2023年粒子物理实验计算软件与技术研讨会

Muon/Pion Identification Based on Machine Learning Algorithm at BESIII

报告人：翟云聪

指导老师：李腾、黄性涛

2023年6月11日

目录

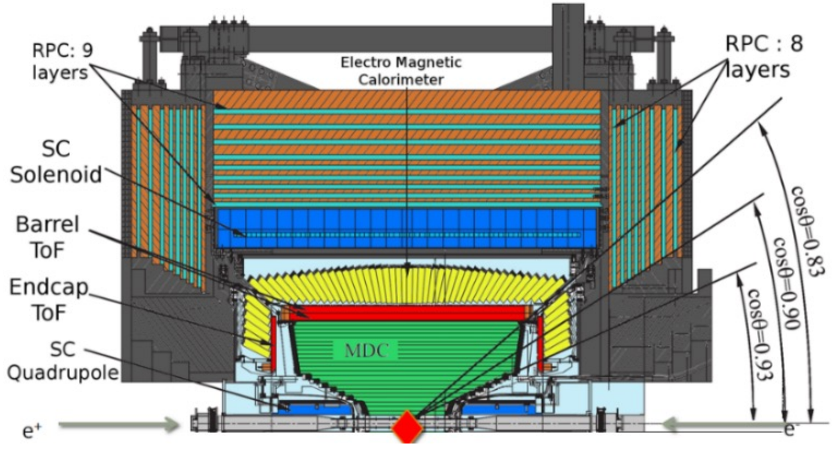
CONTENTS

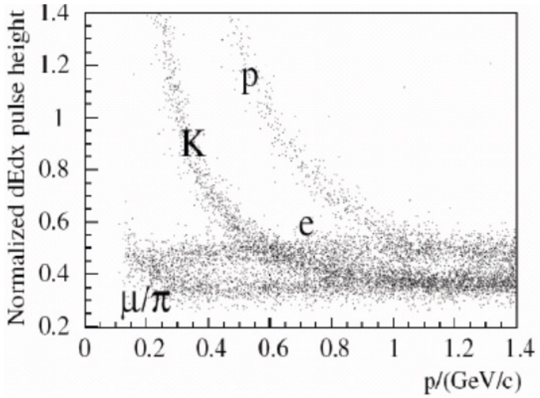
INTRODUCTION

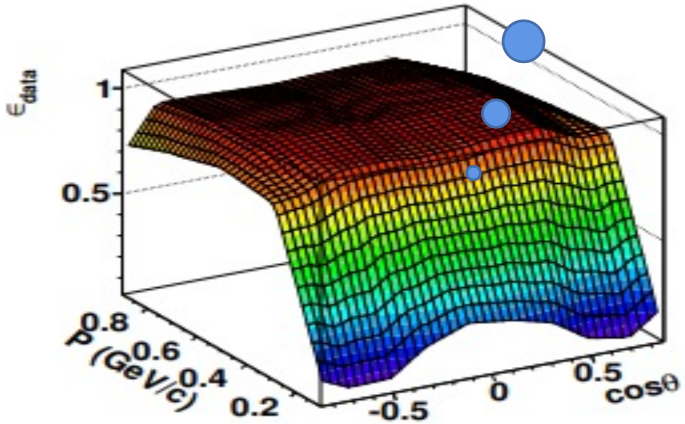
INTRODUCTION

- Particle identification (PID) is one of the most important and commonly used tools for the physics analysis in collider physics experiments.
- For BESIII experiment, traditional methods like the maximum likelihood method are difficult to improve due to the intrinsic correlations between input variables.
  - Especially for very challenging problem: muon/pion separation

Great room for improvement at certain regions







- The muon discrimination efficiency w.r.t. momentum and cos θ by traditional PID software.

INTRODUCTION

- **In recent decades, The data-driven machine learning (ML) has provided a powerful toolbox.**
  - ML based techniques have been rapidly developed and have shown successful applications in HEP experiments .
  - ML have developed rapidly and achieved outstanding results in the field of particle identification. (Hot topic)
  - One of the obvious advantages of applying ML to PID is its capability of combing many correlated variables to solve the most difficult problems for traditional methods
  - Previous studies show that the gradient boosting decision tree (typically BDT) has superior performance
- **Targeting at the muon/pion identification problem at the BESIII experiment, we have developed a new PID algorithm based on the BDT algorithm.**
  - Further improving the performance of traditional PID algorithms and exploring its physical potential

METHODOLOGY

METHODOLOGY

In order to fully explore the PID performance of the detector. Using advanced BDT (XGBoost) , develop a novel muon/pion PID algorithm. (Challenging)

01 Configuration

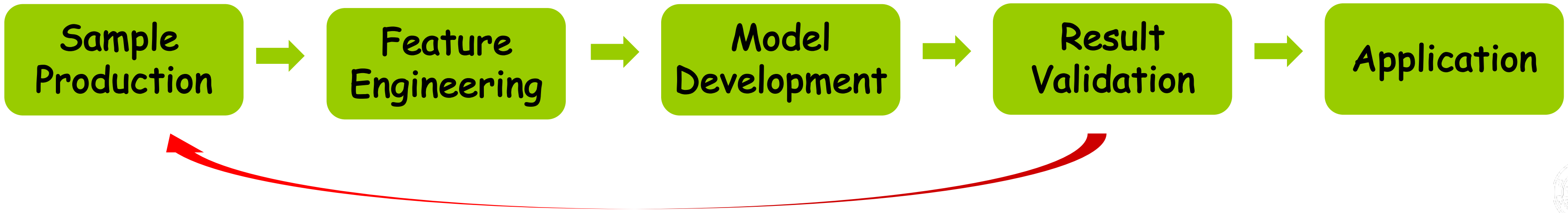
- Based on a data-driven approach, BDT is used as a key technical approach.
- Selected hyper-parameters:
  - max_depth: 8
  - n_estimators: 300

02 Systematic errors

- Systematic error :

$$\Delta\varepsilon = \frac{\varepsilon(Data) - \varepsilon(MC)}{\varepsilon(MC)} \quad (\varepsilon : \text{PID efficiency})$$

- Through detailed cross-validation to evaluate deviations :
  - Different decay processes
  - MC/data



DATA SAMPLE & FEATURE SELECTION

DATA SAMPLE

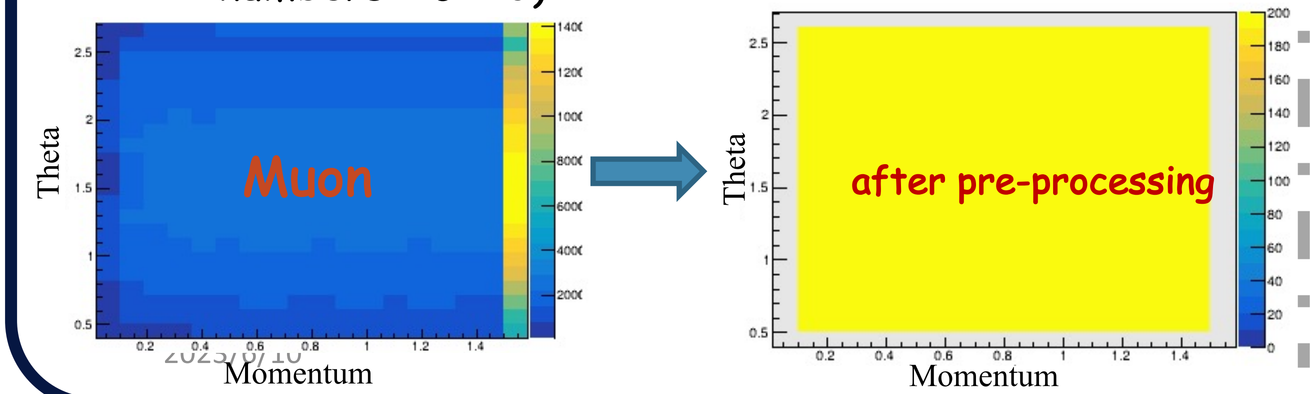
Based on the substantial amount of high-quality Monte Carlo simulation (MC)/real data samples from BESIII, relying on its mature offline software system (BOSS).

Train sample

- ➢ Single muon/pion MC samples
- ➢ High purity and well distribution (Pre-processing)

  – Make sure the distribution of p and $\cos\theta$ is flattened to avoid bias

  – $0.1\ GeV/c < p < 1.5\ GeV/c, -0.88 < \cos\theta < 0.88$ (bin numbers :16*20)



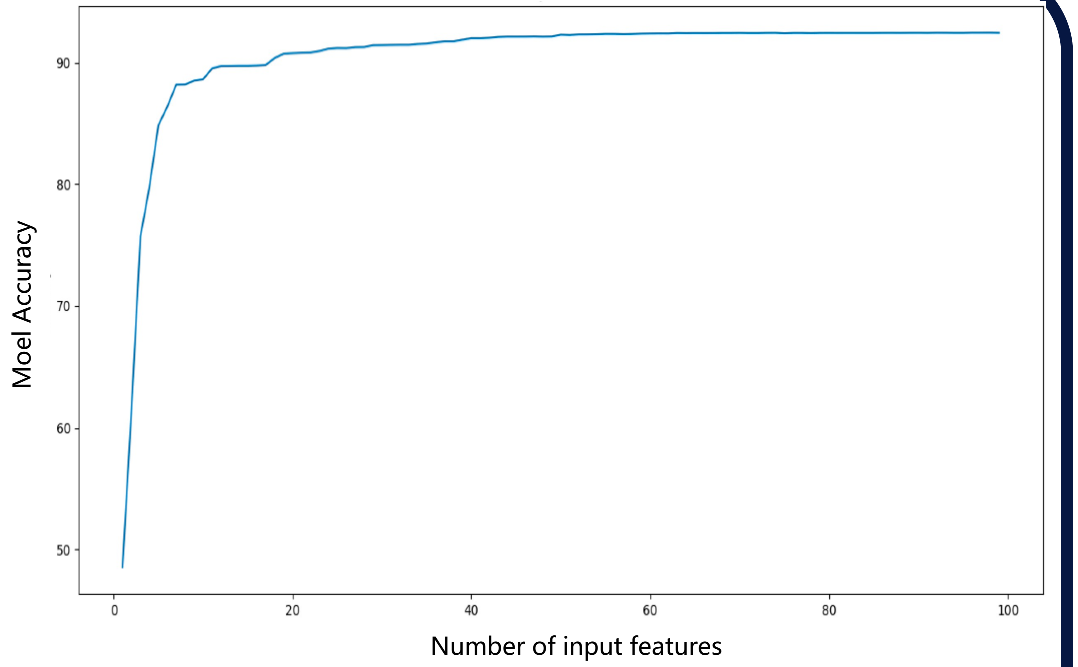
Cross-validation sample

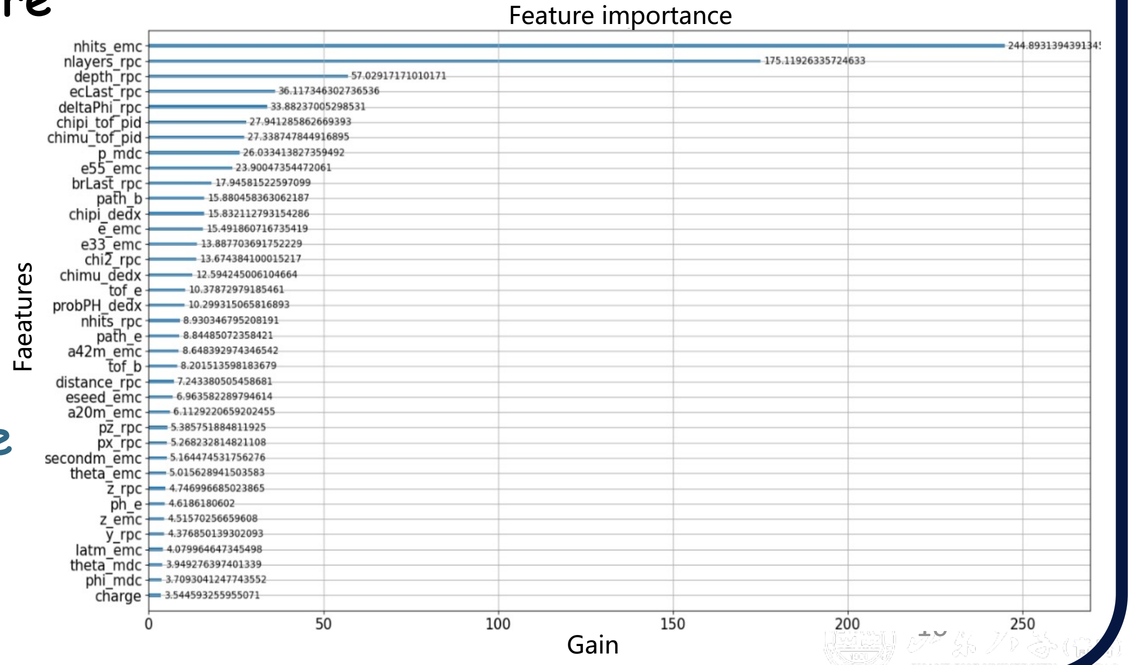
The purity (P) of the μ/π samples : $\frac{N_{sample\ ture}}{N_{sample}}$

- ➢ Different decay processes:
  - – $\psi(2s) \rightarrow \pi^+\pi^- J/\psi \rightarrow \pi^+\pi^-\ \mu^+\mu^-$ (P = 99.13%)
  - – $J/\psi \rightarrow \pi^+\pi^-\ \pi^0 \rightarrow \pi^+\pi^-\gamma\gamma$ (P = 99.37%)
  - – $J/\psi \rightarrow \gamma\ \mu^+\mu^-$ (P = 97.97%)
- ➢ MC/data:
  - – $J/\psi \rightarrow \pi^+\pi^-\ \pi^0 \rightarrow \pi^+\pi^-\gamma\gamma$ (P = 99.37%)
  - – $J/\psi \rightarrow \gamma\ \mu^+\mu^-$ (P = 97.97%)

FEATURE SELECTION

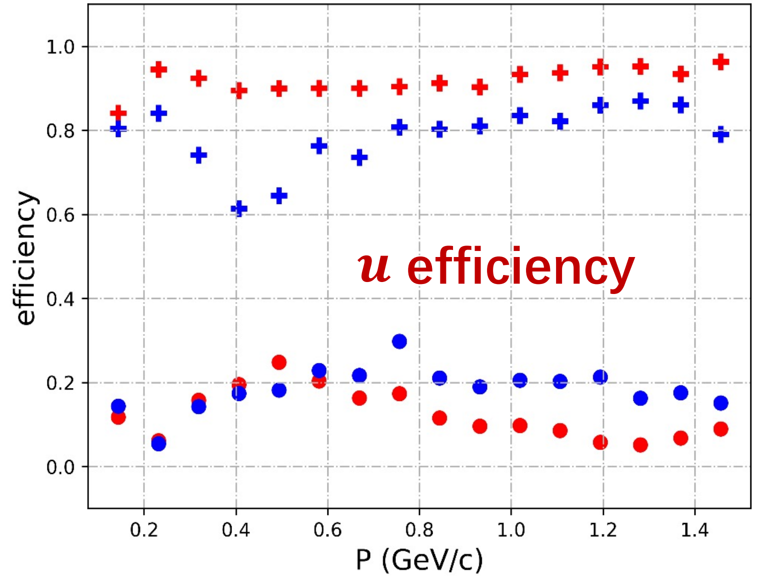
- To extract effective features from a large amount of interrelated sub-detectors information.
- First model trained with all 108 features.
  - Contain MDC, dE/dX, TOF, EMC, MUC information
  - Based on XGBoost (as baseline)
- Features are then selected according to feature importance.
  - Eliminate redundant features to reduce training time
  - Eliminate features that have large MC/Data deviation to suppress systematical error
- Eliminate strongly-correlated features, 37 features are kept

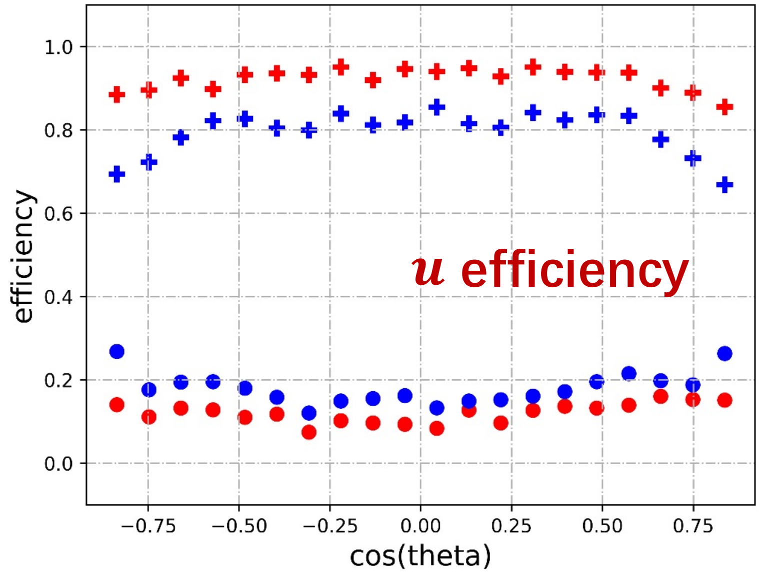


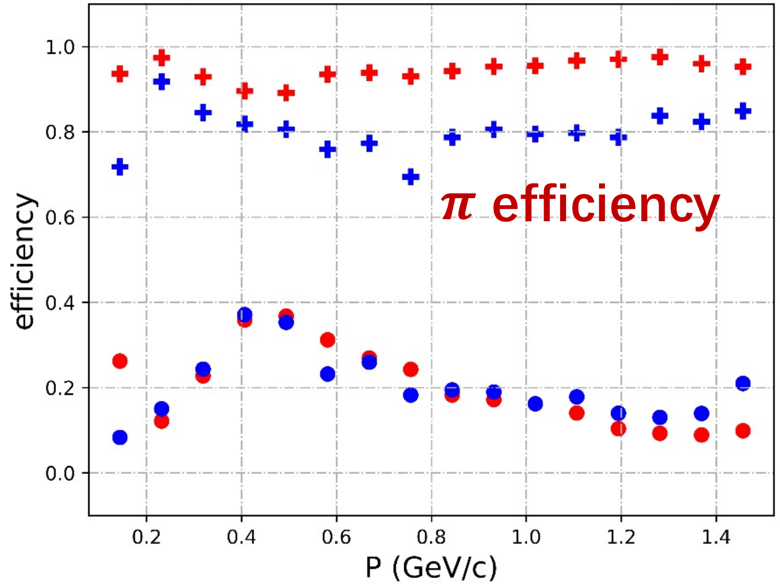


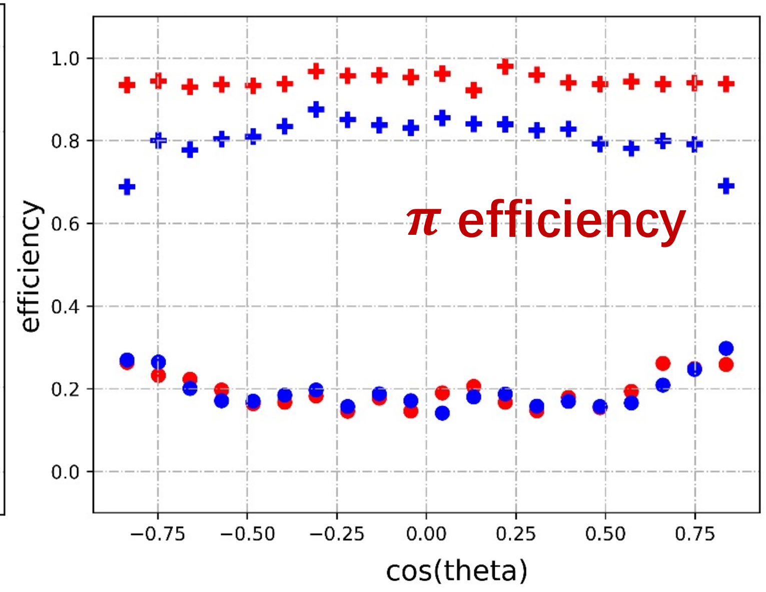
PERFORMANCE ANALYSIS

Comparison with traditional PID algorithm









Signal efficiency

- XGboost model
- Traditional PID algorithm

Background efficiency

- XGboost model
- Traditional PID algorithm

➤ *Signal efficiency*:

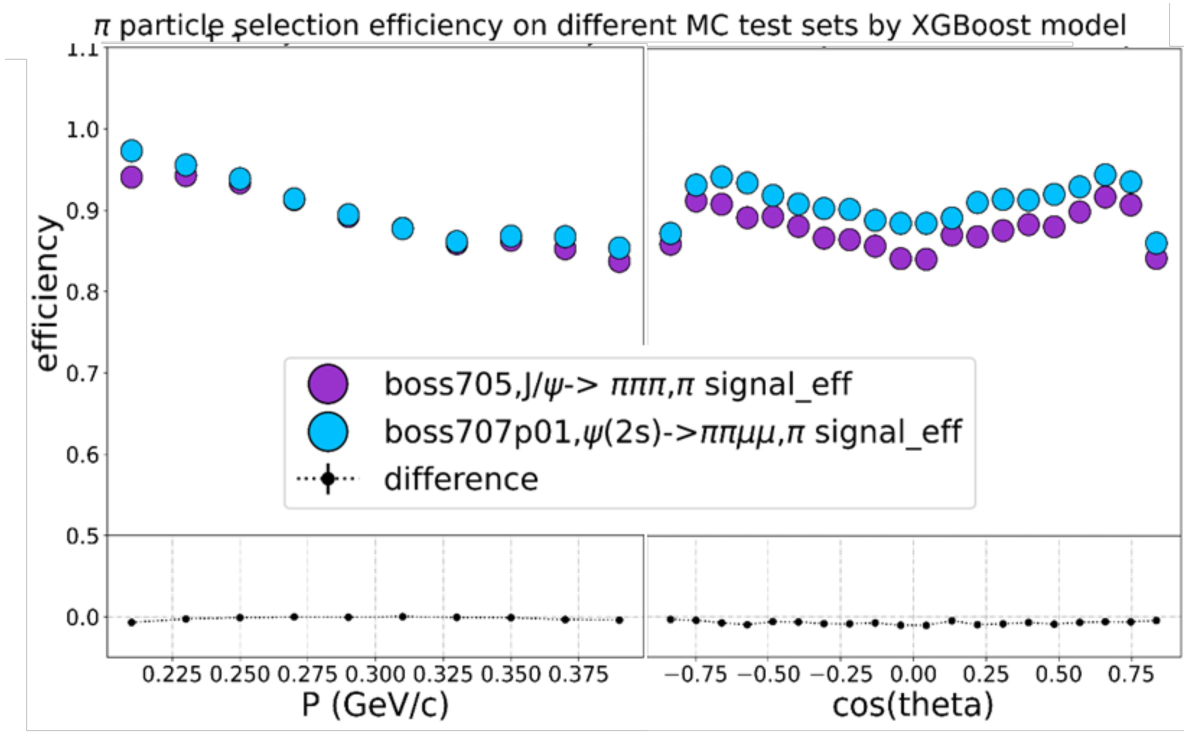
$$\frac{\text{The number of signal selected correctly}}{\text{The total number of signal}}$$

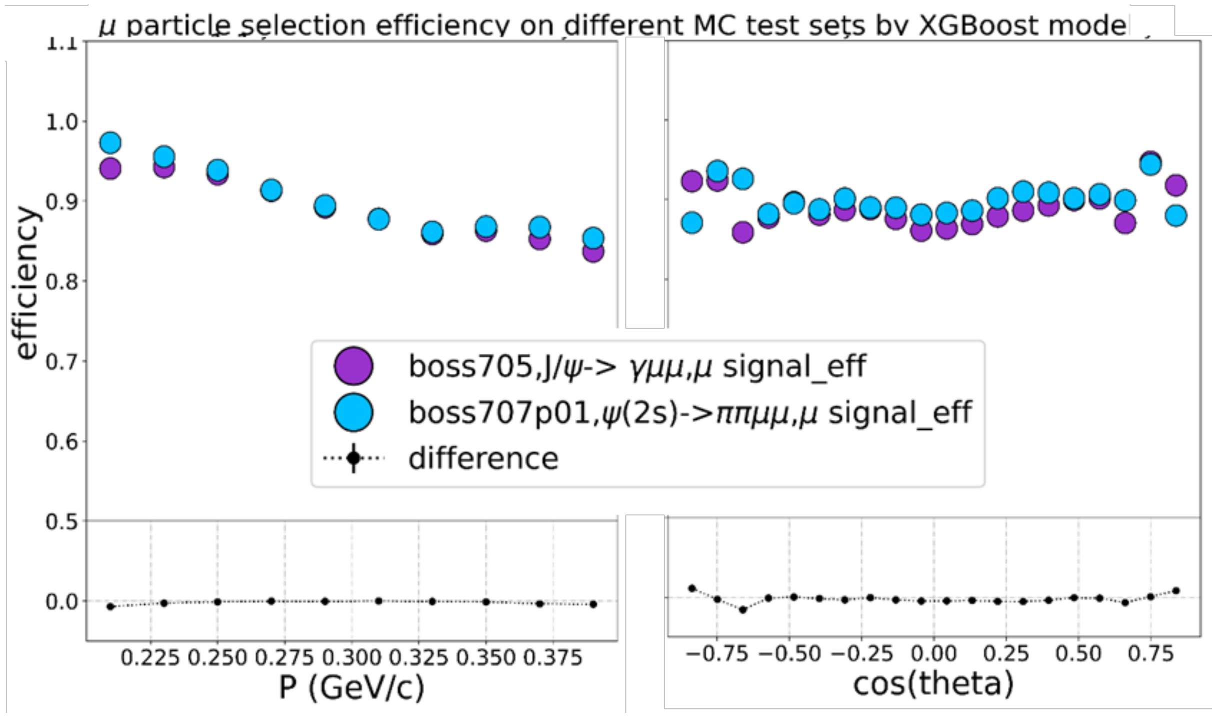
➤ *Background efficiency*:

$$\frac{\text{The number of background misidentified as signal}}{\text{The total number of background}}$$

Cross validation between different decay processes

- To check generalization ability
- To estimate the deviations different decay channels





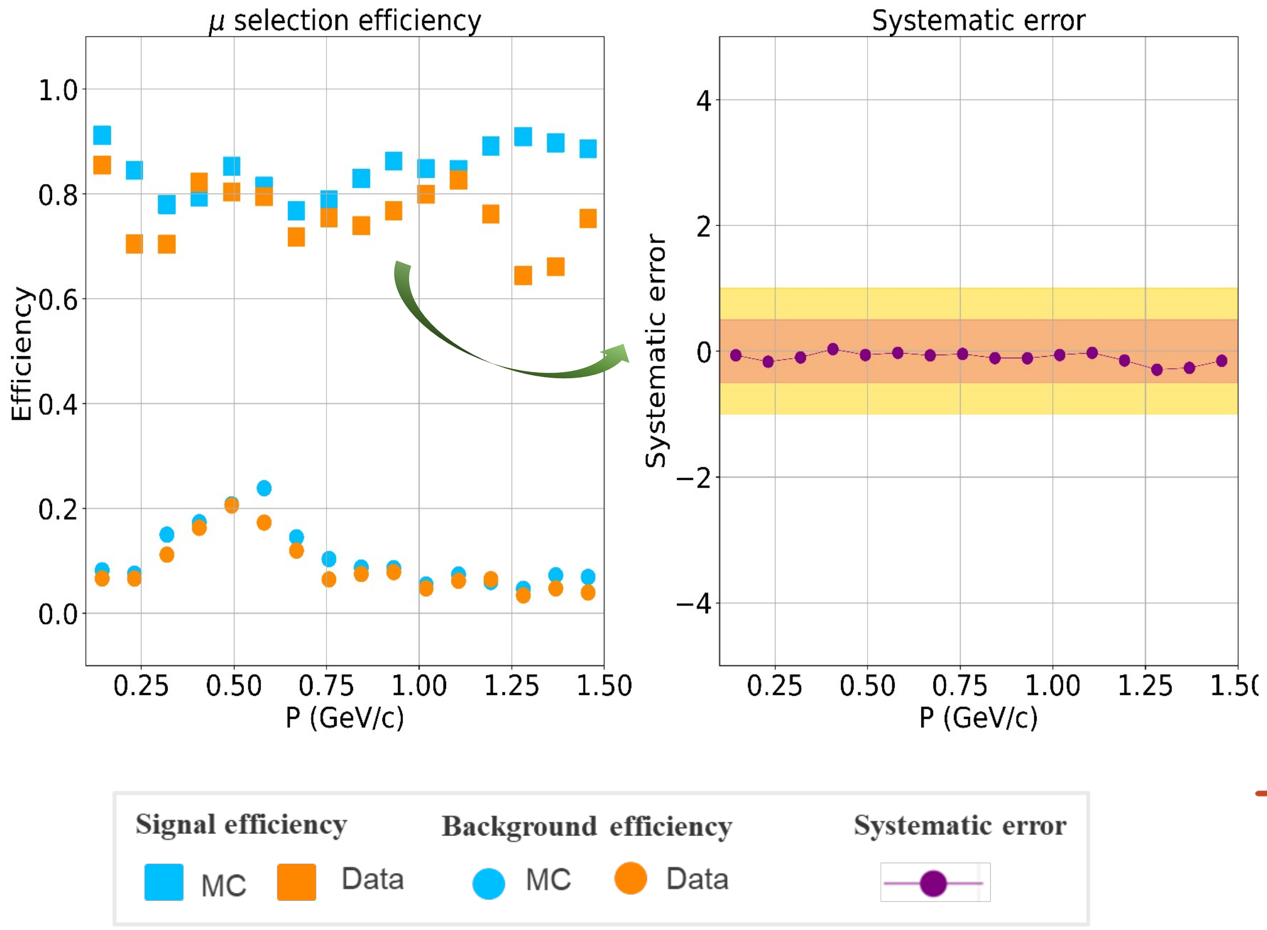
Cross validation between MC and Data

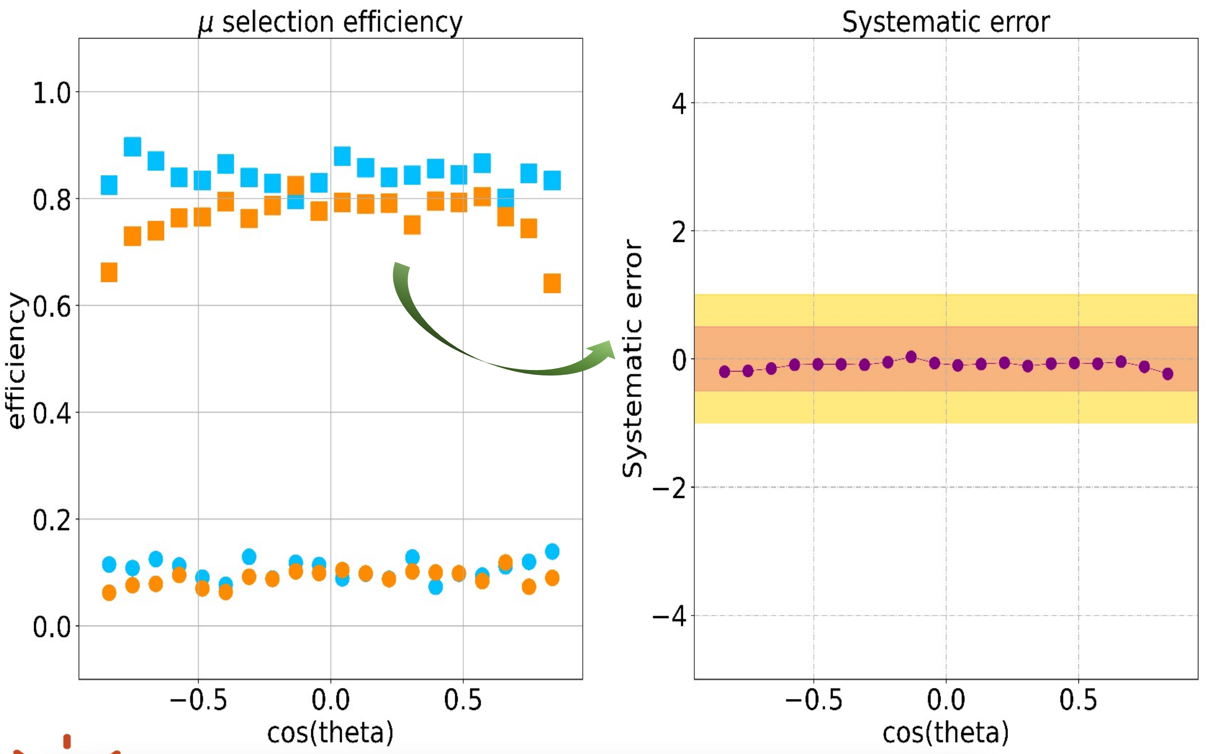
MC/data:

$J/\psi \rightarrow \pi^+\pi^- \pi^0 \rightarrow \pi^+\pi^-\gamma\gamma$ (P = 99.37%)

$J/\psi \rightarrow \gamma\, \mu^+\mu^-$ (P = 97.97%)

– To estimate systematical error





Systematic error :

$$\Delta\varepsilon = \frac{\varepsilon(Data) - \varepsilon(MC)}{\varepsilon(MC)} \quad (\varepsilon : \text{PID efficiency})$$

BOSS Integration

To make the algorithm available to analyzers, a BOSS package is developed

- For easy-to-use, the package is integrated with BESIII Event Data Model
- Based on C-API of XGBoost, and provided similar interface with PID package
- Pre-trained model is integrated, and made transparent to users

Will make available to public once validated

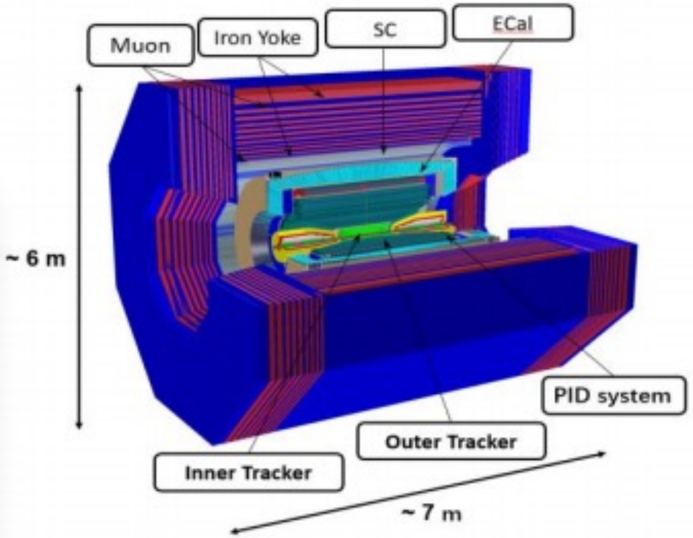
```
#include "DeepParticleID/DeepParticleID.h“

StatusCode AnalysisAlg::execute() {
  // .......
  DeepParticleID* Deeppid = new DeepParticleID(XGBoost);
  Deeppid->calculate(*itTrk);
  float prob_mu = Deeppid->prob(0);
  float prob_pi = Deeppid->prob(1);
  if (prob_mu > prob_pi) {
      //......
  }

  //........
}
```

GlobalPID Algorithms Based on Machine Learning at STCF

GlobalPID Algorithm

超级陶粲装置（STCF）

- The Super Tau Charm Facility (STCF), with a luminosity greater than 0.5×10^{35} cm^{-2} s^{-1} and a center-of-mass energy range of 2-7 GeV, is an important option for China's future accelerator-based particle physics large-scale scientific facility.
- The development and research of the Global Particle Identification (GlobalPID) software algorithm is crucial for achieving the future physics objectives of the STCF experiment.
  - PID software is an important component of The STCF offline software system (OSCAR).
- Building on the experience gained from particle identification work in the early stages of BESIII, the STCF experiment will utilize advanced ML techniques to innovate and develop the GlobalPID algorithm.
  - By integrating all sub-detector information
  - To fully exploit the PID performance of the detector
  - Needed to facilitate the progress of physics analysis work



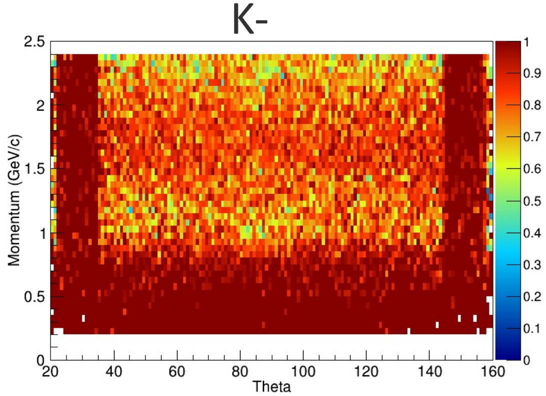
Physics Objectives

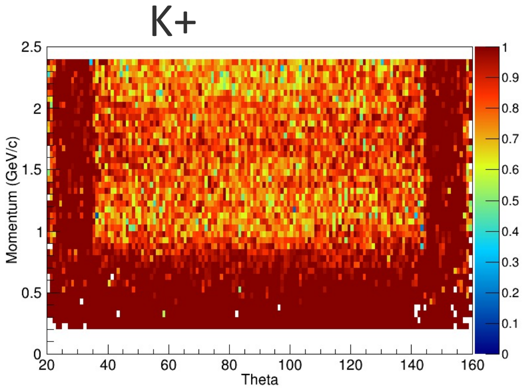
- Searching for new exotic hadronic states
- Studying flavor physics and CP violation physics
- Searching for new physics beyond the standard model at the forefront of high precision

GlobalPID Algorithm

超级陶粲装置（STCF）

- Based on OSCAR simulation and reconstruction results, Tracker/dEdx/RICH/DTOF/ECAL/MUD information have been collected. (Full list of variables please see backup slides)
  - 50000 tracks for each type (e±, μ±, π±, K±, p±)
  - MC single charged track using ParticleGun
  - p∈(0.2, 2.4)Gev/c, θ∈(20°, 160°), phi = 0°
- ML model(based on XGBoost) is trained and optimized to discriminate (e, μ, π, k, P)
- Preliminary results have been obtained. The model and GlobalPID algorithm have been integrated into OSCAR software and is available for analysis and research.



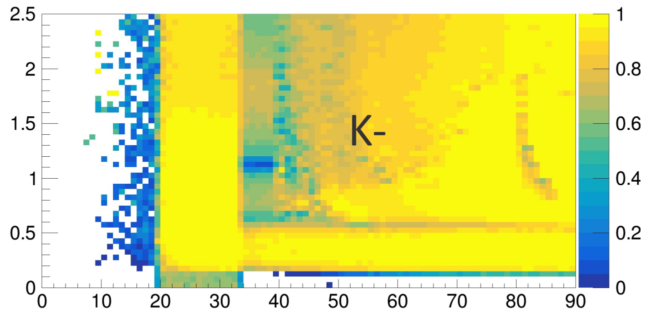


X-axis:Theta（20，160，140bins）

Y-axis: Momentum（0，2.5，50bins）

Color gradient :Efficiency (0,1)

- The PID efficiency of BESIII



Performance needs to be further validated !!

SUMMARY

SUMMARY

✓A muon/pion identification algorithm based on machine learning model (XGBoost) is developed based on the high quality data samples at BESIII and has been integrated into the BOSS.

✓Performance analysis shows XGBoost model provides obviously higher discrimination power than traditional methods.

✓Detailed cross-validation was conducted and an evaluation method for the systematic error of the machine learning model was provided, which can be used by BESIII physics analysts.

- Evaluate deviations between different decay processes
- Evaluate deviations between MC/data

✓ Developed a ML-based GlobalPID algorithm for future STCF experiments.

- Algorithm framework is established
- Integrated into OSCAR software
- Global PID algorithm has preliminary results

THANKS

Backup

Backup 1

SHANDONG UNIVERSITY

- $\pi^{\pm}$ ***selection:*** $J/\psi \rightarrow \pi^{+}\pi^{-}\pi^{0} \rightarrow \pi^{+}\pi^{-}\gamma\gamma$

◆ Good charge Track Selection:

- $|V_z| < 10.0$ cm
- $|R_{xy}| < 1.0$ cm
- $|\cos\theta| < 0.93$
- $N_{\text{good charge}}=2$, total charge=0

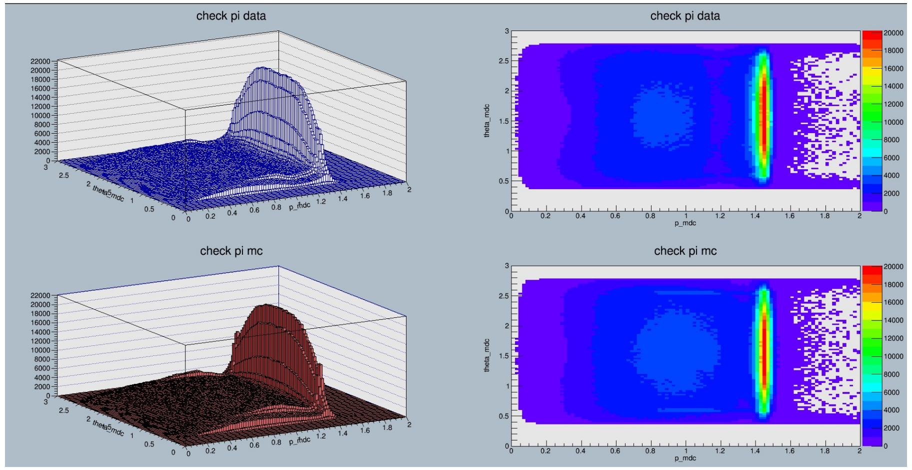
◆ Good photon Selection:

- Eγ>25MeV for barrel EMC(|cosθ|<0.8) or
- Eγ>50MeV for endcap (0.86<|cosθ|<0.92)
- Time spent in emc: 0<t<700ns
- θγ,charge>=10(degree)
- Nγ=2

◆ 4C Kinematic fit:

- |mγγ-0.135|<0.015 GeV
- $\chi^2(\gamma\gamma\pi\pi)<\chi^2(\gamma\gamma KK)$, $\chi^2(\gamma\gamma\pi\pi)<100$

◆ Only one track is used as PID, which needs to meet prob(π)>prob(k) 、 prob(π)>prob(p) and E/P < 0.8.And keep another track information.



Backup 2

- $\mu^{\pm}$ *selection*： $e^{+}e^{-} \rightarrow J/\psi \rightarrow \gamma\mu^{+}\mu^{-}$

◆ **Charge track selection**:

✓ $N_{charge}=2$

◆ **Good photon selection**:

✓ Time spent in emc: 0<t<700ns

✓ Eγ>25MeV for barrel EMC(|cosθ|<0.8) or

✓ Eγ>50MeV for endcap (0.86<|cosθ|<0.92)

✓ Nγ>0

◆ **randomly selected one of the charged traces and compared its momentum with another traces:**

✓ If the momentum of the track is within (1.5, 1.8) GeV, further selection cuts include:

- $|V_z|$ < 10.0 cm
- $|R_{xy}|$ < 1.0 cm
- $|\cos\theta|$ < 0.8
- The energy deposited in the EMC is within (0.05, 0.27) GeV
- The depth in MUC of the track is greater than 40 cm
- prob(mu)>0.001 && prob(mu)>prob(k)&& prob(mu)>prob(e)

✓ If the momentum of the track is less than 1.5 GeV, further selection cuts include:

- $|V_z|$ < 10.0 cm
- $|R_{xy}|$ < 1.0 cm
- $|\cos\theta|$ < 0.93
- Energy deposited in the EMC within (0.03, 0.22) GeV
- prob(mu)>0.001 && prob(mu)>prob(k)&& prob(mu)>prob(e)

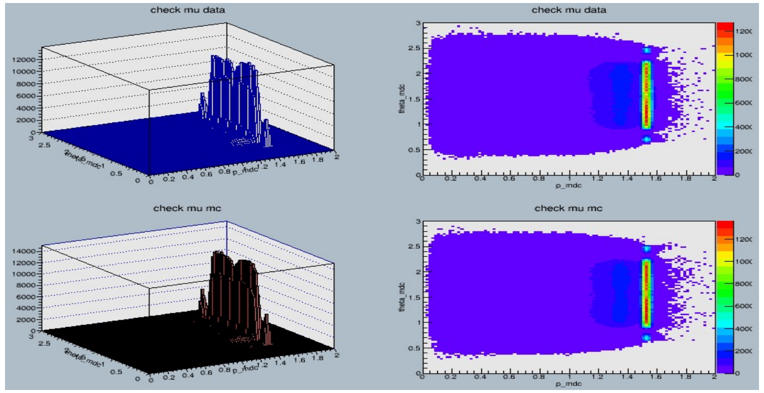
✓ The angle between the photon and the missing momentum is less than 10°.

✓ 4C Kinematic fit:

- The combination with the smallest χ^2 is chosen as the best combination(select good photon)

◆ **Saving that track information without any cut conditions**

◆ **Repeat the three and four steps for the other track in this event**



Backup 3

SHANDONG UNIVERSITY

- *37 features*

  - **MDC:** 'p_mdc', 'theta_mdc', 'phi_mdc', 'charge'
  - **dE/dX:** 'chimu_dedx','chipi_dedx', 'probPH_dedx’,
  - **TOF_B:** 'tof_b', 'path_b'
  - **TOF_E:** 'tof_e', 'path_e', 'ph_e'
  - **TOF:** ' chimu_tof ', 'chipi_tof ‘
  - **EMC:** 'nhits_emc','z_emc', 'theta_emc', 'e_emc', 'eseed_emc', 'e33_emc', 'e55_emc', 'secondm_emc', 'latm_emc','a20m_emc', 'a42m_emc'
  - **MUC:** 'brLast_rpc', 'ecLast_rpc', 'nhits_rpc', 'nlayers_rpc’, 'depth_rpc', 'chi2_rpc', 'y_rpc', 'z_rpc', 'px_rpc', 'pz_rpc','distance_rpc', 'deltaPhi_rpc'

Backup 4

SHANDONG UNIVERSITY

- $\psi(2s) \rightarrow \pi^+\pi^- J/\psi \rightarrow \pi^+\pi^- \mu^+\mu^-$

➢ Good charge Track Selection:

$|V_z| < 10.0$ cm

$|R_{xy}| < 1.0$ cm

$|\cos\theta| < 0.93$

charge $_{\text{every good charge track}} = \pm 1$

$N_{\text{good charge track}} = 4$, total charge=0

➢ Candidates for π :

✓ The momentum of the charged track is required to less than 1 GeV

➢ Candidates for $\boldsymbol{\mu}$:

✓ The momentum of the charged track is required to greater than 1 GeV

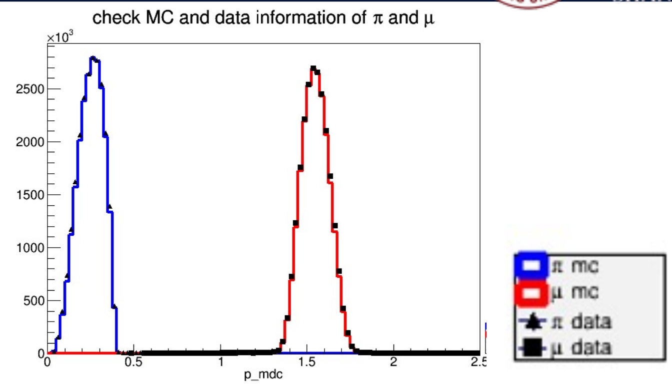
✓ The energy deposited in the EMC of the charged track is less than 0.6 GeV

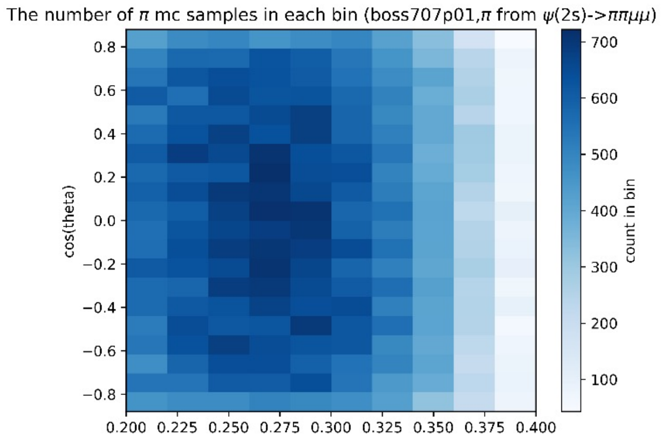
➢ There must be two charged muon candidates,which one is plus and one is minus

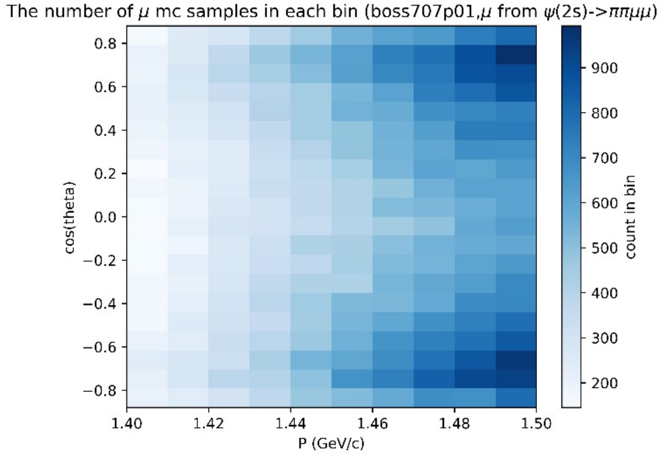
➢ There must be two charged pion candidates,which one is plus and one is minus

➢ 4C Kinematic fit:

Four momentum constrained kinematic fit is perfromed and the χ 2 is less than 200.







Backup 5

SHANDONG UNIVERSITY

	特征量信息	说明
ReconstructinParticle	‘charge’	重建粒子的电荷
	‘mom_x’	
	‘mom_y’	粒子在xyz方向上的动量
	‘mom_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
TrackerRecTrack	‘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一个点（p0）与参考点的距离
	‘helixPar_phi’	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_tanl'	螺旋线倾斜度（pz/pt）
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

	特征量信息	说明
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
DTOFPid(未来增加使用)	‘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'	