A Machine Learning Approach to Local Hadronic Calibration in the Atlas Calorimeters

Peter Loch[1]
Department of Physics
University of Arizona
Tucson, Arizona 85721
USA

June 19, 2020

`Version: 1.00`

Introduction

The principal signal for physics analysis from the ATLAS calorimeters are dynamically formed clusters of readout cells with topologically connected signals, the so-called *topo-clusters*. Their formation and general features are discussed in detail in Ref. [1]. In particular, this reference presents a first complete approach to a local hadronic calibration of these clusters using shape and location information reconstructed the spatial distribution of the cells in the cluster and their individual cell energies. The development of this approach started sometime in Spring 2003 and principally concluded at the time physics data taking started in 2010. Since then only a few minor adjustments, mostly concerning details of the cluster formation, splitting and seeding, and some small extensions were introduced to optimize the performance in higher pile-up scenarios and address shortcomings overserved mainly in experimental data. The methodology applied for the local hadronic calibration (often called LCW - local cell weighting calibration) is characterized by the state-of-art computational methods available at the time of development and the then available computing resources for massive reconstruction campaigns of ATLAS data and Monte Carlo (MC) simulations.

Recent developments in both computing architecture and software now allow for the large scale application of complex numerical methods exploring multidimensional variable spaces without e.g. the need of (binned) tabular look-ups. Developing a calibration employing these machine learning (ML) methods is expected to address some of the shortcomings of the classical implementation in use so far. In particular, the non-closure of the LCW calibration at lower cluster energies is of concern for (mostly) jet and jet substructure measurements. Additional advantages of ML approaches are the possible intrinsically

[1] Presently also associated with CERN, Geneva, Siwtzerland

smooth (non-binned) application, thus avoiding some of the observed problems at bin boundaries in the classic approach.

The goal of this project is to develop and configure a ML framework which can (1) improve the topo-cluster classification used in LCW (see section 5.2 of the topo-cluster paper), (2) develop a calibration function which can shift the topo-cluster signal to the expected *true* energy it repesents after application of the ML-based classification, and finally (3) test the option of single step calibration without explicit classification. This note documents some of the project goals and documents software and data formats.

Simulated data

The data represent the response of the ATLAS calorimeters to single neutral (π^0) and charged pions ($\pi^{\pm}$) from full simulations. The pions have energies between 200 MeV and 2 TeV (logarithmically sampled), and pseudorapidities within $-5 < \eta_\pi < 5$ (full acceptance in ATLAS). Each event contains exactly one pion. There is no pile-up added. The pions emerge from the nominal vertex (0,0,0) in the detector frame of reference. The magnetic fields (both solenoid and toroid) are included in the simulations (toroidal field not really relevant for charged pions). The calorimeter response is represented by the topo-clusters. (can be more than one per pion).

Topo-clusters

The topo-clusters are reconstructed and calibrated as described in the paper. The cluster kinematics is represented by two *signal states*, the (basic) electromagnetic (EM) scale signal and the locally calibrated hadronic LCW-scale. All cluster moments are calculated using EM scale signals of the cells in the topo-cluster. By ATLAS default, the topo-clusters are reconstructed with zero mass, such that

$$E_{clus} = \sqrt{p_{x,clus}^2 + p_{y,clus}^2 + p_{z,clus}^2}$$

on both EM and LCW scale.

Data

Structure

The data from the full simulations are stored in ROOT tuples, in a tree called `ClusterTree`. Each row of numbers in the data contains the signal, moments and information related to the composition for one topo-cluster. As the number of topo-clusters per particle N_{clus} can be larger than one, the full calorimeter signal of each particle is associated with $N_{clus} \geq 1$ rows. Each cluster/row has an index $i_{clus} = 0 \ldots N_{clus} - 1$ for reference. The clusters are sorted by descending EM scale energies, yielding the highest energetic topo-cluster to be identified by $i_{clus} = 0$. A full event is built by collecting the N_{clus} clusters belonging to the given source particle.

Additional data in each row contain the kinematics and particle type information for the pion. As N_{clus} rows are associated with the same pion, this information is repeated in each row that belongs to this pion. The same is true for other data like the event number (a sequence number assigned during the simulation) and a run number (typically fixed, also assigned at simulation time). These two identifiers are probably not really needed.

The one-row-per-cluster storage model allows to store the data using only trivial types. All non-integral numbers are stored in single precision (`float`) to save disk space. Integral numbers are correspondingly stored as `int`. Table 1 contains the description of the data content in terms of leafs in the `ROOT` tree (column names). It also relates the leaf name with the variable symbol and references to the description in the paper, if applicable/available.

Table 1: Tuple structure. There are additional signal quality related moments which for now are not included in the table. The tree name is `ClusterTree`.

ROOT leaf name	Variable/Formula Expression	Comment/Reference
Event Properties		
`runNumber`	n/a	Not used or needed
`eventNumber`	n/a	Not used or needed
Truth Particle Properties		
`truthE`	E_{π} [GeV]	True pion energy
`truthPt`	$p_{T,\pi}$ [GeV]	True pion transverse momentum
`truthEta`	η_{π}	True pion pseudo-rapidity
`truthPhi`	φ_{π}	True pion azimuth
`truthPDG`	n/a	PDG particle code identifies incoming particle: $PDG(\pi^0) = 111$, $PDG(\pi^+) = 211$, $PDG(\pi^-) = -211$
Topo-cluster Properties & Composition		
`nCluster`	N_{clus}	Number of clusters/pion
`clusterIndex`	i_{clus}	Index of cluster for pion
`cluster_nCells`	$N_{cell,clus}^{E>0}$	Number of cells in cluster with $E > 0$
`cluster_nCells_tot`	$N_{cells,clus}$	Total number of cells in cluster
Topo-cluster Kinematics (EM and LCW Scales)		
`clusterECalib`	E_{clus}^{LCW} [GeV]	Calibrated (LCW) cluster energy [section 5.6]
`clusterPtCalib`	$p_{T,clus}^{LCW}$ [GeV]	Calibrated (LCW) cluster transverse momentum [section 5.6]
`clusterEtaCalib`	η_{clus}^{LCW}	Cluster pseudorapidity after LCW calibration [section 5.3, section 5.6]
`clusterPhiCalib`	φ_{clus}^{LCW}	Cluster azimuth after LCW calibration [section 5.3, section 5.6]
`cluster_sumCellECalib`	$\sum_{\{i\|E_{cell,i}^{EM}>0\}}^{N_{cells,clus}} w_{cell,i}^{LCW} \times E_{cell,i}^{EM}$ [GeV]	Sum of calibrated cell energies for cells with energy $E_{cell}^{EM} > 0$
`clusterE`	E_{clus}^{EM} [GeV]	Cluster EM scale energy (basic signal) [section 3.2,eq.(11)]
`clusterPt`	$p_{T,clus}^{EM}$ [GeV]	Cluster EM scale transverse momentum (basic signal scale)

		[section 3.2,from eq.(12)]
clusterEta	η_{clus}	Cluster pseudorapidity (basic signal scale) [section 3.2,eq.(9)]
clusterPhi	φ_{clus}	Cluster azimuth (basic signal scale) [section 3.2,eq.(10)]
cluster_sumCellE	$\sum_{\{i\|E^{EM}_{cell,i}>0\}}^{N_{cells,clus}} w^{geo}_{cell,i} \times E^{EM}_{cell,i}$ [GeV]	Sum of EM scale energies for cells with energy $E^{EM}_{cell} > 0$
Topo-cluster Calibration		
cluster_EM_PROBABILITY	$P^{EM}_{clus} \in [0,1]$	Probability for cluster to be generated by EM shower [section 5.1;section 5.2,eq.(31)]
cluster_HAD_WEIGHT	$P^{EM}_{clus} \times w^{em-cal}_{clus} + (1 - P^{EM}_{clus}) \times w^{had-cal}_{clus}$ with $w^{em(had)-cal}_{clus} = E^{em(had)-cal}_{clus} / E^{EM}_{clus}$	Effective cluster signal weight after hadronic cell weighting calibration [section 5.3;section 5.6,tab.(3)]
cluster_OOC_WEIGHT	$P^{EM}_{clus} \times w^{em-ooc}_{clus} + (1 - P^{EM}_{clus}) \times w^{had-ooc}_{clus}$ with $w^{em(had)-ooc}_{clus} = 1 + E^{em(had)-ooc}_{clus} / E^{em(had)-cal}_{clus}$	Effective cluster signal weight after out-of-cluster correction [section 5.4;section 5.6,tab.(3)]
cluster_DM_WEIGHT	$P^{EM}_{clus} \times w^{em-dm}_{clus} + (1 - P^{EM}_{clus}) \times w^{had-dm}_{clus}$ with $w^{em(had)-dm}_{clus} = 1 + E^{em(had)-dm}_{clus} / E^{em(had)-dm}_{clus}$	Effective cluster signal weight after dead material correction [section 5.5;section 5.6,tab.(3)]
Topo-cluster Truth Expectations		
cluster_ENG_CALIB_TOT	E^{dep}_{clus} [GeV]	Deposited energy in cluster (sum of deposited energies in cluster cells) [section 5.3;eq.(33)]
cluster_ENG_CALIB_OUT_T	E^{ooc}_{clus} [GeV]	Energy deposited in cells outside of the cluster but associated with it[section 5.4;eq.(36)]
cluster_ENG_CALIB_DEAD_TOT	E^{dm}_{clus} [GeV]	Energy deposited in dead material around the cluster [section 5.5]
Topo-cluster Moments: Locations & Directions		
cluster_CENTER_X	x_{clus} [mm]	Cluster center of gravity x coordinate (detector frame of reference) [section 4.1.1]
cluster_CENTER_Y	y_{clus} [mm]	Cluster center of gravity y coordinate (detector frame of reference) [section 4.1.1]

cluster_CENTER_Z	z_{clus} [mm]	Cluster center of gravity z coordinate (detector frame of reference) [section 4.1.1]
cluster_CENTER_MAG	$c = \sqrt{x_{clus}^2 + y_{clus}^2 + z_{clus}^2}$ [mm]	Distance of cluster from nominal vertex (linear space, detector frame of reference) [section 4.1.1,Fig.8]
cluster_CENTER_LAMBDA	λ_{clus} [mm]	Distance of cluster from calorimeter frontface (measured along the principal shower axis) [section 4.1.1]
cluster_FIRST_PHI	$<\varphi>$ [rad]	Energy-weighted first moment of cell φ distribution in cluster [section 4.1.2]
cluster_FIRST_ETA	$<\eta>$	Energy-weighted first moment of cell η distribution in cluster [section 4.1.2]
cluster_DELTA_PHI	$\Delta\varphi$ [rad]	Azimuthal distance of principal shower axis in cluster with respect to its direction from the nominal vertex [section 4.1.2,Fig.8]
cluster_DELTA_THETA	$\Delta\theta$ [rad]	Polar angle distance of principal shower axis in cluster with respect to its direction from the nominal vertex [section 4.1.2,Fig.8]
cluster_DELTA_ALPHA	$\Delta\alpha$ [rad]	Angular distance between principal shower axis in cluster and its direction from the nominal vertex[section 4.1.2,Fig.8]
Topo-cluster Moments: Spatial Extensions & Shapes		
cluster_SECOND_R	$<r^2>$ [mm^2]	Second moment of radial distances of cells to the principal cluster axis [section 4.1.3,eq.16]
cluster_SECOND_LAMBDA	$<\lambda^2>$ [mm^2]	Second moment of distances of cells from cluster center along the principal axis (longitudinal) [section 4.1.3,eq.17]
cluster_LATERAL	$<m_{lat}^2>$	Measure of lateral energy dispersion

		[section 4.1.3,eq.18]
cluster_LONGITUDINAL	$< m^2_{long} >$	Measure of longitudinal energy dispersion [section 4.1.3,eq.19]
cluster_ISOLATION	f_{iso}	Measure for cluster isolation [section 4.2.5,eq.28]
Topo-cluster Moments: Signal Density & Sharing		
cluster_FIRST_ENG_DENS	$< \rho_{cell} >$ [GeV/mm^3]	Energy-weighted first moment of cell signal density distribution in cluster [section 4.2.2]
cluster_SECOND_ENG_DENS	$< \rho^2_{cell} >$ [(GeV/mm^3)^2]	Energy-weighted second moment of cell density distribution in cluster [section 4.2.2]
cluster_ENG_FRAC_EM	f_{emc}	Fraction of energy in EM calorimeter [section 4.2.4,eq.24]
cluster_ENG_FRAC_MAX	f_{max}	Most energetic cell signal fraction in cluster [section 4.2.4,eq.25]
cluster_ENG_FRAC_CORE	f_{core}	Energy fraction in core of cluster [section 4.2.4,eq.26]
cluster_ENG_POS	$E^{EM}_{clus,pos}$ [GeV]	Sum of $E^{EM}_{cell} > 0$ [section 4.2.4,eq.27]
cluster_SIGNIFICANCE	ζ^{EM}_{clus}	Cluster signal significance [section 4.2.1,eq.22]
cluster_CELL_SIGNIFICANCE	$\max\{\zeta^{EM}_{cell}\}$ in cluster	Highest cell signal significance in cluster [section 3.1,eq.2]
cluster_CELL_SIG_SAMPLING	sampling id	Sampling where cell with highest signal significance is located
Topo-cluster Moments: Additional Shapes		
cluster_PTD	$p_T D = \frac{\sqrt{\sum_{i=1}^{N_{cell,clus}} (E^{EM}_{cell,i})^2}}{\sum_{i=1}^{N_{cell,clus}} E^{EM}_{cell,i}}$	Similar to longitudinal fragmentation function in jets; originally described in [2]
cluster_MASS	m^{EM}_{clus} [GeV]	Cluster mass calculated using cells with $E^{EM}_{cell} > 0$ only

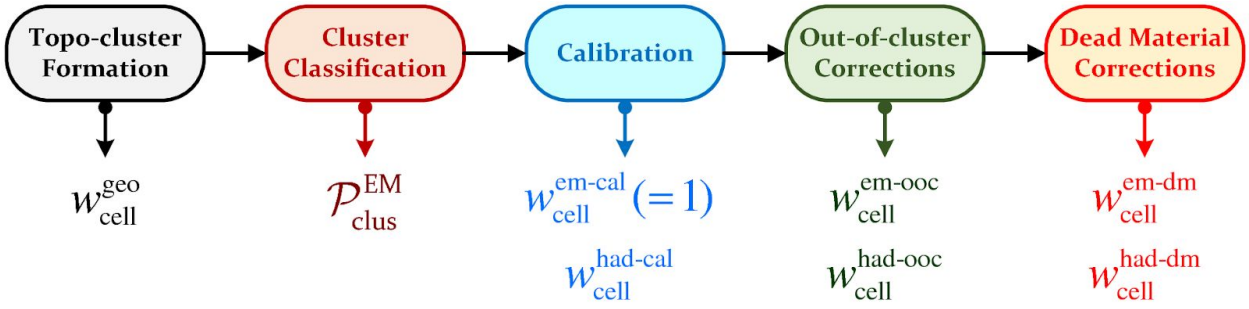
---

Variable content

The variables represented by leafs in the ROOT tree are in many cases identical to the ones introduced in Ref. [1]. In addition, variables expressing the effects of the various calibration and correction steps outlined in that paper are added to the data. The (common) calibration sequence after classification is depicted in Fig. 1 (taken from [1]) . In the LCW procedure all topo-clusters are subject to this sequence. For clusters with $0 < P^{EM}_{clus} < 1$ both the EM and the HAD calibration procedures are applied and the

effective weights used to determine the final energy are the result of a weighted linear interpolation between the two calibrations. The effect of each step of the calibration procedure on the total cluster energy is measured by the ratio of the cluster energy after to the energy before its application. These ratios are stored in the data, see Table 1.

Figure 1: Topo-cluster calibration sequence (from [1]). Note that this schematic shows the nomenclature for the cell signal weights after each step of the procedure, while the text discusses the corresponding cluster-level weights.



The following variables are employed in the respective calculations:

EM calibration (none)	$w_{clus}^{em-cal} = E_{clus}^{em-cal}/E_{clus}^{EM} = 1$
HAD calibration (cell signal weighting)	$w_{clus}^{had-cal} = E_{clus}^{had-cal}/E_{clus}^{EM} = \frac{\sum_{i=1}^{N_{cell,clus}} w_{cell,i}^{had-cal} \times w_{cell,i}^{geo} \times E_{cell,i}^{EM}}{\sum_{i=1}^{N_{cell,clus}} w_{cell,i}^{geo} \times E_{cell,i}^{EM}}$
EM OOC	$w_{clus}^{em-ooc} = E_{clus}^{em-ooc}/E_{clus}^{em-cal}$
HAD OOC	$w_{clus}^{had-ooc} = E_{clus}^{had-ooc}/E_{clus}^{had-cal}$
EM DM	$w_{clus}^{em-dm} = E_{clus}^{em-dm}/E_{clus}^{em-ooc}$
HAD DM	$w_{clus}^{had-dm} = E_{clus}^{had-dm}/E_{clus}^{had-ooc}$

The basic cluster signal $E_{clus}^{EM} = \sum_{i=1}^{N_{cell,clus}} w_{cell,i}^{geo} \times E_{cell,i}^{EM}$ is at the EM scale, and an immediate result of the ter topo-cluster formation. Here w_{cell}^{geo} is the geometric weight of the cell signal contribution to the cluster and E_{cell}^{EM} is the basic cell signal on EM scale. Following the sequence shown in Fig. 1, the application of the EM and HAD calibrations yields the ratio $w_{clus}^{em-cal} = 1$ and $w_{clus}^{had-cal} \geq 1$ (typically) at cluster level. The corresponding signals are then E_{clus}^{em-cal} and $E_{clus}^{had-cal}$, respectively.

At the next step the out-of-cluster corrections are applied. The cluster-level effective ratio of the signal after and before are w_{clus}^{em-ooc} and $w_{clus}^{had-ooc}$. The final step of the LCW procedure is the dead material correction yielding weights w_{clus}^{em-dm} and w_{clus}^{had-dm}. The corresponding evolution of the topo-cluster signal can be schematically summarized:

$$E_{clus}^{EM} \rightarrow P_{clus}^{EM} \times E_{clus}^{em-cal} + (1 - P_{clus}^{EM}) \times E_{clus}^{had-cal} \rightarrow$$
$$P_{clus}^{EM} \times E_{clus}^{em-ooc} + (1 - P_{clus}^{EM}) \times E_{clus}^{had-ooc} \rightarrow$$
$$P_{clus}^{EM} \times E_{clus}^{em-dm} + (1 - P_{clus}^{EM}) \times E_{clus}^{had-dm} \rightarrow E_{clus}^{LCW}$$

Issues with data

June 19, 2020 The initial production has a repetition in the data. The `clusterIndex` leaf has been booked twice (branches #8 and #11), with exactly the same data content. This does not affect the validity of the data, but could lead to confusions concerning the unpacking. A new production will fix this issue.

Working with the data – some hints

Fractional reconstruction validation

The data in the `ROOT` tuple can be used to measure the effectiveness of the individual calibration and correction steps implemented in LCW. The code fragments can be used to compare of the cluster signal after hadronic calibration only with the energy deposited in the cluster. Similarly, the out-of-cluster energy losses and the dead material losses, both estimated by LCW, can be evaluated using the respective truth information:[2]

```
// collect weights
double whad(cluster_HAD_WEIGHT);
double wooc(cluster_OOC_WEIGHT);
double wdmc(cluster_DM_WEIGHT);
// collect reconstructed energies
double eemc(clusterE);
double elcw(clusterECalib);
// calculate the fractional energies
double ehad(whad*eemc);              // hadronic calibration applied
double eooc((wooc-1.)*ehad);         // reconstructed out-of-cluster energy
double edmc((wdmc-1.)*(eooc+ehad)); // reconstructed dead material losses
// collect calibrated energies
double ecalhad(cluster_ENG_CALIB_TOT);      // true energy deposit in cluster
double ecalooc(cluster_ENG_CALIB_OUT_T);    // true out-of-cluster energy
```

[2] This is code appropriate for a `ROOT`-based analysis of the tuple. It should be implemented in the Loop method. Code templates (.h and .C files) are provided by `ClusterTree::MakeClass()`.

```
double ecaldmc(cluster_ENG_CALIB_DEAD_TOT); // true dead material loss
// total true energy assigned to cluster (LCW calibration target)
double ecaltot(ecalhad+ecalooc+ecaldmc);    // calibration reference
```

In this example the following rule should hold (good test to make sure the fractional energies are correctly reconstructed):

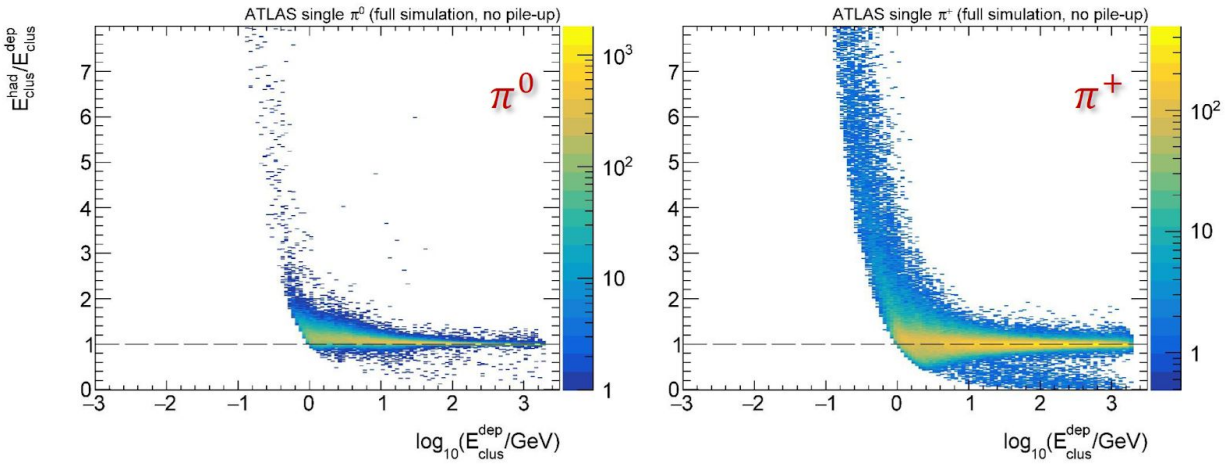
```
(ehad+eooc+edmc)/elcw == 1.; // always!!!!
```

Validation of the fractional energies with truth is done using the following ratios — all of which should yield 1 on average, but unfortunately they do not:

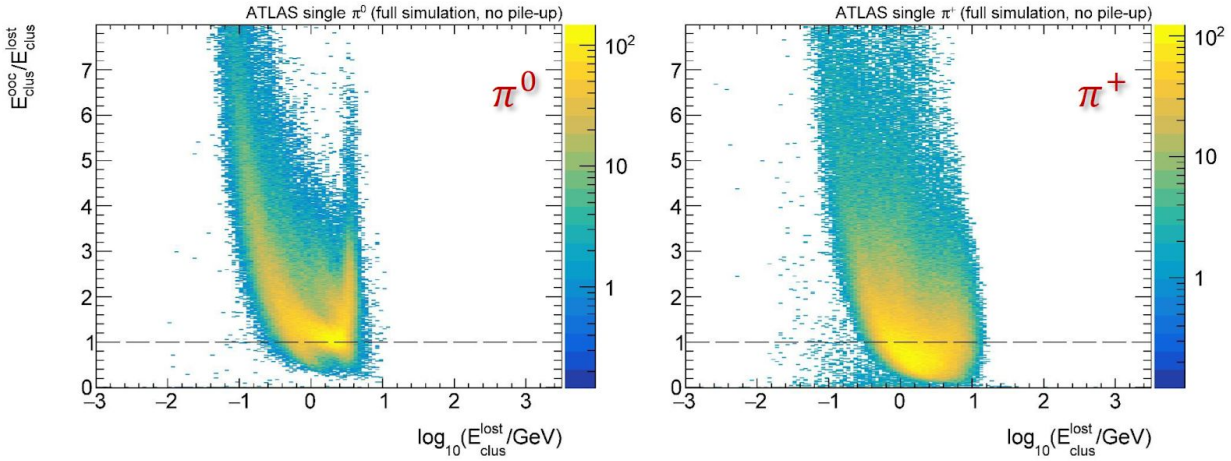
```
double fhad(ehad/ecalhad);
double fooc(eooc/ecalooc);
double fdmc(edmc/ecaldmc);
```

Some initial findings already presented at the US ATLAS Hadronic Final State Forum 2019 (University of Chicago):

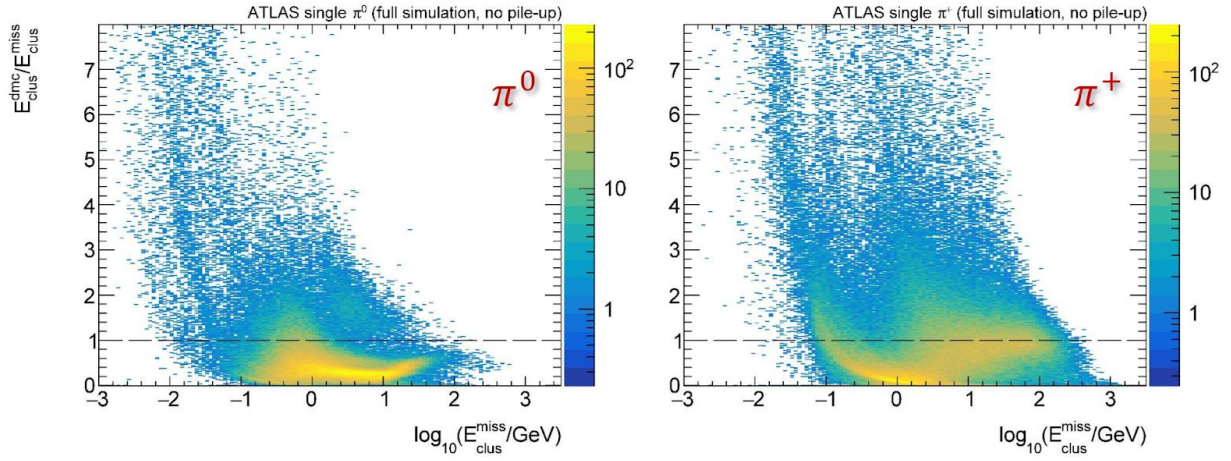
Hadronic calibration only (y-axis value `fhad` should be 1 on average for perfect calibration)



Out-of-cluster correction (y-axis value `fooc` should be 1 on average for perfect correction)



Dead material corrections (y-axis value `fdmc` should be 1 on average for perfect correction)



Bibliography

[1] ATLAS Collaboration, *Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1*, Eur.Phys.J.C 77 (2017) 490

[2] CMS Collaboration., *Performance of quark/gluon separation using pp collision data at* $\sqrt{s}$ *= 8 TeV*, CMS PAS JME-13-002 (2013)