DREAM FOR FUTURE

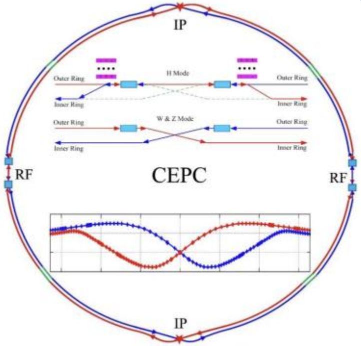
Simulation and Machine Learning for the Dual-Readout Calorimeter

YunJae Lee (Univ. of Seoul)
on behalf of the Dual-Readout Calorimeter team

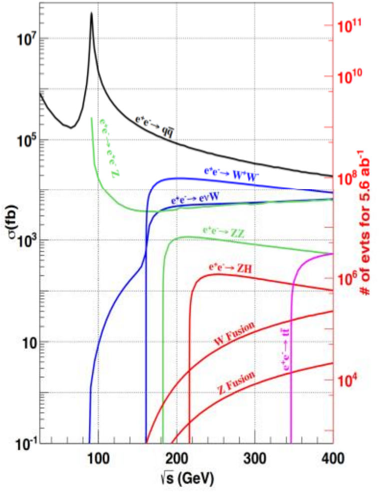
Introduction

- CEPC produces large amount of Higgs and other events with high signal to background ratio.
  - Bosons (W, Z), quarks ($t\bar{t}$, b, c) and leptons (τ)
- Physics objects of those events need high-precision detectors and particle identification capability.
- Dual-Readout calorimeter has excellent energy resolution.
  - It directly measures hadronic and electromagnetic components.
- Machine learning applications at HEP have demonstrated improvements in data processing.
  - Such as particle identification and signal/background discrimination.
- Dual-readout calorimeter with machine learning will maximize the potential of CEPC project.

CEPC baseline



Cross sections for e^+e^- collision

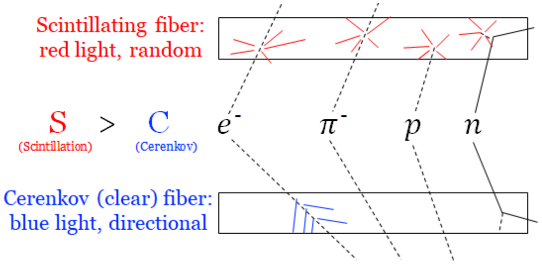


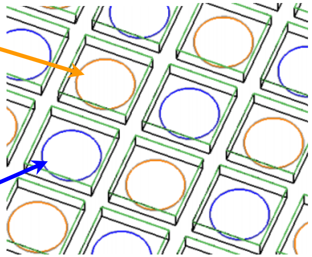
Dual-Readout Calorimeter

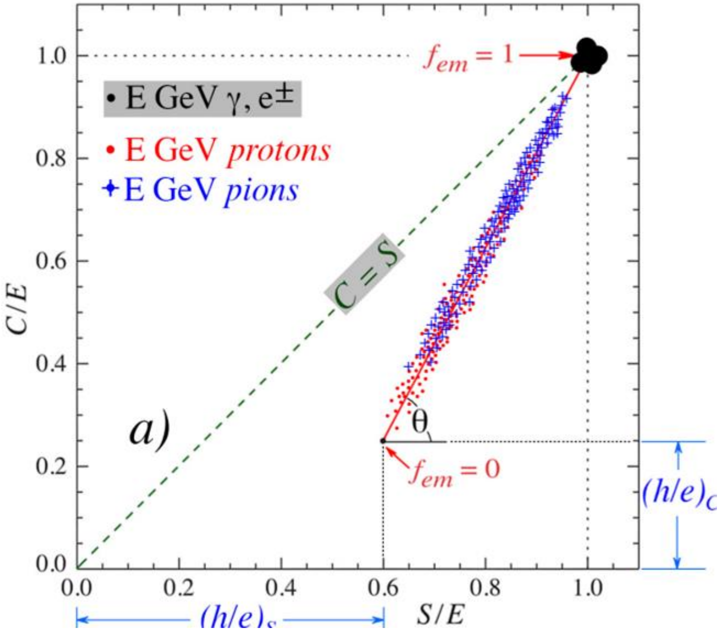
- There are two different, Scintillation and Cerenkov, fibers components.
    - EM shower fraction (f_{em}) is directly measured by scintillation and Cerenkov response.
    - Using f_{em} we measure energy wit igh precision

$$f_{\mathrm{em}} = \frac{(h/e)_C - (C/S)(h/e)_S}{(C/S)[1-(h/e)_S] - [1-(h/e)_C]} \qquad \cot\theta = \frac{1-(h/e)_S}{1-(h/e)_C} = \chi$$

$$\begin{aligned} S &= E\left[f_{\mathrm{em}} + \frac{1}{(e/h)_S}(1-f_{\mathrm{em}})\right] \\ C &= E\left[f_{\mathrm{em}} + \frac{1}{(e/h)_C}(1-f_{\mathrm{em}})\right] \end{aligned} \qquad E = \frac{S-\chi C}{1-\chi}$$



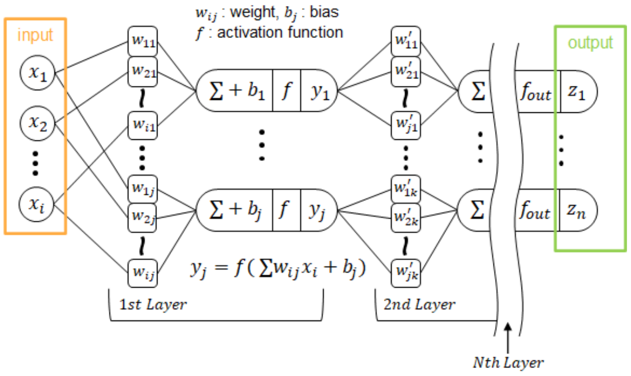


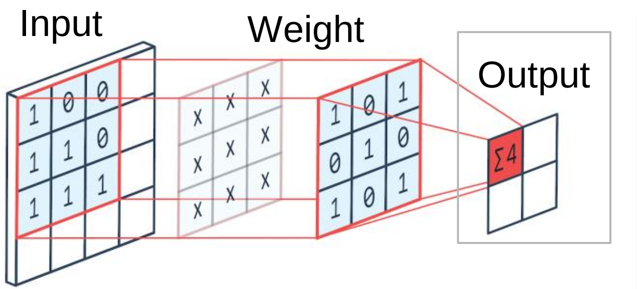


Why Deep Learning?

- Deep learning is one of ML methods, which are based on neural networks
  - Multi-layer application of weights and activation functions provides estimates of output values from input conditions.
  - There are variations, such as convolutional neural networks(CNN) powerful for images processing.
- Deep learning can be used in very general cases.
  - Input can be any dataset. Output can be setup for regression or classification or both.
  - DL is already being applied in HEP researches.
- All infrastructures of CEPC should be ML friendly!
  - Maximizing physics performance can only be achieved by using machine learning methods to solve a wide range of problems.
  - Fast simulation can also be improved by DL methods.

Neural Networks





Convolutional Neural Networks

Goals of Deep Learning Application

- **Our goals**

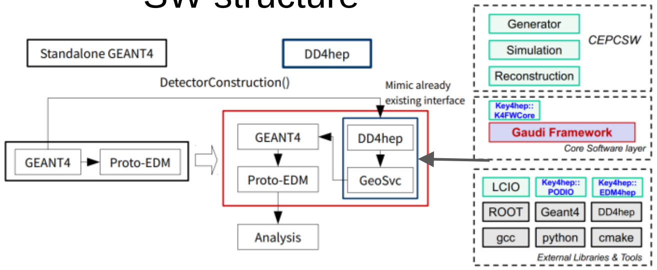
Application areas	Technical goals	Works ongoing
Jet reconstruction	Mass and energy reconstruction.	Applying regression for quark/gluon jet discriminant variables.
Particle identification	Shower object identification, lepton and quark flavor tagging.	Investigating DL methods for τ, quark/gluon jet discrimination, e^-, γ, π^+ shower identification.
Physics analysis	Higgs decay, Z boson decay, QCD measurements, flavor physics.	Validating MC sample production.
Fast simulation(GAN)	GAN based shower generator.	Applying GAN for electron shower image generation.

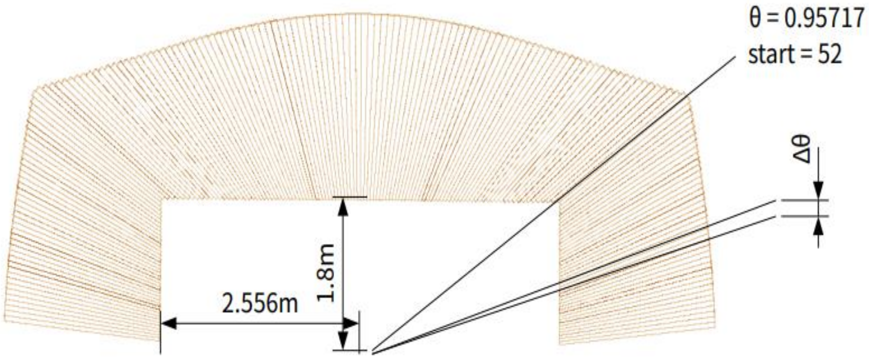
- DL studies and DR calorimeter developments should evolve in parallel.
  - Optimization to ML approaches, including DL, will be considered for maximal performance.
  - For DL, raw energy deposit by fibers is being used.

Software Setup for Simulation

- Detector Simulation
  - For calorimeter developments and pre-analysis of samples.
  - Current setup is DD4HEP with GEANT4 full simulation.
  - Detector description has been implemented in DD4HEP in the process of migration to Key4HEP
  - Developing fast simulation of optical photons in fibers
- Deep learning
  - Keras (Tensorflow) and PyTorch implemented for deep learning.

SW structure

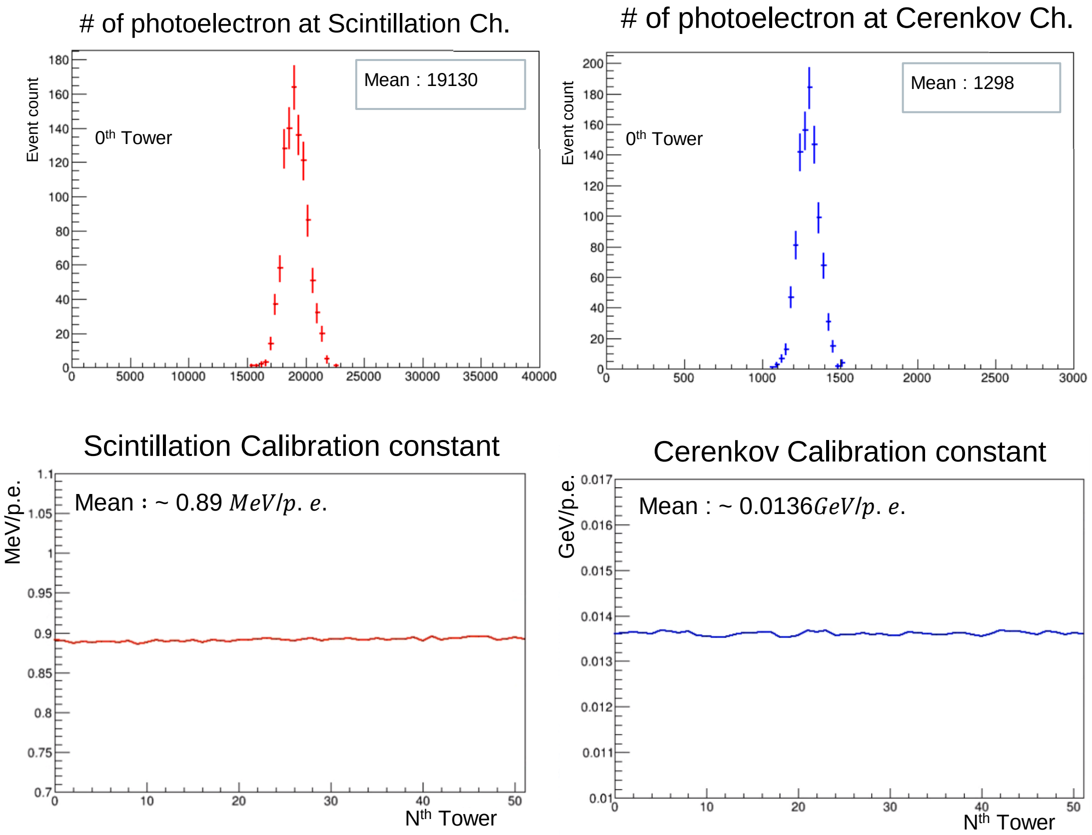




Detector geometry for ongoing studies

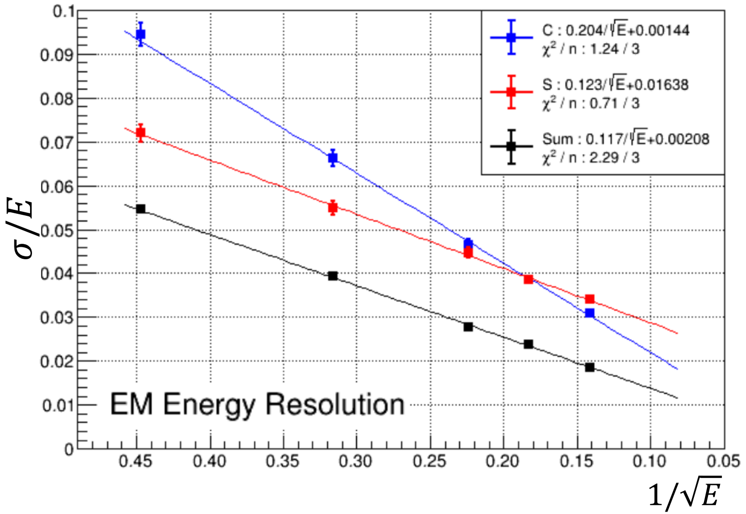
Calibration with electron beam

- Calibration is needed for correct energy measurement.
  - The calibration constants obtained from energy deposit and photoelectron counts.
- Calibration for barrel region has done with 20 GeV electron beam.
- End-cap region is ongoing.

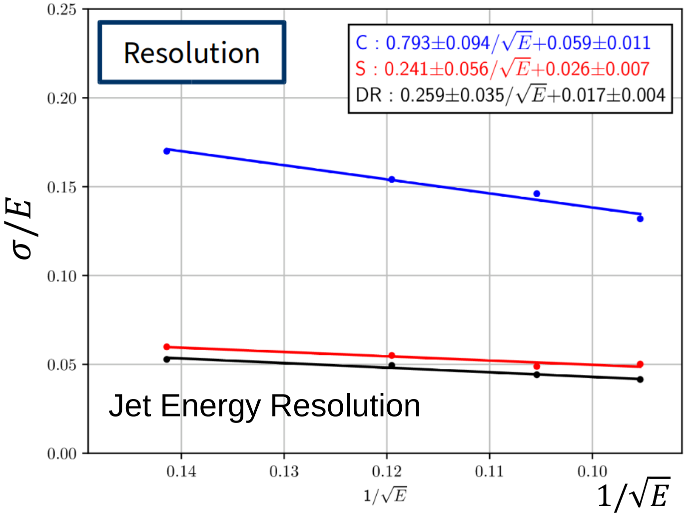


Energy resolution Study

- EM energy resolution :$\sim 11\%/\sqrt{E}$
  - Measured in 5 energy points (5, 10, 20, 30, 50 GeV)
  - Combined channel gives better resolution than single channel.

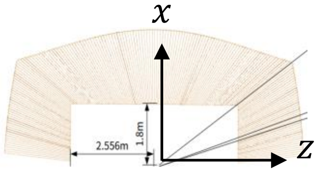


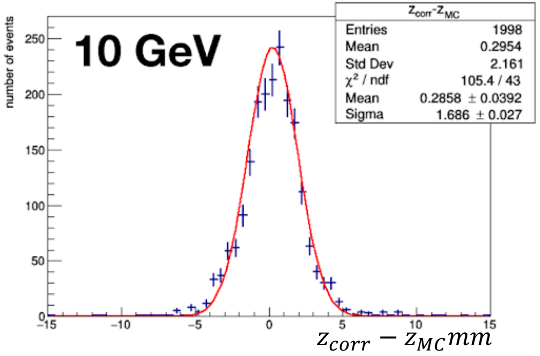
- Jet energy resolution : $\sim 26\%/\sqrt{E}$
  - Measured in 4 energy points (50, 70, 90, 120 GeV)
  - Anti-kt algorithm ($dR = 0.8$) is used for jet clustering.

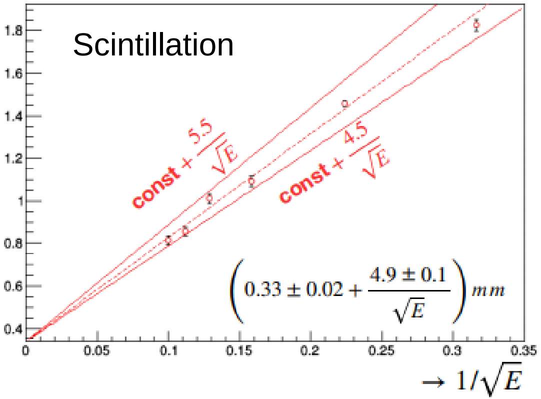


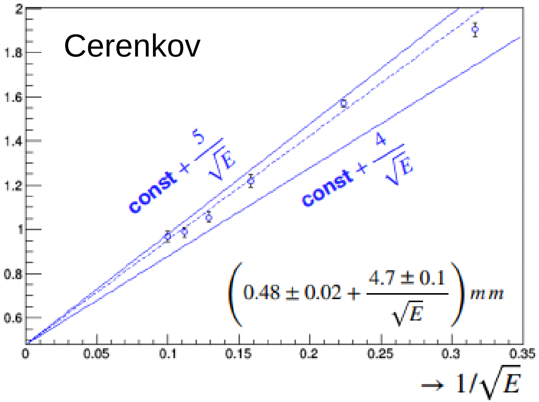
Position resolution Study

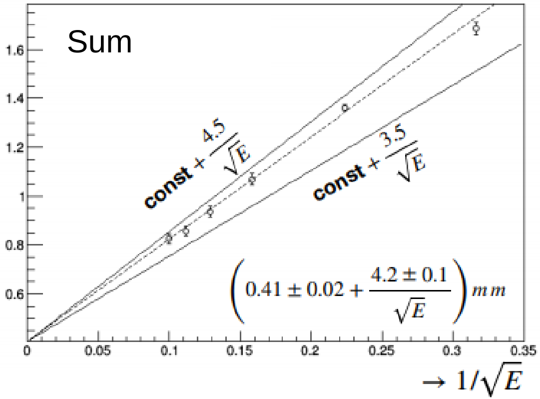
- Position resolution along z direction
  - Tested with e^- beam with energy 10, 20, 40, 60, 80 and 100 GeV in $0 \times 40\ mm$ beam spot which covers from the center of one tower to that of a neighbor tower.
  - Position reconstructed by center of gravity of the energies and compared with generated position.
  - Position resolution : $4.2/\sqrt{E} + 0.41\ mm$.







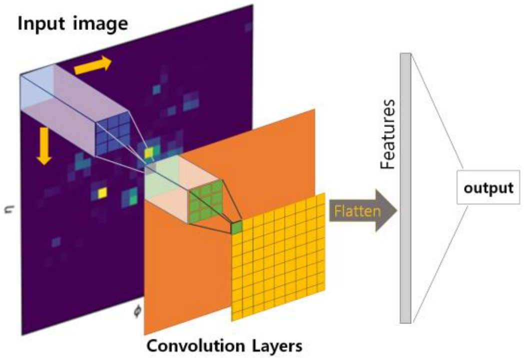




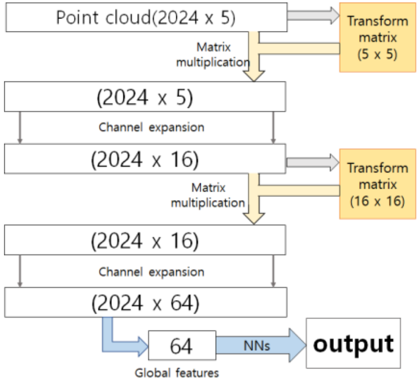
Inputs to Deep Learning

- Data preparation is most critical part while implementing DL.
- Two types of inputs data are considered for variety of models.
  - **Image** : Energy and number of photon image from fibers in the (η, ϕ) area. With CNN implementation.
  - **Point cloud** : list of energy deposit positions (η, ϕ, depth) with scintillation and Cerenkov energies. With PointNet implementation.

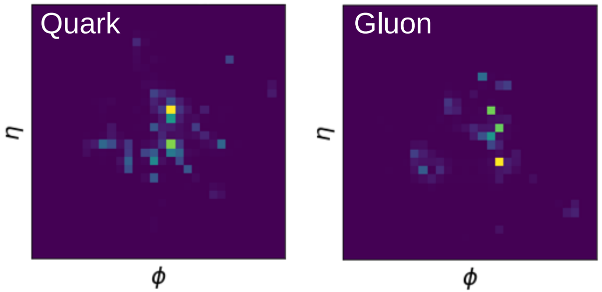
CNN model



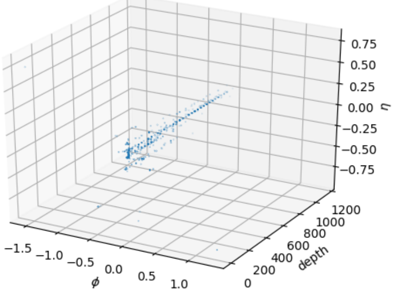
PointNet model



Pixelated energy image

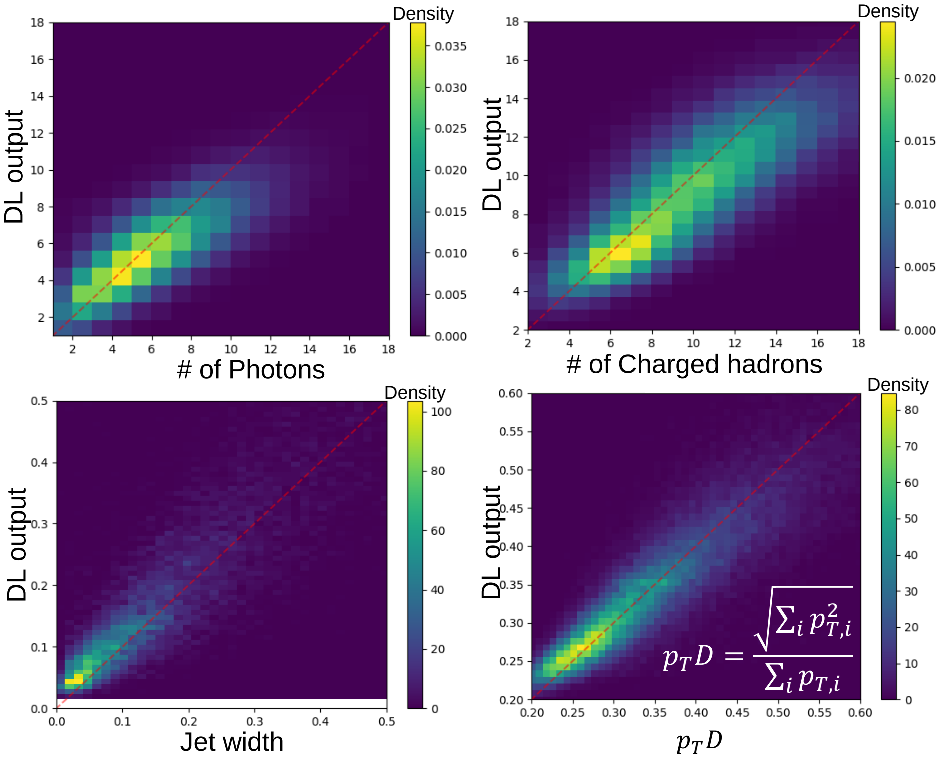


Point cloud



DL Reco. : Jet Variable Regression

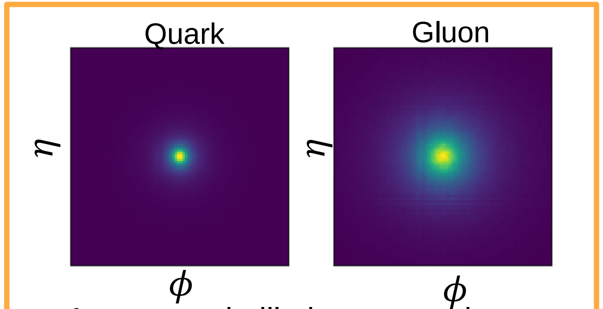
- Regression for jet discriminant variables of quark, gluon jet.
  - Particle multiplicity, jet width, $p_T D$ (jet fragments variable) are used in jet discrimination.
  - Trained with images of 50 GeV u quark and gluon jets.



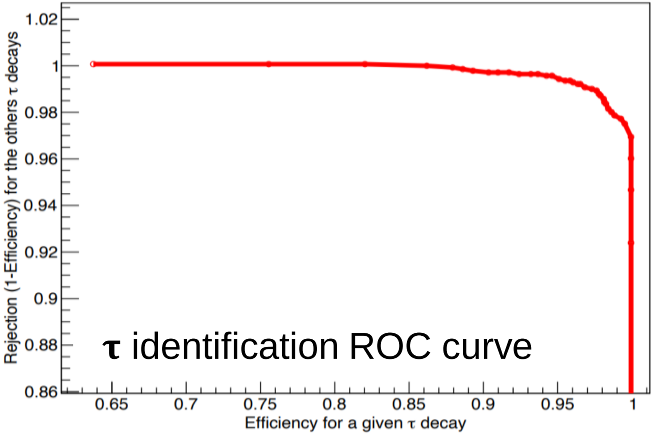
- Jet mass and energy reconstructions are in pipeline.
  - This will improve jet flavor study.

Jet, τ Identification

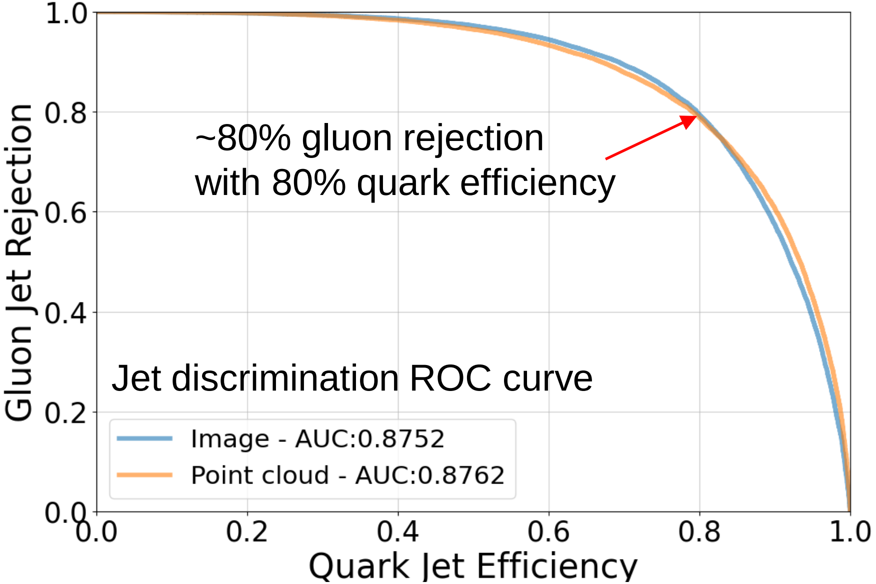
- Classification of quark and gluon jets.
  - Performed for 50 GeV u quark and gluon jets with image and point cloud format.
  - Model structures need to be developed.
- τ identification with deep neural networks also shows promising performance
- Jet flavor tagging is in pipeline.



Average scintillation energy images



τ identification ROC curve

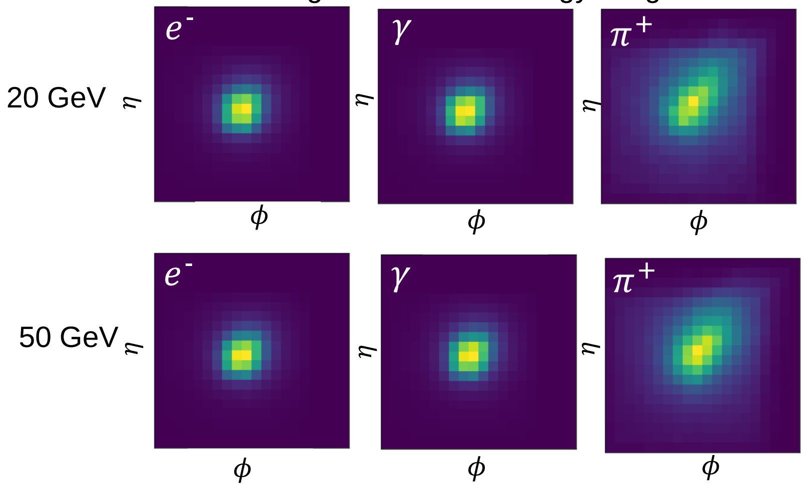


Jet discrimination ROC curve

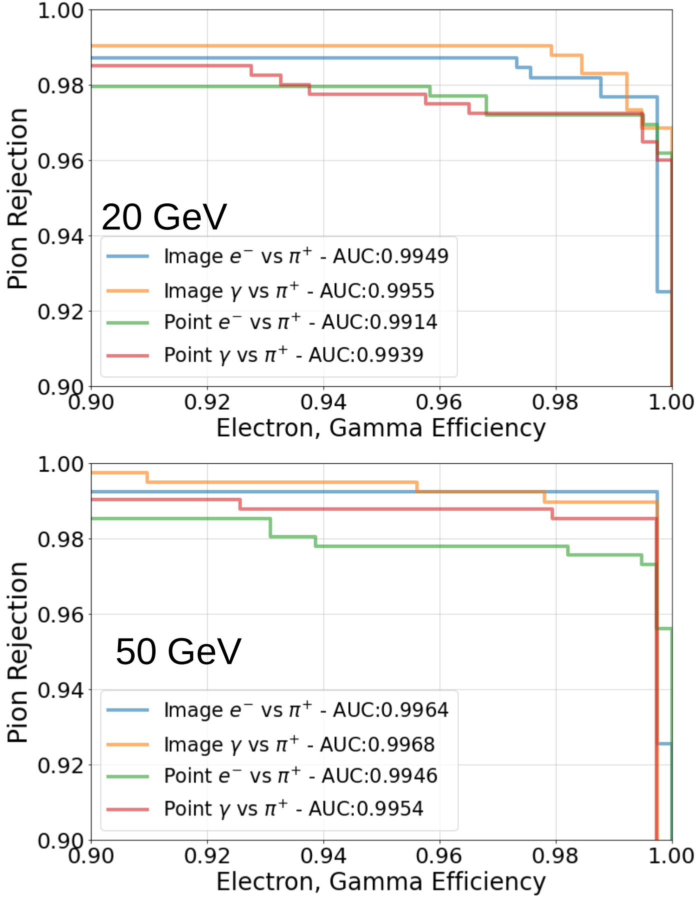
Shower identification

- Discrimination of pion with respect to electron and gamma showers.
  - Performed for 20, 50 GeV showers with image and point cloud format.
  - Applying more statistics is in pipeline.

Average scintillation energy images



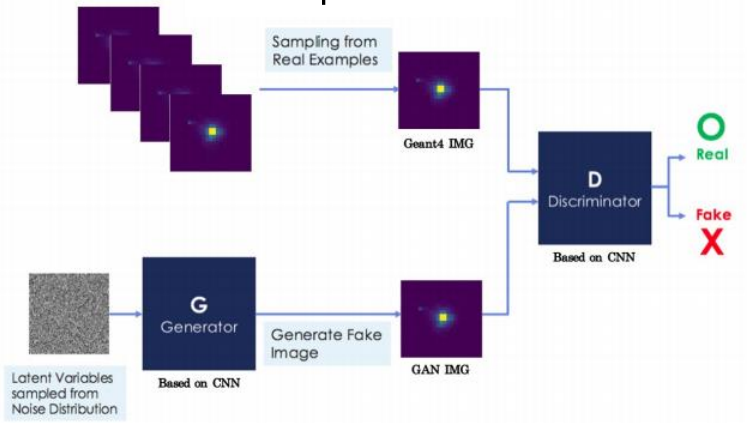
ROC curves



Fastsim with GAN

- Generative Adversarial Networks (GAN) consist of a Generator and a Discriminator network.
  - Generator create fake images trying to not be discriminated as fake.
- GAN can be applied for shower images.
  - Previous study with CMS ECAL showed electron shower can be generated using GAN.
  - It is much faster than fast simulation.
- We can produce larger datasets with GAN shower fastsim.
  - Massive computing facilities are needed for GEANT4 simulation

Concept of GAN



Maximum availability

Institute	Computing facilities
KISTI	400 cores
KNU	300 cores
UOS	300 cores + 8 GPU for DL
SNU	200-300 cores

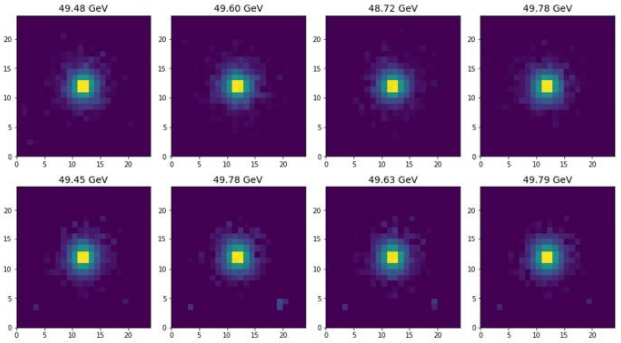
Fake shower generator

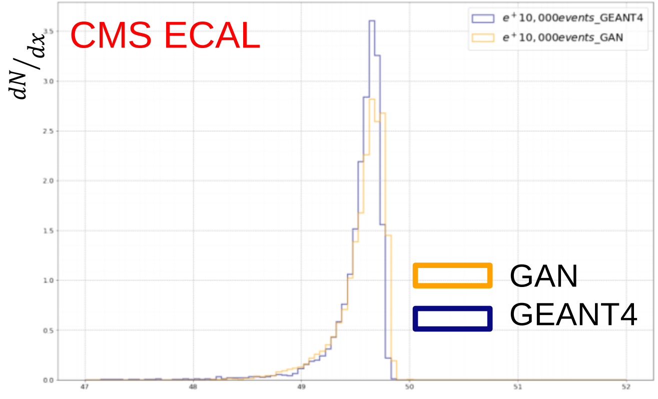
- Previous GAN study about CMS ECAL showed good performance.
  - Trained with 50 GeV electron shower images.
  - GAN total-energy distribution matched to GEANT4 predictions.
- Reproducing GAN result with DR calorimeter GEANT4 simulation.
  - Working on matching GAN images with GEANT4 images of electron showers.
  - Training using larger statistics from GEANT4 fast simulation in the pipeline.

CMS ECAL electron images comparison

GEANT4

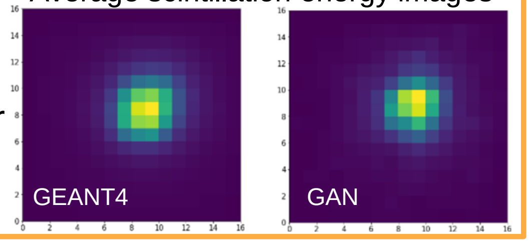
GAN





Undergoing study with DR calorimeter

Average scintillation energy images



Snowmass21

- Snowmass 2021 is a scientific study for future in HEP.
  - https://www.snowmass21.org
- International dual-readout calorimeter R&D team submitted a single letter of interest (LoI) for overall R&D plan in our team.
  - https://www.snowmass21.org/docs/files/summaries/IF/SNOWMASS21-IF6-008.pdf
  - Include world-wide community of the dual-readout calorimeter R&D: Asia, Europe, US.
- Additional seven LoIs related to the dual-readout calorimeter R&D project have been submitted too!
  - All topics of seven LoIs are potentially based on ML application.
  - ML study can be boosted and extended with the Snowmass 2021 campaign, stay tuned!
- Various MC productions such as Multi-jets, Higgs and τ events are underway.

List of (ML-based) LoIs

Topic 1: Feasibility study of combining a MIP Timing Detector with the Dual-Readout Calorimeter at future e^+e^- colliders. (link)

- Collaborators : David Stuart (UCSB), C.S. Moon (KNU), J.H. Yoo (Korea Univ.)

Topic 2: Heavy flavor tagging using machine learning technique with silicon vertex detector and Dual-Readout Calorimeter at future e^+e^- colliders. (link)

- Collaborators: J. Huang (BNL), Q. Hu (LLNL), S.H. Lim (PNU)

Topic 3: τ reconstruction and identification using machine learning technique with Dual-Readout Calorimeter at future e^+e^- colliders. (link)

- Collaborators : M. Murray (U. of Kansas), Y.S. Kim (Sejong Univ.), Y.J. Kwon (Yonsei Univ.)

Topic 4: Sensitivity study of $HZ \to Z\gamma$ with Dual-Readout Calorimeter at future e^+e^- colliders. (link)

- Collaborators : Y. Maravin (Kansas State Univ.), K.W. Nam (Kansas State Univ.)

Topic 5: Multi-object identification with Dual-Readout Calorimeter at future e^+e^- colliders. (link)

- Collaborators : P. Chang (UCSD)

Topic 6: Dual-Readout Calorimeter for the future Electron-Ion Collider. (link)

- Collaborators : S.H. Lim (PNU), H.S. Jo (KNU), Y.S. Kim (Sejong Univ.)

Topic 7: Fast optical photon transport at GEANT4 with Dual-Readout Calorimeter at future e^+e^- colliders. (link)

Summary

The dual-readout calorimeter is simulated with GEANT4 simulation with DD4HEP

Simulation performance studies are performed with GEANT4 full simulation

- EM energy resolution : $\sim 11\%/\sqrt{E}$, Jet energy resolution : $\sim 26\%/\sqrt{E}$.
- Position resolution : $\sim 4.2/\sqrt{E} + 0.41\ mm$

Deep learning application shows promising performance for particle identification.

- Jet discrimination : ~80% gluon rejection with 80% quark efficiency

Mass and energy reconstruction for jets is in pipeline.

GAN shower fastsim is ongoing with electron shower images.

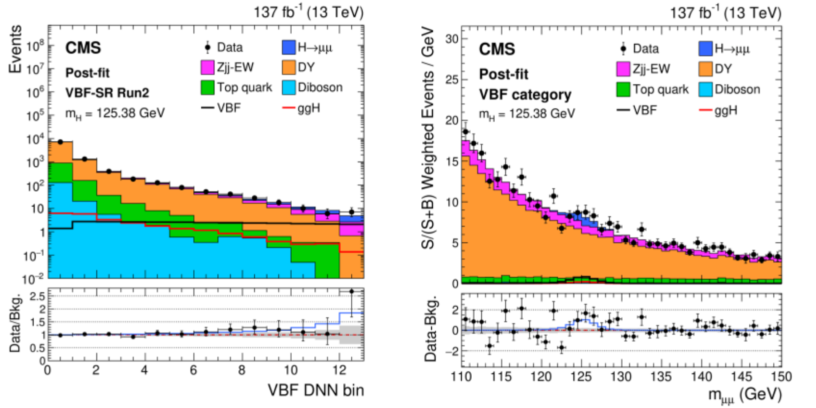
We are progressing various projects for dual-readout calorimeter with ML application

Backups

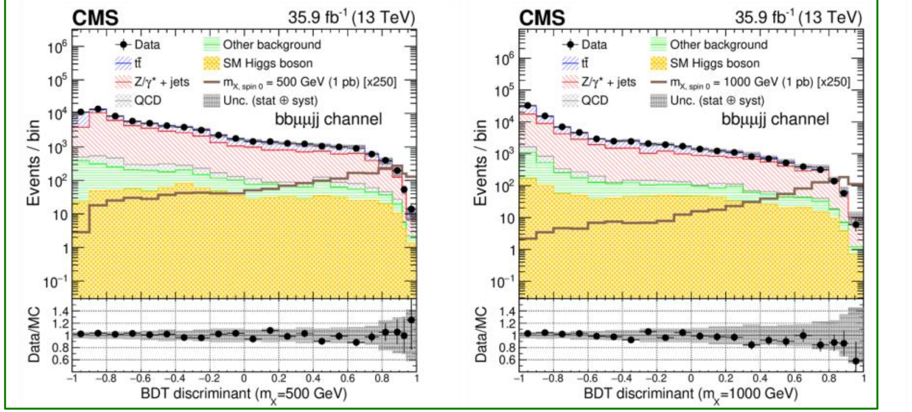
ML Application for Physics Analyses

- Machine learning is widely used for LHC physics analysis with sensible improvements.
    - Most LHC physics analyses apply the ML separately to between object tagging and event level selection
- Our ultimate goal is to apply ML techniques to the CEPC project and maximize the performance.
    - A more comprehensive event level tagging could be performed using raw input from DR calorimeter for maximal performance.

VBF analysis with $H \rightarrow \mu\mu$ decay (HIG-19-006)



Event selection for $HH \rightarrow bbZZ$ decay(HIG-18-013)



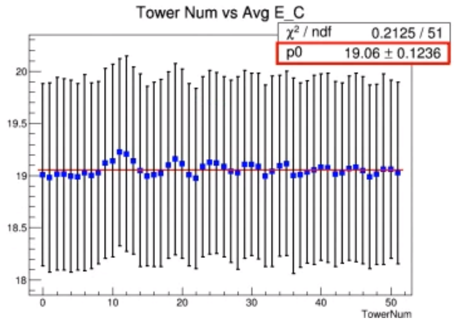
Calibration with 20GeV electron beam

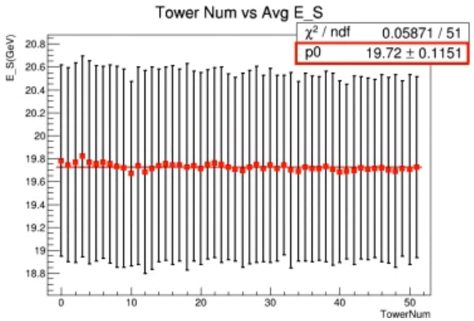
$$Equalization\ constant = \frac{Energy\ deposit}{\#\ of\ photoelectron} = \begin{cases} Scintillation\ channel & = 0.87\ MeV/p.e \\ Cerenkov\ channel & = 12.9\ MeV/p.e. \end{cases}$$

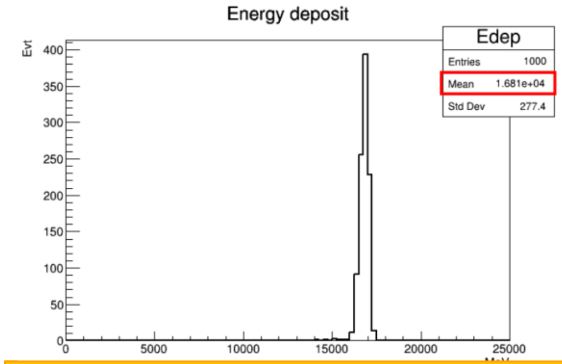
$$\text{Scale factor} = \begin{cases} Scintillation\ channel & = 0.99 \\ Cerenkov\ \ channel & = 0.95 \end{cases}$$

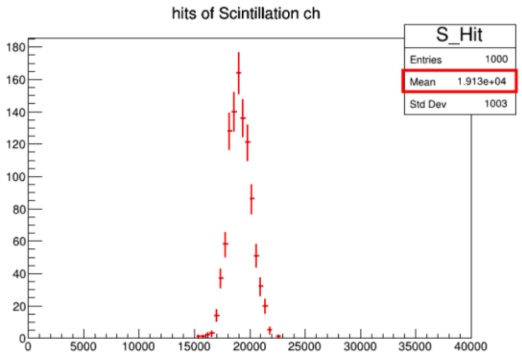
$$Calibration\ constant = eq.constant/Scale\ factor$$

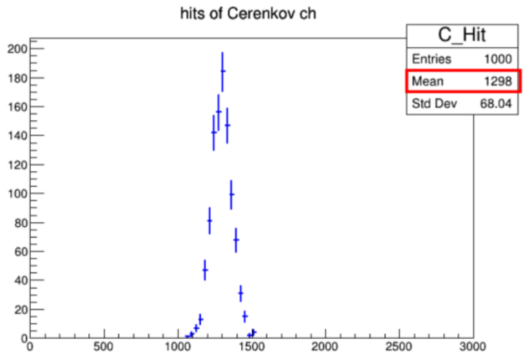
Scale factor





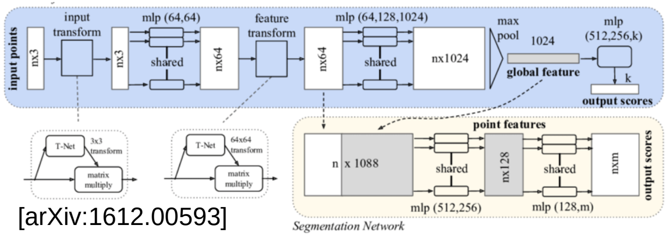






Backups

Model structure



[arXiv:1612.00593]

Image model
(90,90) jet image or (20,20) shower image

64 channels (3,3) kernel Convolution layer
(2,2) Max pooling
Batch normalization
32 channels (3,3) kernel Convolution layer
(2,2) Max pooling
Batch normalization
128 channels (1,1) kernel Convolution layer
Average whole image(128 outputs from channels)
Batch normalization
64 feature Dense layer

Regression or classification

- ReLU activation function, Adam optimizer

Point cloud model
2048*5 channel

Matrix multiply (2048,5) input and transform matrix
16 channels 1D convolution((2048,5) to (2048,16))
Batch normalization
Matrix multiply with transform matrix
64 channels 1D convolution((2048,16) to (2048,64))
Batchnormalzation
2 channels 1D convolution((2048,64) to (2048,2))
Batchnormalzation
Flatten

classification

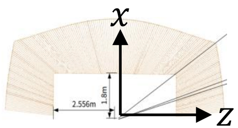
- ReLU activation function, Adam optimizer

Transform matrix{
64 channels 1D convolution((2048,5) to (2048,64))
Batchnormalization
Max pooling over points ((2048,64) to 64)
64 features Dense layer
Batch normalization
5*5 features Dense layer
Batch normalization
Reshape (5,5) for transform matrix}

Transform matrix{
64 channels 1D convolution((2048,16) to (2048,64))
Batchnormalization
Max pooling over points ((2048,64) to 64)
64 features Dense layer
Batch normalization
16*16 features Dense layer
Batch normalization
Reshape (16,16) for transform matrix}

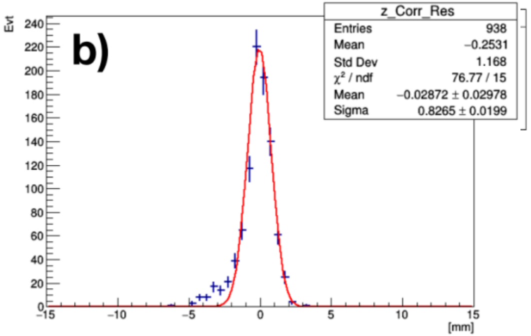
Position resolution Study with $(\theta, \phi) = (0, 1°)$

- Position resolution for z
  - Tested by e^- beam with energy 10, 20, 40, 60, 80 and 100 GeV in $z(0 - 35mm)$.
  - Fastsim for optical photon was implemented
  - z position resolution : $\sim \frac{4.1}{\sqrt{E}} mm + const.$

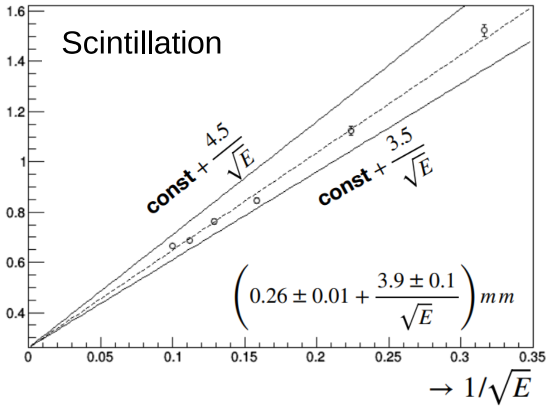


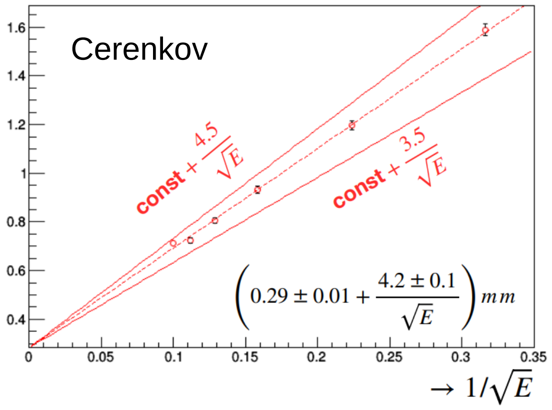
$$z_{reco} = \frac{\sum_i E_i \times z_i}{\sum_i E_i}$$

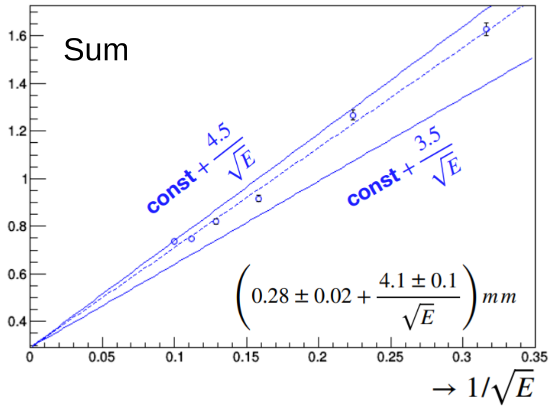
z_i : position of i^{th} SiPM
E_i : calibrated signal count at i^{th} SiP



$z_{corr} - z_{MC} mm$







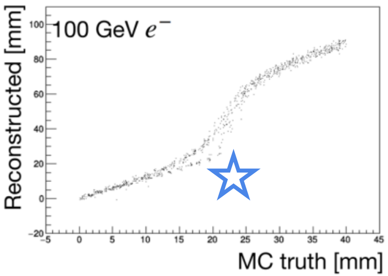
Position resolution

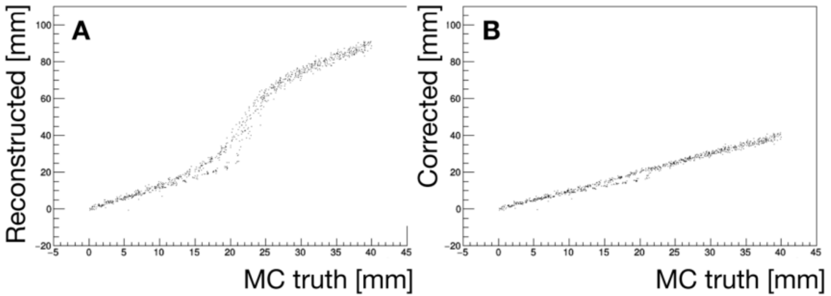
- Position reconstruction
  - Center-of-gravity method
  - Uses the center of gravity of the energies as reconstructed position

$$x_{reco} = \frac{\sum_i E_i \times x_i}{\sum_i E_i}$$

x_i : position of i^{th} SiPM

E_i : signal count at i^{th} SiPM multiplied by calibration constant

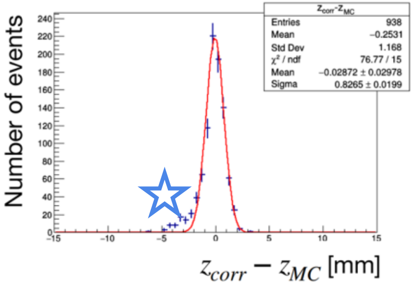


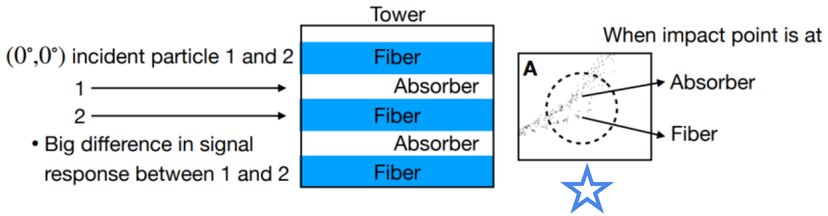


- Correction equation

$$x_{corr} = p_0 + p_1 x + p_2(x_{reco} - p_3)^2 \arctan(p_4(x_{reco} - p_5))$$

- Resolution
  - The σ from the Gaussian fit to $x_{corr} - x_{gen}$ is regarded as the resolution





- Big difference in signal response between 1 and 2