基于机器学习的μ/π粒子鉴别

翟云聪　李腾　**秦小帅**　黄性涛

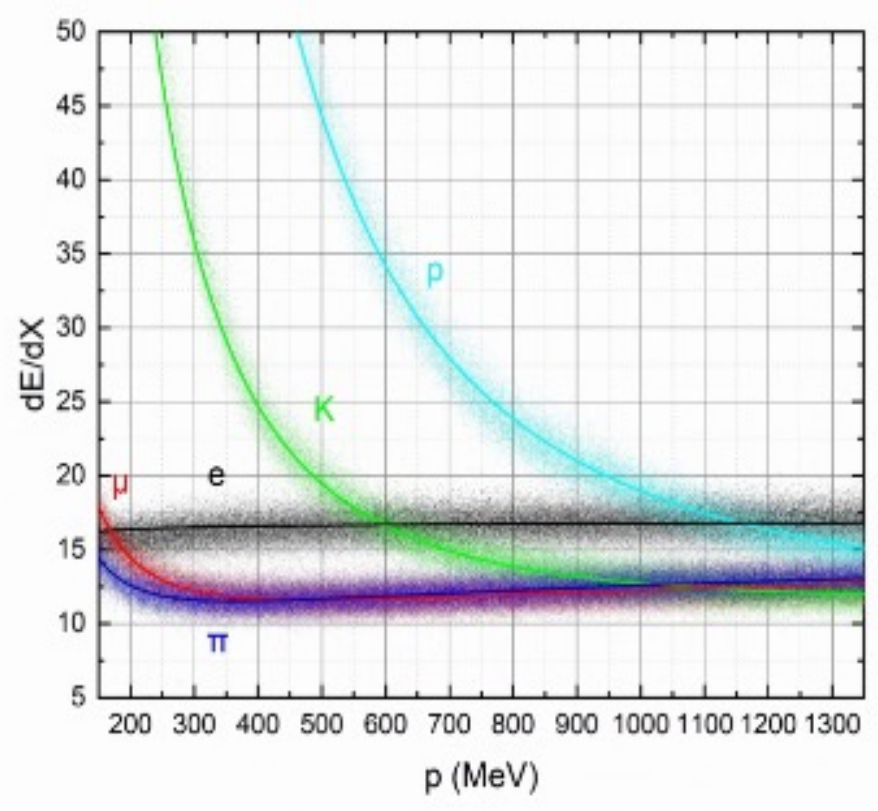
山东大学

BESIII粲强子物理研讨会　兰州

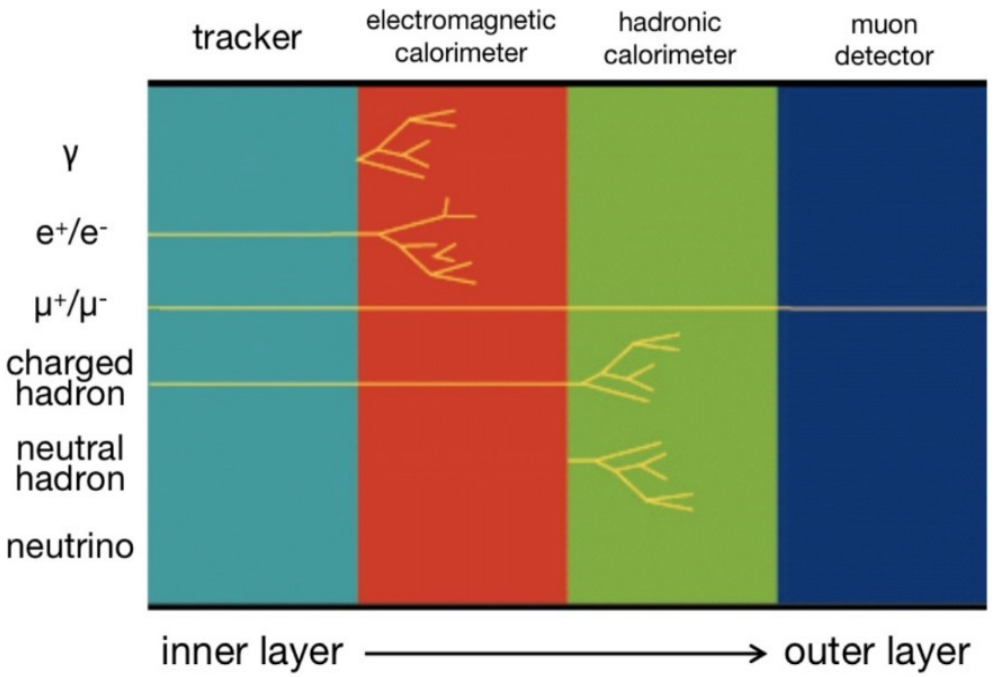
2025-08-07

粒子鉴别

- 粒子鉴别(PID)是高能物理实验中重要的研究方法
  - 借助于动量测量、dE/dx、飞行时间、量能器能量沉积、MUC信息
  - 物理研究要求较高粒子鉴别效率，并压低本底误判



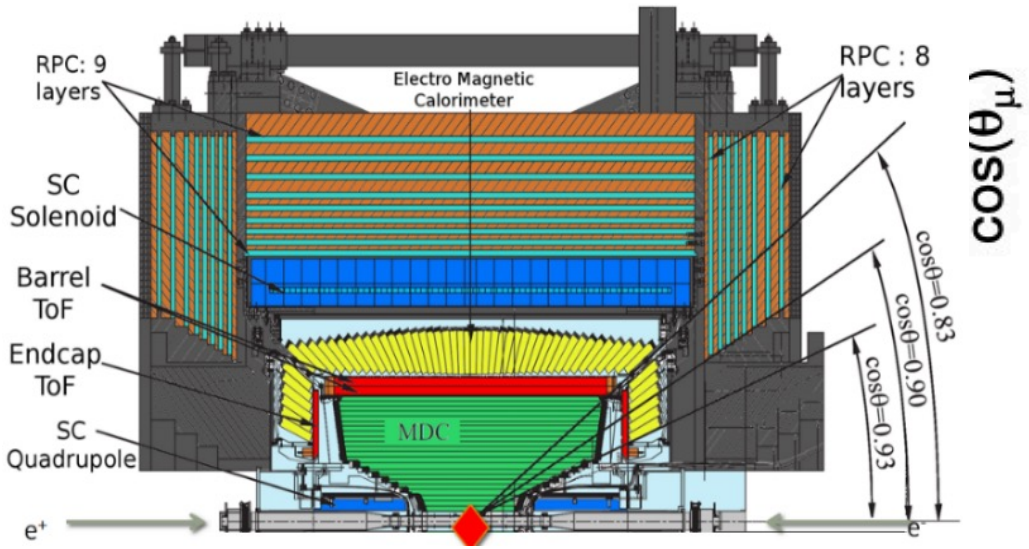
dE/dx 分辨性能.



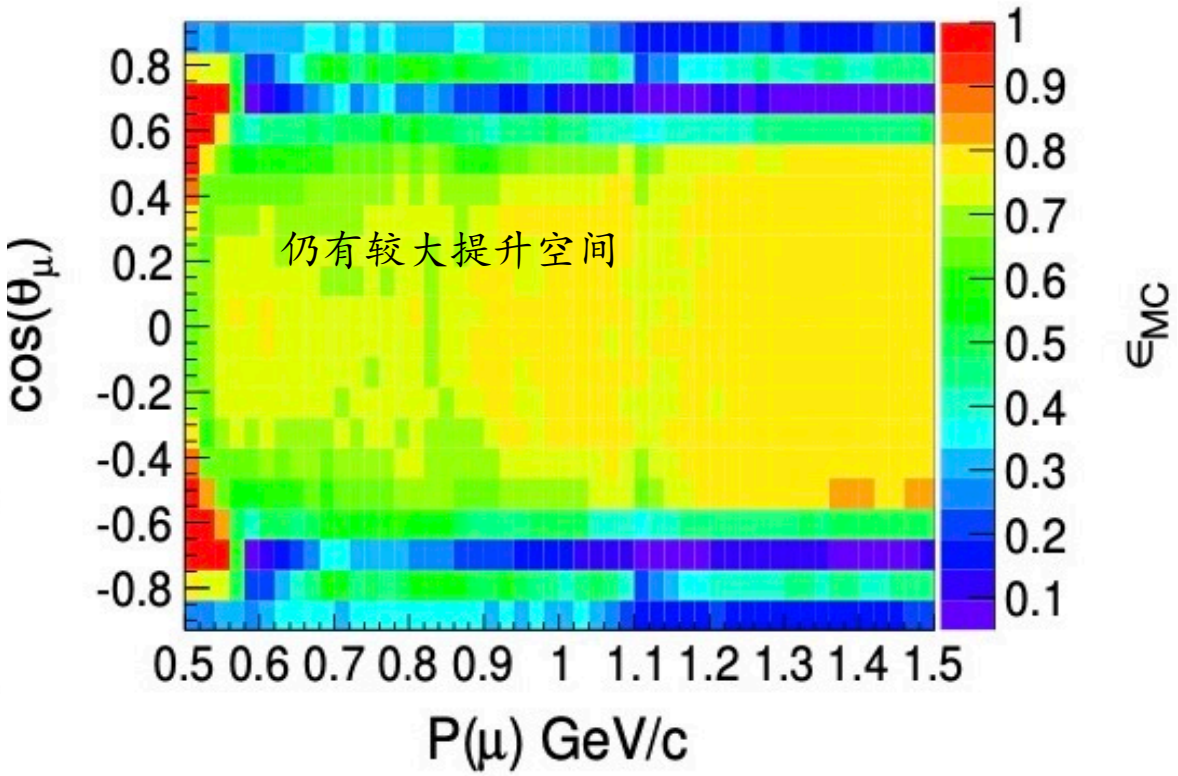
探测器响应

BESIII实验中的传统PID方法

- 基于子探测器的信息构建似然值，最大似然法等，MUC深度等cut
  - 传统的PID方法仍有较大提升空间
- 利用机器学习方法有望提升PID性能
  - 联合来自所有子检测器的底层信息



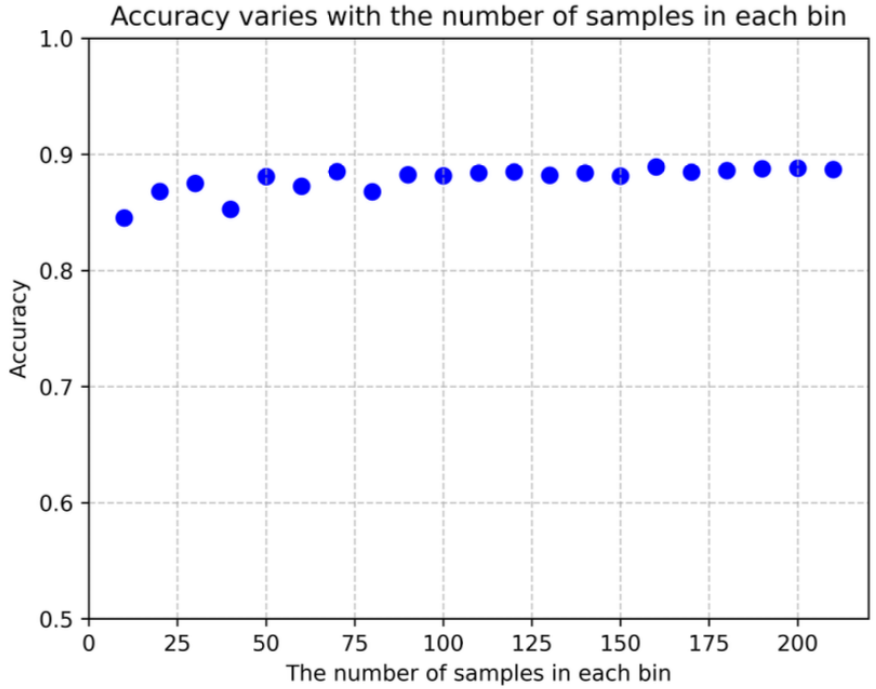
BESIII探测器



使用BESIII 传统PID算法的μ子鉴别效率

数据样本

- 训练样本
  - μ、π单粒子MC样本
  - 0.1 GeV/c < p < 2.0 GeV/c, -0.88 < cosθ < 0.88
  - 将动量和θ谱展平，避免由不同过程的p/θ分布差异造成的偏差
  - 数据控制样本训练结果接近



- 选择控制样本来评估性能和偏差
- 控制样本I:
  - 2021 inclusive MC样本
  - μ/π来自 $\psi(2S) \rightarrow \pi^+\pi^- J/\psi \rightarrow \pi^+\pi^-\mu^+\mu^-$
  - 纯度: 99.13%; 统计量:10w
- 控制样本II:
  - 2019 inclusive MC and 2019 J/ψ 数据样本
  - $J/\psi \rightarrow \gamma\mu^+\mu^-$，$J/\psi \rightarrow \pi^+\pi^-\gamma\gamma$
  - μ子纯度: 97.97%;π介子纯度: 99.37%
  - π介子统计量:6w

特征选择和模型优化

- 为了抑制潜在的系统误差，移除具有较大MC/数据差异的特征
  - 联合来自MDC, TOF, EMC 和 MUC子探测器的信息
  - 递归地消除不太有用的特征和高度相关的特征
  - 108个特征中有37个被保留

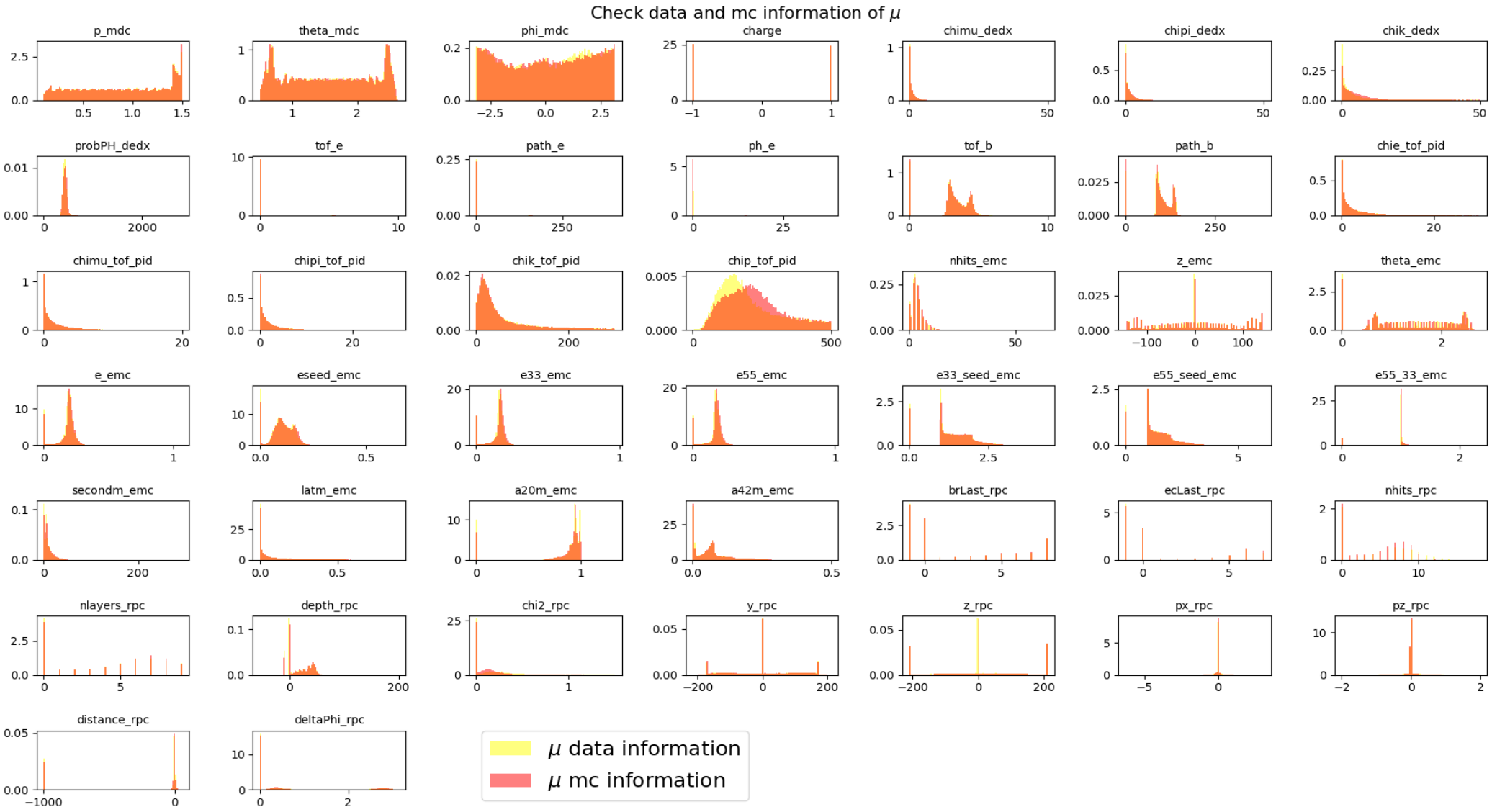
- 研究和测试了各种模型:
  - 基于XGBoost 和 LightGBM 的决策树模型
  - 深度神经网络MLP
  - 支持向量机

特征向量

- 联合来自MDC, TOF, EMC 和 MUC子探测器的信息
- 为了抑制潜在的系统误差，移除具有较大MC/数据差异的特征
- 递归地消除不太有用的特征和高度相关的特征
- 108个特征中保留37个

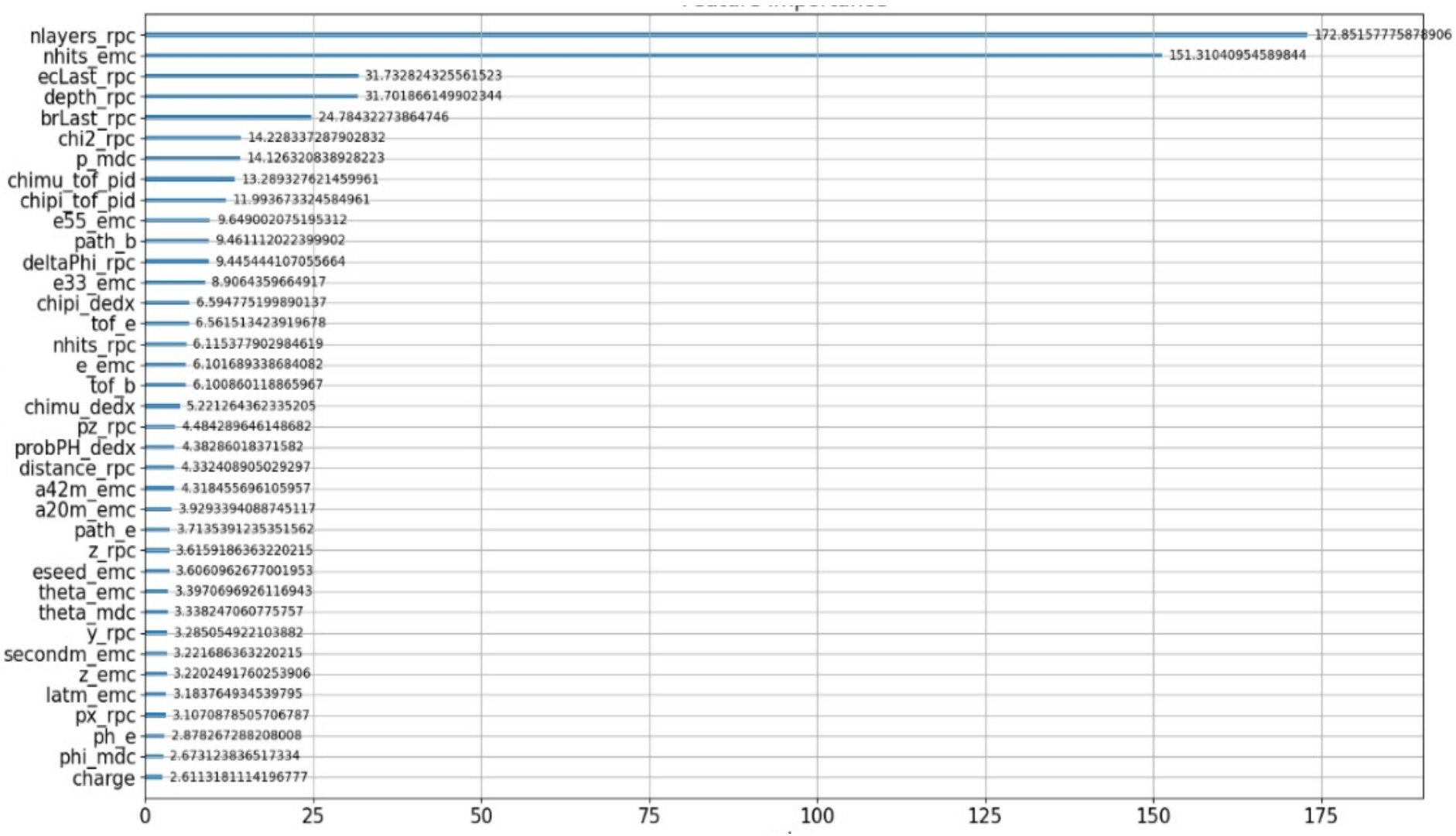
- 37 个特征量：
  - MDC: 'p_mdc', 'theta_mdc', 'phi_mdc', 'charge'
  - dE/dX: 'chimu_dedx','chipi_dedx', 'probPH_dedx',
  - TOF_B: 'tof_b', 'path_b'
  - TOF_E: 'tof_e', 'path_e', 'ph_e'
  - TOF: 'chimu_tof ', 'chipi_tof '
  - EMC: 'nhits_emc','z_emc', 'theta_emc', 'e_emc', 'eseed_emc', 'e33_emc', 'e55_emc', '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'

MC/数据分布对比：μ from $e^+e^- \to J/\psi \to \gamma\mu^+\mu^-$



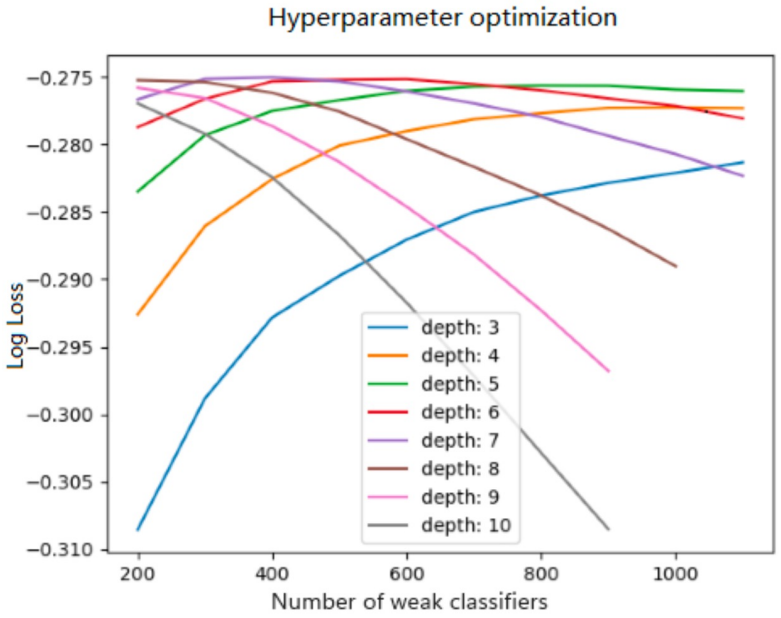
特征向量重要性排序

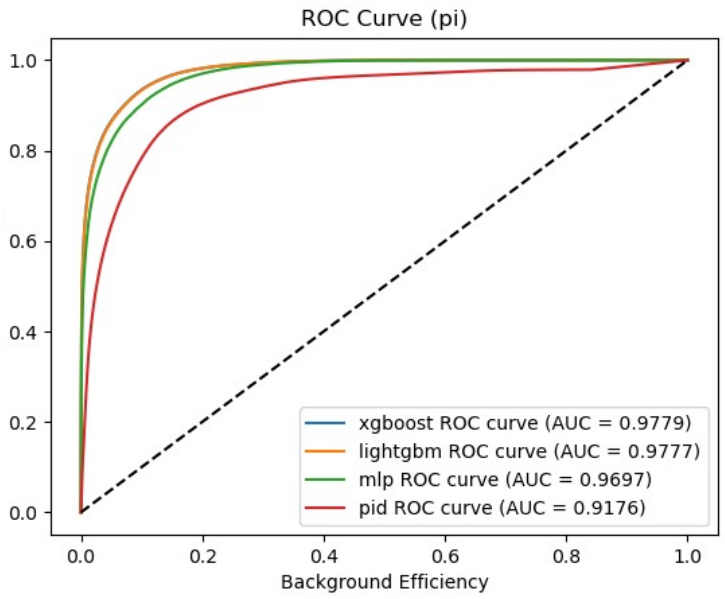
- μ/π鉴别问题中起主要作用的特征量

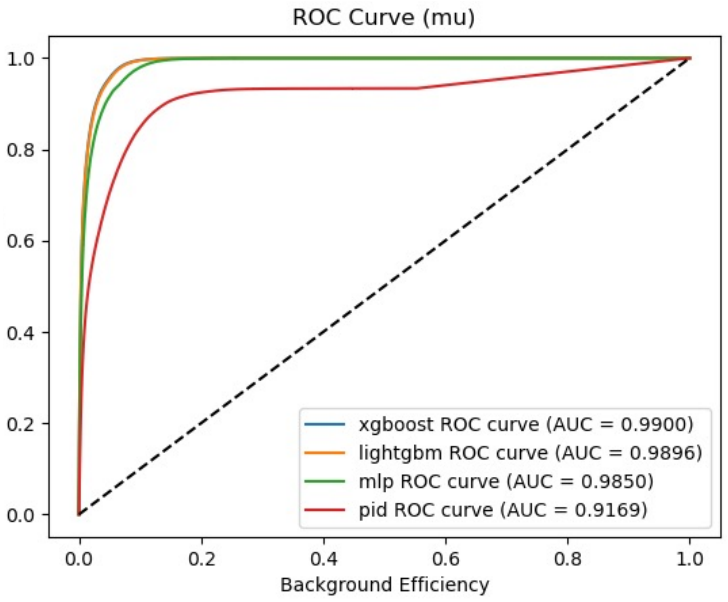


模型选择和优化

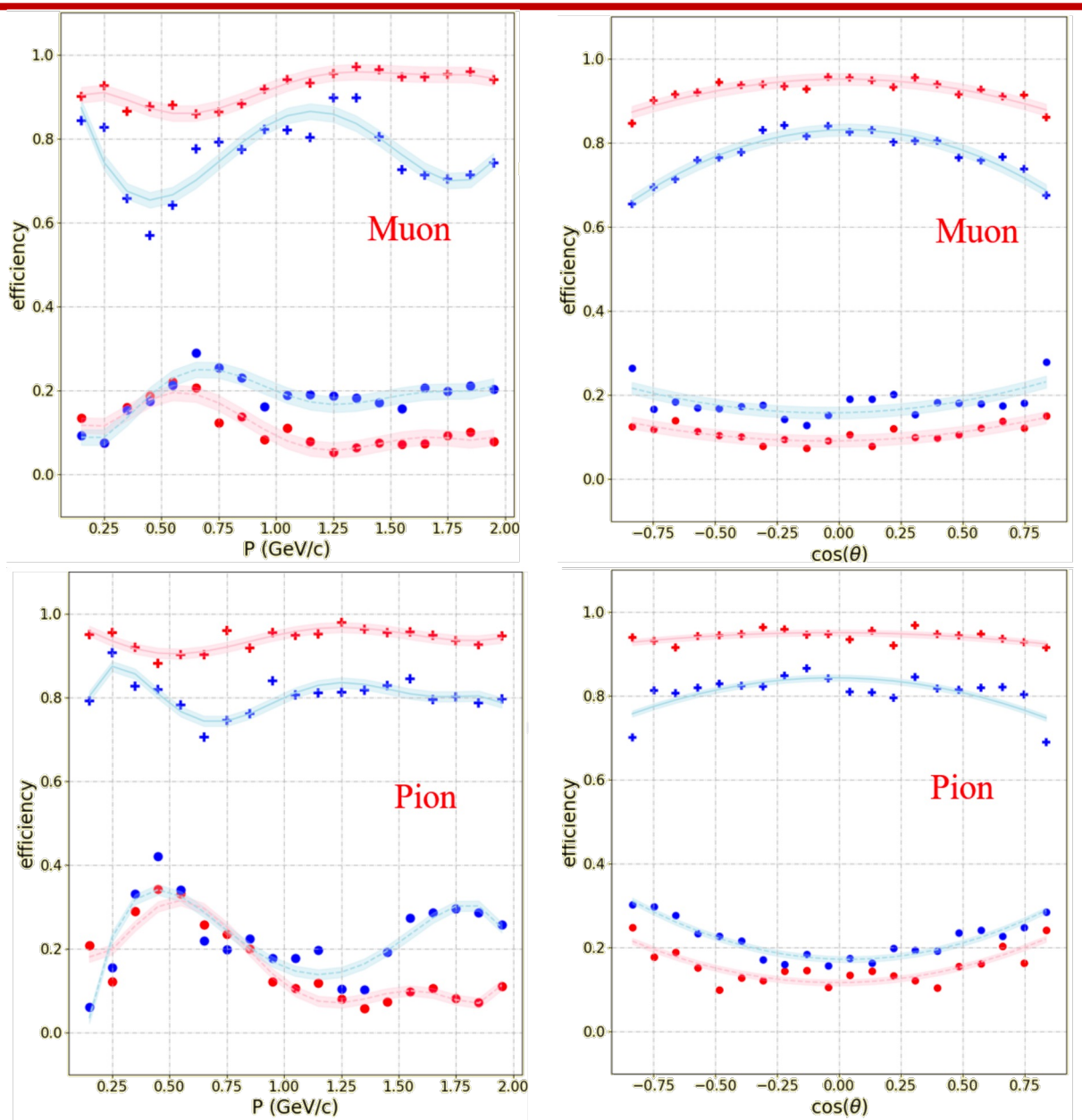
- 基于网格搜索和贝叶斯优化相结合的算法对模型进行优化
  - 深度: 7
  - 分类器个数: 400

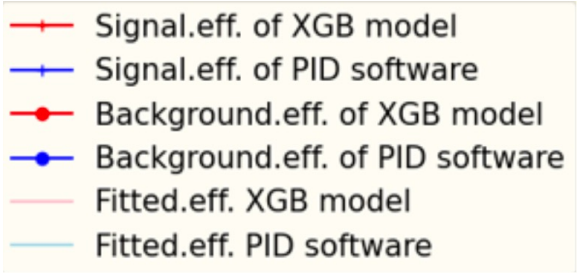






粒子鉴别性能

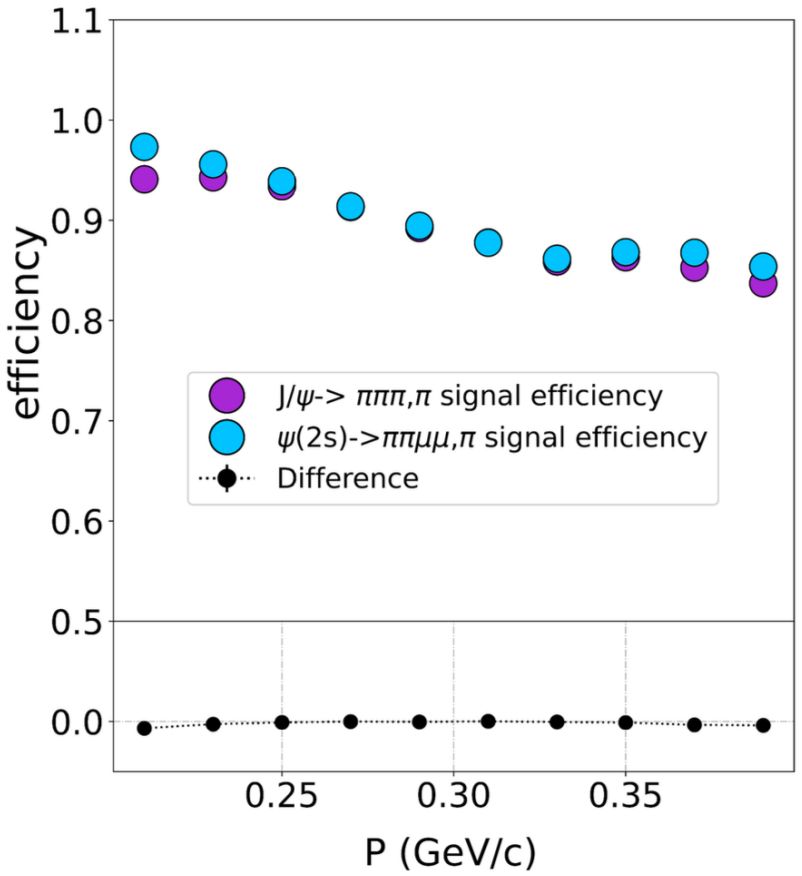


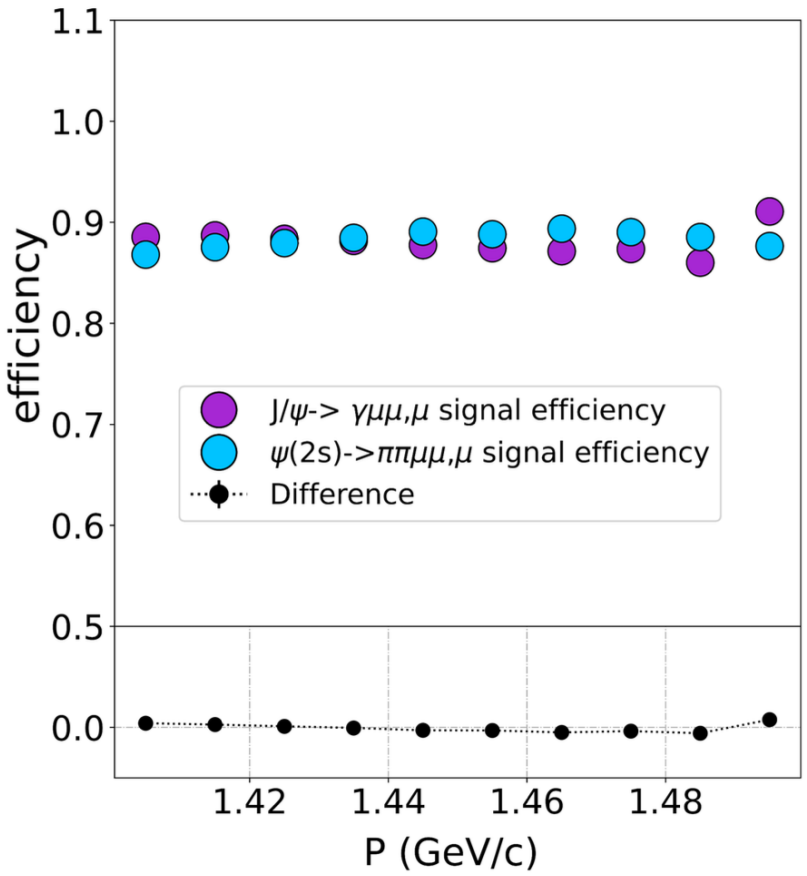


- 信号效率 = 真正选定的信号数/所有信号数
- 本底效率 = 误鉴别为信号的本底数/所有本底数
- 平均μ/π信号效率：XGBoost 模型>90%

交叉验证

- 利用不同衰变道过程进行交叉验证
- 选择效率大致一致，由于p/θ分布的差距而存在一些小差异





物理分析中的应用

Jie Yuan（袁杰） Yadi Wang（王雅迪）

- 数据样本:
  - 2009 J/ψ Inclusive MC, Boss 版本: 7.0.8
- 事例挑选:
  - 选择两条好的带电径迹
  - 传统PID方法 vs XGBoost PID模型
  - 相同的选择标准
    - probπ^+> probμ^+ &&probπ^-> probμ^-

Truth	Decay Process	Number
1	$J/\psi \to \mu^+\mu^-$	924216
2	$J/\psi \to \pi^+\pi^-$	2520
3	$J/\psi \to \pi^+\pi^- f_0(2100)$	134
4	$J/\psi \to K^+K^-$	81
5	$J/\psi \to$ other	83
ALL		927034

	Decay Process	Number
1	$J/\psi \to \mu^+\mu^-$	313403
2	$J/\psi \to \pi^+\pi^-$	2040
3	$J/\psi \to \pi^+\pi^- f_0(2100)$	105
4	$J/\psi \to K^+K^-$	74
5	$J/\psi \to$ other	62
ALL		315684

传统PID方法: 信号效率 :81.03%
本底效率: 33.91%

	Decay Process	Number
1	$J/\psi \to \mu^+\mu^-$	171780
2	$J/\psi \to \pi^+\pi^-$	2232
3	$J/\psi \to \pi^+\pi^- f_0(2100)$	121
4	$J/\psi \to K^+K^-$	69
5	$J/\psi \to$ other	69
ALL		174271

XGBoost PID模型: 信号效率 :88.57%
本底效率: 18.58%

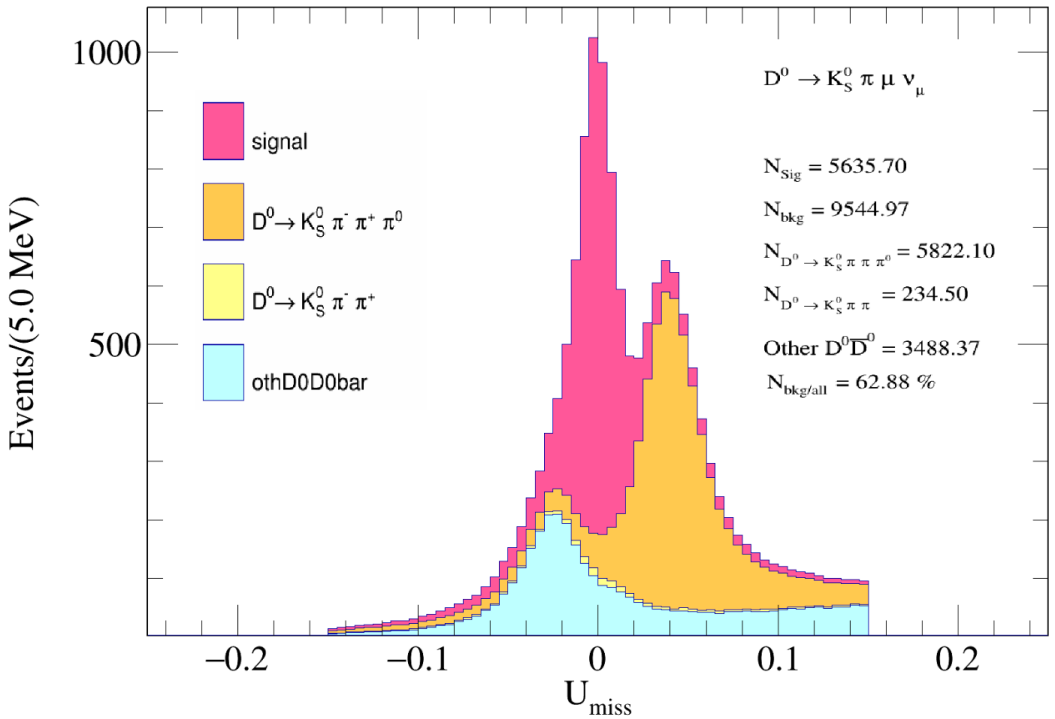
对比传统PID软件包：信号效率提升9.3%，本底效率压低45.20%

$D^0 \to K_S^0 \pi^- \mu^+ \nu_\mu$

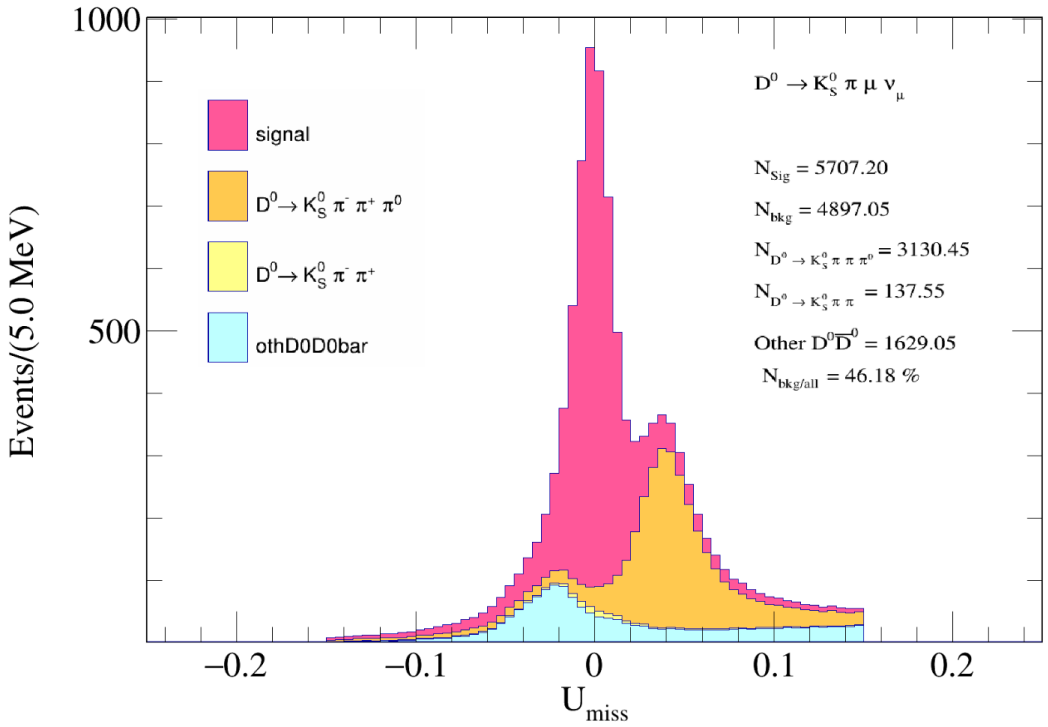
物理分析中的应用

Heng Ma（马衡） Shulei Zhang（张书磊）

◆数据样本: Inclusive MC, Boss 版本: 7.1.0



传统PID方法: 本底效率: 5.91%



XGBoost PID模型:本底效率: 3.35%

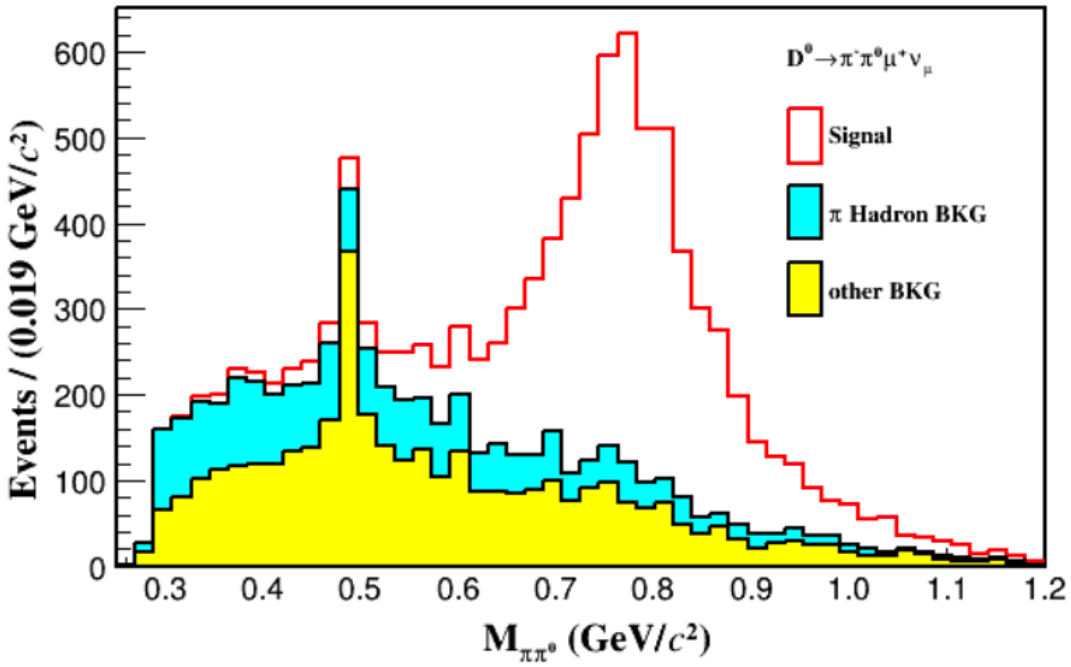
对比传统PID算法：本底效率压低43.3%

$D^0 \to \pi^0 \pi^- \mu^+ \nu_\mu$

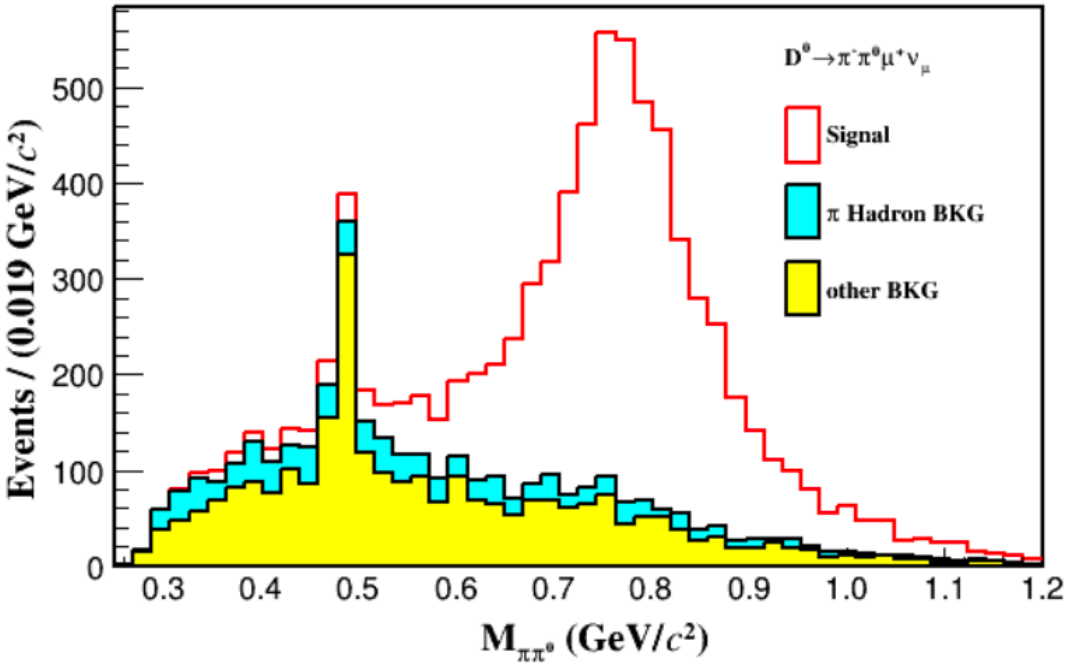
物理分析中的应用

Jialei Shi（石家磊） Shulei Zhang（张书磊）

◆数据样本: Inclusive MC, Boss 版本: 7.1.0



传统PID方法: 本底效率: 1.78%



XGBoost PID模型:本底效率: 0.73%

对比传统PID算法：本底效率压低58.99%

BOSS中 DeepParticleID的使用方法

- 一、环境创建
  - source /workfs2/bes/zhaiyuncong/DeepPID/Setup.sh
- 二、将**DeepParticleID**软件包拷贝到用户自己**workarea**下（以**Boss7.1.2**为例子）
  - /workfs2/bes/zhaiyuncong/workarea/workarea712/DeepParticleID/DeepParticleID-00-00-03/
  - /user/workarea712/DeepParticleID/DeepParticleID-00-00-03
- 三、在运行文件中添加头文件
  - #include "DeepParticleID/DeepParticleID.h"
- 四、创建相关指针并分配内存
  - DeepParticleID*deeppid= new DeepParticleID(XGBoost);
- 五、获取某一条径迹的径迹信息以及各个子探测器信息
  - deeppid->calculate(*itTrk);
- 六、得到该条径迹在**muon/pion**粒子假设下的预测概率 **(muon/pion**二分类问题**)**
  - float prob_mu = deeppid->prob(0);（在Muon假设下的预测概率）
  - float prob_pi = deeppid->prob(1);（在Pion假设下的预测概率）
- 其他：可以设定模型文件路径deeppid->set_path("YOURFILEPATH/xgb.model")

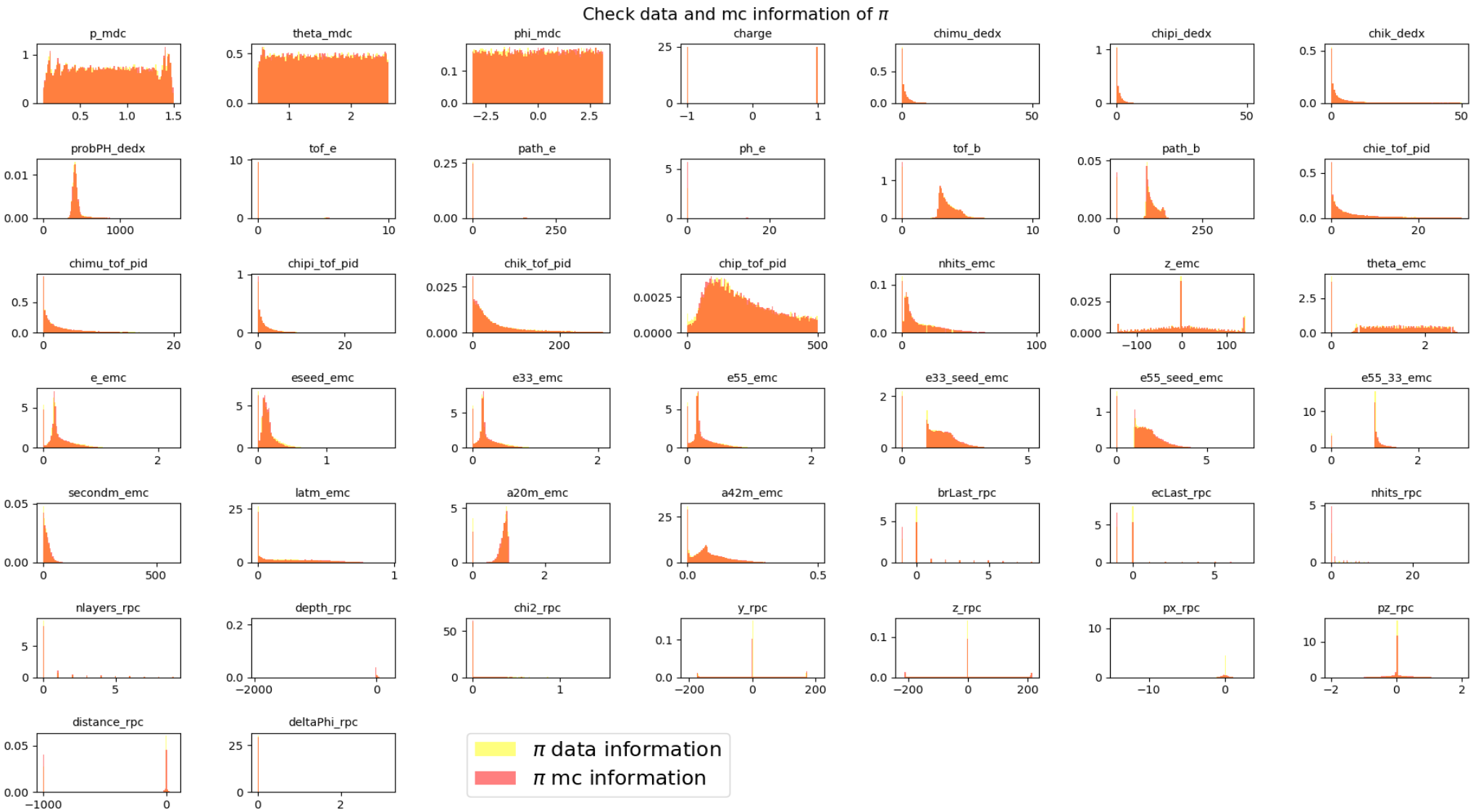
总结

- 基于XGBoost的PID算法在鉴别效率和本底效率上表现优异
- 物理分析中涉及μ、π误判本底的过程会获得较明显的提升
- 欢迎在分析中使用DeepParticleID

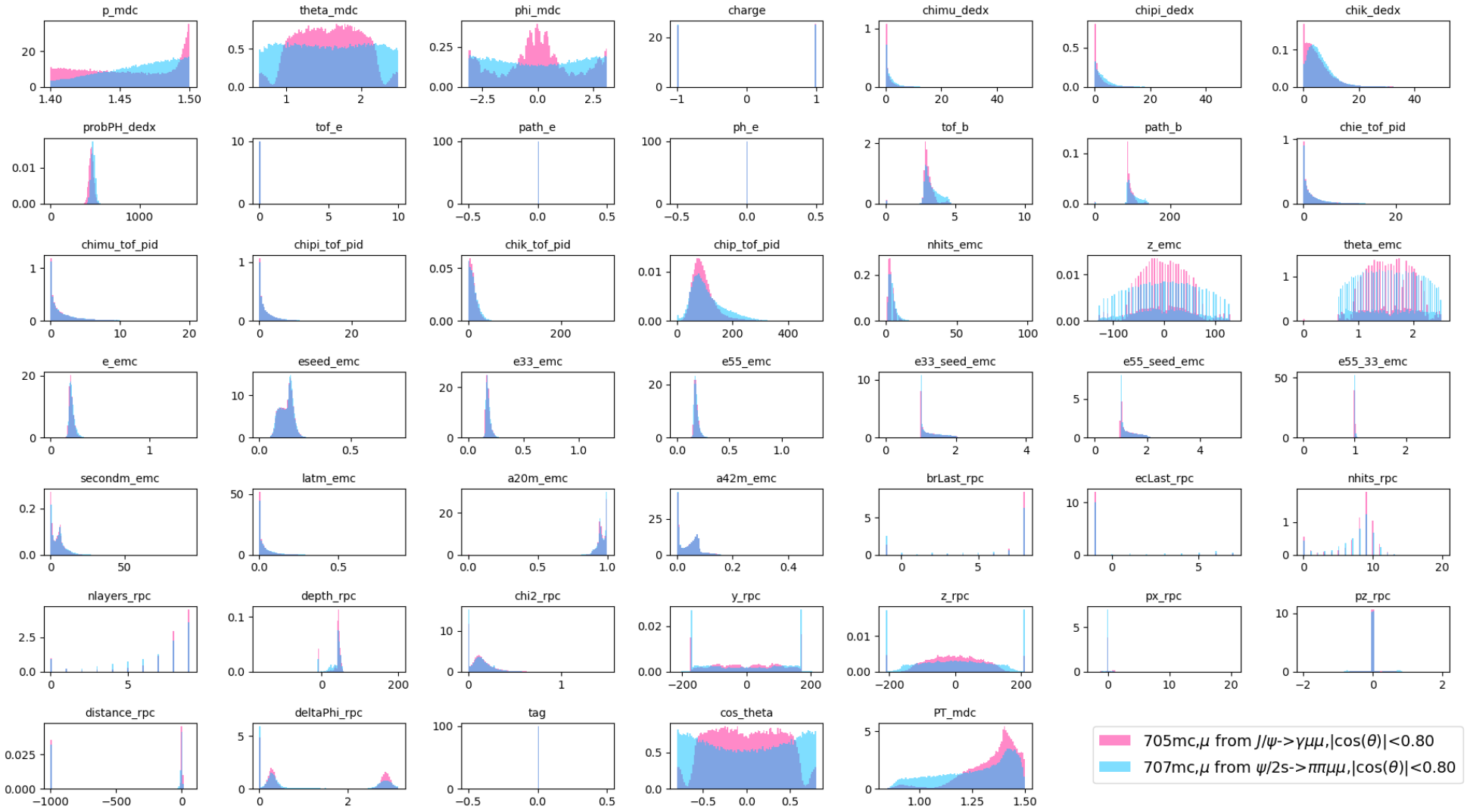
谢谢！

backup

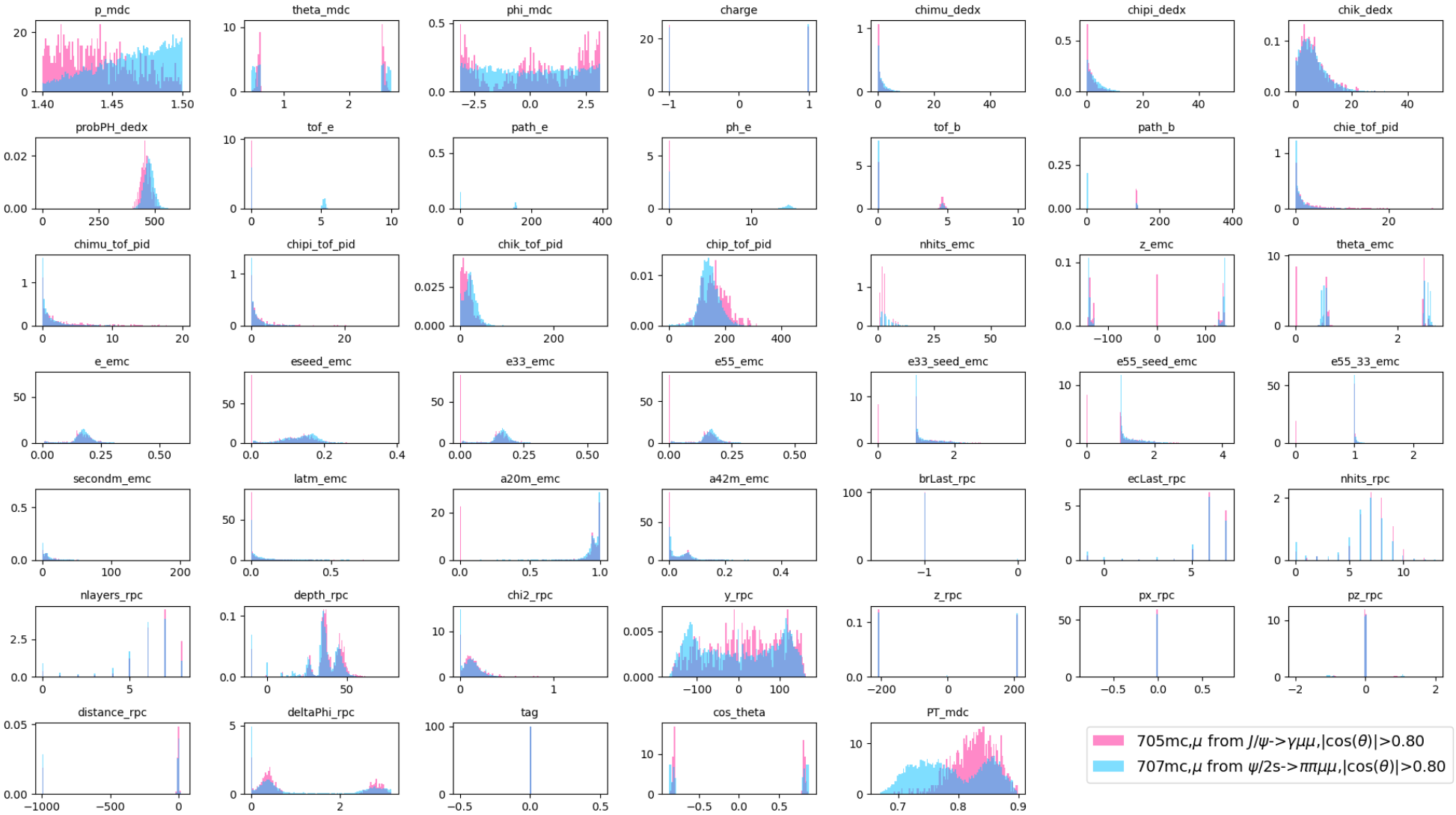
MC/数据分布对比：π from J/ψ → $\pi^+\pi^-\pi^0$ → $\pi^+\pi^-\gamma\gamma$



The comparison plots of MC sample sets distributions for μ from different decay channels at the barrel

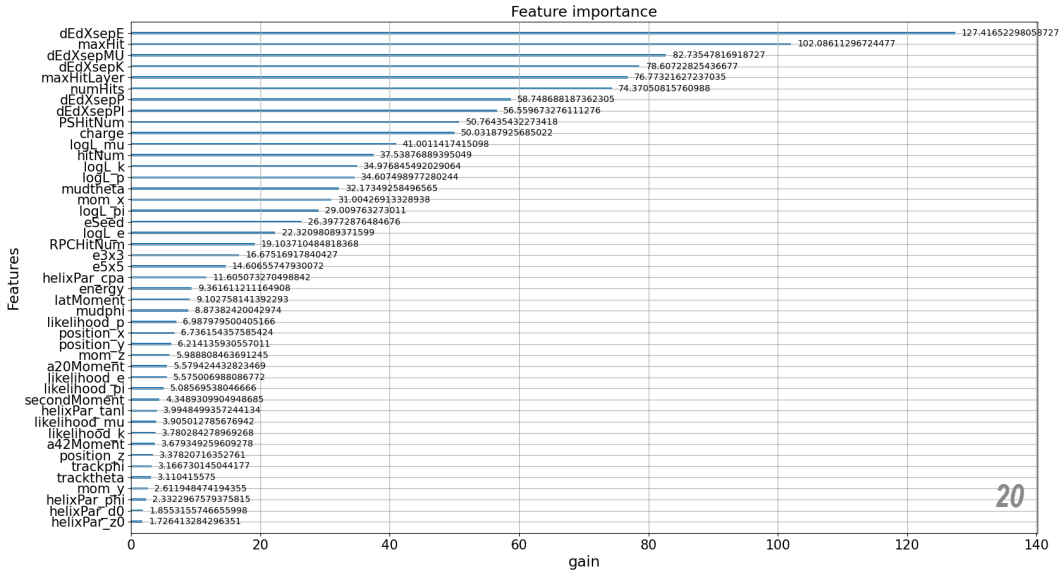


The comparison plots of MC sample sets distributions for μ from different decay channels at the end cap



||----------|特征量信息|----------|说明|

ReconstructinParticle	‘charge’	重建粒子的电荷
	‘mom_x’ ‘mom_y’ ‘mom_z’	粒子在xyz方向上的动量
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’ 'logL_mu' 'logL_pi' 'logL_k' 'logL_p'	粒子分别在五种粒子假设下的可能性