PROCEEDINGS OF SCIENCE

.12Schematic view of (Tibet-III+MD) array(left) and a MD detector structure(right).figure.caption.1

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] **Xu Chen,**[c] **LiuMing Zhai,**[c] **Yu Meng,**[a,b] **Yihuang 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 P/gamma identification. Leveraging Full Monte Carlo simulations, emulating the Tibet ASgamma experiment (Tibet-III + MD), we achieve compelling results. By harnessing the power of AutoML and GNNs, our integrated approach achieves a remarkable 31% enhancement in energy resolution for reconstructed cases above 100 TeV, surpassing the performance of S50 reconstruction. Additionally, our method effectively reduces the cosmic ray background by 30%, 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

PoS(ICRC2023)491

1. Introduction

The Tibet ASgamma experiment, situated at an altitude of 4300 meters in Yangbajing, Tibet, China, covers an area of 65,700 m^2 [1]. Comprising three sub-arrays, namely, the Tibet air-shower array (Tibet-III), air-shower-core detector-grid (YAC-II), and underwater Cherenkov muon detector array (MD) extending over 3,400 m^2 [2, 3], this study focuses on the application of Graph Neural Networks (GNN) and automated machine learning (autoML) trained on simulated data from Tibet-III and MD array.

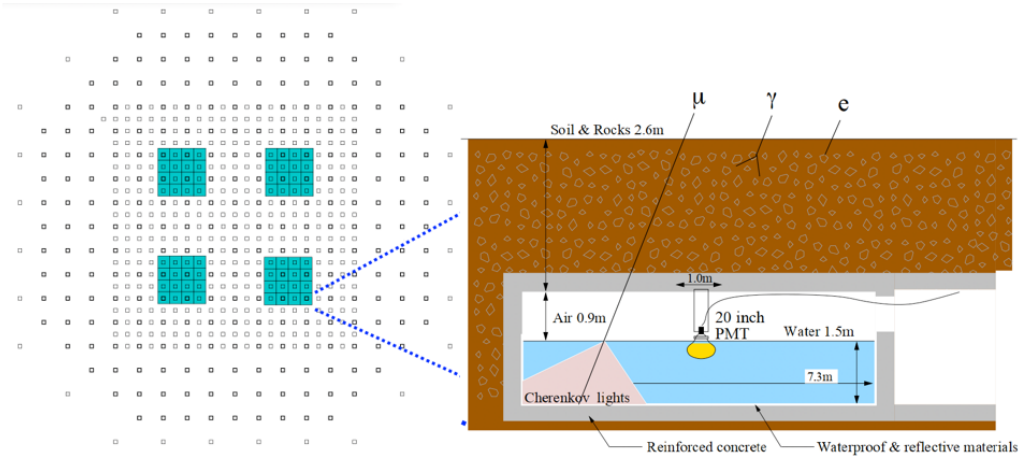

Figure 1: Schematic view of (Tibet-III+MD) array(left) and a MD detector structure(right).

Machine learning (ML) has proven invaluable in particle physics, enabling data collection, physics object reconstruction, identification, and new physics searches [4]. Traditionally, ML methods relied on manually extracted high-level features, employing algorithms like decision trees, support vector machines, and shallow neural networks for regression or classification. However, the landscape has evolved, with deep neural networks, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Graph Neural Networks (GNNs), capable of leveraging low-level features directly obtained from detectors [5]. This shift eliminates the need for laborious manual feature extraction and yields superior results.

Among these advancements, graph neural networks have seen remarkable progress, finding applications in diverse fields like recommendation systems, medical biology, risk control, and combination optimization.

Given the abundance of ML methods, selecting suitable techniques and searching for optimal hyperparameters can become time-consuming. In this context, automated machine learning (autoML) offers a straightforward, robust, and efficient solution, ensuring high fault tolerance.

The Tibet ASgamma experiment's detector arrangement is hexagonal, resulting in different internal and external intervals, making it challenging to directly employ matrix representations. Additionally, the data translation symmetry is not ideal, leading to subpar performance when using convolutional neural networks. However, graph neural networks excel in handling data in non-Euclidean spaces. This article proposes utilizing a graph neural network for feature extraction and

integrating the results with traditionally extracted features. The subsequent event reconstruction and identification will be accomplished through the application of the autoML approach.

2. Graph Neural Network

Graph Neural Networks (GNNs) are powerful tools for representing complex relationships found in various datasets, including social networks, maps, knowledge graphs, and more. In the realm of particle physics, leveraging graphs to represent data offers distinct advantages over traditional table or matrix structures. These advantages include the ability to handle variable-sized data without the need to fill vacant positions with zeros and effectively manage sparse, heterogeneous detector data that may not be easily projected into image representations[6].

Formally, we define a graph $G = (u, V, E)$ with N_v vertices and N_e edges. Here, u represents the global graph features, while $V = v_i$ constitutes the set of nodes, each denoted by v_i, capturing the features of the i-th node. Correspondingly, $E = e_{ij}$ forms the set of edges, with e_{ij} representing the features associated with the edge connecting the i-th and j-th nodes.

In the context of graph neural networks, the computation for the $(l+1)^{th}$ iteration of the graph $G = (u^{l+1}, V^{l+1}, E^{l+1})$ can be outlined as follows:

- Update edge features: $e_{ij}^{l+1} = \phi^e(v_i^l, v_j^l, e_{ij}^l)$, where ϕ^e is a function that aggregates information from adjacent nodes via edges.
- Update node features: $v^{l'}i = \rho(e_{ij}^{l+1})$ for all $j \in N_i$, where ρ is a function that processes the aggregated edge features.
- Update global graph features: $u^{l+1} = \phi^v(v_i^{l'}, v_i^l, u^l)$, where ϕ^v is a function that updates node features and the global graph features.

The choice of functions ϕ^e, ϕ^v, and ρ allows for various graph neural network structures, enabling flexibility in capturing different patterns and dependencies within the data. By iteratively applying these operations, GNNs can effectively learn and represent complex relationships within graph-structured data, making them well-suited for tasks in particle physics and beyond[6–8].

3. Automated Machine Learning

Automated Machine Learning (AutoML) is a methodology designed to simplify the process of selecting, configuring, and optimizing machine learning models. Traditional machine learning requires skilled data scientists or machine learning experts to conduct extensive manual work in model selection and hyperparameter tuning. However, with the rapid advancement of machine learning, AutoML has emerged to offer a convenient pathway for non-experts to leverage machine learning techniques effectively[9].

The primary goal of AutoML is to automate various common machine learning tasks, including data preprocessing, feature engineering, model selection, hyperparameter optimization, and model ensemble. By employing AutoML tools and techniques, users can significantly reduce manual intervention and quickly build high-performing machine learning models. AutoML algorithms

automatically choose appropriate model types, optimize hyperparameters, and further enhance model performance through model fusion techniques[10, 11].

Several popular AutoML tools and platforms include Google's AutoML[12], Microsoft's Azure AutoML[13], Auto-Sklearn[14], H2O.ai[15], autogluon[16], among others. They provide users with automated model selection and optimization capabilities, making machine learning more accessible and user-friendly.

4. Data

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 [17], is employed to generate both gamma events and cosmic ray events.

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 cosmic rays, the model spectrum proposed by M. Shibata et al. [18] is adopted to determine their chemical composition and energy spectrum. The low-energy hadronic interactions are simulated using FLUKA [19], while the high-energy interactions are modeled using EPOS LHC [20]. A significant number of 4×10^9 cosmic ray events are generated to ensure robust and statistically significant results.

Table 1: Parameters used in thet 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.[18]	$0.3 - 10^5$	4×10^9

5. Method

The cut condition employed closely follows the approach used in the Crab study [21], with the omission of Nmu cut condition and the removal of age cut condition to increase the dataset.The traditional method refers to the method in ref[21, 22].

not yet

6. Result and Discussion

not yet

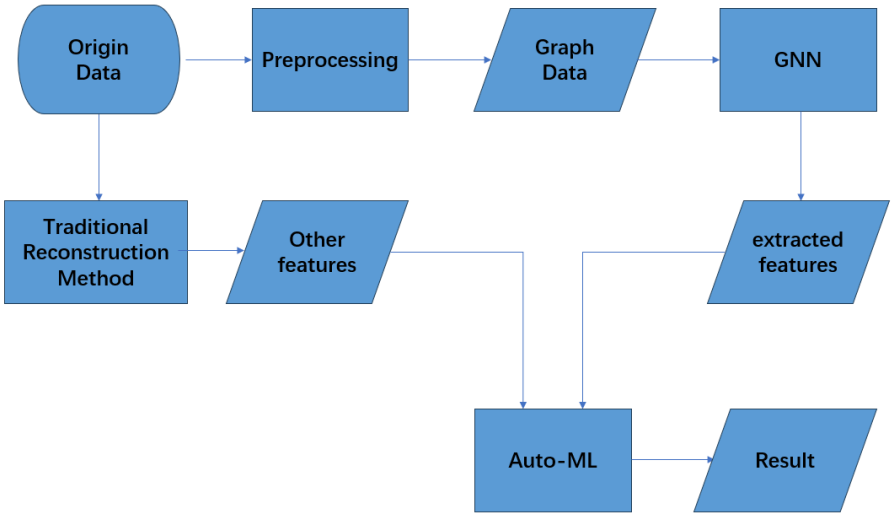

Figure 2: Algorithm flowchart

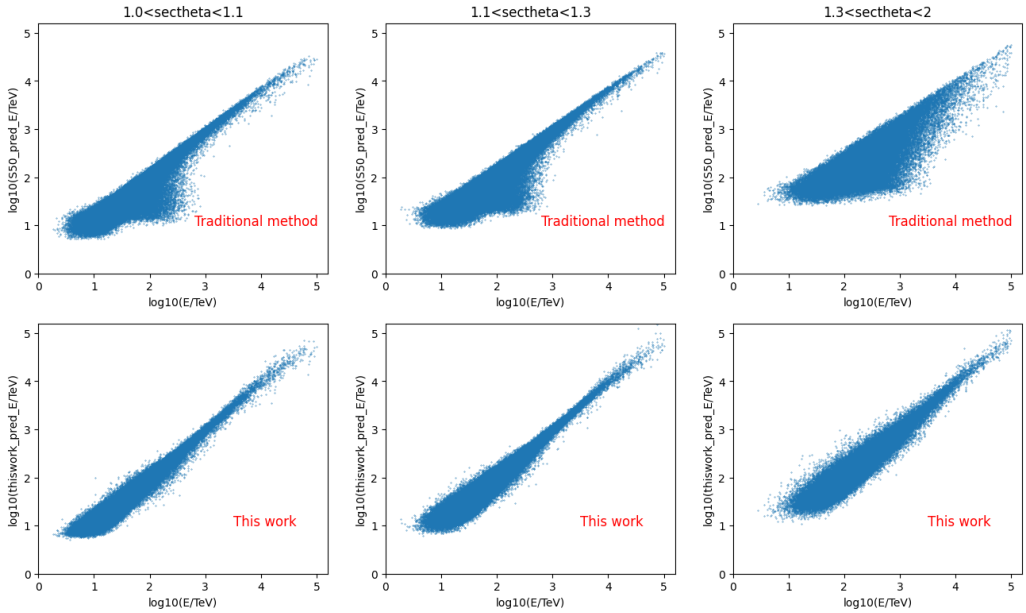

Figure 3: Comparison of energy fitting results between traditional method and this work.

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

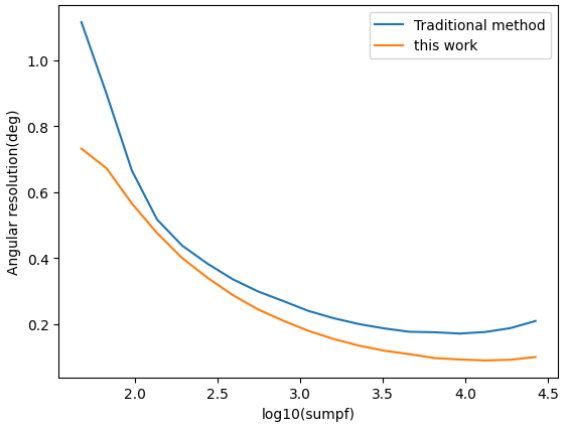

Figure 4: Comparison of arrival directions between traditional method and this work

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] Ekaba Bisong and Ekaba Bisong. Google automl: cloud vision. *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, pages 581–598, 2019.

[13] Deepak Mukunthu, Parashar Shah, and Wee Hyong Tok. *Practical automated machine learning on Azure: using Azure machine learning to quickly build AI solutions*. O'Reilly Media, 2019.

[14] Matthias Feurer, Katharina Eggensperger, Stefan Falkner, Marius Lindauer, and Frank Hutter. Auto-sklearn 2.0: The next generation. *arXiv preprint arXiv:2007.04074*, 24, 2020.

[15] Arno Candel, Viraj Parmar, Erin LeDell, and Anisha Arora. Deep learning with h2o. *H2O. ai Inc*, pages 1–21, 2016.

[16] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020.

[17] 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.

[18] 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.

[19] Alfredo Ferrari, Paola R Sala, Alberto Fasso, and Johannes Ranft. a multi-particle transport code.

[20] 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.

[21] 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.

[22] 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.