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

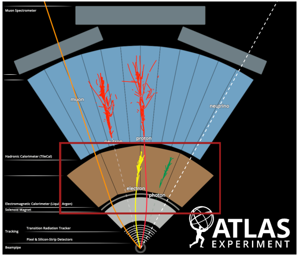
CHEP 2023 - Norfolk

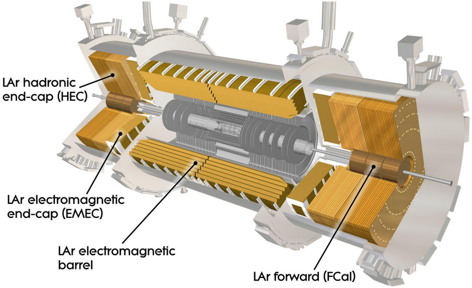
Johann C. Voigt
on behalf of the ATLAS Liquid Argon Calorimeter Group

11 May 2023

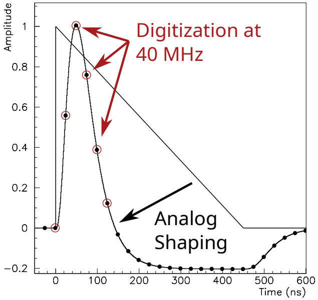
# ATLAS LAr-Calorimeter

- 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



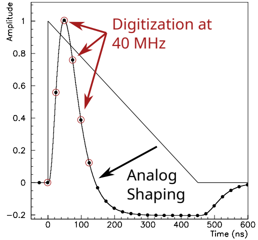

https://cds.cern.ch/record/2770815 [1], https://cds.cern.ch/record/1095928 [2], http://cds.cern.ch/record/1701107 [3]

# Digital energy reconstruction

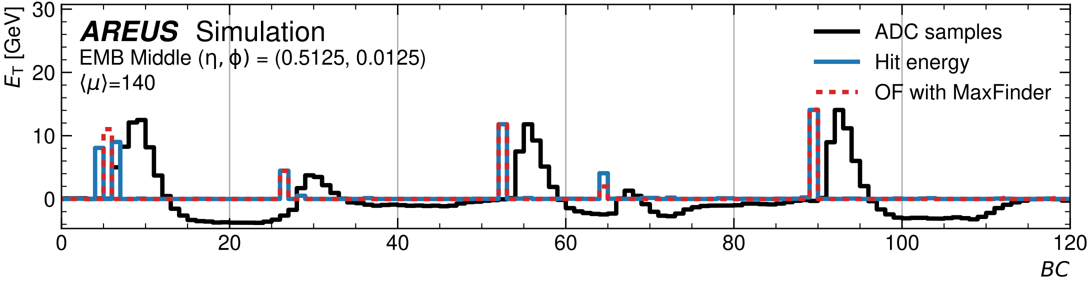
- Digital energy reconstruction with Optimal Filter (OF)

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

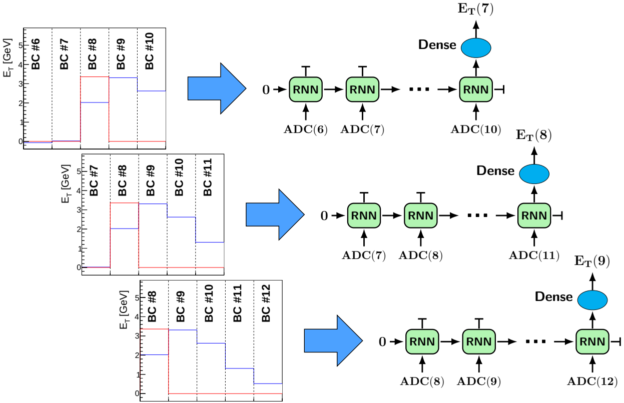
- Overlapping signals require better algorithm
- 556 high-performance FPGAs will be installed for real-time digital signal processing

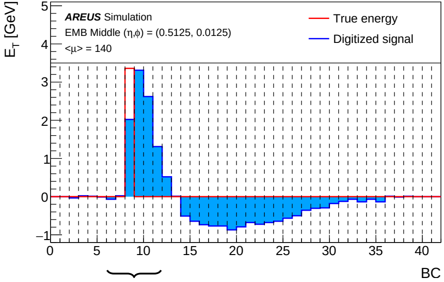



http://cds.cern.ch/record/1701107 [3]



# Recurrent neural network architecture (RNN)



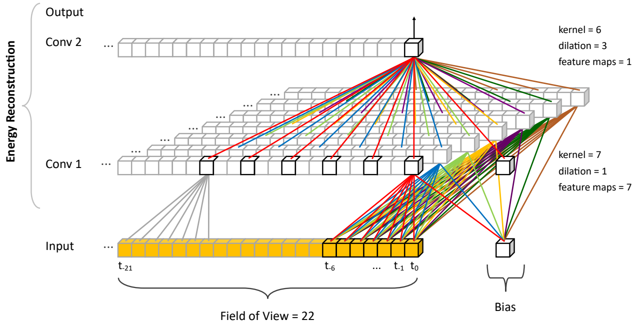




Field of view = 5 BC

- 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

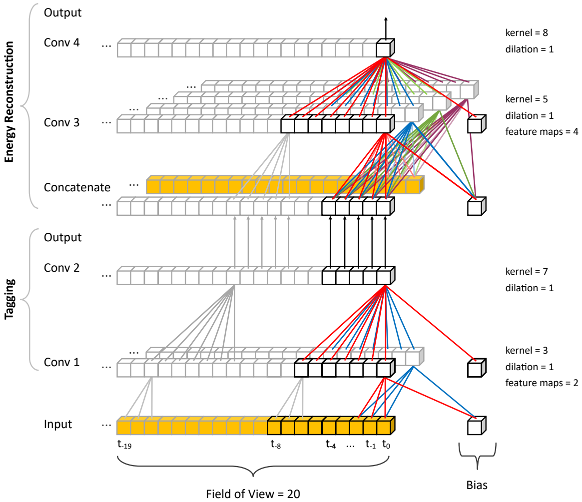
https://doi.org/10.1007/s41781-021-00066-y [4]

# Convolutional neural network architecture (CNN)



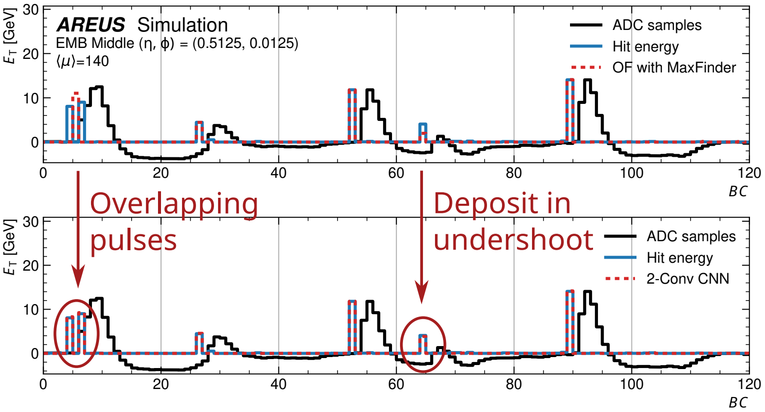

- 2 convolutional layers using ReLU activation for energy reconstruction
- ≈ 100 parameters
- ≈ 20 BC field of view

# Convolutional neural network architecture (CNN)



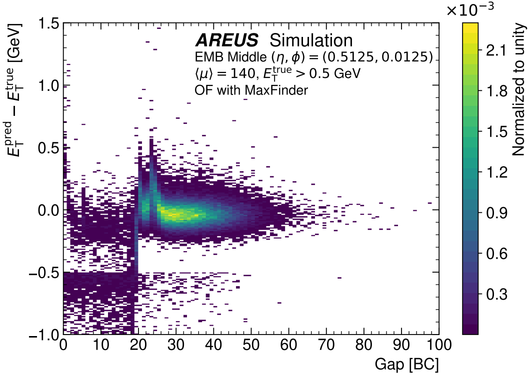

- 2 convolutional layers using ReLU activation for energy reconstruction
- $\approx$ 100 parameters
- $\approx$ 20 BC field of view
- 2 convolutional layers to tag undershoot of previous pulses using sigmoid activation ($\rightarrow$ less hardware friendly)

# Example sequence

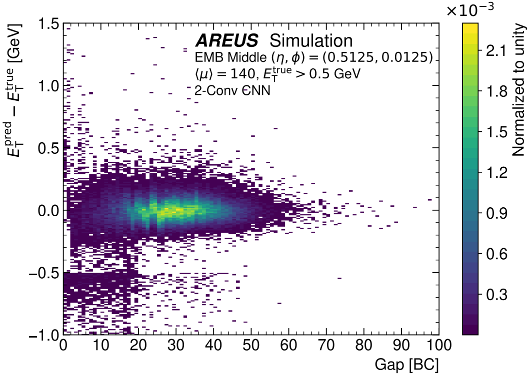



- Trained on signal-enriched simulated detector sequences including pileup
- True energy available as training target
- Network and Optimal Filter performance can be evaluated by comparison with true energy

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




Optimal Filter

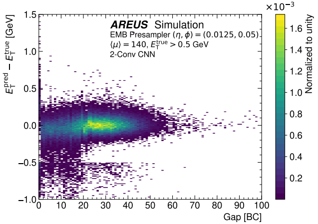



2-Conv CNN

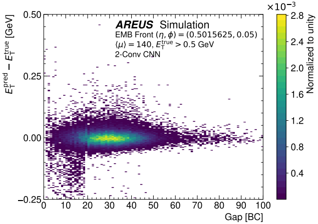
→ Improvements in reconstruction of overlapping pulses (gap $<$ 20 BC)

# CNN performance for different detector regions

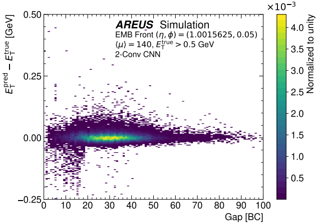
EMB Presampler, $\eta = 0$



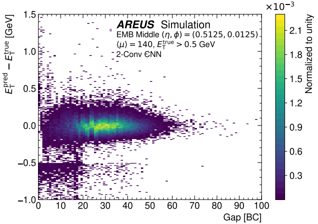

EMB Front, $\eta = 0.5$



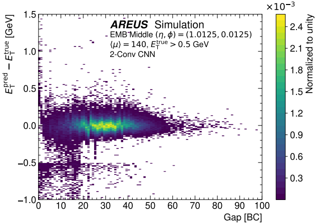

EMB Front, $\eta = 1$



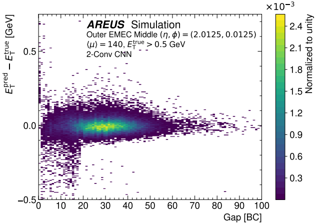

EMB Middle, $\eta = 0.5$




EMB Middle, $\eta = 1$



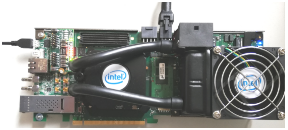

Outer EMEC Middle, $\eta = 2$

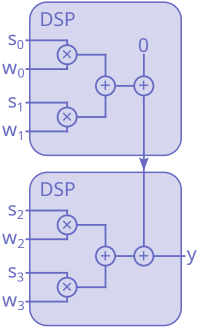



- Same architecture trained for different detector regions

# CNN firmware implementation

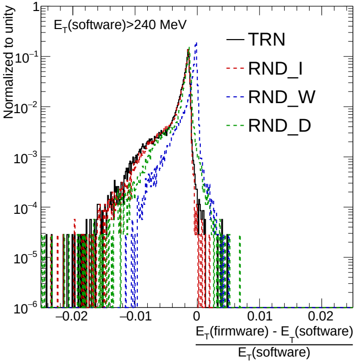
- CNN inference implemented in VHDL
- Model architecture configurable and automatically extracted from Keras output files
- Support multiplexing:
  $\rightarrow$ Design runs at $12\times$ ADC frequency and cyclically processes 12 detector cells
- Development on Intel Stratix-10 FPGA, final design will use Intel Agilex
- Calculation in 18 bit fixed point numbers
- Intel DSPs can multiply two pairs of 18 bit numbers at once
- DSP can be chained for vector multiplications

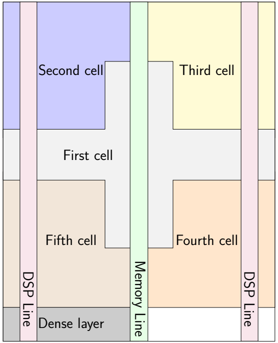




# RNN firmware implementation

1. Prototype in HLS4ML
   - Added Intel/Quartus support for RNN
2. Optimized High Level Synthesis (HLS) implementation:
   - Multiplexing support
   - Study influence of truncation (TRN)/rounding (RND): intermediary results (I), weights (W), in-/output (D)
3. VHDL implementation:
   - Reuse common results between RNN cells
   - Placement constraints
   - Incremental compilation




https://doi.org/10.48550/arXiv.2302.07555 [5]

# 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 | Multiplexing | Detector cells | $f_{\max}$ | ALMs | DSPs |
|---|---|---|---|---|---|---|
| Stratix-10 | RNN (HLS) | 10 | 370 | 393 MHz | 90 % | 100 % |
| | RNN (VHDL) | 14 | 392 | 561 MHz | 18 % | 66 % |
| | 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 % |

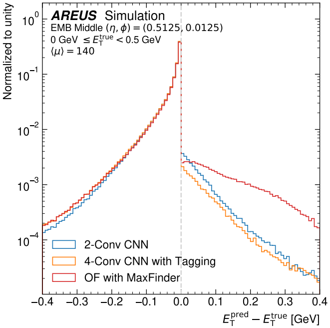
# Summary

- RNNs and CNNs outperform Optimal Filter, especially for overlapping signals
  $\rightarrow$ Study effect of new cell energy reconstruction on photon, electron and jet measurements
- 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

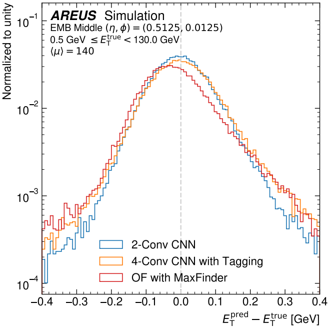
# Backup

# Distribution of deviation from true energy

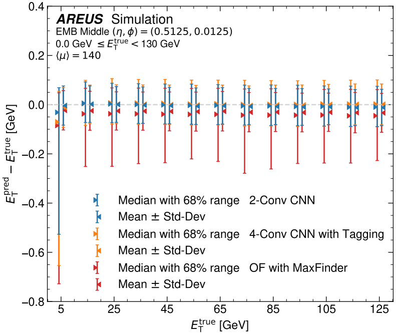
## $E_{\mathrm{T}}^{\mathrm{true}} < 0.5$ GeV



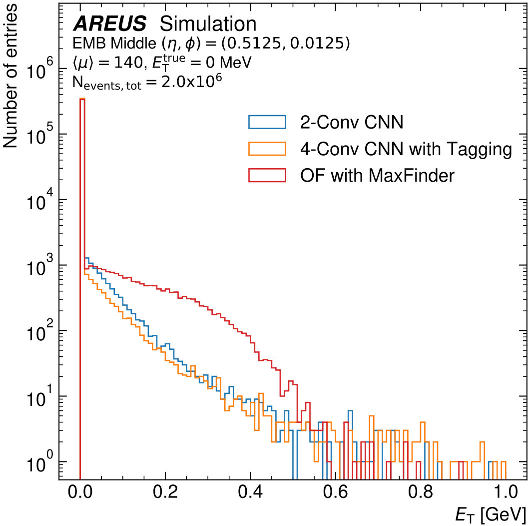

## $E_{\mathrm{T}}^{\mathrm{true}} \geq 0.5$ GeV



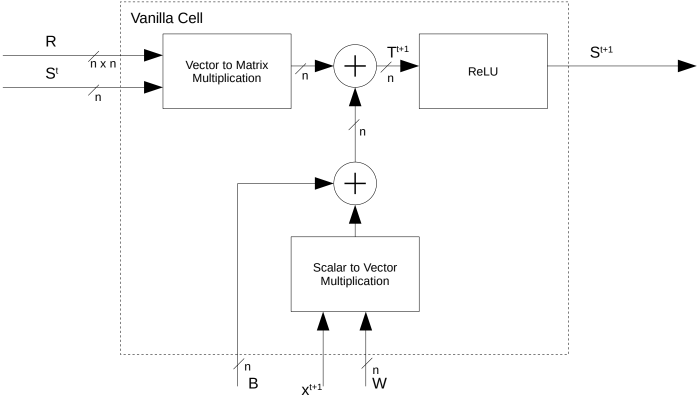

## Median/Mean over $E_{\mathrm{T}}^{\mathrm{true}}$ range

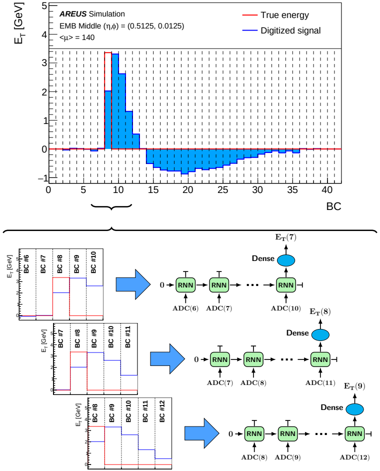


# Prediction in BCs without energy deposit



# RNN cell structure



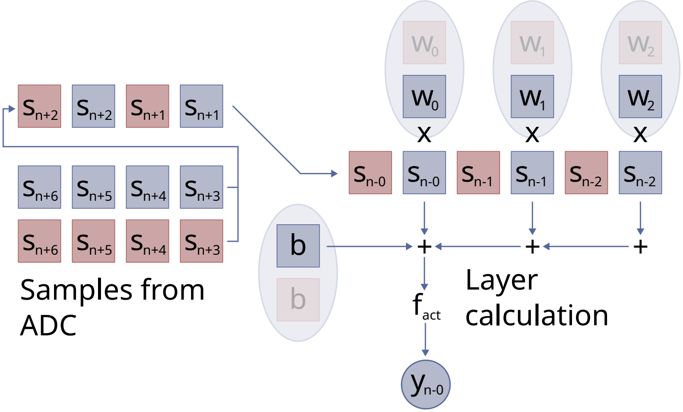




https://doi.org/10.48550/arXiv.2302.07555 [5], https://doi.org/10.1007/s41781-021-00066-y [4]

# 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

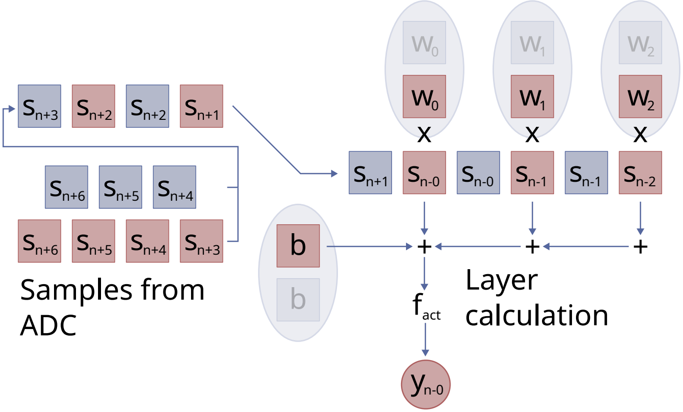
Example for two ADCs:



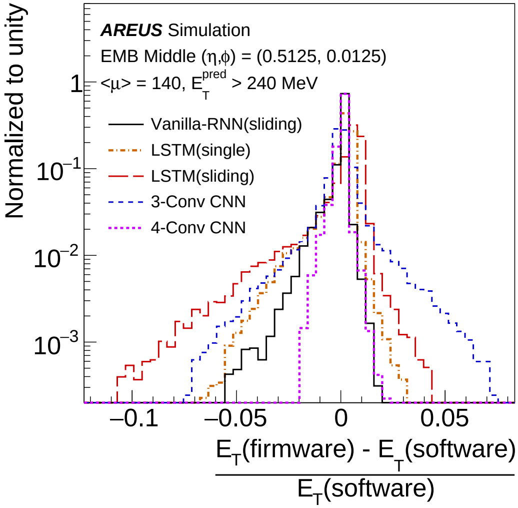
# CNN multiplexing concept

- 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:



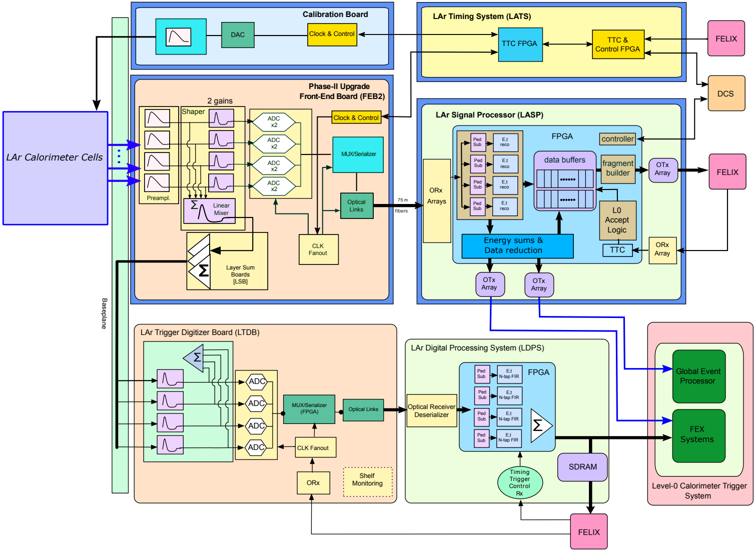
# Relative deviation between firmware and software




https://doi.org/10.1007/s41781-021-00066-y [4]

- Good agreement between firmware and software (for samples with pred. energy above 240 MeV)

# LAr Phase-II readout overview



# Sources I


[1] Sascha Mehlhase. *ATLAS detector slice (and particle visualisations)*. 2021. URL: https://cds.cern.ch/record/2770815.

[2] 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).

[3] ATLAS Collaboration. "Monitoring and data quality assessment of the ATLAS liquid argon calorimeter". In: *JINST* 9.arXiv:1405.3768. CERN-PH-EP-2014-045 (May 2014). Plot available separately: http://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PAPERS/LARG-2013-01/, P07024. 39 p. URL: http://cds.cern.ch/record/1701107 (visited on 05/28/2017).

# Sources II

[4] 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.

[5] 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*. 2023. DOI: 10.48550/ARXIV.2302.07555. URL: https://doi.org/10.48550/arXiv.2302.07555.