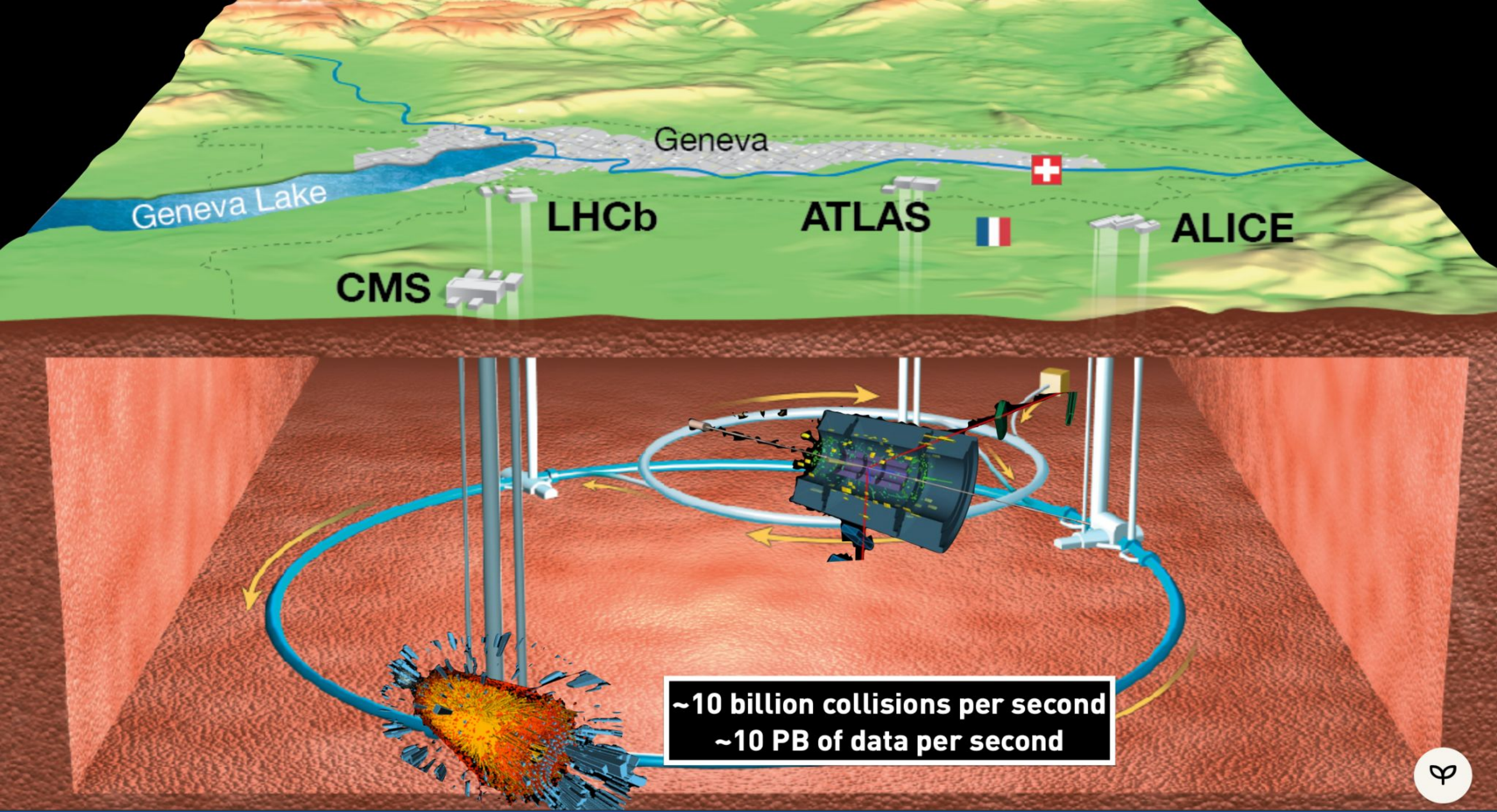
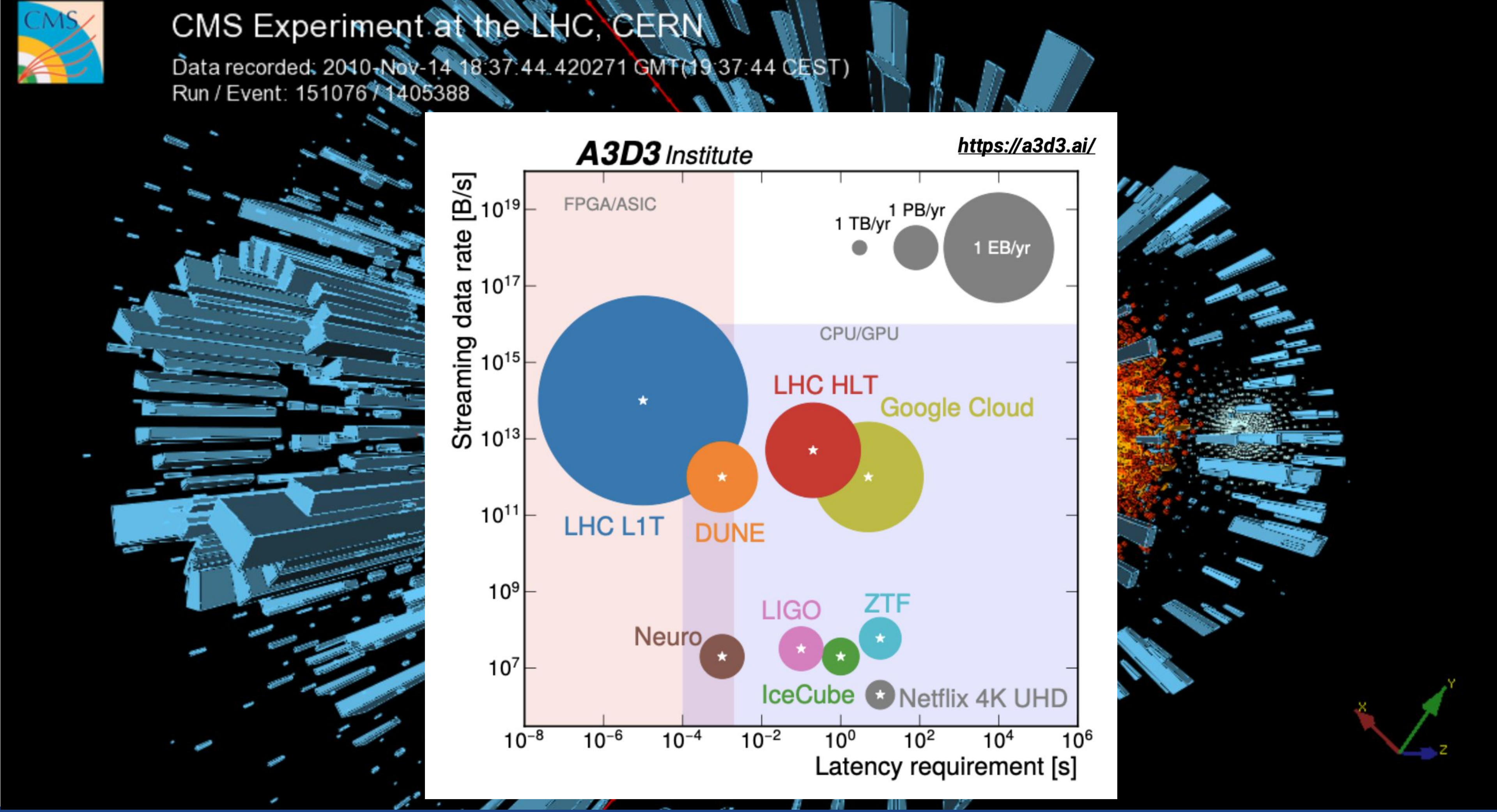
A Convolutional Neural Network for a topological fast selection algorithm of HH → bbbb in FPGAs for the HL-LHC upgrade of the CMS experiment

Maciej Głowacki on behalf of the CMS experiment

7th November 2023

Motivations


Geneva
Geneva Lake
LHCb
ATLAS
ALICE
CMS
~10 billion collisions per second
~10 PB of data per second


CMS Experiment at the LHC, CERN
Data recorded: 2010-Nov-14 18:37:44.420271 GMT(19:37:44 CEST)
Run / Event: 151076 / 1405388
A3D3 Institute
https://a3d3.ai/
Streaming data rate [B/s]
Latency requirement [s]
FPGA/ASIC
CPU/GPU
1 TB/yr
1 PB/yr
1 EB/yr
LHC L1T
DUNE
LHC HLT
Google Cloud
Neuro
LIGO
IceCube
ZTF
Netflix 4K UHD
10^{7}
10^{9}
10^{11}
10^{13}
10^{15}
10^{17}
10^{19}
10^{-8}
10^{-6}
10^{-4}
10^{-2}
10^{0}
10^{2}
10^{4}
10^{6}

LHC

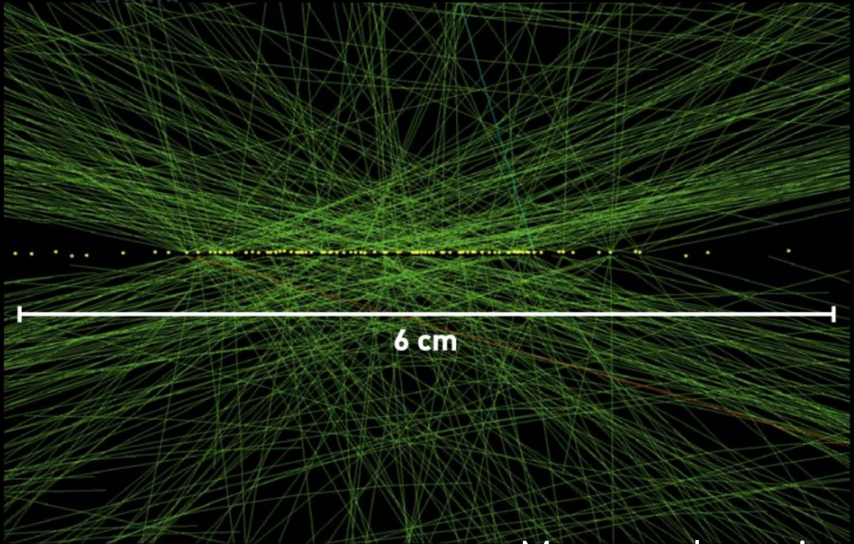
High Luminosity LHC

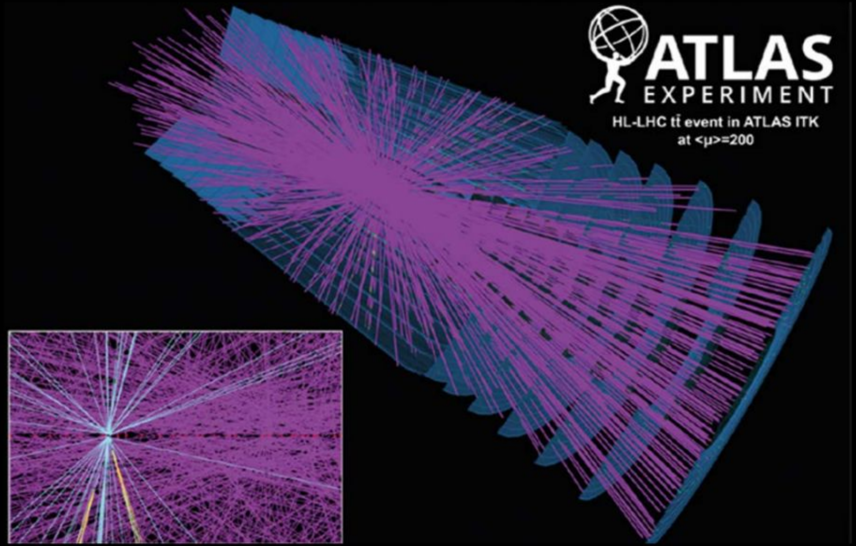
Current

78 vertices
(average 60)

Phase 2

200 vertices
(average 140)



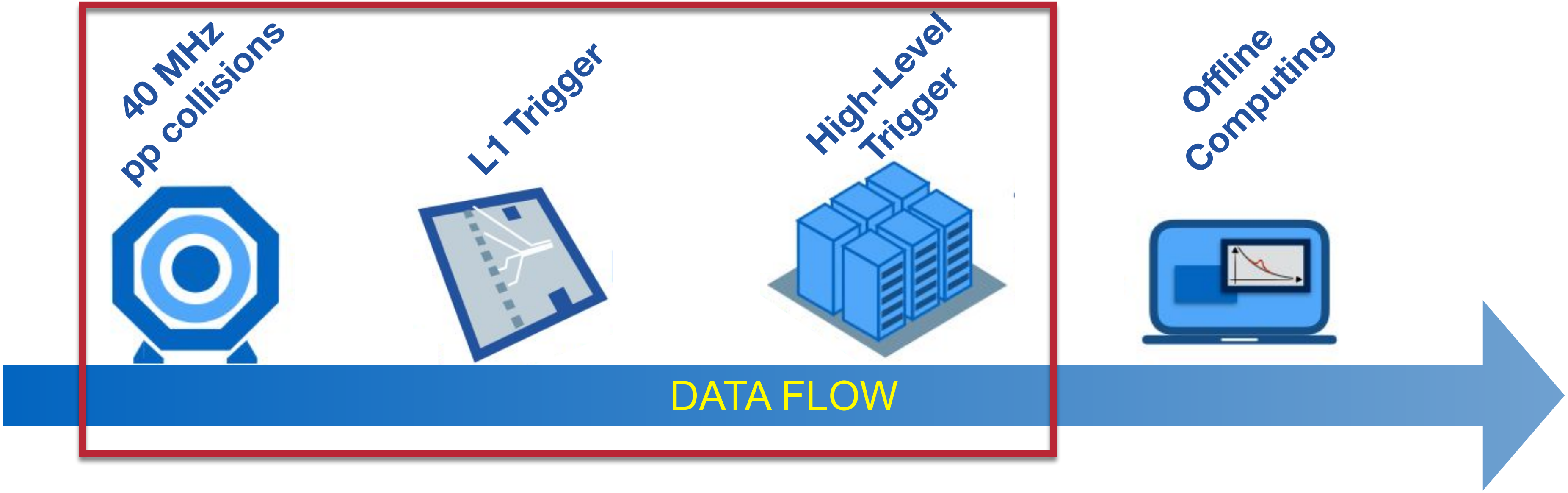


- More complex environment with 200 Pile-Up interactions
- Current approaches not sustainable (acceptance rate would triple)
- Necessity for more sophisticated triggering algorithms
  - **Machine Learning**

Run 3

Run 4+5

Phase 2 Level 1 Trigger



Deploy ML algorithms already at Level 1

Phase 2 Level 1 Trigger

- L1 Phase II Highlights
  - Larger L1 trigger rate / detector readout rate (100 kHz → 750 kHz).
  - Larger L1 trigger latency (3.8 us → 12.5 us) → **more sophisticated algo.**
  - More info at L1 trigger → L1 tracks, higher granularity.
    - Allows for Particle Flow (PF) event reconstruction and PUPPI Pile-Up mitigation.
- Topology targeting trigger already tested in L1 Global Trigger
  - Phase 2 Level-1 Trigger upgrade TDR (4.3.6) showed feasibility of topological Machine Learning triggers, targeting VBF Higgs Boson production with invisible plus dijet final states.
  - Classification performed with reconstructed jet and event-level features is currently being tested during Run 3.
- **This talk** shows the feasibility of topology classification performed on PUPPI candidates in Correlator Level 2 (CTL2).
  - I.e. **not** jet clustering but event-level classification.
- Transition from heuristic, rules based code, written in the classical stack, to Neural Networks (“Software 2.0”).

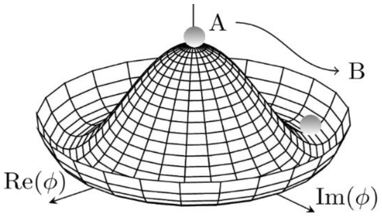
HH → 4b

$$\mathcal{L}_{scalar} = D_\mu\phi^\dagger D^\mu\phi - V(\phi^\dagger\phi) \text{ with } \phi = (\varphi^+ \ \varphi^0)^T \text{ doublet under } SU(2)$$

$$V(\phi^\dagger\phi) = -\mu^2(\phi^\dagger\phi) + \lambda(\phi^\dagger\phi)^2$$

$$\phi(x) = \frac{1}{\sqrt{2}} \exp\left(i\sigma^i\xi(x)\right)\begin{pmatrix} 0 \\ v + h(x)\end{pmatrix} \xrightarrow{EWSB} \phi(x) = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v + h(x)\end{pmatrix}$$

Non-zero vacuum expectation value $v = \mu^2/\lambda$



$\sigma(pp \to HH) \simeq \dfrac{\sigma(pp \to H)}{1000}$

If SM is correct :

→ 4000 HH events during Run-2

... not enough to see HH

$$\mathcal{L}_{scalar} = D_\mu\phi^\dagger D^\mu\phi + \mu^2(\phi^\dagger\phi) - \lambda(\phi^\dagger\phi)^2$$

$$= \frac{v^2}{8}\left(g^2 W^i_\mu W^{i\mu} + g'^2 B_\mu B_\nu - 2g'gB_\mu W^{3\mu}\right)\left(1 + \frac{h}{v}\right)^2$$

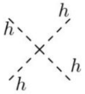
$$+ \underbrace{\frac{1}{2}(\partial_\mu h \partial^\mu h)}_{\text{kinetic term}} - \underbrace{\lambda v^2 h^2}_{\text{mass term}} - \underbrace{\lambda v h^3}_{\text{trilinear coupling}} - \underbrace{\frac{\lambda}{4}h^4}_{\text{quartic coupling}} - \frac{\lambda v^4}{4}$$

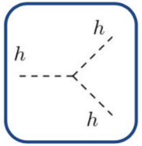
$$m_H = \sqrt{2\lambda}v$$

Mass of the weak bosons

\+

Mass of the fermions through Yukawa couplings

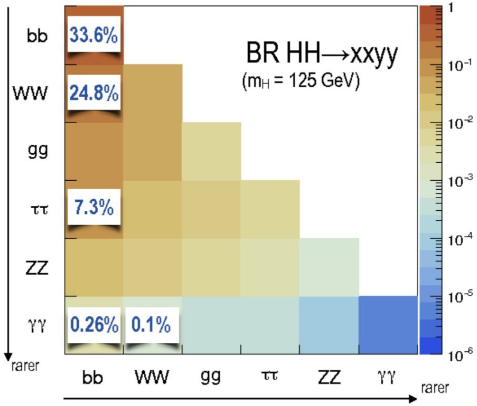




(on our menu today)

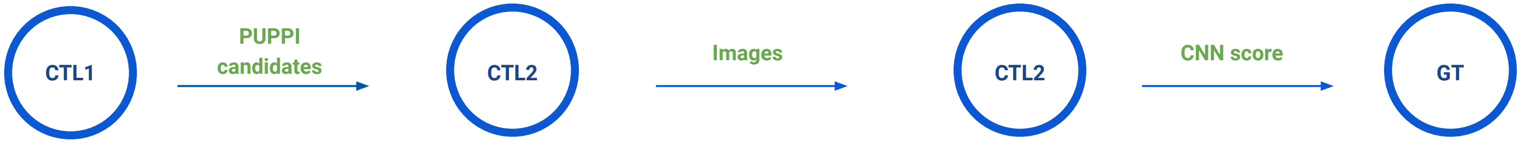
Fully parameterized by **λ**

- Theory value given by v and m_H
- Experimental measurement
  → Test of the SM
  → Probe the shape of the potential
  → Very sensitive to BSM



CNN based topological trigger

Dataflow



Particle Flow reconstruction and pileup mitigation

1. Calorimeter and track information used in PF reconstruction and PUPPI pile-up mitigation.

Histogramming & preprocessing

2. PUPPI candidate P_T binned in the η - ϕ space of the detector to produce 2-D *images* used concurrently by topological classifier and Jet Finding Algorithm [1]. Preprocessing is applied to refine images to serve as input classifier.

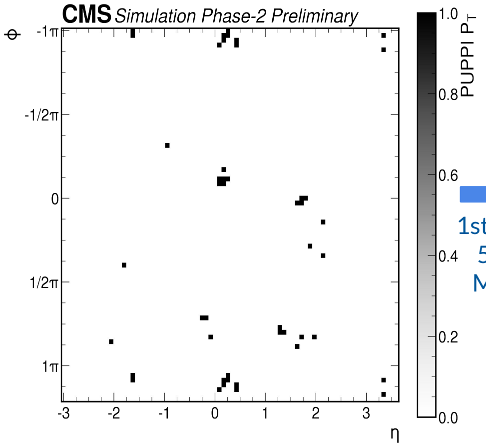
Classification

3. Convolutional Neural Network (CNN) executes its inference procedure from input images.

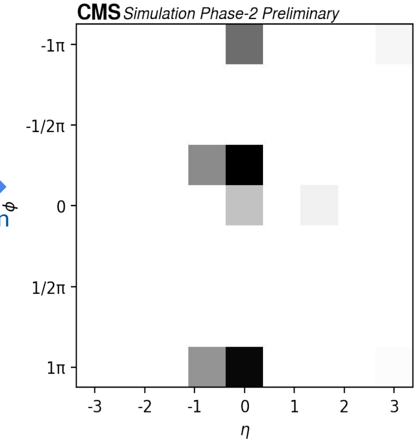
Global Trigger input

4. CNN probability score delivered to GT to be used alongside existing menu bits.

Architecture

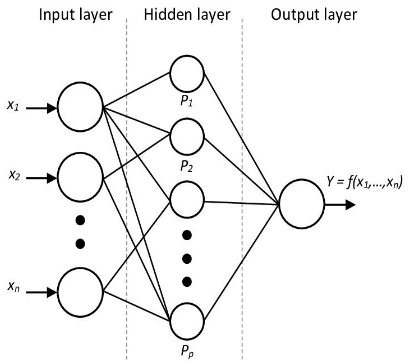


1st convolution
5x5 kernel*
Max Pooling



2nd convolution
2x2 kernel
Flatten

Augment features with HT

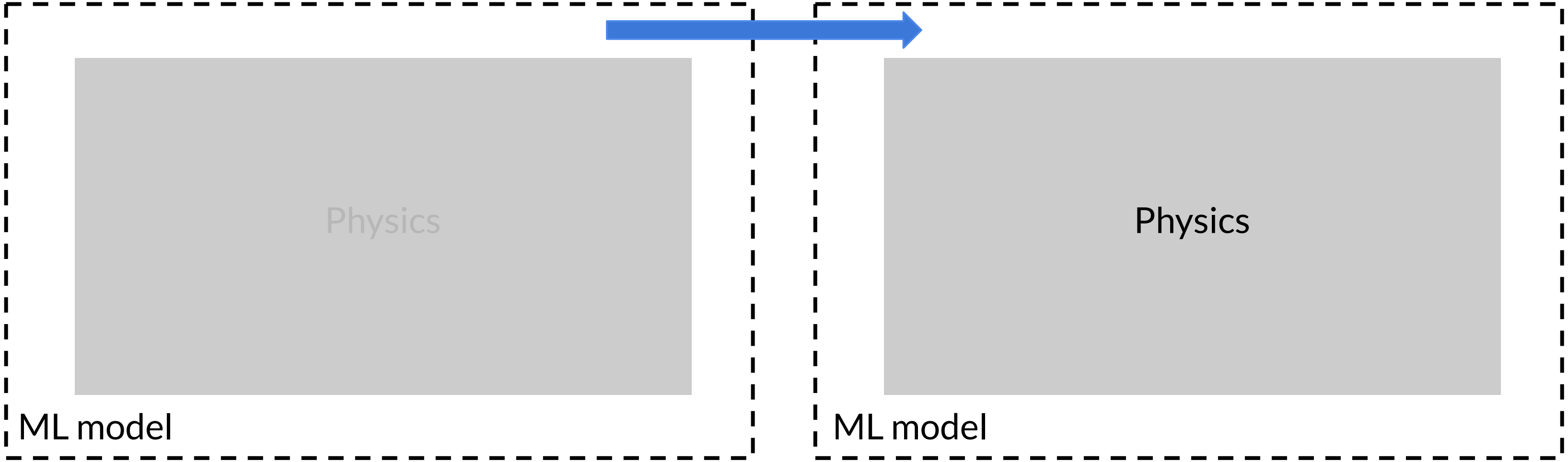


Augment features with reconstructed L1 quantities or regress them during inference procedure.

*Kernel regularisation $\propto \sum(\text{weights}_{\text{kernel}})^2$

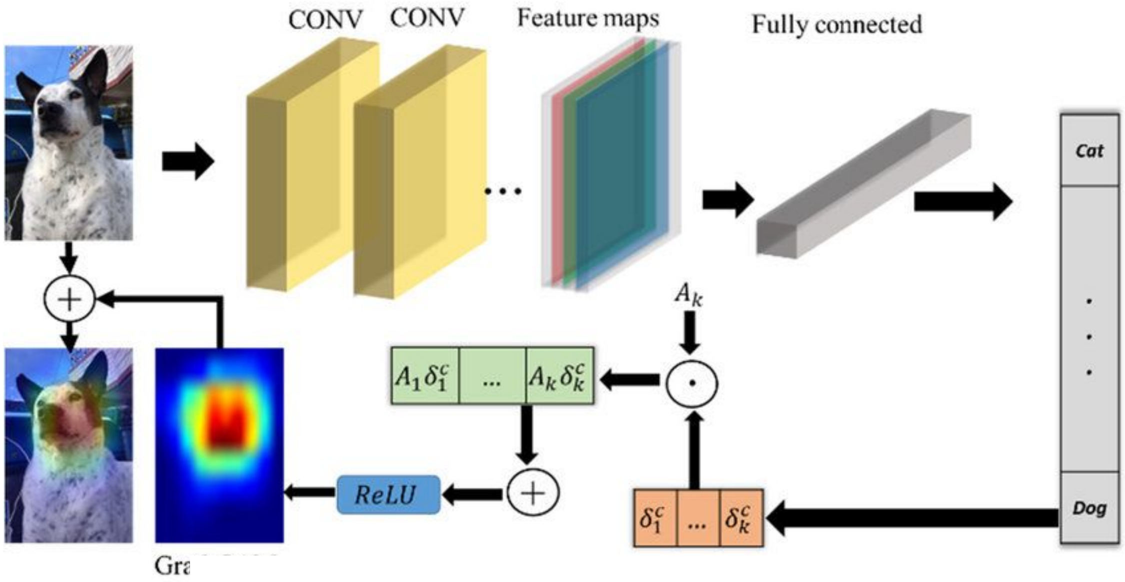
Explainability

- Make physics transparent.
- Condition model on/regress L1T reconstructed quantities.
- Guide model architecture by physics,
    - improve understanding of feature extraction layers.



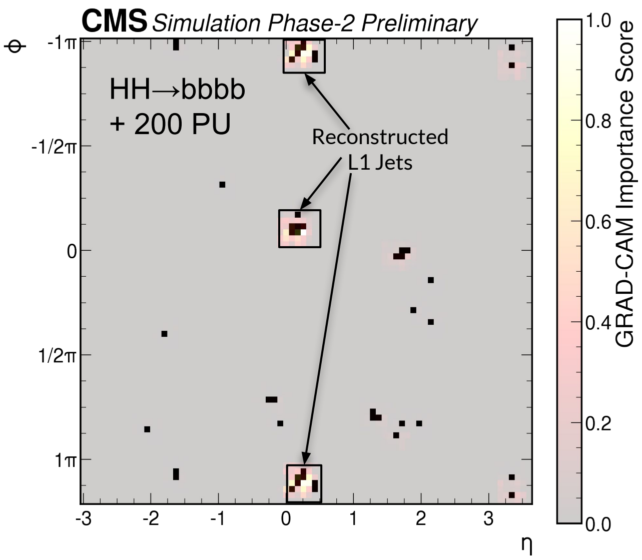
GradCam

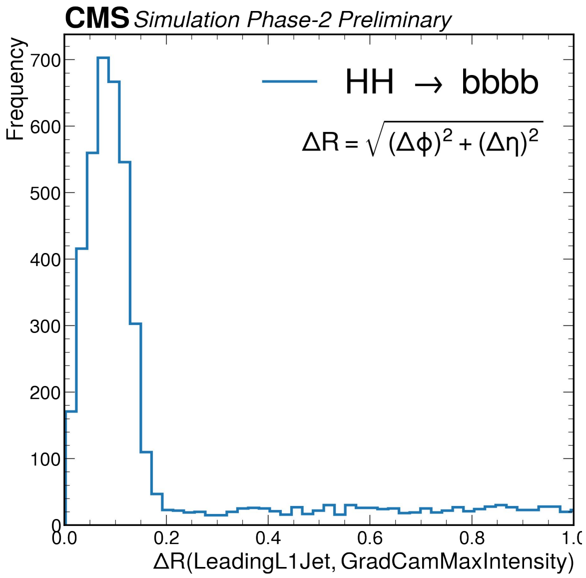
- Pixel-wise Importance scores computed from backpropagated gradients of class scores w.r.t. final feature map activations.
- Gradients are then averaged, producing a weight for each feature map point that represents its importance in the classification decision.
- Visualises areas of high importance in the input on the final classification.

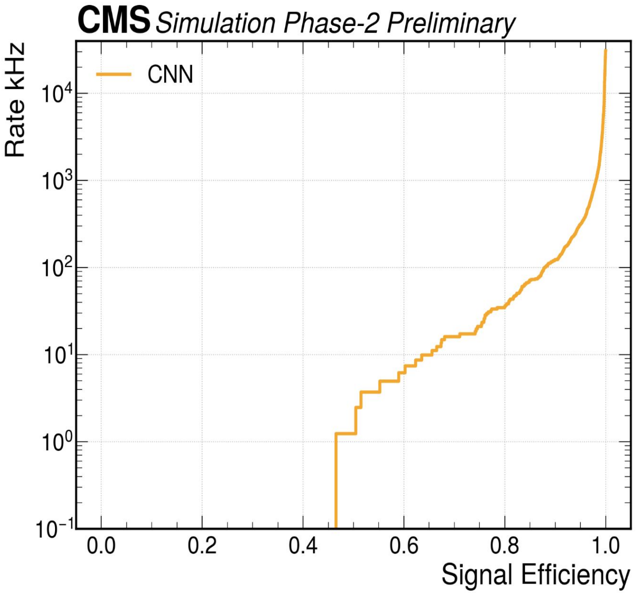


$$\text{Grad-CAM}_c(x, i, j) = \text{ReLU}\left(\sum_k \frac{\partial Y^c}{\partial A_{i,j}^k} A_{i,j}^k\right)$$ [2]

Physics Performance



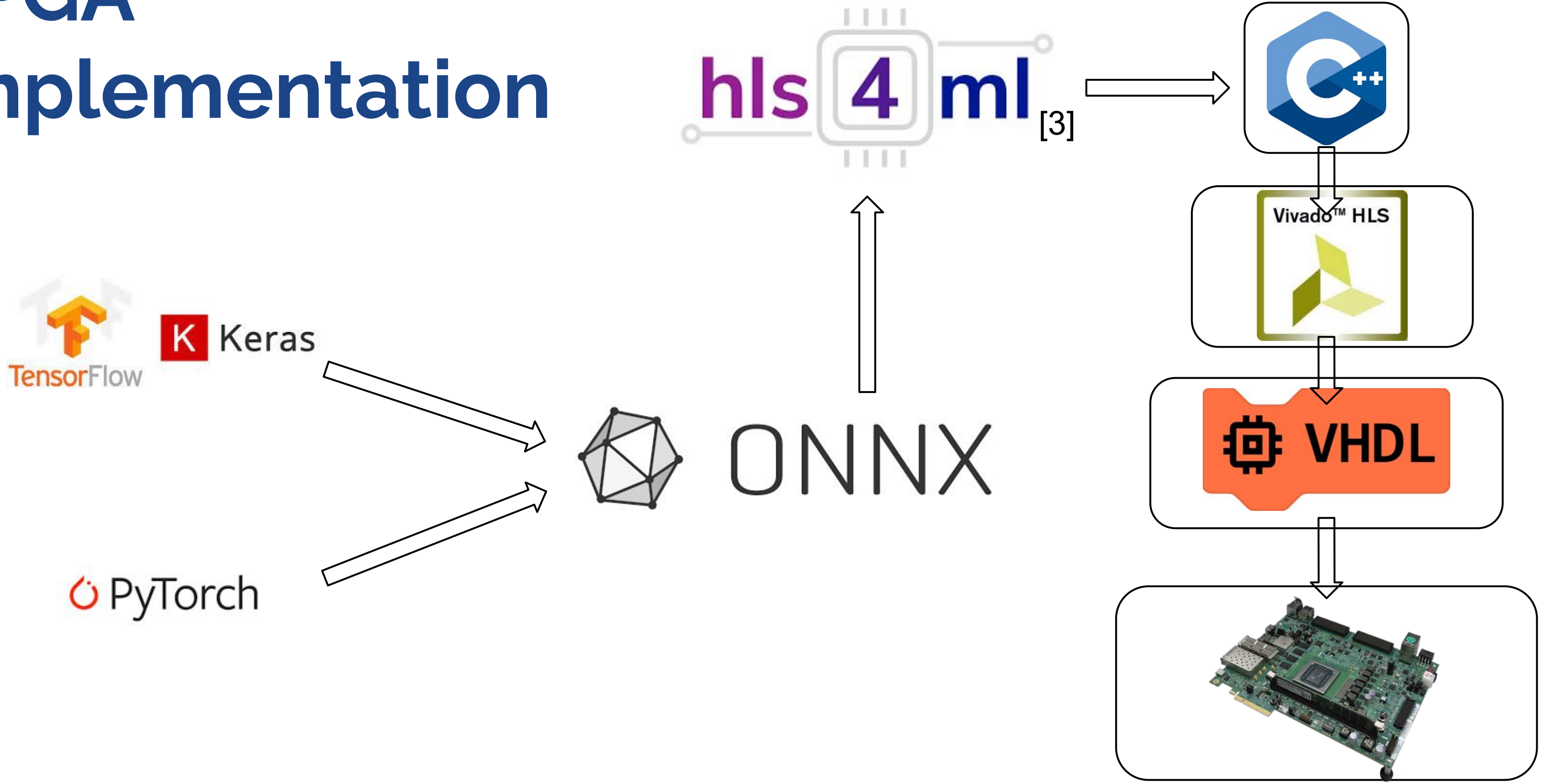




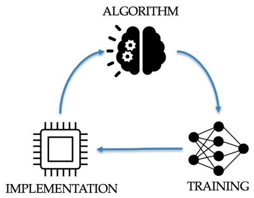
- Grad-CAM scores indicate model is learning to cluster PUPPI candidates into jets.
  - Good agreement seen between highest Grad-CAM intensity and leading Level 1 reconstructed jet.
- 65% efficiency at 10kHz total rate achieved (rate equivalent to existing QuadJetHT L1T path).

Machine Learning on the edge

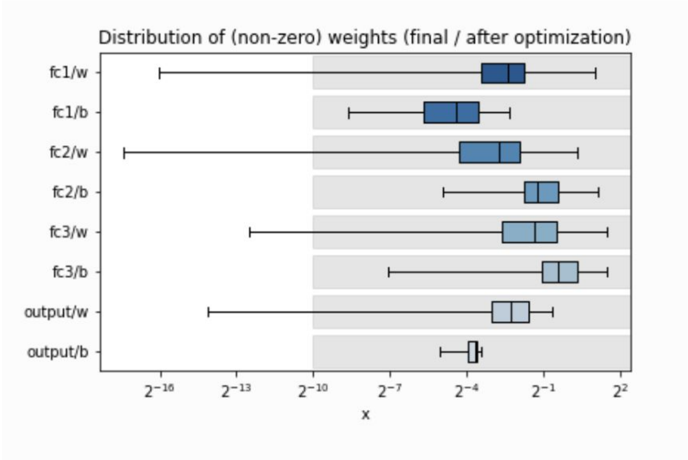
FPGA Implementation



Co-design

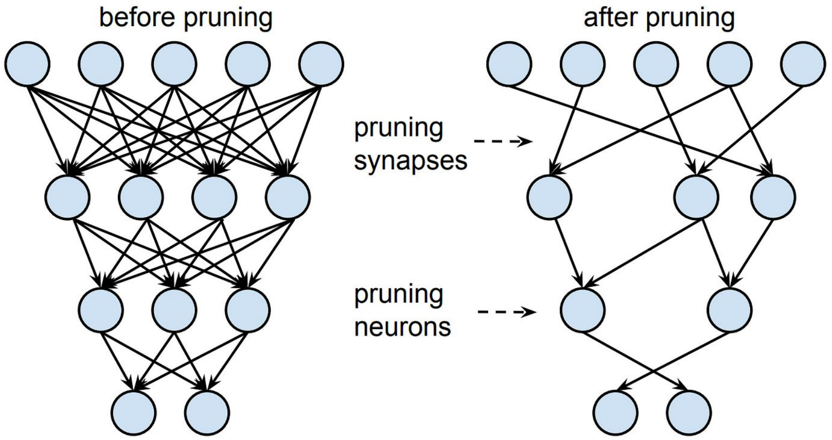


Quantisation



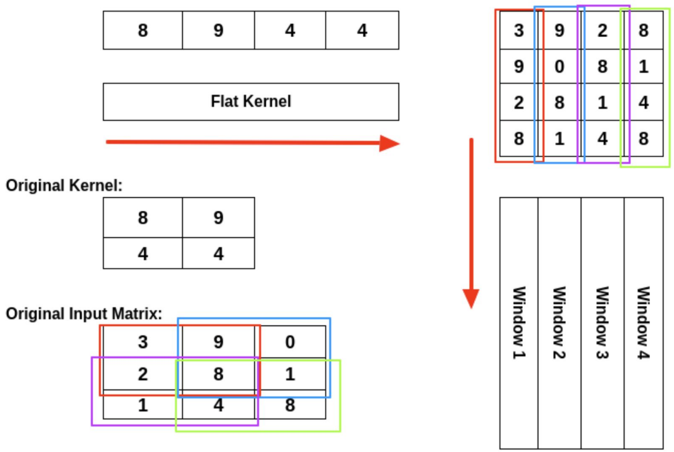
QKeras

Pruning



Resource & Latency usage

- Im2col [4] algorithm used to increase parallelism on each clock cycle
  - Meets initiation and latency requirement of the CTL2 subsystem.



Target Device

Part	xcvu9p-flga2577-2-e
Target clock	2.78 ns
Parallelisation / Reuse Factor	18/1

Resource Usage

	SLR usage	Device usage
LUT	10.6%	3.5%
FF	7.0%	2.3%
DSP	10.5%	3.5%
BRAM	<0.1%	<0.1%

Latency

Clock frequency	360 MHz
Latency	269 ns
Initiation Interval	117 ns

Summary

- Conditions for the CMS L1T during Phase 2:
  - 200 Pile-Up interactions.
  - Reduced latency constraints.
- Addition of tracking information at L1:
  - Full Particle Flow reconstruction.
  - PUPPI Pile-Up mitigation.
- Machine Learning approaches viable:
  - CNN-based topological trigger performs well (efficiency vs rate).
  - Meets latency constraints while maintaining small FPGA footprint.
- Future iterations:
  - Improved model training pipeline with hard negative mining.
  - Location-dependent kernel weights (mirror detector structure with different learnable weights for barrel, endcap, forward regions).
  - Test interface with HLT.

References

[1] **"The Phase-2 Upgrade of the CMS Level-1 Trigger"**, CERN-LHCC-2020-004 ; CMS-TDR-021

[2] **"Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization"**, 10.1007/s11263-019-01228-7

[3] **"Fast inference of deep neural networks in FPGAs for particle physics"**, 10.1088/1748-0221/13/07/p07027, https://github.com/fastmachinelearning/hls4ml

[4] **"Fast convolutional neural networks on FPGAs with hls4ml"**, 10.1088/2632-2153/ac0ea1

Back up

Phase II trigger menus

- Quad jet and HT requirements (reconstructed jets and summed)
- ~50-60% efficiency at ~10 kHz rate

L1 Trigger seeds	Offline Threshold(s) at 90% or 95% (50%) [GeV]	Online Threshold(s) (Barrel) [kHz]	Rate* $\langle PU \rangle = 200$ [kHz]	Additional Requirement(s) [cm, GeV]	Objects plateau efficiency [%]
Single/Double/Triple Lepton (electron, muon) seeds					
Single TkMuon	22	20	12	$\lvert\eta\rvert < 2.4$	95
Double TkMuon	15,7	13,6	1	$\lvert\eta\rvert < 2.4$, $\Delta z < 1$	95
Triple TkMuon	5,3,3	4,2,2	16	$\lvert\eta\rvert < 2.4$, $\Delta z < 1$	95
Single TkElectron	36	32	24	$\lvert\eta\rvert < 2.4$	93
Single TkIsoElectron	28	25	28	$\lvert\eta\rvert < 2.4$	93
TkIsoElectron-StaEG	22, 12	19,8	36	$\lvert\eta\rvert < 2.4$	93, 99
Double TkElectron	25, 12	22,10	4	$\lvert\eta\rvert < 2.4$, $\Delta z < 1$	93
Single StaEG	51	46	25	$\lvert\eta\rvert < 2.4$	99
Double StaEG	37,24	32,20	5	$\lvert\eta\rvert < 2.4$	99
Photon seeds					
Single TkIsoPhoton	36	33	43	$\lvert\eta\rvert < 2.4$	97
Double TkIsoPhoton	22, 12	19,9	50	$\lvert\eta\rvert < 2.4$	97
Taus seeds					
Single CaloTau	150(119)	109	21	$\lvert\eta\rvert < 2.1$	99
Double CaloTau	90,90(69,69)	65,65	25	$\lvert\eta\rvert < 2.1$, $\Delta R > 0.5$	99
Double PuppiTau	52,52(36,36)	36,36	7	$\lvert\eta\rvert < 2.1$, $\Delta R > 0.5$	90
Hadronic seeds (jets,H_T)					
Single PuppiJet	180	121	70	$\lvert\eta\rvert < 2.4$	100
Double PuppiJet	112,112	72,72	71	$\lvert\eta\rvert < 2.4$, $\Delta\eta < 1.6$	100
PuppiH_T	450(377)	363	11	jets: $\lvert\eta\rvert < 2.4$, $p_T > 30$	100
QuadPuppiJets-PuppiH_T	70,55,40,40,400(328)	41,30,19,19,316	9	jets: $\lvert\eta\rvert < 2.4$, $p_T > 30$ safety online cut $p_T > 25$ for jets	100,100

Path	Inclusive acceptance	Loosely presel. evts. acceptance	YR presel. evts. acceptance
QuadJet_70_55_40_40	59%	85%	99%
QuadJet_70_55_40_40_HT320	50%	76%	91%
QuadJet_40_40_40_40_MuJet40	23%	36%	44%
QuadJet_40_40_40_40_MuJet40_HT250	22%	35%	43%
QuadJet_70_55_40_40_HT320 OR QuadJet_40_40_40_40_MuJet40_HT250	52%	79%	94%

Datasets (DAS)

MinBias:

/MinBias_TuneCP5_14TeV-pythia8/Phase2HLTTDRWinter20DIGI-PU200_110X_mcRun4_realistic_v3-v3/GEN-SIM-DIGI-RAW

HH→ bbbb:

/GluGluToHHTo4B_node_SM_TuneCP5_14TeV-madgraph_pythia8/Phase2HLTTDRWinter20DIGI-PU200_110X_mcRun4_realistic_v3-v5/GEN-SIM-DIGI-RAW