Applications of machine learning techniques for the CMS detector at the HL-LHC

Gustavo Gil da Silveira

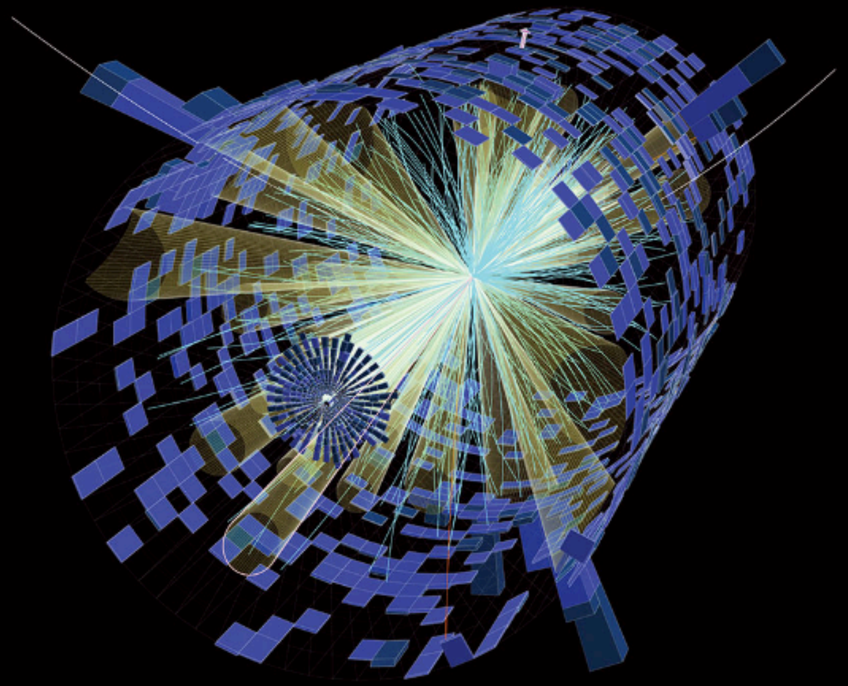
with A. Sznajder, E.G. Brock, V. S. Sousa, J.A.Chinellato
(UFRGS, UERJ, and UNICAMP)

on behalf of the UERJ CMS-Rio group

Workshop da RENAFAE 2021

12–14/julho/2021

Nanosecond Jet Classification at the L1 Trigger for HL-LHC



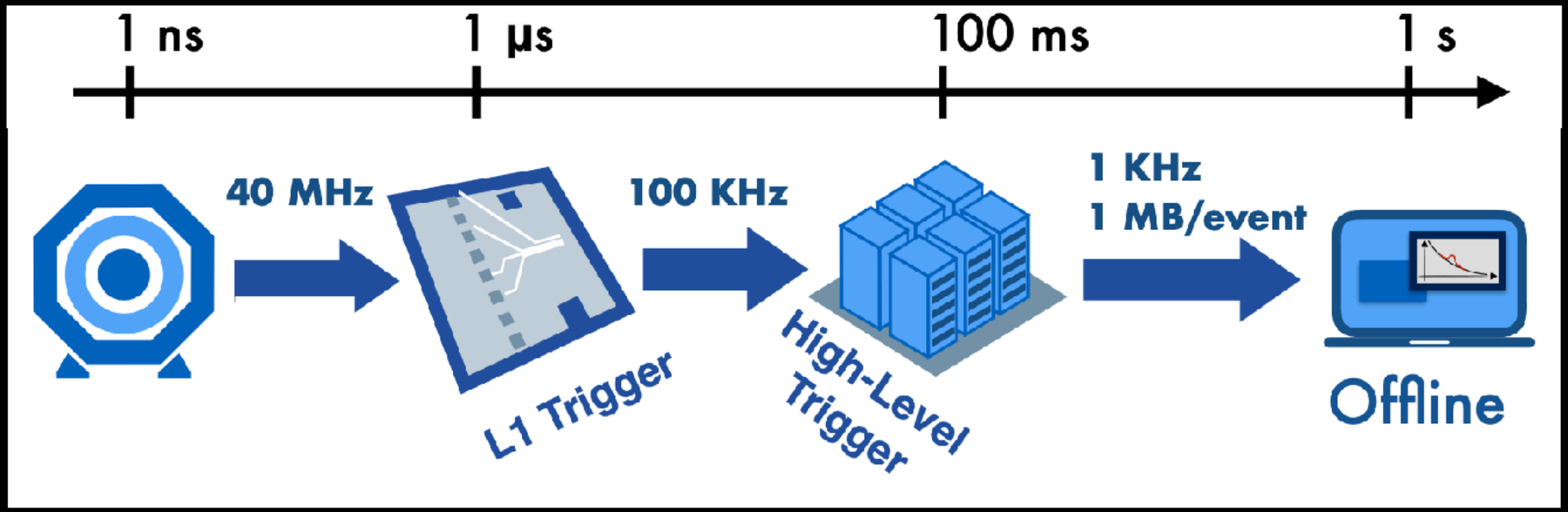
Andre Sznajder
UERJ (Brazil)

In Collaboration with:
M.Pierini(CERN)
T.Aarrestad(CERN)
S.Summers(CERN)
J.Ngadiuba(FNAL)
V.Loncar(CERN)

Work presented and the ML4Jets workshop (Heidelberg July 6-8 ,2021)
https://indico.cern.ch/event/980214/contributions/4413606/

LHC Event Processing

Multiple filtering stages to reduce data rates to manageable levels



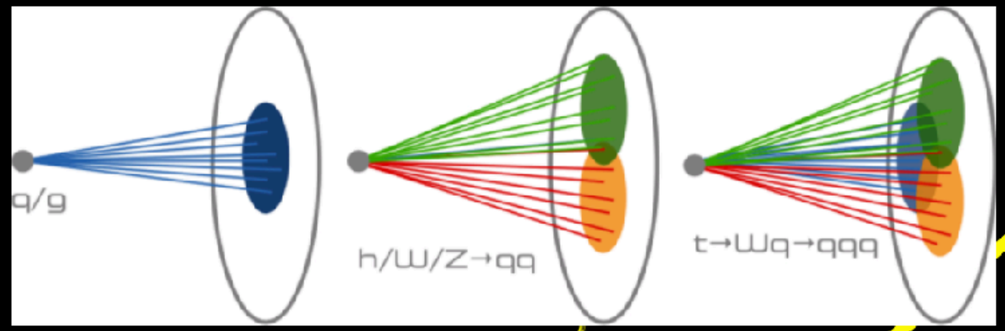
- L1 trigger absorbs O(100 TB/s)
- About 99% of events is rejected (needs high purity trigger)
- Trigger decision to be made in O(µs)

=> Latencies imposes an all FPGA design for L1 !

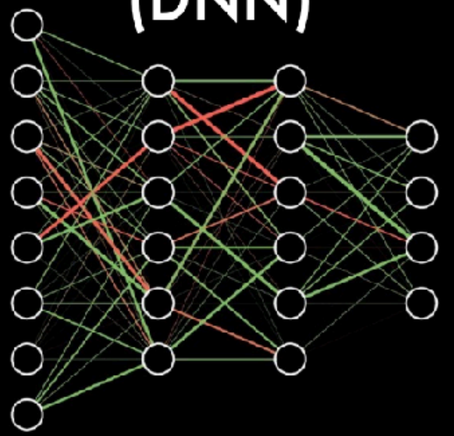
Jet ID at the L1 trigger

Proposal: recast L1 trigger problem into a ML problem !

Jet ID



Deep Neural network (DNN)



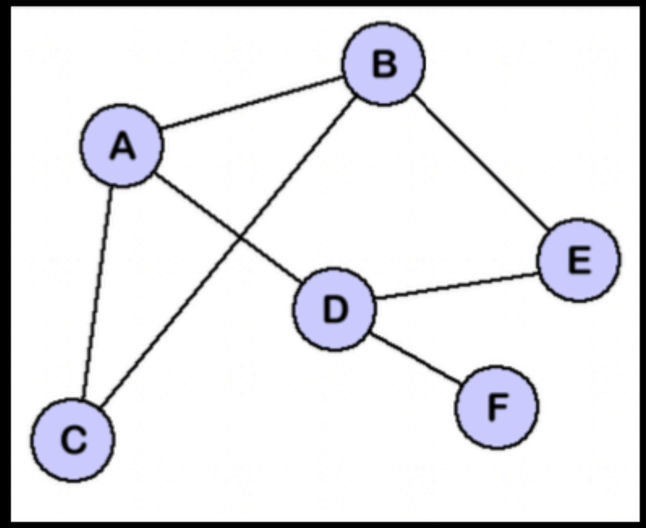
- Feed a DNN with L1 jets constituents and train it as classifier
- While training might take long time, inference is very fast

Challenge: need to implement the DNN in a FPGA to conform with L1 hardware !

Neural Network Models

L1 provides unordered jet constituents
=> Use Graph NN as jet classifier

- constituents (graph nodes) features (Pt,η,ϕ)
- L1 provides fixed #constituents, so use fixed graph size (fully connected)



σ
activation function

W_{ij}
weight matrix

B_i
bias

M
message passing

MLP

$$x'_i = \sigma\left(\sum_j W_{ij} x_j + B_i\right)$$

- Dense layers
- Used as standard candle

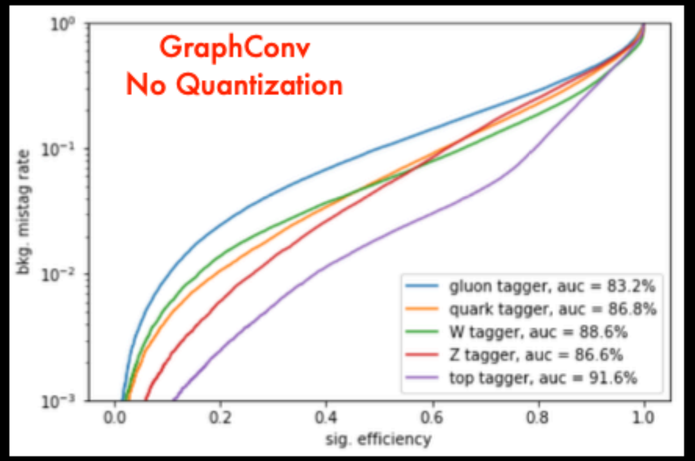
GraphConv

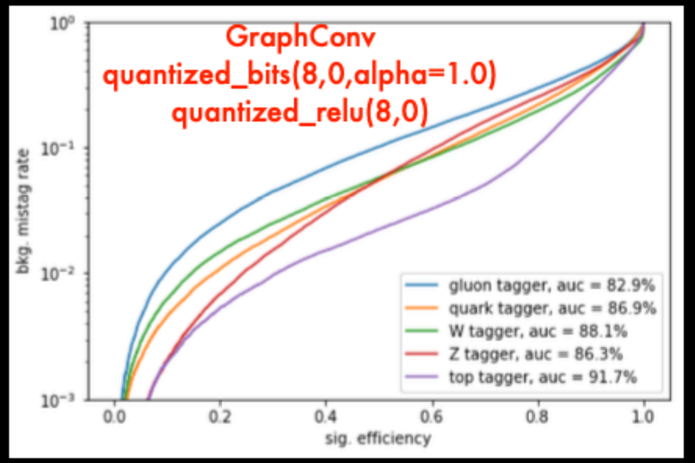
$$x'_i = \sigma\left(\bigoplus_{j \in \mathcal{N}_i} M(x_i, x_j)\right)$$

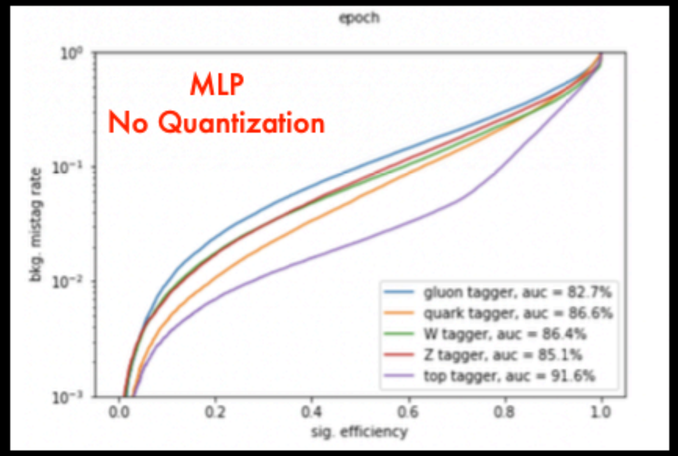
- Graph Convolution applied to graph nodes
- Dense layer for classification

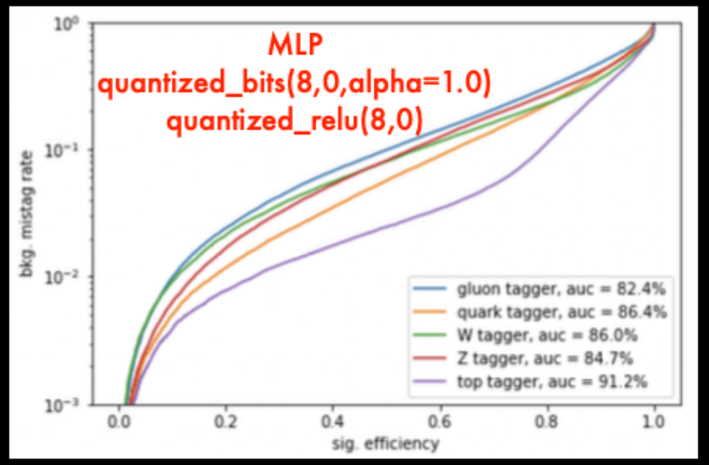
QKERAS - Quantization Aware Training

QKeras is a library for quantization aware training (QAT) of Keras models, to study the quantisation impact on networks. It quantizes layers , weights, bias and activations, and uses limited precision for forward pass.









keras.io

google/qkeras

HLS4ML

After KQERAS study of best model we use HLS4ML package to synthesize the network model into the FPGA firmware.

MODEL	DSP	LUT	FF	BRAM	Latency
MLP (8 bits)	300 (4.4%)	71917 (6.1%)	10555 (0.4%)	1.5 (0.07%)	40.0ns ii:1
GraphConv (8 bits)	43 (0.6%)	80977 (6.8%)	11987 (0.5%)	1.5 (0.07%)	85.0ns ii:13

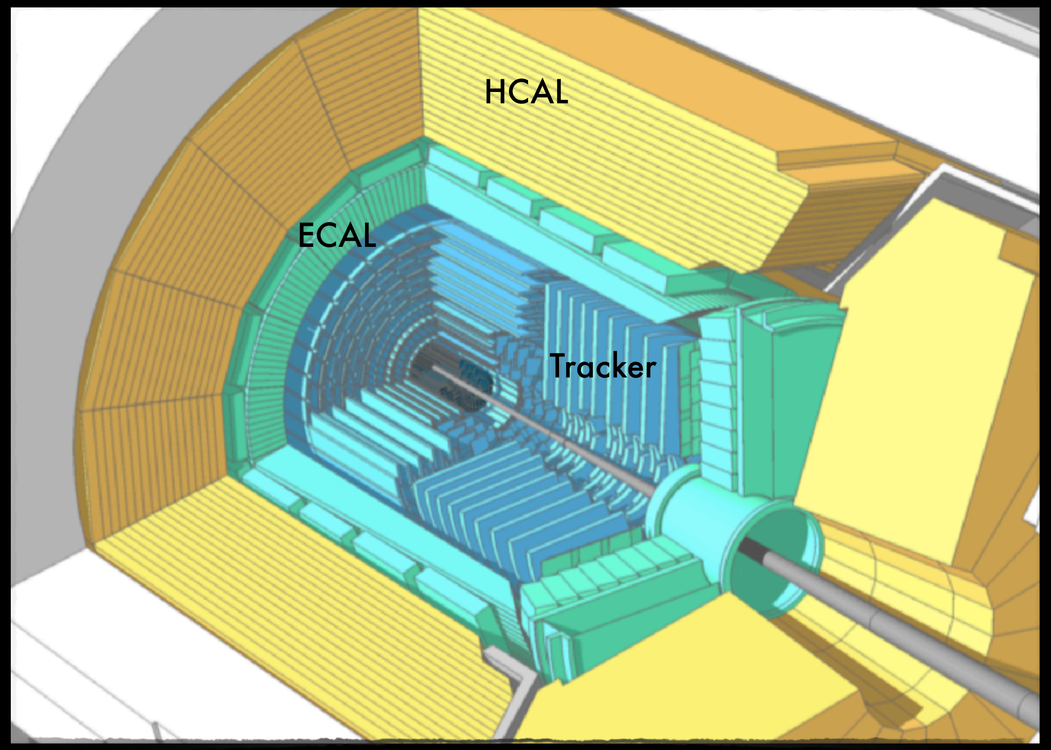
Xilinx™

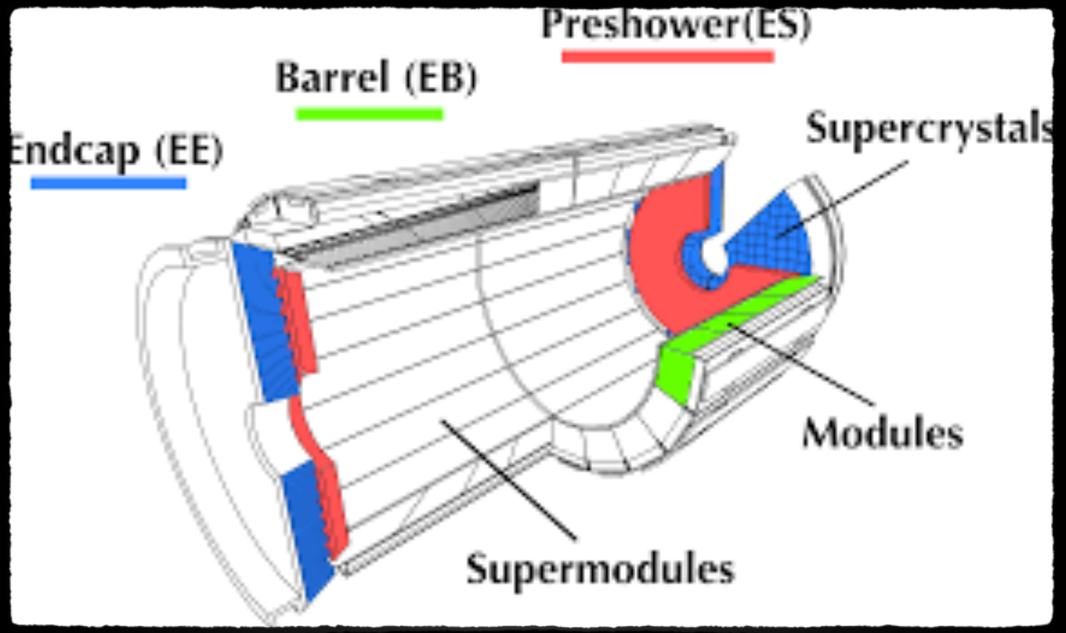
- HLS4ML shows models latency well within L1 trigger O(μs) constraint
- FPGA resource usage compatible with Xilinx Virtex Ultrascale 9+ specs

This work is the first time a Graph Neural Network is implemented on a FPGA for the HL-LHC L1 trigger !!!

Simulation of showers in CMS ECAL with generative adversarial networks

Gustavo Gil da Silveira, Eduardo Brock (UFRGS)



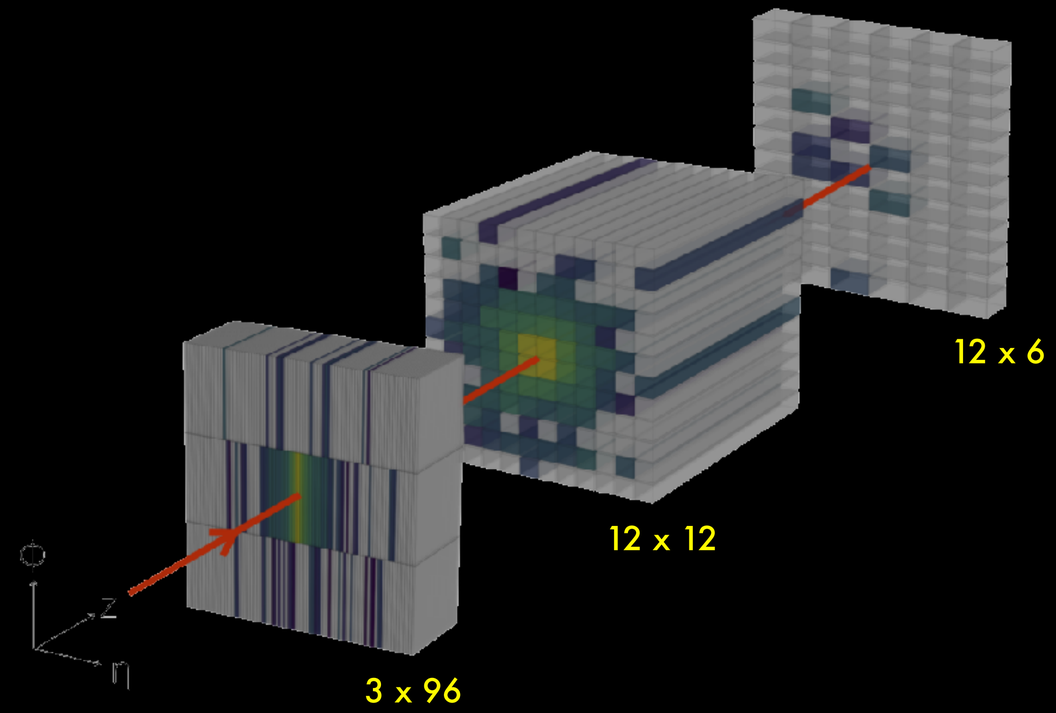


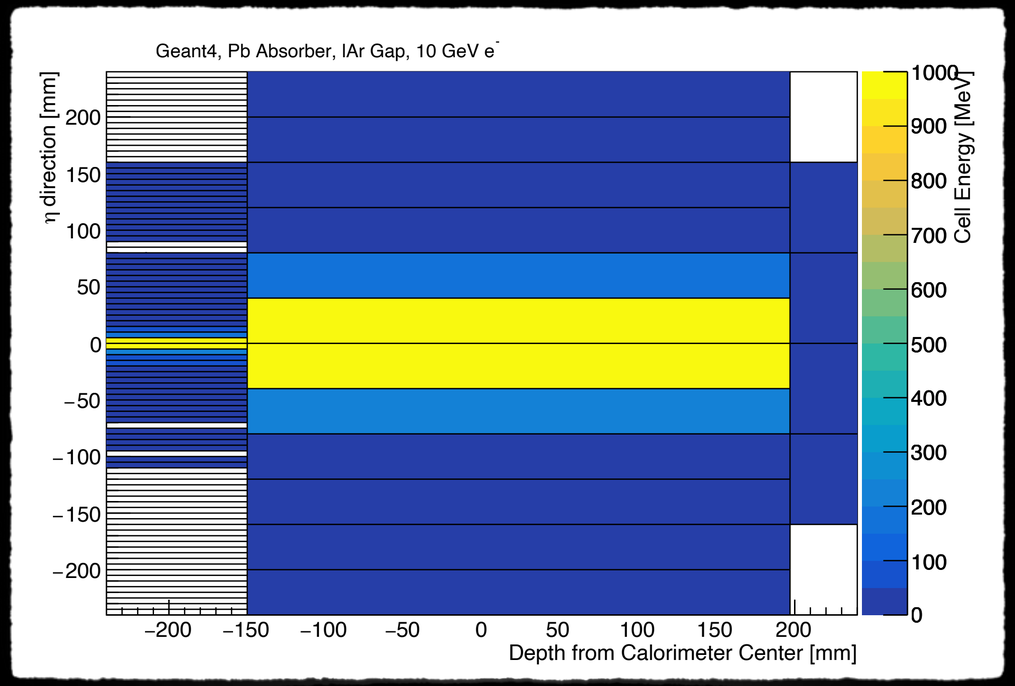
C. Biino, J. Phys.: Conf. Ser. 587 (2015) 012001

M. Paganini, L. de Oliveira, B. Nachman
Phys. Rev. D 97, 014021 (2018)
arXiv:1712.10321v1 [hep-ex]

CaloGAN simulation

- This study is based on the CaloGAN package for LAr of the ATLAS ECAL;
- The strategy slices the sensitive detector into sections to be used as layers for a generative adversarial network (GAN) to be trained.

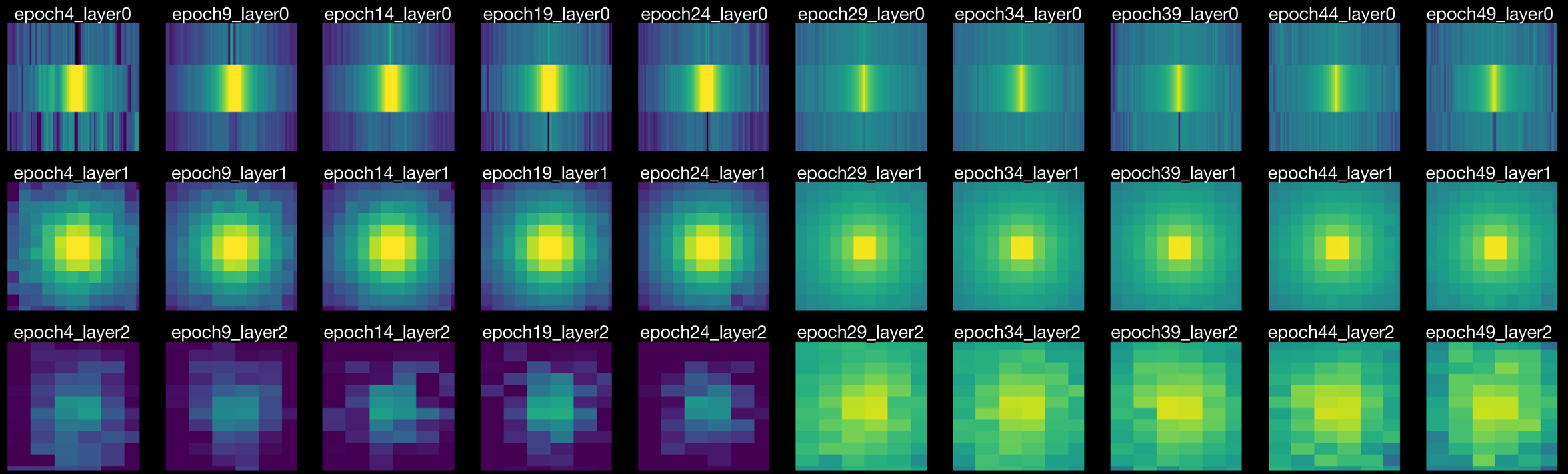




M. Paganini, L. de Oliveira, B. Nachman
Phys. Rev. D 97, 014021 (2018)
arXiv:1712.10321v1 [hep-ex]

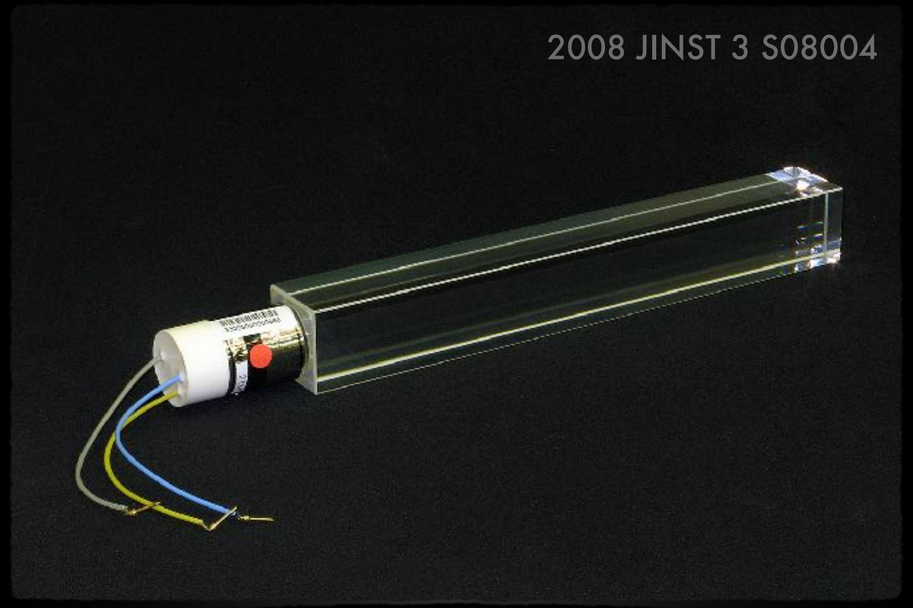
GAN training

- Positrons, photons, and pions are used as incident particles for training;
- Sampling over the 3 layers to train the energy deposition in different granularities;
- GPU training ~3x faster than usual CPU.



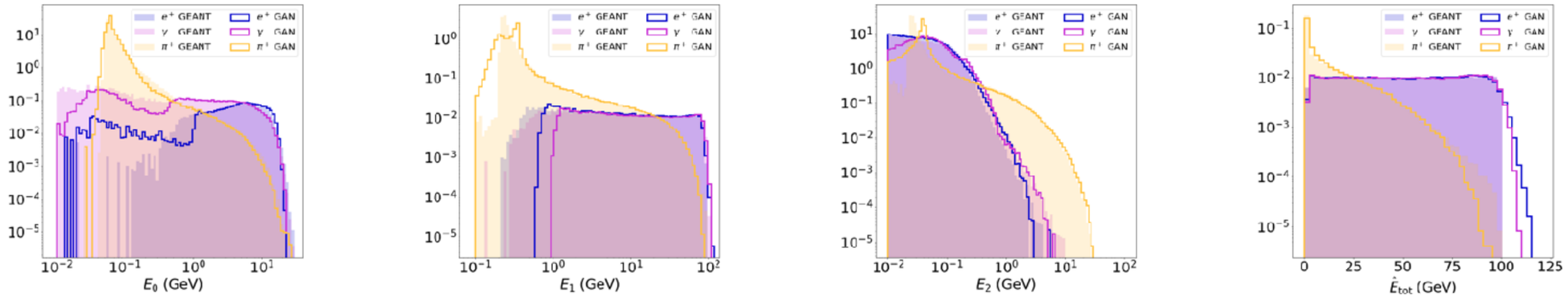
CMS ECAL

- PbWO4 crystals used as sensitive detector, updating the detector length to the 25 cm of the CMS crystals;
  - Pb layer of 3 mm is used as preshower to enlarge shower area.
- The detector description follows a close design of the ECAL modules in the CMS detector:
  - Different layer binning for better description;
  - 10% – 70% – 20% of total length;
- Distributions investigated: Energy ratio, particle fractions per layer, max depth, etc.

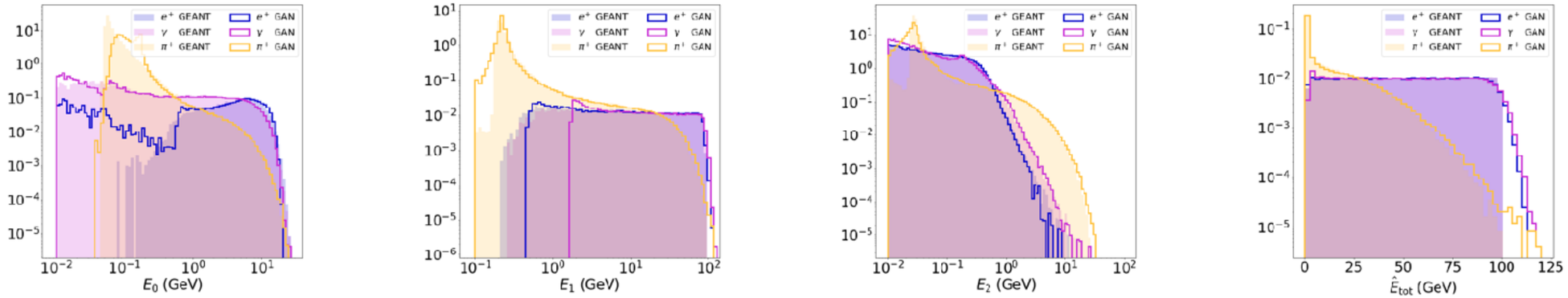


Energy distributions

ATLAS LAr ECAL with 48 cm



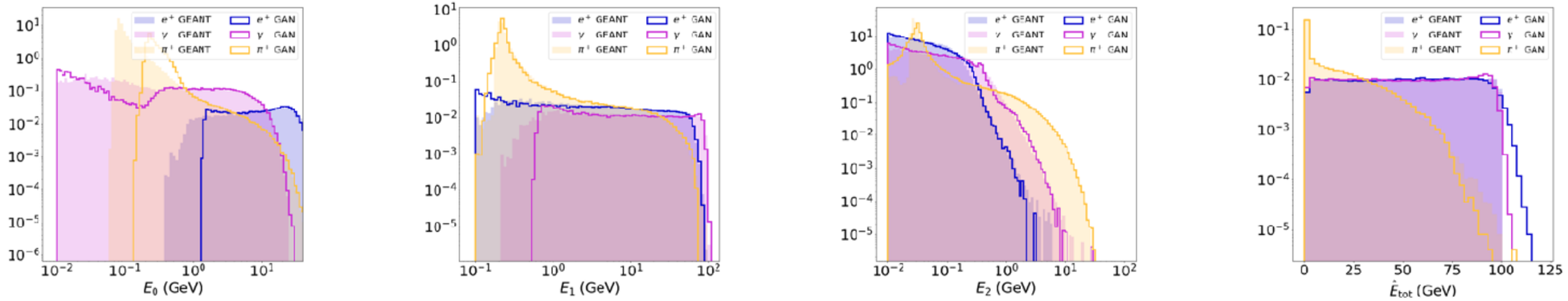
CMS PbWO4 ECAL with 25 cm



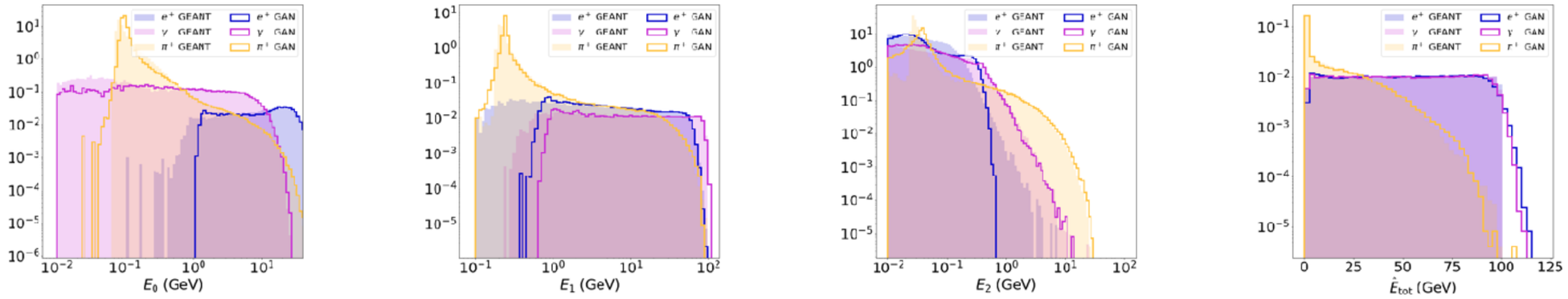
Google Colab: GPU K80 ~8h training
CERN LXBATCH: GPU Tesla T4 ~5h training

Adding the preshower

CMS PbWO4 ECAL with 25 cm + preshower with 20 x 20 crystals in setup #1



CMS PbWO4 ECAL with 25 cm + preshower with 20 x 20 crystals in setup #2



Google Colab: GPU K80 ~8h training
CERN LXBATCH: GPU Tesla T4 ~5h training

Processing power with GRID T3s

- **T3_BR_UFRGS**:
  - Current T3 unit with 768 GPU cores and large storage area;
  - Focus on local development of ML algorithms and GAN training.
- **UNICAMP** (plan): J.A. Chinellato in collaboration between IFGW + FEEC
  - hybrid computational model for large-scale processing based on CPUs + GPUs + FPGAs;
  - Processing cases relative to energy deposition, such as dE/dx electromagnetic cascades, with GPU + FPGA accelerators;
  - Environment for development of machine learning algorithms, e.g. statistical deconvolution, with bayesian approach.

Overview

- Timescale:
  - GNN: tools available and possible tests in LHC Run3 but focus on HL-LHC (2025);
  - Calo: PhD student (4yr) for possible application meant to CMS ECAL or HGCal.
- Estimated costs:
  - Improvement of T3 processing power with new stations: R$ 30.000,00 (FAPERGS)
  - Proposal of T3 with hybrid instance: possible funding from CNPq or FAPESP
  - Budget for in-person work at CERN: R$ 100.000,00 (FAPERGS, CAPES, FAPERJ)
    - 6-month PhD internship at CERN and short-term visits.

Possible extensions

- Synergies:
  - Build closer collaboration with other ML groups in Brazil and at CERN;
  - Cross-experiment simulations for electromagnetic showers with GPU;
  - Train students for future developments and extensions for data analyses.
- Spin-offs:
  - Application of ML simulation for proton reconstruction in CMS-PPS;
  - Simulation of electromagnetic showers in future detectors (ILC);
  - Flexible processing infrastructure to extend additional hardware-level studies.

Closing remarks

- ML applications is a fast growing research field with great potential for high-efficient computing;
- HL-LHC demands new solutions to overcome the enormous amount of data and complex pileup environment in view of data analyses;
  - Trigger and simulation/reconstruction are the main drives of processing time.
- Large-scale computing resources are available at CERN, but preparing a local infrastructure allows more profit collaborations with cross-experiment researchers;
- Improvement of expertise to pursue more sophisticated project in the future;
- Current funding from: