Research Summary

Measurement of the Fake Rate for Hadronic Tau Lepton Decays in the ATLAS Experiment *

Zunran Guo[1]

Supervised by: Timo Dreyer[2]

In consultation with Prof. Stan Lai[2] and Dr. Michel Janus[2]

[1]University of Rochester
[2]II. Institute of Physics, University of Göttingen

August 2017

Abstract

In this research project, a set of analysis on the fake rates for hadronically decaying tau leptons is conducted based on the 2016 data collected by the ATLAS experiment at the LHC (10 times more than the 2015 data), including comparing the re-weighted Monte Carlo simulation and the data, event selection and reweighting, particle truth matching, fake rate measurements with systematic and statistical error analysis, scale factor computing, template fitting to distinguish the distributions of quarks and gluons, and quark and gluon fake rates computing. The analysis results are compared to the previous results from 2015's data.

*This research internship is fully funded by the German Academic Exchange Service Research in Science and Engineering (DAAD RISE).

Contents

1 Introduction

1.1 Physics Background and Purpose of Measuring

The first direct observation of the Higgs boson coupling to leptons was confirmed by the joint effort of the CMS and ATLAS experiments at the Large Hadron Collider (LHC) in 2015 from the $H \to \tau\tau$ signal at a significance of 5.5σ [1]. Due to the high mass of the tau lepton, the branching ratio of the Higgs boson decaying into a $\tau^-\tau^+$ pair is the second highest branching ratio ($6.30 \pm 0.36\%$) of all decays into fermions. This $H \to \tau\tau$ decay is also a good probe for the Yukawa coupling.

To detect final states involving hadronically decaying tau leptons, it is critical to evaluate the performance of the tau lepton reconstruction and identification in the ATLAS experiment. In particular, jets originating from quark and gluon emissions could be falsely identified as hadronically decaying tau leptons. The fraction of misidentified jets (the so-called fake rate) is important to know in order to estimate the background for signal events like $H \to \tau\tau$.

To evaluate the performance of the tau lepton reconstruction and identification, a pure sample of tau candidates originating from quarks or gluons (QCD jets) is probed from the $Z \to ee$ events in Monte Carlo and detector simulation (MC [1]), as well as from $36fb^{-1}$ of data collected by the ATLAS experiment in 2015 and 2016 at the LHC. The decay of a Z^0 boson into an electron positron pair can be easily identified and is therefore "tagged" as a clean channel. Any additional tau lepton candidates found in this event are very likely misidentified QCD jets, instead of real hadronically decaying tau leptons.

In this research project, fake rates of hadronically decaying tau leptons both in MC and in data collected by the ATLAS experiment are calculated and compared. A further analysis on the distribution of the origin of the fake rates, e.g. quark and gluon initiated jets, is also conducted. Previously, only 2015 data were analyzed. Due to the small number of desired events available, the statistical uncertainties of the obtained fake rates were quite high. However, with newly collected 2016 data, about 10 times more events are now available for analysis. It is expected that we are now able to reduce the statistical uncertainties of the fake rates by more than a factor of 3 [2].

1.2 The Large Hadron Collider

The Large Hadron Collider (LHC) (Figure 1) is the world's largest and most powerful particle collider, the most complex experimental facility ever built, and the largest single machine in the world [2]. It was built by the European Organization for Nuclear Research (CERN) between 1998 and 2008 in collaboration with over 10,000 scientists and engineers

[1] The shorthand of MC in the following text means both the Monte Carlo and detector simulation.

[2] Since the events identified by the detector and registered in the specific bins follow Poisson distribution of which the standard error σ is $\sqrt{\lambda}$ where λ is the mean of the total events (A proof on the mean and the standard error of Poisson distribution is included in Appendix A.1 for curious readers.). The new relative uncertainty is now $\frac{\sigma_{New}}{\lambda_{New}} = \frac{\sqrt{\lambda_{New}}}{\lambda_{New}} = \frac{\sqrt{10\lambda_{Old}}}{10\lambda_{Old}} = \frac{\sqrt{10}}{10}\frac{\sqrt{\lambda_{Old}}}{\lambda_{Old}} \approx 0.316\frac{\sigma_{Old}}{\lambda_{Old}} < \frac{1}{3}\frac{\sigma_{Old}}{\lambda_{Old}}$

from over 100 countries, as well as hundreds of universities and laboratories. It lies in a tunnel that is 27 kilometres in circumference, and as deep as 175 meters beneath the France–Switzerland border near Geneva to shield it from the atmospheric radiation [3]. Its first research run (Run I) took place from March 2010 to early 2013 at an energy of 3.5 to 4 teraelectronvolts (TeV) per beam (7 to 8 TeV center-of-mass energy), about 4 times the previous world record for a collider. Afterwards, the accelerator was upgraded for two years. It was restarted in early 2015 for its second research run (Run II), reaching 6.5 TeV per beam (13 TeV total, the current world record).

The aim of the LHC is to allow physicists to test the predictions of different theories of particle physics, including measuring the properties of the Higgs boson and searching for the large family of new particles predicted by supersymmetric theories, as well as other unsolved questions of physics.

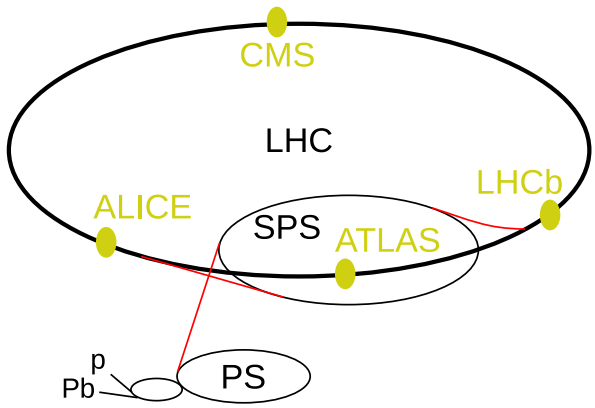


Figure 1: The LHC and the pre-accelerators

1.3 The ATLAS Experiment

The collider has four collision points, around which seven detectors are positioned: ATLAS, CMS, LHCb, ALICE, LHCf, TOTEM, MOEDAL, each designed for certain kinds of research. The ATLAS and CMS experiments are general-purpose detectors that are designed to cover a wide range of physics questions.

The ATLAS (A Toroidal LHC ApparatuS) Experiment (Figure 3) consists of a general-purpose particle detector build around one of the four collision points of the LHC [4]. The four major components of the ATLAS detector are the Inner Detector, the Calorimeter, the Muon Spectrometer and the Magnet System. [5]

1.3.1 Coordinate System

The coordinate system of the ATLAS detector is based on the beampipe, which is defined as the z-axis of the system, where the positive z-direction points anti-clockwise along the LHC ring. The x-axis of the coordinate system is defined to point towards the centre of the LHC ring.

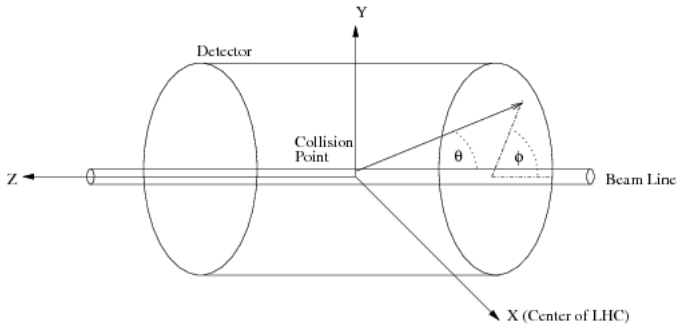


Figure 2: Illustration of the ATLAS and CMS coordinate system. [6]

Alternative coordinates used in the ATLAS experiment are the azimuthal angle ϕ around the z-axis, which is defined with respect to the x-axis, and the pseudorapidity $\eta = -\ln\tan(\theta/2)$, where θ is the polar angle with respect to the z-axis (Figure 2). Distances between two objects reconstructed in the ϕ-η plane are typically given as $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$. [7]

1.3.2 Detector Components

Inner Tracking Detector The inner detector is the first part of ATLAS to see the decay products of the collisions, so it is very compact and highly sensitive. It consists of three different systems of sensors all immersed in a magnetic field parallel to the beam axis. The Inner Detector measures the direction, momentum, and charge of electrically-charged particles produced in each proton-proton collision. [8]

Calorimeters System Calorimeters measure the energy a particle loses as it passes through the detector. It is usually designed to stop entire or absorb most of the particles coming from a collision, forcing them to deposit all of their energy within the detector. Calorimeters typically consist of layers of passive or absorbing high-density material – for example, lead – interleaved with layers of an active medium such as solid lead-glass or liquid argon. Electromagnetic calorimeters measure the energy of electrons and photons as they interact with matter. Hadronic calorimeters sample the energy of hadrons (particles that contain quarks, such as protons and neutrons) as they interact with atomic nuclei. Calorimeters can stop most known particles except muons and neutrinos. [9]

Muon System Muons are particles that usually pass through the Inner Detector and Calorimeter undetected. The muon spectrometer is made up of 4,000 individual muon chambers using four different technologies by 48 institutions in 23 production sites around the world. The muon spectrometer identifies and measures the momenta of muons. [10]

Magnetic System The magnet system bends particles around the various layers of detector systems, making it easier to contain the tracks of particles and to measure the momenta of particles. Two sections consist of magnet system: Central Solenoid Magnet, Barrel Toroid and End-cap Toroids. [11]

1.3.3 Other Experiment Systems

Two system integrated with the detector components are: the Trigger and Data Acquisition System; and the Computing System.

Trigger and Data Acquisition System The trigger system in the latest research run (Run II) consists of a hardware-based first level trigger (Level-1) and a software-based high level trigger (HLT). The Level-1 trigger uses custom electronics to determine Regions-of-Interest (RoIs) in the detector, taking as input coarse granularity calorimeter and muon detector information. The Level-1 trigger reduces the event rate from the LHC bunch crossing rate of approximately 30 MHz to 100 kHz. The decision time for a Level-1 accept is 2.5 μs. The RoIs formed at Level-1 are sent to the HLT in which sophisticated selection algorithms are run using full granularity detector information in either the RoI or the whole event. The HLT reduces the rate from the Level-1 output rate of 100 kHz to approximately 1 kHz on average within a processing time of about 200 ms. [12]

Computing System The computing system analyses the data produced by the ATLAS detector, developing and improving computing software used to store, process and analyze vast amounts of collision data. Data from the ATLAS detector is calibrated, reconstructed and automatically distributed all around the world by the ATLAS Data Management system. The ATLAS Production System then filters through these events and selects the ones needed for a particular type of analysis. With over 130 computing centres worldwide, the ATLAS computing infrastructure and software are constantly evolving. [13]

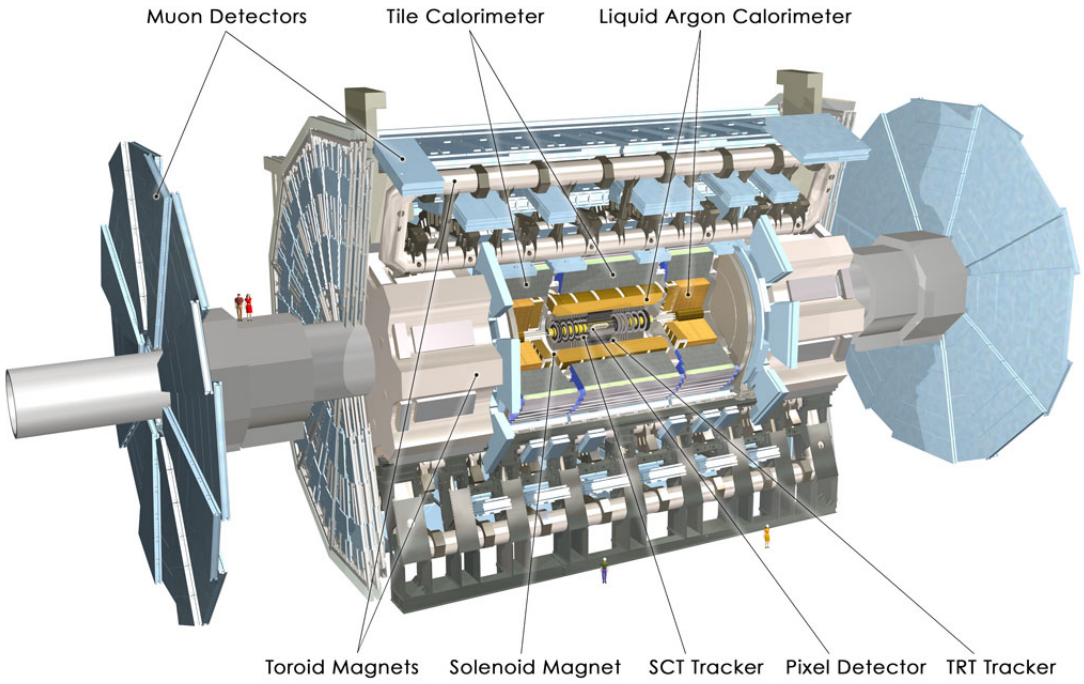


Figure 3: The ATLAS detector [4]

1.4 Hadronic Tau Decays

Hadronic decays of tau leptons produce mainly pions and kaons. These decays are further classified by their numbers of charged tracks, or "prongs", produced in the decay. Due to charge conservation, tau leptons can only decay into an odd number of charged particles. The 1 prong decays occur about 3 times as often as 3 prong decays, since the phase space is larger for decays into fewer particles. In addition, the number of vertices in the 1-prong case is also smaller than the number of vertices in the 3-prong case, hence the higher probability for 1-prong case. Higher prong numbers can also occur, but are neglected in this research, since their branching ratio is very low.

Figure 4 shows schematics for a 1-prong and a 3-prong tau lepton decay. A jet cone containing the decay products of a tau lepton is usually much narrower than a typical cone containing a QCD jet; and the leading particles inside the jet tend to carry higher fractions of the initial momentum of the tau lepton. In addition, the mean lifetime (290.3 $\pm$ 0.5 fs [14]) of the tau lepton makes it possible to reconstruct the secondary vertex from which the jet originates.

(a) 1-prong (b) 3-prong

Figure 4: Hydronic decays of tau leptons

1.5 Tau Lepton Identification

To distinguish QCD jets and real tau lepton decays, a boosted decision tree (BDT) is used (Figure 5). A single decision tree is a binary tree with an if-statement at each node. A decision is made at each node as the τ candidate traverses the tree until an end node is reached. Each end node is labelled either as signal (τ candidate) or background (QCD jets) depending on which category the majority of candidates in this node belong to. A single decision tree can be interpreted as a set of simple rectangular cuts on the parameter space of the observables.

For a BDT, a set of trees (a "forest") is trained and the output of each tree is interpreted as either 1 if the output is "signal" or 0 if the ouput is "background". For each event the ouputs of all trees are calculated and then combined into a single number (usually by averaging the outputs) which acts as a final discriminant. If the number is close to 1, the event is finally classified as "signal-like". If the number is close to 0, the event is finally classified as a "background-like".

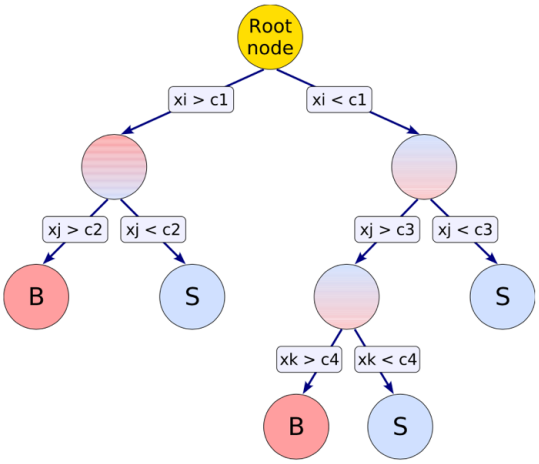


Figure 5: Schematic view of a decision tree. Starting from the root node, a sequence of binary splits using the discriminating variables x_i is applied to the data. Each split uses the variable that gives the best separation between signal and background when being cut on at this node. The same variable may thus be used at several nodes, while others might not be used at all. The leaf nodes at the bottom end of the tree are labeled "S" for signal and "B" for background depending on the majority of events that end up in the respective nodes.

2 Computing Environment Setup

2.1 Remote Login

To maximize the production efficiency, almost all the scripts are executed remotely on a batch system, instead of the on the computer one is using. Many options are available during this research. The login procedure in details can be referred in Appendix A.2.

2.2 ROOT Analysis Framework

Another core requirement in this research is the installation of ROOT [15]. ROOT is a modular scientific software framework mainly used in high energy physics. It provides all the functionalities needed to deal with big data processing, statistical analysis, visualisation and storage. It is mainly written in C++ but integrated with other languages such as Python and R.

One can easily install ROOT on a local computer. The installation process is well detailed on its website.

2.3 Data Format Workflow

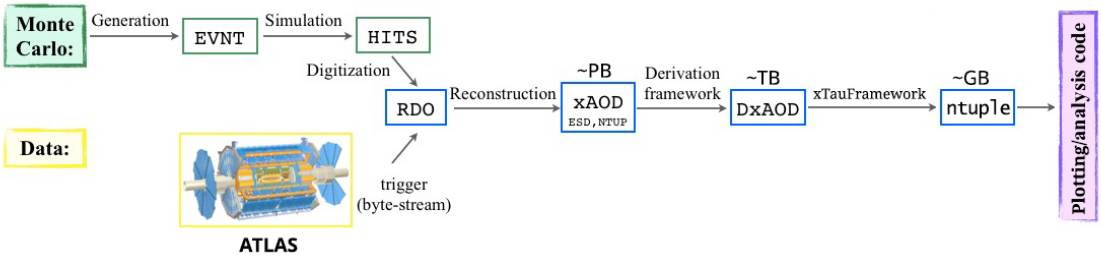


Figure 6: Data Format Workflow

Figure 6 above shows the data format workflow in the ATLAS experiment. HITS files store the simulated energy depositions in the detector simulation from Geant 4. RDO (Raw Data Object) is a C++ object representation of the byte-stream information. xAOD (Analysis Object Data) is a new format for Run II, which is a summary of the reconstructed event, and contains sufficient information for common analyses (tracks, jets, taus, etc.), accessible in Athena or Standalone ROOT. DxAOD is the derived xAODs with refined content and event selections, specific to physics and performance groups. In the next step, the information stored in the DxDAO format is transformed into ntuples, which are ROOT [15] files containing flat `TTree` objects, in which the desired physics variables are stored. Through reading ntuples, each of which has a typical size in the order of gigabytes, the actual analysis is performed and plots are generated.

In this project, the ntuples are stored at `/nfs/dust/atlas/user/tidreyer/ntuple/GRID_170815_TAUP`. The script `download.sh` in this directory could be run to download the ntuples from the Worldwide LHC Computing Grid (WLCG) first via rucio [16].

2.4 Plotting Framework

The repository of the complete research analysis code is hosted at
https://gitlab.cern.ch/zuguo/fakerate-plotting.
The analysis and debugging scripts are in the directory
`fakerate-plotting/trunk/analysis/Zee/analysis`.
The paths to the ntuples used in this research analysis are in the `.xml` files in the directory
`fakerate-plotting/trunk/analysis/Zee/datasets`.
The plotting framework used in this analysis is in the directory
`fakerate-plotting/trunk/plotting`.
This plotting library is written by Christian Grefe who is currently based at CERN, from the University of Bonn.

To set up the computing environment for the plotting framework on the server, one should run `source fakerate-plotting/trunk/analysis/Zee/env.sh` before running any other plotting scripts (e.g. `dataMC.py`, `FakeRates_Syst.py`, `TemplateFit.py`) in this research project.

3 Analysis

3.1 Monte Carlo and Detection Simulation

The MC samples used in this research are generated by POWHEG [17] and interfaced with PYTHIA 8 [18] for parton showering using the AZNLO tune [19] and CTEQ6L1 [20] as the parametrization for parton distribution functions. The $Z \to ee$ process in proton proton collisions with jets are produced from MC in the final state at a center-of-mass energy of 13 TeV. Other processes produced from MC with similar final states are used to estimate the background contribution in the final selection.

Each MC event also contains simulation of *pile-up* [21], which is produced at the LHC due to the fact that for each bunch crossing multiple collisions occur inside the ATLAS detector. The particles produced in different collisions of one bunch crossing and the neighboring crossings would be detected simultaneously with the desired events and therefore affect the reconstruction and identification efficiencies.

After the MC event generation, desired $Z \to ee$ events are further preselected through a list of triggers and criteria.

3.1.1 Event Selection

The $Z \to ee$ events for the analysis are preselected by a single electron trigger with the additional requirement that the particle activating the trigger must also match the leading reconstructed lepton in the events.

Since the low-threshold Level 1 trigger includes an "absolute" isolation criterion, the relative isolation criteria decrease with the transverse momentum of the electron candidate. Therefore the triggers with a p_T threshold of 24 GeV and 60 GeV are only applied to the leading leptons with p_T values below 65 GeV and 135 GeV respectively (Table 1).

The efficiencies of the triggers applied to MC and data for a $p_T(\ell_{lead})$ are then corrected by a scale factor, which corrects the modeling of the effects in MC in comparison to data (Section 3.2.1).

The events are further required to contain at least two electrons and at least one τ candidate. Any event containing muons is rejected.

In accordance with the tag and probe method (Section 3.3.1), the two leading electrons must also pass a medium electron identification criterion. In addition, the two leading electrons must be well isolated and carry opposite charge. The p_T threshold of the leading electron is set at 26 GeV, while the threshold of the sub-leading electron is set at 20 GeV. The reconstructed invariant mass of the two electrons needs to be compatible with the known Z^0 boson mass of about 91.19 GeV within $\pm$ 5 GeV, which is about two times the decay width $\Gamma_Z = 2.4952 \pm 0.0023$ GeV of the Z^0 boson [14].

After this event selection, the leading τ candidate is used in the fake rate measurement.

Data Taking Period	$p_T(\ell_{\text{lead}})$ Cut	Trigger
2015	$< 65\,\text{GeV}$	HLT_e24_lhmedium_L1EM20VH
	$< 135\,\text{GeV}$	HLT_e60_lhmedium
	unlimited	HLT_e120_lhloose
2016	$< 65\,\text{GeV}$	HLT_e26_lhtight_nod0_ivarloose
	$< 135\,\text{GeV}$	HLT_e60_lhmedium_nod0
	unlimited	HLT_e140_lhloose_nod0

Table 1: Single electron triggers applied for the analysis. In addition, a trigger match to the leading lepton is required. Each event needs to pass at least one of these criteria.

3.1.2 Truth Matching

The process of associating a reconstructed object in the detector simulation with the corresponding truth particle is known as truth matching (Figure 7). In this research, the tau analysis tools (TAT) truth matching algorithm of the xTauFramework has been used to match candidates for hadronically decaying tau leptons to quarks and gluons.

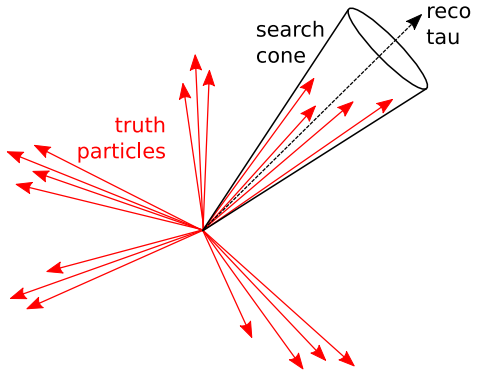


Figure 7: Scheme for the truth matching algorithm. Only particles within a search cone of $\Delta R < 0.2$ for leptons and $\Delta R < 0.4$ for partons are considered in the matching algorithm.

3.2 Data/MC Comparison

Due to the fact that MC sample events generated are more than the data events collected at the ATLAS experiment, before the MC samples are used in the analysis, the luminosity weights are applied to all the events in the sample. The luminosity weights scales the number of events in the MC samples to the expected number in data. In addition to luminosity weights, in this research project, p_T weights of the Z^0 boson are also applied to each bin so as to improve the modeling of the data.

3.2.1 Luminosity Weighting of MC Events

The amount of pile-up during a data-taking period is influenced by many different experimental factors, which makes it nearly impossible to predict the exact distribution.

Therefore, the selected events of MC samples that are simulated with a predicted pile-up distribution are also weighted to match the measured pile-up distribution of the data.

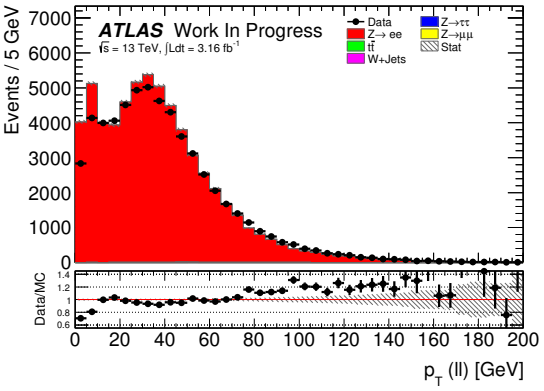
After the distributions of an MC sample is corrected by applying a weight w_i to each event i, the total number of events are then scaled to the expected number of events N^X_{exp} of the given process X in data. This number is given by the product of the integrated luminosity $\int \mathcal{L}dt$ of the data and the cross section σ_X of the process.

Since the total number of modeled events is changed by the weights applied to the MC sample, the expected number N^X_{exp} should be compared to the sum of these weights to obtain the scale factor W^X_{lumi} [3]:

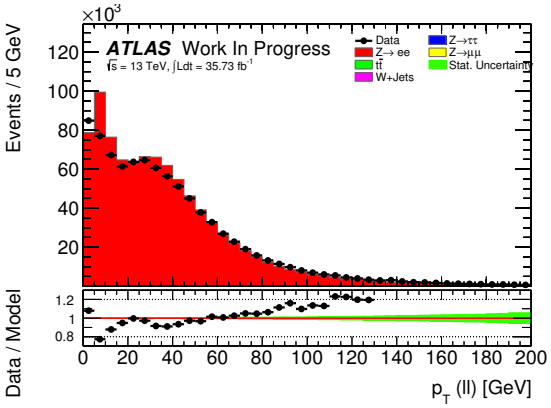
$$W^X_{lumi} = \frac{N^X_{exp}}{N^X_{MC}} = \frac{\sigma_X \cdot \int \mathcal{L}dt}{\sum_i w_i}$$

3.2.2 Z Boson Transverse Momentum Weighting of MC Events

Since the kinematic variables like ΔR of the two leading leptons and the transverse momentum of the leading lepton and the leading τ candidate are correlated (Figure 12, 13a, 13b, 14a, and 14b), a mismodeling of the p_T of the Z^0 boson could also produce a mismodeling in the other kinematic variables. Therefore, in addition to luminosity weighting, the data/MC ratio of the transverse momentum distribution of the Z^0 boson is applied as an additional scale factor to reweight MC events, so as to improve the modeling of the data.



(a) 2015 events only (before reweighting)



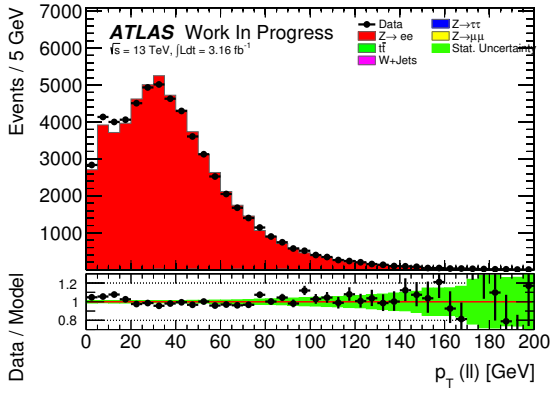
(b) Combined 2015 and 2016 events (before reweighting)

Figure 8: Comparison of the transverse momentum (p_t) of the Z^0 boson as reconstructed from the two leading leptons in the 2015 and 2016 events with τ selection (before reweighting)

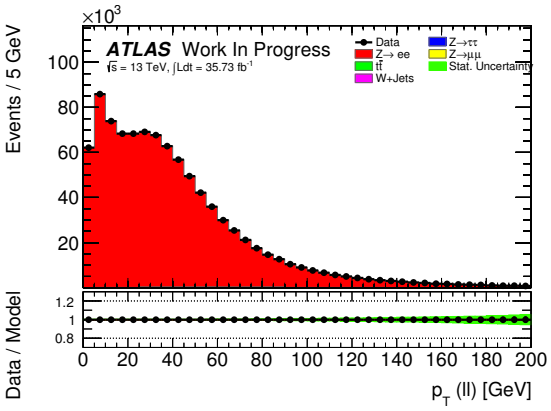
Note that before running `dataMC.py` for data and MC comparison, one need to generate the weights first by running `reweighting.py`. The weights will be stored in

[3] Note that the scale factor for luminosity reweighting is built into the `process.getHistogram()` function across the plotting framework.

`SF_dipleton_pt_vect` in `.C`, `.eps`, `.png`, `.pdf`, and `.root` formats respectively. So are all the plots generated in this project.



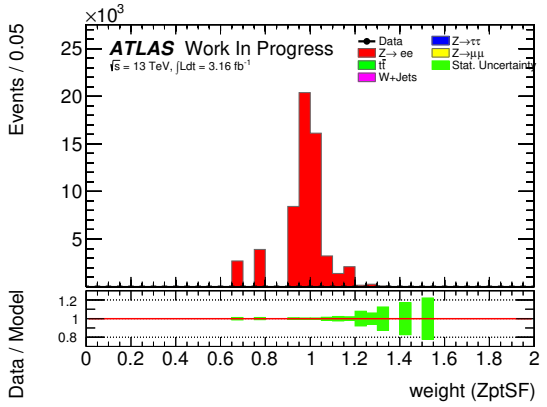
(a) 2015 events only (after reweighting)



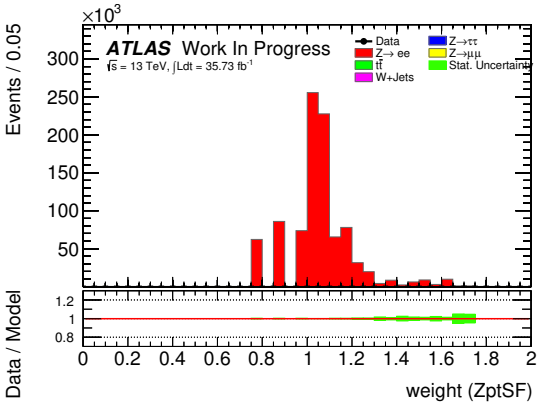
(b) Combined 2015 and 2016 events (after reweighting)

Figure 9: Comparison of the transverse momentum (p_T) of the Z^0 boson as reconstructed from the two leading leptons in the 2015 and 2016 events with τ selection (After reweighting)

Note that Figure 9a is reweighted before τ selection, and Figure 9b is reweighted after τ selection.

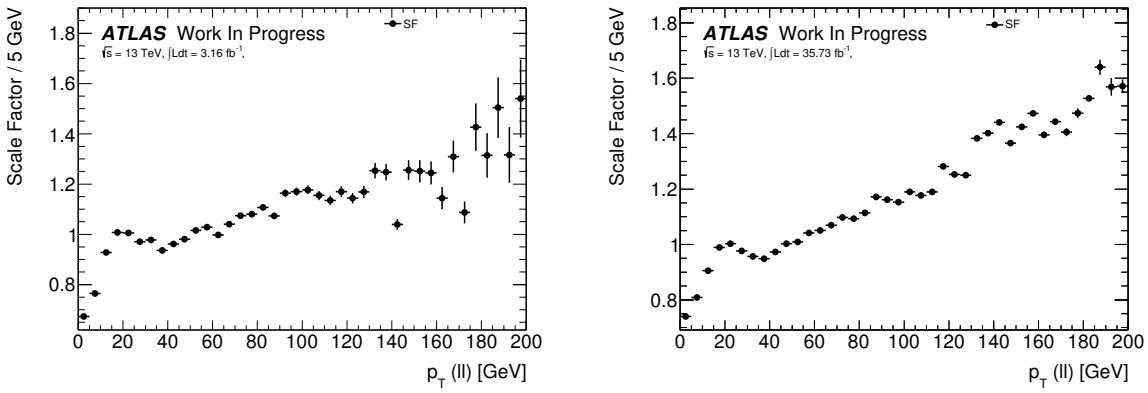


(a) 2015 events only



(b) Combined 2015 and 2016 events

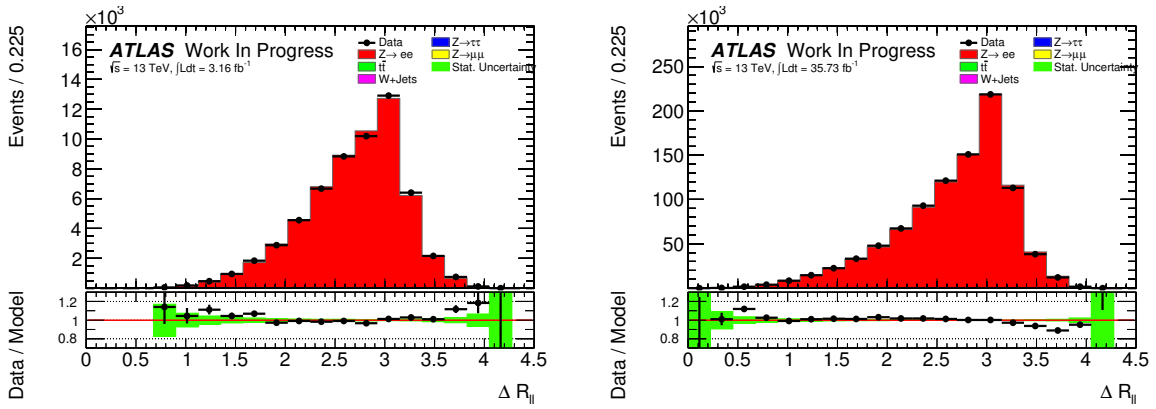
Figure 10: Distribution of applied weights for the MC events that are passing the event selection



(a) 2015 events only

(b) Combined 2015 and 2016 events

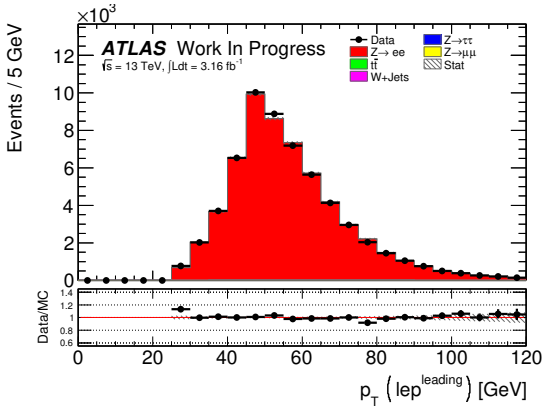
Figure 11: Weights used for the reweighting of the MC samples vs. reconstructed p_T of the Z boson



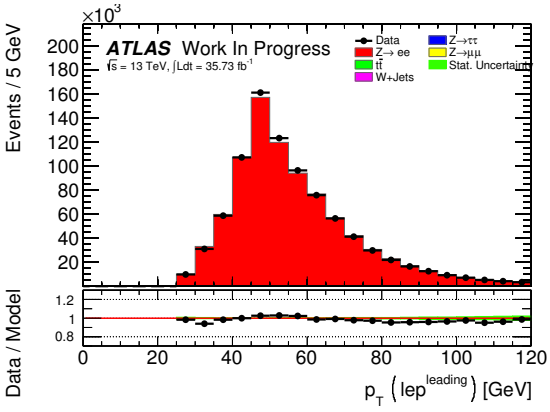
(a) 2015 events only

(b) Combined 2015 and 2016 events

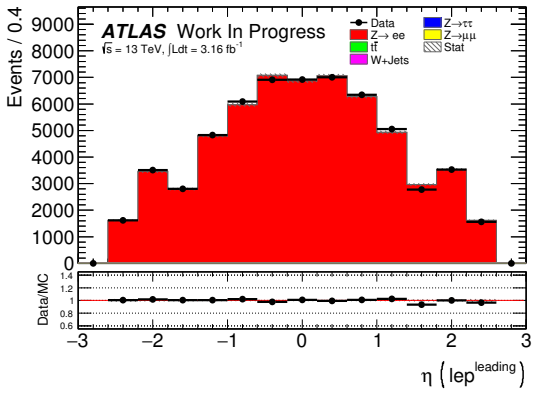
Figure 12: Comparison of data/MC agreement in ΔR of dileptons after the MC events have been reweighted.



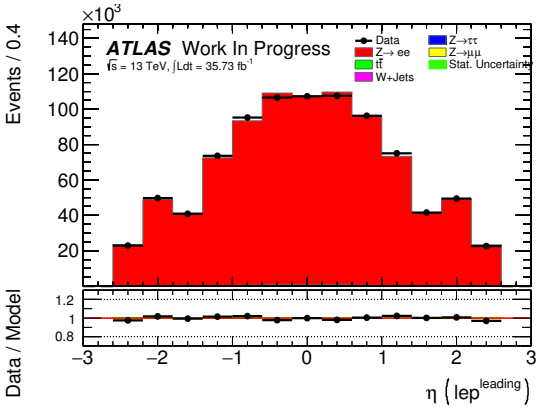
(a) 2015 events only



(b) Combined 2015 and 2016 events

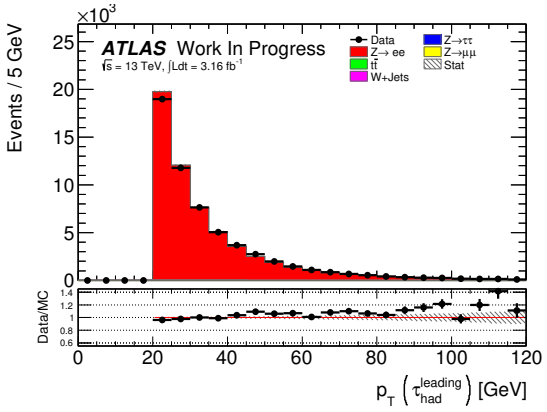


(c) 2015 events only

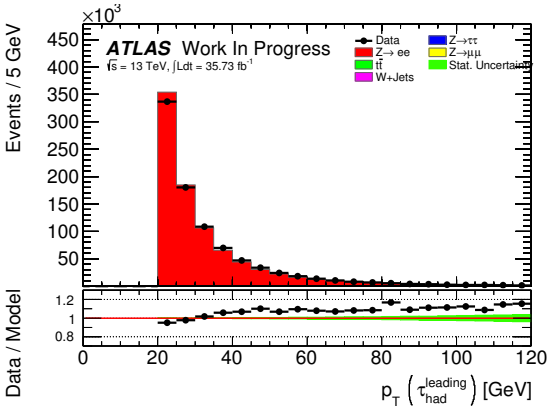


(d) Combined 2015 and 2016 events

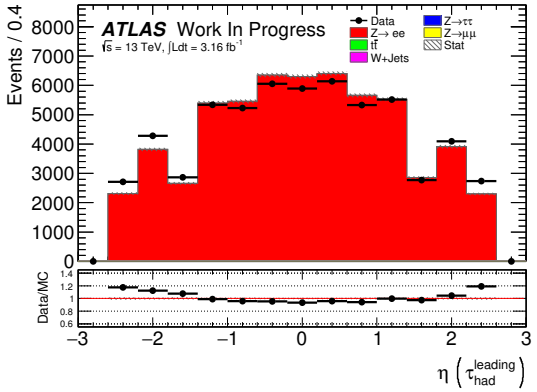
Figure 13: Comparison of the leading τ candidate kinematic variable data/MC agreement after reweighting the MC events. The contributions of MC samples other than $Z \to ee$ are too small to be visible after the tag-and-probe method is applied.



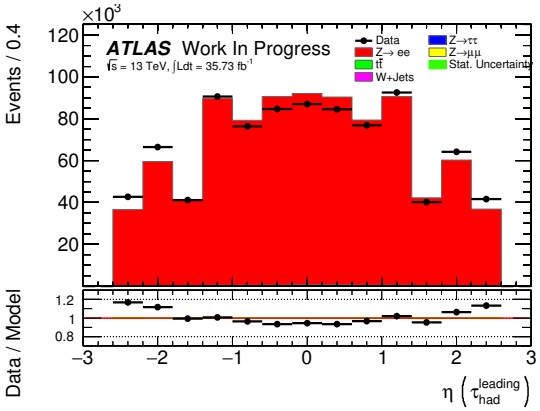
(a) 2015 events only



(b) Combined 2015 and 2016 events

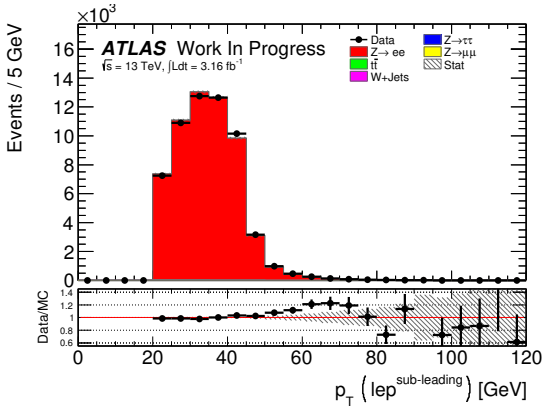


(c) 2015 events only

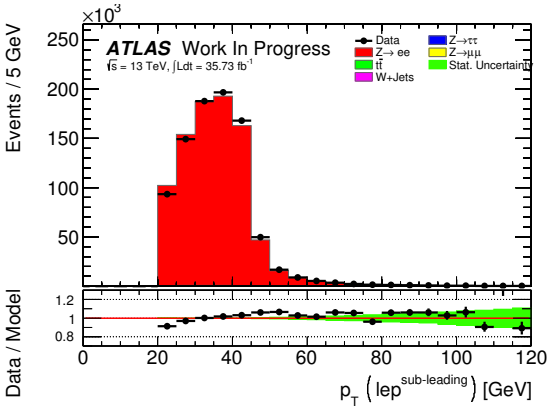


(d) Combined 2015 and 2016 events

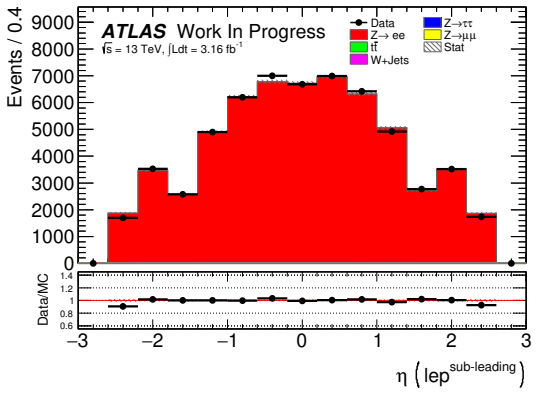
Figure 14: Comparison of the leading lepton kinematic variable data/MC agreement after reweighting the MC events. The contributions of MC samples other than $Z \to ee$ are too small to be visible after the tag-and-probe method is applied.



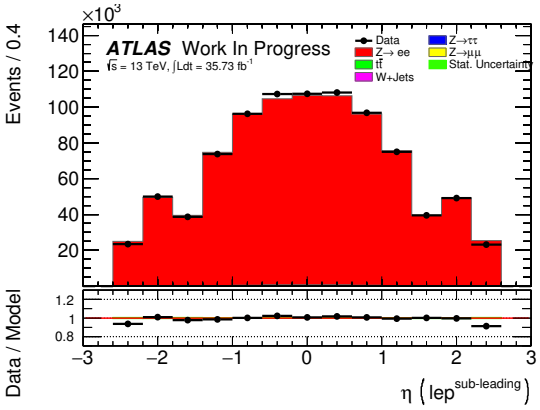
(a) 2015 events only



(b) Combined 2015 and 2016 events



(c) 2015 events only



(d) Combined 2015 and 2016 events

Figure 15: Comparison of data/MC agreement of subleading particles in p_T and η after the MC events have been reweighted. The contributions of MC samples other than $Z \to ee$ are too small to be visible after the tag-and-probe method is applied.

3.3 Fake Rate Measurement

As motivated in Section 1.5, the performance of the tau lepton identification algorithm is very important in all analyses involving hadronic tau lepton decays, including measurements involving $H \to \tau\tau$ decays. One of the major backgrounds for this analysis, beyond the irreducable background from $Z \to \tau\tau$ decays, results from the hadronic jets (QCD jets) that are misidentified as hadronically decaying tau leptons.

One way to quantify this background is the measurement of the fake rate (FR), which is defined as the fraction of QCD jets that pass the tau identification algorithm out of the total number of jets that are reconstructed as τ candidates:

$$FR = \frac{\#\text{jets } (\tau\text{-reco, selection, } \tau\text{-ID})}{\#\text{Jets } (\tau\text{-reco, selection})}$$

3.3.1 Tag and Probe Method

To measure the fake rate, it is critical to select a clean and pure sample, from which the misidentified τ candidates produced from jets can be easily identified. In this analysis, the decay of a Z^0 boson into an electron positron pair is used as the event "tag". Any additional τ candidates in the same event are very likely to be misidentified QCD jets, instead of real tau decays. Figure 16 16 shows one example of such a process.

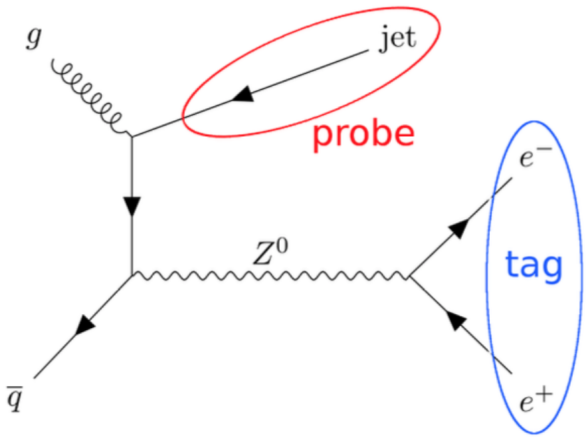


Figure 16: A possible Feynmann diagram for a $Z \to ee$ event with an additional jet. This process acts as a clean and pure sample for the tag and probe method.

3.3.2 Estimation of Systematic Uncertainties

For the systematic uncertainties in this analysis, *up* and *down* variations are applied on the MC samples in `FakeRates_Syst.py` to estimate their effects on the fake rates. The sources are listed below:

Electron Scale Factor To improve the modelling of the data, multiple scale factors are applied to the MC events. Since these scale factors are determined from a data to MC comparison, each of them has an assigned uncertainty, which determines the up and down variation. Systematic uncertainties have been determined for the scale factors associated with the choice of identification requirements on the leading lepton in the event (`MediumLLH`), the modeling of the track reconstruction, the isolation criterion applied to electrons and the choice of the trigger. [7]

Electron Energy Scale The tag-and-probe method relies on the reconstruction of the invariant mass of the two electrons produced in a $Z \to ee$ decay to separate signal from background events. Therefore, the measurement of the energy and momentum of electrons plays a significant role for the amount of background in the selection. The MC modelling of the interpretation of the detector output for the electrons is reflected in two systematic uncertainties for the resolution and the scale of the calibration. [7]

Z Mass Window For the tag-and-probe method a ± 5 GeV window around the Z^0 boson mass is defined, in which the invariant mass of the two electrons is required to fall.

Since the choice of this width is to some extent arbitrary and could affect the fake rate measurement, a variation of the window's width up to ±8 GeV and down to ±4 GeV is considered as an additional source of uncertainty. [7]

Tau Energy Scale The reconstruction of the τ candidate in the event uses the tau energy scale (TES), which is derived from simulated events. [22] The calibration of the TES introduces systematic uncertainties on the measured properties of the τ candidate arising from the simulation of the detector, the choice of the used MC model and the comparison to in-situ measurements. [7]

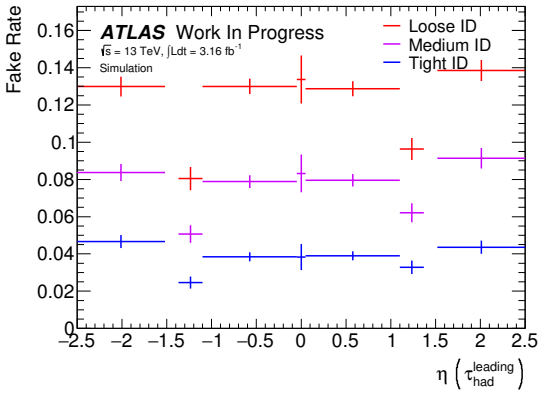
	1 Prong			3 Prong		
Source of Uncertainty	Loose	Medium	Tight	Loose	Medium	Tight
Statistics	8.9 %	14 %	22 %	18 %	30 %	41 %
Z Mass Window	4.1 %	5.9 %	5.7 %	5.1 %	3.8 %	11 %
Electron Energy Resolution	1.7 %	6.8 %	9.4 %	3.5 %	4.5 %	9.8 %
Electron Energy Scale	2.4 %	6.3 %	9.4 %	3.4 %	4.7 %	8.8 %
Electron SF MediumLLH	0.05 %	0.03 %	0.03 %	0.04 %	0.07 %	0.09 %
Electron SF Trigger	0.04 %	0.02 %	0.03 %	0.08 %	0.10 %	0.08 %
Electron SF Reco Track	0.02 %	0.02 %	0.02 %	0.03 %	0.03 %	0.06 %
Electron SF Isolation	0.01 %	0.005 %	0.008 %	0.03 %	0.04 %	0.03 %
Tau Energy Scale In-situ	0.05 %	0.08 %	0.007 %	0.30 %	0.01 %	0.01 %
Tau Energy Scale Model	0.04 %	0.08 %	0.007 %	0.15 %	0.01 %	0.01 %
Tau Energy Scale Detector	0.0 %	0.0 %	0.0 %	0.01 %	0.01 %	0.01 %

Table 2: The effect of the considered uncertainties on the fake rates for the three different working points of the tau identification algorithm for 2015 events. Listed is always the uncertainty of the p_T bin with the largest absolute uncertainty.

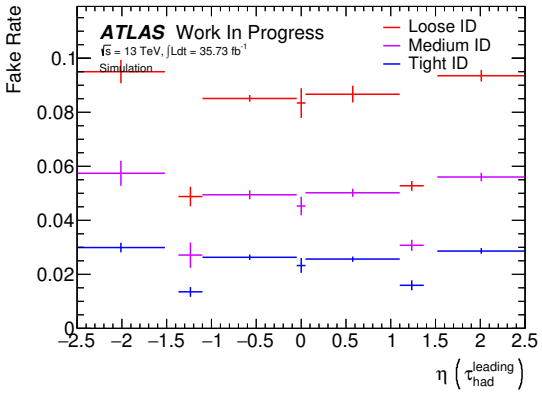
Source of Uncertainty	1 Prong			3 Prong		
	Loose	Medium	Tight	Loose	Medium	Tight
Statistics	0.34 %	0.27 %	0.19 %	0.062 %	0.039 %	0.023 %
Z Mass Window	0.070 %	0.065 %	0.045 %	0.024 %	0.016 %	0.0083 %
Electron Energy Resolution	0.046 %	0.026 %	0.022 %	0.0082 %	0.0046 %	0.0017 %
Electron Energy Scale	0.011 %	0.011 %	0.012 %	0.0035 %	0.0032 %	0.0034 %
Electron SF MediumLLH	0.0030 %	0.0027 %	0.0015 %	0.00063 %	0.00027 %	0.000082 %
Electron SF Trigger	0.0016 %	0.0012 %	0.00060 %	0.00020 %	0.000082 %	0.000054 %
Electron SF Reco Track	0.00079 %	0.00057 %	0.00034 %	0.000084 %	0.000048 %	0.000014 %
Electron SF Isolation	0.0026 %	0.0020 %	0.0010 %	0.00033 %	0.00012 %	0.000049 %
Tau Energy Scale In-situ	0.25 %	0.21 %	0.15 %	0.065 %	0.040 %	0.022 %
Tau Energy Scale Model	0.27 %	0.24 %	0.29 %	0.061 %	0.048 %	0.024 %
Tau Energy Scale Detector	0.034 %	0.052 %	0.10 %	0.018 %	0.017 %	0.016 %

Table 3: The effect of the considered uncertainties on the fake rates for the three different working points of the tau identification algorithm for combined 2015 and 2016 events. Listed is always the uncertainty of the p_T bin with the largest absolute uncertainty.

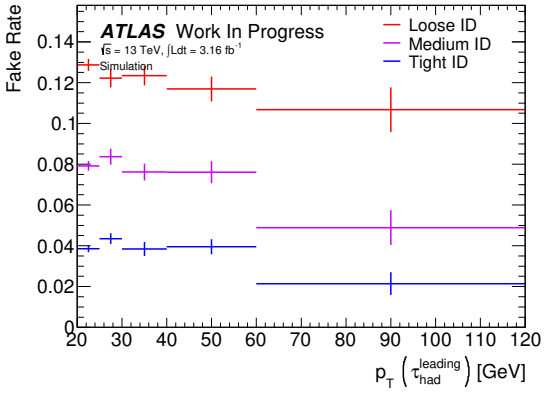
3.3.3 Fake Rates in MC



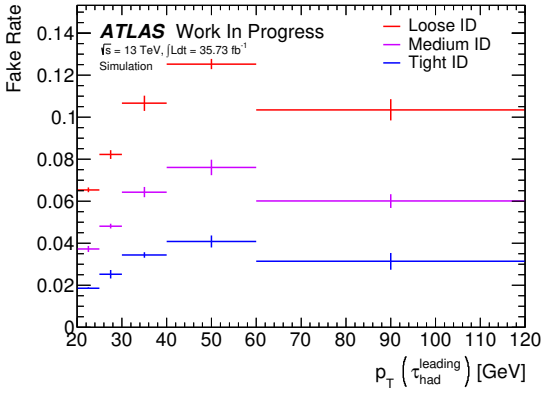
(a) 2015 events only



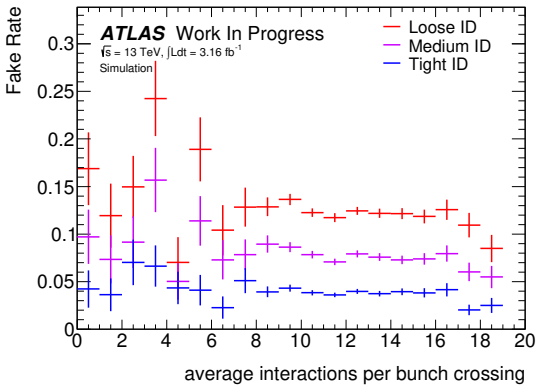
(b) Combined 2015 and 2016 events



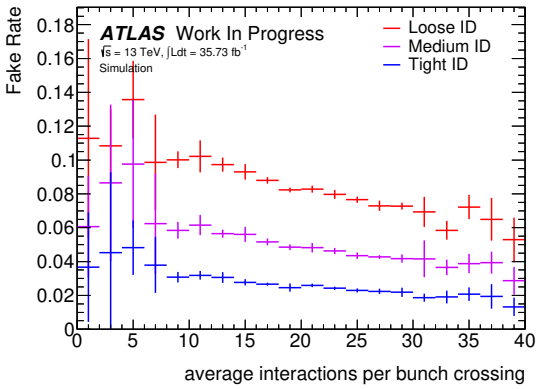
(c) 2015 events only



(d) Combined 2015 and 2016 events

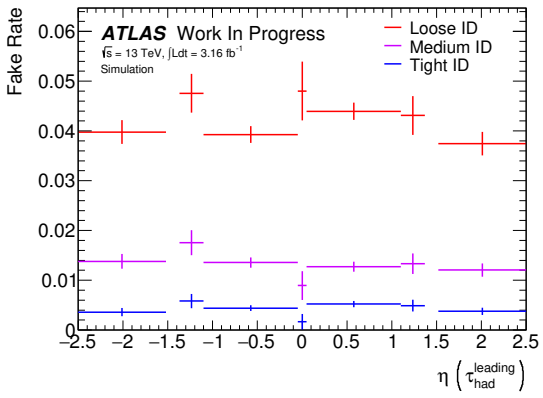


(e) 2015 events only

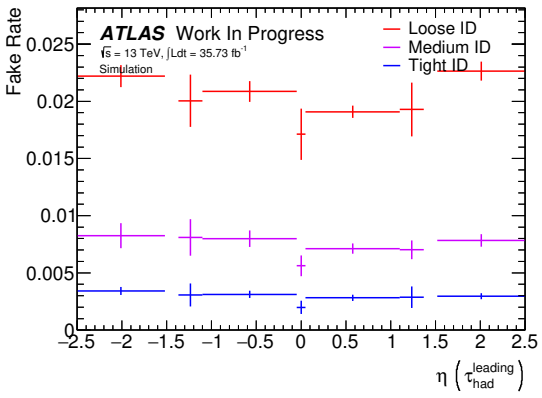


(f) Combined 2015 and 2016 events

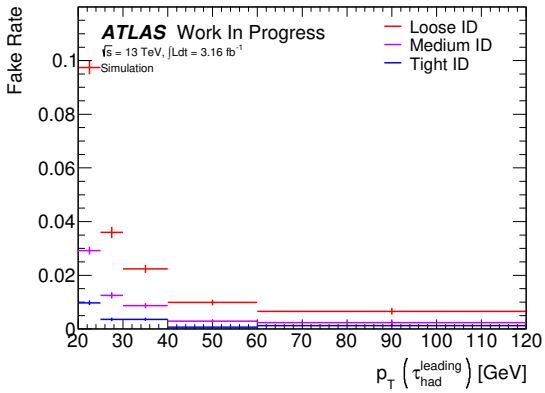
Figure 17: Fake rates determined from $Z \to ee$ MC sample using the tag and probe method for τ candidates with one charged track



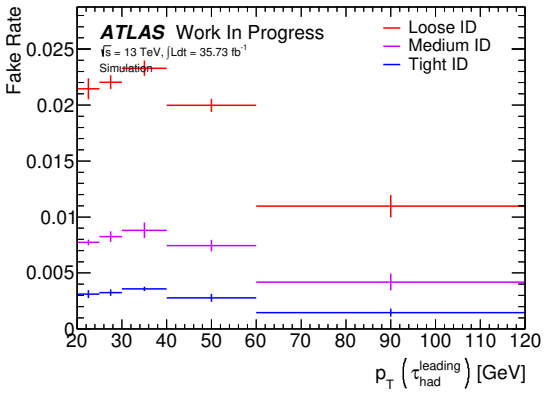
(a) 2015 events only



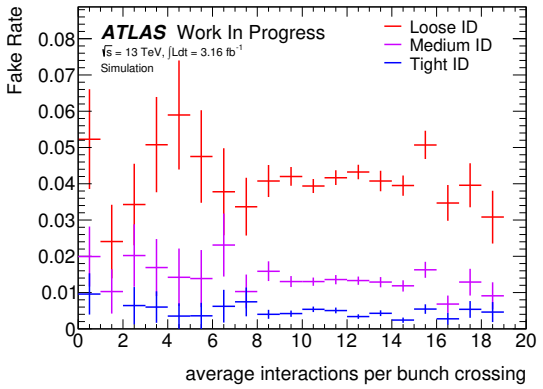
(b) Combined 2015 and 2016 events



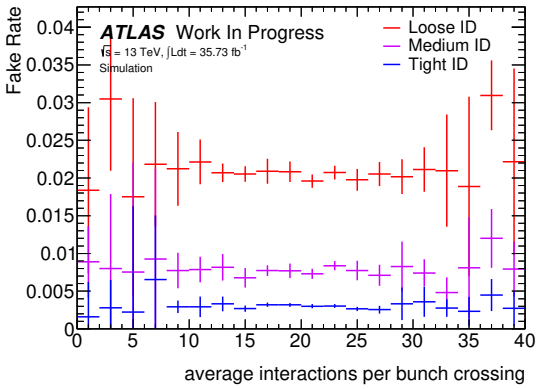
(c) 2015 events only



(d) Combined 2015 and 2016 events



(e) 2015 events only



(f) Combined 2015 and 2016 events

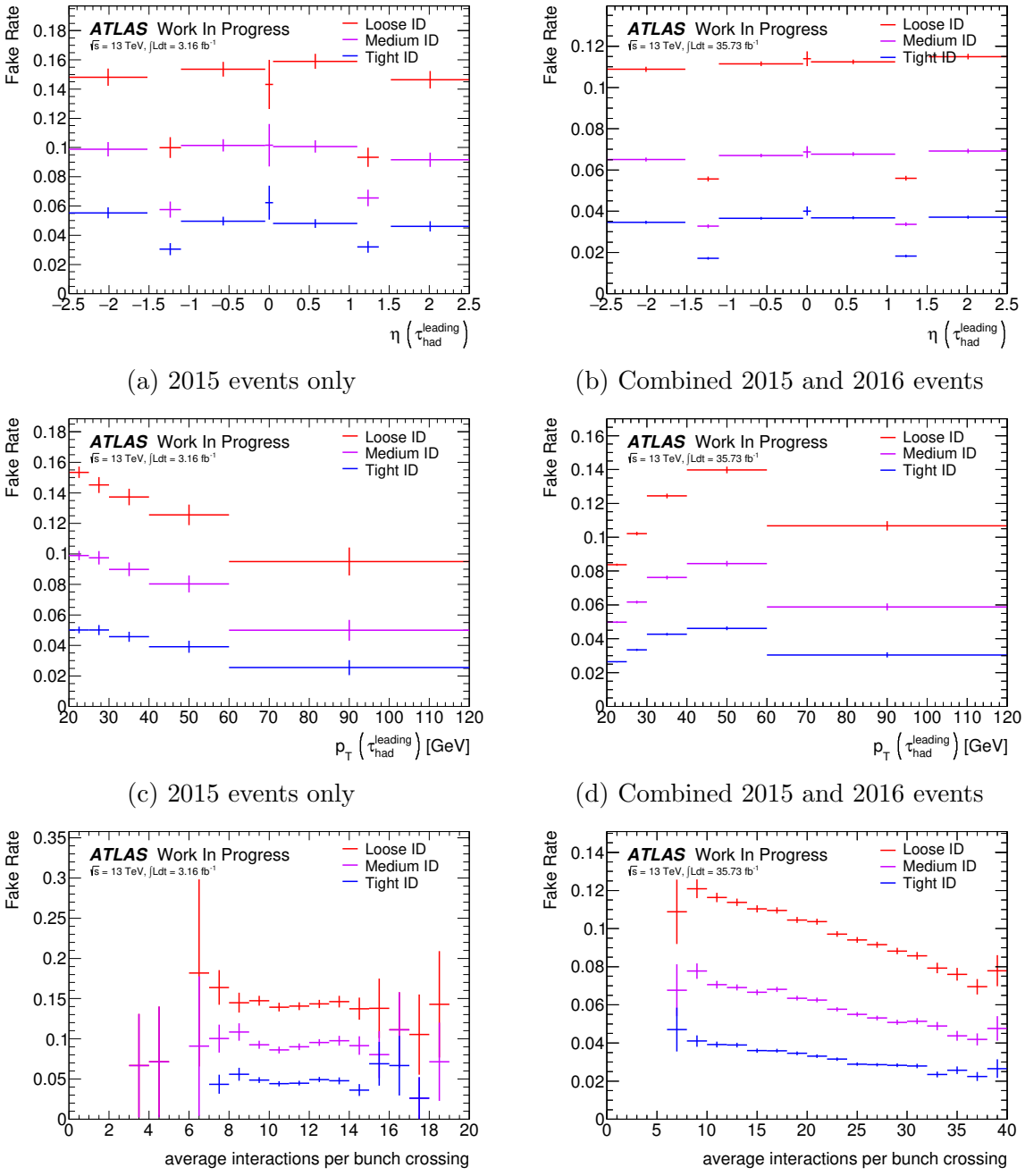
Figure 18: Fake rates determined from $Z \to ee$ MC sample using the tag and probe method for τ candidates with three charged track

From Figure 17 and 18, we see that reasonable symmetric distribution of pseudorapidity of the τ candidates in both 2015 events only and the combined 2015 and 2016 events. However the transverse momentum of the τ candidates reveals a quite different trend for $p_T(\tau) < 40$GeV. In the 2015 events only cases, the trend are going down, particularly obvious for τ candidates with three charged track but not so much with the one charged track. While in the combined events cases, the trend are going up. As for the fake rate dependence on the average number of interactions per bunch crossing, related to the amount of pile-up in the event, the distributions seem relatively flat in both 2015 events

only and the combined 2015 and 2016 events, indicating that the additional overlaid tracks from pile-up do not have strong influence on the fake rates.

Note that in the 2015 events only cases, the average interactions per bunch crossing is only available up to 20.

3.3.4 Fake Rates in Data



(a) 2015 events only

(b) Combined 2015 and 2016 events

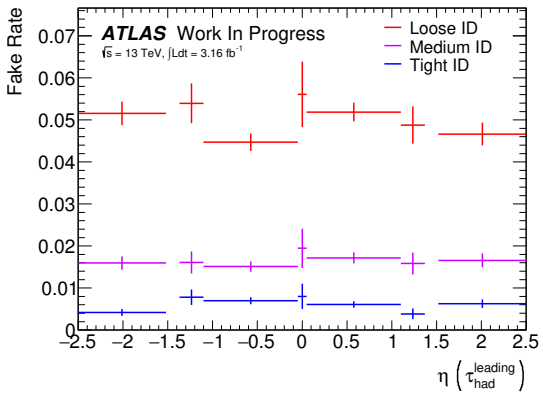
(c) 2015 events only

(d) Combined 2015 and 2016 events

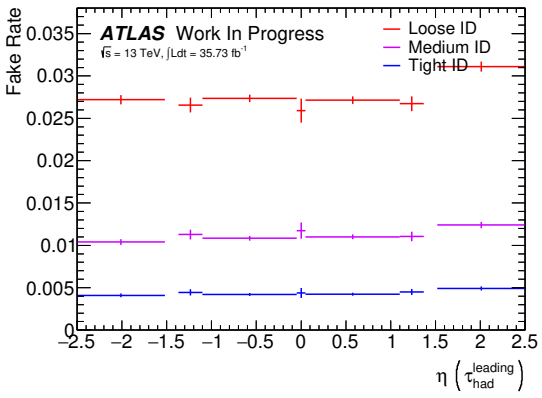
(e) 2015 events only

(f) Combined 2015 and 2016 events

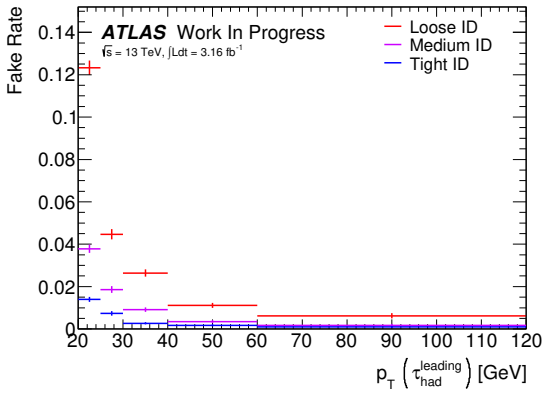
Figure 19: Fake rates determined from $Z \to ee$ data using the tag and probe method for τ candidates with one charged track



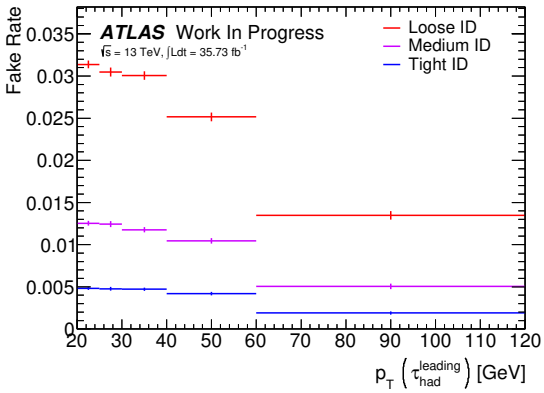
(a) 2015 events only



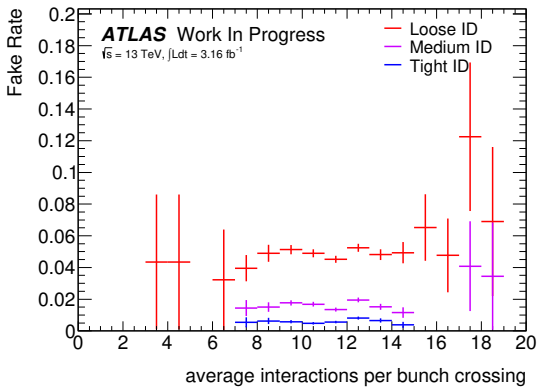
(b) Combined 2015 and 2016 events



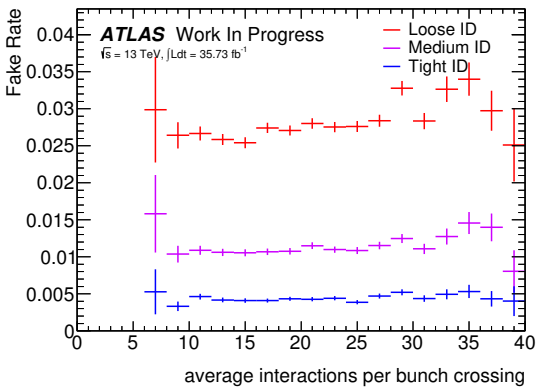
(c) 2015 events only



(d) Combined 2015 and 2016 events



(e) 2015 events only



(f) Combined 2015 and 2016 events

Figure 20: Fake rates determined from $Z \to ee$ data using the tag and probe method for τ candidates with three charged track

From Figure 19 and 20, we see that reasonable symmetric distribution of pseudorapidity of the τ candidates in both 2015 events only and the combined 2015 and 2016 events. However the transverse momentum of the τ candidates reveals a quite different trend for $p_T(\tau) < 40$GeV. In the 2015 events only cases, the trend are clearly going down. While in the combined events cases, the trend are going up. As for the fake rate dependence on the average number of interactions per bunch crossing, related to the amount of pile-up in the event, the distributions seem relatively flat in both 2015 events only and the combined 2015 and 2016 events, indicating that the additional overlaid tracks from pile-up do not

have strong influence on the fake rates.

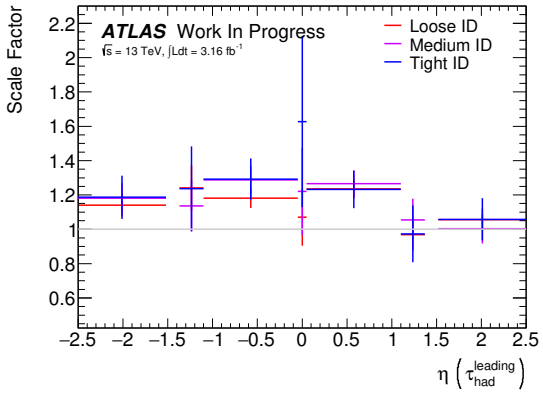
Note that in the 2015 events only cases, the average interactions per bunch crossing is only available up to 20.

3.3.5 Scale Factors for Fake Rates

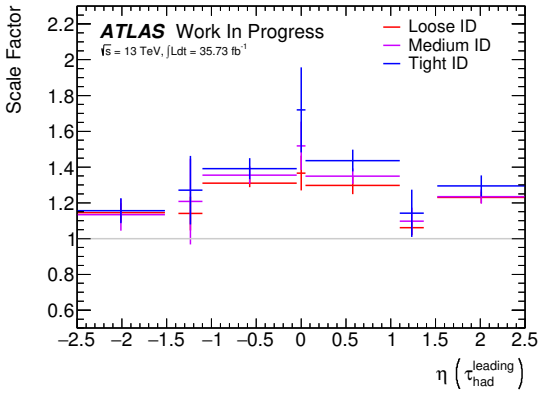
As a metric to compare the fake rates of MC and data, the scale factor s is defined in the following way:

$$s = \frac{FR^{Data}}{FR^{MC}}$$

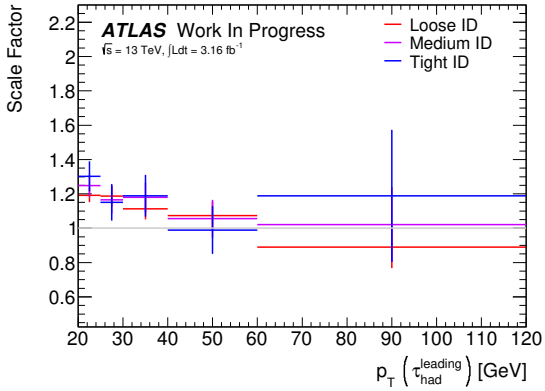
The following figure displays the scale factors calculated from the distributions of Figures 17 to 20. The scale factors are compatible with 1 within 1 or 2 σ in almost all bins of the distributions for 2015 events only cases. However, the distributions of the scale factor differs quite significantly for the combined events cases. The η and the pile-up distributions show that scale factors above 1 dominate in both 2015 events only and the combined events cases.



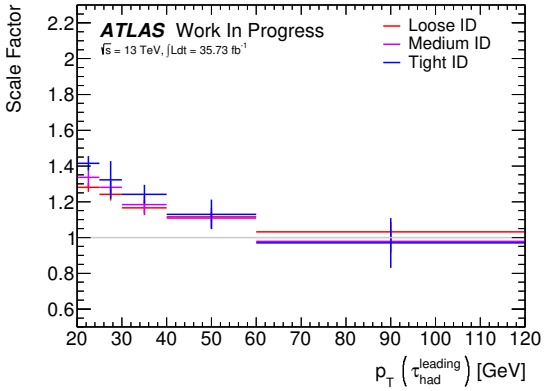
(a) 2015 events only



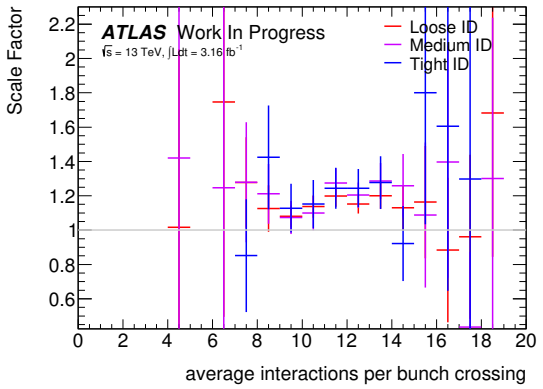
(b) Combined 2015 and 2016 events



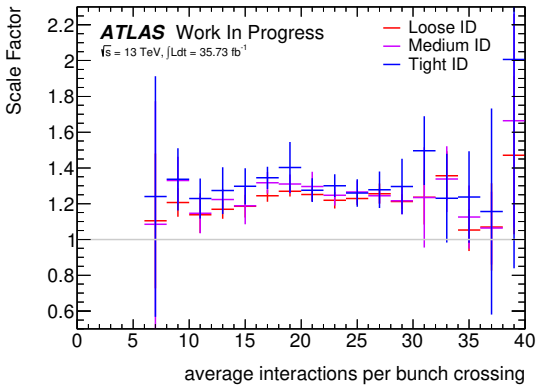
(c) 2015 events only



(d) Combined 2015 and 2016 events

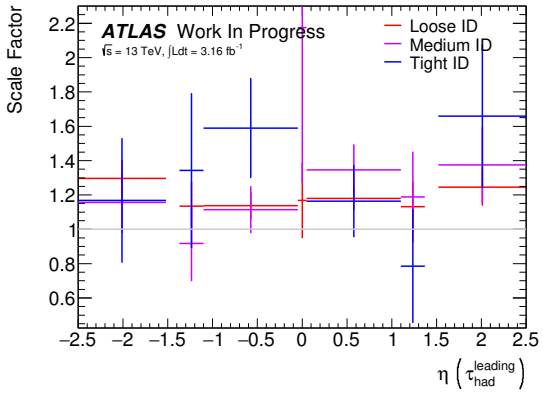


(e) 2015 events only

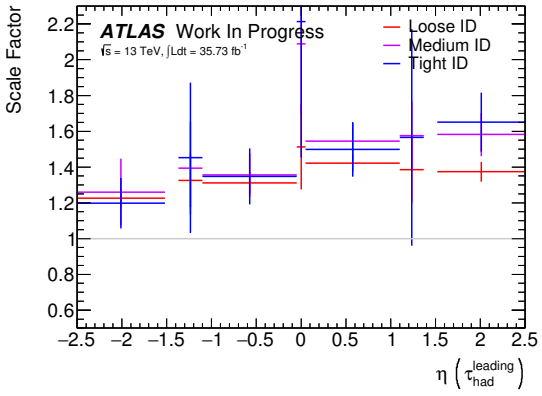


(f) Combined 2015 and 2016 events

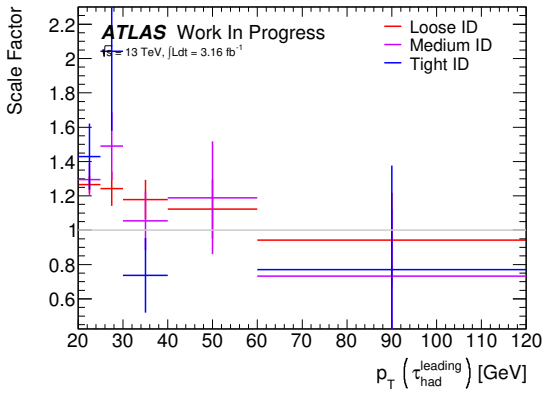
Figure 21: Fake rates determined from $Z \rightarrow ee$ MC sample using the tag and probe method for τ candidates with one charged track



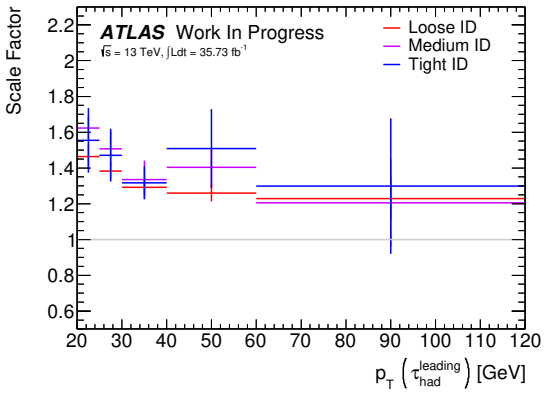
(a) 2015 events only



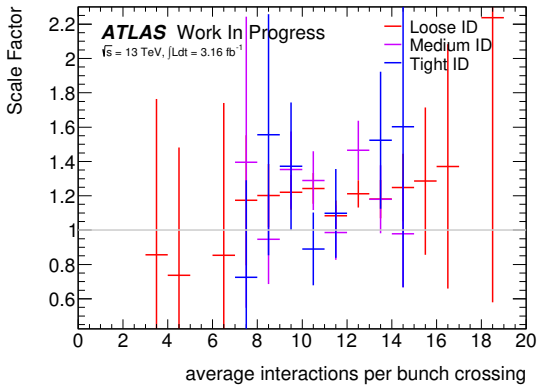
(b) Combined 2015 and 2016 events



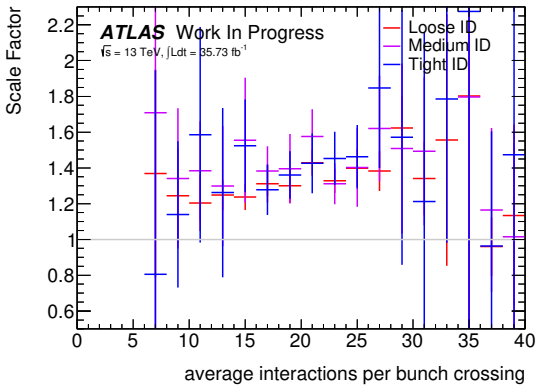
(c) 2015 events only



(d) Combined 2015 and 2016 events



(e) 2015 events only



(f) Combined 2015 and 2016 events

Figure 22: Fake rates determined from $Z \rightarrow ee$ MC sample using the tag and probe method for τ candidates with three charged track

3.4 Extraction of Quark Jet and Gluon Jet Fake Rates

A previous ATLAS study suggests that the differences in the fake rate distribution in different processes is mainly caused by the different ratio of quark to gluon initiated jets [23]. In addition, it is possible to calculate "pure" fake rate distributions FR_q and FR_g of quark and gluon initiated jets from two fake rate measurement FR_i in different regions $i = 1, 2$ with different known quark and gluon fractions q_i and g_i, by solving the linear

system:

$$FR_i = q_i \cdot FR_q + g_i \cdot FR_g$$

Since all jets originate either from a quark or from a gluon, the sum of the quark and gluon fractions is required to be one ($q_i + g_i = 1$). By substituting g_i with $1 - q_i$, we have:

$$FR_i = q_i \cdot FR_q + (1 - q_i) \cdot FR_g$$

Then the measured fake rates FR_1 and FR_2 in regions with different known quark fractions q_1 and q_2 allow the calculation of the pure fake rates FR_q and FR_g:

$$FR_q = \frac{(1 - q_2) \cdot FR_1 - (1 - q_1) \cdot FR_2}{q_1 - q_2} \quad \text{and} \quad FR_g = \frac{q_2 \cdot FR_1 - q_1 \cdot FR_2}{q_2 - q_1}$$

The estimation of uncertainties on these fake rates is derived in Section A.5.

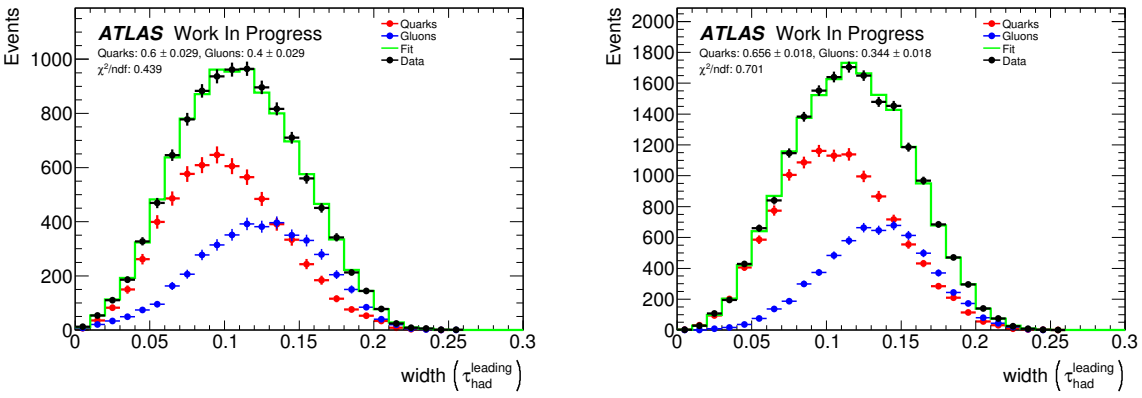
3.4.1 Template Fit

To measure the relative amount of quarks and gluons in the selection, i.e. to determine the values of q_i and g_i, templates of quarks and gluons are obtained from the leading τ candidates in the $Z \to ee$ MC sample by a truth matching algorithm similar to the one referred in Section 3.1.2. The distribution of data with the same event selection criteria is then fitted using these templates.

For the template fit method, a variable is chosen to distinguish the distributions of quarks and gluons. This variable is defined as the weighted average ΔR of all clusters within the jet, where the weights are given by the transverse momenta p_T of the clusters:

$$w = \frac{\sum_i \Delta R^i p_T^i}{\sum_i p_T^i}$$

Note that the ROOT [15] function `TFractionFitter` is used, which implements the fitting method described in [24]. This fitting algorithm takes into account the finite statistics in the MC samples used for the template.



(a) τ candidates with one associated reconstructed track

(b) τ candidates with three associated reconstructed track

Figure 23: Fit of quark and gluon templates to 2015 data only. The templates are obtained from the $Z \to ee$ MC sample using truth matching.

3.4.2 Corrections on Fit Uncertainties

To validate the error estimation on the quark fraction q given by the `TFractionFitter` function, the *pull* of q has been calculated. The pull is defined as follows:

$$\text{Pull}(q_i) = \frac{q_i - \bar{q}}{\sigma_{q_i}}, \tag{1}$$

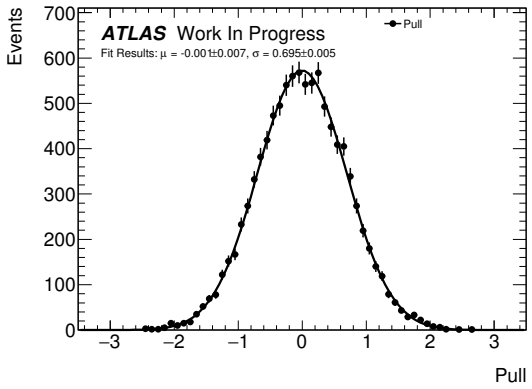
where $\bar{q}$ is the average quark fraction over an ensemble of template fits in 10 000 toy experiments. q_i and σ_{q_i} denote the quark fraction and its uncertainty in a specific experiment out of the 10 000.

For the toy experiments, templates have been extracted from the $Z \to ee$ MC sample as usual, but the fit is performed to a data distribution that is randomly fluctuated in each bin according to the Poisson distribution.

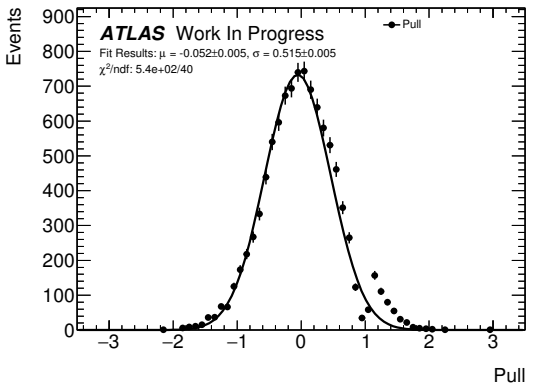
For a correctly estimated error σ_{q_i}, the pull distribution over all toy experiments should yield a Gaussian distribution with a mean value μ of 0 and a standard deviation σ of 1.

The obtained pull distribution for τ candidates with one associated track is shown in Figure 24 and 25 together with a fitted Gaussian. While the shape and mean value matches the expectations, the standard deviation is significantly lower than 1[4]. This corresponds to an *overestimation* of σ_{q_i} by the `TFractionFitter`.

[4]While the exact value of the fitted σ varies between different selections, it is consistently below 1.

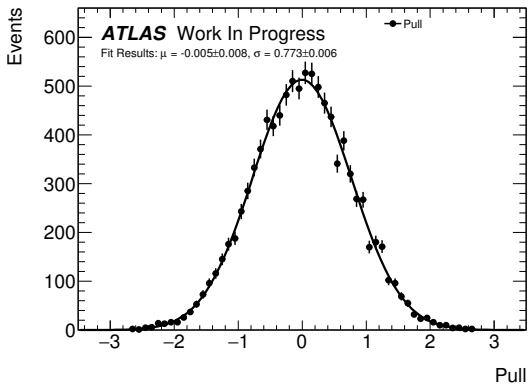


(a) 2015 events only

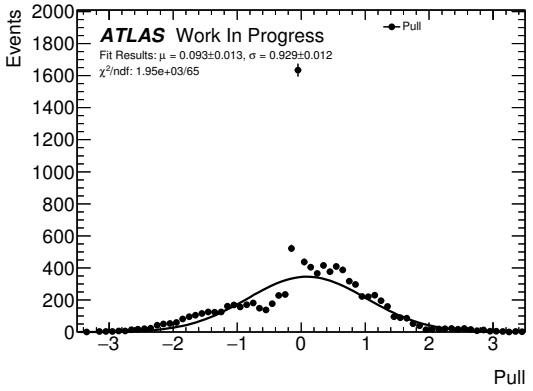


(b) combined 2015 and 2016 events

Figure 24: Pull of the quark fraction given by the `TFractionFitter` for τ candidates with one associated track before correction



(a) 2015 events only



(b) combined 2015 and 2016 events

Figure 25: Pull of the quark fraction given by the `TFractionFitter` for τ candidates with three associated track before correction

To correct this behaviour, a pull distribution is calculated for every performed template fit and the σ obtained from a Gaussian fit to the pull distribution is multiplied onto the σ_q given by the `TFractionFitter`.

3.4.3 Template Fit Results

#Tracks	p_T (τ candidate)	Fitted Quark Fraction	Truth Matching
1	20 - 25 GeV	$0.64 \pm 0.039 \pm 0.15$	$0.40 \pm 0.00005 \pm 0.32$
1	25 - 30 GeV	$0.67 \pm 0.047 \pm 0.17$	$0.53 \pm 0.00011 \pm 0.26$
1	30 - 40 GeV	$0.822 \pm 0.037 \pm 0.073$	$0.67 \pm 0.00012 \pm 0.18$
1	40 - 60 GeV	$0.871 \pm 0.027 \pm 0.018$	$0.816 \pm 0.00017 \pm 0.095$
1	60 - 120 GeV	$0.893 \pm 0.017 \pm 0.024$	$0.916 \pm 0.00033 \pm 0.051$
3	20 - 25 GeV	$0.906 \pm 0.038 \pm 0.012$	$0.445 \pm 0.00005 \pm 0.219$
3	25 - 30 GeV	$0.902 \pm 0.037 \pm 0.022$	$0.551 \pm 0.00007 \pm 0.175$
3	30 - 40 GeV	$0.854 \pm 0.027 \pm 0.026$	$0.670 \pm 0.00005 \pm 0.118$
3	40 - 60 GeV	$0.929 \pm 0.018 \pm 0.008$	$0.802 \pm 0.00006 \pm 0.057$
3	60 - 120 GeV	$0.913 \pm 0.012 \pm 0.026$	$0.914 \pm 0.00007 \pm 0.016$

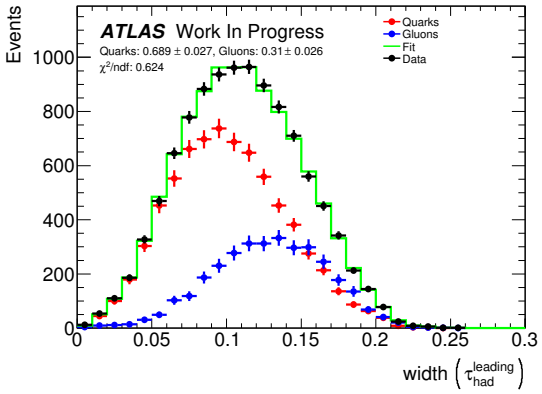
Table 4: Quark fraction q±stat.±syst. in the full $p_T(\ell\ell)$ region as determined by a template fit on 2015 data or from truth matching on the $Z \to ee$ MC sample. The given systematic uncertainties origin from the unmatched τ candidates.

#Tracks	p_T (τ candidate) Region	Fitted Quark Fraction	Truth Matching
1	20 - 25 GeV	$0.705 \pm 0.059 \pm 0.164$	$0.26 \pm 0.00008 \pm 0.46$
1	25 - 30 GeV	$0.523 \pm 0.061 \pm 0.108$	$0.28 \pm 0.00023 \pm 0.50$
1	30 - 40 GeV	$1.000 \pm 0.035 \pm 0.000$	$0.29 \pm 0.00043 \pm 0.48$
1	40 - 60 GeV	$1.000 \pm 0.059 \pm 1.000$	$0.25 \pm 0.0016 \pm 0.57$
1	60 - 120 GeV	$1.000 \pm 0.011 \pm 0.000$	$0.10 \pm 0.0075 \pm 0.53$
3	20 - 25 GeV	$0.801 \pm 0.040 \pm 0.016$	$0.32 \pm 0.00009 \pm 0.38$
3	25 - 30 GeV	$0.642 \pm 0.044 \pm 0.026$	$0.34 \pm 0.00017 \pm 0.39$
3	30 - 40 GeV	$0.721 \pm 0.052 \pm 0.056$	$0.36 \pm 0.00025 \pm 0.39$
3	40 - 60 GeV	$1.000 \pm 0.055 \pm 1.000$	$0.31 \pm 0.00078 \pm 0.43$
3	60 - 120 GeV	$1.000 \pm 0.022 \pm 0.580$	$0.24 \pm 0.0059 \pm 0.47$

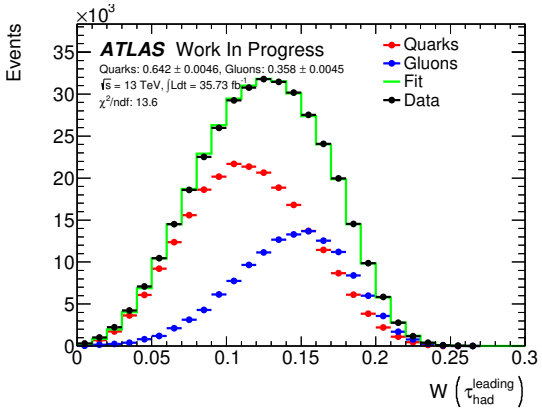
Table 5: Quark fraction q±stat.±syst. in the $p_T(\ell\ell) < 25$ GeV region as determined by a template fit on 2015 data or from truth matching on the $Z \to ee$ MC sample. The given systematic uncertainties origin from the unmatched τ candidates.

#Tracks	p_T (τ candidate) Region	Fitted Quark Fraction	Truth Matching
1	20 - 25 GeV	$0.634 \pm 0.041 \pm 0.076$	$0.60 \pm 0.00012 \pm 0.13$
1	25 - 30 GeV	$0.748 \pm 0.064 \pm 0.103$	$0.717 \pm 0.00017 \pm 0.083$
1	30 - 40 GeV	$0.874 \pm 0.036 \pm 0.026$	$0.802 \pm 0.00013 \pm 0.056$
1	40 - 60 GeV	$0.834 \pm 0.022 \pm 0.013$	$0.891 \pm 0.00016 \pm 0.024$
1	60 - 120 GeV	$0.914 \pm 0.015 \pm 0.003$	$0.955 \pm 0.00025 \pm 0.014$
3	20 - 25 GeV	$0.928 \pm 0.029 \pm 0.218$	$0.574 \pm 0.00010 \pm 0.073$
3	25 - 30 GeV	$0.991 \pm 0.052 \pm 0.140$	$0.677 \pm 0.00010 \pm 0.049$
3	30 - 40 GeV	$0.907 \pm 0.026 \pm 0.007$	$0.759 \pm 0.00006 \pm 0.034$
3	40 - 60 GeV	$0.944 \pm 0.018 \pm 0.097$	$0.848 \pm 0.00006 \pm 0.016$
3	60 - 120 GeV	$0.877 \pm 0.017 \pm 0.040$	$0.928 \pm 0.00007 \pm 0.005$

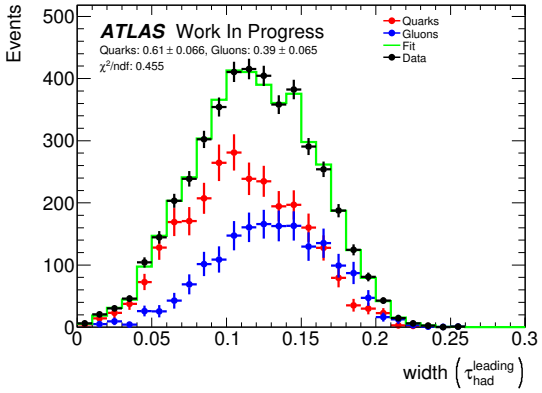
Table 6: Quark fraction $q \pm$stat.$\pm$syst. in the $p_T(\ell\ell) > 25\,\text{GeV}$ region as determined by a template fit on 2015 data or from truth matching on the $Z \to ee$ MC sample. The given systematic uncertainties origin from the unmatched τ candidates.



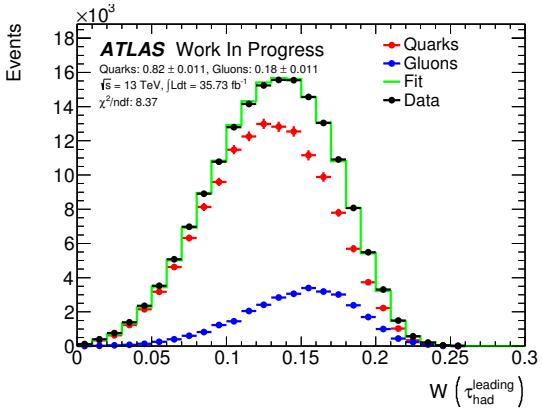
(a) Entire $p_T(\ell\ell)$ region, 2015 events only



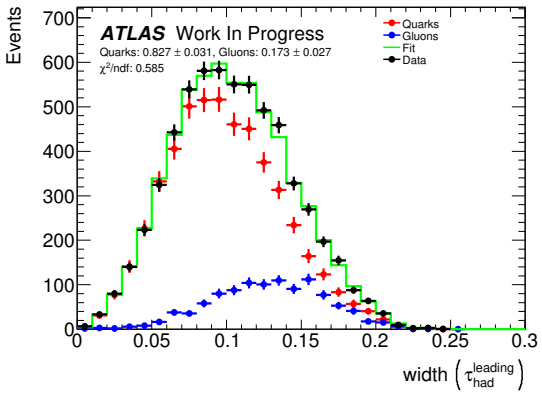
(b) Entire $p_T(\ell\ell)$ region, combined 2015 and 2016 events



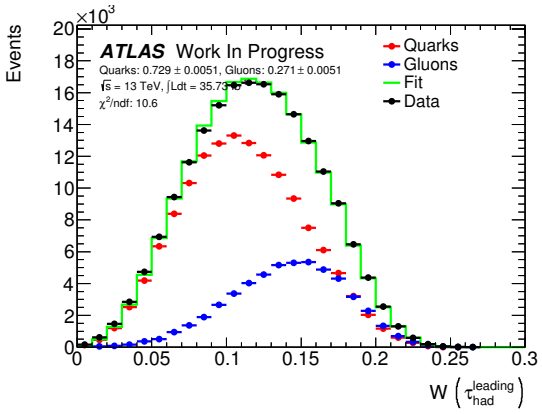
(c) Low $p_T(\ell\ell)$ region, 2015 events only



(d) Low $p_T(\ell\ell)$ region, combined 2015 and 2016 events

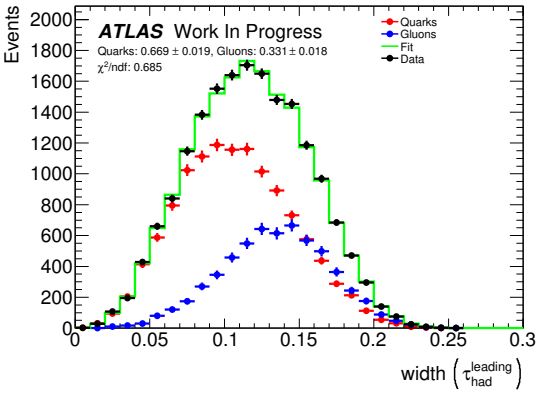


(e) High $p_T(\ell\ell)$ region, 2015 events only

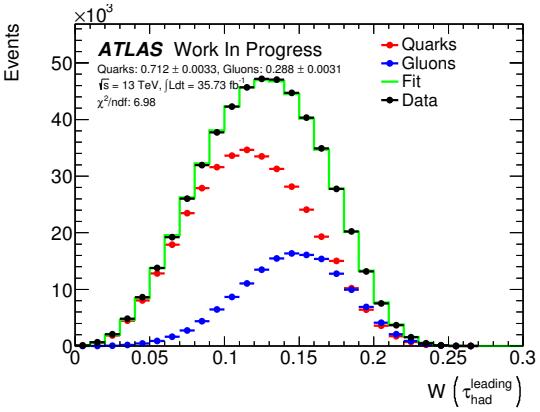


(f) High $p_T(\ell\ell)$ region, combined 2015 and 2016 events

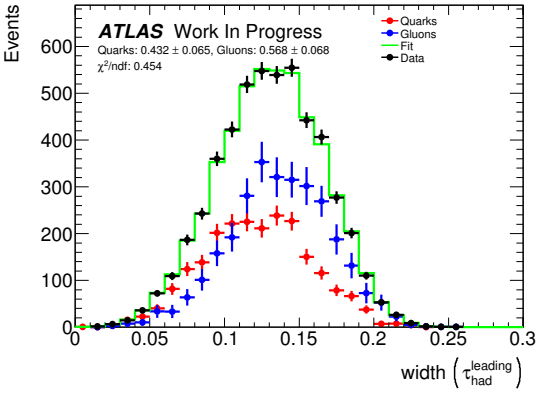
Figure 26: Comparison of fit of the quark and gluon templates in different regions of $p_T(\ell\ell)$ for τ candidates with one associated track. The templates are obtained from the $Z \to ee$ MC sample using truth matching.



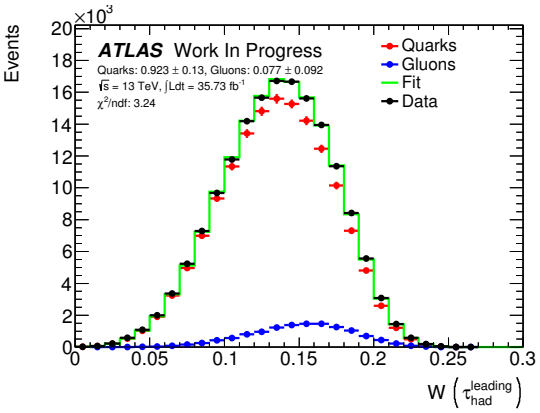
(a) Entire $p_T(\ell\ell)$ region, 2015 events only



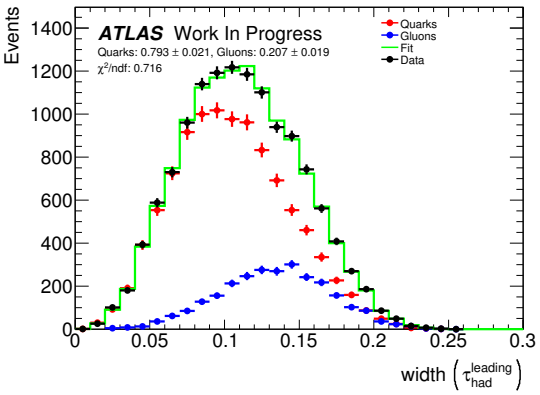
(b) Entire $p_T(\ell\ell)$ region, combined 2015 and 2016 events



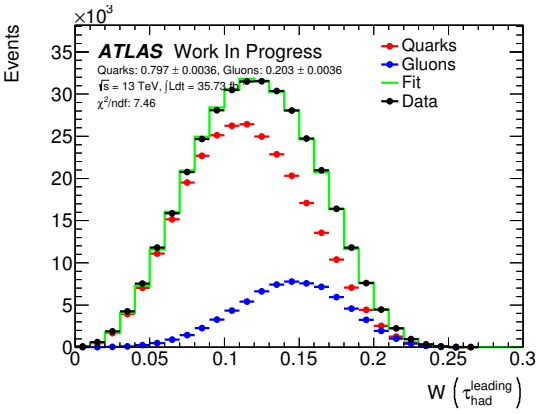
(c) Low $p_T(\ell\ell)$ region, 2015 events only



(d) Low $p_T(\ell\ell)$ region, combined 2015 and 2016 events



(e) High $p_T(\ell\ell)$ region, 2015 events only



(f) High $p_T(\ell\ell)$ region, combined 2015 and 2016 events

Figure 27: Comparison of fit of the quark and gluon templates in different regions of $p_T(\ell\ell)$ for τ candidates with three associated track. The templates are obtained from the $Z \to ee$ MC sample using truth matching.

4 Conclusion

In this research, a set of analysis is conducted based on both 2015 and 2016 data collected by the ATLAS experiment. The fake rates of hadronically decaying Tau leptons are measured from the tagged $Z \rightarrow ee$ MC samples with respect to transverse momentum (p_T), pseudorapidity (η), and the amount of pile-up in the events.

From the scale factors (Section 3.3.5) determined from a $Z \rightarrow ee$ MC sample, for the τ candidates with one charged track, we see an indication of MC mismodeling in low p_T regions, as the scale factors are above one in low p_T regions with an increasing trend for lower p_T values (< 40 GeV). For the τ candidates with three charged tracks, we see that there is also a mismodeling of the MC, but instead of only low p_T regions, the high p_T regions also suggests MC mismodeling, as the scale factors are above one. In fact, for the τ candidates with three charged track, the scale factors across all p_T regions are at least 20% greater than 1.

As a 2011 study [23] suggests, this mismodeling of the fake rate in MC is due to the mismodeling of the fraction of quark/gluon induced jets in the events. A template fit method has been applied to the data to extract the quark/gluon fraction. This measured fraction was compared to the fraction obtained from MC using a truth matching algorithm.

A separation of the data into two $p_T(ll)$ regions has been optimized for a separate measurement of fake rates and quark/gluon fractions in both regions. The results of these measurements in different regions are yet to be computed to obtain pure quark and gluon jet fake rates (Section 3.4).

The 2015 ATLAS data used in this project corresponds to about 3.2 fb^{-1}. While the combined 2016 and 2016 ATLAS data also used in this project corresponds to about 35.7 fb^{-1}, which is roughly more than ten times as much as the 2015 data. With the 2016 data, the statistical uncertainties of the fake rates measured in this project are reduced by more than a factor of 3 (Section 1.1).

This project still has areas for further investigation. For example, the unmatched candidates in the $Z \rightarrow ee$ MC sample in template fit (Section 3.4.1) are considered as a systematic uncertainty on the template fit since the distribution of the template fit variable of these candidates is highly similar to the distribution of the template fit variable of gluons. This is also the reason that we can rule out the suggestion that these unmatched candidates are the pile-ups in the MC. However the true reason that these unmatched candidates exist remains unclear. One hypothesis is that the source is in the upper stream in the data format work flow (Figure 6). In addition to using $Z \rightarrow ee$ process as tagged region, similar processes like $Z \rightarrow ee$ can also be used as tagged region to further increase the data available for analysis. The use of two independent physics processes for the fake rate measurements, which have different quark/gluon fractions, can further reduce the uncertainties on the extracted quark and gluon fake rates.

If our hypothesis holds true, that the tau fake rate distribution in a physics process depends solely on the quark/gluon fraction in this process, it will be possible to distribute

the p_T-binned (and ideally η and $< \mu >$-binned) pure quark and gluon fake rates along with a tool (e.g. a template fitter) to measure the quark fraction in a given selection. The fake rates could then be mixed into the expected total fake rate distribution for any physics process, using the quark/gluon fractions. This would remove the need to estimate the fake contribution in an individual way for every analysis involving hadronically decaying tau leptons. [7]

A Appendix

A.1 Poisson Distribution

The Poisson distribution applies when: (1) the event is something that can be counted in whole numbers; (2) occurrences are independent, so that one occurrence neither diminishes nor increases the chance of another; (3) the average frequency of occurrence for the time period in question is known; and (4) it is possible to count how many events have occurred. [25]

Definition. Let $0 \leq \lambda$. A random variable X has the Poisson distribution with parameter λ if X is $\mathbb{Z}_+$-valued and has the probability mass function

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!} \quad \text{for} \quad k \in \mathbb{Z}_+$$

Proof. The mean

$$E[X] = \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^k}{(k-1)!} = \lambda \sum_{k=1}^{\infty} \frac{e^{-\lambda}\lambda^{k-1}}{(k-1)!} = \lambda \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!} = \lambda$$

To derive the variance, we first derive

$$E[X(X-1)] = \sum_{k=0}^{\infty} k(k-1)\frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=2}^{\infty} = \frac{e^{-\lambda}\lambda^k}{(k-2)!} = \lambda^2 \sum_{j=0}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!} = \lambda^2$$

Now the variance

$$Var(X) = E[X^2] - (E[X])^2 = E[X(X-1)] + E(X) - (E[X])^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

The standard deviation (standard error) is then

$$\sigma = \sqrt{Var(X)} = \sqrt{\lambda}$$

A.2 Remote Login Procedure

A.2.1 Local Computing Cluster

One option is to run jobs on one of the servers located right at the II Institue of Physics of the University of Göttingen. This server is referred as pcatlas 36.

The login Bash script from local Bash shell (terminal on Mac OS or Linux OS) is the following:

```
ssh -Y your_user_name@login.ph2.physik.uni-goettingen.de -p24
ssh -Y pcatlas36
```

where `-p24` means login through port 24 and `-Y` activates the remote window forwarding option.

A.2.2 National Analysis Facility

The other option is to run jobs on the servers of the German National Analysis (NAF) Facility hosted by the German Electron Synchrotron at Hamburg. NAF supports batch job submission. This feature massively accelerates the progress of the analysis and is therefore the reason that NAF is chosen in this research project.

The login Bash script from local is the following:

```
ssh -Y your_user_name@naf-atlas.desy.de
```

To submit batch jobs, one need to write jobs in bash scripts and submit each job with `qsub abc.sh` command, where `abc.sh` is the script of the intended job.

To kill a job, the command is `qdel abc.sh`.

To fast login, one can also create the following file in `~/.ssh/config`, in which one can add

```
Host your_login_alias
    HostName naf-atlas.desy.de
    User your_user_name
    ForwardX11 yes
```

And to login, just type the following in the Bash shell

```
ssh your_login_alias
```

A.3 Binomial Errors for Fake Rates

In general, the error propagation for a ratio $r = a/b$, where both a and b are fluctuating, is given as:

$$\sigma_r^2 = \left(\frac{\partial r}{\partial b}\right)^2 \sigma_b^2 + \left(\frac{\partial r}{\partial a}\right)^2 \sigma_a^2 + 2\frac{\partial r}{\partial b}\frac{\partial r}{\partial a}\mathrm{cov}(a,b). \tag{2}$$

For a fake rate $FR = k/n$, where k is the number of events in a subset of the set of n events, the covariance term becomes $\mathrm{cov}(n,k) = FR \cdot n = \sigma_k^2$ [26]. With this, the following formula can be obtained:

$$\sigma_{FR}^2 = \frac{(1 - 2 \cdot \frac{k}{n}) \cdot \sigma_k^2 + (\frac{k}{n})^2 \cdot \sigma_n^2}{n^2}, \tag{3}$$

which is the formula used in the ROOT [15] function `TH1::Divide()`, when the option `"B"` (for binomial errors) is specified.

In the case of Poisson uncertainties $\sigma_k = \sqrt{k}$ and $\sigma_n = \sqrt{n}$, Equation 3 can be simplified to the usual form of a binomial error estimation:

$$\sigma_{FR}^2 = \frac{\frac{k}{n}(1 - \frac{k}{n})}{n^2} \tag{4}$$

A.4 Error Propagation for Scale Factors

For the scale factor $s = FR^{Data}/FR^{MC}$ between the fake rate in data and the MC fake rate, the usual error propagation (Equation 2) can be applied. Since the two fake rate measurements are independent from each other, the covariance term vanishes and the uncertainty of the scale factor is given by:

$$\sigma_s^2 = \left(\frac{1}{FR^{MC}} \cdot \sigma_{FR^{Data}}\right)^2 + \left(\frac{FR^{Data}}{(FR^{MC})^2} \cdot \sigma_{FR^{MC}}\right)^2 \tag{5}$$

A.5 Error Propagation for Extracted Quark-/Gluon Fake Rates

As discussed in Section 3.4, the fake rate of quark or gluon initiated jets FR_q or FR_g can be estimated from two fake rates FR_i ($i = 1,2$) measured in selections with different fractions of quark initiated jets q_i:

$$FR_q = \frac{(1-q_2)\cdot FR_1 - (1-q_1)\cdot FR_2}{q_1 - q_2} \qquad \text{and} \qquad FR_g = \frac{q_2\cdot FR_1 - q_1\cdot FR_2}{q_2 - q_1} \tag{6}$$

The uncertainty on this fake rate can be estimated by propagating the errors of the measured fractions and fake rates:

$$\sigma^2_{FR_x} = \sum_i \left(\frac{\partial FR_x}{\partial FR_i}\cdot\sigma_{FR_i}\right)^2 + \left(\frac{\partial FR_x}{\partial q_i}\cdot\sigma_{q_i}\right)^2, \tag{7}$$

where $x = q, g$. The necessary differential derivations are given by:

$$\frac{\partial FR_q}{\partial FR_1} = \frac{1-q_2}{q_1-q_2}, \qquad \frac{\partial FR_q}{\partial FR_2} = \frac{q_1-1}{q_1-q_2}, \tag{8}$$

$$\frac{\partial FR_q}{\partial q_1} = (q_2-1)\cdot\frac{FR_1-FR_2}{(q_1-q_2)^2}, \qquad \frac{\partial FR_q}{\partial q_2} = (q_1-1)\cdot\frac{FR_1-FR_2}{(q_1-q_2)^2}, \tag{9}$$

$$\frac{\partial FR_g}{\partial FR_1} = \frac{q_2}{q_2-q_1}, \qquad \frac{\partial FR_g}{\partial FR_2} = \frac{-q_1}{q_2-q_1}, \tag{10}$$

$$\frac{\partial FR_g}{\partial q_1} = q_2\cdot\frac{FR_1-FR_2}{(q_2-q_1)^2}, \qquad \frac{\partial FR_g}{\partial q_2} = -q_1\cdot\frac{FR_1-FR_2}{(q_2-q_1)^2}, \tag{11}$$

References

[1] ATLAS and CMS Collaborations. *Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s}$ = 7 and 8 TeV*. ATLAS-CONF-2015-044, CMS-PAS-HIG-15-002.

[2] CERN. *The Large Hadron Collider*. URL: http://home.cern/topics/large-hadron-collider.

[3] Oliver Sim Brüning et al. *LHC Design Report*. Geneva: CERN, 2004.

[4] ATLAS Collaboration. "The ATLAS Experiment at the CERN Large Hadron Collider". In: *JINST* 3 (2008), S08003.

[5] CERN. *Detector & Technology*. URL: https://atlas.cern/discover/detector (visited on 08/21/2017).

[6] Matthias Schott and Monica Dunford. "Review of single vector boson production in pp collisions at $\sqrt{s}$ = 7 TeV". In: *Eur. Phys. J.* C74 (2014), p. 2916. DOI: 10.1140/epjc/s10052-014-2916-1. arXiv: 1405.1160 [hep-ex].

[7] Timo Dreyer. "Measurement of the Fake Rate for Hadronic Tau Lepton Decays using the ATLAS Experiment". Master's Thesis. II. Institue of Physics, University of Göttingen, Oct. 12, 2016.

[8] CERN. *The Inner Detector*. URL: https://atlas.cern/discover/detector/inner-detector (visited on 08/21/2017).

[9] CERN. *Calorimeter*. URL: https://atlas.cern/discover/detector/calorimeter (visited on 08/21/2017).

[10] CERN. *Muon Spectrometer*. URL: https://atlas.cern/discover/detector/muon-spectrometer (visited on 08/21/2017).

[11] CERN. *Magnet System*. URL: https://atlas.cern/discover/detector/magnet-system (visited on 08/21/2017).

[12] Aranzazu Ruiz-Martinez and ATLAS Collaboration. *The Run-2 ATLAS Trigger System*. Tech. rep. ATL-DAQ-PROC-2016-003. Geneva: CERN, Feb. 2016. URL: http://cds.cern.ch/record/2133909.

[13] CERN. *Software and Computing*. URL: https://atlas.cern/discover/detector/software-computing (visited on 08/21/2017).

[14] K. A. Olive et al. "2017 Review of Particle Physics". In: *Chin. Phys.* C38 (2016), p. 090001.

[15] R. Brun and F. Rademakers. "ROOT: An object oriented data analysis framework". In: *Nucl.Instrum.Meth.* A389 (1997), pp. 81–86.

[16] CERN. *Rucio 1.2.5-1 documentation*. URL: http://rucio.cern.ch/ (visited on 08/22/2017).

[17] P. Nason S. Frixione and C. Oleari. "Matching NLO QCD computations with Parton Shower simulations: the POWHEG method". In: *JHEP* 0711 (2007), p. 070.

[18] S. Mrenna T. Sjostrand and P. Z. Skands. "A brief introduction to PYTHIA 8.1". In: *Comput. Phys. Commun.* 178 (2008), p. 852.

[19] ATLAS Collaboration. "Measurement of the Z/γ^* boson transverse momentum distribution in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector". In: *JHEP* 1409 (2014), p. 145.

[20] J Pumplin et al. "New Generation of Parton Distributions with Uncertainties from Global QCD Analysis". In: *JHEP* 07 (2002), p. 012.

[21] Z Marshall. *Simulation of Pile-up in the ATLAS Experiment*. Tech. rep. ATL-SOFT-PROC-2013-030. Geneva: CERN, Oct. 2013. URL: https://cds.cern.ch/record/1616394.

[22] ATLAS Collaboration. *Determination of the tau energy scale and the associated systematic uncertainty in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector at the LHC in 2012*. ATLAS-CONF-2013-044. 2013.

[23] ATLAS Collaboration. *Measurement of the Mis-identification Probability of τ Leptons from Hadronic Jets and from Electrons*. ATLAS-CONF-2011-113.

[24] Roger Barlow and Christine Beeston. "Fitting using finite Monte Carlo samples". In: *Comp. Phys. Comm.* 77 (1993), pp. 219–228.

[25] University of Massachusetts Amherst. *Statistics. The Poisson Distribution*. Aug. 2007. URL: https://www.umass.edu/wsp/resources/poisson/ (visited on 08/28/2017).

[26] G. Ranucci. "Binomial and ratio-of-Poisson-means frequentist confidence intervals applied to the error evaluation of cut efficiencies". In: (Jan. 2009). arXiv: 0901.4845 [physics.data-an].

Acknowledgement

I would like to thank the German Academic Exchange Service Research in Science and Engineering (DAAD RISE) program for fully funding my research at the University of Göttingen.

I would also like to thank my supervisor Timo Dreyer for his excellent explanation, patience, and help along the way in both the physics and coding parts. The overall framework of this research is derived from the work in his Master's thesis. Timo and I have been working very closely over this summer to revamp the old plotting framework. The ntuples, as shown in Figure 6, are re-derived entirely with the combined 2015 and 2016 data from the `xTauFramework`. Numerous bug fixes are done, including removing a duplicated systematic uncertainties setting to achieve a much better reweighting, removing the runs that are not in the good run number list for the 2015 data, replacing the ineffective truth-matching algorithms, and rewriting the `histogramStore` option to ensure the histograms stored for accelerated reproduction from different plotting scripts do not interfere from each other. The luminosity of the combined data of 2015 and 2016 is recalculated. The names of numerous variables in the plotting framework are updated to match the latest trigger names in event selection and variable names in the core plotting library written by Christian Grefe. New control flags are introduced and some deprecated ones are removed. A new set of debugging tools are also created along the way including the script like `Cutflow.py`. It has been an intense debugging process with lots of trials and failures but we are very glad that we are able to fix all the problems in the end. I have learned tremendously both in research and in life. It was a fantastic experience in the incredible city of Göttingen, where all the history and famous figures in physics and mathematics intertwined together.

In addition, I would also like to thank Prof. Stan Lai and Dr. Michel Janus for their help and support in the bi-weekly research chats. It has been a great experience participating the weekly ATLAS meeting and colloquium and giving updates in the weekly Higgs meeting. This research could not have been successfully conducted with all your help and support.