Neural Network Based Primary Vertex Reconstruction with FPGAs for the Upgrade of the CMS Level-1 Trigger System

C. Brown[1], A. Bundock[3], M. Komm[2], V. Loncar[2], M. Pierini[2], B. Radburn-Smith[1], S. Summers[2], A. Tapper[1] on behalf of the CMS collaboration

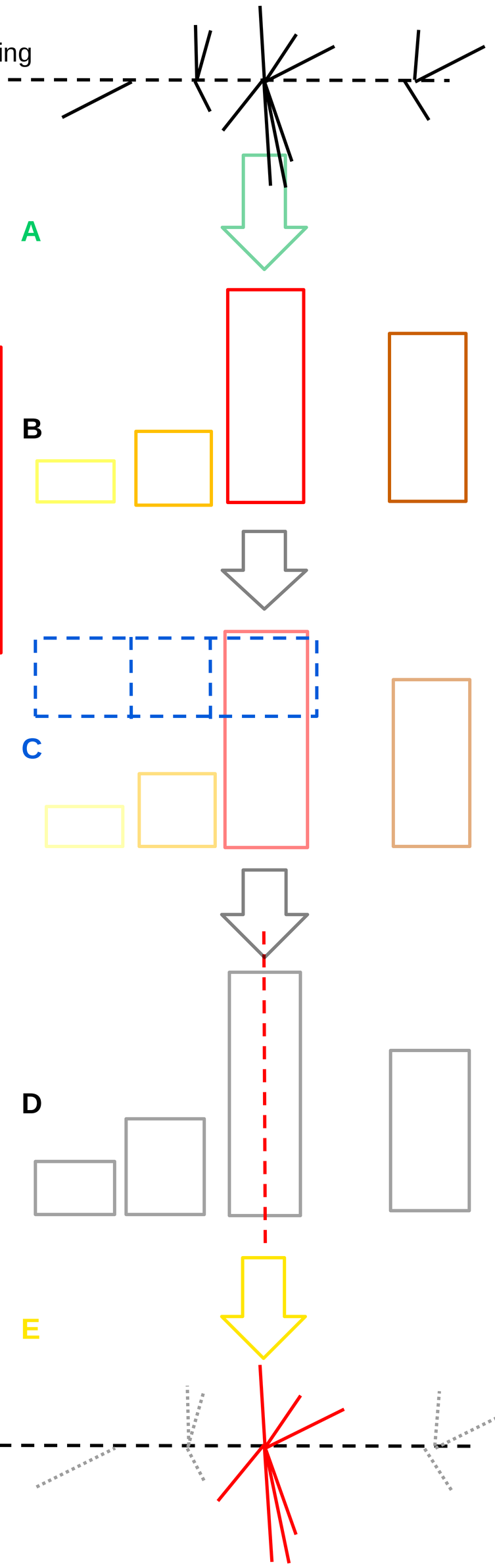
1. Imperial College London, 2. CERN, 3. University of Bristol

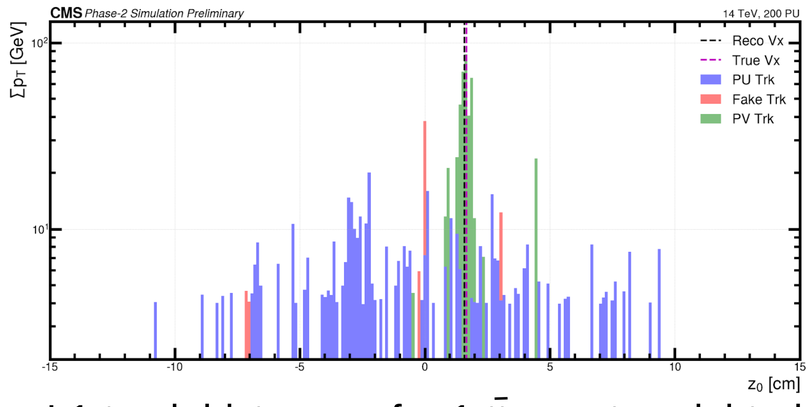
Hi-Lumi LHC [1]
- 5-7x inst lumi
- 10x int lumi
- 200 PU

L1 trigger
- FPGA architecture
- < 12.5 us latency
- L1 track finding

Track finding [2]
- pT > 2 GeV
- Track Isolation & Matching
- Global track algorithms

Primary Vertex (PV)
- Hard scatter
- Associated objects downstream
- < 500 ns latency.

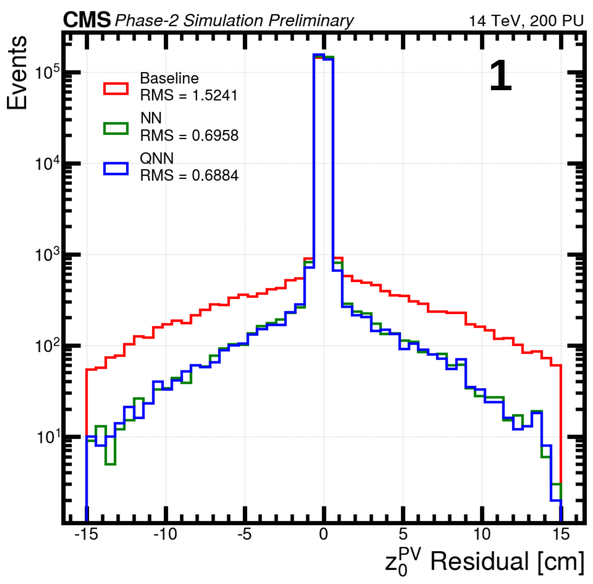


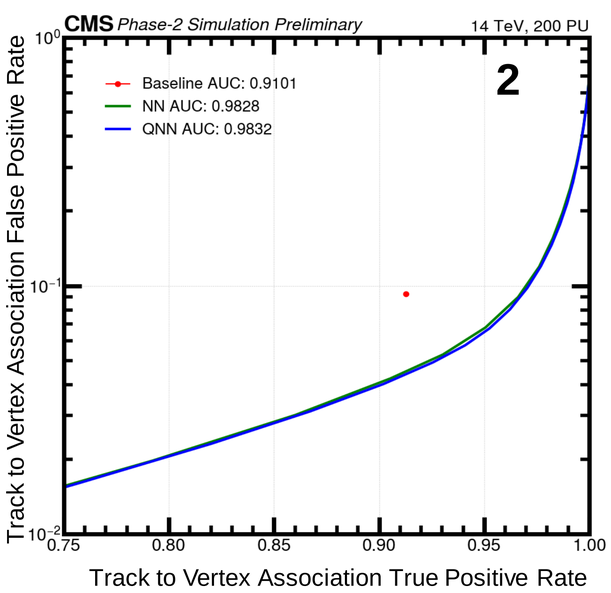


L1 track histogram for 1 $t\bar{t}$ event, weighted by p_T. True PV and baseline reco vertex shown

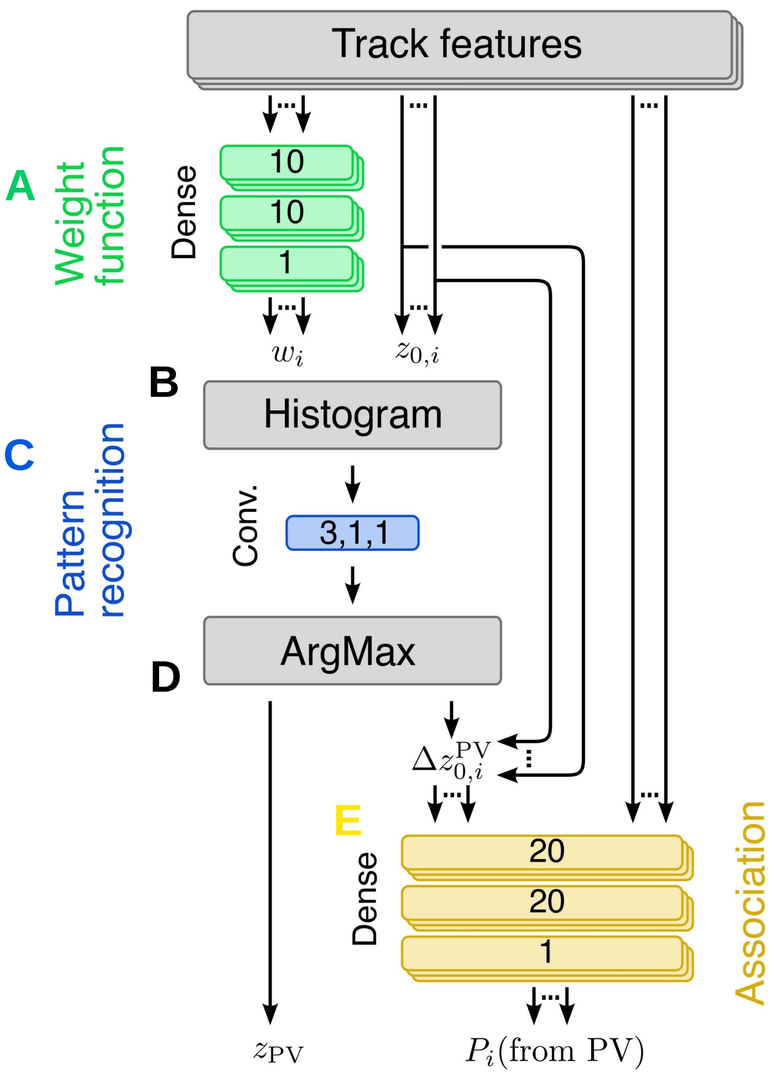
Baseline Approach

A. Weight tracks $\mathbf{p_T}$
B. 256 bin histogram
C. 3 consecutive bins with highest weight
D. Argmax to find peak
E. Δz threshold for assocation depends on **track η**

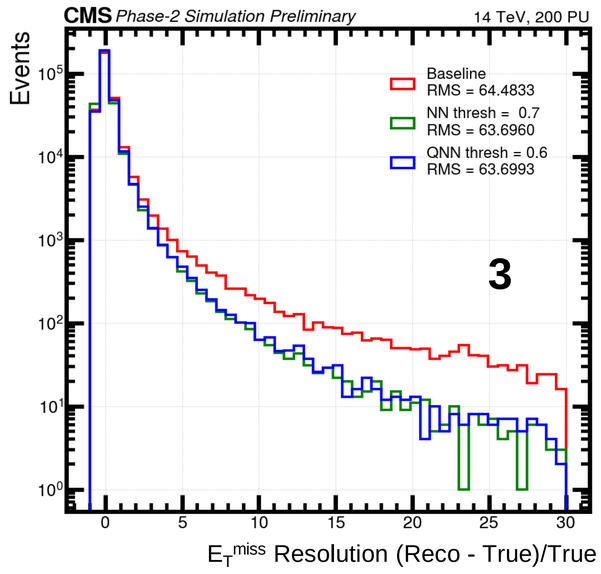




End-to-end Network



2-part loss function, z_{PV}, P(from PV).
Differentiable argmax peak finding
Follows structure of baseline approach; learned **histogram weighting**, learned **pattern recognition**, learned **track-to-vertex association**
Robust to track changes:
fake filtering & **resolution learning**.



Performance

1. PV residual, **long tails reduced** by NN
2. Track-to-vertex ROC, **better discrimination**
3. Track E_T^{miss}, vector sum of PV track p_T, **reduced tails** in NN

Implementation

	Latency (ns)	Initiation Interval (ns)	LUTs %	DSPs %	BRAMs %	FFs %
NN Weight	22	2.7	0.14	1.11	0.00	0.04
QNN Weight	**14**	**2.7**	**0.05**	**0.00**	**0.00**	**0.02**
NN Pattern	58	51	4.27	3.74	5.28	3.22
QNN Pattern	**42**	**35**	**4.43**	**0.00**	**5.28**	**3.15**
NN Assoc.	30	2.7	0.63	5.98	0.00	0.15
QNN Assoc.	**25**	**2.7**	**0.44**	**0.83**	**0.00**	**0.13**

L1 regularisation and pruning
QKeras [3], tuned quantisation.
Converted to FW blocks using **hls4ml** [4]
Implement in existing vertex finding FW.

[1] I. Béjar Alonso *et al, (*Eds). *High-Luminosity Large Hadron Collider (HL-LHC): Technical Design Report V.10* CERN Yellow Reports: Monographs, 2020.

[2] The CMS Collaboration, "The Phase-2 Upgrade of the CMS Level-1 Trigger Technical Design Report," **CMS-TDR-019-002,** Jun 2020

[3] Coelho, C.N., Kuusela, A., Li, S. et al. Automatic heterogeneous quantization of deep neural networks for low-latency inference on the edge for particle detectors. **Nat Mach Intell 3,** 675–686 (2021)

[4] J. Duarte *et al,* "Fast Inference of Deep Neural Networks in FPGAs for Particle Physics", **JINST 13** P07027, 2018