Fellini
Fellowship for Innovation at INFN

H2020 MSCA COFUND
G.A. 754496

INFN
Sezione di
Roma Tor Vergata

Data pre-processing

Michele Faucci Giannelli,
On behalf of the ATLAS collaboration

LPCC Fast Detector Simulation
22-11-2021

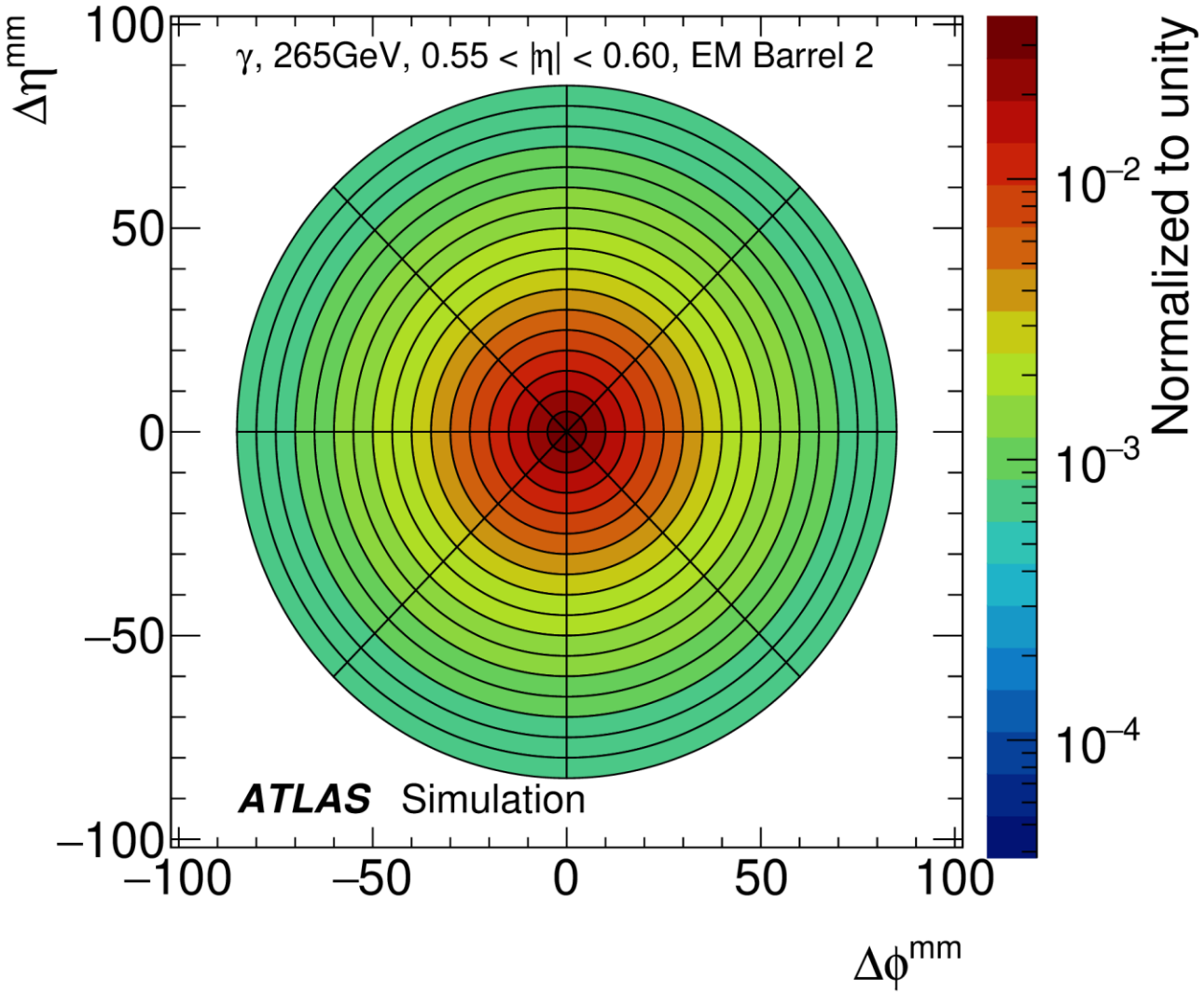
Data pre-processing

- Which particle types and features (energy, angle) are considered?
  - 3 particles were considered: photons, electrons and charged pions
    - Other hadrons only considered for corrections
  - The ATLAS calorimeter was segmented in 100 slices in η
  - For each particle and η slice, 17 energy points (64 MeV – 4 TeV in power of 2) were produced
    - FCSV2 use all samples for e/gamma but start at 256 MeV for pions
    - FastCaloGAN starts at 256 MeV for all particles,
    - VAE and GANs at cell level used only a reduced range (1GeV-1TeV)
    - The cell based GANs only simulated photons in a single η slice
  - Particles are generated at the calorimeter surface pointing to the IP, no position smearing or calorimeter noise is used.
  - Use “FCS hits” which contain the information from the G4steps, AFTER going through the ATLAS calculators, with a granularity of 1mm/5mm.

Data pre-processing

- How large is the training dataset?
  - Each sample has 10k events up to 256 GeV, then the size is reduced to 5k, 3k, 2k and 1k for 0.5, 1, 2 and 4 TeV respectively.
- How to define the structure of the input data (hits, cells, clusters, custom voxels,..)?
  - FCSV2 uses 2D histograms to store highly granular voxelisation in polar coordinates (r, α)
    - 1mm/5mm along r, 8 bins in α (phi-symmetric)
  - FastCaloGAN uses a similar voxelisation strategy but with a coarser granularity in r
  - The cell-based GAN used cells
  - VAE was trained on both cells and hits merged in custom voxels optimised using ML

Example of voxelisation: FCSV2



FastCaloGAN binning

Layer	r edges [mm]	N bins α
PreSamplerB	0,5,10,30,50,100,200,400,600	1
EMB1	0,2,4,6,8,10,12,15,20,30,40,50,70,90,120,150,200	10
EMB2	0,2,5,10,15,20,25,30,40,50,60,80,100,130,160,200,250,300,350,400	10
EMB3	0,50,100,200,400,600	1
TileBar0	0,100,200,400,1000,2000	1

Data pre-processing

- Which data structure is used for the training (1D vector, images, graphs..)?
  - 1D vector,
    - changing from narray to tensor doubled our GPU utilisation!
- Which scaling is used?
  - In GANs the energy in the voxels is normalised to the nominal energy of the sample
  - In VAE each layer is normalised to the total energy and each voxel is normalised to the energy in the layer
- Is the dataset balanced (is the number of events for the different particle properties almost the same?).
  - Not for very high energy

Data pre-processing

- How do you store the preprocessed data?
  - FastCaloGAN uses csv files
  - Other approaches use HDF5
- How are the condition values (energy of the particle, angle,..) encoded?
  - All approached condition on energy of particle,
  - FastCaloGAN trains one GAN per η slice, since events are homogeneous there is limited benefit in adding a conditioning on η
  - VAE for full-detector use conditioning on energy and η