PROCEEDINGS OF SCIENCE

A Deep Learning Method for Event Recognition in CALET Data

Adrien Picquenot,[a,*] **Michela Negro**[a] **and Nicholas Cannady**[b] **and the CALET collaboration**

[a] *Department of Astronomy, Louisiana State University, Baton Rouge, Louisiana, USA*
[b] *NASA/GSFC, Greenbelt, Maryland, USA*

E-mail: apicquenot@lsu.edu

The Calorimetric Electron Telescope (CALET) is a powerful tool to observe cosmic-ray electrons between 1 GeV and 20 TeV. Its 30 radiation-length calorimeter enables total containment of electron-induced showers up to TeV energies, yielding an energy resolution of $\sim 2\%$ for these events. The CALET all-electron spectrum obtained using the first 7.5 years of data closely matches the one produced by the AMS-02, but in tension with DAMPE and Fermi-LAT in the 60 GeV - 2 TeV energy range. To investigate this tension, we explore the possibility of a bias in the event selection by developing an alternative classification method between electrons and protons using machine learning techniques instead of a deterministic algorithm. These unsupervised learning techniques are used to find clustering in the flight data events without training on simulated data. Here we present preliminary results from this analysis, and the performance of the trained method when applied to the simulated dataset.

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

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

*Speaker

PoS(ICRC2025)113

1. Introduction

Cosmic ray electrons (CREs, referring here both to electrons and positrons) are challenging to observe, as they constitute less than 2% of the primary CRs. Separation even becomes increasingly difficult at higher energies, CREs having a softer spectrum than protons. This difference can be explained by the energy loss mechanisms that distinguish between hadronic and leptonic components of CRs. For electrons, the interactions with the interstellar medium (ISM) are purely electromagnetic, with ionization losses being dominant up to MeV energies, bremsstrahlung and adiabatic losses being important around 10 MeV, and inverse Compton scattering (IC) and synchrotron radiation becoming the dominant processes at energies above a few GeV. As low energy CRE are suppressed by the geomagnetic cut-off, instruments in Low Earth Orbit, like CALET, can only observe events above a few GeV.

The Calorimetric Electron Telescope (CALET) is a space experiment installed at the Japanese Experiment Module–Exposed Facility (JEM-EF) on the International Space Station (ISS) and operational since October 2015. Its primary objective is to observe cosmic rays and it is optimized for the measurement of the all-electron spectrum [2]. CALET's capabilities allowed for the first ever significant observations reaching into the TeV region, with an all-electron spectrum ranging from 10 GeV to 3 TeV [5]. An updated spectrum was later published with nearly doubled statistics and the full geometrical acceptance in the high-energy region [6]. The most recent publication presented another increase in statistics by a factor of about 3.4, with 2637 days of flight data ranging from October 13, 2015 to December 31, 2022. The spectrum integrated 7.02 million electrons and positrons, above 10.6 GeV and up to 7.5 TeV [1]. This all-electron spectrum is shown in Fig.1, together with the results from other recent experiments in space (AMS-02, Fermi-LAT, and DAMPE).

When compared with other instruments, the CALET spectrum shows good agreement with AMS-02 data up to 2 TeV. The observed electron spectrum follows an approximate power-law behavior with a cutoff at TeV energies. In this work, we investigate the impact of data selection strategies, particularly the discrimination between electrons and protons by exploring an alternative approach based on unsupervised machine learning.

2. Instrument

CALET consists of two instruments: the Calorimeter (CAL) that we use in this work, and the CALET Gamma-ray Burst Monitor (CGBM). The CAL has a field of view of about 45° from zenith and a geometrical factor of 1040 cm^2 sr for high-energy electrons. and can observe high-energy electrons from 1 GeV to 20 TeV, protons, helium, and heavy nuclei from 10 GeV to 1000 TeV and gamma-rays from 1 GeV to 10 TeV. The CAL consists of three components: a Charge Detector (CHD), a 3 radiation-length thick imaging calorimeter (IMC), and a 27 radiation-length thick total absorption calorimeter (TASC). The CHD is located above the IMC and measures the charge of the incident particle. The IMC induces the start of the shower development for electromagnetic particles while suppressing nuclear interactions in order to maximize the proton rejection power for the electron candidates, and provides the direction of incident particles. The TASC installed below the IMC measures the energy of shower particles caused by the interactions of the incident

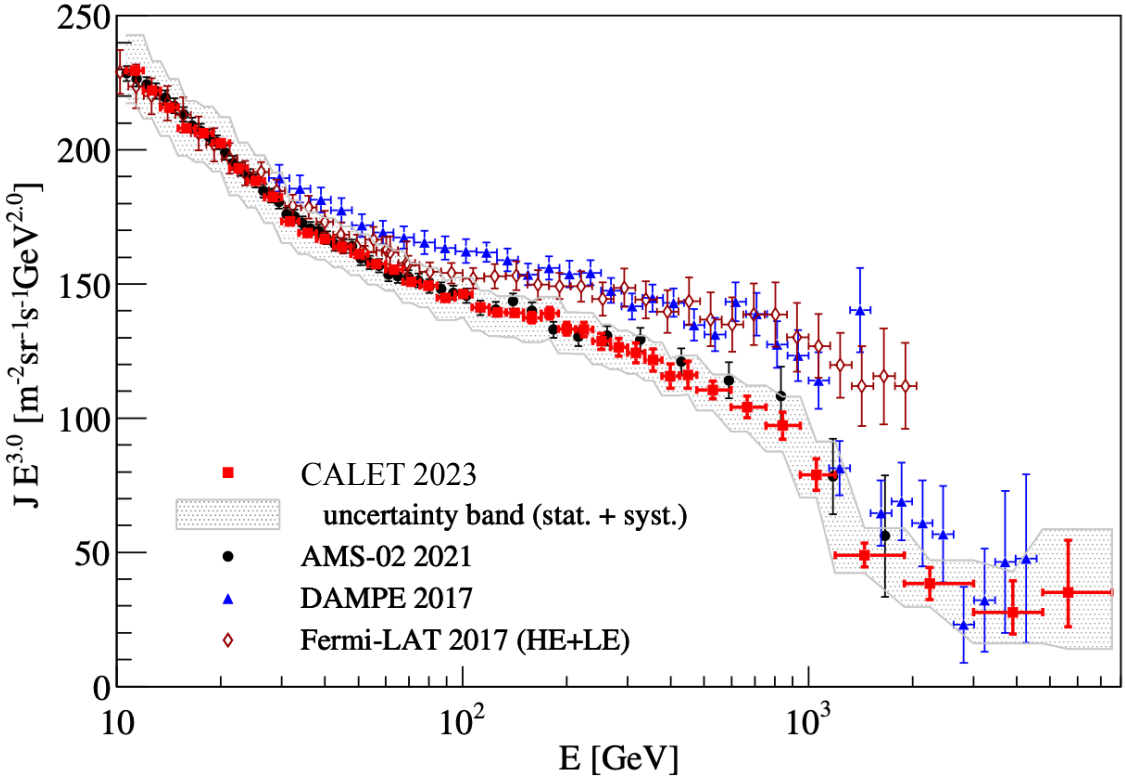

Figure 1: Cosmic-ray electron + positron spectrum observed with CALET from 10.6 GeV to 7.5 TeV, compared with other direct measurements. The gray band indicates the quadratic sum of statistical and systematic errors (not including the uncertainty on the energy scale). Figure from Ref.[1].

particles in the IMC, allowing for a nearly total containment of TeV-electron showers. We are primarily interested in the data gathered with the high-energy (HE) trigger mode from CAL, targeting electrons in the 10 GeV - 20 TeV energy range. More details on the CAL detector and triggers can be found in Ref.[7].

Before any other selection step, a number of precuts are applied both on the simulations and the data to keep only well-reconstructed and well-contained single-charged events.

3. Supervised learning: Boosted Decision Trees

The analysis to obtain the all-electron spectrum from Ref.[1] shown in Fig.1 was carried out on the data collected with the HE trigger. A Monte Carlo (MC) program was used to simulate physics processes and detector response based on the simulation package EPICS [8] (EPICS9.20/COSMOS8.00). Our simulated sample of electrons have 20M events per decade from 200 MeV to 20 TeV and simulated sample of protons have 100M events per decade from 200 MeV to 20 TeV. Both species are simulated with a E^{-1} energy distribution.

The electron selection is performed through boosted decision trees (BTDs) [3], a supervised machine learning algorithm that iteratively add small decision trees to minimize the errors made by the previous trees in predicting the true nature of the events used to train the network. The output is then defined as a weighted combination of the individual trees outputs. BDTs can discriminate

between classes with great predictive accuracy in structured data, but are very sensitive to the initial tuning of hyperparameters such as the learning rate, the number of trees and the tree depth.

The BDT method from Ref.[1] was trained and optimized on MC simulations, using variables describing parameters such as the charge measured by the CHD or the IMC and TASC shower profiles to discriminate between events. It was trained in the 476 - 949 GeV energy range with 9 parameters, and with 13 parameters above 949 GeV. This method yields good results: the contamination ratios of protons are 5% up to 1 TeV, and less than 10% in the 1–7.5 TeV region, while keeping a constant high efficiency of 80% for electrons. Nonetheless, BDTs being highly sensitive to their initial tuning [4], as well as slight discrepancies between data and MC simulation may affect the final result, in a way that is hard to estimate accurately.

In this work, we aim to apply unsupervised machine learning methods to CALET data to extract cosmic-ray electron (CRE) events, with the goal of minimizing reliance on simulations and producing a result that is as data-driven as possible.

4. Unsupervised Learning

Unsupervised learning is a type of machine learning where the algorithm learns patterns and structure from data without being given explicit labels or instructions. For a classification problem, such as discriminating between electrons and protons in a set of observed events, it means that the learning phase is entirely based on intrinsic features of input observables and does not involve labeled data as in supervised learning methods; the algorithm only searches for patterns in the data to base its clustering on.

While it is not strictly necessary to find an optimal representation of the data— a sufficiently powerful algorithm with enough data can learn to disregard uninformative variables— it is still beneficial to guide the process by identifying features that are most relevant for classification, thereby improving efficiency and performance. In particular, working with simulations allows to measure the efficiency and contamination of the method and improve on those performances before running the real data through it.

We identified 8 variables describing the IMC and TASC shower profiles, similar to what was previously done for the BDT analysis. The correlation matrix in Fig.2 indicates that these variables are uncorrelated, supporting the validity of using them together in the analysis. We consider the energy range from 30 GeV to 10 TeV and we perform independent classifications for 10 logarithmic energy bins. We keep the same set of 8 variables for all energy bins considered. We chose these variables for their discriminative capacities based on the MC simulations, and for the similarities in distribution of these variables between the MC simulations and the actual data. Even if our method will be applied to the data directly without a training based on the use of MC simulations, we can still make use of the simulations to evaluate and compare the capabilities of different available algorithms and fine-tune their hyper-parameters. We checked the data-MC agreement for the variables we identified by performing a template fitting according to standard procedures as outlined in previous works [5].

Among the available clustering algorithms, Automatic and Hierarchical Gaussian Mixture Modeling (AutoGMM) [9] performs well in presence of mixture model scenarios with Gaussian-

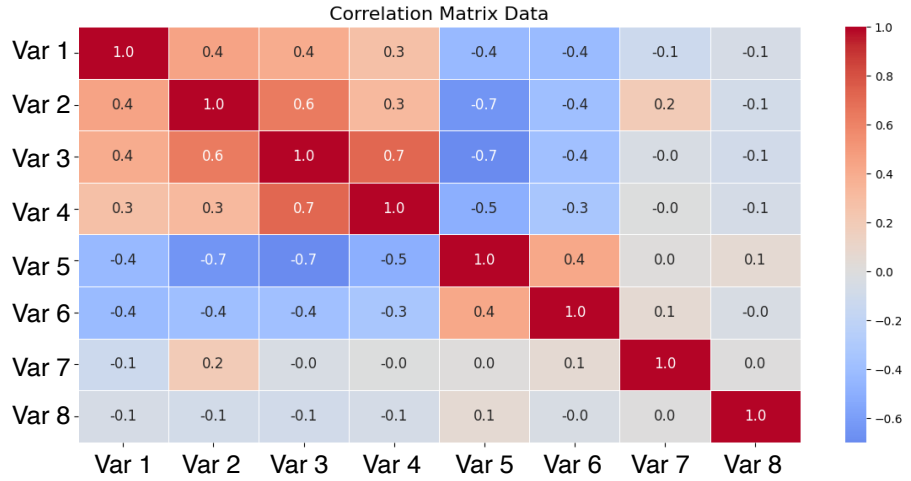

Figure 2: Correlation matrix between the 8 variables we selected for our analysis.

like clusters with moderate dimensions. We first approached this problem utilizing AutoGMM and we present the preliminary studies on simulated data here.

AutoGMM was fed with the nx8 simulated dataset, where 8 is the number of features (the 8 variables we identified) and n the total number of simulated events (both electrons and protons) after pre-cuts. We found that this basic classification test yielded poor results: the AutoGMM algorithm failed to distinguish meaningful particle classes, mislabeling the majority of events as belonging to one class, not reproducing the injected fraction of electrons.

We proceeded with a standard machine learning approach of dimensionality reduction, to reduce the number of features representing the data. Such procedures aim to maintain only the main critical features that characterize the data in a lower dimensional space. The Principal Component Analysis (PCA), is a classic example of such a method, useful when relations among the features are known to be linear. On the other hand, the Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) [10] is excellent in finding non-linear patterns among features and outperforms most of the dimensionality reduction algorithms available in terms of separation of different events in the lower dimensional space. We applied UMAP to reduce the dimensionality of the data to a *latent* space of dimension n for every n between 2 and 7, and applied AutoGMM clustering algorithm on the latent space. We repeat this in every energy bin. To evaluate the performances, we looked at how well the predicted classification would reproduce the injected electron fraction $\frac{N_e}{N_{e+p}}$, where N_e is the number of electrons and N_{e+p} the number of total events. The best results, across all energy bins, were obtained when the latent space dimension was 4. We defined error bars associated to the estimated electron fractions by looking at the span of this fraction obtained when using different n between 2 and 8. In the next section we report the results of this preliminary study.

5. Results and Perspectives

The results of this study are shown in Fig.3. The reconstructed electron fractions obtained with our method between 30 GeV to 10 TeV on MC simulations are consistently close to the true

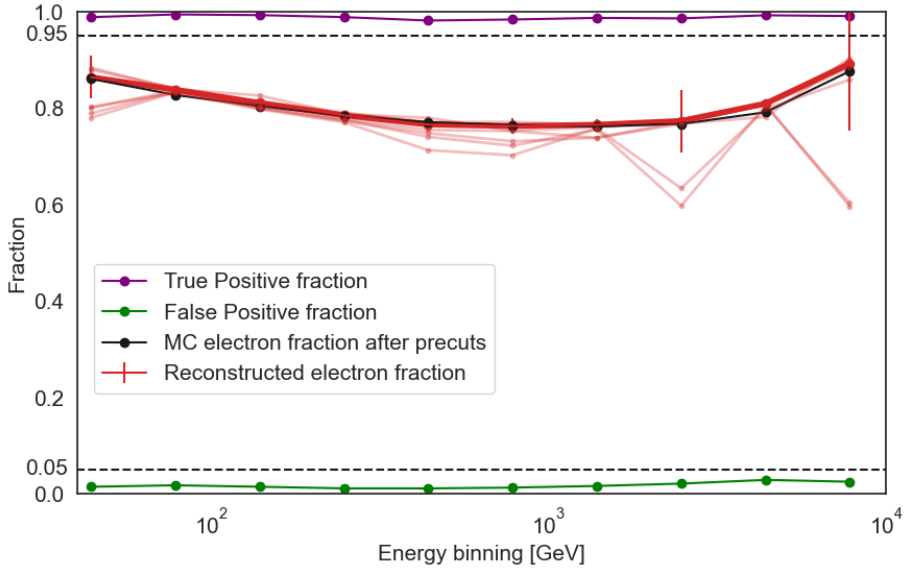

Figure 3: Reconstructed electron fractions on MC simulations in red, compared with the surviving electron fractions after precuts, for ten energy bins between 30 GeV to 10 TeV. The main results are obtained with a latent space of dimension 4, while the results obtained with different dimensions are shown in transparency. Error bars are marking the standard deviation of the results obtained with the different latent space dimensions. We can see that the False Positive fraction of protons mislabeled as electrons stays below 0.05, and that the True Positive fraction of electrons correctly labelled stays above 0.95.

electron fractions for the ten energy bins we defined. The true positive percentage of electrons being correctly labeled as electrons stays significantly above 95%, while the false positive protons being mislabeled as electrons stays below 5%.

Our current preliminary results show that the latent space we use as a representation to perform the clustering accurately captures discriminating features to separate electrons from protons. Our next objective is to ensure that the clustering step, currently performed by AutoGMM, still gives accurate results with lower relative electron statistics. This method constitutes a promising alternative to the BDT classification that does not rely on simulated datasets. In the future, we will continue to develop the pipeline by conducting a systematic study comparing different clustering algorithms. To avoid biasing the final results, the selection of the most effective algorithm will be based on performance evaluated using simulated datasets, before comparing our results with the standard BDT approach and applying our method to the real flight data.

6. References

[1] O. Adriani et al. (CALET Collaboration), "Direct measurement of the spectral structure of cosmic-ray Electrons + Positrons in the tev region with calet on the international space station," Phys. Rev. Lett., 131:191001 (2023)

[2] Shoji Torii and Pier Simone Marrocchesi, "The calorimetric electron telescope (calet) on the international space station," Advances in Space Research, 64(12):2531–2537 (2019)

[3] Friedman, J. H. "Greedy function approximation: A gradient boosting machine." Annals of Statistics, 29(5), 1189–1232 (2001)

[4] https://www.geeksforgeeks.org/machine-learning/how-to-tune-a-decision-tree-in-hyperparameter-tuning/

[5] O. Adriani et al. (CALET Collaboration), "Energy spectrum of cosmic-ray electron and positron from 10 gev to 3 tev observed with the calorimetric electron telescope on theinternational space station," Phys. Rev. Lett., 119:181101 (2017)

[6] O. Adriani et al. (CALET Collaboration), "Extended measurement of the cosmic-ray electron and positron spectrum from 11 gev to 4.8 tev with the calorimetric electron telescope on the international space station," Phys. Rev. Lett., 120:261102 (2018)

[7] Y. Asaoka et al., "n-orbit operations and offline data processing of CALET onboard the ISS," Astropart.Phys., 100:29–37 (2018)

[8] K. Kasahara, "Introduction to Cosmos and some Relevance to Ultra High Energy Cosmic Ray Air Showers," In International Cosmic Ray Conference, vol.1(1995).

[9] T. Athey et al., "Autogmm: Automatic and hierarchical gaussian mixture modeling in python," (2021).

[10] L. McInnes et al., "Umap: Uniform manifold approximation and projection for dimension reduction," (2021).

CALET author list

O. Adriani,[1,2] Y. Akaike,[3,4] K. Asano,[5] Y. Asaoka,[5] E. Berti,[2,6] P. Betti,[1,2] G. Bigongiari,[7,8] W.R. Binns,[9] M. Bongi,[1,2] P. Brogi,[7,8] A. Bruno,[10] N. Cannady,[11] G. Castellini,[6] C. Checchia,[7,8] M.L. Cherry,[12] G. Collazuol,[13,14] G.A. de Nolfo,[10] K. Ebisawa,[15] A. W. Ficklin,[12] H. Fuke,[15] S. Gonzi,[1,2,6] T.G. Guzik,[12] T. Hams,[16] K. Hibino,[17] M. Ichimura,[18] M.H. Israel,[9] K. Kasahara,[19] J. Kataoka,[20] R. Kataoka,[21] Y. Katayose,[22] C. Kato,[23] N. Kawanaka,[24,25] Y. Kawakubo,[26] K. Kobayashi,[3,4] K. Kohri,[25,27] H.S. Krawczynski,[9] J.F. Krizmanic,[11] P. Maestro,[7,8] P.S. Marrocchesi,[7,8] M. Mattiazzi,[13,14] A.M. Messineo,[8,28] J.W. Mitchell,[11] S. Miyake,[29] A.A. Moiseev,[11,30,31] M. Mori,[32] N. Mori,[2] H.M. Motz,[33] K. Munakata,[23] S. Nakahira,[15] J. Nishimura,[15] M. Negro,[12] S. Okuno,[17] J.F. Ormes,[34] S. Ozawa,[35] L. Pacini,[2,6] P. Papini,[2] B.F. Rauch,[9] S.B. Ricciarini,[2,6] K. Sakai,[36] T. Sakamoto,[26] M. Sasaki,[11,30,31] Y. Shimizu,[17] A. Shiomi,[37] P. Spillantini,[1] F. Stolzi,[7,8] S. Sugita,[26] A. Sulaj,[7,8] M. Takita,[5] T. Tamura,[17] T. Terasawa,[5] S. Torii,[3] Y. Tsunesada,[38,39] Y. Uchihori,[40] E. Vannuccini,[2] J.P. Wefel,[12] K. Yamaoka,[41] S. Yanagita,[42] A. Yoshida,[26] K. Yoshida,[19] and W. V. Zober[9]

(CALET Collaboration)

[1] *Department of Physics, University of Florence, Via Sansone, 1 - 50019, Sesto Fiorentino, Italy*
[2] *INFN Sezione di Firenze, Via Sansone, 1 - 50019, Sesto Fiorentino, Italy*
[3] *Waseda Research Institute for Science and Engineering, Waseda University, 17 Kikuicho, Shinjuku, Tokyo 162-0044, Japan*
[4] *JEM Utilization Center, Human Spaceflight Technology Directorate, Japan Aerospace Exploration Agency, 2-1-1 Sengen, Tsukuba, Ibaraki 305-8505, Japan*
[5] *Institute for Cosmic Ray Research, The University of Tokyo, 5-1-5 Kashiwa-no-Ha, Kashiwa, Chiba 277-8582, Japan*
[6] *Institute of Applied Physics (IFAC), National Research Council (CNR), Via Madonna del Piano, 10, 50019, Sesto Fiorentino, Italy*
[7] *Department of Physical Sciences, Earth and Environment, University of Siena, via Roma 56, 53100 Siena, Italy*
[8] *INFN Sezione di Pisa, Polo Fibonacci, Largo B. Pontecorvo, 3 - 56127 Pisa, Italy*
[9] *Department of Physics and McDonnell Center for the Space Sciences, Washington University, One Brookings Drive, St. Louis, Missouri 63130-4899, USA*
[10] *Heliospheric Physics Laboratory, NASA/GSFC, Greenbelt, Maryland 20771, USA*
[11] *Astroparticle Physics Laboratory, NASA/GSFC, Greenbelt, Maryland 20771, USA*
[12] *Department of Physics and Astronomy, Louisiana State University, 202 Nicholson Hall, Baton Rouge, Louisiana 70803, USA*
[13] *Department of Physics and Astronomy, University of Padova, Via Marzolo, 8, 35131 Padova, Italy*
[14] *INFN Sezione di Padova, Via Marzolo, 8, 35131 Padova, Italy*
[15] *Institute of Space and Astronautical Science, Japan Aerospace Exploration Agency, 3-1-1 Yoshinodai, Chuo, Sagamihara, Kanagawa 252-5210, Japan*
[16] *Center for Space Sciences and Technology, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, Maryland 21250, USA*
[17] *Kanagawa University, 3-27-1 Rokkakubashi, Kanagawa, Yokohama, Kanagawa 221-8686, Japan*
[18] *Faculty of Science and Technology, Graduate School of Science and Technology, Hirosaki University, 3, Bunkyo, Hirosaki, Aomori 036-8561, Japan*
[19] *Department of Electronic Information Systems, Shibaura Institute of Technology, 307 Fukasaku, Minuma, Saitama 337-8570, Japan*
[20] *School of Advanced Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku, Tokyo 169-8555, Japan*
[21] *National Institute of Polar Research, 10-3, Midori-cho, Tachikawa, Tokyo 190-8518, Japan*
[22] *Faculty of Engineering, Division of Intelligent Systems Engineering, Yokohama National University, 79-5 Tokiwadai, Hodogaya, Yokohama 240-8501, Japan*
[23] *Faculty of Science, Shinshu University, 3-1-1 Asahi, Matsumoto, Nagano 390-8621, Japan*
[24] *Department of Physics, Graduate School of Science, Tokyo Metropolitan University , 1-1 Minamii-Osawa, Hachioji, Tokyo 192-0397, Japan*
[25] *National Astronomical Observatory of Japan, 2-21-1 Osawa, Mitaka, Tokyo 181-8588, Japan*
[26] *Department of Physical Sciences, College of Science and Engineering, Aoyama Gakuin University, 5-10-1 Fuchinobe, Chuo, Sagamihara, Kanagawa 252-5258, Japan*
[27] *Institute of Particle and Nuclear Studies, High Energy Accelerator Research Organization, 1-1 Oho, Tsukuba, Ibaraki 305-0801, Japan*
[28] *University of Pisa, Polo Fibonacci, Largo B. Pontecorvo, 3 - 56127 Pisa, Italy*
[29] *Department of Electrical and Computer Engineering, National Institute of Technology (KOSEN), Gifu College, 2236-2 Kamimakuwa, Motosu-city, Gifu 501-0495, Japan*

PoS(ICRC2025)113

[30] *Center for Research and Exploration in Space Sciences and Technology, NASA/GSFC, Greenbelt, Maryland 20771, USA*
[31] *Department of Astronomy, University of Maryland, College Park, Maryland 20742, USA*
[32] *Department of Physical Sciences, College of Science and Engineering, Ritsumeikan University, Shiga 525-8577, Japan*
[33] *Faculty of Science and Engineering, Global Center for Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku, Tokyo 169-8555, Japan*
[34] *Department of Physics and Astronomy, University of Denver, Physics Building, Room 211, 2112 East Wesley Avenue, Denver, Colorado 80208-6900, USA*
[35] *Quantum ICT Advanced Development Center, National Institute of Information and Communications Technology, 4-2-1 Nukui-Kitamachi, Koganei, Tokyo 184-8795, Japan*
[36] *Kavli Institute for Cosmological Physics, The University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, USA*
[37] *College of Industrial Technology, Nihon University, 1-2-1 Izumi, Narashino, Chiba 275-8575, Japan*
[38] *Graduate School of Science, Osaka Metropolitan University, Sugimoto, Sumiyoshi, Osaka 558-8585, Japan*
[39] *Nambu Yoichiro Institute for Theoretical and Experimental Physics, Osaka Metropolitan University, Sugimoto, Sumiyoshi, Osaka 558-8585, Japan*
[40] *National Institutes for Quantum and Radiation Science and Technology, 4-9-1 Anagawa, Inage, Chiba 263-8555, Japan*
[41] *Nagoya University, Furo, Chikusa, Nagoya 464-8601, Japan*
[42] *College of Science, Ibaraki University, 2-1-1 Bunkyo, Mito, Ibaraki 310-8512, Japan*

PoS(ICRC2025)113