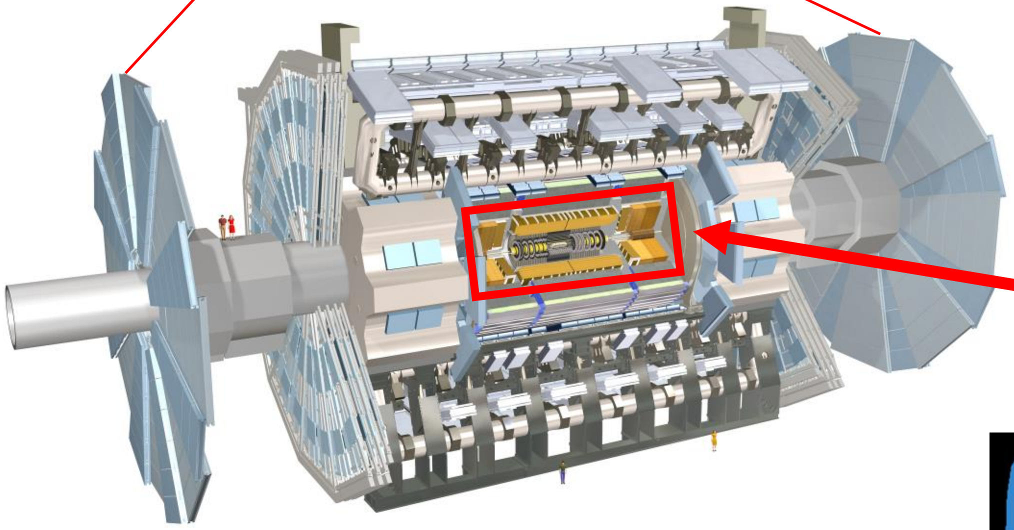
Machine Learning for Real-Time Processing of ATLAS Liquid Argon Calorimeter Signals with FPGAs

Anne-Sophie Berthold, Nick Fritzsche, Rainer Hentges, Alexander Lettau,
Arno Straessner, Johann Christoph Voigt

Fast Machine Learning For Science Workshop

London, 25th September 2023

The ATLAS Detector at the LHC



[1,2]

Large Hadron Collider (LHC):

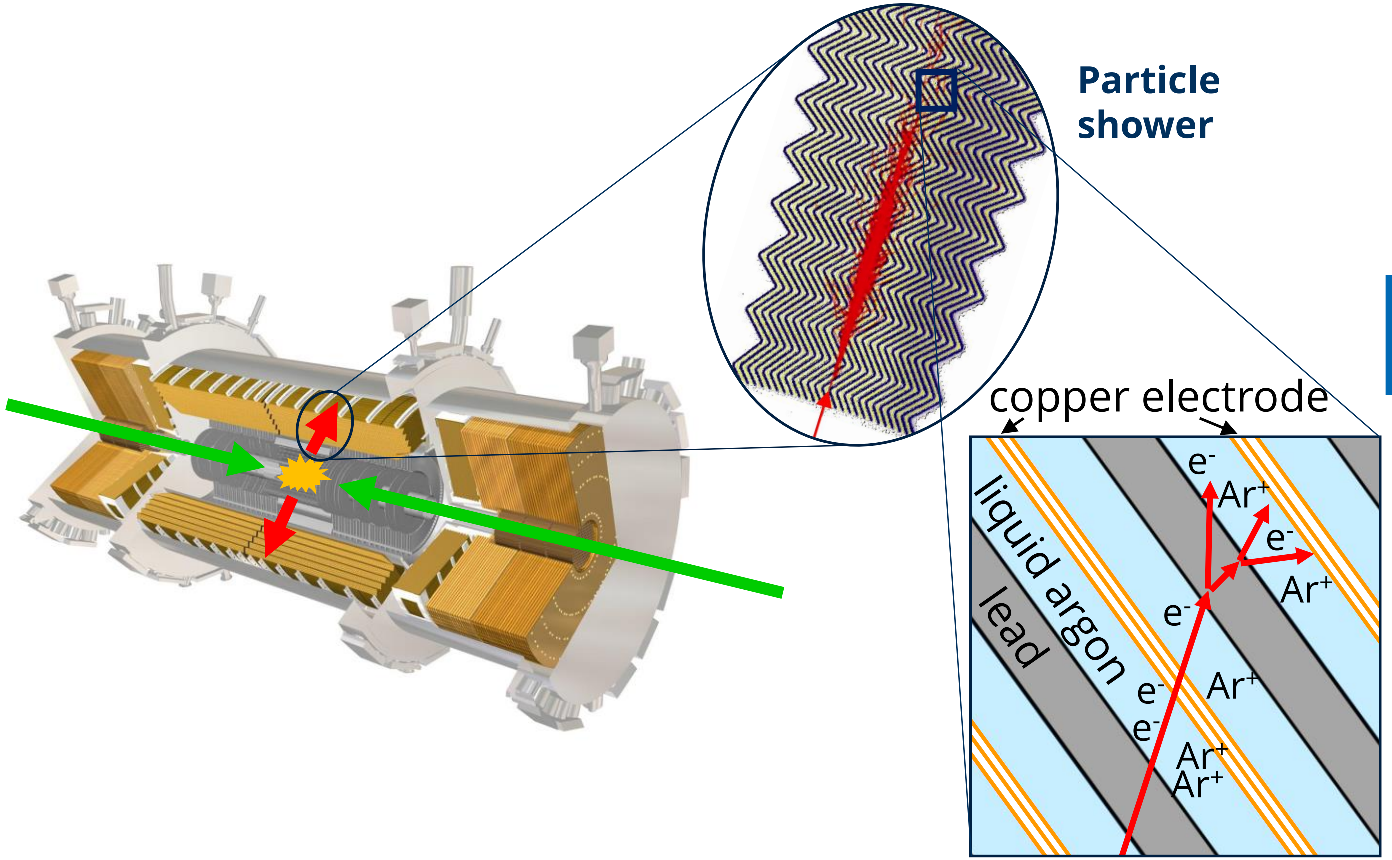
- Proton bunches collide with **25 ns** spacing (40 MHz)
- 2029: Start of **High Luminosity LHC** (HL-LHC) with up to ~7 x nominal luminosity

ATLAS Detector

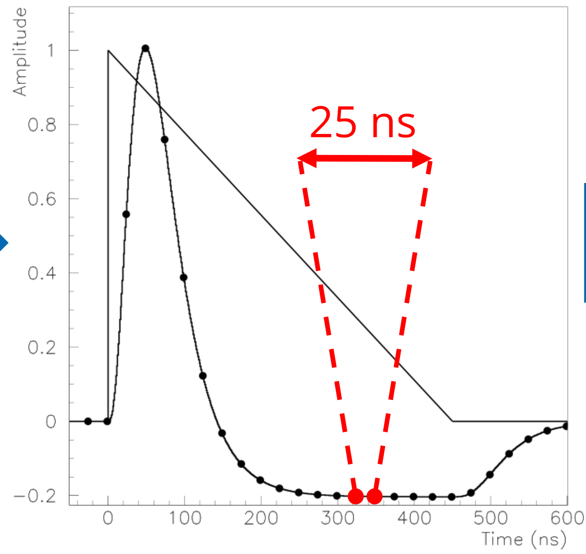
- From ~20 collisions to up to ~200 collisions per bunch crossing (BC) → pileup increases
- 2026-2028: Phase-II upgrade
- Readout electronics of **Liquid-Argon (LAr) calorimeter** need to be improved

[3]

LAr Calorimeter Readout



~182 000
LAr cells



…

[4,5,6]

- Absorber (Pb, Cu, W) and electrodes in accordeon geometry
- Active medium: Liquid Argon (**LAr**) in between

- Free electrons raise **triangular pulse**
- Shaped into **bipolar pulse** and digitized
- **Parameter of interest: Amplitude** proportional to deposited energy

LAr Calorimeter Readout

FPGA – Field Programmable Gate Array

[7]

- Real-time signal processing
- Installation of **556 high-performance FPGAs**
- 250 Tbps total data rate
- 150 ns maximum latency

Optimal Filtering Algorithm (OF)

- calculates deposited energy per cell

$$y_t = \sum_{n=0}^{M-1} x_{t-n} \cdot a_n$$

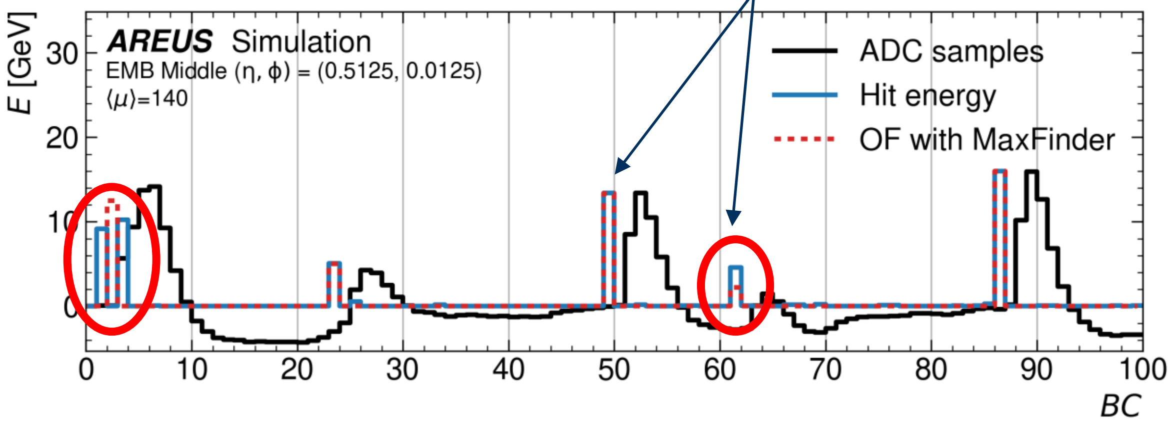
y_t … OF output for bunch crossing (BC) t
$\vec{x}$ … input ADC samples
a_n … OF coefficients
M … OF filter depth

- Trigger system applies additional maximum finder

Trigger

Event Readout

- Good in energy resolution **but**
- Weak in reconstruction of <u>overlapping signals</u>

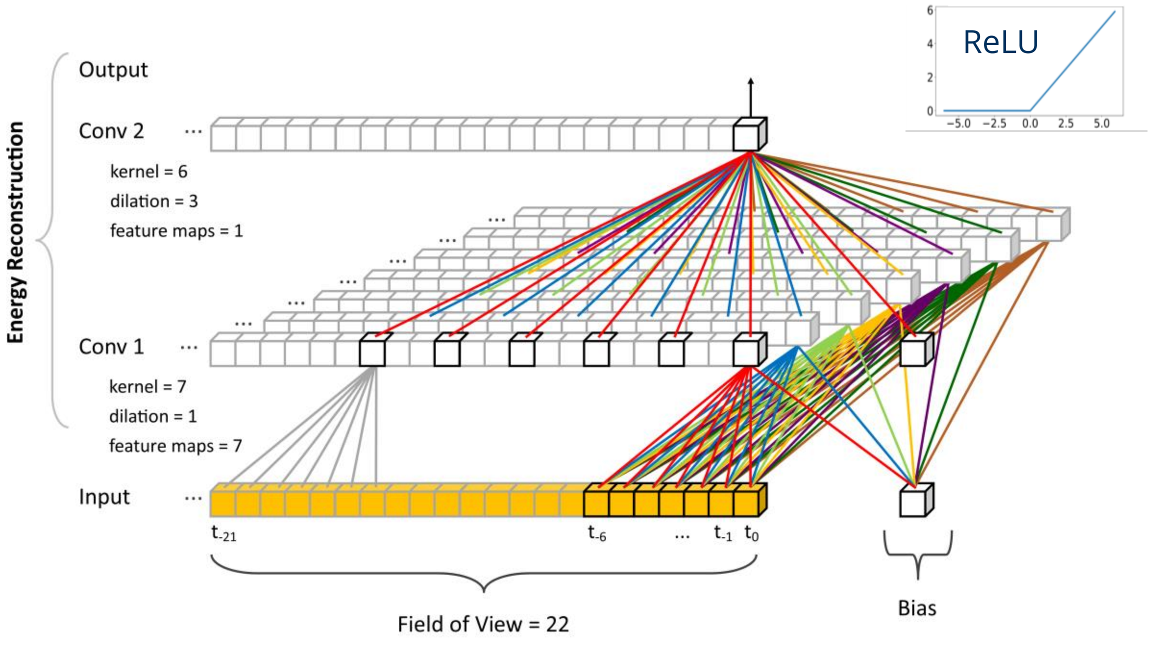


- ➢ Development of Neural Networks to improve energy reconstruction
- ➢ Keep parameters low (≈ 100) due to FPGA resource limits

Convolutional Neural Networks for LAr Readout

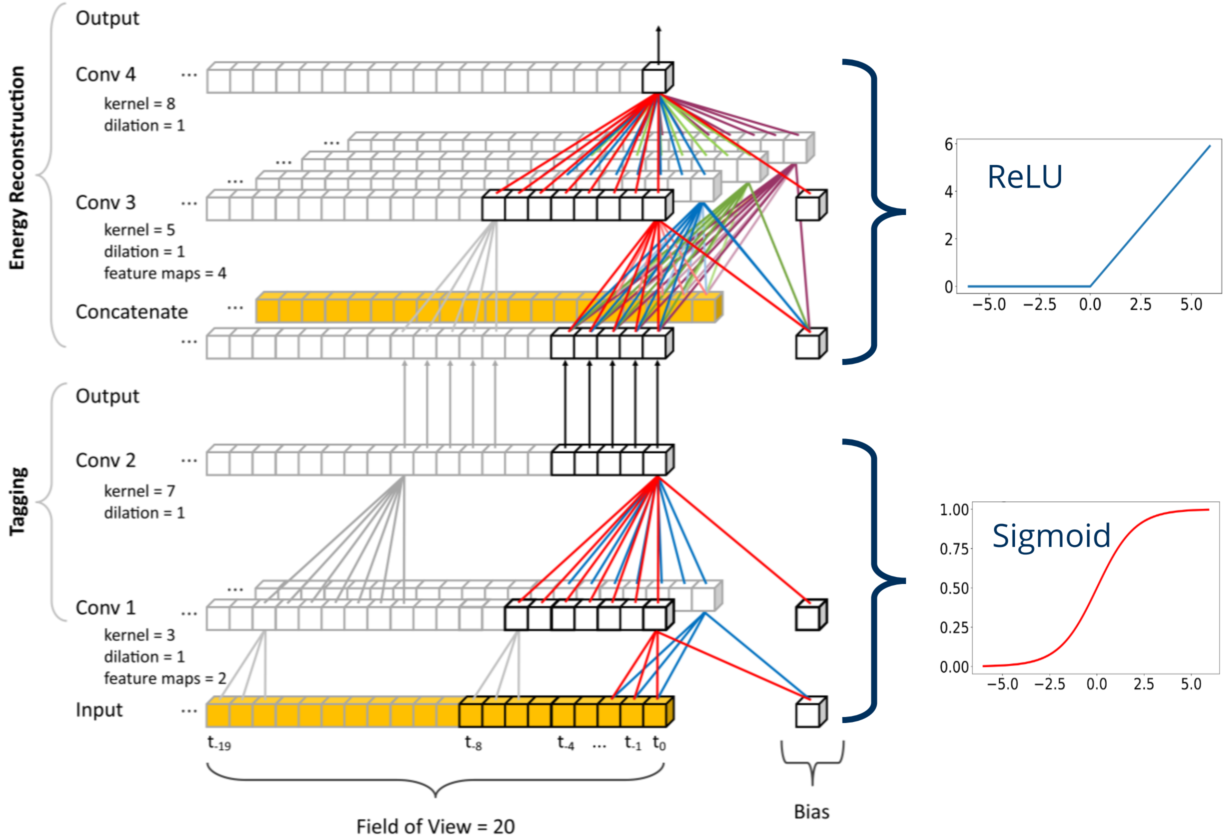
Plain 2-layered CNN (2-Conv CNN)

- Dilation enables larger Field of View (FoV)
- ReLU activation functions
- Output: reconstructed energy



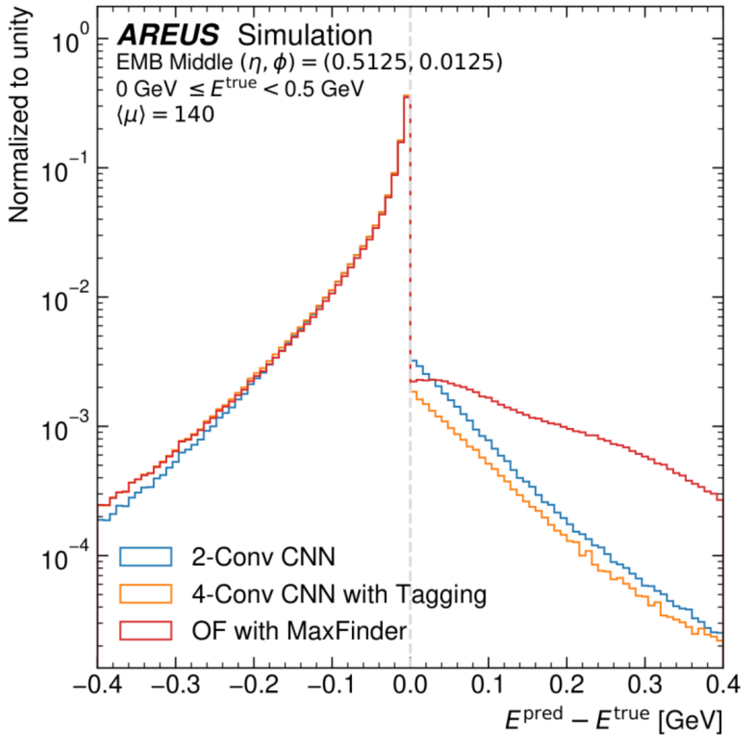
4-layered CNN with Tagging (4-Conv CNN)

- Sigmoid and ReLU activation functions
- Intermediate output tags signal overlaps
- Output: reconstructed energy

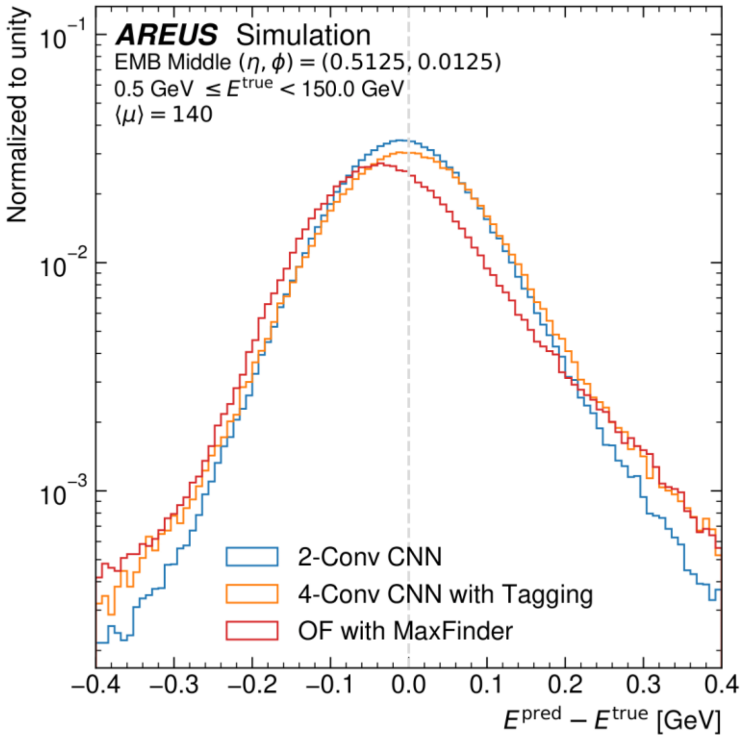


Performance Evaluation: Energy Resolution

$E^{true} < 0.5$ GeV



$E^{true} > 0.5$ GeV



Optimal Filter:

- Larger deviation spread in low energy region
- Negative bias in high energy region

CNNs:

- Improvement in energy resolution, more symmetric and centered distribution
- Performance stable within large energy range

Performance Evaluation: Sequence Comparison

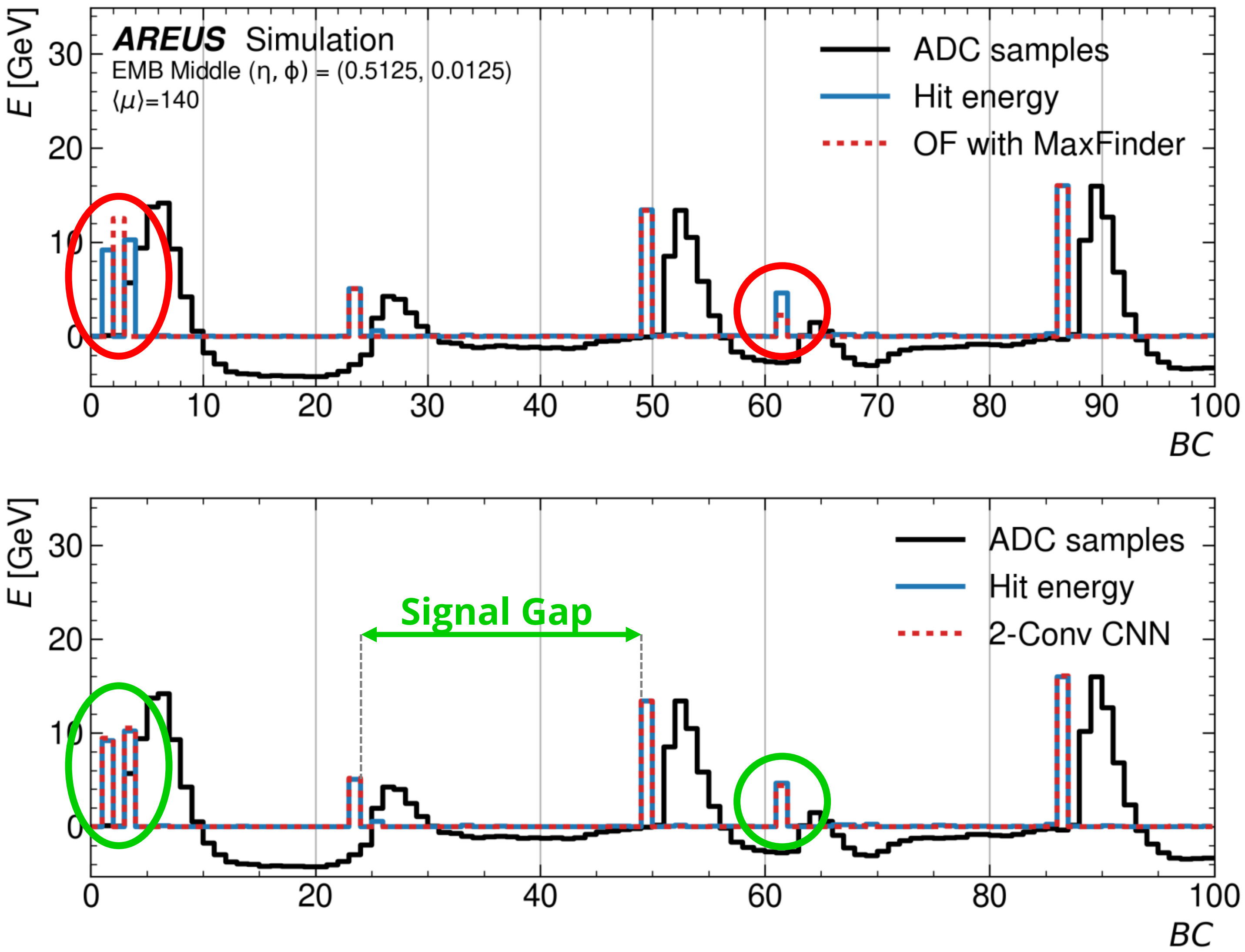
Example Sequence

Optimal Filter

- Close signals cannot be resolved
- Signals within undershoot underestimated

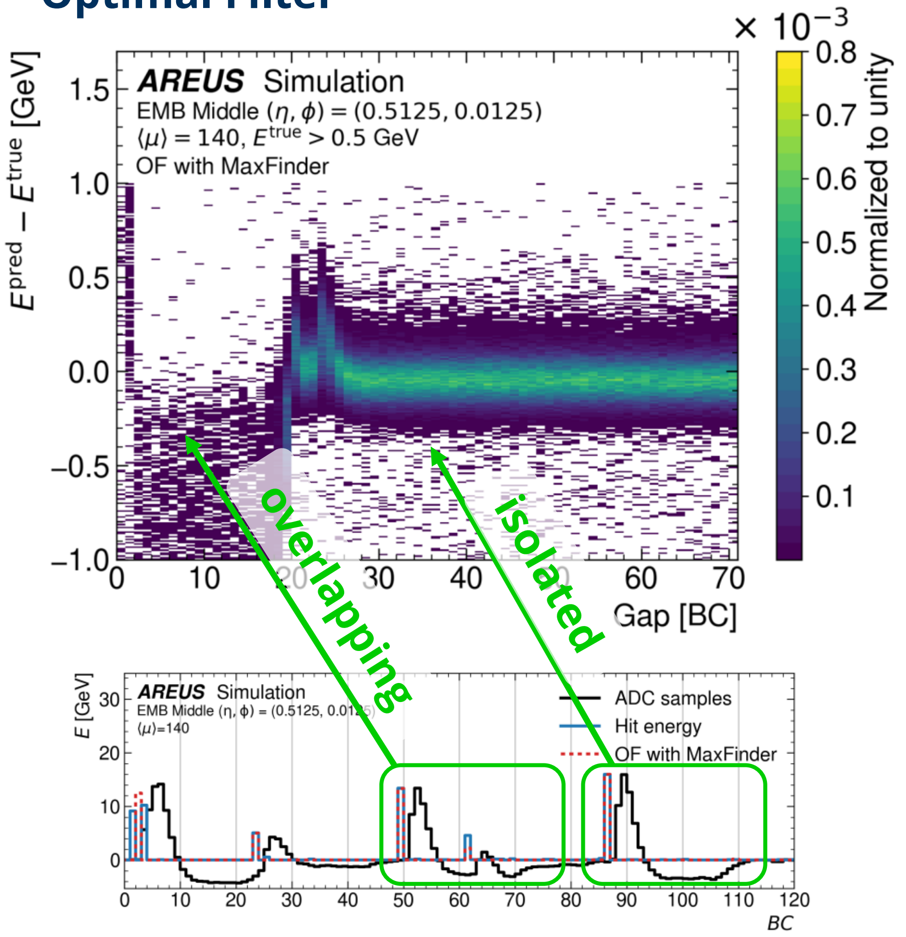
CNN

- Optimized to reconstruct overlapping signals

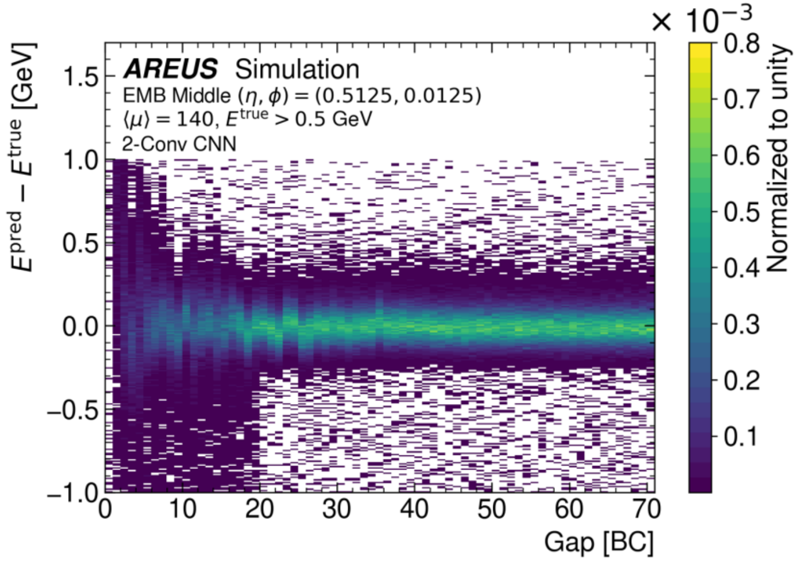


Performance Evaluation : Energy Reconstruction as Function of Gap

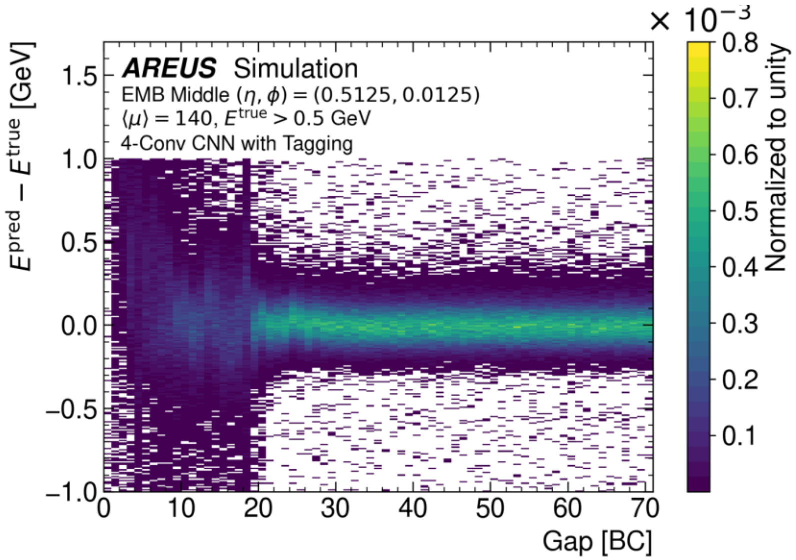
Optimal Filter



2-Conv CNN

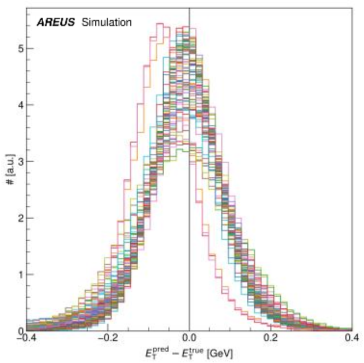


4-Conv CNN with Tagging



- OF struggles with overlapping pulses (gap < 25 BC)
- CNNs show improvement

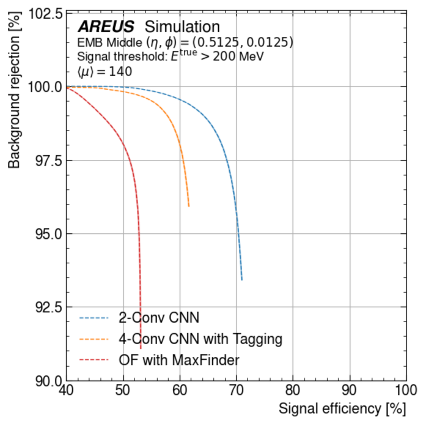
Performance Evaluation - Combining all: Star Plot

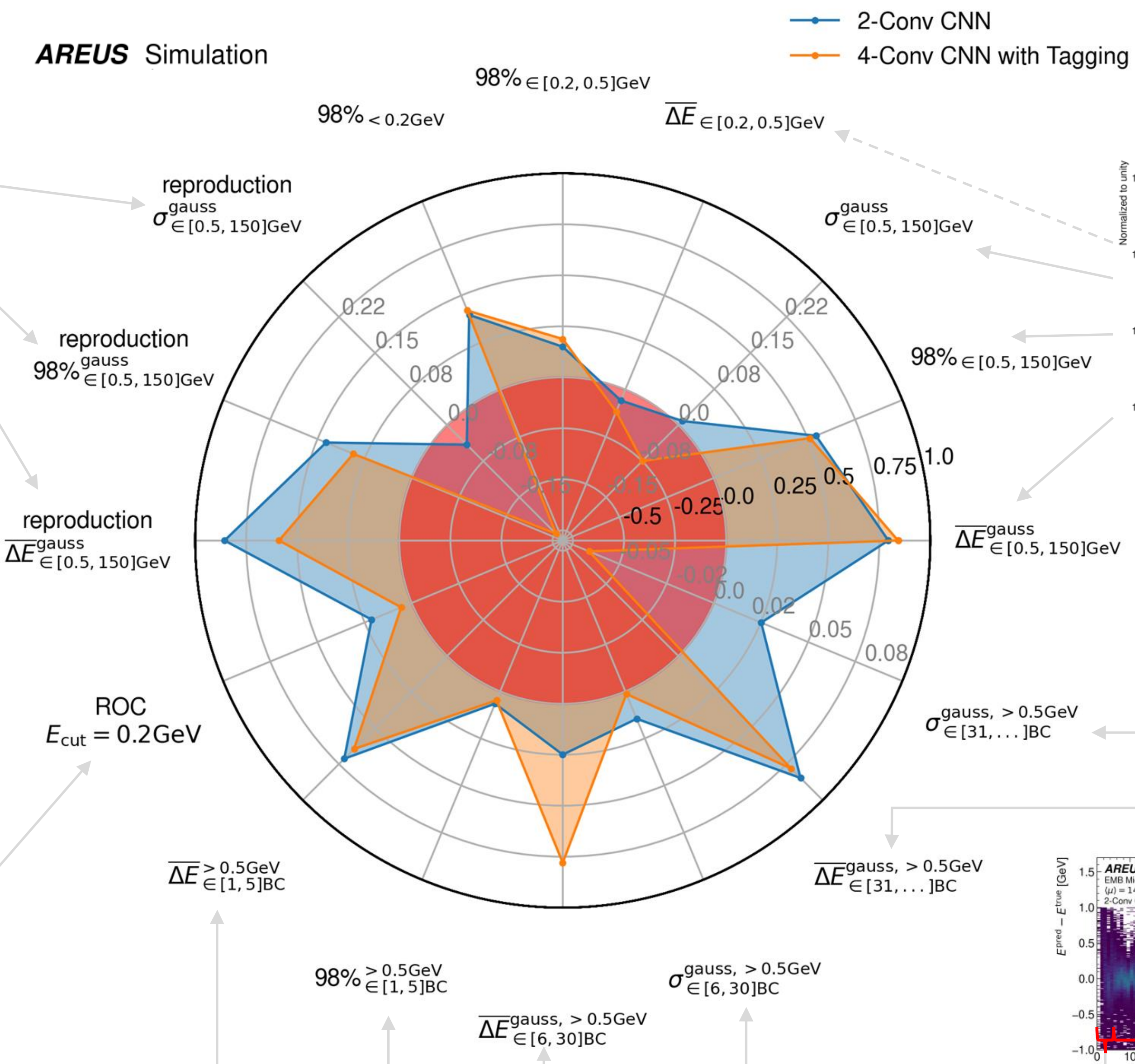


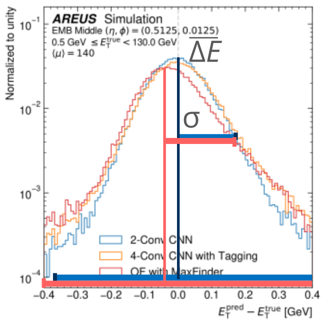
Score indicates CNN performance:

$$sc(X) = 1 - \frac{X\_ANN}{X\_OF}$$

- **Red circle:** (=0) OFMax yield
- **Outer circle:** (=1) best yield
- Inside circle (<0): worse than OF

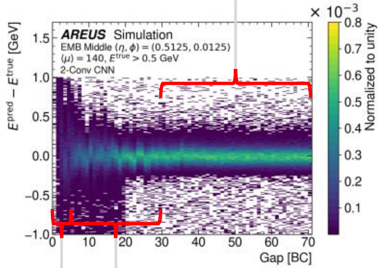






98 % range

- ✓ Performance overview
- *x* No replacement for other plots as details might be hidden



TECHNISCHE UNIVERSITÄT DRESDEN

Firmware Implementation

ANN model from Keras

[8,9]

Converter

FPGA

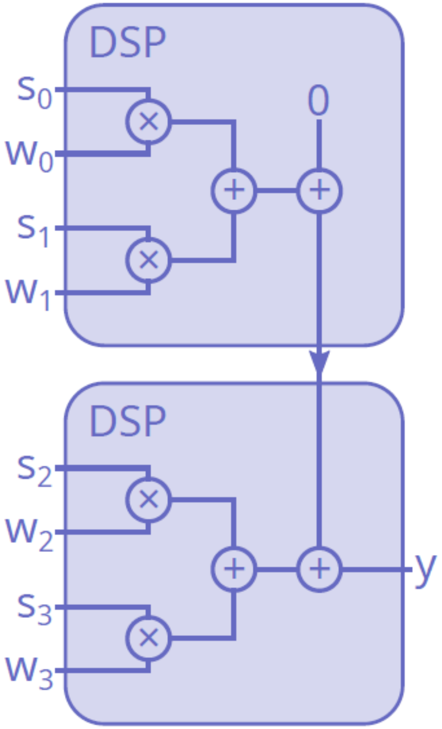
[7]

Integrated circuits, configurable by user after manufacturing

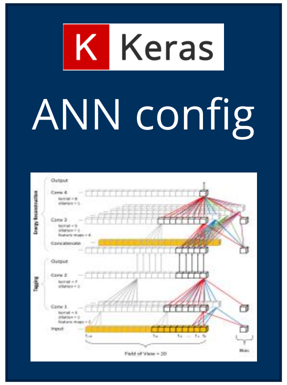
- **DSP:** digital signal processor
- **ALM:** adaptive logic module with lookup table

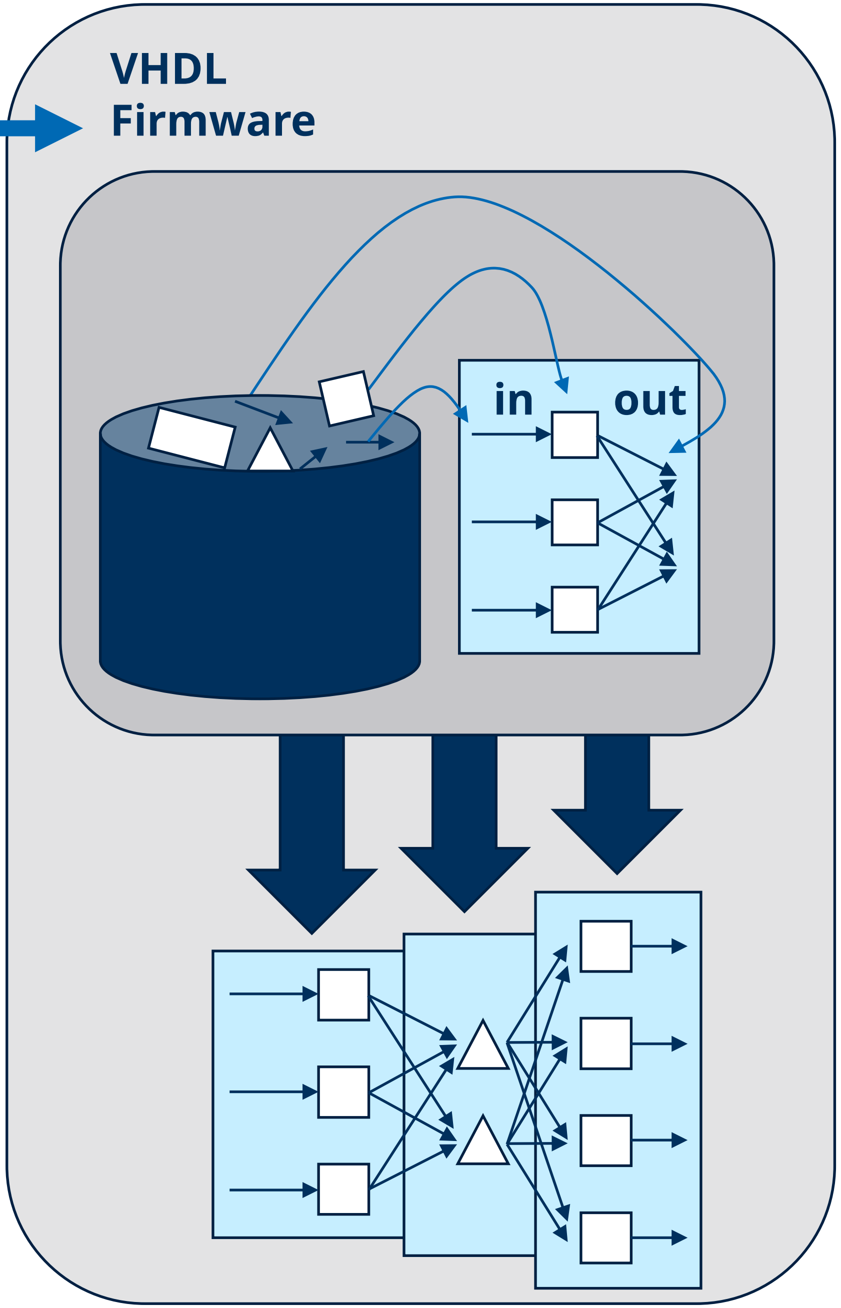
Firmware development on **Intel Stratix-10** FPGA

Final design uses **Intel Agilex**



CNN Firmware Implementation





- Fully configurable CNN network implementation in **VHDL**
  - Layer building blocks with configurable
    - Inputs
    - Outputs
    - Activation functions
  - Chaining of components with configurable
    - Kernel sizes
    - Filters per layer
    - Dilation
- Parameter automatically extracted from Keras output files
- Calculation in 18 bit fixed point numbers
- Supports pipelining and time division **multiplexing**:
  - ➢ Design runs at 12x ADC frequency with cyclic processing of 12 detector cells

FPGA Resource Estimation

- Trigger latency requirement ≈ 150 ns
- Need to process 384 detector cells on 1 FPGA
  - ➢ E.g. 12-fold multiplexing with 33 parallel instances

} ✓ Achieved by all compiled firmwares

- Resource estimates (based on Intel Quartus reports):

FPGA	Network	Multiplexing	Detector cells	$f_{\max}$	ALMs	DSPs
Stratix-10	2-Conv CNN	12	396	415 MHz	8 %	28 %
	4-Conv CNN	12	396	481 MHz	18 %	27 %
Agilex	2-Conv CNN	12	396	539 MHz	4 %	13 %
	4-Conv CNN	12	396	549 MHz	9 %	12 %

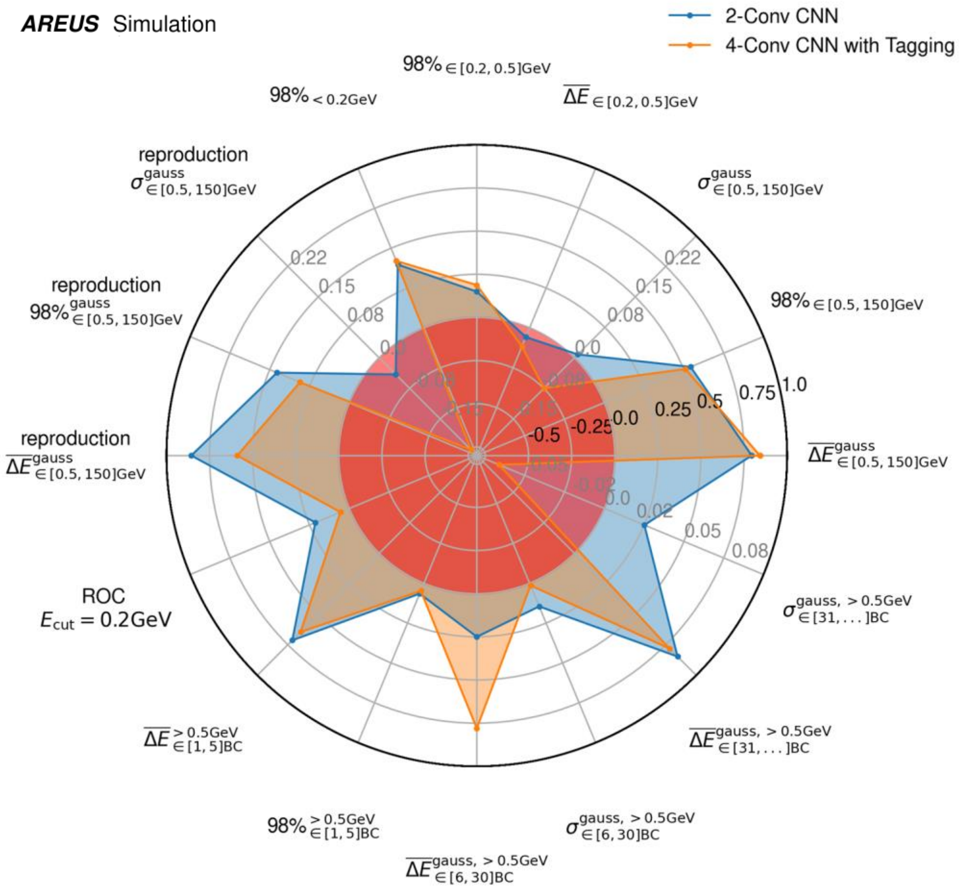
Summary and Outlook

Summary

- CNNs are able to replace Optimal Filtering algorithm
- CNNs show good performance results in energy resolution and especially on signal overlaps
- Firmware implementation of CNNs with VHDL
- Resource requirements regarding latency and bandwidth are satisfied

Outlook

- Further improvements by applying quantization aware training and more CNN parameters ($100 \rightarrow 400$)
- Study for influence of energy reconstruction by CNNs for full event reconstruction
- Further tests on FPGA hardware ongoing



Thank you for your attention!

Sources I

Slide 2:

[1] URL: https://static1.bmbfcluster.de/3/4/3/8_ef6a5eef8f44963/3438meg_22ce2885dae52af.jpg.

[2] Joao Pequenao. *Computer generated image of the whole ATLAS detector.* CERN. Mar. 27, 2008. URL: https://cds.cern.ch/record/1095924 (visited on 05/10/2023).

[3] Peter Vankov, *ATLAS Upgrade for the HL_LHC: meeting the chalenges of a five-fold increase in collision rate.* CERN. Jan. 25, 2012. URL: https://cds.cern.ch/record/1419213/ (visited on 05/10/2023).

Slide 3:

[4] Joao Pequenao. *Computer generated image of the ATLAS Liquid Argon.* CERN. Mar. 27, 2008. URL: https://cds.cern.ch/record/1095928 (visited on 05/17/2023).

[5] Karl Jakobs. Lecture Material. CERN. 2015. URL: https://www.particles.uni-freiburg.de/dateien/vorlesungsdateien/particledetectors/kap8

[6] ATLAS Collaboration. *Monitoring and data quality assessment of the ATLAS liquid argon calorimeter.* CERN. May 13, 2014. URL: https://cds.cern.ch/record/1701107 (visited on 05/24/2023).

Slides 4, 10:

[7] Intel. *Stratix 10 FPGA*. URL: https://newsroom.intel.com/editorials/intels-stratix-10-fpga-supporting-smart-connected -revolution (visited on 04/18/2021).

Sources II

Slides 10, 11:

[8] *Keras Logo*. URL: https://keras.io/ (visited on 05/25/2023)

[9] *Tensorflow Logo*. URL: https://www.vectorlogo1.zone/logos/tensorflow/index.html (visited on 05/25/2023)

Papers related to these slides:

- Georges Aad et al. *Artificial Neural Networks on FPGAs for Real-Time Energy Reconstruction of the ATLAS Lar Calorimeters*. In: Computing and Software for Big Science 5.1 (Oct. 2021) DOI: 10.1007/s41781-021-00066-y. URL: https://doi.org/10.1007/s41781-021-00066-y.

Convolutional Neural Networks (CNNs)

- Convolutional operation with certain **kernel** size
- **Activation function** gives opportunity to classify, weight, cut

$$y_t = A\left(\sum_{i=0}^{n} x_i \cdot w_i + b\right)$$

y_t ... node output for time t

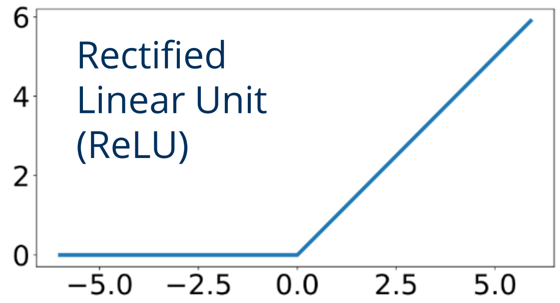
$\vec{x}$ ... input samples

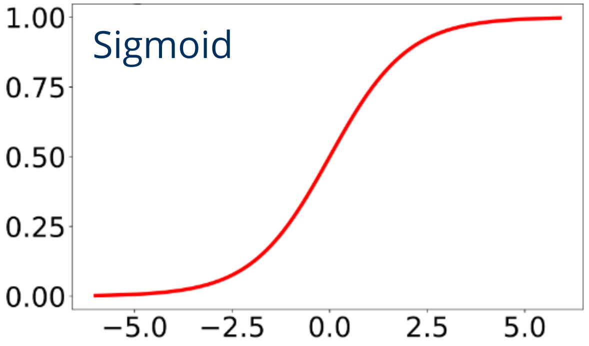
$\vec{w}$ ... weights

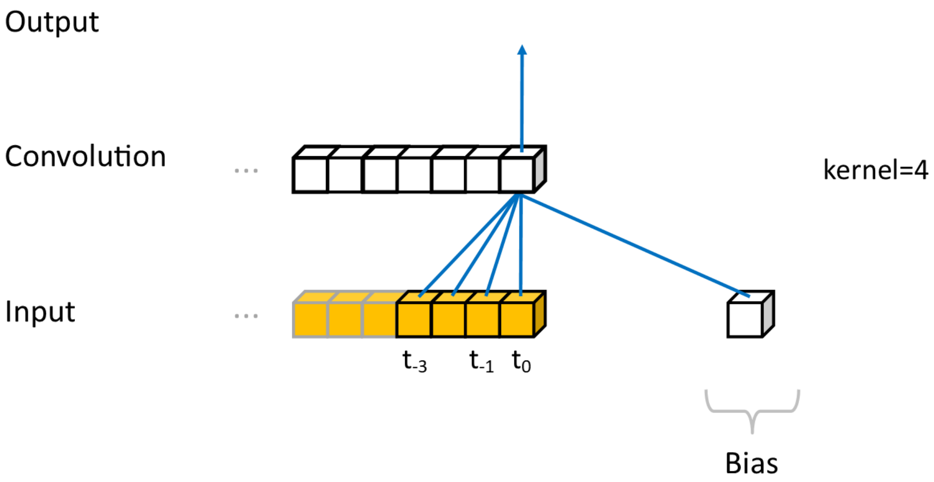
b ... bias term

A ... activation function

→ **similar to OF**

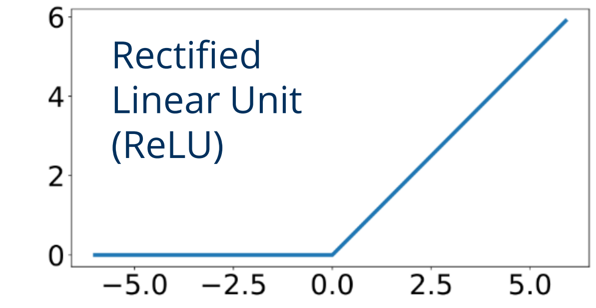


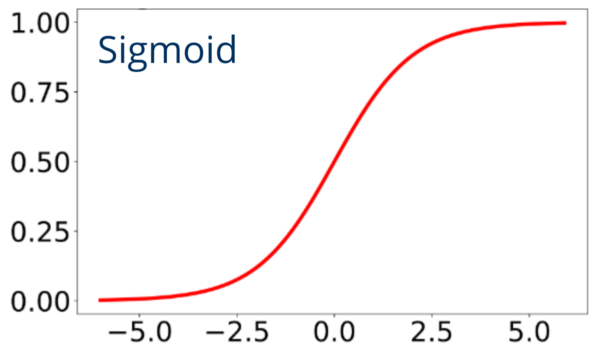


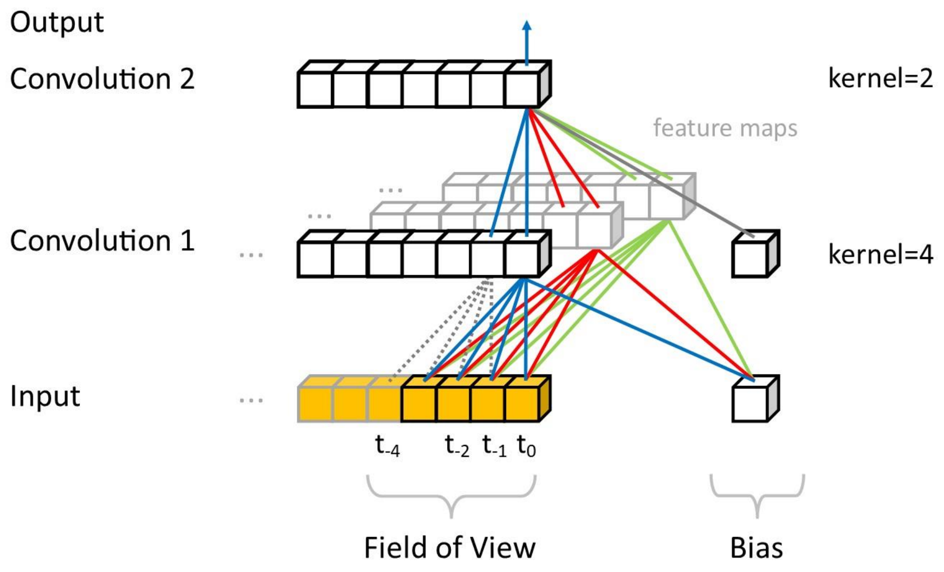


Convolutional Neural Networks (CNNs)

- Convolutional operation with certain **kernel** size
- **Activation function** gives opportunity to classify, weight, cut
- **Feature maps** focus on different properties
- **Training** minimizes difference between output and target

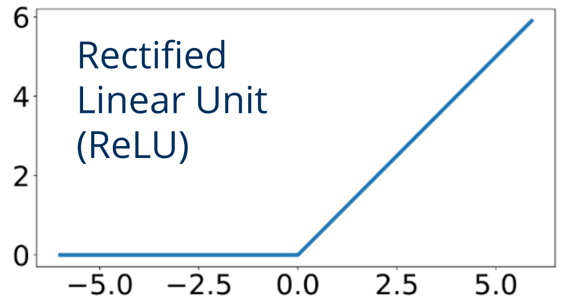


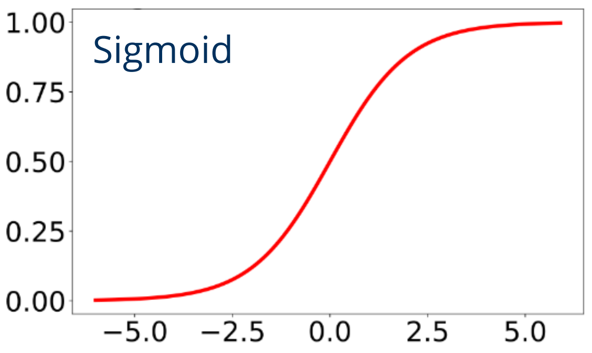


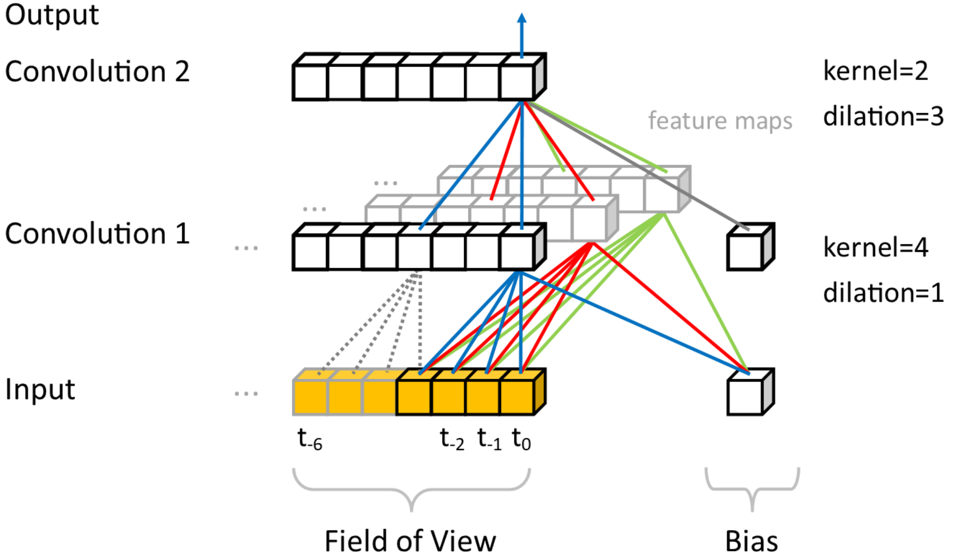


Convolutional Neural Networks (CNNs)

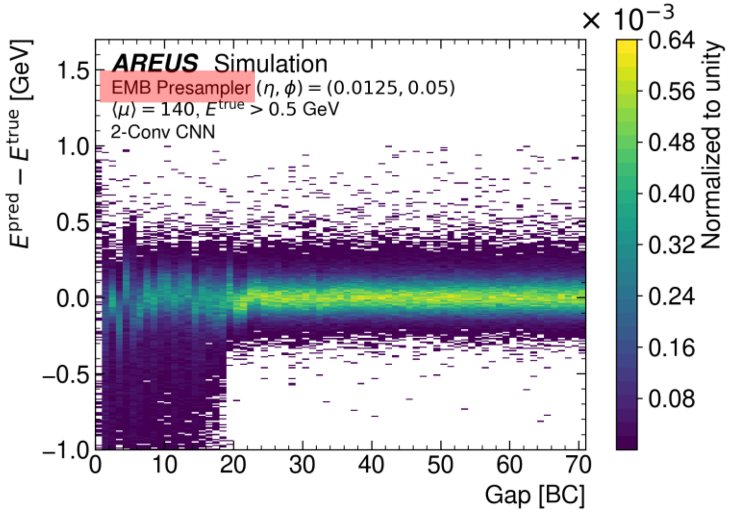
- Convolutional operation with certain **kernel** size
- **Activation function** gives opportunity to classify, weight, cut
- **Feature maps** focus on different properties
- **Training** minimizes difference between output and target
- **Dilation** varies field of view (FoV) without increasing parameters
- **Keep parameters low** (≈ 100 /≈ 400) and FoV realistic (≤ 24) due to FPGA implementation

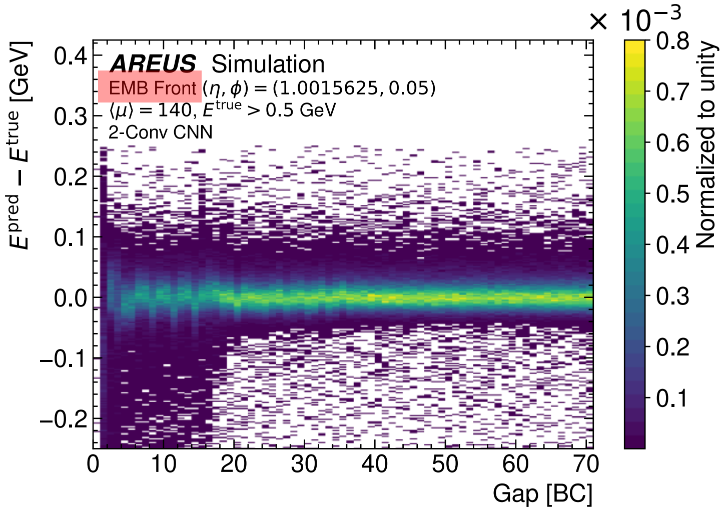


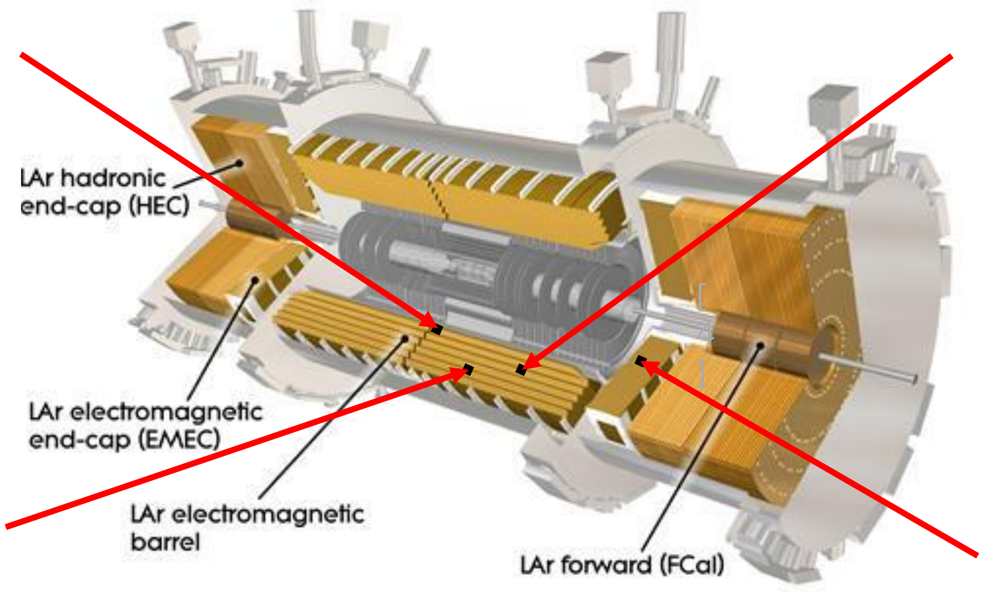


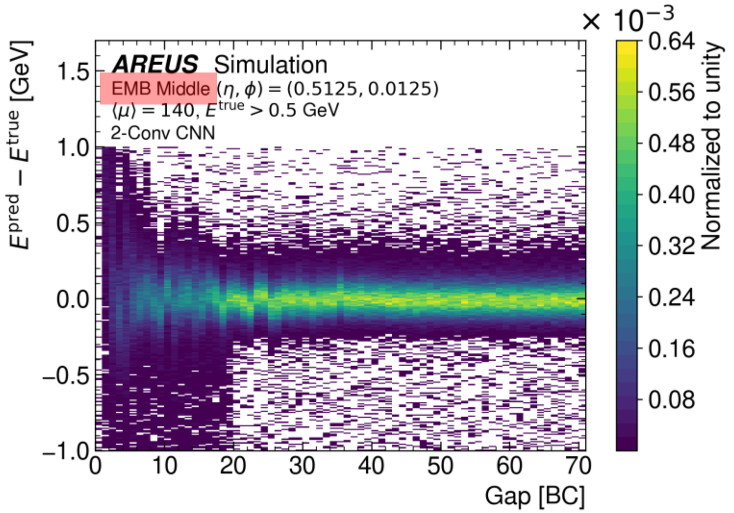


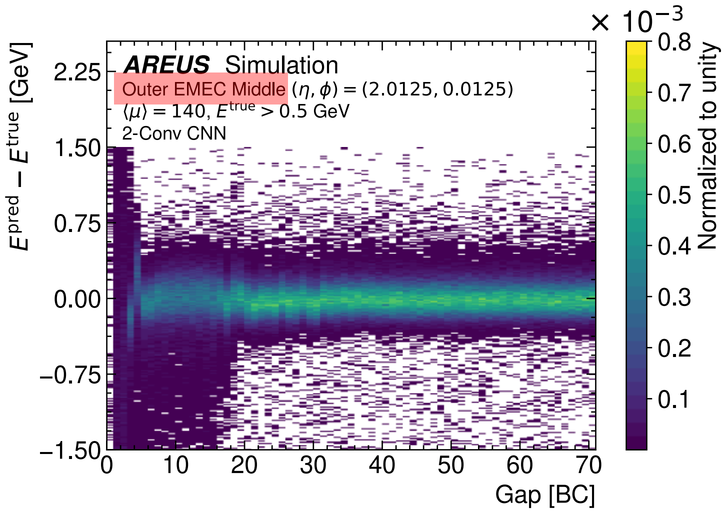
Performance Evaluation: Different Detector Regions





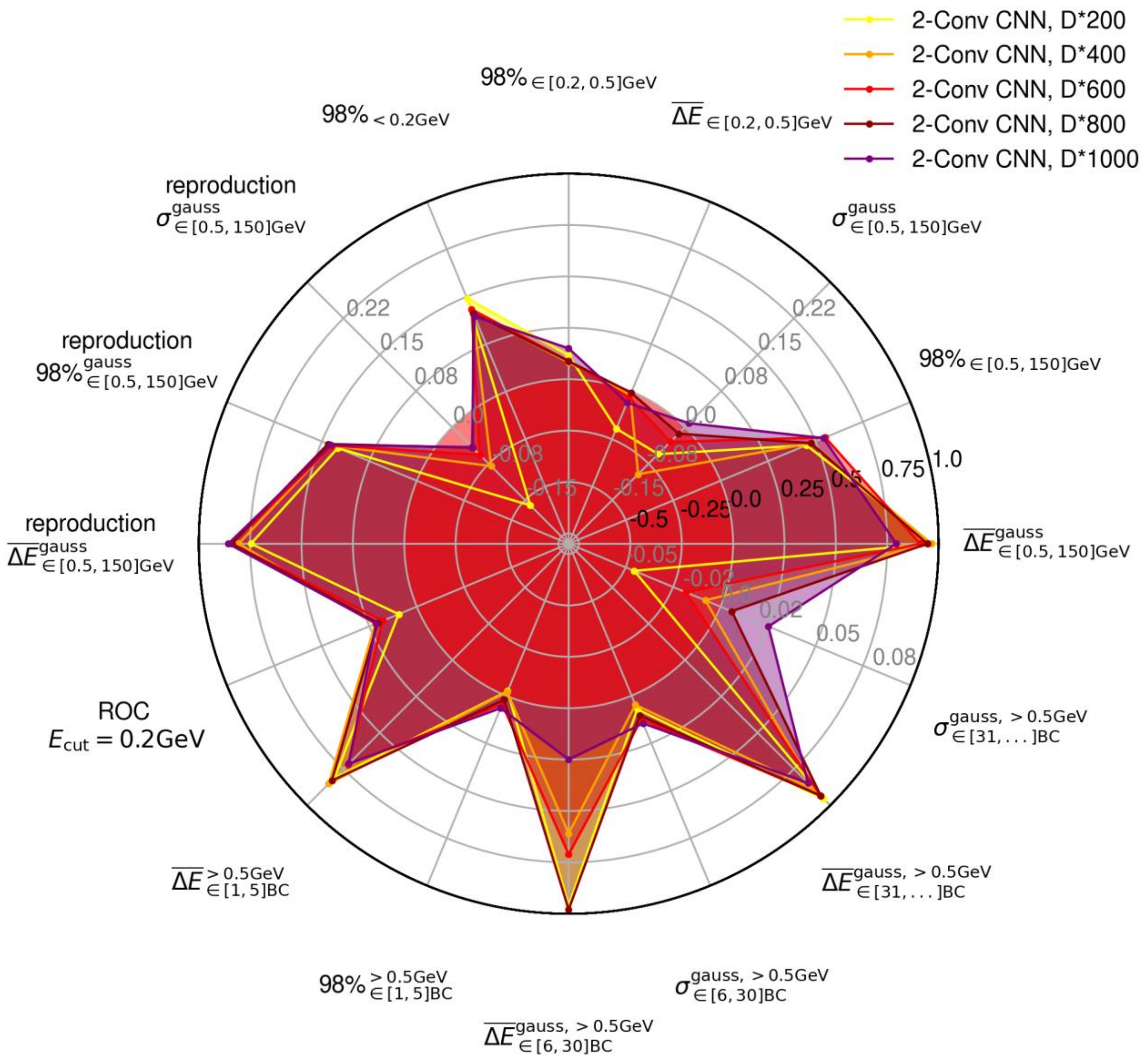






➢ Same architecture trained for different detector regions → shows similar results

Studying Influence of Size of Training Dataset with the Star Plot



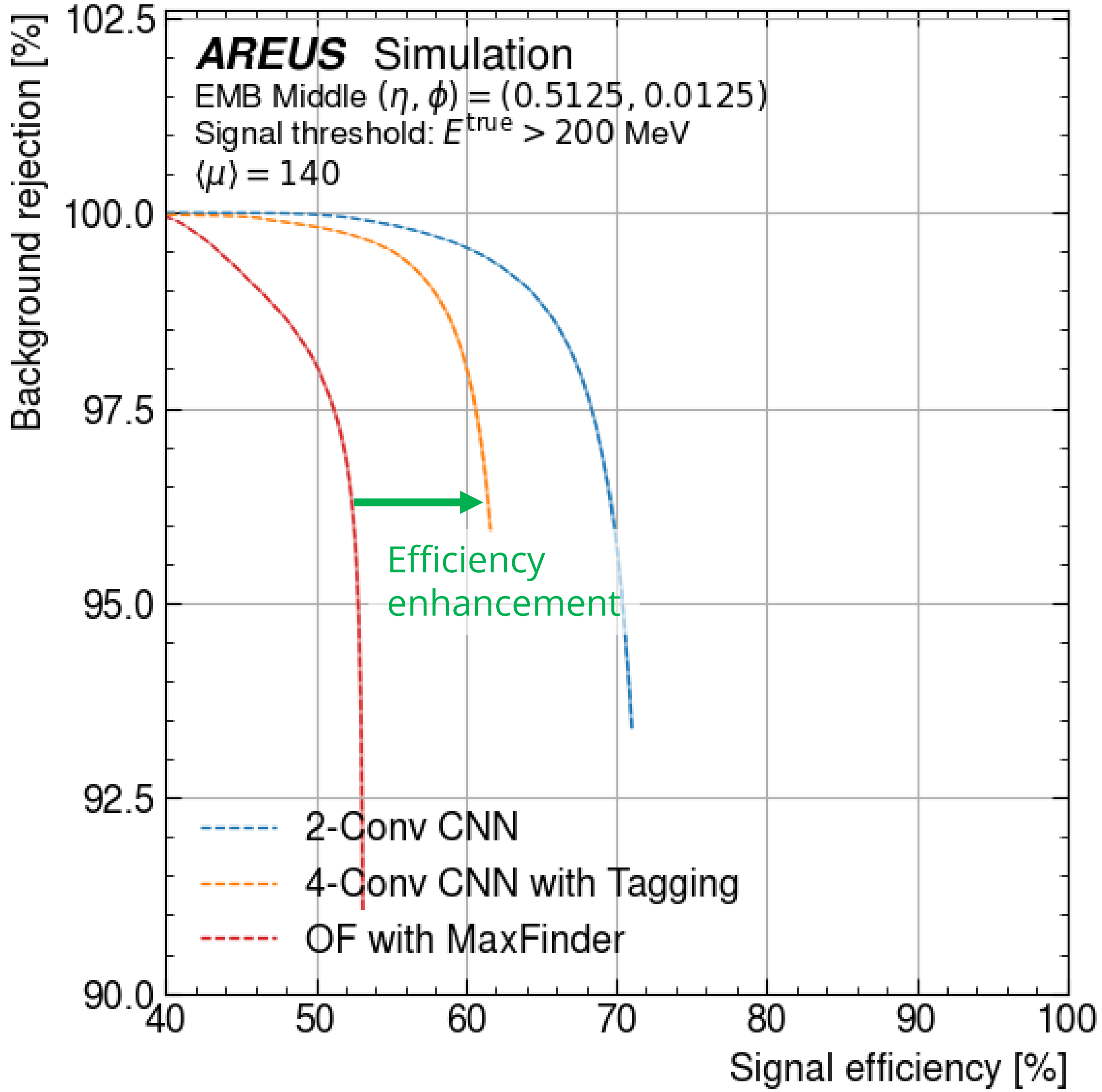
- Training dataset consists of several sub-datasets that hold different scenarios

- Study influence of dataset size by enhancing all sub-datasets D equally:
  - [200, 400, 600, 800, 1000]*D for each scenario

➢ Some scores not affected

➢ For others: at least 600*D for each

Performance Evaluation: Signal Efficiency vs Background Rejection



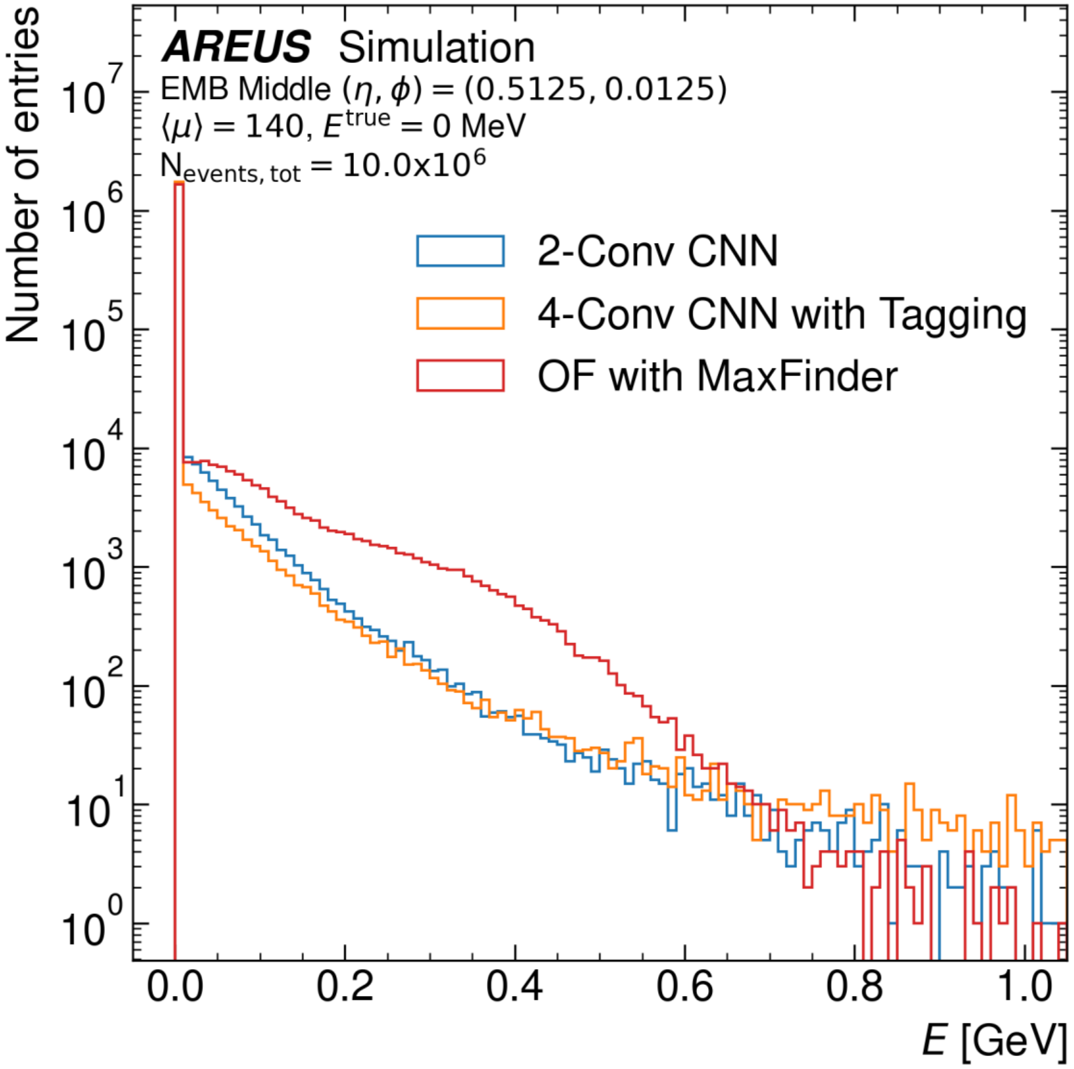
Receiver Operating Characteristic (ROC) Curves

- Indicate detection performance
- Signal efficiency $= \frac{true\ positives}{true\ positives + false\ negatives}$
- Background rejection $= \frac{true\ negatives}{true\ negatives + false\ positives}$
- Dependent on threshold

CNNs reach **higher signal efficiencies** at same background rejection level compared to OFMax

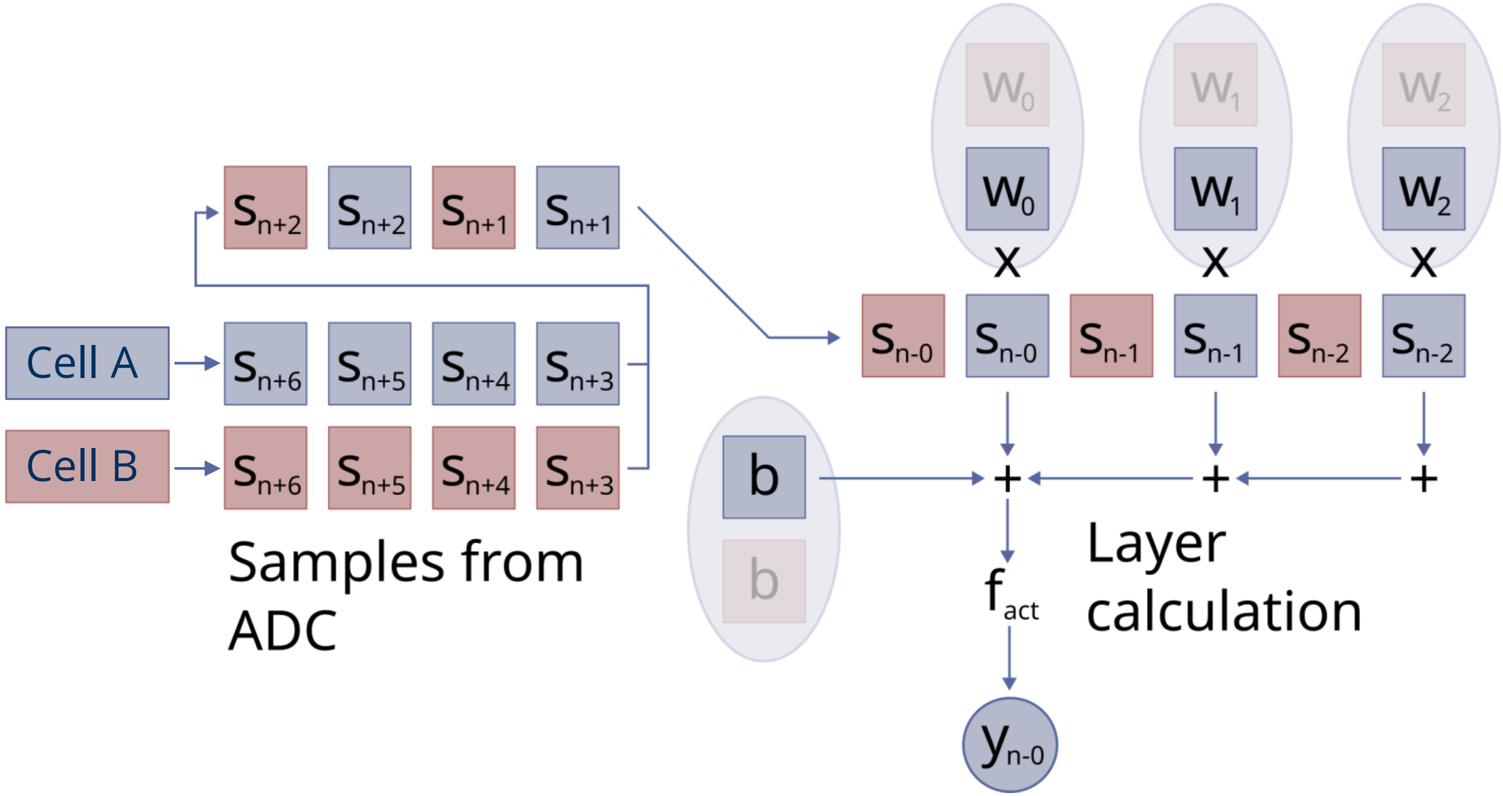
Performance Studies: Fakes

Spectrum of predicted transverse energy in BCs without energy deposition



CNN Firmware Implementation

- Transfer to hardware implemented in **VHDL**
- Time division **multiplexing**:
  - ➢ Design runs at 12x ADC frequency with cyclical processing of 12 detector cells



CNN Firmware Implementation

- Transfer to hardware implemented in **VHDL**
- Time division **multiplexing**:
  - Design runs at 12x ADC frequency with cyclical processing of 12 detector cells