Neural network-based event selection on FPGA for COMET Phase-I

Masaki Miyataki

Osaka University

計測システム研究会2024@東大

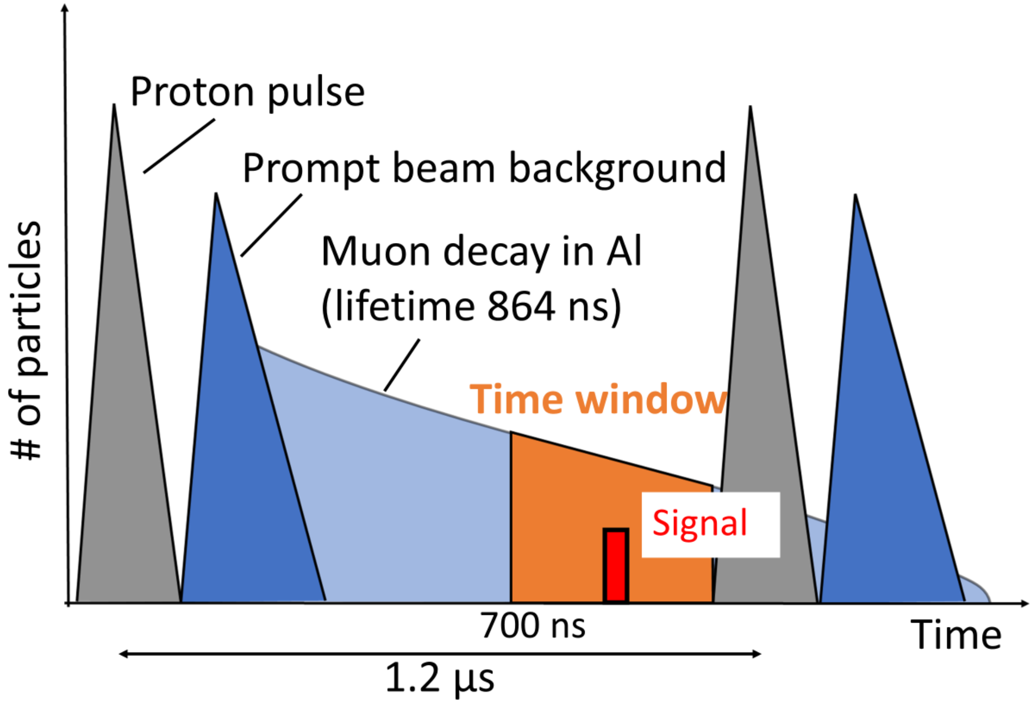
COMET Phase-I @J-PARC

[1] DOI: 10.1093/ptep/ptz125
[2] DOI: 10.1140/epjc/s2006-02582-x

Purpose : **Investigate new physics by searching for charged lepton flavor violating process**

- **μ-e conversion in an Al target**
  - $\mu^- + N(A,Z) \rightarrow e^- + N(A,Z)$
  - Signal : monoenergetic 105 MeV electron
  - Single event sensitivity: 3×10^{-15} [1]

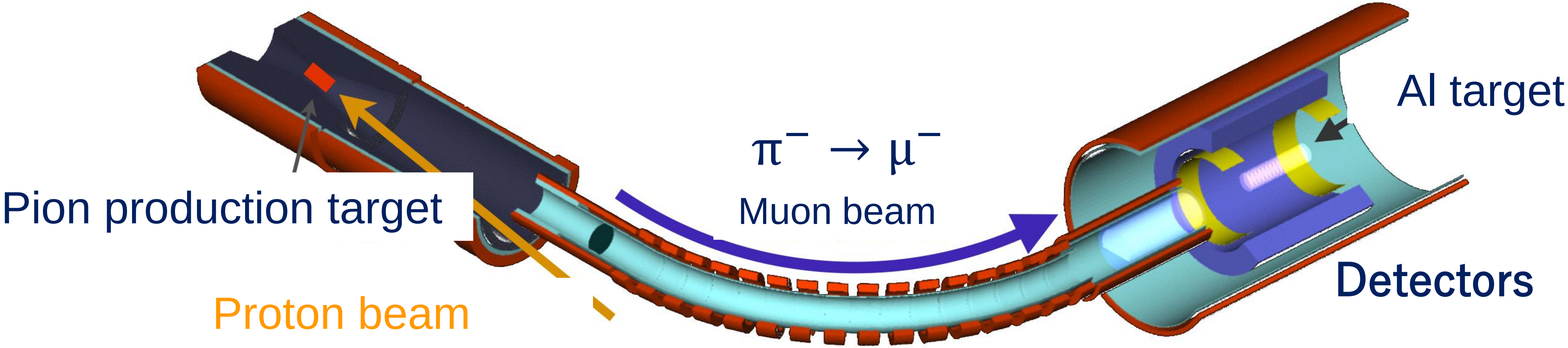
    ⇔100 times better than the current limit [2]



Beam Structure

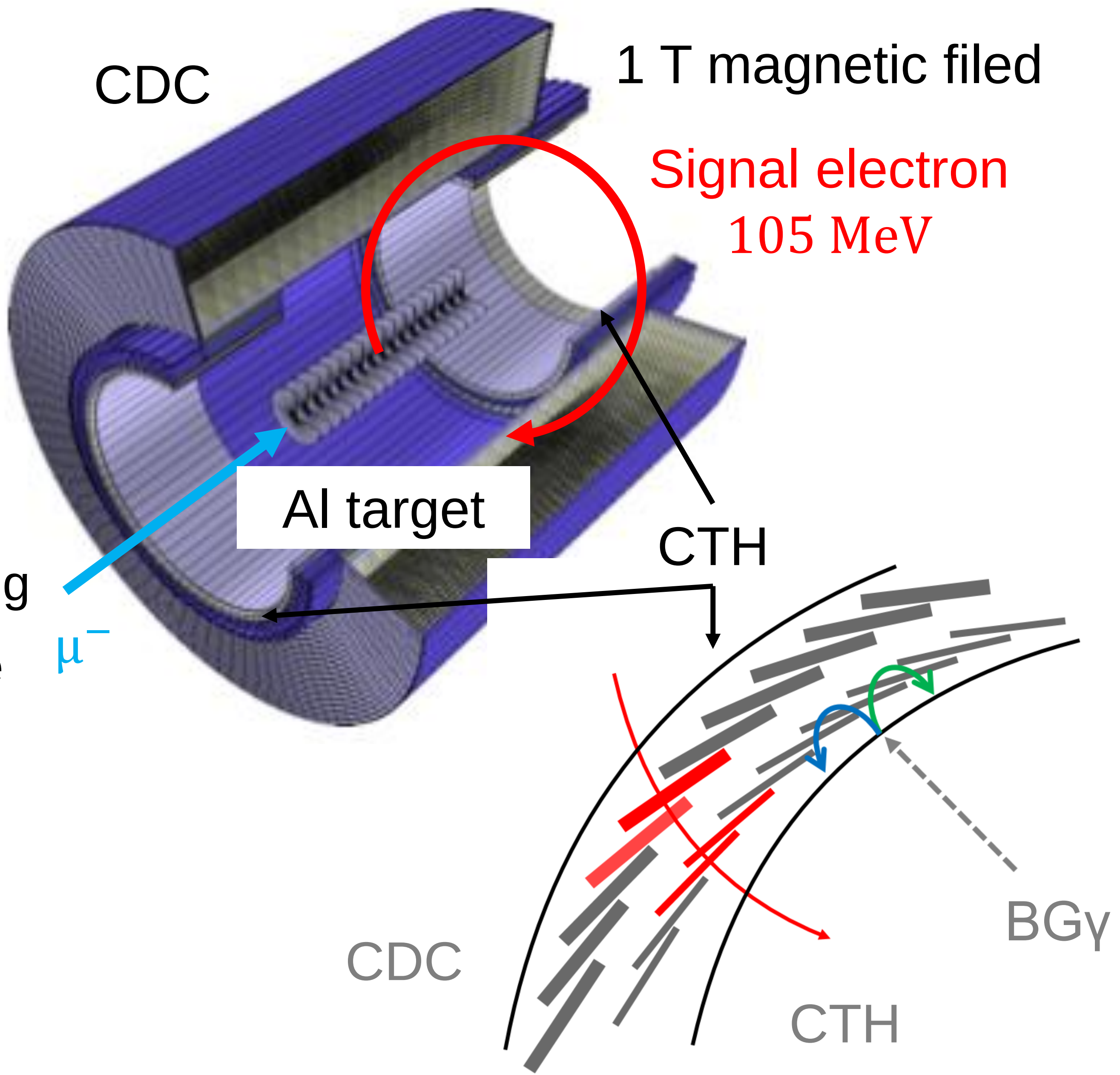
Pulsed proton beam to suppress beam-related backgrounds

Detectors : Cylindrical detector system



Cylindrical Detector System (CyDet)

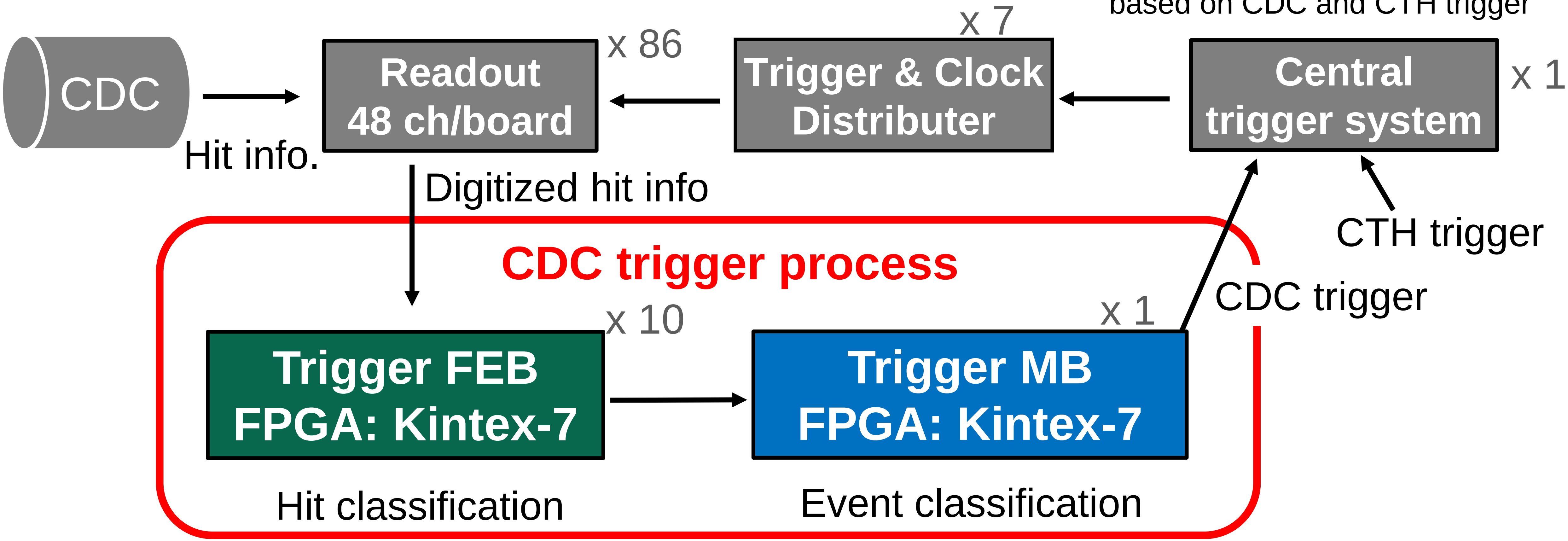
- **CDC (Cylindrical Drift Chamber)**
  - Measure particle momentum
    - Gas mixture He:iC_4H_{10} = 9:1
    - 4986 sense wires
    - 20 layers x ~250 cells/layer
- **CTH (Cylindrical Trigger Hodoscope)**
  - Trigger events and record particle timing
  - 64 pairs of plastic scintillators, with one set placed upstream and downstream
  - 4-fold coincidence suppresses accidental trigger events



Trigger requirements

- **Trigger rate suppression**
  - DAQ trigger rate < 26 kHz
  - Signal efficiency > 90%
- **Fast online event selection**
  - Latency < 6.5μs
- **Implementation to FPGA**
  - Modifiable trigger algorithm

CyDet trigger system



- **Previous research outcomes** [3]
  - GBDT*-based hit classification + Cut-based event classification
  - 96% signal retention efficiency while suppressing trigger rate 13 kHz (CTH trigger rate ~90 kHz)
  - Latency : 3.2 μs

[3]DOI: 10.1109/TNS.2021.3084624

*Gradient Boosted Decision Trees

Motivation to develop new trigger algorithm

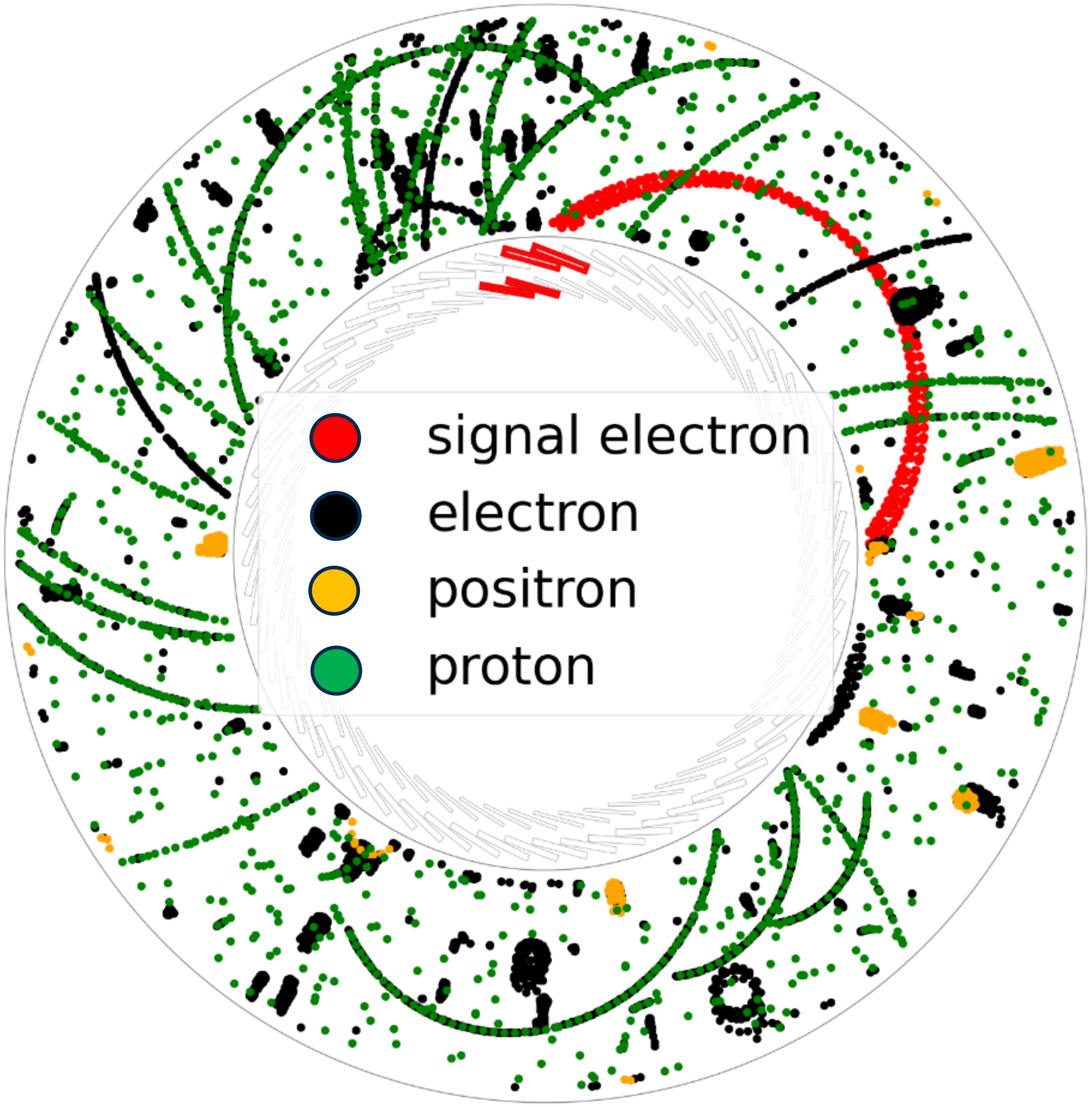
Since the previous research,
the designs of facility and detectors have been changed.

The CTH background trigger rate could be doubled.

The current trigger algorithm cannot achieve high efficiency.
=> **A better trigger algorithm is being studied**.

Event features

Simulated event display in cross–section view in CyDet
Each dot represents MC hits.



Signal characteristics

- **Helical tracks with a signal-like curvature**
- MIP-level hit in a cell

Background characteristics

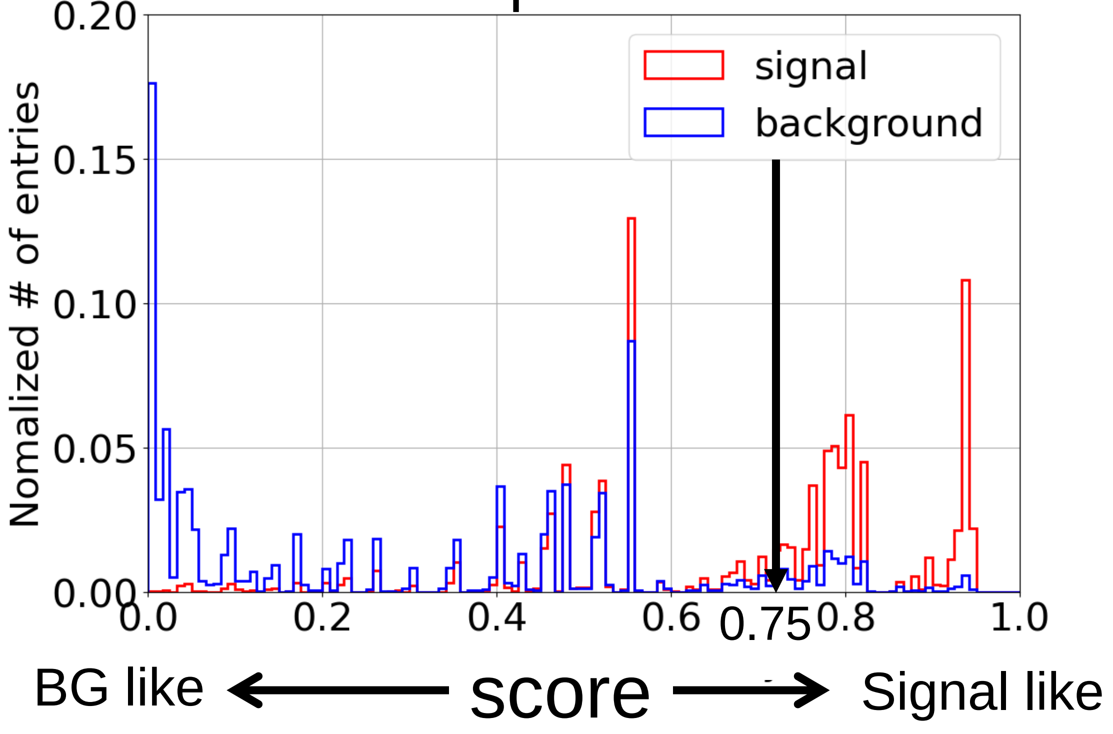
- Low energy electrons
  - Long lived in a small region
- Protons
  - Large curvature
  - Large energy deposit

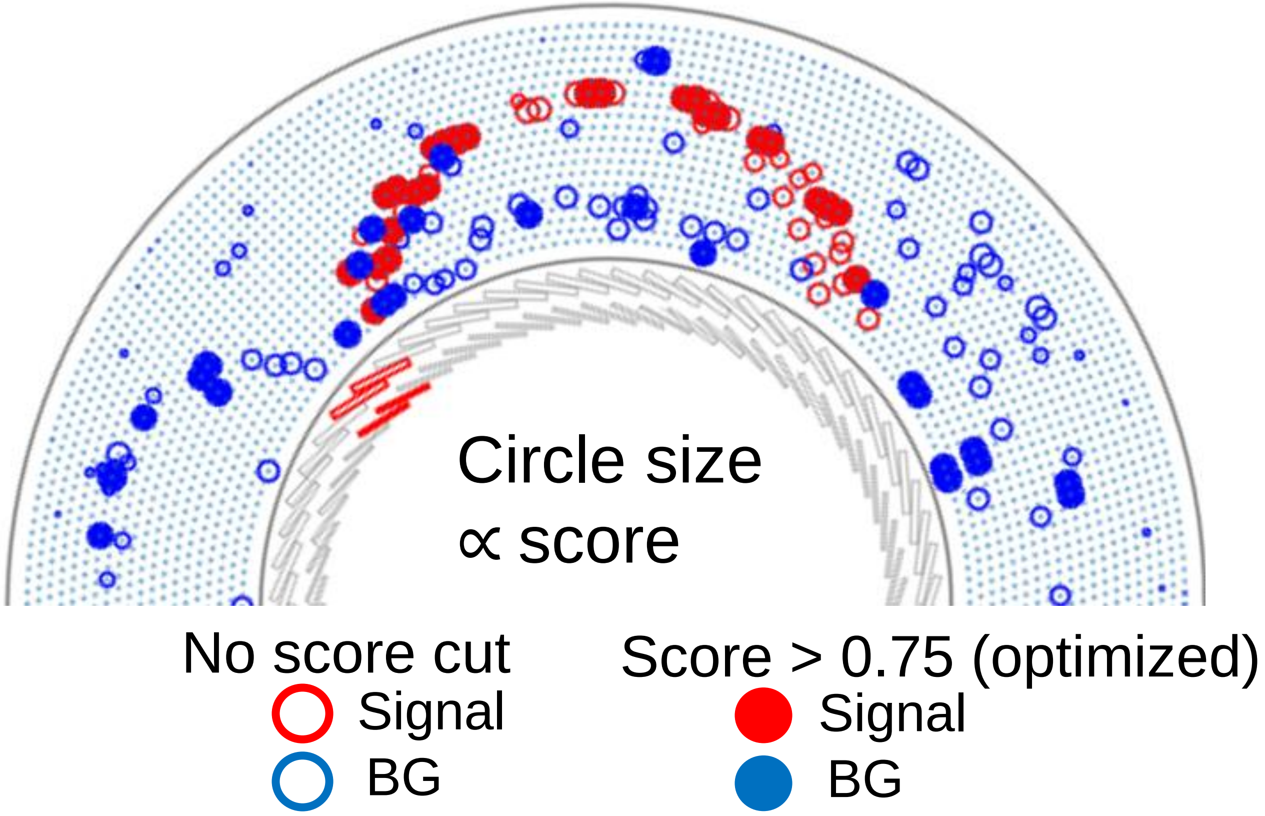
Review : GBDT-based hit classification

GBDT calculates the hit score based on the following local hit features.

1. Energy deposits on the interest wire
2. Energy deposits on the neighboring wires
3. Radial position

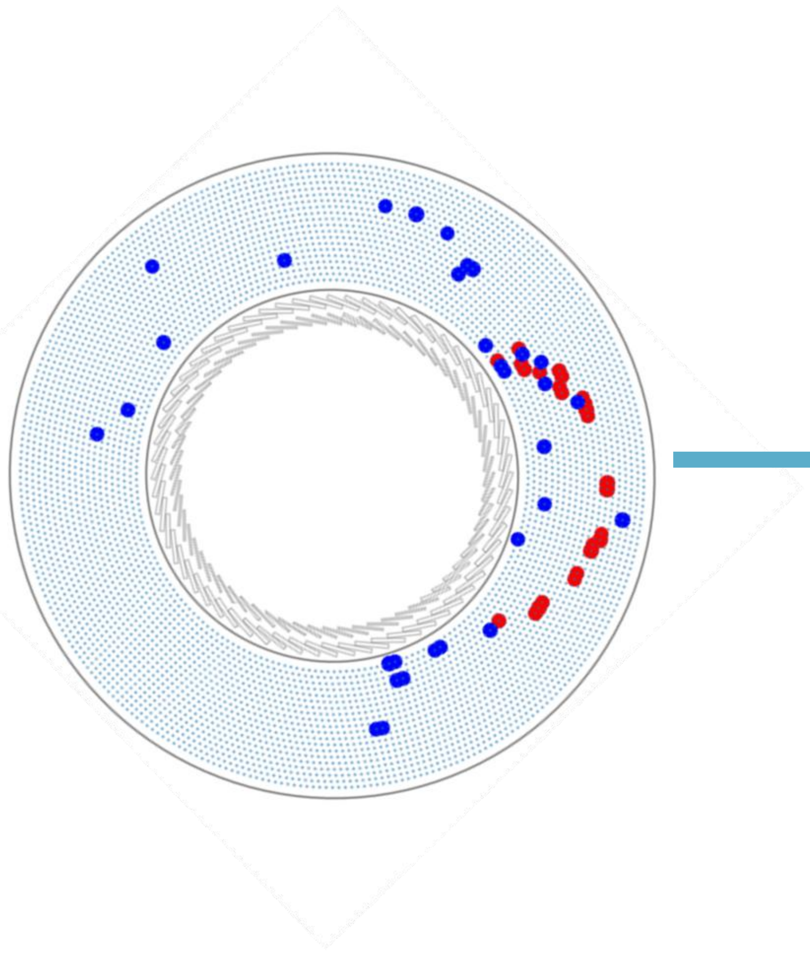
GBDT output distribution

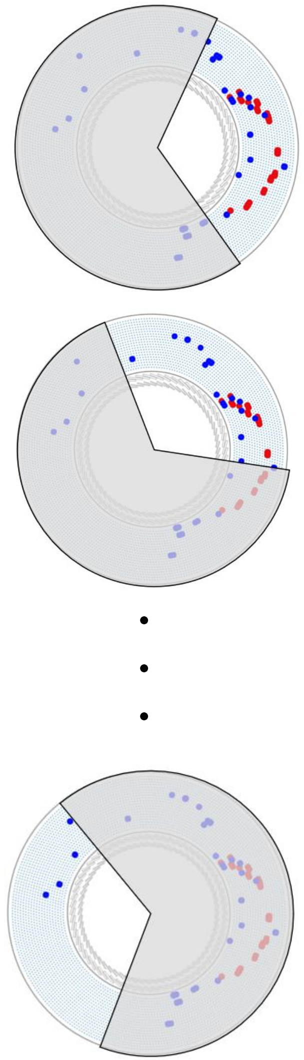




Review :Cut-based event classification

1. Scan each 1/3 area
2. <u>Count the number of the hits</u> that exceeds the score threshold





Generate CDC trigger

The cut-based method <u>doesn't</u> consider the pattern of the helical tracks.

New trigger algorithm

[4] DOI: 10.1088/1748-0221/13/07/P07027

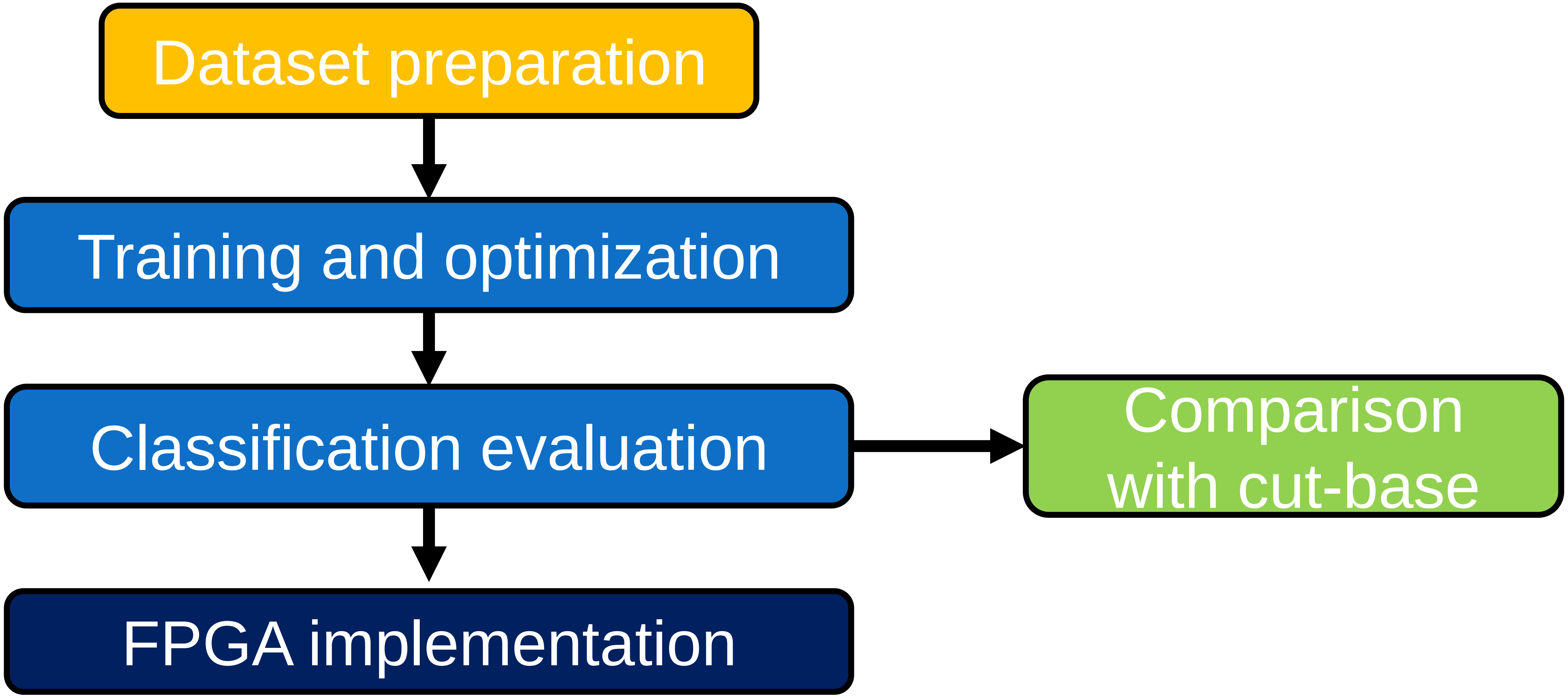
- **Cut-based => Pattern recognition**
  - Conventional algorithms may need
    - massive FPGA resources
    - long processing time
  - Tools such as hls4ml [4] allow us to implement ML-based algorithms on FPGAs.
  - We tested Multilayer Perceptron (MLP).

- **Technical challenges and solutions**
  - Data size limitation
    - Data transfer rate < 3.84 Gbps

      => Use compressed hit data as MLP inputs
  - FPGA resource limitation

    ⇒ Reduce data precision of calculation in MLP (Quantization)

MLP-based algorithm R&D



Dataset preparation

1. Full MC simulation

- **Signal + BG : 3250 samples**
- **Pure BG : 3250 samples**

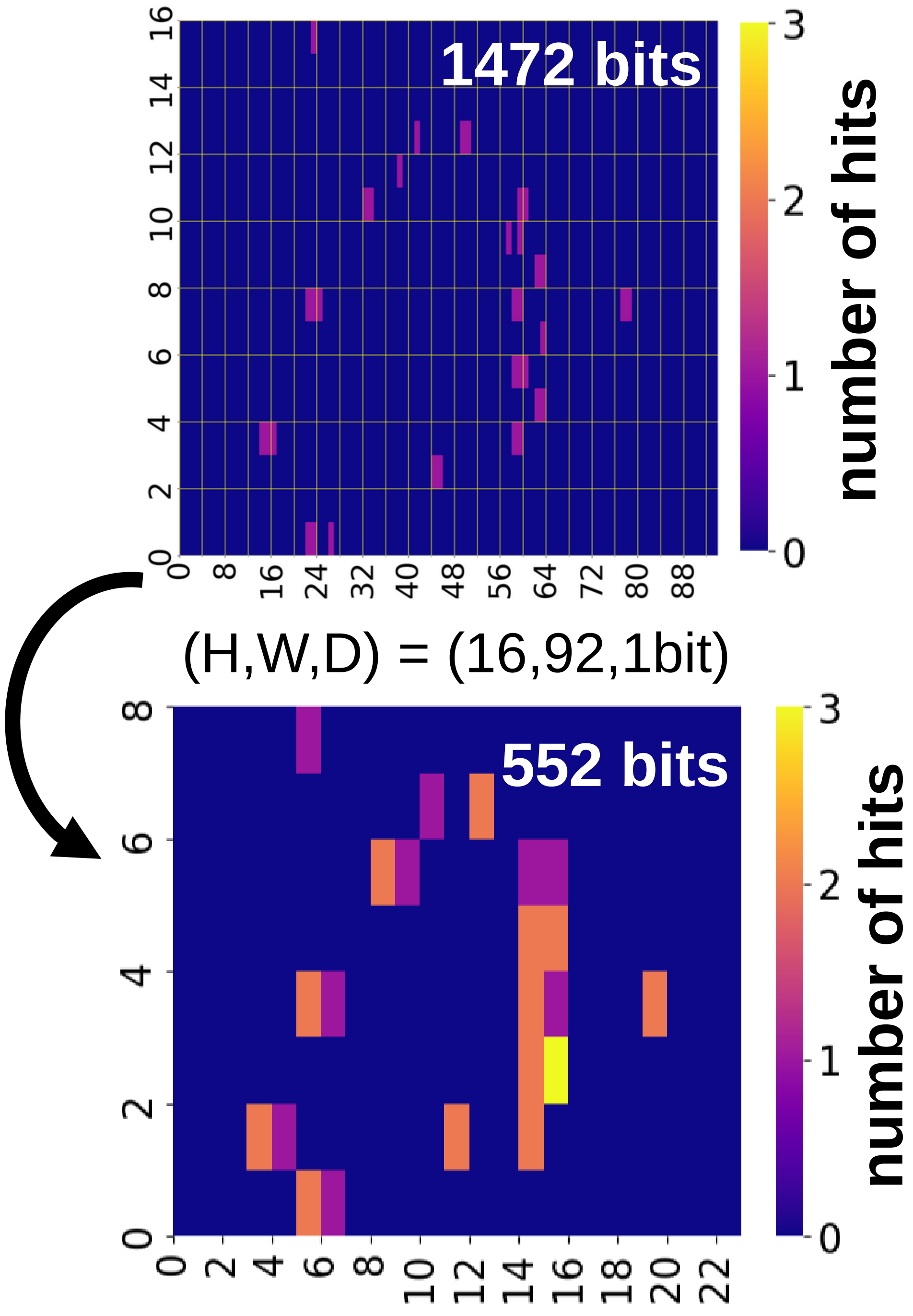
  (real time ~4.5 ms equivalent data)

2. Hit mapping

- hit filtering by GBDT
- 1/3 area extraction

3. hit map compression

- Hit clustering



Training and optimization

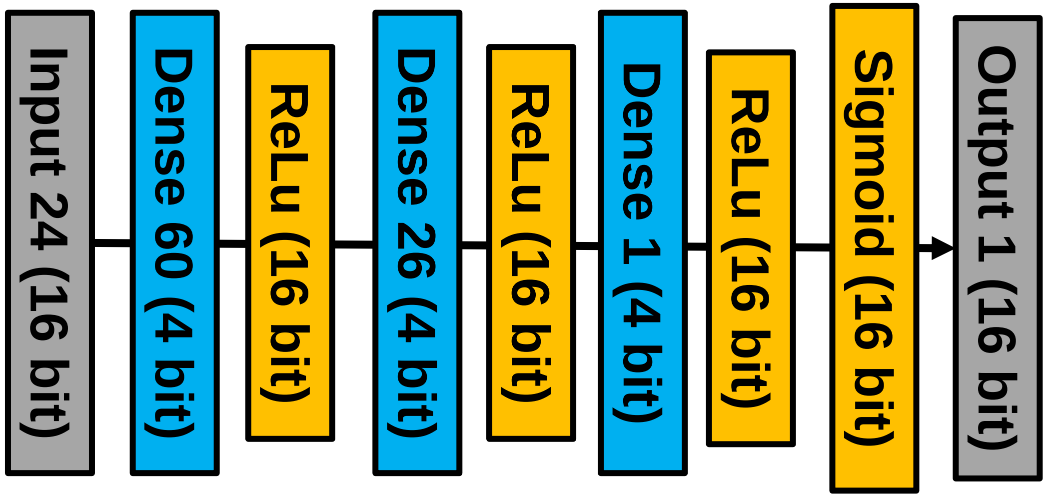
[5] https://doi.org/10.48550/arXiv.2006.10159
[6] https://doi.org/10.1145/3292500.3330701

- Quantized MLP (QMLP) was constructed by using QKeras[5].
  - Quantization (calculation precision reduction) is essential.
    - For our target FPGA, AMD Xilinx Kintex7 xc7k355t-ffg901-1

- Parameters were optimized by using Optuna[6].
  - Hit mapping
    - Score threshold
    - Compression ratio
  - QMLP structures
    - Number of layers
    - Number of neurons
    - Precision

The best QMLP model structure

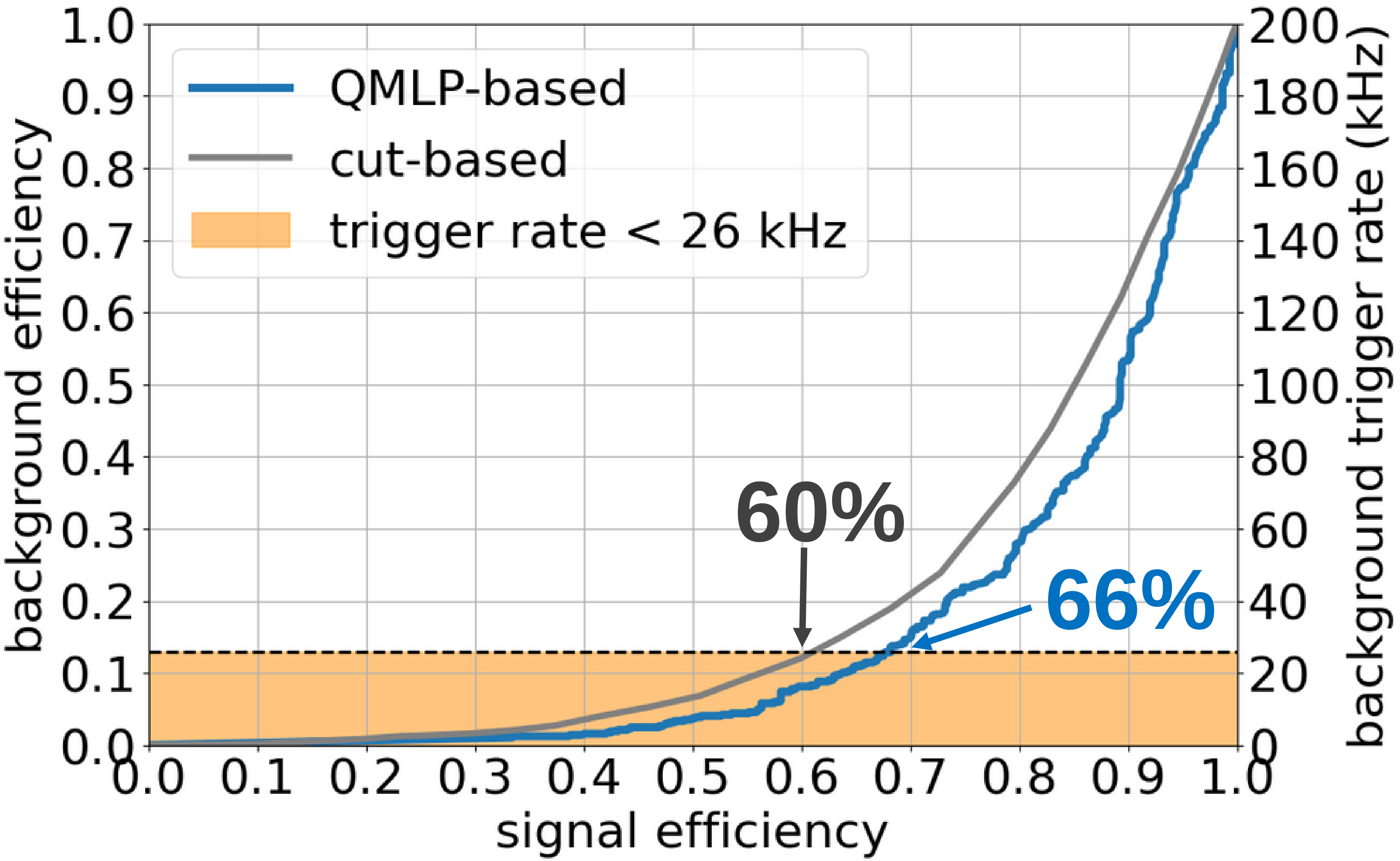
(GBDT score threshold = 0.75)



Number of parameters 3,113

Dense = Fully connected layer

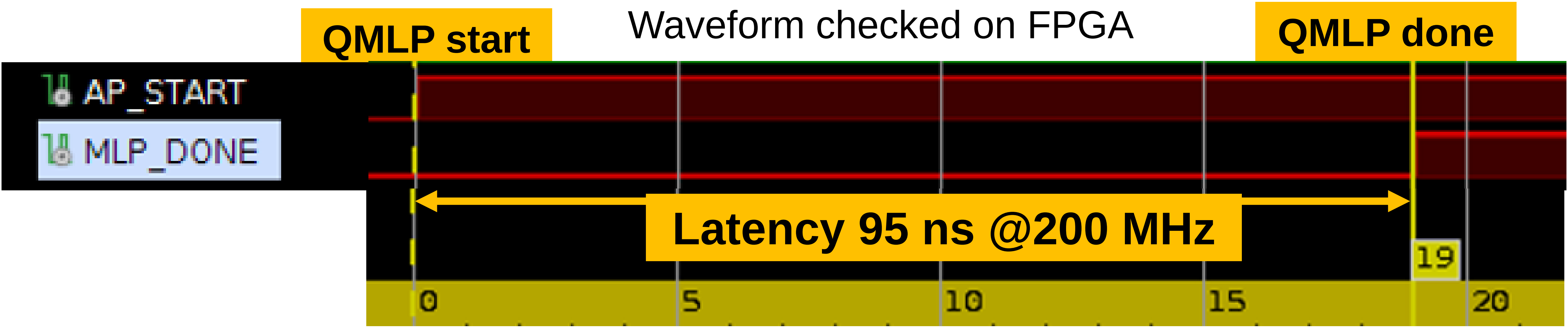
Result



- The signal efficiency is **66% at a trigger rate of 26 kHz**, assuming a CTH trigger rate of 200 kHz.
- **10% better** performance is achieved than the cut-based method.

FPGA implementation

The optimized QMLP was converted into FPGA firmware using hls4ml [4].
The firmware worked in the trigger MB.
The latency was measured with the logic analyzer.



The expected total latency is 3.4 μs.
This satisfies the requirement !!

Summary and prospects

- COMET Phase-I will search for μ-e conversion at J-PARC, Japan.
- An MLP-based trigger algorithm was developed.
  - A model was trained and optimized using MC data.
- It demonstrated **10% better** performance than cut-based method.
  - The signal efficiency is 66% for a trigger rate < 26 kHz.
- The module was successfully implemented on the trigger MB.
  - The expected total **latency 3.4 μs satisfies the requirement**.

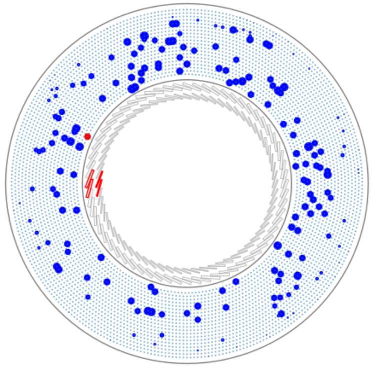
To achieve a signal efficiency > 90%,

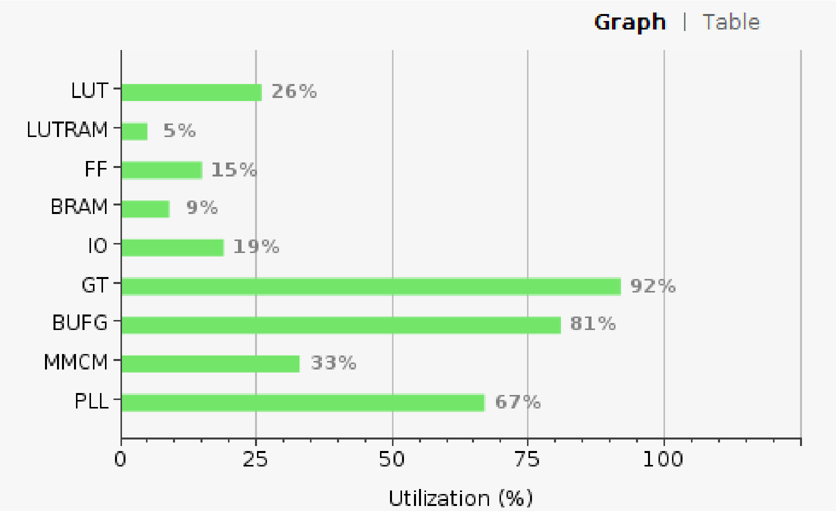
- Increase MC samples
- Redesign the input data structure
- Test other neural networks

backup

Ideas to reach higher efficiency

- Reconsider the definition of "Signal event".
  - CTH 4-fold coincidence + at least a hit in CDC
    - This results in the inclusion of signal electrons that do not have helical trajectories.
  - Study "CTH 4-fold coincidence + at least a hit in 3 layers in CDC
- Improve the GBDT performance
  - Use more local information such as energy deposits of up and down.
  - => Signal hits are always close to other signal hits.
- Redesign the input data structure
  - The compressed hit maps are being flattened into a 1D array for inputs into the MLP.
- Test other neural networks
  - There is still room in both FPGA resource utilization and latency.





COMET Phase-I design updates from the previous study

- Geometry updates
  (finer resolution, made smaller parts and realistic shapes, installation etc)
  - Proton target
  - Proton beam dump
  - Detector/Bridge solenoid vacuum window
  - CyDet cradle installation
  - Detector solenoid yoke installation
  - Comic Ray Veto design/position
  - Radiation shielding
  - CTH design/shield material
  - Etc

- Magnetic field map updates

- Improved physics model (Geant4 updates)

DAQ limitation

			15 14 13 12 11 10 9 8	7 6 5 4 3 2 1 0
Header			Packet Type	Board ID
			Sent Number	
			0 Trigger Time	
			Data Length	
			Trigger Count Upper Bits	
			Trigger Count Lower Bits	
Event Data	Hit Channel Data 0	Header	Channel ID	Length
			Count Over Threshold	
		ADC	Summed ADC Value	
		TDC	TDC Hit 0	
			(TDC Hit 1)	
	Hit Channel Data 1	Header	Channel ID	Length
			Count Over Threshold	
		ADC	Summed ADC Value	
		TDC	TDC Hit 0	
			(TDC Hit 1)	
	Hit Channel Data 2	Header	Channel ID	Length
	...	...	...	
	Hit Channel Data N_{CHdata}	TDC	(TDC Hit 1)	

Data format
Header 12 Byte
Max data 2 Byte x 5 x 48 ch
Total data : below 492 Byte

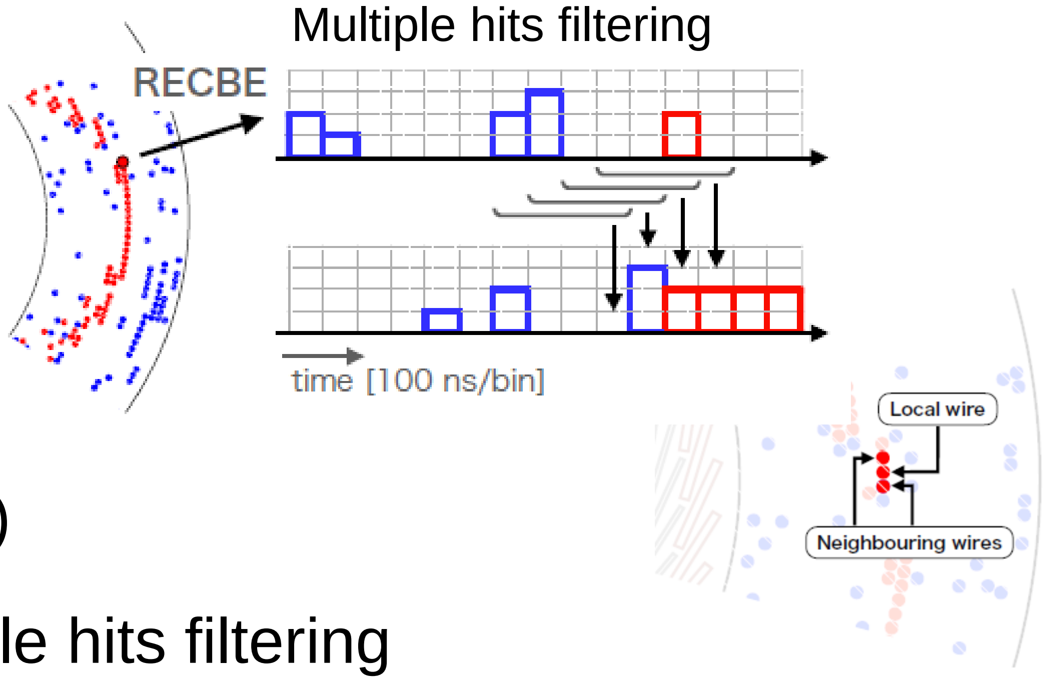
Assumptions:
50% of data comes from CDC
CDC occupancy : 40%
Entire DAQ throughput : 1 Gbps

Acceptable trigger rate 26 kHz

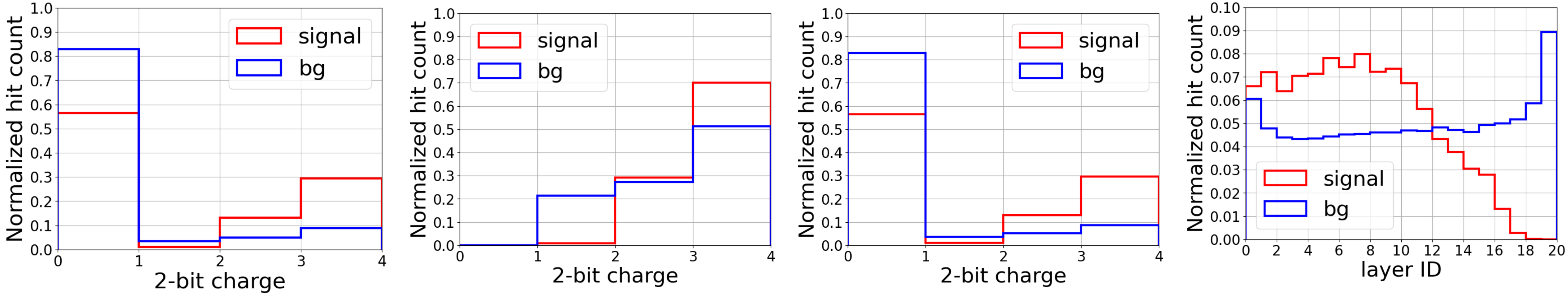
GBDT-based hit classification

Hit classification has 2 steps.

1. Cut multiple hits in the same cell
2. Scoring by GBDT
    - Input features
        - 2-bit charge on the interest wire
        - 2-bit charge on the left and right wires
        - Layer ID (radial distance from CDC center)

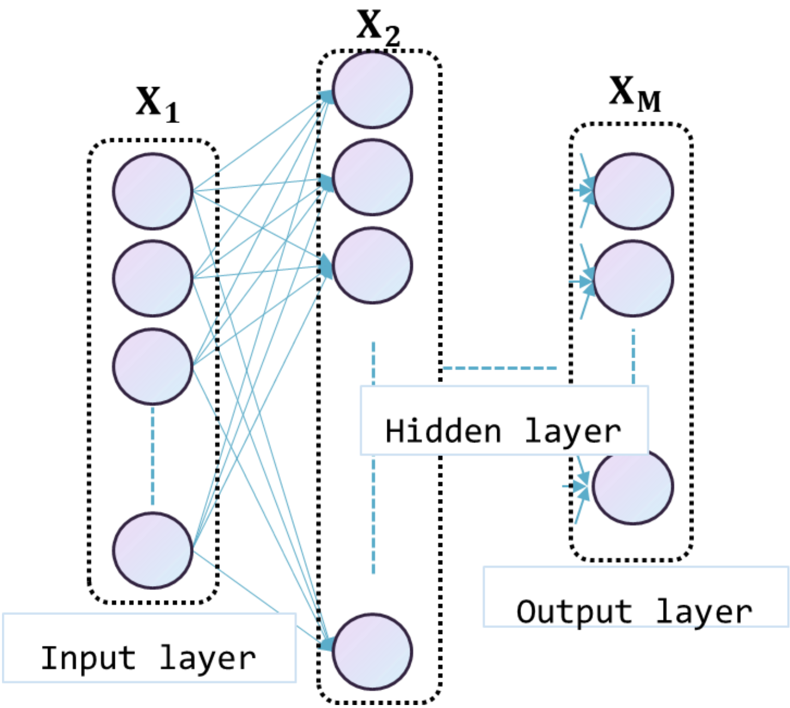


GBDT input features after multiple hits filtering



GBDT was trained by using these variables.

Multilayer Perceptron Implementation technique on FPGA [4]



Activation function — Weight matrix — bias

$$\mathbf{x}_m = g_m(\mathbf{W}_{m,m-1}\,\mathbf{x}_{m-1} + \mathbf{b}_m)$$

Precomputed values stored in BRAM (Block RAM)

DSPs (Digital Processing units) for multiplications

LUTs (Look Up Tables), FFs (Flip-Flops) for additions

- The target FPGA doesn't have enough DSPs to implement MLP.
- More LUTs (222,600) are available than DSPs (1,440) in our FPGA. (AMD Xilinx Kintex7 xc7k355t-ffg901-1)
- LUTs can replace DSPs by reducing the calculation precision (quantization). [7]

[4] DOI: 10.1088/1748-0221/13/07/P07027
[7] DOI: 10.1038/s42256-021-00356-5

Searched parameters

- **Score threshold**
  - 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.75, 0.8, 0.85, 0.9
- **Cluster size (H,W)**
  - (1,6),(1,8),(1,12),(1,48),(2,4),(2,6),(2,12),(2,24),(3,4),(3,8),(3,16),(4,6),(4,12),(6,8)
- **Number of layers**
  - 4 dense layers at maximum
- **Number of neurons**
  - The first layer : Minimum 2, Maximum 64, step 2
  - The second layer onwards : Minimum 2, Maximum 32, step 2
  - Constraint : number of parameters < 4096
- **Precision (bits)**
  - Dense layers : 2, 4, 6
  - ReLu : 3, 4, 8, 16
  - Sigmoid : 2, 4, 6, 8, 10, 16