PROCEEDINGS OF SCIENCE

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

Washington R. de Carvalho Jr.[a,*] **and Lech Piotrowski**[a,]

[a] *Faculty of Physics, Warsaw University,*
ul. Pasteura 5, Warsaw, Poland

E-mail: carvajr@gmail.com

We have developed a radio detection mass discrimination method for cosmic ray events. This method uses supervised machine learning (ML) algorithms, namely random forests (RF), to discriminate between light (p) and heavy (Fe) primary compositions on an event-by-event basis. It bypasses any shower maximum (X_{max}) reconstructions and instead tries to infer the primary composition directly. As features of the random forest we used, for each triggered antenna, the distance to the shower axis, the peak amplitude of the electric field and the spectral slope. To perform the discrimination, the method also needs an estimate of the primary or electromagnetic (EM) energy of the shower along with its uncertainty, which is also taken into account. Initially we used a 2-feature approach, only with the antenna distance and the peak electric field amplitude. Yet, we obtained much better than expected accuracies in these test runs, especially at low zeniths. Even with the restrictive feature set and using a huge primary energy uncertainty of 30%, we obtained an accuracy of 82% at 54°. An analysis of the random forest feature importances uncovered that such good accuracies were possible because the RF was using a large electric field amplitude dependence on the position of X_{max} to perform the discrimination. We describe this amplitude dependence and explain it in detail in our other contribution to this conference. After adding the spectral slope as a third feature, we observed a significant improvement in the discrimination accuracy, which now varied from 81% to 96%, depending on zenith angle (θ). This novel approach may offer particular benefits to radio-only setups like GRAND. This work is Monte Carlo based and uses ZHAireS simulations along with RDSim to generate separate sets of events for training and testing the random forest algorithm.

39th International Cosmic Ray Conference (ICRC2025)
15–24 July 2025
Geneva, Switzerland

ICRC 2025
The Astroparticle Physics Conference
Geneva July 15-24, 2025

[*] Speaker

1. Introduction

We have developed a simple machine learning (ML) approach that uses classification random forests (RF) to discriminate, on an event-by-event basis, the mass composition of cosmic ray (CR) events into two classes: A heavy class, with events that resemble iron induced showers, and a light class, with events that resemble proton showers instead. This method is non-traditional, in the sense that it does not try to infer the composition from X_{max} reconstructions. Instead, it tries to classify the mass composition of each event directly. In that sense, it is similar to a previous method we have developed [1, 2] that used, instead of ML, a χ^2 analysis that compares an event to multiple simulations with different compositions. This χ^2 approach is akin to the one used by LOFAR-like X_{max} reconstruction methods, but instead of reconstructing X_{max} , it also infers the composition directly. One should note that although neither of the two methods perform the traditional mid-step of reconstructing X_{max} , the position of the shower maximum, not the primary composition itself, is responsible for most of the characteristics of the electric field at ground level that make our discrimination possible.

An interesting characteristic of supervised ML algorithms, such as RF algorithms, is that they are not black-boxes. An analysis of the feature importances can be used to understand what characteristics of the data the RF is using to do the discrimination. By performing such an analysis, we were able to establish that the RF was using a very large dependence of the e-field amplitudes on X_{max} (and thus also composition) to perform the mass discrimination. This dependence is much stronger than the inherit average differences in the electromagnetic (EM) energy of events with the same primary energy but different compositions. We explore and explain this dependence, in a semi-quantitative way, in our other contribution to this conference [7].

ML methods require very large data sets for training and testing. The data sets used in this work were created with RDSim [3–5], an extremely fast yet comprehensive Monte-Carlo simulation of the radio emission and its detection. In Section 2, we present a brief description of RDSim along with the superposition model it uses to estimate the radio emission. The latter was recently expanded, in order to estimate the spectral slope at any antenna position. In Section 3, we describe the simulation parameters used in this work, as well as the features and parameters used by the RF. In Section 4, we describe the RF results and also explore the feature importances to better understand how such a simple RF method is able to attain such good accuracies. Finally, in Section 5 we present a short discussion, followed by the conclusions.

2. RDSim

RDSim [3–5] is a fast, flexible and accurate package for the simulation of the radio emission of downgoing EAS and its detection by an arbitrary array. It is based on simple toymodel-like approaches, such as a superposition emission model that disentangles the Askaryan and geomagnetic components of the radio emission. This emission model uses full ZHAireS simulations [6] as input and is an extension of [8]. It can estimate the peak electric field, the polarization and the spectral slope at any position at ground level. One of its biggest advantages is that it can use a single input full simulation to generate multiple events with different arrival directions and energies by scaling the energy of the simulation and rotating the shower in azimuth, taking all relevant effects into account. This means that a huge number of events, with different geometries and energies, can be generated using just a few ZHAIRES input simulations. Due to the large statistics made possible by its speed, ability to reuse input full simulations and capacity to estimate the characteristics of the field at any position, it is especially suited to create the large data sets needed to both train and test ML algorithms, such as the RF used in this work.

For each input full ZHAireS simulation, with just a few antennas along a chosen reference line, the superposition model disentangles the Askaryan and geomagnetic components of the emission and obtains the amplitudes of the peak electric field, for each mechanism separately, as a function of the distance to the core along said reference line. From the shower geometry and a model of the atmosphere (to calculate the Cherenkov angle θ_C), it can obtain the ellipse that represents the Cherenkov ring at ground level, where the radio signal is expected to be highest. Note that any point in this ellipse observes X_{max} at the same angle w.r.t. the shower axis. This is not only true for the Cherenkov ellipse, but also for any similar ellipse that has the same ratio between the major and minor axes. This means that the amplitudes of the Askaryan and Geomagnetic emission mechanisms are approximately the same along any ellipse with the same shape as the Cherenkov ellipse (see also Early-Late corrections below). The model then uses this elliptical symmetry to estimate the net peak electric field and its polarization at any position on the ground: Given an arbitrary observer at a distance r from the center of the ellipse (blue antenna on the right panel of Fig. 1), we use the elliptical symmetry to get the distance R_{eff} along the reference line (red line on the right panel of Fig. 1), where we sample the Askaryan and geomagnetic amplitudes. We then add them up vectorially, using their theoretical polarizations (left panel of Fig. 1), to obtain the net electric field and polarization at the desired observer position (see [8] for more details). The model uses a similar procedure to obtain the spectral slope at any arbitrary position. From the input simulation it obtains the spectral slopes along the reference line and then uses the same elliptical symmetry to estimate the slope for any observer. We also take into account Early-Late effects that arise due to changes in the distance to the emission point (X_{max}) as the position of the observer changes. This is modeled by scaling the amplitudes sampled at the relevant point R_{eff} on the reference line by $D_{\mathrm{eff}}/D_{\mathrm{obs}}$, where D_{eff} is the distance between X_{max} and the point R_{eff} on the reference line and D_{obs} the distance rom X_{max} to the point r on the observer line.

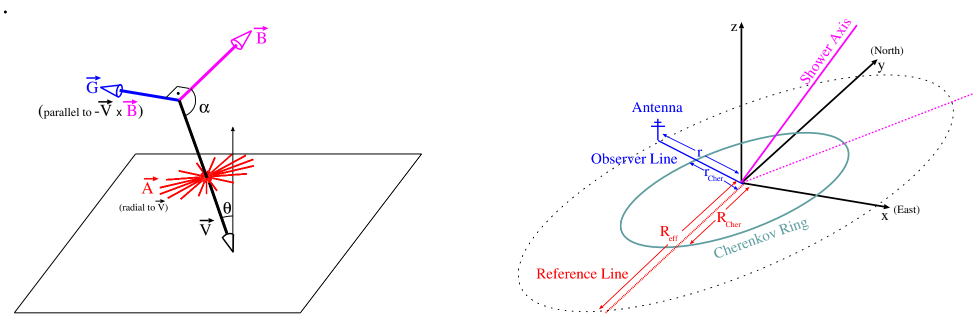

Figure 1: Left: Theoretical polarization of the Askaryan (red) and geomagnetic (blue) emission mechanisms. $\vec{V}$ is the shower axis and $\vec{B}$ the geomagnetic field. Right: Elliptical symmetry at ground level. The reference line is shown in red and the antenna where we want to estimate the electric field is shown in blue, on the observer line. Both figures were extracted from [8].

We have extensively compared the results obtained from this emission model with full simulations and found that even at high zenith angles, where the differences are expected to be greatest, there is a very good agreement. On Fig. 2, we show examples of such comparisons for the peak electric field (left panel) and the spectral slope (middle and right panels).

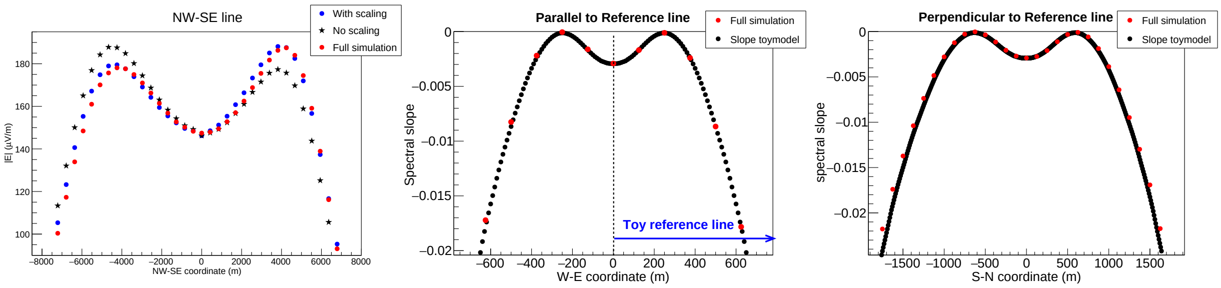

Figure 2: Comparison between full simulations and the results of the superposition model. Left: Peak amplitudes for an 80° shower along the major axis of the elliptical radio footprint, extracted from [4]. One can see that including the Early-Late corrections (“With Scaling”) creates a very good agreement. The maximum difference in this example comparison is 6%. Middle: Spectral slopes of a 66° shower along the EW direction. Right: Same as middle, but along the NS direction.

RDSim takes into account, in a very simple way, the main characteristics of the detector. The antennas, in any arbitrary layout, are assumed to be on a flat plane at ground altitude. The antenna-level trigger is modeled using any arbitrary electric field threshold (net field or per component). Any arbitrary beam pattern for the antennas can also be taken into account by just multiplying the original electric field obtained from the emission model by the beam pattern for the arrival direction w.r.t. X_{max} . Array-level triggers are modeled by requiring an arbitrary minimum number of triggered antennas to consider the whole event as triggered. At the end of the run, all information is saved to a compressed ROOT file, making it possible to implement more complex analyses and triggers outside the main simulation. We have also compared this simple detection model with full simulations of both the shower and the AUGER-RD detector response and have also found that there is a very good agreement, as can be seen on Fig. 3 of [5].

3. Simulation and RF parameters and features

For this work we used full ZHAireS simulations of showers at the GRAND site with energy $E_0 = 1.25$ EeV, a geomagnetic field $|\vec{B}| = 56.4\mu T$ and zeniths from 50° to 82°, in steps of 4°, using Sibyll as the hadronic model. For each zenith we simulated 50 p and 50 Fe showers. These simulations, after applying a band-pass filter between 30 and 80 MHz[1], were then used to create 100 different instances of the superposition emission model per zenith angle. In order to erase the intrinsic difference between the average EM energy of protons and iron, we normalized all toymodels to the exact EM energy of each simulated shower. So every shower now has the same EM energy instead of the same primary energy. As the detector, we used an hexagonal antenna array with spacing of 500 m (1000 m for 82° showers). The trigger threshold was set to 101 $\mu V/m$ per component of the electric field (EW, NS or vertical). No antenna gains were included and no noise was generated for the events.

[1] Initially, when not using the spectral slopes, we used a filter between 50 and 200 MHz.

We then generated 10k RDSim events per zenith angle, with equally distributed core positions in area and equally distributed azimuths. We then added to each event a random 10% Gaussian energy smearing, which is twice the quoted $\sim 5\%$ uncertainty of modern radio EM energy reconstructions (e.g., [9]). This energy smearing, along with the EM energy normalization applied previously, makes our event set mimic the events in an energy bin reconstructed by such EM reconstruction methods, albeit with twice the uncertainty. We then divided each 10k set into two sets of 5k events each, one for training the RF and the other for testing the discrimination accuracy.

For this work we have developed our own discrimination RF code, tailored to our specific application. It is capable of using the Gini impurity or entropy measures and can also calculate feature importances using the shuffling method. It also includes extra tools for analyzing feature importances, performing out-of-bag calculations, inspecting nodes, saving the trained forest along with importances in binary format, multi-thread support along with other amenities useful in our case. In this work, we used as features the triggered antenna distance to the shower axis (D_A), the detected electric field peak amplitude ($|E|$) and the spectral slope (SS) of each triggered antenna. The triggered antennas of each event were ordered with increasing distance to the axis, so the features used were D_{A1}, $|E_1|$, SS_1; D_{A2}, $|E_2|$, SS_2; (...); D_{AN}, $|E_N|$, SS_N, where index 1 refers to the closest antenna, 2 to the second closest and so on. The total number of features is then $N_f = 3N$, where N is the number of triggered antennas in the event with the highest number of antennas. For all events with less antennas, the missing features were substituted by zeros.

The forests used, one for each zenith angle, were composed of 200 trees and used the Gini impurity measure. The maximum depth for nodes was set to 100 and the minimum samples per leaf to 10. In order to reduce over-fitting, each tree used a subset of all features. This value was set differently for each zenith angle. Although the more traditional approach is to set this value to around $\sqrt{N_f}$, we found that in our case we obtained the best results, with minimal over-fitting, for values much higher than that, averaging around $\sim 0.8N_f$. We also varied the bootstrap size, starting with $0.1N_E$ events, where $N_E = 5000$ is the total number of events in the training data set, and only increased the size as needed to achieve good and reasonably stable accuracies on the test data set. Most forests used a bootstrap with around $0.2N_E$ random events per tree. We also used an extra variable that removed the last N_R features from the analysis (the ones related to the furthest antennas). Since most events have many fewer antennas than the event with most antennas, the last features tended to be dominated by zeros. The actual value of N_R we used varied with zenith. But, in general, small values of N_R not only gave the best results, but also tended to slightly diminish the time needed for training. That said, we did not exhaustively optimized these parameters and there is still room for modest improvements in our discrimination accuracy results, if a more thorough sweep of the RF parameter phase-space is performed.

4. Results

We have started our analysis by only using the distance D_A to the axis and the peak amplitude $|E_A|$ as RF features. At this stage, we were baffled by the good accuracies obtained by such a simple method, even when using a huge 30% energy uncertainty in test runs. An analysis of the feature importances hinted that proton showers tended to be much brighter than iron ones inside the Cherenkov cone, especially close to the core. This motivated us to revisit the Radio Lateral

Distribution Function (RLDF) and investigate this apparently very large amplitude dependence on X_{max} /composition, which the RF was using for the discrimination (see our other contribution [7]).

Our discrimination accuracy results can be seen on Fig. 3. When using only the amplitudes and distances with the normal 10% EM energy smearing, our maximum accuracy was ∼ 87% at $\theta = 50°$, which tended to decrease with increasing zenith (blue line on Fig. 3). When using only the spectral slopes and distances as features, we observed the opposite trend, with a maximum accuracy of ∼ 96% at $\theta = 78°$, which then tended to decrease with decreasing zenith, reaching a minimum of ∼ 79% at $\theta = 58°$, and then increasing again to ∼ 83% at $\theta = 50°$. The observed opposite trends when using only the slope and only the amplitude is fortuitous, as using one or the other, depending on the zenith region, can provide good accuracies over the whole zenith range studied. We then proceeded to use all features, i.e., the distances and both the slopes and amplitudes, at the same time (green line on Fig. 3). This, as expected, tended to significantly increased our discrimination accuracies, in general leading to greater accuracies than using either feature separately. So, the final results for our discrimination accuracies varied from ∼ 81% at $\theta = 58°$ to an amazing ∼ 96% at $\theta = 82°$.

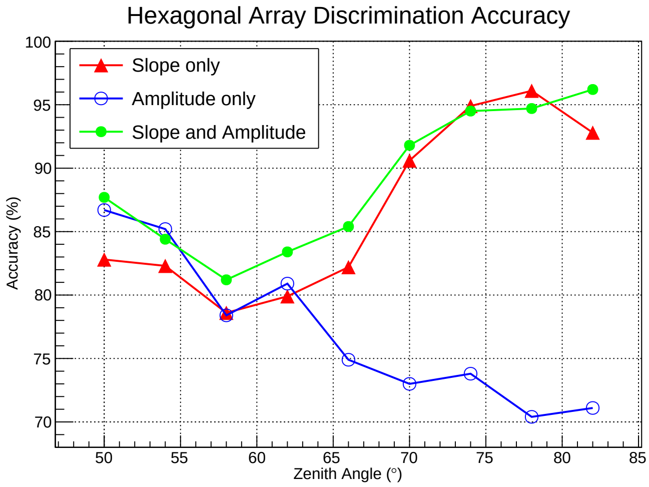

Figure 3: RF discrimination accuracy vs zenith angle.

Looking at the feature importances when using both the slope and the amplitudes at the same time corroborated the zenith behavior of using one or the other feature set separately. At $\theta = 82°$ (top panel of Fig. 4), the most important features are the spectral slopes in the region where the slope overlap between proton and iron was smallest, with no contribution from the amplitude related features. On the other hand, at $\theta = 62°$ (bottom panels of Fig. 4), an analysis of the feature importances shows that the RF is using both, the slopes and the amplitudes, depending on the distance to the antenna. At lower antenna distances, where the difference in amplitude between p and Fe is maximum, the most important feature is indeed the amplitude of the closest antenna (bottom left panel). In the mid-distance range, where the p-Fe slope overlap is smallest, the RF now mostly uses the slopes for the discrimination (bottom right panel). At the largest distances, the amplitude becomes important again, maybe due to the fact that proton footprints tend to be smaller than their Fe counterparts (bottom left panel).

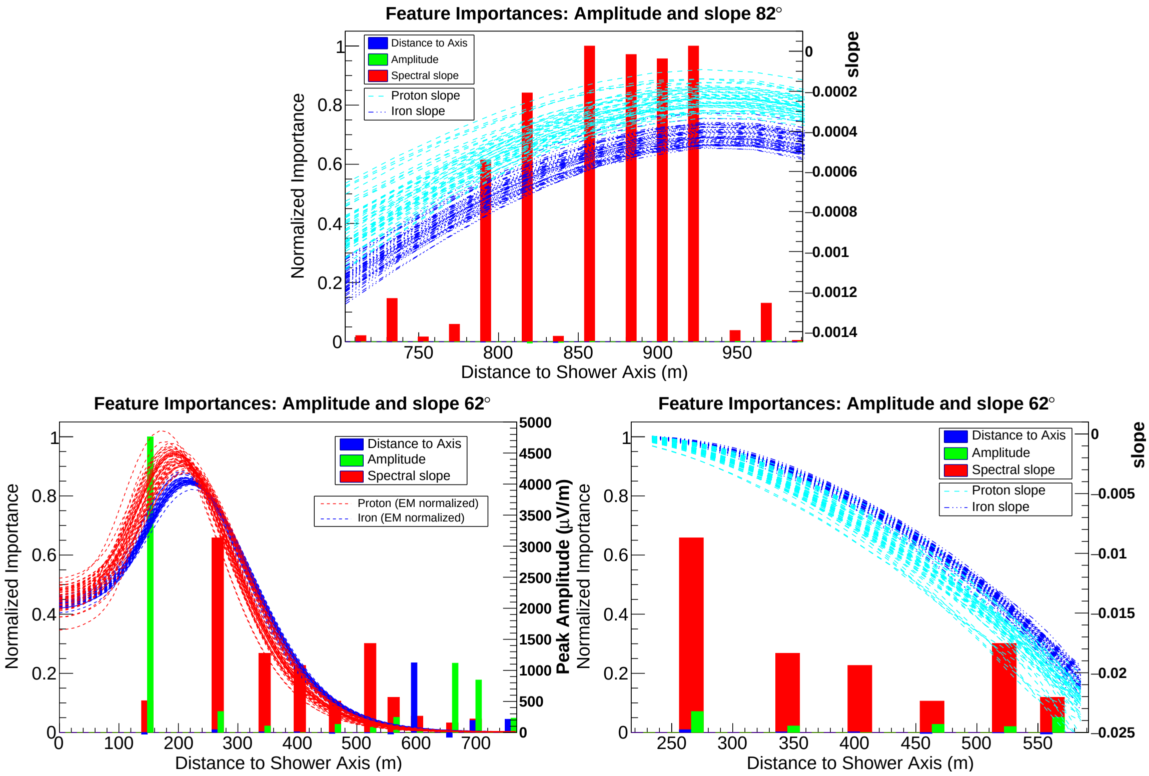

Figure 4: Top: Feature importances at $\theta = 82°$, normalized to the most important feature, as a function of the average antenna distance to the shower axis relating to each feature. Also shown is the spectral slope as a function of antenna distance for all input ZHAires simulations at the same zenith. Bottom left: Same as top, but for $\theta = 62°$and showing the amplitudes obtained from all simulations, normalized by the EM energy of each shower. Bottom right: Same as bottom left, but zoomed to the mid-distance region and showing the spectral slopes from the full simulations instead of amplitudes.

5. Discussion and conclusions

Although our results show incredibly good discrimination accuracies, varying from 81 to 96%, it is prudent to point out some possible caveats. The first one is that we have not yet included noise in this analysis. But as this method is geared more towards higher energy showers, we don't expect a large effect, especially on the amplitudes. However, the galactic noise, which is much higher at lower frequencies, may slightly alter the spectral slopes of antennas far away from the Cherenkov ring, possibly degrading the slope discrimination power slightly. Also, one should note that the quoted accuracies are for the specific data set we used in this work, i.e., the specific 100 ZHAireS showers per zenith angle we used to create the 10k RDSim events. If we were to use a different set of simulations, the p-Fe slope, amplitude and X_{max} overlaps could change slightly, impacting the accuracies. Another point is that the shower-to-shower fluctuations in RDSim come exclusively from the multiple simulations used to create instances of the emission model. While we have 10k events per zenith, and RDSim is capable of rotating the showers and change the observables accordingly, we still only have 100 different showers and thus only 100 different values of X_{max} . Using a different input simulation set could, in principle, slightly change the X_{max} overlap of p and Fe and also impact accuracies. In this aspect, the choice of hadronic model used in the simulations is relevant too, as it will change the X_{max} distributions. Yet another caveat is that we don't really know with precision how closely the simulations resemble real showers in terms of amplitudes, spectral slopes and X_{max} distributions. So, one should think of a "simulation uncertainty" that could impact the results in the case of real measurements. Finally, the detector array used in this work is very dense. Using a less dense array, such as the 1 km spacing array proposed for GRAND, will in general lead to fewer triggered antennas and fewer available features. We did some very preliminary tests by doubling the antenna spacing and (tentatively) found a small decrease in accuracy (<10%), mostly at the lower zenith angle range. All that said, since we are starting with such high accuracies, it is very unlikely that all these factors could completely destroy the method and render it useless.

We have developed a method that uses classification RFs to discriminate between light (p) and heavy (Fe) CR primary compositions on an event-by-event basis. Although the discrimination power comes from observables that depend on X_{max} , the method infers the composition directly, without reconstructing X_{max} itself. As features of the RF we used the distance to the shower axis, the peak amplitude of the electric field and the spectral slope of each triggered antenna. Our results show very good discrimination accuracies, between 81 and 96%, even when normalizing the emission of each shower by its own EM energy and including an extra 10% energy uncertainty to the generated events.

By performing a feature importance analysis, we uncovered a very large e-field amplitude dependence on X_{max} (see [7]), even when accounting for differences in the EM energy of each shower. This analysis also show that the discrimination power of the method at high zenith angles is mostly due to the spectral slopes, while at lower zenith angles both the slopes and amplitudes contribute to the discrimination power, with a slightly higher contribution from the latter.

Although in this work we have not accounted for several sources of uncertainty and biases, we are confident that the impact of these factors will not have a huge impact on the method, although the discrimination accuracies may decrease as these are taken into account in the future. But we start with such high accuracies that even a very significant decrease in accuracies will still keep the method useful.

References

[1] W. Carvalho Jr. and J. Alvarez-Muñiz, EPJ Web of Conferences **216**, 02005, (2019)

[2] W. R. Carvalho Jr, J. Alvarez-Muñiz, Astropar. Phys. **109**, 41-49, (2019)

[3] Washington R. de Carvalho Jr. and Abha Khakurdikar, PoS (ECRS) **079**, (2022)

[4] Washington R. de Carvalho Jr. and Abha Khakurdikar, PoS (ARENA2022) **055**, (2022)

[5] W. R. de Carvalho Jr., A. Khakurdikar and J. Hörandel, PoS (ICRC2023) **1097**, (2023)

[6] J. Alvarez-Muñiz, W.R. Carvalho, E. Zas, Astropart. Phys. **35**, 325 (2012)

[7] Washington R. de Carvalho Jr., PoS (ICRC2025) **213**, (2025)

[8] J. Alvarez-Muniz, W. Carvalho Jr., H. Schoorlemmer, E. Zas, Astropar. Phys., **59**, 29, (2014)

[9] F. Schlüter and T. Huege, JCAP01(**2023**),008, (2023)