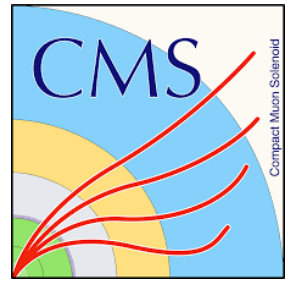
CERN openlab



Boosting Online Recalibration of Physics Objects for the 40 MHz Scouting System at CMS

Leyla Naz Candoğan

Supervisors:

Emilio Meschi, Rocco Ardino

Introduction

Data Acquisition and Triggering in CMS

- Bunch crossings at 40MHz frequency
- ~500 TB/s of raw data
- Two-level trigger system: Level-1 (L1) and High-Level (HLT) triggers.

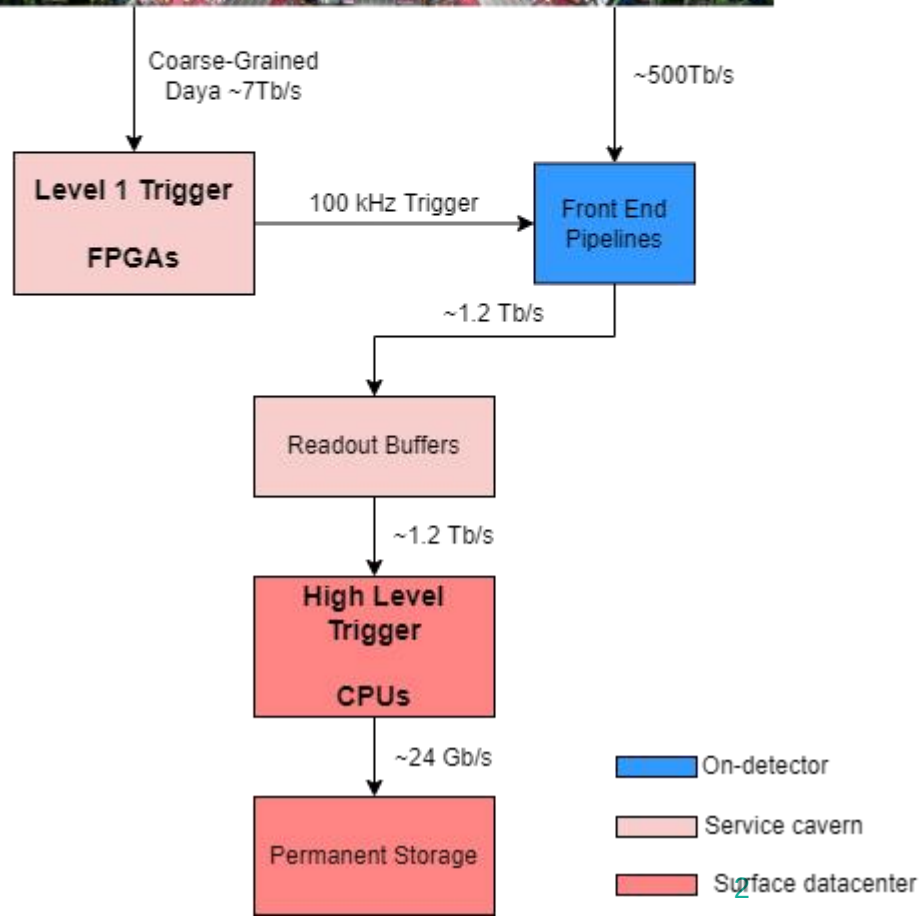
L1: custom hardware field programmable gate array (FPGA) boards

- Recorded event rate reduced to ~100 kHz

HLT: farm of processors running a version of the full event reconstruction software

- Data rate to around 1 kHz

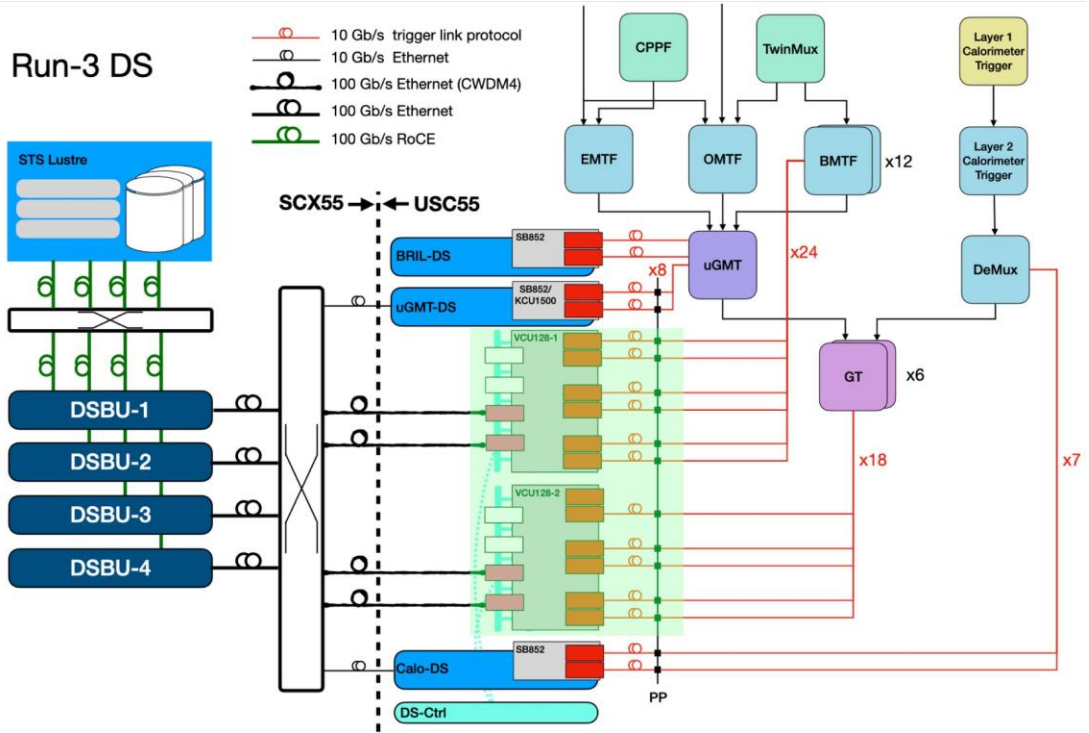
- Bias induced by trigger system limits the beyond standard model research

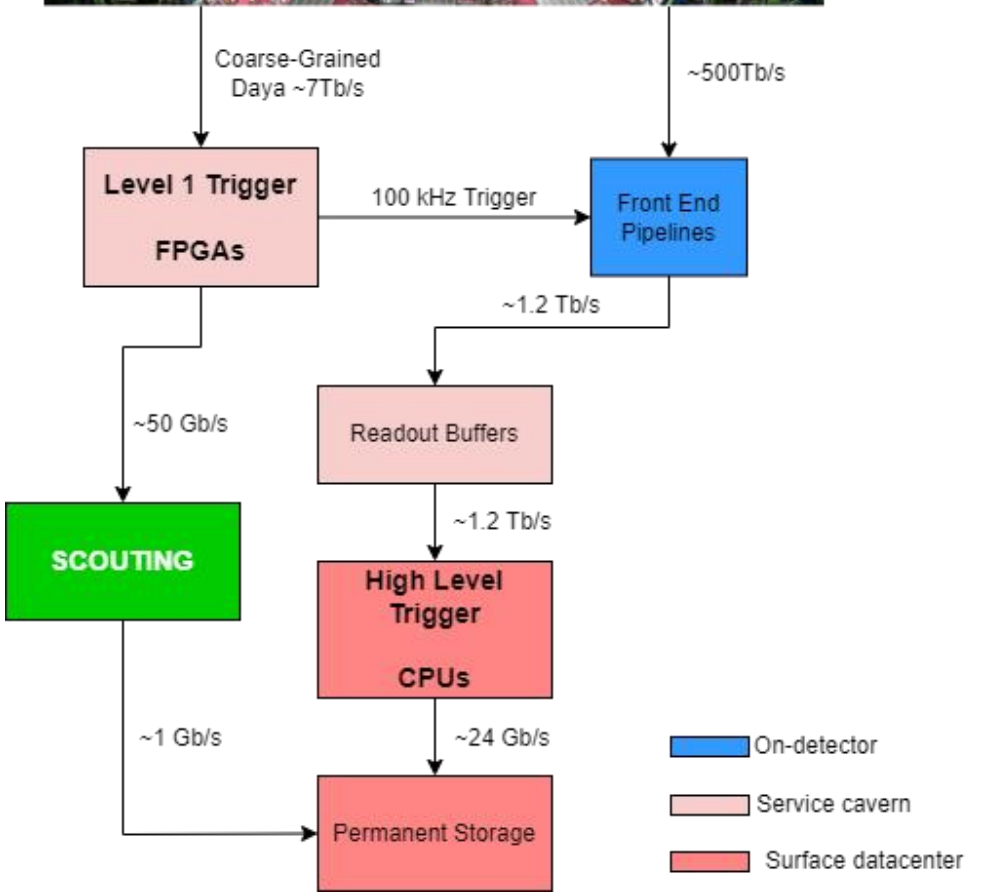


Introduction

L1 Scouting System

- Performs semi-real time analysis by collecting and storing L1 reconstructed primitives at full bunch crossing rate (40 MHz)
- Focus of the project: L1 muons from Global Muon Trigger (μGMT) board

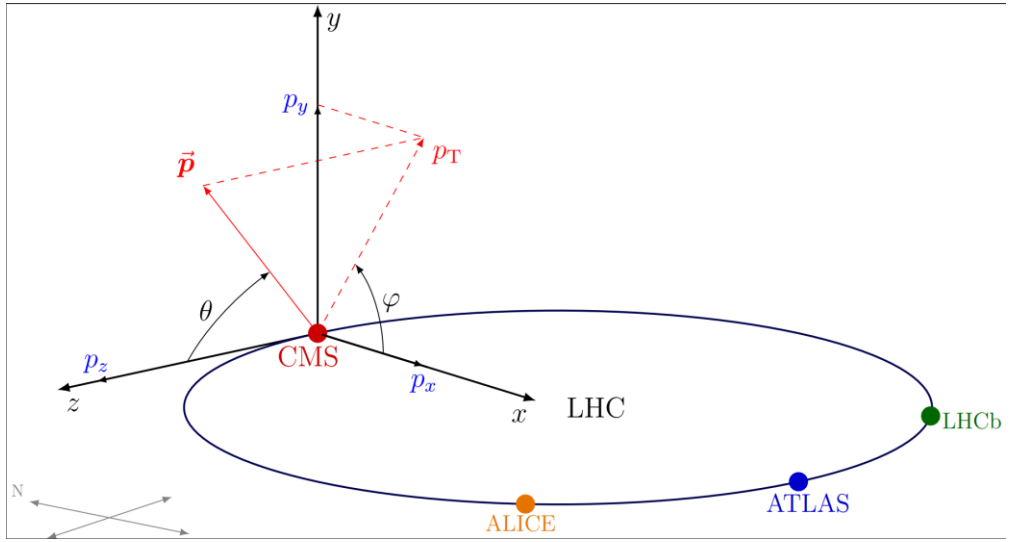


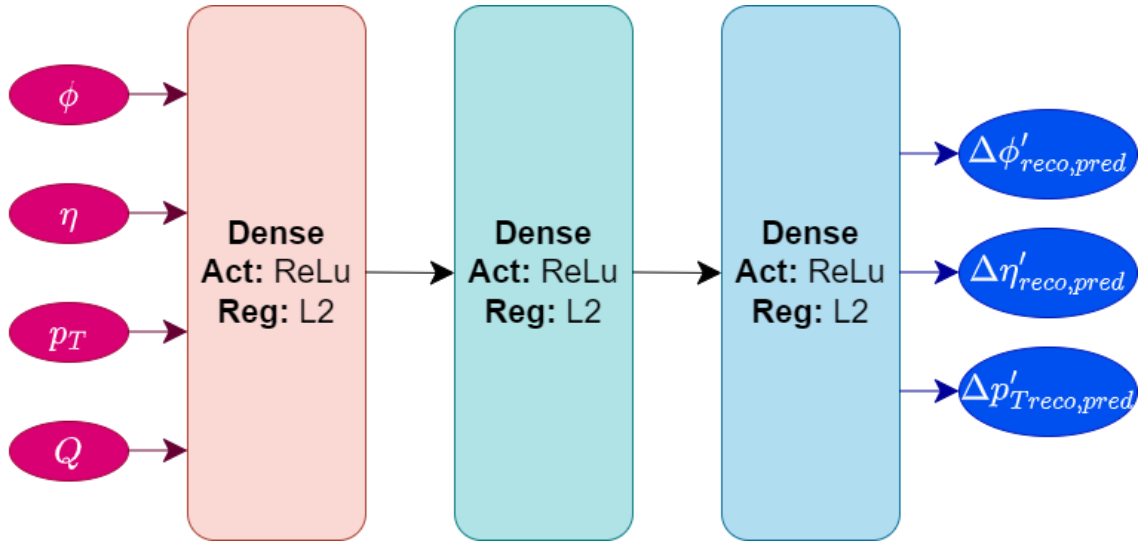


Introduction

The Goal

- Integrating Neural Network (NN) on Xilinx VCU128 board for uGMT scouting
- NN for online recalibration of phi, eta and pT values of uGMT muons
- **Input:** Phi, eta, pT, charge values of uGMT objects
- **Output:** Correction terms to input parameters





Introduction

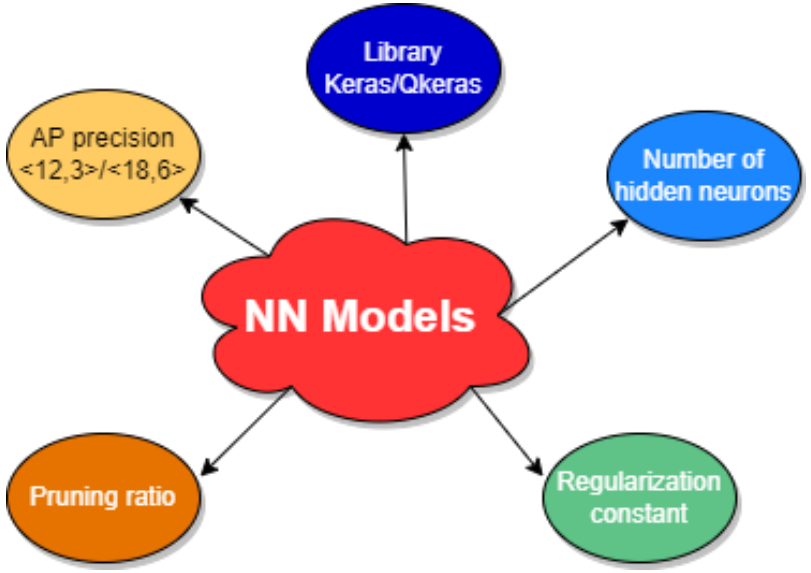
Tools: hls4ml

- Python library to generate HLS package of machine learning models
- Compatible with **Keras/Tensorflow/QKeras**, PyTorch, Onnx
- Supported Neural network architectures:
  - **Fully Connected NNs (multi-layer perceptron)**
  - Convolutional NNs (1D/2D)
  - Recurrent NN/LSTM, in prototyping
- Key configuration options:
  - **Configurable fixed point precision (relying on ap_fixed<X,Y>)**
  - **Target clock period for application**
  - DSP reuse factor
- https://fastmachinelearning.org/hls4ml/

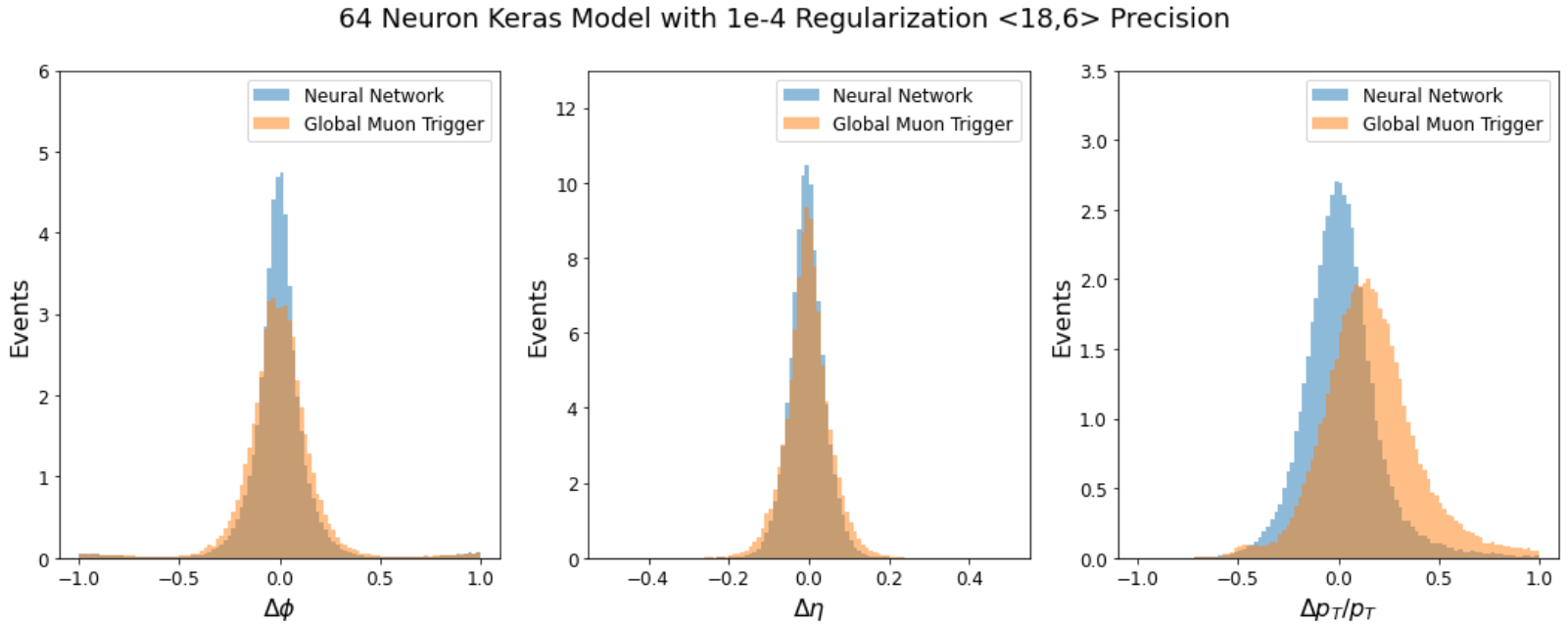
Software Results

Model Configurations and Performances

- Comparison of different models



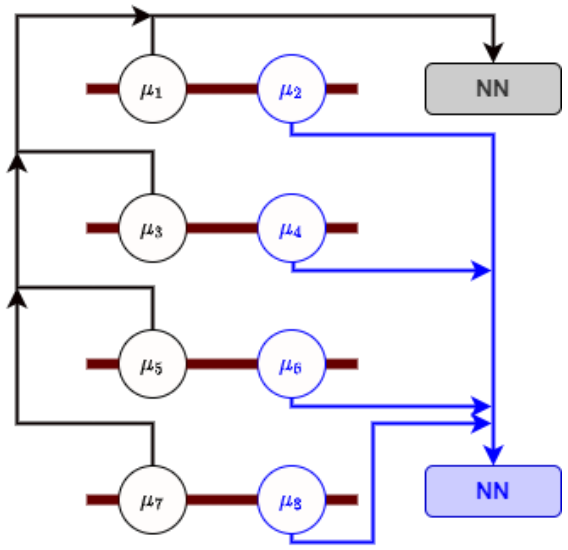
- Performance metrics:
  - Logarithmic hyperbolic cosine loss
  - Offline reco quantities vs L1 quantities
  - FWHM values

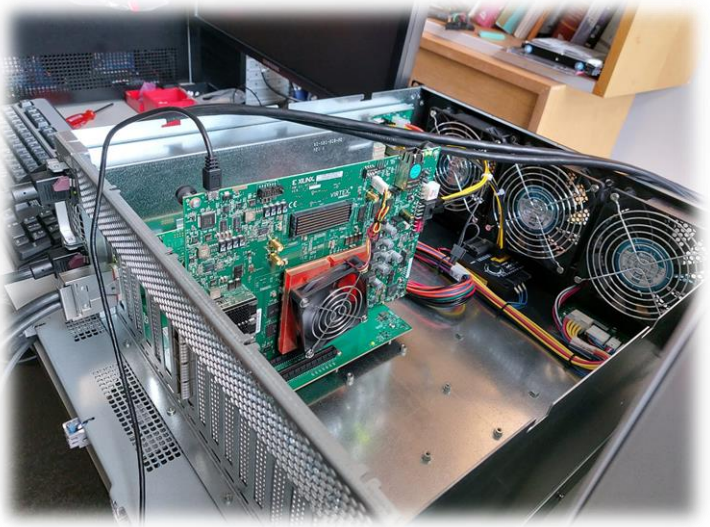


	Mean	Sigma	FWHM	Data in Core (%)	Mean	Sigma	FWHM	Data in Core (%)	Mean	Sigma	FWHM	Data in Core (%)
Model 1	0.0002	0.092	0.217	58.36	-0.004	0.039	0.092	66.54	0.0229	0.157	0.369	70.16

Hardware Implementation

Available Resources	VCU128
DSP	9024
FF	2.607.360
LUT	1.303.680

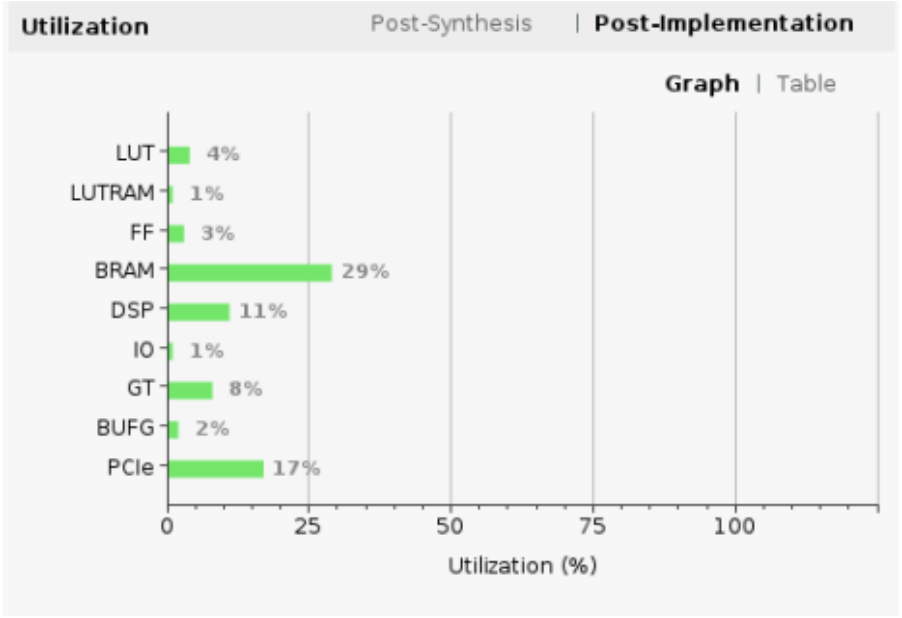


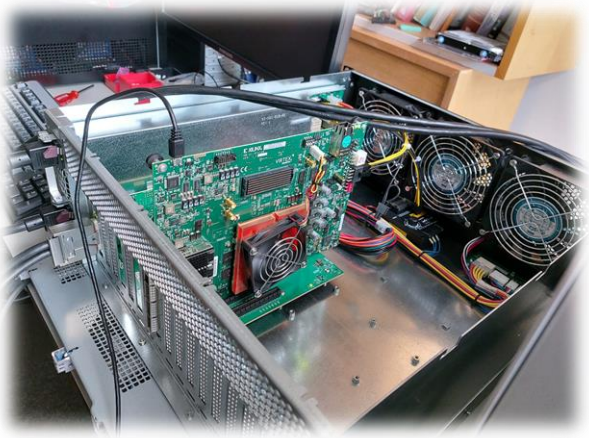


- Try to reduce the number of NN:
  - To optimize resource utilization,
  - To eliminate timing violations,
  - To ease routing

Hardware Implementation

- L1 uGMT muons generated on hardware for testing and development
- 2 Neural Networks added to the firmware
  - QKeras Model
  - 3 Layer
  - 32 Neurons
  - Pruning: 0.5
  - Precision: <18, 6>
  - 4 L1 muons per NN
- Project steps:
  - Vivado simulation for proof of concept
  - Hardware implementation
  - Validation in test setup
- Results reproduce software predictions accurately





Summary

- **Goal:** Integrating NN in uGMT scouting firmware for online recalibration of uGMT muons

1. Exploring different NN models to predict correction terms to L1 muons kinematic quantities
2. Model optimization
3. Integrating NN to the uGMT L1 scouting firmware
4. Validation with simulation and hardware test setup

QUESTIONS?