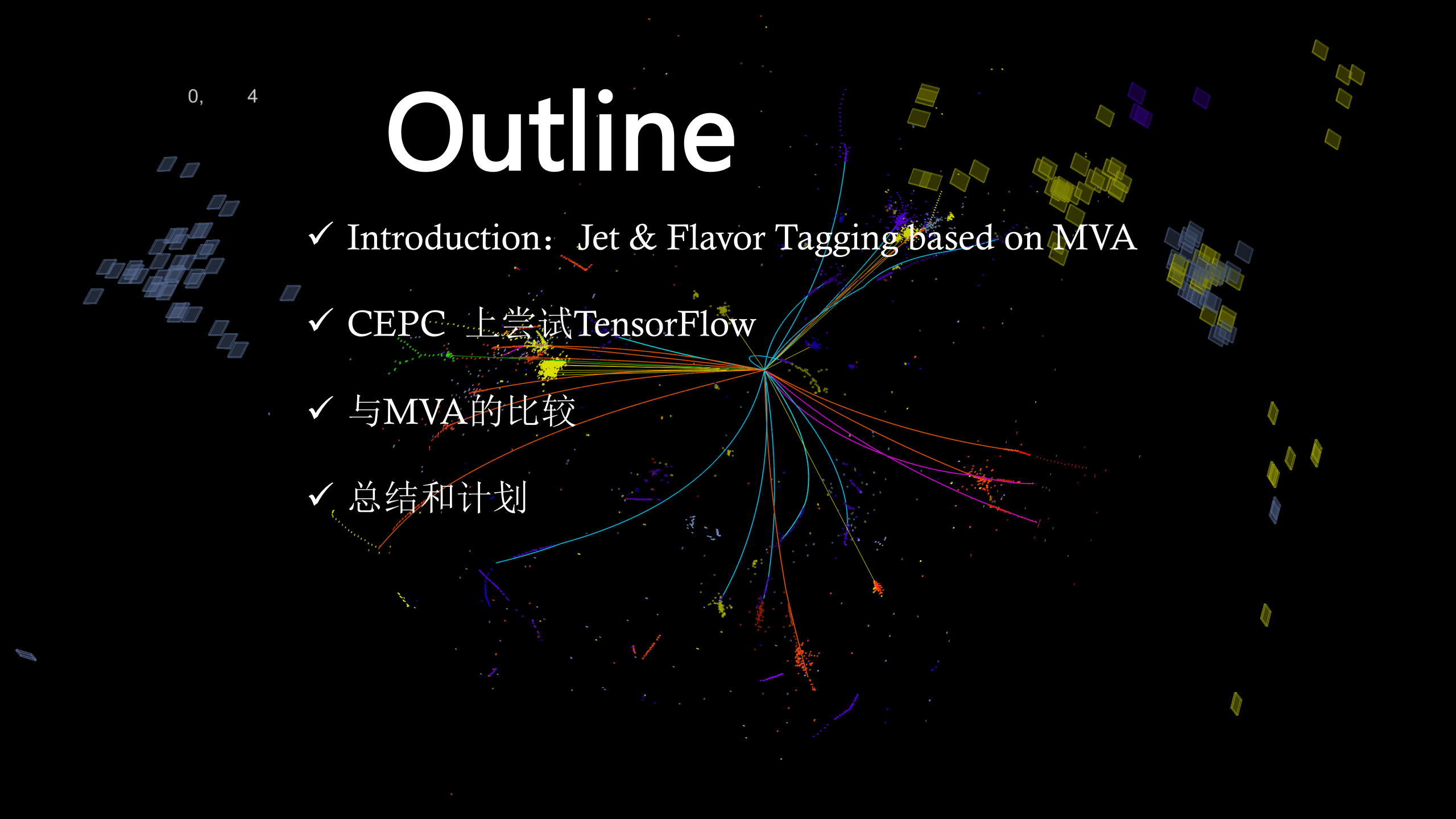
基于TensorFlow平台的喷注味道鉴别尝试

张冰洋 汪璐 李刚

中国科学院高能物理研究所



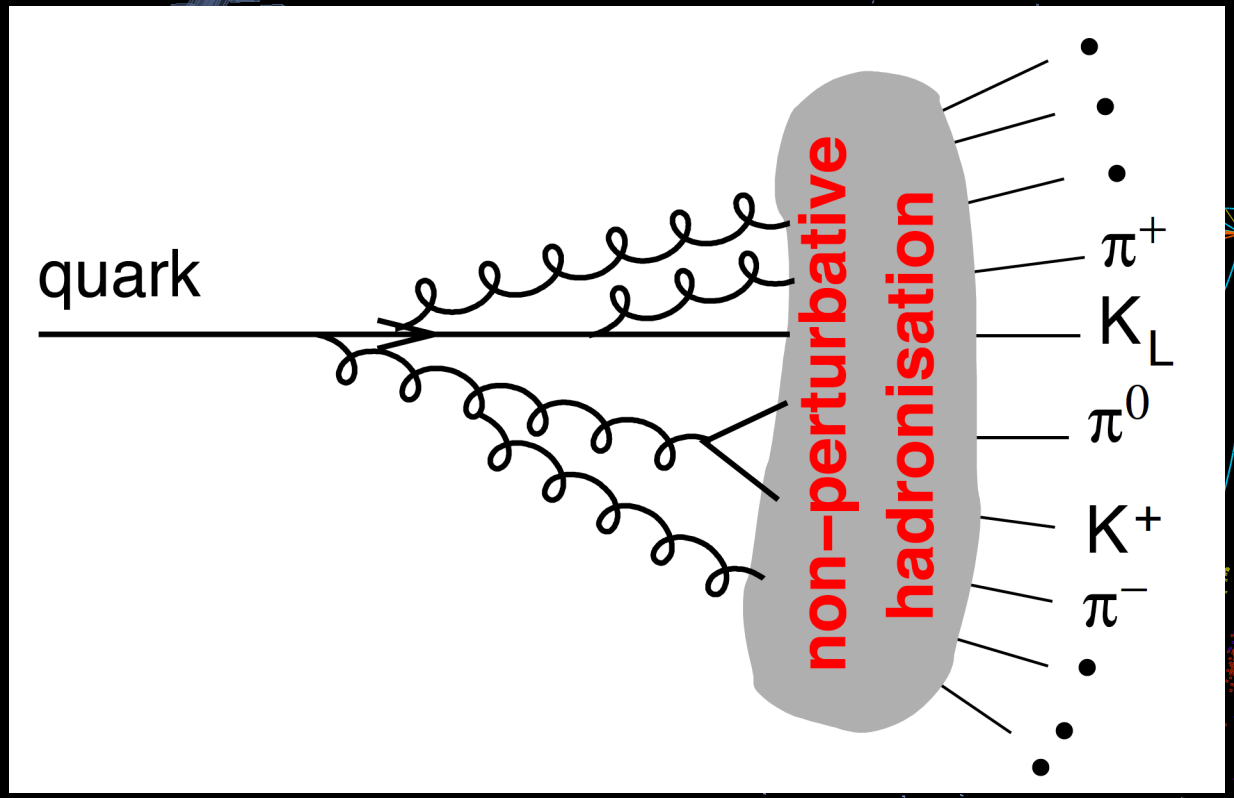
Outline

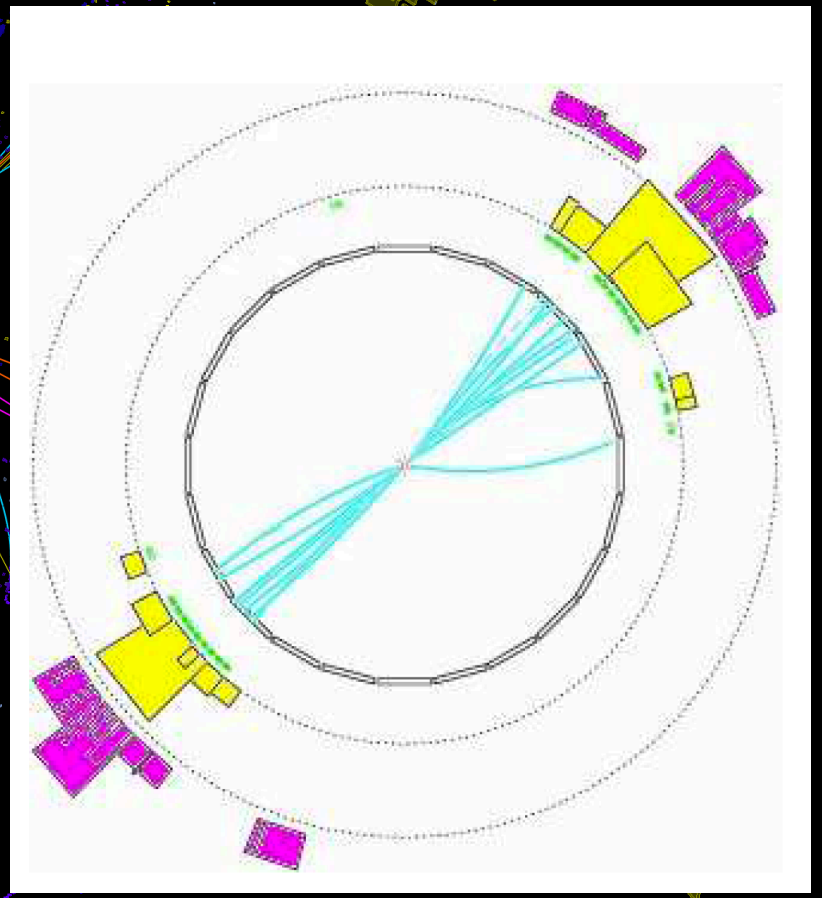
- ✓ Introduction：Jet & Flavor Tagging based on MVA
- ✓ CEPC 上尝试TensorFlow
- ✓ 与MVA的比较
- ✓ 总结和计划

Jet?

High-energy partons lead to collimated bunches of hadrons Jets [0, 4]

Leading parton radiates gluons uniformly distributed azimuthally around jet axis



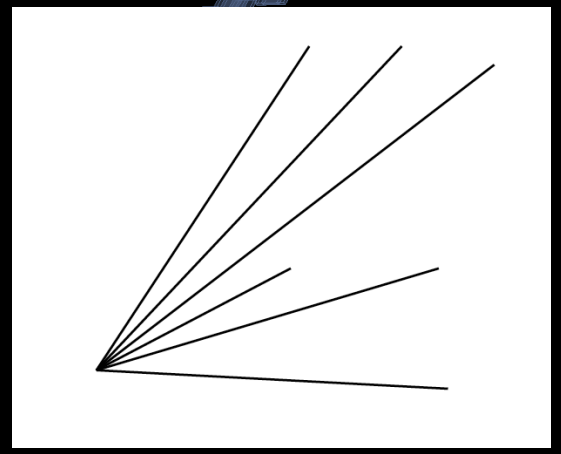
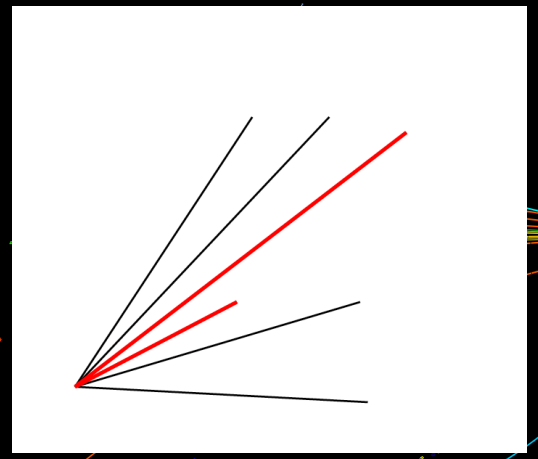
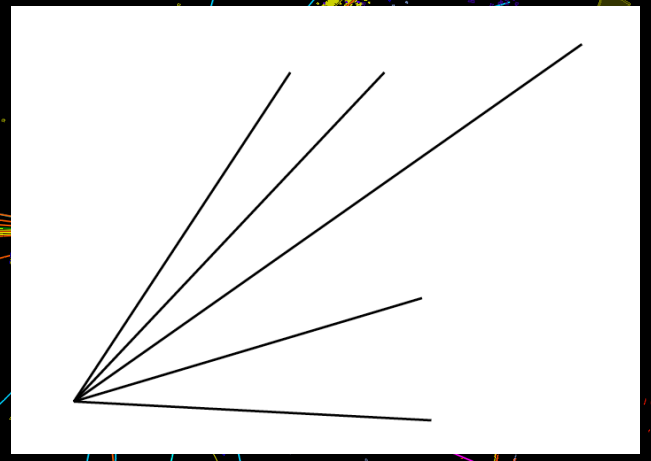
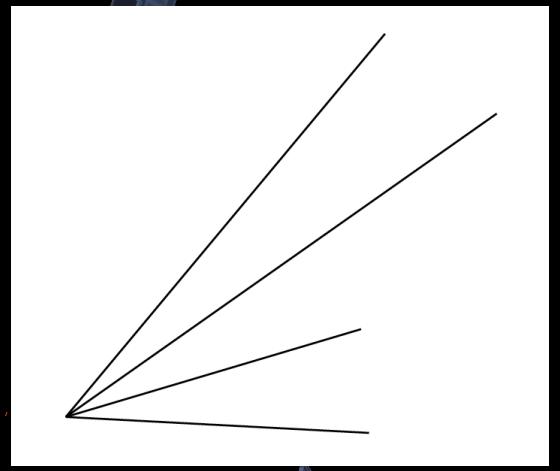
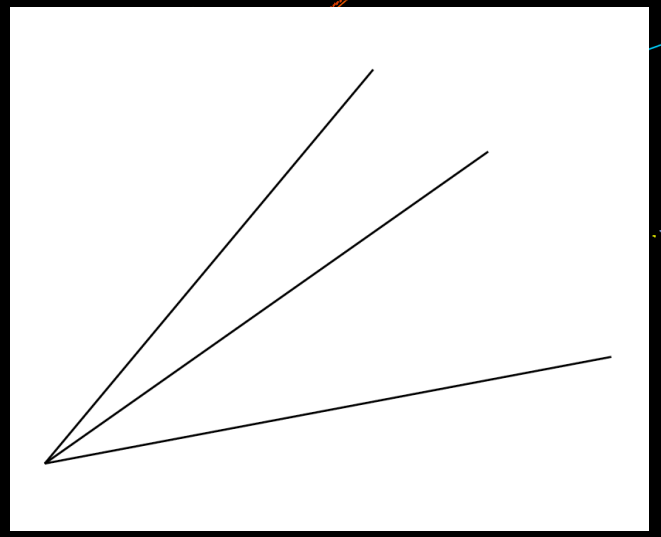
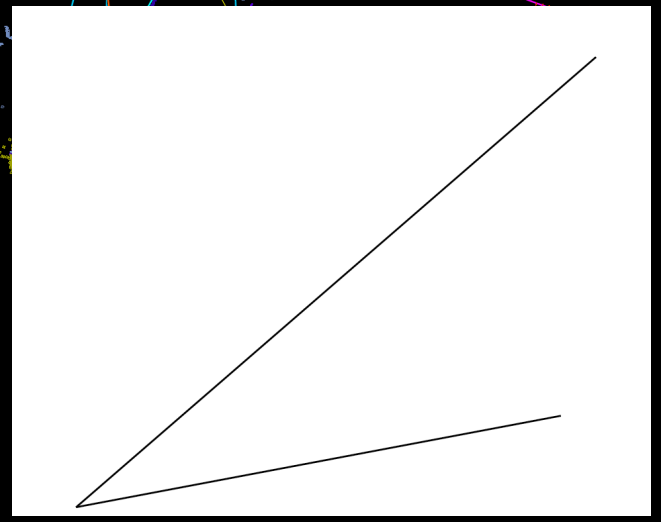


✓ Jet (definitions) provide central link between experiment and theory

✓ Jets are an input to almost all analyses

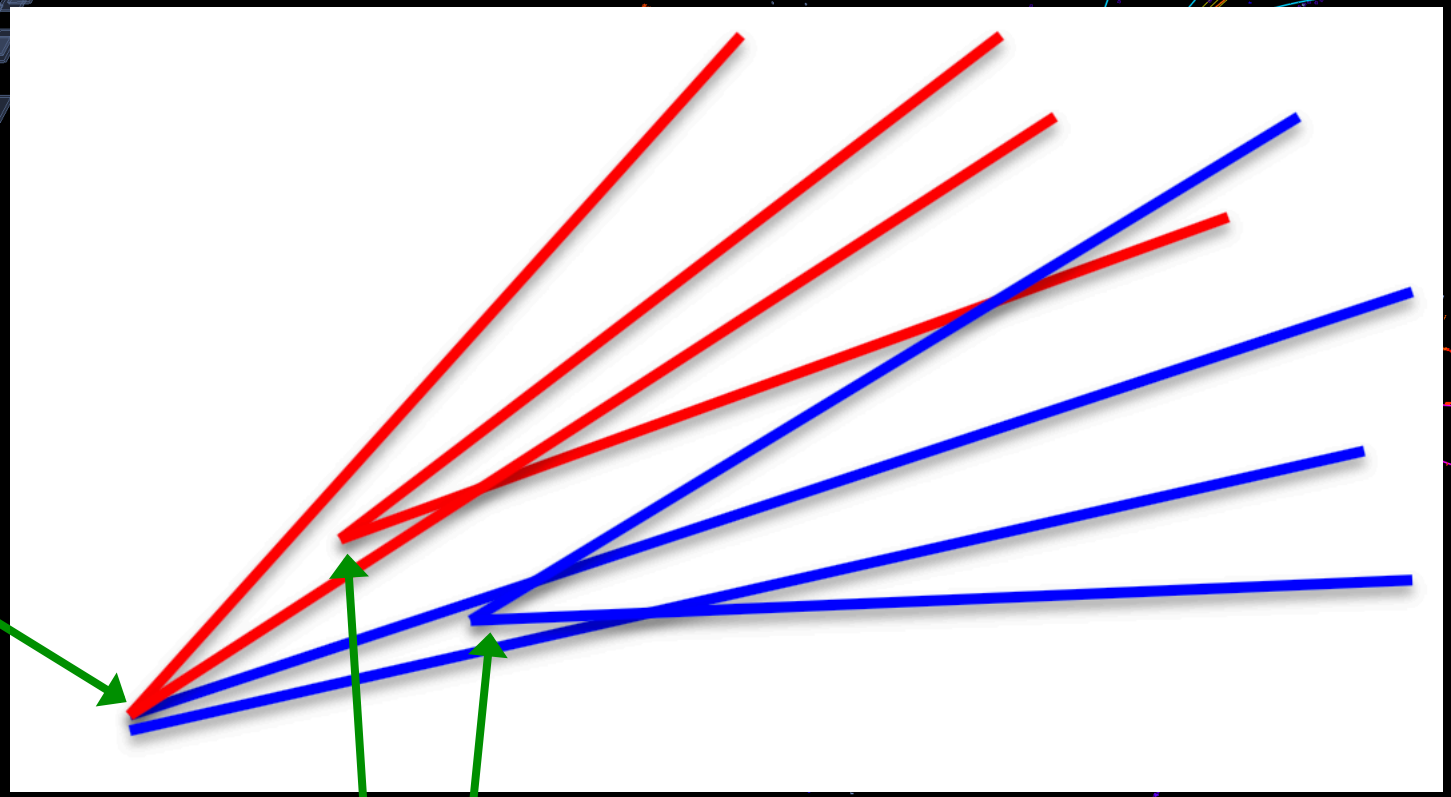
Two main classes of jet-clustering algorithms

- Sequential recombination and Cone

Jet Flavor Tagging – identify the flavor of a jet

b-tagging



- Benefit from precise vertex measurement, CEPC better than 5μm
- And PFA (global info.)
- And multivariate tools (BDT)

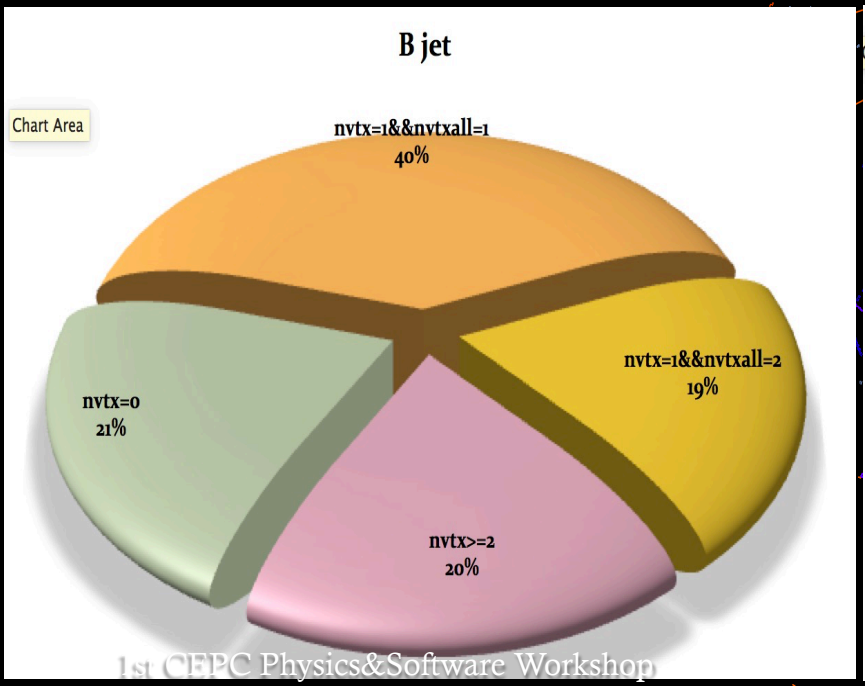
- Rather good performance at CEPC

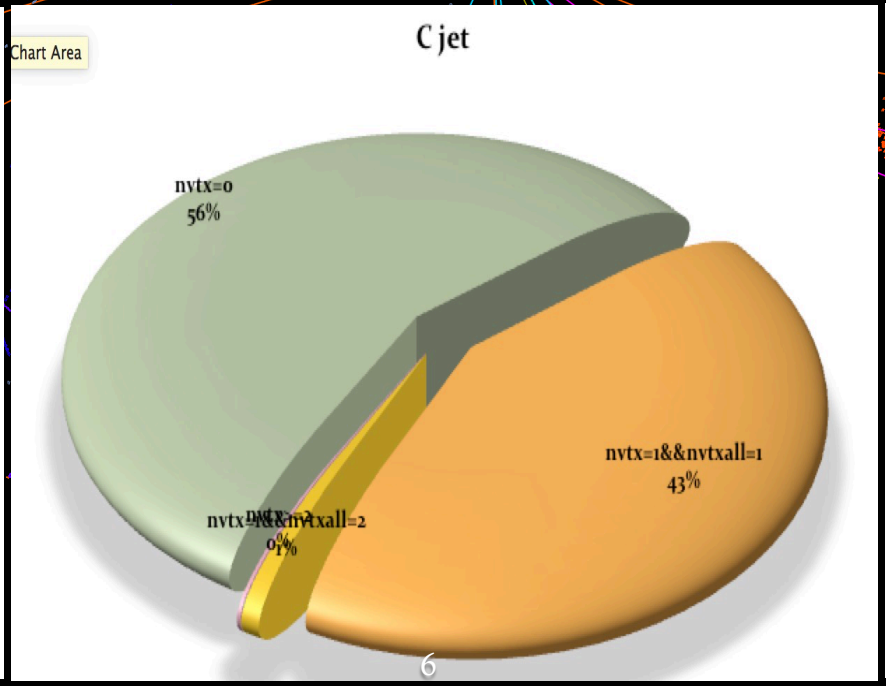
More ambitious:

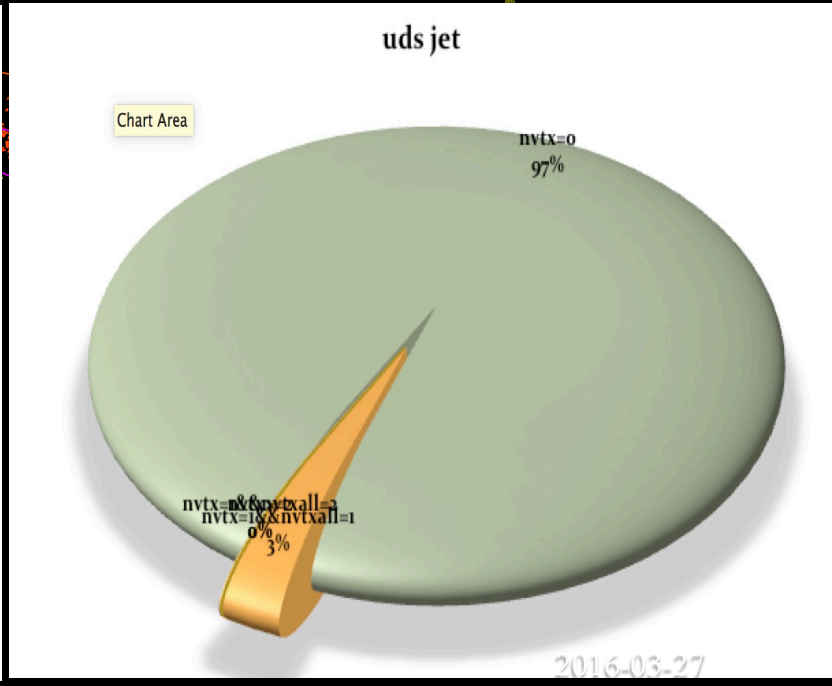
- Separate gluon and uds
- Jet charge
- Jet sub-structure …

Categorize jet based # of VTX→four independent MVA jobs – What ILC does

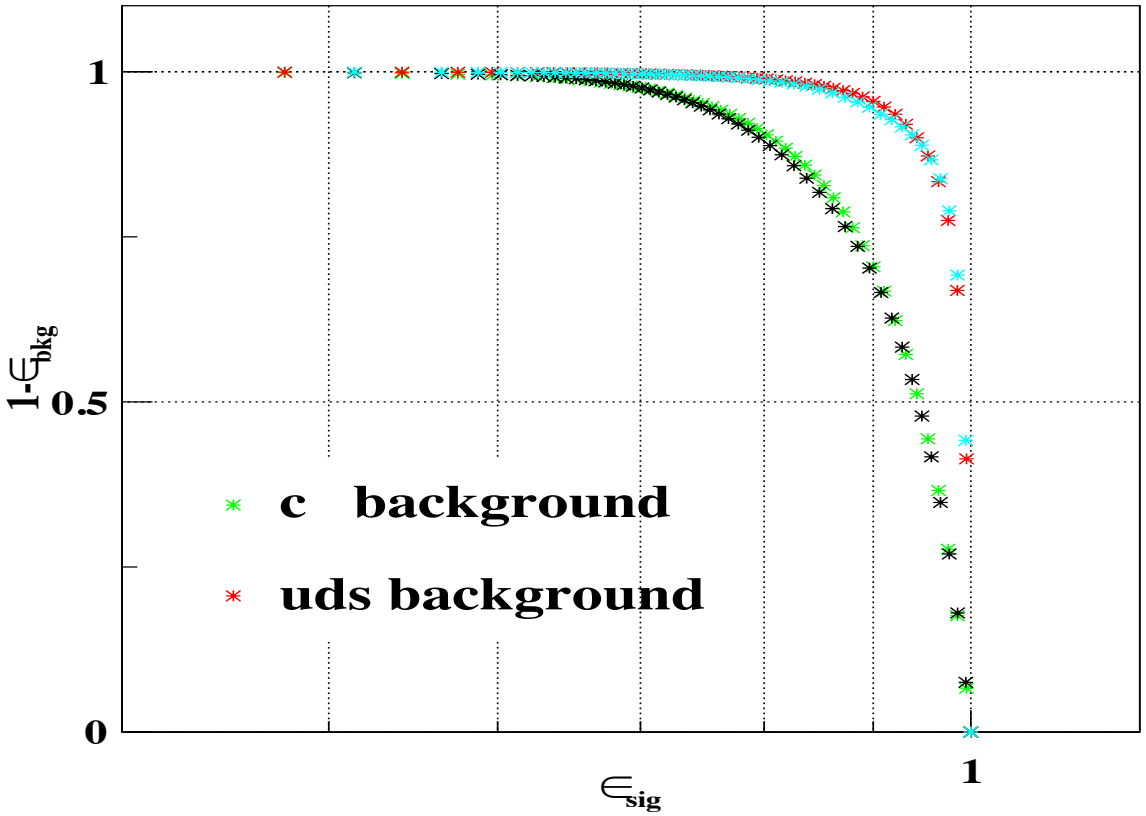
	Total	nvtx==0	nvtx=1&& Nvtxall==1	nvtx==1&&nvtxall==2	Nvtx>=2
B	395567	83099	156094	76239	80135
C	396692	223238	169400	3392	662
uds	393310	382522	10511	171	106

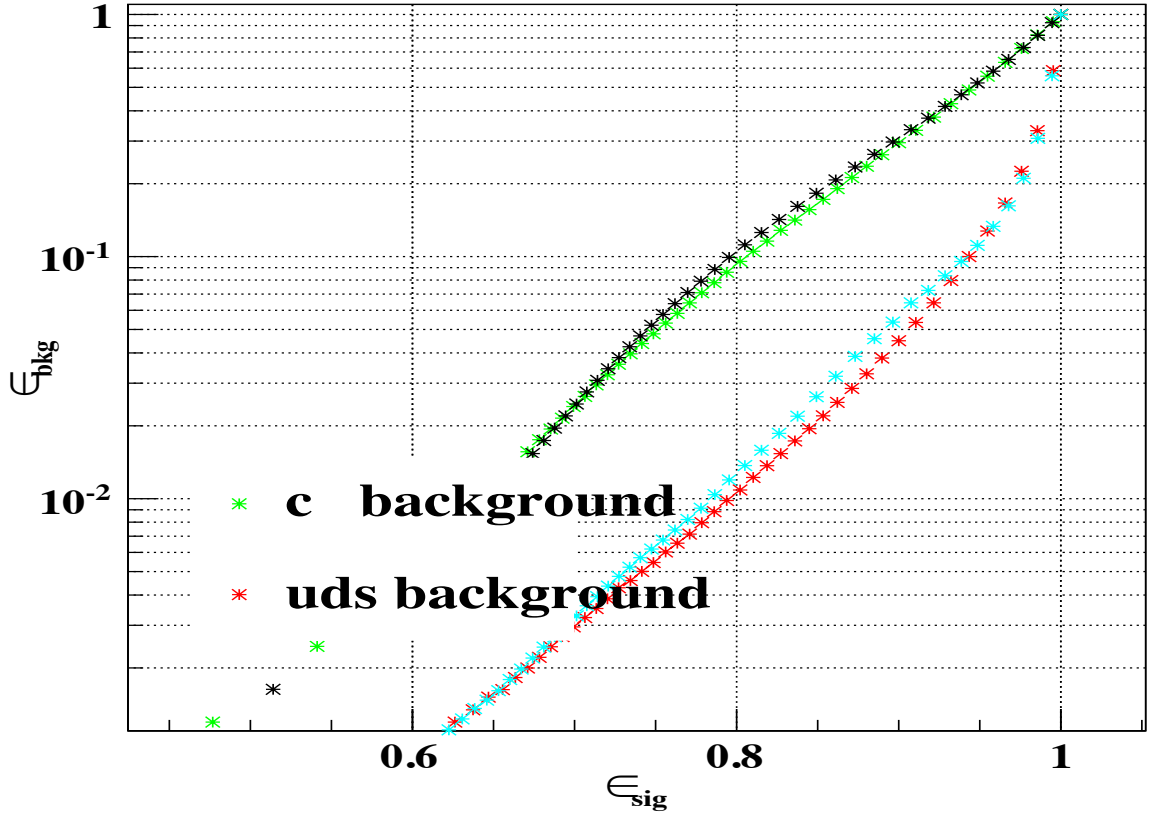






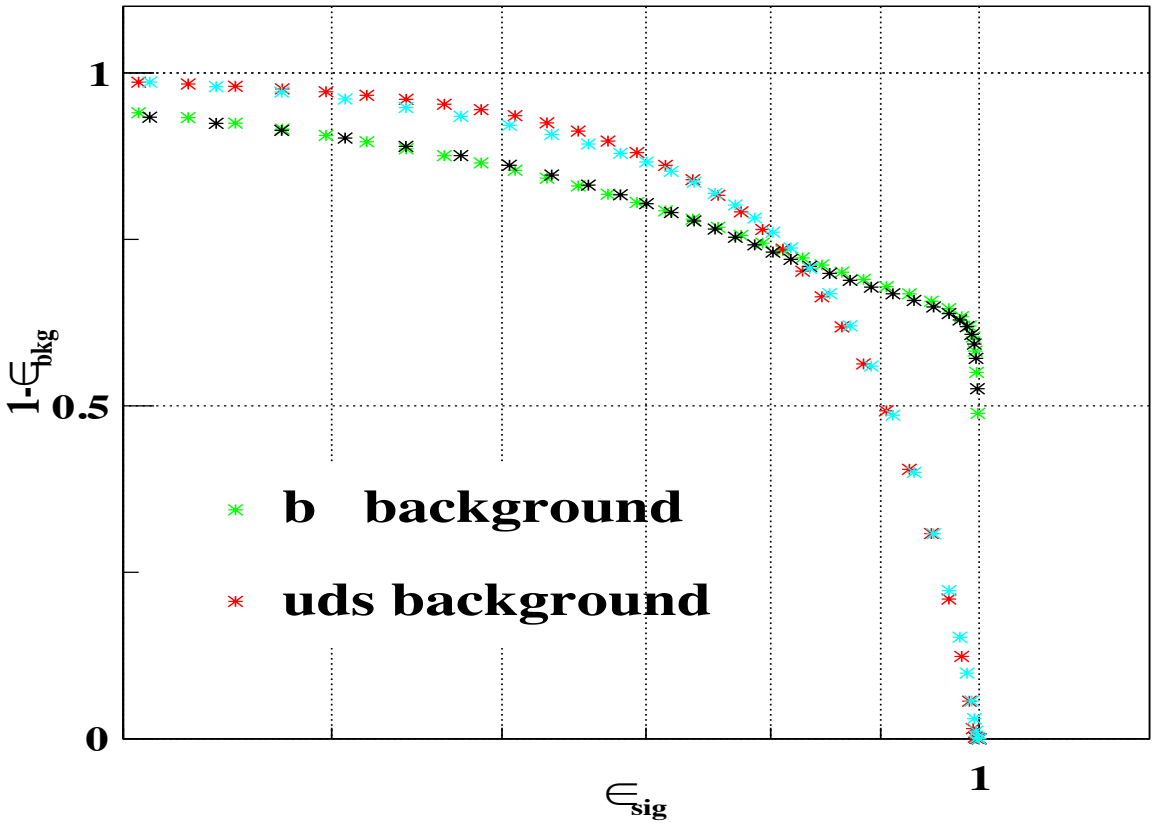
BTag

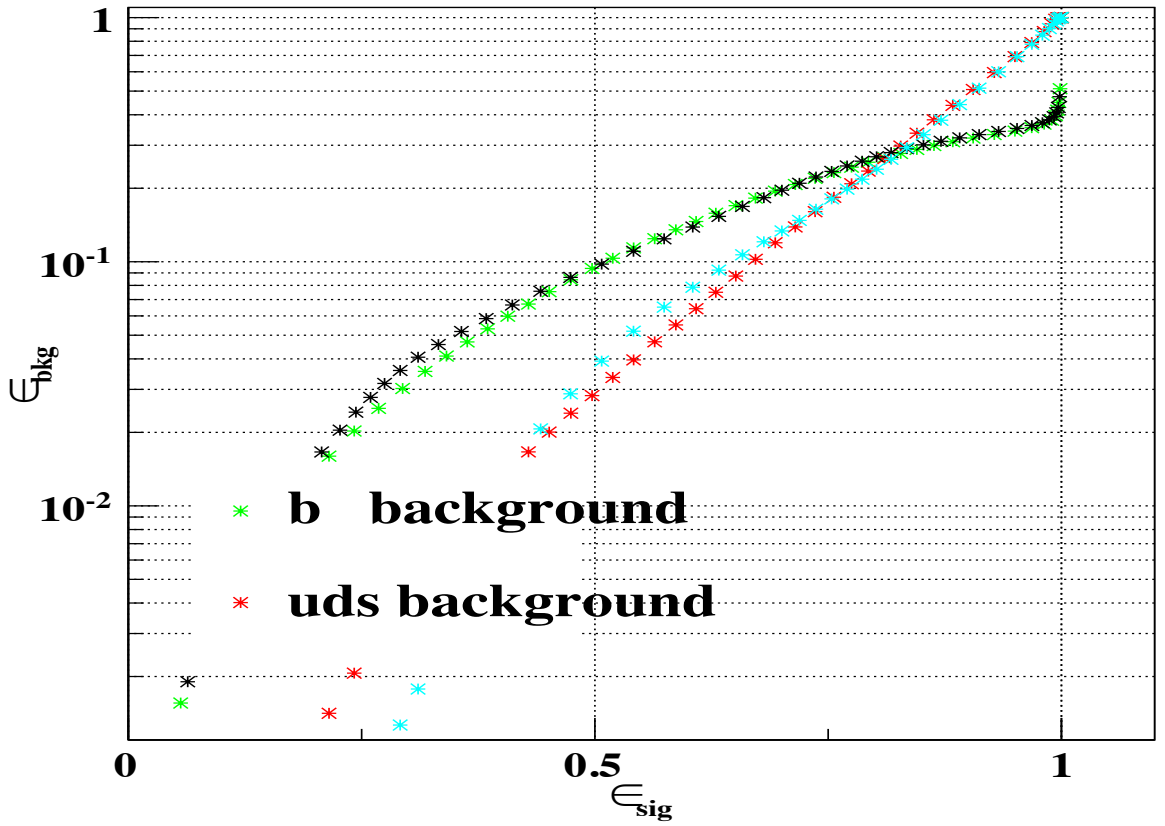




~60 variables in total, but 20-30 used in each category

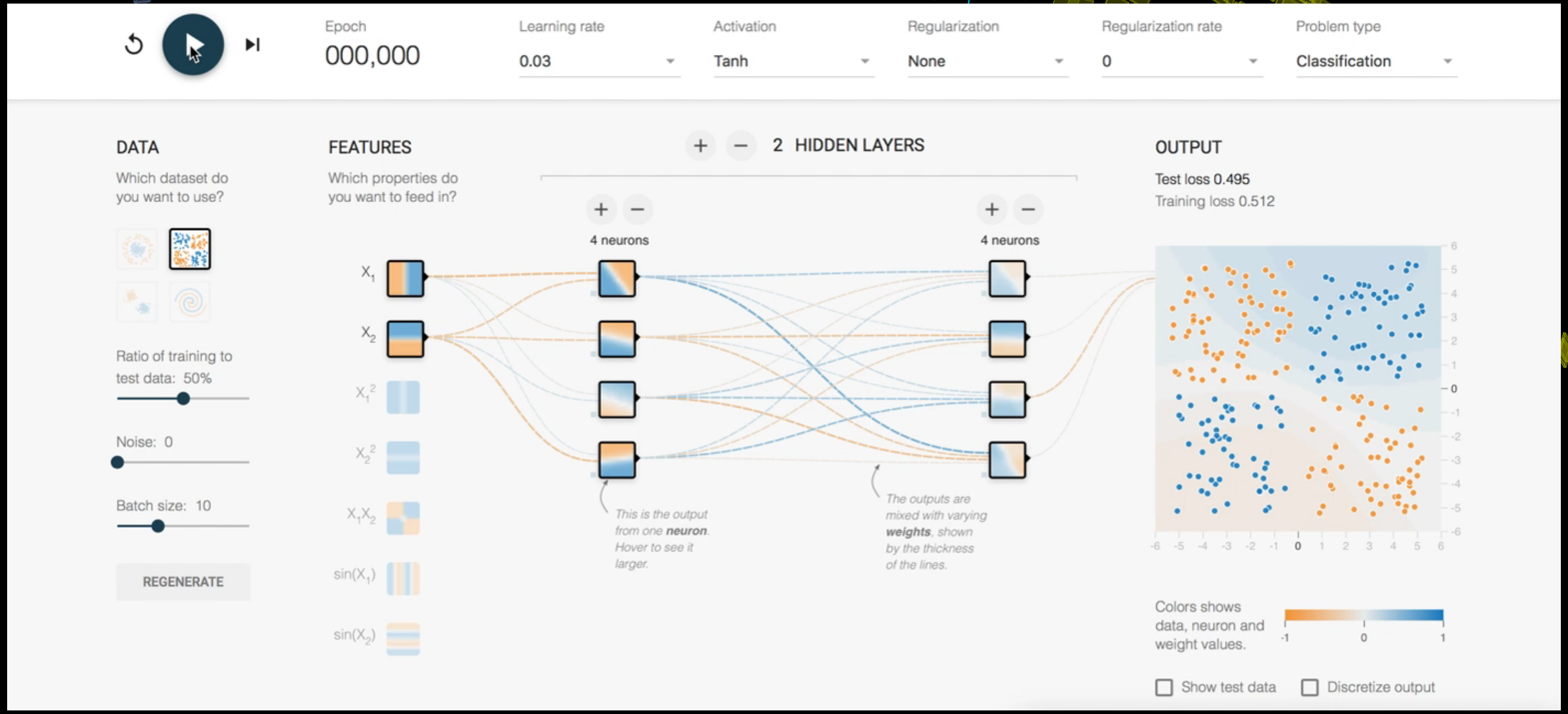
CTag


1
0.5
0
1
$1-\epsilon_{bkg}$
ϵ_{sig}
b background
uds background


1
10^{-1}
10^{-2}
0
0.5
1
ϵ_{bkg}
ϵ_{sig}
b background
uds background

0, 4

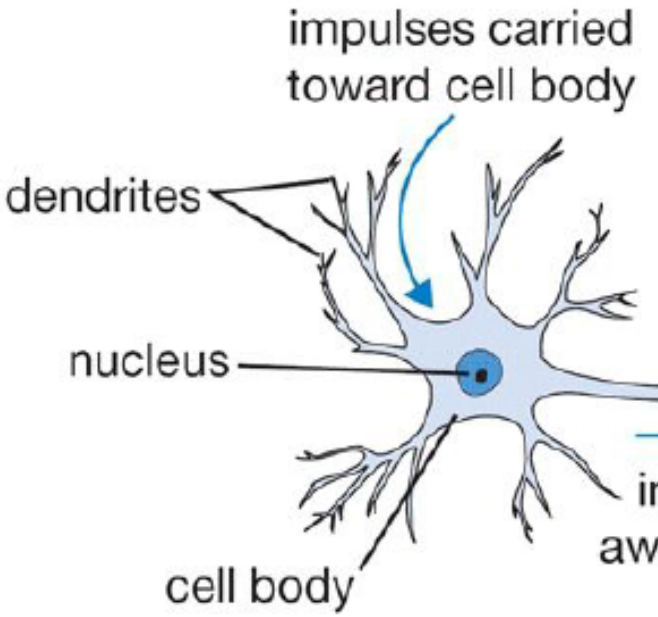
Trying TensorFlow

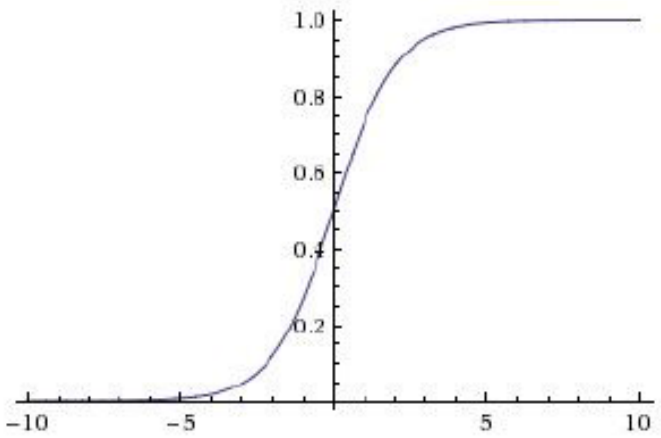


现状：

- 借鉴计算机视觉处理图像的方法，将喷注粒子的各物理量映射为二维图片的像素值，并采用卷积神经网络结构。在相同的信号效率下，其结果相比于传统方法本底排除能力有明显的提高。【Boosted Jet Tagging with Jet-Images and Deep Neural Networks】

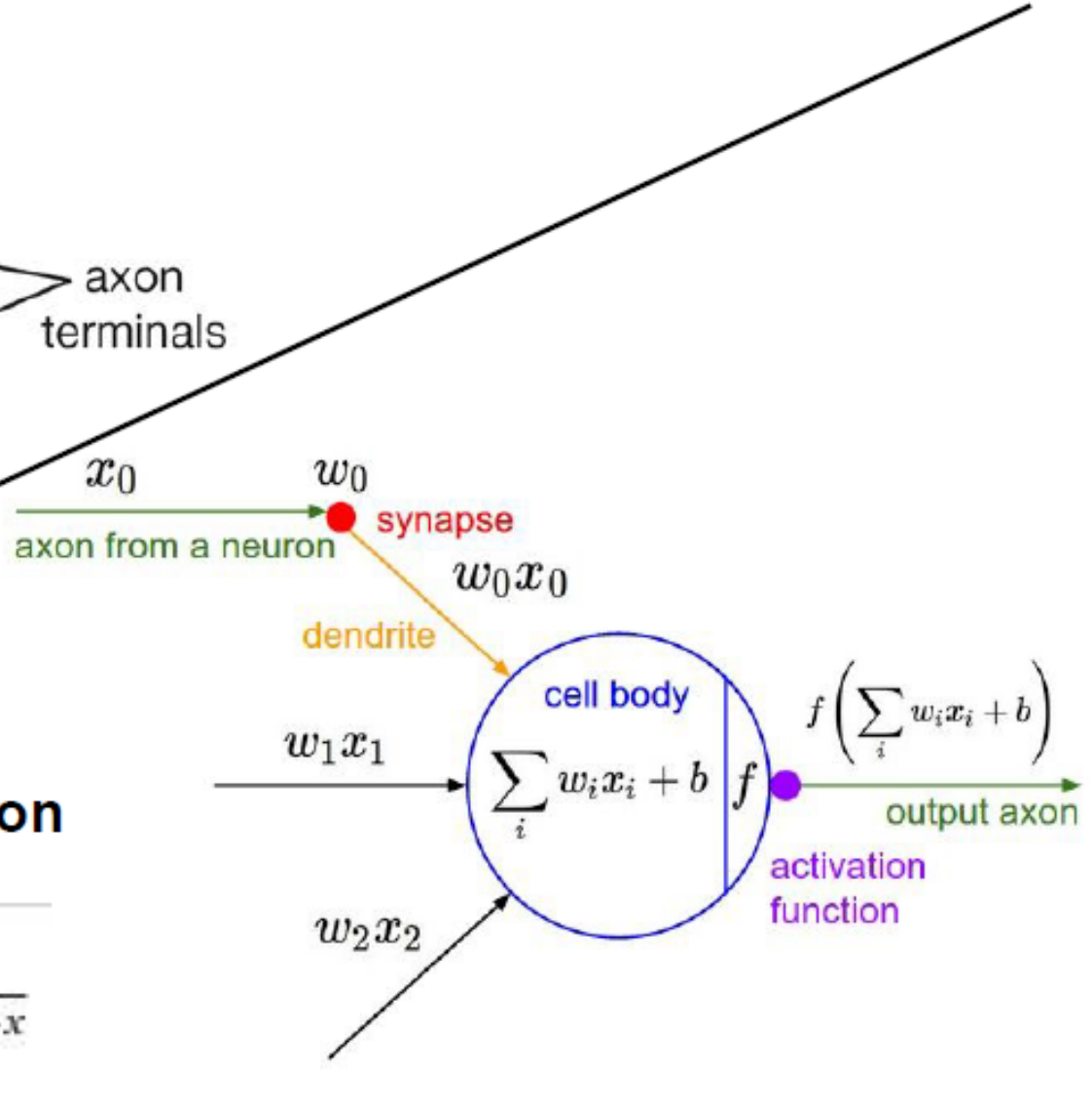
- 采用LSTM网络结构，鉴别效率相比于传统方法有显著的提高。国内暂时还没有此方面的研究进展。【Jet Flavor Classification in High-Energy Physics with Deep Neural Networks】



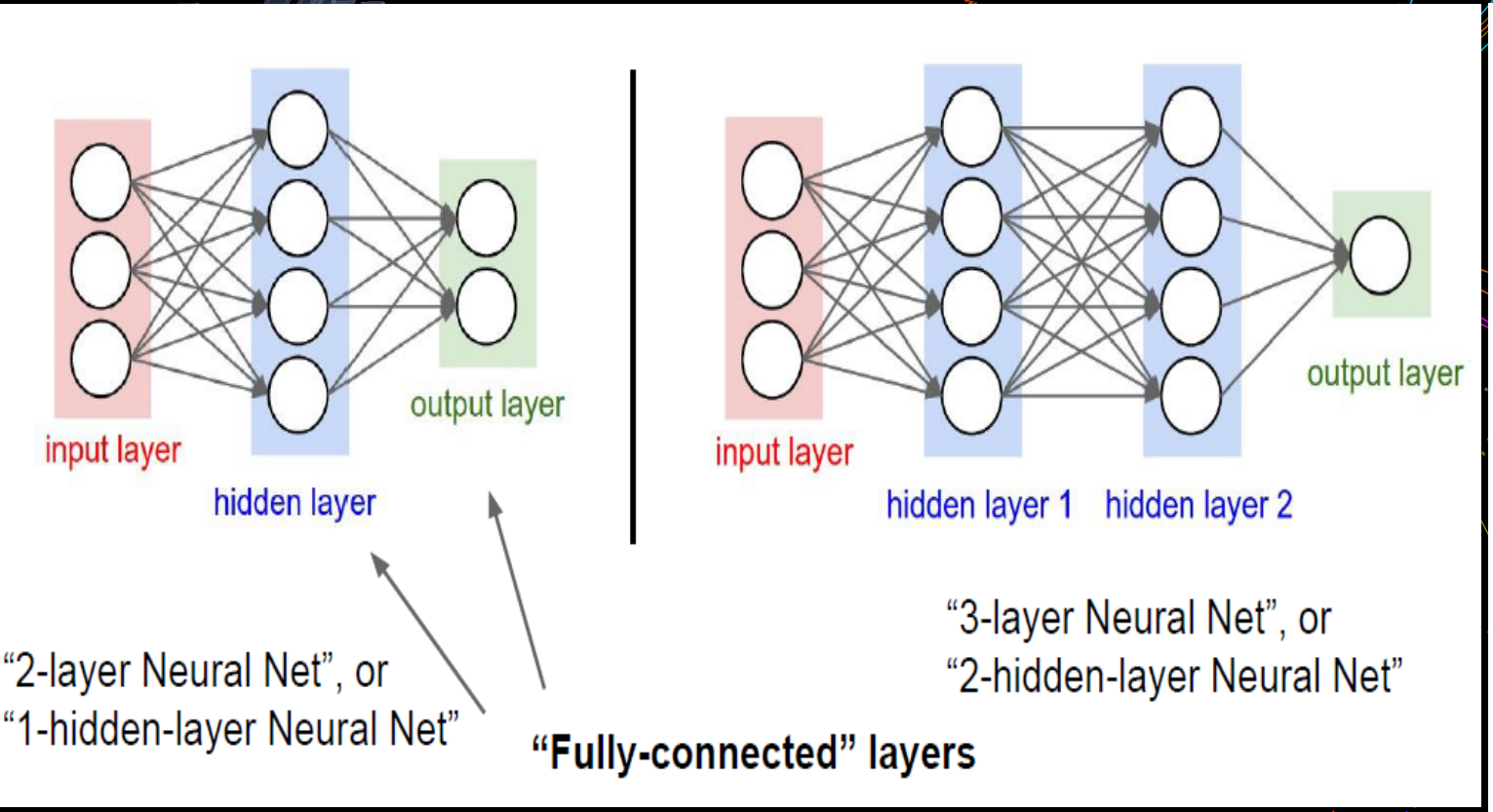


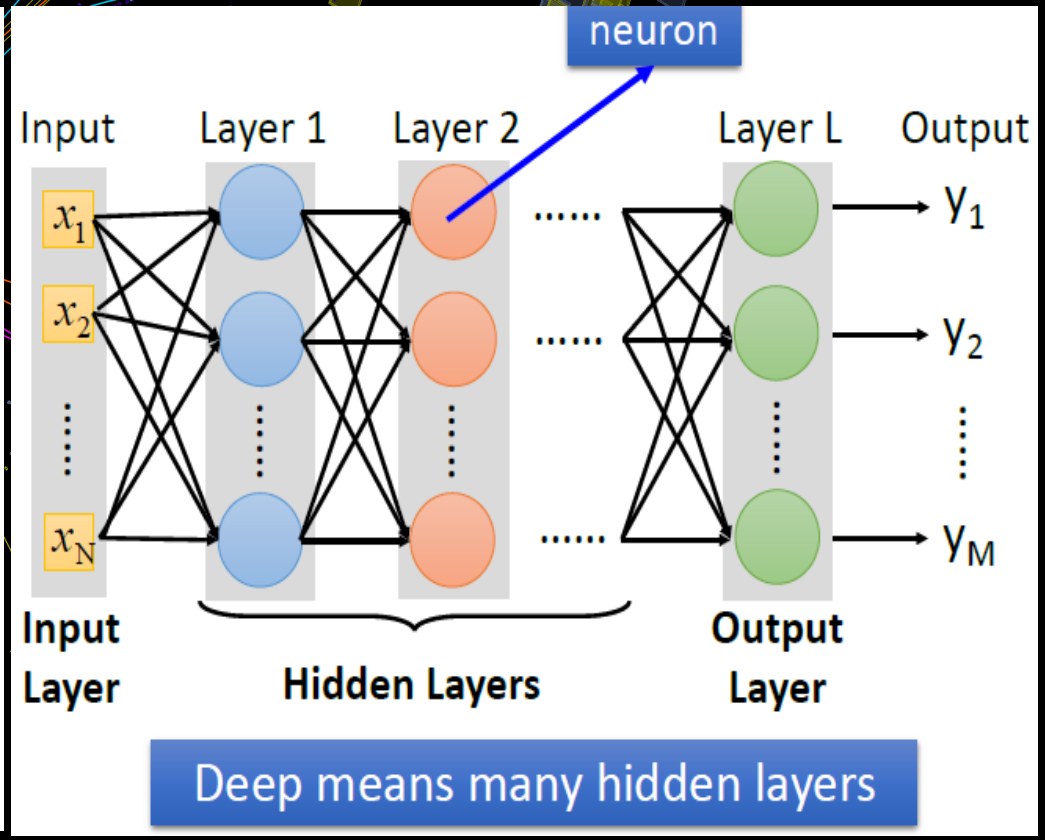
sigmoid activation function

$$\frac{1}{1+e^{-x}}$$



深度神经网络结构


input layer
hidden layer
output layer
input layer
hidden layer 1
hidden layer 2
output layer
"2-layer Neural Net", or
"1-hidden-layer Neural Net"
"3-layer Neural Net", or
"2-hidden-layer Neural Net"
"Fully-connected" layers


neuron
Input
Layer 1
Layer 2
Layer L
Output
x_1
x_2
x_N
y_1
y_2
y_M
Input Layer
Hidden Layers
Output Layer
Deep means many hidden layers

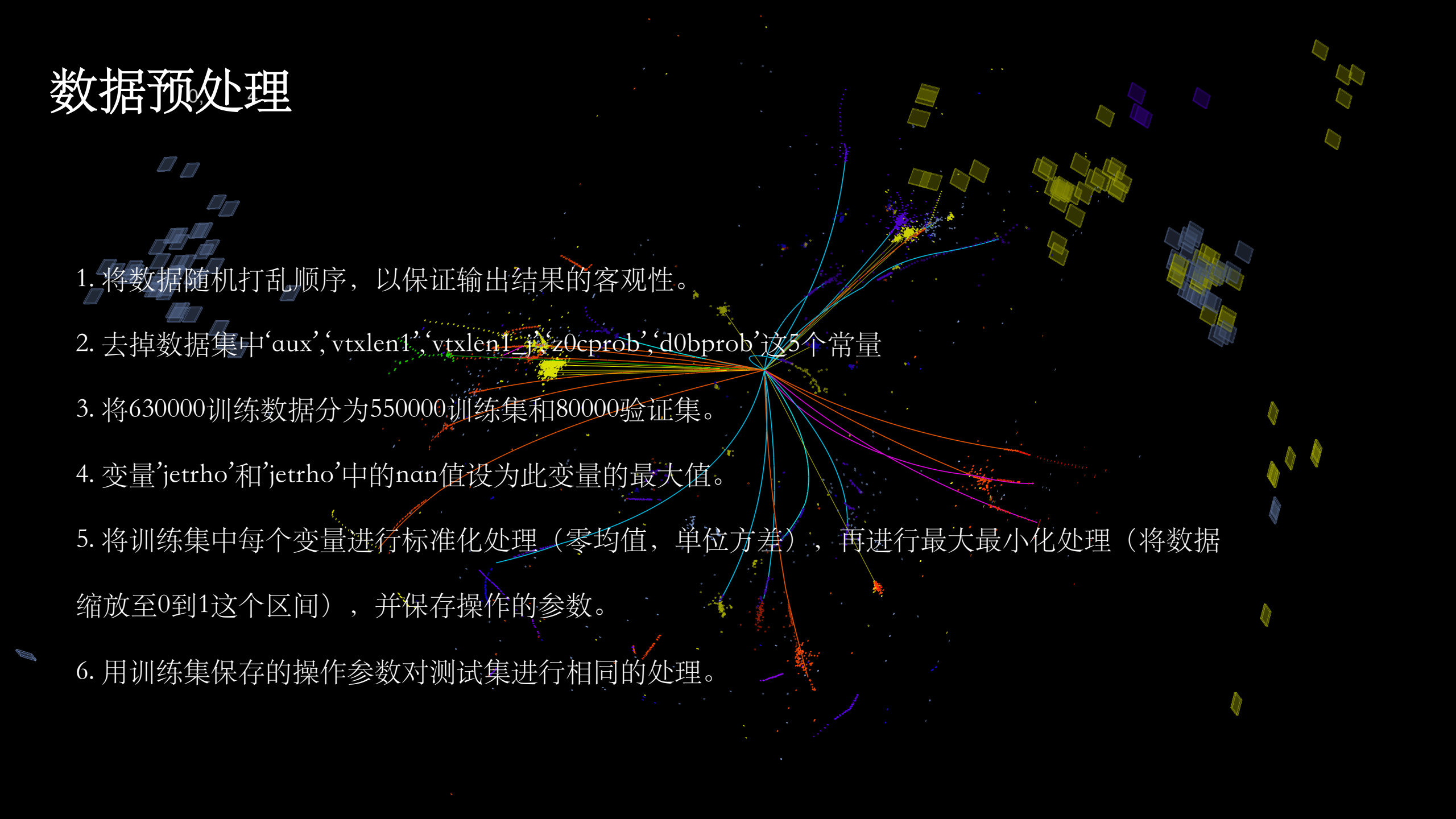
0, 4

使用全部变量作为输入

```
['aux' 'ntrkwitho' 'ntrk' 'nvtxall' 'vtxmassal' 'vtxlen12a' 'vtxlen12a'
 '1vtxprob' 'nvtx' 'jete' 'jetcosine' 'jeteta' 'vtxlen1' 'vtxlen2'
 'vtxlen12' 'vtxlen1_j' 'vtxlen2_j' 'vtxlen12_' 'vtxsig1' 'vtxsig2'
 'vtxsig12' 'vtxsig1_j' 'vtxsig2_j' 'vtxsig12_' 'vtxdirang' 'vtxdirang'
 'vtxdirang' 'vtxdirang' 'vtxdirang' 'vtxdirang' 'vtxmom' 'vtxmom1'
 'vtxmom2' 'vtxmom_je' 'vtxmom1_j' 'vtxmom2_j' 'vtxmass' 'vtxmass1'
 'vtxmass2' 'vtxmasspc' 'vtxmult' 'vtxmult1' 'vtxmult2' 'vtxprob'
 'trk1d0sig' 'trk2d0sig' 'trk1z0sig' 'trk2z0sig' 'trk1pt' 'trk2pt'
 'trk1pt_je' 'trk2pt_je' 'jprobr' 'jprobz' 'jprobr5si' 'jprobz5si'
 'sphericit' 'Fd' 'jetrho' 'trkmass' 'nmuon' 'nelectron' 'd0bprob'
 'd0cprob' 'd0qprob' 'z0bprob' 'z0cprob' 'z0qprob']
```

采用全连接型深度神经网络模型，目的是将喷注中的底夸克，璨夸克和其他粒子正确分类。

数据预处理



1. 将数据随机打乱顺序，以保证输出结果的客观性。
2. 去掉数据集中‘aux’,‘vtxlen1’,‘vtxlen1_j’,‘z0cprob’,‘d0bprob’这5个常量
3. 将630000训练数据分为550000训练集和80000验证集。
4. 变量’jetrho’和’jetrho’中的nan值设为此变量的最大值。
5. 将训练集中每个变量进行标准化处理（零均值，单位方差），再进行最大最小化处理（将数据缩放至0到1这个区间），并保存操作的参数。
6. 用训练集保存的操作参数对测试集进行相同的处理。

网络参数

0, 4

二.网络结构

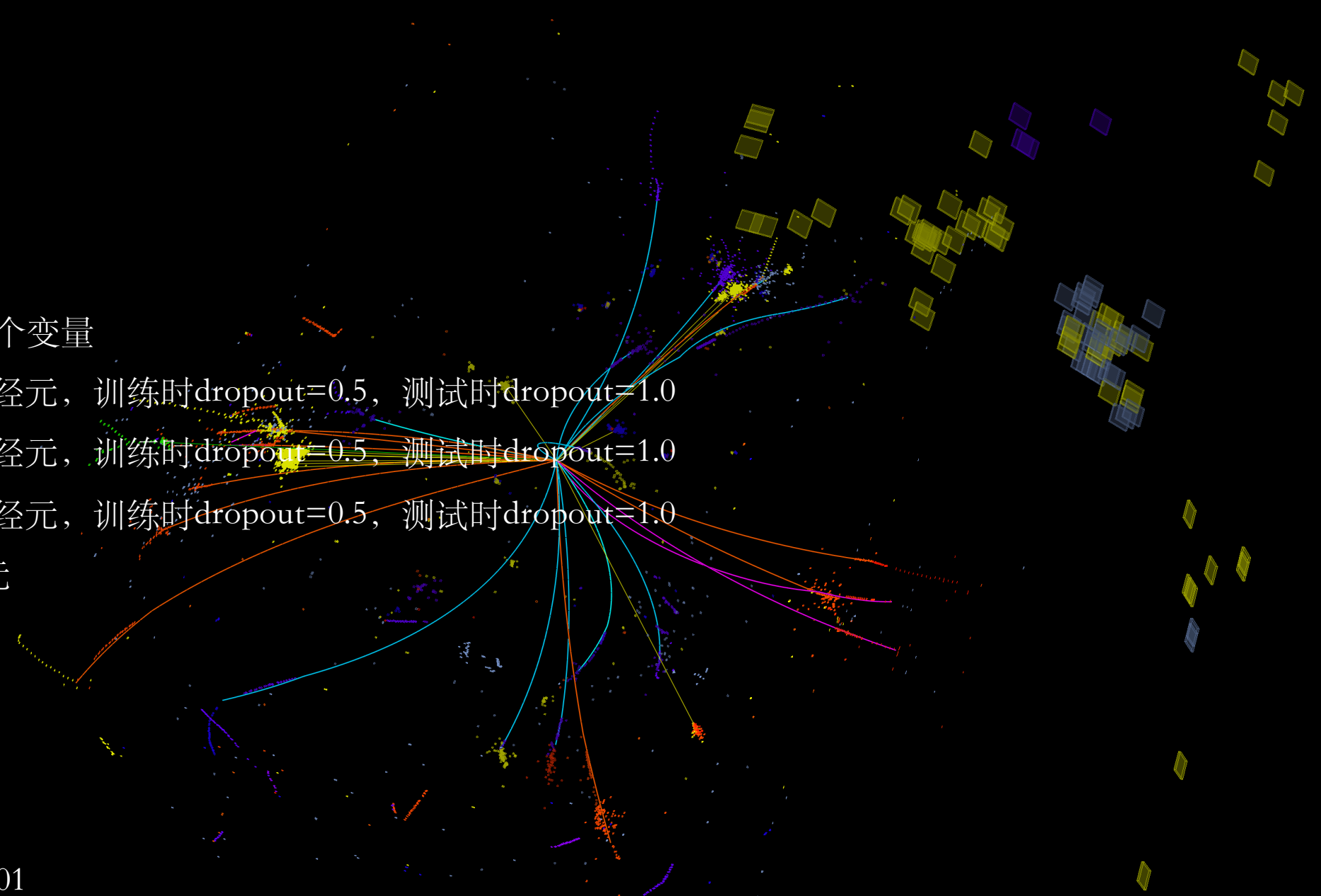
1.输入层：一事例的63个变量

2.隐藏层1：256个神经元，训练时dropout=0.5，测试时dropout=1.0

3.隐藏层2：256个神经元，训练时dropout=0.5，测试时dropout=1.0

4.隐藏层3：256个神经元，训练时dropout=0.5，测试时dropout=1.0

5.输出层：3个神经元



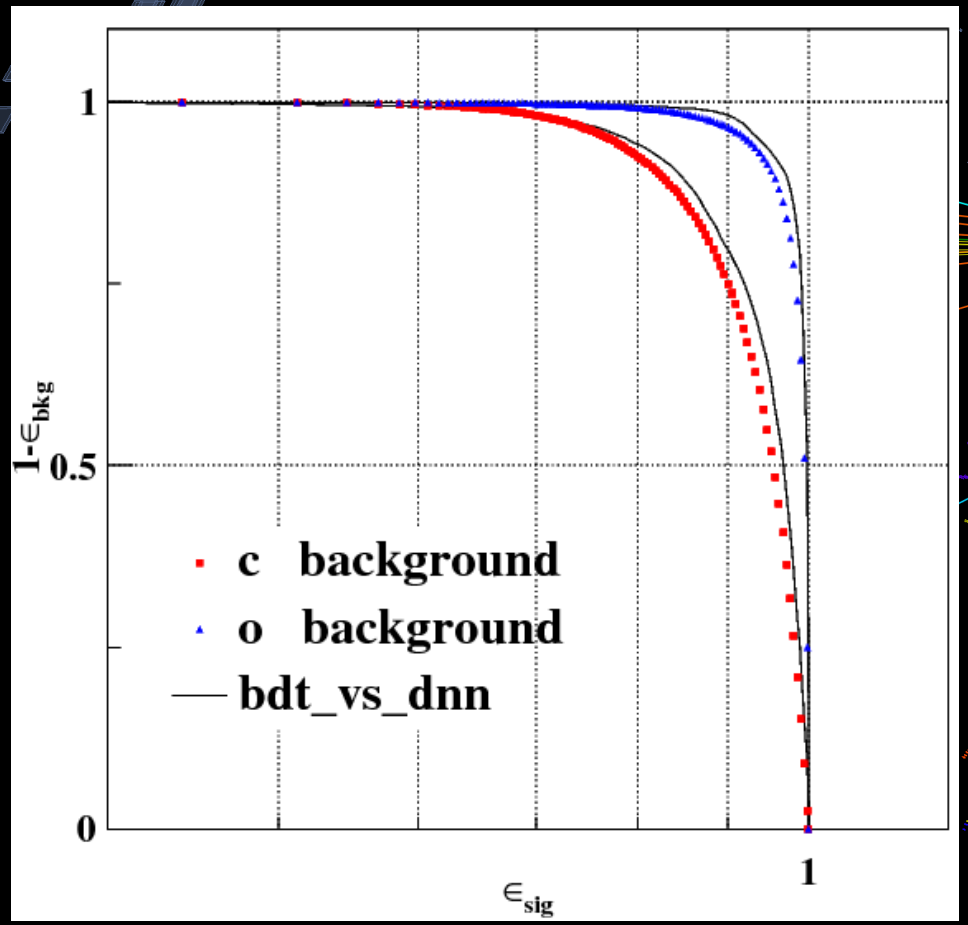
三：网络参数

1. batch size=100

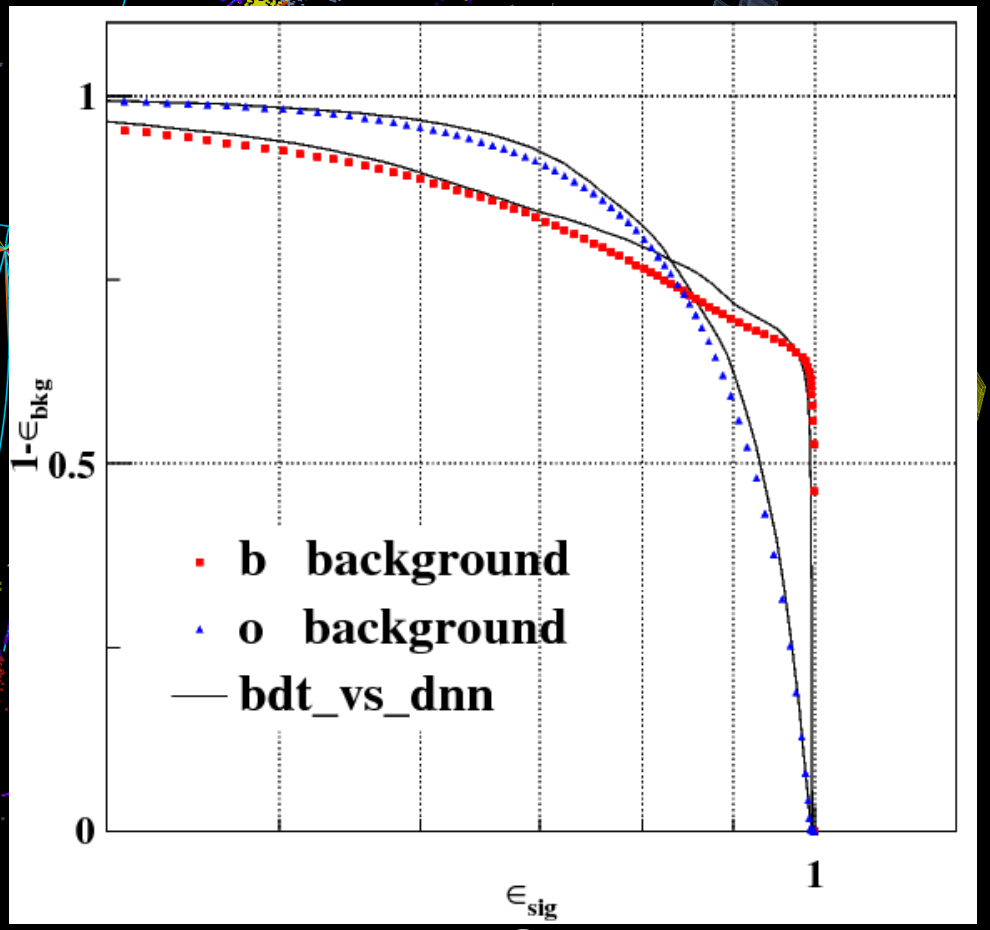
2. 迭代次数：500000

3. learning decay:0.0001

全连接神经网络与BDT算法对比：ROC 曲线，100k 样本



Btag



CTag

0, 4

全连接神经网络与BDT算法的信号效率-本底排除率对比

BTag

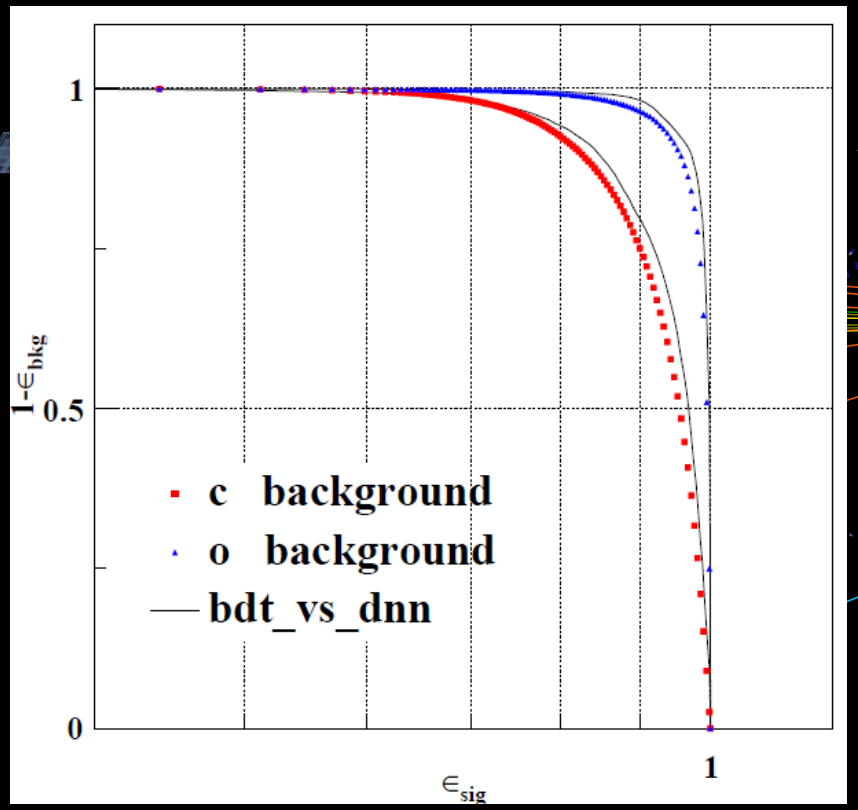
E_{sig}(b)	1-E_{bkg}(c, dnn)	1-E_{bkg}(c, bdt)	1-E_{bkg}(o, dnn)	1-E_{bkg}(o, bdt)
0.8	0.94	0.92-0.93	0.99	0.99
0.9	0.79-0.80	0.75-0.76	0.98	0.96-0.97
0.95	0.62-0.65	0.52-0.55	0.93-0.94	0.90-0.91

CTag

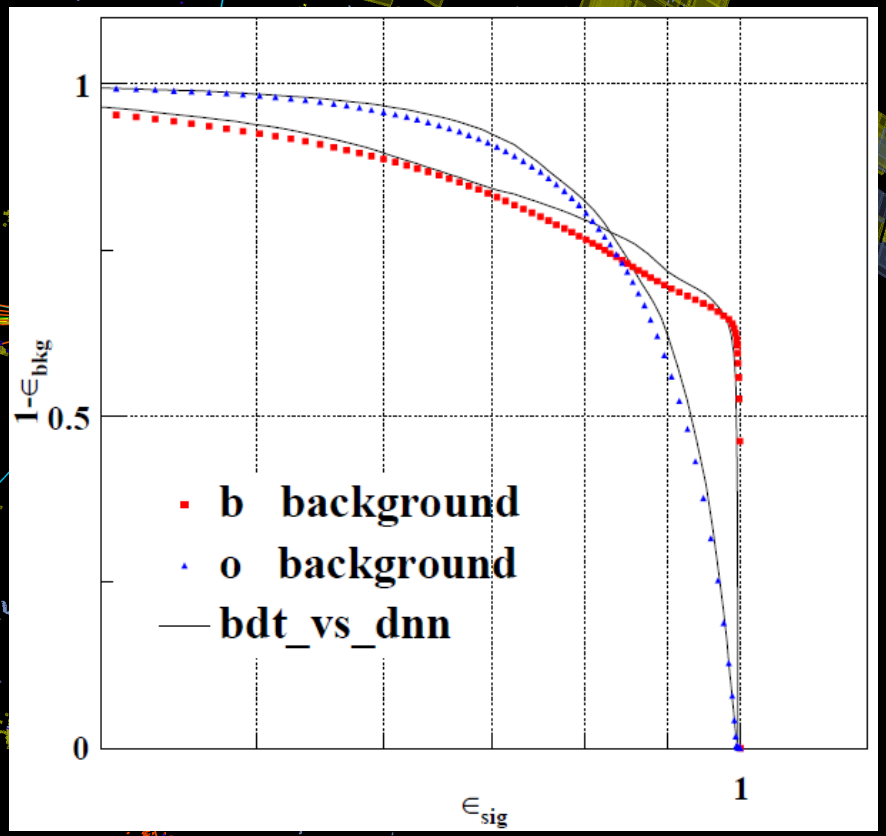
E_{sig}(c)	1-E_{bkg}(b, dnn)	1-E_{bkg}(b, bdt)	1-E_{bkg}(o, dnn)	1-E_{bkg}(o, bdt)
0.8	0.8	0.77	0.83	0.81
0.9	0.72	0.7	0.62	0.59
0.95	0.68	0.67	0.4	0.38

全连接神经网络与BDT算法对比：ROC 曲线，550k 样本

0, 4



b signal

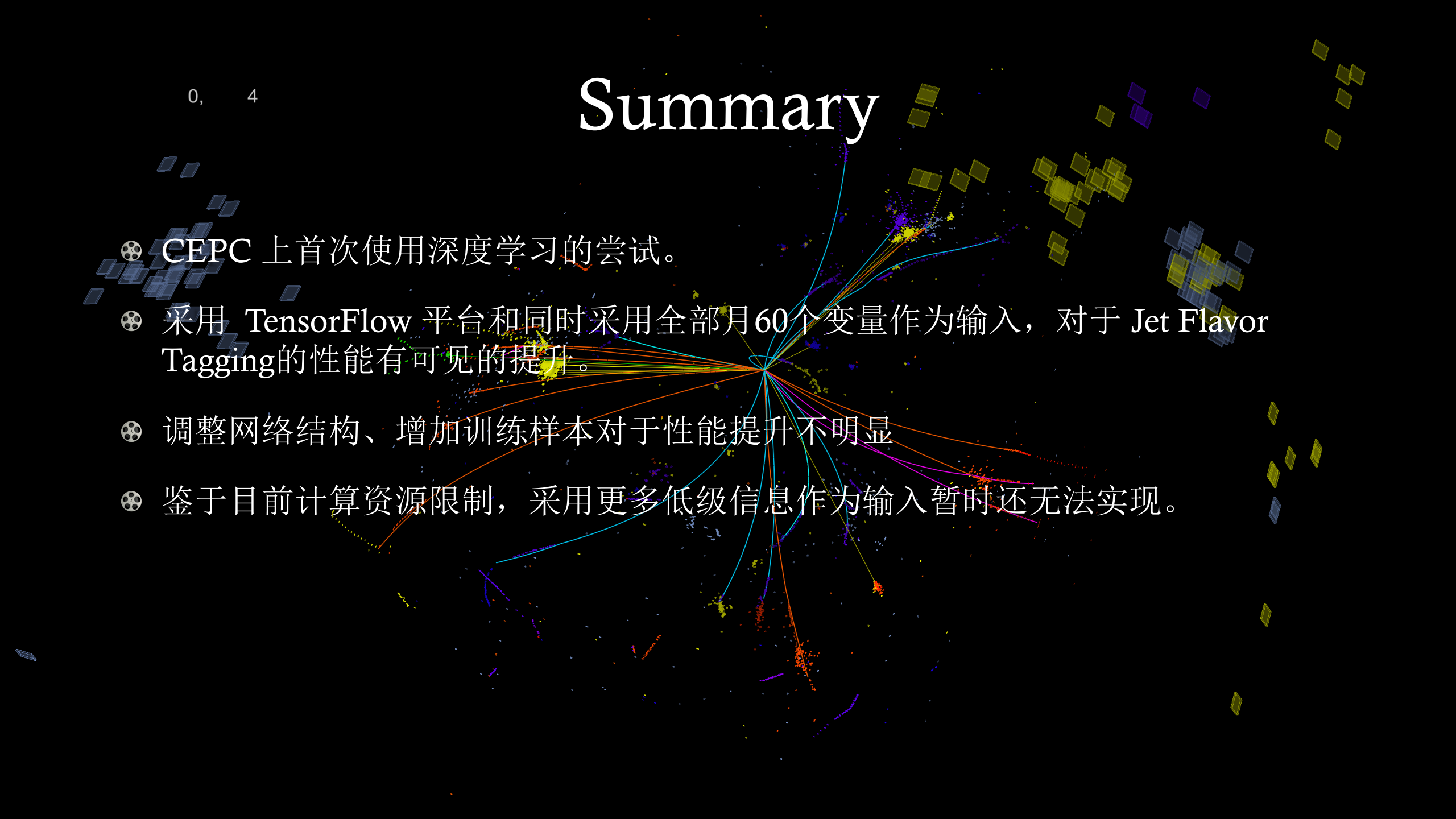


c signal

0, 4

Summary

- CEPC 上首次使用深度学习的尝试。
- 采用 TensorFlow 平台和同时采用全部月60个变量作为输入，对于 Jet Flavor Tagging的性能有可见的提升。
- 调整网络结构、增加训练样本对于性能提升不明显
- 鉴于目前计算资源限制，采用更多低级信息作为输入暂时还无法实现。



0, 4

计划和展望

- ✓ DL 在高能数据处理中非常有前景：模式识别和分类，可以引入业界的最新发展
- ✓ 采用 Jet Flavor Tagging 作为入口，掌握DL 工具和 TensorFlow 等平台的使用，积累经验。
- ✓ 未来设想
  - ☐ 尝试区分 gluon 和 uds jet
  - ☐ 尝试用 DL 做 CEPC 的 track finding：数据集已经准备好了：有难度
  - ☐ 尝试用 DL 在 CEPC 和 BESIII 的 PID上的性能：相对可行
- ✓ 有关的NSFC项目申请，建立大点的硬件平台
- ✓ 积累经验
- ✓ 培养人才