Machine Learning Approaches for the Energy Reconstructions in the CMS HCAL

Hui Wang (王徽)
Rutgers University

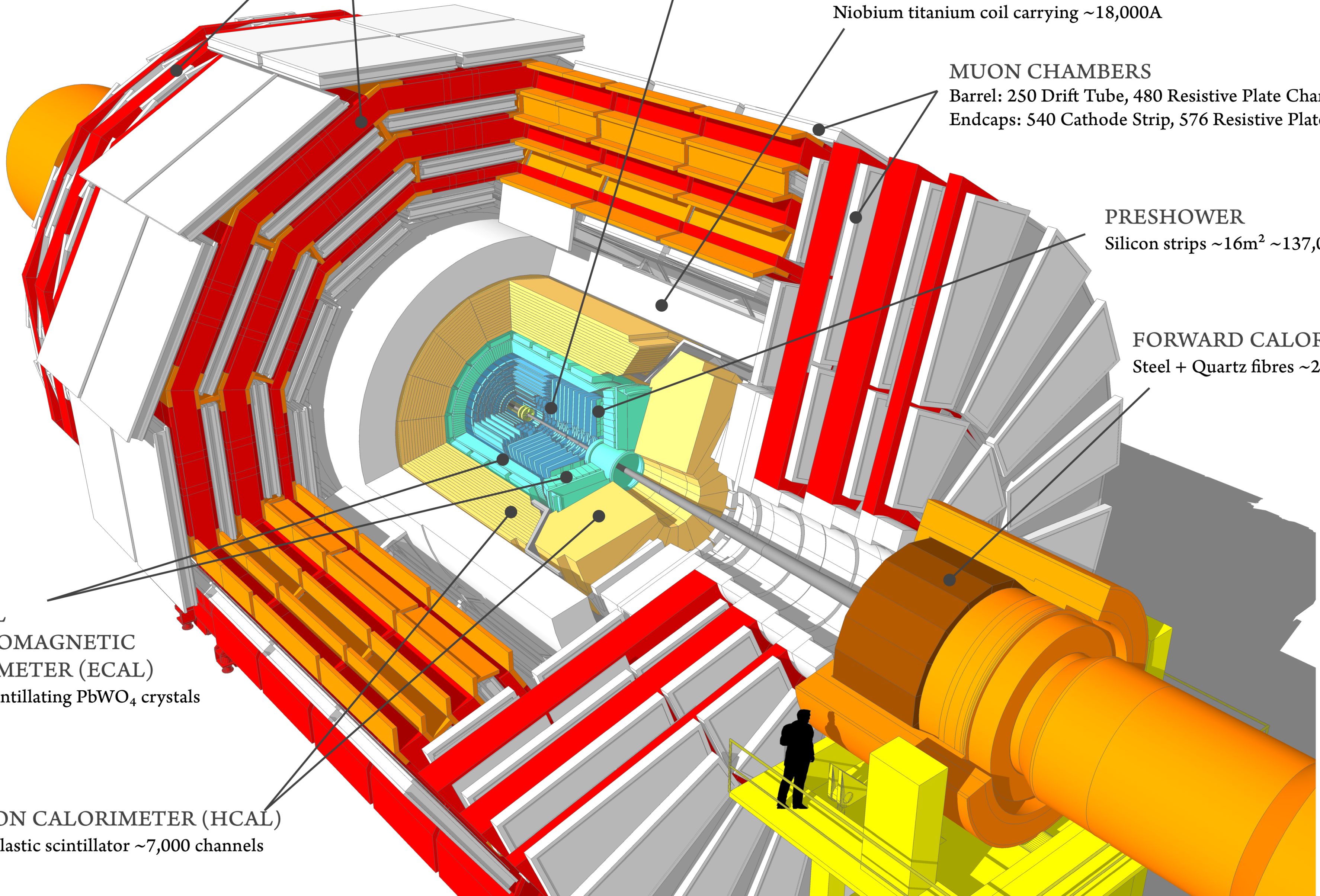
IHEP Seminar - Nov 30, 2022

Outline

- The CMS detector
- Introduction to the HCAL energy reconstruction
- Current analytical methods and their disadvantages
- Common ML architectures: DNN, CNN and RNN
- Their applications on HCAL energy reconstruction
- HCAL calibration
- ML performance

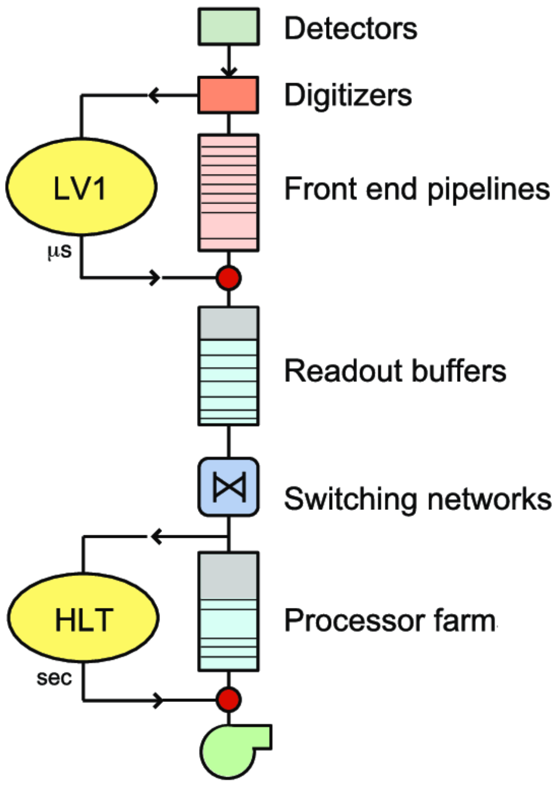
CMS DETECTOR

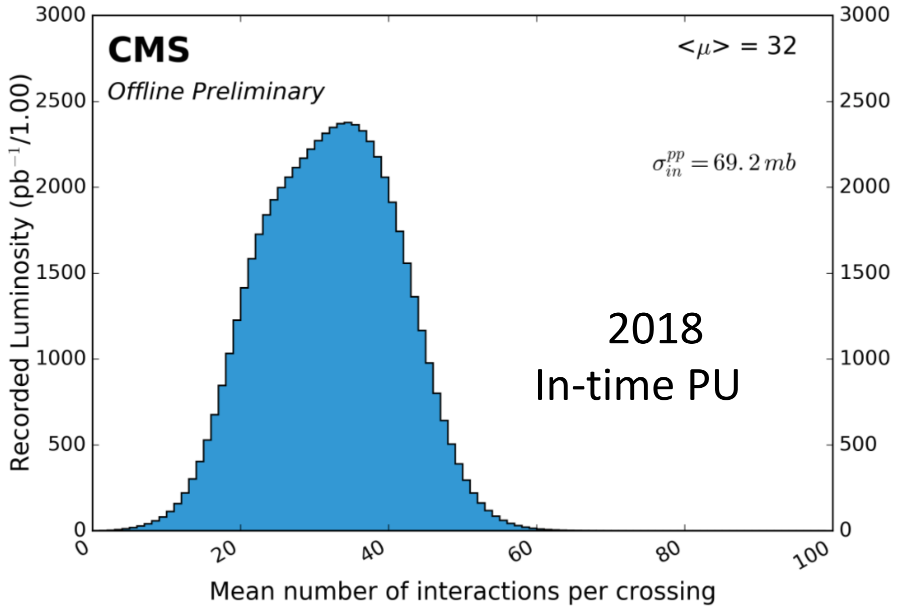
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T



Trigger System and Pileup

- Two-level trigger system
  - Reduce the event rates from 40 MHz to ~1kHz
  - While keeping most of the interesting events
- Level-1 trigger (L1T)
  - Custom electronics
  - Reduce rate to 100 kHz
- High-level trigger (HLT)
  - Processor farm
  - Rate reduce to ~1k Hz

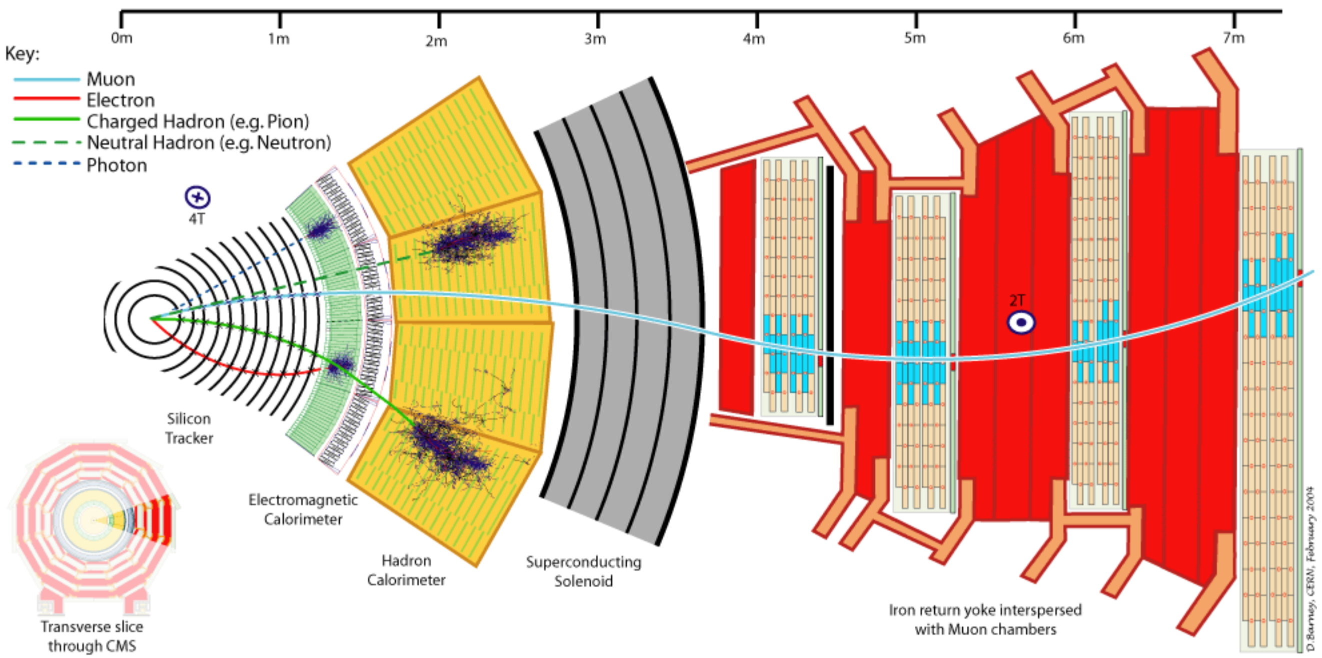




pileup (PU)

- The PP interactions in addition to the collision of interest
- In-time PU and out-of-time PU

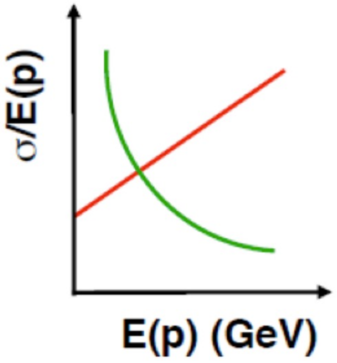
Event Reconstruction



- Particle Flow (PF) Algorithm
  - Runs on HLT and offline reconstruction
  - Synthesizes information from all sub-detectors and reconstructs particles based on their signatures
    1. Muon
    2. Electron and Photon
    3. Charged and Neutral Hadron
- Then PF particles are clustered as jets
  - Usually anti-k_T algorithm in CMS
- Last global quantities of an event
  - e.g. missing transvers momentum p_T^{miss}, aka MET usually a manifest of neutrinos, but may also from BSM :P

HCAL is Important

(Charged) particle resolution is dominated by tracker at low energy, calorimeters at high energy



Tracking

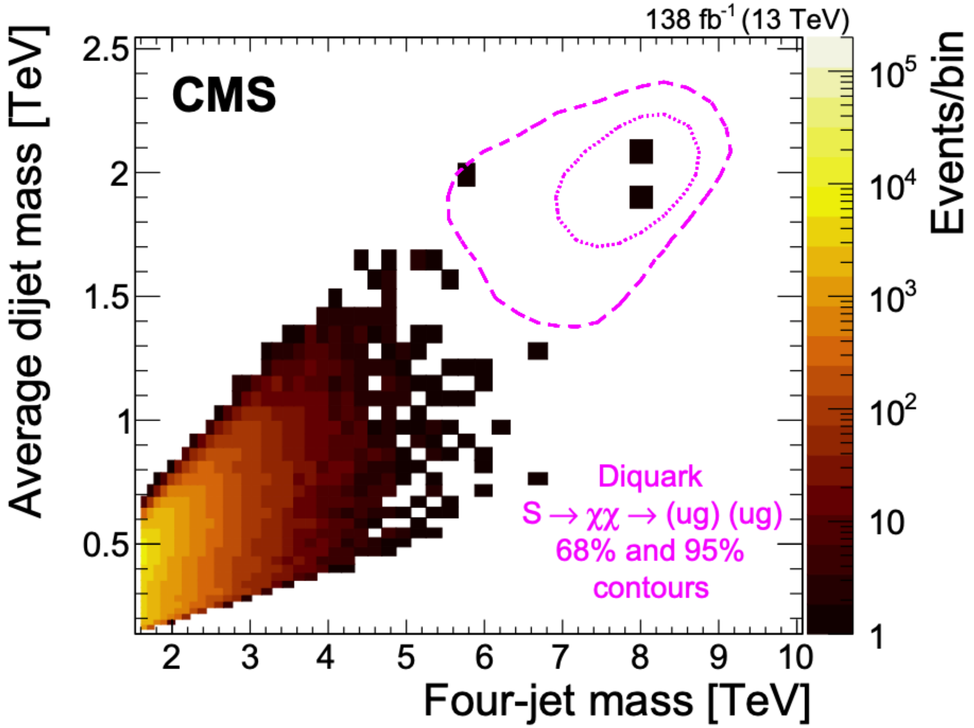
$$\frac{\sigma(p)}{p} = ap \oplus b$$

Calorimetry

$$\frac{\sigma(E)}{E} \approx \frac{a}{\sqrt{E}}$$

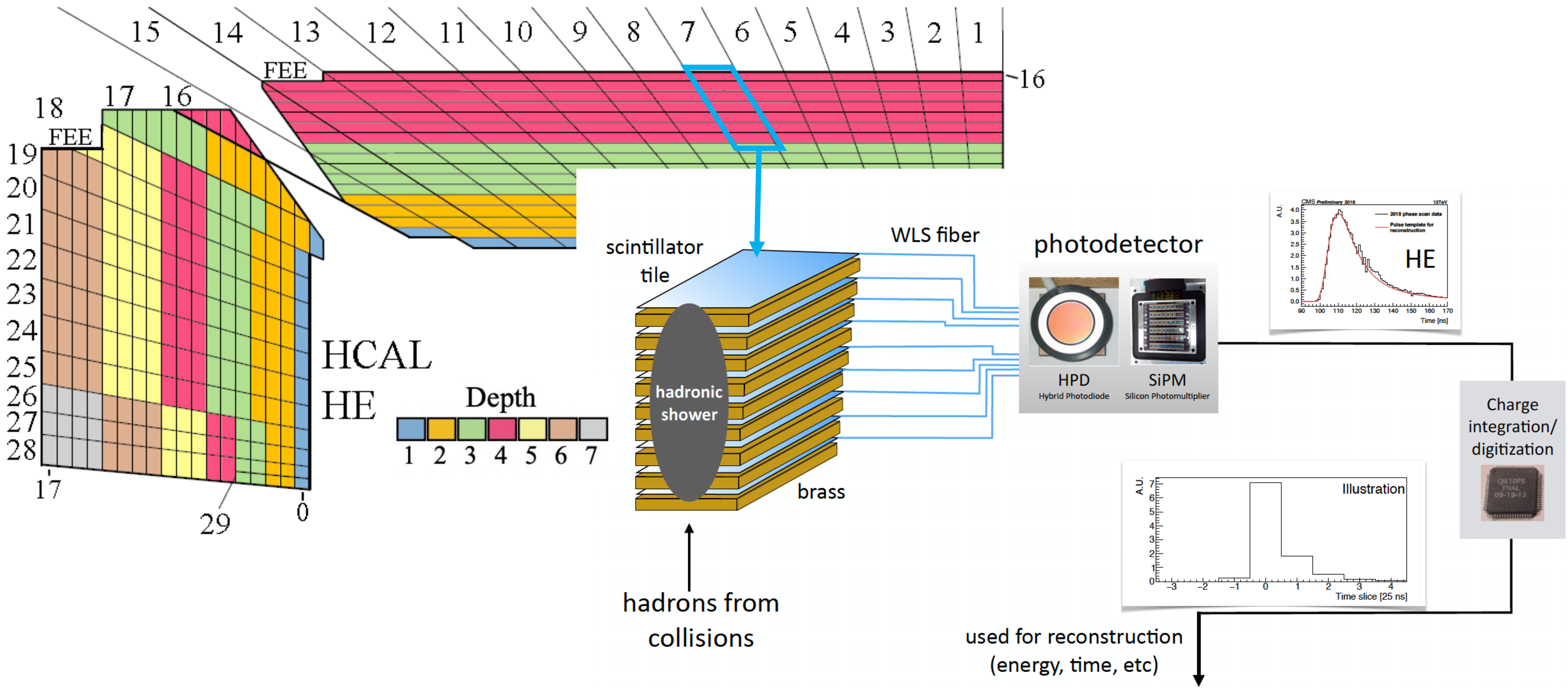
HCAL is important for:

- L1T (tracker is not involved in L1T)
- Hadrons, especially neutral hadrons
- Lepton identification (H/E) and isolation
- Physics analyses
  - e.g. di-jet resonance searches

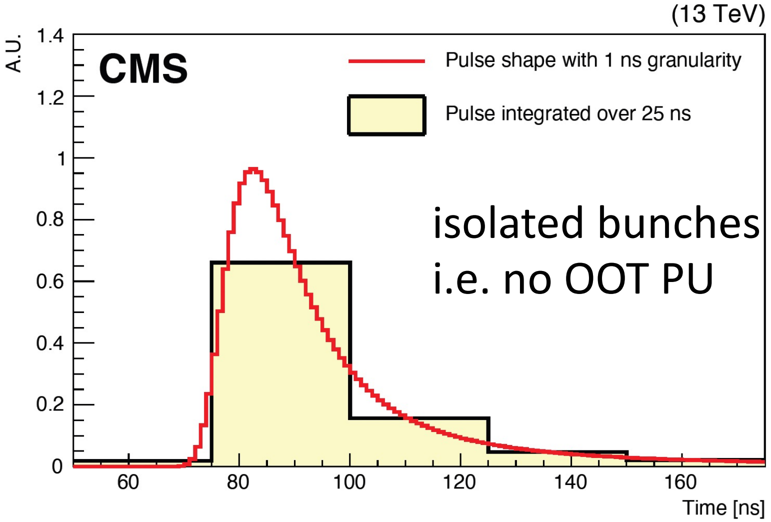


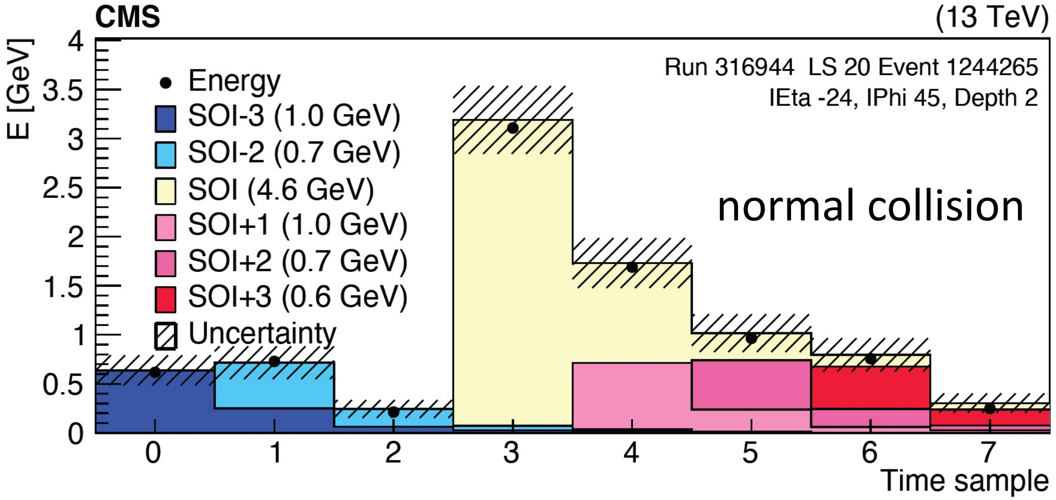
CERN news and arXiv:2206.09997

HCAL Readout Chain



HCAL Energy Reconstruction





- Pulse shape
  - Digitized charge as a function of time
  - Measured with 1 ns granularity in isolated bunches
  - 8 time samples/slices (TS) in the buffer each TS = 25 ns
  - Sample of interest (SOI): 75-100 ns ~60% total charge
  - SOI+1: ~20% total charge
- First reco algorithm: Method 0
  - Used in Run1 (50ns bunch spacing)
  - OOT PU almost negligible
  - [(SOI) + (SOI+1)] * scale factors
- Pulse fitting algorithms
  - In use since Run2 (25 ns bunch spacing)
  - 2016-2017: Method 2 (3) offline (HLT)
  - from 2018: MAHI both offline and HLT

Method 2

- Fit up to 3 pulses (SOI-1, SOI and SOI+1) to 8 TS
- Minimize χ^2 using MIGRAD algorithm in Minuit

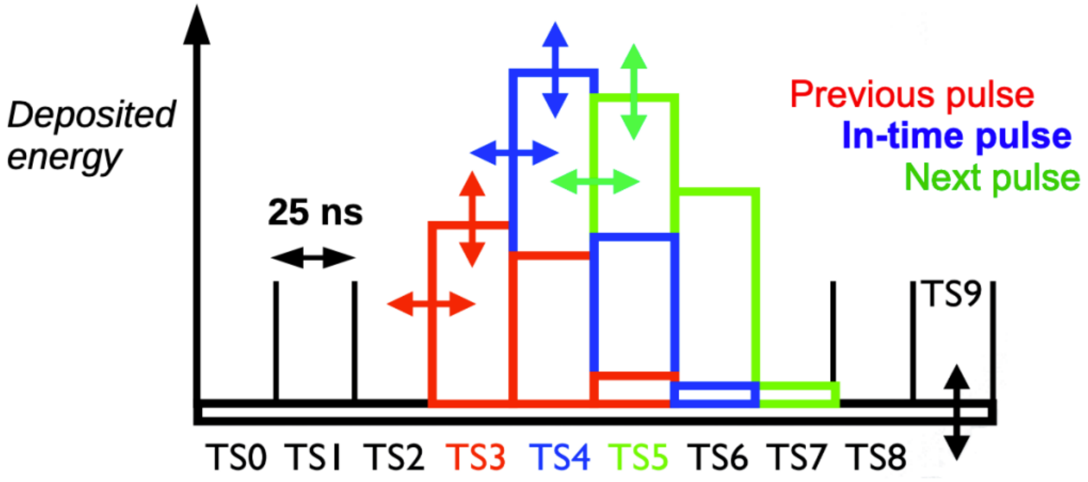
$$\chi^2 = \sum_{i=0}^{7} \frac{(TS_i - A_i)^2}{\sigma_{p,i}^2} + \sum_{j=0}^{2} \frac{(t_j - <t>)^2}{\sigma_t^2} + \frac{(\text{ped} - <\text{ped}>)^2}{\sigma_{\text{ped}}^2}.$$

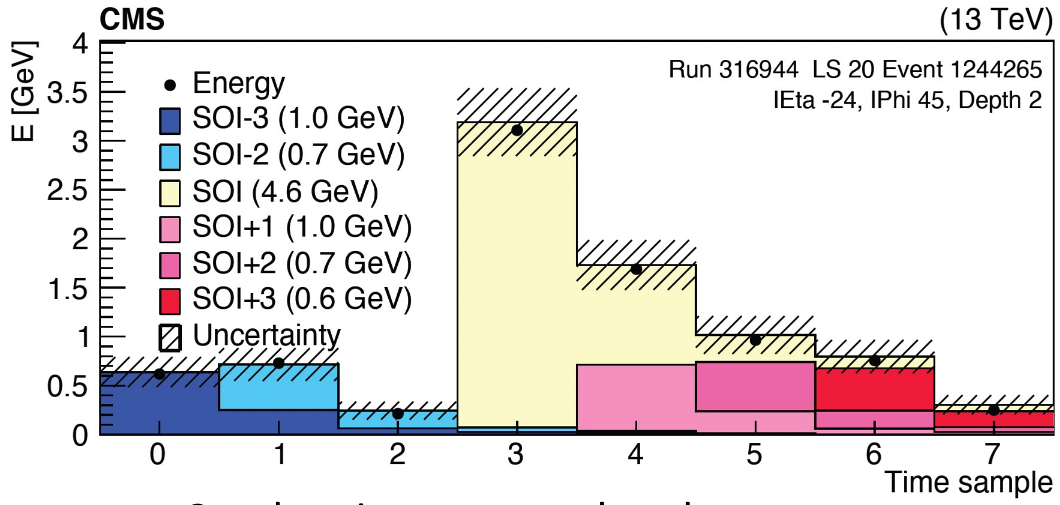
TS_i: net charge of the ith TS. A_i : pulse amplitude. t_j : pulse arrival time

Ped: pedestal noise. i.e. a floating baseline of SiPM and QIE leak current

Disadvantages:

1. Too slow. Can only run in offline reco
2. Only fit up to 3 pulses
   Bad performance at low energy
3. Sometimes fitting unstable
   Force to fit only 1 pulse when OOT PU is small (energy > 20 GeV)
   → a “kink” in the output spectrum





3 pulses is not enough at low energy

MAHI

- Minimization At HCAL, Iteratively (MAHI)
- Fit 8 pulses to 8 TS
- Matrix based minimization with Non-Negative Least Square (NNLS) algorithm

$$\chi^2 = \left(\sum A_i \vec{p_i} - \vec{q}\right)^{\mathrm{T}} \left(\Sigma_{\mathrm{d}} + \sum A_i^2 \Sigma_{p,i}\right)^{-1} \left(\sum A_i \vec{p_i} - \vec{q}\right)$$

A_i : pulse amplitude. p_i : pulse shape. q : net charges of 8TS

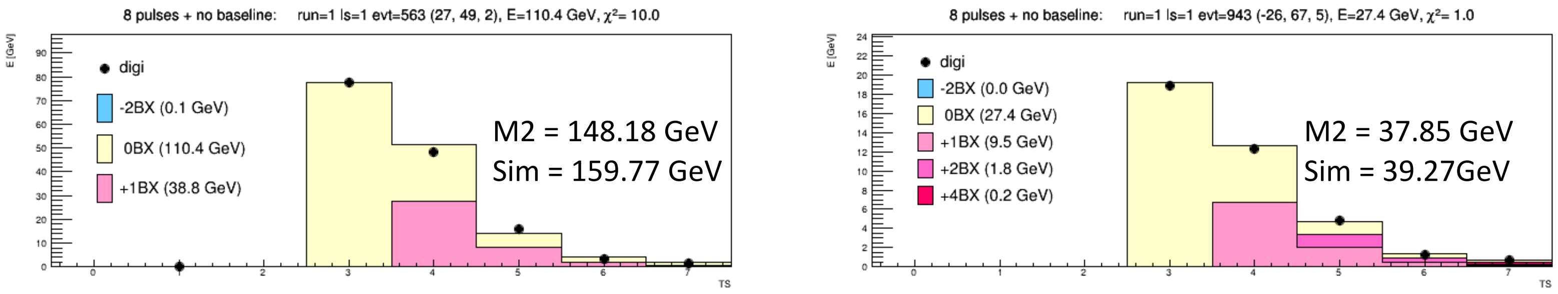
$\sum d$: quadratic sum of uncertainties. $\sum p, i$: pulse shape uncertainty

- ~10 times faster than M2: can be used in HLT

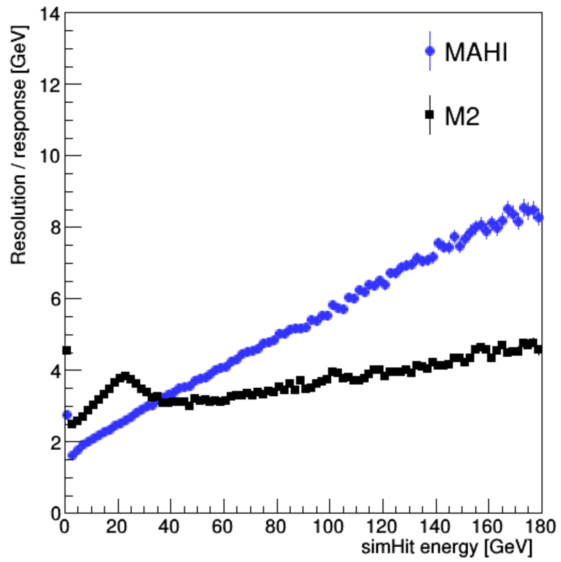
Disadvantage:

Cannot fit for pulse arrival time

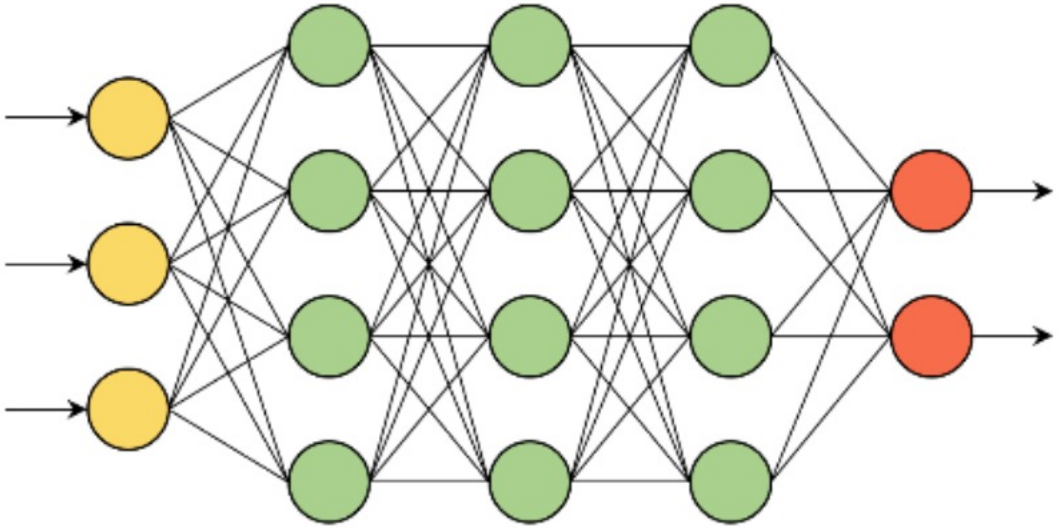
Bad performance at high energy



MAHI bad performance for late arrival pulses (MC sample without PU)



Feedforward Neural Network



Universal approximation theorem

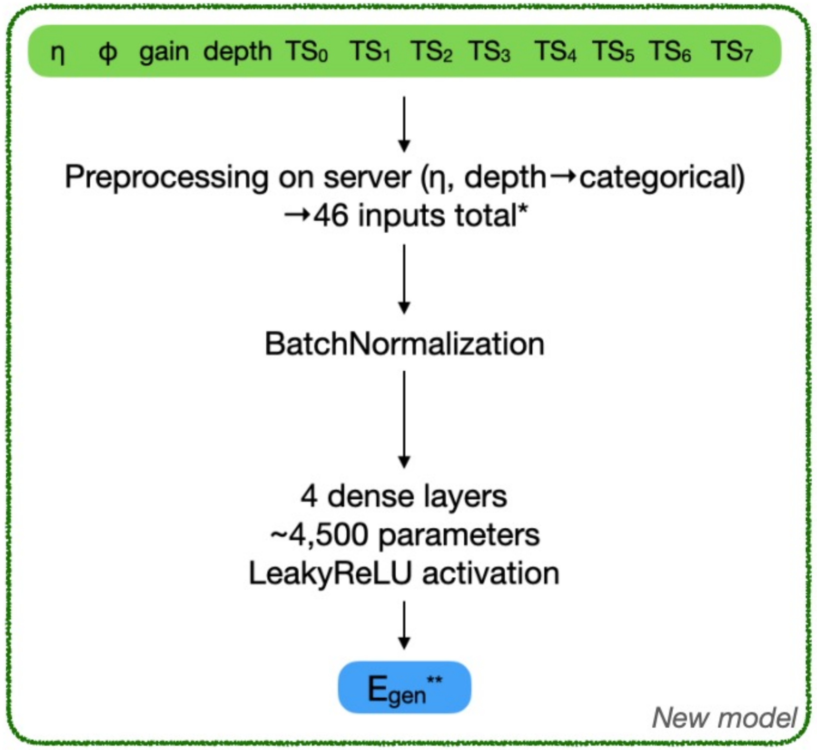
[K. Hornik, M. Stinchcombe, and H. White. 1989]

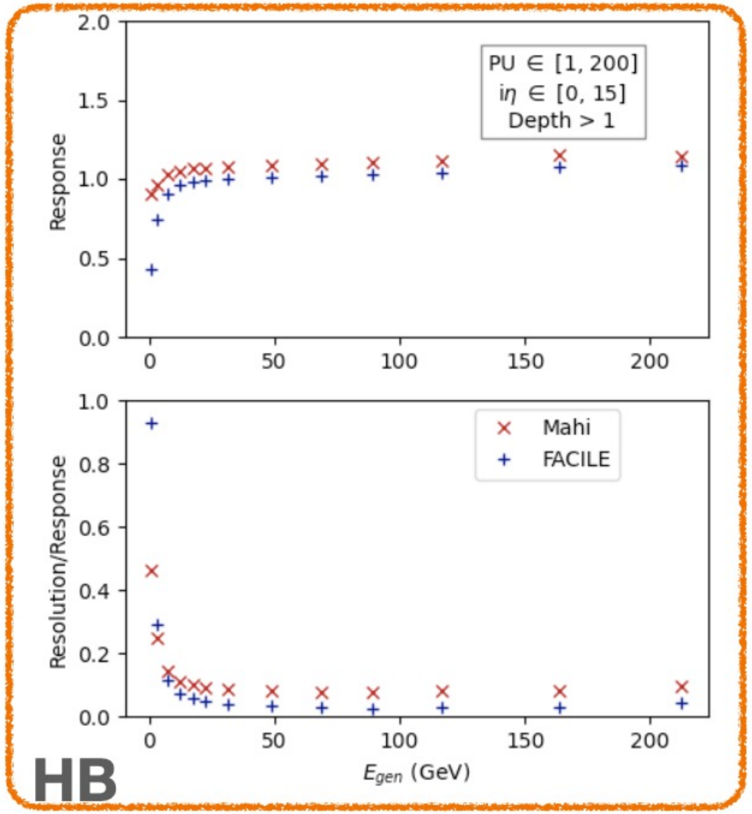
Feedforward Neural Network with as few as one hidden layer is able to approximate any measurable function

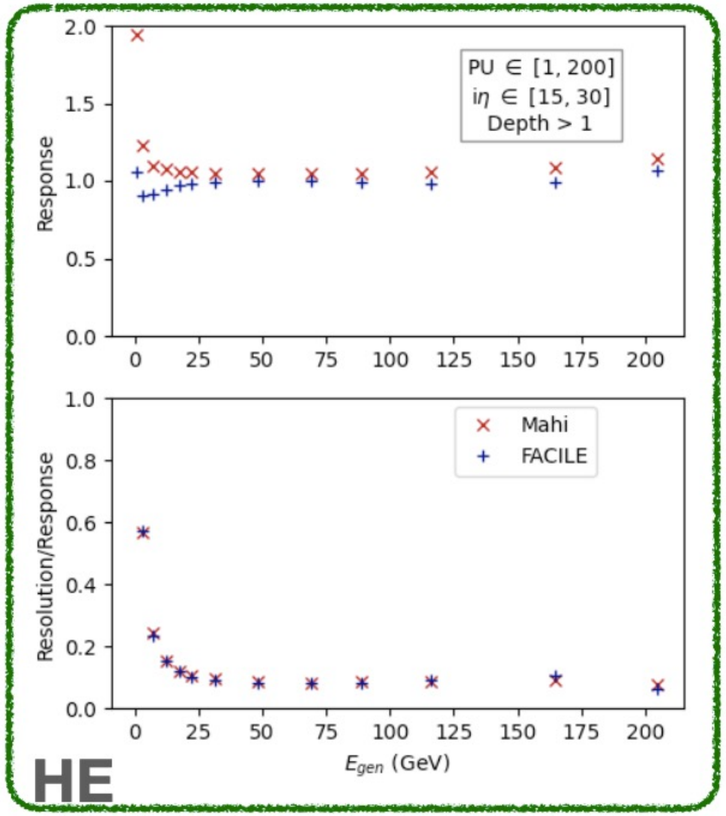
- Perceptron: a single neuron
  $y = f(\sum w_i x_i + b)$
  y: output. x: inputs. w: weights. b: a bias term.
  f: nonlinear activation function
  e.g. Sigmoid, ReLU, etc.
- Dense layer: outputs fully connected as inputs to the next layer

- Loss function: evaluate the predictions
  e.g. Mean Squared Error (MSE) for regression
  Loss (MSE) = $\frac{1}{n}\sum(y_{\text{pred}} - y_{\text{truth}})^2$
- Training: find weights and bias terms that minimize the loss function
  e.g. stochastic gradient decent (SGD)

HCAL Reco with DNN





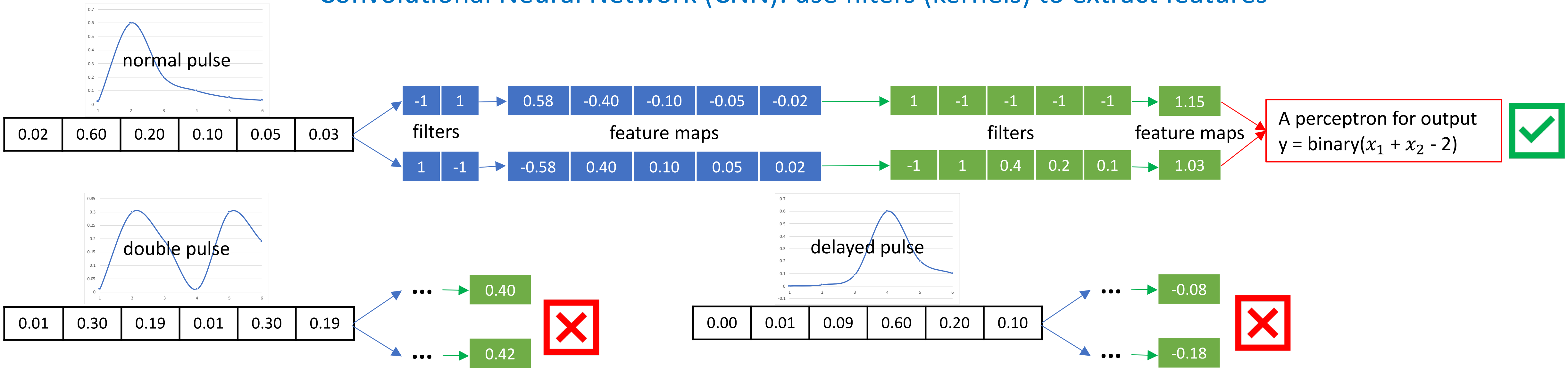


[FACILE]

Goal: a lite architecture with performance similar to MAHI, and can run on FPGA for L1 trigger

Convolutional Neural Network

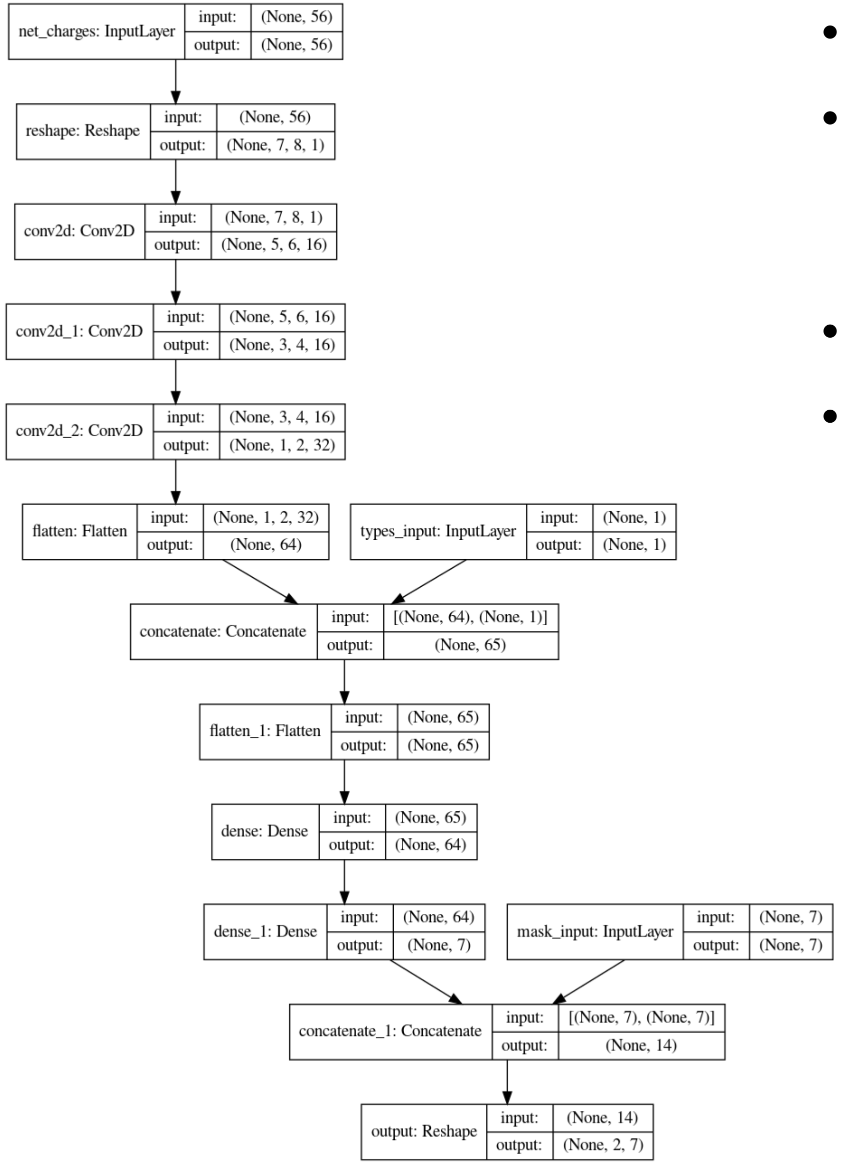
Convolutional Neural Network (CNN): use filters (kernels) to extract features



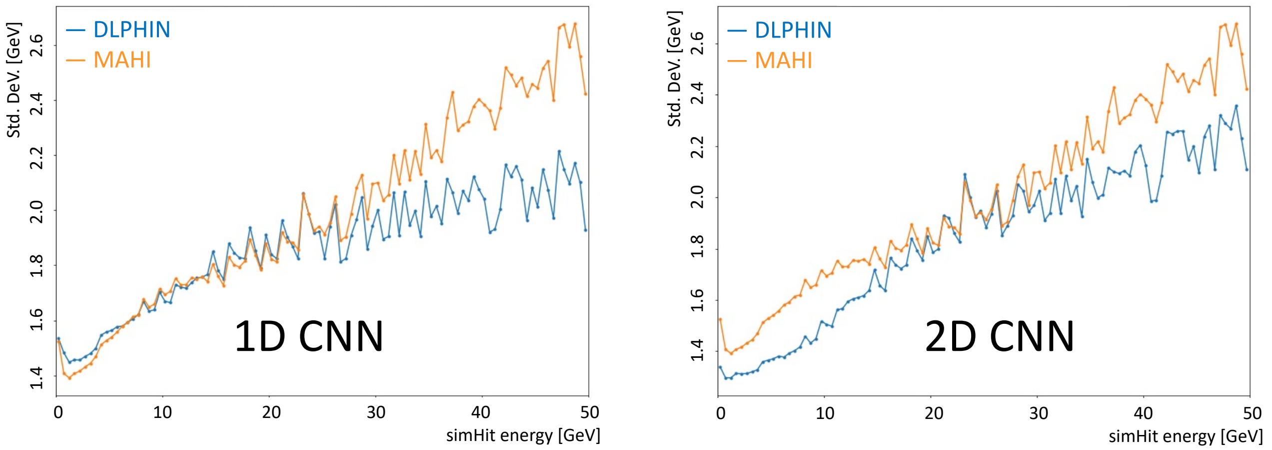
Example: select signal (normal pulse) from common backgrounds (double pulse, delayed pulse, etc)

1. Extract low-level features of rising and falling
2. Extract high-level features: the location and multiplicity of the low-level features
3. A simple perceptron for output. 1: signal, 0: background

HCAL Reco with CNN



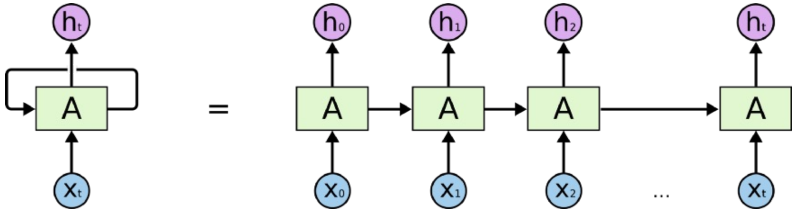
- DLPHIN = Deep Learning Processes for HCAL INtegration
- Architecture evolved from 1D CNN to 2D CNN
  - Dim. 1: net charges of 8TS
  - Dim. 2: depth → exploit correlations among channels in a tower
- More than 3 times faster than MAHI reco (both on CPU)
- Better performance from upstream to downstream
  - channel-level → single particle-level → jet-level

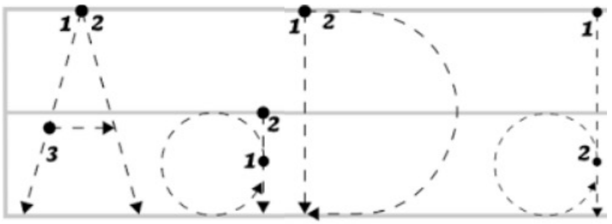


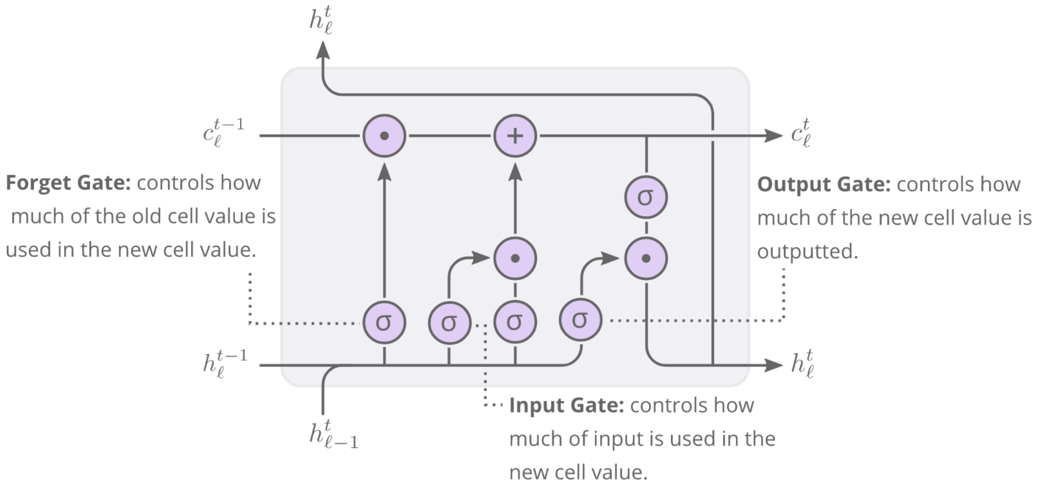
channel-level resolution

Recurrent Neural Network

Recurrent Neural Network (RNN): add a time dimension to better process sequential inputs







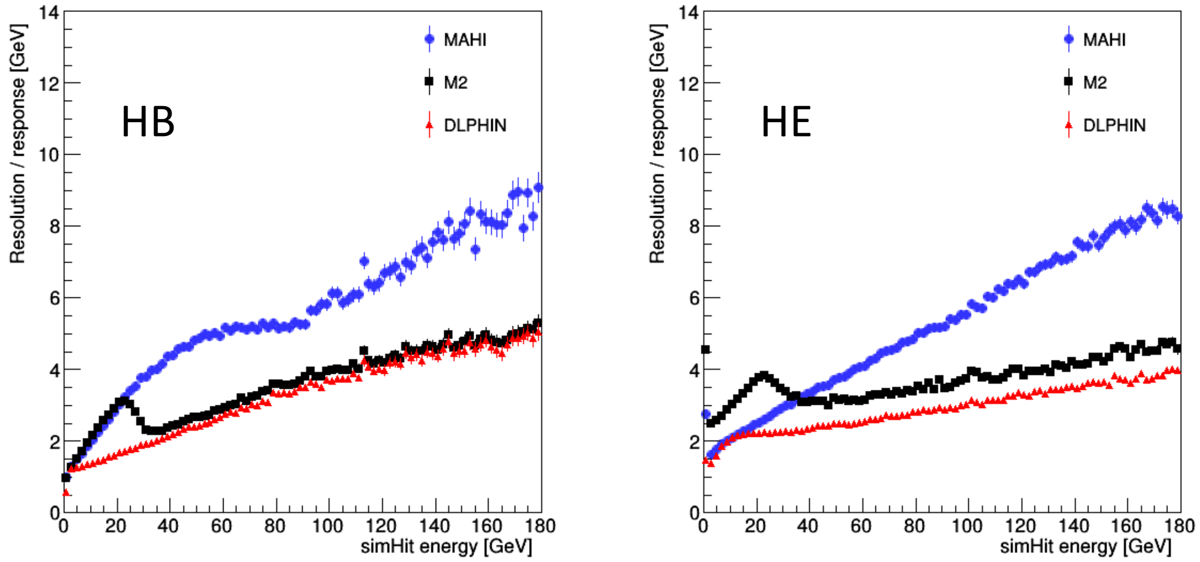
- Feedforward: process all inputs x_i at once. $y = f(\sum w_i x_i + b)$
- RNN: process x_i in time steps; each time step has a state h_t that is updated recurrently. $h_t = f(w_h h_{t-1} + w_x x_t)$
  Output $y_t = f\left(w_y h_t\right)$

- Example: handwriting recognition, letter "a" vs "d"
  - Feedforward: inputs from static image
  - RNN: sequential inputs as "strokes"

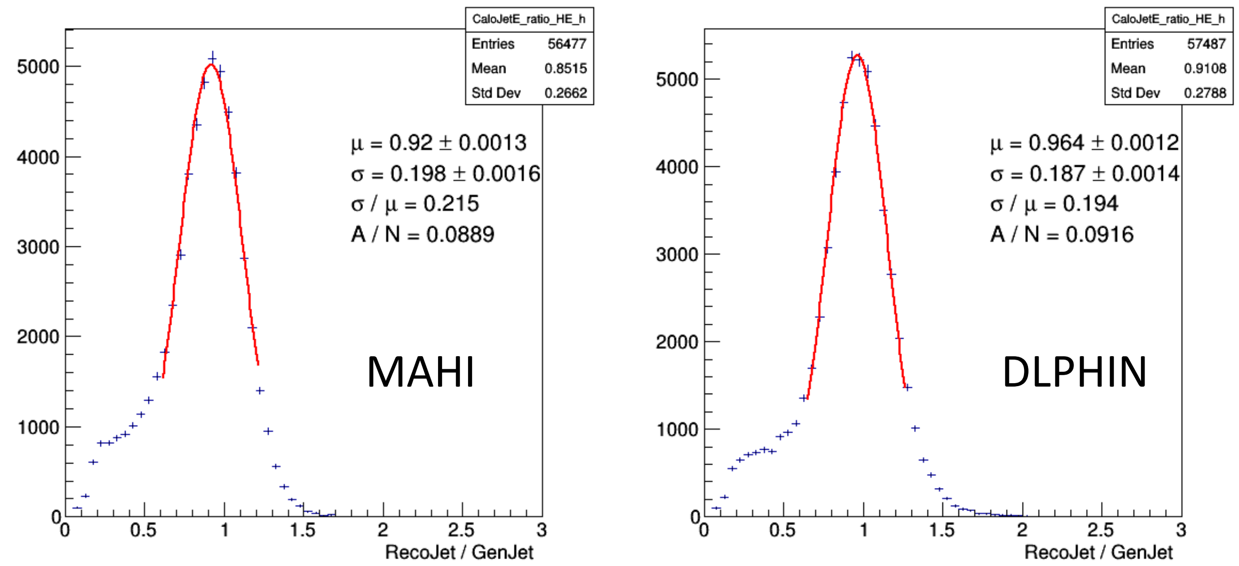
- A common issue for vanilla RNN: vanishing gradient problem
- Most used solution: Long Short Term Memory (LSTM)
  - Add gates to RNN units to control what info is passed through
- Charge inputs of 8TS are also sequential inputs
  - Tried LSTM in DLPHIN, only several percent improvement
  - Keep the 2D CNN for a lite and fast architecture

DLPHIN performance

HCAL channel-level resolutions



Particle-level ($\pi^{\pm}$) resolutions



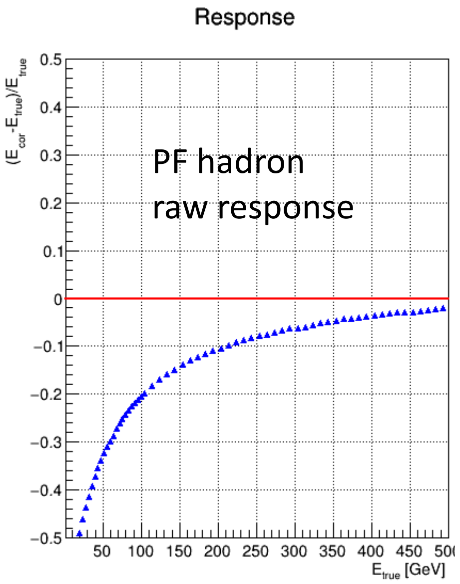
- HCAL channel-level resolutions: 1TeV pion-gun MC. Compare reconstructed energy to simulation energy
  DLPHIN has better resolution than both MAHI and M2
- Single particle-level ($\pi^{\pm}$) resolutions: 50GeV pion-gun MC. Match calorimeter jets to generated pions
  DLPHIN resolution ~10% better than MAHI

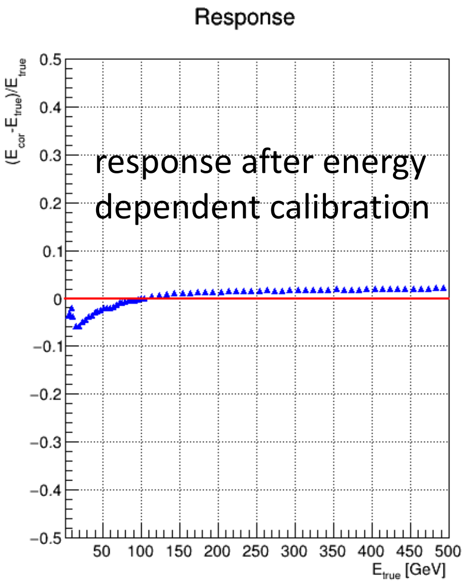
PF hadron calibrations

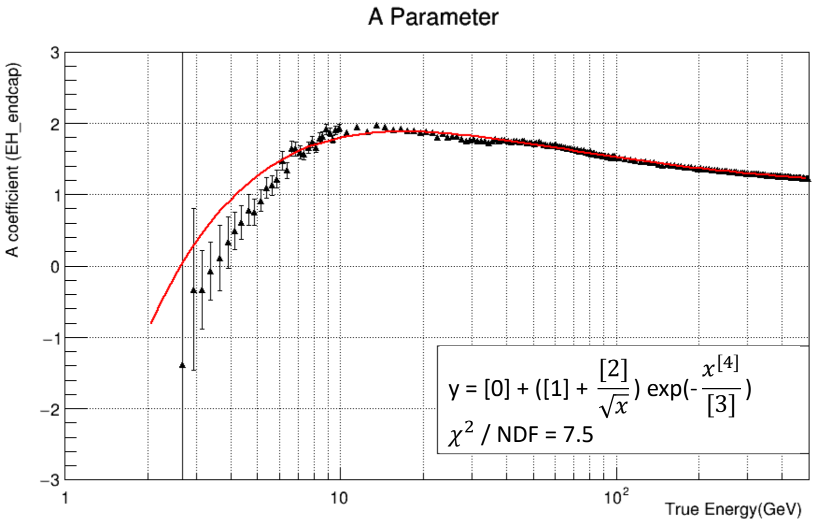
- HCAL is an under-compensating calorimeter
  - EM component of hadron shower smaller at low energy → Low response at low energy
- Need particle-level calibration
  - HCAL calibrated with 50 GeV pion (ECAL energy negligible) for channel energy
  - Then energy dependent calibration on PF hadrons
- PF hadron (start showering in ECAL) as an example

$$\mathrm{E_{corr}} = \mathrm{A(E)} * \mathrm{E_{raw}^{ECAL}} + \mathrm{B(E)} * \mathrm{E_{raw}^{HCAL}} + \text{offset}$$

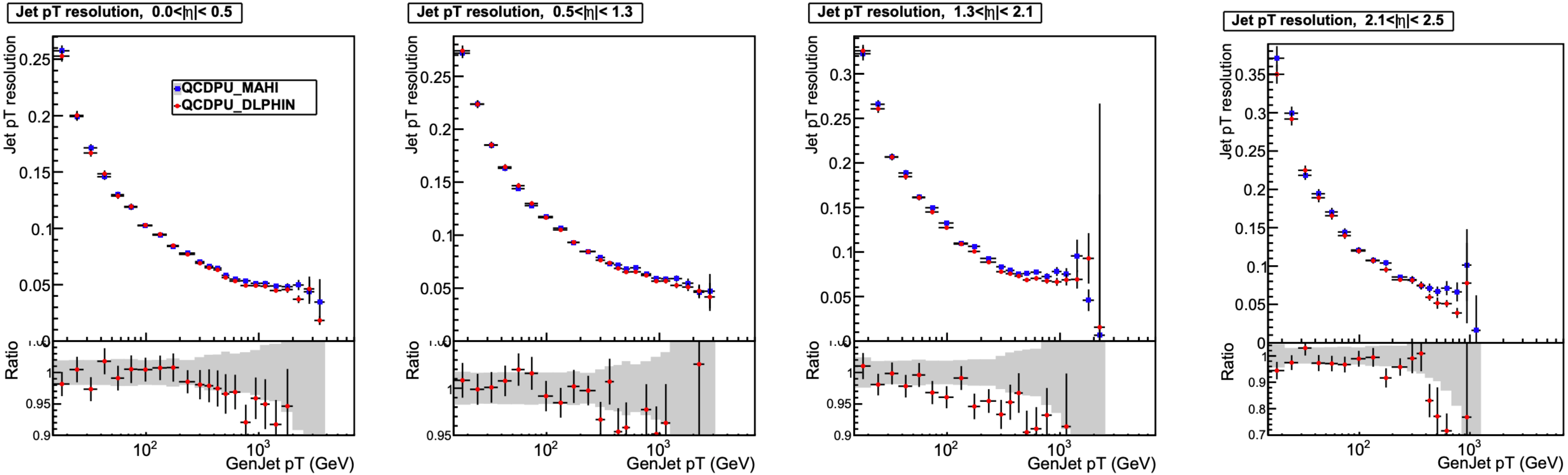
  - Currently the parameters A(E), B(E), etc are based on function fitting
  - Plan to use ML on this step, expect big improvements on some bad fittings
- Downstream tests (e.g. jet resolution) have to be after PF calibrations







DLPHIN performance



- PF jet-level resolutions: QCD MC (flat p_T 15-3000 GeV). Match PF jets to generator level jets
- PF jet performance dominated by tracker at low energy
- DLPHIN resolution 5% / 10% better than MAHI for HB / HE at high energy

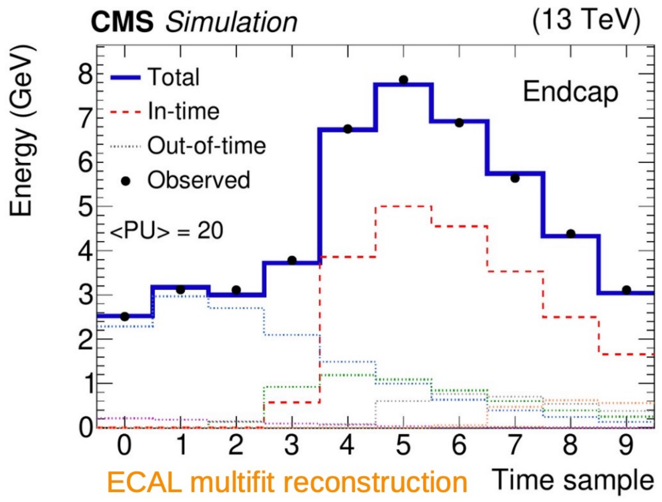
Summary and Outlook

- Introduced some common ML architectures and the physics behind the design
  - For HCAL reco: DLPHIN showed better performance and speed than traditional fitting algos

- Other possible applications of DLPHIN
  - A lite version on FPGA for L1T (DLPHIN/FACILE collaboration)

    Will bring huge improvements on hadrons and leptons
  - ECAL energy reconstruction

    Currently also use traditional fitting
  - Use in future detectors in CEPC, FCC etc.



Thanks for your attention!

Backup slides

dijet resonance search

The analysis considers the four leading jets in an event

- **4-jet mass** is unambiguously defined, but any event has three possible di-jet pairings from which to obtain **average 2-jet mass**
- Minimizing $\boldsymbol{\Delta R = |(\Delta R_1 - 0.8)| + |(\Delta R_2 - 0.8)|}$, where $\Delta R_{1,2}$ are the ΔR between the two jets of each combination, was found to be the optimal metric, yielding the highest expected signal significance for both the resonant and non-resonant analyses

Additional cuts are made on the resulting pairs to reduce background and "wrong combinations" of jets:

- A mass asymmetry requirement: $\boldsymbol{|m_1 - m_2| / (m_1 + m_2) < 0.1}$
- An angular requirement between the pairs: $\boldsymbol{\Delta\eta < 1.1}$
- An angular requirement between jets in a pair: $\boldsymbol{\Delta R_{1,2} < 2.0}$

Also working on an ML approach for dijet pairing

Background functions

- Both analyses use smoothly falling functions to model the background in the usual "bump hunt" approach
- Three forms are considered:

Dijet-3p:

$$\frac{d\sigma}{dm_{4j}} = \frac{p_0(1 - m_{jj}/\sqrt{s})^{p_1}}{(m_{jj}/\sqrt{s})^{p_2}}$$

PowExp-3p:

$$\frac{d\sigma}{dm_{4j}} = \frac{p_0}{(m_{jj}/\sqrt{s})^{p_1}} e^{-p_2(m_{jj}/\sqrt{s})}$$

ModDijet-3p:

$$\frac{d\sigma}{dm_{4j}} = \frac{p_0(1 - (m_{jj}/\sqrt{s})^{1/3})^{p_1}}{(m_{jj}/\sqrt{s})^{p_2}}$$

Systematic uncertainties

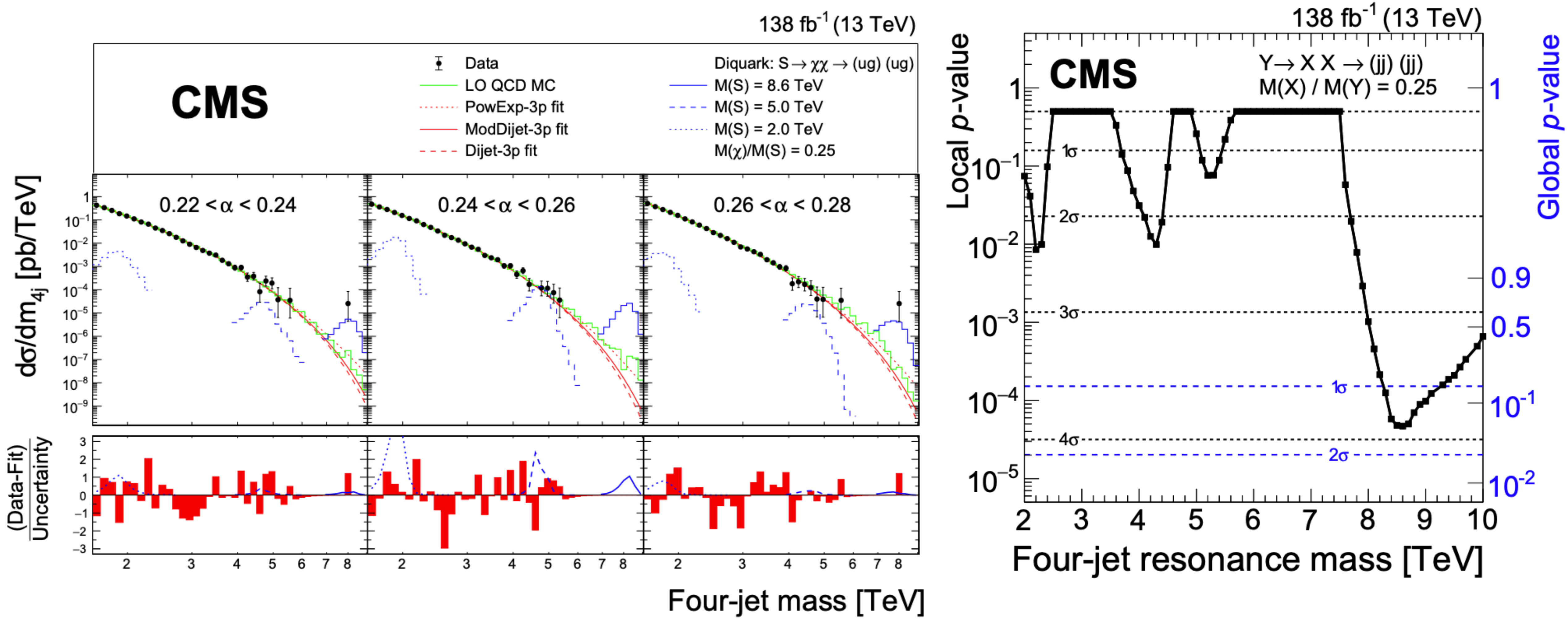
- The two searches have the same sources of systematic uncertainty:
  - **Signal:**

Systematic Uncertainty Source	Nominal Value	Uncertainty
Jet Energy Resolution	no smearing	10% of RECO resolution
Jet Energy Scale	no shift	$\pm 2\%$ shift of m_{jj} or m_{4j}
Luminosity	137.5 fb^{-1}	$\pm 1.6\%$

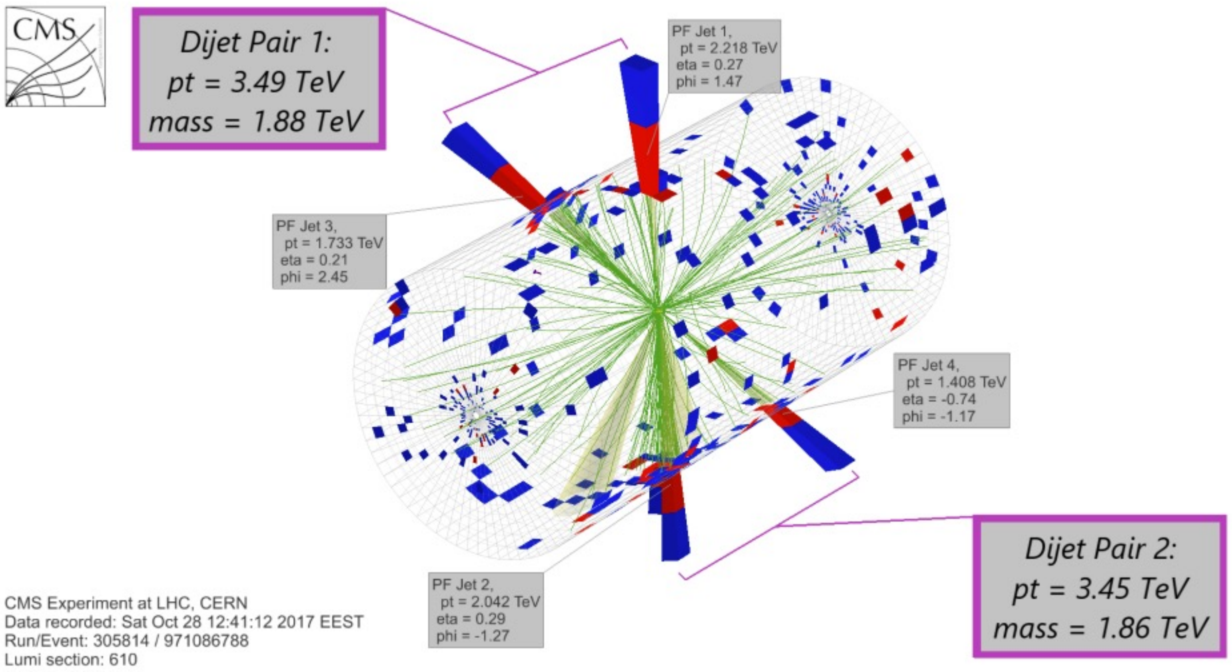
  - **Background** (*dominant*):
    Uncertainty from the fit (including effect from envelope method)

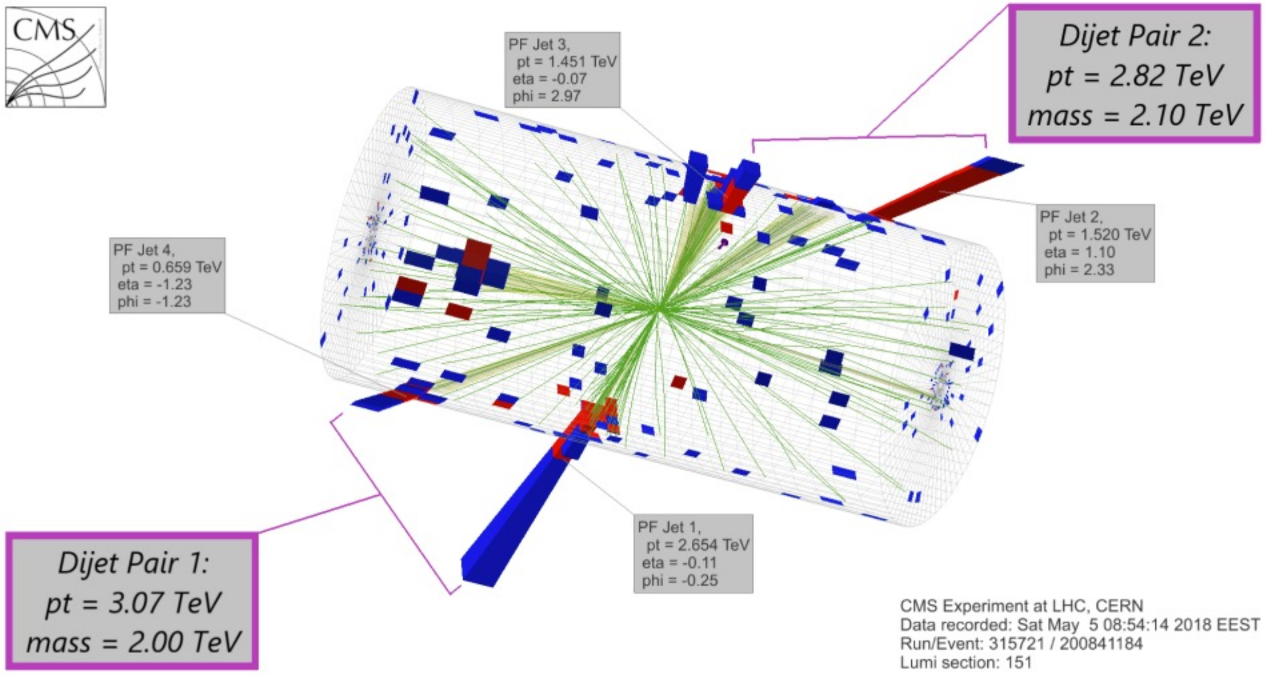
- All other sources (e.g. PDFs) are found to be neligible.

dijet resonance search

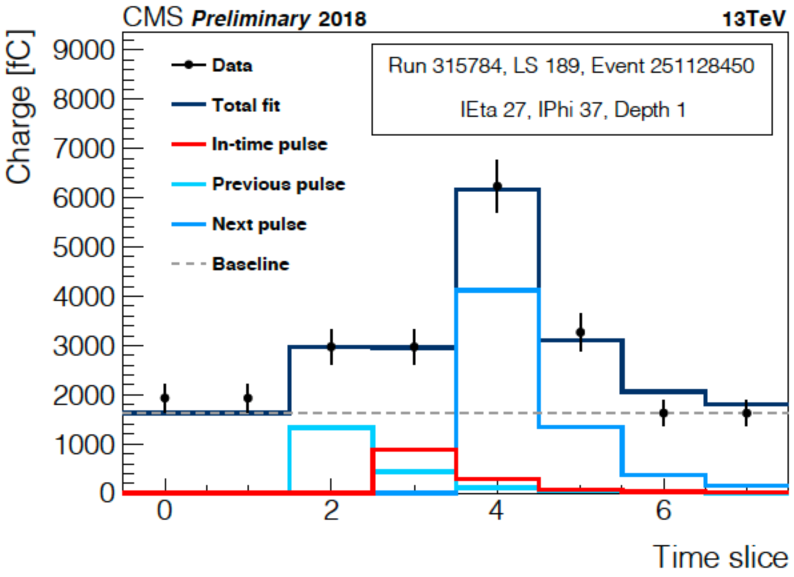


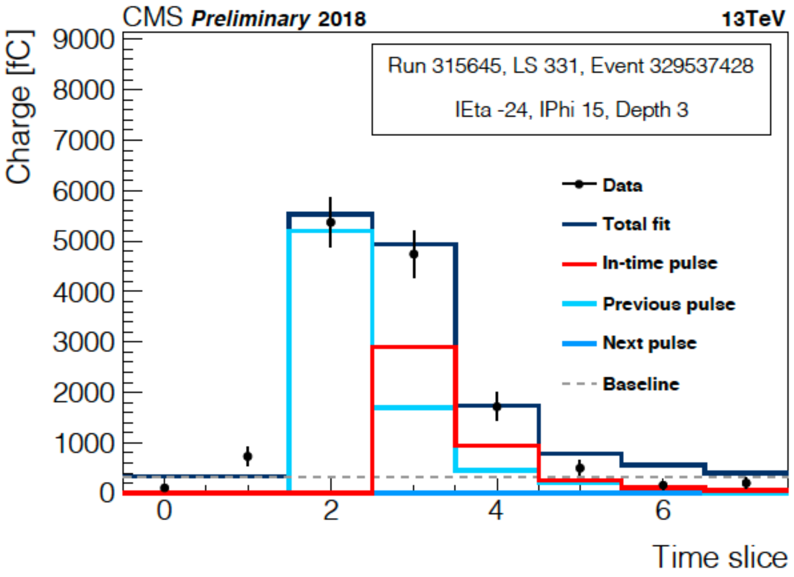
Event displays

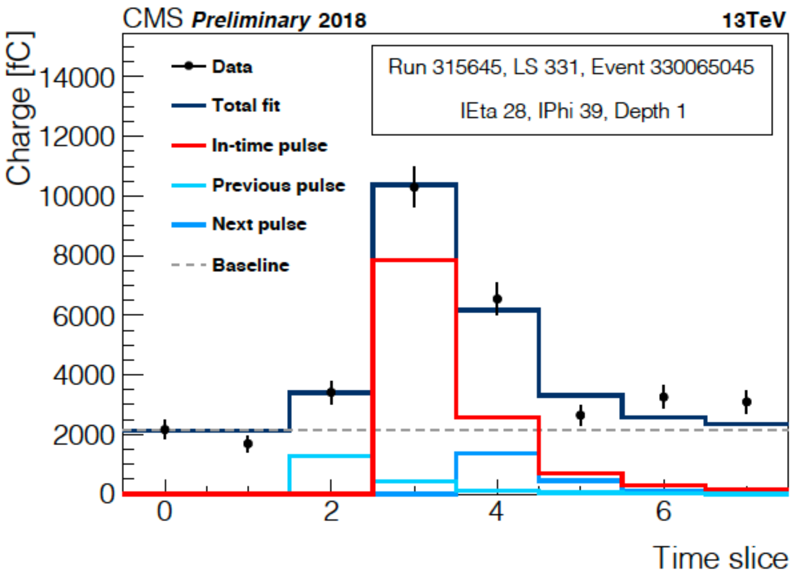


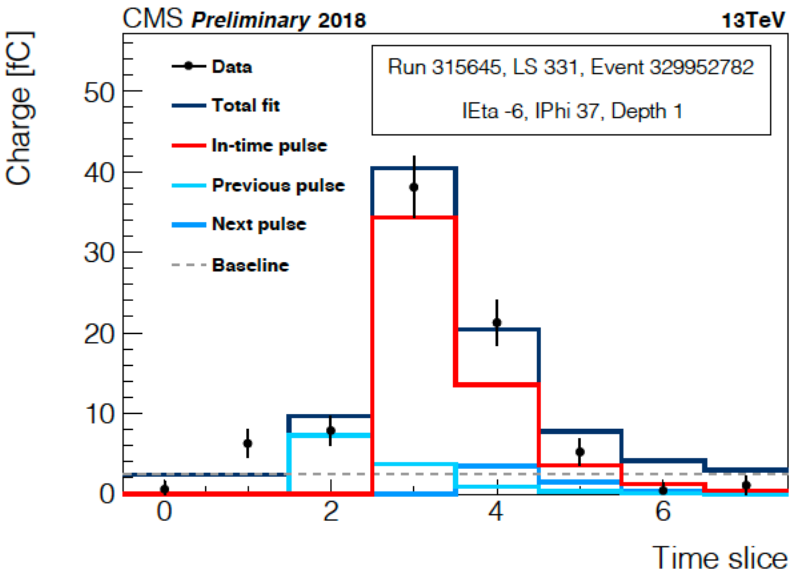


Method 2


CMS Preliminary 2018
13TeV
Run 315784, LS 189, Event 251128450
IEta 27, IPhi 37, Depth 1
Data
Total fit
In-time pulse
Previous pulse
Next pulse
Baseline
Charge [fC]
Time slice


CMS Preliminary 2018
13TeV
Run 315645, LS 331, Event 329537428
IEta -24, IPhi 15, Depth 3
Data
Total fit
In-time pulse
Previous pulse
Next pulse
Baseline
Charge [fC]
Time slice


CMS Preliminary 2018
13TeV
Run 315645, LS 331, Event 330065045
IEta 28, IPhi 39, Depth 1
Data
Total fit
In-time pulse
Previous pulse
Next pulse
Baseline
Charge [fC]
Time slice

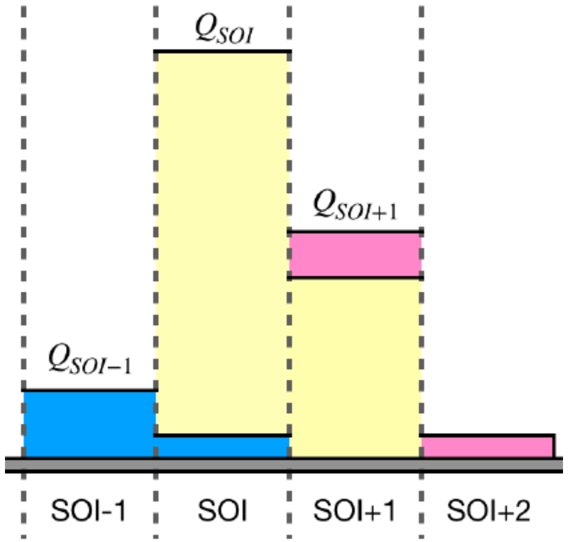

CMS Preliminary 2018
13TeV
Run 315645, LS 331, Event 329952782
IEta -6, IPhi 37, Depth 1
Data
Total fit
In-time pulse
Previous pulse
Next pulse
Baseline
Charge [fC]
Time slice

Method 3

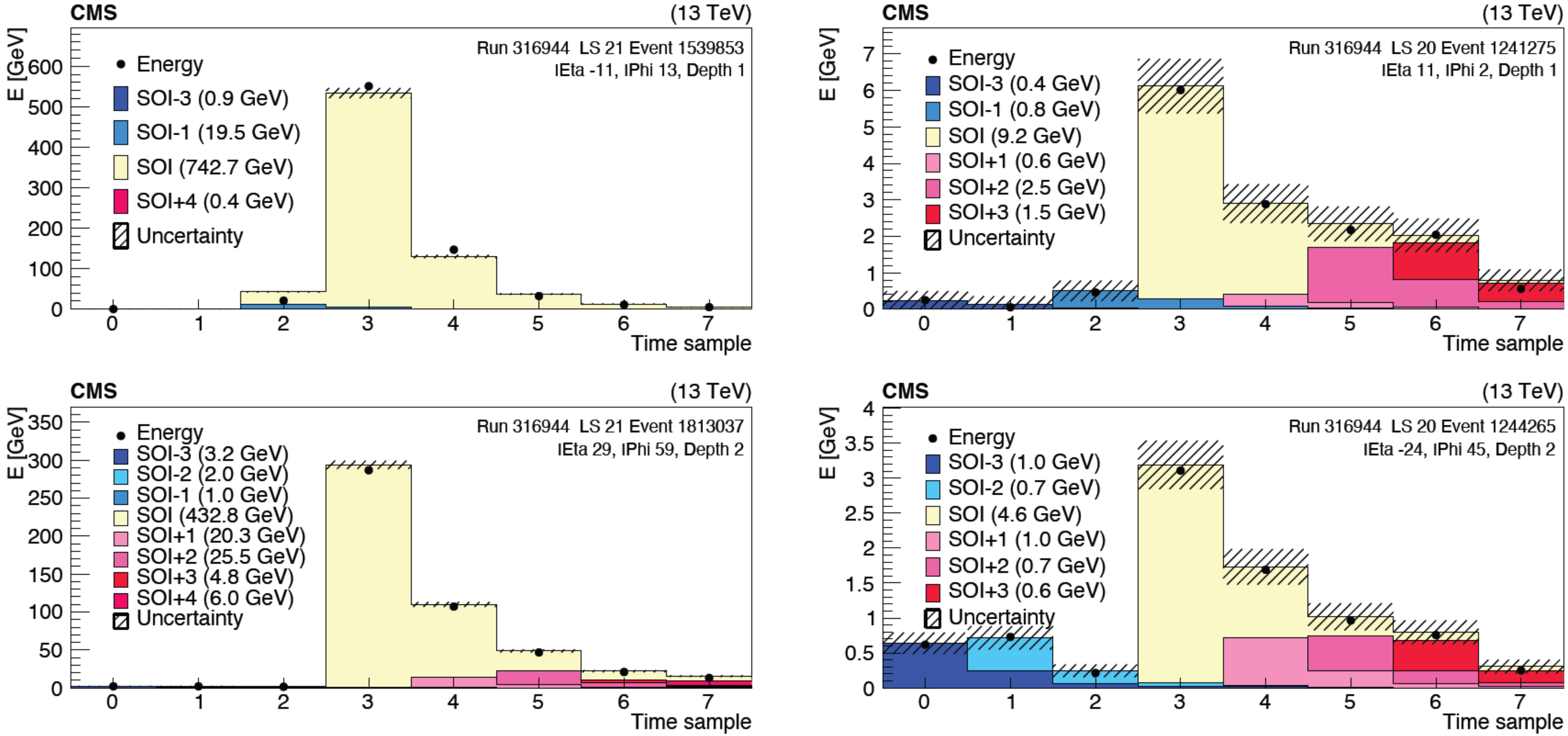
- Online version of M2, used in 2016 and 2017
- Fit 3 pulses (SOI - 1, SOI and SOI + 1) to only 3 TS
- Drop the arrival time term
- Use constant baseline term
- Fitting → solving linear equation

$$\begin{bmatrix} TS_{SOI-1} \\ TS_{SOI} \\ TS_{SOI+1} \end{bmatrix} = \begin{bmatrix} f_0 & 0 & 0 \\ f_1 & f_0 & 0 \\ f_2 & f_1 & f_0 \end{bmatrix} \begin{bmatrix} A_{SOI-1} \\ A_{SOI} \\ A_{SOI+1} \end{bmatrix} + \begin{bmatrix} B \\ B \\ B \end{bmatrix}$$

f0, f1 and f2 are the premeasured fractions of the pulse template in +0, +1 and +2 TS



MAHI


CMS
(13 TeV)
Run 316944 LS 21 Event 1539853
IEta -11, IPhi 13, Depth 1
Energy
SOI-3 (0.9 GeV)
SOI-1 (19.5 GeV)
SOI (742.7 GeV)
SOI+4 (0.4 GeV)
Uncertainty
E [GeV]
Time sample
CMS
(13 TeV)
Run 316944 LS 20 Event 1241275
IEta 11, IPhi 2, Depth 1
Energy
SOI-3 (0.4 GeV)
SOI-1 (0.8 GeV)
SOI (9.2 GeV)
SOI+1 (0.6 GeV)
SOI+2 (2.5 GeV)
SOI+3 (1.5 GeV)
Uncertainty
E [GeV]
Time sample
CMS
(13 TeV)
Run 316944 LS 21 Event 1813037
IEta 29, IPhi 59, Depth 2
Energy
SOI-3 (3.2 GeV)
SOI-2 (2.0 GeV)
SOI-1 (1.0 GeV)
SOI (432.8 GeV)
SOI+1 (20.3 GeV)
SOI+2 (25.5 GeV)
SOI+3 (4.8 GeV)
SOI+4 (6.0 GeV)
Uncertainty
E [GeV]
Time sample
CMS
(13 TeV)
Run 316944 LS 20 Event 1244265
IEta -24, IPhi 45, Depth 2
Energy
SOI-3 (1.0 GeV)
SOI-2 (0.7 GeV)
SOI (4.6 GeV)
SOI+1 (1.0 GeV)
SOI+2 (0.7 GeV)
SOI+3 (0.6 GeV)
Uncertainty
E [GeV]
Time sample

Processing time

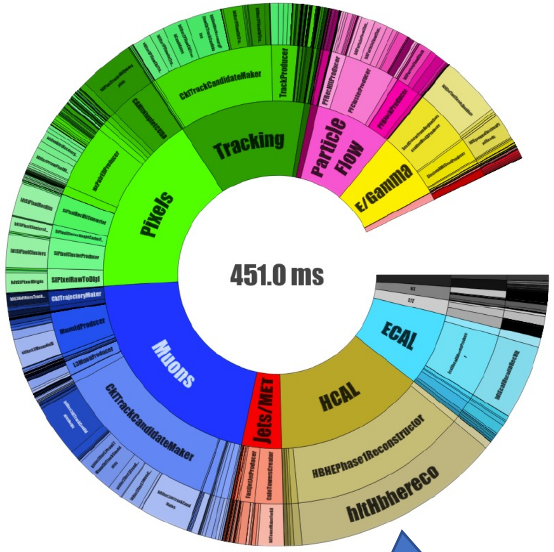
```
50.62  SimpleHBHEPhase1Algo::reconstruct(HBHEChannelInfo const&, HcalRecoParam const*, HcalCalibrations const&, bool) [40]
48.76  MahiFit::phase1Apply(HBHEChannelInfo const&, float&, float&, float&, bool&, float&) const [41]
48.35  MahiFit::doFit(std::array<float, 4ul>&, int) const [42]
29.34  MahiFit::minimize() const [43]
22.40  MahiFit::nnls() const [44]
14.23  DLPHIN::DLPHIN_run(HcalDbService const&, edm::SortedCollection<HBHEChannelInfo, edm::StrictWeakOrdering<HBHEChannel
Info> > const*, edm::SortedCollection<HBHERecHit, edm::StrictWeakOrdering<HBHERecHit> >*) [45]
```

↑ Total CPU time [s]

↑ Process name

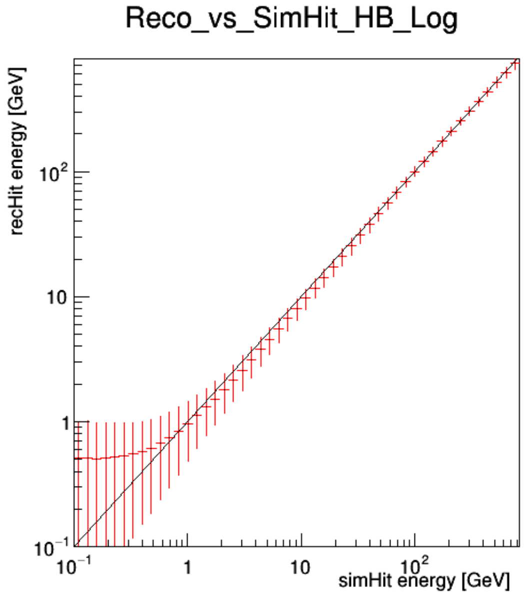
Processing time profiled with IgProf
CMSSW_12_4_3, 1000 events in 2022C_JetHT data
DLPHIN processing time < 30% of MAHI (both on CPU)

HCAL on CPU used to cost ~15% of total HLT time
DLPHIN on CPU can achieve ~negligible (<5%) of total HLT time, like MAHI on GPU

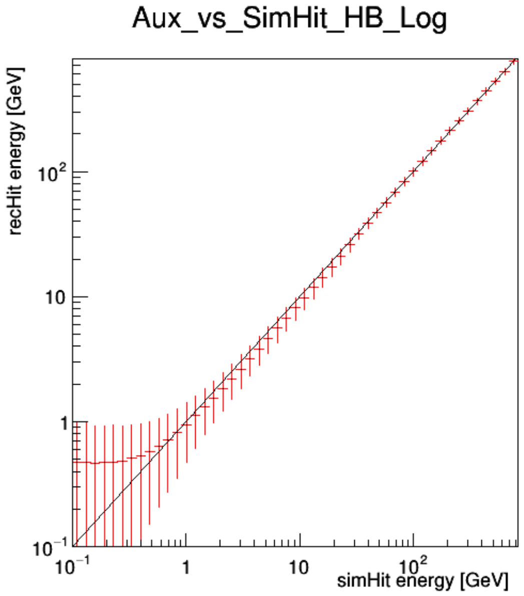


MAHI on CPU

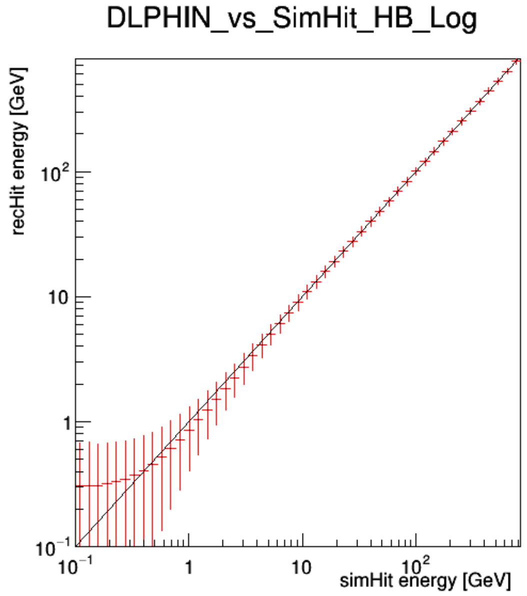
recHit resolution in HB



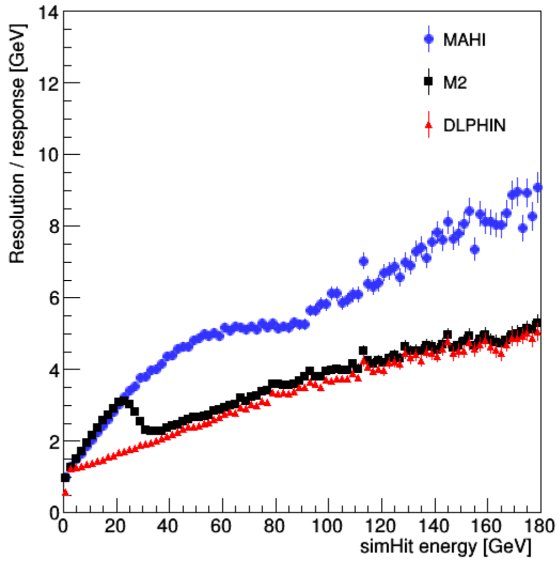
MAHI



M2



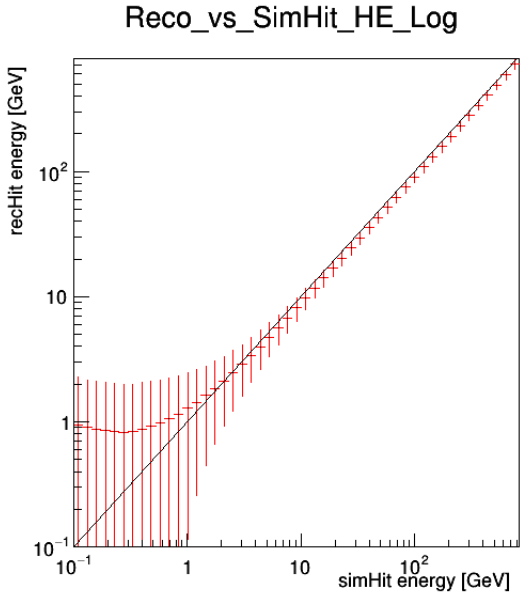
DLPHIN



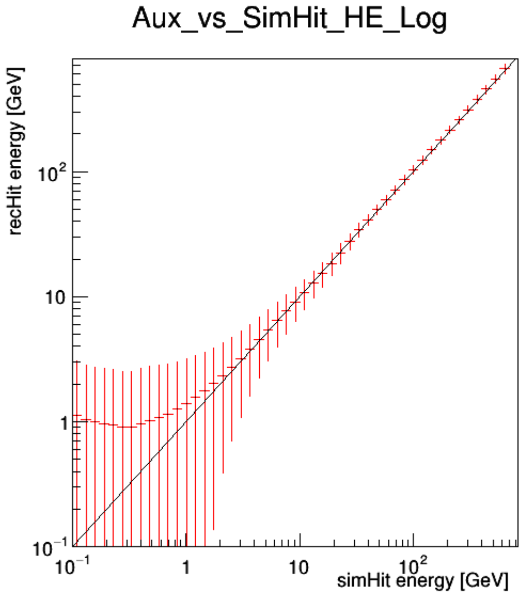
Normalized resolutions

Resolutions vs simHits with UL 2018 pion-gun sample (realistic PU)
M2 forced to fit 1 pulse for HPD charge > 100 fC (~20 GeV), hence the kink
HB only had 1 depth in Run2. DLPHIN is expected to be even better in Run3

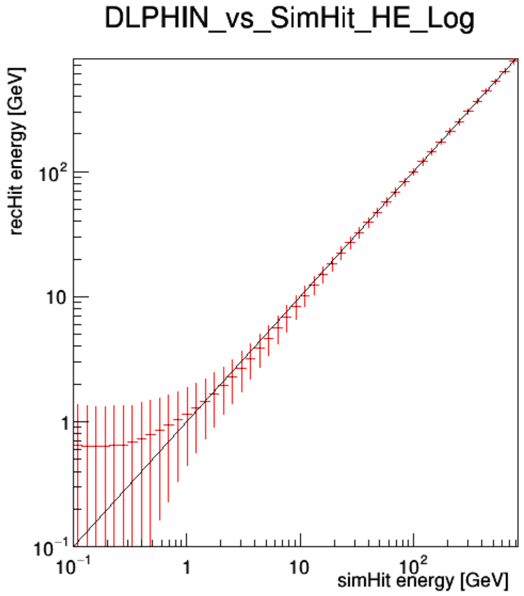
recHit resolution in HE



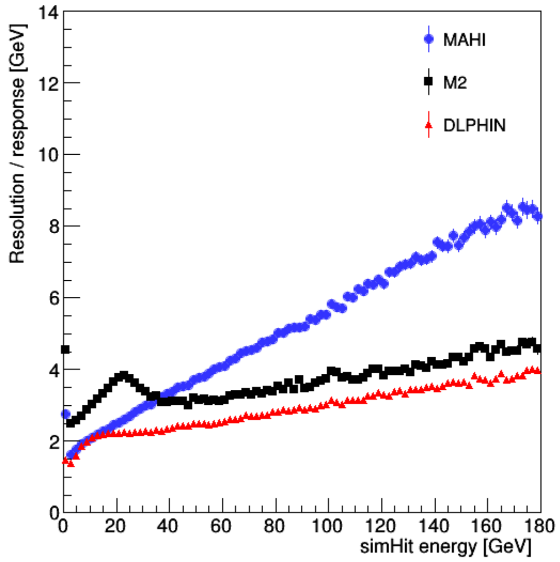
MAHI



M2



DLPHIN



Normalized resolutions

Resolutions vs simHits with UL 2018 pion-gun sample (realistic PU)
M2 forced to fit 1 pulse for SiPM charge > 25k fC (~20 GeV), hence the kink