CEPC上基于DeepSets模型的喷注鉴别算法研究

报告人：廖立波

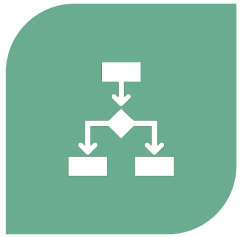
合作人：李刚，宋维民，王书栋，张兆领

2023.7.10

目录

背景介绍

机器学习框架

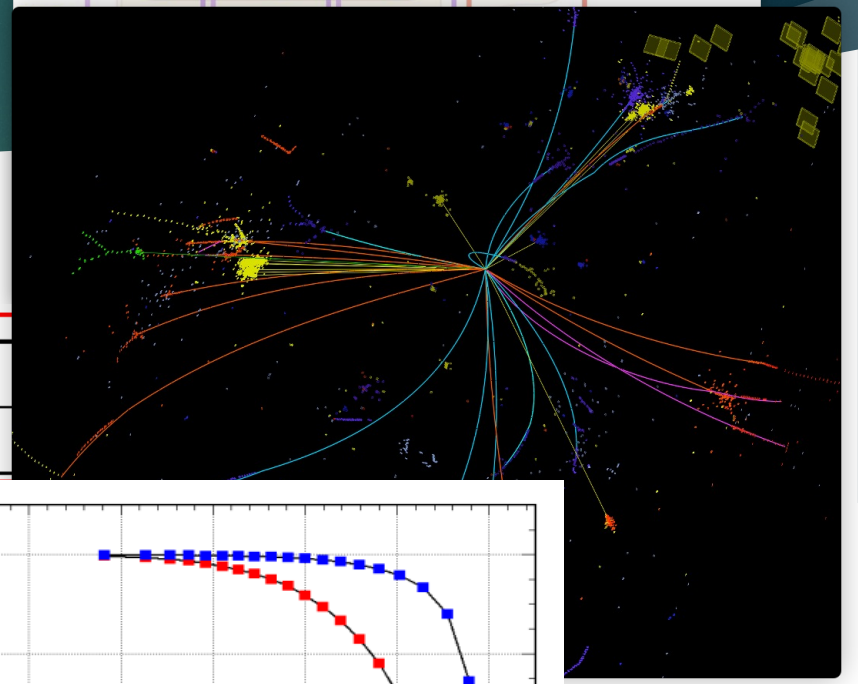


结果

总结和展望

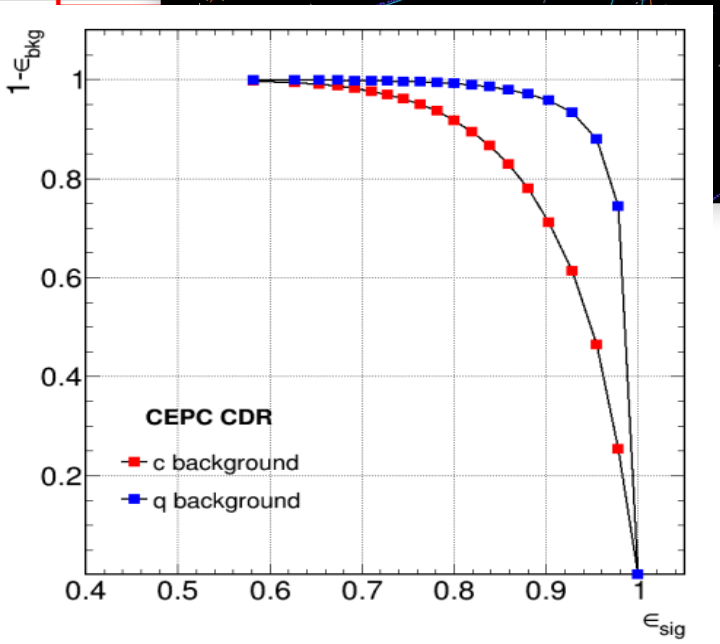
新想法

背景介绍



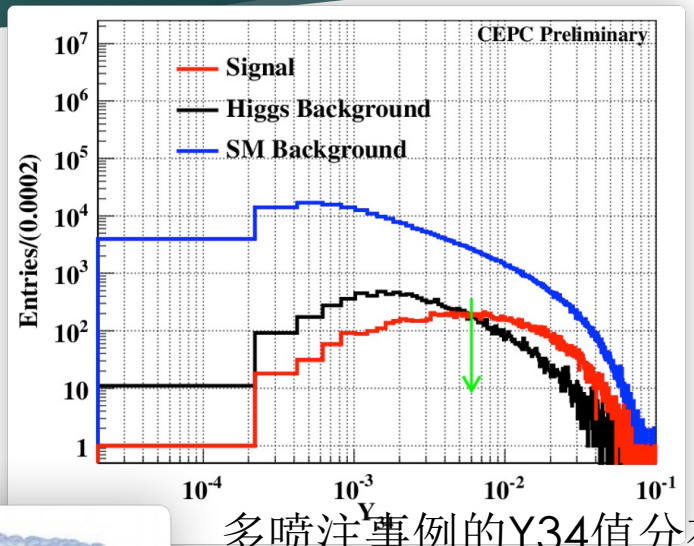
Mode	$b\bar{b}$
BR (%)	57.8

- 2012年发现希格斯玻色子后，进入精确测量时代
- 需要新机器，发现新物理
- 喷注是精确测量中一个重要的物理对象
- 如何鉴别喷注？
  - 手动cut→多变量分析（TMVA）→深度学习
- 基于CEPC基准探测器的结果
  - b夸克喷注鉴别：80%效率、90%以上本底排除率
  - c夸克喷注鉴别：60%效率、60%本底排除率

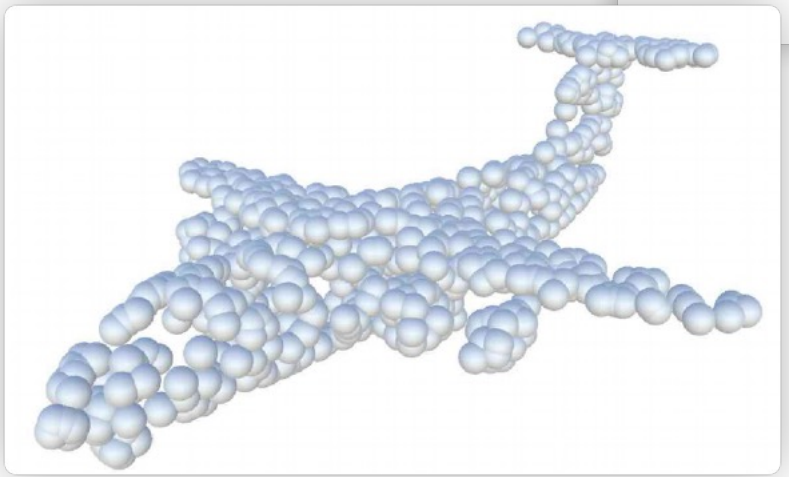


机器学习框架

- 传统喷注鉴别算法（高级变量）
  - 优点：基于相关物理规律和性质，容易解释
  - 缺点：经验依赖较强，很难发现某些特征
- 基于图的神经网络算法
  - 不变性 $f(x) = f(x') = f(g(x))$
  - 等变性 $f(g(x)) = g(f(x))$

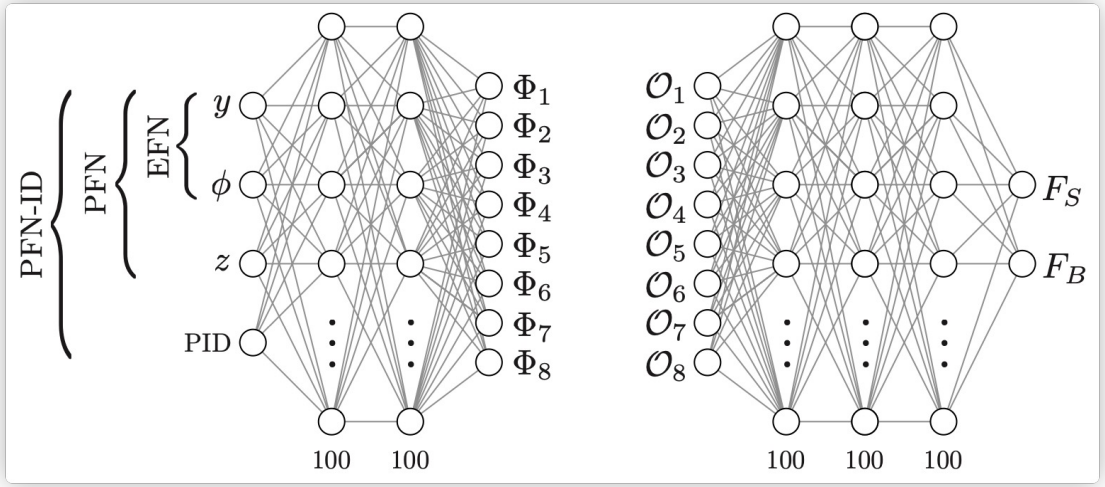


多喷注事例的Y34值分布



机器学习框架

- DeepSets特点
  - 粒子顺序交换不变
  - 充分提取多个有效特征
  - 提取出的特征作为输入用一个三层全连接层的神经网络进行分类

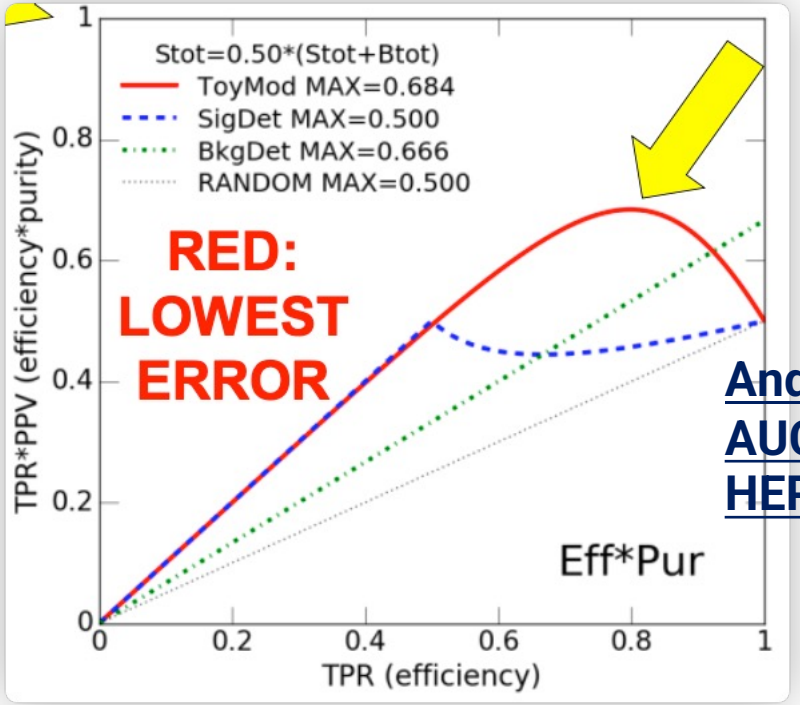


Energy flow networks: deep sets for particle jets

评价指标

- 效率（recall）
- 纯度（precision $\sigma^2 \propto \frac{1}{\epsilon\rho}$
- 效率*纯度 [0,1]
- 精度
- AUC/ROC
- …

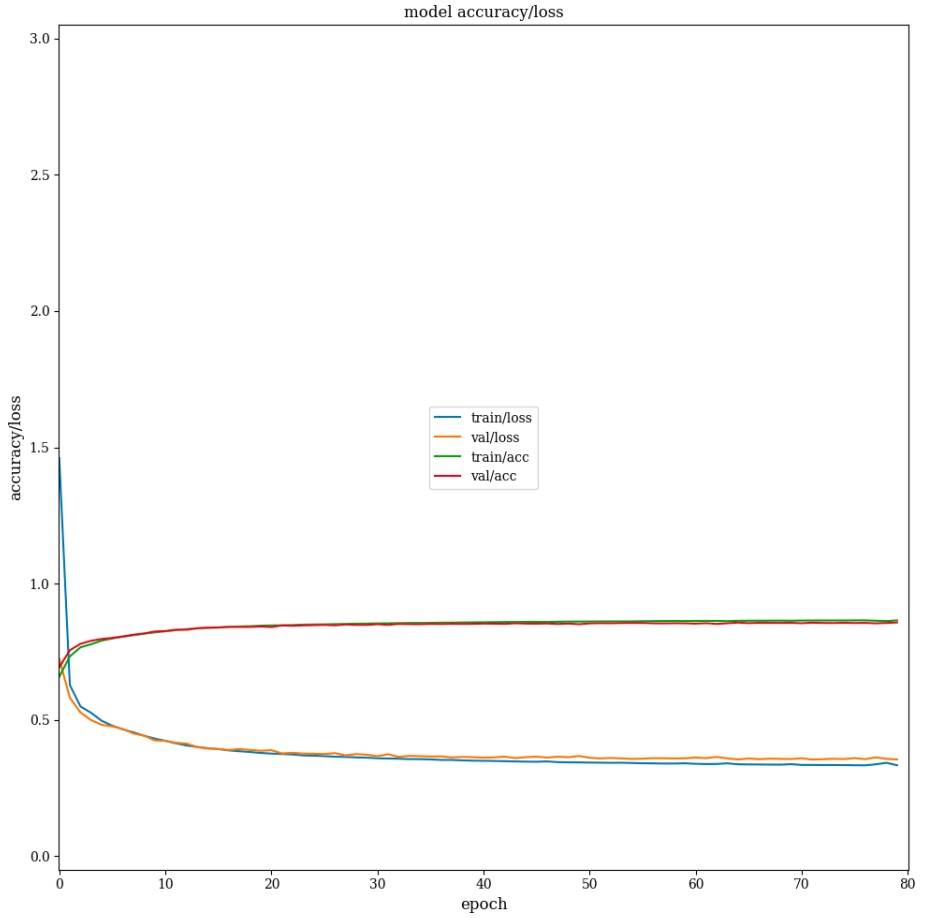
TP (S_{sel}) FP (B_{sel}) FN (S_{rej}) TN (B_{rej})	TP (S_{sel}) FP (B_{sel}) FN (S_{rej}) TN (B_{rej})	TP (S_{sel}) FP (B_{sel}) FN (S_{rej}) TN (B_{rej})
$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$	$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$	$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} = 1-\mathrm{FPR}$
HEP: "efficiency" $\epsilon_s = \frac{S_{sel}}{S_{tot}}$	HEP: "purity" $\rho = \frac{S_{sel}}{S_{sel}+B_{sel}}$	HEP: "background rejection" $1-\epsilon_b = 1-\frac{B_{sel}}{B_{tot}}$



Andrea Valassi : ROC's, AUC's and alternatives in HEP and other domains

数据集

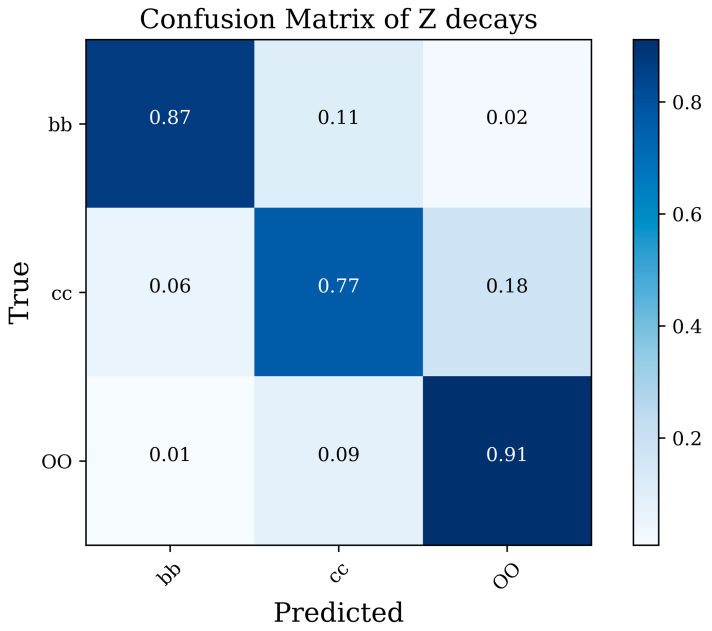
- 基于CEPC基准探测器的Z玻色子衰变全模拟数据集
- 味道鉴别
  - 每种味道的喷注90万（$b, c, o = uds$）
  - 用eekt算法聚类成两个喷注
- 训练:验证:测试 = 7:1.5:1.5
- 使用particle level的变量（避免特征工程）

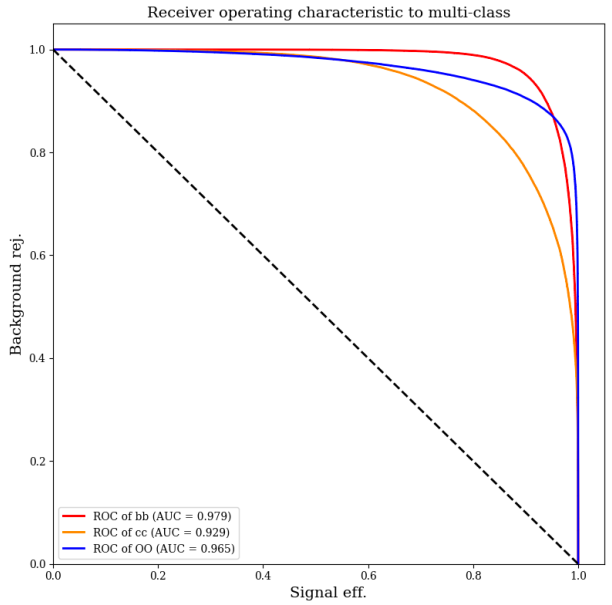


结果

CEPC CDR baseline (machine learning)

算法	PFN	DNN	BDT	GBDT	gcforest	XGBoost
精度	0.850	0.788	0.776	0.794	0.785	0.801





- 与传统的决策树和XGBoost算法相比，平均精度提高约6%
- 混淆矩阵的对角元表示各个分类的准确率。

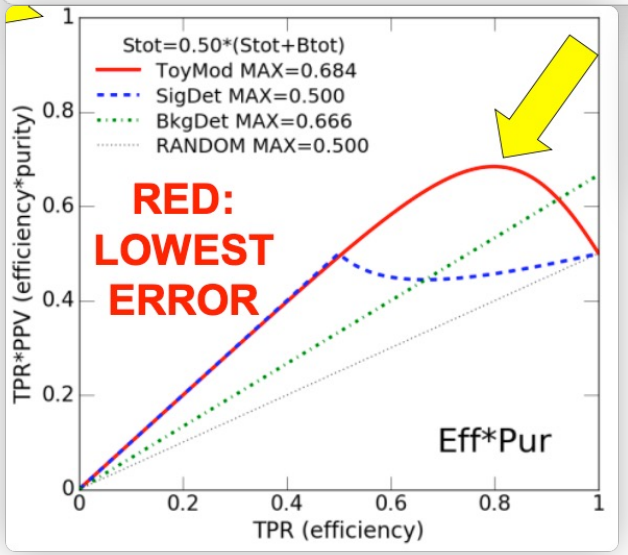
本底排除率

鉴别效率	CEPC 基准	PFN
b 喷注 80%	90% 以上	99%
c 喷注 60%	60%	97%

结果

- 从物理分析角度评价喷注鉴别
  - 与CEPC基准探测器的结果相比，提升明显，尤其是对于c夸克喷注
  - 能有效降低统计误差

tag	$\epsilon_S(\%)$	$\epsilon \times \rho$		
		LCFIPlus	XGBoost	PFN
b	80	-	0.747	0.763
	90	0.72	0.713	0.752
c	60	0.36	-	0.485
	70	-	-	0.497
	80	-	0.345	0.467
	90	-	0.292	0.402



$$\sigma^2 \propto \frac{1}{\epsilon\rho}$$

$$\sqrt{0.467/0.345} = 1.16$$

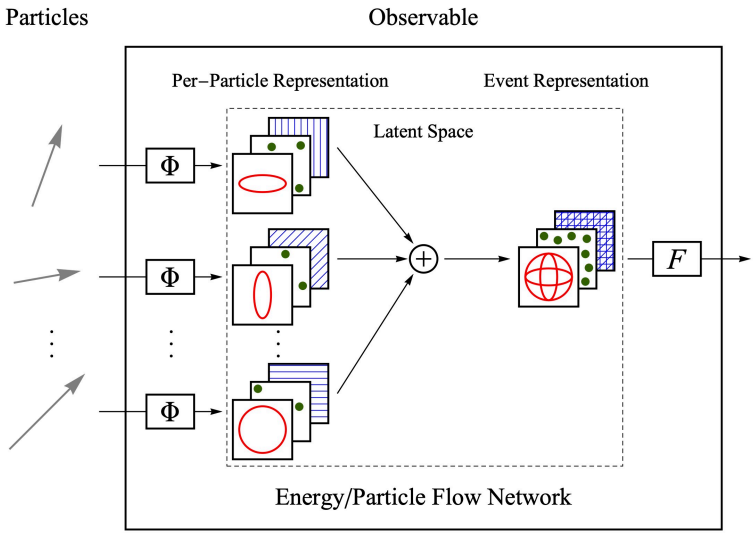
总结和展望

- 总结：初步研究表明，基于DeepSets模型的夸克喷注标记问题取得了明显的性能提升。

- 未来的研究计划主要关注以下两方面：
  - 一是对数据信息进行进一步发掘，以及对模型的超参数进行深入优化，期望性能可以得到的进一步提高；
  - 二是尝试对完整事例进行标记，该结果预期能对三种玻色子性质的精确测量有较大提升，降低工作量和时间成本；
  - 三是将这一工具投入到CEPC的模拟分析中，产生一批物理结果。

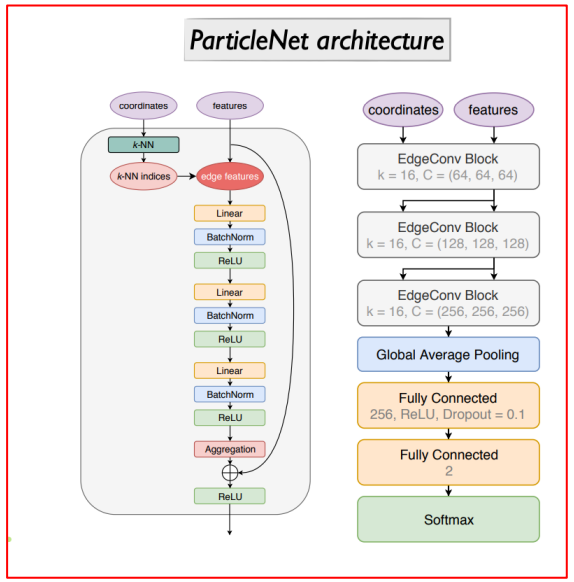
深度学习方法的新发展

Energy flow networks: deep sets for particle jets



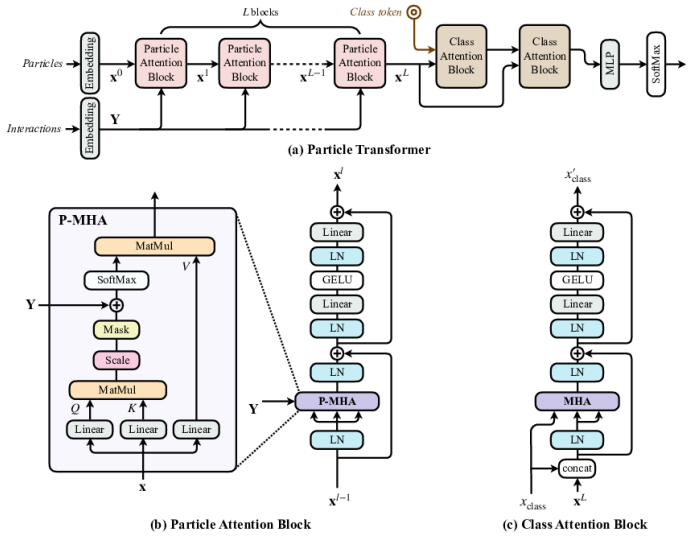
Energyflow networks: 2019

Jet tagging via particle clouds



ParticleNet: 2020

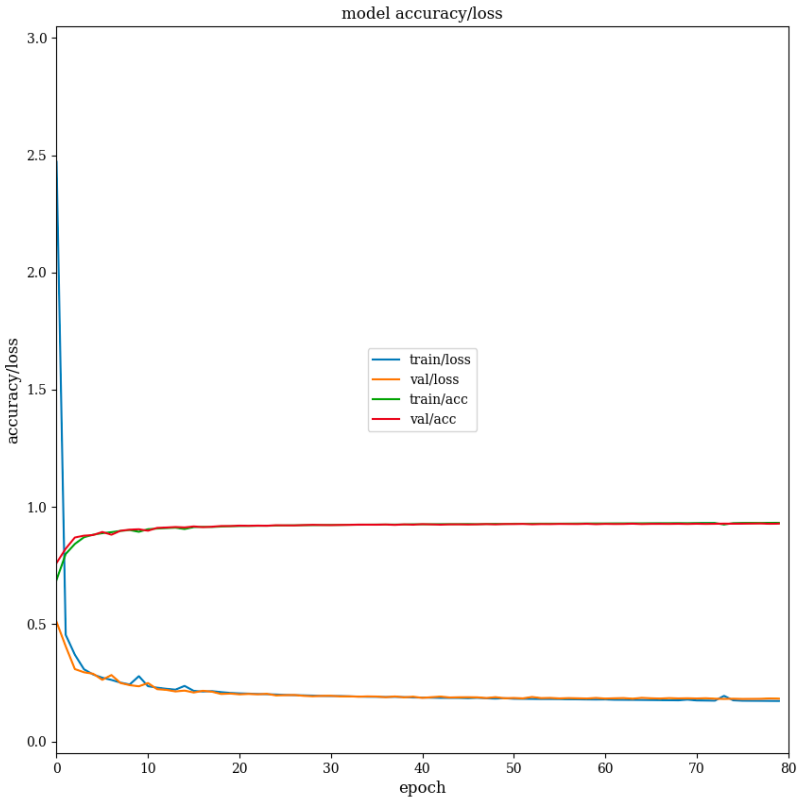
Particle Transformer for Jet Tagging

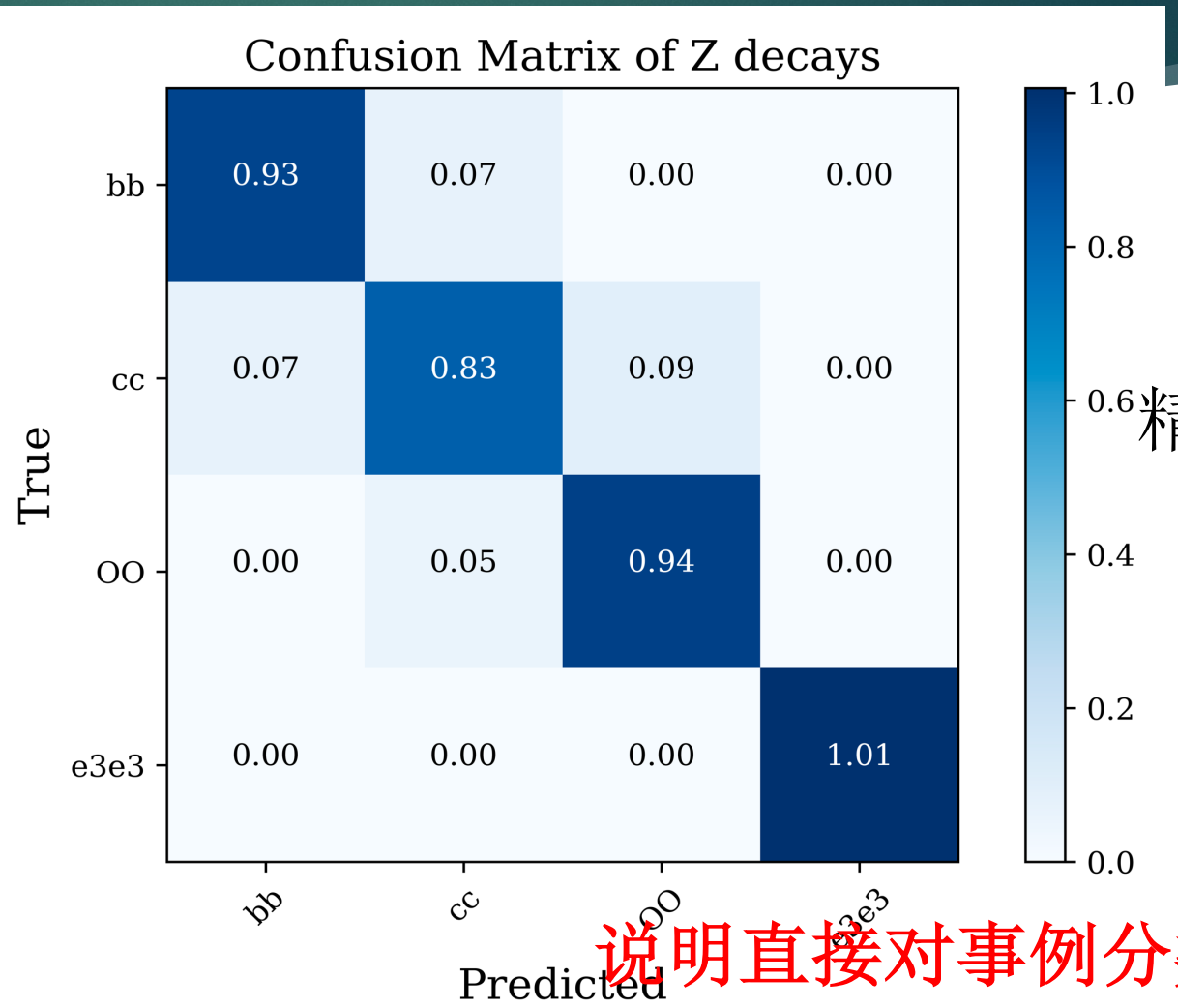


ParticleTransformer: 2022

我们的新想法

一个💡：对事例直接分类？





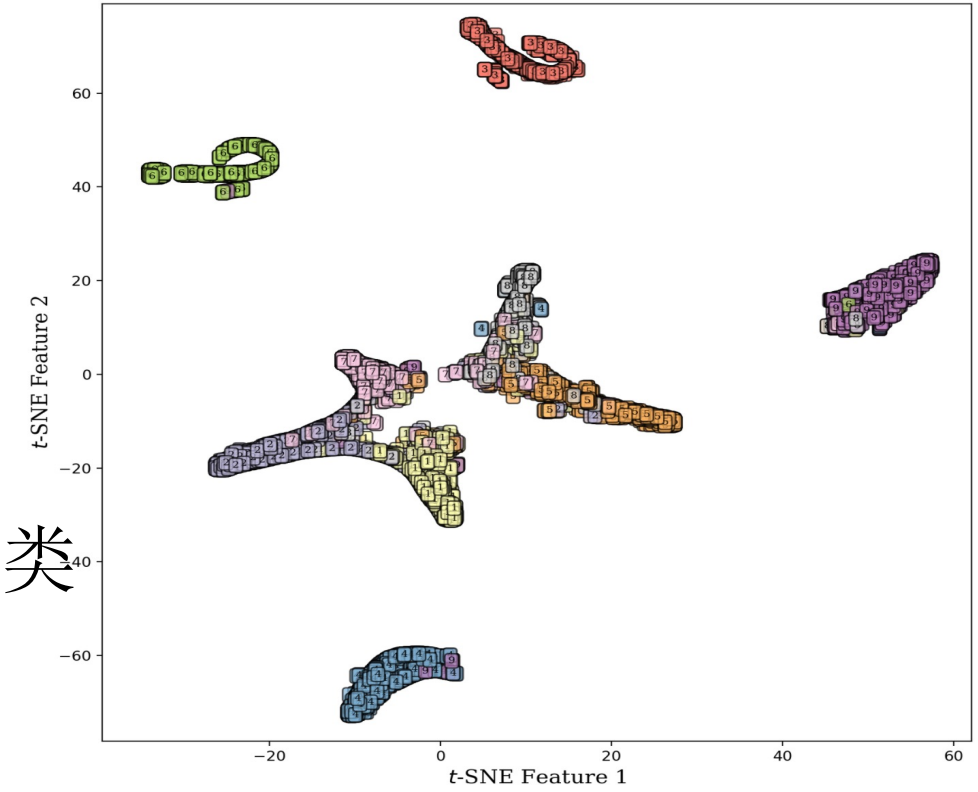
精度=0.925

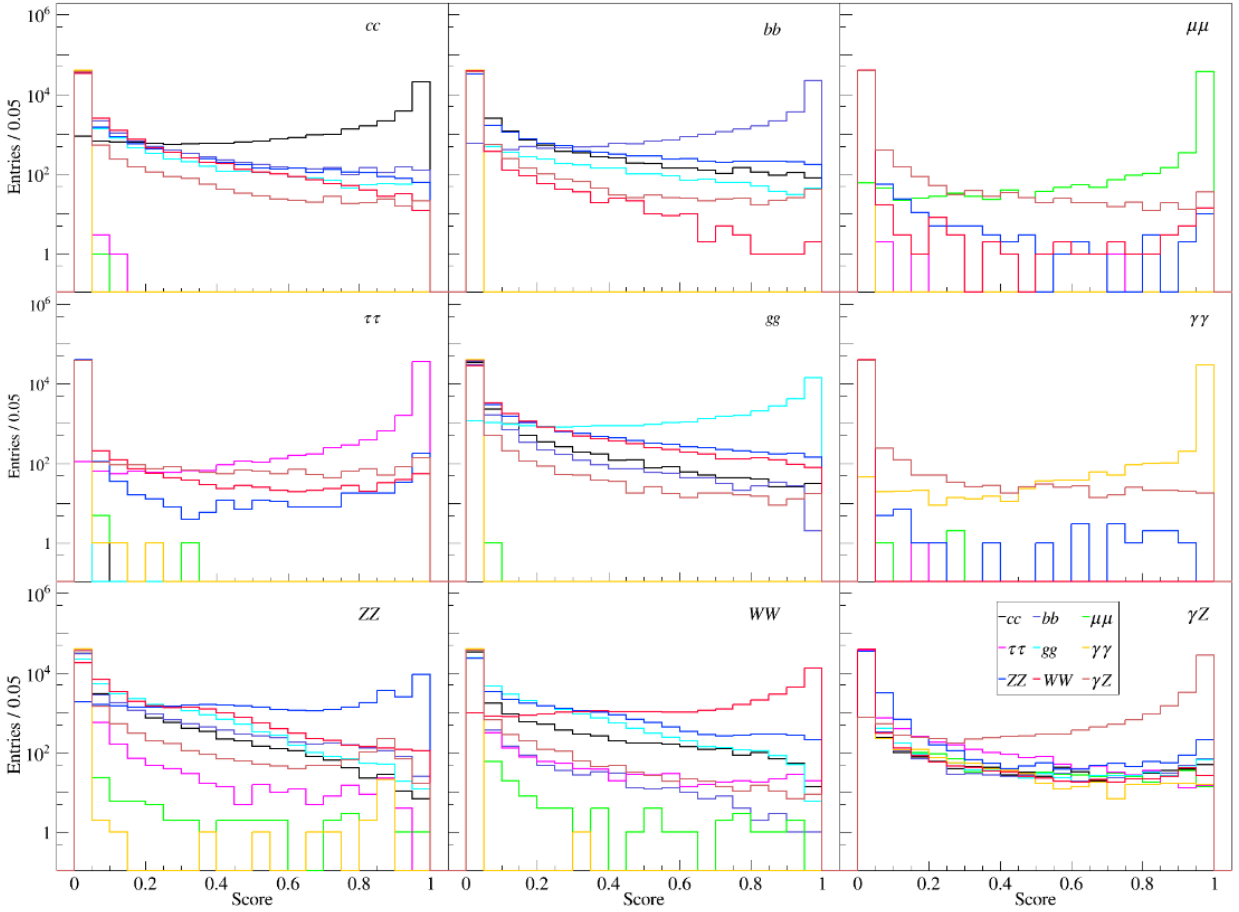
说明直接对事例分类有效！

我们的新想法

能不能更进一步？ 当然可以！

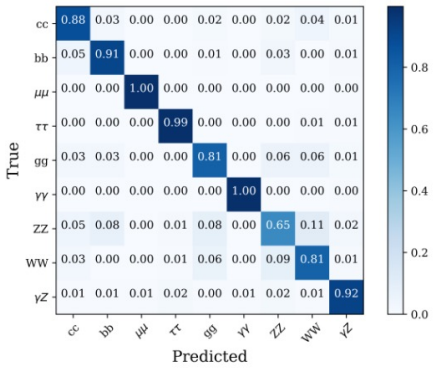
9分类



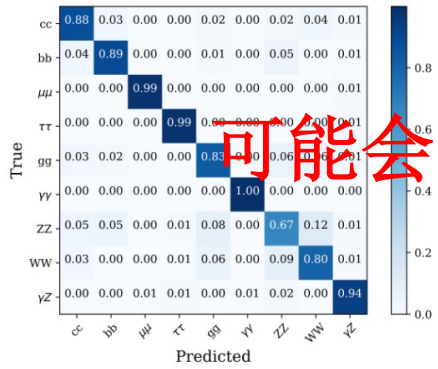


更多尝试

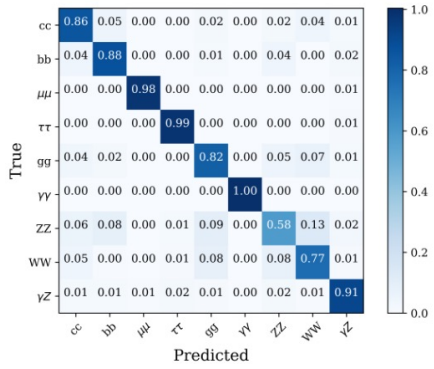
尝试更多分类

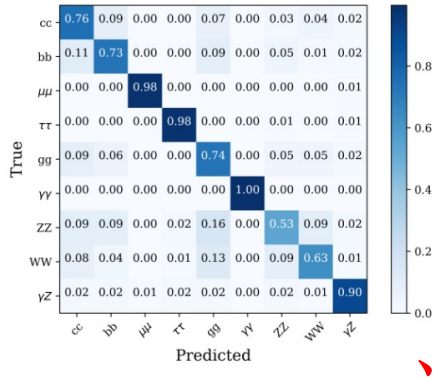


(a)

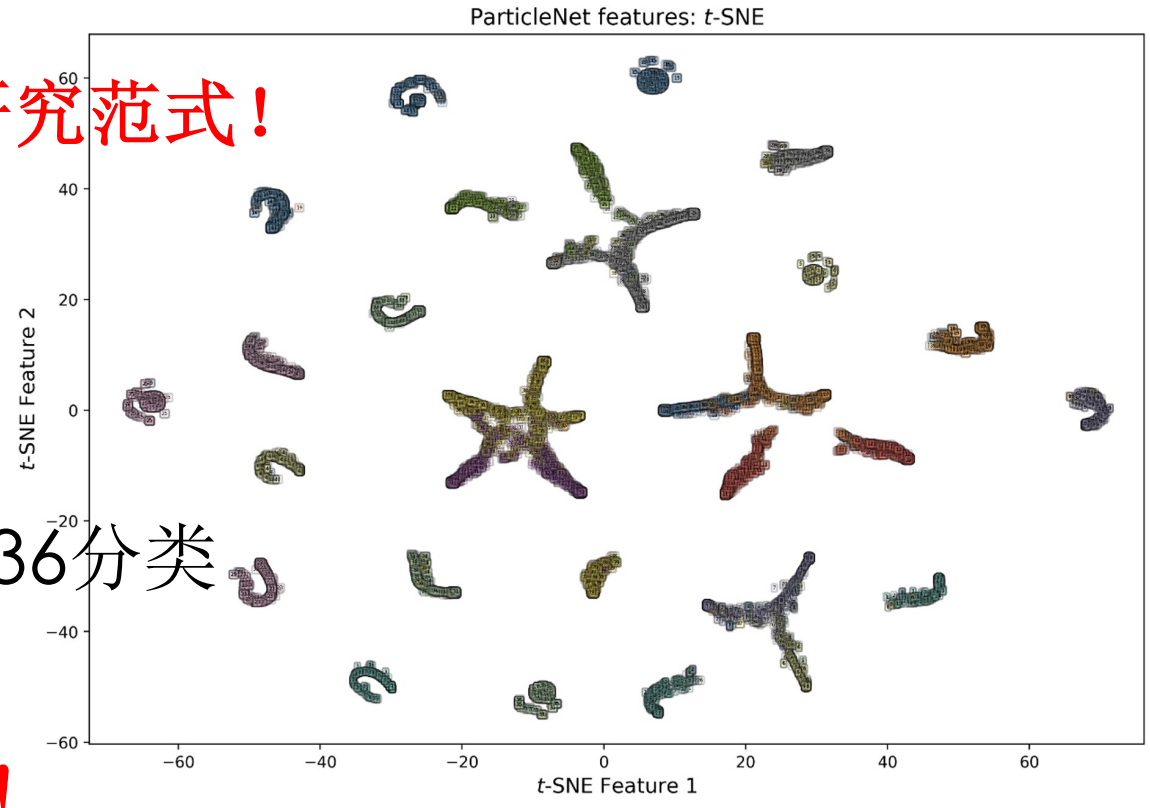


(b)





可能会改变研究范式！



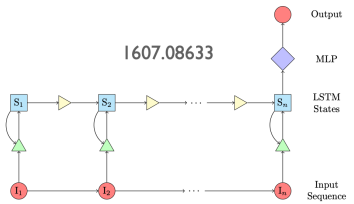
36分类

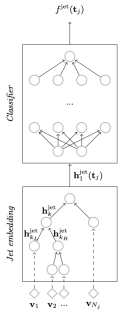
Chinese Physics C Vol. 46, No. 11 (2022) 113001

谢谢！

backup

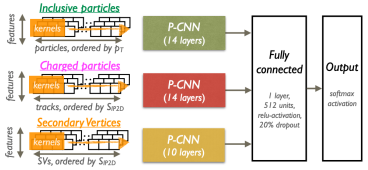
1607.08633

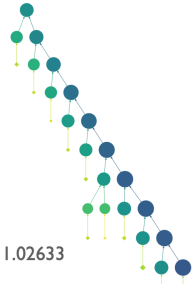




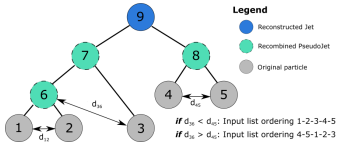
1702.00748

"DeepAK8"
CMS-DP-2017-049



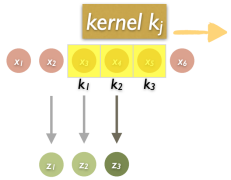


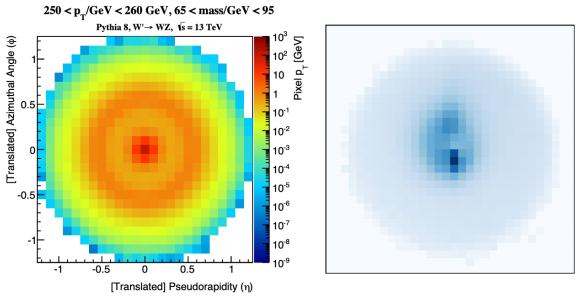
1711.02633



1711.09059

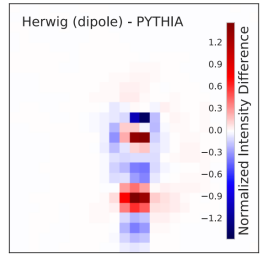
P-CNN:



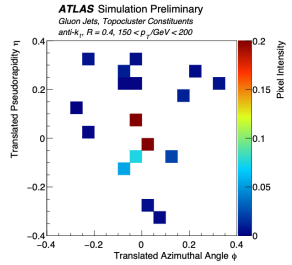


1511.05190

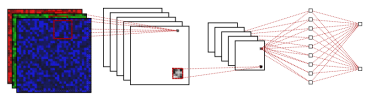
1603.09349



1609.00607



ATL-PHYS-PUB-2017-017



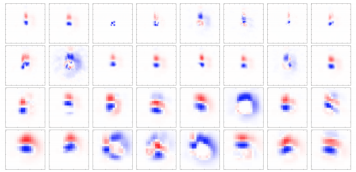
red = transverse momenta of charged particles

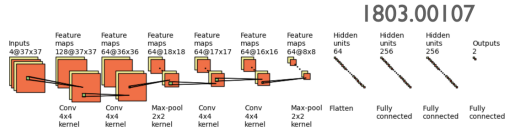
green = the transverse momenta of neutral particles

blue = charged particle multiplicity

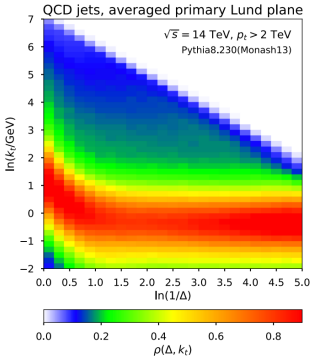
1612.01551

1701.08784





1803.00107



1807.04758

输入变量	描述
$\log E$	粒子能量的对数
$\cos\theta$	粒子极角的余弦
ϕ	粒子的方位角
PID	经鉴别后的粒子 ID
D_0	粒子次级顶点到束流的垂直距离
Z_0	粒子次级顶点到束流对撞点的水平距离
$\log M$	粒子动量长度的对数
$sigD_0$	粒子次级顶点到束流的垂直距离比上该项的误差
$sigZ_0$	粒子次级顶点到束流对撞点的水平距离比上该项的误差
δR	粒子与喷注之间的角距离
$\phi \cdot \sin\theta$	粒子方位角与极角正弦的乘积
$\delta\theta$	粒子与喷注之间的极角差
$\delta\phi$	粒子与喷注之间的方位角差
q	粒子的电荷