# Machine Learning on FPGAs for Real-Time Processing for ATLAS Liquid Argon Calorimeter

**Lauri Laatu**, Georges Aad, Nemer Chiedde, Robert Faure, Emmanuel Monnier, Nairit Sur

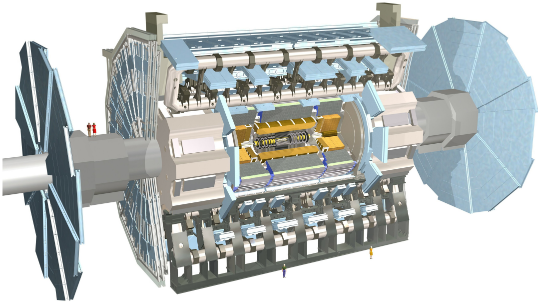
17.10.2022

# Content

# The ATLAS Experiment at the Large Hadron Collider (LHC)

## General purpose detector

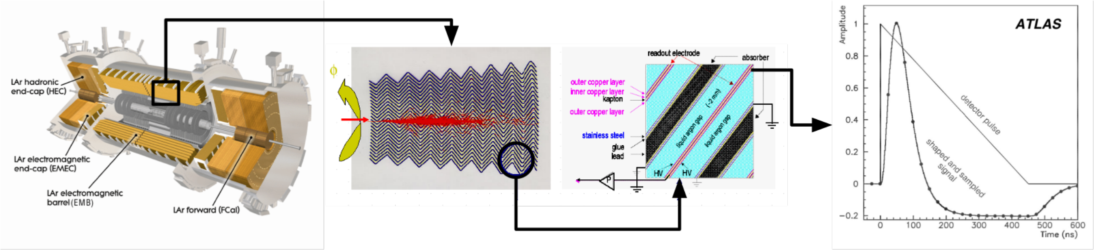
- The ATLAS Experiment is one of the general purpose detectors at the LHC
  - Consists of a tracker, electromagnetic and hadronic calorimeters and muon detectors
- Proton-proton collisions every 25ns (40MHz) referred to as bunch crossings (BCs)
  - Real-time event selection from 40MHz to store events at 10kHz



# Liquid Argon Calorimeter

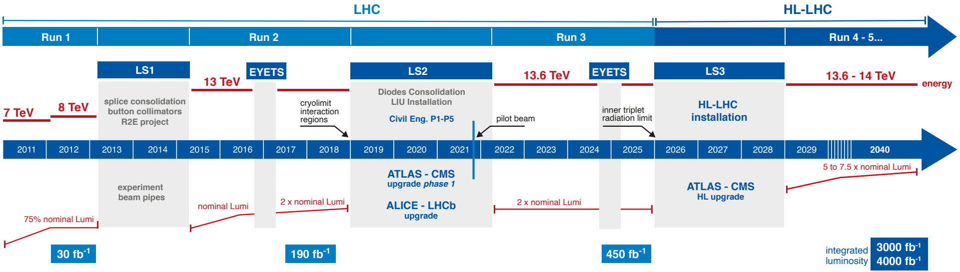
## Energy reconstruction in the LAr calorimeter

- Liquid Argon Calorimeter (LAr) mainly measures the energy deposited by electromagnetically interacting particles
  - Consisting of $\approx$ 182 000 calorimeter cells
- Passing particles ionize the material
  - Bipolar pulse shape with total length of up to 750 ns (30 BCs)
  - Pulse is sampled and digitized at 40MHz
- Energy reconstruction is done real-time and used in triggering decision
  - Using the digitized samples from the pulse



# The Phase-II Upgrade of the LHC

## Upgrade of the ATLAS experiment



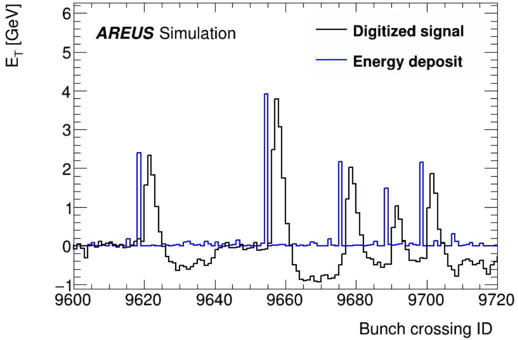

- The High Luminosity LHC (HL-LHC) is an important milestone for particle physics
  - Increase the luminosity to study rare processes
  - Increase the collision rate to up to 200 simultaneous p-p collisions (pileup) per bunch crossing (BC)
- The detectors will be upgraded to cope with the high collision rate at the HL-LHC
  - In particular the ATLAS calorimeter readout electronics will be completely replaced

# Energy Reconstruction

## Energy reconstruction in the LAr calorimeter

- Current energy reconstruction uses optimal filtering algorithm with maximum finder (OFMax)
  - Using five samples around pulse shape peak is used in Phase-II studies
  - Assuming perfect pulse shape
- High pileup leads to higher rate of overlapping pulse shapes
  - Distorted bipolar shape → significantly decreased performance of OFMax

- Energy is computed real-time at 40MHz
  - Using specialized boards based on FPGAs
  - For Phase-II one FPGA processes 384 channels
  - Latency requirement of 125 ns
- Phase-II electronics with high-end FPGAs
  - Increased computing capacity
  - Improved online energy reconstruction using machine learning based methods
- Constraints from running on FPGAs
  - Latency, frequency and occupancy
  - Small networks needed

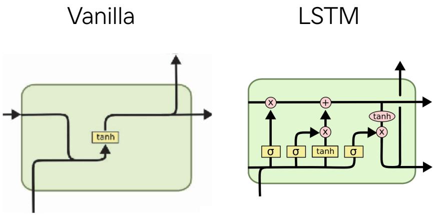


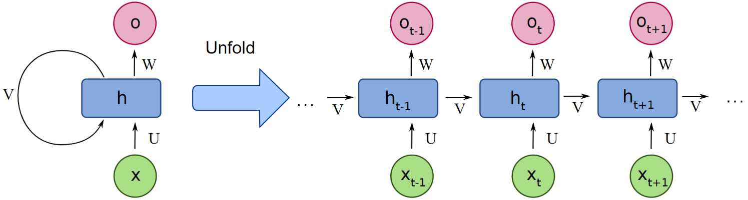
# Table of Contents

# RNN Architecture

## Timeseries processing

- Recurrent Neural Networks (RNNs) are designed to process time series data
- RNNs consists of neural network layers that process by combining new time input with past processed state
- Vanilla RNN is the smallest RNN structure
- Long Short-Term Memory (LSTM) network for efficiently handling past information

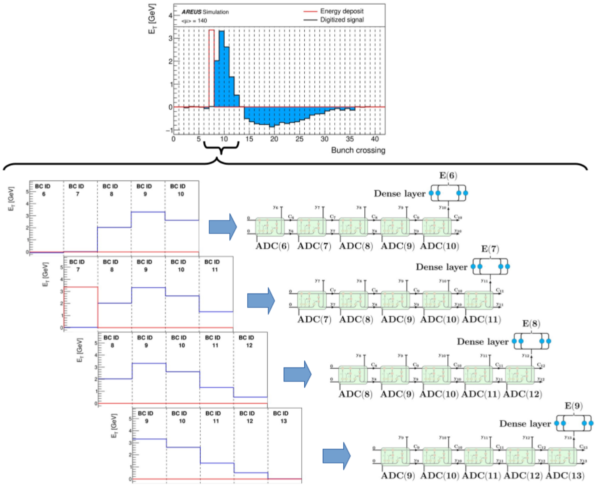


# RNNs for Energy Reconstruction

Using a many-to-one and many-to-many networks for energy reconstruction

- Use digitized samples as inputs for the recurrent network
- Sliding window
  - Full sequence split into overlapping subsequences with a sliding window
  - One energy prediction per subsequence
  - Network receives limited amount of data from the past
  - Possible for Vanilla RNN and LSTM
- Single cell
  - Use the LSTM cell to process all digitized samples in one continuous chain instead of a sliding window
  - Full history of events available
  - Possible only for LSTM

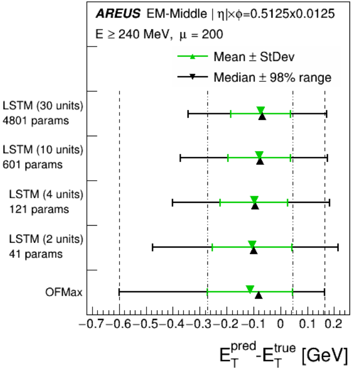


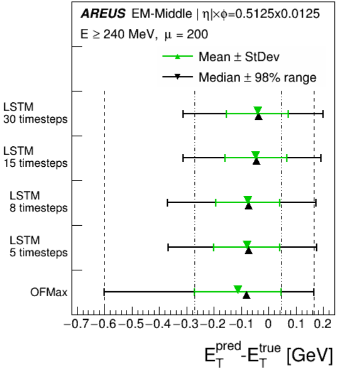
# Table of Contents

# Network Optimization

Find the smallest well performing network, example for sliding window LSTM

- Use standard deviation and 98% range to compare energy resolution
  - Non-gaussian distribution of the energy resolution
- Optimization of the energy resolution while keeping the network size under control
  - Vary the network parameters: internal dimension (units), sliding window size (timesteps)
  - Network trained with simulated data of a single LAr calorimeter cell using the AREUS software

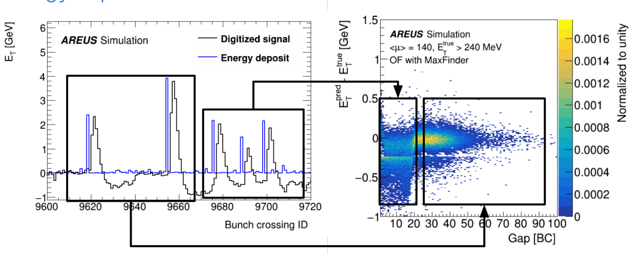


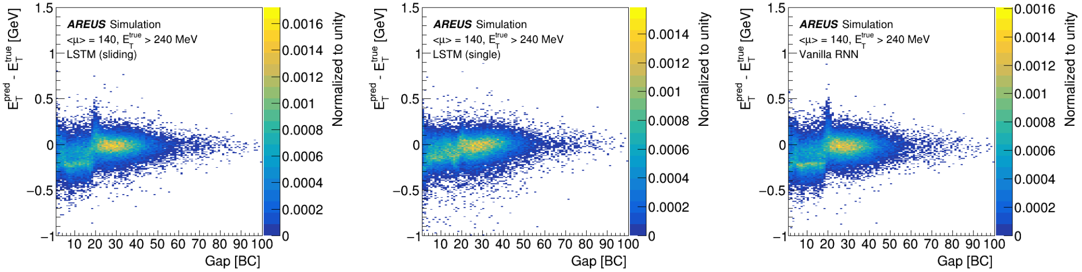



# RNN Performance

Resolution as a function of gap to previous energy deposit in BCs

- Vanilla 89 params, LSTM 496 params
- Clear performance decrease with OFMax at low gap
- All RNNs perform better with overlapping events

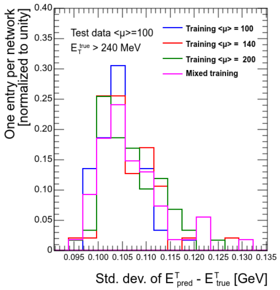


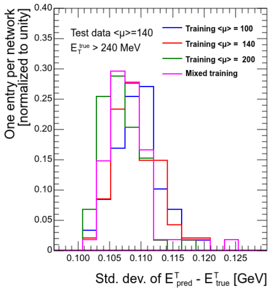



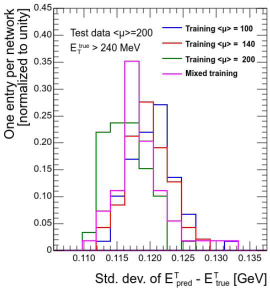
# Network Robustness

## Against pileup ($\mu$) for Vanilla RNN

- Resilience against varying pileup (simultaneous p-p collisions per BC)
- Train 276 models with different pileup rates, cross evaluate
- The networks show resilience against varying pileup

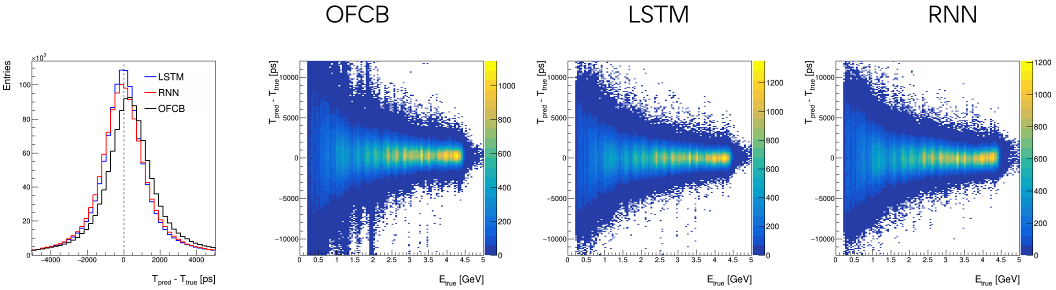


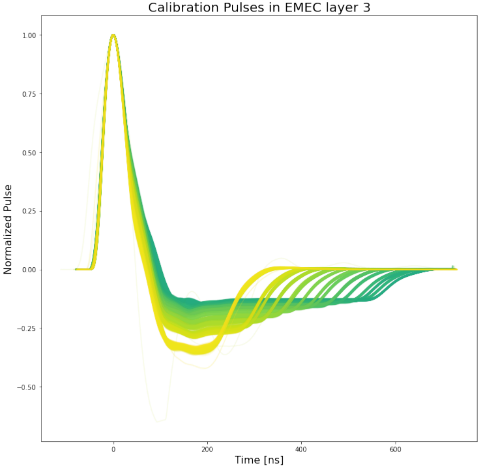
# Reconstructing Energy and Timing

- Time shift in pulse shape is also computed by OF (OFCB)
- This value is used when determining the quality of the pulse
- It could possibly also be used in discovering long-lived particles
- Adding timing computation to RNN adds only few extra parameters
- RNNs reconstruct the phase shift better than OF



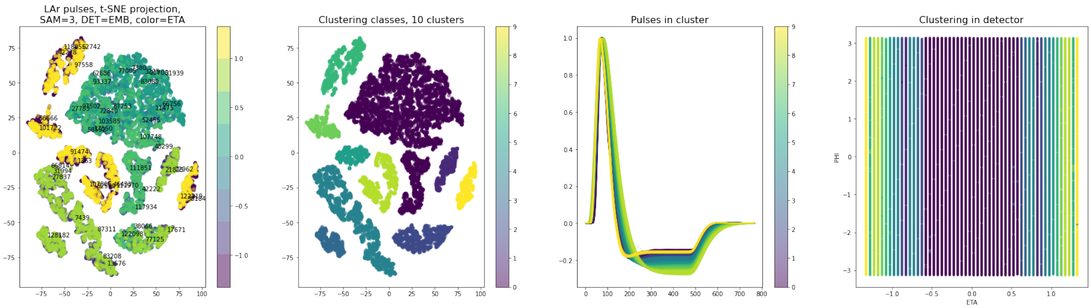
# Reconstruction for the Full Detector

- LAr consists of barrel (EMB) and endcaps (EMEC) which have 4 layers each
- Significantly different pulse shapes for different parts of the detector
- Example of 10138 pulses in EMEC layer 3
  - The color denotes the abs(ETA) value
- One NN training will not perform well for the full detector, nor is 182k NNs feasible
- It is essential to find a way to reduce the amount of NNs while keeping high accuracy



# Reconstruction for the Full Detector

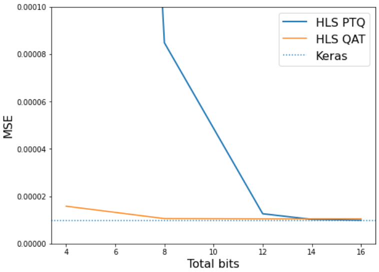
- Use t-SNE for dimensionality reduction for LAr calibration pulses to acquire 2D representation
- DBSCAN unsupervised clustering to group LAr cells with different pulse shapes
- Able to distinguish real differences in pulse shape with good ETA separation

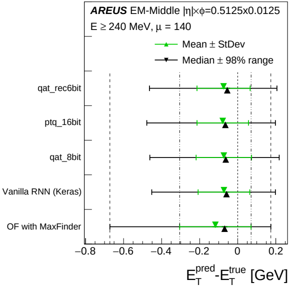


# Quantization Aware Training

## Optimizing NNs for firmware

- FPGAs operate with fixed point of arbitrary bitwidth instead of 32bit floating point numbers
- Using lower bitwidth numbers reduces the resource usage
- Quantizing NNs after training (PTQ) with floating point variables decreases the accuracy
- It is possible to mitigate this effect with quantization aware training (QAT)
- Simulation results from High Level Synthesis (HLS) implementation of RNNs show that the required bitwidth can be halved by using QAT

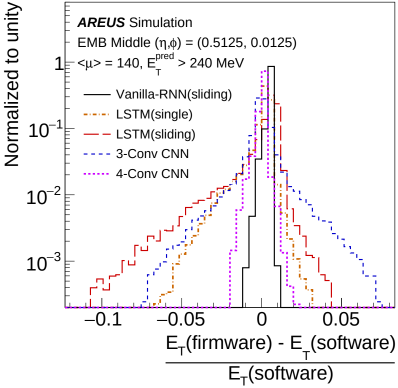


# Firmware Implementation

## Running in the FPGA

- Single FPGA processing of 384 cells requires special implementation
- Multiplexing implemented to serialize several parallel networks
  - Run 10 parallel networks, each computing 37 RNN cells within the 25 ns input interval
- HLS does not achieve required latency for Phase-II specifications
- VHDL implementation based on the HLS acquires a latency of 121 ns using 28x14 multiplexing




| | N networks x multiplexing | ALM | DSP | FMax | latency |
|---|---|---|---|---|---|
| **target** | **384 channels** | **30%*** | **70%*** | **-** | **125 ns** |
| HLS (no multiplexing) | 384x1 | 226% | 529% | - | 322 ns |
| HLS optimized | 37x10 | 23% | 100% | 414 MHz | 302 ns |
| VHDL optimized | 28x14 | 18% | 66% | 561 MHz | 121 ns |

*based on experience with the phase I upgrade

# Table of Contents

# Conclusion

Energy reconstruction using recurrent neural networks

- Energy reconstruction with RNNs overperform legacy algorithms in Phase-II conditions
  - Better energy resolution overall
  - Better recovery of energy resolution with overlapping signals
- Clustering to reduce the amount of required NNs
- Implemented and validated in firmware and mostly fulfills the LAr real-time processing requirements
- Next steps: performance evaluation in full detector simulation
- Paper published available ▸ Here