A machine learning approach to identify the air shower cores for the GRAPES-3 experiment

Medha Chakraborty

DLCP-2022
On behalf of GRAPES-3 collaboration

July 6, 2022

Outline

- Motivation
- GRAPES-3 experiment
- Shower reconstruction
- Manual cuts
- ML in analysis
- Results

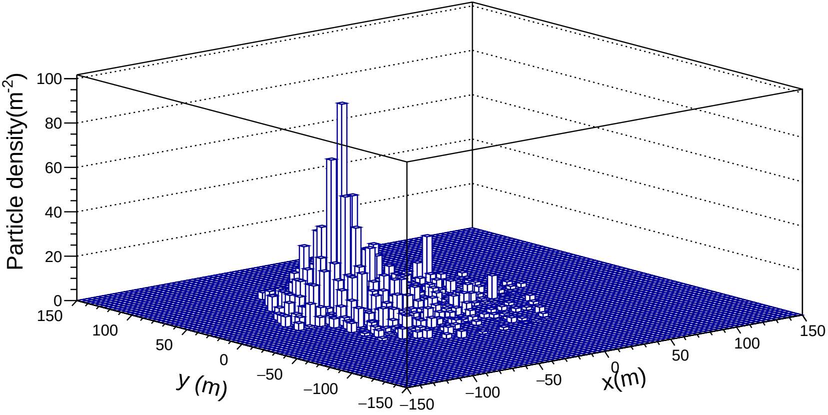
Motivation

- One of the primary objectives of GRAPES-3 is to measure cosmic ray energy spectrum, composition and sources to probe century old mysteries of CR acceleration and propagation.
- GRAPES-3 also performs several energy dependent analysis, like angular resolution, anisotropy.
- Such analysis get affected by mis-reconstructed air showers.
- This work describes the identification and removal of such showers.

The GRAPES-3 experiment

- Location: Ooty, India (11.4°N, 76.7°E, 2200 m asl)
- ~400 plastic scintillators spread over 25000 m^2 with 8 m inter detector separation
- Trigger: L0: 3 line coincidence, L1: at least 10 detectors hit.
- Observables: particle densities and relative arrival times
- Statistics: ~3 million showers per day
- Muon telescope covering 560 m^2
- Energy range: 1 TeV - 10 PeV

Shower profile



Shower reconstruction using NKG function

$$\rho_i = \frac{N_e}{2\pi r_m^2}\frac{\Gamma(4.5-s)}{\Gamma(s)\Gamma(4.5-2s)}(\frac{r_i}{r_m})(1+\frac{r_i}{r_m})^{s-4.5}$$

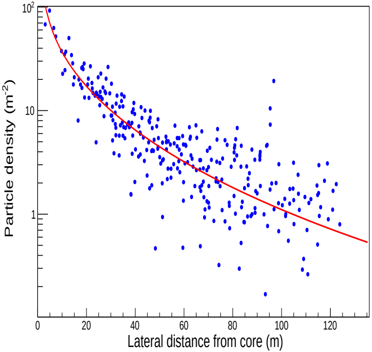
ρ_i : expected density at i-th detector

r_i : distance of i-th detector from shower core (X_c, Y_c)

N_e : Shower size

s : Shower age

r_m: Moliere radius 103 m at Ooty



Mis-reconstructed cores

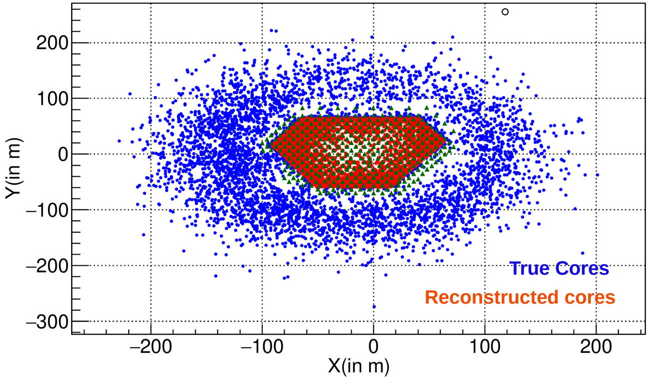


Figure: Mis-reconstructed shower cores for 100-158 TeV showers

Shower simulation using CORSIKA

- Hadronic interaction generator FLUKA below 80 GeV and SIBYLL above this
- Proton : 1 TeV - 10 PeV with spectral index -2.5
- Detector response is calculated and reconstructed
- Total number of showers $\sim 5 \times 10^8$

Presence of contamination

Thrown upto a distance beyond which L1-trigger fraction is less than 1%.

S: Both true and reconstructed cores inside.

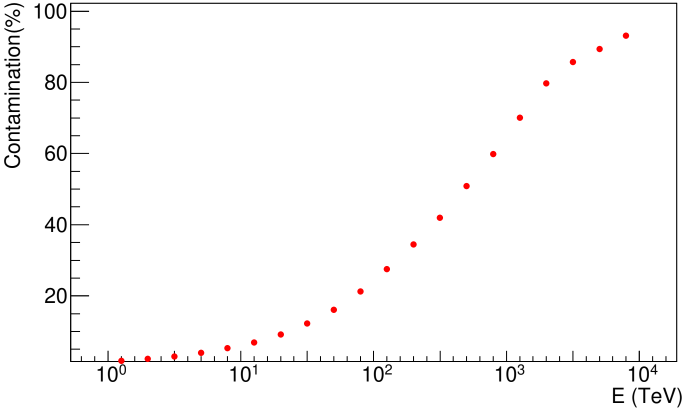
B: True cores outside but reconstructed cores inside

Energy(TeV)	Distance(in m)
1-10	100
10-15.8	110
15.8-25.1	120
25.1-39.8	130
39.8-63.1	140
63.1-251.2	300
251.2-398.1	450
398.1-1584.8	500
1584.8-2511.9	650
2511.8-3981.1	700
3981.1-6309.6	750
6309.6-10000	800

Contamination, $C = \frac{B}{S+B} \times 100\%$

Initial contamination

- $\theta \leq 25^\circ$, reconstructed cores within the array
- Successful NKG fit



Contaminated showers

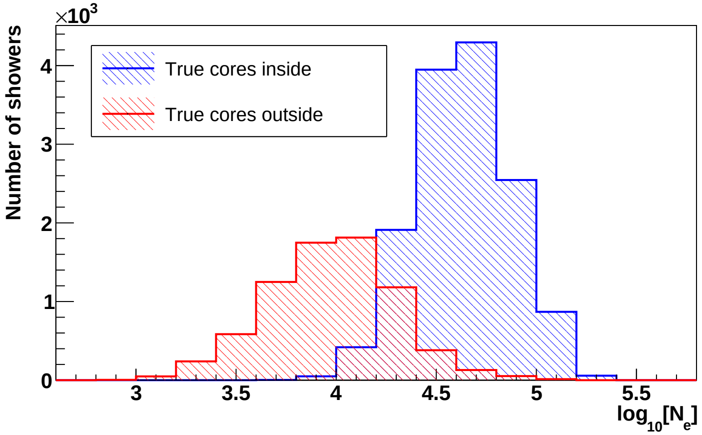
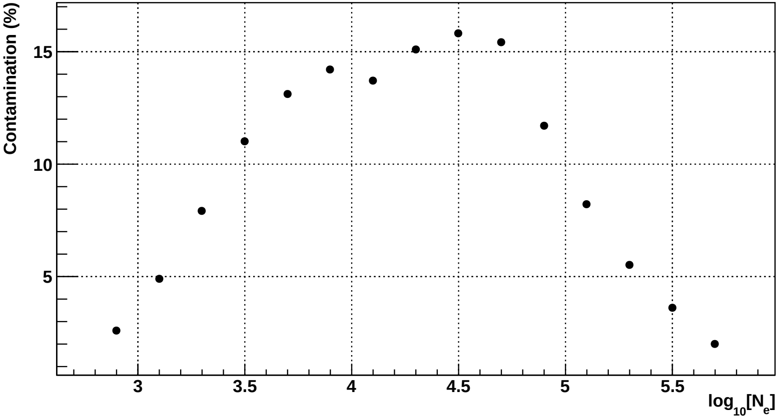


Figure: Shower size distributions for $158\,TeV \leq E \leq 251\,TeV$

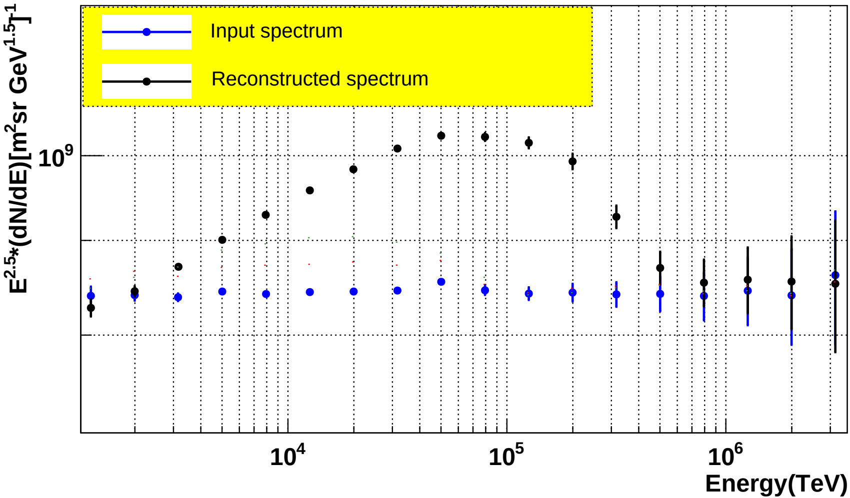
Energy reconstruction is performed using shower size. These showers are interpreted as low energy showers.

Contamination with shower size



Maximum contamination shifts to intermediate shower size. Reconstructed energy is a function N_e, so this affects any energy dependent analysis.

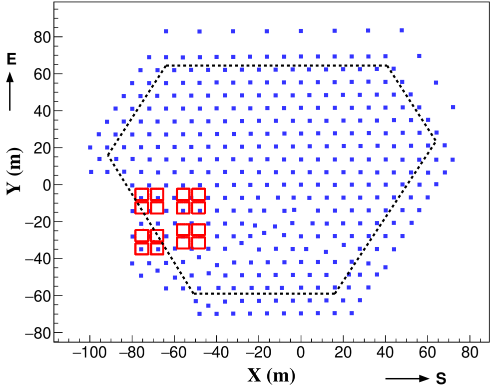
Effects on energy spectrum



Study performed on simulated dataset shows that unfolded energy spectrum does not match with expected spectrum.

Variables

1. PSumRatio: $PSumRatio = PSumOut/PSumIn$
   PSumIn: Sum of particle densities inside
   PSumOut: Sum of particle densities outside



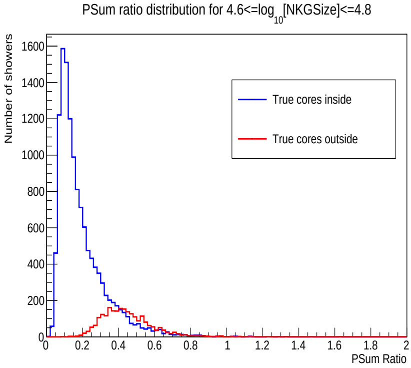
Variables

- LnNKGP : best functional value obtained for negative log likelihood function used for NKG fit.
- Age : Developmental stage of shower, obtained from NKG fit
- Age err : Error on Age parameter
- ChiSq1 : ChiSq1 of the planar fit for direction reconstruction
- LnCErr : Error in constant term of NKG function

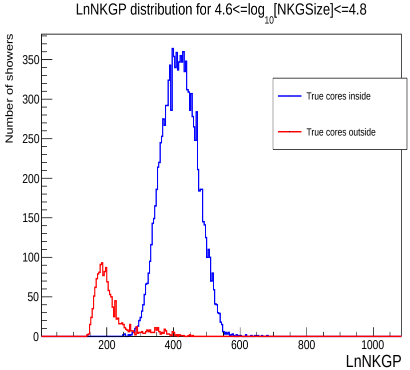
Variables are divided into logarithic N_e bins of width $10^{0.2}$.
Cuts are devised on the above variables manually and using machine learning.

Variables

$4.6 \leq log_{10}[N_e] \leq 4.8$



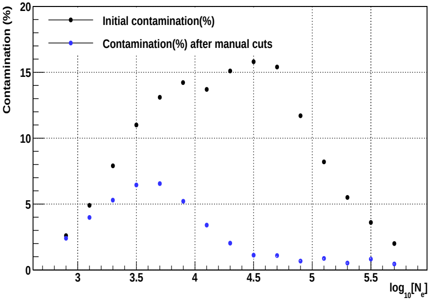
(a) PSumRatio

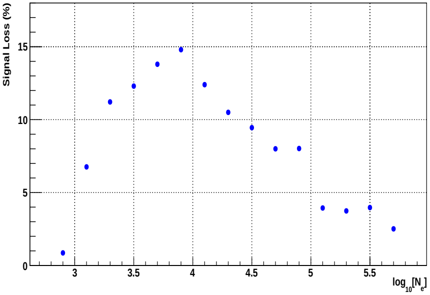


(b) LnNKGP

Method 1: Applying cuts manually

The cuts are applied chronologically on the variables by calculating signal loss, contamination and signal significance ($S/\sqrt{S+B}$) at each step

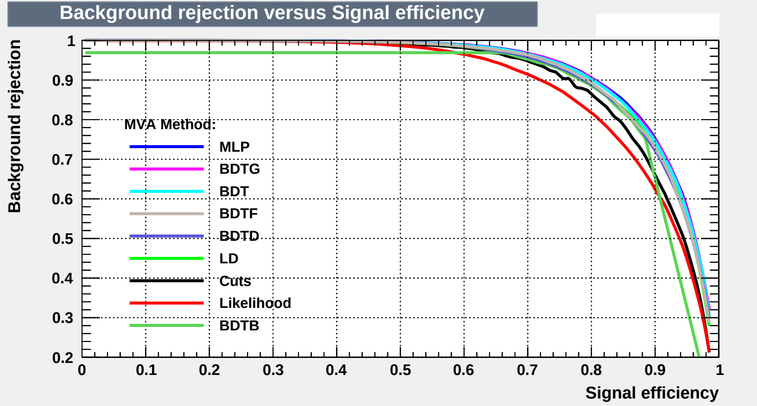




This was repeated for all other size bins and all other variables.

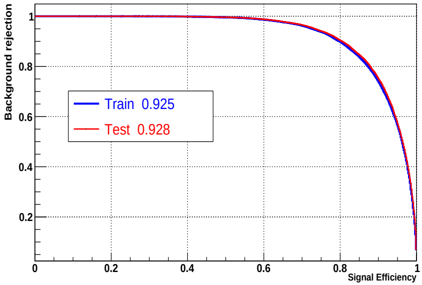
Method 2: Analysis with machine learning

- Tedious to deal with several variables in manual cuts.
- Machine learning was used for this purpose.
- Method: BDT-G, maximum area of ROC curve
- Simulated data divided into two equal halves for training and testing

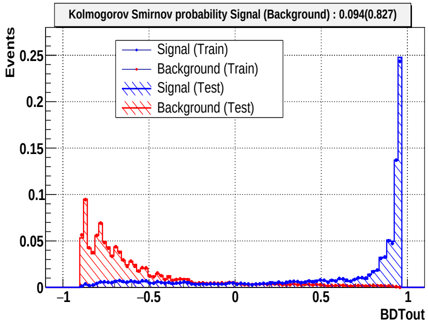


TMVA package of ROOT was used.

$4.0 \leq log_{10}[N_e] < 4.2$



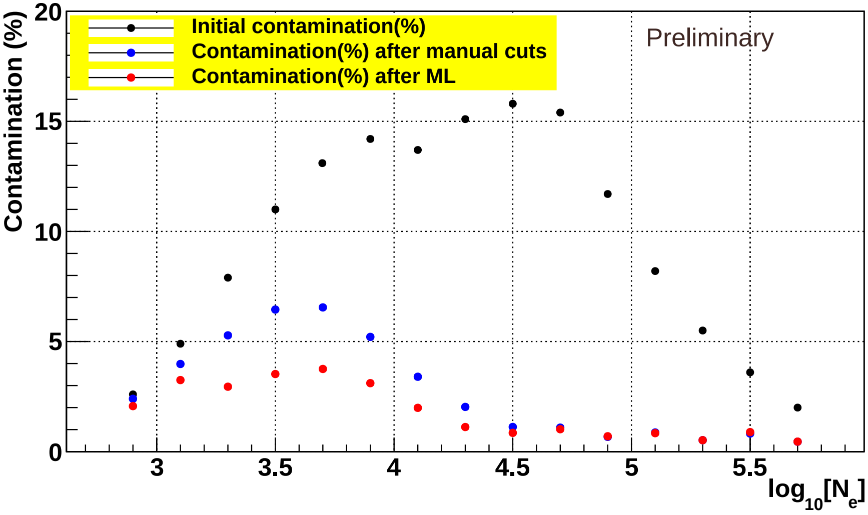
(a) The ROC curve matches well for train and test



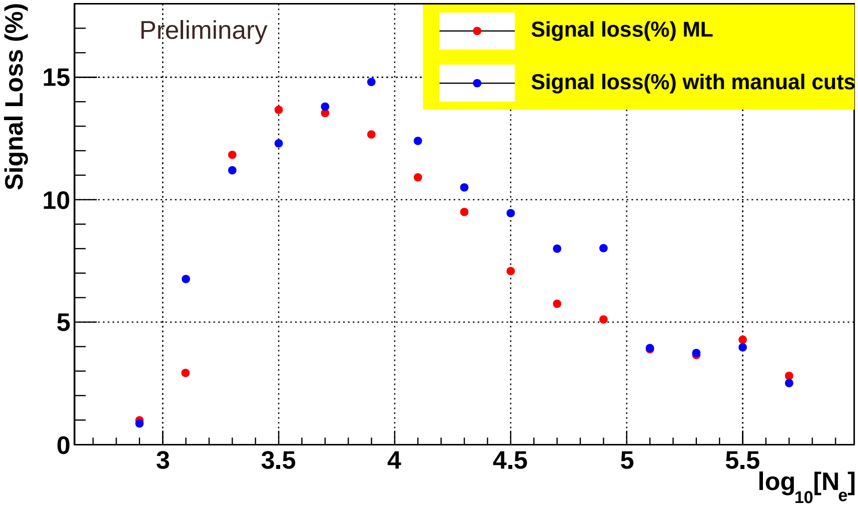
(b) KS test

BDT-G parameters adjusted to obtain the maximum area within ROC. BDT output variable shows clear separation

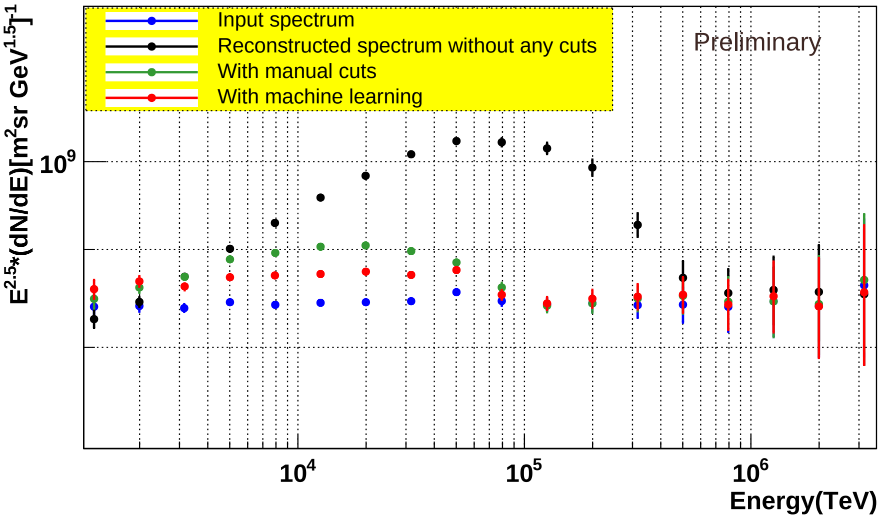
Results: Contamination



Results: Signal loss



Improvements in energy spectrum



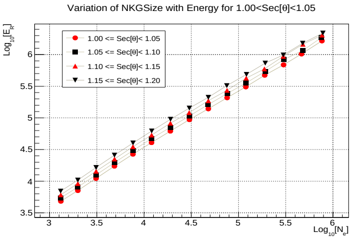
Machine learning reduces the deviation in energy spectrum measurements.

Summary

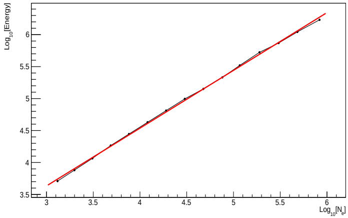
- Mis-reconstructed showers were identified using manual cuts and BDT-G.
- Better measurement of energy spectrum can be achieved using machine learning
- This approach will improve energy estimation of any energy dependent analysis

Backup

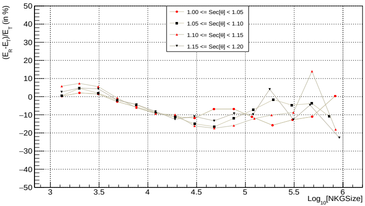
Energy reconstruction



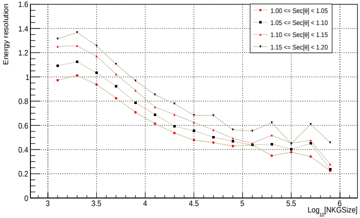
(a) Median energy plotted with median N_e for different zenith bins



(b) Fitted linearly to find $E_R(N_e, \theta)$ for $1.0 \leq sec(\theta) < 1.05$

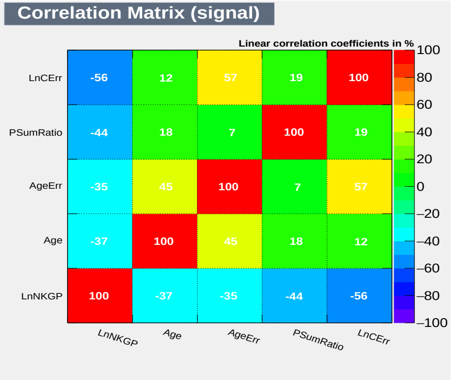


(c) Accuracy of energy callibration

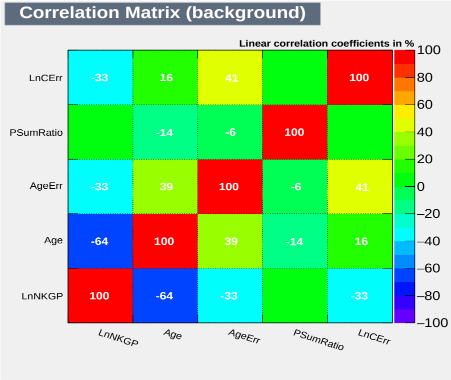


(d) Precision of callibration

Correlation



(a) Signal

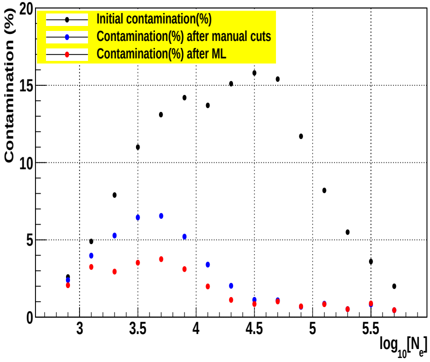


(b) LnNKGP

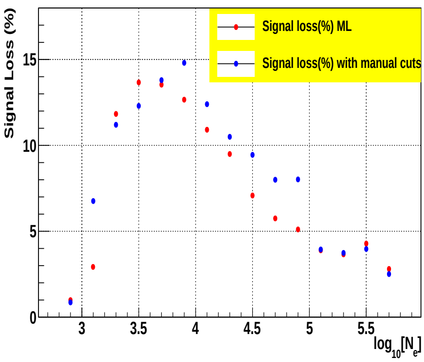
Variable Importance

Variables	Importance
PSumRatio	0.48
LnNKGP	0.32
Age	0.14
LnCErr	0.04
AgeErr	0.02

Results: Contamination



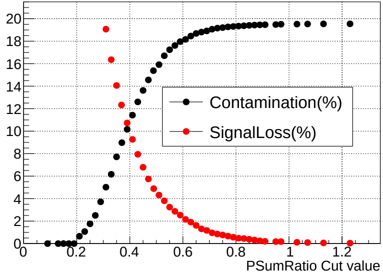
(a) Contamination (%)

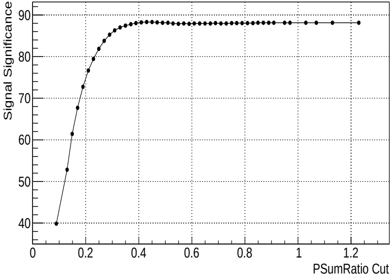


(b) Signal loss (in %)

Method 1: Applying cuts manually

The cuts are applied chronologically on the variables. Contamination, signal loss and signal signifcance are studied at every step. For $4.6 \leq log_{10}[N_e] \leq 4.8$. Eg: The cut deviced for PSumRatio is shown.

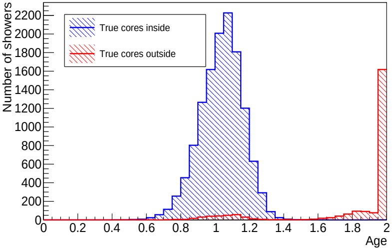




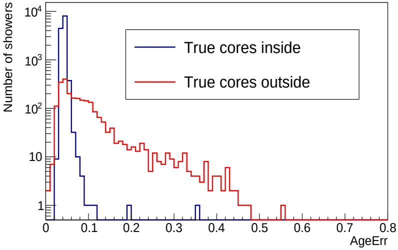
This was repeated for all other size bins and all other variables.

Variables

$4.6 \leq log_{10}[NKGSize] \leq 4.8$



(a) Age



(b) AgeErr