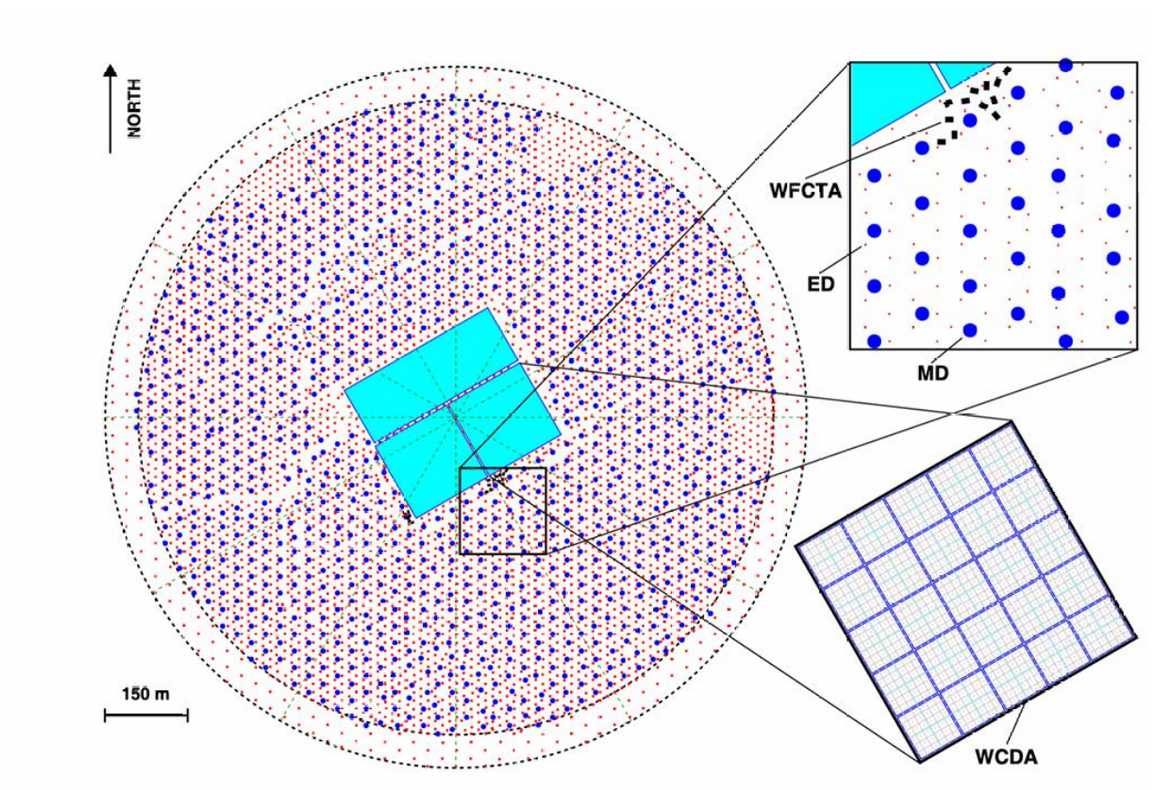
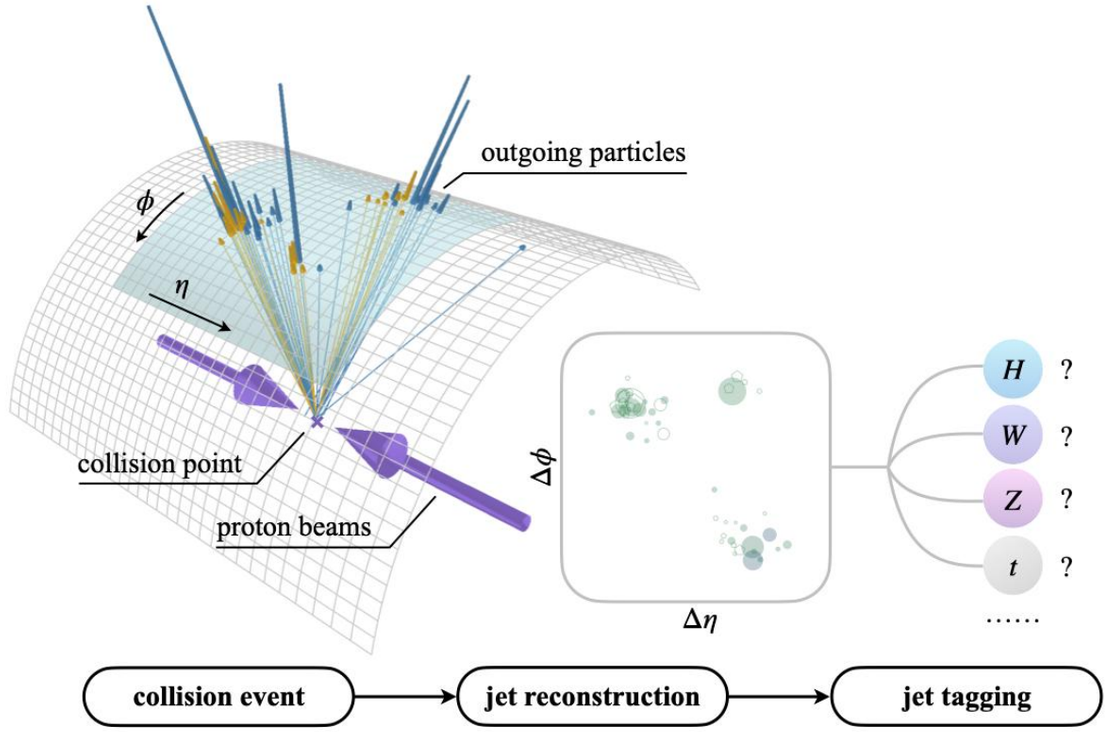
机器学习方法在LHAASO和CEPC上的应用

汇报人：杨超

中国科学院高能物理研究所
Institute of High Energy Physics Chinese Academy of Sciences

工作概述



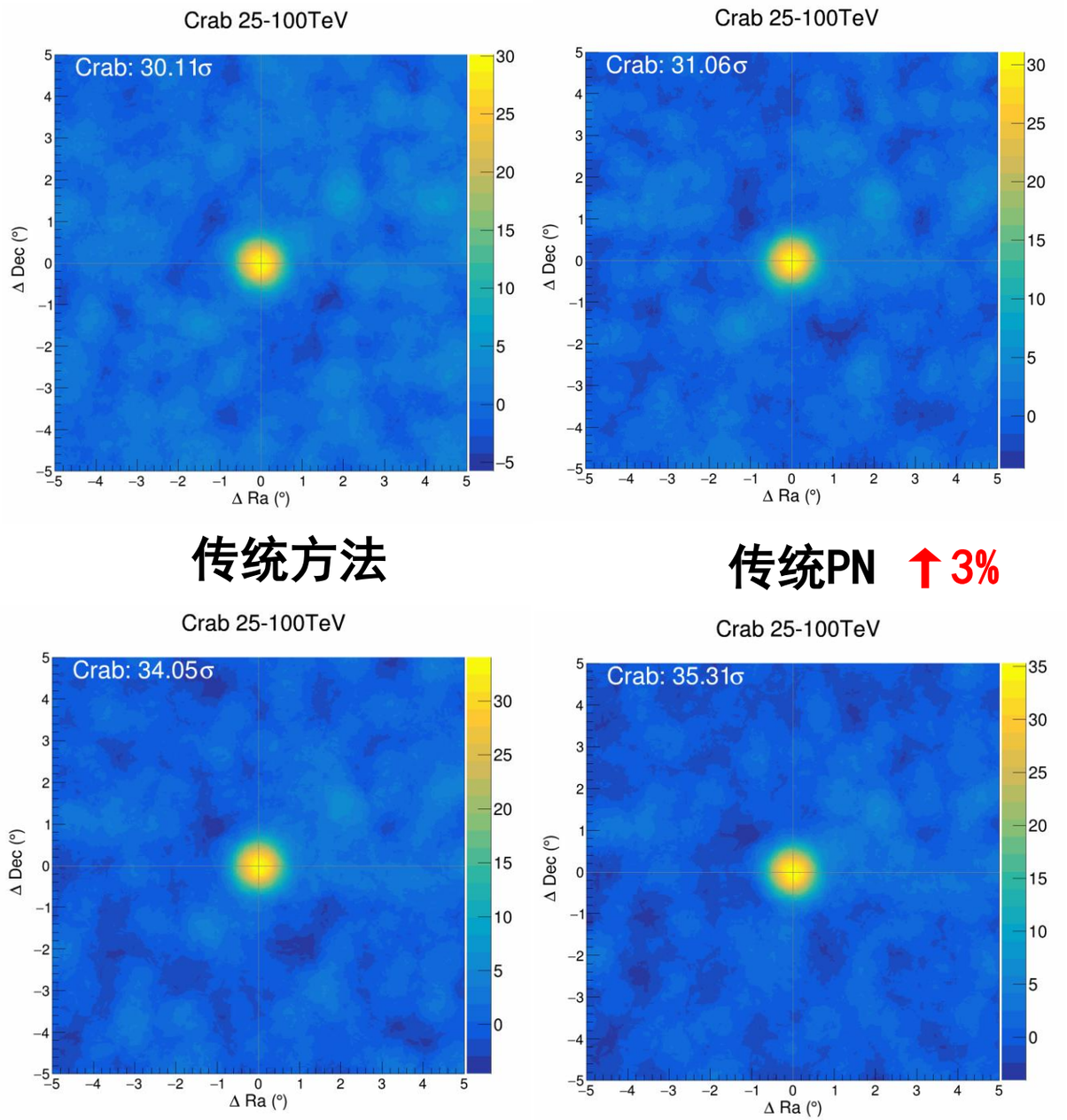


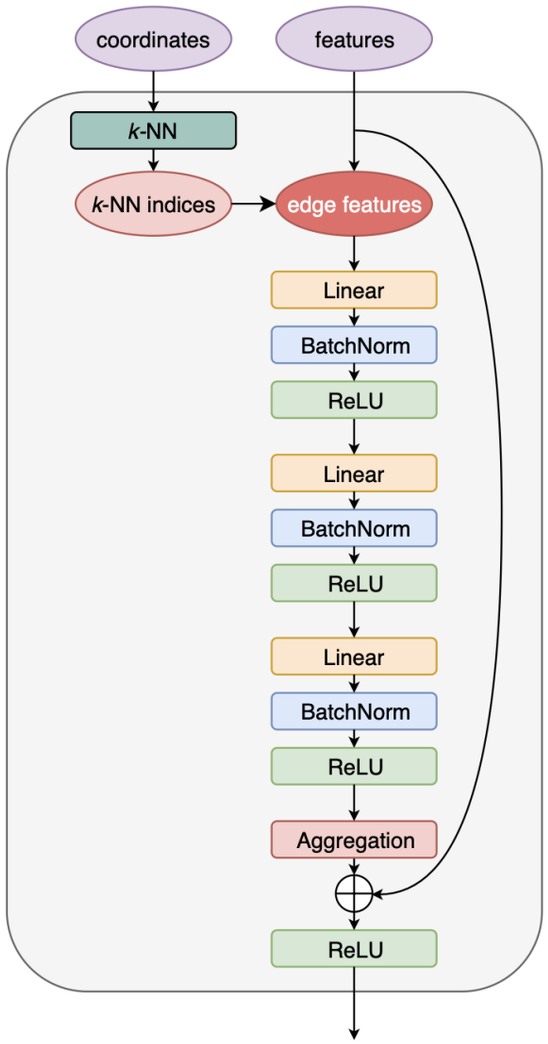
- LHAASO-KM2A上$\boldsymbol{\gamma/p}$鉴别
- 深度学习模型代替KM2A中光子追迹模拟

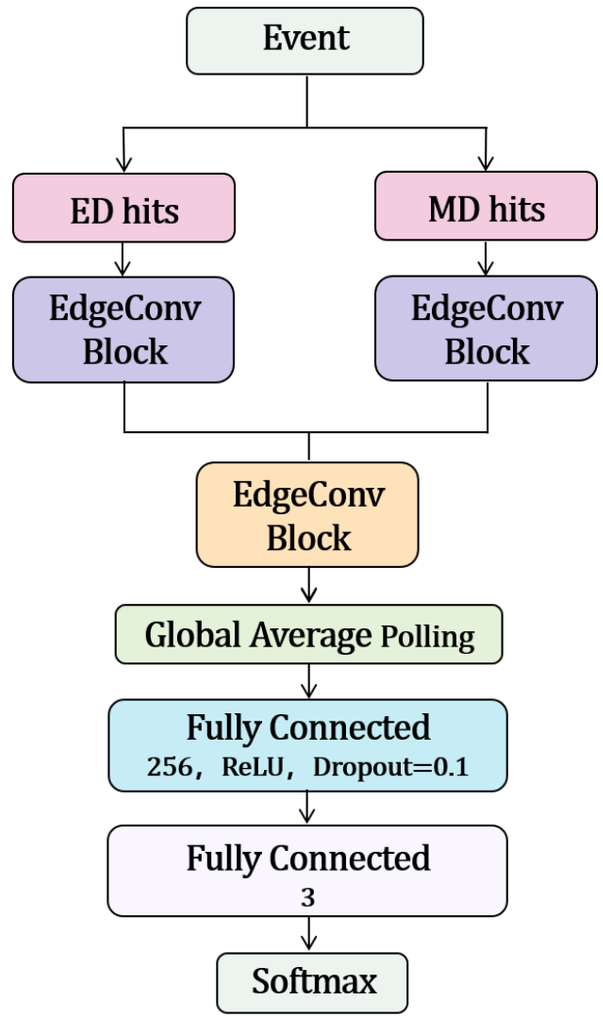
- From Leaves to the Tree（深度学习用于末态粒子逐级重建）

γ/p鉴别--模型改造

- 为了充分利用两种探测器的信息，我们基于ParticleNet设计了探测器分流输入的模型Dual Stream PN　Crab 0101~0130



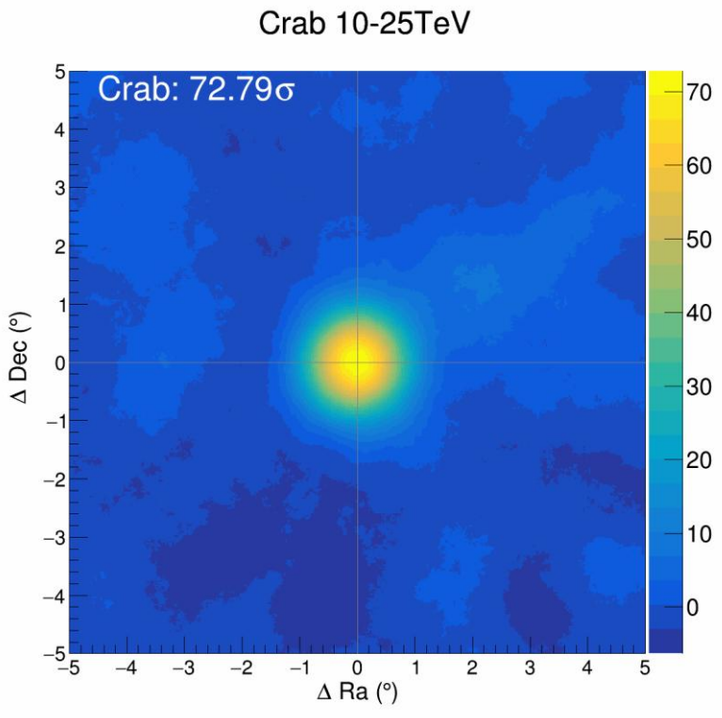


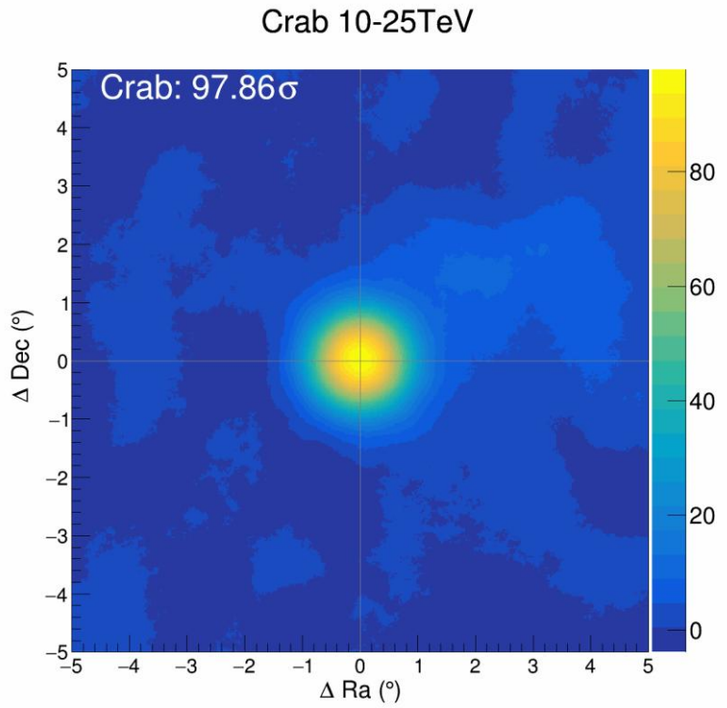


γ/p鉴别--模型改造

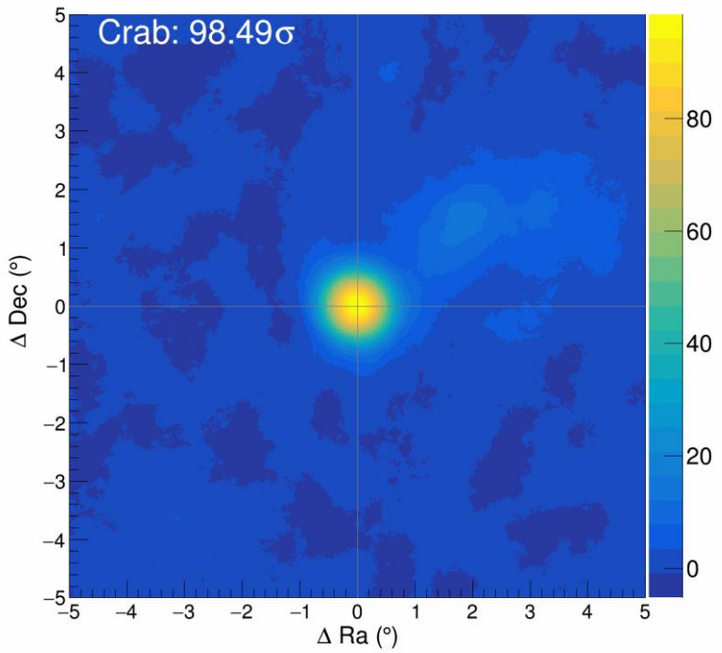
- 基于Dual Stream PN，我们预测了Crab 2024全年的实验数据

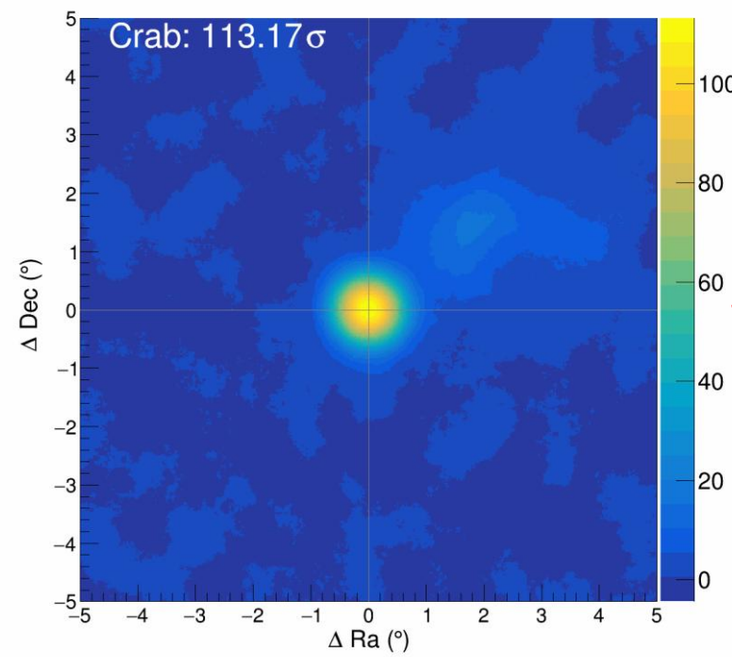
	Nsig	Nb	显著性
ML (10-25)	12977	12703	97.86
Baseline (10-25)	15989	40625	72.79
ML (25-100)	3977	291	113.2
Baseline (25-100)	3381	330	98.49





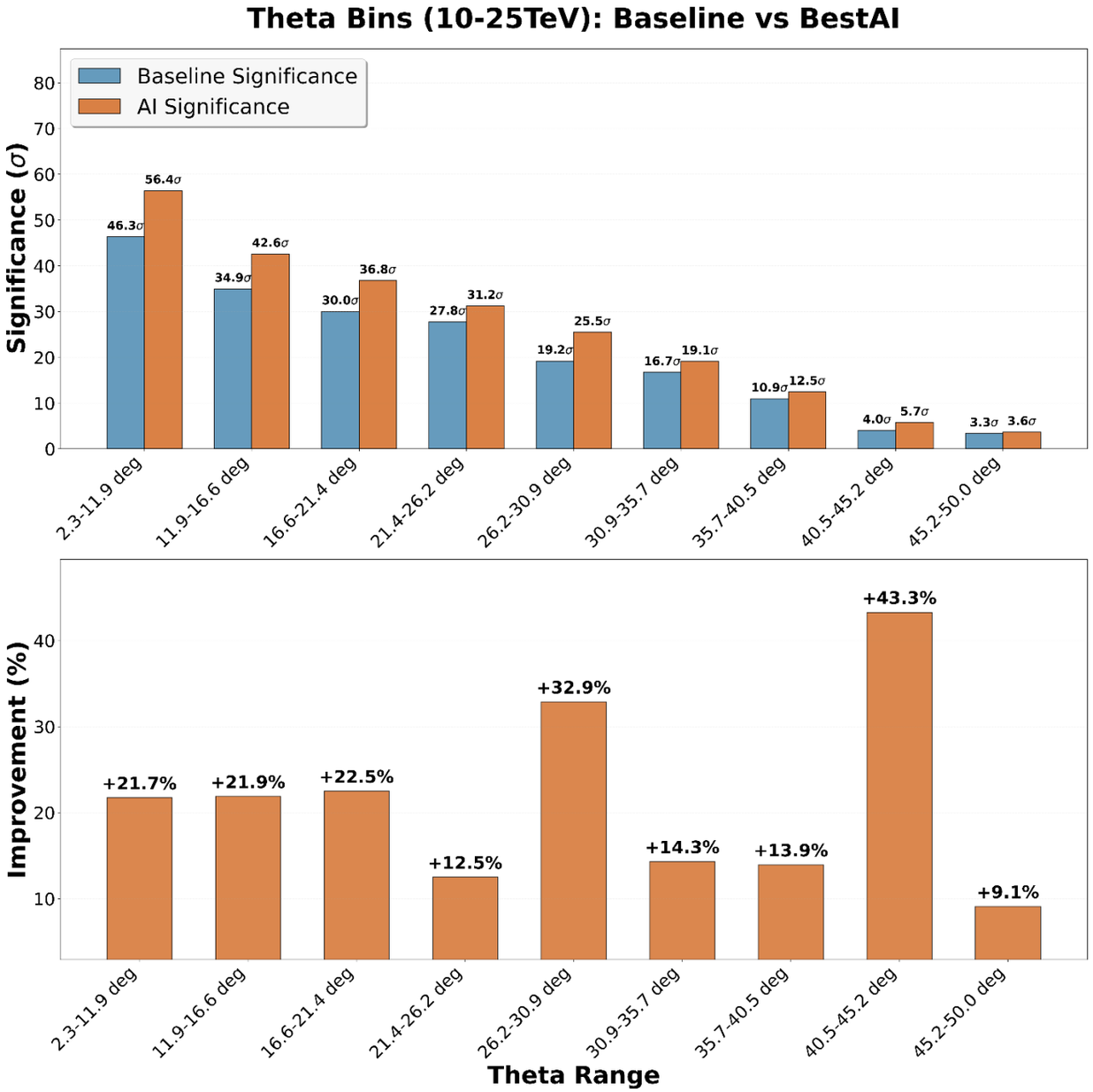
↑34.4%

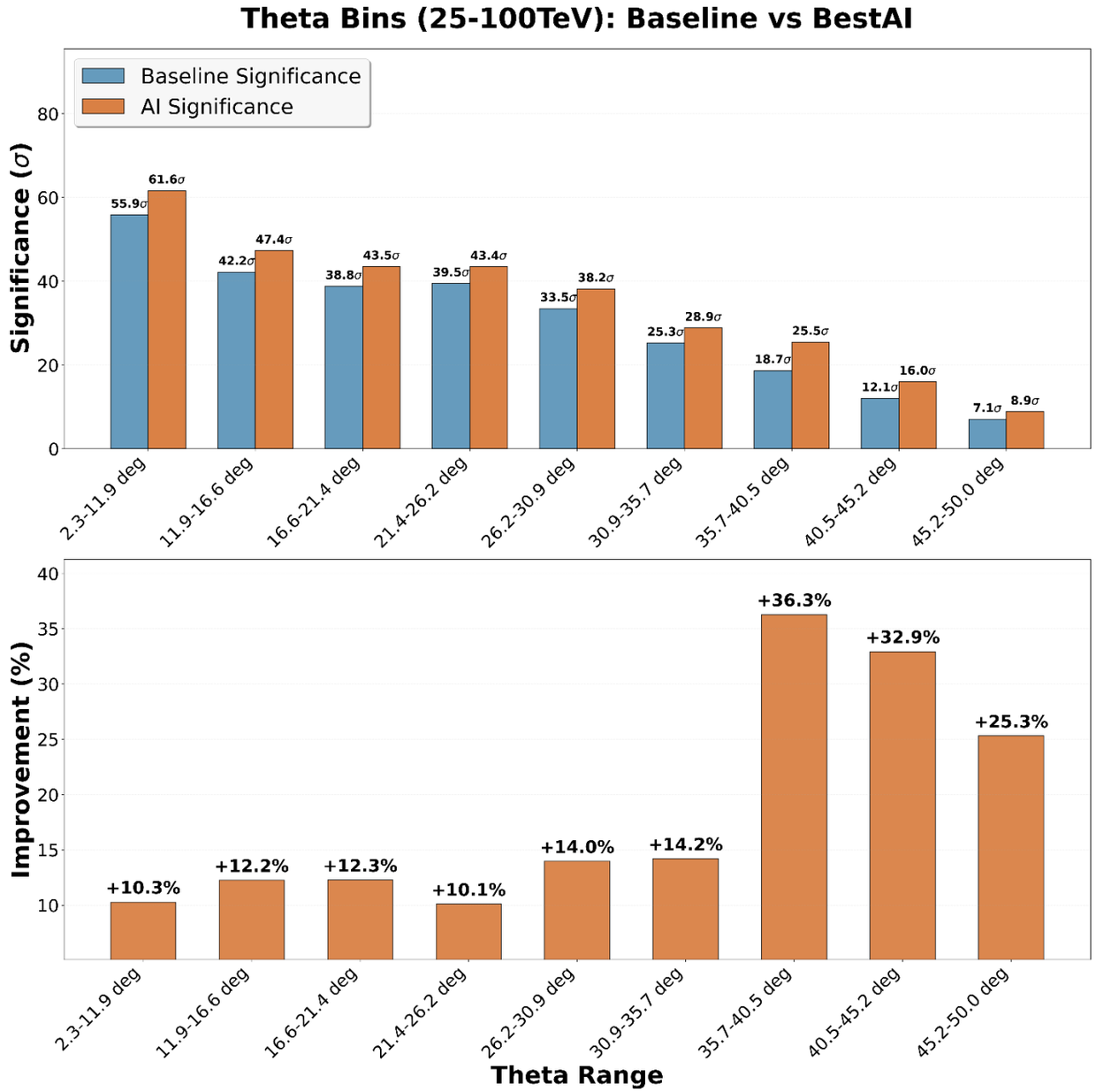




↑14.9%

γ/p鉴别--模型改造

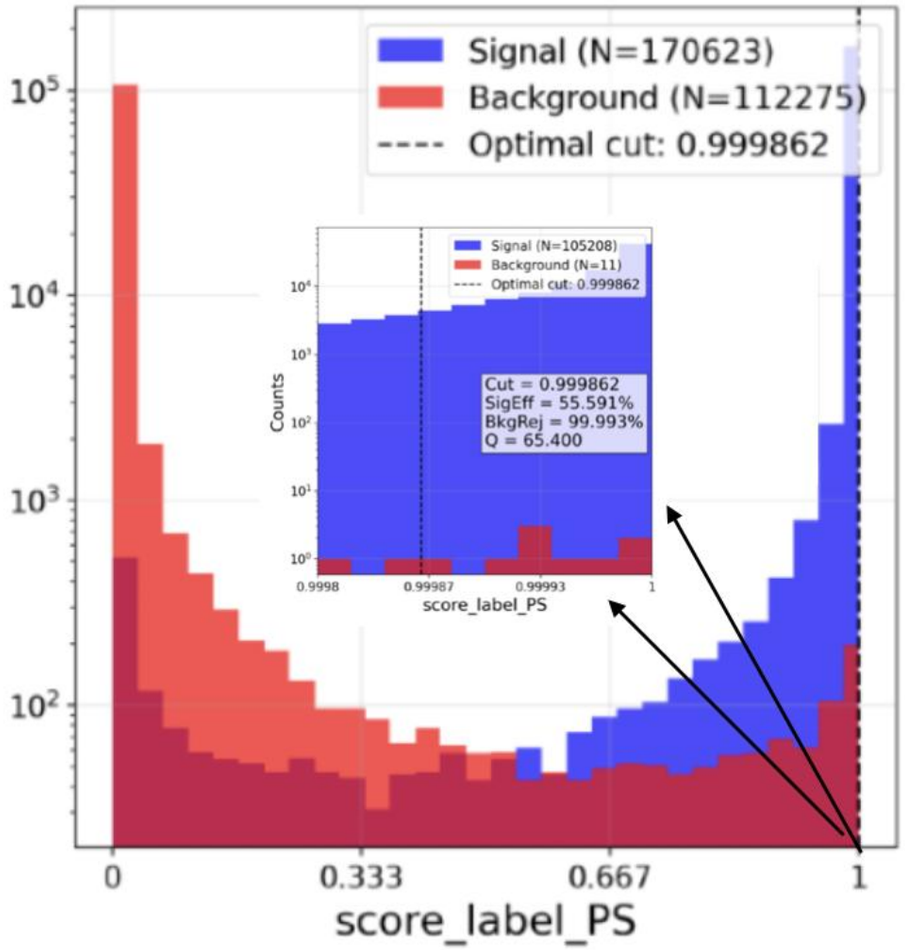


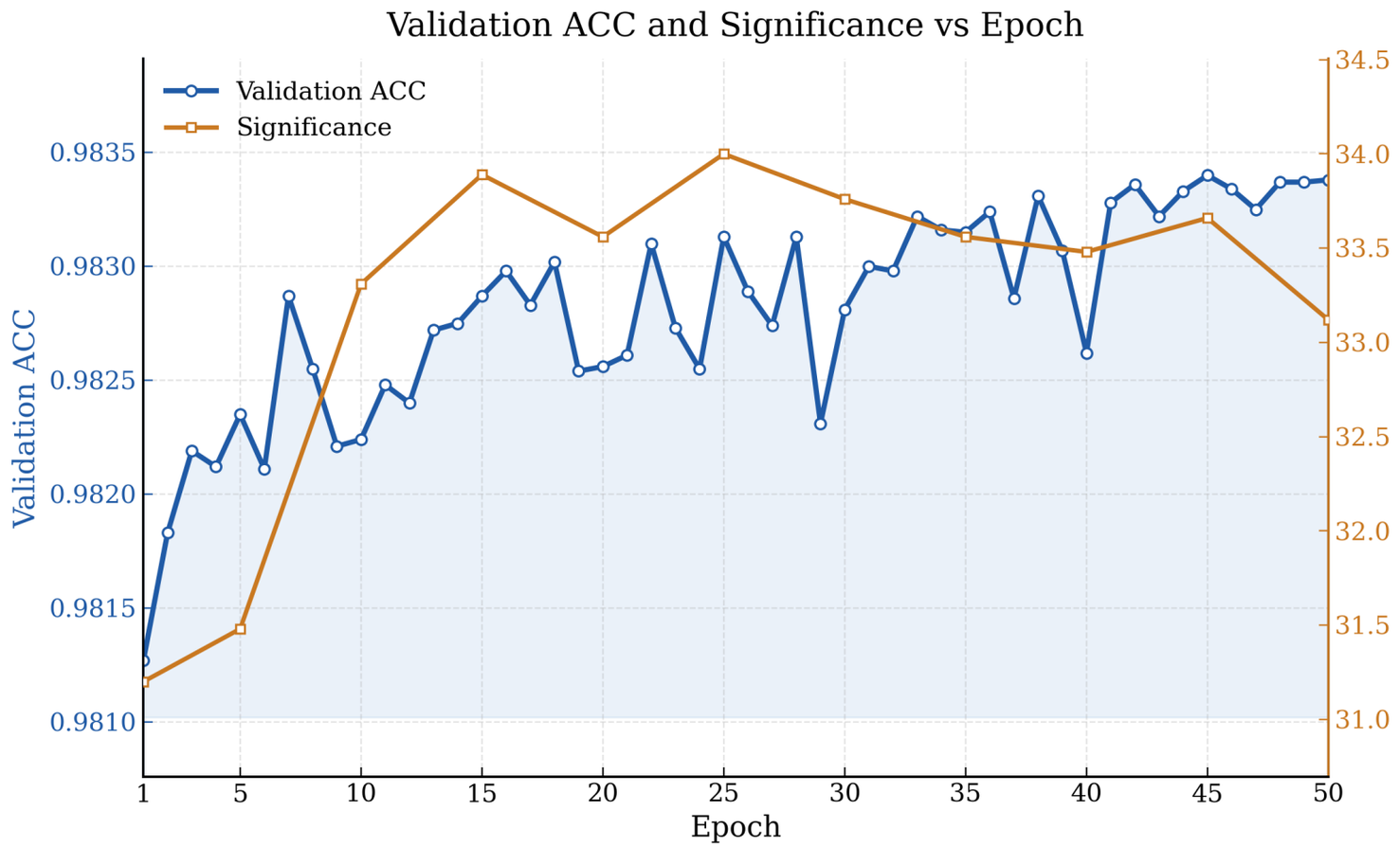


- 利用**Dual Stream PN**对重建天顶角进行differential分析
- 25~100TeV，大天顶角事例提升比较明显，短板呈现在小天顶角事例

γ/p鉴别--半监督学习

MC上提升>50%

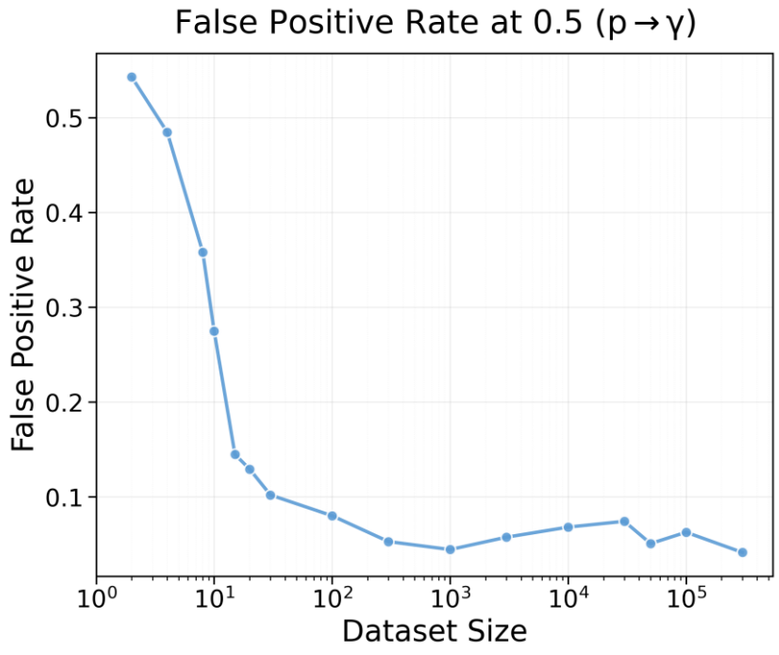


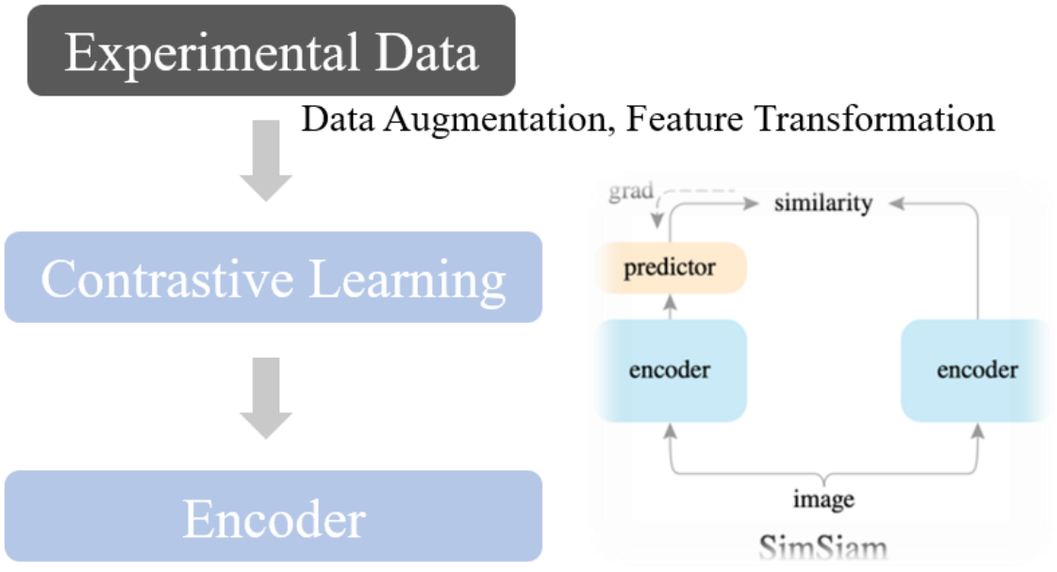


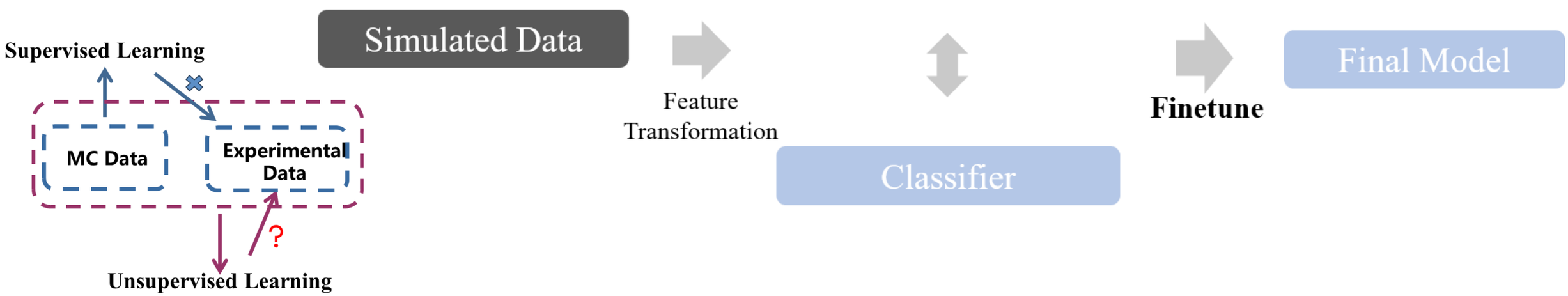
模型对MC过拟合示意图

- 由于蒙卡和Data的差异，有监督模型在实验数据上性能会打折
- 随着模型训练，模型在MC验证集上的泛化性能继续提高，但在实验数据上的效果开始下降

γ/p鉴别--半监督学习

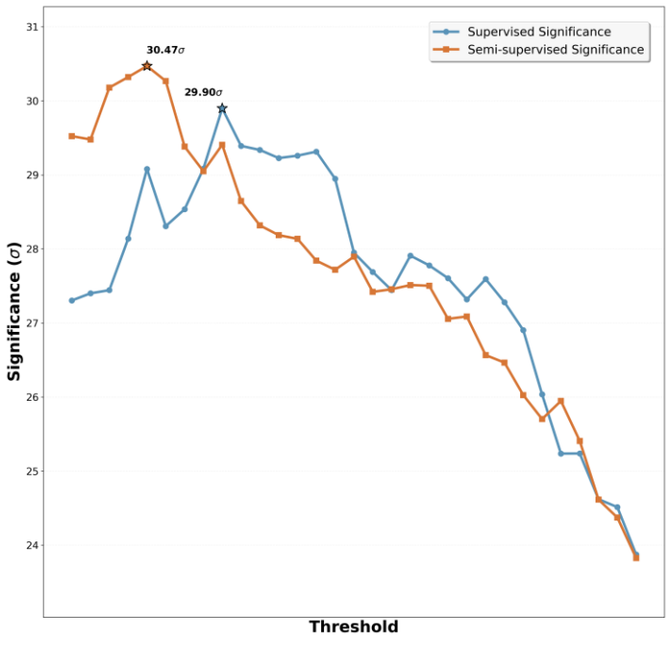


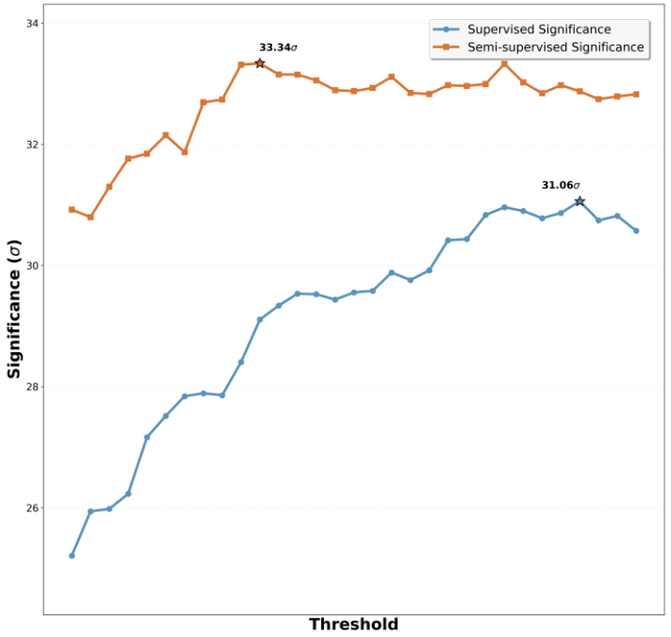


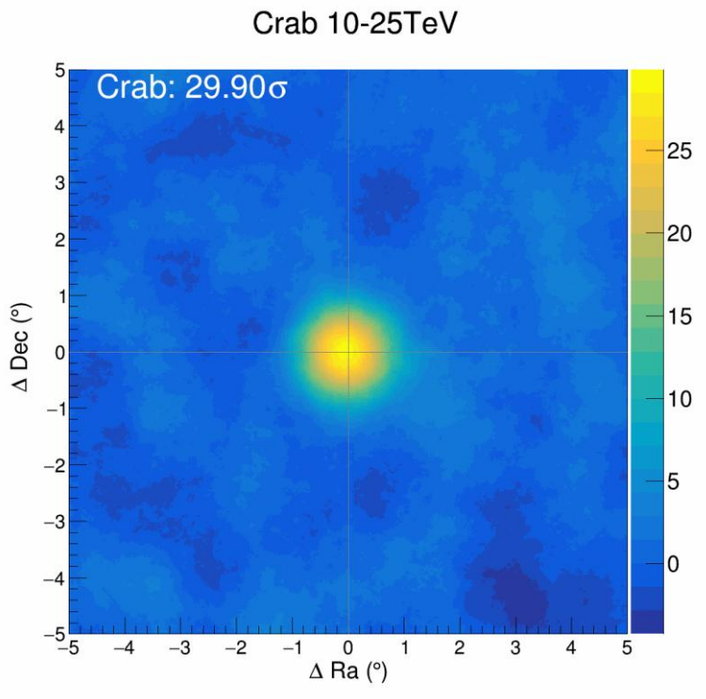


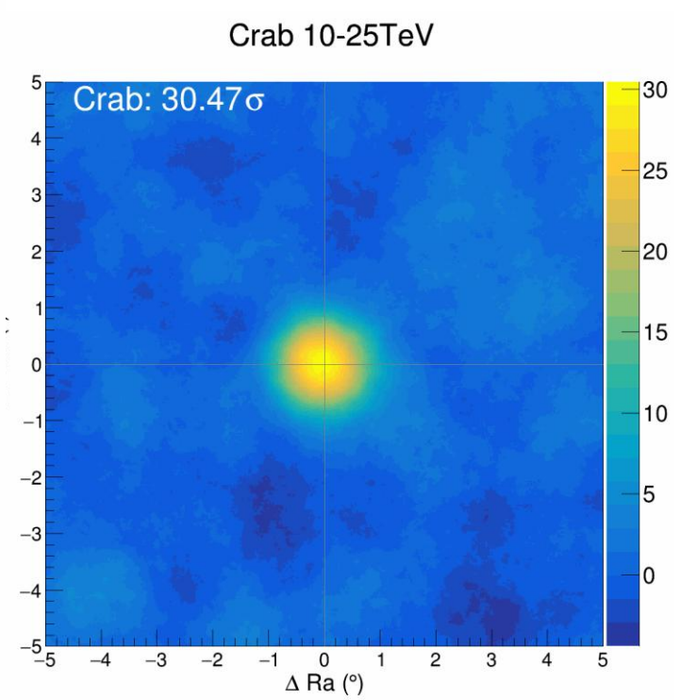
- 由于蒙卡和Data的差异，有监督模型在实验数据上性能会打折
- 让模型先经过大量实验数据的预训练，下游再用有标签的模拟数据进行微调，期待增强模型在实验数据上的泛化能力

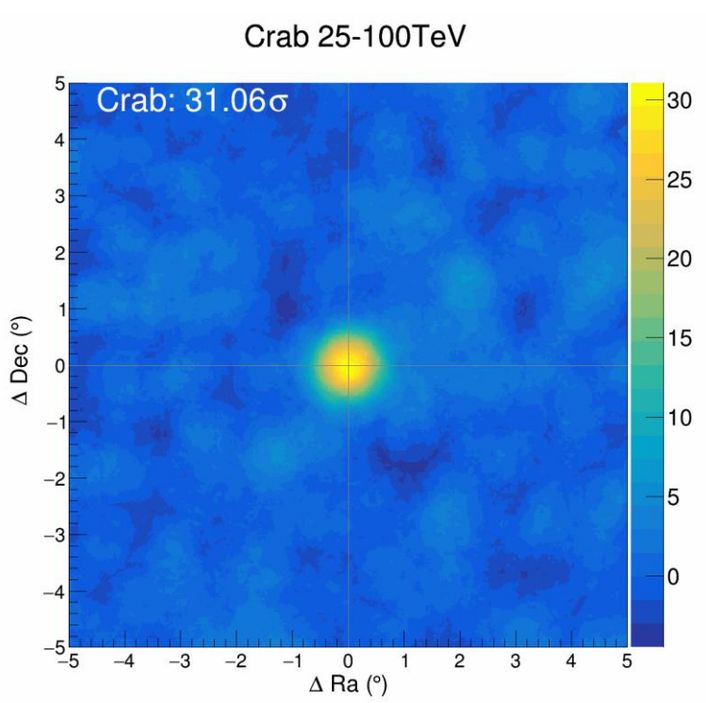
γ/p鉴别--半监督学习

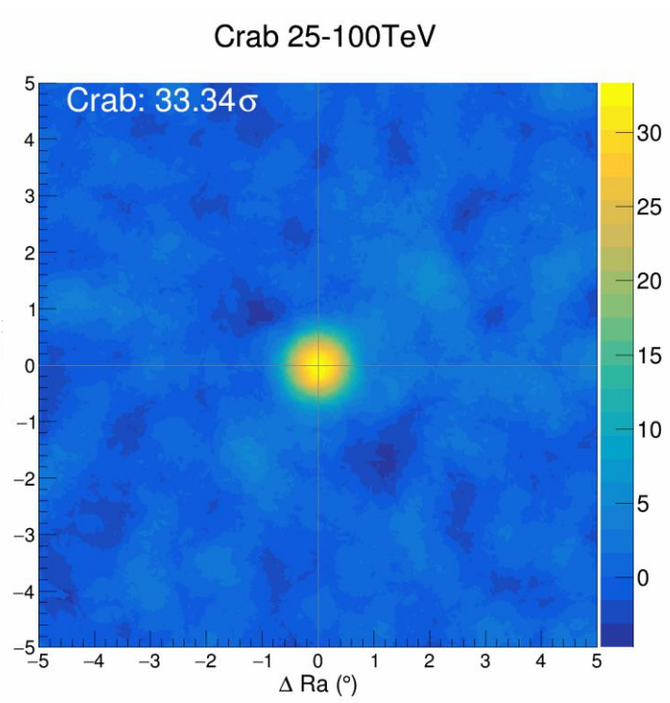








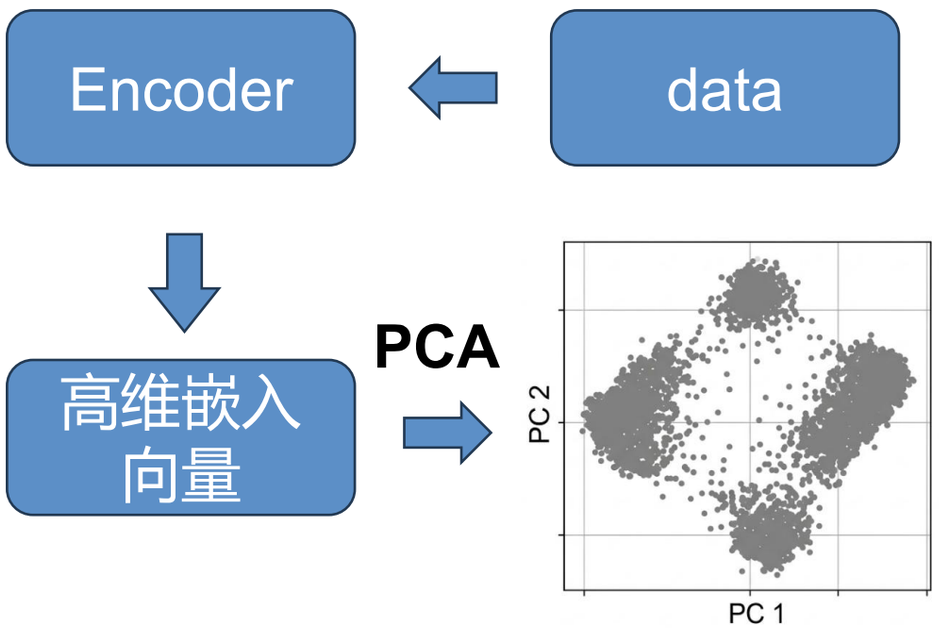




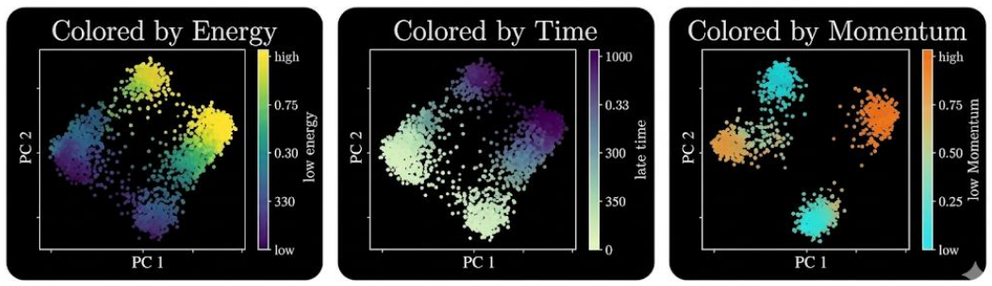
有监督学习 **半监督学习**

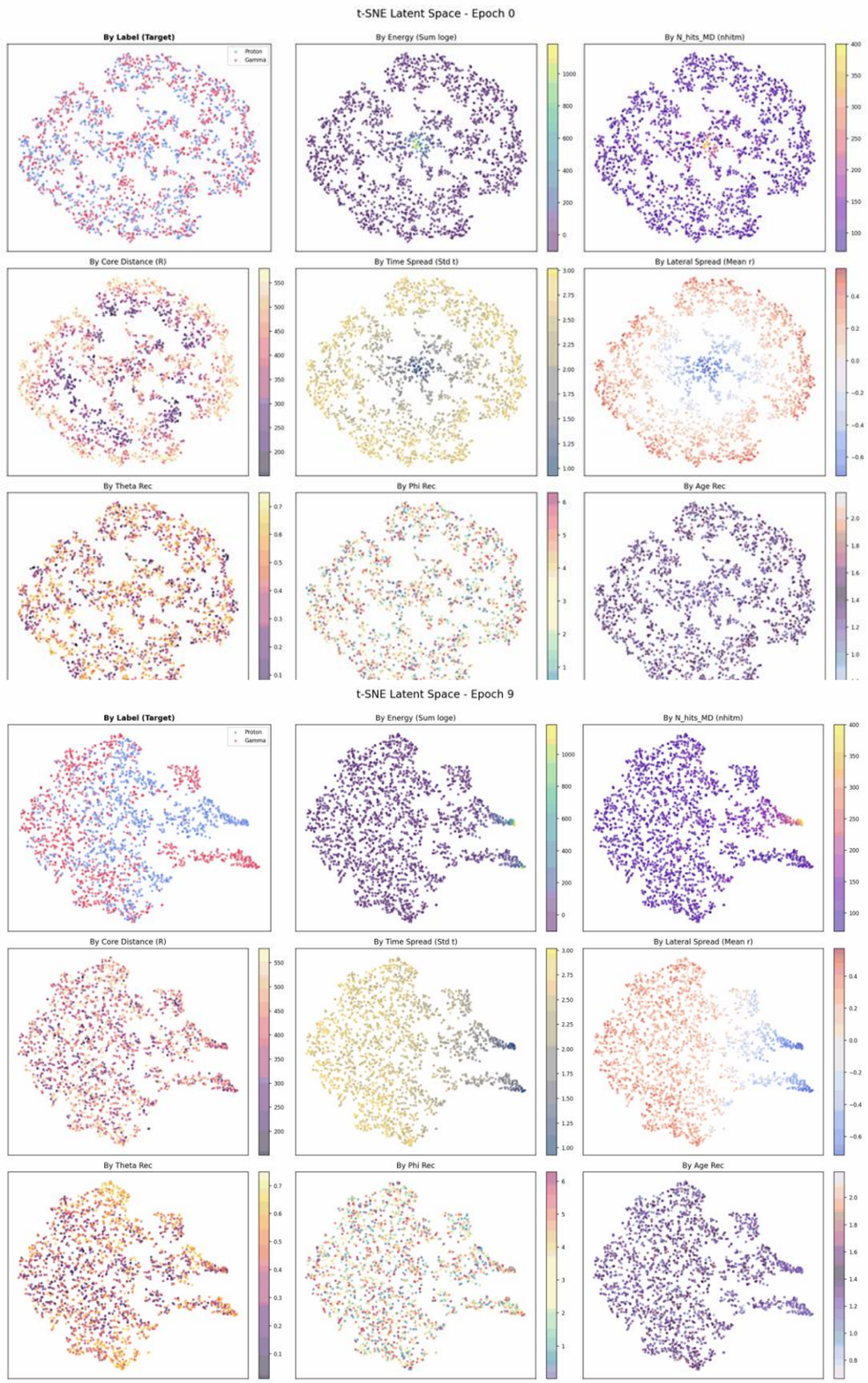
- 半监督模型在实验数据上的性能超越了同等参数量与训练数据量的有监督模型。

γ/p鉴别--半监督学习



用特征上色

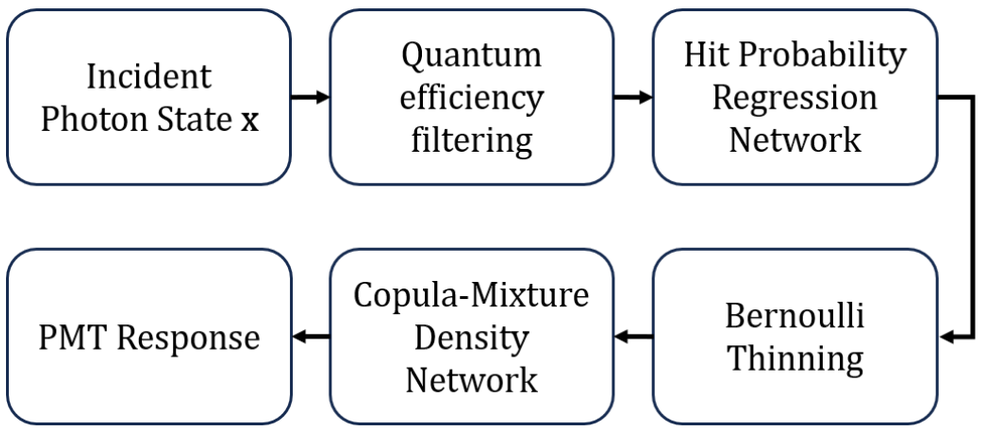


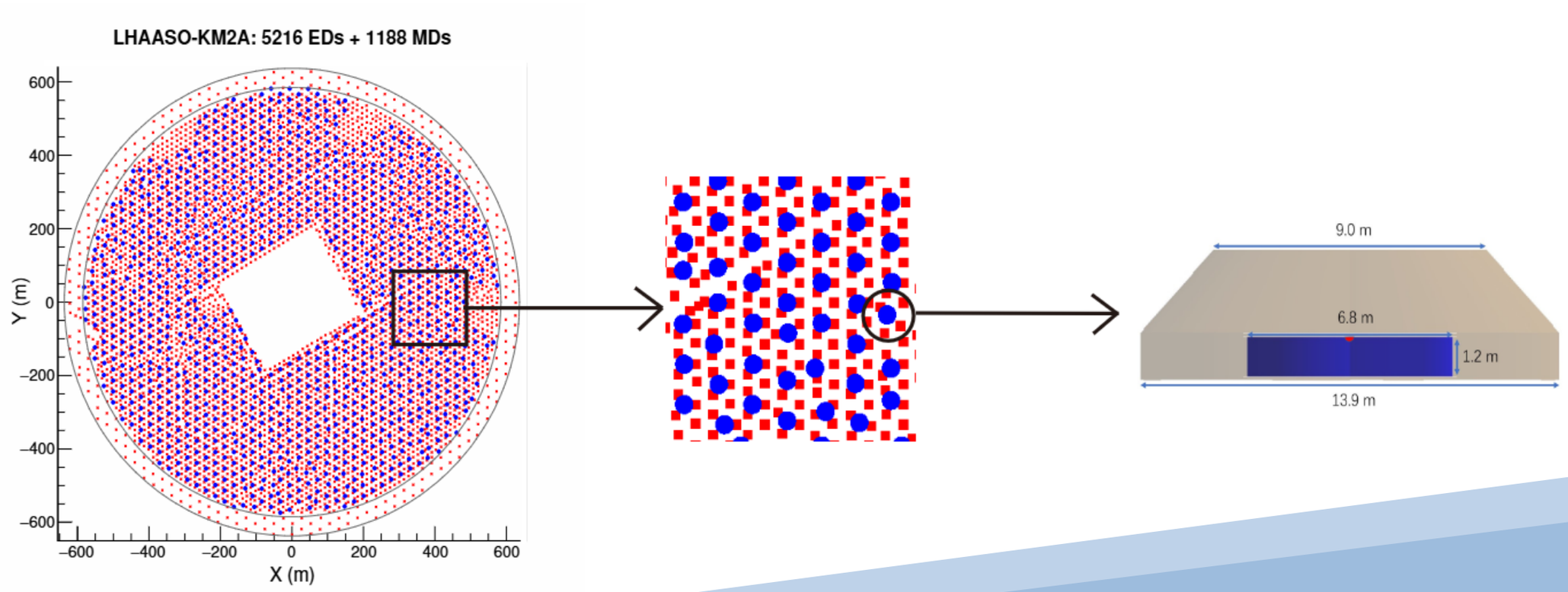


- 为了更好地理解模型预训练过程中是如何利用特征信息的，我们可以考虑对编码器输出的高维语义向量先进行PCA降维，再用不同特征进行上色

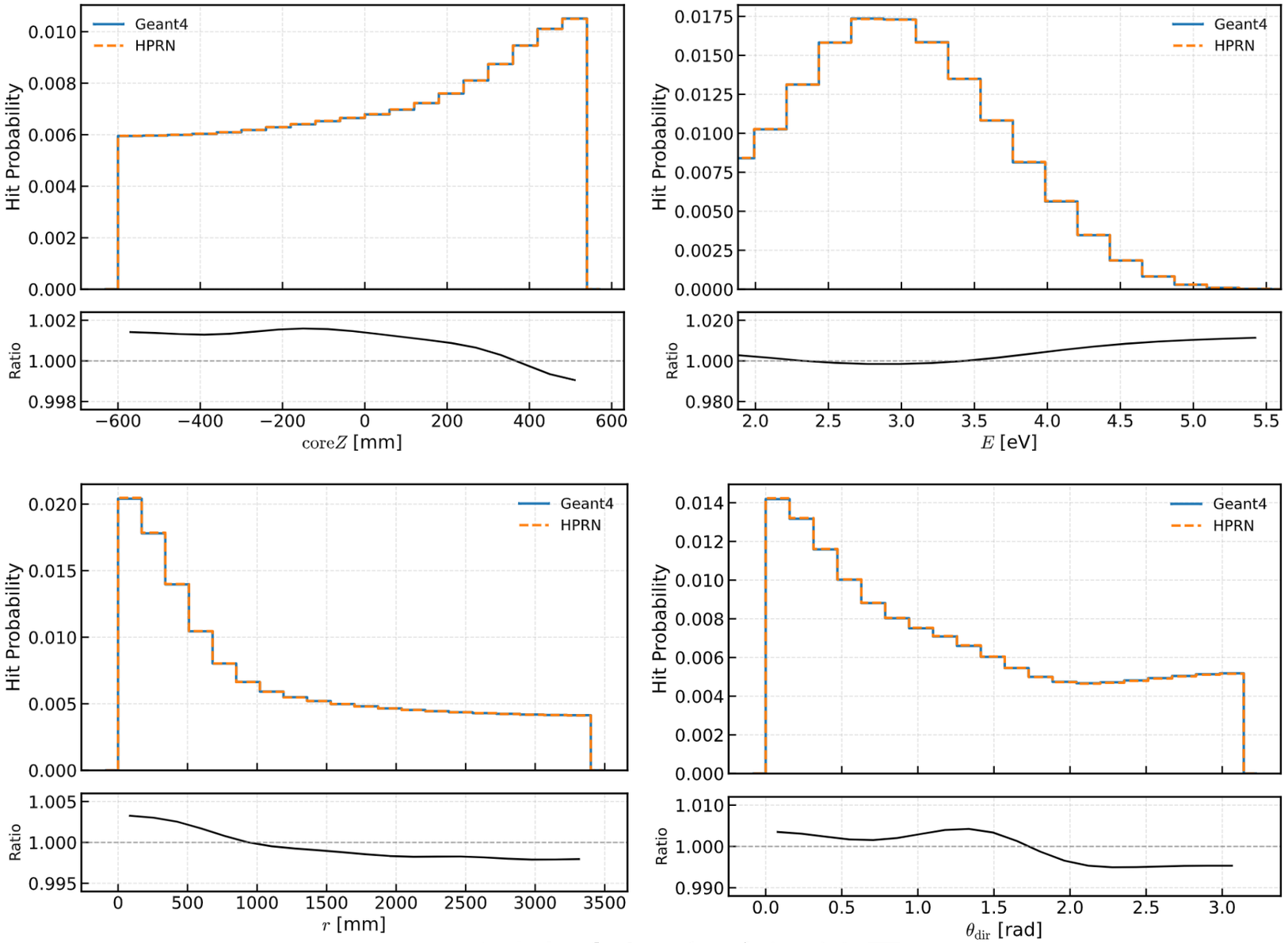
光子追迹模拟

- 在MD中，单个高能次级带电粒子会产生 ~10^4 个光学光子，Geant4对大量光子进行追迹模拟耗费大量时间和计算资源。
- 完全利用神经网络来学习输入光子状态到击中概率与相应分布的映射，**将模拟过程转化为推理计算。**
- 通过将光子击中概率预测与PMT响应预测解耦，实现高精度的两阶段快速模拟。

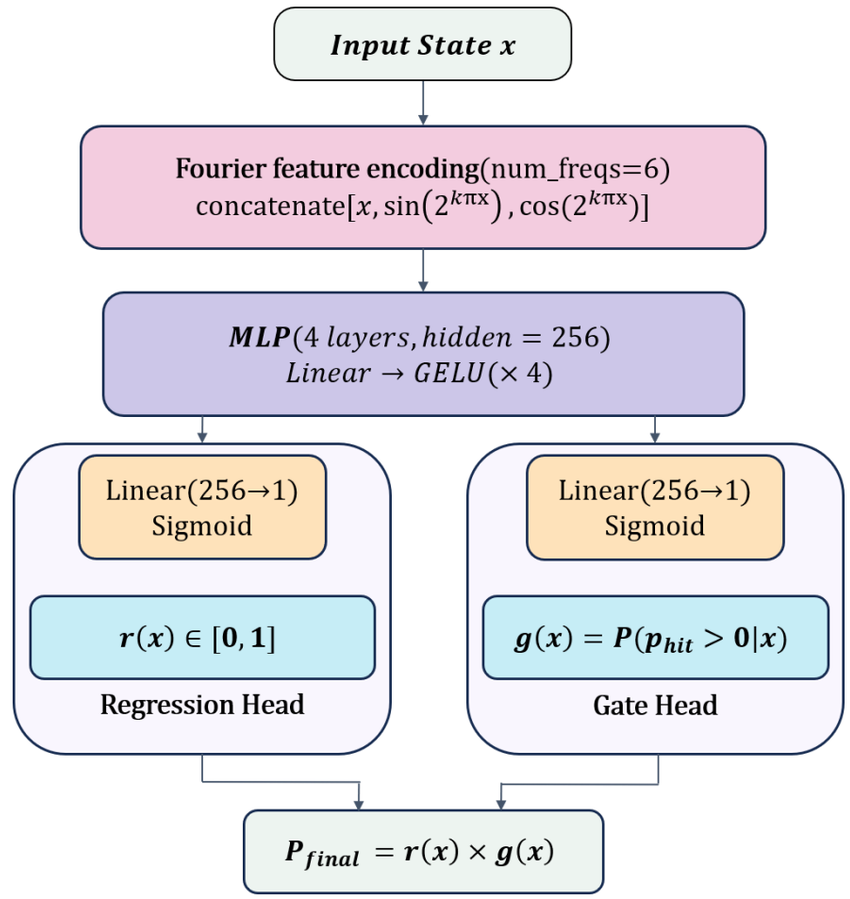




光子追迹模拟



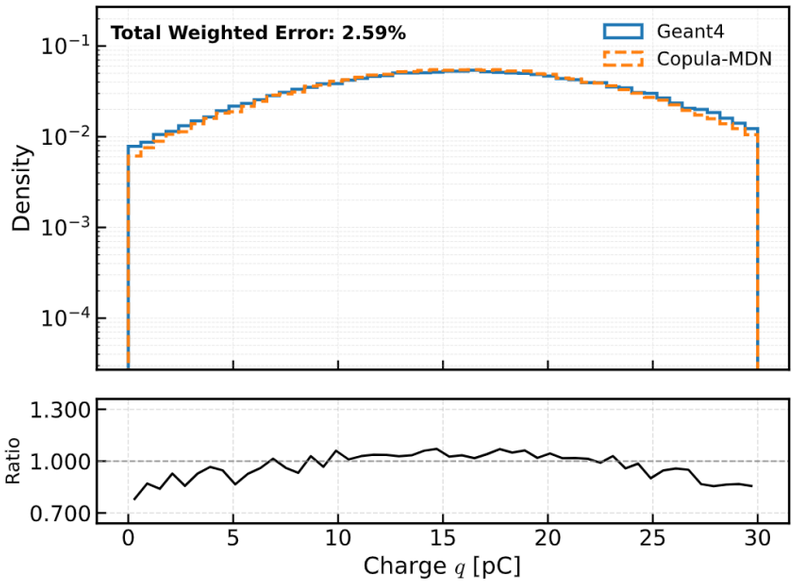
预测击中概率验证结果

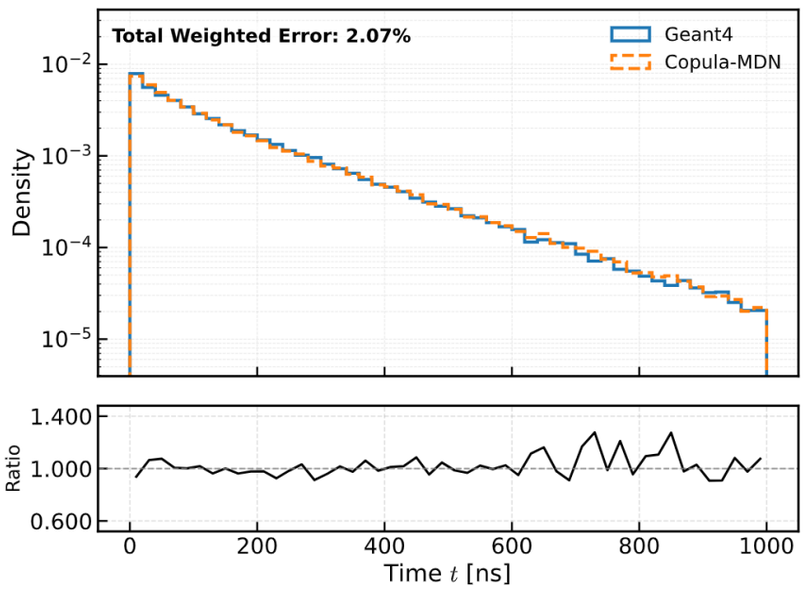


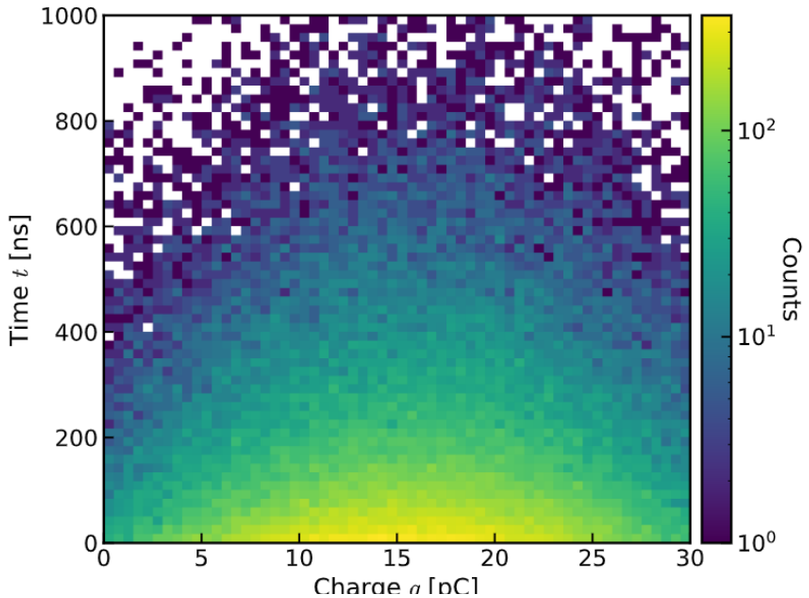
光子击中概率回归稀疏网络

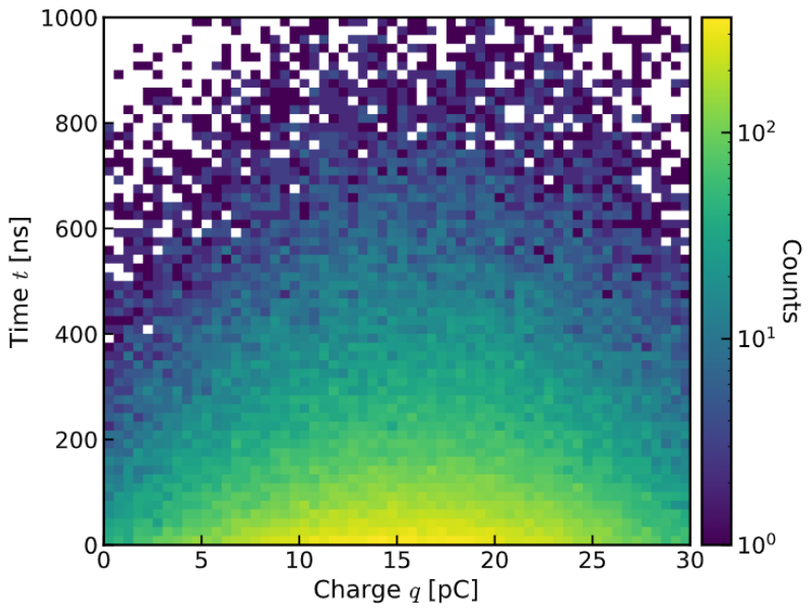
- 为了应对大量光子无法击中PMT的稀疏预测任务，我们设计了0概率门控
- 击中概率回归的平均误差控制在1%以内

光子追迹模拟

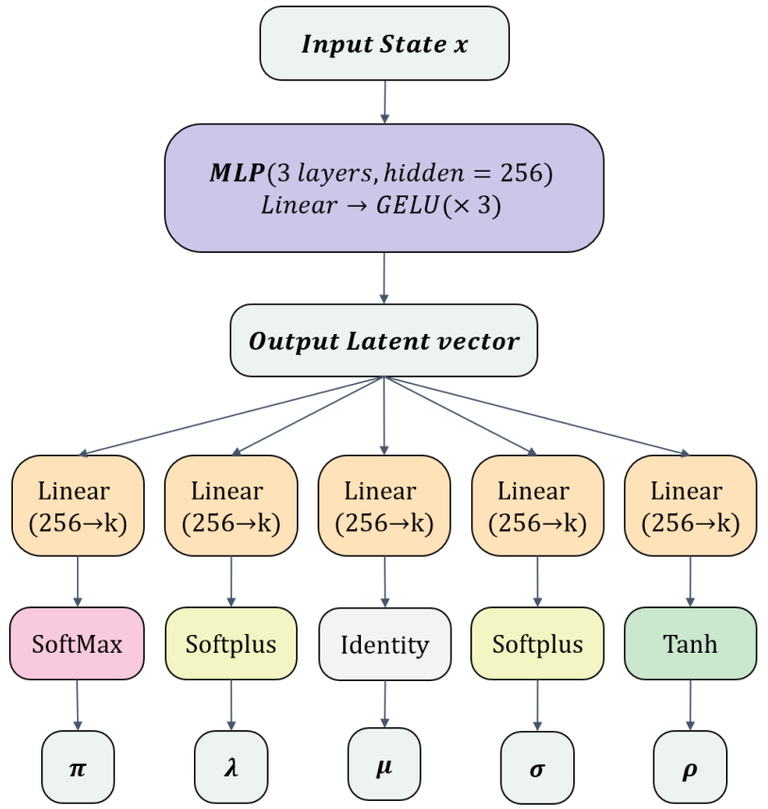








预测PMT响应验证结果



PMT响应分布建模网络（MDN）

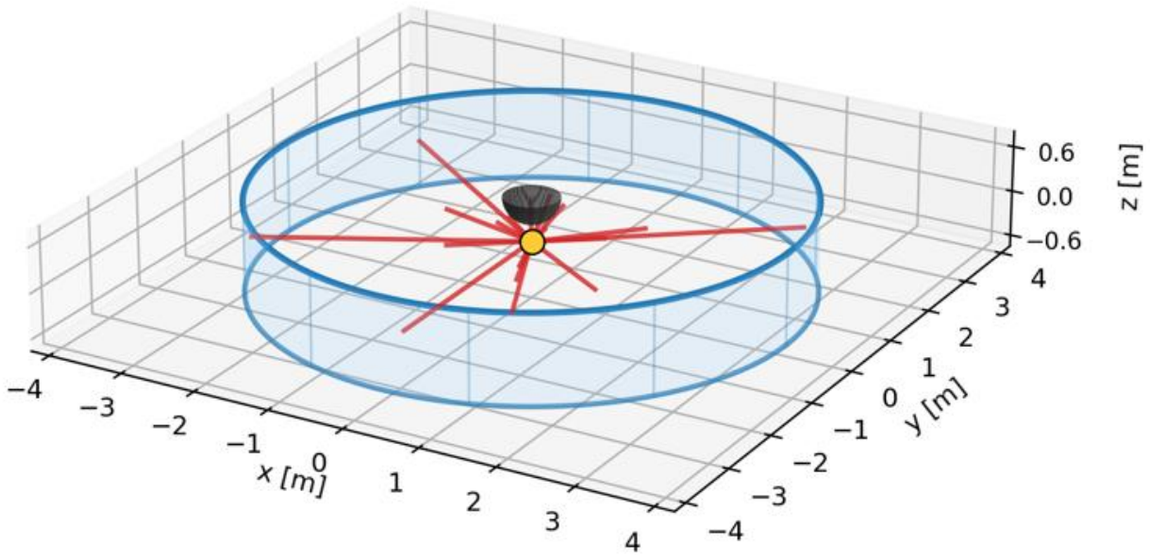
Table 1: Runtime comparison for processing 10^9 optical photons, excluding common overheads (photon-state generation and QE filtering).

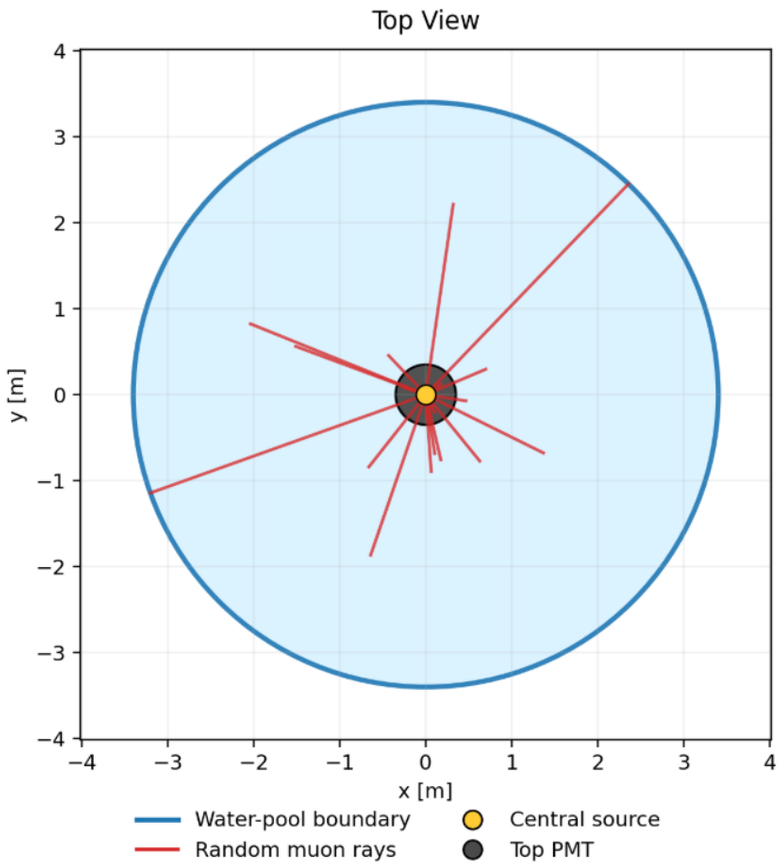
Hardware	Geant4 [s]	Two-stage model [s]
CPU	87599	729.55
GPU	—	12.87

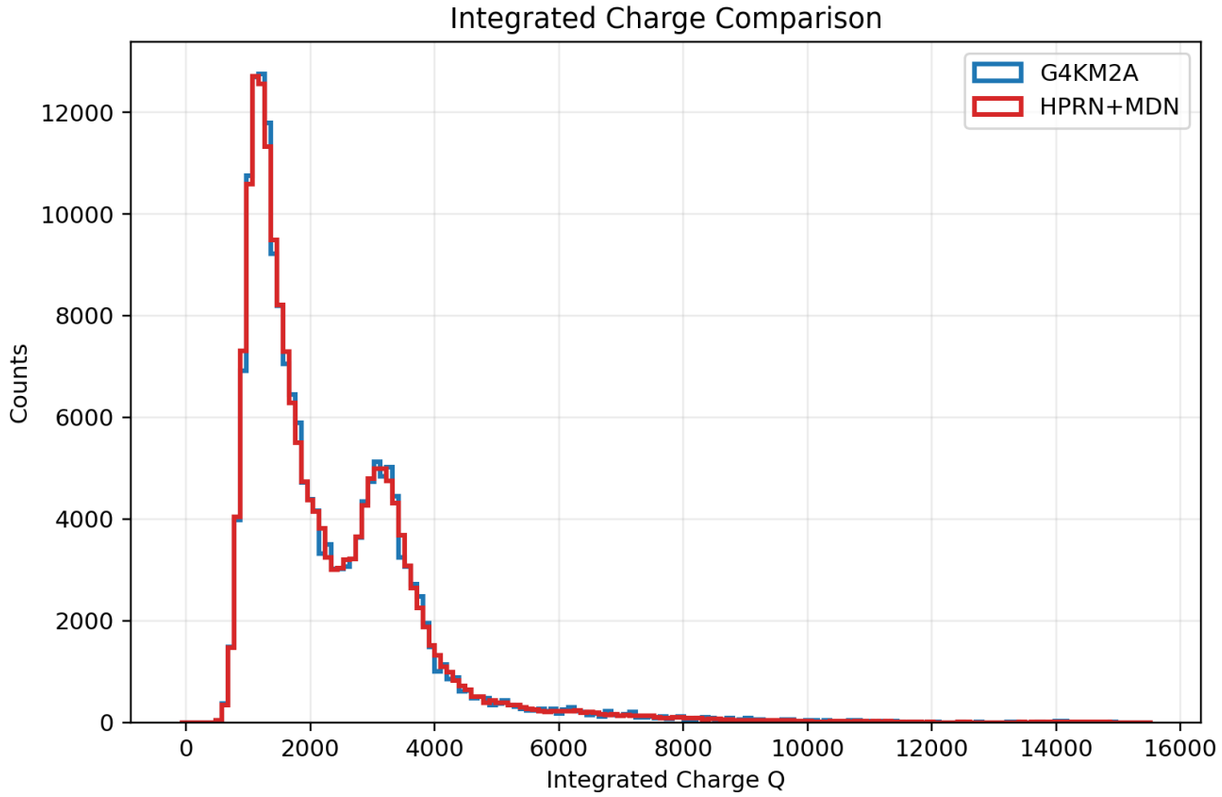
Hardware details: CPU: Intel(R) Xeon(R) CPU E5-2660 v4 @ 2.00GHz; GPU: NVIDIA Tesla V100-SXM2-32GB.

- 在总体误差在2~3%的误差下，深度学习相比Geant4在单CPU上实现了~120倍的提速，在GPU上推理模型则能实现~6800倍提速。

光子追迹模拟

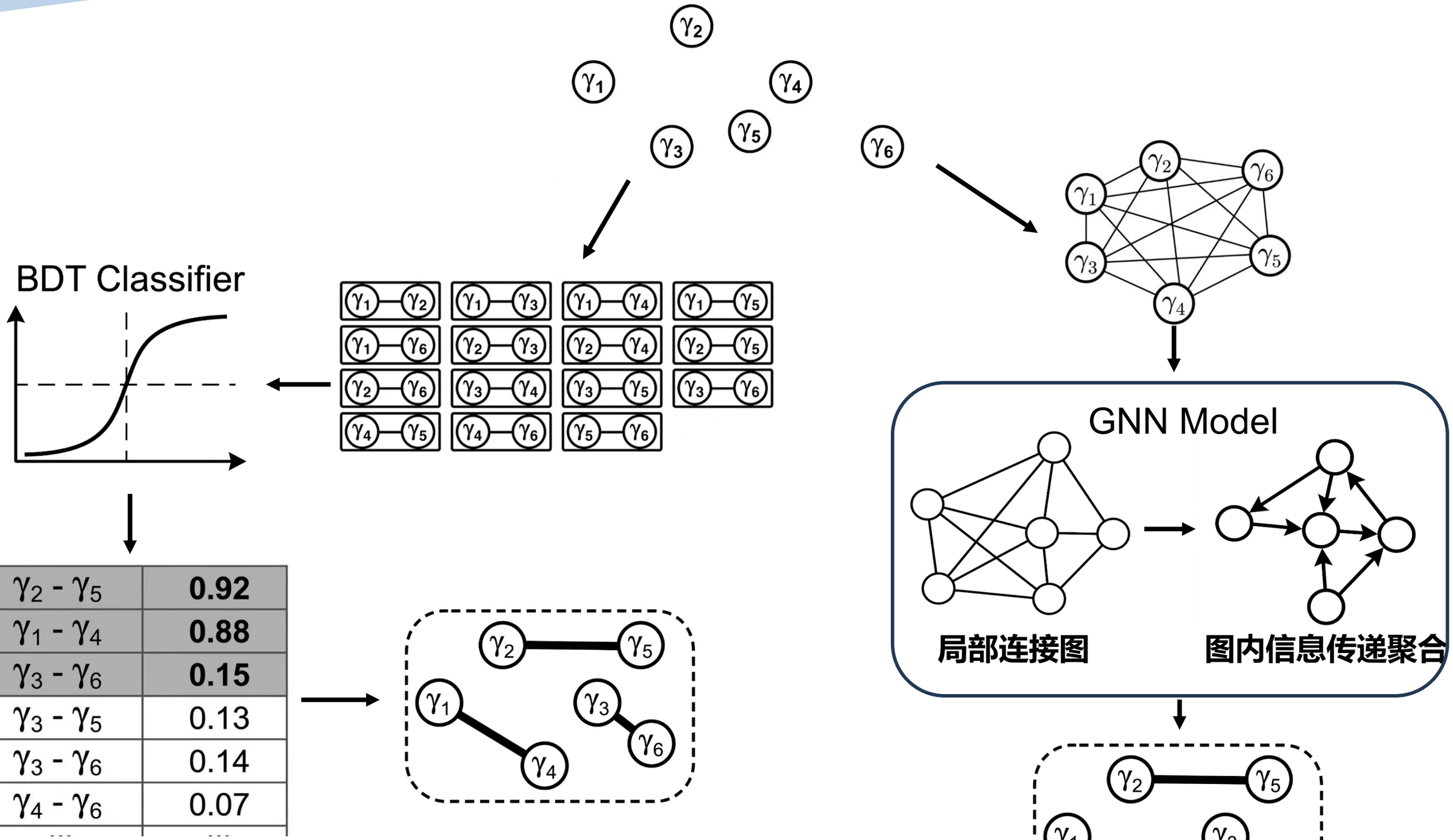






- 为验证深度学习替代光子追迹在上级模拟的可行性，我们以MD中心为起点随机发射20w个高能缪子，统计其PMT上积分电荷的分布
- 两阶段深度学习模型得到的积分电荷分布与G4KM2A真值符合得比较好

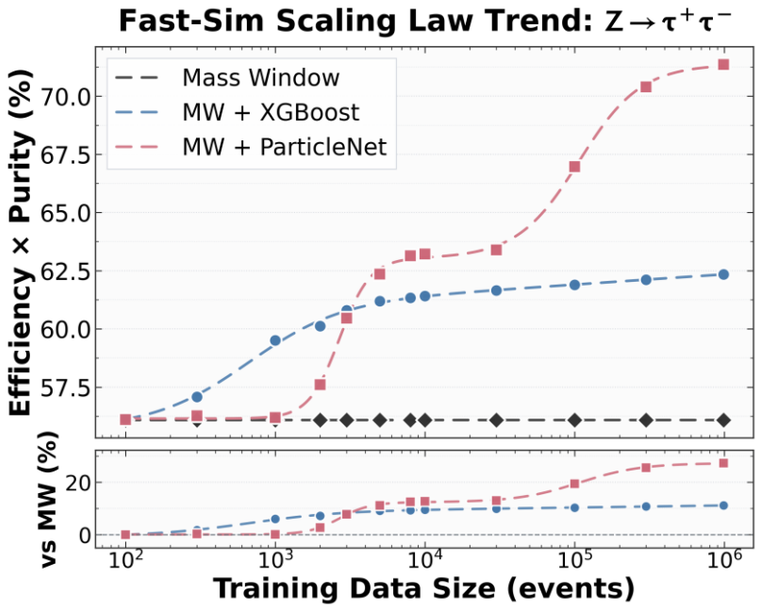
From Leaves to the Tree--衰变末态双光子重建

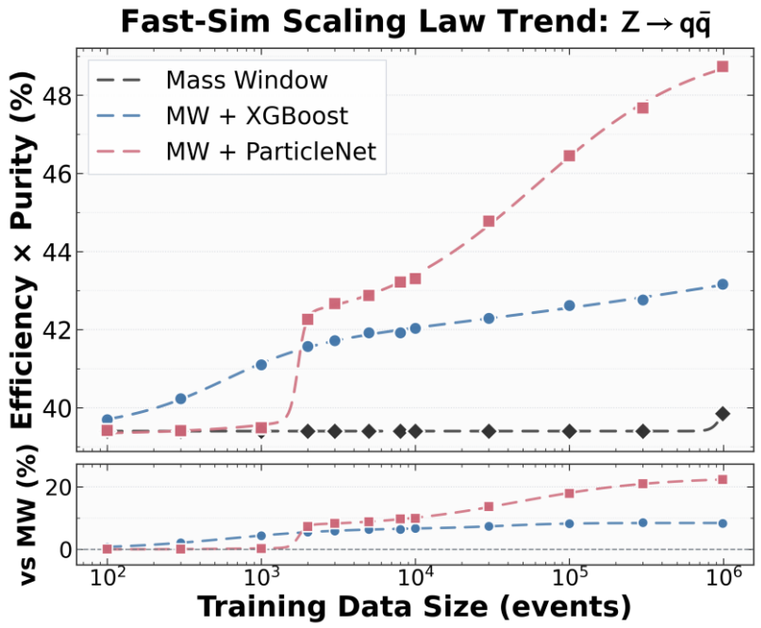


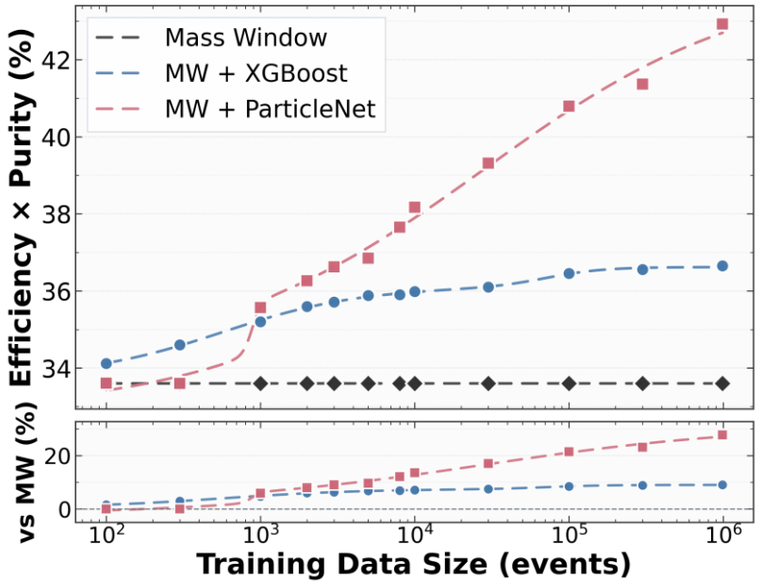
γ_2 - γ_5	**0.92**
γ_1 - γ_4	**0.88**
γ_3 - γ_6	**0.15**
γ_3 - γ_5	0.13
γ_3 - γ_6	0.14
γ_4 - γ_6	0.07

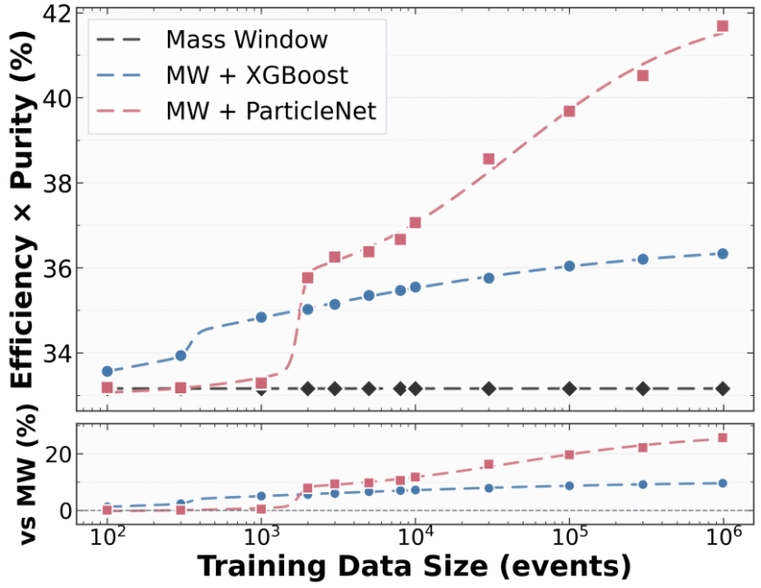
- 传统的BDT等机器学习方法只能依据两光子的局部联合特征进行分类，但深度学习方法可以让模型学习每个光子的领域信息，在整个事例末态的全局特征指导下进行分类。

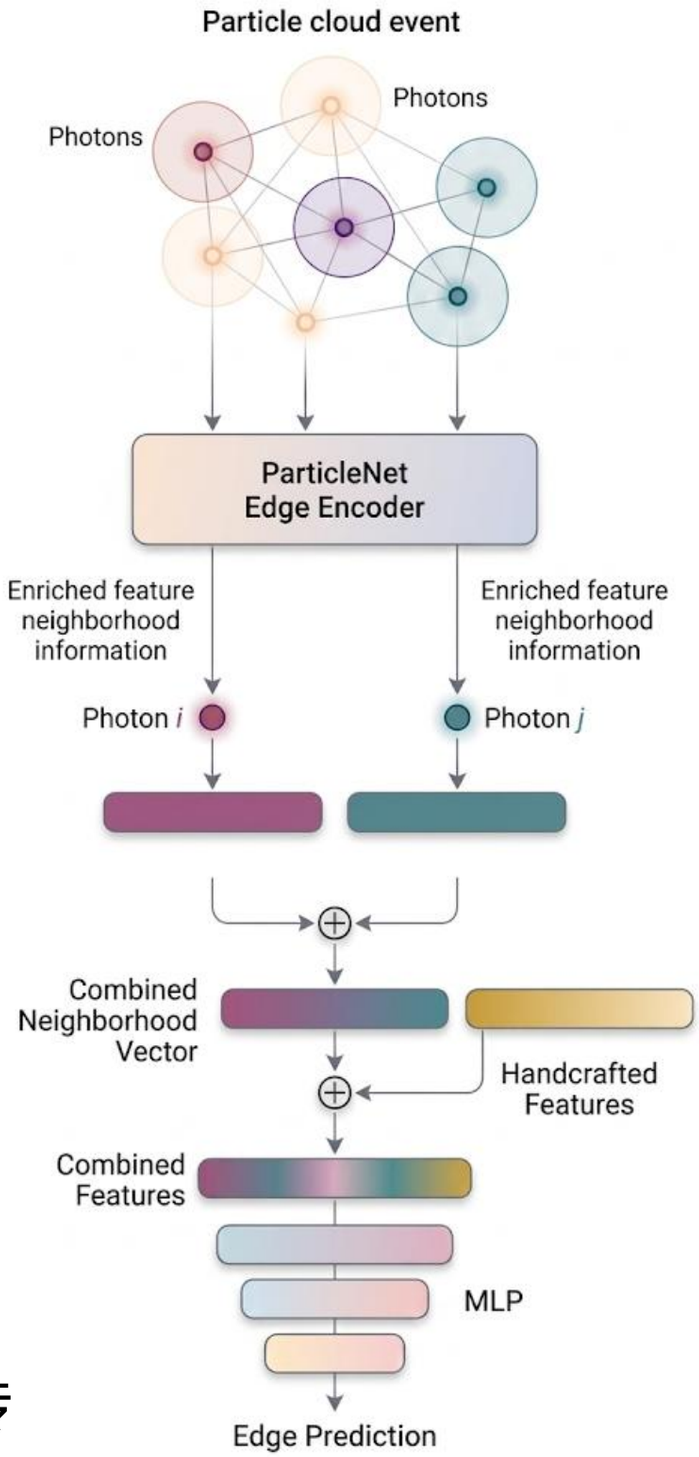
From Leaves to the Tree--衰变末态双光子重建









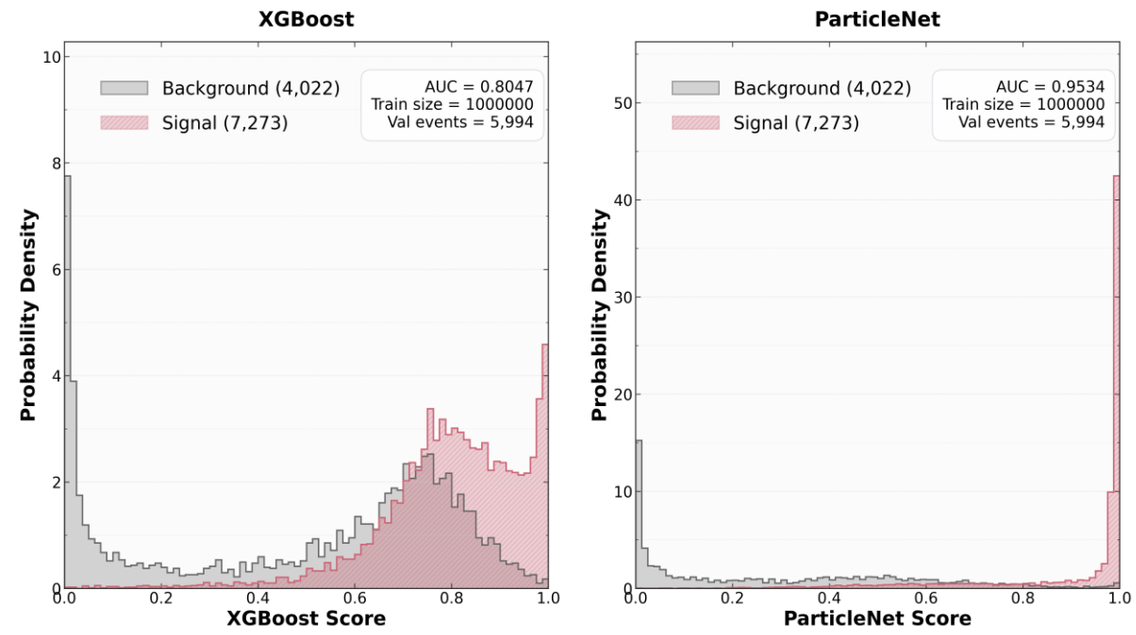


Fast Simulation的结果

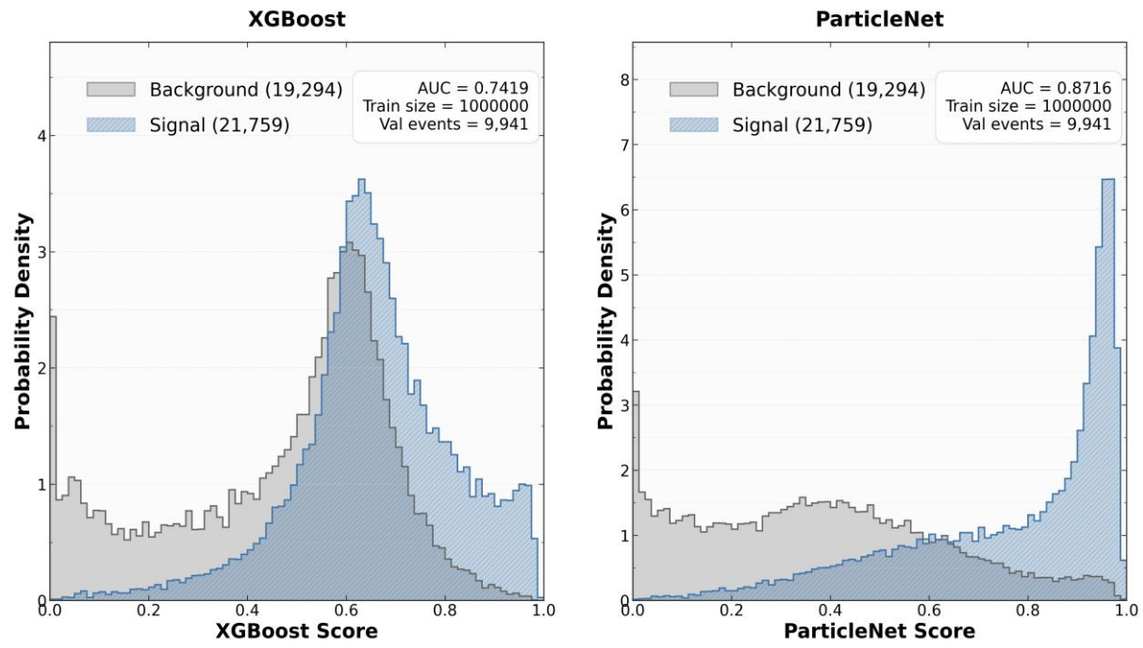
- 通过学习全图的光子邻域信息再做分类，深度学习模型比传统的机器学习模型能够更好地理解事例末态粒子之间的关系，重建性能得到大幅提升。

From Leaves to the Tree--衰变末态双光子重建

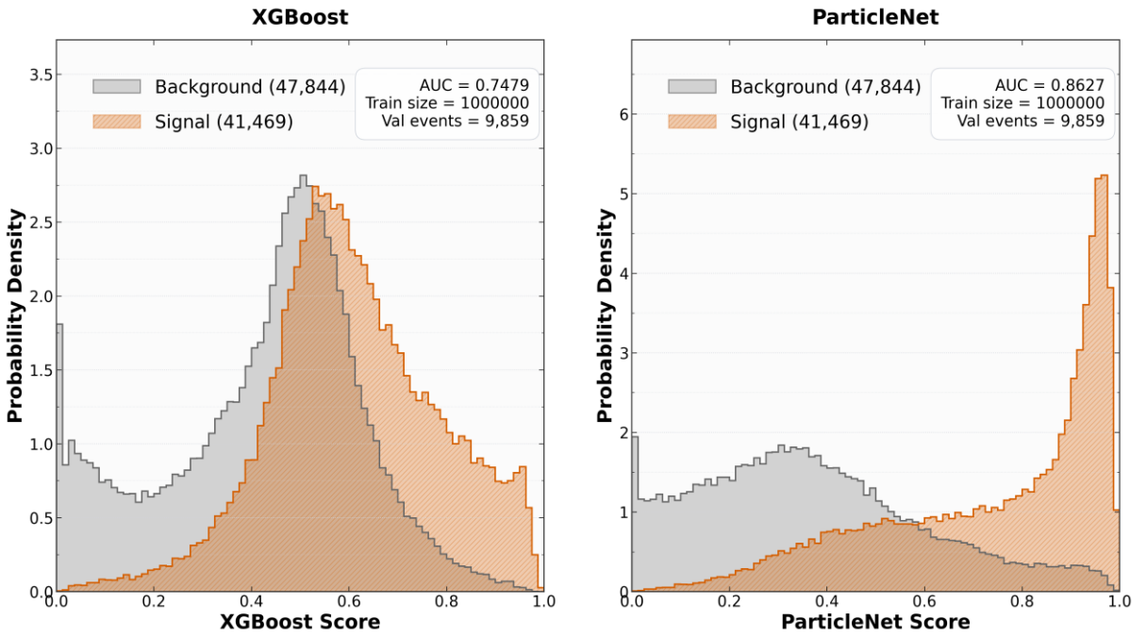
$Z\to\tau^+\tau^-$ Score Distributions



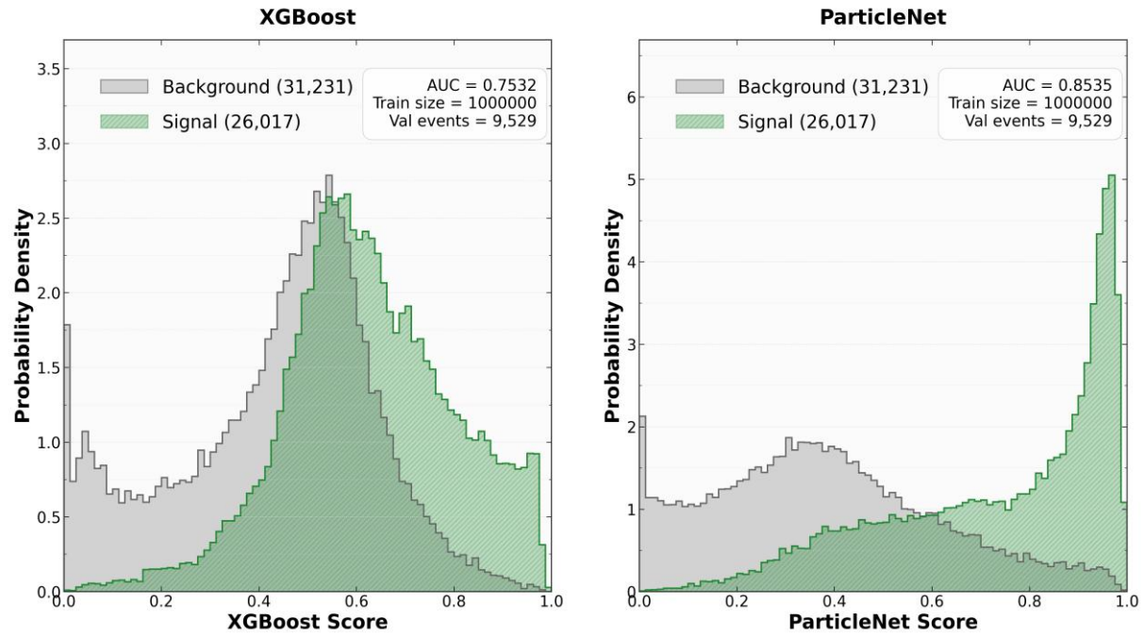
$Z\to q\bar{q}$ Score Distributions



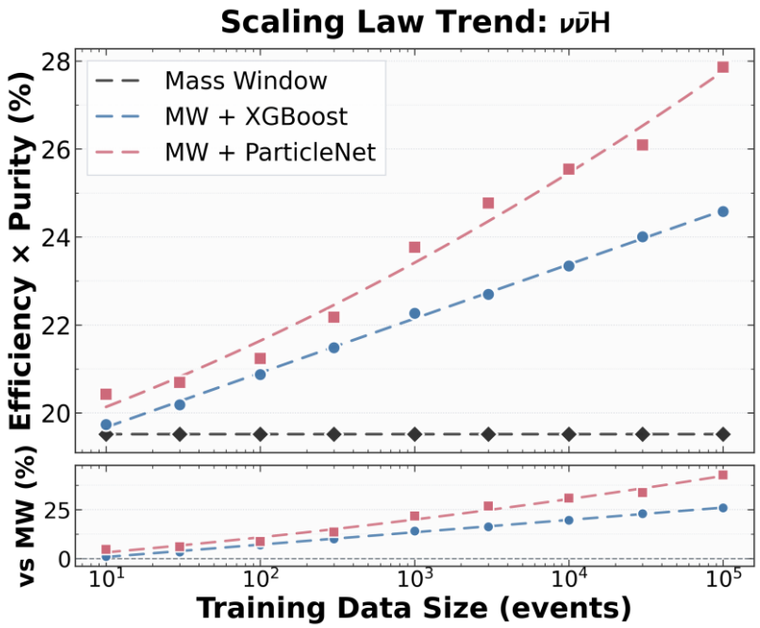
ZH Score Distributions

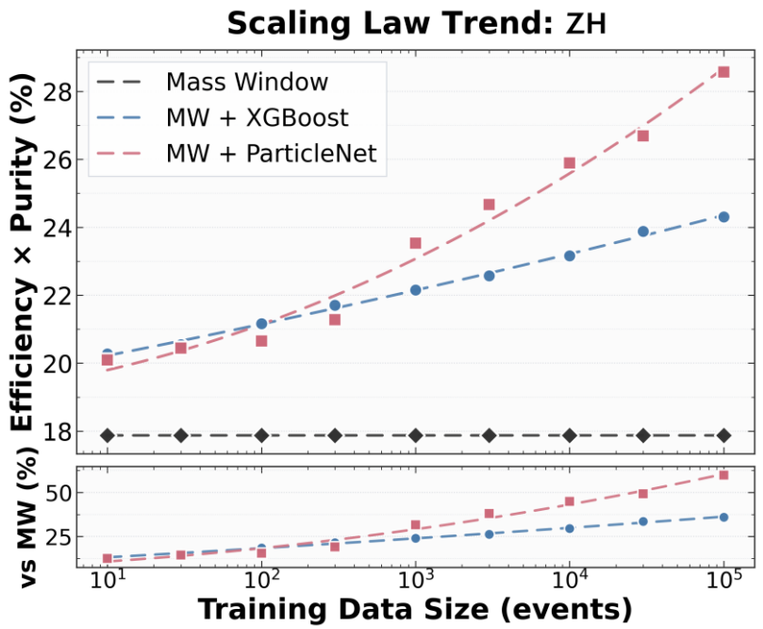


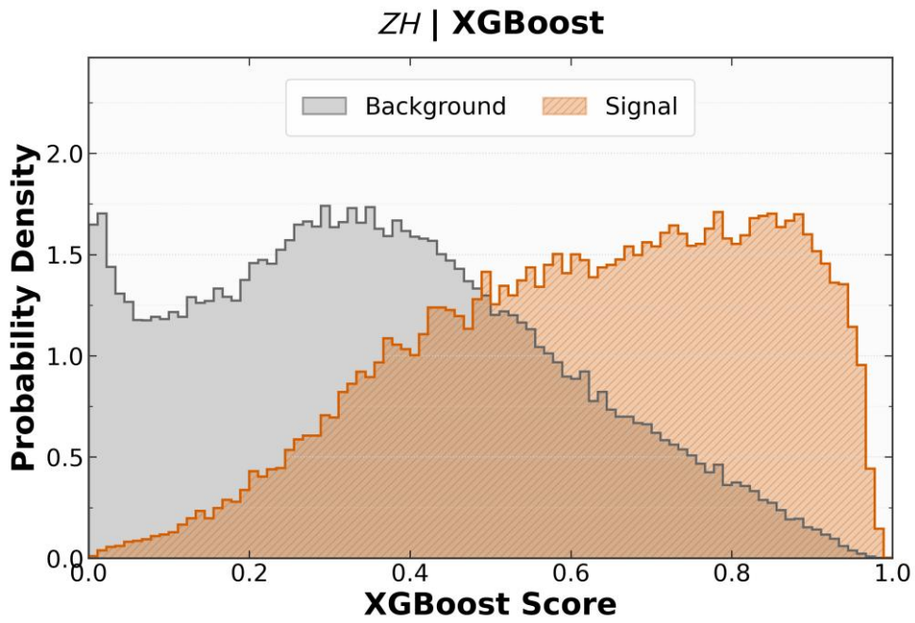
$\nu\bar{\nu}H$ Score Distributions

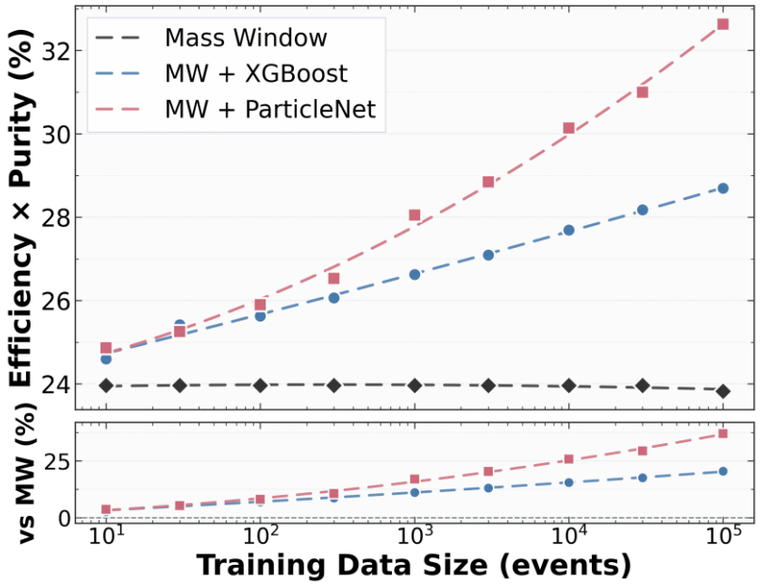


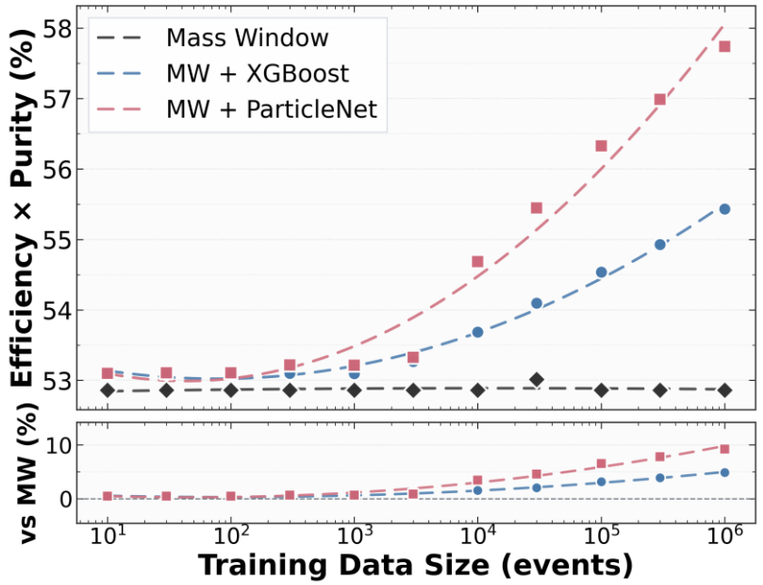
From Leaves to the Tree--衰变末态双光子重建

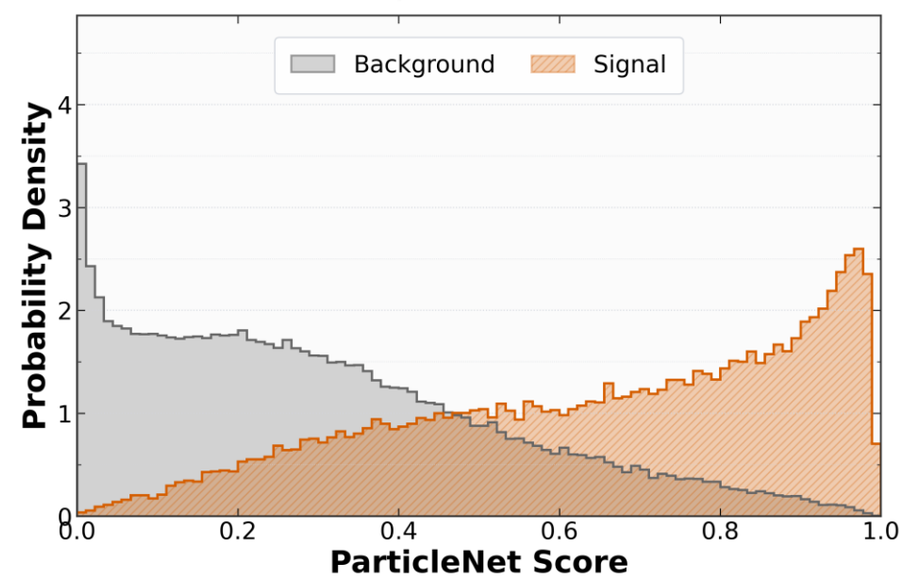










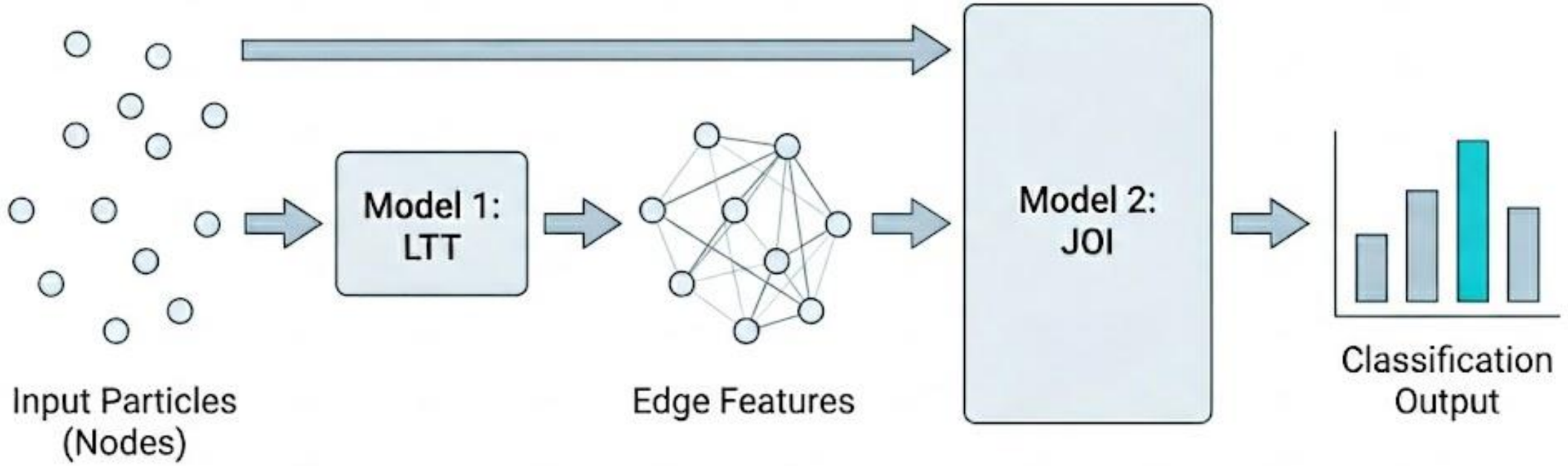


Full Simulation的结果

- 对Full Simulation, 重建难度更大，同等数据量下模型性能相比于Fast Simulation下降，模型收敛速度下降。

未来工作

- 研究双光子重建scaling law（Fast vs Full）
- 深度融合GNN与Transformer架构，探究局部与全局注意力在重建问题的影响
- 尝试对带电末态粒子进行重建
- 验证From Leaves to the Tree方法对其他机器学习工作的性能提升（如JOI）



Thanks for your listening!