

人工智能在LHAASO中的需求、 现状和前景

张笑鹏（高能物理研究所 粒子天体中心）
for LHAASO AI Group

SCE2025
2025.8.26 长春

内容提要

- LHAASO 背景和科学成果介绍
- AI 能够为 LHAASO 带来哪些提升?
- 目前进行的探索和初步结果
- 未来展望

LHAASO实验介绍

阵列布局和探测器

科学数据和算力

主要科学成果

LHAASO基本情况

高海拔宇宙线观测站

(LHAASO, Large High Altitude Air Shower Observatory)
是以宇宙线观测研究为核心的国家重大科技基础设施

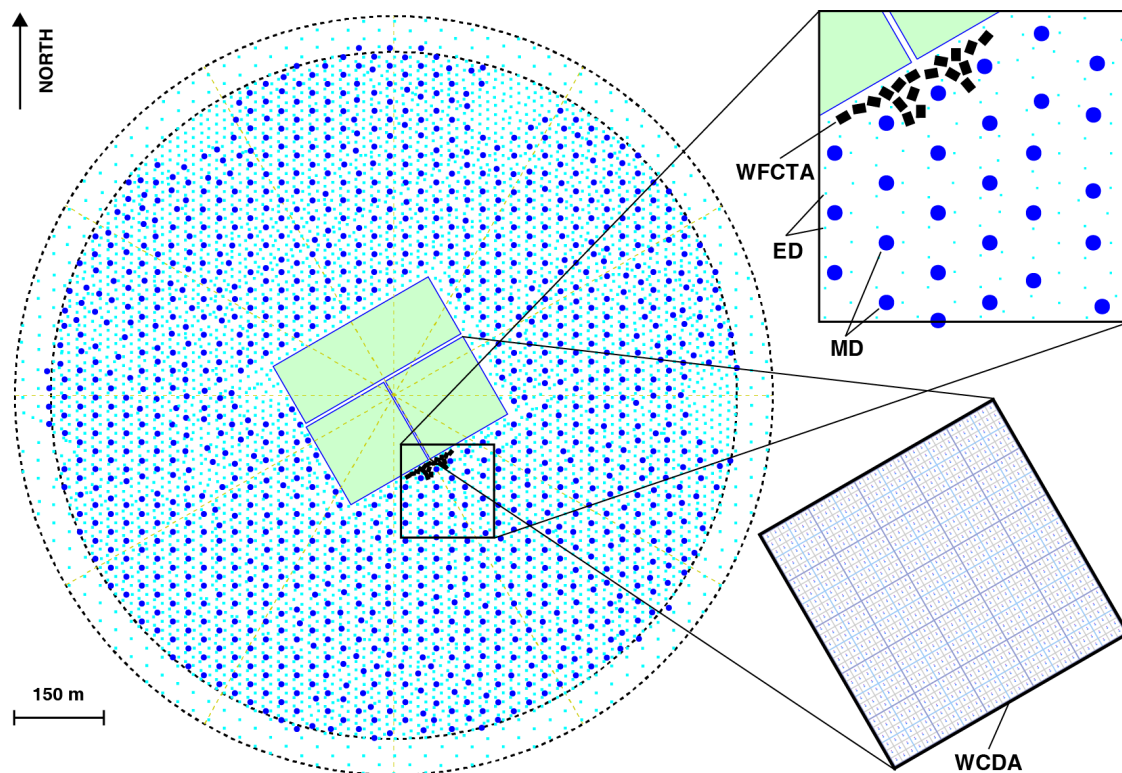
- 灵敏度最高的甚高能伽马巡天望远镜
- 灵敏度最高的超高能伽马天文探测器
- 测量精度最高的膝区宇宙线探测器

位于四川稻城海子山，海拔4410 m
2021年7月 全面建成并投入运行
2023年5月 通过国家验收

科学目标：

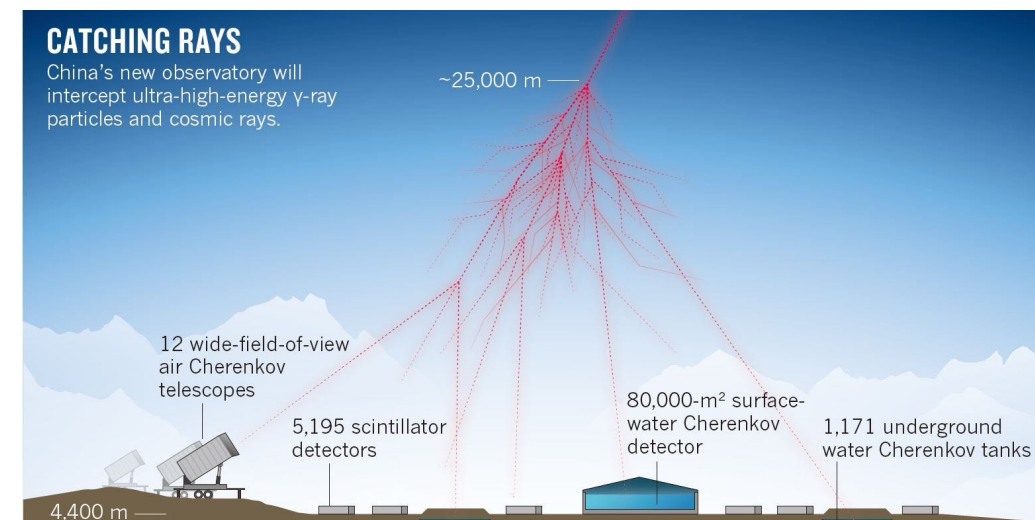
- 寻找宇宙线起源
- 伽马射线源的巡天普查
- 探索新物理前沿

阵列布局

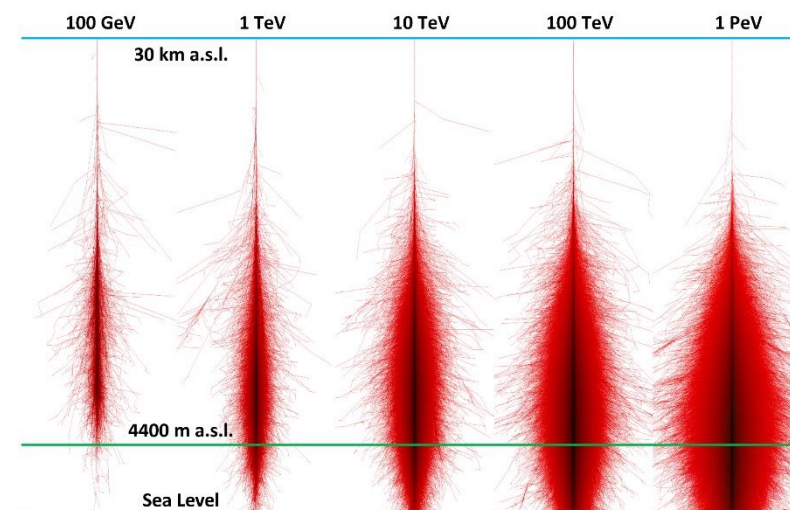


四种类型探测器，构成三大阵列
 占地面积：1.3 km²
 每年产生数据量：~10 PB

探测原理：原初宇宙线和广延大气簇射



不同能量空气簇射的纵向与横向发展 (伽马)



PeV最佳观测高度！

三个子阵列 四种探测器

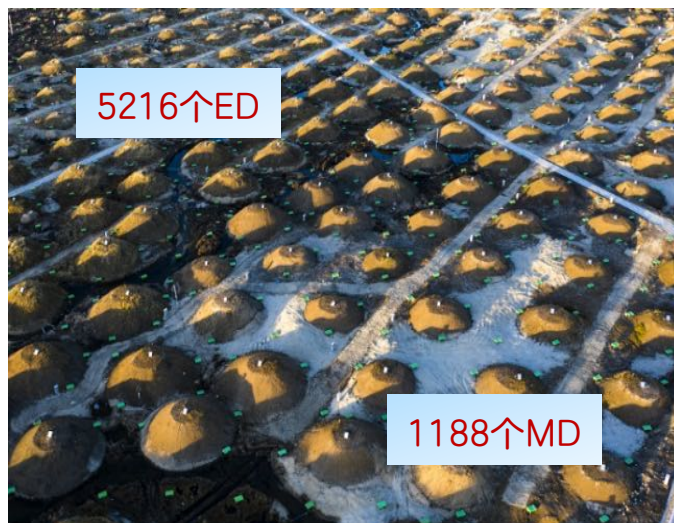
最灵敏的超高能伽马探测装置

地面粒子探测器阵列

KM2A

电磁粒子探测器阵列

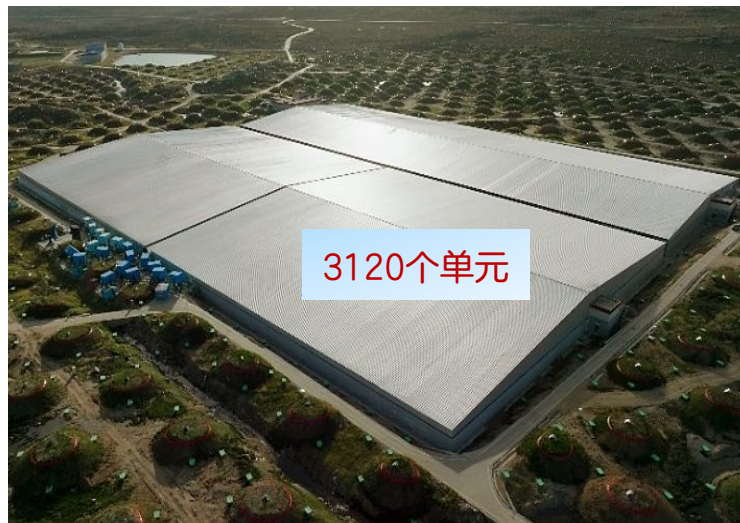
缪子探测器阵列



最灵敏度的甚高能伽马巡天望远镜

水切伦科夫探测器阵列

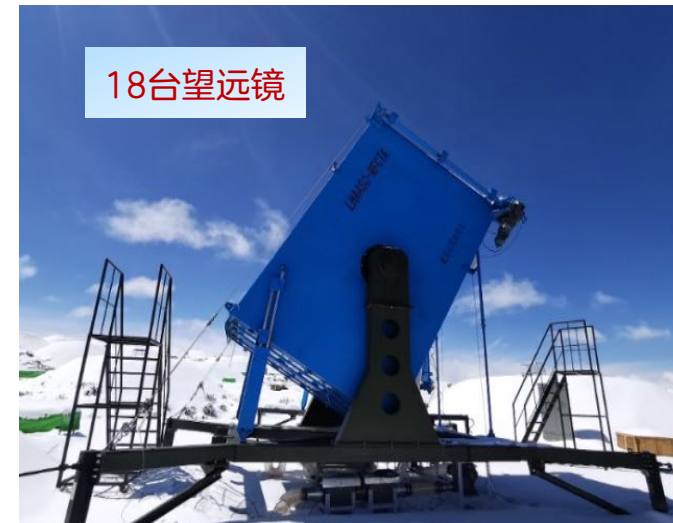
WCDA



能区最宽的高能宇宙线探测装置

广角切伦科夫望远镜阵列

WFCTA



时钟分配系统

数据获取系统

数据处理平台

离线软件系统

通用系统

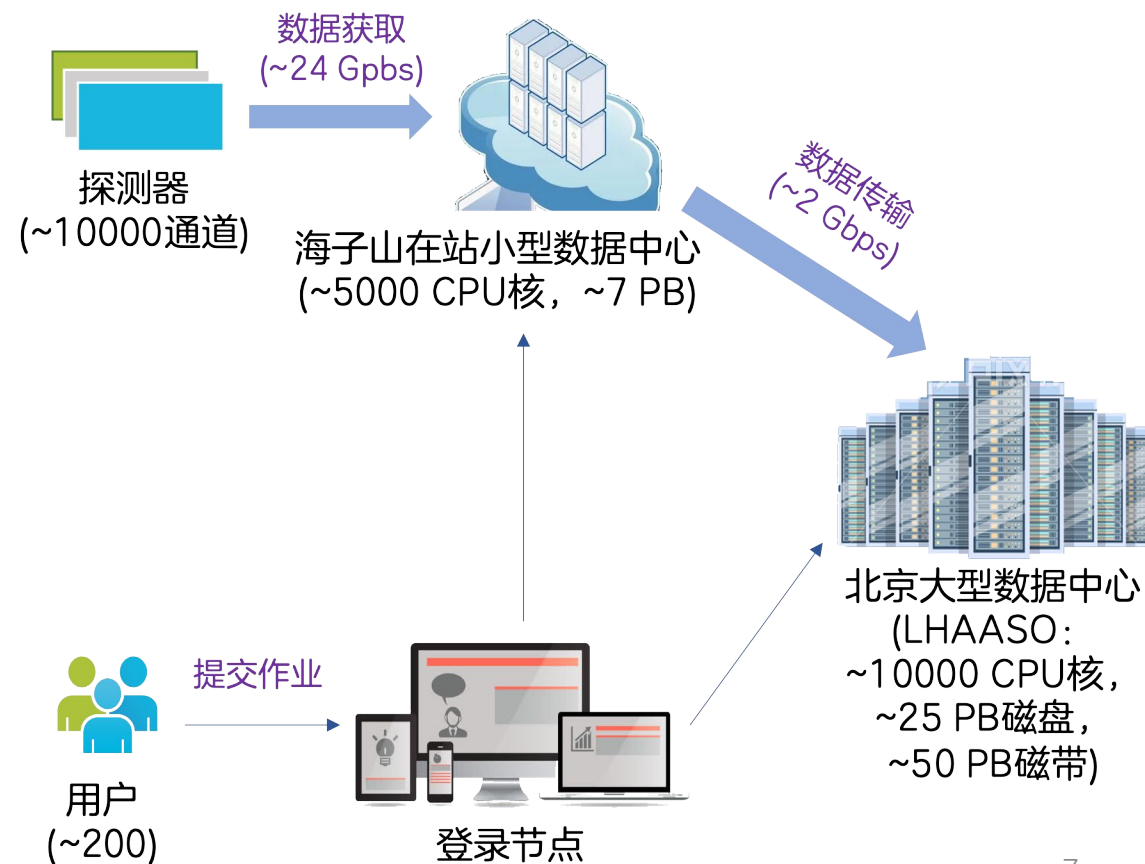
四项技术突破：新型时钟分配系统、超大型光电倍增管、新型硅光电管相机、海量数据获取及数据传输系统

科学数据与算力资源

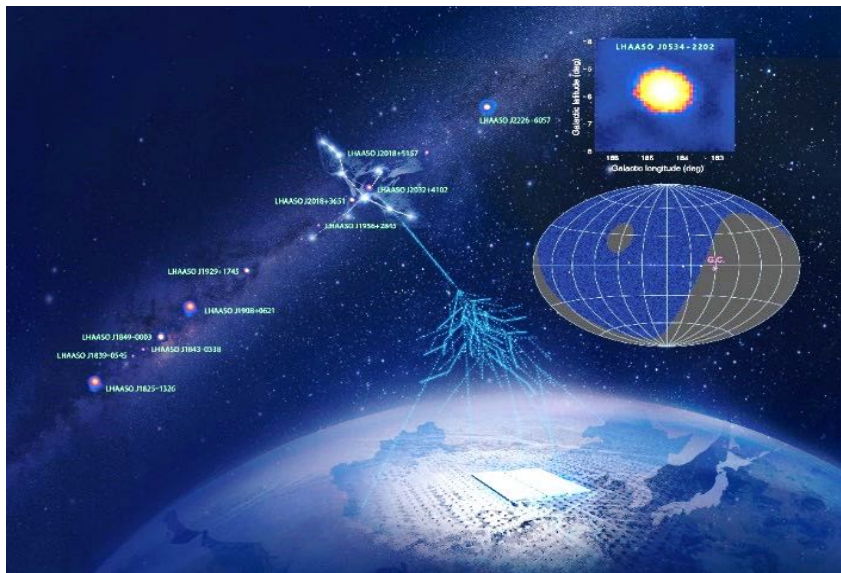
LHAASO每年产生约 **11PB** 科学数据。

高能所计算中心提供数据传输、存储、计算等软硬件支持。

科学数据类别		事例数 (亿)	存储量 (TB)
WCDA	触发	10000	2000
	无触发	1000000 (hit)	1500
	模拟	2	1000
KM2A	触发	1450	800
	模拟	1	300
WFCTA	观测	1	150
	模拟	0.05	1500
总计			7250



LHAASO科学成果

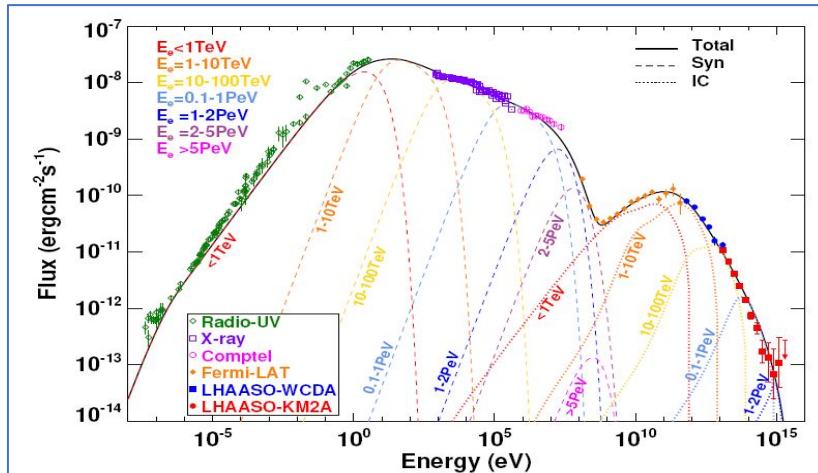
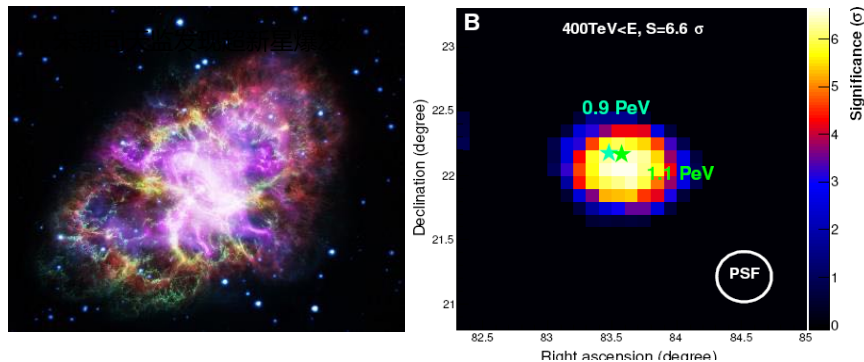


发现首批“拍电子伏加速器”和最高能量光子，
开启“超高能伽马天文学”时代

Nature 594, 33 (2021)

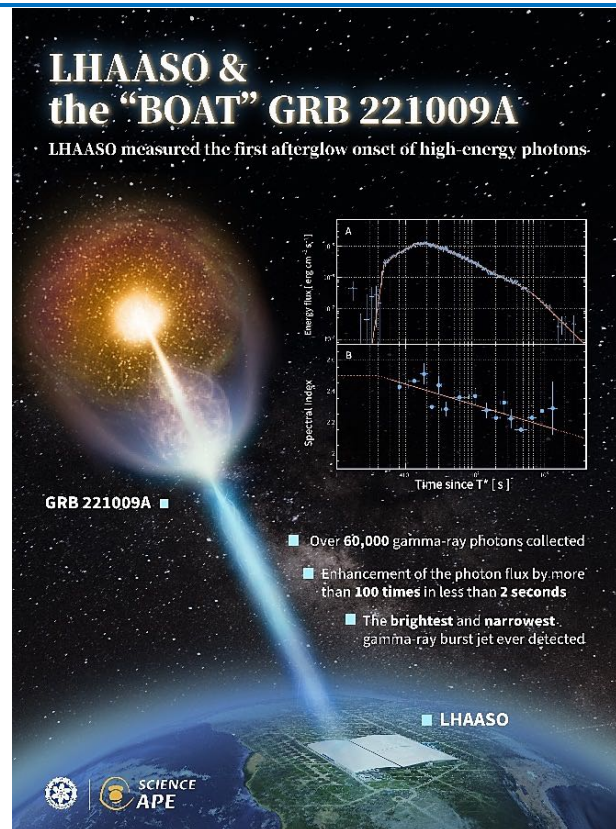
2021年度中国十大科技进展

The LHAASO Collaboration



测定标准烛光蟹状星云的超高能亮度，发
现拍电子伏电子加速器，挑战理论极限

Science 373, 425 (2021)



完整记录大质量恒星死亡瞬间万亿电
子伏特伽马射线暴全过程

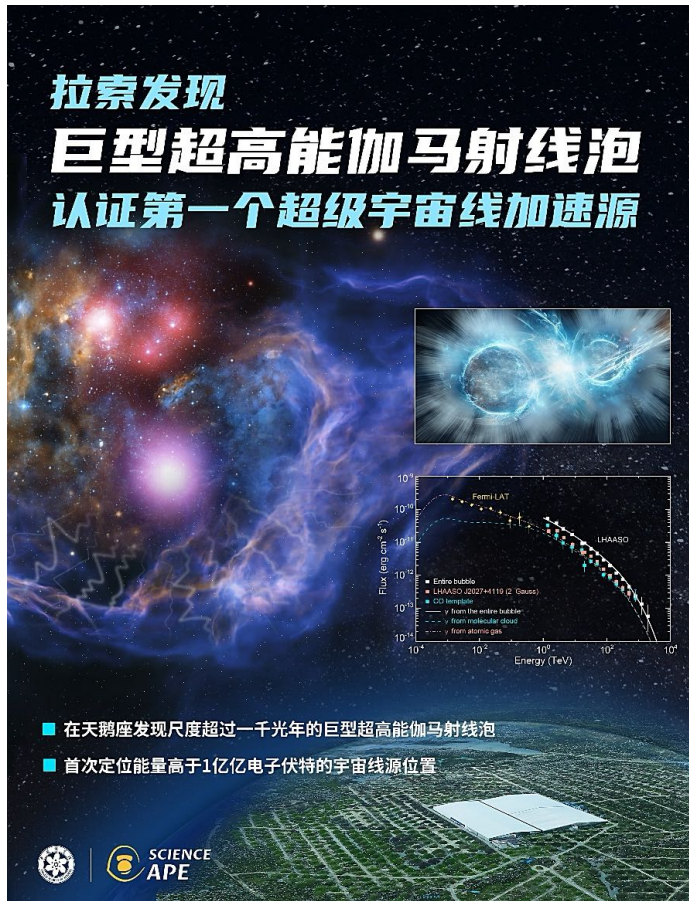
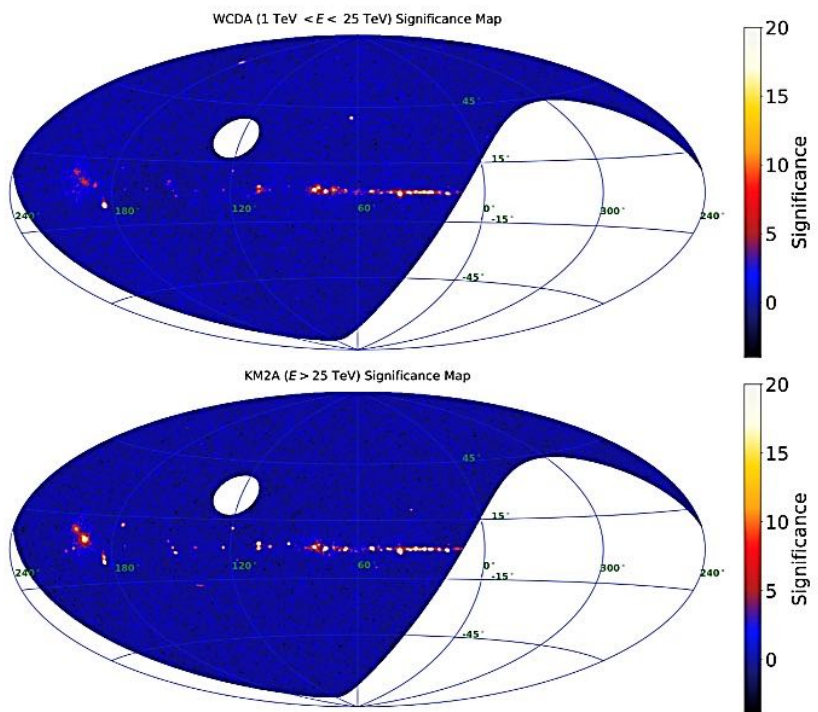
Science 380, 1390 (2023)

2023度中国科学十大进展

精确测量迄今最亮伽马射线暴高能辐
射能谱

Science Advances 9, eadj2778 (2023)

LHAASO科学成果

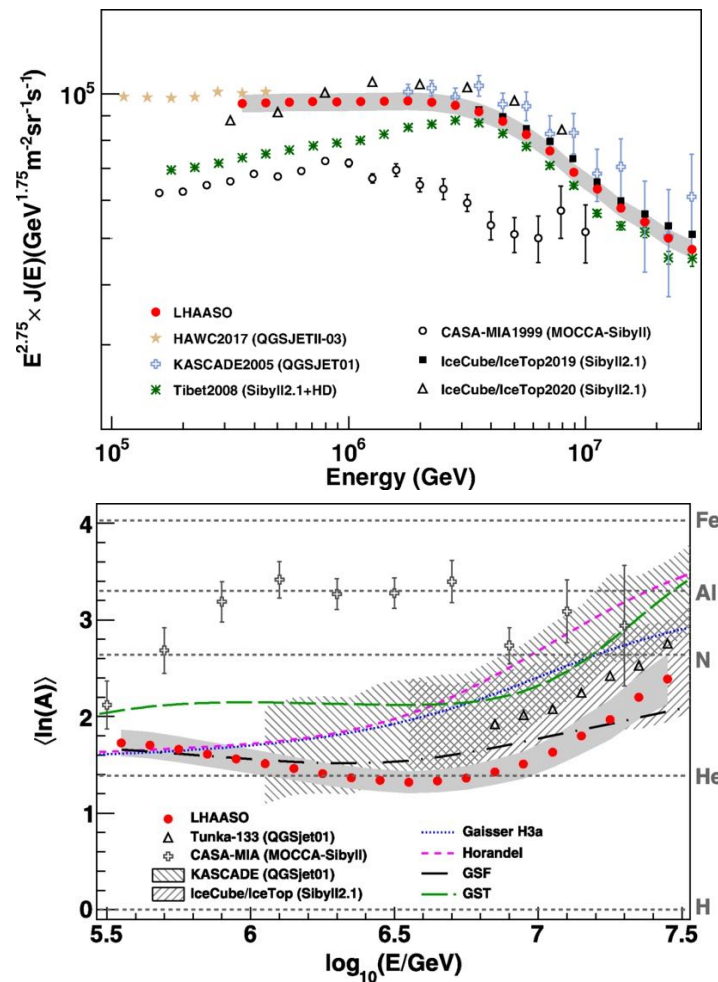


发现巨型超高能伽马射线泡，认证

第一个超级宇宙线加速源

2024度科技日报十大科技新闻

Science Bulletin 69, 449 (2024)



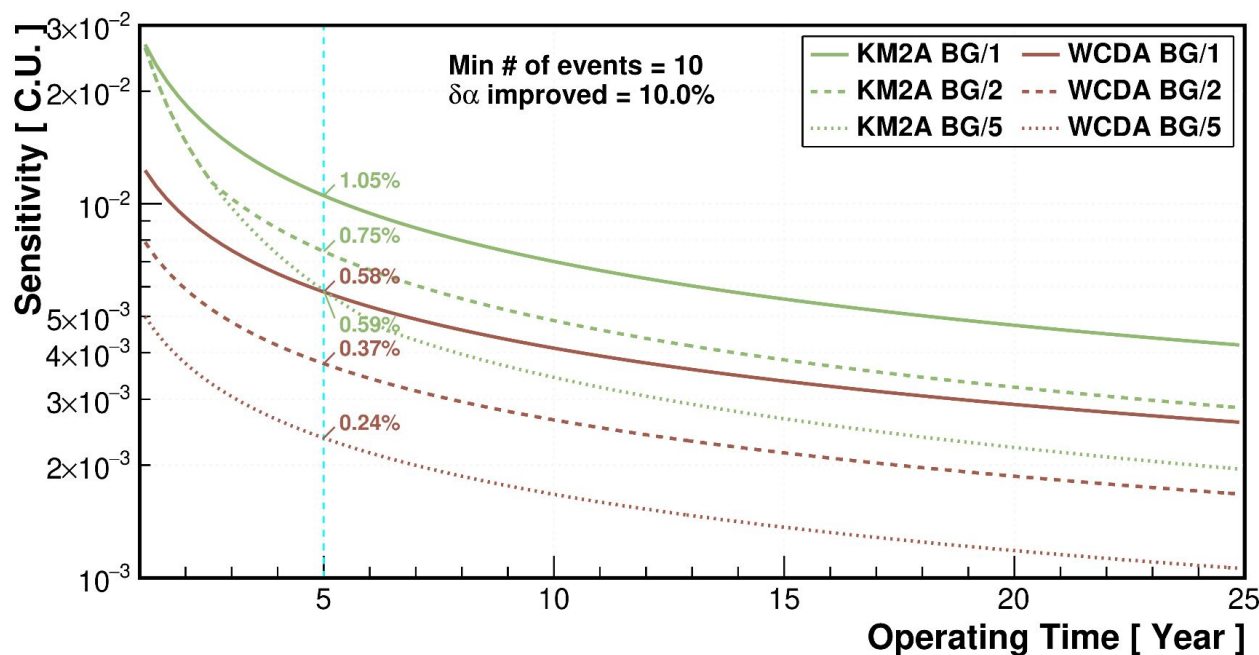
精确测定宇宙线全粒子谱和平均核子数

PRL 132, 131002 (2024)

发布LHAASO第一期源表，包含32个新源

arXiv: 2305.17030; ApJS 271, 1 (2024)

LHAASO对AI的核心需求



- LHAASO积分灵敏度逐渐走缓，迫切需要通过数据处理提高背景排除能力和角分辨
- 基于分布和统计的传统方法已经走到极致
- 与人工智能技术结合将是一个重要的发展方向！

LHAASO- AI 需求

- 背景排除和成分鉴别
- 事例重建
- 其他

LHAASO的AI需求

- 背景事例排除

质子-伽马区分

- 宇宙线成分鉴别

H, He, CNO, MgAlSi, Fe

- 事例重建

芯位、方向、能量

- 天体源分析

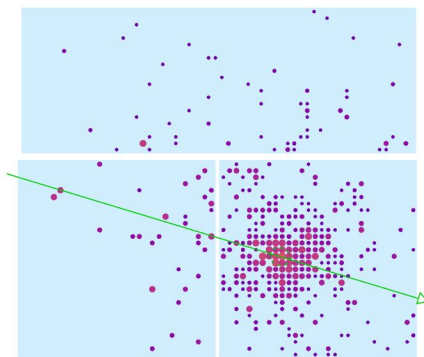
位置、能谱、扩展度、多源解析、多
波段多信使分析、时变分析、各向异性分析

- 其他

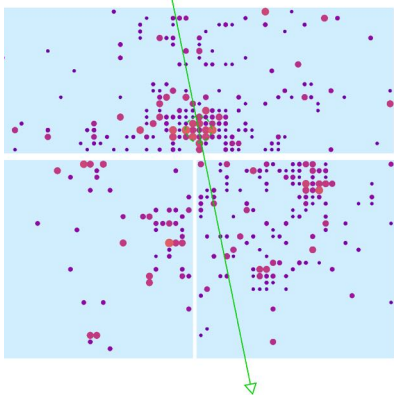
- 噪声过滤
- 快速模拟
- 探测器标定
- 智能运维
-

背景排除

- 根据原初粒子种类的不同，地面阵列观测到的大气簇射分为**强子簇射**和**电磁簇射**两类
- 伽马天文是寻找宇宙线起源问题的重要手段，有效鉴别**伽马射线**和**强子背景**是伽马天文学的前提和基础
- 利用AI技术提高宇宙线背景排除能力，对于LHAASO提高灵敏度，增强对伽马源的观测能力至关重要



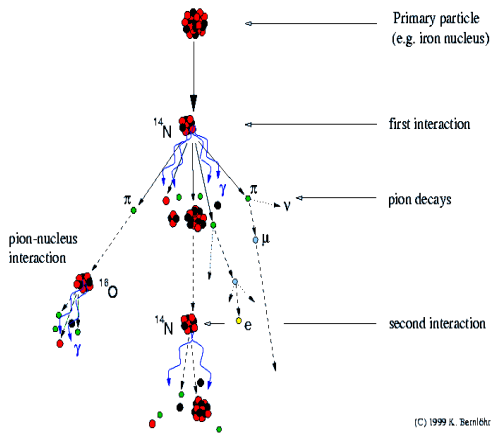
伽马事例



宇宙线事例

强子簇射

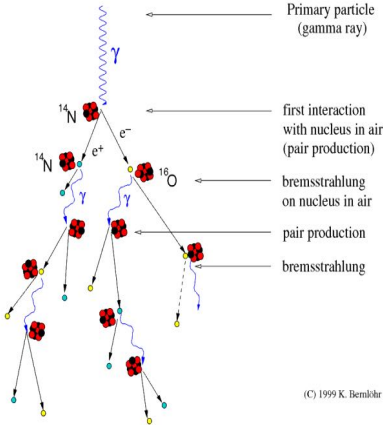
Development of cosmic-ray air showers



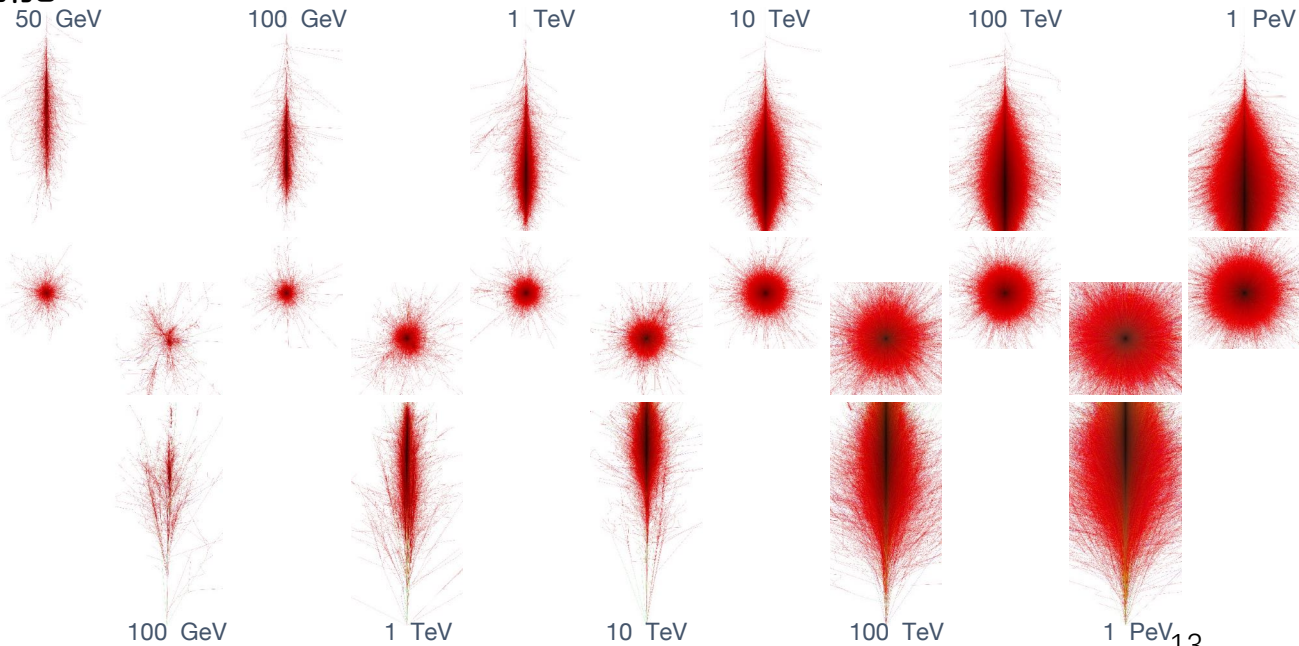
(C) 1999 K. Bernlöhr

电磁簇射

Development of gamma-ray air showers



(C) 1999 K. Bernlöhr



伽马

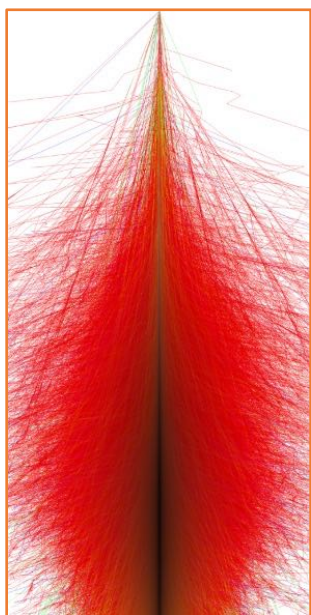
观测面

宇宙线

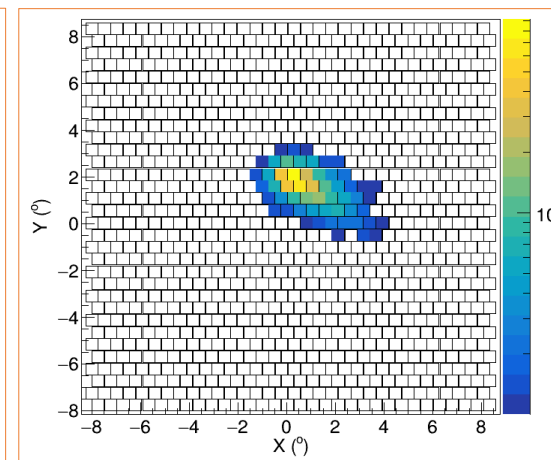
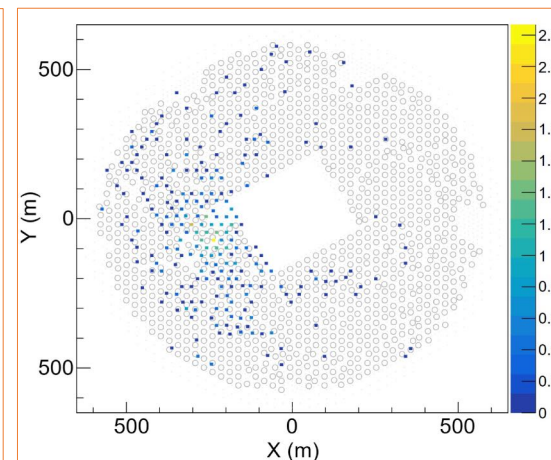
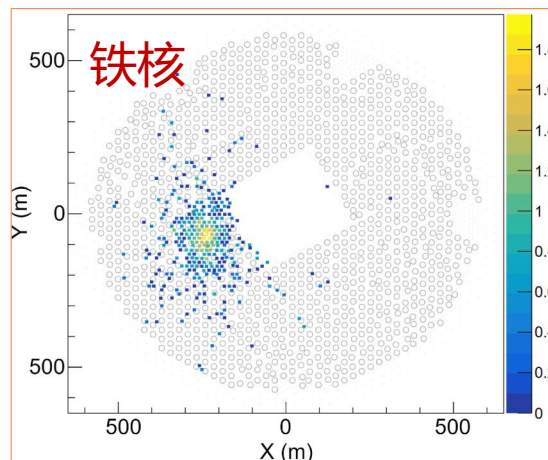
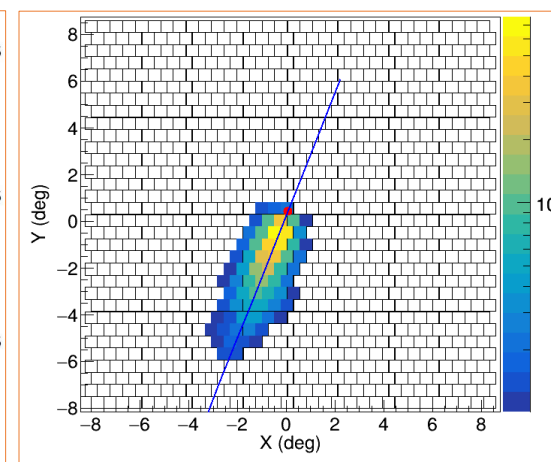
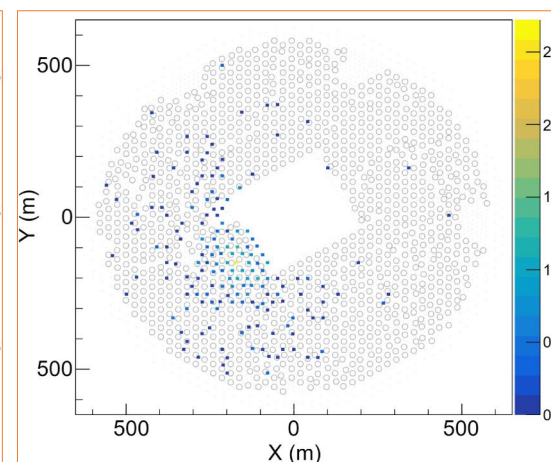
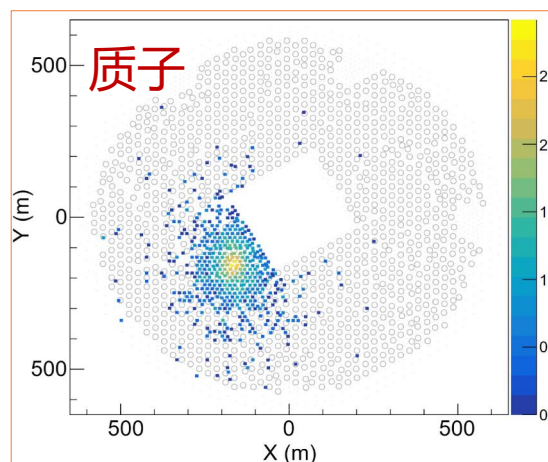
宇宙线成分鉴别

宇宙线成分鉴别对宇宙线成分能谱、各向异性等研究十分重要，传统方法鉴别能力的提升已达瓶颈

质子/1 PeV



铁核/1 PeV

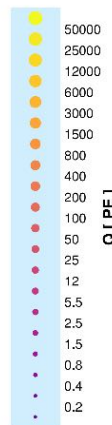


<https://www.ikp.kit.edu/corsika/>

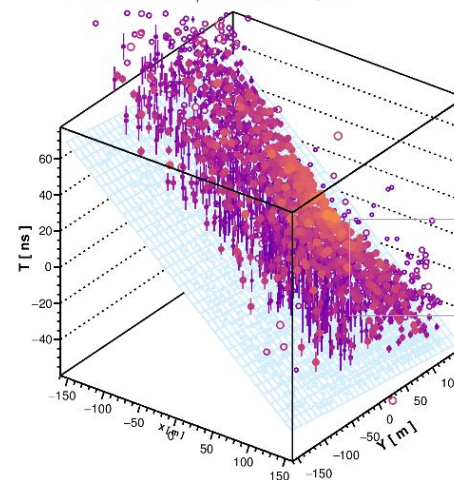
事例重建

- 事例重建是为了通过次级粒子的信息反推原初粒子的信息（**能量、芯位、方向**等）。
- 簇射前锋面近似一个抛物面。通过重建的芯位进行曲面修正，可以得到准确的方向重建。
- 方向重建精度直接影响阵列的观测性能，能量较低或芯位落在阵列之外事例的方向重建精度较差。

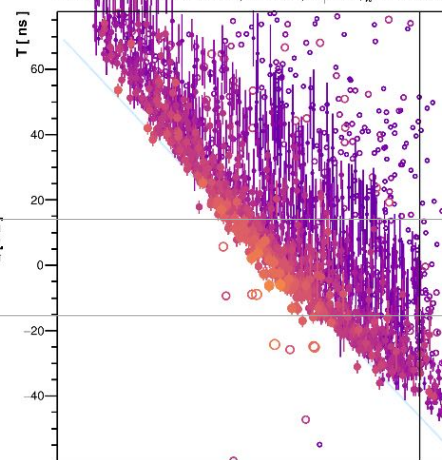
20221009/132156/0.207964261: $\theta=6.36\pm0.05^\circ$, $\phi=6.58\pm0.87^\circ$



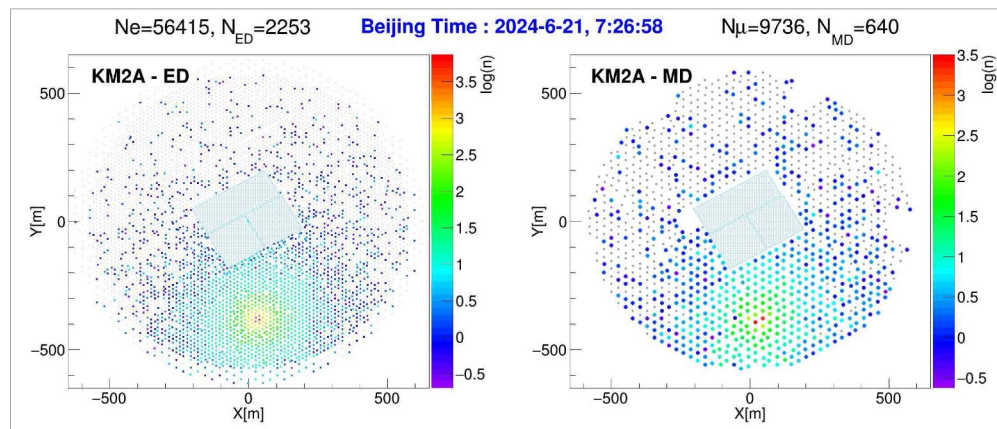
20221009/132156/0.207964261: $\theta=6.36\pm0.05^\circ$, $\phi=6.58\pm0.87^\circ$



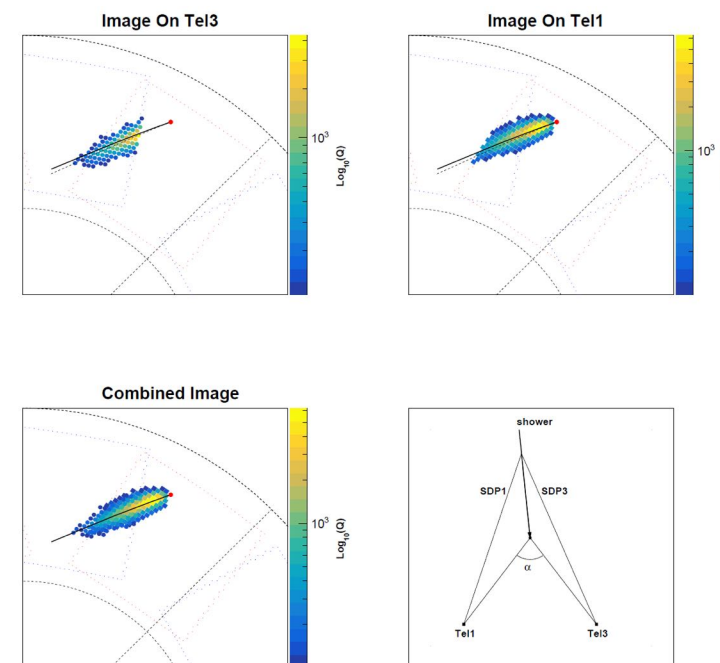
147125/171218/12931: nHit=2024, nFit=1558, $\Delta\alpha=0.11^\circ$, $\chi^2=427.3/1556$



WCDA



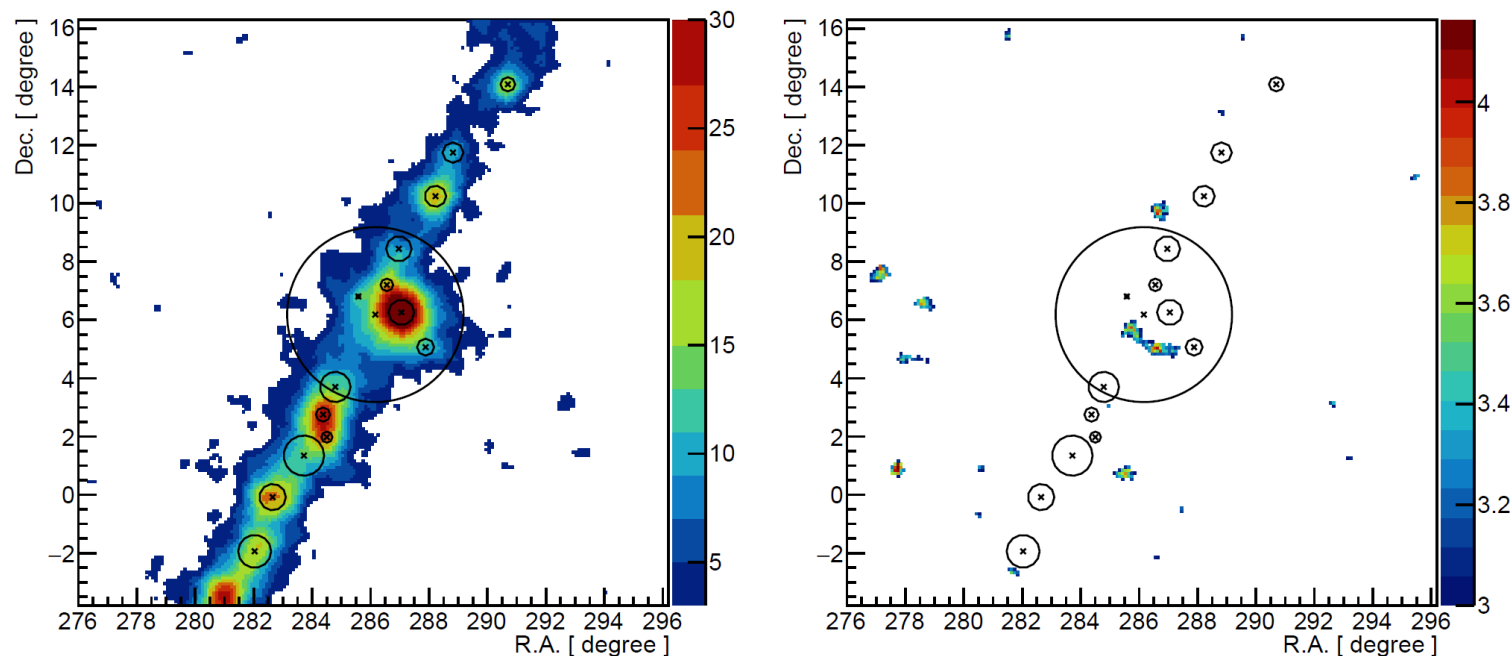
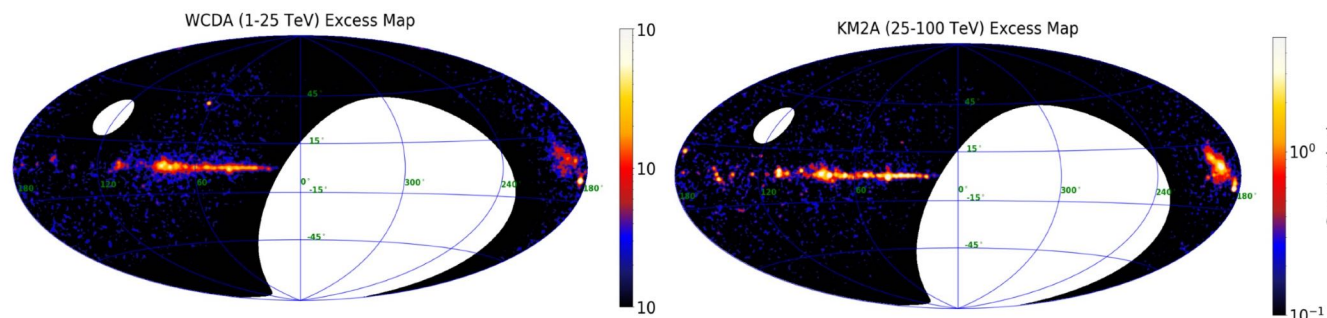
KM2A



天体源分析

目前LHAASO已经发布了第一期伽马射线源表，发现了大量新的甚/超高能伽马射线源

AI在多个重叠源+泡状结构+弥散辐射的解析方面具有优势



LHAASO- AI 现状

- 团队建设
- 数据集
- 模型
- 目前结果

团队和整体进度

主要工作和参与者

	WCDA	KM2A	WFCTA
背景排除/粒子分类	● ★ ▲ ◆ ❖	● ★ ▲	★
能量重建	▲	● ★ ▲	
方向重建	▲ ◆	● ★ ▲	
无监督基座训练	●		
实验数据准备		●	
仿真数据准备	●	●	●
数据平台+可视化		●	

AI模型研发进展

	WCDA	KM2A	WFCTA
背景排除/粒子分类	○	○	○
能量重建		△	
方向重建		△	

LHAASO数字化智能化改造项目 (AI相关部分)

高能所

* 白云翔、姚志国、程耀东

高能所

● 姚志国、胡世聪、黄勇、吕洪魁、张笑鹏、黄志豪、阮曼奇、朱永峰、李哲、刘祥廷、张易于、张航畅、席绍强、尹丽巧、张兵、陈松战

之江实验室

★ 陈华曦、徐宸原、王智、刘欢迎、刘畅

自动化所

▲ 朱炳科、王潇潇、史千秋、杨辰

智谱AI

❖ 李玮巍、黄子赫、刘倩颖

上海交大

◆ 周浩、唐睿仪

数据集

观测数据（无/弱标签，规划中）

WCDA

事例层级：

触发时刻（MJD）

重建角度（ θ , ϕ ）、芯位（ x_c , y_c ）、鉴别特征参量（ p_{in} , $compactness$, $f5w$ 等）

hit层级：

探测器ID（位置）、时间、电荷量

KM2A

事例层级：

触发时刻（MJD）

角度（ θ , ϕ ）、芯位、簇射参数（ $size$, age ...）

hit层级：

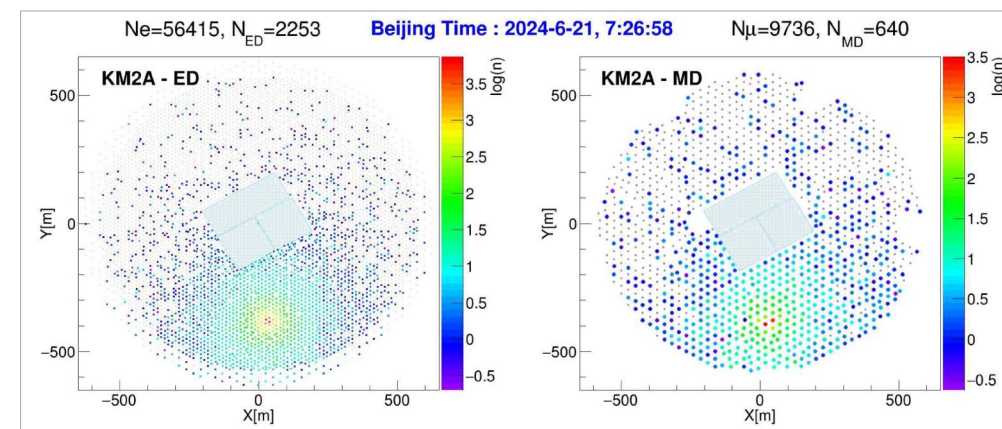
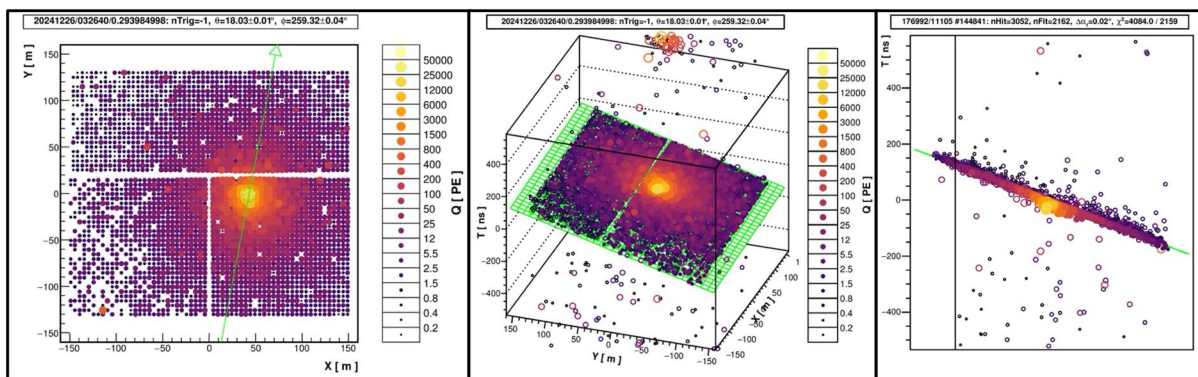
探测器ID（位置）、类型（ED, MD）、时间、电荷量（ $anode$, $dynode$ ）

WCDA和KM2A的实验数据在产生后均经过解码、重建、事例筛选、标定等标准处理流程，保存为ROOT格式

1) 电荷量 Q 的二维分布

2) 电荷量 Q 和相对时间 t 的二维分布

3) 侧视图



数据集

模拟数据（有标签）

- CORSIKA+Geant4模拟EAS过程和探测器响应
- 融入传统重建方法结果
- 格式与观测数据相似

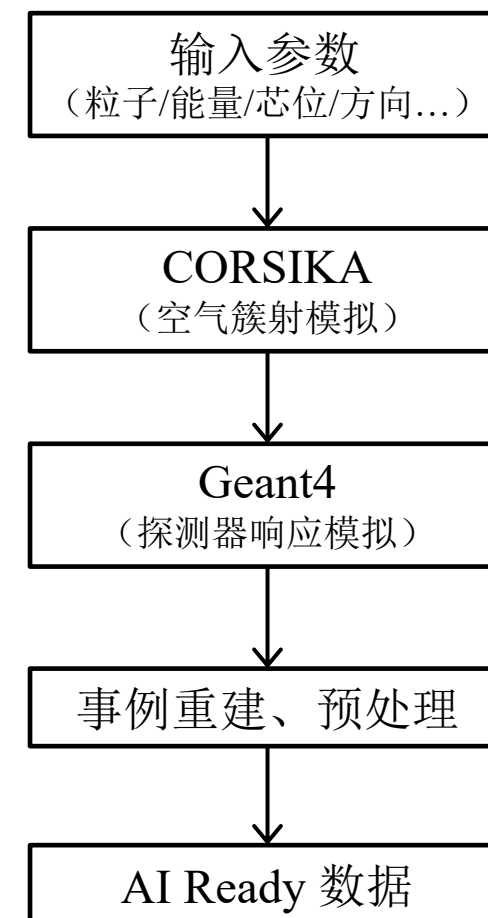
正在进行：

- 更大规模数据量的产生
- 标准化数据制备和管理工具的开发

目前的数据集规模

阵列	类别	事例数 (10^6)	文件大小 (GB)
WCDA	宇宙线	115	446
	伽马射线（各向同性）	38	162
	伽马射线（跟踪Crab）	140	600
KM2A	宇宙线	20	210
	伽马射线	7	41

数据集制作流程

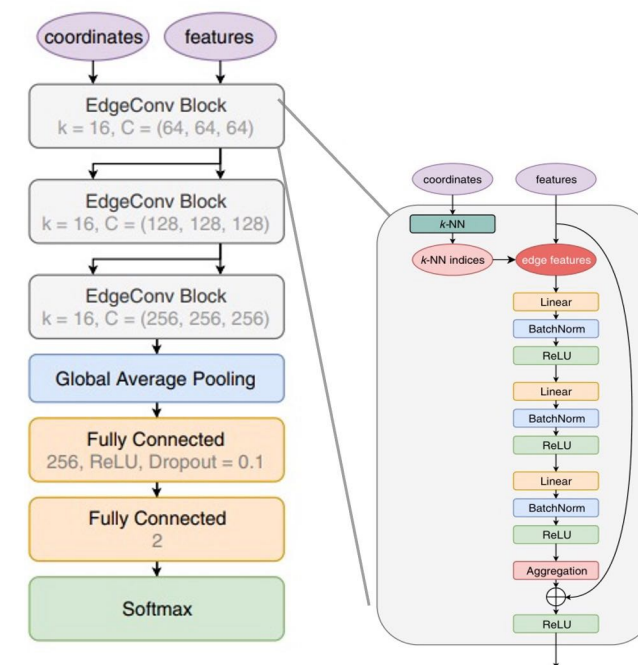
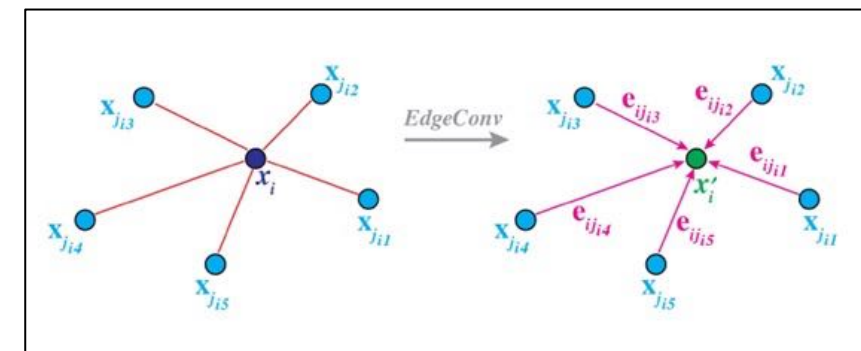
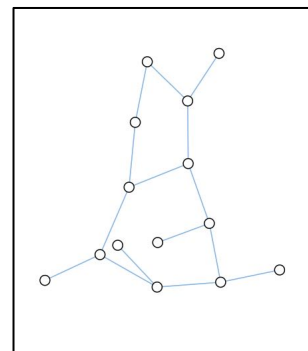
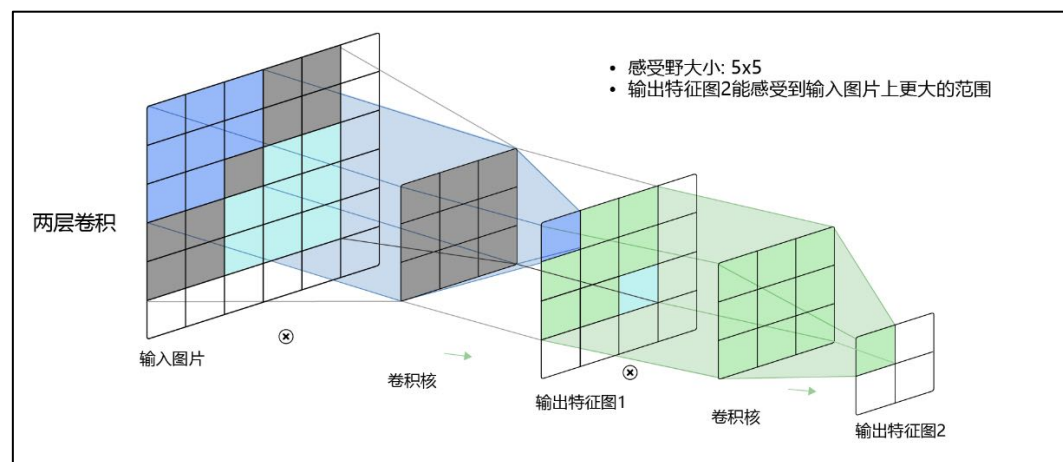


模型

传统模型

- CNN
- GNN及其变种

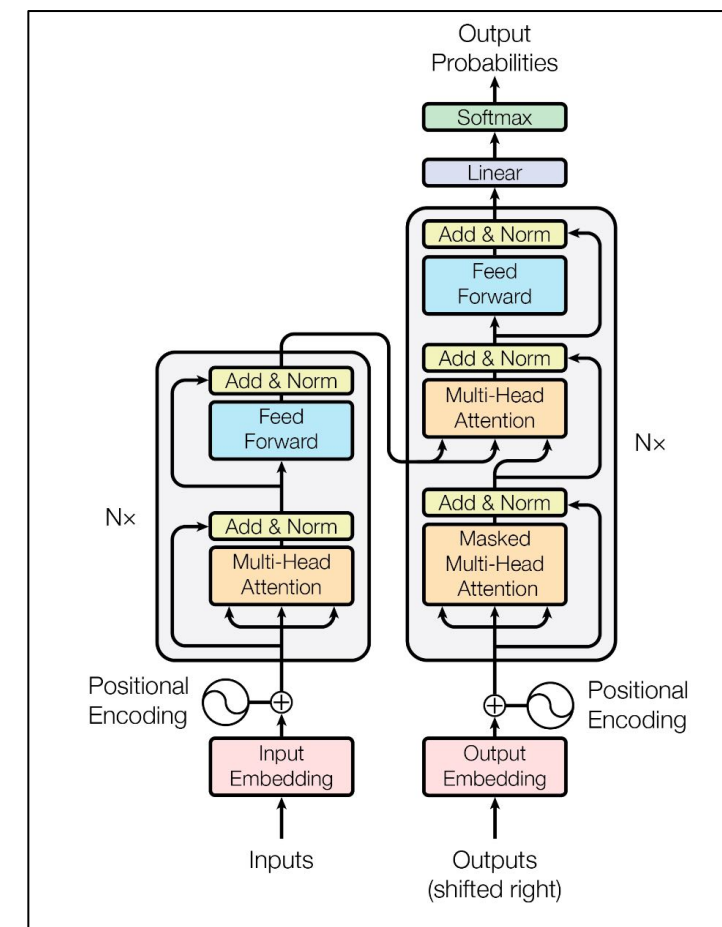
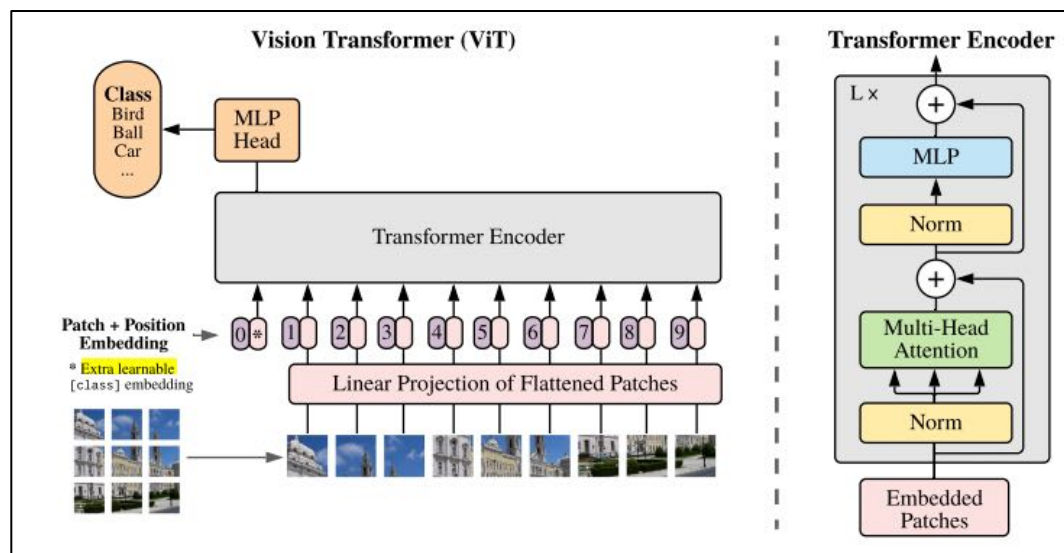
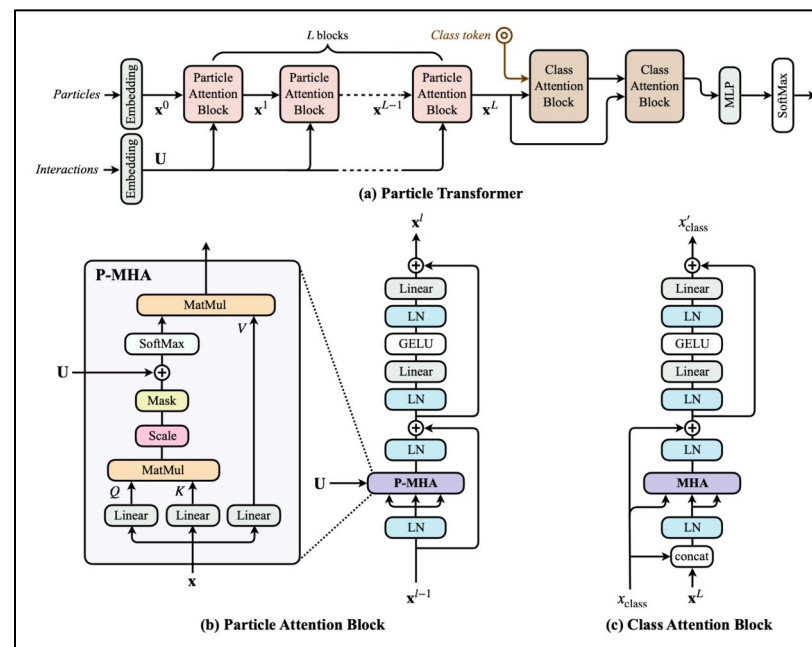
DGCNN/ParticleNet



模型

Transformer及其变体

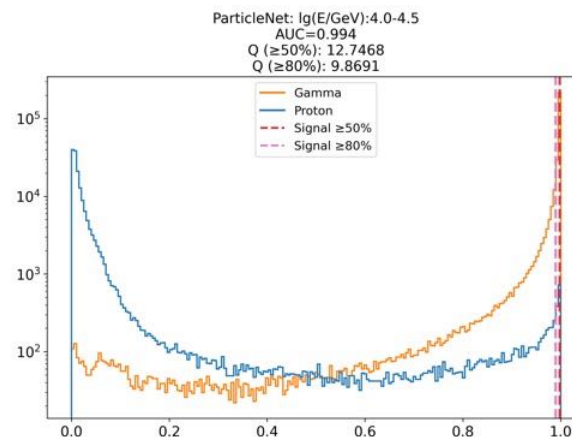
- Vision Transformer (ViT)
- Particle Transformer
- LLM/多模态大模型



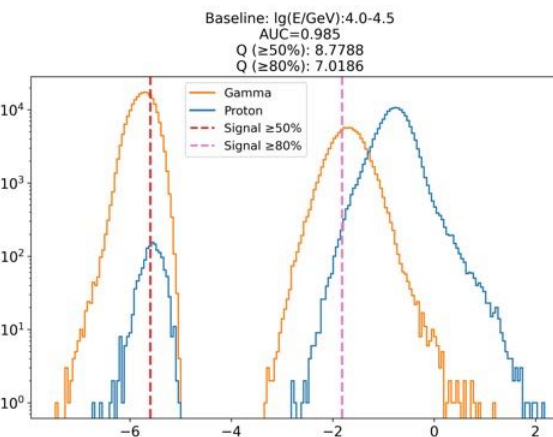
初步结果 – KM2A背景排除任务

$$Q = \frac{\epsilon_{\gamma}}{\sqrt{\epsilon_{CR}}}$$

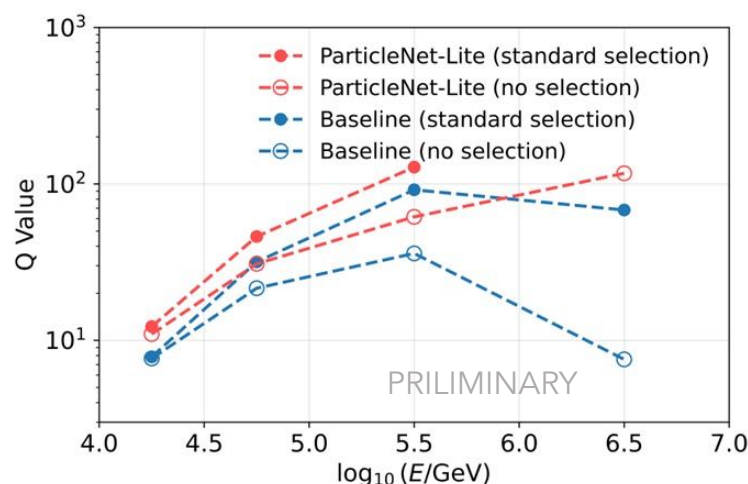
- 采用ParticleNet, Transformer, ViT等模型
- 相比传统方法, 对质子-伽马的区分能力有较大改善, 在10–1000TeV能区Q因子提50%–200%



ParticleNet

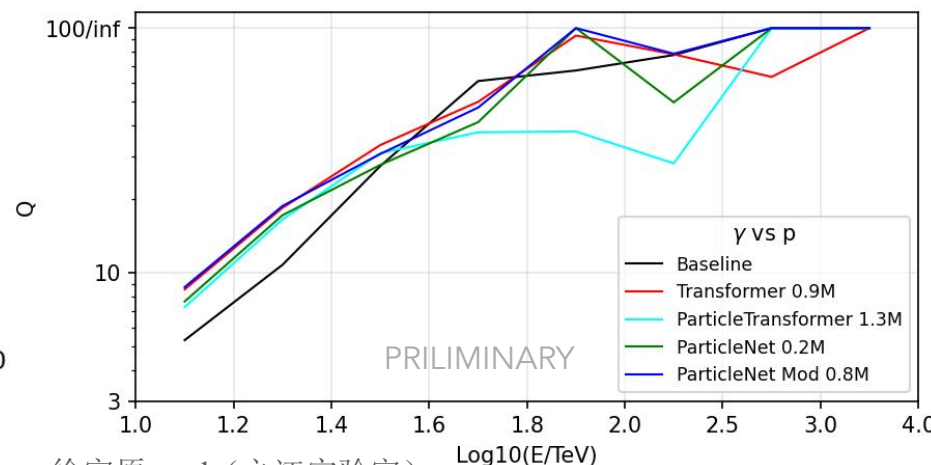


Baseline



吕洪魁 et.al (高能所)

The LHAASO Collaboration



徐宸原 et.al (之江实验室)

伽马 vs 宇宙线	10 ⁴ ~10 ^{4.5} GeV	10 ^{4.5} ~10 ⁵ GeV
ViT	ACC: 0.9270 Q因子: 17.6378	ACC: 0.9536 Q因子: 103.0970
HiT	ACC: 0.9417 Q因子: 16.7268	ACC: 0.9880 Q因子: 129.0220
GLaM微调	ACC: 0.9827 Q因子: 19.7903	ACC: 0.9928 Q因子: 170.0124

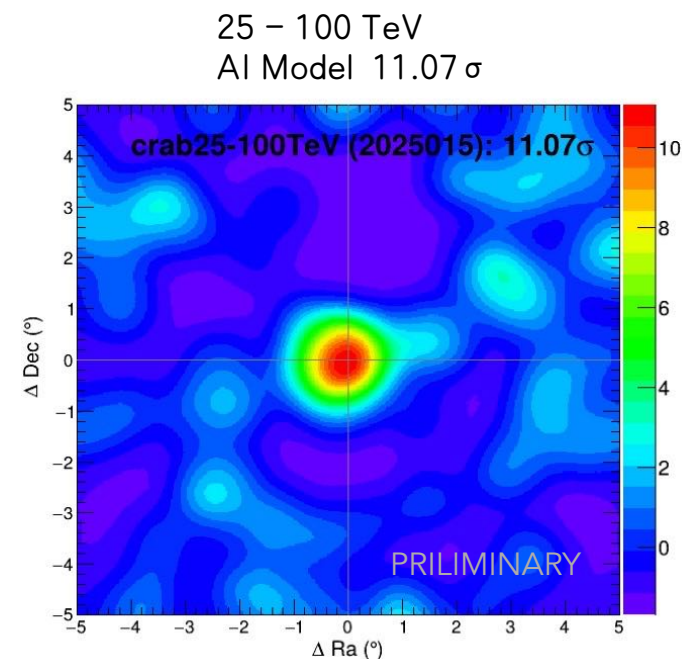
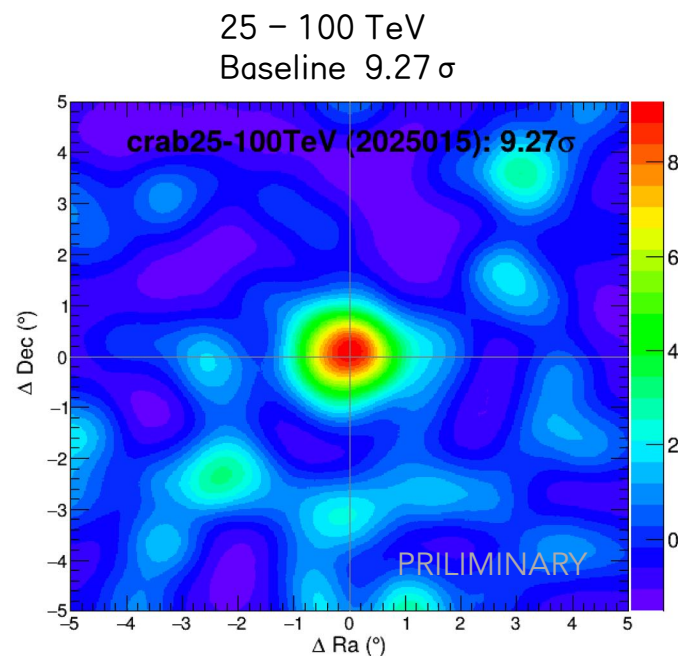
PRILIMINARY

史千秋 et.al (自动化所)

在实验数据上的应用尝试

- 将训练后的模型应用于KM2A对宇宙“标准烛光”Crab Nebula 的观测数据
- 经过时间和电荷标定, 事例重建, 坐标转换等, 得到与模拟数据相同的输入变量
- 显著性提高约10–20%

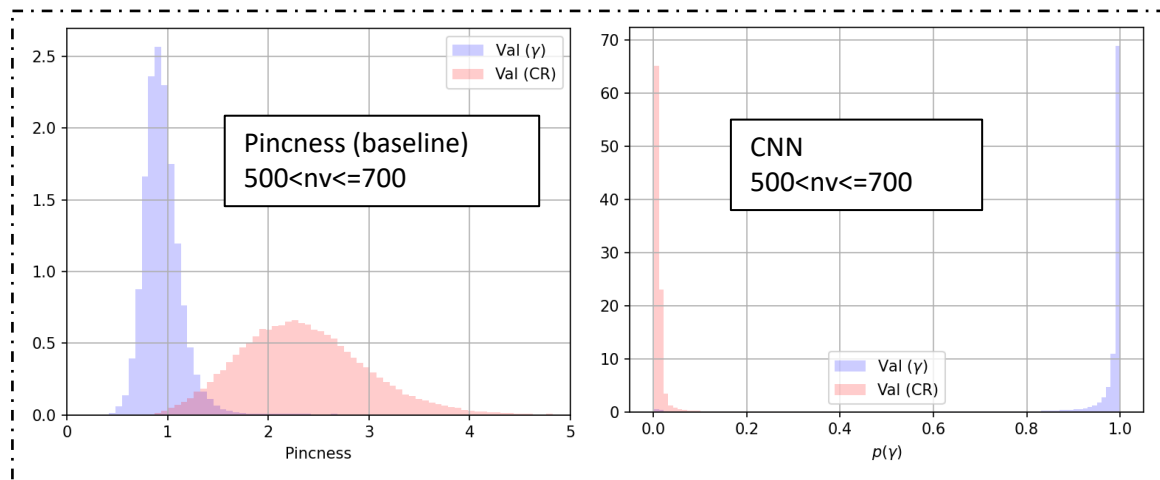
与预期差距较大, 需适当调整策略, 解决模拟–真实迁移问题!



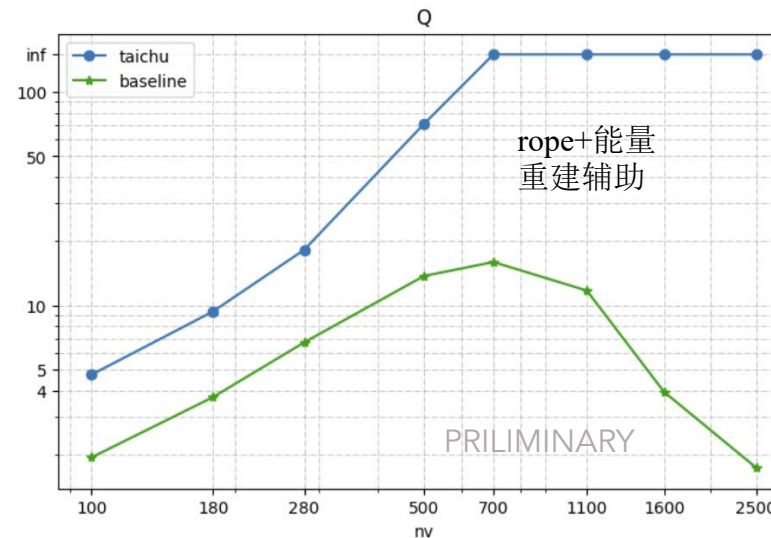
张笑鹏 et.al (高能所)

WCDA背景排除任务

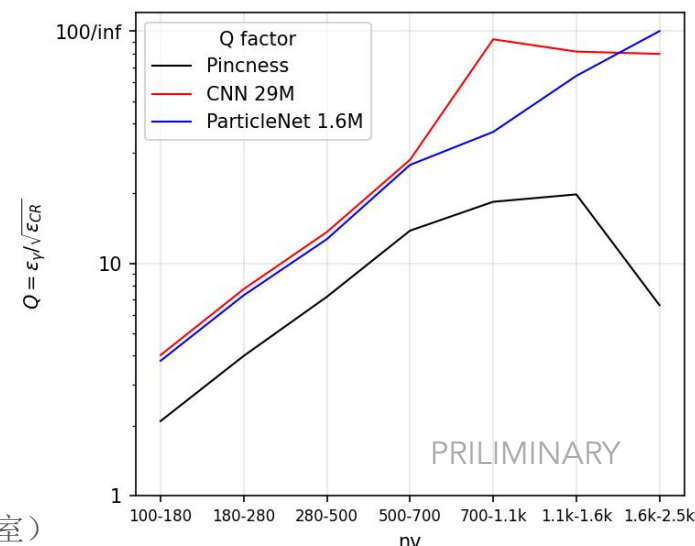
- 采用CNN, ParticleNet, ViT等模型
- 在模拟数据上, Q因子相比传统方法 (Pinness) 在全部能段内有50%以上的提升



徐宸原 et.al (之江实验室)



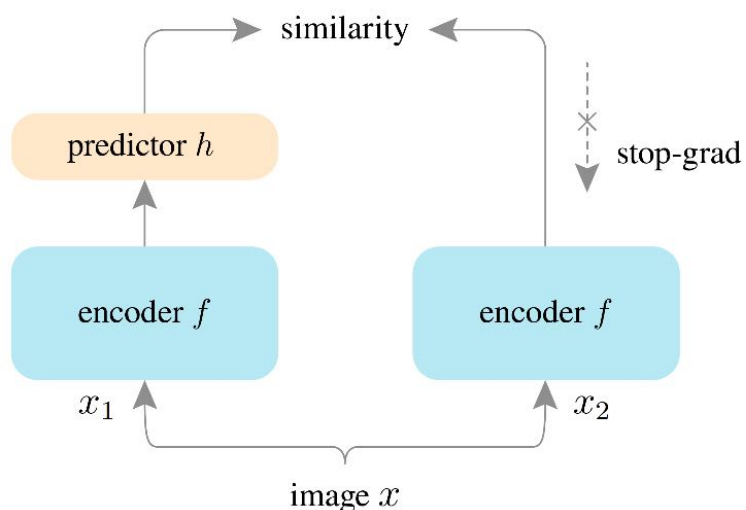
王潇潇 et.al (自动化所)



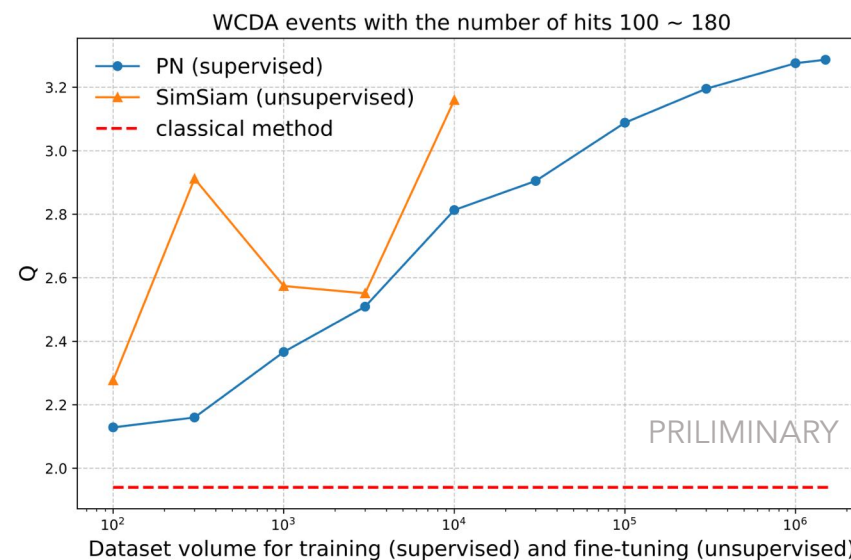
WCDA背景排除任务 自监督学习

朱永峰 et.al (高能所)

自监督学习SimSiam [1]:
数据增强后的样本 x_2 , 与原样本 x_1 通过同一
编码器后, 接预测器将数据增强前后的相似
度拉近



使用 3×10^6 无标签数据训练编码器,
再使用少量有标签数据微调 $\Rightarrow$ 效果优
于直接在有标签数据训练



[1] X. Chen and K. He, "Exploring Simple Siamese Representation Learning," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

WCDA背景排除任务 大语言模型/多模态模型

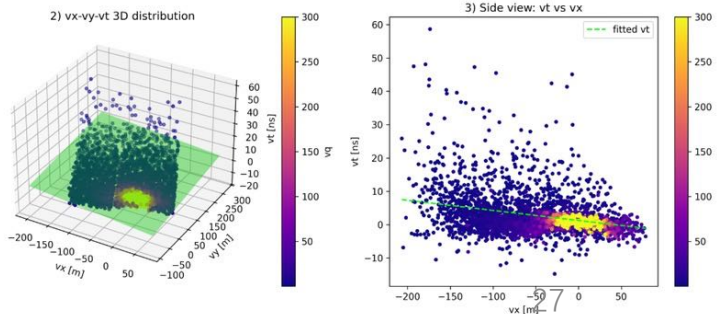
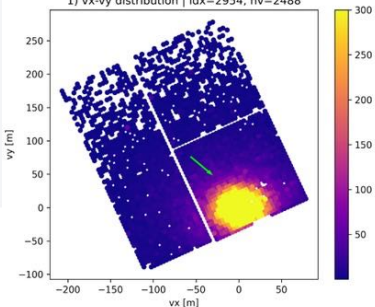
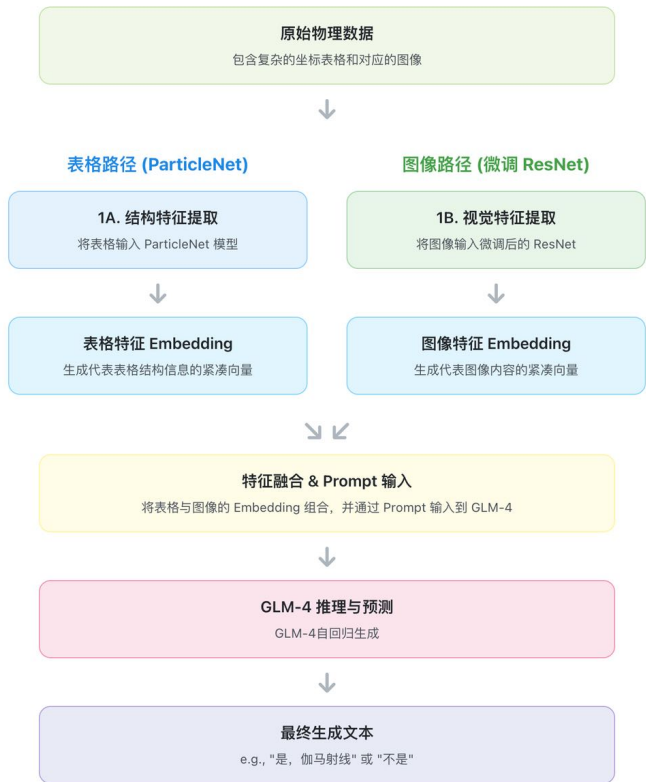
李玮巍 et.al (智谱AI)

- 基于LLM/图像融合多模态模型的解决方案，表现出良好的潜力，是未来探索的重点方向之一

	GLM-4-0414		particlenet-T		视觉	
	Acc	Q	Acc	Q	Acc	Q
NV_ALL	0.7325	1.5043	0.8069	1.4474	0.8349	1.9950
NV_100-180	0.8552	2.4500	0.8388	1.5300	0.8679	2.1089
NV_180-280	0.9262	4.1844	0.8922	1.8733	0.9213	2.582
NV_280-500	0.9568	7.3599	0.8790	1.8774	0.9095	2.5877
NV_500-700	0.9647	7.6529	0.8213	1.5939	0.8498	2.1969
NV_700-1100	0.9807	8.6656	0.7832	1.4384	0.8104	1.9826
NV_1100-1600	0.9580	8.8138	0.6808	1.3020	0.7044	1.7946
NV_1600-2500	1.0000	-	0.6301	1.1399	0.652	1.5712
NV_2500-4500	0.9149	2.9721	0.4113	0.6704	0.4256	1.324

使用相同数据只训练一轮时LLM表现最好，表明LLM有更好的数据利用效率，可能具有更好的泛化性能。

未来技术方案（之一）：
专家小模型+融合视觉信息的多模态大模型

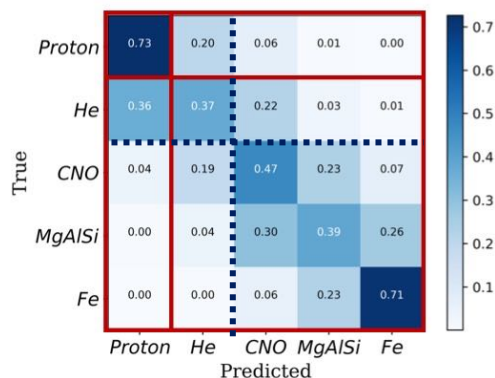


宇宙线成分鉴别 (KM2A)

基于ParticleNet模型尝试了两种分类任务:

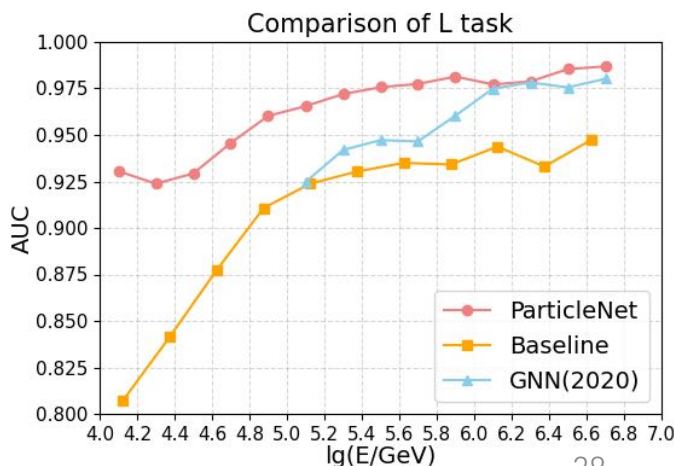
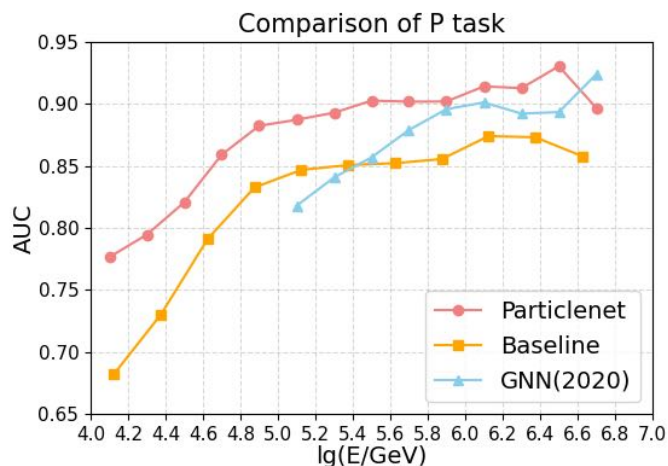
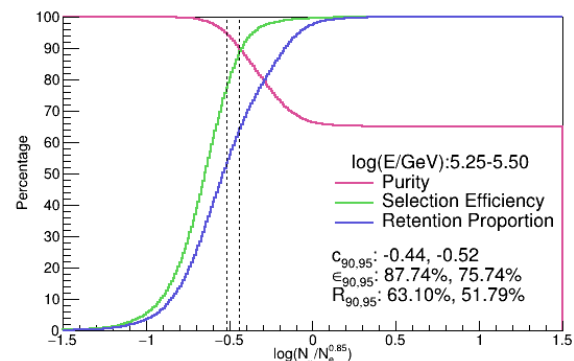
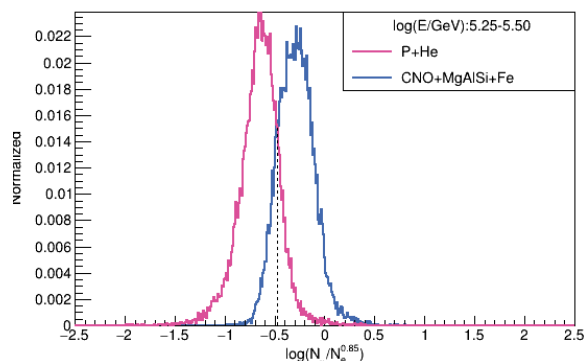
P任务 (挑选质子)、L任务 (挑选质子+He核)

- 根据重建能量按照logE分组
- 按照芯位和天顶角进行事例筛选
- 粒子的空间坐标作为顶点
- hit信息作为特征, 包括坐标, 探测器种类, 电荷量, 时间



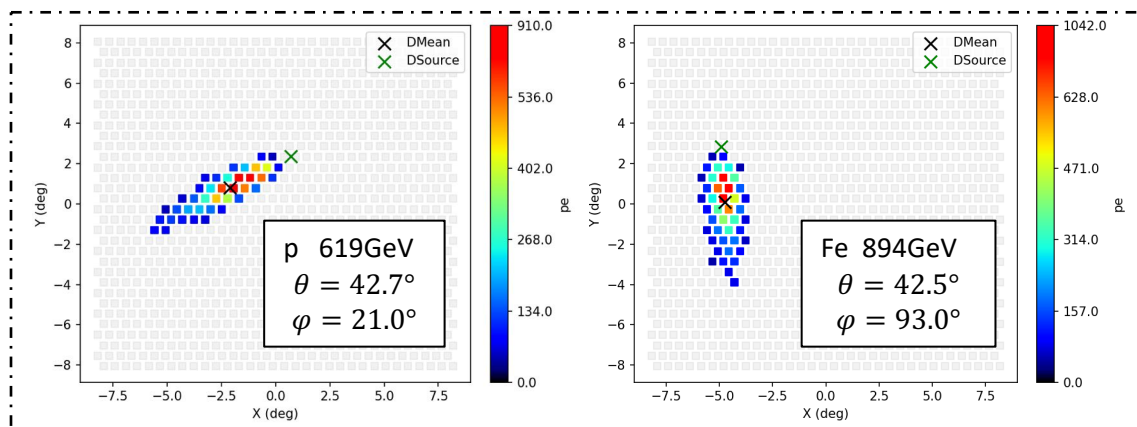
baseline:

$$f(N_e, N_\mu) = \log\left(\frac{N_\mu}{N_e^{0.86}}\right)$$



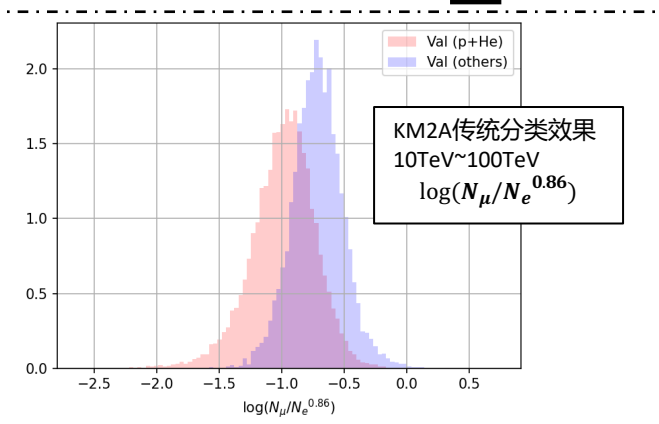
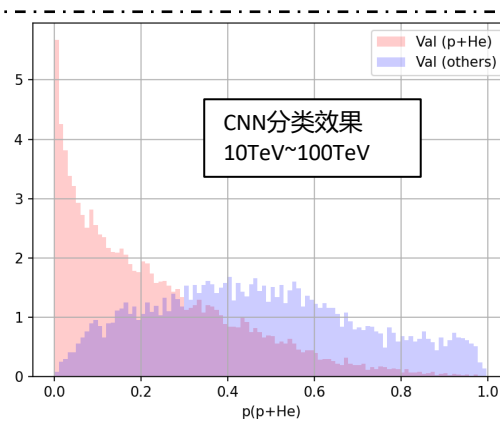
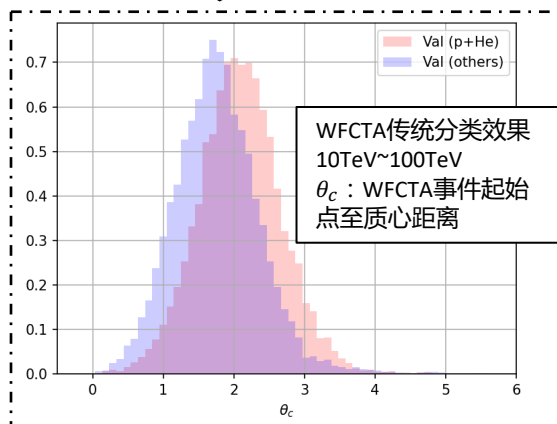
宇宙线成分鉴别 (WFCTA)

徐宸原 et.al (之江实验室)



肉眼已无法辨别WFCTA的p/Fe

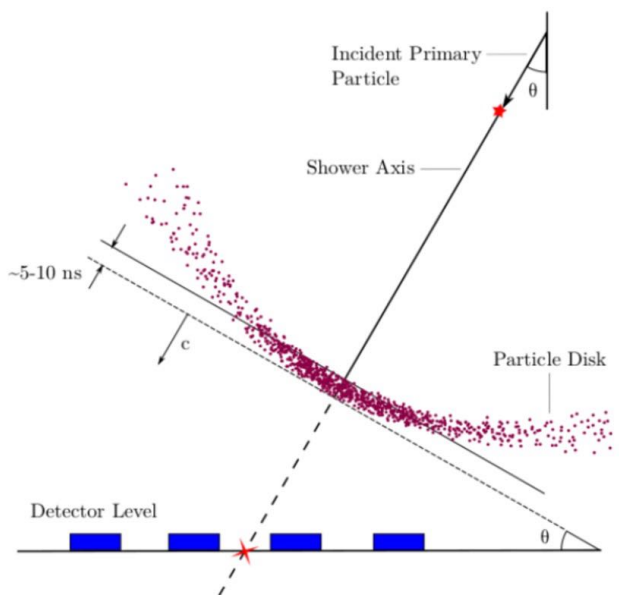
只靠WFCTA信息有多少效果?



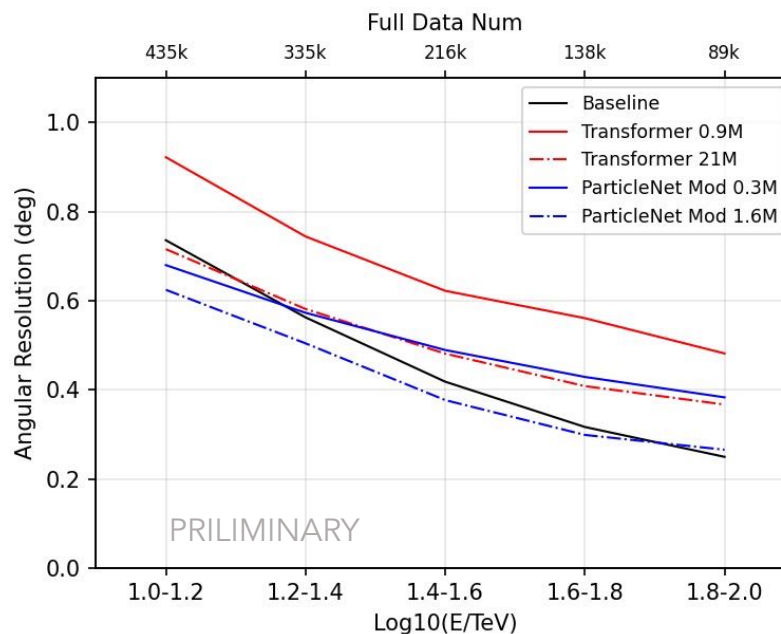
方向重建 (KM2A)

- 地面粒子实验通过探测到次级粒子的**位置**、**到达时间**等信息，反推原初粒子的能量、方向
- 传统方向重建为最小二乘拟合

$$\chi^2 = \sum [Ct_i - Ct_0 - (lx_i + my_i - nz_i) - \alpha dr_i]^2$$

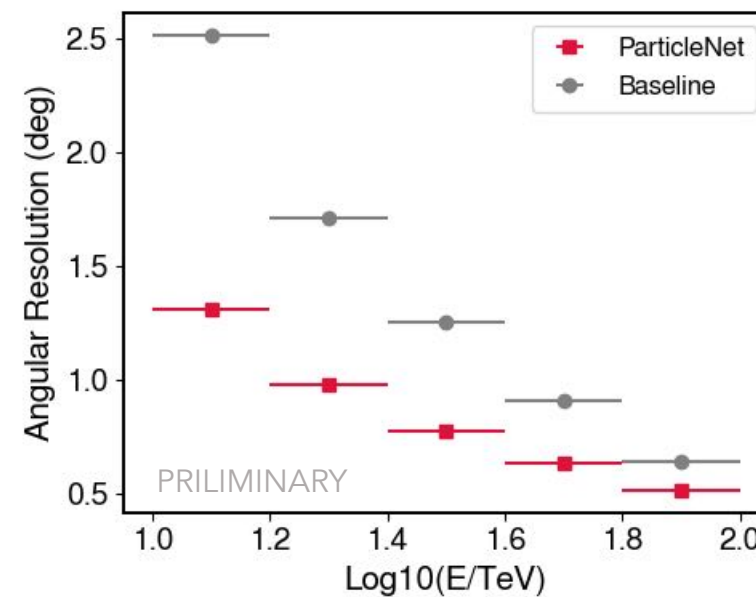


- 当前性能表现最好的模型是略作改动的ParticleNet
- 对伽马射线 (10–40 TeV) 有**10–20%**的角分辨提升
- 对铁核 (10–100 TeV) 有**20–50%**的提升



徐宸原 et.al (之江实验室)

- 探测器的角分辨与其**观测灵敏度**成正比，提升角分辨对LHAASO非常重要！
- 目前已观察到，相比分类任务，方向重建对**模型参数量**和**训练数据量**有更大的需求



张笑鹏 et.al (高能所)

能量重建 (KM2A)

测量宇宙线能谱特别是“膝”区能谱，是研究宇宙线起源中的一个关键问题。但在地面实验中，能量、成分和强相互作用模型互相纠缠，使得准确的能量估计变得复杂困难。

基于ParticleNet模型的能量重建，在几乎整个“膝”区内的能量分辨率和bias都优于传统方法。

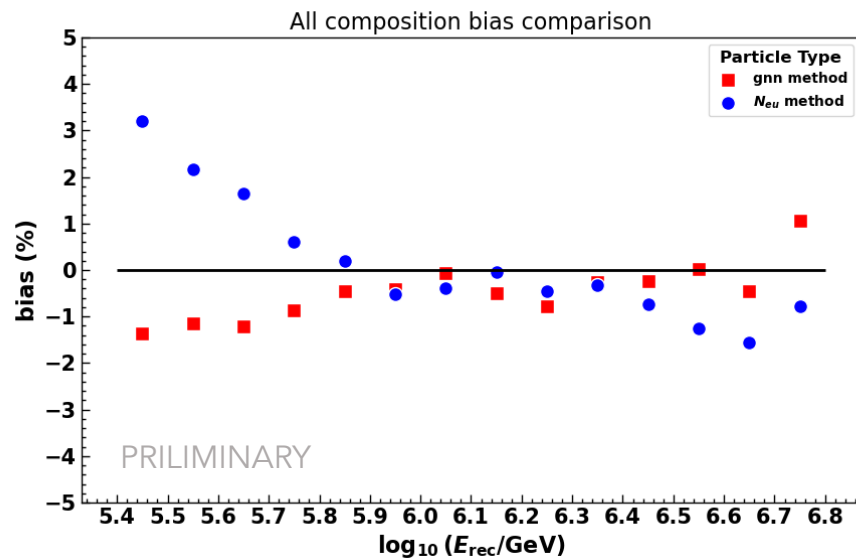
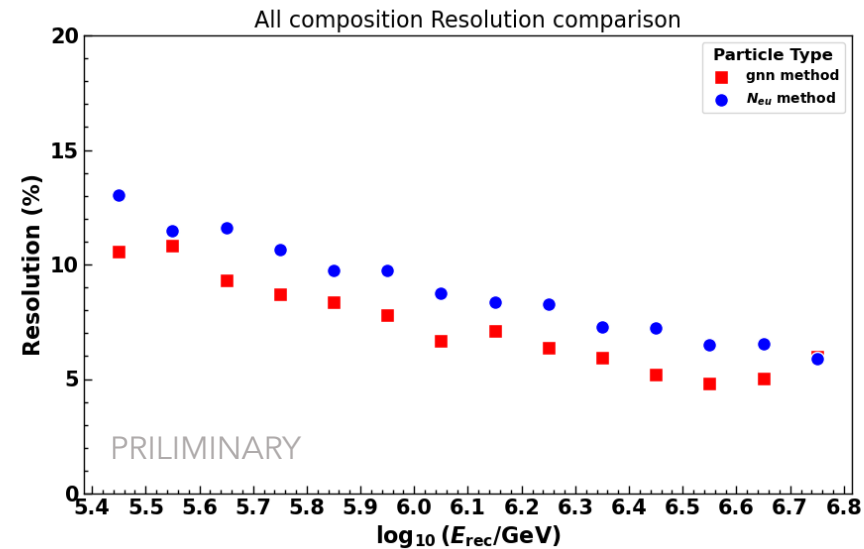
- 五种主要成分：P, He, CNO, MgAlSi, Fe
- 能量100TeV ~ 1PeV, 强相互作用模型QSJETII
- 事例筛选：芯位分距离阵列中心200 ~ 500m, 重建天顶角 $\leq 35^\circ$
- 按照能谱幂率-2为事例添加权重
- 目标变量logE, 损失函数MSE

baseline (传统方法):

$$N_{e\mu} = \frac{N_e}{A_{ED}} + \frac{25 \cdot N_\mu}{A_{MD}}, \quad E_{rec} = b N_{e\mu},$$

$$b = 548 \text{ GeV}$$

(H.Y. Zhang et. al, PRD **106**, 123028)



LHAASO- AI 总结&展望

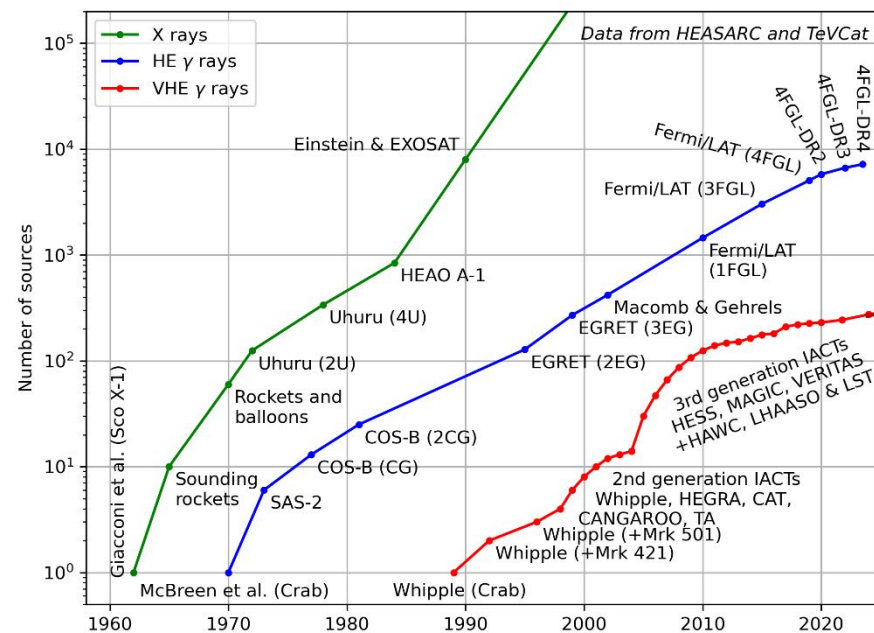
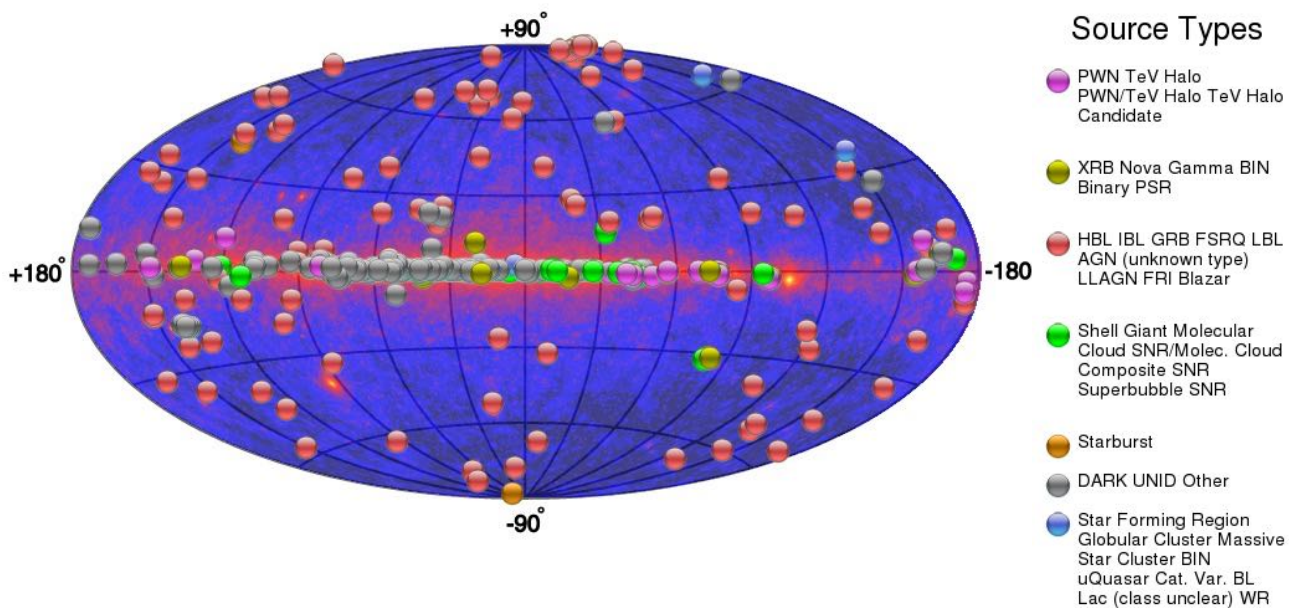
- 我们希望AI对LHAASO不仅作为助手帮助科研人员，更能够帮助提升装置性能，直接作用于科学产出
- 已经开展了一些探索性质的工作，显示出良好的效果，但仍有很大的提升空间
- 继续着力提升现有模型（ViT, ParticleNet ...）及其变体在WCDA/KM2A上的背景排除能力，特别是在真实数据上的表现
- 自监督/弱监督学习或许能有效缓解模拟和实验数据的不一致性难题
- 大语言模型/多模态模型在分类、多源解析、时变分析等任务上可能有更高上限

AI助力LHAASO取得更多科学发现

随着观测数据的积累，以及新技术的加持，甚/超高能伽马天文将进入新时代

引入AI将使LHAASO继续保持国际领先

终极目标：解决宇宙线起源这一世纪之谜



2030
(1000
sources)

2027
(start
using AI)

github.com/sfegan/kifune-plot

感谢关注!