Machine Learning Methodologies for CEPC AHCAL

--Zhen Wang

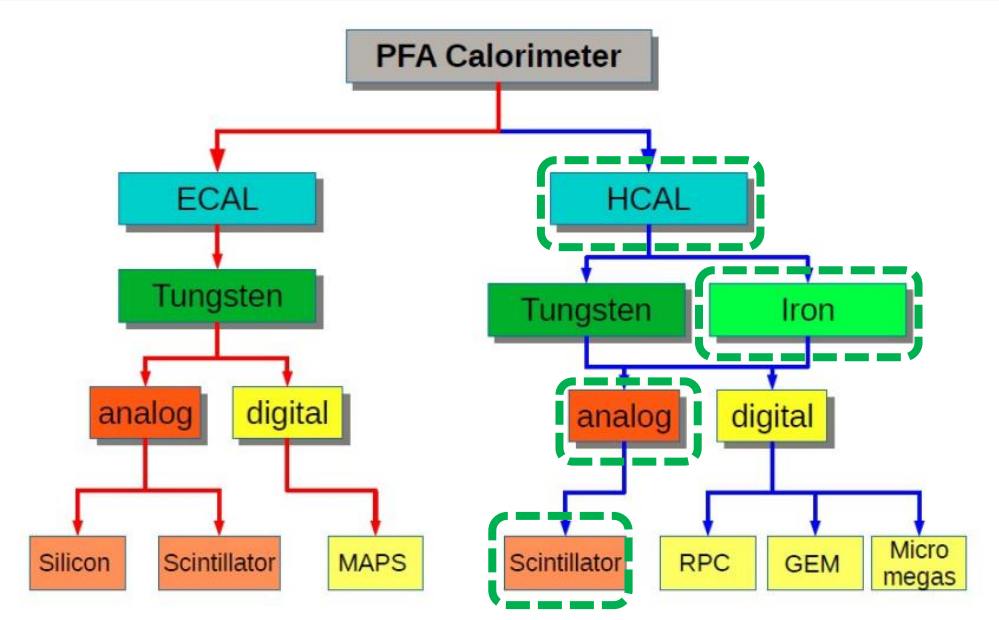
On behalf of AHCAL algorithm team

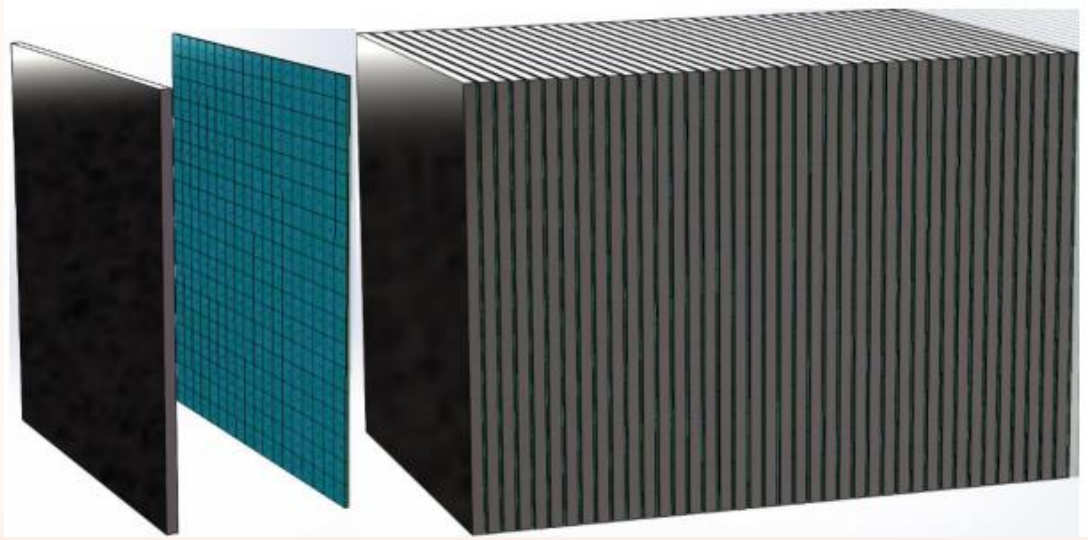
(which only consists of two poor guys)

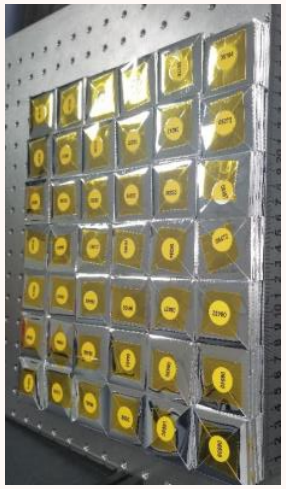
CEPC A(nalog) H(adronic) CAL(orimeter)

- **Particle Flow Algorithm oriented: high granularity**
- **40 sampling layers**

PFA based technologies



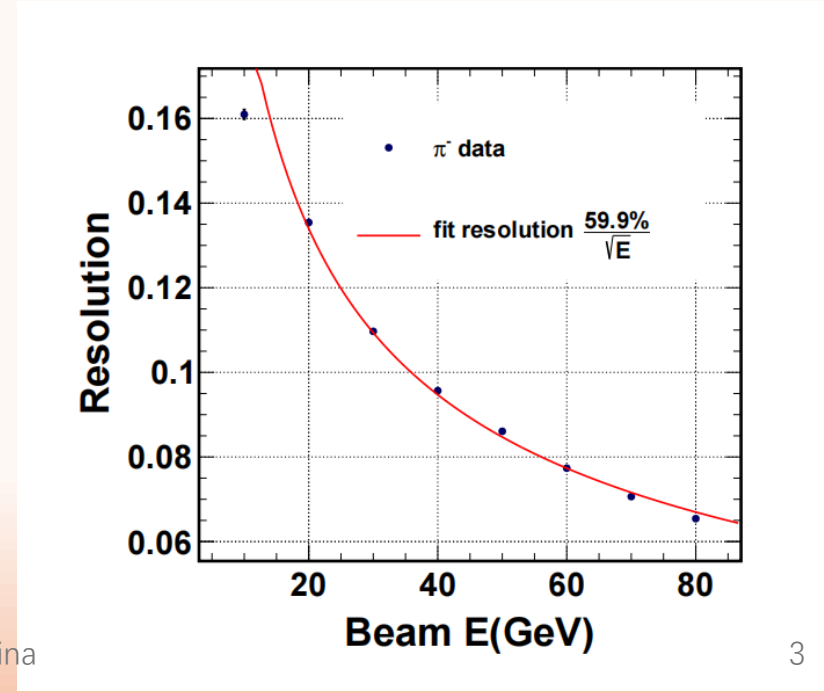




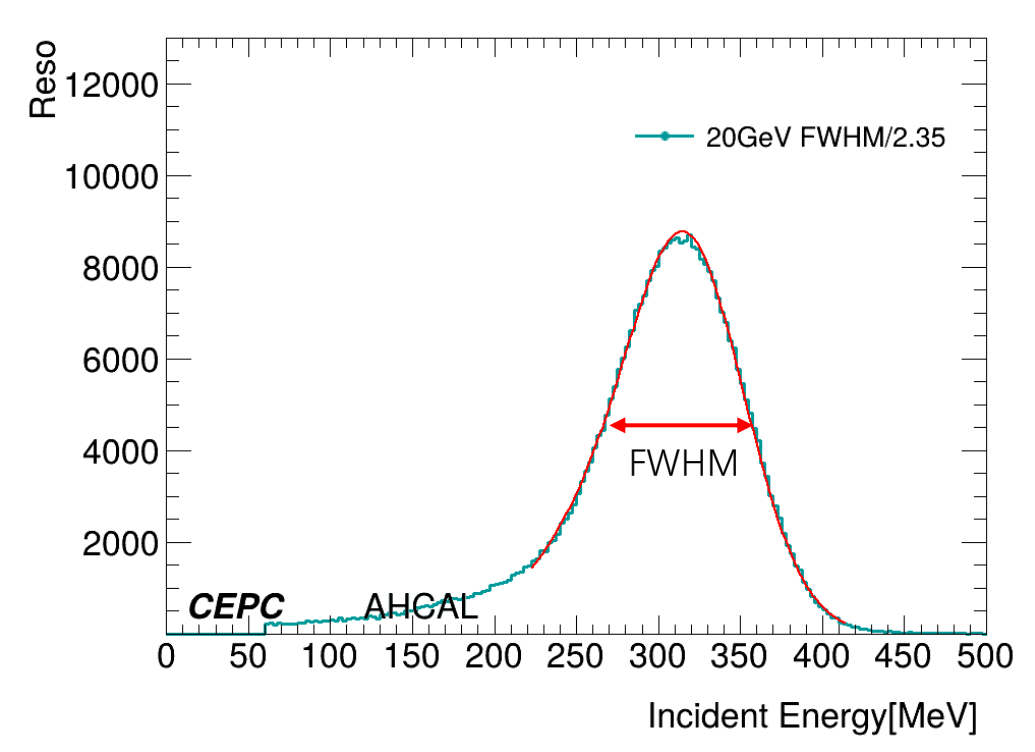
40 [layers] x **18** x **18** [cells/layer]

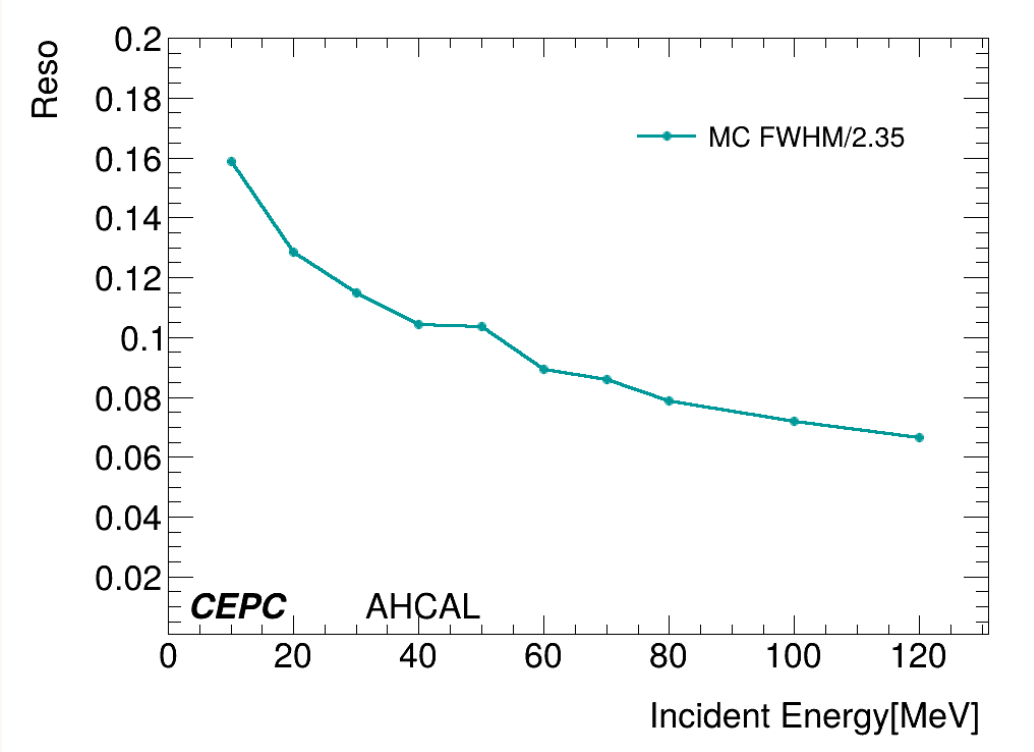
Motivation

- Two beam tests @ CERN for AHCAL prototype
- Analysis of beam test data ongoing …
- Energy resolution meets the targets ($60\%/\sqrt{E}$)



Energy resolution (MC)

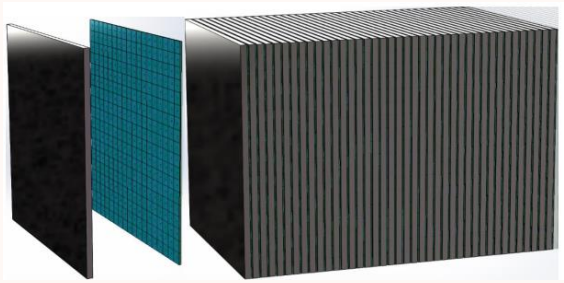




MC simulation provided by Hongbin with different incident energies (10-120 GeV)

Raw MC simulation **without** any selections/cuts

M(ulti) L(ayer) P(erceptron)

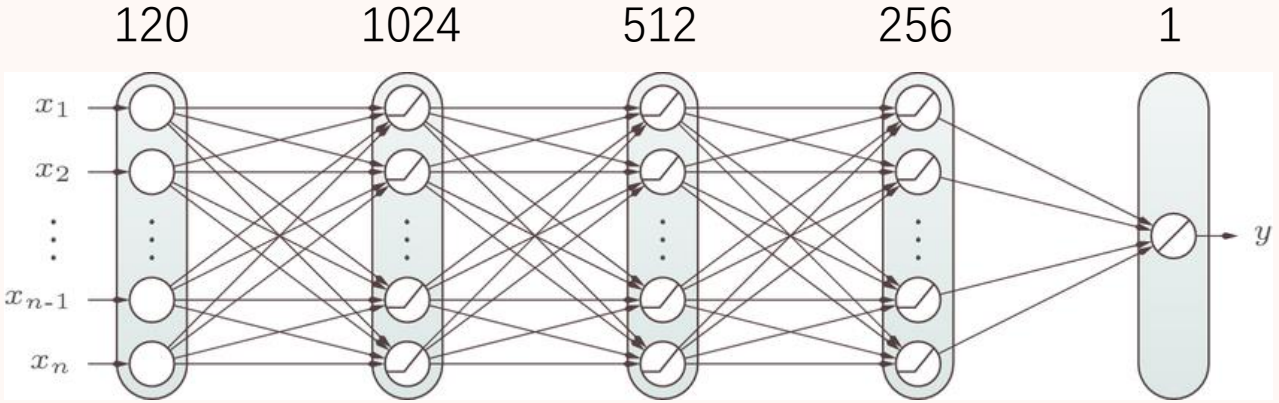


E_{layer}

N_{layer}^{hits}

$\frac{E}{N}$

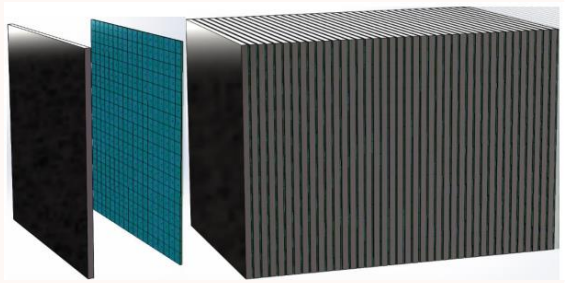
Dropout: 0.456



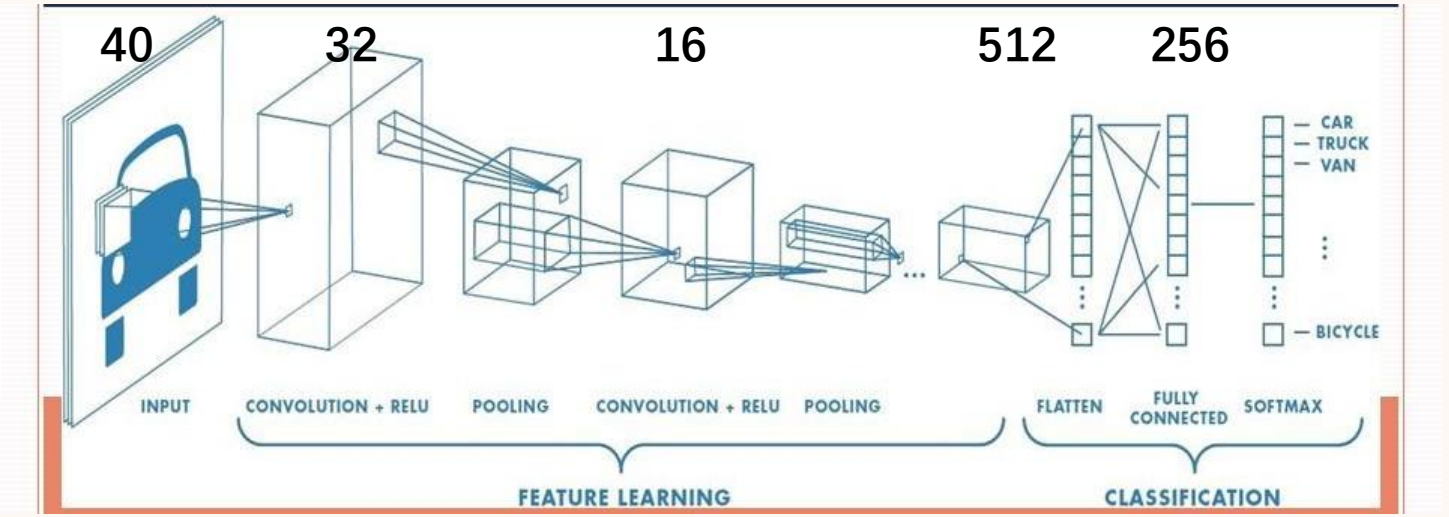
Predicted Energy

- Features: layer energy, layer hits, average hit energies of all 40 layers
- Loss: relative error loss $(y_{pred} - y_{true})^2 / y_{true}^2$
- Targets: truth incident energy

C(onvolutional) N(eural) N(etwork)



40 x 18 x 18

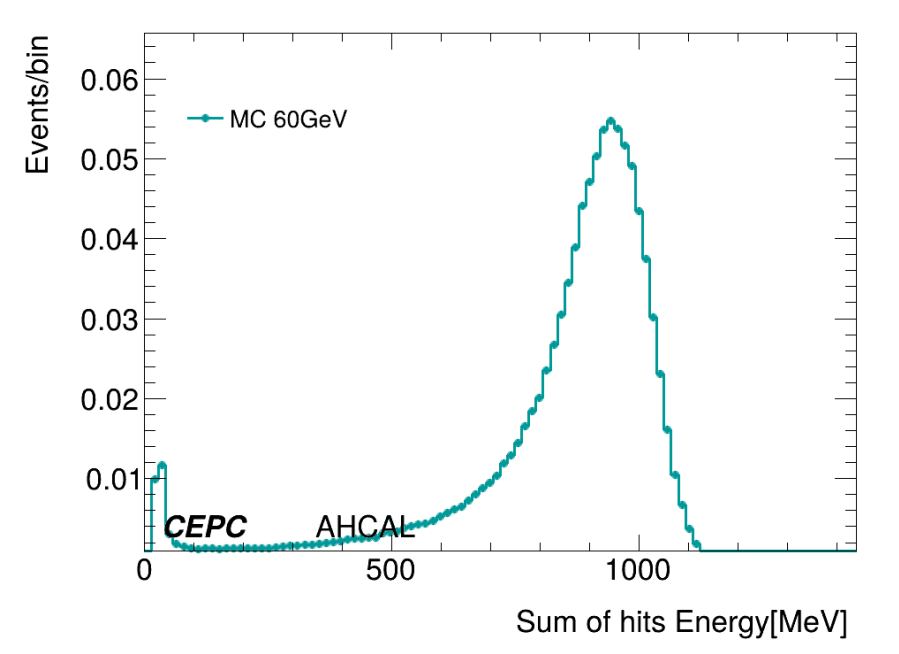


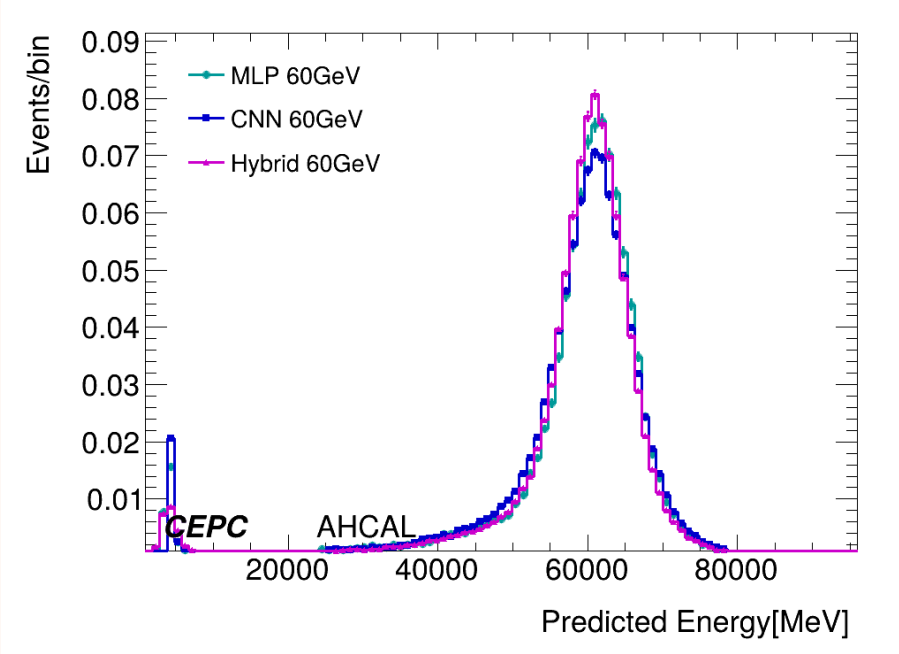
1

Predicted Energy

- **Two convolutional layers with two max pooling layers with RELU activation**
- **Loss: relative error loss $\left(y_{pred} - y_{true}\right)^2 / y_{true}^2$**
- **Fully connected layers using 512,256,1 structure**

Energy distribution @ 60GeV

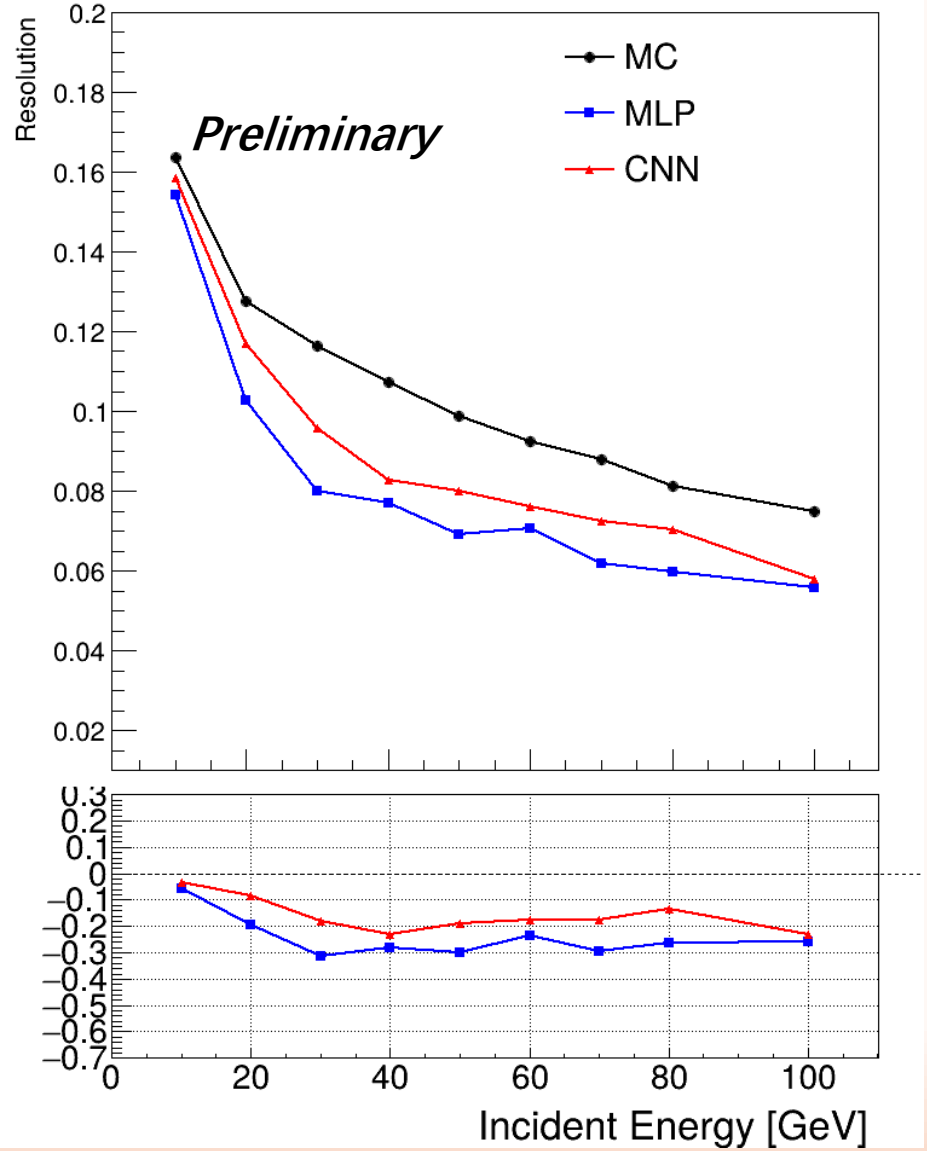




- Predicted energy located at ~60 GeV as expected
- Smaller RMS got with all models
- Events with small energies are still kept

Energy resolution

- General improvements observed ~ 20%
- MLP seems to be better
- CNN extracts the topological powers
- Should study effects on **linearity**



Hybrid model

E_{layer}

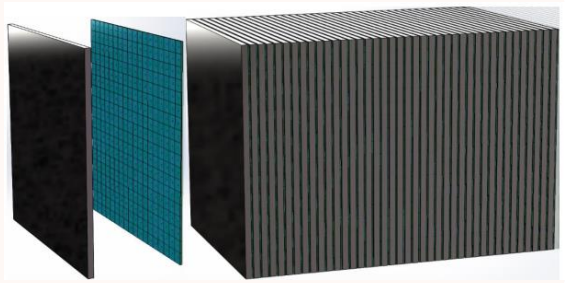
N_{layer}^{hits}

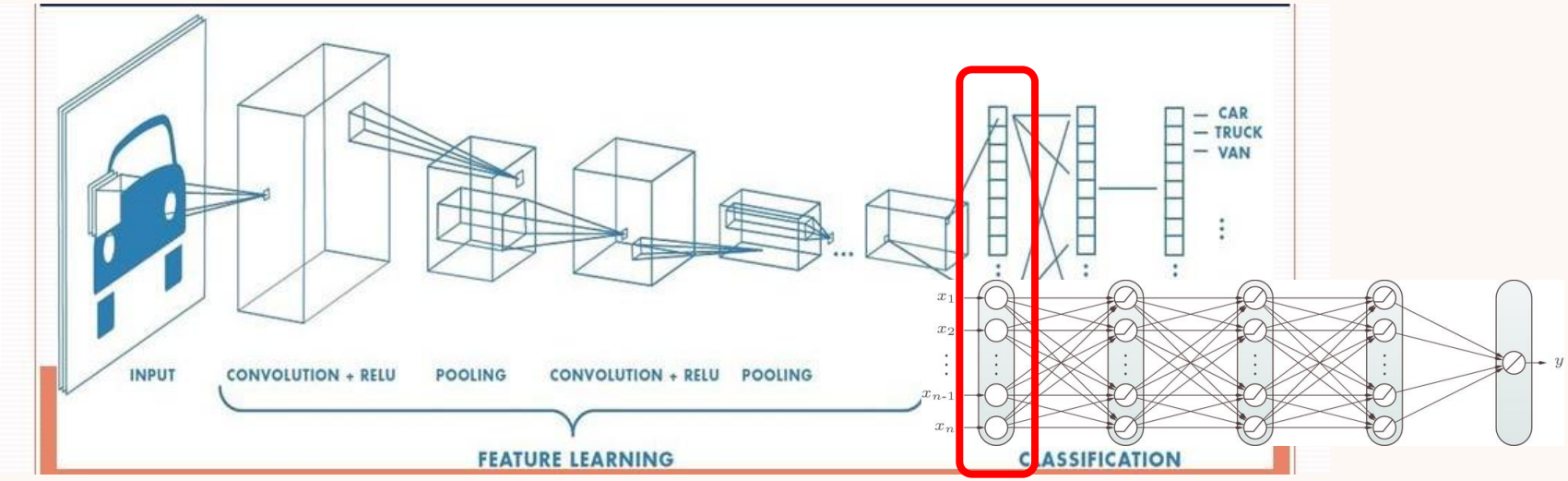
$\frac{E}{N}$

Concatenate **120 features**

CNN Part

32 16 **256** + **120**



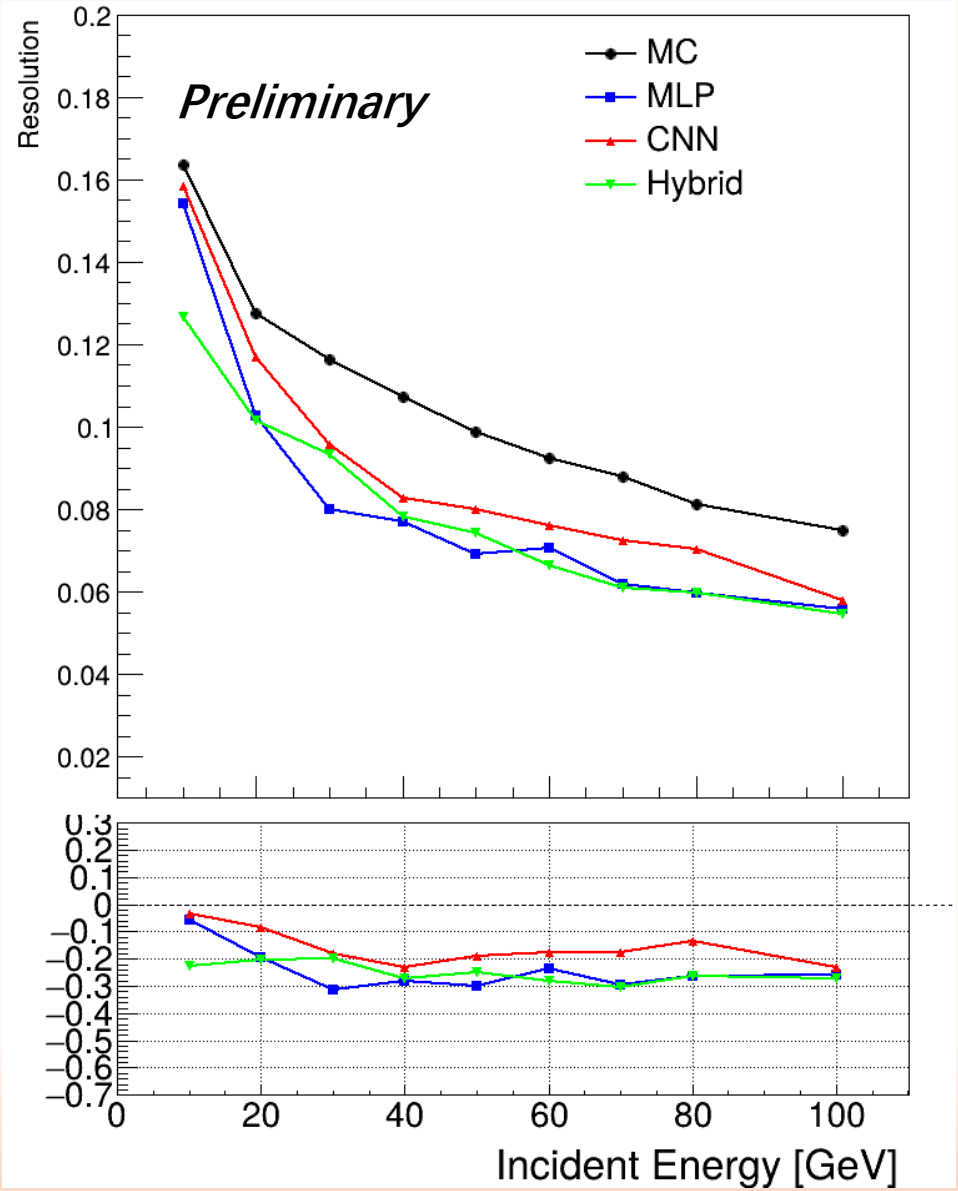


- CNN hyperparameters are reduced to avoid overtraining
- Better results are expected

Energy resolution

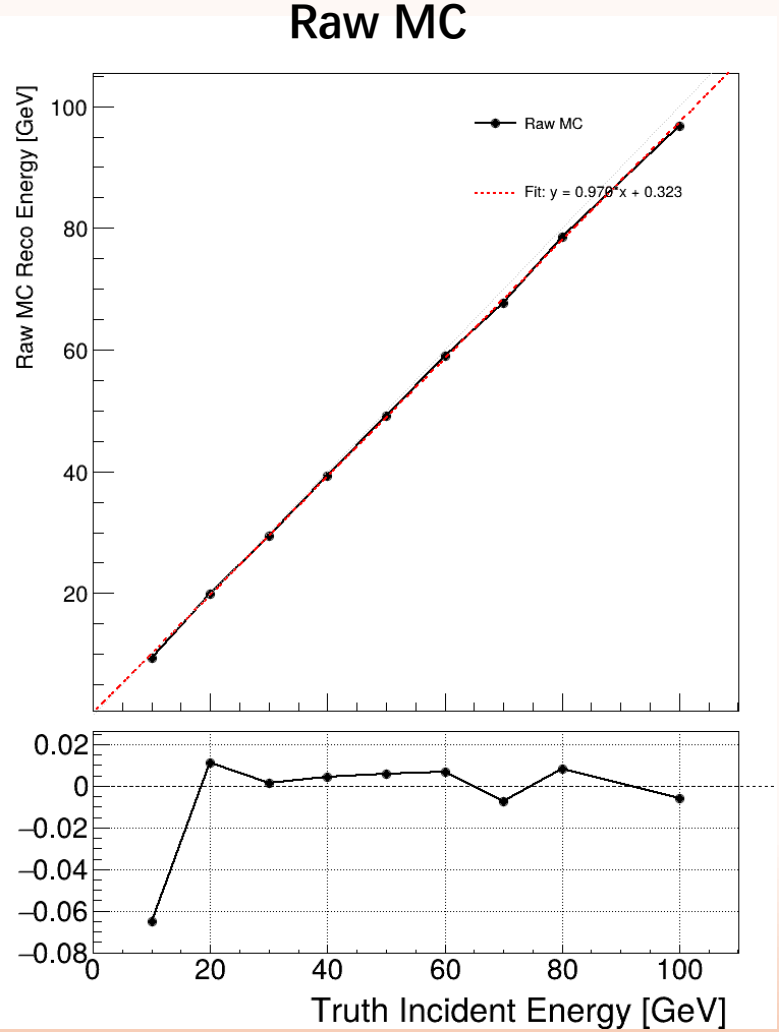
- **All DL Methods Excel:** MLP, CNN, and Hybrid all surpass direct summation
- **Optimal Architecture:** Hybrid Model Achieves **20-30%** Improvement

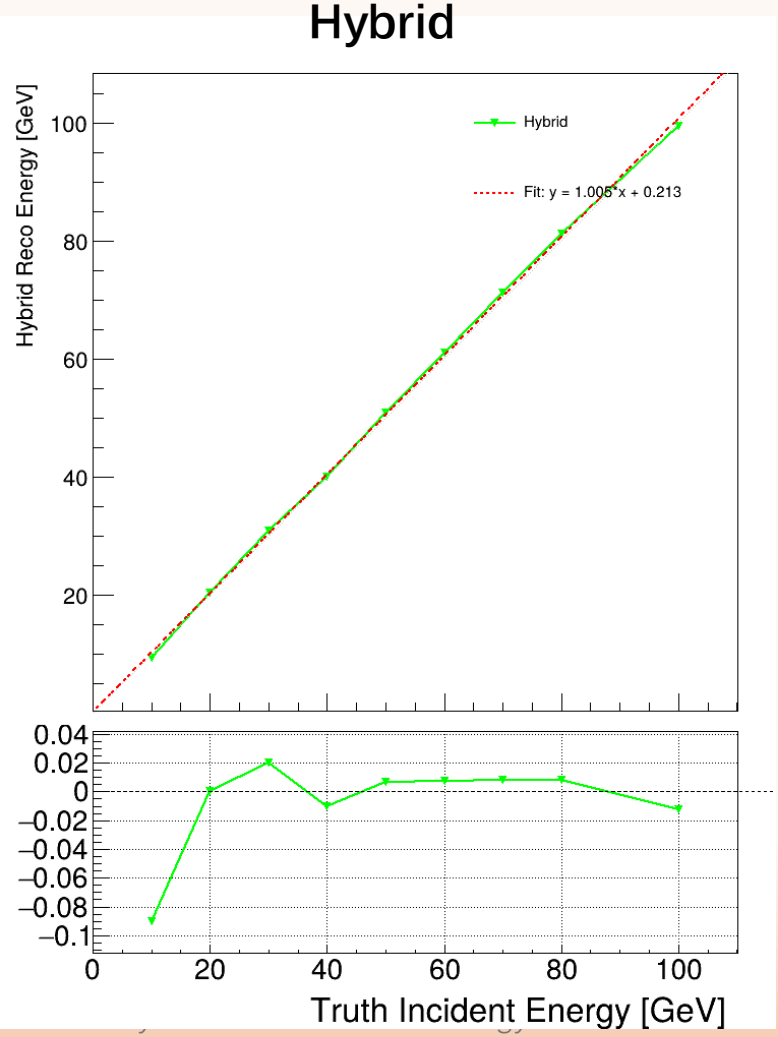
Should examine the linearity right now !



Energy linearity

- Quick check: **linear fit** on reco/predicted energies





Summation of hits energies are multiplied by a factor just for comparison

Acceptably worse than raw summation of hit energies

Summary

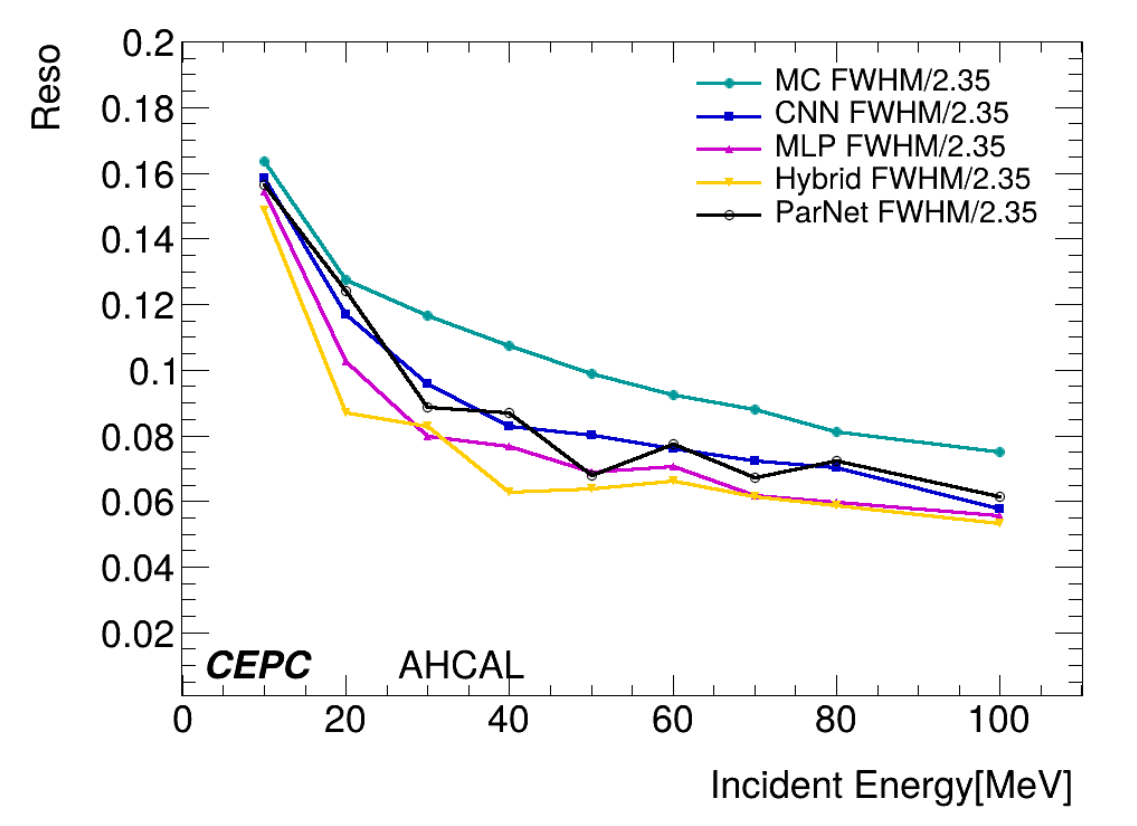
- Deep neural networks developed for regression tasks
- General improvements observed in all models
- **Hybrid** model gives generally **better performance**
- Small energy regions should be taken care of

- Plans:
  - Try different architectures like **GNN, transformers**
  - Optimization of hyperparameters (depends on devices)
  - Test beam data should be blinded until satisfied with MC results

Discussion

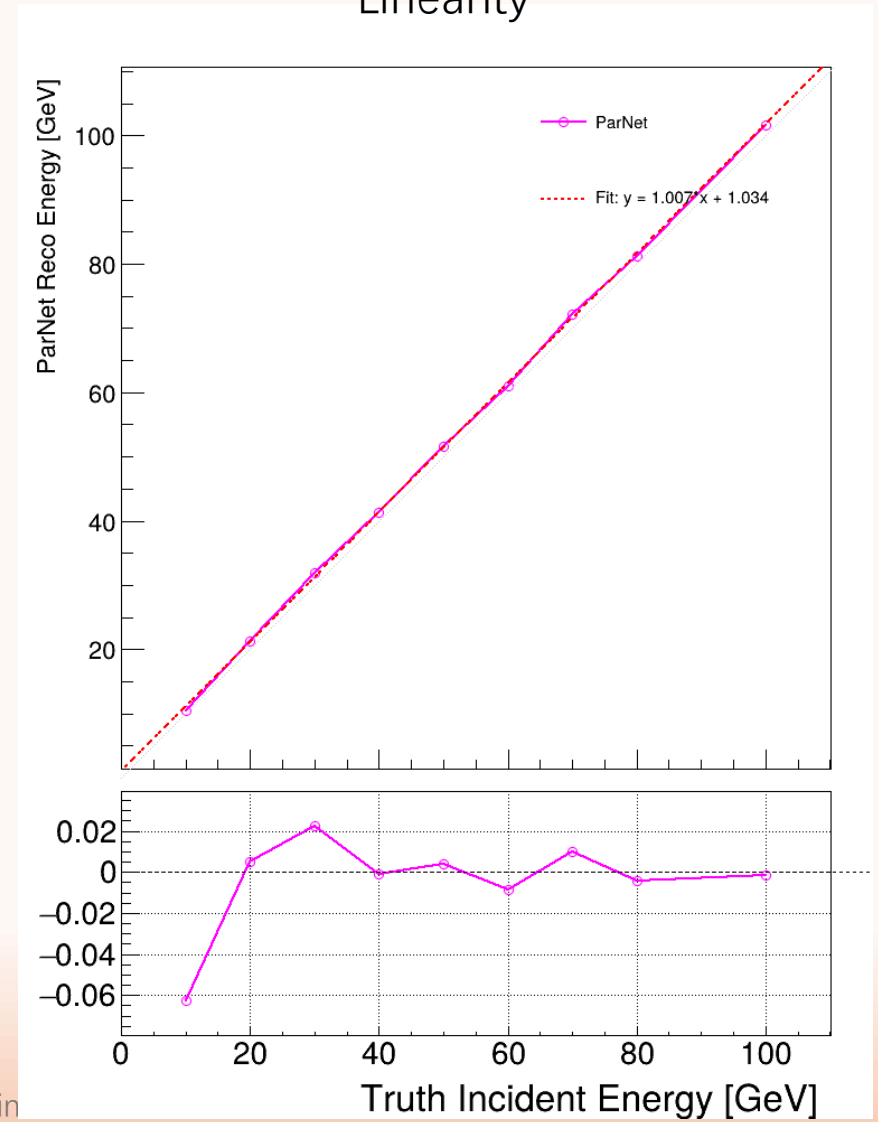
- Better ways for optimization ?
- Grid Search? Optuna? RL? Pre-training? Proper initialization?
- Compare the performance of different level information?

Backup DGCNN

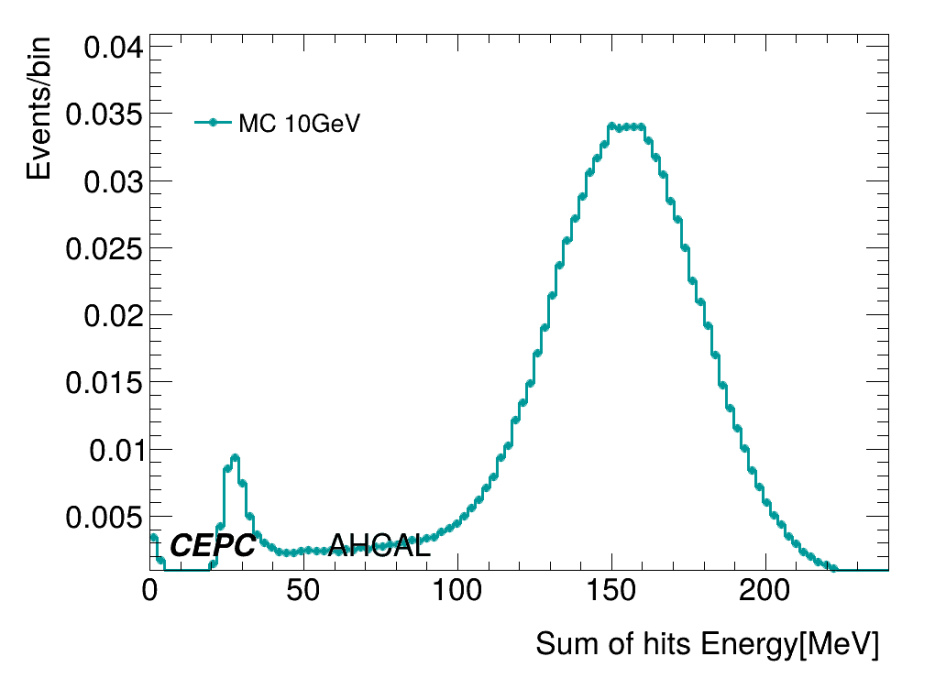


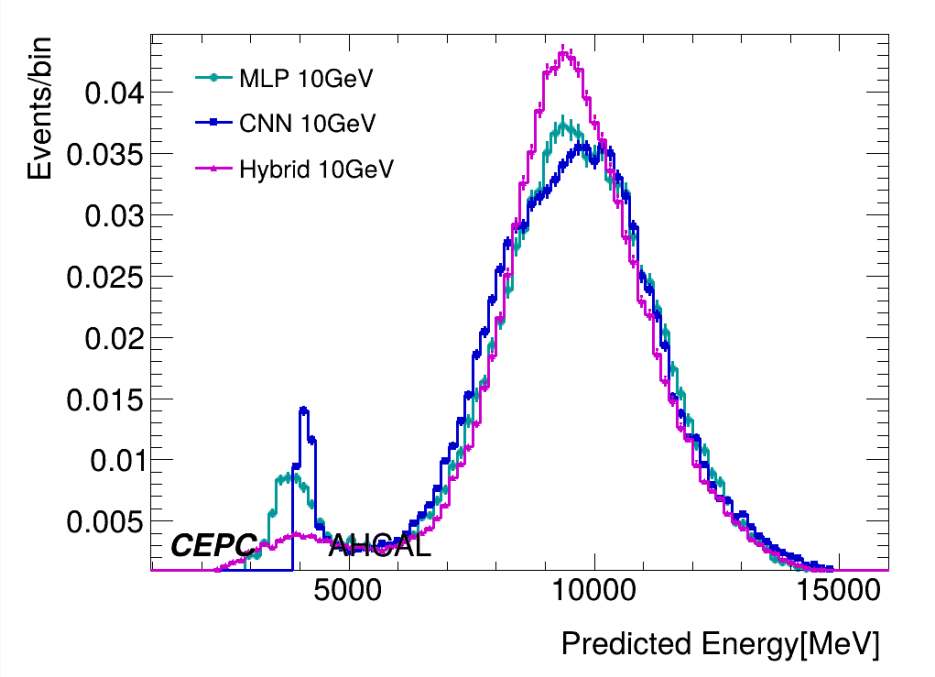
Optimization needed

Linearity

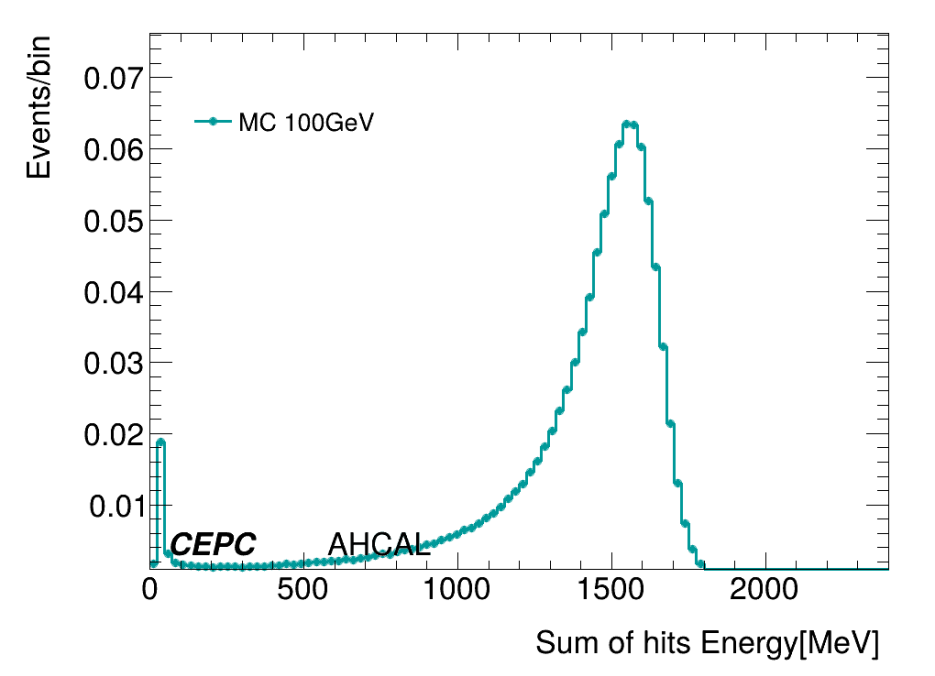


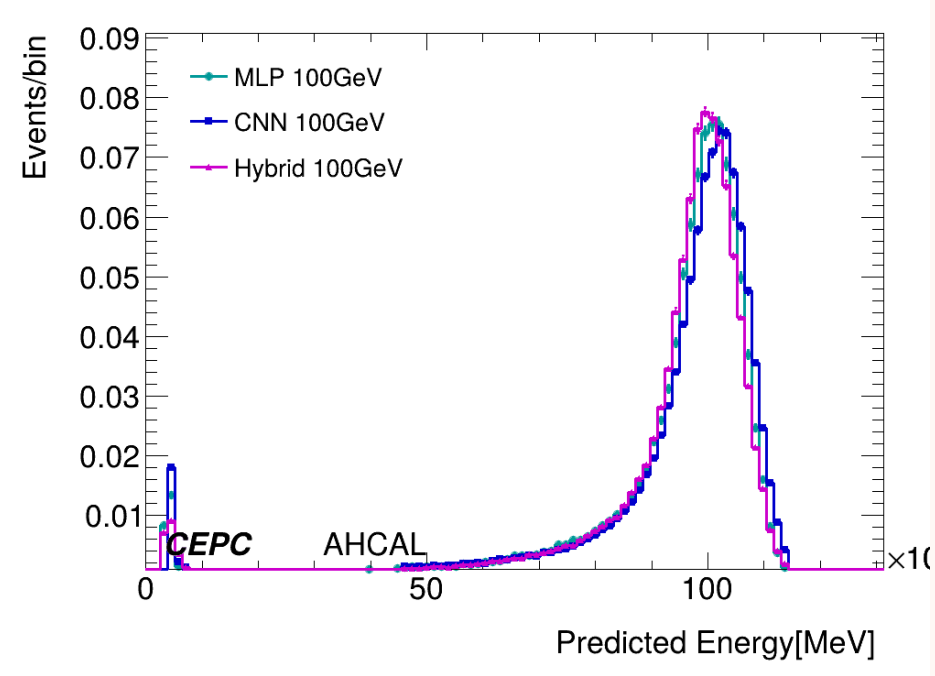
Backup energies



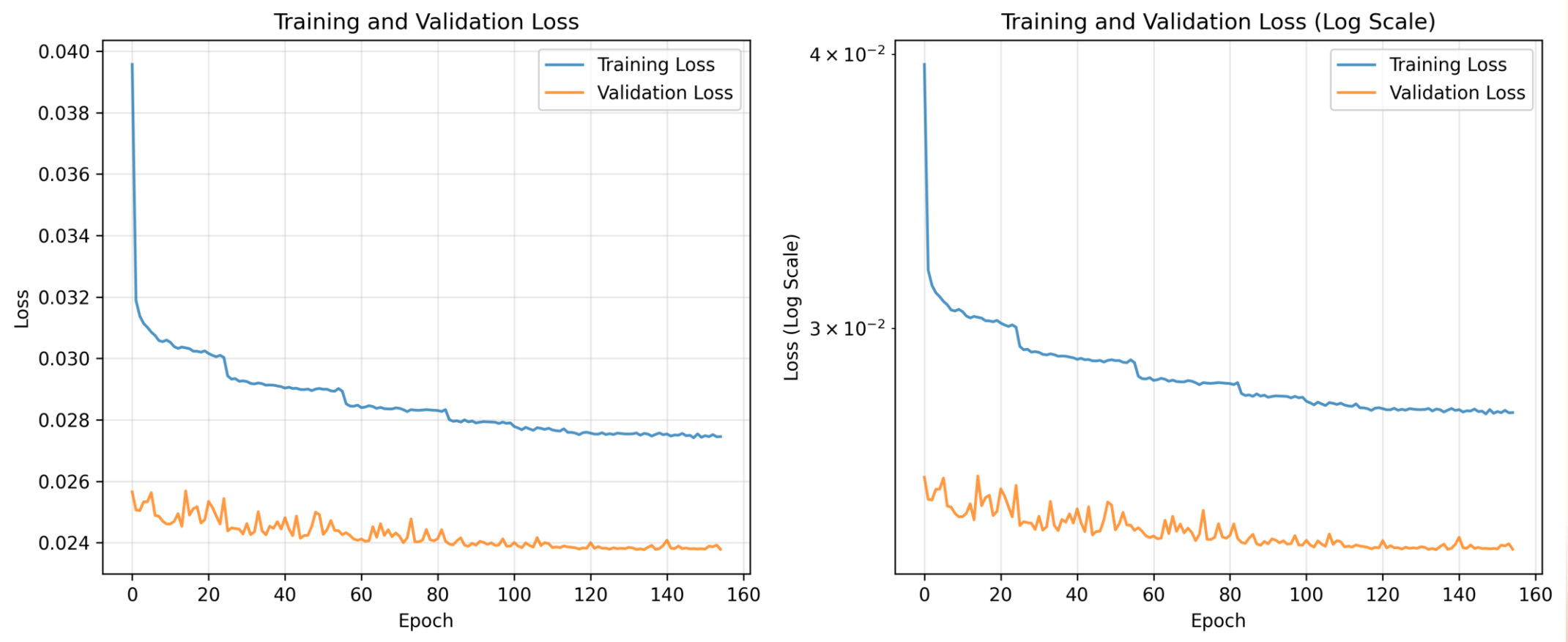


Backup energies

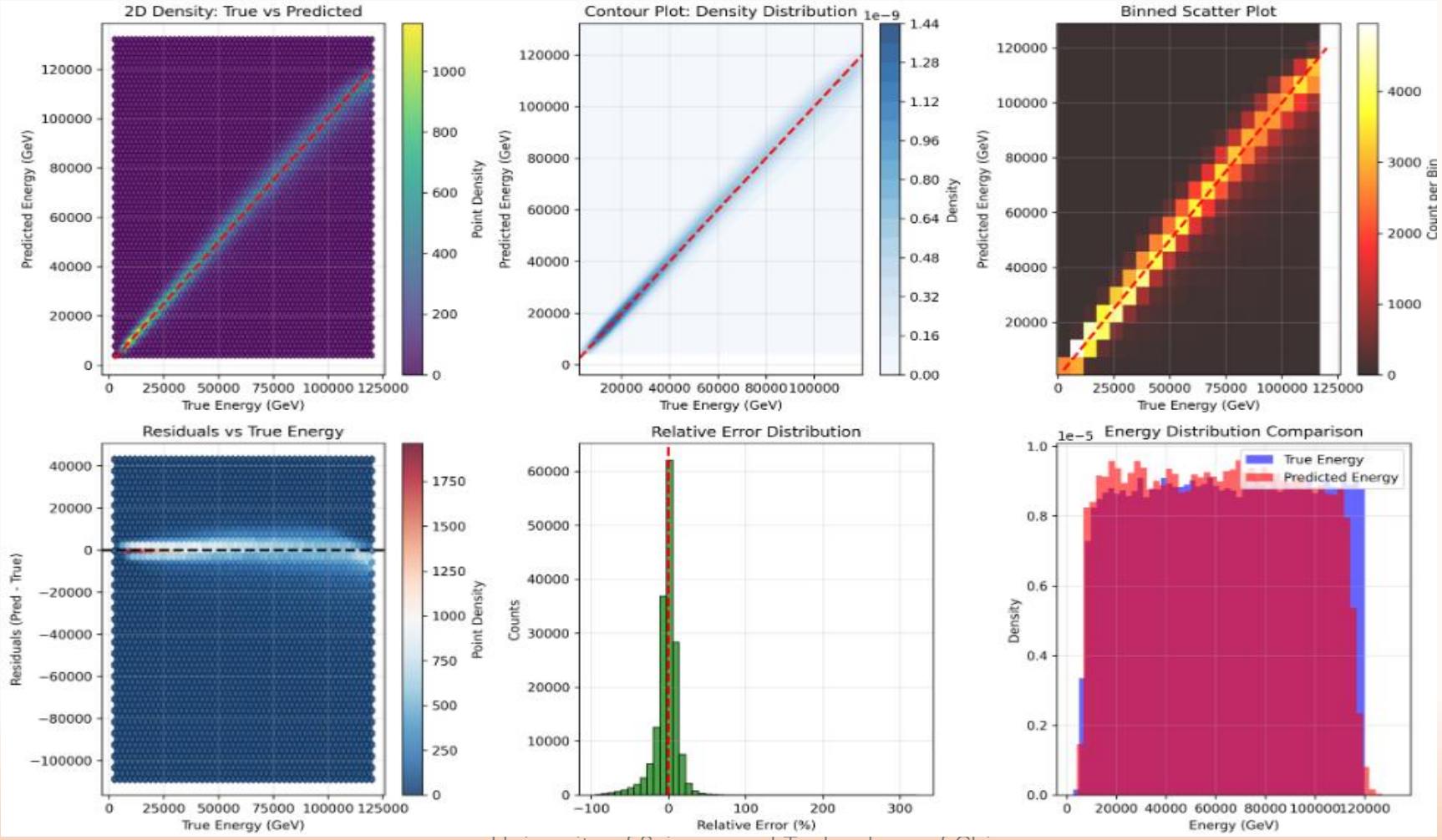




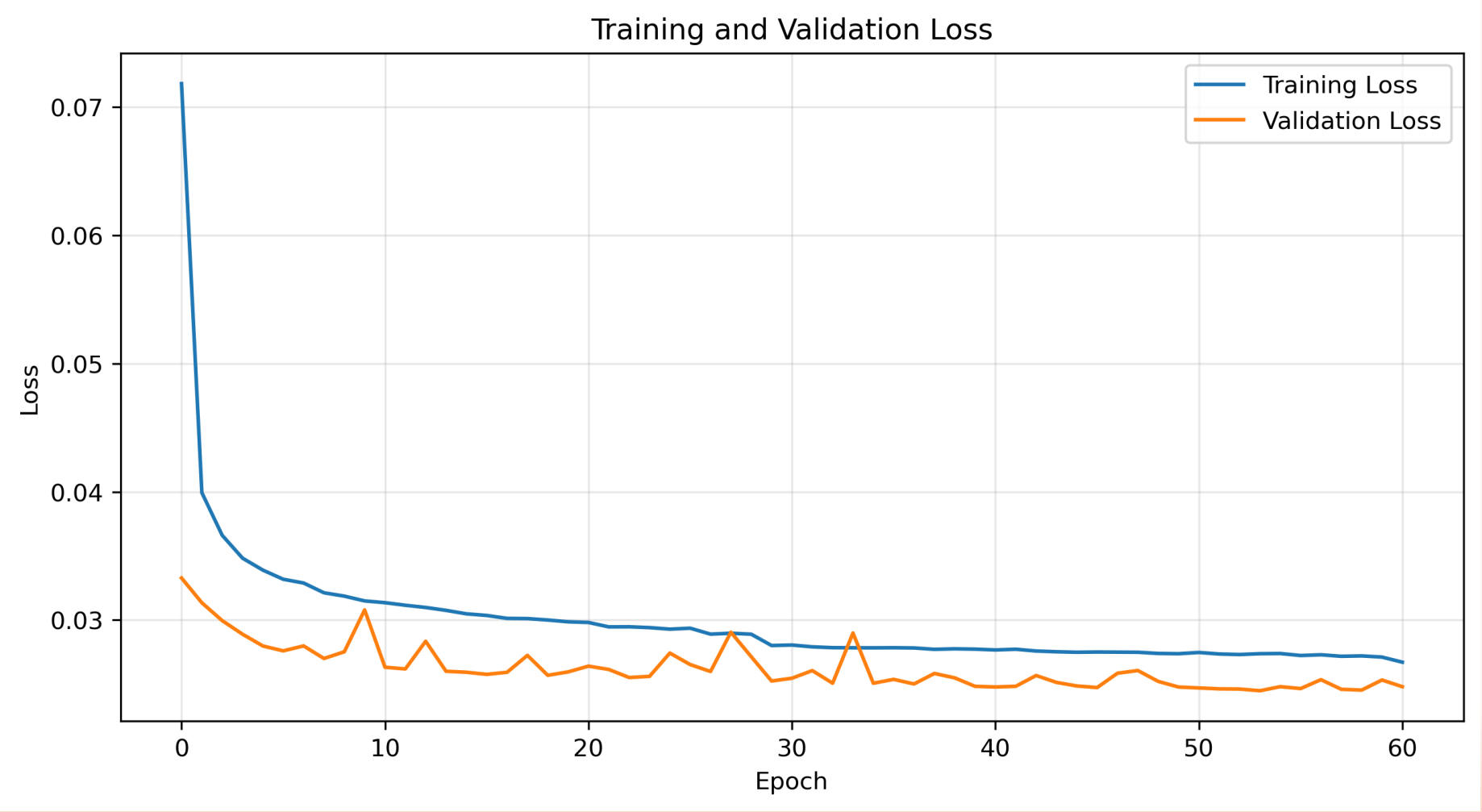
Backup MLP loss curve



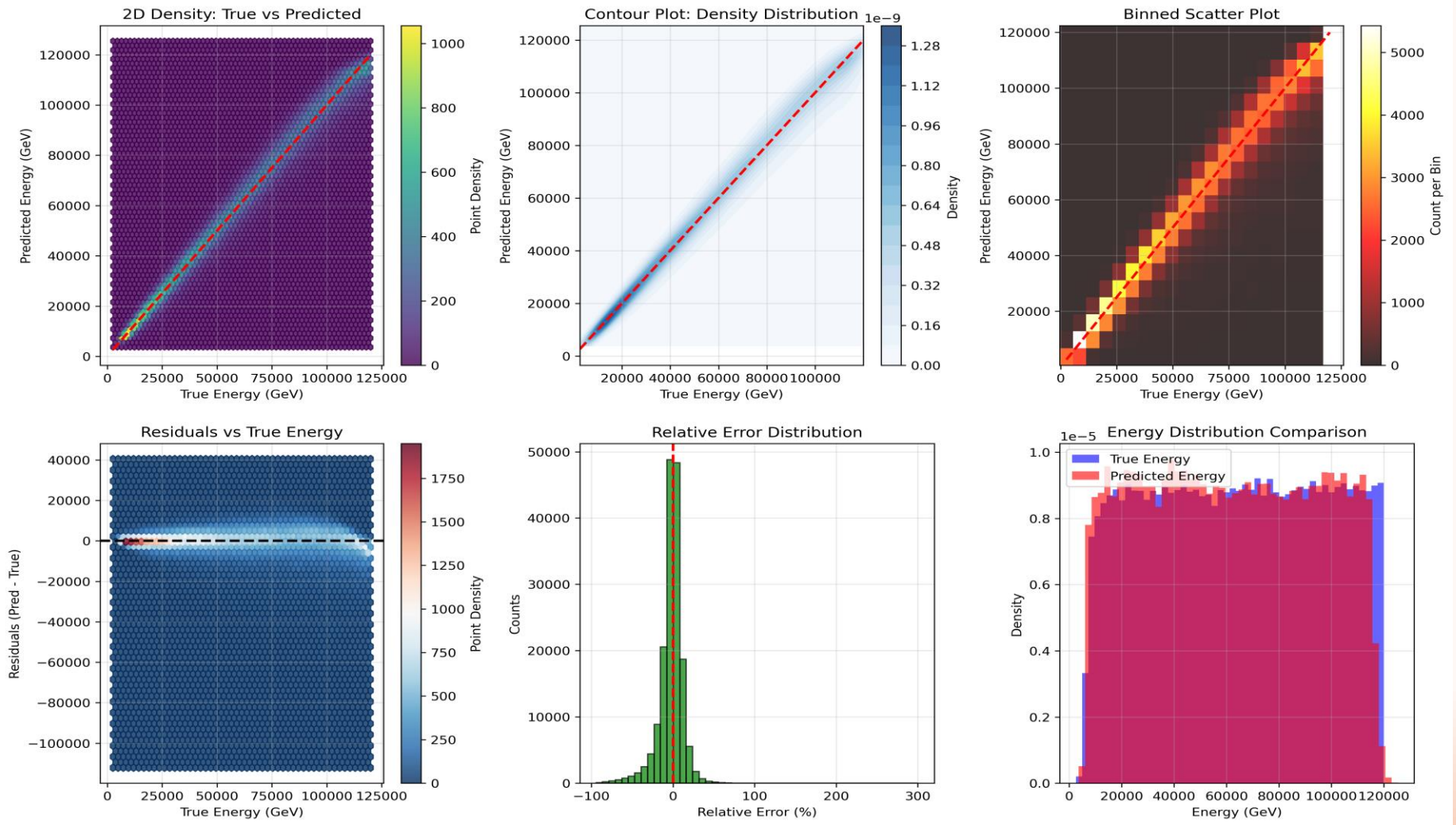
Backup MLP tests



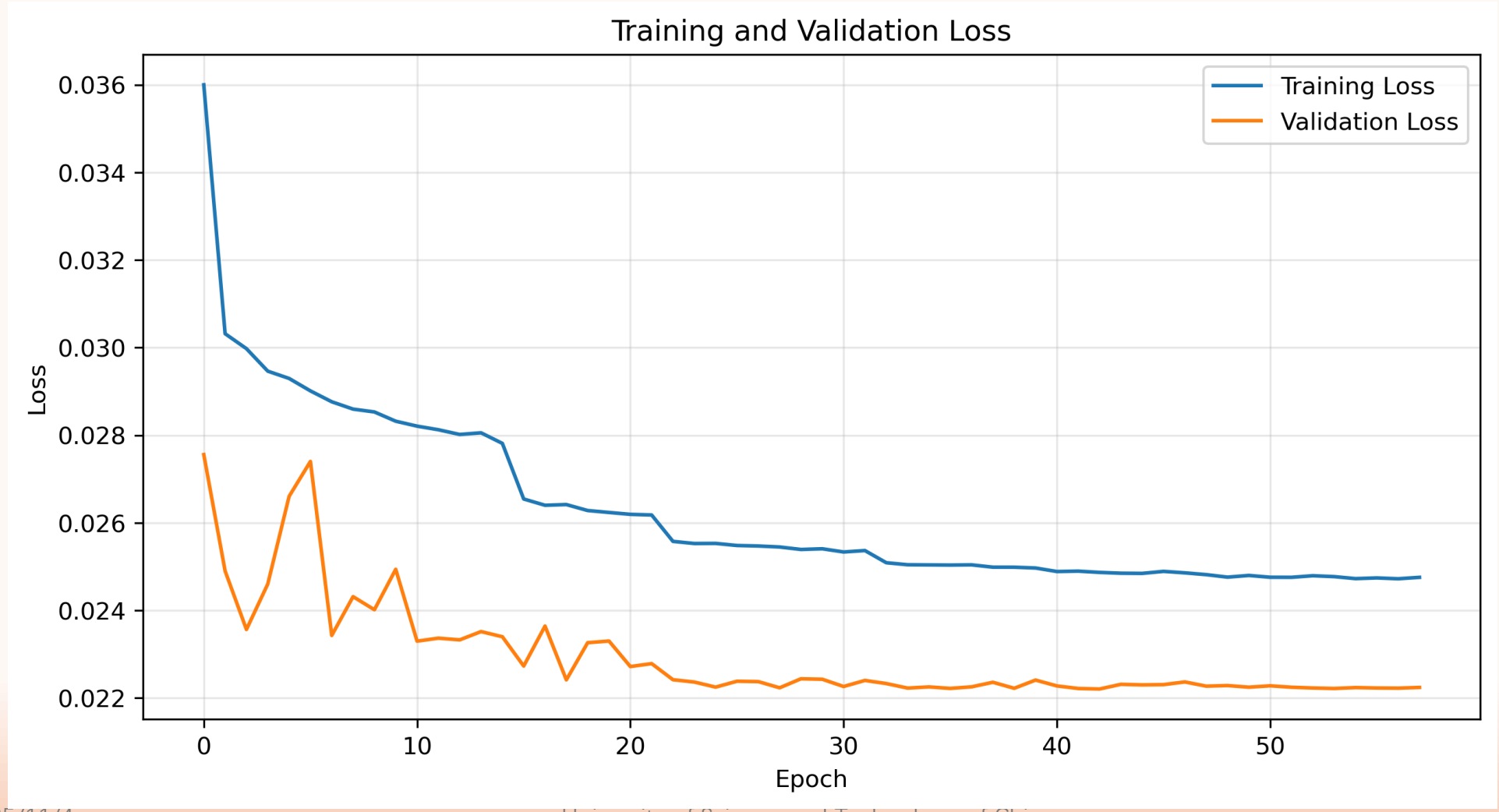
Backup CNN loss curve



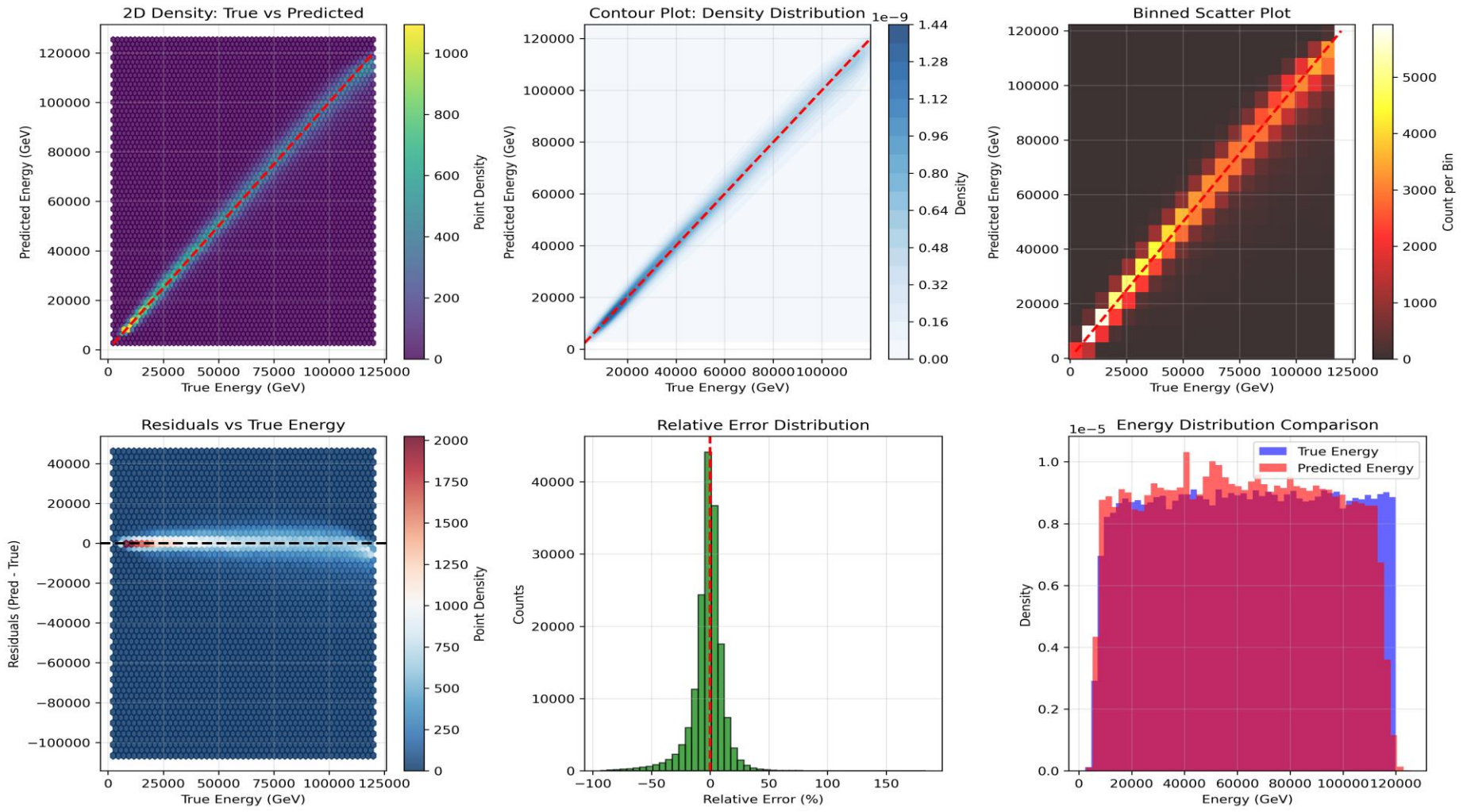
Backup CNN tests



Backup Hybrid loss curve



Backup Hybrid tests



Backup PFA

- The Particle Flow Algorithm (PFA) reconstructs a list of low-level particles (called PFA particles) and associates every detector hit with these particles. For each physics event, all physics objects that are input to physics analyses are then identified or reconstructed from this unique list of PFA particles.