Optimisation of embedded neural networks for the energy reconstruction of the Liquid Argon Calorimeter cells of ATLAS

-

CPPM Seminar 3rd year PhD
27.10.2025

Raphaël Bertrand

CENTRE DE PHYSIQUE DES PARTICULES DE MARSEILLE
CPPM

ATLAS EXPERIMENT

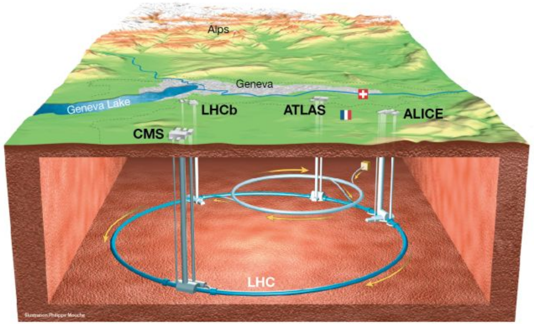
cnrs

amU Aix Marseille Université

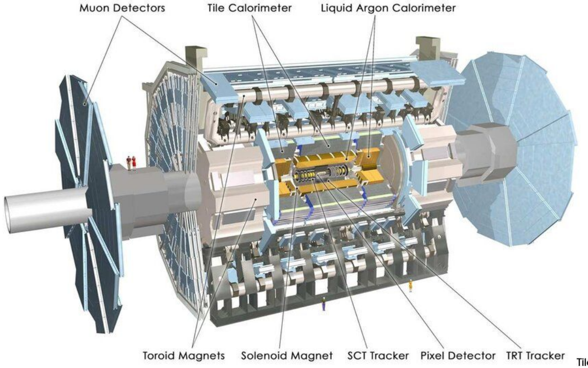
Introduction

Experimental Context

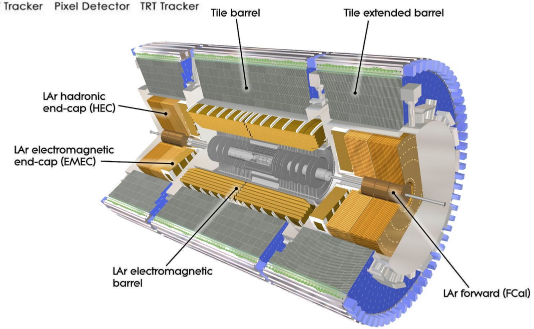
- **Large Hadron Collider (LHC)**
  - Proton-proton collider at 13.6 TeV
  - Protons accelerated via superconducting magnets
  - Collisions at 40 MHz



- **ATLAS detector**
  - General-purpose experiment
  - Very high data rate
    - ➢ On-the-fly event selection required

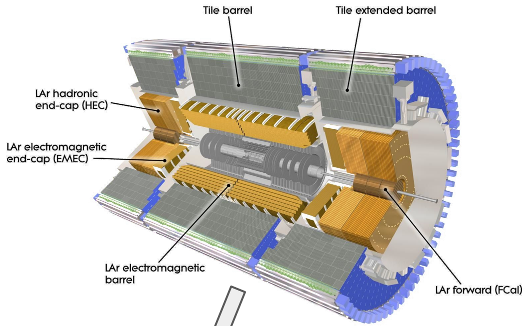


- **Liquid Argon (LAr) Calorimeter**
  - ATLAS sub-detector for energy measurement ($e^{-/+}$, γ)
  - Sampling in active LAr alternating with inactive metal (Cu, Pb, W)
    - ➢ Accordion shaper absorbers for EMB and EMEC
    - ➢ Ionization signal from particle interactions

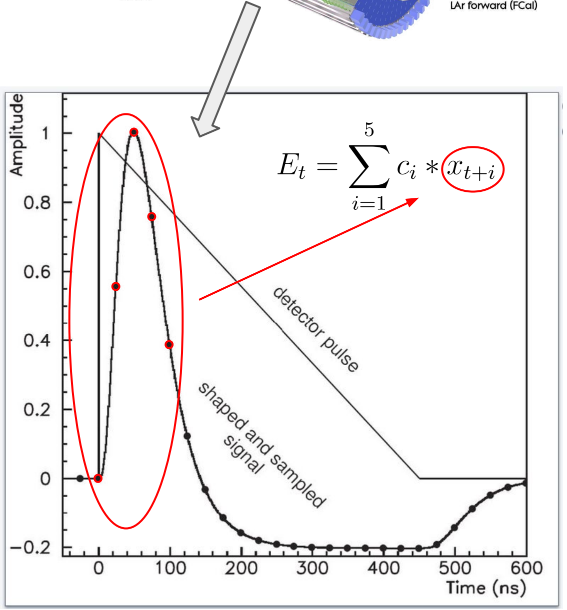


Signal processing and energy reconstruction

- **Electronic signal** produced
  - **Amplitude ∝ true deposited energy (E^{true})**
  - **Spans ~625 ns** (25 proton-proton **B**unch **C**rossings)
  - **Shaped**, **sampled** and **digitised at 40 MHz**

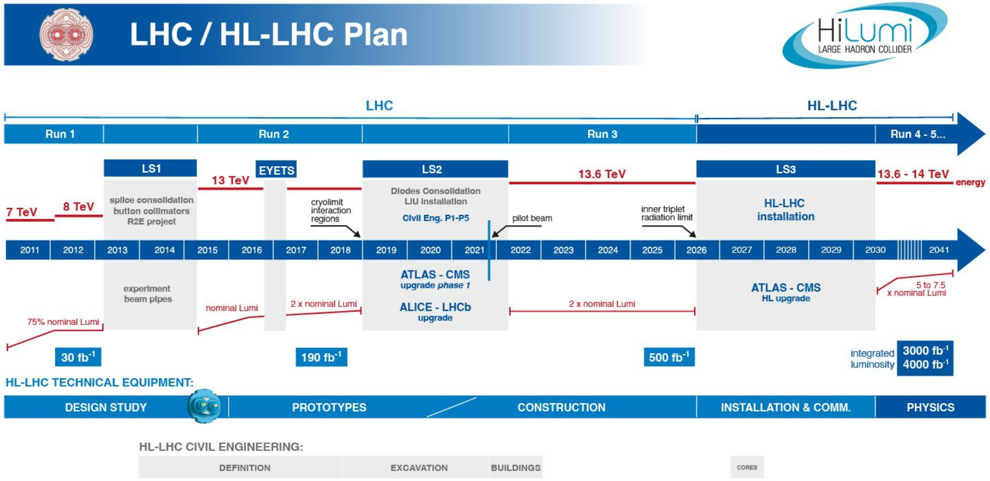


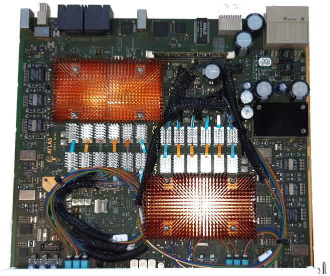
- **Energy reconstruction** with optimal filtering (OF) algorithm
  - **Weighted sum** of samples around the pulse peak
  - **Max finder/Timing cut** to select the correct BC



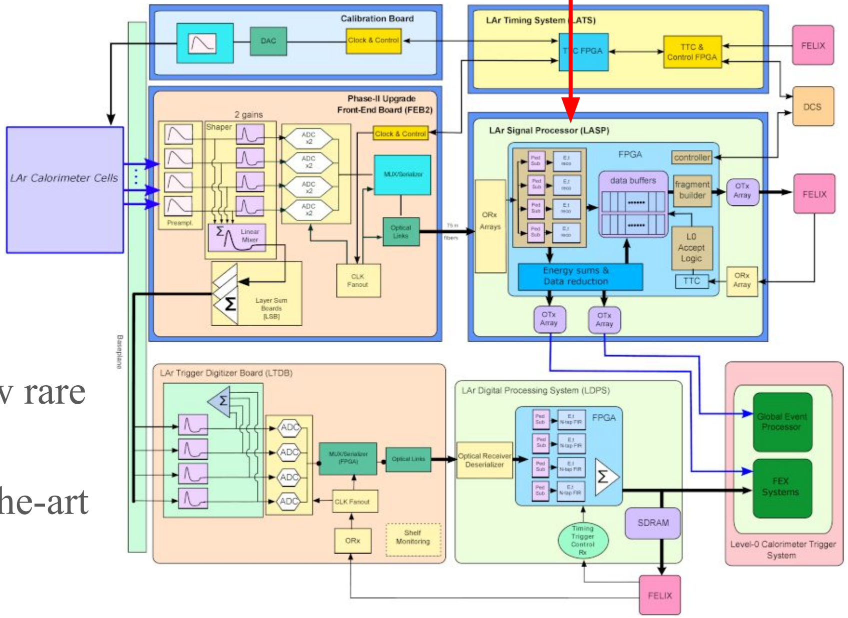
- Reconstruction algorithm requirements :
  - **Online** computation (per BC)
  - **Max latency : ~125 ns** (used in trigger system)
  - **Fit in FPGAs** : **O(500)** Multiply-Accumulate operations (**MAC units**)
    - ➢ 5 MAC units required to implement OF
  - **384 channels per FPGA** (many algorithm instances needed)

HL-LHC schedule and ATLAS Phase-II Upgrade





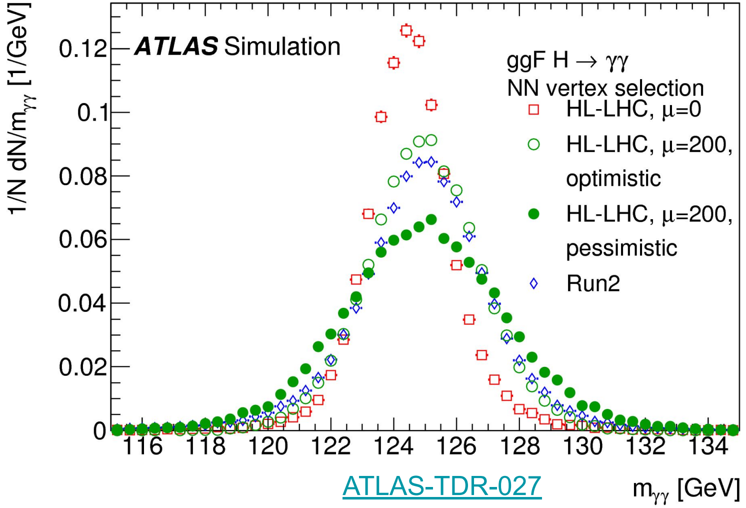
LASP board demonstrator



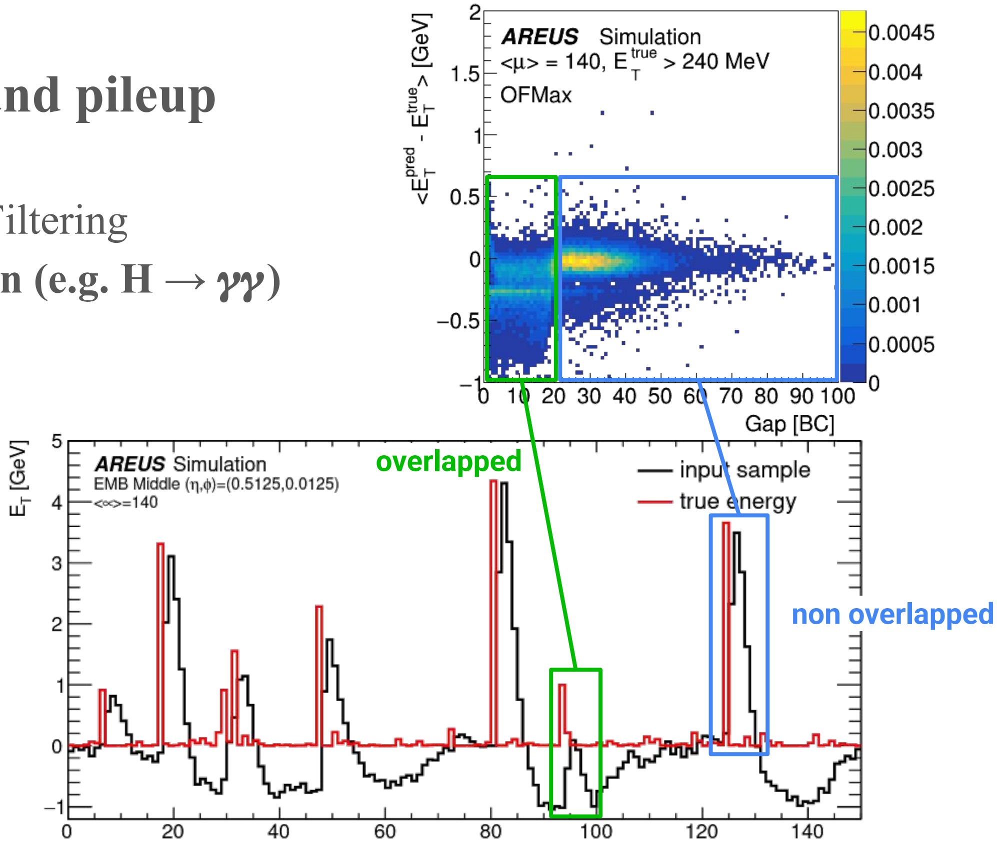
- HL-LHC ⇒ Luminosity ↗
- HL-LHC is needed to study Higgs properties and detect new rare processes
- Off-detector readout board (LASP) will carry two state-of-the-art **FPGAs for energy computation**
    - **Opportunity to embark more complex algorithms**

Impact of high luminosity

- HL-LHC ⇒ **Increased luminosity and pileup**
  - **Increased rates of overlapping pulses**
  - ↳ **Degraded performance** of Optimal Filtering
  - **Significant impact on energy resolution (e.g. $H \rightarrow \gamma\gamma$)**



ATLAS-TDR-027



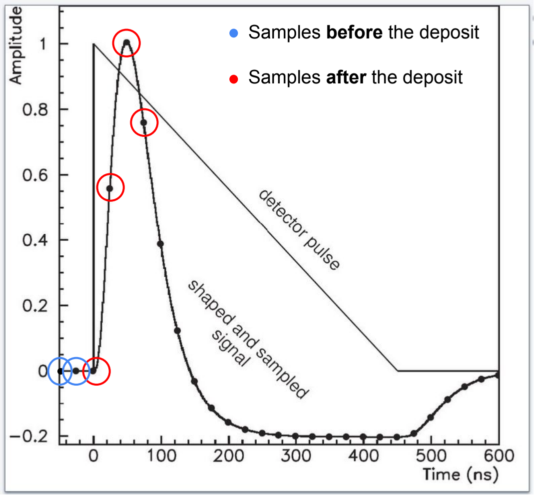
Simulated pileup with additional injected pulses at higher energy

Neural network approaches as energy reconstruction algorithms

Neural networks

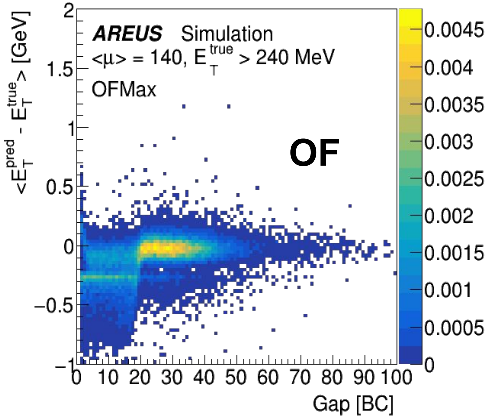
- Exploit samples before the energy deposit to **correct overlapping pulses**
- Several architectures tested : RNN, RNN+Dense, CNN, Dense

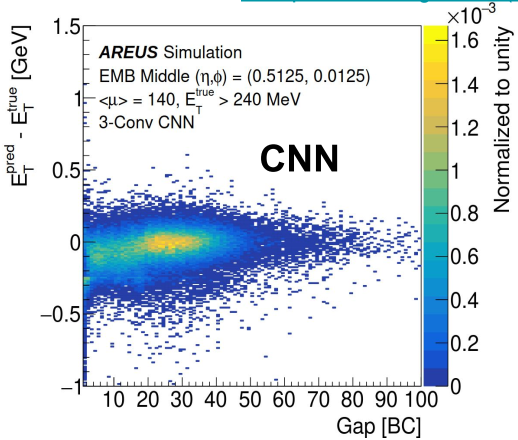
- **Samples** from **before** and **after** the energy deposit are used :
  - **After the energy deposit** (similar to OF inputs)
    - ➢ Capture the pulse amplitude
  - **Before the energy deposit** (additional inputs)
    - ➢ Correct for pulse distortions from previous deposits

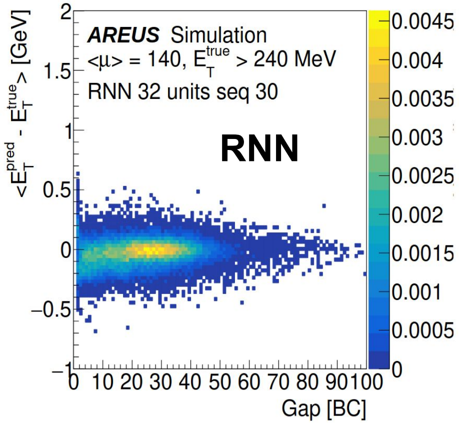


- Preliminary studies done with high rate of pulse overlap
  - **Neural networks can correct for overlapping pulses**
    - The correction is **dependent on the size** of network

Comput.Softw.Big Sci. 5 (2021), s41781-021-00066-y







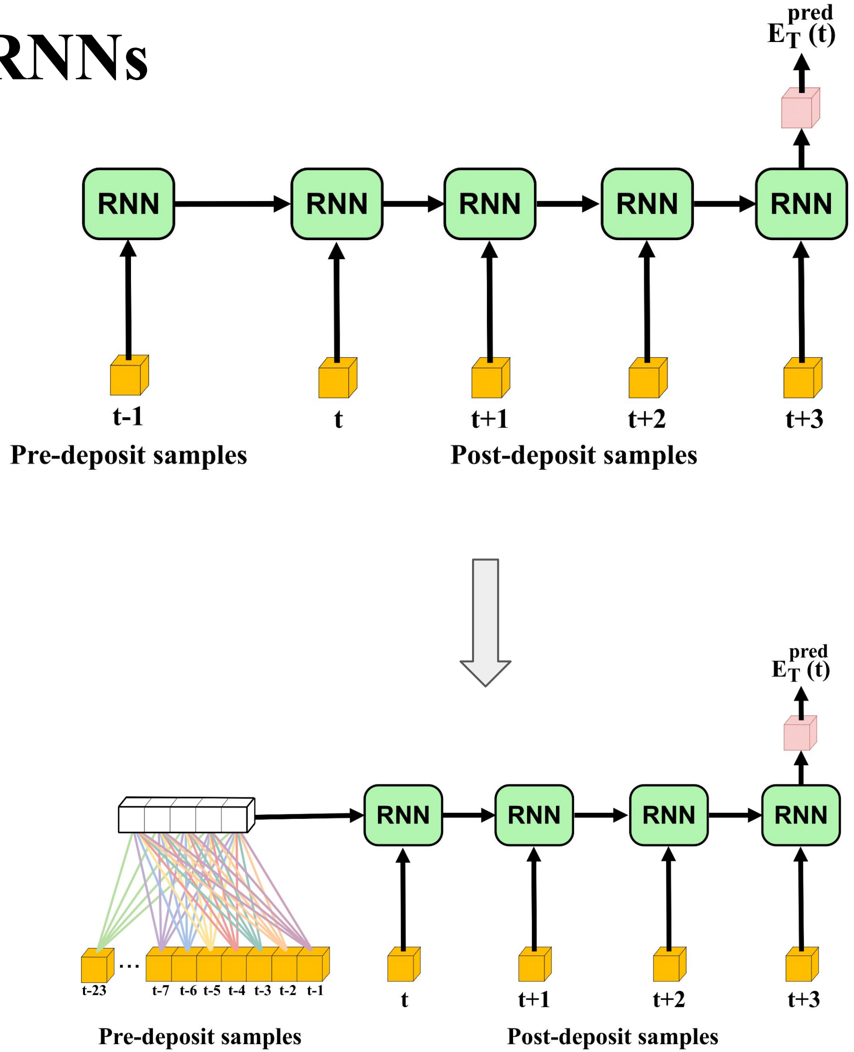
Neural network architecture - RNNs

- Architecture **RNN**
  - Multiple **RNN cells sharing the same parameters**
    - ➢ **One cell per sample**
  - One dense layer to concatenate output from last cell
    - ➢ **Return predicted energy**
- The RNN architecture assigns **equal importance to all samples**
- Start computations at first sample
  - **Good latency**
- FPGA implementability :
  - Number of **Multiply-Accumulate Operations** (**MAC units**)

- Architecture **Dense+RNN** → **additional dense layer** before RNN cells
  - Computation for **pre-deposit samples**
  - **Less MAC units** for a dense layer

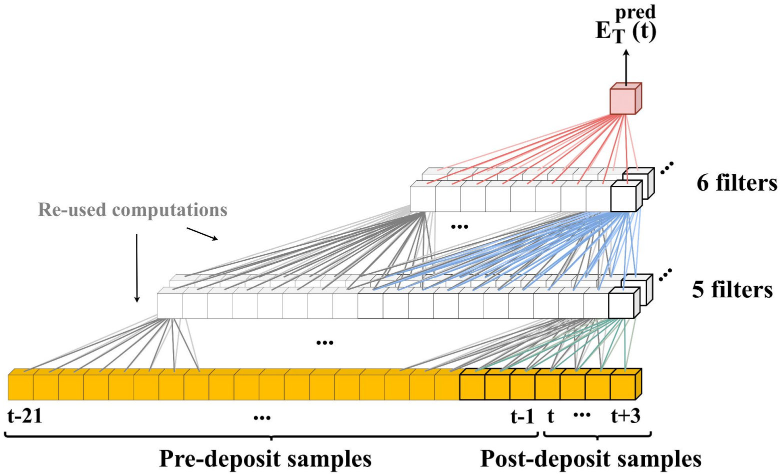
⇒ MAC units $_{\text{RNN layer}}$ ∝ units^2 x nb of samples

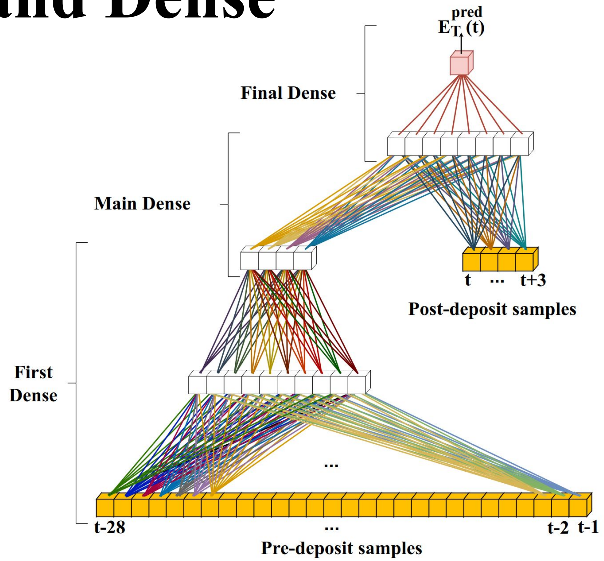
⇒ MAC units $_{\text{Dense layer}}$ ∝ units x nb of samples



Neural network architecture - CNN and Dense

- Architecture CNN ⇒ **Three CNN layers**
- Uncertain latency compliance
- FPGA implementability :
  - **Re-used computations**





- Architecture **Dense**
  - One block of dense for **aggregation of pre-deposit samples**
  - One dense layer to add a layer of computations before concatenation
  - One dense layer to concatenate output from last cell
    - ➢ **Return predicted energy**
- **Good latency**
- **Smaller number of MAC units** with dense layers

Neural networks hyperparameters tuning

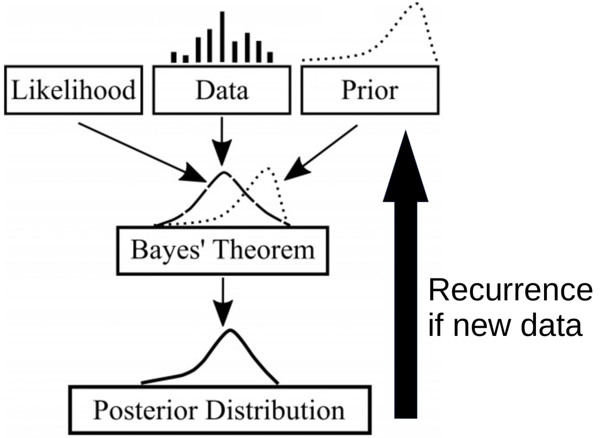
bayesian optimization

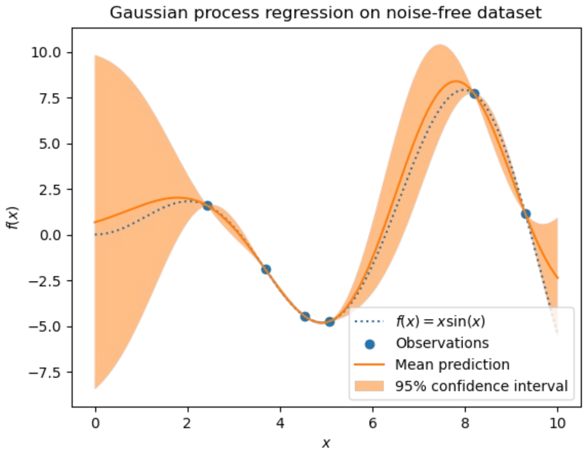
Bayesian optimisation

- Goal : **Find the best parameters** to maximize/minimize a **performance function** while **evaluating the function as few times as possible**

- **Initialization** with several random points

- **Iterations** to find the best parameters space
  - **Interpolation** between points
    - Based on a gaussian kernel with associated uncertainty
  - **Acquisition function** to determine where to evaluate next
    - Balance between **exploration** and **exploitation**
  - **Evaluation** of the performance function **at the chosen point**





Bayesian optimisation applied on energy reconstruction

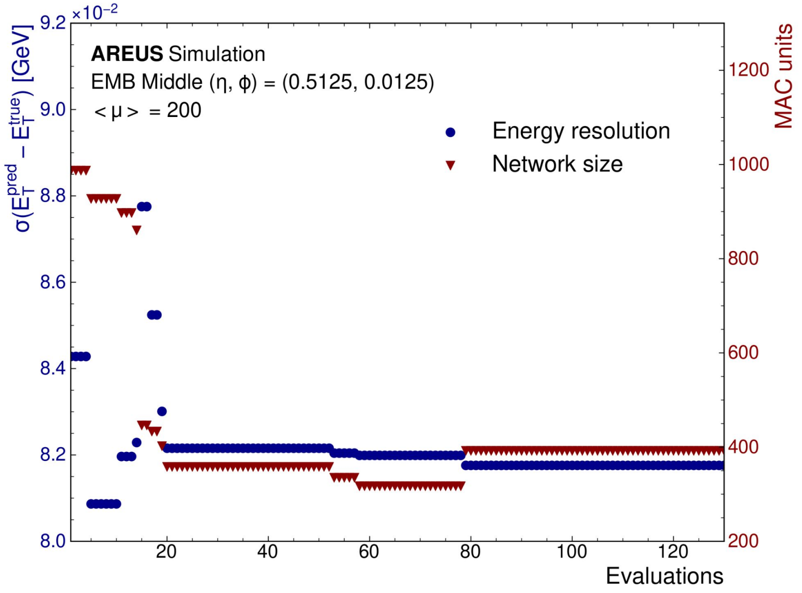
- Optimisation on both performance and hardware to fit in FPGAs
    - **Energy resolution** (σ [MeV])
    - **Number of MAC units** (M)
- Hyperparameters to be tuned (e.g. for the Dense architecture) :
    - **Number of samples** (before the energy deposit)
    - **Number of units** for the intermediate layers

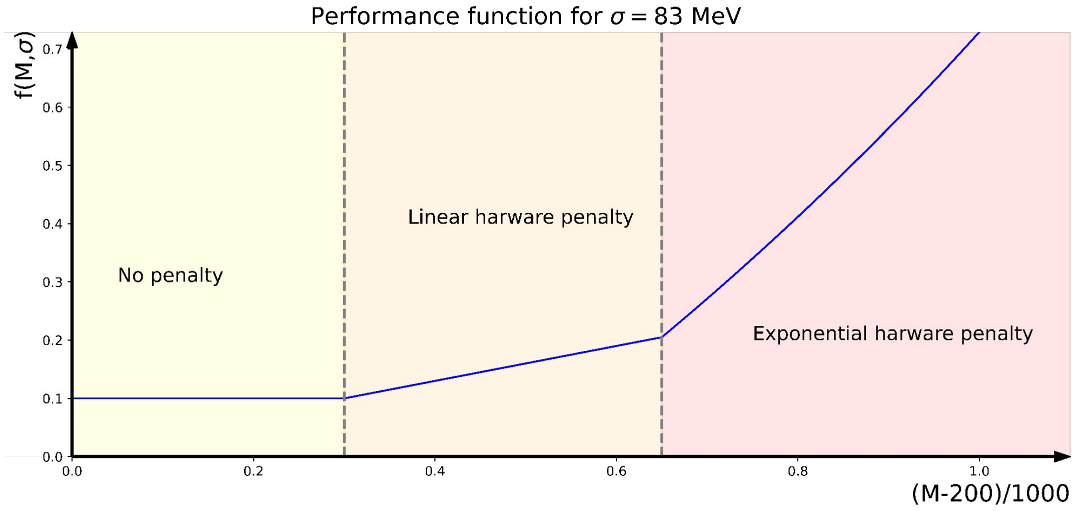
Performance function used for the bayesian optimization :

$$f(M,\sigma) = \frac{\sigma - 70}{130} \; for\ M \leqslant 500$$

$$f(M,\sigma) = f(500,\sigma) + a * \frac{M-500}{1000} \; for\ M \in]\,500\,;850\,]$$

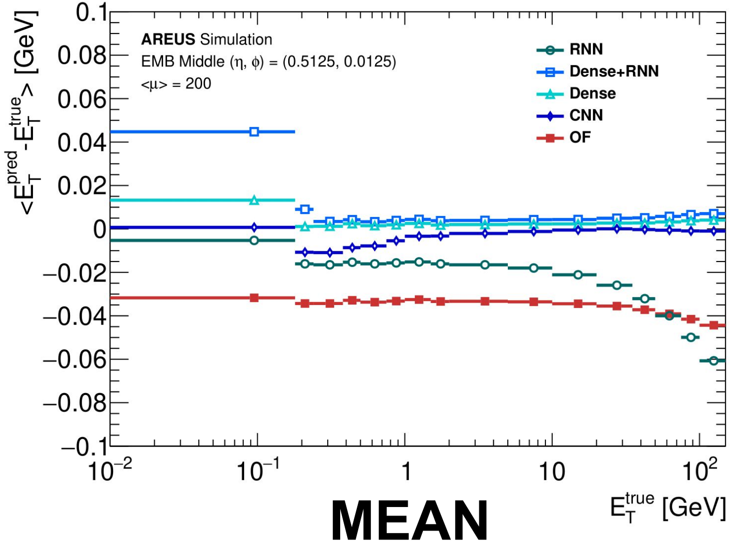
$$f(M,\sigma) = f(850,\sigma) + b * e^{\frac{M-850}{1000}} - 1 \; for\ M > 850$$





Energy scale and resolution as function of true energy

- **Better energy scale** of $E_T^{pred} - E_T^{true}$ **for Dense+RNN, Dense and CNN** architectures compared to OF
  - RNN energy scale falling with higher E_T^{true}
  - OF energy scale not centered around 0
- **Better energy resolution** of $E_T^{pred} - E_T^{true}$ **for Dense+RNN, Dense and CNN** architectures compared to OF
  - Visible for the whole energy range

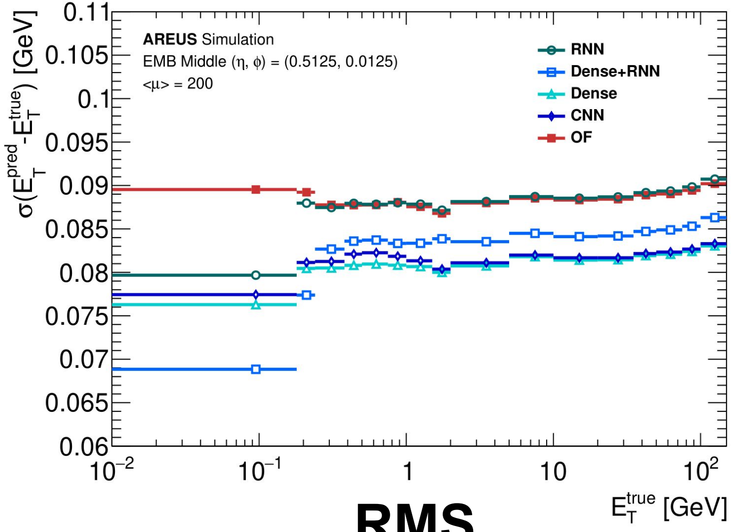


MEAN

MAC units
368
240
392
419
5

Pre-print
(submitted to EPJC)

Public code



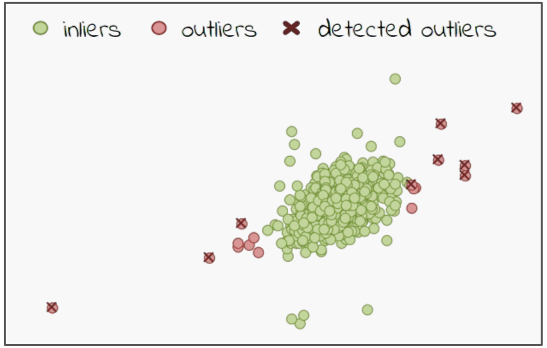
RMS

Uncertainty prediction using neural network

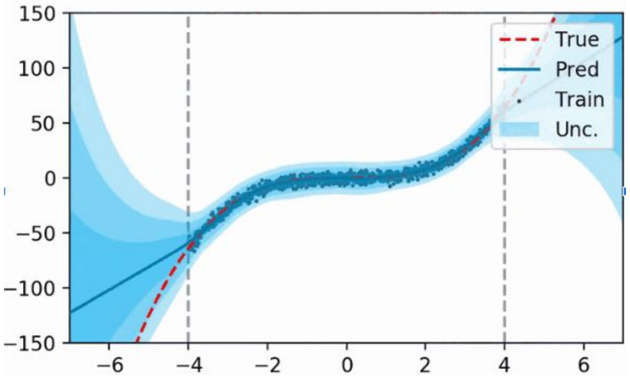
with deep evidential regression

Deep evidential regression (DER)

- NNs are trained to minimize their prediction errors
  - Unknown accuracy of the model for individual prediction
  - It would be interesting to **know when the model is more likely to fail (or the opposite)**
- **Model the energy prediction as a distribution**
  - Mean of the distribution → **energy prediction**
  - Standard deviation of the distribution → **uncertainty**



- Differentiate uncertainties :
  - **Epistemic**
    - ➢ Lack of knowledge, model uncertainty
    - ➢ Can be reduced
  - **Aleatoric**
    - ➢ Inherent to data
    - ➢ Cannot be reduced



[Deep Evidential Regression, Amini et al.](#)

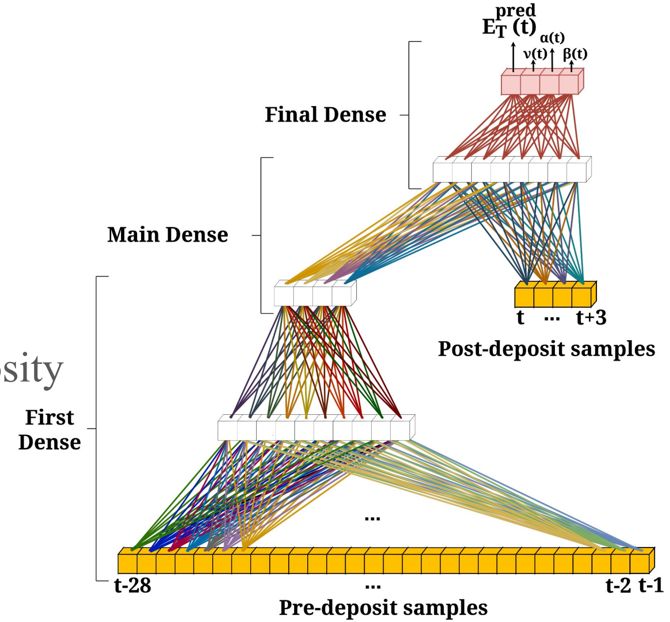
Deep evidential regression (DER)

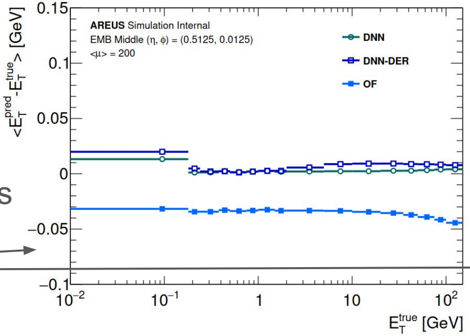
- Normale-Inverse Gamma distribution to describe mean and uncertainty
    - **4 parameters** (γ,ν,α,β) rather than one
        - Uncertainty computation
- **DER applied to LAr cells energy reconstruction**
    - Would allow to take into account instantaneous luminosity changes or bunch train structure
- Adapted to the Dense architecture
    - **Use of bias** compared to the other architectures
    - **Still possible to implement in FPGA**
        - **416 MAC units**
- Training loss function : ~~MSE~~
    - Likelihood + Regularisation

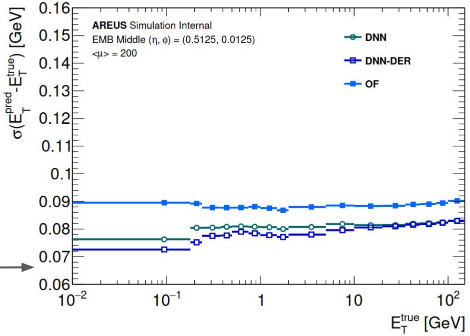
fit the distribution

increase uncertainty on large errors

- **Similar performance**

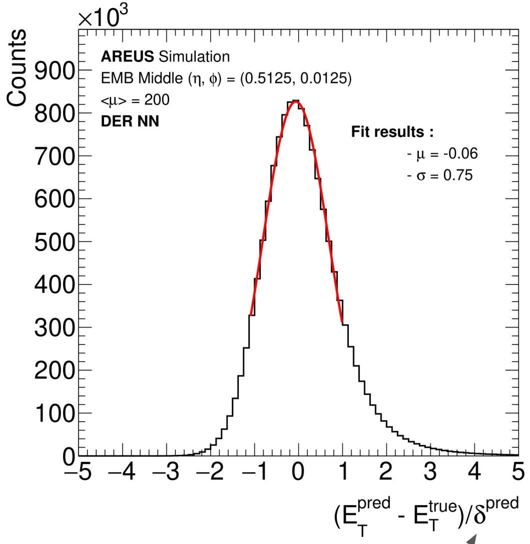




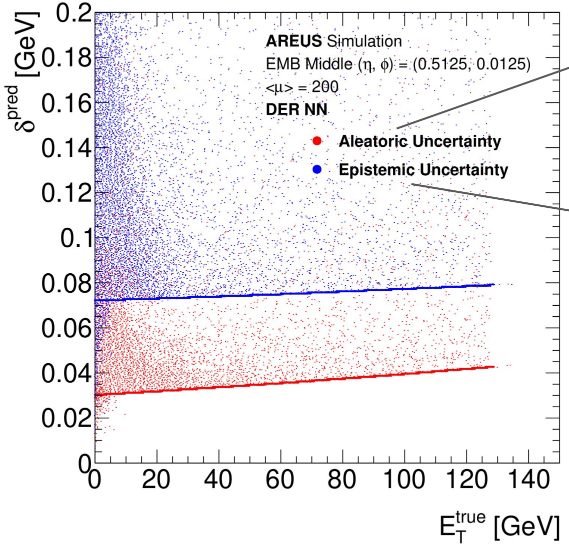


Uncertainty prediction

- Overall **good pull distribution**
  - Estimated uncertainty comparable to $E_T^{pred} - E_T^{true}$
  - **Slightly biased**
    - Right tails
    - Uncertainty overestimated by 25%



predicted uncertainty



Data uncertainty

$$\mathbb{E}[\sigma^2] = \frac{\beta}{\alpha - 1}$$

Model uncertainty

$$\mathrm{Var}[E_\mathrm{T}^\mathrm{pred}] = \frac{\beta}{\nu(\alpha - 1)}$$

- Epistemic and aleatoric uncertainties are constant
- **Epistemic uncertainty is dominant** in the total uncertainty

Conclusion

- Online **energy reconstruction for LAr cells** performed **using neural networks**
  - **Pre-print on arxiv :** **Optimised neural networks for online processing of ATLAS calorimeter data on FPGAs**
- Four neural network architectures were tested and optimized
  - **CNN**, **RNN**, **RNN+Dense** and **Dense**
- **Hyperparameter tuning** performed using bayesian optimization
  - **Balance between performance and size of the network** to fit in FPGAs
  - **NNs outperform OF**
- **Uncertainty** on energy prediction using deep evidential regression
  - **Accurate** uncertainty prediction
  - **Possible to implement in FPGAs**

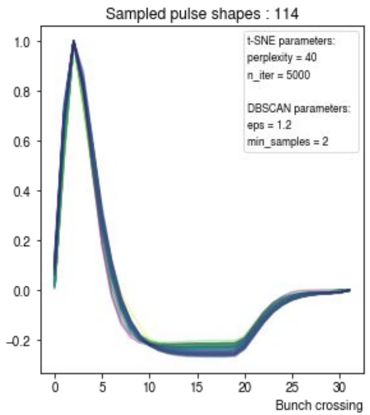
BACKUP

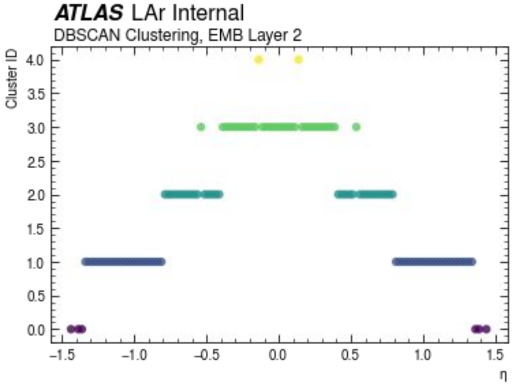
Bayesian optimisation code

- **Parallel iterations**
  - **Multiple models (with different parameters) trained simultaneously**
    - **Min distance** defined to avoid to train models on the same peak/minimum
      - Fixed
      - Decreasing
      - At lowest
    - **Exploration parameter** (to favour high-uncertainty regions)
      - Fixed
      - Fixed
      - Decreasing
- **Noise considerations**
  - Choice 1 : **multiple models trained at the same parameters, only the best one is considered**
  - Choice 2 : multiple models trained at the same parameters, all used to train the Gaussian Process Regressor
- **Integers considerations** :
  - **Only integer parameters are proposed**
  - Two iterations can't be on the same set of parameters
- **Kernel choice : Matern 5/2**
  - Twice differentiable
  - **More realistic** than usual RBF/Gaussian kernel (infinitely differentiable)

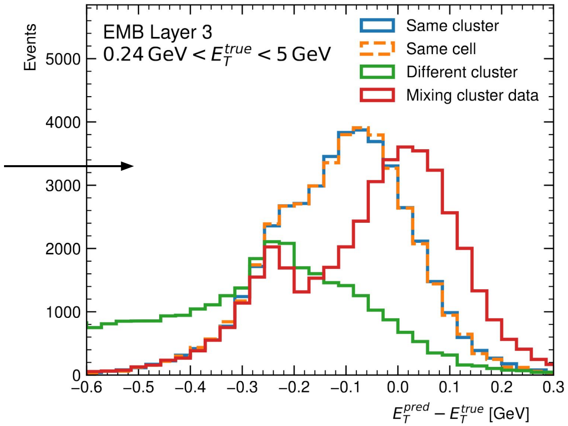
Cells grouping

- Training one network per cell is not feasible
- LAr calorimeter cells grouped
  - One neural network ⇔ several cells
  - 182,468 cells → **few hundreds groups**
- Grouping performed using techniques known as t-sne and dbscan
  - Groups are trained layer per layer
  - The grouping recovers the detector symmetry





Single cell training performance fully recovered →



Lauri Laatu PhD Thesis