ML algorithms @ FPGA in the CMS Level-1 Trigger

Sven Bollweg, Karim El-Morabit, Finn Labe, Johannes Haller,
Gregor Kasieczka, **Artur Lobanov**, Lars Emmrich, Matthias Schroeder
Uni Hamburg, Institut für Experimentalphysik

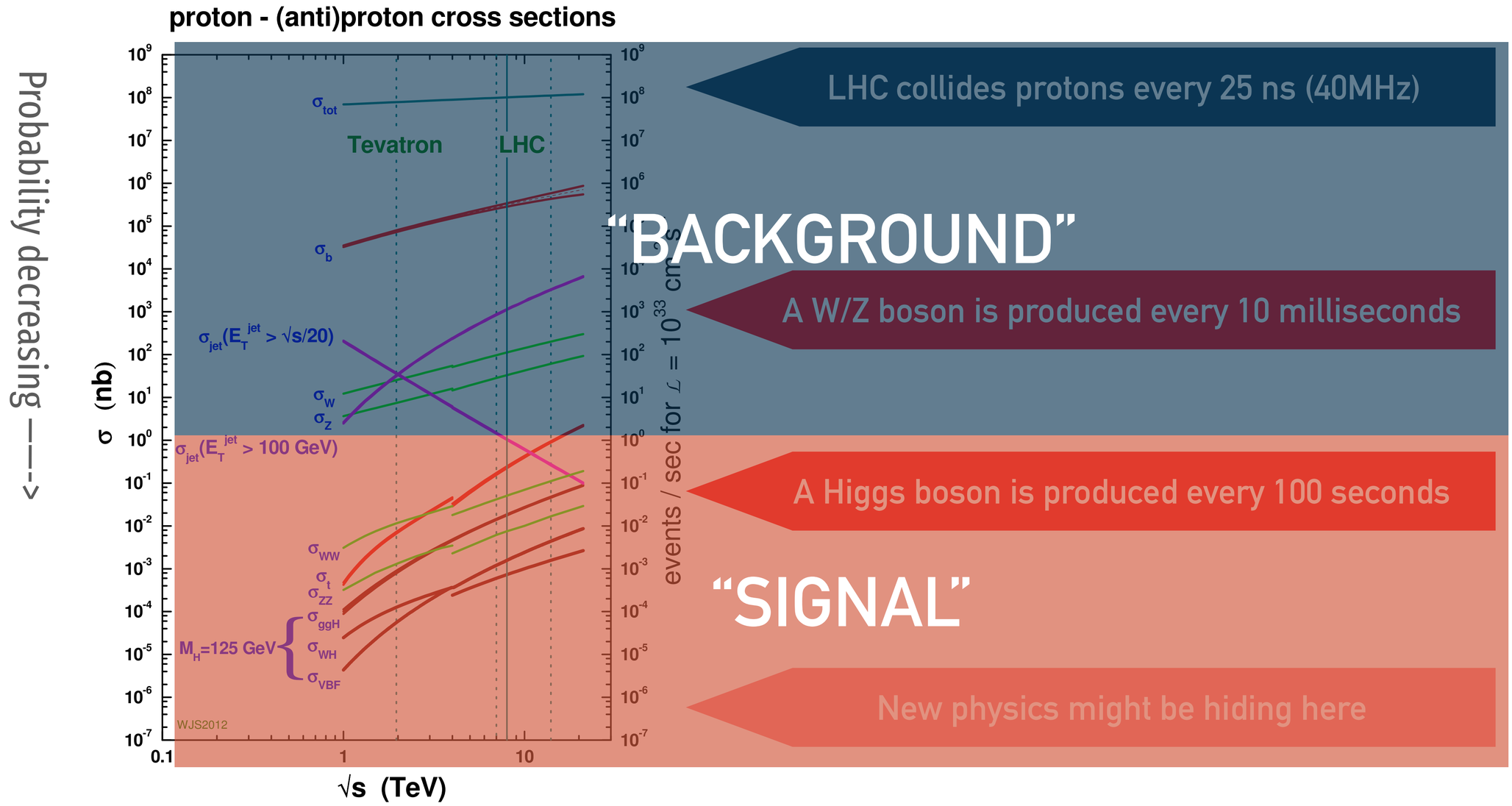
DESY AI Roundtable | 25.11.2022

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

CLUSTER OF EXCELLENCE
QUANTUM UNIVERSE

Searching for the needle in the LHC Haystack

Artur Lobanov |
ML @ L1 Trigger in CMS |
DESY AI Roundtable 11.2022 |

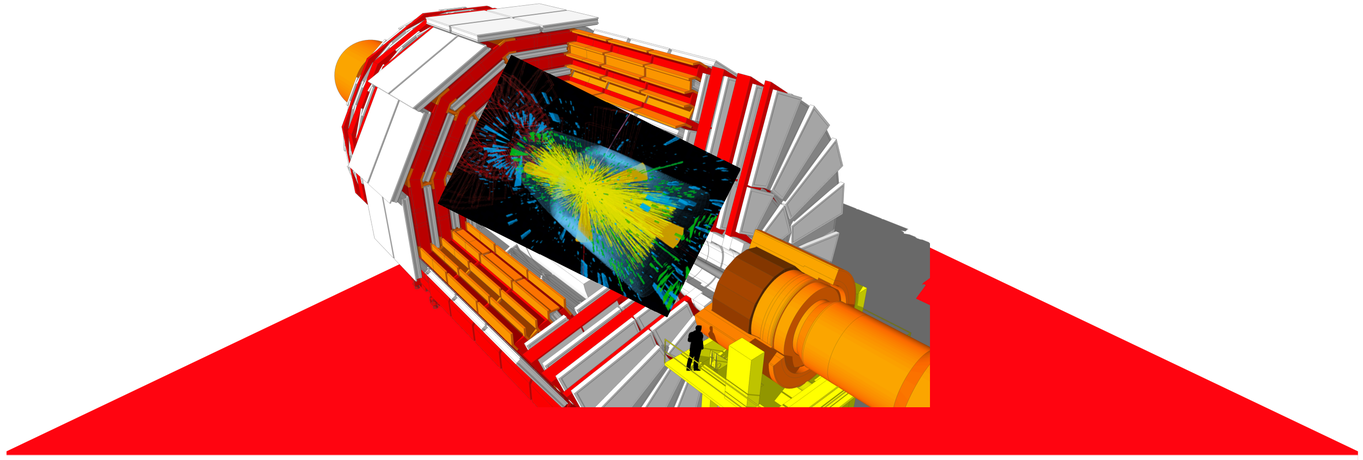


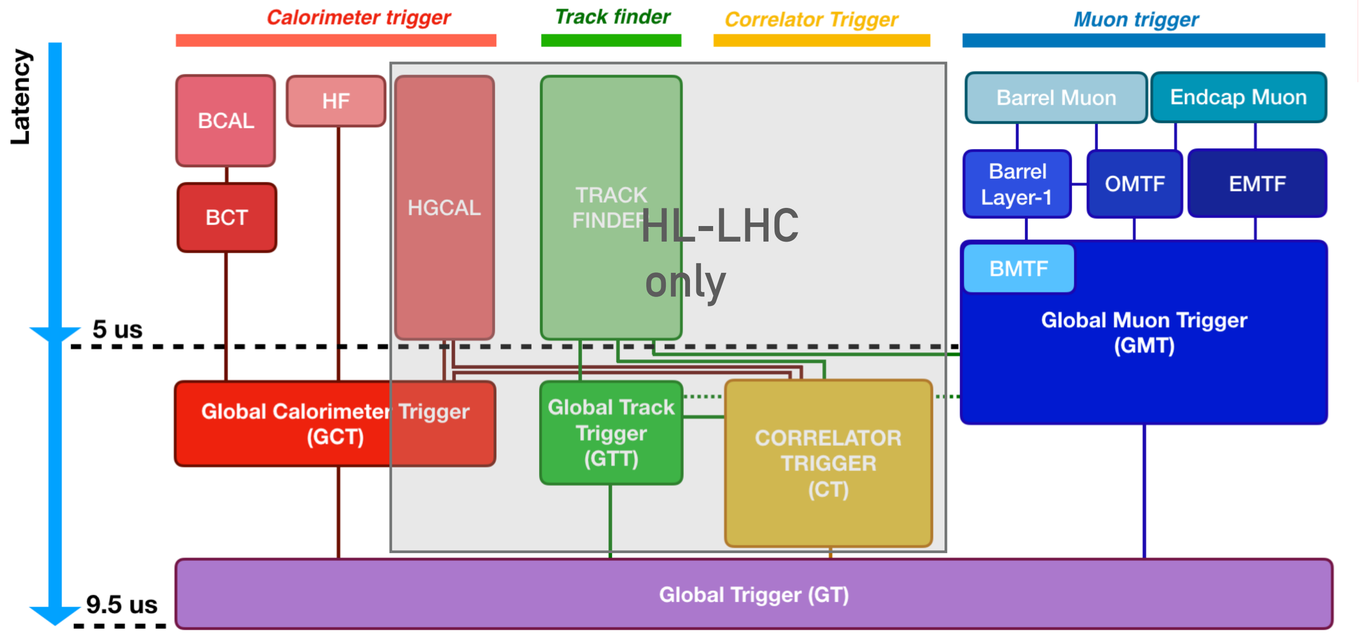
CMS Level-1 Trigger

Detector data "in"
@40MHz

Processing data
and reconstructing
physics objects
~9us

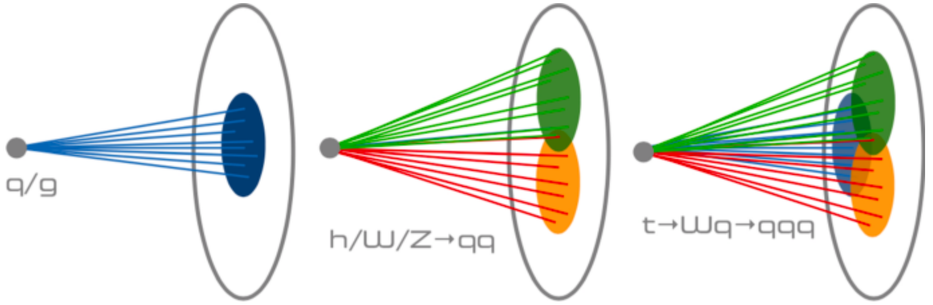
Decision on event ~1us

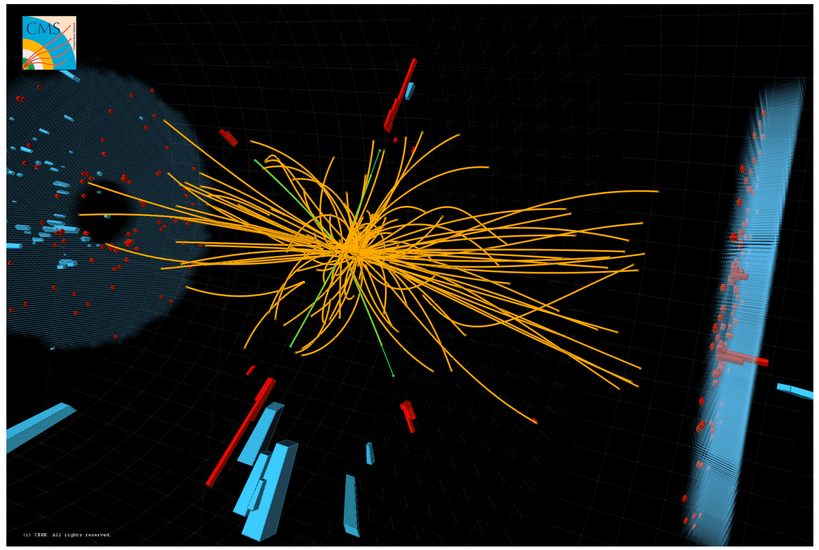




ML@L1T: Jet and Event Classification

- **Our group's subjects of studies:**
  - Jet identification based on jet constituents (in CT)
  - Event classification based on topology (in GT)
  - **Inherently both are based on the object "topology"**
- **Jet classification:**
  - PU vs light vs heavy-flavour jet etc.
- **Event classification:**
  - Go beyond simple correlations and learn kinematics using Machine Learning (ML)
  - Separate signal(s) vs. background ("MinBias")
- **ML-approach effectiveness already proven "offline"**





ML Topology Triggers

- Traditional L1 triggers: 1-4 particles, filter on energy + kinematical correlations
  - Mostly general purpose, recently more signal-targeted (e.g. B->mumu)
- **ML approaches based on ~full event information = all detected "particles" (@L1)**
  - **Target inaccessible signal-phase**: soft final states, unusual signatures etc.

- **Classifier:** *"supervised ML"*
  - Event classification: signal vs background
  - Model-dependent
  - High purity

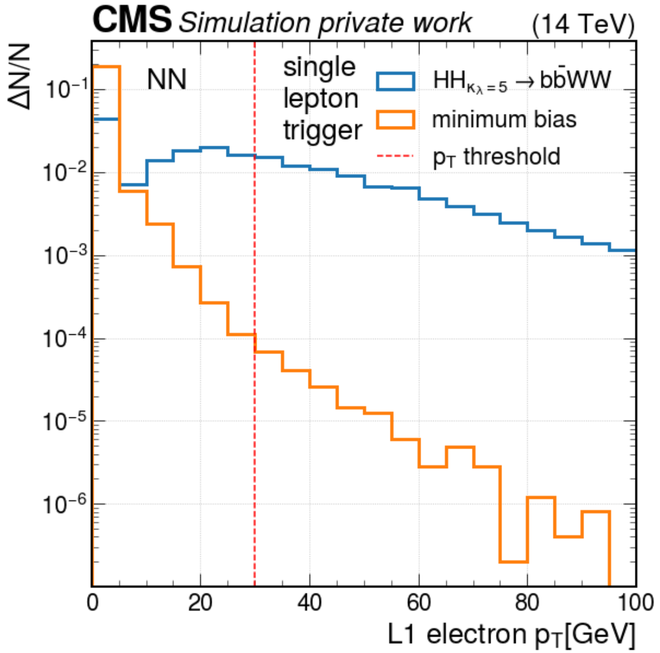
... ML-powered traditional trigger

- **Anomaly detection:** "*unsupervised ML*"
  - Event classification: reject background-like events
  - Model-independent
  - Low purity

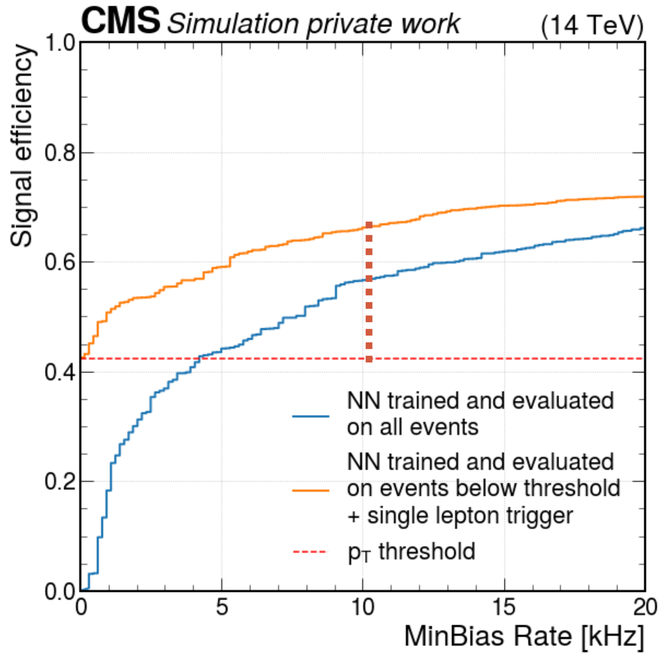
... novel approach (impossible w/o ML)!

ML Topology Trigger vs Standard Approach

- Benchmark signal: HH > bbWW (semi-leptonic) -> soft decay products
- NN should be complementary to “existing” L1 Trigger menu, e.g. single lepton triggers
  - Train/evaluate NN trigger only on phase-space not covered by single lepton
  - NN added efficiency: **> 25% at 10kHz -> 60% total gain (wrt 30GeV single ele.)**



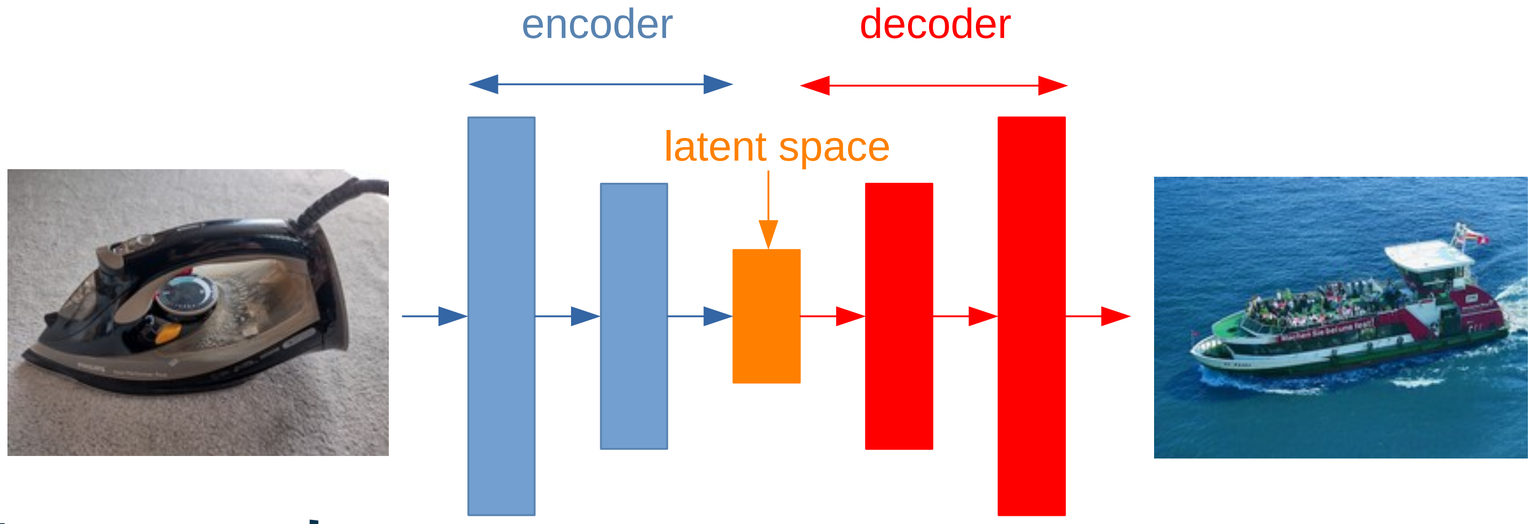
Lepton pT coverage by NN vs standard trigger



Single electron at 30 GeV = 30kHz

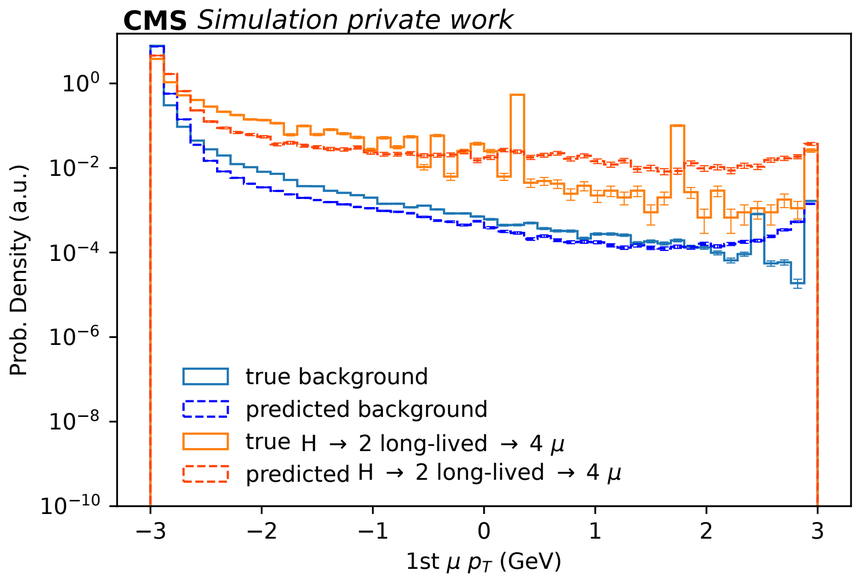
Anomaly Detection @ L1T

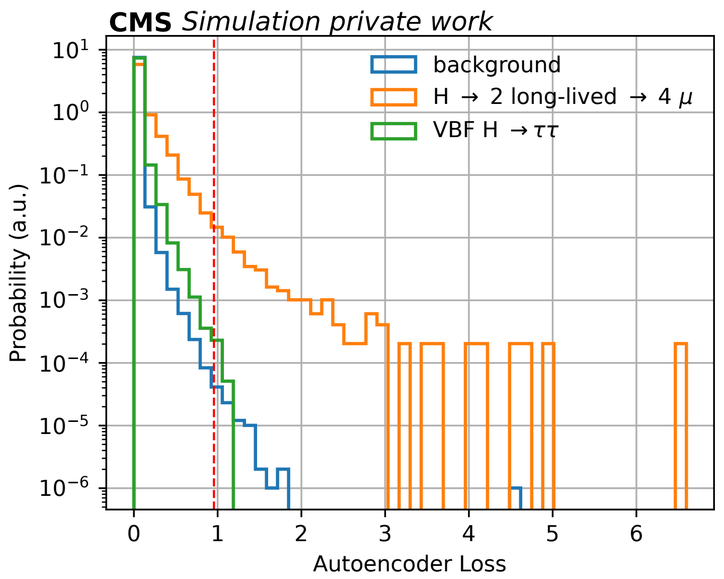
- **New approach in triggering: detect anomalies with ML model taught on background only**

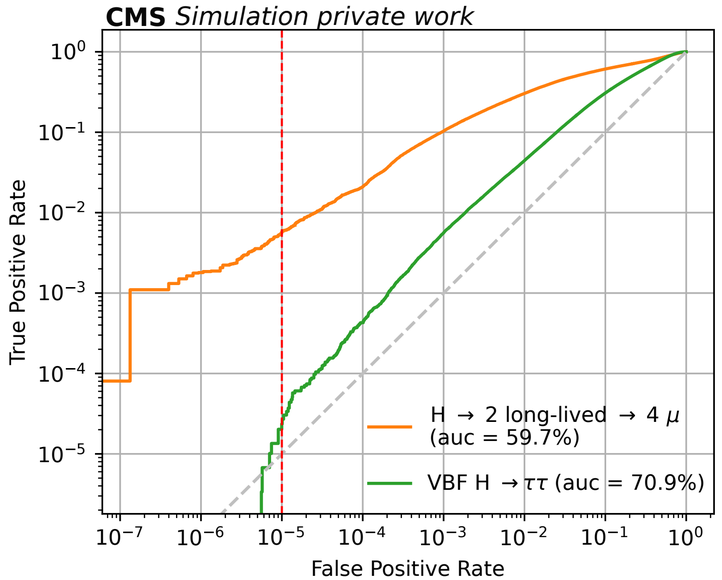


- **Based on ML auto-encoders:**
  - Encoder compresses input, decoder reconstructs the input from the latent space
  - Trained with mean squared error (MSE) loss of input and output
    - Good reconstruction performance for data similar to the training set
    - **Bad reconstruction for data different to the training set**

Anomaly Trigger





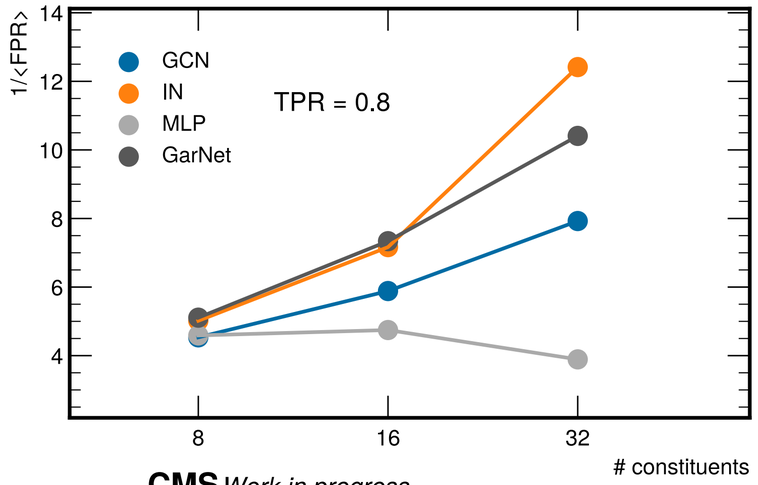


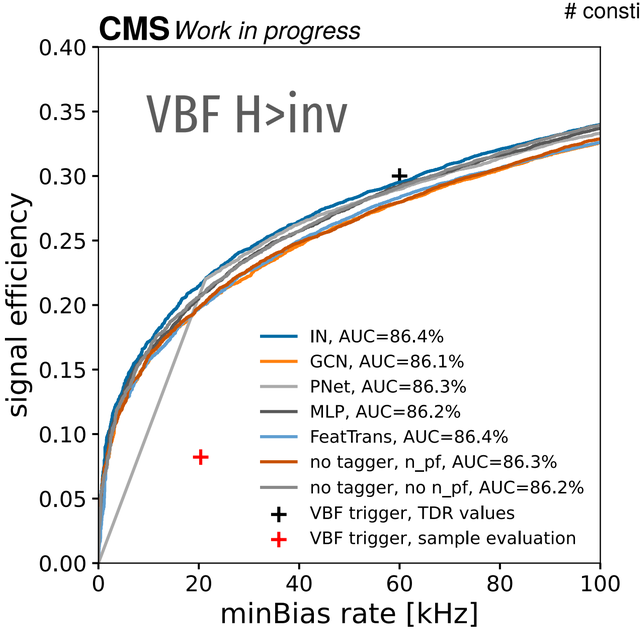
- Validate autoencoder (AE) by checking reconstructed variable distributions
- Use “AE loss” as discriminating variable on trigger level
- Background will dominate: low trigger rate → low false positive rate (and signal eff)

Jet Identification

- Contributed to **new study of Graph-based NNs for Jet Tagging** using synthetic dataset w/ HLS4ML team
  - GNNs profit from larger N of constituents
- In CMS L1T: investigating Jet tagging in “CT” system
  - **JetID could be used for “simple” jet triggers or as input to GT Topo Triggers**
  - Studied different jet ID problems for low-pT jets (untriggered)
  - Looking into NN Topo trigger for VBF H>inv
    - Similar to HH approach: low-level feature NN
    - **Gain acceptance wrt the L1 menu VBF seed!**

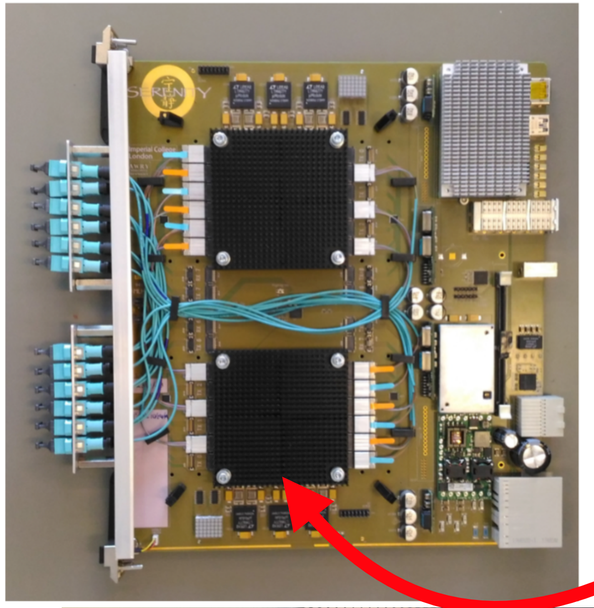
HLS4ML (WIP)



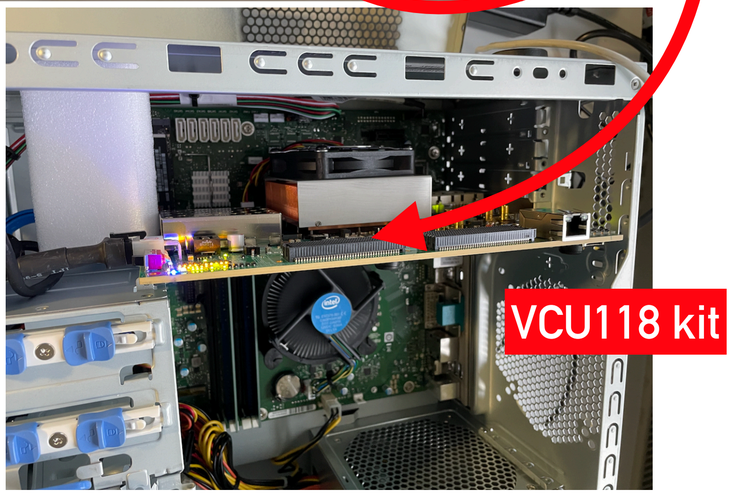


Hardware Demonstration

- L1T **algorithms run on powerful FPGAs,** e.g. Xilinx VU9P
  - Fast I/O (25Gb/s) for L1T data transfer
  - **Large FPGA "memory"** useful for storing complex algorithms, e.g. Neural Network weights
- While CMS host board "Serenity" in R&D, use commercial "development kit" for demonstrator setup:
  - **VCU118 kit hosts same FPGA** as Serenity and provides fast interfaces to PC (optics or PCIe)
  - Using the setup to **test & run algorithms in a realistic FPGA environment**



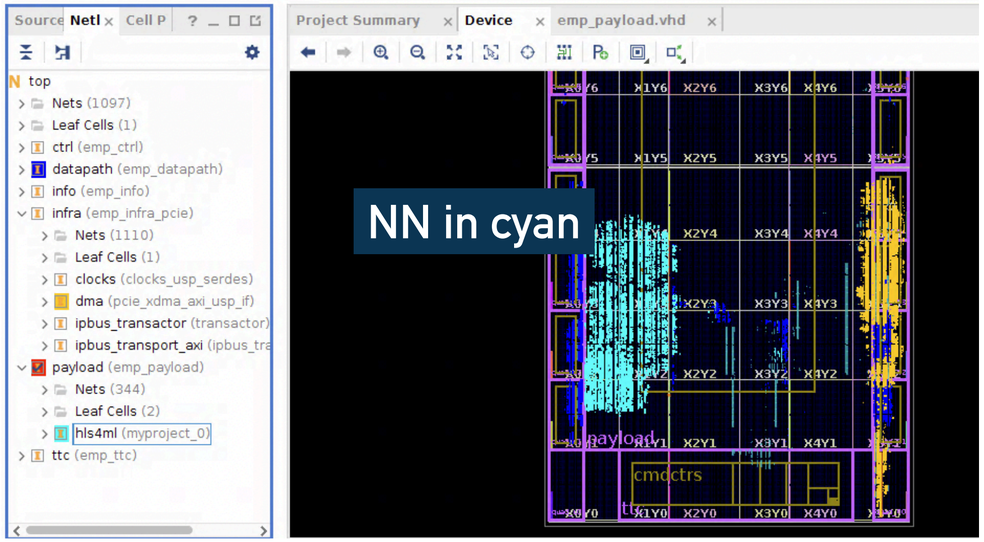
Serenity board for CMS L1T



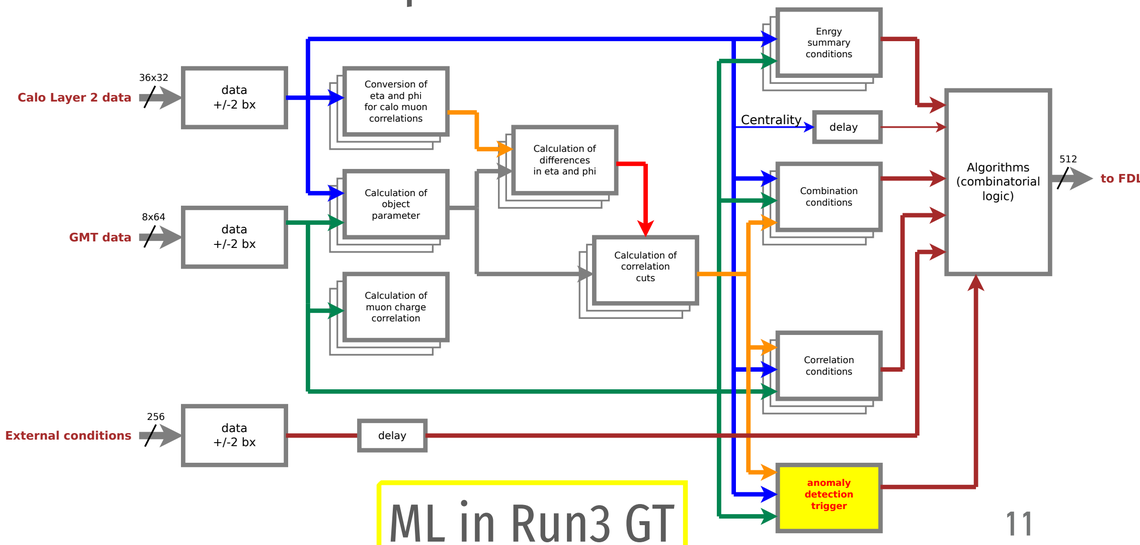
VCU108 setup @UHH

ML@FPGA RESOURCES

- Collaborating on algorithm development and hardware demonstration:
  - HLS4ML team (M. Pierini et al)
  - CMS Global Trigger teams (H. Sakulin, M. Jeitler)
- **FPGA implementation of NN with** HLS4ML in the CMS L1T firmware architecture
- **Performance and resource usage promising!**
  - **Latency ~50ns** -> good for Run3 already!
  - **Resources: ~ few %** for of FPGA
- **Targeting first tests of Topo and Anomaly Triggers for LHC Run3 soon!**



NN implementation in firmware



ML in Run3 GT

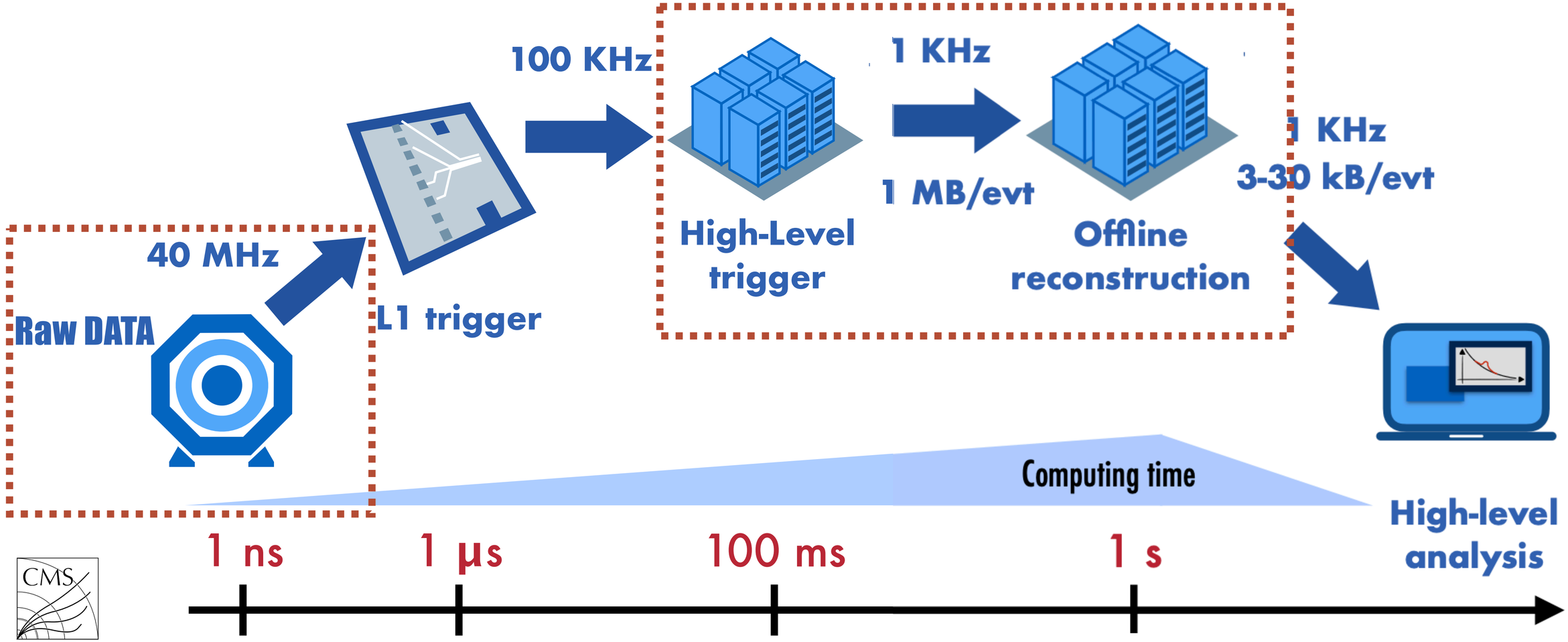
SUMMARY

Summary

- *ML enhances physics sensitivity throughout HEP experiments' data flow*
  - ML arrives in "online" (trigger) systems of e.g. the CMS experiment
- Performing **proof-of-concept of ML algorithms for the CMS L1T** in several areas:
  - **Topology trigger**: promising performance for various benchmark signals
  - **Jet identification**: benchmarks promising, exploring "realistic" CMS datasets
  - **Anomaly Detection**: advancing this novel approach in trigger systems
- First **hardware demonstrations achieved** in HL-LHC system (w/ CERN teams)
- **Targeting first real implementation of Topo & Anomaly Triggers in Run3 already!**

Outlook: ML@FPGA + ASIC?

- ML@FPGA also potential as co-processor -> larger ML models in HLT/Reconstr.
- ML@ASIC -> potential to revolutionise HEP experiment design? (cf. HGCAL ASIC)



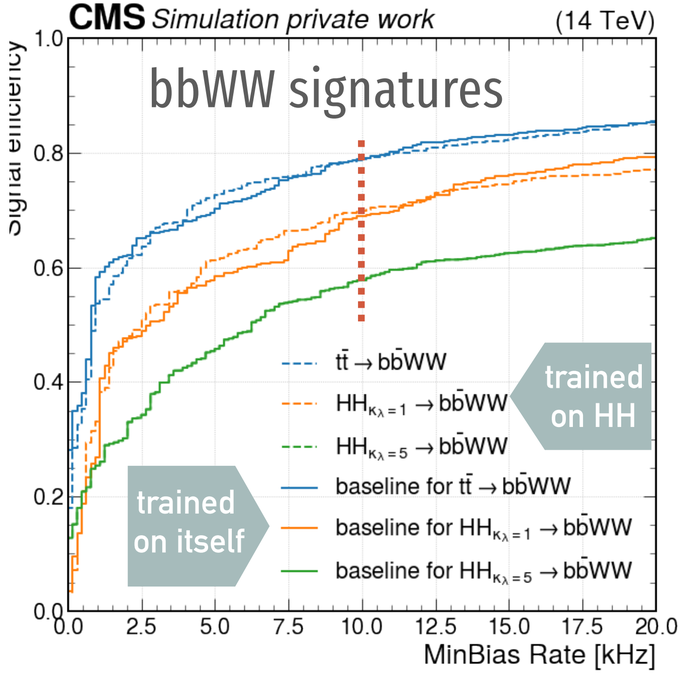
TEAM: ML @ L1-TRIGGER IN CMS

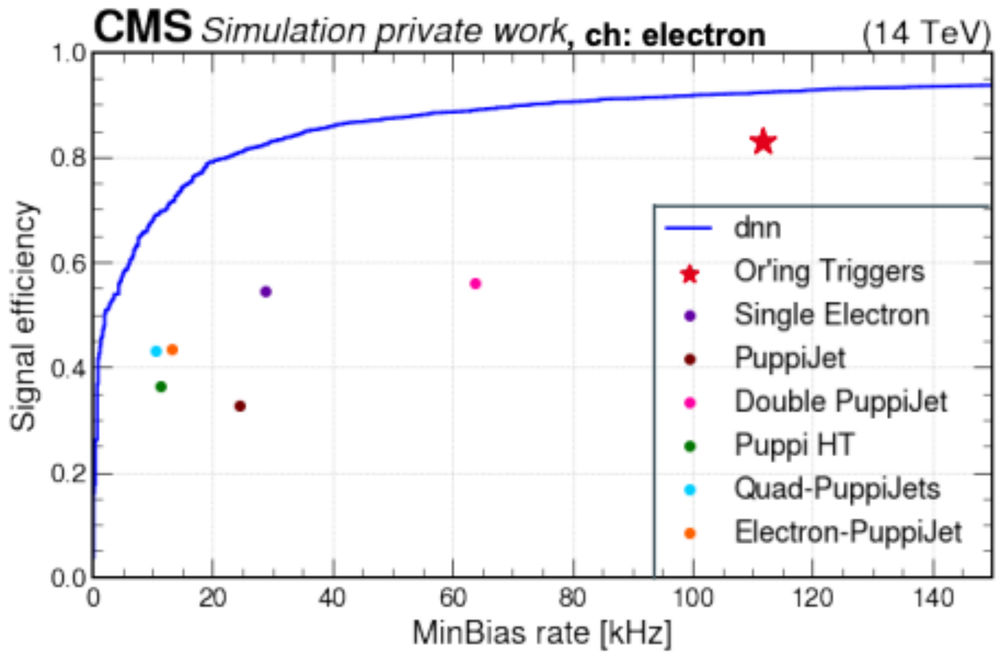
Project lead	Artur Lobanov (postdoc)
PIs	Johanes Haller, Gregor Kasieczka (Prof)
Higgs expert	Matthias Schroeder (Staff)
Topo trigger	Finn Labe (PhD), Ihor Komarov (MSc ✔), Karla Kleinboelting (BSc ✔)
Jet identification	Philipp Rincke (MSc), Karim El-Morabit (pd)
Anomaly detection	Sven Bollweg (PhD), Lars Emmrich (BSc), KEM

BACKUP

Multiple signals with one NN?

- Can one use one NN topo trigger for processes with similar signatures?
  - E.g. HH > bbWW (SM and BSM), ttbar (bWbW), HH > hadronic
- **NN trained on similar processes performs similar to NN trained on the signal itself**
  - Hints that NN largely learns background minbias [—> anomaly detection!]



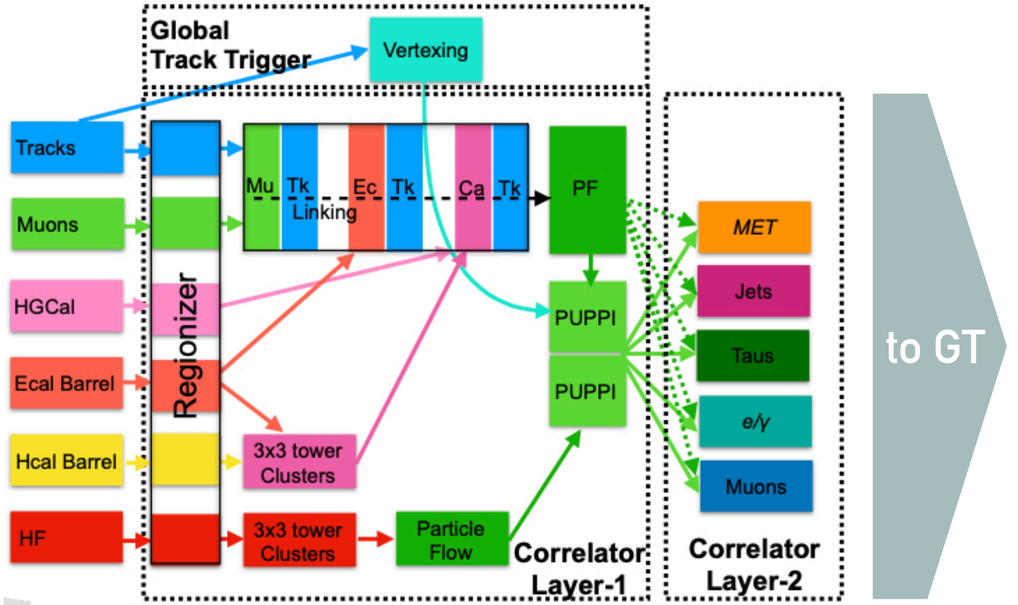


ML trigger
vs standard L1 seeds:
efficiency vs rate

-> ML best

Correlator and Global Trigger

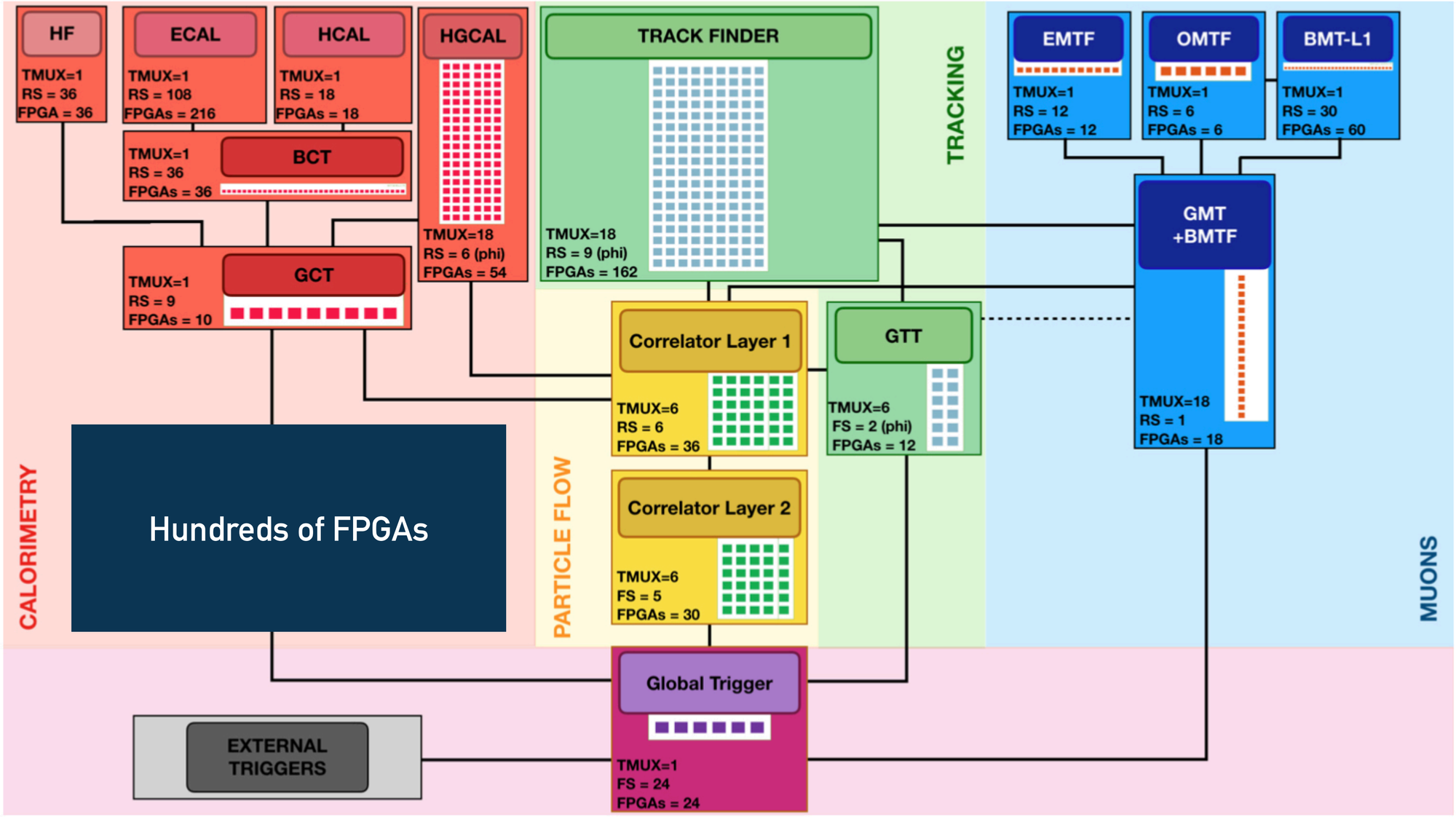
- **Correlator Trigger** (CT, new in L1T)
  - Using **Particle Flow** to reconstruct and identify all particles using all sub-detectors
  - Outputs: e/y/mu/taus/jets and MissingET
  - Latency: ~3 µs (ID: <1 µs)
- **Global Trigger (GT)**
  - Receives objects from all L1T systems
  - Computes correlations or other algorithms
  - Latency: ~1 µs



Correlator Trigger architecture

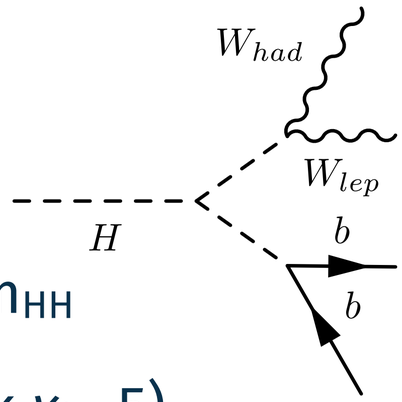
- Powerful FPGAs and increased latency enable the use of complex/expensive algorithms
  - **Bringing Machine Learning to the L1 Trigger!**

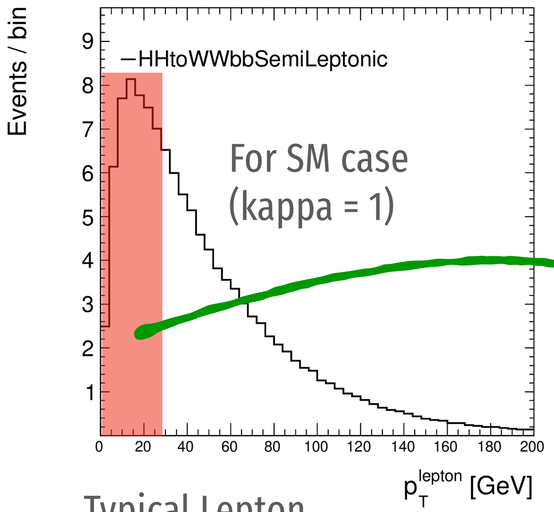
FPGAs: WORKHORSE OF THE CMS L1T



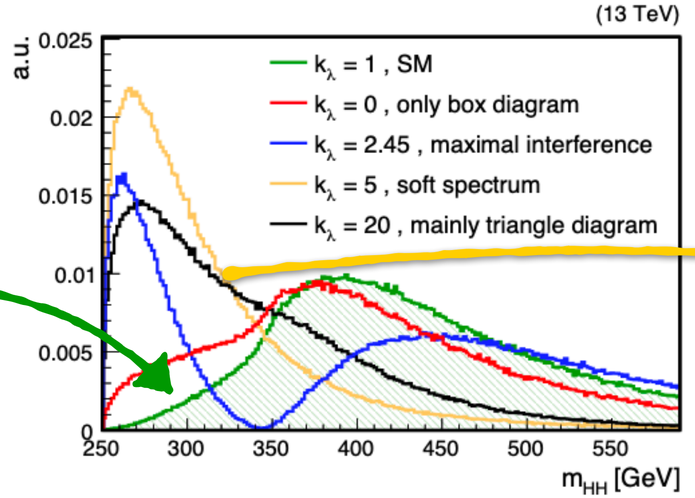
HH FOR HL-LHC

- Two ML L1T algorithms already shown as PoC in L1T TDR for HL-LHC
- Our **target signal: HH** – one of the showcases for HL-LHC
- An indirect handle for the analysis sensitivity is the HH invariant mass: m_{HH}
- Low m_{HH} likely results in **softer objects -> trigger limited region** (see kink κ_λ ~5)

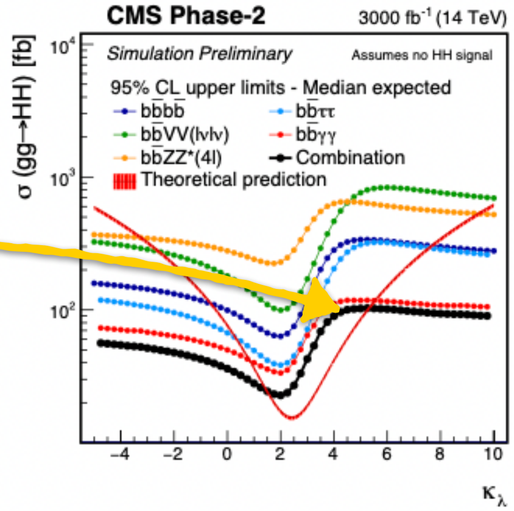




Typical Lepton threshold



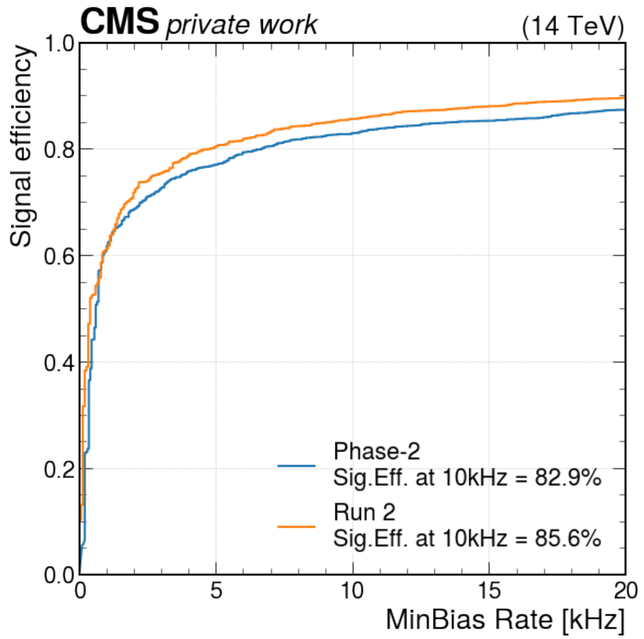
m_{HH} for different kappa_lambda

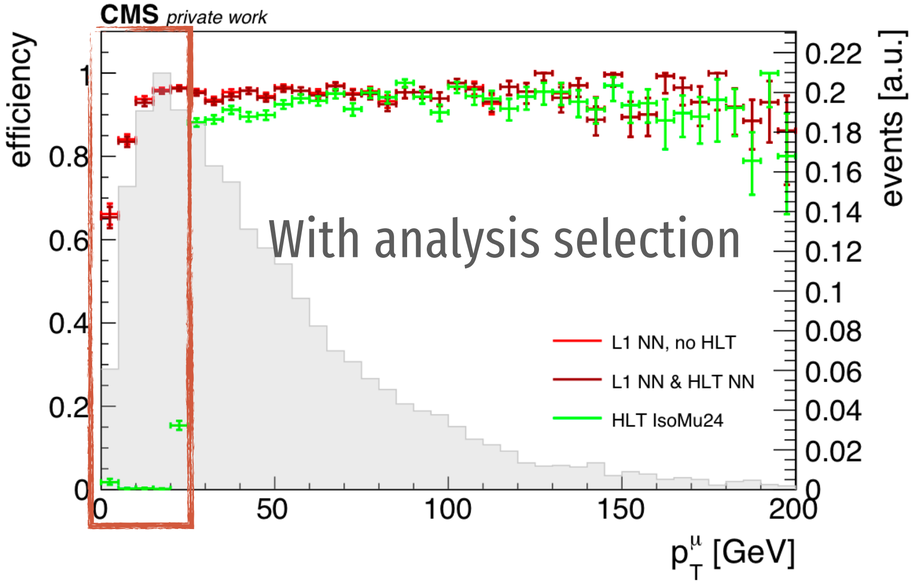


From HL-HLC Yellow Report CMS-FTR-18-019

Towards usage of ML TopoTriggers

- Estimating effect of NN topo trigger @ L1 on analysis [Reusing existing Run-2 setup]
- **NN performance with Run-2 inputs similar to HL-LHC:**
  - Larger PU <> better trigger resolution
  - **Prospect of using ML TopoTrigger for Run3?**



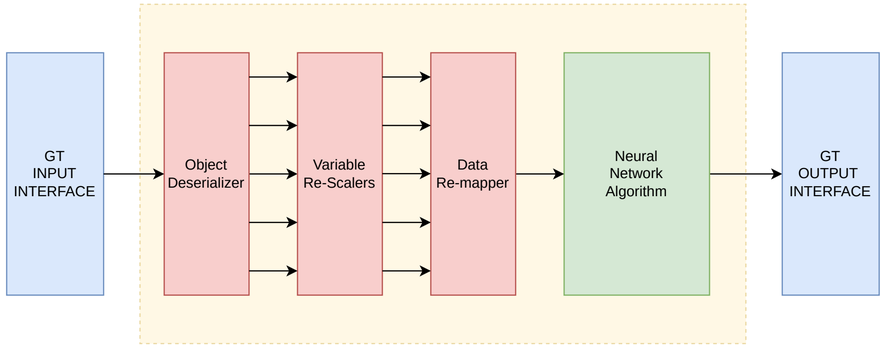


- Next: evaluating L1 trigger efficiency using reco objects as “HLT” proxy
- **Clear added efficiency from replacing L1 lepton seed with NN at HLT/reco**

HARDWARE DEMO

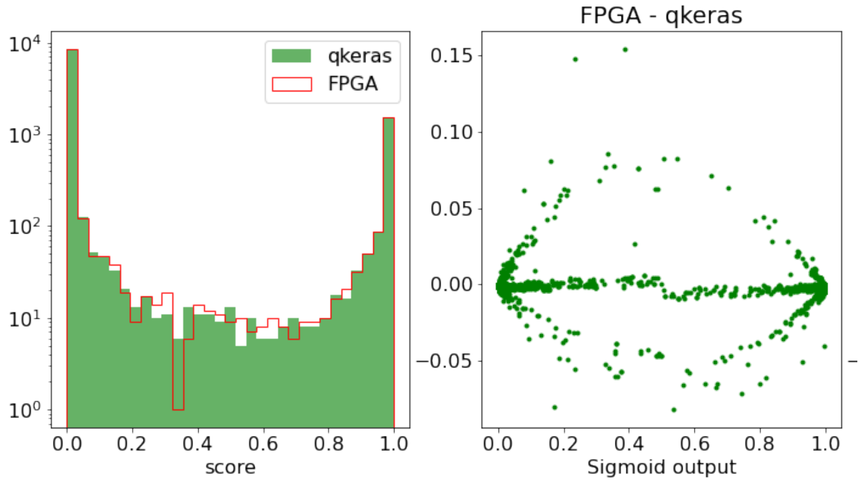
PHASE2 GT INTEGRATION

- Collaboration with P2GT team for integration of NN in demonstrator
- **Gabirele Bortolato implemented the NN algos in the P2GT FW for Serenity** (in EMP-FWK)
- Resource/latency/performance as expected
- **Agreement for (Q)Keras/HLS/FPGA inference of NN trigger algorithm**
  - Towards emulation in CMSSW using HLS4ML?



NN label	LUT	FF	DSP	latency[ns]	Latency[clk]
3 layers	3484	1858	0	33.33	16
2 layers	3059	2046	0	29.17	14
1 layer	3845	2887	13	25.00	12
4 nodes	1444	1308	4	22.92	11

Table 4: 480MHz target clock, implementation numbers



RUN3 GT DEMO

- **First discussions with uGT team about demonstration in Run3 uGT test crate**
  - NN resource usage @40MHz: 1% of Virtex7 / II = 25 ns / latency = 50 ns – OK!
- **Profiting from preparatory work for Anomaly Trigger (synergy!)**
  - **uGT firmware ready to integrate NN algorithm** [Herbert Bergauer]
  - Next steps:
    - Train NN on Run3 samples
    - Integrate NN IP into uGT FW
    - Emulation in CMSSW?
    - Implement FW in test crate
    - TEST rates @P5