Machine learning approach to estimating the mass composition of cosmic rays

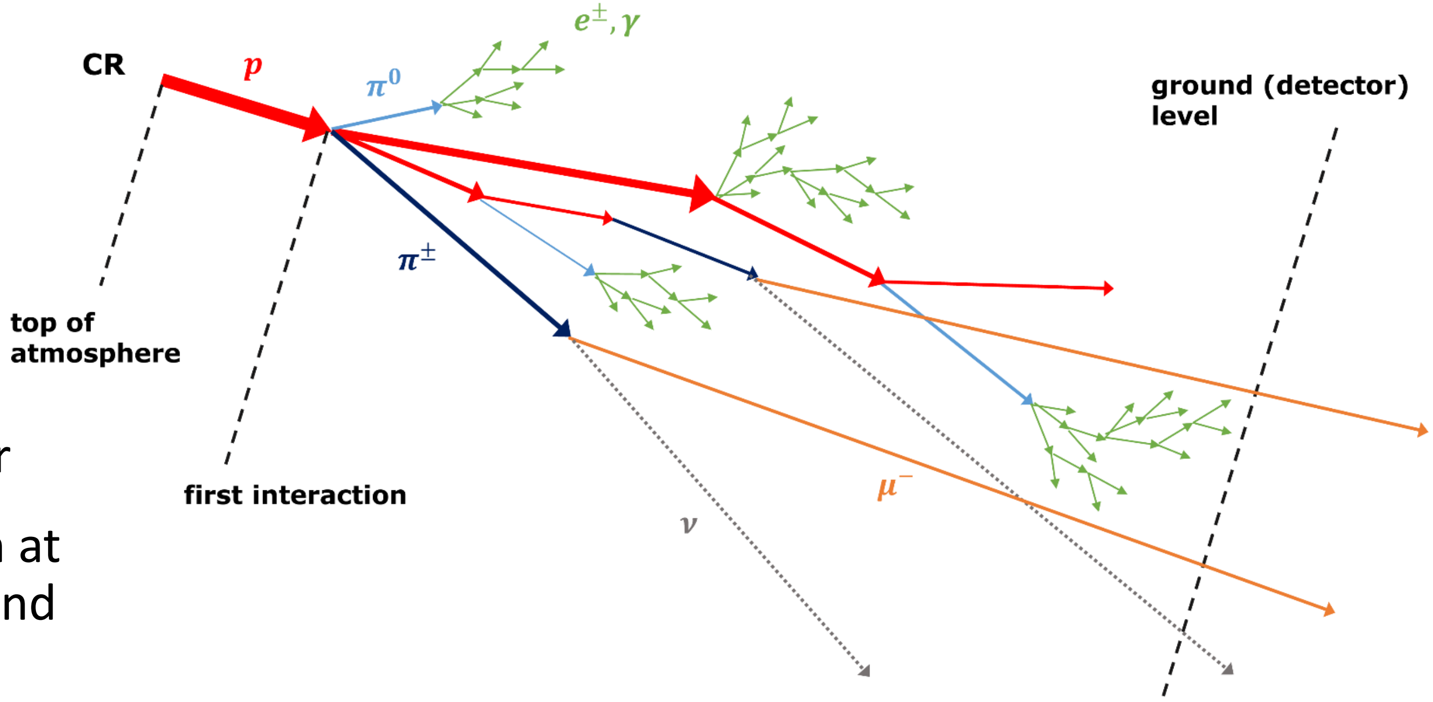
dr. Gašper Kukec Mezek

Mentor: prof. dr. Andrej Filipčič

ASTRO@TS 2019 (June 24th 2019)

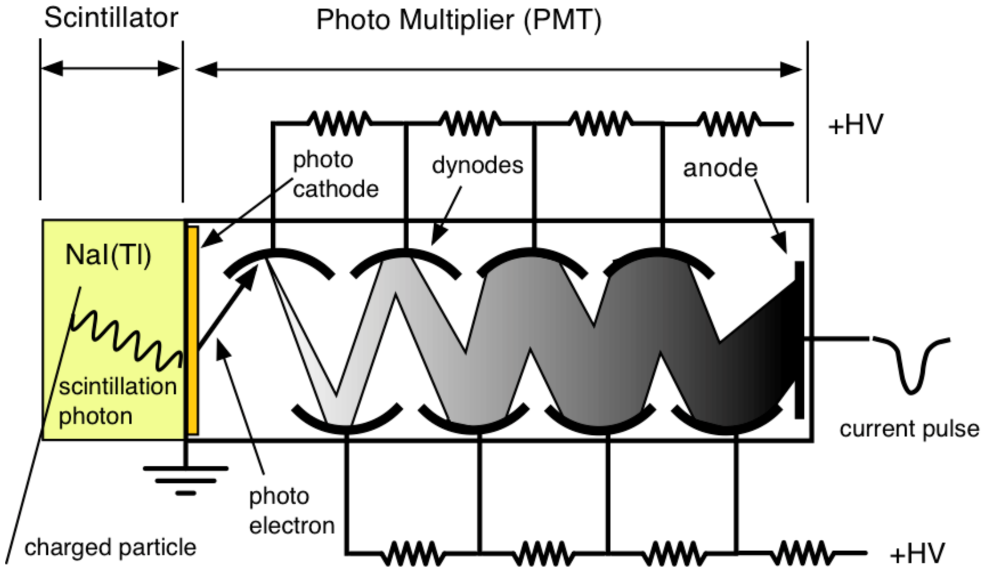
Cosmic rays and mass composition

- Cosmic rays (CR): Charged particles arriving to Earth from extraterrestrial sources
- Ultra-high energy cosmic rays (UHECR): CR with energies above $\sim 10^{18}\ eV$
- Extensive air shower (EAS): Cascade of secondary particles after interaction of UHECR with atmospheric nuclei
- Mass composition studies:
  - Type of initial particle dictates the evolution of the extensive air shower
  - Motivation: Interaction cross-section at extreme energies, uncover sources and acceleration processes of UHECR
  - Main drawback: Cross-section at the highest energies extrapolated from LHC measurements (max $E_{LHC} \sim 10^{17}\ eV$ in the laboratory reference frame)



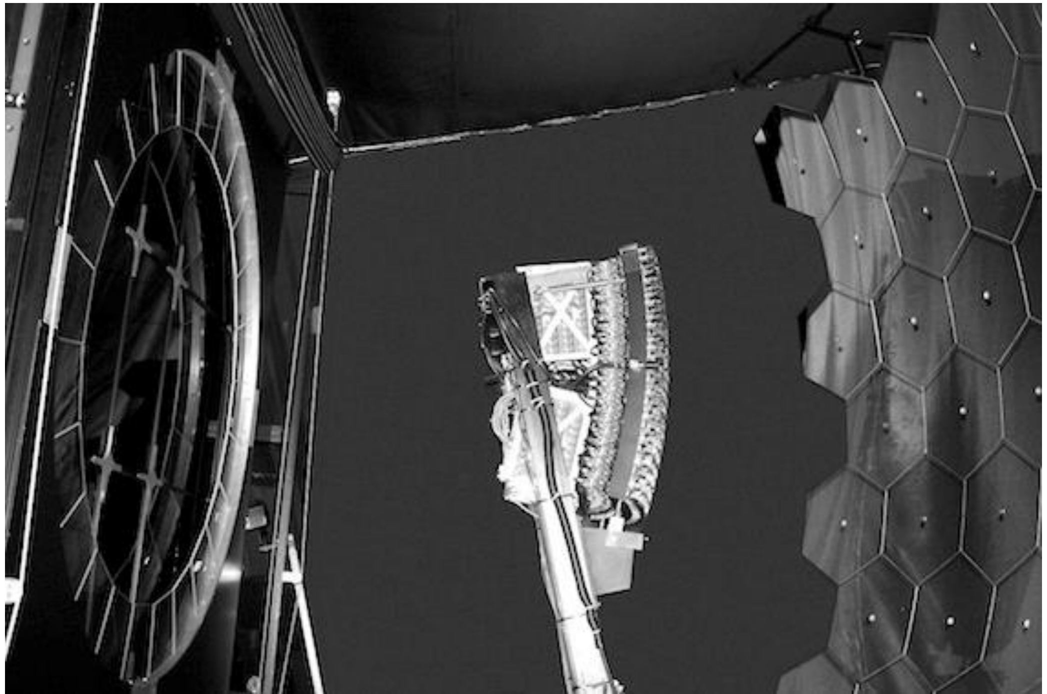
Detection of UHECR

- Various detection systems:
  - Water Cherenkov stations: filled with water, photomultipliers detect produced Cherenkov light
  - Scintillation detectors: production of luminescence in a material excited by ionizing radiation
  - Fluorescence telescopes: observing deexcitation from nitrogen molecules in the UV wavelength range



wanda.fiu.edu/teaching/courses/Modern_lab_manual/scintillator.html

www.flickr.com/photos/134252569@N07



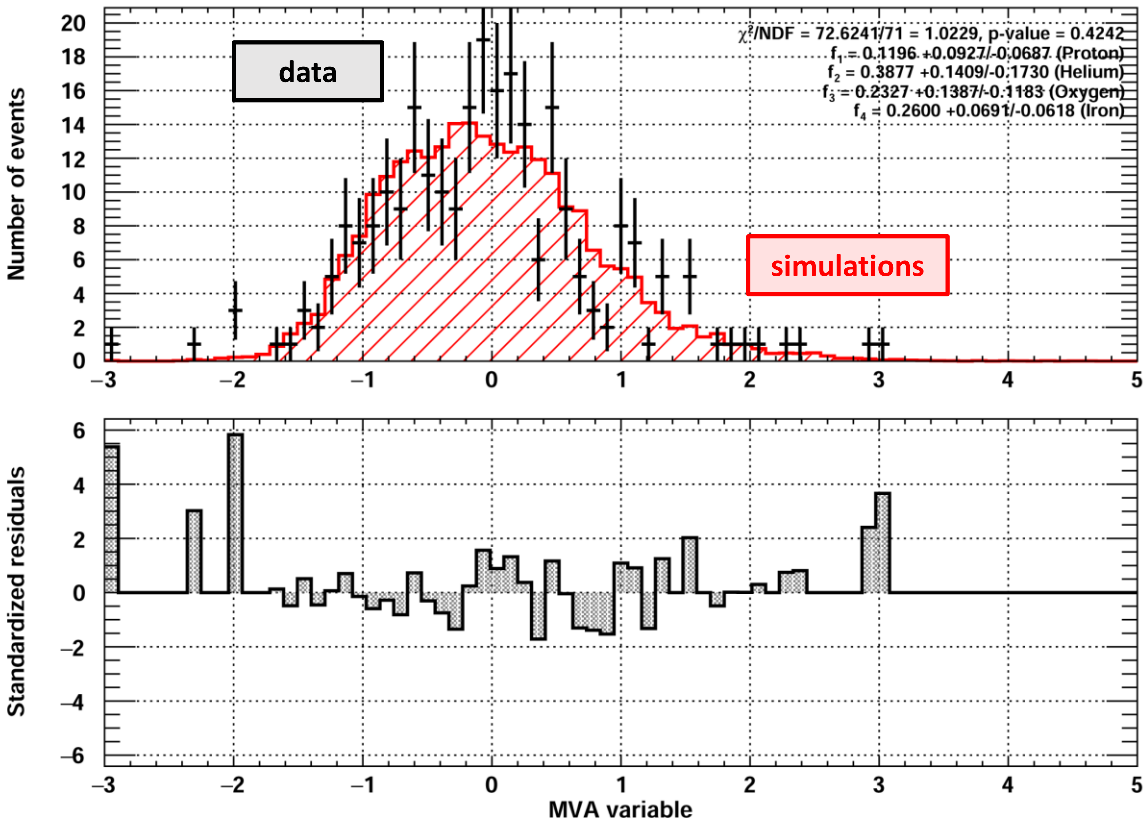
Machine learning and multivariate analysis

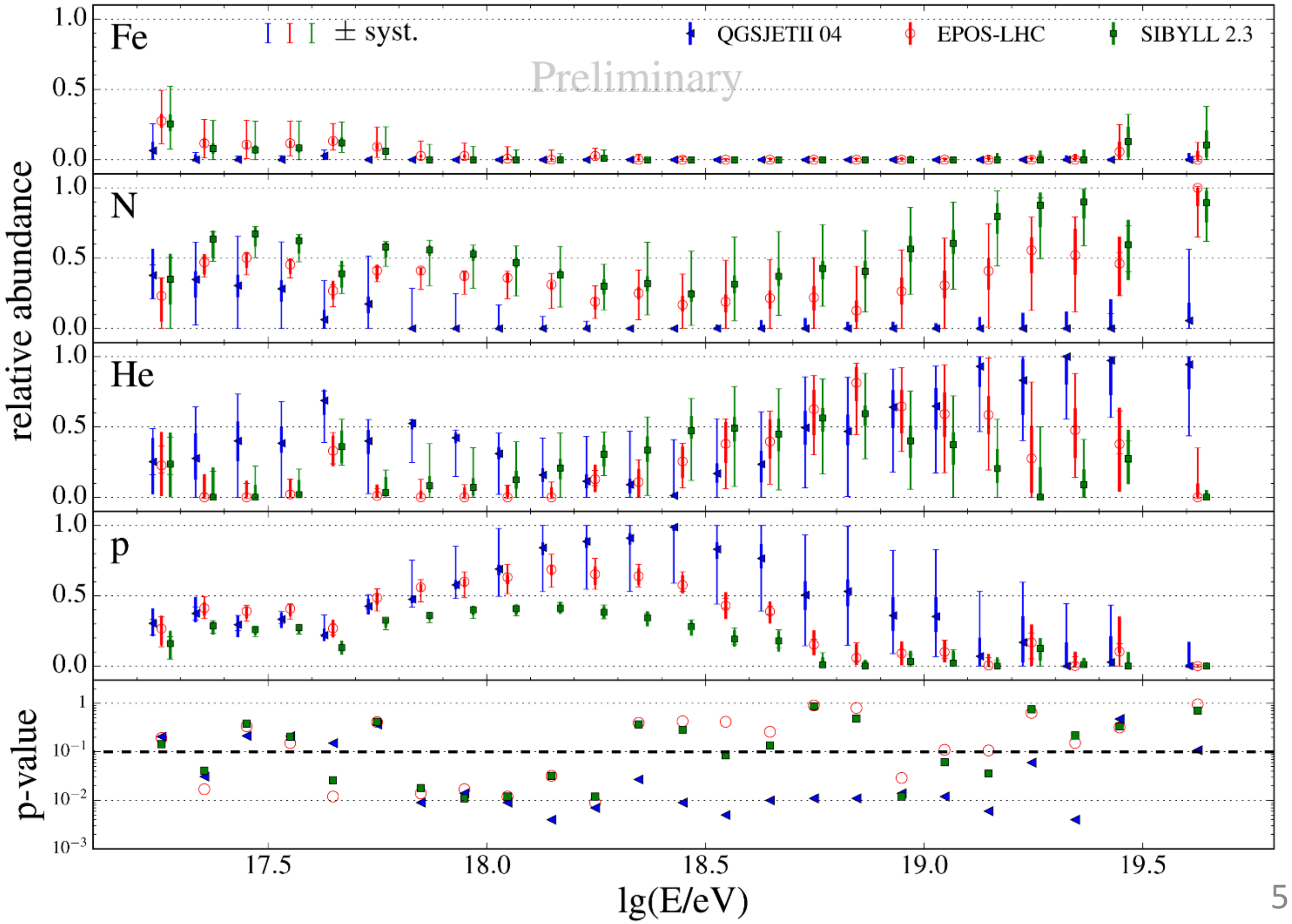
- Machine learning: Using computer algorithms, which learn from data without being explicitly programmed
- Multivariate analysis (MVA): Combining multiple input features (variables) in order to improve separation strength between different classes
- Complementary detection system for UHECR → mass composition can be estimated by combining EAS properties

	Statistical approach	Event-by-event approach
Description	• Split the data set into subsets using the same constraints as for simulations • Perform distribution fitting or parameterization to extract mass composition information	• Use simulations in an MVA analysis to classify between different particle types • Apply classification cuts to individual events for an event-by-event classification
Strengths	• Simple to implement • Only one step in the MVA analysis • Works even when separation strength is weaker	• True determination of particle type for each event separately
Weaknesses	• Only gives elemental fraction values for included particle types (generalization)	• Difficult to implement • Many steps in the MVA analysis (for multiple classes) • Requires good separation strength for all classes

Why statistical approach?

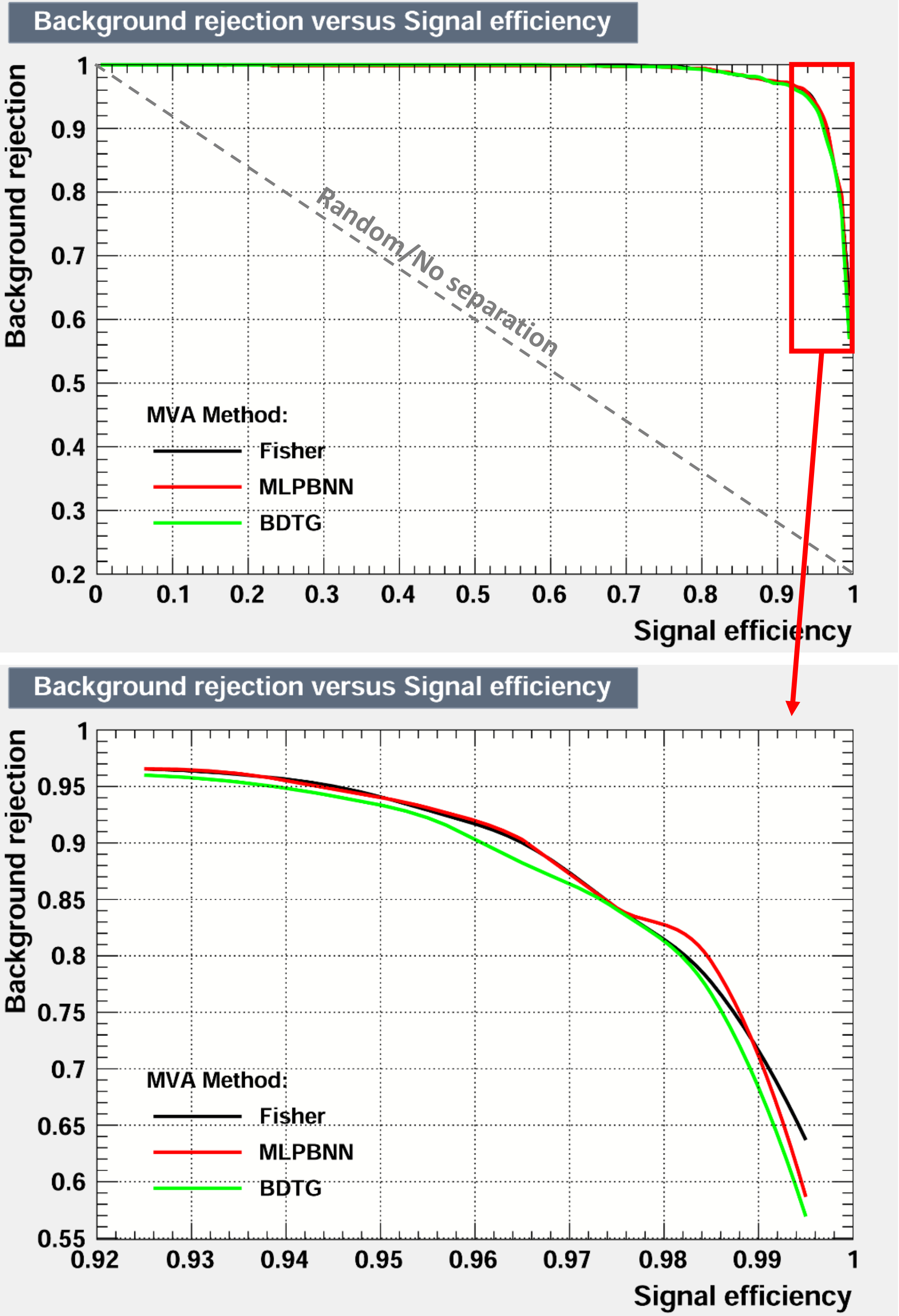
- Much easier to implement than event-by-event identification
- Can extract elemental fractions through distribution fitting (maximum likelihood)
- Direct comparison to results (ex. Pierre Auger Observatory [PoS(ICRC2017), PRD 90 (2014) 122006])





Multivariate analysis steps

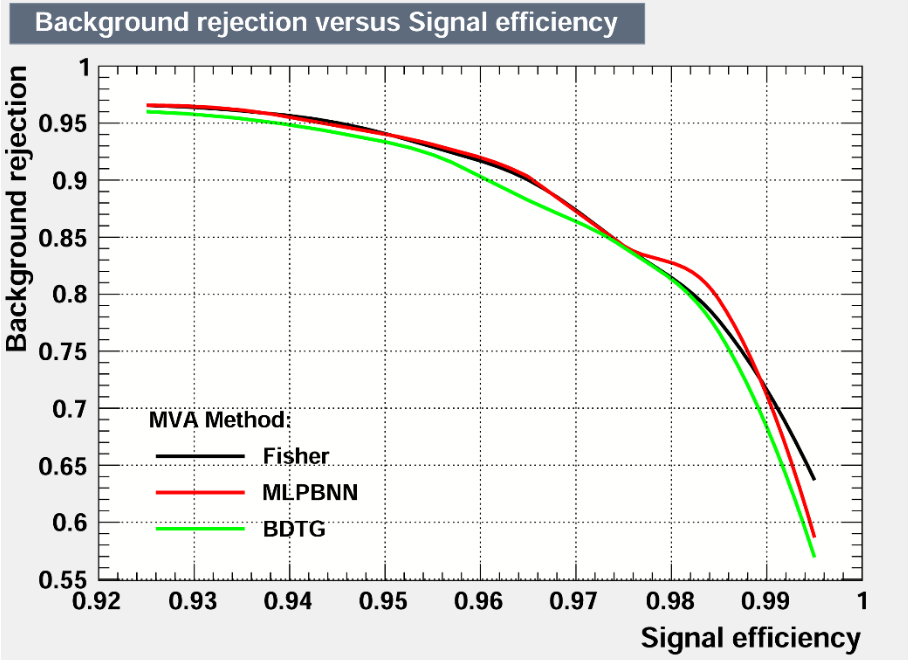
- The simulations need to be split into:
  - MVA training set: Training the MVA method, performing distribution fitting
  - Cross-validation set: Estimate stability of method on events not used during training
- Analysis follows these steps:
  1. Perform treatment of simulations and data
  2. Select input features (variables) and the MVA method
  3. Train and test the MVA method (determines separation strength)
  4. Apply MVA method on all data sets to get the output MVA variable distribution
  5. Perform MVA variable distribution fitting

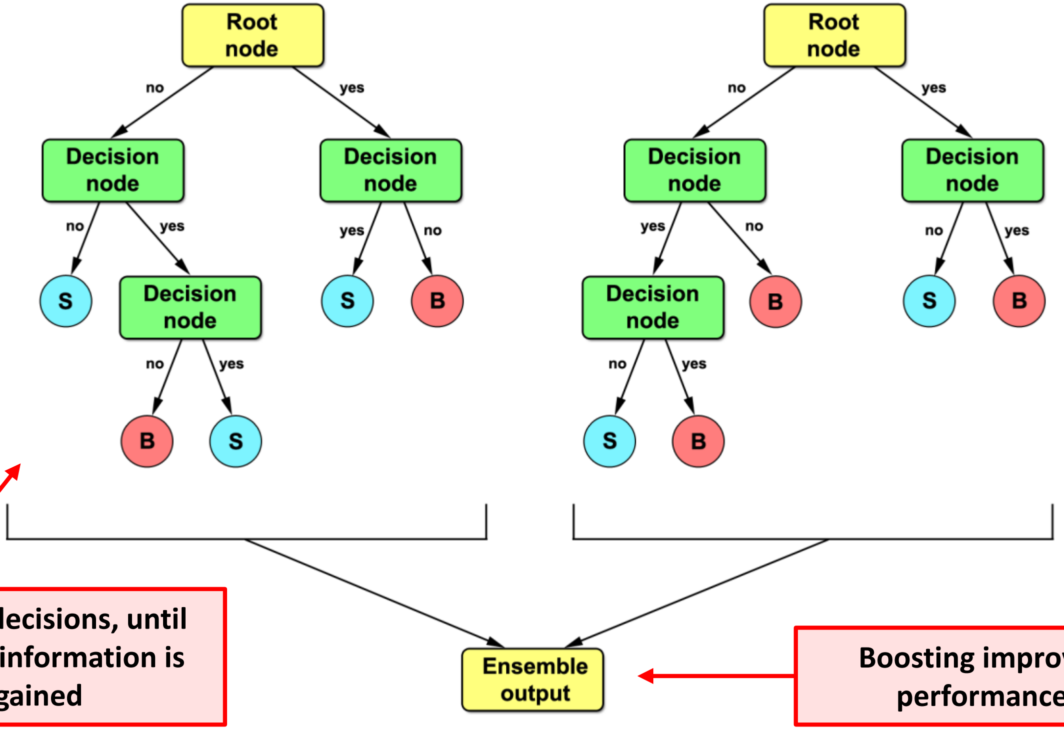


Multivariate analysis methods

- MVA methods determine the separation strength

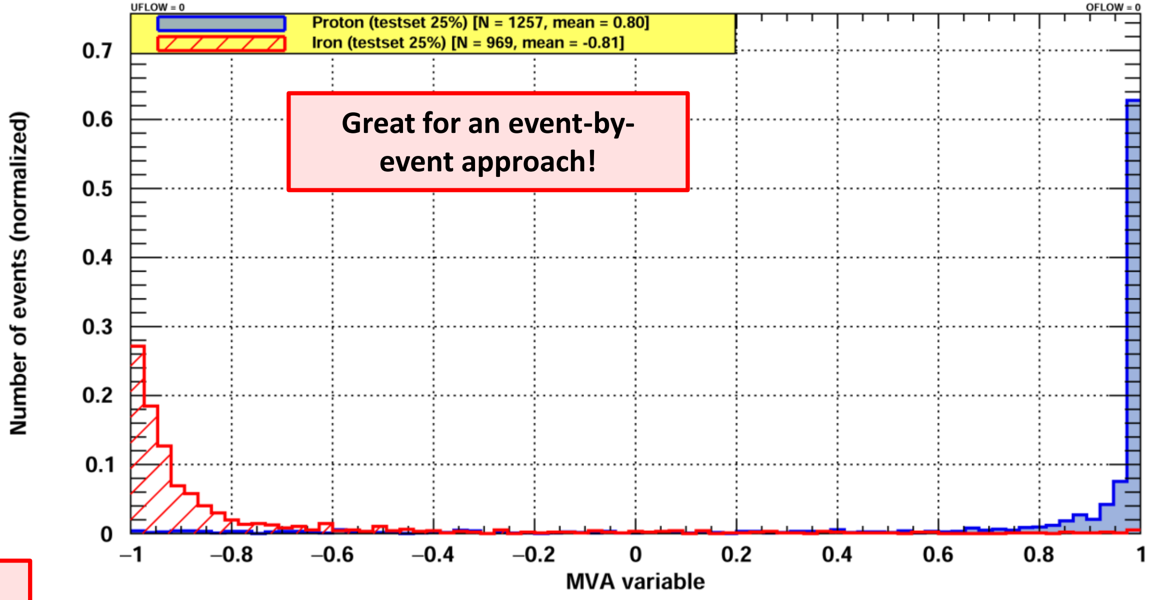
MVA method	No or linear correlations	Non-linear correlations	Training speed
Boosted decision trees (BDT)	Fair	Good	Fast
Multi-layer perceptrons (ANN)	Good	Good	Slow
Fisher linear discriminants	Good	Bad	Fast





Yes/no decisions, until no new information is gained

Boosting improves performance

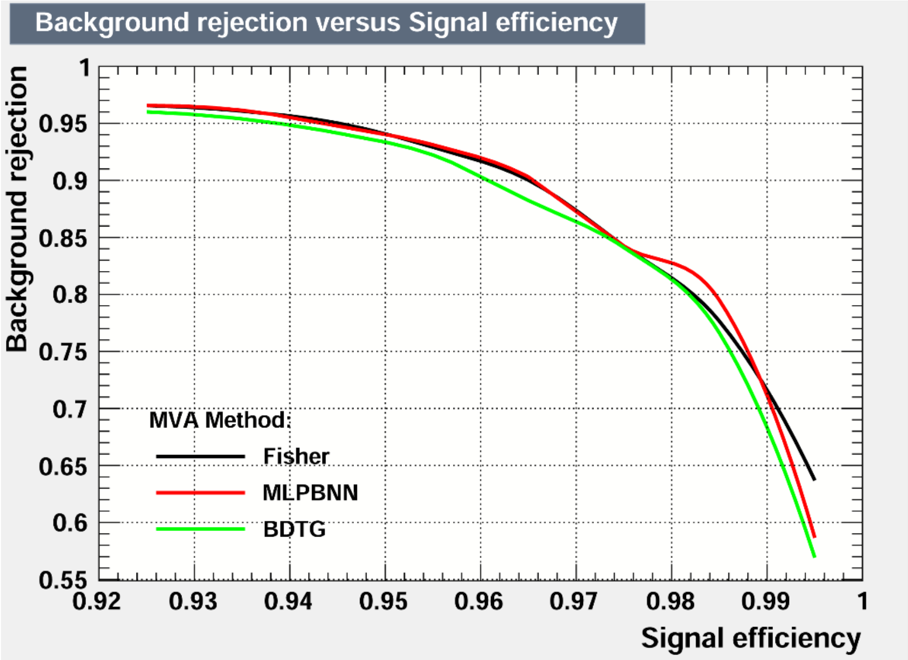


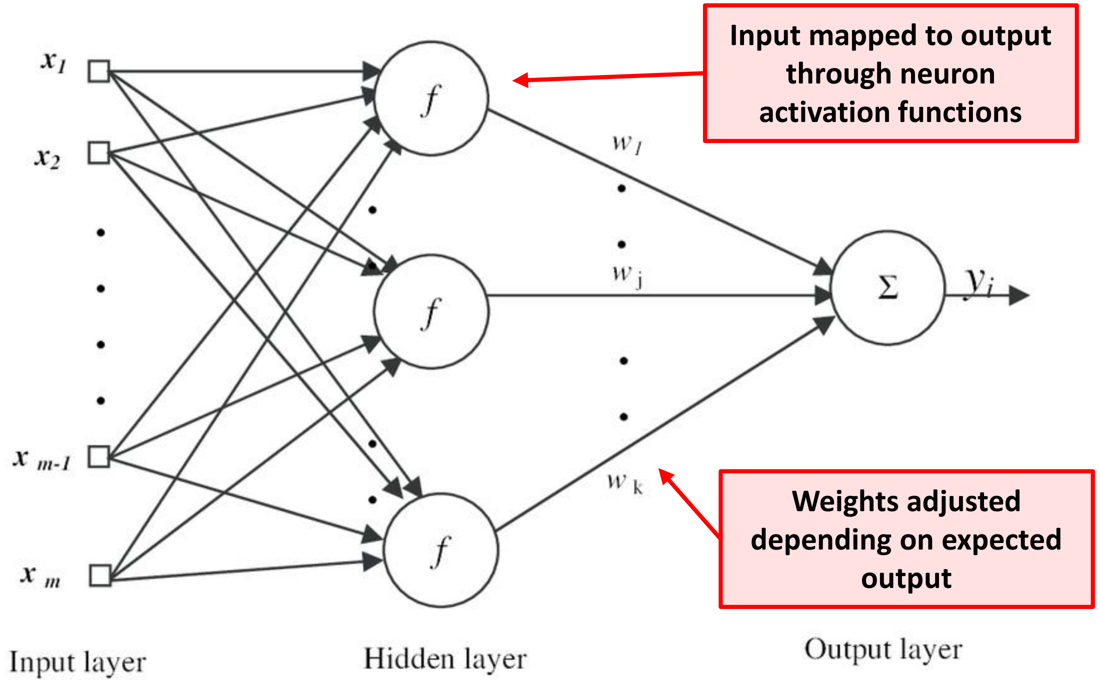
Great for an event-by-event approach!

Multivariate analysis methods

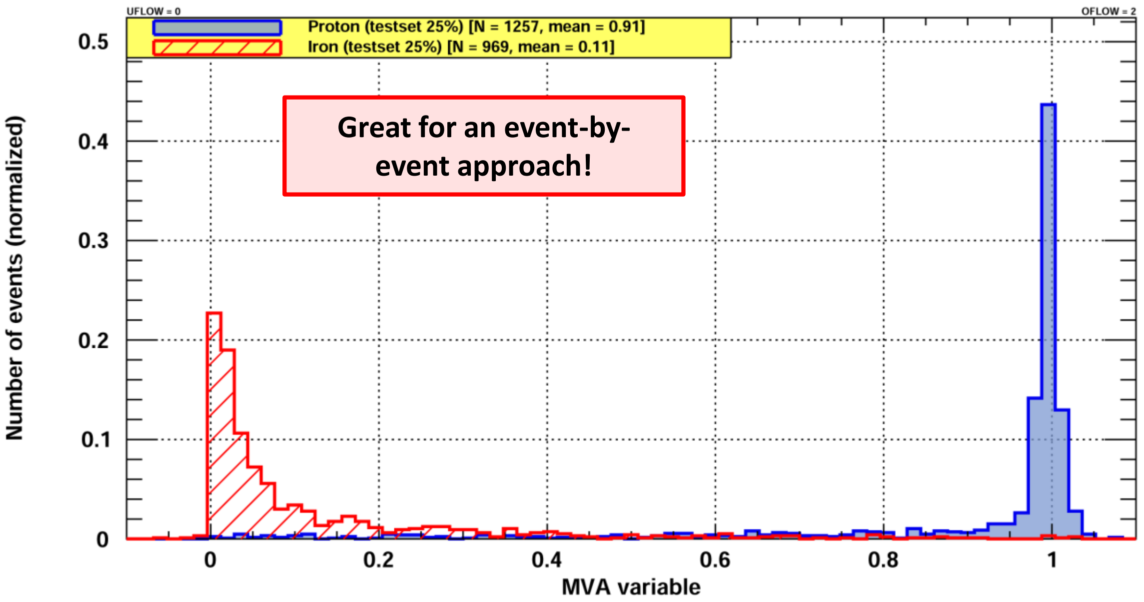
- MVA methods determine the separation strength

MVA method	No or linear correlations	Non-linear correlations	Training speed
Boosted decision trees (BDT)	Fair	Good	Fast
Multi-layer perceptrons (ANN)	Good	Good	Slow
Fisher linear discriminants	Good	Bad	Fast





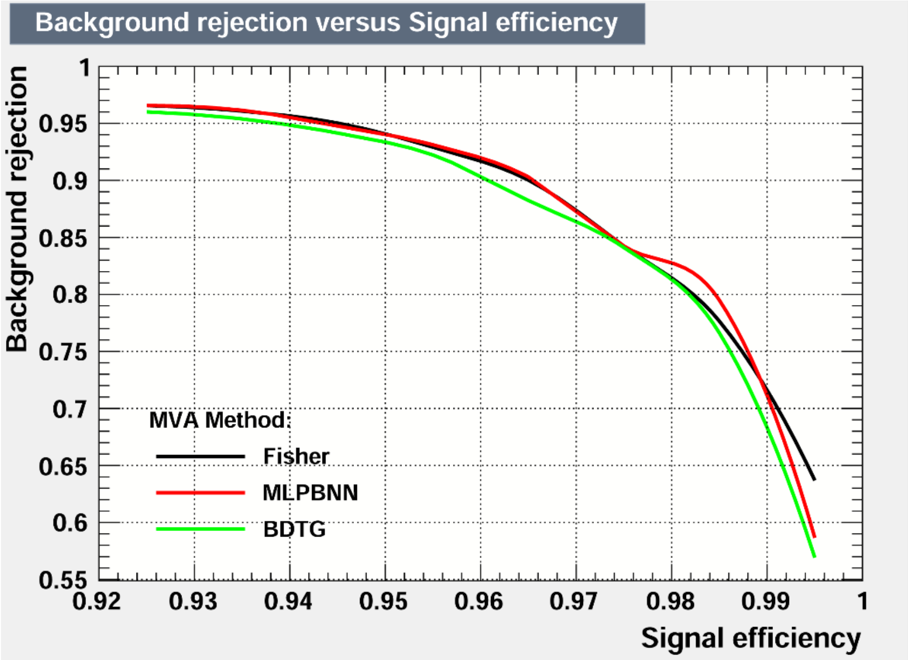
Pettern recognition and machine learning, Springer, 2016

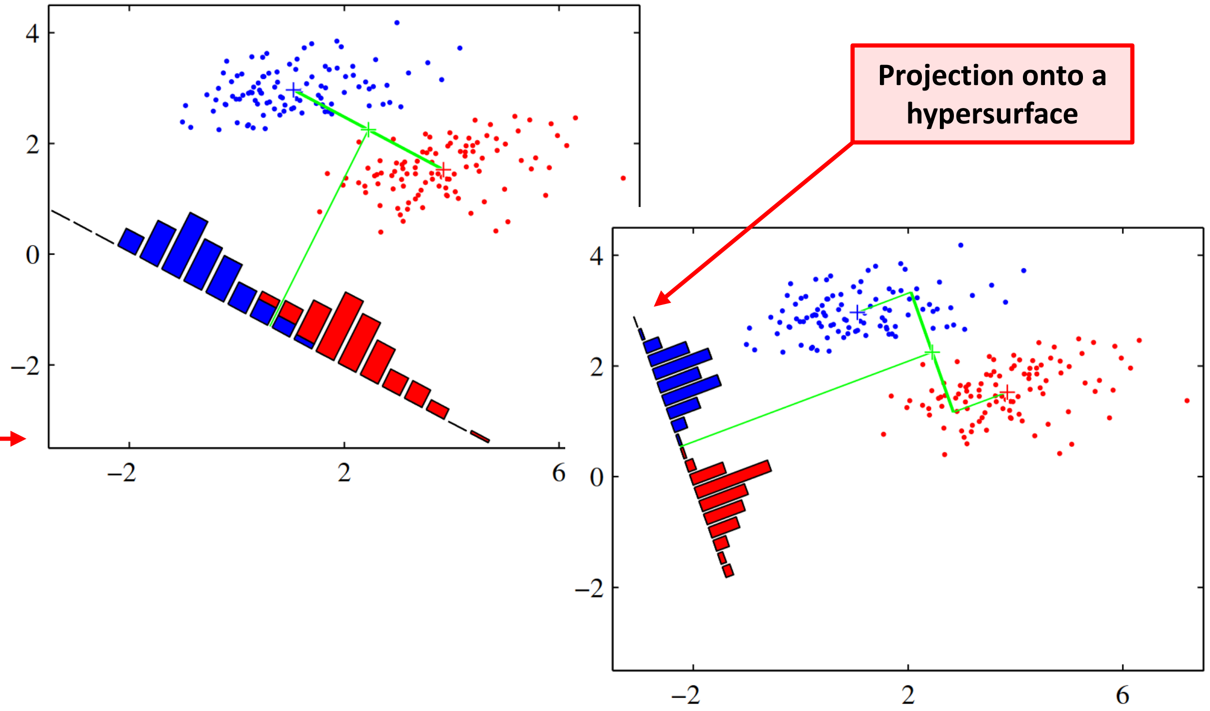


Multivariate analysis methods

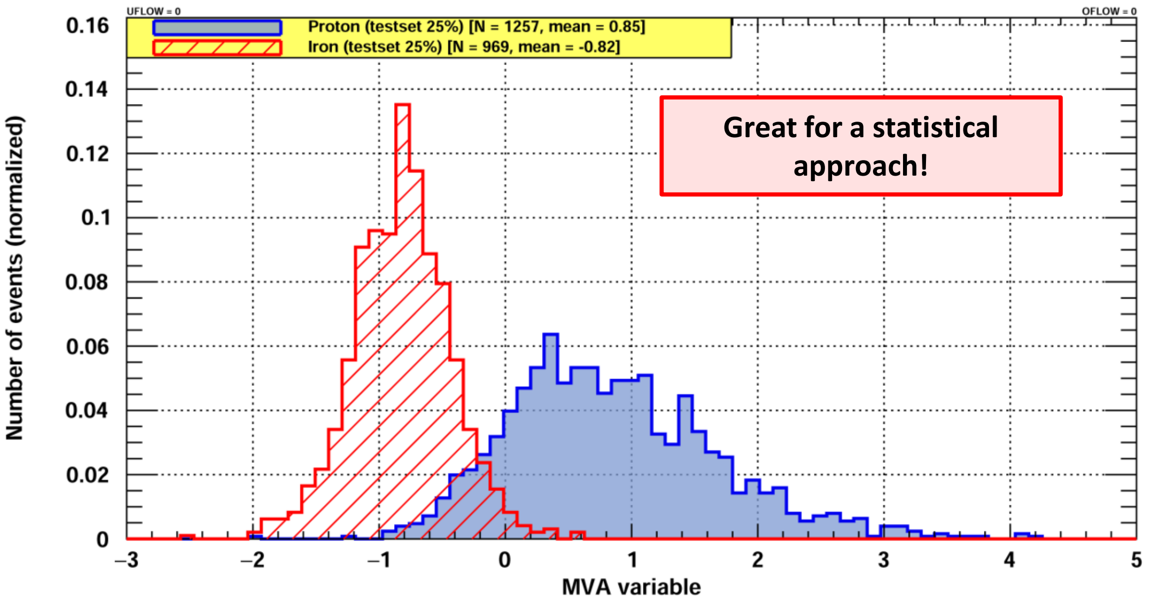
- MVA methods determine the separation strength

MVA method	No or linear correlations	Non-linear correlations	Training speed
Boosted decision trees (BDT)	Fair	Good	Fast
Multi-layer perceptrons (ANN)	Good	Good	Slow
Fisher linear discriminants	Good	Bad	Fast





Projection onto a hypersurface



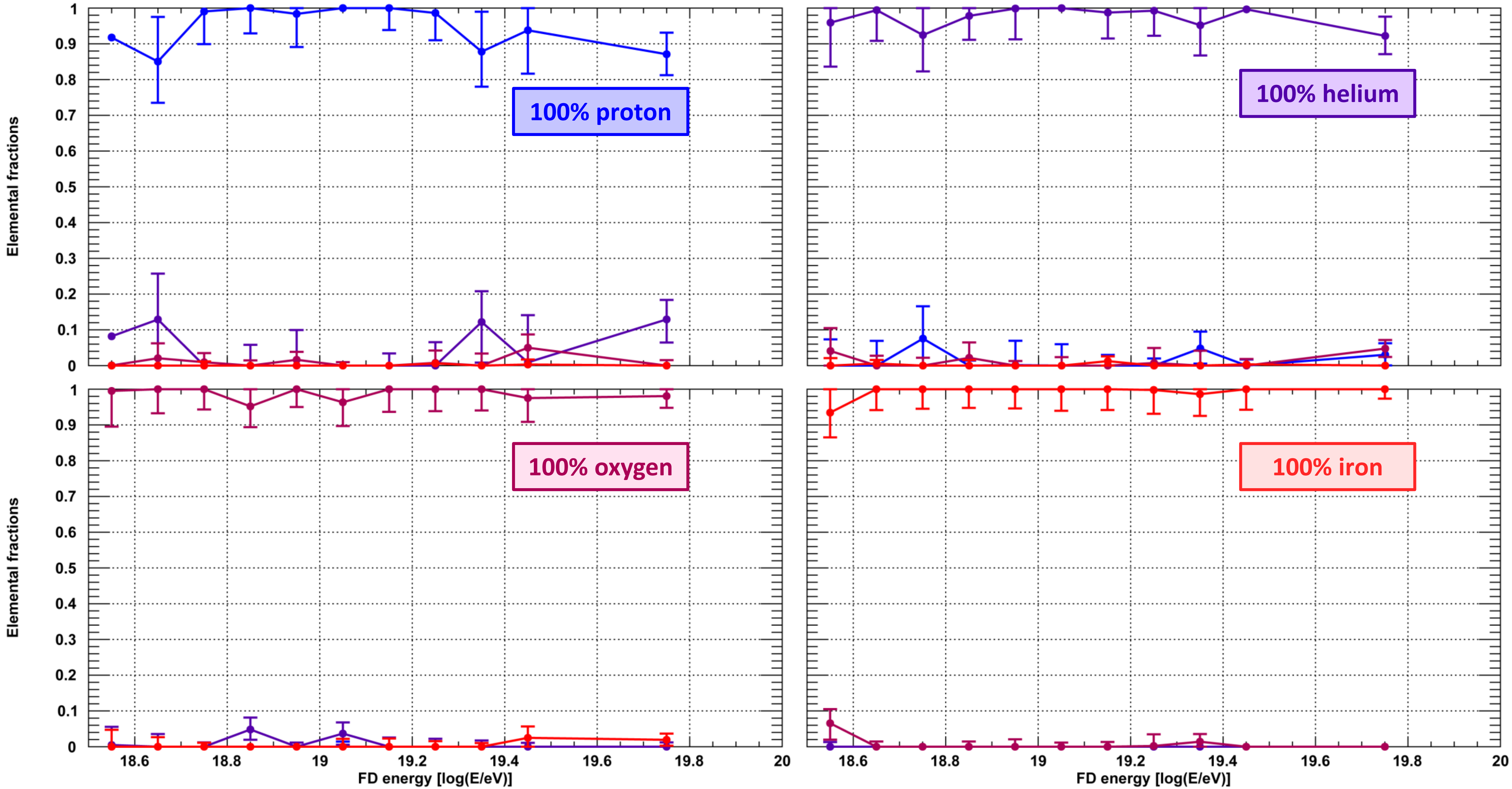
Great for a statistical approach!

www.researchgate.net/figure/Schematic-diagram-of-the-radial-basis-function-neural-network-for-one-output-45_fig1_258196355

Analysis of simulation samples

EPOS-LHC, Fisher

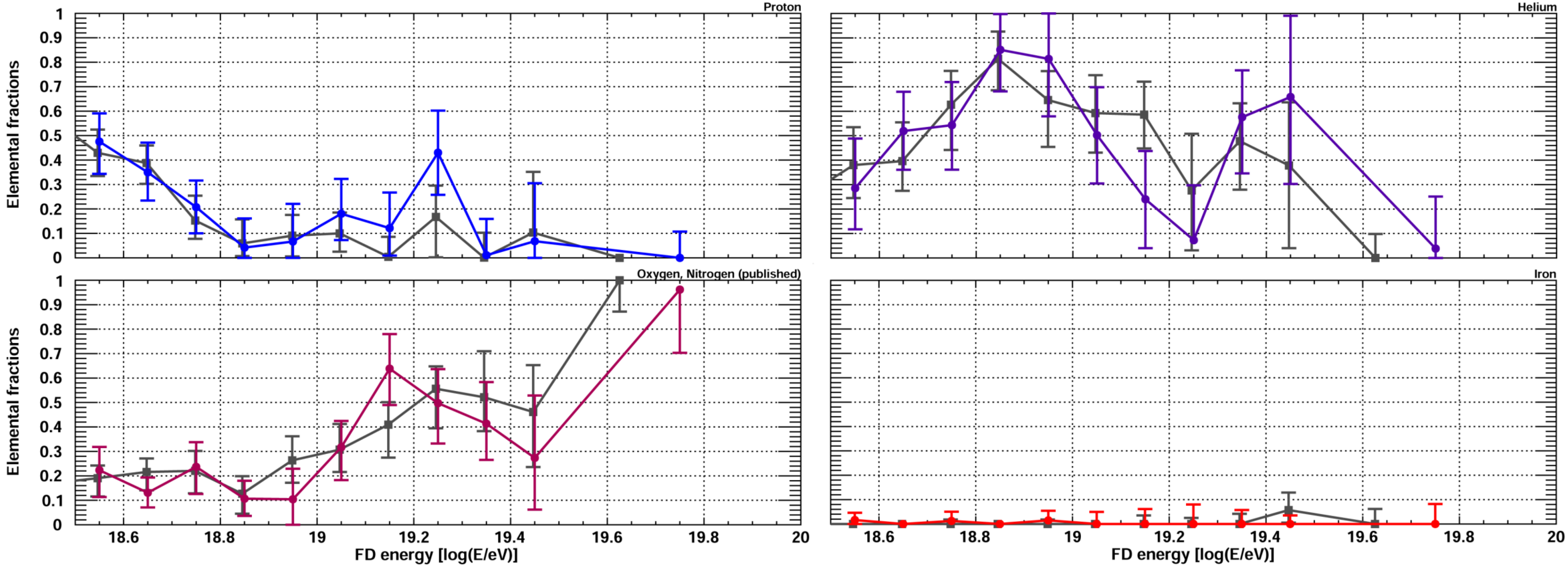
- Determine analysis method stability from the cross-validation simulation sample



Mock data set

EPOS-LHC, Fisher

- Mock data set (colors) imitates the published Pierre Auger Observatory mass composition (grey) [PoS(ICRC2017), PRD 90 (2014) 122006]
- Determines the performance on a mixed composition data set
- The same approach can then be applied to data (not the scope of this talk)



Thank you for your attention!

Backup slides

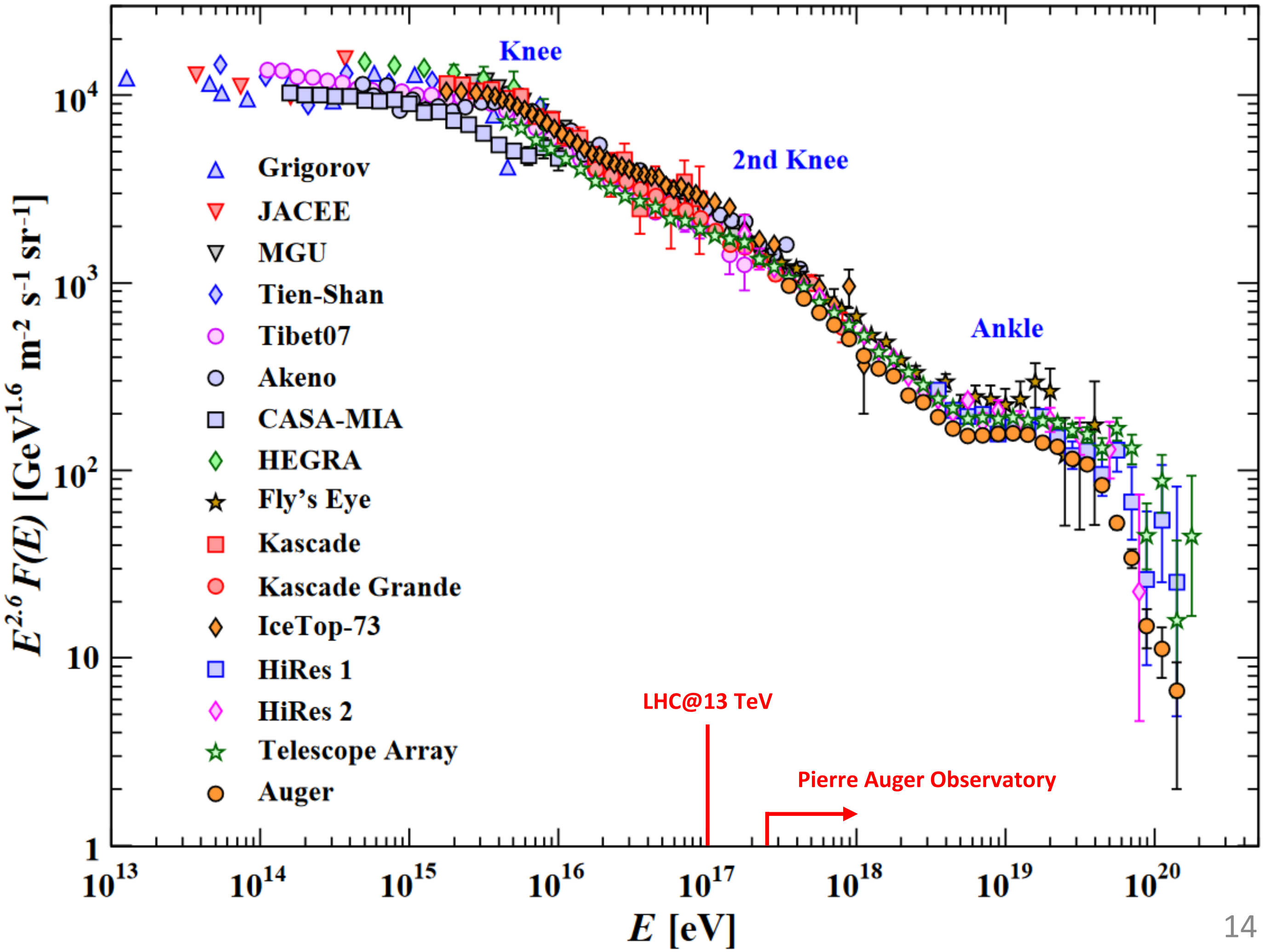
Introduction

- Cosmic rays (CR): Charged particles arriving to Earth from extraterrestrial sources
- Ultra-high energy cosmic rays (UHECR): CR with energies above $\sim 10^{18}\ eV$
- Energy spectrum features:
  - Knees – Exhaustion of galactic sources of CR
  - Ankle – Domination of extragalactic sources or GZK effect
  - GZK effect – Abrupt drop at the highest energies, scattering of protons and neutrons on cosmic microwave background (CMB) photons

$$p + \gamma_{CMB} \longrightarrow n + \pi^+$$

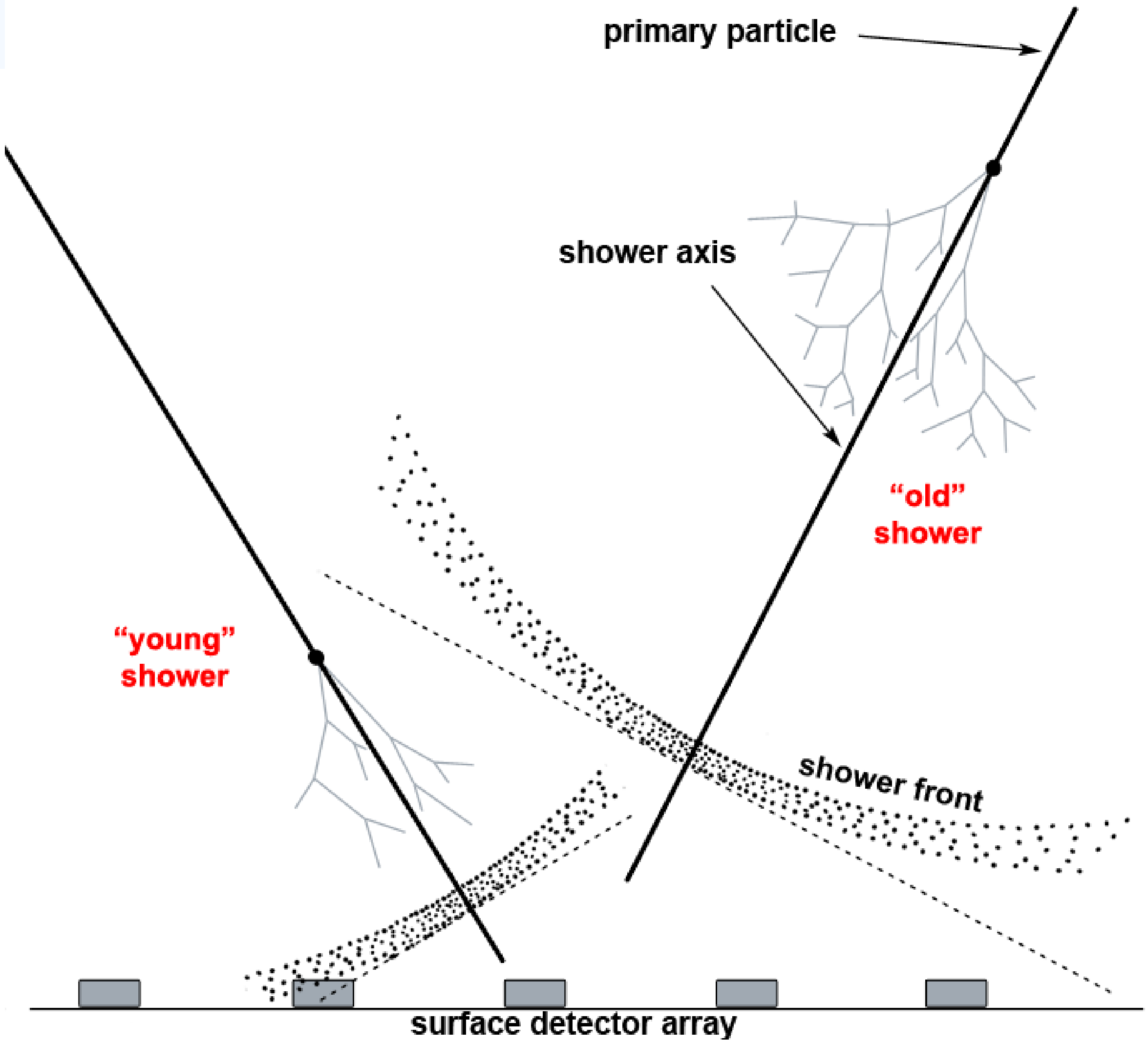
$$n + \gamma_{CMB} \longrightarrow p + \pi^-$$

[Chin. Phys. C40 (2016) 100001]



Extensive air showers

adapted from www.lip,pt/~jespada/Research/develop.jpg



- Extensive air shower (EAS): Cascade of secondary particles after interaction of UHECR and atmospheric nuclei
- Main EAS parts:
  - Electromagnetic part (electrons, positrons, photons)
  - Hadronic part (hadrons and mesons)
  - Weakly interacting shower remnants (muons and neutrinos)
- Primary particle determines the evolution of the EAS:
  - EAS develops higher in the atmosphere for heavier particles (larger interaction cross section)
  - EAS develops lower in the atmosphere for lighter and weakly interacting particles

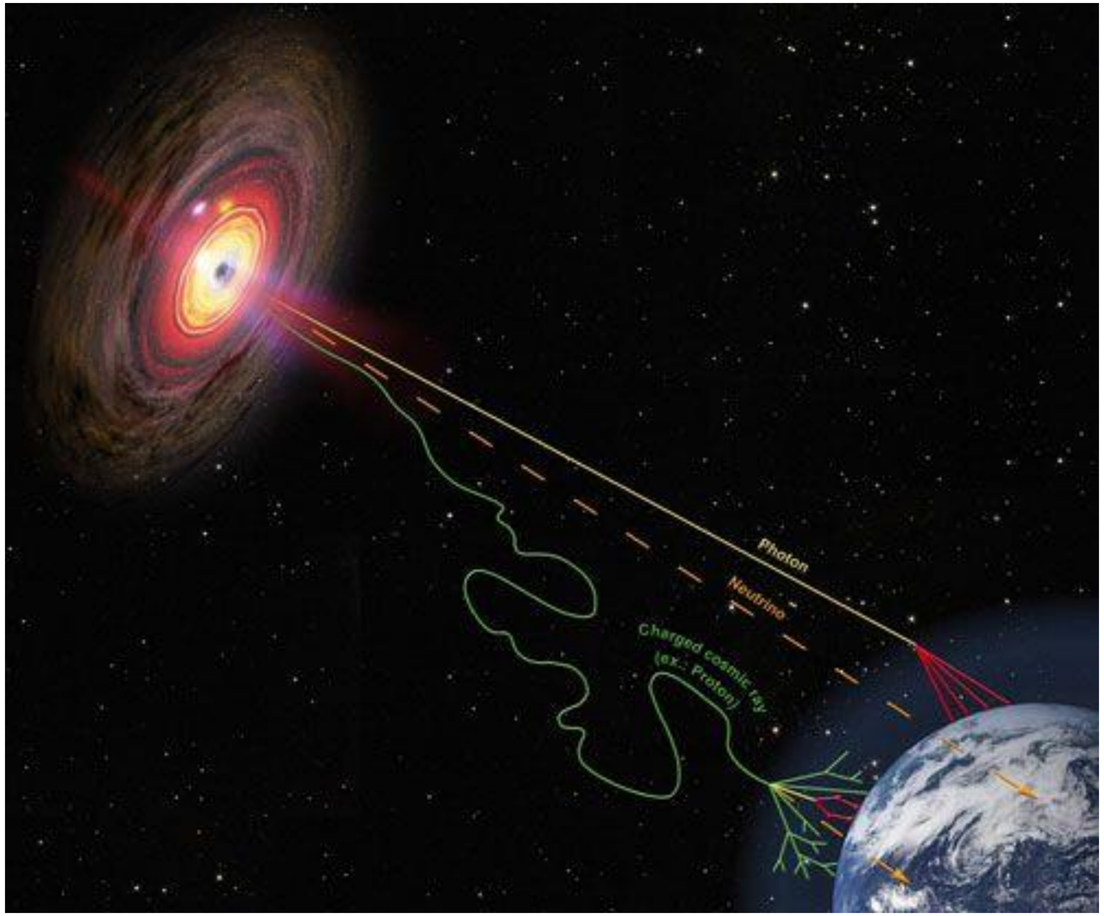
Mass composition of UHECR

- Mass composition studies: Determine mass and charge of UHECR
- Motivations for performing mass composition studies:
  - Discrimination between hadronic interaction models
  - Backtracking of light UHECR with energies $> 10^{19} eV$ to their sources

$$\Delta\alpha = \frac{Zec}{E}\int_0^L B(x)\sin(\varphi(x))\,dx$$

  - Acceleration processes that produce UHECR
  - Cosmic magnetic field strength
  - Identifying energy spectrum features
- Main drawback: Mass composition highly dependent on hadronic interaction models (extrapolated cross-sections)

steemit.com/science/@shehzad/understanding-cosmic-rays-in-a-simple-way

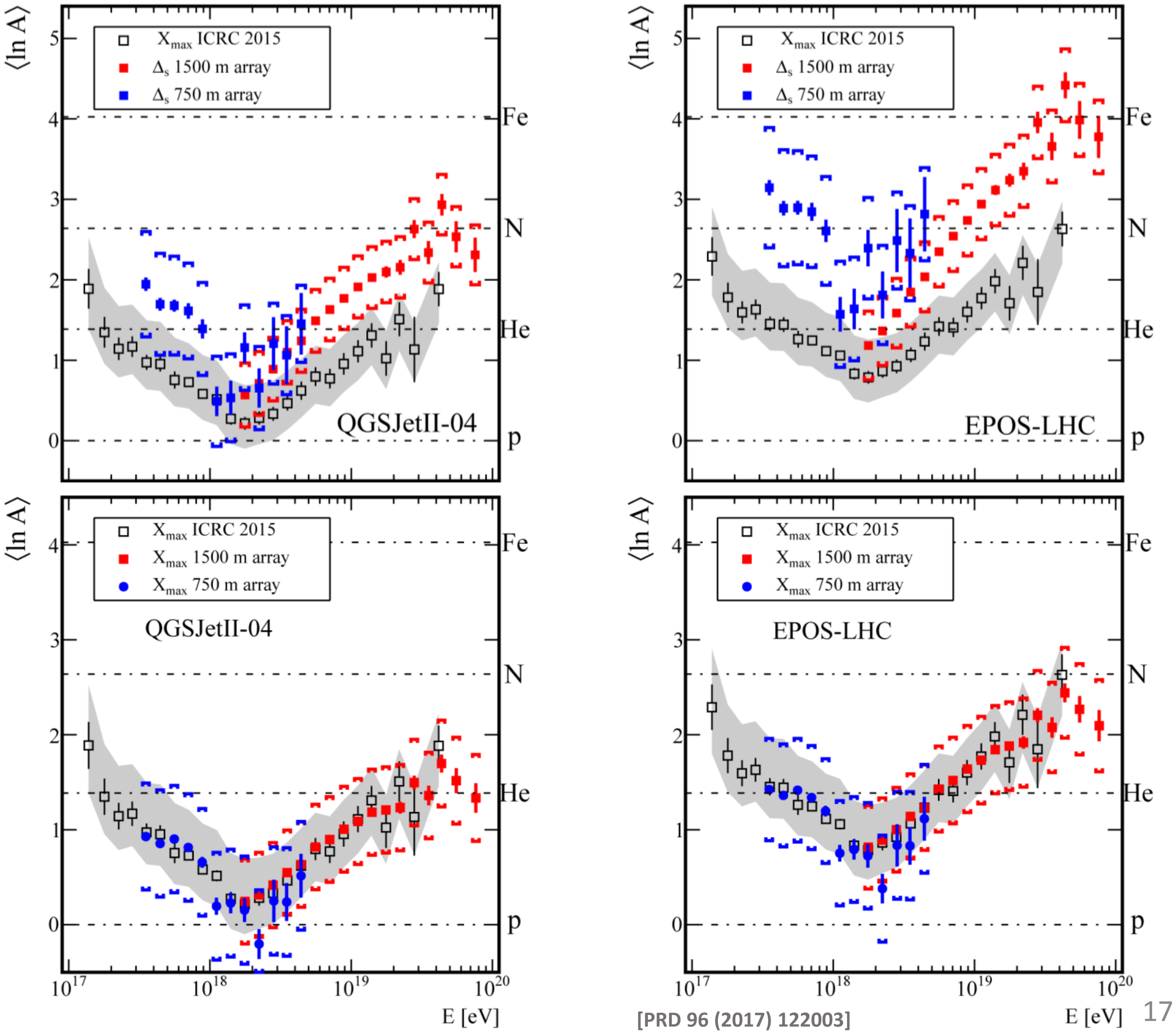


Existing mass composition results

- Results of SD-only Delta method [PRD 96 (2017) 122003]
- Our conversion of risetime $t_{1/2}$ to Δ_R is based on this work
- Average mass estimator:

$$\langle \ln A \rangle = \ln 56 \cdot \frac{\langle \Delta_s \rangle_p - \langle \Delta_s \rangle_{data}}{\langle \Delta_s \rangle_p - \langle \Delta_s \rangle_{Fe}}$$

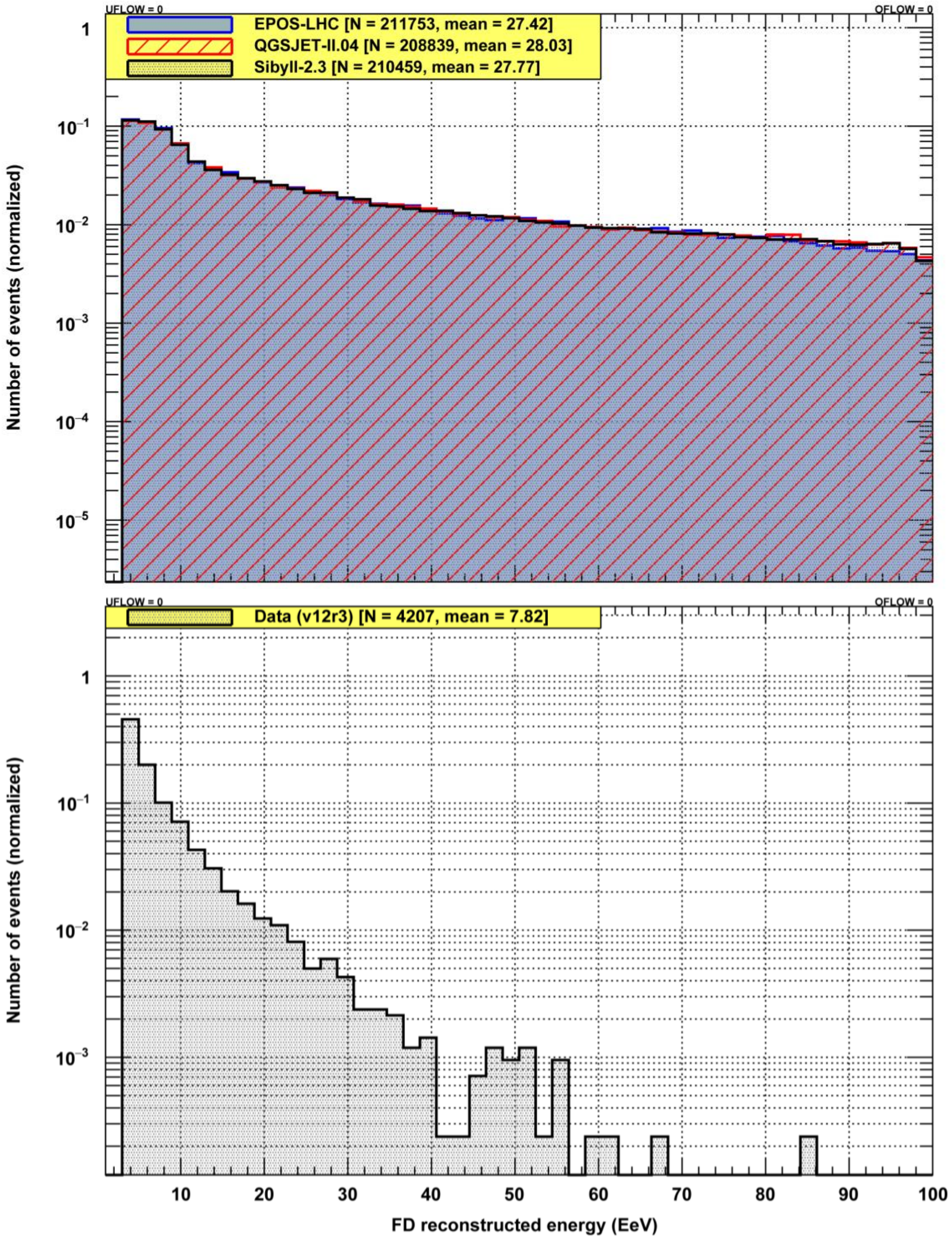
- Results from the Delta method is then calibrated with X_{max} analysis results
- Discrepancy explained as the inability of hadronic interaction models to predict muonic content



[PRD 96 (2017) 122003]

Simulations and data

- Simulations from the Napoli shower library:
  - Three hadronic interaction models (EPOS-LHC, QGSJET-II.04 and Sibyll-2.3)
  - Four primary particle masses (proton, helium, oxygen and iron)
  - Energies between $10^{18.5} eV$ and $10^{20.0} eV$
- Data from the Pierre Auger Observatory:
  - Hybrid events with SD and FD measurements
  - Covering measurements between 1.12.2004 and 31.12.2015
  - Energies between $10^{18.5} eV$ and $10^{20.0} eV$
- Both sets taken through selection cuts, taking only high quality hybrid events



Simulations and data

- Simulations are split into three sets for the MVA analysis:
  - MVA training set: Training the MVA method, determining elemental fractions after the MVA analysis
  - Cross-validation set: Estimating the stability of the analysis method with simulation events, that were not used during MVA method training
  - AugerMix set: A controlled mock data set that aims to imitate previously published mass composition results [PoS(ICRC2017), PRD 90 (2014) 122006]
- Size of the cross-validation set is $^{1}/_{3}$ of the MVA training set
- AugerMix mock data set has the same number of events as Pierre Auger Observatory data

Energy $[\log(E/eV)]$	Number of data events	
	FD-only	SD+FD
18.5 – 18.6	1108	824
18.6 – 18.7	840	627
18.7 – 18.8	583	463
18.8 – 18.9	471	370
18.9 – 19.0	359	259
19.0 – 19.1	281	214
19.1 – 19.2	193	139
19.2 – 19.3	134	106
19.3 – 19.4	110	80
19.4 – 19.5	66	45
19.5 – 20.0	62	45
	Unlimited zenith angle	Limited to $\theta = [0°, 60°]$

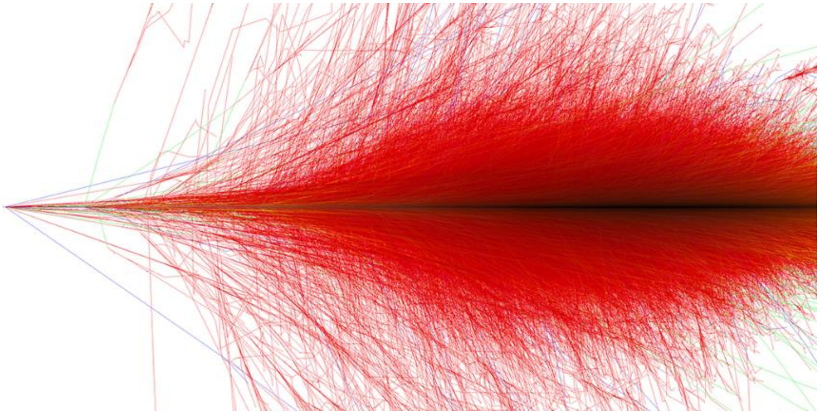
Analysis observables

- Taking mass composition sensitive observables:
  - **Depth of shower maximum (X_{max})**
  - SD signal at $1000\ m$ from the shower axis (S_{1000})
  - Risetime at $1000\ m$ from the shower axis (t_{1000})
- S_{1000} and t_{1000} depend on zenith angle θ – convert to relative observables ΔS_{38} and Δ_R

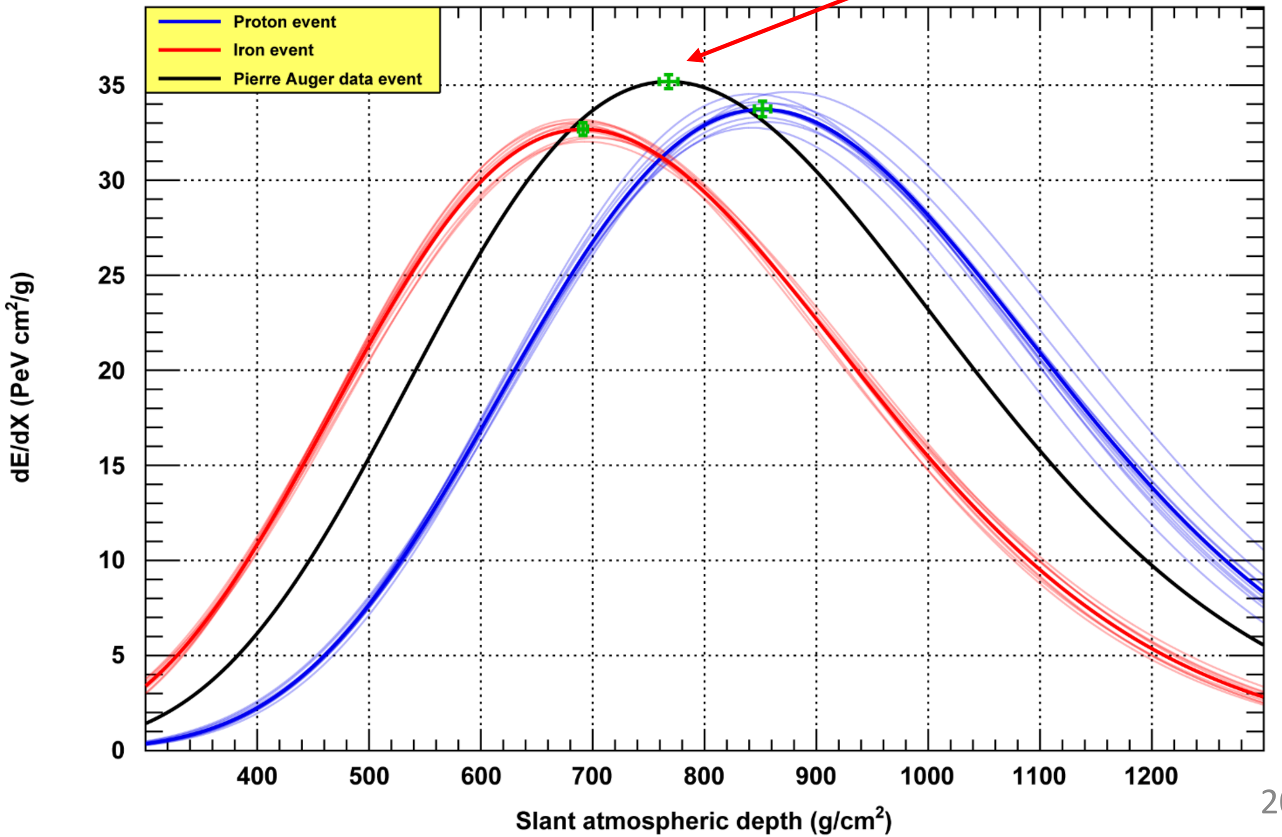
Depth at which the EAS reaches the maximum number of secondary particles

Heavy < Light

www-zeuthen.desy.de/~jknapp/fs/proton-showers.html



Depth of shower maximum



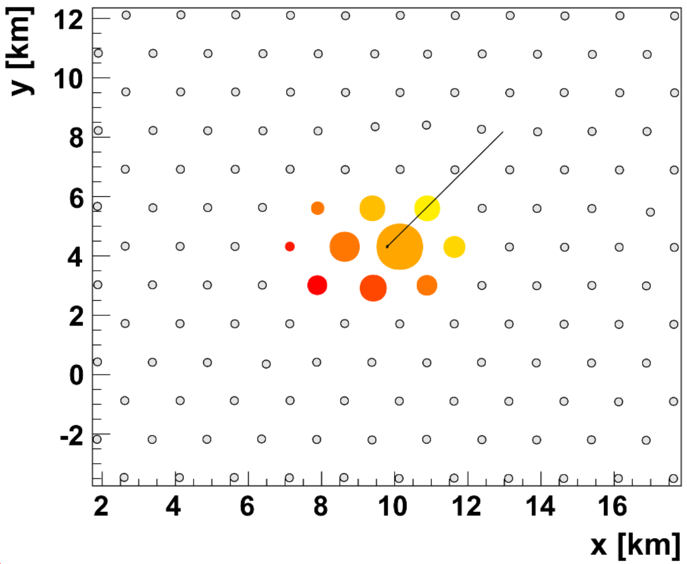
Analysis observables

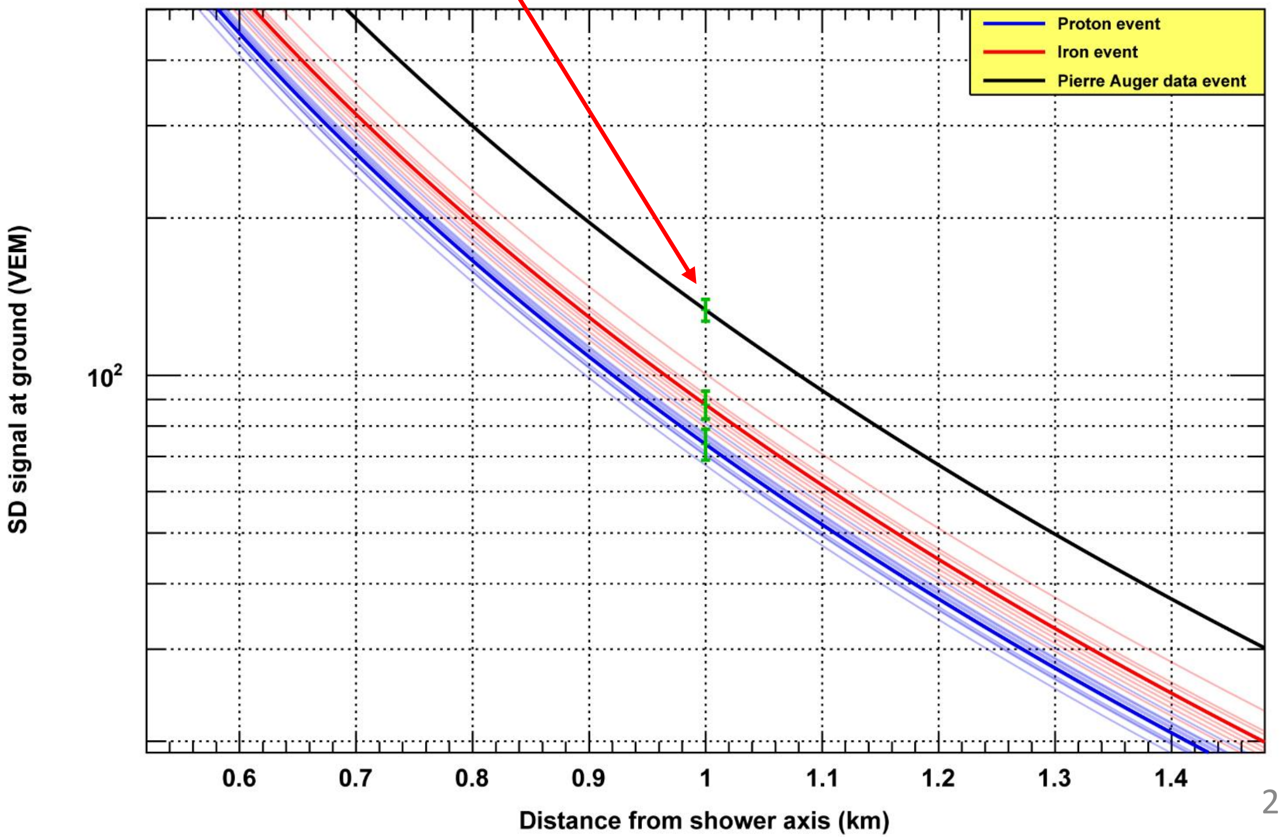
- Taking mass composition sensitive observables:
  - Depth of shower maximum (X_{max})
  - **SD signal at $1000\ m$ from the shower axis (S_{1000})**
  - Risetime at $1000\ m$ from the shower axis (t_{1000})
- S_{1000} and t_{1000} depend on zenith angle θ – convert to relative observables ΔS_{38} and Δ_R

Distribution of SD station signals around the shower axis

Heavy > Light

Signal at $1000\ m$ from shower axis



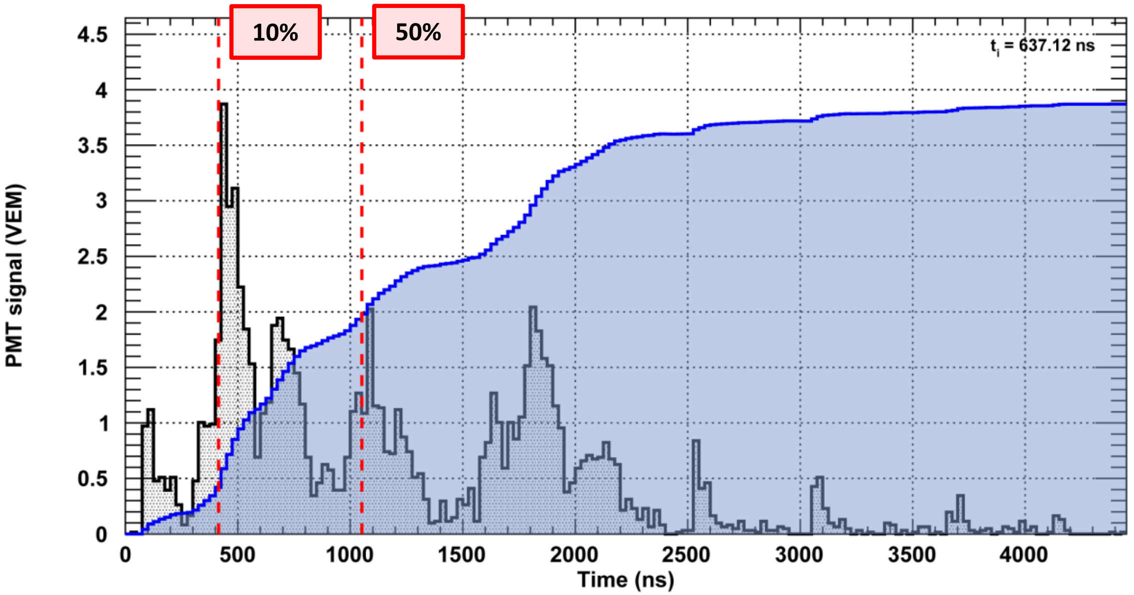


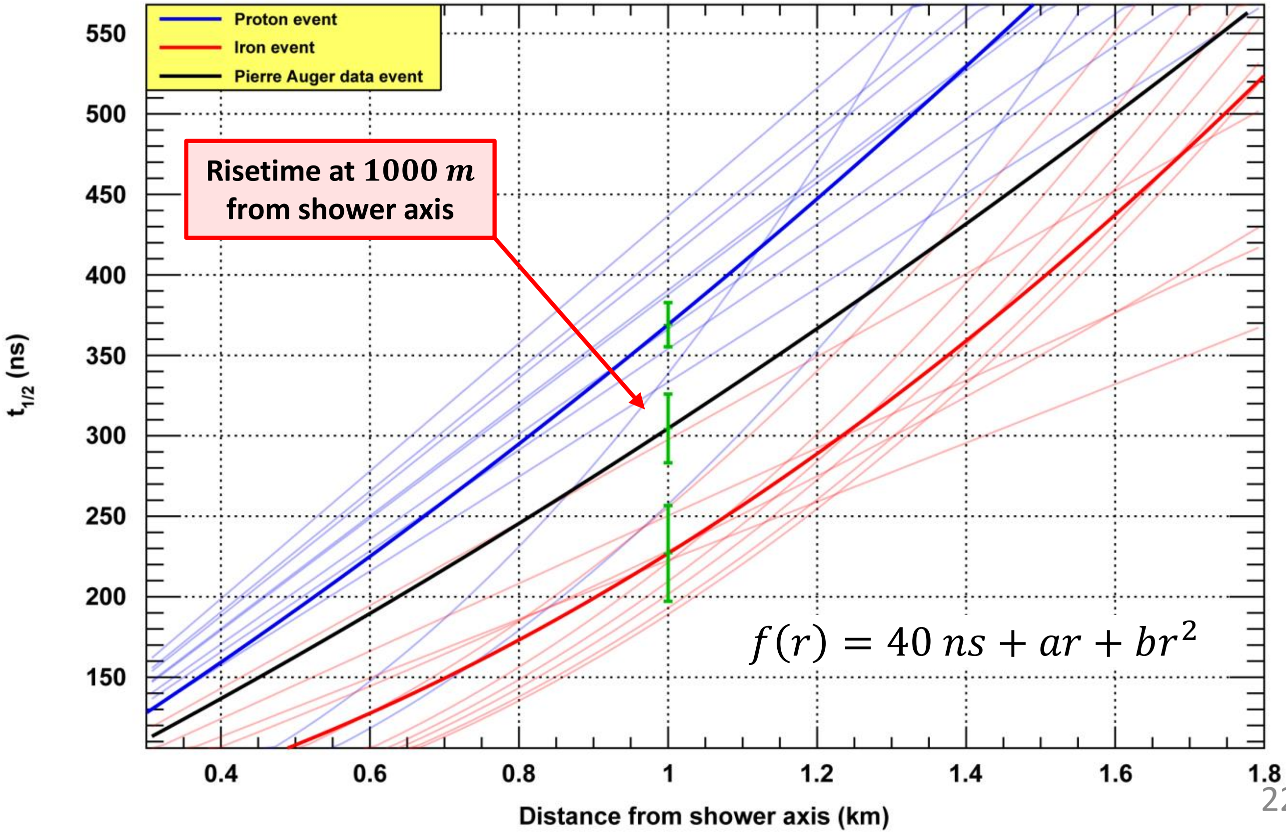
Analysis observables

- Taking mass composition sensitive observables:
  - Depth of shower maximum (X_{max})
  - SD signal at $1000\ m$ from the shower axis (S_{1000})
  - **Risetime at $1000\ m$ from the shower axis (t_{1000})**
- S_{1000} and t_{1000} depend on zenith angle θ – convert to relative observables ΔS_{38} and Δ_R

Muon versus electromagnetic content in SD station signals – shower age indicator

Heavy < Light





Analysis observables - ΔS_{38}

- Removing zenith angle dependency from S_{1000} to get S_{38}

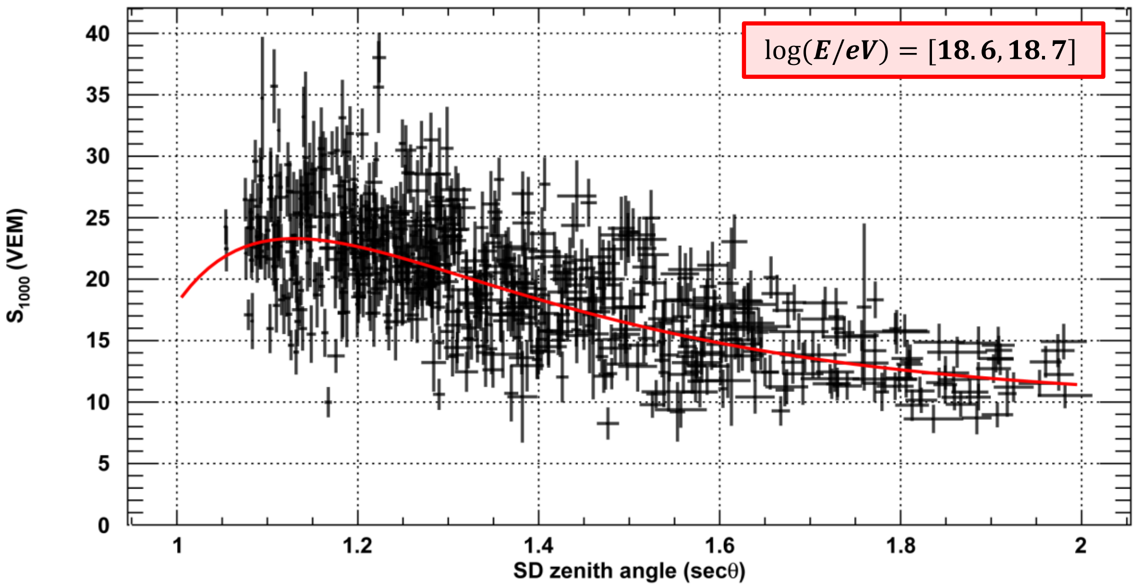
$$S_{38} = \frac{S_{1000}}{f_{CIC}(\theta)}$$

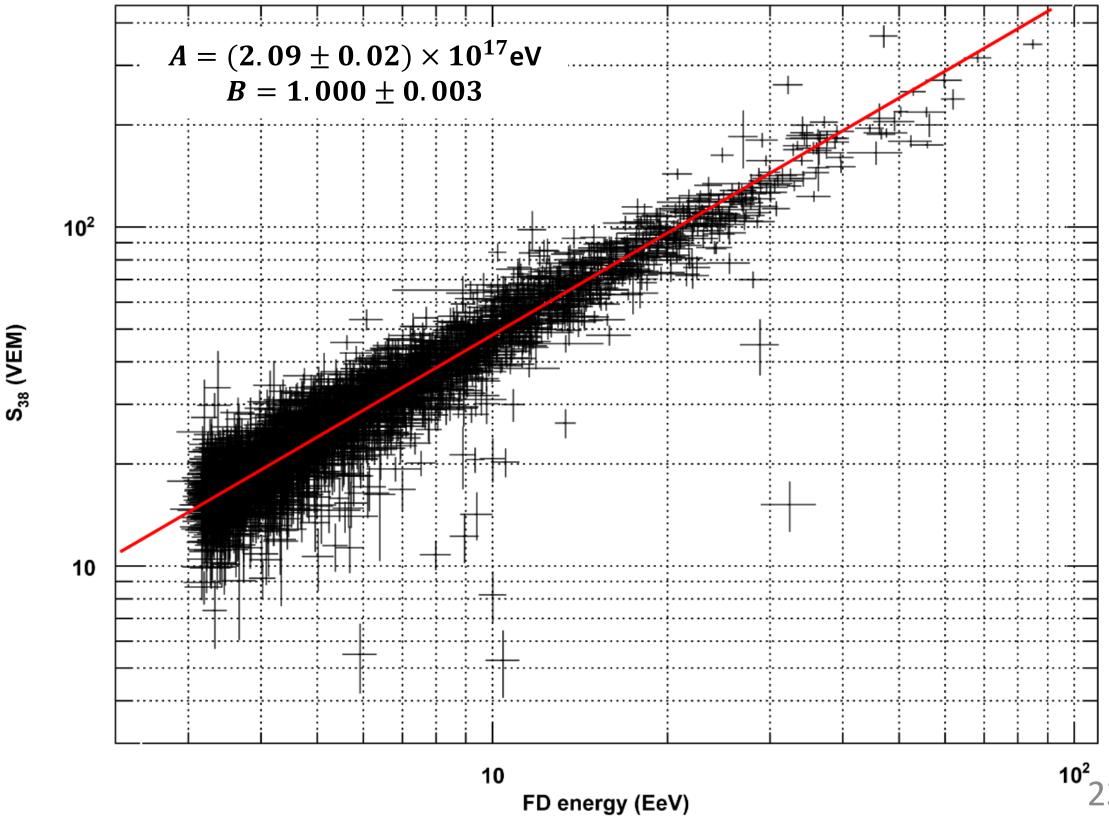
$$f_{scale}(\theta) = S \cdot f_{CIC}(\theta) = S \cdot (1 + ax + bx^2 + cx^3)$$

$$x = \cos^2\theta - \cos^2(38°)$$

- S_{38} values determined for each of the 11 energy bins
- Relative observable from a power-law fit

$$\Delta S_{38} = S_{38} - \left(\frac{E_{FD}}{A}\right)^{1/B}$$





Analysis observables - ΔS_{38}

- Removing zenith angle dependency from S_{1000} to get S_{38}

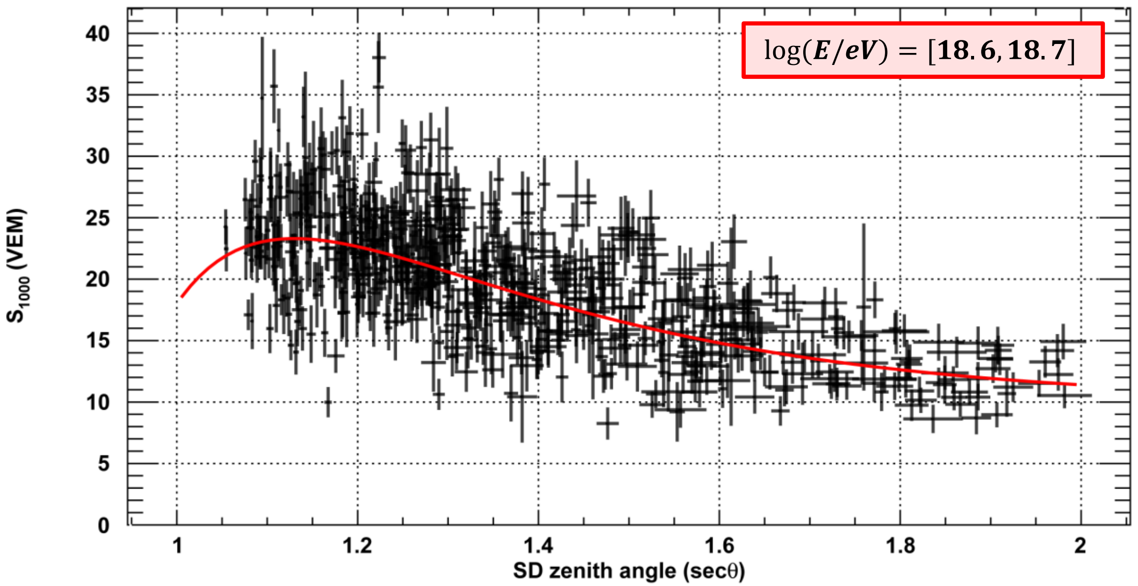
$$S_{38} = \frac{S_{1000}}{f_{CIC}(\theta)}$$

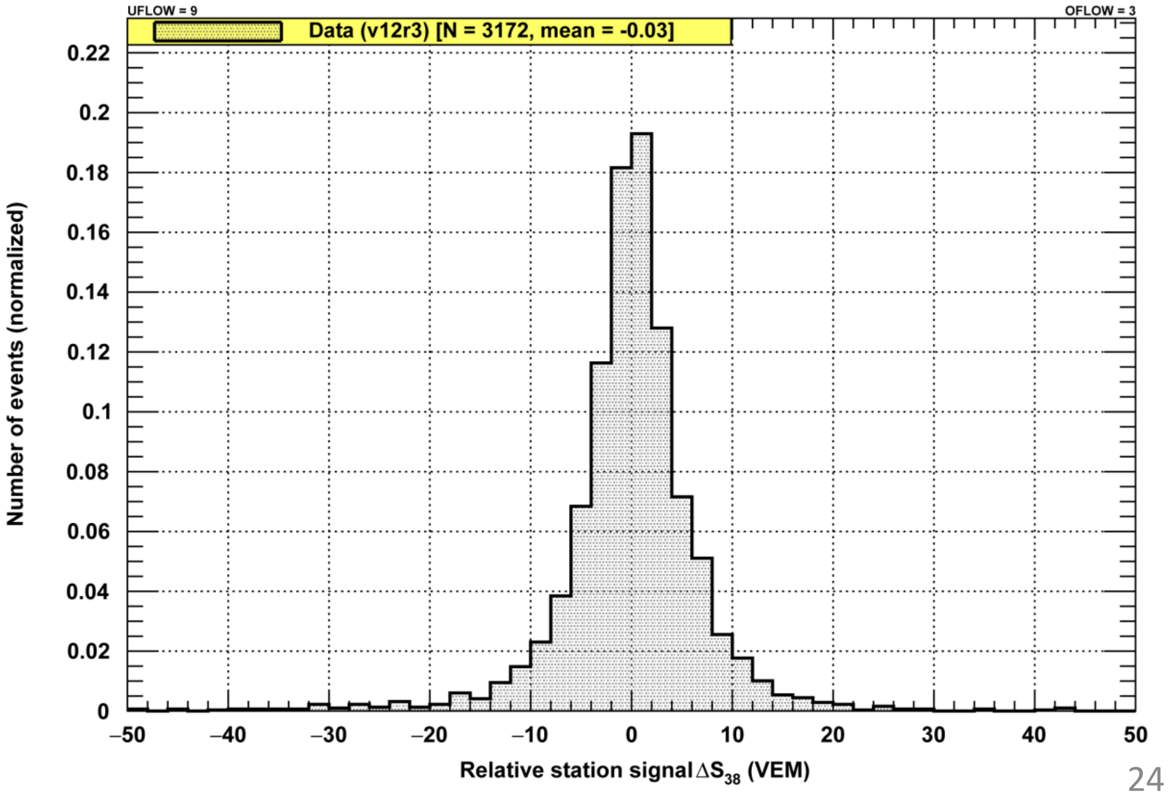
$$f_{scale}(\theta) = S \cdot f_{CIC}(\theta) = S \cdot (1 + ax + bx^2 + cx^3)$$

$$x = \cos^2\theta - \cos^2(38°)$$

- S_{38} values determined for each of the 11 energy bins
- Relative observable from a power-law fit

$$\Delta S_{38} = S_{38} - \left(\frac{E_{FD}}{A}\right)^{1/B}$$





Analysis observables - Δ_R

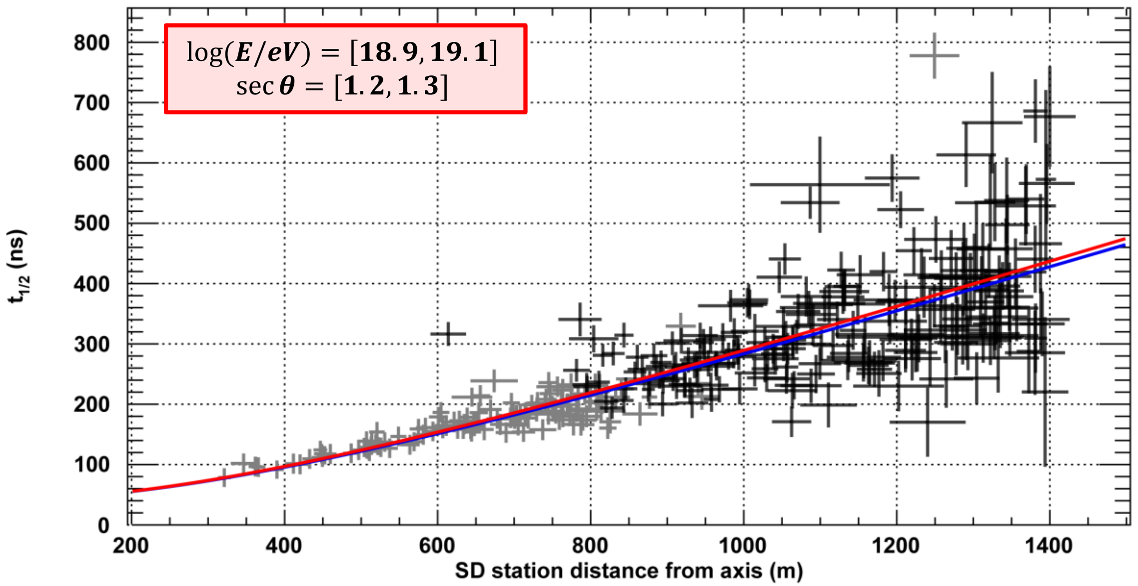
- Removing distance from shower axis dependency by fitting benchmark functions

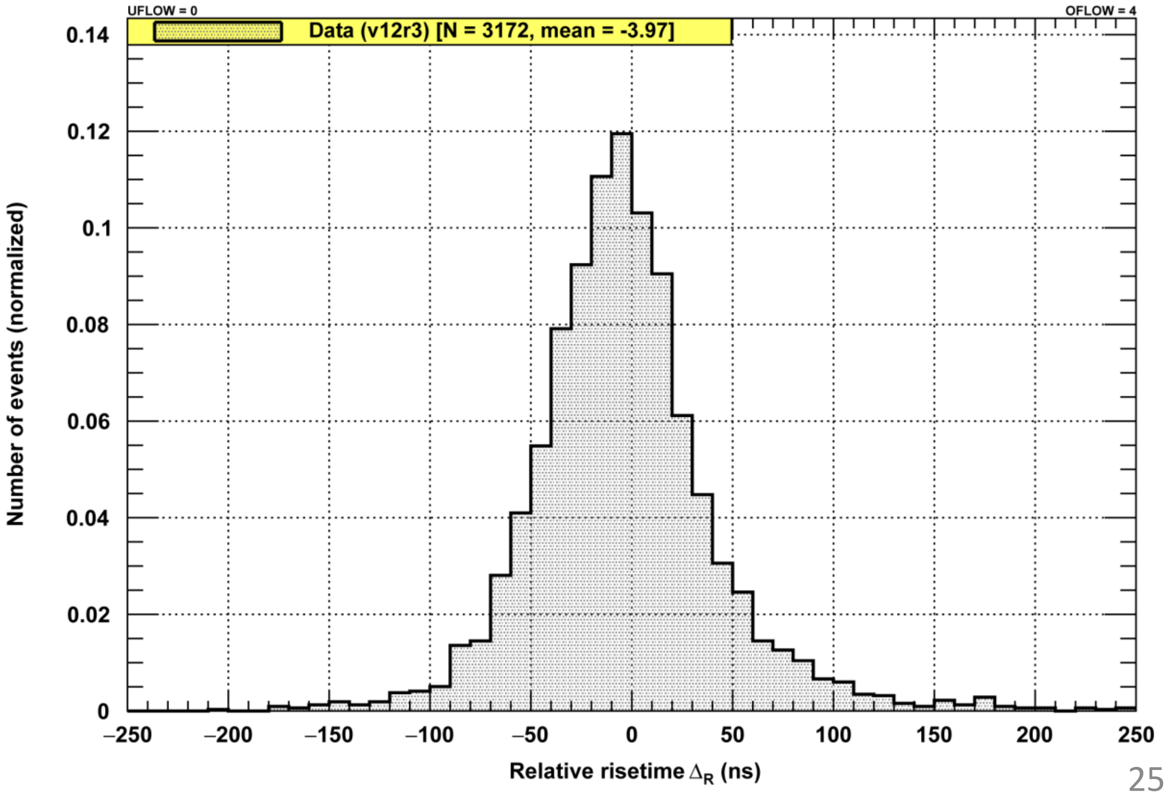
$$t_{1/2}^{bench,HGsat} = 40\ ns + \sqrt{A^2 + B^2 r^2} - A$$
$$t_{1/2}^{bench} = 40\ ns + M\left(\sqrt{A^2 + B^2 r^2} - A\right)$$

- Benchmark functions determined for 10 zenith angle bins and a reference energy bin – removing zenith angle dependence

- Combine station relative risetimes Δ_i into a relative risetime observable Δ_R

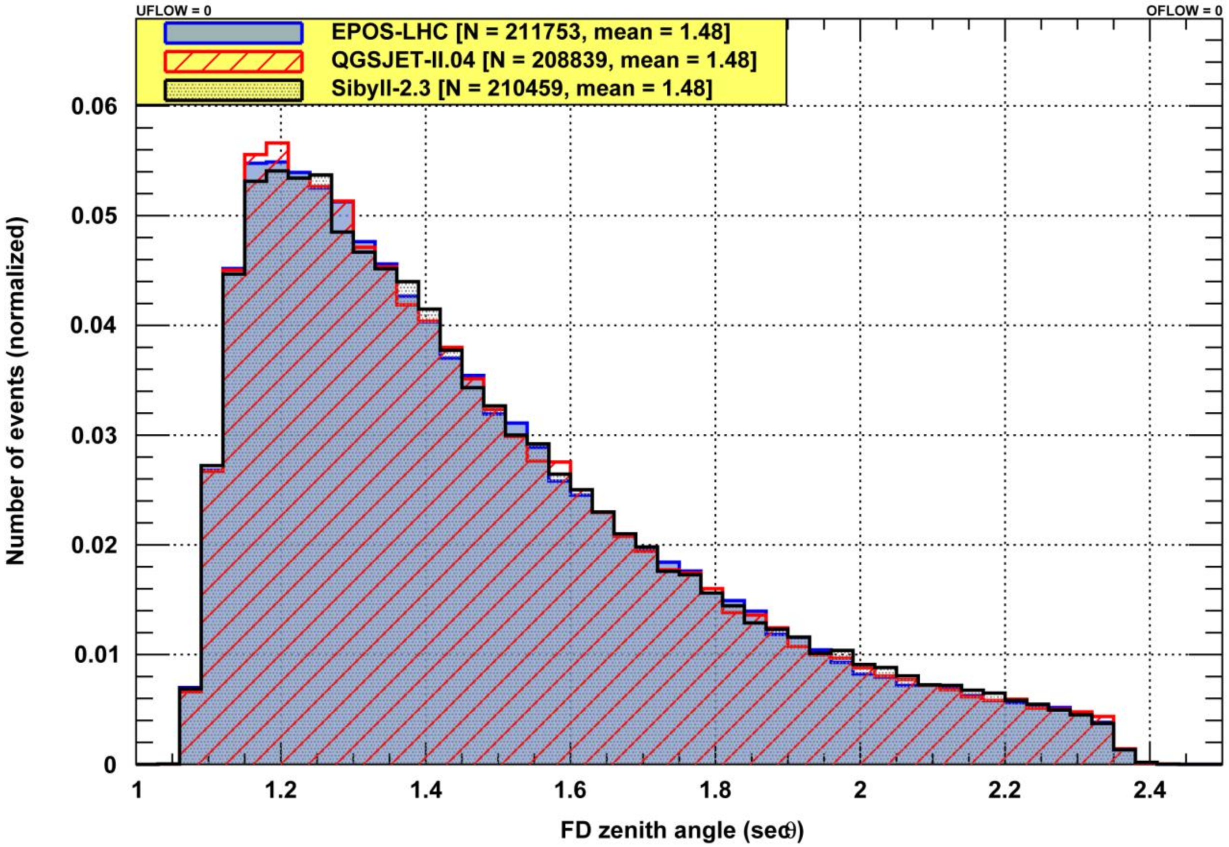
$$\Delta_i = t_{1/2} - t_{1/2}^{bench} \qquad \Delta_R = \frac{1}{N}\sum_{i=1}^{N}\Delta_i$$

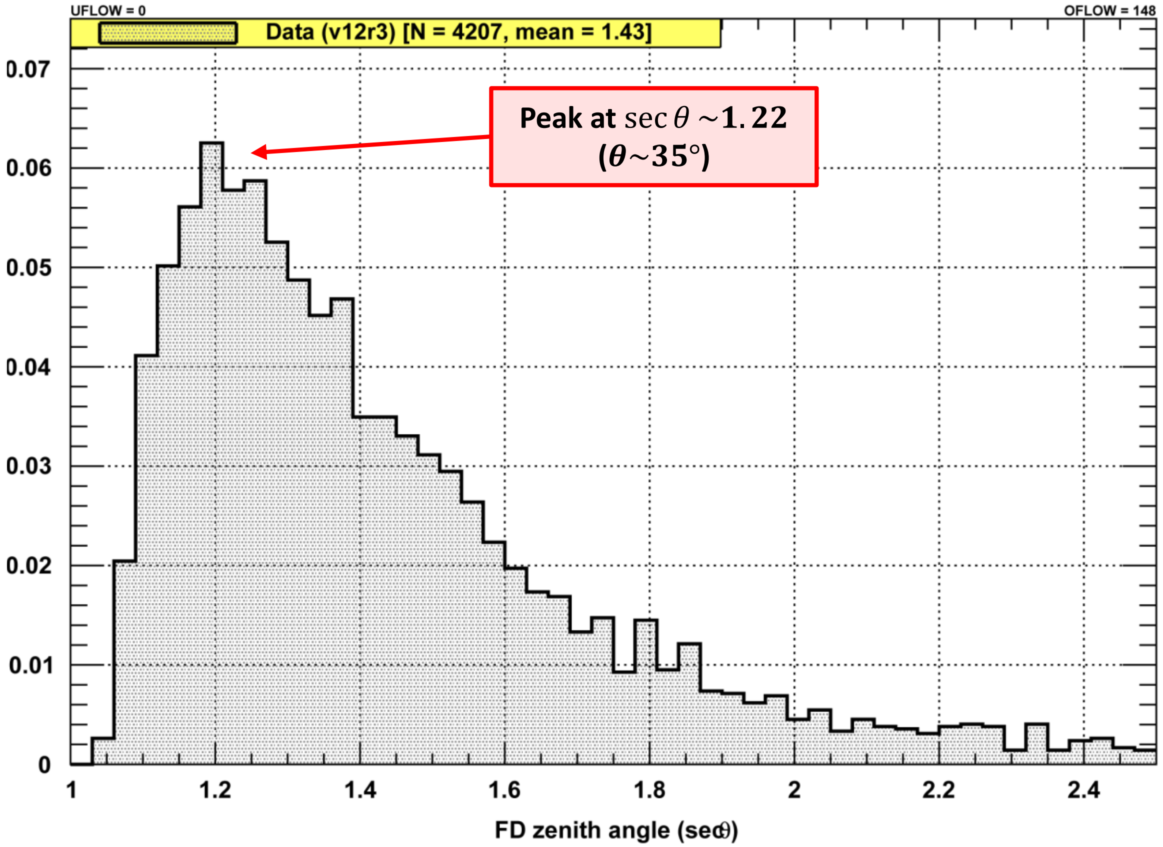




Simulations and data

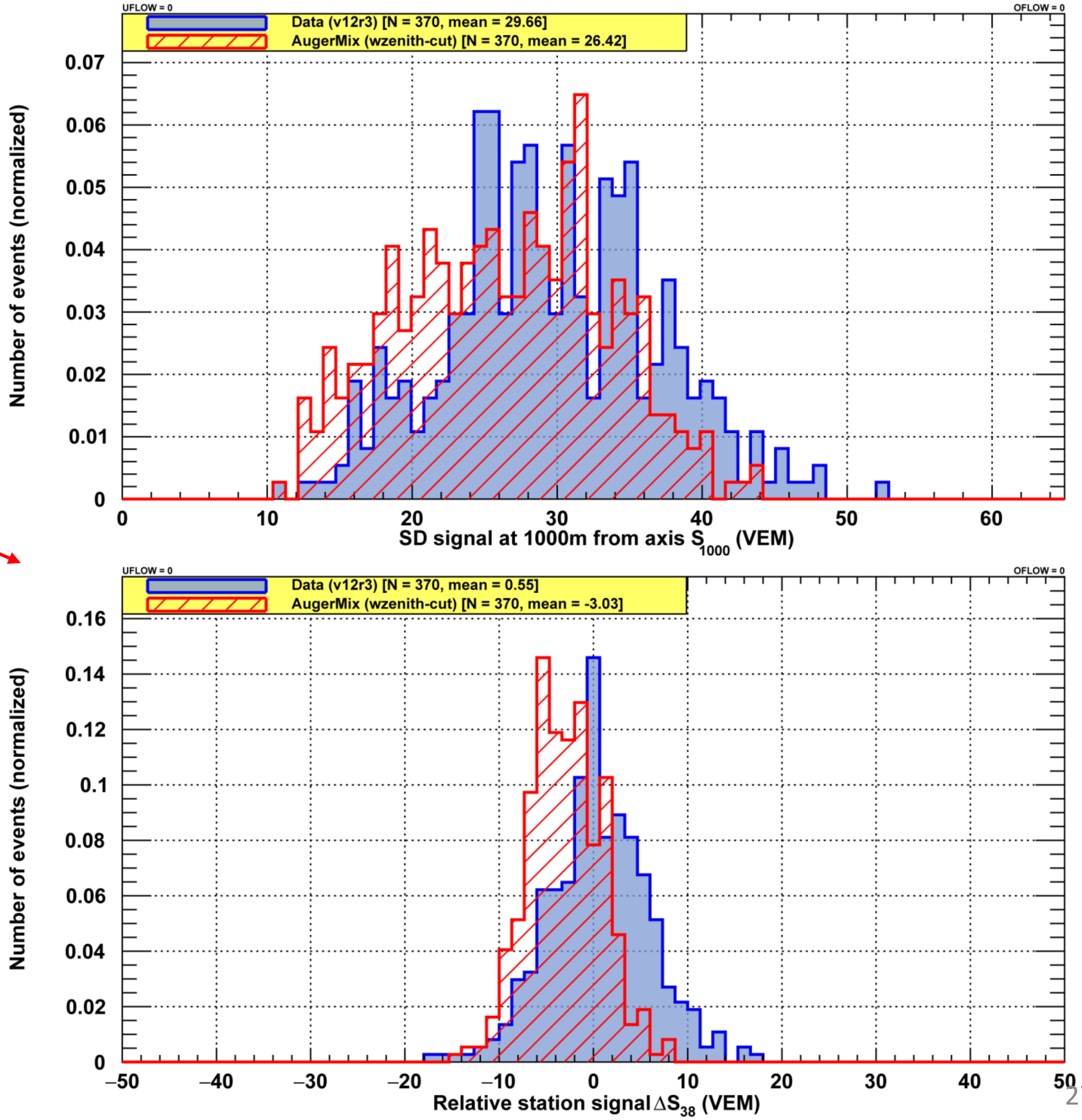
- Zenith angle ($\sec\theta$) distributions of simulations and data





Analysis observables

- SD station signal:
  - S_{1000}
  - ΔS_{38}
- Comparison between Pierre Auger data and AugerMix mock data set
- Larger S_{1000} corresponds to heavier mass composition

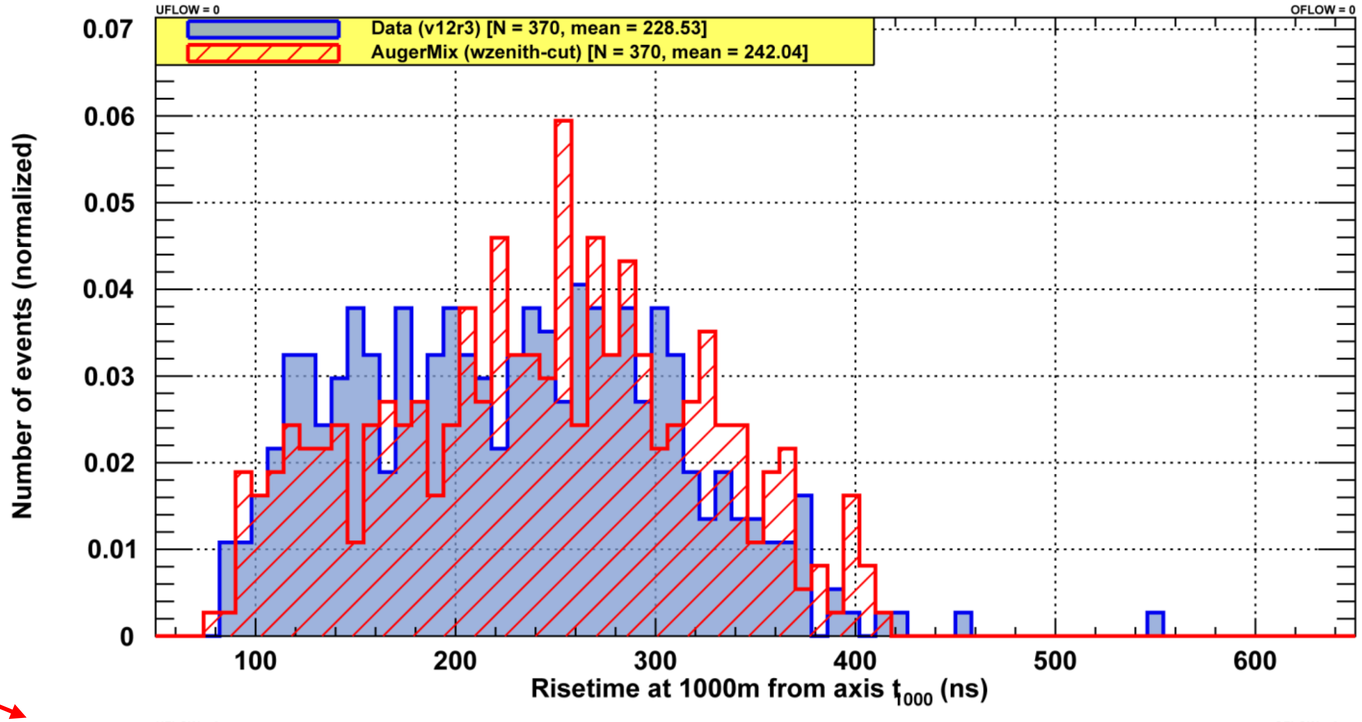


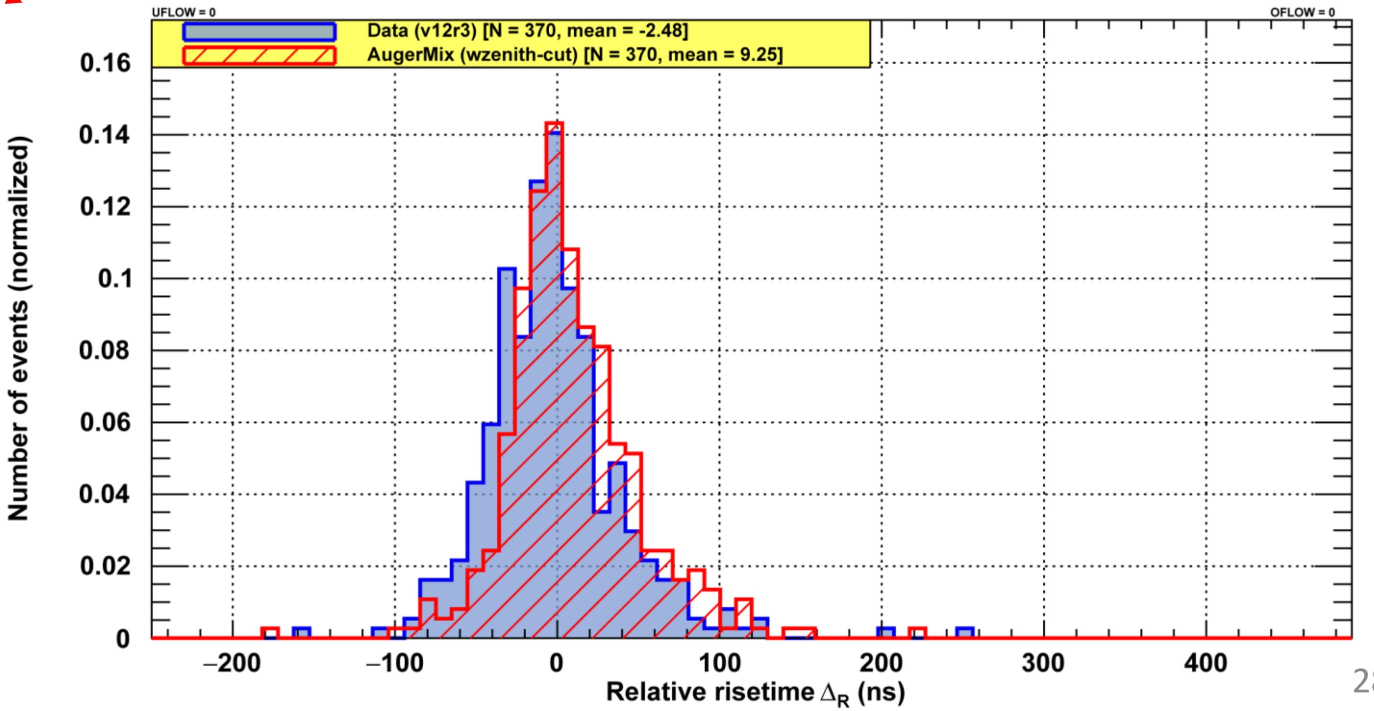
EPOS-LHC
$\log(E/eV) = [18.8, 18.9]$

Analysis observables

- SD risetime:
  - t_{1000}
  - Δ_R
- Comparison between Pierre Auger data and AugerMix mock data set
- Shorter t_{1000} corresponds to heavier mass composition

EPOS-LHC
$\log(\boldsymbol{E/eV}) = [\mathbf{18.8}, \mathbf{18.9}]$





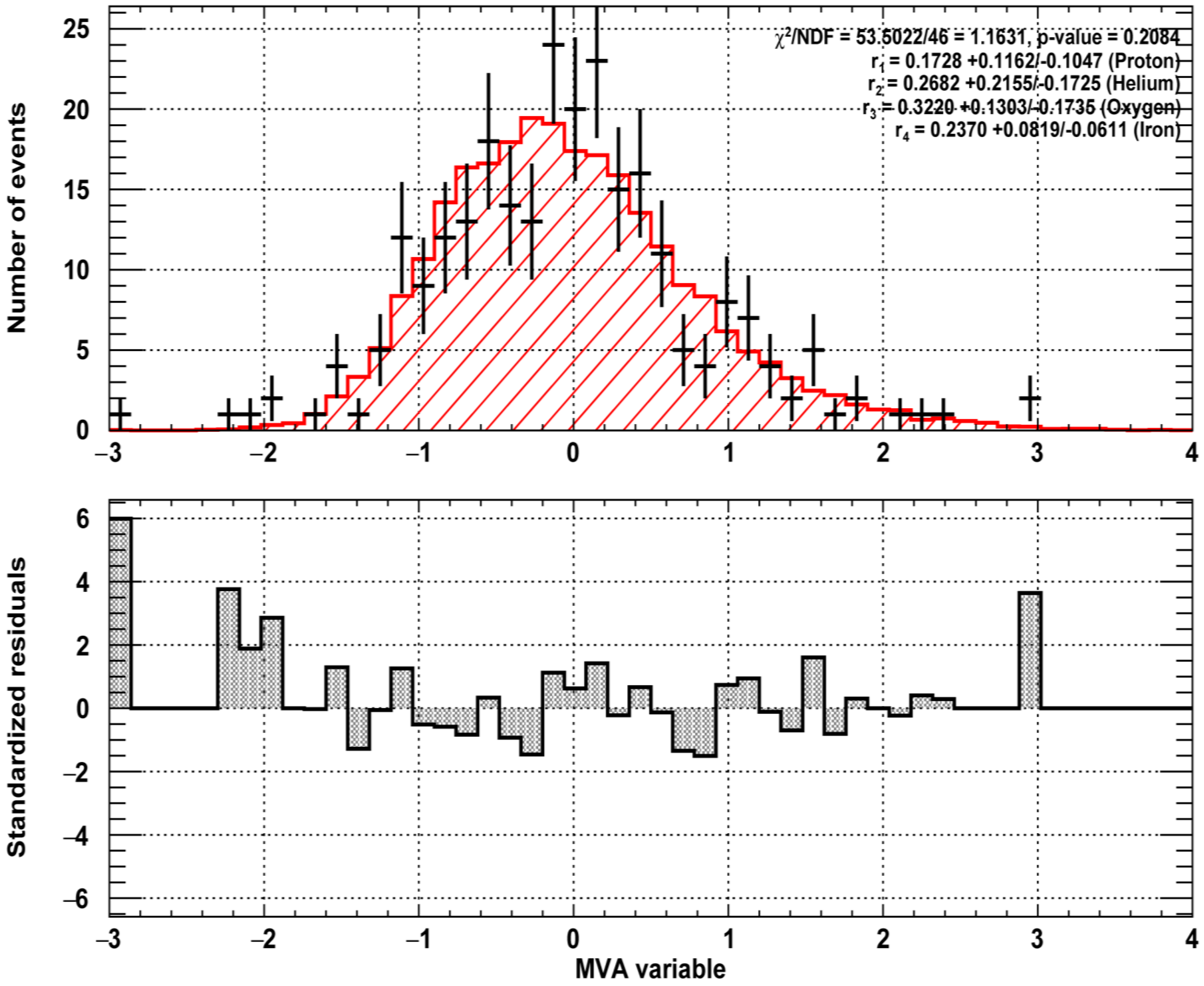
Distribution fitting procedure

- For determining elemental fractions perform distribution fitting:
  - Combine simulation MVA distributions of individual elements into H_{sim}

$$H_{sim} = \sum_{i=1}^{N} f_i \cdot H_i$$

  - Fit H_{sim} to H_{data} with a maximum likelihood fitting approach (finite distributions with Poissonian statistics)
  - Fitting parameters f_i are limited between 0 and 1
  - Standardized residuals give comparison between simulations and data

$$R_i = \frac{n_i - m_i}{\sqrt{n_i}}$$



Distribution fitting procedure

- For determining elemental fractions perform distribution fitting:
  - Combine simulation MVA distributions of individual elements into H_{sim}

$$H_{sim} = \sum_{i=1}^{N} f_i \cdot H_i$$

  - Fit H_{sim} to H_{data} with a maximum likelihood fitting approach (finite distributions with Poissonian statistics)
  - Fitting parameters f_i are limited between 0 and 1
  - Standardized residuals give comparison between simulations and data

$$R_i = \frac{n_i - m_i}{\sqrt{n_i}}$$