TECHNISCHE UNIVERSITÄT DRESDEN

Identification of boosted hadronic tau pair decays at ATLAS

Bachelor-Arbeit
zur Erlangung des Hochschulgrades
Bachelor of Science
im Bachelor-Studiengang Physik

vorgelegt von

FRANZISKA KATHRIN SCHOGER
geboren am 27.03.1995 in Starnberg

Institut für Kern- und Teilchenphysik
Fachrichtung Physik
Fakultät Mathematik und Naturwissenschaften
Technische Universität Dresden
2016

Eingereicht am 03. Juni 2016

1. Gutachter: Prof. Dr. Arno Straessner
2. Gutachter: Prof. Dr. Michael Kobel

Summary

The Standard Model of Particle Physics is a very successful theory but it has several shortcomings. These can be solved in extended theories, which mostly predict more particles with higher masses than those, that were already found. A heavy particle can decay in leptons and quarks through intermediate steps, leading to a boost for the decay products. Tau leptons that decay hadronically are used for measuring standard model processes but also to find new physics at the ATLAS physics programm. But the standard reconstruction for tau leptons cannot deal with highly boosted topologies so a new reconstruction for di-taus was introduced. This di-tau reconstruction provides nearly no protection against jets, which are produced by fragmented quarks and gluons. To differentiate between di-taus and quarks and gluon jets an identification algorithm is introduced. It distinguishes between signal and background with the help of a multivariate analysis, the boosted decision trees. To provide a good discrimination between signal and background a list of identification variables is introduced and their separation efficiency tested. Furthermore the parameters of the Boosted Decision Trees are adjusted to optimize the performance in signal and background efficiency.

Zusammenfassung

Das Standardmodell der Teilchenphysik ist bis jetzt eine sehr erfolgreiche Theorie, allerdings gibt es mehrere Fragen, die diese Theorie nicht beantworten kann. Weiterführende Theorien, die das beheben, sagen meist auch mehr Teilchen mit höheren Massen voraus. Diese können dann über Zwischenschritte in Leptonen und Quarks zerfallen. Durch diese Zwischenschritte werden die Zerfallsprodukte nicht mehr mit Impulsen antiparallel zueinander erzeugt, sondern schließen einen kleineren Winkel ein. Hadronisch zerfallende Tau Leptonen werden beim Physik Programm bei ATLAS genutzt um Prozesse im Standardmodell zu vermessen, aber auch um neue Physik zu finden. Allerdings verliert die herkömmliche Rekonstruktion von Tau Leptonen bei hoch geboosteten Topologien ihre Effizienz. Aus dem Grund wurde eine neue Rekonstruktion für solche Objekte, sogenannte Di-Taus, eingeführt. Diese rekonstruiert aber nicht nur Tau Leptonen, sondern auch viele Jets, die durch Quarks und Gluonen erzeugt wurden. Um Di-Taus von Quark und Gluon Jets zu unterscheiden wird ein Identifikationsalgorithmus eingeführt. Durch eine multivariate Anlayse, den Boosted Decision Trees, wird Signal und Untergrund getrennt. Um eine möglichst gute Trennung herzustellen werden Identifikationsvariablen eingeführt und ihre Separationsfähigkeit getestet. Weihterhin werden die Parameter der Boosted Decision Trees variiert um eine optimale Performance zu erhalten.

Contents

1 Introduction

In December 2015 the experiments ATLAS [1] and CMS [2] at the Large Hadron Collider (LHC) [3] announced that they had found an excess over the expected value of photon pairs being produced by collisions at a invariant mass of 750 GeV[1] [4, 5]. If this excess proves to be a new particle and not just a statistical fluctuation, it will not be part of the Standard Model of Particle Physics, because since 2012 [6, 7] all particles belonging that theory have been found. Now the analysis of the new data collected in 2016 is awaited expectantly.
But not only photons are used to find new particles. In the ATLAS physics program hadronically decaying tau leptons are used in measurements of standard model processes, Higgs boson searches, searches for new physics and many more [8]. To measure these processes a working tau reconstruction and identification is needed. The conventional single tau reconstruction falters at boosted topologies with a transverse momentum of more than 500 GeV [9]. Boosted topologies are topologies, where the particle, which is decaying, already has a large transverse momentum, which it passes on to its decay products, in addition to the momentum they gain from the decay itself. A new reconstruction algorithm has already been build [9], the goal of this thesis is to provide an identification algorithm.
The Standard Model of Particle Physics, its problems and possible extensions are discussed in chapter 2 and in chapter 3 the LHC and the ATLAS detector are introduced briefly. In chapter 4 tau leptons are described in more detail as well as their reconstruction. Also different types of clustering algorithms to form the jets used in reconstruction and identification will be explained and the source of background considered in this thesis is introduced.
In chapter 5 the identification variables implemented in this thesis are motivated and explained. In chapter 6 the method used for the identification is introduced and in chapter 7 the results are presented. In the last chapter, chapter 8, a summary and an outlook is given.

[1] In this thesis natural units are used, where every unit is expressed in powers of the unit of the energy, see Appendix B.

2 The Standard Model of Particle Physics and Extensions

2.1 The Standard Model of Particle Physics

The Standard Model (SM) [10, 11, 12, 13, 14, 15, 16, 17, 18] is, until now, the most successful theory regarding elementary particles and their interactions. It predicts a range of particles, which can be seen in figure 2.1. The final building block, the Higgs boson, was found in 2012, with characteristics in accordance with the SM [6, 7].
The particles proposed by the SM can be divided into two groups: the fermions, particles with a half-integer spin, which are again divided into leptons and quarks, and the bosons, particles with an integer spin.
The interactions covered by the SM are described as a gauge Quantum Field Theory (QFT). There is the electromagnetic interaction coupling to particles with an electric charge, the charged leptons and the quarks. It is mediated by a neutral and massless boson, the photon. The strong interaction couples to particles with a color charge (red, green or blue), the quarks and is mediated by eight different bosons, the gluons, which carry a colour and a anticolour charge. The theory describing the strong interaction is the Quantum Chromodynamics (QCD), which is based on an SU(3) local gauge symmetry. Free quarks have never been observed, only colourless objects, the hadrons have been found. This effect is called colour confinement. Hadrons are divided into two groups: the mesons, consisting of a quark and an antiquark with colour and anticolour charge, and the baryons, consisting of three quarks with three different colour charges.
The third interaction covered by the SM is the weak interaction. It is combined with the electromagnetic interaction to the electroweak interaction, which is described as a $\mathrm{U}(1)_Y \times \mathrm{SU}(2)_L$ local gauge symmetry. It couples to particles with a Hypercharge Y and/or a weak Isospin which all left handed (L) fermions have. The electroweak interaction is mediated by the $W^{\pm}$ and the Z^0 bosons and the photon. [19]
In experiments one finds that the $W^{\pm}$ and the Z^0 gauge boson are massive. But introducing a mass term into the Lagrangian density leads to its loss of invariance under local gauge transformations, which is the concept all interactions are based on. To solve that problem the Higgs mechanism [21, 22, 23, 24, 25] was introduced. The Higgs field, a scalar field, couples

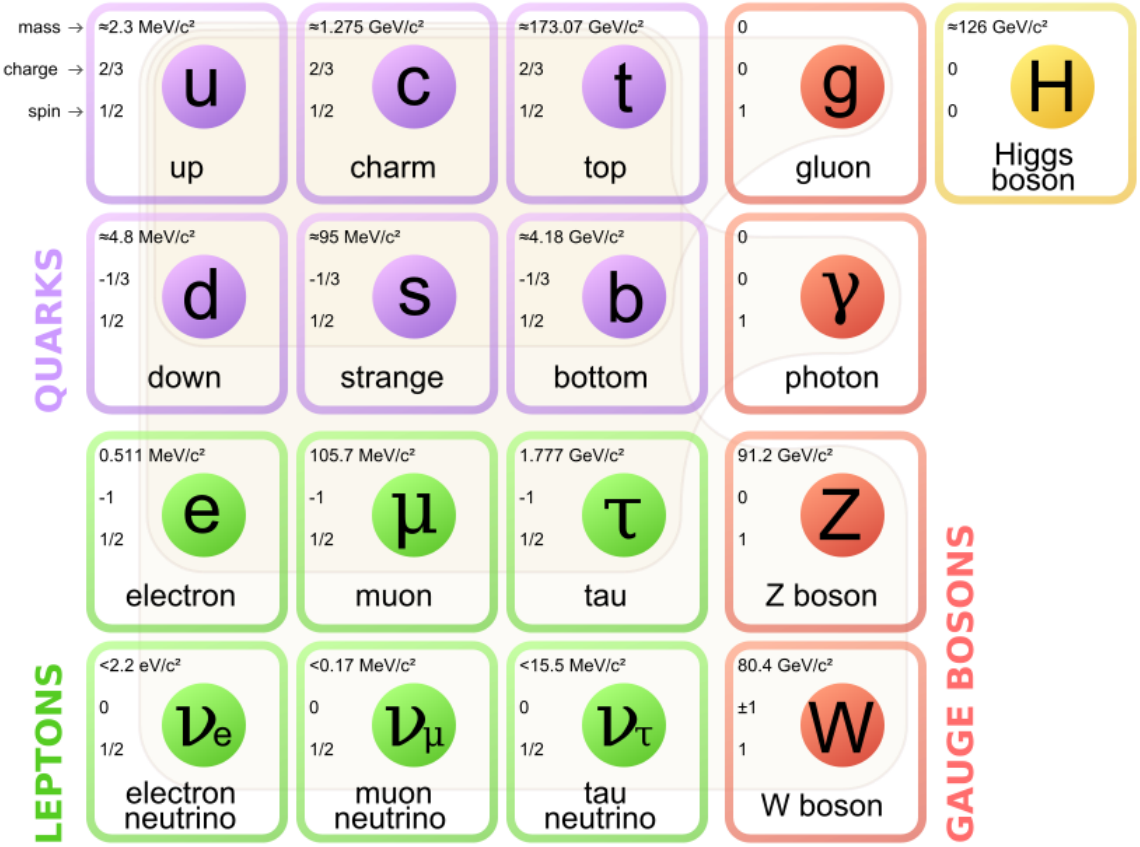


Figure 2.1: Particles of the SM and their properties.[20]

to massive particles and through spontaneous symmetry breaking (choosing the non zero vacuum state), the heavy particles gain their mass. The excitation of the Higgs field is the Higgs Boson. [19]
The Higgs mechanism ensures that the invariance under local gauge transformations is not violated.

2.2 Problems of the SM and Possible Extensions

The SM leads to results in good agreement with experiments until now, but there are several problems, which cannot be solved within the SM.
One shortcoming is that it doesn't cover gravity, the fourth fundamental force. But it would only be favourable to have one theory covering all fundamental forces, called the Grand Unified Theory [26].
Also there is the dark matter [27], which makes up about a quarter of the universe, but interacts, as far as we see, only through gravity and the weak force. The SM does not provide a particle which is heavy enough or exists often enough in the universe to make up the dark matter.

A further problem is that through the observed neutrino oscillations it has been proven, that at least two of the neutrinos have to have a mass, but that is not so in the SM. And even though the SM describes three fundamental forces it does not explain where they originate from. [28]
Because of these and several more shortcomings new theories have been introduced that solve some of these problems, like Supersymmetry, where every particle of the SM gets a supersymmetric partner. The extension with the smallest set of free parameters is called Minimal Supersymmetric Extension of the SM (MSSM) [28], where three neutral Higgs bosons h, H, and A and two charged Higgs bosons $H^{\pm}$ are introduced. The masses of these Higgs bosons depend on the choice of the free parameters of the theory. They can be chosen in such a way, that h is the lightest of them and behaves SM like. Over a wide range in the parameter space the decay rates in third generation fermions, like tau leptons, are comparatively high.

3 The LHC and the ATLAS Detector

3.1 The LHC

The Large Hadron Collider (LHC) [3] is situated at CERN, the European Center for Nuclear Research. It is a ring accelerator with a circumference of 26.7 km, and consists of two superconducting rings that were build between 170 and 45 m underground into the tunnel that already existed from the Large Electron Positron Collider (LEP) [29]. Two transfer tunnels link the LHC to the CERN accelerator complex, which acts as injector [3]. There, hadrons but also heavy ions, like lead, can be accelerated in up to five stages and brought to collision (see also figure 3.1).
Hydrogen atoms are stripped off their electrons, so that only the protons remain. These are accelerated in the linear accelerator LINAC2 to an energy of 50 MeV, then they are injected into the BOOSTER, a circular accelerator, where they gain an energy of 1.4 GeV. This is followed by two more synchrotrons, the proton synchrotron (PS), accelerating the protons to 25 GeV and the super proton synchrotron (SPS) pushing them to 450 GeV. Then the proton beam is sent to the LHC rings where it can be accelerated even further. Thereby one beam is circled clockwise, the other anticlockwise. [30]
At four points the beams can be brought to collision, there the four experiments are situated: ATLAS [1] and CMS [2], the two general purpose detectors, ALICE [32], the detector specialised for studies of lead-lead ion collisions and LHCb [33] specialised for physics concerning b-Quarks.
The LHC aims to discover new physics beyond the SM. But the exploration of rare events needs high event rates $^{\mathrm{d}N}/_{\mathrm{d}t}$ which depend on the luminosity L and the physical cross section of the process σ:

$$\frac{\mathrm{d}N}{\mathrm{d}t} = L \cdot \sigma. \tag{3.1}$$

With peak luminosities of $10^{-34}\,{}^{1}/_{\mathrm{cm}^2\mathrm{s}}$ and center of mass energies up to 14 TeV this can be achieved. [3]

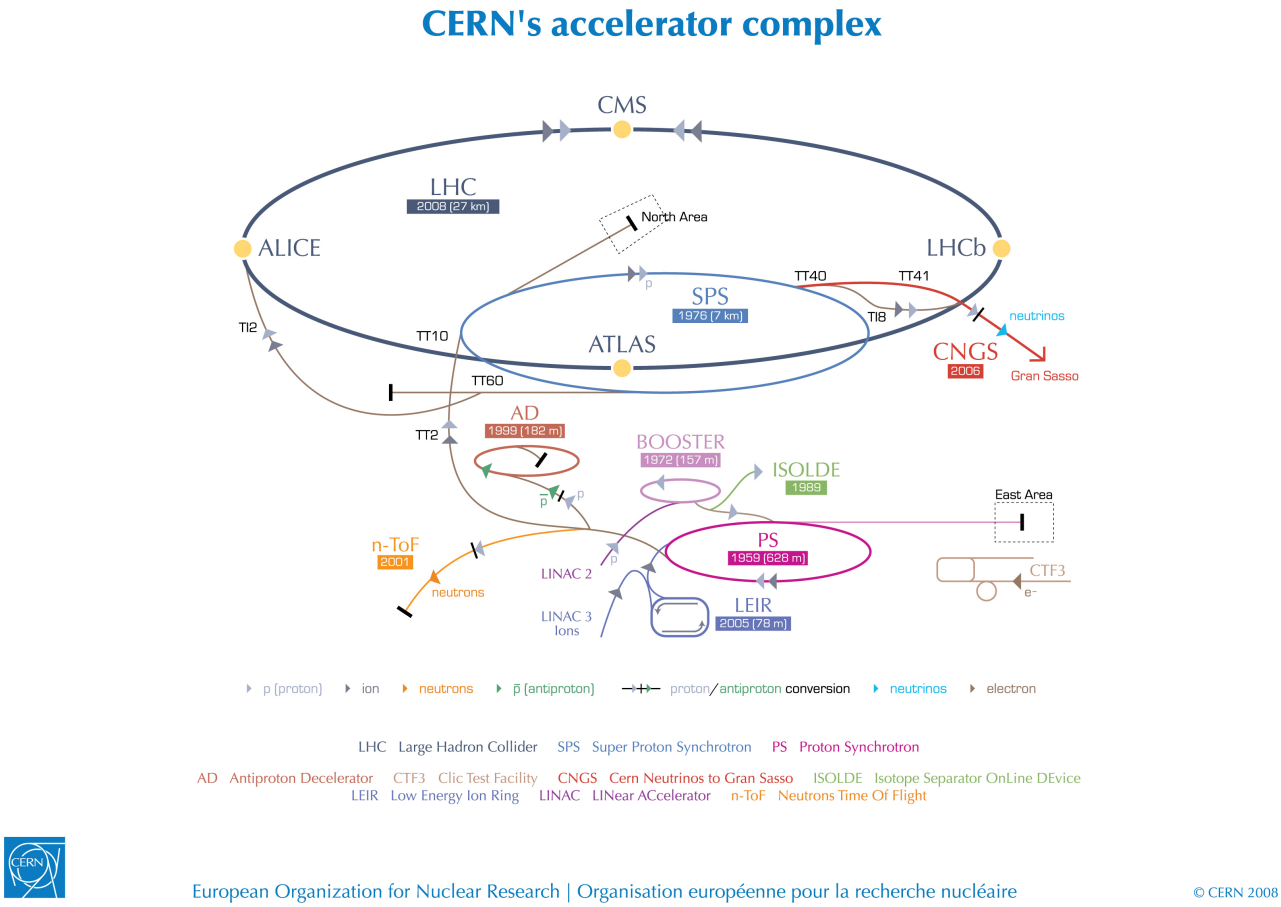


Figure 3.1: The accelerator complex at CERN. [31]

3.2 The ATLAS Detector

A Toroidal LHC Apparatus (ATLAS) [1] is one of the experiments at the LHC ring. The detector is about 25 m high and 44 m long and weights about 7000 t. It is forward-backward symmetric with respect to the interaction point and build in several layers, like displayed in figure 3.2. There are the inner detector, the calorimeters, the muon spectrometers and the forward detectors.

The **inner detector** was build for pattern recognition, momentum, vertex and charge measurement and electron identification. Therefore pixel and silicon microstrip trackers and straw tubes of transition radiation trackers were build in the detector. The whole inner detector is immersed in a 2 T magnetic field to bend the tracks of charged particles and so enables the measurement of the charge and momentum of particles with an electric charge.

The next layers are the **calorimeters** for exact position and energy measurements. The electromagnetic calorimeter consists of liquid argon detectors (LAr), the hadronic calorimeter of scintillator tiles in the barrel and LAr in the end caps. The forward calorimeter also consists of LAr and is build to measure electromagnetic as well as hadronic interactions. Particles interacting with the calorimeters produce showers (a cascade of particles). In the electromagnetic

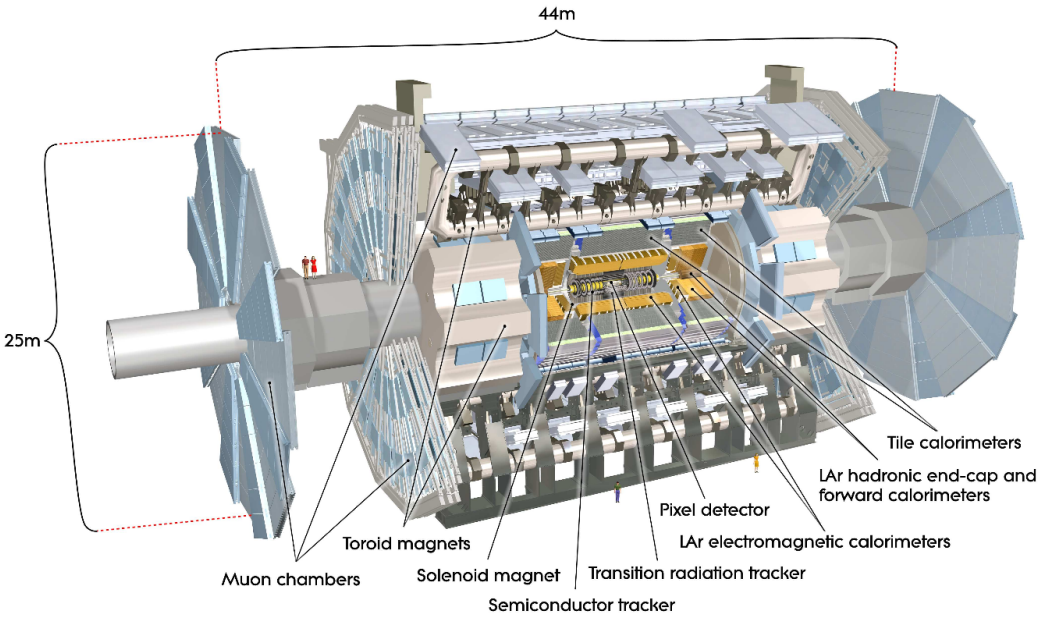


Figure 3.2: Schematic view of the ATLAS detector. [1]

calorimeter electromagnetic showers by photons or electrons are produced, in the hadronic calorimeter hadronic showers by e.g. protons, neutrons or pions are produced. The size of the calorimeters was planned in such a way that all showers are well contained and that there is no punch-through to the next layer.

The next layer is the outermost layer, the **muon chambers**. They consist of a long barrel with two inserted end-cap magnets, which can bend the muon tracks within a large volume. A precise momentum measurement can be achieved in the three layers of tracking chambers.

The **forward detectors** are three smaller systems, which cover the forward region of the ATLAS detector. Two of those measure the luminosity, the third system measures the centrality of heavy ion collisions.

Combining all information from the different layers of the detector, the type of the particles reaching the detector and their momentum can be determined. From that the processes that happened during the collisions can be reconstructed.

But there is too much data to store collected by the detector. To reduce the amount and filter out the interesting events there is a **trigger system**. The L1 trigger searches for high transverse momentum muons, electrons, and jets, amongst others, and large missing transverse or total energy. Thus, it reduces the amount of data. Furthermore the L1 defines regions of interest, which the second trigger L2 uses. L2 uses the full resolution of the detector within those regions to select which events pass this step. The last trigger is an event level trigger, which uses offline event processing methods. After these steps the event rate is reduced from about 40 MHz to 200 Hz. [1]

ATLAS uses a coordinate system, where the origin is the nominal interaction point. The z-axis is defined parallel to the beam direction, therefore the x-y-plane is the plane perpendicular to

the beam direction. The positive y-axis is defined as pointing upwards, the positive x-axis as pointing to the center of the LHC ring.
Most of the time x and y coordinates are unhandy to use. They are transformed to the azimuthal angle ϕ, measured around the beam axis and the rapidity y for massive objects or the pseudorapidity η for massless objects. These are defined as:

$$\eta = -\ln\left(\tan\left(\frac{\theta}{2}\right)\right), \tag{3.2}$$

$$y = \frac{1}{2}\ln\left(\frac{E + p_z}{E - p_z}\right), \tag{3.3}$$

where θ is the polar angle, the angle relative to the beam axis. Using these definitions, $\eta \approx y$ for objects where $E \gg m$.
All transverse parameters, like the transverse momentum p_T or the transverse energy E_T, are defined in the x-y-plane. Distances ΔR are defined in the pseudorapidity-azimuthal plane as:

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}. \text{ [1]} \tag{3.4}$$

4 Tau Leptons and Reconstruction

4.1 Overview over Tau Leptons

The tau lepton is the heaviest of the leptons. It has an electric charge of $|Q| = 1\,\mathrm{e}$ and a mass of $1.777\,\mathrm{GeV}$ [34]. It is very short lived, having a mean life time of $2.9 \cdot 10^{-13}\,\mathrm{s}$ [34], in most cases decaying before it can reach the active regions of the detector. Therefore it can only be detected through its decay products. It can decay either leptonically ($\tau^- \to l^- \overline{\nu}_l \nu_\tau$[1], where l^- marks either e^- or μ^-) or hadronically ($\tau^- \to$ hadrons ν_τ). In 65% of all cases it decays hadronically, out of these decays in 72% one charged pion is produced, in 22% three charged pions [8]. These decays are called 1 prong and 3 prong respectively:

$$1 \text{ prong} : \tau^- \to \pi^- \nu_\tau (n \cdot \pi^0), \tag{4.1}$$

$$3 \text{ prong} : \tau^- \to \pi^- \pi^- \pi^+ \nu_\tau (n \cdot \pi^0). \tag{4.2}$$

In the majority of the hadronic decays left a charged kaon is present.
In over 3 quarters of all hadronic decays only up to one neutral pion is produced. The hadrons produced in the decay make up the visible decay products, they are refered to as $\tau_{\text{had-vis}}$. [8] Hadronically decaying tau leptons are used to measure SM processes but also to find new particles in the ATLAS physics program.

4.2 Clustering Algorithms

To form jets, which are used in the reconstruction and identification, clustering algorithms are needed. Thereby individual jets are derived from the calorimeter energy deposits. This is done by defining two distances, d_{ij}, the distance between the two entities (particles or pseudojets) i and j calculated as

$$d_{ij} = \min(k_{ti}^{2p}, k_{tj}^{2p}) \frac{\Delta_{ij}^2}{R^2}, \tag{4.3}$$

[1] The antitau τ^+ decays analogously in every decay mode, where all particles are charge conjugated and particles become antiparticles and vice versa.

and d_{iB}, the distance between the entity i and the Beam B defined as

$$d_{iB} = k_{ti}^{2p}, \tag{4.4}$$

with $\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$, y_i the rapidity, ϕ_i the azimuth and k_{ti} the transverse momentum of the particle i. The parameter p controls the relative power of the energy versus geometrical scales. For the anti-k_t [35] algorithm $p = -1$, for Cambridge Aachen [36] $p = 0$ and for the k_t algorithm [37] $p = 1$ was chosen.
In the clustering process one searches for the smallest of theses distances. If it is d_{ij}, the entities i and j are combined, if it is d_{iB} i is going to be called a jet and removed from the list of entities. Then the distances are recalculated and the procedure is repeated until there are no more entities left.
R is the distance parameter, defining the radius of the jets. The anti-k_t algorithm produces a cone around a particle with high transverse momentum as long as there is no other such particle within a close range. If there is a particle closer than R they will form one jet. For the other algorithms the forms are more complex. [35]

4.3 Single Tau Reconstruction and Identification

The single tau reconstruction starts with the energy deposits that were reconstructed as jets by the anti-k_t algorithm with a distance parameter of 0.4. Thereby the events must have a primary vertex with at least three associated tracks and the jet must have at least a p_{T} of 10 GeV and a $|\eta| < 2.5$ [8]. Tracks are associated with the tau candidate if they are within the core region ($R < 0.2$) and satisfy quality cuts. Tracks that satisfy these quality cuts but are in the isolation region ($0.2 < R < 0.4$) are called isolation tracks. Also π^0 belonging to the tau candidate are reconstructed.
To differentiate between jets, formed by fragmented quarks and gluons, and tau signal several identification variables are introduced. With those boosted decision tree algorithms are trained, separately for 1 prong and 3 prong decays. Requirements on the score were made in such a way, that the resulting efficiency is independent of the tau p_{T}. [8]

4.4 Di-Tau Reconstruction

As long as the topologies are not boosted the single tau reconstruction works very well. But it falters for highly boosted topologies, like $A \to Zh \to ll + \tau\tau$. Here the Higgs boson decaying into two τ already has a transverse momentum which leads to a smaller angle between the two

tau leptons. If the angle and therefore the distance between the taus is to small ($\Delta R < 0.4$) they will not be reconstructed as two jets but as one. Therefore a new reconstruction algorithm was introduced, which deals with the two taus as one object, the di-tau object, by forming a anti-k_t jet of radius 1. Within this jet anti-k_t subjets are searched for with a radius of 0.2. For a jet only those are considered di-tau candidates, which fulfil following criteria:

- The jet should have at least two subjets.
- The leading and subleading subjet should have at least one track each.

Furthermore the jet should have a p_{T} greater that 15 GeV and $|\eta| < 2.5$.
For the subjets a core region is defined with a radius of $\Delta R < 0.1$. Then there are three regions within a jet: the core region within a subjet, the subjet region and the isolation region outside the subjets.
Tracks found in the jet can be divided in three groups:

- Tracks: they are within a subjet and satisfy quality criteria.
- Isolation tracks: they are outside of the subjets but satisfy quality criteria.
- Other tracks: they fail the quality criteria.

The di-tau four-momentum is calculated as the sum of the four-momenta of the leading and subleading subjet. [9]

4.5 QCD Jets and Pileup

When identifying hadronically decaying tau leptons the main source of background is naturally QCD jet background. These are high energetic jets produced through the fragmentation of gluons and quarks [8]. Thereby high energetic quarks or gluons produced in the collision fragment through strong interaction into more quarks and gluons, which can fragment even further, forming bunches. The quarks combine to hadrons which make up the measured jet.
Additionally to the desired collision there are many more collisions taking place in the same time slot and are therefore measured together. These can come from other particles in the proton interacting, other protons in the same bunch interacting or protons of another bunch interacting, arriving faster than the measuring time of the detector. The number of primary interaction vertices in addition to the desired one is referred to as pileup μ.
The di-tau reconstruction algorithm provides nearly no protection against QCD jets. In the following chapters an identification algorithm is introduced to differentiate between di-taus and QCD jets.

5 Identification Variables

To distinguish di-tau signal form QCD jet background different variables are introduced, which are described and discussed here. Most of them also have an equivalent for single tau identification, which can be found in [8], some of them were already used in [9].
To test the variables and efficiencies discussed in the following chapters simulated events are used as signal as well as background input. They are described together with their p_{T} distribution in Appendix A.
Not all variable distribution can be displayed here, those missing can be found in appendix C.

Subjet Core Momentum Fraction

f_{core} is defined as the fraction of the transverse momentum of particles measured in the core region of a subjet over the transverse momentum measured in the whole subjet:

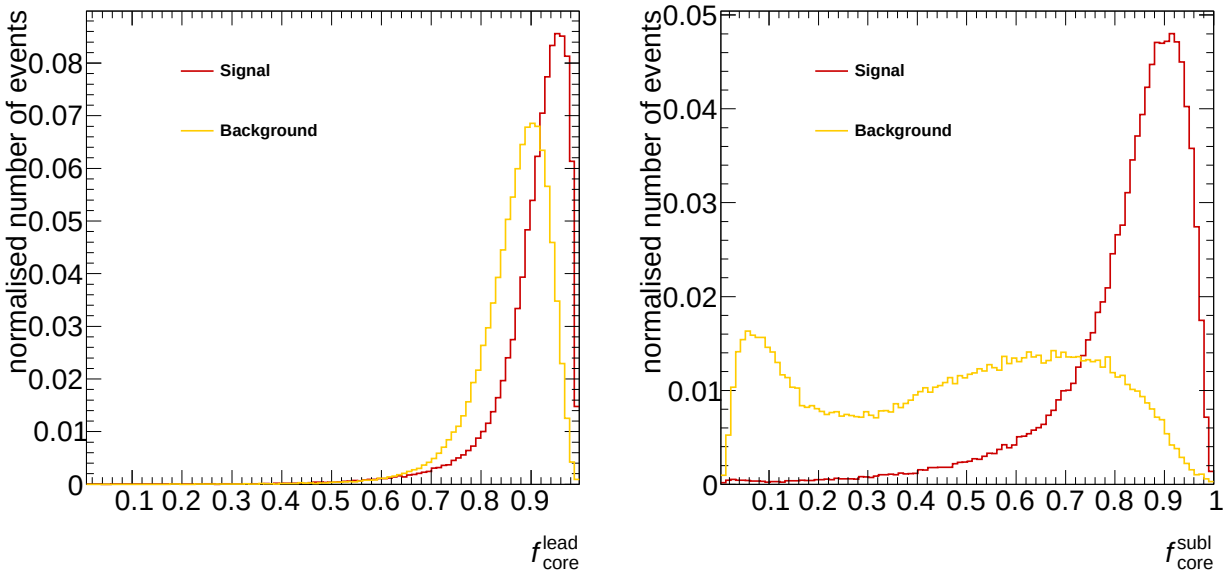
$$f_{\mathrm{core}} = \frac{\sum_{cells}^{\Delta R<0.1} p_{\mathrm{T}}}{\sum_{cells}^{\Delta R<0.2} p_{\mathrm{T}}}, \tag{5.1}$$

where p_{T} is the transverse momentum registered by the calorimeter cells. It is defined for the leading as well as for the subleading subjet individually.

For the leading subjet both signal and background have a narrow distribution where the background is only slightly wider. This behavior results from the fact, that almost all energy is deposited in the core region of the subjet as expected from a jet. The same behaviour can be seen for the signal in the subleading subjet (see figure 5.1). Here the distribution is also wider than for the leading subjet which can be explained by the fact that the subleading tau has less energy than the leading tau and therefore its decay products are less collimated. Completely different is the distribution for the background in the subleading subjet. Here the object found has no narrow distribution. The energy is not primarily in the core region but is spread over the whole subjet, allowing the assumption that it results from no independent jet.

Subjet Momentum Fraction

f_{subjet} is the transverse momentum of a subjet compared to the transverse momentum of the



(a) Distribution for the leading subjet.

(b) Distribution for the subleading subjet.

Figure 5.1: Normalised distributions of the f_{core} variables. Yellow represents QCD jet background, red di-tau signal.

whole jet:

$$f_{\text{subjet}} = \frac{p_{\text{T}}^{\text{subjet}}}{p_{\text{T}}^{\text{jet}}}. \tag{5.2}$$

Again, it is introduced separately for the leading and subleading subjet. Also the sum of leading and subleading Subjet Momentum Fraction can be defined:

$$f_{\text{subjets}} = \frac{p_{\text{T}}^{\text{leading subjet}} + p_{\text{T}}^{\text{subleading subjet}}}{p_{\text{T}}^{\text{jet}}}. \tag{5.3}$$

It is the ratio of the p_{T} of the two expected tau leptons and the p_{T} of the whole jet.

As QCD background events have no distinct di-jet structure most of their energy is found in the leading subjet leaving the subleading subjet with only a small fraction of the overall p_{T}. That is different for the signal events. Through the distinct di-jet structure no subjet stands out. For $f_{\text{subjet}}^{\text{subl}}$ the distribution can be seen in figure 5.2, the distributions for the two other f_{subjet} variables can be found in figure C.1.
For f_{subjets} the distributions for signal and background are similar. The background is only slightly wider and has its maximum at a slightly lower fraction. For both signal and background most of the energy can be found in the leading two subjets, but the background has overall more subjets which all carry some energy (see below) leading to a maximum shifted to smaller values.

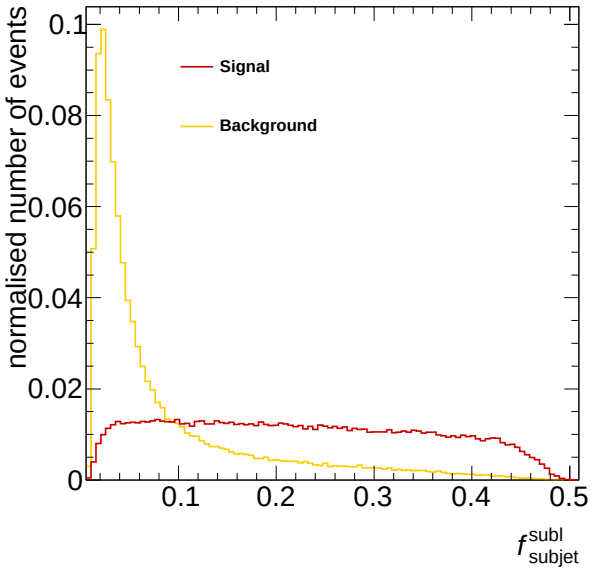


Figure 5.2: Normalised Distribution of $f^{\text{subl}}_{\text{subjet}}$. Yellow denotes QCD jet background, red di-tau signal.

Subjet Energy Fraction

E_{frac} is similar to the Subjet Energy Fraction defined above. They are introduced following similar considerations, namely that for QCD jets the leading subjet carries most of the energy, which is different for signal. Here the energy deposited in a subjet is divided by the energy deposited in the leading subjet:

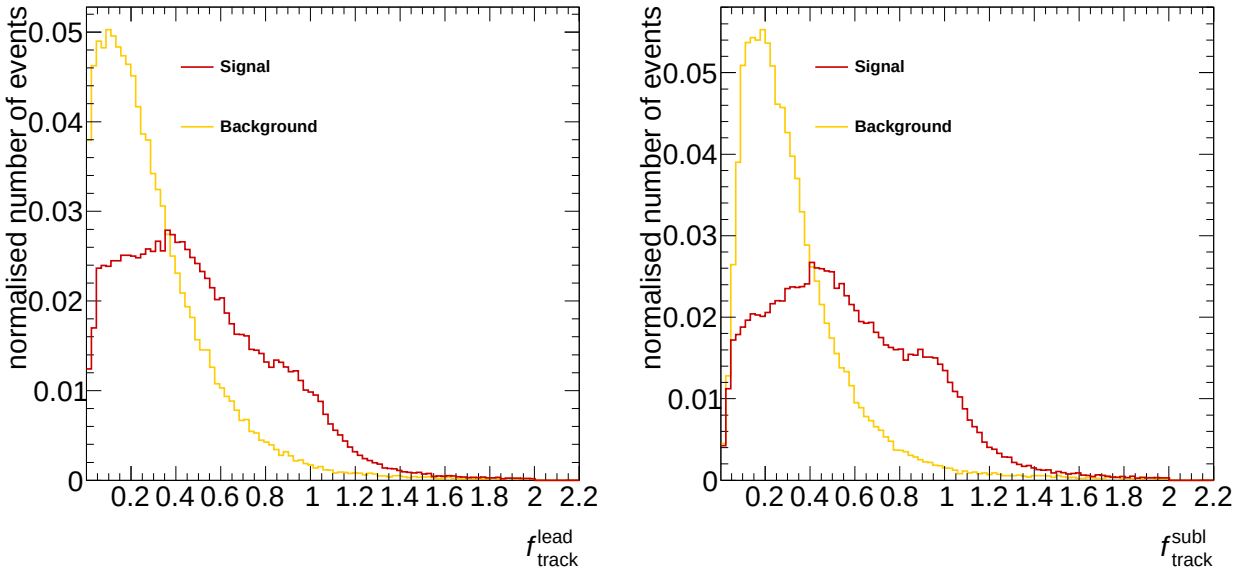
$$E_{\text{frac}} = \frac{E^{\text{subjet}}}{E^{\text{leading subjet}}}. \tag{5.4}$$

Therefore this is calculated only for the subleading and subsubleading (third) subjet. The distributions can be found in figure C.2.
Contrary to the QCD jet background the di-tau signal distribution only has a slight slope towards smaller values, whereas the background peaks at nearly zero for $E^{\text{subl}}_{\text{frac}}$. For the third subjet the distributions are similar for signal and background, they provide mostly the information whether there is a third subjet.

Leading Track Momentum Fraction

f_{track} is defined as the transverse momentum of the leading track inside a subjet divided by the transverse momentum of this subjet:

$$f_{\text{track}} = \frac{p_{\text{T}}^{\text{leading track}}}{p_{\text{T}}^{\text{subjet}}}. \tag{5.5}$$



(a) Distribution for the leading subjet.

(b) Distribution for the subleading subjet.

Figure 5.3: Normalised distributions of the f_{track} variables. Yellow represents QCD jet background, red di-tau signal.

In figure 5.3 it can be seen, that for background events it is highly unlikely that the p_{T} of one track makes up the whole p_{T} of a subjet whereas for signal events this is possible. Considering the decay modes of tau leptons this is understandable. If the tau decays into one pion (and one ν_τ) this pion has to carry most of the subjet's p_{T}.

Isolation Track Momentum Fraction

$f_{\text{isolation track}}$ is calculated by summing up all the transverse momenta of the isolation tracks in the jet and dividing it by the transverse momentum of the whole jet:

$$f_{\text{isolation track}} = \frac{\sum^{\text{jet}}_{\text{isolation tracks}} p_{\text{T}}}{p^{\text{jet}}_{\text{T}}}. \tag{5.6}$$

It shows therefore how much the isolation tracks contribute to the overall p_{T}.

As can be seen in figure 5.4 for background events the distribution is wider, meaning that isolation tracks contribute more to the overall p_{T} than for signal events. This is because the background events have generally more isolation tracks than signal events (see below).

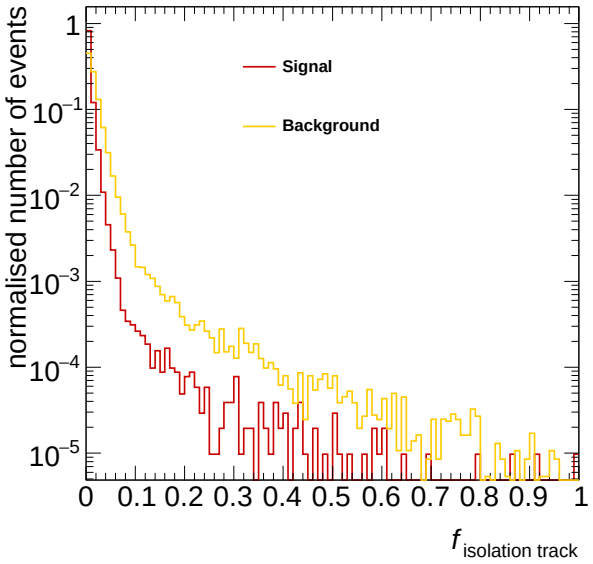


Figure 5.4: Normalised distributions of $f_{\text{isolation track}}$. Yellow represents QCD jet background, red di-tau signal.

Cluster Energy Fraction

f_{clusters} is defined as the energy of the clusters within an ellipse around the leading and subleading subjet but not those within the subjets divided by the energy of all clusters within the jet.

$$f_{\text{clusters}} = \frac{\sum_{\text{clusters}}^{\text{ellipse}} E}{\sum_{\text{clusters}}^{\text{jet}} E}. \tag{5.7}$$

Thereby the centres of two subjets are the two focal points of the ellipse. Here it is constructed in such way, that the sum of the distances of a point inside the ellipse to the focal points is always smaller than the distance between the two focal points plus two times the subjet radius 0.2. In that way it can be ensured that the two subjets are always enclosed by the ellipse. The clusters inside the two subjets were not used.

It can be seen that the distribution for background events is wider than the one for signal events. For signal events it is expected that there is little energy deposited in the space between two subjets because it is assumed that the decay products of the two tau leptons do not interact. For QCD jets it is expected that the object results from one jet, and therefore there is no real difference between a subleading subjet and the isolation region. This expectation is consistent with the observed distribution.

Maximum Track Distance

R_{max} is the maximal distance of a track, associated with a subjet, to the axis of the subjet. It

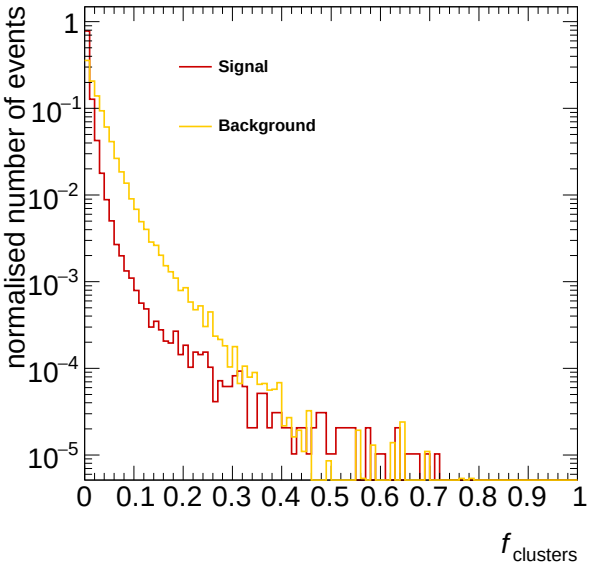


Figure 5.5: Normalised distributions of f_{clusters}. Yellow represents QCD jet background, red di-tau signal.

can again be defined for the leading and subleading subjet separately. While the jets of tau decays are expected to be collimated and so the maximal distance within a jet to be small, QCD jets are expected to cover a wider range and therefore the tracks are expected to have a larger distance from the subjet axis. Especially for the subleading subjet it is likely to find tracks at the rim of the subjet for QCD jets. For di-tau events this is not so as can be seen in figure C.3.

Weighted Track Distance

They are defined as the p_{T}-weighted sum of the track distances to their associated subjet axis:

$$R_{\text{track}} = \frac{\sum \Delta R \cdot p_{\text{T}}}{\sum p_{\text{T}}}. \tag{5.8}$$

Thereby the summation range differs for the implementations, in table 5.1 they are specified. These variables can be motivated analogue to the maximum track distance. For signal events one expects that the tracks are close to the axis of the subjet while for background events they are spread all over the subjet. Especially for the subleading subjet this behaviour can be observed. For the leading subjet this is not so obvious because the tracks with the highest p_{T} are closest to the center. But still the distribution for background is wider and its peak is shifted slightly to higher values. The distributions are displayed in figure C.4 and C.5.
Also, as shown in figure 5.6, it is more likely for background events to find isolation tracks close to a subjet whereas the probability to find isolation tracks does not depend much on the distance for signal events. That is because QCD jets are less collimated than tau decays.

Table 5.1: Different implementations of the Weighted Track Distance variable.

Variable	tracks used	maximum distance to the subjet axis
R_{track}	within leading and subleading subjet	0.2
$R_{\text{track}}^{\text{core}}$	within leading and subleading subjet	0.1
$R_{\text{track}}^{\text{all}}$	within all subjets	0.2
$R_{\text{isolation track}}$	isolation tracks outside the leading and subleading subjet	0.4
$R_{\text{tracks}}^{\text{lead}}$	within leading subjet	0.2
$R_{\text{tracks}}^{\text{subl}}$	within subleading subjet	0.2
$R_{\text{tracks}}^{\text{core lead}}$	within leading subjet	0.1
$R_{\text{tracks}}^{\text{core subl}}$	within subleading subjet	0.1

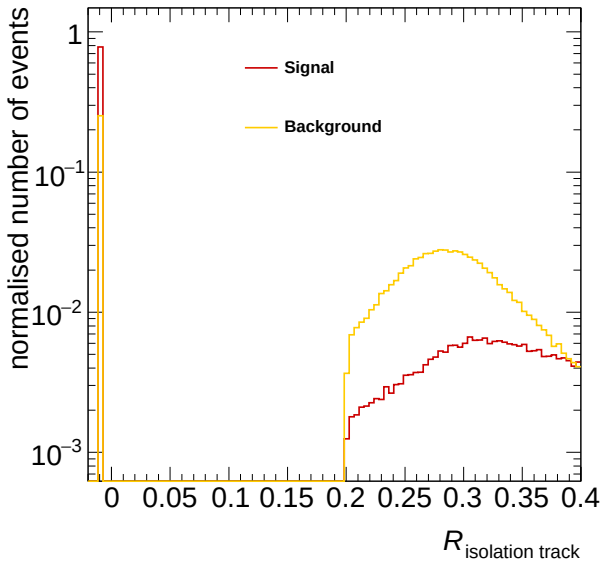


Figure 5.6: Normalised distributions of $R_{\text{isolation track}}$. Yellow represents QCD jet background, red di-tau signal.

The peak at the value -0.01 is the default value for those cases where there are no isolation tracks. This variable provides mostly the information, whether there are isolation tracks or not.

Subjet Distance

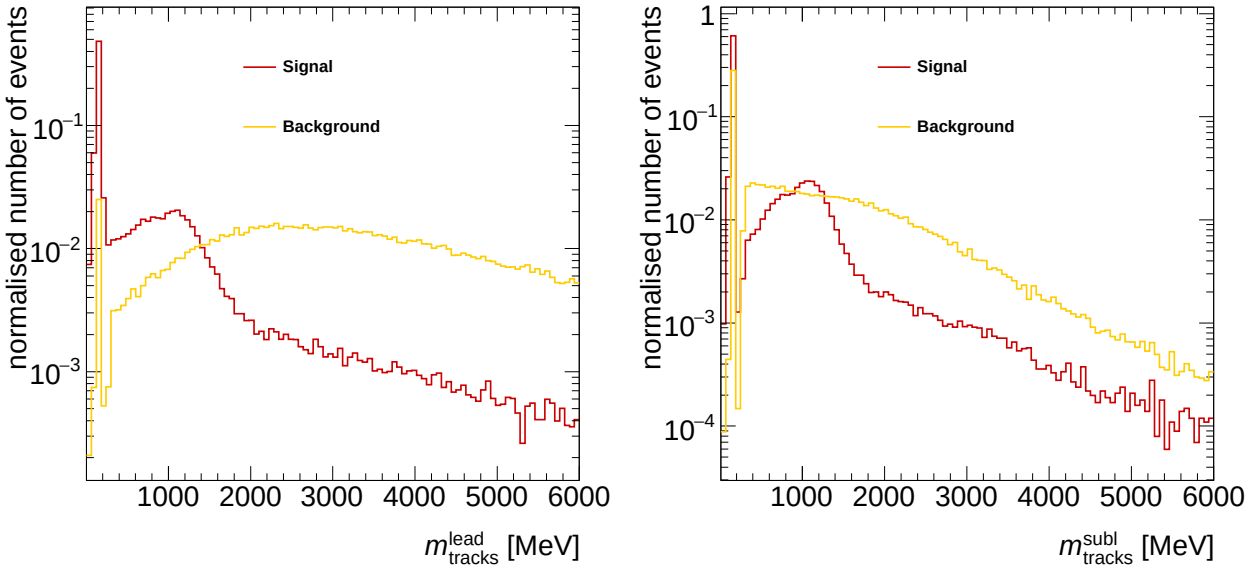
R_{subjets} is defined as the distance between a subjet and the leading subjet. Here it is calculated for the subleading subjet ($R_{\text{subjets}}^{\text{subl}}$) and the subsubleading subjet ($R_{\text{subjets}}^{\text{subsubl}}$).
It has to be considered that the distance of the subjets may depend on the physics of the process the particles are derived from. But in general it can be said that the distance between the leading and subleading subjet has a narrow peak while background events have a peak at small distances which then merges into a plateau. This can be seen in figure C.6.

Track Mass

m_{track} is the invariant mass of the sum of all tracks within a predefined region. There are different implementations of this variable, shown in table 5.2.

Table 5.2: Different implementations of the Track Mass variable.

Variable	region used
m_{track}	leading and subleading subjet
$m_{\text{track}}^{\text{core}}$	core region of the leading and subleading subjet
$m_{\text{track}}^{\text{all}}$	all subjets and all isolation tracks
$m_{\text{tracks}}^{\text{lead}}$	leading subjet
$m_{\text{tracks}}^{\text{subl}}$	subleading subjet
$m_{\text{tracks}}^{\text{core lead}}$	core region of the leading subjet
$m_{\text{tracks}}^{\text{core subl}}$	core region of the subleading subjet



(a) Distribution for the leading subjet.

(b) Distribution for the subleading subjet.

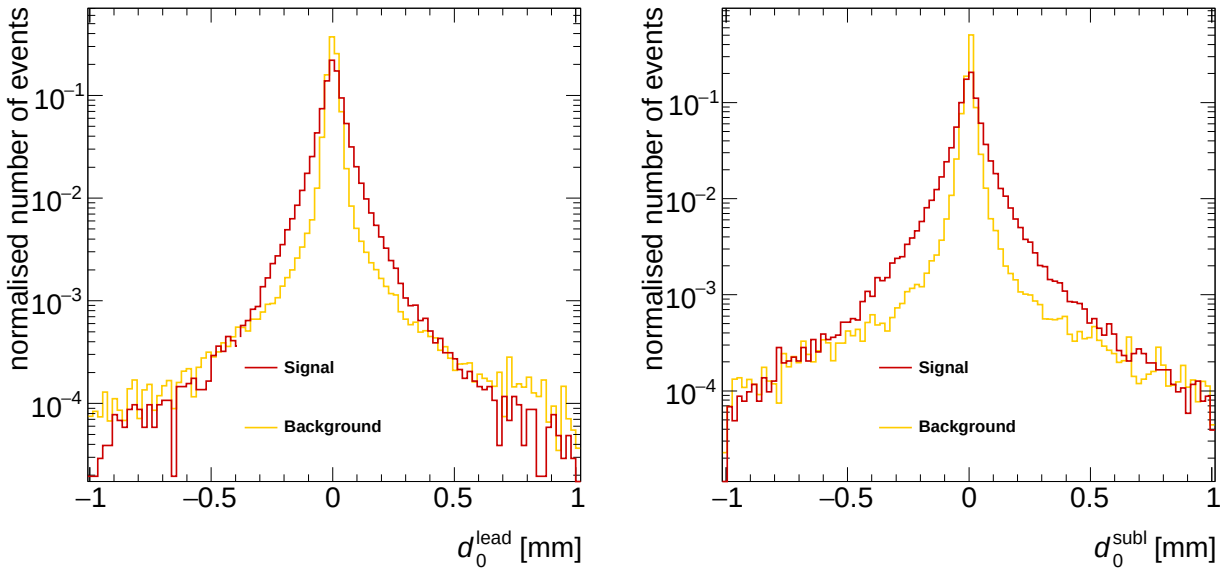
Figure 5.7: Normalised distributions of two m_{tracks} variables. Yellow represents QCD jet background, red di-tau signal.

As can be seen in figure 5.7 the distributions for signal in the leading and subleading subjet are almost the same but the distributions for background events differ. There the mean of the distribution is at higher values for the leading subjet. For the subleading subjet the distribution declines always. Because most of the energy is in the leading subjet it has a higher invariant mass than the subleading subjet which has almost none.
The first peak found in all distributions is at about 140 MeV which is the mass of a pion [34].

The peaks result from those cases where there is only one track in a subjet. The mass of one track is set by definition to the pion mass in the event reconstruction.
Further distributions can be seen in figure C.7. It is noticeable that the mass distributions of the signal events peak at higher values than that for background events for variables including more subjets. But those variable distributions may depend on the physics of the process, so they should be avoided.

Leading Track Impact Parameter

The impact parameter d_0 is the closest distance in the transverse plane of the track, which has the highest p_T within the core region of the subjet, to the primary vertex. This is an addition to the impact parameter cut. It is calculated for the leading and subleading subjet.
As shown in figure 5.8 the distributions for signal events are wider than those for background



(a) Distribution for the leading subjet.

(b) Distribution for the subleading subjet.

Figure 5.8: Normalised distributions of the d_0 variables. Yellow represents QCD jet background, red di-tau signal.

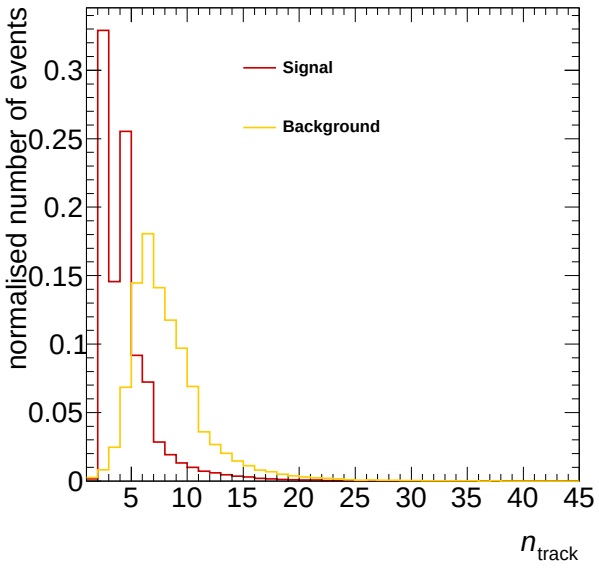
events in the leading as well as in the subleading subjet. That results from the fact that the tracks measured in the detector have their origin in the secondary vertex, the tau decay vertex, for di-tau events. So their distance to the primary vertex is higher. For QCD jets more tracks originate from the primary vertex.

Number of Tracks

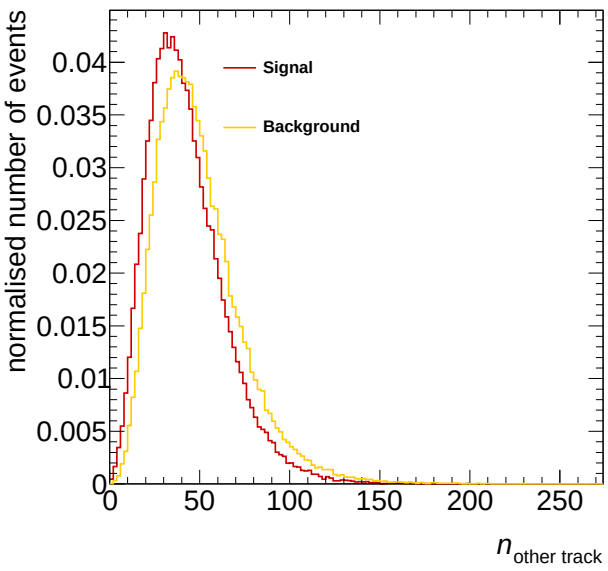
n_{track} counts the number of tracks. Thereby different tracks are included for the six implementations. These are shown in table 5.3.

Table 5.3: Different implementations of the number of tracks variable.

Variable	tracks used
n_{track}	tracks within all subjets
$n_{\text{othertrack}}$	all other tracks
$n_{\text{isolation track}}$	all isolation tracks
$n^{\text{ellipse}}_{\text{isolation tracks}}$	all isolation tracks within an ellipse around the first two subjets[1]
$n^{\text{lead}}_{\text{tracks}}$	all tracks within the leading subjet
$n^{\text{subl}}_{\text{tracks}}$	all tracks within the subleading subjet



(a) Distribution of n_{track}.



(b) Distribution of $n_{\text{othertrack}}$.

Figure 5.9: Normalised distributions of numbers of tracks. Yellow represents QCD jet background, red di-tau signal.

In the distribution for n_{track} (figure 5.9 (a)) one can see the decay modes of the tau leptons. In the signal distributions there are most of the time two or four tracks in all subjets which corresponds to 1 prong - 1 prong or 1 prong - multiprong (and vice versa) decays. Multiprong - multiprong decays are seldom (see table 7.1), consequently there is no third peak. In the background distribution no such features can be seen. This behaviour can also be observed regarding the leading and subleading subjet alone, which can be found in figure C.9. In figure 5.9 (b) it can be seen that background events usually have more other tracks than signal events. Also background events have more isolation tracks than signal events resulting from the fact that signal events consist of two jets with no interaction between those, background events consist of only one QCD jet. This can be found in figure C.8.

[1] The ellipse was calculated in the same way as for f_{clusters} defined above.

Number of Subjets

There are three implementations that count the number of subjets found in the jet. n_{Subjets} counts the number of subjets found by the reconstruction, where the requirement of a minimum of one track per subjet was applied. $n_{\text{Subjets}}^{\text{anti-}k_t}$ counts all subjets found by the anti-k_t algorithm, $n_{\text{Subjets}}^{\text{CA}}$ counts all subjets found by the Cambridge/Aachen algorithm.

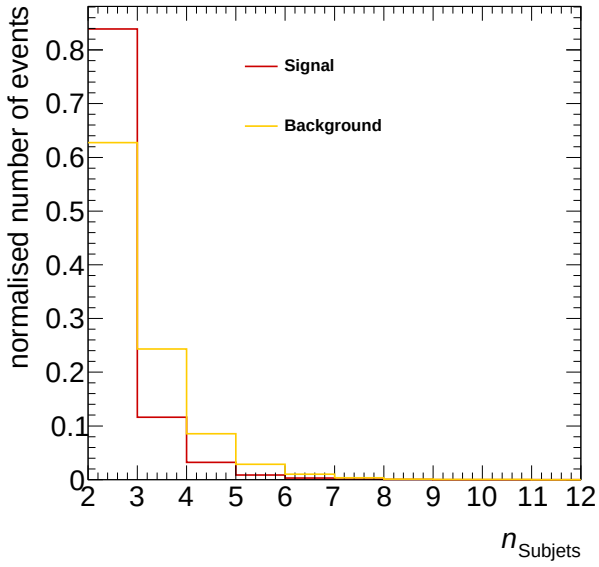


Figure 5.10: Normalised distributions of n_{Subjets}. Yellow represents QCD jet background, red di-tau signal.

In the background the reconstruction algorithm finds more subjets than in the signal. Because the signal jet should contain only the decay products of two tau leptons only two subjets should be found. That is different for QCD background, where any number of subjets could be found. This is conform with the distributions displayed in figure 5.10 and C.10.

Massdrop Tagger

The Massdrop variables were derived from the FastJet software package [38]. Thereby the last step of the clustering is undone, the jet is again divided into two parts. The part with the higher mass is labelled as j_1 and the part with the lower mass is labelled as j_2, j refers to the jet before the splitting. Then μ and y are:

$$\mu_{\text{massdrop}} = \frac{m_{j_1}}{m_j} \tag{5.9}$$

$$y_{\text{massdrop}} = \min(p_{\text{T}j_1}, p_{\text{T}j_2}) \frac{\Delta R^2_{j_1,j_2}}{m_j}, \tag{5.10}$$

Only if there is a sufficient massdrop $\mu < 0.9$ and at the same time the splitting sufficiently symmetric $y > 0.01$ the results are used, otherwise they are set to zero.

μ is the ratio of the mass of one part of the jet and the mass of the whole jet. For signal events one expects that there is a significant drop in mass since in that step the jet containing two tau leptons is again divided in two jets each containing one tau. For background events most of the energy is found in the leading subjet so by dividing the jet in two subjets no high drop in mass is expected. The distributions can be found in figure C.11.

6 Boosted Decision Trees

To effectively distinguish signal from background based on many variables a multivariate data analysis is needed. Here Boosted Decision Trees (BDTs) [39] by TMVA [40] are used, as it is common for such analyses.
Decision trees are binary structures where repeated yes/no decisions are made until a stop criterion is fulfilled. Before a decision tree can be applied it has to be trained. The events used for training should be independent of the events the tree is planned to operate on. The algorithm by TMVA divides the used event sample in two halfs, one for training and one for testing. This is done in a random way, the events sorted to the training or the test sample change every time. When training the tree, decisions are made in each step on the one variable, which has the highest discrimination power at that time. The events are then sent into one node or the other, depending on the output of the decision. The classification of the nodes is derived from the the ratio of signal to background training events in each node.
Boosted decision trees go a step further and extend the one tree to a forest. There events that were misclassified in the previous tree get an higher event weight in the following tree forcing the tree to concentrate more on these events so that they are classified correctly. In the end, the trees are combined into a single event classifying algorithm, which is obtained from the weighted average of the individual trees.
As an output each event gets a score depending on how it has been classified in the different trees. The scores range from -1 to 1, where events only classified as background get a score of -1 and events only classified as signal get a score of 1. When a cut is introduced in the BDT score, everything above that cut will be considered signal, everything below will be considered background. By checking the ratio of true signal events that have been classified as signal the signal efficiency for that cut can be calculated, respectively the background efficiency for true background events classified as signal.
A ROC-Curve (Receiver Operating Characteristic) can be derived from the BDT score. Usually the BDT score of a test sample, which is statistically independent of the training sample, is used. There for each signal efficiency the corresponding cut value in the BDT score is determined and for that cut value the associated background efficiency is calculated. This results in a dependency of the background efficiency on the signal efficiency. It is called a ROC-Curve. It has proven to be reasonable to plot the inverse background efficiency versus the signal efficiency, as it is done here.

Boosted decision trees are more stable against statistical fluctuations in the training sample and lead to a performance boost compared to the individual tree. If an event is misclassified in one tree it can still be correctly classified in other trees in the BDT. But for one single decision tree this is not possible. Events can get only a score of -1 or 1.
The BDT can be trained with different options. Here following options were used:

- 100 Trees were to train.
- The number or grid points within a variable range to find the optimal cut was set to 100.
- The separation type is Gini Index, the separation index is defined as $p(p-1)$, with p the purity of the node.
- The boosting type is AdaBoost [41] (short for adaptive boost), where the boost weight α for each event is calculated as

  $$\alpha = \frac{1 - \text{err}}{\text{err}}.$$

  err is the misclassification rate of an event in the previous tree. The learning process of the trees can be slowed down by introducing the parameter β, so that the boost weight is not α but α^β instead. Here β was chosen as 0.1.
- Nodes with a purity > 0.5 are considered signal, otherwise background.
- The first stop criterion is the minimal number of events in a node, which was here chosen as 2% of the number of all events.
- The second stop criterion is the maximal depth of each tree, which was here set to 20 layers.

7 Results

Through adjusting the BDT parameters and reducing the set of variables a BDT could be trained, which leads to a very good separation of signal and background events. With a signal efficiency of 80% an inverse background efficiency of more than 10^3 could be achieved, which means that while 80% of signal events are correctly classified less than 1 out of 1000 background events is misclassified (compare figure 7.1).

7.1 BDTs versus Simple Cut Algorithms

BDTs are very well suited for differentiating between signal and background events. They perform far better than the simpler cut algorithms, like the Rectangular Cut Optimisation by TMVA, which can be seen in figure 7.1. That is because simple cut algorithms only classify events as signal or background depending on their satisfaction or failure of the cut ensemble applied. BDTs, however, assign each event a score depending on how it has been classified in the different trees with different sets of cuts.

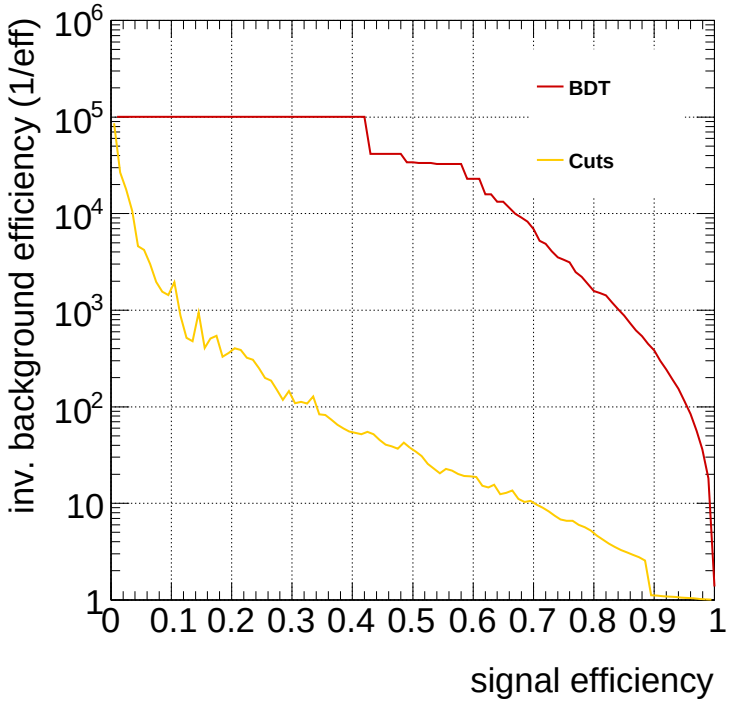


Figure 7.1: Comparison of the performance of rectangular cut optimisation and BDT through ROC-Curves.

7.2 Adjustment of the BDT Parameters

To find the best adjustment for the BDT parameters they are varied step by step. A higher tree number also enhances the performance because with more trees more events that are difficult to classify can be classified correctly. A high minimal node size and low maximal depth of the tree hinders the performance due to less possibilities within a tree to make cuts. By changing these two parameters it has always to be considered which is the predominant stop criterion. If the other criterion is changed in a range where it is still dominated by the first criterion there will be no effect. Changing the number of grid points to find the optimal cut has only a small effect in the tested range but generally it can be said that with a higher number of grid points tested the best cut is more likely to be found.
But by changing the options to get the optimum from the BDT overtraining has to be avoided. Overtraining is the effect of training a method too well on the given sample, letting it recognize all fluctuation within the sample. When it is used on a different sample many events are misclassified because they do not follow the same fluctuations. This can be avoided by lowering the complexity of the BDT, which leads to restrictions on the adjustments of the BDT parameters.
Further restrictions on the parameters come from minimizing statistical fluctuations. When training the same BDT with different random selected events statistical fluctuations can be observed around the mean in the ROC-curve like shown in figure 7.2. They appear due to the

statistical fluctuations within the samples, which results in different cuts being made every time. As can be seen in figure 7.3 the influence on the signal efficiency depending on the

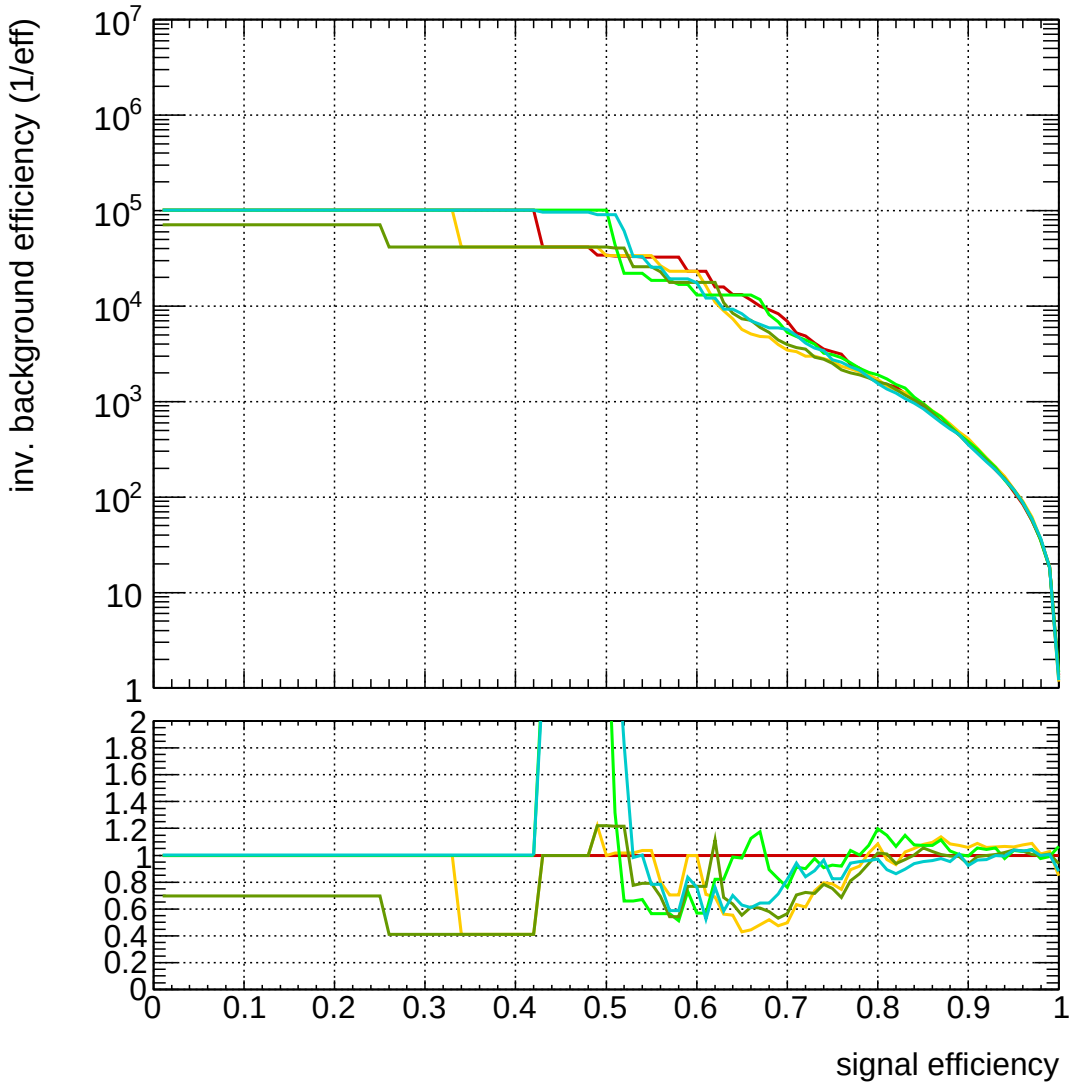
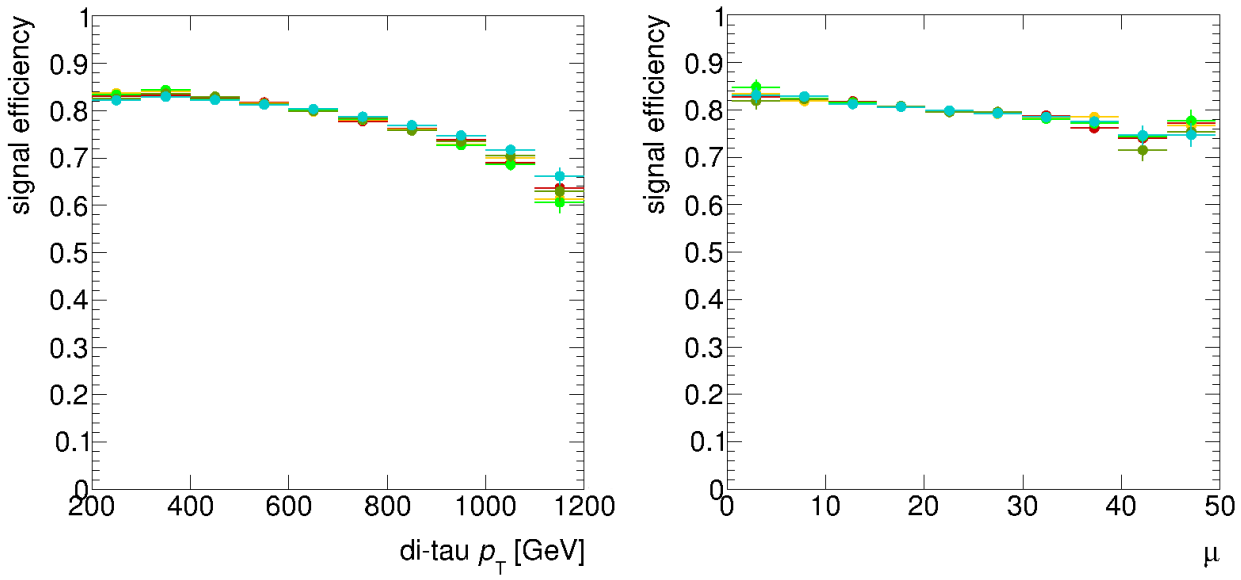


Figure 7.2: ROC-curves for the same training with different input samples. The parameters for training are those described in chapter 6, the set of variables is the final set, see below.

transverse momentum of the di-tau object or on the pileup is small. The fluctuations can be reduced by increasing the minimal node size.

So considering all these points the options are chosen as presented in chaper 6, leading to a BDT score as shown in figure 7.4. One can see that the results for training and test sample are in good agreement meaning that no overtraining occurred. The statistical fluctuations observed in the ROC-curves could be forced down to a maximum of 20% around the mean.



(a) Signal efficiency as a function of the di-tau p_T.

(b) Signal efficiency as a function of μ.

Figure 7.3: All Curves were trained identically, only the selection of the training events was varied every time.

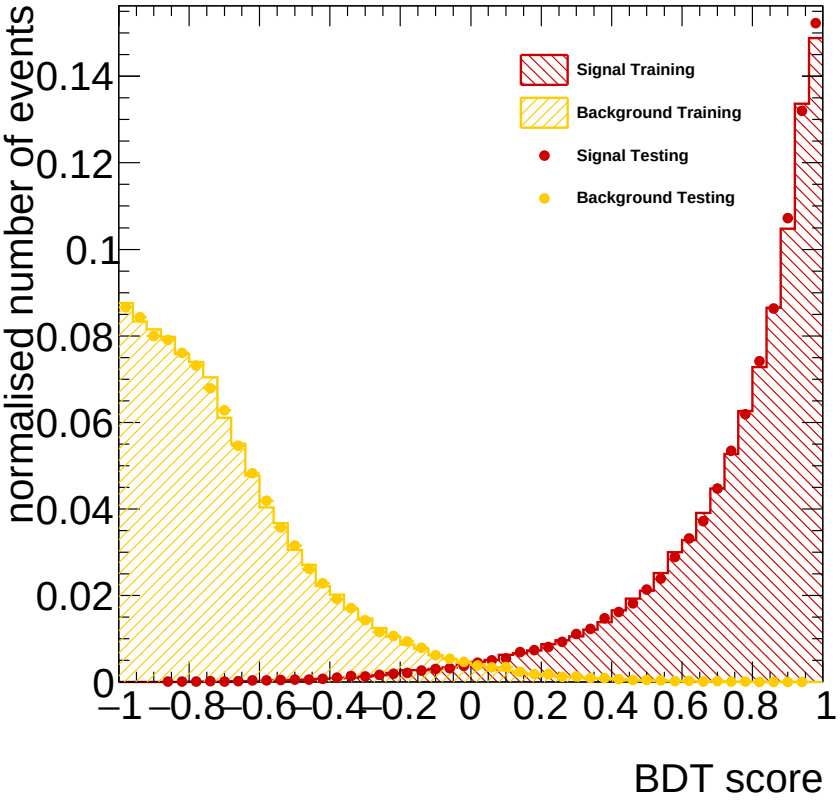


Figure 7.4: BDT score of the final training. Yellow represents QCD jet background, red di-tau signal.

7.3 Reduction of the ID Variables

To provide a simpler identification algorithm, where the efficiency does not depend on the di-tau p_{T} and the variables are not correlated the number of variables used for the analysis has to be reduced. A cut on $n_{\mathrm{tracks}}^{\mathrm{lead}}$ and $n_{\mathrm{tracks}}^{\mathrm{subl}}$ is applied, where both have to be smaller than 5. It has a signal efficiency of 97% and a background efficiency of 3%, those have to be multiplied with the efficiencies displayed here to derive the absolute value.
Furthermore it must be determined which variables perform best. This is done by giving all variables to the BDT. The effect a variable has is examined by removing this variable from the BDT. Depending on whether the BDT performs better, same or worse by comparing the ROC-Curves, the variables are left in the training or removed. The best training with 33 variables can be derived by removing all variables which hinder the performance of the BDT. The final training with 15 variables can be derived by removing all variables which do not influence the BDT, but correlations of the variables have to be taken into account. Also the dependency of the variables on the di-tau p_{T} has to be considered, which leads to the set of variables displayed in table 7.2 and 7.3. In figure 7.5 it can be seen that the final training with about a third of the 43 variables performs nearly as good as the one with all variables. The best training, which uses ten variables less than the original training performs as good as the one with 43 variables. The complete list of variables including the subsets for best training and final training can be found in table 7.2 and 7.3.

In figure 7.6 one can see that most correlated variables could be eliminated. The highest correlation is left between n_{track} and n_{subjets}, which is understandable because in each subjet tracks can be found, so a higher number of subjets results in a higher number of tracks.
Additionally the efficiency as a function of the di-tau p_{T} and the pileup was examined. It is aimed that the signal efficiency is as independent from these variables as possible. Flat distributions show that the ID variables are mostly independent from those two entities and therefore that this aim had been reached. As shown in figure 7.7 no completely flat distributions could be achieved. But it has to be considered, that for a p_{T} higher than 1200 GeV only few events are left (see Appendix A) so the distribution for points higher provide no statistical significance. Also for μ higher than 40 only small statistical significance can be provided. In the regions with high statistics the distributions are quite flat.

Taking a look at the background efficiencies depicted in figure 7.8 it can be seen that they are quite flat, too. Also they are at a quite low level for every working point and will not rise higher than 10^{-2}, corresponding to a maximum of 1 background event out of 100 being misclassified.

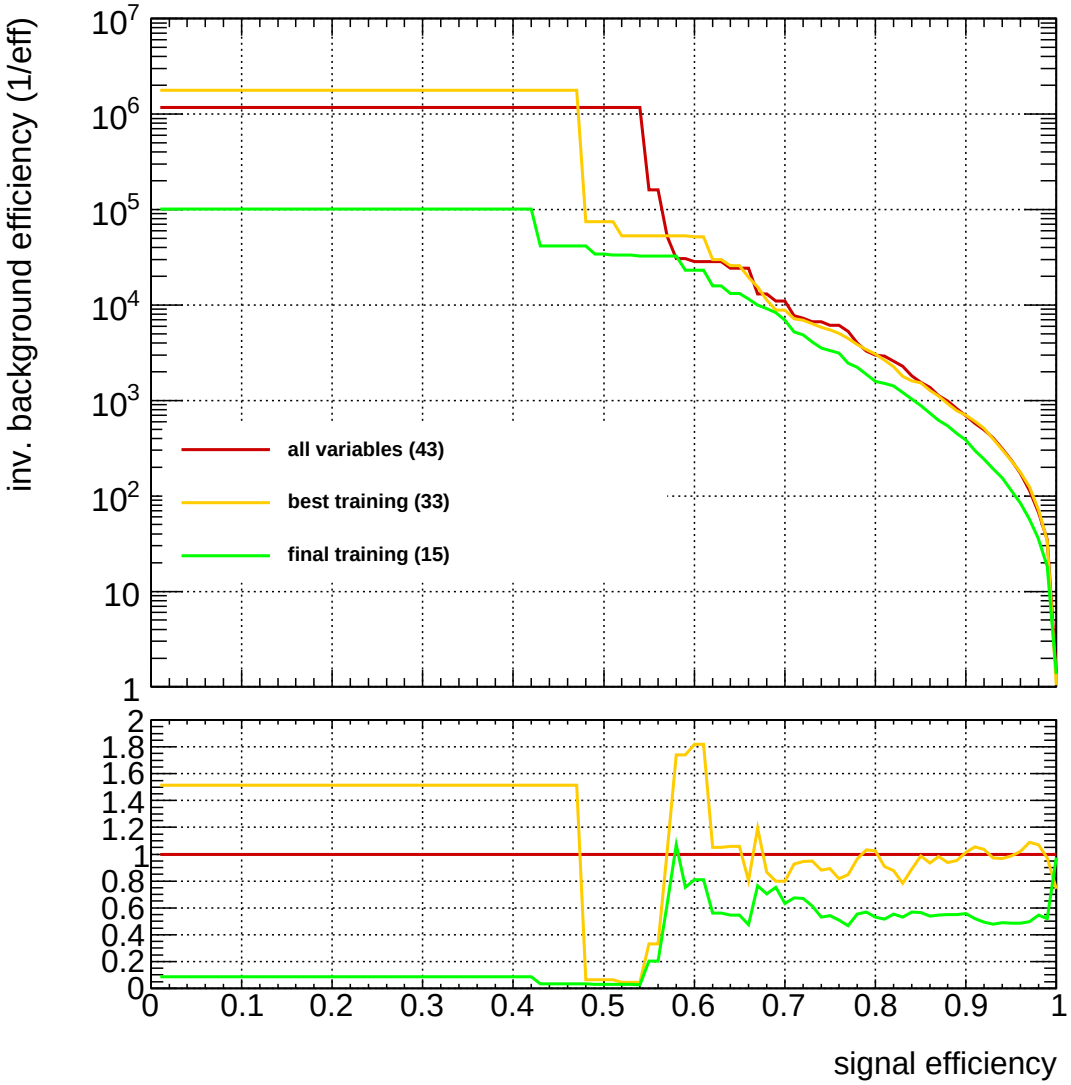
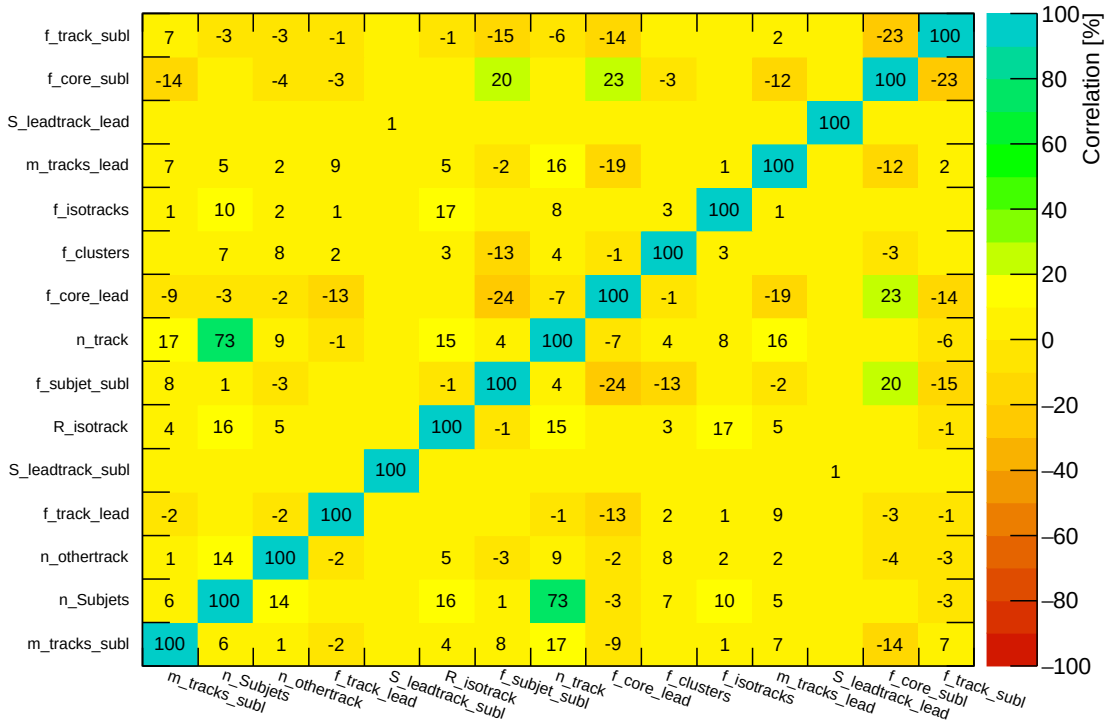
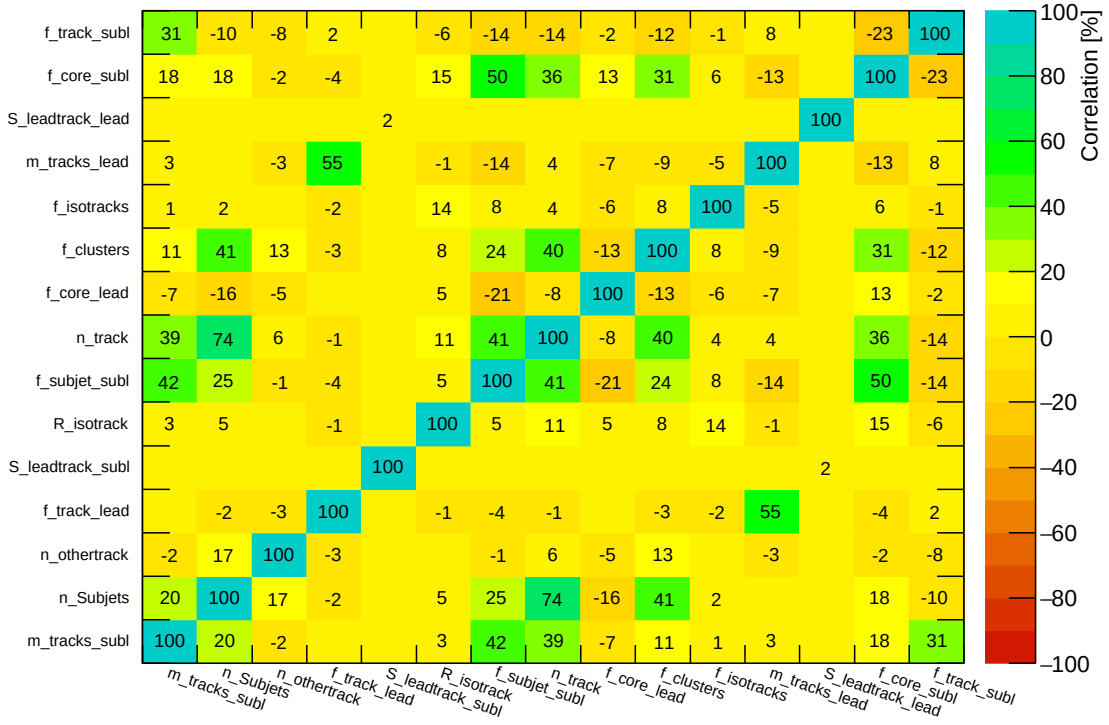


Figure 7.5: ROC-curves for the three different variable sets which are explained in table 7.2 and 7.3.

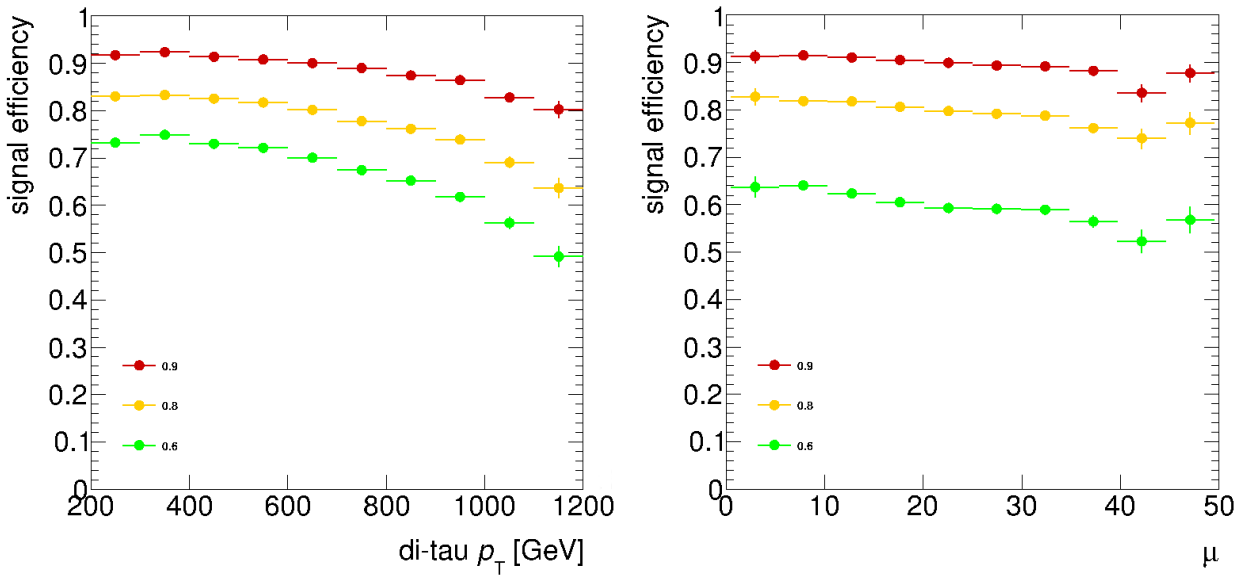


(a) Correlation matrix for signal events.



(b) Correlation matrix for background events.

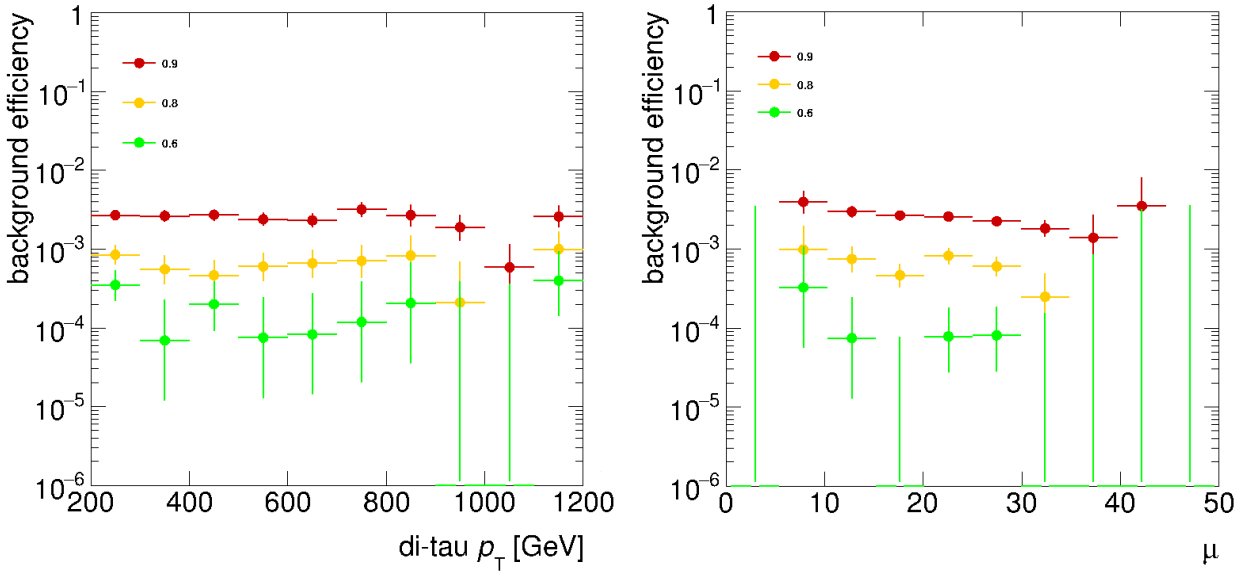
Figure 7.6: Correlation matrices for di-tau signal and QCD jet background.



(a) Signal efficiency as a function of the di-tau p_T.

(b) Signal efficiency as a function of the pile up μ.

Figure 7.7: The signal efficiency is shown for the three working points loose, medium and tight corresponding to signal efficiencies of 0.9, 0.8 and 0.6 respectively.



(a) Background efficiency as a function of di-tau p_T.

(b) Background efficiency as a function of the pile up μ.

Figure 7.8: The background efficiency is shown for the three working points loose, medium and tight corresponding to signal efficiencies of 0.9, 0.8 and 0.6 respectively.

7.4 Separation according to Decay Channels

Finally it is possible to train individual BDTs for 1 prong and 3 prong decays, as it is done in the single tau ID [8]. But in this case it is more complicated. Separating the sample in 1 prong and 3 prong decays does not lead to two BDTs to train but to four, one for 1 prong - 1 prong, one for 1 prong - multiprong, one for multiprong - 1 prong and one for multiprong - multiprong decays in the leading respectively subleading subjet. Just training separate trees for the different decays results in no performance boost of the BDT. Here all variables would have to be tested again, searching for the best. Also the training parameters have to be varied to get the optimal result for each decay mode. Furthermore for efficiently doing that one should take a step back and review the track reconstruction as there are many more multiprong - multiprong decays reconstructed as there should be and to few 1 prong - 1 prong decays as shown in table 7.1.

Table 7.1: Branching ratios considering a di-tau object if only hadronically decaying taus are examined.

Decay in leading subjet - subleading subjet	expected branching ratio	observed branching ratio in reconstruction
1 prong - 1 prong	0.59	0.36
1 prong - multiprong	0.18	0.20
multiprong - 1 prong	0.18	0.26
multiprong - multiprong	0.05	0.18

Table 7.2: All variables and where they have been used. To be continued in table 7.3.

Variable	best training	final training
$f_{\text{core}}^{\text{lead}}$	✓	✓
$f_{\text{core}}^{\text{subl}}$	✓	✓
$f_{\text{subjet}}^{\text{lead}}$	✓	
$f_{\text{subjet}}^{\text{subl}}$	✓	✓
f_{subjets}	✓	
$E_{\text{frac}}^{\text{subl}}$		
$E_{\text{frac}}^{\text{subsubl}}$		
$f_{\text{track}}^{\text{lead}}$	✓	✓
$f_{\text{track}}^{\text{subl}}$	✓	✓
$f_{\text{isolation tracks}}$	✓	✓
f_{clusters}	✓	✓
$R_{\text{max}}^{\text{lead}}$		
$R_{\text{max}}^{\text{subl}}$		
R_{track}	✓	
$R_{\text{track}}^{\text{core}}$	✓	
$R_{\text{track}}^{\text{all}}$	✓	
$R_{\text{isolation track}}$	✓	✓
$R_{\text{tracks}}^{\text{lead}}$		
$R_{\text{tracks}}^{\text{subl}}$		
$R_{\text{tracks}}^{\text{core lead}}$	✓	
$R_{\text{tracks}}^{\text{core subl}}$	✓	
$R_{\text{subjets}}^{\text{subl}}$	✓	
$R_{\text{subjets}}^{\text{subsubl}}$	✓	
m_{track}	✓	
$m_{\text{track}}^{\text{core}}$	✓	
$m_{\text{track}}^{\text{all}}$	✓	
$m_{\text{tracks}}^{\text{lead}}$	✓	✓
$m_{\text{tracks}}^{\text{subl}}$	✓	✓
$m_{\text{tracks}}^{\text{core lead}}$	✓	
$m_{\text{tracks}}^{\text{core subl}}$	✓	
$S_{\text{leadtrack}}^{\text{lead}}$	✓	✓
$S_{\text{leadtrack}}^{\text{subl}}$	✓	✓
n_{track}	✓	✓
$n_{\text{isolation track}}$	✓	
$n_{\text{isolation track}}^{\text{ellipse}}$	✓	
$n_{\text{other track}}$	✓	✓
$n_{\text{tracks}}^{\text{lead}}$		
$n_{\text{tracks}}^{\text{subl}}$		

Table 7.3: All variables and where they have been used. Continuation of table 7.2.

Variable	best training	final training
n_{Subjets}	✓	✓
$n_{\text{Subjets}}^{\text{anti-}k_t}$		
$n_{\text{Subjets}}^{\text{CA}}$		
μ_{massdrop}	✓	
y_{massdrop}	✓	

8 Summary and Outlook

In this thesis a working ID for hadronically decaying, boosted di-taus could be presented. It uses BDTs filled with 15 identification variables. Those were found through introducing and testing 43 variables, searching for a smaller set with firstly nearly as good performance when comparing the ROC-Curves, secondly reduced correlations between the variables, thirdly reduced dependencies of these variables on the di-tau p_{T} and finally nearly no dependencies of the ID efficiency from the pileup. This goal could be achieved. The parameters of the BDT training were varied so that the best performance could be found without overtraining it.
Furthermore the sensitivity of BDTs to statistical fluctuation in the training sample has been examined. It was found that also the ROC-Curve fluctuates with these variations in the training sample, the effect on the signal efficiency depending on p_{T} or μ is small.
At last it was examined whether there is a performance boost if the sample is split according to the decay modes (and therefore the number of tracks). But until now no improvement could be observed. More effort and time has to be put into evaluating the identification variables again and changing the options of the BDT algorithm for each of the four split samples to optimize the training. But it has to be kept in mind that by splitting the sample set less statistic is provided. To train on the same amount of events and therefore provide the same amount of statistics the samples have to be much larger.
The next step would be to test the ID on real data. The results for simulation and data have to be compared to check if the data was correctly modelled. This could be done by comparing the distributions of the identification variables for simulated events and experimental data.
Also it is important to test the ID with other types of background like $t\bar{t}$ decays. If it is not effective against other types of background further variables have to be introduced to differentiate between signal and background.

A Event samples

The samples used in this thesis are listed in table A.1. For the signal Monte Carlo 15 $G \rightarrow hh \rightarrow bb\tau\tau$ samples with Graviton masses of 1800 GeV, 2000 GeV and 2250 GeV are used. The event generation was done with MadGraph5_aMC@NLO [42] the showering with Pythia8 [43]. Monte Carlo 15 jets with a p_{T} of the leading jet between 400-800 GeV, 800-1300 GeV and 1300-1800 GeV were used for background. For generating the events as well as showering Pythia8 was used.
All events were reconstructed by the di-tau reconstruction [9]. Thereby event weights were inserted so that the di-tau p_{T} distribution of the background samples matches that of the signal samples. The weighted distribution can be seen in figure A.1.

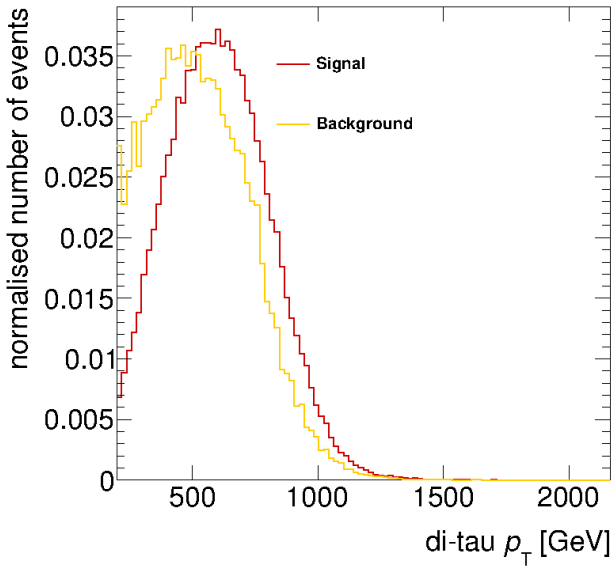


Figure A.1: Di-tau p_{T} distributions of the signal and background samples.

Table A.1: Monte Carlo Samples.

Signal
mc15_13TeV.303367.MadGraphPythia8EvtGen_A14NNPDF23LO_RS_G_hh_bbtt_hh_c10_M2000.recon.ESD.e4438_s2608_r6869/
mc15_13TeV.303366.MadGraphPythia8EvtGen_A14NNPDF23LO_RS_G_hh_bbtt_hh_c10_M1800.recon.ESD.e4438_s2608_r6869/
mc15_13TeV.303368.MadGraphPythia8EvtGen_A14NNPDF23LO_RS_G_hh_bbtt_hh_c10_M2250.recon.ESD.e4438_s2608_r6869/
Background
mc15_13TeV.361024.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ4W.recon.ESD.e3668_s2576_s2132_r6869/
mc15_13TeV.361025.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ5W.recon.ESD.e3668_s2576_s2132_r6869/
mc15_13TeV.361026.Pythia8EvtGen_A14NNPDF23LO_jetjet_JZ6W.recon.ESD.e3569_s2608_s2183_r6869/

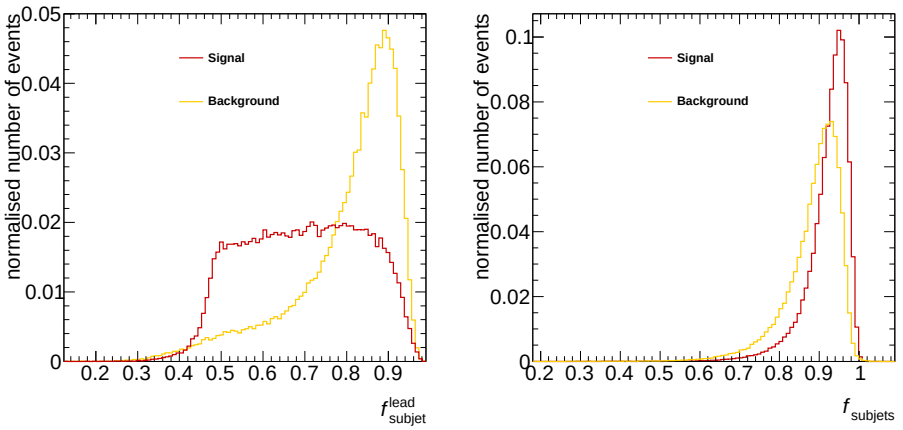
B Units

In this thesis natural units are used as it is common in high energy physics. Thereby the speed of light in vacuum, the reduced Planck constant and the Boltzmann constant were set to 1 ($c = \hbar = k_\mathrm{B} = 1$). As a result all other units of entities can be expressed in powers of the units of the energy. Using electronvolts or gigaelectronvolts as units for the energy proved handy.

$$1\,\mathrm{eV} = 1.6022 \cdot 10^{-19}\,\mathrm{J} \tag{B.1}$$

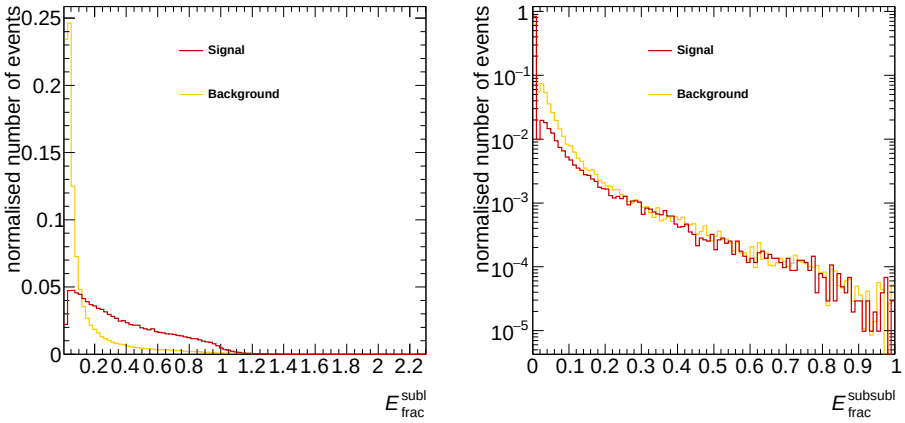
C Further variable distributions

Here the distributions of variables, which are not in the final training, are displayed.



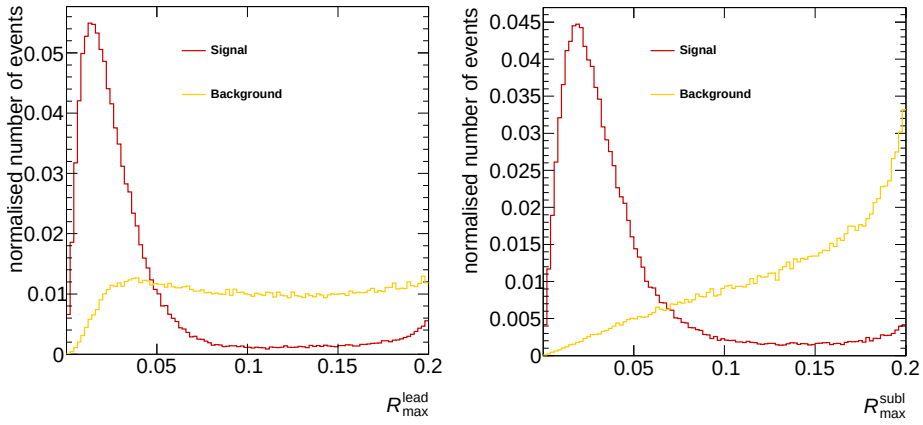
(a) Distribution for the leading subjet. (b) Distribution for the sum of leading and subleading subjet.

Figure C.1: Normalised distributions of the f_{subjet} variables. Yellow represents QCD jet background, red di-tau signal.



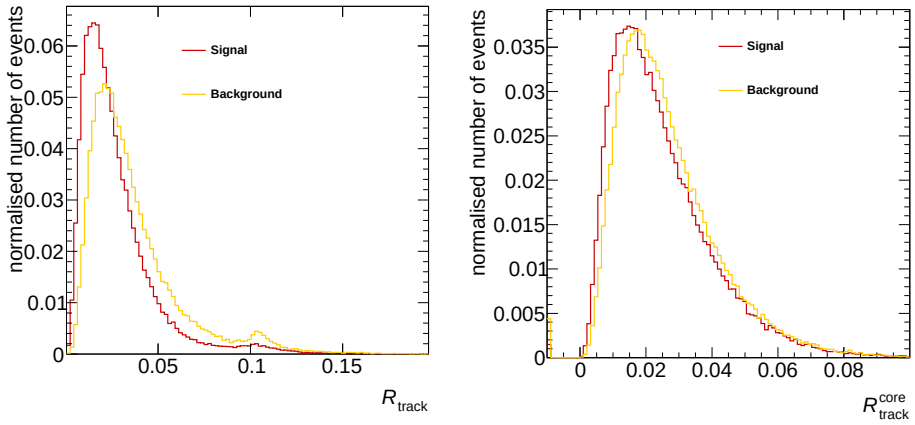
(a) Distribution for the subleading subjet. (b) Distribution for the subsubleading subjet.

Figure C.2: Normalised distributions of the E_{frac} variables. Yellow represents QCD jet background, red di-tau signal.

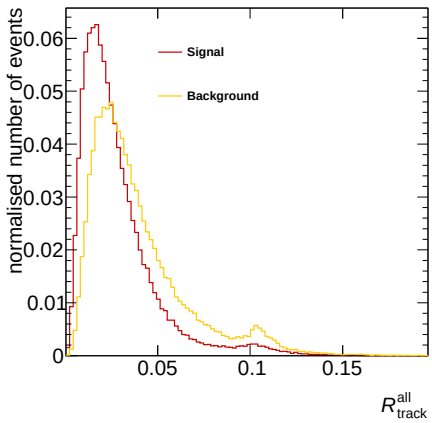


(a) Distribution for the leading subjet. (b) Distribution for the subleading subjet.

Figure C.3: Normalised distributions of the R_{max} variables. Yellow represents QCD jet background, red di-tau signal.

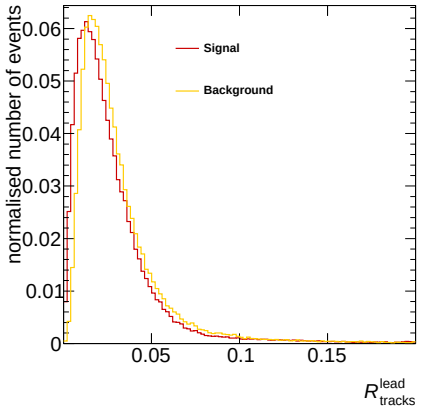


(a) Distribution for the leading and subleading subjet. (b) Distribution for the core region of leading and subleading subjet.

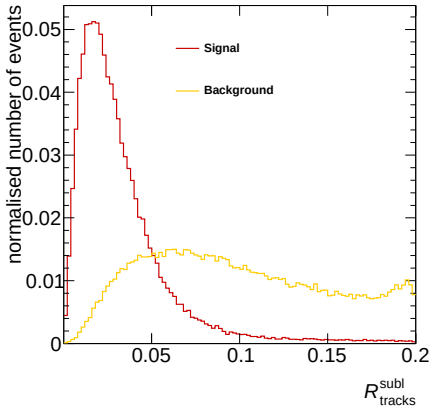


(c) Distribution for all subjets.

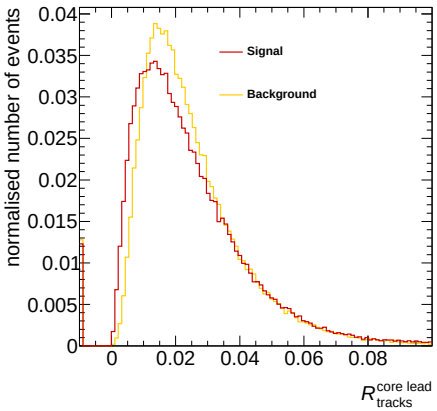
Figure C.4: Normalised distributions of the R_{track} variables. Yellow represents QCD jet background, red di-tau signal.



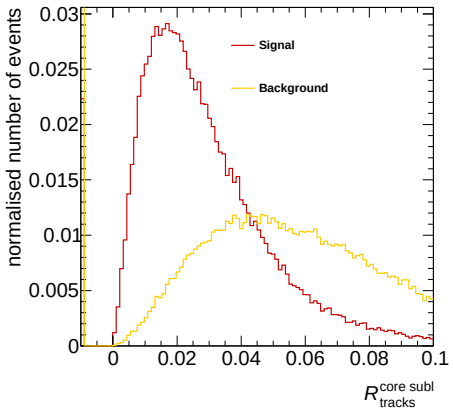
(a) Distribution for the leading subjet.



(b) Distribution for the subleading subjet.

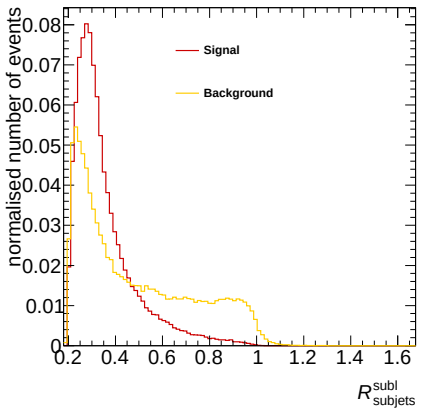


(c) Distribution for the core region of the leading subjet.

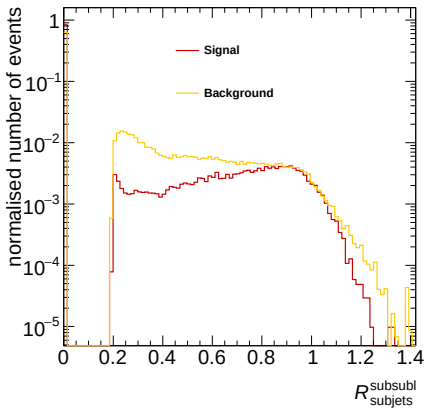


(d) Distribution for the core region of the subleading subjet.

Figure C.5: Normalised distributions of the R_{track} variables separate for leading and subleading subjet. Yellow represents QCD jet background, red di-tau signal.

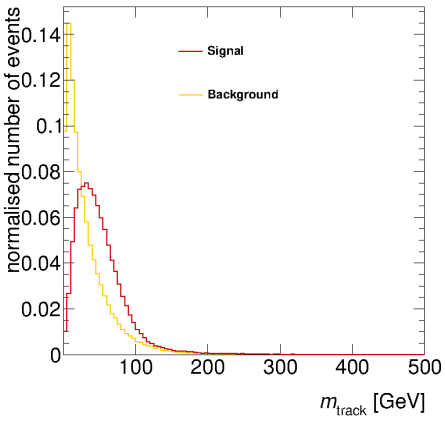


(a) Distribution for the subleading subjet.

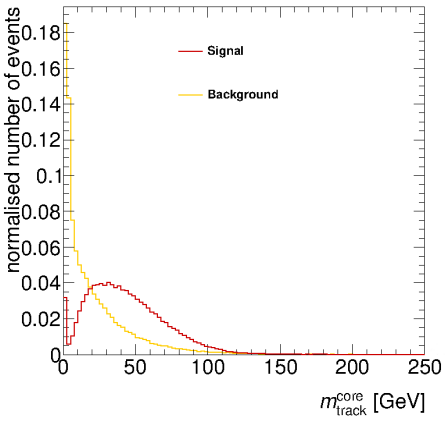


(b) Distribution for the subsubleading subjet.

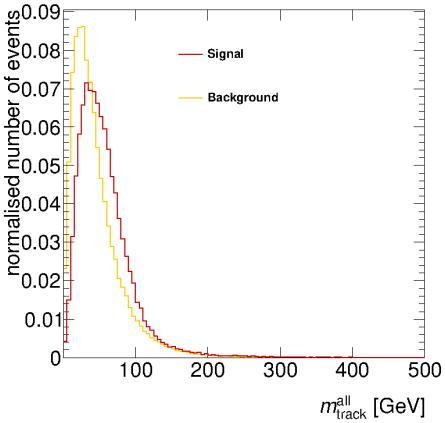
Figure C.6: Normalised distributions of the R_{subjets} variables. Yellow represents QCD jet background, red di-tau signal.



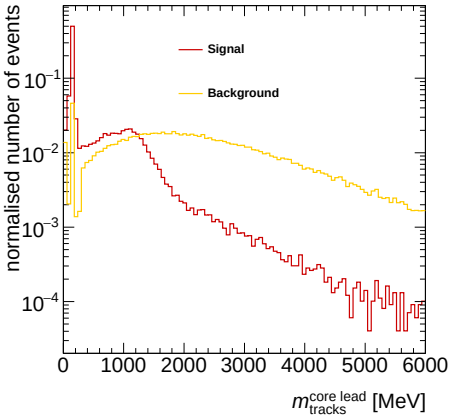
(a) Distribution for the leading and subleading subjet.



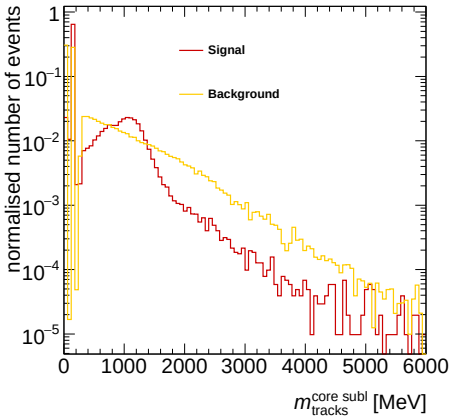
(b) Distribution for the core region of the leading and subleading subjet.



(c) Distribution for all tracks within subjets and all isolation tracks.

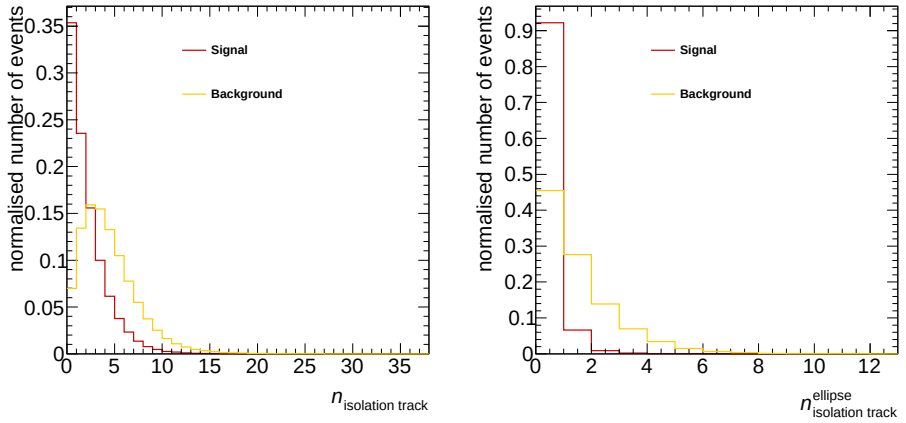


(d) Distribution for the core region of the leading subjet.



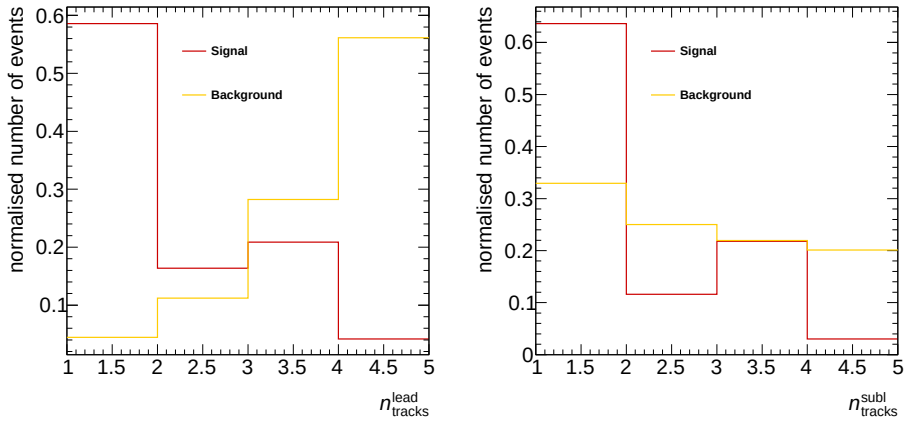
(e) Distribution for the core region of the subleading subjet.

Figure C.7: Normalised distributions of the m_{track} variables separate for leading and subleading subjet. Yellow represents QCD jet background, red di-tau signal.



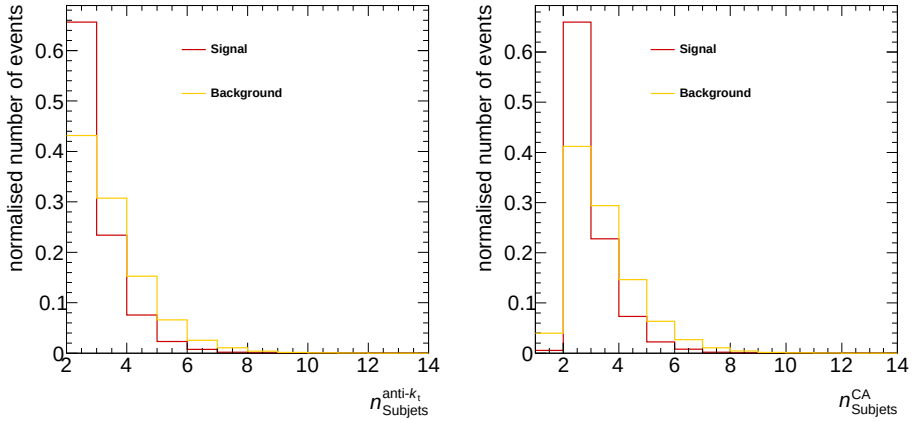
(a) Distribution for all isolation tracks. (b) Distribution for the isolation tracks within an ellipse around the first two subjets.

Figure C.8: Normalised distributions of the $n_{\text{isolation tracks}}$ variables. Yellow represents QCD jet background, red di-tau signal.



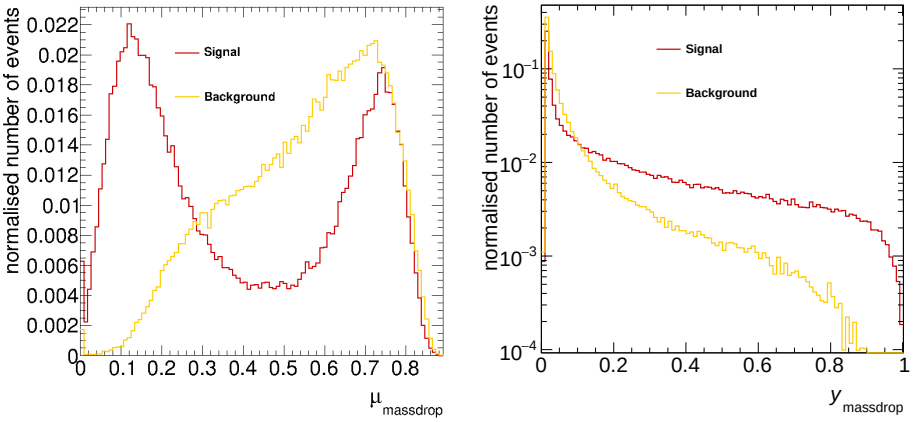
(a) Distribution for the leading subjet. (b) Distribution for the subleading subjet.

Figure C.9: Normalised distributions of the n_{tracks} variables separate for leading and subleading subjet. Yellow represents QCD jet background, red di-tau signal.



(a) Distribution for the anti-k_t subjets. (b) Distribution for the Cambridge Aachen subjets.

Figure C.10: Normalised distributions of the n_{subjets} variables without a track cut. Yellow represents QCD jet background, red di-tau signal.



(a) Distribution for μ_{massdrop}. (b) Distribution for y_{massdrop}.

Figure C.11: Normalised distributions of the massdrop variables. Yellow represents QCD jet background, red di-tau signal.

Bibliography

[1] The ATLAS Collaboration. The ATLAS Experiment at the CERN Large Hadron Collider. *Journal of Instrumentation*, 3(08):S08003, 2008.

[2] The CMS Collaboration. The CMS experiment at the CERN LHC. *Journal of Instrumentation*, 3(08):S08004, 2008.

[3] L. Evans and P. Bryant. LHC Machine. *Journal of Instrumentation*, 3(08):S08001, 2008.

[4] The ATLAS Collaboration. Search for resonances decaying to photon pairs in 3.2 fb^{-1} of pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. Technical Report ATLAS-CONF-2015-081, CERN, Geneva, 2015.

[5] The CMS Collaboration. Search for new physics in high mass diphoton events in proton-proton collisions at $\sqrt{s} = 13$ TeV. Technical Report CMS-PAS-EXO-15-004, CERN, Geneva, 2015.

[6] The ATLAS Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys.Lett. B716 (2012), arXiv:1207.7214v2 [hep-ex]*, 2012.

[7] The CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys. Lett. B 716 (2012) 30, arXiv:1207.7235v2 [hep-ex]*, 2012.

[8] The ATLAS Collaboration. Identification and energy calibration of hadronically decaying tau leptons with the ATLAS experiment in pp collisions at $\sqrt{s} = 8\,\text{TeV}$. *Eur. Phys. J. C75 (2015) 303, arXiv:1412.7086v2 [hep-ex]*, 2015.

[9] D. Kirchmeier. Reconstruction and Identification of Boosted Tau Pair Topologies at ATLAS. Master's thesis, Technische Universität Dresden, 2015.

[10] S. L. Glashow. Partial-Symmetries of Weak Interaction. *Nuclear Physics* **22**, 1961.

[11] S. Weinberg. A model of leptons. *Phys. Rev. Lett.*, 19:1264–1266, Nov 1967.

[12] A. Salam. Weak and Electromagnetic Interactions. *Nobel Symposium 8, Elementary Particle Theory: Relativistic Groups and Analyticaly*, 1968.

[13] G. t'Hooft. Renormalizable Lagrangians for Massive Yang-Mills Fields. *Nuclear Physics* **B 35**, pages 167–188, 1971.

[14] G. t'Hooft and M. Veltman. Regularization and Renormalization of Gauge Fields. *Nuclear Physics* **B 44**, pages 189–213, 1972.

[15] H. Fritzsch, M. Gell-Mann, and H. Leutwyler. Advantages of the Color Octet Gluon Picture. *Physics Letters* **47 B**, 1973.

[16] H. David Politzer. Reliable Perturbative Results for Strong Interactions? *Phys. Rev. Lett.*, 30:1346–1349, 1973.

[17] David J. Gross and Frank Wilczek. Ultraviolet behavior of non-abelian gauge theories. *Phys. Rev. Lett.*, 30:1343–1346, 1973.

[18] Steven Weinberg. Non-abelian gauge theories of the strong interactions. *Phys. Rev. Lett.*, 31:494–497, 1973.

[19] M. Thomson. *Modern Particle Physics*. Cambridge University Press, 2015.

[20] MissMJ. Standard model of elementary particles. `https://commons.wikimedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg`, Copy licence `http://creativecommons.org/licenses/by/3.0/legalcode`, 2014, call date 26.05.16.

[21] F. Englert and R. Brout. Broken symmetry and the mass of gauge vector mesons. *Phys. Rev. Lett.*, 13:321–323, 1964.

[22] Peter W. Higgs. Broken symmetries, massless particles and gauge fields. *Phys. Lett.*, 12:132–133, 1964.

[23] Peter W. Higgs. Broken symmetries and the masses of gauge bosons. *Phys. Rev. Lett.*, 13:508–509, 1964.

[24] Peter W. Higgs. Spontaneous symmetry breakdown without massless bosons. *Phys. Rev.*, 145:1156–1163, 1966.

[25] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble. Global Conservation Laws and Massless Particles. *Phys. Rev. Lett.*, 13:585–587, 1964.

[26] Howard Georgi and S. L. Glashow. Unity of all elementary-particle forces. *Phys. Rev. Lett.*, 32:438–441, 1974.

[27] D. N. Spergel et al. First year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Determination of cosmological parameters. *Astrophys. J. Suppl.*, 148:175–194, 2003.

[28] A. Djouadi. The Anatomy of Electro-Weak Symmetry Breaking. II: The Higgs bosons in the Minimal Supersymmetric Model. *Phys.Rept.459:1-241,2008, arXiv:hep-ph/0503173v2*, 2005.

[29] *LEP design report.* CERN, Geneva, 1984. Copies shelved as reports in LEP, PS and SPS libraries.

[30] CERN. The accelerator complex. `http://home.cern/about/accelerators`, 2016, call date 31.05.2016.

[31] C. Lefèvre. The CERN accelerator complex. Complexe des accélérateurs du CERN. 2008.

[32] The ALICE Collaboration. The ALICE experiment at the CERN LHC. *Journal of Instrumentation*, 3(08):S08002, 2008.

[33] The LHCb Collaboration. The LHCb Detector at the LHC. *Journal of Instrumentation*, 3(08):S08005, 2008.

[34] K.A. Olive et al.(Particle Data Group). 2015 Review of Particle Physics. `http://pdglive.lbl.gov/Viewer.action`, 090001 (2014) and 2015 update, call date 01.06.2016.

[35] M. Cacciari, G. P. Salam, and G. Soyez. The anti-k_t jet clustering algorithm. *JHEP 0804:063,2008, arXiv:0802.1189v2 [hep-ph]*, 2008.

[36] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber. Better jet clustering algorithms. *JHEP*, 08:001, 1997.

[37] S. Catani, Y. L. Dokshitzer, M. H. Seymour, and B. R. Webber. Longitudinally invariant K_t clustering algorithms for hadron hadron collisions. *Nucl. Phys.*, B406:187–224, 1993.

[38] M. Cacciari, G. P. Salam, and G. Soyez. FastJet user manual. *Eur. Phys. J. C* **72** *(2012) 1896, arXiv:1111.6097 [hep-ph]*, 2012.

[39] Y. Freund and R. Schapire. Experiments with a New Boosting Algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference*, 1996.

[40] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, and H. Voss. TMVA: Toolkit for Multivariate Data Analysis. *PoS*, ACAT:040, 2007.

[41] Y. Freund and R. Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(SS971504):119–139, 1997.

[42] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *JHEP07(2014)079, arXiv:1405.0301v2 [hep-ph]*, 2014.

[43] T. Sjöstrand, S. Mrenna, and P. Z. Skands. A Brief Introduction to PYTHIA 8.1. *Comput. Phys. Commun. 178 (2008) 852, arXiv: 0710.3820 [hep-ph]*, 2008.

List of Figures

List of Tables

Erklärung

Hiermit erkläre ich, dass ich diese Arbeit im Rahmen der Betreuung am Institut für Kern- und Teilchenphysik ohne unzulässige Hilfe Dritter verfasst und alle Quellen als solche gekennzeichnet habe.

Franziska Schoger
Dresden, Juni 2016