昆明LHAASO合作组会议

利用人工神经网络及决策树研究WCDA质子伽马区分

南开大学/高能所

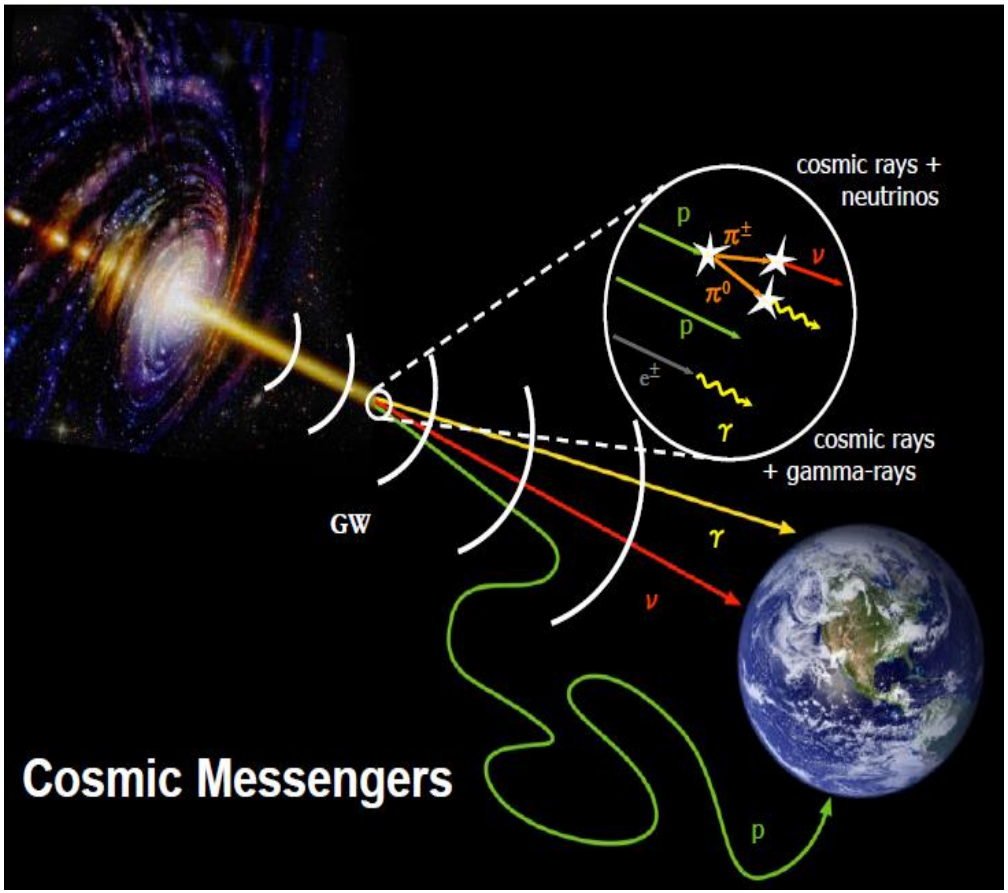
廖文英

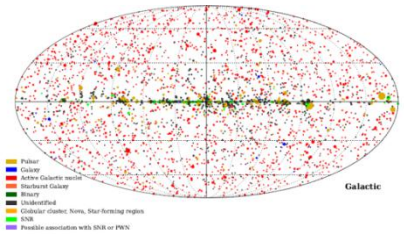
昆明，17-20/01/2017

内容提要

- WCDA伽马/质子区分背景介绍
- 特征敏感参数及结果
- 人工神经网络及决策树多参量分析
- 总结

研究背景和意义





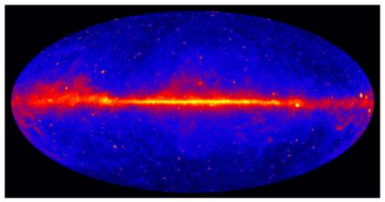


Figure 1.5 Counts Map > 1 GeV with five years of *Fermi*-LAT data. Image credit: NASA (see Appendix A)

LHAASO-WCDA 主要科学目标(伽马天文)：

- VHE γ 源巡天扫描(100GeV – 30TeV)
  - 河外源/耀变源；
  - GRB；
  - 河内源；
  - 其他.

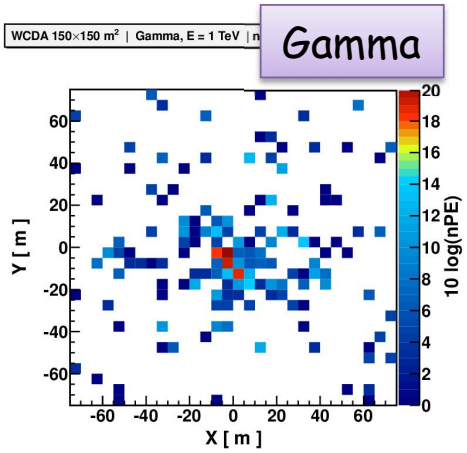
LHAASO-WCDA特点：

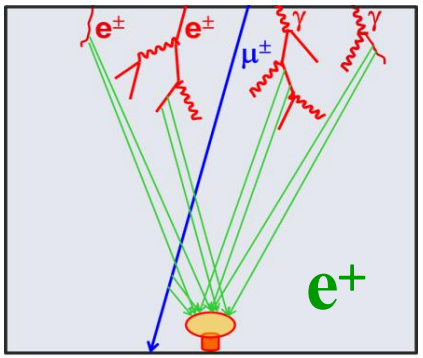
- 全天候；
- 宽视场；
- 低阈能；
- 高灵敏度；

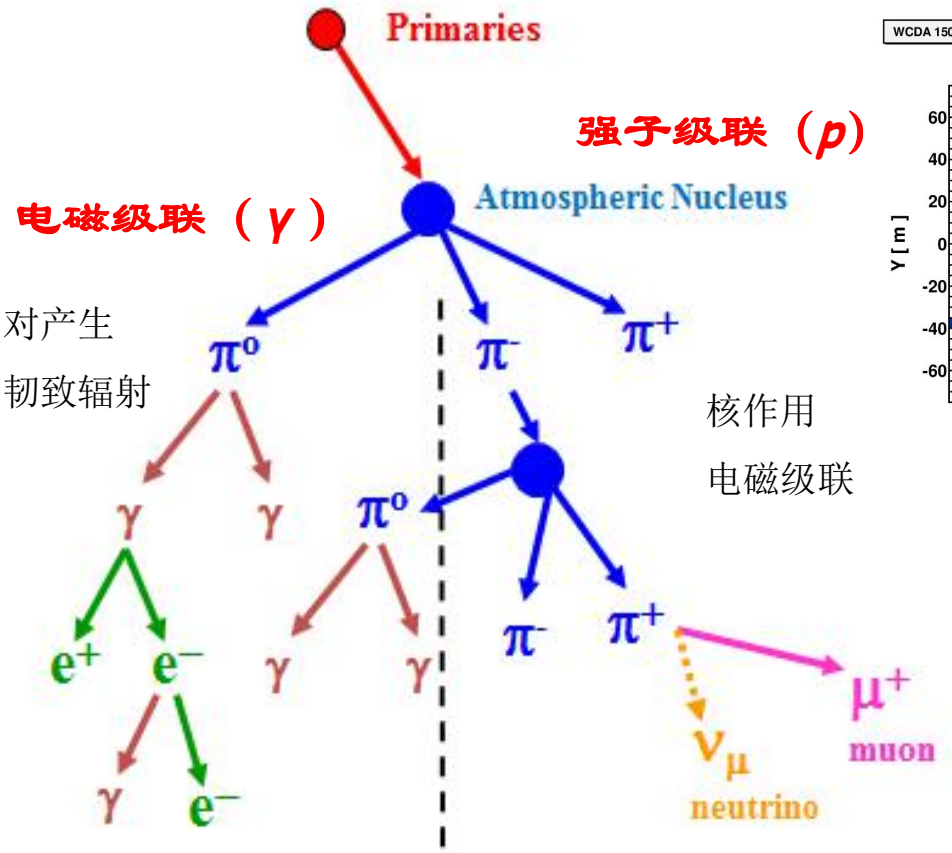
信号： γ 线；
背景： 其他宇宙线（质子p）；

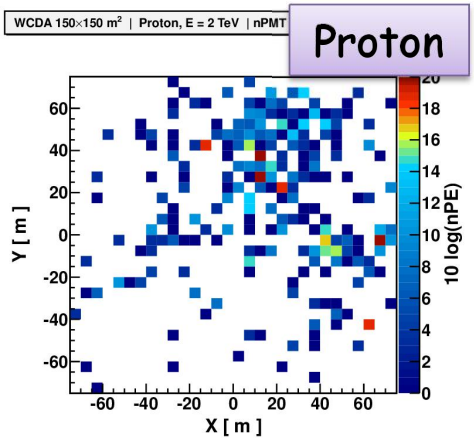
排除宇宙线本底噪声（γ/p 区分）至关重要

伽马/质子区分原理及方法





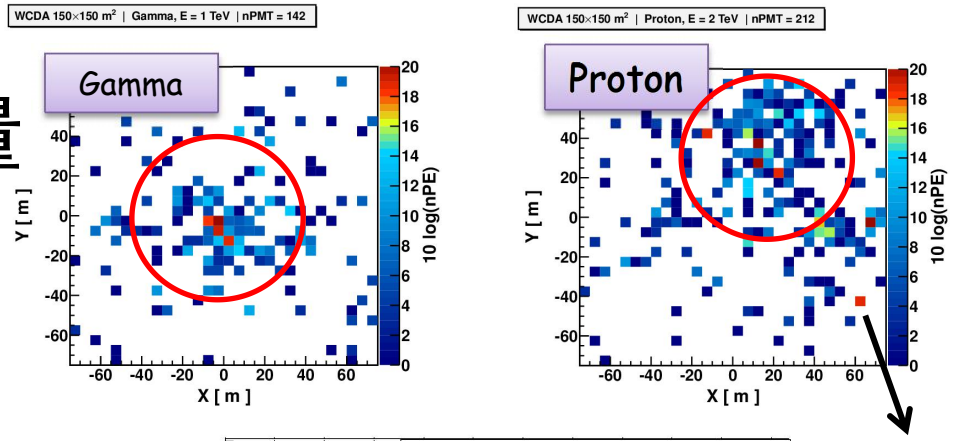




μ子与次芯粒子外产生大信号

次级粒子分布：	单芯对称；集中分布；	多芯或单芯不对称；杂乱分布；
主要成分：	e± ; γ ;	e±; γ; 核子; **μ子；**
首次作用点：	离地面较低；	离地面较高；

伽马/质子区分敏感参量

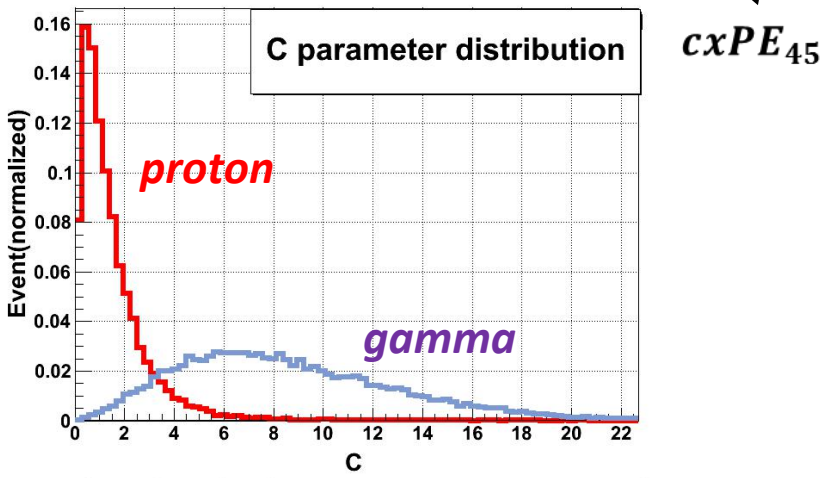


利用μ子信息和次芯结构

1、 $Compatness(C)=\frac{nFit}{cxPE_{45}}$

$nFit$: 参与重建着火PMT个数；

$cxPE_{45}$: 重建芯位45m之外着火PMT最大光电子数；

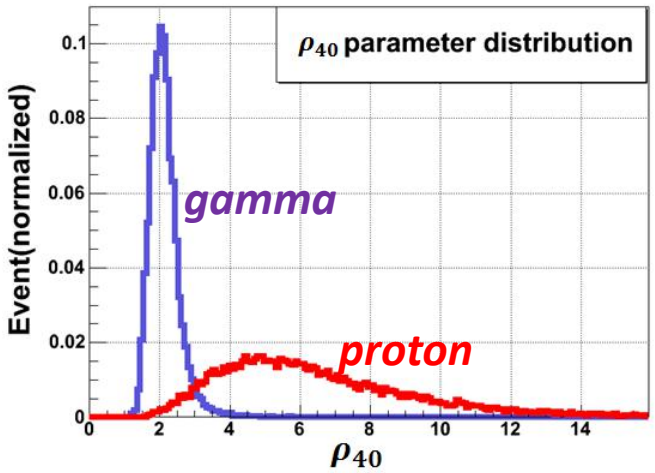


$cxPE_{45}$

2、 $\text{Density out}(\rho_{40})=\frac{\sum PE_{40}}{\sum PMT_{40}}$

$\sum PE_{40}$： 芯位40m之外所有光电子数；

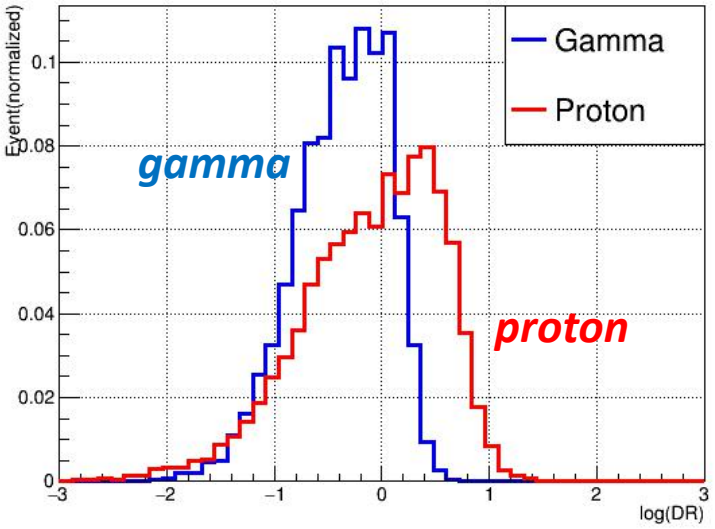
$\sum PMT_{40}$： 芯位40m之外所有着火PMT个数；



伽马/质子区分敏感参量

3、 Density raito(DR）

$$\mathrm{DR} = \frac{\sum PE_{50\_out} / \sum PMT_{50\_out}}{\sum PE_{10\_in} / \sum PMT_{10\_in}}$$

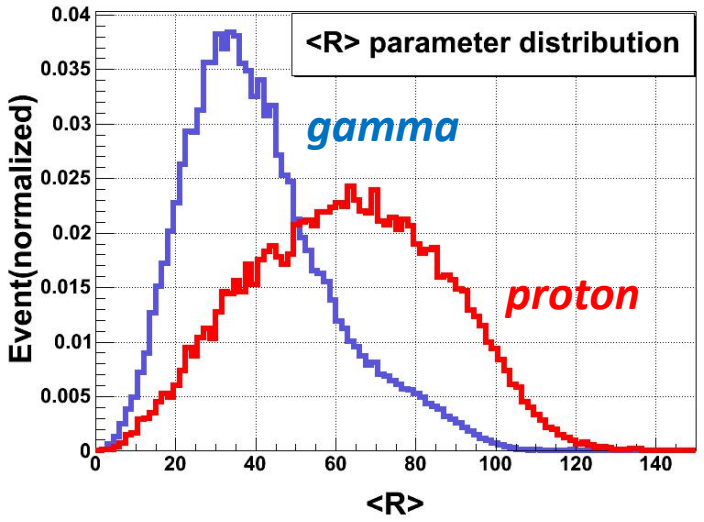


利用簇射横向分布情况

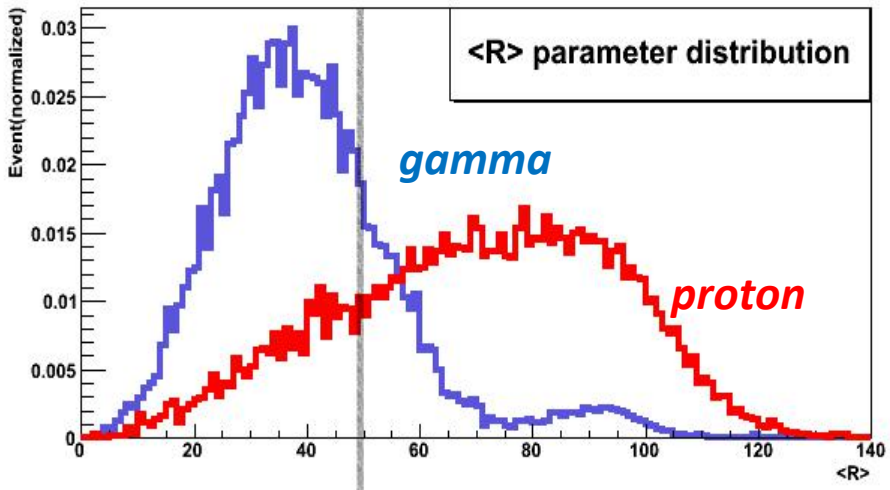
4、 $\langle R \rangle = \frac{\sum PE_i R_i}{\sum PE_i}$ （平均横向扩展半径）

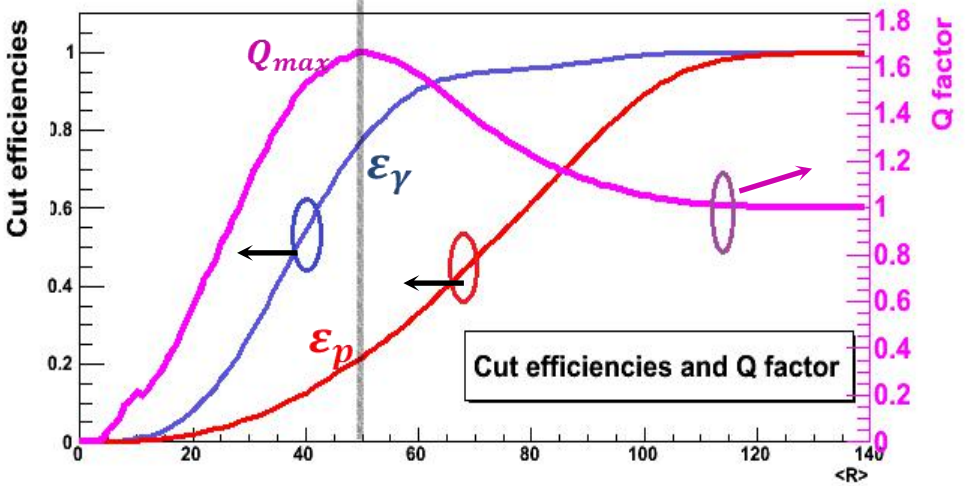
PE_i: 第i个着火PMT上光电子数；

R_i ： 第i个着火PMT 到芯位距离；



评价参数——Q 因子





$$Q = \frac{\varepsilon_\gamma}{\sqrt{\varepsilon_p}}$$

ε_γ:伽马事例保留效率

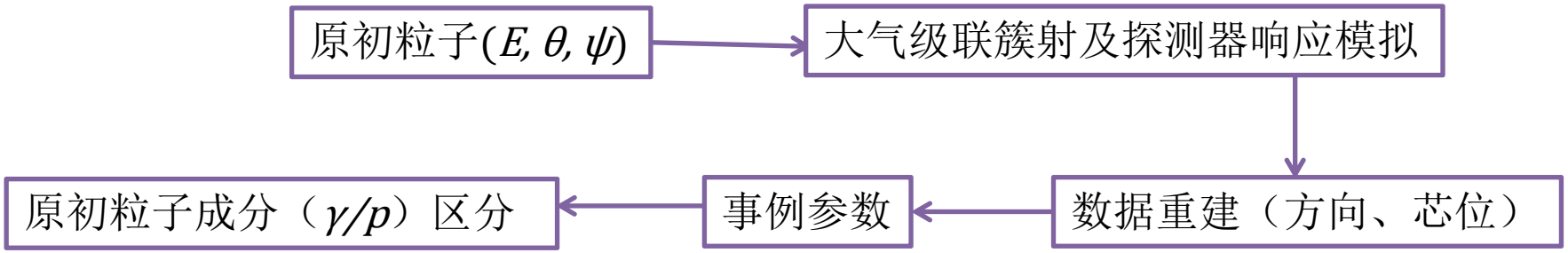
ε_p: 背景质子事例挑选剩余率

<R>=50 为最优挑选效率

Q_{max}=1.68
ε_γ=78%
ε_p=20%

排除80%背景噪声，同时保证78%的信号保留

伽马/质子区分步骤



蒙特卡罗数据模拟

- Site: YBJ @ 4300 m a.s.l.
- Code: Corsika 6720 + QGSJET-II (*was EPOS*) + GHEISHA
- Primary: point source (γ)
  Spectrum & Flux: Crab measured by *HEGRA (astro-ph/0407118)* ➔2.05×10^{-6} $(E/GeV)^{-2.62}$ $cm^{-2}s^{-1}GeV^{-1}$.
  Energy:
  6 segments: 10-20-50-100 GeV-1-10-100 TeV
  Event ratios: 0.05-0.1-0.5-1.0-0.5-1.0
- Primary: background(p)
  Spectrum & flux: *J.R. Hoerandel, Astroparticle Physics 19 (2003) 193-220*
  Energy: same, but min energy = min(10,1.1×A).
- Energy cuts:
  50 (hadron), 50 (muon), 0.3 (electron), 0.3 (photon / pion) MeV

事例筛选

筛选标准：

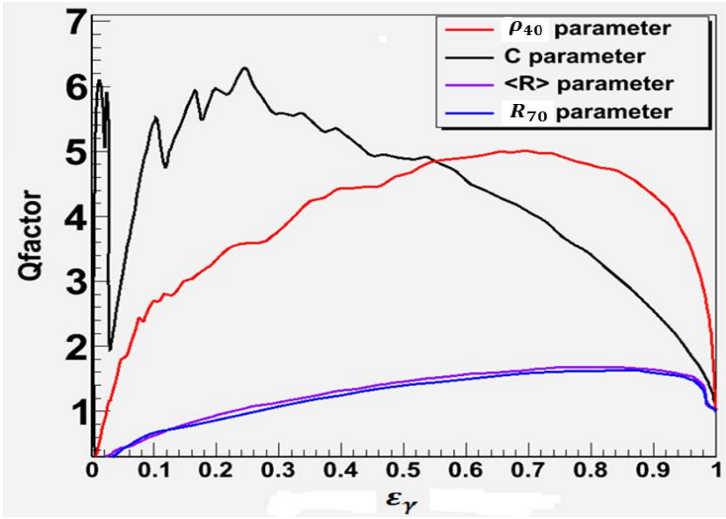
- 重建芯位位于阵列中心的
- 150x150 m^2范围内。
- 重建天顶角θ：0~45°
- 重建方位角ψ：0~360°
- 伽马事例保留>50%

单参量质子伽马区分结果

Energy = ~1 TeV

	ρ_{40}	C	$<R>$	DR
Q因子	5.012	4.91	1.6	1.49
ε_γ	69.8%	53.8%	77.5%	79.43%
ε_p	1.9%	1.2%	2.1%	0.46%

参量ρ_{40}和C：背景排除率98%



参量$<R>$和 DR：背景排除率70%~80%

nFit	$Events(\gamma)$	Events(p)	$< E_\gamma > (TeV)$	$< E_p > (TeV)$	$Q(DR)$	$Q(< R >)$	Q(C)	$Q(\rho_{40})$
10 – 20	162694	741041	0.16	0.53	1.23	1.43	2.19	2.18
20 – 50	293989	694448	0.28	0.67	1.31	1.55	2.52	2.57
50 – 100	192625	355761	0.62	1.22	1.47	1.62	4.23	4.46
100 – 200	153233	262216	1.25	2.35	1.64	1.63	6.48	8.42
200 – 400	154457	202647	2.57	4.58	1.69	1.56	12.16	15.54
400 – 800	17378	159789	5.99	10.16	1.58	1.49	21.75	17.61
> 800	34676	96004	17.64	27.49	1.43	1.44	12.69	9.24

缺点：在低能区域（<1TeV），区分能力有限。

各参量的Q因子和挑选效率

	C parameter			ρ_{40} parameter		
nFit	Q 因子	ε_γ	ε_p	Q 因子	ε_γ	ε_p
[10,20]	2.19	71.97%	10.81%	2.18	77.43%	12.67%
[20,50]	2.52	62.00%	6.07%	2.57	74.98%	8.53%
[50,100]	4.23	53.41%	1.59%	4.46	62.10%	1.94%
[100,200]	6.48	52.39%	0.65%	8.42	73.71%	0.77%
[200,400]	12.16	50.97%	0.18%	15.54	51.71%	0.11%
[800,~]	21.75	50.40%	0.05%	17.61	65.13%	0.14%

	<R> parameter			DR parameter		
nFit	Q 因子	ε_γ	ε_p	Q 因子	ε_γ	ε_p
[10,20]	1.43	69.99%	24.10%	1.13	96.92%	74.11%
[20,50]	1.62	78.08%	23.17%	1.18	95.49%	65.07%
[50,100]	1.60	81.46%	26.08%	1.26	84.58%	45.09%
[100,200]	1.53	82.50%	29.06%	1.44	79.43%	30.46%
[200,400]	1.49	84.27%	31.85%	1.55	85.52%	30.29%
[800,~]	1.47	79.03%	28.78%	1.53	79.50%	26.84%

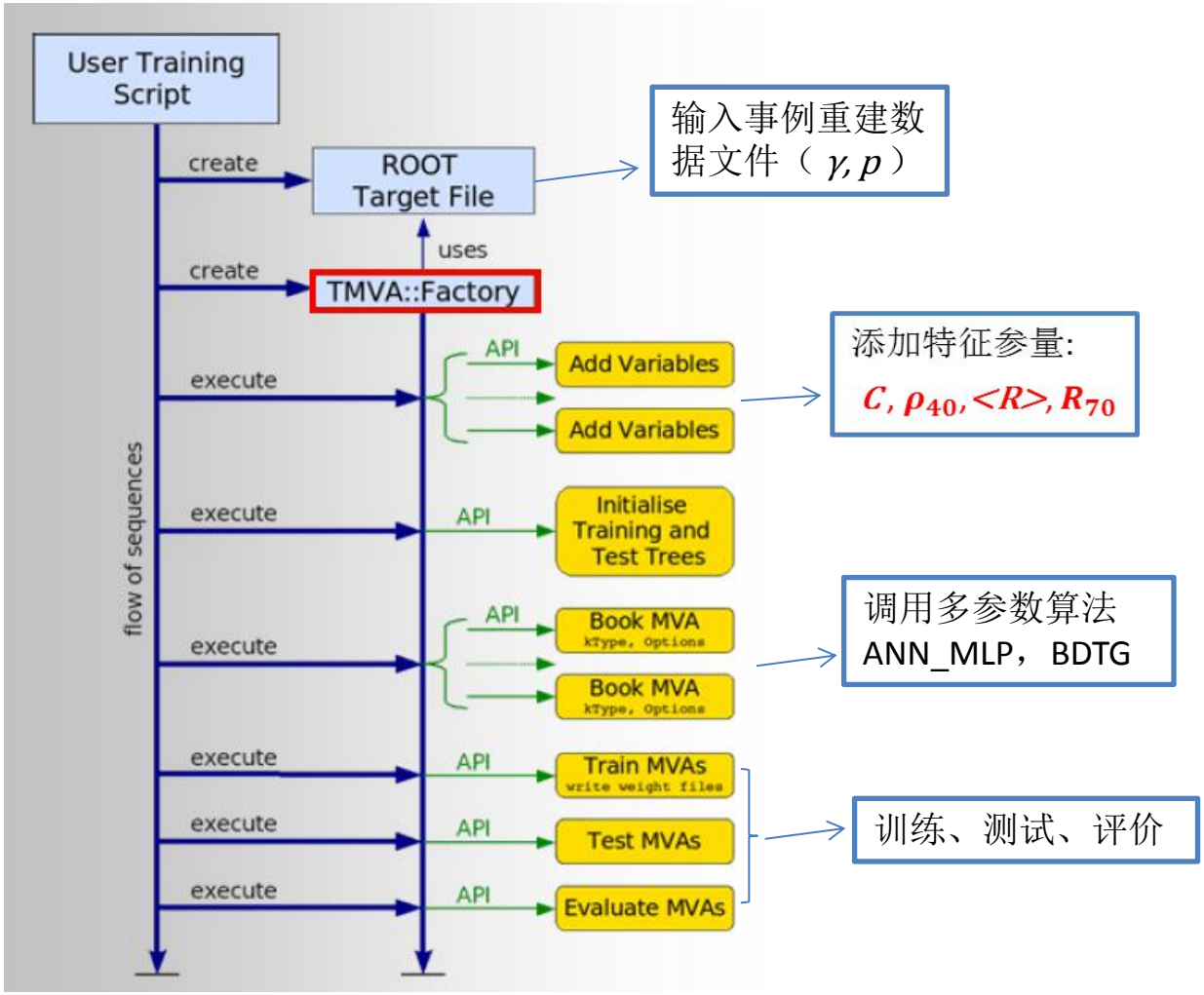
低能区域
质子剩余率
较高。

多参数分析——人工神经网络, 决策树

ROOT多参量分析包 (v5.14)——**TMVA**

TMVA 多参数分析软件

分析步骤：

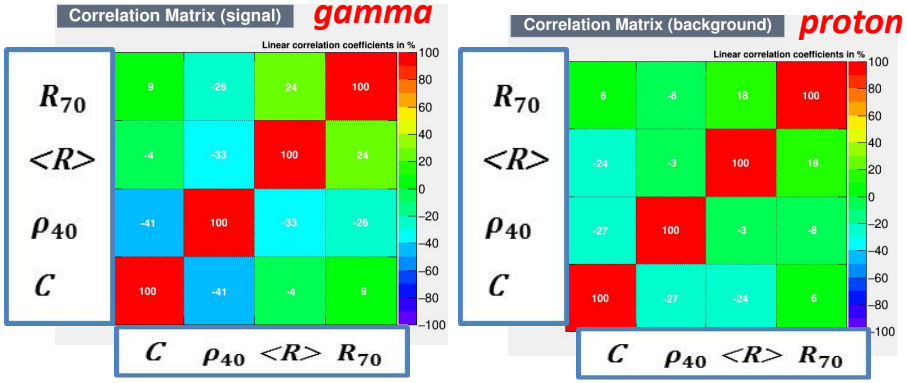


多参数分析算法：

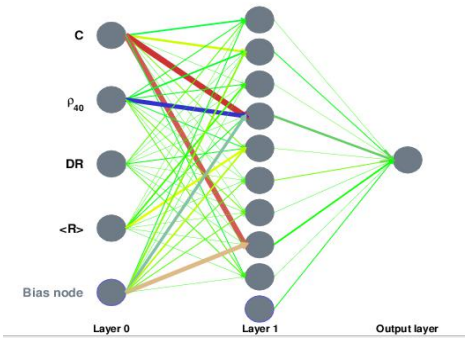
- 人工神经网络(ANN)
- 决策树 (BDTG)

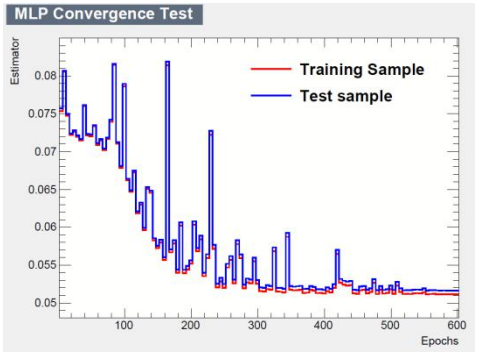
人工神经网络多参数分析结果

1、 输入参数关联矩阵



2、 神经网络及训练测试收敛情况





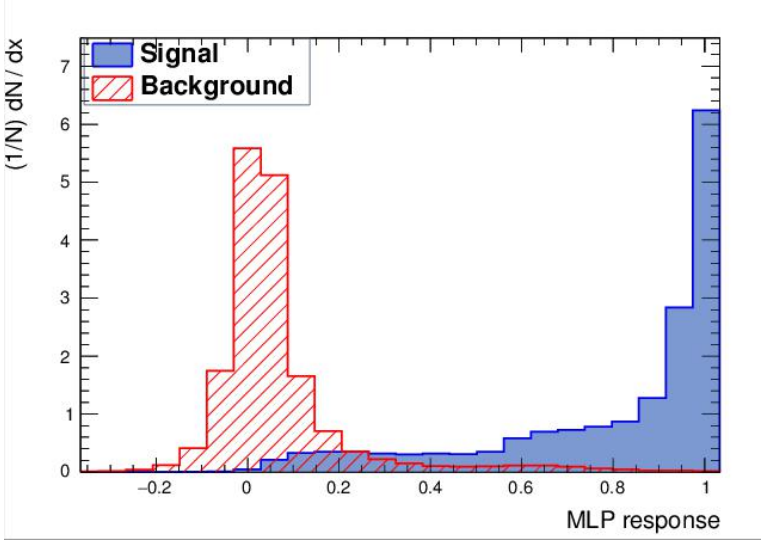
C parameter

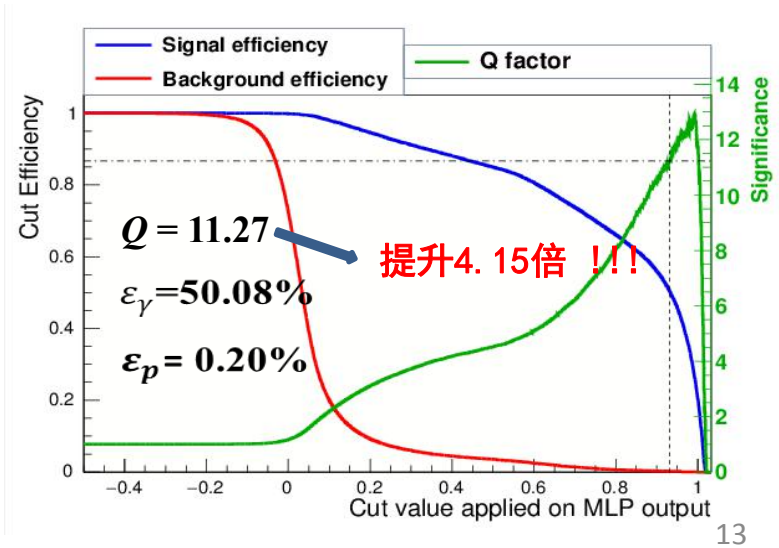
$Q = 2.19$

$\varepsilon_\gamma = 71.97\%$

$\varepsilon_p = 10.81\%$

3、 ANN算法输出结果（10<*nFit*<20）





人工神经网络及决策树优缺点

人工神经网络

优点

- 分类的准确度高
- 并行分布处理能力强
- 联想记忆的功能
- 对噪声神经的自适应、自组织性及容错性能力强

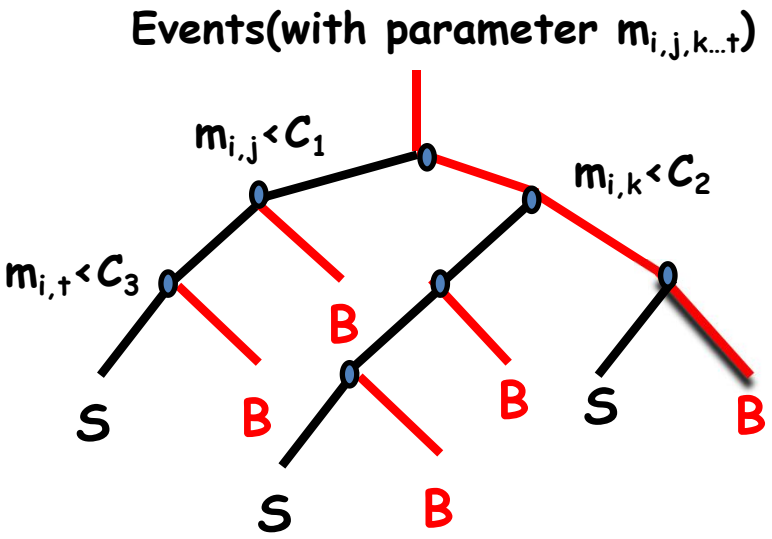
缺点

- 较长的收敛和训练时间
- 黑盒，内部规则可理解性差
- 性能依赖网络结构设计和学习算法

决策树（BDTG）

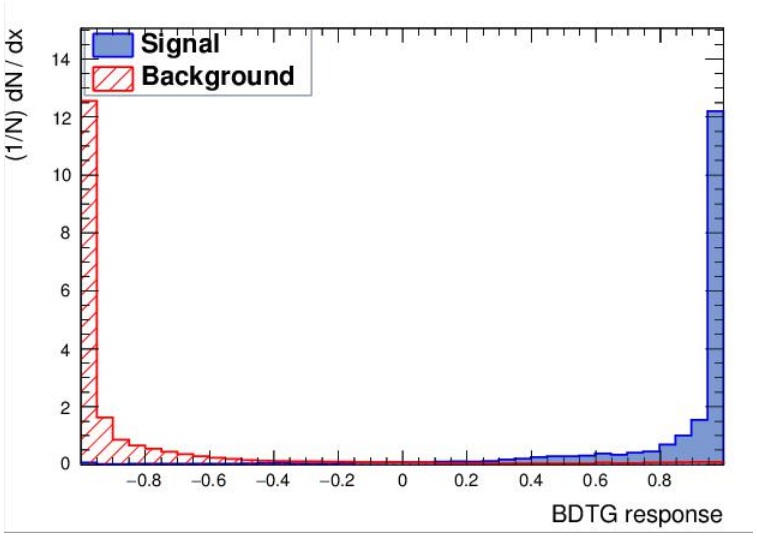
- 精度较高
- 数据训练和构造时间短
- 白盒，数据规则可视化，容易理解
- 忽略数据集中属性之间的相关性
- 各类别样本数量不一致时，结果偏向于更多数值的类型
- 过度拟合
- 多变量组合发现规则
- 不同决策树分支之间的分裂不平滑

决策树多参数分析结果

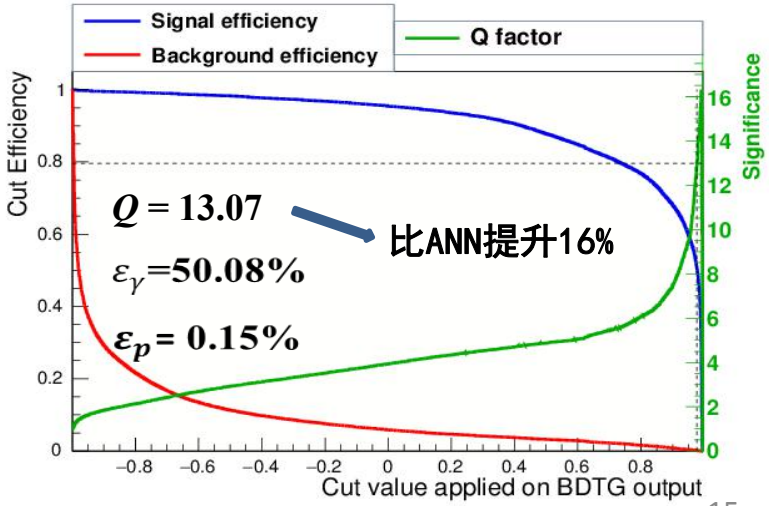


决策树示意图

BDTG算法输出结果（10<nFit<20）



ANN
- Q = 11.27
- ε_γ = 50.08%
- ε_p = 0.20%



人工神经网络及决策树运行时间

人工神经网络ANN

Elapsed time for training with 185993 events: 767 sec

Elapsed time for evaluation of 185993 events: 0.515 sec

事例数越多，
运行时间越长。

决策树BDTG

Elapsed time for training with 185993 events: 222 sec

Elapsed time for evaluation of 185993 events: 0.476 sec

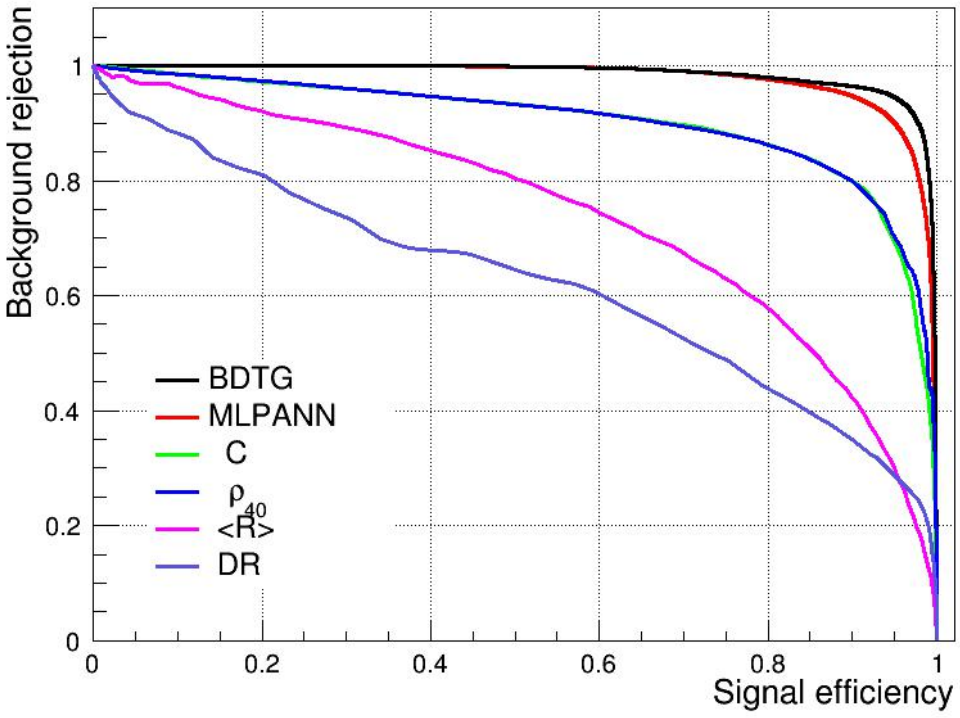
运行时间比较：

训练时间>>测试和评价时间

人工神经网络>决策树

多参数分析与单参数区分结果对比

10<nFit<20：不同区分方法结果对比



	BDTG	ANN	C
Qmax	13.07	11.27	2.19
ε_γ	50.08%	50.08%	71.97%
ε_p	0.15%	0.20%	10.81%

多参数分析提高原因：

背景事例排除效率增加而信号丢失相对较少。

人工神经网络多参数分析结果及对比

nFit	ANNMLP			BDTG		
	ε_γ	ε_p	Q	ε_γ	ε_p	Q
10~20	50.08%	0.20%	11.27	50.08%	0.15%	13.07
20~50	51.23%	0.16%	12.63	50.03%	0.15%	13.12
50~100	63.02%	0.16%	15.96	51.95%	0.08%	17.92
100~200	61.34%	0.05%	26.59	57.00%	0.05%	24.32
200~400	56.57%	0.04%	30.16	64.14%	0.04%	30.86
400~800	59.55%	0.05%	25.73	51.81%	0.03%	32.04
800~3600	63.03%	0.15%	16.29	51.01%	0.04%	26.37

- 在低能区域，利用多参数分析法极大的增加**Q**值，大大提高背景排除能力（**89% → 99.8%**）。
- 决策树的区分能力略优于人工神经网络

总结

- LHAASO-WCDA具有很好的质子伽马区分能力。

- 利用簇射muon信息和次芯结构能有效进行质子伽马区分。

  Energy >1TeV: 参数C和 ρ_{40} 的Q因子 > 5，排除98%背景事例。

- 利用神经网络及决策树多参数分析的Q因子可提高4倍，优于单参数分析，极大的提升了WCDA在低能区域的质子伽马区分能力。

- 决策树的质子伽马区分能力略优于人工神经网络，但其运行时间快3倍。

- 下一步计划优化和细化多参数分析，计算利用该方法下WCDA的灵敏度。

谢谢！