Gamma/Proton discrimination for LHAASO-WCDA

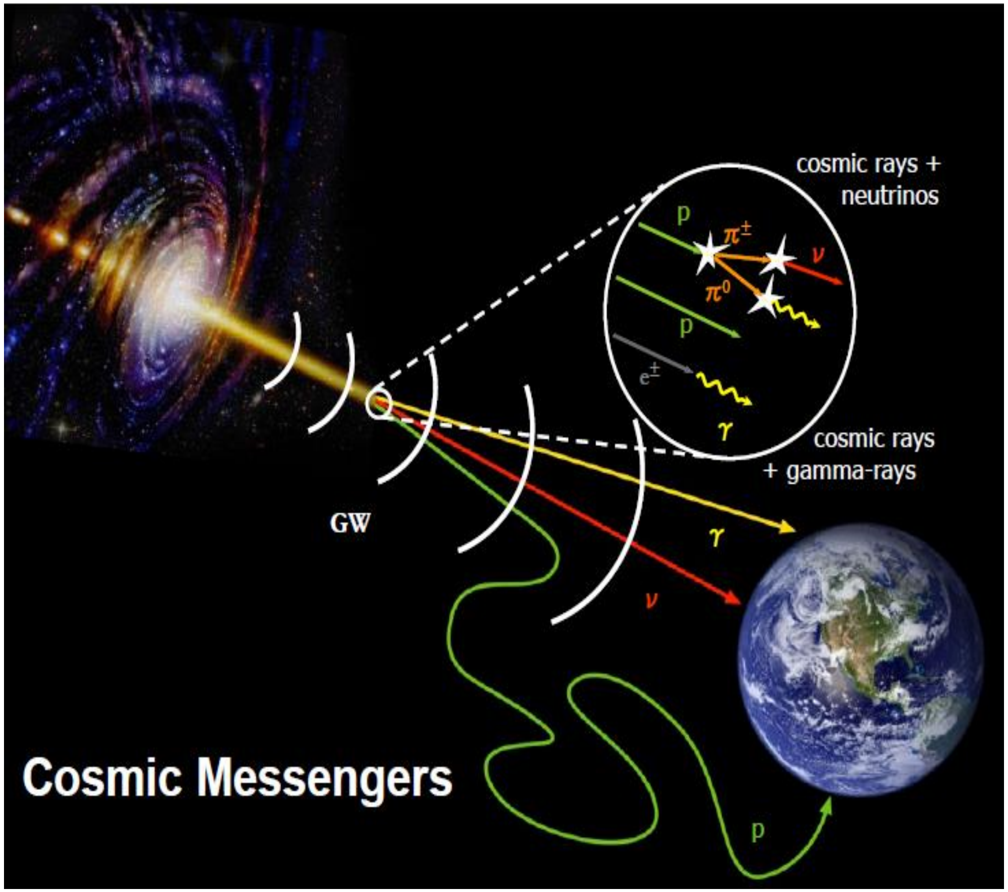
WANG XIAOJIE

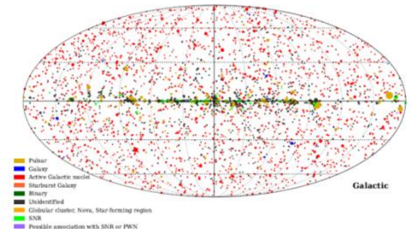
Nanjing , 13-15/04/2019

Outline

- Gamma/Proton background introduction
- Methods
- Parameters & results
- Multivariate analysis
- Summary

Background





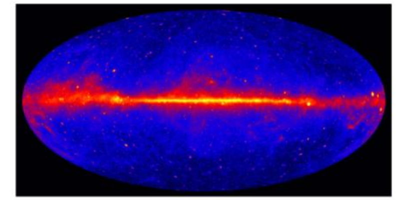


Figure 1.5 Counts Map > 1 GeV with five years of *Fermi*-LAT data. Image credit: NASA (see Appendix A)

LHAASO-WCDA scientific goals (gamma ray astronomy) :

- VHE γ-ray sources full sky survey(100GeV – 30TeV)
  - ➢ extragalactic sources/transenit sources ;
  - ➢ GRB ;
  - ➢ galactic source

LHAASO-WCDA :

- ➢ **full sky ;**
- ➢ **Wide FOV ;**
- ➢ **Low threshold ;**
- ➢ **High sensitivity ;**

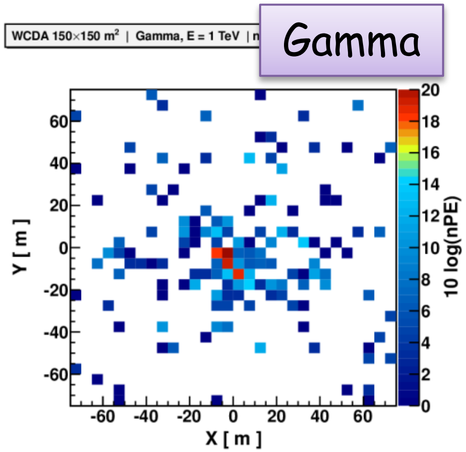
Signal : gamma-ray ;
background : hadrons ;

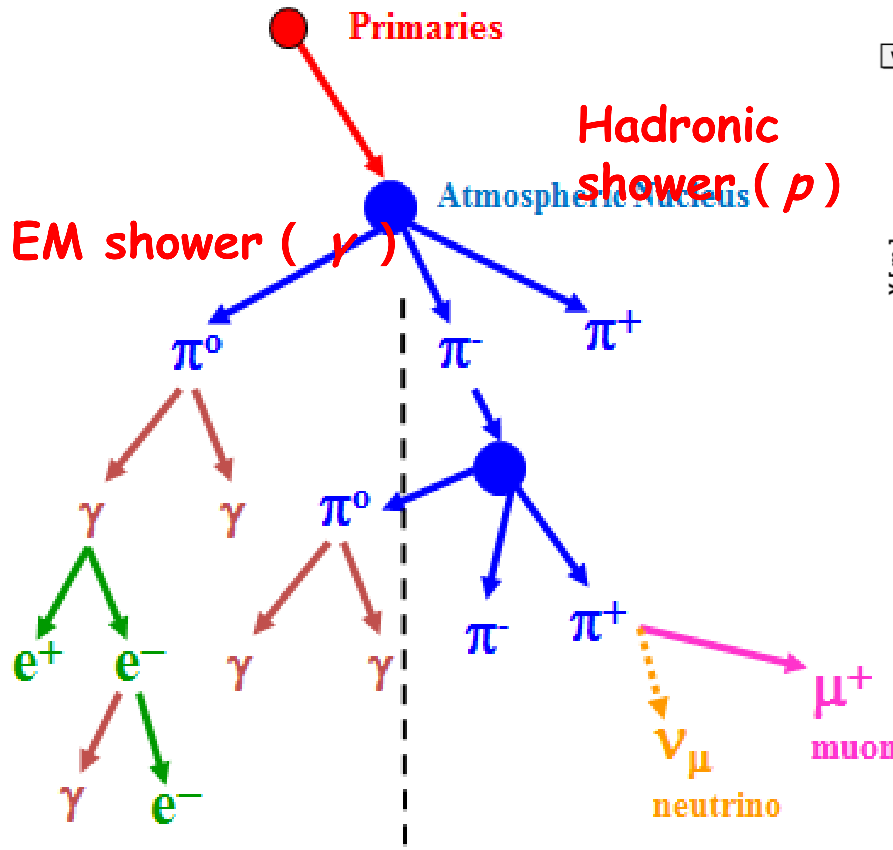
Background rejection

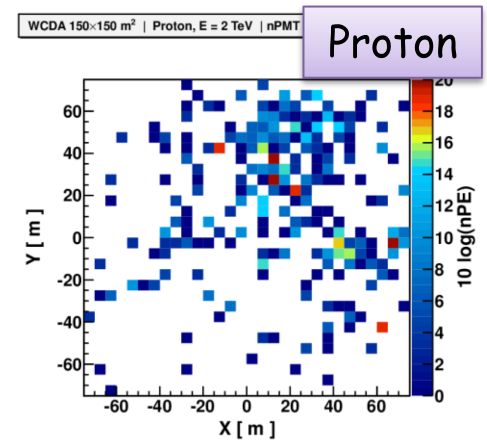
(*gamma/proton discrimination*)

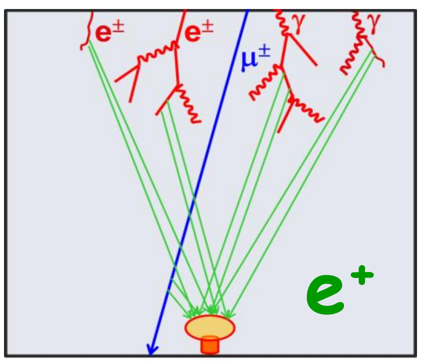
is very important

Mechanism & methods



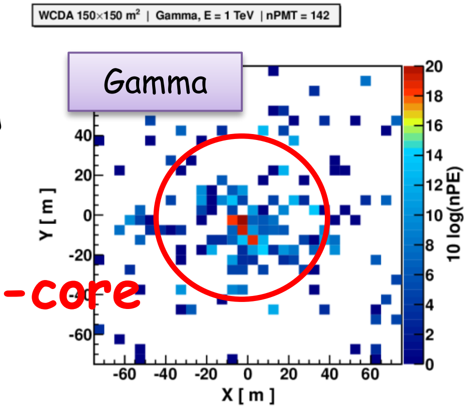


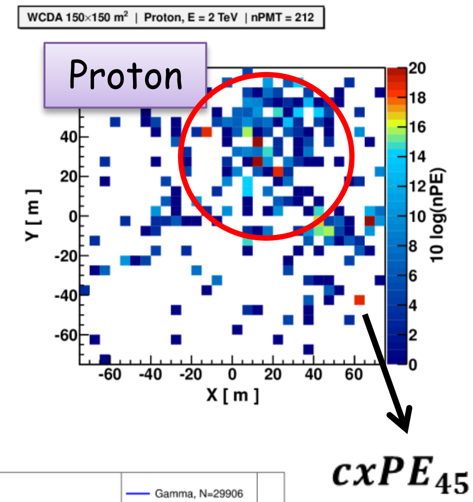




Secondary particles :	Single core, concentrated	Sub-core ;
Main compenante :	e± ; γ ;	e±; γ; hadron; μ ;
First altitude :	Low altitude ;	High altitude ;

Sensitive parameters



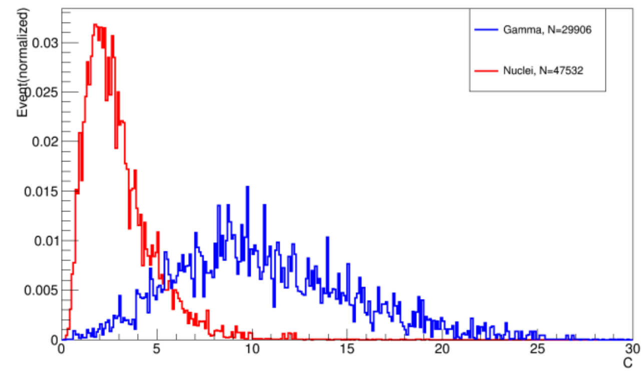


Take advantages of muons & sub-core

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

$nFit$: *number of PMT involved in reconstruction*

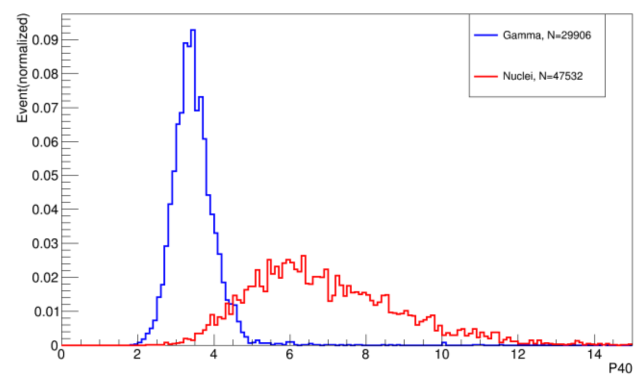
$cxPE_{45}$: maximum PE count outside the reconstruction core with a distance of 45m ;



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

$\sum PE_{40}$: sum number of nPE outside core within 40m

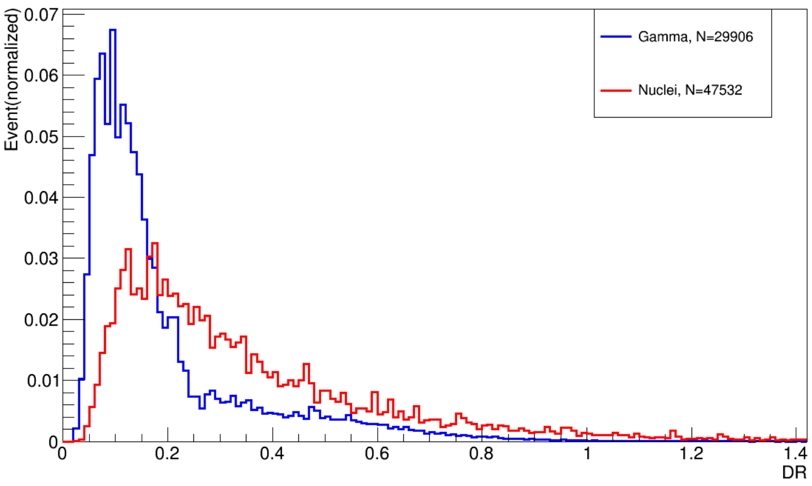
$\sum PMT_{40}$: sum number of fired PMTs outside shower core with distance of 40m



伽马/质子区分敏感参量

3、 Density raito(DR)

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



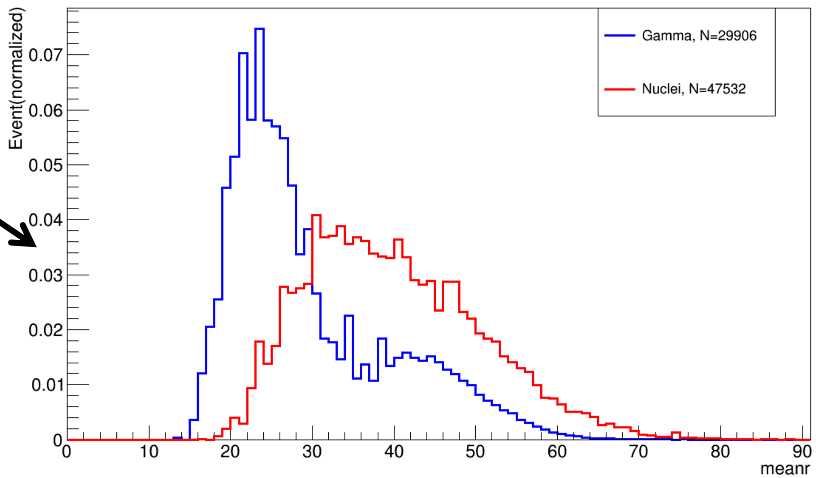
Use the lateral distribution of shower

4、 <R> = $\frac{\sum PE_i R_i}{\sum PE_i}$

PE_i: number of nPE of the ith fired PMT ;

R_i : distance between the ith fired PMT

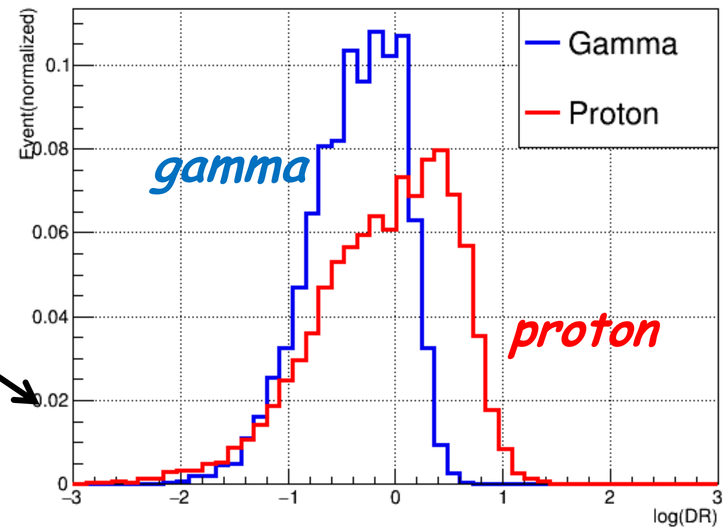
to shower core



伽马/质子区分敏感参量

5、 r70

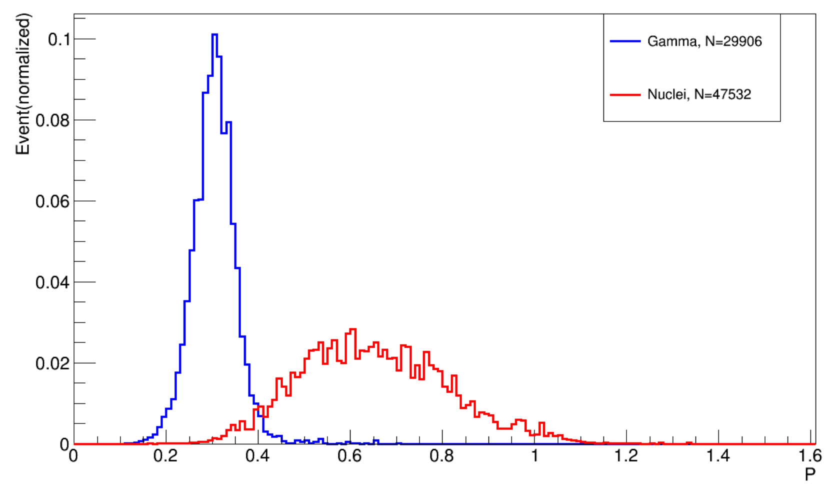
Minimum radio which contained 70% PE charge of the shower



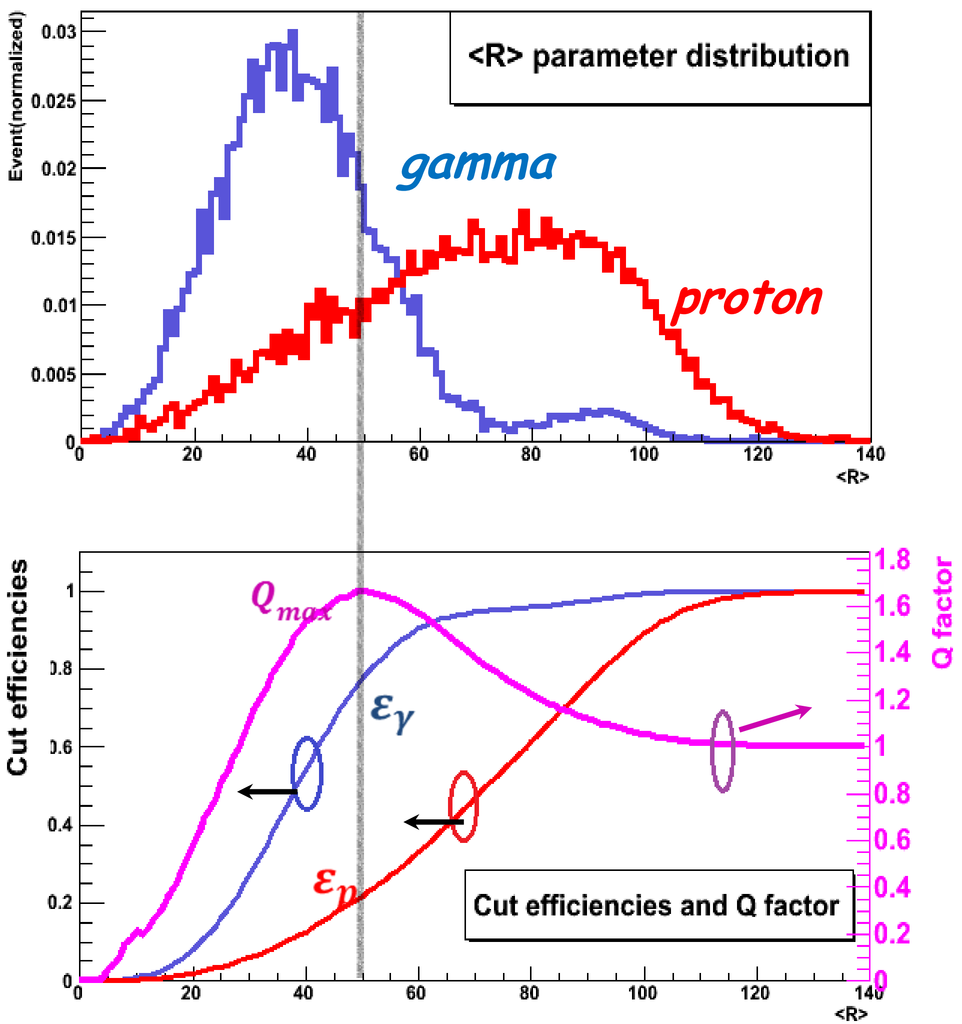
6、 P = $\frac{1}{N}\sum_{i=0}^{N}\frac{(\zeta_i-\langle\zeta_i\rangle)^2}{{\sigma_{\zeta_i}}^2}$

$\zeta_i = \log 10(Q_{\mathrm{eff},i})$

$\langle\zeta_i\rangle$: average ζ_i in an annulus



Judgment parameter——Q factor



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

ε_γ: keeping ratio of gamma

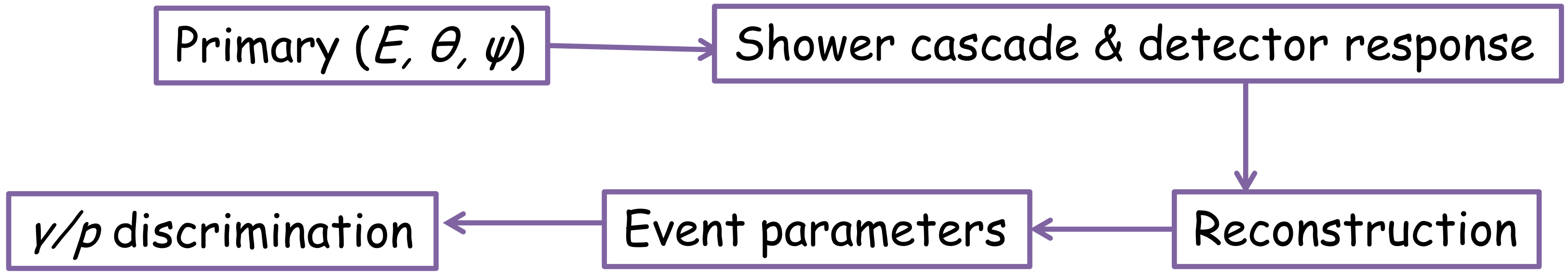
ε_p: 1-rejection ratio of proton

For example,

Q_{max}=1.68
ε_γ=78%
ε_p=20%

Background rejection ratio is 80% , keeping 78% of signal

G/P separation procedure



MC simulation

- 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

Event cuts condition

cuts :

- reconstruction core fall in 150x150 m^2
- *recθ* : 0~45 °
- *recψ* : 0~360 °
- Gamma keeping ratio >50%

Single parameter result

	P	*C*	<*R*>	DR
Q因子	2.91	3.25	1.22	1.04
ε_γ	67.65 %	66.02 %	59.33 %	91.42 %
ε_p	5.4%	4.1%	23.6%	76.0%

Energy = ~2 TeV

P &*C* : rejection ratio 95%

<*R*> & **DR** : rejection ratio 20%~70%

npmt	E_gam	<R>			DR			Compactness			Pincness		
10-47	189.017	1.00002	99.9679	99.9319	0.999931	99.9797	99.9732	1.04423	73.8223	49.9789	1.0217	97.6283	91.3077
47-71	543.876	1.04676	91.3689	76.191	1.00035	99.8994	99.7294	1.48189	76.9148	26.9395	1.21538	85.9487	50.0097
71-112	860.003	1.09559	82.3736	56.5296	1.00004	99.8885	99.7699	1.87352	74.8762	15.9724	1.4466	80.4579	30.9343
112-172	1341.22	1.1621	66.3033	32.5524	1.02324	93.1197	82.8194	2.59154	63.0289	5.91511	1.95906	75.2315	14.747
172-254	2165.21	1.22231	59.3376	23.5665	1.04885	91.4216	75.9745	3.24649	66.0249	4.13606	2.90978	67.6586	5.40663
254-354	3609.94	1.35336	50.7232	14.047	1.14121	50.5401	19.6128	4.46821	51.0002	1.30279	4.16882	77.2624	3.43487
354-467	5591.14	1.74479	50.2809	8.30457	1.48468	52.3975	12.4554	7.71384	50.0282	0.420619	8.76408	65.6019	0.560299
467-581	8881.78	1.84221	52.7288	8.19248	1.59034	52.2343	10.7878	5.71224	51.2606	0.805292	18.2973	50.4109	0.0759058
581-680	15867.2	1.94035	51.0675	6.92676	1.39635	60.211	18.5936	5.44534	56.1482	1.06322	9.87803	76.8524	0.605304
680-900	27195.7	1.73014	51.9086	9.00148	1.34749	58.9049	19.1095	5.58073	51.7693	0.860523	4.40111	69.5774	2.49926

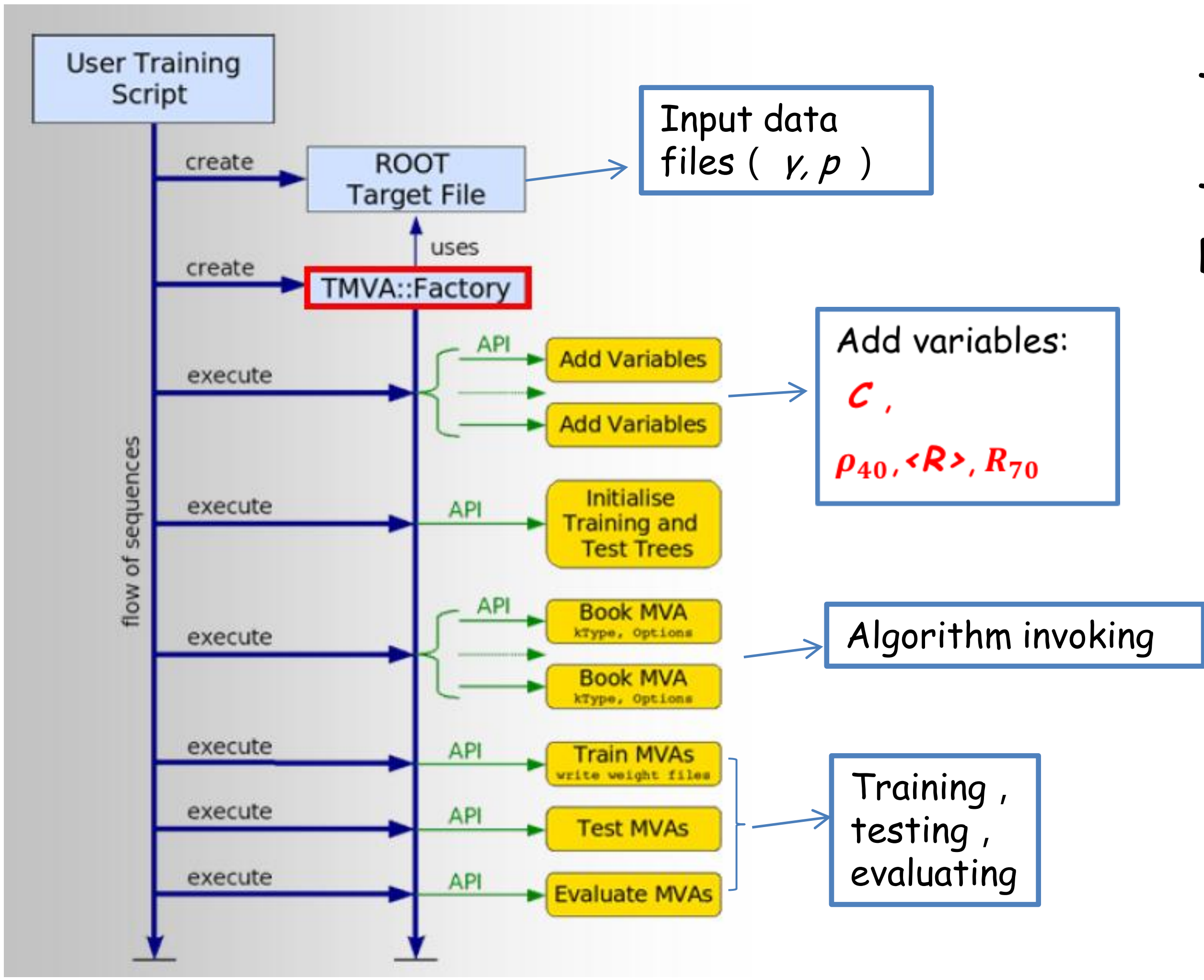
weakness : poor performance at low energy range (<2TeV)

Multivariate analysis

ROOT toolkit (v5.14)——**TMVA**

TMVA

分析步骤：



TMVA:

Toolkit for Multivariate Data Analysis with ROOT

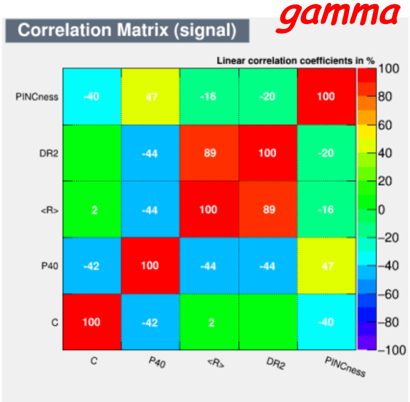
Algorithms :

- **Artificial Neural Network (ANN)**
- **Boosting decision trees (BDTG)**

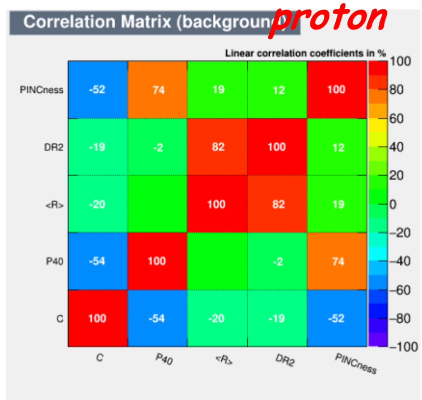
Artificial Neural Network (ANN)

1、 input variable correlations

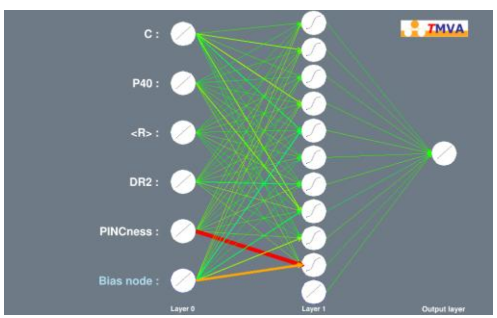
gamma

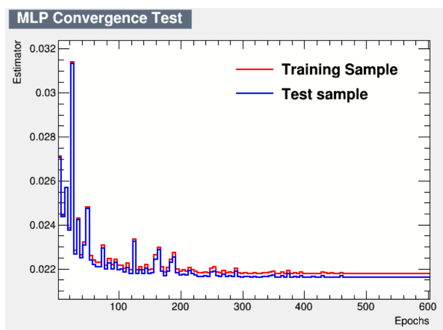


proton



2、 neural network layout & convergence test





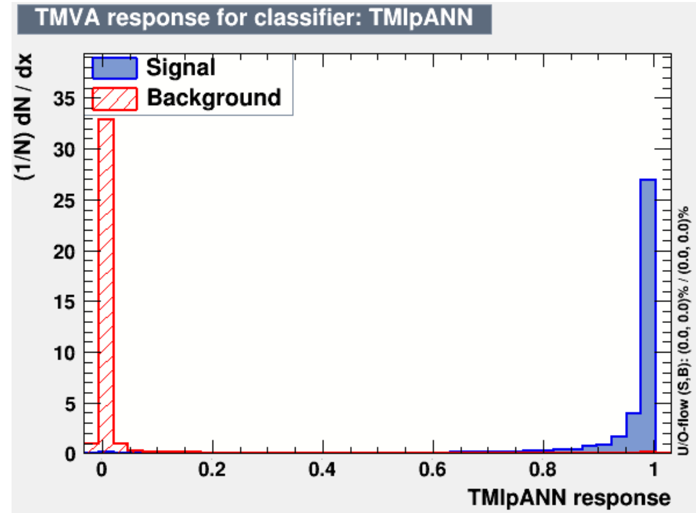
C parameter

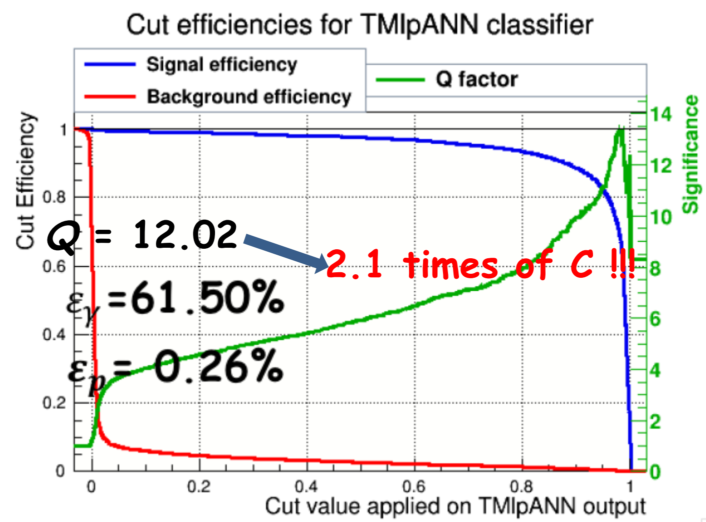
Q = 5.71

ε_γ = 51.26%

ε_p = 0.81%

3、 ANN result (467<*npmt*<581)



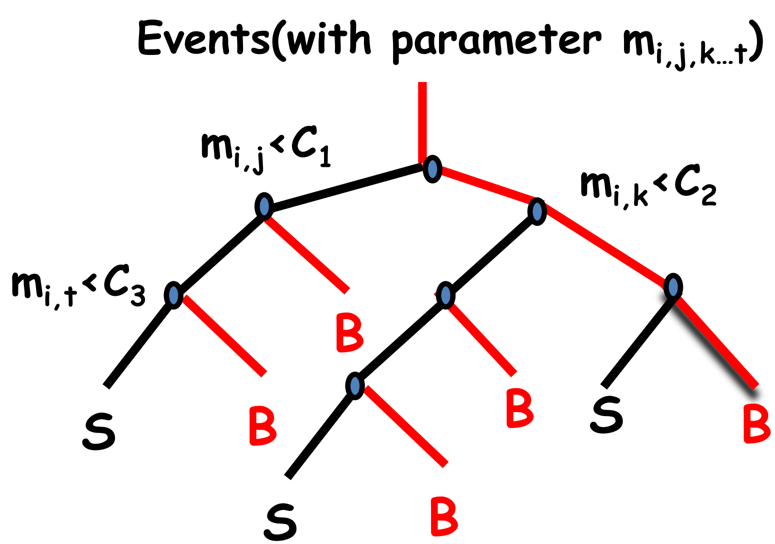


Q = 12.02 → 2.1 times of C !!!

ε_γ =61.50%

ε_p = 0.26%

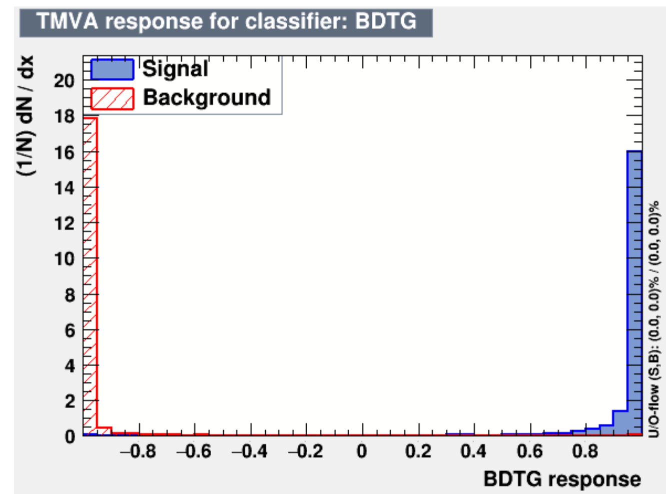
决策树多参数分析结果

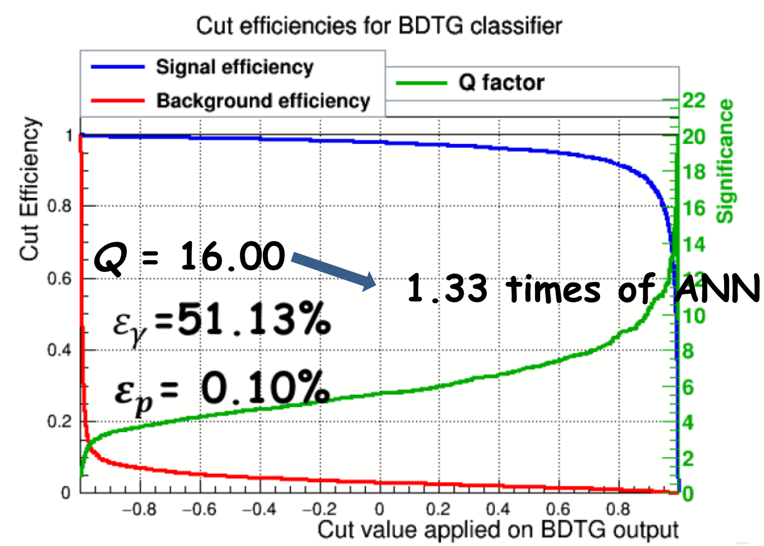


决策树示意图

ANN
- Q = 12.02
- ε_γ = 61.50%
- ε_p = 0.26%

BDTG result (467<*npmt*<581)





Single parameter v.s. multivariate analysis

Single parameter:

1.04423	73.8223	49.9789	1.0217	97.6283	91.3077
1.48189	76.9148	26.9395	1.21538	85.9487	50.0097
1.87352	74.8762	15.9724	1.4466	80.4579	30.9343
2.59154	63.0289	5.91511	1.95906	75.2315	14.747
3.24649	66.0249	4.13606	2.90978	67.6586	5.40663
4.46821	51.0002	1.30279	4.16882	77.2624	3.43487
7.71384	50.0282	0.420619	8.76408	65.6019	0.560299
5.71224	51.2606	0.805292	18.2973	50.4109	0.0759058
5.44534	56.1482	1.06322	9.87803	76.8524	0.605304
5.58073	51.7693	0.860523	4.40111	69.5774	2.49926

Multivariate:

1.45301	0.579734	0.159191	1.53959	0.551961	0.12853
1.80456	0.626471	0.12052	1.98333	0.654642	0.108948
2.32802	0.605486	0.0676449	2.53814	0.625198	0.060674
3.29678	0.560947	0.028951	3.59167	0.559532	0.0242693
4.93119	0.521227	0.0111725	5.26314	0.595072	0.0127835
7.24098	0.501201	0.00479103	8.09207	0.502552	0.00385693
9.61128	0.554136	0.00332408	11.1532	0.501237	0.00201969
12.0243	0.61504	0.0026163	16.0013	0.511289	0.00102099
11.8993	0.763823	0.00412041	14.5283	0.515502	0.00125901
11.8684	0.691859	0.00339821	16.7056	0.587248	0.00123571

Multivariate analysis has higher background rejection ratio---higher Q value

ANN v.s. BDTG method

npmt		MLPANN			BDTG		
10-47	189.017	1.45301	0.579734	0.159191	1.53959	0.551961	0.12853
47-71	543.876	1.80456	0.626471	0.12052	1.98333	0.654642	0.108948
71-112	860.003	2.32802	0.605486	0.0676449	2.53814	0.625198	0.060674
112-172	1341.22	3.29678	0.560947	0.028951	3.59167	0.559532	0.0242693
172-254	2165.21	4.93119	0.521227	0.0111725	5.26314	0.595072	0.0127835
254-354	3609.94	7.24098	0.501201	0.00479103	8.09207	0.502552	0.00385693
354-467	5591.14	9.61128	0.554136	0.00332408	11.1532	0.501237	0.00201969
467-581	8881.78	12.0243	0.61504	0.0026163	16.0013	0.511289	0.00102099
581-680	15867.2	11.8993	0.763823	0.00412041	14.5283	0.515502	0.00125901
680-900	27195.7	11.8684	0.691859	0.00339821	16.7056	0.587248	0.00123571

- BDTG is better than ANN

Running time

ANN

Elapsed time for training with 185993 events: 767 sec

Elapsed time for evaluation of 185993 events: 0.515 sec

BDTG

Elapsed time for training with 185993 events: 222 sec

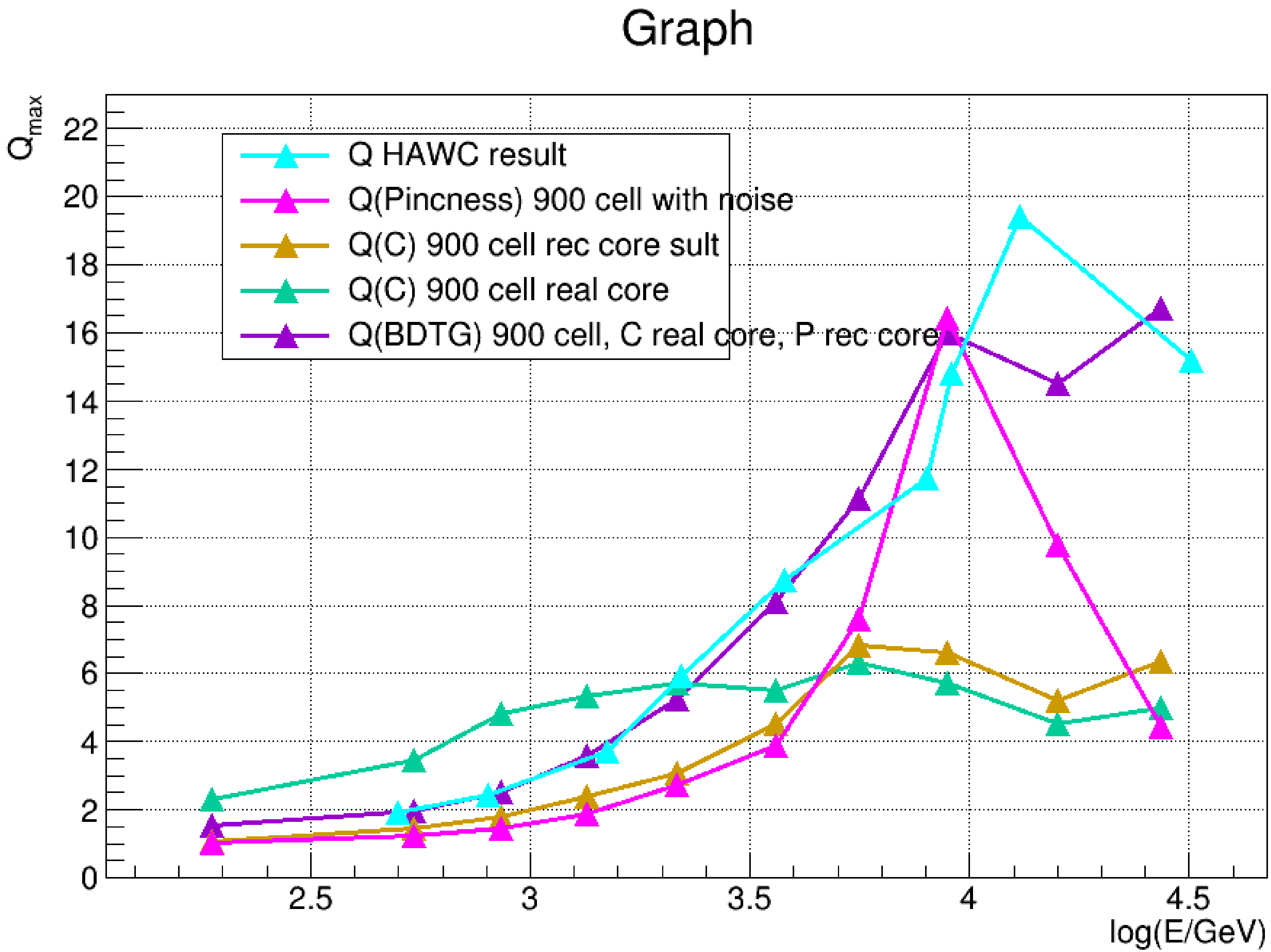
Elapsed time for evaluation of 185993 events: 0.476 sec

More events ,
Longer time。

Comparison ：

Training time >>test and evaluate time

ANN > BDTG



Graph
Q HAWC result
Q(Pincness) 900 cell with noise
Q(C) 900 cell rec core sult
Q(C) 900 cell real core
Q(BDTG) 900 cell, C real core, P rec core
Q_{max}
0
2
4
6
8
10
12
14
16
18
20
22
2.5
3
3.5
4
4.5
log(E/GeV)

Summary

- LHAASO-WCDA can separate gamma/proton well。
- Energy >2TeV: $\boldsymbol{Q\ factor}$ of C & P $> \mathbf{3}$ ，排除85%背景事例。
- Multivariate analysis can improve Q factor value obviously.
- There are still some problems need to be resolved
- BDTG is better than ANN both in performance and running velocity.

THANK YOU !

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