UNIVERSITÀ DEGLI STUDI DI MILANO
FACOLTÀ DI SCIENZE E TECNOLOGIE

Improving electrons and photons ATLAS data to simulation agreement using machine learning techniques

Relatore: prof. Leonardo Carminati

Correlatore: dott. Ruggero Turra

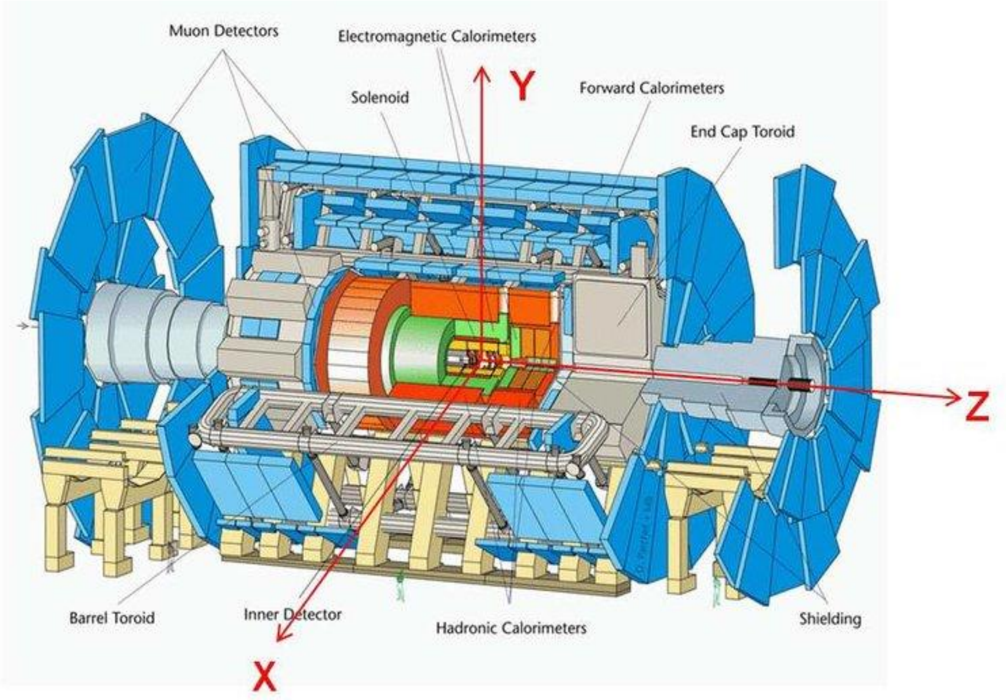
Candidato: Tommaso Cozzi

Matricola: 944205

Discussione di Laurea
Tesi Triennale in Scienze e Tecnologie Fisiche
14 Dicembre 2022

The ATLAS detector

- ATLAS performs precision measurements of Standard Model processes and searches for new physics.



ATLAS layout

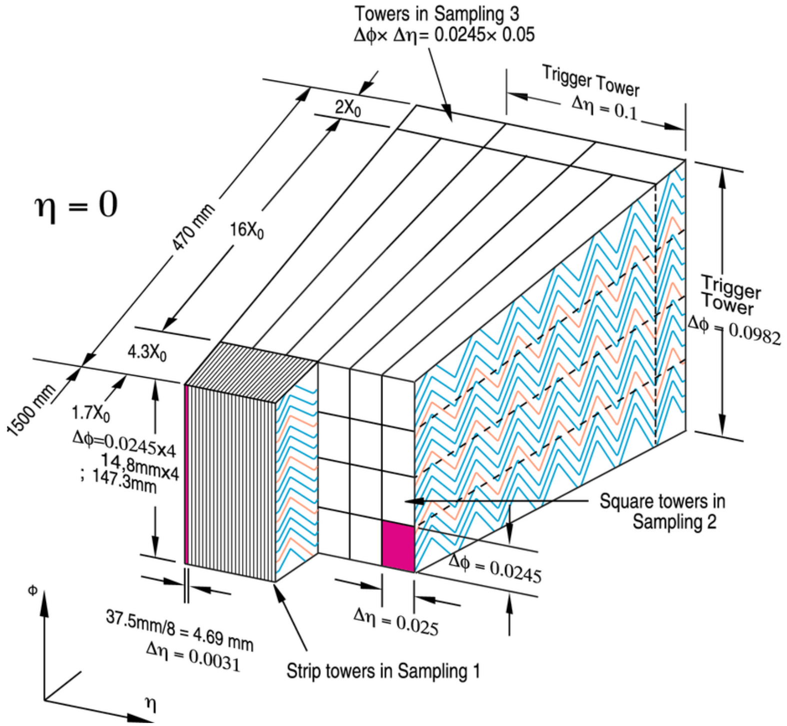
- Cylindrical symmetry, frame of reference: $(z,\ \phi,\ \eta)$ where

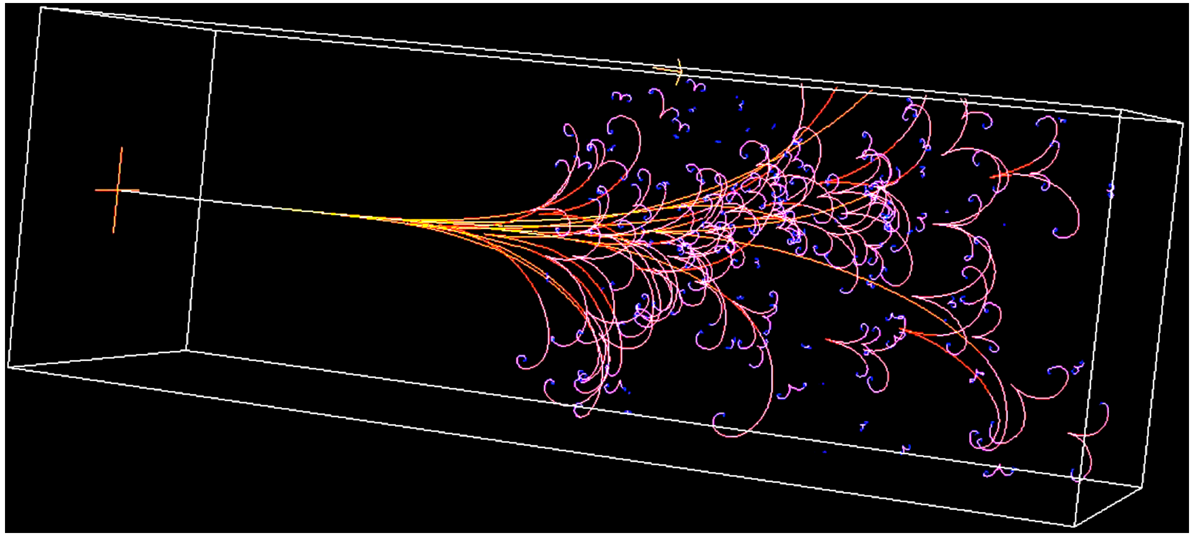
$$\eta = -\ln\left(\tan\left(\frac{\theta}{2}\right)\right)$$

- Inner Detector (ID): measures the direction and momentum of charged particles, detects secondary vertices.
- Calorimeter system: absorbs the incoming particles, the energy released in the detector is transformed in measurable signal.
  - Electromagnetic calorimeter (EMC):
  - Hadronic calorimeter (HC)
- Muon Spectrometer: measures the muon trajectory and momentum.
- Magnet system:
  - Solenoid up to $B = 2\ T$
  - Toroids (2 end-caps, 1 barrel) up to $B = 3.5\ T$

Calorimeters

- The incident particle interacts with the detector material and produces showers of secondary particles with progressively degraded energy.
- A photon appears as a cluster of energies measured in topologically connected cells.



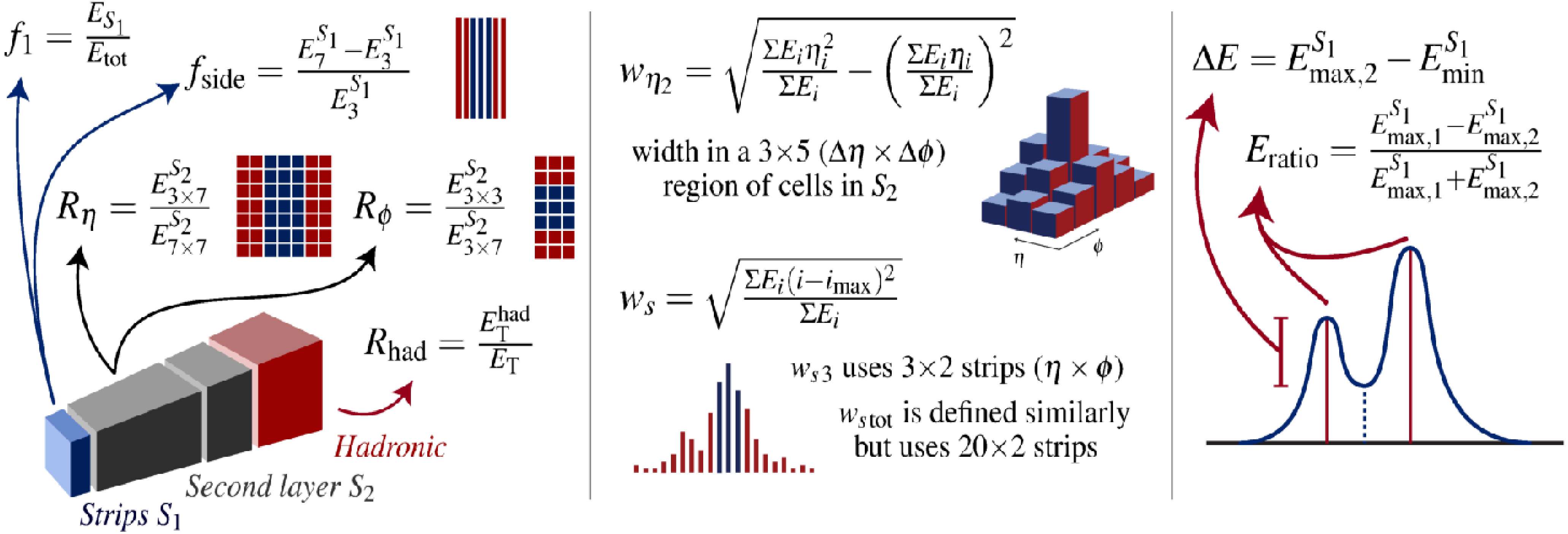


Particle shower in a calorimeter

- The EMC measures mainly electrons and photons, while the HC measures jets of hadrons.
- Physics analyses make heavy use of accurate Monte Carlo simulations of the detector response.
- Usually achieved by exploiting sophisticated Monte Carlo codes (GEANT4).
- A residual difference is usually observed between data and simulations.

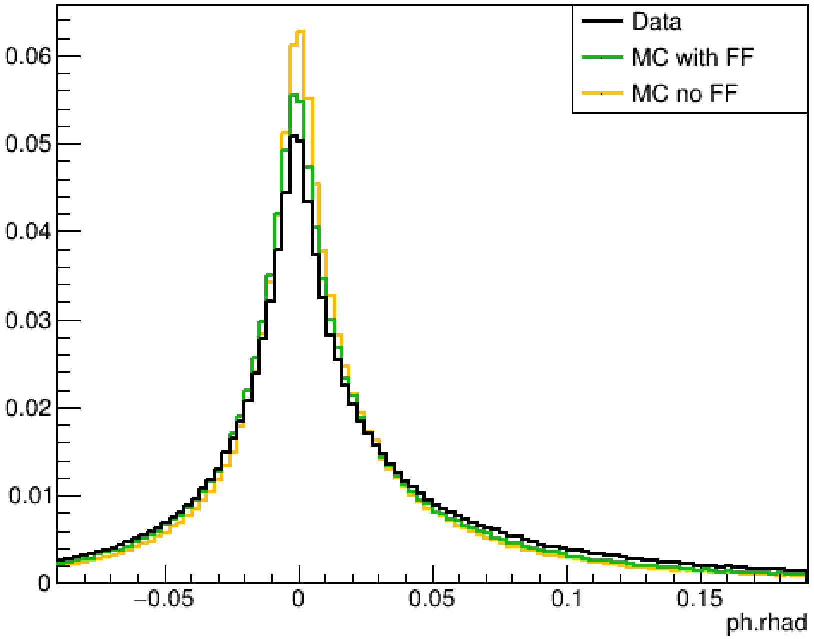
Shower shapes

- The shower shapes are the observables that describe the shower developement in the calorimeter allowing to distinguish between photons and hadrons.
- A photon is defined as a cluster of energy fulfilling specific requirements on 9 shower shapes.



Shower shapes description.

Shower shapes



Data, MC with and without FF distributions for the R_{had} shower shape

- In the case of photons, to account for the differences between data and simulations, correction factors called ”Fudge Factors” (FF) have been introduced.
- The FFs correct each shower shape separately.

- The purpose of this thesis is to calculate a unique weight per event that corrects all the shower shapes.

- The work consists in three steps:
  - Firstly, find the weight training a Boosted Decision Tree to discriminate MC and data events in a sample of photons $Z \to \ell\ell\gamma$ candidates;
  - Evaluate the amount of background in the data sample;
  - Finally, perform a new ‘background aware’ training, adding background events with a negative weight to data.

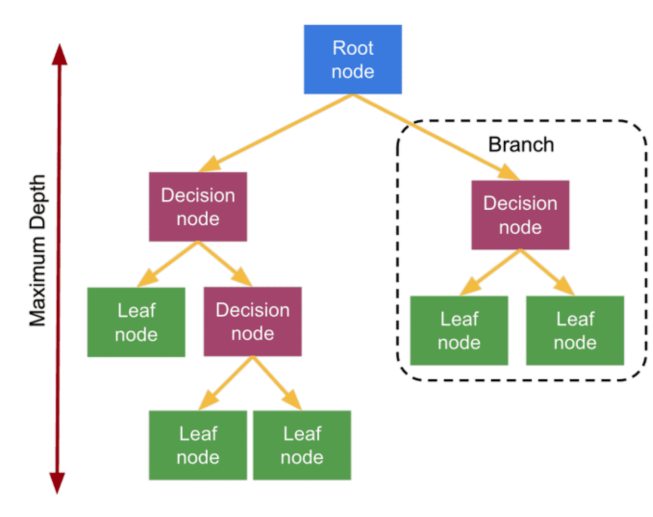
Machine learning techniques

- The training dataset contains the input variables x_i (with its multiple features x_{ij}) and the correct output (y_i^{true}). The input x_i has to be mapped into a target output $y_i^{prediction}$ with a fixed model.

- The predictive performance of the model is evaluated with the Objective Function:

$$Obj(\vec{\vartheta}) = \mathcal{L}(\vec{\vartheta}) + \Omega(\vec{\vartheta})$$

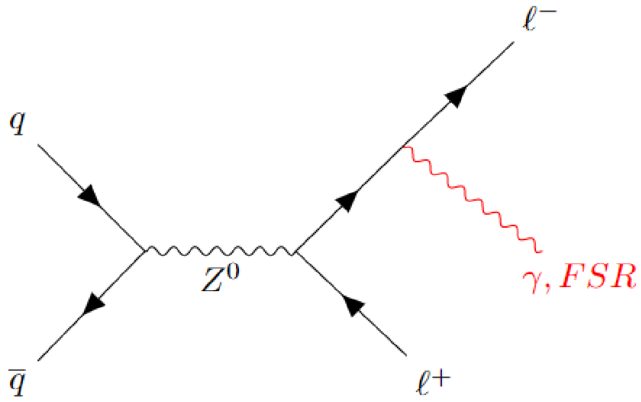
- A Decision Tree is a flowchart structure where the internal nodes represent a condition on a feature of the input variables.



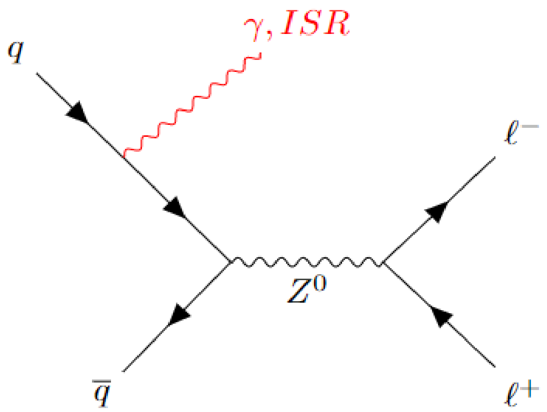
- Gradient Boosting: Trees are added iteratively, choosing, at each step, the Tree that optimizes the Objective Function.

The radiative Z boson decay

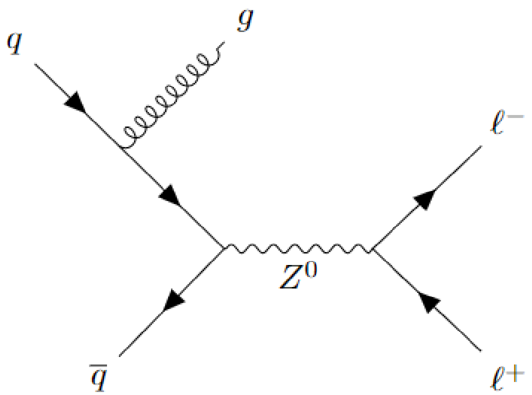
- Need to extract a pure sample of photons in data without using the photon identification criteria.



Final State Radiation. A photon is radiated by a lepton after the Z boson decay.



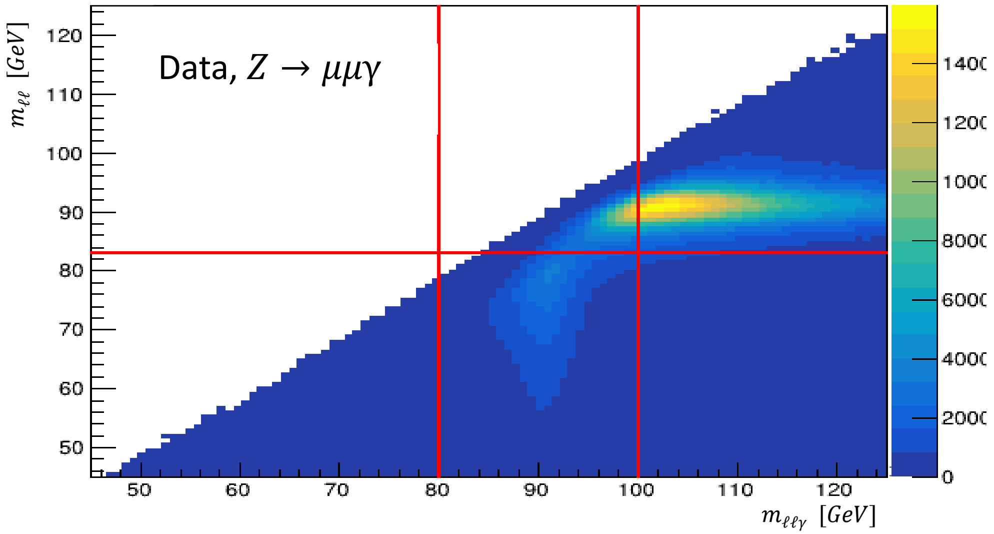
Initial State Radiation. A photon is radiated by the interacting quark quark.

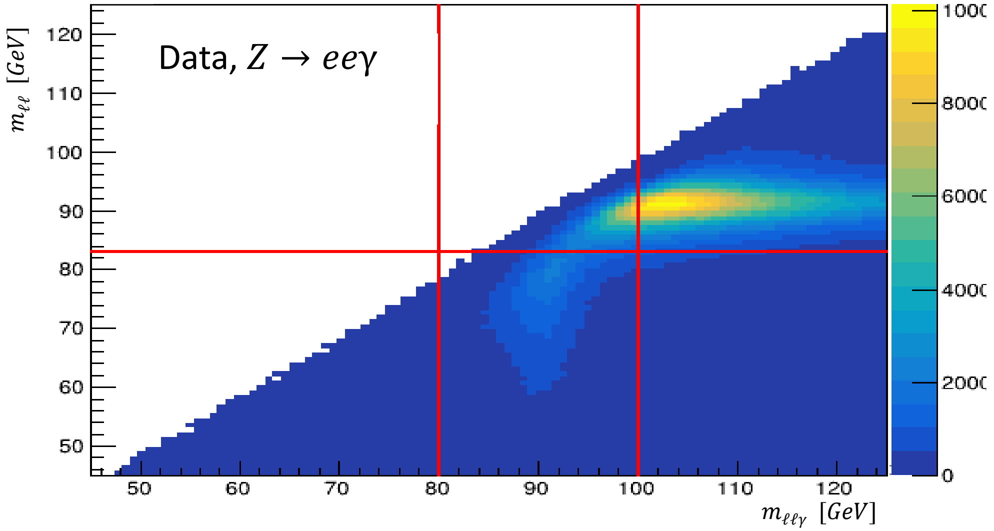


Z+jet production. A gluon is radiated by the interacting quark and it will give rise to an hadronic jet.

- To increase the purity of the photon sample focus on FSR only.
- Only events with $\mu^{\pm}$ or $e^{\pm}$ in the final state are considered.

Event selection





Selections:

- $80 \text{ GeV} < m_{\ell\ell\gamma} < 100 \text{ GeV}$
- $m_{\ell\ell} < 83 \text{ GeV}$
- $p_T^{\gamma} > 10 \text{ GeV}$
- $\Delta R(\gamma, \ell_{1,2}) > 0.4$

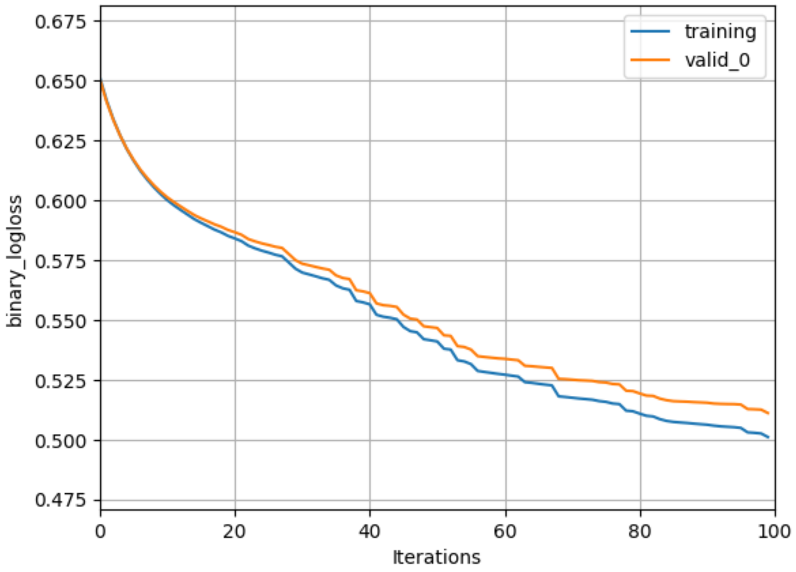
Selection	$Z \to ee\gamma$	$Z \to \mu\mu\gamma$
Total events	2912988	4242106
$80\text{GeV} < m_{\ell\ell\gamma} < 100\text{GeV}$	845164	1264823
$m_{\ell\ell} < 83\text{GeV}$	507224	794922
$p_T^{\gamma} > 10\text{GeV}$	373836	591917
$\Delta R(\gamma, \ell_{1,2}) > 0.4$	297835	474008

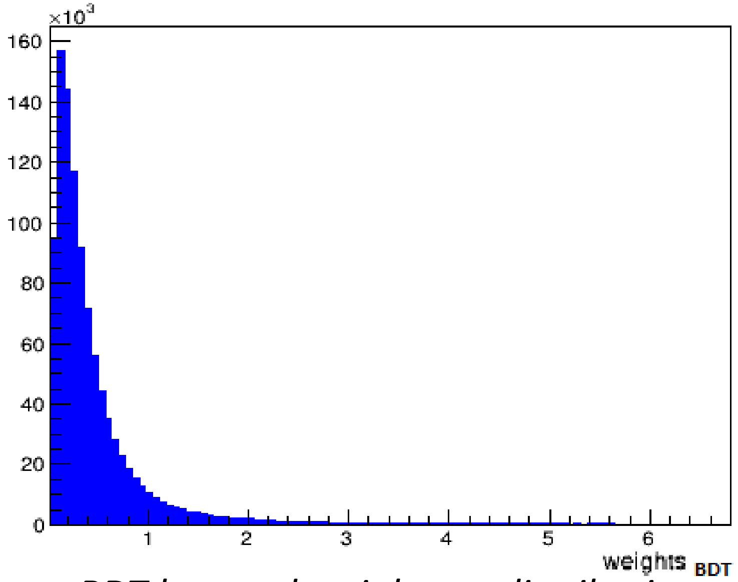
BDT

- BDT is trained to distinguish between photon candidates in MC and data $Z \to \ell\ell\gamma$ samples.
- All the shower shapes have been used as input variables together with η_γ.
- The early stopping has been set to 5 iterations.
- From the algorithm prediction y^{pred} it's possible to obtain the weight to apply to the MC samples.

$$y^{pred} = P[data|x]$$

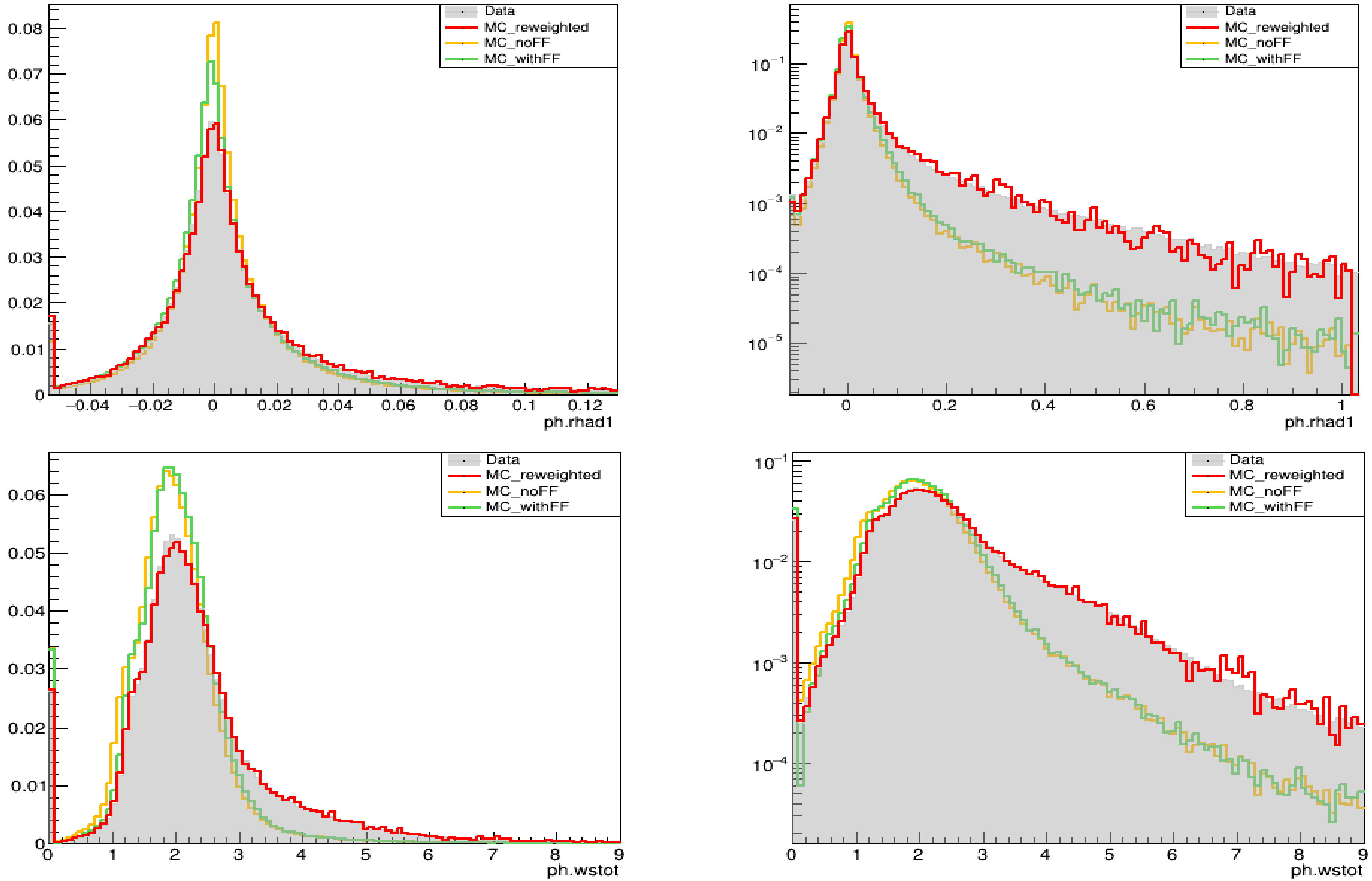
$$weights_{BDT} = \frac{P[data|x]}{P[MC|x]} = \frac{y^{pred}}{1 - y^{pred}}$$





BDT loss and weights$_{BDT}$ distribution.

MC reweighting

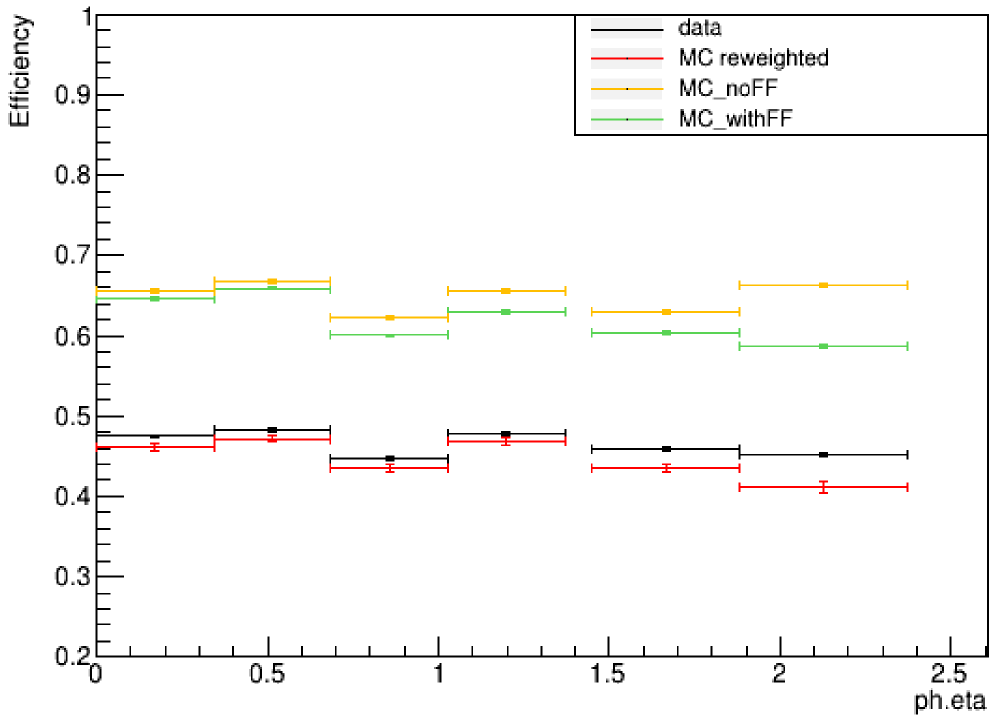


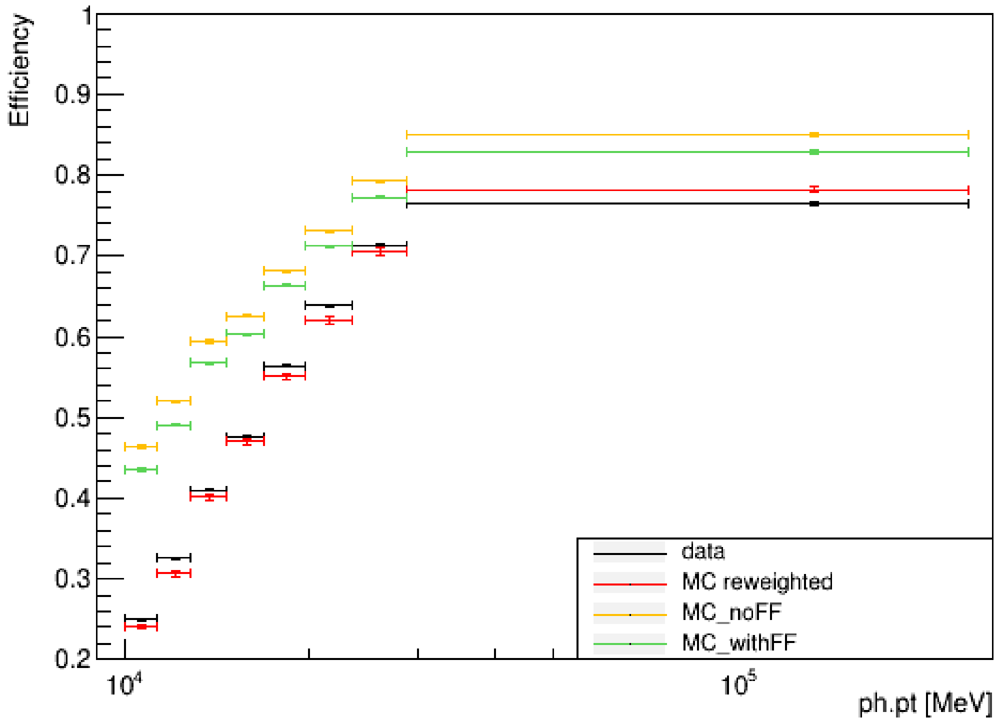
R_{had1} and $w_{s,tot}$ shower shapes are shown. Grey histogram represents the data, yellow and green histograms are the official MC, while the red one is the reweighted MC. Left plots are in linear scale, qhile right ones are in logarithmic scale on y-axis.

In backup the other shower shapes are collected.

Efficiency

- Efficiency: $\epsilon = \frac{N_{pID}}{N_{tot}}$ where N_{pID}=number of photons fulfilling the selection, N_{tot}=total number of photons.





Tight photon identification efficiency as functions of $|\eta_\gamma|$ *and* $p_T{}^\gamma$*. Black line is the data efficiency, yellow and green are the official MC efficiency and the red one is the reweighted MC efficiency.*

- The BDT based weight makes the MC efficiency match the data efficiency.
- Non negligibile background contribution in the data sample (lower efficiency).

Background subtraction

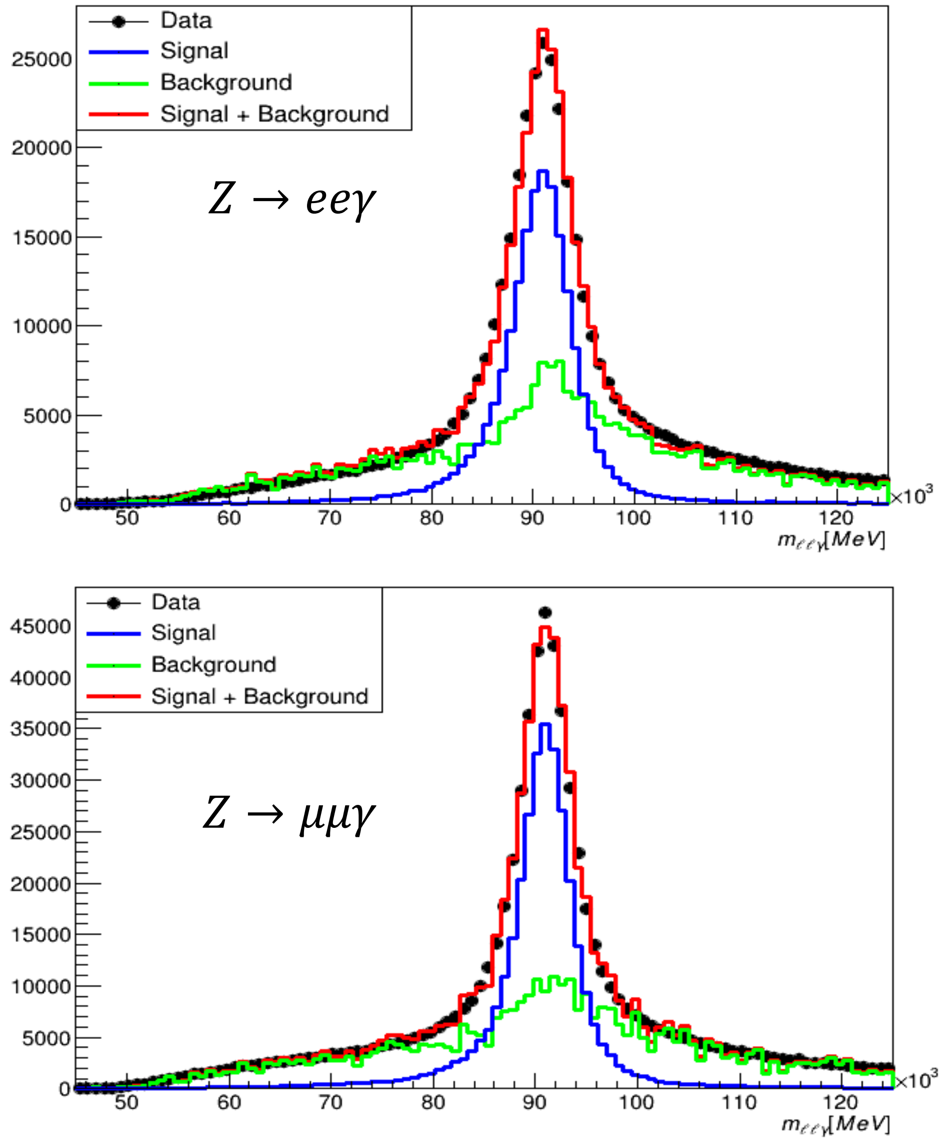
- Background template: MC $Z\ell\ell + j$ sample
- Signal template: MC $Z \to \ell\ell\gamma$ sample
- Data: data $Z \to \ell\ell\gamma$ sample

- Model: $P_{tot} = n_{sig} \cdot P_{sig} + n_{bkg} \cdot P_{bkg}$

Decay mode	$\mathbf{n_{sig}}$	$\mathbf{n_{bkg}}$
$Z \to ee\gamma$	190388.36	234125.54
$Z \to \mu\mu\gamma$	317966.18	365646.82

- n_{bkg} $Z\ell\ell + j$ MC events are added to the data sample with a negative weight:

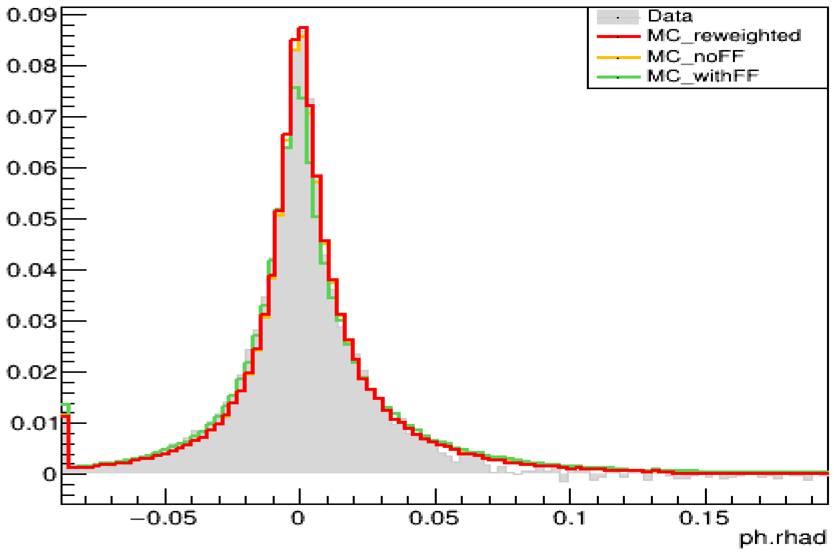
$$W_{bkg} = -\frac{w_j}{\sum_j w_j} \cdot n_{bkg}$$

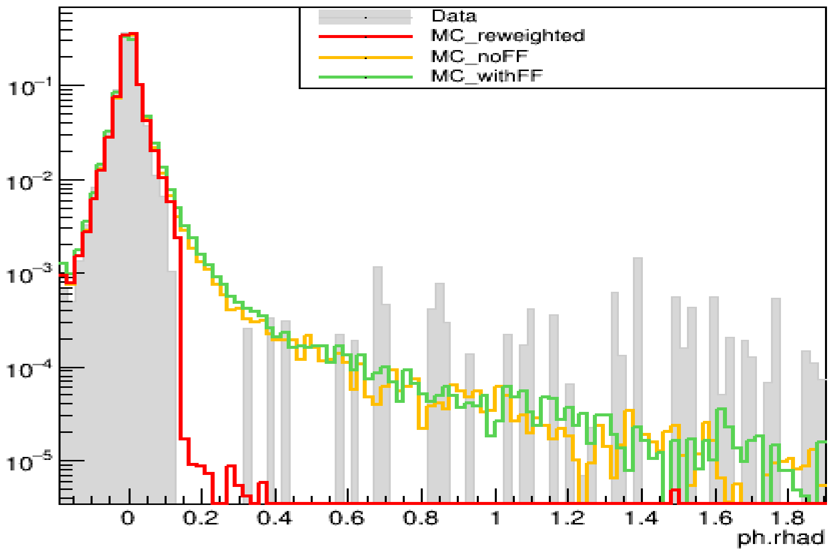


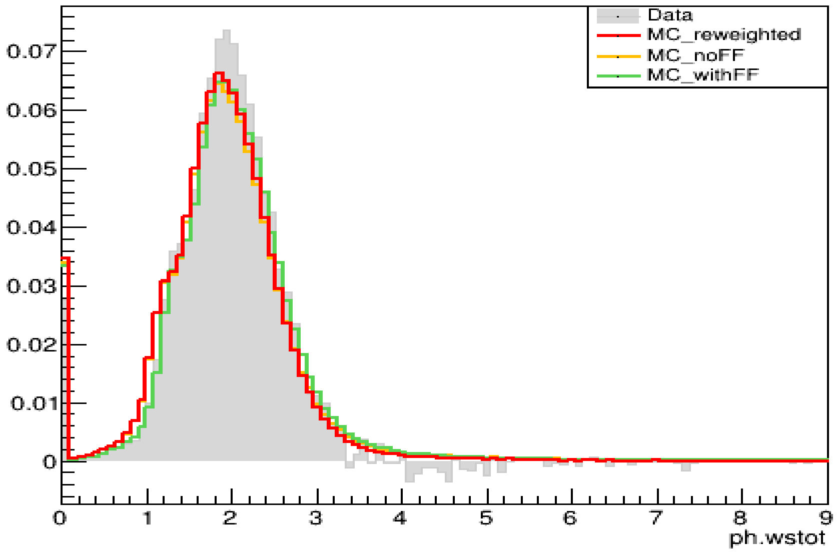
Fit to data of the model (red line).

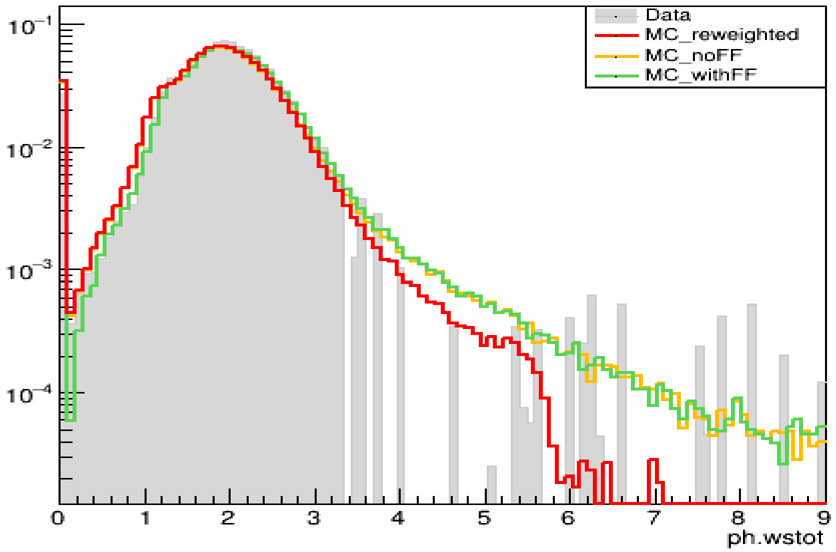
Background aware MC reweighting

R_{had} and $w_{s,tot}$ shower shapes are shown. Grey histogram represents the data, yellow and green histograms are the official MC, while the red one is the reweighted MC . Left plots are in linear scale, qhile right ones are in logarithmic scale on y-axis.



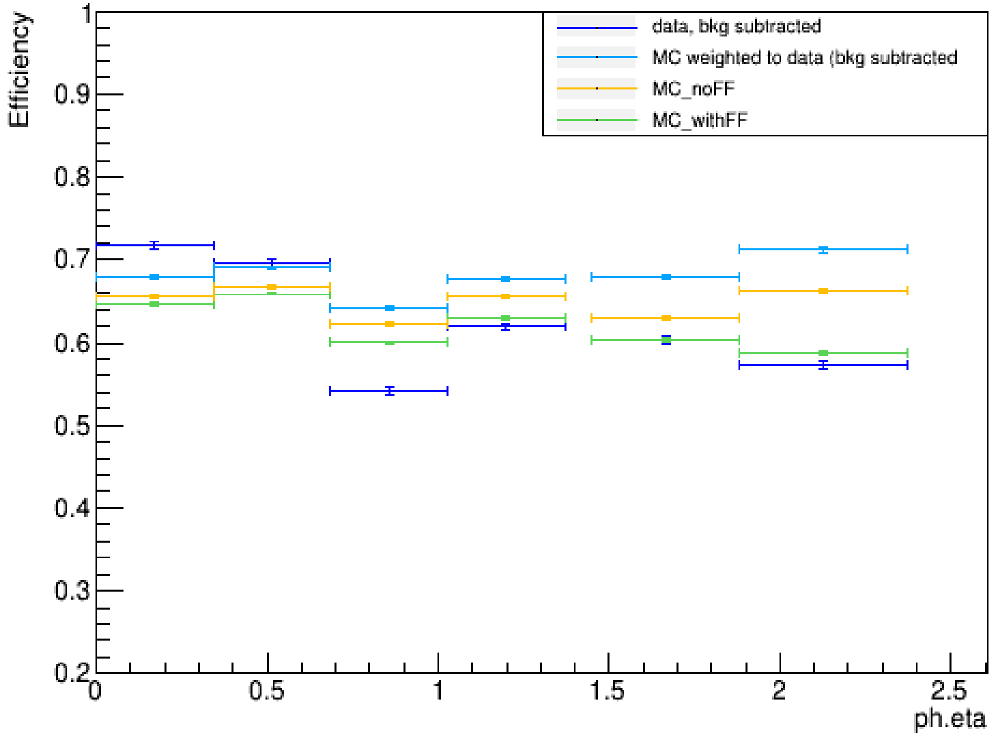


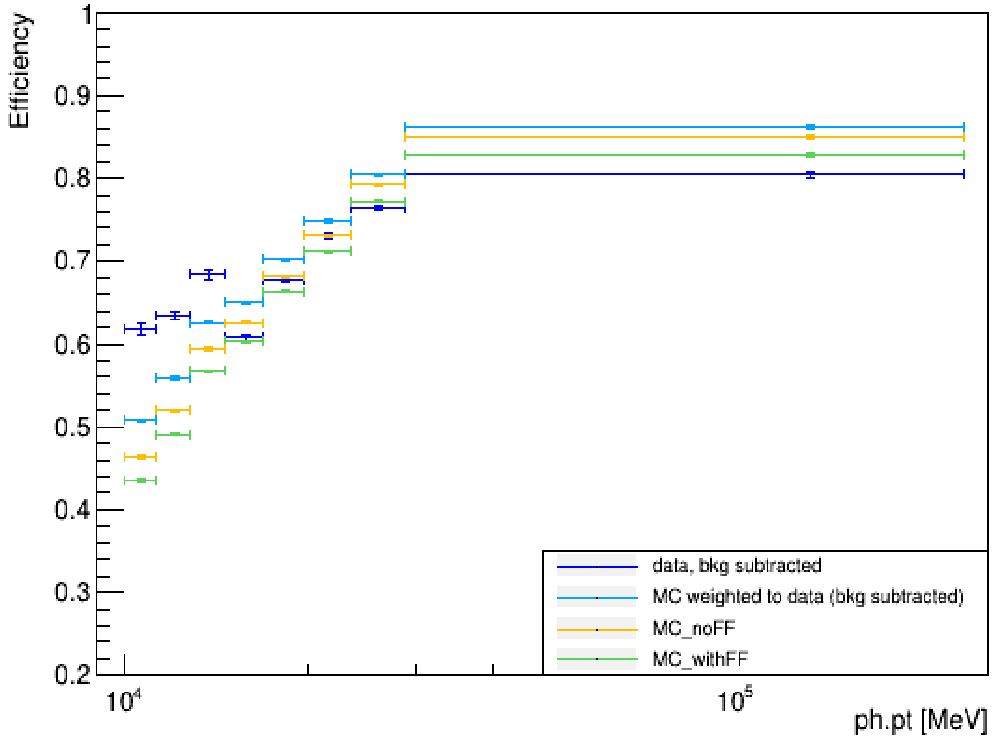




In backup the other shower shapes are collected.

Background aware MC reweighting: efficiency





Tight photon identification efficiency as functions of $|\eta_\gamma|$ *and* $p_T{}^\gamma$. *Blue line is the data efficiency, yellow and green are the official MC efficiency and the light blue one is the reweighted MC efficiency.*

- The efficiency in data after the background subtraction is higher (as expected).
- Data and reweighted MC agreement is not fully stisfying yet.
- The reweighted MC efficiency describes the data efficiency better than the official MC only at low $p_T{}^\gamma$ and $|\eta_\gamma|$.

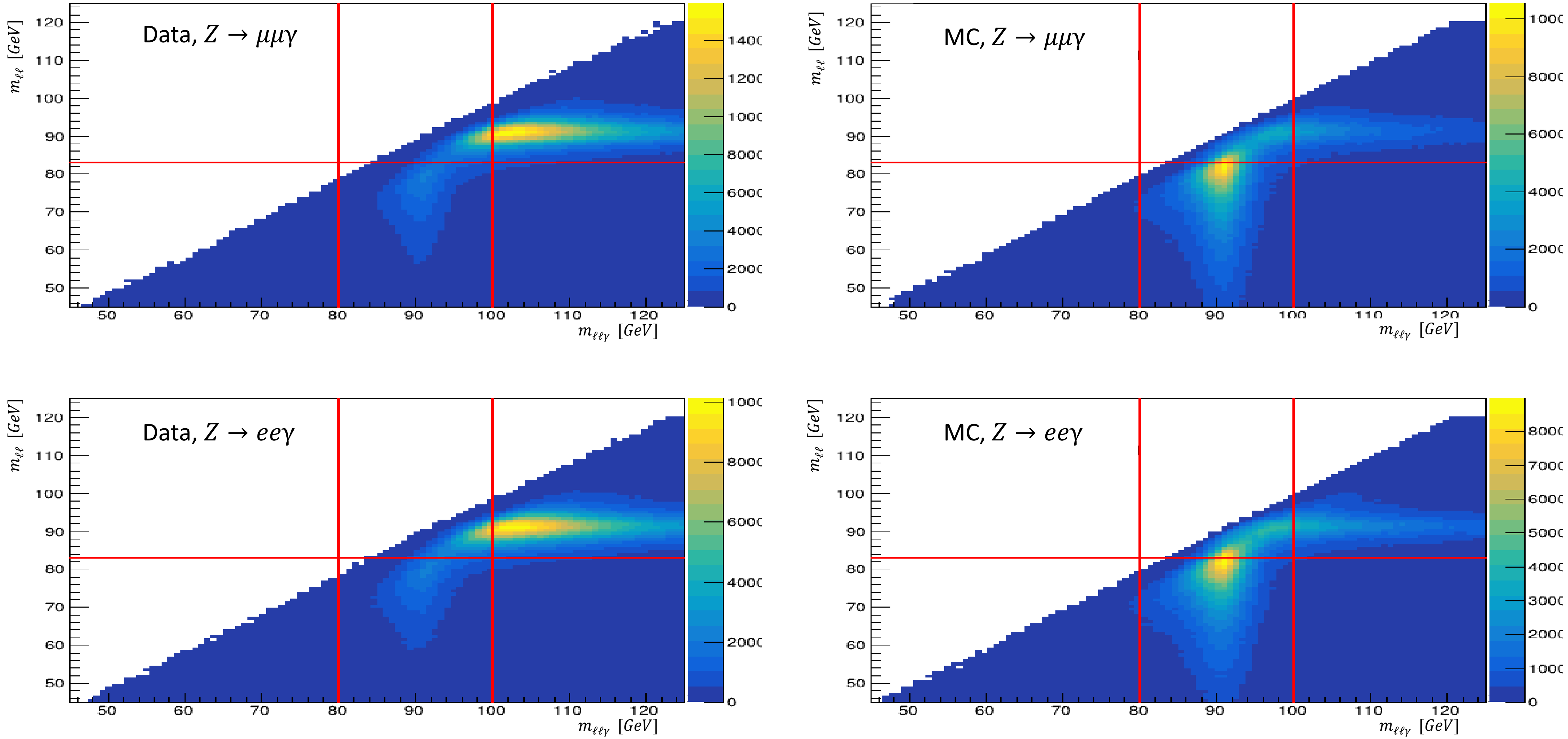
Conclusions

- In this thesis a new approach to simulation correction has been introduced using a BDT algorithm trained to distinguish data from MC events in a sample of photons from $Z \to \ell\ell\gamma$ decay.
  - The BDT based weight improves the agreement between data and simulation in the shower shape variables and makes their efficiencies closer.
  - The data sample is contaminated by $Z\ell\ell + j$ background and these weights can not be applied directly to $Z \to \ell\ell\gamma$ MC samples.
- The amount of background has been evaluated from a ‘signal + background’ fit to data.
  - Estimated purity: $\sim 46.5\%$ in the muonic channel, and $\sim 44.8\%$ in the electronic channel.
- A new BDT has been trained adding MC background events with negative weights to the data sample.
  - The new weight doesn’t improve the agreement between data and simulations.
- The method is promising although it has to be investigated further to improve its performance.

Thank you for your attention

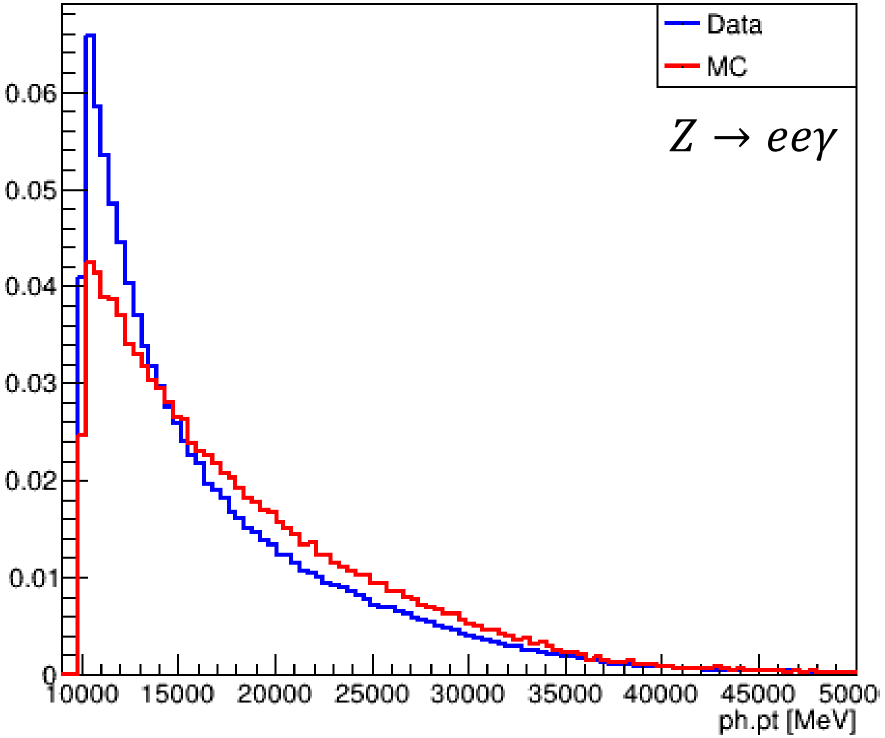
Backup slides

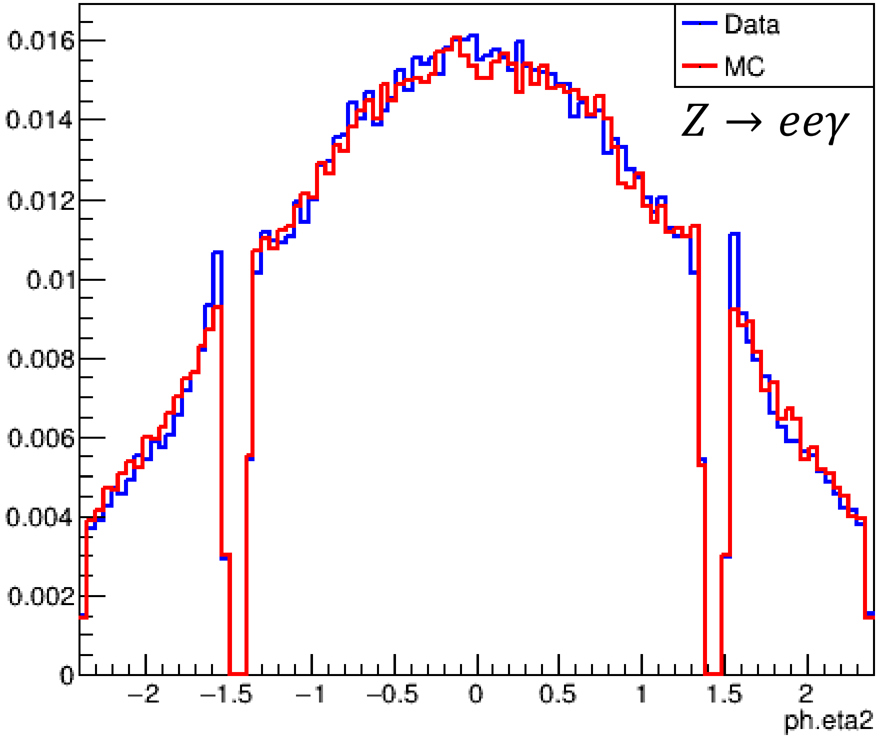
Event selection



Invariant masses of the two lepton (y-axis) and of two leptons and a photon (x-axis). The yellow spots represent the ISR and the FSR.

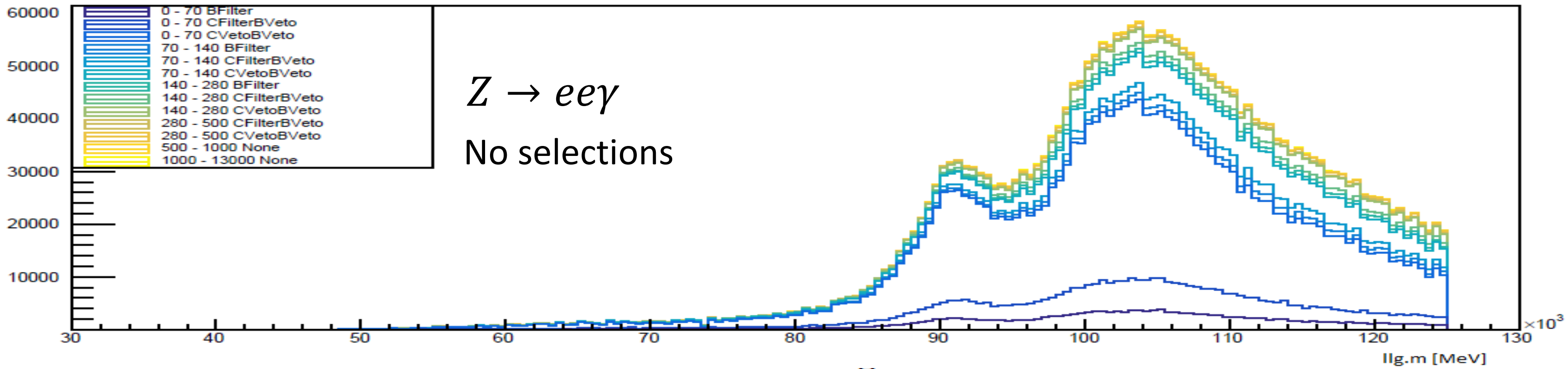
Event selection

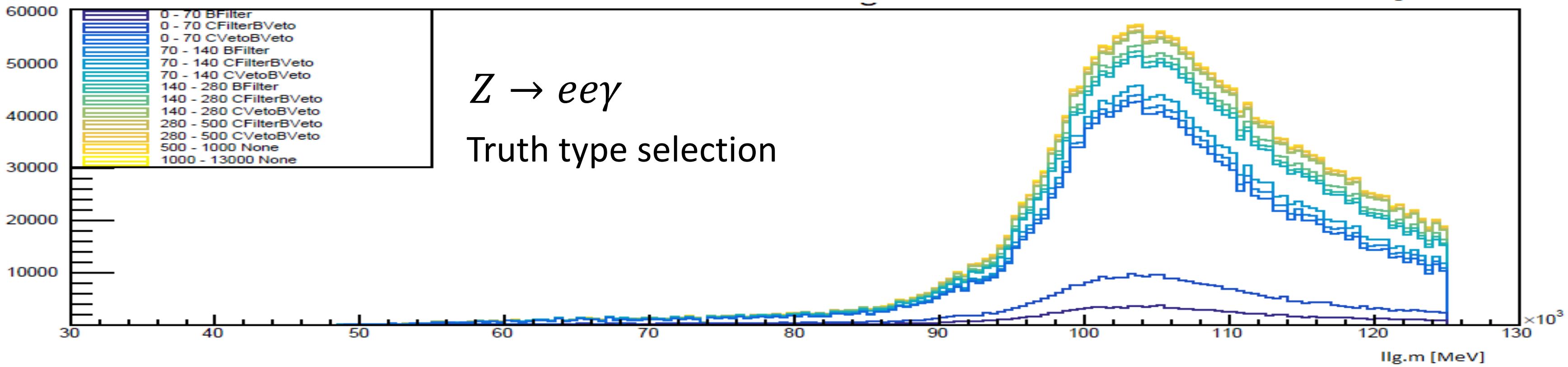


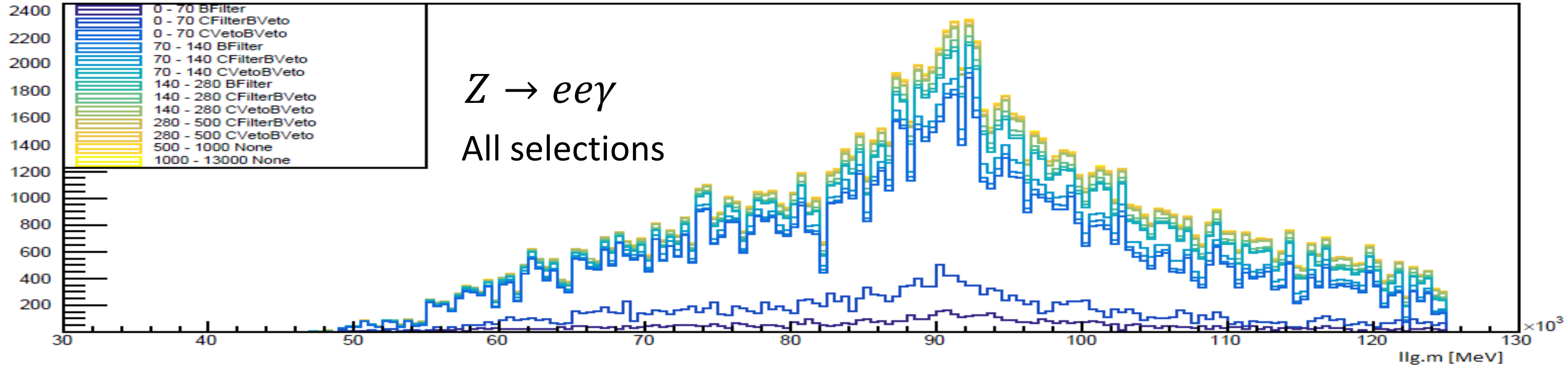


Photon transverse momentum and photon pseudorapidity distributions in data and MC $Z \rightarrow ee\gamma$ samples after the selections.

Event selection

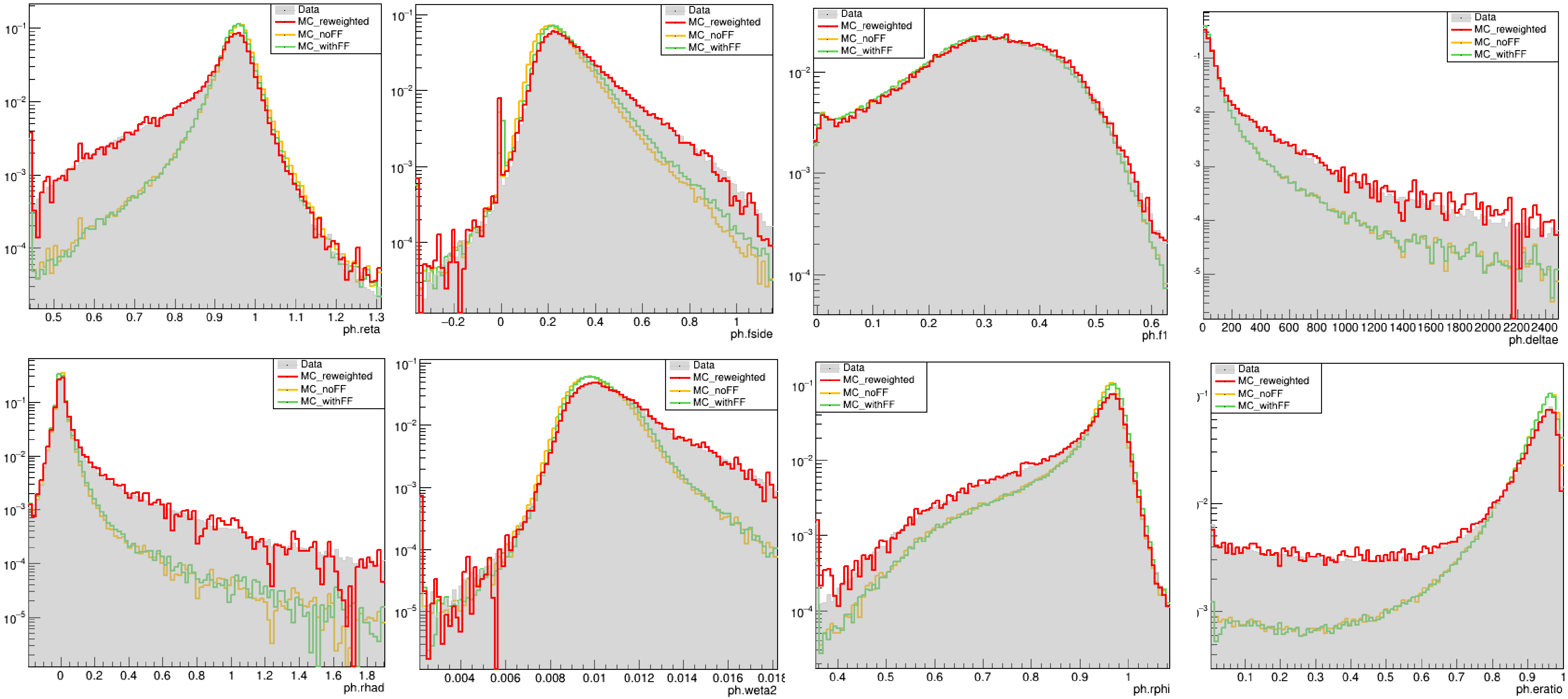




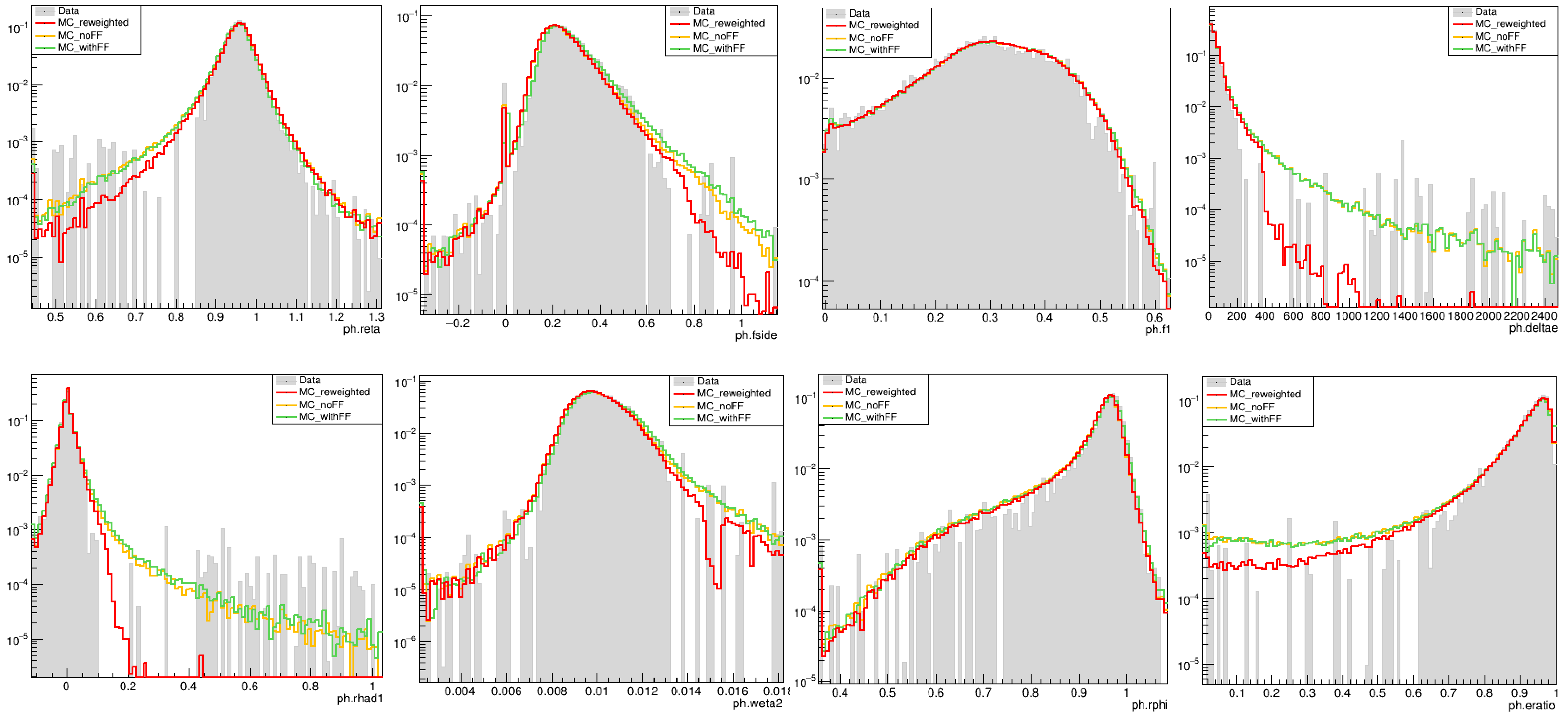


Invariant mass $m_{\ell\ell\gamma}$ in the $Z\ell\ell + j$ sample.

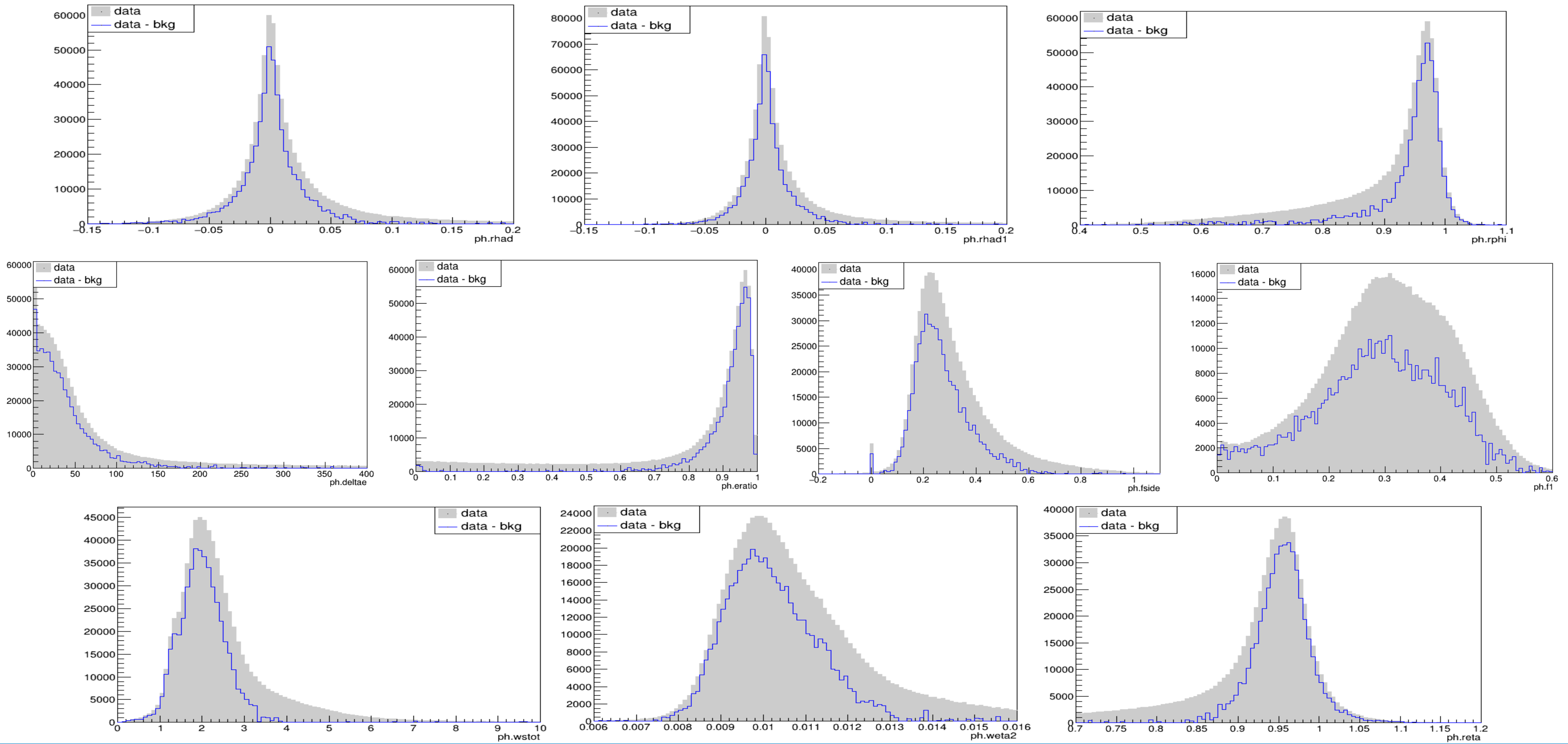
MC reweighting



Background aware MC reweighting

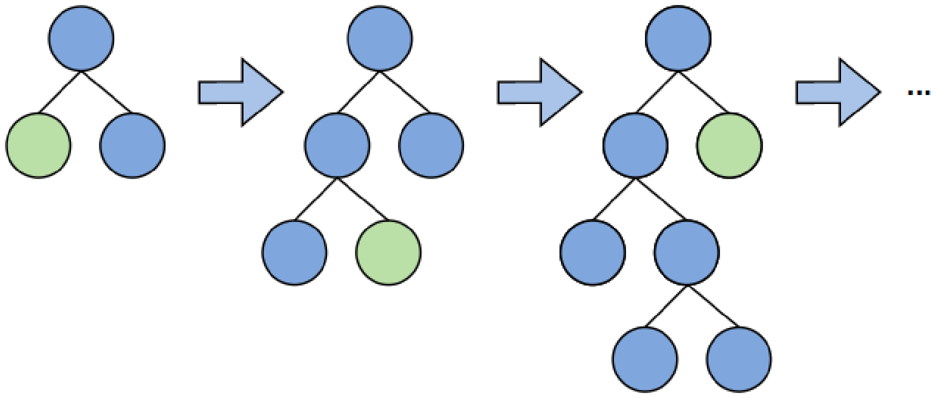


Background subtraction



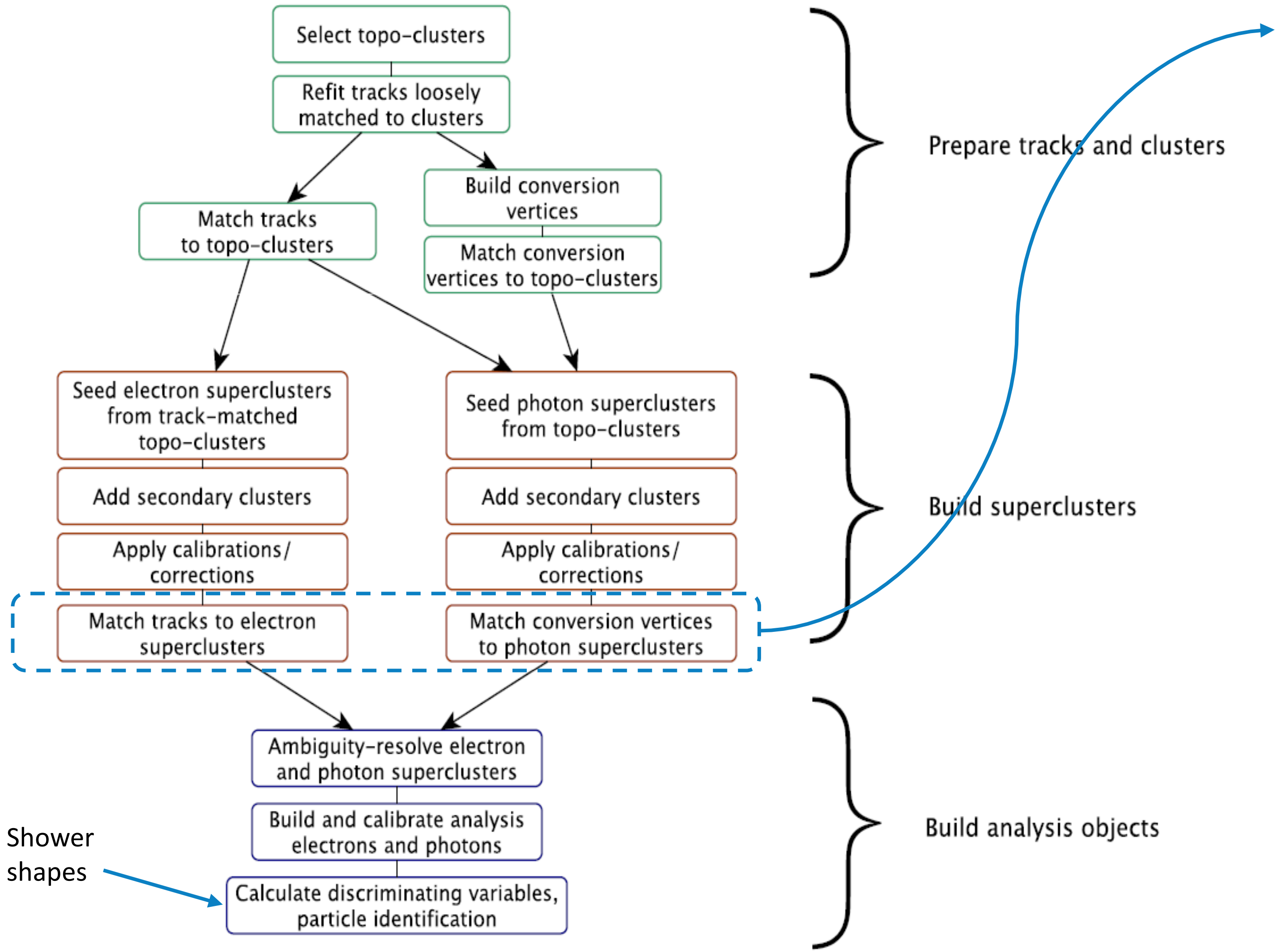
LightGBM

- Light Gradient Boosting Machine (LightGBM) is the machine learning implementations used for this work. It is based on the Gradient Boosting Decision Tree model but it operates with a leaf-wise growth.

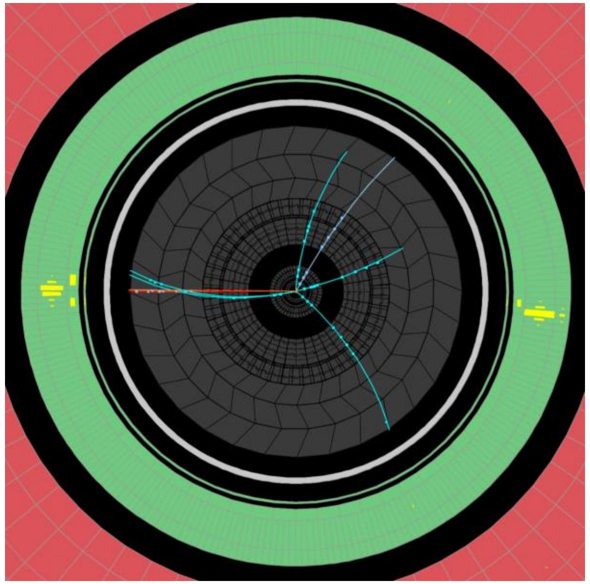


Leaf-wise growth.

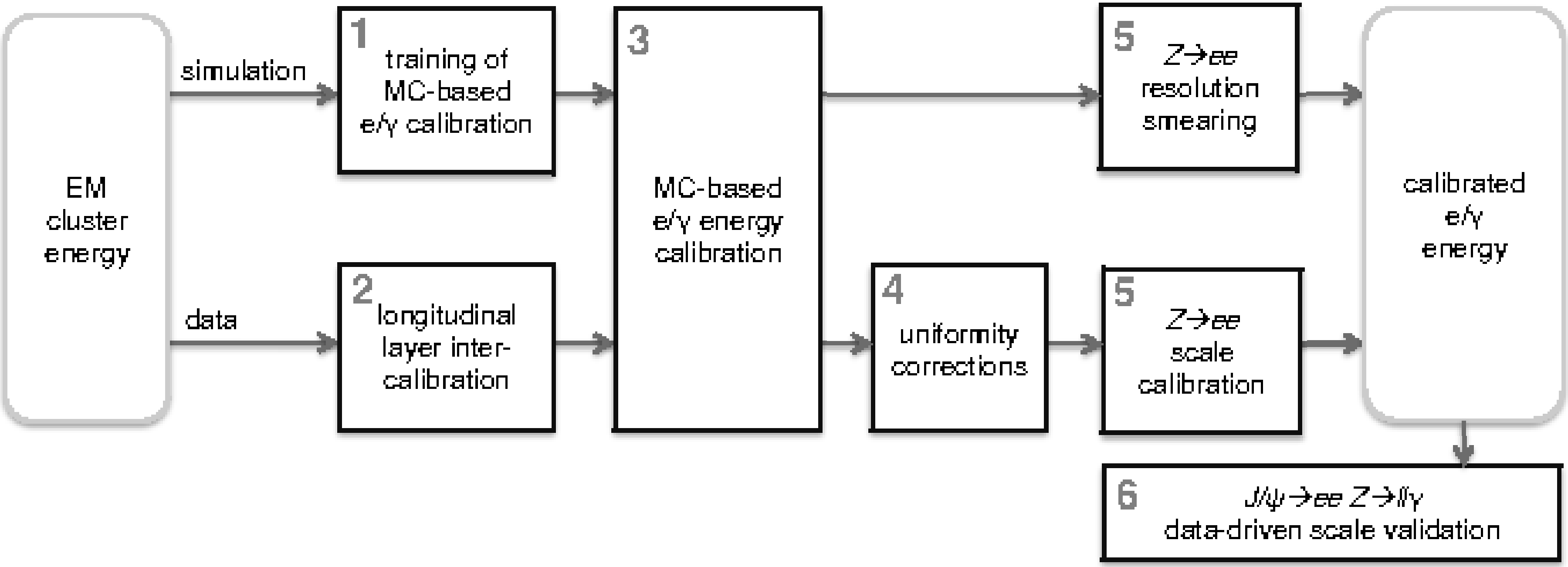
Electron and photon reconstruction



- Electron: object built from energy deposit in calorimeter and a matched track.
- Converted photon (left): supercluster matched to a conversion vertex.
- Unconverted photon (right): supercluster with no associated tracks or conversion vertices.



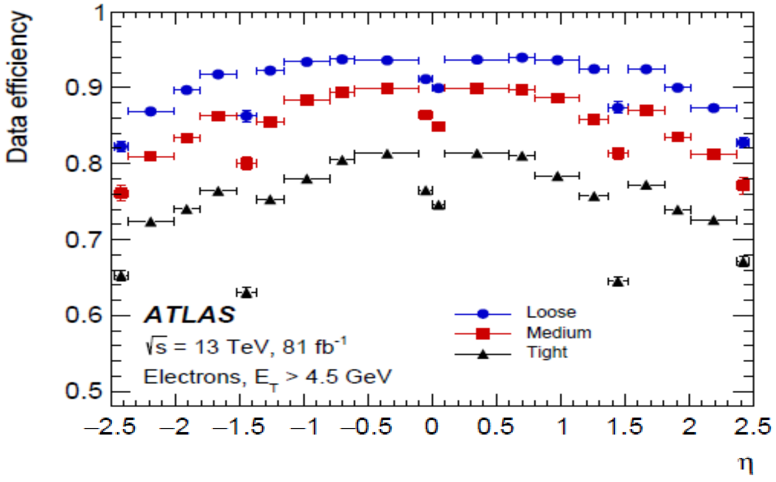
Electron and photon energy energy calibration

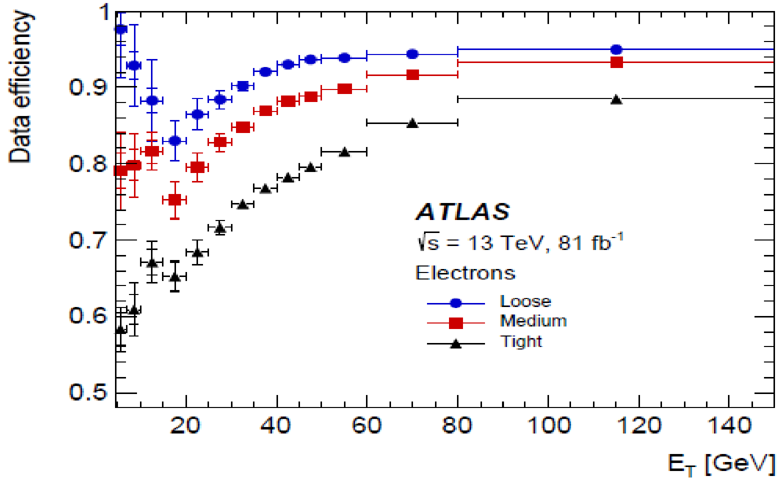


Electron identification

- It's based on a likelihood discrimination to separate isolated electrons from photon conversion, hadron misidentification and heavy flavor decays.
- It uses variables measured in the ID and in the calorimeters.
- It gives better background rejection for a given signal efficiency than a 'cut-based' algorithm.
- Measured in data with 'tag and probe' method: J/Ψ (low E_T) and Z (high E_T) decays to electrons.

- Three working points are defined:
  - Loose (efficiency $\sim 93\%$)
  - Medium (efficiency $\sim 88\%$)
  - Tight (efficiency $\sim 80\%$)





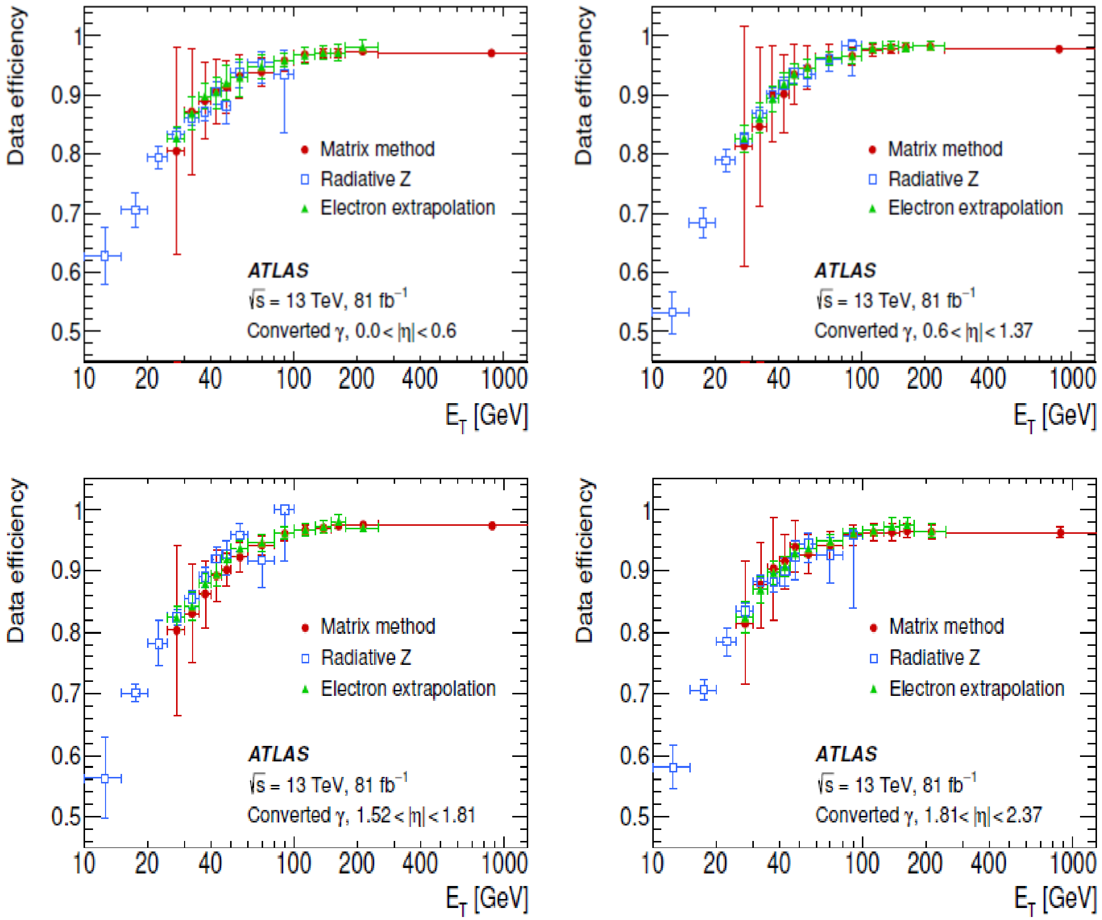
The electron identification efficiency in $Z \rightarrow ee$ events in data for the Loose, Medium and Tight operating points.

Photon identification

- The purpose is to select prompt photons and reject the background (e.g. $\pi^0 \rightarrow \gamma\gamma$).
- 9 discriminating varibles based on energy in cells of electromagnetic and headronic calorimeters are used.

- Three methods:
    - Radiative Z
    - Matrix method
    - Electron extrapolation from $Z \rightarrow ee$

- Loose and Medium: discriminating variables in the HCalo and in the EMCalo middle layer; used by triggers.
- Tight: tighter cuts on discriminating variables; uses also EMCalo strip layer; used for offline analysis.



Loose photon identification efficiency, for converted photons computed with three different methods.

Electron and photon isolation

- Eliminates the remaing background particles misidentified as electrons or photons.
- Remains high energy jets that can be rejected requiring low hadronic activity around the candidate.
- Two variables are defined: E_T^{coneXX} and p_T^{coneXX} which are calculated in a cone centered around the candidate cluster with a radius $\Delta R = XX/100$.

- Three standard selections are defined imposing different thresholds on E_T^{coneXX} and p_T^{coneXX}: *Loose, Tight, Calorimeter-only Tight.*

Efficiency

- In this work the Tight photon identification efficiency is used to assess the goodness of the reweighting applied on the MC samples.
- The efficiceny is computed as:

$$\epsilon = \frac{N_{pID}}{N_{tot}}$$

- In the case of the background subtraction, the number of signal events is given by the fit algorithm.
- The error are computed as:

$$\sigma_N = \sqrt{\sum N^2}$$

- In the case of the background subtraction, the error on N is given by the fit algorithm.