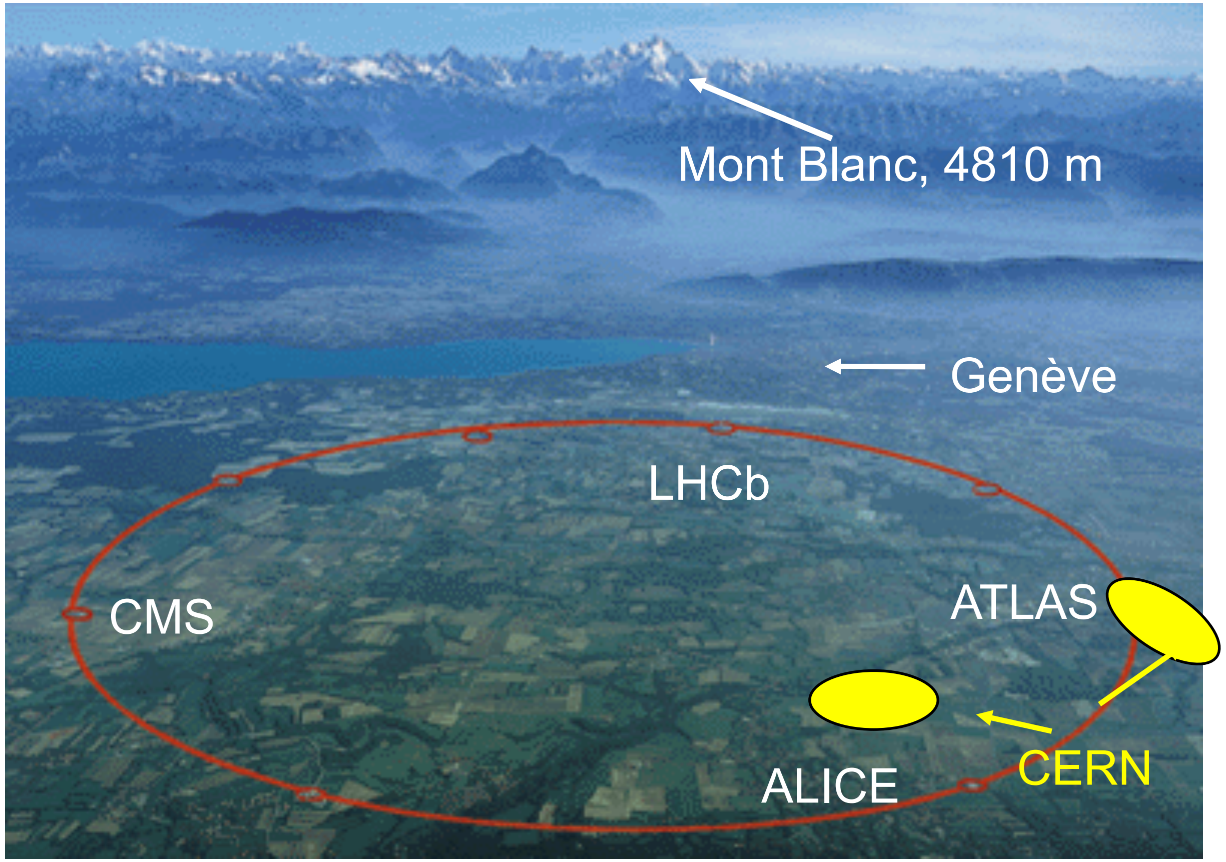
Reconstruction of $Z^0 \rightarrow e^+e^-$ in the ATLAS experiment

(and machine learning for electron identification)



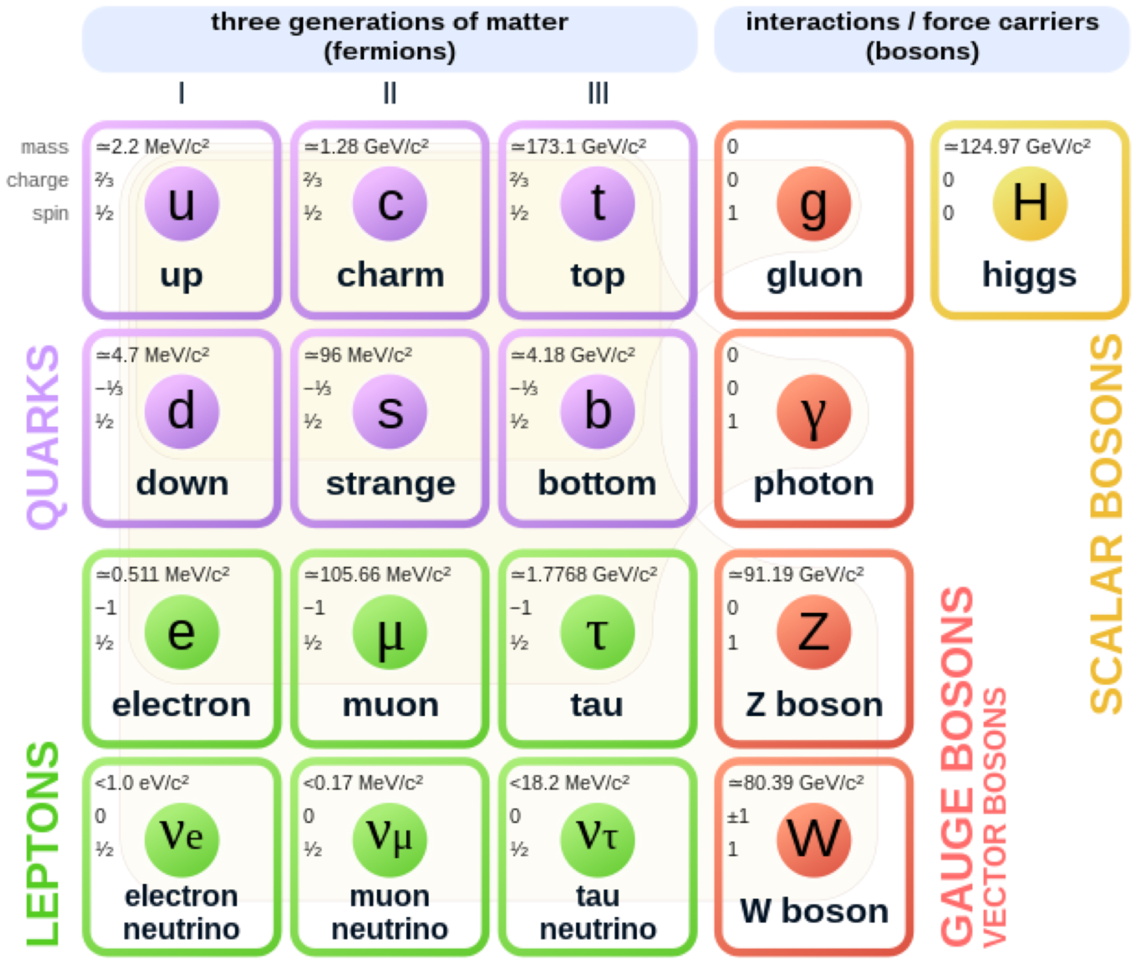
Andrew Oliver

Supervisor : Frédéric Derue
Laboratoire de Physique Nucléaire et de Hautes Energies (LPNHE), Paris

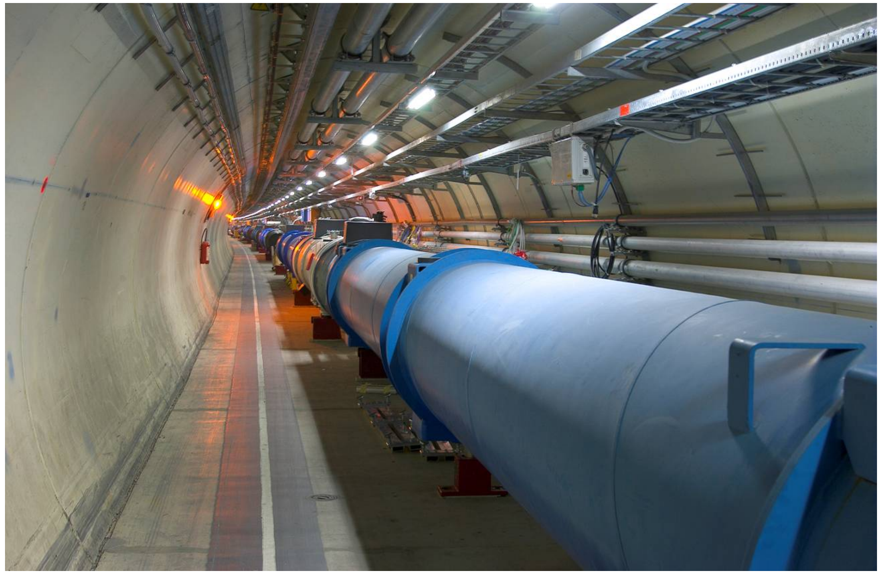
Standard model of particle physics

- Combines multiple theories
- Matter is comprised of fermions (spin-1/2):
  - Quarks feel all interactions
  - Leptons don’t feel the strong force
  - 3 generations
- Spin-1 bosons mediate interactions:
  - Gluons: strong
  - Photon: EM
  - $W^{\pm}$, Z^0: Weak

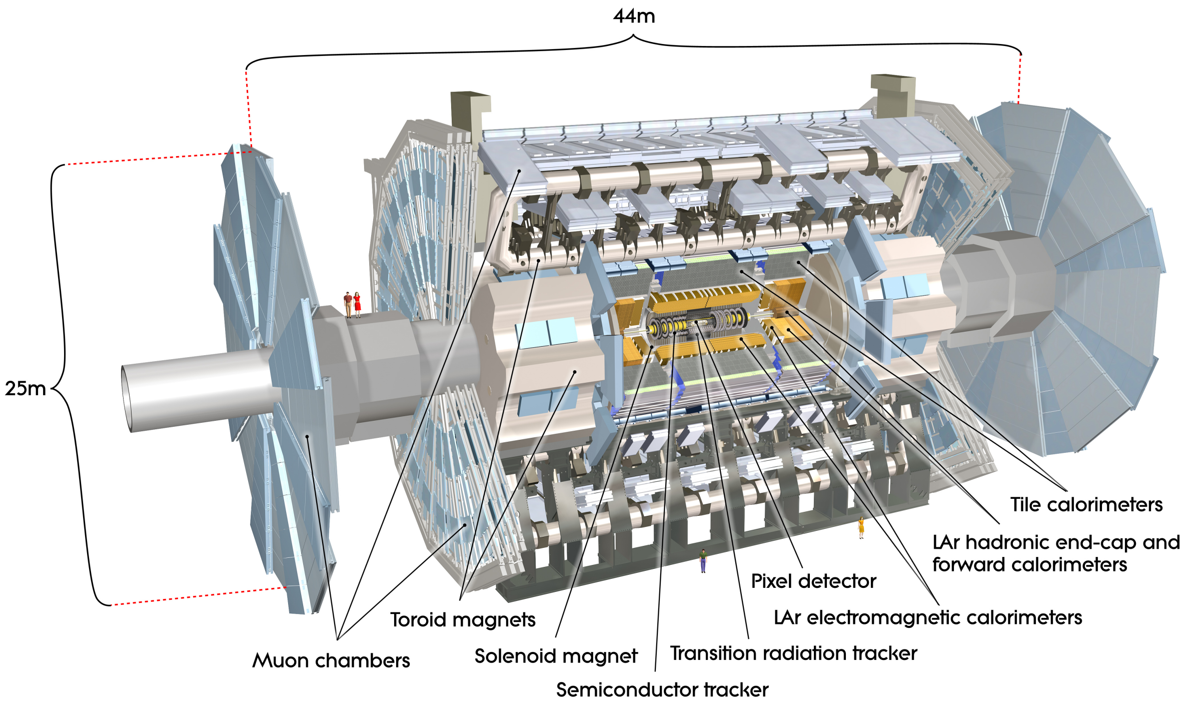
  $\left.\begin{array}{l}\text{Photon: EM}\\ W^{\pm}, Z^0\text{: Weak}\end{array}\right\}$ Electroweak
- Spin-0 Higgs explains the origins of particle masses
- Some limitations : Does not include gravitation, dark Matter and energy, asymmetry matter-antimatter, ...
  → need accelerators at order ~TeV



LHC and ATLAS



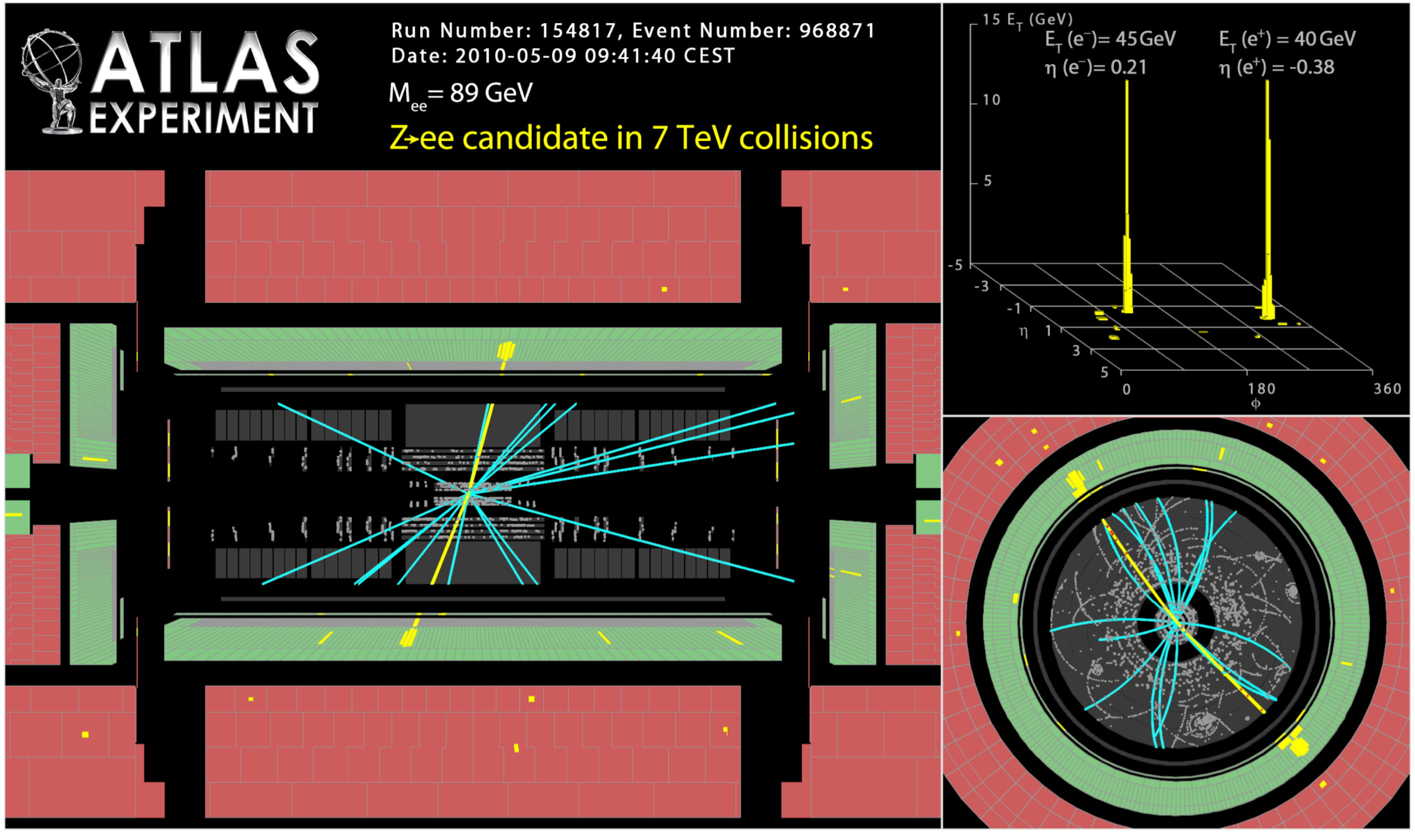
- The Large Hadron Collider (LHC) is the largest particle accelerator (proton-proton and ions) :
  - Situated near Geneva at CERN
  - In a 27 km circumference ring 100 m below ground
  - Energy of collisions : $\sqrt{s}$=7-8 TeV during Run 1 in 2011-2012, $\sqrt{s}$=13TeV during Run 2 in 2015-2018, and $\sqrt{s}$=13.6 TeV in the just starting Run 3
  - Detectors are located at each interaction point
- This internship used data from the ATLAS experiment:
  - General purpose experiment that studies pp collisions
  - Multiple aims, such as studying the Higgs and testing the SM
- The ATLAS detector has a layered structure:
  - Inner detector (ID) measures charged particle tracks and momenta
  - Calorimeters measure particles (electrons, photons, jets) energies and positions
  - Muon Spectrometer targets muons, which pass through the other sections


44m
25m
Tile calorimeters
LAr hadronic end-cap and forward calorimeters
Pixel detector
LAr electromagnetic calorimeters
Toroid magnets
Muon chambers
Solenoid magnet
Transition radiation tracker
Semiconductor tracker

Reconstruction of $Z^0 \rightarrow e^+e^-$

Selection of events and reconstruction of the Z^0 boson

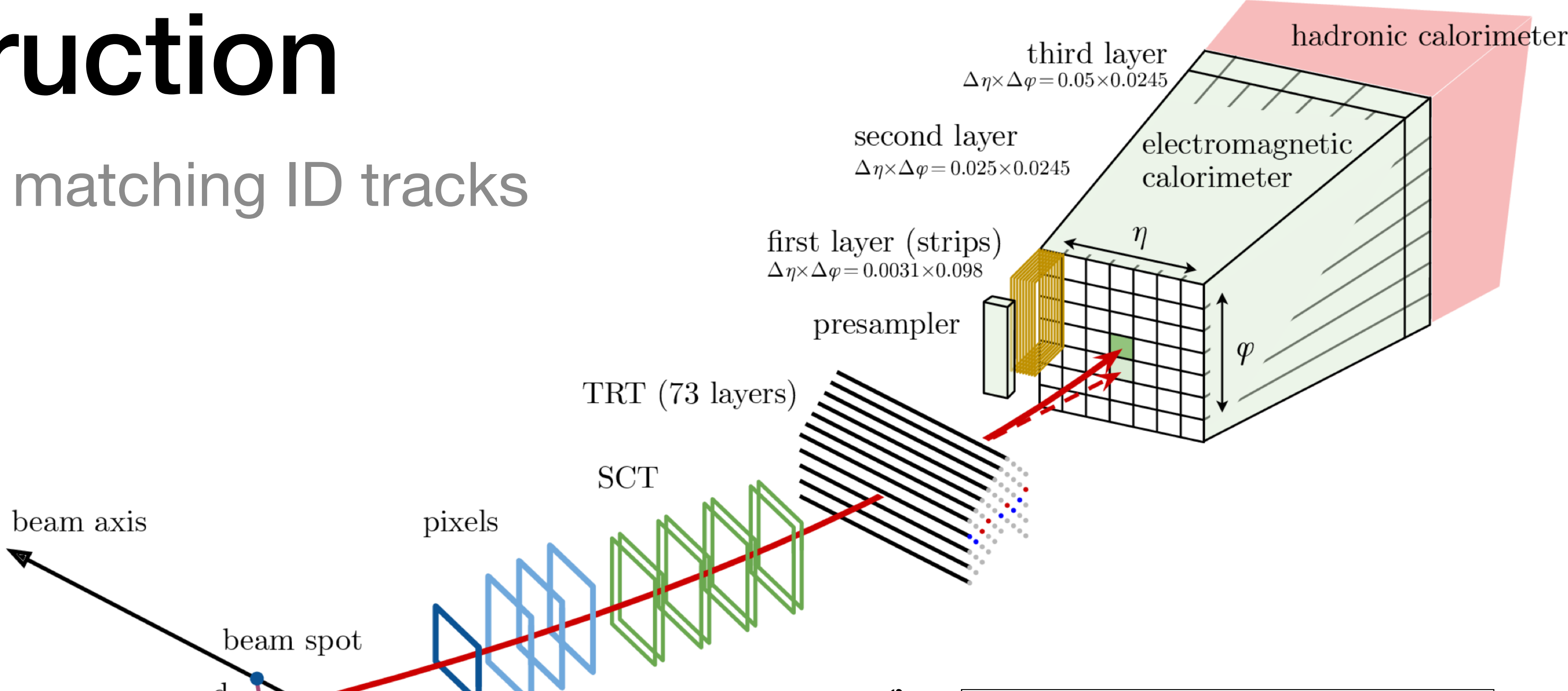
- $Z^0 \rightarrow e^+e^-$ is an excellent channel for calibration:
  - Electrons and positrons have a clean signature in the detector
  - LHC produces large quantities of Z^0 bosons
  - The Z^0 is well known so can be used as a 'standard candle'
- Data samples
  - Collision data 2015-2018 (Run 2)
    - contains ~10 million events, preselected to get at least one electron
  - Monte Carlo simulated sample
    - $Z^0 \rightarrow e^+e^-$ events were generated
    - They were fed into a simulation of the ATLAS detector to generate signals
    - Use of a simulation allows access to the 'truth' information about the particles
    - Distributions for different particle types can be created
- Analysis framework
  - ROOT (v6.20) program developed by CERN for data analysis, visualisation and storage
  - It is used by members of the ATLAS team
  - Code in C/C++



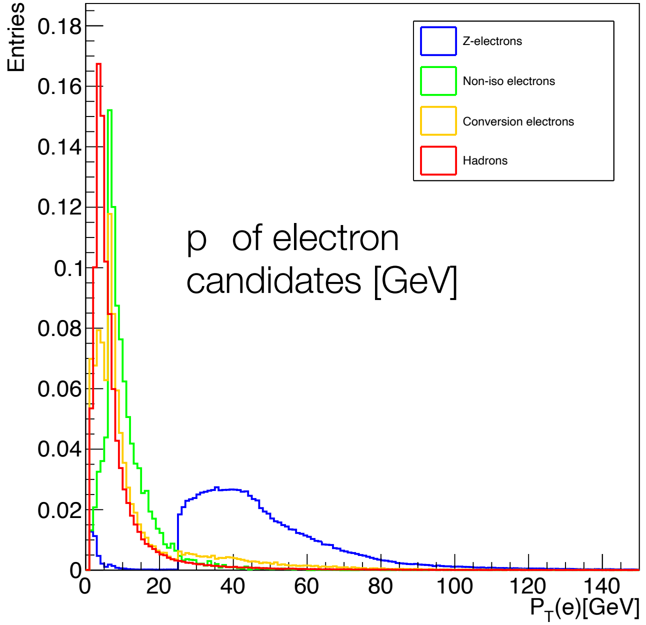
Electron reconstruction

Electrons are reconstructed by matching ID tracks to calorimeter clusters

- Each electron candidate is described by different variables:
  - Kinematic variables are related to physical quantities of the candidates
  - Discriminating variables describe the passage of candidates through the detector and the formation of EM showers
- Electron candidates can be classified using ‘truth’ information
  - We are searching for prompt electrons from the decay of the Z^0
  - Non-isolated electrons appearing in jets can appear as a result of heavy quark decay
  - Electrons can be produced by the conversion of photons into e^+e^- pairs
  - Other particles such as hadrons in light-flavour jets can mimic the signal of electrons



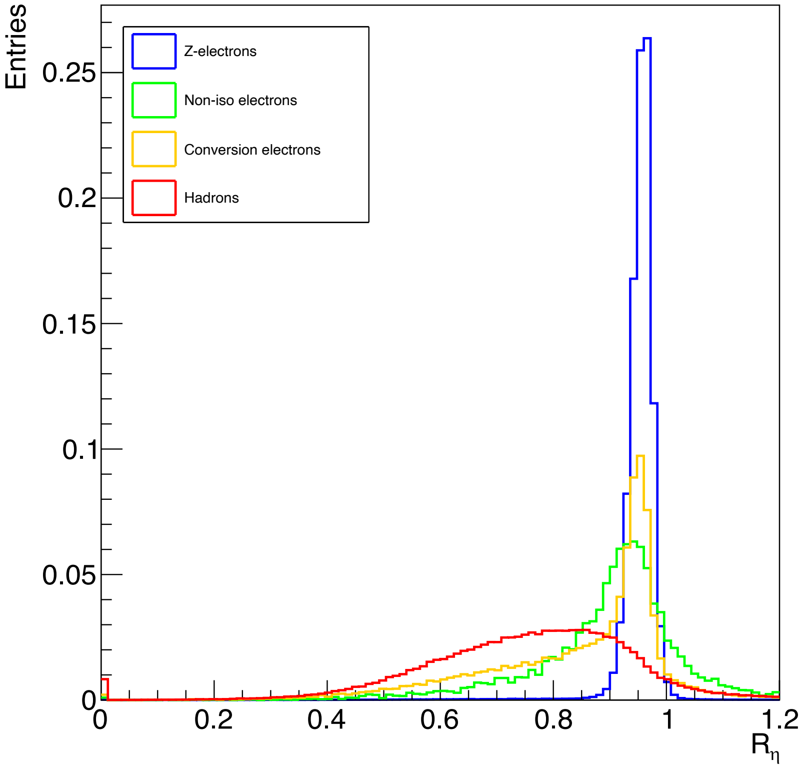
hadronic calorimeter
third layer
$\Delta\eta\times\Delta\varphi = 0.05\times0.0245$
second layer
$\Delta\eta\times\Delta\varphi = 0.025\times0.0245$
electromagnetic calorimeter
first layer (strips)
$\Delta\eta\times\Delta\varphi = 0.0031\times0.098$
η
presampler
φ
TRT (73 layers)
SCT
beam axis
pixels
beam spot
d_0
insertable B-layer



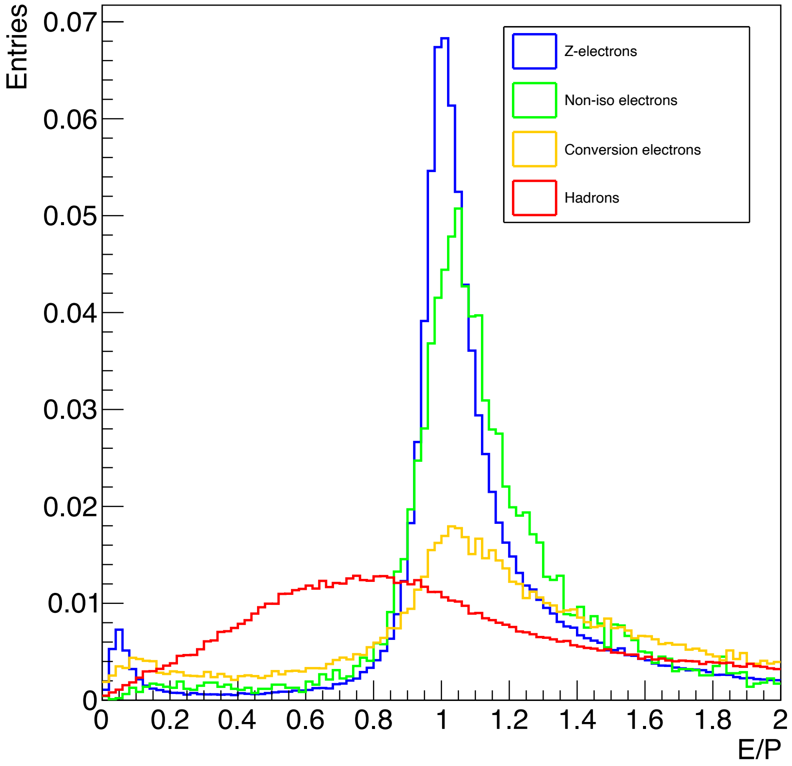
p of electron candidates [GeV]

Electron identification

Electrons are identified using the information of tracks, their thin EM shower development and the spatial & energy matching of ID/Calorimeter information



Lateral development Rη of electron candidates



E/p of electron candidates

Selection efficiencies

Definition of efficiencies as used with MC simulated samples*

Variable	Cut
p_T	$p_T \geq 25$ GeV
R_{had}	$R_{had} \leq 0.05$
R_η	$R_\eta \geq 0.87$
$\omega_{\eta 2}$	$\omega_{\eta 2} \leq 0.0125$
ω_{tot1}	$\omega_{tot1} \leq 3$
E/p	$E/p \geq 0.75$

- Rectangular cuts were performed on 6 variables:
  - Trivial cut performed so that $p_T \geq 25$ GeV
  - Other cuts applied to 5 discriminating variables
- Success of the selections is determined using the following quantities:
  - Electron identification efficiency:

$$\epsilon_e = \frac{N_e^{sel}}{N_e}$$

  - Hadron rejection factor:

$$r_{had} = \frac{1}{\epsilon_{had}} = \frac{N_{had}}{N_{had}^{sel}}$$

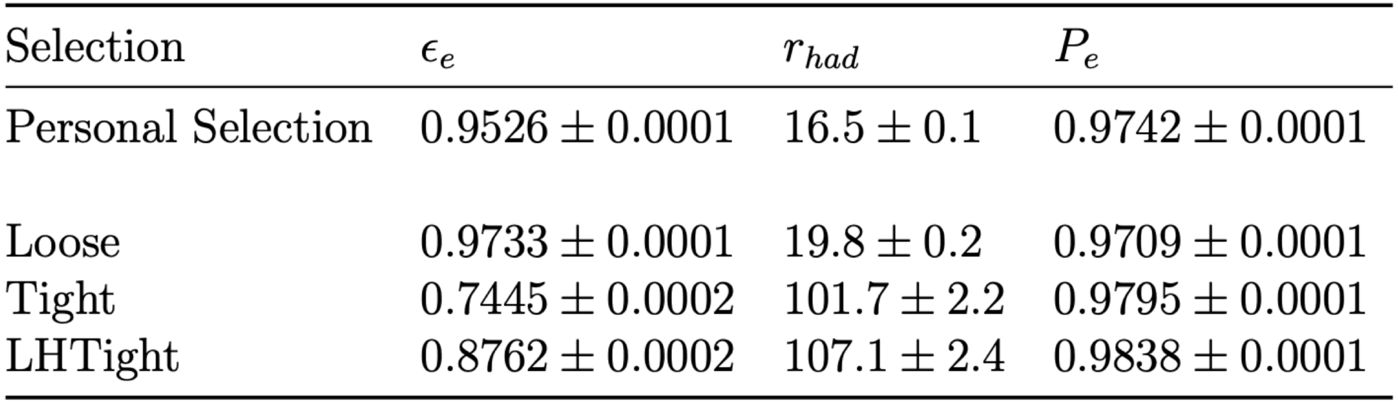
  - Electron purity:

$$P_e = \frac{N_e^{sel}}{N_{sel}}$$

* It is also possible to use the tag-and-probe method on real data (see backup)

Evaluation of selections

- The rectangular cut based selection was applied to the simulated events

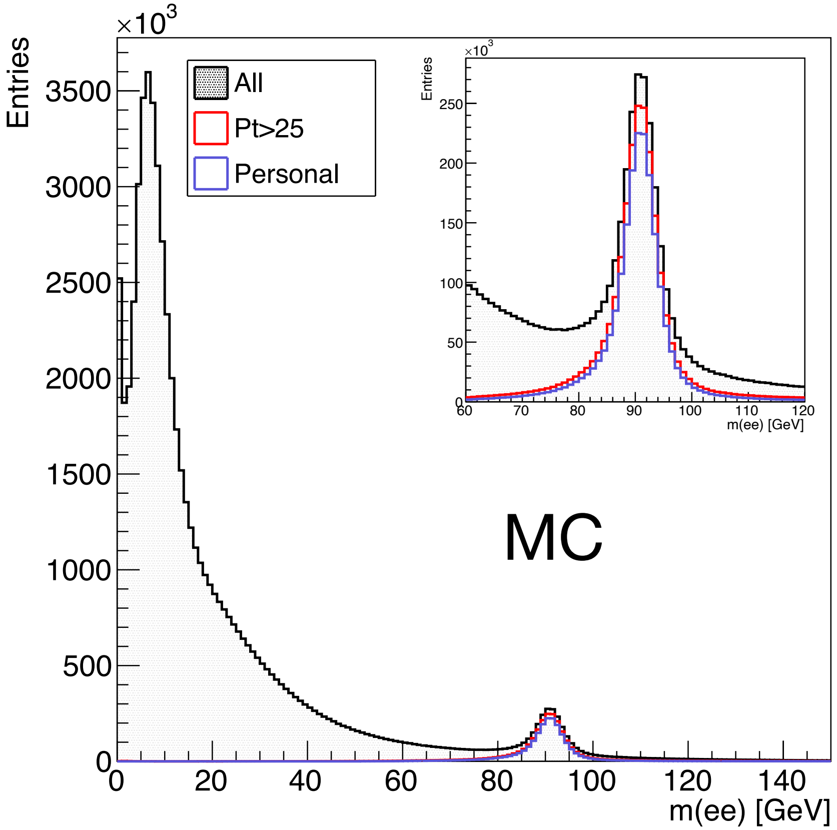


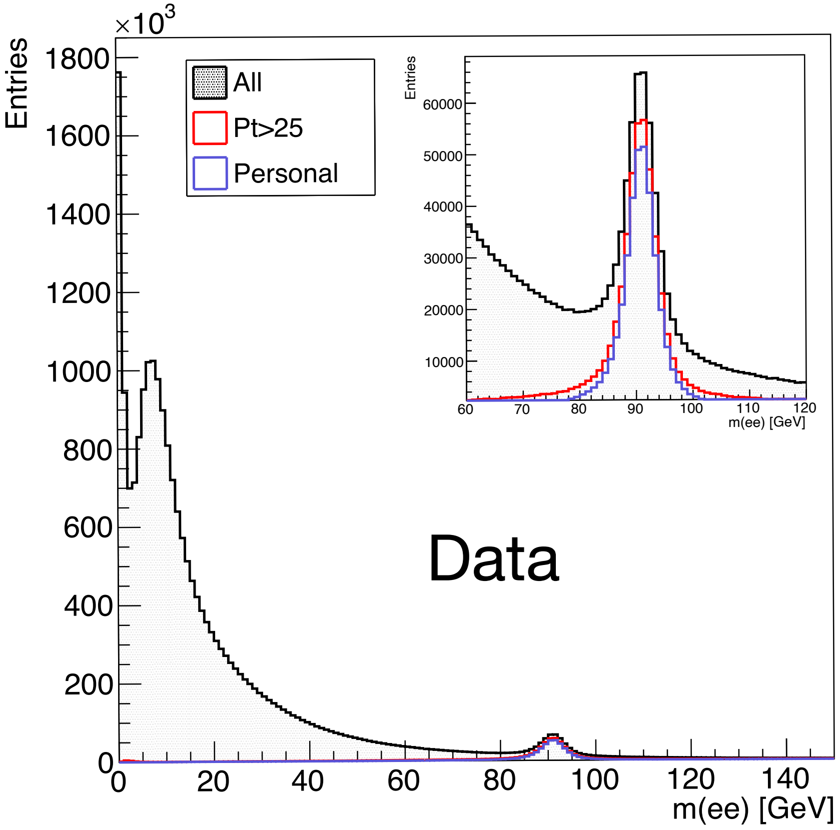
Selection	ϵ_e	r_{had}	P_e
Personal Selection	0.9526 ± 0.0001	16.5 ± 0.1	0.9742 ± 0.0001
Loose	0.9733 ± 0.0001	19.8 ± 0.2	0.9709 ± 0.0001
Tight	0.7445 ± 0.0002	101.7 ± 2.2	0.9795 ± 0.0001
LHTight	0.8762 ± 0.0002	107.1 ± 2.4	0.9838 ± 0.0001

- Also included are reference selections provided by ATLAS:
  - Loose selections target high efficiency at the cost of a lower rejection factor
  - Tight selections aim to reject as many hadrons as possible, but have a lower identification efficiency
  - LHTight is a likelihood based selection that follows the Tight criteria but has a higher efficiency
  - The cuts based selection from this analysis seems to roughly correspond to a Loose ATLAS selection
- All uncertainties are purely statistical

Z^0 reconstruction

Selection of two electrons passing identification, with opposite charge signs, reconstruction of their invariant mass





- Clean sample of Z^0 events with low background contamination can be obtained ‘easily’
- Such a sample can be used to study the reconstruction of the detector, like the electron energy calibration or to measure the electron identification efficiency

Electron identification with MVA methods

… or how to use shape of distributions and correlations between all discriminating variables

- It has become increasingly common to use Machine learning methods in HEP
- This analysis is an example of a classification problem, which is a form of supervised learning
- A typical analysis consists of two main steps: training and application.
- The training phase consists of using samples with known background and signal composition (e.g. a MC simulation with truth values) to train and test classifiers.
  - ➢ The chosen classifier is optimised to maximize Signal-Background separation.
- The application phase involves using the identified classifier to classify unknown samples.

Frameworks

Two different frameworks : TVMA and Scikit-Learn

- TMVA:
  - Uses ROOT code
  - Highly used in the HEP community for many years
  - Was used first in this internship- the methods used are based on those provided in this package
- Scikit-Learn:
  - Python libraries for Machine Learning (numpy) and many others
  - Used in code written as a Python notebook (.ipynb) in a Jupyter notebook
- Jupyter notebook:
  - Web-based interactive computing platform
  - Can be deployed on laptop, server or cloud sites (LPNHE, CC-IN2P3, CERN),
  - The notebook combines live code, equations, narrative text, visualisations
  - Used for MVA analysis but also as an end-user analysis Python alternative to ROOT

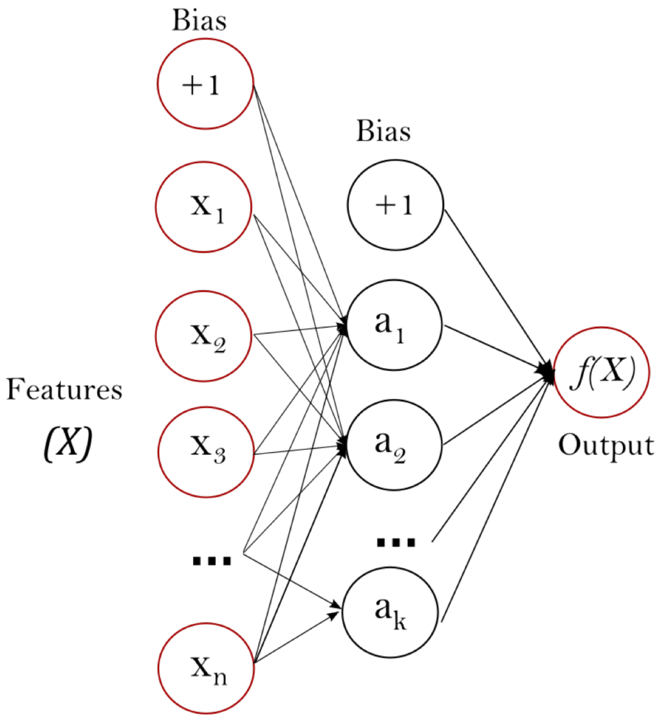
MVA Methods

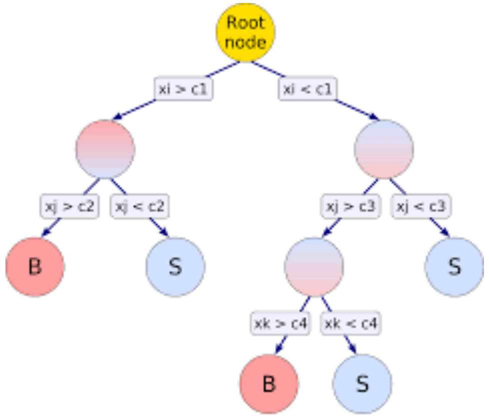
Only a small amount of MVA methods were chosen

$$L_{S(B)} = \prod_{i=1}^{n} P_{S(B),i}(x_i),$$

$$d_L = \frac{L_S}{L_S + L_B}$$

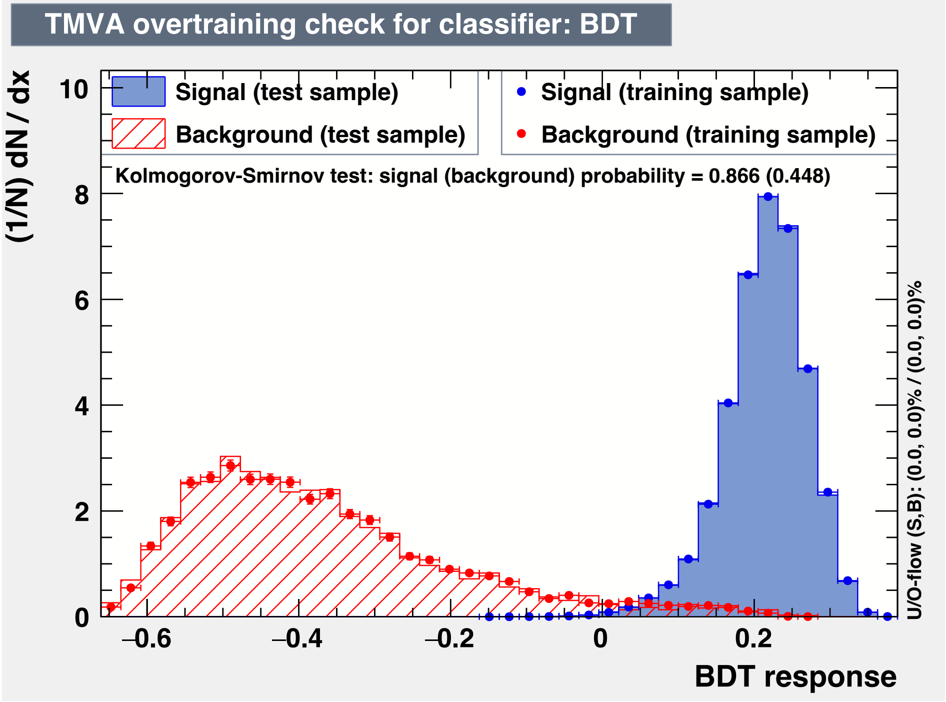
- Likelihood:
  - Uses likelihood ratio between signal and background event PDFs
  - Equivalent to a Naïve Bayes estimator since correlations are ignored
- Fisher’s Linear Discriminant (LDA):
  - Introduces a projection of the variables that aims to maximise the ratio between class separation over in class variance
  - A form of linear discriminant analysis
- Artificial Neural Network (ANN / MLP):
  - Based upon biological neurons in the brain
  - Artificial neurons receive input signals that are weighted to determine its output
  - Trained through backpropagation
- Boosted Decision Tree (BDT):
  - Uses many individual decision trees
  - The trees are combined (boosted) so that an objective function is minimised

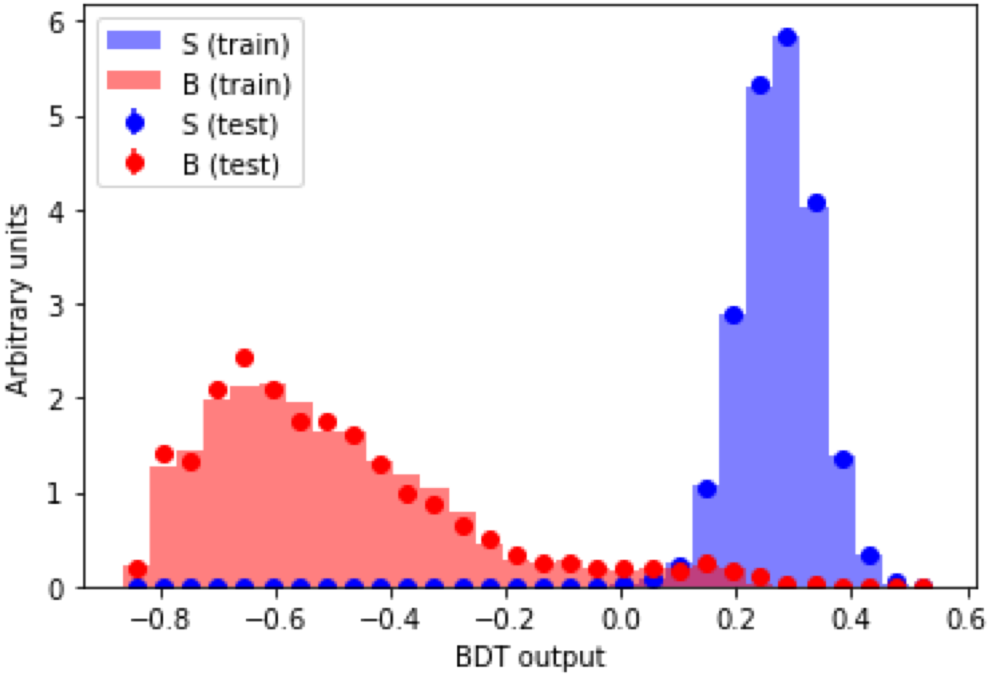




Algorithms implementations can be different in the two frameworks, e.g for BDT AdaBoost for TMVA, XGBoost in the Jupyter notebook

MVA method outputs



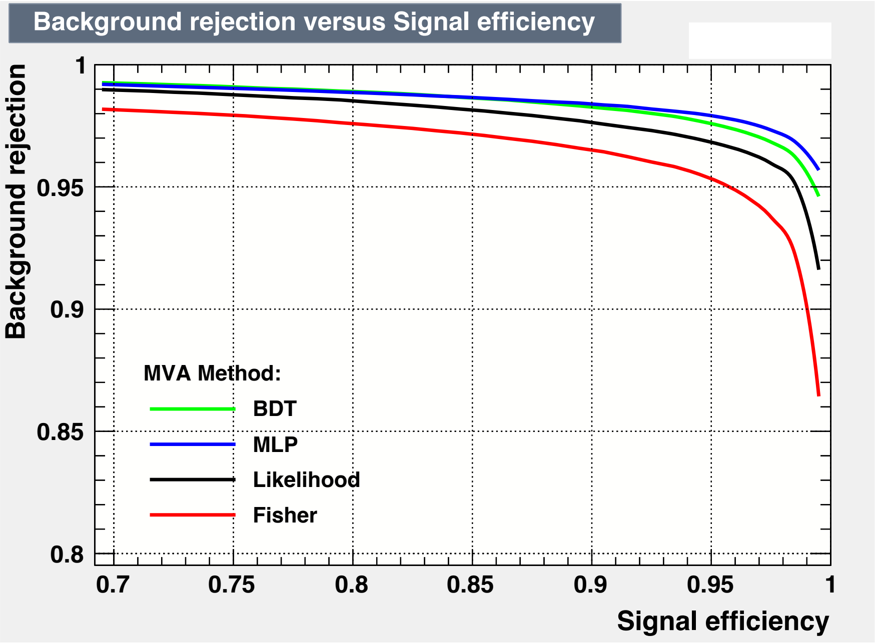


- TMVA produces outputs multiple graphs:
  - Input variables are shown, as well as their correlation functions
  - Outputs (above, left) and other plots like probability are rarity distributions are created for each classifier
- These are not so readily available in Python, but some Scikit-Learn classifiers have a ‘decision function’ attribute that produces a similar output (above, right)

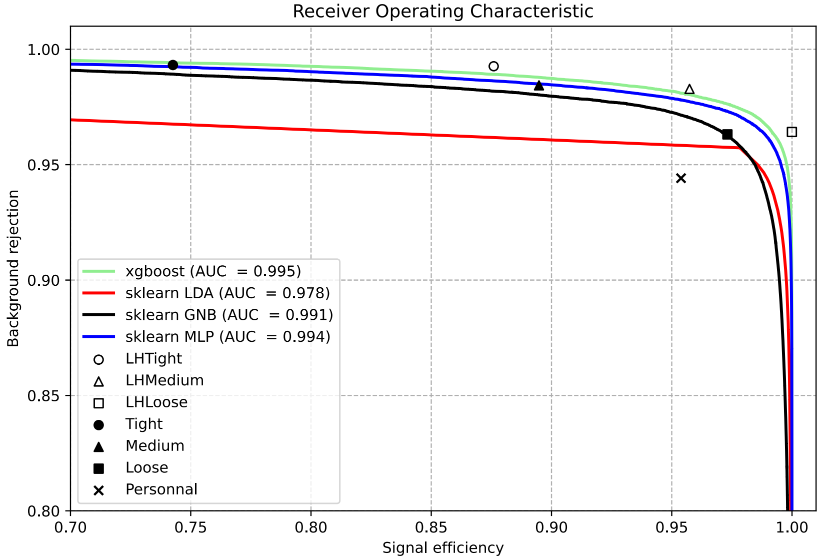
Evaluation of methods

- Methods are evaluated by using Receiver Operating Characteristic (ROC) curve:
  - Plots signal efficiency vs background rejection
  - A better method will have a curve that is closer to (1,1)
  - This can be examined by looking at the area under the curve (AUC): higher is better, max is 1
- ROC curves were plotted and AUC calculated
  - Similar results obtained with both frameworks

MVA Method	AUC	
	TMVA	Scikit-Learn
Likelihood	0.989	0.991
LDA	0.983	0.978
MLP	0.993	0.994
BDT	0.993	0.995



- Scikit-Learn plot shows the previous methods (Personal + ATLAS Loose/Tight) added for reference



- ➢ BDT and Neural Networks perform better
- ➢ XGBoost BDT is much faster than the others!
- ➢ ATLAS official results perform better has trained on better background samples than what used for this internship

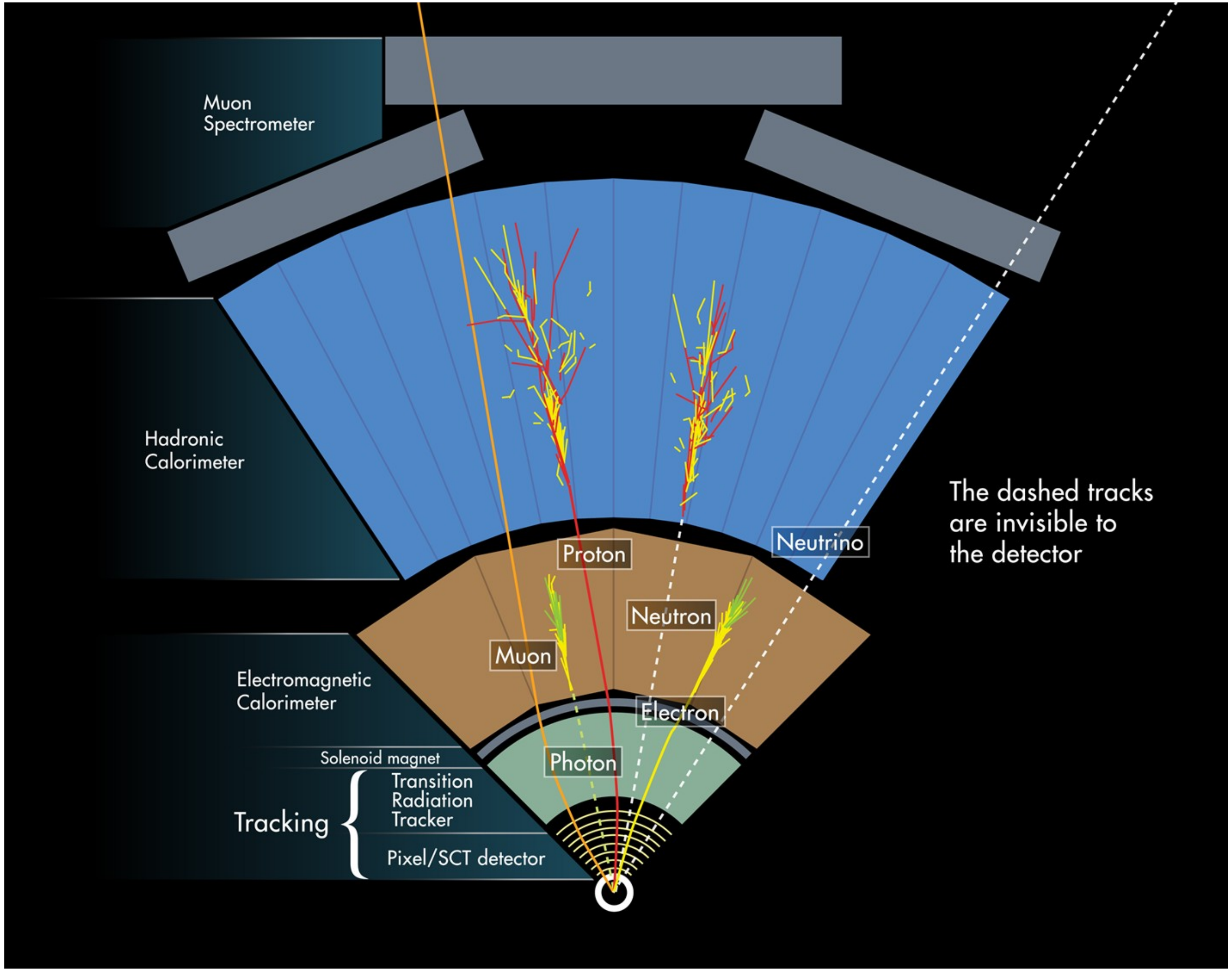
Conclusions

- The work completed on this internship has been generally successful
- Electrons were selected using rectangular cuts in a ROOT program and $Z^0 \rightarrow e^+e^-$ was reconstructed :
  - This selection produced similar results to the Loose reference criterion provided by ATLAS
  - Having never used ROOT before, this was a valuable learning experience
- Two MVA frameworks were evaluated:
  - TMVA is strongly linked to ROOT, has been used for years in HEP and has mature tools that quickly get results
  - Scikit-Learn is more recent, based on Python and can be used in a Jupyter notebook.
- Analysis framework based on Jupyter notebook was developed during this internship
  - Using python – it is more suitable for future internships since it is known by more L3 and M1 students
  - Can incorporate figures, comments, hyperlinks : easier to start with
- MVA methods were used to identify electrons in Z events:
  - The methods were evaluated, with the best method found to be the XGBoost BDT in the Python based environment
  - There is more that can be done with these methods, optimisation in particular, but this needs time and larger background samples
- The remaining time in the internship will be spent trying to apply the MVA methods to top quark events : the standard selection of events in the Jupyter notebook is already done
- This internship has been an excellent opportunity to understand the life of a researcher and practice skills that are relevant in multiple environments

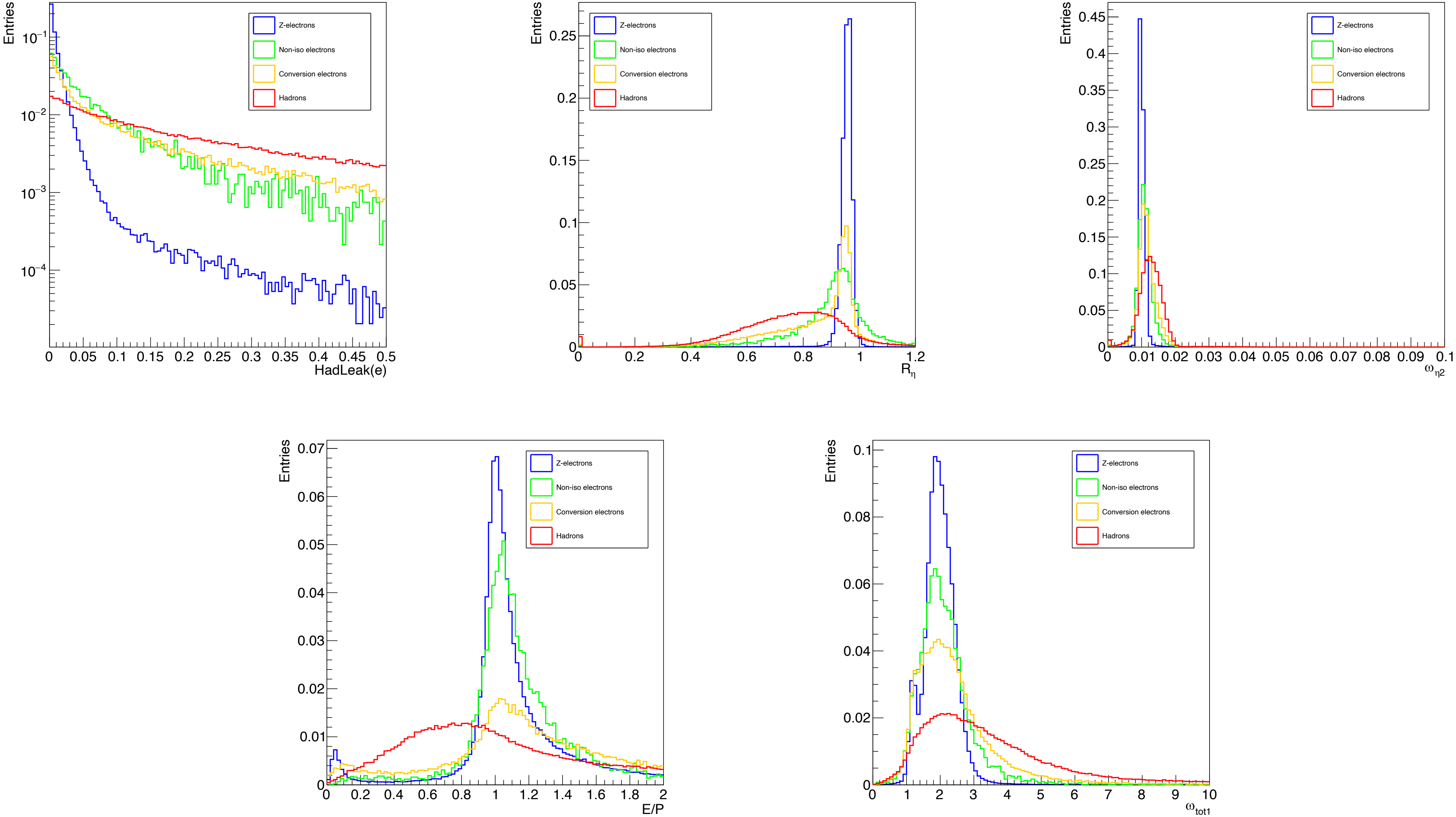
Thank you for listening

Time for questions

Particle identification



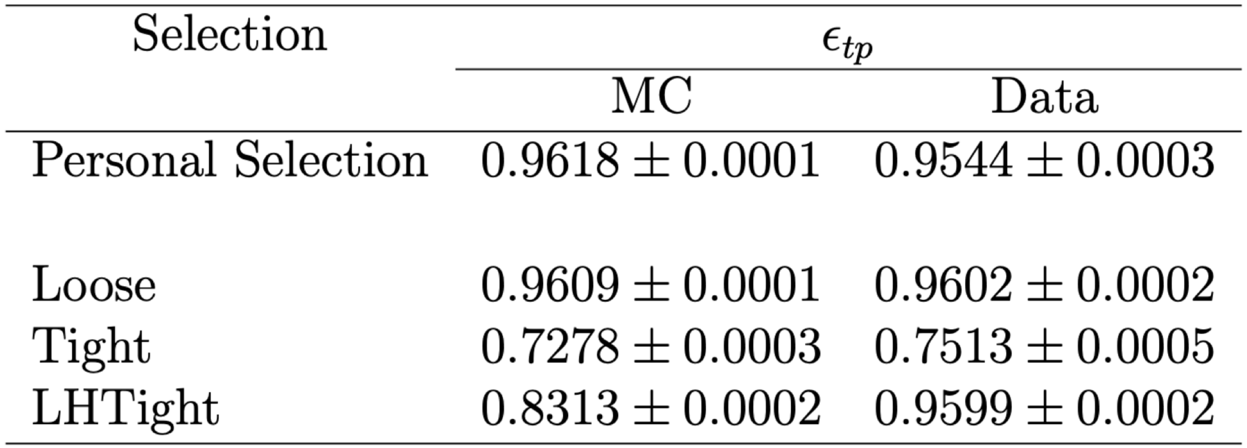
The dashed tracks are invisible to the detector

Entries
Z-electrons
Non-iso electrons
Conversion electrons
Hadrons
10^{-1}
10^{-2}
10^{-3}
10^{-4}
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
HadLeak(e)
Entries
0.25
0.2
0.15
0.1
0.05
0
0.2
0.4
0.6
0.8
1
1.2
R_η
Entries
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.1
$\omega_{\eta 2}$
Entries
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
E/P
Entries
0.1
0.08
0.06
0.04
0.02
0
1
2
3
4
5
6
7
8
9
10
ω_{tot1}

Tag and probe efficiency

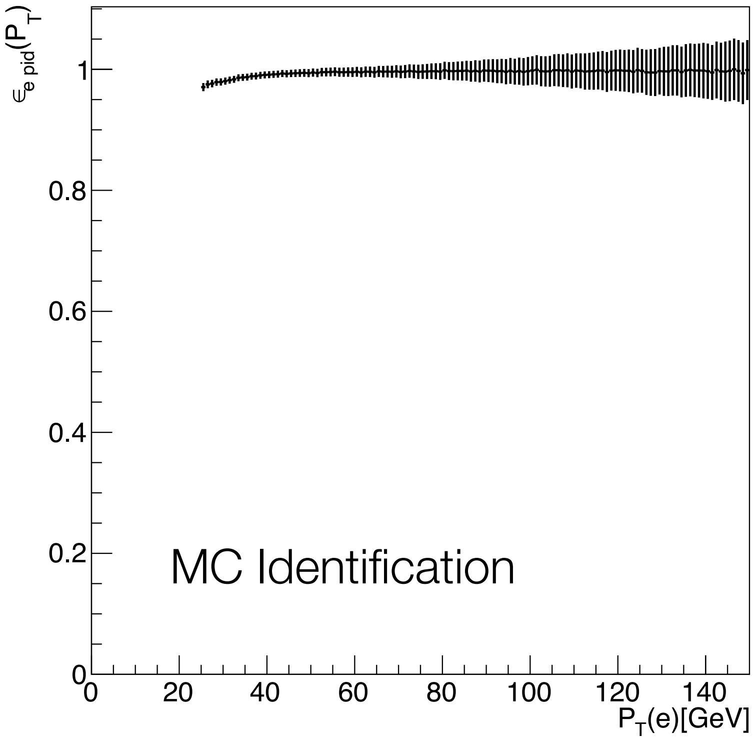
Definition of efficiencies as used with real data

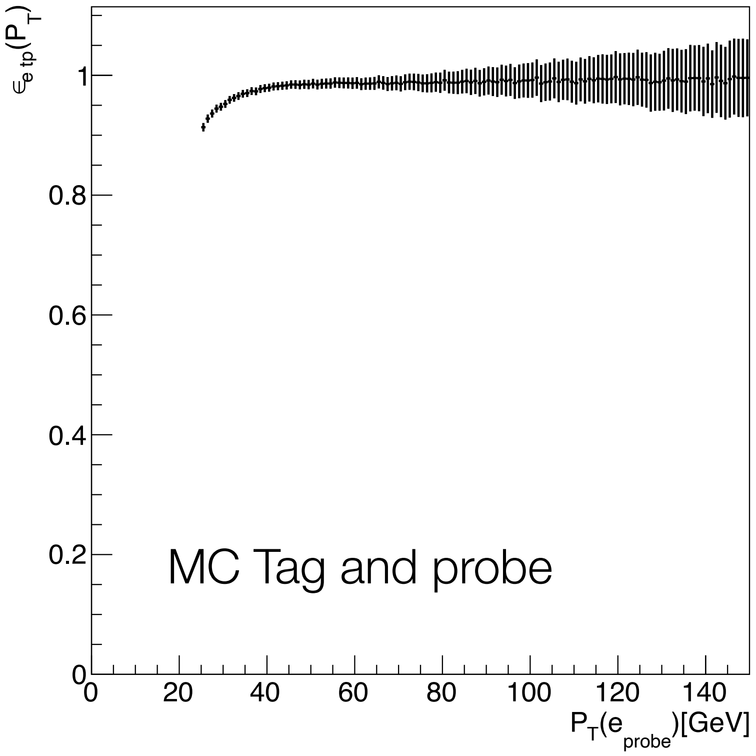
- The other methods for evaluation can only be performed on the simulated data since truth information is required
- An alternative is to use tagged and probed electrons:
  - Electron pairs are evaluated and electrons satisfying the ATLAS Tight criteria are ‘tagged’
  - If the pair has an invariant mass that corresponds to the Z^0 mass, the other electron is ‘probed’
  - The probe electrons are treated as true prompt electrons to calculate the tag and probe efficiency

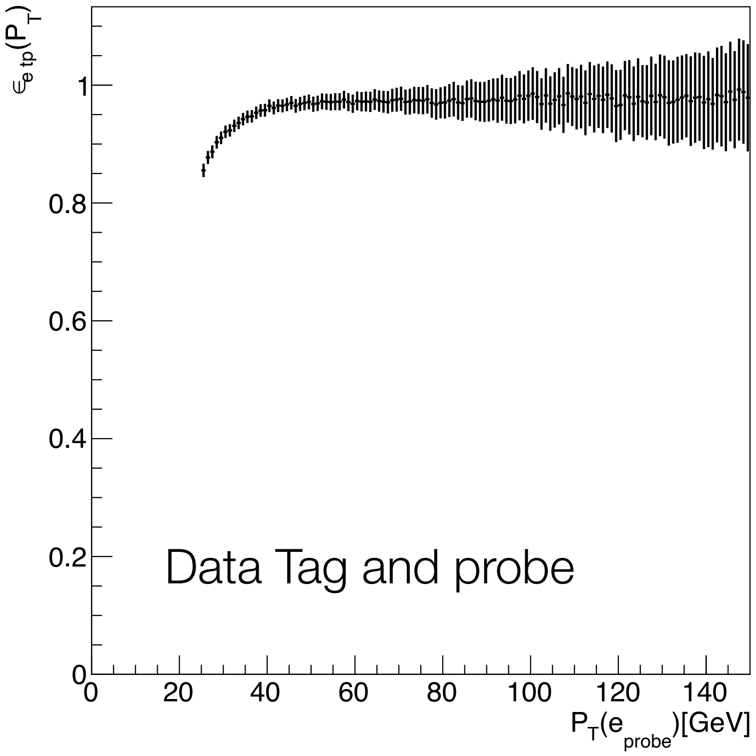


Selection	ϵ_{tp}	
	MC	Data
Personal Selection	0.9618 ± 0.0001	0.9544 ± 0.0003
Loose	0.9609 ± 0.0001	0.9602 ± 0.0002
Tight	0.7278 ± 0.0003	0.7513 ± 0.0005
LHTight	0.8313 ± 0.0002	0.9599 ± 0.0002

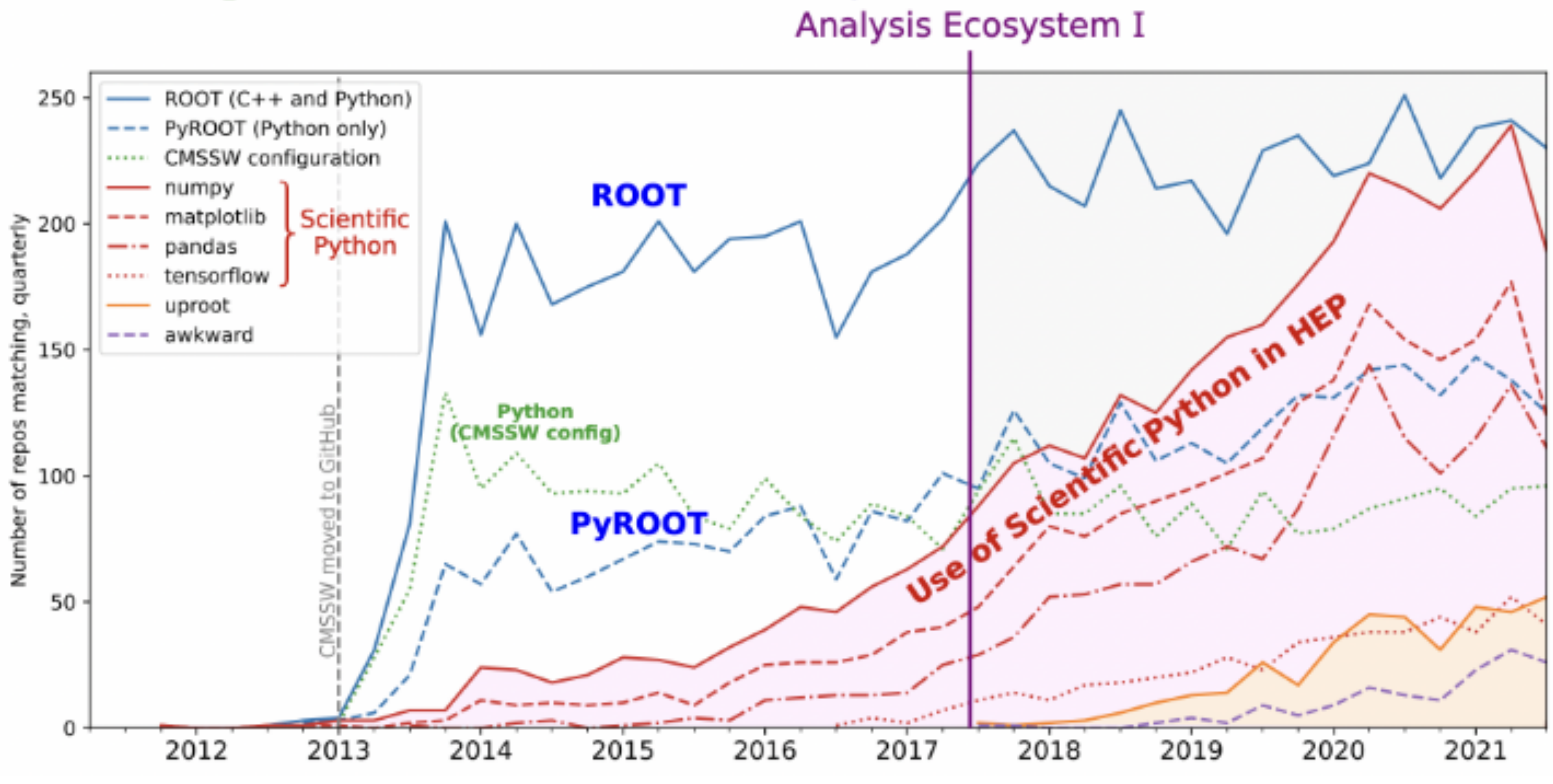
Efficiencies (personal)

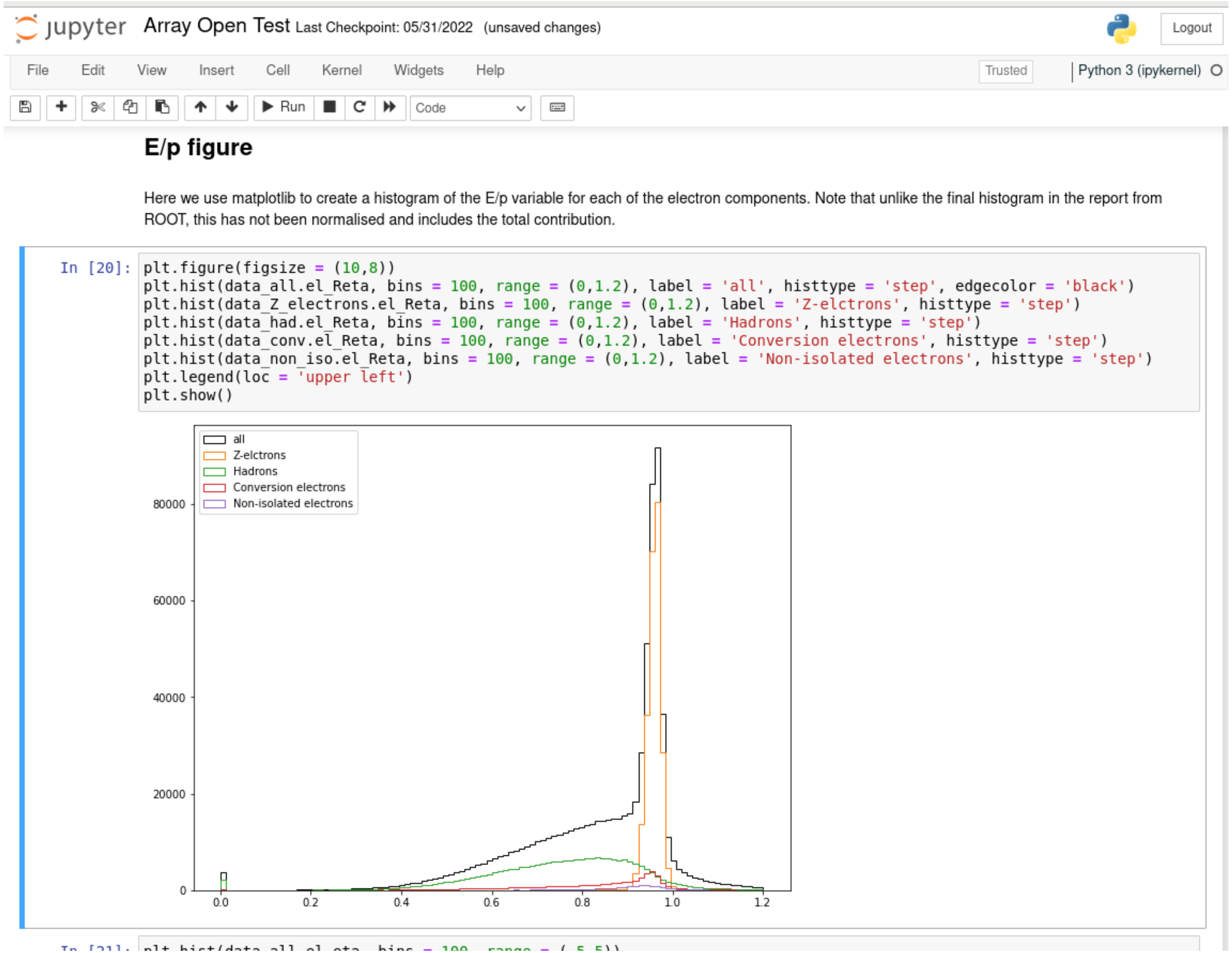






Source: “**import XYZ**” matches in GitHub repos for users who fork CMSSW.





E/p figure

Here we use matplotlib to create a histogram of the E/p variable for each of the electron components. Note that unlike the final histogram in the report from ROOT, this has not been normalised and includes the total contribution.

```
In [20]: plt.figure(figsize = (10,8))
         plt.hist(data_all.el_Reta, bins = 100, range = (0,1.2), label = 'all', histtype = 'step', edgecolor = 'black')
         plt.hist(data_Z_electrons.el_Reta, bins = 100, range = (0,1.2), label = 'Z-elctrons', histtype = 'step')
         plt.hist(data_had.el_Reta, bins = 100, range = (0,1.2), label = 'Hadrons', histtype = 'step')
         plt.hist(data_conv.el_Reta, bins = 100, range = (0,1.2), label = 'Conversion electrons', histtype = 'step')
         plt.hist(data_non_iso.el_Reta, bins = 100, range = (0,1.2), label = 'Non-isolated electrons', histtype = 'step')
         plt.legend(loc = 'upper left')
         plt.show()
```

```
In [21]: plt.hist(data_all.el_eta, bins = 100, range = (-5,5))
```

Studies of top quarks

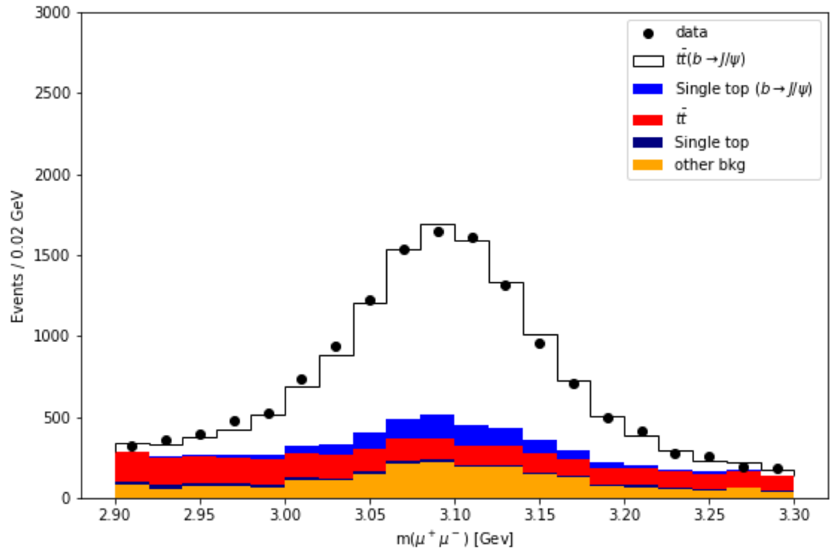
Development of an analysis framework with Jupyter notebook
Application of MVA methods from Scikit-Learn to another area

I took part in the Top LHC France meeting (2 days) in May 2022. Here, French theoreticians and members of ATLAS and CMS and gave talks which included some on MVA applications to top quark studies

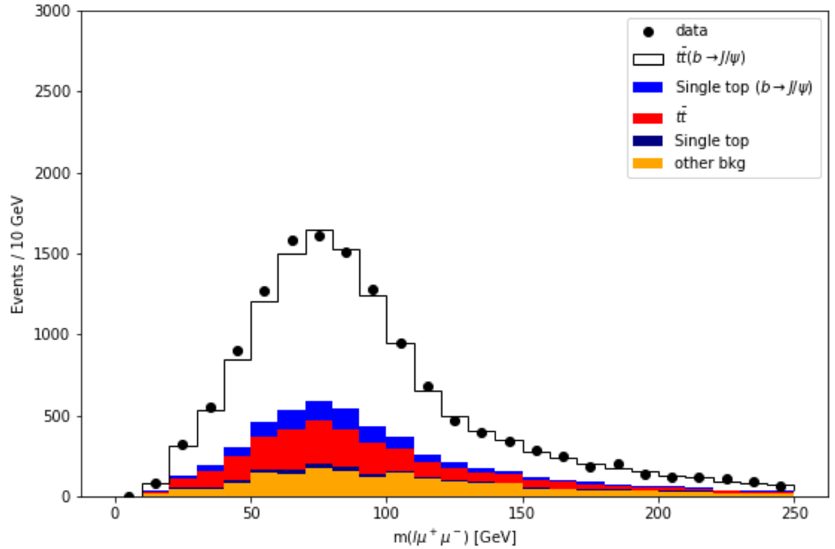
Context

- Top is heaviest particle in SM:
  - Strong Higgs coupling → role in spontaneous EW symmetry breaking ?
  - Standard decay is t→Wb
- The semileptonic decay channel:
  - Top- antitop pair each decay to W and b
  - One W decays to a quark pair, the other to a charged lepton and neutrino
  - Signal in the detector is four jets and one charged lepton, with missing transverse energy from the neutrino
- My supervisor is working on measuring the top mass using semileptonic channel where one of the b jets produces a J/ψ→μμ :
  - The J/ψ is studied by observing its decay into two muons
  - The top is not directly reconstructed, instead a 'proxy' is used : the invariant mass made from the lepton from the W and the two muons from the J/ψ

The final selections and plots can now be done using the Jupyter notebook



Invariant mass of muon pairs [GeV]

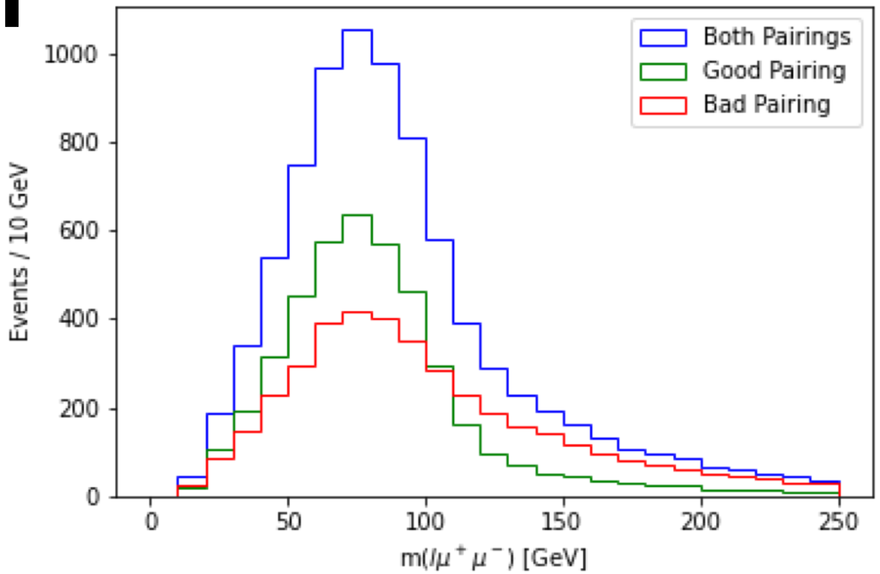


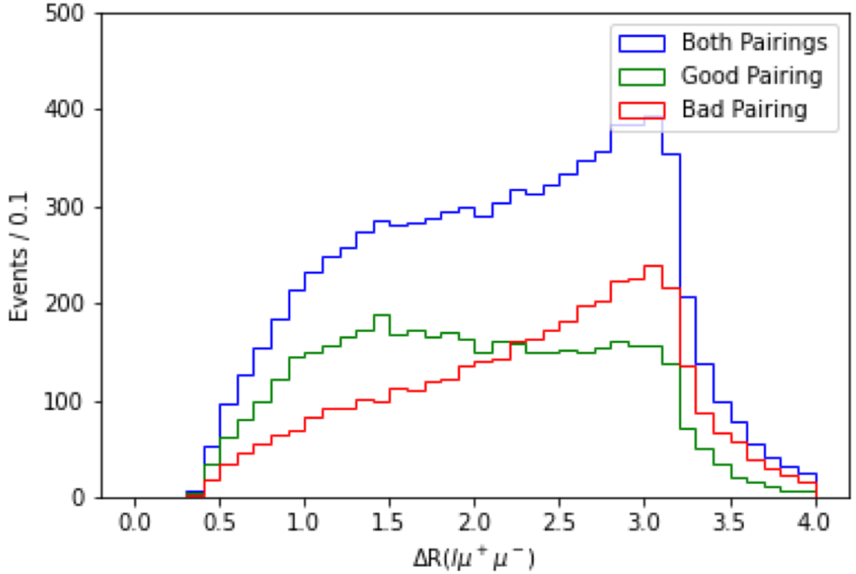
Invariant mass of lepton+muon pairs [GeV]

Optimization of the selection

- The samples used are from my supervisor:
  - Using both MC simulation and real data
  - The simulation has to be normalised to the luminosity of the data so that the contributions can be compared
- Plot of the angular separation has a shape that indicates that the lepton and and J/ψ are emitted back to back:
  - This is due to a 'bad' pairing where the lepton and J/ψ are from different top quarks
- Bad pairings produce a less Gaussian mass peak so introduce more uncertainty on the top mass
- We want to select only events with good pairing:
  - Low number of available events (~10k) mean that cuts on distance between lepton and the two muons is not suitable
  - Need to use more sophisticated methods and more variables

MVA methods to be tested within this new framework … that will be for the very end of the internship


Events / 10 GeV
$m(l\mu^+\mu^-)$ [GeV]
Both Pairings
Good Pairing
Bad Pairing


Events / 0.1
$\Delta R(l\mu^+\mu^-)$
Both Pairings
Good Pairing
Bad Pairing