Universität Bonn

Physikalisches Institut

Data/Monte-Carlo Comparison for Tau Lepton Reconstruction with Decay Mode Identification in the ATLAS Detector

Robert Beckmann

Dieser Forschungsbericht wurde als Masterarbeit von der Mathematisch-Naturwissenschaftlichen Fakultät der Universität Bonn angenommen.

Angenommen am:	20.10.2014
1. Gutachter:	Prof. Dr. Klaus Desch
2. Gutachter:	Prof. Dr. Jochen Dingfelder

Acknowledgements

In the first place I would like to thank Peter Wagner for his help and support during the year of this work and Christian Limbach for the answers to technical questions, explanations related to the reconstruction algorithms and most importantly motivating humour. I do also thank Till Nattermann and Steffen Schaepe for the help in setting up the comparison. Special thanks go to Klaus Desch for the opportunity to write this thesis in his working group and to Philip Bechtle. Their advices were a great help that guided me through the research phase of this study.

In the twilight zone between physics and real life, I have to thank Nicklas Denis for all kinds of discussions and counsel. Most of all, I thank Katrin Weiser and Mogli. They are the greatest support I can imagine.

Contents

1 Introduction **1**

2 The ATLAS detector at the Large Hadron Collider **3**
2.1 The Large Hadron Collider . . . 3
2.2 The ATLAS Experiment . . . 3
2.2.1 Coordinate System at ATLAS . . . 3
2.2.2 Inner Detector . . . 4
2.2.3 Calorimeter System . . . 4
2.2.4 Muon Spectrometer . . . 5
2.2.5 Luminosity, Trigger and Pile Up . . . 6

3 Reconstruction and Identification of Tau Leptons **9**
3.1 Tau Decays in ATLAS . . . 9
3.2 Reconstruction and Identification of Taus in ATLAS . . . 10
3.2.1 Reconstruction of Taus with tauRec . . . 11
3.2.2 Separating Taus from Jets . . . 11
3.3 Reconstruction of Decay Particles with cell-based . . . 12
3.4 Recovering Misclassified with PanTau . . . 14

4 Data/Monte-Carlo Comparison **17**
4.1 Experimental and Simulated Datasets . . . 17
4.2 Corrections Applied to Monte Carlo . . . 21
4.3 Selecting $Z \rightarrow \mu\mu$ Events . . . 22
4.4 Evaluation via χ^2 Test . . . 23
4.5 Systematic Uncertainties . . . 28
4.6 Results . . . 29
4.6.1 PanTau BDT Variables . . . 31
4.6.2 Cell-based π^0 ID BDT Variables . . . 35
4.6.3 Tau ID Variables . . . 38

5 Summary and Conclusion **41**

A Appendix **43**
A.1 PanTau . . . 43
A.2 Cell-based . . . 57
A.3 Tau ID Variables . . . 65

Bibliography **73**

List of Figures **75**

List of Tables **81**

CHAPTER 1

Introduction

In the first run of the Large Hadron Collider (LHC) from 2010 to 2012, protons have been collided at energies up to 8 TeV. The two major experiments at the LHC (ATLAS and CMS) have continuously recorded and analysed the particles created in such collisions. A major achievement of this research was the discovery of a Higgs-like particle in 2012. For the proposition of the Higgs particle 50 years ago, Peter Higgs [1] and François Englert [2] have been awarded the Nobel Prize in Physics in 2013. This discovery concludes the last gap in the current theory of particle physics, the *Standard Model*. The $H \rightarrow \tau\tau$ channel is one of the decay channels, where the Higgs boson has been discovered. Due to the heavy mass of tau leptons, the coupling to the Higgs boson is stronger than for the lighter leptons. This channel is specifically interesting for the measurement of the CP eigenstate of the Higgs boson, which can be done with tau polarisation studies.

However, it is already known that the standard model has some shortcomings. There are several approaches in theoretical particle physics, that try to fill the gaps in that model – one of them is *supersymmetry*. Physics, that is not explained in the standard model, is typically referred to as *physics beyond the standard model*. The tau lepton is the heaviest lepton currently known and is again of special interest in these research areas. Many searches for supersymmetry include tau leptons and therefore benefit from an excellent reconstruction of these particles. A barrier for this is the short lifetime of the tau lepton, it decays before it can be detected. To reconstruct a tau lepton therefore means to reconstruct its decay particles. In the detector, they form a collimated particle jet – a tau jet.

In ATLAS, a new reconstruction algorithm chain has been developed, that follows a particle flow approach. That means, single particles are identified and followed throughout the detector. The benefit of that is a highly improved energy resolution and the identification of the decay modes of the tau lepton. While the improved resolution is beneficial for all kinds of tau analyses, the decay mode classification is especially important for polarisation studies.

The algorithms have been developed using simulated events. It is important to check whether the assumptions made in these simulations do apply to experimental data. In this study a Data/Monte-Carlo

comparison is presented. Variables used in the new tau reconstruction are compared in events with QCD[1] jets. These are quark or gluon initiated jets. They feature the same constituent particles as tau jets, but are available in much higher statistics. Also, QCD jets represent the major background for tau identification.

[1] Quantum Chromodynamics

CHAPTER 2

The ATLAS detector at the Large Hadron Collider

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is a particle collider of the European Organization for Nuclear Research (CERN). It is built in the reused tunnel of the former Large Electron-Positron Collider at the Franco-Swiss border near Geneva. This thesis deals with data taken from the LHC in 2012 at a centre of mass energy of $\sqrt{s} = 8\,\text{TeV}$.

2.2 The ATLAS Experiment

The ATLAS experiment is one of seven experiments at the LHC. It is a general purpose detector that aims at a wide range of particle physics measurements. The detector basically consists of three different detection systems that are described in the following sections [4].

2.2.1 Coordinate System at ATLAS

The coordinate system used at ATLAS has its origin in the centre of the detector. The x axis points towards the centre of the LHC, the y axis points upwards and the z axis points along the beam line towards the LHCB experiment. Since the detector is cylindrical, conveniently cylindrical coordinates are used. The azimuthal angle ϕ is defined in a range of $]-\pi;\pi]$ as the angle to the x axis in the transverse (x-y) plane, where positive values are towards the positive y axis. The polar angle θ is defined in the range $[0;\pi]$ as the angle to the z axis. Mostly, θ is replaced by the pseudorapidity $\eta = -\ln(\tan(\theta/2))$, because differences in η are approximately invariant under a longitudinal boost. In the η-ϕ space, distances are measured with $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$.

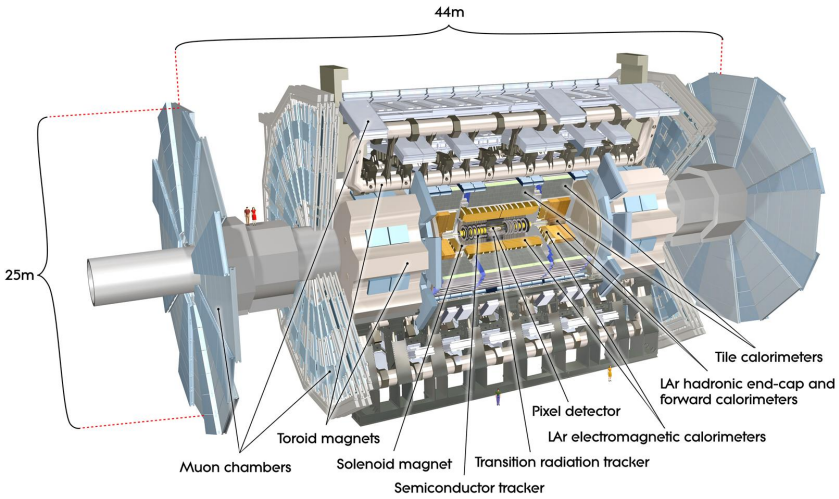


Figure 2.1: The ATLAS detector [3]

2.2.2 Inner Detector

The inner detector is the innermost part of the ATLAS detector. It is responsible for the tracking of charged particles which contributes to momentum measurement, particle identification and vertex reconstruction. This system itself consists of three distinct detectors. Going from the beam spot outwards, the first one is a silicon based pixel detector. Then comes a silicon based strip detector. The last one is the transition radiation tracker (TRT), which consists of straw tubes that are filled with a xenon gas mixture. Besides tracking of charged particles, the TRT also detects transition radiation, which can be used to distinguish electrons from charged pions. This is possible, because these particle create different amounts of transition radiation. The inner detector region is enclosed in a solenoid magnet to provide a magnetic field of 2 T for momentum measurements.

2.2.3 Calorimeter System

In the calorimeter system most of the particles loose their whole energy by initialising particle showers. Therefore, they consist of much more material than the inner detector. The calorimeters do also have active parts, that measure the energy left in the system. While this energy measurement is usually[1] less precise than the momentum measurement of the inner detector, it is the only measurement taken for neutral particles. The calorimeter system consists of two subsystems, the electromagnetic and the hadronic calorimeter. In the electromagnetic calorimeter (Ecal), particles are detected that start electromagnetic showers (like electrons or photons). In the hadronic calorimeter, hadronic showers are measured. However, particles can, of course, start a hadronic shower already before the hadronic calorimeter. The

[1] This is not true for charged particles with very high momentum. The precision of the track momentum measurement decreases with momentum, since the track curvature gets lower, while the calorimeter measurement increases with higher momentum.

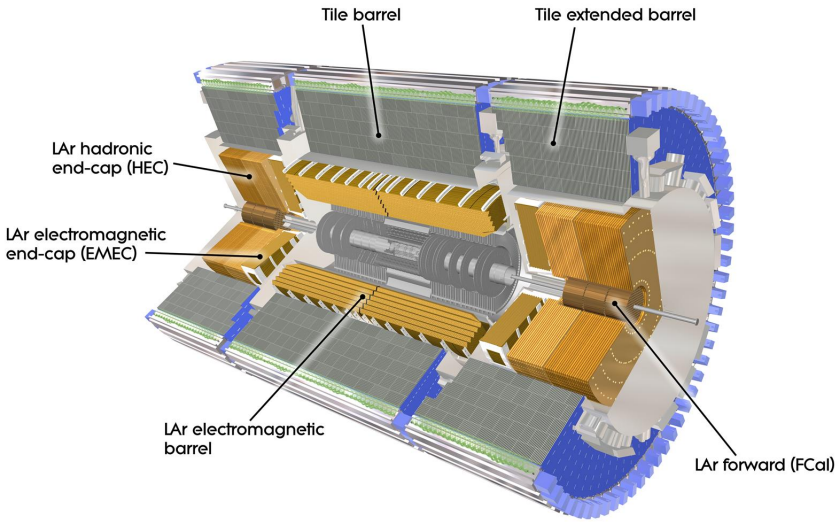


Figure 2.2: Overview of the calorimeter system [3]

whole system is shown in figure 2.2.

The Ecal consists of a barrel part, which is a cylinder around the beam axis, and two end-cap parts next to it. These consist of disc shaped detectors. The Ecal has accordion shaped layers of lead as absorption material and liquid argon (LAr) gaps in which the energy of the showers is measured.

A module of the barrel Ecal is shown in figure 2.3. There you can see the three layers of the calorimeter and their different granularities. The first sampling has a very fine granularity in the η direction.

The barrel part of the hadronic calorimeter (Hcal) is a tile calorimeter that uses iron as absorption material and scintillating tiles as active material for the energy measurement. It consists of a central barrel part and two extended barrels. Their thickness provides a good protection against punch-through of particles into the muon spectrometer and containment of hadronic showers. For the end-caps and forward[2] calorimeters LAr technology has been chosen. This has been done because LAr calorimeters provide the radiation hardness needed close to the beam pipe. Copper and tungsten is used as absorption material. Tungsten has been chosen for parts of the forward calorimeters because of the high density of the material.

2.2.4 Muon Spectrometer

Muons are minimal ionising particles (MIPs) that are able to traverse the whole ATLAS detector. Other charged particles are usually stopped at latest in the hadronic calorimeter. Therefore, the muon spectrometer is placed behind the calorimeter system to provide momentum measurement and identification specialised to muons. The large toroidal magnets and two end-cap magnets create a magnetic field of 3.9 T and 4.1 T, respectively. This is needed for muon momentum measurements. The muon spectro-

[2] Close to the beam pipe at large pseudorapidity

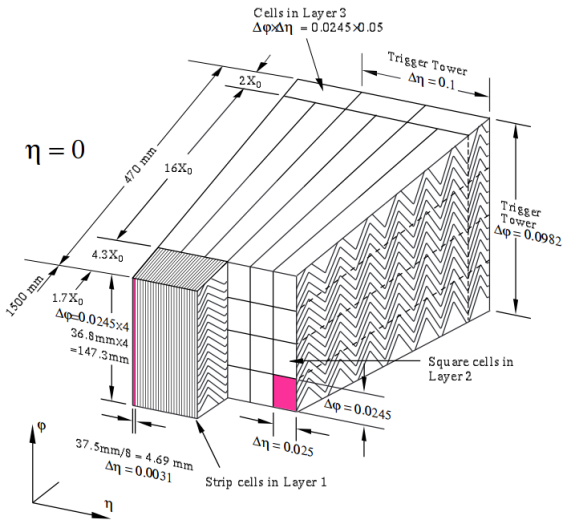


Figure 2.3: The module of the LAr electromagnetic barrel. The first layer has a fine granularity in the η direction. [3]

meter covers a range of $|\eta| < 2.7$. However, the muon trigger system covers a pseudorapidity range up to $|\eta| < 2.4$.

2.2.5 Luminosity, Trigger and Pile Up

The luminosity $\mathcal{L}$ of a collider is a measurement for the frequency of particle interactions. It is defined as

$$\mathcal{L} = \frac{1}{\sigma_{\text{mb}}} \dot{N},$$

where σ_{mb} is the *minimum bias cross section* and $\dot{N}$ the interaction rate. The minimum bias cross section refers to interactions on which only minimal requirements are set (to be detectable by ATLAS). This is lower than the total cross section, since many of the proton-proton interaction are outside the acceptance range of ATLAS. The luminosity is a collider property and depends on the number of protons in a bunch, the revolution frequency of the collider and the dimension and alignment of the proton bunches. Instead of the luminosity, often the integrated luminosity $L = \int \mathcal{L}\, \mathrm{d}t$ is used. Given the cross section σ of a process, one can calculate the expected number of interactions N as

$$N = L\sigma. \tag{2.1}$$

While the minimum bias cross section is in the range of mb, most of the interactions considered by that are not of interest for physics research.

At ATLAS, a trigger system selects the events of interest, while all other events are ignored. This is done in three steps. The Level-1 trigger (L1) decides in about 2 µs and reduces the frequency of events from 40 MHz to 75 kHz. This is done with only a subset of information from the calorimeter and

muon spectrometer. The Level-2 trigger (L2) uses the full resolution of all detector systems at regions of interest. These regions were selected by the Level-1 trigger. After this step about 1 kHz of events are left. The last step is the event filter (EF). A full reconstruction and analysis of the data is done. About 200 Hz of events do fulfil the requirements of the EF and are saved.

While it is typically important to have a high luminosity to gain statistics for analysis, this can also lead to problems. The luminosity can be increased either by increasing the frequency at which proton bunches collide or by condensing the proton bunch itself. An increased frequency is problematic, since the detector systems need a certain time to recover from a collision event. Condensing the proton bunch does lead to a higher number of interaction per collision. The problem with that is that most of these events are not of interest, but overlay the interaction of physical value. These additional events are called pile up. For $\sqrt{s} = 7\,\text{TeV}$ there are on average 3–19 interactions per bunch crossing. For $\sqrt{s} = 8\,\text{TeV}$ this increases to 6–39 interactions.

In run 1 (2010–2012) the ATLAS detector has recorded $4.57\,\text{fb}^{-1}$ usable integrated luminosity at $\sqrt{s} = 7\,\text{TeV}$ in 2010 and 2011 and $20.3\,\text{fb}^{-1}$ at $\sqrt{s} = 8\,\text{TeV}$ in 2012.

CHAPTER 3

Reconstruction and Identification of Tau Leptons

3.1 Tau Decays in ATLAS

With a mass of 1.777 GeV the tau lepton is the heaviest lepton known and the only one heavy enough to decay into hadrons. Due to its heavy mass it has also a very short life time of about 291 fs. The tau decays before it can be detected by ATLAS. Therefore, only the decay products of a tau can be seen in the detector.

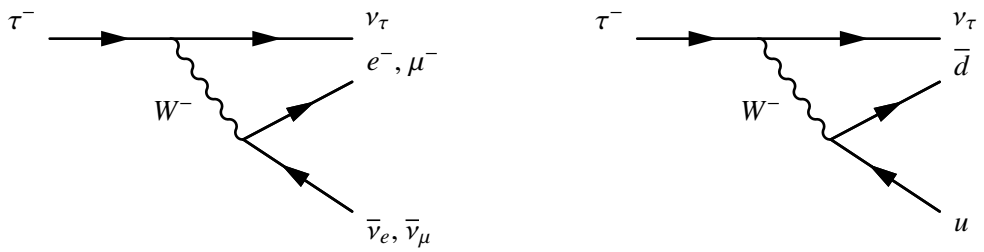
Similar to the β decay, the tau lepton decays via the weak interaction. The tau decays into a tau-neutrino (ν_τ) under radiation of a virtual W boson. The W boson can then decay leptonically or hadronically (figure 3.1). Leptonically, the W boson decays either into $e + \overline{\nu}_e$ or $\mu + \overline{\nu}_\mu$. The hadronic decay of the W boson lead to a quark pair. Table 3.1 shows the branching fractions of the tau decay modes. The suppression of the leptonic decays compared to the hadronic decay is due to the colour charge of quarks. The hadronic decay gives rise to a variety of different modes. The simplest one is the formation of a charged pion ($\overline{u}d$). Decays with multiple pions are also possible, e.g. via resonances like ρ^- or a_1^-, and even happen to appear more often. The tau lepton is also heavy enough to produce strange particles (kaons). However, they are treated as pions by the reconstruction algorithms since they are rare. The hadronic decays appear as a so called tau jet in the detector, consisting of charged and neutral hadrons that travel in approximately the same direction.

The reconstruction of taus in the ATLAS detector only handles its decay particles. A direct detection is not possible due to its short decay length of 87 µm (in the tau rest frame). The leptonic decays are difficult since there are 2 neutrinos produced and the charged lepton cannot easily be tagged as a tau decay product. Therefore only the hadronic decays are considered. The neutrino that appears in the hadronic modes can't be detected. Thus, the reconstructed tau in ATLAS does only consider the pions of its hadronic decay. This object is often referred to as *visible tau*, $\tau_{\text{vis}}^{\text{had}}$ or simply as tau.

	Decay	Branching Fraction	Notation
Leptonic	$e + \overline{\nu}_e + \nu_\tau$	17.83 %	
	$\mu + \overline{\nu}_\mu + \nu_\tau$	17.41 %	
Hadronic	$\pi^- + \pi^0 + \nu_\tau$	25.52 %	1p1n
	$\pi^- + \nu_\tau$	10.83 %	1p0n
	$\pi^- + 2\pi^0 + \nu_\tau$	9.30 %	1pXn
	$2\pi^- + \pi^+ + \nu_\tau$	8.99 %	3p0n
	$2\pi^- + \pi^+ + \pi^0 + \nu_\tau$	2.70 %	3pXn
	$\pi^- + 3\pi^0 + \nu_\tau$	1.05 %	1pXn
	Other modes	6.37 %	

Table 3.1: Branching fractions of the tau decay modes [5]. The decay modes considered for substructure reconstruction are listed explicitly in the hadronic section.

The reconstruction of the single decay particles of a tau is also called *tau substructure reconstruction*. However, not all decay modes are considered for the substructure reconstruction. A shorthand notation for the decay mode is $X\text{p}Y\text{n}$, where X denotes the number of charged particles and Y the number of neutral pions. The considered decay modes and their notation are listed in table 3.1.



(a) Leptonic decay

(b) Hadronic decay

Figure 3.1: Leptonic and hadronic tau decays

3.2 Reconstruction and Identification of Taus in ATLAS

This section gives an overview of how the hadronic tau decays are reconstructed and identified in ATLAS. In run 1 the standard tau reconstruction algorithm was tauRec. It is seeded from jets, recalculates their properties optimised for taus and identifies real taus from background jets. A new particle-flow based approach has been developed during run 1 and will be the standard tau reconstruction in run 2. It is seeded from tauRec taus, but goes beyond that in reconstructing the individual decay particles. Particle-flow based means to follow a single particle throughout the detector and combine the measurements from all detector systems. Algorithms that implement the tau substructure reconstruction in ATLAS are *cell-based* and *PanTau*. These, as well as the tauRec algorithm, are explained in the following.

3.2.1 Reconstruction of Taus with tauRec

The tauRec algorithm uses jets reconstructed with the anti-k_t algorithm that pass $p_{\mathrm{T}} \geq 10\,\mathrm{GeV}$ and $|\eta| \leq 2.5$ as tau seeds. For taus there is a dedicated vertex association algorithm called Tau Jet Vertex Association (TJVA). This algorithm calculates the best candidate for a primary vertex of the tau and is more robust against pile up than the method used for jets. With respect to this vertex a so called intermediate axis is calculated and the four-momentum of the tau is recalculated. The mass of a tau object is set to 0.

Up to now the algorithm only uses calorimeter measurements. Since a tau is charged, the tau-jet will also contain tracks. Only tracks that are within the core cone ($\Delta R \leq 0.2$ to the intermediate axis) and pass the cuts

- $p_{\mathrm{T}} \geq 1\,\mathrm{GeV}$,
- Number of pixel hits ≥ 2,
- Number of pixel hits + number of SCT hits ≥ 7,
- $|d_0| \leq 1\,\mathrm{mm}$ and
- $|z_0 \sin\theta| \leq 1.5\,\mathrm{mm}$

are associated to the tau, where d_0 is the transversal distance and z_0 the longitudinal distance of the closest approach to the tau vertex. Tracks in a wider region of $0.2 < \Delta R \leq 0.4$, called the isolation annulus, are used to separate taus from jets, but do not count for the number of tracks of a tau.

3.2.2 Separating Taus from Jets

Tau and QCD jets both consist of a collimated bunch of charged and neutral hadrons. Taus are on average narrower than jets. To separate taus from the jet background a number of variables are used, that compare the shape and the tracks around the tau candidate. These variables are evaluated using a multi-variate technique, the Boosted Decision Tree (BDT) method. Since BDTs are also used in tau substructure reconstruction, it is explained in more detail.

Boosted Decision Trees

Boosted decision trees can be used in particle identification to distinguish signal and background (in this case real taus from fake taus). A Monte-Carlo dataset is needed, where the right classification is already known. The dataset is then split into two samples, a *training* and a *test* sample. With the training sample the BDT is created as follows.

The classification into signal and background is done via a set of variables. The variable with the strongest separation power is used first. A cut value on that variable is chosen that splits the sample into two subsamples. The cut value is chosen such, that one of the subsamples is most signal-like and

the other most background-like. This procedure is repeated with the other separation variables. If a subsample contains only signal or background, the procedure is stopped and the subsample is called "signal" or "background" respectively. This is also done if the number of particles in a subsample becomes too low.

Up to know, this has been an explanation of a decision tree. Boosting means, that the weight of misclassified particles is increased. A misclassified background particle would then, e.g. be counted twice in the separation variable distribution and the counting of signal and background particles in the sample and subsamples. Since this can change the chosen cut values on the separation variables, the boosted decision tree can lead to different results. The boosting is repeated until the desired number of decision trees is created. Typical values are 500 or 1000 decision trees.

The so trained BDT must then be tested with the test sample. To decide whether a particle is signal or background, it is filled into each tree. Depending on how often the particle is classified as signal or background, a BDT score is calculated. This score lies in the range of -1 for a most background-like particle and 1 for a most signal-like particle. It can happen, that the BDT is *overtrained*, i.e. it is specifically good at separating signal from background for the training sample, but fails at an independent sample. This is checked by comparing the score distributions for the training and the test sample. If the BDT is not overtrained, both distribution will be approximately equal.

Tau Identification

In our case tau identification means to separate taus from jets. The variables used to do that are listed in table 3.2. The variable `tau_calcVars_corrCentFrac` for example quantifies, which fraction of the energy is in the central region ($\Delta R < 0.1$) of the tau. For a narrow object this would be large, for a wider object it would be low. Figure 3.2 shows the distribution of this variable for taus and jets. It can be clearly seen that for taus the distribution peaks at high values. For jets, the distribution is much flatter and does contain more lower values. Fractions above 1 are caused by the pile up correction applied to this variable. While in principle the tau identification is still separate from substructure reconstruction, a few ID variables are taken from there (called π^0 variables in table 3.2).

3.3 Reconstruction of Decay Particles with cell-based

This short description of the cell-based algorithm follows the master thesis of Benedict Winter [7].

To reconstruct the decay mode of a hadronically decaying tau lepton means to count the number of charged and neutral pions in a tau jet.

Charged pions leave tracks in the inner detector and energy in the electromagnetic and hadronic calorimeter. They are usually stopped in the hadronic calorimeter. In most cases (98.8 % [5]) neutral pions decay instantly into two photons. Since they have no charge, they do not create a track but leave a shower in the electromagnetic calorimeter. It is also possible, that one of the photons or both convert into an electron-positron pair. Since they also shower electromagnetically and are collimated, they leave

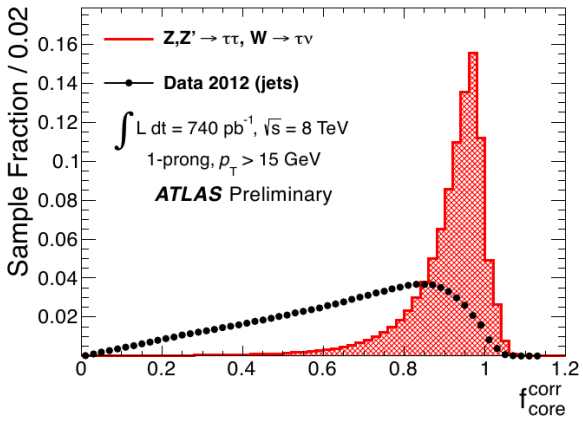


Figure 3.2: Distribution of the TauID variable `tau_calcVars_corrCentFrac`. The red area is for a Monte Carlo tau sample, the dots is a multi-jet sample taken from data. [6]

Variable name	**Description**
Any prong	
`tau_calcVars_corrCentFrac`	Pile-up corrected E_{T} fraction in central region ($\Delta R < 0.1$)
`tau_calcVars_corrFTrk`	Pile-up corrected leading track momentum fraction
`tau_seedCalo_trkAvgDist`	p_{T} weighted ΔR sum
1-prong only:	
`tau_ipSigLeadTrk`	Impact parameter significance of leading track
`tau_seedCalo_wideTrk_n`	Number of tracks in isolation annulus
multi-prong only:	
`tau_massTrkSys`	Invariant mass of track system
`tau_trFlightPathSig`	Transverse flight path significance
`tau_seedCalo_dRmax`	Maximal ΔR between tau axis and tau candidate track
π^0 variables (Any prong):	
`tau_ptRatio`	Ratio of substructure p_{T} and tauRec p_{T}
`tau_pi0_n`	Number of π^0s
`tau_pi0_vistau_m`	Mass of Lorentz vector made from track system and π^0s

Table 3.2: Variables used for Tau ID. 1-prong variables are used only for 1-prong taus, multi-prong variables are used for taus with 2 or more prongs. For a more detailed description, see [6].

η range	E_{T} cut	1-prong BDT cut	3-prong BDT cut
0.0–0.8	$\geq 1500\,\mathrm{MeV}$	≥ -0.06	≥ 0.43
0.8–1.4	$\geq 1700\,\mathrm{MeV}$	≥ -0.14	≥ 0.51
1.4–1.5	$\geq 1800\,\mathrm{MeV}$	≥ 0.01	≥ 0.48
1.5–1.9	$\geq 1500\,\mathrm{MeV}$	≥ -0.10	≥ 0.66
1.9–9.9	$\geq 1300\,\mathrm{MeV}$	≥ -0.01	≥ 0.65

Table 3.3: The cuts on the BDT score and E_{T} of neutral clusters

a similar signature in the calorimeter as a single photon.

While the number of charged pions can be counted by the number of tracks, the counting of neutral pions is more difficult. An algorithm that reconstructs neutral pions in a tau decay is cell-based. The substructure reconstruction is done in two steps:

- Subtraction of the energy the charged pions left in the electromagnetic calorimeter
- Reclustering of the remaining energy and identification of neutral pions

The subtraction is done by estimating the energy deposition in single cells of the electromagnetic calorimeter from the energy deposit in the hadronic calorimeter and the track momentum. The remaining energy is then reclustered. These clusters cannot simply be taken as neutral pions, since there are also energy clusters coming from pile up, noise or the imperfect energy subtraction. Therefore the second step is the π^0 identification. In order to do that, cell-based uses a boosted decision tree. The variables used in the BDT are listed in table 3.4. A neutral cluster is counted as a neutral pion, if it is passing η dependent cuts on E_{T} and the BDT score. The cuts valid for the used datasets are listed in table 3.3.

The decay mode classification of cell-based is then done according to the number of charged pions and the number of identified neutral pions in the core cone $\Delta R < 0.2$. One charged pion and two neutral pions would, e.g. be classified as a 1pXn tau.

3.4 Recovering Misclassified with PanTau

The decay mode classification in cell-based is done by looking at each neutral cluster in a tau separately. PanTau combines the information about all the constituents of a tau to recover misclassified taus. There are three tests implemented in PanTau that can migrate the classification between two decay modes. These are

- 1p0n-vs.-1p1n,
- 1p1n-vs.-1pXn and
- 3p0n-vs.-3pXn.

Variable name	Description
NPosCells_EM1	Number of cells with positive energy in Ecal1
NPosCells_EM2	Number of cells with positive energy in Ecal2
AsymmetryWRTTrack	Asymmetry of energy distribution in Ecal1 with respect to the extrapolated track position
ENG_FRAC_EM	Fraction of energy in EM calorimeter accordion
ENG_FRAC_CORE	Sum of the energy fractions in the most energetic cells per sampling
SECOND_R	Second moment in distance to the shower axis
CENTER_LAMBDA_helped	Distance to the shower centre from the calorimeter front face measured along the shower axis
Abs_FIRST_ETA	First moment in pseudorapidity
log_SECOND_ENG_DENS	Logarithm of the second moment in energy density
EcoreOverEEM1	Energy in the innermost Ecal1 cells normalized to total energy in Ecal1
secondEtaWRTClusterPosition_EM1	Second moment in pseudorapidity in Ecal1 with respect to the cluster position
secondEtaWRTClusterPosition_EM2	Second moment in pseudorapidity in Ecal2 with respect to the cluster position

Table 3.4: Cell-based π^0-ID BDT variables, the descriptions are taken from [7].

The tests are done via the evaluation of BDTs. Obviously a tau candidate cannot enter all of these test. Which test is chosen depends on the complete tau composition including neutral clusters that are tagged as non-π^0. A shorthand notation for this is $Rijk$, where R stands for "reconstructed", i for the number of charged constituents, j for the number of π^0-identified neutral constituents and k for the number of non-π^0 neutral constituents. However, not all possible combinations are considered. High multiplicities of neutral constituents are summed up with an "X". All considered compositions are listed in table 3.6 together with the test evaluated for that tau.

The variables that are used in the BDTs are listed in table 3.5.

Variable name	Description
1p0n-vs.-1p1n	
Neutral_PID_BDTValues_BDTSort_1	Highest π^0-BDT score found in all neutral PFOs
Neutral_Ratio_1stBDTEtOverEtAllConsts	Ratio of E_T in highest π^0-BDT score neutral E_T of all core constituents
Combined_DeltaR1stNeutralTo1stCharged	Distance in ΔR between the leading neutral and leading charged PFO
Charged_JetMoment_EtDRxTotalEt	Sum of E_T weighted distance of charged PFOs to the tau axis
Neutral_Shots_NPhotonsInSeed	Number of photons expected in tau candidate
1p1n-vs.-1pXn	
Neutral_PID_BDTValues_BDTSort_2	Second-highest π^0-BDT score found in all neutral PFOs
Neutral_HLV_SumM	Invariant mass of all neutral PFOs
Neutral_Ratio_EtOverEtAllConsts	Ratio of energy in neutral PFOs divided by energy of all core PFOs
Basic_NNeutralConsts	Number of neutral PFOs
Neutral_Shots_NPhotonsInSeed	Number of photons expected in tau candidate
3p0n-vs.-3pXn	
Neutral_PID_BDTValues_BDTSort_1	Highest π^0-BDT score found in all neutral PFOs
Neutral_Ratio_EtOverEtAllConsts	Ratio of energy in neutral PFOs divided by energy of all core PFOs
Charged_StdDev_Et_WrtEtAllConsts	Ratio of standard deviation of charged PFOs E_T values and E_T of all core PFOs
Neutral_Shots_NPhotonsInSeed	Number of photons expected in tau candidate

Table 3.5: PanTau BDT variables. The description is taken from [8]. PFOs are *Particle Flow Objects*. These are the technical objects in substructure reconstruction. They can either be a charged ($\pi^\pm$) or neutral. Neutral PFOs can be further distinguished into π^0-identified and non-π^0 PFOs.

Composition	Number of $\pi^\pm$ constituents	Number of π^0 constituents	Number of non-π^0 constituents	BDT test
$R100$	1	0	0	—
$R10X$	1	0	≥ 1	1p0n-vs.-1p1n
$R110$	1	1	0	1p0n-vs.-1p1n
$R11X$	1	1	≥ 1	1p1n-vs.-1pXn
$R1XX$	1	≥ 2	≥ 1	1p1n-vs.-1pXn
$R300$	3	0	0	—
$R30X$	3	0	≥ 1	3p0n-vs.-3pXn
$R3XX$	3	≥ 1	≥ 1	3p0n-vs.-3pXn

Table 3.6: The complete tau composition determines in which BDT test the tau is filled.

CHAPTER 4

Data/Monte-Carlo Comparison

The main background for tau identification are quarks or gluons, that hadronise and form jets. These *QCD* jets and tau jets both consist of charged and neutral hadrons. Their main difference is the number of tracks (1 or 3 for tau jets) and their shape. The tau substructure reconstruction is even seeded from jet finder algorithms, that are developed for reconstruction of QCD jets. Due to the similarity of both objects, the substructure reconstruction should also work on jets. Of course, the reconstructed decay modes are not really "decay" modes and the selection of particles which are taken into account for reconstruction is optimised for the tau case.

To select taus in a data sample, $Z \rightarrow \tau\tau$ events are used in which one tau decays hadronically. To gain a high purity, the other tau has to decay into a muon. This is done because particles that traverse the calorimeter system and reach the muon spectrometer are with a very high purity muons and can be easily identified. The process $Z \rightarrow (\tau \rightarrow \mu)(\tau \rightarrow \text{hadrons})$ only happens in $\approx 23\,\%$ of the cases. To reduce background contribution, identification requirements on the hadronically decaying tau would be made. This further decreases the statistics.

A Data/Monte-Carlo comparison on jets – as done in this study – has the advantage of higher statistics, since $Z(\rightarrow \mu\mu)$ + jets is easier to select and happens more often. In the production process of a Z boson, quarks and gluons can appear and form jets. Some example feynman diagrams for this process are shown in figure 4.1. In addition, the behaviour of the reconstruction on background can be studied. However, it should be kept in mind that usually the hadronisation modelling is not that good. This could mean, that jets are composed differently and their constituents have, e.g. a different p_{T} spectrum in Monte Carlo.

4.1 Experimental and Simulated Datasets

The used datasets are listed in table 4.2. The data is taken from period B of the 2012 LHC run at a centre of mass energy of 8 TeV. On the experimental data a GoodRunsList (GRL) is applied. Every run that

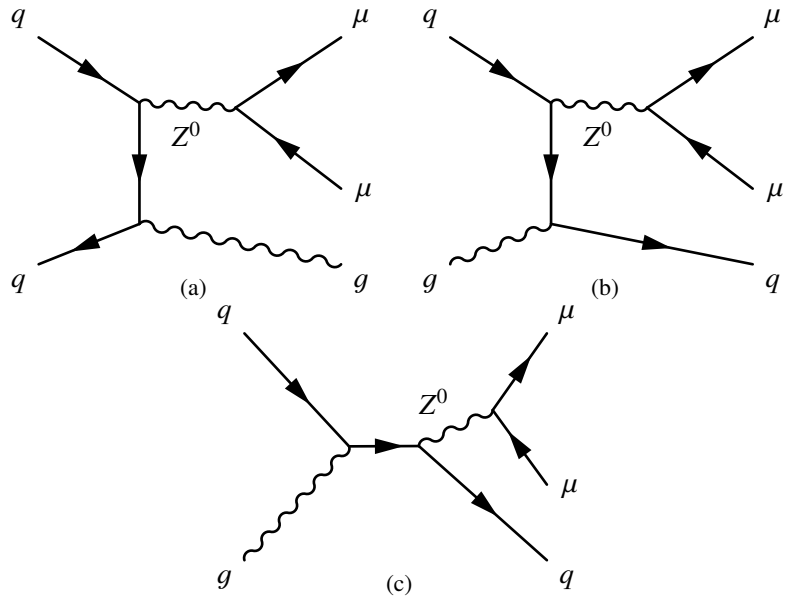


Figure 4.1: Z boson production processes

can be used for physics analysis is listed in a GRL. All other events should be skipped since they could contain unphysical information, e.g due to a detector system failure. The experimental dataset delivers $L = 5094.68\,/\text{pb}$ integrated luminosity. This value is calculated with the official ATLAS Luminosity Calculator, that has the used data taking period and the GRL as input. However, there are events missing in the actually used dataset files, which diminishes the integrated luminosity to

$$L_{\text{act}} = L \cdot \frac{N_{\text{act}}}{N_{\text{doc}}} = 4877.91\,/\text{pb}\,. \tag{4.1}$$

Here $N_{\text{act}} = 15\,857\,180$ is the number of events the analysis runs over and $N_{\text{doc}} = 16\,561\,872$ is the documented number of events for the dataset.

The Monte Carlo datasets do only cover signal events, because the background is expected to be extremely low. Six different datasets are used to cover different number of partons (gluons or quarks) forming jets. Unfortunately, the Monte Carlo datasets have two filters applied. They cut on muons and taus and are more stringent than the intended selection of this study. The first one is called `TauForZtautauFilter`, the other one `MuonTauLooseFilter`. They both require an isolated muon with a lower and an upper p_{T} cut and a tau with 1 to 3 tracks and a lower p_{T} cut. The detailed settings are listed in table 4.1. For TauID variables, other datasets have been used that are not affected from these filters. They could not be used for PanTau and cell-based variables since they did not contain the needed information.

An event passes the filters if one of the filter requirements are fulfilled. There are also filters applied on data, but these are much less restrictive. However, since we compare them with the simulated ones, the filters in table 4.1 have to be applied on data as well.

The Monte-Carlo datasets represent different physical processes, that have different cross sections.

TauForZtautauFilter	
A muon with	$p_T > 25\,\text{GeV}$
	$p_T < 45\,\text{GeV}$
	`ptCone40` < 0.06
	`etCone40` < 0.2
A tau with	$p_T > 20\,\text{GeV}$
	`numTrack` ≥ 1
	`numTrack` ≤ 3
MuonTauLooseFilter	
A muon with	$p_T > 22\,\text{GeV}$
	$p_T < 40\,\text{GeV}$
	`ptCone40` < 0.06
	`etCone40` < 0.2
A tau with	$p_T > 15\,\text{GeV}$
	`numTrack` ≥ 1
	`numTrack` ≤ 3

Table 4.1: These are the requirements of the filters applied on Monte Carlo. The cuts on `ptCone40` and `etCone40` are isolation requirements. They are defined as `ptCone40` $= \frac{\sum_{\text{tracks}\in\Delta R<0.4} p_T}{p_T^{\text{track}}}$ or E_T respectively. It means that the ratio of transversal momentum/energy of other tracks in a cone of $\Delta R < 0.4$ over the one of the muon/tau track may not be larger than the values given in the table.

The number of expected events N_{exp} can be calculated using the cross section σ according to equation (2.1). The physics process is simulated event by event, i.e. the amount of events is not necessarily related to the physical cross section. To calculate how many events are expected from all simulated datasets they have to be scaled according to their cross section, number of events and integrated luminosity of the experimental dataset. The scale factor is

$$\frac{L_{\text{act}} \cdot \sigma}{N_{\text{calc}}},$$

with the integrated luminosity L, the cross section σ and the actual number of events in the dataset N_{calc}. Since the cross section was only available for the unfiltered dataset, it has been calculated from the number of events in the filtered and unfiltered samples by

$$\sigma = \sigma_{\text{unfiltered}} \frac{N_{\text{filtered}}}{N_{\text{unfiltered}}}.$$

Table 4.3 shows the cut flow, i.e. the number of events that pass each cut step. After the last cut step, data and Monte Carlo are in good agreement. The cuts themselves are explained in section 4.3. Since this study focuses on shape comparisons of variable distributions, the simulated distributions are normalized to data.

Data:
user.pmalecki.TauPi0Rec_D3PD.periodB.physics_Muons. PhysCont.DESD_SGLMU.repro14_v01.v06-02/
MC for PanTau and cell-based variables:
user.mahansen.TauPi0Rec_D3PD.147113.AlpgenPythia_Auto_P2011C_ZmumuNp0. recon.DESD_SGLMU.e1880_s1581_s1586_r4767.v06-00/
user.mahansen.TauPi0Rec_D3PD.147114.AlpgenPythia_Auto_P2011C_ZmumuNp1. recon.DESD_SGLMU.e1880_s1581_s1586_r4767.v06-00/
user.natterma.TauPi0Rec_D3PD.147115.AlpgenPythia_Auto_P2011C_ZmumuNp2. recon.DESD_SGLMU.e1880_s1581_s1586_r4767.v06-00/
user.mahansen.TauPi0Rec_D3PD.147116.AlpgenPythia_Auto_P2011C_ZmumuNp3. recon.DESD_SGLMU.e1880_s1581_s1586_r4767.v06-00/
user.mahansen.TauPi0Rec_D3PD.147117.AlpgenPythia_Auto_P2011C_ZmumuNp4. recon.DESD_SGLMU.e1880_s1581_s1586_r4767.v06-00/
user.natterma.TauPi0Rec_D3PD.147118.AlpgenPythia_Auto_P2011C_ZmumuNp5incl. recon.DESD_SGLMU.e1880_s1581_s1586_r4767.v06-00/
MC for TauID variables:
mc12_8TeV.147113.AlpgenPythia_Auto_P2011C_ZmumuNp0. merge.NTUP_TAU.e1880_s1581_s1586_r3658_r3549_p1344/
mc12_8TeV.147114.AlpgenPythia_Auto_P2011C_ZmumuNp1. merge.NTUP_TAU.e1880_s1581_s1586_r3658_r3549_p1344/
mc12_8TeV.147115.AlpgenPythia_Auto_P2011C_ZmumuNp2. merge.NTUP_TAU.e1880_s1581_s1586_r3658_r3549_p1344/
mc12_8TeV.147116.AlpgenPythia_Auto_P2011C_ZmumuNp3. merge.NTUP_TAU.e1880_s1581_s1586_r3658_r3549_p1344/
mc12_8TeV.147117.AlpgenPythia_Auto_P2011C_ZmumuNp4. merge.NTUP_TAU.e1880_s1581_s1586_r3658_r3549_p1344/
mc12_8TeV.147118.AlpgenPythia_Auto_P2011C_ZmumuNp5incl. merge.NTUP_TAU.e1880_s1581_s1586_r3658_r3549_p1344/

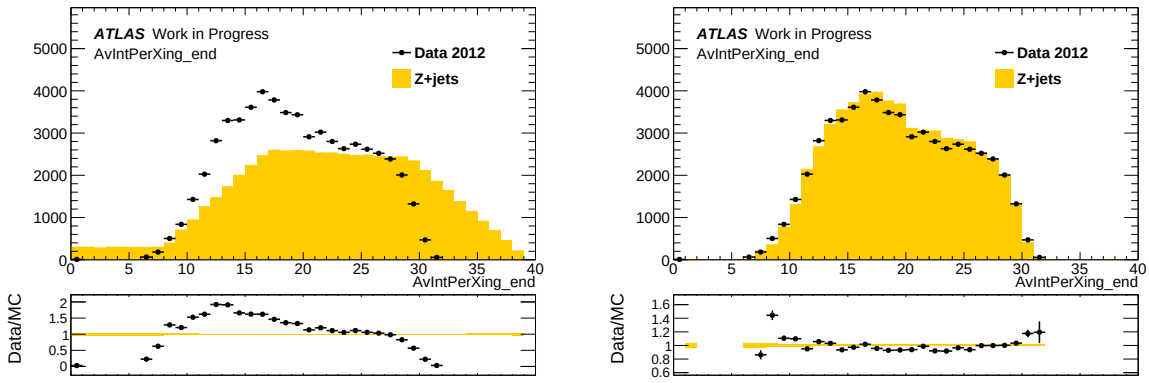
Table 4.2: The used datasets for experimental and simulated data.

4.2 Corrections Applied to Monte Carlo

When two proton bunches collide in the ATLAS detector the numerous partons generate more than just one collision. The quantity that characterises how many interactions take place per collision of a proton bunch is called *average interactions per crossing* or simply μ. It is defined as

$$\mu = \frac{L\sigma_{\text{mb}}}{fn},$$

where L is the luminosity, σ_{mb} the minimum bias cross section, f the revolution frequency of the LHC and n the number of proton bunches. This quantity is not always constant, but depends on the state of the collider. However, its distribution differs between simulated and experimental data. To correct this, a pile up re-weighting is applied on Monte Carlo events. Depending on μ, this gives every event a certain weight. If the μ of the event is overrepresented, it will get a low weight and vice versa. Without re-weighting (figure 4.3a), the distribution overestimates the high region above 30. With the correction applied (figure 4.3b) experimental and simulated distributions match. The tool that performs this correction is developed by the ATLAS AnalysisSoftwareGroup and described in [9].



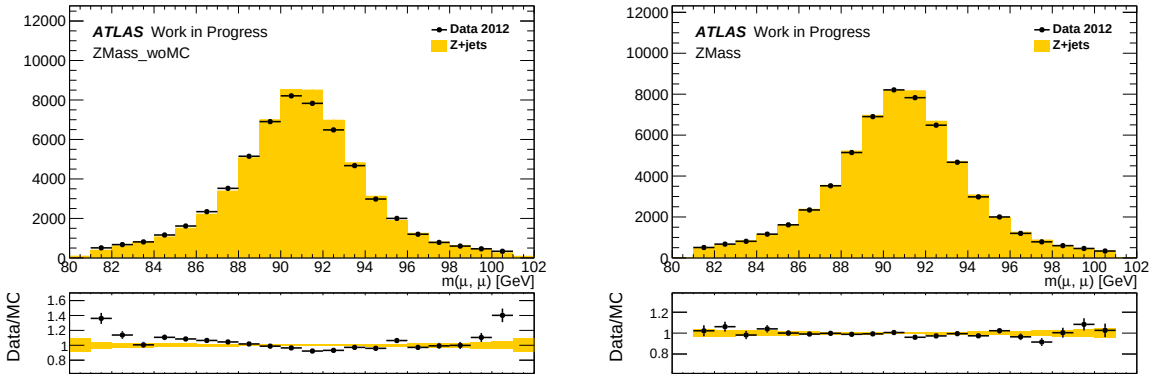
(a) without re-weighting

(b) with re-weighting

Figure 4.2: Average interactions per crossing with and without pile-up re-weighting.

There are three corrections applied concerning muons. One of them is the correction of the muon momentum. The assumed resolution of the momentum measurement in Monte Carlo is a little overestimated. A tool, developed by the ATLAS Muon Working Group (Muon WG), is used to smear out the muon momentum resolution. It uses the transverse momentum measurements from the inner detector, the muon spectrometer and the combined measurement to provide a smeared-out combined measurement value. It also uses the orientation of the muon in η and ϕ. A random number generator is used to apply this smearing statistically distributed. However, to guarantee the same smearing for same muons, the event number and muon index is used as random seeds. This correction is directly visible when comparing the mass of the di-muon system with and without correction in figure 4.3.

The second correction is called Muon Efficiency Corrections. The efficiency to reconstruct a muon is



(a) without muon corrections

(b) with muon corrections

Figure 4.3: Reconstructed Z boson mass with and without muon corrections.

not totally equal in experimental and simulated data. Scale factors are provided by the ATLAS Muon Working Group that compensate this. A third tool, called TrigMuonEfficiency, corrects the efficiencies of the di-muon trigger in Monte Carlo. However these factors are small compared to the momentum correction and thus have less effect. The tools are described in [10] and [11].

4.3 Selecting $Z \to \mu\mu$ Events

In table 4.4 you can see an overview of the event selection applied in this study. The selection follows the standard selection of Z+jets events proposed by the ATLAS Standard Model Working Group (SM WG) [12].

The first selection applied is the di-muon trigger `mu18_tight_mu8_EFFS`. This trigger selects events with multiple muons, after the single muon trigger `mu18_tight` is confirmed. The second step is to skip events with *bad jets*. These jets are measured energy in the calorimeter that is caused by hardware defects, beam background or cosmic rays.

Now the actual selection for two opposite charged muons is done. For this study, muons reconstructed by the STACO algorithm have been used. STACO combines a track in the inner detector with a muon spectrometer track using a statistical method. These muons can be selected by using muons from the `StacoMuonCollection` and requiring them to be `combined`.

The muon selection requires at least 20 GeV p_{T} and $\eta < 2.4$ since this is the coverage of the muon trigger system. In jets, that originate from a b quark, muons can appear. Therefore muons overlapping with jets within $\Delta R < 0.4$ are not considered. Track quality requirements have been applied to ensure a low number of misreconstructed tracks at a high efficiency. These are listed in table 4.4.

Only events that have exactly two of those muons with opposite charge are considered, since the muons from a Z boson decay are opposite charged. In addition the mass of the di-muon system has to be around the Z boson mass (81 GeV $< m(\mu,\mu) <$ 101 GeV). The cut on the di-muon system mass is a

	Cut	Data	MC
Filtered datasets	Passed GRL	15 426 748	570 794
	Passed Filter	3 106 770	429 070
	Passed Trigger	350 802	268 197
	Passed MuonSelection	203 417	190 950
	Passed ZmassCut	175 141	174 533
	Passed GoodVertex	172 557	173 971
	Passed TauSelection	58 915	59 958
Unfiltered datasets	Passed GRL	15 426 748	5 657 941
	Passed Trigger	3 607 983	2 572 224
	Passed JetCleaning	3 602 553	2 569 983
	Passed MuonSelection	2 083 731	1 894 994
	Passed ZmassCut	1 805 956	1 708 814
	Passed GoodVertex	1 792 371	1 695 836
	Passed TauSelection	83 638	87 014

Table 4.3: This table shows the number of events that pass each cut for data and Monte Carlo. Due to a bug, the number of events after the jet cleaning is not available for the filtered datasets. You can still estimate the effect by comparing with the unfiltered cut flow.

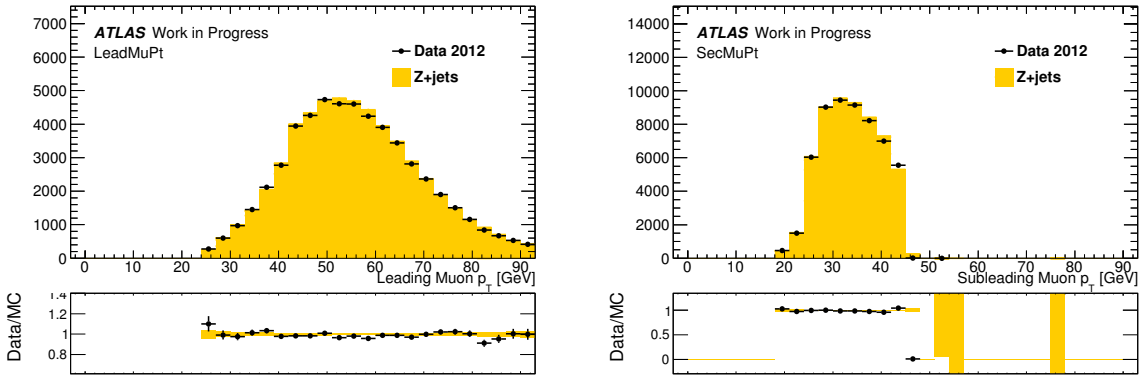
bit more restrictive than the one proposed in [12]. This has been done to reduce background contribution in the tails since outside of the chosen region the Data/Monte-Carlo agreement is slightly worse.

The tau candidates need to have $p_{\mathrm{T}} > 24\,\mathrm{GeV}$ and a pseudorapidity $\eta < 2.5$. In hadronic tau decays only 1 or 3 charged particles appear. Therefore taus are required to have 1 or 3 associated tracks. A cut on the jet vertex fraction of `jvtxf>0.5` has been applied. The jet vertex fraction is the fraction of tracks in the jet, that comes from the associated vertex. A cut on `jvtxf` reduces possible pile-up contribution. After this event selection and the requirements on tau candidates, only the candidate with the highest p_{T} is used for the comparison.

Figure 4.4 shows the transverse momenta of the leading and the subleading muon. The distributions show good agreement between data and Monte Carlo. In figure 4.5 the mass and transverse momentum of the di-muon system is shown. The Z boson mass peak is clearly visible with a good agreement, especially in the central region. The p_{T} spectrum shows a little kink at low p_{T}. In figure 4.6 p_{T} and η of the tau candidate are shown. In the selected kinematic region the distributions again show good agreement.

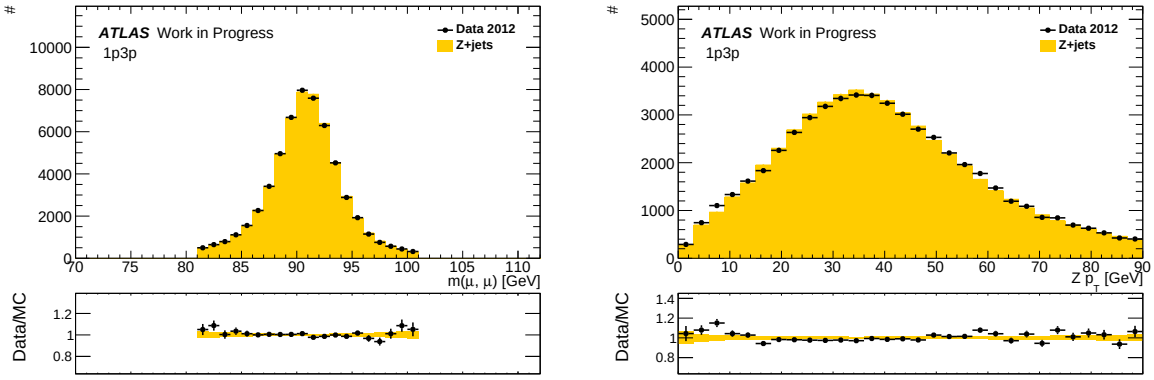
4.4 Evaluation via χ^2 Test

The main aspect of this study lies on the variables that PanTau uses in its BDTs to migrate the reconstructed decay modes from cell-based. These variables are not only looked at for all tau candidates that pass the selection, but also depending on their kinematics. The tau candidates are grouped in bins of p_{T} and η, which are also referred to as *dependency variables*. This enables us to study the modelling of the



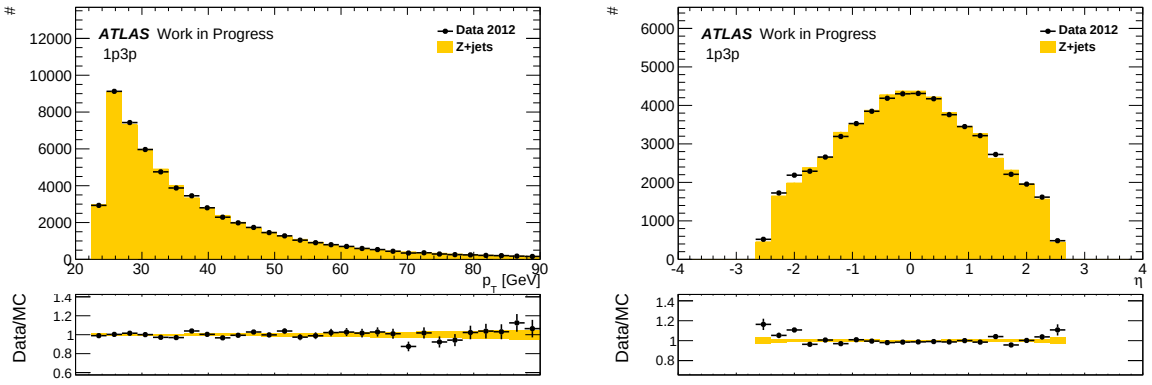
(a) (b)

Figure 4.4: p_{T} of the leading and subleading muon



(a) (b)

Figure 4.5: The mass and p_{T} of the Z boson



(a) (b)

Figure 4.6: p_{T} and η of the tau candidate

Trigger `mu18_tight_mu8_EFFS`
Jet Cleaning Skip events were a jet is tagged a `isBadLooseMinus`.
Good vertex Skip events in which the primary vertex has `nTracks`≤ 2.

Muons
`Tau D3PD prefix: mu_staco_*`
Exactly two muons that pass:
$p_T > 20\,\text{GeV}$
$|\eta| < 2.4$
`isCombinedMuon == 1`
Remove muons overlapping with jets within $\Delta R < 0.4$
Track quality cuts:
N(pixel hits) + N(pixel dead) ≥ 1
N(SCT hits) + N(SCT dead) ≥ 5
N(pixel holes) + N(SCT holes) ≤ 2
TRT quality cuts:

```
if 0.1 < abs(eta) < 1.9:
    if not ( (nTRTHits + nTRTOutliers > 5) and \
             (nTRTOutliers < 0.9*(nTRTHits + nTRTOutliers))):
        return False
return True
```

Di-muon system
$81\,\text{GeV} < m(\mu,\mu) < 101\,\text{GeV}$

Tau Candidates
`Tau D3PD prefix: tau_*`
$p_T > 24\,\text{GeV}$
$|\eta| < 2.5$
`tau_numTrack == 1 || tau_numTrack == 3`
Remove candidates that cannot be matched to a jet within $\Delta R < 0.6$
`jvtxf > 0.5`

Table 4.4: Selection for Zmumu+jets events

variables for certain kinematic regions. Please note that p_T and η are not uncorrelated, since particles of the same momentum have a lower p_T at high η.

The χ^2 test functionality of ROOT has been used to evaluate the agreement of the variable distributions. It is based on a χ^2 method developed by Gagunashvili [13], that improves the common Pearson's χ^2 test for comparing unweighted (data) with weighted (Monte Carlo) histograms. The χ^2 gives a so called p-value. It describes how probable it is, to get a test statistic at least as extreme as observed, given that the hypothesis of identity is true. That means a low p-value indicates bad agreement, a high one good agreement. Since the p-value is a probability it has values between 0 and 1. The used method is described below, referring to Gagunashvili.

The χ^2 test formulates the hypothesis of identity of an unweighted and a weighted histogram with r bins as follows. The number of events or weights in the ith histogram bin is n_i for the unweighted and w_i for the weighted histogram. Their sum gives the total number of events $N = \sum n_i$ or weights

$W = \sum w_i$ respectively. The true probability of belonging to bin $r \in [1; r]$ is p_i with $\sum_{i=1}^{r} p_i = 1$. For the unweighted histogram n_i is a Poisson-distributed variable with mean value at Np_i. For the weighted histogram w_i is normal distributed with a mean at Wp_i and variance σ_i^2. The method estimates the variance by the sum of squares of weights s_i^2 in that bin.

Given that the hypothesis of identity is valid, the maximum likelihood estimator for the true probability p_i is

$$\hat{p}_i = \frac{Ww_i - Ns_i^2 + \sqrt{(Ww_i - Ns_i^2)^2 + 4W^2 s_i^2 n_i}}{2W^2}, \tag{4.2}$$

and the χ^2-value is

$$\chi^2 = \sum_{i=1}^{r} \frac{(n_i - N\hat{p}_i)^2}{N\hat{p}_i} + \sum_{i=1}^{r} \frac{(w_i - W\hat{p}_i)^2}{s_i^2}. \tag{4.3}$$

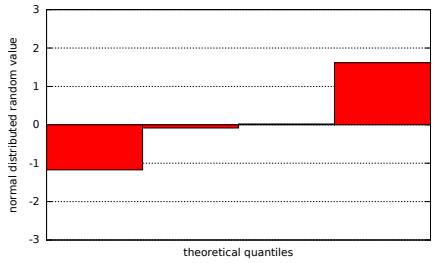
A high difference between the estimated expected events ($N\hat{p}_i$ and $W\hat{p}_i$) and the observed events (n_i and w_i) leads to a higher χ^2. With the assumption that the above χ^2 is approximately Chi-Squared distributed, one can calculate the p-value by integration of the Chi-Squared distribution for the number of degrees of freedom $ndf = r - 1$ from χ^2 to ∞ . This method has restrictions on the estimated expected events. It requires at least 1 expected event for an unweighted histogram bin (data) and at least 10 expected events for a weighted histograms bin (Monte Carlo). Bins with lower statistics than that can have heavy effects on the calculated p values. For the variable `Charged_JetMoment_EtDRxTotalEt` of the 1p0n-vs.-1p1n test, the p value of the low p_T bin goes down from 0.208 719 to $3.064\,41 \times 10^{-77}$, if low statistics bins are included for the χ^2 test. The other bins, however, are much less affected. To avoid such an influence on the χ^2 test, such bins are removed before applying the test.

In order to make the p-values of a variable comparable between the dependency variable bins, the bin borders are chosen for the same data statistics in every bin. This also means, that the bins cannot simply be adjusted to the detector geometry. The only way of changing the bins is by adjusting the number of bins. Plots of the statistics in each dependency bin can be found in the appendix in figure A.1, A.22 and A.39.

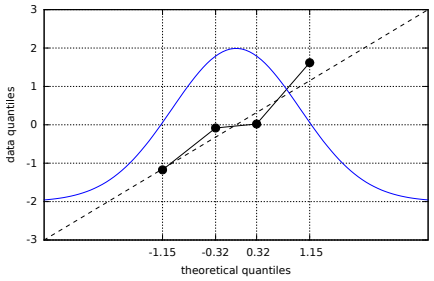
The statistics heavily differs between the three tests of PanTau. This has an immediate effect on the χ^2 test. For lower statistics the relative error is higher, which leads to better p values. The p values have a higher sensitivity for higher statistics. This does, of course, also apply to the p value plots of the cell-based π^0 ID variables and the Tau ID variables. If there are p values missing in the respective plots, this means that the p value was below 1×10^{-300}. The χ^2 test also provides normalised residuals for further analysis. Normalised residuals are a measure for the deviation between the two distributions for a single bin in terms of the error (sigma).

With the help of the residuals and a Q-Q plot, it can be determined, which bins lead to a bad p value. In a Q-Q plot the quantiles of the residual distribution is plotted against the quantiles of a test distribution, in this study a gaussian distribution is assumed. An example of this is shown in figure 4.7. The residuals are ordered by value (figure 4.7a) and the normal distribution is cut into equally probable bins[1] (figure

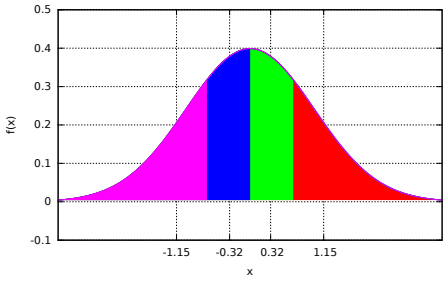
[1] Into bins of equal integral over the distribution



(a) Four normal distributed random values



(b) Q-Q plot



(c) Binned normal distribution

Figure 4.7: Figure (a) shows four normal distributed random values. They are sorted by value. Figure (c) shows the normal distribution in four equally probable bins. On the x axis the bin means are marked. Figure (b) shows the resulting Q-Q plot. The random values on the y axis are plotted against the normal distribution bin means on the x axis. The normal distribution is shown in blue for orientation.

4.7c). A specific bin is then represented by its mean value (on the x axis). If the residuals are normal distributed, each of them should lie in the range of on the mean values of the normal distribution parts. In the Q-Q plot, the residuals are then plotted against the mean values of the normal distribution (figure 4.7b). Ideally, this would lead to a straight line. This is approximately the case for the normal distributed random values shown here as an example.

Due to the sensitivity of the χ^2 test to statistics effects and disagreement in single bins, the resulting p values should be evaluated with caution. The test should rather be used as a hint to disagreement and to visualise trends. A check of the distributions is still necessary.

4.5 Systematic Uncertainties

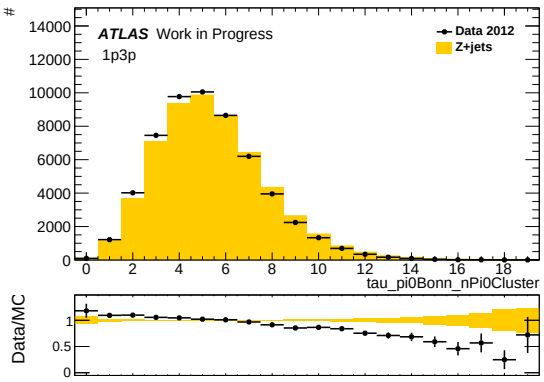
Sources of mismodelling relevant for this study are:

- jet hadronisation
- cell-based charged pion subtraction
- modelling of the calorimeter energy measurement

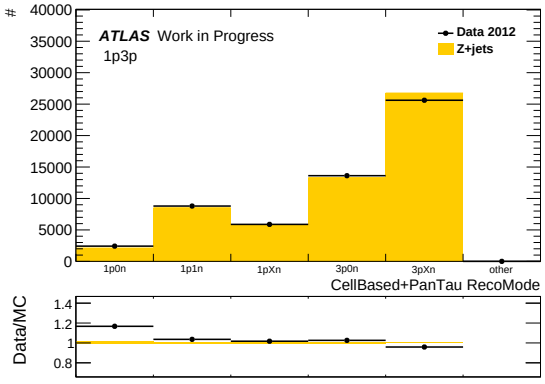
Several approaches have been made to estimate the systematic uncertainties on the decay mode classification. The number of Pi0Clusters (`nPi0Cluster`) is the number of neutral clusters in a reconstructed tau that pass a cut on $p_T > 500\,\text{MeV}$ after the subtraction of the charged pion's energy. No π^0 identification has been applied on these objects. Figure 4.8a shows that the number of neutral clusters is overestimated in Monte Carlo. This could lead to a higher amount of decay modes with neutrals, which is what can be seen in figure 4.8b. The mismodelling visible here is typically sourced in the jet hadronisation modelling, but also the cell-based charged pion subtraction can play a role here. To estimate the effect of this, event weights have been calculated so that the distributions of `nPi0Cluster` match. The impact of this re-weighting on the decay mode classification can be seen in figure 4.8c. The effect is rather small, but especially in the 3 prong decay modes the agreement gets a bit better.

In Benedict Winter's master thesis [7] it is noted, that the energy deposit of the charged hadron in the first layer of the Ecal varies heavily. In this layer neutral pions are reconstructed. Therefore the energy of the neutral pions does have a large uncertainty from the subtraction of the energy of the charged hadron in the Ecal. This can have an effect on the decay mode classification, since there is a E_T cut on identified pions. The effect of a mismodelling of the cluster energy has been quantified in the following way. The number of Pi0Clusters that would pass the π^0 identification has been calculated for nominal E_T (`nPi0_nom`), for a E_T that is 20 % higher (`nPi0_up`) and for a E_T that is 20 % lower (`nPi0_down`). If the difference between them (`nPi0_up` − `nPi0_nom` or `nPi0_nom` − `nPi0_down`) is not zero, the decay mode was changed accordingly. Also migration between 1p0n and 1pXn was taken into account, if the difference was greater than 1. In figure 4.9a, the decay mode classification by cell-based+PanTau is shown. The red striped areas mark the uncertainties estimated from this method.

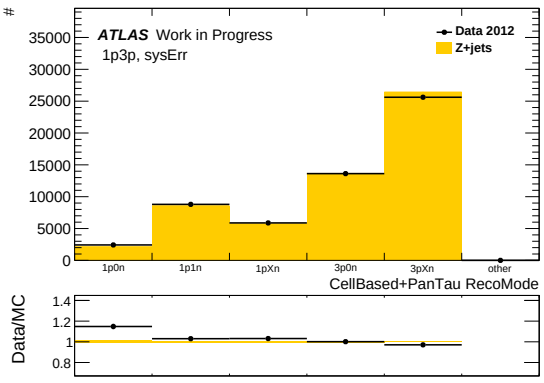
The systematic uncertainties obtained from the last method do cover the disagreement between data and Monte Carlo. However the justification of this is not clear. Figure 4.9b shows the summed p_T of π^0 identified neutrals. The distributions show a good agreement. Their means only differ by (272 ± 14) MeV at values of ≈ 14.9 GeV. This would lead to a p_T uncertainty of ≈ 1.8 %, which would have a vanishing effect on the decay mode classification.



(a)



(b) without re-weighting



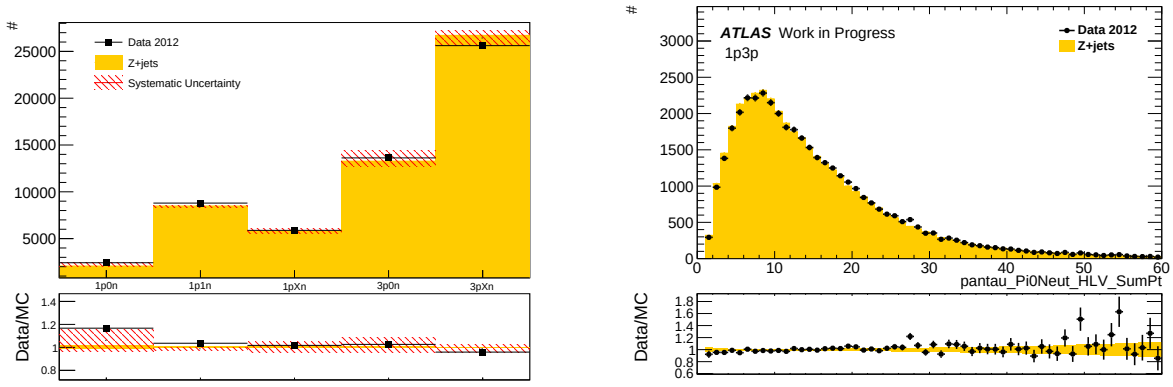
(c) with re-weighting

Figure 4.8: Figure (a) shows the distribution of `nPi0Cluster`. From this, systematic uncertainties have been obtained for the decay mode classification. The effect of this can be seen in figure (b) and (c).

4.6 Results

One main result of the tau substructure reconstruction is the decay mode classification. The distribution for this is shown in figure 4.9a and the p value plots in figure 4.10. The p values indicate the most disagreement for the high η and low p_T regions.

In the following sections, selected distributions and p value plots are shown. They are divided into PanTau BDT variables, cell-based BDT variables and TauID variables. A more complete set of plots can be found in the appendix.



(a) (b)

Figure 4.9: Figure (a) shows the decay modes reconstructed by cell-based+PanTau with the systematic uncertainties from the decay mode shifting method. Figure (b) shows the distribution of `Pi0_sumPt`.

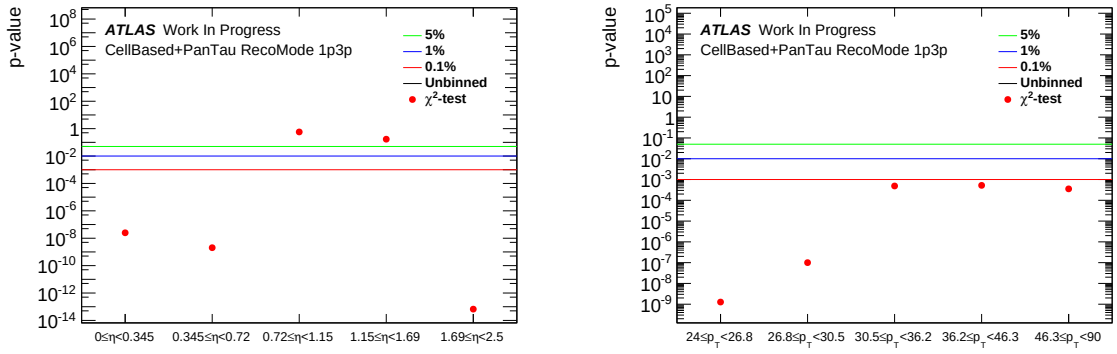
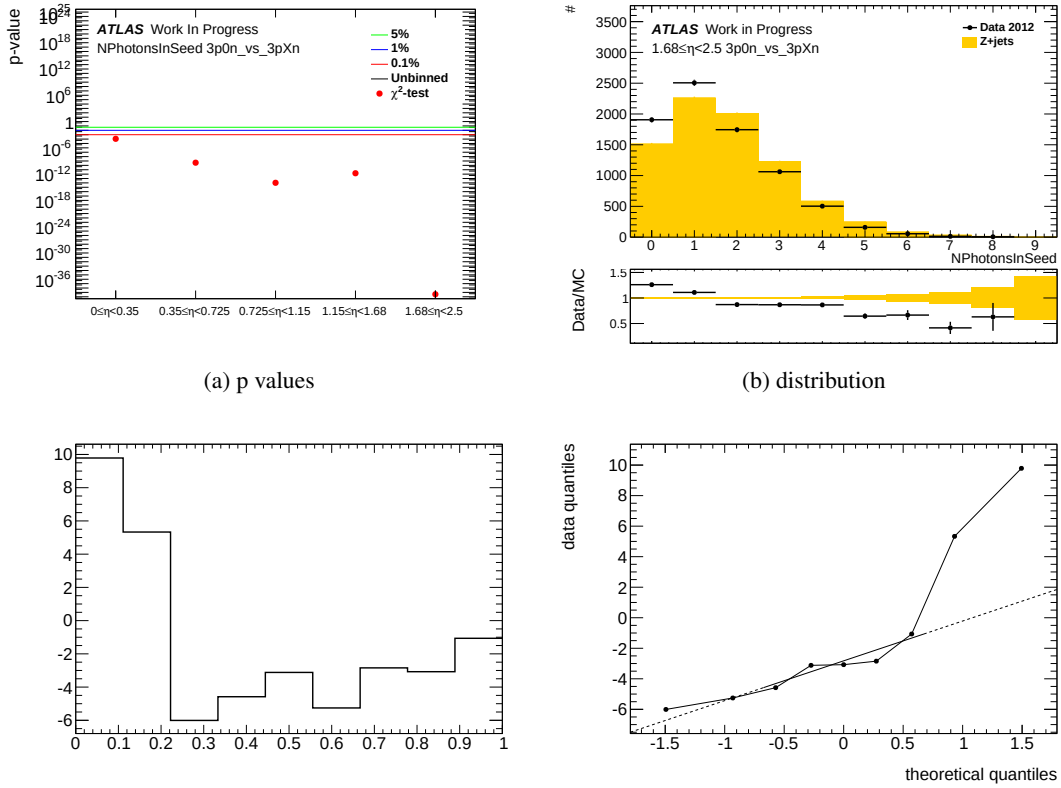


Figure 4.10: P value plots for the decay mode classification of Cell-based+PanTau

4.6.1 PanTau BDT Variables



(a) p values

(b) distribution

(c) residuals

(d) Q-Q plot

Figure 4.11: Figure (a) shows the p values of the BDT variable `NPhotonsInSeed` for the 3p0n-vs.-3pXn. Positive and negative values of η are taken into account. Figure (b) shows the distribution for the high η bin that has the lowest p value. Figure (c) shows the normalised residuals of non-empty bins used for the χ^2 test, the x axis is arbitrary. Figure (d) shows the Q-Q plot that compares the quantiles of the residuals with those of the theoretical gaussian distribution.

Figure 4.11a shows an example of the resulting p value plots. It shows the p values of the 3p0n-vs.-3pXn BDT variable `NPhotonsInSeed` depending on the absolute value of η. The x axis is labelled with the dependency bins. You can see that for high pseudorapidity the agreement becomes worse. The lines drawn in this plot mark prominent p value levels. Every p value above the red 0.1 % line is in agreement within the statistical errors. A black line in these plots mark the p value of the inclusive distribution (not binned in η). For `NPhotonsInSeed`, the line is not drawn, because the p value is too low. The unbinned p value tends to be lower due to higher statistics than in the binned distributions. The disagreement in the high η bin will, of course, also affect the unbinned p value. The title in the plot gives information about the tested variable and the kind of tau candidates that are taken into account. In this case, all tau candidates in the 3p0n-vs.-3pXn test.

Figure 4.11b shows the distribution of `NPhotonsInSeed` for the high η bin. The x axis is labelled with the plotted variable, the title indicates the kind of tau candidates included in this plot. In this case,

tau candidates in the 3p0n-vs.-3pXn test in the high η bin. Heavy disagreement is visible, especially for 0 to 3 `NPhotonsInSeed`.

In figure 4.11c and 4.11d, the residuals and the Q-Q plot of the `NPhotonsInSeed` distribution in the high η bin is displayed. Only non-empty bins are shown. The highest residuals are in the first three bins. In the Q-Q plot it is visible that these high residuals come at a higher frequency than expected from a gaussian distribution (The arm at the top right). Hence, the deviation in these bins lead to a bad p value. The rise of the dashed line is also steeper than $y = x$. This indicates that the residuals are more scattered than expected from a gaussian; the errors are too small for the hypothesis of identity.

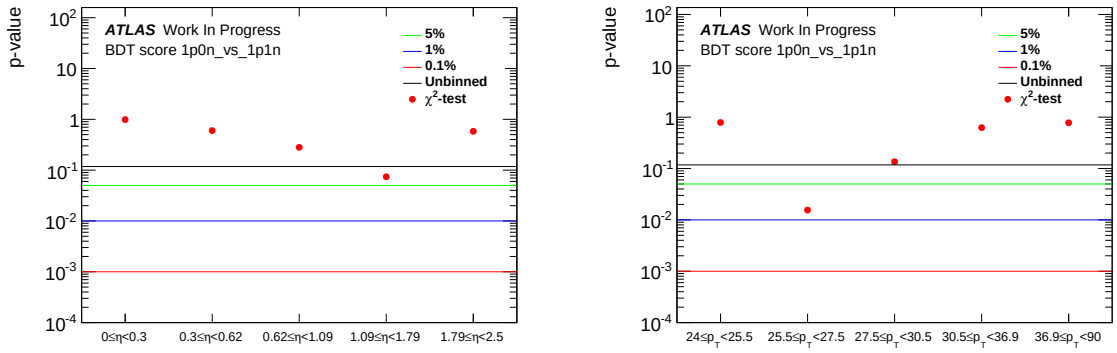


Figure 4.12: P value plots for the BDT scores of PanTau's 1p0n-vs.-1p1n test. p_{T} is in GeV.

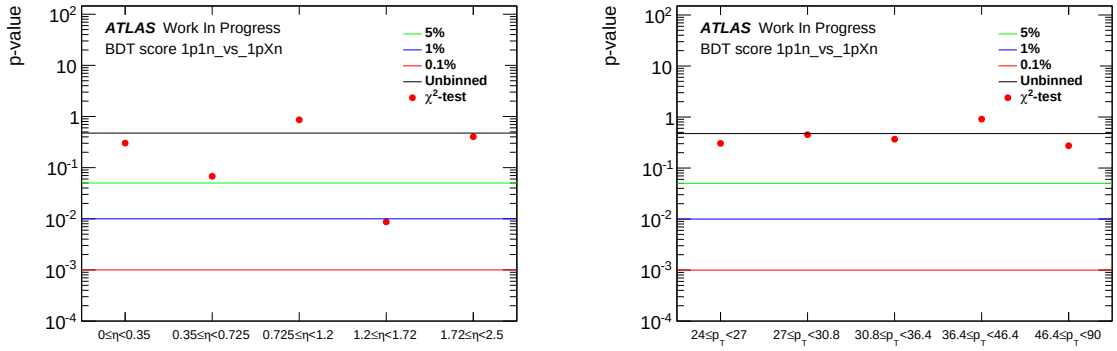


Figure 4.13: P value plots for the BDT scores of PanTau's 1p1n-vs.-1pXn test. p_{T} is in GeV.

The p value plots for the 1p0n-vs.-1p1n and the 1p1n-vs.-1pXn BDT score are shown in figure 4.12 and 4.13. They show no significant disagreement in the whole kinematic range. Figure 4.14 shows the p values of the 3p0n-vs.-3pXn BDT score. Here disagreement for high η and low p_{T} is visible. Note, that the χ^2 test is more sensitive in this case due to higher statistics than for the other BDT scores. The same behaviour is also visible in the BDT variables `BDTSort1` and `NPhotonsInSeed` (figure 4.15 for η). Figure 4.16 shows the BDT score and the distributions of `NPhotonsInSeed` and `BDTSort1` for the lowest and highest η bin. For `NPhotonsInSeed` and `BDTSort1` the disagreement is much worse for

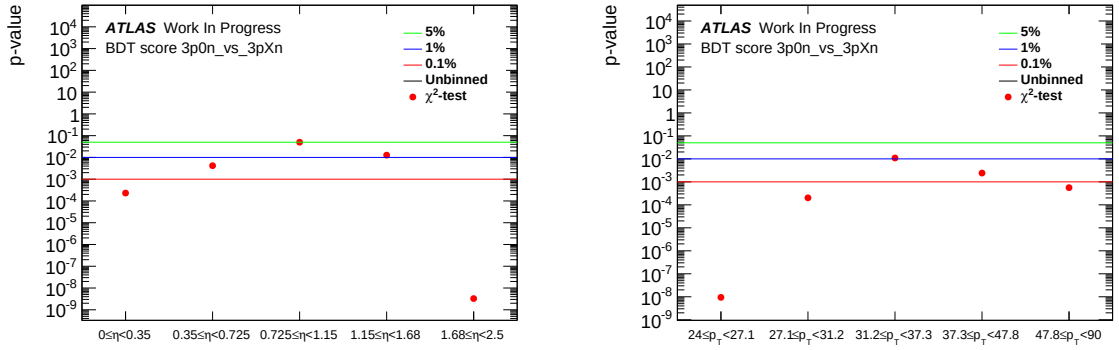
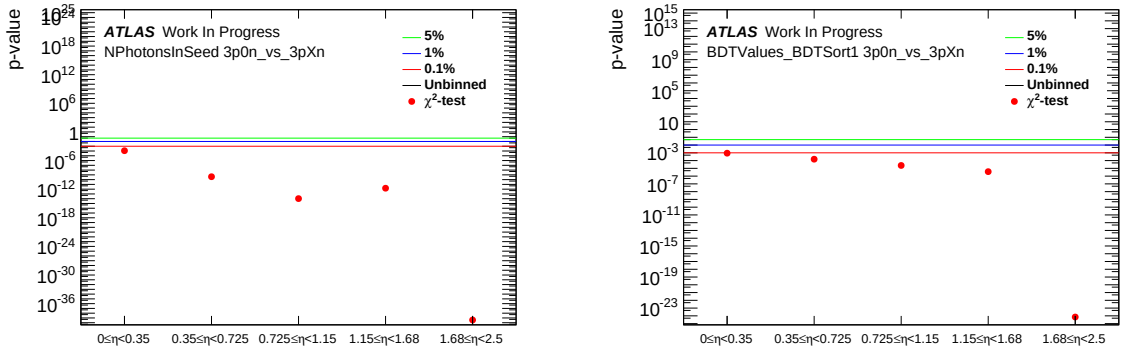


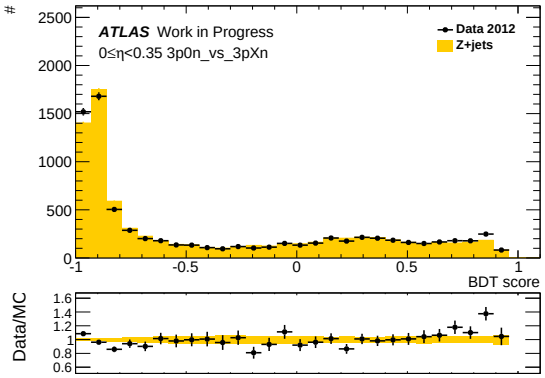
Figure 4.14: P value plots for the BDT scores of PanTau's 3p0n-vs.-3pXn test. p_{T} is in GeV.

high η, as indicated by the p value plot. However, the BDT score itself has a reasonable agreement. Inclusive distributions for all variables used in the PanTau BDTs are shown in the appendix in figure A.2, A.3 and A.4.

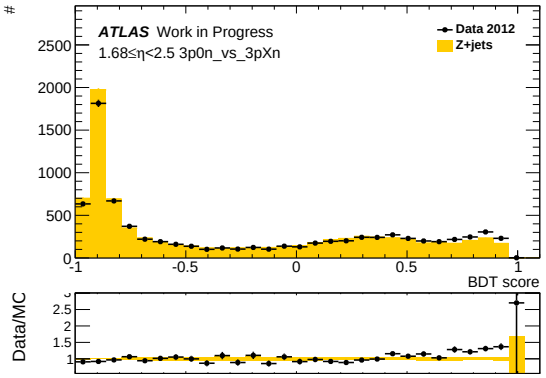


(a) (b)

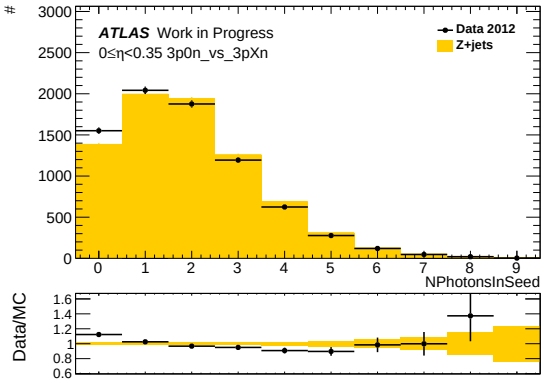
Figure 4.15: P value plots for the variables `NPhotonsInSeed` and `BDTSort1` of PanTau's 3p0n-vs.-3pXn test for η.



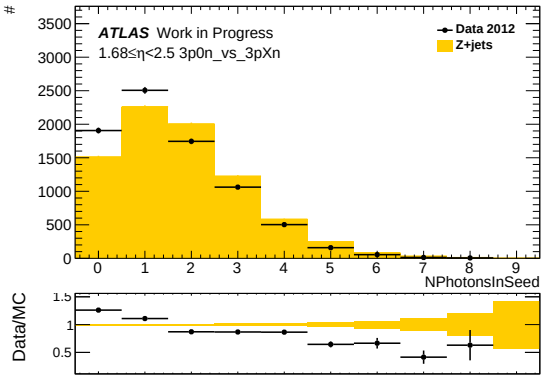
(a) 3p0n-vs.-3pXn BDT score for low η



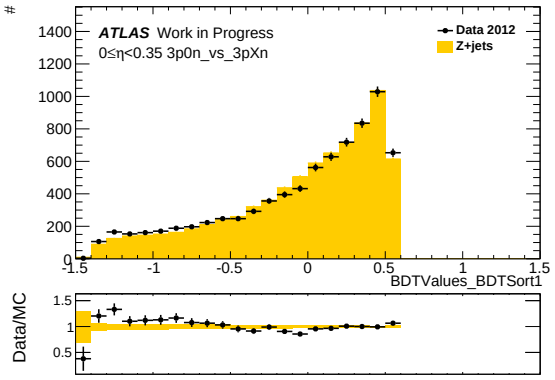
(b) 3p0n-vs.-3pXn BDT score for high η



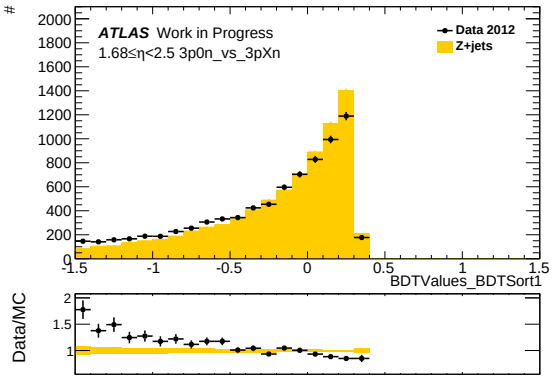
(c) `NPhotonsInSeed` for low η



(d) `NPhotonsInSeed` for high η



(e) `BDTSort1` for low η



(f) `BDTSort1` for high η

Figure 4.16: Figures (a-b) show the 3p0n-vs.-3pXn BDT score for the low and high η bin. Figures (c-f) show the variables that shows most disagreement in the high η bin.

4.6.2 Cell-based π^0 ID BDT Variables

In addition to the PanTau BDTs, also plots of the cell-based π^0 ID BDT variables have been made. These are not per tau candidate, but per Pi0Cluster. Since there are many Pi0Clusters in a tau, these plots have much higher statistics. This increased statistics is problematic for the χ^2 test of some variables. The relative errors are so small that the p value is calculated to 0 for every dependency bin. Of course, this happens preferably for variables that show some kind of disagreement and is also caused by the finite precision in computer calculations. Unfortunately, this is the case for most of the variables. A workaround to this has been done by using χ^2/ndf instead of the p value. Here, a higher value indicates a higher disagreement.

Inclusive distributions of selected variables used in the cell-based π^0 ID BDT are shown in figure 4.20. There are slopes visible in the Data/MC plot of most variables, with the exception of the variables `AsymmetryWRTTrack`, `FIRST_ETA`, and only a smaller slope at low values of `SECOND_R`, `secondEtaWRTClusterPosition_EM1/2`. The variable `ENG_FRAC_CORE` in figure 4.20b shows heavy disagreement and a strong slope in the Data/MC plot. However, the agreement of the BDT score is in contrast to that in an acceptable range (figure 4.18). The data distributions tend a bit more to extreme values at -1 and 1. The χ^2/ndf plots for the BDT score are shown in figure 4.17. They indicate disagreement for high η. This can also be seen for many of the variables used in the BDT. As an example, the χ^2/ndf plots for `AsymmetryWRTTrack` and `ENG_FRAC_CORE` are shown in figure 4.19. This is also visible in the distributions of the BDT score themselves. Figure 4.18 shows the distribution of the BDT score for the η bins with the best and worst χ^2/ndf. The agreement gets much worse for high η.

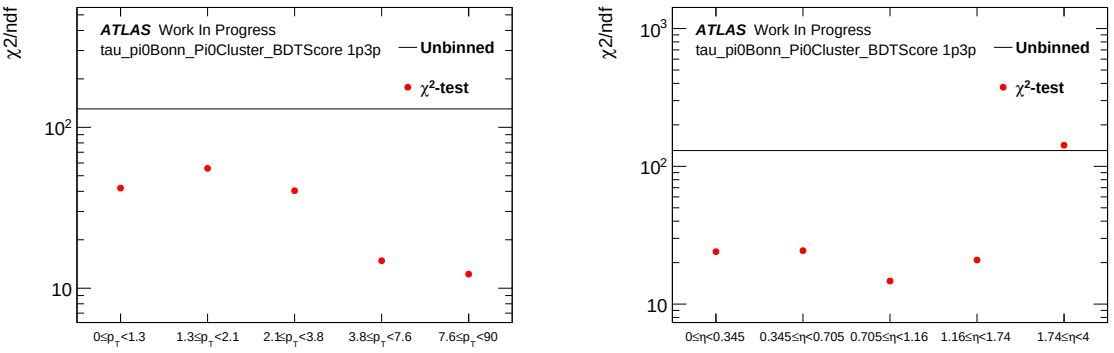
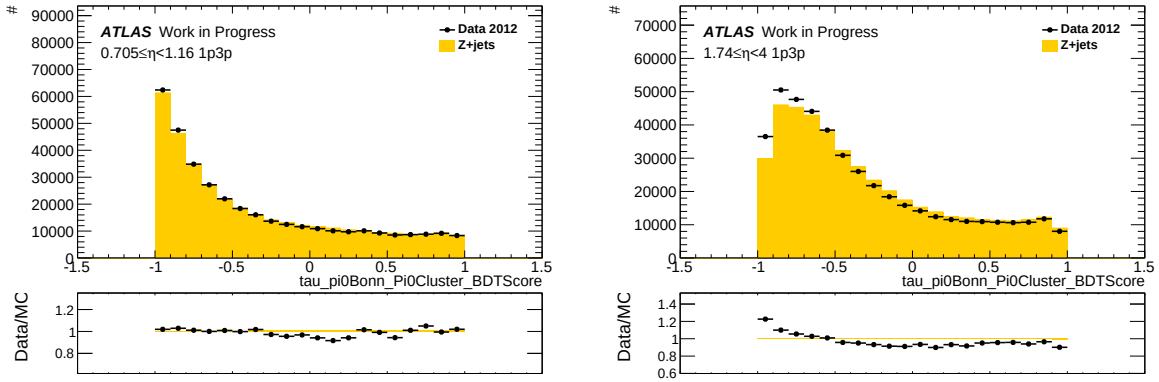


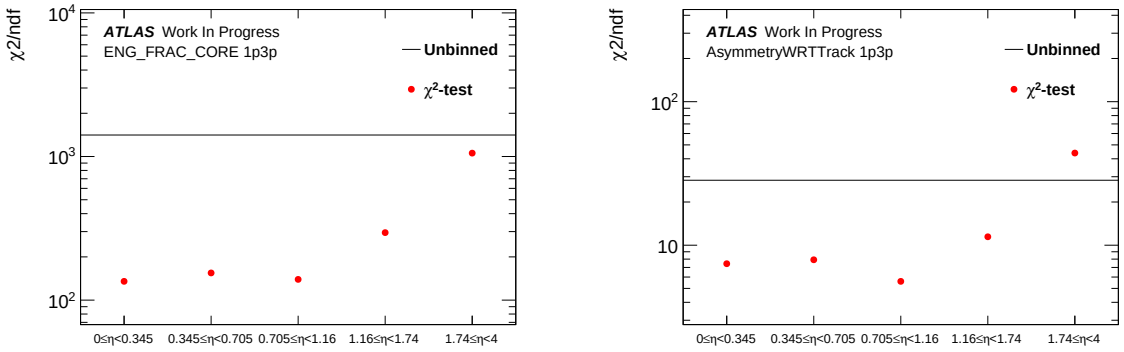
Figure 4.17: The χ^2/ndf plots for the BDT score of the cell-based π^0 ID BDT test.



(a) BDT score with best χ^2/ndf

(b) BDT score with worst χ^2/ndf

Figure 4.18: The BDT score distribution of the cell-based π^0 ID BDT test for the η bin with the best χ^2/ndf (a) and the worst (b).



(a) `ENG_FRAC_CORE`

(b) `AsymmetryWRTTrack`

Figure 4.19: χ^2/ndf plots for two variables that indicate disagreement at high p_{T}.

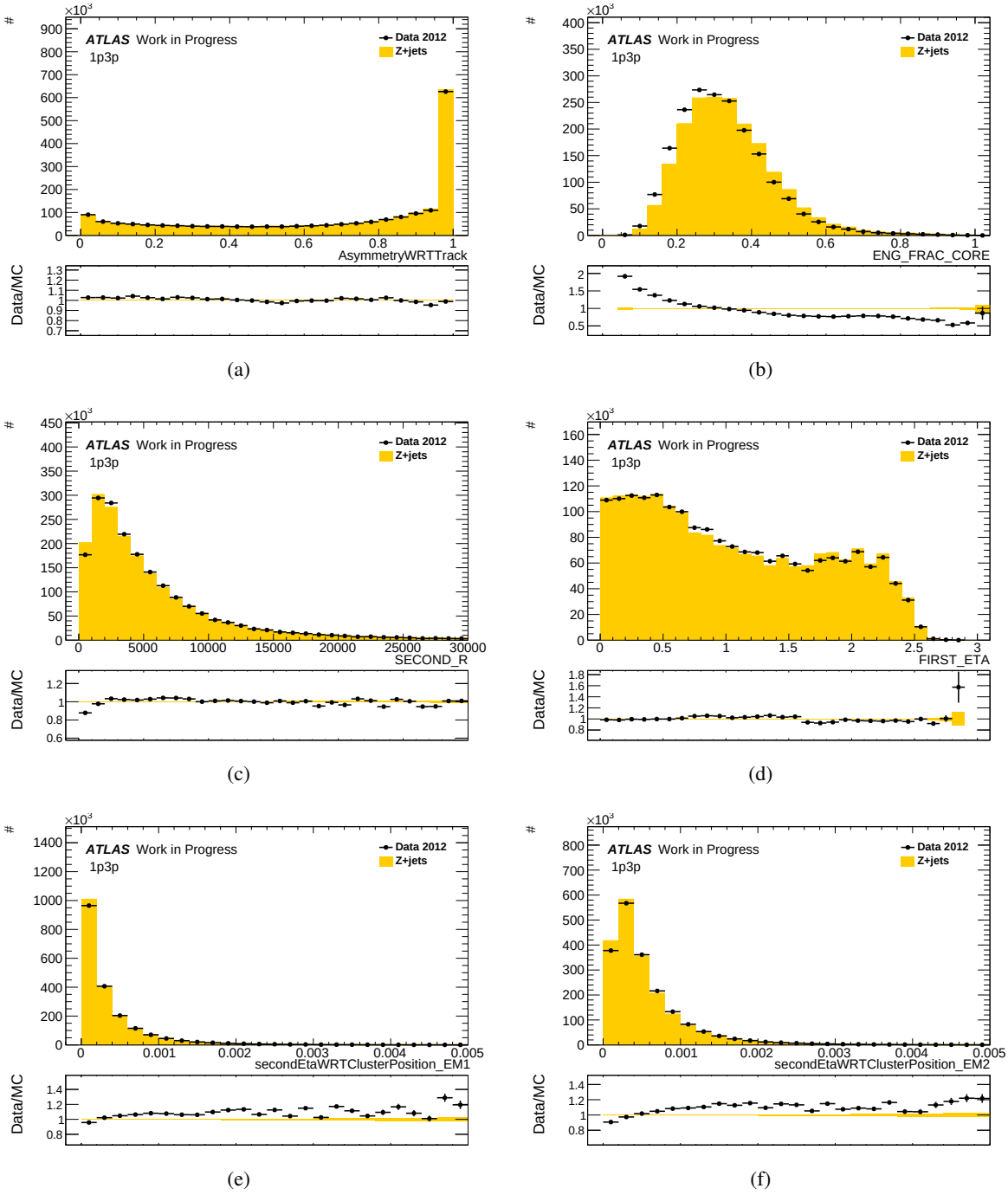
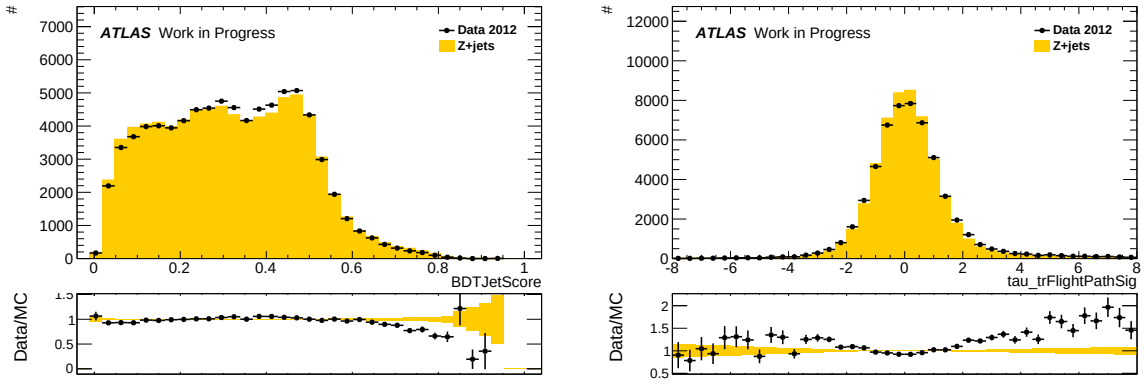


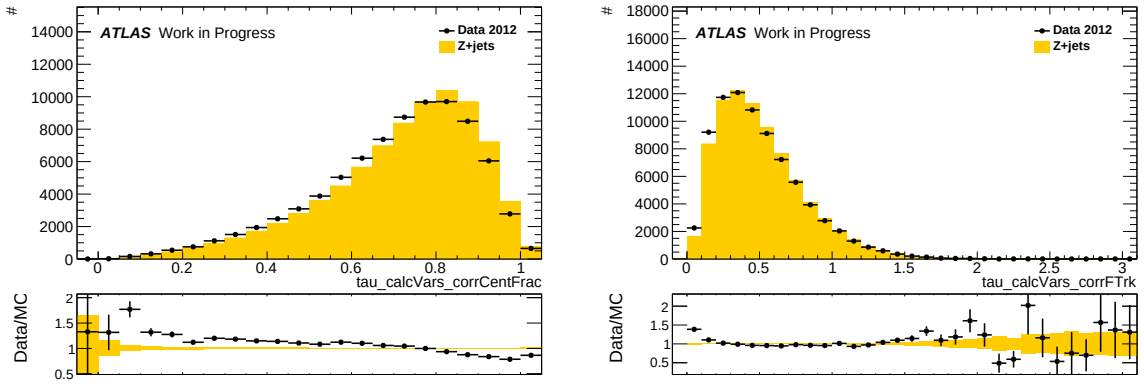
Figure 4.20: Inclusive distributions of selected variables used for the π^0 ID of cell-based.

4.6.3 Tau ID Variables

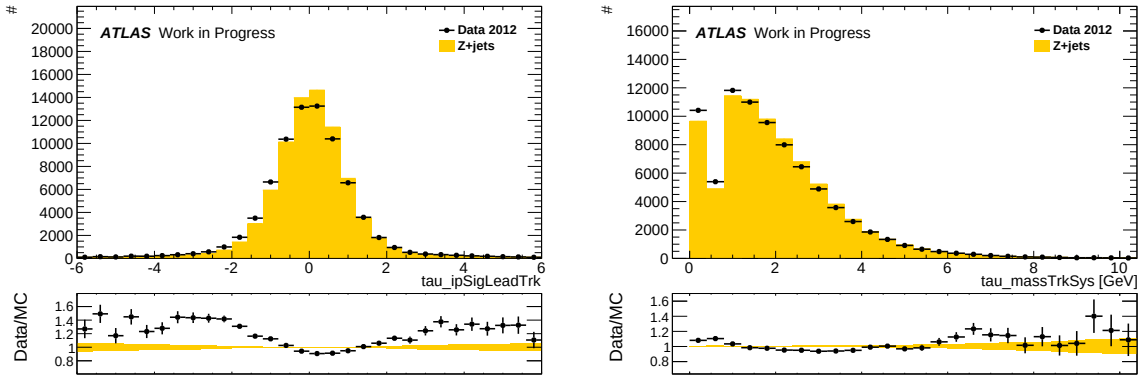
At last the variables used for tau identification were studied. Inclusive distributions for selected variables used in the TauID BDT are shown in figure 4.21. The BDT score itself shows a rather good agreement except for high values. It can be seen that the distribution flattens out at high values, as these mark tau-like jets. The peak shape of the variables `ipSigLeadTrk` and `trFlightPathSig` is significantly wider in data. The variable `corrCentFrac` shows slope in the Data/MC plot. `corrFTrk` and `massTrkSys` have some disagreement at low values. A low p value for the high p_{T} bin can be seen for the variables `corrCentFrac` (figure 4.22b), `pi0_vistau_m` (figure 4.22c) and `trkAvgDist` (figure 4.22d). However, the p value plot of the BDT score in figure 4.22a does not show that for the high p_{T} bin. All inclusive distributions and p value plots are listed in the appendix A.3.



(a)

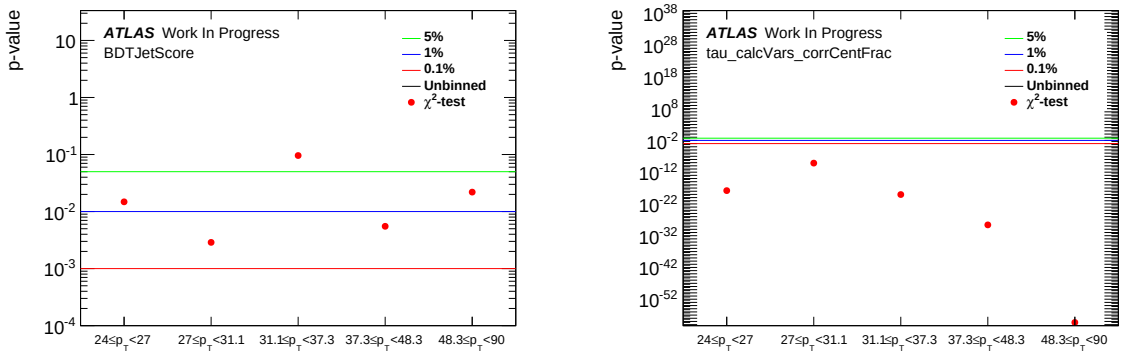


(b)



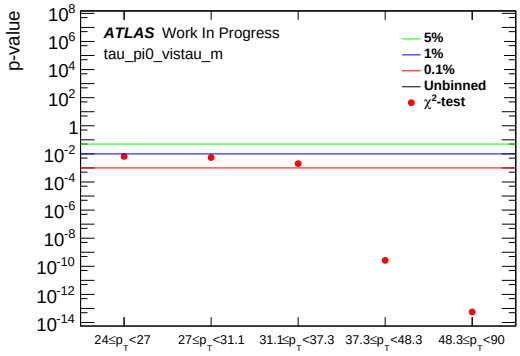
(c)

Figure 4.21: Inclusive distributions of the TauID BDT score and selected variables.

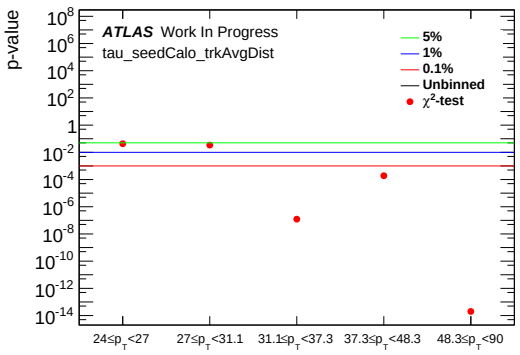


(a) BDTJetScore

(b) `corrCentFrac`



(c) `pi_vistau_m`



(d) `trkAvgDist`

Figure 4.22: Figure (a) shows the p values TauID BDT score for p_{T} bins. Figures (b-c) show variables with disagreement in the high p_{T} bin. p_{T} is in GeV.

CHAPTER 5

Summary and Conclusion

A Data/Monte-Carlo comparison of variables used in the BDTs of PanTau, cell-based and for TauID has been presented for jets in $Z \rightarrow \mu\mu$ events. The distribution agreement has been evaluated with a χ^2 test. This has been done inclusively as well as binned in p_{T} and η of the tau or Pi0Cluster respectively.

For the PanTau 3p0n-vs.-3pXn BDT, disagreement for high η and low p_{T} has been seen. In two of the variables used, `NPhotonsInSeed` and `BDTSort1` the same behaviour has been discovered. However, in general the variables used in PanTau show a reasonable agreement. It has been shown that many of the variables used for the cell-based π^0 ID BDT show disagreement. However the resulting disagreement of the BDT score itself is moderate, except for high η. For the TauID BDT score, the p value plots indicate a reasonable agreement for all η and p_{T} bins. But again, many variables used in that BDT do show significant disagreement.

The χ^2 test has proven to be a valuable method in indicating disagreeing distributions and trend visualisation. However, the sensitivity of the test for low statistics bins can be problematic since results become less reliable.

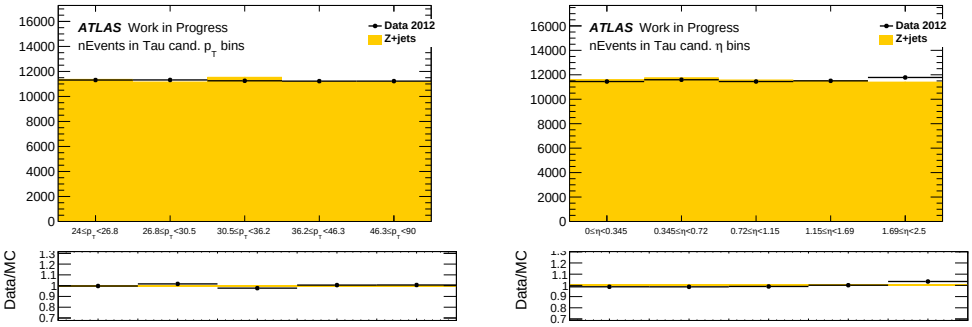
The comparison presented here will be useful for the development of the tau reconstruction development. It helps in choosing the right variables for a reliable reconstruction. Since the reconstruction algorithms have been developed using simulated events, it also validates proper operation on data.

APPENDIX A

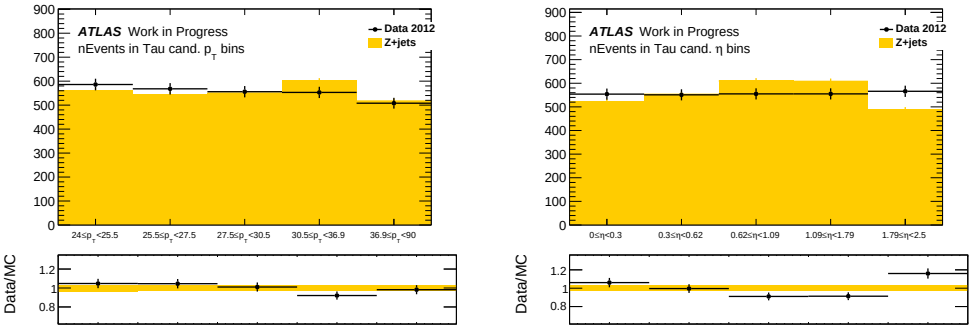
Appendix

Here all p value/χ^2/ndf plots are shown as well as inclusive distributions of all variables. Plots of the statistics per dependency bin are also shown.

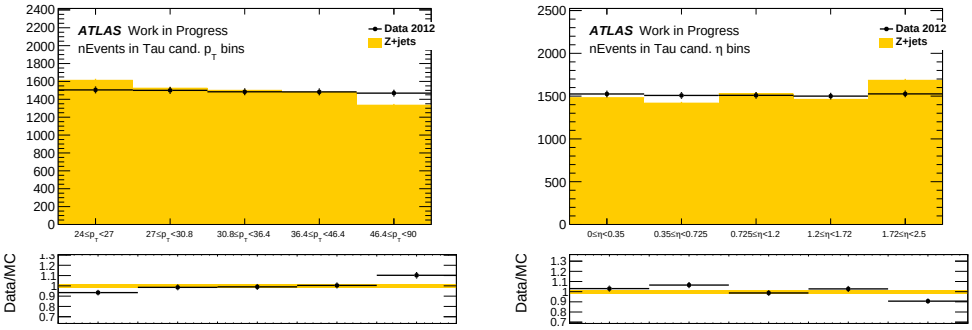
A.1 PanTau



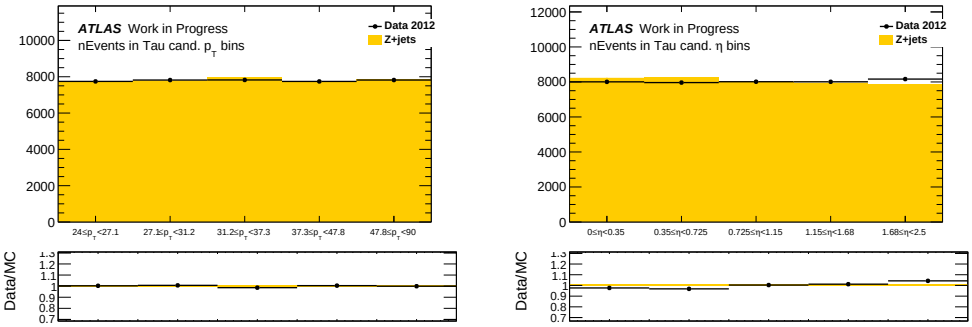
(a) All tau candidates



(b) 1p0n vs. 1p1n

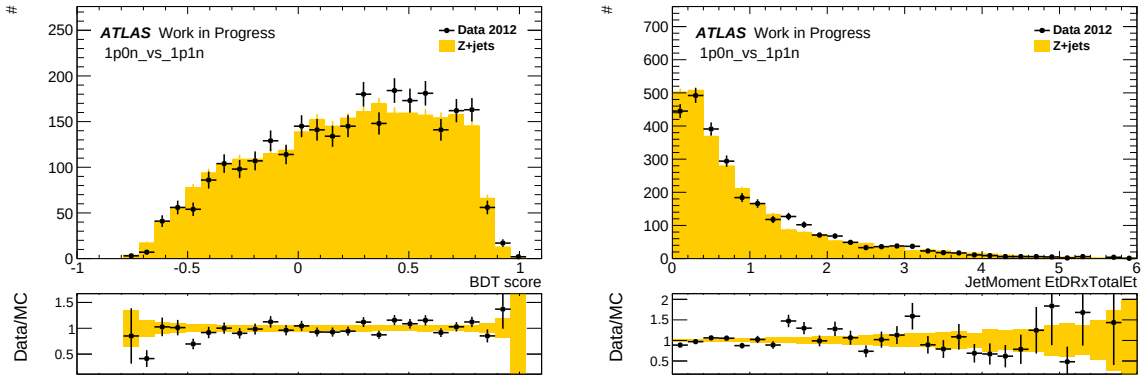


(c) 1p1n vs. 1pXn

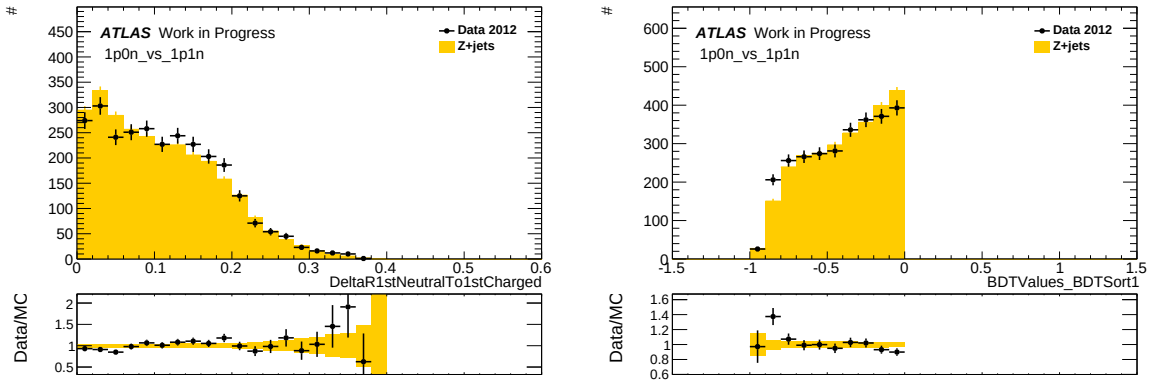


(d) 3p0n vs. 3pXn

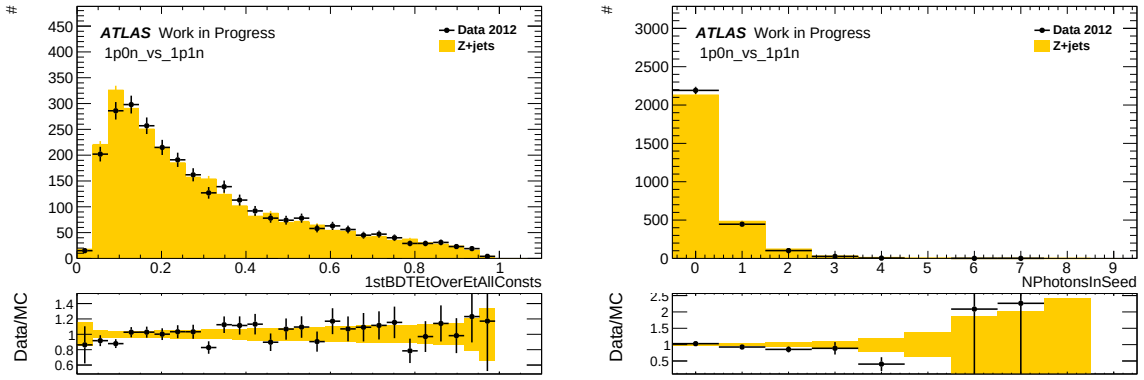
Figure A.1: The number of events in the p_T and η bins for all tau candidates and per BDT



(a)

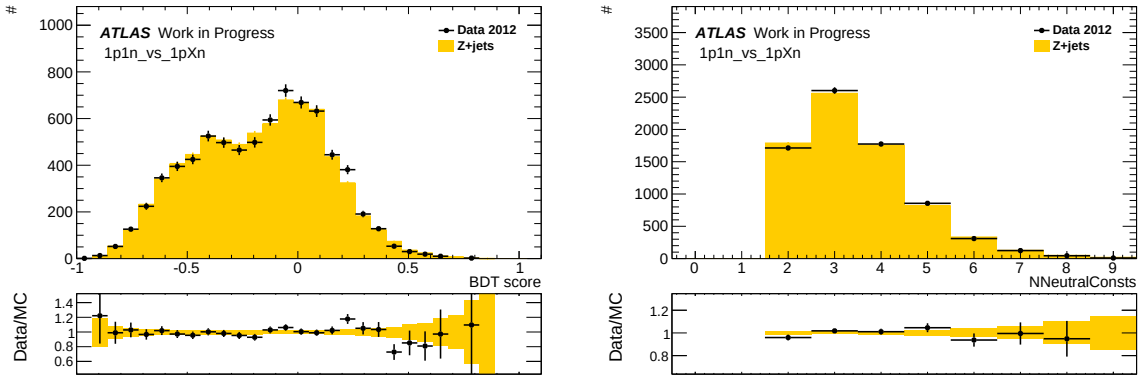


(b)

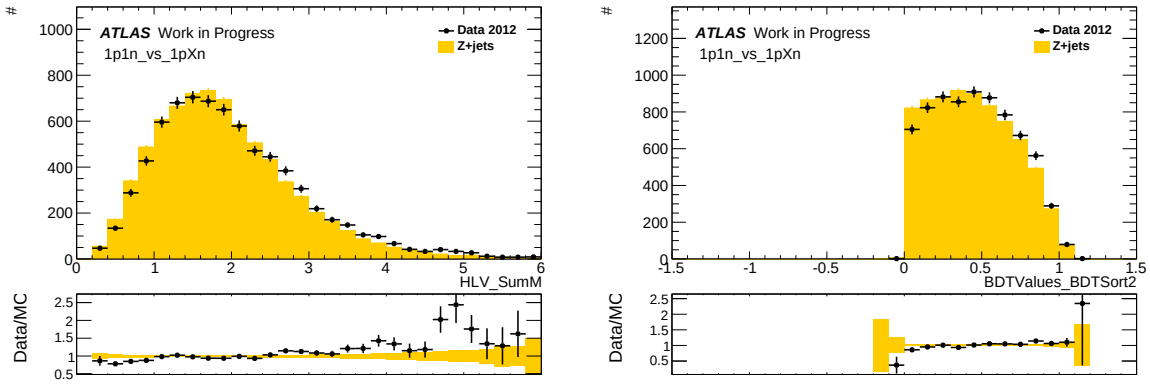


(c)

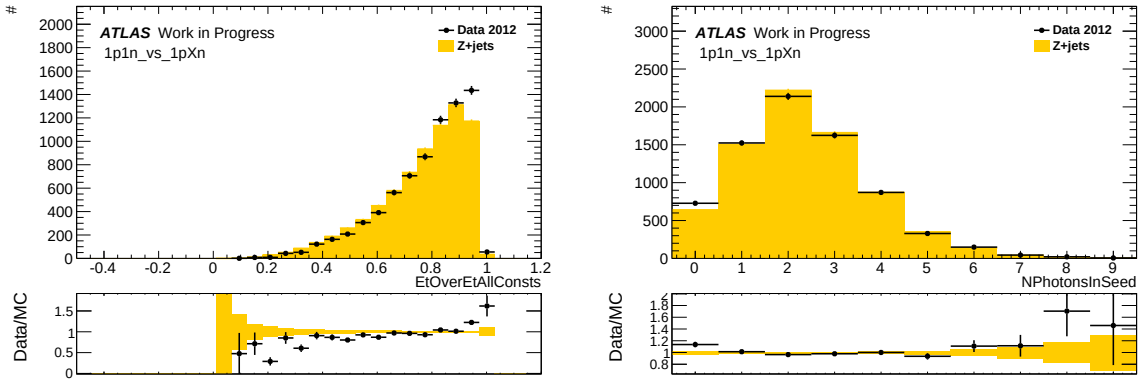
Figure A.2: Inclusive distributions of the variables used for the 1p0n-vs.-1p1n BDT.



(a)

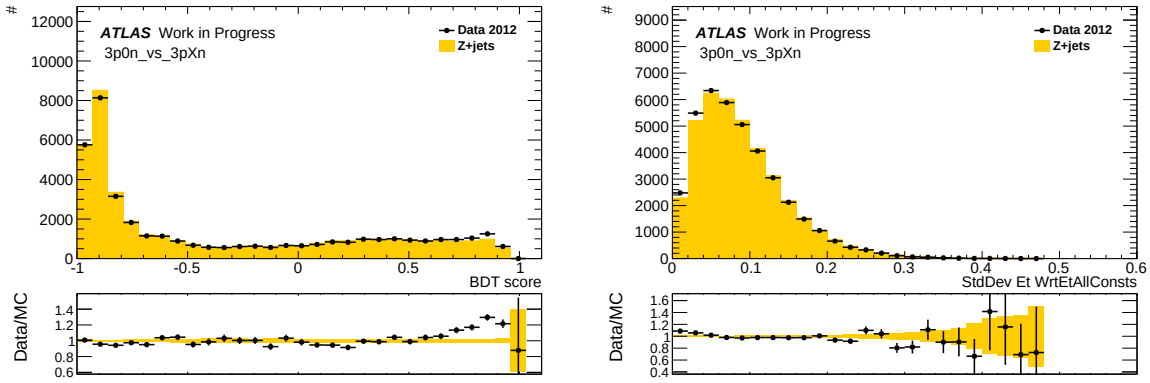


(b)

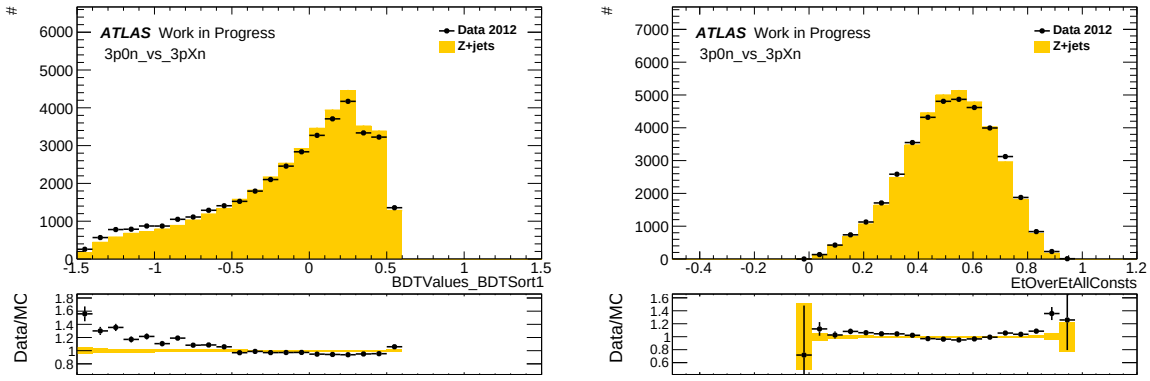


(c)

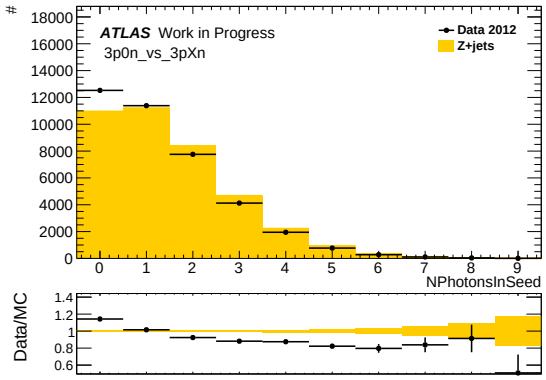
Figure A.3: Inclusive distributions of the variables used for the 1p1n-vs.-1pXn BDT.



(a)



(b)



(c)

Figure A.4: Inclusive distributions of the variables used for the 3p0n-vs.-3pXn BDT.

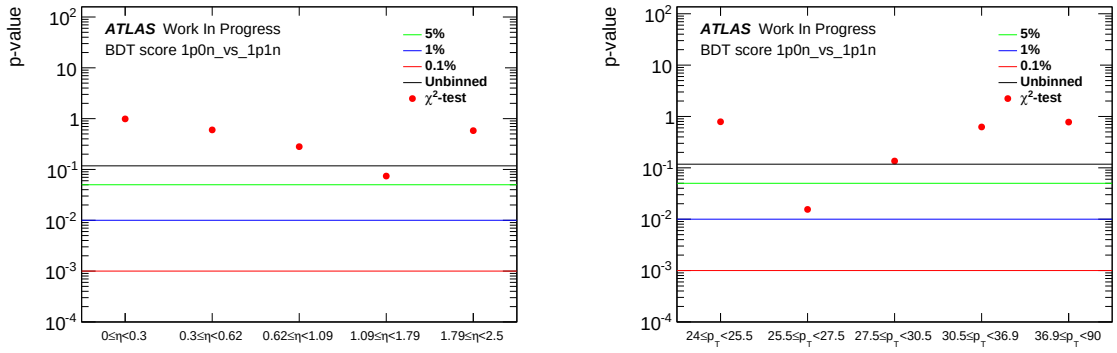


Figure A.5: P value plots for the BDT score of the 1p0n-vs.1p1n test. p_{T} is in GeV, η bins doalso include negative values.

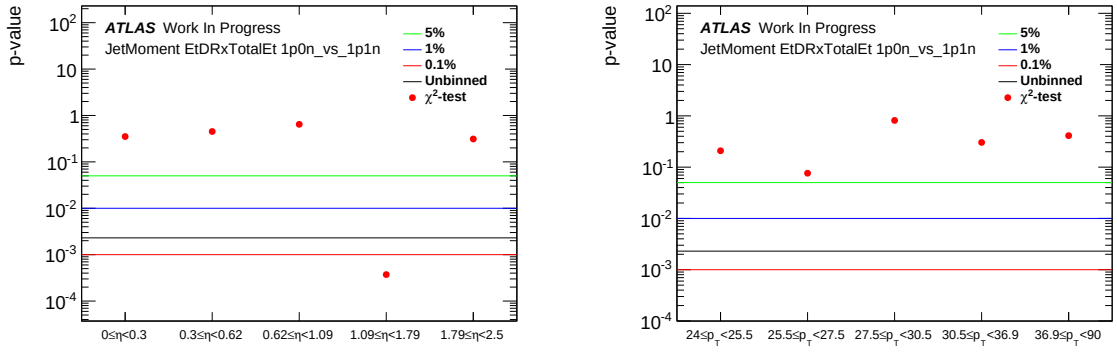


Figure A.6: P value plots for the variable `EtDRxTotalEt` of the 1p0n-vs.-1p1n test. p_{T} is in GeV, η bins do also include negative values.

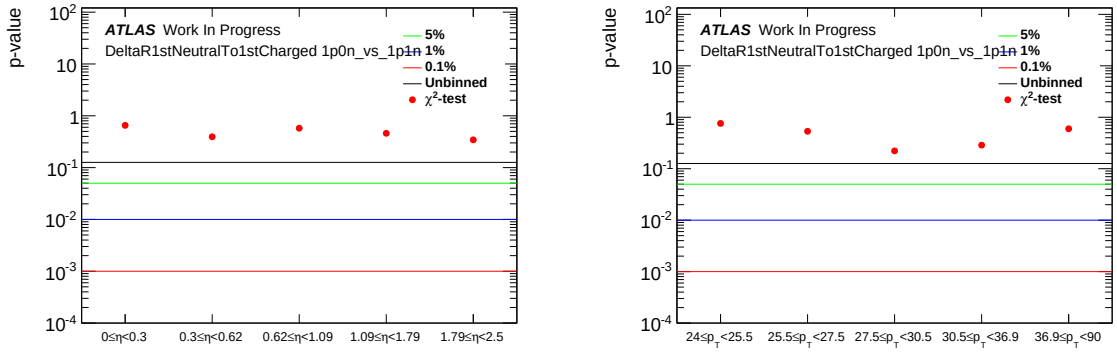


Figure A.7: P value plots for the variable `DeltaR1stNeutralTo1stCharged` of the 1p0n-vs.-1p1n test. p_{T} is in GeV, η bins do also include negative values.

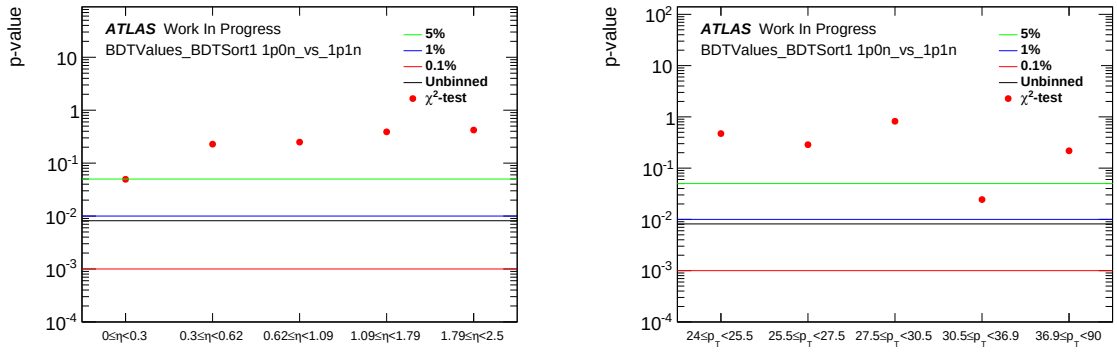


Figure A.8: P value plots for the variable `BDTSort1` of the 1p0n-vs.-1p1n test. p_{T} is in GeV, η bins do also include negative values.

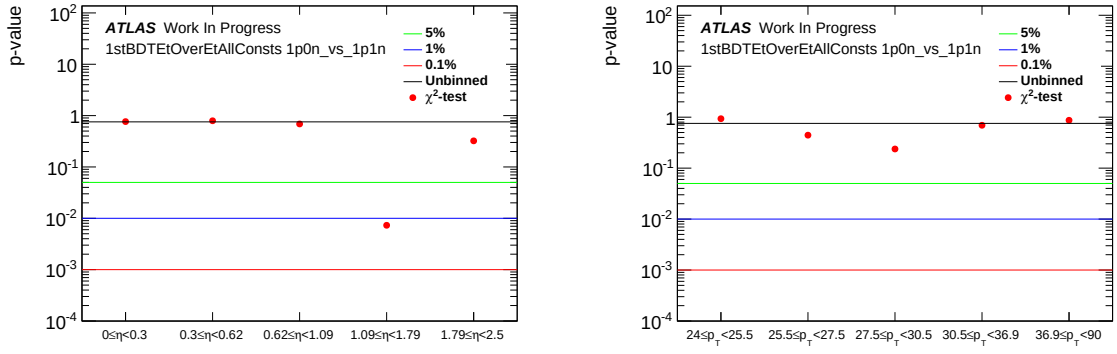


Figure A.9: P value plots for the variable `1stBDTEtOverEtAllConsts` of the 1p0n-vs.-1p1n test. p_{T} is in GeV, η bins do also include negative values.

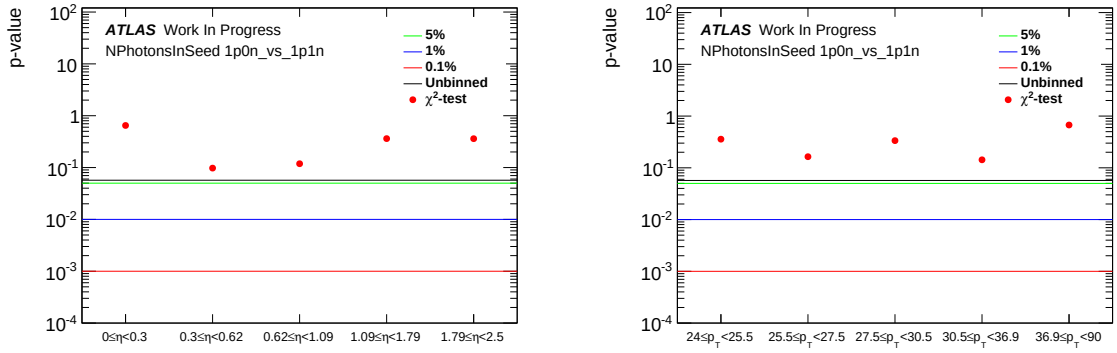


Figure A.10: P value plots for the variable `NPhotonsInSeed` of the 1p0n-vs.-1p1n test. p_{T} is in GeV, η bins do also include negative values.

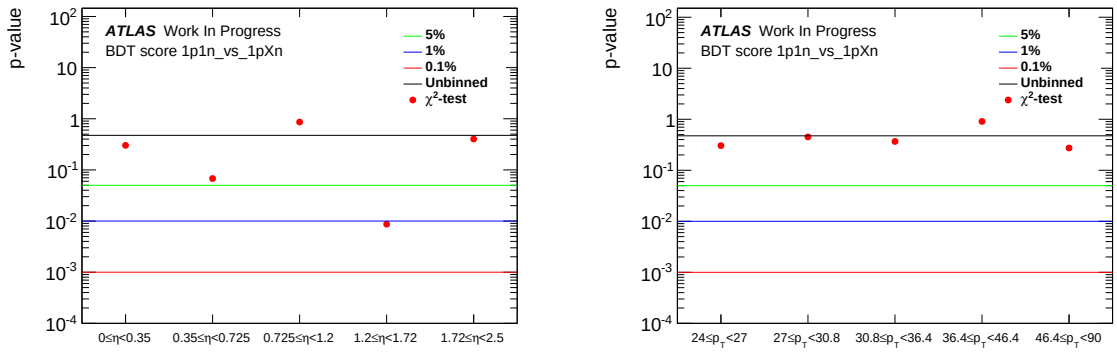


Figure A.11: P value plots for the BDT score of the 1p1n-vs.1pXn test. p_{T} is in GeV, η bins do also include negative values.

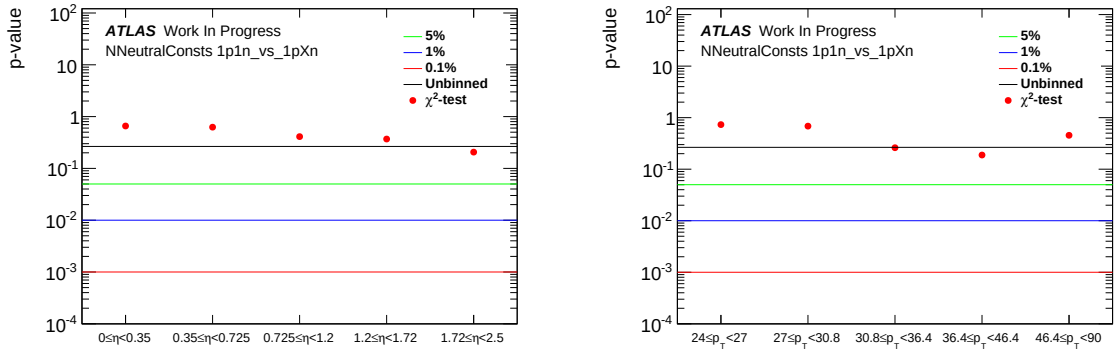


Figure A.12: P value plots for the variable `NNeutralConsts` of the 1p1n-vs.-1pXn test. p_{T} is in GeV, η bins do also include negative values.

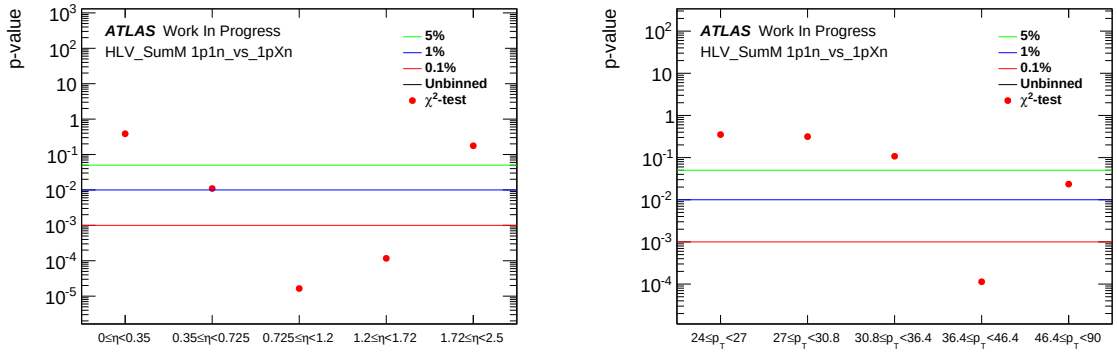


Figure A.13: P value plots for the variable `HLV_SumM` of the 1p1n-vs.-1pXn test. p_{T} is in GeV, η bins do also include negative values.

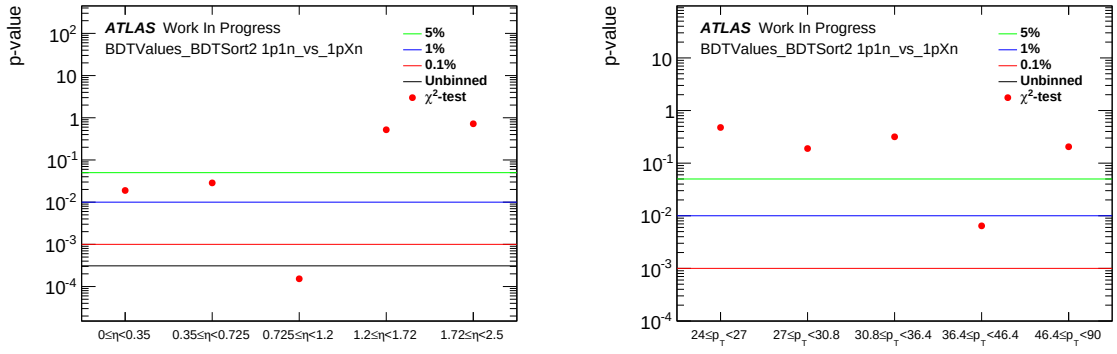


Figure A.14: P value plots for the variable `BDTSort2` of the 1p1n-vs.-1pXn test. p_{T} is in GeV, η bins do also include negative values.

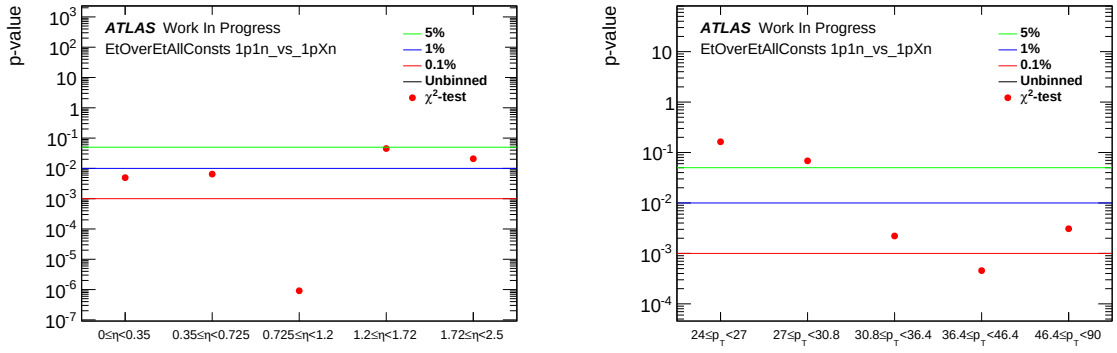


Figure A.15: P value plots for the variable `EtOverEtAllConsts` of the 1p1n-vs.-1pXn test. p_{T} is in GeV, η bins do also include negative values.

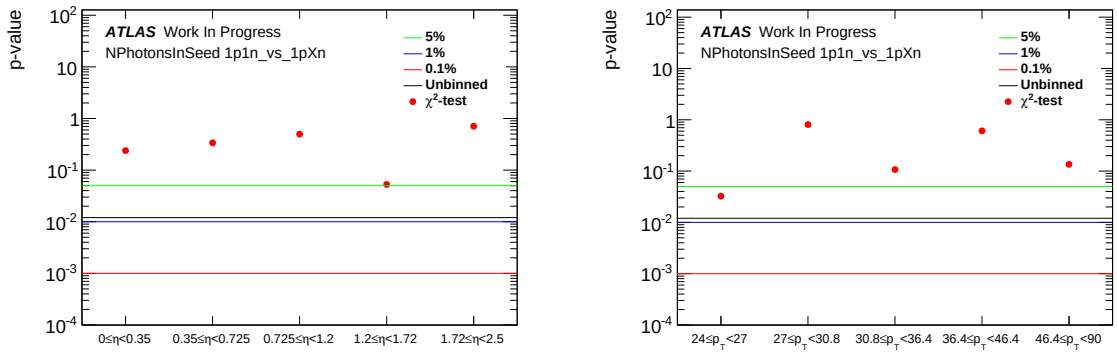


Figure A.16: P value plots for the variable `NPhotonsInSeed` of the 1p1n-vs.-1pXn test. p_{T} is in GeV, η bins do also include negative values.

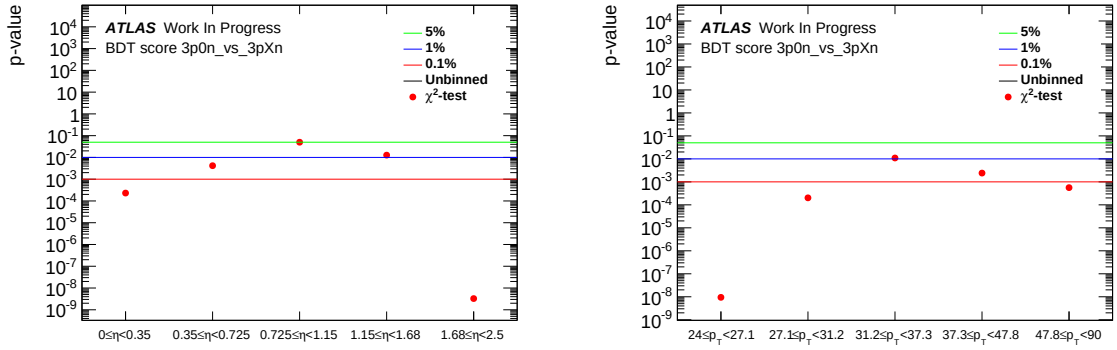


Figure A.17: P value plots for the BDT score of the 3p0n-vs.-3pXn test. p_{T} is in GeV, η bins do also include negative values.

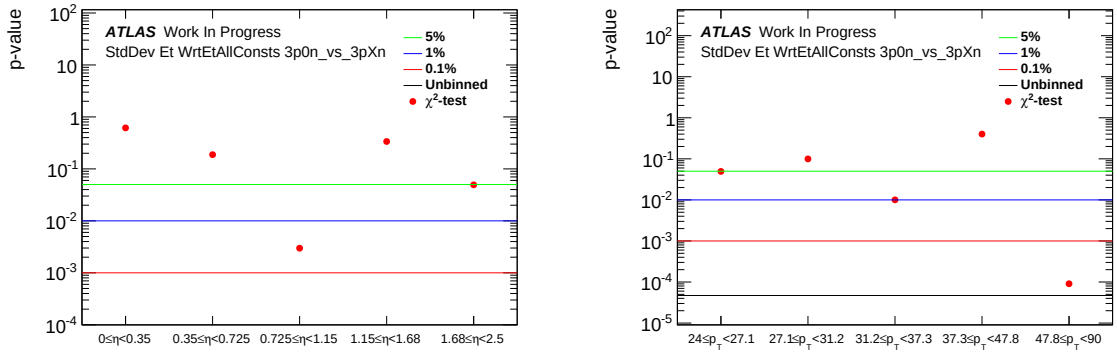


Figure A.18: P value plots for the variable `StdDev_Et_WrtEtAllConsts` of the 3p0n-vs.-3pXn test. p_{T} is in GeV, η bins do also include negative values.

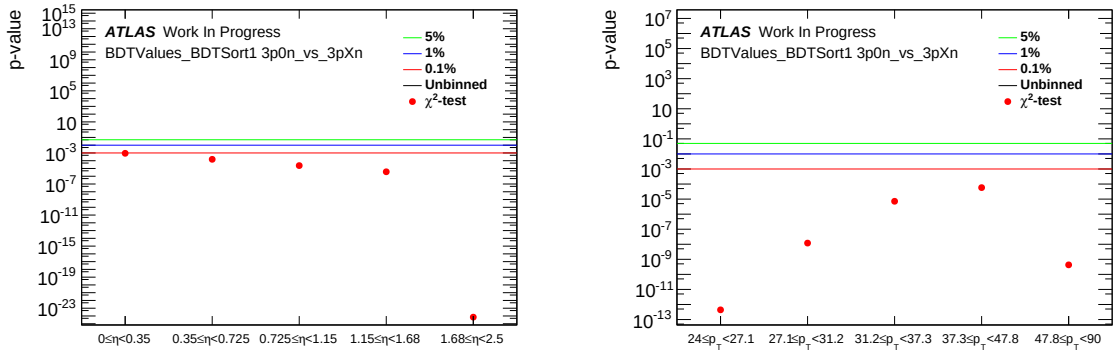


Figure A.19: P value plots for the variable `BDTSort1` of the 3p0n-vs.-3pXn test. p_{T} is in GeV, η bins do also include negative values.

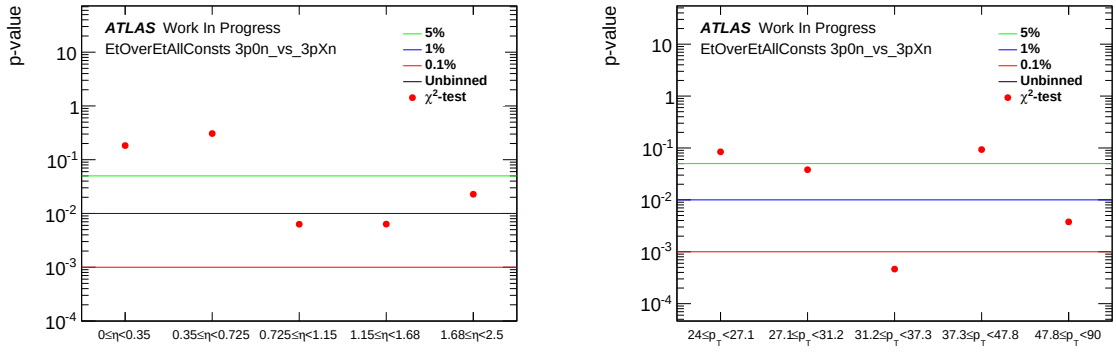


Figure A.20: P value plots for the variable `EtOverEtAllConsts` of the 3p0n-vs.-3pXn test. p_{T} is in GeV, η bins do also include negative values.

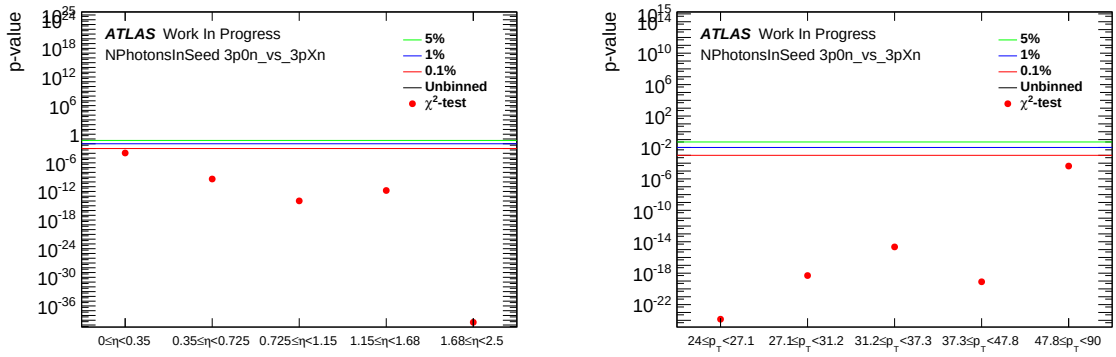
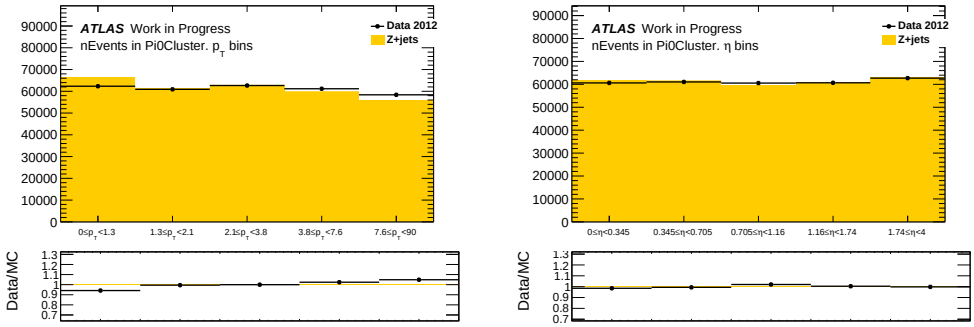


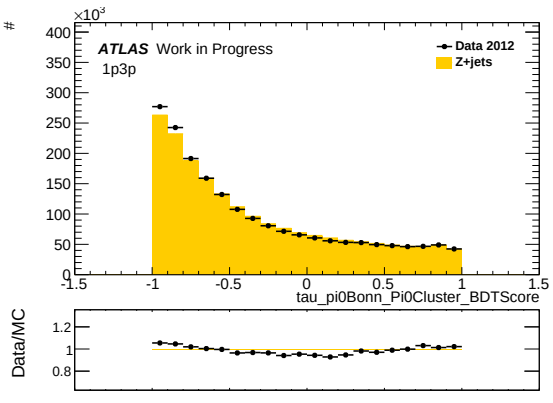
Figure A.21: P value plots for the variable `NPhotonsInSeed` of the 3p0n-vs.-3pXn test. p_{T} is in GeV, η bins do also include negative values.

A.2 Cell-based



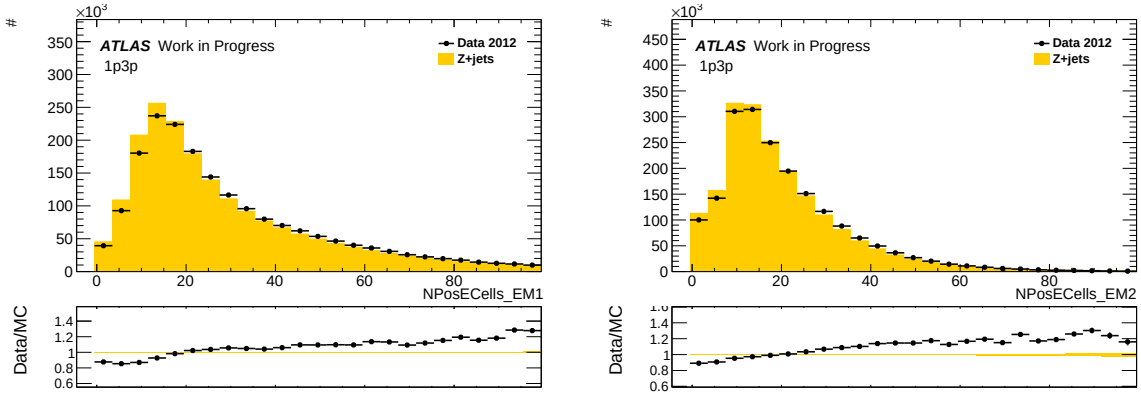
(a) All tau candidates

Figure A.22: The number of events in the p_T and η bins for all Pi0Clusters

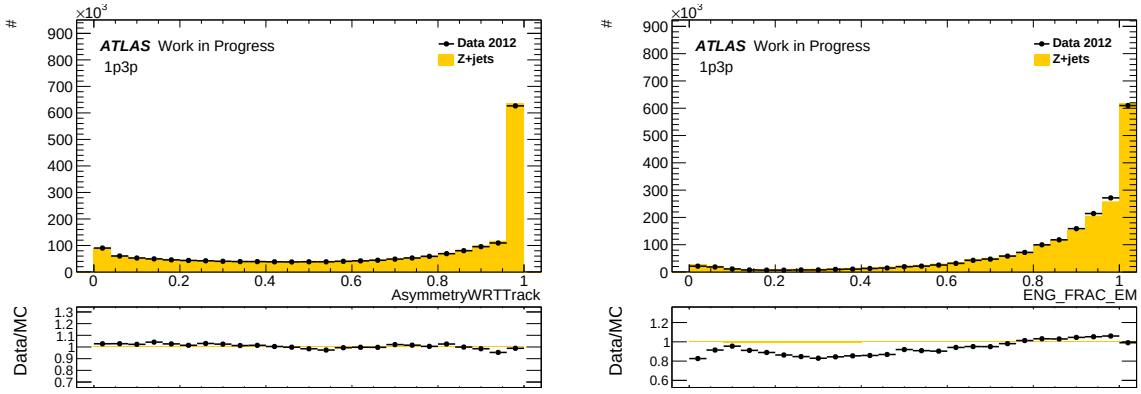


(a)

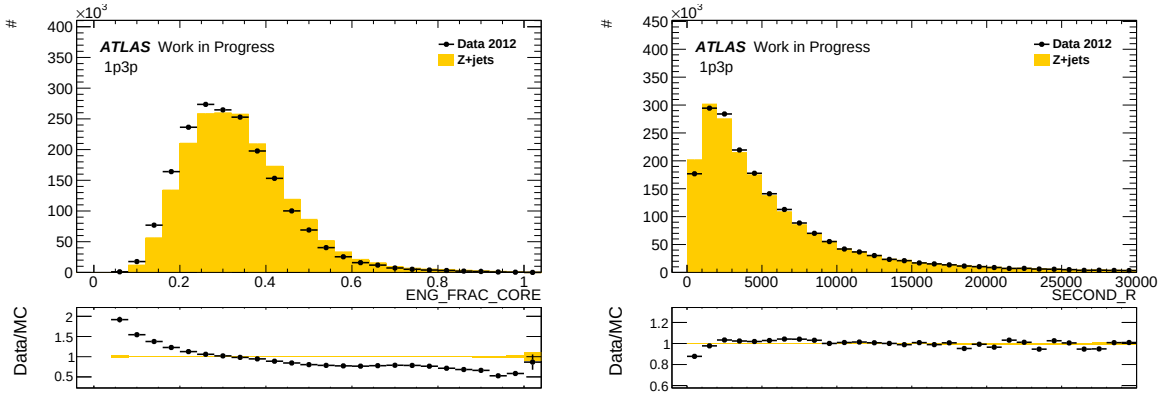
Figure A.23: Inclusive distribution π^0 ID BDT score of cell-based.



(a)

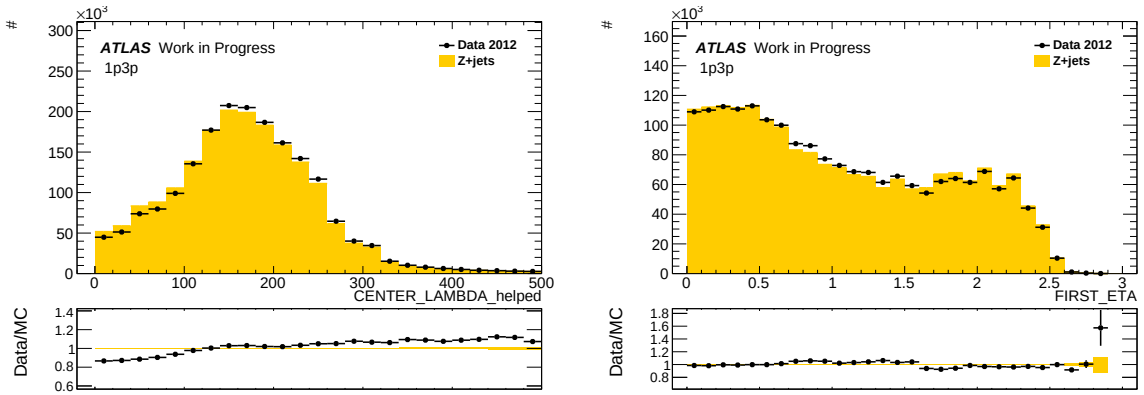


(b)

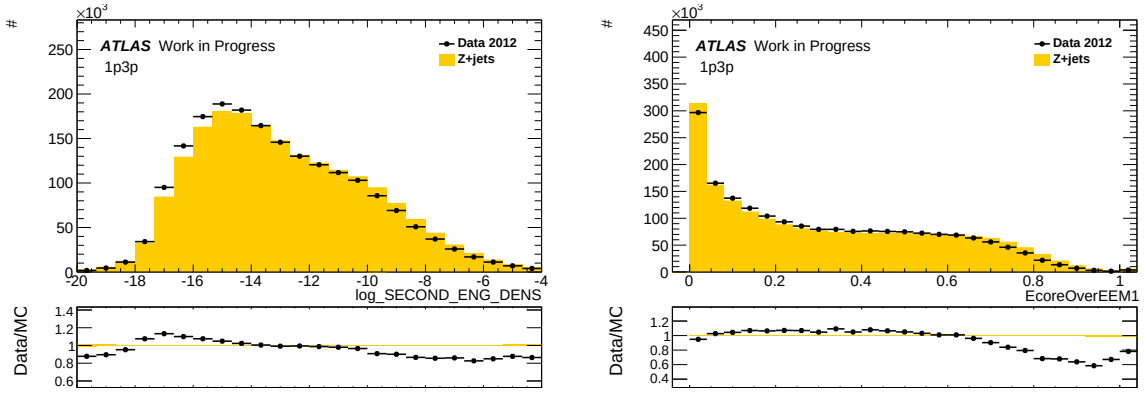


(c)

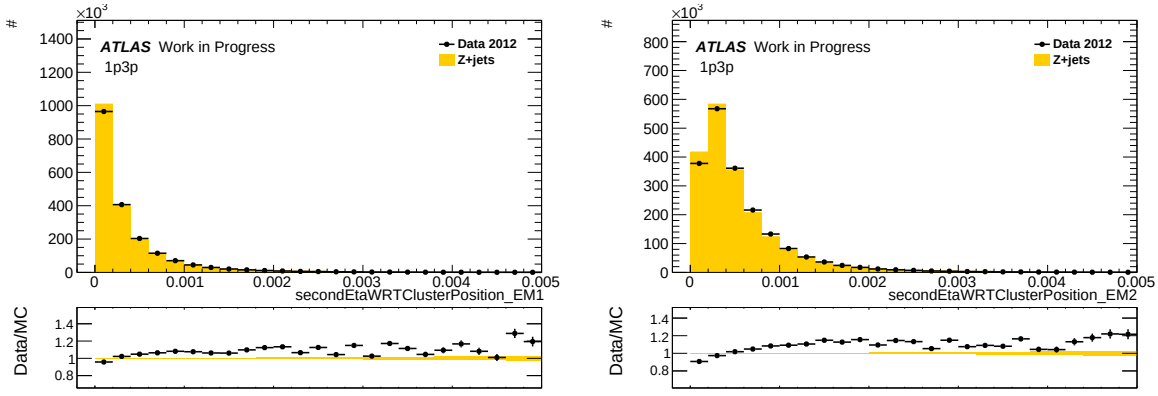
Figure A.24: Inclusive distributions of the variables used for the π^0 ID of cell-based.



(a)

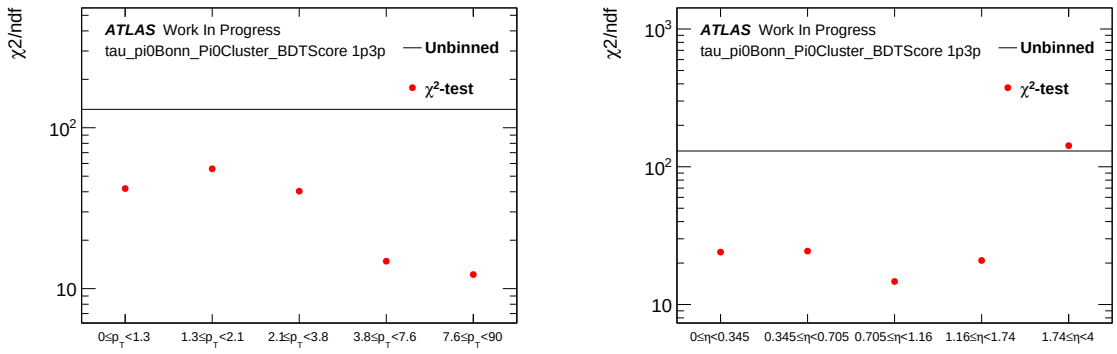


(b)



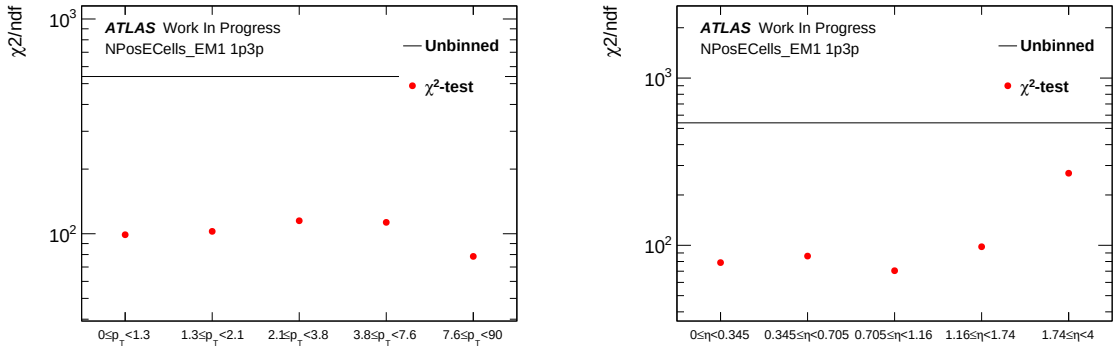
(c)

Figure A.25: Inclusive distributions of the variables used for the π^0 ID of cell-based.



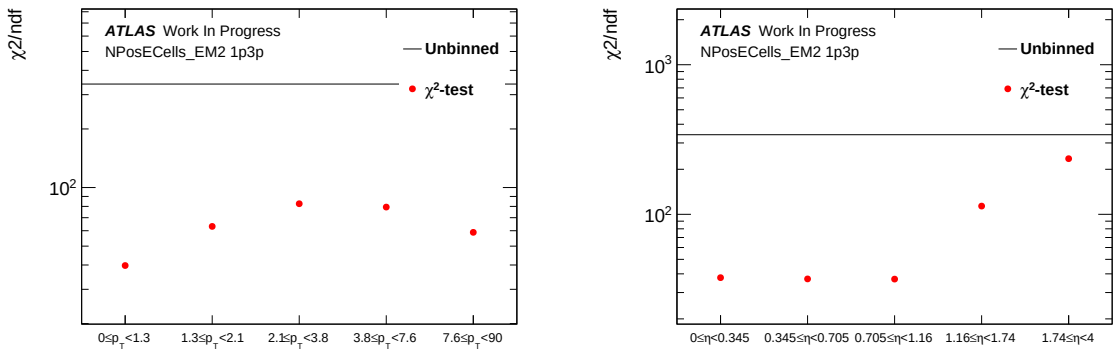
(a)

Figure A.26: χ^2/ndf plots of the BDT score of the π^0 ID of cell-based. Pi0Cluster p_{T} is in GeV, η bins do also include negative values.



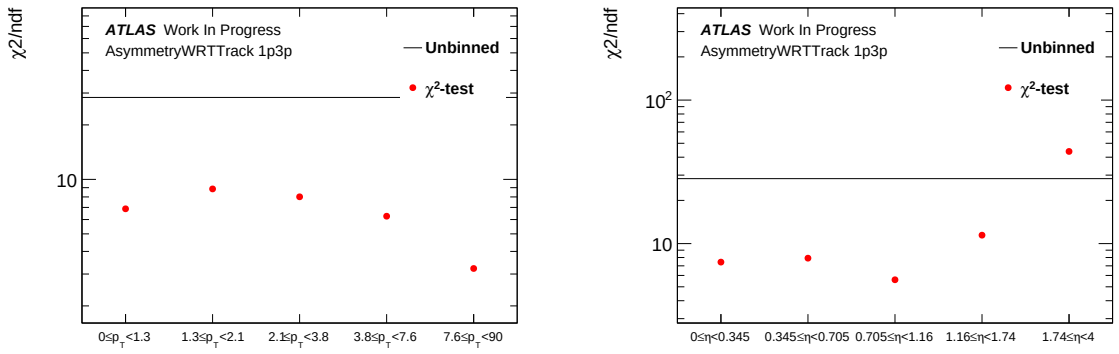
(a)

Figure A.27: χ^2/ndf plots of the variable `NPosECells_EM1` used for the π^0 ID of cell-based. Pi0Cluster p_{T} is in GeV, η bins do also include negative values.



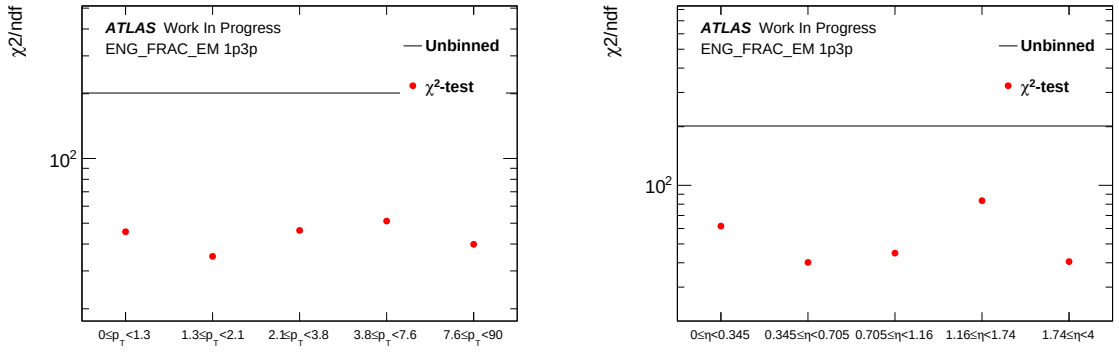
(a)

Figure A.28: χ^2/ndf plots of the variable `NPosECells_EM2` used for the π^0 ID of cell-based. Pi0Cluster p_{T} is in GeV, η bins do also include negative values.



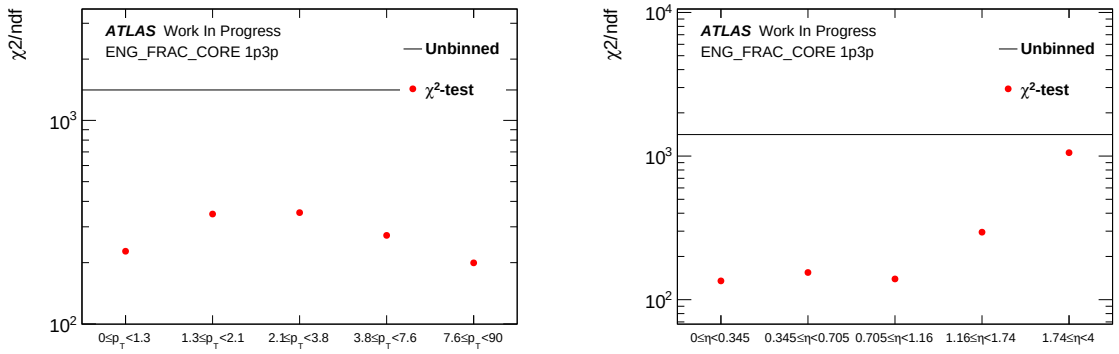
(a)

Figure A.29: χ^2/ndf plots of the variable `AsymmetryWRTTrack` used for the π^0 ID of cell-based. Pi0Cluster p_T is in GeV, η bins do also include negative values.



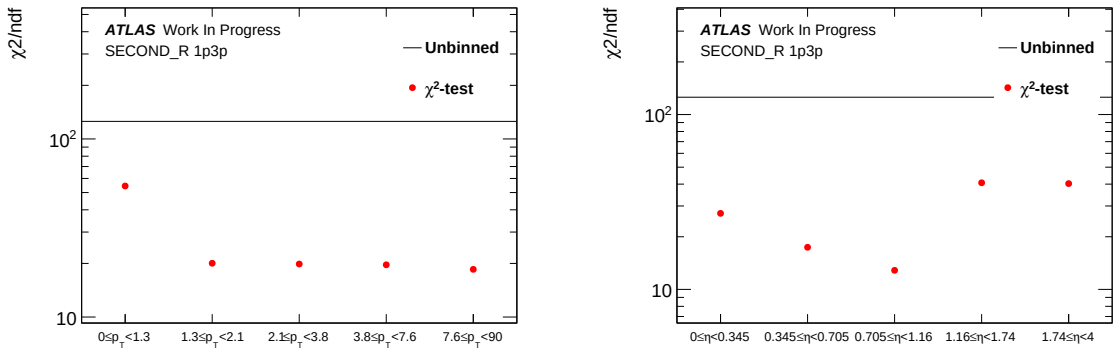
(a)

Figure A.30: χ^2/ndf plots of the variable `ENG_FRAC_EM` used for the π^0 ID of cell-based. Pi0Cluster p_T is in GeV, η bins do also include negative values.



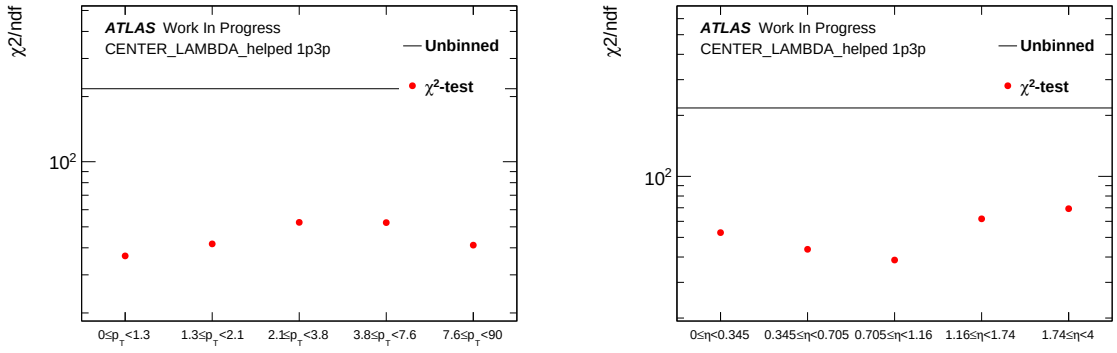
(a)

Figure A.31: χ^2/ndf plots of the variable `ENG_FRAC_CORE` used for the π^0 ID of cell-based. Pi0Cluster p_T is in GeV, η bins do also include negative values.



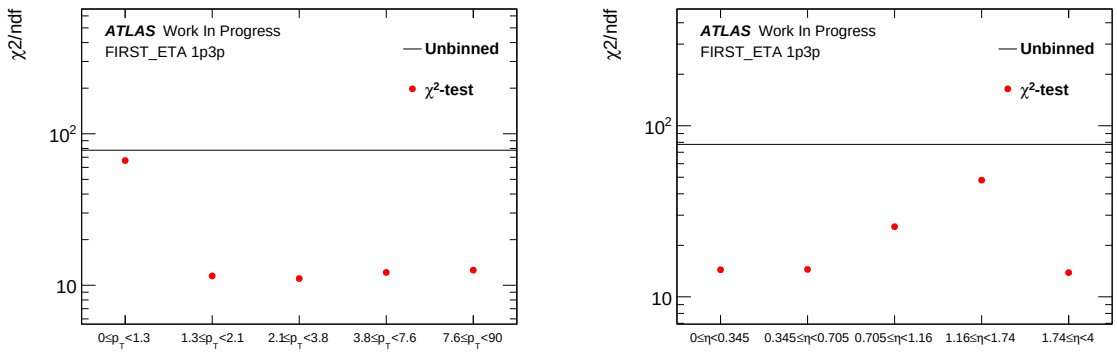
(a)

Figure A.32: χ^2/ndf plots of the variable `SECOND_R` used for the π^0 ID of cell-based. Pi0Cluster p_{T} is in GeV, η bins do also include negative values.



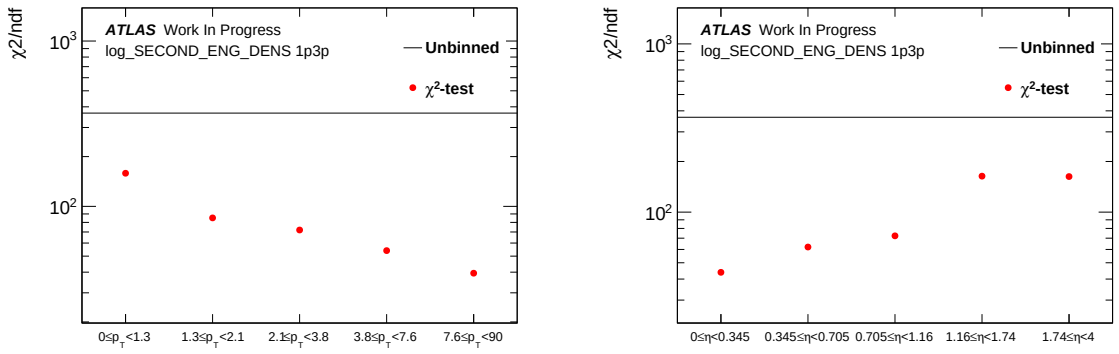
(a)

Figure A.33: χ^2/ndf plots of the variable `CENTER_LAMBDA_helped` used for the π^0 ID of cell-based. Pi0Cluster p_{T} is in GeV, η bins do also include negative values.



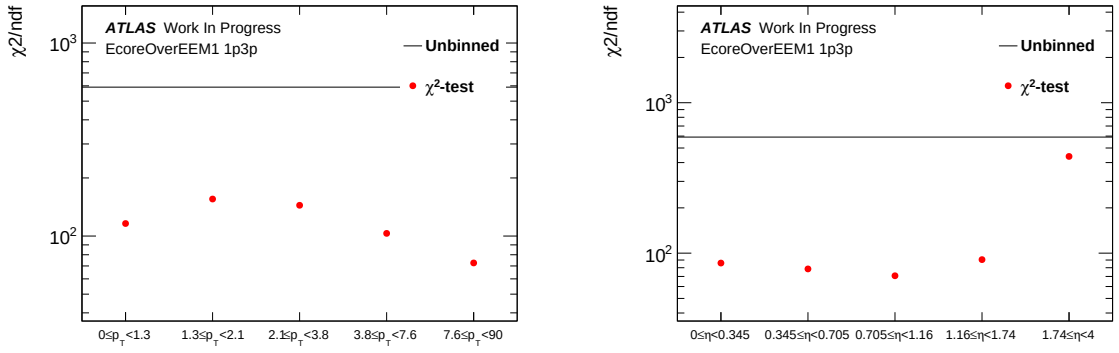
(a)

Figure A.34: χ^2/ndf plots of the variable `FIRST_ETA` used for the π^0 ID of cell-based. Pi0Cluster p_{T} is in GeV, η bins do also include negative values.



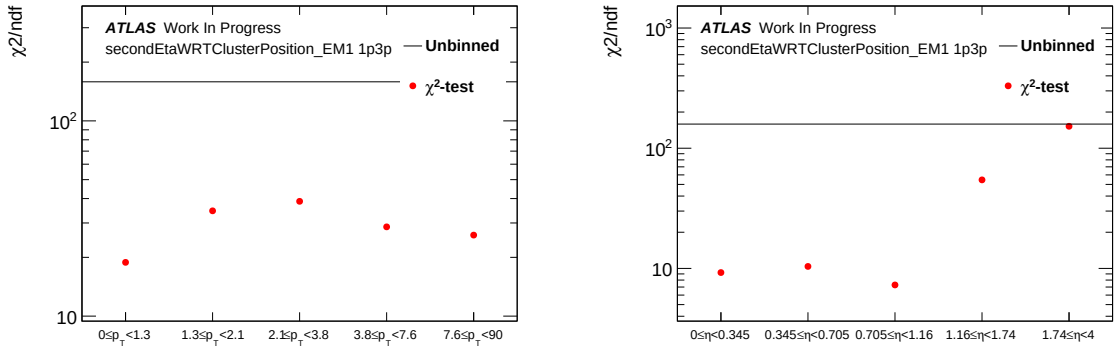
(a)

Figure A.35: χ^2/ndf plots of the variable `log_SECOND_ENG_DENS` used for the π^0 ID of cell-based. Pi0Cluster p_{T} is in GeV, η bins do also include negative values.



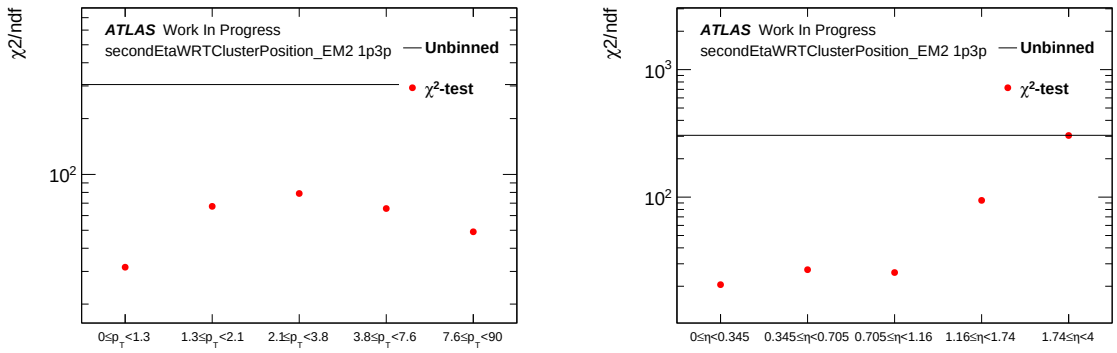
(a)

Figure A.36: χ^2/ndf plots of the variable `EcoreOverEEM1` used for the π^0 ID of cell-based. Pi0Cluster p_{T} is in GeV, η bins do also include negative values.



(a)

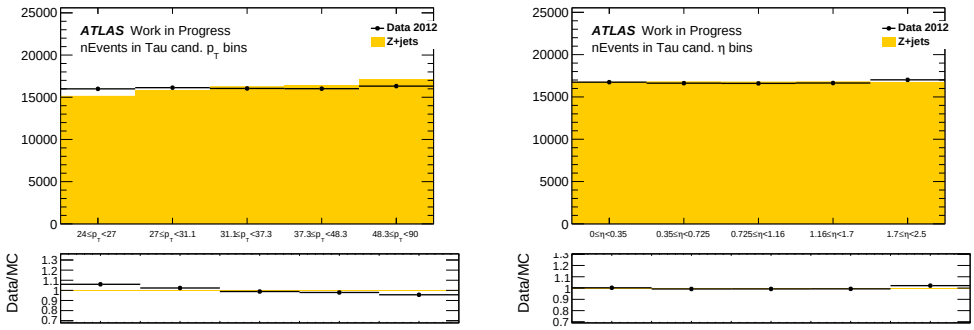
Figure A.37: χ^2/ndf plots of the variable `secondEtaWRTClusterPosition_EM1` used for the π^0 ID of cell-based. Pi0Cluster p_{T} is in GeV, η bins do also include negative values.



(a)

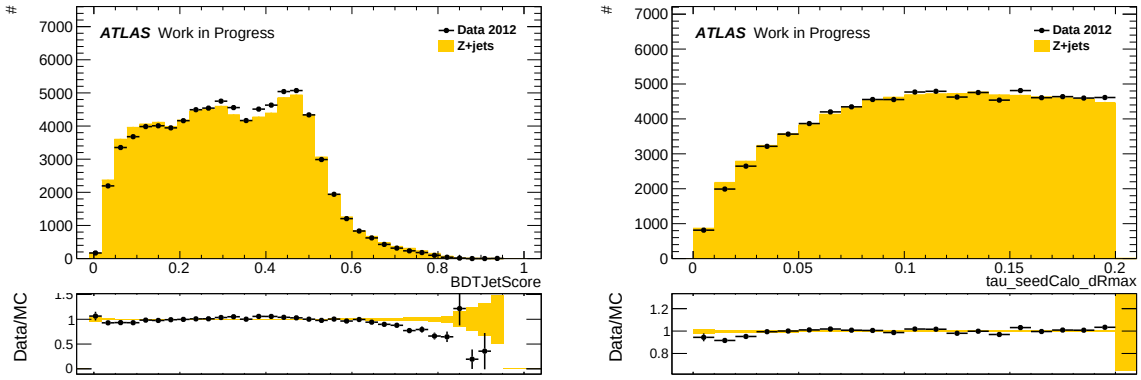
Figure A.38: χ^2/ndf plots of the variable `secondEtaWRTClusterPosition_EM2` used for the π^0 ID of cell-based. Pi0Cluster p_{T} is in GeV, η bins do also include negative values.

A.3 Tau ID Variables

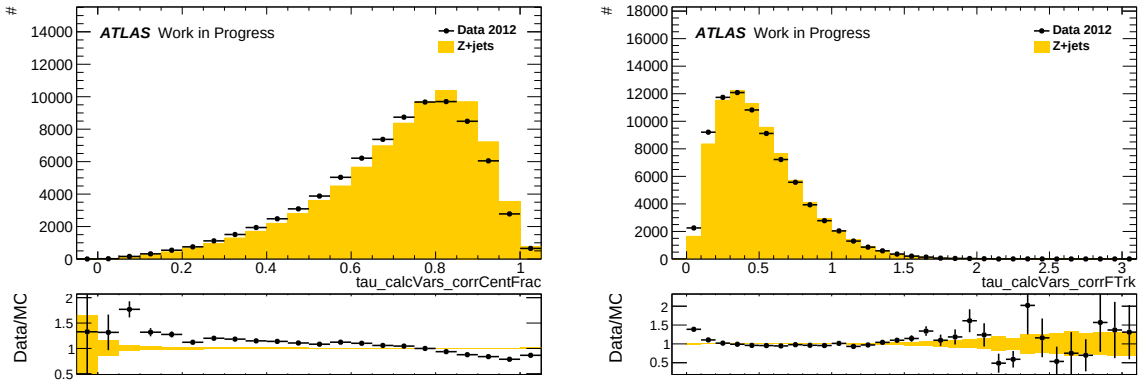


(a) All tau candidates

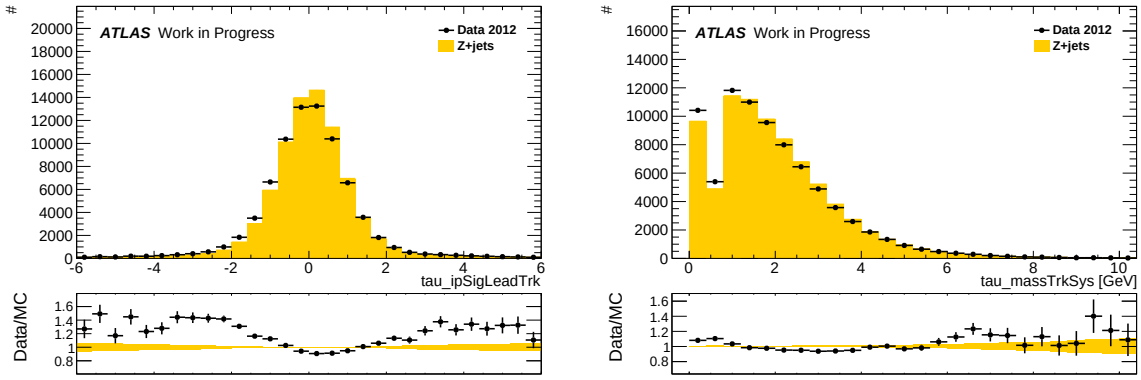
Figure A.39: The number of events in the p_T and η bins for all tau candidates.



(a)

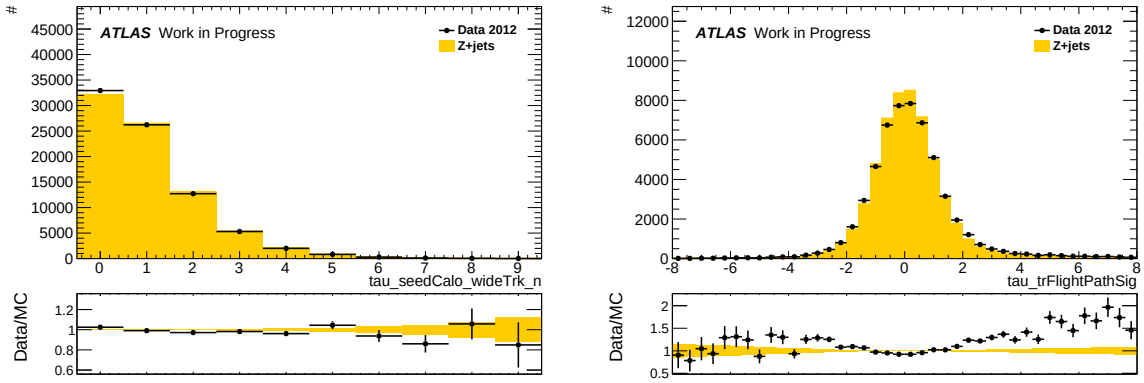


(b)

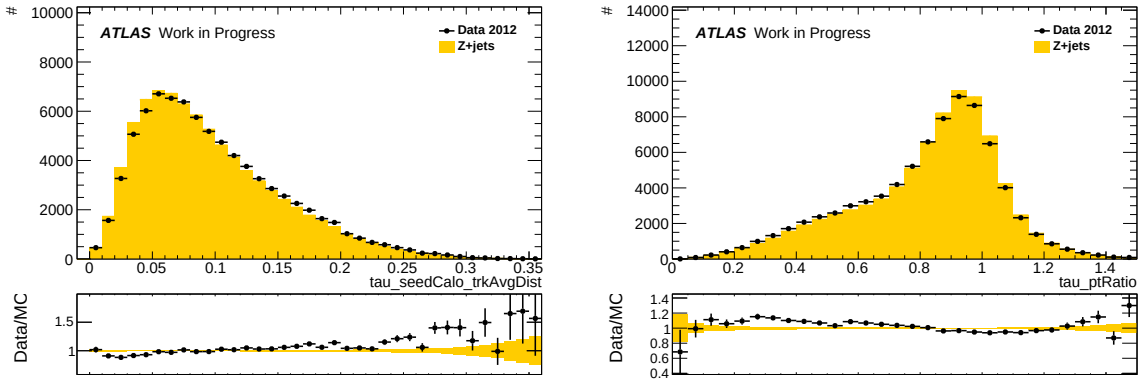


(c)

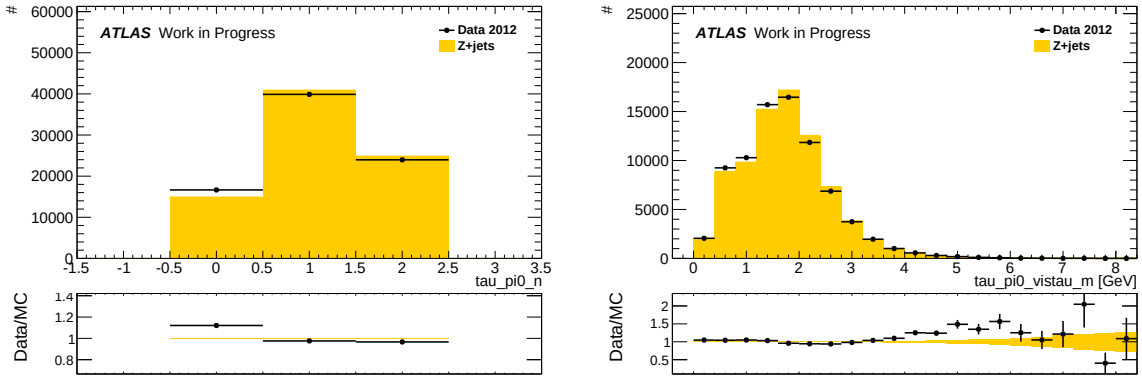
Figure A.40: Inclusive distributions of the TauID BDT score and variables used for that.



(a)

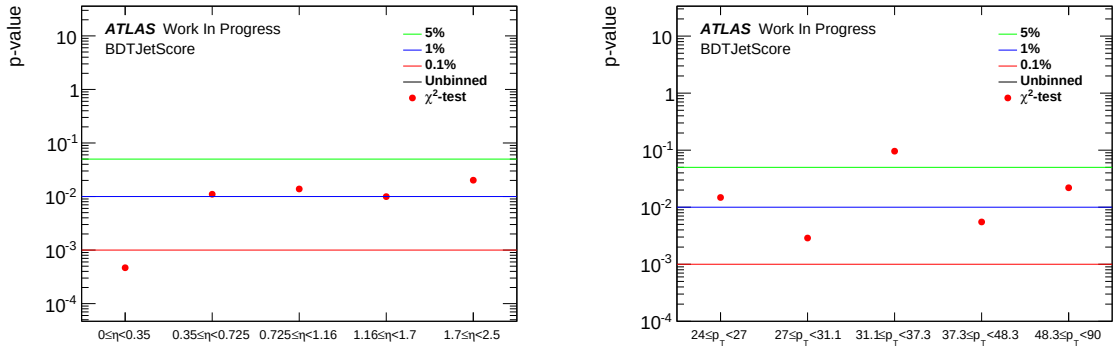


(b)



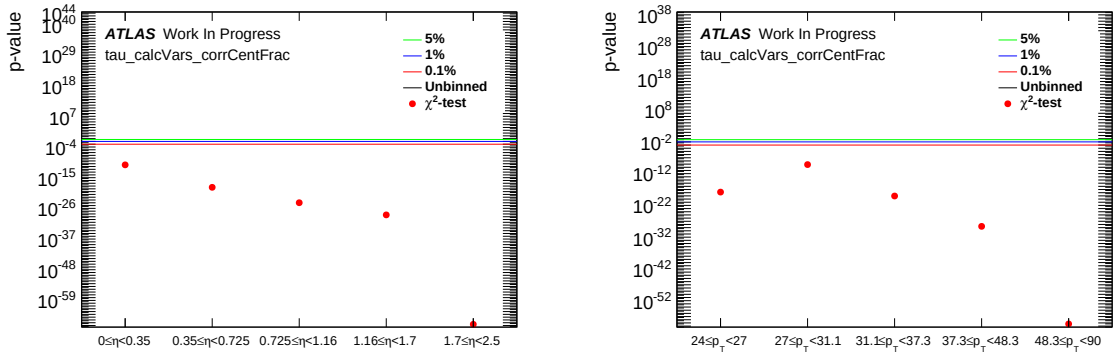
(c)

Figure A.41: Inclusive distributions of the variables used for the TauID BDT.



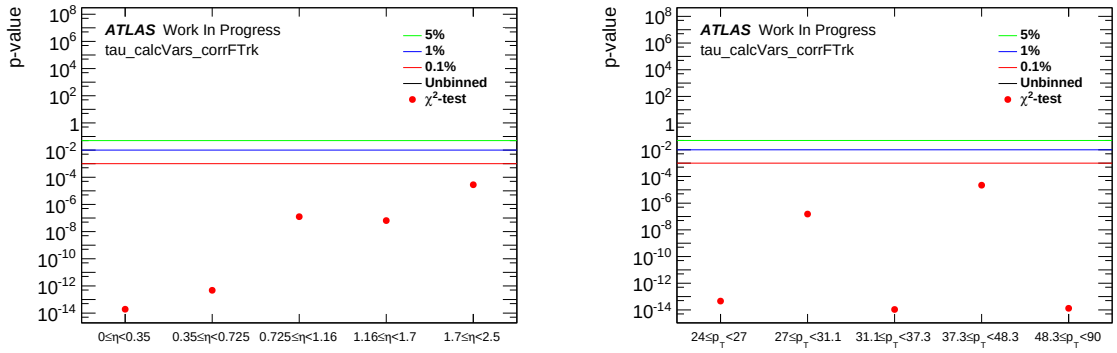
(a)

Figure A.42: P value plots of the TauID BDT score. p_{T} is in GeV, η bins do also include negative values.



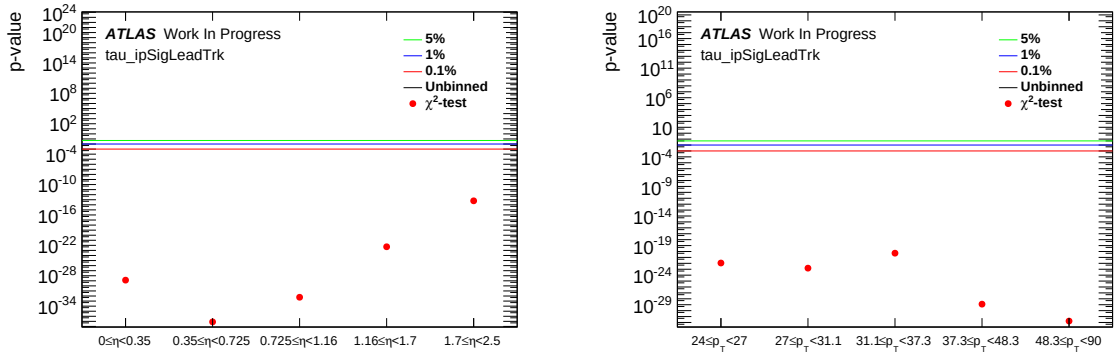
(a)

Figure A.43: P value plots of variable `corrCentFrac` used for the TauID BDT. p_{T} is in GeV, η bins do also include negative values.



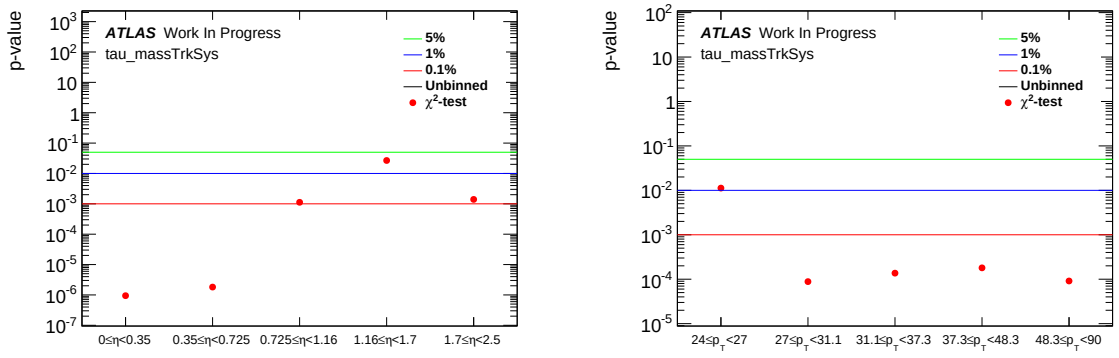
(a)

Figure A.44: P value plots of variable `corrFTrk` used for the TauID BDT. p_{T} is in GeV, η bins do also include negative values.



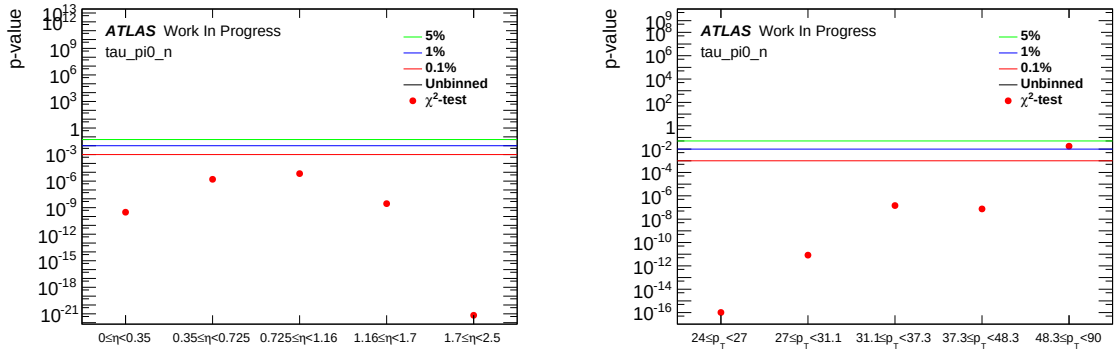
(a)

Figure A.45: P value plots of variable `ipSigLeadTrk` used for the TauID BDT. p_{T} is in GeV, η bins do also include negative values.



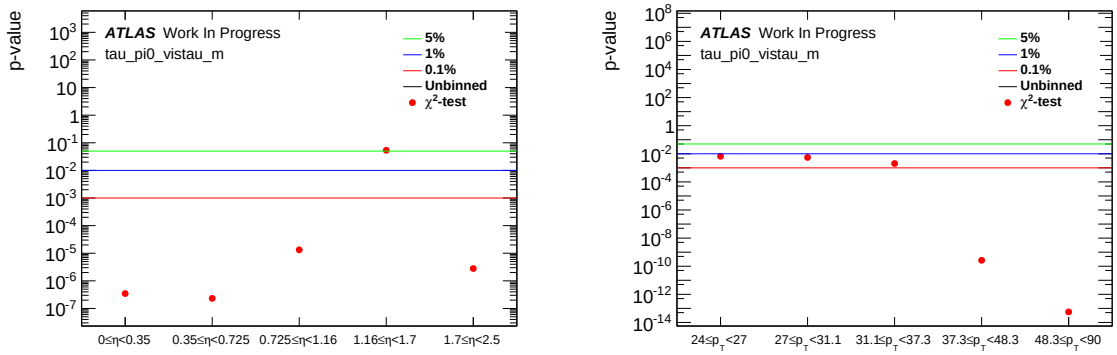
(a)

Figure A.46: P value plots of variable `massTrkSys` used for the TauID BDT. p_{T} is in GeV, η bins do also include negative values.



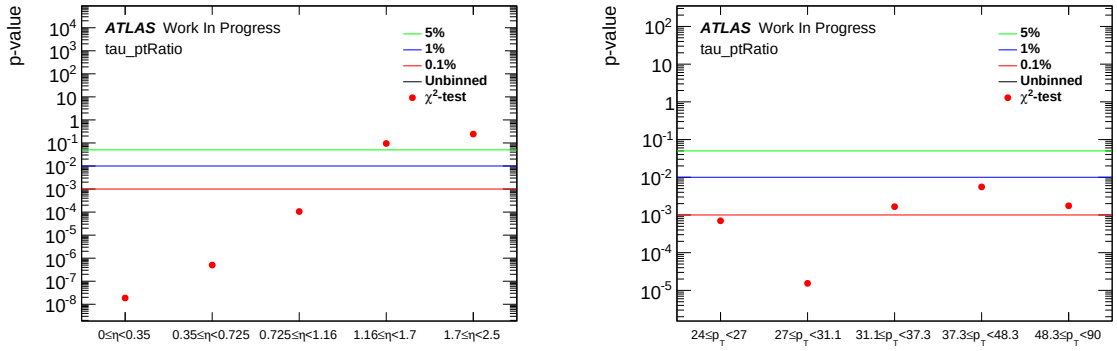
(a)

Figure A.47: P value plots of variable `pi0_n` used for the TauID BDT. p_{T} is in GeV, η bins do also include negative values.



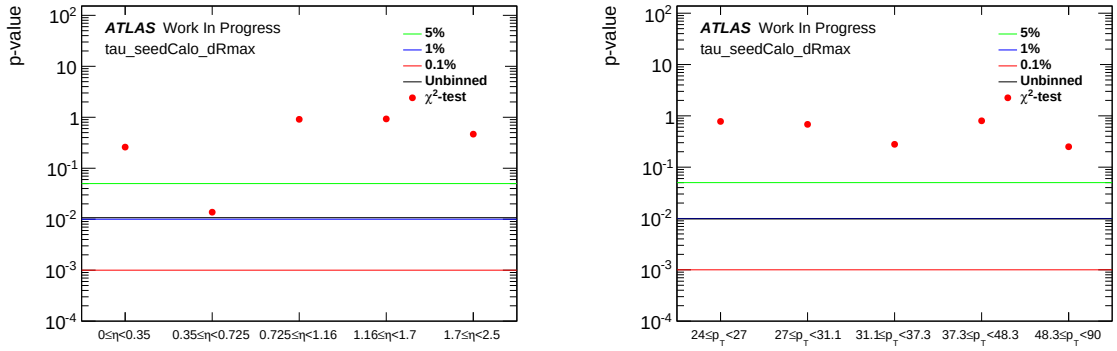
(a)

Figure A.48: P value plots of variable `pi0_vistau_m` used for the TauID BDT. p_{T} is in GeV, η bins do also include negative values.



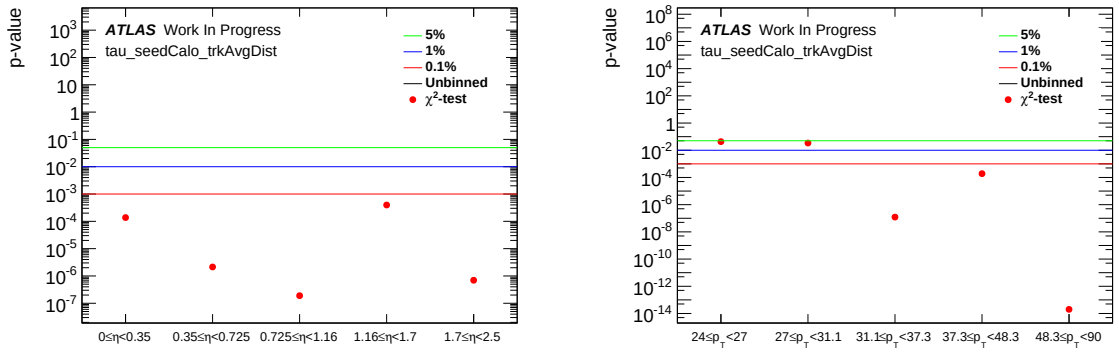
(a)

Figure A.49: P value plots of variable `ptRatio` used for the TauID BDT. p_{T} is in GeV, η bins do also include negative values.



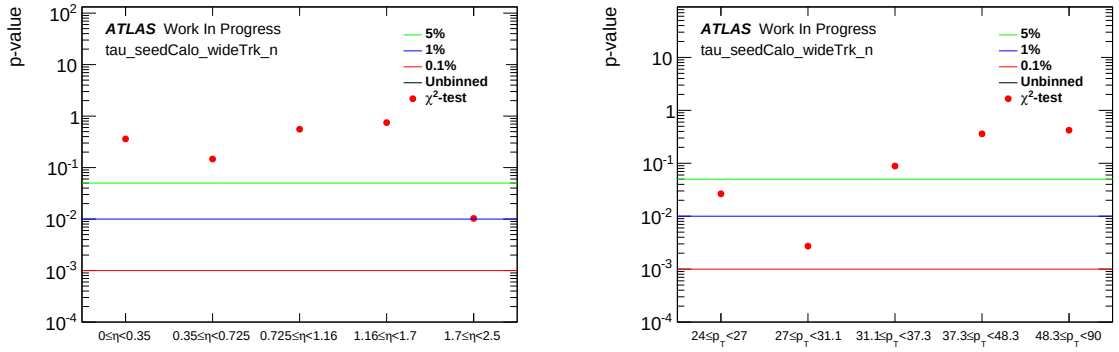
(a)

Figure A.50: P value plots of variable `seedCalo_dRmax` used for the TauID BDT. p_{T} is in GeV, η bins do also include negative values.



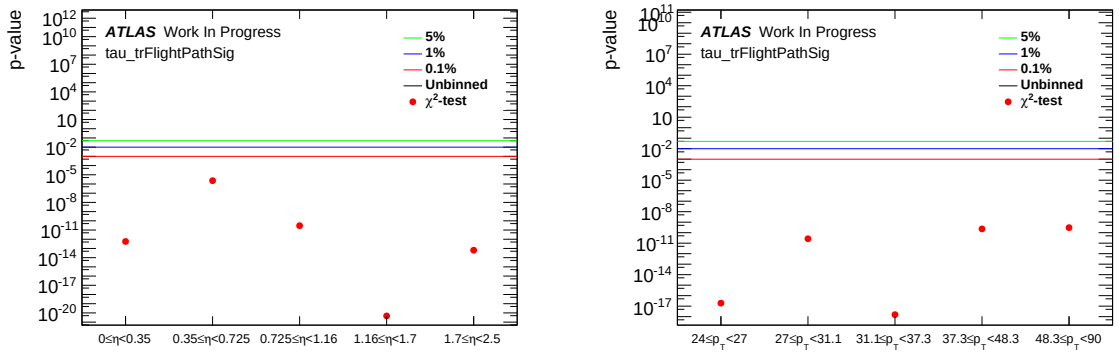
(a)

Figure A.51: P value plots of variable `seedCalo_trkAvgDist` used for the TauID BDT. p_{T} is in GeV, η bins do also include negative values.



(a)

Figure A.52: P value plots of variable `seedCalo_wideTrk_n` used for the TauID BDT. p_{T} is in GeV, η bins do also include negative values.



(a)

Figure A.53: P value plots of variable `trFlightPathSig` used for the TauID BDT. p_{T} is in GeV, η bins do also include negative values.

Bibliography

[1] P. W. Higgs, "Broken symmetries, massless particles and gauge fields", *Phys.Lett.* 12 (1964) 132–133, DOI: 10.1016/0031-9163(64)91136-9.

[2] F. Englert and R. Brout, "Broken Symmetry and the Mass of Gauge Vector Mesons", *Phys.Rev.Lett.* 13 (1964) 321–323, DOI: 10.1103/PhysRevLett.13.321.

[3] ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider", *J. Instrum.* 3 (2008), Also published by CERN Geneva in 2010 S08003. 437 p.

[4] ATLAS Collaboration, "ATLAS: Detector and physics performance technical design report. Volume 1" (1999).

[5] Particle Data Group, K. Olive et al., "Review of Particle Physics", *Chin. Phys. C* 38 (2014) 090001, URL: http://pdg.lbl.gov.

[6] "Identification of the Hadronic Decays of Tau Leptons in 2012 Data with the ATLAS Detector", tech. rep. ATLAS-CONF-2013-064, CERN, July 2013.

[7] B. T. Winter, "Reconstruction of neutral pions in hadronic tau lepton decays in the ATLAS detector" (2012) 82, URL: http://hep1.physik.uni-bonn.de/fileadmin/Publications/ATLAS_Analysis/Dipl/winter.pdf.

[8] C. Limbach, "PhD thesis draft (to be published)", *University of Bonn* (2014).

[9] ATLAS AnalysisSoftwareGroup, *ExtendedPileupReweighting*, 2014, URL: https://twiki.cern.ch/twiki/bin/view/AtlasProtected/ExtendedPileupReweighting.

[10] ATLAS Muon Combined Performance, *Guidelines for Analyses of 2012 Data*, 2014, URL: https://twiki.cern.ch/twiki/bin/view/AtlasProtected/MCPAnalysisGuidelinesData2012.

[11] ATLAS Muon Trigger Signature Group, *TrigMuonEfficiency*, 2014, URL: https://twiki.cern.ch/twiki/bin/view/Atlas/TrigMuonEfficiency.

[12] ATLAS Standard Model Working Group, *SMWZZjetsZxsec2012*, 2013, URL: https://twiki.cern.ch/twiki/bin/view/AtlasProtected/SMWZZjetsZxsec2012.

[13] N. Gagunashvili, “Comparison of weighted and unweighted histograms” (2006), URL: http://arxiv.org/pdf/physics/0605123.pdf.

List of Figures

2.1 The ATLAS detector [3] 4
2.2 Overview of the calorimeter system [3] 5
2.3 The module of the LAr electromagnetic barrel. The first layer has a fine granularity in the η direction. [3] 6

3.1 Leptonic and hadronic tau decays 10
3.2 Distribution of the TauID variable `tau_calcVars_corrCentFrac`. The red area is for a Monte Carlo tau sample, the dots is a multi-jet sample taken from data. [6] 13

4.1 Z boson production processes 18
4.2 Average interactions per crossing with and without pile-up re-weighting. 21
4.3 Reconstructed Z boson mass with and without muon corrections. 22
4.4 p_{T} of the leading and subleading muon 24
4.5 The mass and p_{T} of the Z boson 24
4.6 p_{T} and η of the tau candidate 24
4.7 Figure (a) shows four normal distributed random values. They are sorted by value. Figure (c) shows the normal distribution in four equally probable bins. On the x axis the bin means are marked. Figure (b) shows the resulting Q-Q plot. The random values on the y axis are plotted against the normal distribution bin means on the x axis. The normal distribution is shown in blue for orientation. 27
4.8 Figure (a) shows the distribution of `nPi0Cluster`. From this, systematic uncertainties have been obtained for the decay mode classification. The effect of this can be seen in figure (b) and (c). 29
4.9 Figure (a) shows the decay modes reconstructed by cell-based+PanTau with the systematic uncertainties from the decay mode shifting method. Figure (b) shows the distribution of `Pi0_sumPt`. 30
4.10 P value plots for the decay mode classification of Cell-based+PanTau 30

4.11 Figure (a) shows the p values of the BDT variable NPhotonsInSeed for the 3p0n-vs.-3pXn. Positive and negative values of η are taken into account. Figure (b) shows the distribution for the high η bin that has the lowest p value. Figure (c) shows the normalised residuals of non-empty bins used for the χ^2 test, the x axis is arbitrary. Figure (d) shows the Q-Q plot that compares the quantiles of the residuals with those of the theoretical gaussian distribution. 31
4.12 P value plots for the BDT scores of PanTau's 1p0n-vs.-1p1n test. p_{T} is in GeV. 32
4.13 P value plots for the BDT scores of PanTau's 1p1n-vs.-1pXn test. p_{T} is in GeV. 32
4.14 P value plots for the BDT scores of PanTau's 3p0n-vs.-3pXn test. p_{T} is in GeV. 33
4.15 P value plots for the variables NPhotonsInSeed and BDTSort1 of PanTau's 3p0n-vs.-3pXn test for η. 33
4.16 Figures (a-b) show the 3p0n-vs.-3pXn BDT score for the low and high η bin. Figures (c-f) show the variables that shows most disagreement in the high η bin. 34
4.17 The χ^2/ndf plots for the BDT score of the cell-based π^0 ID BDT test. 35
4.18 The BDT score distribution of the cell-based π^0 ID BDT test for the η bin with the best χ^2/ndf (a) and the worst (b). 36
4.19 χ^2/ndf plots for two variables that indicate disagreement at high p_{T}. 36
4.20 Inclusive distributions of selected variables used for the π^0 ID of cell-based. 37
4.21 Inclusive distributions of the TauID BDT score and selected variables. 39
4.22 Figure (a) shows the p values TauID BDT score for p_{T} bins. Figures (b-c) show variables with disagreement in the high p_{T} bin. p_{T} is in GeV. 40

A.1 The number of events in the p_T and η bins for all tau candidates and per BDT 44
A.2 Inclusive distributions of the variables used for the 1p0n-vs.-1p1n BDT. 45
A.3 Inclusive distributions of the variables used for the 1p1n-vs.-1pXn BDT. 46
A.4 Inclusive distributions of the variables used for the 3p0n-vs.-3pXn BDT. 47
A.5 P value plots for the BDT score of the 1p0n-vs.1p1n test. p_{T} is in GeV, η bins doalso include negative values. 48
A.6 P value plots for the variable EtDRxTotalEt of the 1p0n-vs.-1p1n test. p_{T} is in GeV, η bins do also include negative values. 48
A.7 P value plots for the variable DeltaR1stNeutralTo1stCharged of the 1p0n-vs.-1p1n test. p_{T} is in GeV, η bins do also include negative values. 48
A.8 P value plots for the variable BDTSort1 of the 1p0n-vs.-1p1n test. p_{T} is in GeV, η bins do also include negative values. 49
A.9 P value plots for the variable 1stBDTEtOverEtAllConsts of the 1p0n-vs.-1p1n test. p_{T} is in GeV, η bins do also include negative values. 49
A.10 P value plots for the variable NPhotonsInSeed of the 1p0n-vs.-1p1n test. p_{T} is in GeV, η bins do also include negative values. 49

A.11 P value plots for the BDT score of the 1p1n-vs.1pXn test. p_{T} is in GeV, η bins do also include negative values. 51
A.12 P value plots for the variable `NNeutralConsts` of the 1p1n-vs.-1pXn test. p_{T} is in GeV, η bins do also include negative values. 51
A.13 P value plots for the variable `HLV_SumM` of the 1p1n-vs.-1pXn test. p_{T} is in GeV, η bins do also include negative values. 52
A.14 P value plots for the variable `BDTSort2` of the 1p1n-vs.-1pXn test. p_{T} is in GeV, η bins do also include negative values. 52
A.15 P value plots for the variable `EtOverEtAllConsts` of the 1p1n-vs.-1pXn test. p_{T} is in GeV, η bins do also include negative values. 52
A.16 P value plots for the variable `NPhotonsInSeed` of the 1p1n-vs.-1pXn test. p_{T} is in GeV, η bins do also include negative values. 53
A.17 P value plots for the BDT score of the 3p0n-vs.-3pXn test. p_{T} is in GeV, η bins do also include negative values. 55
A.18 P value plots for the variable `StdDev_Et_WrtEtAllConsts` of the 3p0n-vs.-3pXn test. p_{T} is in GeV, η bins do also include negative values. 55
A.19 P value plots for the variable `BDTSort1` of the 3p0n-vs.-3pXn test. p_{T} is in GeV, η bins do also include negative values. 56
A.20 P value plots for the variable `EtOverEtAllConsts` of the 3p0n-vs.-3pXn test. p_{T} is in GeV, η bins do also include negative values. 56
A.21 P value plots for the variable `NPhotonsInSeed` of the 3p0n-vs.-3pXn test. p_{T} is in GeV, η bins do also include negative values. 56
A.22 The number of events in the p_T and η bins for all Pi0Clusters 57
A.23 Inclusive distribution π^0 ID BDT score of cell-based. 57
A.24 Inclusive distributions of the variables used for the π^0 ID of cell-based. 58
A.25 Inclusive distributions of the variables used for the π^0 ID of cell-based. 59
A.26 χ^2/ndf plots of the BDT score of the π^0 ID of cell-based. Pi0Cluster p_{T} is in GeV, η bins do also include negative values. 60
A.27 χ^2/ndf plots of the variable `NPosECells_EM1` used for the π^0 ID of cell-based. Pi0Cluster p_{T} is in GeV, η bins do also include negative values. 60
A.28 χ^2/ndf plots of the variable `NPosECells_EM2` used for the π^0 ID of cell-based. Pi0Cluster p_{T} is in GeV, η bins do also include negative values. 60
A.29 χ^2/ndf plots of the variable `AsymmetryWRTTrack` used for the π^0 ID of cell-based. Pi0Cluster p_{T} is in GeV, η bins do also include negative values. 61
A.30 χ^2/ndf plots of the variable `ENG_FRAC_EM` used for the π^0 ID of cell-based. Pi0Cluster p_{T} is in GeV, η bins do also include negative values. 61
A.31 χ^2/ndf plots of the variable `ENG_FRAC_CORE` used for the π^0 ID of cell-based. Pi0Cluster p_{T} is in GeV, η bins do also include negative values. 61

A.32 χ^2/ndf plots of the variable SECOND_R used for the π^0 ID of cell-based. Pi0Cluster p_T is in GeV, η bins do also include negative values. 62
A.33 χ^2/ndf plots of the variable CENTER_LAMBDA_helped used for the π^0 ID of cell-based. Pi0Cluster p_T is in GeV, η bins do also include negative values. 62
A.34 χ^2/ndf plots of the variable FIRST_ETA used for the π^0 ID of cell-based. Pi0Cluster p_T is in GeV, η bins do also include negative values. 62
A.35 χ^2/ndf plots of the variable log_SECOND_ENG_DENS used for the π^0 ID of cell-based. Pi0Cluster p_T is in GeV, η bins do also include negative values. 63
A.36 χ^2/ndf plots of the variable EcoreOverEEM1 used for the π^0 ID of cell-based. Pi0Cluster p_T is in GeV, η bins do also include negative values. 63
A.37 χ^2/ndf plots of the variable secondEtaWRTClusterPosition_EM1 used for the π^0 ID of cell-based. Pi0Cluster p_T is in GeV, η bins do also include negative values. 63
A.38 χ^2/ndf plots of the variable secondEtaWRTClusterPosition_EM2 used for the π^0 ID of cell-based. Pi0Cluster p_T is in GeV, η bins do also include negative values. 64
A.39 The number of events in the p_T and η bins for all tau candidates. 65
A.40 Inclusive distributions of the TauID BDT score and variables used for that. 66
A.41 Inclusive distributions of the variables used for the TauID BDT. 67
A.42 P value plots of the TauID BDT score. p_T is in GeV, η bins do also include negative values. 68
A.43 P value plots of variable corrCentFrac used for the TauID BDT. p_T is in GeV, η bins do also include negative values. 68
A.44 P value plots of variable corrFTrk used for the TauID BDT. p_T is in GeV, η bins do also include negative values. 68
A.45 P value plots of variable ipSigLeadTrk used for the TauID BDT. p_T is in GeV, η bins do also include negative values. 69
A.46 P value plots of variable massTrkSys used for the TauID BDT. p_T is in GeV, η bins do also include negative values. 69
A.47 P value plots of variable pi0_n used for the TauID BDT. p_T is in GeV, η bins do also include negative values. 69
A.48 P value plots of variable pi0_vistau_m used for the TauID BDT. p_T is in GeV, η bins do also include negative values. 70
A.49 P value plots of variable ptRatio used for the TauID BDT. p_T is in GeV, η bins do also include negative values. 70
A.50 P value plots of variable seedCalo_dRmax used for the TauID BDT. p_T is in GeV, η bins do also include negative values. 70
A.51 P value plots of variable seedCalo_trkAvgDist used for the TauID BDT. p_T is in GeV, η bins do also include negative values. 71

A.52 P value plots of variable `seedCalo_wideTrk_n` used for the TauID BDT. p_T is in GeV, η bins do also include negative values. 71
A.53 P value plots of variable `trFlightPathSig` used for the TauID BDT. p_T is in GeV, η bins do also include negative values. 71

List of Tables

3.1 Branching fractions of the tau decay modes [5]. The decay modes considered for substructure reconstruction are listed explicitly in the hadronic section. 10
3.2 Variables used for Tau ID. 1-prong variables are used only for 1-prong taus, multi-prong variables are used for taus with 2 or more prongs. For a more detailed description, see [6]. 13
3.3 The cuts on the BDT score and E_{T} of neutral clusters 14
3.4 Cell-based π^0-ID BDT variables, the descriptions are taken from [7]. 15
3.5 PanTau BDT variables. The description is taken from [8]. PFOs are *Particle Flow Objects*. These are the technical objects in substructure reconstruction. They can either be a charged ($\pi^{\pm}$) or neutral. Neutral PFOs can be further distinguished into π^0-identified and non-π^0 PFOs. 16
3.6 The complete tau composition determines in which BDT test the tau is filled. 16

4.1 These are the requirements of the filters applied on Monte Carlo. The cuts on `ptCone40` and `etCone40` are isolation requirements. They are defined as $\texttt{ptCone40} = \frac{\sum_{\text{tracks}\in\Delta R<0.4}\, p_{\mathrm{T}}}{p_{\mathrm{T}}^{\text{track}}}$ or E_{T} respectively. It means that the ratio of transversal momentum/energy of other tracks in a cone of $\Delta R < 0.4$ over the one of the muon/tau track may not be larger than the values given in the table. 19
4.2 The used datasets for experimental and simulated data. 20
4.3 This table shows the number of events that pass each cut for data and Monte Carlo. Due to a bug, the number of events after the jet cleaning is not available for the filtered datasets. You can still estimate the effect by comparing with the unfiltered cut flow. . . 23
4.4 Selection for Zmumu+jets events 25