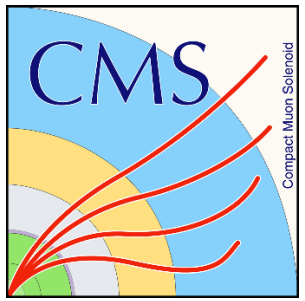
Deep Learning for the Level-1 ME0 Trigger in the CMS Experiment

HEO WooHyeon

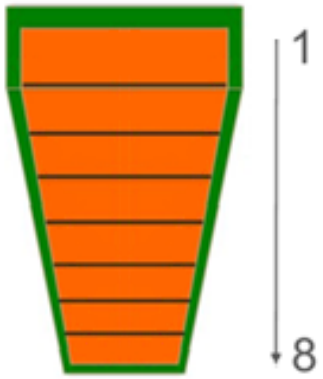
University of Seoul

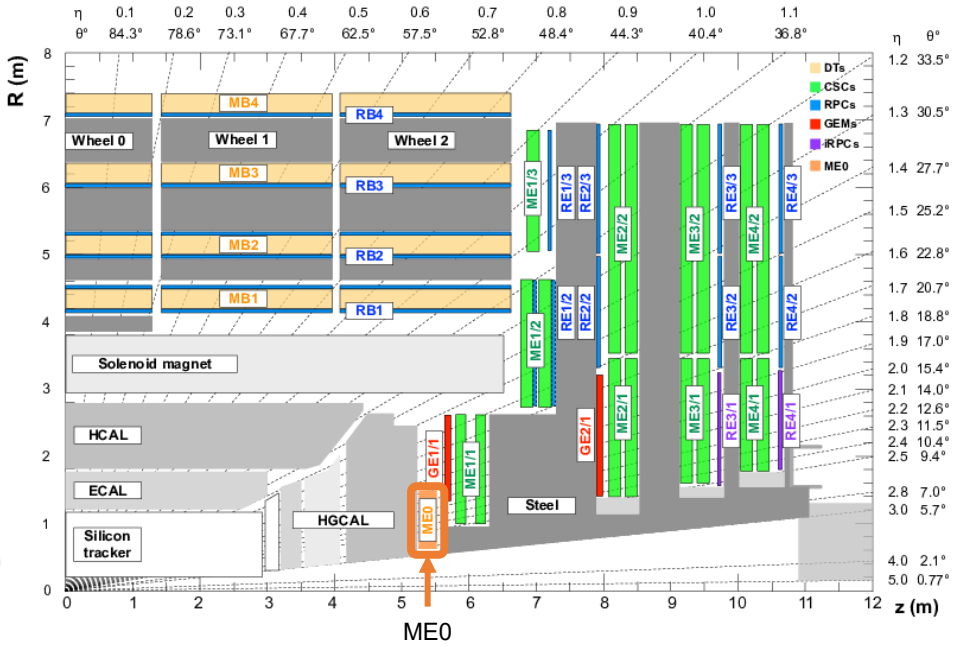
2025 NSRI Workshop

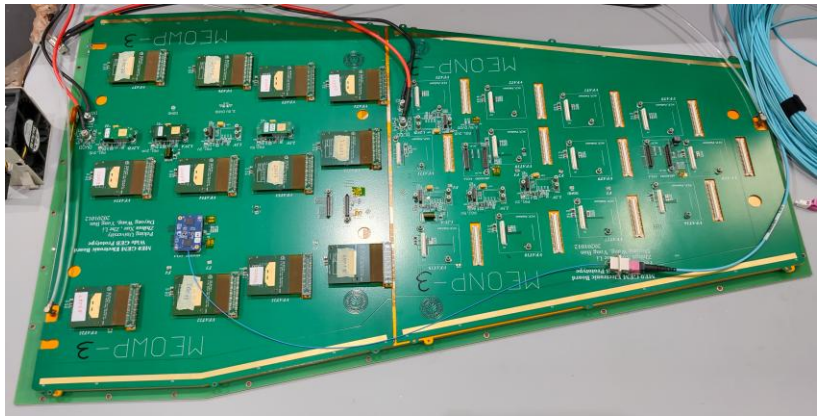


ME0

- In phase-2 upgrade of CMS, ME0 will be installed at the endcap as a part of the Muon system
- Feature[1]:
  - 6-layers of triple Gas Electron Multiplier (GEM) Chamber
    → Stack
  - 18 Stacks will be installed for each disk
  - Inner radius ≈ 0.6 m / Outer radius ≈ 1.5 m
  - Covers **2.0 < |η| < 2.8**, Δφ = 20°
    → The only muon detector above |η| = 2.4
  - Consists of 8 partitions along the η direction (iη) and 384 strips (374 for iη = 1) along the φ direction
- Due to the high background environment of ME0, it is important to trigger on proper targets



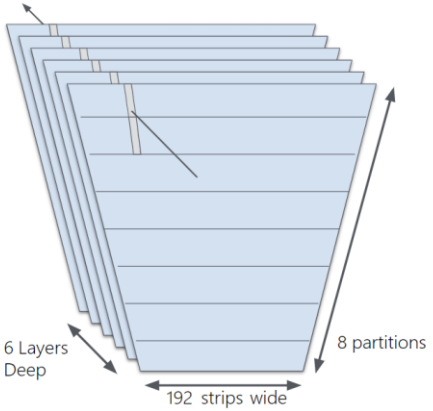


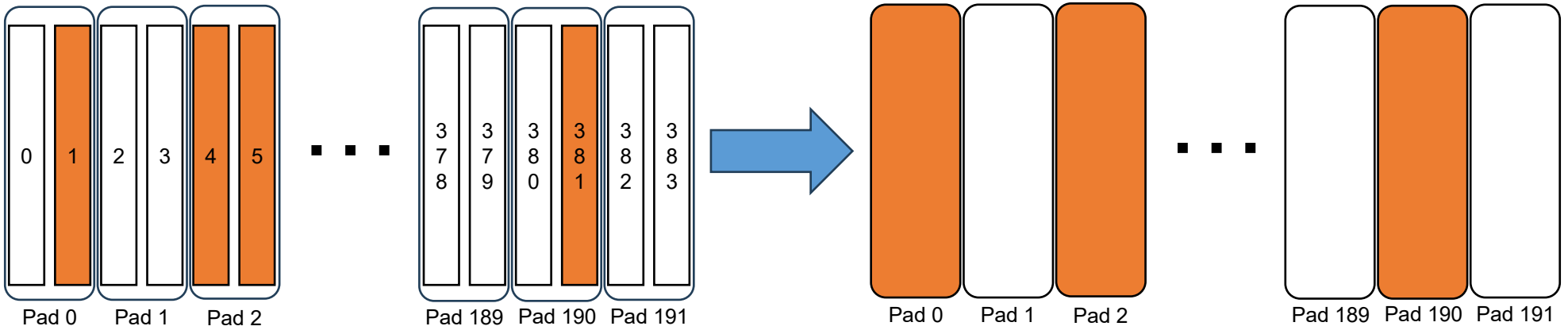


[1] “The Phase-2 Upgrade of the CMS Muon Detectors”, CMS Report (2017)

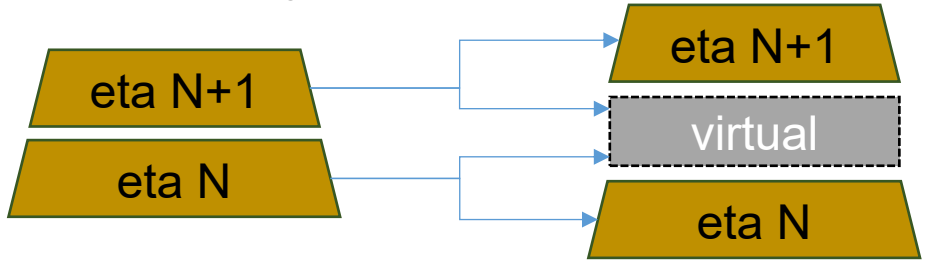
ME0 Stub Finder

1. Pre-Processing Data
   1) Pad Strip : PadStrip(N) <= Strip(2N) Or Strip(2N+1)
      → Reducing the processing time





   2) Combined eta partitions :
      Original 8 eta Partitions + 7 virtual partitions (combined data of two adjacent eta partitions)
      → Able to detect a track crossing two eta partitions



ME0 Stub Finder

Used Sample :

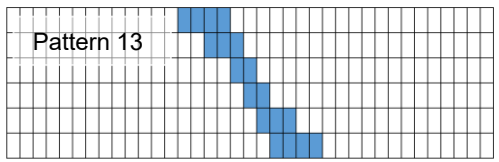
For efficiency : 50,000 events, each containing 8 randomly generated muons with uniform p_T=1–200 GeV and |η|=2.0–2.8, along with an average of 200 additional pile-up collisions per bunch crossing (BX)

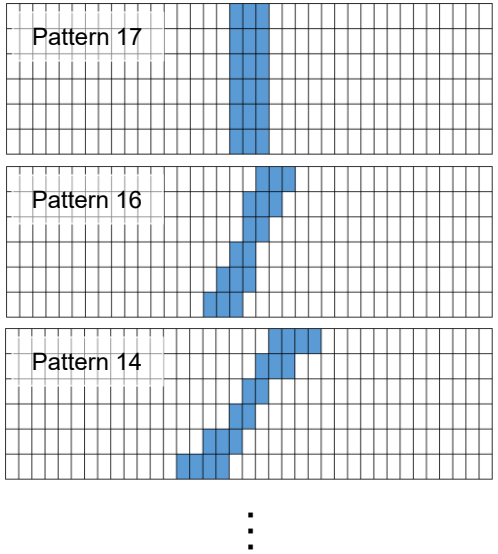
For Minbias rate : 50,000 events, only pile-up collisions, with an average of 200 per BX

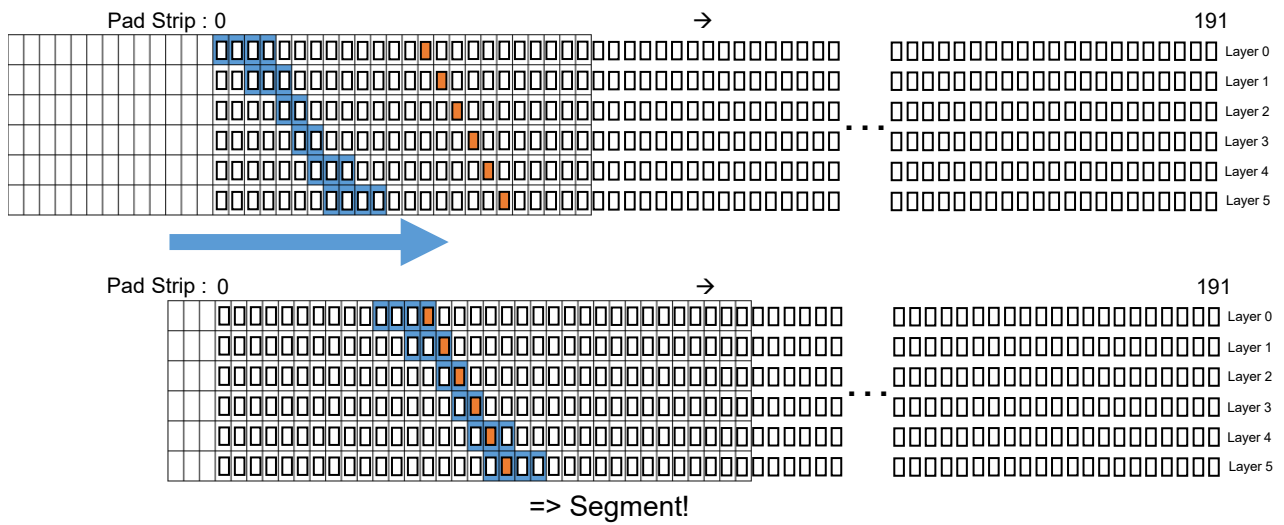
2. Scanning Pad Data with pattern masks

Pattern 1, 3, 5, … , 15
are mirrored patterns of
Pattern 2, 4, 6, … , 16

ex)







- Segments that satisfy certain conditions are sent to the Endcap Muon Track Finder (EMTF)
  - A segment must have hits in at least 4 layers (Minimal requirement)
  - 27 Bits per segment (4 : Eta / 10 : Phi / 9: Bending Angle / 4 : Quality)
  - Position and Bending Angle are obtained by linear regression
- Simulation result with minimal requirement:
  - Efficiency = 99.17 %
  - Minbias rate per chamber = 179.7 MHz

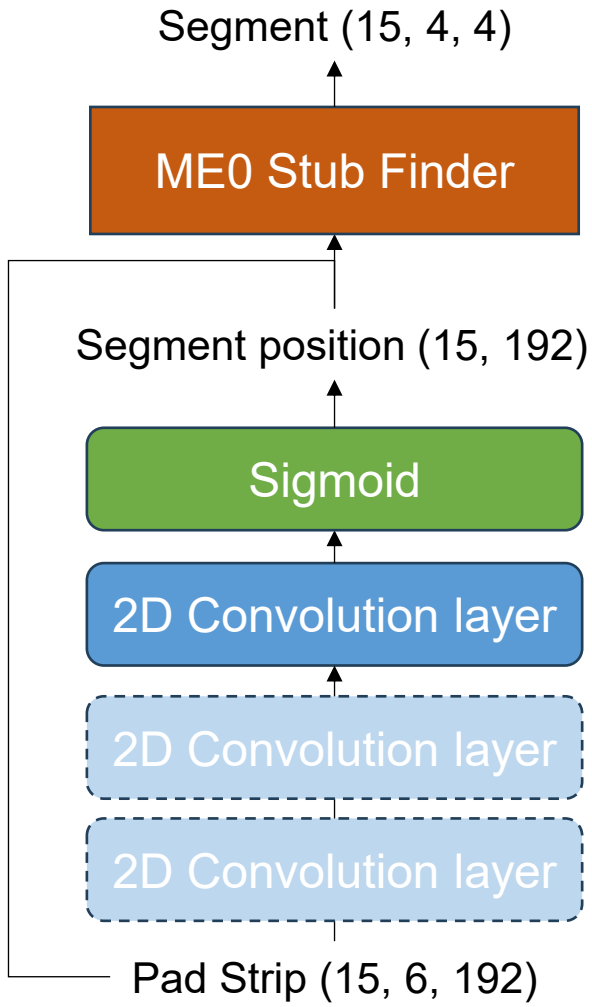
$$\text{Muon Efficiency} = \frac{(\#\text{ of matched muon track})}{(\#\text{ of total muon track})}$$

$$\text{Minbias rate per chamber} = \frac{(\#\text{ of unmatched segment})}{(36\text{ chambers}) \times (25\text{ ns})} \times (\text{Fill Factor})$$

Concerning number of segments to send EMTF / most of them are effect of Pile-up → need to filter the segments from Pile-Ups

ME0 Stub Finder with ML

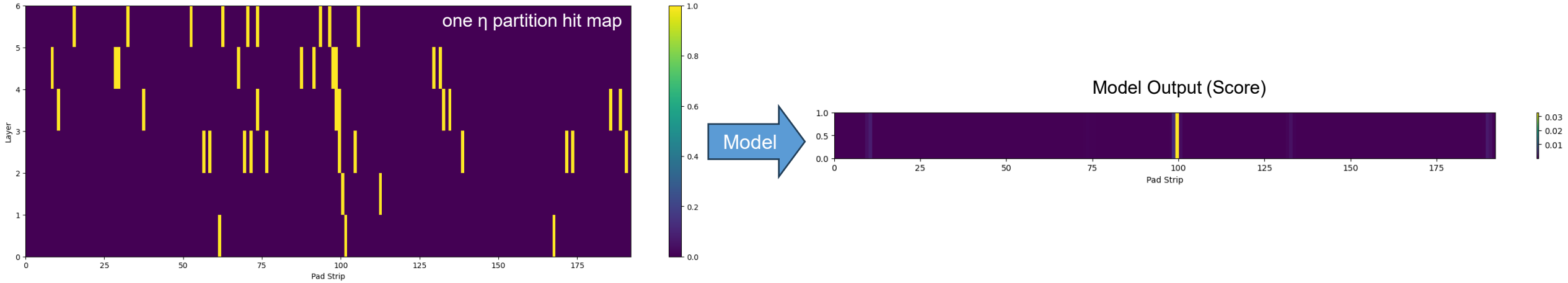
- Study for ME0 Stub Finder (ME0SF) using CNN
  - CNN show great performance in pattern finding problem
- Model :
  - 1-3 layers of 2D Convolution filter + ReLU between CNNs
  - Set to be light to meet the maximum processing time of ME0 Trigger system
- Input Data : (15, 6, 192)
  - Pad Strip data for 15 η partitions (8 original η partitions + 7 virtual partitions)
  - Training data : muons with p_T = 1-10 GeV and average 200 Pile-Up (~100,000 events)



ME0 Stub Finder with ML

- Output Data : (15, 192)
  - Segment strip position
    - 15 vectors corresponding to strip-wise segment position for 15 partitions
    - Each vector has 192 dimensions and n^{th} dimension correspond to n^{th} pad strip
    - The score of 0 to 1 will be given representing how a segment is likely be at that pad strip
  - ME0SF will only run on the strips specified by the model, while the standard algorithm scans all strips

    → Potential to decrease the processing time for ME0SF

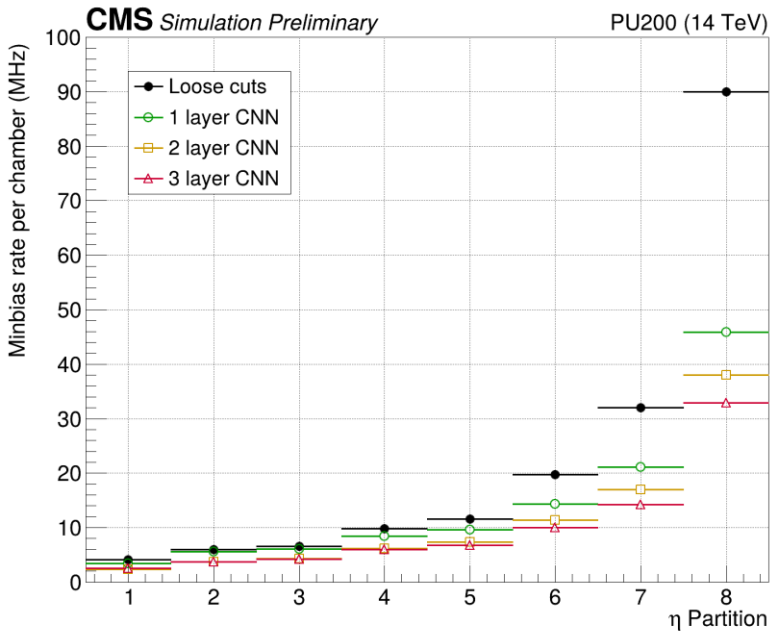


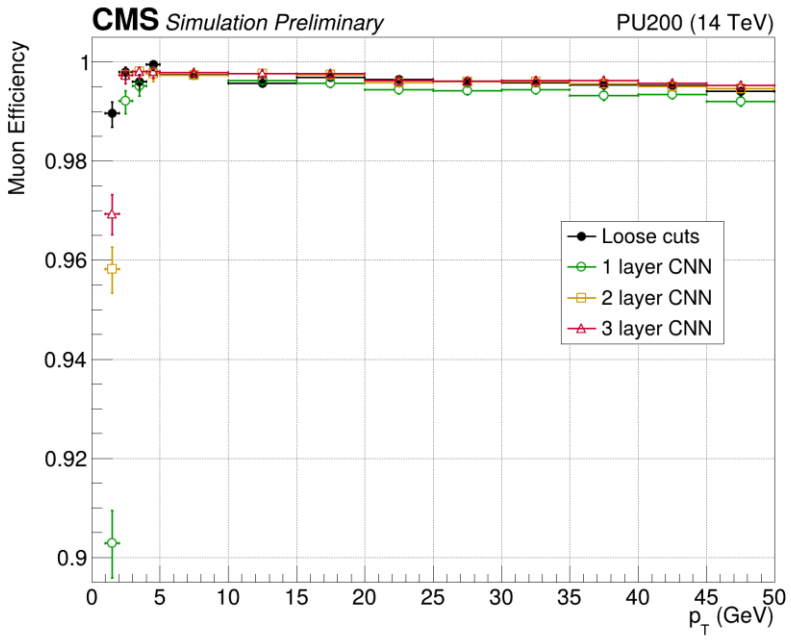
Result (I)

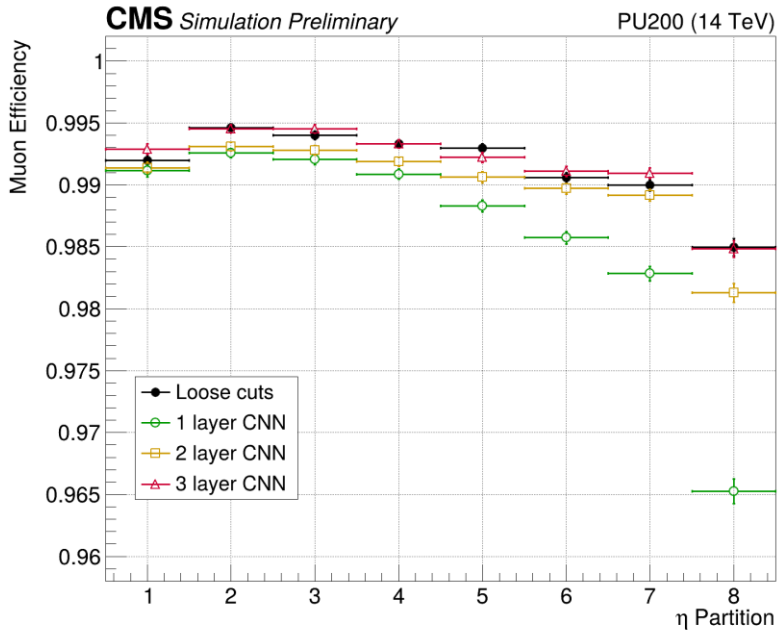
Used Sample :

For efficiency : 50,000 events, each containing 8 randomly generated muons with uniform p_T=1–200 GeV and |η|=2.0–2.8, along with an average of 200 additional pile-up collisions per bunch crossing (BX)

For Minbias rate : 50,000 events, only pile-up collisions, with an average of 200 per BX







Overall performance	Loose cut*	1 layer CNN	2 layers CNN	3 layers CNN
Muon Efficiency	99.17%	98.70%	99.04%	99.21%
Minbias rate per chamber	179.7 MHz	114.3 MHz	90.10 MHz	80.32 MHz

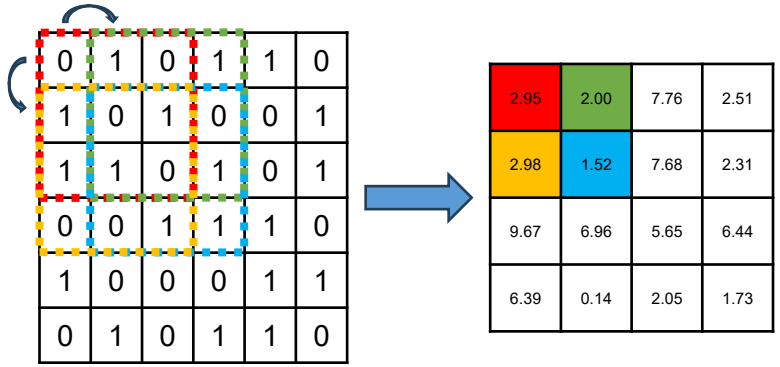
* “Loose cut” indicate the standard ME0SF implementation with a minimal segment requirement, defined as having hits in at least 4 layers. The “loose cut” is also applied to CNN assisted ME0SF

$$\text{Muon Efficiency} = \frac{(\#\text{ of matched muon track})}{(\#\text{ of total muon track})}$$

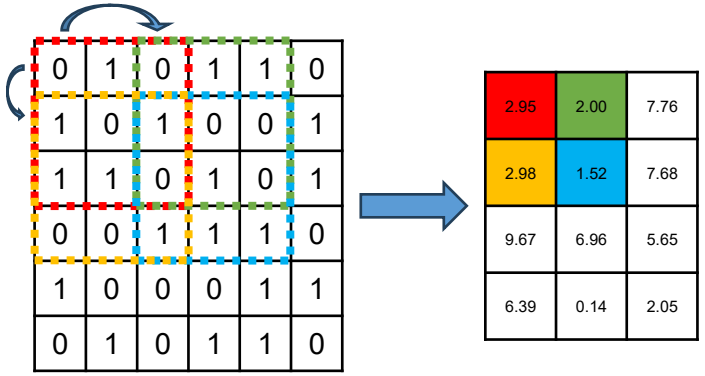
$$\text{Minbias rate per chamber} = \frac{(\#\text{ of unmatched segment})}{(36\text{ chambers}) \times (25\text{ ns})} \times (\text{Fill Factor})$$

Optimization - Stride

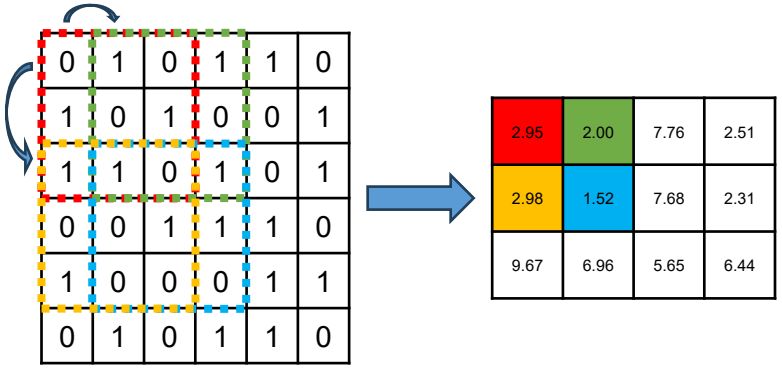
- In convolutional neural networks (CNNs), stride controls how the filter moves across the image:
  - The CNN filters move one pixel at a time if stride is 1
  - Otherwise, they skip (stride - 1) pixels



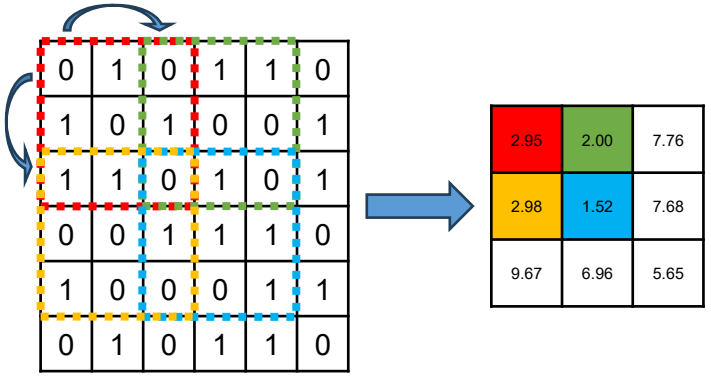
Stride = (1, 1)



Stride = (1, 2)



Stride = (2, 1)

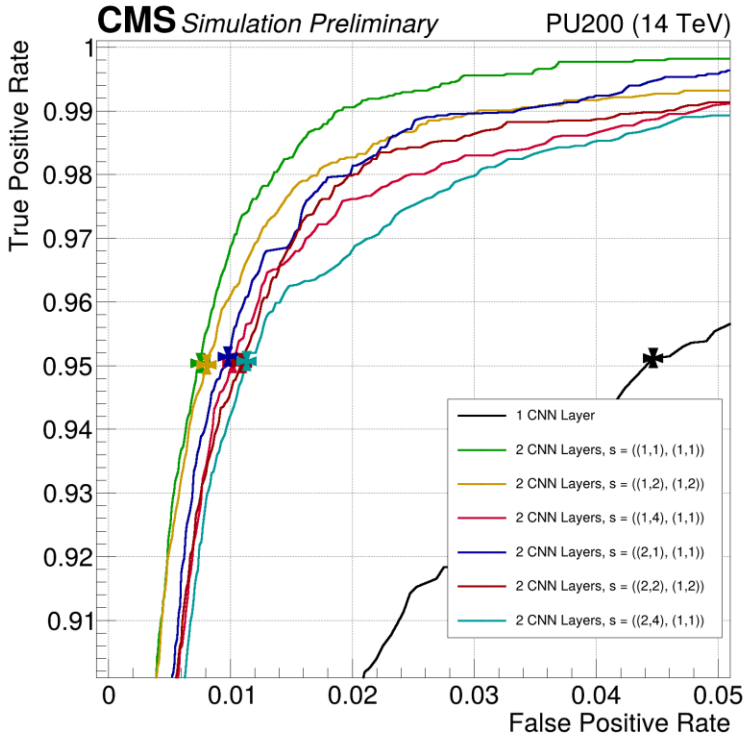


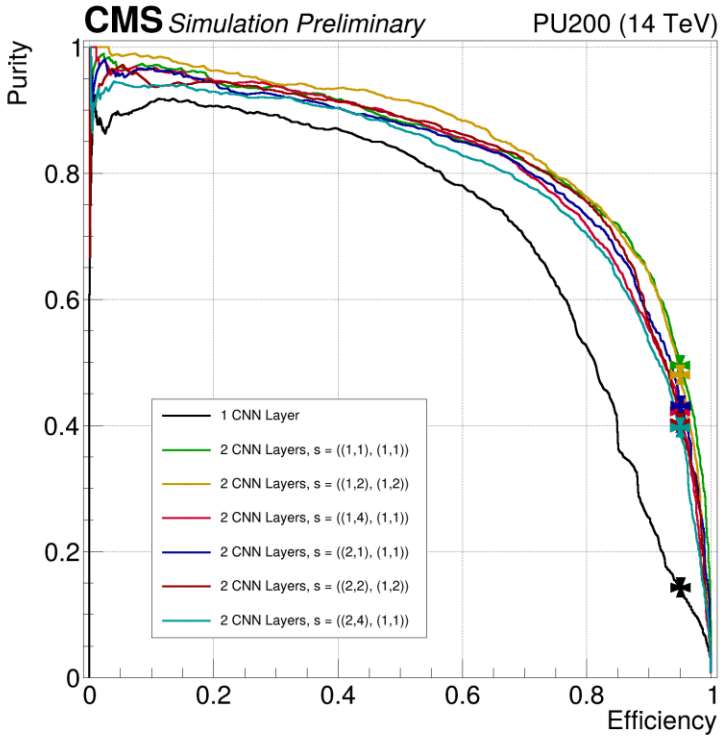
Stride = (2, 2)

Optimization - Stride

Used Sample :

10,000 events, each containing 8 randomly generated muons with uniform p_T=1–10 GeV and |η|=2.0–2.8, along with an average of 200 additional pile-up collisions per bunch crossing (BX)





- Optimize the number of stride for each CNN layer to reduce processing time while preserving the performance
- Tested with the 2 layers model with N_{kernel}=5, kernel size = ((3, 5), (6, 5)) or ((3, 5), (3, 5) for vertical stride = 2)
- Quantization-aware training is not applied

$$\text{Efficiency} = \frac{\text{True Positive}}{\text{True Positive + False Nagative}}$$

$$\text{Purity} = \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

* Each point indicates 95% efficiency at 1-10 GeV samples

Stride	1 layer CNN	(1, 1), (1, 1)	(1, 2), (1, 2)	(1, 4), (1, 1)	(2, 1), (1, 1)	(2, 2), (1, 2)	(2, 4), (1, 1)
Efficiency	0.9512	0.9504	0.9501	0.9504	0.9514	0.9504	0.9506
Purity	**0.1428**	**0.4952**	**0.4802**	**0.4223**	**0.4312**	**0.4042**	**0.3963**

Processing Latency

- Processing Latency result from Vitis-hls synthesis
  - Target Device = xcvu13p-flga2577-2-e
  - Target Clock Period = 2.778 ns
  - Minimum Requirement : Max Latency < 1 μs
  - Using Model of N_{kernel}=5, Kernel size = (3, 5), (6, 5) or ((3, 5), (3, 5) for vertical stride = 2) / N_{kernel} = 5 / 8 bits quantization

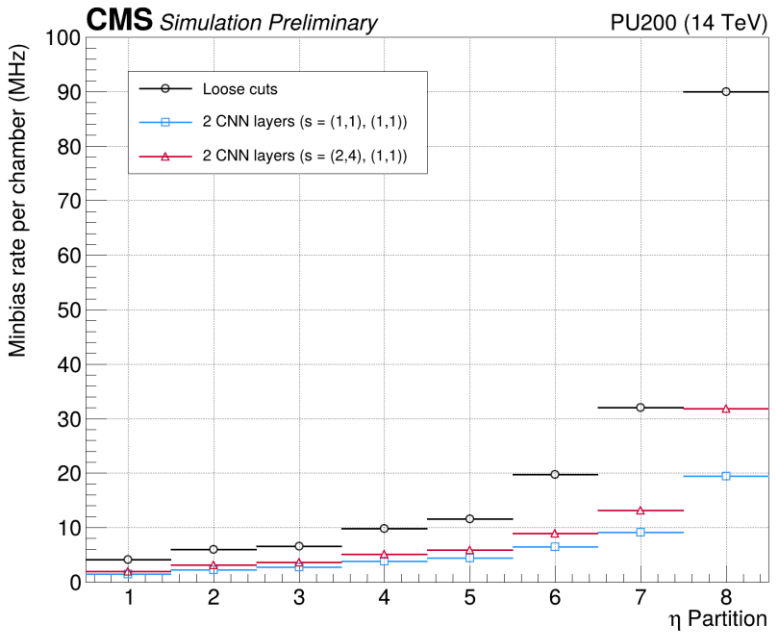
Model	Max Latency (cycle / absolute)	FF (Total / Utilization SLR %)	LUT (Total / Utilization SLR %)	Efficiency	Purity
1 layer CNN	33 / 91.67 ns	228,751 / 26%	1,277,557 / 295%	0.9512	0.1428
2 layers CNN (Stride = (1,1), (1,1))	1,359 / 3.775 μs	463,788 / 53%	1,057,236 / 244%	0.9504	0.4952
2 layers CNN (Stride = (1,2), (1,2))	640 / 1.778 μs	211,722 / 24%	434,080 / 100%	0.9501	0.4802
2 layers CNN (Stride = (1,4), (1,1))	351 / 0.975 μs	128,373 / 14%	252,951 / 58%	0.9504	0.4223
2 layers CNN (Stride = (2,1), (1,1))	784 / 2.178 μs	238,345 / 27%	530,045 / 122%	0.9514	0.4312
2 layers CNN (Stride = (2,2), (1,2))	350 / 0.972 μs	111,982 / 12%	224,469 / 51%	0.9504	0.4042
2 layers CNN (Stride = (2,4), (1,1))	**207 / 0.575 μs**	**73,438 / 8%**	**127,650 / 29%**	**0.9506**	**0.3963**

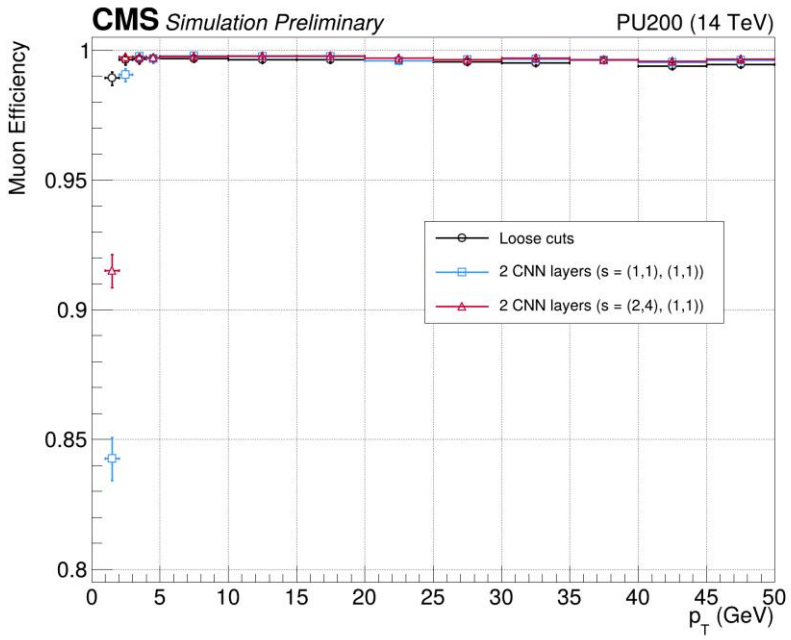
Result (II)

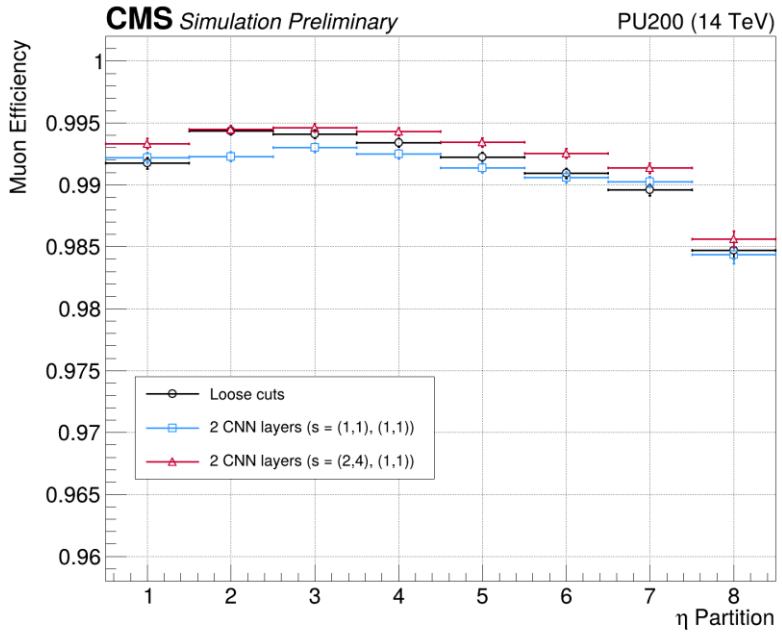
Used Sample :

For efficiency : 50,000 events, each containing 8 randomly generated muons with uniform p_T=1–200 GeV and |η|=2.0–2.8, along with an average of 200 additional pile-up collisions per bunch crossing (BX)

For Minbias rate : 50,000 events, only pile-up collisions, with an average of 200 per BX







Overall performance	Loose cut*	2 CNN layers (s = (1,1), (1,1))	2 CNN layers (s = (2,4), (1,1))
Muon Efficiency	99.17%	99.11%	99.28%
Minbias rate per chamber	179.7 MHz	49.66 MHz	73.55 MHz

* “Loose cut” indicate the standard ME0SF implementation with a minimal segment requirement, defined as having hits in at least 4 layers. The “loose cut” is also applied to CNN assisted ME0SF

$$\text{Muon Efficiency} = \frac{(\#\text{ of matched muon track})}{(\#\text{ of total muon track})}$$

$$\text{Minbias rate per chamber} = \frac{(\#\text{ of unmatched segment})}{(36\text{ chambers}) \times (25\text{ ns})} \times (\text{Fill Factor})$$

Summary

- CNN Models are trained to filter the pile-up induced segments in ME0SF
- Stride size is optimized to minimize the processing time while maintain the performance
  - There was no critical difference in the performance after applying the strides
  - The processing time and resource usage decreased inversely proportionally to the stride size of 1st CNN layer
- CNN Models effectively reduced the minbias rate while preserving efficiency even for high η or low p_T
  - Performance drop was observed after applying strides to CNN layers, but it still able to remove > 50% of minbias rate

Next Plan

- Further optimization for N_{kernel}, kernel size and quantization precision
- Get the full processing time of ME0SF combined with ML models

Back Up

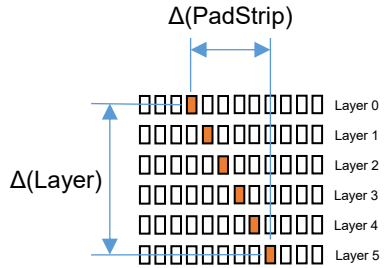
Performance

- Performance of ME0SF combined with Models combined
- Sample
  - For efficiency : 50,000 events, each containing 8 randomly generated muons with uniform p_T=1–200 GeV and |η|=2.0–2.8, along with an average of 200 additional pile-up collisions per bunch crossing (BX)
  - For Minbias rate : 50,000 events, only pile-up collisions, with an average of 200 per BX
- Matching rule :
  - A segment is considered matched with a muon track if

    Eta position match : $\left|(\eta\ Partition)_{MuonTrack} - (\eta\ Partition)_{segment}\right| \leq 1$

    Strip position match : $\left|(PadStrip)_{MuonTrack} - (PadStrip)_{segment}\right| \leq 5$

    Bending angle match : $\left|(Bending\ Angle)_{MuonTrack} - (Bending\ Angle)_{segment}\right| \leq 0.4$

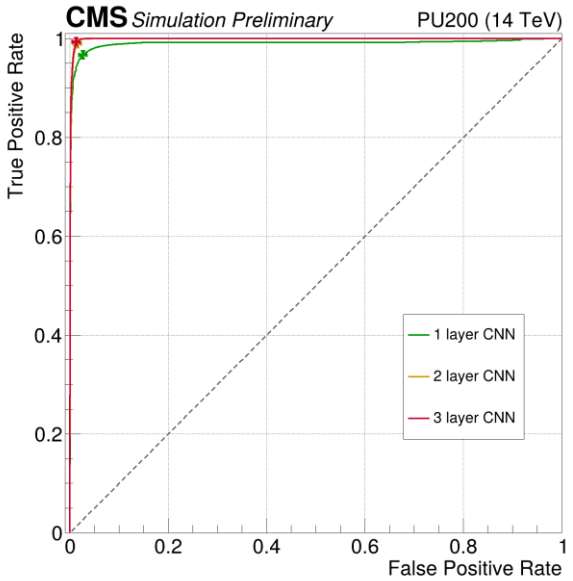


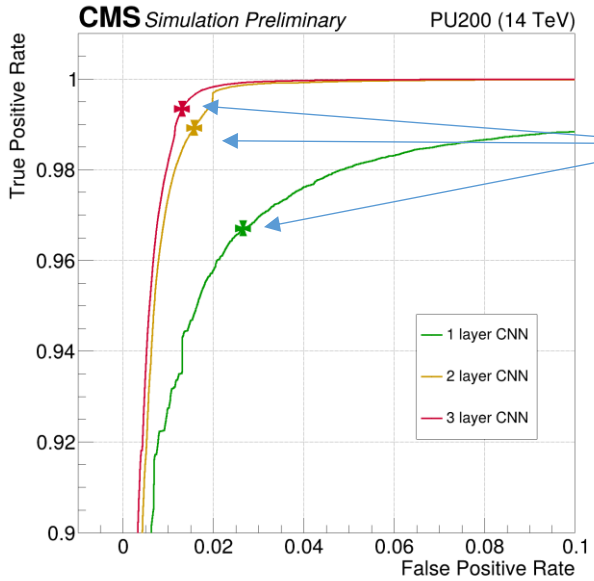
➔ Bending Angle = Δ(PadStrip)/Δ(Layer)

- Muon Efficiency $= \dfrac{(\#\text{ of matched muon track})}{(\#\text{ of total muon track})}$
- Minbias rate per chamber $= \dfrac{(\#\text{ of unmatched segment})}{(36\text{ chambers}) \times (25\text{ ns})} \times (\text{Fill Factor})$

Fill Factor ≈ 0.7710

Overall performance





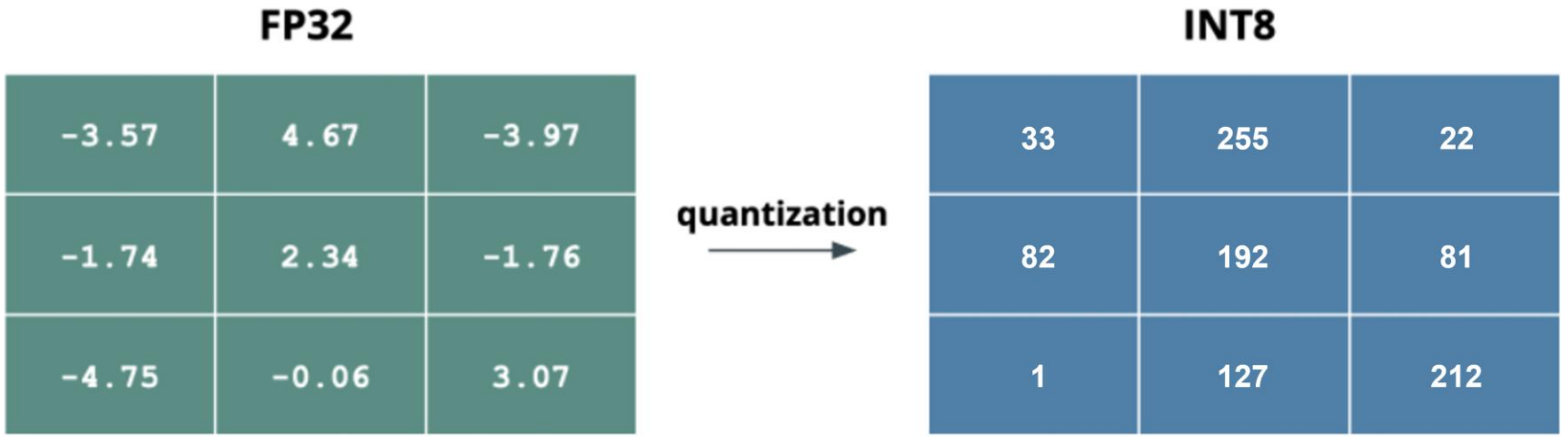
Best Discriminator Points

	Loose cut*	1 layer CNN	2 layers CNN	3 layers CNN
Efficiency	99.19 %	98.70 %	99.04 %	99.21 %
Minbias rate per chamber	179.7 MHz	114.3 MHz	90.10 MHz	80.32 MHz

For every case of CNN, Minbias rate is reduced by 1/3 to 1/2 of the one from the standard ME0SF while preserving ~99% of Efficiency, even for the 1-layer CNN

* “Loose cut” indicate the standard ME0SF implementation with a minimal segment requirement, defined as having hits in at least 4 layers. The “loose cut” is also applied to CNN assisted ME0SF

Quantization



$$y = W \cdot x + b$$

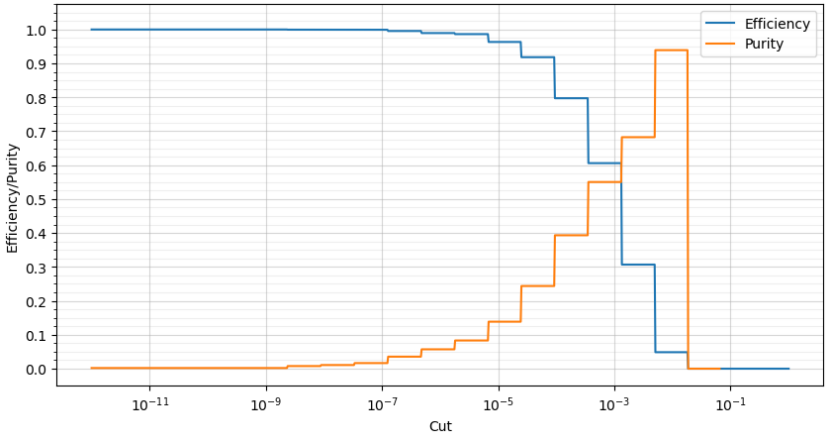
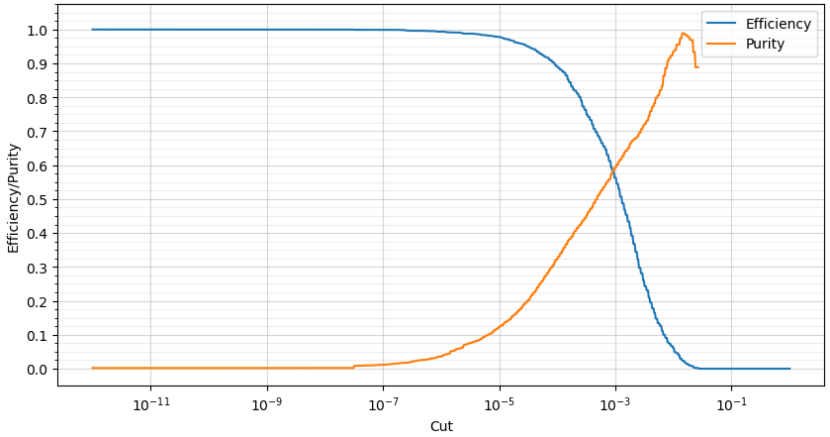
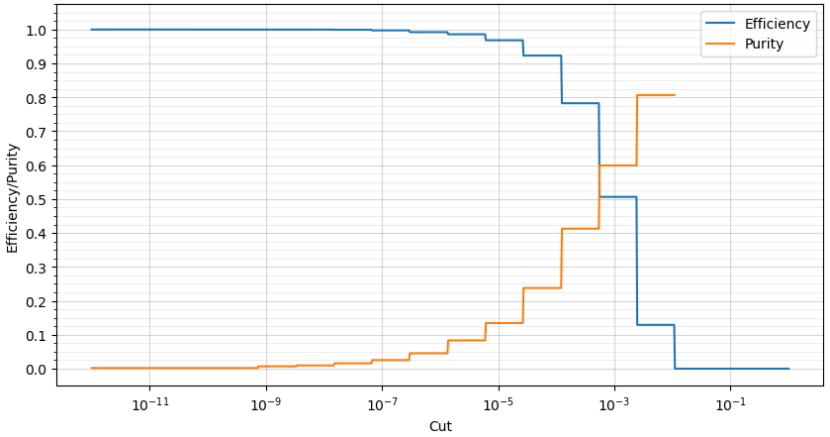
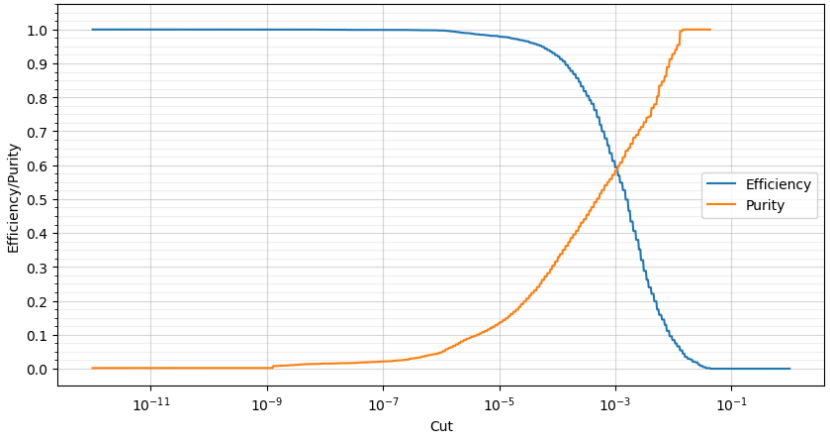
$$y = s_W(W_q - z_W) \cdot s_x(x_q - z_x) + b$$

$$y_{int} = (W_q - z_W) \cdot (x_q - z_x) + \left\lfloor \frac{b}{s_W s_x} \right\rceil$$

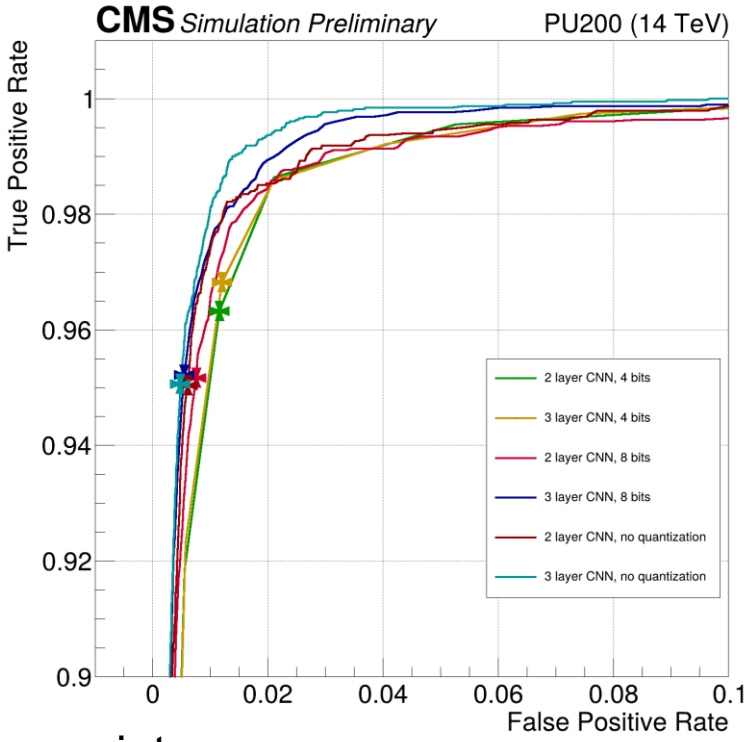
- Technique to simplifies the computation of deep learning model reducing memory usage and increasing processing speed
  - Reducing the number of bits in weight
  - Pruning multiple layer of network
- To fit in the required processing time, quantizing the model is essential

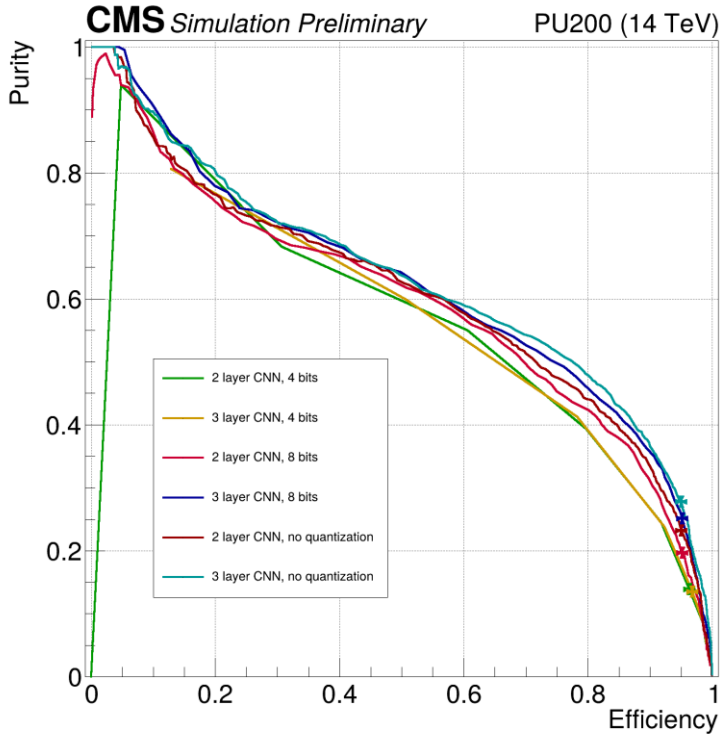
Quantization-aware Training

- Quantization-aware training for 2 and 3 layers CNN model with 4 and 8bits quantization
- Kernel size = (3, 5) for intermediate layer and (6, 5) for final layer, $N_{kernel} = 5$

	4bits Quantization	8bits Quantization
2 Layers CNN		
3 Layers CNN		

Quantization-aware Training

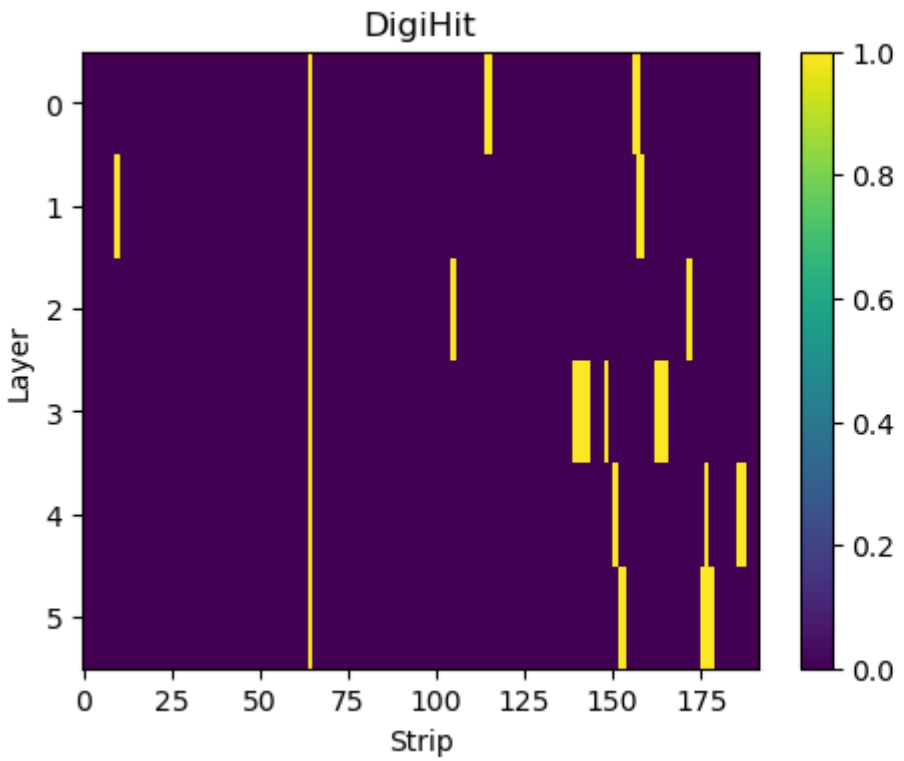




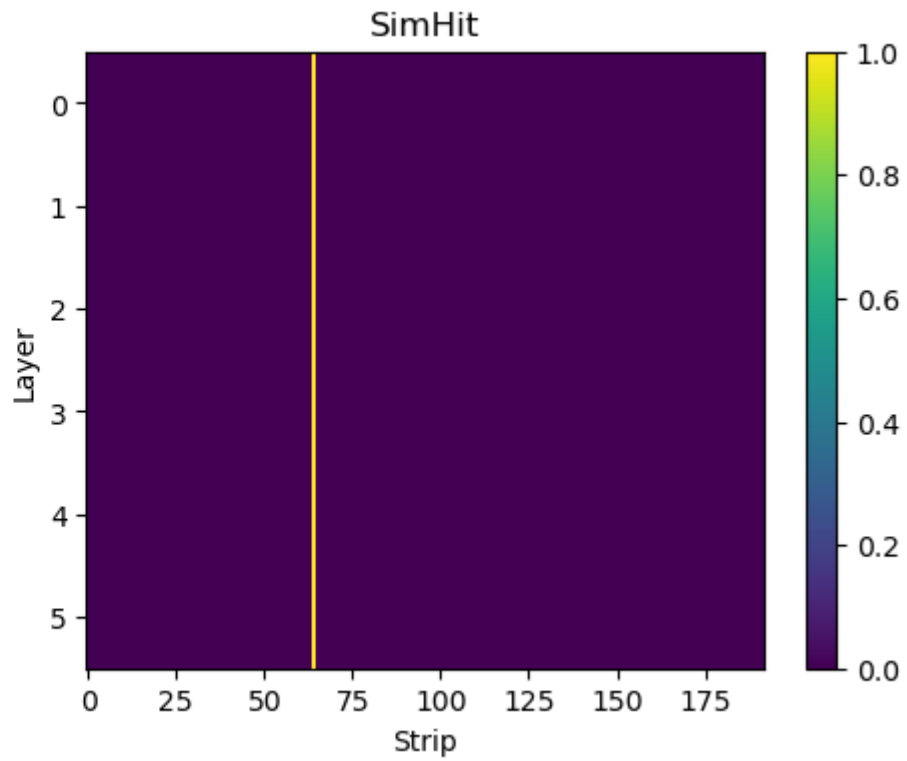
Using 95% efficiency point

	2 layer, 4bits	3 layer, 4bits	2 layer, 8bits	3 layer, 8bits	2 layer (no quantization)	3 layer (no quantization)
Efficiency	0.9632	0.9682	0.9515	0.9522	0.9504	0.9506
Purity	**0.1385**	**0.1346**	**0.1965**	**0.2508**	**0.2323**	**0.2777**

ML Out Example

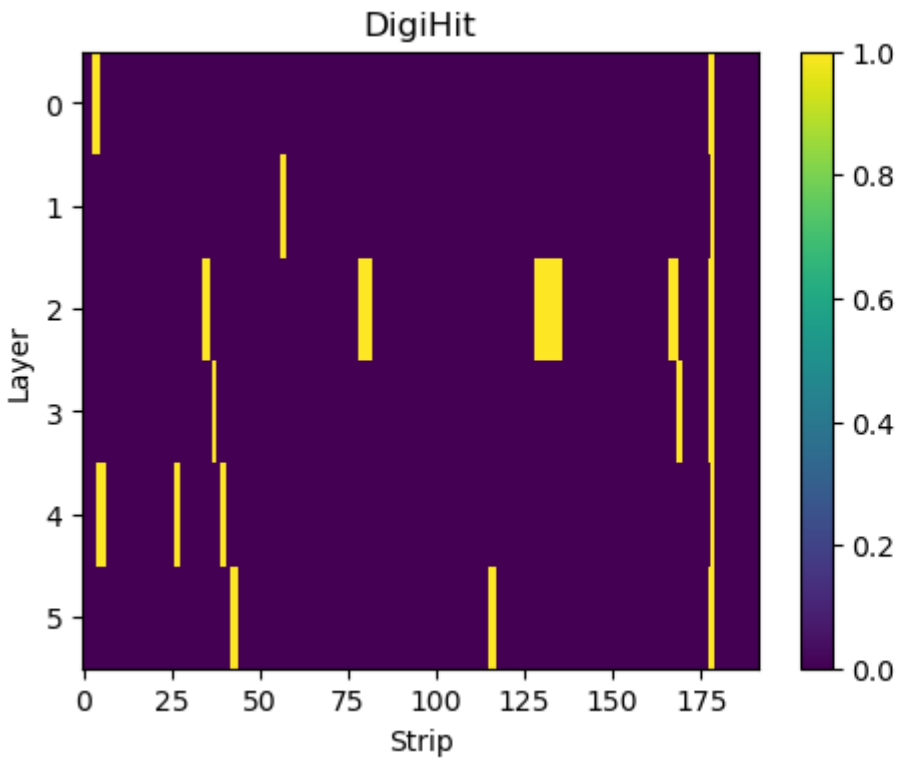


Seg Position from ML : 64, 65, 63

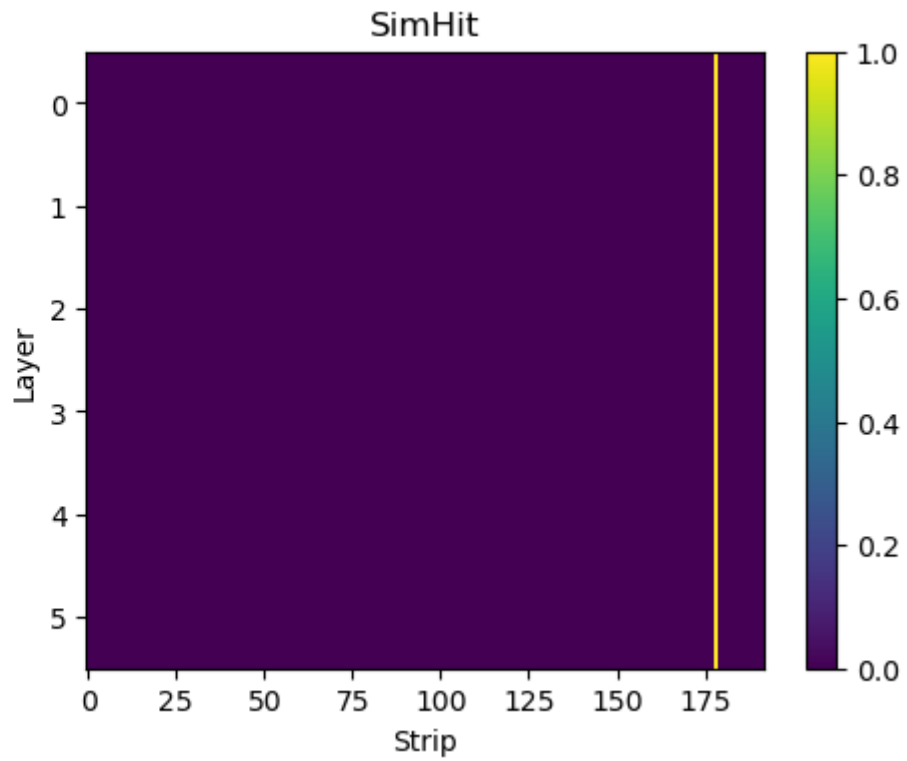


Real Track Position : 64

ML Out Example

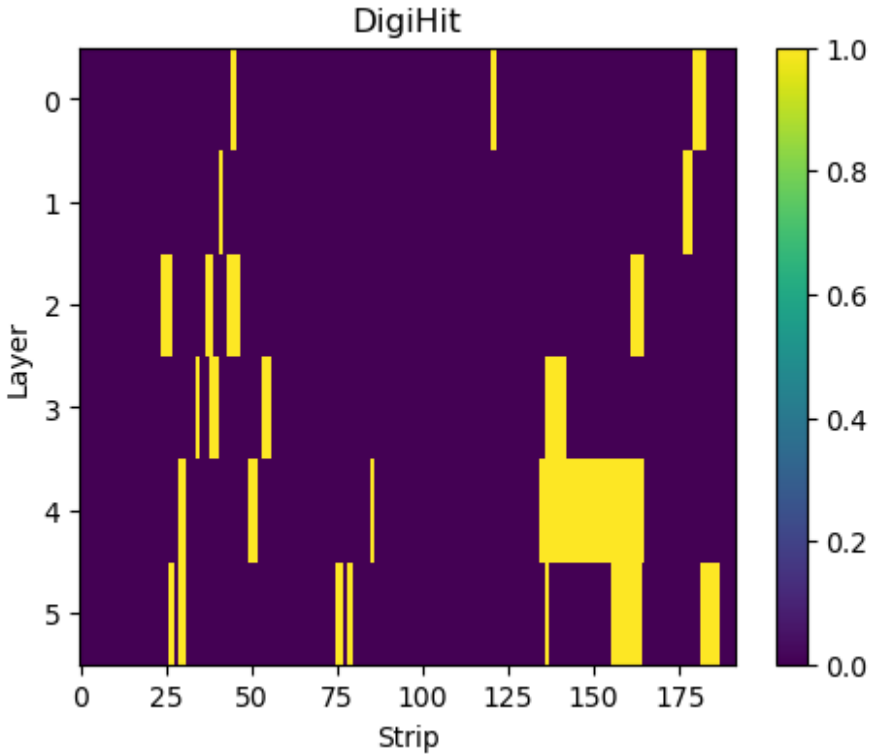


Seg Position from ML : 178, 177, 168

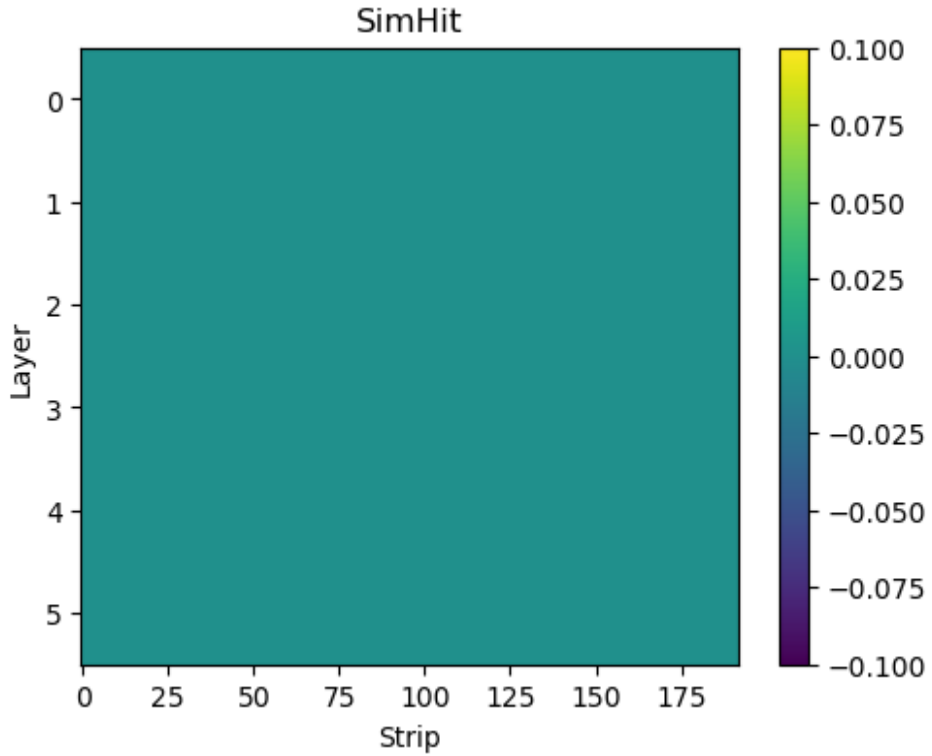


Real Track Position : 178

ML Out Example

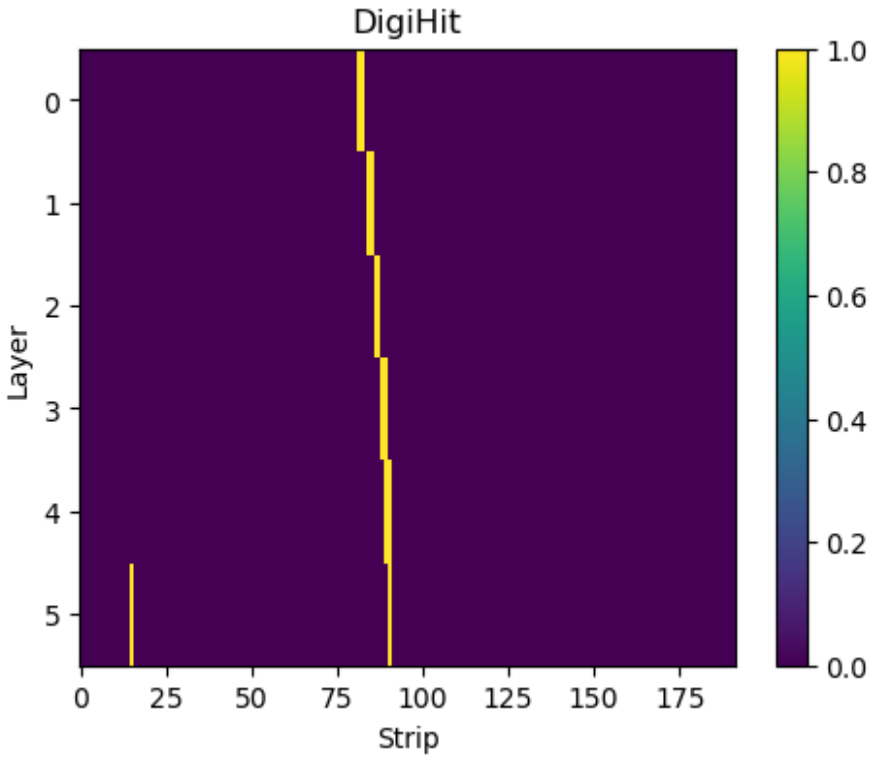


Seg Position from ML :

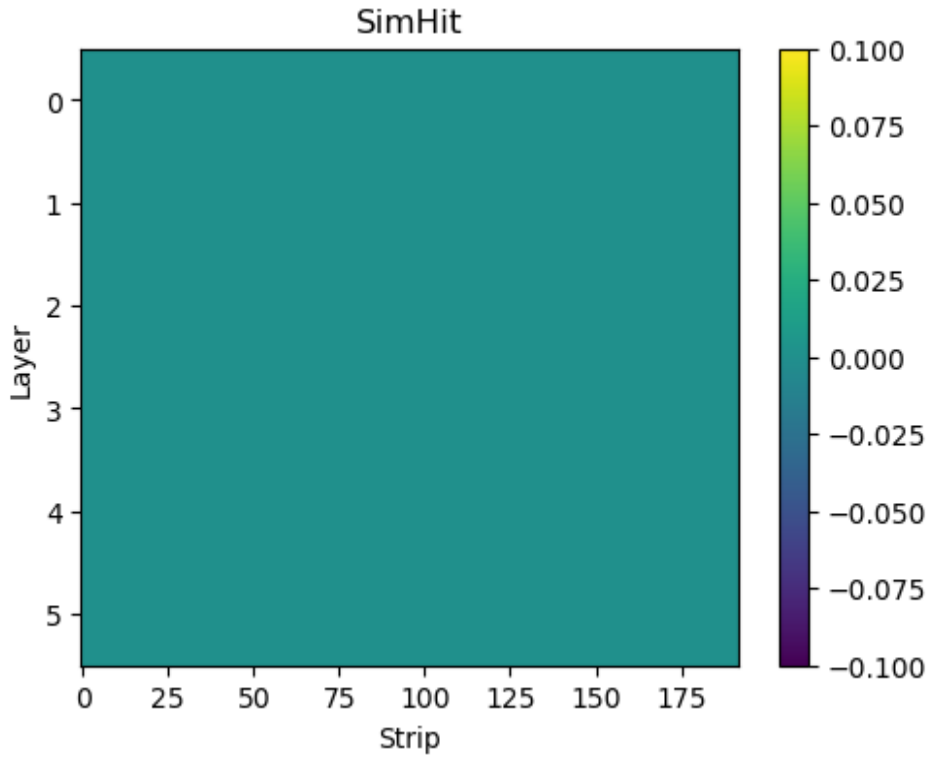


Real Track Position :

ML Out Example



Seg Position from ML : 88, 87



Real Track Position :