Fellini
Fellowship for Innovation at INFN

H2020 MSCA COFUND
G.A. 754496

INFN
Sezione di
Roma Tor Vergata

Lessons learned on pre-processing in FastCaloGAN

Michele Faucci Giannelli

03-06-2021

FastCaloGAN in one slide

- Use the 4500 single-particle samples produced by ATLAS as part of the effort to create a new Fast Calorimeter Simulation (FCSV2)
  - 3 particles, 100 η slices, 15 energy points per slice
  - Energies from 256 MeV to 4 TeV (in powers of two)
- Define a GAN for each particle in each η slice → 300 GANs
- Train the GANs on voxelised hits (see backup for hit definition)
  - Because cell structure is not homogeneous and would require different GAN architectures, voxelisation allow to reduce the differences
- Select best epoch based on total energy distribution of each sample
- Convert the selected generator into LWTNN
- Simulate hits in Athena inverting voxelisation
- Compare with Geant4 using high-level observables for single-particle and di-jet samples after reconstruction
- More information:
  - Pub note: ATL-SOFT-PUB-2020-006
  - Presentations at IML

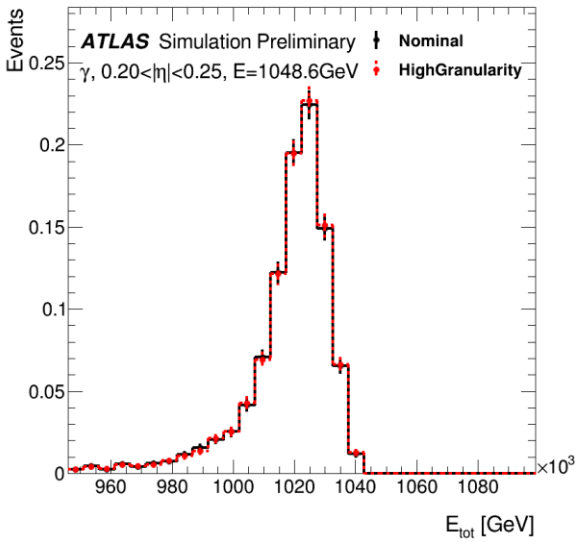
The training sample

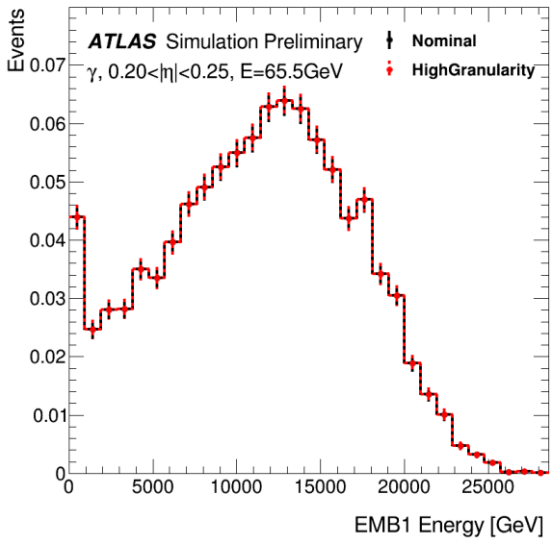
- The single-particle samples have:
  - detailed hits (i.e. with a step << cell size) to better map ATLAS cells, the position is stored
  - without z-vertex spread
  - without noise
  - without parts of the electronic cross talk
- Events generated by momentum
  - But used the Ekin for the simulation
- The statistics in the high energy region is lower than at low energies
  - 10k events up to 256 GeV
  - Drop to 1k events for 4 TeV samples

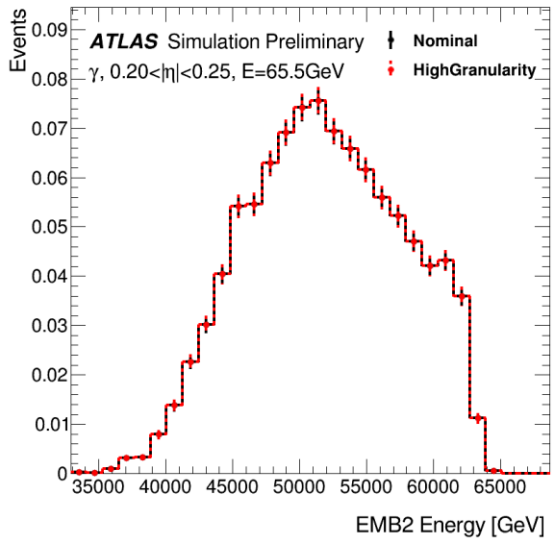
Voxelisation

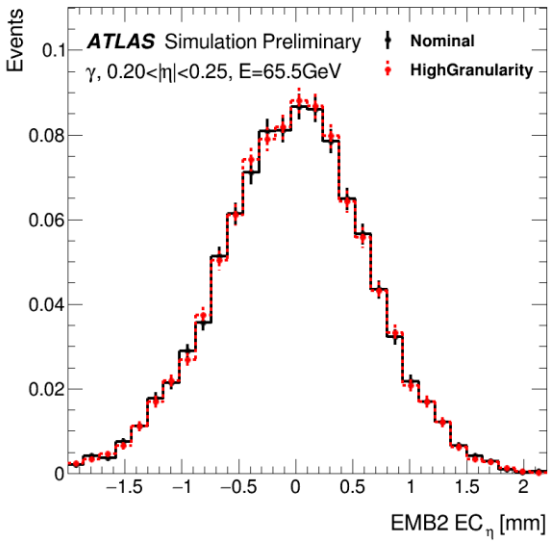
- Hits are transformed from ATLAS (x,y,z) coordinates to cylindrical (r, α, R) coordinates
  - R is not use, as hits are grouped in layers
- GAN cannot be trained on hits, so they are grouped in areas in the (r, α) plane in each layer
  - each volume in the (r, α, layer) space is called a voxel
- We store the energy in each voxel in csv files
  - Simple to read with *pd.read_csv()*

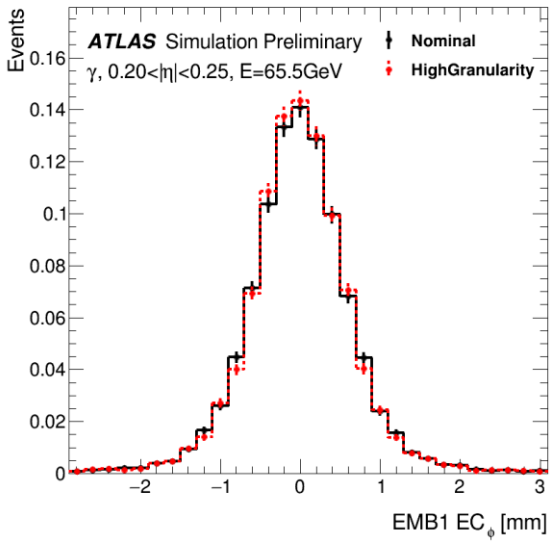
Validation

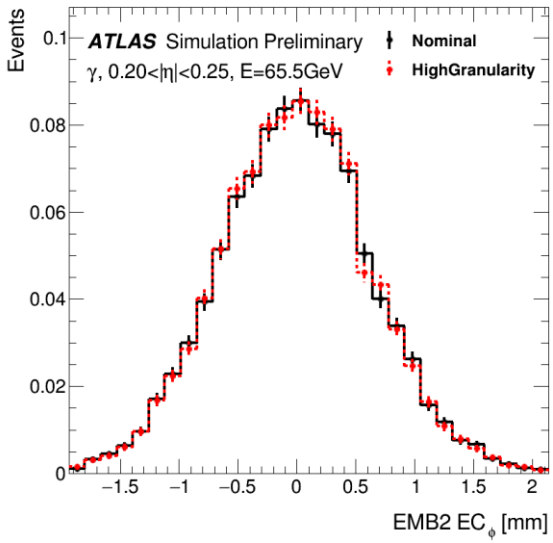

ATLAS Simulation Preliminary
Nominal
HighGranularity
γ, 0.20<|η|<0.25, E=1048.6GeV
Events
E_tot [GeV]
×10^3


ATLAS Simulation Preliminary
Nominal
HighGranularity
γ, 0.20<|η|<0.25, E=65.5GeV
Events
EMB1 Energy [GeV]


ATLAS Simulation Preliminary
Nominal
HighGranularity
γ, 0.20<|η|<0.25, E=65.5GeV
Events
EMB2 Energy [GeV]


ATLAS Simulation Preliminary
Nominal
HighGranularity
γ, 0.20<|η|<0.25, E=65.5GeV
Events
EMB2 EC_η [mm]


ATLAS Simulation Preliminary
Nominal
HighGranularity
γ, 0.20<|η|<0.25, E=65.5GeV
Events
EMB1 EC_φ [mm]


ATLAS Simulation Preliminary
Nominal
HighGranularity
γ, 0.20<|η|<0.25, E=65.5GeV
Events
EMB2 EC_φ [mm]

Lesson learned: speed

- Processing 4500 samples takes time and is very I/O intensive
  - Compiling code and not running on EOS speed things up by about a factor 10
- Running the voxelisation for the whole detector takes several days on HTCondor
  - Another good reason to keep the GAN separate by slice, it's much faster to assess the performance of different voxelisations
- Optimisation of voxelisation is absolutely not trivial, many iterations are needed
  - Unfortunately some effect only visible after training, simulation and reconstruction, which takes a lot of time
  - More pre-training validation is needed

Lesson: abstract the voxelisation

- We have several scripts (training, validation, plotting, TF2LWTNN conversion, simulation, …)
- At the beginning we had the voxelisation hardcoded but it was a nightmare to change it
- We settled on a XML in which we define the bins in r and alpha in each layer for each particle in different detector regions
  - Used everywhere, from voxelisation to simulation in Athena

```
<Bin pid="22" etaMin="0" etaMax="130">
    <Layer id="0"  r_edges="0,5,10,30,50,100,200,400,600"                                  n_bin_alpha="1" />
    <Layer id="1"  r_edges="0,2,4,6,8,10,12,15,20,30,40,50,70,90,120,150,200"              n_bin_alpha="10"/>
    <Layer id="2"  r_edges="0,2,5,10,15,20,25,30,40,50,60,80,100,130,160,200,250,300,350,400" n_bin_alpha="10"/>
    <Layer id="3"  r_edges="0,50,100,200,400,600"                                          n_bin_alpha="1" />
```

Further manipulation

- Before feeding the csv information to the GANs, two normalisations are applied
- The energy in the voxels is normalised to the sample true energy
  - This allow to focus on the shape rather than the absolute value of the energy
- The conditional parameter (momentum of particle) is normalised to the highest energy (4TeV) to that the labels are in the range (0,1]