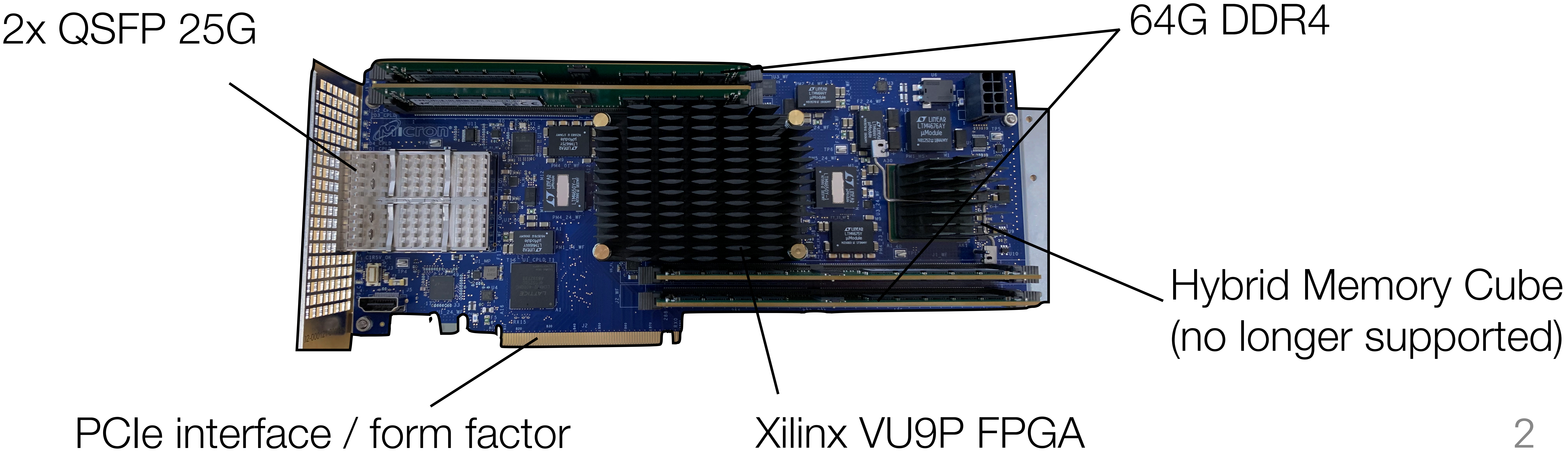
FPGA-based Machine Learning Inference for CMS with the Micron Deep Learning Accelerator

Thomas James, Emilio Meschi, Ema Puljak

9th Mar 2021

Technology: Micron Deep Learning Accelerator

- SB852 PCIe board, Xilinx VU9P, 2x QSFP
  - 64GB DDR4, PCIe x16 Gen3 to host
- Firmware: Proprietary Inference Engine, scalable and programmable solution to deep learning inference offers ~Tera MAC/s



Technology: Micron Deep Learning Accelerator

- Board configured with Micron Deep Learning Compiler

Keras

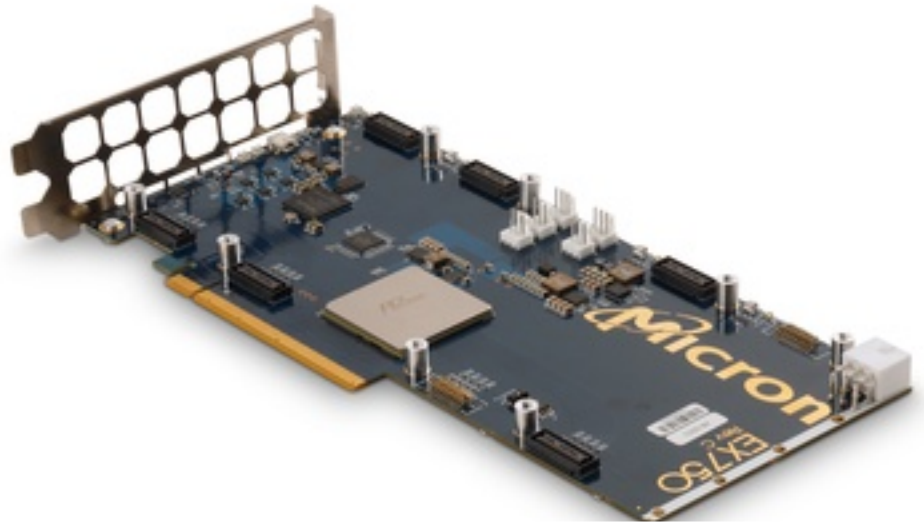
TensorFlow

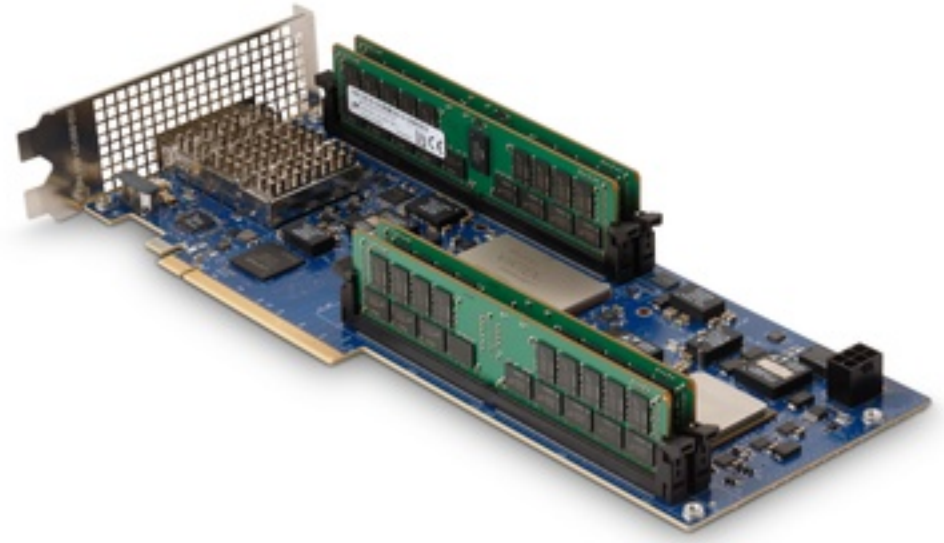
Caffe2

PyTorch

ONNX

Micron Deep Learning Compiler*

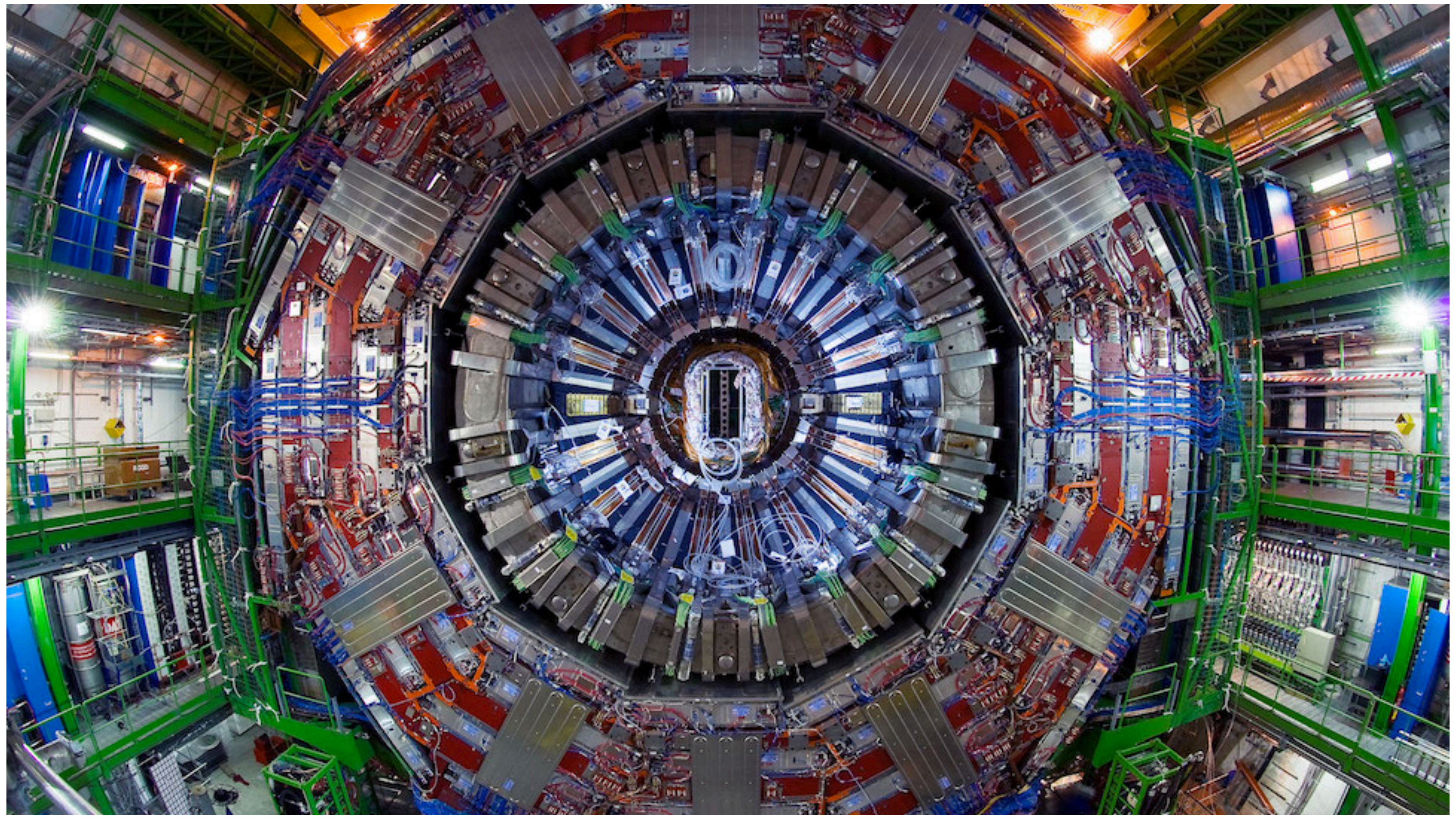




*User friendly API; reports diagnostics of interest e.g latency, precision, bandwidth

No need to write VHDL to run most DL models on FPGA

Application: The CMS Detector at the LHC

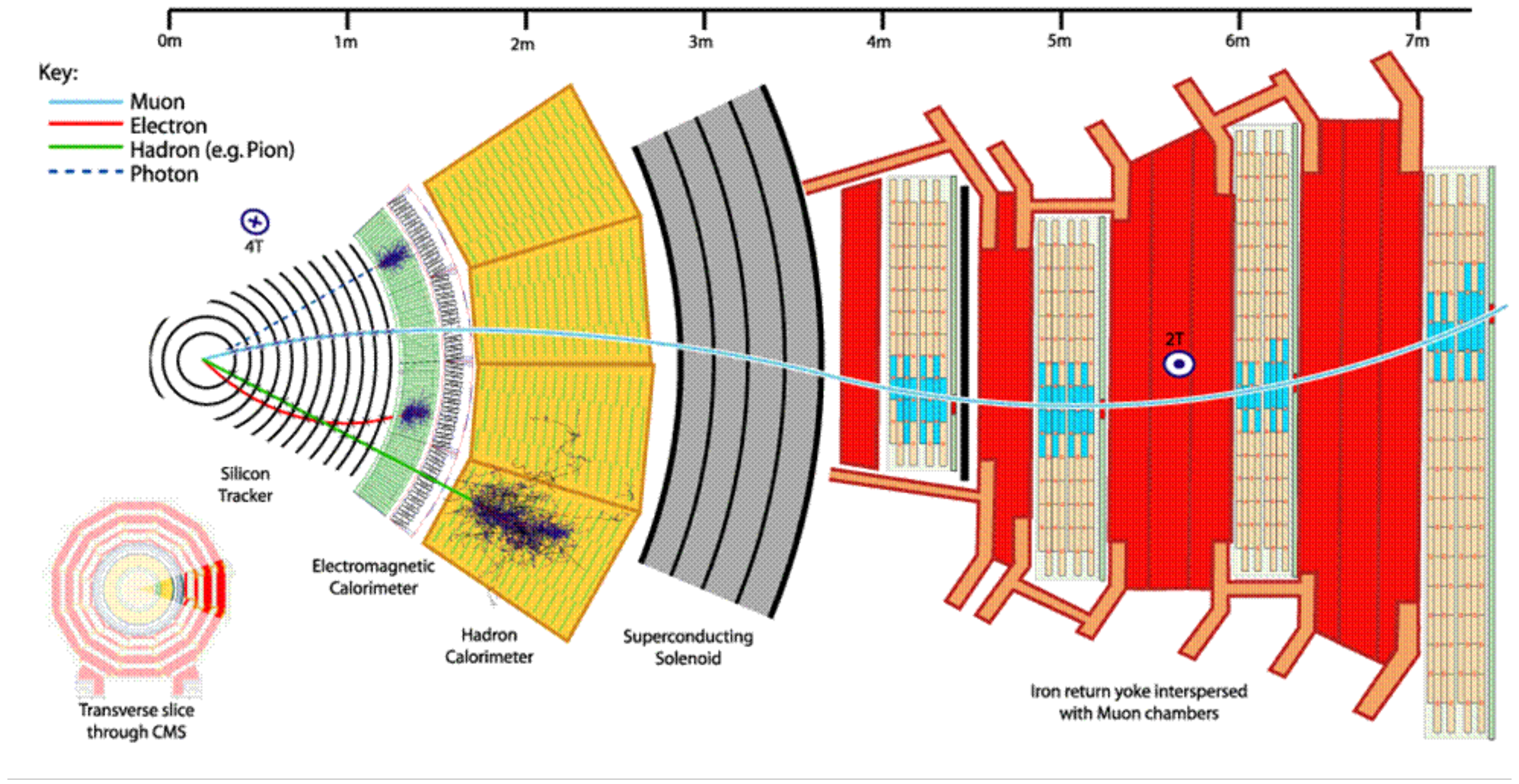


- 2.4 billion collisions / second
  - In CMS ~ 100M sensors
  - Produce ~ 1.5 MB @ 40 MHz, ~500 Tb/s
  - Impossible to read out (or store) all data
  - Need fast 'trigger' to select *interesting* collisions for analysis

- Two layered:
  - Level 1: Fixed latency of 3.2 microseconds -> ASICs and FPGAs required
  - High Level Trigger: Flexible latency ~100 ms compute / event -> CPUs/GPUs

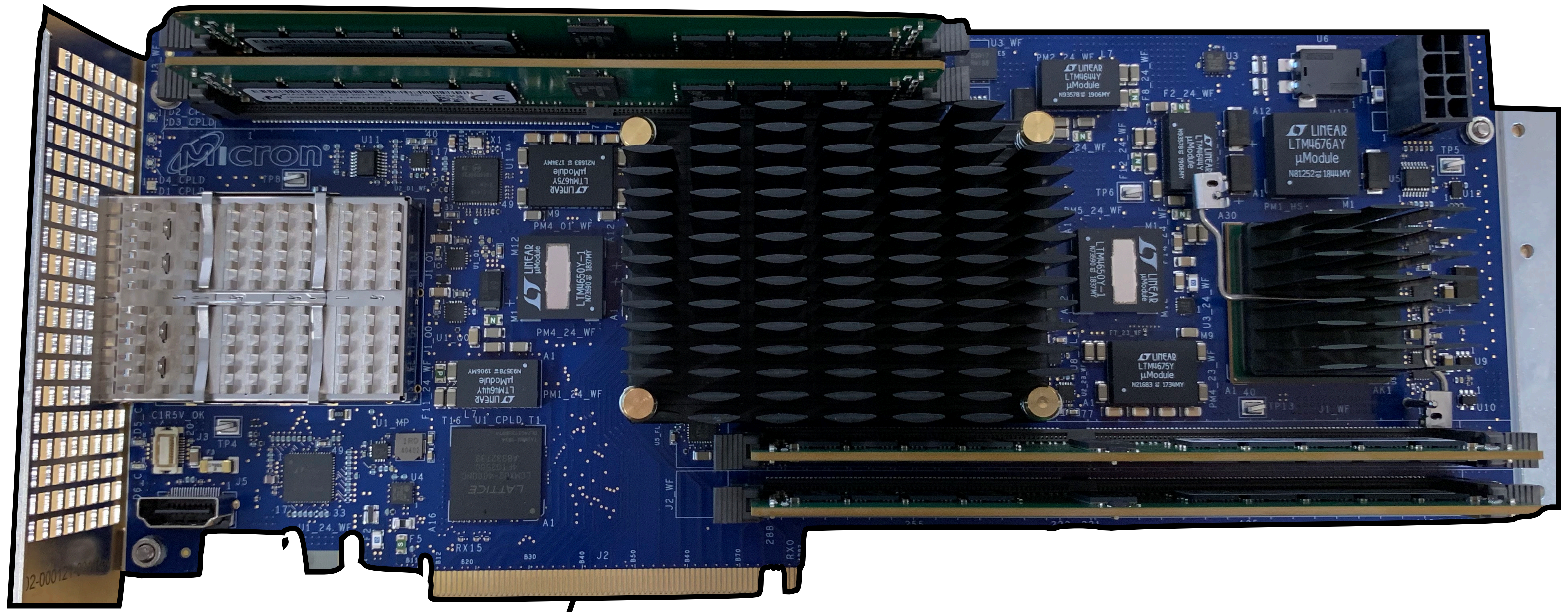
Extension to CMS TDAQ: 40 MHz Scouting

- **Acquire L1 trigger data at full bunch crossing rate**
  - subset of detector information, limited resolution
- **Allows for analysis of certain topologies at full rate**
  - » semi real-time analysis and/or
  - » storing of tiny event record



- Demonstrated for first time at end of 2018
- Current plans to scout objects from the Global Muon Trigger, Barrel Muon Trigger & Calorimeter Trigger at LHC Run 3

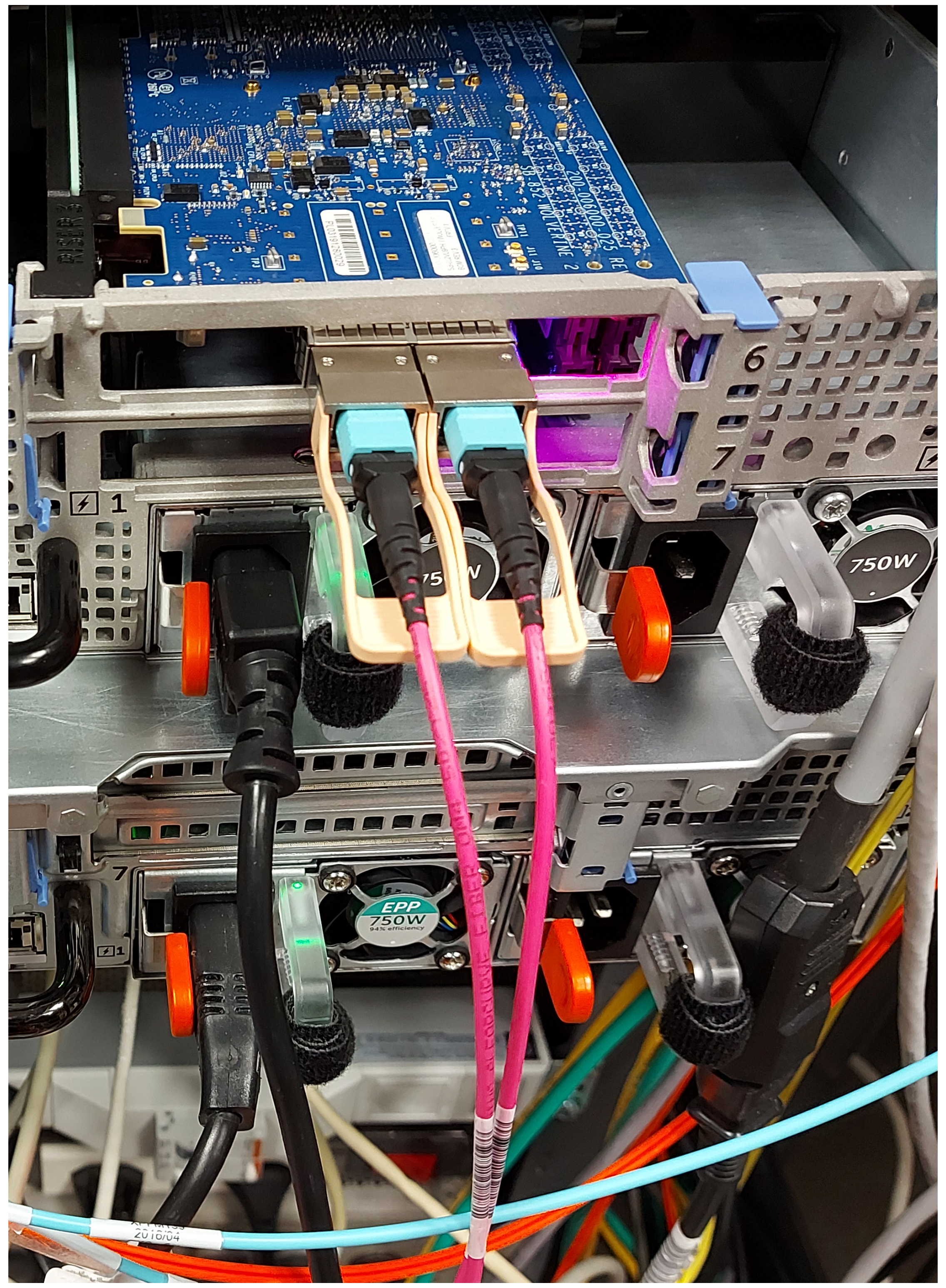
CMS 40 MHz Scouting with SB-852

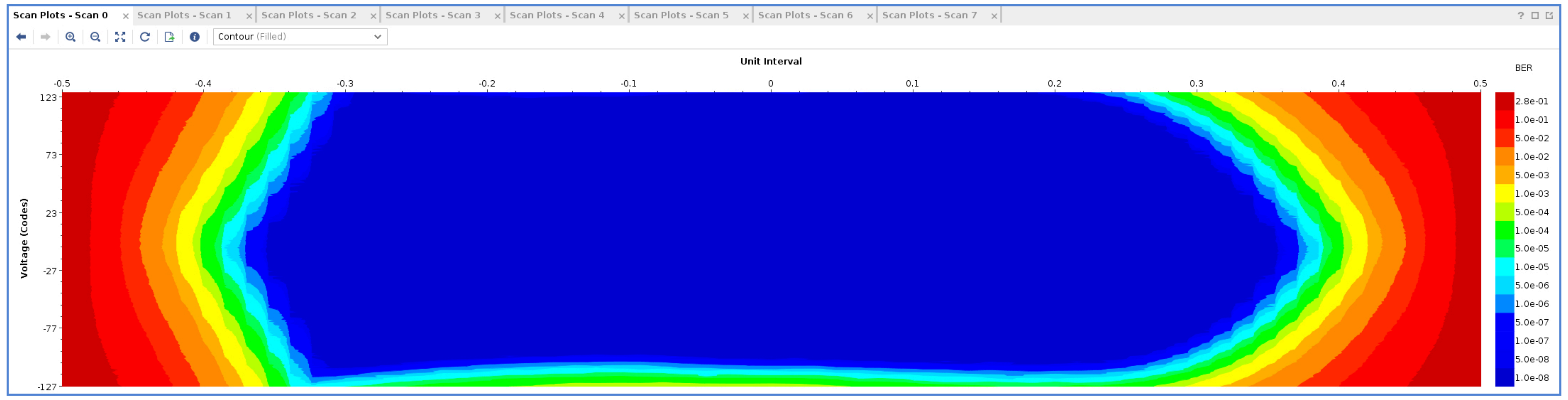


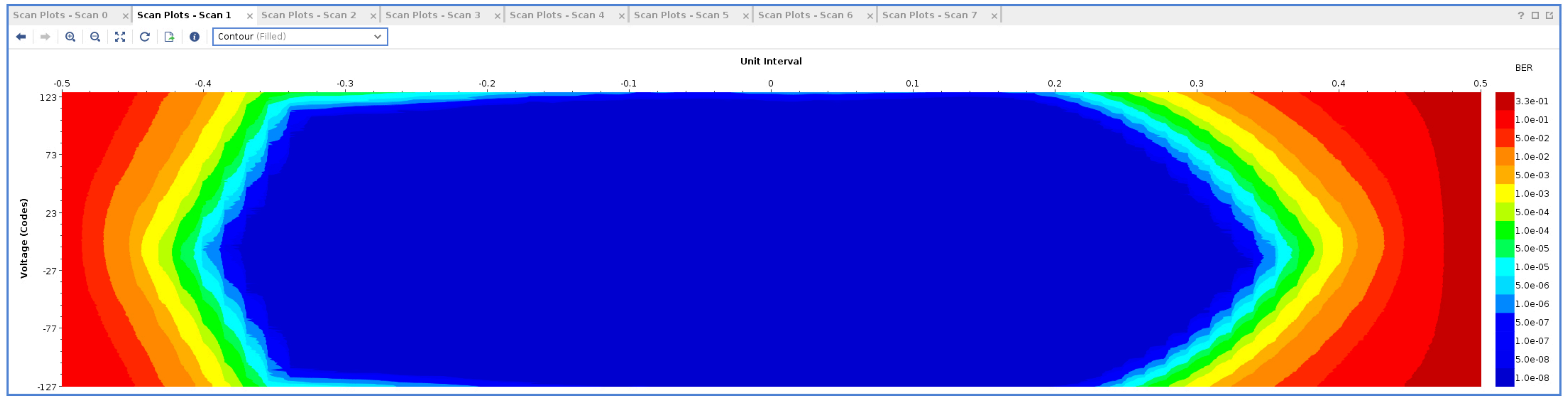
- Micron SB-852 for optical input -> DMA to PC
- Perform machine inference with Micron DLA after firmware ZS

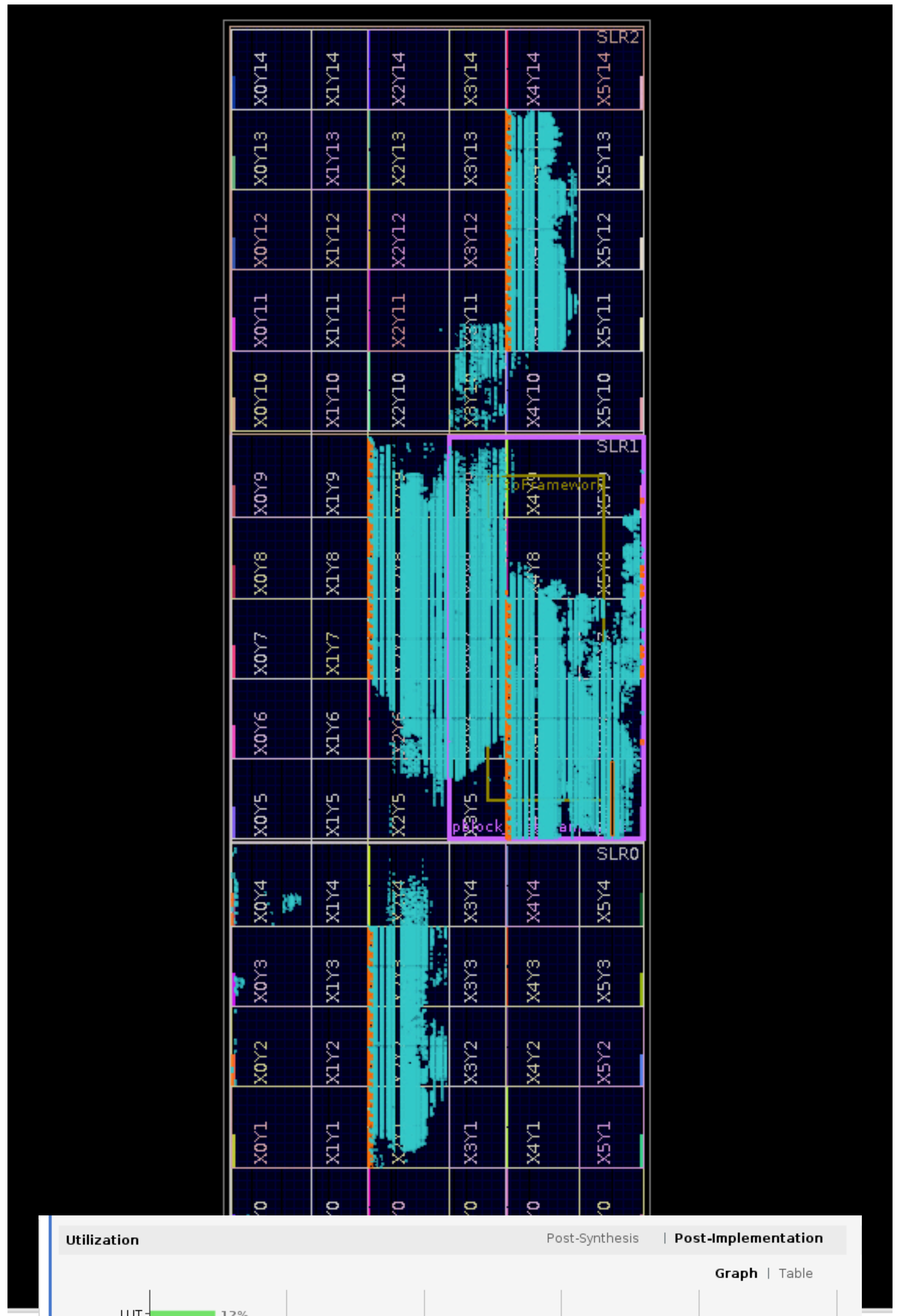
Extension to CMS TDAQ: 40 MHz Scouting

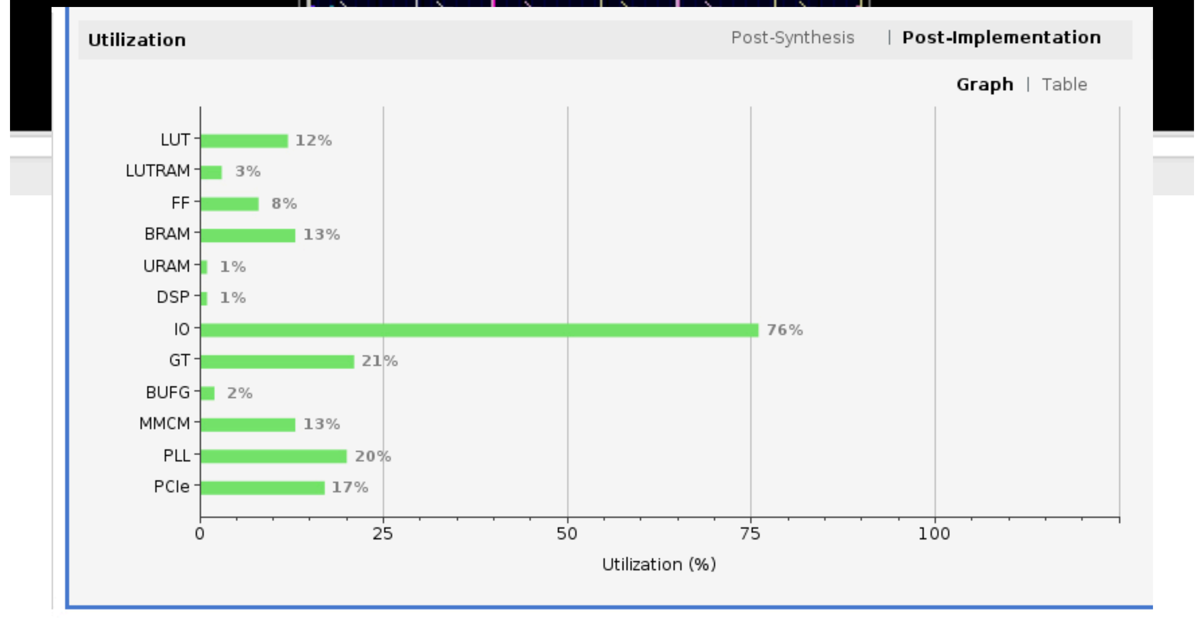
- **Firmware ported and developed for SB852**
- Optical link interface implemented and tested





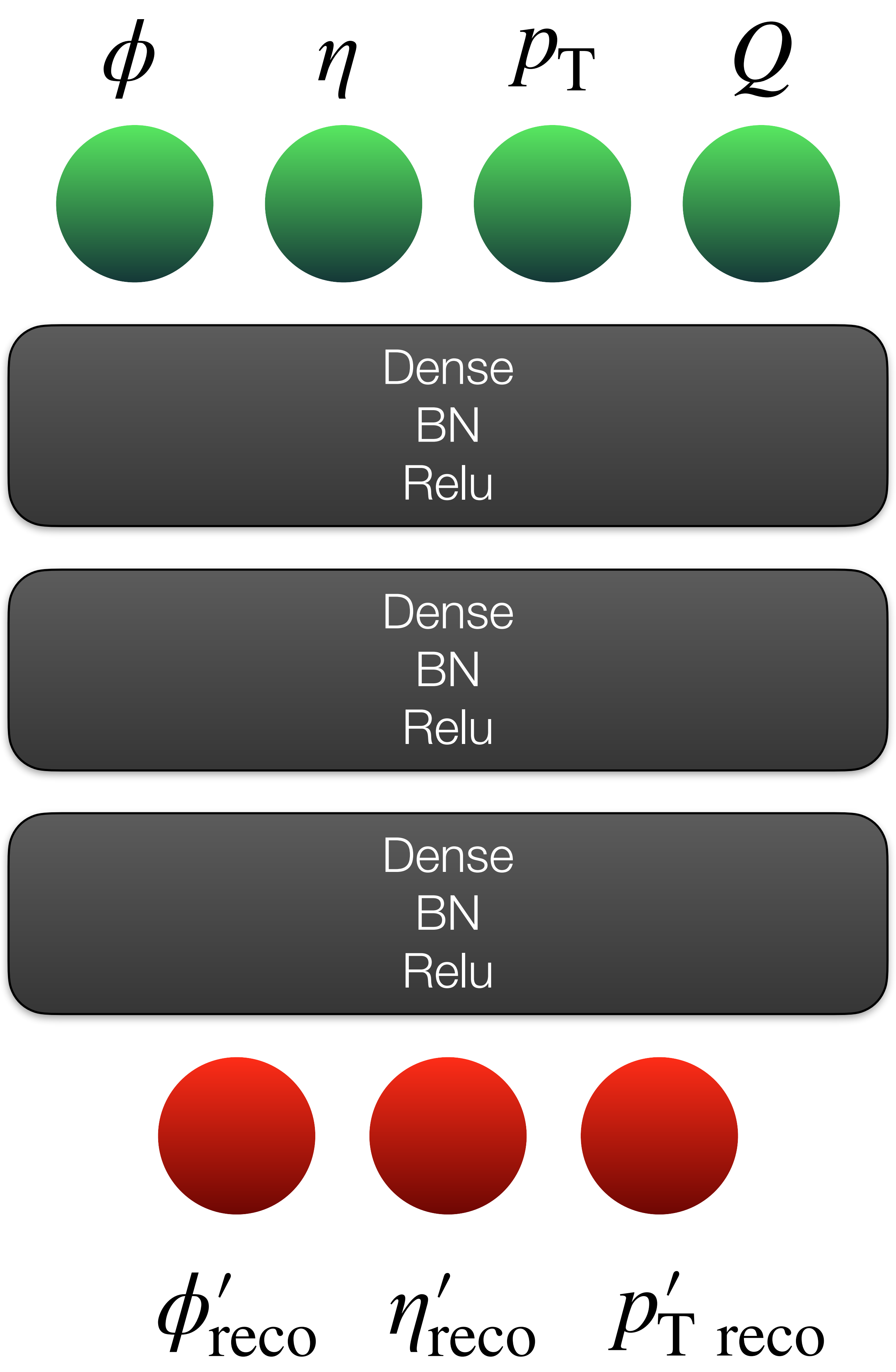






Why ML for scouting?

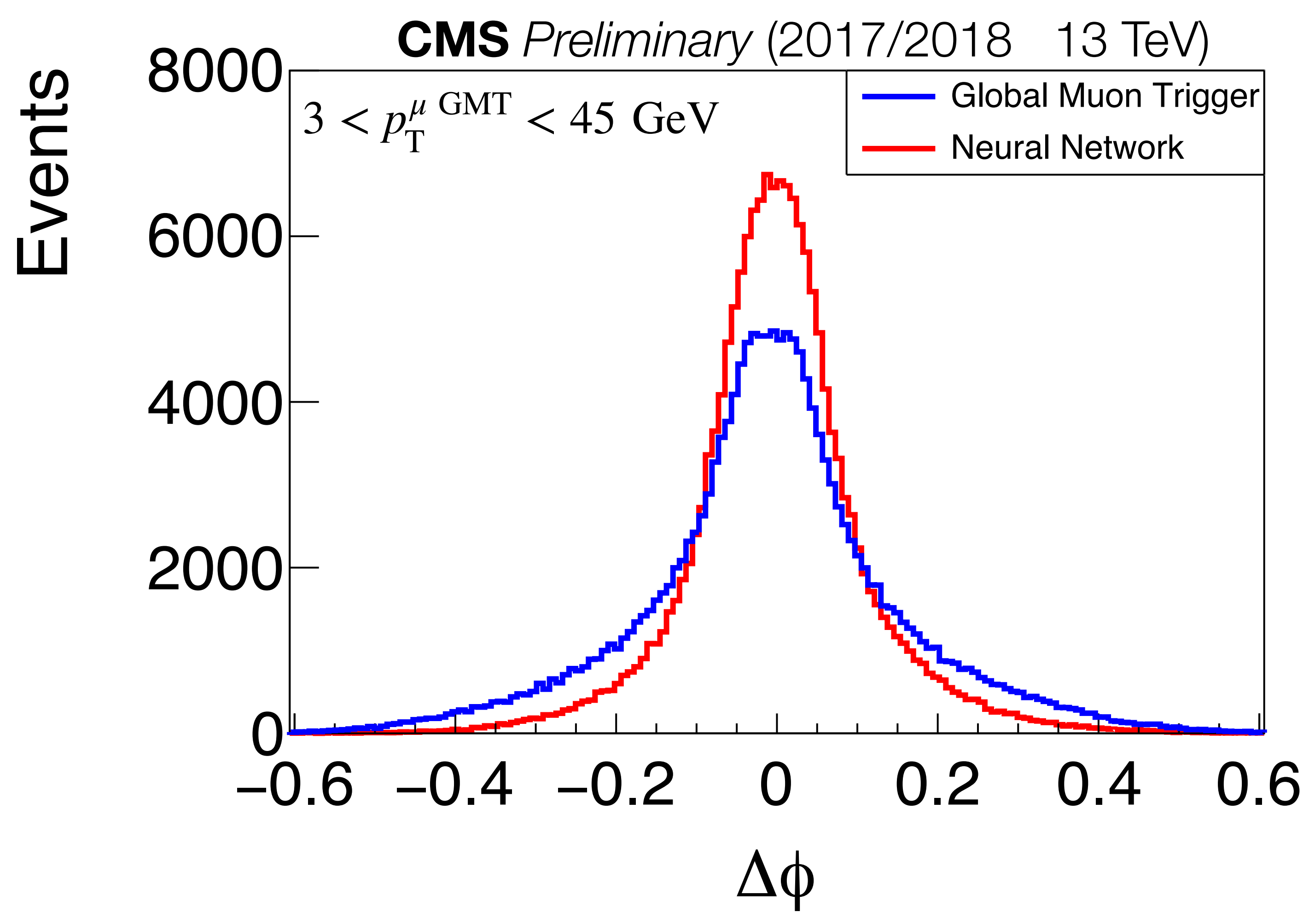
- Trigger objects calibrated for best efficiency at threshold
  - Not for best physics analysis
- But we have full offline reco & trigger objects for Zero Bias and Triggered events
- Can we use the offline objects as target to correct the parameters of the trigger level objects?
- Use of classical neural networks to 'correct' L1 information e.g muon helix parameters
- **Inputs** - L1 objects e.g GMT muons: **Target** - Offline fully reconstructed objects

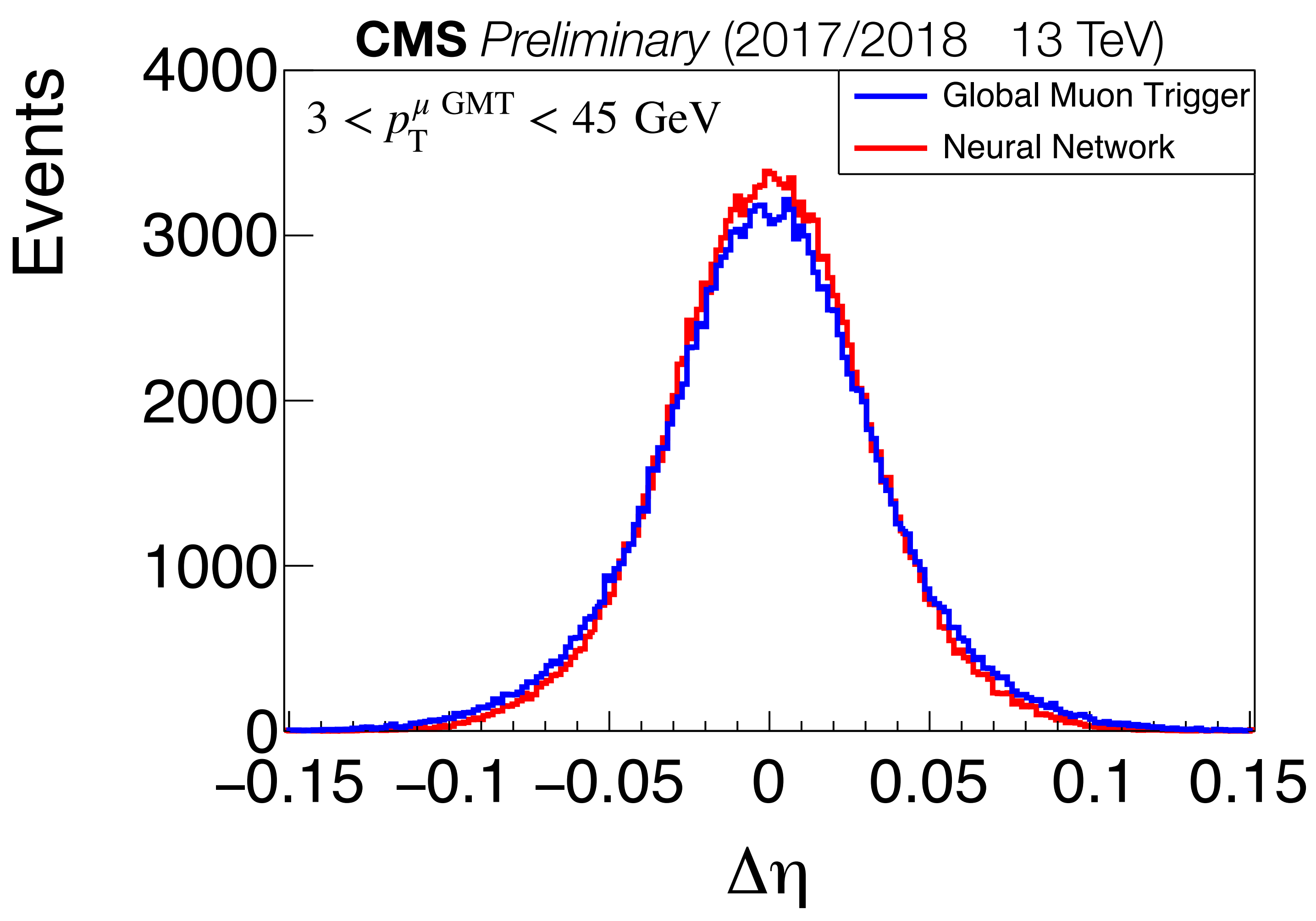


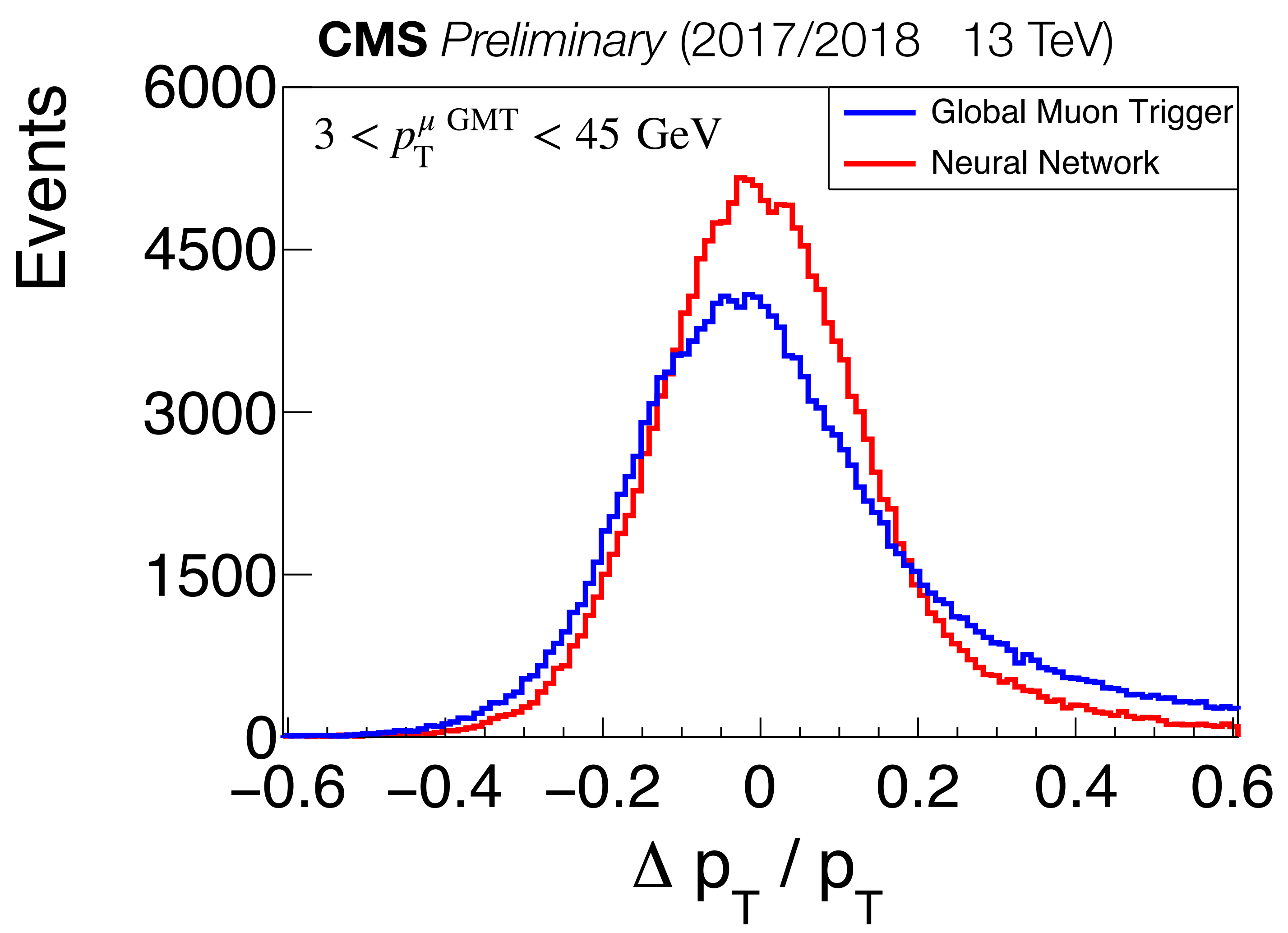
Trained with *Zero-bias* dataset

GMT with Neural Network

- Able to achieve ~2x improvement in track parameter resolution for some interesting areas of phase-space





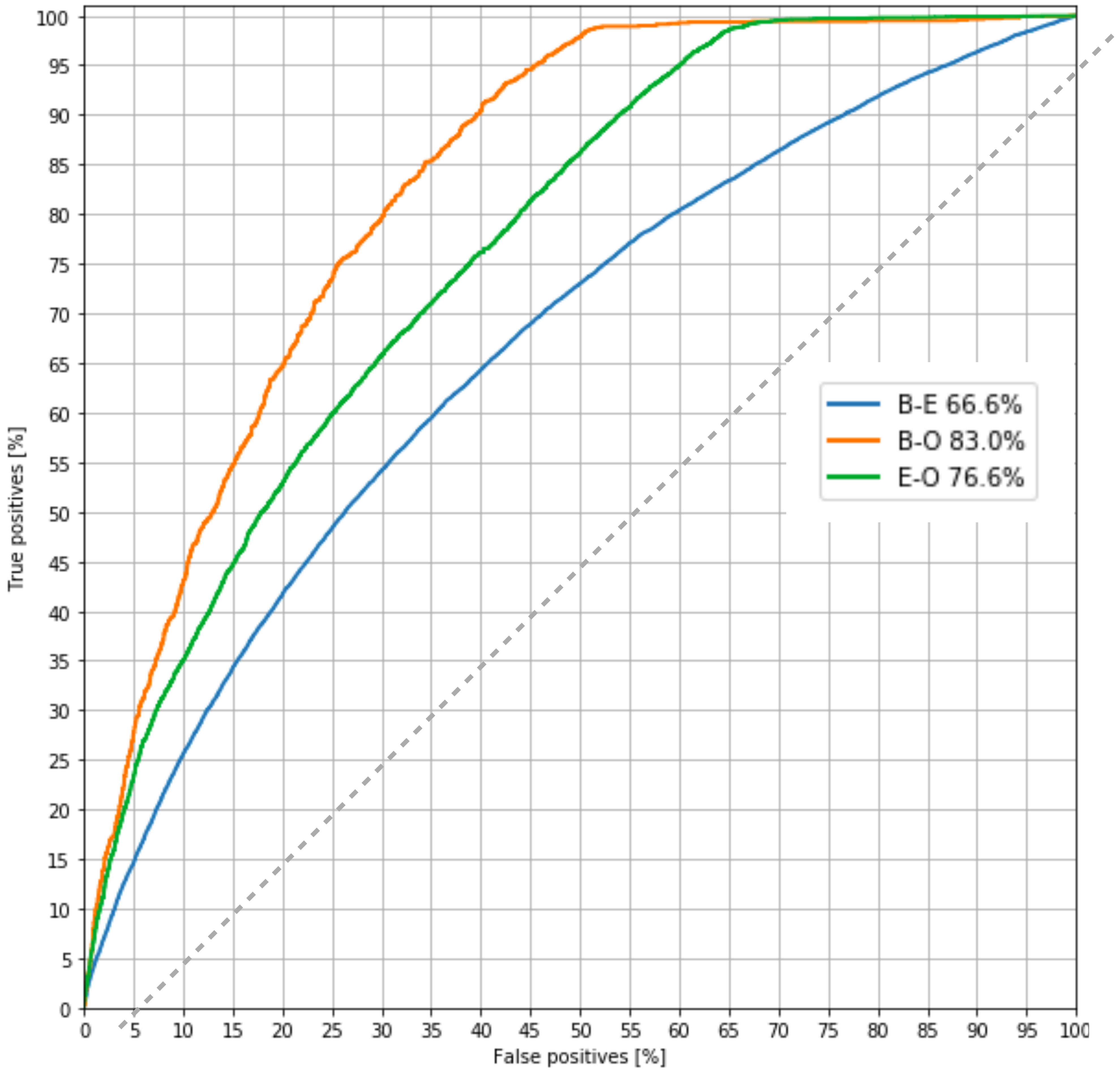


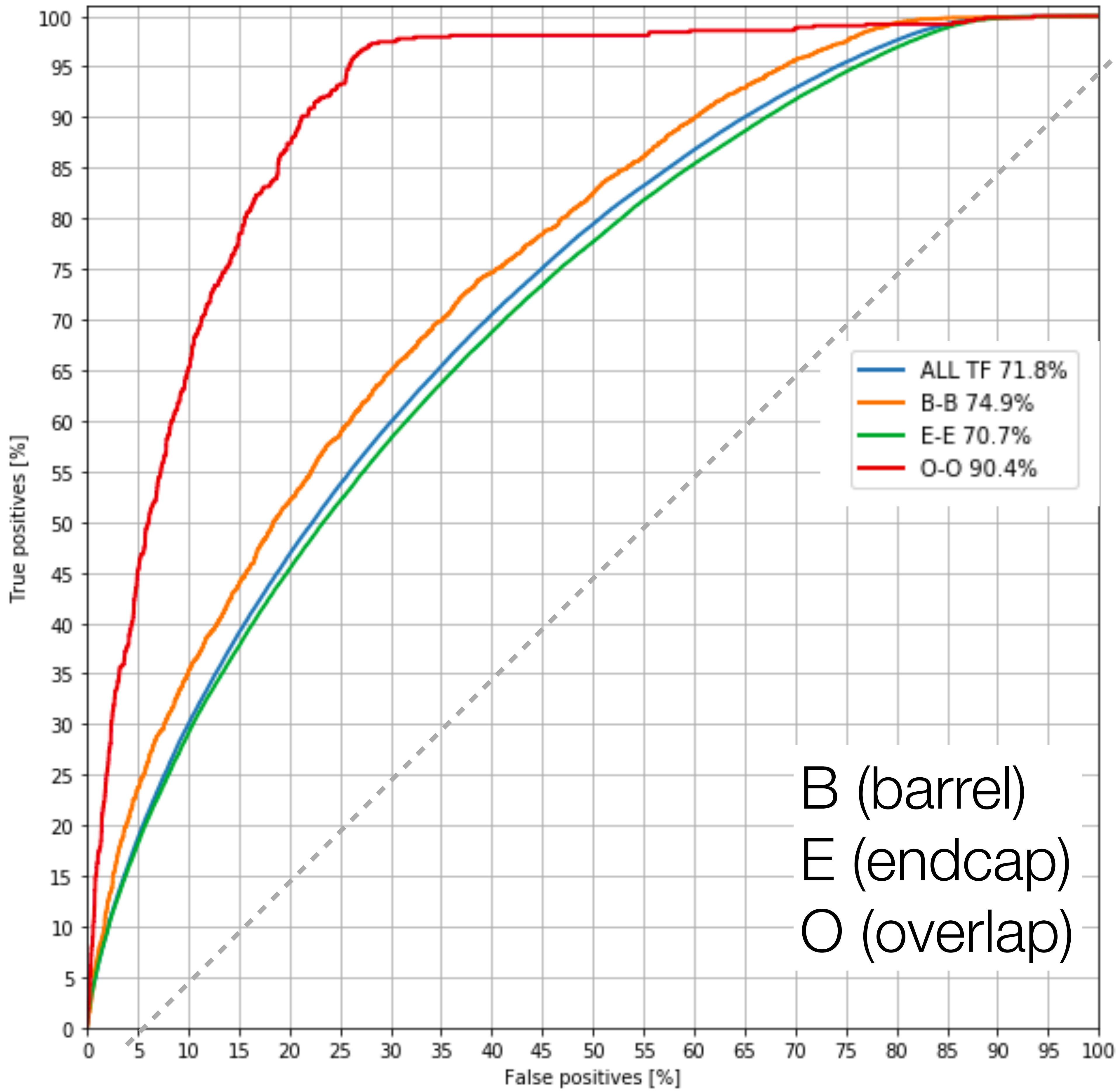
- Applied a neural network to the L1 muons to improve their accuracy for real-time analysis in the L1 Scouting system. Produced with Zero Bias data.

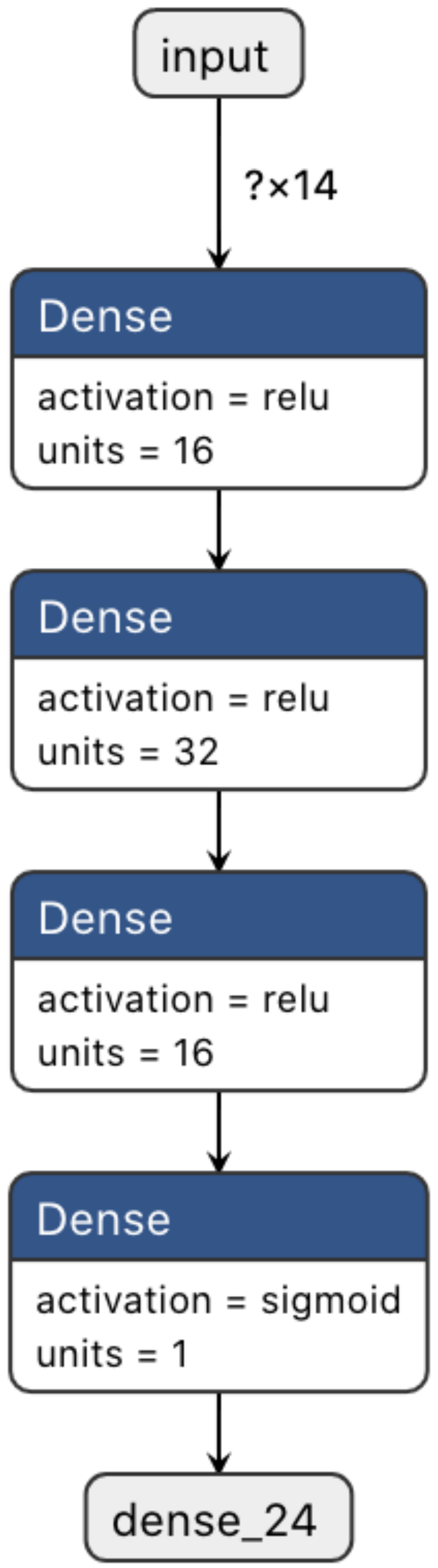
- $\Delta\eta$, $\Delta\varphi$, Δp_T is the difference between the prediction (or GMT) values, and the offline muon tracks for matched muons ($\Delta R<0.1$ at 2nd muon station).

Fake muon pair classifier

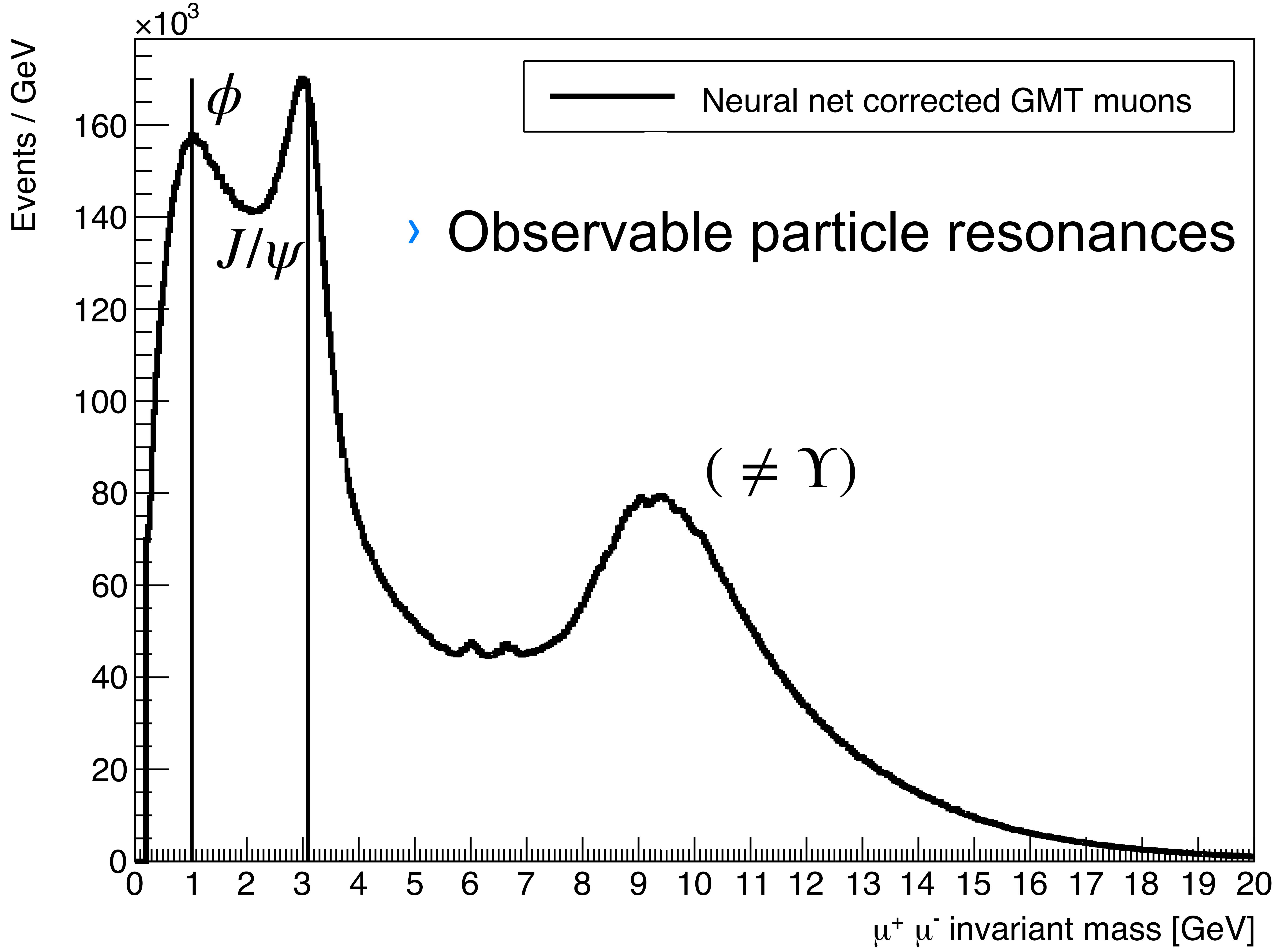
- Train a DNN with ZB data to predict fake muon pairs
- Can use to improve purity of di-muon sample







GMT Scouting data from 2018 re-calibrated with NN



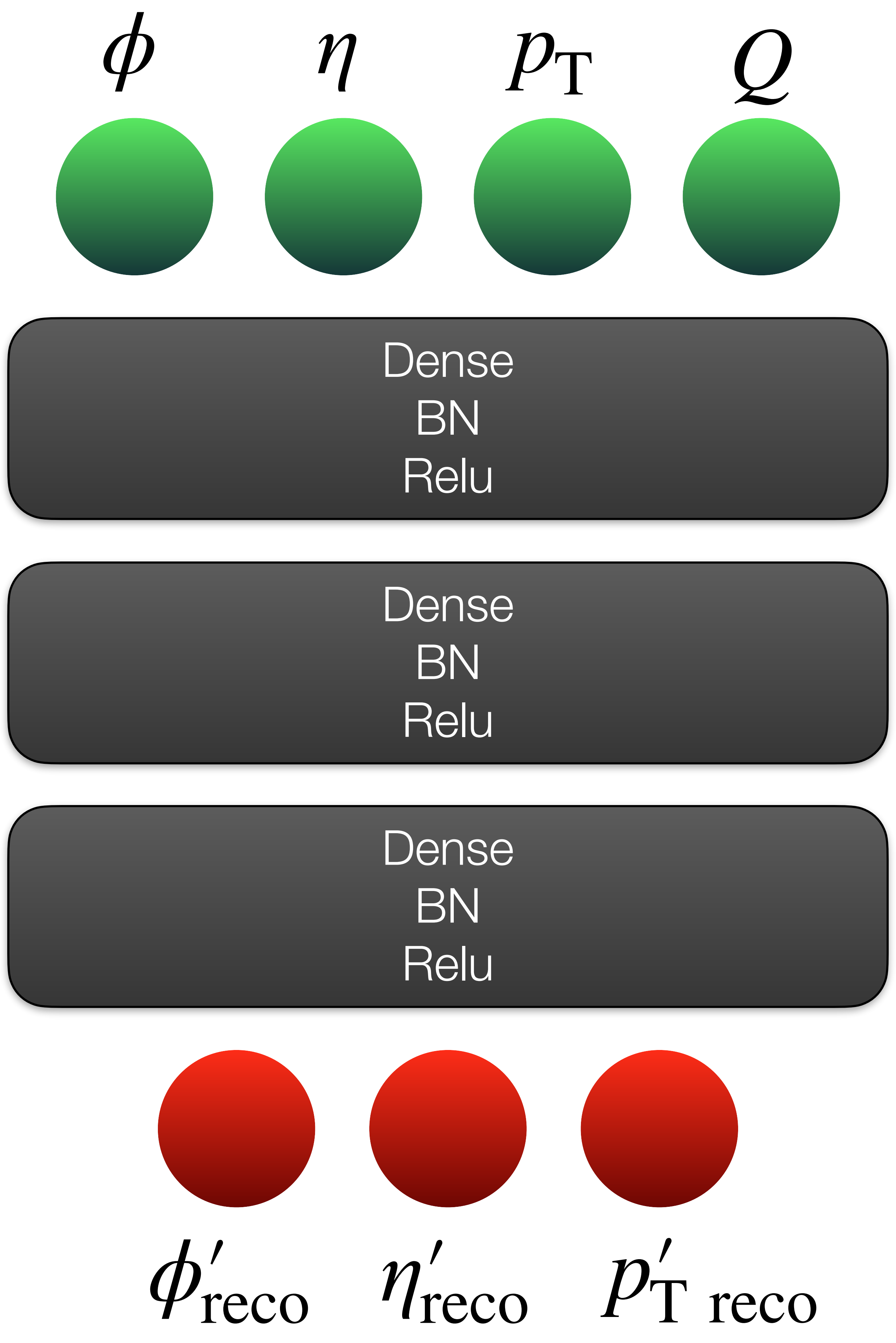
- » GMT muons; Barrel-Barrel pairs
- » Muon p_T 0-15 GeV
- » NN re-calibrations & fake pair prediction applied
- » Duplicate removal $d\phi < 0.1$ rad
- » Equivalent results obtained for endcap, overlap, and mixed TF pairs

Scouting: Latency and precision

- Close to latency target:
- Majority of latency from data/weights transfer RAM/FPGA:
  - Batching implemented to remove this bottleneck

Micron hardware	Latency / inference
SB-852 - 4 cluster	1.3 μs
Target	1.0 μs

Precision |hw - py. sw|	Frac. < 1% diff
Model w/ integer inputs, no batch norm	99%

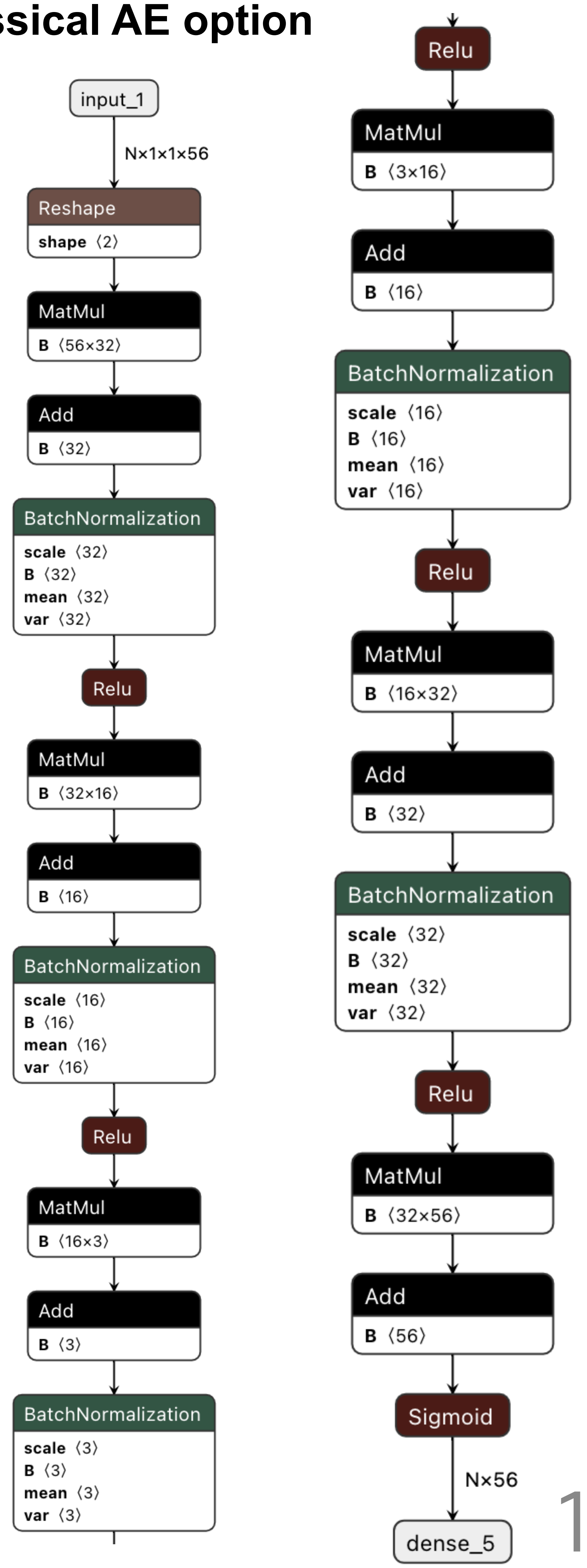


Autoencoder for anomaly detection in L1 Trigger

- Anomalous events -> may be new physics candidates
- Model/theory independent
- Train on Standard Model 'QCD' background
- Inputs: fixed size arrays/images of up to 10 jets, 4 muons & 4 electrons & MET (each with 3 parameters)
- Test with simulated Beyond Standard Model events e.g new massive vector bosons, unusual Higgs decays

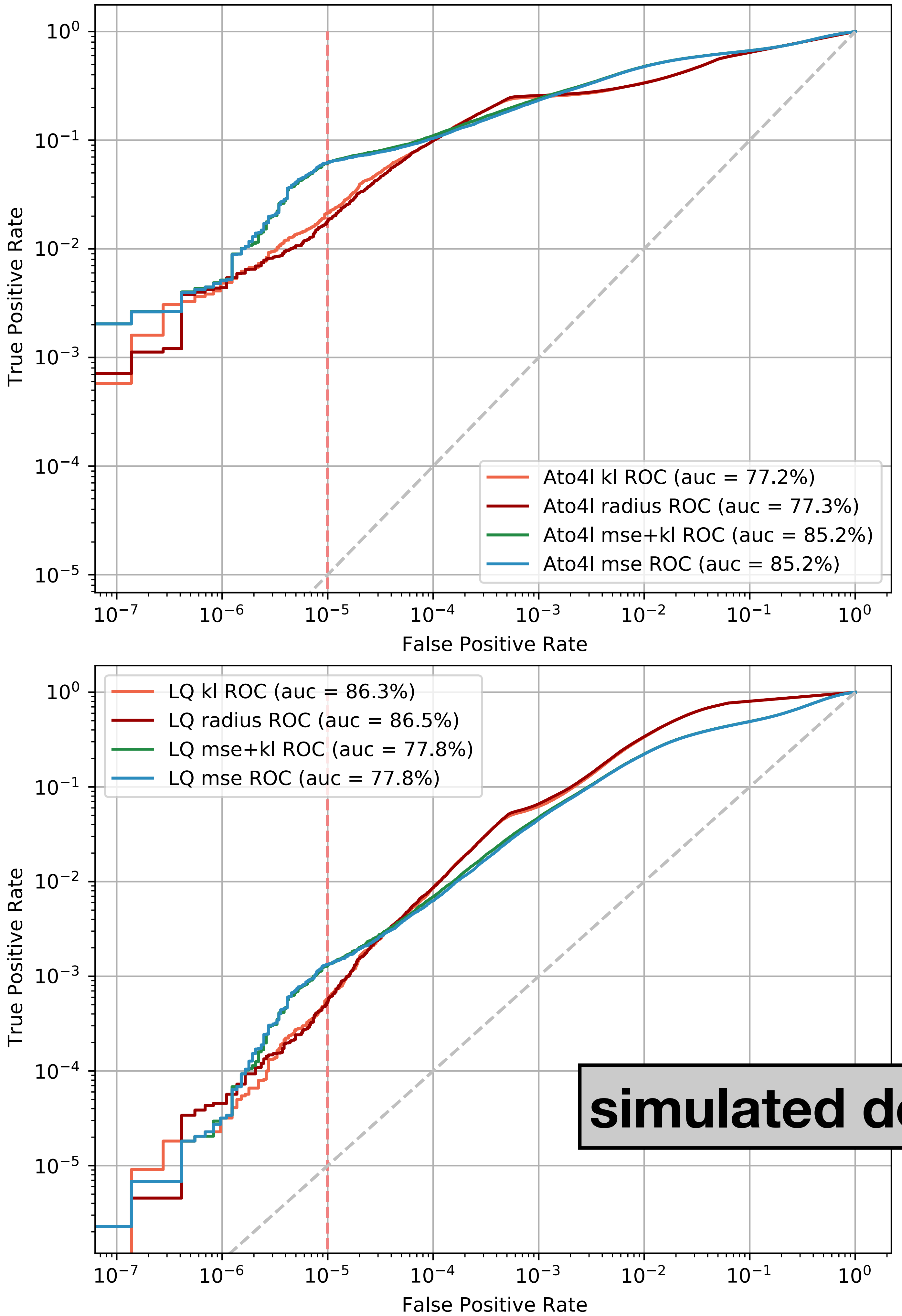
- Option to run in scouting system (no strict latency requirement)
- Developing both classical, convolution, and variational auto-encoders and comparing all approaches

Classical AE option

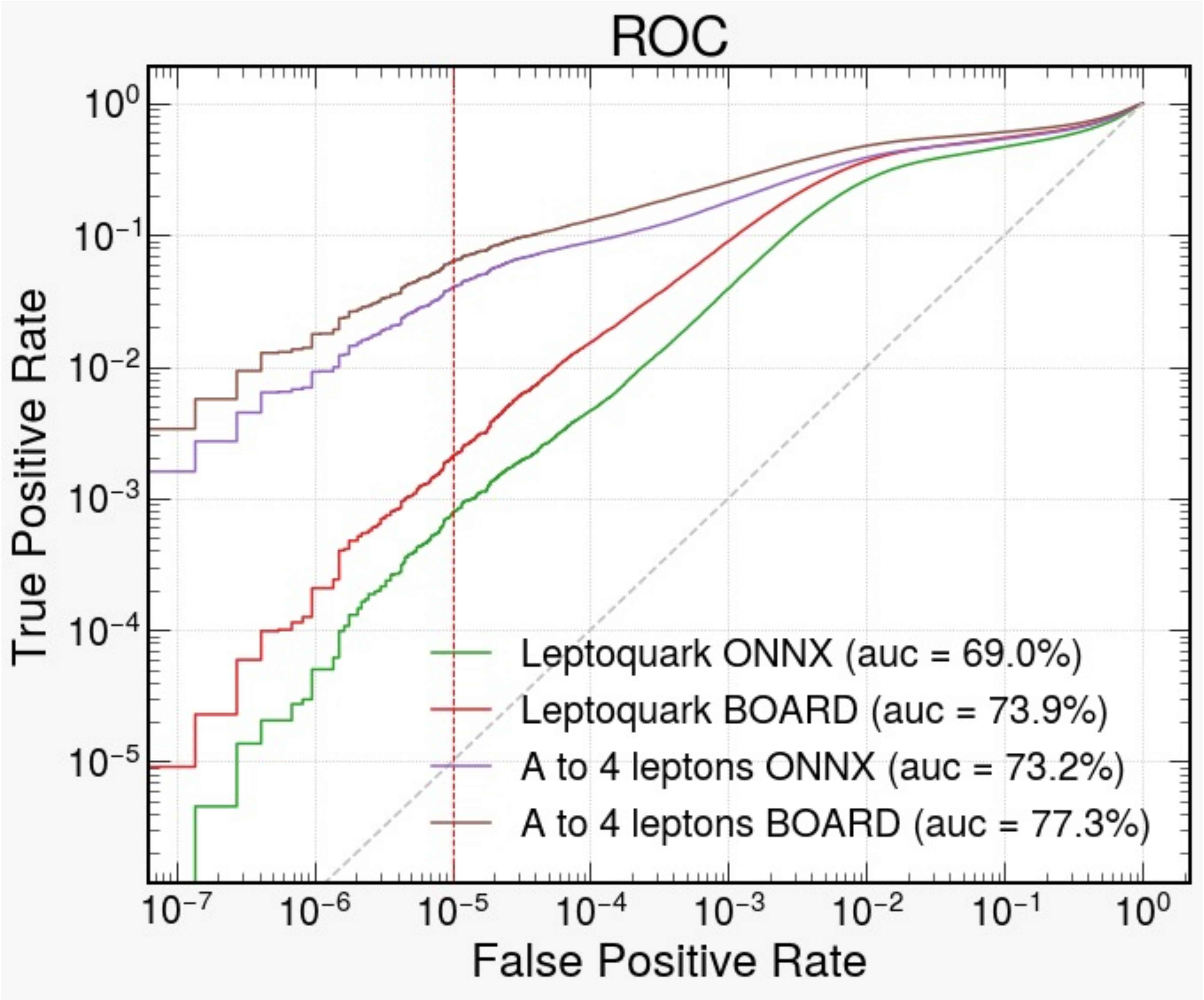


Autoencoder for anomaly detection in L1 Trigger

- Successfully ran on SB-852 - ironing out precision differences - latency ~625ns



simulated delphes data



Summary and next steps

CERN openlab

- Challenging year (for everybody), but progress made on all fronts
- Run 3 delayed Q1 2022
- Focus: testing and integrating aspects of the scouting system & associated hardware, firmware, software - use of monthly ‘Global Runs’ with cosmic muons to test system in situe
- Performance of deep-learning driven anomaly detection algorithm being evaluated for use at LHC Run 3
- Thank you to the team at Micron for the great collaboration - looking forward to see what 2021 brings!

CERN

Micron®

CMS

Compact Muon Solenoid

CERN openlab