USING BOOSTED DECISION TREES FOR TAU IDENTIFICATION IN THE ATLAS EXPERIMENT

by

Jennifer Godfrey

BSc., University of the Fraser Valley, 2006

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN THE DEPARTMENT
OF
PHYSICS

SIMON FRASER UNIVERSITY
Spring, 2009

APPROVAL

Name: Jennifer Godfrey

Degree: Master of Science

Title of thesis: Using Boosted Decision Trees for Tau Identification in the ATLAS Experiment

Examining Committee: J. Steven Dodge (Chair)

Dugan O'Neil
Senior Supervisor

Michel Vetterli
Supervisor

Howard Trottier
Supervisor

Levon Pogosian
Examiner

Date Approved: February 9, 2009

Abstract

The ATLAS detector will begin taking data from p-p collisions in 2009. This experiment will allow for many different physics measurements and searches. The production of tau leptons at the LHC is a key signature of the decay of both the standard model Higgs (via $H \to \tau\tau$) and SUSY particles. Taus have a short lifetime ($c\tau = 87$ μm) and decay hadronically 65% of the time. Many QCD interactions produce similar hadronic showers and have cross-sections about 1 billion times larger than tau production. Multivariate techniques are therefore often used to distinguish taus from this background. Boosted Decision Trees (BDTs) are a machine-learning technique for developing cut-based discriminants which can significantly aid in extracting small signal samples from overwhelming backgrounds. In this study, BDTs are used for tau identification for the ATLAS experiment. They are a fast, flexible alternative to existing discriminants with comparable or better performance.

To my parents.

Acknowledgments

I would like to thank my family for their amazing support and patience as September after September has found me “still in school”. To my parents, brothers Bryan and Dean, and grandparents: I love you and appreciate you. Having a strong family is a great blessing to me.

I am fortunate to work with a wonderful supervisor. Dugan, thank you for teaching me about research and giving me the opportunity to work with you. I’ve learned that the time you spend and that many of the things that you have helped me with or taught me are beyond what should reasonably be expected from a supervisor. Thank you for correcting my spelling mistakes, teaching me shell scripting, and not getting mad when I bought a new computer only to discover that it was the monitor that was broken. Thanks also to the other HEP supervisors: Mike Vetterli for teaching me about detectors and Bernd Stelzer for helpful comments and suggestions.

I owe much thanks to the ATLAS tau group, which has some great people working in it. Among them are those who led us as conveners over the course of my degree: Elzbieta Richter-Was, Wolfgang Mader, and Yann Coadou. Stan Lai has enthusiastically answered so many random email questions and, along with Nico Meyer, Anna Kaczmarska, and Marcin Wolter, greatly assisted me with tau reconstruction issues. I also thank Łukasz Janyst for help in the BDTAnalysis program design and for the programming help via instant messenger.

I thank the students in the SFU HEP group, from “oldest” to “youngest”: Dag, for ROOT help, in-office soccer games, and various forms of “older brother” style teasing and support; Zhiyi for making my macros work and me laugh; Doug, our computing guru, for designing the BDTAnalysis package and teaching me a lot about programming; Travis for the company and random facts about everything; Rogayeh for discussions about world issues outside of physics (they do exist); Michel for teaching me French (or at least trying to); Noel for reconstructing the calibration study samples; Sarah for increasing both the female and the Abbotsford population of the office; and Suvayu for the fun conversations and for the future computer help. I also thank the post-docs. Jyothsna has fed me and kept me company at CERN (and got lost with me in Torino) and helped me with many tau-related issues. She really is “always available” for us students. I thank Teresa for help with the error calculations and for many very helpful discussions about physics and other things.

On a personal level, I have had great community and emotional support throughout university. I thank all my roommates past and present. A special thanks to my little community here who has kept me company, sane, fed, and dancing during the last stretch of this degree: Sara (for good talks), Mirjam and David, Laleh (always a bright smile), Scott (trouble), Shannon (walking buddy), and Dave (supplier of food and good breaks).

Finally, I thank my creator God for every reason I have to thank those listed above. Thank you for grace and love and the life you’ve given me. Thank you also for making the world so interesting to study.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

With the start-up of the Large Hadron Collider (LHC), particle physics is entering a new era. The LHC will collide protons at a centre-of-mass energy over 7 times that of the current highest energies. The discovery potential for new physics and opportunities to verify the standard model are enormous. Detectors have been designed and built with this range of possibilities in mind. One general purpose detector situated on an LHC collision point is A Toroidal LHC ApparatuS (ATLAS).

In order to meet the physics goals of the ATLAS experiment, it is important to correctly identify particles within the detector. Taus are one type of particle worth detecting, as they are a signature for several interesting physics processes. Unfortunately, they are also difficult to distinguish from a very high jet background.

The Boosted Decision Tree (BDTs) algorithm is a multivariate technique, which creates a specialized set of criteria over many variables to extract small signals from large, similar backgrounds. BDTs have been employed in tau identification and it will be shown that they make improvements over baseline discriminants.

This thesis documents studies in tau identification using BDTs. Chapter 1 introduces the standard model and motivates these studies. Chapter 2 provides details on the LHC and ATLAS detector. Chapter 3 gives further background on the tau lepton and its reconstruction in ATLAS while Chapter 4 discusses identification methods. The details of how BDTs are built are given in Chapter 5. The application of BDTs to taus, including optimization studies, are shown in Chapter 6. Finally, some concluding remarks are given in Chapter 7.

1.1 The Standard Model

The standard model of particle physics is very well tested and extremely successful. Rather than being a single theory in itself, the standard model is a collection of several complementary theories which together describe interactions between the elementary particles that make up matter. Three fundamental interactions (two of which have been unified) are described by corresponding theories and incorporated into the standard model. Each interaction involves an exchange of force carrying bosons (integer spin[1] particles) between fermions (half-integer spin particles).

The elementary particles whose bound states make up all known matter in the universe fall into one of two categories: leptons or quarks. The sets of leptons and quarks are each organized into three generations. They are shown with their charges[2] and masses[3] in Table 1.1.

There are four fundamental forces in nature. Each is mediated by bosons as force carriers and is described by a physical theory. The forces, from strongest to weakest, are (relative strength in brackets):

- **Strong force** (1): interacts at short range (within the radius of a nucleon) and is mediated by gluons. Particles which have colour charge take part in strong interactions. Strong interactions are governed by the theory of Quantum Chromodynamics (QCD).

- **Weak force** (10^{-3}): interacts at short range, to about 10^{-18} m. It is mediated by 3 bosons: $W^{\pm}$ and Z^0. All particles may undergo weak interactions. It is described by Flavordynamics (the Glashow-Weinberg-Salam theory).

- **Electromagnetic force** (10^{-14}): of infinite range and mediated by the photon γ. All particles with an electric charge are affected by the electromagnetic force. It is described by the theory of Electrodynamics.

[1] Spin is an intrinsic internal quantum number carried by the particle

[2] all charges are in units of the electron charge e

[3] The unit of energy used in particle physics is the electron-volt (eV), which has a value of 1.602×10^{-19} Joules. Because this is a very small unit, MeV (10^6 eV), GeV (10^9 eV), and TeV (10^{12} eV) are more common in subatomic physics. Likewise, masses are measured in units of MeV/c^2. Often c is set to 1 and masses are quoted in units of MeV (or GeV).

Table 1.1: The leptons and quarks in the standard model [1]. There is an antiparticle with equal mass and opposite charge associated with each lepton and quark listed above. Quarks are also coloured, which is a label given to specify that there are three distinct types of quark for each flavour.

Generation	**Leptons - Spin $\frac{1}{2}$**			**Quarks - Spin $\frac{1}{2}$**		
	Flavour	Mass (GeV/c^2)	Charge	Flavour	Mass[a] (GeV/c^2)	Charge
I	ν_e electron neutrino	$< 1 \times 10^{-8}$	0	u up	$< 3 \times 10^{-2}$	$\frac{2}{3}$
	e electron	0.511×10^{-3}	-1	d down	$< 6 \times 10^{-2}$	$-\frac{1}{3}$
II	ν_μ muon neutrino	$< 2 \times 10^{-3}$	0	c charm	1.27	$\frac{2}{3}$
	μ muon	0.106	-1	s strange	0.104	$-\frac{1}{3}$
III	ν_τ tau neutrino	< 0.02	0	t top	171	$\frac{2}{3}$
	τ tau	1.78	-1	b bottom	4.20	$-\frac{1}{3}$

[a] Quark masses are dependent upon the scale and scheme of the measurements. u-, d-, and s-quark masses are estimates of so-called "current-quark masses", in a mass-independent subtraction scheme such as $\overline{\mathrm{MS}}$ at a scale $\mu \approx 2$ GeV. The c- and b-quark masses are the "running" masses in the $\overline{\mathrm{MS}}$ scheme. The t-quark mass is a direct observation of top events.

- **Gravity** (10^{-43}): of infinite range but very weak compared to the other three forces. The force of gravity does not play a role in elementary particle physics and is not incorporated in the standard model.

The masses of the mediating bosons are shown in Table 1.2.

The electron is the lightest lepton. There are two other lepton generations, or flavours, (μ and τ) which are similar to the electron but heavier. For each lepton flavour there is a particle of charge –1 and a neutral, weakly interacting particle called a neutrino. Of the three charged leptons, only the electron is stable.

A major difference between quarks and leptons is that quarks carry colour charge and leptons do not. The result is that unlike leptons, quarks are affected by the strong force which can bind quarks together to form hadrons (colourless[4] bound states of quarks). Or-

[4] All hadrons are colourless. They either contain 3 quarks each with a unique colour (red, green, and blue), or 2 quarks with a colour-anticolour configuration.

Table 1.2: The bosons which mediate the forces.

Name	Mass (GeV/c^2)	Charge
Electromagnetic Force - Spin 1		
γ photon	0	0
Weak Force - Spin 1		
W^-	80.4	-1
W^+	80.4	+1
Z^0	91.2	0

Strong Force - Spin 1		
Name	Mass (GeV/c^2)	Charge
g gluon	0	0

dinary matter consists of protons, neutrons and electrons. The protons and neutrons are bound states of quarks (*uud* and *udd* respectively) held together by the strong force. Processes such as α decays and jet hadronization (see Section 4.2) are due to the strong force.

Leptons only interact electromagnetically and via the weak force. The electromagnetic and weak forces were unified by Salam, Glashow, and Weinberg in what is known as the electroweak force. This unification is appropriate for energies above 100 GeV.

There is one outstanding aspect of the standard model that has yet to be verified. A mechanism called electroweak symmetry breaking is responsible for giving masses to particles in the standard model [9]. Without this phenomenon, all particles within the standard model should be massless, which is contrary to observation. However, a consequence of electroweak symmetry breaking is the existence of a spin 0 particle, the Higgs boson, which has never been observed. The detection of the Higgs is essential to validate the standard model and finding or excluding the Higgs is among the most important goals of the ATLAS experiment.

The Higgs can decay to two fermions $f\bar{f}$, two photons $\gamma\gamma$, two gluons gg, and to the weak mediating bosons W^+W^- and ZZ if the mass of the Higgs is high enough to allow the decay. The probability for each decay depends upon the coupling strength of the Higgs to the vertex of the decay. The coupling strength is proportional to the masses or square of the masses (depending on the process) of the decay products, so that the Higgs decays preferentially to particles of a higher mass. This is a significant motivation for ATLAS to have strong tau identification capabilities, as the tau is the heaviest lepton and the Higgs preferentially decays to taus over other leptons.

1.1.1 Extensions to the Standard Model

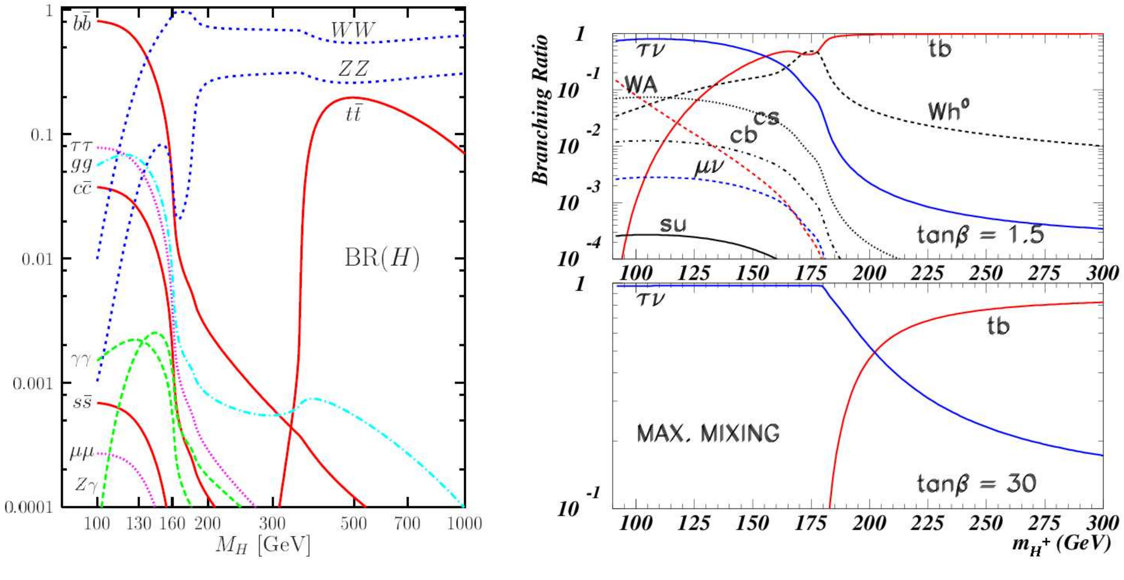
Despite the experimental success of the standard model, there are convincing reasons to believe that the standard model is not complete. There are more than 20 parameters in the standard model (mainly fundamental masses and mixing angles) that cannot be predicted by it and must be found through experimentation. This requirement for so much external input implies that there are aspects of particle physics that cannot be explained or predicted by the standard model. Furthermore, the standard model has no candidate for dark matter or any framework to incorporate gravity at the quantum scale. It also cannot explain why significantly more matter than antimatter is observed in the universe when equal amounts of both should have been created during the big bang. For this reason, many people believe that there must be a larger overarching theory which describes how fundamental matter interacts. If this is the case, the standard model must exist as a limiting case or approximation to the larger theory.

One popular extension of the standard model is "supersymmetry" (SUSY). It is necessary in some formulations of string theory and a quantum theory of gravity is compatible with it. Furthermore, it may also hold the explanation for matter-antimatter asymmetry [10]. In this theory, every fermion (half-integer spin particle), has a corresponding supersymmetric "twin" boson (integer-spin). Likewise, for every boson in the standard model, there is a corresponding supersymmetric fermion superpartner. The lightest supersymmetric particle is a candidate to explain dark matter. Searches are being conducted for these extra supersymmetric particles in order to verify such a theory. There are many types of supersymmetry theories. One of these theories is the Minimal Supersymmetric extension of the Standard Model (MSSM), which also unifies the strong and electroweak interactions at very high energies. In this model, there are five Higgs bosons: $H^{\pm}$ and three neutral: h, H, and A.

1.1.2 Physics Motivation to Study Taus

Tau properties have been studied extensively and many are known to high precision. The majority of current experimental publications on the tau, including many branching fraction measurements, are made at B factories such as those used by the BaBar and Belle collaborations. The ATLAS detector in a *p*-*p* collision environment is not optimal for these types of precision measurements and this is not the goal.

Even though certain precision measurements will not be the focus of the experiment, taus play a significant role in the physics goals of the ATLAS program both as a decay product of the Higgs boson and a probe for new physics. As the heaviest lepton, the tau couples to the Higgs boson more strongly than the other leptons both in the standard model (SM) and Minimal Super Symmetric Model (MSSM). Figure 1.1(a) shows that the branching ratio $H \rightarrow \tau^+\tau^-$ for a low mass standard model Higgs ($m_h < 120$ GeV) is smaller only to that of $H \rightarrow b\bar{b}$. These are the only two direct fermion decay modes which are expected to be distinguishable from background. The $H \rightarrow b\bar{b}$ events have an irreducible larger background from $gg \rightarrow b\bar{b}$, however. Furthermore, unlike *b*-quarks, taus only interact by the electroweak force, so higher order QCD corrections are not required, making calculations cleaner.



(a)

(b)

Figure 1.1: (a) Branching ratios for main decays of the SM Higgs boson [3] and (b) branching ratios for heavy charged Higgs boson in the SUSY model [4].

Taus will also be important to many beyond-the-standard-model searches such as the following Higgs searches [11]:

- neutral Higgs bosons from the MSSM couple most strongly to the heaviest lepton and the heaviest down-type quark. Even though the Higgs branching ratio to the b-quark is ~ 90 % and only ~ 10 % to the tau, the tau provides the cleanest signal.
- $H^{\pm} \to \tau\nu_{\tau}$ is one of the most promising decay channels for charged Higgs discovery if $m_H < m_t$, as shown by the branching fractions in Figure 1.1(b). Even for $m_H > m_t$, the more plentiful $H^{\pm} \to tb$ channel has a large irreducible background, unlike $H^{\pm} \to \tau\nu_{\tau}$. The polarization of the tau can be used to further distinguish signal from background, which will make $H^{\pm} \to \tau\nu_{\tau}$ valuable for extending the discovery potential of the charged Higgs.

1.2 Experimental Particle Physics

Many particles are short lived and must first be created in the laboratory before they can be studied. Furthermore, one cannot see elementary particles; they must be observed indirectly. Their properties are measured according to their interactions with matter. With these interactions, particle identification, energies, and decay properties are inferred. These two requirements, creation and observation, are the driving motivation behind large particle accelerators and detectors such as the LHC and ATLAS. In this system, two beams of protons are accelerated to very high energies in the LHC and collide head-on in the middle of the ATLAS detector.

High energies are required for two main reasons. The first reason is to be able to create massive particles. During beam collisions, two particles of mass m_1 and m_2 may interact and annihilate to form a third particle. The mass m of this third particle may exceed $m_1 + m_2$ if the initial particles have sufficient energy, due to Einstein's equation $E = mc^2$. Secondly, high energies are required to study features at small distances. The wavelength of a particle decreases with momentum by $\lambda = \frac{h}{p}$, where h is Planck's constant. A smaller wavelength allows the interaction to resolve small distances.

The actual interactions in ATLAS are not fixed at 14 TeV, however. A proton of high energy contains not only two u and one d valence quarks, but many gluons and additional "sea" quark-antiquark pairs as well. Each parton[5] within the proton carries some fraction x of the total proton momentum. When two partons from opposing beams collide, the total center of mass energy of the collision is dependent on the sum of the two momentum fractions and sets the maximum mass of any newly created particle. The fraction of momentum carried by a parton can range from very low to nearly 1. The proton-proton collisions that occur within ATLAS therefore cover a wide range of center of mass energies, with 14 TeV being the theoretical maximum energy. It is because the z component of the parton energy is unknown that the typical energy measurement in ATLAS is only in the direction transverse to the beamline (denoted by E_T). While the original energy boost along z is unknown, the transverse energy and momentum should always be conserved in every interaction.

Every collision within the LHC produces many secondary particles. Different types of particles require different detection strategies. The size and variety of components within the ATLAS detector reflects these needs.

[5] a quark or gluon

Chapter 2

Experimental Setup

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC), which will begin full operations in 2009, is designed to collide proton beams with a center-of-mass energy of 14 TeV. This will be the highest energy achieved thus far by a hadron collider. The LHC will also periodically accelerate lead ions for a center-of-mass collision energy of 5.5 TeV per nucleon pair.

The LHC is located on the France-Switzerland border in a tunnel 50 to 175 m underground. To save on construction costs, the LHC reused the accelerator tunnel from the Large Electron-Positron Collider (LEP). This pre-defined tunnel radius limited the maximal energy the proton beam could achieve for a given magnet strength.

The protons pass through a series of accelerators before reaching the LHC. They begin in the Linac2 (Linear Accelerator), then are injected into the PSB (Proton Synchrotron Booster), then the PS (Proton Synchrotron), and then the SPS (Super Proton Synchrotron). Finally, protons with an energy of 450 GeV are injected from the SPS into the LHC. The layout of these accelerators is shown in Figure 2.1.

In total, there are 4 collision points and 6 experiments on the LHC ring:

- A Toroidal LHC ApparatuS (ATLAS), a general-purpose experiment utilizing the *p*-*p* beam,

- Compact Muon Solenoid (CMS), the second general-purpose experiment utilizing the *p*-*p* beam,

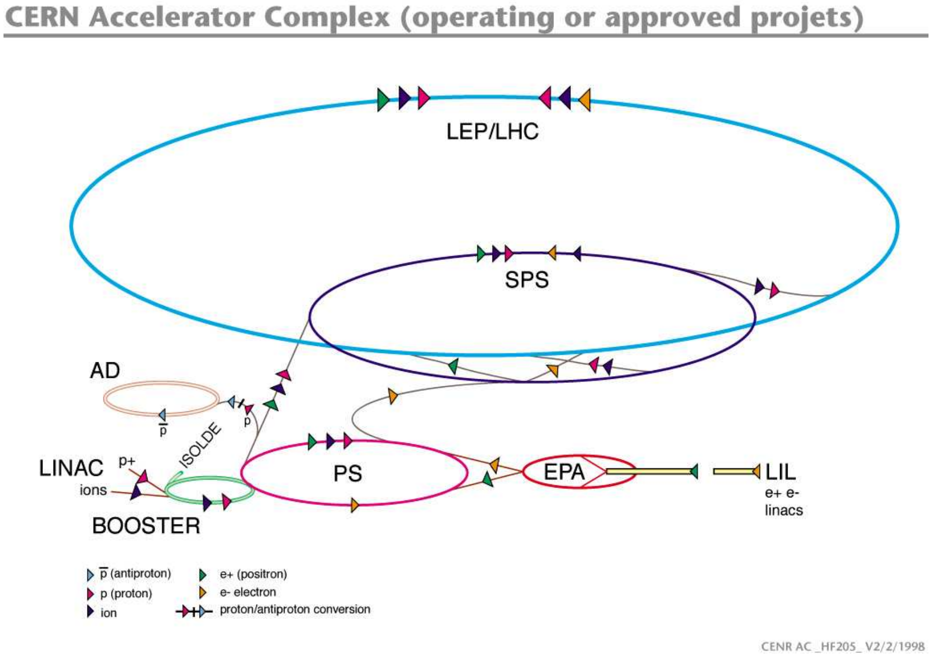


Figure 2.1: Illustration of the CERN accelerator complex. Copyright CERN [5].

- Large Hadron Collider forward (LHCf) consists of 2 detectors situated at ±140 m from the ATLAS interaction point. It will study very forward particles to compare and verify MC simulations of cosmic ray events,

- A Large Ion Collider Experiment (ALICE), which will use the lead beam collisions to study the quark-gluon plasma,

- LHCb, which will use b-quarks created in proton collisions to study matter-antimatter asymmetry,

- TOTal Elastic and diffractive cross section Measurement (TOTEM), which will also study forward particles to measure the LHC luminosity and total cross section, elastic scattering, and diffractive processes.

Figure 2.2 shows the geographical location of the LHC and the four collision points.

The two counter-rotating beams collide discretized packets, or bunches, of protons every 25 ns. The LHC tunnel contains dipole magnets which have two cores with opposite magnetic fields. In each of these cores, a proton beam circulates in the direction opposite to

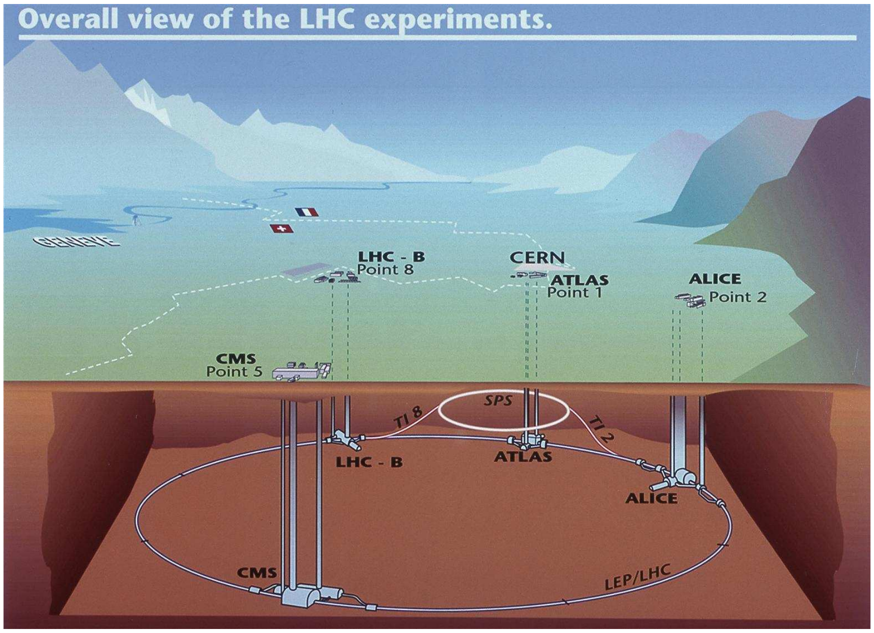


Figure 2.2: Overall view of the LHC experiments, located on the France-Switzerland border. Copyright CERN [6].

the beam in the other core, held in its path by the dipole magnets. At the collision point, the beams are angled towards each other and collide with a half crossing angle of 142.5 μrad.

CMS and ATLAS search for events and signatures which are very rare compared to ordinary events. The luminosity of the LHC, which is related to the number of collisions that can be produced in a detector per cm^2 per second, is therefore important. It is defined as

$$L = f\frac{N_1 N_2}{A}$$

where f is the frequency of bunch crossings, N_i is the number of protons in bunch i, and A is the effective cross-sectional area of the beams (a cross-section that takes the angles of the beam collision into account). The peak luminosity at the LHC will be 10^{34} cm^2 s^{-1}.

2.2 The ATLAS Experiment

Surrounding one of these four collision points on the LHC is “A Toroidal LHC ApparatuS” (ATLAS), one of two general purpose detectors on the LHC ring. Its design, assembly, calibration, and overall preparation have been a collaborative effort among the 37 nations and 2500 scientists who contribute to the ATLAS project.

The ATLAS detector is situated in an underground cavern and completely surrounds the LHC collision point with many layers of different detectors arranged progressively through the radial direction. It is 25m high, 44 m long, and weighs approximately 7000 tonnes. The detector is comprised of several subsystems which include the inner detector, the calorimeter system, and the muon system. A computer engineered view of the detector is shown in Figure 2.3.

2.3 Geometry of the ATLAS Detector

The Cartesian coordinates of the ATLAS detector are defined with the axis origin at the interaction point. The beam direction defines the z-axis, the positive x-axis points into the center of the LHC ring, and the positive y-axis points upwards.

The azimuthal angle is measured around the beam axis such that $\tan(\phi) = \frac{p_x}{p_y}$. The polar angle θ is the angle from the beam axis. However, a quantity called the rapidity is favoured to express angle with respect to the beam line because differences in rapidity are invariant under Lorentz boosts. It has the further advantage that particle production is roughly independent of the rapidity. A massless approximation to rapidity is the pseudorapidity, defined as $\eta = -\ln\left(\tan\left(\frac{\theta}{2}\right)\right)$. Angular separation between points is defined by

$$\Delta R = \sqrt{(\phi' - \phi)^2 + (\eta' - \eta)^2} = \sqrt{\Delta\phi^2 + \Delta\eta^2}$$

The main regions of the detector are the barrel (with $|\eta| < 2.5$) and the two end-caps ($|\eta| \geq 2.5$) on either side.

2.4 Detector Requirements

While the ATLAS detector was designed to detect many possible physics processes, its design was heavily influenced by the need to observe or exclude the Higgs boson over its entire allowed mass range. The dominant decay modes of the Higgs depend on its mass, which is unknown. The multiplicity of events and high backgrounds also require several

Higgs channels to be measured in order to verify results. This results in a very broad set of required detector capabilities. At low mass, for example, the $H \to \gamma\gamma$ channel is important, and requires the ability to identify the outgoing photons amidst background. If the Higgs has a higher mass, $H \to Z^*Z$ with both Z's decaying to leptons is a clean channel to identify. This is a strong motivator for the stand-alone muon system

Along with direct Higgs decay products, objects associated with the Higgs production are used for Higgs identification. As shown in Figure 2.4(a), the WW and ZZ fusion production of the Higgs requires the tagging of forward jets, requiring full calorimeter coverage of the detector. The $t\bar{t}$ fusion production of the Higgs shown in Figure 2.4(b), which results in a ttH final state, requires secondary vertexing capabilities in the tracker to identify the b-jets from the t decays and from $H \to bb$. Because all these processes are relatively rare, a high luminosity is required.

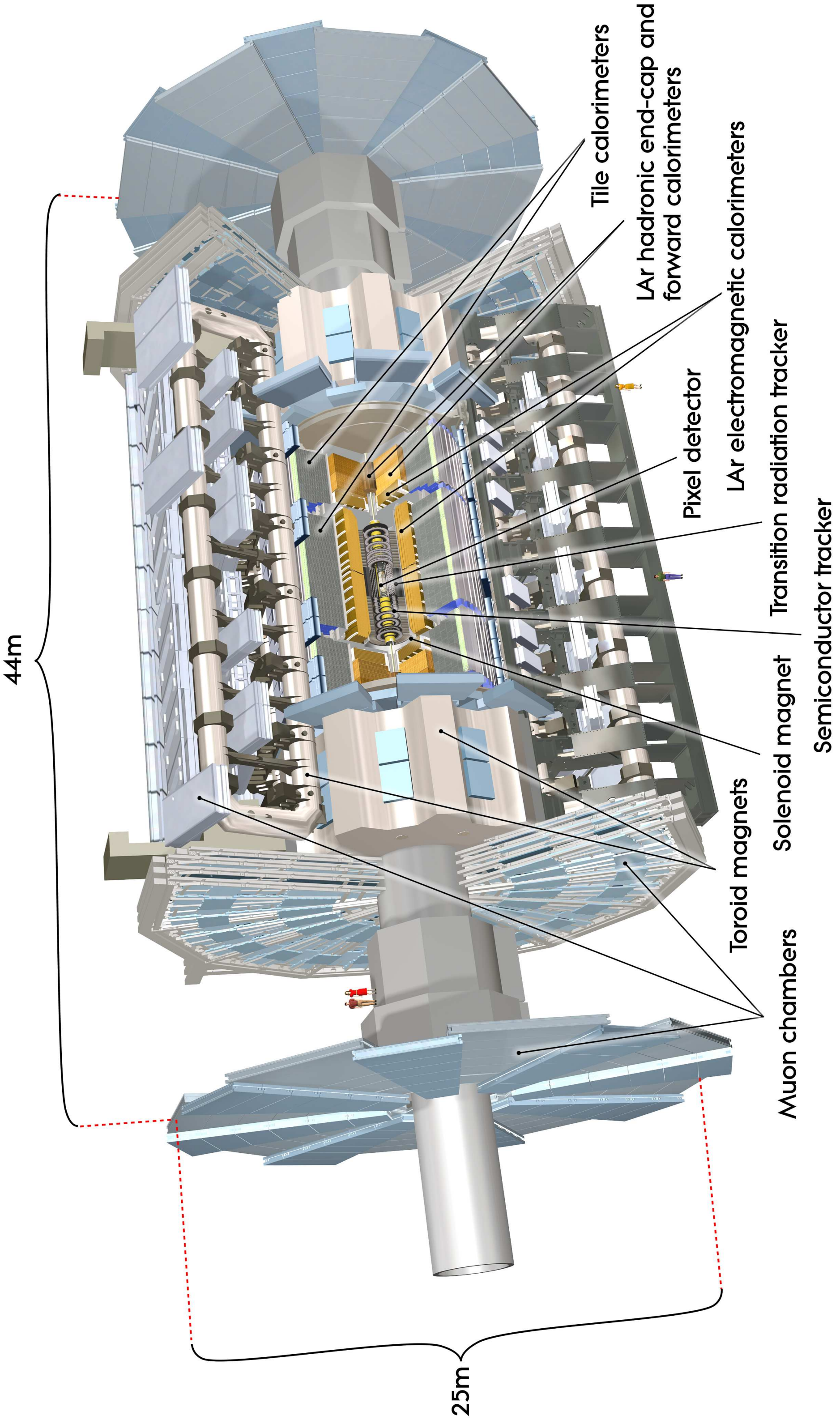
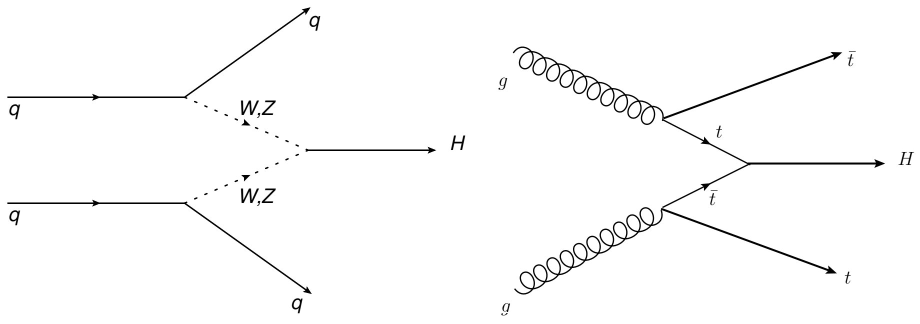

Tile calorimeters
LAr hadronic end-cap and forward calorimeters
Pixel detector
LAr electromagnetic calorimeters
Transition radiation tracker
Semiconductor tracker
Solenoid magnet
Toroid magnets
Muon chambers
44m
25m

Figure 2.3: Cut-away view of the ATLAS detector. ATLAS Experiment Image: Copyright CERN, [7].



(a) WW and ZZ fusion production of the Higgs boson.

(b) $t\bar{t}$ fusion production of the Higgs boson.

Figure 2.4: Two production channels of the Higgs boson.

The following are general requirements that the ATLAS detector has been designed to fulfill [7]:

- Electronics are to be fast and radiation hard and a high detector granularity is required to cope with LHC energies and luminosity.
- Coverage in all directions (ϕ and pseudo-rapidity η).
- Excellent calorimetry: electromagnetic for electron and photon identification and full hadronic coverage for jet and missing transverse energy measurements.
- Good muon identification and energy resolution with good charge identification up to large p_T.
- Efficient triggering on processes of interest with acceptable background rejection.

2.5 Detector Components

Descriptions of subdetectors relevant to these tau identification studies are given below [7, 12].

2.5.1 The Inner Detector

The paths of charged particles can be reconstructed by piecing together discrete space points in detectors as they pass through. This process, called tracking, uses data from the inner detector (ID), which is comprised of three complimentary but separate sub-systems: a pixel detector, a silicon SemiConductor Tracker (SCT), and a Transition Radiation Tracker (TRT). An overview of the ID is shown in Figure 2.5 and a more structural view is in Figure 2.6. The general function of each system is to record a "hit" when a particle passes through a certain point of the detector. These hits, when fit together, form the particle's trajectory. The inner detector is immersed in a 2T magnetic field. Charged particles are deflected by the magnetic field and their momentum and charge can be found using the curvature of the tracks. With 1000 particles expected to traverse the region covering $\eta < 2.5$ every 25 ns, high granularity is required to separate tracks from different particles.

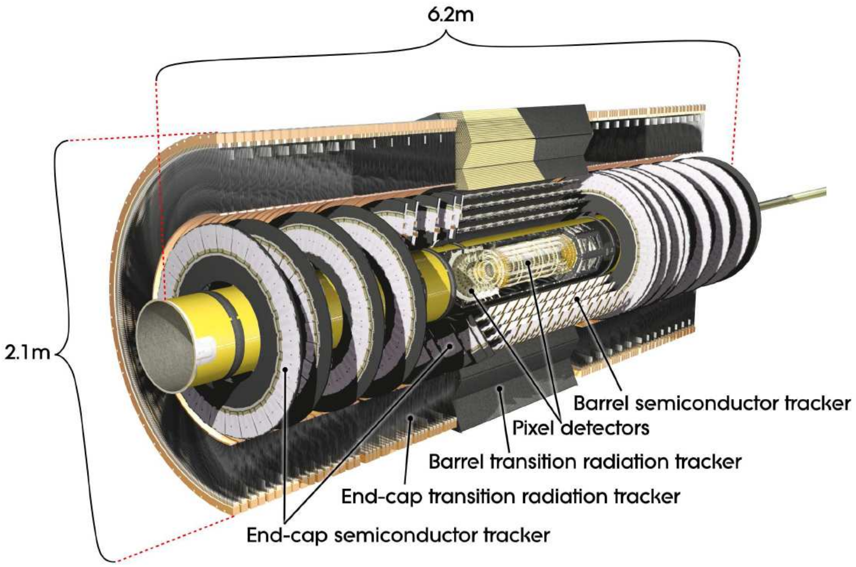

6.2m
2.1m
Barrel semiconductor tracker
Pixel detectors
Barrel transition radiation tracker
End-cap transition radiation tracker
End-cap semiconductor tracker

Figure 2.5: Cut-away view of the ATLAS inner detector. Copyright: The ATLAS Experiment at CERN, [7].

The pixel detector and the SCT are collectively referred to as the precision tracker. They are arranged in concentric cylinders around the beam axis in the barrel region and in disks perpendicular to the beam in the end-cap region. The purpose of these detectors is to provide the fine granularity required for impact parameter and vertexing measurements.

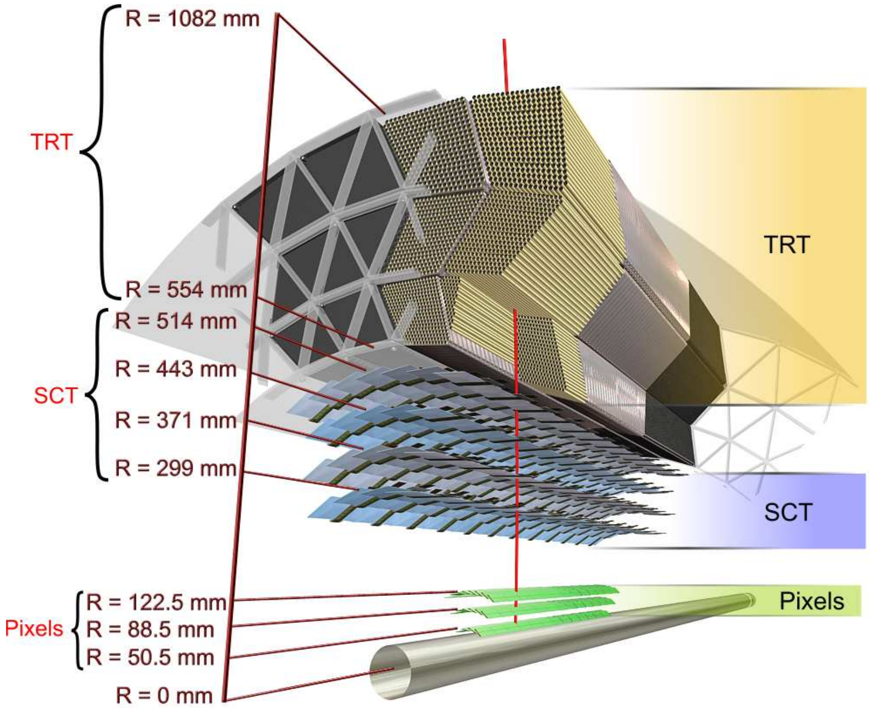
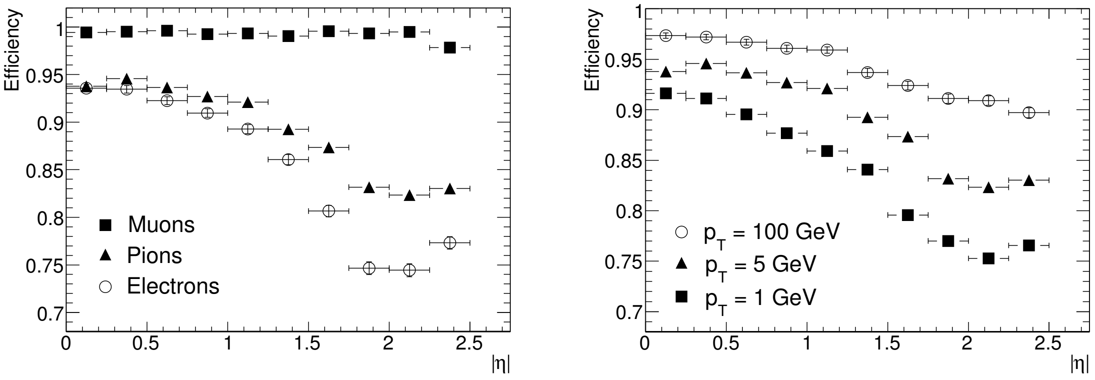


Figure 2.6: Drawing showing the sensors and structural elements traversed by a charged track of 10GeV p_T in the barrel inner detector (η=0.3). The track traverses successively the beryllium beampipe, the three cylindrical silicon-pixel layers with individual sensor elements of 50x400 μm^2, the four cylindrical double layers (one axial and one with a stereo angle of 40 mrad) of barrel silicon-microstrip sensors(SCT) of pitch 80μm, and approximately 36 axial straws of 4 mm diameter contained in the barrel transition-radiation tracker modules within their support structure. Copyright: The ATLAS Experiment at CERN, [7].

The size of this detector is limited by cost and by the desire to limit the amount of material placed in front of the calorimeters. The pixel detector is the innermost detector. It covers the vertex region and has the highest granularity. Located directly outside the beampipe, each pixel in the barrel region is 50μm wide in $R-\phi$ and 400 μm long. Each track crosses 3 pixel layers on average. The intrinsic precision of the barrel is 10 μm (in $R-\phi$) and 115 μm (in z), while it is 10 μm (in $R-\phi$) and 115 μm (in R) in the end-cap disks. The middle tracker in the ID is the silicon SCT. It consists of four double layers of stereo silicon strips. That is, each set of stereo strips contains one strip aligned in the azimuthal direction and one rotated by 40 mrad with respect to the first one. This configuration provides twodimensional capabilities. The end-cap region is constructed similarly, with a set of strips running radially and a second set of stereo strips at an angle of 40 mrad to the first.

The TRT, which is a set of straw tube trackers, is located outside of the SCT. The straw (Polyimide) drift tubes are 4mm in diameter and are filled with gas (70% Xe, 27% CO_2, and 3% O_2) and contain an anode (tungsten plated with gold) running through the center. The straws are 144 cm long in the barrel region and extend parallel to the beam pipe. The TRT registers a large number of hits, typically 36 per track. This high number of hits compensates for its lower intrinsic precision of 130 μm per straw. It is also used for electron identification by detecting transition radiation photons.

Track reconstruction efficiencies for muons, pions, and electrons are shown in Figure 2.7(a) and reconstruction efficiencies for pions as a function of $|\eta|$ for several values of transverse momentum are shown in Figure 2.7(b).



(a) Track reconstruction efficiencies as a function of $|\eta|$ for muons, pions and electrons with $p_T = 5$ GeV. The inefficiencies for pions and electrons reflect the distribution of material in the inner detector as a function of $|\eta|$. Figure from [7].

(b) Track reconstruction efficiencies as a function of $|\eta|$ for pions with $p_T = 1$, 5 and 100 GeV. Figure from [7].

2.5.2 Calorimetry

The energy of particles in ATLAS is measured by a system of calorimeters in the barrel and end-cap regions. The calorimeter system has two main components: the electromagnetic (EM) and hadronic calorimeters. The barrel region of the EM calorimeter provides fine granularity for precision measurements while the barrel and endcap, along with the hadronic calorimeters provide appropriate jet and missing transverse energy measurements. A view of the calorimeter system is shown in Figure 2.7.

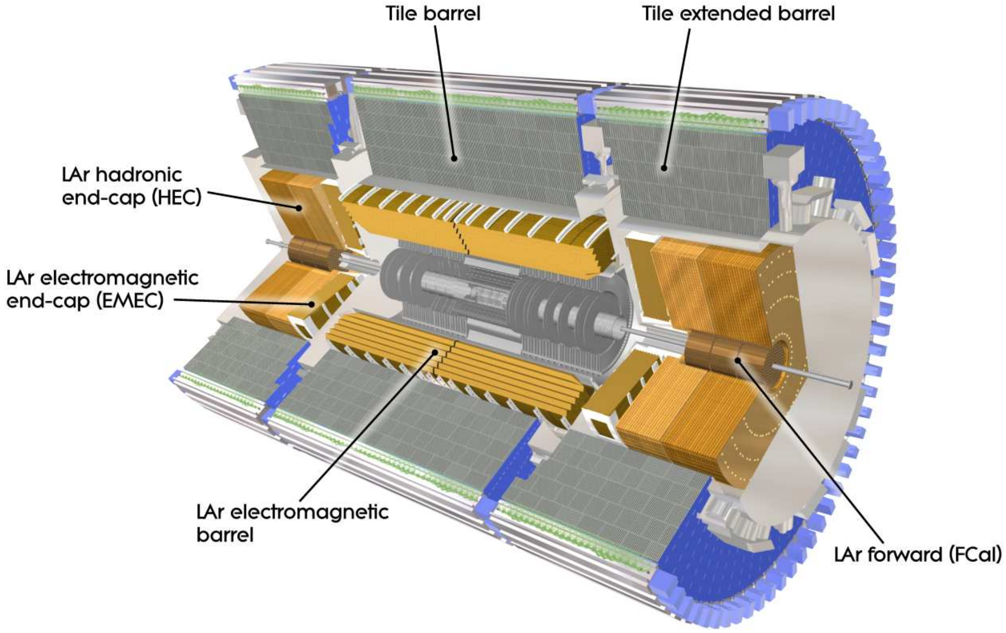

Tile barrel
Tile extended barrel
LAr hadronic end-cap (HEC)
LAr electromagnetic end-cap (EMEC)
LAr electromagnetic barrel
LAr forward (FCal)

Figure 2.7: Cut-away view of the ATLAS calorimeter system. Copyright: The ATLAS Experiment at CERN, [7].

The EM calorimeters are located outside of the inner detector. They consist of one barrel ($|\eta| < 1.475$) and two end-cap components ($1.375 < |\eta| < 3.2$), each divided into 3 layers. Each is constructed of layers of lead in an accordion shape with LAr sandwiched in between. The accordion shape allows for multiple absorbing plates without the need for azimuthal supports and gaps for readout cables, which would cause dead regions. Immediately before the EM calorimeter is a thin LAr layer, a presampler detector which covers the barrel region from $|\eta| < 1.52$ and the endcap up to $|\eta| < 1.8$. The first layer of the

EM calorimeter is finely segmented in η (0.003×0.1 in $\Delta\eta \times \Delta\phi$), which enhances precision measurements of the position of individual particles, such as the two photons from π^0 decay. It is referred to as the η strip layer elsewhere in this thesis. Angular resolution in the central region of the EM calorimeter is best in the strip layer and ranges in the last two layers from 0.025×0.1 in $\Delta\eta \times \Delta\phi$ and ranges to 0.050×0.025 in the third layer. This granularity is fine enough to detect the position of electrons and photons precisely in the area covered by the ID, which is important for π^0 reconstruction. The total thickness of the EM calorimeter is $> 22 (> 24)$ radiation lengths X_0 in the barrel(end-cap) region, where a radiation length is the amount of material traversed for an electron's energy to be reduced to $\frac{1}{e}$ of its original energy on average. The geometry of the barrel component of the EM calorimeter is shown in Figure 2.8. The energy response for hadronic tau decays with a reconstructed electromagnetic subcluster is shown in Figure 2.9.

The hadronic calorimeter system consists of the tile calorimeter in the barrel region ($|\eta| < 1.0$ and extended barrels from $0.8 < |\eta| < 1.7$), and the LAr calorimeter in the end-cap ($1.5 < |\eta| < 3.2$). The tile calorimeter is a sampling calorimeter, segmented into 3 layers. The absorber material is steel and it uses scintillating tiles as the active material. The thicknesses of the 3 layers are 1.5, 4.1, and 1.8(1.5, 2.6, and 3.3) interaction lengths λ for the barrel(extended barrel), where an interaction length is the average distance travelled by a hadron before interacting. The hadronic end-cap calorimeter is made of 2 wheels of 32 wedges each per end-cap. Each wedge is a series of copper plates with LAr gaps.

The LAr forward calorimeter is located radially inside the endcap calorimeters. Its three modules in each endcap include one copper module optimized for EM measurements and two tungsten layers for hadronic measurements. The forward calorimeter is approximately 10 interaction lengths deep.

2.5.3 ATLAS Software

The ATLAS detector provides readout in the form of many electronic signals. These signals must be converted into quantities that characterize the particles in a process called reconstruction. A software framework called Athena has been built to handle the reconstruction and aid in physics analysis. Monte Carlo sample sets have been generated to simulate physics and the detector response in the ATLAS detector. The simulation begins with the

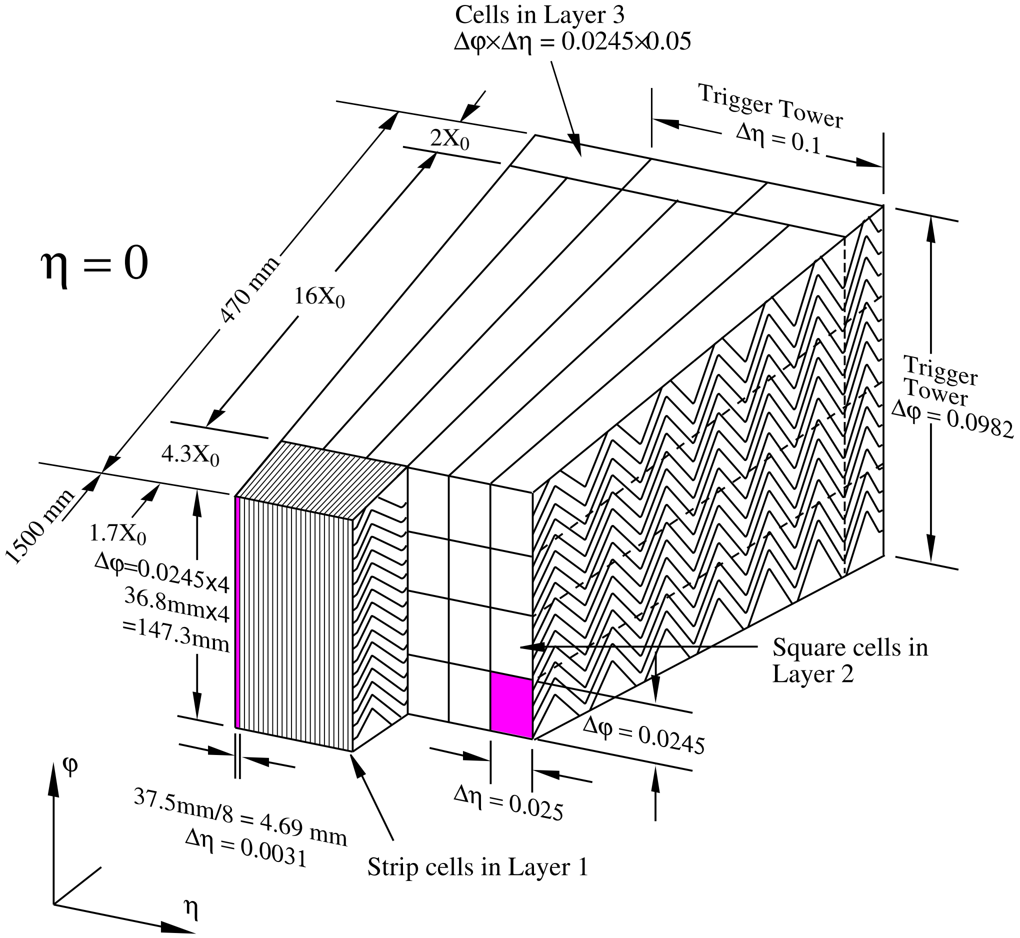


Figure 2.8: Sketch of a barrel module of the EM calorimeter, where the different layers are clearly visible with the ganging of electrodes in ϕ. The granularity in eta and phi of the cells of each of the three layers and of the trigger towers is also shown. Figure from [7].

generation of the interactions by a program such as Pythia [13]. The energy depositions in each detector due to these events are simulated and these depositions are turned into the electronic signals that each detector would put out in a process called digitization. This is achieved using software called Geant4 [14] . Either these simulated MC digitized signals or real signals from data are fed to reconstruction.

Reconstructed data contain particle kinematic and identification information, which can be used to do physics analysis.

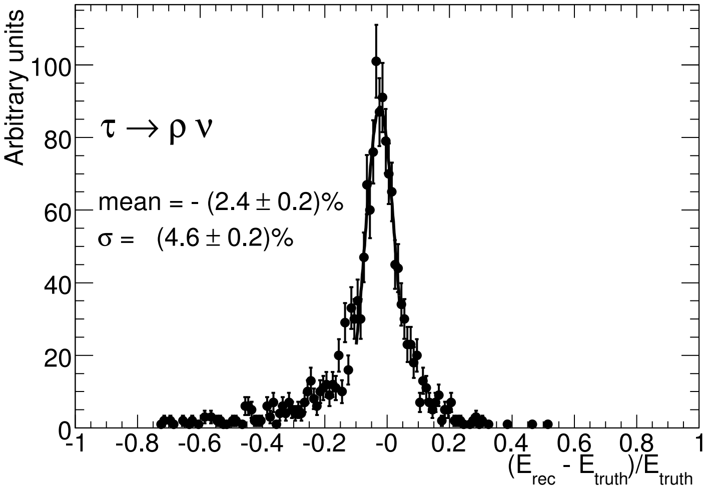


Figure 2.9: The energy response obtained for the visible energy from $\tau \rightarrow \rho\nu$ events using candidates with one π^0 subcluster. Figure from [8].

Chapter 3

The Tau Lepton

3.1 Introduction to Taus

In 1975, a team at the Stanford Linear Accelerator announced the observation of

> "64 events of the form
>
> $$e^{+} + e^{-} \rightarrow e^{\pm} + \mu^{\mp} + \geq 2 \text{ undetected particles}$$
>
> for which we have no conventional explanation"

when the center-of-mass energies reached at least 4 GeV. It was concluded that

> "the signature e-μ events cannot be explained by the production and decay of any presently known particles or as coming from any of the well-understood interactions which can conventionally lead to an e and a μ in the final state" [15].

The explanation turned out to be the tau, which falls in to the third and final of the known lepton generations. The $e-\mu$ events observed by Perl et al were explained by $e^{+}e^{-} \rightarrow \tau^{+}\tau^{-}$ production followed by a subsequent tau decay:

$$\tau^{\pm} \rightarrow \mu^{\pm}\nu_{\tau}\nu_{\mu}$$

$$\tau^{\mp} \rightarrow e^{\mp}\nu_{\tau}\nu_{e}$$

Table 3.1: Leading decay modes for τ^- [1]. $h^\pm$ stands for $\pi^\pm$ or $K^\pm$.

Leptonic Decay Modes	
Decay Mode	Branching Fraction
$e^- \bar{\nu}_e \nu_\tau$	17.85 ± 0.05 %
$\mu^- \bar{\nu}_\mu \nu_\tau$	17.36 ± 0.05 %
Total Leptonic	35.21 %

Leading Hadronic Decay Modes	
Decay Mode	Branching Fraction
$\pi^- \nu_\tau$	10.91 ± 0.07 %
$\pi^- \pi^0 \nu_\tau$	25.52 ± 0.10 %
$\pi^- \pi^+ \pi^- \nu_\tau$	8.99 ± 0.06 %
$\pi^- \pi^0 \pi^0 \nu_\tau$	9.27 ± 0.12 %
$h^- \omega \nu_\tau$	1.99 ± 0.08 %
$\pi^- \pi^+ \pi^- \pi^0 \nu_\tau$	2.70 ± 0.08 %
$\pi^- 3\pi^0 \nu_\tau$	1.04 ± 0.07 %
Total Hadronic	64.79 %

As with all charged leptons, the tau is a spin $\frac{1}{2}$ particle with a charge of –1. It is the heaviest lepton, with its measured mass averaged to 1776.84 ± 0.17 MeV [1], about 3500 times the mass of the electron.

Unlike the electron, which is stable, the tau has an average mean lifetime of $290.5 \pm 1.0 \times 10^{-15}$ sec ($c\tau \sim 87\mu m$) [1]. It has two general decay modes: leptonic and hadronic. The leading decay modes are listed in Table 3.1. To conserve charge, all decay modes have an odd number of charged daughters in the final state and while modes with 5 charged daughters do exist, the branching fraction is very small (~ 0.001 %). Similarly, taus do decay to charged hadrons other than $\pi^\pm$ (such as $K^\pm$), but the branching fractions are significantly smaller than the modes listed in Table 3.1. Hadronic taus are often classified by the number of stable charged decay products (“prongs”) and in practice they are referred to as “1-prong” or “3-prong”.

3.2 Tau Reconstruction

The lifetime of the tau is small enough that the leptonic decay modes are difficult to distinguish from prompt primary leptons. The tau reconstruction program in ATLAS is therefore limited to hadronically decaying taus. All hadronic decay modes with a significant branching fraction are composed of one or more $\pi^\pm$ and a ν_τ, and may include one or more π^0, which further decays by $\pi^0 \rightarrow \gamma\gamma$. For detection, these 3 decay components require the use of the tracking system and the EM and hadronic calorimeters for $\pi^\pm$ and the EM calorimeter for π^0. The daughter neutrinos limit the reconstructed tau energy and require a good E_T^{miss} resolution if the invariant mass of the object decaying into the tau is to be measured.

Taus are reconstructed in the ATLAS offline reconstruction software. The tau reconstruction algorithm uses objects which have been previously reconstructed, such as tracks and clusters of energy depositions. The tau reconstruction relies on these other objects and is therefore considered a higher level reconstruction algorithm. Tracks, for example, are formed from hits using the ATLAS track reconstruction software (described in Chapter 2.2) and these are then read in by the tau reconstruction software. For good tau reconstruction and identification, it is important to be able to reconstruct impact parameters and count isolated charged tracks accurately.

Historically, two reconstruction algorithms, one calorimeter seeded (named TauRec) and one track seeded (named Tau1P3P), have been used. These have recently been merged so that both algorithms are run and tau candidates are classified as track seeded only, calo seeded only, or calo+trk seeded (seeded by both). The details of these algorithms and the order in which they run are discussed in Sections 3.2.1 and 3.2.2.

Each algorithm provides a list of discriminating variables useful for tau identification, which will be discussed in Chapter 4. Many discriminating variables are unique to one particular algorithm so that the calo+trk candidates benefit from a more extensive list of available variables. The details of the reconstruction algorithms have been documented within ATLAS [8, 2].

Along with tau reconstruction, the reconstruction of π^0 subclusters within the tau decay is possible due to the high granularity of the electromagnetic calorimeter [8]. A topological clustering algorithm (see Section 3.2.2) is used to identify electromagnetic subclusters with $E_T > 1$ GeV separated from the impact point of tracks in the middle layer by $\Delta R < 0.0375$. It further requires the subclusters to deposit more than 10% of their energy in the strip+presampler, which helps distinguish against hadronic $\pi^\pm$. Further corrections were also are made to reduce the energy due to close or overlapping $\pi^\pm$.

3.2.1 The Track Seeded Algorithm (Tau1P3P)

A good quality track is a seed to this algorithm. The algorithm then searches for additional tracks within $\Delta R < 0.2$ of this track seed and which pass the “Associated Track” criteria. Candidates with more than 8 tracks and multi-track candidates whose charge is not $|Q| = 1$ are excluded.

The track seeded algorithm searches for tracks with $p_T > 6$ GeV which pass the track quality criteria in Table 3.2, where

- N_{Si} is the number of hits in the SCT.
- normalized χ^2 is a function expressing the overall fit of the tracks.
- Pixel Hits N_{pixel} is the number of hits in the pixel detector.
- B-layer Hits N_{blay} is the number of hits in the inner-most pixel layer.
- High/Low Threshold Hit Ratio $N^{HT}_{TRT}/N^{LT}_{TRT}$ is the ratio of high to low threshold hits in the TRT, which helps in the identification of electrons.

Table 3.2: Track quality criteria for seed tracks and default associated tracks (track-based candidates), and loose tracks (calo-based candidates) [2].

Track Criteria	Seed Track	Associated Track	Loose Track
p_T (GeV) $>$	6	1	1
$\lvert\eta\rvert <$	2.5	2.5	2.5
impact parameter d_0 (mm) $<$	1	1	1.5
Silicon Hits $N_{Si} \geq$	8	8	6
TRT Hits $N_{TRT} \geq$	10	no cut	no cut
normalized $\chi^2 <$	1.7	1.7	3.5
Pixel Hits $N_{pixel} \geq$	no cut	1	1
B-layer Hits $N_{blay} \geq$	no cut	1	no cut
High/Low Threshold Hit Ratio $N^{HT}_{TRT}/N^{LT}_{TRT} <$	no cut	0.2	no cut

If only 2 tracks associated to a candidate are found, the algorithm loosens the associated track criteria by dropping the χ^2 and $N^{HT}_{TRT}/N^{LT}_{TRT}$ criteria in the hopes of matching a third track.

The direction of the candidate is defined by the direction of the track seed at the primary vertex in the case of a 1-prong candidate; it is the p_T- weighted barycenter of the tracks for a 3-prong candidate.

The energy of these candidates is calculated using an energy flow algorithm [2, 8, 16]. This algorithm is designed to model tau energy well and underestimate energy from QCD jets, the largest background to taus (see Chapter 4). It is defined particularly to exploit the fact that the tau is expected to deposit a charged hadronic energy component from one or more $\pi^{\pm}$'s and a neutral electromagnetic component from any additional π^0. Unlike a typical jet, a hadronic tau decay contains few if any neutral components and no neutral hadronic components, as the daughter π^0's decay into photons before interacting with the detector. A tau decay therefore deposits minimal neutral hadronic energy and its track(s) tend to have a higher transverse momentum than a jet in the same energy range.

The energy from a tau jet ideally consists of the energy from charged tracks plus the neutral electromagnetic energy depositions from photons decaying from π^0's:

$$E_T^{\text{Ideal}} = E_T^{\text{em}} + \sum p_T^{\text{track}}. \tag{3.1}$$

Energy depositions from decay daughters often overlap each other or leak outside of small energy clusters and corrections are therefore needed. The energy flow algorithm is defined similarly, but includes terms to correct for leakage and overlapping showers. The full energy flow calculation is

$$E_T^{\text{eflow}} = E_T^{\text{emcl}} + E_T^{\text{neuEM}} + \sum p_T^{\text{track}} + \text{res}E_T^{\text{chrgEM}} + \text{res}E_T^{\text{neuEM}}. \tag{3.2}$$

This definition does not allow for neutral hadronic components to the transverse energy, which results in an underestimation of the energy of a jet.

The energy of the reconstructed tau candidate is categorized according to deposition in order to define terms in E_T^{eflow}:

- E_T^{emcl}: the transverse energy summed over all cells labelled as purely photonic, which is found by a clustering algorithm. It is seeded by a cell in the electromagnetic calorimeter with $E_T > 0.2$ GeV and includes all cells from the first three samplings of the LAr calorimeter within $\Delta\eta \times \Delta\phi = 0.0375 \times 0.0375$ of the seed. This energy is mainly deposited from the decay of the π^0 daughter into photons. Only clusters isolated from tracks and which pass a hadronic leakage cut are used;

- E_T^{chrgEM}: the charged transverse energy in the EM calorimeter. That is, the total scalar sum of energy deposited in the electromagnetic calorimeters within a window of $\Delta\eta \times \Delta\phi = 0.0375$ of a valid track. There are no charged electromagnetic daughters in a hadronic tau decay but the charged hadronic daughters deposit energy in the electromagnetic calorimeters. Likewise, E_T^{chrgEM01} is the same scalar sum, but only for cells within the first two EM layers. The energy deposited by the $\pi^{\pm}$ in the first two layers of the electromagnetic calorimeter is minimal. These are used in defining the correction terms;

- E_T^{neuEM}: the neutral electromagnetic transverse energy. This is the energy in the electromagnetic calorimeter that was not flagged either as E_T^{emcl} or E_T^{chrgEM}. In the case of a tau, it is comprised mainly of energy from γ and π^0's that either did not seed isolated clusters and did not overlap any tracks or whose electromagnetic energy depositions were beyond that of the cluster radius. There is also a contribution from the leakage of the charged hadronic component that showered early and was outside the narrow cone defined for E_T^{chrgEM}.

- $\sum p_T^{\mathrm{track}}$: the sum of the transverse track momentum;

- E_T^{chrgHAD}: the charged hadronic transverse energy. It includes the energies from all cells in the hadronic calorimeter within $\Delta R < 0.2$ of a valid track and is flagged for use with the correction terms;

Two correction terms are also defined for the energy flow algorithm:

- $\mathrm{res}E_T^{\mathrm{chrgEM}}$ corrects for leakage of γ showers into cells flagged for E_T^{chrgHAD}. In the case of 3 prong decays and 1 prong decays in which $E_T^{\mathrm{chrgEM}}/p_T^{\mathrm{track}} < 0.05$, the correction is $\mathrm{res}E_T^{\mathrm{chrgEM}} = max(0., E_T^{\mathrm{chrgEM}} - 0.7 \times \sum p_T^{\mathrm{track}})$. For 1 prong decays in which the charged track and neutral electromagnetic showers are minimally overlapped, such that $E_T^{\mathrm{chrgEM}}/p_T^{\mathrm{track}} > 0.05$ and $E_T^{\mathrm{chrgHAD}}/p_T^{\mathrm{track}} > 0.4$, the correction is res$E_T^{\text{chrgEM}} = min(E_T^{\text{chrgEM}}, 2.5 \times E_T^{\text{chrgEM01}}$. For those 1 prong decays in which $E_T^{\text{chrgEM}}/p_T^{\text{track}} > 0.05$ but $E_T^{\text{chrgHAD}}/p_T^{\text{track}} < 0.4$, then res$E_T^{\text{chrgEM}} = max(0., E_T^{\text{chrgEM}} - 0.65 \times p_T^{\text{track}})$. This has been defined empirically for late hadronic showers in which little energy is deposited in the first layers of the electromagnetic calorimeter.

- resE_T^{neuEM} corrects for double counting from leakage of energy from charged hadronic particles in the electromagnetic calorimeter which was deposited outside the associated track cone. It is defined as res$E_T^{\text{neuEM}} = -0.1 \times p_T^{\text{track}}$ and is applied in the third category above: when $E_T^{\text{chrgEM}}/p_T^{\text{track}} > 0.05$ but $E_T^{\text{chrgHAD}}/p_T^{\text{track}} < 0.4$, as long as res$E_T^{\text{neuEM}} + E_T^{\text{neuEM}} > 0$.

If no calorimeter seeds are found, the energy flow defines the default energy for this candidate.

3.2.2 The Calorimeter Seeded Algorithm (TauRec)

Once a track seeded candidate is established, the track seeded algorithm searches for topoclustered jets within $\Delta R < 0.2$ of the candidate and calls the calorimeter seeded algorithm if any are found. Once the track seeded algorithm is finished constructing candidates for the event, the calorimeter seeded algorithm is run to search for further calorimeter seeds that do not lie within $\Delta R < 0.2$ of the track seeds. It uses topoclustered jets in a cone algorithm of radius $\Delta R = 0.4$ as seeds.

The topological clustering (4-2-0) scheme used for tau reconstruction is seeded by a cell with energy $> 4\sigma$ in calorimeter noise. Cells directly neighbouring these seed cells which have energy $> 2\sigma$ in calorimeter noise are then added to the cluster, as are the cells neighbouring these ones which are also above 2σ in noise. Finally, once this base cluster is formed, all cells directly adjacent to the cluster are also added to the cluster [2, 17].

A cone jet algorithm forms a cone of uniform radius in $\eta - \phi$ space, centered by the energy distribution within the cone. In the case of a topocluster cone algorithm, the cone jet is seeded by clusters with p_T greater than some threshold (1 GeV in ATLAS). All objects (ie. clusters) within $\Delta R < 0.4$ are then combined with this seed. The four-momentum of the clusters within the cone are added and a cone center is defined as the center of the combined four-momenta. A cone of $\Delta R < 0.4$ is formed around this new direction and objects within the new cone are again used to calculate a new four-momentum. A new cone center is again defined and the process repeats until the direction of the cone does not change.

Topoclustered cone jets with $p_T > 10$ GeV and $|\eta| < 2.5$ are considered seeds to the calo-seeded tau reconstruction algorithm. Topoclustering resolution in the forward region ($|\eta| \gtrsim 2.5$) is poor due to the decrease in linear distance between jets and the larger cell size in $\eta - \phi$ [17].

The energy of the calo-seeded tau candidate is calculated as a sum of cell-weights derived from MC simulations (H1-style calibration). They are a function of cell energy density and position. These weights are then multiplied by an additional factor which is tuned to an energy scale for the tau using an MC energy distribution and is often less than 1.

Tracks within $\Delta R < 0.3$ of the calo-seed are associated to the tau candidate if they meet the quality requirements in Table 3.2. The direction of the tau is defined as the E_T-weighted barycenter of the calorimeter cells which seeded the candidate.

3.2.3 Reconstruction Performance

Overall tau reconstruction efficiency for either algorithm is 98% and efficiency for reconstruction by both seeds is 74% for candidates above 10 GeV. The performance of the algorithms can be seen in Figure 3.1.

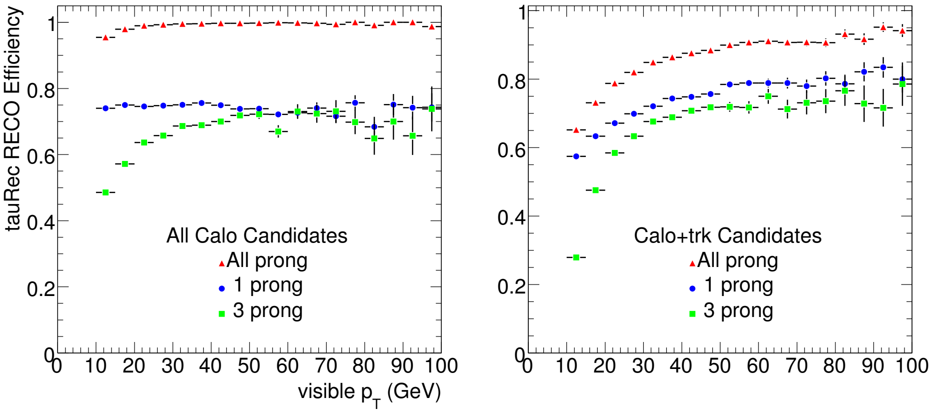


Figure 3.1: Reconstruction efficiency for all calo seeds (left) and calo+trk seeds (right) for $Z \rightarrow \tau\tau$ events. The “All prong” plot does not require any track matching; the 1(3) prong plots require exactly 1(3) true MC stable charged daughter(s) and 1(3) reconstructed tracks.

Chapter 4

Tau Identification

In ATLAS, the identification of taus is a separate stage of tau candidate analysis which occurs after tau reconstruction. Reconstruction, as described in Section 3.2, uses tracks and calorimeter clusters to construct objects which look like taus, the outcome of which is called a "tau candidate". Identification then selects the candidates which are most tau-like. The goal of tau reconstruction in ATLAS is to maximize the efficiency for finding true hadronic taus. Reconstruction is perfect when every tau that decays hadronically in the ATLAS detector is reconstructed as a tau candidate by reconstruction software. The trade-off is that a high rate of other objects faking taus are also reconstructed as tau candidates. Recall, for example, that calorimeter seeded tau candidates are identified by a jet algorithm with little requirements other than a minimum energy deposit. Most jets with a minimum energy should then be reconstructed as a calorimeter seeded candidate. Further background rejection therefore occurs during the identification process, in which tau candidates are analyzed by some criteria and scored according to how likely it is that they are really a tau.

4.1 Tau Production

Taus will be produced through electroweak decays of bosons, such as $W \to \tau\nu_\tau$ and $Z \to \tau\tau$ and possibly $H \to \tau\tau$. The expected product of cross-section and branching ratios for tau production from W and Z bosons, which are the most important samples of taus produced, for p-p collisions at $\sqrt{s} = 14$ TeV is 5540 pb and 458 pb respectively.

4.2 Jet Background

The major source of background to taus is jets, which is a general name for sprays of particles produced through a process called hadronization. Jets are produced through QCD processes, mainly qq, qg, and gg interactions. Isolated partons undergo fragmentation as they radiate gluons which form quark-antiquark pairs. Hadronization occurs when these quarks combine to produce stable colourless hadrons. The result is a stream of collinear hadrons with total momentum approximately the same as the original outgoing quark. A jet can contain any hadron kinematically allowed, but it is dominated $\pi^{\pm}$ and π^{0}. It is this stream of similar particles, many with tracks, that can be difficult to distinguish from a hadronic tau decay. In fact, it is possible for a jet to contain only a few pions so that their content is identical to a hadronic tau.

The most plentiful source of jets at the LHC energies will be from dijet events. These occur when two partons are scattered during a collision and subsequently hadronize, resulting in two back to back jets. Unlike the mediator in electroweak interactions, gluons interact with each other. Higher order corrections to cross-sections are significant and this, along with parton distribution function (PDF) uncertainties, is why the exact rate of expected dijet production is not known to high precision. The dijet cross-section for p-p collisions at $\sqrt{s} = 14$ TeV is expected to be around 1.9×10^{10} pb, more than 10^{6} times larger than the expected production due to W and Z decays. This overwhelming rate of background objects, many of which look very much like hadronic tau decays, requires a strong tau identification program for ATLAS.

4.3 Tau Identification Variables

A jet is produced through the hadronization process and contains many particles within a localized area in $\eta - \phi$ (ΔR) space. During shower development, some of these original particles also decay. The hadronic tau jet, however, starts with a weak decay so that the resulting daughter particles have a smaller opening angle, which contributes to the fact that they tend to have narrower particle showers than jets. Furthermore, taus usually have fewer charged tracks and carry larger fractions of momentum in the leading tracks than jets in the same energy range do.

Variables which exploit these differences have been defined and are calculated automatically by the reconstruction algorithm. Some variables are calculated only for one of the two types of seeded algorithms while some are calculated for both types of seeds (the results may not be identical due to slightly different inputs or definitions). A candidate that has been seeded by both algorithms therefore possesses the maximum amount of discriminating information available. Below is a set of variable descriptions and definitions, largely adapted from documentation for each algorithm, as indicated by the citations in each entry. Variable distributions for a signal sample of taus from a $Z \to \tau\tau$ decay and a background of jets from dijet events over a range of p_T are provided with the definitions.

4.3.1 Calorimeter Seeded Algorithm

The calorimeter seeded algorithm reconstructs the following discriminating variables relevant to the studies in this work:

- **The centrality fraction**

  The centrality fraction quantifies the characteristic that taus carry the majority of their energy in the central region of the jet cone by

  $$C_{frac} = \frac{\sum_i E_{T,i}}{\sum_j E_{T,j}}, \tag{4.1}$$

  where the indices i and j run over all calorimeter cells in a cone around the cluster barycentre with $\Delta R < 0.1$ and $\Delta R < 0.4$, respectively, and $E_{T,i}$ and $E_{T,j}$ denote the transverse cell energies [2]. The centrality fraction distribution is shown in Figure 4.1.

- **The electromagnetic radius R_{em}**

  To exploit the smaller transverse shower profile in tau decays, the electromagnetic radius R_{em} is used, defined as

  $$R_{\text{em}} = \frac{\sum_{i=1}^{n} E_{T,i}\sqrt{(\eta_i - \eta_{\text{cluster}})^2 + (\phi_i - \phi_{\text{cluster}})^2}}{\sum_{i=1}^{n} E_{T,i}}, \tag{4.2}$$

  where i runs over all cells in the electromagnetic calorimeter in a cluster with $\Delta R <$ 0.4. The quantities η_i, ϕ_i, and $E_{T,i}$ denote their position and transverse energy. Cells

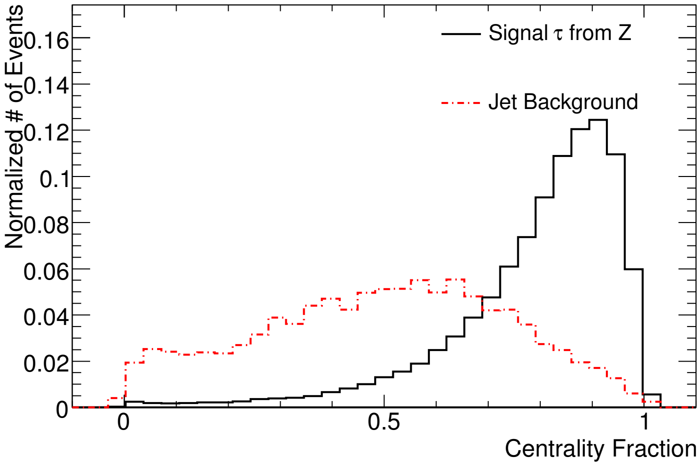


Figure 4.1: The centrality fraction.

may have different sizes depending on the layer and their η value. The size varies from $\Delta\eta \times \Delta\phi = 0.003 \times 0.1$ in the η-strip region of the barrel to 0.025×0.025 for the second calorimeter layer. Furthermore, the segmentation ranges from this lower limit of 0.003×0.1 in the central regions to 0.1×0.1 in $2.5 < |\eta| < 3.2$. This change in segmentation in η leads to a dependence of the performance on η as the resolution decreases in the forward region. This variable shows good discrimination power at low $\mathrm{E_T}$ but becomes less effective at higher $\mathrm{E_T}$ as the jets become narrower with high $\mathrm{E_T}$ [8]. The variable distribution can be seen in Figure 4.2.

- **The hadronic radius** R_{had}

  The hadronic radius R_{had} is defined analogously:

$$R_{\text{had}} = \frac{\sum_{i=1}^{n} E_{T,i} \sqrt{(\eta_i - \eta_{\text{cluster}})^2 + (\phi_i - \phi_{\text{cluster}})^2}}{\sum_{i=1}^{n} E_{T,i}}, \tag{4.3}$$

  where i runs over all cells in the hadronic calorimeter in a cone of $\Delta R < 0.4$ [2]. Like the EM radius, the narrower shower profile of the tau results in a smaller hadronic radius than that of a typical jet. The variable distribution can be seen in Figure 4.3.

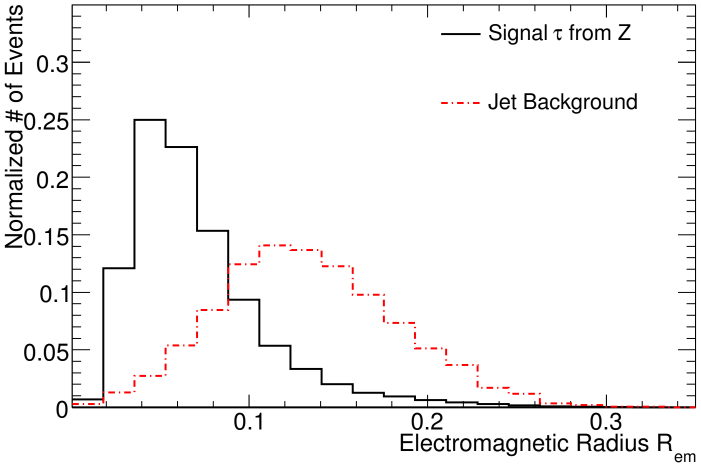


Figure 4.2: The electromagnetic radius.

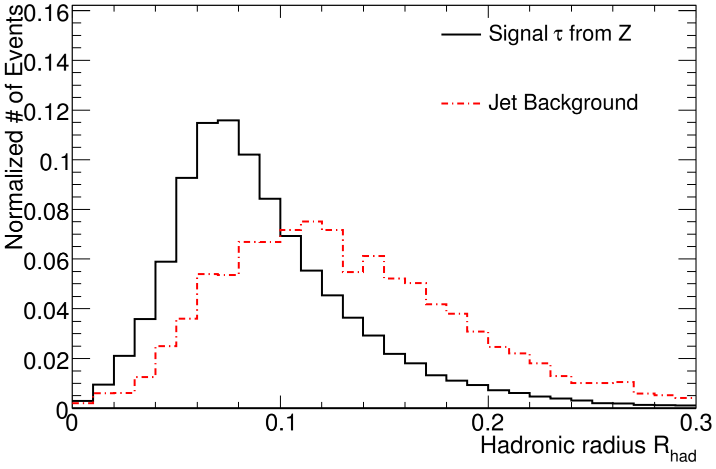


Figure 4.3: The hadronic radius.

- **Isolation in the calorimeter**

  Jets built on clusters from hadronic tau decays are well collimated and therefore a tight isolation criterion can be used. Here a ring of $0.1 < \Delta R < 0.2$ was chosen as the isolation region and the quantity

$$\Delta E_{\mathrm{T}}^{12} = \frac{\sum_i E_{T,i}}{\sum_j E_{T,j}}, \tag{4.4}$$

is calculated, where the indices i and j run over all electromagnetic calorimeter cells in a cone around the cluster axis with $0.1 < \Delta R < 0.2$ and $\Delta R < 0.4$, respectively, and $E_{T,i}$ and $E_{T,j}$ denote the transverse cell energies.

Like R_{em}, the ΔE_T^{12} distribution shows an $\mathrm{E_T}$ dependence and becomes narrower with increasing $\mathrm{E_T}$. This variable also depends on the event type and is expected to be less effective for events with higher hadronic activity, like e.g. $t\bar{t}$ events. The variable distribution can be seen in Figure 4.4.

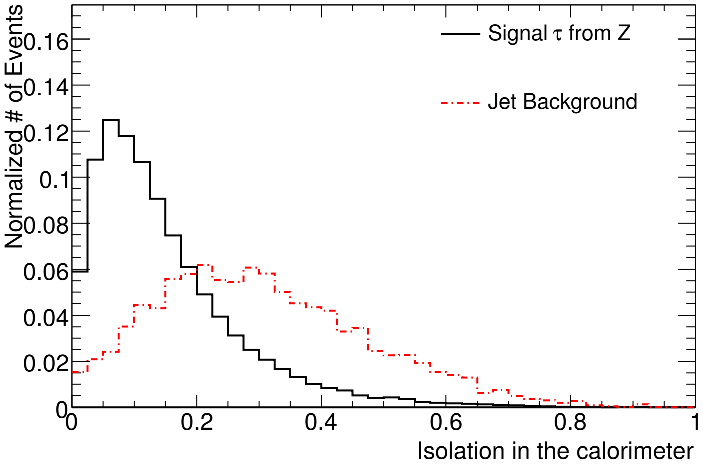


Figure 4.4: The isolation in the calorimeter.

- **Transverse energy width in the η strip layer**

  The transverse energy width $\Delta\eta$ is defined as

$$\Delta\eta = \sqrt{\frac{\sum_{i=1}^{n} E_{Ti}^{\mathrm{strip}} \left(\eta_i - \eta_{\mathrm{cluster}}\right)^2}{\sum_{i=1}^{n} E_{Ti}^{\mathrm{strip}}}}, \qquad (4.5)$$

  where the sum runs over all strip cells in a cone with $\Delta R < 0.4$ around the cluster barycentre and E_{Ti}^{strip} is the corresponding strip transverse energy [2]. This is denoted as stripWidth2 in Chapter 6. The variable distribution can be seen in Figure 4.5.

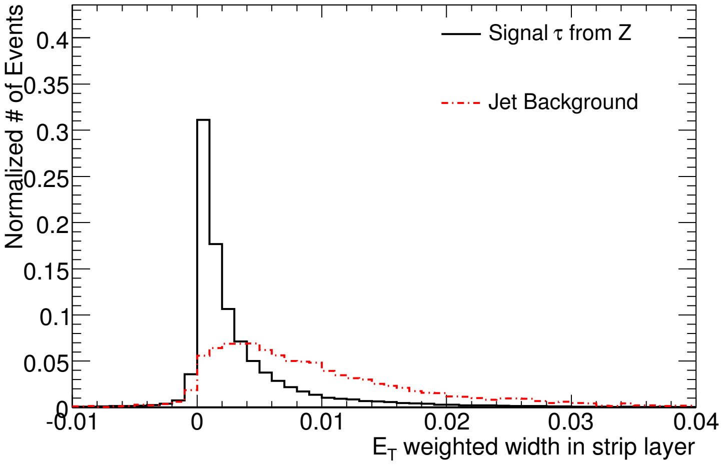


Figure 4.5: The transverse energy width in the η strip layer.

- **Charge of the tau candidate**

  The charge of a tau candidate is defined as the sum over the charge(s) of the associated track(s). A tau is more likely to be reconstructed with the correct charge $|Q| = 1$ than a jet. The misidentification of the charge on the level of a few percent shows almost no $\mathrm{E_T}$ dependence [2]. The charge distribution can be seen in Figure 4.6.

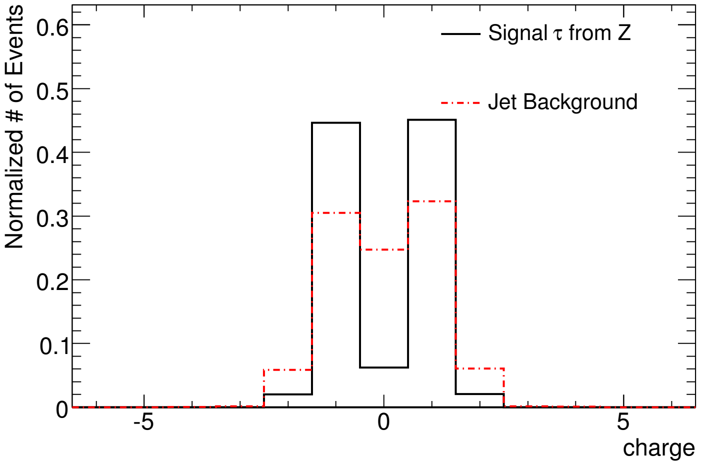
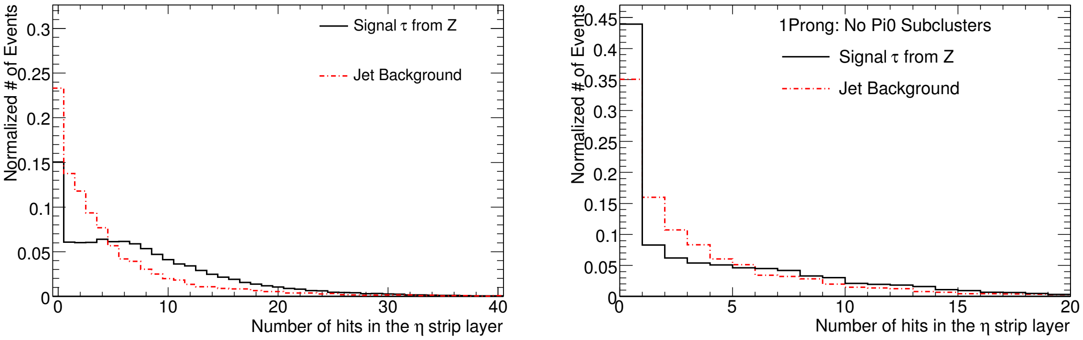


Figure 4.6: The charge calculated by the associated tracks.

- **Number of hits in the η strip layer**

  This is the number of hits in the η direction in the finely segmented strip detector, N_{strip}, in the first layer of the electromagnetic barrel calorimeter. Cells in the η strip layer within $\Delta R < 0.4$ around the cluster barycentre are counted as hits if the energy deposited exceeds 200 MeV. In contrast to jets, a significant fraction of 1 prong tau leptons deposit nearly no energy in the η strip layer ($\tau^{\pm} \rightarrow \pi^{\pm}\nu$ decays have only a hadronic component) and the number of corresponding hits is small [8]. The variable distribution can be seen in Figure 4.7.



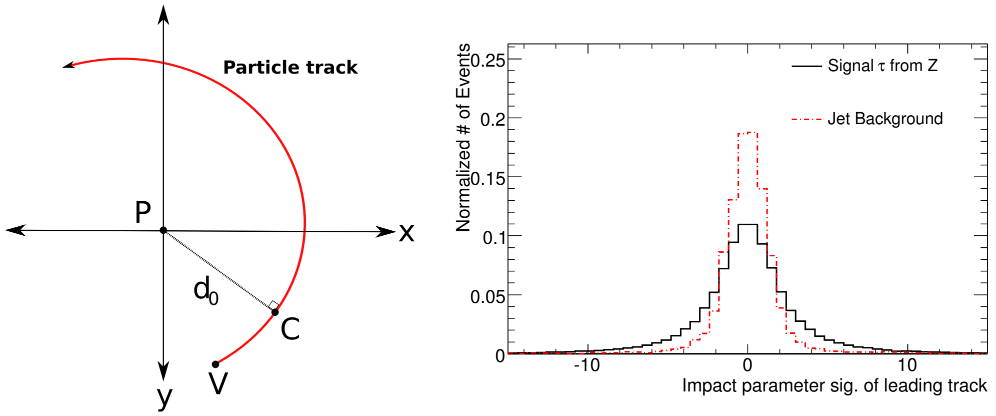
(a) All calo+trk seeded candidates

(b) Calo+trk seeded candidates which have 0 reconstructed π^0 subclusters

Figure 4.7: Number of hits in the η strip layer.

- **Lifetime signed pseudo impact parameter significance of leading track**

  The impact parameter of the leading track may be calculated for hadronic tau decays even in the case of 1 prong decays, where a secondary vertex cannot be calculated. At present only a 2-dimensional impact parameter, also called the pseudo impact parameter, is used. It is defined as the distance from the beam axis to the point of closest approach of the track in the plane perpendicular to the beam axis, as shown in Figure 4.8(a). The lifetime of the tau ($c\tau = 87\mu m$) is significant enough for a detectable separation between the primary vertex and its decay vertex. A jet does not have such a separation, so the impact parameter from the leading track in the tau decay is expected to be larger and have higher resolution. From this information and from the jet axis, a quantity denoted as lifetime signed pseudo impact parameter significance, defined as $\mathrm{sig}_{d_0} = d_0/\sigma^2_{d_0}$ where σ is the impact parameter resolution, is calculated [8]. The variable distribution can be seen in Figure 4.8(b).



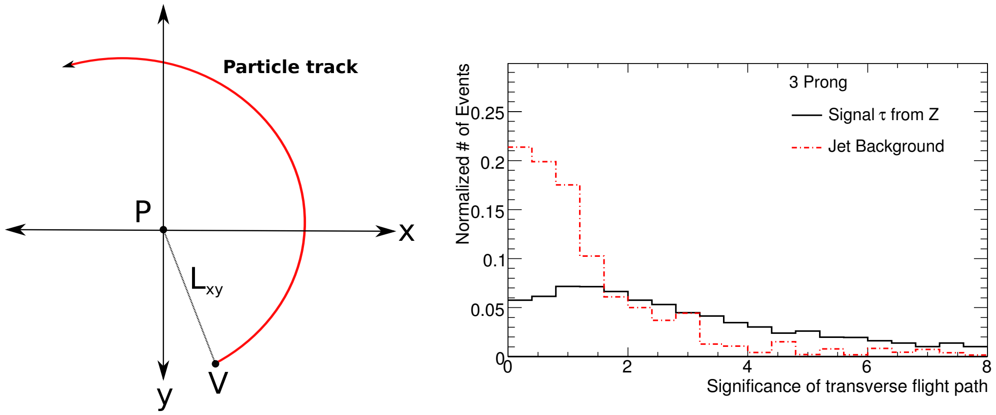
(a) Pseudo impact parameter, adapted from [18].

(b) Distribution of pseudo impact parameter significance of leading track.

Figure 4.8: Definition of pseudo impact parameter in the $x-y$ plane is shown in (a). P indicates the primary vertex of the collision, V is the secondary vertex, and d_0 is the impact parameter: the closest approach perpendicular the track to the z-axis. The impact parameter significance of the leading track is show for signal and background in (b).

- **Transverse flight path significance L_{xy}/σ_{Lxy}**

  For τ candidates with more than one loose track (associated tracks assigned by the calorimeter seeded algorithm), a vertex is reconstructed from the loose tracks using the adaptive vertex fitter [19]. The transverse flight path significance is defined by the transverse displacement of this vertex L_{xy} with respect to the primary vertex, divided by its uncertainty σ_{Lxy}. When the primary vertex resolution of the event is worse than the uncertainty of the beam spot, then the beam spot position and uncertainty is used instead of the primary vertex. As with the pseudo impact parameter significance, the transverse flight path significance is expected to be higher for taus than for jets. This variable is available for all candidates with more than one reconstructed track. The variable distribution can be seen in Figure 4.9.



(a) Transverse flight path.

(b) Transverse flight path significance distribution for 3 prong candidates.

Figure 4.9: The transverse flight path is shown in (a), where P is the primary interaction point and V is the secondary vertex, as in 4.8(a). Part (b) shows the transverse flight path significance distribution.

- **Number of tracks in small ring**

  The number of tracks with $p_T > 1$ GeV within the region between a smaller and larger radius in $\eta - \phi$ space is used for discrimination. The default radii used in this study are $\Delta r = 0.07$ and $\Delta R = 0.35$ [20]. This variable is named nTracksdrdR within the algorithm and in the analysis in Chapter 6. Hadronic tau jets tend to have fewer tracks within this ring than other jets do. The variable distribution can be seen in Figure 4.10.

- **$\boldsymbol{E_T}$ over $\boldsymbol{p_T}$ of the leading track**

  A large fraction of the energy is expected to be carried by the leading track of a tau decay and, consequently, the ratio of the cluster transverse energy $\mathrm{E_T}$ to the transverse momentum of the leading track p_{T1} ($\frac{\mathrm{E_T}}{p_{T1}}$) is expected to be small (close to 1). Values above one are expected from tau decay modes involving additional π^0s

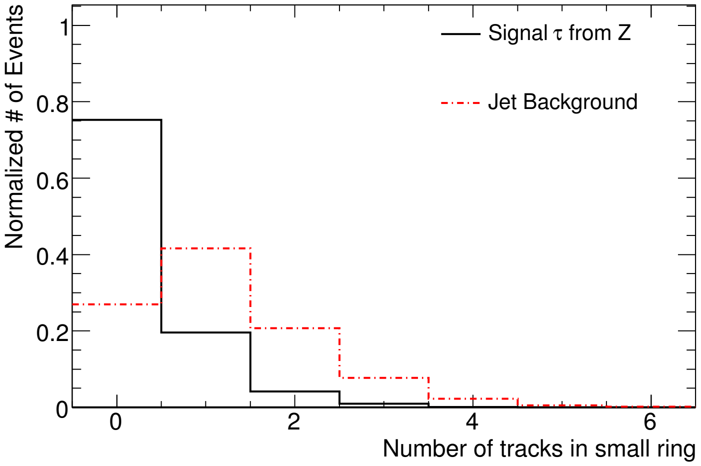


Figure 4.10: Number of tracks in small ring.

and for three-prong decays. This provides discrimination against QCD jets, which are expected to have a more uniform distribution of p_T among the tracks. They are also expected to have more additional neutral particles than hadronic taus. The E_T dependence is modest for tau decays but more pronounced for QCD jets, which tend to become more signal like with higher E_T [8]. The variable distribution can be seen in Figure 4.11.

The following quantities are also calculated by the calorimeter seeded algorithm. They are not directly used as discriminating variables, but are used in expressions (generally ratios) with other quantities to form further discriminating variables (see Chapter 6):

- **Total transverse momentum of associated tracks**

  The scalar sum of the transverse momentum of each track attributed to the tau candidate: $\sum p_T^{track}$.

- **Total transverse energy of tau candidate**

  The total transverse energy, E_T, measured by the calorimeters as described in Section 3.2.2.

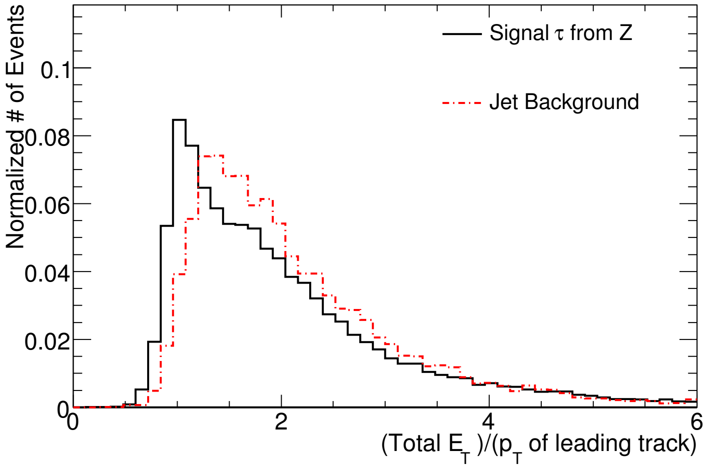


Figure 4.11: E_T over p_T of the leading track: E_T/p_{T1}.

- **Electromagnetic transverse energy of tau candidate**

  The electromagnetic transverse energy using MC based cell-weights (H1-style calibration). The electromagnetic portion includes the Presampler + EM1 + EM2 layers.

- **Hadronic transverse energy of tau candidate**

  The hadronic transverse energy calibrated using MC based cell-weights (H1-style calibration). The hadronic portion includes the cryo + EM3 + TILE1 + TILE2 + TILE3 layers.

4.3.2 Track Seeded Algorithm

The track seeded algorithm reconstructs the following discriminating variables relevant to the studies in this work:

- **Number of isolated tracks**

  Number of tracks, N_{trk}^{core}, in the isolation cone from the seed track [8]. The default cone is $0.2 \leq \Delta R_{iso} \leq 0.4$. Tau decays tend to have fewer charged tracks which are closer together than the tracks from jets, as seen in Figure 4.12.

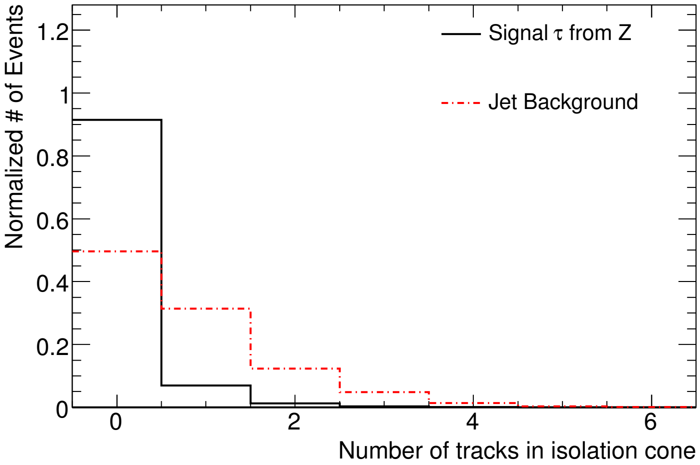


Figure 4.12: The number of isolated tracks.

- **The width of the energy in strips**

  The width of the energy deposition in the strips, $(\Delta\eta)^2$, calculated as the variance in the η coordinate, weighted by the transverse energy deposition in a given strip [8]:

$$(\Delta\eta)^2 = \frac{\sum(\Delta\eta^{\tau_{1p3p},strip})^2 \cdot E_T^{strip}}{\sum E_T^{strip}} - \frac{(\sum\Delta\eta^{\tau_{1p3p},strip} \cdot E_T^{strip})^2}{(\sum E_T^{strip})^2}. \qquad (4.6)$$

  Referred to as rWidth2Trk3P within the analysis in Chapter 6.

- **The width of tracks**

  The width of tracks, weighted with their transverse momenta, calculated as the variance (for candidates with more than one track) [2]:

$$TrackWidth = \frac{\sum(\Delta\eta^{\tau_{1p3p},track})^2 \cdot p_T^{track}}{\sum p_T^{track}} - \frac{(\sum\Delta\eta^{\tau_{1p3p},track} \cdot p_T^{track})^2}{(\sum p_T^{track})^2}. \qquad (4.7)$$

  The variable distribution can be seen in Figure 4.13

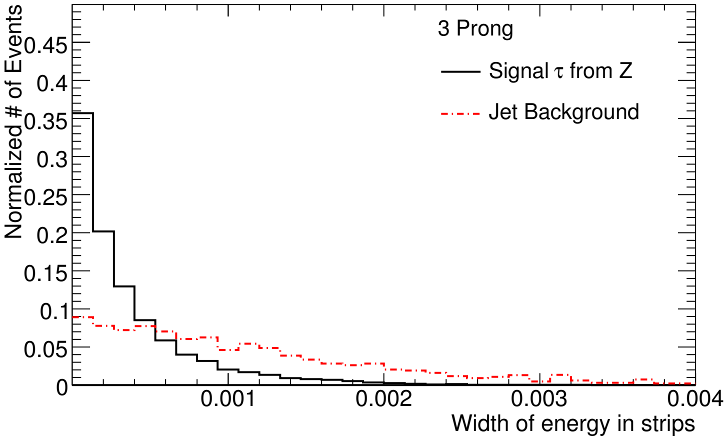


Figure 4.13: The width of the tracks for 3 prong candidates.

- **Invariant mass of the tracks,** M_{trk}

  Invariant mass of the tracks system (for candidates with more than one track) [2]. For true taus, this should not exceed the tau mass. Jets do not have such an upper limit on the invariant mass. Referred to as massTrk3P in Chapter 6.

  The variable distribution can be seen in Figure 4.14

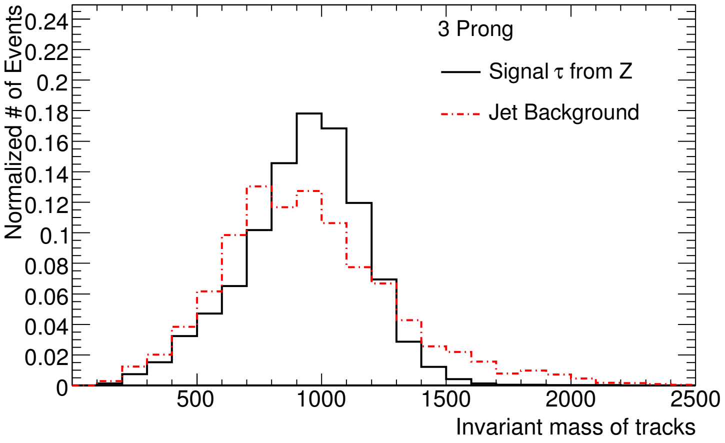
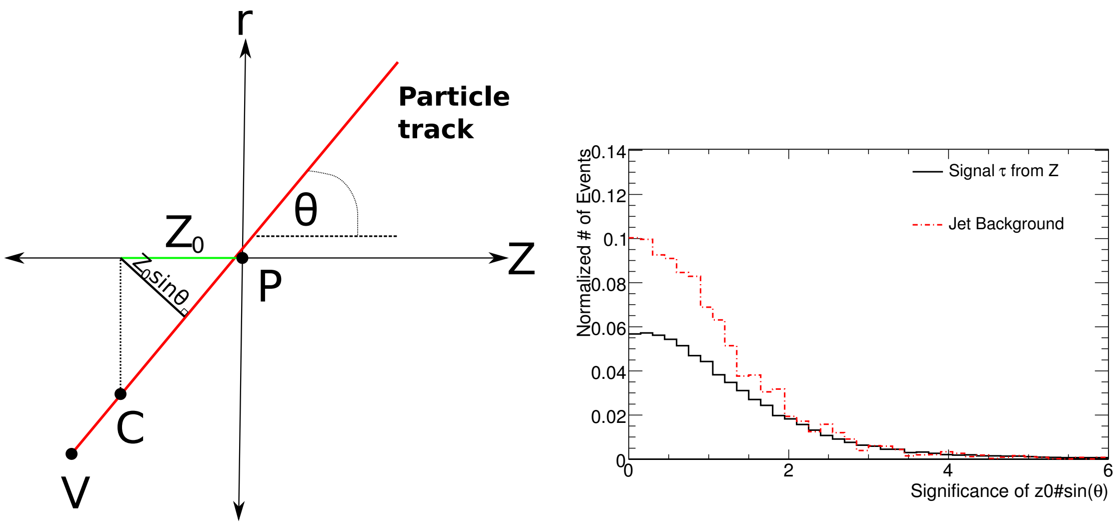


Figure 4.14: The invariant mass of the tracks for 3 prong candidates.

- **$\mathbf{Z_0}\sin(\theta)$ significance of leading track**

  The significance of $Z_0\sin(\theta)$ of the leading track, where Z_0 is the distance in the transverse plane between the track and the reconstructed primary vertex. It is multiplied by $\sin(\theta)$ to obtain the projection of Z_0 along the line perpendicular to the track [8]. See Figure 4.15(a) for clarification. The significance is defined as $\frac{Z_0\sin(\theta)}{\sigma_{Z_0\sin(\theta)}}$, where $\sigma_{Z_0\sin(\theta)}$ is the uncertainty in $Z_0\sin(\theta)$. As with the impact parameter, this is expected to be higher for taus than for jets. The variable distribution is shown in 4.15(b).



(a) $Z_0\sin(\theta)$, adapted from [18].

(b) Distribution of $\frac{Z_0\sin(\theta)}{\sigma_{Z_0\sin(\theta)}}$ of the leading track.

Figure 4.15: The definition of $Z_0\sin(\theta)$ is shown in (a). C is the point of closest approach in the $x-y$ plane to the primary vertex P. The points C, P, and the secondary vertex V correspond exactly with the points in Figure 4.8(a). The distribution of the significance of $Z_0\sin(\theta)$ is shown in (b).

- **Number of π^0 Subclusters**

  Isolated electromagnetic subclusters with no associated track are used to identify subclusters due to π^0 decays. The distribution is shown in Figure 4.16.

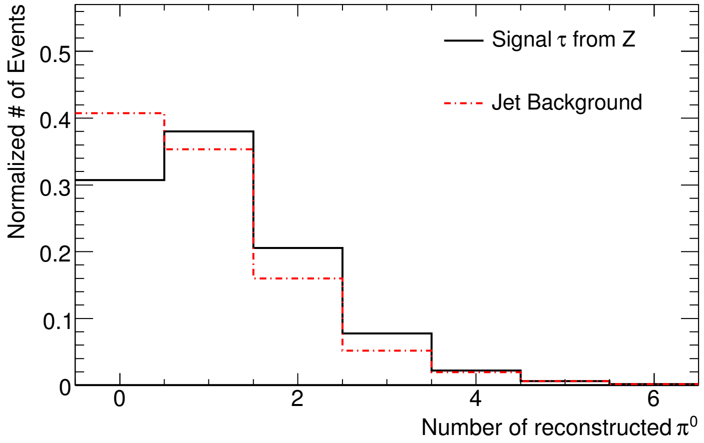


Figure 4.16: The number of π^0 subclusters associated with the tau candidate.

4.4 Current Identification Algorithms

Tau identification focuses on identifying taus and rejecting background objects from the set of reconstructed candidates. This process is not specific to the type of event from which a candidate originates. Therefore, only quantities that express the properties of the tau are used in identification. No event information (such a $\not{E}_T$) is used. Furthermore, the background is so plentiful and too many jets are so similar to taus that any one variable cannot be used to distinguish the two. The general strategy in the identification algorithms is to use a technique to combine several variables in a defined way to create a single, superior variable. This combination strategy is what is meant by a "multivariate" technique.

The baseline identification tool for taus is currently a logarithmic likelihood multivariate technique which is derived from probability density functions (pdfs) of signal and background Monte Carlo (MC) samples [21].

The classifier is defined as

$$d = \frac{\mathscr{L}_S}{\mathscr{L}_B + \mathscr{L}_S}, \tag{4.8}$$

where $\mathscr{L}_{S(B)}$ is the likelihood that the tau candidate is truly signal(background). The likelihoods are found using the pdf for each variable x_k:

$$\mathscr{L}_{S/B} = \prod_{k=1}^{k=\text{nVars}} p_k^{S/B}(x_k), \tag{4.9}$$

where $p_k^{S/B}$ is the probability density function for each variable for signal/background. Note that the likelihood defined in Equation 4.9 clearly does not take any correlations into account. This is one disadvantage to the likelihood multivariate discriminant. The discriminant d peaks sharply at 0 and 1 and a transformed discriminant is therefore used:

$$d^{'} = -\ln\left(\frac{1}{d} - 1\right) = \sum_{k=1}^{k=\text{nVars}} \ln \frac{p_k^S(x_k)}{p_k^B(x_k)}. \tag{4.10}$$

This discriminant is available for all candidates formed by the calorimeter seeded algorithm (both alone and calo+trk seeded).

The optimization and performance of the likelihood has been documented in detail in [8, 21]. Other identification methods, including a cut method and neural networks (track-based algorithm), also exist in the ATLAS tau program. The performance of the logarithmic likelihood method is shown in Chapter 6 in comparison to boosted decision tree performance.

Chapter 5

Boosted Decision Trees

In high energy physics, the general technique for identifying events or objects of interest amongst a high background environment involves applying a set of criteria on eligible candidates. The criteria are chosen to minimize the number of true background candidates (background faking the interesting object) while maximizing the number of true signal candidates (actual object of interest) which pass. These criteria are generally referred to as cuts. However, with such a technique some true signal candidates are bound to fail the series of cuts and some background candidates will pass. This results both in a loss in signal events and in an impure final state.

A decision tree uses this multiple cut technique in a sophisticated way in order to salvage the signal candidates which would otherwise be lost and remove background that would normally pass. It is designed to maximize signal and background separation by recycling events that both pass and fail cuts for further analysis. Decision tree (DT) algorithms do not require any a priori assumptions about input variable correlations. Furthermore, the training time for building boosted decision trees is relatively short, making them easier to study and develop than other multivariate methods such as neural networks. Boosted Decision Trees (BDTs) have already been shown to be effective in high energy physics, such as in the first evidence for single top quark production analysis by the D0 Collaboration [22]. The similarity of the jet background to the tau signal at the LHC is a motivation to study the use of boosted decision trees in combining many weak classifying variables into a stronger discriminant for tau identification in ATLAS.

5.1 Decision Trees

A decision tree is a machine-learning technique which combines several weak classifiers to create a more powerful multivariate discriminant. It is a way of organizing and choosing the cuts applied to a candidate depending on whether it passed or failed the previous cuts.

In general, a decision tree is a structure of cuts organized into nodes. A node is the decision point in the tree in which a variable and cut value are provided and the candidate is determined to either pass or fail it. The pass or failure determines which node the candidate will encounter next. As seen in Figure 5.1, a tree begins at a primary (root) node and branches off to two secondary nodes corresponding to the pass or fail of the root node cut. Each of these nodes carries a cut qualification and a tested candidate would again advance from this node either to the left or right daughter node. Each cut path eventually stops at some terminal node or “leaf” possessing a classifier value which will be assigned to the candidate. Consequently, any event that fails a certain cut will not be thrown away immediately as background, but continue to be analyzed. All events are given a decision tree score between 0 and 1, which is described later in this section.

Decision tree training uses a set of known signal and background training events, each with weight w_i, to build a tree structure of cuts node by node. The function used in this method to quantify the separation between signal and background at any given node is called the Gini index, defined as

$$Gini = 2p(1-p),$$

where the purity p is

$$p = \frac{S}{S+B} = \frac{\sum\limits_s w_s}{\sum\limits_s w_s + \sum\limits_b w_b}$$

and $S(B)$ is the *weighted*[1] total number of signal (background) events which landed on the node during training. Thus, the Gini index for a node is 0 (minimal) when the purity is either 1 or 0 (pure signal or pure background) and maximized when the purity is 0.5 (maximally mixed sample).

[1] Events in a signal or background sample are often given different weights, which correspond to cross-sections. Further weighting is introduced during boosting iterations, as explained in Section 5.2.

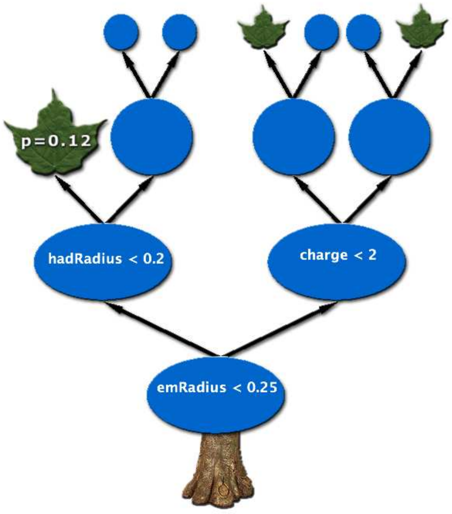


Figure 5.1: Visual representation of a single decision tree. Blue circles represent nodes and green leaves are the terminal leaves. In this case, the root node cuts on emRadius < 0.25. Candidates with a value greater than this travel left and a cut on hadRadius is made. Those candidates with hadRadius ≤ 0.2 (failing the node cut) would be assigned a decision tree value $D_i = 0.12$ according to the leaf purity.

Beginning with an initial root node containing all the events, the variable and corresponding cut which maximizes separation is calculated and executed. The maximum separation is defined as the maximum change in the Gini index between the mother node and the two daughter nodes. That is, for each potential cut, the weighted change in Gini index

$$\Delta Gini = Gini_{Mother} - f_L \cdot Gini_L - f_R \cdot Gini_R$$

is calculated, where $L(R)$ represents the left(right) daughter node and $f_{L(R)}$ is the weighted fraction of events in the daughter node. The cut corresponding to the highest $\Delta Gini$ is chosen. Following this initial branching, the best cuts according to the Gini separation are calculated and executed for both the left and right daughter nodes. The tree then grows as the cut selection process continues at each node. A node is no longer split when a split would result in less than a minimum number of events landing in the node. Such an unsplit node is the “leaf” node described earlier. Each leaf node is assigned a purity value p as defined above.

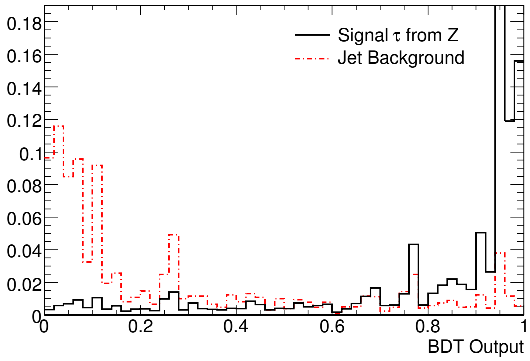


Figure 5.2: Decision tree output for 1 tree (no boosting). Background events peak to the left and signal events peak to the right.

A test event works through the cut conditions beginning with the root node and follows the path along the tree according to pass (right daughter) or fail (left daughter) until it lands on a terminal node. The decision tree result (or classifier value) $D(i)$ for an event tested on a single tree is equal to the purity of the leaf on which the testing terminates. See Figure 5.2 for the distribution of a decision tree purity value for a single tree.

5.2 Boosting

While a single tree on its own improves upon a simple cut-based analysis, boosting significantly increases the performance of this single tree. It also helps smooth distributions which may otherwise appear spiky due to features of a specific training sample when limited statistics are used. In general, the boosting process uses the training results of the first tree to increase the weights of candidates that were misclassified. A new tree is then trained using these weights. Boosting effectively re-weights candidates that the previous tree classified incorrectly in order to increase their importance during the next training. Terminal leaves are labelled either background or signal leaves according to a set purity threshold (often 0.5). Misclassification occurs when a candidate of one type (signal or background) terminates on a leaf of opposite classification.

Using this boosting method, many trees are then trained with new weights calculated after each retraining. In this study, boosting has been applied according to the (discrete)

AdaBoost method [23]. Once a single tree T_m (where m denotes the tree number) is trained, the boosting parameter assigned to this tree is given by

$$\alpha_m = \beta \ln\left(\frac{1-\varepsilon_m}{\varepsilon_m}\right), \tag{5.1}$$

where β is a boosting parameter studied in Section 6.2.1 and ε_m is the misclassification error rate of the m^{th} tree defined as the weighted fraction of misclassified candidates

$$\varepsilon_m = \frac{\sum_i w_{m_i} \times M_m(i)}{\sum_i w_{m_i}}$$

using the binary value $M_m(i) = 1$ for candidates misclassified by T_m and 0 otherwise and i as a sum over all candidates in the training sample, each with weight w_{m_i}. Training candidates for the tree T_{m+1} are then assigned the weight

$$w_{m+1_i} = w_{m_i} \times e^{\alpha_m \times M_m(i)}.$$

Consequently, the weights of correctly classified candidates are not changed. Because $\varepsilon_m > 0.5$ and thus α is always positive, this increases the weight of misclassified candidates, giving them higher priority in the calculation of tree T_{m+1}.

During the testing phase, the final decision tree result assigned to candidate i is

$$D(i) = \sum_{m=1}^{M_{tree}} \alpha_m D_m(i).$$

Because using an average of many trees also reduces statistical fluctuations that may creep in due to a limited training sample size, boosting makes decision trees more stable.

5.3 Tree Parameters

As is clear from the description above, several user defined parameters affect training and boosting of decision trees. These include:

- **Minimum leaf size**: threshold number of candidates in a node, below which node splitting is forbidden, converting the node into a leaf

- **NBoosts**: number of boosting cycles
- **AdaBoost parameter β**: controls strength of candidate re-weighting as in Equation (5.1)
- **BoostingPurityLimit**: purity threshold to classify a leaf as signal or background

The effect of these parameters on tau ID is shown in Section 6.2.1.

Chapter 6

Tau Identification Using Boosted Decision Trees

This chapter defines the complete list of variables used for tau identification. It then explores several BDT parameters which can be varied to increase performance. Final results are shown for the tau identification and background rejection. Some extra studies which look towards early data within ATLAS are also presented.

6.0.1 Samples Used

This study uses Monte Carlo simulated events for the training and assessment of BDT performance. The events are all reconstructed from version 14.2.10 of the ATLAS Athena software. The taus come from a $Z \rightarrow \tau\tau$ sample as signal and jets from QCD dijet events of various E_T ranges as background. The MC QCD dijet events used are in the E_T ranges from below the tau reconstruction threshold to above 560 GeV to simulate the background and are weighted by cross-section. The weighted E_T is shown in Figure 6.1.

The signal taus used include calo and calo+trk seeded candidates and have been matched to true tau leptons generated by Monte Carlo in the event. Likewise, the jet background consists of reconstructed calo and calo+trk seeded tau candidates which have been matched to jets reconstructed at truth level using a cone algorithm with $\Delta R < 0.4$ (Cone4TruthJets). The matching requirement for both signal and background is that the candidate lies within a cone of $\Delta R < 0.2$ from the corresponding truth object.

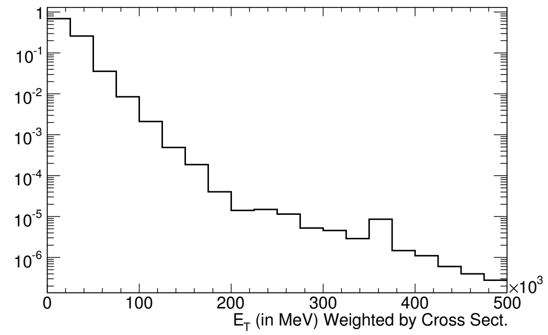


Figure 6.1: Reconstructed background E_T weighted by cross section for calo+trk candidates.

6.1 Discriminating Variables

This study develops BDTs separately for tau candidates seeded by both the calorimeter and track algorithm (calo+trk seeded candidates) and for those seeded only by the calorimeter seeded algorithm (calo only candidates). The BDT developed for the calo only candidates uses all the variables listed below from the calorimeter seeded algorithm. The BDT for the calo+trk seeded candidates uses all the track seeded algorithm variables listed in addition to the calorimeter algorithm variables. This capitalizes on the more extensive list of discriminating variables calculated for the candidates seeded by both algorithms.

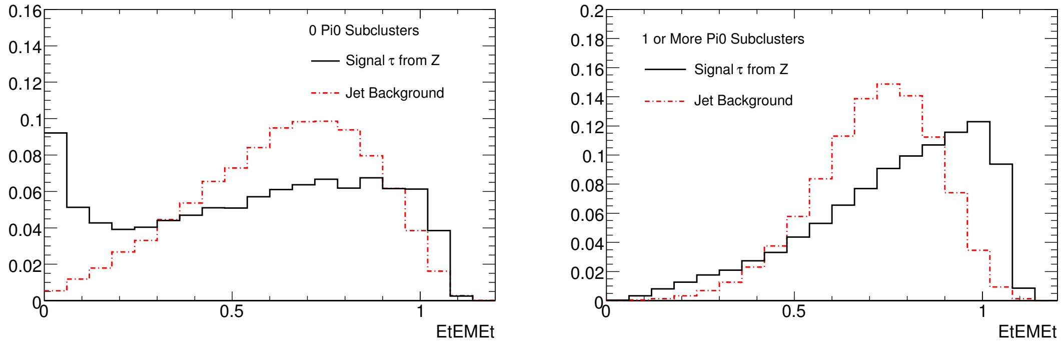
The variables below defined by the calorimeter seeded reconstruction algorithm and discussed in Section 4.3.1 are used in this analysis. They are referred to in this analysis by the shortened names in bold:

- **centralityFraction**: the centrality fraction.
- **emRadius**: the electromagnetic radius R_{em}.
- **hadRadius**: the hadronic radius R_{had}.
- **isolationFraction**: isolation in the calorimeter.
- **stripWidth2**: transverse energy width in the η strip layer.
- **charge**: charge of the tau candidate.

- **numStripCells**: number of hits in the η strip layer.
- **ipSigLeadTrack**: lifetime signed pseudo impact parameter significance of leading track.
- **nTracksdrdR**: number of tracks in small ring.
- **trFlightPathSig**: transverse flight path significance $\boldsymbol{L_{xy}/\sigma_{Lxy}}$.
- **etOverPtLeadTrack**: $\boldsymbol{E_T}$ over $\boldsymbol{p_T}$ of the leading track.

Additional variables defined in the following way are calculated manually using information from the calorimeter seeded algorithm:

- **EtEMEt**: the fraction of electromagnetic transverse energy, $\frac{E_T(EM)}{E_T}$. A significant portion of energy deposited by hadronically decaying taus with 1 or more reconstructed π^0 subclusters is electromagnetic. In this case, the tau often deposits a greater fraction of its energy in the electromagnetic calorimeter than background jets do. Note that the electromagnetic energy $E_T(EM)$ is calibrated using a MC weighting and that the total energy E_T includes an extra, global calibration factor on top of this. The result is that the fraction $\frac{E_T(EM)}{E_T}$ may sometimes be greater than 1 due to the calibration differences. This distribution is shown in Figure 6.2.
- **dRmin**: the smallest separation between associated tracks in a cone with $\Delta R < 0.2$ for candidates with more than one reconstructed track. Hadronic tau decays are more collimated than jets and therefore the tracks are generally closer together. This results in a smaller separation between tracks. The distribution is shown in Figure 6.3.
- **dRmax**: the largest separation between associated tracks in a cone with $\Delta R < 0.2$ for candidates with more than 1 reconstructed track. As with dRMin, this value is expected to be smaller for hadronic taus than for background jets. The distribution is shown in Figure 6.4.
- **etEMSumPTtracks**: the electromagnetic transverse energy divided by the total transverse energy from tracking, $\frac{E_T(EM)}{\sum p_{T_{Track}}}$. Many hadronic tau decays with no π^0 daughters leave little to no energy in the electromagnetic calorimeters. Furthermore, as discussed in Section 4.3, a larger portion of the total transverse tau energy is carried



(a) Trk+calo candidates with 0 reconstructed π^0 subclusters

(b) Trk+calo candidates with 1 or more reconstructed π^0 subclusters

Figure 6.2: Fraction of electromagnetic transverse energy $\frac{E_T(EM)}{E_T}$.

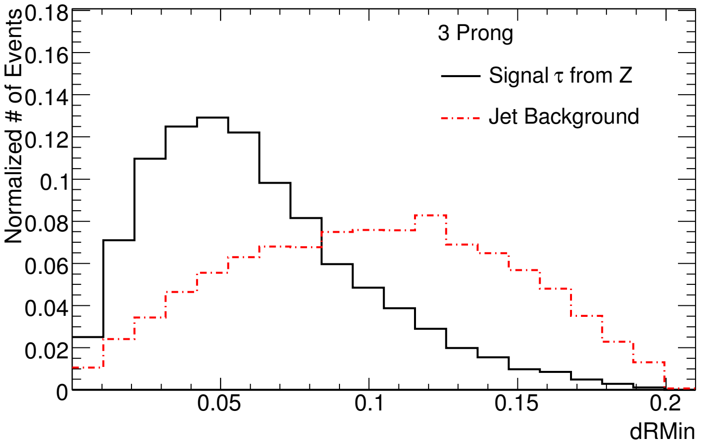


Figure 6.3: The smallest separation between associated tracks in cone $\Delta R < 0.2$ for 3 prong calo+trk candidates.

by charged candidates and registered in the total p_T of the tracks. This results in a larger denominator and smaller value on average in $\frac{E_T(EM)}{\sum p_{T_{Track}}}$ for taus. This is shown for candidates both with and without any reconstructed π^0 subclusters in Figure 6.5.

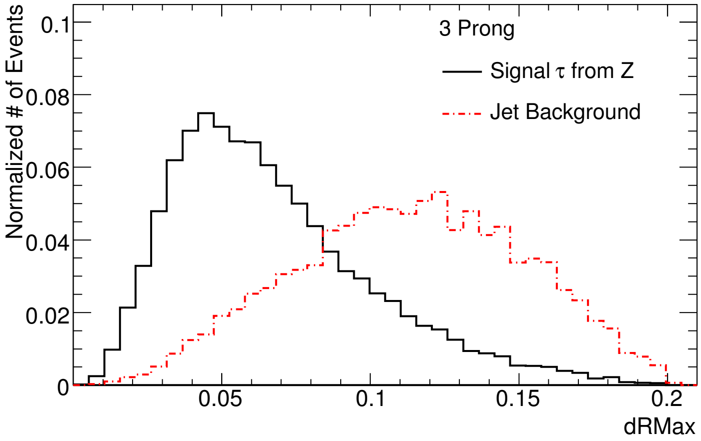
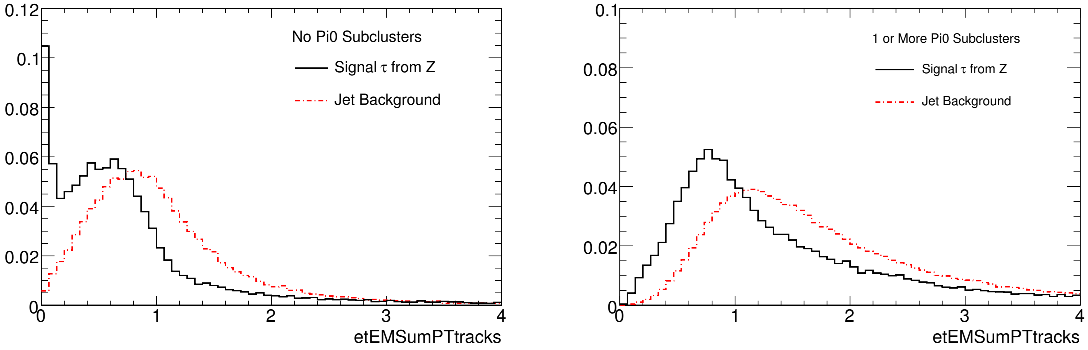


Figure 6.4: The largest separation between associated tracks in cone $\Delta R < 0.2$ for 3 prong calo+trk candidates.



(a) Calo+trk candidates with no reconstructed π^0 subclusters.

(b) Calo+trk candidates with 1 or more reconstructed π^0 subclusters.

Figure 6.5: Distributions of etEMSumPTtracks.

- **etHadSumPTtracks**: the hadronic transverse energy divided by the total transverse energy from tracking, $\frac{E_T(Had)}{\sum p_{T_{Track}}}$. As with etEMSumPTtracks, the denominator is expected to be larger for hadronic taus. However, all hadronic taus have a hadronic decay component and this variable has less discriminating power than etEMSumPTtracks, which can take advantage of those without an electromagnetic component.

The distribution of etHadSumPTtracks is shown in Figure 6.6.

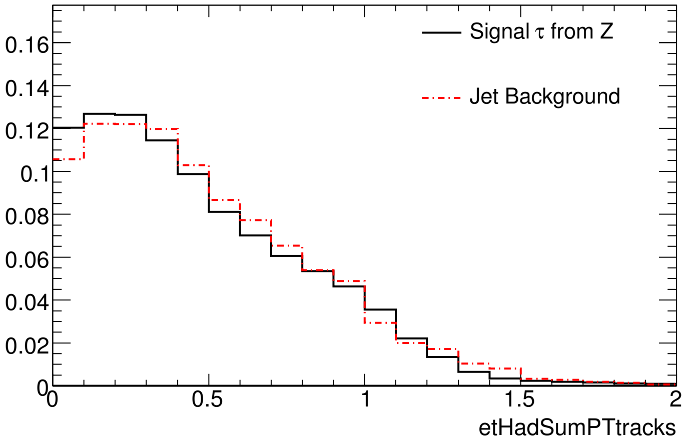
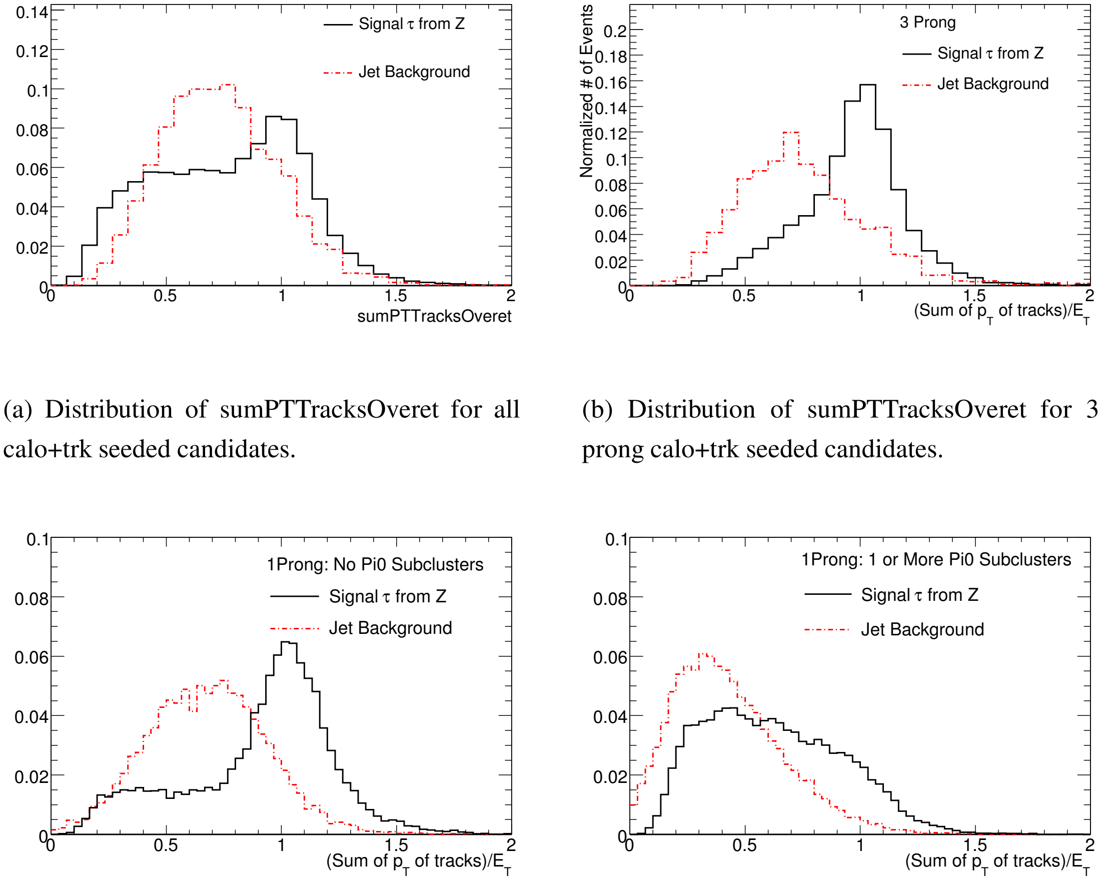


Figure 6.6: Distribution of etHadSumPTtracks.

- **sumPTTracksOveret**: the ratio of p_T calculated from tracking to total transverse energy found by the calorimeter, $\frac{\sum p_{T_{Track}}}{E_T}$. Charged decay products carry a larger fraction of the energy of a hadronic tau decay than of a jet. Therefore the fraction of the E_T which is also carried by total track transverse momentum is higher for taus than jets, as shown in Figure 6.7.

  Figure 6.7(a) shows that the total signal distribution appears to have two maxima centered near 0.5 and 1. This is due to several hadronic decay modes of the tau. As seen in Table 3.1, roughly $\frac{2}{3}$ of all three prong decay modes do not contain a π^0. In this case, the charged daughters should carry all of the visible transverse energy of the tau decay. For the $\frac{1}{3}$ of 3 prong decays that do contain a π^0 daughter, it is one of four visible decay products and one would expect that the charged daughters to carry $\frac{3}{4}$ of the visible energy, on average. Figure 6.7(b) shows that for 3 prong taus, the distribution of sumPTTracksOveret peaks near 1 and has a low tail due to the decays with a π^0 daughter.

  For 1 prong tau decays with no π^0 daughter, the charged daughter should carry most of the visible transverse energy of the tau decay, as shown in Figure 6.7(c). In the case of 1 prong tau decays with a π^0 daughter, one would expect the charged daughters to carry roughly half of the visible transverse energy of the tau decay, which is consistent with Figure 6.7(d).



(a) Distribution of sumPTTracksOveret for all calo+trk seeded candidates.

(b) Distribution of sumPTTracksOveret for 3 prong calo+trk seeded candidates.

(c) Distribution of sumPTTracksOveret for those 1 prong calo+trk seeded candidates which have no reconstructed π^0 subcluster.

(d) Distribution of sumPTTracksOveret for 1 prong calo+trk seeded candidates which have 1 or more reconstructed π^0 subclusters.

Figure 6.7: Distribution of sumPTTracksOveret.

The following additional variables calculated by the track seeded algorithm and discussed in Section 4.3.1 are used in this analysis. They are referred to in this analysis by the shortened names in bold:

- **rWidth2Trk3P**: the width of tracks momenta.
- **massTrk3P**: the invariant mass of the tracks system.
- **trFlightPathSig**: the significance of transverse flight path.
- **numPi0**: the number of π^0 Subclusters.

One additional variable is calculated manually using the following information from the track seeded algorithm:

- **EteflowOverEt**: the ratio of E_T calculated by the energy flow algorithm and E_T from the calorimeter. As discussed in Section 3.2.1, the energy flow algorithm calculates the transverse tau energy well, but underestimates the transverse jet energy. This ratio is therefore higher for taus than for jets. The distribution of EteflowOverEt is shown in Figure 6.8.

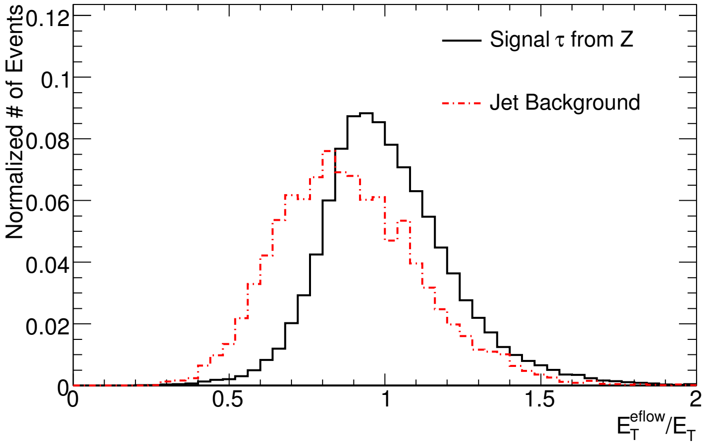


Figure 6.8: Distribution of EteflowOverEt for calo+trk candidates.

6.1.1 Variable Correlations

As many of the above variables describe the same physical quantities using varied definitions, it is expected that some discriminating variables will be correlated. While it is not necessary to know the variable correlations when choosing discriminating variables, it can be valuable to understand correlations in order to simplify trees. For example, one may reduce the variable list by removing one of a set of highly correlated variables without a significant loss in performance. This improves training time and reduces the number of variables one has to understand when dealing with systematic uncertainties. This method was used when defining a list of safe variables described in Section 6.4.1. Correlation scatter plots are shown in Figure 6.9 and the linear correlation matrix calculated with TMVA [24] is in Figure 6.10.

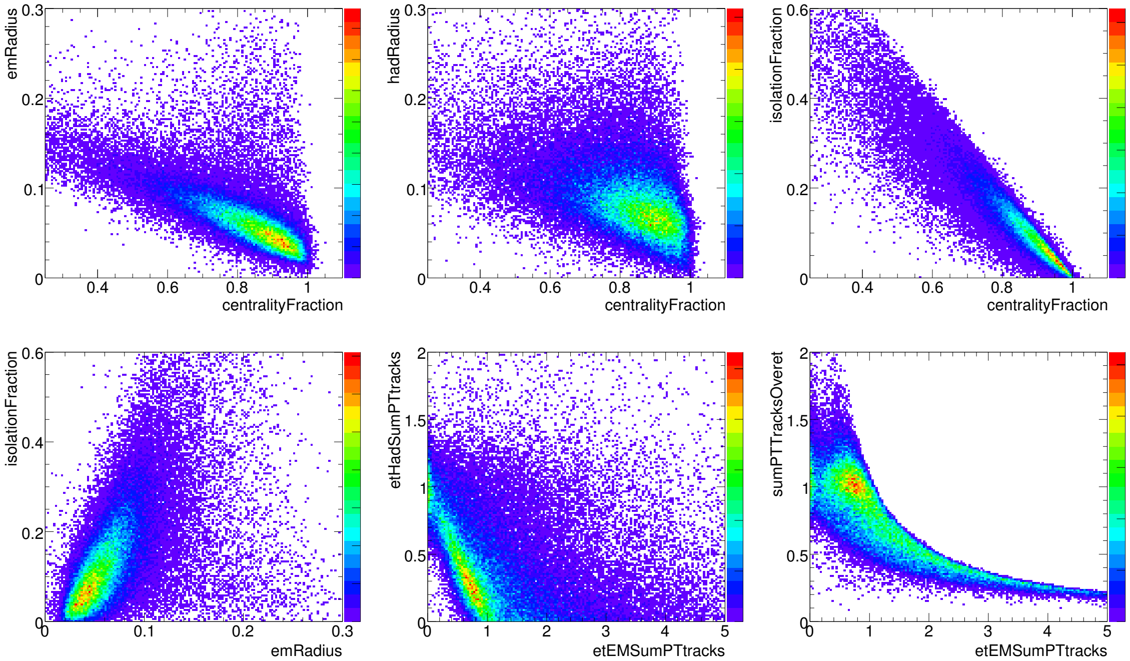
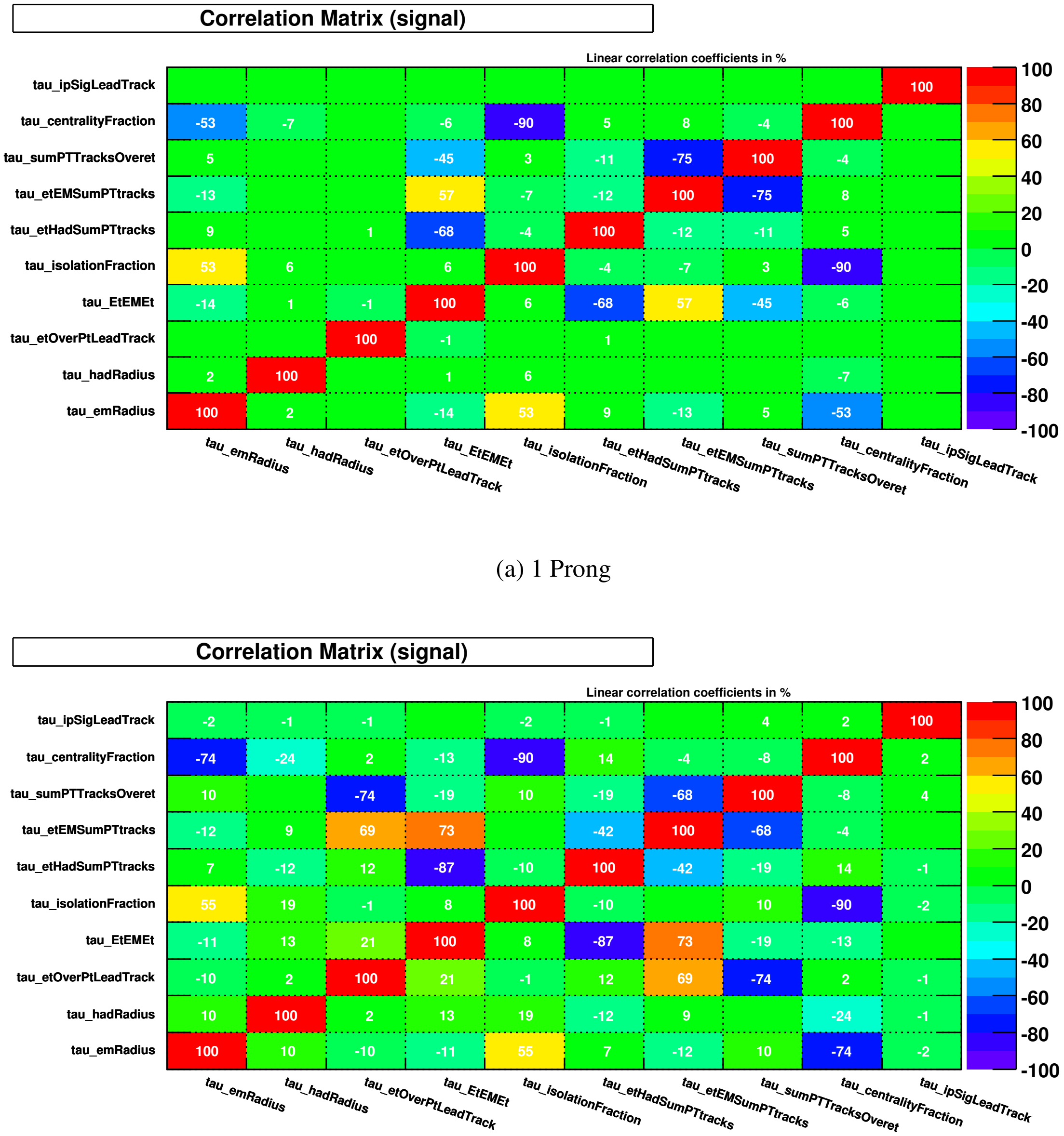


Figure 6.9: Variable correlation plots for signal samples of all calo+trk seeded candidates.



(a) 1 Prong

(b) 3 Prong

Figure 6.10: Linear correlations of a subset of the discriminating variables for truth matched reconstructed calo+trk seeded taus from $Z \rightarrow \tau\tau$ (by prong type), calculated with TMVA.

6.2 Developing BDTs for Tau ID

Boosted Decision Trees for tau identification are built using the algorithm described in Chapter 5 and the variables listed in Section 6.1. Training samples of pure signal and pure background objects are used to build the trees. Once a BDT is built, it is applied to samples of pure signal and pure background which are independent of the training samples and used for BDT evaluation. Each testing object is evaluated and assigned a BDT output score. An example of the BDT output for signal and background can be seen in Figure 6.11.

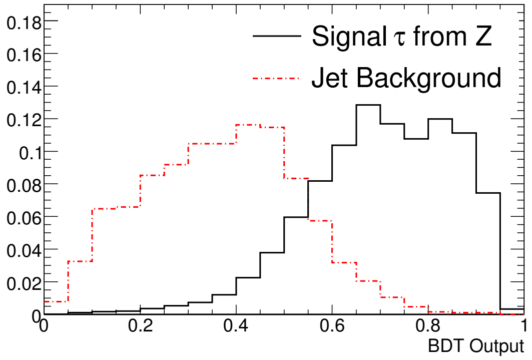


Figure 6.11: Example of BDT output assigned to an independent sample of signal and background objects. Background objects have a lower score on average than true taus do. A cut on this score can be used to distinguish taus from background.

The criterion for evaluating the performance is the relative background rejection rate for given signal efficiencies. The signal and background efficiencies are defined as:

$$\varepsilon^{Signal} = \frac{\text{\# of matched reco taus passing cut which have } n \text{ tracks}}{\text{\# of MC taus which have } n \text{ charged daughters}}$$

$$\varepsilon^{Jet} = \frac{\text{\# of matched tau candidates passing cut which have } n \text{ tracks}}{\text{\# of MC Jets}}$$

where n may be 1 or 3 and is the number of tracks. In the case of MC, n is the number of the true charged stable daughters in MC. In the case of reconstruction, n is the number of reconstructed tracks. Note that the signal sample requires that the number of reconstructed tracks match the number of MC charged stable daughters. For the purpose of these performance plots, the cut to be passed is either on a BDT or likelihood classifier value. The kinematic binning is by the true visible transverse energy (neutrino momentum is excluded) in the case of the signal sample and is done by reconstructed transverse energy in the case of background rejection.

The rejection is then

$$Rejection = \frac{1}{\varepsilon} - 1$$

To quantify performance as a single number, the ratio of the areas below and above the background rejection vs signal efficiency curve was calculated. Because of the unstable behaviour of the curves at low efficiency and the unfixed y-axis limits, the bounds $\varepsilon > 0.1$ and Rejection $< 10^4$ were used in the area calculations. The area below the curve increases with higher performance resulting in an increased ratio. Figure 6.12 illustrates these areas.

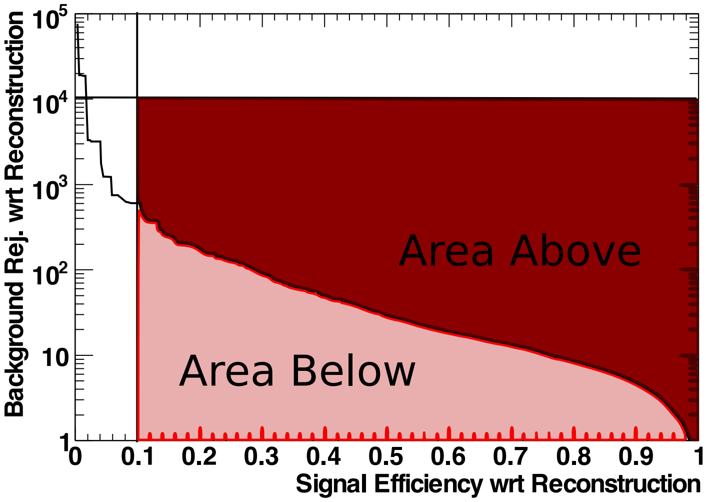


Figure 6.12: The background rejection vs signal efficiency curve is shown. Performance can be evaluated by this curve. To summarize the performance as a single number, the ratio of the areas below and above the curve within the bounding box, as shown, is also used as an evaluation criterion.

Tuning parameters and object definitions may improve the performance of boosted decision trees. Overall optimization is achieved both through physics optimizations and algorithm optimizations.

For physics optimizations, improvements in signal efficiency and background rejection might be made by specializing on different categories of taus. While decision trees are calculated to optimize signal and background separation node by node, they do not predict how a cut at a certain node will affect the possible separation of one or more nodes later. Therefore, if a certain initial cut naturally separates candidates into streams of similar tau types but does not produce the maximum Gini separation, it will not be chosen. One must implement such initial separations by hand. In this case, a specialized boosted decision tree is calculated specifically for candidates that all share further selected qualities in common. Generally referred to as "binning", BDTs may be calculated individually for taus which fall into a common energy range and hadronic decay type (for example: 1 prong, 3 prong, and those including π^0 daughters) to compare performances.

The tuning of the BDT parameters in 5.2 is discussed below.

6.2.1 Tuning Decision Tree Parameters

There are very few parameters that can be tuned while training BDTs. To measure performance, BDTs were trained for varying parameter values to study the effects of the parameter. Unless otherwise stated, the values of the parameters which are not being varied were held at MinLeafSize=100, nBoosts=10, and AdaBoost $\beta = 0.2$.

The MinLeafSize controls the minimum number of training events that must land on a leaf. If less than the minimum number would have landed in the leaf if the chosen cut had been made, the splitting stops. The choice of MinLeafSize is a balance between keeping the parameter high enough so that the leaves are statistically significant but low enough that a suitable number of cuts are made before the splitting stops. Furthermore, it is possible for the MinLeafSize performance to correlate to the size of the training sample. For this reason, the MinLeafSize choice was studied using both the full unbinned training sample available and the 1 prong 10–30 GeV set as an example of a smaller training set. The results can be seen in Table 6.1 and Figures 6.13–6.14.

One sees that the MinLeafSize does not affect performance significantly within this

Table 6.1: Relative background rejection rates as a function of the MinLeafSize and signal efficiencies. Numbers are with respect to reconstruction, not MC truth, and are therefore useful only for comparing relative parameter performance.

MinLeafSize	30% eff	40% eff	50% eff	60% eff	70% eff	90% eff
20	200	110	60	40	30	10
50	220	110	70	40	20	10
80	200	110	60	40	20	10
100	210	110	60	40	20	10
120	200	110	60	40	20	10
150	220	110	60	40	20	10
200	200	100	60	40	20	10

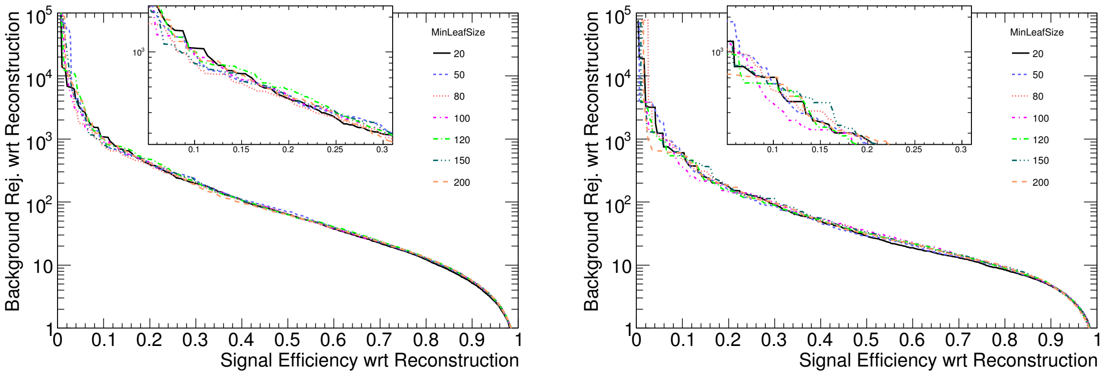
Unbinned sample

MinLeafSize	30% eff	40% eff	50% eff	60% eff	70% eff	90% eff
20	90	50	30	20	10	0
50	90	50	30	20	10	0
80	90	50	30	20	10	0
100	90	60	40	20	10	0
120	90	50	30	20	10	0
150	110	60	30	20	10	0
200	100	50	30	20	10	0

1 prong 10–30 GeV candidates

range. The MinLeafSize affects training time, as a smaller leaf size requires more nodes to be made and results in larger and deeper trees. This is an argument to keep the MinLeafSize relatively high. For further studies it was therefore decided to study the unbinned sample with a MinLeafSize of 100 and that any smaller training sample (due to E_T and/or prong binning) would use a MinLeafSize of 150.

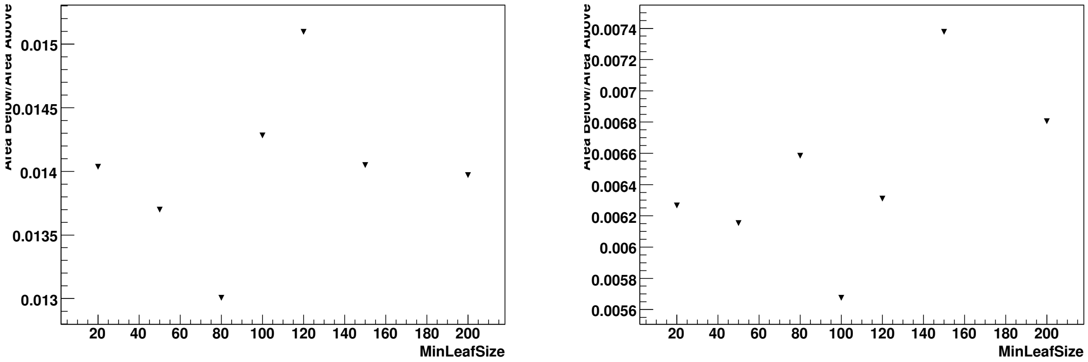
The performance as a function of number of boosting cycles can be seen in Table 6.2 and Figures 6.15–6.16. Figure 6.15 shows that the performance improves with more boosts in



Trained and tested on unbinned sample.

Trained and tested on 1 prong 10–30 GeV candidates.

Figure 6.13: Results for various minimum leaf size settings for trees with 10 boosting cycles.



Trained and tested on unbinned sample.

Trained and tested on 1 prong 10–30 GeV candidates.

Figure 6.14: Ratio of area below and above the curve of background rejection vs signal efficiency wrt. MinLeafSize in Figure 6.13.

the large sample. The 10–30GeV bin sample, however, did not gain significant performance when increasing the boosting to more than 10 times. Based on this study, 50 boosting cycles were chosen for the unbinned sample. The performance as a function of the AdaBoost β parameter training with 50 boosting cycles is shown in Figure 6.17. With this increase in boosts, a β value of 0.2 is sufficient.

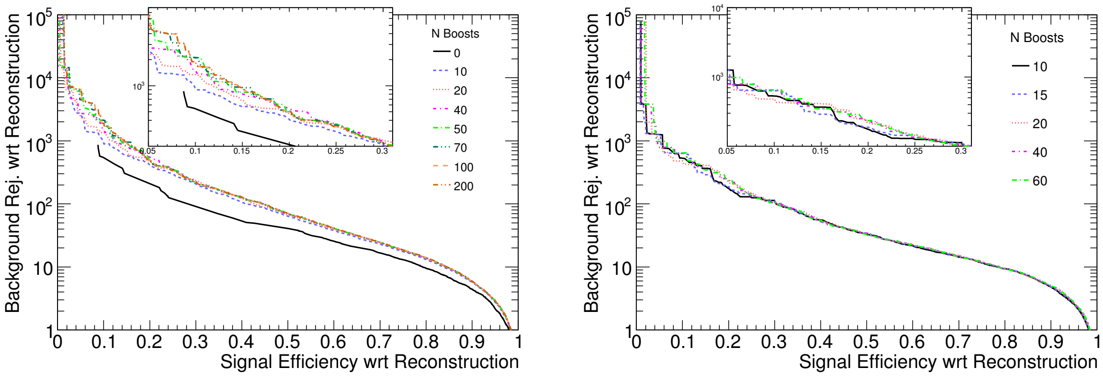
Table 6.2: Relative background rejection rates as a function of number of boosts and signal efficiencies for calo+trk candidates. Numbers are with respect to reconstruction, not MC truth, and are therefore useful only for comparing relative parameter performance.

nBoosts	30% eff	40% eff	50% eff	60% eff	70% eff	90% eff
0	70	51	41	25	17	4
5	170	100	60	40	20	10
10	210	110	60	40	20	10
20	240	130	70	40	20	10
40	230	130	70	40	20	10
50	220	130	70	40	20	10
70	220	130	70	40	20	10
100	220	130	70	40	20	10
200	220	130	70	40	20	10

Unbinned sample

nBoosts	30% eff	40% eff	50% eff	60% eff	70% eff	90% eff
10	110	60	30	20	10	0
15	110	50	30	20	20	0
20	100	60	30	20	20	0
40	100	50	30	20	10	10
60	100	50	30	20	10	10

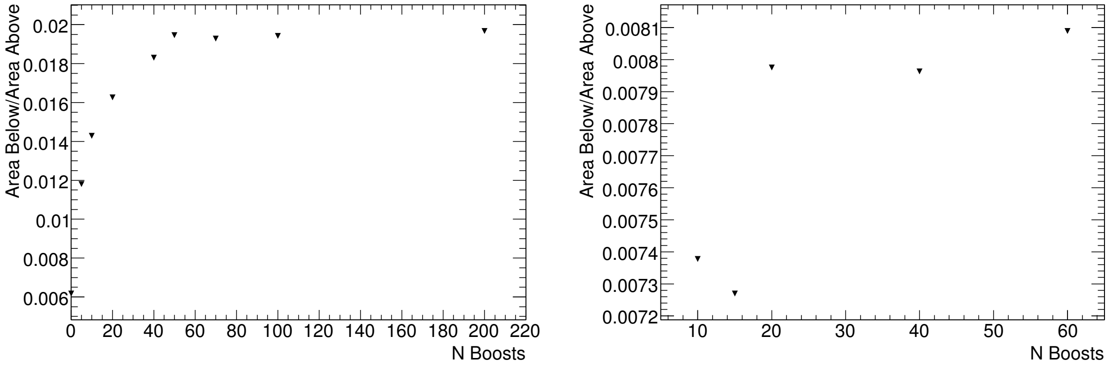
1 prong 10–30 GeV candidates



Trained and tested on unbinned sample.

Trained and tested on 1 prong 10–30 GeV candidates.

Figure 6.15: Performance for various number of boosts. Numbers are with respect to reconstruction, not MC truth, and are therefore useful only for comparing relative parameter performance.



Trained and tested on unbinned sample.

Trained and tested on 1 prong 10–30 GeV candidates.

Figure 6.16: Ratio of area below and above the curve of background rejection vs signal efficiency wrt. the number of boosting cycles in Figure 6.15.

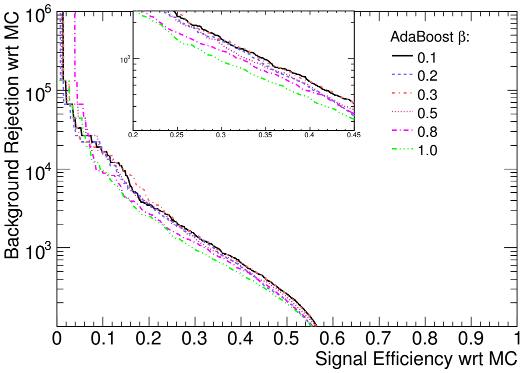


Figure 6.17: Performance for various AdaBoost β values for calo+trk candidates. Trees trained with 50 boosts. Numbers are with respect to MC truth.

6.2.2 Binning by E_T

Taus of different energy ranges show slightly different characteristics. In general, they tend to have narrower showers with increasing energy, which affects the distribution of most discriminating variables. Jets also become narrower with higher energy, but less so than taus do. The dependence on E_T can be seen in Figures 6.18 and 6.19, in which several discriminating variables are plotted in different E_T ranges.

Figure 6.20 shows that the unbinned training performs better than one binned by this E_T range (10–30, 30–60, 60–80, 80–80+ GeV).

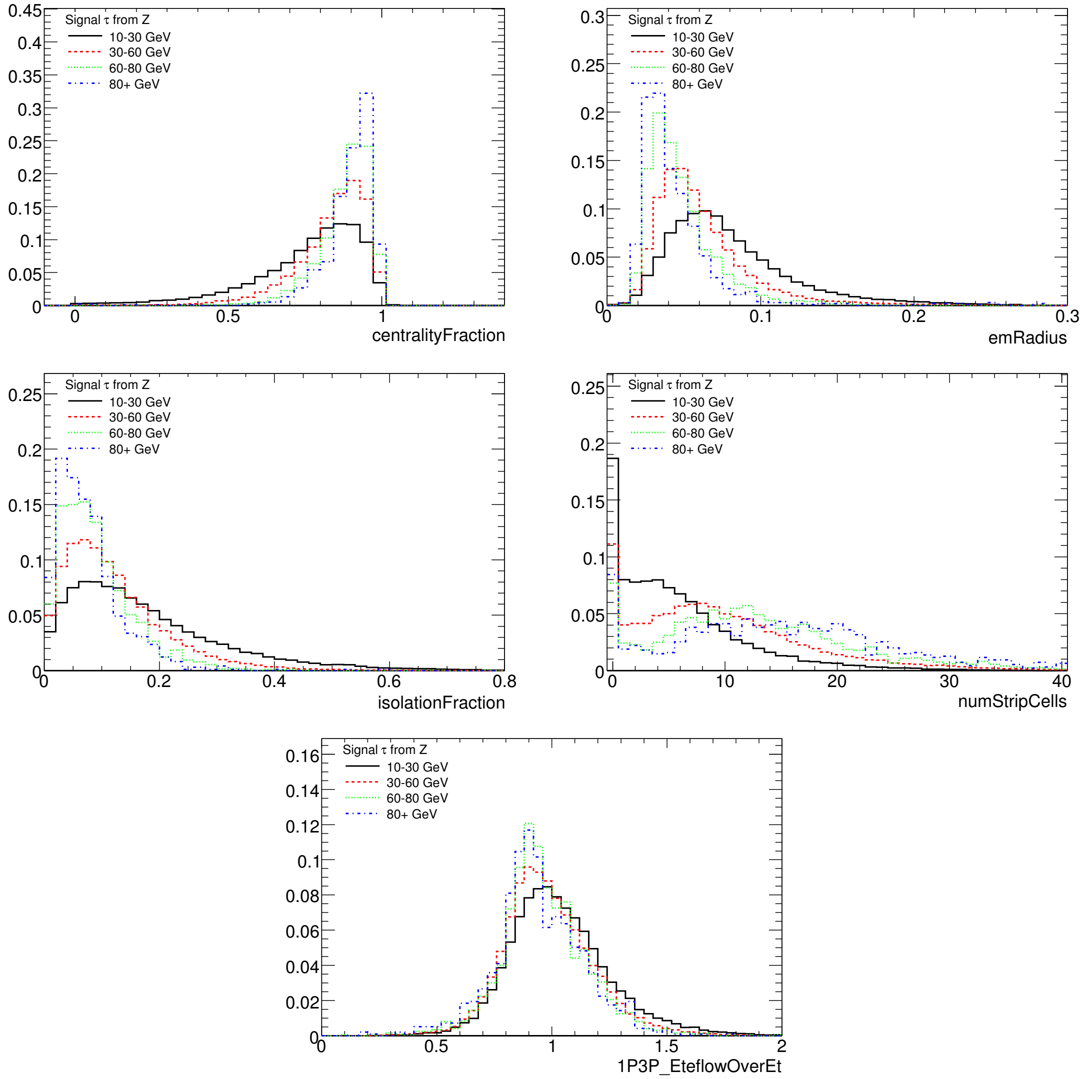


Figure 6.18: Variable distributions for MC taus reconstructed as calo+trk seeded candidates by E_T bin.

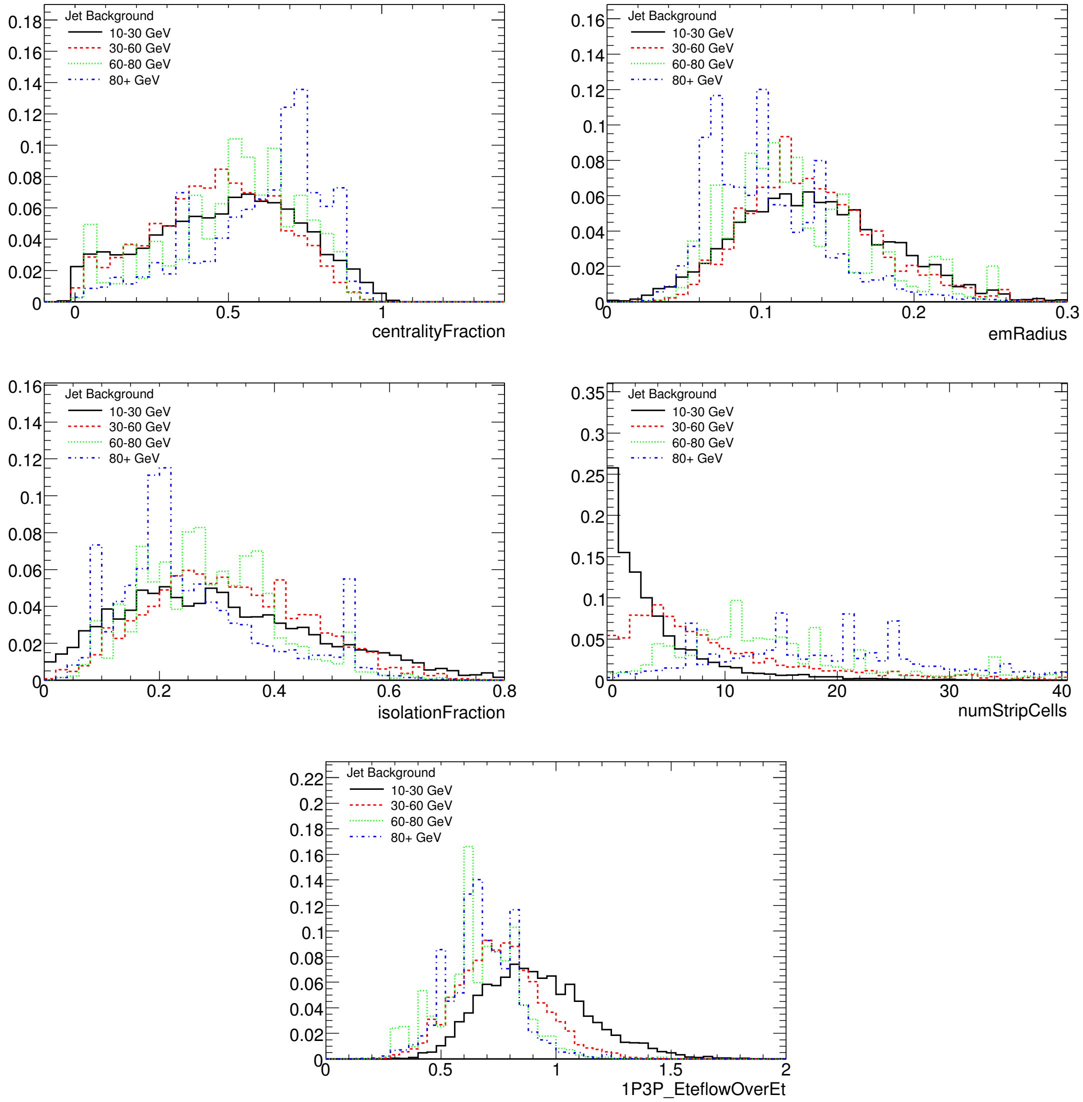


Figure 6.19: Variable distributions for QCD dijets reconstructed as calo+trk seeded tau candidates by E_T bin.

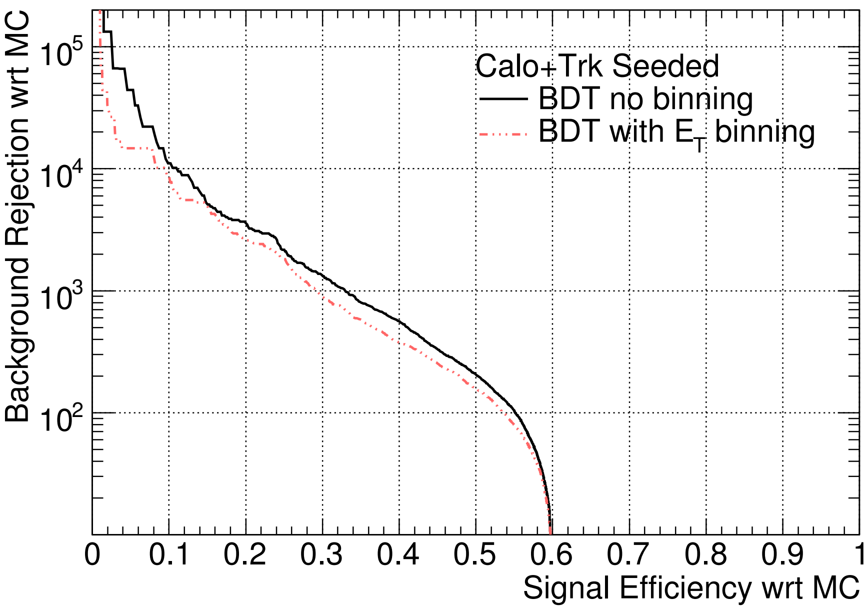


Figure 6.20: BDTs with 20 boosts each. The unbinned trees are trained on all E_T and prong ranges and the binned set contains 4 BDT's for the E_T ranges 10,30,60,80,80+ GeV with no prong binning.

6.2.3 Binning by Track Multiplicity

The distribution of the number of charged tracks can be seen in Figure 6.21. By definition, 3 prong taus have a minimum of 3 visible decay products which are most often $\pi^{\pm}$. They therefore tend to be wider, their track momentum carries a greater fraction of the measured calorimeter energy, and they show an increased proportion of hadronic energy compared to 1 prong taus. The number of charged tracks therefore affects the shape of discriminating variables. This can be seen in Figures 6.22-6.23. Note that the truth matching for the signal further requires that the reconstructed candidate have the same number of charged tracks as the true tau has stable charged daughters. The 3-prong bin is trained and tested (for performance) on true 3-prong taus. In practice, it will be applied to any candidate with > 1 charged track. Figure 6.24 shows, however, that in this case an unbinned tree performs better than one binned by prong.

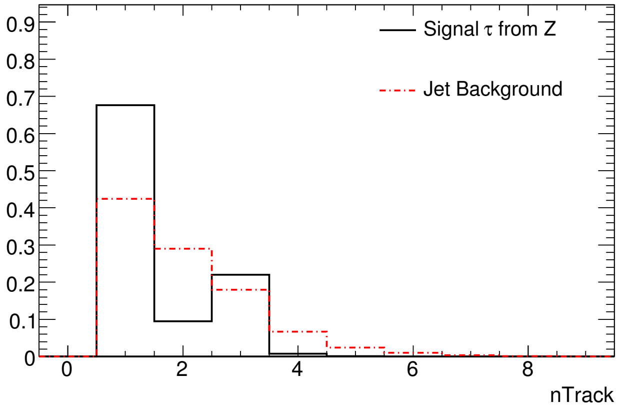


Figure 6.21: Distribution of charged tracks for calo+trk seeded candidates, as calculated by the calo seeded algorithm.

Figure 6.25 summarizes the binning performance for calo seeded candidates. As with the calo+trk candidates, the unbinned tree performs the best.

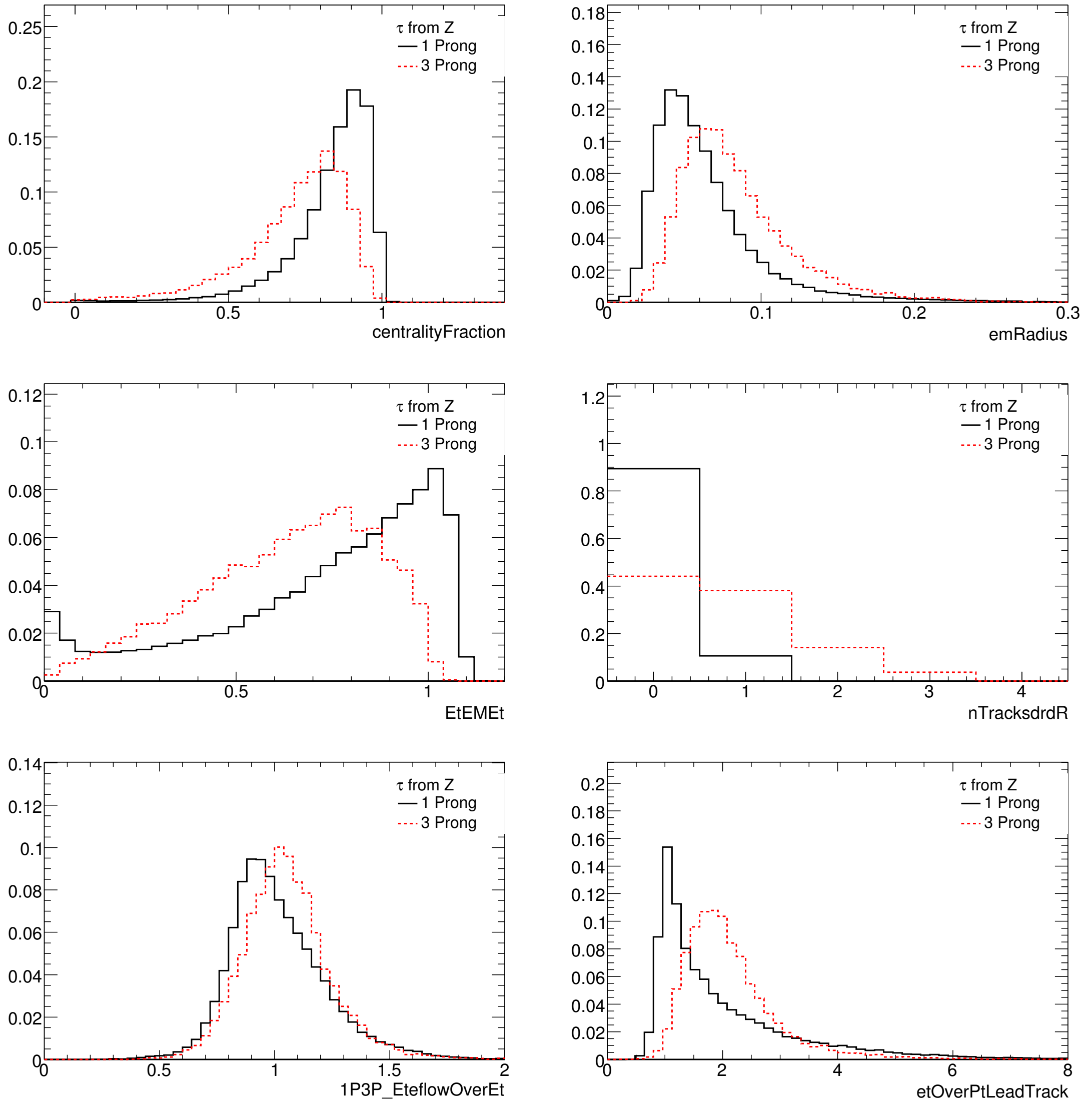


Figure 6.22: Sample of distributions by prong bin for calo+trk seeded signal taus.

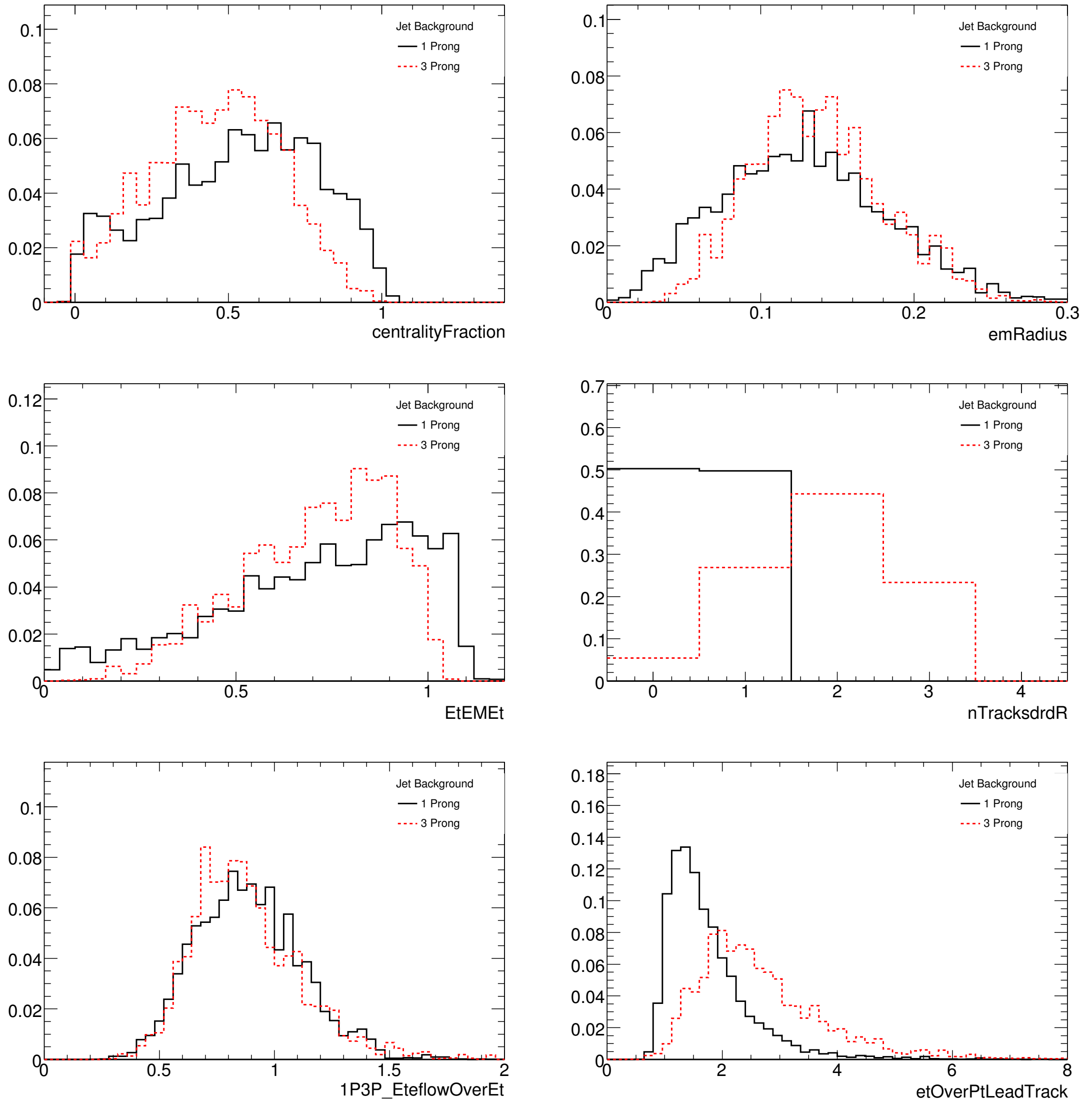


Figure 6.23: Sample of distributions by prong bin for calo+trk seeded background jets.

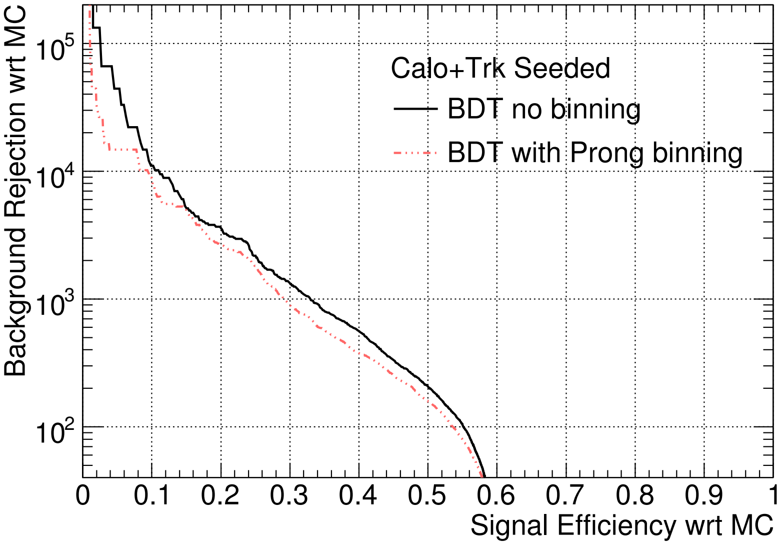


Figure 6.24: BDT with 20 boosts each for calo+trk seeded candidates. The unbinned trees are trained on all E_T and prong ranges. For trees with binning, two BDTs are trained, one each for 1-prong and 3-prong candidates.

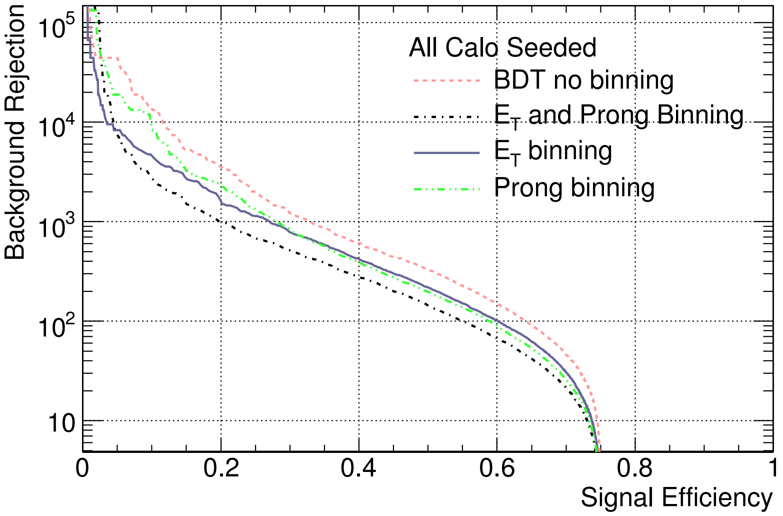
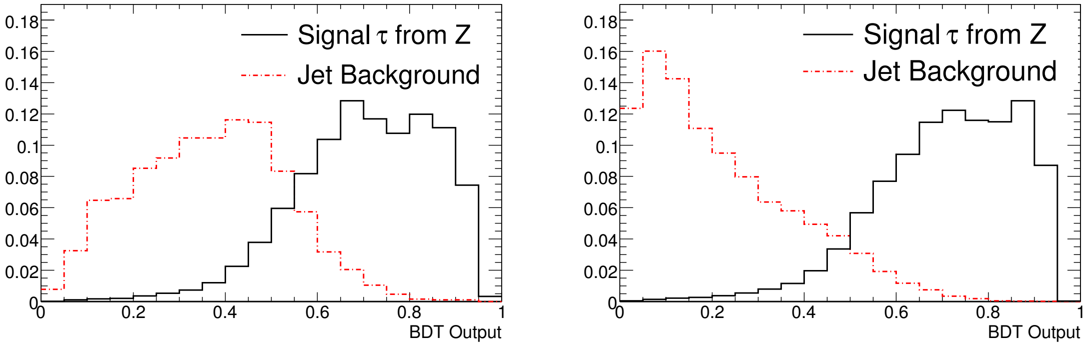


Figure 6.25: BDTs trained and tested on calo seeded candidates. All trainings use 20 boosts, $\beta = 0.2$, and MinLeafSize=100. E_T bins are 10, 30, 60, 80, 80+GeV and prong bins are 1 and 3 prong.

6.3 Performance of BDTs

Following the studies in the previous section, the classifier output for both seed types is shown in Figure 6.26.

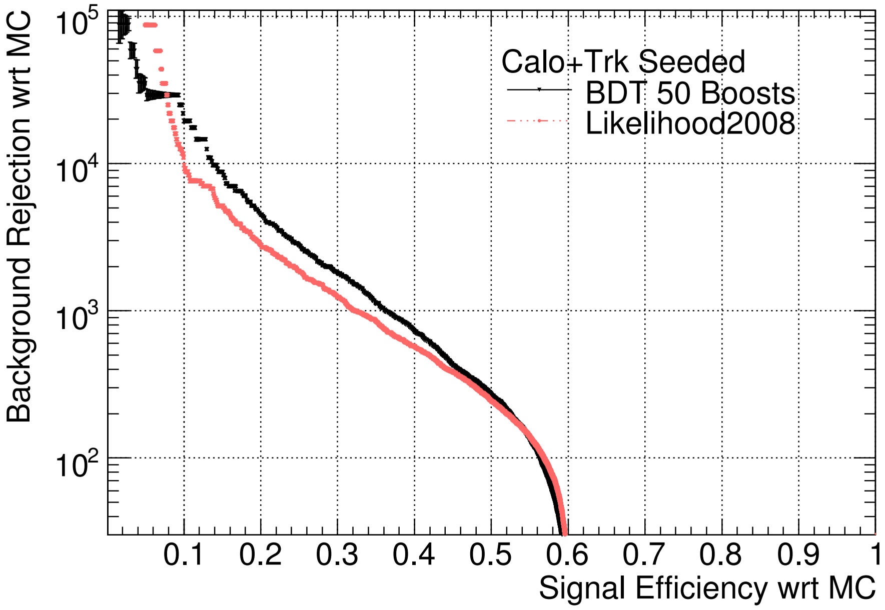


Calo+trk seeded candidates

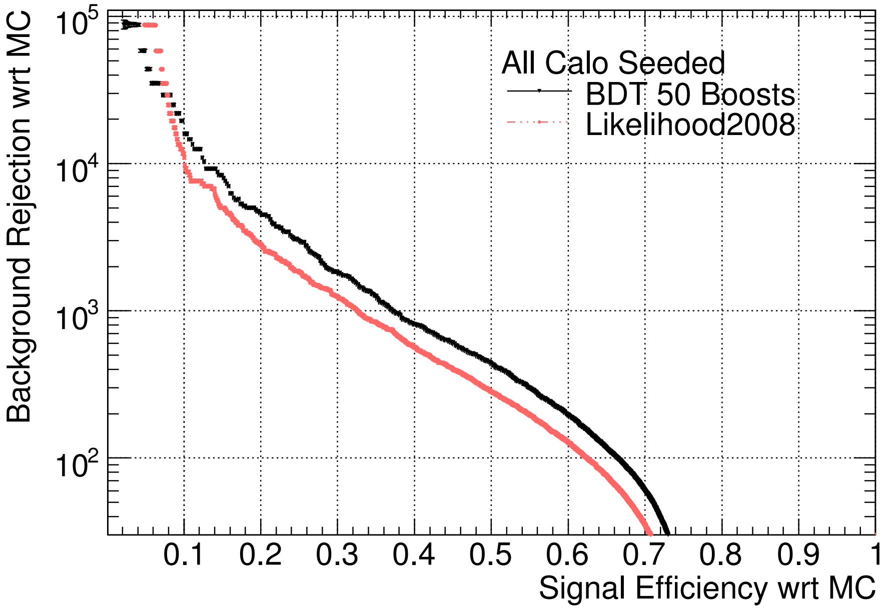
All calo seeded candidates

Figure 6.26: BDT output for calo+trk (a) and all calo (b) candidates using the unbinned BDTs with 50 boosts.

Figures 6.27 and 6.28 compare the performance of BDTs for the calo+trk and calo seeded candidates to the standard likelihood2008 results. Figure 6.27 also includes error bars for statistical uncertainty (details in Appendix A). It is evident that with the number of events used for testing, the statistical errors are small compared to the performance of the BDT. The final parameters used were MinLeafSize=100 and AdaBoost $\beta = 0.2$, with 50 boosting cycles and no binning by either E_T or track multiplicity. These studies show that BDTs need not be binned by E_T or track multiplicity to be as good as or better than the standard likelihood (which is binned) in most regions.

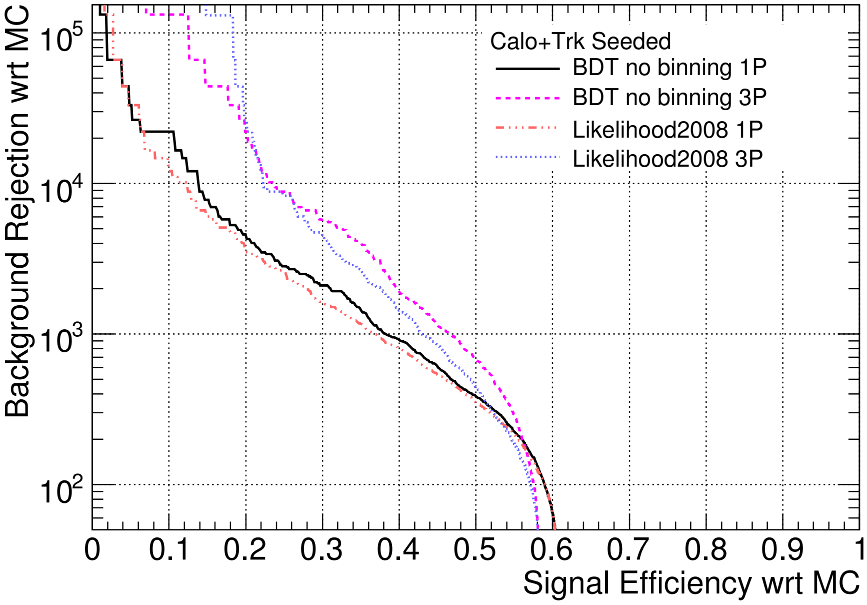


(a)

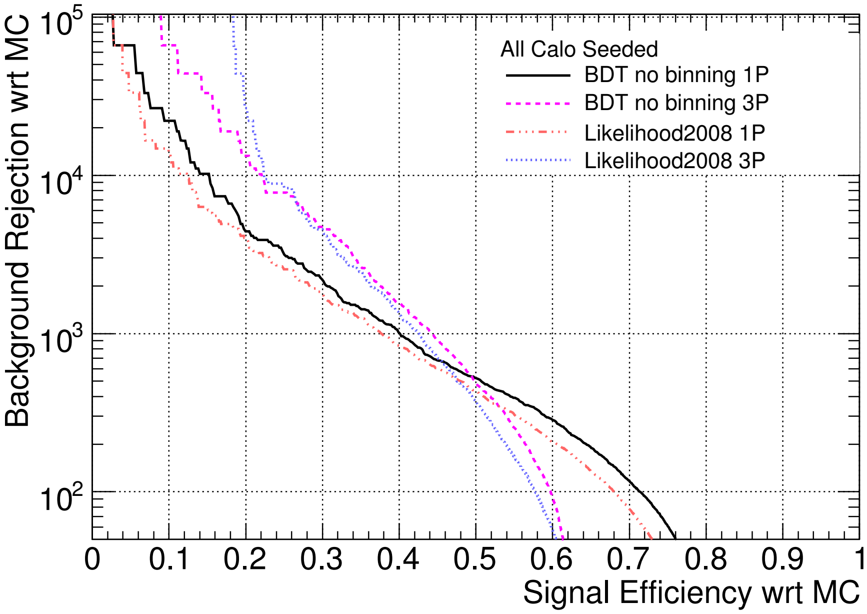


(b)

Figure 6.27: BDT performance on calo+trk (a) and all calo (b) candidates. BDT training uses 50 boosts, $\beta = 0.2$, and MinLeafsize=100. Not binned by E_T or prong. Error bars according to statistical uncertainty are included, but are too small to see.



(a)



(b)

Figure 6.28: BDT performance on calo+trk (a) and all calo (b) candidates by prong. BDT training uses 50 boosts, $\beta = 0.2$, and MinLeafsize=100. Training not binned by E_T or prong.

The background rejection for specific signal efficiencies is given in Table 6.3.

Table 6.3: Signal efficiencies with corresponding background rejection and BDT cut for final BDT results shown in this section.

Signal Eff.	30%	40%	50%	60%	70%
	1 Prong				
Bkg Rejection	2100	910	390	70	30
BDT Cut	0.74	0.68	0.61	0.39	0
	3 Prong				
Bkg Rejection	5800	1900	690	50	50
BDT Cut	0.70	0.64	0.53	0	0

Calo+trk seeded candidates

Signal Eff.	30%	40%	50%	60%	70%
	1 Prong				
Bkg Rejection	2200	1000	520	290	120
BDT Cut	0.81	0.76	0.71	0.66	0.58
	3 Prong				
Bkg Rejection	4600	1600	480	90	10
BDT Cut	0.70	0.64	0.56	0.40	0

All calo seeded candidates

6.4 Systematic Effects

A strategy for evaluating the errors in multivariate techniques has been demonstrated by the D0 experiment in the handling of neural network (NN) systematic errors [25]. In the case of D0, in which the experiment has been running for many years, the main factor in determining the accuracy of the training model is statistics. Ensembles of individual variables fluctuated by statistical uncertainty are generated to assess acceptance error on NN cuts. The algorithm is reassessed for each variable ensemble so that for each assessment only one variable is fluctuated. The systematic error for a given neural network cut is the quadratic sum of the RMS of events passing this cut for each variable ensemble.

This type of assessment is appropriate for an analysis that has been properly calibrated and in which the Monte Carlo variable distributions model those from data very well. However, for an experiment such as ATLAS, which is just starting up, statistical errors will not be the dominant error. Instead, choosing well modeled variables and estimating the effects due to miscalibrations or other problems related to understanding the detector will be of more importance.

6.4.1 Safe Variables for Early Running

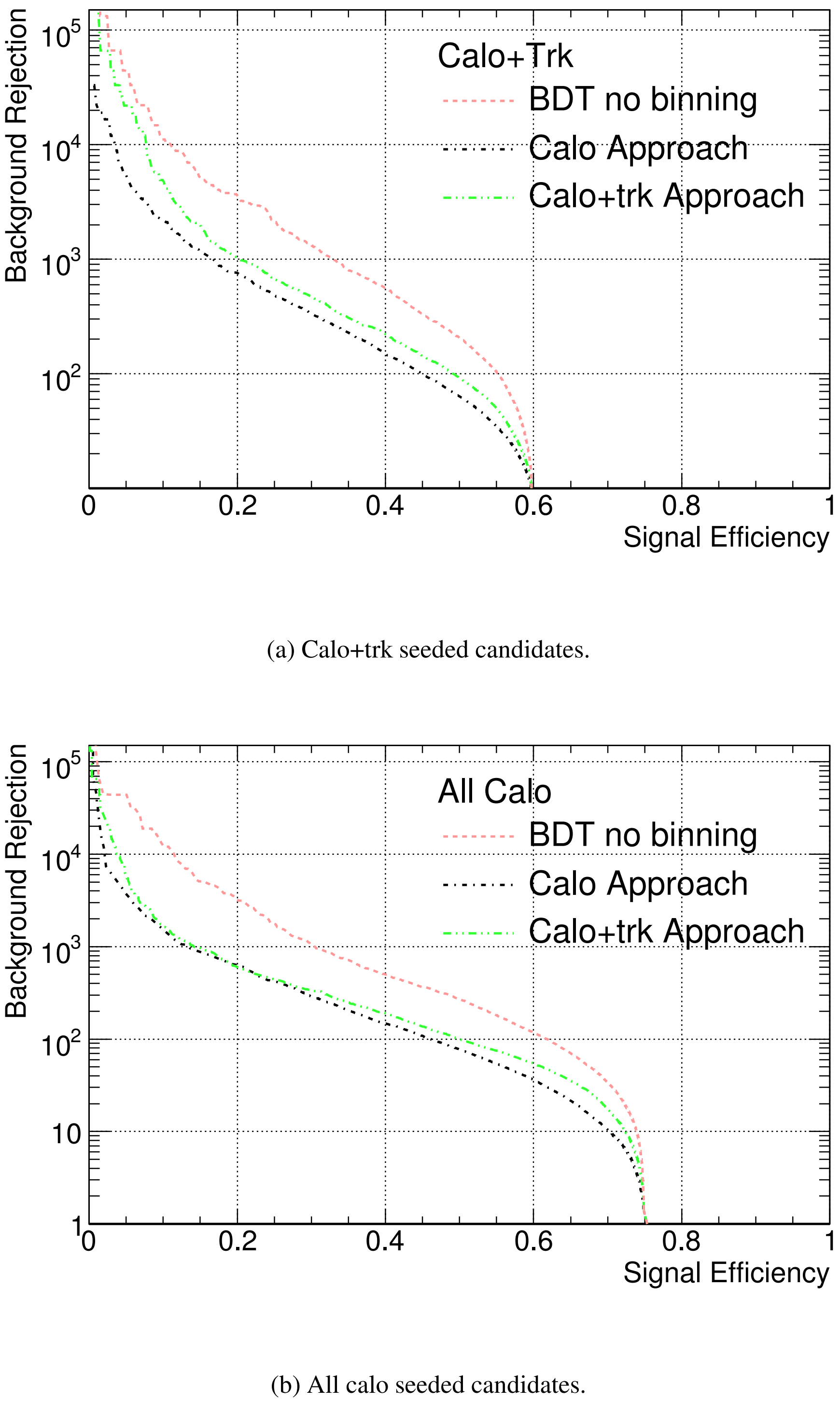
With collisions in ATLAS beginning very soon, one must think practically about how ATLAS subdetectors will perform in early running and with what accuracy discriminating variables will be modeled. Two lists of variables which are expected to be well modeled in early data have been defined by the tau group in ATLAS. This includes an approach which uses information from the calorimeter as well as a more aggressive approach which includes additional tracking variables. Variables which use information from one subdetector should be favoured over those that use several subdetectors or which rely heavily on hadronic calibration.

An example of the reasoning involved in these decisions is as follows. The centrality fraction relies on energy measurements from both the electromagnetic and hadronic calorimeters. It is also highly correlated to the variable emRadius, which depends only on the electromagnetic calorimeter. Therefore, emRadius was chosen and the centrality fraction was excluded from the safe variable list.

The two approaches include the following variables, which were previously defined in 6.1.

Calorimeter Approach	**Additional Track Variables**
emRadius	$\frac{E_T(Had)}{\sum p_{T_{Track}}}$
isolationFraction	$\frac{E_T(EM)}{\sum p_{T_{Track}}}$
strip width (stripWidth2)	W^{Track} (rWidth2Trk3P)
$\frac{E_T(EM)}{E_T}$ (EtEMEt)	etOverPtLeadTrack

The performance of BDTs using this reduced list is shown in Figure 6.29. In both the calo+trk and all calo seeded categories, performance is degraded when moving to the safe variable list. The calo+trk seeded BDTs improve performance when moving to the calo+track variables approach, but the BDTs for the calo candidates do not. The calo+trk candidates have better quality tracks than the calo only candidates, so it is expected that tracking variables would be more significant in this case. For a signal efficiency of 30%, the calo+trk BDTs decrease in background rejection from 1227 to 335 (reduced to 27% of the original rejection) when moving to the calo plus track approach. Likewise, the calo only candidates drop from a background rejection of 1312 to 476 (reduced to 26% of the original rejection) when moving to the calo plus track safe variable approach.



(a) Calo+trk seeded candidates.

(b) All calo seeded candidates.

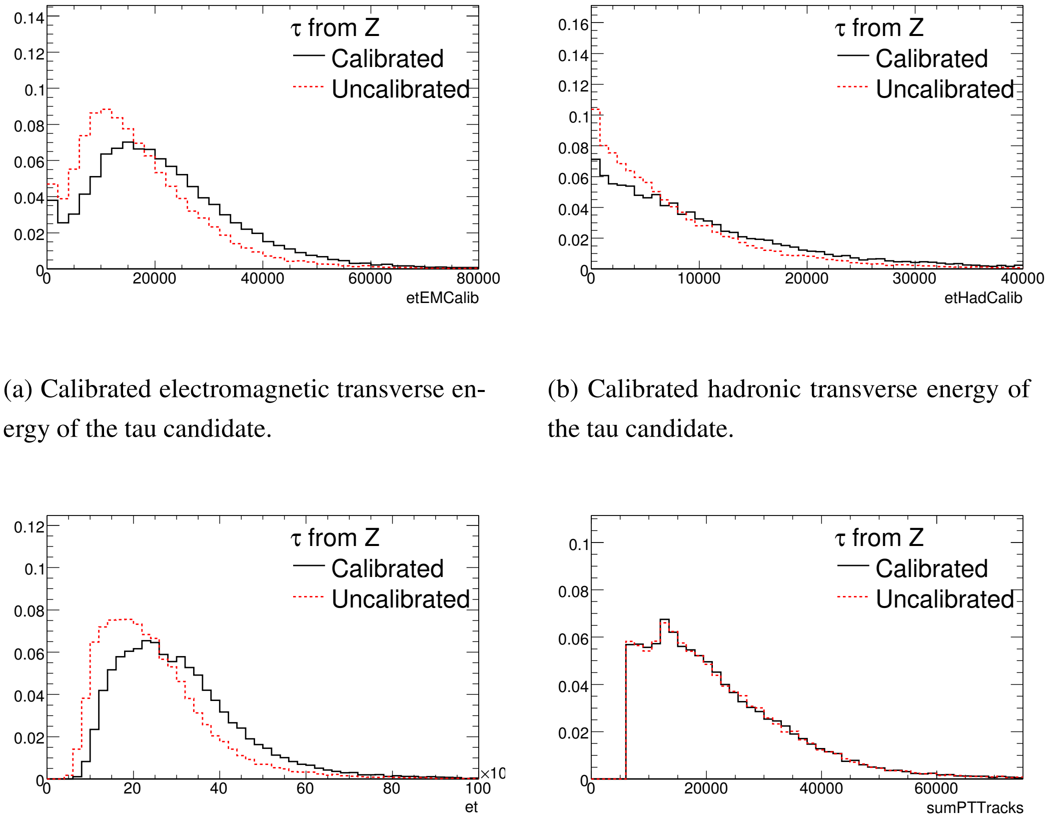
Figure 6.29: BDT performance comparison for safe variable lists. BDTs trained on the full variable list (pink), calorimeter approach (black), and calo+trk approach. All sets use 20 boosts.

6.4.2 Calibration Studies

To estimate the performance of BDTs in an environment in which calibration may not be reliable, MC samples were reconstructed with all energy calibrations at the EM scale. The reason for this is twofold. First of all, it provides an opportunity to study the change in discriminating variables and multi-variate performance in an environment in which the energy scale is incorrect. Secondly, it is expected that the EM scale will be calibrated faster than the hadronic scale in early running. This calibration shift is then a model for one possible early running situation. In this study, the "calibrated" sample refers to candidates calibrated in the default manner (as described in Section 3.2), while "uncalibrated" refers to the samples in which all energy calibration is at the EM scale. All of the following studies were performed using the calo+trk seeded candidates.

The variables etEMCalib and etHadCalib are normally calibrated using MC based cell-weights (H1-style calibration). Using only an EM calibration decreases the average of the distributions for both signal and background, as shown in Figures 6.30(a), 6.30(b), 6.31(a), and 6.31(b). The total E_T of the tau uses this same calibration, with an extra factor applied to scale the energy appropriately for tau decays (see Section 3.2.2). The overall E_T of the candidate also decreases for both signal and background when using the EM scale calibration, as seen in Figures 6.30(c) and 6.31(c). Figures 6.30(d) and 6.31(d) show no significant change to the sum of the transverse momentum calculated from the tracks, as expected.

Variables which are ratios of energy measurements (as described in Section 6.1) are shifted according to the change in energy detailed above. These changes are shown in Figure 6.32 for signal and 6.33 for background. Variables that describe the width of the candidate shower change very little, as shown in Figures 6.34 and 6.35.



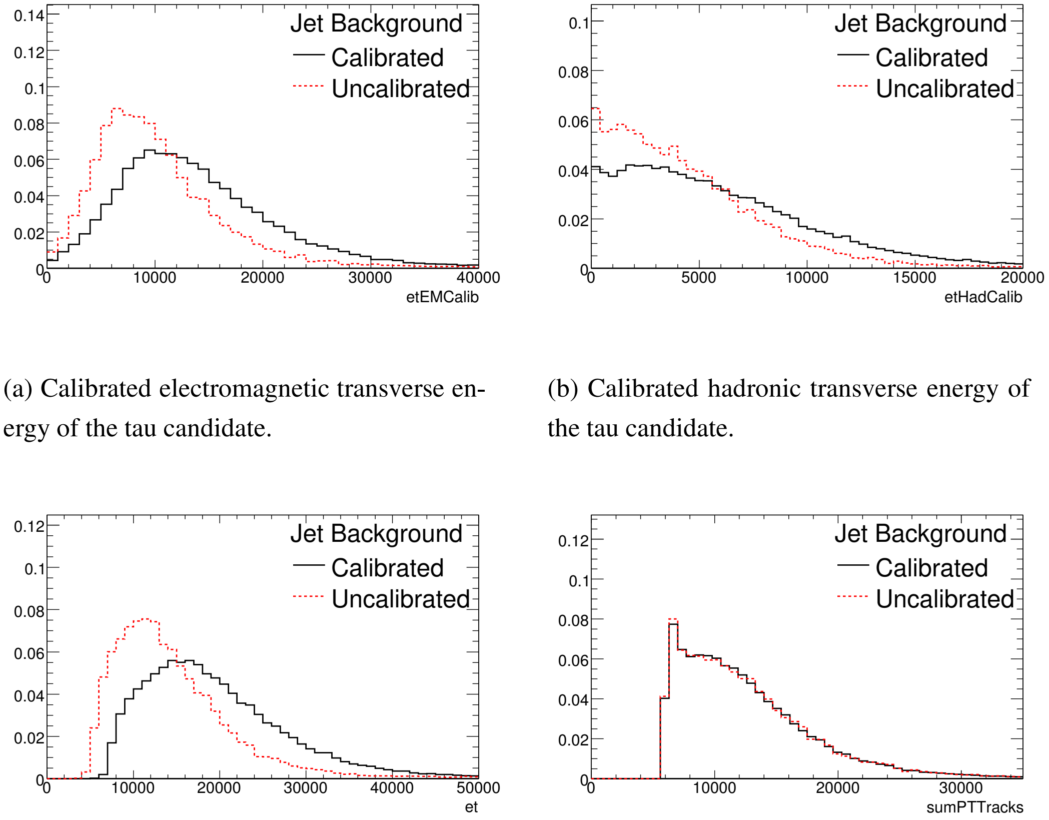
(a) Calibrated electromagnetic transverse energy of the tau candidate.

(b) Calibrated hadronic transverse energy of the tau candidate.

(c) Total transverse energy of the tau candidate.

(d) Sum of the transverse energy of the tracks associated with the tau candidate.

Figure 6.30: Variable distributions for signal taus comparing calibration schemes.



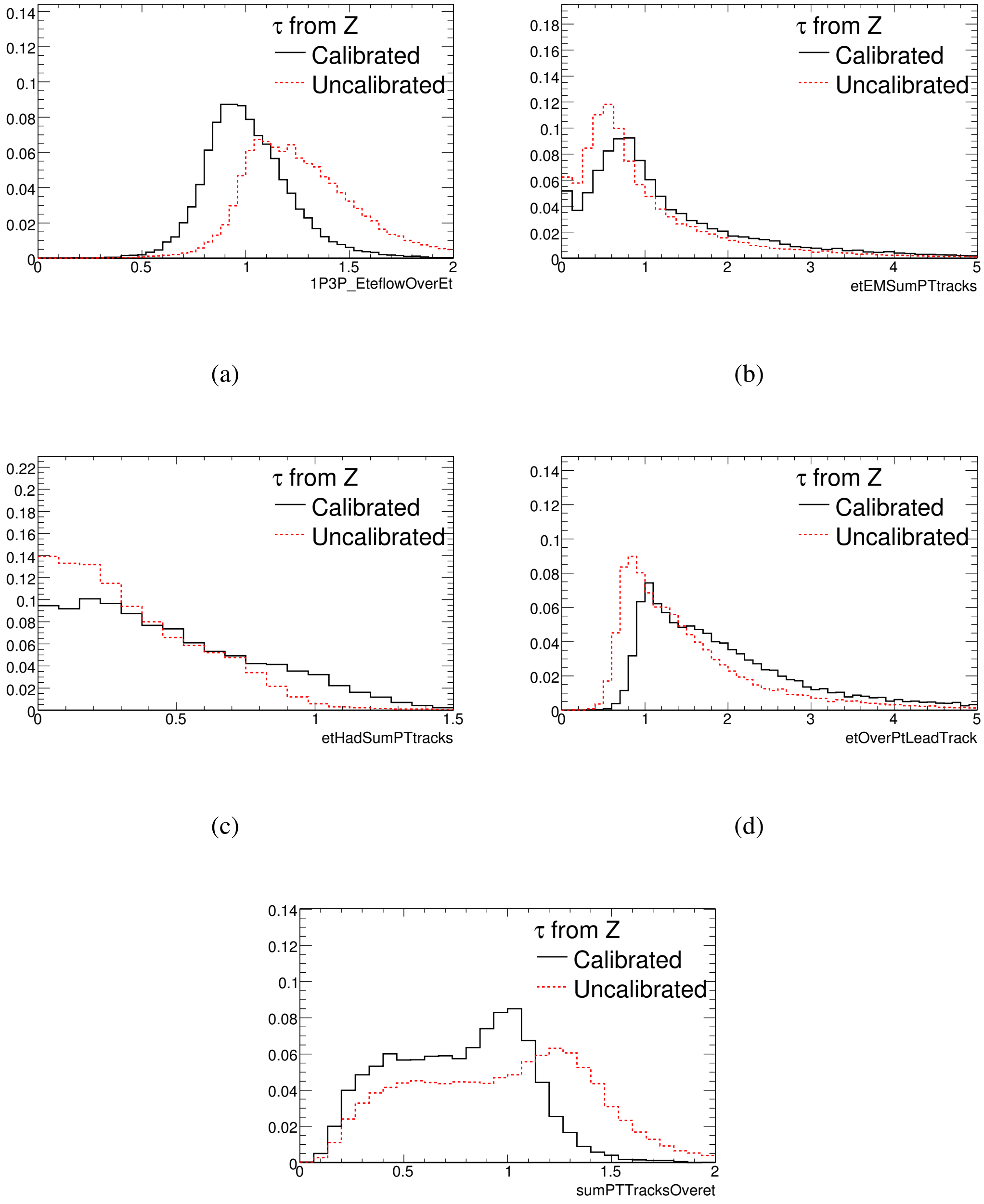
(a) Calibrated electromagnetic transverse energy of the tau candidate.

(b) Calibrated hadronic transverse energy of the tau candidate.

(c) Total transverse energy of the tau candidate.

(d) Sum of the transverse energy of the tracks associated with the tau candidate.

Figure 6.31: Variable distributions for background tau candidates comparing calibration schemes.



(a)

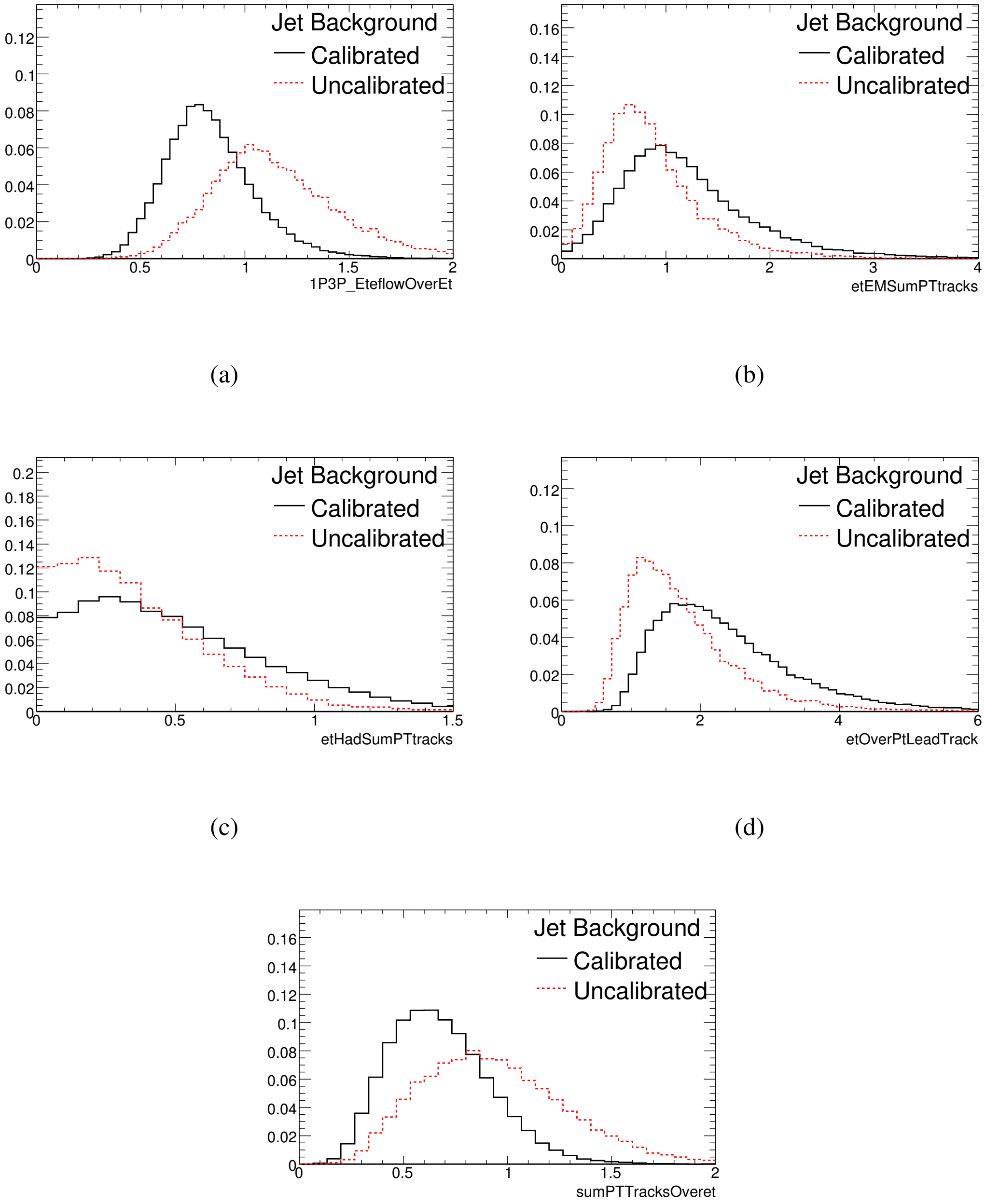
(b)

(c)

(d)

(e)

Figure 6.32: Variable distributions for signal taus comparing calibration schemes.



(a)

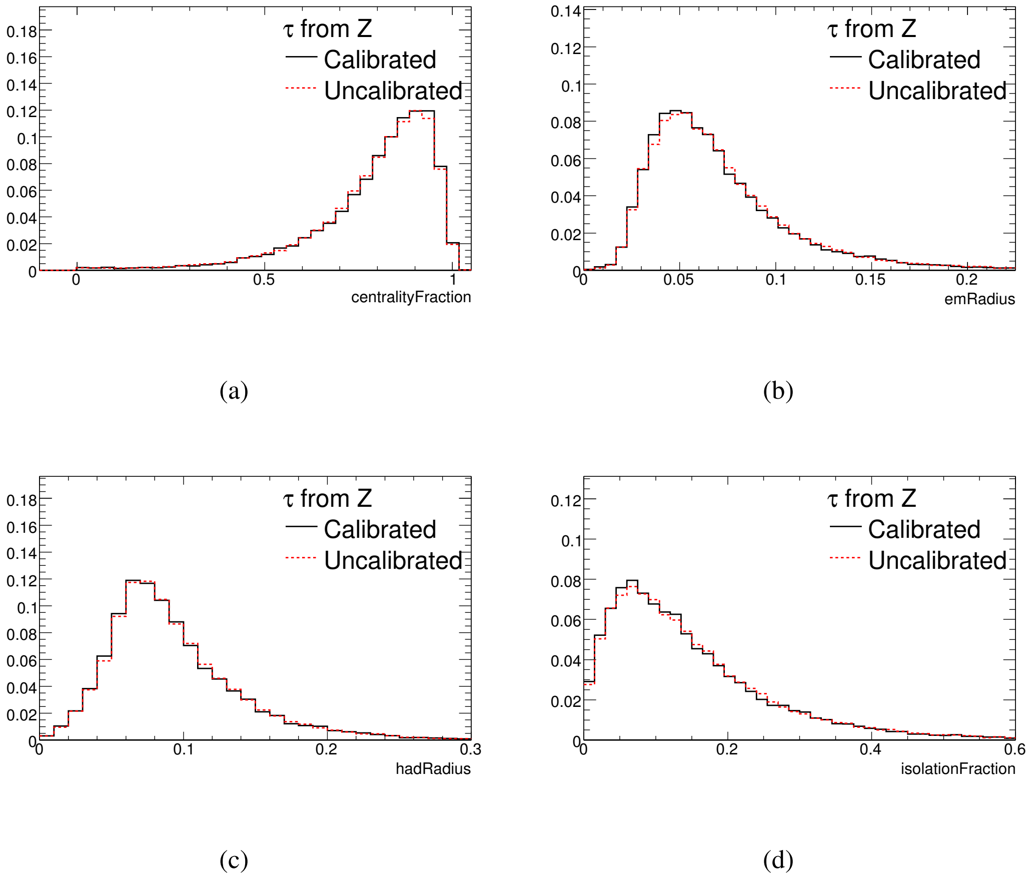
(b)

(c)

(d)

(e)

Figure 6.33: Variable distributions for background tau candidates comparing calibration schemes.



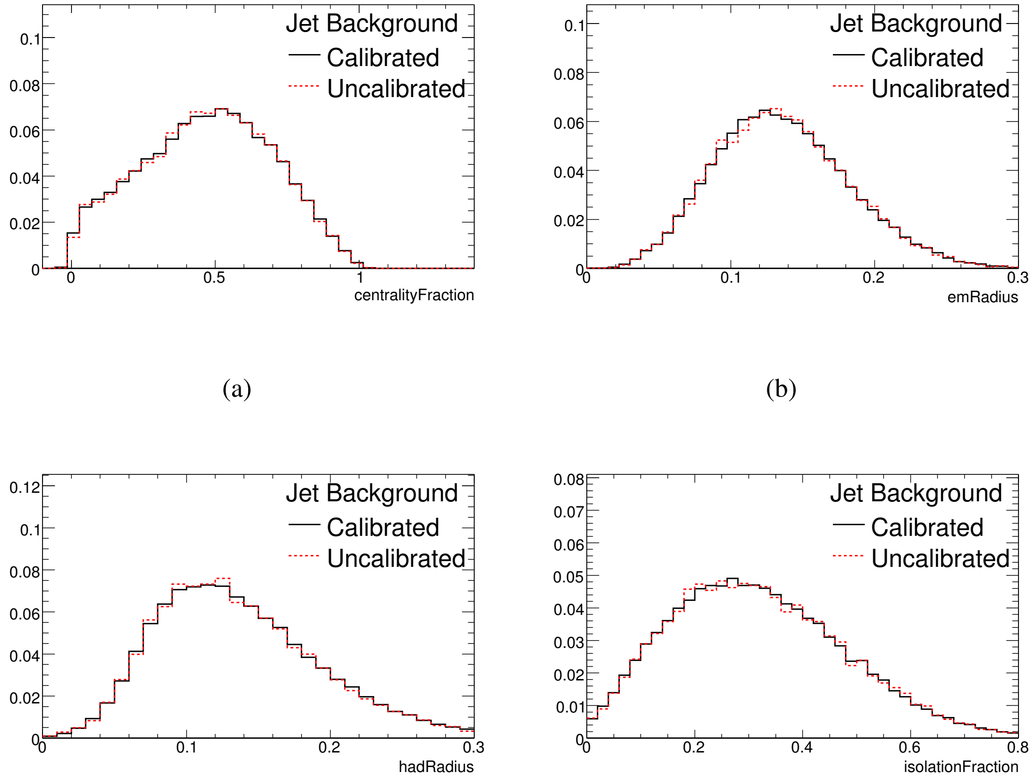
(a)

(b)

(c)

(d)

Figure 6.34: Width type variable distributions for signal taus comparing calibration schemes.



(a)

(b)

(c)

(d)

Figure 6.35: Width type variable distributions for background tau candidates comparing calibration schemes.

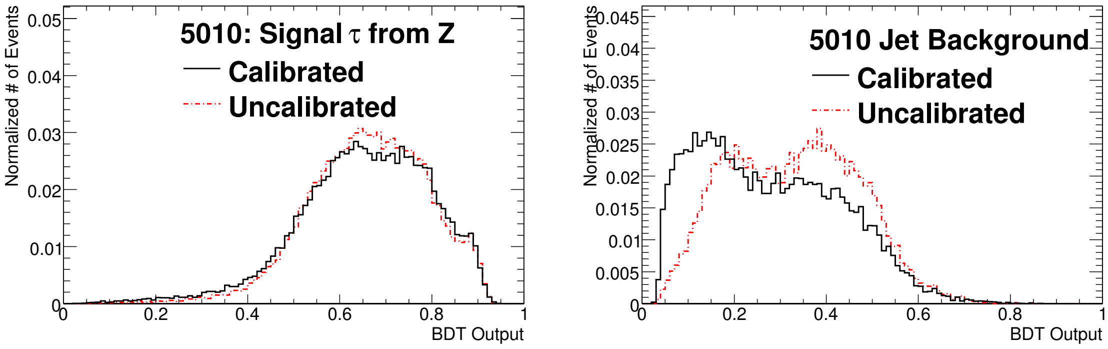
Table 6.4: Training statistics for calibrated sample. The background types available through ATLAS are binned by energy. Each background event is weighted by cross-section.

Sample	# of Calib Training Events	# of Calib Testing Events	# of Uncalib Testing Events	Cross-section (pb)
Signal				
$Z \rightarrow \tau\tau$	21537	35184	36154	246
Background Dijet Events				
17-35 GeV	57235	28120	53400	1.38×10^{9}
70-140 GeV	323545	-	-	5.88×10^{6}
140-280 GeV	57659	-	-	3.08×10^{5}
280-560 GeV	23925	-	-	1.25×10^{4}

A boosted decision tree was trained on the calibrated sample using the number of training events listed in Table 6.4. The training was of 50 boosting cycles using an Adaboost $\beta = 0.2$ and MinLeafSize=100. It was then applied both to calibrated and uncalibrated independent testing samples. The background testing sample was comprised entirely of a sample in the range 17-35 GeV (labelled "5010") to avoid a weighted background. This simplifies error calculations as well as direct comparisons between two different samples.

The signal and background BDT scores are shown in Figures 6.36(a) and 6.36(b) respectively. The uncalibrated background sample shows a shift towards higher BDT output values for a BDT score of up to about 0.6. The background shift to higher BDT scores can be expected by the variable shifting. For example, the maximum of the variable "fraction of energy flow to total E_T" (6.33(a)) shifts to just above 1. Similarly, the fraction of the sum of the track momentum to the total E_T shifts significantly towards higher values. Both of these shifts cause the background to appear more tau-like.

Throughout the range of signal efficiencies from 0.59 to 0.39 the calibrated sample performs slightly better than the uncalibrated sample. This corresponds to BDT cuts from about 0 to 0.63. The performance of this region can be seen in Figure 6.37. The errors are calculated using a Bayesian method, the details of which are in Appendix A.



(a) Signal samples.

(b) Background samples.

Figure 6.36: Boosted decision tree output for calibrated and uncalibrated samples.

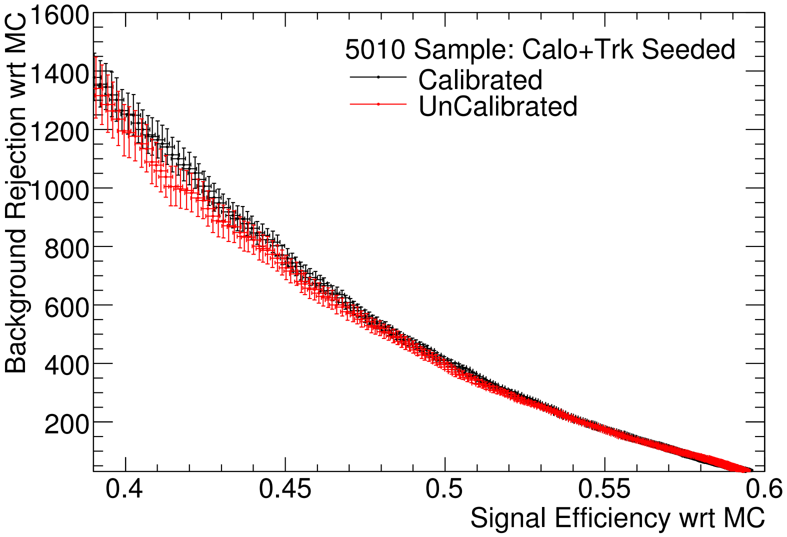


Figure 6.37: Performance of the BDT trained on the calibrated sample and applied to the calibrated and uncalibrated testing set are overlayed for BDT cuts corresponding from 0 to 0.63.

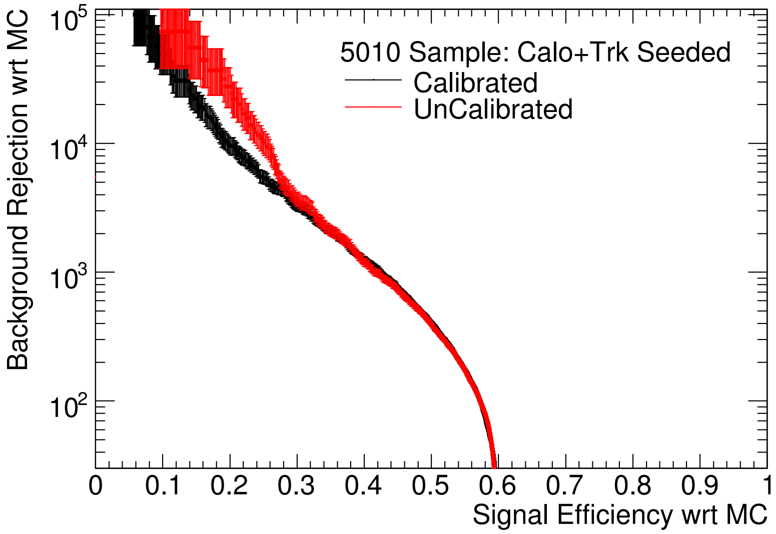


Figure 6.38: Background rejection vs signal efficiency of the BDT trained on the calibrated sample and applied to the calibrated and uncalibrated testing set are overlayed for the entire BDT cut range.

In the region in which very few background events pass the BDT cut (the high background rejection region), the uncalibrated sample performs better. This is shown in Figure 6.38. Because of the small number of background events remaining after cuts above 0.63, it is questionable as to whether this difference in performance is statistically significant.

Signal efficiencies with corresponding background rejection and BDT cut for the calibrated and uncalibrated testing results are quantified in Table 6.5. The column corresponding to 15% signal efficiency is provided to demonstrate how few testing candidates pass cuts in the high rejection region, which leads to high uncertainty. Table 6.6 shows the signal efficiencies which correspond to a particular BDT cut for both the calibrated and uncalibrated sample. While the signal efficiencies do lie outside the error range for several cuts, the differences are small. One can expect only small deviations in expected signal efficiency in such a calibration situation.

Overall, the uncalibrated sample shows a shift towards higher values in the BDT output of the background and little change in the signal sample. The result is that over the range of most BDT cuts, a signal efficiency quoted for a given BDT cut which was calculated on

Table 6.5: Signal efficiencies with corresponding background rejection and BDT cut for the calibrated and uncalibrated testing results for calo+trk seeded candidates

Signal Eff.	15%	30%	40%	50%
	Calibrated; # MC Jets: 399836			
Bkg Rejection	22212^{+5687}_{-4856}	3250^{+302}_{-285}	1252^{+72}_{-69}	403^{+13}_{-13}
BDT Cut	0.771	0.68	0.62	0.548
# Jets Pass	18	123	319	990
	Uncalibrated; # MC Jets: 221304			
Bkg Rejection	55325^{+33259}_{-23666}	3687^{+450}_{-455}	1195^{+90}_{-86}	396^{+17}_{-17}
BDT Cut	0.764	0.679	0.625	0.557
# Jets Pass	4	60	185	557

Table 6.6: Signal efficiencies for the calibrated and uncalibrated samples corresponding to the same BDT cut for calo+trk seeded candidates

BDT Cut	0.74	0.68	0.62	0.549
Signal Efficiencies				
Calibrated	20.07% ±0.18	30.05%±0.21	40.08% ±0.22	49.97% ±0.23
Uncalibrated	19.21%±0.17	29.90% ±0.20	40.83 ±0.22	50.95% ±0.22

a calibrated sample will yield the same signal efficiency for an uncalibrated sample. Signal acceptance expectations do not have to be adapted for an uncalibrated sample. However, one should expect a slightly higher background acceptance for an uncalibrated sample up to BDT cut of about 0.63.

6.5 Implementation in ATLAS Software

BDTs for tau identification have been implemented in TauDiscriminant, the tau identification package within the ATLAS Athena software package. This package runs on every reconstructed tau candidate and provides multi-variate discrimination scores for background rejection. Details on the package have been given in a talk by Marcin Wolter [26].

Chapter 7

Summary and Outlook

ATLAS will begin taking data on *p*-*p* collisions in 2009. Understanding and identifying standard model particles is an important aspect of the experiment both to ensure proper functioning of the detector and for precision measurements. This is especially necessary during early data taking. Taus are one of these standard model particles whose identification and properties must be well-understood in ATLAS. They are especially difficult to identify because jets look very similar and have a very high cross-section. Taus are also signatures for the Higgs decay as well as for some SUSY events and are therefore of particular importance.

Boosted decision trees have been shown to be valuable for tau identification. They are a fast and flexible alternative to existing discriminants and perform better than baseline discriminants without requiring samples to be divided into E_T and prong bins. The use of safe variables exclusively was also studied and shows a degradation in background rejection when using the calo plus track approach from 1227 to 335 for calo+trk seeded candidates. A calibration study was also performed and showed that the expected signal efficiency does not change through these calibration changes but that background acceptance may increase up to moderate BDT cuts. An implementation has been included in the ATLAS software in the TauDiscriminant package.

Appendix A

Uncertainty Calculations Using Bayes' Theorem

The uncertainties in this analysis were calculated by a method developed by Marc Paterno and implemented as the BayesDivide method in the ROOT analysis software [27]. A summary of the method, derived from the documentation, is included for completeness [28].

Often in high energy physics analysis either a Poisson or binomial error calculation is used. However, these both break down in certain boundary regions in the case of efficiency uncertainties, where k events out of a total N pass a cut. The BayesDivide method is designed specifically to calculate uncertainties in efficiency values when the value of k is close to 0 or N.

For example, in a Poisson distribution the uncertainty δ_k in k is $\sqrt{k}$ and δ_N in N is $\sqrt{N}$ so that the uncertainty in efficiency is given by

$$\begin{aligned} \delta_{\varepsilon'} &= \varepsilon' \sqrt{\left(\frac{\delta_k}{k}\right)^2 + \left(\frac{\delta_N}{N}\right)^2} \\ &= \sqrt{\frac{k^2(N+k)}{N^3}}. \end{aligned} \tag{A.1}$$

When no events pass the cut, $k = 0$, which results in an error of $\delta_{\varepsilon'} = 0$. Likewise, $k = N$ results in $\delta_{\varepsilon'} = 1 \pm \sqrt{2/N}$, which is greater than 1. Neither of these results is reasonable, so it appears that the boundaries of the possible values for k are not treated properly by this method.

Similarly, the binomial method would estimate that

$$\begin{aligned}\sigma_k &= \sqrt{var(k)} \\ &= \sqrt{\varepsilon(1-\varepsilon)N}.\end{aligned} \tag{A.2}$$

The error in efficiency is found by dividing both sides of the equation by N to obtain

$$\delta_{\varepsilon'} = \frac{1}{N}\sqrt{k(1-k/N)}. \tag{A.3}$$

When $k = 0$ or $k = N$, the error is zero, which cannot be correct.

For this reason, a Bayesian approach was used. We want to know $P(\varepsilon|k,N,I)$, the probability that given an observed measurement of k events passing out of N (with prior information I) the efficiency is truly ε. Bayes' Theorem states that

$$P(\varepsilon|k,N,I) = \frac{P(k|\varepsilon,N,I)P(\varepsilon|N,I)}{Z}. \tag{A.4}$$

That is, the probability that ε is the true efficiency is equal to the probability that k events will pass for an efficiency ε times the probability that, prior to any measurements, the efficiency will be ε, all divided by a normalization constant. A binomial distribution is assumed so that

$$P(k|\varepsilon,N,I) = \frac{N!}{k!(N-k)!}\varepsilon^k(1-\varepsilon)^{N-k} \tag{A.5}$$

and all physically reasonable efficiencies, prior to considering the data, were given equal probability:

$$P(\varepsilon|N,I) = \begin{cases} 1 & \text{if } 0 \le \varepsilon \le 1, \\ 0 & \text{otherwise.} \end{cases} \tag{A.6}$$

The normalization constant is found by

$$\int_{-\infty}^{\infty} P(\varepsilon|k,N,I)d\varepsilon = 1. \tag{A.7}$$

The full solution becomes

$$P(\varepsilon|k,N,I) = \frac{\Gamma(N+2)}{\Gamma(k+1)\Gamma(N-k+1)}\varepsilon^k(1-\varepsilon)^{N-k}, \tag{A.8}$$

which gives the probability distribution as a function of efficiency for given values of k and N. The shortest 68.3% confidence interval of this probability function is then used to characterize the uncertainty. This effectively results in a 1 standard deviation Gaussian error. The confidence interval does not have a closed-form solution, so the error is calculated numerically by the BayesDivide method.

Bibliography

[1] ed. L. Alvarez-Gaumé et al. Review of particle physics. *Physics Letters B*, 667:489–517, 2008.

[2] A. Christov et al. Performance of the tau reconstruction and identification algorithm with release 14.2.10. *ATLAS Physics Communication*, ATL-COM-PHYS-2008-196, 2008.

[3] A. Djouadi. The anatomy of electro-weak symmetry breaking. i: The higgs boson in the standard model. hep-ph/0503172, 2005.

[4] K. Assamagan, Y. Coadou, and A. Deandrea. Atlas discovery potential for a heavy charged higgs boson. *EPJ Direct*, 4:9, 2002.

[5] LHC Project Illustration. http://doc.cern.ch//archive/electronic/cern/others/PHO/photo-lhc//lhc-pho-1991-001.jpg.

[6] ATLAS Detector Photos. http://atlas.ch/photos/index.html.

[7] ATLAS Collaboration, G. Aad et al. *The ATLAS Experiment at the CERN Large Hadron Collider*, 2008. https://twiki.cern.ch/twiki/pub/Atlas/ AtlasTechnicalPaper/Published_version_jinst8_08_s08003.pdf.

[8] ATLAS Collaboration. Expected performance of the atlas experiment, detector, trigger and physics. CERN-OPEN-2008-020, 2008.

[9] Otto Nachtmann. *Elementary Particle Physics*. Springer-Verlag, 1990.

[10] David Griffiths. *Introduction to Elementary Particles,* 2$^{\text{nd}}$ *Revised Ed.* Wiley-VCH, 2008.

[11] A. Kazmarska. Tau leptons as a probe for new physics at lhc. *Nuclear Physics B (Proc. Suppl.)*, 169:351–356, 2007.

[12] *ATLAS Detector and Physics Performance Technical Design Report Volume I and II*, 1999. http://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/TDR/access.html.

[13] Torbjorn Sjostrand et al, 2000. http://home.thep.lu.se/ torbjorn/pythiaaux/past.html.

[14] Geant4 Collaboration. http://www.geant4.org/geant4.

[15] M. Perl. *Physics Review Letters*, 35:1489, 1975.

[16] D. Froidevaux, P. Nevski, and E. Richer-Was. Energy-flow studies for hadronic τ-decays using dc1 data samples. *ATLAS Physics Communication*, ATL-COM-PHYS-2005-024, 2005.

[17] The ATLAS Collaboration. Jet reconstruction performance. *CSC Note*, CERN-OPEN-2008-020.

[18] ATLAS Twiki. http://atlas-computing.web.cern.ch/atlas-computing/links/ nightlyDevDirectory/AtlasOffline/latest_doxygen/InstallArea/doc//TrkParameters/html// classTrk_1_1Perigee.html.

[19] G. Piacquadio, K. Prokofiev, and A. Wildauer. Atlas primary vertex reconstruction. *ATLAS Note*, in Preparation, 2008.

[20] Stan Lai. Private communication, 2008.

[21] A. Kalinowski and K. Benslama. Tau identification with the logarithmic likelihood method. *ATLAS Physics Communication*, 2008. ATL-COM-PHYS-2008-196.

[22] V.M. Abazov. et al. (DØ Collaboration). Evidence for production of single top quarks. *Phys. Rev. D*, 78, 2008.

[23] Y. Freund and R.E. Schapire. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, 1999.

[24] A. Hoecker et al. Toolkit for multivariate data analysis, 2005-2007. http://tmva.sourceforge.net/.

[25] S. Protopopescu and P. Svoisky. Calculation of the τ identification neural network systematic uncertainty for p17 data.

[26] Marcin Wolter. Tau identification using multivariate techniques in atlas. XII International Workshop on Advanced Computing and Analysis Techniques in Physics Research, 2008. http://acat2008.cern.ch.

[27] Rene Brun and Fons Rademakers. Root, an object-oriented data analysis framework. http://root.cern.ch/.

[28] Marc Paterno. Calculating efficiencies and their uncertainties, 2003. http://home.fnal.gov/ paterno/images/effic.pdf.