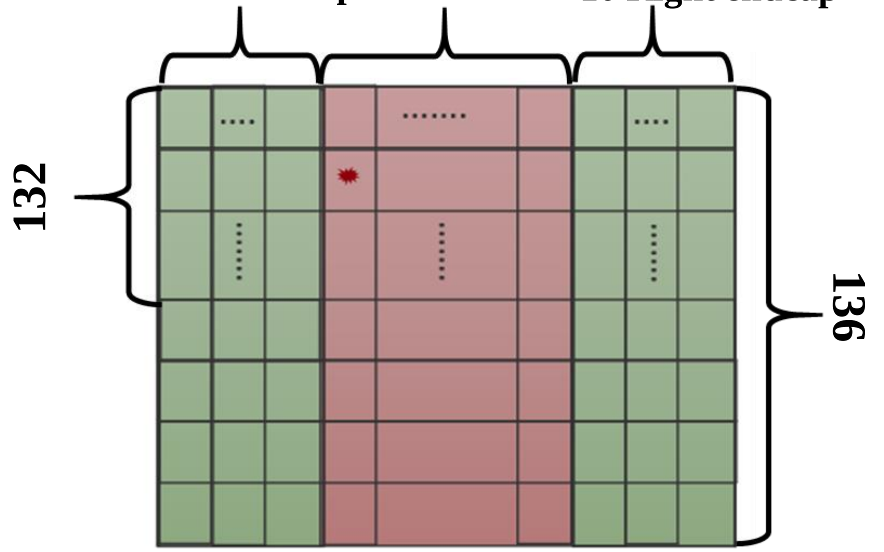
- The signal efficiency for pions (the background no more than 3%)

Identification For Neutral Particles

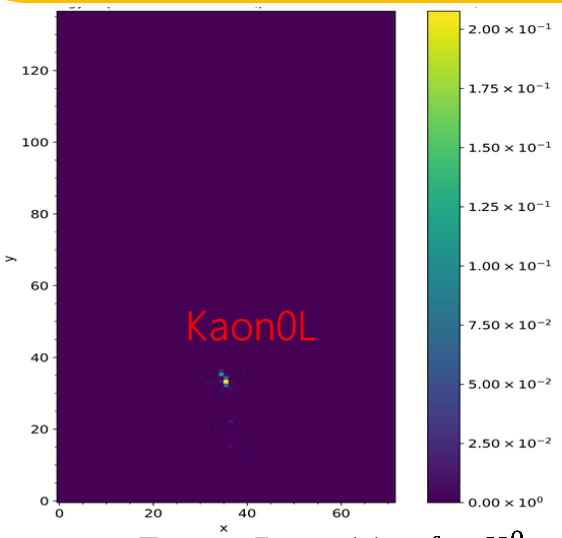
II

Data Sample

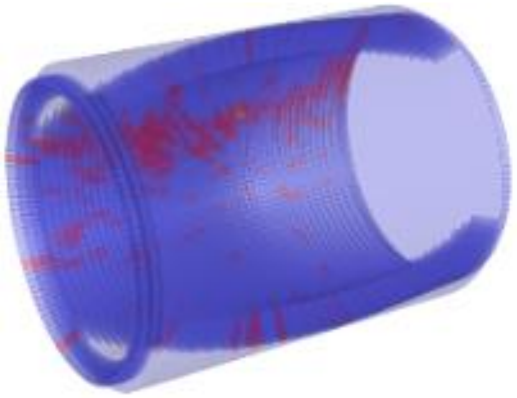
- **Energy deposition pixel map（71*136）：**
  - X-coordinate： Position information
    - Left endcap / Right endcap（0-9/61-70)
    - Barrel（10-60)
  - Y-coordinate： CrystalID
  - Value： Energy deposition inside the crystal



- Energy deposition pixel map



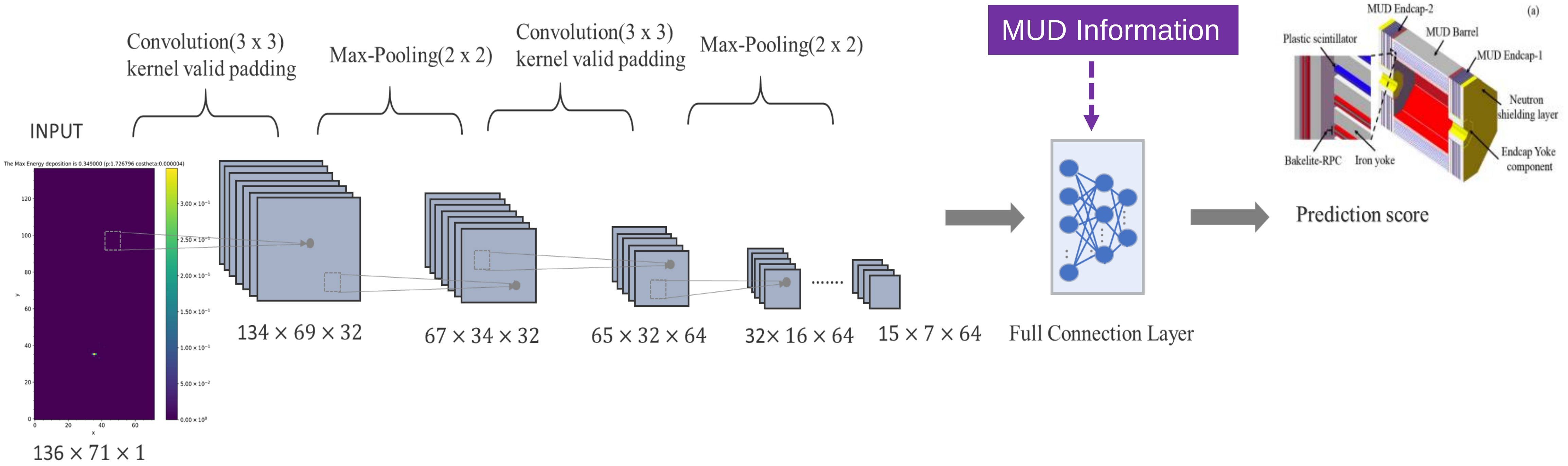
- Energy Deposition for K_L^0



- Energy Deposition in ECAL

- **Neutral Particle Data Sample：**
  - $\gamma/K_L/n$
  - Generated by ParticleGun
  - 100,000(Each type)
  - $P \in (0, 2.0)$ Gev/c, $\theta = 90°, \varphi = 0°$

CNN



- The initial implementation of a global neutral particle discriminator based on CNN
- CNN consists of alternating convolutional and pooling layers, ending with fully connected layers
  - * Convolutional Layer: Use convolutional kernels to extract new hidden features
  - * Pooling Layer: Reduce data dimensionality, prevent overfitting, and reduce resource usage
  - * Fully Connected Layer：Add MUD information in the future

Performance

■ **Analyzing the energy deposition distribution in ECAL** (preliminary)

•***Neutron***:

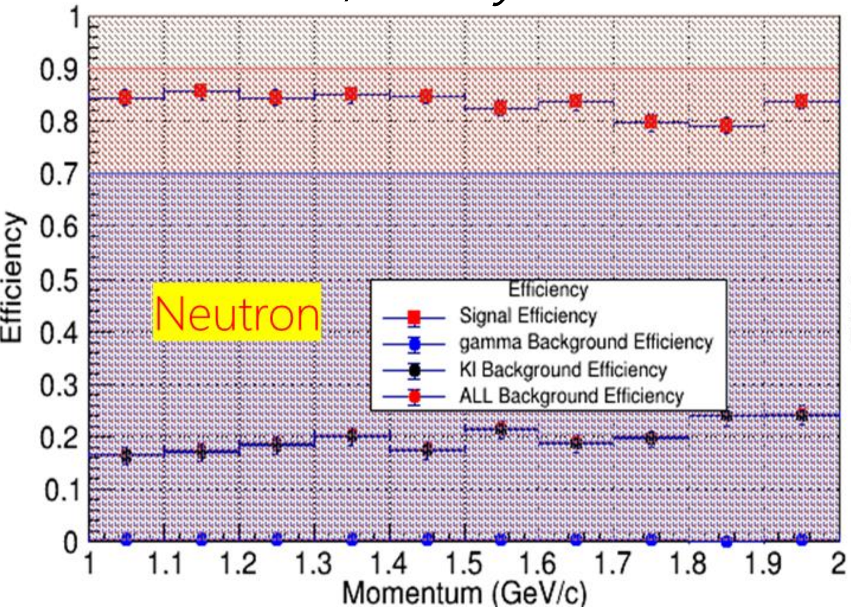
- *Signal efficiency is controlled to be above 80%.*
- *Background misidentification ~20%, mainly for KL*

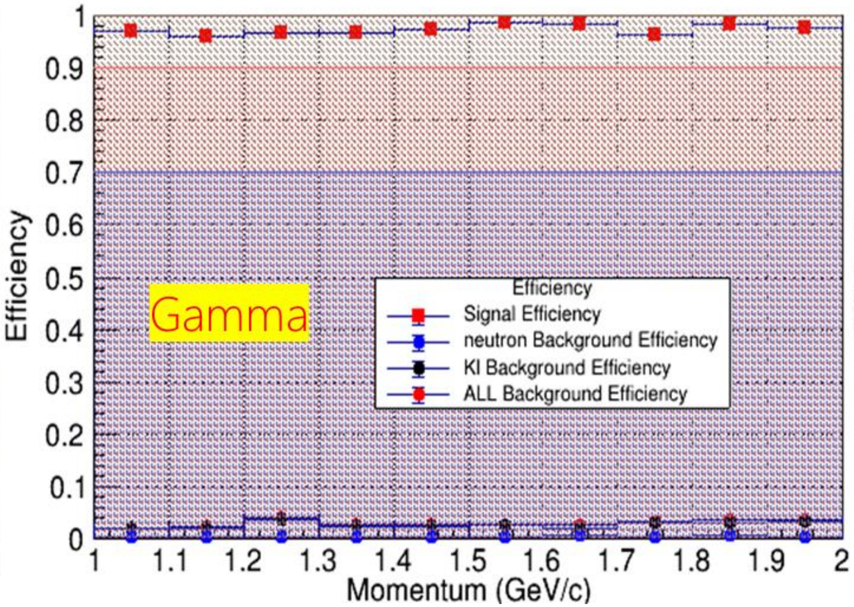
•***Gamma***:

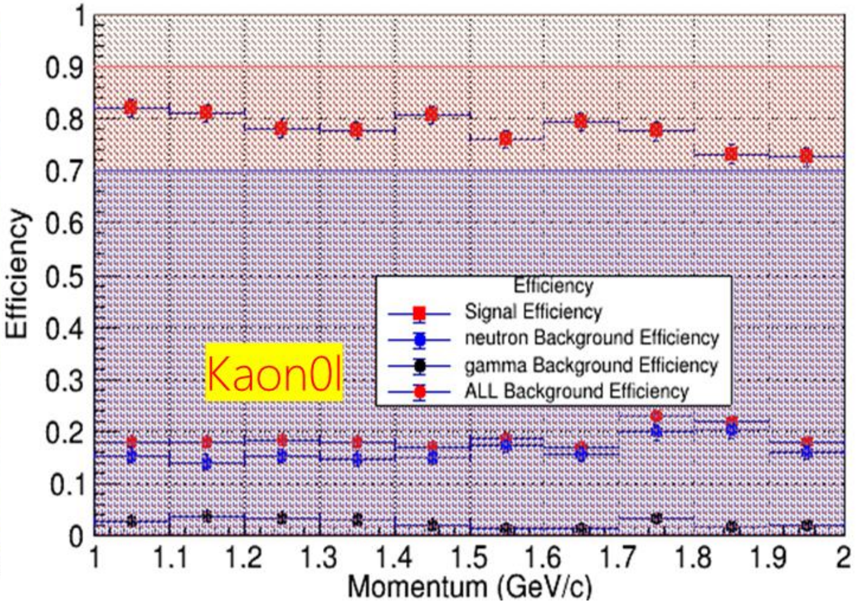
- *Good photon discrimination performance*
- *Signal efficiency > 90%*

•***K_L***:

- *Signal efficiency >70%*
- *Background misidentification ~20%, mainly for Neutron*



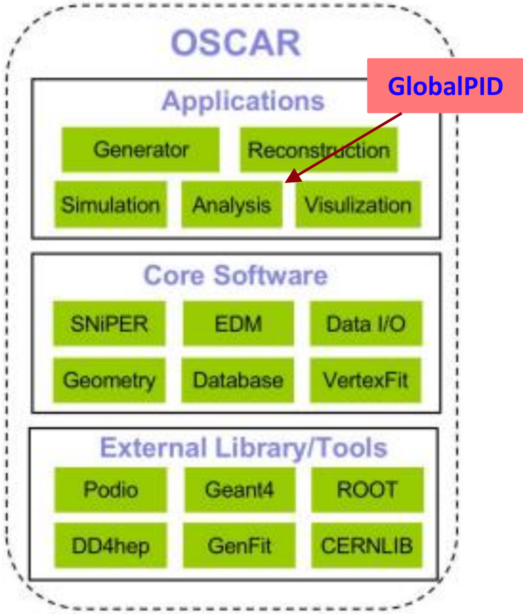


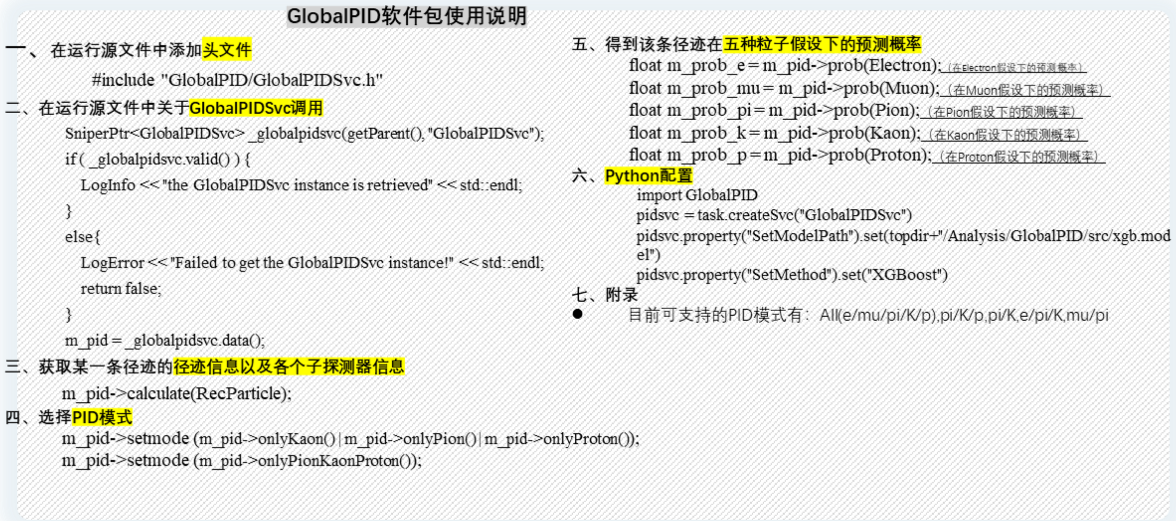


The neutron and K_L identification capability still needs improvement

GlobalPID Software

- The BDT model and GlobalPID algorithm have been integrated into OSCAR software and is available for analysis and research.
  - For the identification of charged particles
  - Pre-trained model is integrated, and made transparent to users
  - Based on C-API of XGBoost, Provided simple interface and user manual for users





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.model")
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

- Development of the ML-based software packages for hadron and neutron particle identification.
- The GlobalPID packages integration : All the software packages will be transferred into the ONNX framework.

Summary

- To fully exploit the performance of the STCF detector, a novel GlobalPID algorithm based on machine learning is developing.
- Based on a data-driven method, BDT is used as a baseline to discriminate charged particles at STCF.
  - Extract features from many correlated variables(integrating all sub-detector information)
  - Provides charged particle identification performance in different PID modes
  - Drive the fast simulation work
- Integrate PID system information and use CNN to achieve hadron discrimination.
- A global neutral particle identifier based on CNN is initially implemented.
- Preliminary results for the identification of charged and neutral particles have been obtained, but need to be further checked and validated.
- The GlobalPID software package has been completed for charged particles identification and is available for analysis and research.

More study is needed to do:

- Add time response and MUD information to neutral particle identification
- Further study the variables used for PID
- Try other machine learning techniques
- Upgradation and result verification for GlobalPID software package

THANKS

Backup

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’	

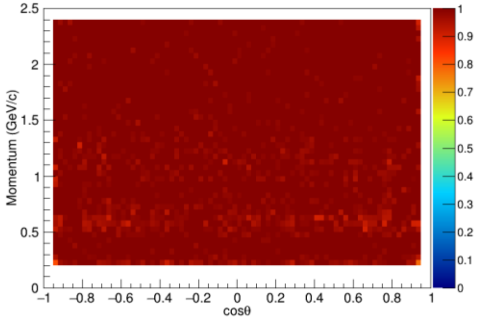
	特征量信息	说明
DEDX	'dEdXsepE/MU/PI/K/P'	基于五种粒子假设下的chi2值
RecECALShower	‘numHits’	在ECAL里的击中数目
	‘energy’，	重建粒子的能量
	‘eSeed’	种子的能量
	‘e3x3’	3*3晶体内的能量沉积
	‘e5x5’	5*5晶体内的能量沉积
	‘position.x’	Shower的x坐标
	‘position.y’	Shower的y坐标
	‘position.z’	Shower的z坐标
	‘secondMoment ’	二阶矩阵
	‘LateralMoment ’	横向矩阵
	‘ZernikeMoment{2,0} ‘	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

Backup 2

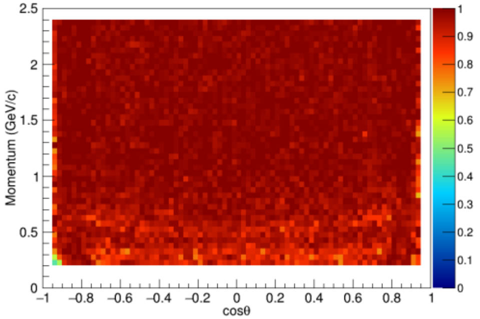
SHANDONG UNIVERSITY

- ## *Efficiency distribution*

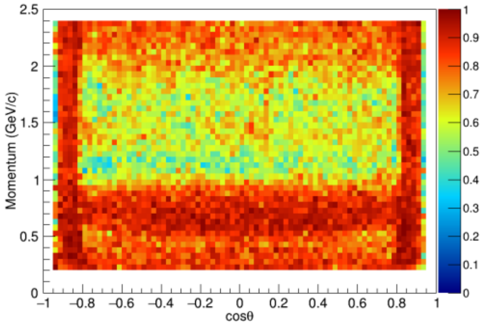
E-



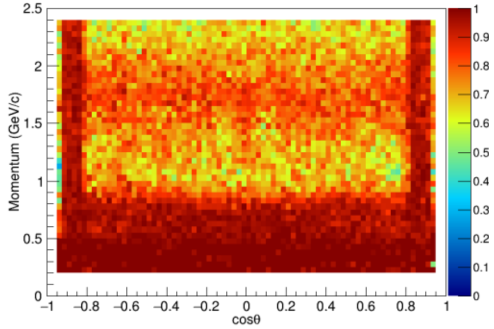
Mu-



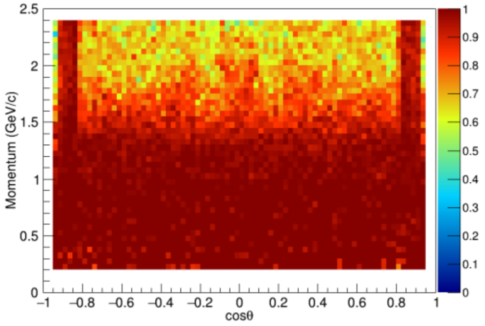
Pi-



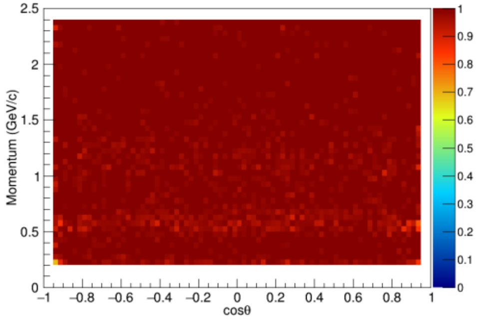
K-



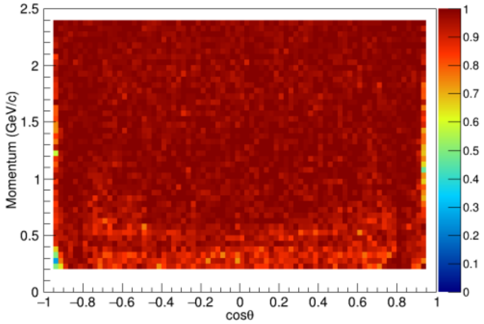
P-



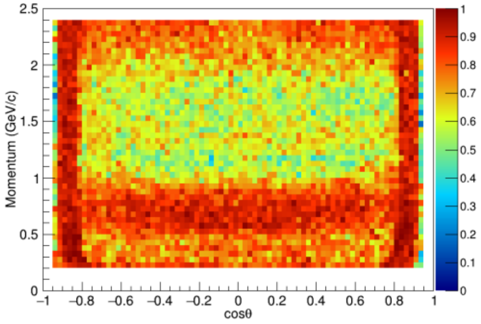
E+



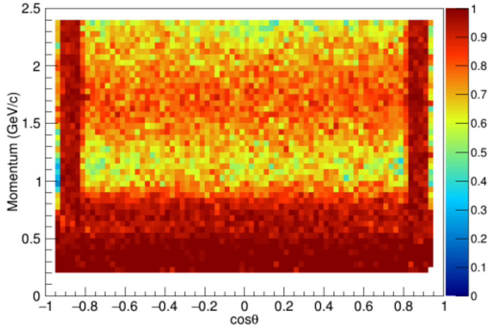
Mu+



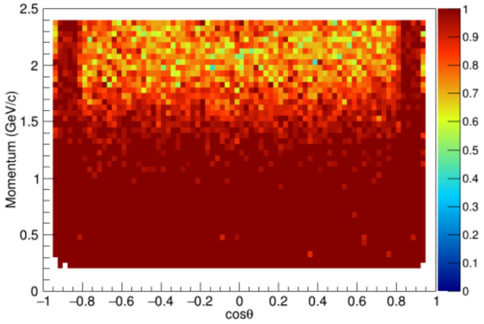
Pi+



K+

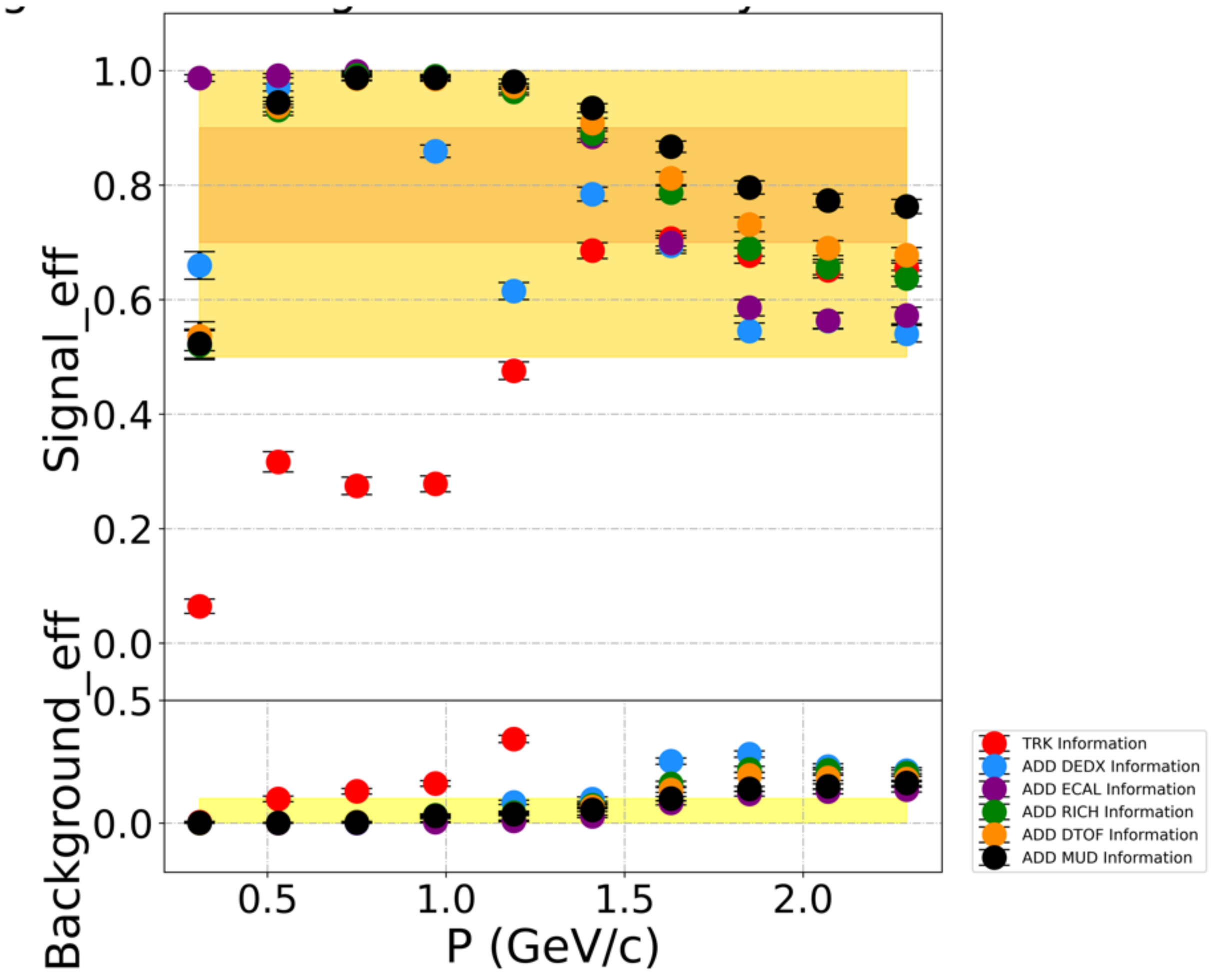


P+

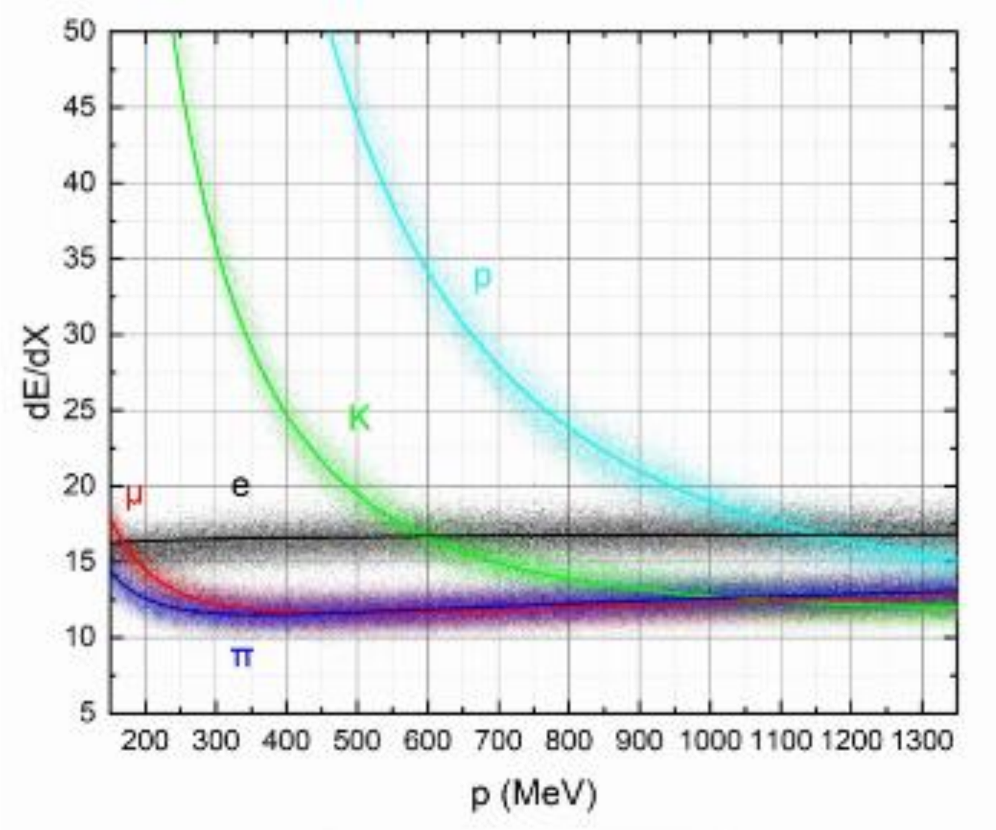


SHANDONG UNIVERSITY

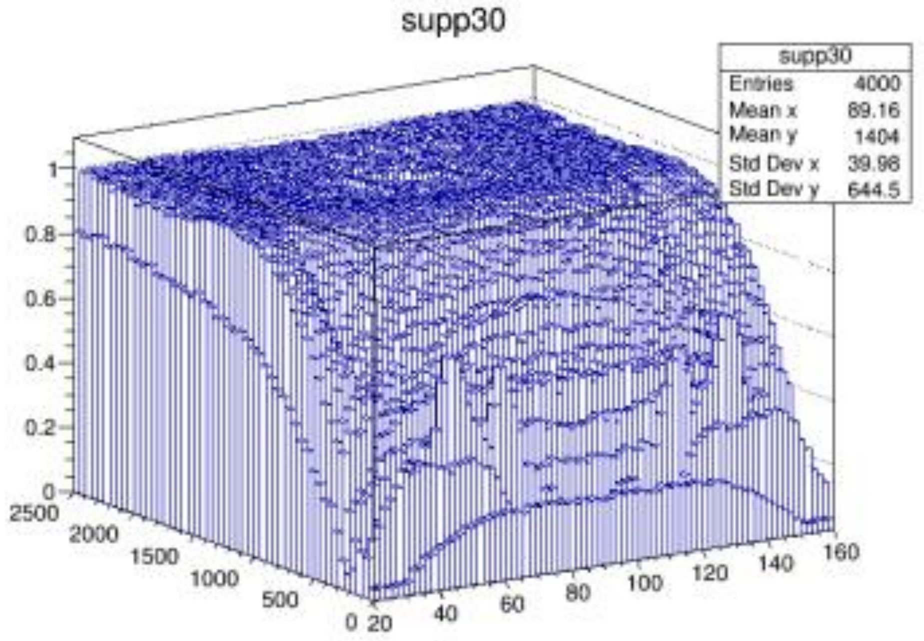
- ***The signal efficiency of Proton***



- *DE/dx Sepa.*



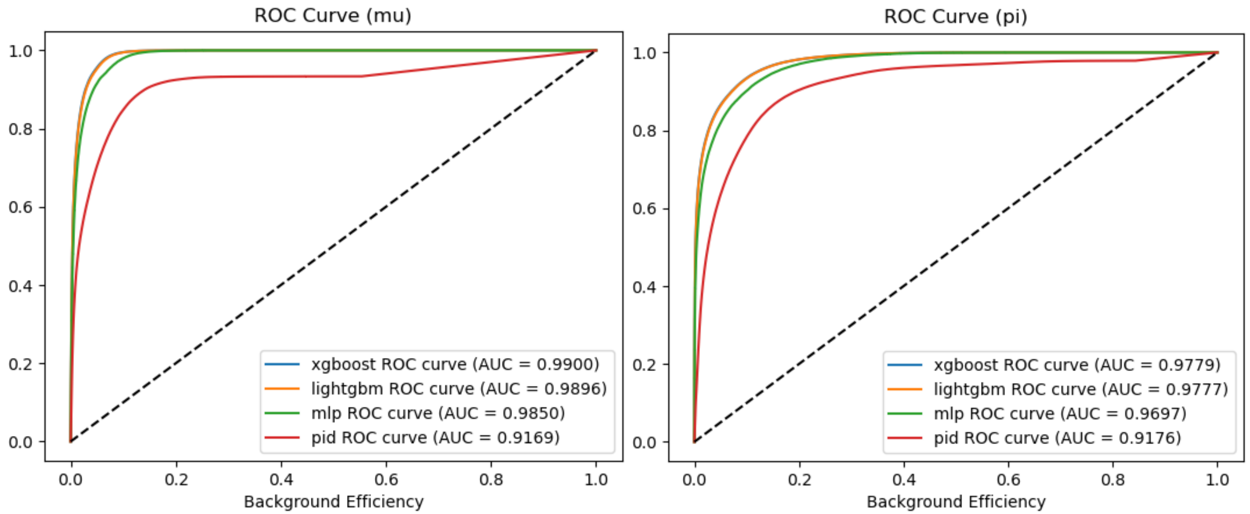
- *MUD Sepa.*



Backup 5

SHANDONG UNIVERSITY

- 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