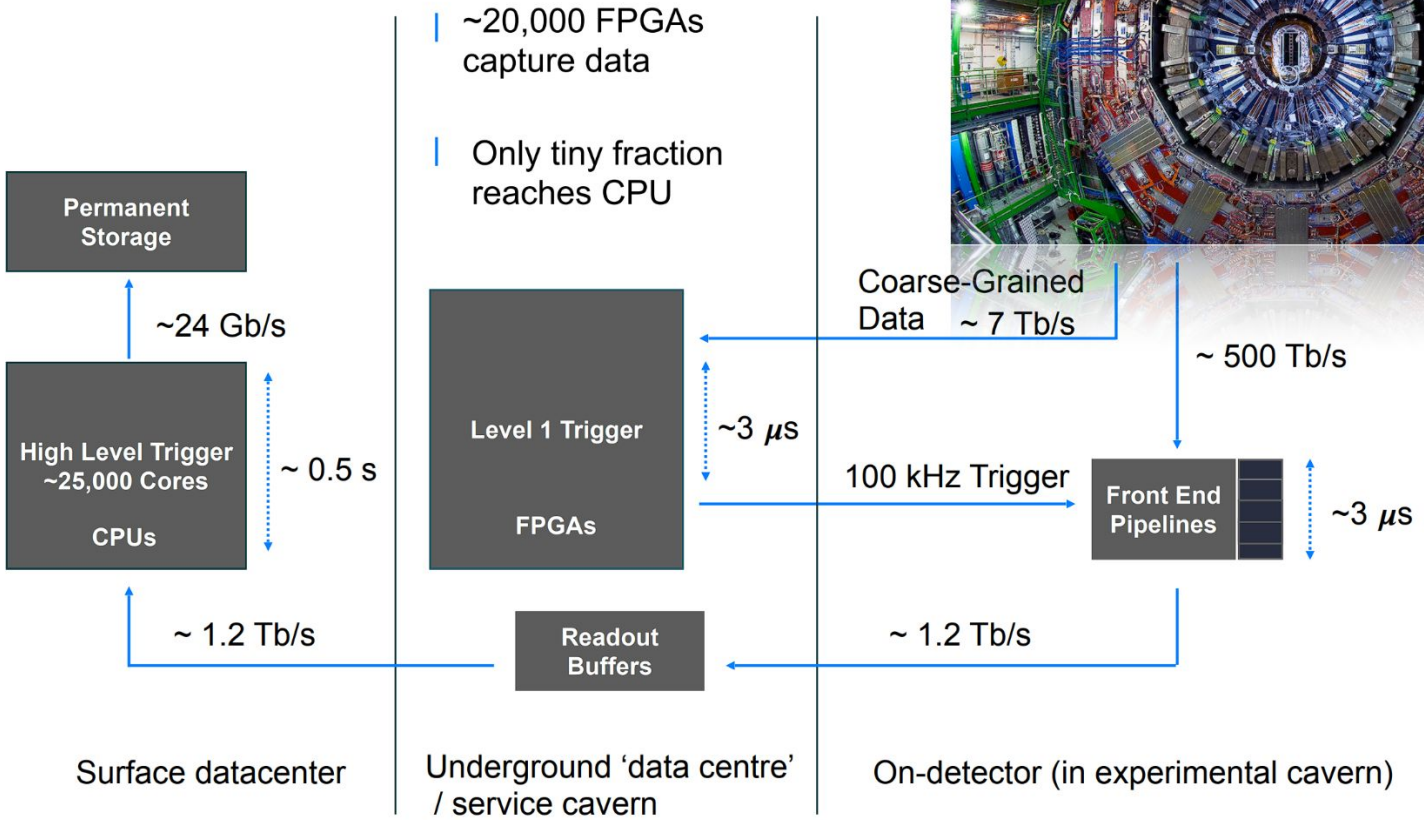
FPGA-Accelerated Neural Network Inference for Ultra-Low-Latency Recalibration and Classification of Physics Objects at 40 MHz within CMS

Student: ***Diptarko Choudhury***

Supervisors: ***Rocco Ardino*** *and* ***Thomas Owen James***

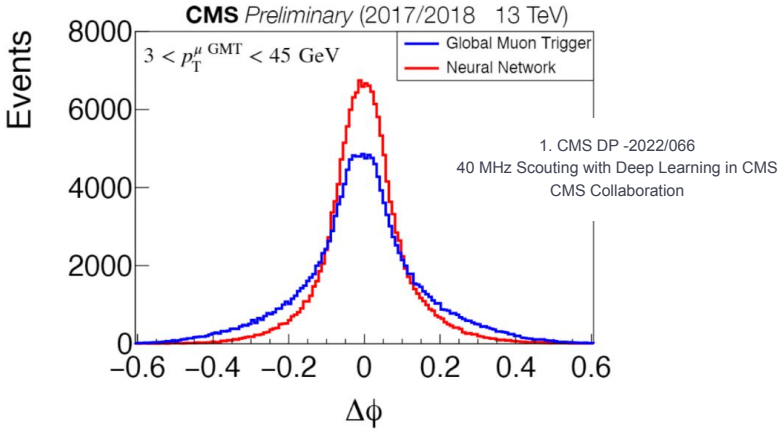
Openlab Summer Student Programme 2023

Data acquisition at CMS

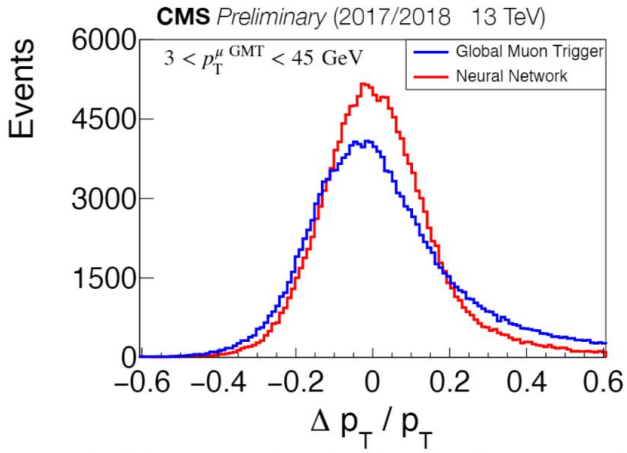


Advantage of Machine Learning

- Physics based algorithms are based on assumption and often lead to biases.
- Physics based algorithms might not be suitable for Beyond Standard Model search.
- Often ML based algorithms are lighter than their physics based counterparts.
- ML based algorithms outperform physics based ones buy a large margin of error.



Particle angular position perpendicular to beam



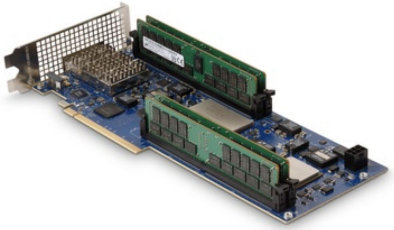
Particle momentum in direction transverse to beam

Aim of the project

- Design the smallest possible neural network with the highest possible resolution for Global Muon Trigger (μGMT) muon primitives recalibration.

- Design the smallest possible neural network with the highest possible accuracy for μGMT muon pair fake/real classification.

- Quantize and Deploy both the models to Micron SB852 and Xilinx VCU128 FPGA boards without any loss of performance.



Methodology

We are looking for the best possible model with the smallest possible memory and computational footprint. Hence we choose the following setup:-

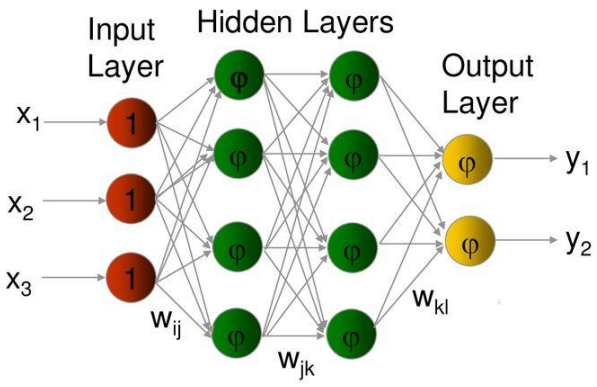
1. Find the best possible model/ensemble with the highest scores on the evaluation metrics disregarding the computational footprint. This will be called the Oracle/Teacher from now on.

2. Distill the oracle to the smallest model. We keep reducing the size of the student until we either lose a lot of performance or the footprint is so low that it doesn’t matter anymore.

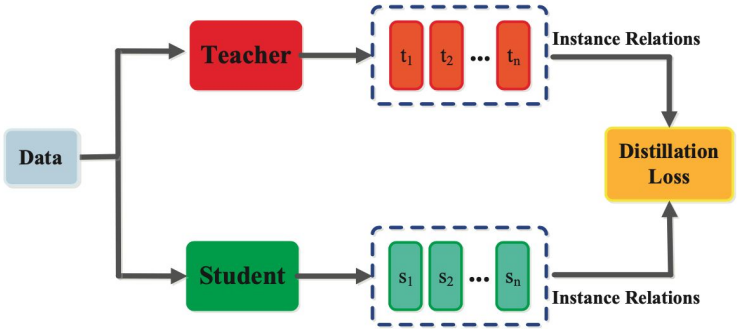
Model Architectures

- ➔ Teacher (Over parameterized)
  - ◆ 256 Neurons 4 hidden layers Multi-layer perceptron
  - ◆ Total params: 204 K
- ➔ Student(Distilled Under parameterized)
  - ◆ 8 Neuron 4 hidden layers multi-layer perceptron
  - ◆ Total params: 419
- ➔ Baseline
  - ◆ 16 Neuron 4 hidden layers multi-layer perceptron
  - ◆ Total params: 1219

	Factor
★ Theoretical compression(Parameters):	486
★ Compression from baseline(Parameters):	2.9
★ Theoretical compression(Flops):	1024
★ Compression from baseline(Flops):	4



Multilayer Perceptron

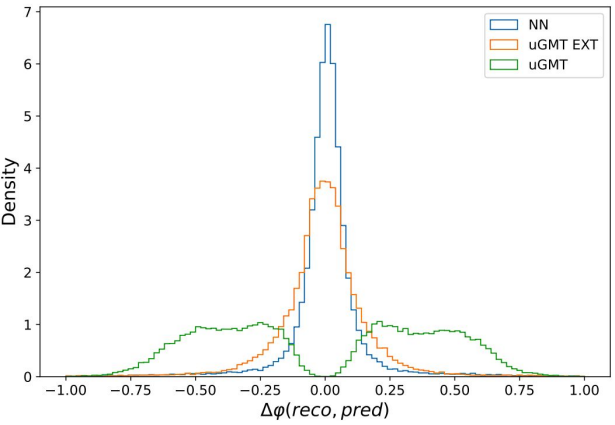


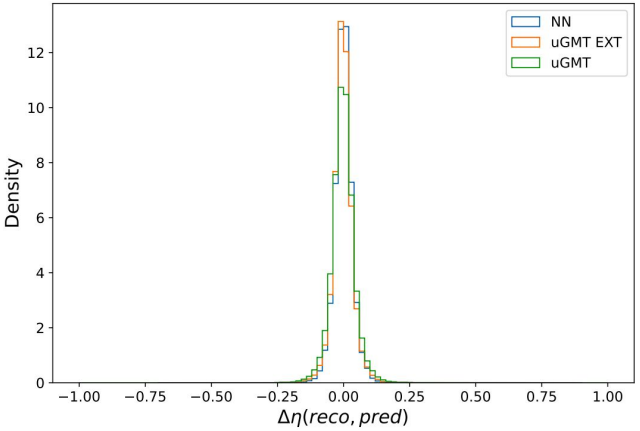
Knowledge Distillation

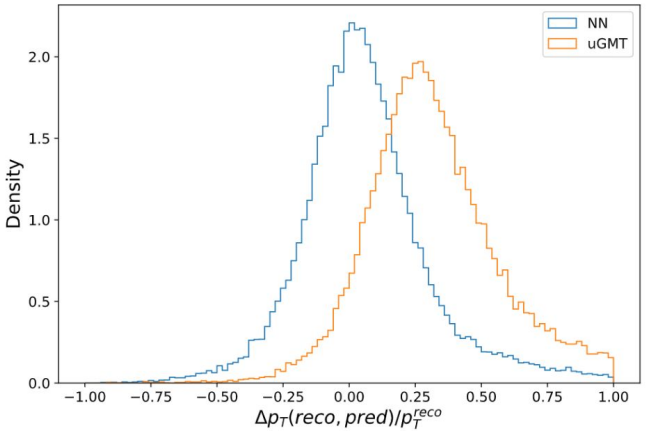
Results

µGMT muons recalibration on FPGA devices for L1 Trigger at CMS

Resolution plots for recalibration



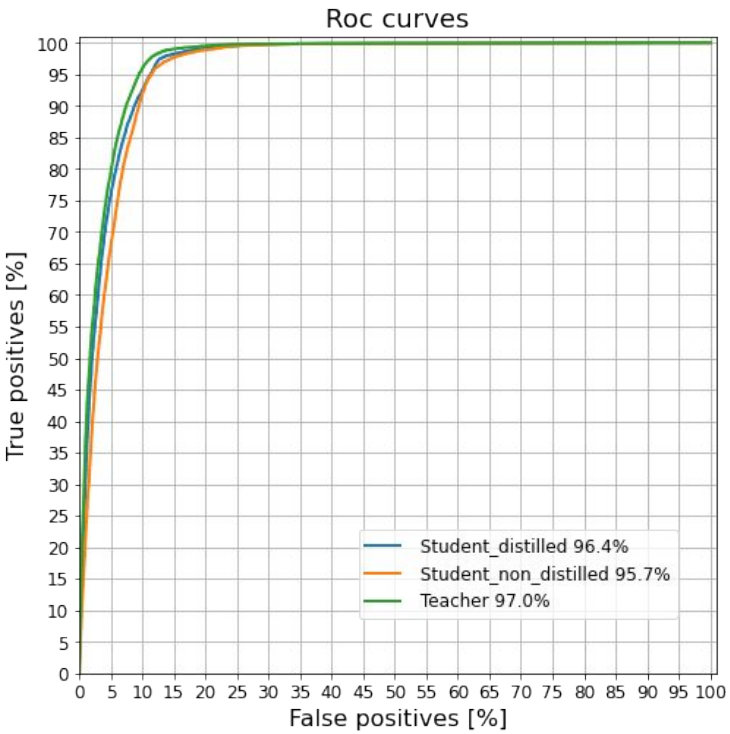




FWHM	Teacher	Student	Baseline
φ	0.120	0.124	0.140
η	0.061	0.063	0.0662
p_T	0.413	0.417	0.433

Muon pair classification on FPGA devices for L1 Trigger at CMS

ROC Curve



Model	Area Under ROC curve
Student (Not distilled)	95.7%
Teacher	97.0%
Student Distilled	96.4%

Implementation using HLS4ML and Resource Usage

DSP Reuse Factor: Number of times a multiplier is used to do a computation = 4

VU9P FPGA	DSP	Flip Flops	Look Up Table	BRAMS
Available	9024	2.6 M	1.3 M	2160
Used	72 (0.79%)	5677 (0.21%)	11.3 K (0.87%)	0

The model architecture for both the tasks are exactly same.

QKeras

Vivado™ HLS

hls4ml

AMD XILINX

https://fastmachinelearning.org/hls4ml/

Summary

- Exploring different NN models to predict correction terms to L1 muons kinematic quantities and classification of true/fake muon pairs.
- Train the oracle model that performs the best for both our tasks with Quantization Aware Training (QAT).
- Distilling the oracle to the student network with Quantized weights.
- Testing the student on both the tasks and validating with simulation and hardware setup.
- Integrating both Neural Networks to the μGMT L1 scouting firmware