PROCEEDINGS OF SCIENCE

Classification of cosmic ray components using deep learning methods for LHAASO-KM2A

Weiyan Zhang,[a] **Xiaopeng Zhang,**[b,c] **Hongkui Lv**[b,c] **and Min Zha**[b,c,*]

[a] *Hebei Normal University,*
050024 Shijiazhuang, Hebei, China
[b] *Key Laboratory of Particle Astrophyics, Institute of High Energy Physics, CAS,*
100049 Beijing, China
[c] *TIANFU Cosmic Ray Research Center,*
610213 Chengdu, Sichuan, China

E-mail: zhangxp@ihep.ac.cn

LHAASO-KM2A is a pivotal facility for studying cosmic rays through extensive air shower detection. However, accurately classifying cosmic ray components (e.g., protons, helium nuclei, and heavy nuclei) remains challenging due to overlapping shower signatures and background noise. In this proceeding, we proposes a deep learning-based method to enhance the classification accuracy of cosmic ray components using KM2A simulation data. Current results demonstrate that the proposed method achieves a higher classification accuracy compared to the conventional method.

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

*Speaker

PoS(ICRC2025)323

1. Introduction

The origin of cosmic rays (CRs) has been an enigma in astrophysics for over a century. A critical step towards solving this puzzle is the precise determination of their chemical composition, i.e., the relative abundance of different nuclei as a function of their energy. A particularly revealing feature in the all-particle cosmic ray energy spectrum is the "knee", a distinct steepening of the power-law flux that occurs at an energy of approximately 4 PeV[1]. Precise localization of the knees of different chemical compositions is key to explore the hidden physics. Besides, the mass and charge of a primary cosmic ray dictate its acceleration efficiency in astrophysical sources and its trajectory through galactic and intergalactic magnetic fields. A precise measurement of how the composition evolves through the knee is therefore a primary scientific objective for understanding the origin, acceleration and propagation of CRs[2].

The Large High Altitude Air Shower Observatory (LHAASO) is a new-generation, multi-component instrument specifically designed to tackle the challenges of ultra-high-energy (UHE) gamma-ray astronomy and cosmic ray physics[3]. Situated at an altitude of 4410 meters in Sichuan, China, its location allows for the observation of extensive air showers (EAS)—the cascade of secondary particles produced when a primary CR interacts with the atmosphere—closer to their point of maximum development for showers of energies near the knee region. LHAASO employs a hybrid detection strategy, combining the Kilometer Square Array (KM2A), the Water Cherenkov Detector Array (WCDA), and the Wide Field-of-view Cherenkov Telescope Array (WFCTA) to provide comprehensive measurements of different EAS components. This study focuses on the capabilities of the KM2A.

LHAASO-like ground-based observatories infer the properties of the primary particles indirectly by measuring the characteristics of the EAS, which makes the components discrimination a rather challenging task. The traditional method of composition analysis relies on the number of detected muons (N_μ) relative to the number of electromagnetic particles (N_e)[4]. However, the performance of this approach is limited since it compresses the rich measurement (space and time) of a shower front into just two integrated quantities, inevitably losing a vast amount of potentially discriminating information.

In recent years, deep learning techniques have emerged as a powerful tool in nearly every filed of physics, including astroparticle physics. Experiments like the Pierre Auger Observatory and IceCube have successfully applied machine learning techniques in event reconstruction and classification[5, 6]. The classifying of cosmic-ray groups has also been carried out at LHAASO-KM2A with a Graph Neural Network (GNN)[7]. This work proposes to extend this paradigm to LHAASO by applying a Dynamic Graph Convolutional Neural Network (DGCNN) to exploit the full granularity of EAS data. The specific architecture chosen is ParticleNet[8], a model that has achieved state-of-the-art results in jet tagging at the LHC by treating particle collections as "particle clouds". The high density and hybrid nature of the LHAASO-KM2A provides an exceptionally rich and high-resolution particle cloud for each EAS event, making it an ideal system for this GNN-based approach.

Table 1: Total number of events in the full dataset.

	10 TeV - 100 TeV	100 TeV - 1 PeV	1 PeV - 10 PeV
Proton	2.09×10^6	5.79×10^5	9.64×10^4
He	1.62×10^6	5.74×10^5	9.73×10^4
CNO	1.11×10^6	5.63×10^5	9.74×10^4
MgAlSi	8.77×10^5	5.54×10^5	9.75×10^4
Fe	6.50×10^5	5.44×10^5	9.73×10^4

2. Method

2.1 Dataset

The analysis presented in this paper relies on simulated data of LHAASO-KM2A, which covers an area of 1.3 km^2 and composed of two types of detectors - 5216 electromagnetic particle detectors (EDs) and 1188 muon detectors (MDs), designed to detect electromagnetic and muonic components in EAS secondaries, respectively. The entire array is synchronized with sub-nanosecond precision using the White Rabbit protocol, ensuring accurate timing for shower front reconstruction.

To train and evaluate the composition discrimination algorithms, extensive Monte Carlo (MC) simulations are required, as the primary particle type is unknown in real data. The simulation process involves a two-step chain[9]. In the first step, the development of extensive air showers in the atmosphere is simulated using the CORSIKA software package with the hadronic model QGSJET II-04. The simulation has five primary components - protons (H), helium (He), CNO, MgAlSi, and iron (Fe). Each component covers primary energy ranging from 10 TeV to 10 PeV with a power law spectral of index -2, and categorized into three sub-groups by the magnitude of energy. In the second step, the secondary particles generated are propagated through a detailed model of the KM2A detectors using a Geant4-based simulation framework, producing simulated event data in ROOT files with the same format as the experimental data. Reconstruction of the simulated data was carried out using the same algorithm (described in Ref[10]) as applied to the experimental data, to get the reconstructed core position, energy, and direction. Other key variables of hits are time, charge and coordinates of each detector. For each event, the original coordinates (x, y, z) are also transformed into the shower disc coordinate system (u, v, w) to align their core and axis.

Simulation events are then filtered using following criteria:

- 200 m < $R_{\rm core}$ < 500 m, where $R_{\rm core}$ is the distance from the reconstructed shower core to the array center
- $\theta < 35^\circ$, where θ is the reconstructed zenith angle

The size of the datasets, after event selection, are listed in Table 1. The selected events are split into train, test, and evaluation data sets, with a ratio of 3:1:1.

Table 2: Input features of particles used in the network.

Variable	Definition
(u, v, w)	Transformed coordinates of the detector
N_{pe}	Number of photoelectrons generated
t	Arriving time of a secondary particle
mode	Detector type (0 for EDs, 1 for MDs)

2.2 Model and Training

The architecture of ParticleNet is shown in Figure 1, which is developed from the DGCNN. The central operation of DGCNN is the edge convolution or EdgeConv[11]. For each point in the cloud, EdgeConv constructs a local graph by identifying its k-nearest neighbors. It then applies a shared multi-layer perceptron (MLP) to the features of the central point and the relative features of its neighbors. The results are then aggregated by a mean operation to produce an updated feature vector for the central point.

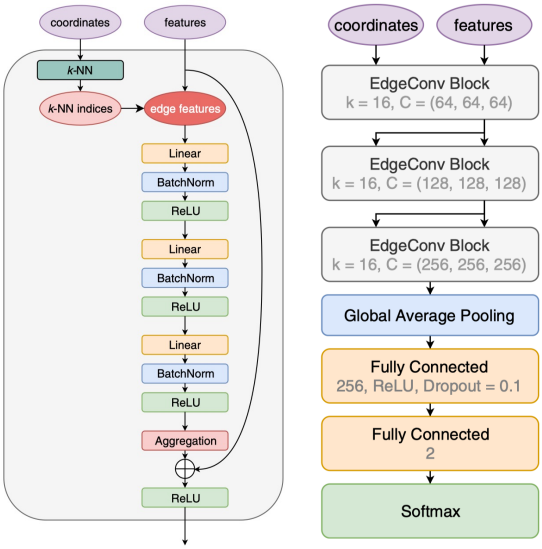

Figure 1: The architecture of ParticleNet[8]

The specific network we used consists of three EdgeConv blocks, one global average pooling block, and two fully-connected blocks followed by a softmax function to output the classification probabilities. The input features fed into the network are listed in Table 2. In the first EdgeConv block, the nearest neighbors are determined based on the physical coordinates of the detectors. In subsequent blocks, the neighbors are re-calculated in the high-dimensional feature space learned by the previous layer.

2.3 Baseline approach

The traditional method for cosmic ray composition analysis, which serves as the baseline for our comparison, is founded on the well-understood differences in how light and heavy nuclei interact

with the atmosphere. Heavy nuclei, such as iron, possess a significantly larger inelastic cross-section than light nuclei like protons. This causes them to initiate their air showers at higher altitudes on average. A shower from a heavy primary begins earlier and develops more rapidly, compared to a proton-induced shower of the same total energy. Consequently, its electromagnetic component suffer more attenuation, while $\pi\pm$ components will have more opportunity to decay into penetrating muons. This leads to a higher N_μ and a lower N_e by the time the shower reaches the ground. A simple cut on a parameter defined as

$$f(N_e, N_\mu) = \frac{N_\mu}{N_e^{0.85}}$$

is then constructed using these two integrated variables to distinguish between primary types.

3. Performance

To quantitatively evaluate the effectiveness of the ParticleNet model against the traditional baseline, we define two binary classification tasks: Proton identification (P task) aims to select proton-initiated showers from all other components, and the light component identification (L task) aims to isolate the light-mass component (proton and helium). The ratio and spectrum index in the datasets differ from the actual flux of cosmic rays, to address this, all events are weighted based on their primary energies according to the Horandel model[12]. The trained ParticleNet model and the baseline classifier were applied to an independent test set.

The resulting Receiver Operating Characteristic (ROC) curves for the two classification tasks are shown in Figure 2(a) and Figure 2(b), respectively. In these curves, the x axis refers to the false positive rate, indicating the background retention rate. The y axis refers to as the true positive rate, indicating the signal efficiency. The more the curve bends towards the top left corner, the better the classification performance. So one can clearly see that the ParticleNet model is better than the baseline for both tasks.

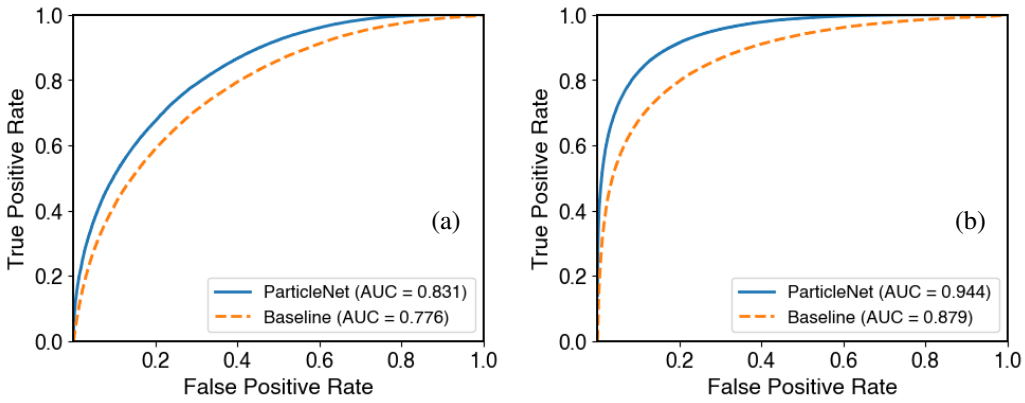

Figure 2: ROC curves for P task (a) and L task (b).

The area under the ROC curve, or AUC value, is commonly used as a quantitative metric of a classifiers's performance. We partitioned the test dataset into a series of energy bins to assess its performance as a function of the primary energy. Then AUC value for each bin is calculated

independently. The results are presented in Figure 3(a) for the P task and Figure 3(b) for the L task. These plots clearly demonstrate that ParticleNet consistently and significantly outperforms the traditional N_μ/N_e method across the entire energy range.

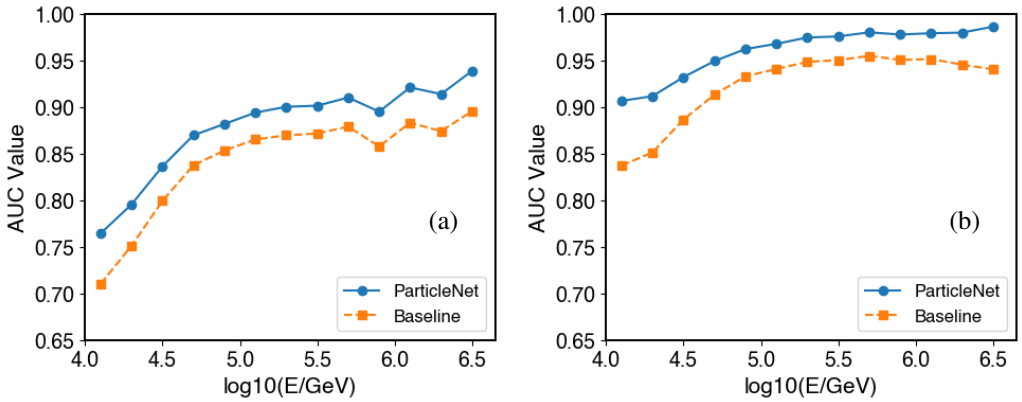

Figure 3: AUC values in different energy bins for P task (a) and L task (b).

4. Discussion

The results presented demonstrate that the ParticleNet model provides a dramatic improvement in cosmic ray composition discrimination over the traditional method. Although, like most applications of deep learning technologies in physics, it is challenging to provide an accurate and in-depth explanation for this performance improvement, it can be speculated that GNN networks possess the ability to abstract features from rich, high-dimensional data that assist in identifying different components' characteristics.

It should be noted that these results are still preliminary. After optimizing the network structure, fine-tuning hyperparameters, and increasing the data volume, the identification performance is expected to be further improved. In the future, more metrics directly related to physical measurements will be introduced, and models will be applied to experimental observation data to evaluate the capabilities of different models in solving cosmic ray composition and energy spectrum problems.

5. Summary

In this proceeding we investigate the application of a deep learning methodology for cosmic ray composition analysis for the LHAASO-KM2A experiment. Motivated by the idea of representing extensive air shower events as "particle clouds", we built up a ParticleNet architecture, a type of Dynamic Graph Convolutional Neural Network. Each point in the cloud represents a hit detector, characterized by its position, signal time, measured charge, and detector type. The network is trained on the KM2A simulated data.

The performance of this GNN-based approach was benchmarked against a conventional method based on the correlation between N_μ and N_e. On the crucial tasks of identifying primary protons

and the combined light component (protons and helium), the new method demonstrated a stable and substantial improvement.

References

[1] G.V. Kulikov and G.B. Khristiansen, *On the size spectrum of extensive air showers*, *Sov. Phys. JETP* **35** (1959) 441.

[2] Z. Cao, F. Aharonian, Axikegu, Y.X. Bai, Y.W. Bao, D. Bastieri et al., *Measurements of All-Particle Energy Spectrum and Mean Logarithmic Mass of Cosmic Rays from 0.3 to 30 PeV with LHAASO-KM2A*, *Physical Review Letters* **132** (2024) 131002.

[3] H. He and For the LHAASO Collaboration, *Design of the LHAASO detectors*, *Radiation Detection Technology and Methods* **2** (2018) 7.

[4] W.D. Apel, J.C. Arteaga-Velázquez, K. Bekk, M. Bertaina, J. Blümer, H. Bozdog et al., *Kneelike Structure in the Spectrum of the Heavy Component of Cosmic Rays Observed with KASCADE-Grande*, *Physical Review Letters* **107** (2011) 171104.

[5] Pierre Auger Collaboration, A. Abdul Halim, P. Abreu, M. Aglietta, I. Allekotte, K. Almeida Cheminant et al., *Inference of the mass composition of cosmic rays with energies from* $10^{18.5}$ *to* 10^{20} *eV using the pierre auger observatory and deep learning*, *Physical Review Letters* **134** (2025) 021001.

[6] N. Choma, F. Monti, L. Gerhardt, T. Palczewski, Z. Ronaghi, P. Prabhat et al., *Graph Neural Networks for IceCube Signal Classification: 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018* (2019) 386.

[7] C. Jin, S.-z. Chen, H.-h. He and f.t.L. Collaboration), *Classifying cosmic-ray proton and light groups in LHAASO-KM2A experiment with graph neural network **, *Chinese Physics C* **44** (2020) 065002.

[8] H. Qu and L. Gouskos, *Jet tagging via particle clouds*, *Physical Review D* **101** (2020) 056019.

[9] Z. Cao, F. Aharonian, Q. An, Axikegu, Y.X. Bai, Y.W. Bao et al., *LHAASO-KM2A detector simulation using Geant4*, *Radiation Detection Technology and Methods* (2024) .

[10] F. Aharonian, Q. An, . Axikegu, L.X. Bai, Y.X. Bai, Y.W. Bao et al., *Observation of the Crab Nebula with LHAASO-KM2A - a performance study*, *Chinese Physics C* **45** (2021) 025002.

[11] Y. Wang, Y. Sun, Z. Liu, S.E. Sarma, M.M. Bronstein and J.M. Solomon, *Dynamic Graph CNN for Learning on Point Clouds*, *ACM Transactions on Graphics* **38** (2019) 146:1.

[12] J.R. Hörandel, *On the knee in the energy spectrum of cosmic rays*, *Astroparticle Physics* **19** (2003) 193.