

机器学习在江门中微子实验事例重建中的应用

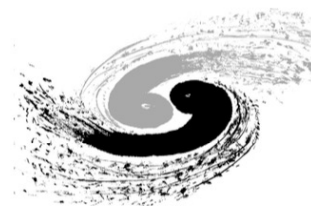
李紫源 罗武鸣 钱圳 尤郑昀

中山大学

中科院高能物理研究所

**on behalf of the JUNO collaboration
& Machine Learning group**

2021 年 8 月 18 日



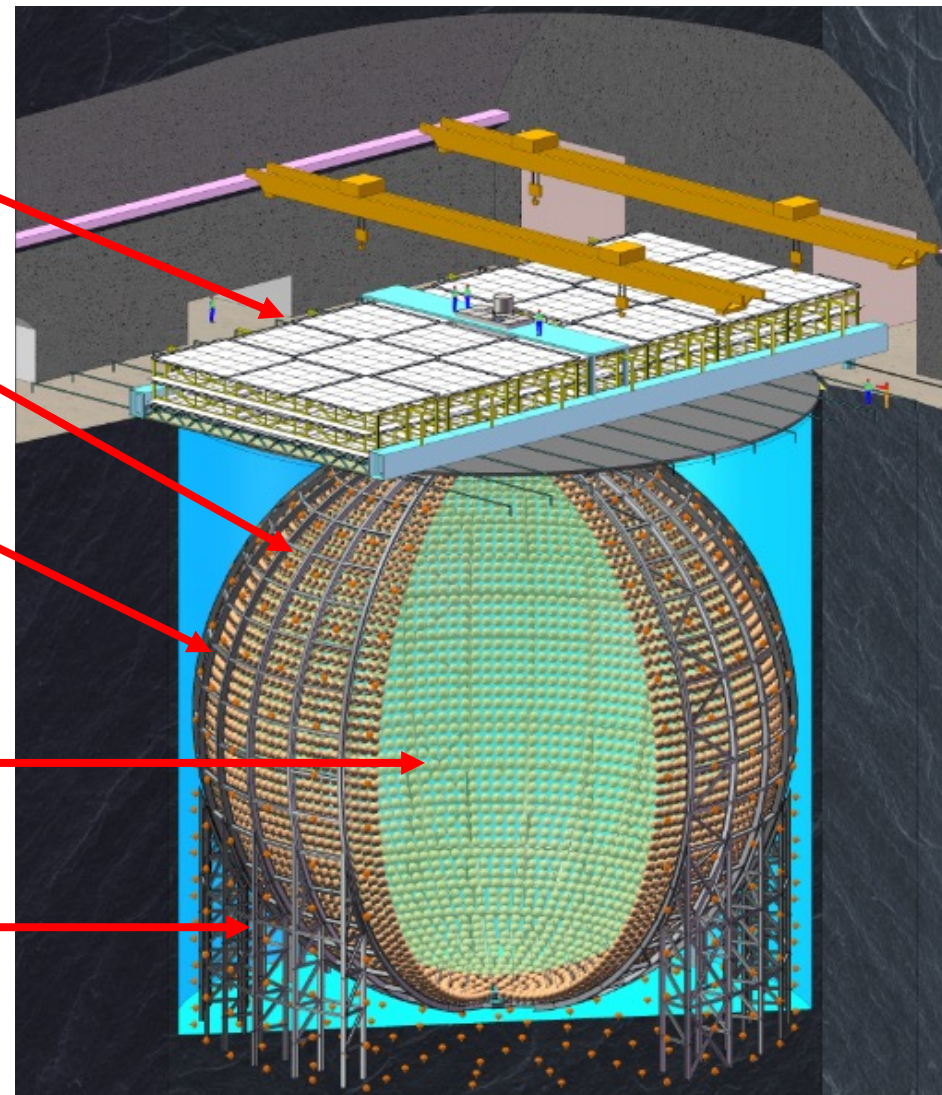
- 江门中微子实验
- JUNO中的几种机器学习重建算法
 - 提升决策树 (BDT)
 - 全连接神经网络 (FC-DNN)
 - 卷积神经网络 (CNN)
 - 图神经网络 (GNN)
- 能量与顶点重建性能
- 总结与展望

• 江门中微子实验 (JUNO)

Jiangmen **U**nderground **N**eutrino **O**bservatory

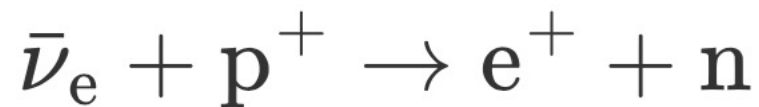
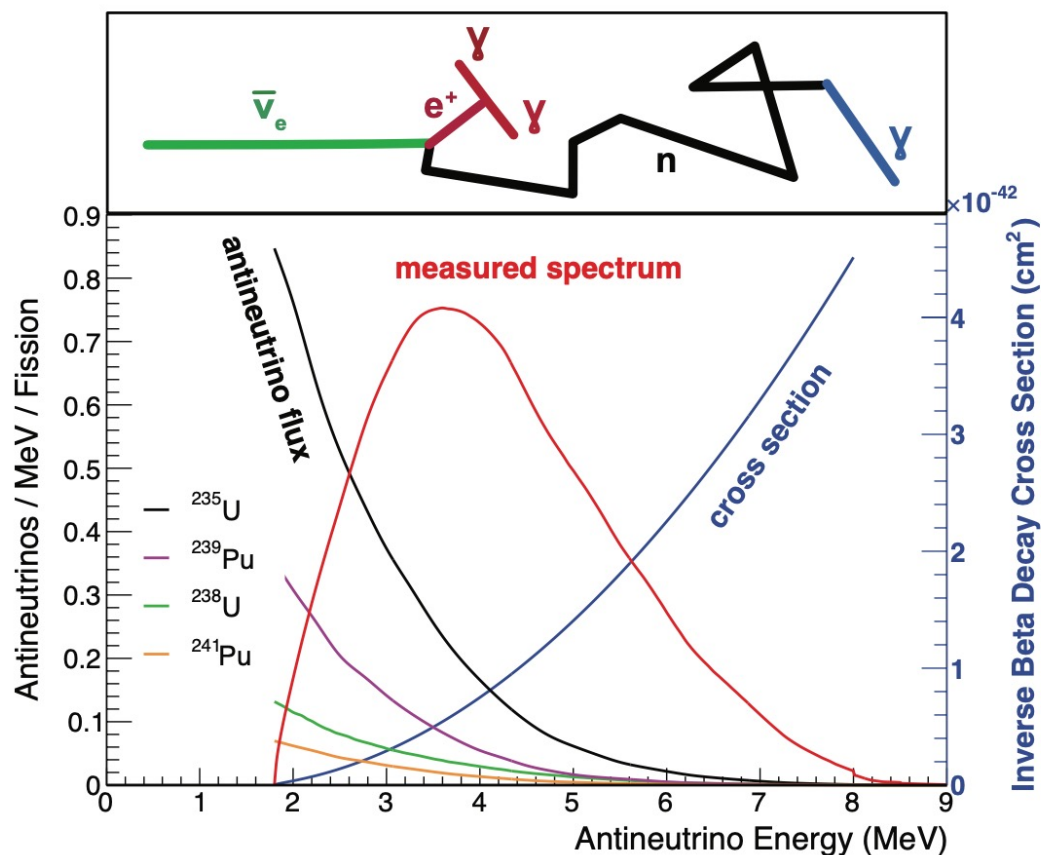
- 广东省江门市
- 700 米岩层覆盖
- 接受来自阳江核电站、台山核电站的中微子源
- 物理意义：
 - 测定三代中微子质量顺序
 - 精确测量振荡参数

- 顶部径迹探测器
- 中心探测器
 - 有机玻璃球壳
 - 不锈钢支架
- 光电倍增管 (PMT)
 - 17600 支 20 英寸 PMT
 - 25600 支 3 英寸 PMT
 - 覆盖率78 %
- 液体闪烁体
 - 20 千吨
- 水切伦科夫探测器
 - 35 千吨
 - 2400 支 20 英寸 PMT



• JUNO中微子探测方法

- 使用反 β 衰变 (IBD) 探测单个中微子能量
- 进而测量中微子能谱, 进而去推测中微子质量排序
- 对能量分辨率要求极高



正电子会迅速湮灭在液闪中, 随后能量沉积出大量光子
光子将被探测器外侧球面上的 PMT 探测

*Neutrino Physics with JUNO,
J. Phys. G 43 (2016) 030401*

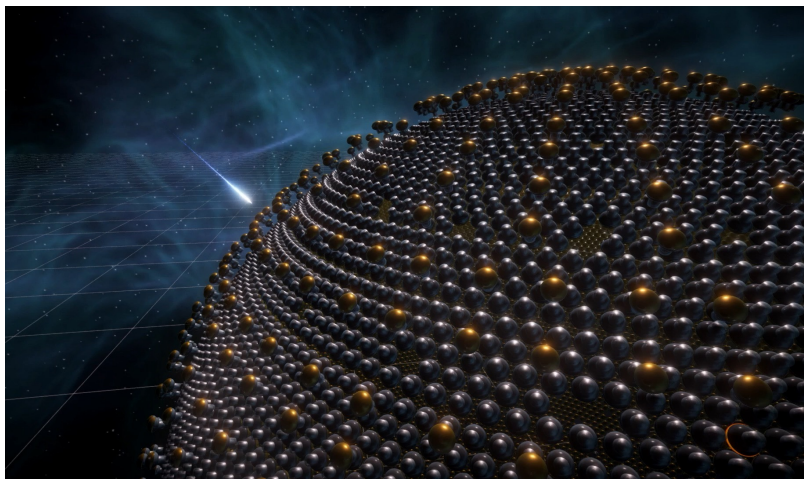
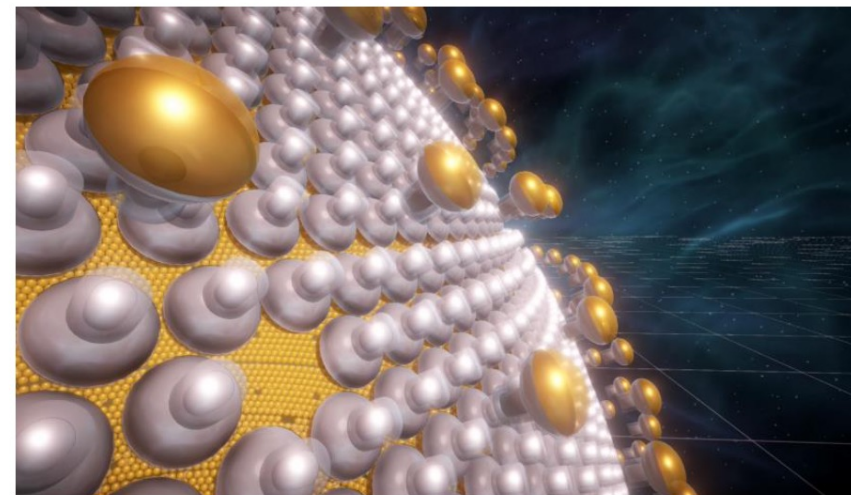
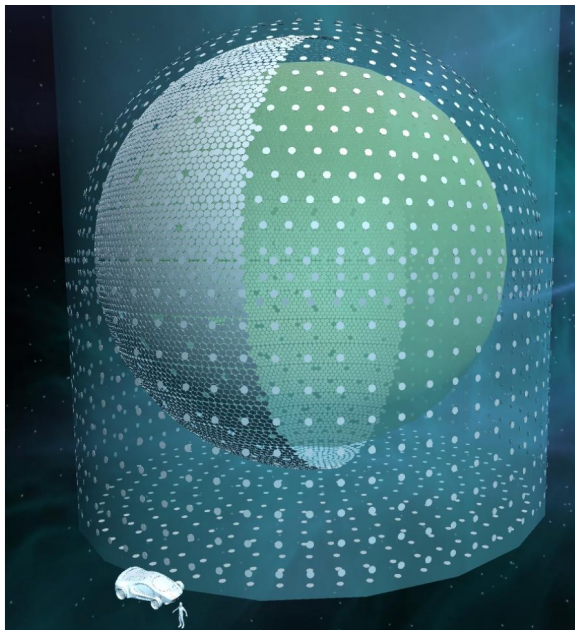
• JUNO 探测器建模与事例显示

探测器建模

JUNO 探测器（左），PMT 排布（右）

建模与显示可以观察
光子在液闪中的传播

K. Li, Z. You, Y. Zhang, et al., NIMA 908 (2018) 43-48
S. Zhang, J.S. Li, Y.J. Su, et al., NST 32, 21 (2021)



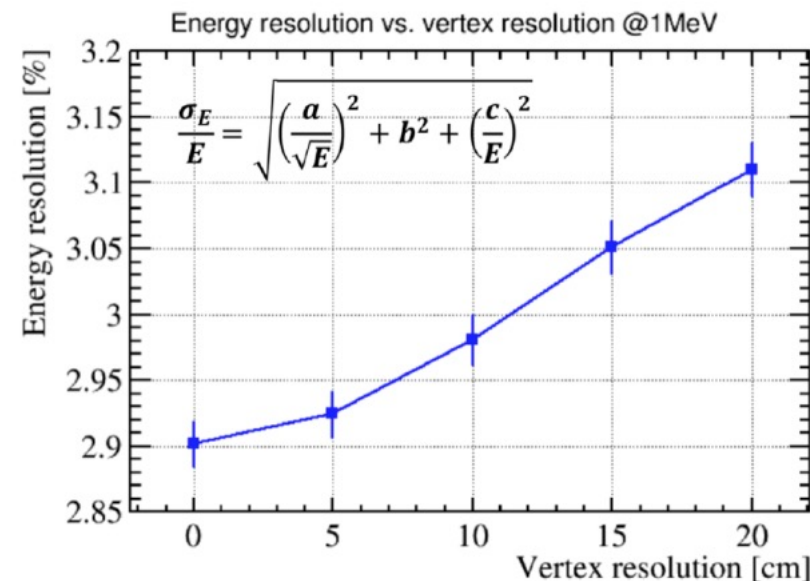
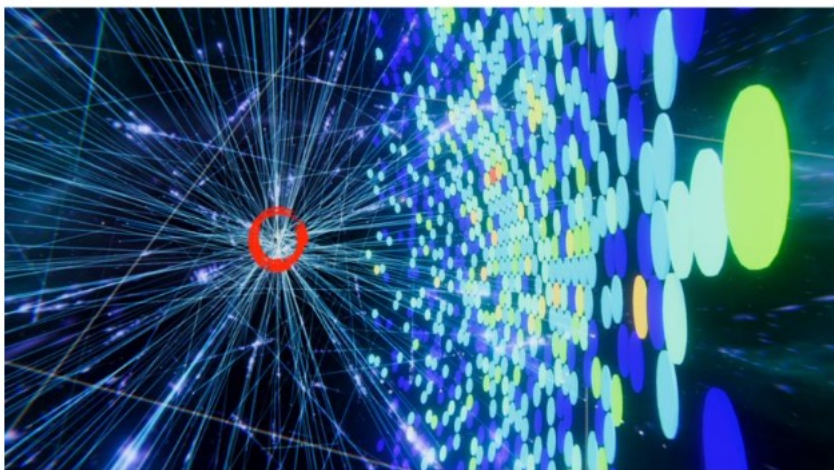
事例显示

- 软件：
 - Unity 3D
 - ROOT Display

Z. You, K. Li, Y. Zhang, et al., JINST 13 (2018) T02002
J. Zhu, Z. You, Y. Zhang, et al., JINST 14 (2019) T01007

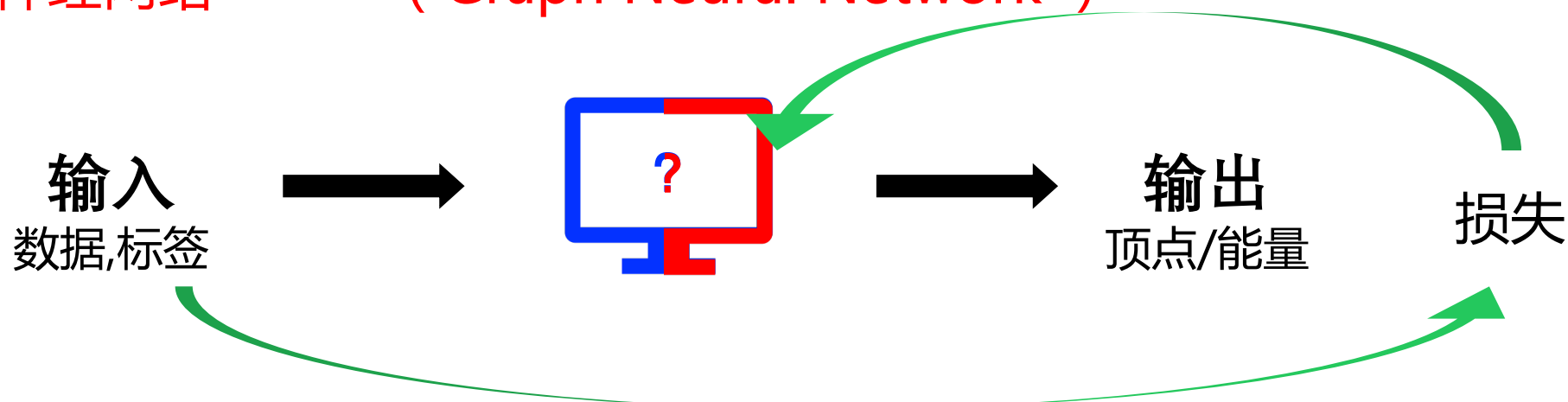
• JUNO中的事例重建

- 输入信息
 - 某个时间段内, 每个 PMT 上接受到的电荷, 与光子的击中时间
- 输出信息
 - 能量沉积 E 大小 (要求能量分辨率 < 3%)
 - 能量沉积平均位置 (顶点) (X, Y, Z) (要求顶点分辨率 < 12 cm)
- 重建算法
 - 传统算法 (基于物理模型、飞行时间、极大似然法)
 - 机器学习 (用数据训练模型, 使其具备预测能力)



• 基于机器学习重建

- 输入：PMT 的电荷与击中时间、聚合特征
- 输出/标签：顶点位置 (x, y, z)、能量 E
- 使用机器学习中的回归功能
- 机器学习算法：
 - 提升决策树 (Boost Decision Tree)
 - 全连接神经网络 (Fully Connected Deep Neural Network)
 - 卷积神经网络 (Convolutional Neural Network)
 - 图神经网络 (Graph Neural Network)



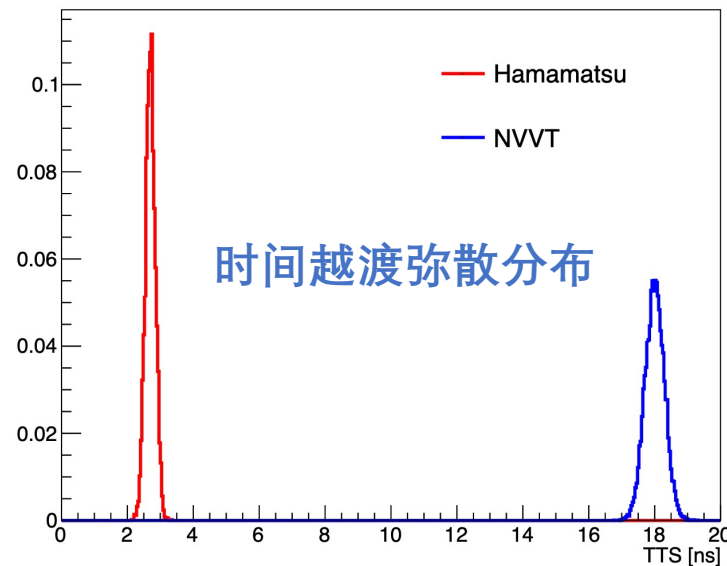
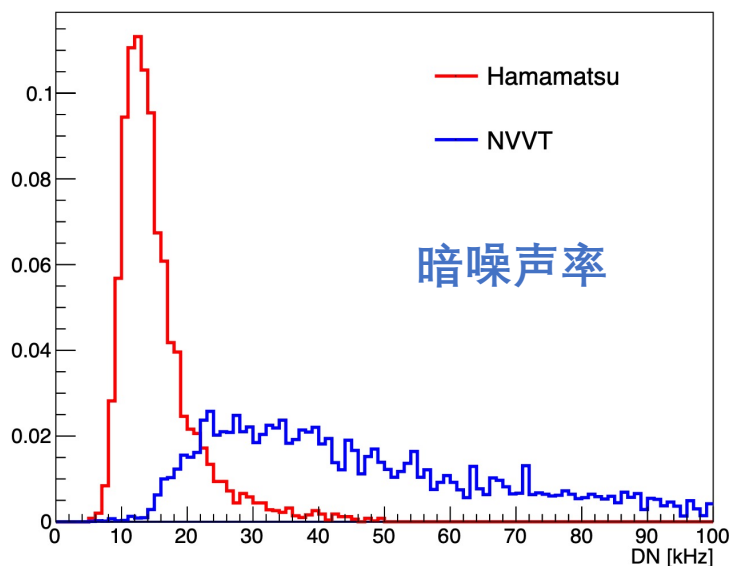
Z. Qian, V. Belavin, et al, N.I.M.A, Ser. 1010 (2021) 165527 (Machine Learning group in JUNO)

数据集

- 机器学习模型的训练和测试使用的数据来自 JUNO 离线软件 **SNiPER** 蒙卡产生

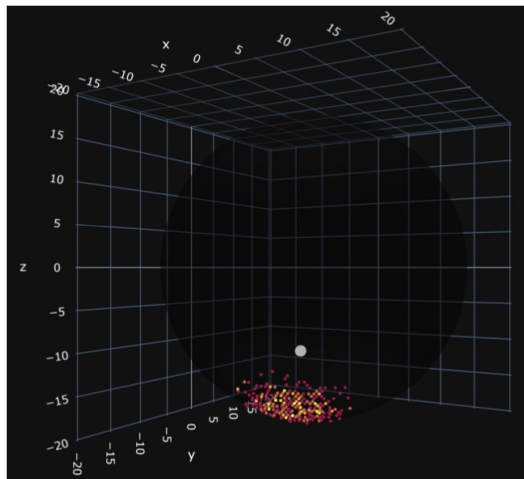
T. Lin, J. Zou, et al, J. Phys. Conf. Ser. 898 (2017) 4, 042029

- **训练数据集** : 5M 个正电子事例, 能量在 0-10 MeV 均匀分布
- **测试数据集** : 11 组 $\times$ 10k 正电子事例, 能量为 0, 1, $\dots$ 10 MeV
- **位置分布** : 均匀的在液体闪烁体中分布
- **PMT两种效应** : 暗噪声 (DN) 与时间越渡弥散 (TTS) 人工加入到数据中
- **四组数据** : **1.** w/ TTS&DN, **2.** w/ TTS, **3.** w/DN, **4.** w/o TTS&DN

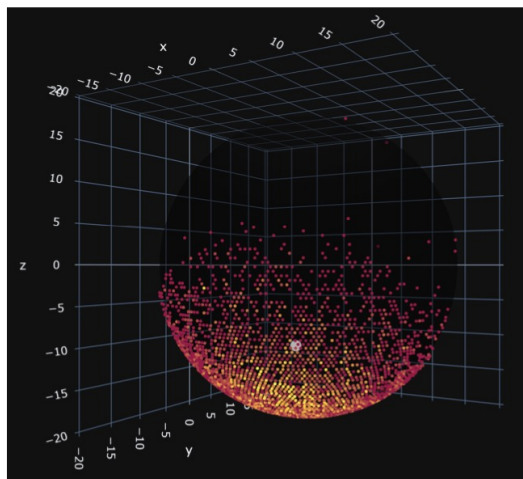


1. PMT 接受到的时间与电荷

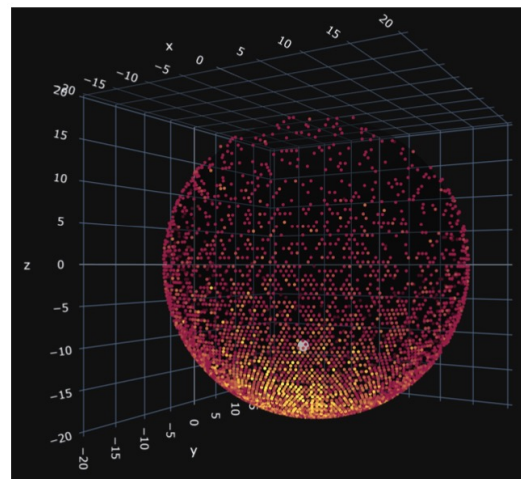
单事例电荷随时间演化



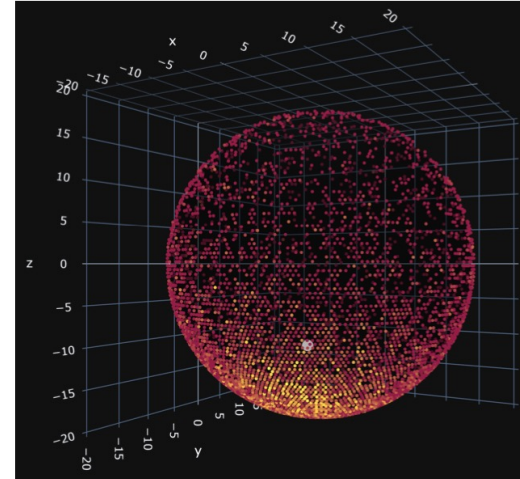
5th ns



55th ns



105th ns



155th ns

2. 聚合特征

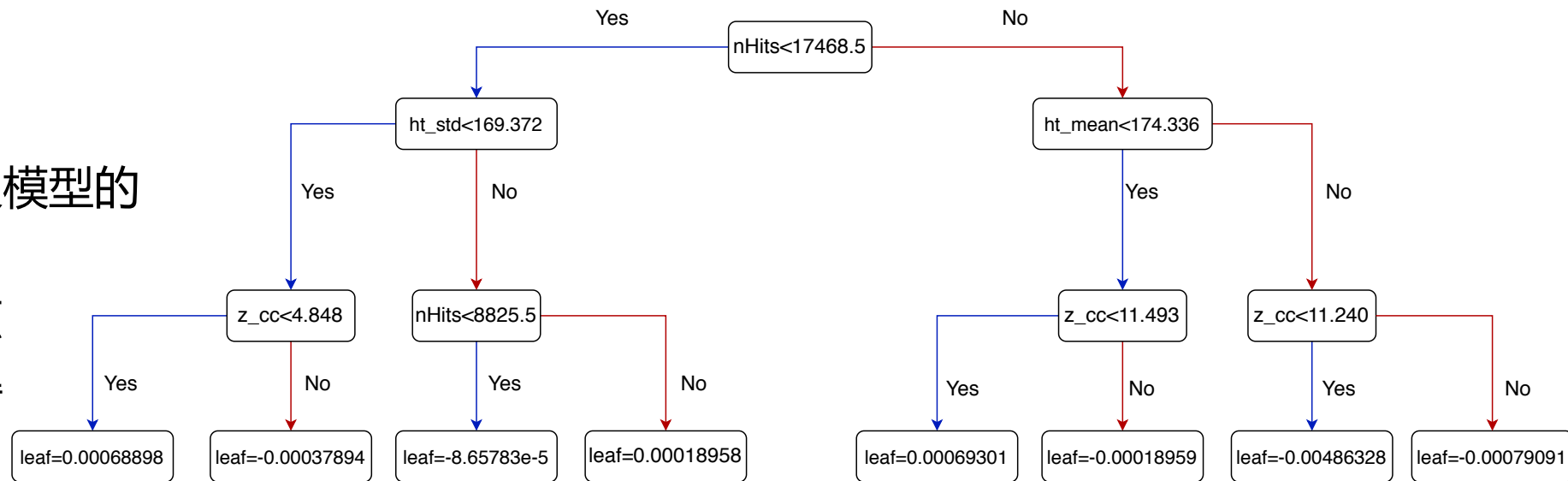
- 第一击中时间的平均值与方差
- 总的电荷
- 电荷的重心 (x, y, z) , (r)

附录见更多信息

提升决策树 (BDT)

- 输入：总电荷数，电荷 (x, y, z)、(r) 的重心，第一击中时间的平均值和方差

- 决策树是多个简单二分叉模型的树形排列的集合
- 叶节点**是每个分支的终点
- 输出值**从叶节点选择而得

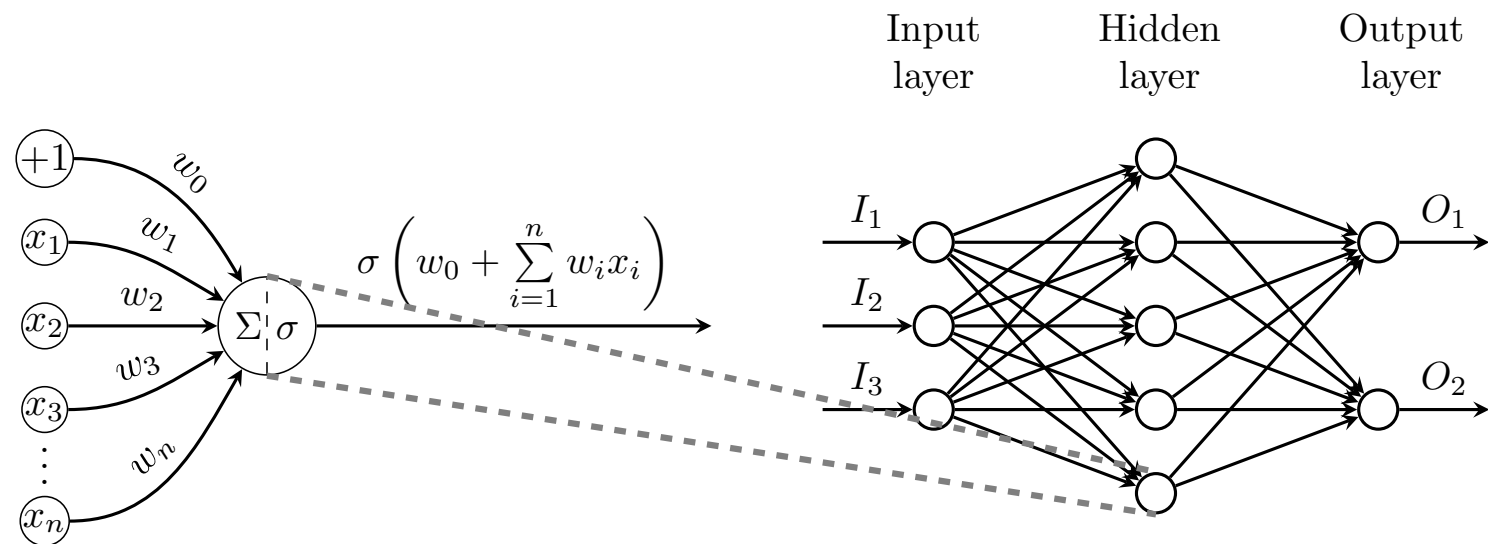


四层树的示意图

- 提升决策树**：多个树被按顺序构建，多个树共同组成最终模型
 - 每一个子树将学习上一个树学习到的残差去修正错误
- 最终输出由所有树输出结果相加而得
 - 模型最大树深为 10，树的个数约150

● 全连接深度神经网络 (FC-DNN)

- 输入：总电荷数，电荷 (x, y, z) 、 (r) 的重心，第一击中时间的平均值和方差



神经元与神经网络示意图

- 全连接神经网络由一系列 f_w 线性函数作为“神经元”排列组成
- 其构筑理念受到生物大脑神经网络功能的运作启发

- 为了发挥所选聚合特征输入的全部潜力，我们对多个超参数进行了优化：

- 网络的深度与宽度
 - 深7, 宽32
- 激活函数
- 权重初始化
- 学习策略
- 层标准化
- 优化器
- 批大小
- 学习率
- 损失函数

● 卷积神经网络 (CNN)

- **输入**：PMT 的电荷、第一击中时间
- 卷积神经网络引入卷积层处理一维或多维数据，最常用于分析视觉图像

- **卷积层**：

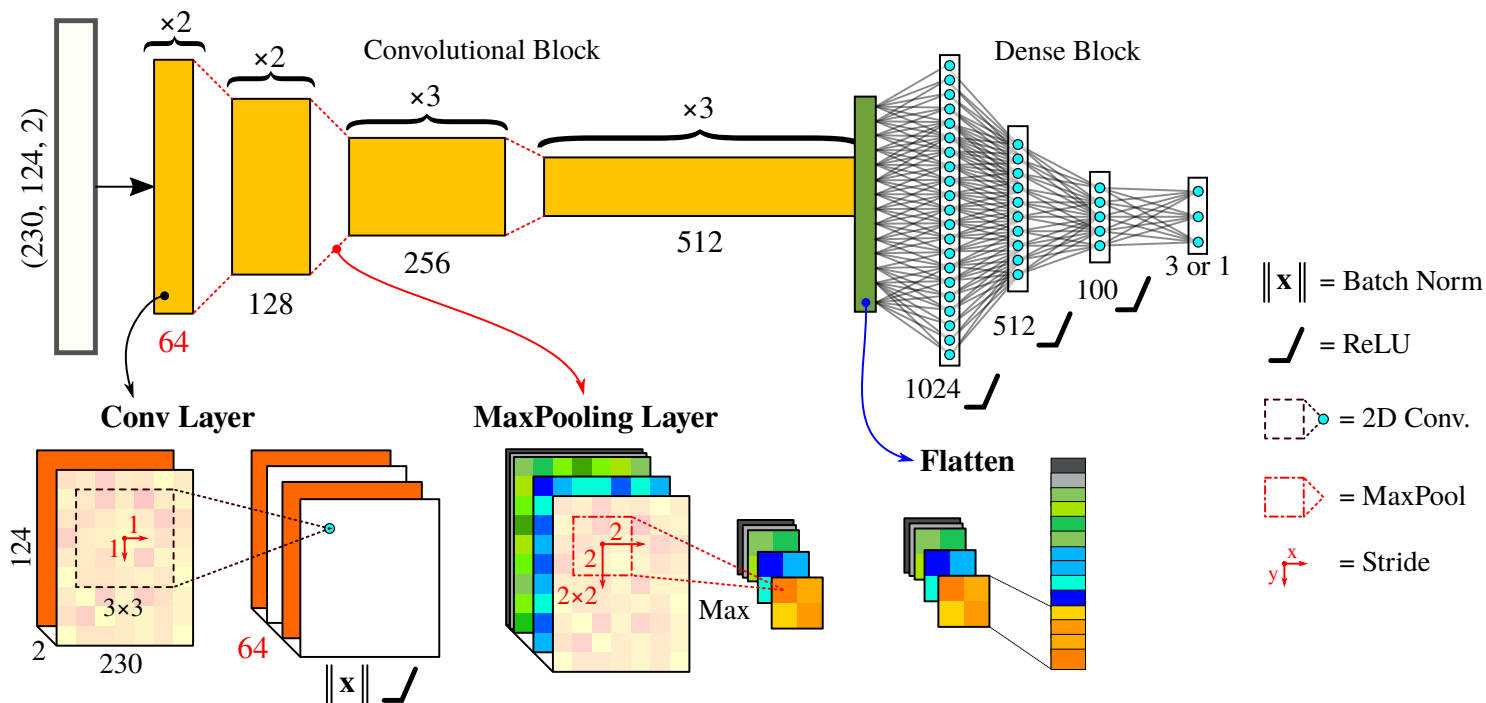
- 有效特征保留
- 权值共享

- **池化层**：

- 数据降维
- 防止过拟合

- **全连接层**

- 增加拟合能力
- 结果输出



两大特点：

1. 能够有效的将大数据量的图片降维成小数据量
2. 能够有效的保留图片特征，符合图片处理的原则

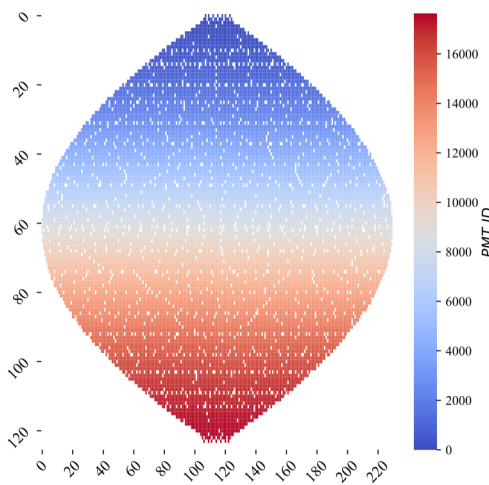
VGG-J——JUNO重建使用的网络之一

输入优化与网络优化

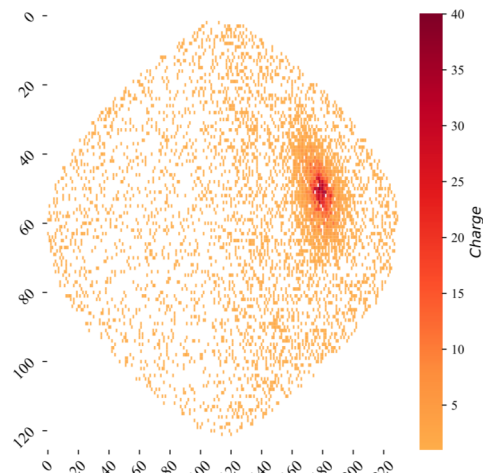
输入：

- 根据JUNO探测器结构设计图像映射方法
- 将击中信息转化为二维图像

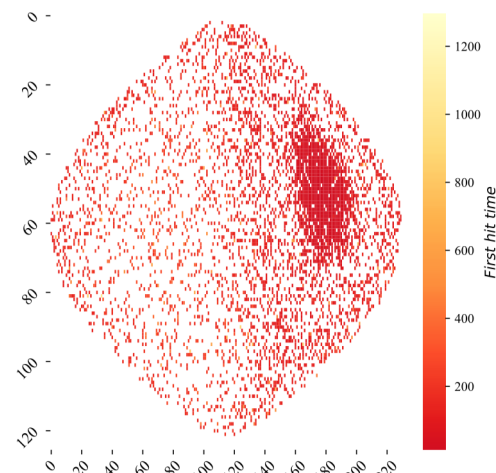
探测器-图像映射关系



PMT击中电荷图像

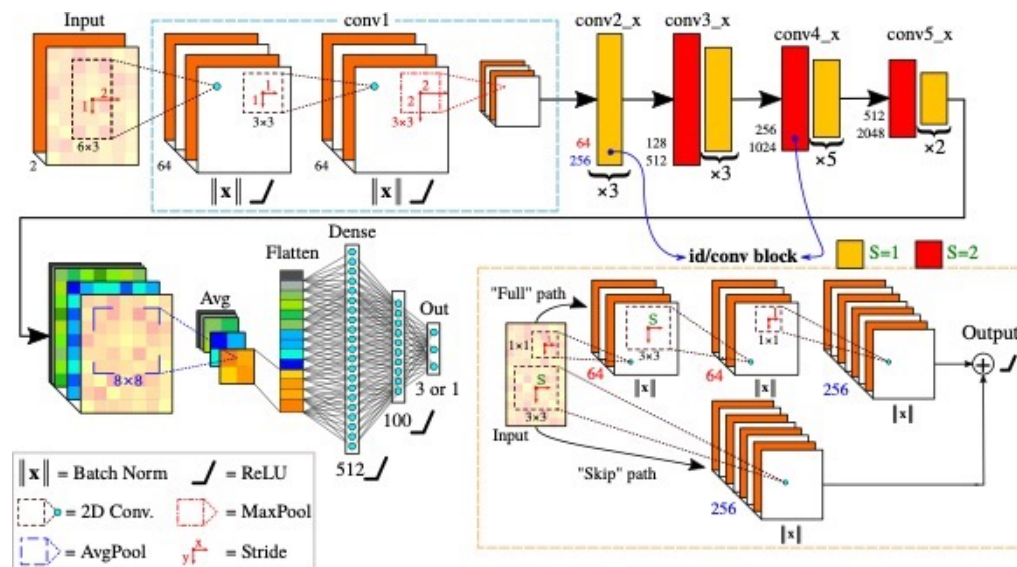


PMT第一击中时间图像



网络：

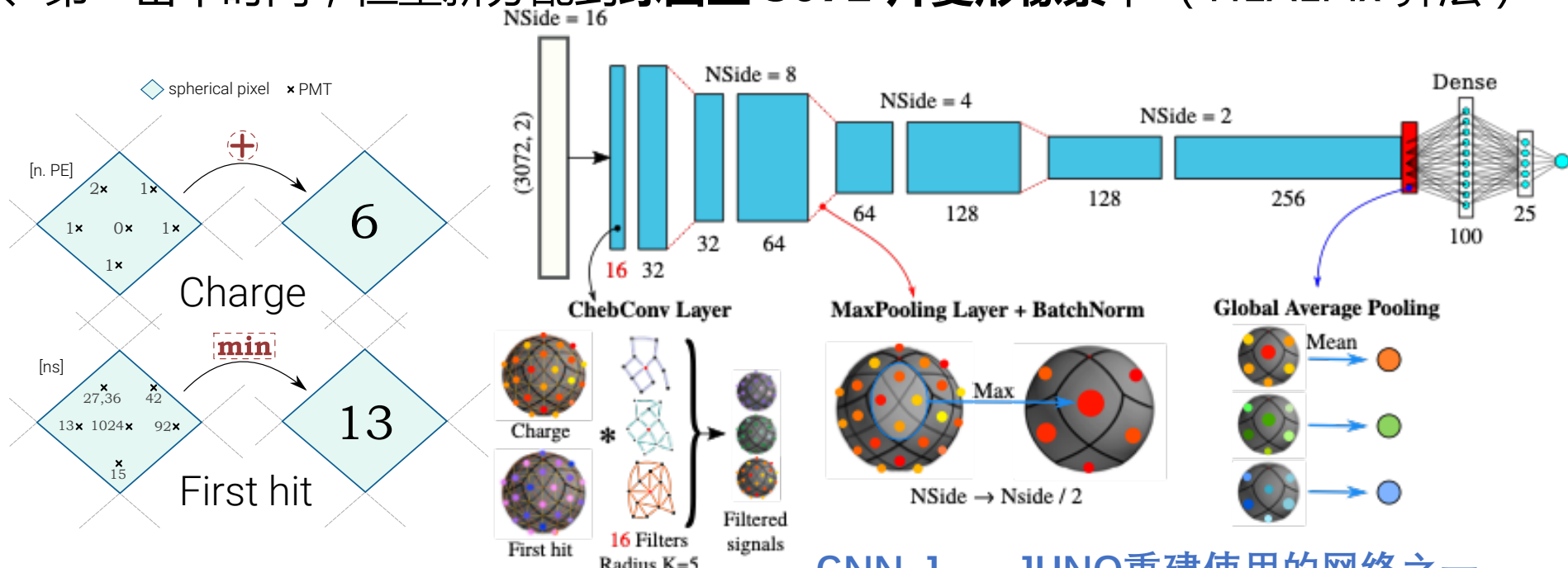
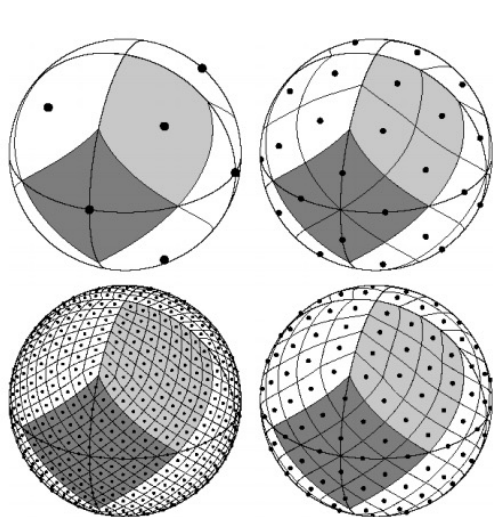
- 结合时下热门的“ResNet-50”网络
- 引入残差学习的概念，进一步加深网络，充分利用输入的信息



ResNet-J——JUNO重建使用的网络之一

图神经网络 (GNN)

- 输入：PMT 的电荷、第一击中时间，但重新分配到球面上 3072 片菱形像素中（HEALPix 算法）



GNN-J——JUNO重建使用的网络之一

- 图神经网络：球面上的像素可作为图神经网络中的节点；卷积可以在图上实现

优势：

- 球面符合探测器几何构造
- 不存在经过 2 维投影到平面的变形

劣势：

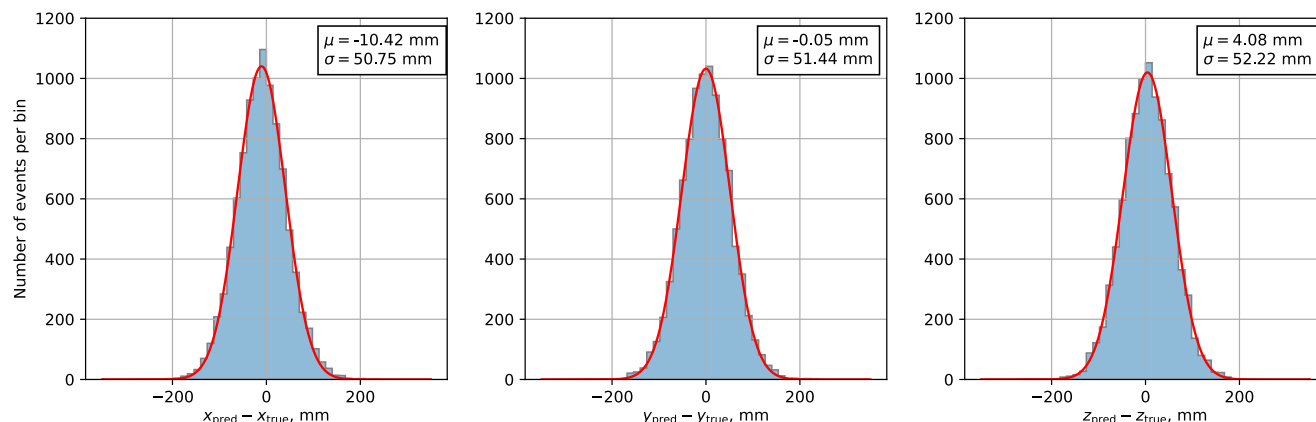
- 图神经网络计算效率不如 2D CNN
- 像素不多

重建性能的评估

- **测试集**：一批模型训练中没有见过的数据；11 组 × 10k 正电子事例，能量为 0, 1, ... 10 MeV
- 训练完成后，我们将测试集交给模型并分析结果：
 - 预测值用高斯拟合，标准差给出分辨率

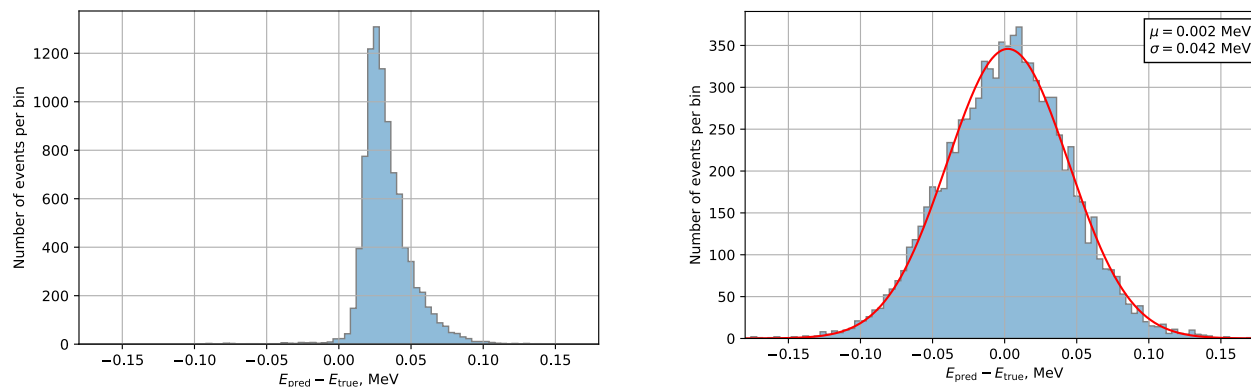
- **顶点**：能量沉积的平均位置

$$\vec{\mathbf{r}}_{\text{dep}} = \frac{\sum_i^{N_s} \vec{\mathbf{r}}_{s,i} E_{\text{dep},i}}{\sum_i^{N_s} E_{\text{dep},i}}$$



顶点重建三个方向的预测值分布

- **能量**：液闪中沉积能量
 - 发现在样本能量边缘的能量重建，结果没法被高斯拟合

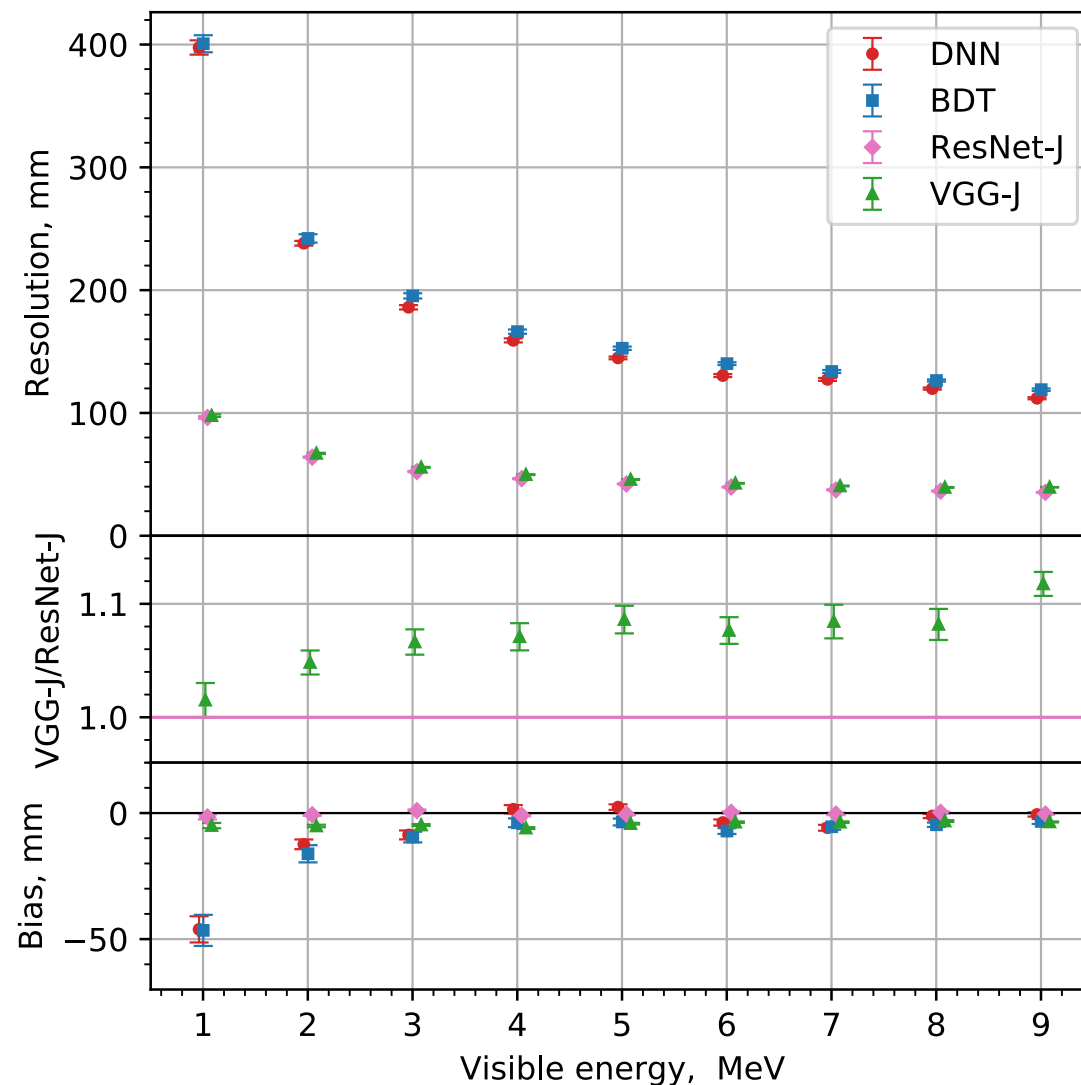


能量重建的分布，右图：边缘，左图：非边缘

顶点重建性能

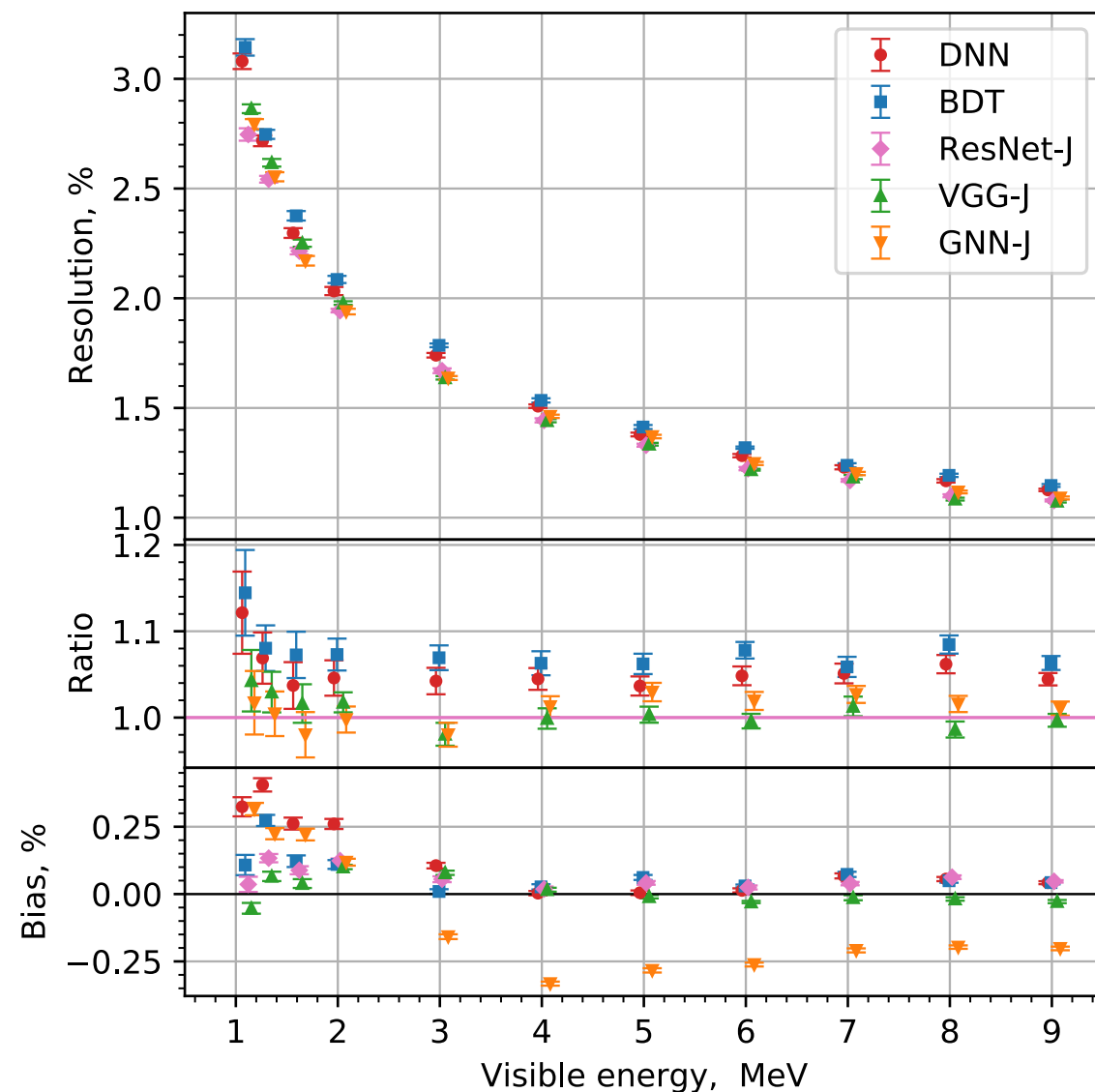
- 顶点重建的在半径方向上的**分辨率与偏差**：

- 复杂模型更有效；简单模型无法提供好的顶点重建表现
- CNN中网络优化策略有效，更深的网络性能更好；
- 卷积神经网络在 1.022 MeV 处分辨率为 10 cm，偏差小于 2 cm，满足实验要求



能量重建性能

- 能量重建的在半径方向上的**分辨率与偏差**：
 - 简单模型能提供相当的表现
 - 复杂模型能获得更好的表现
 - 卷积神经网络在 1.022 MeV 处分辨率小于 3%，满足实验要求



● 计算性能

模型	预测速度, sec/100k events	训练参数个数	参数所占内存, MB	训练时间, min/1M events
提升决策树 BDT	<1	最大树深为 10 树个数约150	17	5
全连接神经网络 FCDNN	<1	6625	0.073	1000
卷积神经网络 VGG-J	155	2.6×10^7	100	1543
卷积神经网络 ResNet-J	235	3.8×10^7	146	840
图神经网络 GNN-J	110	3.5×10^5	4.2	265

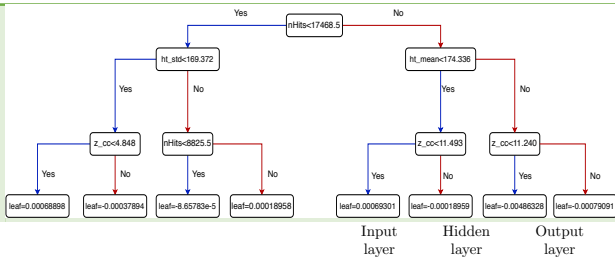
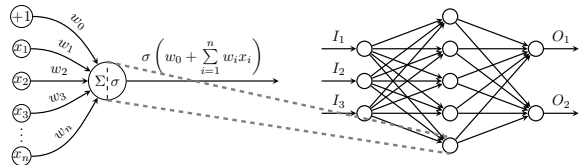
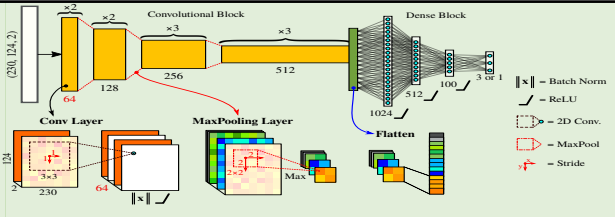
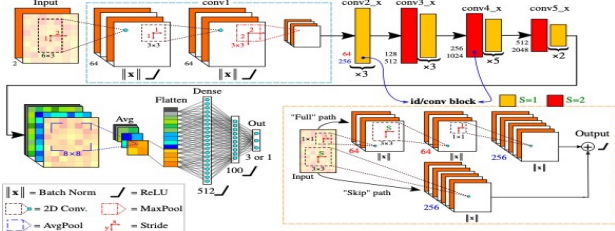
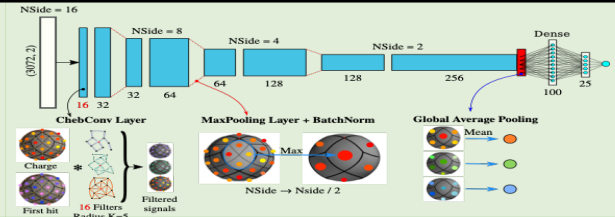
数据预处理时间不包括在内

• 总结与展望

- 在 JUNO 实验的顶点和能量重建上实现了多种机器学习方法
 - **简单模型**：提升决策树 **BDT**，全连接深度神经网络 **FC-DNN**
 - **复杂模型**：平面卷积神经网络 **2D CNN**，图神经网络 **GNN**
- 机器学习方法在能量和顶点重建上能满足实验要求
 - **顶点**：10 cm @ 1.022 MeV, 偏差小
 - **能量**：<3 % @ 1.022 MeV, 偏差小
- 机器学习方法具有高速重建的特点
 - **简单模型**：上千事例/秒/CPU 核
 - **复杂模型**：上百事例/秒/GPU 卡
- **展望**：
 - **提升性能**：优化模型，球面的机器学习方法具有提升空间（GNN-17613）
 - **增强可行性**：研究刻度数据训练方法、迁移学习等方法，使模型不完全依赖于蒙卡数据

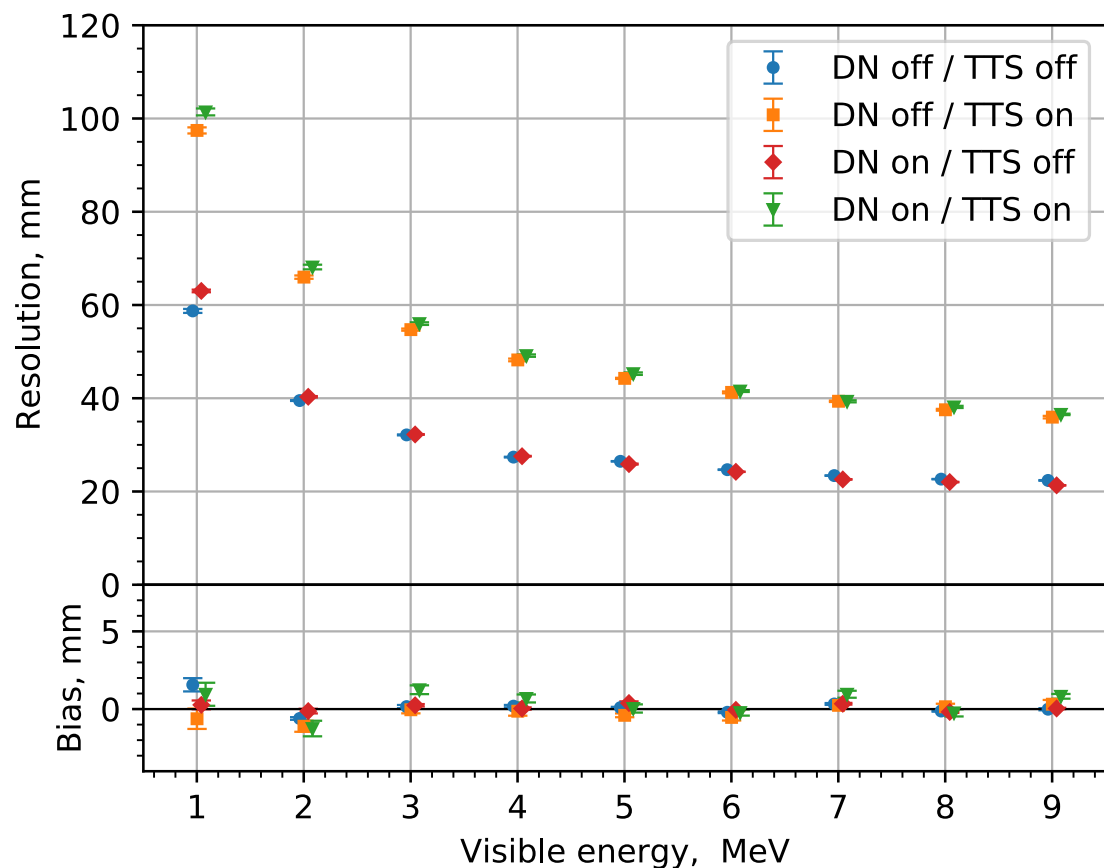
谢谢！

模型总结

模型	输入	特点	图示
BDT	7 个聚合特征	简单、快	
FCDNN	7 个聚合特征	简单、快、深	
VGG-J	17613 个 PMT (电荷与击中时间)	利用全部测量信息	
ResNet-J	17613 个 PMT (电荷与击中时间)	进一步加深网络	
GNN-J	球面 3072 个像素 (PMT的电荷与击中时间)	球面输入、球面卷积	

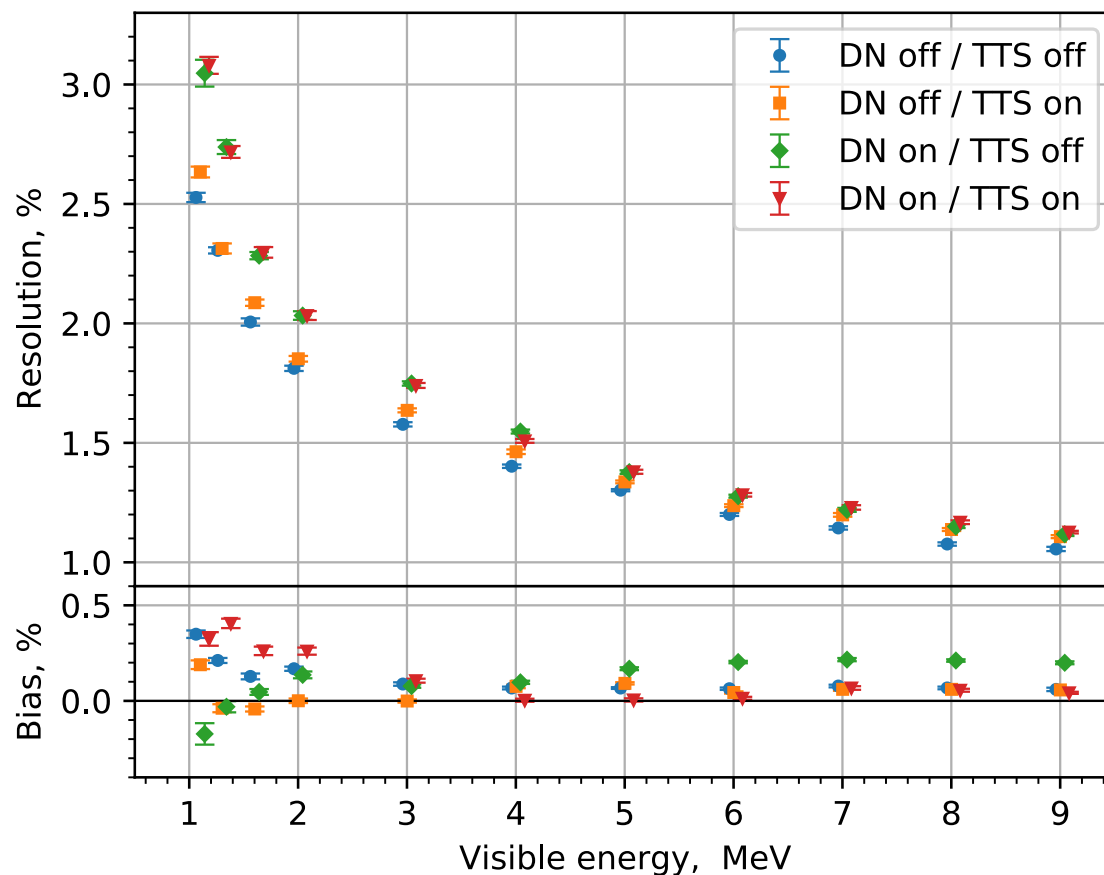
暗噪声与时间越渡弥散的影响

顶点重建 z 方向的重建分辨率与偏差 (ResNet-J 模型)



- 顶点重建受时间越渡弥散 (TTS) 影响更多

能量重建的分辨率与偏差 (DNN模型)



- 能量重建受暗噪声 (DN) 影响更多