Université de Montréal

Training a convolutional neural network with real data for electron identification in ATLAS

Presented by : Olivier Denis
Supervisor : Jean-François Arguin
June 19th 2023

Overview

1) The importance of electrons for LHC physics

2) Electron identification in ATLAS

3) Convolutional neural networks
   - Architecture
   - Training with Monte Carlo

4) Experimental data sample
   - Can we improve even further?
   - Sample purity
   - MC vs Data

5) Conclusion

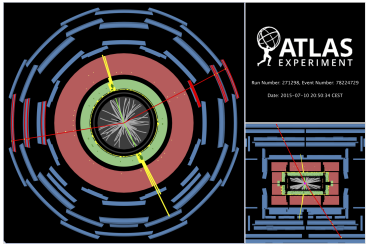
The importance of electrons for LHC physics

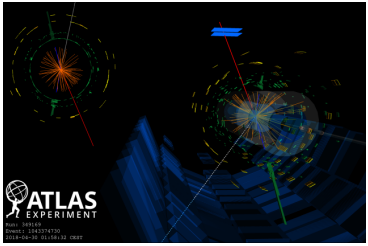
The importance of electrons for LHC physics

- Particles of second or third generation, as well as force carrier, are unstable.
- They decay into particles of first generation.
- Electrons are first generation particles $\Longrightarrow$ they are particularly important for analysis with leptons in the final state.

Examples of processes with electrons in the final state

- Vector bosons decay : $W \longrightarrow e\nu$ and $Z \longrightarrow ee$;
- Higgs bosons principal decays : $H \longrightarrow W^+W^-$, $H \longrightarrow ZZ$, $H \longrightarrow \tau^+\tau^-$, $H \longrightarrow Z\gamma$;
- Some beyond the Standard Model (BSM) phenomenon : SUSY, vector-like quark, BSM Higgs, etc.





Electron identification in ATLAS

Electron identification in ATLAS

The current algorithms used for electron identification (e-ID) in ATLAS are the Likelihood and the DNN.

Background classes

- Charge flip (CF);
- Photon conversion (PC);
- Heavy flavor (HF) ex.: $B \to eX$;
- Electromagnetic light flavor (LF$e\gamma$) ex.: $\pi^0 \to \gamma\gamma$;
- Hadronic light flavor (LFhad) : $\pi^{\pm}$ faking electrons.

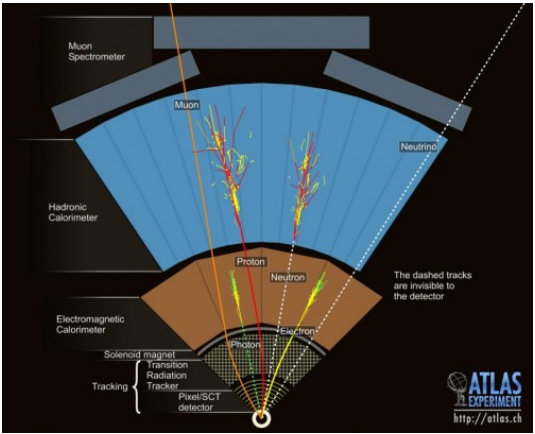


Figure 1: Typical signature of particles in the ATLAS detector

Convolutional neural networks

Convolutional neural networks - Architecture

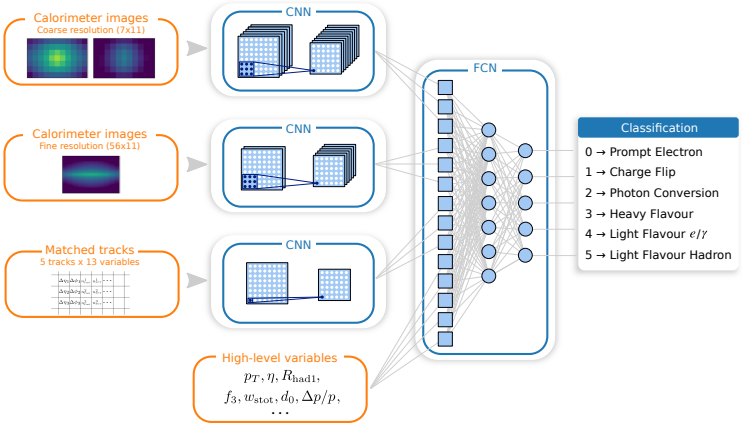


Figure 2: Global neural network architecture.

- Multi convolutional neural network (multi-CNN).
- Takes same high level variables (HLVs) than the LLH and DNN as input.
- Takes calorimeter images, tracks, and some other HLVs as additional input.
- Makes use of more detail information than its conventional counterpart, which allows it to identify electrons more efficiently

- Discriminating **performance** is assessed by comparing signal and background efficiencies (rate of acceptance) at various probability cuts.
- At 70% signal efficiency, signal purification ($\epsilon_{sig}/\epsilon_{bkg}$) can reach up to 2000 and **outperform conventional methods** by a factor 10.
- Pretty good improvements with MC trained CNN.

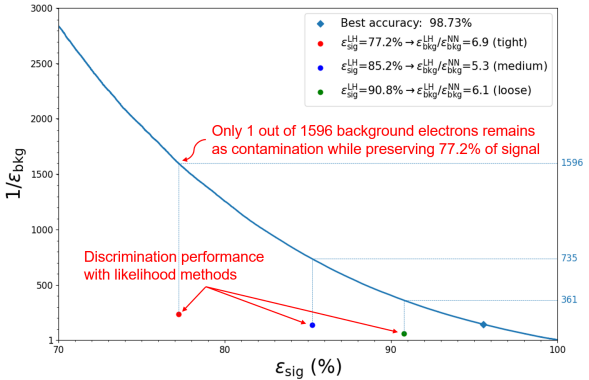


Figure 3

Experimental data sample

Can we improve even further?

- Experimental data is necessarily closer to reality than MC.
- Data is not labeled $\implies$ impossible to use for training without some preprocessing.
- The most common background is light flavor hadrons faking electrons (LF)
- We design a sample pure in LF

Experimental data sample - Sample purity

Class	Fractions (%)
Signal	2.8e-03
CF	0.000
HF	0.120
PC	0.504
LFeγ	27.572
LFhad	71.711
Other	0.089
Total LF	**99.283**

Table 1: Distribution of each class in the MC sample.

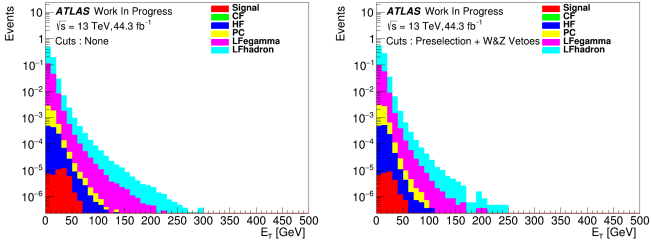


Figure 4: Monte Carlo sample composition before (a) and after (b) applying a preselection and cuts to veto W and Z bosons.

- Purifying cuts are applied to the initial data sample.
- The last set of cuts make sure to remove signal at high E_{T}.
- We obtain a sample extremely pure in LF as shown in table 1.

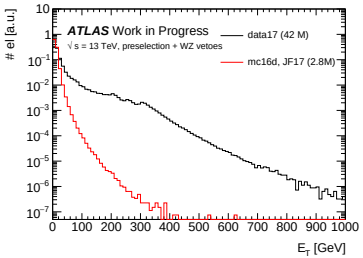
Experimental data sample - MC vs Data

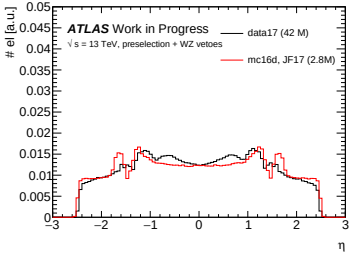
Comparing Data and MC:

- Compare the distribution of the HLVs in each sample.
- Samples have different E_{T} and η spectrum
- Narrow E_{T} and η bins and normalize to 1 for fair comparison.

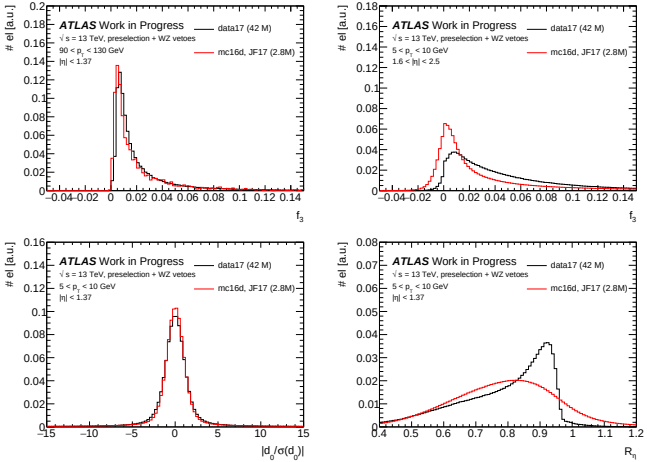
Example HLVs:

- f_3 : Ratio of the energy in the third layer to the total energy in the EM calorimeter.
- $d_0/\sigma(d_0)$: Significance of transverse impact parameter defined as the ratio of d0 to its uncertainty.
- R_η : Ratio of the sum of the energies of the cells contained in a $\eta \times \phi = 3 \times 7$ rectangle (measured in cell units) to the sum of the cell energies in a 7×7 rectangle, both centred around the most energetic cell.





Experimental data sample - MC vs Data



- Most HLVs have the same distribution between MC and Data (see left plots for example).
- However, there is a significative difference in the distributions for HLV like f_3, R_η (see right plots) or $\Delta\phi_{res}$, especially at low E_{T}.
- Suggests that Monte Carlo simulation is imperfect at lower energy $\Longrightarrow$ there is room for further improvements in e-ID.

Conclusion

Conclusion

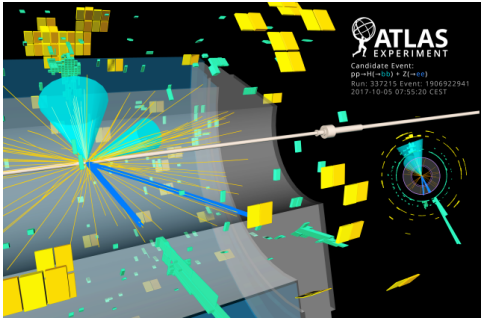


Figure 6: A beautiful event display of a Higgs decaying into 2 b quark and a Z which decays into two electrons.

- e-ID is crucial for many physics analysis in ATLAS.
- The CNN can improve e-ID performance.
- Data and MC have significant differences, especially at low E_{T}.
- Training in data should provide even better e-ID performance.