Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

FSP CMS
Erforschung von
Universum und Materie

Bundesministerium
für Bildung
und Forschung

Real-time ML event classification with FPGAs at the LHC

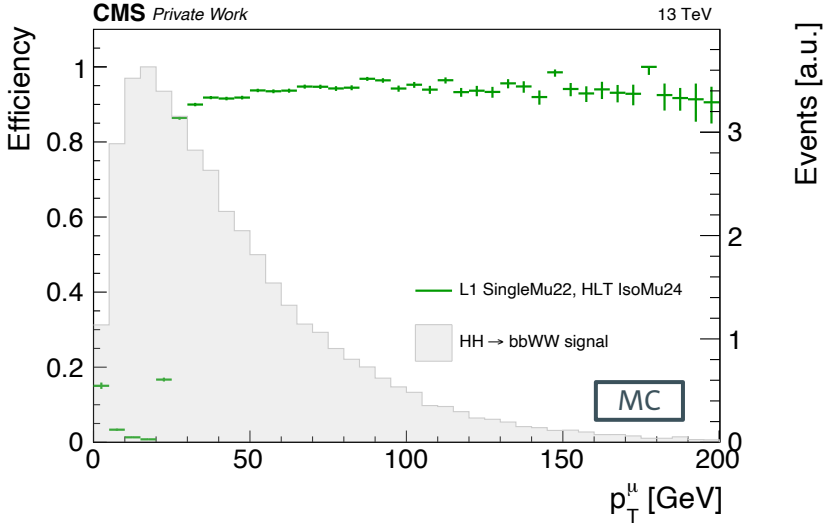
Finn Labe on behalf of the UHH ML@L1 team

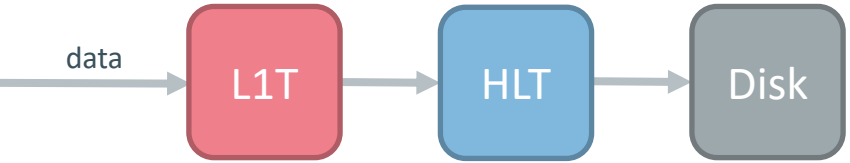
Deep Learning Roundtable at DESY | 08.12.2023

Why do we need ML in the trigger?

- **Trigger performance** decides what data is available for **offline analysis**!

- Sensitivity limited by trigger thresholds
  - Example: **HH → bbWW (single lepton)**
  - Possible solution: **ML@L1**

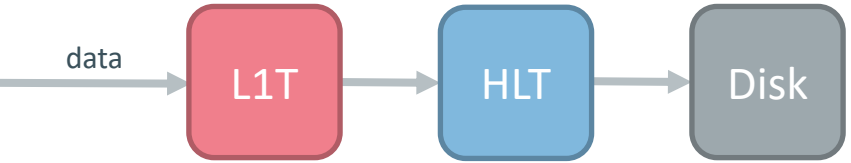


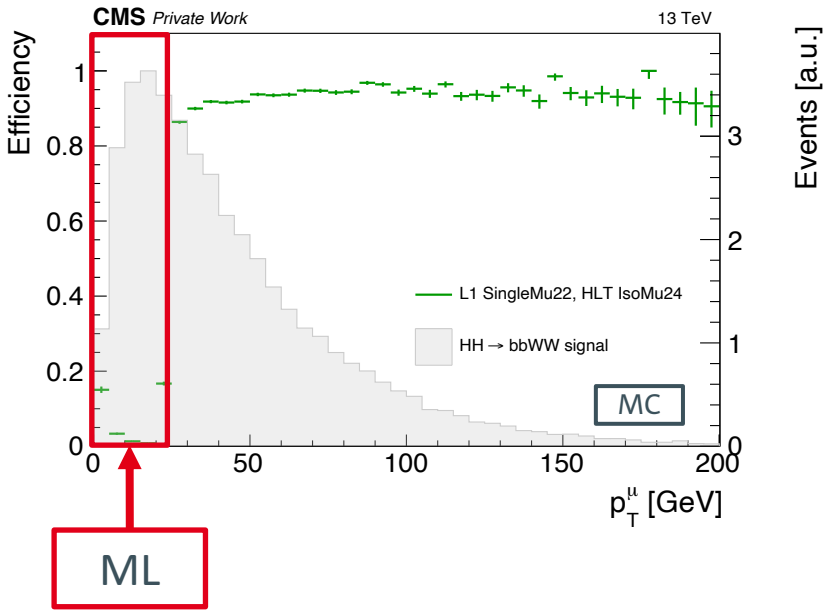


Why do we need ML in the trigger?

- **Trigger performance** decides what data is available for **offline analysis**!

- Sensitivity limited by trigger thresholds
  - Example: **HH → bbWW (single lepton)**
  - Possible solution: **ML@L1**

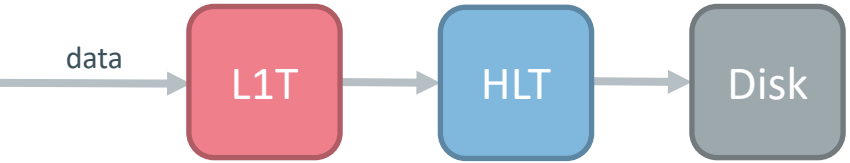


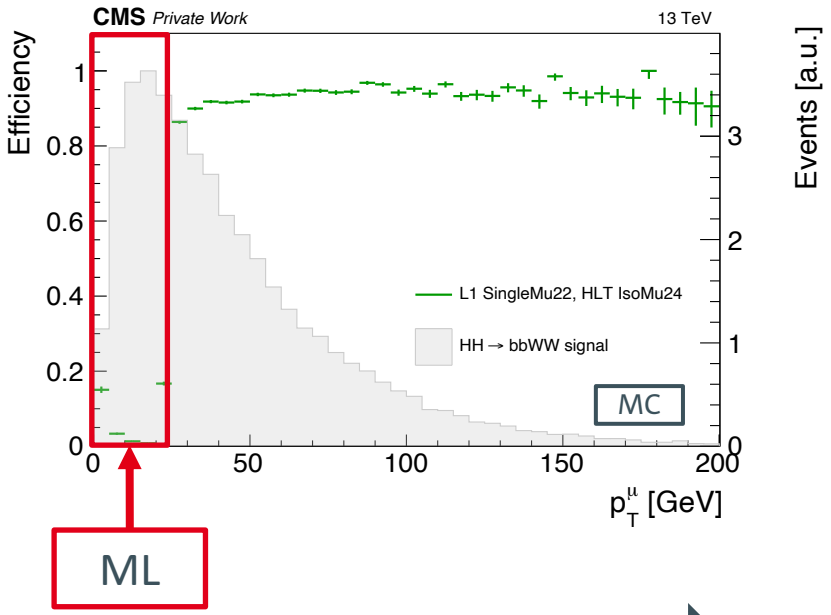


Why do we need ML in the trigger?

- **Trigger performance** decides what data is available for **offline analysis**!

- Sensitivity limited by trigger thresholds
  - Example: **HH → bbWW (single lepton)**
  - Possible solution: **ML@L1**

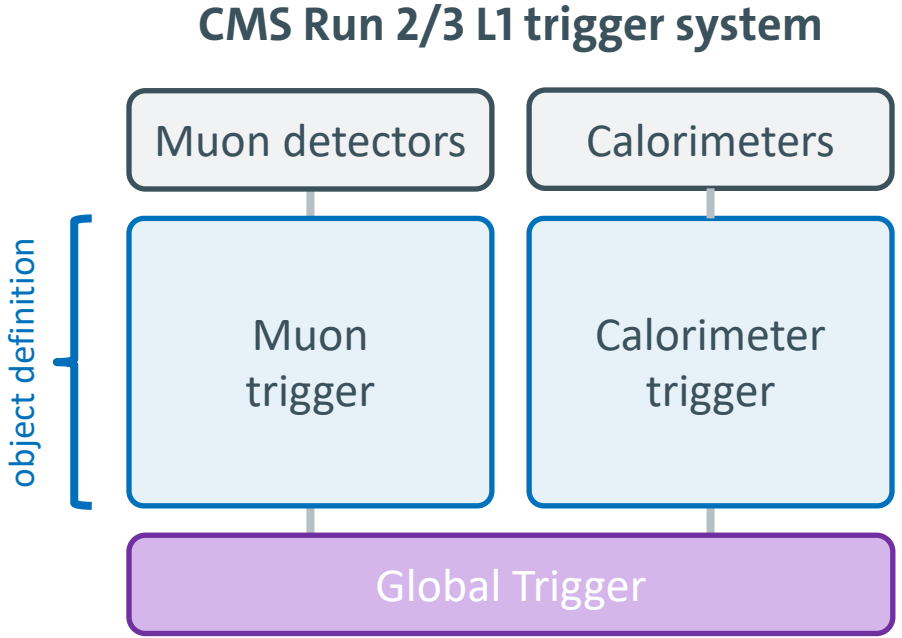




Focus: CMS L1 trigger system

Event classification in the L1 trigger

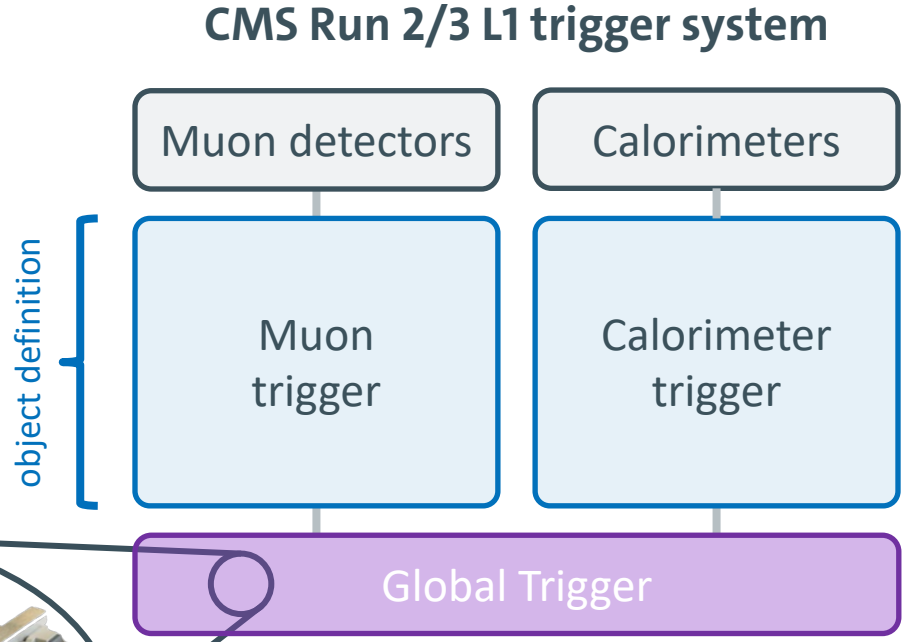
- Event classification in **global trigger**
  - Uses **objects** defined further upstream in L1 trigger system
- **Binary network**, cut on output node defines L1 trigger condition



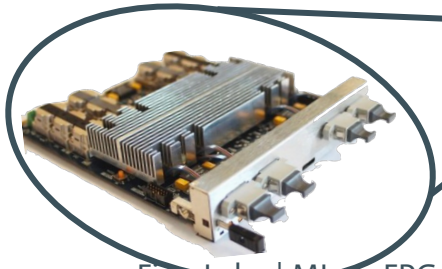
Total L1 rate
100 kHz (Run 2/3)
750 kHz (HL-LHC)

Event classification in the L1 trigger

- Event classification in **global trigger**
  - Uses **objects** defined further upstream in L1 trigger system
- **Binary network**, cut on output node defines L1 trigger condition



L1 trigger built from **F**ield-**P**rogrammable **G**ate **A**rrays



Total L1 rate
100 kHz (Run 2/3)
750 kHz (HL-LHC)

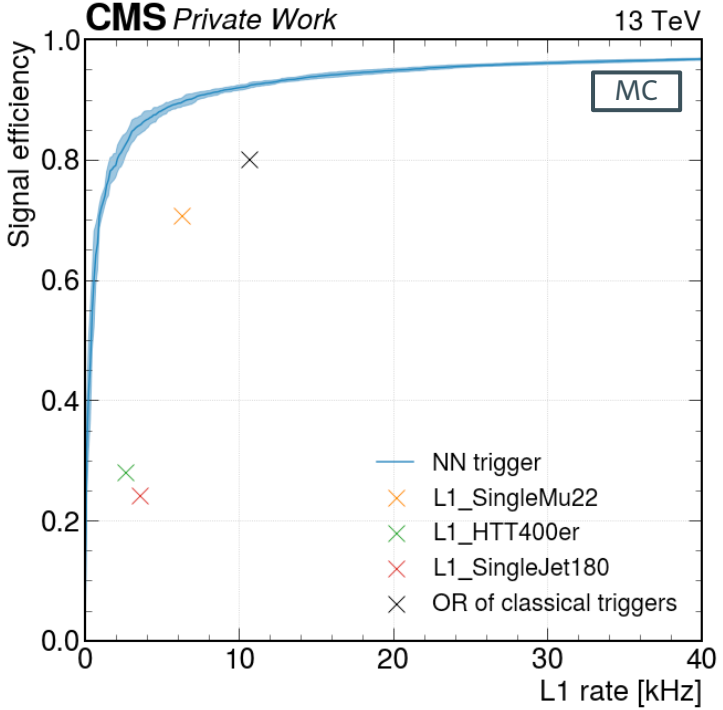
Trigger NN development

*Most common low energy LHC events

- Supervised training: simple NN
  - Discriminating signal from **MinBias***
  - Using p_T, η, φ of L1 objects

- Comparing to cut-based triggers:
  - NN outperforms at any given rate

ML@L1 is promising!



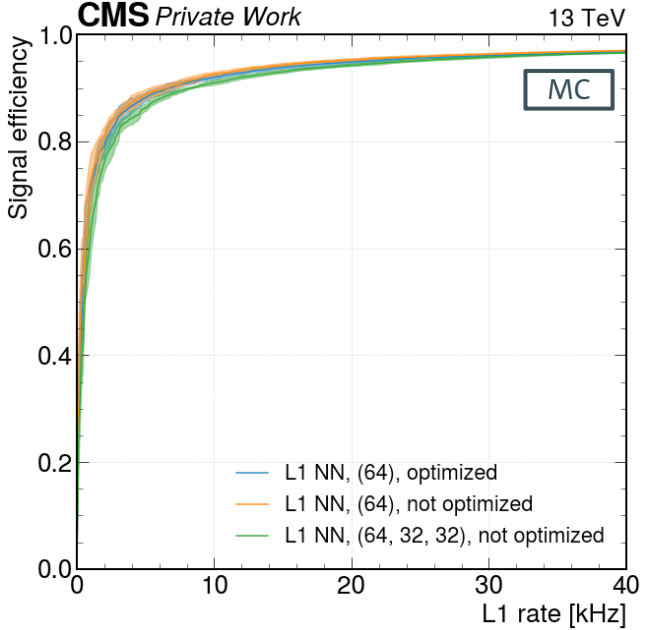
Neural Network on FPGA

- Conversion for FPGA using hls4ml library

- Compressing network to fit FPGA
  - Single (hidden) layer network
  - **Pruning**: removing connections
  - **Bit precision**: less bits per weight

optimization

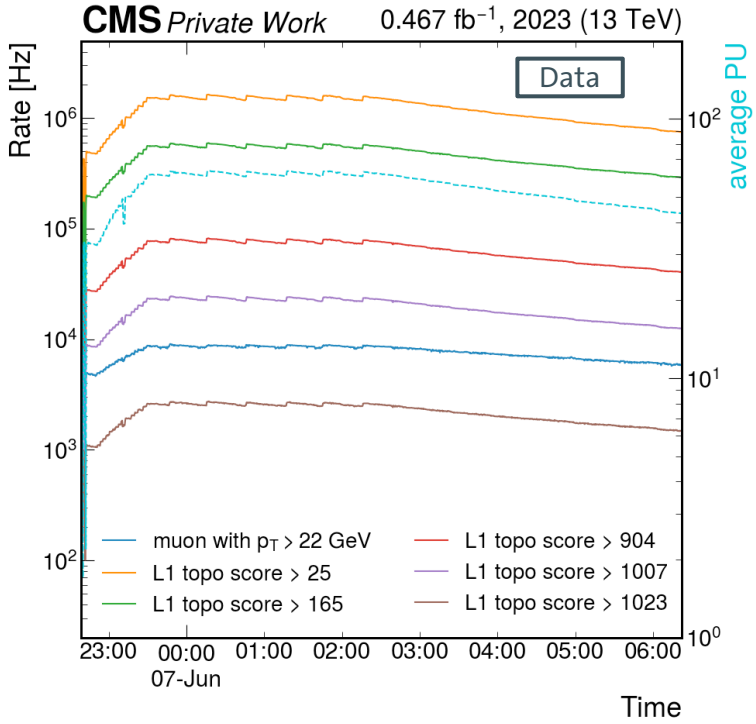
- **75 ns** latency (target < 100 ns)
- **2%** of an FPGAs resources (6 available)



Integration tests

*in so-called test crate ➜ no detector readout

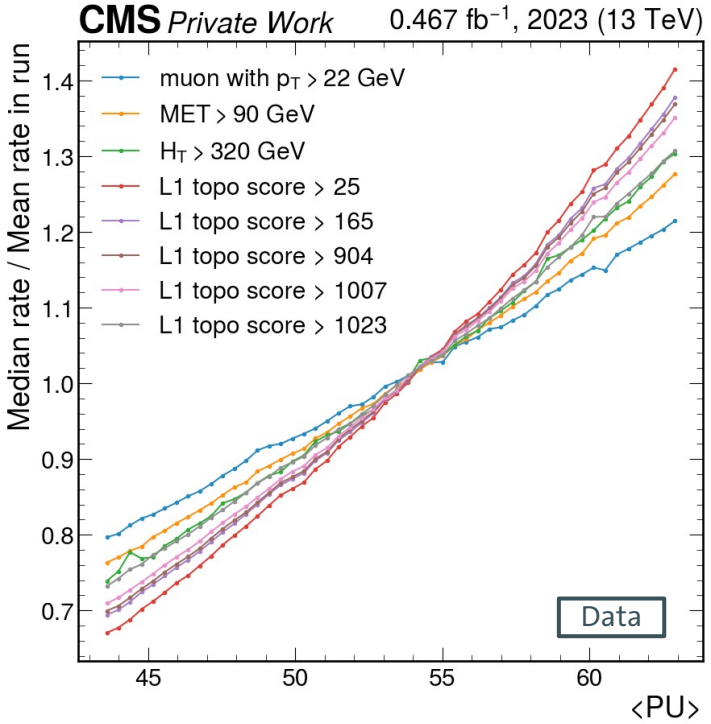
- Testing NN trigger in CMS operation*
  - Probe stability over wide threshold range



Integration tests

*in so-called test crate ➜ no detector readout

- Testing NN trigger in CMS operation*
  - Probe stability over wide threshold range
- PU dependency comparable to traditional cut-based triggers

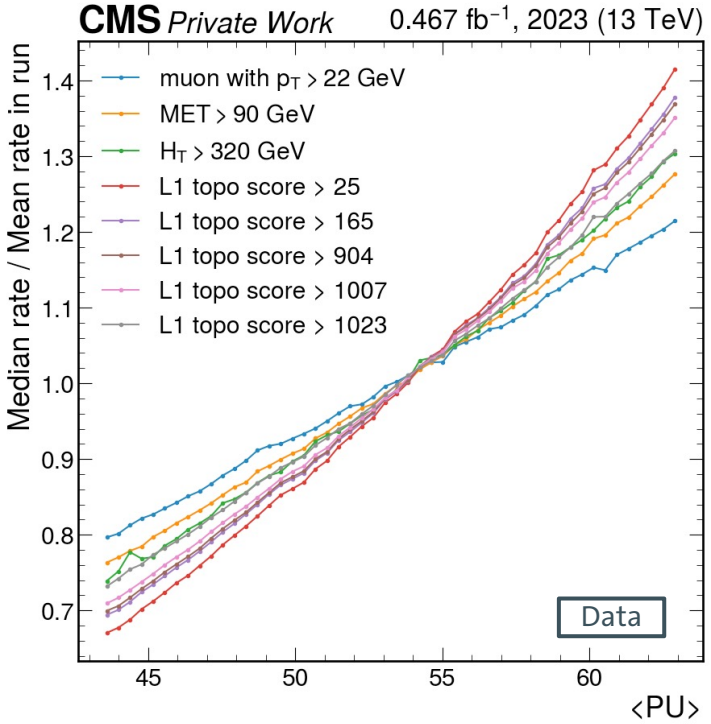


Integration tests

*in so-called test crate → no detector readout

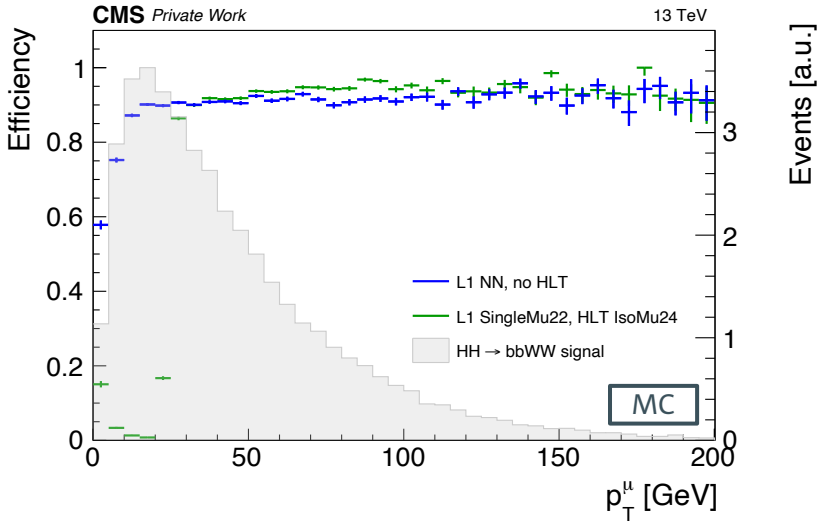
- Testing NN trigger in CMS operation*
  - Probe stability over wide threshold range
- PU dependency comparable to traditional cut-based triggers

Integration results very important to us: this kind of trigger is feasible!



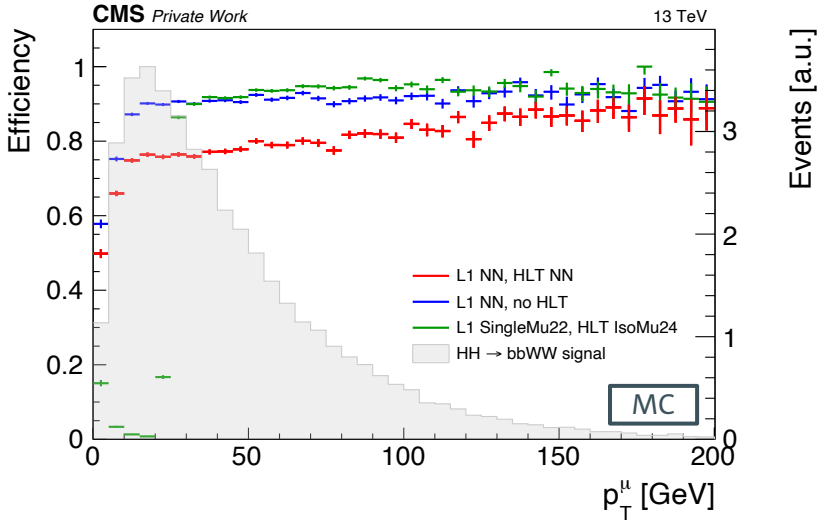
Increasing analysis sensitivity

- Studying expected **sensitivity gain** from NN-based L1 algorithms
  - Simulation-based integration in run 2 $\mathbf{HH \rightarrow bbWW}$ analysis
- L1 NN targeting **10 kHz** L1 rate
  - Resulting **pure rate** much lower



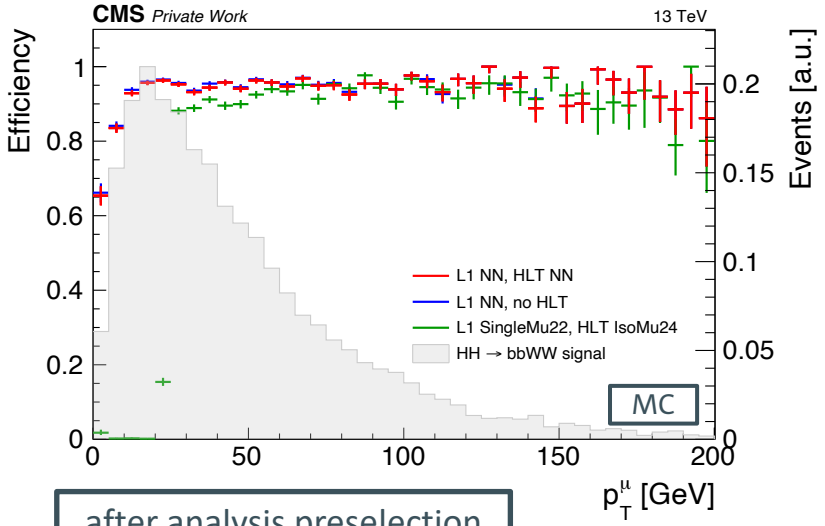
Increasing analysis sensitivity

- Studying expected **sensitivity gain** from NN-based L1 algorithms
  - Simulation-based integration in run 2 $\mathbf{HH \rightarrow bbWW}$ analysis
- L1 NN targeting **10 kHz** L1 rate
  - Resulting **pure rate** much lower
- How to handle **HLT strategy**?
  - Use cut-based trigger
  - Train **another neural network**



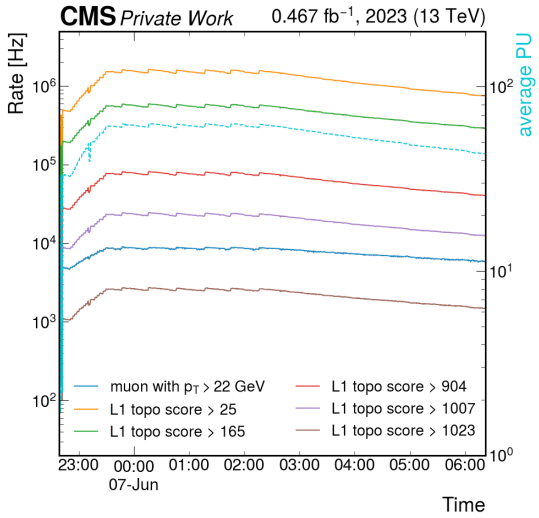
Increasing analysis sensitivity

- Studying expected **sensitivity gain** from NN-based L1 algorithms
  - Simulation-based integration in run 2 $\mathbf{HH \rightarrow bbWW}$ analysis
- L1 NN targeting **10 kHz** L1 rate
  - Resulting **pure rate** much lower
- How to handle **HLT strategy**?
  - Use cut-based trigger
  - Train **another neural network**



Summary & outlook

- **Supervised ML@L1** promising for otherwise trigger-limited signals
- Integration of NNs in the CMS level 1 trigger demonstrated!
- Interesting open questions
  - Can we **generalize** supervised ML in the trigger system?
  - **Efficiency** measurement



Another approach is anomaly detection in the L1T:
UHH with AXOL1TL and CICADA

The UHH ML@L1 team

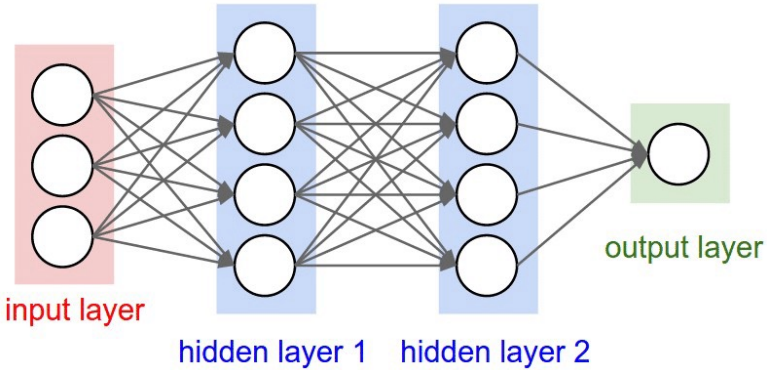
Finished thesis and left the group

Project lead	Artur Lobanov (= AL, postdoc)
PIs	Johannes Haller, Gregor Kasieczka (Profs)
Higgs Expert	Matthias Schroeder (Staff)
Supervised ML triggers	Finn Labe (PhD), Shahin Sepanlou (MSc), Ihor Komarov (MSc), Salome Fresenbet (BSc), Karla Kleinbölting (BSc, MSc), AL
Anomaly detection at L1	Sven Bollweg (PhD), Lars Emmerich (BSc), Susan Sefidrawan (BSc), Karim El Morabit (postdoc), AL
ML jet identification	Philipp Rincke (MSc), Karim El Morabit (postdoc), AL
L1T Menu	Run3: Sven Bollweg (PhD), AL Phase2: Daniel Hundhausen (PhD), Matteo Bonnanomi (postdoc), AL
L1 Run3 DQM	Mathis Frahm (PhD), AL

Backup

L1 neural network details

- Binary fully connected network



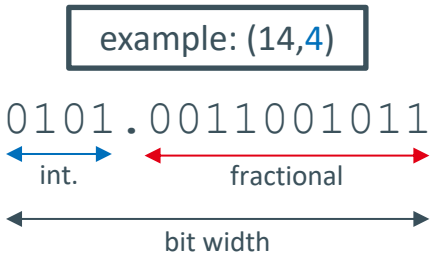
- **ReLu** activation, BCE loss

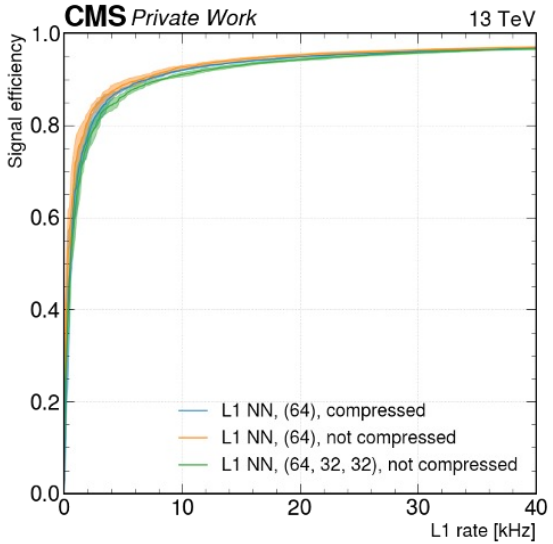
- 26 input variables
  - MET, MET φ
  - 4 jets: p_T, η, φ
  - 2 muons: p_T, η, φ
  - 2 “egammas”: p_T, η, φ

Electrons and photons indistinguishable without tracker

Compression details

- Lowering **bit precision** to (6,1) for inputs & weights
  - Compression performed using **qkeras** package





- Pruning away 80% of node connections

$\mathbf{HH \rightarrow bbWW}$ analysis details

The following offline cuts were used to study the trigger:

- Exactly one muon with CutBasedIdTight, PFIsoTight and $|\eta| < 2.4$
- At least three AK4 jets fulfilling tight PuPPi criteria
- At least one of them fulfilling loose DeepJet b-tagging criteria
- Distance between lepton and jets $\Delta R > 0.2$
- Transverse mass $M_T\left(\ell + E_T^{\text{miss}} + b_{\text{lep}}\right) > 60$ GeV

In the offline analysis, an event classification DNN is next.

Rate check using run 3 data

- Training & evaluating using Run 3 “ZeroBias” sample
  - Signal efficiency at **5 kHz** of rate is below **500 Hz pure rate**