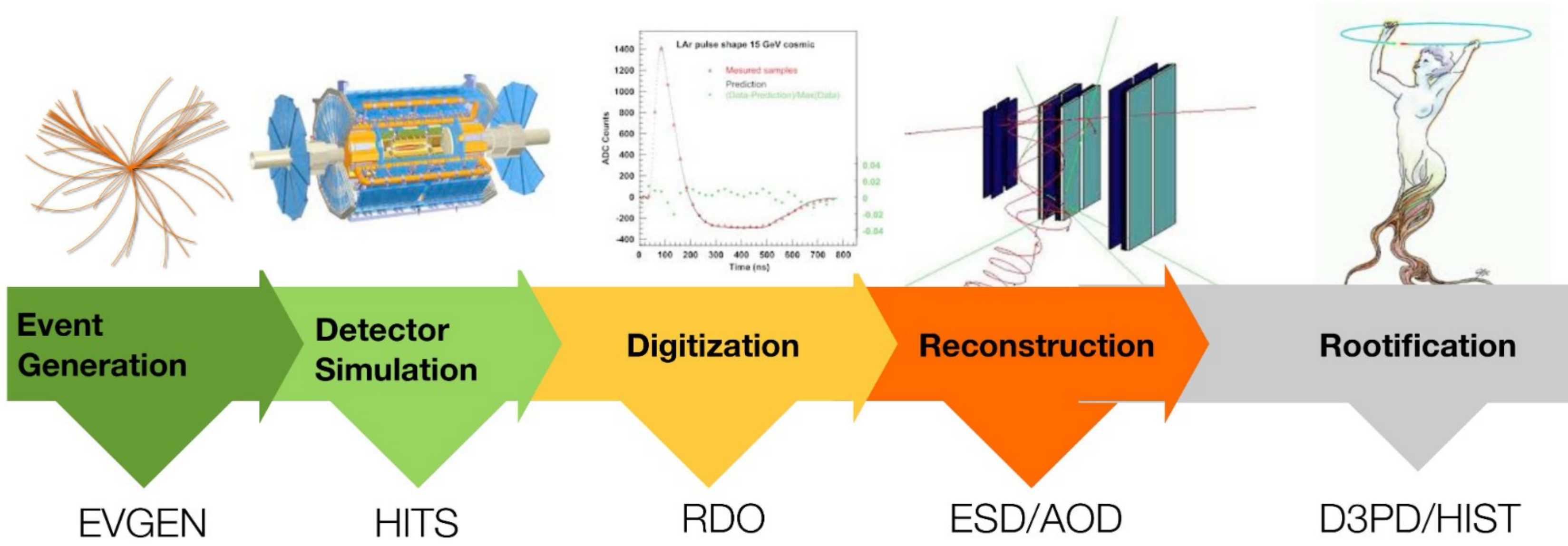
ATLAS
Fast Calorimeter Simulation
(FastCaloSim)

Jana Schaarschmidt (University of Washington)

IRIS HEP Meeting 31. March 2021

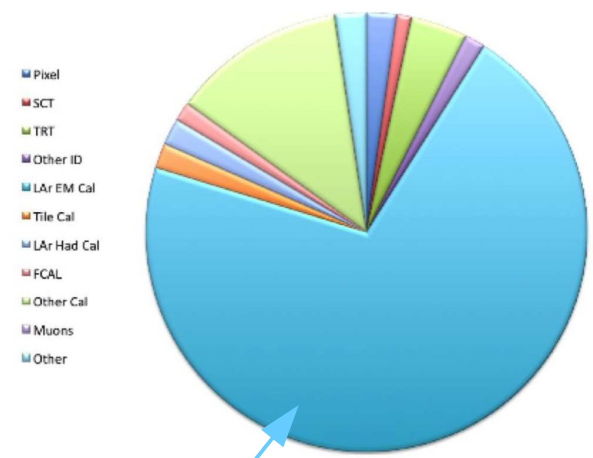
Motivation



Sample	Generation	Simulation	Digitization	Reconstruction
Minimum Bias	0.0267	551.	19.6	8.06
$t\bar{t}$ Production	0.226	1990	29.1	47.4
Jets	0.0457	2640	29.2	78.4
Photon and jets	0.0431	2850	25.3	44.7
$W^{\pm} \to e^{\pm}\nu_e$	0.0788	1150	23.5	8.07
$W^{\pm} \to \mu^{\pm}\nu_\mu$	0.0768	1030	23.1	13.6
Heavy ion	2.08	56,000	267	-

EPJ C 70 (2010) 823 (2010)

ttbar simulation:



(Z. Marshall)

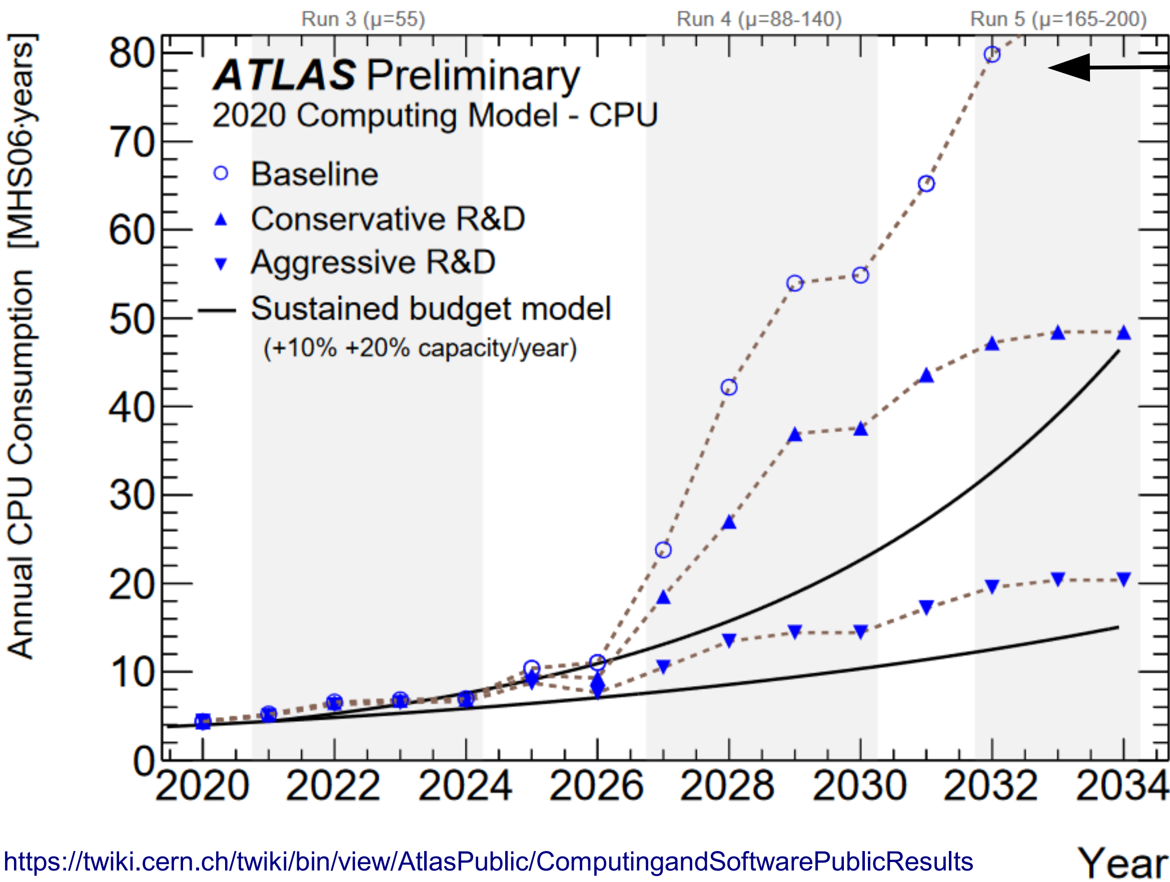
The bulk of the simulation time is spent in the EM calo

→ Replacing the calorimeter simulation is therefore the first priority for an efficient fast simulation approach

Fast Simulation in ATLAS

- A fast calorimeter simulation (FastCaloSim) is used in ATLAS since 2010 (arXiv:1005.4568)
- The tool that combines FastCaloSim with Geant4 in the rest of the detector is called AF2 (AtlFastII)
- About 50% of all simulations in ATLAS were done with AF2, and it was used in countless publications

  → It is a big challenge to improve over this already existing tool!
- Fast Simulation is a paramount component for computing plans for HL-LHC



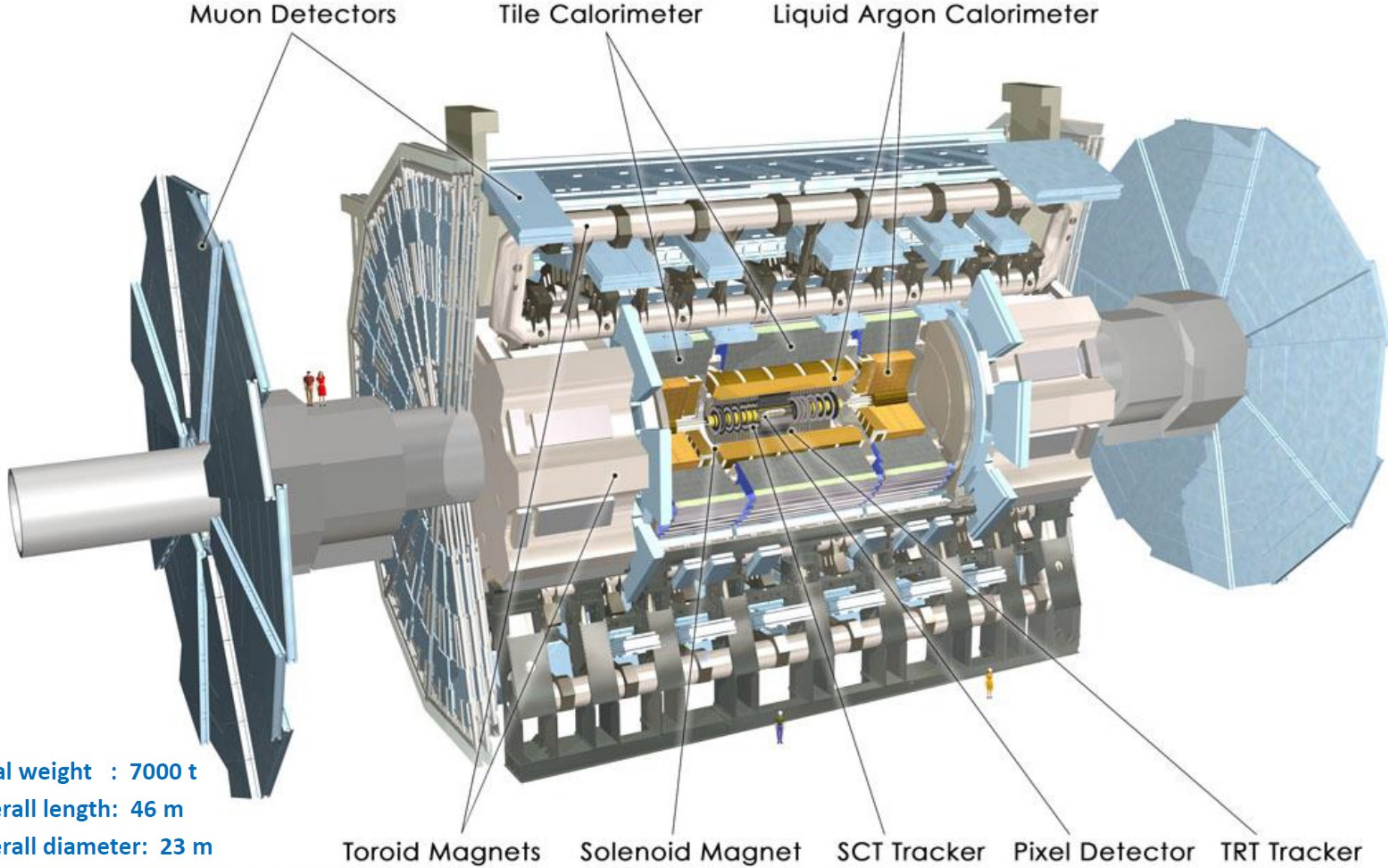
https://twiki.cern.ch/twiki/bin/view/AtlasPublic/ComputingandSoftwarePublicResults

Baseline:
60% FastCaloSim
40% FullSim

Conservative:
75% FastCaloSim
25% FullSim
Faster G4
Faster reconstruction

Aggressive:
90% FastCaloSim+Fast Inner Detector Sim
10% FullSim
Faster G4
Faster reconstruction

A multi-purpose detector system



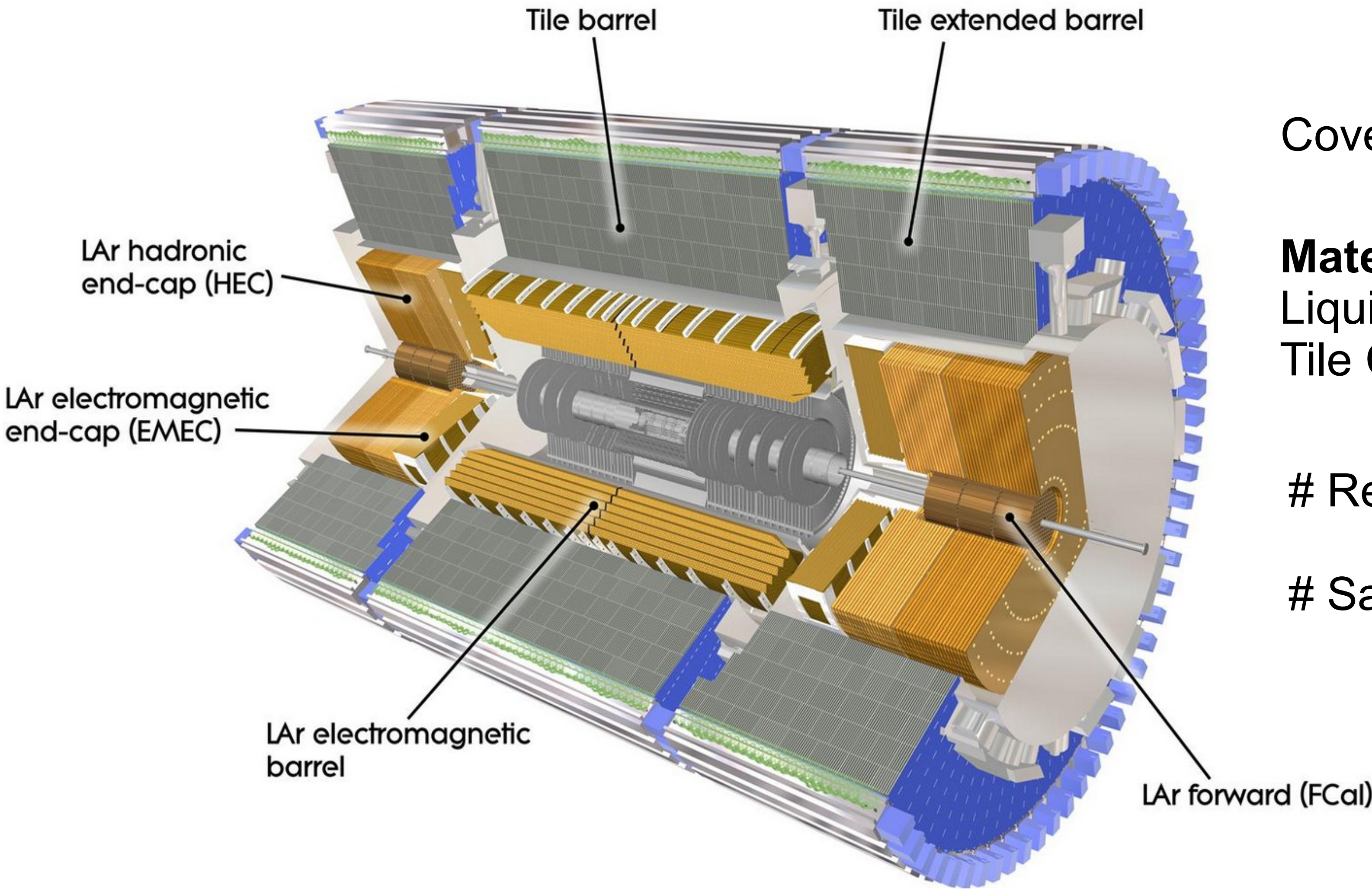
Total weight : 7000 t

Overall length: 46 m

Overall diameter: 23 m

Magnetic field: 2T solenoid + (varying) toroid field

The ATLAS Calorimeter System



Covering $|\eta|<4.9$

Materials:
Liquid Argon + Lead, or copper or tungsten
Tile Cal: Steel + plastic

Readout channels: ~190 k in total

Samplings (layers of cells): 24 in total

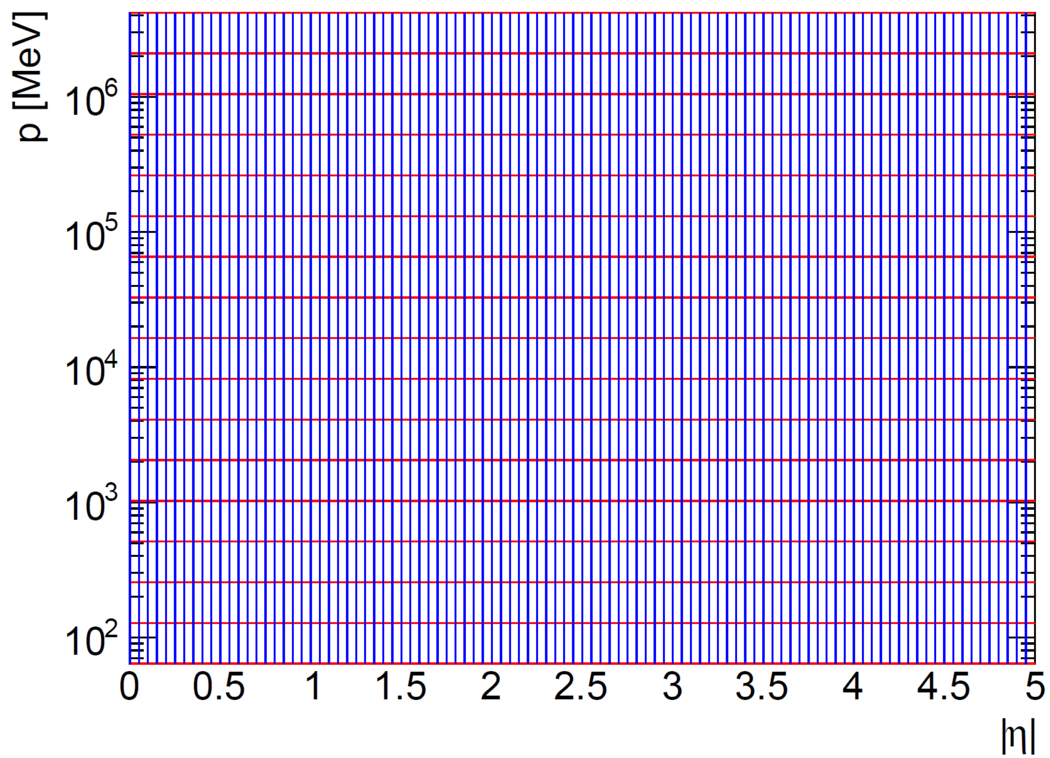
System	EM Barrel	EM EndCap	Hadronic EndCap	Forward (FCAL)	TileCal
#Channels	110k	64k	5.6k	3.5k	9.8k

Crucial for photons & electrons, jets and missing energy reconstruction

FastCaloSim Principle and G4 Input Samples

Parametrized calorimeter energy response of single particles, based on the Geant4 simulation, derived on a fine grid of energy and eta, separated into longitudinal and lateral components.

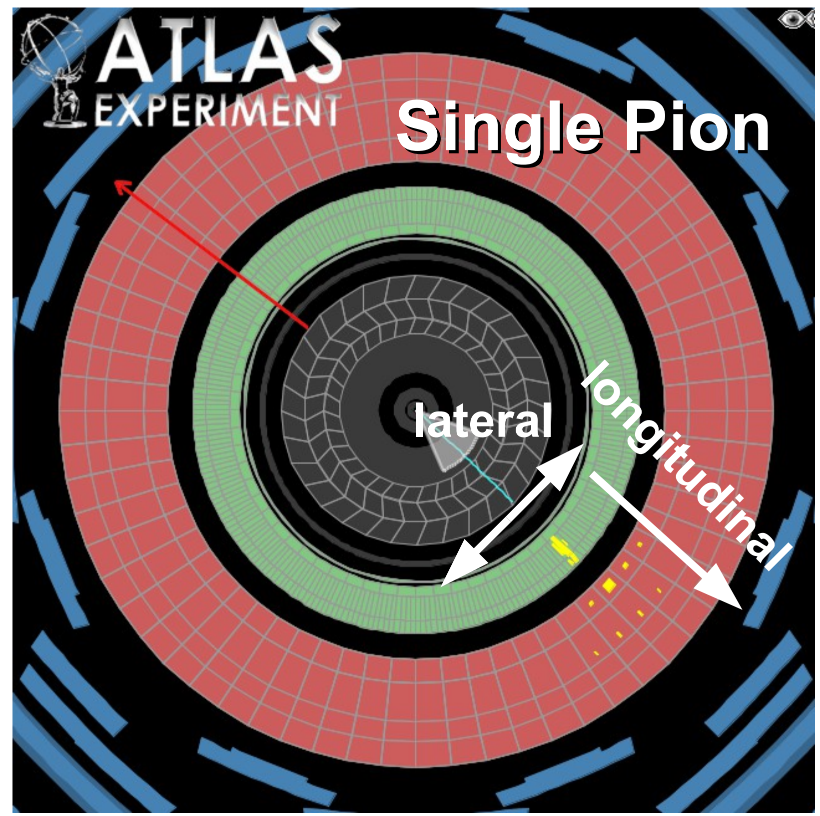
- Eta grid: 100 bins in size of 0.05 covering 0-5.0
- Energy grid: 17 discrete points from 64 MeV – 4.2 TeV (log spacing)



Photons: For the photon showers
Electrons(+/-): For the electron showers
Charged Pions: For all hadronic showers*

*other hadrons simulated to derive corrections

→ 1700 parametrization slices per particle,
5100 parametrization slices in total



Particles generated with the particle gun on the calorimeter surface, no calorimeter noise, no primary vertex smearing, no cross talk, custom G4 hit merging scheme.

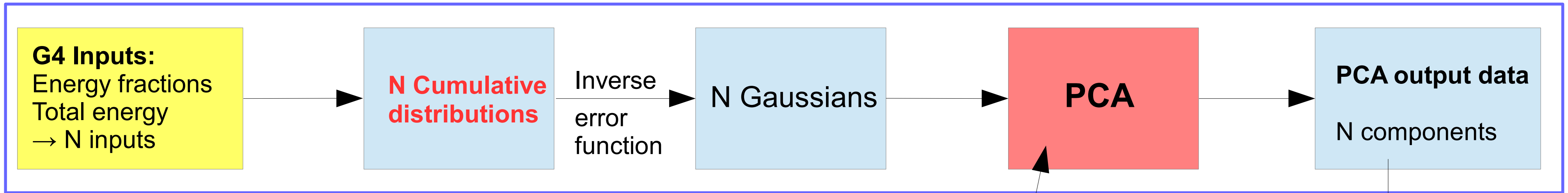
Longitudinal Energy Parametrisation

Energy deposit in each calorimeter layer along the shower axis and total energy

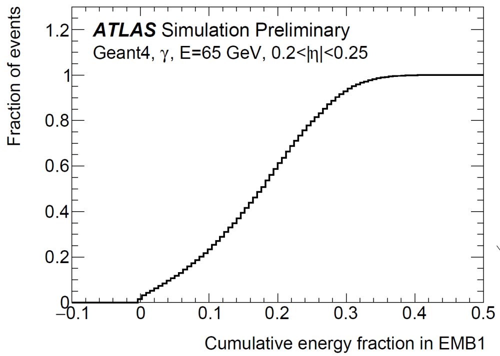
Problem: The energy deposits in the various layers are correlated with each other

Transformation to uncorrelated set of variables with **principal component analysis**, to reduce complexity

1st PCA chain:

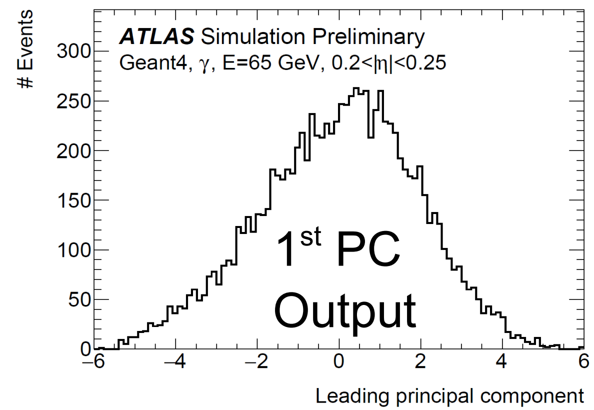


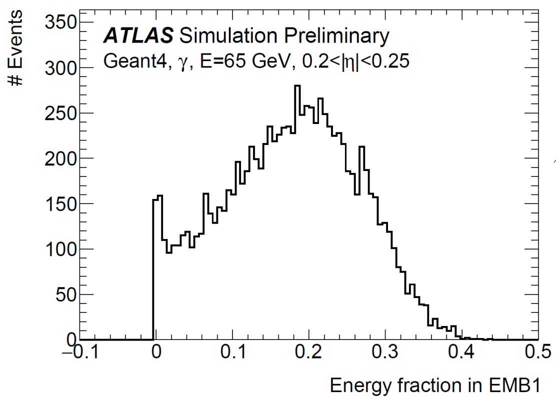
Example:
Photons 50 GeV



PCA performed using
TPrincipal class from Root
https://root.cern.ch/doc/master/classTPrincipal.html

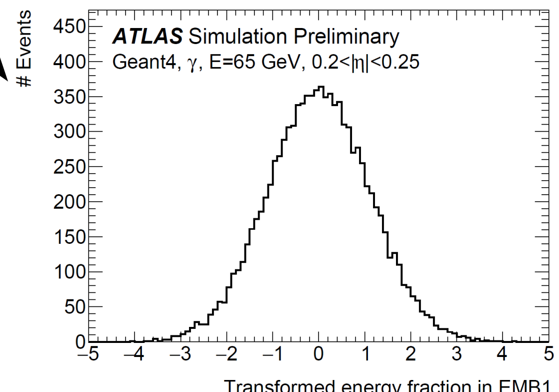
N outputs





Cumulative

Gaussianize



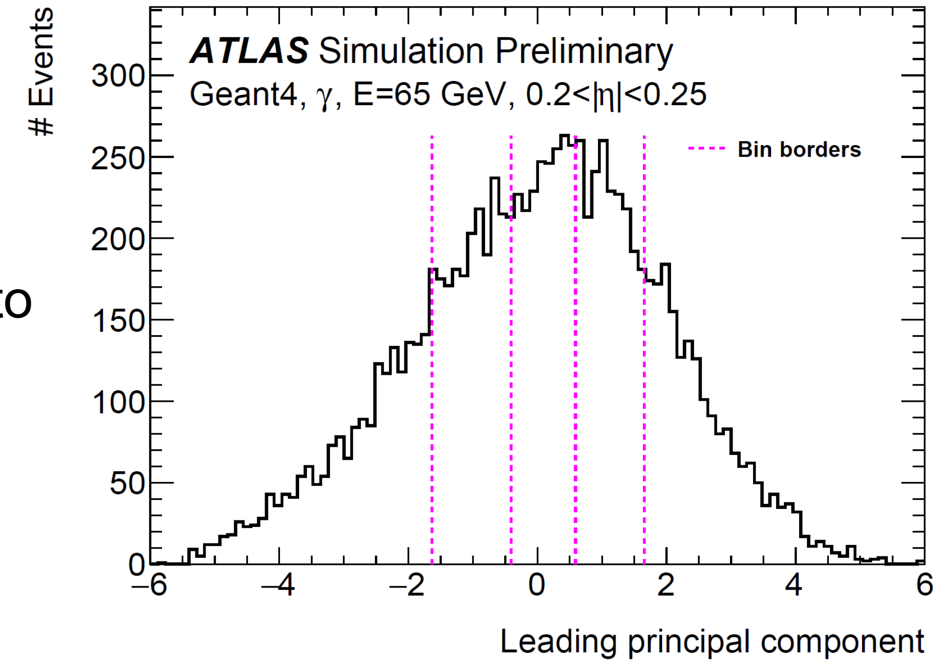
First principle component is that eigenvector of the covariance matrix with the largest eigenvalue (variance)

Longitudinal Energy Parametrisation

The first and second principal component are used to divide the input data into quantiles.

These **“PCA bins”** are also used to derive the shape parametrisation.

The 1st PCA chain is used only to derive this binning.

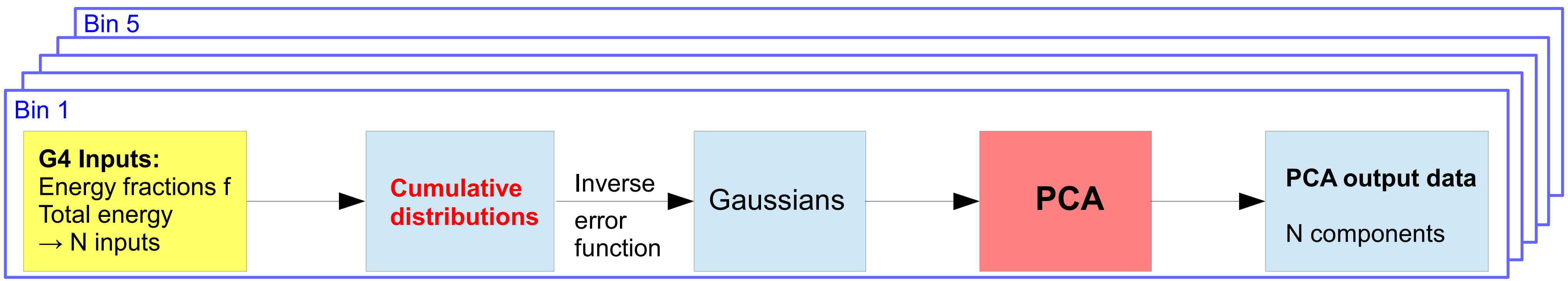


Optional 2nd binning in the 2nd component (not shown)

→ typically 5-10 bins

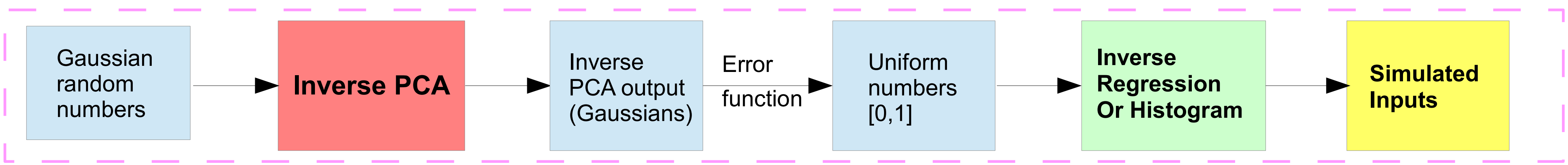
Showers get classified in these „PCA bins“

2nd PCA chain (another PCA, but now in each of the “PCA bins“ from the 1st transformation):



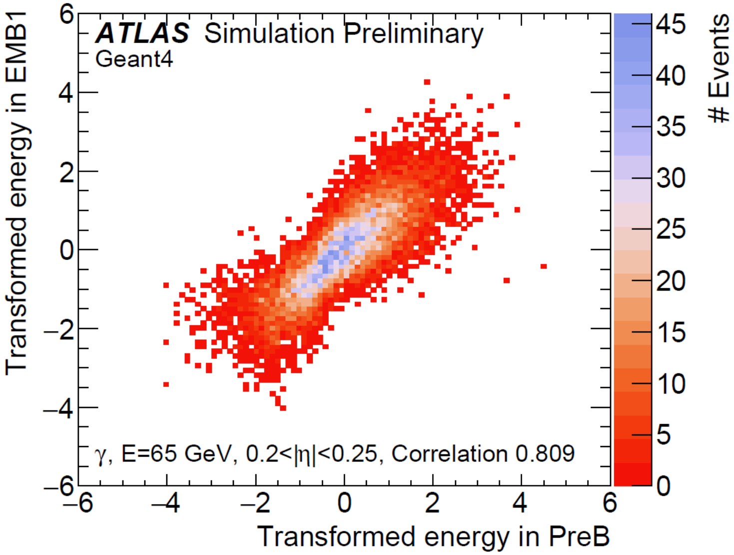
Output is a set of N linearly uncorrelated, Gaussian-like distributions

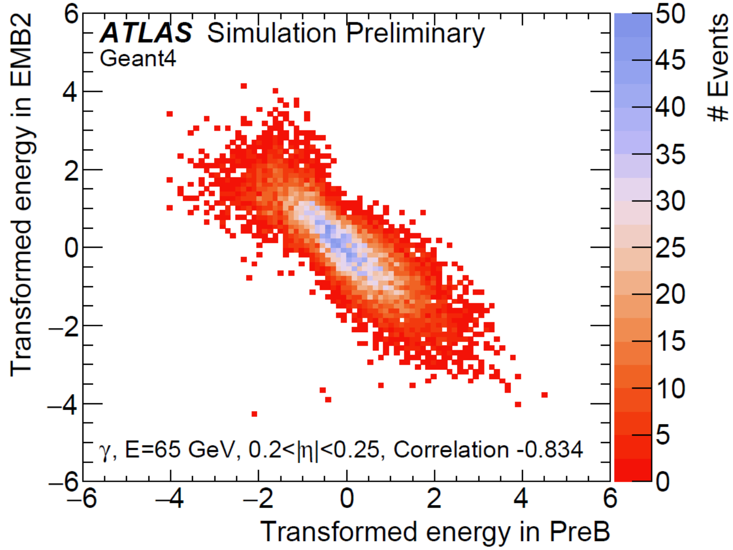
For **simulation** a „PCA bin“ is picked randomly, and then the chain is performed back-wards:

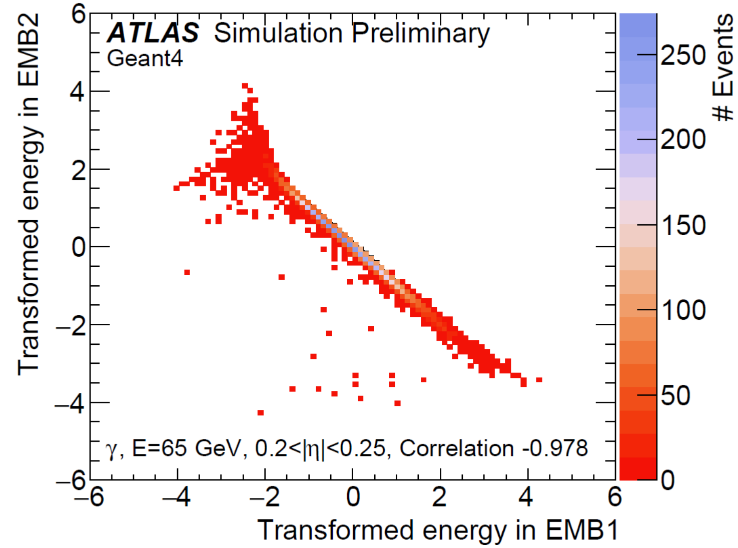


Longitudinal Energy Parametrisation: Decorrelation

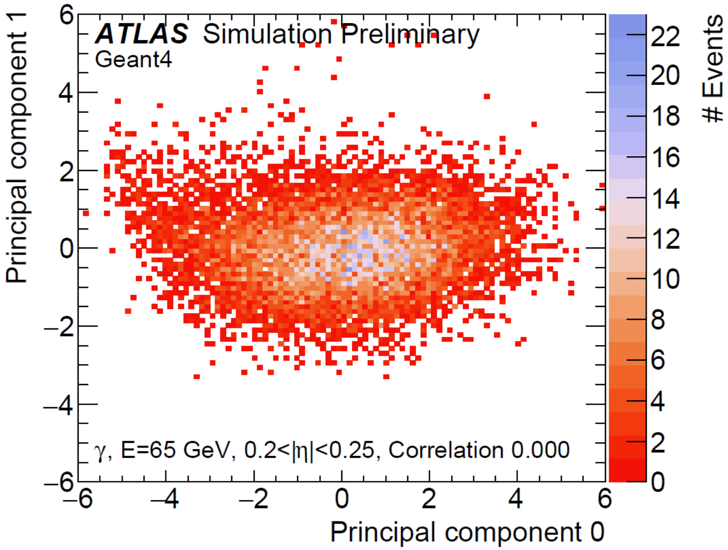
Correlations between energies before PCA rotation, here for 65 GeV photons $0.2<|\eta|<0.25$:

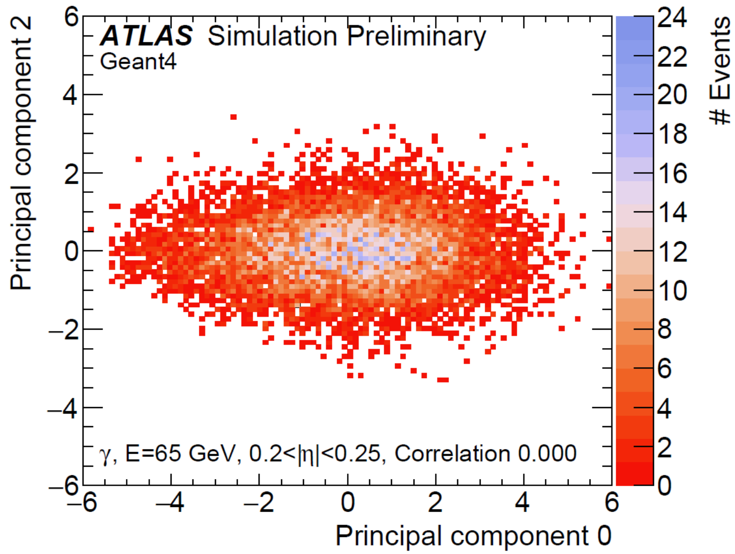


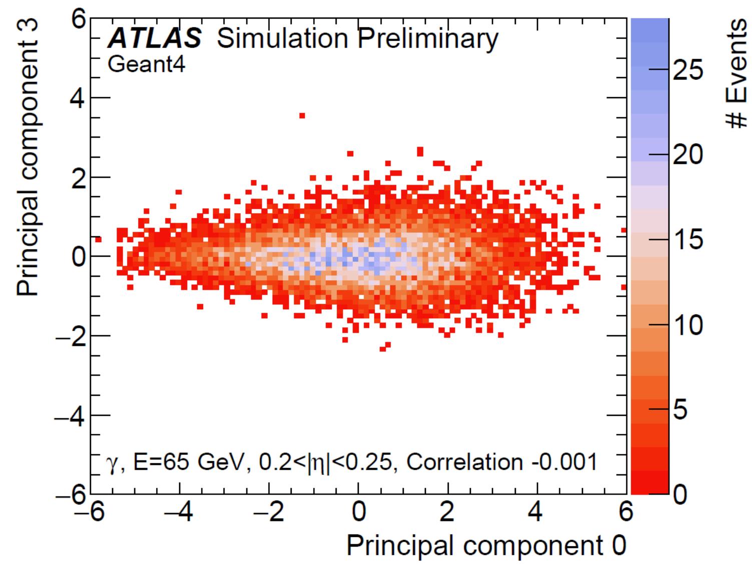




Correlations between energies after PCA rotation:

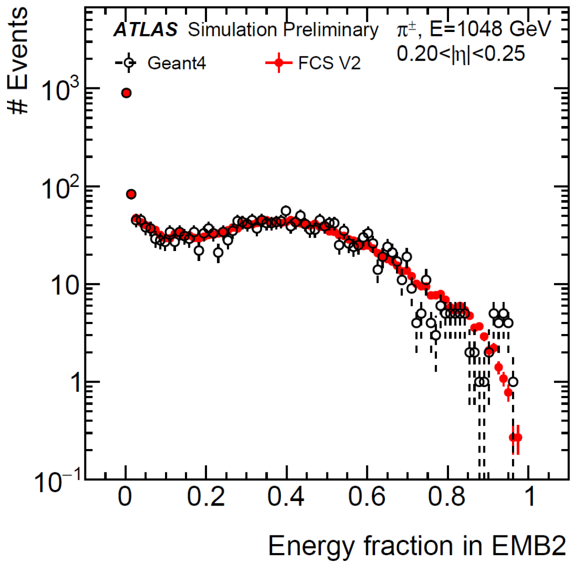


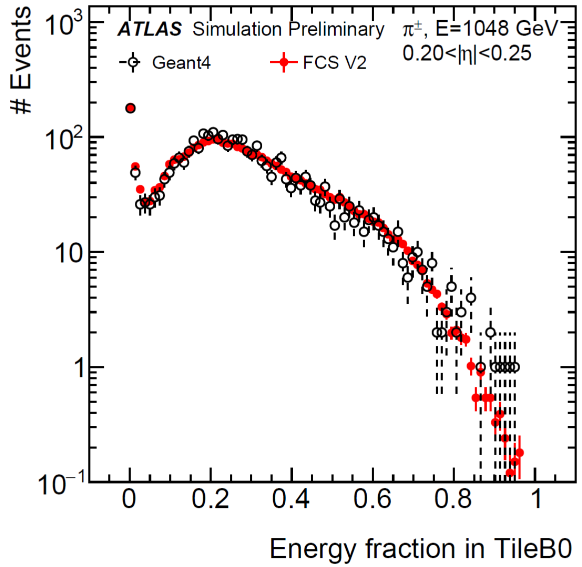


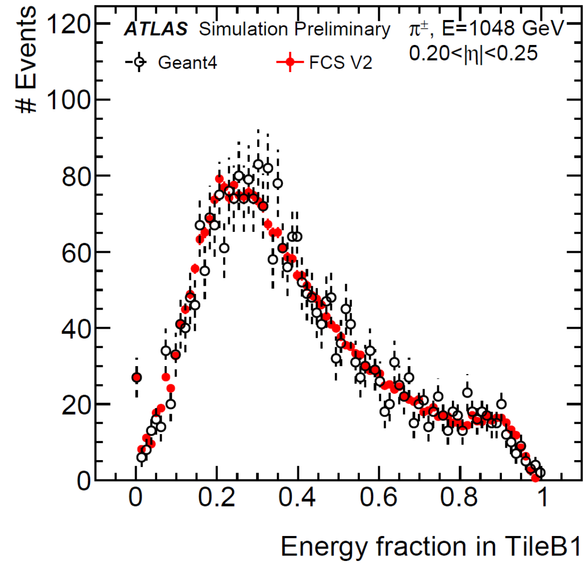


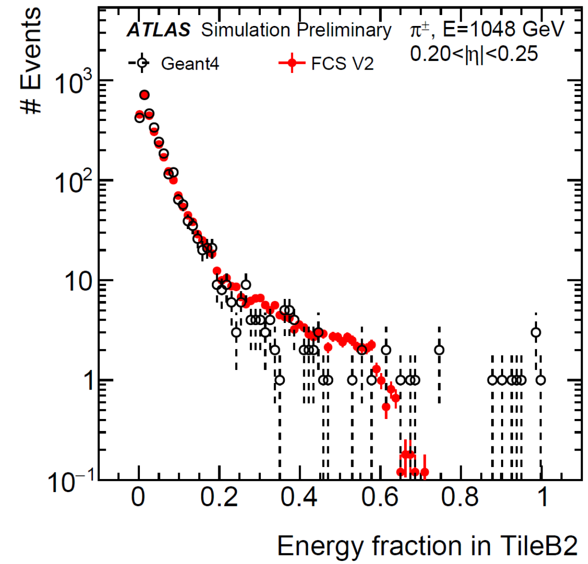
Longitudinal Energy Parametrisation: Validation

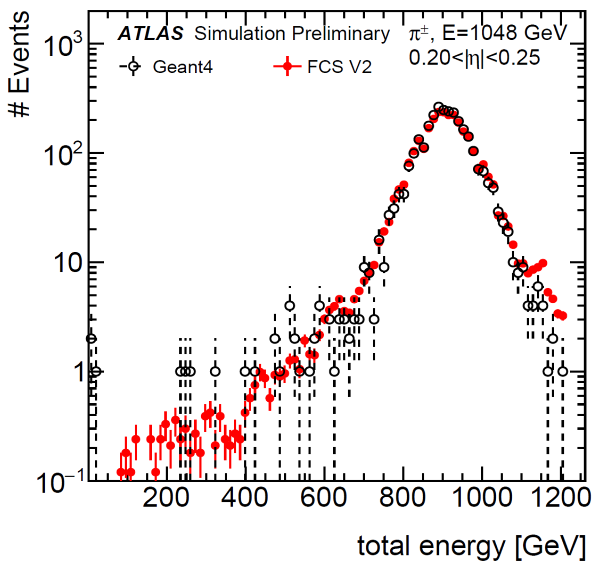
G4
Toy simulation











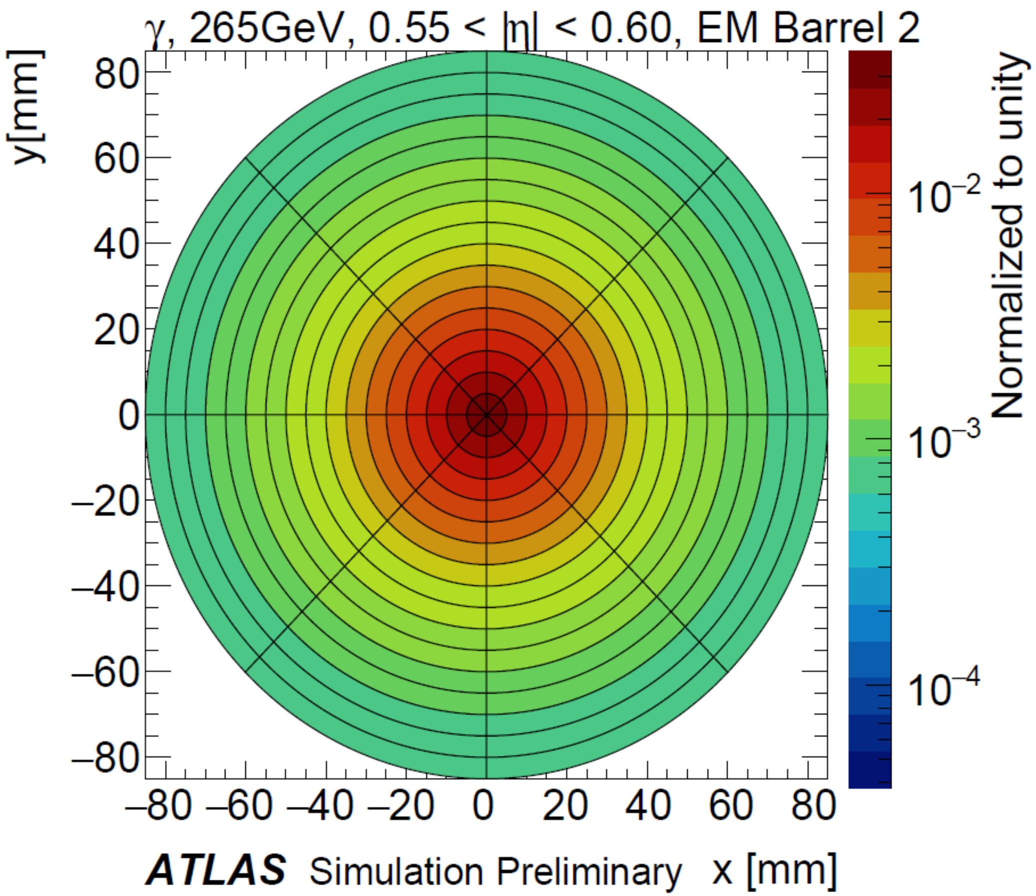
- Energy parametrization efficiently validated with a toy simulation (=performing PCA chain backwards), ie. no object reconstruction

- Plots illustrate one specific parametrization slice: **1 TeV pions, eta 0.2**

  → All energy fractions and total energy very decently modelled!

- 5 PCA bins sufficient for most parametrization slices

Lateral Energy Parametrisation („Shape“)



2D histograms hold hit energy averaged over many showers, integrated over a certain radial distance and 8 bins in the angular direction.

Binning is coarser than the G4 hit granularity, but finer than the calorimeter cell granularity.

We store one such histogram per particle, energy point, eta bin, layer, PCA bin (ie. ~100k histograms).

These 2D histograms are derived during the parametrization step, and then loaded into memory for the simulation step.

2D energy histogram treated like a PDF

→ Randomly sample hit positions from that PDF

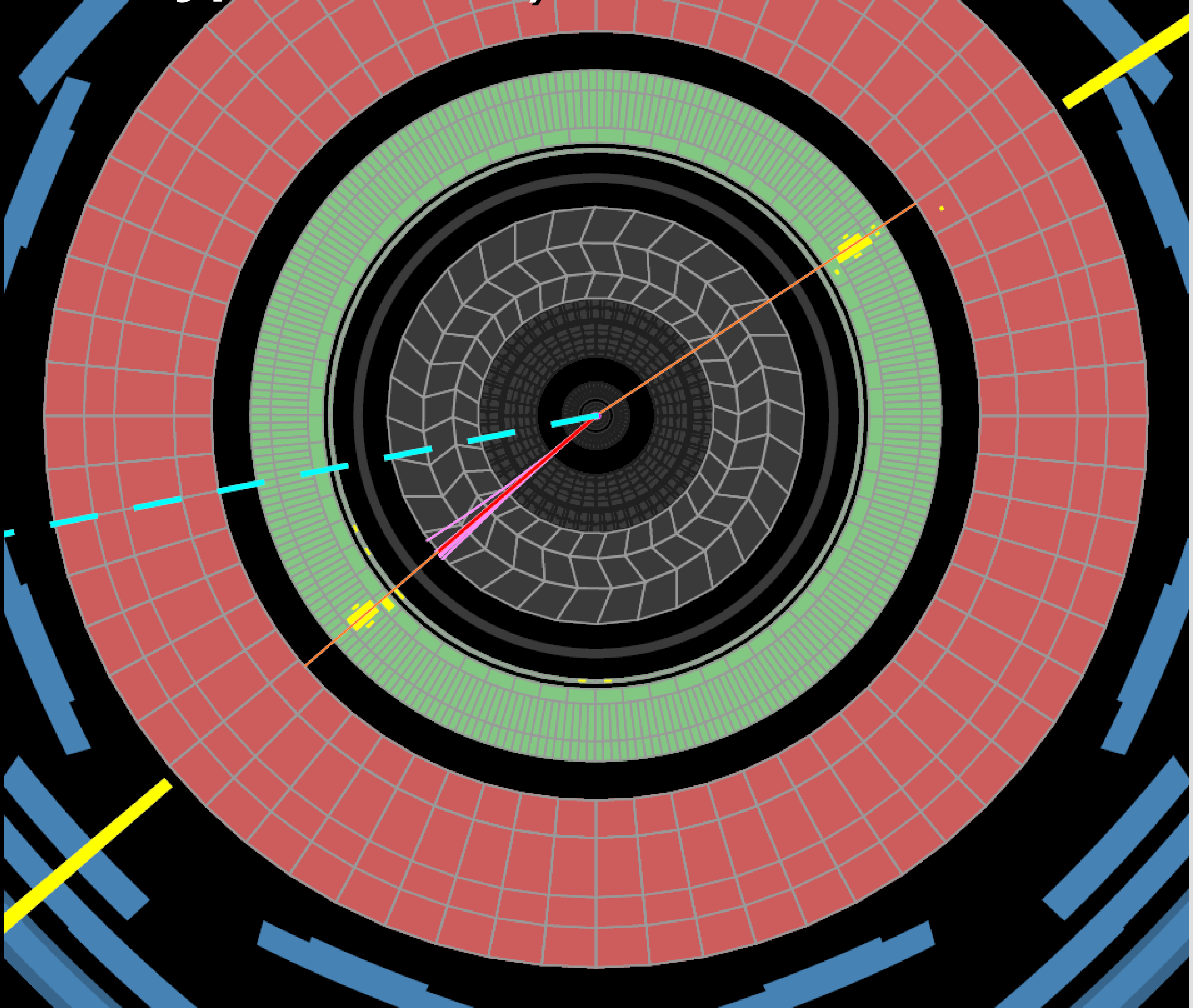
→ Distribute energy to hits*: $E_{hit} = E_{layer} / N_{hits}$

*In a refinement, we weight the hit energies to assure the outer-most tails also get populated with energy

The **number of hits is an important parameter to model energy fluctuations** a good starting point is to calculate it from the expected energy resolution:

$\mathrm{sqrt}(N_{hits})/N_{hits} = \alpha / \mathrm{sqrt}(E)$ (sampling term of the energy resolution, α depends on the layer)

H → γγ simulated with FastCaloSimV2 (Prototype version)



ATLAS EXPERIMENT

FCS V2, H -> γγ MC

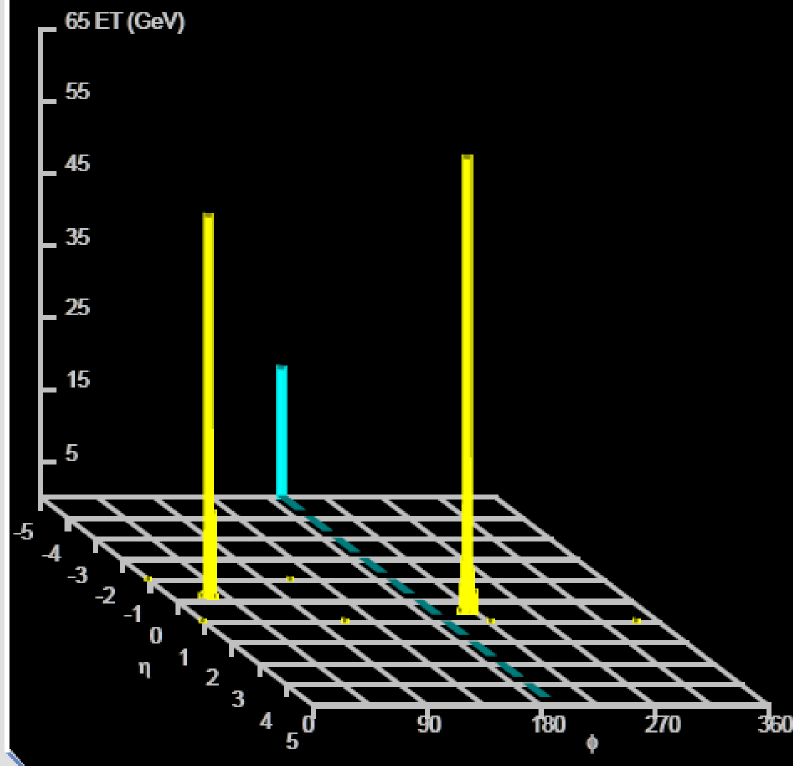
Reconstructed photon

Reconstructed track

MET

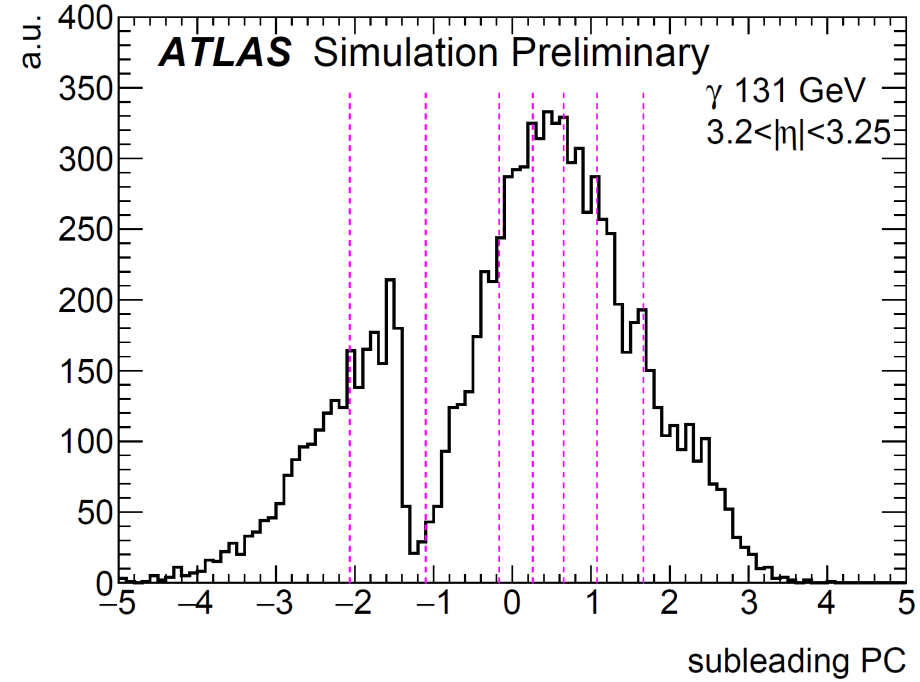
Simulated charged particle

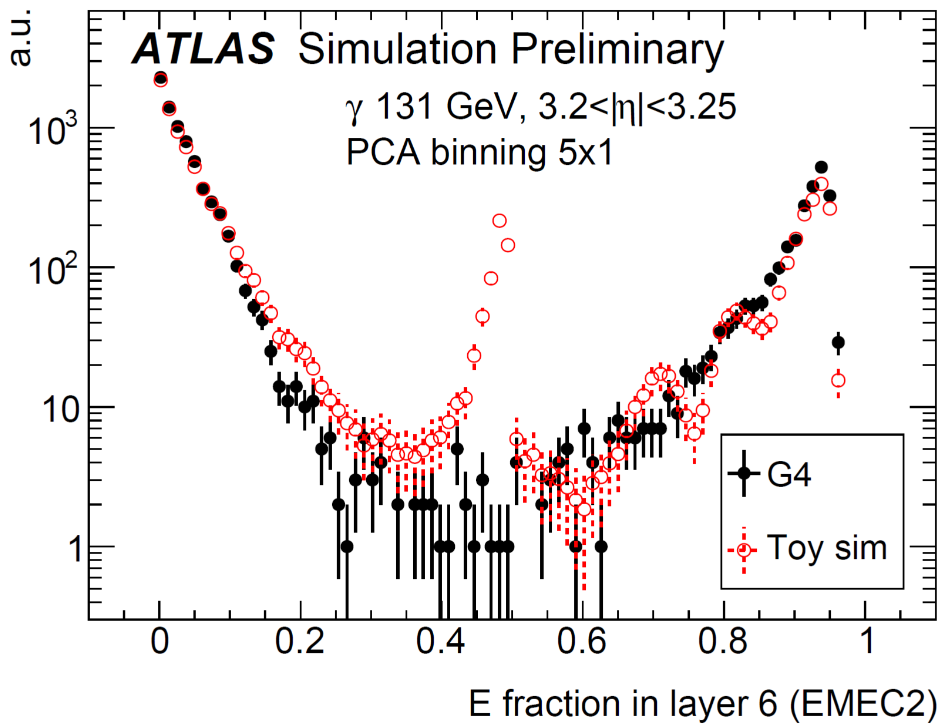
Simulated neutral particle

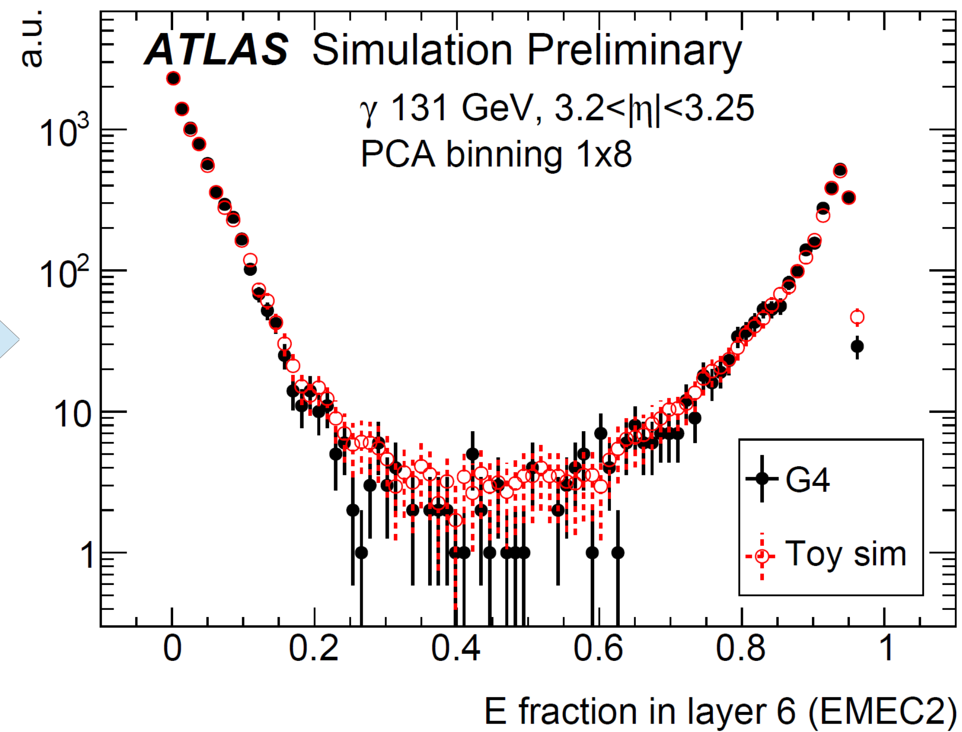


New developments: PCA Binning Optimisation

- Performance of the longitudinal energy parametrisation not great in a few specific slices, namely at eta 0.0, 2.5 and 3.2
- These etas correspond to **transitions of calo material or geometry**, but affect only electrons and photons
- In these cases leading PCs have **non-Gaussian features**
- Developed algorithm that tested many pre-defined PCA binning configurations, best one was selected based on chi2 test of the PCA outputs after 2nd PCA transformation
- **60 parametrization slices** have been optimized and improved that way







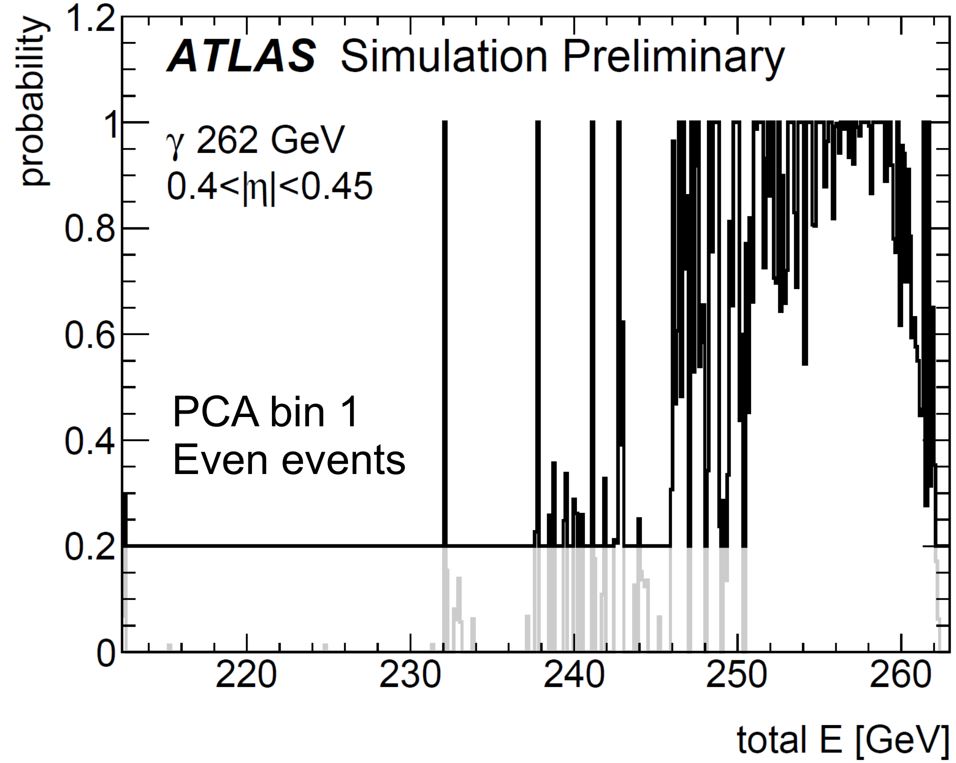
New developments: Probabilistic reweighting of the total energy

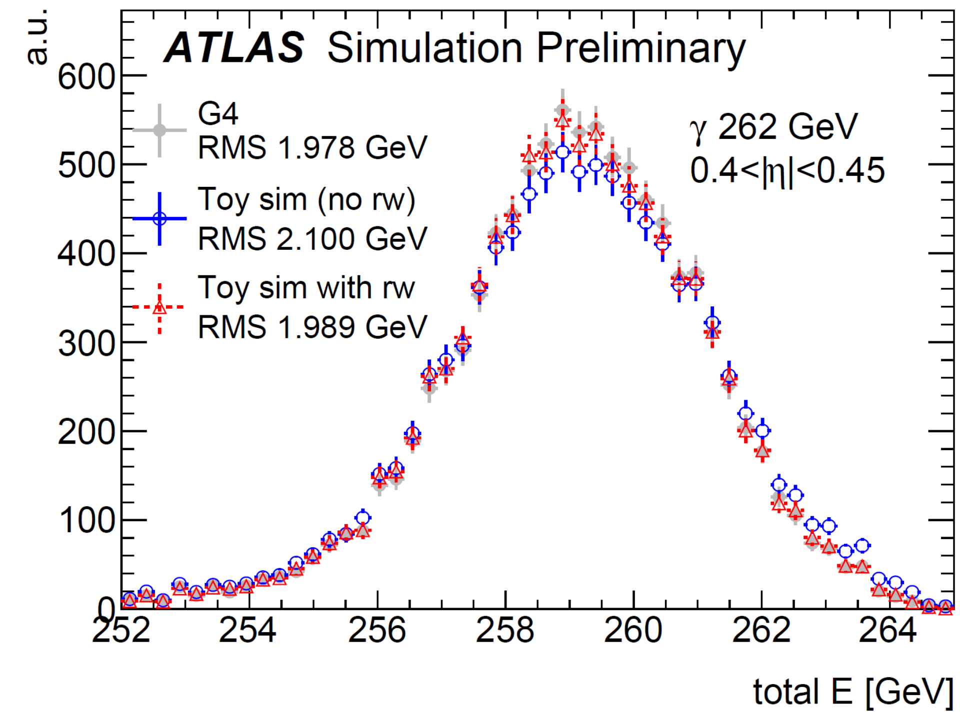
- Total energy resolution in FCS larger than expected by G4 (both in tail and core of the distribution)
- Effect is small, but consistently present in all parametrization slices (electrons, photons, pions)
- **Probabilistic reweighting of energy during simulation** such that it resembles G4 better:

  Step1: Create probability histogram from toy-simulated FCS energy / expected G4 energy and store it

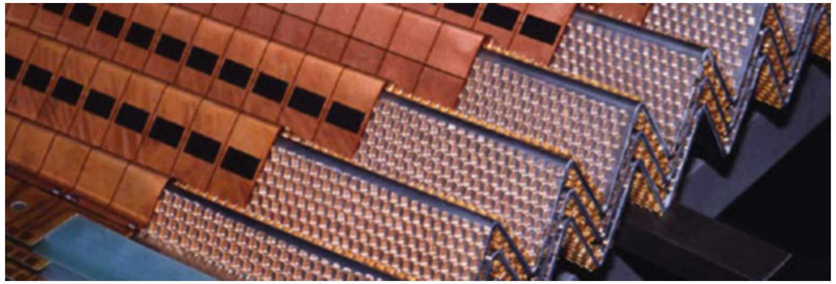
  Step2: During (actual) simulation, compare the simulated total energy value with a uniform random number RAN between 0 and 1:
  If RAN < probability → Simulation result is accepted
  If RAN > probability → Repeat simulation

- The beauty of this method is that the carefully constructed longitudinal correlations are retained
- Lower probability boundary value is tweaked for each parametrisation slice such that the reweighting results in the best possible agreement with G4





New developments: Energy dependence on Phi Impact Position



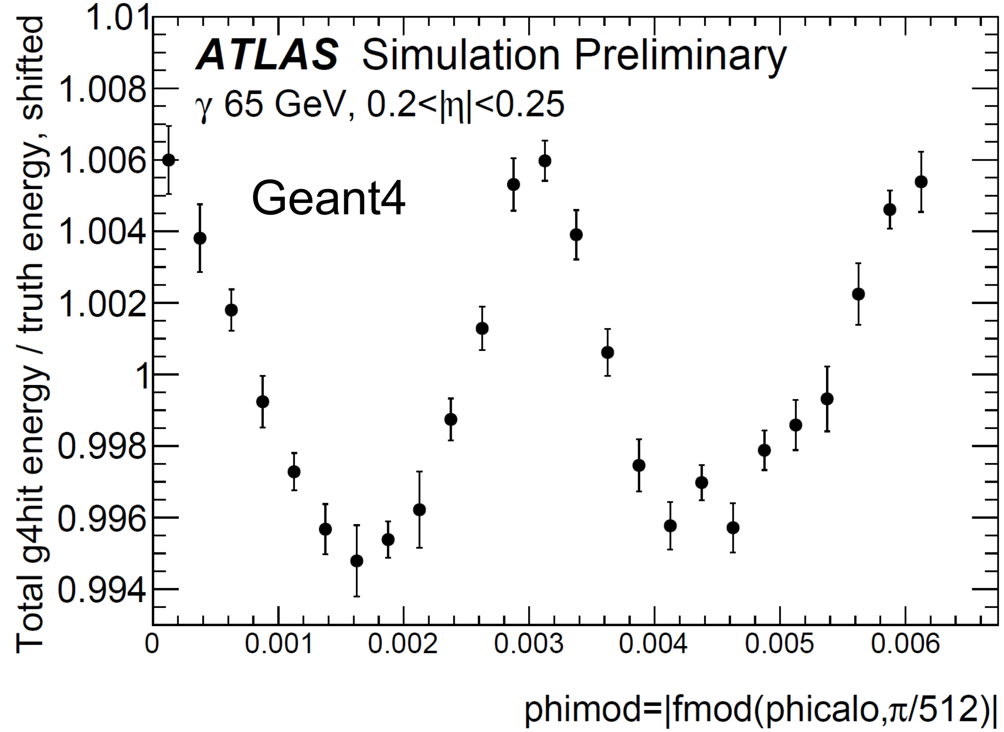
In the ATLAS EM calo, the **cells are „accordion“ shaped** (to provide crackless phi symmetry). This difficult geometry leads to an **oscillation structure of the energy response** depending on the phi impact position („phimod“).

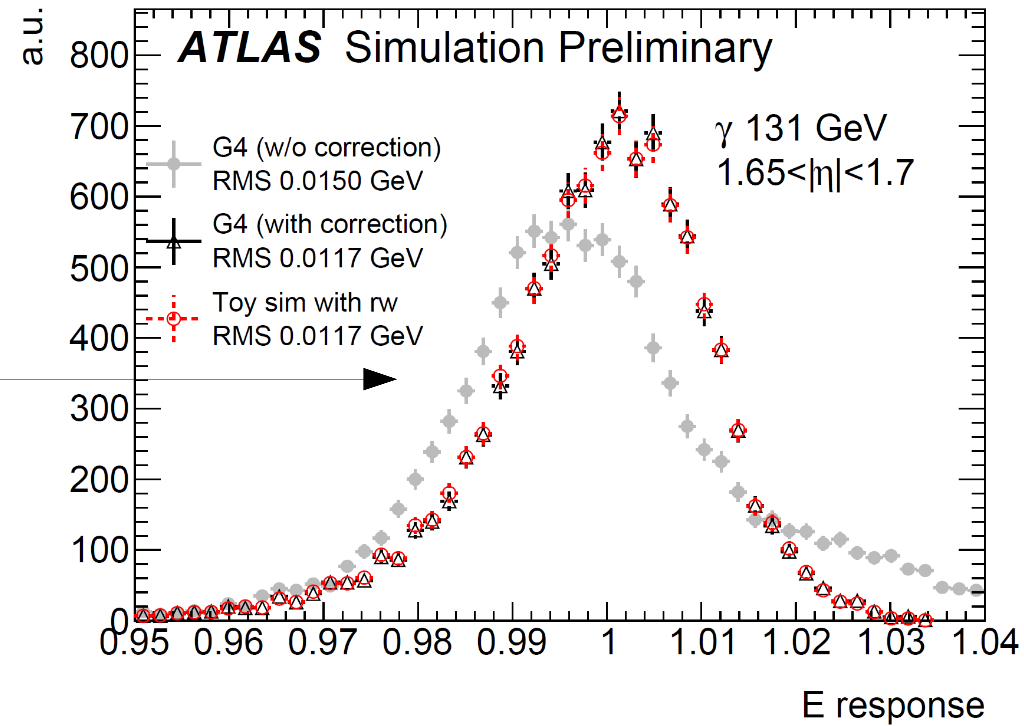
This behaviour is simulated in G4, but **FastCaloSim knows nothing about it** (because FCS uses a simplified geometry where cells are approximated by cuboids).

The energy calibration based on G4 then corrects for that oscillation, that introduces a **widening of the FCS resolution**

→ We have introduced a **new correction**, that **removes the phimod dependence in the G4 input samples** from which the FCS parametrization is obtained

The FCS resolution is forced to the „corrected G4 resolution“ via the probabilistic reweighting during the energy simulation



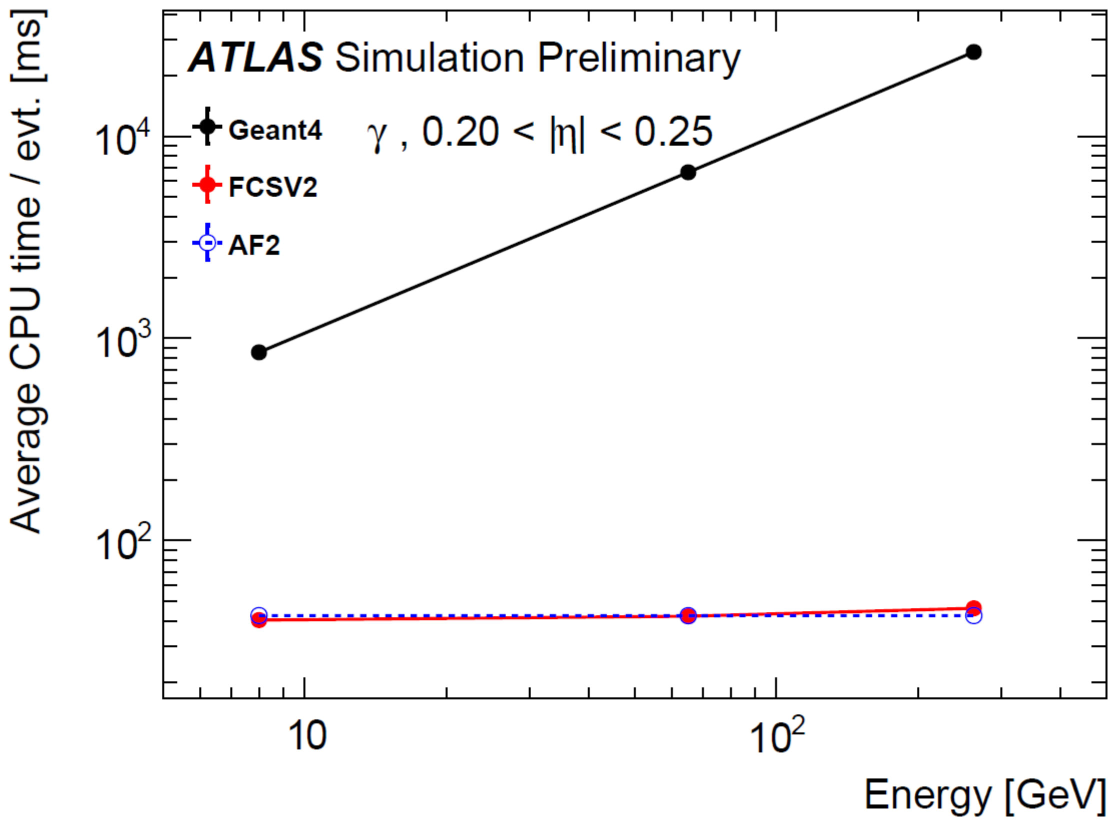


AtlFast3: Configuration and Performance

AF3 is the new ATLAS Fast Simulation tool that is comprised of the following components:

- FastCaloSimV2 (for electrons and photons, low and high energetic hadrons)
- FastCaloGAN (for hadrons with medium energies (8-265 GeV) – see backup)
- Geant4 (very soft hadrons in the calo (E_{kin}<0.4 GeV), all particles in muon system and inner detector)
- Muon Punch Through simulation (for hadrons - see backup)

CPU time for single photons (from 2018):



CPU time for ttbar events (from 2020, no pile-up):

G4 + FCSV2: 26.419 +/- 1.197 sec

G4 itself takes 26.32 +/- 1.111 sec
FCSV2 takes 0.025 sec (estimated separately)

(G4+FastCaloGAN or AF3 would perform similarly)

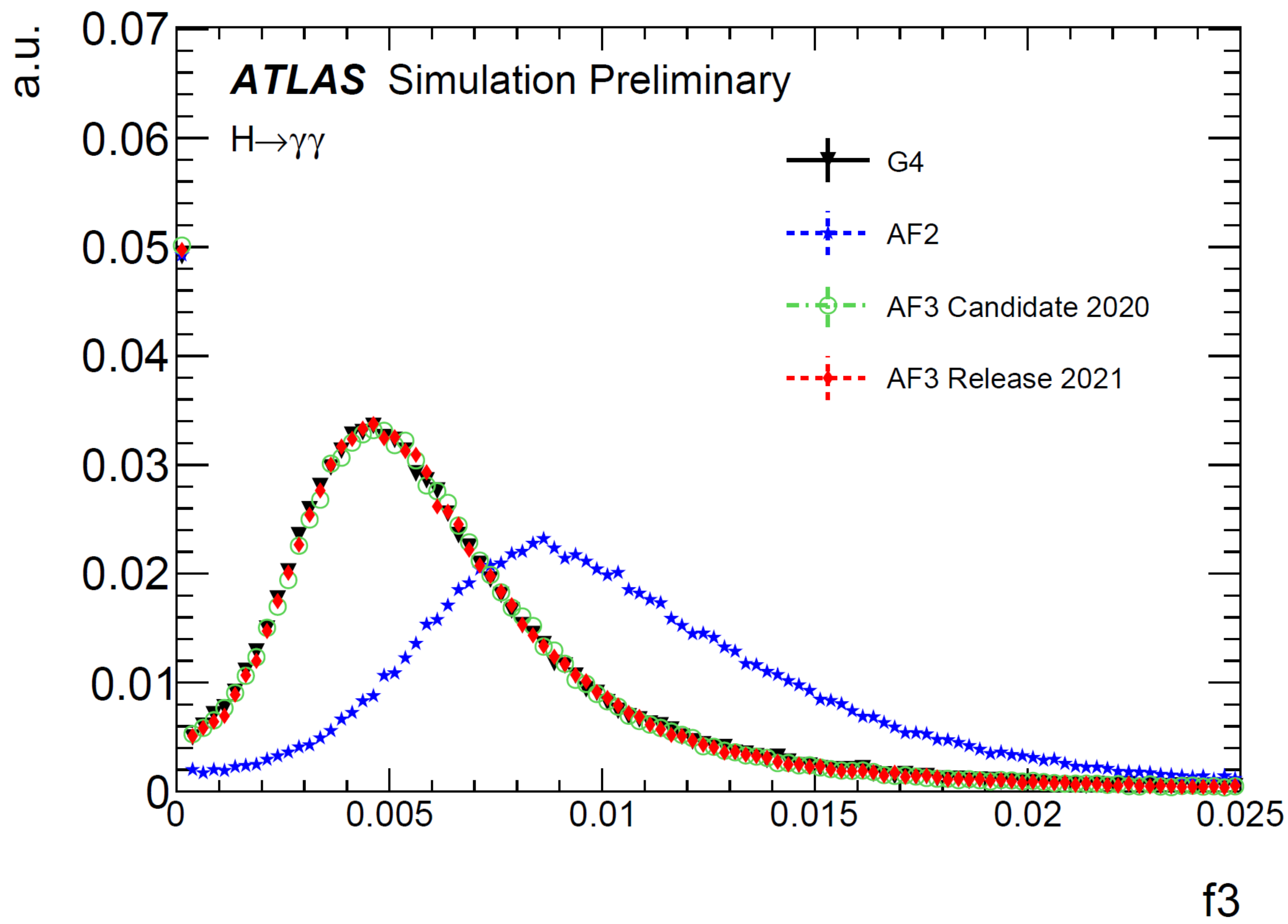
Full G4: 228.89 +/- 10.1 sec

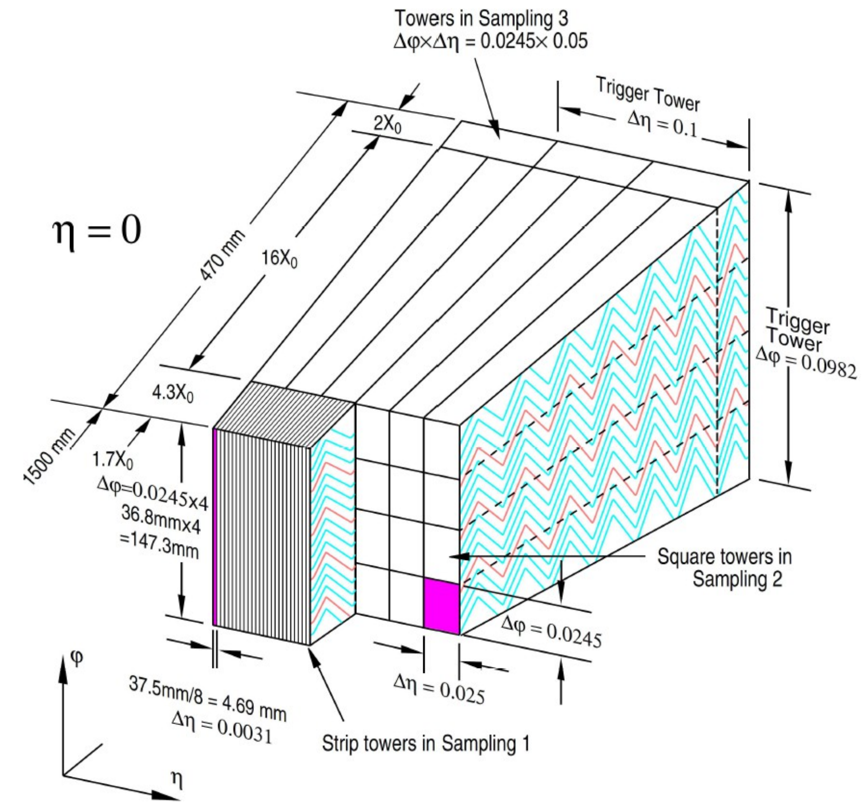
AF2: 27.18 +/- 1.6 sec

→ **Factor 8-10 speed-up compared to Full G4**

CPU time is still totally dominated by G4, the fast simulation itself is no bottle-neck.
Not sufficient for run-4, need to further reduce G4 usage (ie. fast tracking, fast chain).

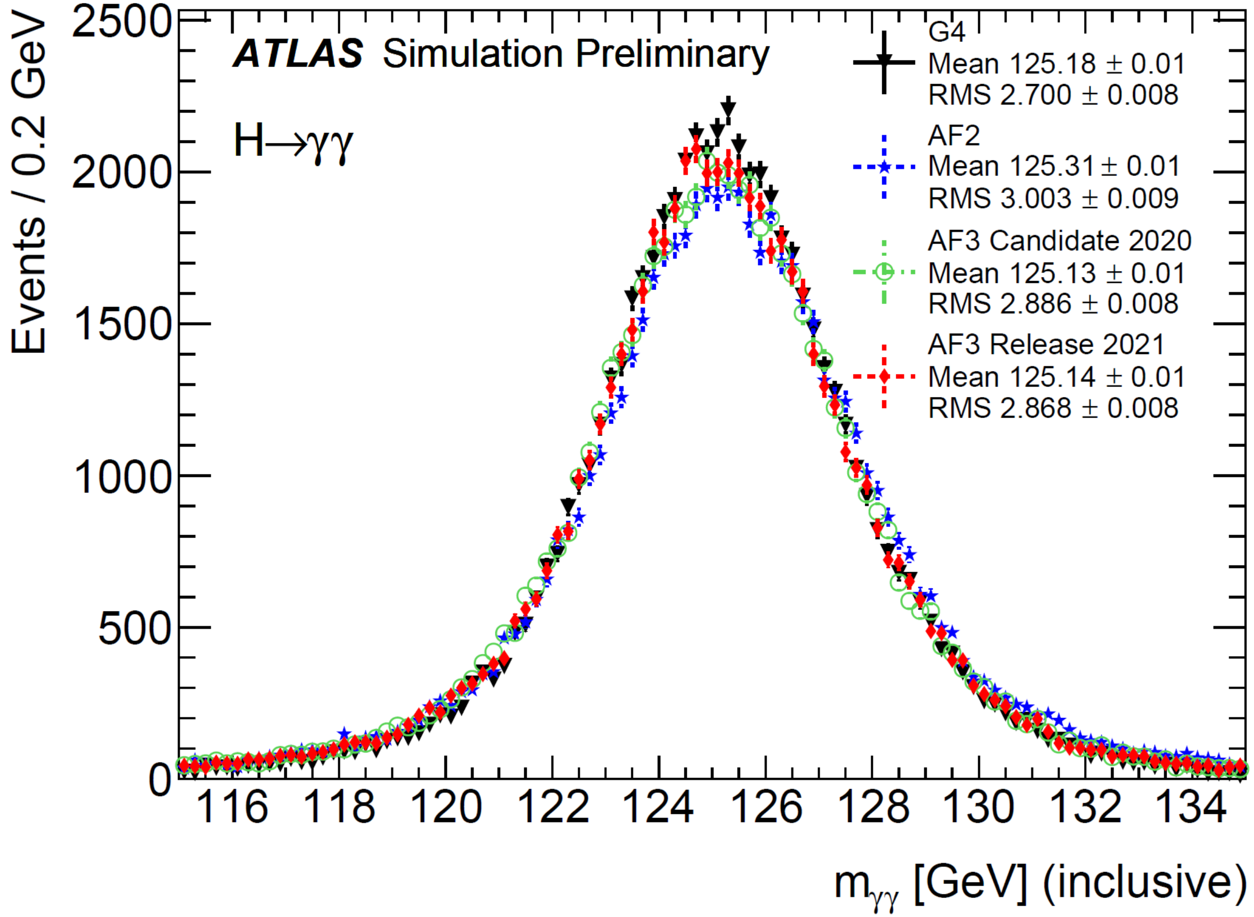
Latest validations (after Athena simulation and reconstruction)

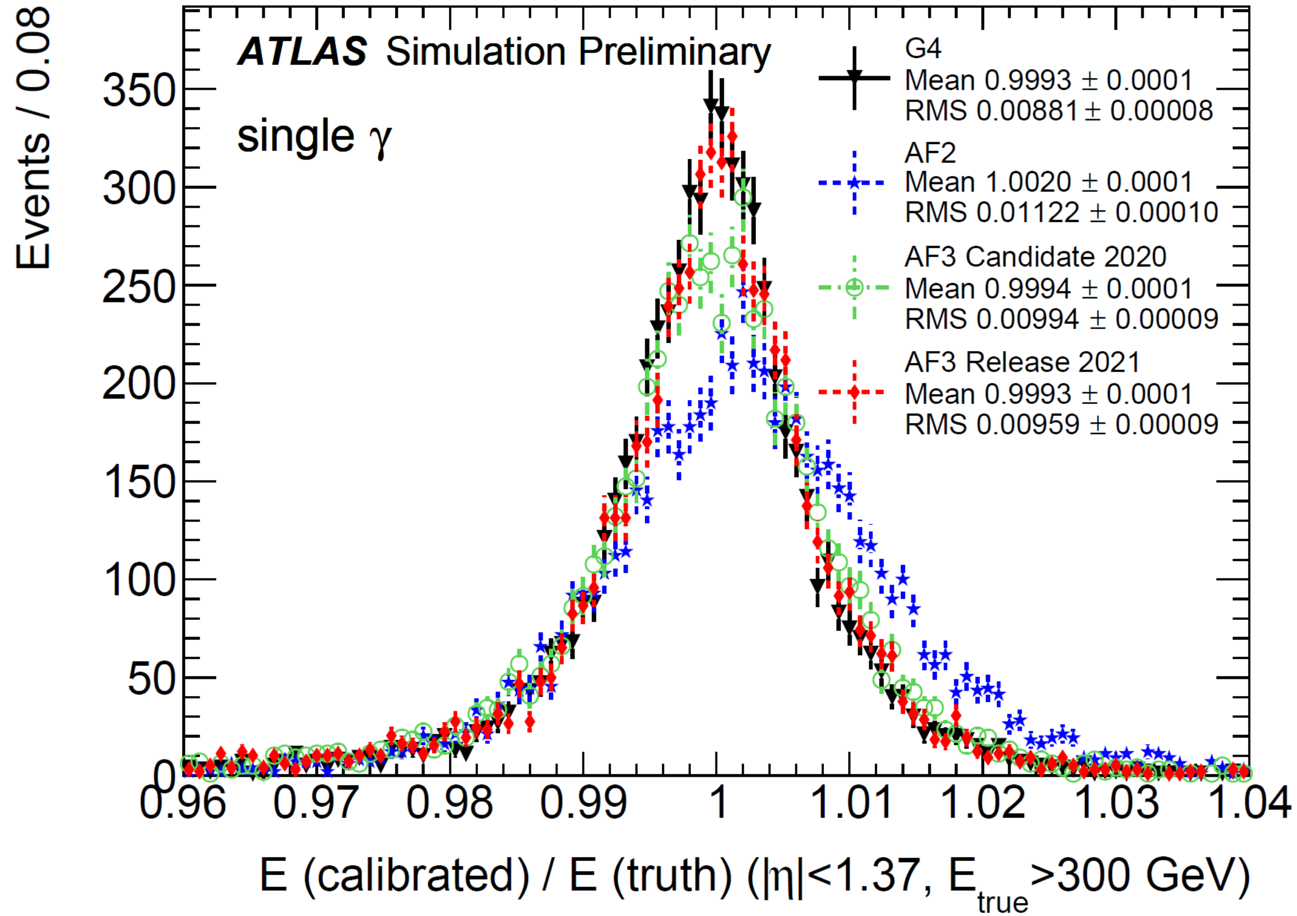




f3 is the energy fraction deposited in the third calorimeter sampling
AF3 reproduced G4 perfectly, while AF2 significantly differs.
AF2 had problems modelling this energy fraction, in addition it was based on an older G4 version

Latest validations (after Athena simulation and reconstruction)





- Latest AF3 version gives best agreement with G4 (2021 version includes the resolution correction)
- Energy calibration is important. In ATLAS we use a BDT-based calibration (arXiv:1908.00005):
  - Each sample displayed here is using a dedicated calibration trained with the respective simulator, except for AF3 2021 that is calibrated based on the AF3 simulation from 2020
  - Potential improvements expected from dedicated energy calibration trained on latest AF3 version

AF3 Release and Future Plans

AF3 release is imminent. Will be used to resimulate ~7 billion AF2 events with AF3!

Another release is planned for LHC Run-3 (early 2022)
(Updated G4 version, improvements for FastCaloSim and FastCaloGAN)

Potential FCSV2 improvements:

- Phi modulation correction on hit level could bring resolution even closer to G4
- Correlated fluctuations between cells to improve pion and jet modelling (see backup), could potentially also improve egamma shower shape variables in first sampling
- Mismodellings in Barrel-Endcap transition region ($1.37<|\eta|<1.52$), more studies needed

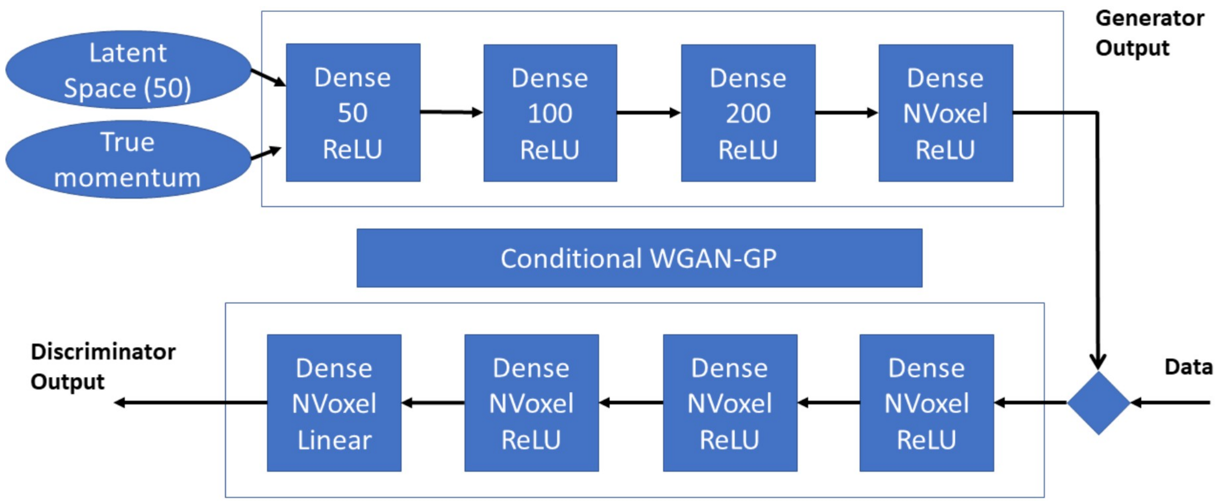
Publication plans:

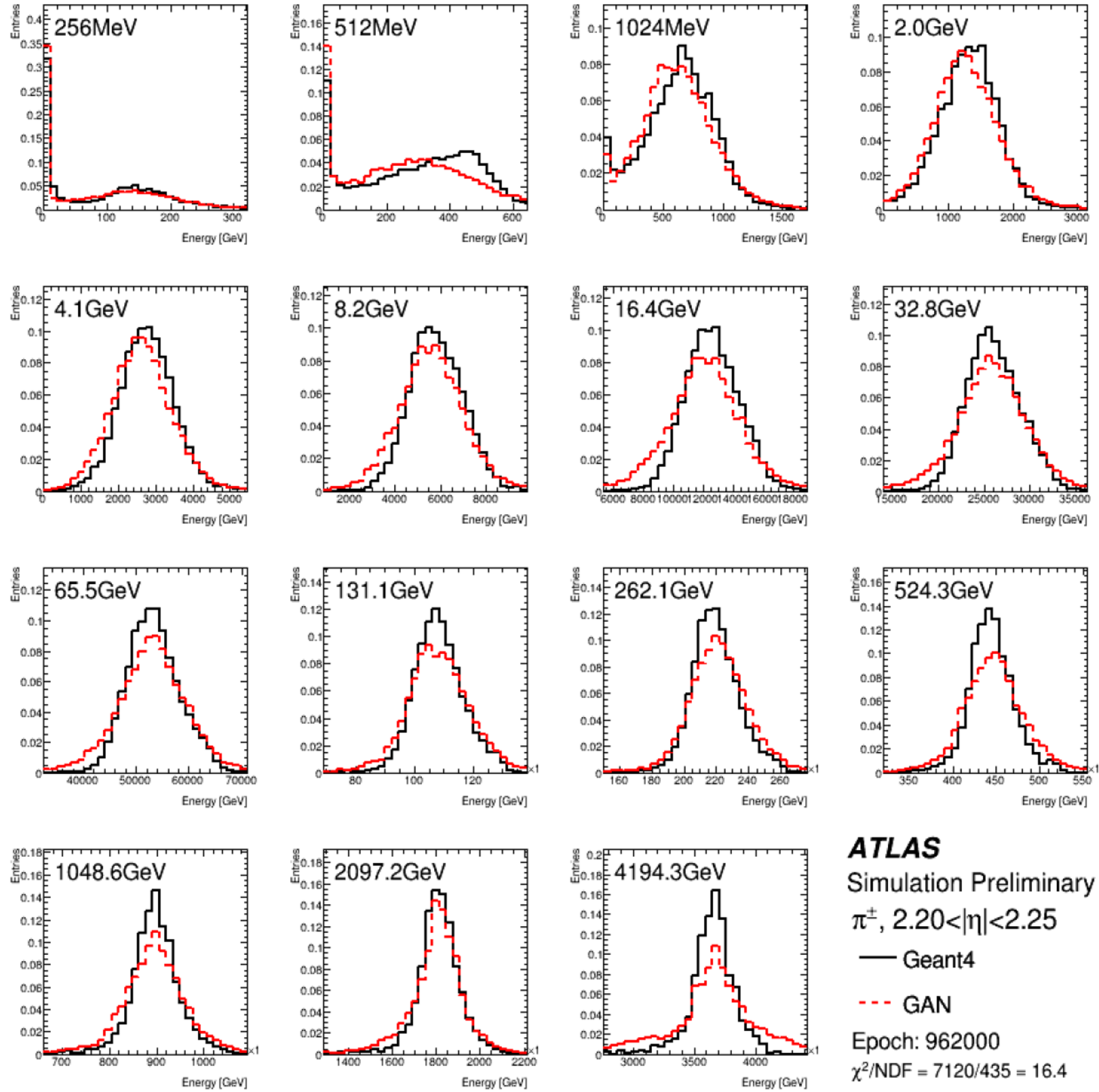
- vCHEP paper submitted (~10 pages, to become public in May)
- AF3 paper planned

Thanks US ATLAS Computing for the support! 💙

Backup

FastCaloGAN





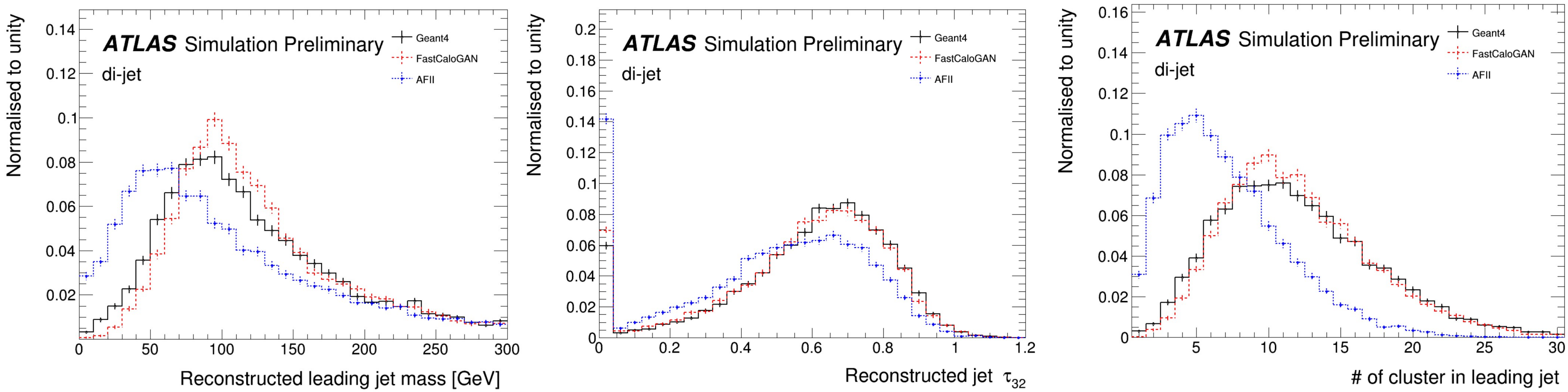
Generative Adversarial Network used to generate entire calo showers, trained on G4

GAN trained for each eta bin, inclusive in energy (100 GANs per particle)

GANs exists also for photons/electrons, but we found it competetive and often better than FCS only for pions

Simulation PUB note: http://cdsweb.cern.ch/record/2746032

FastCaloGAN



In the **latest AF3 configuration**, we use a **hybrid** FastCaloSim/FastCaloGAN approach for pions:

Up to 8 GeV: FastCaloSim
16 GeV – 265 GeV: FastCaloGAN
524 GeV – 4 TeV: FastCaloSim

Not straightforward to evaluate the best thresholds, since single pions are parametrized but ultimately jets are what's needed in physics analysis, and each jet consist of many constituents, each with different energy

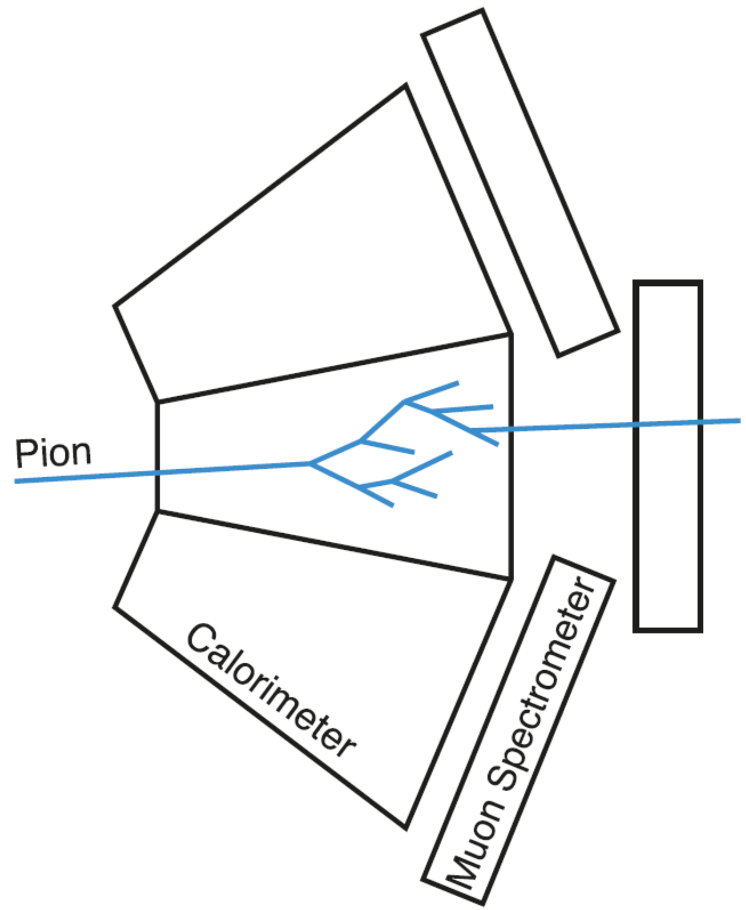
Muon Punch Through Simulation (Principle)

For high energetic hadrons, particle showers are sometimes not fully contained in the calorimeter. Instead, they exit the calorimeter and enter the muon system, where they could be reconstructed as fake muons.

FastCaloSim (or FastCaloGAN) cannot simulate such effects

AF3 includes a new dedicated simulation of muon punch through, that creates such signatures depending on the initial particle energy, type and eta.

The created punch though particles are then passed back to G4, which takes care of the simulation in the muon system.

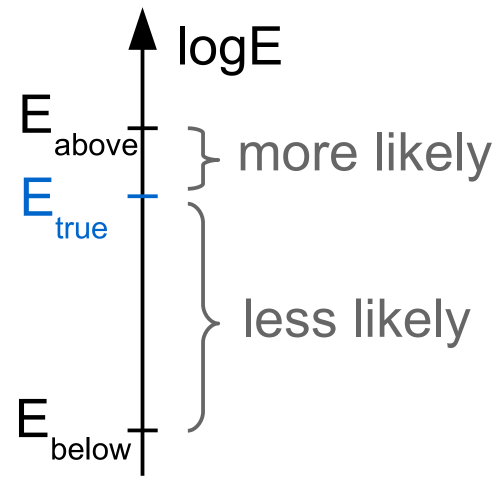


Interpolation between energy points

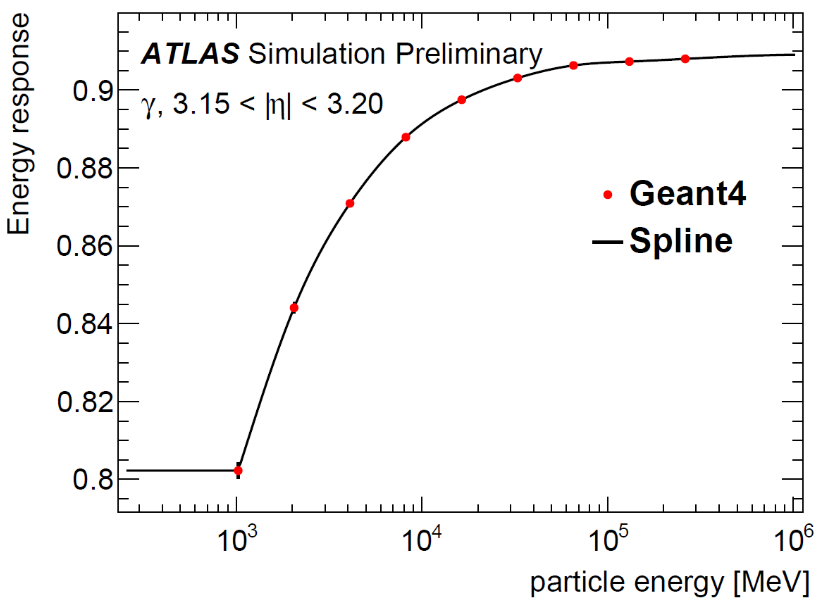
- G4 inputs simulated for fixed energies (17 points on a logarithmic scale, 64 MeV – 4 TeV)
- To simulate a truth particle with any energy value E_{true} :
  - The parametrisation picked is determined randomly, depending on log E_{true}:

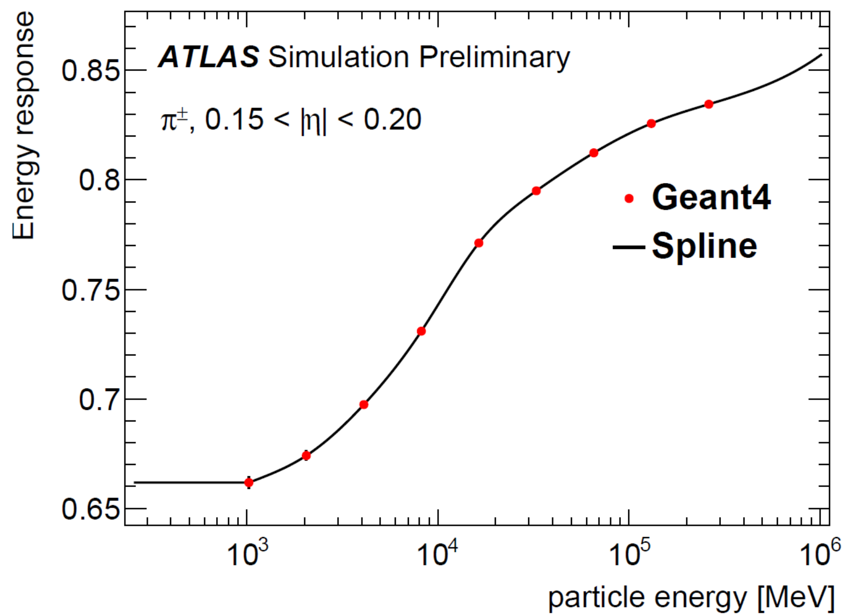
    - throw random number r [0,1]

    - if ($(\log E_{true} - \log E_{below}) \ / \ (\log E_{above} - \log E_{below})$) > r $\rightarrow$ choose above point
    otherwise, choose below point



  - The total energy response is interpolated using a spline:



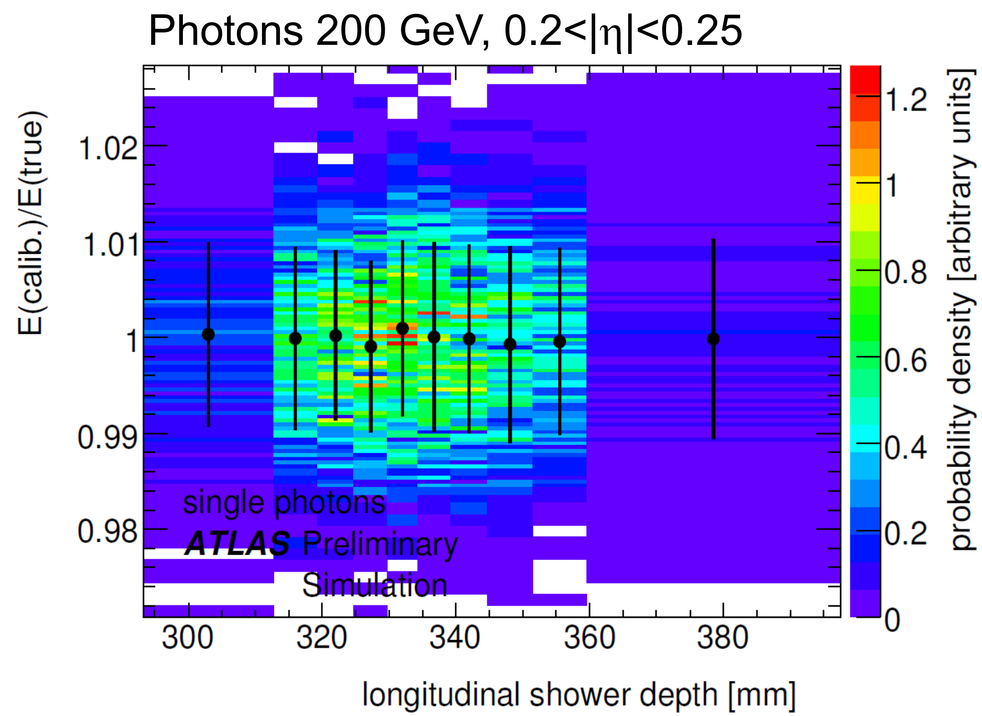


The „old FastCaloSim“ (implemented in AF2)

ATL-PHYS-PUB-2010-013

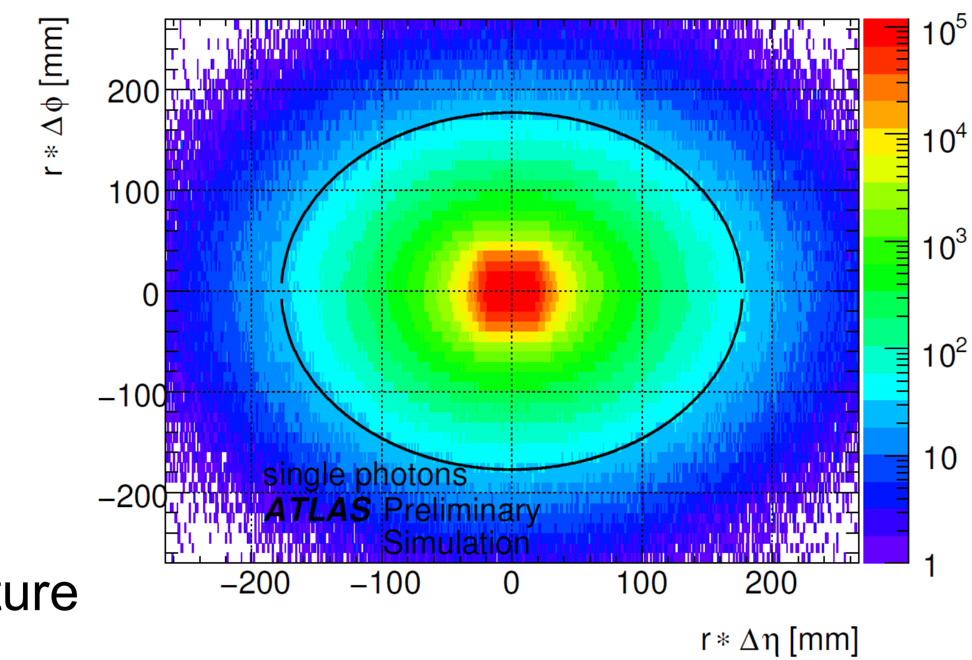
This software has been used when we say „fast simulation“ in our physics papers. It uses Geant4 in the inner detector and muon system, and FastCaloSim in the calorimeter system.

About 50% of all simulations in ATLAS are done with AtlFASTII.



Longitudinal energy parametrisation:

- 2D histograms of energy vs. longitudinal shower depth (=distance of the deposit from the calo surface), for total energy and energy fraction per layer
- Correlations between the deposits in each layer stored explicitely in correlation matrices

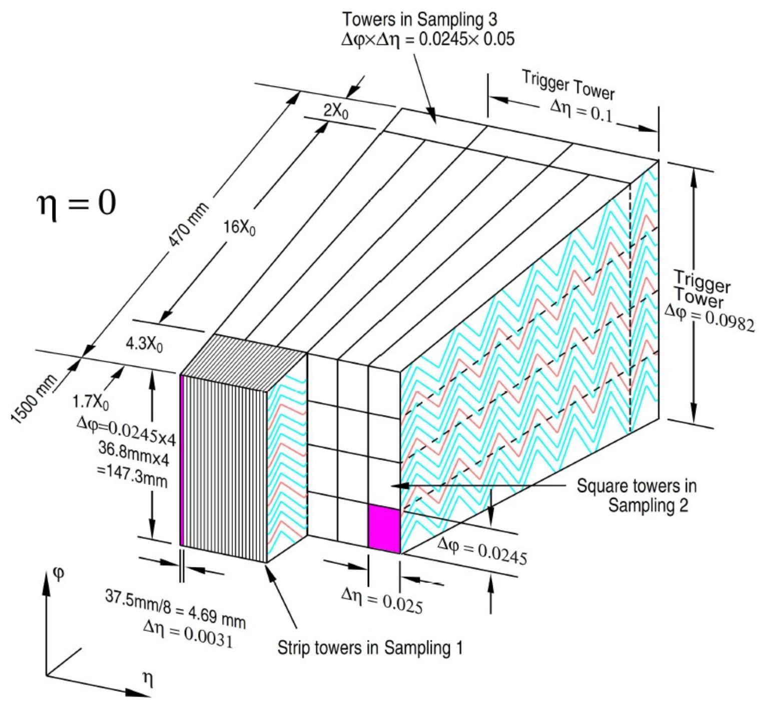


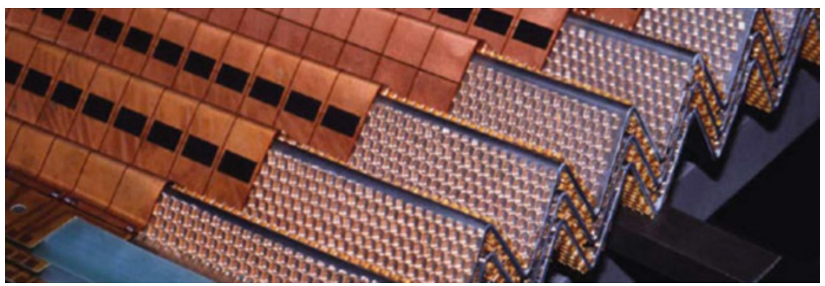
Lateral shower shape parametrization:

- Radial symmetric function centered around the impact point of a particle in the calo layer (=3rd order polynomial function)
- Parameters obtained from a fit to the Geant4 single particle lateral shape in each calo layer, for each particle type, energy, |η|, shower depth bin
- Good average shower description, poor modelling of substructure

The ATLAS Calorimeter System: Different Geometries

EM Calo

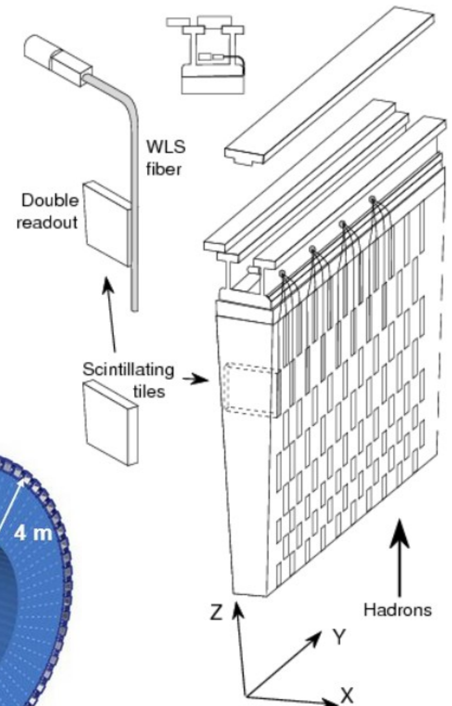


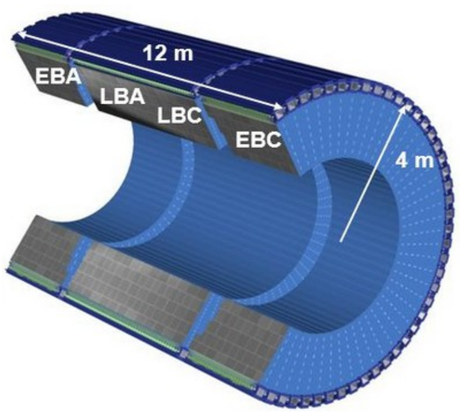


- 3 layers + presampler, with very different cell size and granularity
- Samplings folded in an accordion structure

Tile Cal

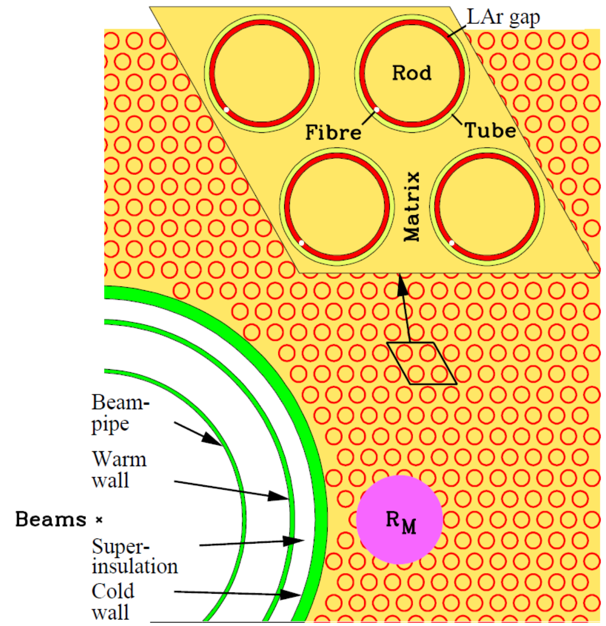
Layers of Tiles

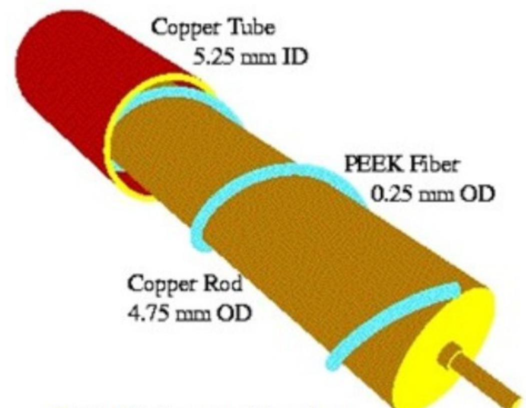




FCAL

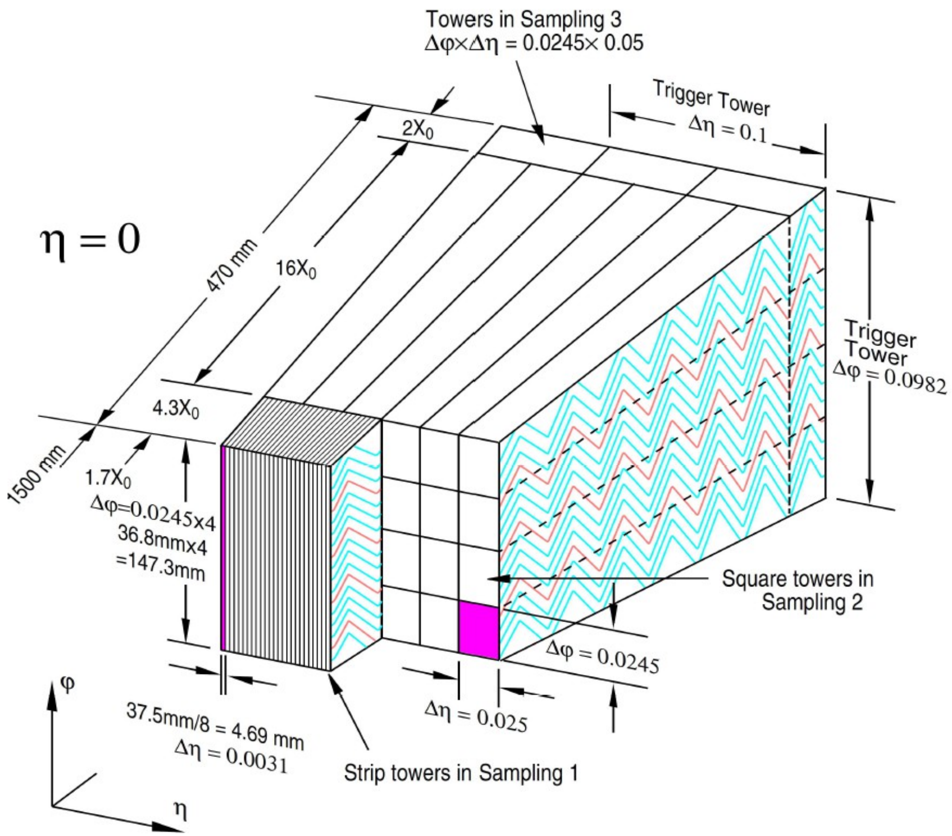
Tubes in a honeycomb structure

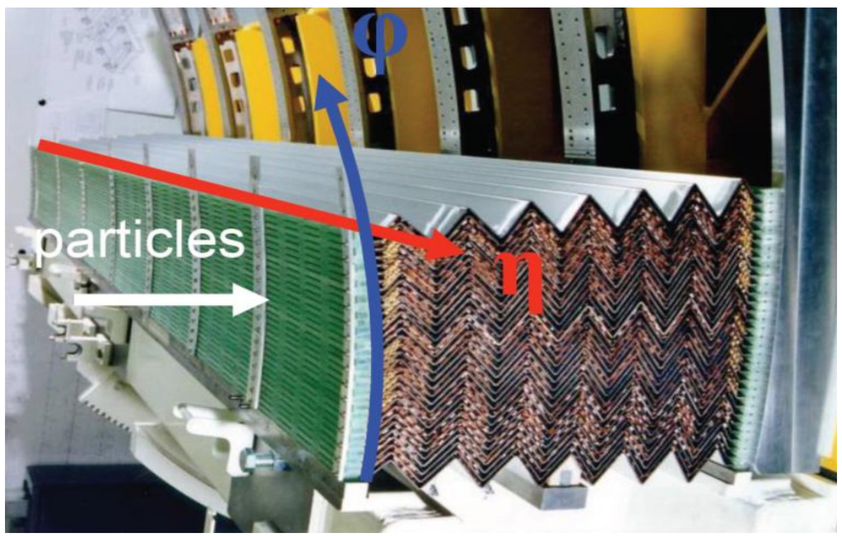




FCAL1 Anode Structure

Lateral Shape Parametrization: Simplified Geometry





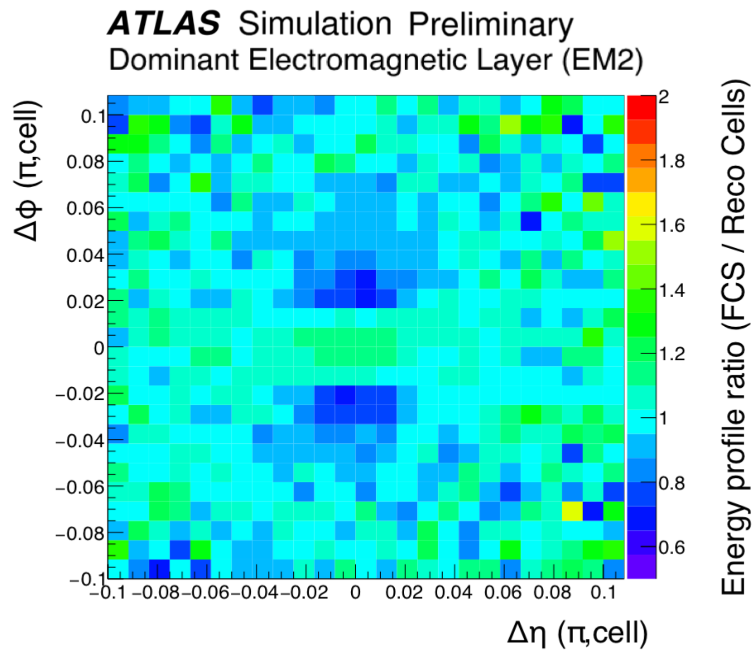
The simulated hit energies sampled from the 2D histograms have to be assigned to the physical calorimeter cells.

An important speed-up for this hit-to-cell assignment is a simplification of the geometry in the EM calo:

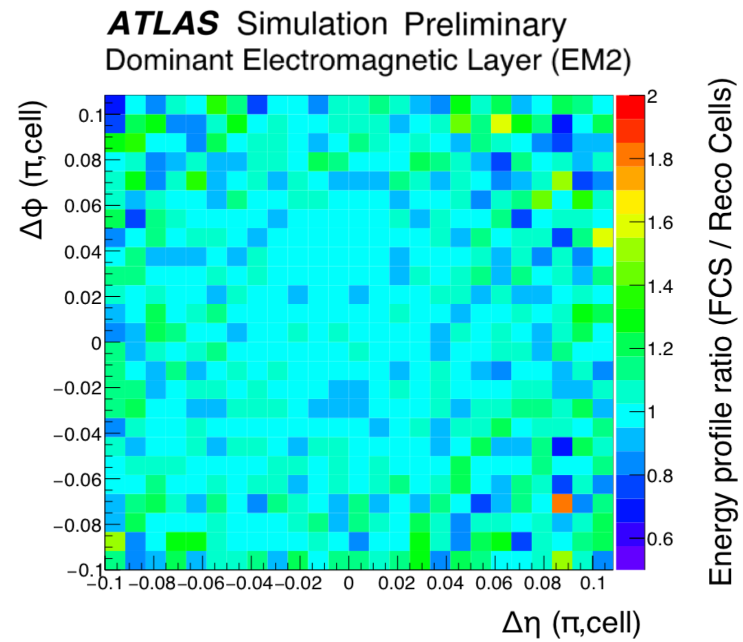
Accordion shaped cells are approximated by cuboids

This simplification however comes with a price. The shower shapes now deviate from the G4 simulation, and the energy distribution needs to be corrected by randomly deplacing hits in phi direction to re-emulate this accordion structure.

Biased:

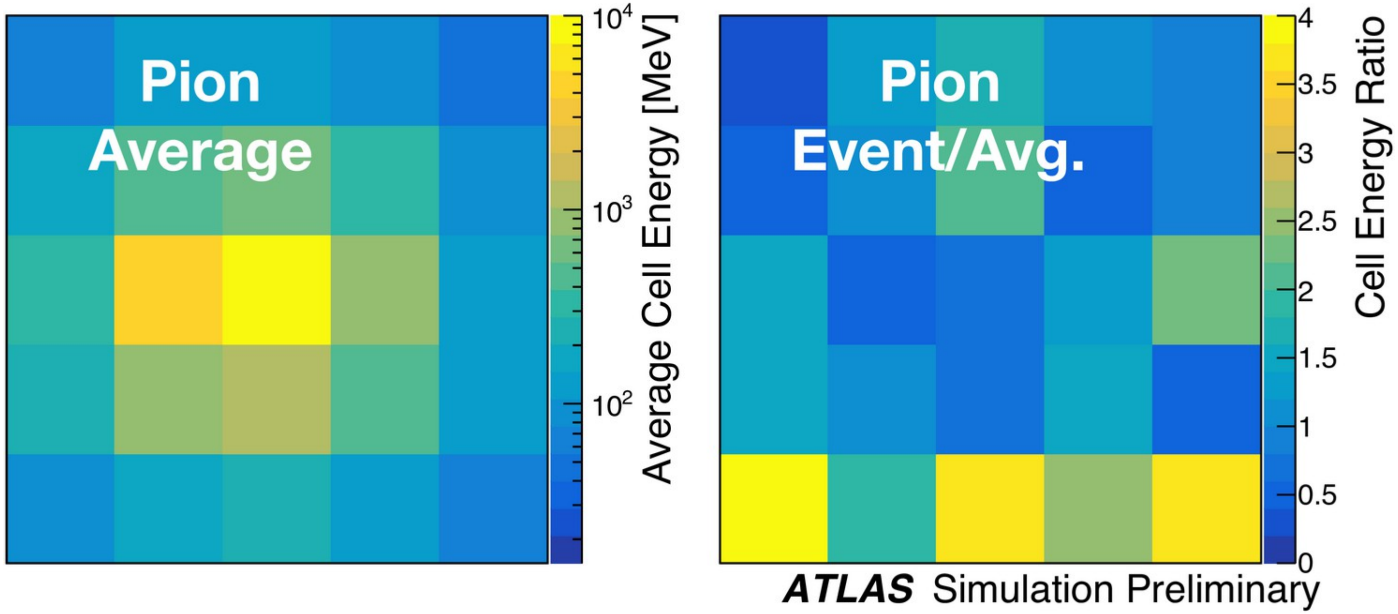


Corrected:



Correlated Fluctuations

ATL-SOFT-PROC-2020-027



Left: Pion lateral energy distribution averaged over 10k showers in a grid of 5x5 cells

Right: Energy distribution of a single shower compared to this average → ratio strongly deviates from 1

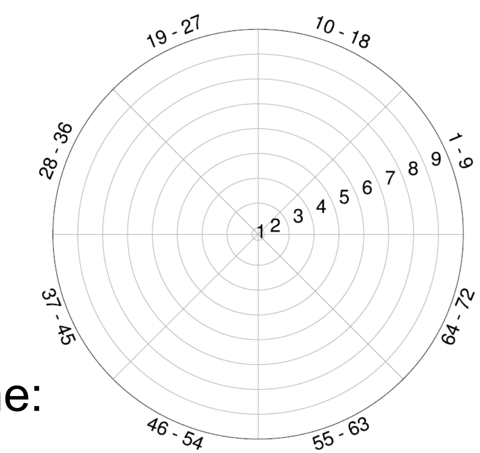
Can we model the correlations of energy deposits between each cell?

This is for cells in the same calorimeter layer, in a given slice of eta and input energy.
Correlations between samplings are incorporated in the longitudinal parametrization

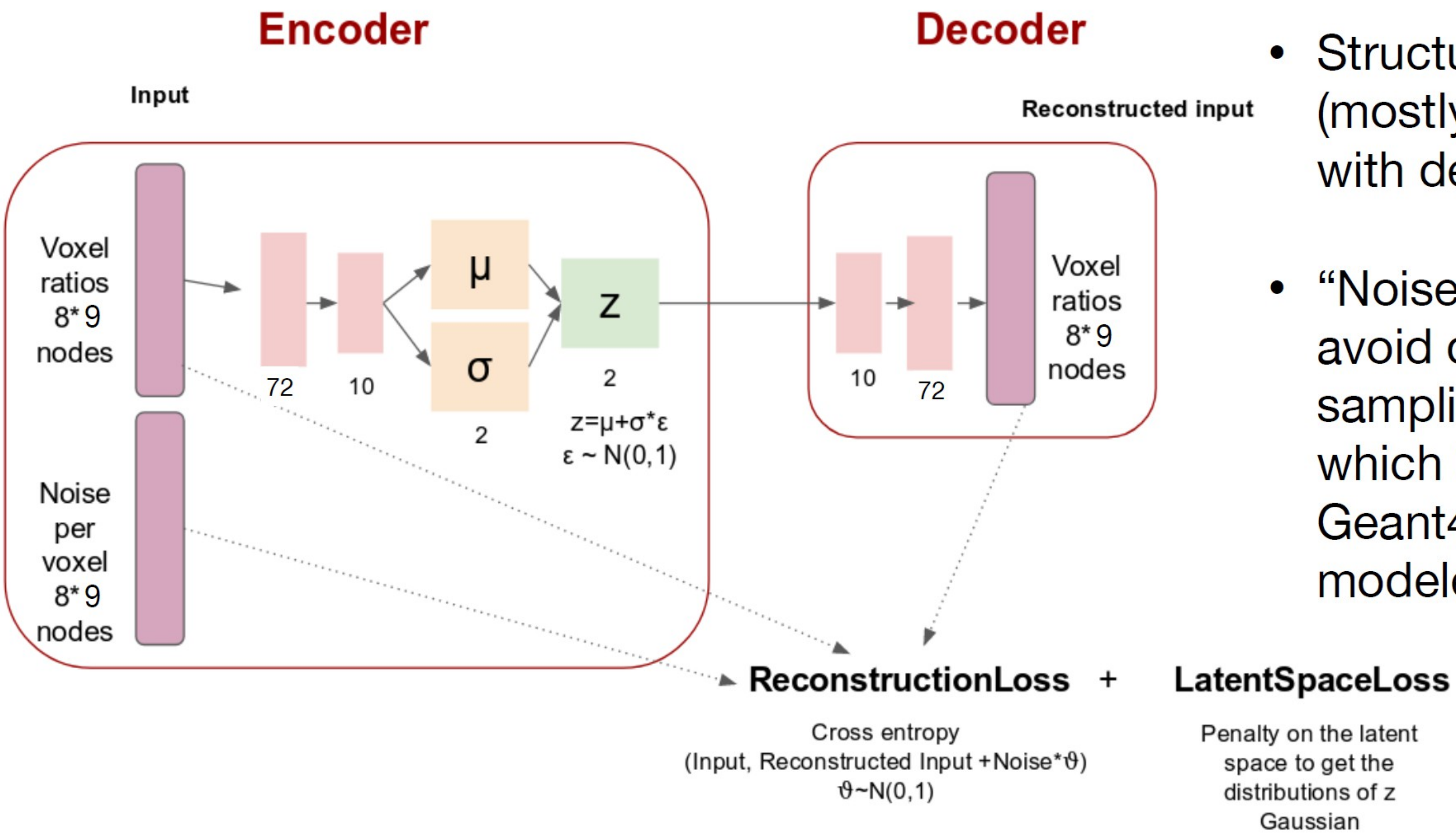
We have found better results are obtained in a finer grid than 5x5 cells.

→ We converged on a grid of 8x9 voxels in R-alpha plane (~half the cell size) for the following studies.

Voxel definition in R-alpha plane:



Modelling Correlated Fluctuations with a VAE

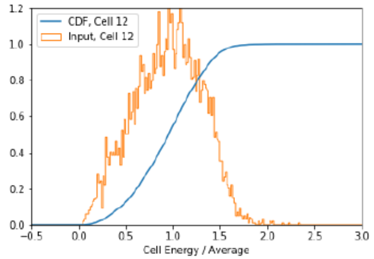


- Structure that we use is (mostly) standard VAE with dense layers
- “Noise term” added to avoid double counting sampling fluctuation, which is present in Geant4 and already modeled by drawing hits

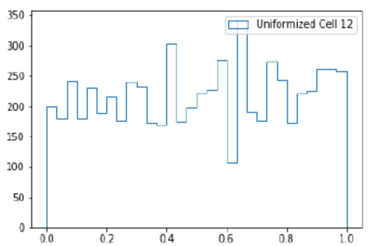
ATL-SOFT-SLIDE-2020-008

Modelling Correlated Fluctuations with a Multi-dimensional Gaussian

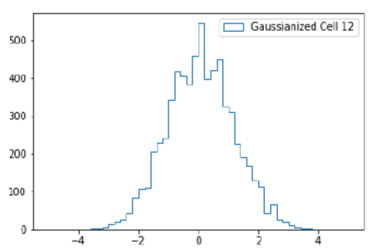
1. **Construct CDF from input distribution**

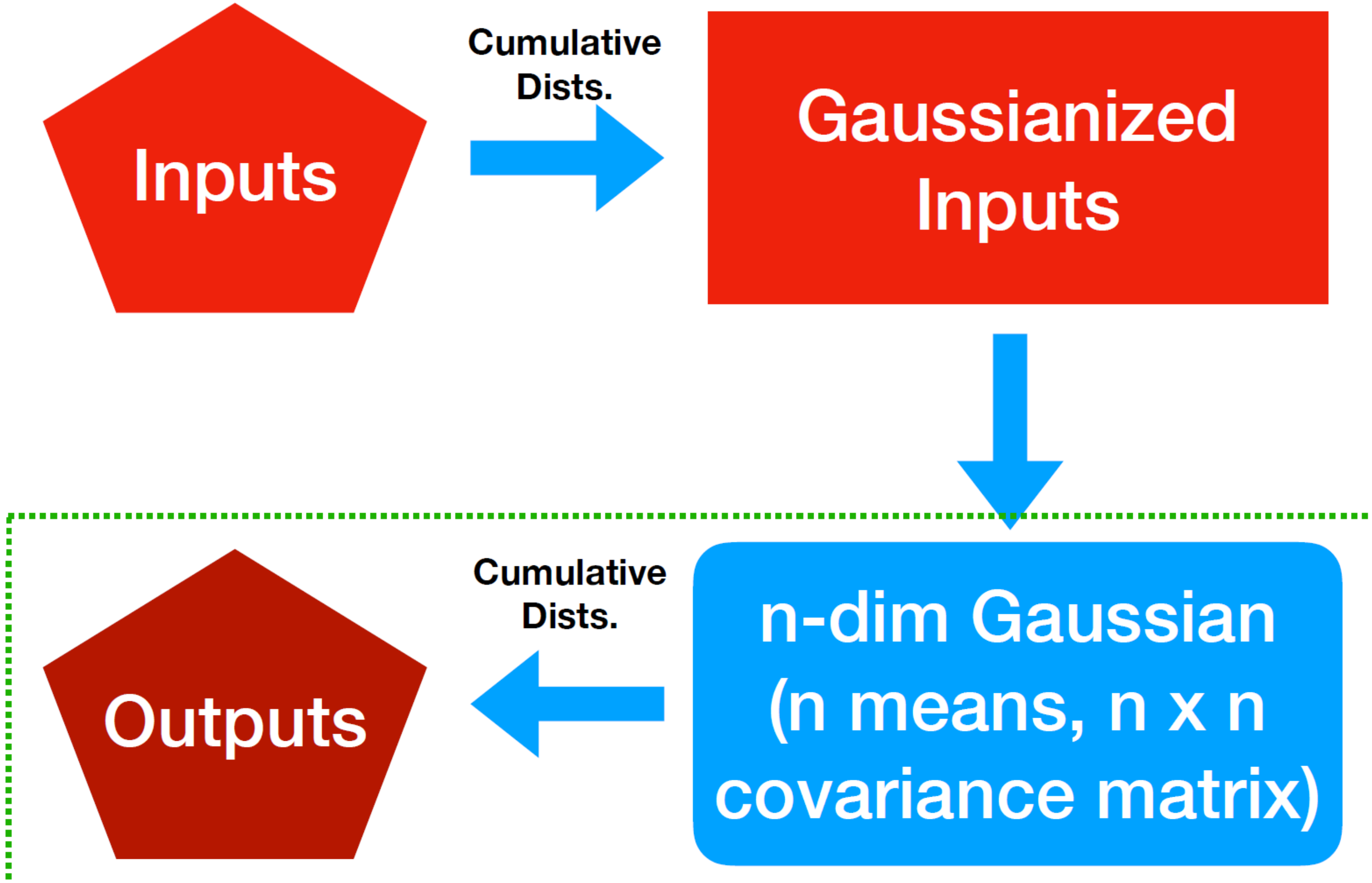


2. **Uniformize by sampling from CDF (CDF(x) for each x value in input)**



3. **erfinv(uniform) = Gaussian**



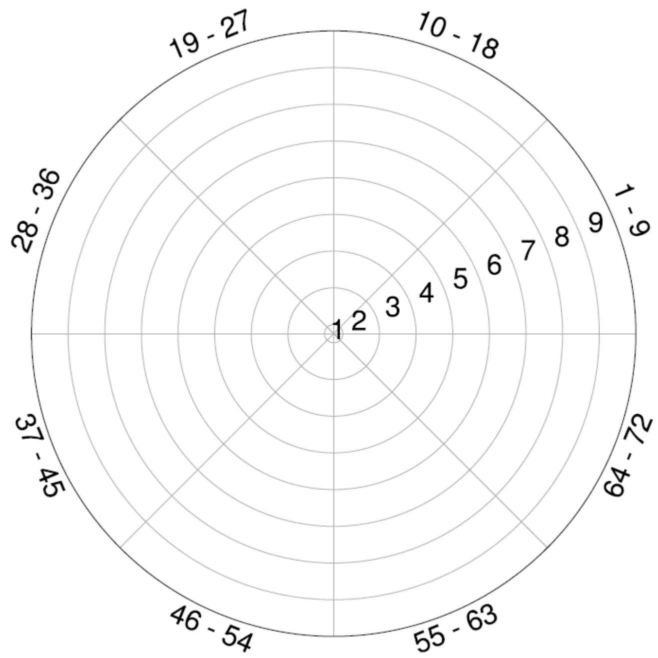


ATL-SOFT-SLIDE-2020-008

Modelling Correlated Fluctuations: Results

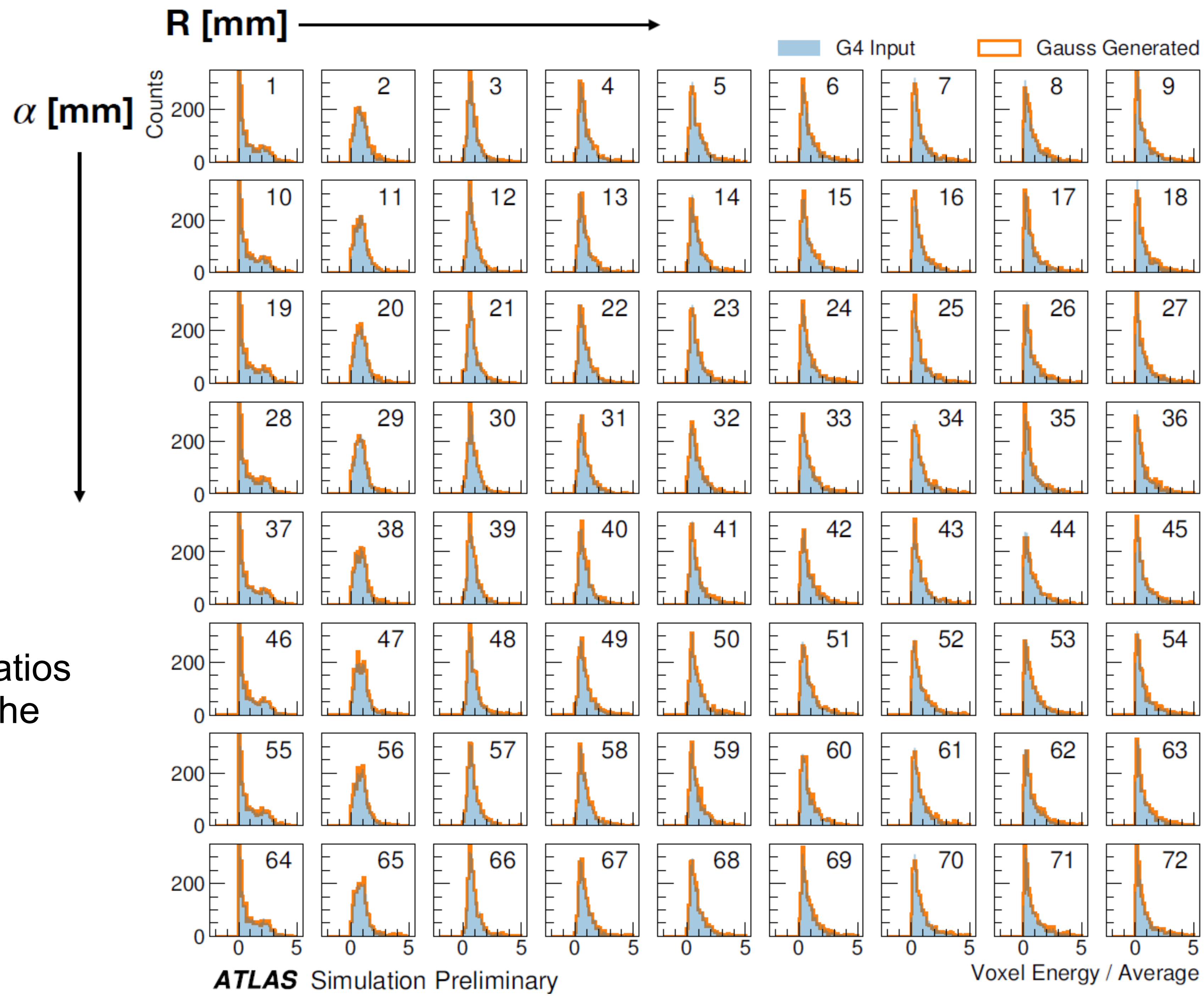
ATL-SOFT-PROC-2020-027

Voxel definitions:



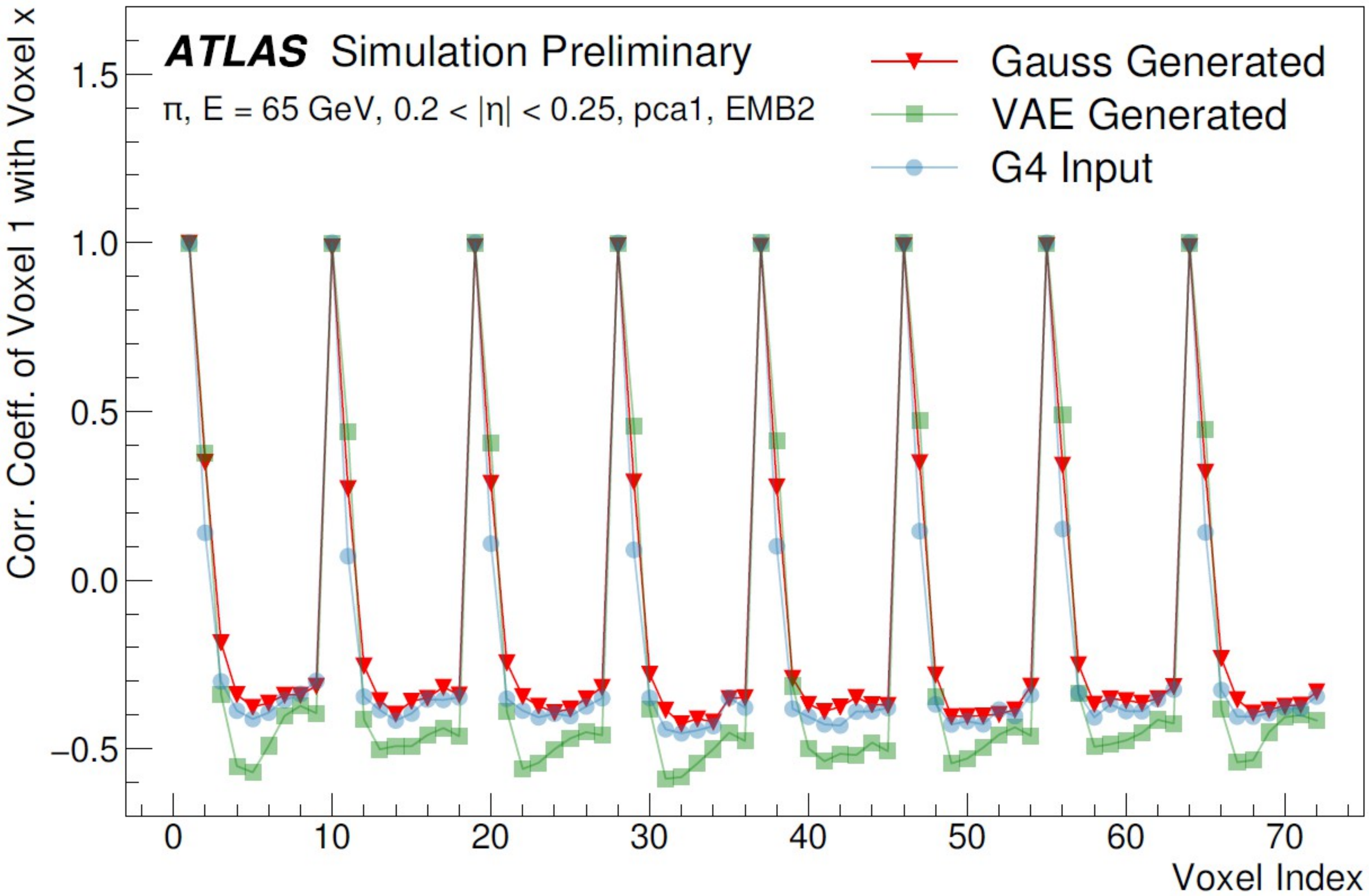
Example of input Geant4 energy ratios and energy ratios generated with the Gaussian method

Distribution in each voxel matches extremely well!



Modelling Correlated Fluctuations: Results

ATL-SOFT-PROC-2020-027

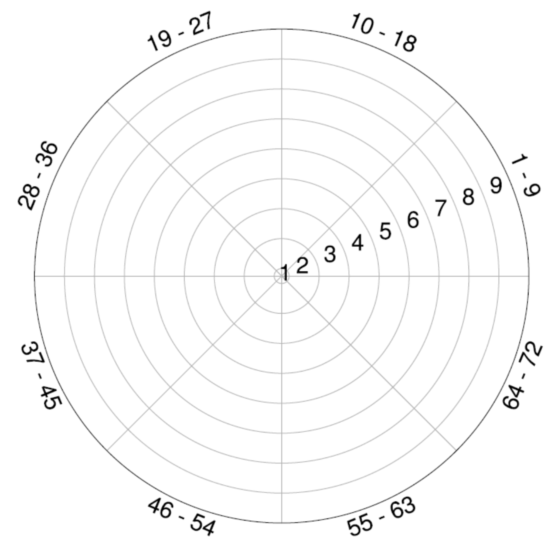


Correlation coefficient of voxel 1 (center voxel) with each of the other voxels

Points with a coefficient of 1 are due the core (0-5 mm in R) voxels in each alpha bin, which are identical and set to the average value across the core in the training input.

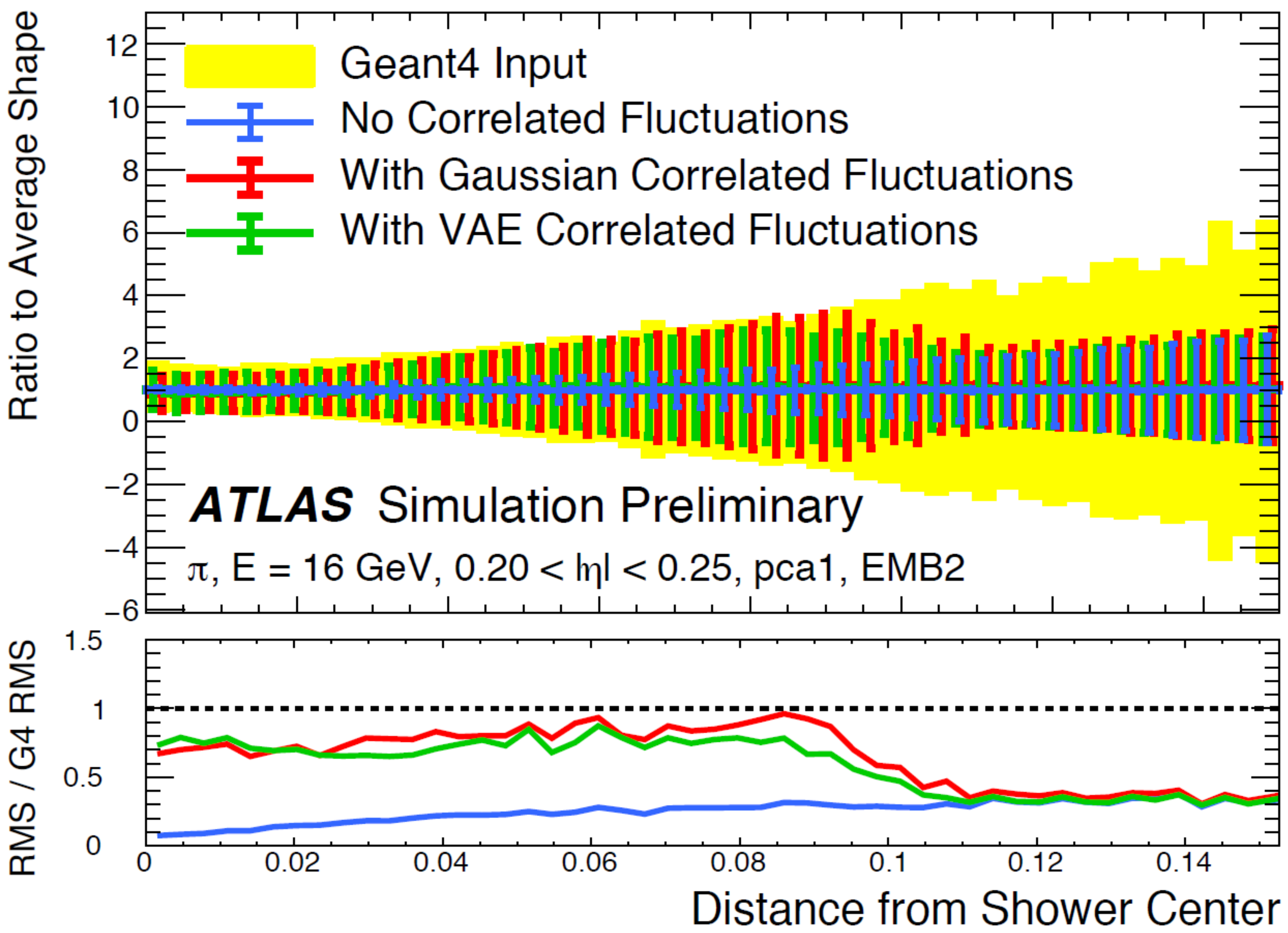
Agreement with Geant4 is good for both the VAE and Gaussian method

Reminder of voxel definitions:



Modelling Correlated Fluctuations: Results

ATL-SOFT-PROC-2020-027



RMS of the fluctuation of the average shape as a function of distance from the shower center.

Shown are for 16 GeV pions, $0.2 < |\eta| < 0.25$ in EMB2.
PCA bin 1 is chosen because it has showers with significant energy in EMB2.

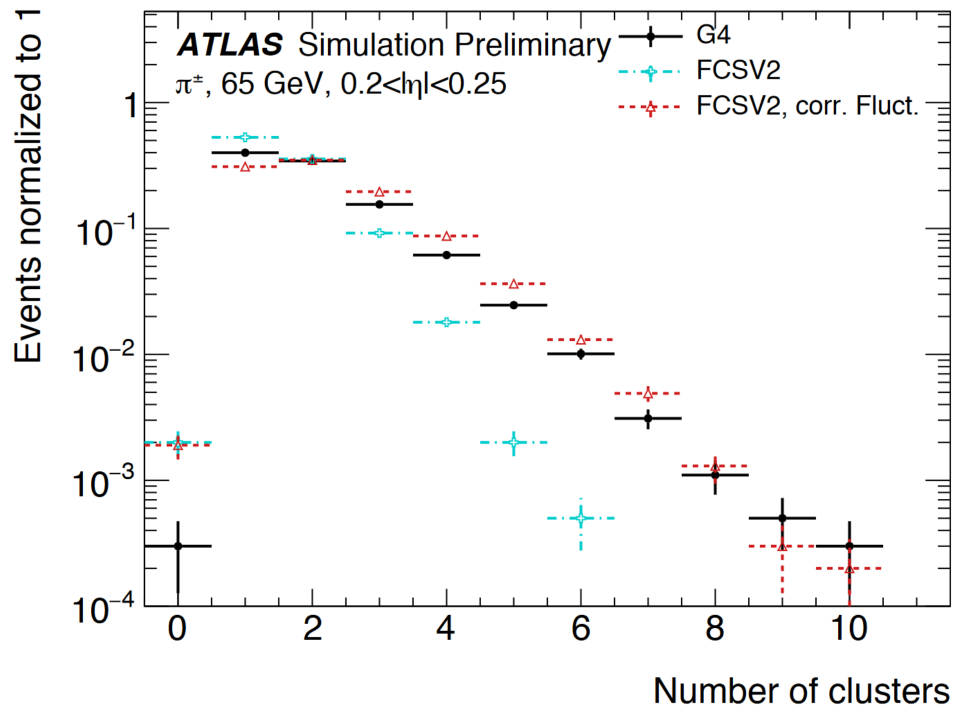
For FCS without the fluctuation models, the RMS is much smaller than in Geant4.

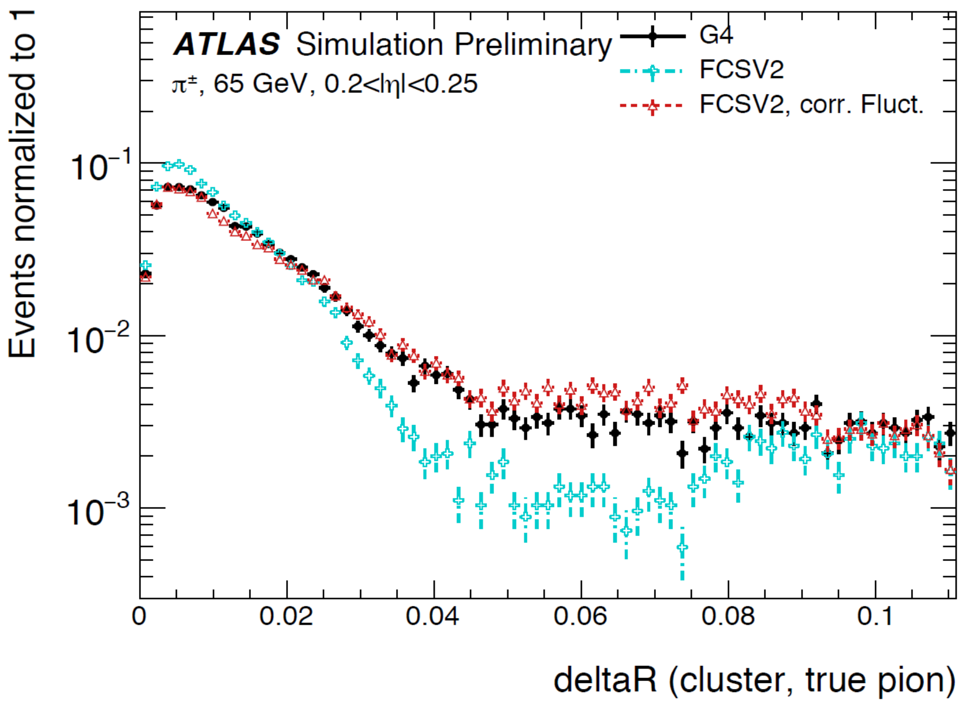
Both the VAE and Gaussian method agree much better, though correlated fluctuations are only applied in a limited range.

This validation is done on „closure test“ level, without passing events through reconstruction, but using a reduced parametrization and calling the fast simulation methods.

Modelling Correlated Fluctuations: Results

ATL-SOFT-PROC-2020-027



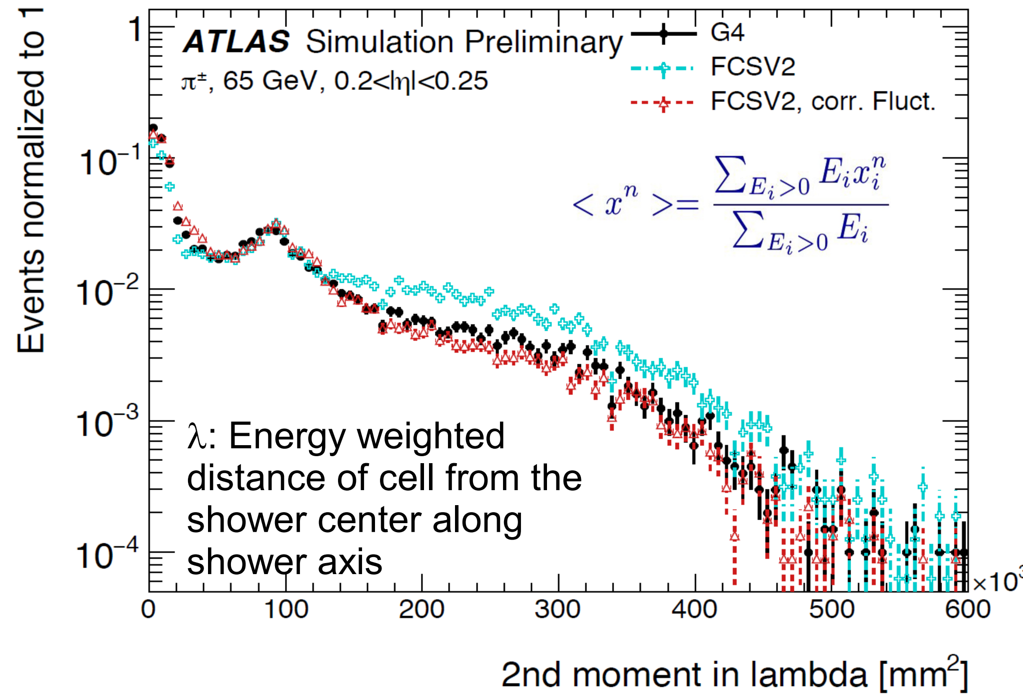


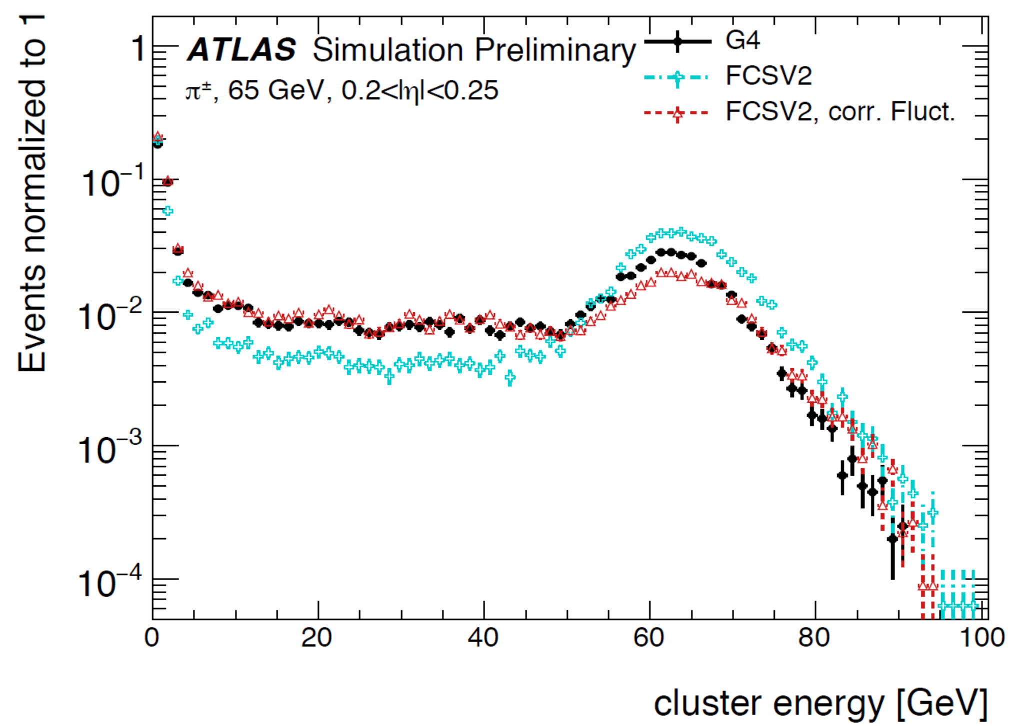
These plots are made after the full MC chain is run in Athena:

FastCaloSim
→ Digitization
→ Reconstruction

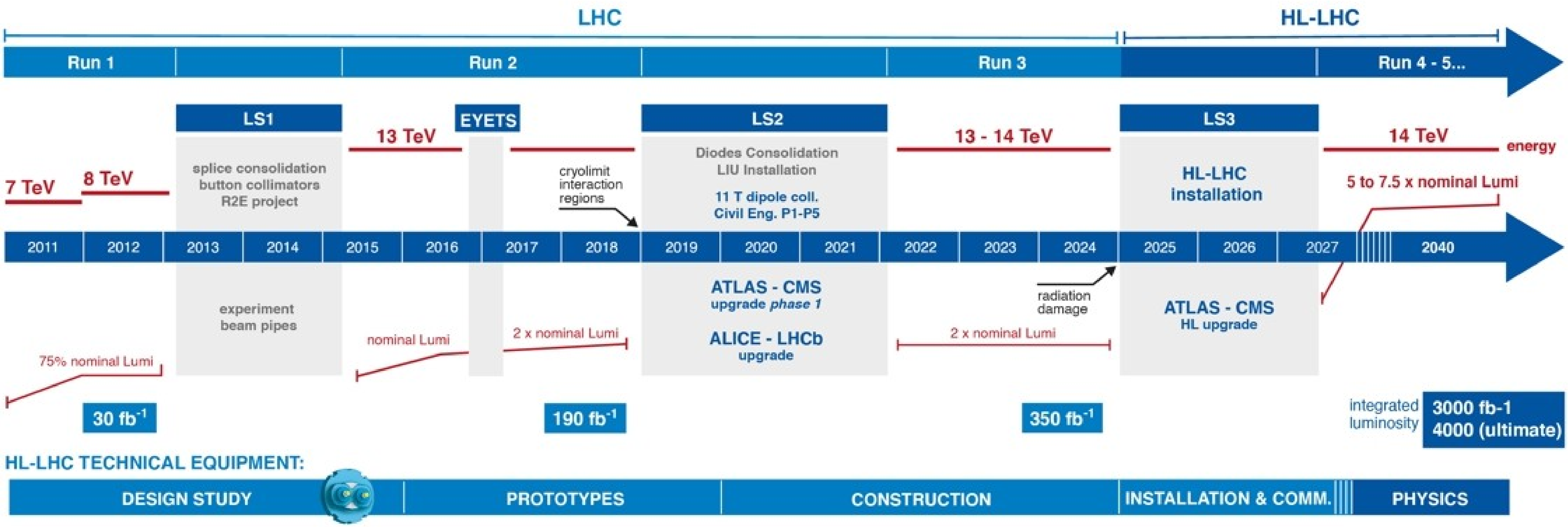
Ultimate proof of concept would be validations of boosted objects and looking at jet sub-structure variables (TO DO)

Here the Multi-Gaussian is used.





LHC / HL-LHC Plan



- High Luminosity LHC expected to run 2027-2040 (after run-3 will end in 2024)
- HL-LHC will deliver 3000/fb (or more) at $\sqrt{s}$=14 TeV
- Several detector upgrades planned (new electronics readout, new inner tracker, new trigger, …)
- Expect average pile-up of $\langle\mu\rangle$=200 interactions

Energy resolution in the calorimeter

$$\frac{\Delta E}{E} = \frac{\alpha}{\sqrt{E}} \oplus \beta \oplus \frac{\gamma}{E}$$

α: Sampling term (choice of active/passive material, fluctuations in number of charged particles passing through active layers)

β: Constant term (cracks, dead material, dominant at high energies)

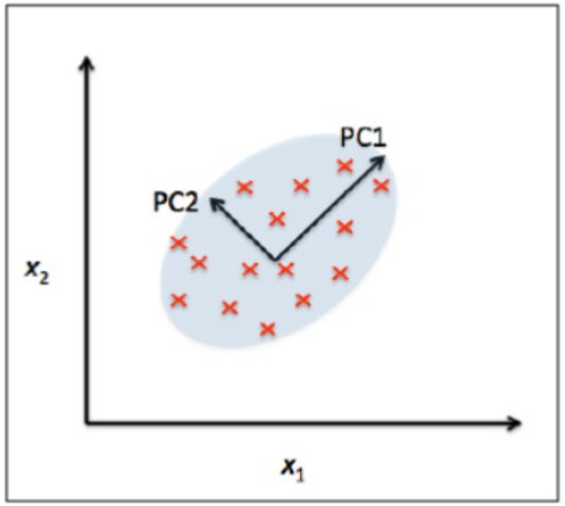
γ: noise term (electronics, dominant at low energies)

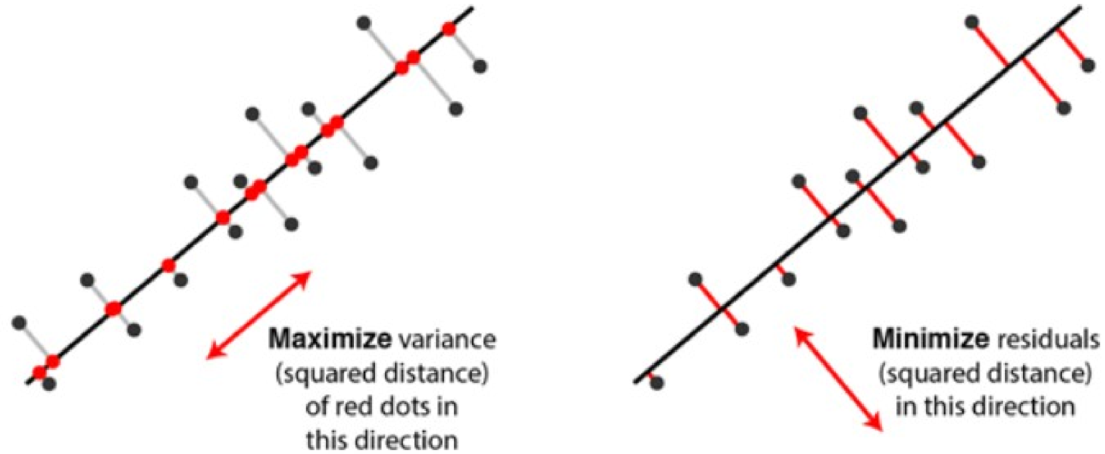
ATLAS calorimeter design resolution:

	Resolution
EM Barrel	$\frac{\sigma_E}{E} = \frac{10\%}{\sqrt{E}} \bigoplus 0.7\%$
EM End-Cap	$\frac{\sigma_E}{E} = \frac{10\%}{\sqrt{E}} \bigoplus 0.7\%$
HEC	$\frac{\sigma_E}{E} = \frac{50\%}{\sqrt{E}} \bigoplus 3\%$
FCAL	$\frac{\sigma_E}{E} = \frac{100\%}{\sqrt{E}} \bigoplus 10\%$

Gentle Introduction to Principal Component Analysis (PCA)

PCA is based on linear algebra and a widely used technique, applied in many fields from neuroscience to computer graphics. In the language of ML it is an unsupervised, non-parametric technique for dimensionality reduction of complex datasets.





PCA identifies a list of principle axes, and then ranks them according to the amount of variance:
First principal component expresses the most amount of variance.
Each additional component is then orthogonal, and expresses less variance.
→ The most interesting dynamics occur for the first k dimensions.

The PCA is unique, ie. there are no parameters to tweak for this method.

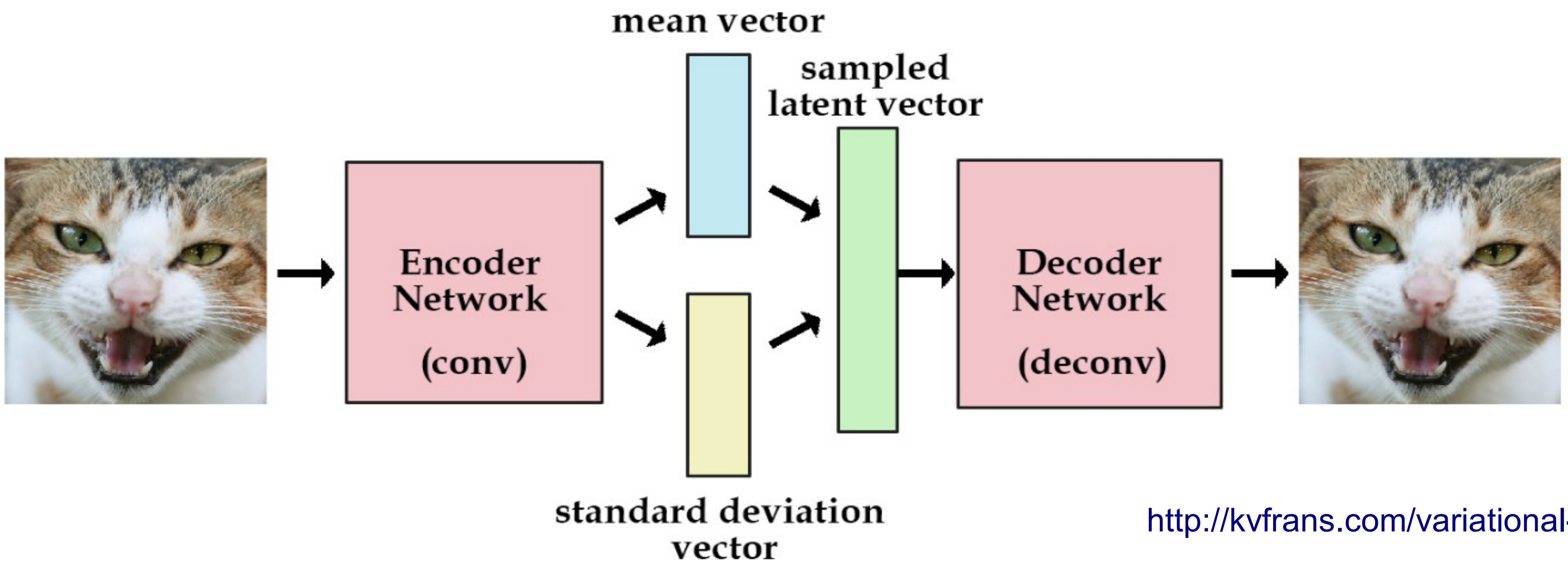
However, in a first step it often makes sense to transform the data into appropriate coordinates, (kernel transformation, eg. Gaussian transformation). This step is parametric.

Variational Auto Encoder (VAE)

A VAE consists of two connected neural networks, an **encoder** and a **decoder**, and a **loss function.**

- The **encoder** maps each input to a point and some spread in a lower dimensional **latent space** (using efficient compression)
- The **decoder** neural network maps from the latent space back to the “input space”, with the goal of reconstructing the given input.

To go from a standard encoder (that for example can memorize cat images) to a **generative model** (that can generate a new cat image), add a constraint in the encoder to force it to generate latent vectors that roughly follow a given prior. Therefore the loss function has two components:
- generative loss (measures how accurately the network reconstructs an image)
- latent loss (Kullback-Leibler divergence, measures how well the latent variables matches the prior)



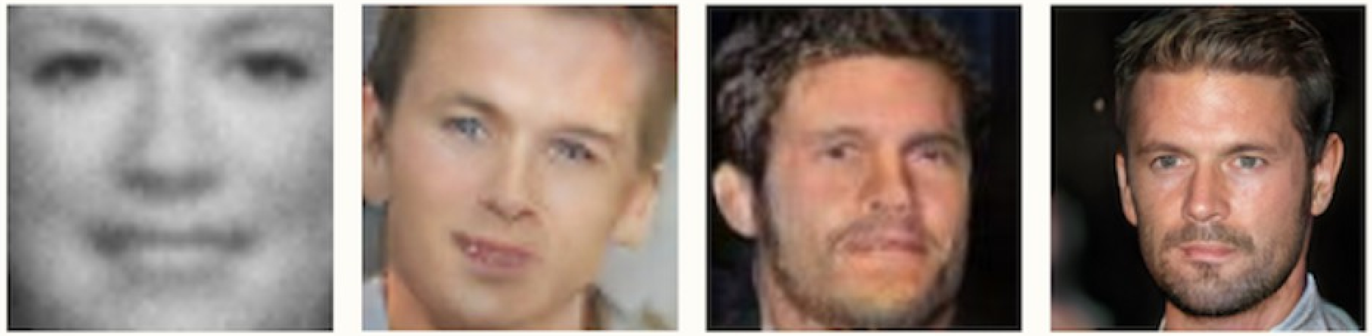
http://kvfrans.com/variational-autoencoders-explained/

Generative adversarial networks (GAN)

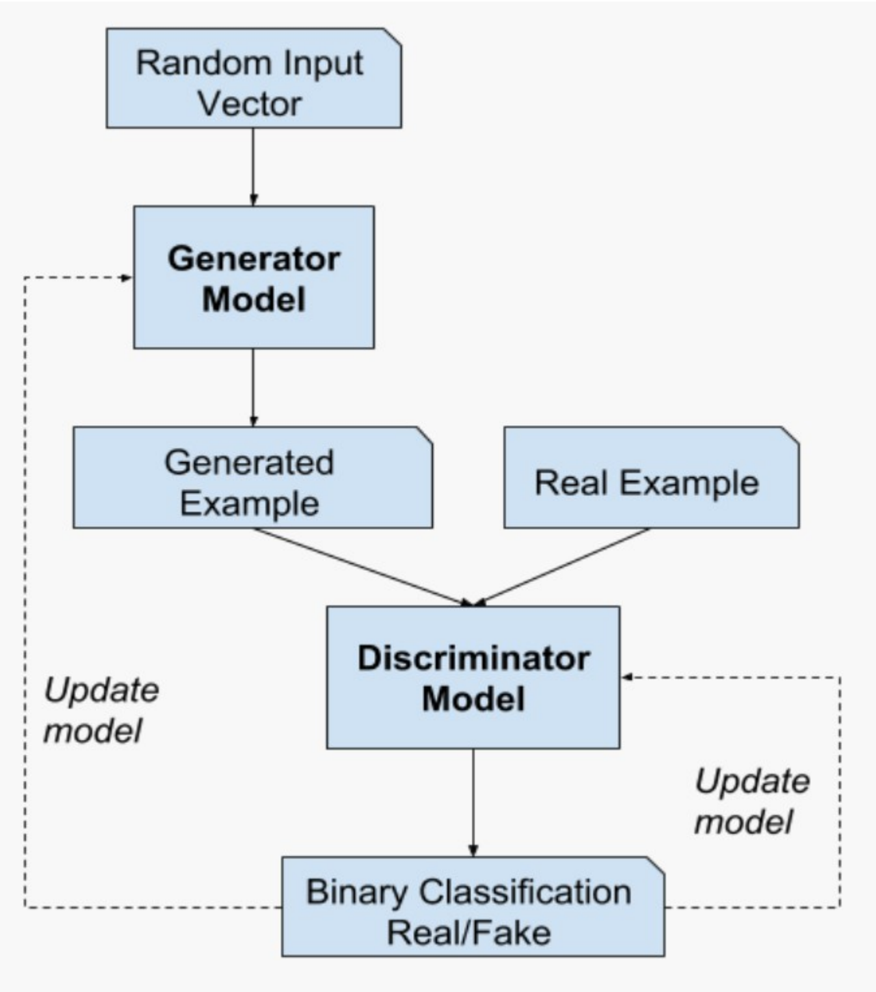
Very popular generative technique

Tremendous progress over the past years in creating for example artificial photos

Goodfellow 2014 https://arxiv.org/abs/1406.2661



Example of the Progression in the Capabilities of GANs From 2014 to 2017. Taken from The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation, 2018.



Example of the Generative Adversarial Network Model Architecture

The GAN model architecture involves two sub-models:

- **Generator** model for generating new data instances

- **Discriminator** model for classifying whether generated examples are real, from the domain, or fake, generated by the generator model

The generator output is connected directly to the discriminator input. Through back-propagation, the discriminator's classification provides a signal that the generator uses to **update** its weights.

Proof of concept to apply this method to shower simulation: CaloGAN 2018 https://arxiv.org/pdf/1712.10321.pdf