Event-by-event primary composition discrimination method using supervised machine learning

Washington Rodrigues de Carvalho Jr.

Faculty of Physics, Warsaw University, Poland

carvajr@gmail.com

GRAND collaboration meeting 2025, Warsaw
June 3rd, 2025

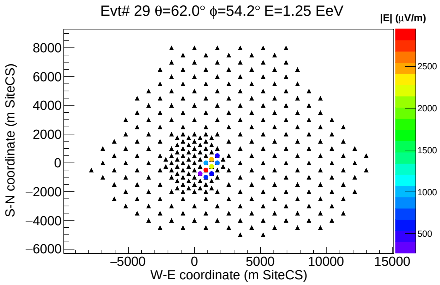
Simple Machine Learning (ML) discrimination approach

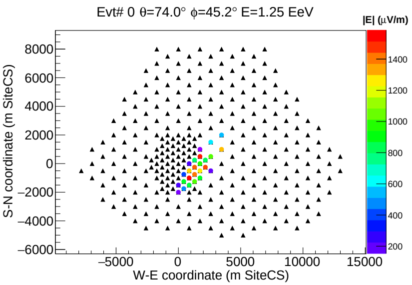
- Discriminates between heavy (Fe) and ligth (p) primary composition on an event-by-event basis
- Bypasses any X_{max} reconstruction and infers composition directly:
  - Similar to Astropart.Phys **109**, 41-49, 2019, but using ML
- Uses Random Forests (RF):
  - Simple approach.
  - Implemented my own RF code to really understand the algorithms
  - Not a black-box! Will also try to understand what is important for the discrimination
- Input data: RDSim simulations on a generic hexagonal array
  - Uses triggered antenna positions, peak amplitudes and spectral slopes
  - Also a restricted set without spectral slopes on GP300 (old layout B)
- Still preliminary!!

UNIVERSITY OF WARSAW
Faculty of Physics

RDSim

- Fast and comprehensive Monte-Carlo simulation of the radio emission and its detection.
- Takes into account the main characteristics of the detector.
  - Trigger setups, thresholds and antenna patterns
- Radio emission model based on a superposition “toymodel” that disentagles the Askaryan and Geomagnetic components





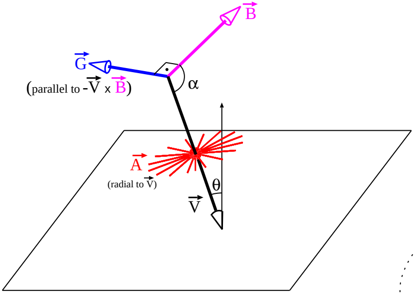
Radio emission: Superposition "toymodel"

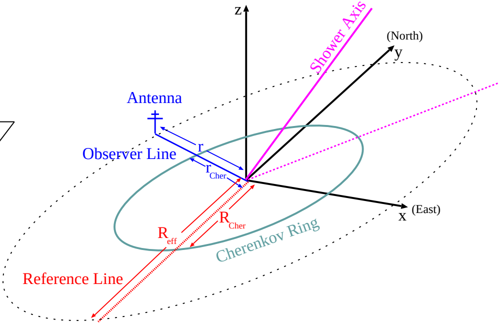
- Based on theoretical polarizations and elliptical symmetry
- Disentangles the Askaryan and geomagnetic components to estimate the electric field in any position on the ground
- Input: Full ZHAireS simulations with specific arrival directions and just a few antennas on a line
- Toymodel can now be rotated to use simulations of a fixed azimuth angle for multiple arrival directions (takes into account $\sin\alpha$, etc...)
- Early/Late effects and electric field linear scaling with energy included
- NEW: the spectral slope can now be estimated at any position
- Can sweep the phase space with much fewer input simulations

UNIVERSITY OF WARSAW
Faculty of Physics

Radio emission: Superposition "toymodel"

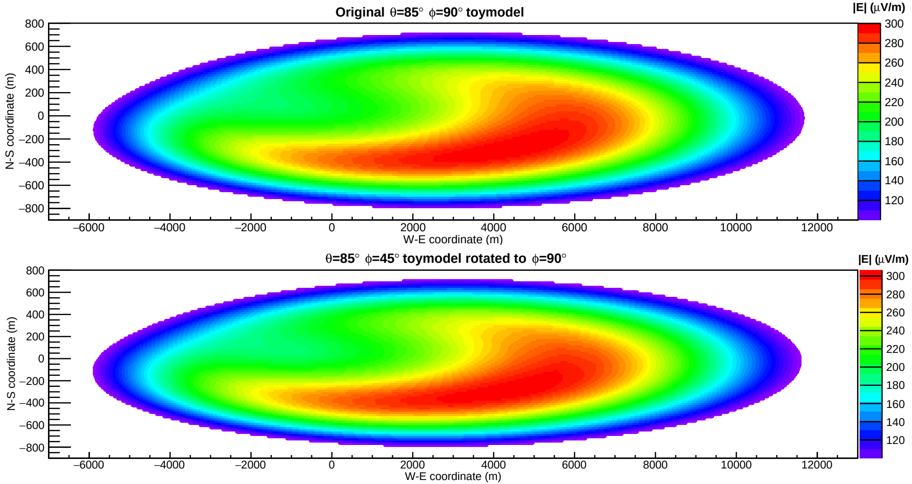
•



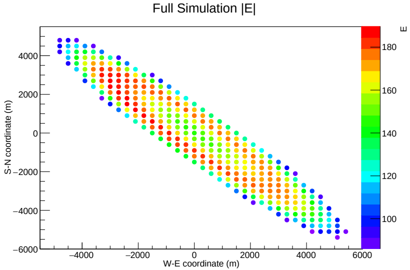


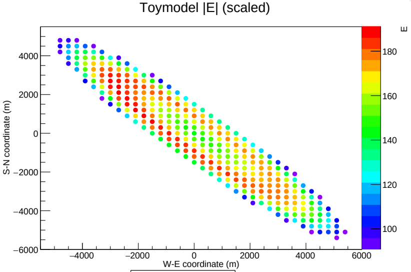
Example rotation: $\theta = 85^\circ$ from NW to W

Maximum difference between rotated toymodel and dedicated toymodel $\sim 2\%$

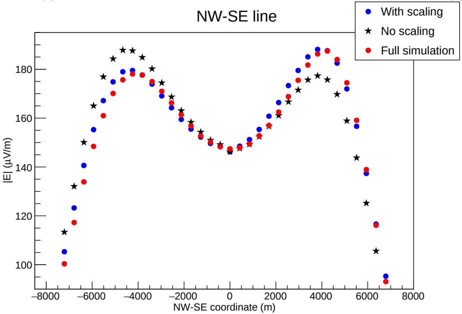


Toymodel p 1EeV 80°: $|\vec{E}|$ comparison to full simulation



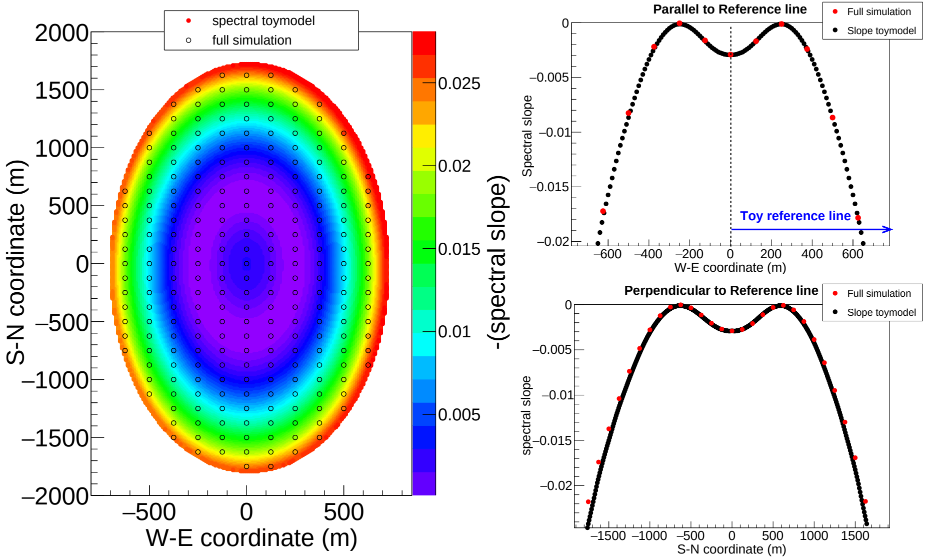


max. diff. $\sim 6\%$



UNIVERSITY OF WARSAW
Faculty of Physics

Toymodel p1.25EeV 66°: Slope comparison to full simulation

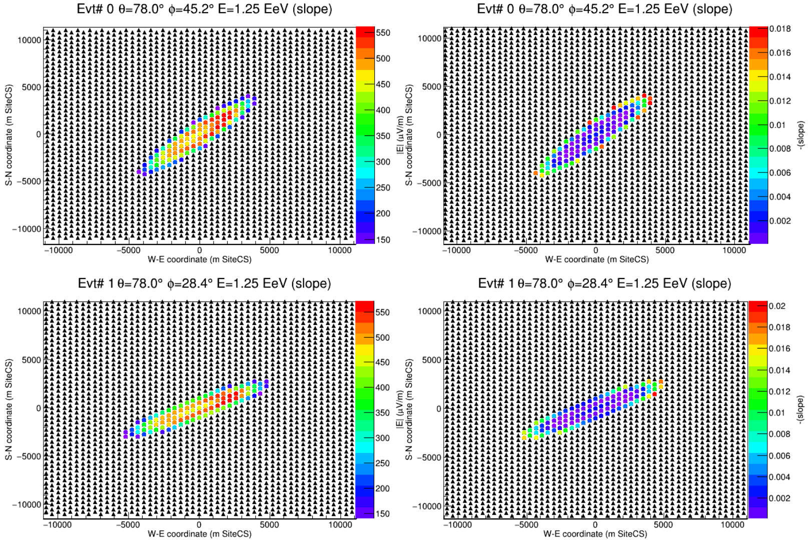


RDSim simulation parameters

- 50 p and 50 Fe input full simulations with E_0=1.25 EeV per zenith
- A total of 100 "Toymodels" were created per zenith and normalized to the exact EM energy of each fully simulated shower
  - Now every shower has the exact same EM energy
  - Erases EM energy dependence on composition
- Zeniths: 50° to 82° in steps of 4° (analyzed separately)
- Hexagonal Array with "infill" distance ("outlier" distance for 82°)
- Antenna threshold of 101 μV/m per component
- Minimum of 5 triggered antennas
- Bandwidth: 30 MHz - 80 MHz (for now)
- Horizon antenna gains not included yet (for now)
- For each zenith, simulated enough events to get ∼10k triggered events
- Created a train and a test file with ∼5k events each
- A Gaussian energy smearing of 10% was added to each event
  - Twice the quoted 5% for Felix's and Tim's E_{EM} reconstruction method
  - Mimics the energy uncertainty of a single energy bin

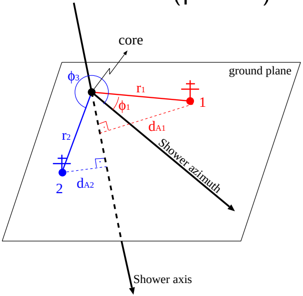
Event examples: $|\vec{E}|$ and spectral slope

$|\vec{E}|$ — Spectral slope



Features

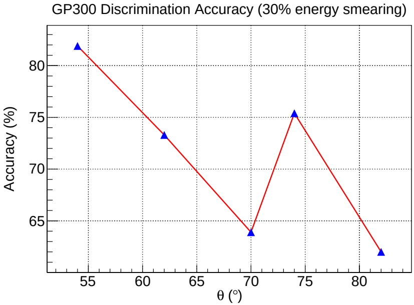
- Triggered antennas are ordered with increasing distance to the axis
- For each antenna *i* we used:
  - The distance d_{Ai} to the shower axis, the peak amplitude $|E_i|$ and the spectral slope SS_i
  - Features: $d_{A1}, |E_1|, SS_1, d_{A2}, |E_2|, SS_2, ..., d_{Ai}, |E_i|, SS_i$
  - The number of features is $3\times$ the number of antennas triggered by the event with the most antennas
  - For events with less antennas, missing features are subtituted by zeros
  - Primary composition also saved (p or Fe)



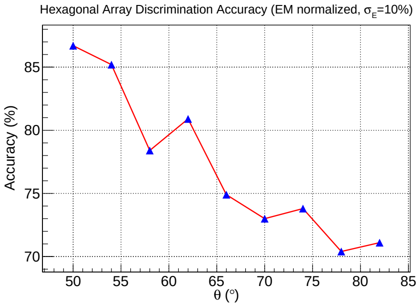
Old results using only distance and amplitude

- Very Good accuracies for such a simple method
- Accuracies tend to decrease with increasing zenith
- Analysis of the feature importances: proton showers seemed to be brighter than Fe near the core on most geometries

30-250 MHz (GP300B)

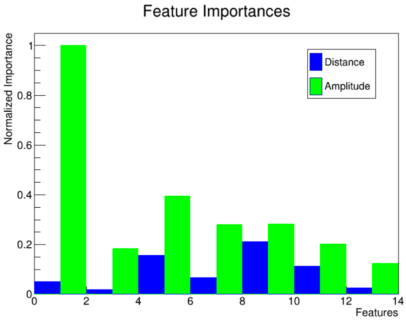


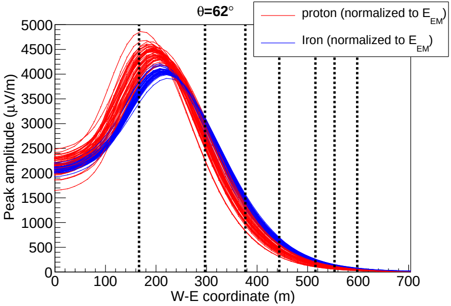
50-200 MHz (Hex array)



Old results using only distance and amplitude

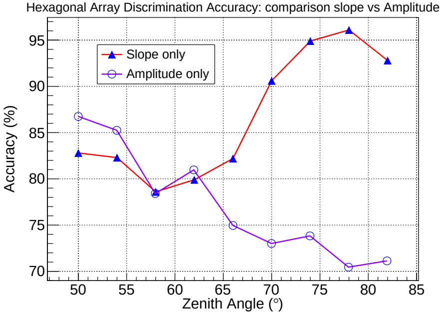
- Very Good accuracies for such a simple method
- Accuracies tend to decrease with increasing zenith
- Analysis of the feature importances: proton showers seemed to be brighter than Fe near the core on most geometries

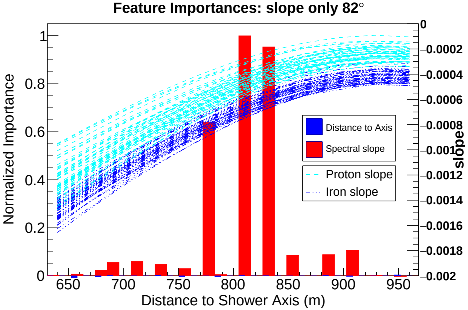




New results using only distance and spectral slope

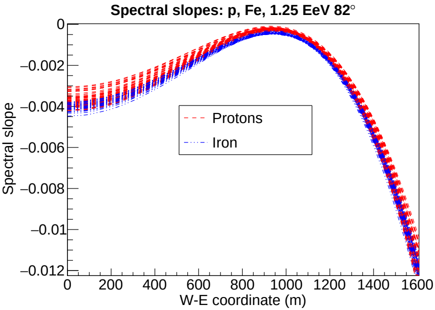
- The effect of the energy uncertainty in the slope is negligible
- Almost perfect discrimination at high zeniths!
- Accuracies tend to decrease with decreasing zenith
- Analysis of the feature importances: Most important features tend to be in regions where there is a smaller overlap between p and Fe

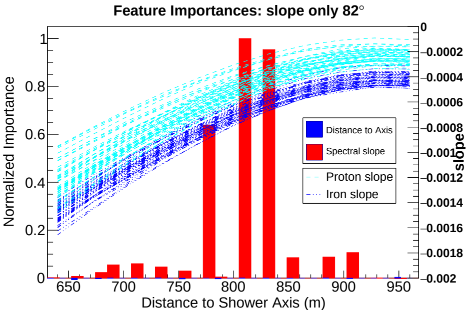




New results using only distance and spectral slope

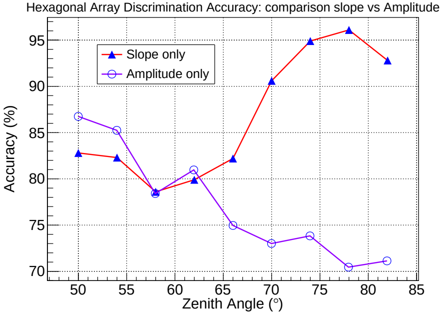
- The effect of the energy uncertainty in the slope is negligible
- Almost perfect discrimination at high zeniths!
- Accuracies tend to decrease with decreasing zenith
- Analysis of the feature importances: Most important features tend to be in regions where there is a smaller overlap between p and Fe

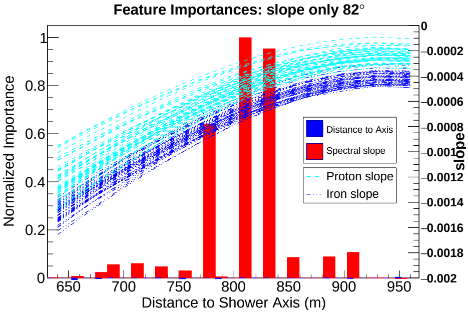




New results using only distance and spectral slope

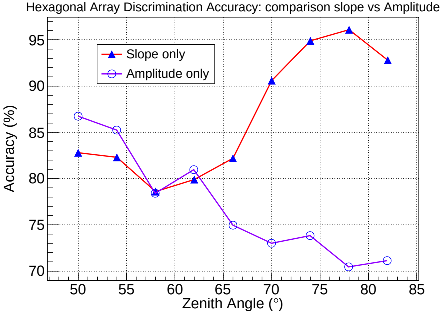
- The effect of the energy uncertainty in the slope is negligible
- Almost perfect discrimination at high zeniths!
- Accuracies tend to decrease with decreasing zenith
- Analysis of the feature importances: Most important features tend to be in regions where there is a smaller overlap between p and Fe

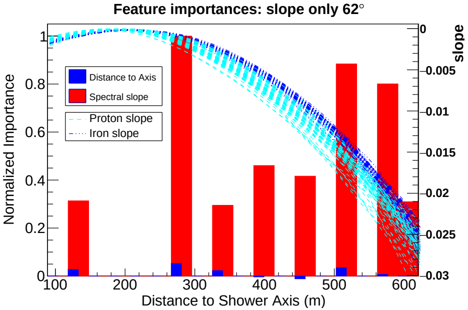




New results using only distance and spectral slope

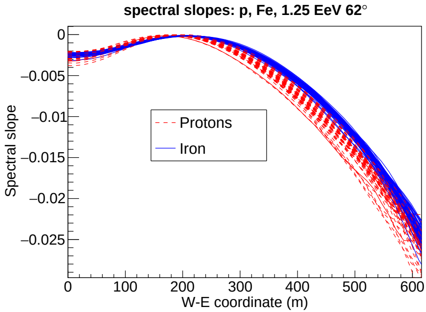
- The effect of the energy uncertainty in the slope is negligible
- Almost perfect discrimination at high zeniths!
- Accuracies tend to decrease with decreasing zenith
- Analysis of the feature importances: Most important features tend to be in regions where there is a smaller overlap between p and Fe

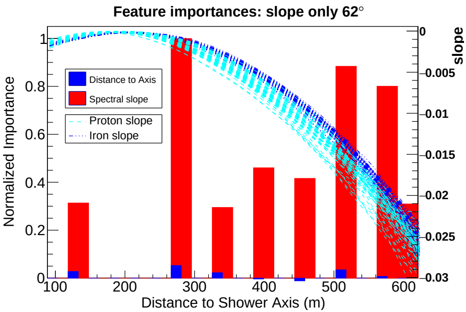




New results using only distance and spectral slope

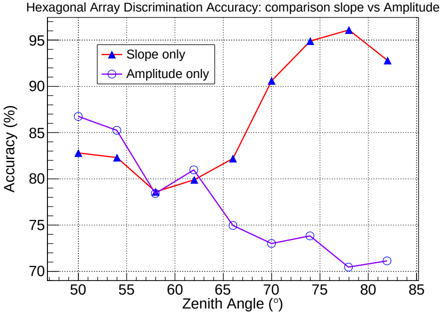
- The effect of the energy uncertainty in the slope is negligible
- Almost perfect discrimination at high zeniths!
- Accuracies tend to decrease with decreasing zenith
- Analysis of the feature importances: Most important features tend to be in regions where there is a smaller overlap between p and Fe

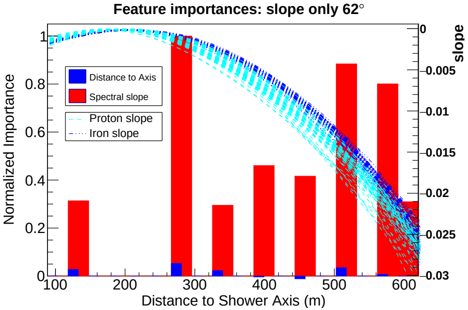




New results using only distance and spectral slope

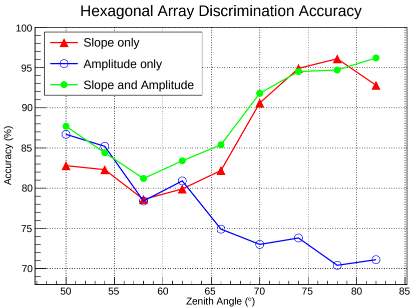
- The effect of the energy uncertainty in the slope is negligible
- Almost perfect discrimination at high zeniths!
- Accuracies tend to decrease with decreasing zenith
- Analysis of the feature importances: Most important features tend to be in regions where there is a smaller overlap between p and Fe

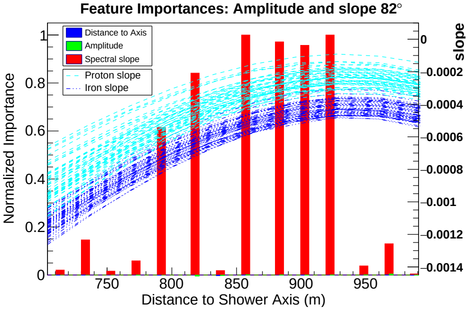




New results using distance, amplitude and spectral slope

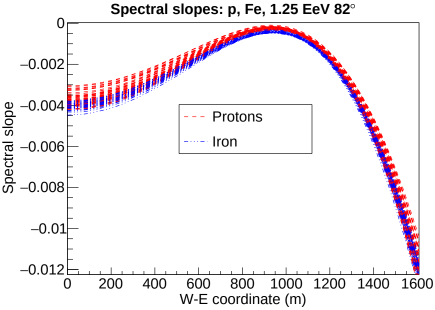
- We get the best of both worlds in all zenith regions!
- Accuracies only decrease to $\sim 81\%$ around 60°
- Most important features tend to be:
  - High zenith: In regions where the slope overlap is smaller
  - Low zenith: In regions where the amplitude overlap is smaller

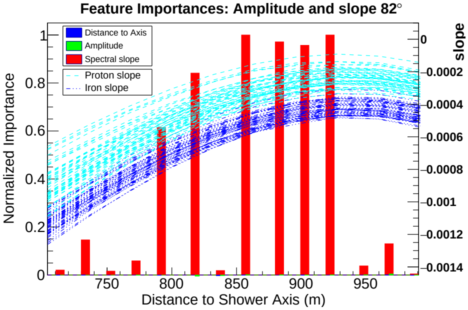




New results using distance, amplitude and spectral slope

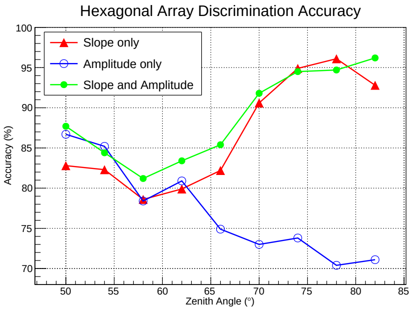
- We get the best of both worlds in all zenith regions!
- Accuracies only decrease to $\sim 81\%$ around 60°
- Most important features tend to be:
  - High zenith: In regions where the slope overlap is smaller
  - Low zenith: In regions where the amplitude overlap is smaller

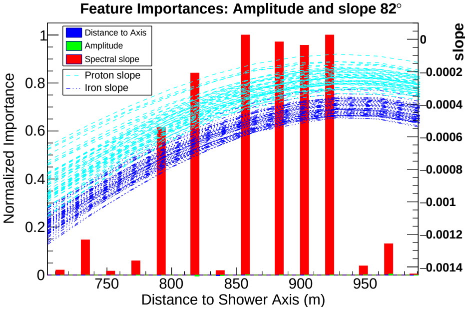




New results using distance, amplitude and spectral slope

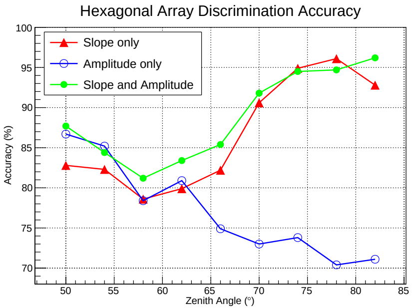
- We get the best of both worlds in all zenith regions!
- Accuracies only decrease to $\sim 81\%$ around 60°
- Most important features tend to be:
  - High zenith: In regions where the slope overlap is smaller
  - Low zenith: In regions where the amplitude overlap is smaller

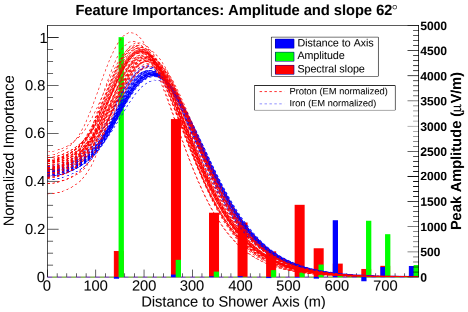




New results using distance, amplitude and spectral slope

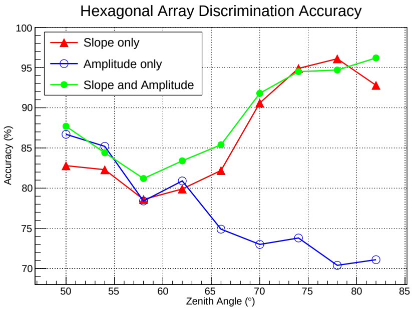
- We get the best of both worlds in all zenith regions!
- Accuracies only decrease to $\sim 81\%$ around 60°
- Most important features tend to be:
  - High zenith: In regions where the slope overlap is smaller
  - Low zenith: In regions where the amplitude overlap is smaller

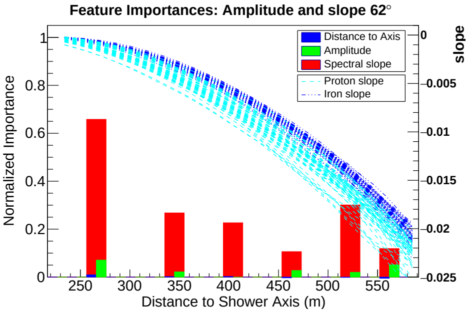




New results using distance, amplitude and spectral slope

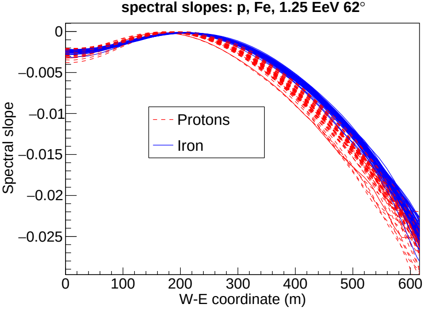
- We get the best of both worlds in all zenith regions!
- Accuracies only decrease to $\sim 81\%$ around 60°
- Most important features tend to be:
  - High zenith: In regions where the slope overlap is smaller
  - Low zenith: In regions where the amplitude overlap is smaller

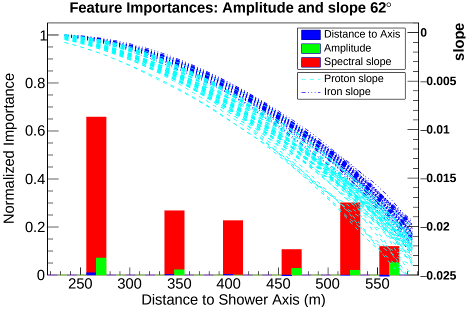




New results using distance, amplitude and spectral slope

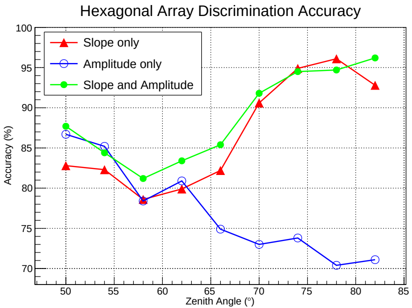
- We get the best of both worlds in all zenith regions!
- Accuracies only decrease to $\sim 81\%$ around 60°
- Most important features tend to be:
  - High zenith: In regions where the slope overlap is smaller
  - Low zenith: In regions where the amplitude overlap is smaller

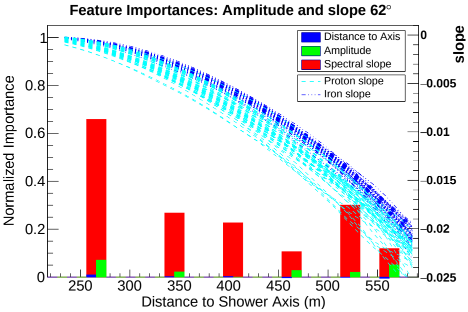




New results using distance, amplitude and spectral slope

- We get the best of both worlds in all zenith regions!
- Accuracies only decrease to $\sim 81\%$ around 60°
- Most important features tend to be:
  - High zenith: In regions where the slope overlap is smaller
  - Low zenith: In regions where the amplitude overlap is smaller





Too good to be true? Caveats: The devil's advocate

- Amazing accuracies: between 81 and 96%! But...
- Noise not included yet!
  - Slopes should be sensitive to noise
  - Could in principle degrade the slope discrimination stregth
- Quoted accuracies are for **MY** sample
  - Simulated 10K events per zenith, but based on only 100 "Toymodels"
    - No full shower-to-shower fluctuations (10k events but only 100 $\neq X_{max}$)
  - Accuracies could vary for different sets, depending on X_{max} overlaps
  - Sensitive to hadronic model used: different X_{max} distros and overlaps
- Real showers: How well do the simulations resemble **REAL** showers?
- Huge and dense array (Infill distance) means many triggered antennas
  - What's the impact of using smaller, less dense arrays?
- Used 30-80 MHz only. Using 50-200 MHz can lead to thinning artifacts on the slopes at low zeniths
  - Can be corrected by lowering thinning on simulations
  - Or "analytically" using a "Cut&Fit" method (backup slides)

UNIVERSITY OF WARSAW
Faculty of Physics

Conclusions

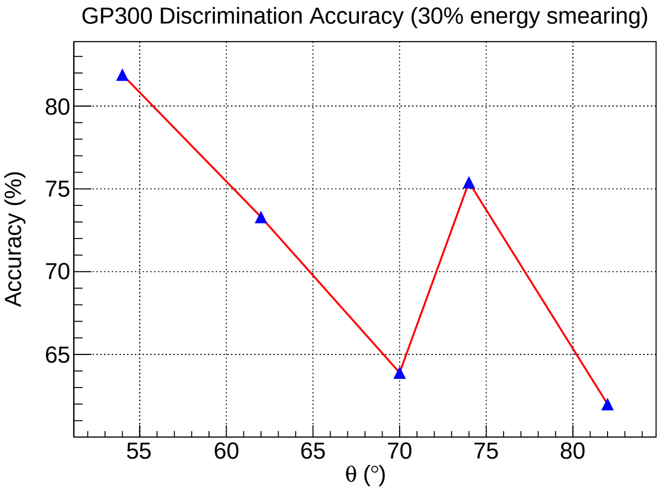
- The spectral slope LDF, just as the amplitude LDF, has a strong correlation with X_{max} and thus also primary composition
- This slope dependence on X_{max} could have the same physical origins as the amplitude dependence on X_{max}
  - Especially the loss of coherence relating to lower densities during shower development. Very clear at high zeniths
  - More study needed to fully understand the origins of this dependence
- Using spectral slopes as RF features significantly increases discrimination accuracies, especially at high zenith angles
- Very promising results
  - Using both the amplitudes and slopes leads to incredibly high discrimination accuracies of 81-96%! Even without RF optimization
- The impact of other factors, such as noise and hadronic model, still need to be addressed
  - But we are starting with such high accuracies, that I find very unprobable that including more effects will destroy the method

Questions?

Other applications of Radio...

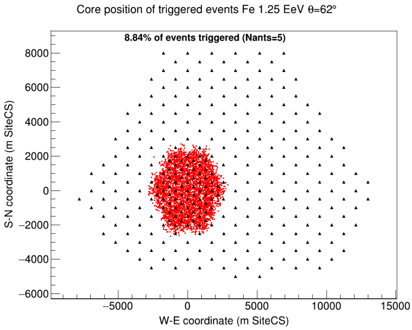
BACKUP

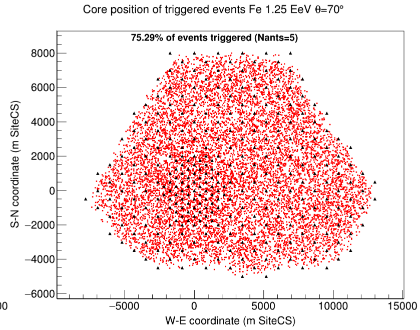
Minimum accuracy around 70°: GP300 change of regime



Minimum accuracy around 70°: GP300 change of regime

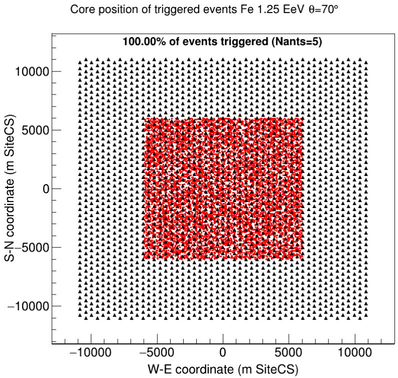
- 62°: Only triggers inside Infill
- 70°: Trigger over the whole array
    - "Effective" antenna distance d increases significantly ($d_{infill} \rightarrow d_{outliers}$)
    - Footprint not properly sampled at 70° (footprint too small)
    - Larger zeniths are better sampled, leading to an increase in accuracy

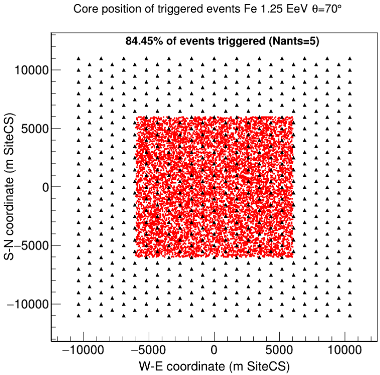




"Fake" array tests at 70°

- Infill spacing: Accuracy $\geq 69.7\%$
- GP300: Accuracy $\geq 61.3\%$
- Outlier spacing: Accuracy $\geq 59.9\%$

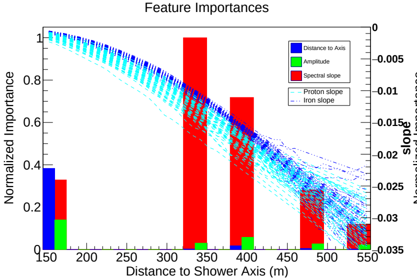


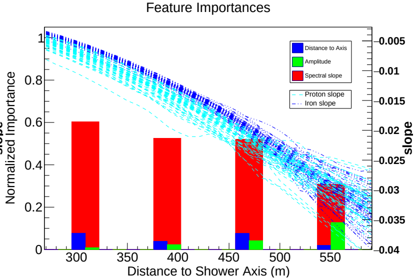


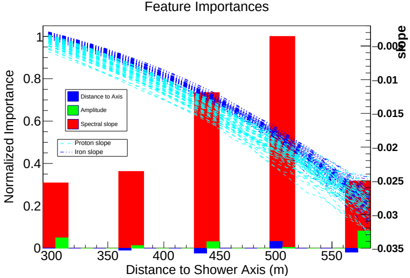
Random Forest parameters

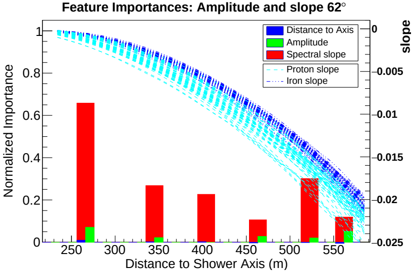
- $N_{trees} = 200$: Number of threes in the forest
- $D_{max} = 100$: Maximum Tree Depth
- $S_{min} = 10$: Minimum number of samples is a node (tested range 5-12)
- $boot_{size}$: Ratio between the number of events in the boostrap and the full train dataset (saves time)
- N_{Fsub}: Number of features in the random feature subset (N_{add})
- $\sigma_E = 0.1$: RMS of Gaussian energy smearing (tested 10-40% range)
- N_{remove}: Number of farthest antennas removed from the features

Feature importances and SLOPE LDF: 50 to 62°

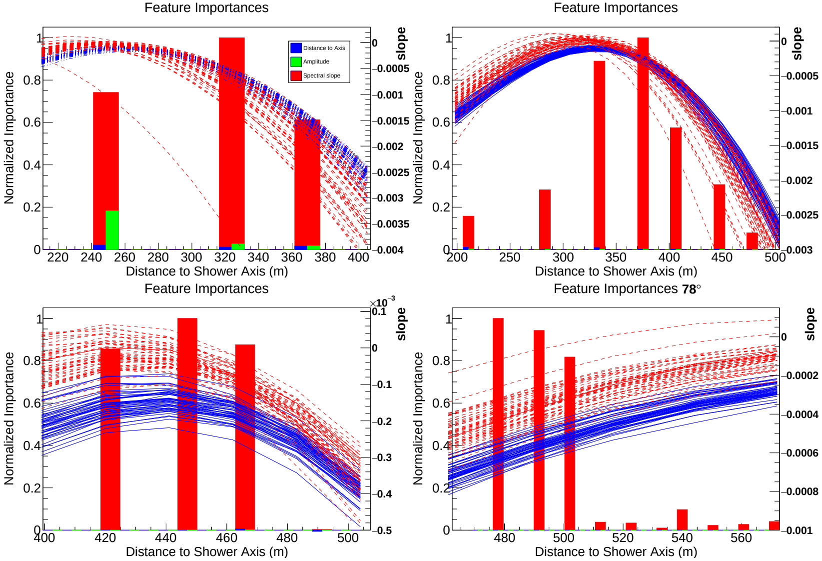




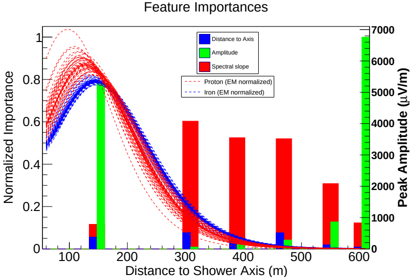


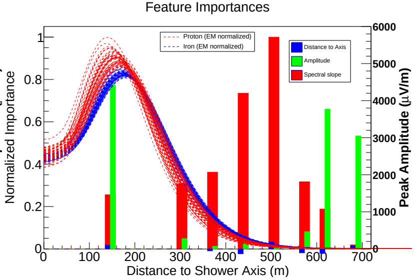


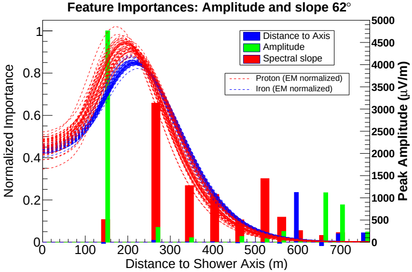
Feature importances and SLOPE LDF: 66 to 78°



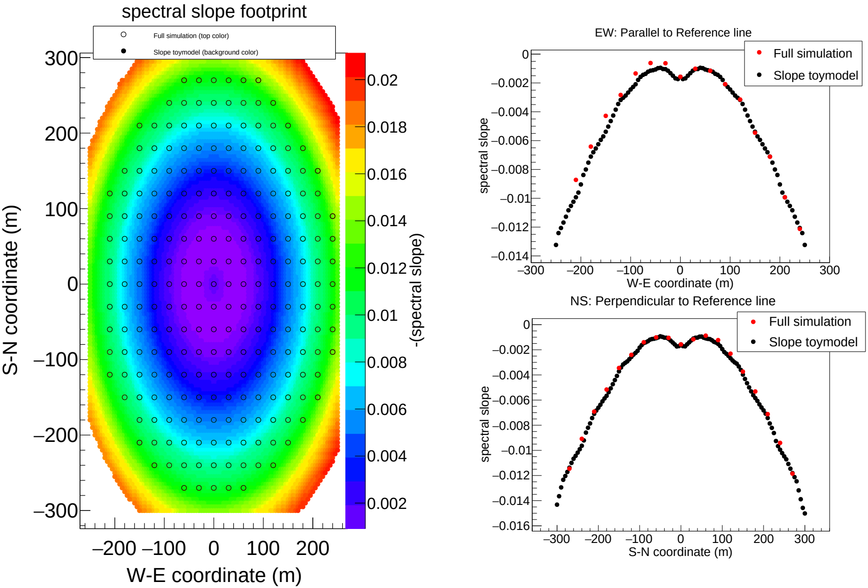
Feature importances and amplitude LDF: 50 to 62°



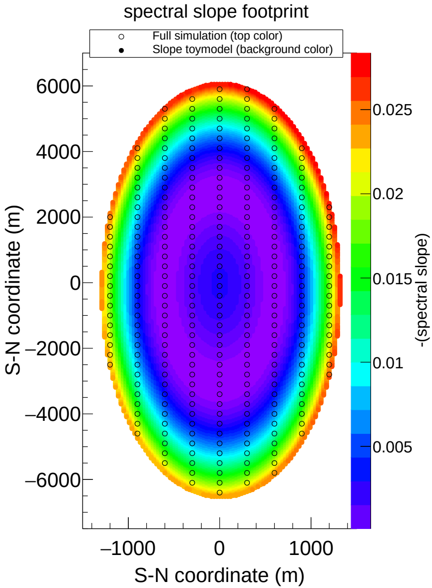


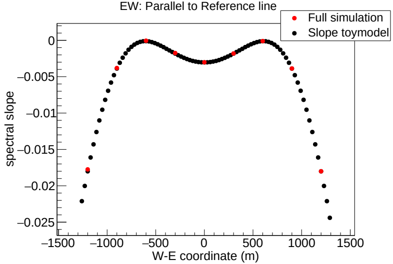


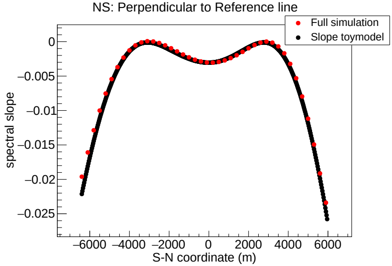
Toymodel p1.25EeV 30°: Slope comparison to full simulation



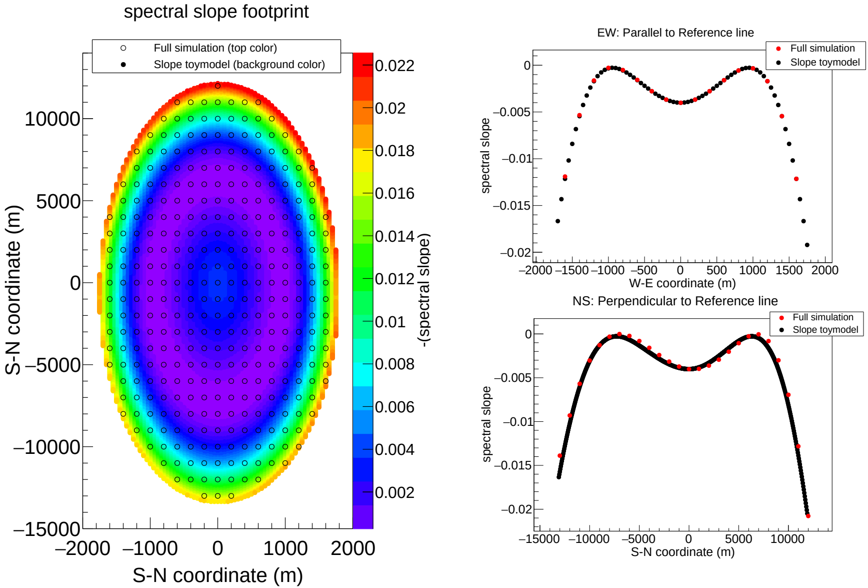
Toymodel p1.25EeV 78°: Slope comparison to full simulation



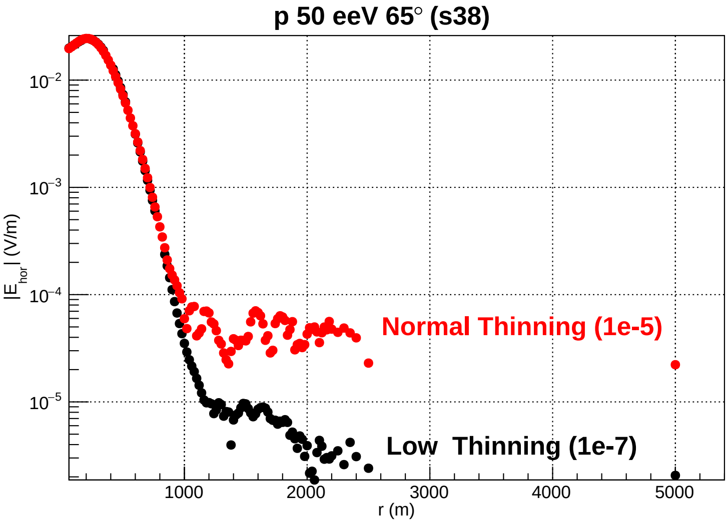




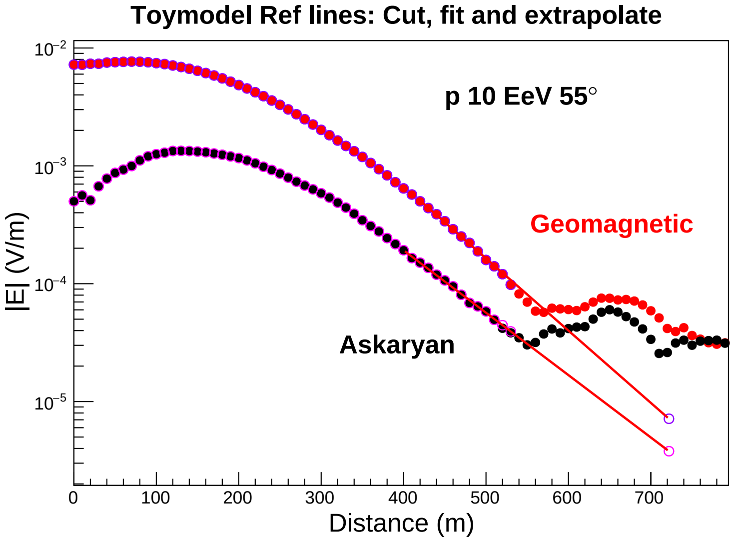
Toymodel p1.25EeV 82°: Slope comparison to full simulation



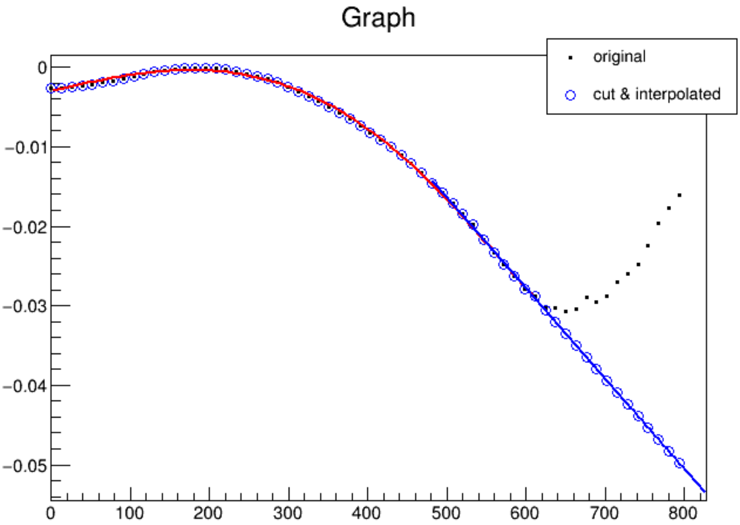
Thinnning artifacts: amplitude (Very relevant for deep ν's!)



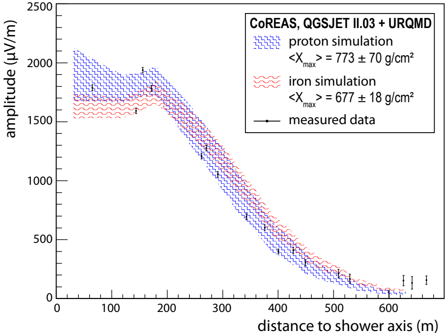
Fixing thinnning artifacts: amplitude Cut&Fit

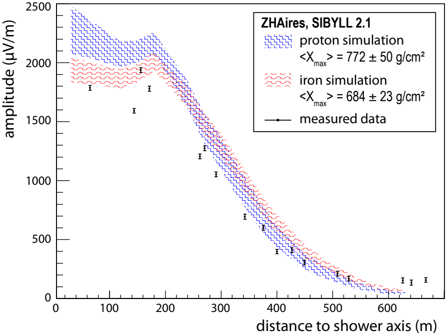


Fixing thinnning artifacts: slope Cut&Fit



Hadronic model dependence?





Tim Huegue, arXiv:1310.6927, Braz. J. Phys., 44, 5, 520-529, (2014)

Spectrum fit example

- Several fit options: But mostly used pol2 fit in Log
- Slope obtained from the linear parameter