BESIII实验粒子鉴别方法研究

报告人：陈正元

chenzhengyuan@ihep.ac.cn

2023年粒子物理实验计算软件与技术研讨会

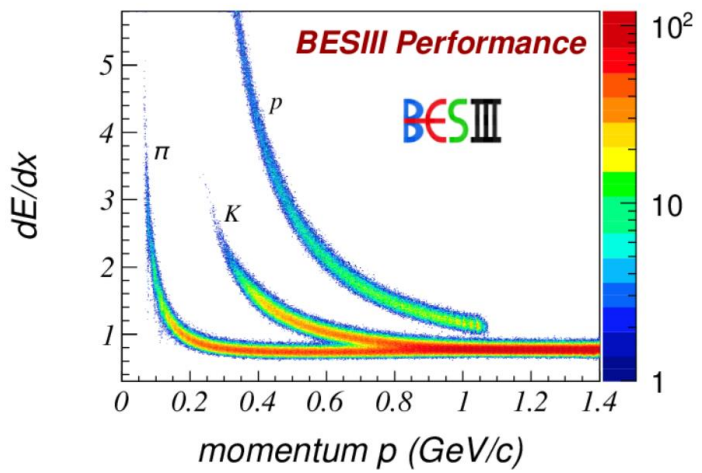
2023-06-10, 山东大学青岛校区

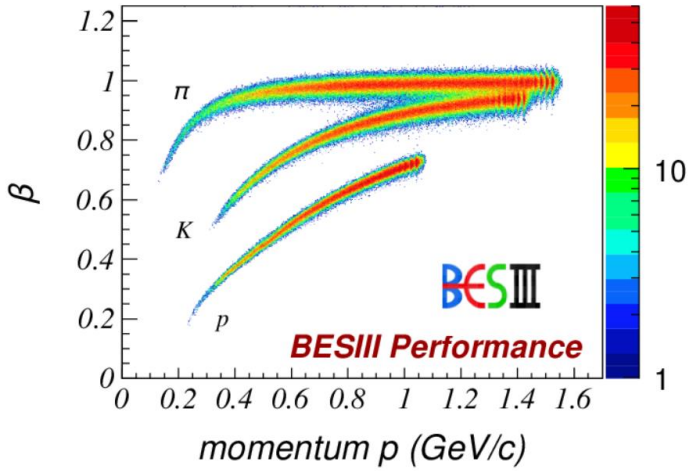
报告大纲

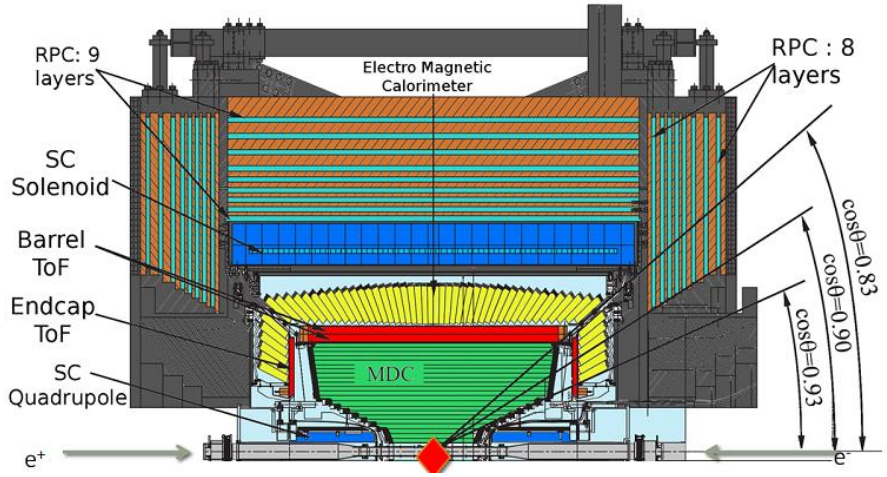
- 简介
- 真实数据和模拟样本
- 机器学习模型训练
- 粒子鉴别效率和系统误差
- 提升决策树部署和物理中的应用
- 总结

简介

- 优秀的粒子鉴别能力是北京谱仪（Beijing Spectrometer, BESIII）实验探测性能的基本要求之一
- BESIII实验的强子鉴别效率在高动量区域（P>1.0GeV/c）较低
- 利用全部子探测器的多个测量信息构造粒子鉴别特征量并进行加权是一个繁琐且复杂的过程
- 机器学习（Machine Learning, ML）方法在处理复杂多变量问题时表现出了强建模能力和自适应性，逐渐成为高能物理领域的热点研究方向之一
- 利用机器学习方法提升BESIII实验的粒子鉴别能力，对最大程度发挥探测器性能有重要意义



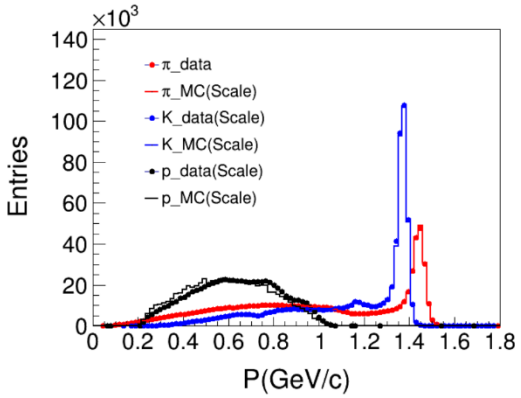


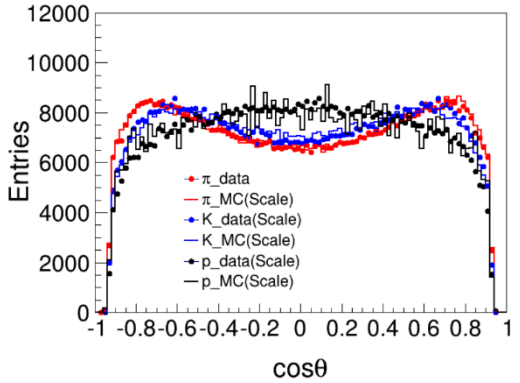


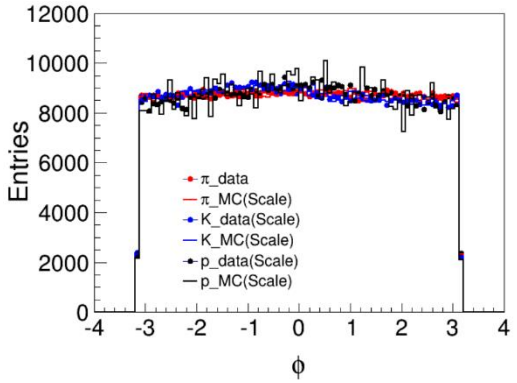
真实数据和模拟样本

- 数据样本质量：高统计量，高纯度，大的动量和立体角覆盖范围
- 获取方式：J/ψ真实数据和模拟样本(round11, round12)，物理过程
  - BOSS版本7.0.8
  - 衰变链
    1. $J/\psi \to \pi^+\pi^-\pi^0, \pi^0 \to \gamma\gamma$
    2. $J/\psi \to K_s^0 K^\pm \pi^\mp, K_s^0 \to \pi^+\pi^-$
    3. $J/\psi \to p\bar{p}\pi^+\pi^-$
- 多变量分析工具包（The Toolkit for Multivariate Analysis，TMVA 6.20）

强子样本的动量，cosθ和ϕ的分布







基于inclusive MC和signal MC估计强子样本纯度

rowNo	decay initial-final states	iDcyIFSts	nEtr	nCEtr
1	$J/\psi \dashrightarrow \pi^0\pi^+\pi^-$	0	29013986	29013986
2	$J/\psi \dashrightarrow \pi^0\pi^+\pi^-\gamma^f$	1	800372	29814358
3	$J/\psi \dashrightarrow \pi^0\pi^+\pi^-\gamma$	7	15127	29829485
4	$J/\psi \dashrightarrow \pi^+\pi^-\gamma$	3	11244	29840729
5	$J/\psi \dashrightarrow \pi^+\pi^-\gamma^F\gamma$	2	7331	29848060
6	$J/\psi \dashrightarrow \pi^0\pi^+\pi^-\gamma^f\gamma^f$	6	7180	29855240
7	$J/\psi \dashrightarrow \pi^0\pi^0\pi^+\pi^-$	12	5746	29860986
8	$J/\psi \dashrightarrow \pi^+\pi^-\gamma\gamma^f$	10	4372	29865358
9	$J/\psi \dashrightarrow \pi^0\pi^+\pi^-\gamma^F$	11	4326	29869684
10	$J/\psi \dashrightarrow e^+e^-\gamma^f\gamma^f$	5	4218	29873902
11	$J/\psi \dashrightarrow \pi^+\pi^-K^+K^-$	9	3221	29877120
12	$J/\psi \dashrightarrow \pi^+\pi^-K^+K^-$	15	2830	29879950
13	$J/\psi \dashrightarrow e^+e^-\gamma$	4	1560	29881440
14	$J/\psi \dashrightarrow K^+K^-$	14	1264	29882704
15	$J/\psi \dashrightarrow \mu^+\mu^-\gamma^f$	19	667	29883371
16	$J/\psi \dashrightarrow e^+e^-\pi^+\pi^-\gamma^F$	25	391	29883762
17	$J/\psi \dashrightarrow \pi^0\pi^0\pi^-K^+$	13	201	29883963
18	$J/\psi \dashrightarrow \pi^+\pi^-\gamma\gamma$	30	178	29884141
19	$J/\psi \dashrightarrow \pi^0\pi^+\pi^-\gamma\gamma^f$	17	158	29884299
20	$J/\psi \dashrightarrow \pi^0\pi^0\pi^+K^-$	24	155	29884454
rest	$J/\psi \dashrightarrow$ others (84 in total)	—	1249	29885703

$\text{Purity}_\pi > 99.9\%$

rowNo	decay initial-final states	iDcyIFSts	nEtr	nCEtr
1	$J/\psi \dashrightarrow \pi^-K_S^0K^+$	1	2999751	2999751
2	$J/\psi \dashrightarrow \pi^+K_S^0K^-$	0	2982326	5982077
3	$J/\psi \dashrightarrow \pi^-K_S^0K^+\gamma^f$	2	11971	5994048
4	$J/\psi \dashrightarrow \pi^+K_S^0K^-\gamma^f$	3	11814	6005862
5	$J/\psi \dashrightarrow \pi^+\pi^-K^+K^-$	4	1619	6007481
6	$J/\psi \dashrightarrow \pi^-K_S^0K^+\gamma$	5	1558	6009039
7	$J/\psi \dashrightarrow \pi^+\pi^+\pi^-\pi^-$	10	158	6009197
8	$J/\psi \dashrightarrow \pi^0\pi^+\pi^+\pi^-\pi^-$	7	124	6009321
9	$J/\psi \dashrightarrow \pi^+\pi^-K^+K^-$	9	65	6009386
10	$J/\psi \dashrightarrow \pi^+\pi^+\pi^-K^-\gamma^F$	17	39	6009425
rest	$J/\psi \dashrightarrow$ others (29 in total)	—	221	6009646

$\text{Purity}_K > 99.4\%$

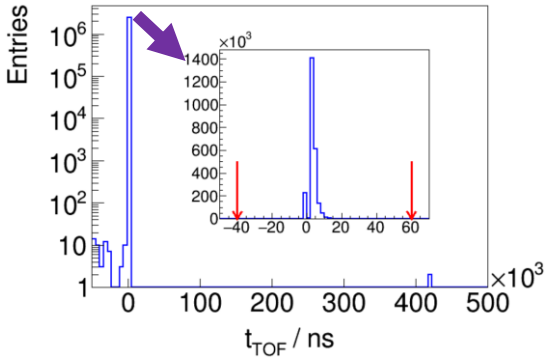
rowNo	decay initial-final states	iDcyIFSts	nEtr	nCEtr
1	$J/\psi \dashrightarrow \pi^+\pi^-p\bar{p}$	0	19712045	19712045
2	$J/\psi \dashrightarrow \pi^+\pi^-p\bar{p}\gamma^f$	1	24597	19736642
3	$J/\psi \dashrightarrow \pi^+\pi^-p\bar{p}\gamma^F$	3	1287	19737929
4	$J/\psi \dashrightarrow \pi^+\pi^-p\bar{p}\gamma$	2	1050	19738929
5	$J/\psi \dashrightarrow \pi^+\pi^-p\bar{p}\gamma$	5	43	19738972
6	$J/\psi \dashrightarrow e^+e^-p\bar{p}\gamma^F$	4	26	19738998
7	$J/\psi \dashrightarrow \pi^0p\bar{p}$	6	24	19739022
8	$J/\psi \dashrightarrow \mu^+\mu^-p\bar{p}\gamma^F$	11	15	19739037
9	$J/\psi \dashrightarrow \pi^+\pi^-p\bar{p}\gamma^f\gamma^f$	14	10	19739047
10	$J/\psi \dashrightarrow e^+e^-p\bar{p}\gamma^F\gamma^f$	30	7	19739054
rest	$J/\psi \dashrightarrow$ others (37 in total)	—	51	19739105

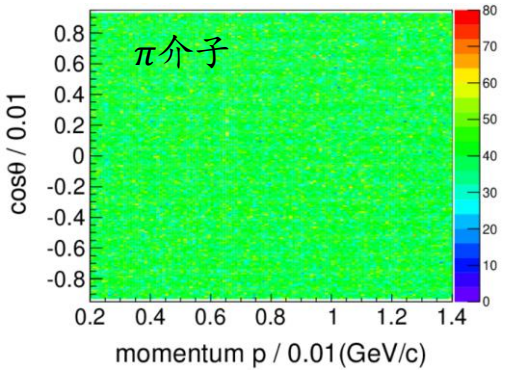
$\text{Purity}_p > 99.6\%$

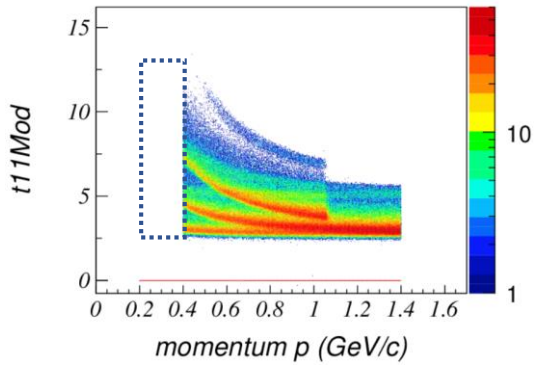
机器学习模型训练：数据预处理

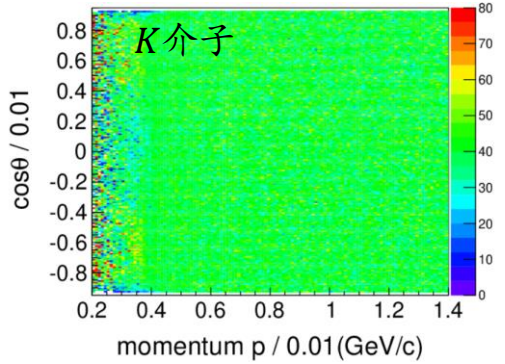
- 提升数据质量，方便机器学习算法“成功学习”
- 数据预处理内容
  1. 数据清洗
     - 移除错误样本，保证数据的一致性
     - 将丢失测量值设为0
     - 把动量低于0.4GeV/c粒子的TOF，EMC和MUC的测量信息设为0
  2. 数据转换
     - 将动量和cosθ分布转换成均匀分布
     - 数据归一化
  3. 类别平衡
     - 使三种强子样本的训练数目大致相当
- 数据集的划分

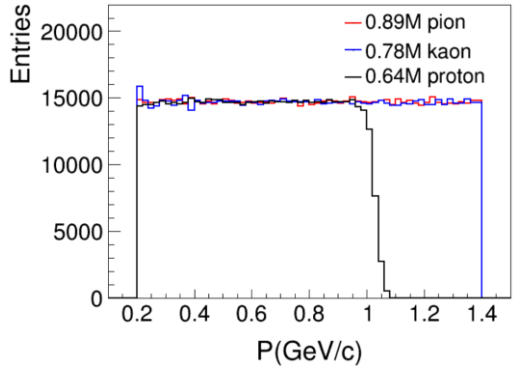
数据预处理后的动量和cosθ分布

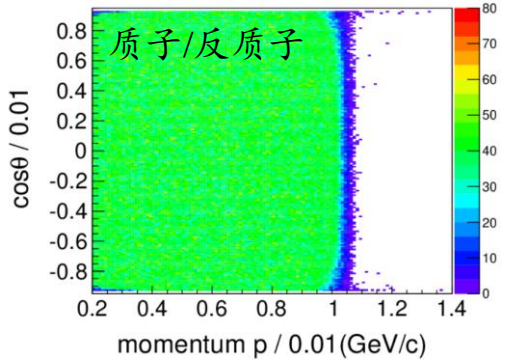




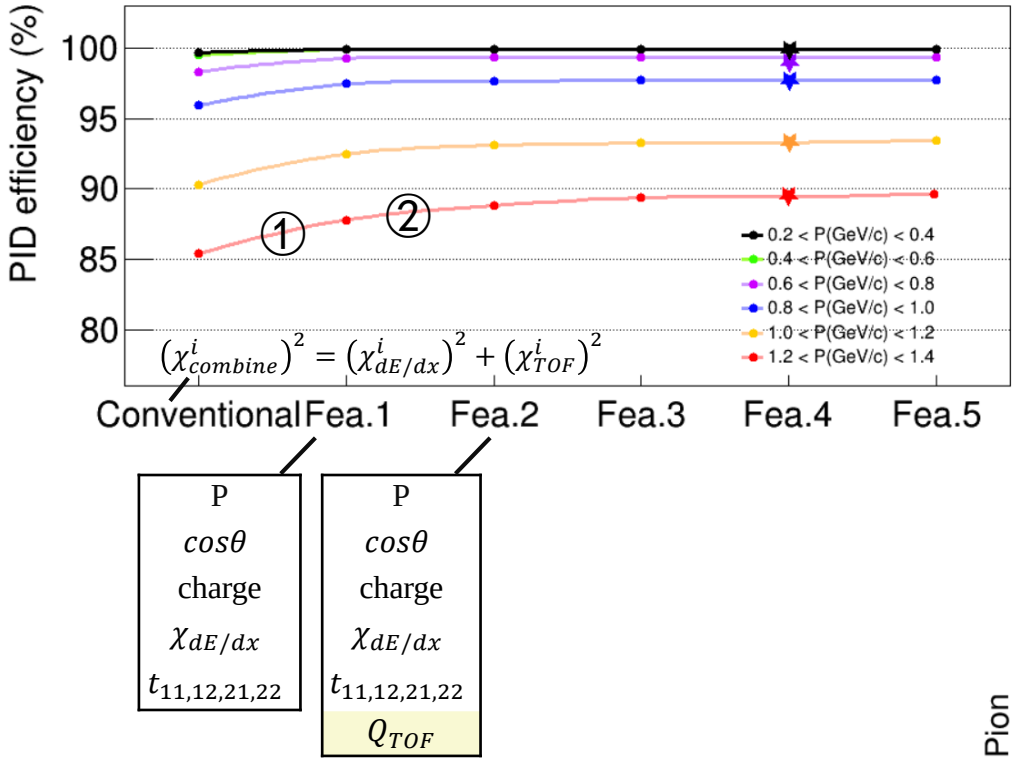








机器学习模型训练：特征挑选



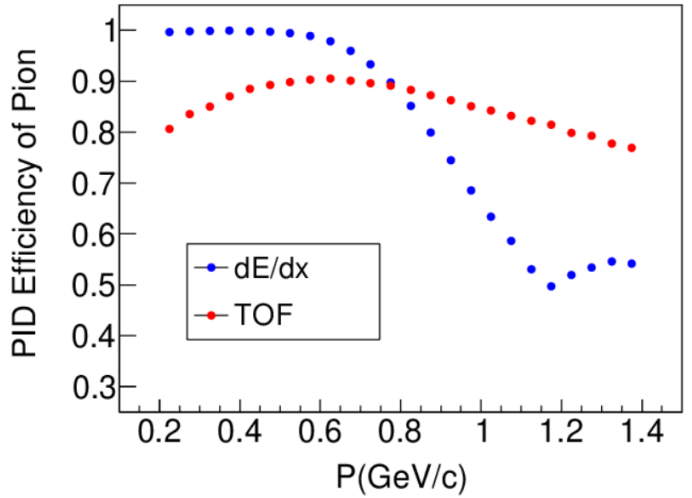
- 从探测器的众多测量信息中挑选一组包含粒子鉴别信息最多的变量集合用于模型的构建
- 综合使用三种常用方法：过滤法，包裹法和嵌入法
- 基于探测器原理
- 鲁棒性较好的提升决策树作为挑选算法
- 挑选标准：粒子鉴别效率，特征之间的相关性和重要性

粒子鉴别效率 $= \frac{n}{N}$

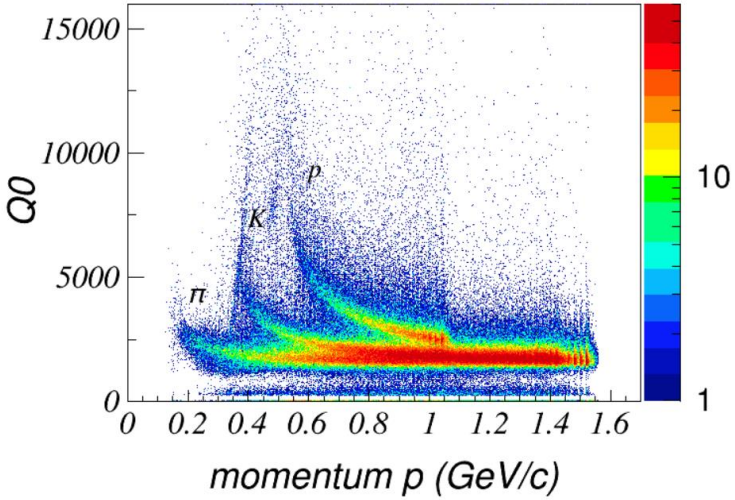
n: 被鉴别正确的强子数目

N: π介子，K介子，质子和反质子的总数目

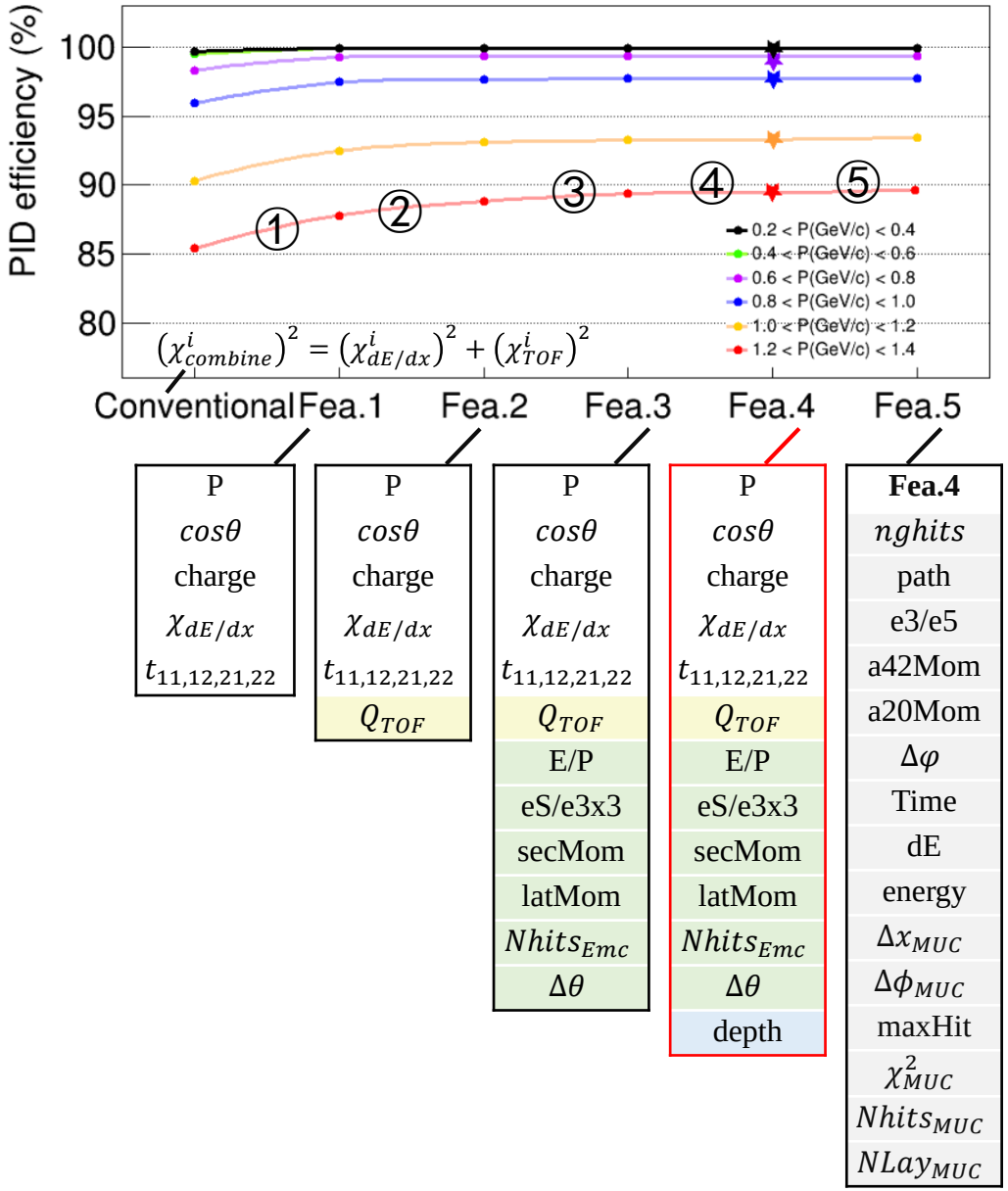
① dE/dx 和 TOF 权重



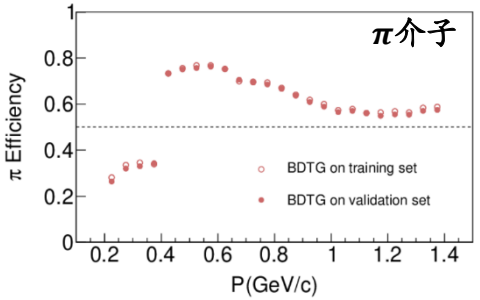
② TOF的脉冲幅度信息随动量的变化

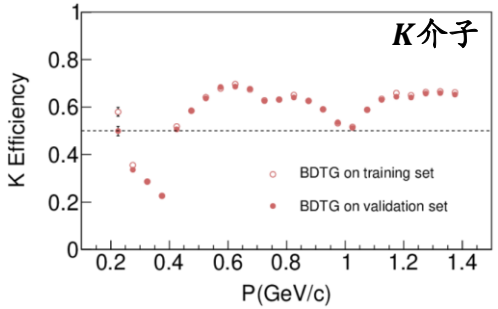


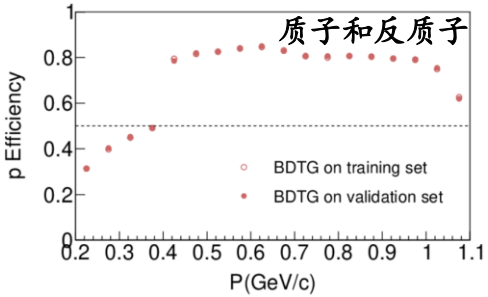
机器学习模型训练：特征挑选



③ 电磁量能器的测量信息对强子鉴别有帮助







④ 缪子探测器的穿透深度信息

⑤ 更多测量信息的加入对模型强子鉴别效率的提升没有帮助

➢ 18个终选特征

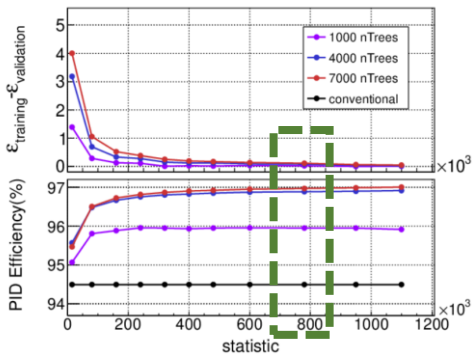
- 动量
- $cos\theta$
- 电荷
- $\chi^{\pi}_{dE/dx}$
- $\chi^{K}_{dE/dx}$
- $\chi^{p}_{dE/dx}$
- t_{11}
- t_{12}
- t_{21}
- t_{22}
- Q_0
- E/p
- eSeed/e3x3
- 横距
- 二次矩
- $Nhits_{EMC}$
- $\Delta\theta$
- 穿透深度

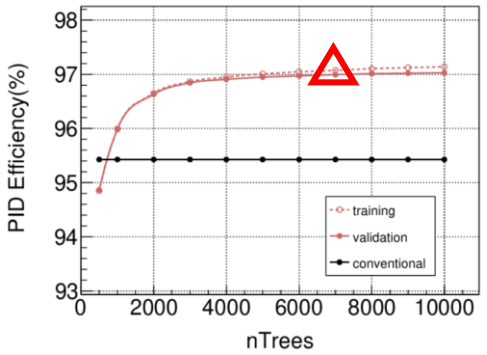
机器学习模型训练

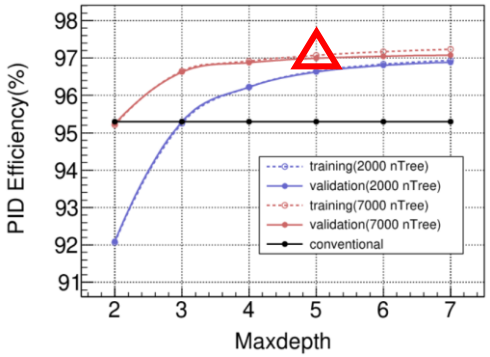
- 目标：高强子鉴别性能和低过拟合程度
- 关键：充足的统计量，合适的机器学习算法以及优化的超参数

提升决策树

统计量充足

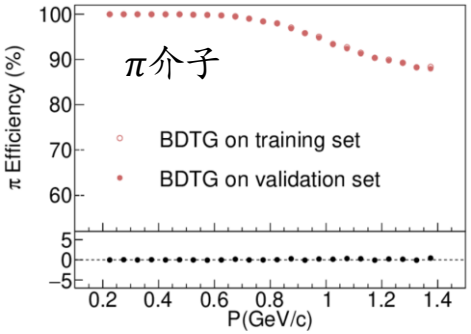


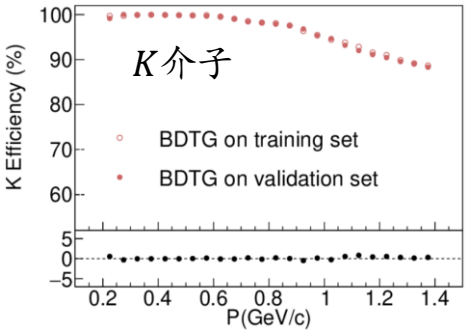




- nTrees = 7000
- Maxdepth = 5
- Shrinkage = 0.01
- UseBaggedGrad = True
- BaggedSampleFraction = 0.5
- SeparationType = 基尼指数
- ncuts = 50
- MinNodeSize = 1%
- NodePurityLimit = 0.5

无明显过拟合现象

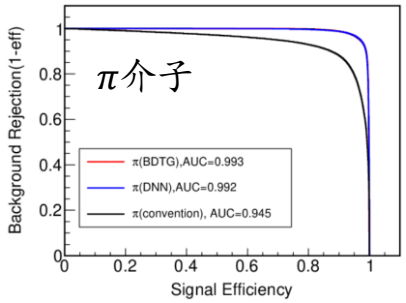


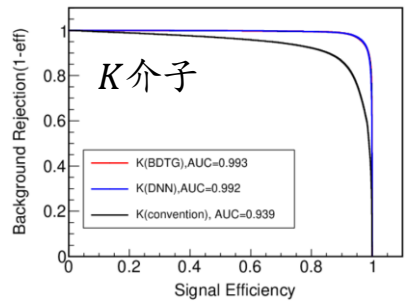


深度神经网络

- 网络架构
  - 隐藏层：64(Relu)，256(Relu)，256(Relu)，128(Relu)
  - 输出层：sigmoid函数$f(x)=\frac{1}{1+\exp(-x)}$
- 损失函数：$\mathrm{L}(Y,f(x))=\sum_{i=0}^{N}(Y-f(x))^2$
- 训练策略
  - 第一阶段：学习率=0.1, batchSize=64, 收敛=20
  - 第二阶段：学习率= 10^{-4}, batchSize=128, 收敛=50
  - 第三阶段：学习率=10^{-6} , batchSize=128, 收敛=10

受试者工作特征（ROC)曲线

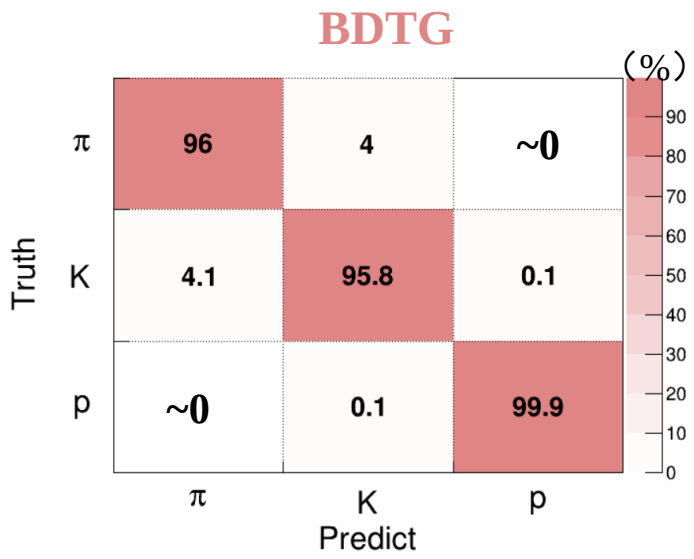


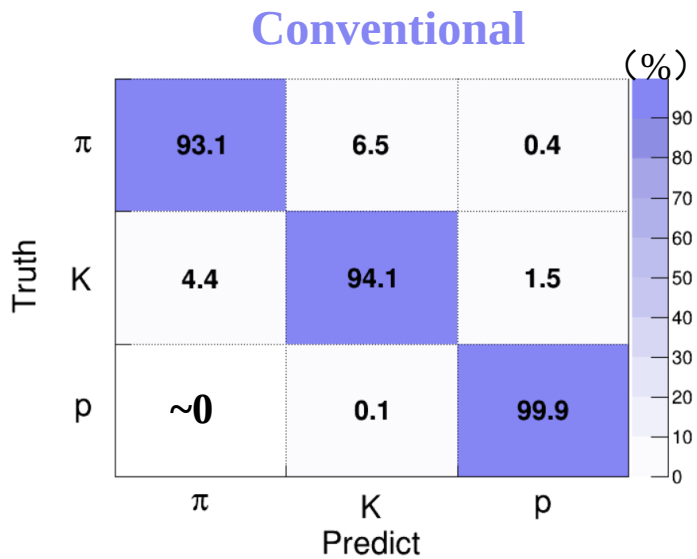


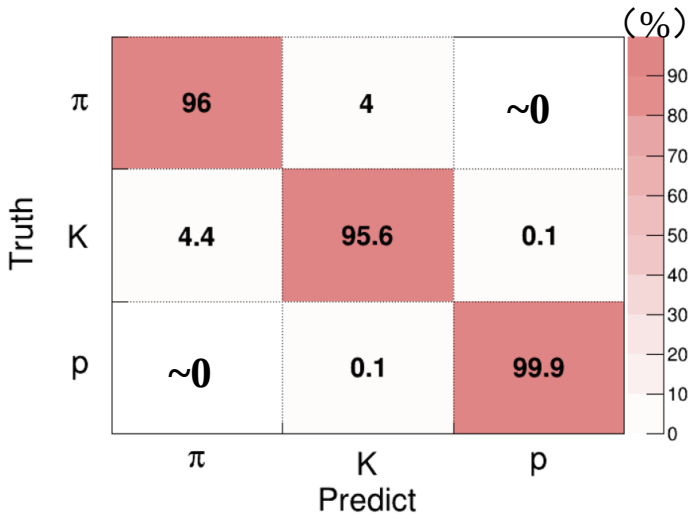
粒子鉴别效率和系统误差

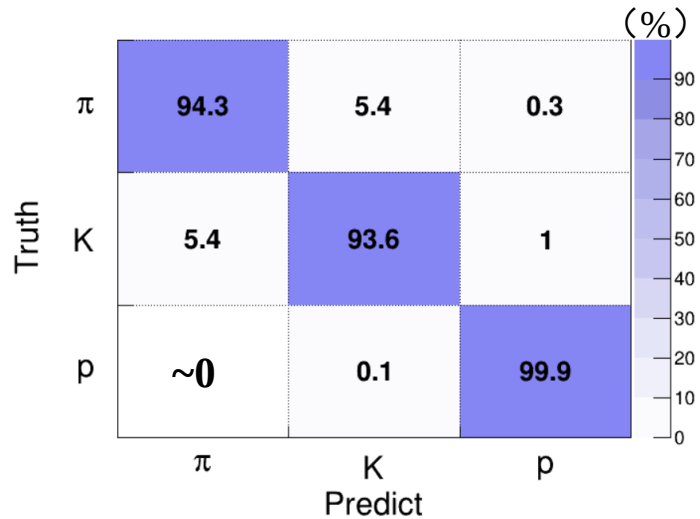
真实数据

模拟样本







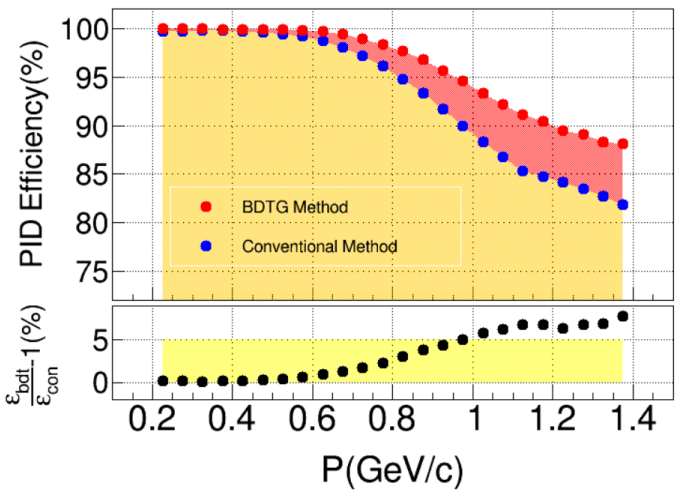


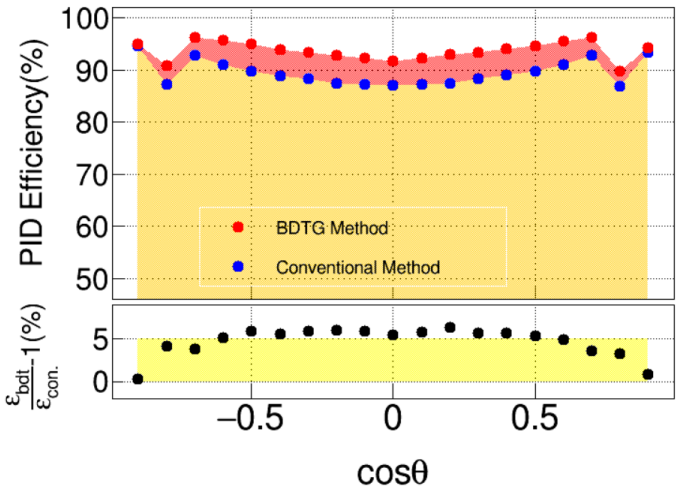
- 提升决策树模型
- 真实数据和模拟样本分开训练
- 混淆矩阵
- 对于真实数据和模拟样本，均有：
  - 粒子鉴别效率提高
  - 误判率降低

粒子鉴别效率和系统误差

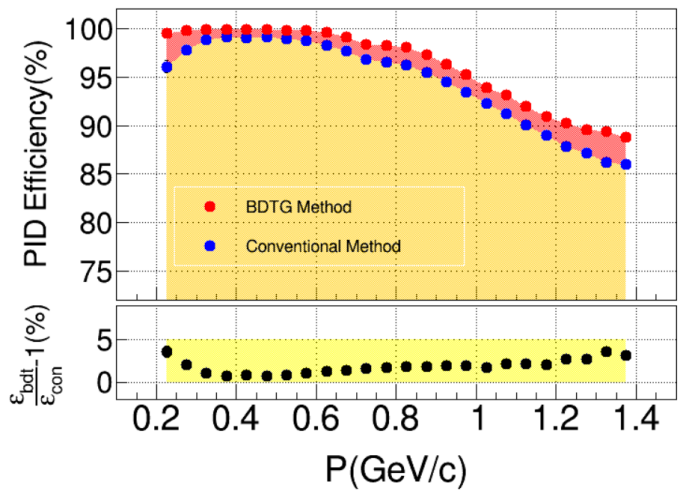
■ 测试集上粒子鉴别效率随动量和cosθ的变化

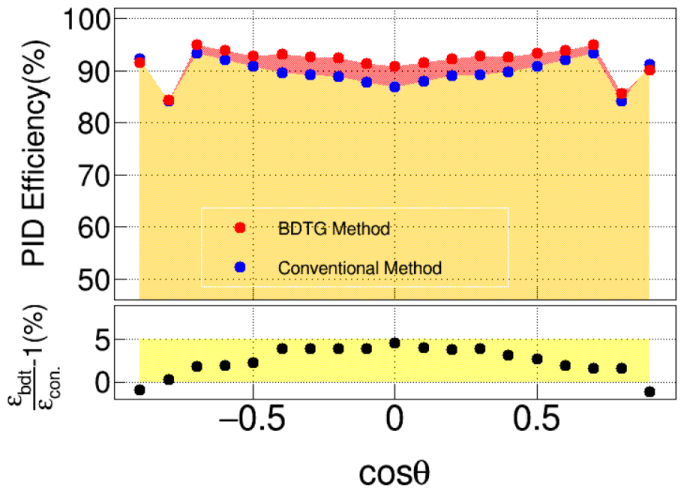
π介子



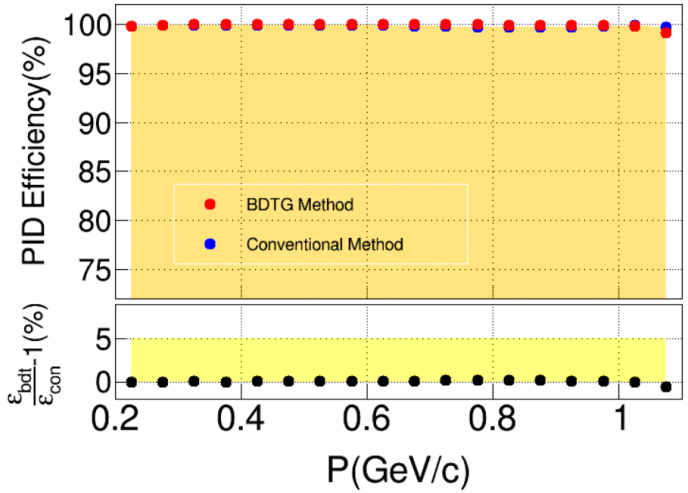


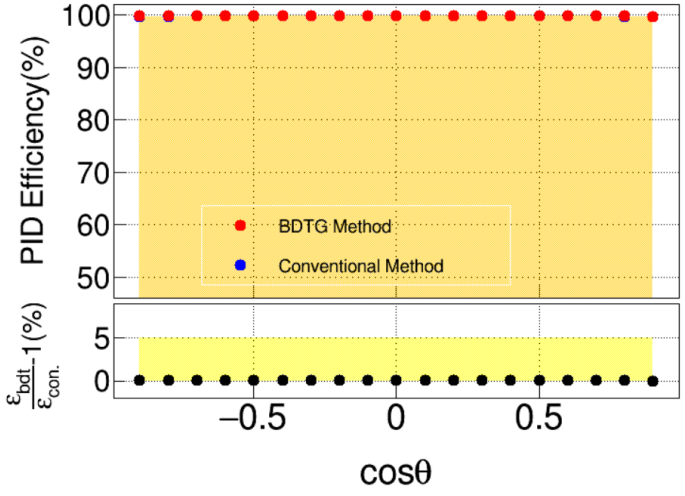
K介子





质子和反质子



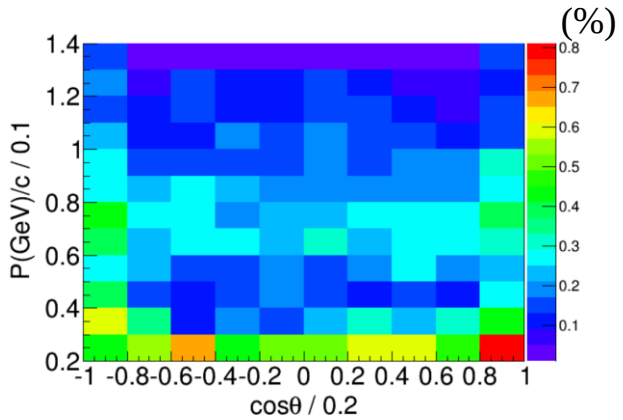


粒子鉴别效率和系统误差

- 模型性能物理过程无关性的验证
- 事例挑选：$J/\psi \to K^+K^-\pi^0$

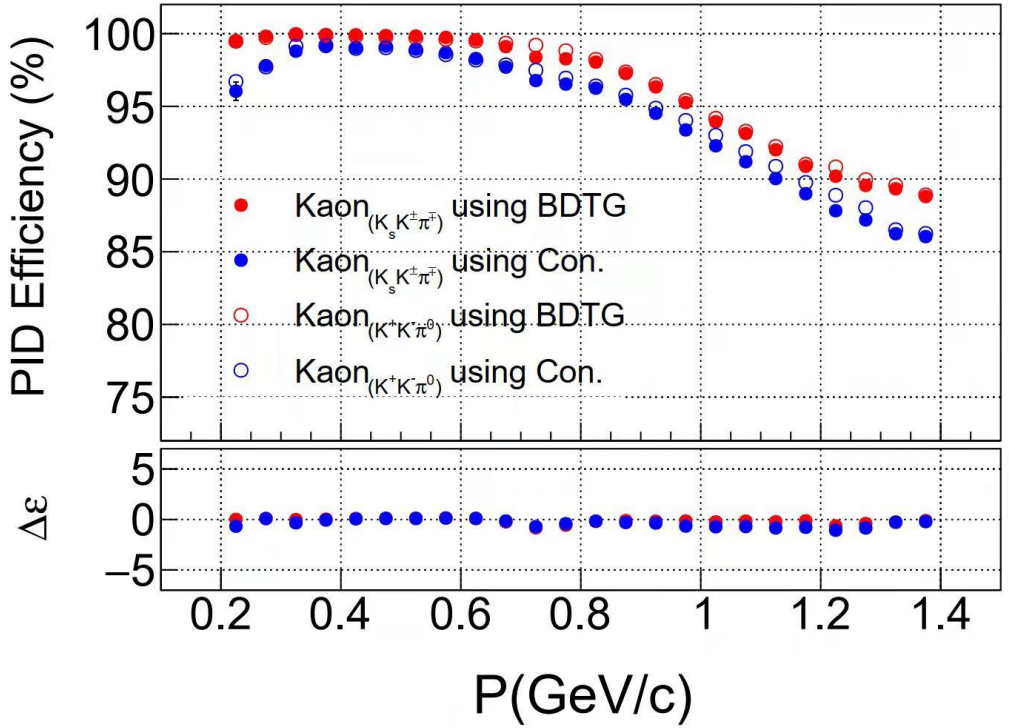
rowNo	decay initial-final states	iDcyIFSts	nEtr	nCEtr
1	J/ψ --> $\pi^0K^+K^-$	0	3580333	3580333
2	J/ψ --> $K^+K^-\gamma$	2	32209	3612542
3	J/ψ --> $\pi^0K^+K^-\gamma^f$	3	12484	3625026
4	J/ψ --> $\pi^0K^+K^-\gamma$	4	5133	3630159
5	J/ψ --> $\pi^0\pi^0\pi^+K^-$	8	3336	3633495
6	J/ψ --> $K^+K^-\gamma\gamma^f$	1	1347	3634842
7	J/ψ --> $\pi^0\pi^0K^+K^-$	6	804	3635646
8	J/ψ --> $K^0_L\pi^+K^-$	14	362	3636008
9	J/ψ --> $\pi^0\pi^+\pi^-$	5	304	3636312
10	J/ψ --> [illegible]	[illegible]	[illegible]	3636562
11	J/ψ --> $K^+K^-\gamma^f$	19	209	3636771
12	J/ψ --> $K^+K^-\gamma^f$	7	199	3636970
13	J/ψ --> $\pi^0\pi^+\pi^-\gamma^f$	17	161	3637131
14	J/ψ --> $e^+e^-K^+K^-\gamma^F$	10	140	3637271
15	J/ψ --> $\pi^0K^0_L\pi^+K^-$	15	134	3637405
16	J/ψ --> $\pi^0\pi^0\pi^+\pi^-$	12	83	3637488
17	J/ψ --> $\pi^0K^+K^-\gamma^F$	24	73	3637561
18	J/ψ --> $K^0_L\pi^+K^-\gamma$	16	58	3637619
19	J/ψ --> $\pi^0\pi^0\pi^0\pi^+\pi^-$	21	44	3637663
20	J/ψ --> $K^+K^-\gamma^f\gamma^f$	28	39	3637702
rest	J/ψ --> others (55 in total)	—	444	3638146

$Purity_K > 99.8\%$



K介子样本的本底率随动量和cosθ的变化

模型的性能与衰变过程没有依赖

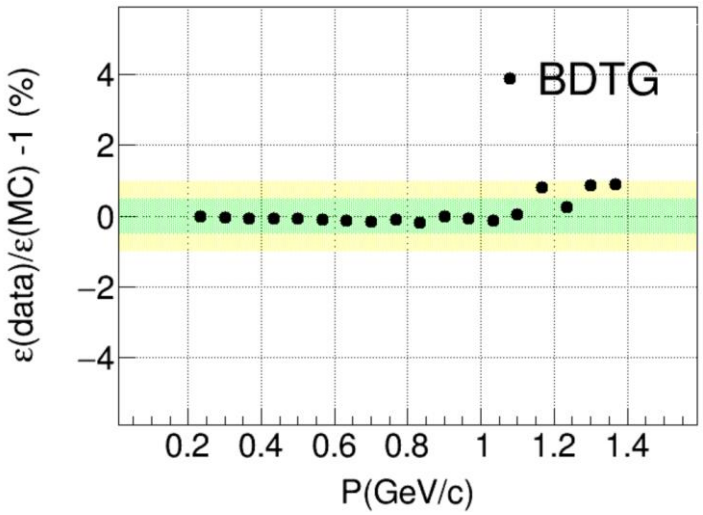


粒子鉴别效率和系统误差

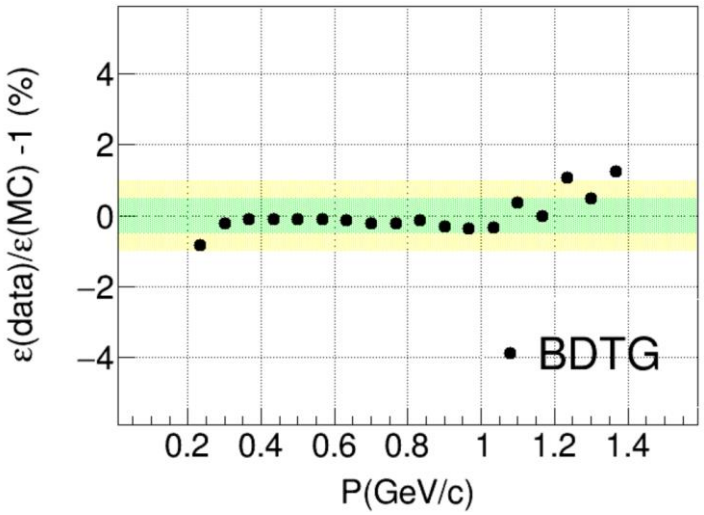
■ 系统误差： $\Delta\varepsilon = \frac{\varepsilon(data)-\varepsilon(MC)}{\varepsilon(MC)}$

* ε：粒子鉴别效率

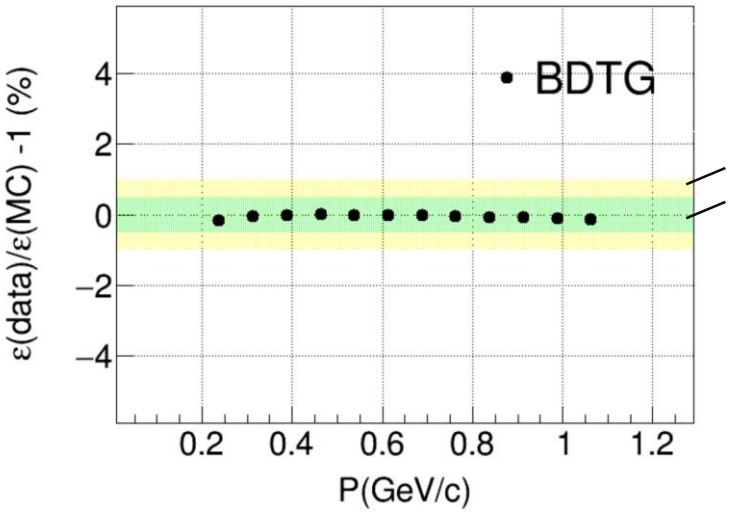
π介子



K介子



质子和反质子



±1.0%

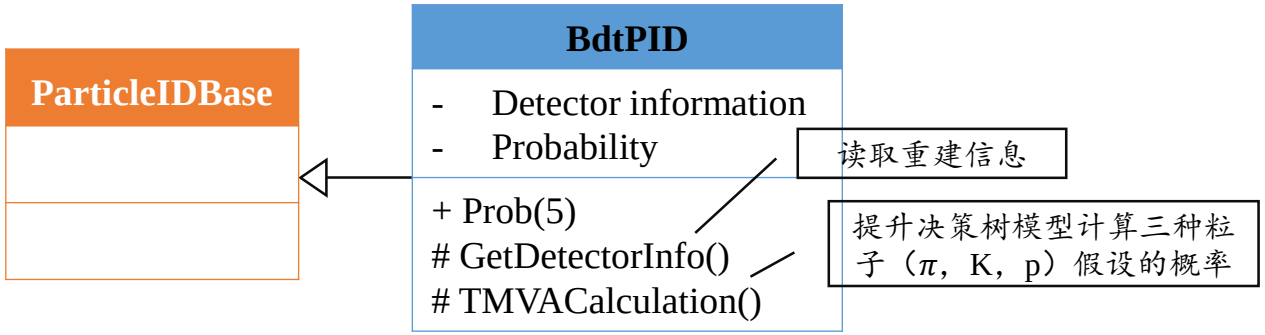
±0.5%

■ 用于粒子鉴别的提升决策树的系统误差：

* π介子和K介子鉴别的系统误差总体小于1%
* 质子鉴别的系统误差总体小于0. 2%

提升决策树部署和物理中的应用

■ BDT模型在BESIII粒子鉴别算法的实现

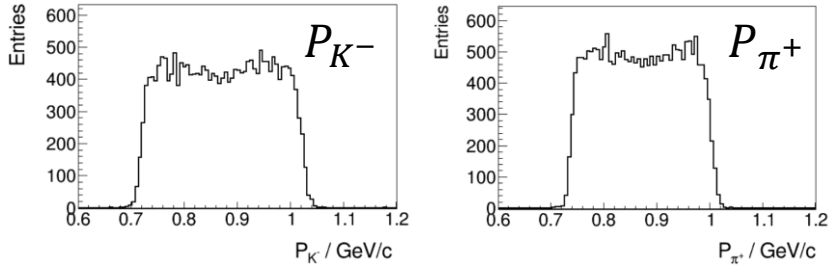


■ BDT模型的使用

```
ParticleID *pid = ParticleID::instance();
for(int i = 0; i < nGood; i++) {
        EvtRecTrackIterator itTrk = evtRecTrkCol->begin() + iGood[i];
        pid->init();
        …
        pid->usePidSys(pid->useDedx() | pid->useTofCorr()); // use PID sub-system
        // pid->usePidSys(pid->useBdt()); // use decision tree model
        pid->identify(pid->onlyPionKaonProton()); // seperater Pion/Kaon/Proton
        …
        pid->calculate();
        …
        if((pid->probPion() < pid->probProton()) || (pid->probPion() < pid->probKaon())) continue;
        …
}
```

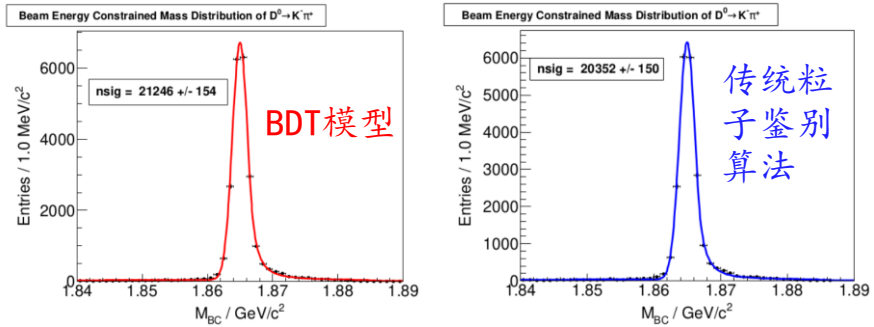
■ $D^0 \to K^-\pi^+$过程的粒子鉴别效率

* 模拟产生了3万个$\psi(3770) \to D^0\overline{D^0}, D^0 \to K^-\pi^+$事例



* 事例挑选
  - 好带电径迹挑选
  - 粒子鉴别（传统粒子鉴别算法，BDT模型）

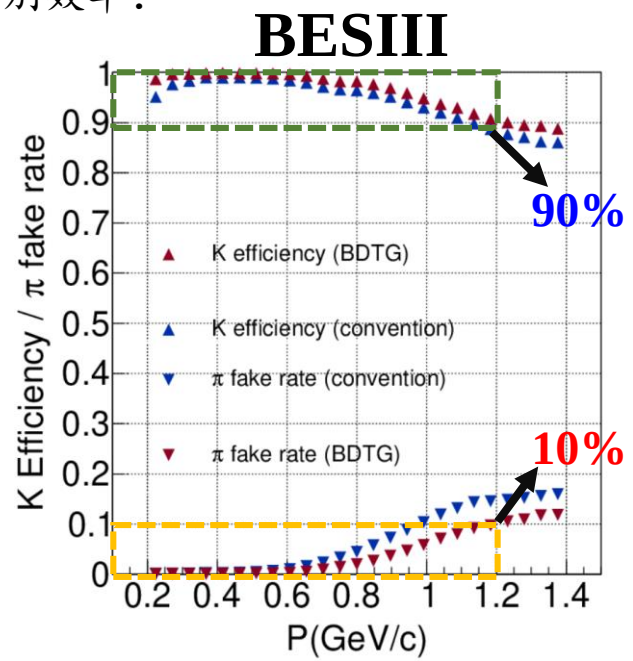
* 束流能量约束下的D^0不变质量分布

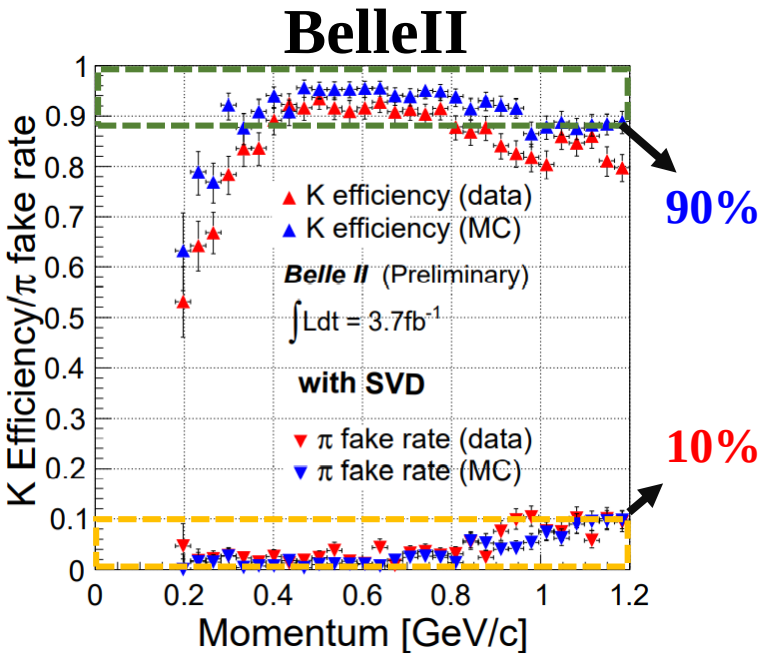


BDT模型将该过程的粒子鉴别效率提升了4.4%

总结

- 利用BESIII实验积累的海量J/ψ真实数据和对应产生的模拟样本，通过物理过程获取了高纯度的强子样本
- 通过机器学习方法高效联合BESIII实验四个子探测器的强子鉴别信息，能够显著提高BESIII实验高动量区域（$P > 1.0GeV/c$）的强子鉴别效率
- 相比于传统的粒子鉴别算法，BDT模型的粒子鉴别效率：
  - π介子提升了~8%@1.4GeV/c
  - K介子提升了~3%@1.4GeV/c
  - 质子/反质子保持约100%的鉴别效率
- 采用真实数据和模拟样本分开训练的方式，可以有效控制BDT模型的系统误差：
  - π介子和K介子总体小于1%
  - 质子/反质子的总体小于0.2%

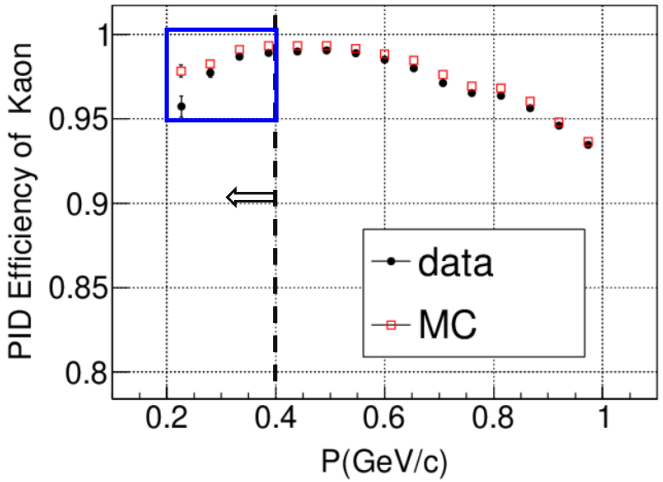


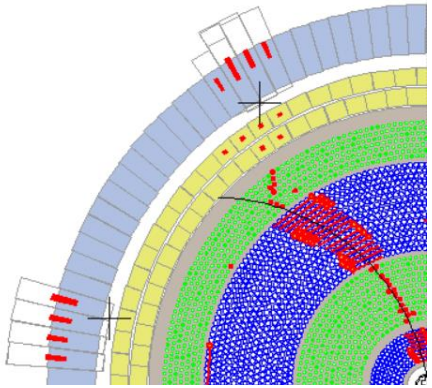


Thanks for your attention

附录：K介子的衰变

- K介子的衰变造成低动量区域鉴别效率偏低



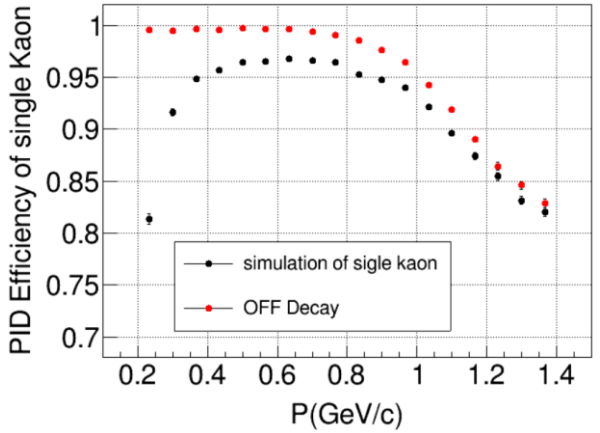


"kink" topology:

$K^{\pm} \longrightarrow \mu^{\pm}\nu_{\mu}$ (64%)

$K^{\pm} \longrightarrow \pi^{\pm}\pi^{0}$ (21%)

…



- 低动量区域的粒子鉴别仅使用dE/dx信息，高动量区域的粒子鉴别联合dE/dx和TOF