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

PUNCH4NFDI TA5 - XFEL Joint Workshop on Machine Learning and Data Processing on FPGAs

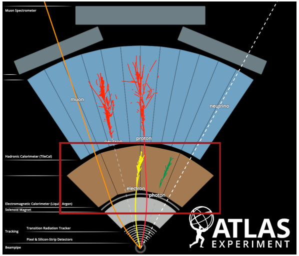
Johann C. Voigt

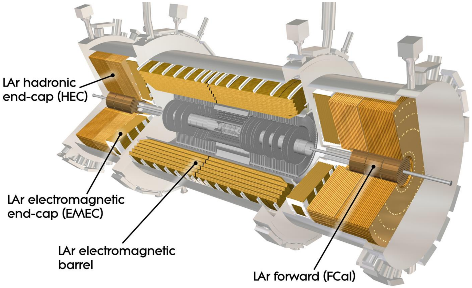
16 June 2023

Gefördert vom Bundesministerium für Bildung und Forschung

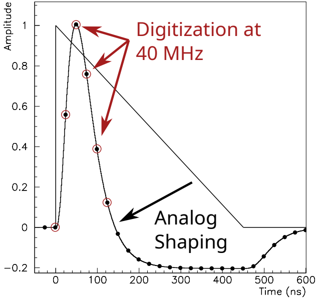
ATLAS LAr-Calorimeter

- LHC provides ≈ 50 proton-proton collisions per bunch crossing (BC) $\widehat{=}$ every $25\,\mathrm{ns}$ $\widehat{=}$ $40\,\mathrm{MHz}$
- 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



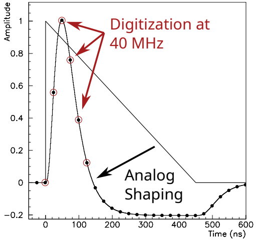
https://cds.cern.ch/record/2770815 [1], https://cds.cern.ch/record/1095928 [2], http://cds.cern.ch/record/1701107 [3]

Digital energy reconstruction

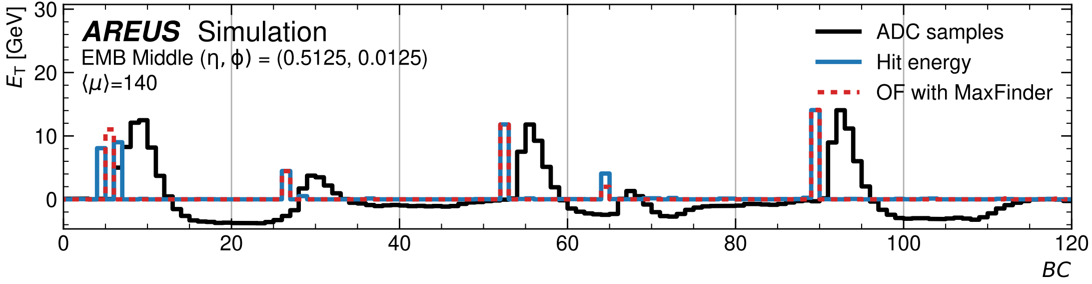
- Digital energy reconstruction with Optimal Filter (OF)

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

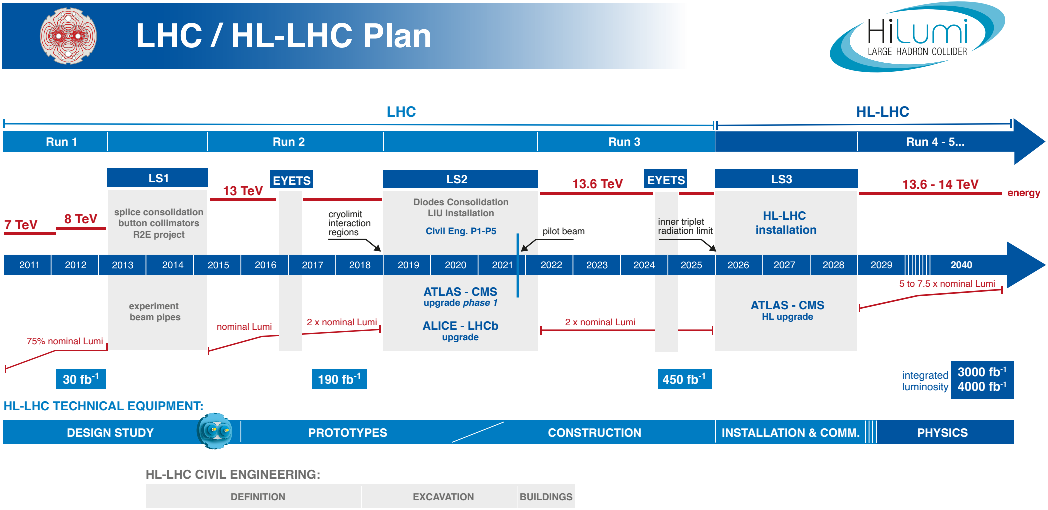
- Overlapping signals require better algorithm
- 556 high-performance FPGAs will be installed for real-time digital signal processing



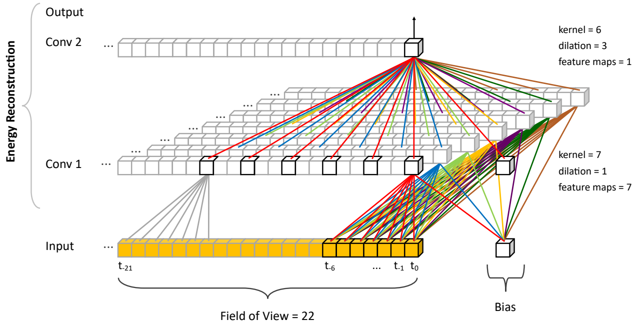
http://cds.cern.ch/record/1701107 [3]



LHC Timeline

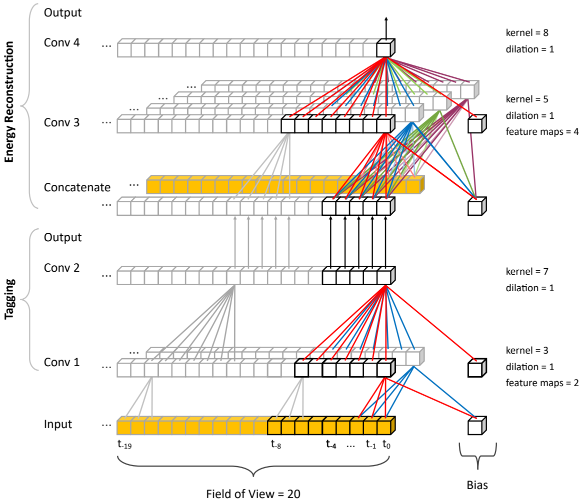


Convolutional neural network architecture (CNN)



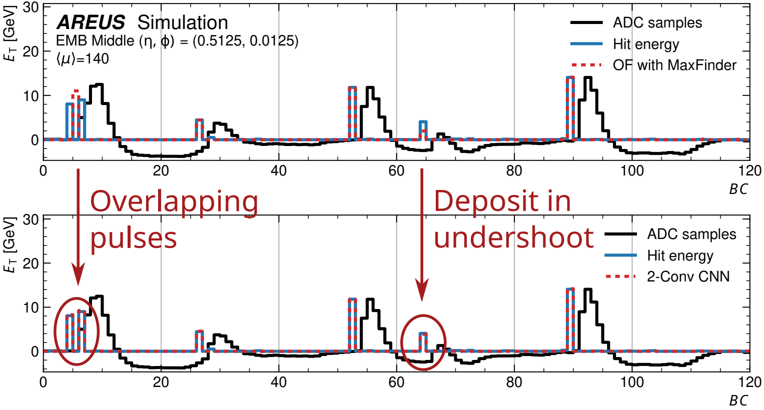
- 2 convolutional layers using ReLU activation for energy reconstruction
- $\approx$ 100 parameters
- $\approx$ 20 BC field of view

Convolutional neural network architecture (CNN)



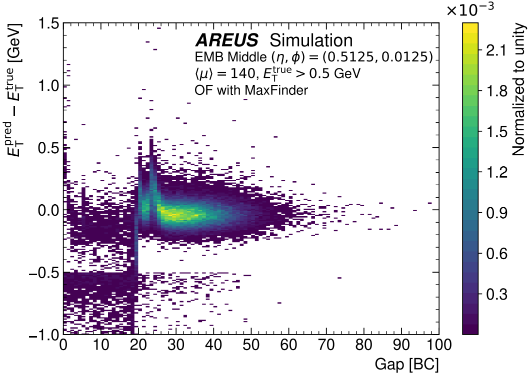
- 2 convolutional layers using ReLU activation for energy reconstruction
- $\approx$ 100 parameters
- $\approx$ 20 BC field of view
- 2 convolutional layers to tag undershoot of previous pulses using sigmoid activation ($\rightarrow$ less hardware friendly)

Example sequence

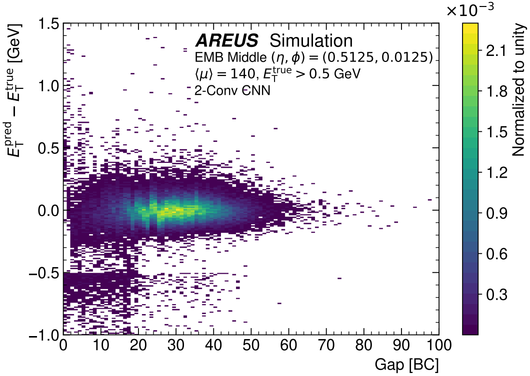


- Trained on signal-enriched simulated detector sequences including pileup
- True energy available as training target
- Network and Optimal Filter performance can be evaluated by comparison with true energy

Energy reconstruction performance as a function of gap between 2 pulses



Optimal Filter

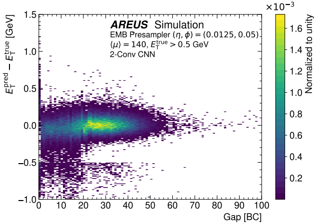


2-Conv CNN

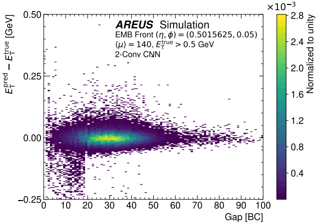
$\rightarrow$ Improvements in reconstruction of overlapping pulses (gap < 20 BC)

CNN performance for different detector regions

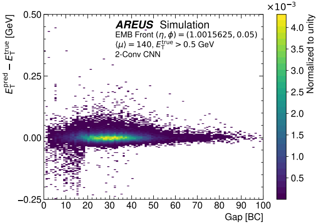
EMB Presampler, $\eta = 0$



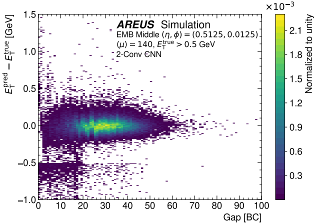
EMB Front, $\eta = 0.5$



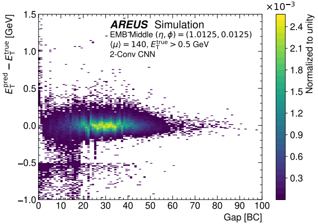
EMB Front, $\eta = 1$



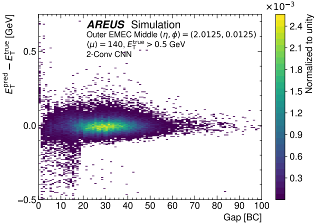
EMB Middle, $\eta = 0.5$



EMB Middle, $\eta = 1$



Outer EMEC Middle, $\eta = 2$



- Same architecture trained for different detector regions

Decisions for firmware implementation

Low level	High level
• Allows more customized optimization	• More compact code

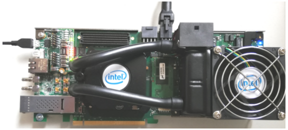
Custom implementation	Framework
• More specialized/customizable	• More professional code

Separated architecture/weights	Weight dependent structure
• Weights can be loaded at runtime, no recompilation necessary • Easier structure of calculation (vector multiplication)	• Allows pruning

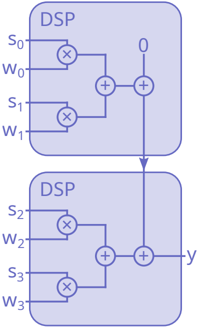
Further considerations

- FPGA vendor/model dependence
- Need for custom features?

CNN firmware implementation



- CNN inference implemented in VHDL
- Model architecture configurable and automatically extracted from Keras output files
- Support multiplexing:
- 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

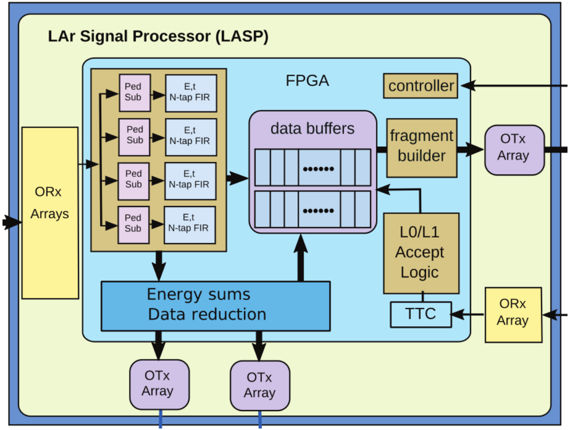


Network size

Initial constraints:

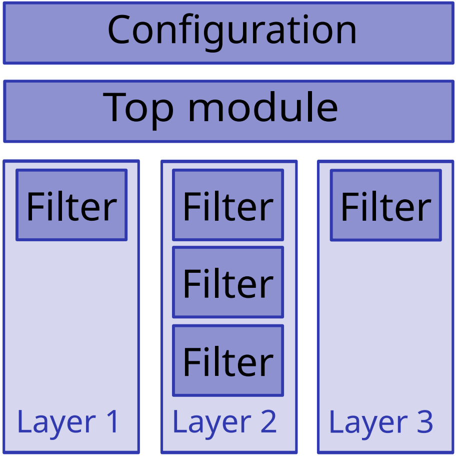
- Initial plan: Up to 512 detector cells per FPGA
- FPGA-Model: Stratix 10 with 5720 DSPs $\rightarrow$ up to 11440 multipliers in 18 bit chained mode
- Input data arrives at 40 MHz, firmware planned to run at at least 320 MHz $\rightarrow$ can process 10 detector cells cyclical
- Worst case scenario requires 52 parallel neural network firmware instances
- Can estimate possible network size based in available multipliers
  $\rightarrow \approx 100$ parameters per network to leave some margin and resources for other systems
- Constraints relaxed significantly now: 384 cells, 480 MHz ($12\times$ multiplexing) and larger FPGA (Intel Agilex with 12792 DSPs)
  $\rightarrow$ Can consider larger networks in the future

LASP framework



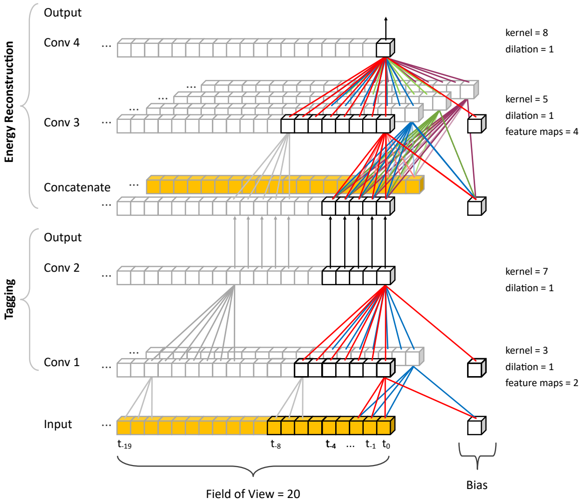
- CNNs to be embedded in larger firmware project (LASP)
- Existing framework provides easy setup for simulation with UVVM checker and compilation
- Gitlab project with established workflow
- CI triggers automated checks of simulations and lightweight compilation for every merge request

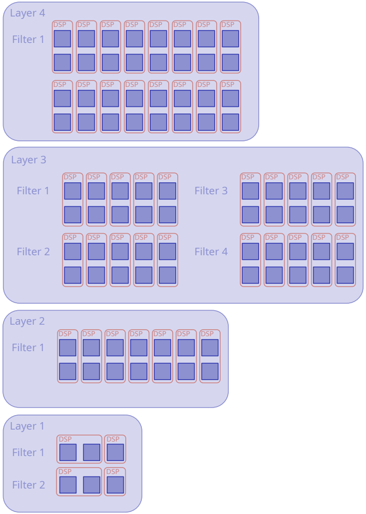
Firmware hierarchy



- Configuration generated from Keras files
- Top module generates *Filter* modules depending on architecture
  $\rightarrow$ Network architecture flexible in code, but fixed at compile time
- Weights stored in RAM per *Filter* instance

DSP assignment

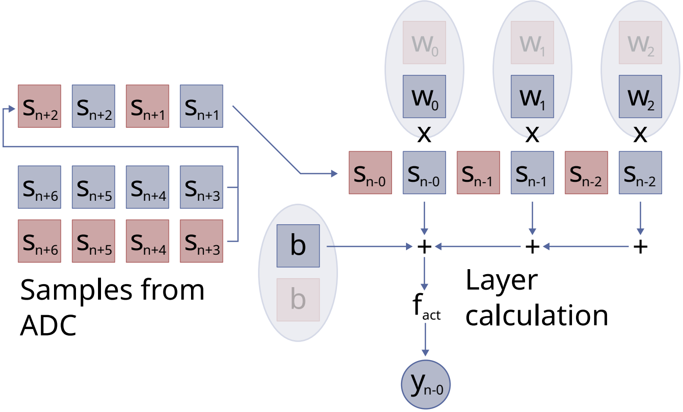




CNN multiplexing concept

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

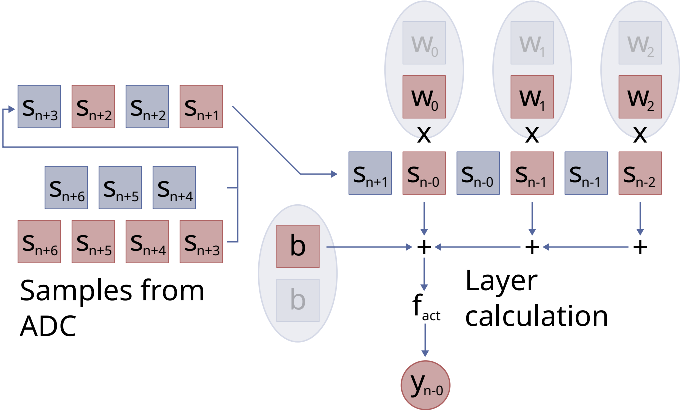
Example for two ADCs:



CNN multiplexing concept

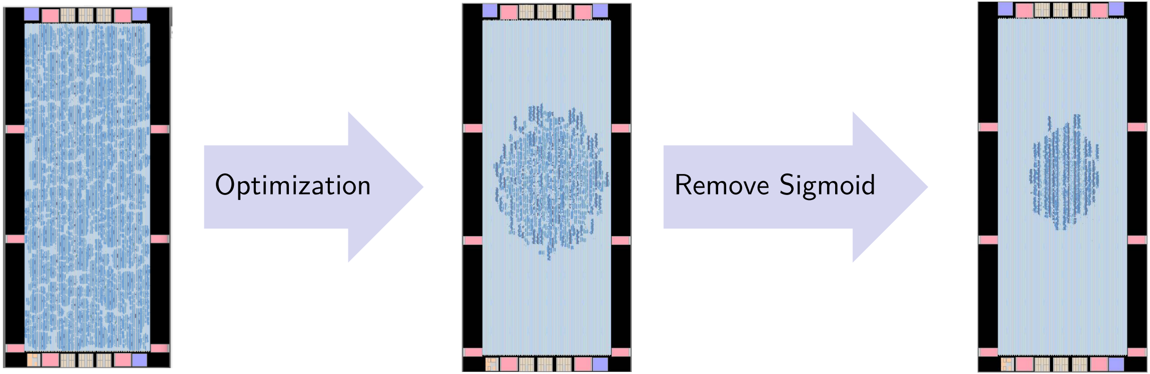
- One FPGA needs to fit 33 CNN instances
- Each instance uses 12× multiplexing
  → Design needs to run at 12× the ADC frequency: 480 MHz

Example for two ADCs:



Moving complexity to software

- Initially stored weigts in logical order like Keras
- Timings of DSP chain require different order
  $\rightarrow$ Significant ALM overhead for reordering of weights on FPGA
- Optimized version reorders weights in software preprocessing $\rightarrow$ 68 % reduction in ALM usage



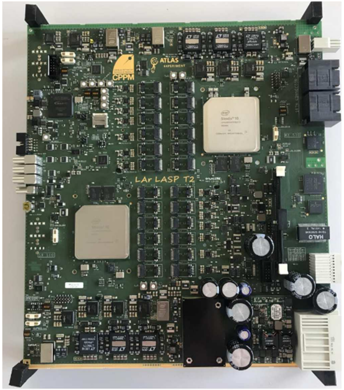
FPGA resource estimation

- 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	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

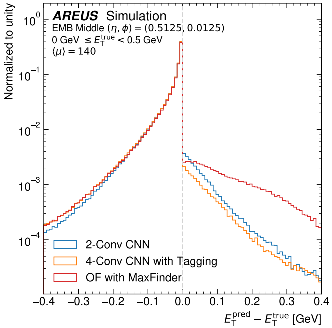
- CNNs outperform Optimal Filter, especially for overlapping signals
  $\rightarrow$ Study effect of new cell energy reconstruction on photon, electron and jet measurements
- VHDL implementation of CNNs with low latency, fitting target FPGA and can run at required clock frequency
- Tests on FPGA hardware ongoing
- Planning to release CNN code as open source in future



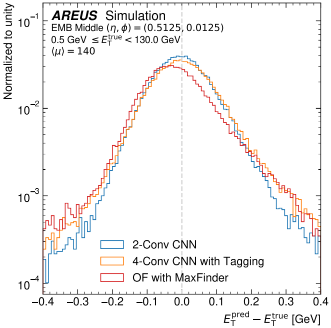
Backup

Distribution of deviation from true energy

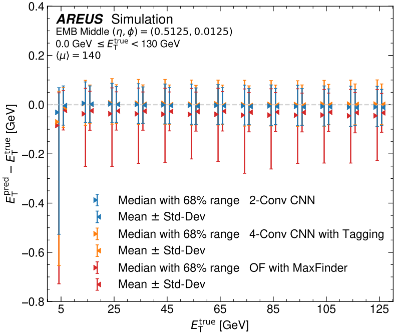
$E_T^{\mathrm{true}} < 0.5$ GeV



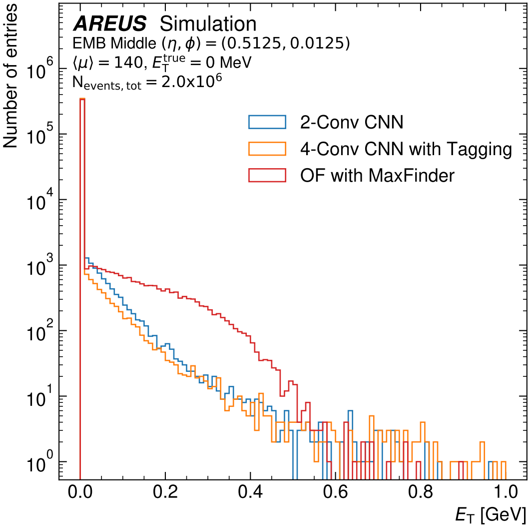
$E_T^{\mathrm{true}} \geq 0.5$ GeV



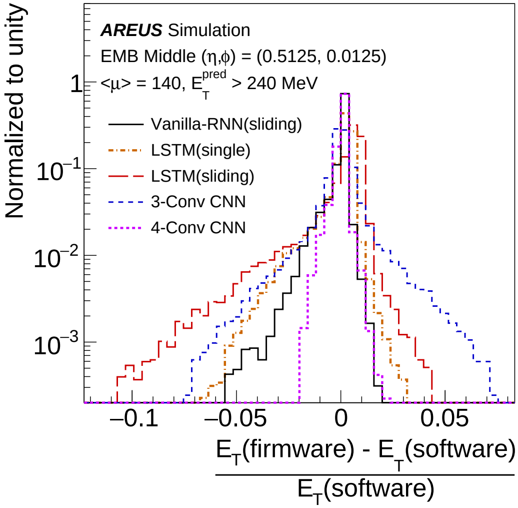
Median/Mean over E_T^{true} range



Prediction in BCs without energy deposit



Relative deviation between firmware and software



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

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

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.