ML for particle reconstruction in the CMS detector

Yuliia Maidannyk

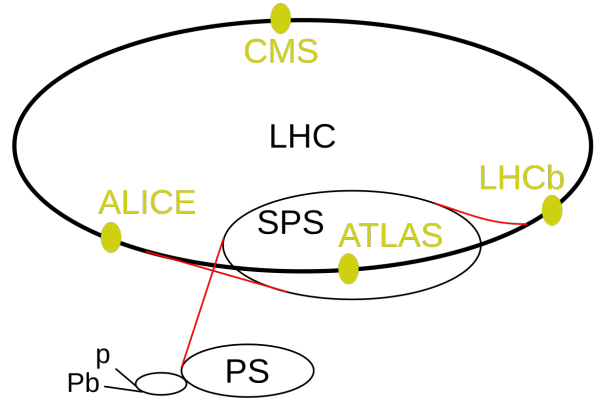
PHENIICS Fest

09 June 2026

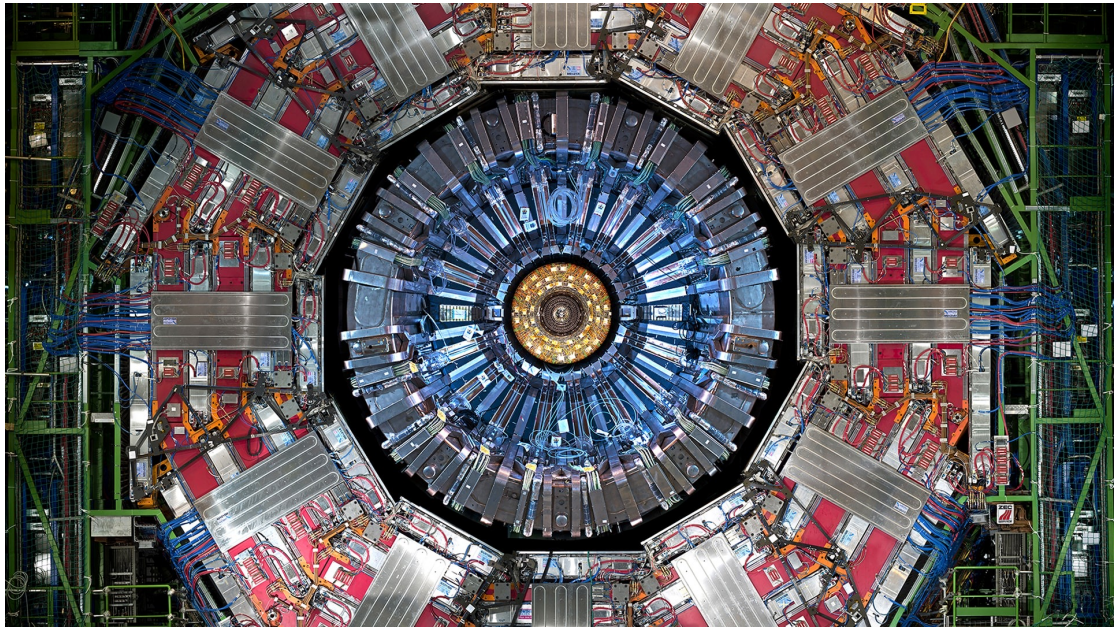
cea irfu

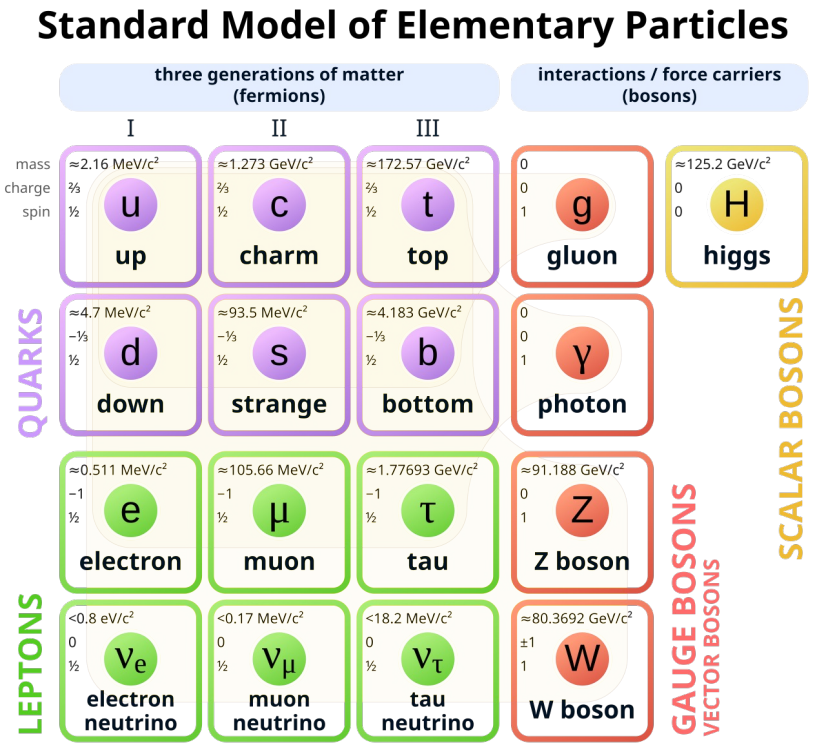
Co-funded by the European Union

The CMS detector



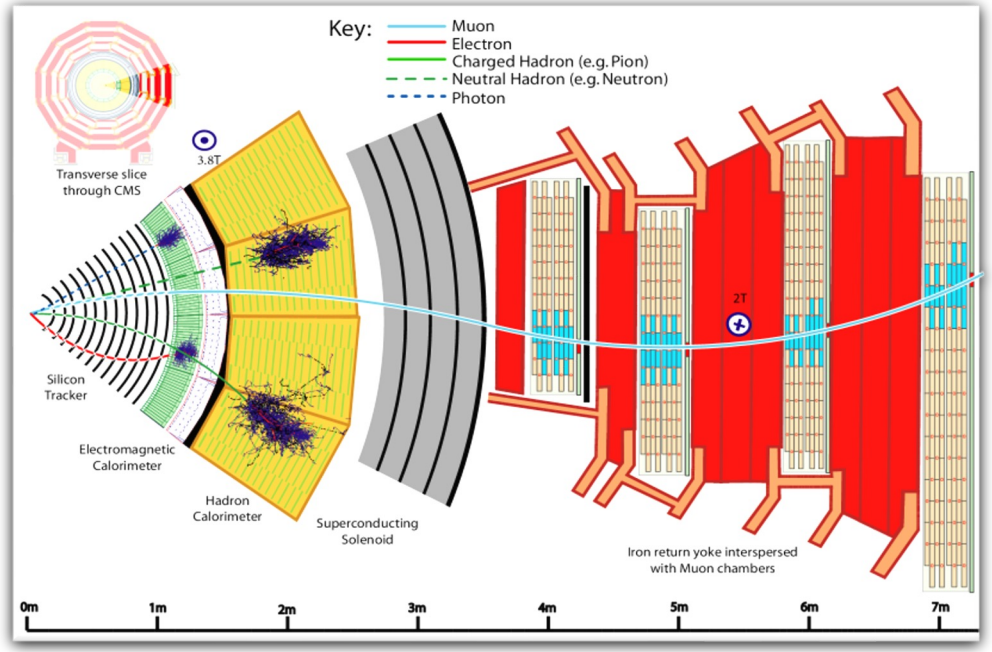
- 1 of 4 experiments at the LHC complex at CERN
- pp collisions at 13.6 TeV at the rate of 40 MHz
- Probing the Standard Model and Beyond the Standard model physics
- Large variety of research areas. My group: a lot of Higgs physics

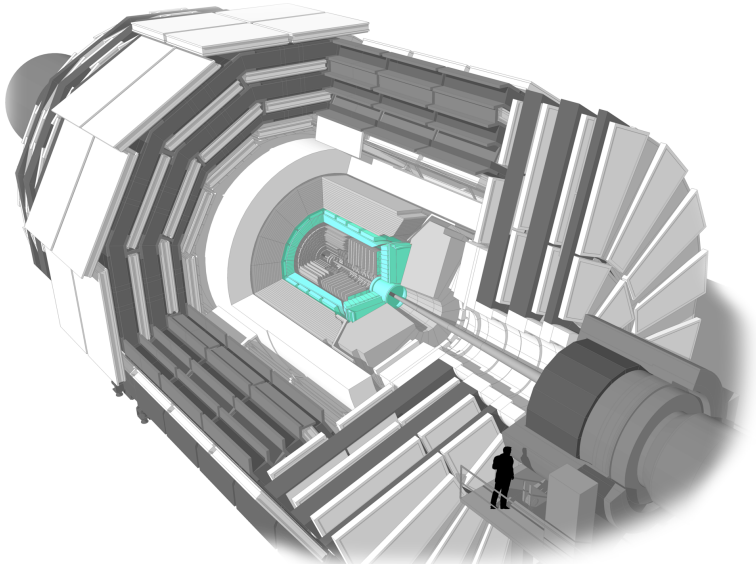




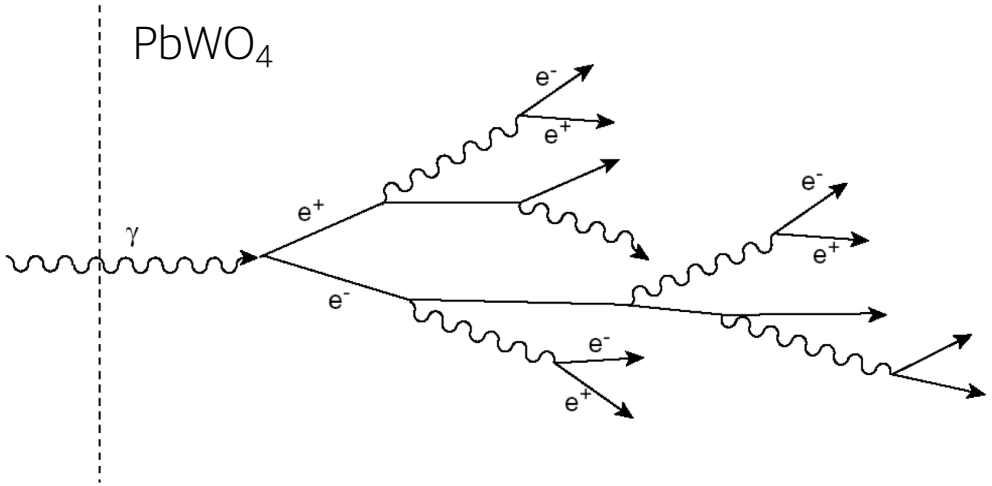
The CMS subdetectors: ECAL

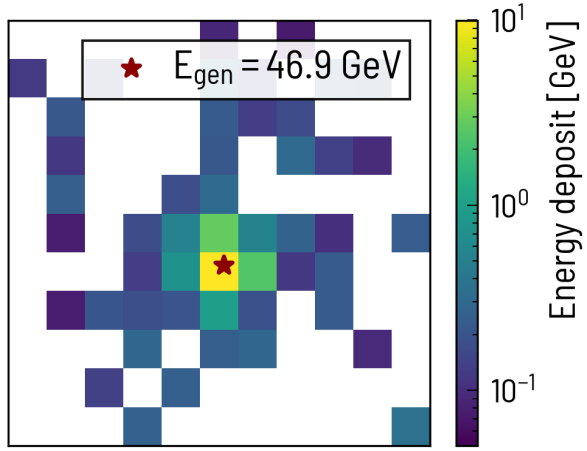
- Layers of subdetectors: tracker, calorimeters, muon chambers
- ECAL (EM calorimeter): **responsible for detecting e/γ**
- Central in many physics analyses
- e/γ create EM showers depositing their energy in **cluster-like formations**





ECAL barrel region
single layer of 61,200 $PbWO_4$ crystals





Calorimeter readout

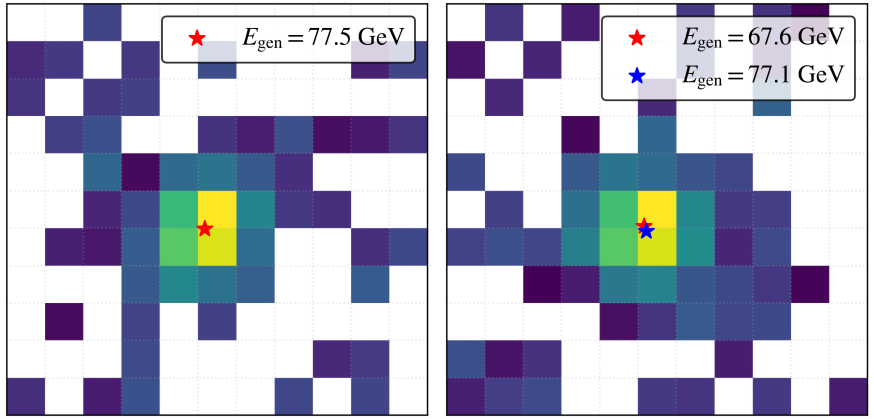
Improving the reconstruction with ML

Aim: reconstruct e/γ energy and momentum from the calorimeter readout (individual crystal E)

Currently using a geometrical algorithm called **PFClustering**

✓ high efficiency on low-E and isolated particles

✗ struggles to resolve overlapping showers



1-photon shower

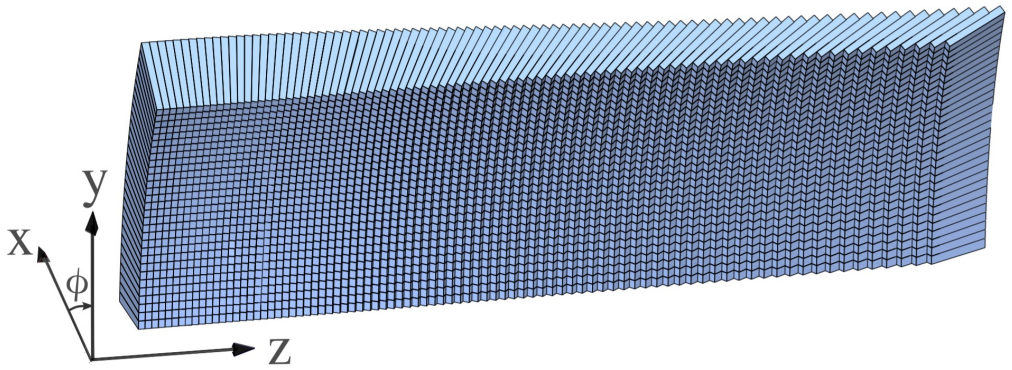
2 overlapping photon showers

Aim of my thesis: develop ML model that improves upon PFClustering

Data generation

Toy calorimeter simulation

Segment of real ECAL (30x85 crystals)



Two datasets for training/validation/testing:

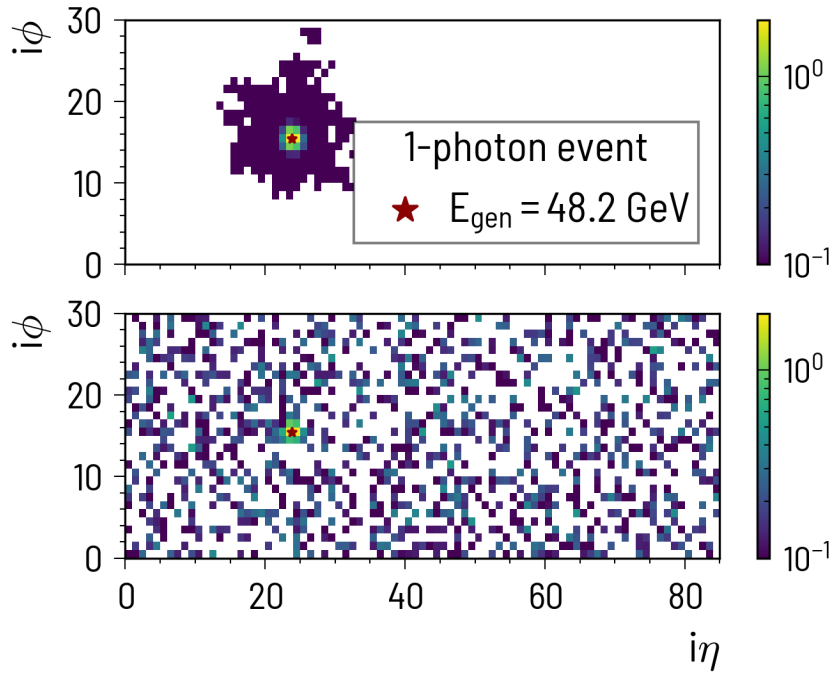
- 1-photon sample with isolated showers in 1-100 GeV range
- 2-photon sample with overlapping showers in 1-100 GeV range

Record energy deposit in each crystal and photon information

$$(E_{gen}, x_{gen}, y_{gen})$$

Add a Gaussian noise that mimics readout electronics

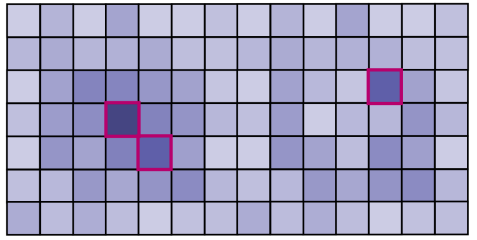
$$\sigma^2 = (0.03)^2 E_{xtal} + (0.0035)^2 (E_{xtal})^2 + (0.167)^2$$



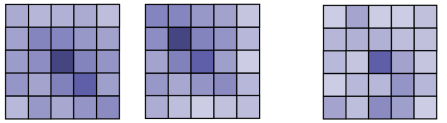
2-step approach

The network should be scalable to arbitrarily large detectors

Showers are localised so we can create 7×7 seed windows around each crystal > 0.5 GeV



Identify E_{xtal} > 0.5 GeV



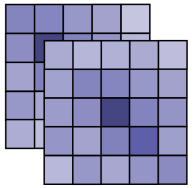
7x7 candidate seed windows

Each window is a candidate that may contain a true cluster. Process seed windows in two steps

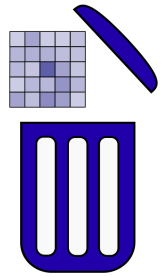
1. Classifier

CNN-based classifier between signal and background

Neighbouring windows show the same cluster: message passing is a network feature that ensures information exchange between them



Selected windows of the same event are processed together



Filter background with SeedFinder

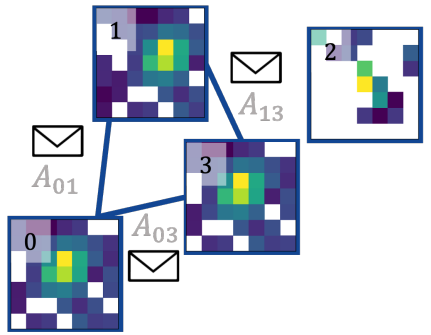
2. Regressor

Regresses $(E_{\text{reco}}, x_{\text{reco}}, y_{\text{reco}})$ from selected windows using **message passing** with either a…

- **GNN:** adjacency matrix of Euclidean distances between seed windows

or

- **Transformer (GAT):** attention mechanism



1-step approach: ClusTEX

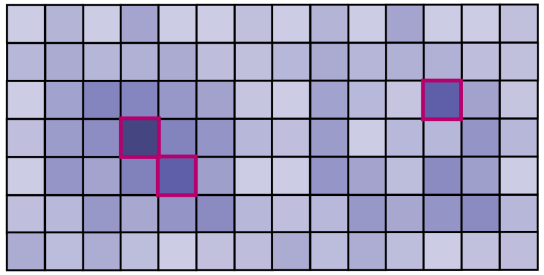
Classification + regression in a single network

Transformers are a powerful tool: remove the classifier altogether

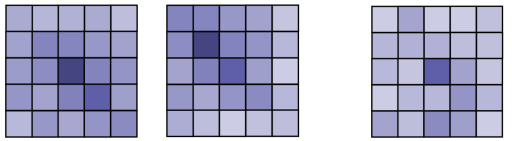
Same starting point: 7x7 seed windows around each crystal > 0.5 GeV

Use the fact that each detector region is independent and form local neighbourhoods of candidates

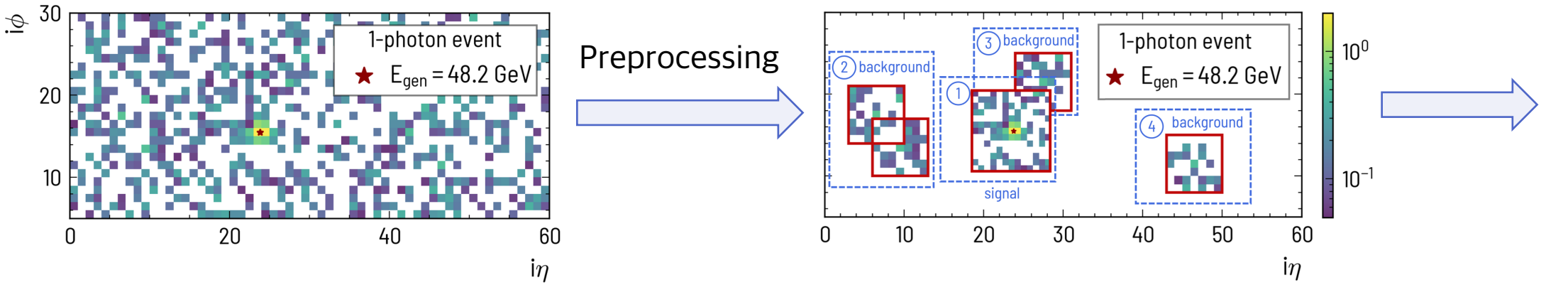
Each local neighbourhood becomes a separate input to the network



Identify E_{xtal} > 0.5 GeV



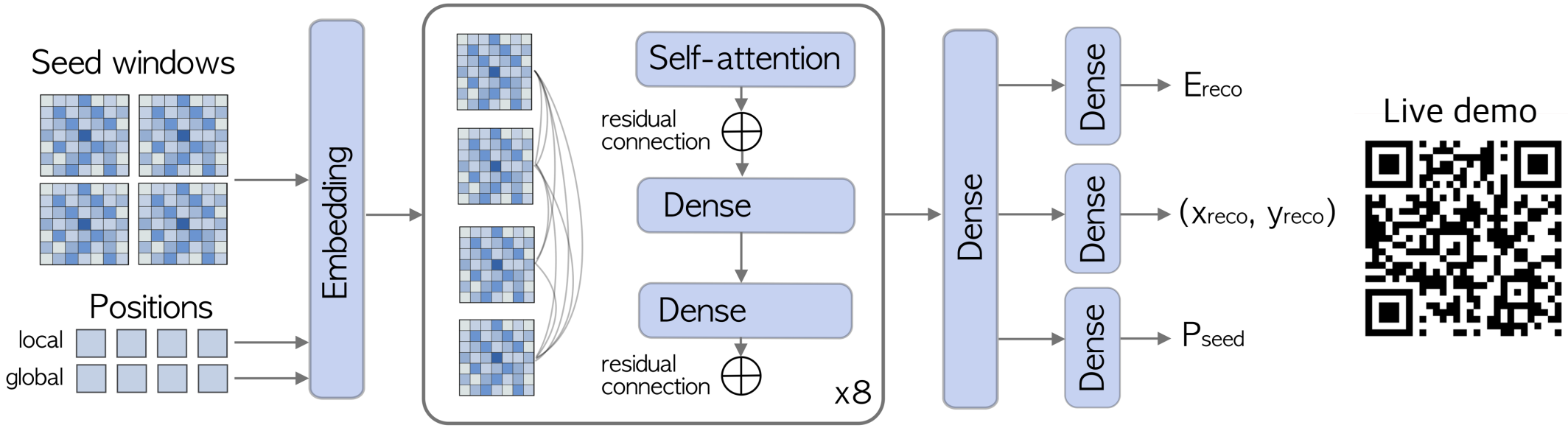
7x7 candidate seed windows



ClusTEX architecture

arXiv (accepted to EPJC)

Transformer that takes groups of calorimeter windows as input and predicts the energy and position of the particle



7x7 windows of detector hits + additional information

Input features

encoding / message passing / decoding

Primary particle $(E_{reco}, x_{reco}, y_{reco})$ + signal/background discrimination

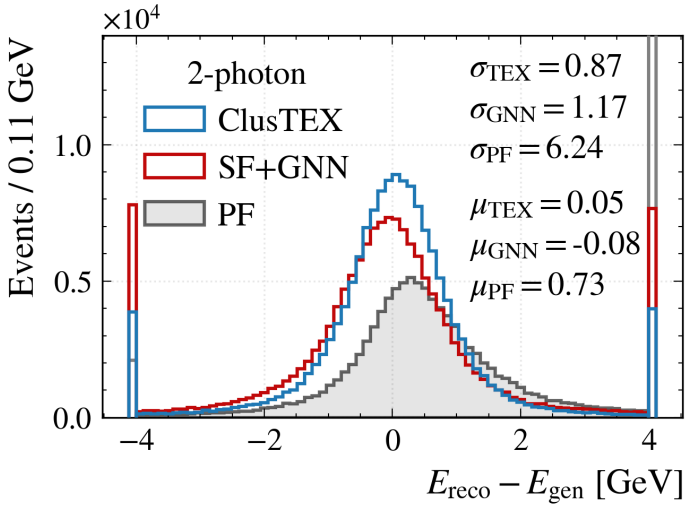
Output

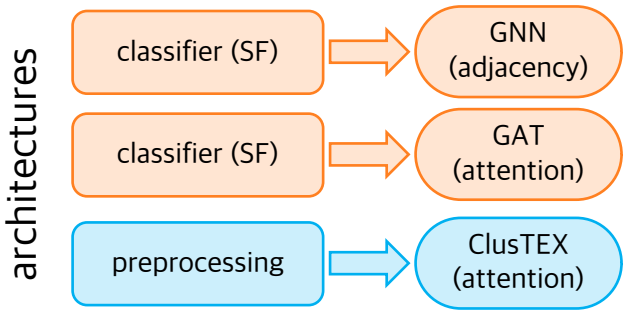
Results for toy calorimeter

Single-step transformer wins all-around

- Outperforms SF+GAT, SF+GNN and PFClustering on σ_E
- Improves σ_x over the PFClustering baseline
- Improves background rejection compared to PFClustering
- Improves 2-photon signal efficiency

All while removing a CNN dedicated to candidate classification





	1-photon sample			
Metric	ClusTEX	SF+GAT	SF+GNN	PFClustering
Efficiency (%)	99.7	99.8	99.8	99.8
σ_E (GeV)	0.55	0.58	0.57	0.59
$\sigma_{i\phi}$ (crystal)	0.03	0.03	0.03	0.05
$\sigma_{i\eta}$ (crystal)	0.04	0.04	0.04	0.06
Splitting (%)	0.08	0.17	0.30	–
N_{bkg}/100k samples	73	26	102	312k

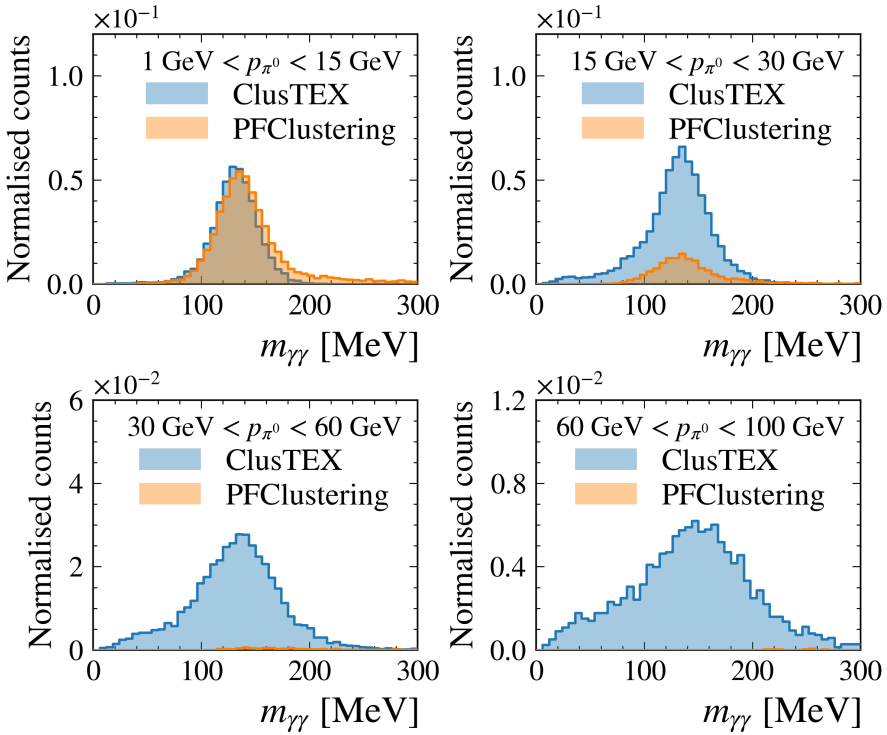
	2-photon sample			
Metric	ClusTEX	SF+GAT	SF+GNN	PFClustering
Efficiency (%)	98.7	98.3	98.3	82.0
σ_E (GeV)	0.87	1.10	1.17	6.24
$\sigma_{i\phi}$ (crystal)	0.04	0.04	0.04	0.09
$\sigma_{i\eta}$ (crystal)	0.04	0.04	0.04	0.09
Splitting (%)	0.06	0.31	0.59	–
N_{bkg}/100k samples	99	84	178	311k

1-step | 2-step attention/GNN | Baseline

π^0 reconstruction

100,000 events of $\pi^0 \to \gamma\gamma$ decay simulated in toy calorimeter and used for evaluation only
π^0 distributed uniformly in energy range 1-100 GeV
Final-state γ become more collimated with increasing $|\mathbf{p}_{\pi^0}|$

Check invariant mass reconstruction $m_{\gamma\gamma} = \sqrt{2E_1^{\text{reco}}E_2^{\text{reco}}(1-\cos\theta_{12})}$ in four $|\mathbf{p}_{\pi^0}|$ bins



ClusTEX significantly improves the reconstruction efficiency for $p_{\pi^0} > 15$ GeV

ClusTEX is now able to reconstruct pions with $p_{\pi^0} > 30$ GeV

Inline with the improved performance for overlapping showers

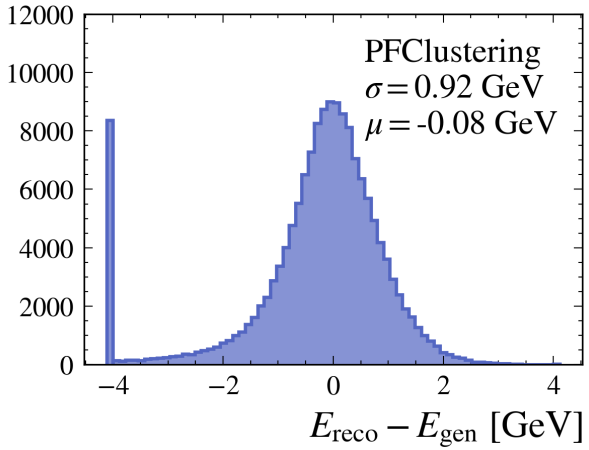
First results for the full calorimeter simulation

Official CMS software is built on a full Monte Carlo simulation of the detector

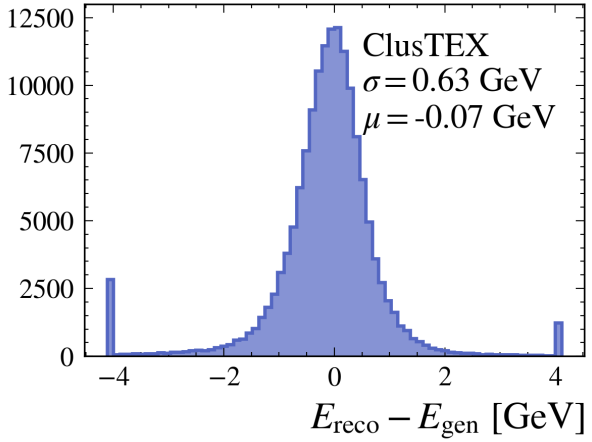
Material in front of ECAL, B-field, non-responsive crystals, irradiation damage, electronic readout problems…

We test our approach on a sample of single unconverted photons in 1-100 GeV range

Traditional approach



Our approach



Metric	1-photon sample	
	ClusTEX	PFClustering
Efficiency (%)	99.7	99.8
σ_E (GeV)	0.55	0.59

PFClustering σ_E is much worse than in the toy simulation, ClusTEX almost recovers nominal performance

Summary

Motivation

Boost the performance of e/γ reconstruction for complex topologies (overlapping showers, pileup, and detector effects)

Approach

Architectures based on GNNs and attention

Results

- ML models significantly improve performance for complex topologies
- Attention mechanism has better interaction encoding capabilities than GNNs
- ClusTEX outperforms previous methods in the toy calorimeter
- Promising results in full simulation

Outlook

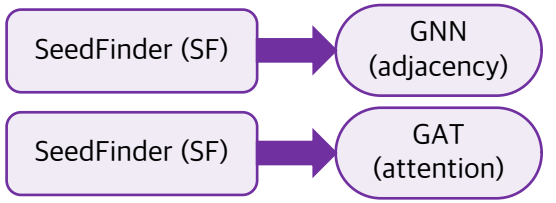
Integration of transformer-based clustering in the full reconstruction chain of CMS

Resources

- Paper uploaded to arXiv, accepted to EPJC
- Simulation dataset publicly available on Zenodo

BACKUP

Results for toy calorimeter

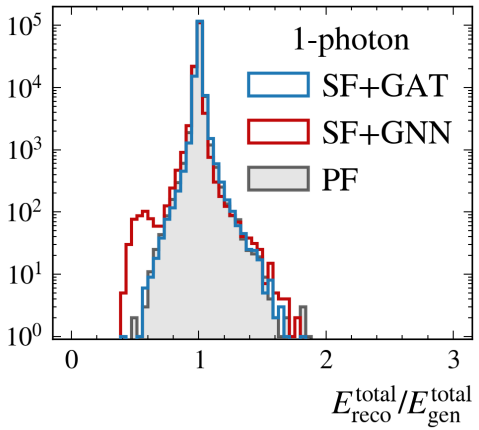


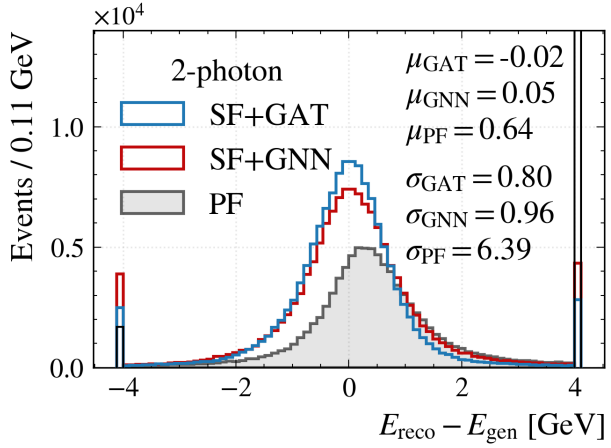
Toy calorimeter: the ML approaches improve upon PFClustering in every metric for the 2-photon sample while preserving the performance on isolated photons

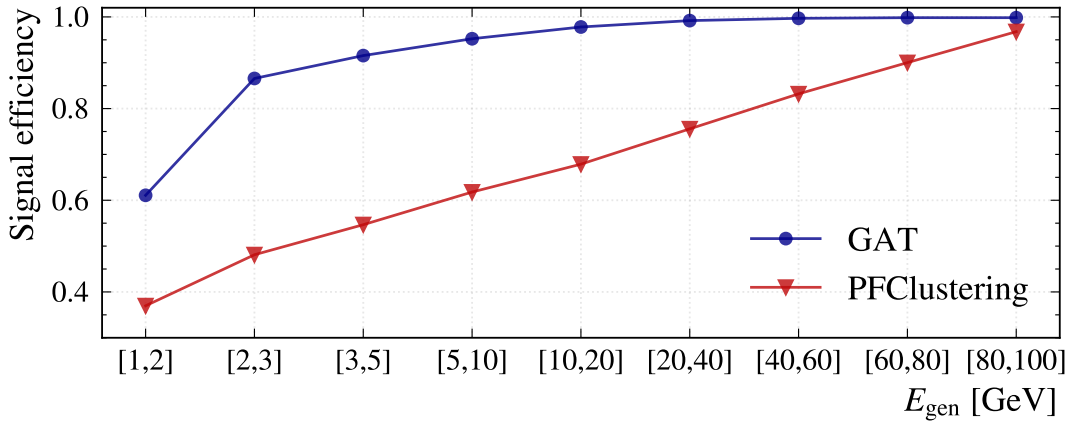
Transformer-based clustering wins all around

Metric	1-photon sample SF+GAT	SF+GNN	PFClustering
Efficiency (%)	99.7	99.7	99.8
σ_E (GeV)	0.53	0.54	0.61
σ_x (crystal)	0.02	0.02	0.04
Splitting (%)	0.05	0.56	–
$N_{\text{background}}$/100k samples	31	96	323k

Metric	2-photon sample SF+GAT	SF+GNN	PFClustering
Efficiency (%)	98.6	98.6	82.0
σ_E (GeV)	0.80	0.95	6.39
σ_x (crystal)	0.03	0.03	0.09
Splitting (%)	0.10	0.33	–
$N_{\text{background}}$/100k samples	35	106	323k



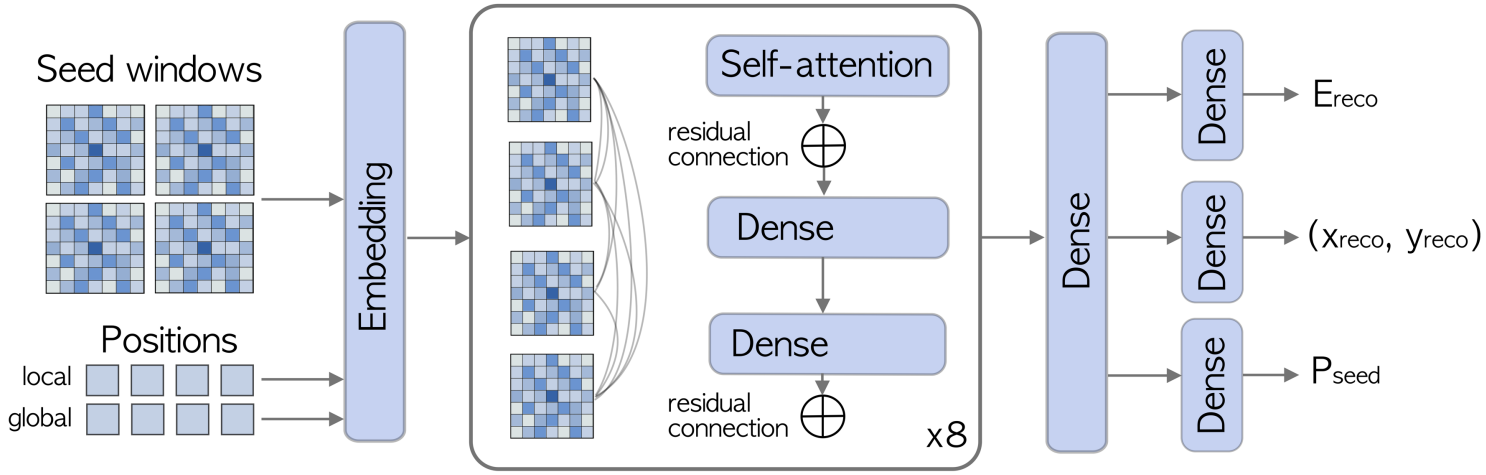




ClusTEX architecture

Real ECAL: non-trivial geometry, η-dependent irradiation, dead cells, …

Key ingredient: geometrical awareness injected with a positional encoding block



Each input seed window i is characterised by its

- Local position with respect to other windows in the graph
- Global position within ECAL (flattened)

$$\mathbf{x}^i_{\text{local}} = \left(\frac{x_i - x_a}{d_{\text{effective}}}, \frac{y_i - y_a}{d_{\text{effective}}}\right) \qquad x^i_{\text{global}} = x_i + y_i \cdot S_{\text{detector}}$$

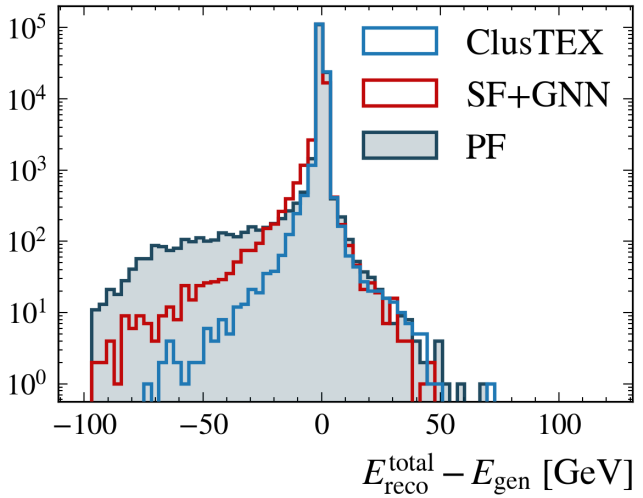
$d_{\text{effective}} = 13$: effective size of the local graph

$S_{\text{detector}} = 85$: # of crystals in $i\eta$

- Embeddings of $\mathbf{x}^i_{\text{local}}$ are concatenated with input seed window embeddings $\boldsymbol{h}_i$
- Embeddings of x^i_{global} are injected with a summation

$$\tilde{\boldsymbol{h}}_i = \boldsymbol{h}_i \mid \boldsymbol{f}\left(\mathbf{x}^i_{\text{local}}\right) \qquad \boldsymbol{z}_i = \tilde{\boldsymbol{h}}_i + \boldsymbol{g}\left(x^i_{\text{global}}\right)$$

1% dead crystals



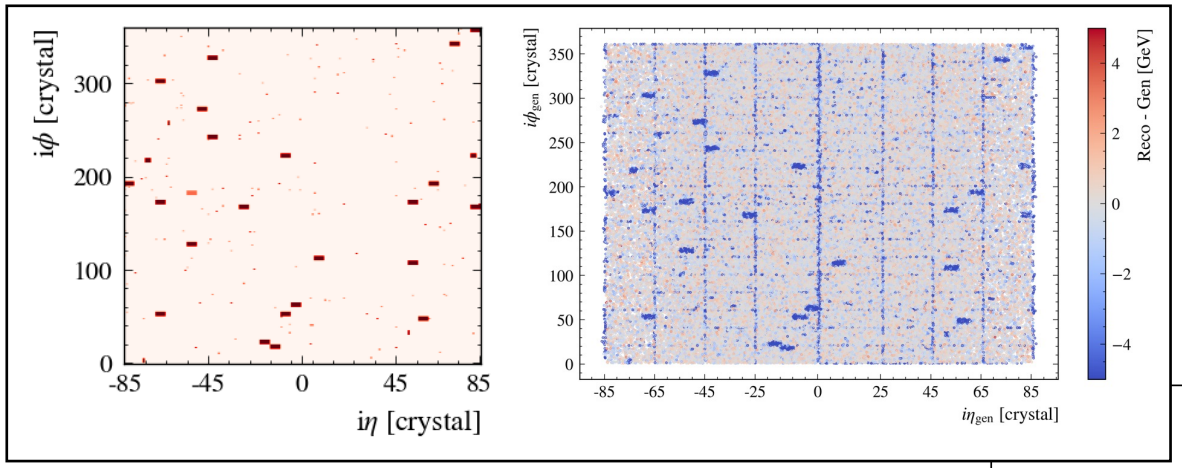
CMSSW developments: preliminary studies

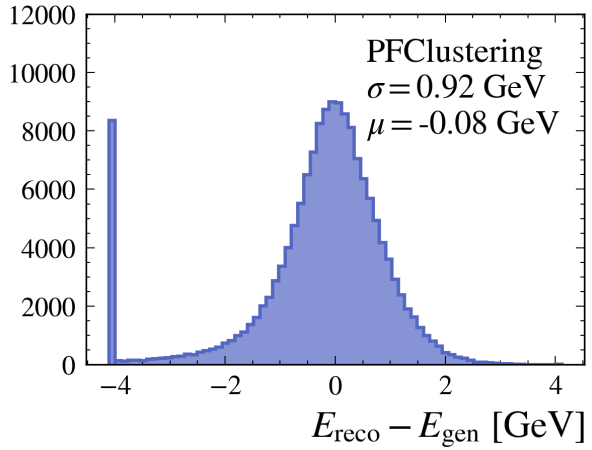
With full CMS Run3 MC obtained a sample of single unconverted photons in 1-100 GeV range in the barrel

Good test of accuracy of cluster energy / position

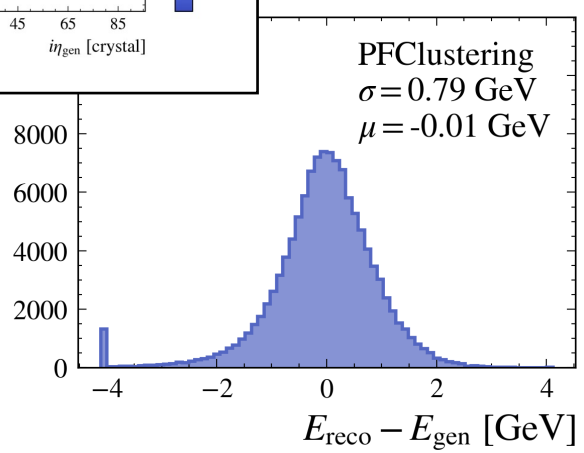
✓ Establish the PFClustering baseline

Underflow in $E_{\text{reco}} - E_{\text{gen}}$ inside modular gaps and dead trigger towers





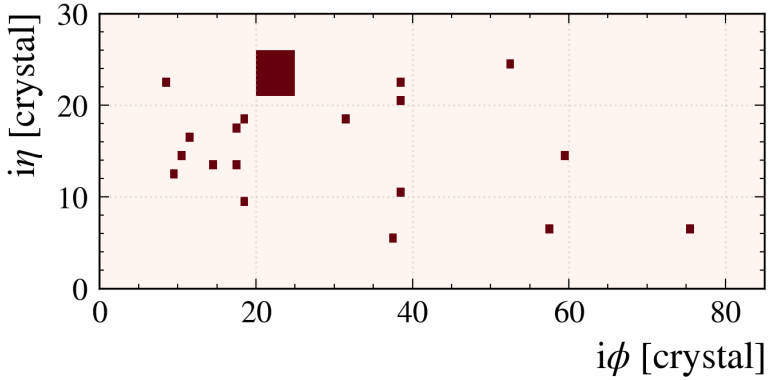
Removing events in module gaps [0, ±25, ±45, ±65, ±85] and dead trigger towers

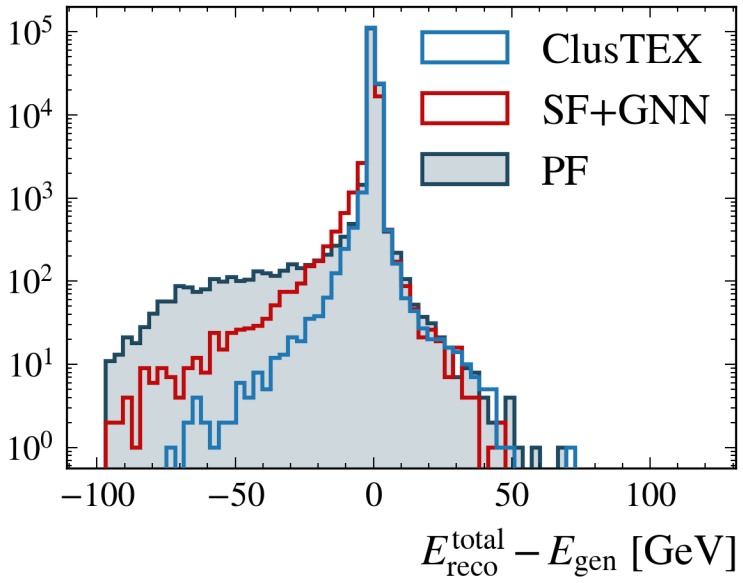


Test of the concept: behaviour in the presence of non-responsive crystals

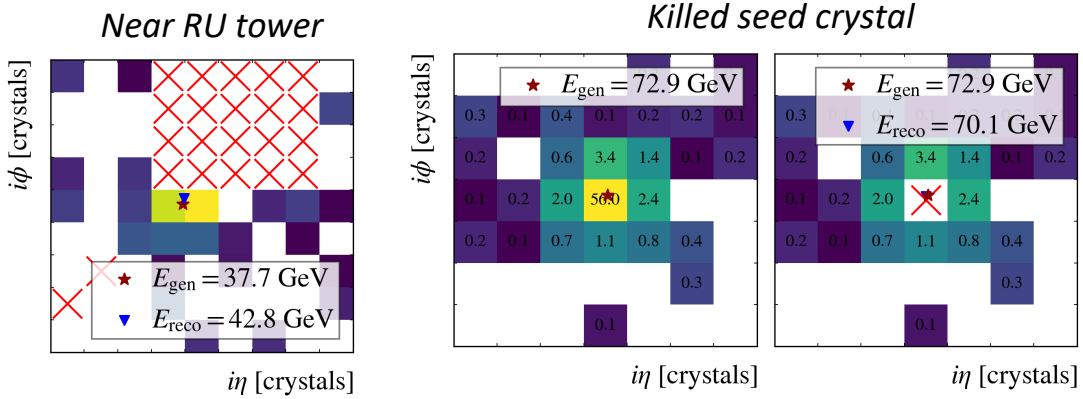
Introduce non-responsive regions

Kill ~1% of individual cells and one 5x5 readout tower





ClusTEX recovers the lost generated energy better than other approaches



Algorithm 1 Clustering graph formation

Parameters tuned on simulation:

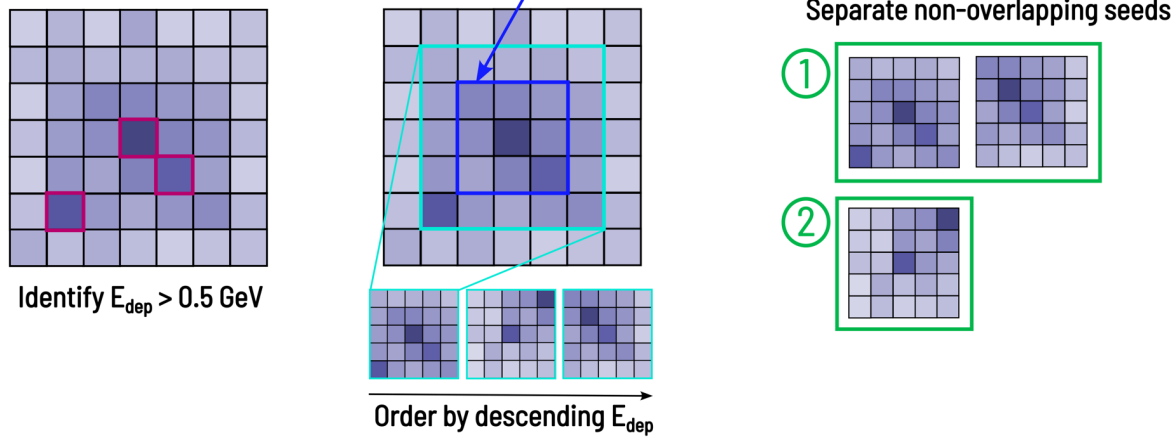
$E_{\text{dep}}^{\text{thresh}} = 0.5\,\text{GeV}$ $\quad d_{\text{overlap}} = 5$ $\quad d_{\text{window}} = 7$

Initialisation:

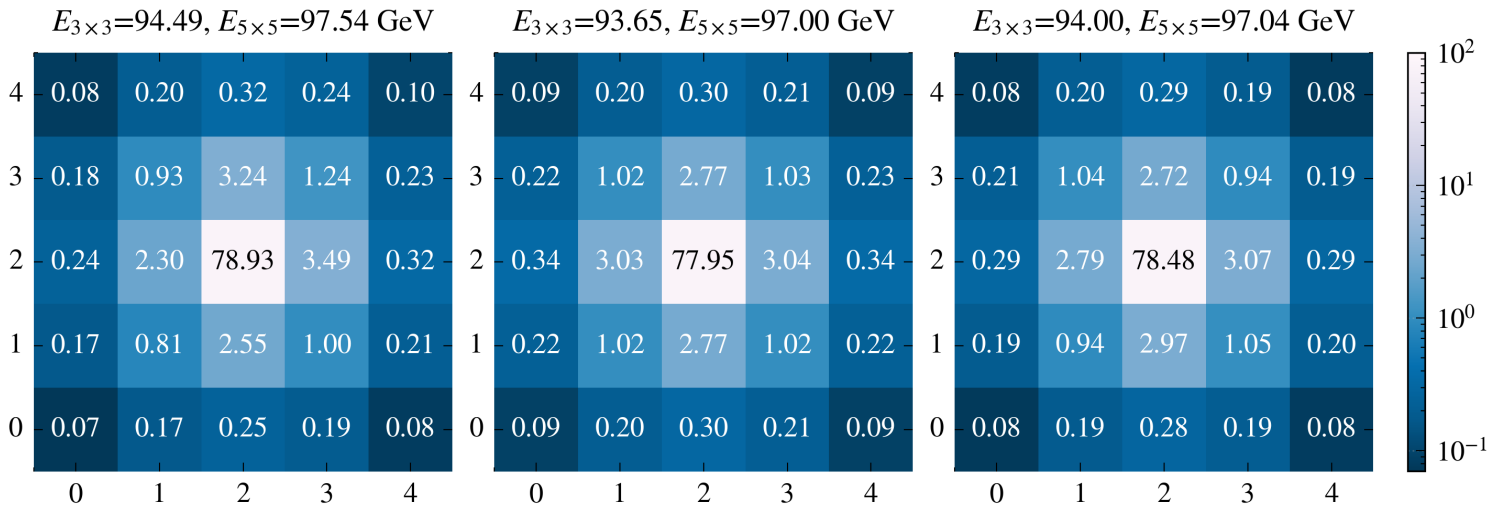
Select all cells with deposited energy above $E_{\text{dep}}^{\text{thresh}}$ as seed candidates. Order them in descending order of deposited energy.

Iterative procedure while seed candidates remain:

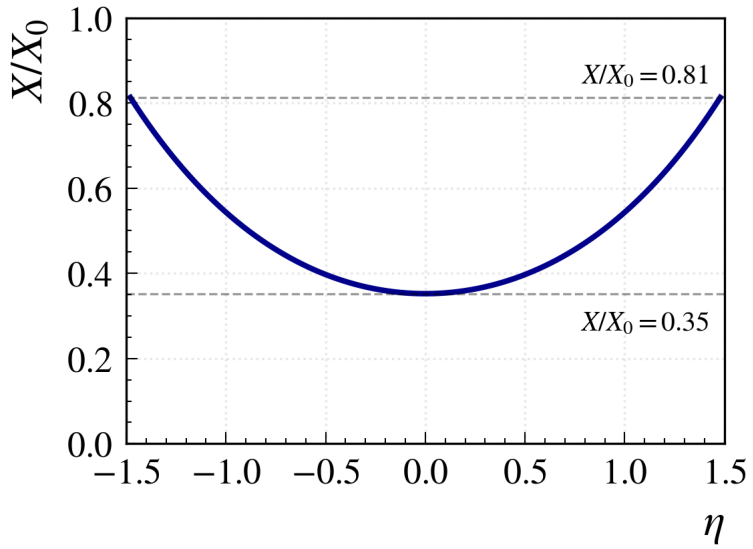
Step 1: Select the highest-E_{dep} remaining seed candidate and define it as the anchor.
Step 2: Construct an overlap window of width d_{overlap} centred on the anchor.
Step 3: Identify all seed candidates contained within this overlap window. For each identified candidate, construct an input window of width d_{window} centred on the candidate.
Step 4: Pass all input windows to the network to obtain the seed probability, relative position $(x_{\text{reco}}, y_{\text{reco}})$ and reconstructed energy E_{reco}.
Step 5: Eliminate all seed candidates contained within the overlap window from the candidate list.

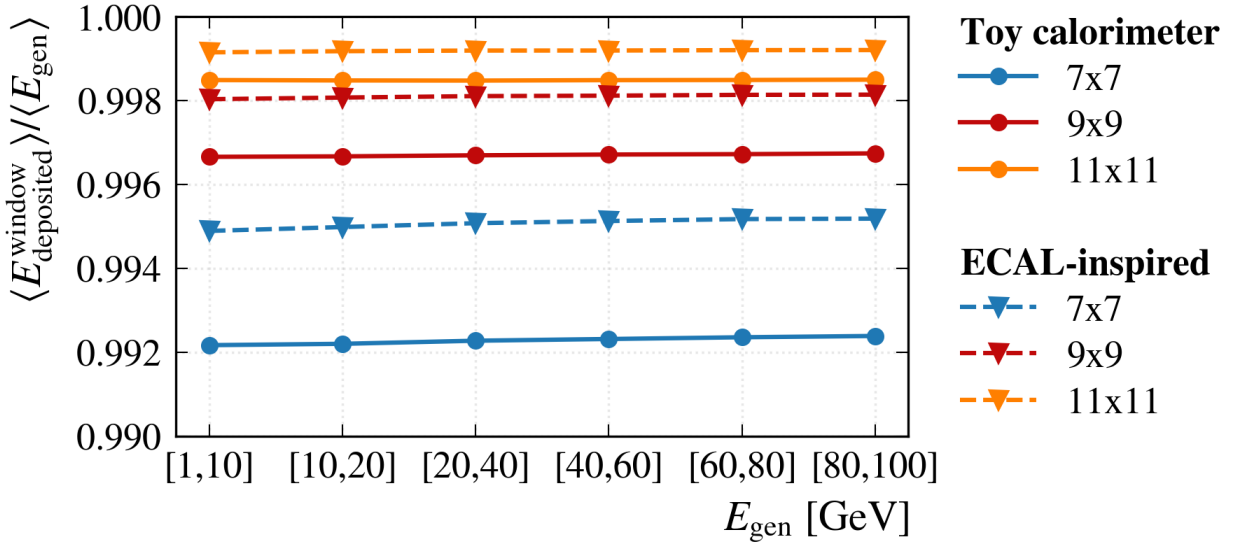


Average energy deposited by a 100-GeV electron beam



Left: ECAL test beam data
Middle: toy calorimeter simulation
Right: realistic geometry simulation

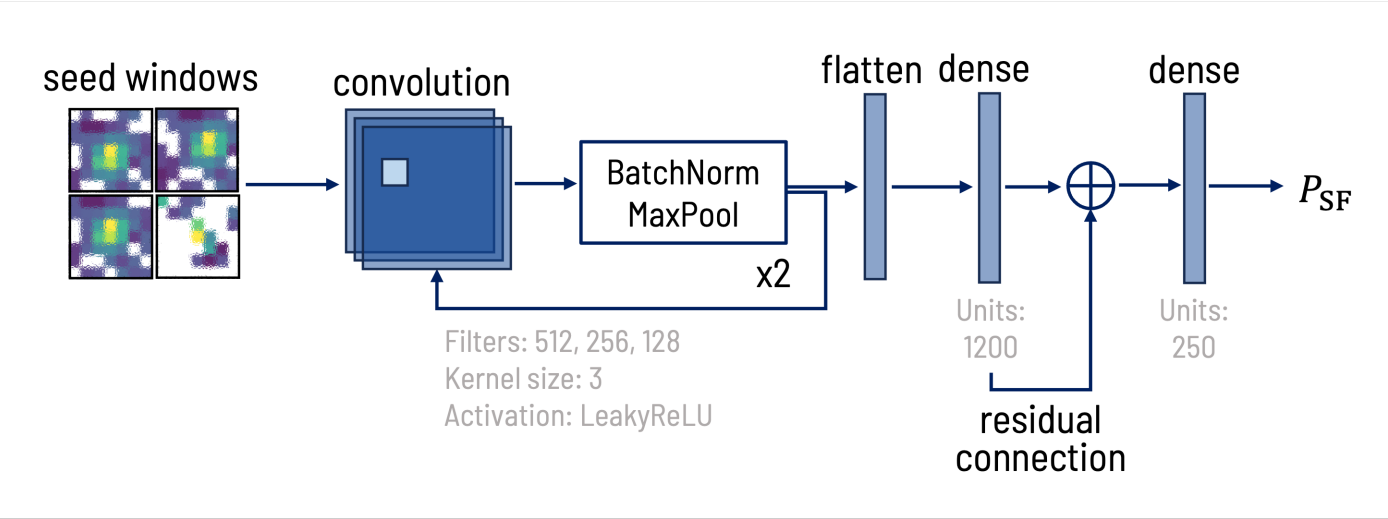


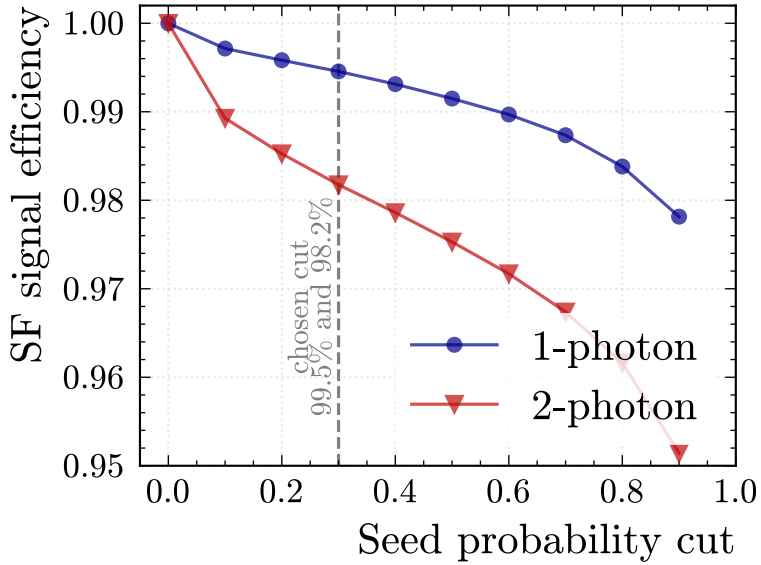


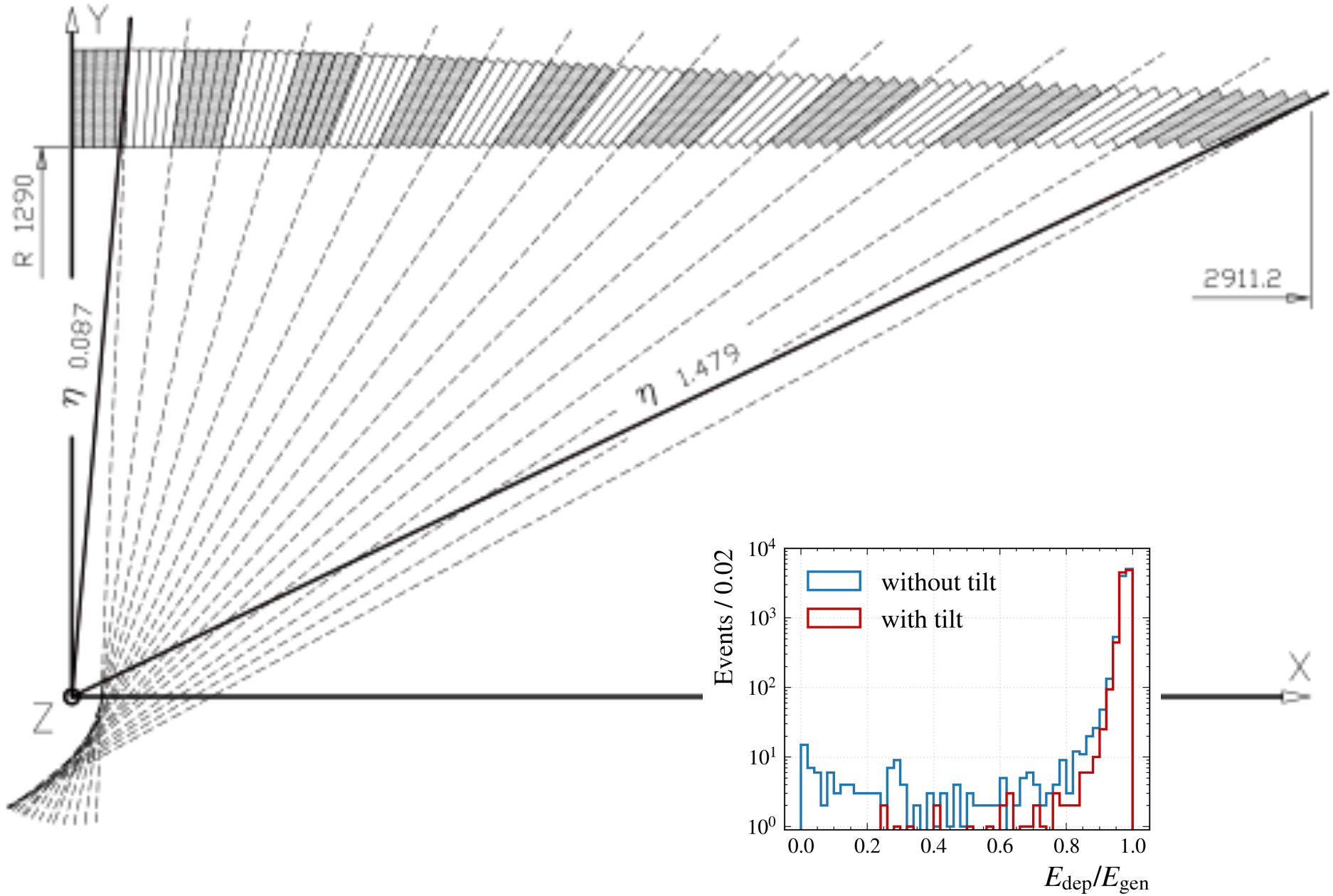
CNN-based network that assigns each candidate window a likelihood that it is centred on the true seed

- Reduces the number of windows passed to the regression stage
- Beneficial for both inference cost and reconstruction stability in events that contain many local maxima

On an event-by-event basis, candidate windows passing a threshold on P_SF are sorted in descending order and the top 4 candidates are forwarded to the regression network.






Y
R 1290
η 0.087
η 1.479
2911.2
Z
X

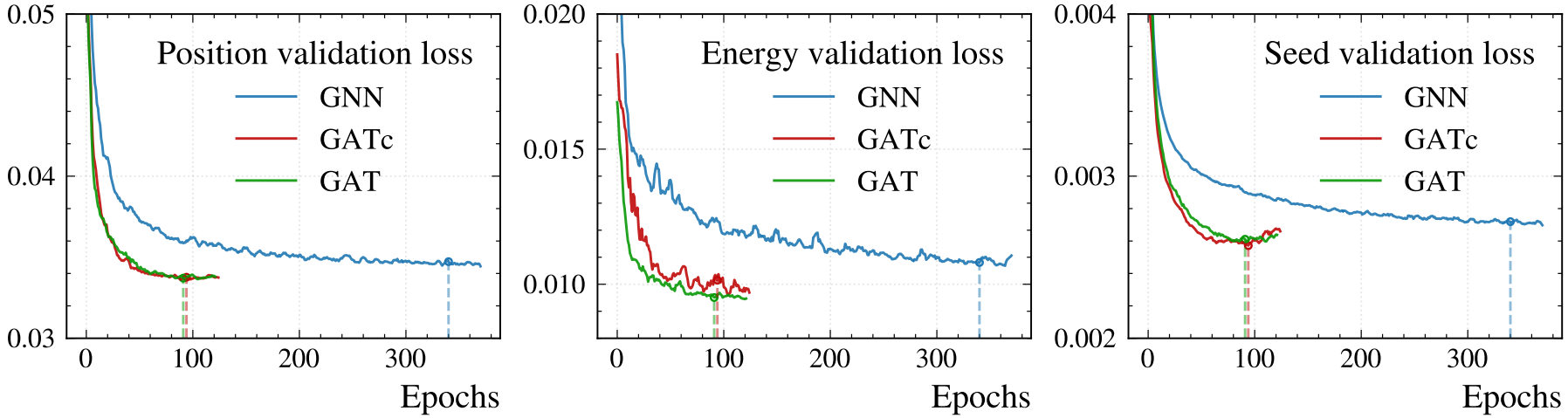
without tilt
with tilt
Events / 0.02
10^4
10^3
10^2
10^1
10^0
0.0
0.2
0.4
0.6
0.8
1.0
E_{dep}/E_{gen}

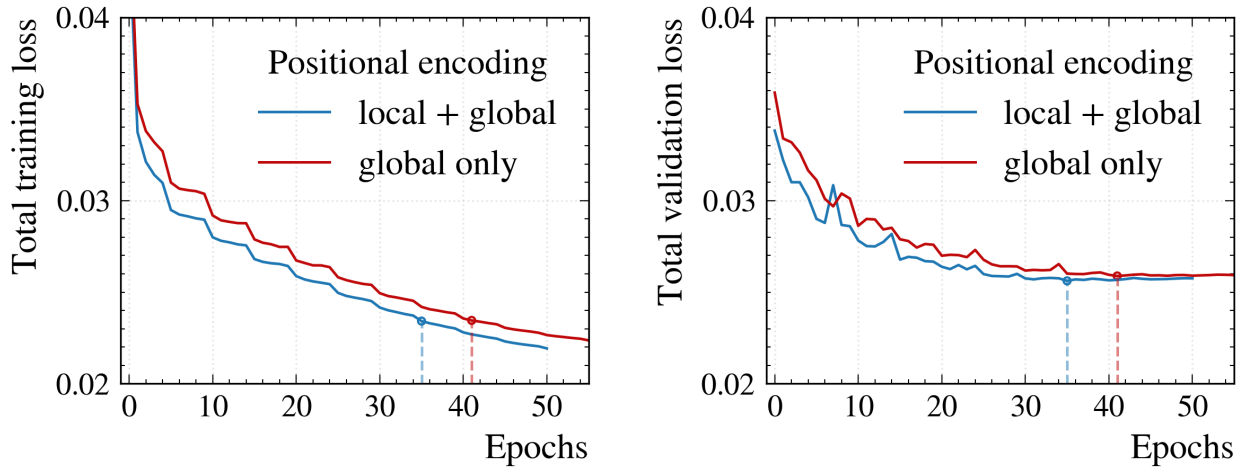
$$\mathcal{L} = \lambda_{\text{energy}} \cdot \mathcal{L}_{\text{energy}} + \lambda_{\text{position}} \cdot \mathcal{L}_{\text{position}} + \lambda_{\text{seed}} \cdot \mathcal{L}_{\text{seed}}$$

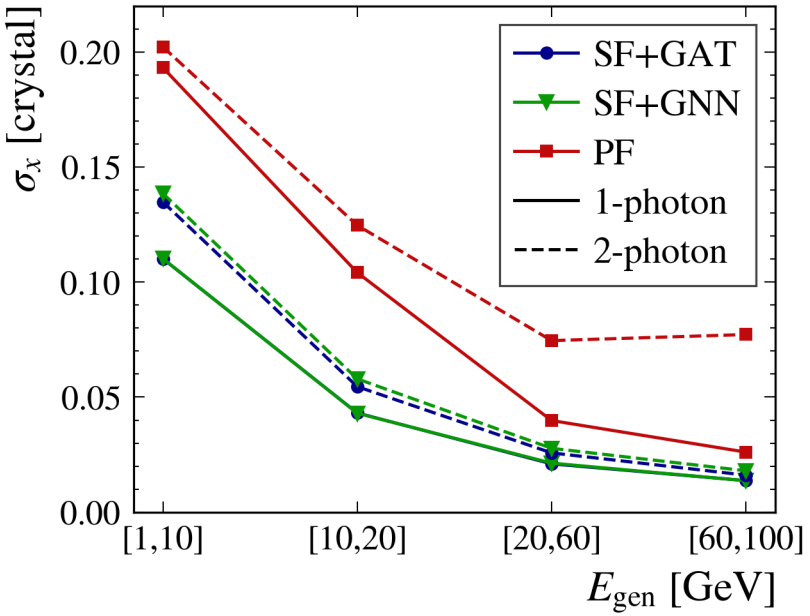
$$\mathcal{L}_{\text{position}} = \frac{1}{N \cdot N_{\text{batch}}} \sum_{i=1}^{N_{\text{batch}}} \sum_{j=1}^{N} \frac{1}{2} \left(|x_{ij}^{\text{pred}} - x_{ij}^{\text{true}}| + |y_{ij}^{\text{pred}} - y_{ij}^{\text{true}}| \right)$$

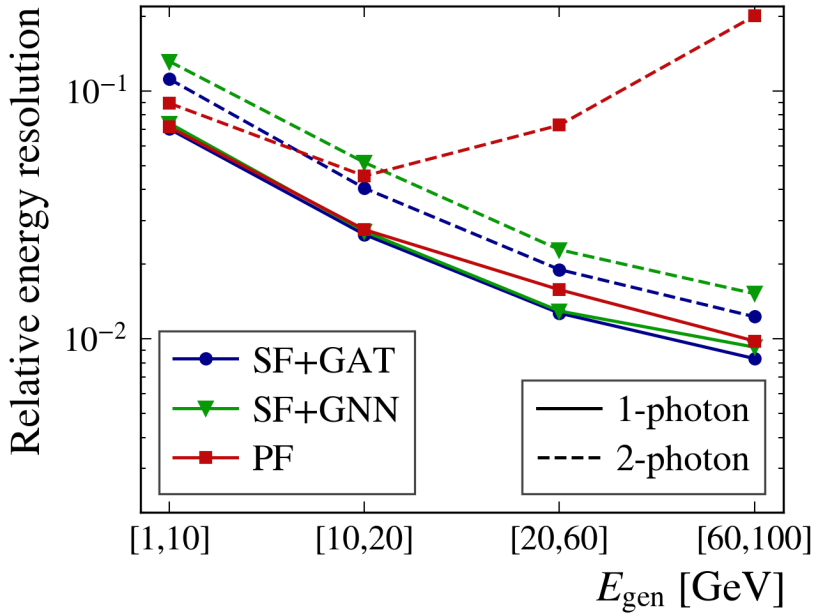
$$\mathcal{L}_{\text{energy}} = \frac{1}{N \cdot N_{\text{batch}}} \sum_{i=1}^{N_{\text{batch}}} \sum_{j=1}^{N} |E_{ij}^{\text{pred}} - E_{ij}^{\text{true}}|$$

$$\mathcal{L}_{\text{seed}} = -\frac{1}{N \cdot N_{\text{batch}}} \sum_{i=1}^{N_{\text{batch}}} \sum_{j=1}^{N} \alpha (1 - P_{ij}^{\text{seed}})^{\gamma} \log (P_{ij}^{\text{seed}})$$

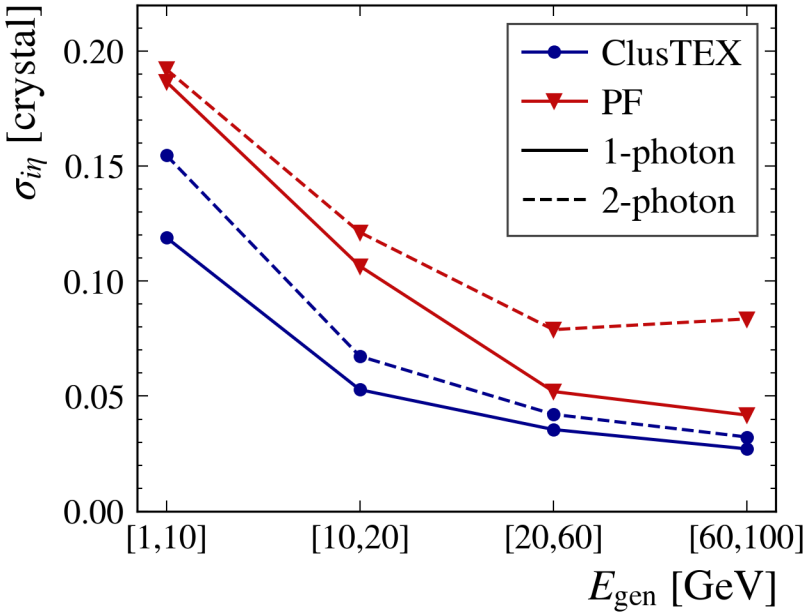


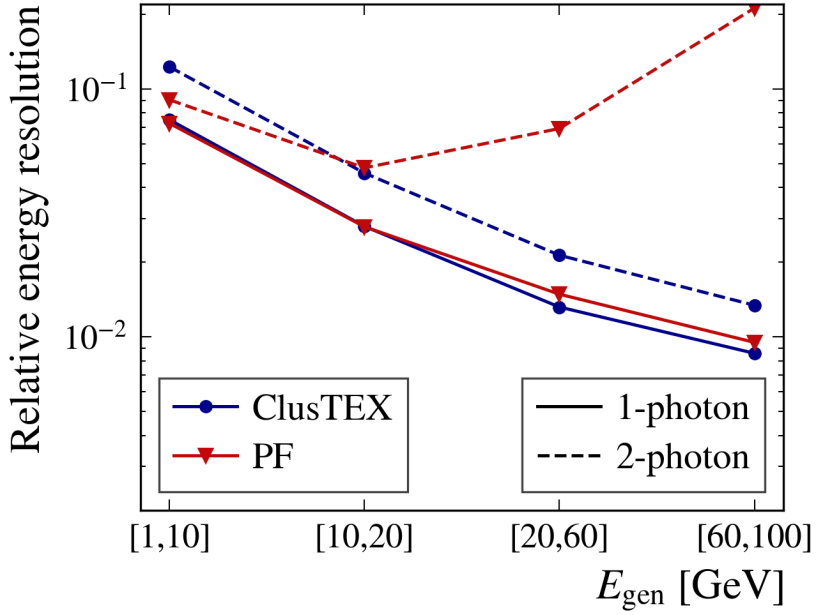






Toy calorimeter simulation





ECAL-inspired simulation