PoS PROCEEDINGS OF SCIENCE

Machine learning applications on event reconstruction and identification for the Tibet ASgamma experiment

Kongyi Hu,[a,b,*] **Jing Huang,**[a] **Ding Chen,**[c] **Ying Zhang,**[a] **LiuMing Zhai,**[c] **Xu Chen,**[c] **Yu Meng,**[a,b] **Yihuan Zou**[a,b] **and Yanlin Yu**[a,b]

[a] *Key Laboratory of Particle Astrophysics, Institute of High Energy Physics, Chinese Academy of Sciences,*
Beijing 100049, China

[b] *University of Chinese Academy of Sciences,*
Beijing 100049, China

[c] *National Astronomical Observatories, Chinese Academy of Sciences,*
Beijing 100012, China

E-mail: hukongyi@ihep.ac.cn

In this paper, we present a cutting-edge approach that combines Graph Neural Networks (GNNs) with AutoML for reconstructing ground-based cosmic ray (CR) observational data. Our novel method accurately estimates primary cosmic ray energy and enhances proton/gamma identification. We have achieved convincing results by using full Monte Carlo simulation data to simulate the Tibet ASgamma experiment (Tibet III+MD). By utilizing the powerful functions of AutoML and GNNs, our integrated approach achieves a significant improvement of 31% energy resolution in data reconstruction above 100 TeV, surpassing the performance of traditional methods in reconstructing the primary energy and arriving direction of the particles. Additionally, our method effectively reduces the cosmic ray background by 30% compared to traditional methods, while preserving the crucial gamma events. The outstanding accuracy of our GNN-based energy reconstruction is further amplified through AutoML, which enables the assimilation of critical information, such as air shower size, secondary cosmic ray lateral distributions, density distributions on the detector, core position, zenith angle distributions, and more. Beyond cosmic ray observation, our versatile machine learning approach holds promise for tackling a wide range of particle physics and astrophysics challenges, making substantial contributions to these fields and paving the way for exciting future advancements.

38th International Cosmic Ray Conference (ICRC2023)
26 July - 3 August, 2023
Nagoya, Japan

ICRC2023 38th International Cosmic Ray Conference The Astroparticle Physics Conference

[*] Speaker

1. Introduction

The Tibet ASgamma experiment, positioned at 4300 meters altitude in Yangbajing, Tibet, China, covering a total area of 65,700 m^2 [1]. It consists of three sub-arrays: Tibet air-shower array (Tibet-III), air-shower-core detector-grid (YAC-II), and underwater Cherenkov muon detector array (MD) covering a total area of 3,400 m^2 [2, 3].

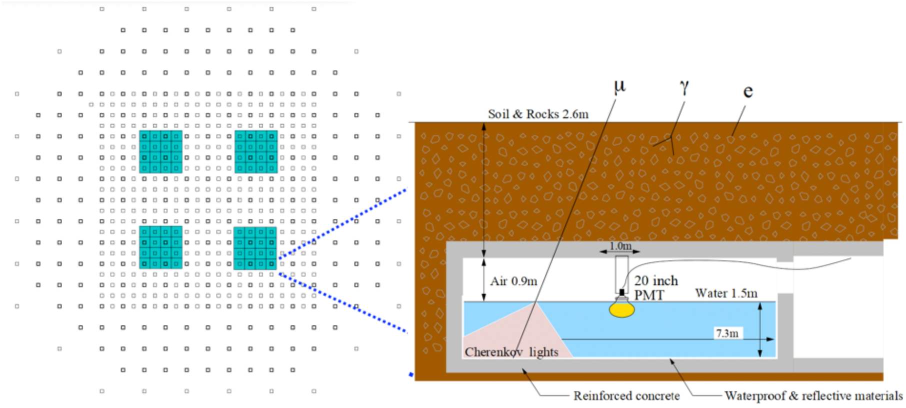

Figure 1: Schematic view of (Tibet-III+MD) array(left) and a MD detector structure(right).

Machine learning (ML) has transformed particle physics, aiding data collection, physics object reconstruction, identification, and new physics searches [4]. Traditional ML methods depended on manually derived high-level features and algorithms like decision trees, support vector machines, and shallow neural networks. However, advancements, particularly in deep neural networks like CNNs, RNNs, and GNNs, leverage raw detector data directly [5], bypassing laborious feature extraction and yielding superior outcomes.

Graph neural networks, especially, have made significant strides, finding applications across domains such as recommendation systems, medical biology, risk control, and optimization.

With a profusion of ML techniques, selecting the right methods and optimal hyperparameters can be time-intensive. AutoML emerges as a robust, efficient solution for this challenge, ensuring fault tolerance.

The hexagonal detector configuration of the Tibet ASgamma experiment introduces internal and external variations, rendering direct matrix representations challenging. Furthermore, data translation symmetry is imperfect, hindering the performance of convolutional neural networks. However, graph neural networks excel in non-Euclidean data spaces. This article proposes utilizing GNNs for feature extraction and combining their outcomes with traditionally derived features. Event reconstruction and identification will be achieved through the autoML approach.

2. Graph Neural Network

Graph Neural Networks (GNNs) excel at capturing intricate relationships in diverse datasets, such as social networks, maps, and knowledge graphs. In the realm of particle physics, graph representations offer advantages over traditional matrices, adeptly handling variable-sized data without unnecessary zero-padding. These graphs effectively manage sparse, heterogeneous detector data that may not seamlessly translate into images[6].

Formally, a graph $G = (u, V, E)$ is defined with N_v vertices and N_e edges. u denotes overall graph features, $V = v_i$ comprises node sets (v_i represents the i-th node's features), and $E = e_{ij}$ signifies edges (e_{ij} holds edge features between the i-th and j-th nodes).

For graph neural networks, the computation in the $(l+1)^{th}$ iteration of graph $G = (u^{l+1}, V^{l+1}, E^{l+1})$ is as follows:

Edge feature update: $e^{l+1}ij = \phi^e(v^l_i, v^l_j, eij^l)$, where ϕ^e aggregates information from adjacent nodes via edges. Node feature update: $v^{l'}i = \rho(e^{l+1}_{ij})$ for all $j \in N_i$, with ρ processing aggregated edge features. Global graph feature update: $u^{l+1} = \phi^v(v^{l'}_i, v^l_i, u^l)$, as ϕ^v updates node and global graph features. Choice of ϕ^e, ϕ^v, and ρ yields varied GNN structures, accommodating diverse patterns and data dependencies. These operations iteratively equip GNNs to learn and represent complex relationships within graph-structured data, effectively addressing tasks in particle physics and beyond[6–8].

3. Automated Machine Learning

Automated Machine Learning (AutoML) simplifies model selection, configuration, and optimization, streamlining the machine learning process. Unlike traditional methods that demand expert intervention, AutoML empowers non-experts to harness machine learning proficiently[9].

AutoML aims to automate common tasks like data preprocessing, feature engineering, model selection, hyperparameter tuning, and ensemble creation. By utilizing AutoML tools, manual effort is minimized, expediting the development of high-performing models. These tools automatically identify suitable models, fine-tune hyperparameters, and enhance performance through model fusion[10, 11].

4. Monte Carlo simulation

The extensive air showers (EAS) development in the atmosphere and the response in the Tibet hybrid experiment array have been comprehensively studied using full Monte Carlo (MC) simulation. The widely-used simulation code, CORSIKA [12], is employed to generate both gamma events and cosmic ray events. And all detector responses are calculated using Geant4[13].

For the gamma events, the primary particle's energy ranges from 300 GeV to 100 PeV, with a spectral index of -2.0. In total, 10^9 gamma events are generated to capture a broad range of energy levels.

Regarding the primary cosmic-ray composition model, the model spectrum proposed by M. Shibata et al. [14] is adopted to determine their chemical composition and energy spectrum. The low-energy hadronic interactions are simulated using FLUKA [15], while the high-energy interactions are modeled using EPOS LHC [16]. A significant number of 4×10^9 cosmic ray events are generated to ensure robust and statistically significant results.

Table 1: Parameters used in the CORSIKA air shower simulation

Primary type	Spectral	Energy range(TeV)	Events
Gamma rays	Power law with index -2	0.3 – 10^5	10^9
Cosmic rays	M. Shibata et al.[14]	0.3 – 10^5	4×10^9

5. Method

The cut condition employed closely follows the approach used in the Crab study [17], with the omission of N_μ cut condition and the removal of age cut condition to increase the dataset.The traditional method refers to the method in ref[17, 18].

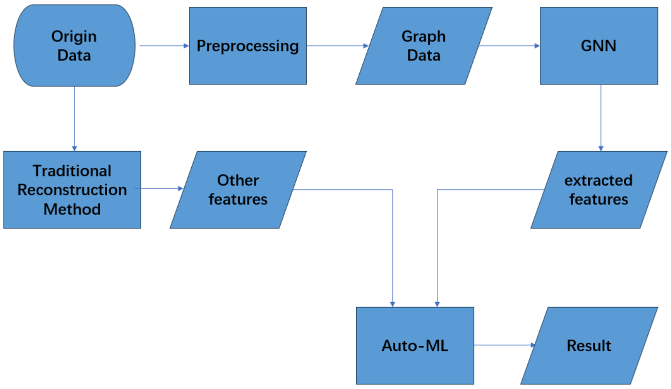

Figure 2: Algorithm flowchart

The algorithm flowchart for our method showed in Figure.2. First, the Monte Carlo simulation data is transformed into graph-structured data based on the relative positions of the detectors. The detector data are converted into graph data, and a GNN is utilized for feature extraction from the graph data. In the case of data reconstruction, two additional fully connected layers are added after the pooling layer of GNN. The extracted features are combined with the features obtained from the Monte Carlo simulation data using traditional method. Subsequently, the combined features are fed into an autoML system to obtain the reconstruction or discrimination results.

6. Result

6.1 Event Reconstruction

For reconstructing (Tibet III+MD) raw experimental data, it is discussed in two parts, angle-resolved and energy-resolved. The angle resolution and energy resolution obtained in this work are compared with the traditional method at different energies.

Comparing the angular resolution obtained by this work with the traditional method, as shown in the figure.3, the angular resolution obtained by this work is better than the traditional method in each energy range.

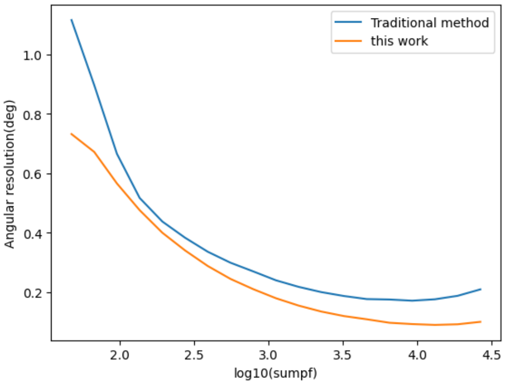

Figure 3: Comparison of arrival directions between traditional method and this work

For the reconstruction of energy, it can be seen in the figure.4 that compared with the traditional method, the image obtained by the reconstruction of this work is narrower, and as can be seen in the table.2, the energy resolution we obtained is at different energies and different zenith angles are superior to traditional methods. And we achieve a 31% improvement in energy resolution for event about 100 TeV in all zenith angle, compared to traditional methods.

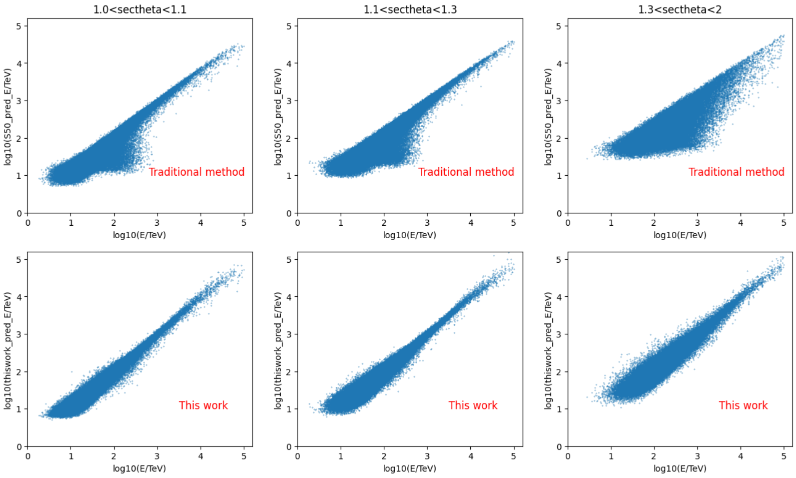

Figure 4: Comparison of energy fitting results between traditional method and this work.

Table 2: Comparison of energy resolution between traditional method and this work

Energy	about 50 TeV			about 100 TeV		
sec(theta)	1.0-1.1	1.1-1.3	1.3-2.0	1.0-1.1	1.1-1.3	1.3-2.0
Traditional method	0.33	0.47	0.87	0.20	0.31	0.72
This work	0.22	0.30	0.44	0.17	0.23	0.39

6.2 Gamma/Proton Identification

For different energy ranges, we define Signal-to-Noise Ratio (SNR) and use the results from this work and traditional methods to calculate the optimal cut conditions that maximize SNR. Then, we compare the SNR values across all energy ranges. Gamma/Proton identification of this work is better than our traditional methods at each energy bin, especially for high energy ranges.At about 100 TeV, this work achieves a third higher SNR than conventional methods, roughly reducing the cosmic ray background by 30% imporvement compare to traditional methods, while preserving the same gamma events.

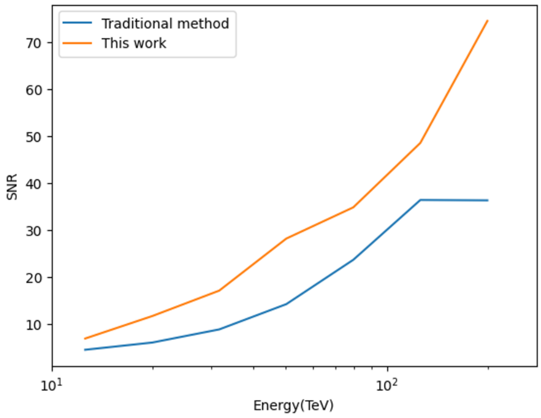

Figure 5: Comparison of SNR between traditional method and this work.

7. Discussion

The integration of Graph Neural Networks (GNNs) and Automated Machine Learning (AutoML) holds remarkable potential for enhancing event reconstruction and identification in the Tibet ASgamma experiment. This innovative fusion promises significant improvements in processing cosmic ray (CR) observational data.

By harnessing the capabilities of GNNs and AutoML, our combined approach achieves a substantial 31% enhancement compare to traditional methods in energy resolution for data reconstruction beyond 100 TeV. This outperforms conventional techniques in primary energy reconstruction and particle arrival direction estimation. Notably, our method reduces cosmic ray background by 30% while effectively preserving essential gamma events compare to traditional methods. The success owes much to GNN-based energy reconstruction, further enhanced by AutoML's ability to assimilate critical details like air shower size, secondary cosmic ray distributions, detector density patterns, core position, zenith angles, and more.

GNNs excel in handling complex relationships within graph-structured data, as illustrated by the hexagonal detector configuration of the Tibet ASgamma experiment. This sets them apart from traditional matrix representations and bridges gaps left by convolutional neural networks.

Moreover, AutoML emerges as a robust tool to navigate the array of machine learning techniques available. It simplifies model selection, configuration, and optimization, making machine learning more accessible and efficient for non-experts. Automation streamlines data preprocessing, feature engineering, model selection, hyperparameter tuning, and ensemble creation, reducing manual intervention and accelerating model development.

The success achieved in the Tibet ASgamma experiment extends beyond, offering insights for particle physics and astrophysics. The combined power of GNNs and AutoML demonstrated here lays the groundwork for future advances in a variety of research domains, showcasing their potential for addressing the challenges of event reconstruction and identification.

Acknowledgements

The collaborative experiment of the Tibet Air Shower Arrays has been conducted under the auspices of the Ministry of Science and Technology of China and the Ministry of Foreign Affairs of Japan. This work was supported in part by a Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Culture, Sports, Science and Technology, and by Grants-in-Aid for Science Research from the Japan Society for the Promotion of Science in Japan. This work is supported by the National Natural Science Foundation of China under Grants No. 12227804, No. 12275282, No. 12103056 and No. 12073050, and the Key Laboratory of Particle Astrophysics, Institute of High Energy Physics, CAS. This work is also supported by the joint research program of the Institute for Cosmic Ray Research (ICRR), the University of Tokyo.

References

[1] M Amenomori, XJ Bi, D Chen, SW Cui, LK Ding, XH Ding, C Fan, CF Feng, Zhaoyang Feng, ZY Feng, et al. Multi-tev gamma-ray observation from the crab nebula using the tibet-iii air shower array finely tuned by the cosmic ray moon's shadow. *The Astrophysical Journal*, 692(1):61, 2009.

[2] M Amenomori, XJ Bi, D Chen, TL Chen, WY Chen, SW Cui, LK Ding, CF Feng, Zhaoyang Feng, ZY Feng, et al. Search for gamma rays above 100 tev from the crab nebula with the tibet air shower array and the 100 m2 muon detector. *The Astrophysical Journal*, 813(2):98, 2015.

[3] J Huang, LM Zhai, D Chen, M Shibata, Y Katayose, Ying Zhang, JS Liu, Xu Chen, XB Hu, XY Zhang, et al. Performance of the tibet hybrid experiment (yac-ii+ tibet-iii+ md) to measure the energy spectra of the light primary cosmic rays at energies 50–10,000 tev. *Astroparticle Physics*, 66:18–30, 2015.

[4] Matthew Feickert and Benjamin Nachman. A Living Review of Machine Learning for Particle Physics. *arXiv e-prints*, page arXiv:2102.02770, February 2021.

[5] Savannah Thais, Paolo Calafiura, Grigorios Chachamis, Gage DeZoort, Javier Duarte, Sanmay Ganguly, Michael Kagan, Daniel Murnane, Mark S. Neubauer, and Kazuhiro Terao. Graph Neural Networks in Particle Physics: Implementations, Innovations, and Challenges. *arXiv e-prints*, page arXiv:2203.12852, March 2022.

[6] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81, 2020.

[7] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[8] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1):4–24, 2020.

[9] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[10] Jonathan Waring, Charlotta Lindvall, and Renato Umeton. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial intelligence in medicine*, 104:101822, 2020.

[11] Xin He, Kaiyong Zhao, and Xiaowen Chu. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622, 2021.

[12] Dieter Heck, Johannes Knapp, JN Capdevielle, G Schatz, T Thouw, et al. Corsika: A monte carlo code to simulate extensive air showers. *Report fzka*, 6019(11), 1998.

[13] S. Agostinelli, J. Allison, K. Amako, J. Apostolakis, H. Araujo, P. Arce, , et al. Geant4—a simulation toolkit. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250–303, 2003.

[14] M. Shibata, Y. Katayose, J. Huang, and D. Chen. CHEMICAL COMPOSITION AND MAXIMUM ENERGY OF GALACTIC COSMIC RAYS. *The Astrophysical Journal*, 716(2):1076–1083, June 2010.

[15] Giuseppe Battistoni, Till Boehlen, Francesco Cerutti, Pik Wai Chin, Luigi Salvatore Esposito, Alberto Fassò, Alfredo Ferrari, Anton Lechner, Anton Empl, Andrea Mairani, et al. Overview of the fluka code. *Annals of Nuclear Energy*, 82:10–18, 2015.

[16] T Pierog, Iu Karpenko, Judith Maria Katzy, E Yatsenko, and Klaus Werner. Epos lhc: Test of collective hadronization with data measured at the cern large hadron collider. *Physical Review C*, 92(3):034906, 2015.

[17] M Amenomori, YW Bao, XJ Bi, D Chen, TL Chen, WY Chen, Xu Chen, Y Chen, SW Cui, LK Ding, et al. First detection of photons with energy beyond 100 tev from an astrophysical source. *Physical review letters*, 123(5):051101, 2019.

[18] K Kawata, TK Sako, M Ohnishi, M Takita, Y Nakamura, and K Munakata. Energy determination of gamma-ray induced air showers observed by an extensive air shower array. *Experimental Astronomy*, 44:1–9, 2017.