机器学习在HERD项目中的应用

权征

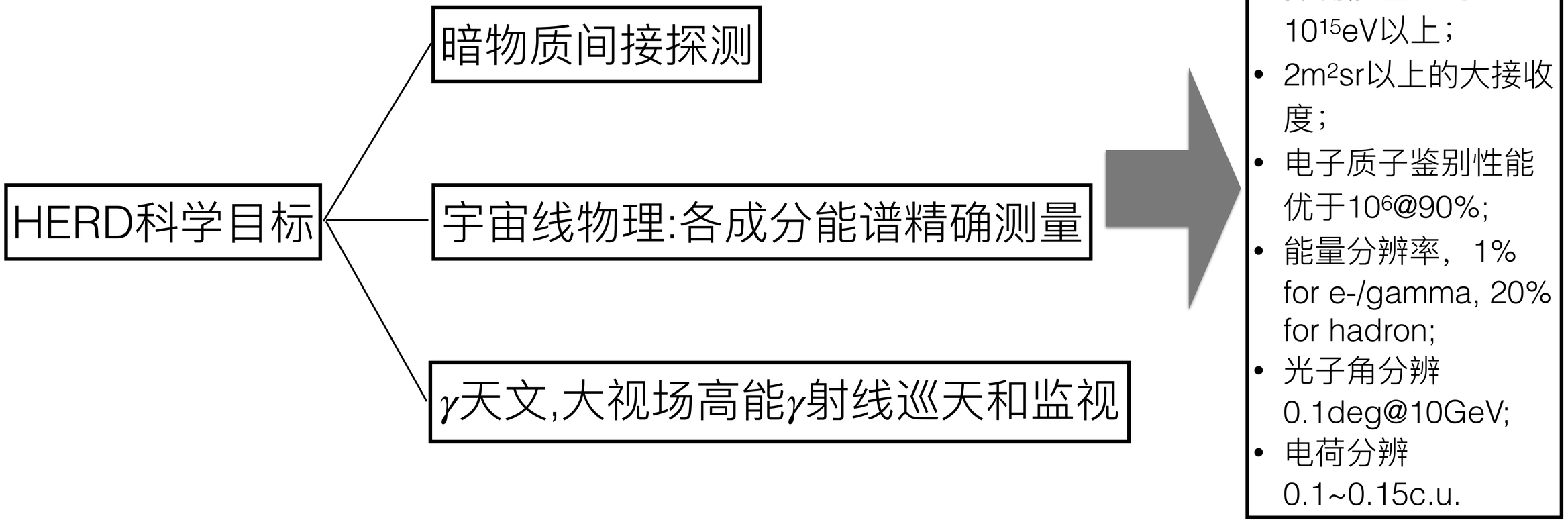
2022.9.19

粒子天体物理中心

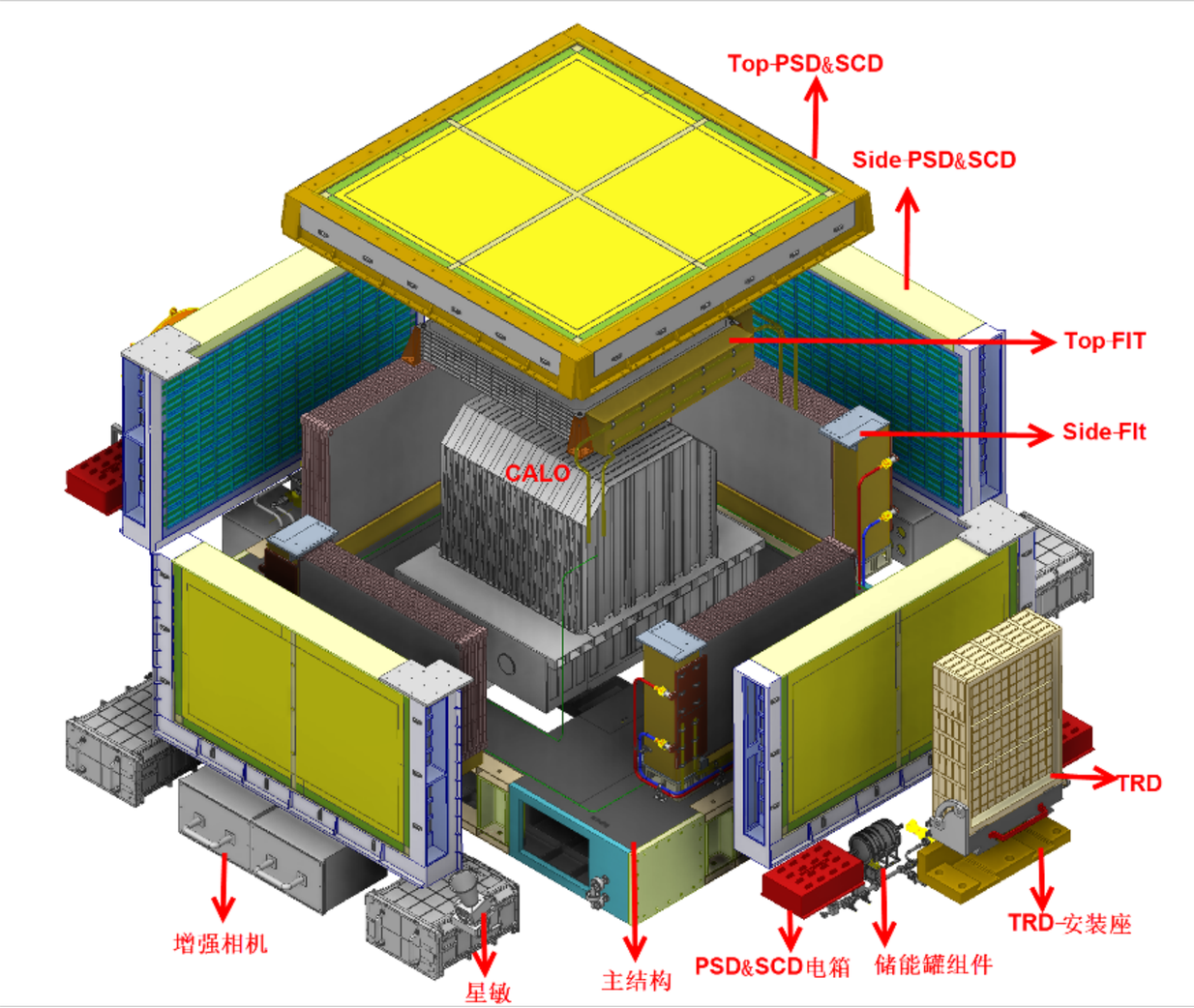
目录

High Energy cosmic-Radiation Detection facility

高能宇宙辐射探测设施：由中国主导的大型国际合作空间天文和粒子天体物理实验，计划安装在中国空间站，在轨运行寿命10年。



HERD基本结构



- 三维成像量能器(CALO):

能量测量、径迹测量、e/p鉴别；

- 硅电荷探测器(SCD):

径迹测量、电荷测量；

- 塑闪探测器(PSD):

电荷测量、带电粒子/光子在线甄别；

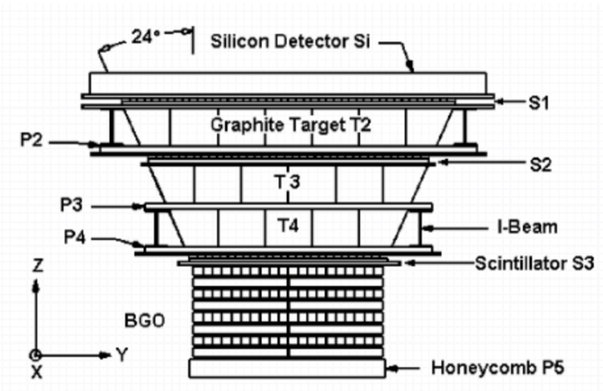
- 闪烁光纤径迹探测器(FIT):

径迹测量；

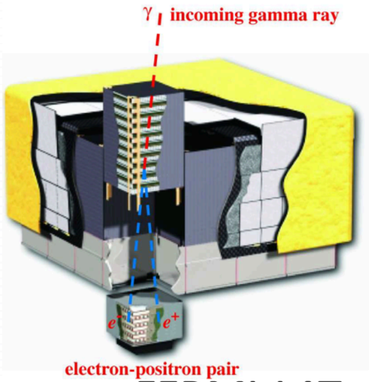
- 穿越辐射探测器(TRD):

TeV能区标定。

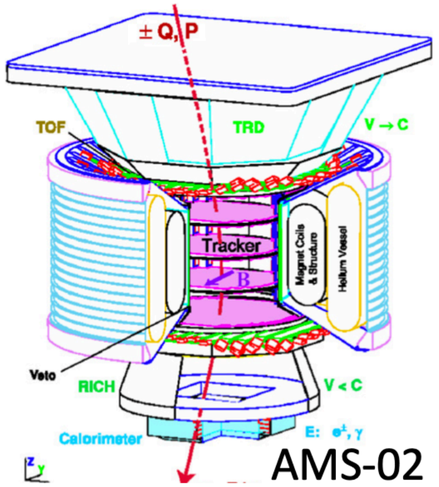
空间物理实验



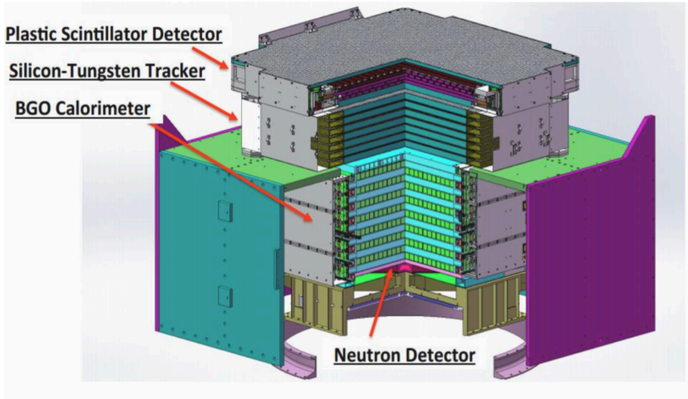
ATIC



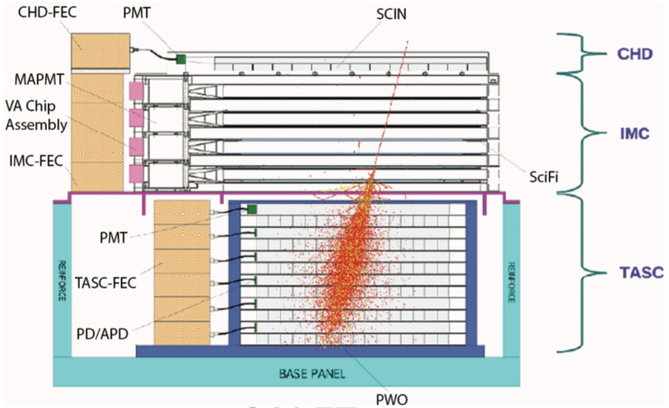
FERMI-LAT



AMS-02



DAMPE



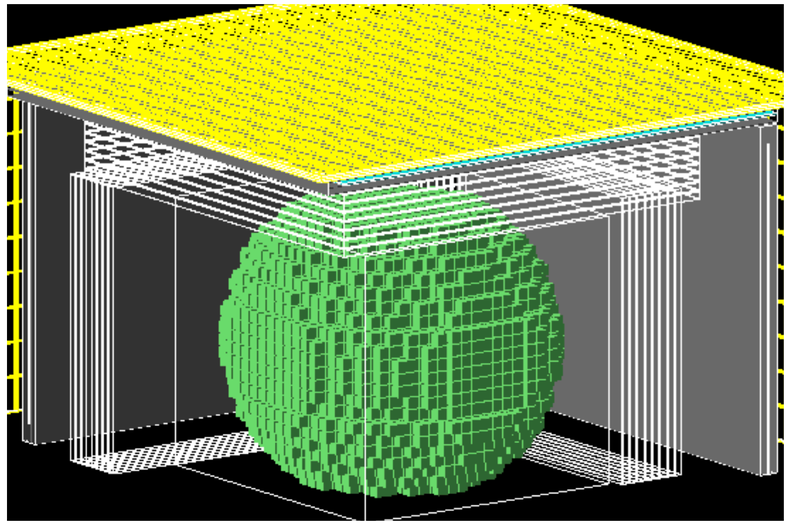
CALET

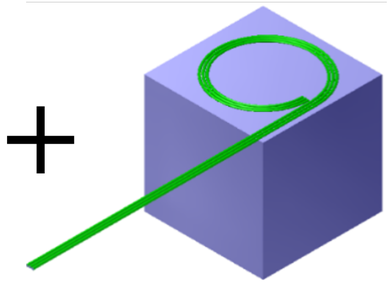
二维XY正交长条量能器：单面灵敏、几何接收度小、读出路数少

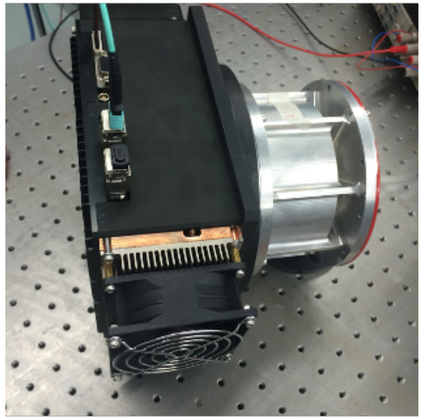
三维成像量能器：五面灵敏、真三维、读出路数多、几何接收度大一个量级

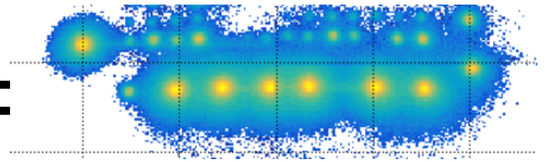
三维成像量能器(CALO)

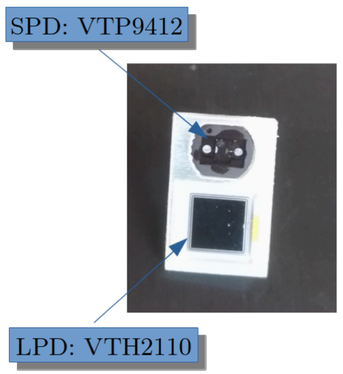
7500块3cm边长LYSO立方体，厚度55X_0，3λ_I

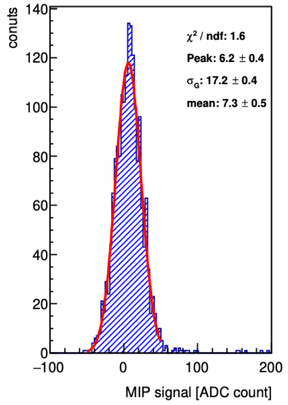






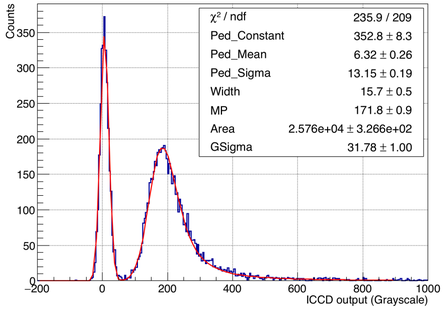






双读出系统：

增强相机IsCMOS + 光电二级管photo-diode



机器学习在HERD项目中主要应用

HERD的绝大多数性能都来自三维成像量能器，希望尝试用机器学习尽可能挖掘量能器的潜力。由于HERD量能器允许各个角度的入射，一些只适用于正入射和小角度斜入射的传统方法失去了作用，需要深入研究对原方法优化推广到大角度斜入射，或者寻找其它替代方案。

➡径迹重建: 主成分分析法(PCA)、神经网络、卷积神经网络；
➡Clustering：DBSCAN、卷积神经网络；
➡能量重建：利用理论公式3D拟合、卷积神经网络；
➡PID：MVA-BDT、神经网络、卷积神经网络；
➡增强相机数字化：图像生成。

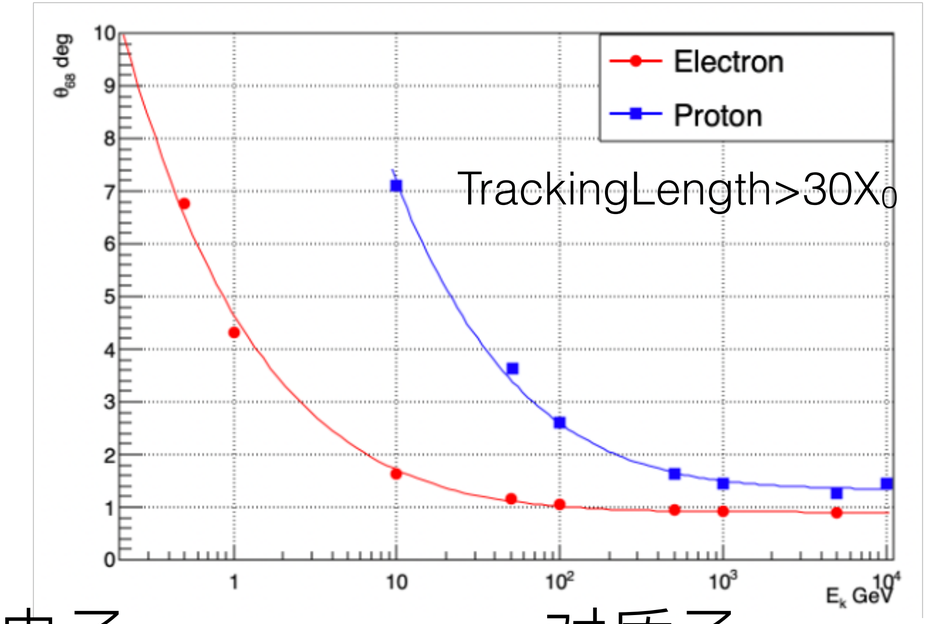
CALO径迹重建

- CALO的簇射主轴可以提供相对粗略的径迹信息，作为全局寻迹的初始条件；
- 另外，在外围探测器无法给出有效信息的情况下(例如受簇射反冲和泄漏的影响严重)作为唯一的粒子径迹信息。

重建方法	描述	优点	缺点
分层求重心线性拟合	投影到两个平面，对量能器每一层求重心(重心法或左右能量比)，再线性拟合	算法简单	只适用于小角度入射的情况
簇射形貌拟合	根据簇射纵向横向发展理论公式去拟合簇射形貌	得到主轴的同时还能得到入射位置、簇射极大位置、粒子初始能量等信息	算法复杂，特别是对斜入射的情况
主成分分析法(PCA)	根据cell位置和能量建立协方差矩阵求特征向量	适用于斜入射	算法本身不区分主轴指向正反，易受噪点干扰，要求簇射完整性
机器学习(CNN)	卷积神经网络建立回归任务	理论上可以获得所有想要的粒子信息，可重建丢失较多信息的事例	

主成分分析法重建簇射主轴：各向同性入射

主成分分析法(PCA)是降维的主流方法，将每个产生能量沉积的晶体单元看作一个样本点（权重为能量），投影簇射主轴子空间，保留最大差异性。

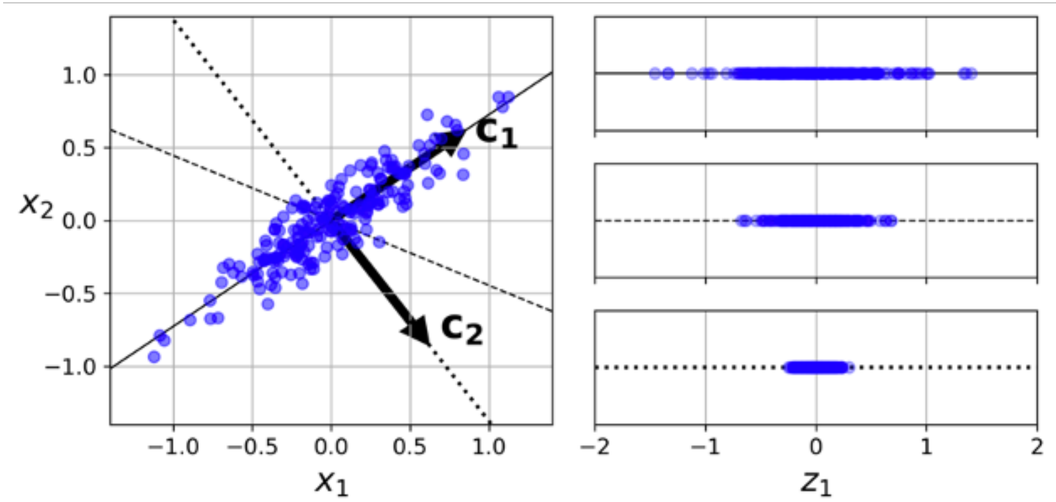


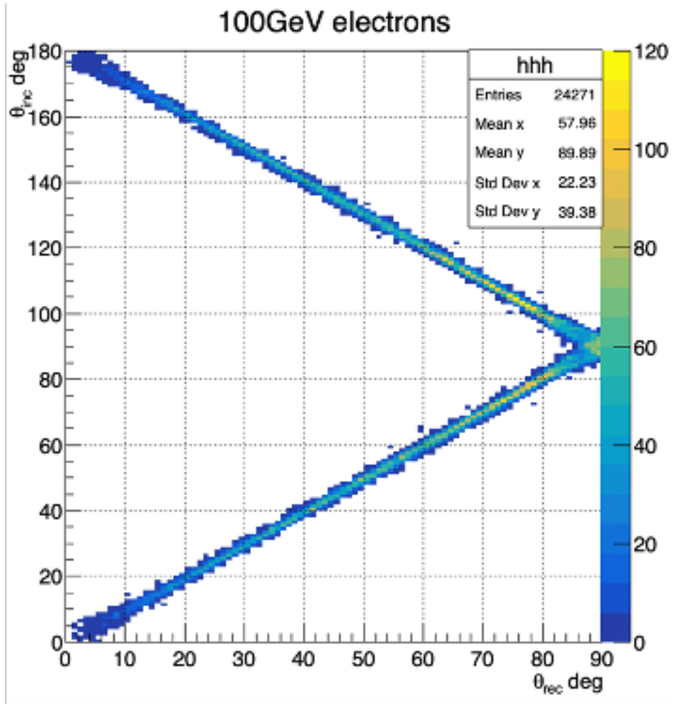
对电子

$$\theta_{68} = \sqrt{\frac{4.57^2}{E_k(GeV)} + 0.904^2}(deg)$$

对质子

$$\theta_{68} = \sqrt{\frac{2.226^2}{E_k(GeV)} + 1.325^2}(deg)$$





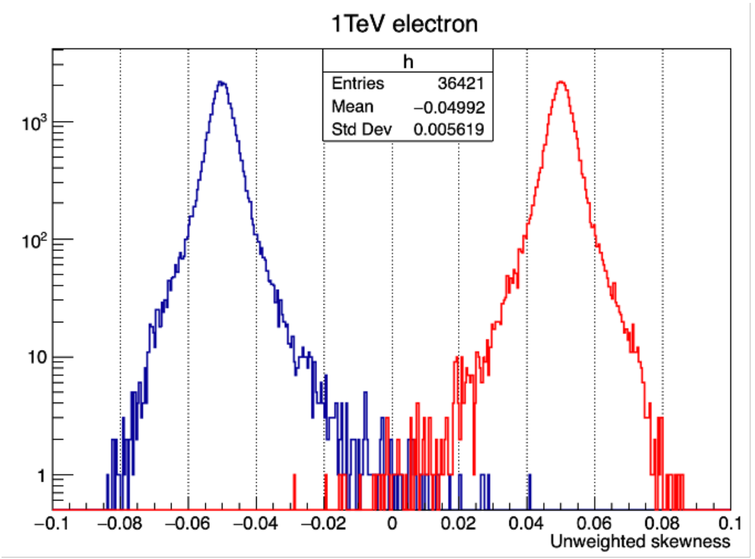
PCA无法区分主轴指向

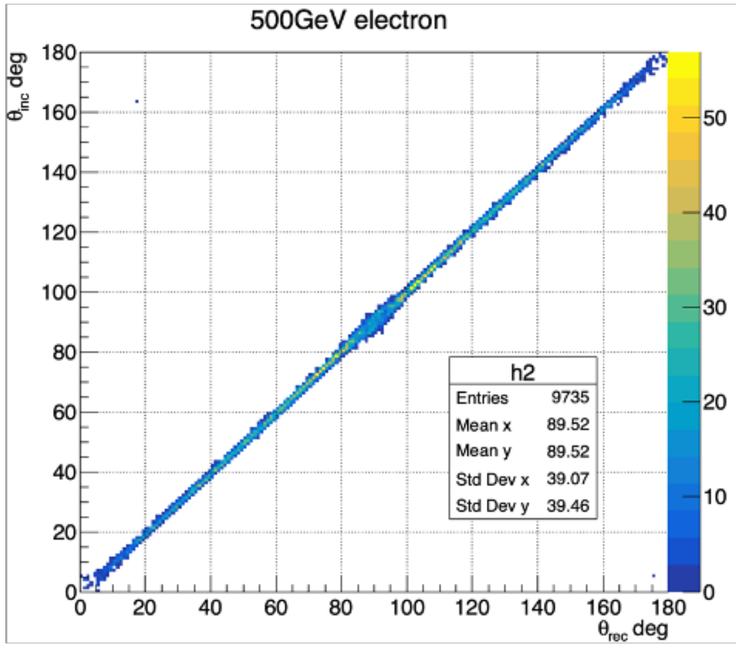
簇射主轴方向甄别

对电子，只需要利用电磁簇射纵向发展特征——上升快下降慢的不对称性就能获得较好的效果。

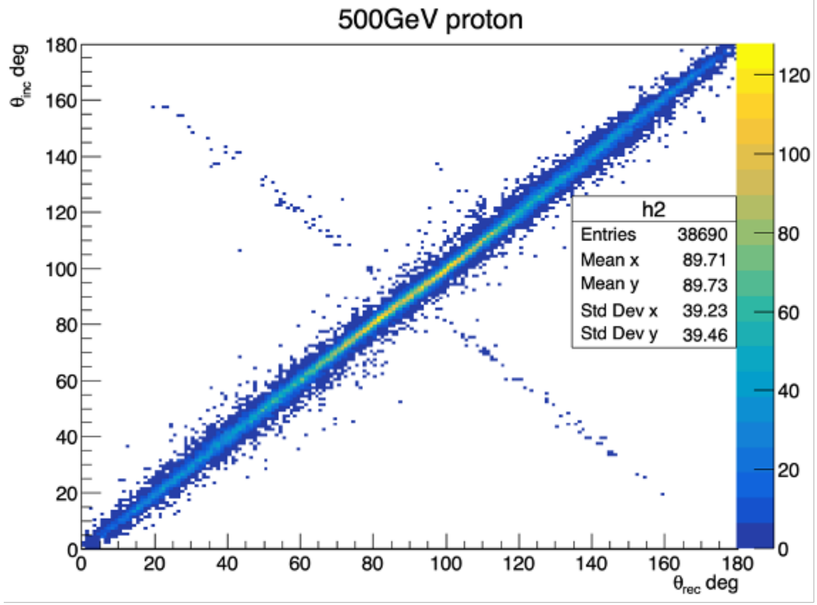
$$var = \frac{\sqrt{n}\sum (X_i - \bar{X})^3}{(\sum (X_i - \bar{X})^2)^{3/2}}$$

轴向无权偏度





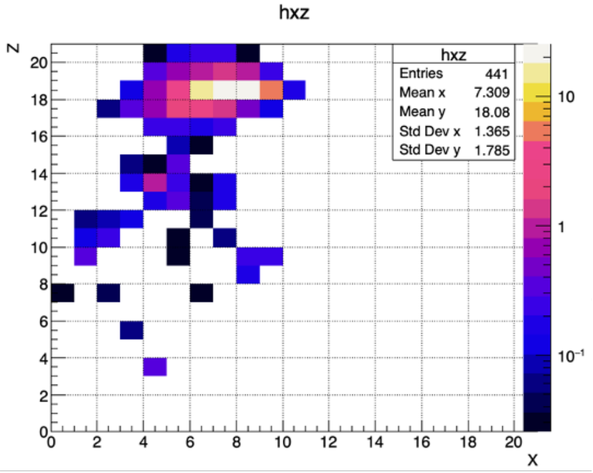
对质子，强子簇射涨落大，除了用以上变量外，还需要加入更多描述纵向发展的信息



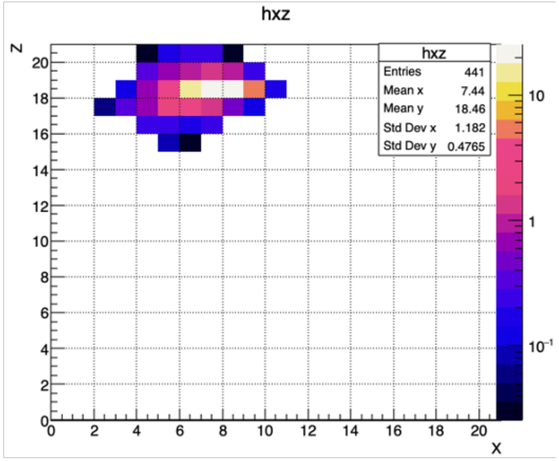
使用4个变量的BDT的效果：对10GeV～20TeV质子(index=-2.7)，误判为0.3%@94%效率

聚类(Clustering)：无监督学习

聚类算法的必要性：轨道环境复杂，本底计数率高，量能器中可能存在多个簇射团；读出系统涨落产生噪点。而PCA算法要求干净的主簇射团。



一个50GeV电子被一个2GeV质子+一个500MeV质子污染



运行DBSCAN聚类算法后得到干净的簇射团

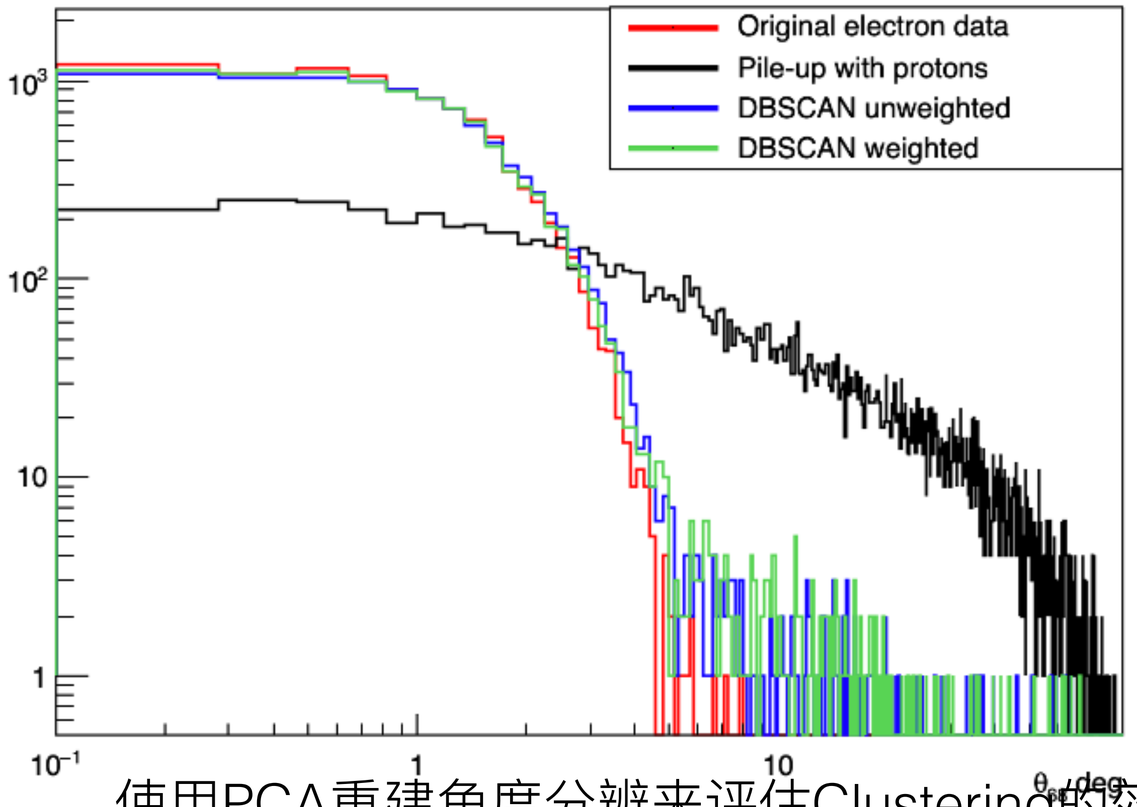
DBSCAN：Density-Based Spatial Clustering of Applications with Noise

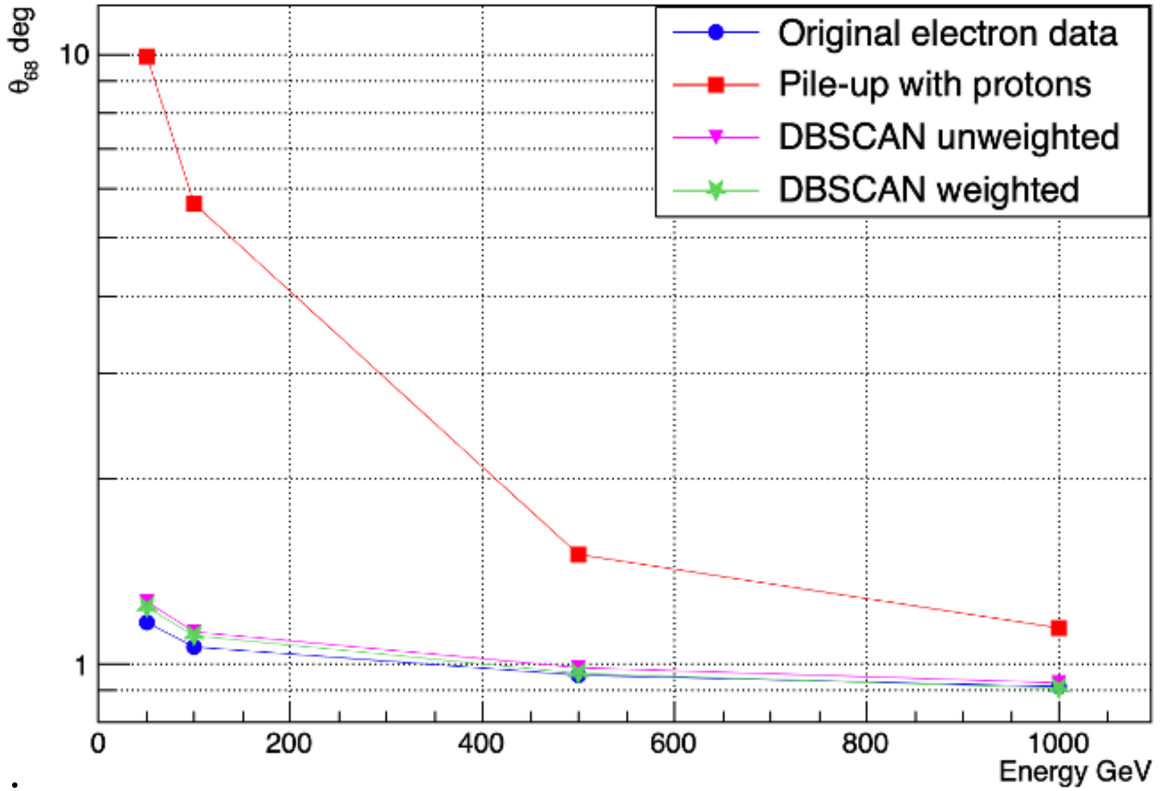
—种基于密度检测的聚类算法，可检测噪点，适用于密度平缓变化的情况

只需要两个参数：eps(扩展cluster的特征距离)，N(构成cluster的最小实例数)

对HERD CALO的模拟结果表明，eps=1, N=5 是较合适的参数。

Clustering的效果



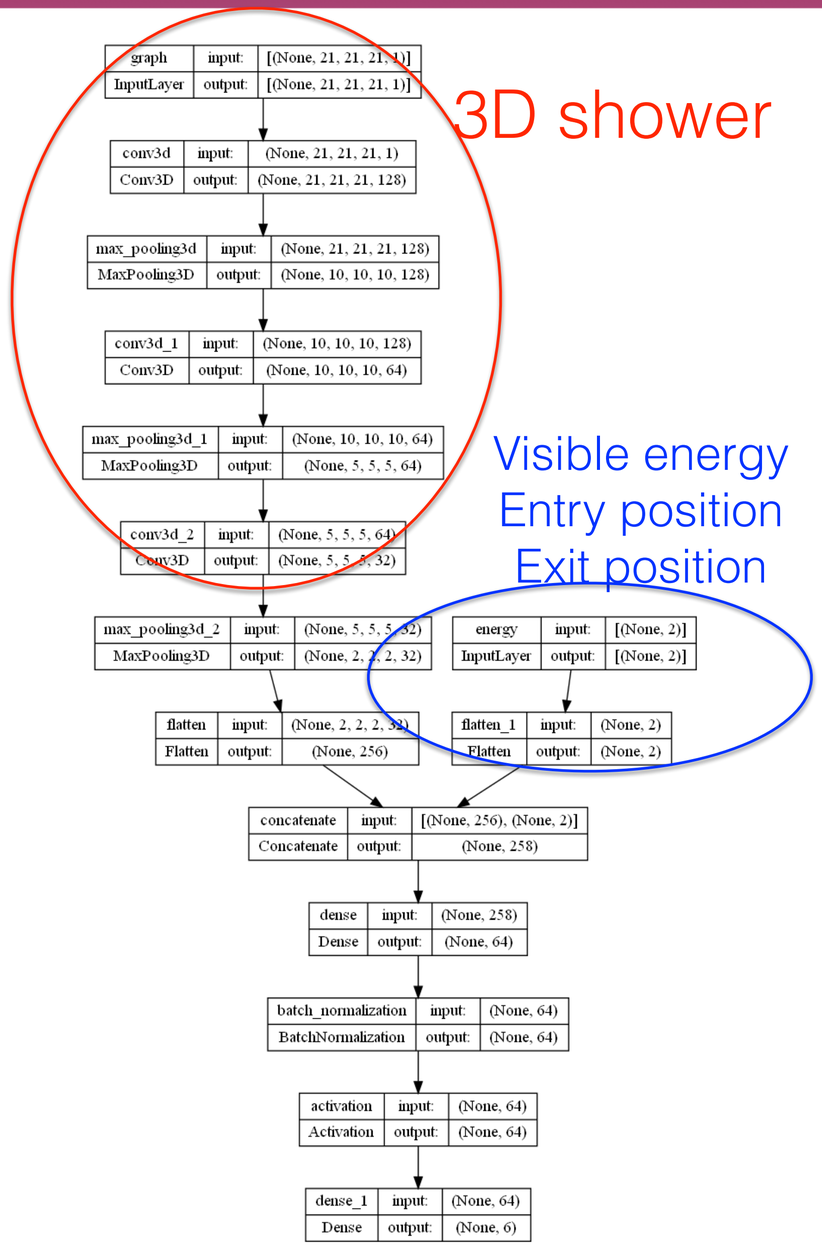


使用PCA重建角度分辨来评估Clustering的效果：

- 噪声存在的情况下，角分辨会变得很差，给径迹重建带来困难；
- 运行Clustering算法后，得到干净的簇射团，角分辨与无污染的事例接近；
- 有部分污染与主簇射团重叠的事例重建效果较差。

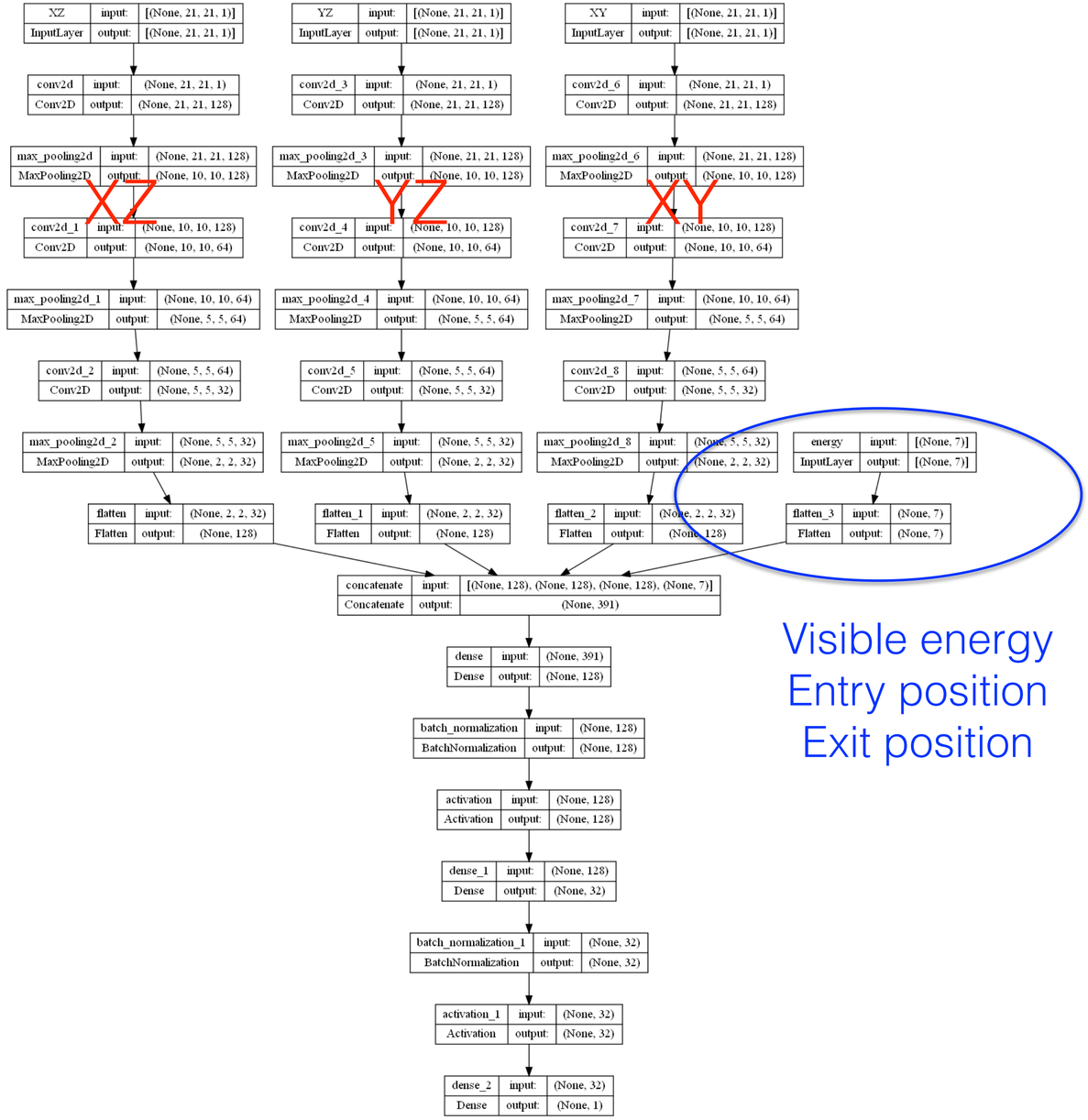
后期的研究需要考虑：阈值的选取以及其对能量重建的影响

CNN方法应用于粒子径迹和能量重建



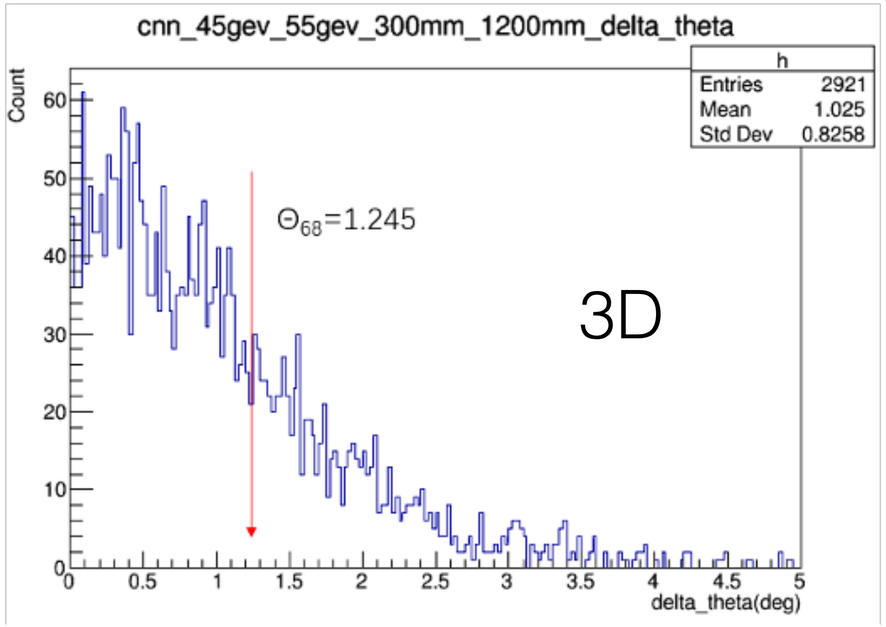
3D shower

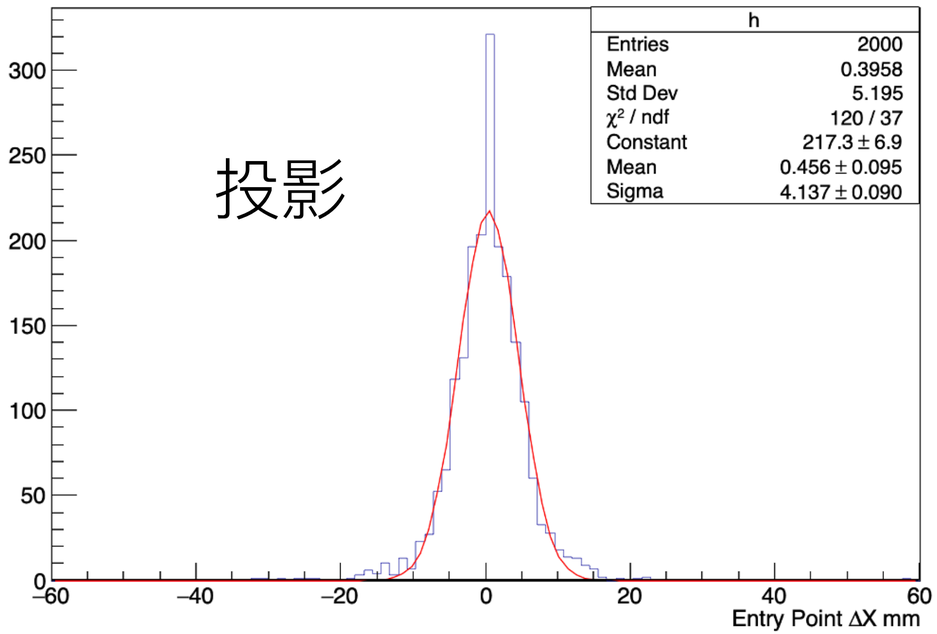
Visible energy
Entry position
Exit position

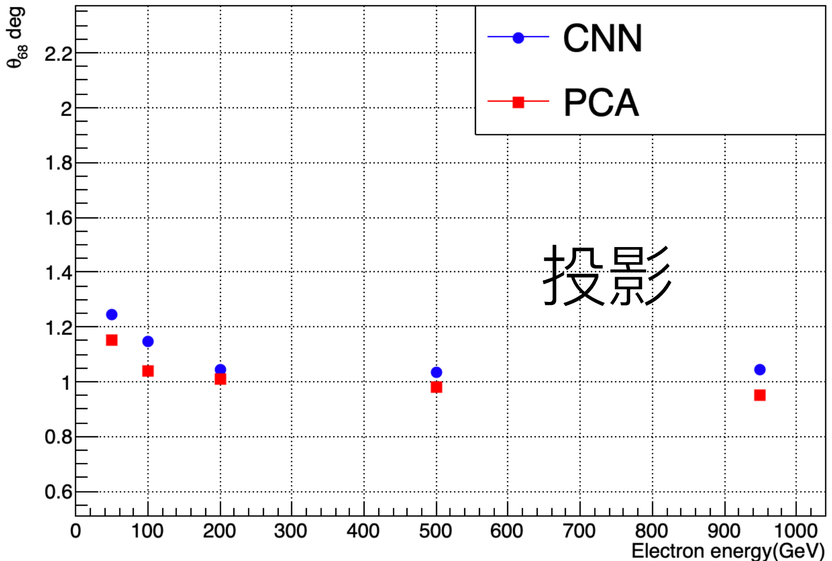


Visible energy
Entry position
Exit position

CNN重建径迹







CNN径迹重建性能与PCA很接近

CNN重建能量初步尝试

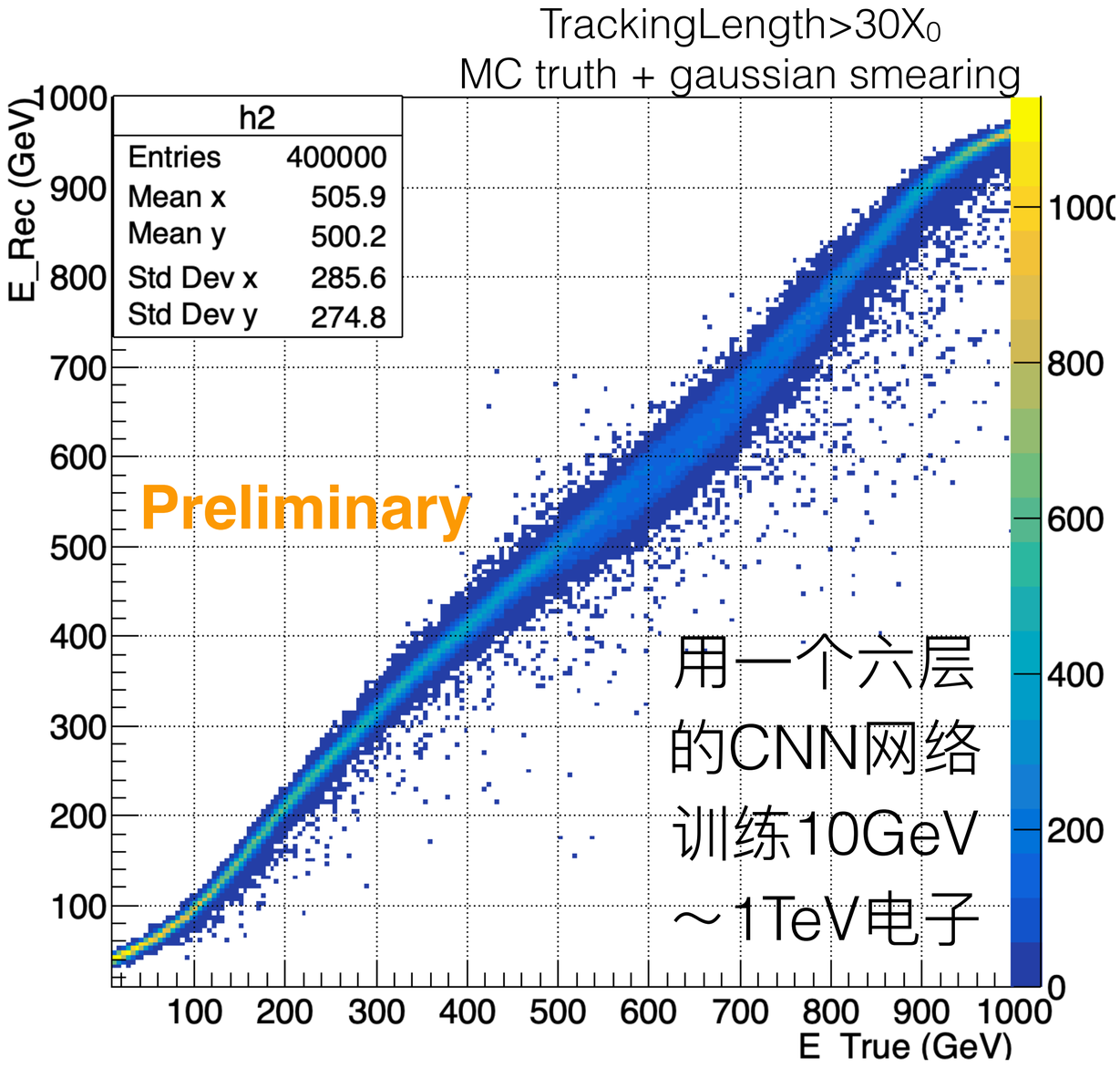
电磁簇射理论公式：

$$\frac{dE}{dt}(t) = E_0 \frac{(\beta t)^{\beta T_0} \beta e^{-\beta t}}{\Gamma(\beta T_0 + 1)}$$

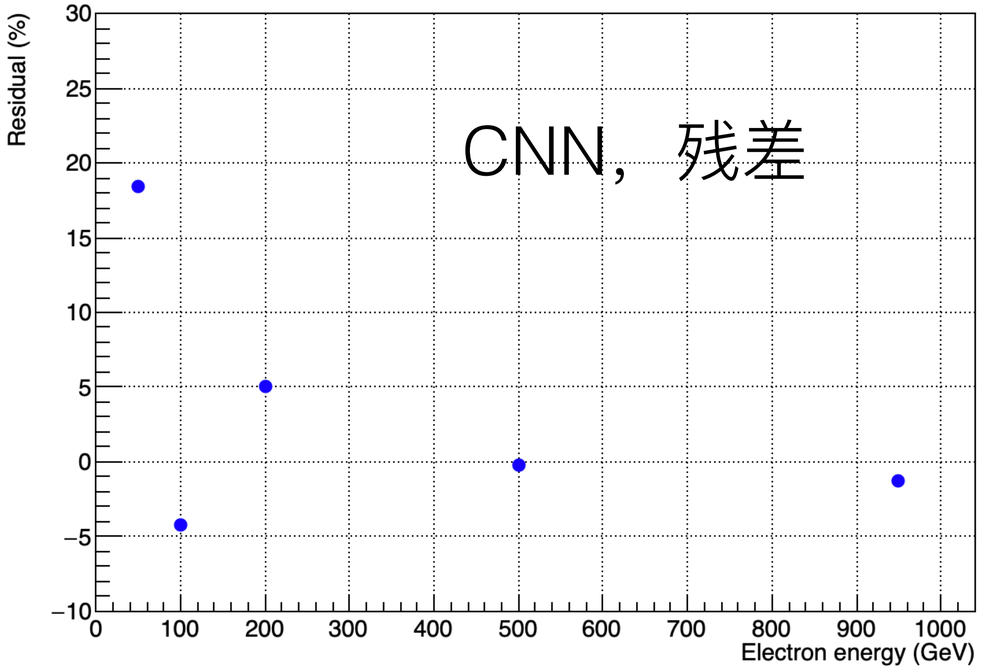
$$\frac{d^2 E_{layer}(x,y)}{dxdy} = \frac{E_{layer}}{\pi}\left[\frac{pR_C^2}{(x^2+y^2+R_C^2)^2} + \frac{(1-p)R_T^2}{(x^2+y^2+R_T^2)^2}\right]$$

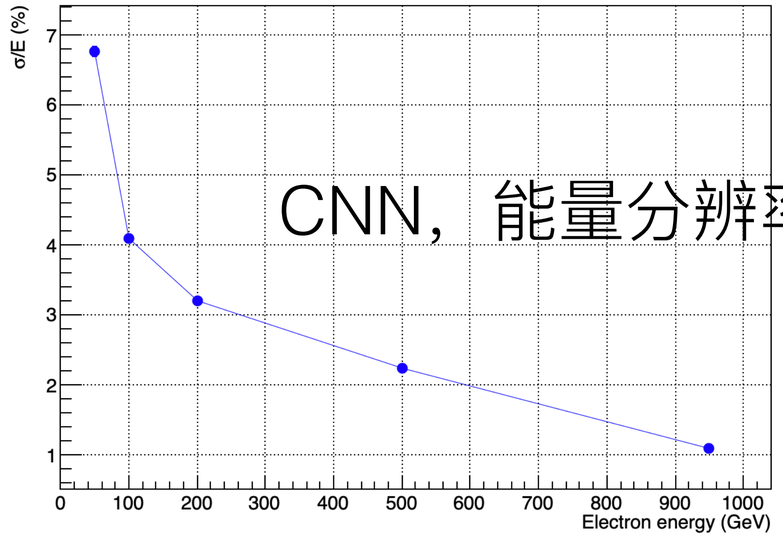
对于各向同性入射粒子，用理论公式去拟合簇射会变得很复杂；

CNN: 自动寻找簇射形貌的特征并建立簇射图像与粒子信息的对应关系



CNN重建能量初步尝试

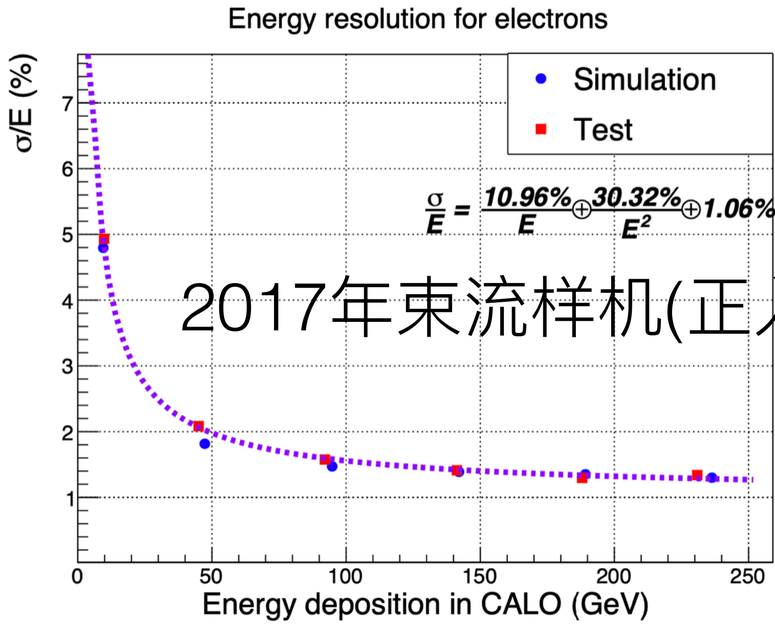




初步尝试用简单CNN网络训练模拟数据重建10GeV～1TeV电子能量：

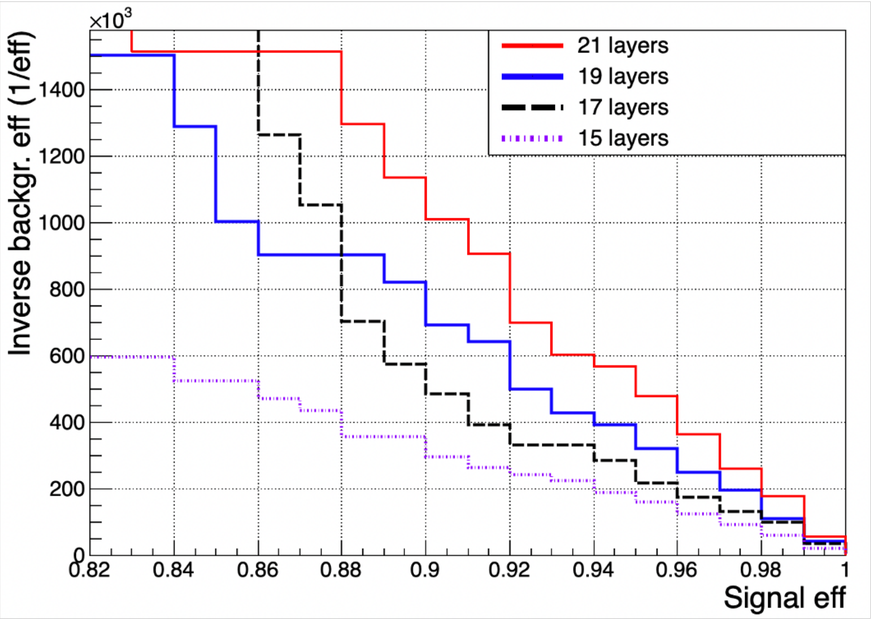
- 训练得到的模型还无法适用于全能段；
- 能量分辨率与样机(厚度26X_0)束流实验及模拟正入射相比有差距

需要进一步优化

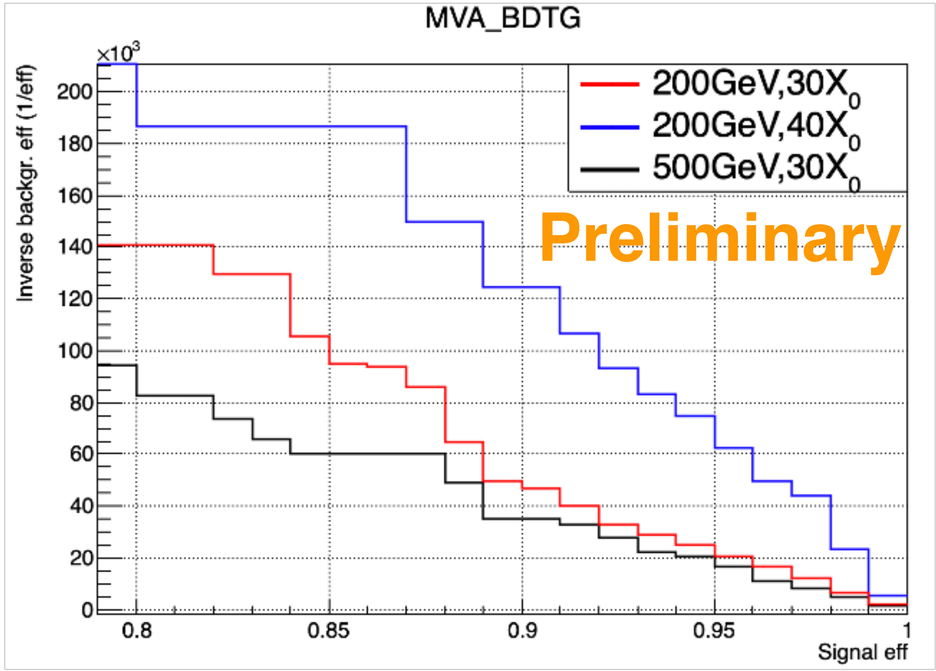


e/p鉴别

多变量分析决策树方法：从正入射到各向同性入射



分层获取描述簇射形貌的参数，总共使用了47个变量，得到e/p鉴别性能为10^6@90%eff.，1TeV



在PCA获取主轴的基础上，使用各个方向上方差、偏度、峰度等共14个变量，尽量避免沿轴向人为分层带来的几何误差，>40X_0挑选条件得到的鉴别性能为1.2×10^5@90%eff. 200GeV该挑选条件下效率为～50%

从BDT到CNN或DNN：期待更好的鉴别性能

未来工作计划

中短期：

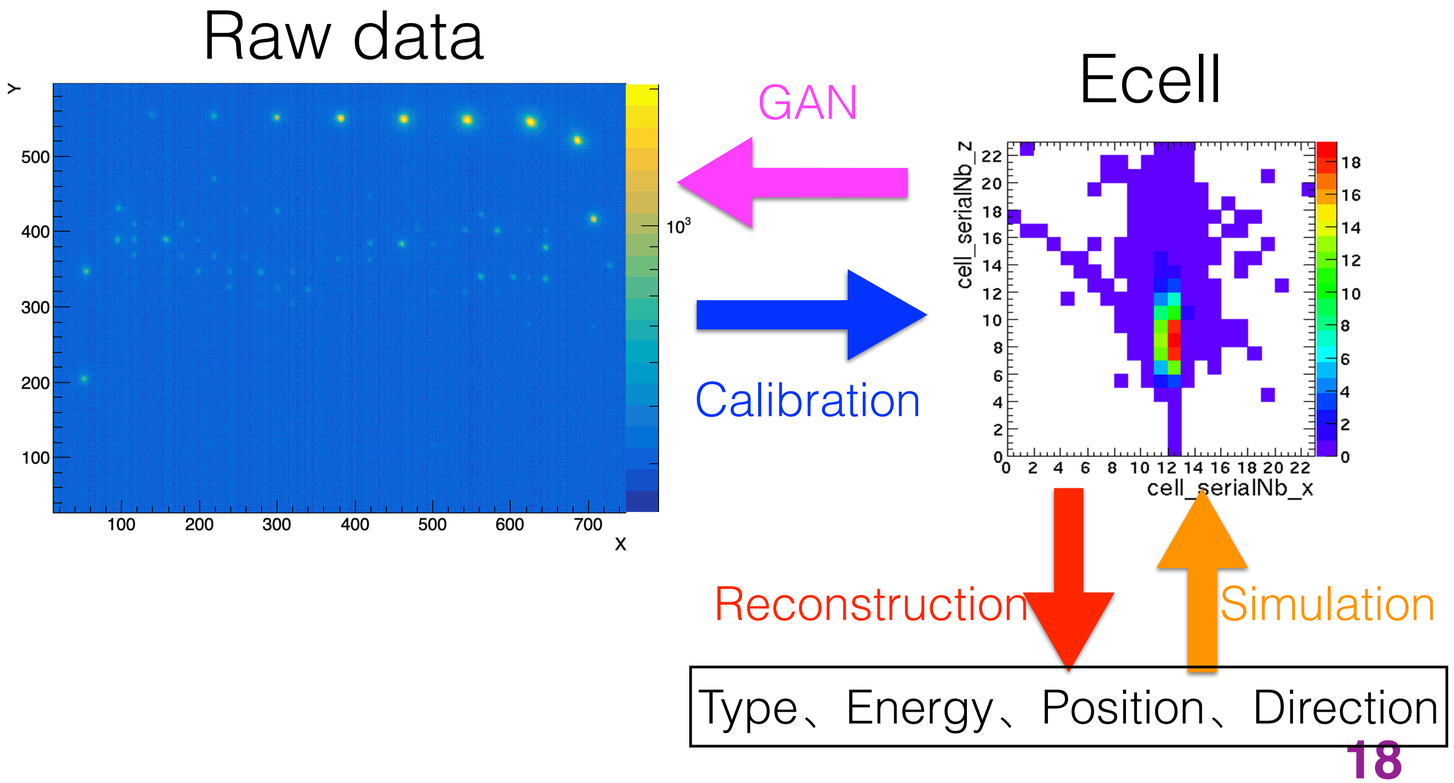
- 优化径迹重建和能量重建的网络(电子、质子重建)；
- 利用通过训练模拟数据得到的模型去分析束流实验数据，验证模型的泛化能力及准确性；

长期：

- 研究DNN和CNN应用于PID；
- 研究神经网络Clustering算法；

未定：

- 增强相机数字化



人力投入

目前为起步阶段：1职工+1学生
对机器学习的新技术进一步了解后，明确项目对这些技术的需求，然后再投入人力

总结

- 针对HERD三维成像量能器的数据分析方法的研究，机器学习占有重要地位，可更深入挖掘在径迹重建、能量重建、粒子鉴别方面的潜力；
- 研究并实现了一些相对成熟的算法，包括PCA重建径迹、DBSCAN聚类、BDT做e/p鉴别等等；
- 初步研究了CNN在径迹重建和能量重建上的应用。

谢谢各位老师，请指正