Hadronic Shower Substructure Reconstruction with Graph Neural Networks

14th Annual Meeting of the Helmholtz Alliance "Physics at the Terascale"

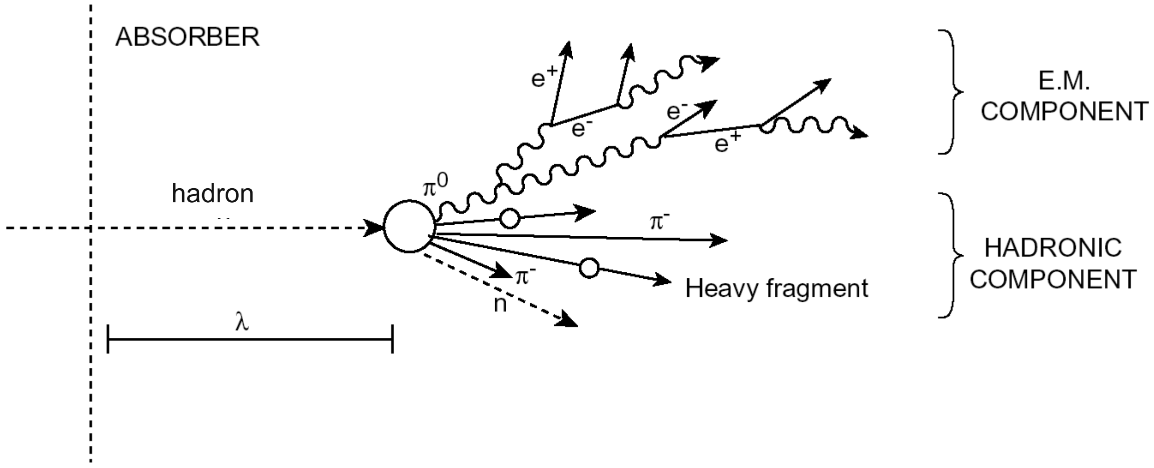
Vladimir Bocharnikov (DESY)
23 Nov 2021

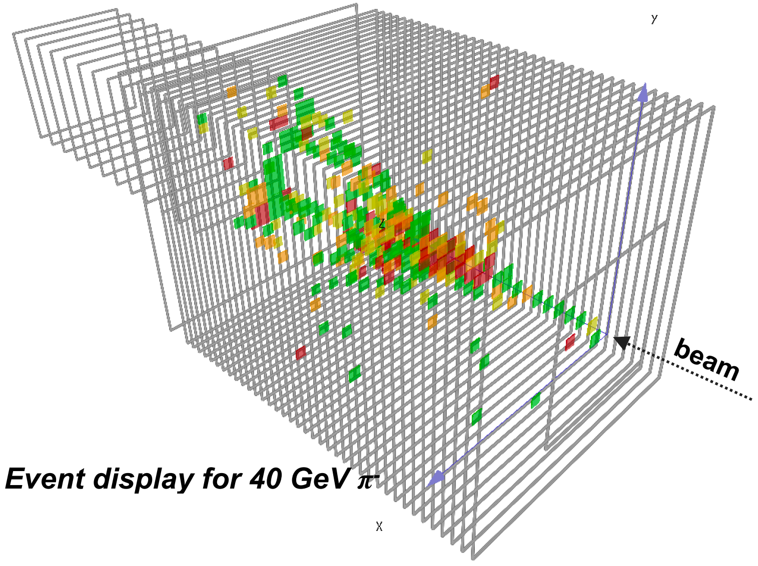
HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

Hadronic showers

General properties

- Hadronic shower development is rather complex:
  - Narrow EM core component from π^0/η
  - Surrounding halo dominated by charged hadrons
  - Large event-by-event fluctuation of EM/HAD ratio
  - Response to EM and HAD components is different in non-compensating calorimeters
  - Invisible energy as binding energy, nuclear recoil, neutrinos + late component
  - ➡ Limited hadronic energy resolution
  - ➡ Detailed simulation is challenging
- Highly granular calorimeter prototypes
  - Imaging capabilities provide detailed calorimetric images
  - Real test beam data for crosschecks and development of data-driven algorithms

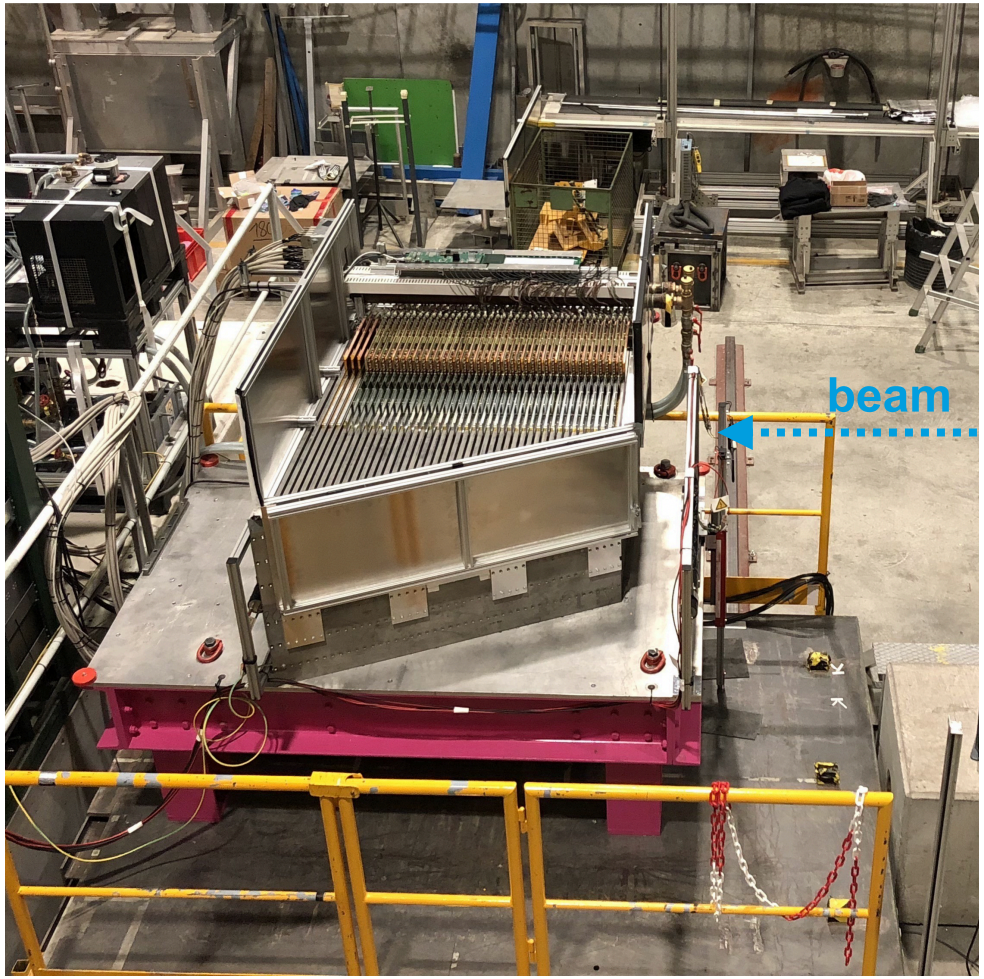




Event display for 40 GeV π^-

CALICE AHCAL

Test beam prototype.



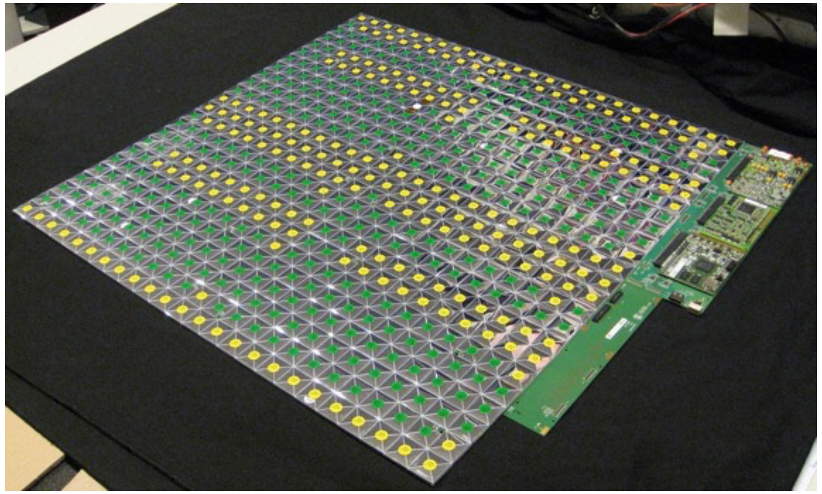
***39* active layers** of *24x24* scintillator tiles (*3x3* cm^2 each) with individual SiPM readout. Active layers alternate with ~*2 cm* steel absorber.

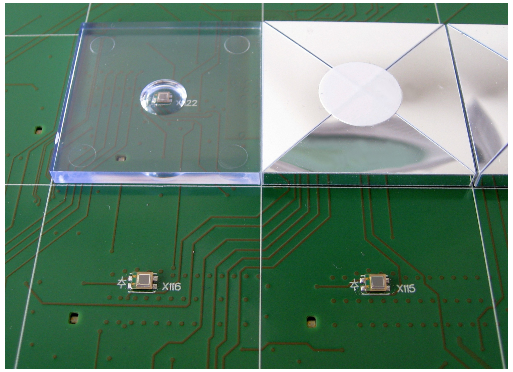
In total: ***~22000* channels** (*<1‰ dead channels*), ***~4 λ***, ***~38X0***

Beam particles: muons, electrons, **pions**

Energy range: ***10-200 GeV*** *in 10-40 GeV steps*

O(1M) hadron events per energy point

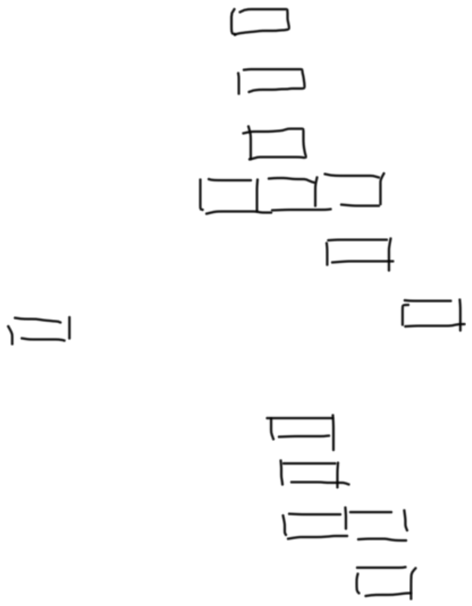




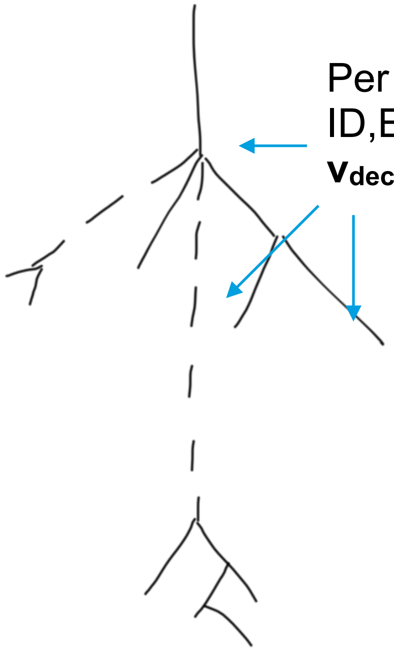
Calorimeter vision for hadronic showers

Ultimate goal and general approach

Set of hits in highly granular calorimeter



Particle interaction tree



Per particle:
ID,E, $\mathbf{p}$, $\mathbf{v}_{\mathbf{prod}}$, $\mathbf{v}_{\mathbf{decay}}$

Potential applications of hit to secondary particle association:

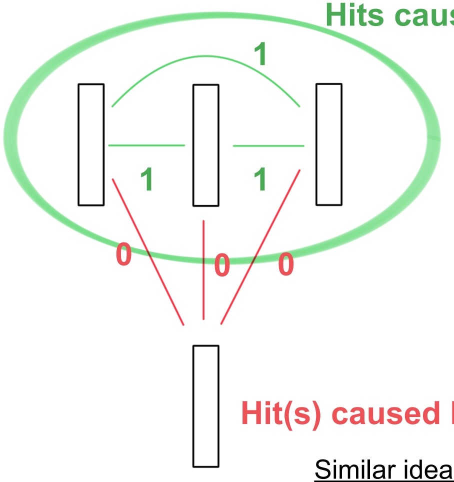
- Shower separation algorithms:
  - Recombination of secondaries between overlaid showers
- Validation of simulation performance:
  - Comparison of global physical distributions
  - Shower description on single event basis is possible

Graph representation of calorimeter event

First steps

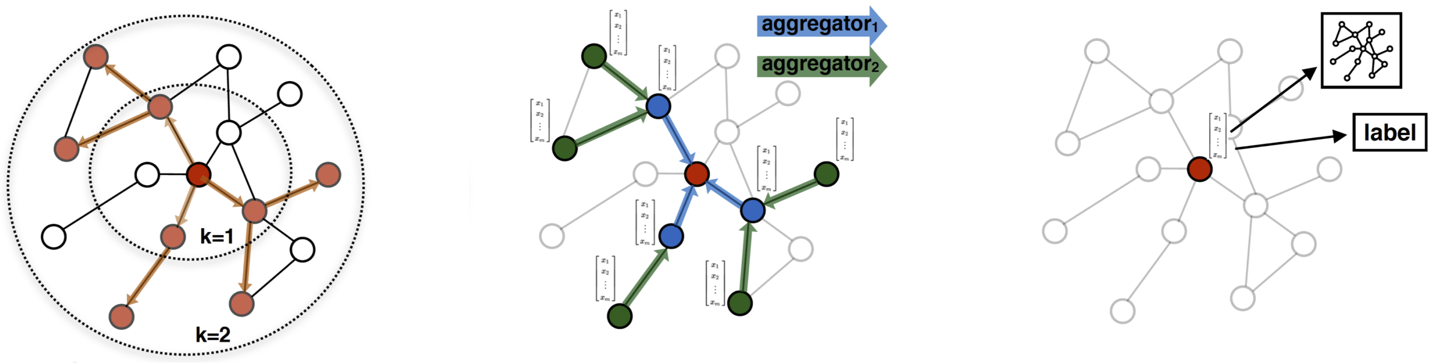
Event graph:

- Nodes - calorimeter hits
- Node features - position, energy, (time)
  - — Edges - neighbours within distance < R_{max} (Radius graph)
  - — Edge weights - 1 if pair of hits belong to same **fundamental object** (e/m sub-shower, track), otherwise 0
- ML **objective** - **predict edge weights** given the radius graph of event



Similar idea is discussed for CMS tracker and HGCAL

GraphSAGE (**SA**mple and aggre**G**at**E**) architecture (Graph neural network model (GNN)):

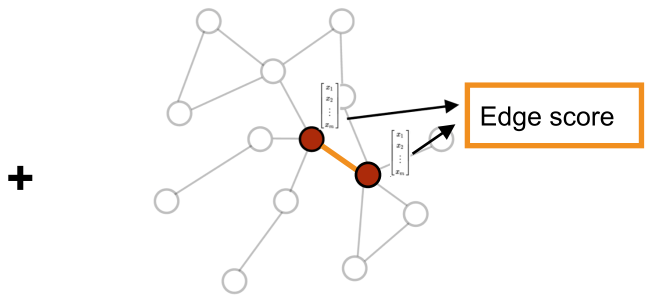


Sample neighbourhood of graph nodes

Aggregate feature information from neighbours

Get graph context embeddings for node using aggregated information

Custom output layer



Predict edge score for each pair of connected nodes using embedded features

Truth information from Monte-Carlo

Algorithm to find truth e/m objects

Simulations

Geant4 (v10.03.p02) QGSP_BERT_HP using CALICE AHCAL geometry

Pure energy deposition in cells (before digitalisation and reconstruction)

Truth electromagnetic sub-shower definition:

"**Electromagnetic**" particles: $e^{\pm}, \gamma, \pi^0, \eta$

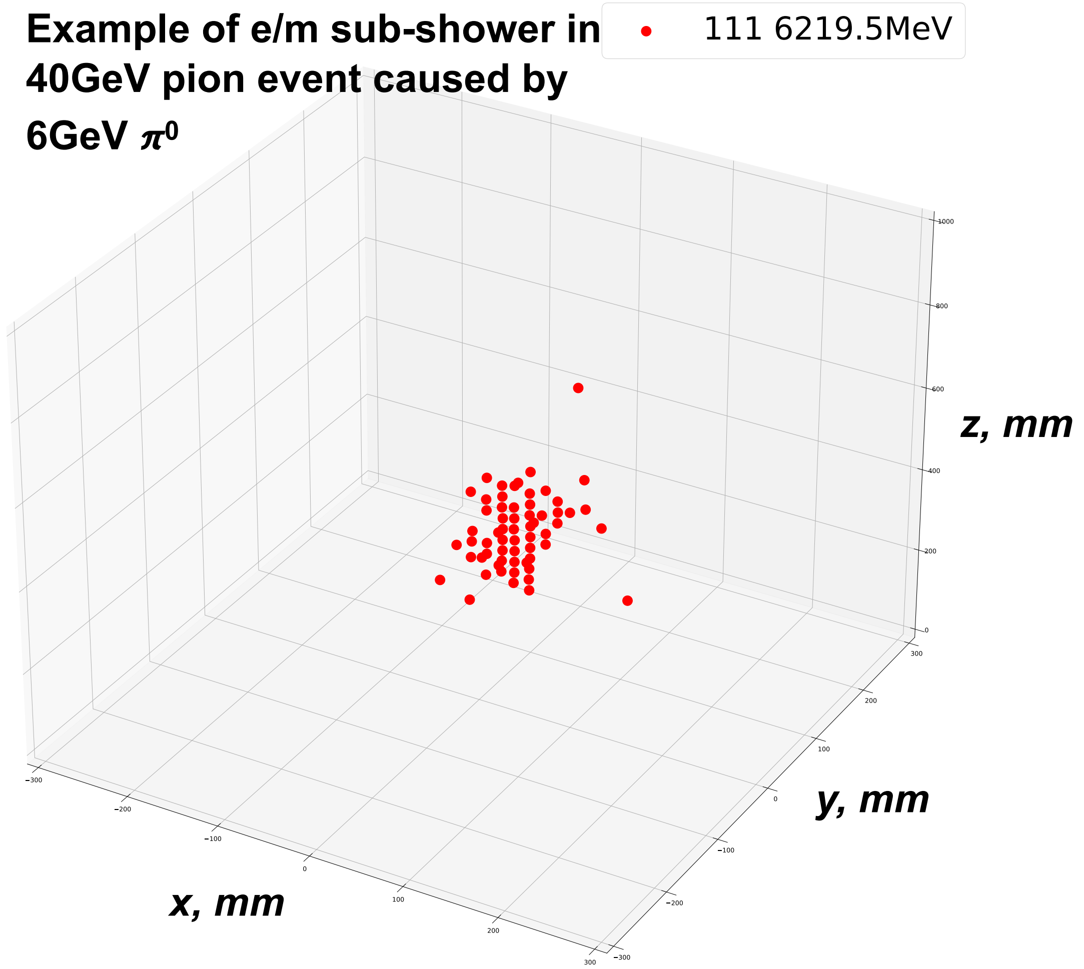
Energy threshold - *0.1GeV* (arbitrary now)

If MC particle is "electromagnetic", all it's "electromagnetic" daughters compose e/m shower are removed from further consideration

Corresponding simulated hits compose sub-shower,

0.5MIP cut: $E_{hit}>0.25MeV$

Example of e/m sub-shower in 40GeV pion event caused by 6GeV π^0



MC history for **ionising particles** is more complicated to easily define individual objects (tracks). Work in progress

Datasets and model parameters

Edge score model

Train&test dataset:

- ~6000 MC event graphs (50/50 split)
  - Pure energy deposition in calorimeter cells (before digitalisation and reconstruction)
  - **10-100 GeV pion** samples

➡ Radius graphs with calorimeter hit nodes (x,y,z,E_{hit}) R_{max} = ***59 mm***

Model:

<u>**GraphSAGE**</u> GNN

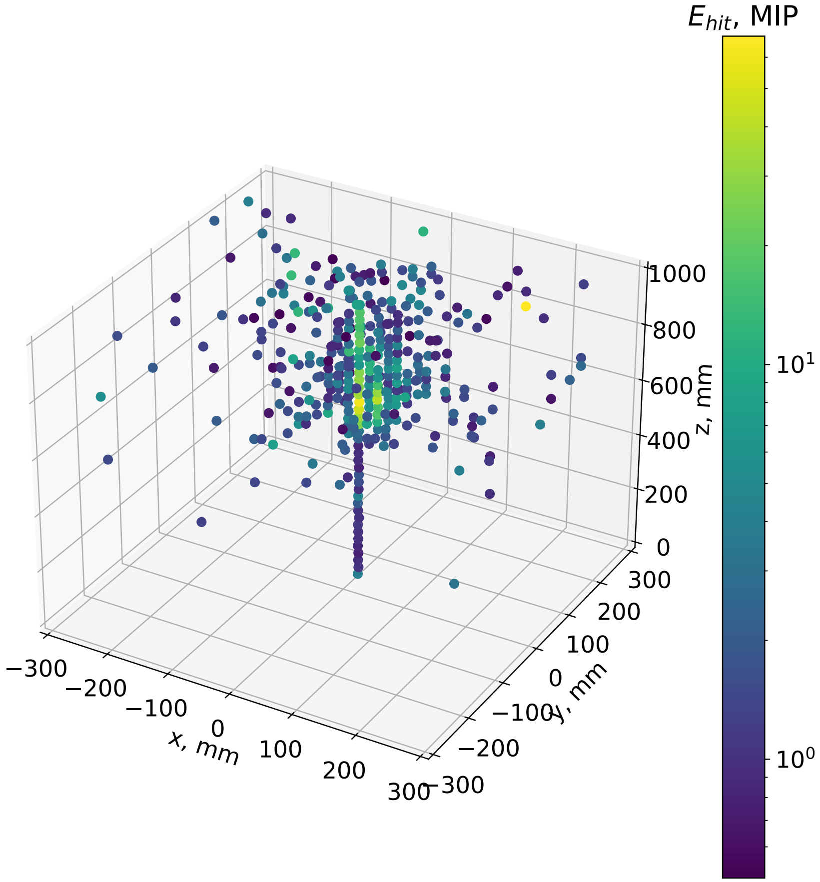
8 layers with 16 hidden channels + 1 linear output layer to convert node embeddings to edge scores

Prediction of edge scores

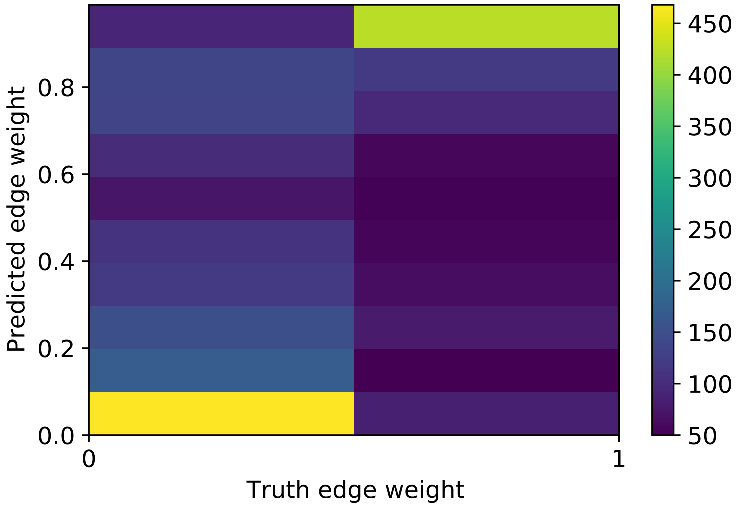
Binary cross entropy loss

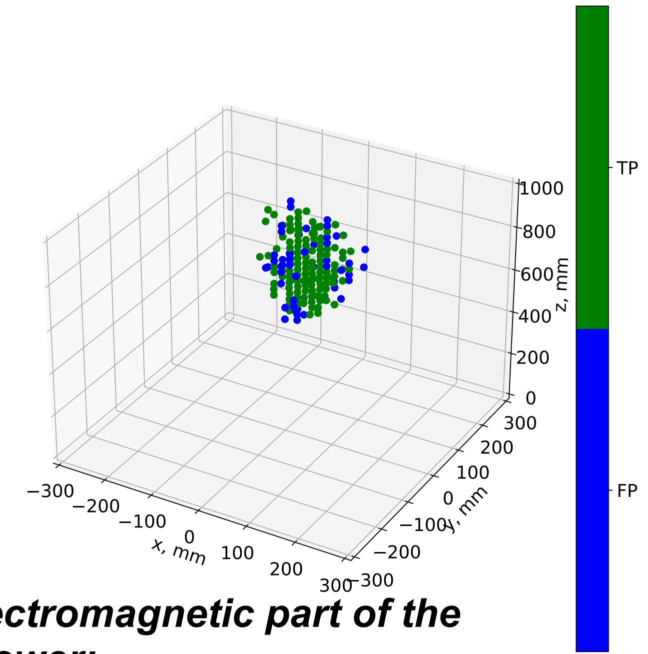
Example of output for test event

Preliminary results for single test event

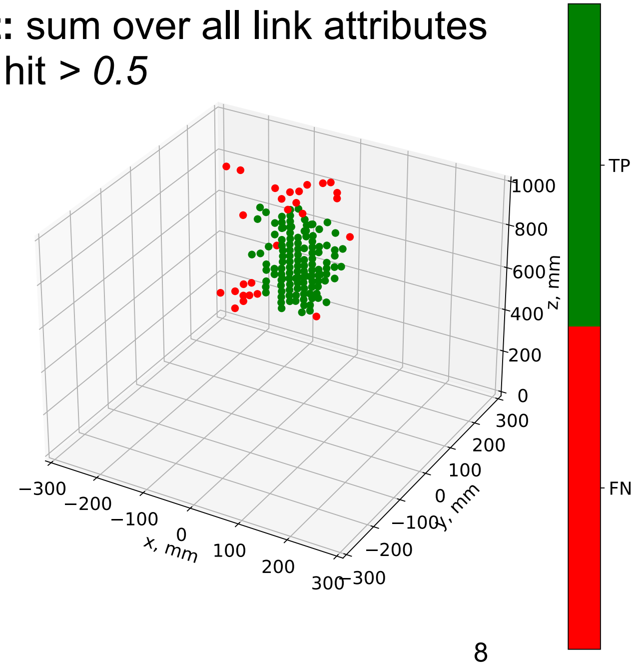


2650 graph edges





Electromagnetic part of the shower:
Cut: sum over all link attributes per hit > *0.5*

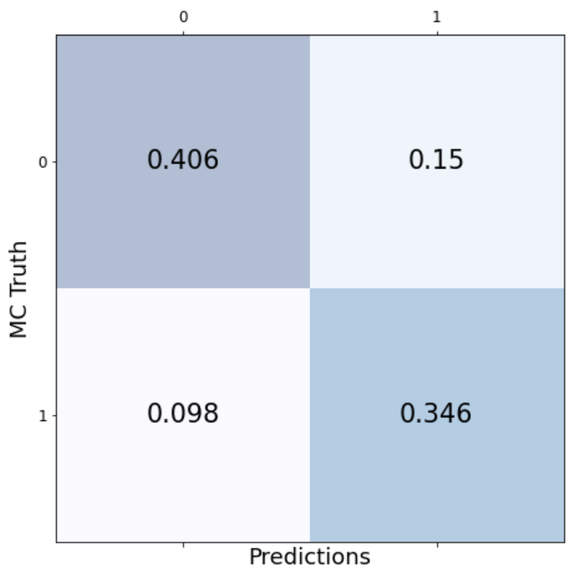


Work in progress…

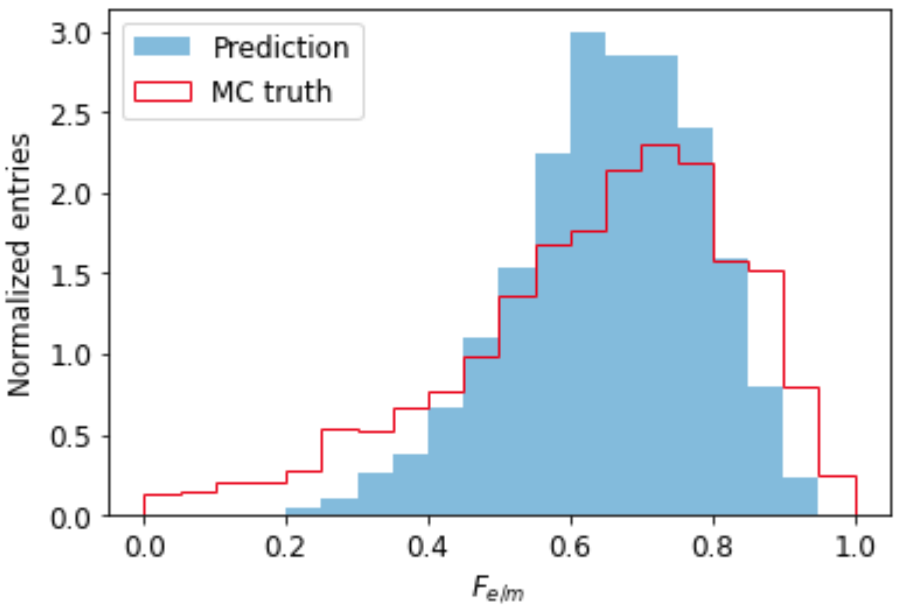
Electromagnetic fraction of hadronic showers

Preliminary results for 10,20,30,40,60,80 GeV pions

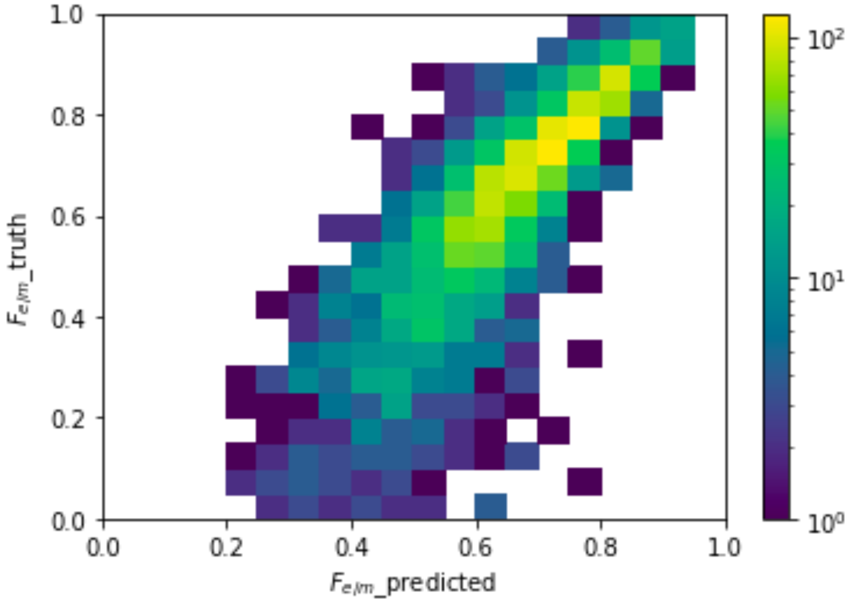
Hit classification



Electromagnetic fraction



Prediction vs truth correlation



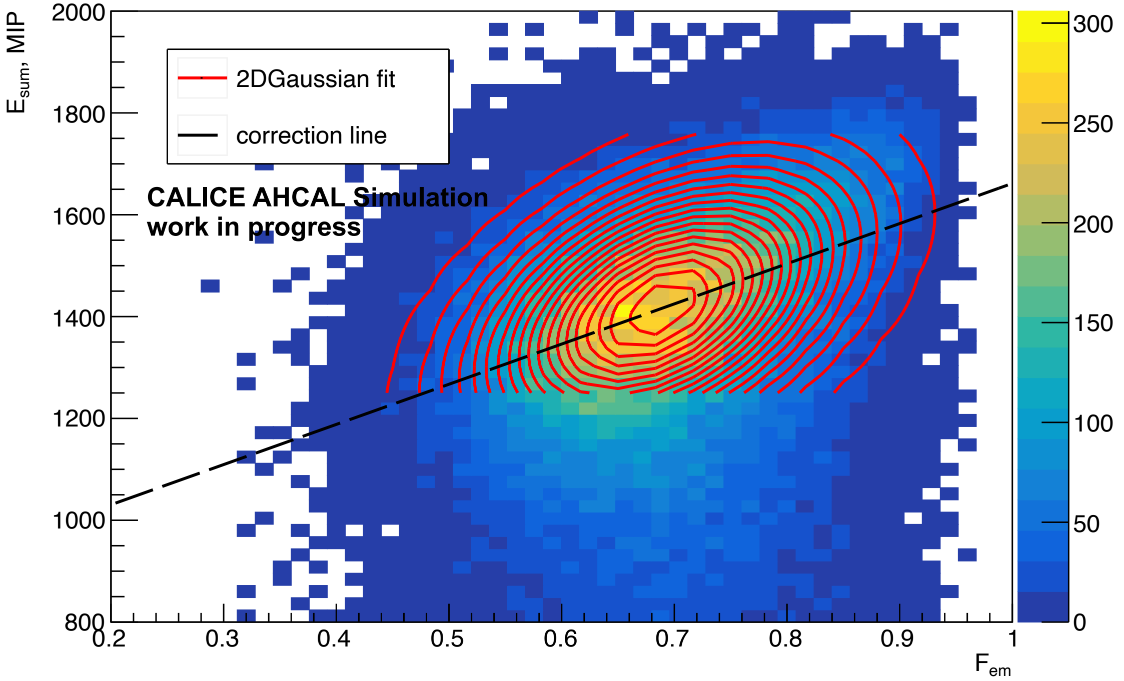
- Higher MPV for F_{em} than expected
  - ➡ Non-e/m contributions to the hits are not taken into account
- Less pronounced tails for F_{em} prediction than for MC truth

W o r k i n p r o g r e s s …

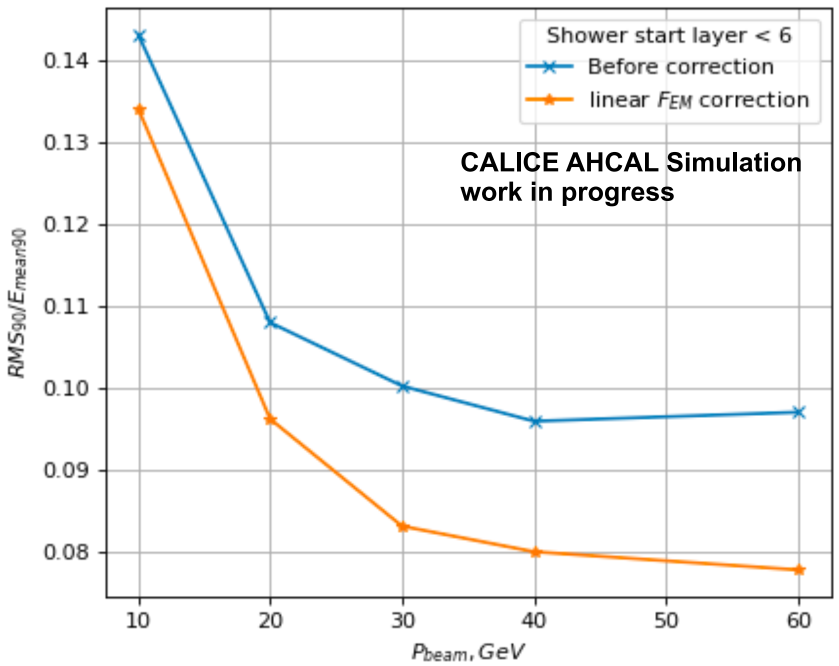
Energy correction

Simple example of using e/m fraction reconstructed by GNN

Correlation example for 40 GeV pion



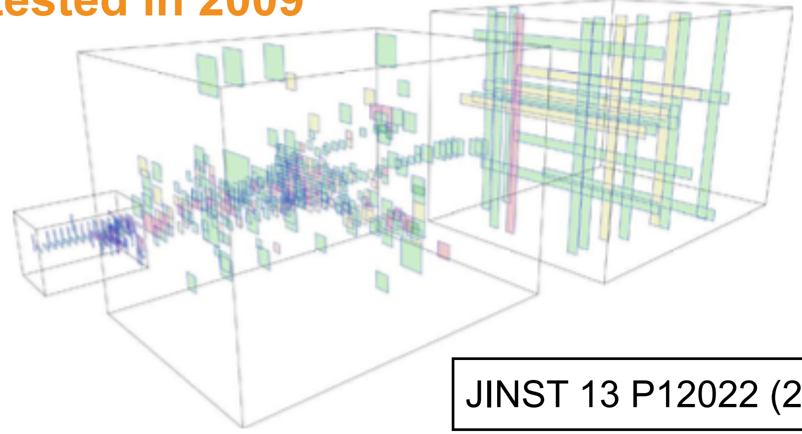
Energy resolution estimation



- Well pronounced correlation between E_{sum} and F_{em} observed for all energies
- For each energy point simultaneous gaussian fit is performed to extract the correction line
- Simple linear correction gives resolution improvement of ~6-20% ➡ to be compared with existing energy reconstruction methods
- Tests on test beam data are ongoing
- Promising resolution improvement for more complex compensation algorithms using reconstructed EM information

Software compensation method

Example for CALICE combined setup ECAL+AHCAL+Tailcatcher tested in 2009

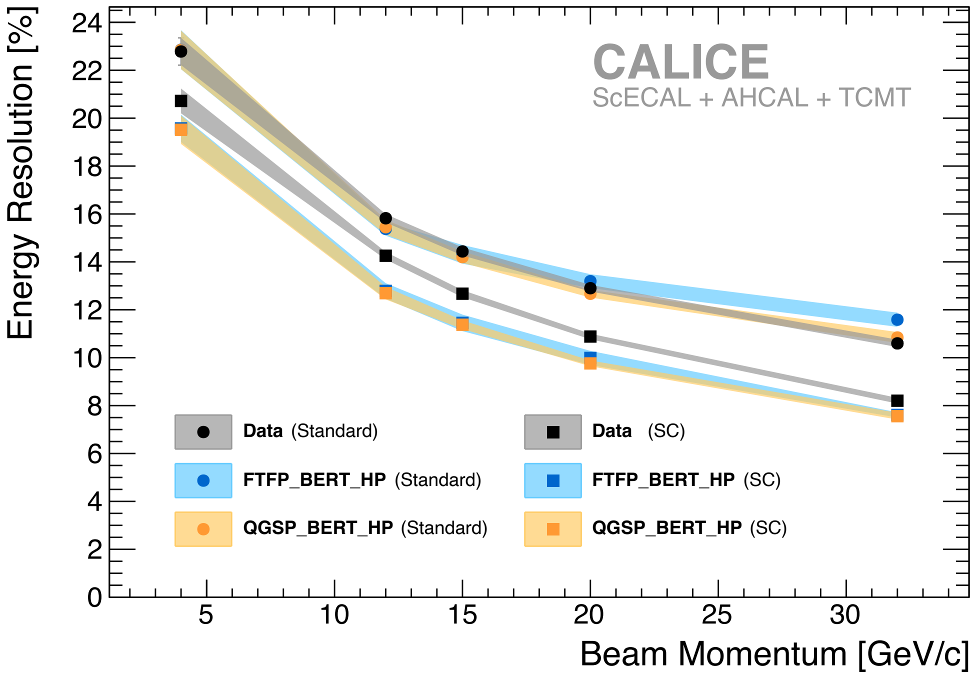


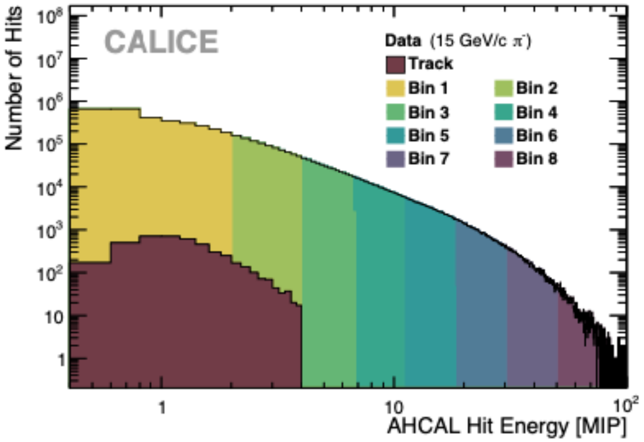
JINST 13 P12022 (2018)

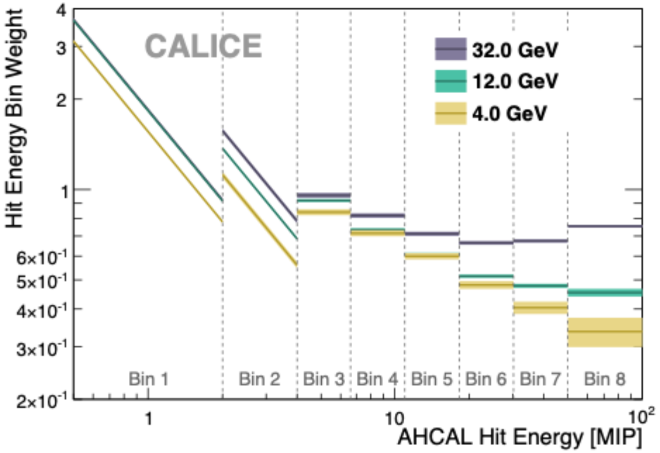
- h/e response compensation by assigning energy-dependent weights to hit energies (⇒local energy density)
    - Higher weights for **low energy hits** - dominated by **HAD** component
    - Lower weights for **high energy hits** - dominated by **EM** component
- 8 bins for hit energies
    - Polynomial fit to get energy dependent weight for each bin

➡ Energy resolution improvement 10-20%

- Disadvantages: limited to fit energy range, polynomial dependence has no physics motivation, additional topological information of hit context is not used







Energy reconstruction using predicted EM information

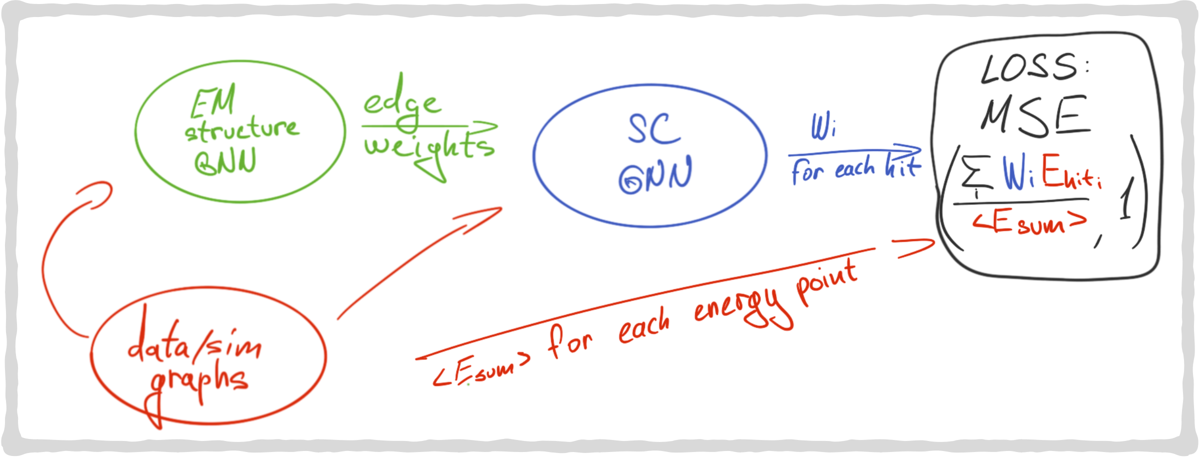
Outlook

Ongoing experiment:

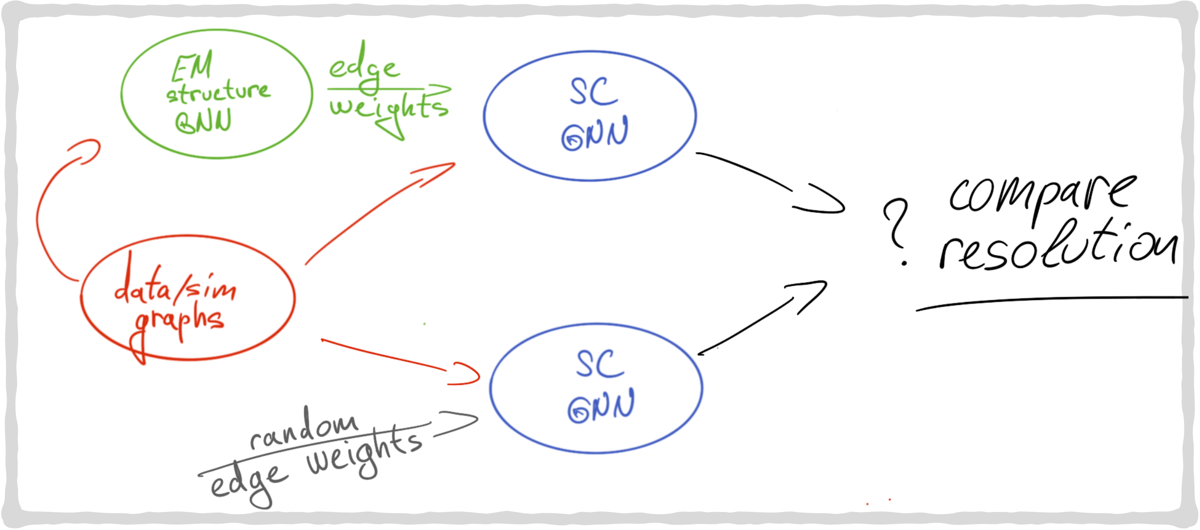
- Test if use of predicted edge weights improves the energy resolution
- Almost same GNN as for EM structure prediction:
  - 1 GraphSAGE layer replaced with ARMAConv (capable to exploit edge attributes during message passing), output has shape [N_{nodes}]
  - Train using predicted EM edge weights
  - Compare resolution for the test sample using predicted EM attributes or random edge weights

Work in progress…

Training:



Experiment:



Towards distinct secondary particle reconstruction

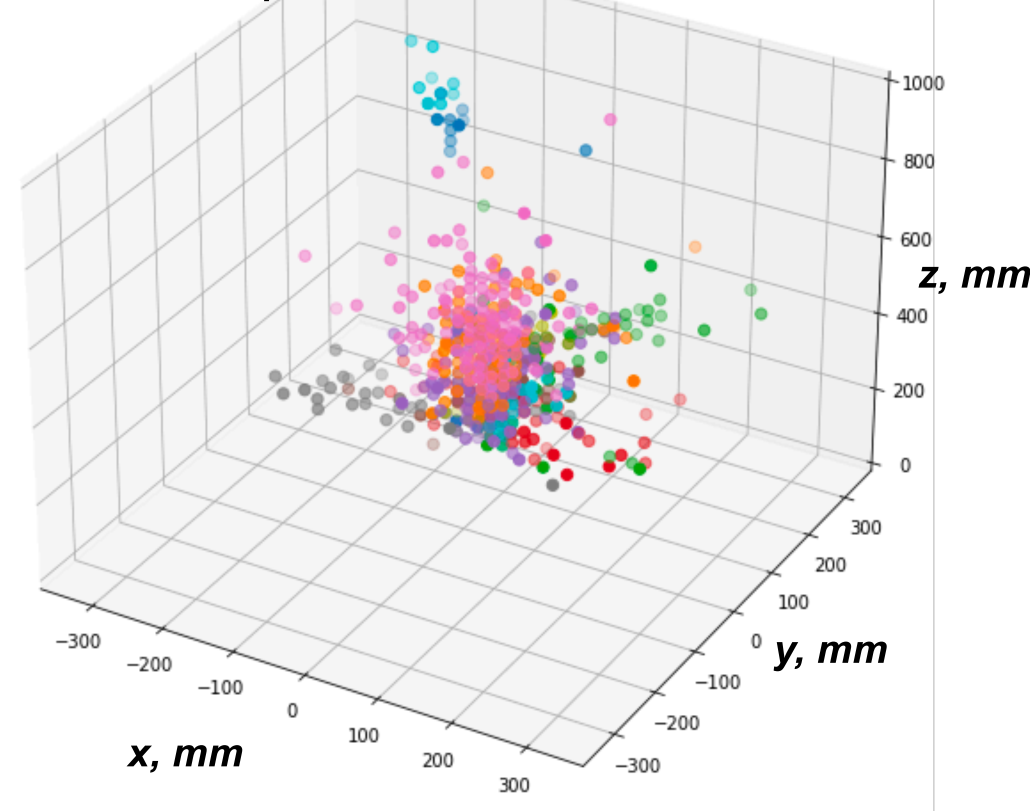
Another outlook

Motivation:

- In HAD showers we can have many EM sub showers at first HAD interaction (overlaid) and later in the had cascade (displaced)
- Further look into the structure of EM fraction:
  - Reconstruct distinct particle components
    - No easy rule-based algorithm to merge overlaid sub showers on MC truth level ➡ go unsupervised!
    - Test Bayesian Gaussian Mixture model with Dirichlet process on point clouds from calorimeter events
      - SKlearn implementation is tested, own flexible Pyro implementation is planned

  ➡ Tune training dataset for substructure GNN
      - e.g. energy thresholds (some EM sub showers have topology closer to ionising tracks)

Event display for MC truth EM fraction of 40 GeV π^-.
Colour code represents different EM sub showers:



Applying Bayesian GM to EM component of had showers

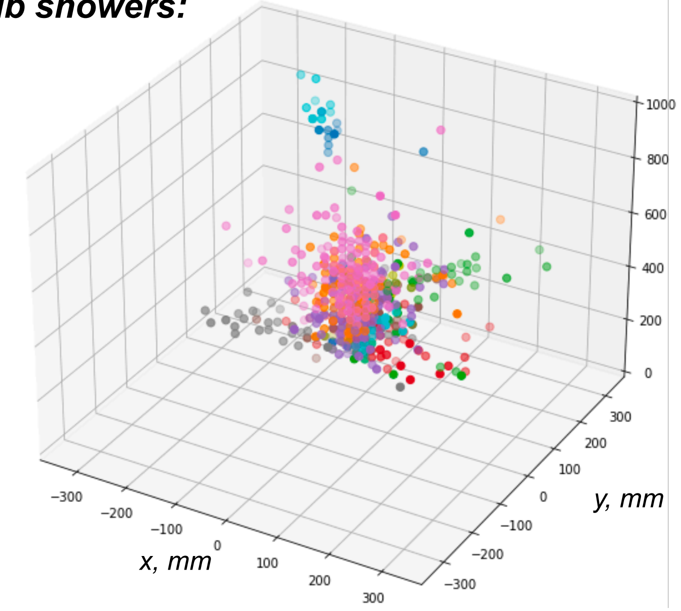
Truth EM component

- SKlearn implementation can handle only scatter plots
- To keep hit energy information, artificial scatter plot is produce:
  - 10 points per MIP
  - uniformly distribute within cell volume: ±15mm,±15mm,±1mm
  - Normalise coordinates: *(-0.36m,0.36m) (-0.36m,0.36m) (0m,1m)*

- Max number of components = 10,
- Object size can be optimised by modifying covariance prior
- Clusters can be filtered by likelihood and energy density

MC truth EM fraction of 40 GeV MC π^-.

Colour code represents different EM sub showers:

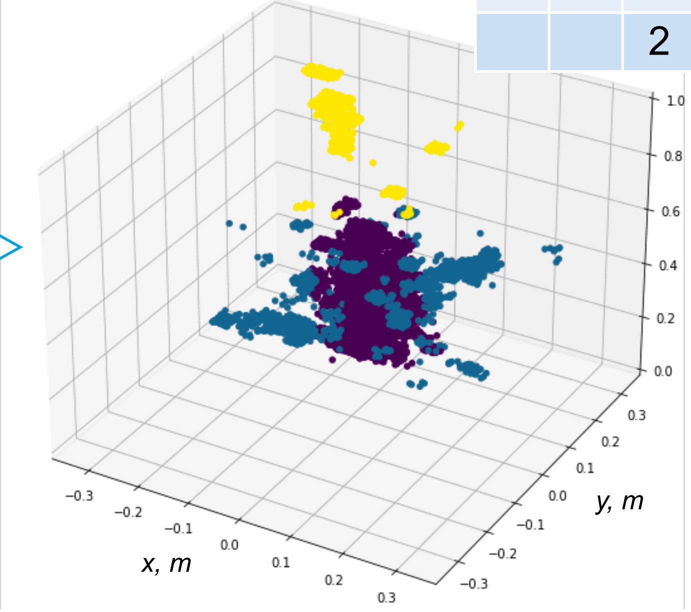


1st BGM iteration:

3 Gaussian components

Cov prior 1st iter

2		
	2	
		2

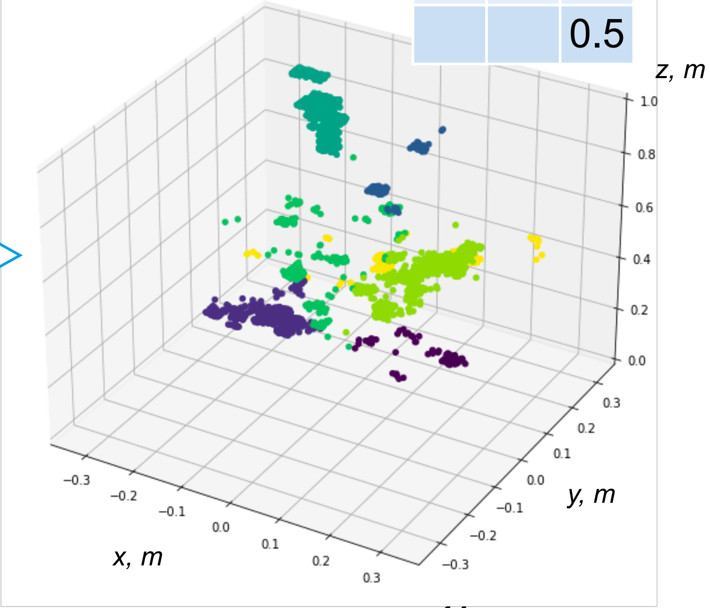


2nd BGM iteration:

7 components

Cov prior 2d iter

0.5		
	0.5	
		0.5



Applying Bayesian GM to EM component of had showers

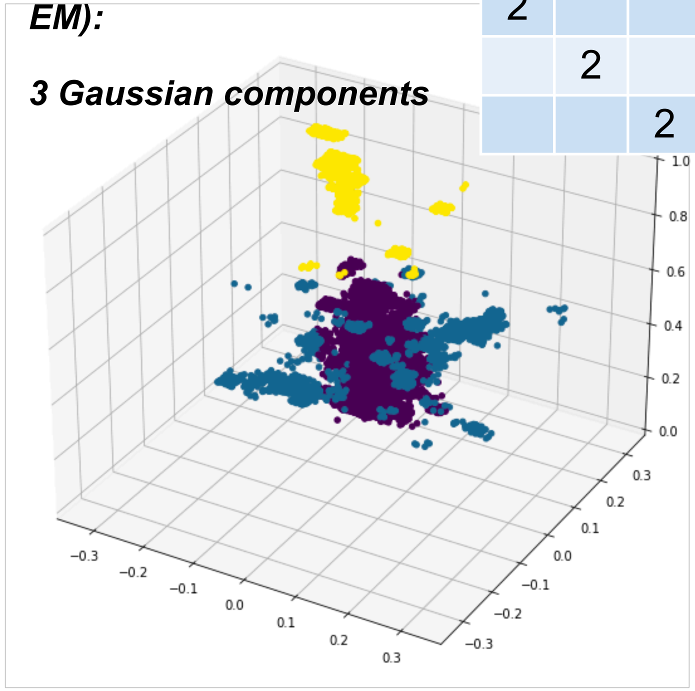
Truth vs reco EM component

1st BGM iteration (truth EM):

3 Gaussian components

Cov prior truth EM

2		
	2	
		2

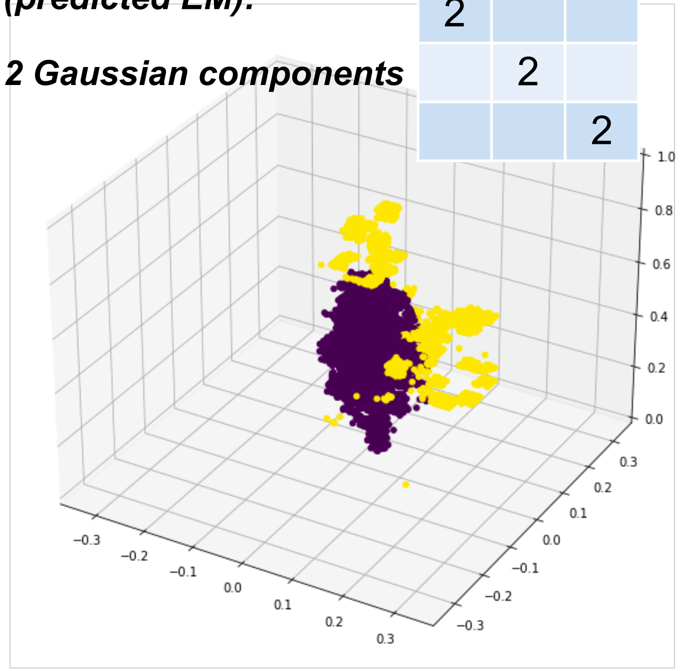


1st BGM iteration (predicted EM):

2 Gaussian components

Cov prior 1st iter

2		
	2	
		2

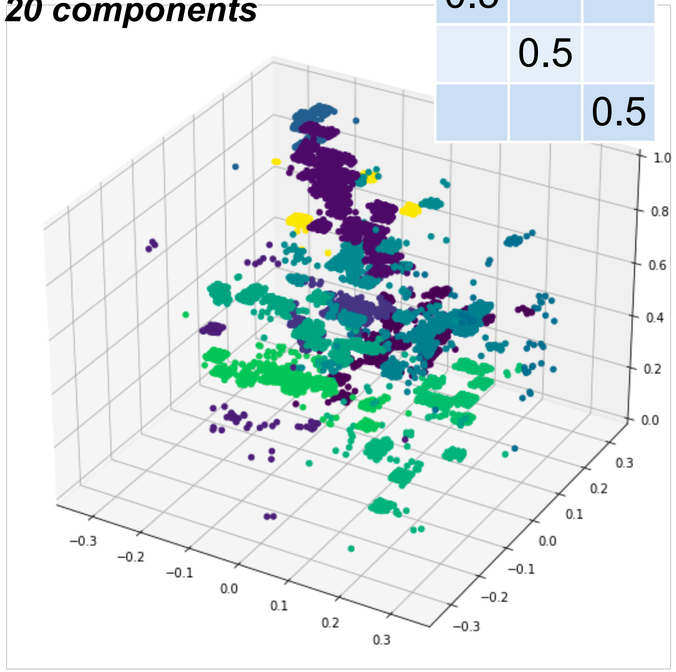


2nd BGM iteration (rest of event) :

20 components

Cov prior 2d iter

0.5		
	0.5	
		0.5

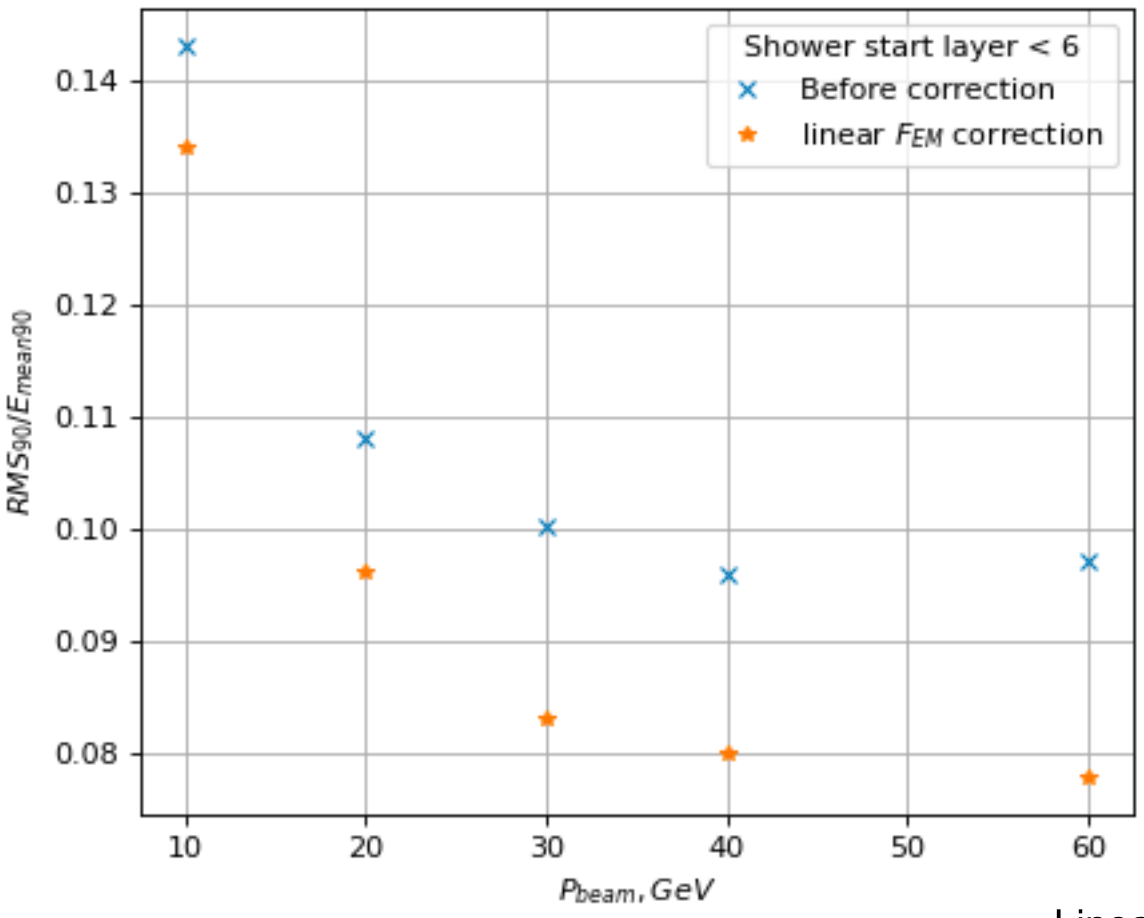


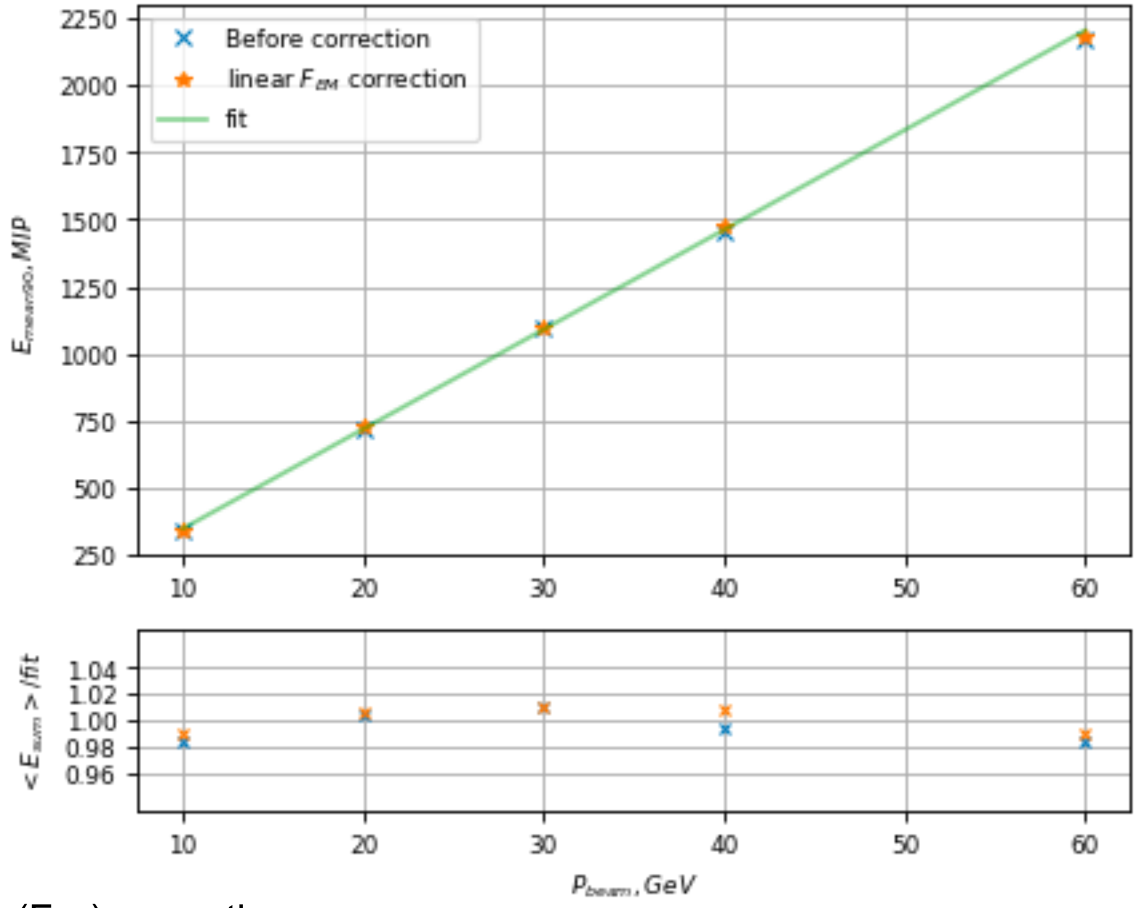
- Visual similarity for main gaussian component
  - Hints of agreement for E_{sum} and $E_{density}$ on several hundred events between truth and predicted EM fraction
- Smaller clusters are more challenging
- ➡ Room for improvement

Summary

- Single hadronic shower substructure can be reconstructed using imaging capabilities of highly granular calorimeters
- GNN reconstruction of electromagnetic components shows promising results
    - Reconstructed EM information can be used to improve hadronic energy resolution
    - EM structure-aware software compensation model is under development
- Prospects of distinct particle reconstruction are discussed

Backup





- Linear $E_{sum}(F_{em})$ correction:

$E_{cor} = C^*E_{sum}$

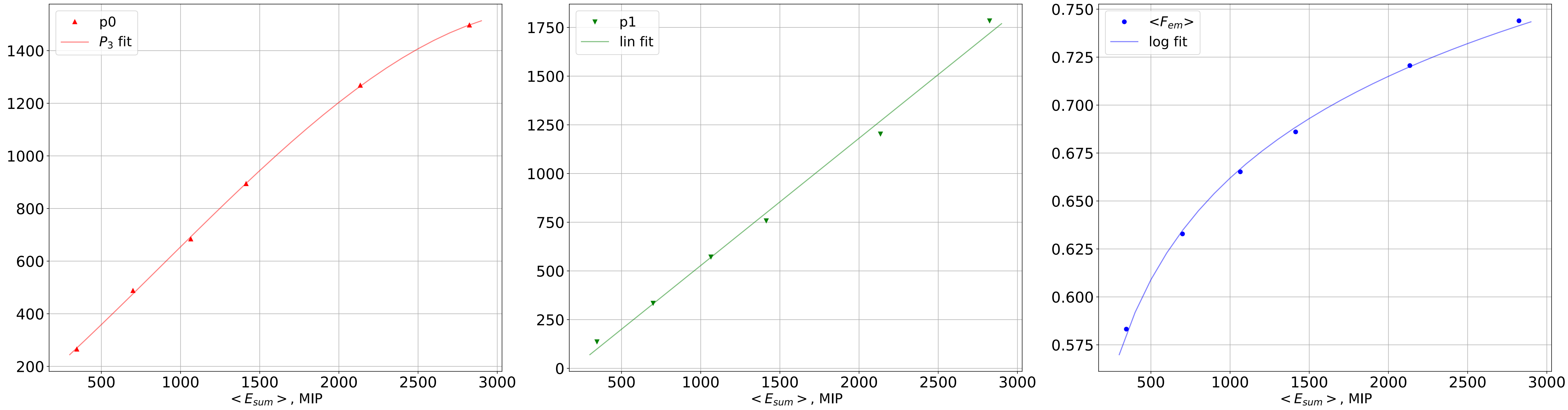
$C = <F_{em}>/(p_1 \cdot F_{em} + p_0)$

Unified correction

Getting P_{beam}-independent correction

Work in progress…

Correction parameters as a function of $<E_{sum}>$:



- p_0,p_1 and $<F_{em}>$ are calculated for each event from the observed energy using resulting fits
  - More energy points need to be included to check the overfitting
  - Parameter uncertainties are not taken into account
  - Performance decrease for resolution ~3%

Dealing with background clusters

- Quality metrics (optimised on several events)
  - likelihood > 2 (first guess)
  - energy density in ellipsoid [MIP/mm^3] > 20 (first guess)