ML/IA usage for LAr crosstalk studies and analysis for LL/ALP search

Nov 21, 2023

Edmar de Souza , Bertrand Laforge, Jose Ocariz, Artur Oudot,
Fred Derue, Mateus Hufnagel, Marton Sandes, Jose Seixas .

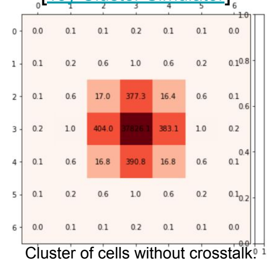
Overview

1. Crosstalk effects in object reconstruction
2. Software implementation for data access
3. Data driven strategy for crosstalk mitigation
4. Ongoing investigation using Lorenzetti Showers Framework
5. Additional ML activities

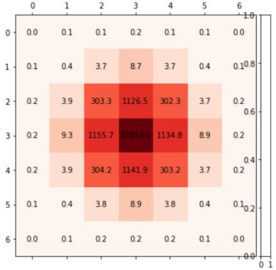
1. Crosstalk effect at LAr Calorimeter

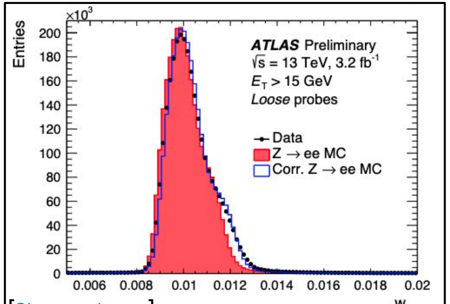
- The crosstalk effect is the charge sharing by means of nearby cabling induction, capacitance coupling and/or direct electrical contact.
- In the LAr calorimeter, it happens directly on its electrodes and between the signal readout chain.
- It mostly impacts signal **timing** reconstruction and its **energy** <u>deposition pattern</u> (shower shapes) in a cluster of cells.

[Toy Cluster Simulator]

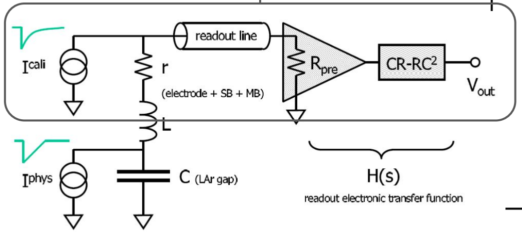


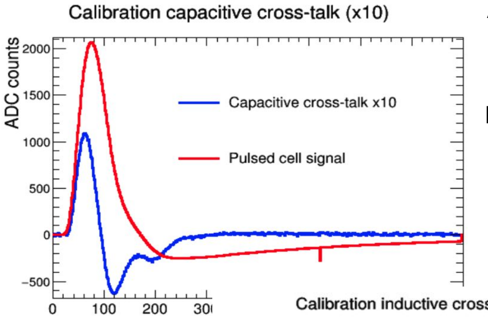
Cluster of cells without crosstalk.





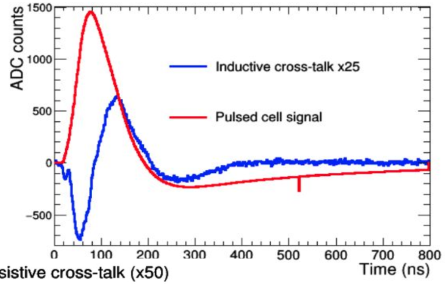
[Shower shapes]



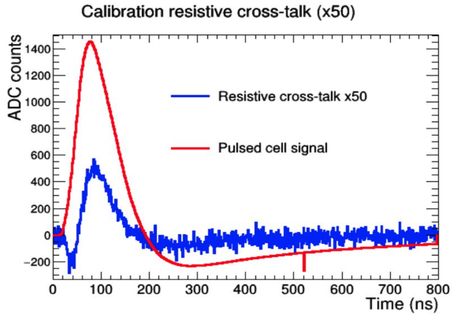


ATLAS Work in progress

Cell is pulsed in EMB1, cross-talk is retrieved from closest neighbor cell in η.



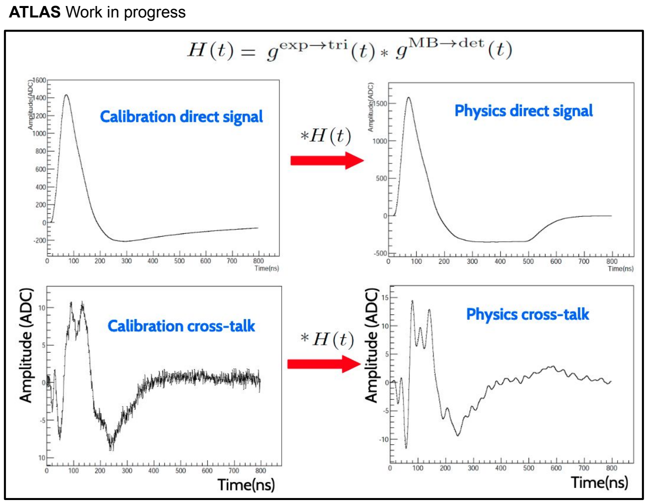
Cell is pulsed in EMB2, cross-talk is retrieved from closest neighbor cell in η.

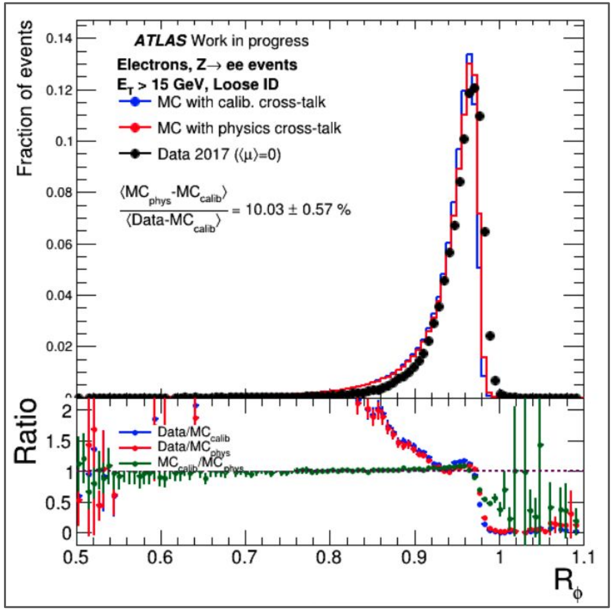


Cell is pulsed in EMB2, cross-talk is retrieved in EMB1 neighbor.

1. Crosstalk effect at LAr Calorimeter

- Complex to be simulated because its effect may be sensitive to 1st and 2nd neighboring.
- Highly non-linear and geometry dependent.
- Current Monte Carlo simulation account a <u>cell-based</u> representation of the crosstalk signal (EM2 only).
- Crosstalk signal from physics was implemented. It gives a better representation when compared to data but has its limitations.





1. Crosstalk: Impact on photon reconstruction

Topoclusters being reconstructed, inside the circles in magenta, arrive later to the calorimeter with respect to the photon candidate, between 5 and 20 ns,with approximately 0.1% of the photon energy.

- The probability of reconstructing crosstalk topoclusters increases with the energy of the photon.
- The reconstruction rate of these topoclusters increases with the energy of the photon candidate, compared to the reconstructing rate of pileup topoclusters, which remains stable.

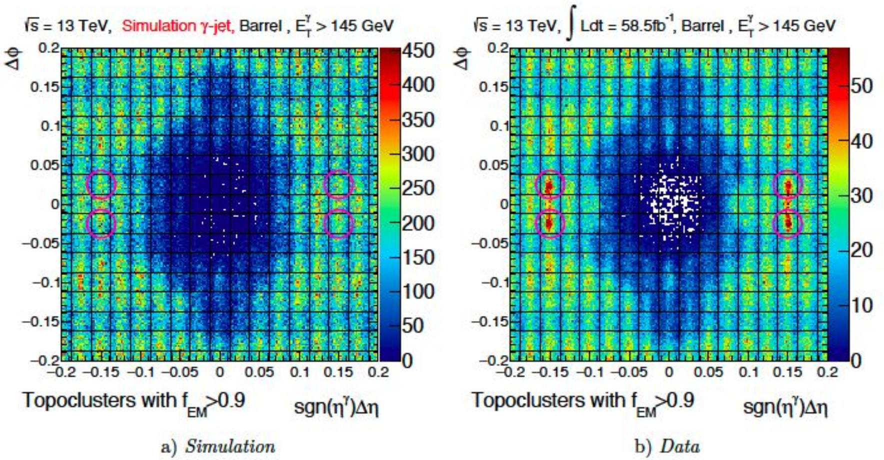
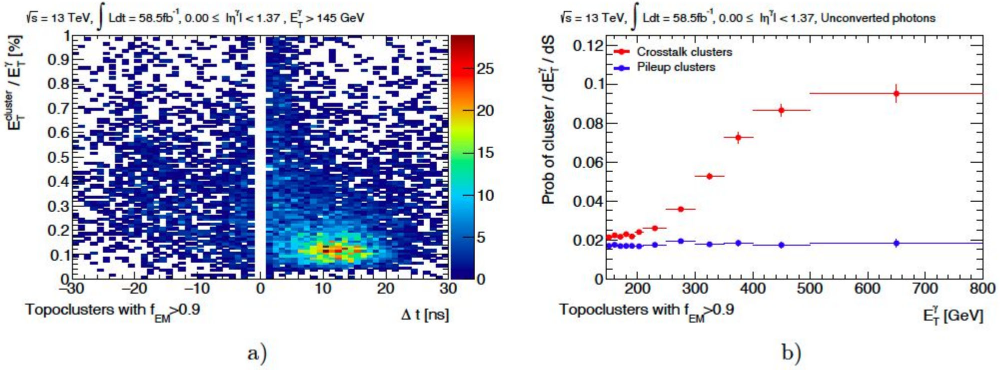


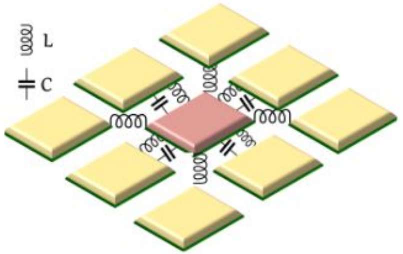
Figure D.6: Topoclusters around tight isolated photon candidates in the barrel with $E_T^\gamma > 145$ GeV in the $(sgn(\eta^\gamma)\Delta\eta, \Delta\phi)$ plane. Simulation is shown in (a) and data in (b). The topoclusters inside the regions in magenta are selected for the presented analysis.



[Luis Pascual's thesis]

1. Crosstalk: Impact on photon reconstruction

- MC can't fully represent crosstalk at LAr Calorimeter.
- Previous investigations were carried out with low mu collisions data. Focused on the analytical model for crosstalk and some tests with machine learning.
- Concept proof: Can machine learning approach deal with it?
- **Now: What do we know?**



- It affects:
  - Energy and shower shape variables.
  - timing distribution.
- ML tools can correct xtalk effects in supervised approach.
- So, could be possible to use a **data driven** strategy for cross-talk correction?
  - Correction single cell is not possible (not enough constraints to optimize the model).
  - But going <u>from 1 cell</u> to a <u>cluster level</u> can provide a feasible solution.
    - Cluster is produced by a <u>single particle</u>, with <u>single time</u>.

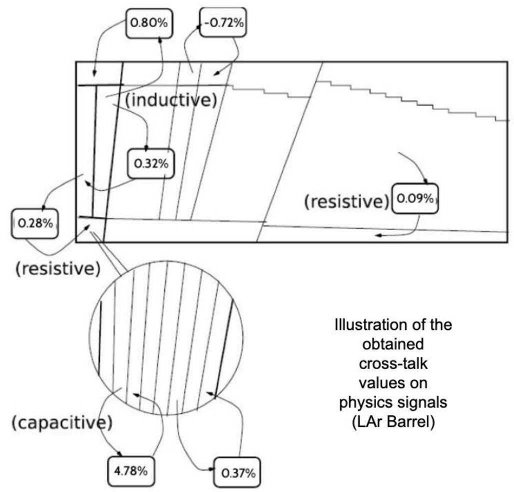
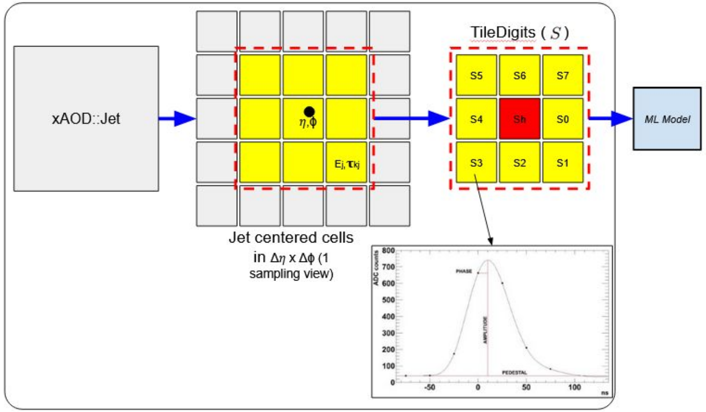


Illustration of the obtained cross-talk values on physics signals (LAr Barrel)

2. Software Implementation for Data Access

- **EventReader** data dumper package (gitlab).
    - Developed for cross-talk studies in LAr Calorimeter, but is designed for a variety of physics objects studies.
    - Access physics objects (xAOD::Electrons, xAOD::Photons) and retrieve its associated **CaloCluster** and its cells.
    - Get **xAOD::Jets** container and map every cell in a $\Delta\eta$ x $\Delta\phi$ region around the Jet.
    - Iterate over its **CaloCells**, mount a HWID map and dump their content
        - E, t, geometry descriptors, online and offline calibration constants from DB, etc.
        - Scan the **RawChannel** container for matching HWID and dump their content too.
        - A custom index number was created to maintain the structured data linked in the flat ntuple.
        - Also, there is a **Channel** level to keep Tile PMT's associated to a given CaloCell, and its **Digits** container samples.
    - Every output option are customizable through job options flags (python).
    - The tool is constantly being updated to match studies that require low level calorimetry data, in association with physics objects.
    - It produces a flat NTuple at the output, that can be easily handled in ML studies.

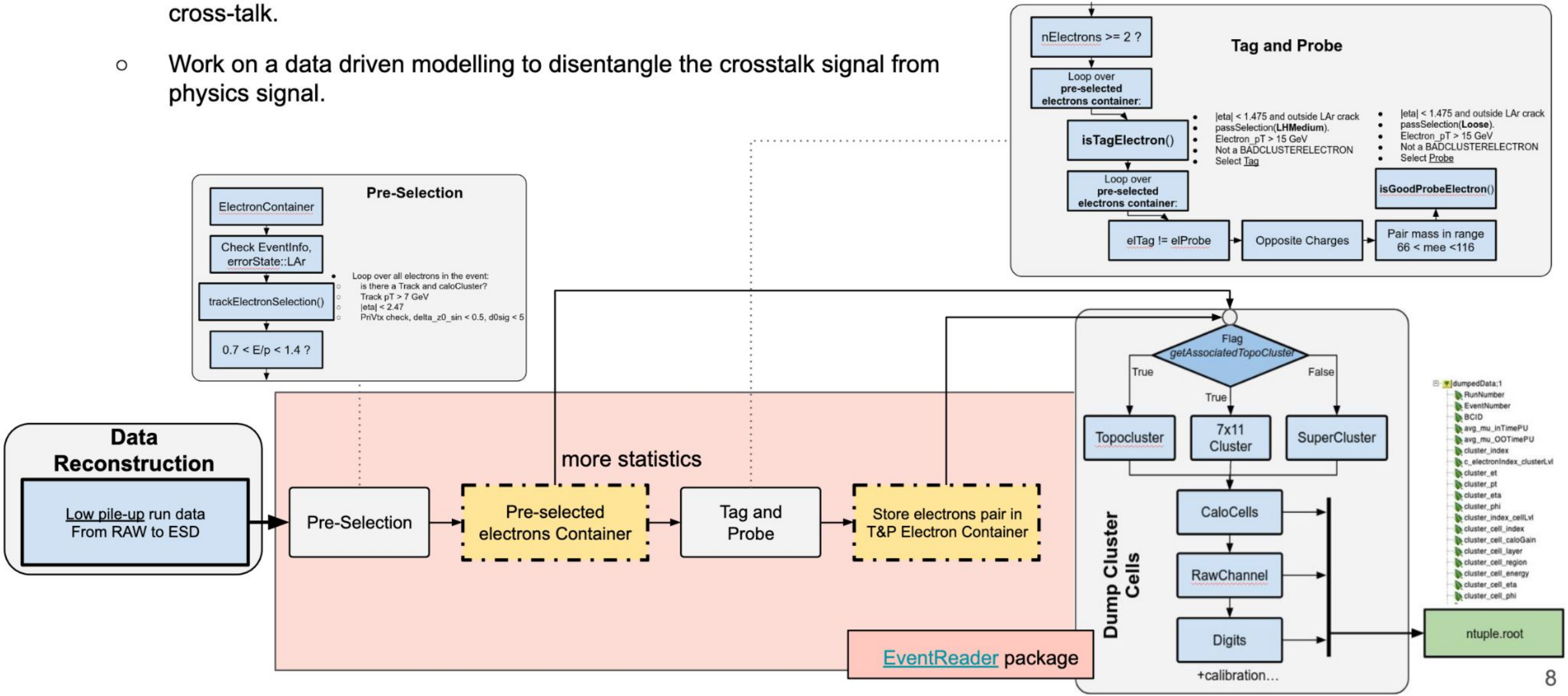


EventReader dumper representation for Jets

Mateus's presentation

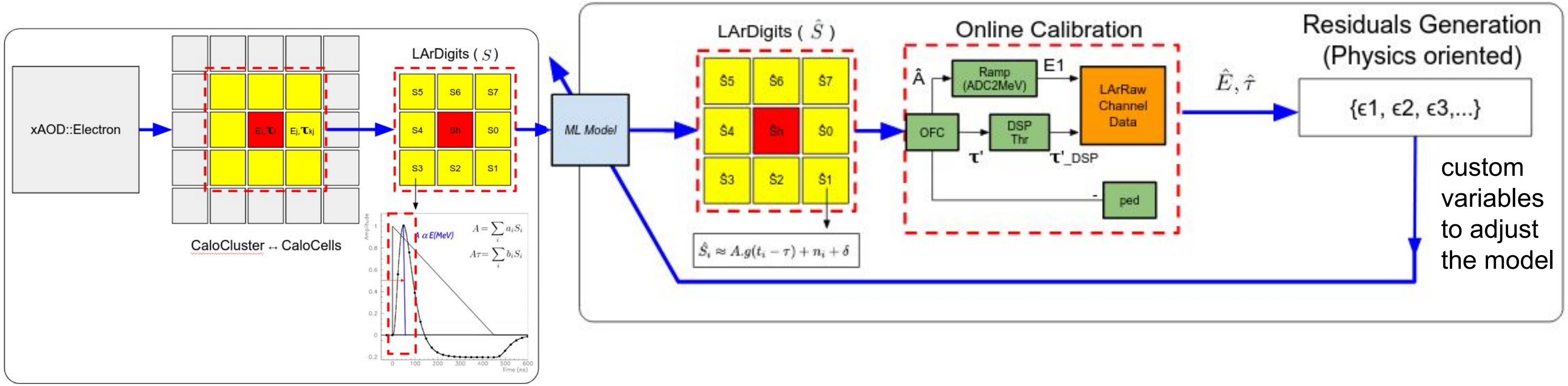
2. Software Implementation for Data Access

- This step roadmap is organized as following:
  - Using low mu data to have a readout signal composed by noise, signal and cross-talk.
  - Work on a data driven modelling to disentangle the crosstalk signal from physics signal.



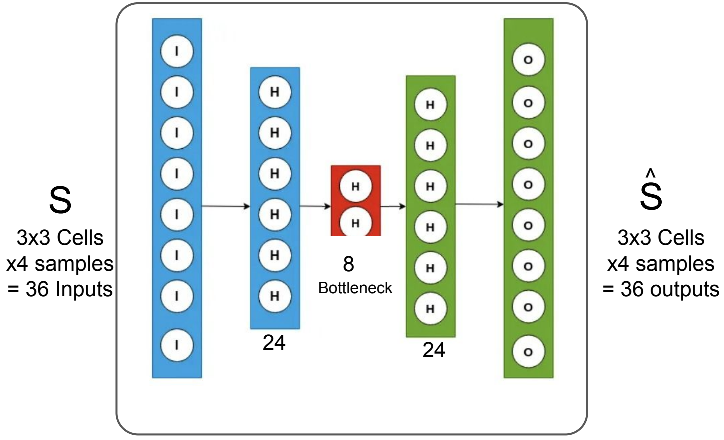
3. Data driven strategy for crosstalk investigation

- ML input will get as input a cluster of signal samples 3x3 set of cells (ηxϕ), from EMB2.
  - simplest cluster of cells and expected a 1st neighbor crosstalk.
  - the signal samples <u>are available only for higher cell energies</u> depositions, related to noise.
  - smaller structures increases (for now) statistics of a closed set of cells.
- Also, the reconstructed energy and time will feedback the model, so calibration 'on the fly' should be performed.
- Variables directly affected by cross-talk, like estimated time, should be part of a custom loss function, which leads the model towards the expected solution.
- Cells η, ϕ and energy resolution (ln e) can be used as inputs to guide the model to better correlate the signals in the cluster to its neighboring cells.

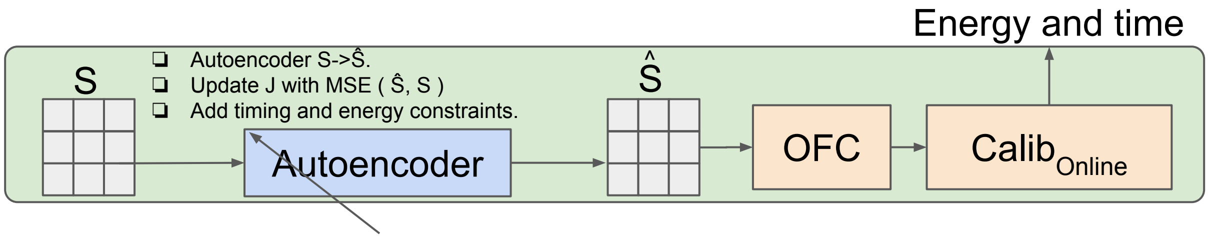


3. Data driven strategy for crosstalk investigation

- **First approach methodology:**
  - Since there aren't any truth value to the model be trained towards it, the solution space is very large and, consequently, hard to achieve in a non-supervised training.
  - Based on that, starting from a simple model in a semi-supervised design allow to control the model variables.
  - This way, complexity can be added as improvement on training process are being made.
  - The chosen topology here is the <u>Autoencoder</u>:
    - This experiment proposes to <u>learn the set of input samples S</u>, like a supervised training.
      - From this, the reconstructed **time** and **energy** are monitored.
    - As a **second** step, the <u>energy</u> and <u>time</u> should be included in the training process, aiming to improve the samples representation at the output and, therefore, the post reconstructed variables.



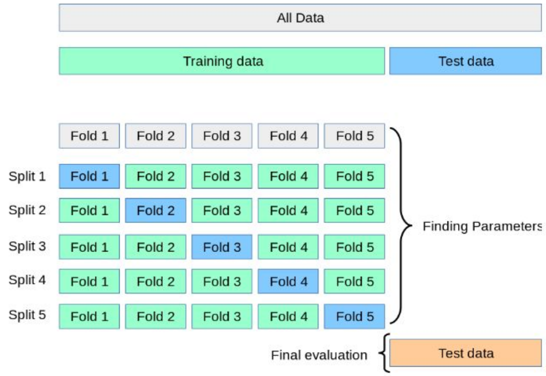
Autoencoder representation



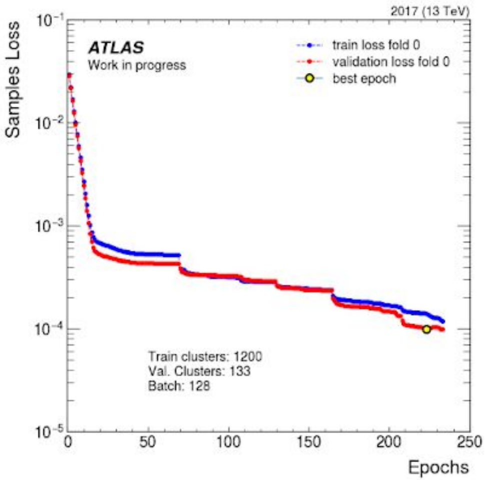
3. Data driven strategy for crosstalk investigation

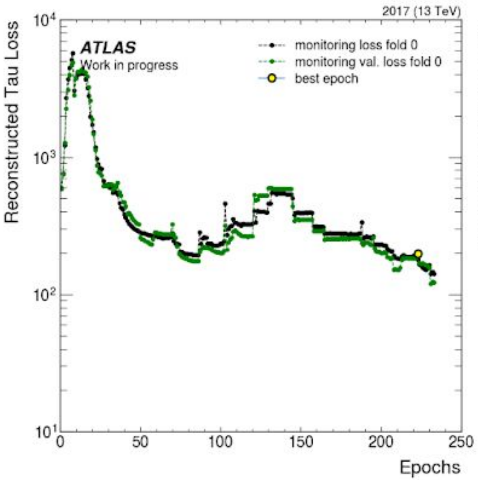
ML Training methodology:

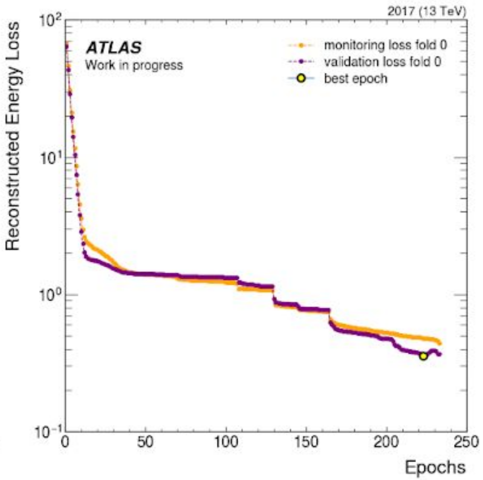
- Using cross-validation strategy to understand statistical fluctuations from data sets;
- Multiple ML model initialization to evaluate/detect issues with local minima;
- Evaluating loss function and training curves to monitor overfitting;
- Ongoing investigation about bottleneck autoencoder size
- Pytorch-based implementations.



Partial results. Just a crosscheck on minimizing the optimization problem:







[LAr Meeting: Mateus Hufnagel]

4. Ongoing investigation using Lorenzetti Showers Framework

The Lorenzetti Showers (LZT) simulation is an integrated software framework that provides complete calorimeter information.

The framework includes:

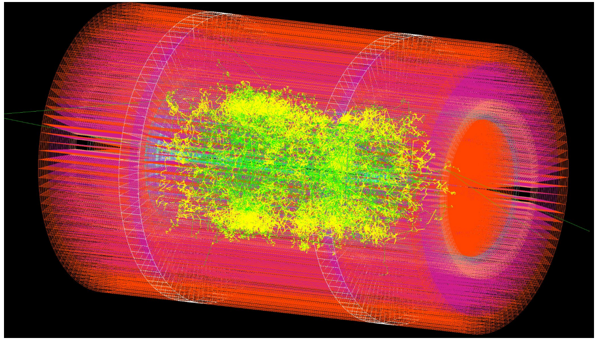
- Cell readout values and configurable sensor pulse-shapes;
- Different energy estimation methods for handling low and high pileup operation conditions;
- Providing a user-friendly, flexible, user-oriented, and low-level calorimeter simulation framework for the scientific community.
- Repository: https://github.com/lorenzetti-hep/lorenzetti

https://doi.org/10.1016/j.cpc.2023.108671


ELSEVIER
Computer Physics Communications
Volume 286, May 2023, 108671

Lorenzetti Showers - A general-purpose framework for supporting signal reconstruction and triggering with calorimeters ☆, ☆☆

M.V. Araújo [a], M. Begalli [b], W.S. Freund [a], G.I. Gonçalves [b], M. Khandoga [c], B. Laforge [c], A. Leopold [d], J.L. Marin [e], B.S-M. Peralva [b], J.V.F. Pinto [a], M.S. Santos [e], J.M. Seixas [a], E.F. Simas Filho [e], E.E.P. Souza [c]



4. Lorenzetti Showers Framework

- Implement barrel, Extended Barrel, Tile, Hadronic EndCap (HEC) and forward calorimeter;

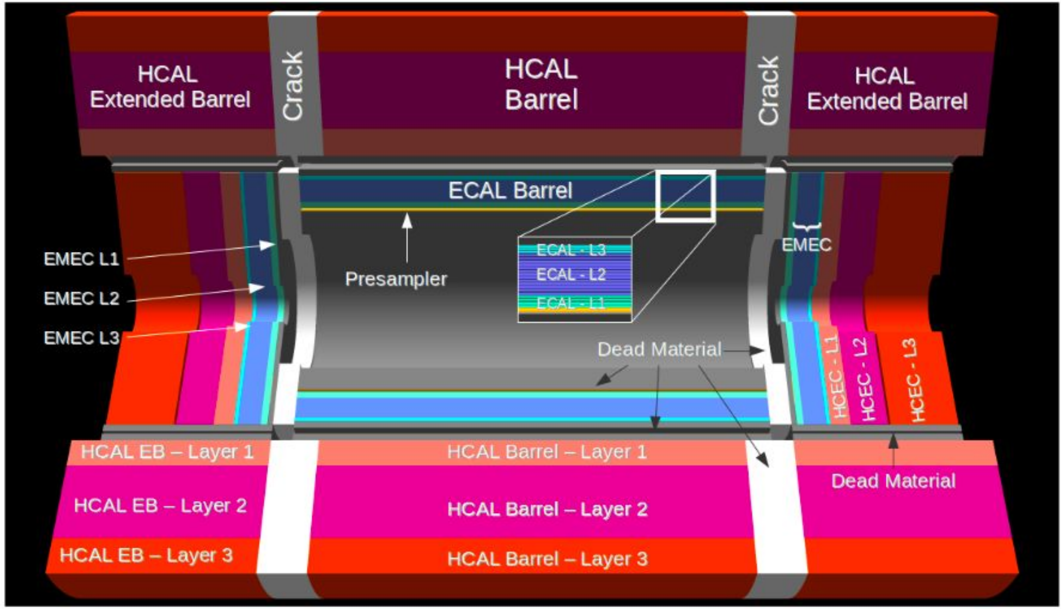
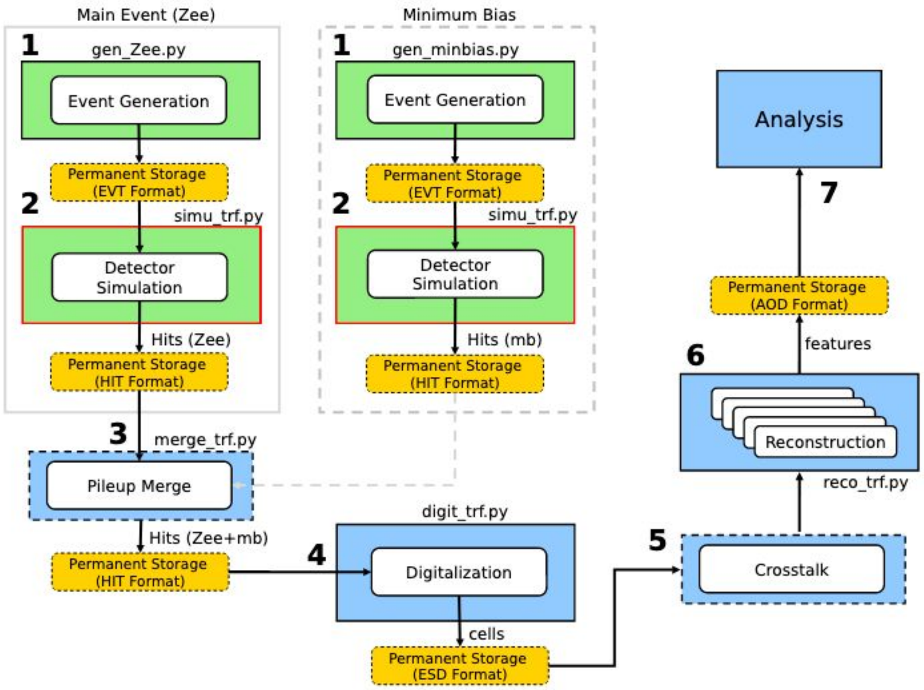


Table 1: Coverage regions in η, cell granularity and sampling layers used in the simulated calorimeter.

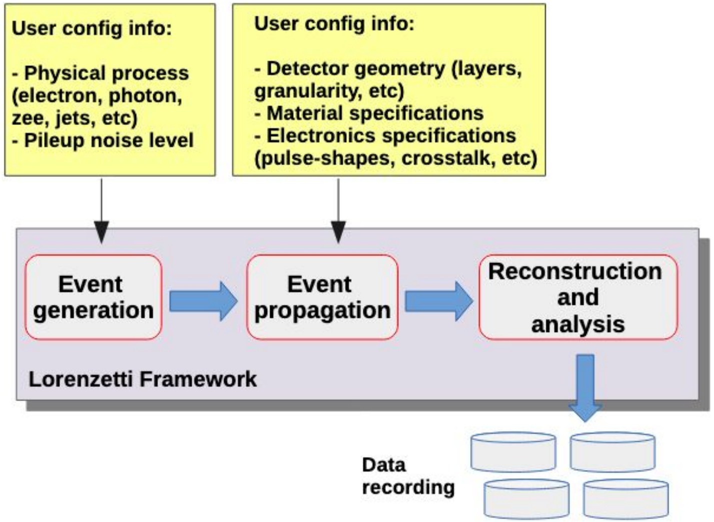
Layer	Sampling	Coverage	Granularity ($\Delta\eta \times \Delta\phi$)
Presampler	Barrel	$0.00 < \|\eta\| < 1.58$	0.025×0.1
	End-Cap	$1.50 < \|\eta\| < 1.80$	0.025×0.1
Electromagnetic Calorimeter			
Layer 1	Barrel	$0.00 < \|\eta\| < 1.55$	0.003×0.1
	End-Cap	$1.37 < \|\eta\| < 1.80$	0.003×0.1
		$1.80 < \|\eta\| < 2.00$	0.025×0.1
		$2.00 < \|\eta\| < 2.37$	0.006×0.1
		$2.37 < \|\eta\| < 3.20$	0.1×0.1
Layer 2	Barrel	$0.00 < \|\eta\| < 1.50$	0.025×0.025
	End-Cap	$1.35 < \|\eta\| < 2.50$	0.025×0.025
		$2.50 < \|\eta\| < 3.20$	0.1×0.1
Layer 3	Barrel	$0.00 < \|\eta\| < 1.58$	0.05×0.1
	End-Cap	$1.35 < \|\eta\| < 2.50$	0.05×0.025
		$2.50 < \|\eta\| < 3.20$	0.1×0.1
Hadronic Calorimeter			
Layer 1	Barrel	$0.00 < \|\eta\| < 1.09$	0.1×0.1
	Extended Barrel	$0.94 < \|\eta\| < 1.77$	0.1×0.1
	End-Cap	$1.50 < \|\eta\| < 2.50$	0.1×0.1
		$2.50 < \|\eta\| < 3.20$	0.2×0.2
Layer 2	Barrel	$0.00 < \|\eta\| < 1.09$	0.1×0.1
	Extended Barrel	$0.85 < \|\eta\| < 1.41$	0.1×0.1
	End-Cap	$1.50 < \|\eta\| < 2.50$	0.1×0.1
		$2.50 < \|\eta\| < 3.20$	0.2×0.2
Layer 3	Barrel	$0.85 < \|\eta\| < 0.72$	0.2×0.1
	Extended Barrel	$0.85 < \|\eta\| < 1.41$	0.2×0.1
	End-Cap	$1.50 < \|\eta\| < 2.50$	0.1×0.1
		$2.50 < \|\eta\| < 3.20$	0.2×0.2

4. Lorenzetti Showers Framework - Software Architecture

Simulation Chain:



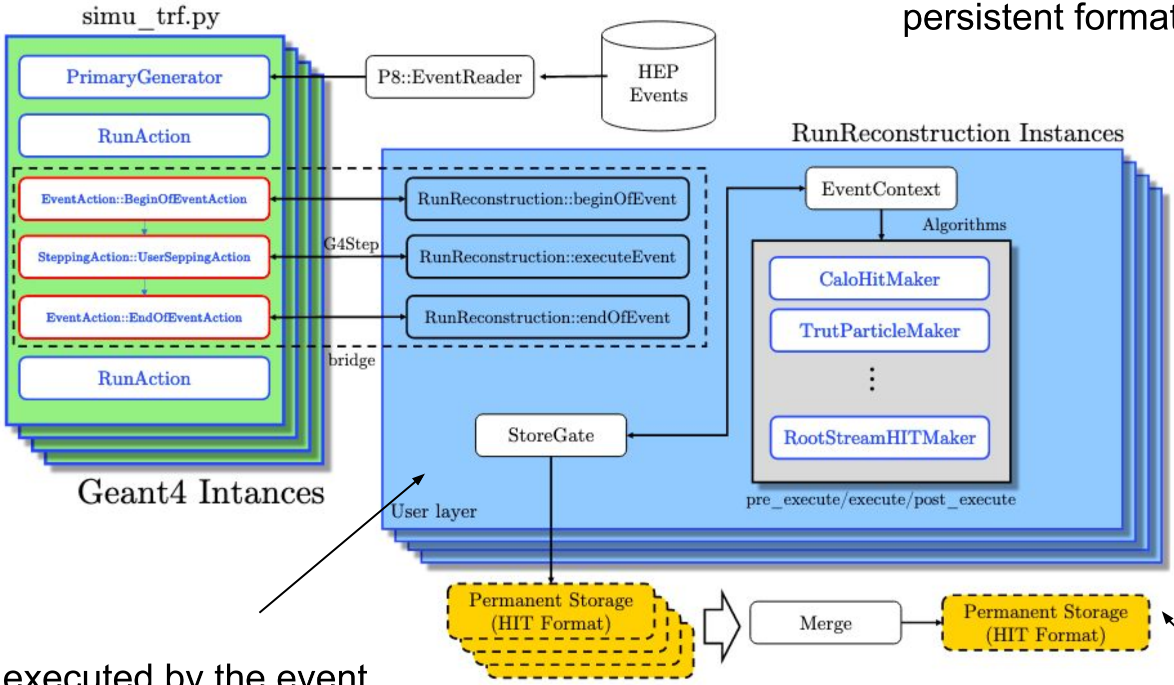
Generation	Simulation	Reconstruction
Pythia	Geant4	Gaudi-Inspired



4. Lorenzetti Showers Framework - Software Architecture

Main steps for running the Geant4 simulation

The information generated by the reconstruction algorithms is stored in persistent format (HIT).

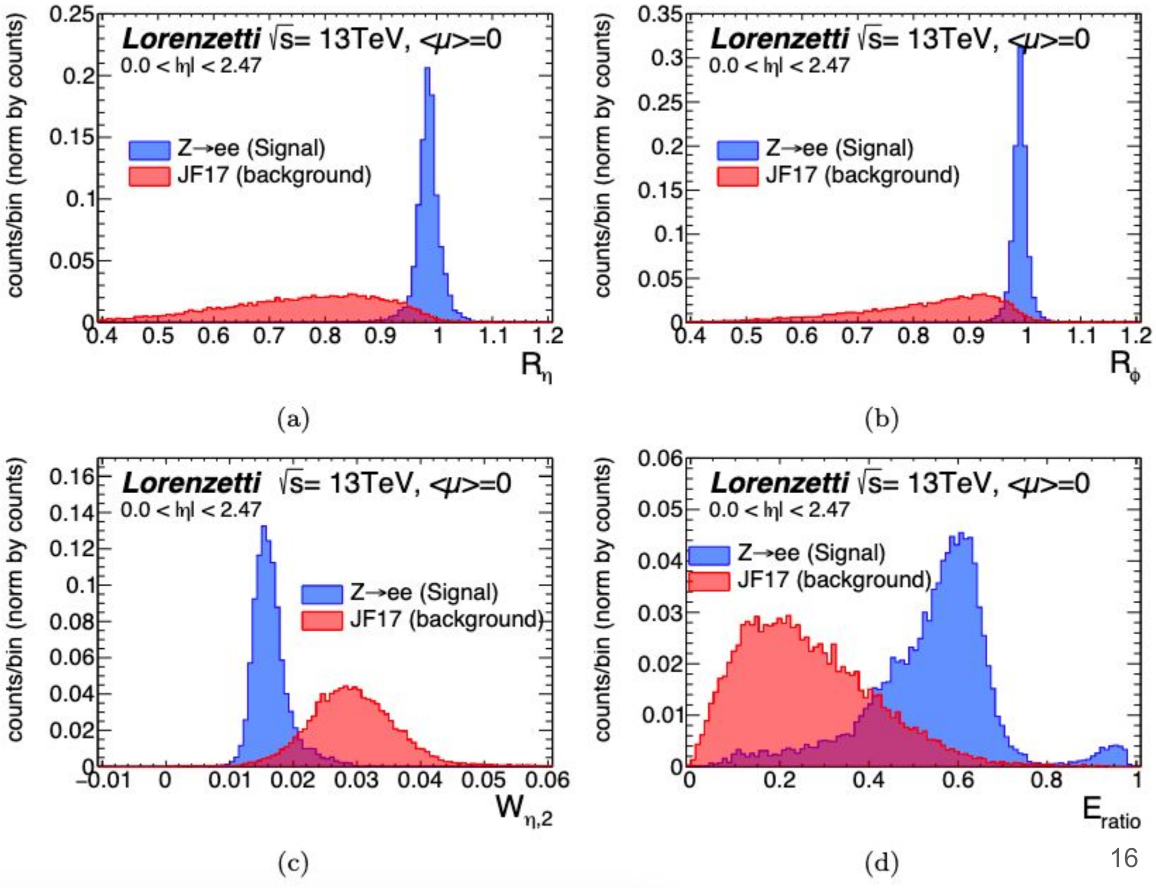


Main components executed by the event manager (`RunReconstruction`)

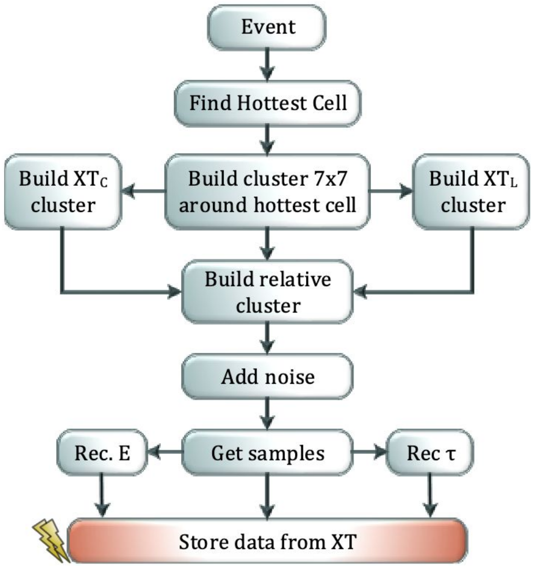
4. Reconstruction for High Level Variables for Shower Description

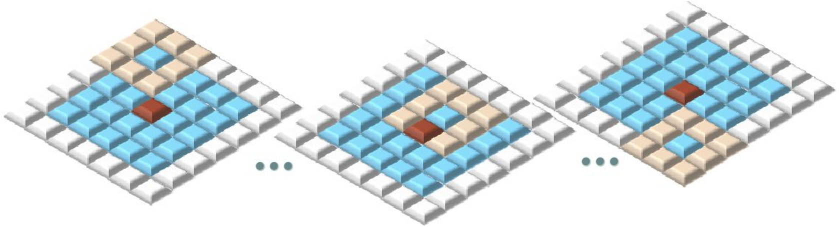
Shower variables computed for the energy clusters from Zee (blue) and jets (red)

- High level variables produced by LZT, can be used in ML designs with calorimetry information, for particle discrimination purposes;
- It is also possible to access cell-level information, which allows investigation of energy estimation methods, topological mappings, etc.

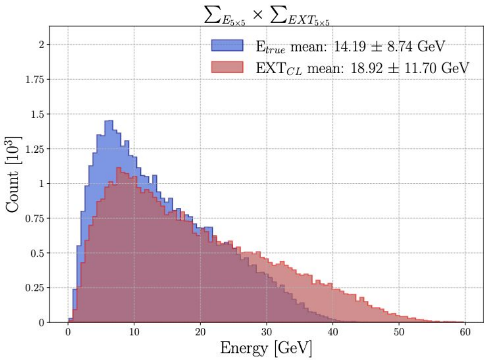


4. Ongoing investigation using Lorenzetti Showers Framework

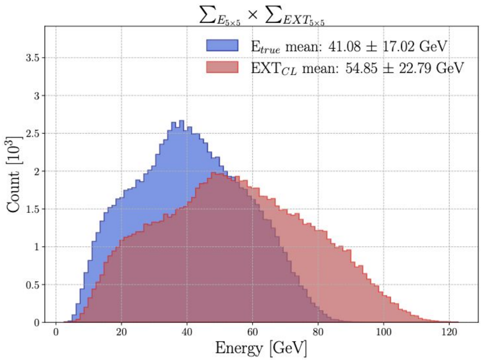




- LZT [10,50] GeV



- LZT [50, 100] GeV

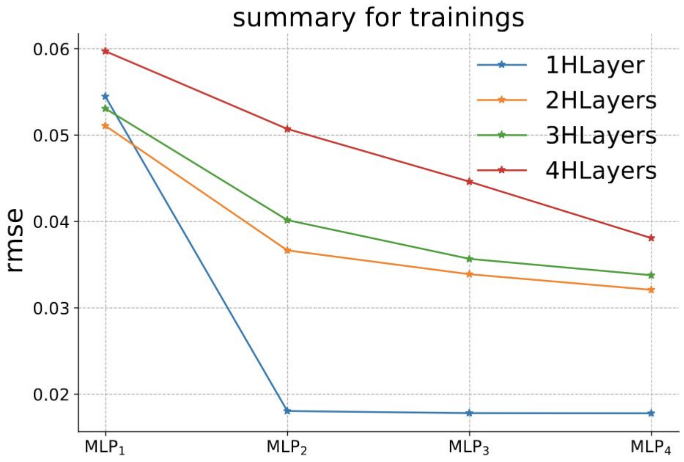


- Producing additional clusters to verify the differences in the inductive and capacitive crosstalk contributions, based on an analytical model.
- Evaluated different energy ranges and also impact on the cell cluster time reconstruction.

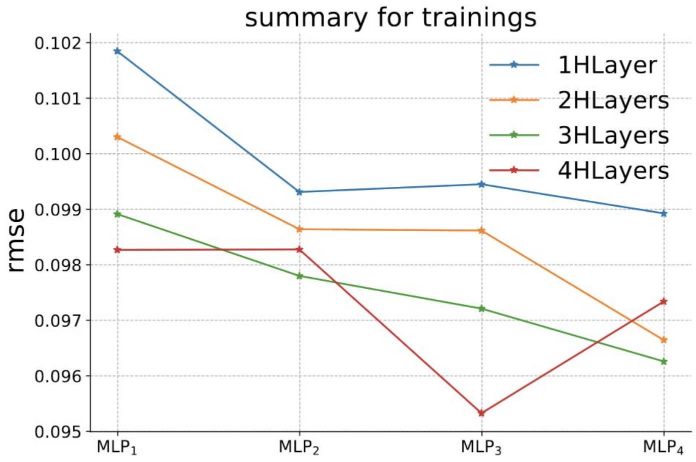
4. Ongoing investigation using Lorenzetti Showers Framework

Evaluating neural network architectures for energy regression.

- Objectives:
    - Include crosstalk model in LZT and build ML models to mitigate crosstalk effects.
    - Evaluating neural network architecture and training methods.
- 10-fold cross-validation with 15 initializations for each model (local minima investigation)
- ~100k events were used for each configuration.
- Different types of architecture, numbers of hidden neurons, and activation functions were evaluated.



rmse (Energy)

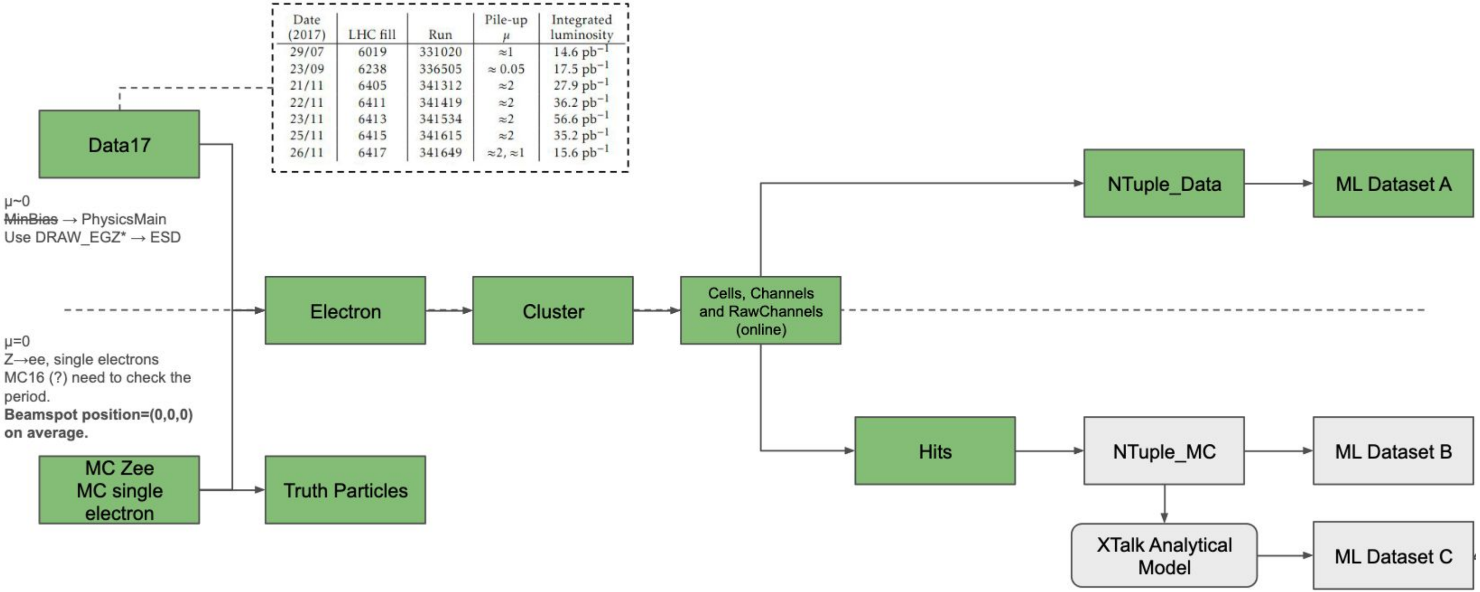


rmse (τ)

5. Additional ML activities

CrossTalk investigation:

- Investigation with Lorenzetti framework will help guide studies with ATLAS LAr data.
- Planning to use applied models and evaluated architectures built with Lorenzetti's data collections.
  - It will help save time and evaluation of many parameters.
- Some studies with different optimization fitness functions to consider time distribution in each cluster of cells can be used.
- It is also desired to use official MC ATLAS productions to complete analyses.



Date (2017)	LHC fill	Run	Pile-up μ	Integrated luminosity
29/07	6019	331020	≈ 1	14.6 pb^{-1}
23/09	6238	336505	≈ 0.05	17.5 pb^{-1}
21/11	6405	341312	≈ 2	27.9 pb^{-1}
22/11	6411	341419	≈ 2	36.2 pb^{-1}
23/11	6413	341534	≈ 2	56.6 pb^{-1}
25/11	6415	341615	≈ 2	35.2 pb^{-1}
26/11	6417	341649	$\approx 2, \approx 1$	15.6 pb^{-1}

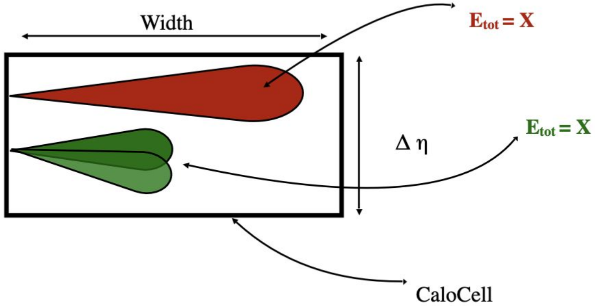
5. Additional ML activities

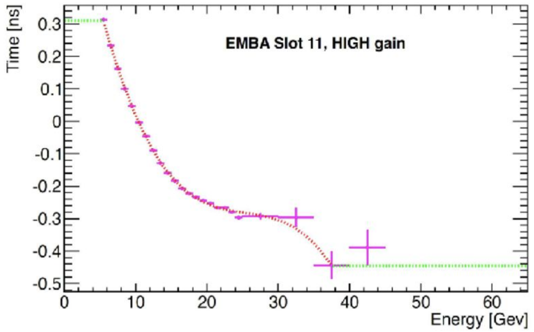
Long Lived Axion-Like Particles:

- Ongoing investigations aimed at improving the reconstruction of closed by photon pairs.

$$h \to Za \quad a \to \gamma\gamma$$

- Highly collimated photon-jet pair.
- Time information can be useful to help with boosted/non-boosted discrimination, since shower depth can be energy and time dependent.
  - But time distributions are largely affected by cross-talk. Therefore, crosstalk mitigation may help in building better features in LLP search.
- We want to combine Calo and time information to build ML models for LLP/ALP analysis.
  - Try to identify photon pairs produced in the calorimters and regress their mass .





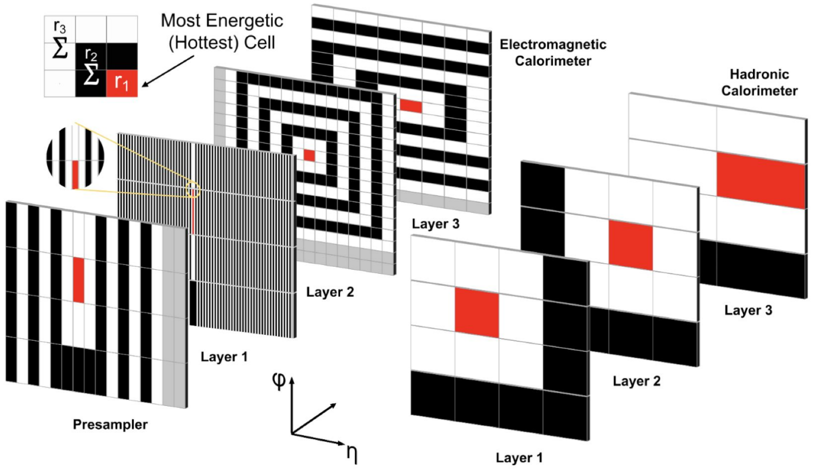
Nevis ATL-LARG-INT-2012-003 : figure 13

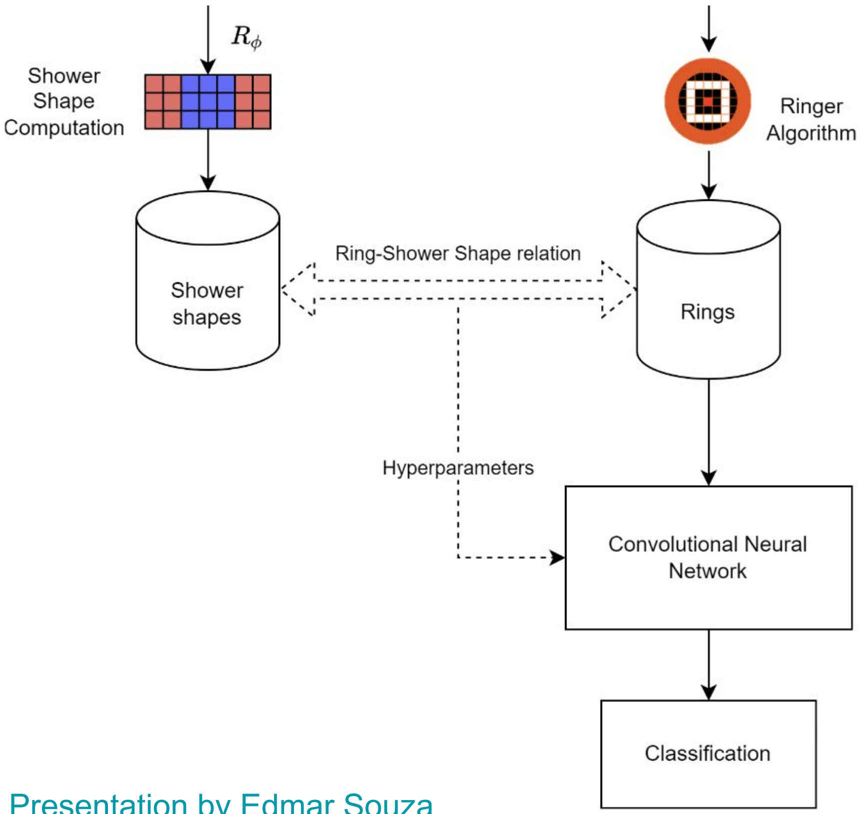
[Presentation by Artur Oudot]

5. Additional ML activities

Improvements on boosted photon pairs identification,

- Investigating ML approaches to improve boosted photon pairs identification.
- Under development some calorimetry feature extraction strategies to describe the lateral and longitudinal shower development.
    - Combining different features as input for deep neural networks.





Presentation by Edmar Souza

5. Final Comments

- Strategies using ML for crosstalk investigation ongoing in the context for two PhD theses.
  - Increase of computing resources usage in the last months, mainly in the data preparation and ML model tunings.
  - Will take advantage the output for this studies to contribute in LL/ALP analysis.
  - ATLAS QT ongoing in LAr community.
  - Some analysis to do in the next months, for fine tuning optimization.

- LL/ALP search (UEH - Long-lived ALP to photons - **ANA-EXOT-2023-04**)
  - Search for displaced, pointing and collimated photons that come from the decay of a long-lived particle.
  - Under development match strategy based on MC production.
  - Some features implemented in Athena Rel24.
    - Planning to include but more high level features in the next months, towards Run3 analysis.
    - ML exercises coming soon, in the context of 1 PhD these.