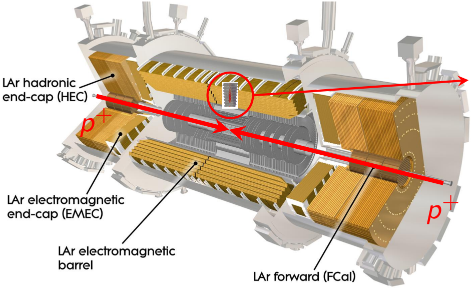
# Machine Learning for Real-Time Processing of ATLAS Liquid Argon Calorimeter Signals with FPGAs

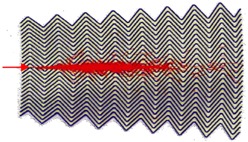
DIS2024 - Grenoble

Johann C. Voigt
on behalf of the ATLAS Liquid Argon Calorimeter Group

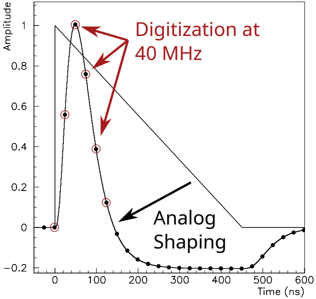
9 April 2024

- LHC provides $\approx 50$ proton-proton collisions per bunch crossing (BC) $\widehat{=}$ every 25 ns $\widehat{=}$ 40 MHz
- 140-200 simultaneous collisions at High Luminosity LHC (HL-LHC) from 2029 onwards
- Higher pileup and higher trigger rate require replacement of LAr calorimeter electronics



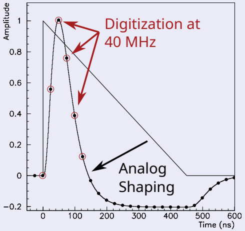



$\approx 182\,000$ detector cells



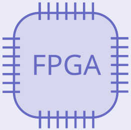

https://cds.cern.ch/record/1095928 [1], https://www.particles.uni-freiburg.de/dateien/vorlesungsdateien/particledetectors/kap8 [2], https://cds.cern.ch/record/331061 [3]

# Digital energy reconstruction



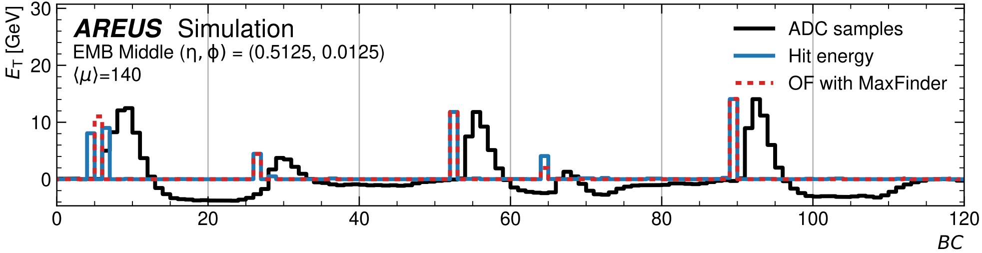

182k cells
@ 40 MHz

## Digital energy reconstruction (HL-LHC)



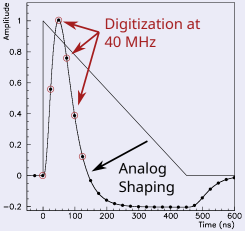

- 556 Intel Agilex-7 FPGAs for real-time processing
- Optimal Filter (OF)

$$E_t = \sum_{i=1}^{5} c_i \cdot x_{t+i}$$



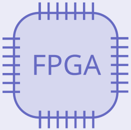

https://cds.cern.ch/record/331061 [3], https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LArCaloPublicResultsUpgrade [4]

# Digital energy reconstruction




182k cells
@ 40 MHz

## Digital energy reconstruction (HL-LHC)




- 556 Intel Agilex-7 FPGAs for real-time processing
- Optimal Filter (OF)

$$E_t = \sum_{i=1}^{5} c_i \cdot x_{t+i}$$

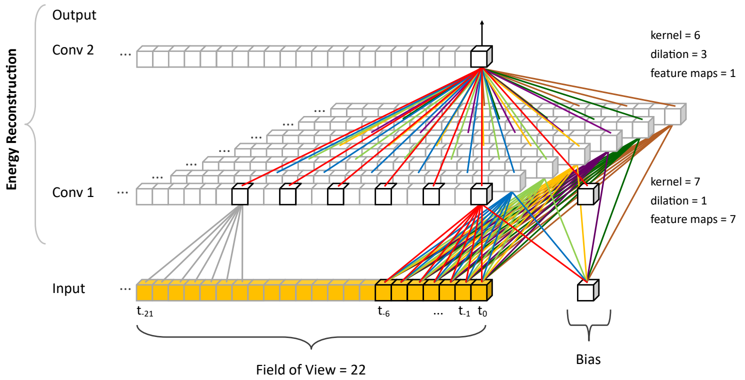
## Evaluating artificial neural networks (ANN) for cell level energy reconstruction

- Recurrent neural networks (RNN)
- Convolutional neural networks (CNN)

https://cds.cern.ch/record/331061 [3], https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LArCaloPublicResultsUpgrade [4]

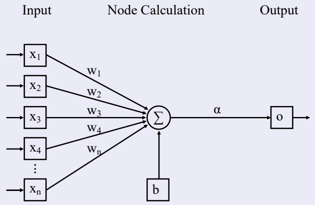
# Convolutional neural network architecture (CNN)

Energy prediction for every BC




Continuous input from one detector cell

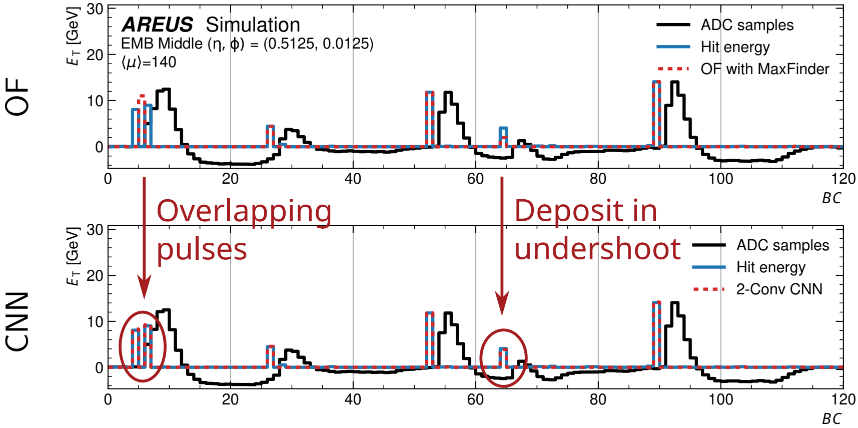
## ANN node




- Two convolutional layers
- ReLU activation
- 100 to 400 parameters
- 22 BC field of view

https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LArCaloPublicResultsUpgrade [4], [5]

# Example sequence

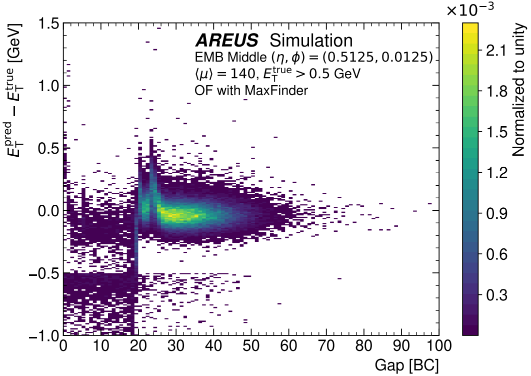



- Input: Signal enriched simulated detector sequences including pileup
- True energy available as training target
- OF/ANN output compared to true energy

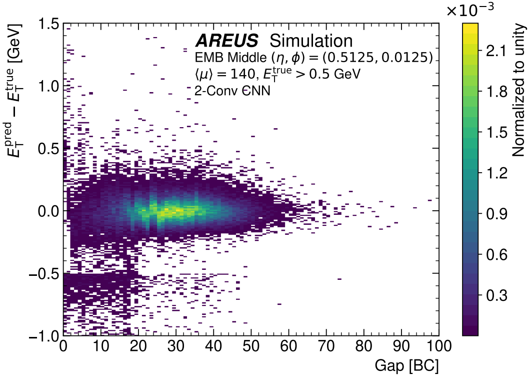
▶ ANNs show improved performance for overlapping pulses

https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LArCaloPublicResultsUpgrade [4]

# Energy reconstruction performance as a function of gap between 2 pulses




Optimal Filter

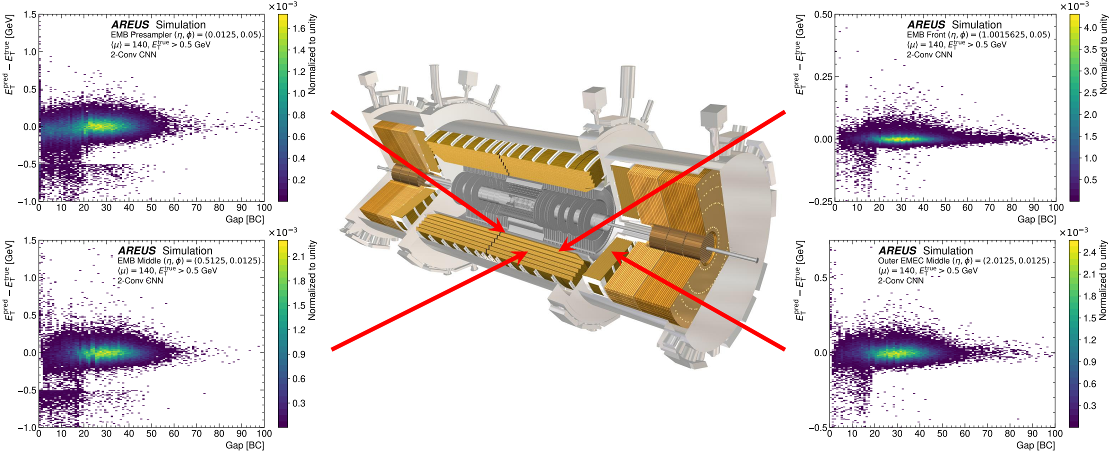



2-Conv CNN

- Improvements in reconstruction of overlapping pulses (gap $< 20$ BC)

https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LArCaloPublicResultsUpgrade [4]

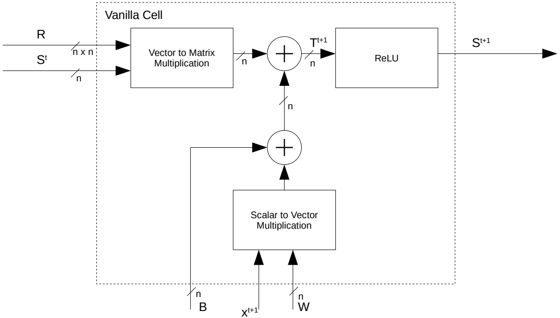
# Performance for different detector regions



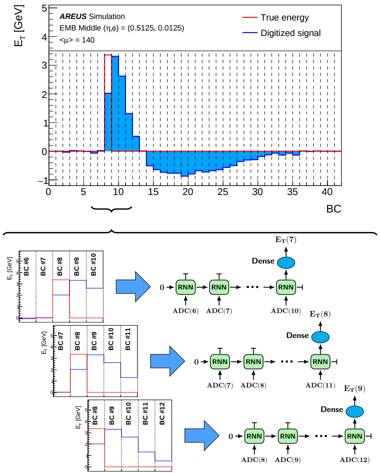

- Same architecture trained for different detector regions
- Identified clusters of similar cells based on pulse shapes for future trainings

https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LArCaloPublicResultsUpgrade [4], https://cds.cern.ch/record/1095928 [1]

# Recurrent neural network architecture (RNN)



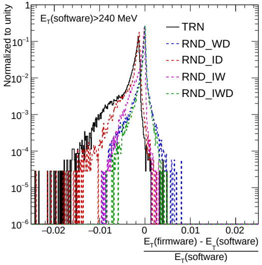

- Vanilla RNN with sliding window
  - Calculate output at every bunch crossing (BC)
  - Based on limited slice out of input sequence
- 5 cells with 8 internal dimensions → 304 multiplications
- Field of view of 5 BC can be extended using dense layer as input

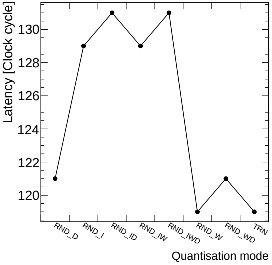


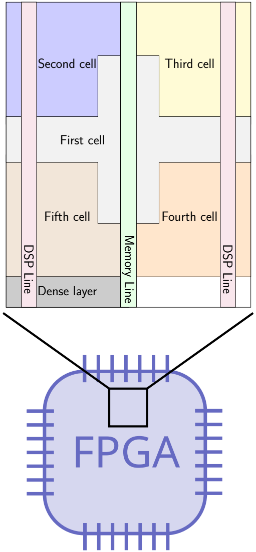

https://doi.org/10.48550/arXiv.2302.07555 [6]
https://doi.org/10.1007/s41781-021-00066-y [7]

# RNN firmware implementation

- Prototype in hls4ml
  - Added Intel/Quartus support for RNN
- Optimized High Level Synthesis (HLS) implementation:
  - Multiplexing support
  - Study influence of truncation/rounding
- VHDL implementation:
  - Reuse common results between RNN cells
  - Placement constraints
  - Incremental compilation



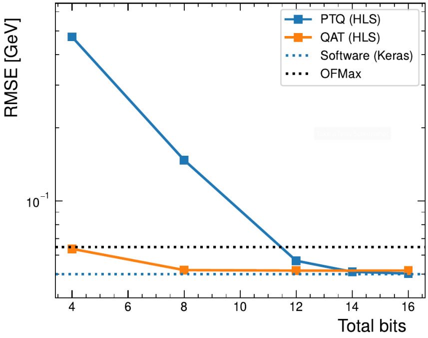




https://doi.org/10.48550/arXiv.2302.07555 [6]

# Quantization aware training

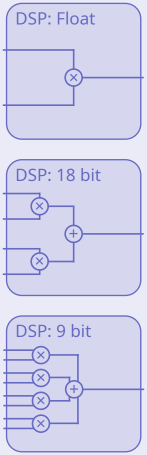
- FPGA more efficient for lower bit widths
- Balance between resources and accuracy
- QKeras allows quantization already during training
- Quantization aware training (QAT) yields better results than post training quantization (PTQ)

Example for RNN



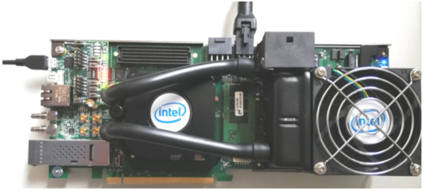

https://cds.cern.ch/record/2875588 [8]

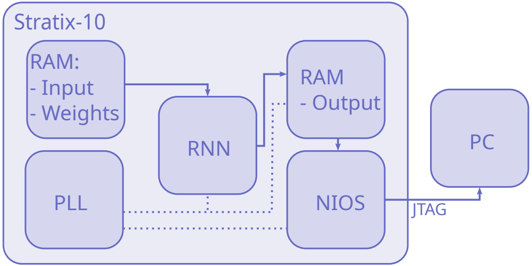
## DSP modes

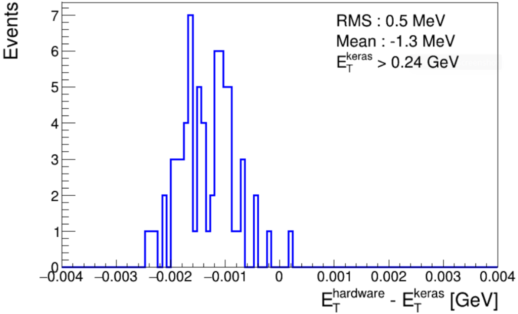


# Testing RNN firmware on hardware

- Test project for Stratix-10 development kit
- Data extracted over JTAG link using NIOS processor
- Bit exact match with firmware simulation
- Resolution compared to Keras as expected



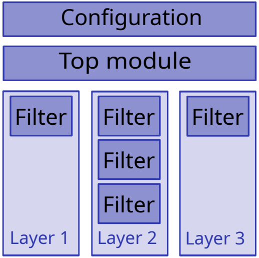


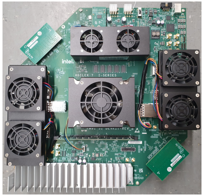



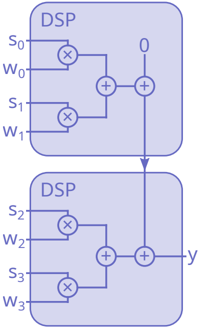

https://cds.cern.ch/record/2884186 [9]

# CNN firmware implementation

- CNN inference implemented in VHDL
- Model architecture configurable and automatically extracted from Keras output files
- Support multiplexing
  - Design runs at 12$\times$ ADC frequency
  - Cyclically processes 12 detector cells
- Development on Intel Stratix-10 and Agilex-7 FPGA
- Calculation in 18 bit fixed point numbers
- DSP can be chained for multiply-accumulate operations





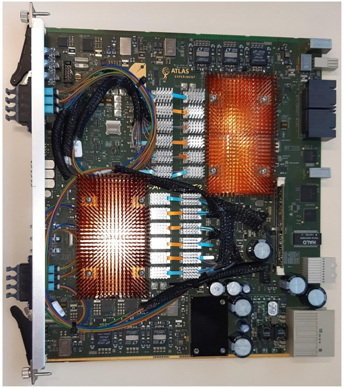
# FPGA resource estimation

- Latency requirement by ATLAS trigger of ≈ 150 ns met by all VHDL implementations
- All VHDL compilation targets can process required number of 384 detector cells
  - E.g. 12-fold multiplexing with 33 parallel instances
- Resource estimates based on Intel Quartus reports

| FPGA | Network | Multiplex. | Detector cells | $f_{\mathrm{max}}$ | ALMs | DSPs |
|---|---|---|---|---|---|---|
| Stratix-10 | RNN (HLS) | 10 | 370 | 393 MHz | 90 % | 100 % |
| | RNN (VHDL) | 14 | 392 | 561 MHz | 18 % | 66 % |
| | CNN (100 param.) | 12 | 396 | 415 MHz | 8 % | 28 % |
| Agilex | CNN (100 param.) | 12 | 396 | 539 MHz | 4 % | 13 % |
| | CNN (400 param.) | 12 | 396 | 510 MHz | 19 % | 50 % |

# Summary and outlook

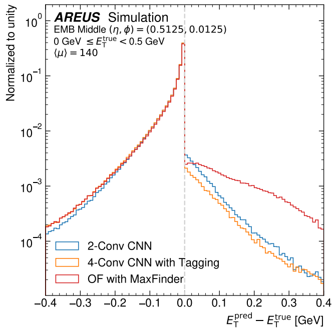
- RNNs and CNNs outperform Optimal Filter, especially for overlapping signals
  - First studies on effects of new cell energy reconstruction on photon, electron and jet measurements ongoing
- VHDL implementation of RNNs and CNNs with low latency available
- RNNs and CNNs fit target FPGA and run at required clock frequency
- Tests on FPGA hardware ongoing
- Integration with off-detector electronics firmware progressing



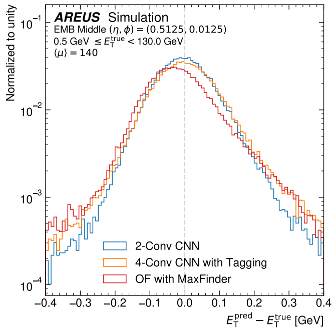
# Backup

# Distribution of deviation from true energy

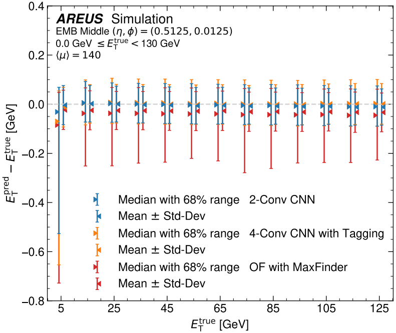
## $E_T^{\mathrm{true}} < 0.5$ GeV




## $E_T^{\mathrm{true}} \geq 0.5$ GeV

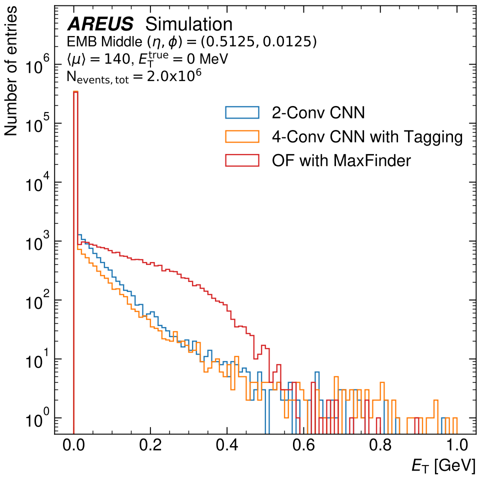



## Median/Mean over $E_T^{\mathrm{true}}$ range




https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LArCaloPublicResultsUpgrade [4]

# Prediction in BCs without energy deposit

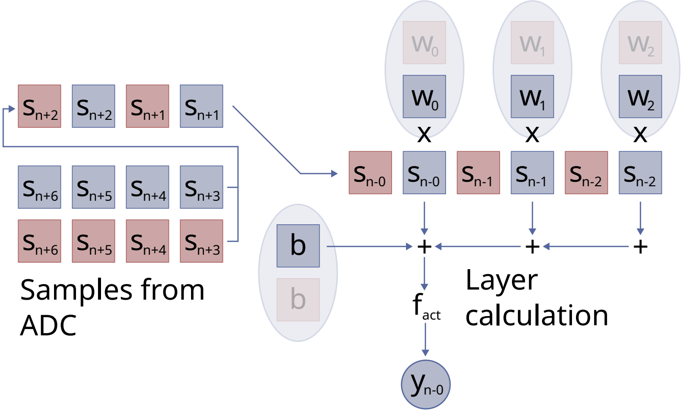



https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LArCaloPublicResultsUpgrade [4]

# CNN multiplexing concept

- One FPGA needs to fit 33 CNN instances
- Each instance uses 12× multiplexing
  → Design needs to run at 12× the ADC frequency: 480 MHz

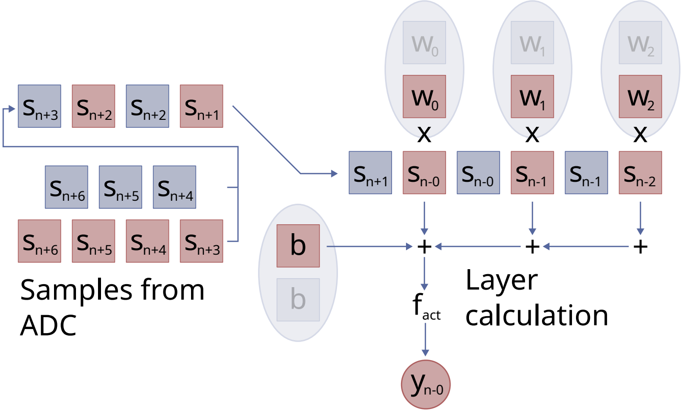
Example for two ADCs:



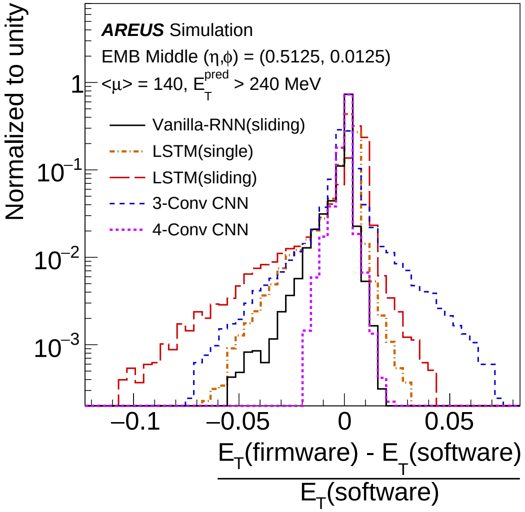
# CNN multiplexing concept

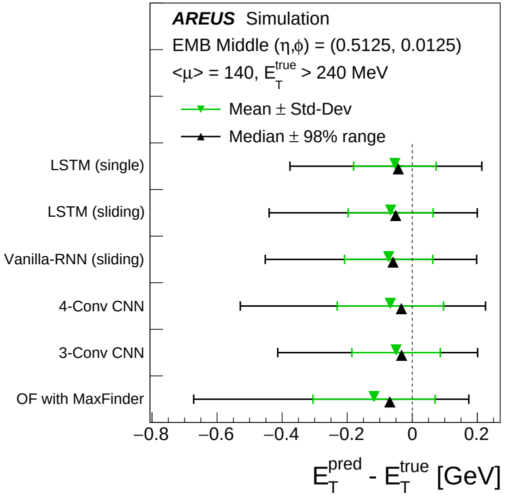
- One FPGA needs to fit 33 CNN instances
- Each instance uses 12× multiplexing
  → Design needs to run at 12× the ADC frequency: 480 MHz

Example for two ADCs:



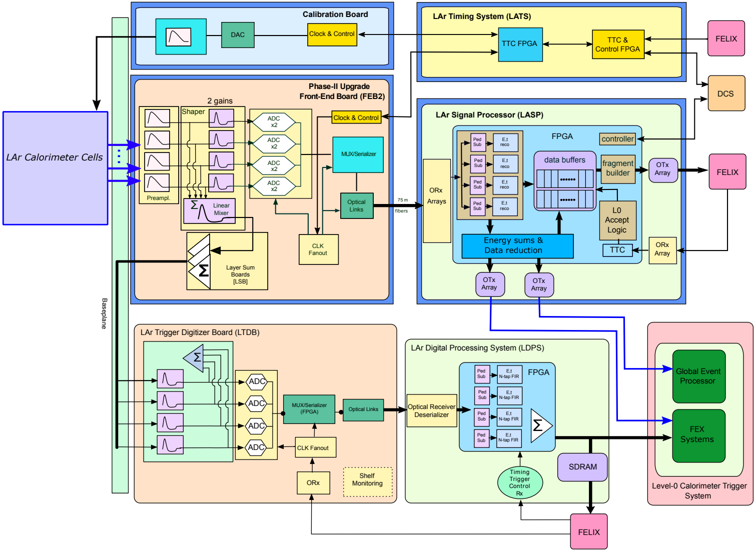
# Comparison of RNN/CNN energy resolution




https://doi.org/10.1007/s41781-021-00066-y [7]

# LAr Phase-II readout overview



# Sources I


[1] Joao Pequenao. *Computer generated image of the ATLAS Liquid Argon*. CERN. Mar. 27, 2008. URL: https://cds.cern.ch/record/1095928 (visited on 03/29/2021).

[2] Karl Jakobs. *Particle Detectors 2015. Chapter 8*. 2015. URL: https://www.particles.uni-freiburg.de/dateien/vorlesungsdateien/particledetectors/kap8 (visited on 03/21/2024).

[3] LHC Experiments Committee, LHCC. *ATLAS liquid-argon calorimeter: Technical Design Report*. Technical design report. ATLAS. Geneva: CERN, 1996. URL: https://cds.cern.ch/record/331061 (visited on 04/06/2021).

[4] ATLAS LAr Calorimeter Group. *Public Liquid Argon Calorimeter Plots on Upgrade*. URL: https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LArCaloPublicResultsUpgrade.

# Sources II


[5] Anne-Sophie Berthold. "Simulation Studies of Convolutional Neural Networks for the Real-Time Energy Reconstruction of ATLAS Liquid-Argon Calorimeter Signals at the High-Luminosity LHC". TU Dresden, IKTP, Dec. 21, 2023.

[6] Georges Aad et al. "Firmware implementation of a recurrent neural network for the computation of the energy deposited in the liquid argon calorimeter of the ATLAS experiment". In: (Feb. 15, 2023). DOI: https://doi.org/10.48550/arXiv.2302.07555. arXiv: 2302.07555v1 [physics.ins-det]. URL: https://doi.org/10.48550/arXiv.2302.07555.

[7] 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.

# Sources III


[8] Lauri Antti Olavi Laatu. "Development of artificial intelligence algorithms adapted to big data processing in embedded (FPGAs) trigger and data acquisition systems at the LHC". Presented 03 Oct 2023. Aix-Marseille U., 2023. URL: https://cds.cern.ch/record/2875588.

[9] Nemer Chiedde. "Implementation of embedded artificial intelligence algorithms in the readout system of the ATLAS liquid argon calorimeter". Presented 21 Nov 2023. PhD thesis. Aix-Marseille Université, 2023. URL: https://cds.cern.ch/record/2884186 (visited on 03/21/2024).