Studies of top tagging identification methods and development of a new heavy object tagger

Dissertation

zur Erlangung des Doktorgrades

des Fachbereichs Physik

der Universität Hamburg

vorgelegt von

Tobias Lapsien

aus Dortmund

Hamburg

2016

Gutachter der Dissertation:	Prof. Dr. Johannes Haller Dr. Andreas Meyer
Gutachter der Disputation:	Prof. Dr. Erika Garutti Prof. Dr. Günter Sigl Dr. Alexander Schmidt
Datum der Disputation:	18. April 2016

Abstract

At the Large Hadron Collider (LHC), precision tests of the standard model of particle physics and searches for new phenomena are performed. To make optimal use of the proton-proton collisions delivered by the LHC and its increasing collision rate, both the detectors and the reconstruction algorithms have to be optimized. The identification of heavy quarks is a key component in many measurements. This thesis describes a hardware and a software project which both aim at improving the identification of heavy quarks.

In the first part of this thesis, the Phase 1 upgrade of the CMS pixel detector is introduced. One of the main motivations of the replacement of the Pixel detector is the improved b jet identification at large collision rates. The Phase 1 upgrade involves several production and calibration steps. An x-ray calibration procedure has been developed and the corresponding experimental setup is described. Measurements show that the calibration of the pixel modules is temperature independent and can be performed at room temperature. The stability of the setup is tested in order to fulfill the requirements for mass production of the pixel modules. A method to stabilize the calibration is introduced which is shown to reduce the systematic uncertainty.

In the second part, algorithms to identify heavily boosted top quarks ("top tagger") are described and their performance is compared. The OptimalR HEP top tagger and the shower deconstruction tagger show a better performance than existing tagging algorithms. They can be used in Run II with increased centre-of-mass energies of 13 and 14 TeV. It is also shown that existing top tagging algorithms can be improved by the usage of multivariate analysis methods. New algorithms are commissioned using CMS data with a centre-of-mass energy of 8 TeV, corresponding to an integrated luminosity of $19.7\,\mathrm{fb}^{-1}$. In order to validate these new algorithms in data, two selections are made to measure the efficiency and mistag rate. The selections aim at enriching one sample with $\mathrm{t\bar{t}}$ events and one with QCD multijet events.

In the last part of this thesis, a novel top tagging algorithm is introduced. It is capable of identifying top quarks over a large p_T range by using a p_T dependent jet size and a method to reconstruct subjets based on a mass jump criterion. The algorithm includes the possibility to remove soft and collinear radiation from the jet. The algorithm is infrared and collinear safe, has a reasonable computational time, and shows a good performance with respect to existing algorithms. The performance of the new algorithm is also investigated on 8 TeV data.

Kurzfassung

Am Large Hadron Collider (LHC) werden Präzisionsmessungen des Standardmodells der Teilchenphysik und Suchen nach neuer Physik durchgeführt. Um die Proton-Proton Kollisionen des LHC und die erhöhte Kollisionsrate optimal nutzen zu können, müssen Detektorkomponenten und Rekonstruktionsalgorithmen optimiert werden. Die Identifikation von schweren Quarks ist eine Schlüsselkompente für viele Messungen. Diese Arbeit beschreibt ein Hardware-und ein Softwareprojekt, welche darauf zielen, die Identifikation von schweren Quarks zu verbessern.

Im ersten Teil dieser Arbeit wird das Phase-1-Upgrade des CMS Detektors vorgestellt. Eine der Hauptmotivationen für den Neubau des Pixel-Detektors ist die daraus resultierende verbesserte b-Quark-Identifikation bei hohen Kollisionsraten. Das Phase-1-Upgrade umfasst verschiedene Produktions-und Kalibrationsprozesse. Eine Röntgen-Kalibration wurde entwickelt und der dazugehörige Versuchsaufbau wird beschrieben. Messungen zeigen, dass die Kalibration der Pixelmodule unabhängig von der Temperatur ist und dass diese bei Raumtemperatur durchgeführt werden kann. Die Stabilität des Versuchaufbaus wird getestet, um die Anforderungen für eine Massenproduktion der Pixelmodule zu erfüllen. Eine Methode zur Stabilisierung der Kalibration wird erläutert, welche die systematischen Unsicherheiten deutlich verringert.

Im zweiten Teil dieser Arbeit werden neue Algorithmen zur Identifikation von Top-Quarks ("Top-Tagger") beschrieben und deren Leistungsfähigkeit verglichen. Die Leistungsfähigkeit des OptimalR-HEP-Top-Taggers und die des Shower-Deconstruction-Taggers übertrifft die Leistungsfähigkeit von existierenden Top-Tagging-Algorithmen. Diese Algorithmen können in Run II des LHCs mit einer erhöhten Schwerpunktsenergie von 13 bzw. 14 TeV benutzt werden. Weiterhin wird gezeigt, dass eine Verbesserung des Top-Tagging durch die Benutzung von mulivariaten Methoden und der Benutzung von Substruktur-Variablen möglich ist. CMS-Daten mit einer Schwerpunktsenergie von 8 TeV und einer integrierten Luminosität von $19.7\,\mathrm{fb}^{-1}$ werden benutzt, um diese neuen Algorithmen zu valideren. Dafür werden zwei Selektionen zur Messung der Effizienz und der Fehlidentifikatinsrate benutzt. Diese beiden Selektionen reichern je einen Datensatz mit $t\bar{t}$-Ereignissen und einen anderen mit QCD Multijet Ereignissen an.

Im letzten Teil dieser Arbeit wird ein neuer Top-Tagging-Algorithms beschrieben, welcher imstande ist, Top-Quarks über einen großen p_{T} Bereich mit Hilfe eines p_{T}-abhängigen Jet-Distanz-Parameters zu identifizieren. Dieser Algorithmus benutzt eine Methode zur Rekonstruktion von Subjets, die auf einem Massensprung Kriterium basiert. Weiche Abstrahlungen werden zusätzlich während des Cluster-Prozesses entfernt. Der Algorithmus ist infrarot-und kollinear sicher, benötigt wenig Rechenzeit und zeigt eine gute Leistungsfähigkeit im Vergleich zu bereits existierenden Algorithmen. Die Leistungsfähigkeit wird auch in 8 TeV-Daten untersucht.

Contents

1 Introduction

What are we made of? Where do we come from? Why do we exist? How did everything start? These are questions that humans asked for centuries. The idea that matter consists of discrete units, not divisible further, is very old and appeared already in the ancient Greek culture around the 4^{th} century BC. Hundreds of years later, in the 1800s John Dalton used the concept of atoms in order to explain that elements always interact in ratios of integer numbers [1]. More than 100 years later, in 1906 J. J. Thomson received the Nobel prize for the discovery of a subatomic particle, the electron [2]. A little later, 1909, the nucleus was observed by Hans Geiger and Ernest Marsden [3] in the so-called Rutherford experiment and 1913 Niels Bohr proposed a model in which the electron circulates around the nucleus [4].

In 1919 Ernest Rutherford found that the nucleus consists of further particles and is not an elementary particle. In 1935 James Chadwick got the Nobel prize for the discovery of the neutron [5]. The same time, Albert Einstein predicted other existence of the photon with the photoelectric effect and was awarded the Nobel prize in 1922 [6, 7]. At this point, the world was described by four particles proton, neutron, electron and photon.

After 1947, new particles were discovered which were called Mesons. Together with the baryons it was found that these particles consists of more fundamental particles, the quarks. In 1968, evidence of their physical existence was provided by the Stanford Linear Accelerator Center in the USA [8]. The gauge bosons W and Z were discovered at CERN in Switzerland 1983 [9, 10]. Later the most heavy quark, the top quark was discovered 1995 at Fermilab in the USA [11].

All these fundamental particles and their interaction are described by the standard model of particle physics. In 1999, Gerard 't Hooft and Martinus Veltman received the Nobel prize for the renormalization of the electroweak interaction [12]. Another Nobel prize was given to Frank Wilcek for the discovery of the asymptoic freedom of quarks [13]. And in 2008, Makoto Kobayashi was awarded the Nobel prize for the discovery of the origin of the broken symmetry which predicts the existence of at least three families of quarks in nature [14]. However, more then 50 years ago, the Higgs mechanism was proposed to give mass to the gauge bosons and to fermions. It predicts the existence of an additional particle, the Higgs boson. Finally it was discovered in 2012 by the experiments ATLAS and CMS at the LHC in Switzerland [15]. Peter Higgs and Francois Englert got the Nobel prize in 2013 for the prediction of the Higgs particle. The discovery of the Higgs boson completed the standard model.

In summary, several Nobel prizes were given to particle physics. Over the last decades particle physics changed our way of seeing the world, which makes this fields one of the most interesting fields in physics.

However, the standard model is not able to describe all phenomena that are observed in nature. Different models were developed in order to explain these effects and the LHC, the world's biggest particle accelerator, is ready to test these theories and exclude them or proof them.

One of the detectors that can test the standard model is the CMS detector. In order to provide most precise measurements, the detector has to be maintained and upgraded. Therefore several components of the detector have to be exchanged and updated after some period of running. The innermost part of the CMS detector, the pixel detector will be exchanged completely in the shutdown 2016/2017. The preparation of this exchange involves several challenges, like mass production and calibration of the pixel detector components. In the first part of this thesis, the upgrade of the detector will be explained and the x-ray calibration of pixel modules is shown. An internal circuit of the pixels readout chip is calibrated with known x-ray flourescences. The temperature dependence of the calibration procedure and its stability for the mass production is investigated.

The heaviest known particle in the stanadard model of particle physics is the top quark. It plays a special role in the standard model since it gives the highest correction to the Higgs mass and it can be the key to discover new physics. Possible new particles are expected to be heavy since they were not discovered yet. In many theories, these heavy particles interact most strongly with the next heavy particles, the top quarks. This makes the identification of the top quark in the detector extremely important. Top quarks decay into a W boson and a b quark, where the W boson can further decay hadronically or leptonically. The emphasis on this thesis is the identification of boosted hadronically decaying top quarks, which are detected as jets in the detector. The identification of top quarks is challenging since its decay products are merged within one single jet for high transverse momentum of the top quark. Top tagging algorithms are tools to identify boosted hadronically decaying top quarks and reject at the same time QCD multijet events, using substructure techniques. A major background for top tagging are QCD multijet events, which have a similar signature in the detector. In this thesis several top quark identification methods are tested, with a focus on the increased centre-of-mass energy of the LHC. In the last part of the thesis a novel method to identify top quarks is introduced. The advantage of the new approach is the identification of top quarks over a large energy range.

The structure of this thesis is explained in the following. In Chapter 2 the theoretical background is given, introducing the standard model of particle physics. Properties, production and decay of the top quark are explained. Chapter 3 describes the LHC, the CMS detector and its subdetectors. The upgrade of the pixel detector and the x-ray calibration of the pixel modules for mass production is shown in Chapter 4. The simulation of events and the simulation of the detector are explained in Chapter 5. In Chapter 6 the reconstruction and identfication of particles and higher level objects with the corresponding algorithms is presented. Chapter 7 summarizes existing methods for identifying hadronically decaying boosted top quarks. Common methods and new approaches as for example the OptimalR HEP top tagger and the shower deconstruction tagger are explained. The identification of top quarks using a multivariate analysis methods is presented in Chapter 8. Several substructure variables were investigated in order to achieve the best signal to background ratio. The comparison of top tagging algorithms on generator level is shown in Chapter 9. In Chapter 10 the top tagging methods are validated on 8 TeV data, collected in 2012 with the CMS detector. Different event generators are compared to data. In chapter 11 a new top tagging algorithm is developed, described, tested, optimized and validated in 8 TeV data in detail. Chapter 12 concludes the thesis.

2 Theory

Already thousands of years ago, in the 5th century BC, humans imagined that matter is made out of elementary particles. Nowadays, we have a successful model that describes the smallest parts of our universe, the elementary particles, and can even give predictions about the biggest part, our visible universe.

2.1 The Standard Model of particle physics

The Standard Model (SM) of particle physics describes all known elementary particles and their interactions, the strong interaction, the weak interaction and the electromagnetic interaction. However, the gravitational interaction cannot be described by the SM. Nevertheless its influence is negligible in particle physics because of the small gravitational coupling constant. Gravity can be described in great detail with the theory of general relativity.

The SM of particle physics is an extremely successfully theory and has been tested with very high precision in various experiments.

Two different types of particles are described by the SM. The fermions with spin 1/2 make up all visible matter. Particles with integer spin values are called bosons. Gauge bosons are responsible for the interactions between the fermions. The weak interaction is mediated by the exchange of W and Z bosons, the photon mediates the electromagnetic force and the gluon the strong force.

Fermions are categorized into leptons and quarks [16]. They are divided into these two categories because they are interacting differently with the gauge bosons. Leptons are colorless and interact weakly and electromagnetically, since quarks carry a color charge they interact in addition strongly. Furthermore, there are three generations of each type, grouped according to their masses. The Higgs field gives mass to the fermions and the associated particle is the Higgs boson. The main sources of information for this section are Ref. [17–19]. An overview of all fermions and bosons is given in Figure 2.1.

2.1.1 Gauge theory

The SM is a renormalizable, local gauge invariant theory with the underlying gauge groups

$$\mathrm{SU}(3)_\mathrm{C} \times \mathrm{SU}(2)_\mathrm{L} \times \mathrm{U}(1)_\mathrm{Y}, \tag{2.1}$$

where the index C stands for color charge, the index L for left handed particles and the index Y for hypercharge. The SU(3) gauge group is responsible for the strong force which acts on particles with color charge, the $\mathrm{SU}(2)_\mathrm{L} \times \mathrm{U}(1)_\mathrm{Y}$ represents the unification of the weak and electromagnetic forces after spontaneous symmetry breaking. The weak force acts just on left-handed fermions and right-handed anti-fermions, and the electromagnetic

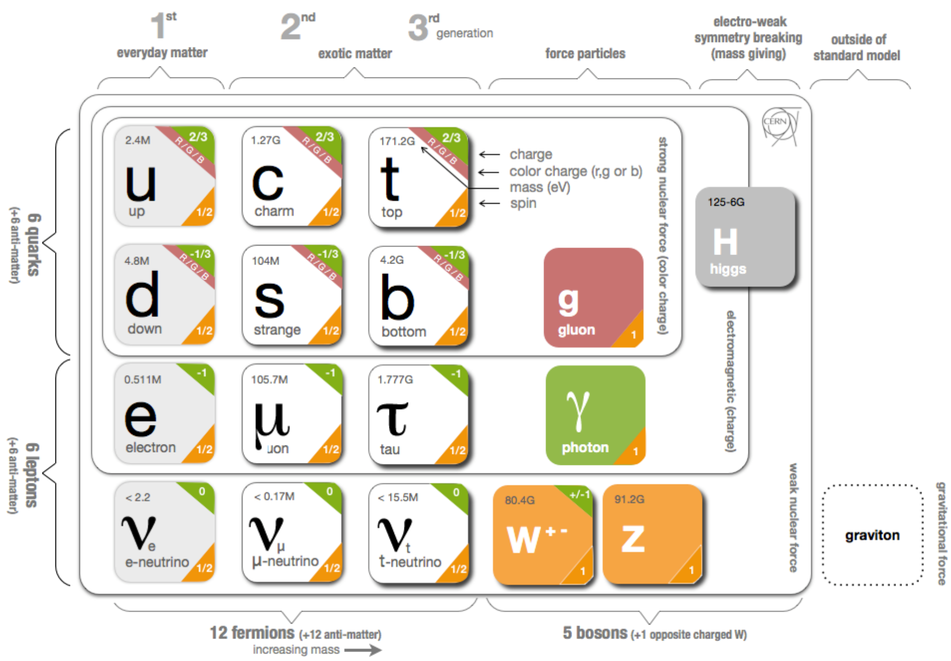


Figure 2.1: Particle content of the standard model of particle physics [20].

force on particles that carry an electric charge.

2.1.2 Electromagnetic interaction

The electromagnet interaction is described by the exchange of photons. For free half integer particles with a mass m the Dirac equation can be used to obtain their equations of motion

$$(\not{p} - m)\Psi(p) = 0, \tag{2.2}$$

$$(i\not{\partial} - m)\Psi(x) = 0. \tag{2.3}$$

Where Equation (2.2) is the Dirac equation in the momentum space and Equation (2.3) the Dirac equation in position space. That leads to the quantized Dirac fields and the Lagrangian density

$$L = \overline{\Psi}(x)(i\not{\partial} - m)\Psi(x). \tag{2.4}$$

Quantum electrodynamics (QED) has to be invariant under U(1) transformations, meaning the Lagrangian has to be invariant under the transformation $\Psi(x) \to \exp(i\alpha(x))\Psi(x)$. Now the transformation to the free Dirac Lagrangian (Equation (2.4)) are applied. And the gauge invariance is broken

$$\partial_\mu \Psi \rightarrow \exp(i\alpha(x))\partial_\mu \Psi + i\exp(i\alpha(x))\overline{\Psi}\partial_\mu \alpha(x). \quad (2.5)$$

The last term is breaking the gauge invariance, therefore the derivation operator is modified to

$$D_\mu = \partial_\mu + ieA_\mu. \quad (2.6)$$

Here A_μ is the four-vector potential of the photon. By changing the operator and making the theory gauge invariant the last missing term in the Lagrangian of QED, which is the interaction term, is introduced. After introducing a four-vector-potential $A^\mu = (\Phi(x), \vec{A}(x))$ of the photon, the antisymmetric tensor can be built

$$F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu = -F^{\nu\mu}. \quad (2.7)$$

The Lagrangian with the field tensor can be written as

$$L = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu}. \quad (2.8)$$

Summarizing, a Lagrangian that describes the motions and a Lagrangian of the photon were found. By requiring the theory to be invariant under U(1) transformations the interaction between the fermion field Ψ and the photon field A_μ are introduced to the Lagrangian.

The full Lagrangian for QED can then be written as

$$L = -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \sum_n \overline{\Psi}_n(i\gamma^\mu D_\mu - m_n)\Psi_n, \quad (2.9)$$

where n is a particle index.

2.1.2.1 Running couplings and renormalization

Since an electrically charged particle is always surrounded by virtual photons and e^+e^- pairs are created and annihilated, a so-called vacuum polarization [21] appears. Responsible for this effect is the attraction of the opposite charge carrier to the particle, which tend to screen the charge of the particle, this effect is similar to the polarization of a dielectric medium after placing it next to a charge. Vacuum polarization gives rise to an effective charge $q(r)$ that depends on the distance r and on the momentum transition Q. The β function can be defined

$$\beta(r) = -\frac{dq(r)}{d\ln(r)}, \quad (2.10)$$

which describes the change of the charge with the distance. In QED the β function is positive.

Charge screening leads to the concept of running couplings.

One can calculate the Matrix element for a diagram in leading order (see. Figure 2.2

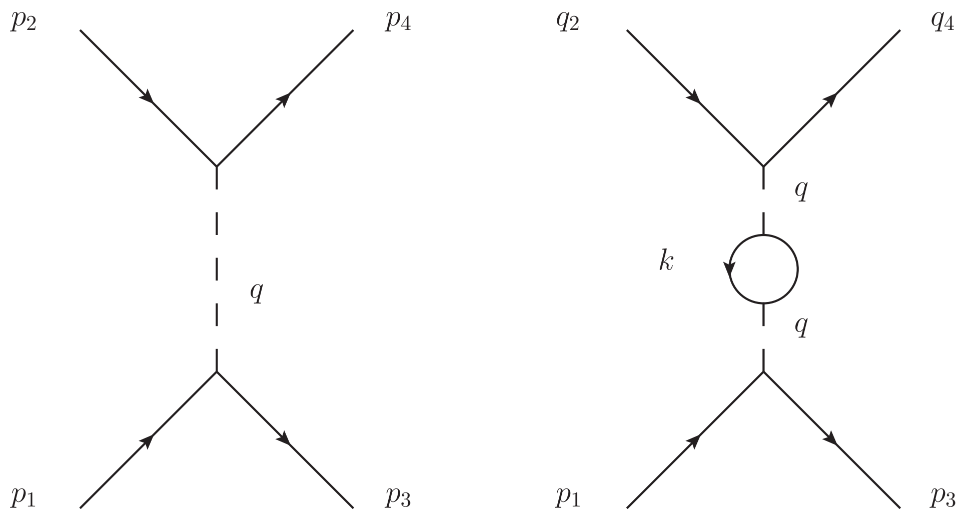


Figure 2.2: Feynman diagram for electron-muon scattering in leading order (LO) (left) and with an additional virtual loop (right)

left)

$$M = -g_e[\overline{u}_3\gamma^\mu u_1]\frac{g_{\mu\nu}}{q^2}[\overline{u}_4\gamma^\nu u_2]. \tag{2.11}$$

The matrix element for the NLO loop diagram for vacuum polarization in Figure 2.2 (right) can be calculated with

$$M = -g_e[\overline{u_3}\gamma^\mu u_1]\frac{-iI_{\mu\nu}}{q^4}[\overline{u}_4\gamma^\nu u_2]., \tag{2.12}$$

where $I_{\mu\nu}$ is

$$I_{\mu\nu} = -g_e^2\int\frac{d^4k}{2\pi^4}\frac{\mathrm{Tr}[\gamma_\mu(\not{k}+m_e)\gamma_\nu(\not{k}-\not{q}+m_e)]}{[k^2-m_e^2][(k-q)^2-m_e^2]}. \tag{2.13}$$

By adding a loop diagram one thus effectively modifies the coupling constant of the electromagnetic force by

$$\frac{g_{\mu\nu}}{q^2} \to \frac{g_\mu\nu}{q^2} - \frac{-iI_{\mu\nu}}{q^4}. \tag{2.14}$$

Taking all possible Feynman diagrams into account the coupling constant is not constant and the modification can be interpreted as a loop correction. The integral $I_{\mu\nu}$ is unfortunately not finite. The integral in Equation (2.13) can be divided into two parts

$$I_{\mu\nu} = -ig_{\mu\nu}q^2I(q^2) + q_\mu q_\nu J(q^2), \tag{2.15}$$

where the second term is not contributing to the matrix element. In the first term $I(q^2)$ is

$$I(q^2) = \frac{g_e^2}{12\pi^2}\left(\int_{m_e}^{\infty}\frac{dz}{z} - 6\int_0^1 dz\, z(1-z)\ln\left[1-\frac{q^2}{m_e^2}z(1-z)\right]\right). \tag{2.16}$$

The first term of Equation (2.16) is logarithmically divergent. By imposing a cutoff scale μ the integral becomes finite

$$\int_{m_e^2}^{\infty} \rightarrow \int_{m_e^2}^{\mu^2}\frac{dz}{z} = \ln(\frac{\mu^2}{m_e^2}) \tag{2.17}$$

This method is called cutoff regularization [22] and is mathematically used in order to isolate singularities. Unfortunately, cutoff regularization is not Lorentz invariant, however there are several regularization schemes that are invariant under Lorentz transformation like dimensional regularization [23].
The Ward identities [24] proved that the higher order vertex corrections cancel exactly the effect of self energies to all orders. Only the vacuum polarization diagrams contribute to the coupling constant and the effective charge. Consequently, these diagrams are the only ones that can cause ultraviolet (UV) divergences, which can be absorbed by introducing an arbitrary scale μ.
One can write the running coupling constant as a function of the momentum transfer Q

$$\alpha(Q^2) = \alpha(0)\left(1+\frac{\alpha(0)}{3\pi}f\left(\frac{Q^2}{m_e^2}\right)+O(\alpha^2)\right). \tag{2.18}$$

Here the cutoff scale μ is absorbed in $\alpha(0)$ and has to be replaced by a measurement, for QED $\alpha(0)$ becomes the fines structure constant $\alpha = 1/137$.
Considering all possible loop contributions the coupling constant becomes

$$\alpha(Q^2) = \frac{\alpha(0)}{1-(\alpha(0)/3\pi)\ln(Q^2/m_e^2)}. \tag{2.19}$$

However the given formula is not exact but a leading-log approximation.

2.1.3 The strong interaction

The strong interaction describes the force between particles that carry color charge. Only quarks participate in the strong interaction, since these are the only fermions that carry a color charge. The gluon interacts with itself since it carries as opposed to other bosons its own charge. The theory of the strong interaction has to be invariant under $SU(3)_C$ gauge transformations and describes the confinement [25] of quarks as well as asymptotic freedom [26, 27] of quarks at short distances and high momentum transfers.
Starting as in QED and generalizing the gauge principle for a SU(N) group with the generators of the gauge group t_i,

$$[t_a, t_b] = i f_{\text{abc}} t_c, \tag{2.20}$$

here f_{abc} is the tensor for the SU(N) algebra.
The gauge group SU(N) has $N^2 - 1$ free generators.
To make the Lagrangian invariant under SU(N) transformations one introduces a field as it was done in QED, the gluon field

$$G_\mu(x) = G^a_\mu(x)t^a. \tag{2.21}$$

And also here the partial derivative has to be changed

$$\partial_\mu \to D_\mu = \partial_\mu + ig_\mathrm{s} G_\mu(x). \tag{2.22}$$

Here g_s is the coupling constant for the strong interaction. One can now write the Lagrangian as before the QED Lagrangian:

$$L_\mathrm{QCD} = \sum_\mathrm{Flavors} \overline{\Psi}_f(x)(i \not{D} - m_f)\Psi_f(x) - \frac{1}{4} G^a_{\mu\nu}(x) G^{\mu\nu,a}(x) \tag{2.23}$$

The $G^a_{\mu\nu}$ represents the gluon fields and for the case $N = 3$ one obtains $3^2 - 1 = 8$ gluons.

2.1.3.1 Renormalization and running couplings

As for QED higher order diagrams contribute to the coupling constant of QCD. Because of the self coupling of the gluon also Feynman diagrams with gluon loops have to be considered. In contrast to the quark loops, gluon loops give a negative contribution (anti-screening) to the β function.
The formula for NLO diagrams with one loop to calculate the running coupling in QCD is given through

$$\alpha_S(Q^2) = \frac{\alpha_S(\mu^2)}{1 + \beta_0 \alpha_S(\mu^2) \ln(Q^2/\mu^2)}, \tag{2.24}$$

$$\text{where } \beta_0 = \frac{11N_c - 2n_f}{12\pi}, \tag{2.25}$$

here N_c is the number of colors and n_f the number of flavors. Since $\beta_0 > 0$ the coupling constant becomes zero for high momentum transfers, which is called the asymptotic freedom. This effect is responsible for the quarks behaving like free particles at short distances and large momentum transfers, which was also observed in deep inelastic scattering. In Figure 2.3 the measurement of the running coupling in QCD of various experiments is shown. The theoretical predictions are in good agreement with the measurements of the various experiments.

2.1.3.2 Perturbative QCD

Since perturbative quantum field theory describes scattering processes where the initial and final states consist of free particles, perturbative theory can just be applied at short distances or high momentum transfers (asymptotic freedom). For low energies the coupling constant becomes large and makes therefore perturbative calculations impossible. Since in the experiment just color neutral objects are detected, it is difficult to connect the theoretical calculations to the experiments. With perturbative calculations one can obtain

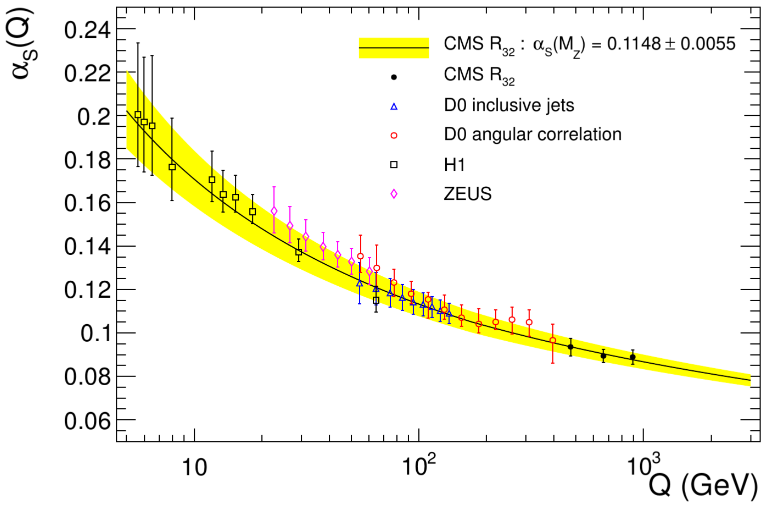


Figure 2.3: Measurement of the running coupling α_S of QCD as a function of Q in various experiments [28].

the coefficients β_i in the expanded β function

$$\frac{d\alpha(Q^2)}{d\ln(Q^2)} = \beta(\alpha) = -(\beta_0\alpha^2 + \beta_1\alpha^3 + \beta_2\alpha^4 + ..), \tag{2.26}$$

that was introduced before. In general, one can calculate an observable O when it is expanded in powers of the coupling constant α_S with perturbation theory

$$O = \sum_{n=0}^{\infty} c_n(\mu)\alpha_S^n(\mu), \tag{2.27}$$

here $c_n(\mu)$ are the coefficients that can be calculated from the evaluation of Feynman diagrams with perturbation theory. The observable O can be differentiated by μ for N terms

$$\frac{d}{d\ln\mu^2}\sum_{n=0}^{N} c_n(\mu)\alpha_S^n(\mu) \sim \mathcal{O}(\alpha_S^{N+1}(\mu)). \tag{2.28}$$

For large values of μ the coupling constant α_S becomes smaller than one and the dependency of the observable O on μ decreases with the number N of terms included [29].

2.1.3.3 Factorization theorem

In order to calculate the cross section in proton-proton collisions the factorization theorem can be used, where the partonic cross section, calculated for short distances with

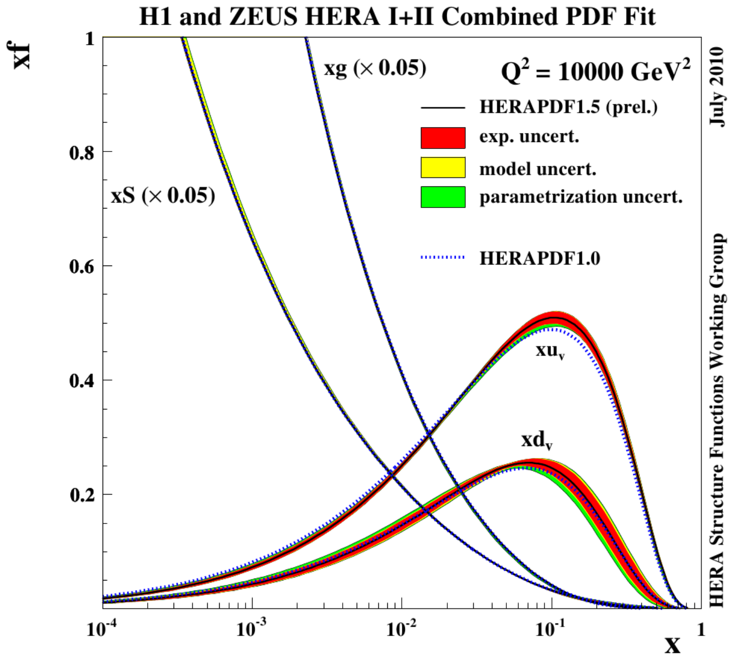


Figure 2.4: Parton distribution functions (PDFs) of the proton. The PDFs are shown as a function of the proton momentum transfer x for the valence quarks u_v and d_v, gluons g and sea quarks S [31].

perturbation theory and the parton distribution functions (PDFs) can be factorized,

$$\sigma_{pp} = \sum_{i,j} \int_0^1 \int_0^1 \sigma_{ij}(x_1, x_2, Q^2) f_i^{p1}(x_1, Q^2) f_j^{p2}(x_2, Q^2) dx_1 dx_2. \tag{2.29}$$

Here $f(x_i, Q)$ are the PDFs which give the probability that a parton i receives the momentum fraction x of the proton at an energy scale Q and σ_{ij} is the cross section for a specific final state. The PDFs have been measured in great detail in deep-inelastic scattering processes at the electron-proton collider HERA [30]. Perturbative calculations and the measurement of the PDFs are connecting QCD theory with experimental results.

2.1.3.4 Soft and collinear singularities

After renormalization, the UV divergences are absorbed. However, other singularities can appear in theoretical calculations since the observables in the experiments are not the quarks themselves but hadrons. In order to connect the theory and the experimental measurements the observables have to be collinear and infrared (IRC) safe. In the total cross section σ_{tot} the virtual and real singularities cancel exactly

$$\sigma_{\text{tot}} = \underbrace{\int_n |M_n^B|^2 d\Phi_n}_{\text{Born contribution}} + \underbrace{\int_n |M_n^V|^2 d\Phi_n}_{\text{virtual contribution}} + \underbrace{\int_{n+1} |M_{n+1}^R|^2 d\Phi_{n+1}}_{\text{real contribution}}. \tag{2.30}$$

In order to achieve the same cancellation for the differential cross section the generic observable $O(X; p_1, ..., p_n)$ should not change the matrix element $|M_{n+1}^R|^2$ since here the cancellation takes place,

$$\frac{d\sigma}{dX} = \int_n |M_n^B|^2 O(X; p_1, ..., p_n) d\Phi_n + \int_n |M_n^V|^2 O(X; p_1, ..., p_n) d\Phi_n + \int_{n+1} |M_{n+1}^R|^2 O(X; p_1, ..., p_{n+1}) d\Phi_{n+1}. \tag{2.31}$$

An observable does not change the matrix element if it is collinear and infrared safe, therefore it must remain unchanged in the limit of a collinear splitting, or in the emission of an infinitely soft particle. The following conditions have to be fulfilled for an IRC safe observable O:

$$O(X; p_1, ..., p_n, p_{n+1} \to 0) \to O(X; p_1, ..., p_n)\text{: infrared safe,} \tag{2.32}$$

$$O(X; p_1, ..., z \cdot p_n, (1-z) \cdot p_{n+1}) \to O(X; p_1, .., p_n)\text{: collinear safe,} \tag{2.33}$$

$$\text{where } z \in [0..1].$$

2.1.4 The electroweak interaction

The weak interaction is called weak since the coupling constant is small in comparison to the coupling constant of the strong interaction and therefore the lifetime of the decaying particles is long in comparison to lifetimes of other decays.
Typical decays caused by the weak interaction are

- β decay: $n \to p e^- \overline{\nu}_e$,
- μ decay: $\mu^- \to e^- \nu_\mu \overline{\nu}_e$,
- $\pi^\pm$ decay: $\mu^\pm$ + corresponding neutrino.

The typical lifetimes of particles decaying weakly are $O(10^{-8}\text{s})$, whereas the ones for the strong interaction are $O(10^{-23}\text{s})$ and for the electromagnetic force $O(10^{-16}\text{s})$. The electroweak interaction causes a CP-violating phase, which is one of the parameters of the CKM matrix (see following section).
Since the W boson couples just to left handed particles, as experimentally observed by P-violation [32], one introduces the following projectors

$$L = \frac{1-\gamma_5}{2}, \tag{2.34}$$

$$R = \frac{1+\gamma_5}{2}. \tag{2.35}$$

Here γ_5 is a product of the γ matrices $\gamma_5 = i\gamma_0\gamma_1\gamma_2\gamma_3$. Applying these operators to a particle, one gets the left handed e_L^- and the right handed e_R^- component of a particle, respectively

$$Re^- = e_\mathrm{R}^-, \tag{2.36}$$

$$Le^- = e_\mathrm{L}^-, \tag{2.37}$$

$$e_\mathrm{L}^- + e_\mathrm{R}^- = e^-. \tag{2.38}$$

In the SM neutrinos are left handed, without a right handed component.
Since the electroweak interaction acts differently on right and left handed particles and is therefore P-violating, one writes right handed particles as a singlet under SU(2) transformation and left handed particles as a doublet under SU(2) transformations.
As in the former sections, one has to modify the partial derivative

$$D_\mu = \partial_\mu - igt^a A^a_\mu - ig'YB_\mu, \tag{2.39}$$

to ensure gauge invariance. Here g and g' are two different coupling constants. One can change the representation to the mass eigenstates [33–35] and rewrite the fields $A^{1,2,3}, B$ with the boson fields $W^\pm, Z, A$, furthermore one can identify the $U(1)_{\mathrm{EM}}$ with the following definitions

$$Q = T^3 + Y, \tag{2.40}$$

$$e = \frac{gg'}{\sqrt{g^2 + g'^2}}, \tag{2.41}$$

$$g = \frac{e}{\sin\theta_W}. \tag{2.42}$$

Here T^3 is the third component of the weak isospin, Q is the electric charge, e the elementary charge and θ_W the Weinberg angle. One has to note that the hypercharge is different for right and left handed particles. One can then write the covariant derivative as

$$D_\mu = \partial_\mu - \frac{ig}{\sqrt{2}}(W^+_\mu \sigma^+ + W^-_\mu \sigma^-) - i\frac{g}{\cos\theta_W} Z_\mu (T^3 - \sin^2(\theta_W)\cdot Q) - \underbrace{ieA_\mu Q}_{U(1)_{EM}}. \tag{2.43}$$

Here the combinations of Pauli matrices $\sigma^\pm$ are the generators of the SU(2).
One can see that given by the definitions in 2.40 - 2.42 one can unify the electroweak and the electromagnetic interaction by using a $\mathrm{SU}(2)\times\mathrm{U}(1)$ gauge group and identify the $W^\pm$, Z and photon field.
The W boson can decay into a lepton and the corresponding neutrino or in a up-type and a down type quark. For the quark-anti-quark decay the decay width of the W boson is proportional to the squared CKM matrix element (see Section 2.1.5) multiplied with three for the possible quark colors. The Z boson decays into a fermion and its antiparticle. Equation (2.43) shows that each vertex factor includes the weak isospin T_3. For this reason its coupling to the fermions depends on the chirality of the fermion.

2.1.5 Symmetry breaking and Higgs mechanism

Whereas all known fermions have masses, the Lagrangian does not include mass terms so far. One can add a mass term trivially to the Lagrangian

$$L_{mass} = -\overline{\Psi} m \Psi. \tag{2.44}$$

In Section 2.1.4 it was described that the electroweak interaction distinguishes between right and left handed particles, therefore one writes the wave functions as

$$\Psi = \Psi_\mathrm{L} + \Psi_\mathrm{R}, \tag{2.45}$$
$$\overline{\Psi} = \overline{\Psi}_L + \overline{\Psi}_R. \tag{2.46}$$

Inserting this into the Lagrangian one gets

$$-\overline{\Psi}m\Psi = -(\overline{\Psi}R + \overline{\Psi}L)m(L\Psi + R\Psi) \tag{2.47}$$
$$= -\overline{\Psi}(R + L)m\Psi \tag{2.48}$$
$$= -\overline{\Psi}_\mathrm{R}m\Psi_\mathrm{L} - \overline{\Psi}_\mathrm{L}m\Psi_\mathrm{R}. \tag{2.49}$$

However, since the left handed and right handed wave functions transform differently under SU(2) and U(1), the Lagrangian is not gauge invariant.
Another method to generate mass is spontaneous symmetry breaking, whereby a new field is introduced that transforms as a doublet under SU(2)

$$\Phi = \begin{pmatrix} \Phi_u \\ \Phi_d \end{pmatrix}. \tag{2.50}$$

One can parametrize $\Phi(x)$ as the follows

$$\Phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h^0(x) \end{pmatrix}. \tag{2.51}$$

with the vacuum expectation value

$$\Phi_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}. \tag{2.52}$$

One calls Φ the Higgs field and h_0 is the actual Higgs boson [34,36,37].
Since the vacuum expectation value is not zero the gauge symmetry is broken spontaneously. Three of the four generators of the SU(2) $\times$ U(1) are broken and lead to three massive bosons and one massless boson. In the Lagrangian, two more gauge invariant terms appear after adding the Higgs field

$$L \supset (D_\mu\Phi)^\dagger D^\mu\Phi - V(\Phi), \tag{2.53}$$

where $V(\Phi)$ is the potential of the Higgs field and D_μ is the partial derivative of Equation (2.43). In the term $(D_\mu\Phi)^\dagger D^\mu\Phi$ the gauge fields appear squared and one can identify the mass terms of the gauge bosons as

$$m_W = \frac{vg}{2}, \tag{2.54}$$
$$m_Z = \sqrt{g^2 + g'^2}\frac{v}{2}, \tag{2.55}$$
$$m_A = 0. \tag{2.56}$$

By introducing the Yukawa coupling, one can now also give masses to the fermions,

$$L_{\text{Yukawa}} = -\overline{\Psi}_{\text{L}}\Phi\Psi_{\text{R}}Y_m + \text{h.c.} \ . \tag{2.57}$$

Here Y_m is the coupling constant for the Yukawa coupling. After spontaneous symmetry breaking the Lagrangian becomes

$$L_{\text{Yukawa}} = -\overline{\Psi}_{\text{L},2}\frac{vY_\Psi}{\sqrt{2}}\Psi_{\text{R}} - \overline{\Psi}_{\text{L},2}\frac{h^0 Y_\Psi}{\sqrt{2}}\Psi_{\text{R}} + \text{h.c.} \ . \tag{2.58}$$

Here $\Psi_{\text{L},2}$ is the second component of the doublet and furthermore one can identify the mass term for fermions

$$m_\Psi = \frac{vY_\Psi}{\sqrt{2}}. \tag{2.59}$$

The three mass generations can be written as matrices, where the mass of the up-type quarks is derived by the conjugated scalar field

$$M_u = \frac{vY_u}{\sqrt{2}}, \tag{2.60}$$

$$M_d = \frac{vY_d}{\sqrt{2}}. \tag{2.61}$$

Here the two matrices in general do not have to be diagonal. Once the representation is changed from the gauge eigenstates to the mass eigenstates one introduces the Cabbibo-Kobayashi-Maskawa (CKM) matrix $V_{\text{CKM}} = V_u V_d^\dagger$ [38] describing the transition from up-type quarks (charge=2/3 e) to down-type quarks (charge=-1/3 e), by interaction with a charged W boson,

$$V_{\text{CKM}} = \begin{pmatrix} V_{\text{ud}} & V_{\text{us}} & V_{\text{ub}} \\ V_{\text{cd}} & V_{\text{cs}} & V_{\text{cb}} \\ V_{\text{td}} & V_{\text{ts}} & V_{\text{tb}} \end{pmatrix} \tag{2.62}$$

$$= \begin{pmatrix} 0.97425 \pm 0.00022 & 0.2253 \pm 0.0008 & (4.13 \pm 0.49) \cdot 10^{-3} \\ 0.225 \pm 0.008 & 0.986 \pm 0.016 & (41.1 \pm 1.3) \cdot 10^{-3} \\ (8.4 \pm 0.6) \cdot 10^{-3} & (40.0 \pm 2.7) \cdot 10^{-3} & 1.021 \pm 0.032 \end{pmatrix}. \tag{2.63}$$

The CKM matrix is hierarchic and has entries around one on the diagonal [39], meaning that transitions within a quark family (for example u→d) are much more likely than the transition to other families (for example u→b).

After electroweak symmetry breaking the full Lagrangian can be written as:

$$\begin{aligned} L = & \underbrace{-\frac{1}{4}B_{\mu\nu}B^{\mu\nu} - \frac{1}{4}W^a_{\mu\nu}W_a^{\mu\nu} - \frac{1}{4}G^b_{\mu\nu}G_b^{\mu\nu}}_{(1)} + \\ & \underbrace{\overline{\Psi}_L\gamma^\mu iD_\mu\Psi_L}_{(2)} + \underbrace{\overline{\Psi}_R\gamma^\mu iD_\mu\Psi_R}_{(3)} + \\ & \underbrace{|D_\mu\Phi|^2}_{(4)} - \underbrace{\frac{m_h^2}{v^2}\left(\overline{\Phi}\Phi - \frac{v^2}{2}\right)^2}_{(5)} - \\ & \underbrace{\frac{\sqrt{2}}{v}\left(\overline{\Psi}_L\Phi M^f_-\Psi_R + \overline{\Psi}_L\Phi M^f_+\Psi_R + \text{h.c.}\right)}_{(6)} \end{aligned} \tag{2.64}$$

Here (1) describes the kinetic energy and the self-interactions of the three fields. Here $B_{\mu\nu}$ is the field tensor of the $U(1)_Y$, $W_{\mu\nu}$ the gauge field of the $SU(2)$ and $G_{\mu\nu}$ the gluon field of the $SU(3)$. The fourth term (2) describes the kinetic energy of the the left-handed fermions with the force fields. The fifth term (3) describes the same for the right-handed fermions. The next term (4) describes the masses and coupling of the electroweak bosons (W,Z) and finally the last two terms ((5) and (6)) describe the masses and couplings of the Higgs boson and the fermions.

2.2 Jets

Since quarks and gluons are affected by the color confinement, they can not be observed directly. At a sufficiently high energy, colored partons can be separated and form colorless hadrons. The final states consist usually of many particles that hadronize and shower in the detector.

In order to simplify final states, final state hadrons can be clustered into to so-called jets [40–42], which are simpler objects. Jets have been studied extensively already in CMS [43] and ATLAS [44] and perturbative QCD calculations are in good agreement with the measurement of the cross section as illustrated in Figure 2.5. The theoretical prediction of the cross section is in good agreement for many orders of magnitude. As it will be shown, there are different methods and algorithms to reconstruct jets. One important property that these algorithms have to fulfill is infrared and collinear safety (IRC) as explained in Section 2.1.3.4. The jet algorithms have to provide the same results when additional soft radiation is present or if a collinear splitting takes place as shown in Figure 2.6. Infrared safety means therefore that the addition of soft particles must not change the four-momentum of jets found by the algorithm. Collinear safety means that the splitting of random particles has to have no influence on the four-momentum of jets. Not all possible jet algorithms fulfill these requirements, leading to infinite order corrections and therefore rendering them useless for high-precision studies.

Furthermore, in order to reconstruct jets one has to define the distance parameter of a jet R_0, which defines the conesize of the jet. Ideally, a jet should contain all radiation that is

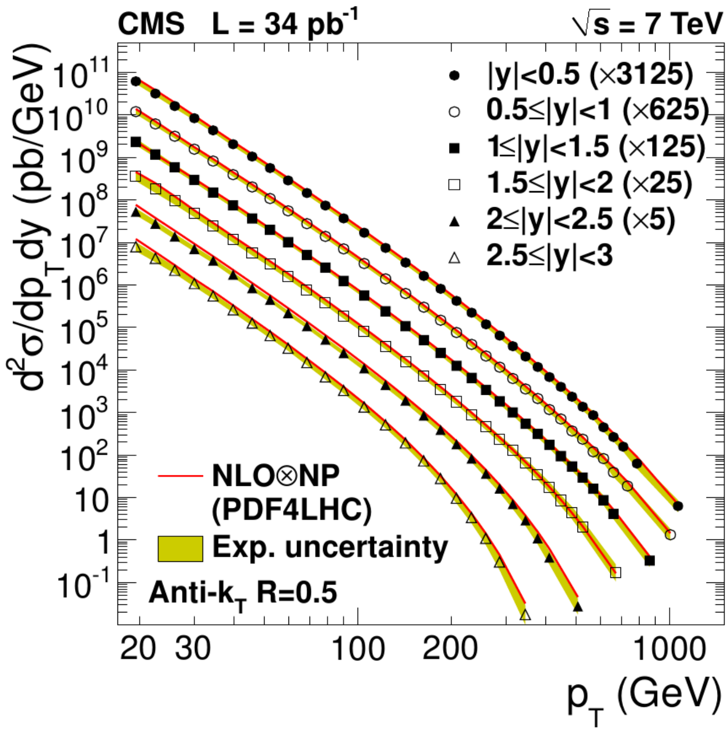


Figure 2.5: Inclusive jet differential cross section measurement as a function of the transverse momentum measured by CMS for different values of rapidity. The theoretical predictions have good agreement with the measurements within the experimental uncertainties [43].

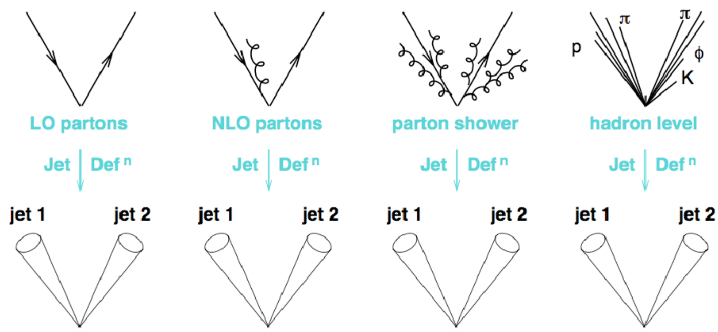


Figure 2.6: Infrared and collinear safe jet reconstruction. The jet reconstruction results in the same jets for leading order partons, NLO partons, parton shower and on hadron level always the same jets [45].

emmited by the parton, as well as a minimum of contamination from the underlying event (UE). The underlying event is an interaction of additional partons of the hard scattering. Through this process the number of particles in the hadronization process increases. In total three effects can contribute to the momentum of a jet: perturbative radiation, non-perturbative hadronization and the UE. In Ref. [46] the following relations were found between the distance parameter R_0 of the jet and the change of the transverse momentum δp_{T}

$$\text{perturbative radiation: } \delta p_{\mathrm{T}} \sim p_{\mathrm{T}} \cdot \ln R_0, \tag{2.65}$$

$$\text{hadronization: } \delta p_{\mathrm{T}} \sim \frac{1}{R_0}, \tag{2.66}$$

$$\text{UE: } \delta p_{\mathrm{T}} \sim R_0^2. \tag{2.67}$$

Minimizing the $\sum \delta \langle p_{\mathrm{T}} \rangle^2$ gives an optimal radius for the jets, as seen in Figure 2.7 for quark jets. As expected, the contributions from perturbative radiation and hadronization decrease with increasing R_0, whereas the contributions from the UE increases.

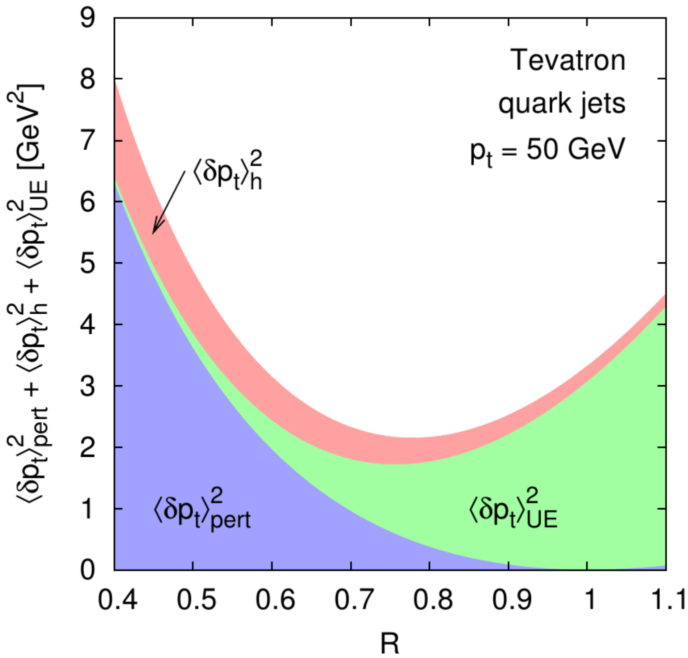


Figure 2.7: Minimum $\sum \delta < p_{\mathrm{T}} >^2$ as a function of the jet distance parameter R_0 for quark jets with a transverse momentum of 50 GeV at the Tevatron [46].

2.3 The top quark

The top quark as the heaviest particle, with a mass of $173.34 \pm 0.27(\text{stat}) \pm 0.71(\text{syst})$ GeV [47], plays a special role. It was discovered in 1995 at Fermilab [11].

The top quark has an relatively short lifetime of about 5×10^{-25}s, caused by its high mass. This lifetime is shorter than the average hadronization time, which makes the top quark the only quark that decays before it hadronizes.

The top quark plays an important role in the hierarchy problem [48], because of its high mass and therefore strong coupling to the Higgs boson. It is contributing the largest corrections to the Higgs mass via loop corrections. Due to its role in electroweak symmetry breaking, the top quark plays an important part in many beyond the SM (BSM) models.

In the following, top quark production at the LHC will be described as well as the decay of the top quark.

2.3.1 Production

At the LHC there are two main production channels, the top quark can be either produced together with its anti particle through QCD, or singly in electroweak processes. Pair production ($t\bar{t}$) is the dominant process at the LHC. It can be produced in leading order (LO) by the four Feynman diagrams shown in Figure 2.8. In proton-proton collisions the strong interaction plays a major role for producing the top quark pairs. There are two different ways of producing a top quark pair. The first is gluon fusion and the second is quark anti-quark annihilation. In order to produce top quark pairs, at least the energy of two times the top quark mass is required. At a centre of mass energy of 8 TeV the gluon fusion process is dominant, due to the strongly rising gluon density at small x (compare Figure 2.4).

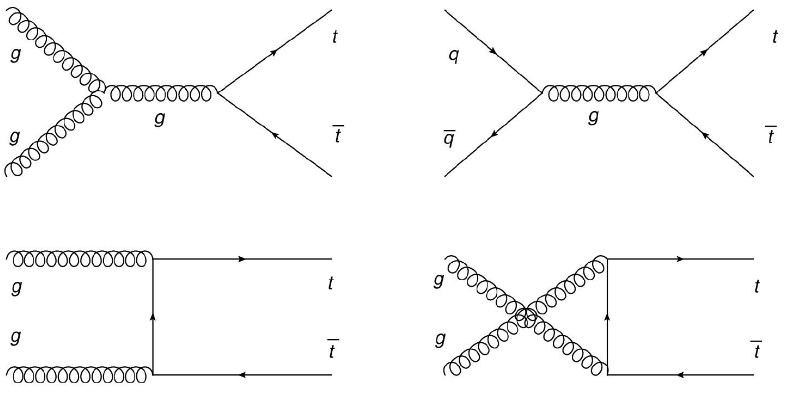


Figure 2.8: $t\bar{t}$ production at the LHC

The inclusive top quark pair cross section, calculated for $\sqrt{s} = 8\,\text{TeV}$ [49] is

$$\sigma_{t\bar{t}} = 239.1^{+9.2}_{-14.8}(\text{scales})^{+6.1}_{-6.2}(\text{PDFs})\text{pb}. \tag{2.68}$$

For Run II with increased centre of mass energy of $\sqrt{s} = 14\,\text{TeV}$ the top quark pair cross section will increase approximately by a factor of 3.9 to [49]

$$\sigma_{t\bar{t}} = 933.0^{+31.8}_{-51.0}(\text{scales})^{+16.1}_{-17.6}(\text{PDFs})\text{pb}. \tag{2.69}$$

2.3.2 Decay

The top quark decays via the weak force to a W boson and a b quark, since the CKM matrix element V_{tb} is approximately 1. The W boson decays further, either leptonically

into a lepton and the corresponding neutrino, or hadronically into two light quarks. Since the quarks are color charged, the W boson has six possibilities to decay hadronically ($3 \cdot \overline{\mathrm{u}}\mathrm{d}$ and $3 \cdot \overline{c}s$) and three possibilities to decay leptonically into the three existing leptons. Consequently, in 67% of the cases it decays hadronically and in 33% of the cases leptonically.
In this thesis the focus will be on the reconstruction of hadronically decaying top quarks, since the reconstruction is more challenging, especially for boosted top quarks. However, in Section 10.1.1 the reconstruction of leptonically decaying top quarks in a $\mathrm{t\bar{t}}$ system will be presented. In a $\mathrm{t\bar{t}}$ system the decay channel is called all-leptonic when both top quarks decay leptonically, semileptonic when one top quark decays leptonically and the other hadronically and full hadronic when both decay hadronically. The branching ratios for all possible final states are given in Table 2.3.2.

W^- decay \ W^+ decay	$u\bar{d}$, $c\bar{s}$	$e^+\nu_e$	$\mu^+\nu_\mu$	$\tau^+\nu_\tau$
$\overline{u}d$, $\overline{c}s$	45.7%	7.3%	7.3%	7.3%
$e^-\overline{\nu}_e$	7.3%	1.2%	1.2%	1.2%
$\mu^-\overline{\nu}_\mu$	7.3%	1.2%	1.2%	1.2%
$\tau^-\overline{\nu}_\tau$	7.3%	1.2%	1.2%	1.2%

Table 2.1: Branching ratios for all possible final states of the $\mathrm{t\bar{t}}$ system, taken from [50].

The decay products of the hadronically decaying top quark form hadrons, which are reconstructed as jets inside the detector. As long as the momentum of the top quark is small enough, the three decay products can be reconstructed in three separate jets. For higher energies of the top quark these jets are collimated in one large jet and in order to reconstruct the decay products, substructure methods have to be used. In Figure 2.9 the two scenarios can be seen (high momentum and low momentum of the top quark).

2.4 Physics beyond the standard model

Unfortunately the standard model is not able to explain all phenomena. Astronomical observations, like galactic rotation curves, predict the existence of dark matter, that make up around 28.3% of the matter in the universe, whereas the bayronic matter described by the standard model makes up just 5%. The rest is dark energy. These phenomena that are not described by the SM indicate that the SM is only a low energy approximation of a more fundamental theory.

2.4.1 Search for new physics with top quarks

The identification of top quarks is especially interesting for searches for new physics. Many models predict heavy resonances that decay mostly into top quark pairs. Some examples of these models are including colorons [51–53], leptophobic topcolor Z' [54], extended gauge theories with massive color-singlet Z' bosons [55–57], axigluons [58, 59], or models in which a pseudoscalar Higgs boson may couple strongly to top quarks [60]. Furthermore, extensions of Randall-Sundrum models [61] that predict more than three space dimensions and therefore Kaluza-Klein excitations of gluons [62] or gravitons [63] can couple as well

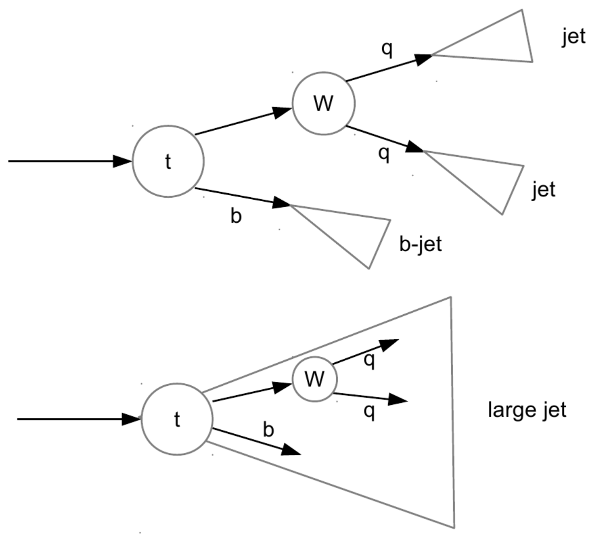


Figure 2.9: Top quark decay: resolved scenario (top), fully merged scenario (bottom).

to top quark pairs.
The heavier these heavy resonances are, the more boosted are the top quarks, consequently the decay products of the top quark are closer together and the identification of these top quarks becomes more difficult.
In the models cited above, the Z' boson decays most likely into $\mathrm{t\bar{t}}$. The invariant $\mathrm{t\bar{t}}$ mass distribution reconstructed from di-jet events for the full hadronic channel measured with the CMS experiment can be seen in Figure 2.10. In order to reconstruct the invariant mass, a top quark identification algorithm is used, but no b quark identification algorithms. A hypothetical Z' boson resonance with a mass of 2 TeV is shown. The reduction of the non-top-multijet background due to improvement of top tagging algorithms will improve the limits on the Z' boson mass.

Other extensions of the SM predict heavy charged gauge bosons [65, 66], e.g. a W' boson, decaying into a bottom and top quark. In order to reconstruct the W' mass in the hadronic channel, top quark identification methods using substructure techniques were used in [67].

Further models predict the existence of excited quarks [68,69]. Excited third generation quarks preferentially couple to third-generation SM quarks, like the top quark. Searches for b^* quarks in the gb decay mode with $gb \to b^* \to Wt$ were published in [70]. In the all-hadronic channel of this analysis, top tagging strategies were used.

Simplified models explaining dark matter with particles [71–73] were tested with the CMS experiment in [74]. Dark matter particles can be produced in association with top quark pairs, and are therefore another area where top tagging methods can be applied.

Consequently, the identification of boosted top quarks is an important key in order to search for new particles and physics beyond the SM.

Furthermore, the identification of top quarks can be used in order to perform precision

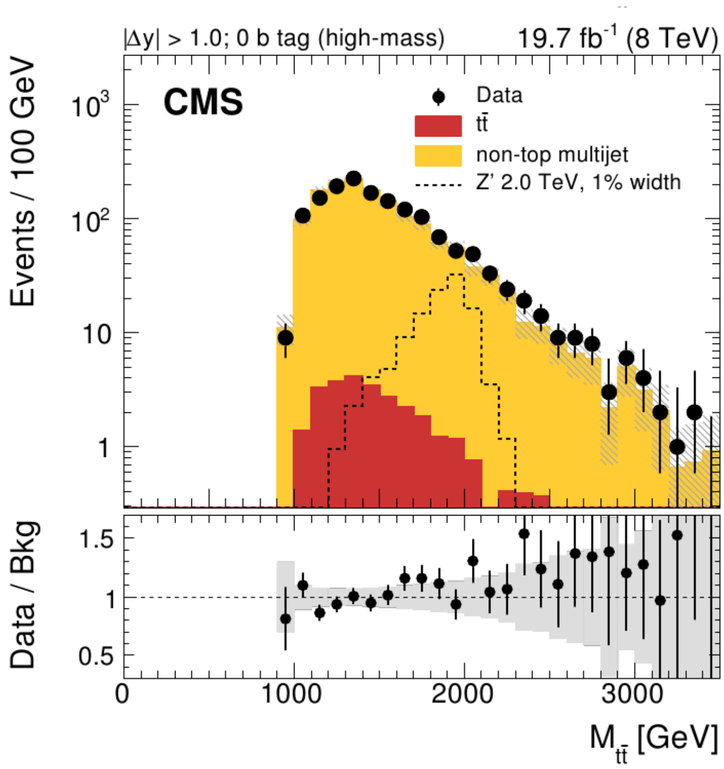


Figure 2.10: Invariant $t\bar{t}$ mass distribution reconstructed from di-jet events in the full hadronic channel, measured with the CMS experiment and used for searches for new physics [64].

tests of the standard model as in Ref. [75–77]. In Ref. [75] and [77] boosted top quarks were used in order to measure the differential cross section of $t\bar{t}$ processes. In Figure 2.11 the normalized differential cross section $1/\sigma \cdot d\sigma/dp_{\mathrm{T}}$ is shown as a function of the particle level top quark p_{T} (left) and of the generator top quark p_{T} for the combination of the electron and muon channel at a centre-of-mass energy of $\sqrt{s} = 8\,\mathrm{TeV}$ measured at the CMS experiment. The measurement is in good agreement with the prediction. Tagging algorithms are used to identify hadronically decaying top quarks.

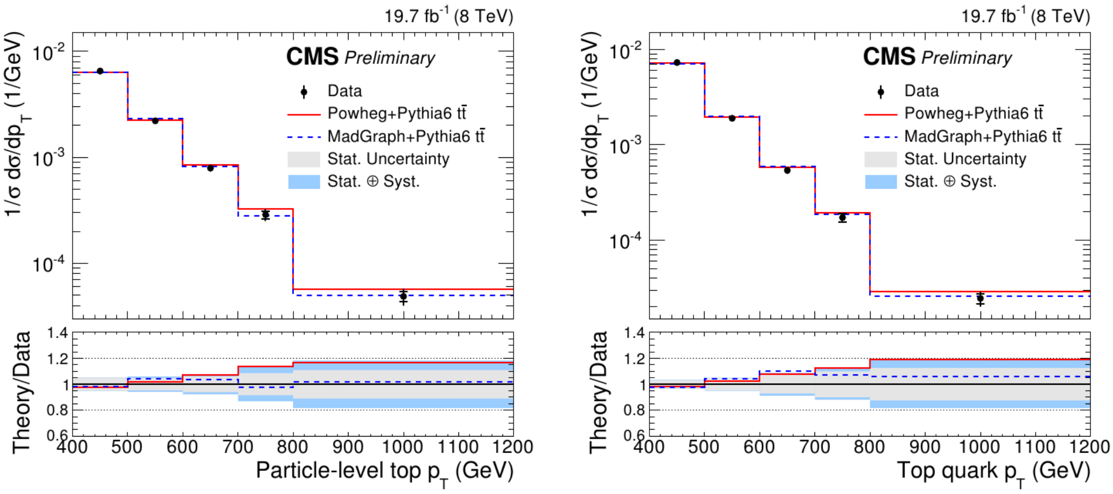


Figure 2.11: Normalized differential $t\bar{t}$ cross section for the combination of the muon and electron channel as a function of the particle level top quark p_{T} (left) and of the generator top quark p_{T} (right) [77].

3 Experimental Setup

3.1 The Large Hadron Collider

The Large Hadron Collider (LHC) [78], located at CERN in Geneva, Switzerland, accelerates protons and collides them with a design centre-of-mass energy of up to 14 TeV. To reach this energy, protons are accelerated in several steps. First they are accelerated linearly up to 50 MeV, afterwards they enter the Proton Synchrotron Booster (PSB), the Proton Synchrotron (PS) and the Super Proton Synchrotron (SPS). After these steps they reach an energy of about 450 GeV. Finally the protons are injected into the LHC and are accelerated to their final collision energy. Besides protons the LHC is capable to collide heavy ions as well. An overview of the full LHC accelerator complex can be found in Figure 3.1.

In order to keep the protons on their track inside the LHC ring extremely strong magnets are required to bend the trajectory of the protons. These magnets are superconducting and are cooled with liquid helium to a temperature of 1.9 K.

In 2010/2011 the LHC achieved a centre-of-mass energy of 7 TeV and in 2012 a centre-of-mass energy of 8 TeV with a peak instantaneous luminosity of $0.76 \cdot 10^{34}\text{cm}^{-2}\text{s}^{-1}$ and a proton bunch spacing of 50 ns (Run I). In 2015 Run II started with a centre-of-mass energy of 13 TeV and the proton bunch spacing was decreased to 25 ns.

The LHC itself is a superconducting synchrotron with a circumference of 27 km. To reach high luminosities the LHC accelerates bunches consisting of $1.7 \cdot 10^{11}$ protons. The proton bunches collide at four interaction points where the main experiments are based, namely LHCb [79], ALICE [80], ATLAS [81] and CMS [82]. The CMS detector will be described in more detail in Section 3.2.

To analyze the data and measure the cross section σ of a physics process the integrated luminosity L_{int} of the LHC has to be known. L_{int} can be derived by integrating the instantaneous luminosity L

$$L_{\text{int}} = \int L dt. \tag{3.1}$$

The number of expected events N for a specific process can be then calculated by

$$N = \sigma \cdot L_{\text{int}}. \tag{3.2}$$

The LHC is designed to achieve high values of the instantaneous luminosity to increase the discovery potential for heavy particles. The design instantaneous luminosity is $10^{34}\,\text{cm}^{-2}\text{s}^{-1}$.

3.2 The CMS detector

The Compact Muon Solenoid (CMS) [83, 84] is a general purpose detector based at one of the four interaction points at the LHC. Over 3200 people from 200 different institutes are

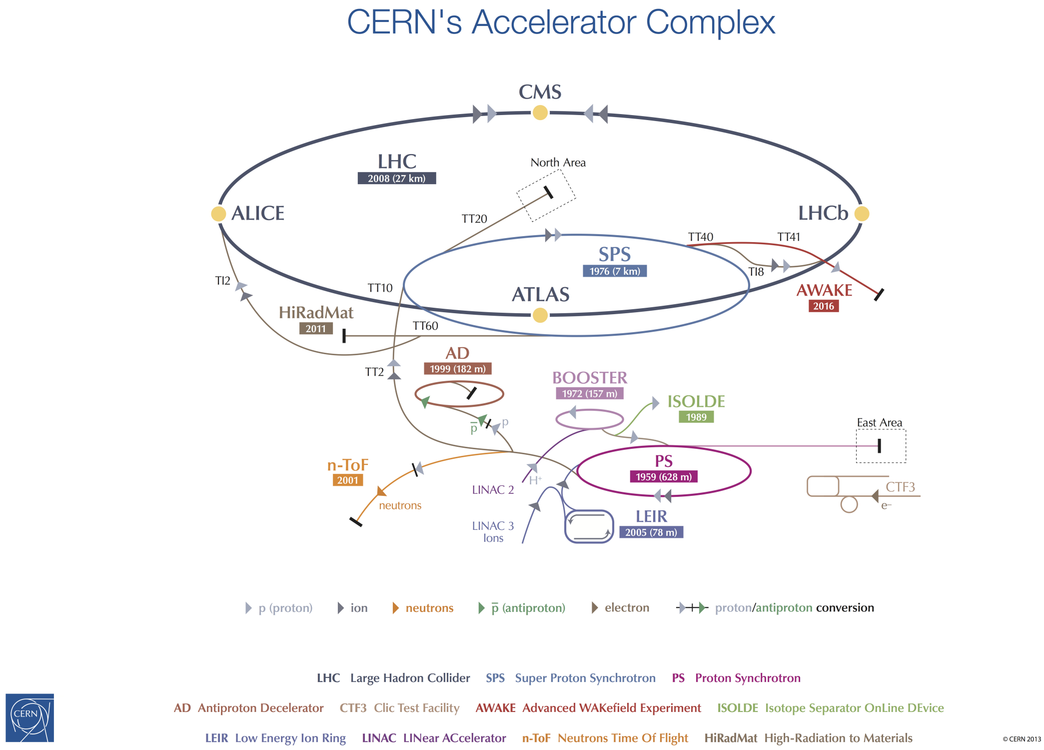


Figure 3.1: The LHC accelerator complex [78].

working at CMS. The detector weighs 14000 tons, has a length of 21 m and a diameter of 15 m. One of the main goals of CMS was to find the Higgs boson, which was discovered on July 31$^{\text{st}}$ of 2012 by CMS and ATLAS [15]. Other goals of CMS are tests of the standard model and the search for new physics.

To analyze proton collisions it is necessary to identify the created particles and measure their momenta and properties as precisely as possible. The momentum of charged particles can be measured by the deflection in the very strong magnetic field of about 3.8 Tesla. The detector has an onion-like structure with a tracking system in the centre, surrounded by the electromagnetic and hadronic calorimeters and the outermost part is the muon detector. In between the calorimeters and the muon detector the superconducting solenoid is localized in order to provide the magnet field. A schematic drawing of the CMS detector can be seen in Figure 3.2.

3.2.1 Coordinate conventions and kinematic variables

To describe the processes in the CMS experiment it is required to define a coordinate system. In CMS the z-axis is defined along the anti-clockwise beam, the x-axis points to the centre of the LHC and the y-axis points upwards. The azimuthal angle ϕ is given by the angle between the x-and the y-axis. The case $\phi = 0$ corresponds therefore to the x-axis. The polar angle θ is the angle between the positive z-axis and the y-axis. The centre of the coordinate system is the collision point. In Figure 3.3 the coordinate system

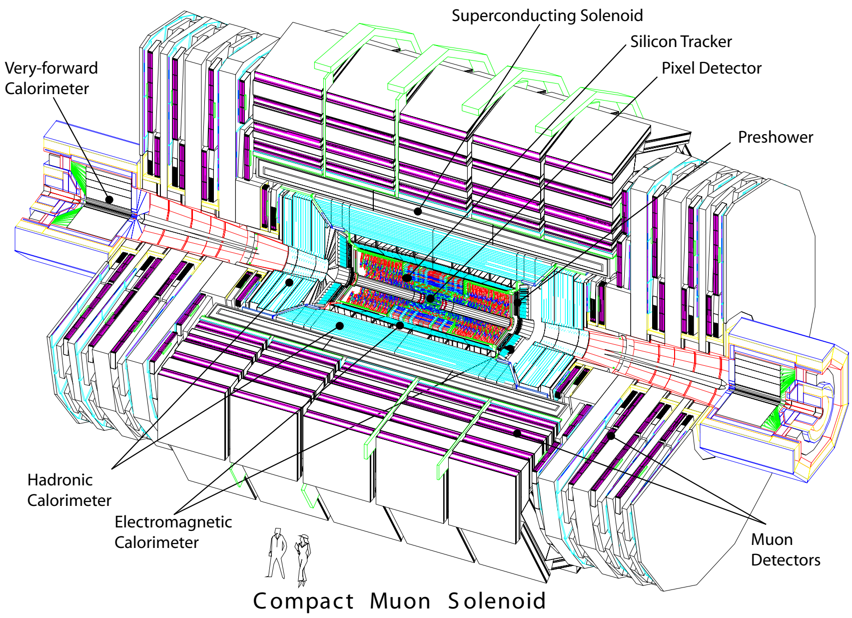


Figure 3.2: Overview of the CMS detector [83].

is shown.

Instead of θ, the pseudorapidity η is used, since differences in the pseudorapidity are invariant under Lorentz boots in z-direction. The pseudorapidity η is given by

$$\eta = -\ln\left(\tan\left(\frac{\theta}{2}\right)\right). \tag{3.3}$$

Here $\eta = 0$ is equal to the y-axis and $\eta \to \infty$ corresponds to the z-axis. Another widely used Lorentz invariant quantity to measure the distance between objects is given by

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

where $\Delta\phi$ is the difference in ϕ between two objects and $\Delta\eta$ the η difference, respectively.

Since in LHC collisions the hard interaction takes place between the partons of the protons and their initial fraction of the proton momenta is unknown, the total momentum

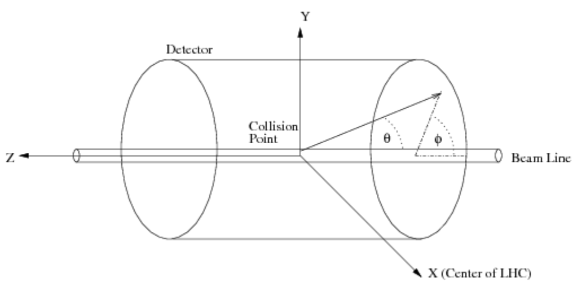


Figure 3.3: Coordinate system for describing physics processes [85].

can not be used in order to describe the momentum balance in the final state. However, it is known that the initial particles have almost no momentum in x and y. One can introduce a variable that is almost independent of the initial momenta of the protons and consequently more dependent on the physics process, the transverse momentum p_{T}, which is given by

$$p_{\mathrm{T}} = \sqrt{p_x^2 + p_y^2}. \tag{3.5}$$

3.2.2 The silicon tracker

The silicon tracker is the innermost part of the detector and is installed almost directly around the beam line, aiming to measure the tracks of charged particles as precisely as possible. Primary as well as secondary vertices used for b quark identification as explained in Section 7.9 are reconstructed with the tracks found by the silicon tracker [86]. The silicon tracker has two parts; the innermost part is the pixel detector which is surrounded by the strip tracker. The pixel detector consists of 3 barrel layers and 2 discs on each side with in total 66 million channels. For more details on the pixel detector see also Section 3.3. The strip detector consists of the Tracker Inner Barrel (TIB), the Tracker Inner Disks (TID), the Tracker Outer Barrel (TOB), and the Tracker End Caps (TEC) with in total 10 million channels. A schematic drawing of the complete tracker can be seen in Figure 3.4.

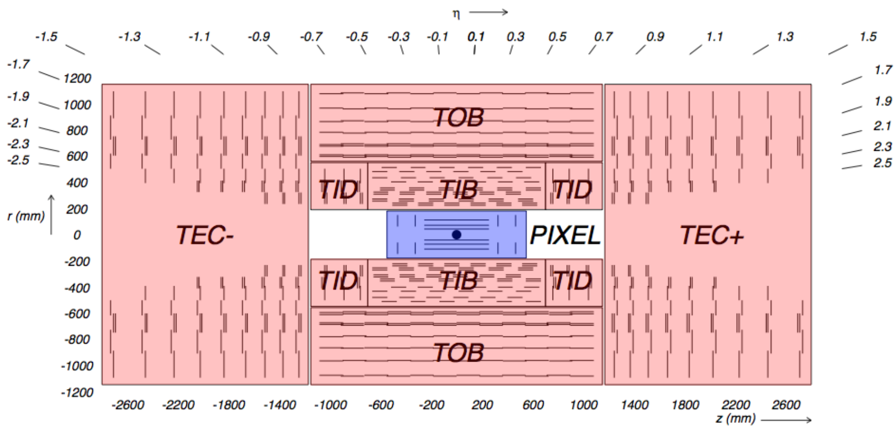


Figure 3.4: Schematic drawing of the silicon tracker. The pixel detector is located in the center. Each line indicates one layer in the silicon tracker [83].

3.2.3 The electromagnetic calorimeter

In order to measure the energy deposit of photons and electrons the silicon tracker is surrounded by the electromagnetic calorimeter (ECAL) [87], which is made of 75848 scintillating lead tungstate (PbWO4) crystals. Since the ECAL is a homogeneous calorimeter the crystals act at the same time as an absorber and a scintillator. Furthermore, these crystals provide a high density, a short radiation length X_0 of 0.89 cm, a small Moliere radius R_m of 2.2 cm and a fast response (80% of the light is emitted within 25 ns).

The ECAL covers pseudorapidities up to 3.0. It is divided in to two sections which differ by the read out. In the barrel section (EB) the signal is read out by avalanche photo diodes whereas the endcaps (EE) are read out by vacuum phototriodes. The EB covers a pseudorapidity range between $|\eta| < 1.479$ and the EE $1.479 < |\eta| < 3.0$. The ECAL preshower (ES) is located in front of the EE and covers $1.653 < |\eta| < 2.6$. A schematic drawing of the ECAL is shown in Figure 3.5. The relative energy resolution for electrons is given by [88]

$$\left(\frac{\sigma(E)}{E}\right)^2 = \left(\frac{2.8\%}{\sqrt{E[\mathrm{GeV}]}}\right)^2 + \left(\frac{12\%}{E[\mathrm{GeV}]}\right)^2 + (0.3\%)^2. \tag{3.6}$$

The stochastic effects of the shower development is described by the first term, the electric

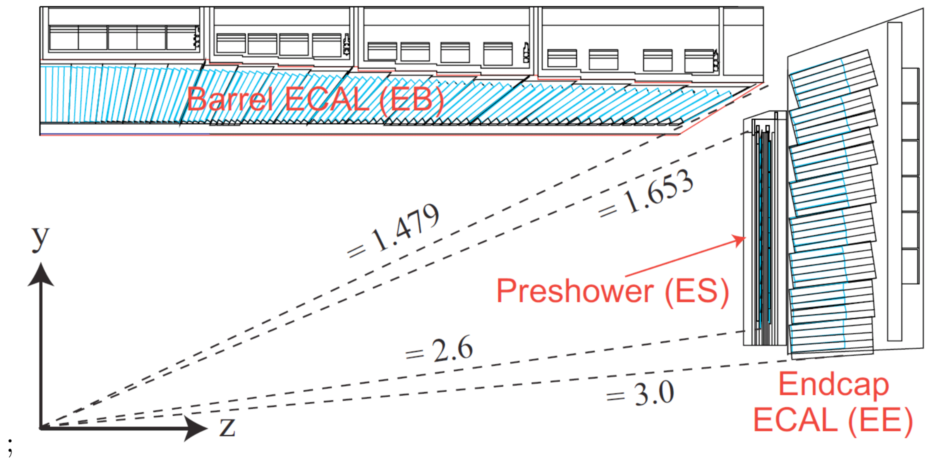


Figure 3.5: Schematic drawing of the electromagnetic calorimeter [83].

noise is parametrized by the second term and the third term describes non-uniformities and non-linearities in the detector response.

3.2.4 The hadronic calorimeter

In comparison to the ECAL the hadronic calorimeter (HCAL) [89] is a sampling calorimeter consisting of absorbers and active material. The absorbers are made out of brass whereas the active material are plastic scintillators connected via wavelength-shifters which are read out by photodiodes.

The HCAL also consists of several parts: The barrel (HB) covers a pseudorapidity range of $0 < |\eta| < 1.4$, the endcaps (HE) $1.3 < |\eta| < 3.0$ and the forward calorimeter (HF) $3.0 < |\eta| < 5.0$. In addition the outer calorimeter (HO), which is located between the magnet and the muon chambers, catches particles that are not absorbed in the HB. A schematic drawing of the HCAL can be seen in Figure 3.6. The energy resolution of the HCAL combined with the ECAL was obtained in beam tests with pions [83] and can be parametrized as

$$\frac{\sigma(E)}{E} = \sqrt{\frac{1.2}{\sqrt{E}} + 0.069^2}. \tag{3.7}$$

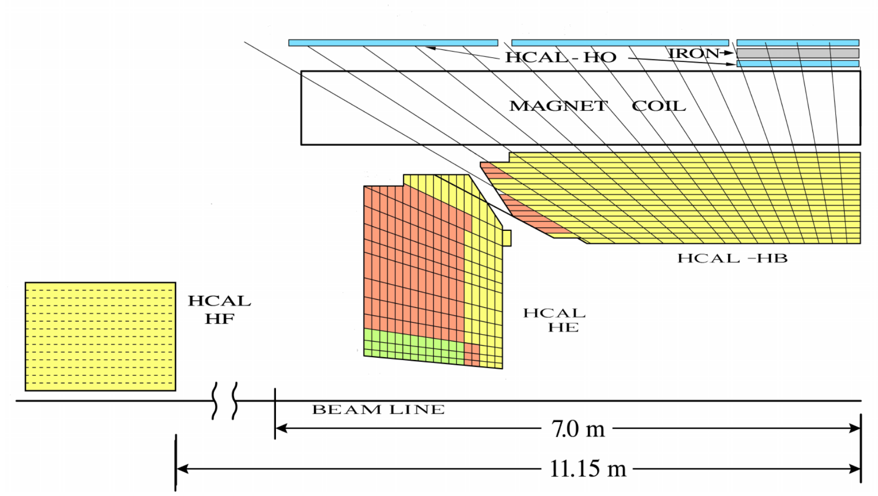


Figure 3.6: Schematic drawing of the hadronic calorimeter [83].

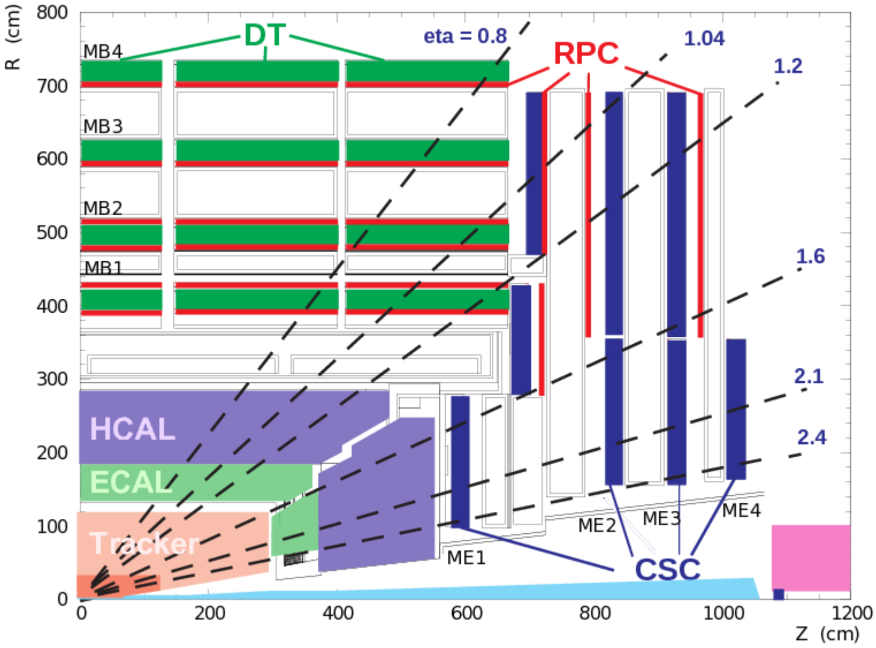
Here the energy E is given in GeV. Just the HCAL has approximately a resolution of $120\%/\sqrt{E[\mathrm{GeV}]}$ [90] for single pions.

3.2.5 The magnet system

In CMS a homogeneous magnetic field of 3.8 T is provided by a superconducting solenoid. The high magnetic field is generated with a current of 18 kA. The magnetic field is used to determine the momenta of charged particles traveling through the detector with bent tracks. Since the tracker and the calorimeters are inside the solenoid to reduce the absorption of particles inside the magnet, the dimensions of the solenoid are relatively large with an inner diameter of 6 m and a length of 12.5 m. Furthermore, the muon system is interleaved with a 10000 t iron yoke in order to extend the magnetic field and close the magnetic lines. Consequently, there is a magnetic field outside the solenoid with an inverted direction.

3.2.6 The muon system

Since muons are not stopped in the calorimeters due to their high mass, they have to be detected in a separate system, called the muon system [91]. The muon system consists of drift tubes (DTs), cathode strip chambers (CSCs), and resistive plate chambers (RPCs), all parts together cover in total a pseudorapidity range up to 2.4. The four barrel layers, consisting of 70 drift tubes, and are located at radii of 4.0 m, 4.9 m, 5.9 m and 7 m. Each hit measurement has a precision around 200 μm. The 234 cathode strips build up one endcap disc; in total there are four discs. The RPCs are located in the barrel region as well as in the endcap region. A schematic drawing of the muon system is shown in Figure 3.7.



;

Figure 3.7: Schematic drawing of the muon detector [83].

3.2.7 The trigger system

At the LHC proton bunches collide with a rate of 20 MHz (Run I), increased to 40 MHz (Run II). In CMS the amount of data is reduced by two steps of triggers [92], the hardware Level 1 Trigger (L1T) and the High Level Trigger (HLT). Most of the events are rejected by the L1T and just 1/1000 of the data is kept. Since the L1T decision has to be made within 3.2 μs, the L1T is based on simple transverse energy cuts of calorimeter and muon objects. Events passing the L1T are analyzed by the HLT. In comparison to the L1T, the HLT is a software trigger and the trigger decision is made in a huge processor farm. The general goal is to reject an event as soon as possible, therefore very sophisticated algorithms are used for the object reconstruction. To reject an event as fast as possible the detector components are used in different steps. First the information of the calorimeters and the muon systems are used and in a last step also the information from the tracker is taken into account. After the HLT the data is reduced drastically to a few hundred Hz [93].

The so-called trigger paths are different conditions that can be applied to the L1T and HLT. For instance, the triggers can select only events with a muon reconstructed with a certain p_{T} threshold or with a certain amount of hadronic activity.

3.3 The pixel detector

The innermost part of the CMS detector is the silicon pixel detector [86], which will be explained in more detail since the first part of this thesis emphasizes the upgrade and the x-ray calibration of the pixel detector. The pixel detector consists of three barrel layers, located at radii of 4.4 cm, 7.4 cm and 10.2 cm, and two discs at each side at a longitudinal distance of 34.5 cm and 46.5 cm for a pseudorapidity η coverage up to 2.5. In total the pixel detector has 66 million pixels with the pixels having a size of 150 μm $\times$ 100 μm and is designed to achieve an optimal vertex position resolution, a very good impact parameter resolution and three spatial points for track reconstruction. The layout of the pixel detector can be seen in Figure 3.8.

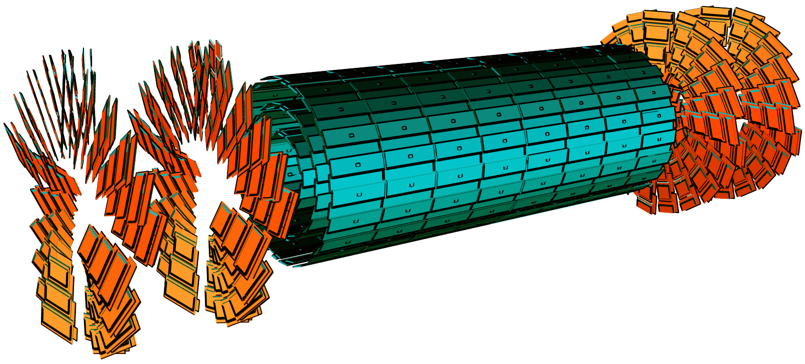


Figure 3.8: Schematic drawing of the current pixel detector [83].

A basic unit of the pixel detector is a module. A module consists of a silicon sensor, 16 readout chips (ROCs), a high density interconnect (HDI) and the signal and power cables. Since the pixel modules consists of a separate sensor and readout chip, bump bonded together, the detector is called "hybrid" detector. At the bottom of a module the base strips are located, responsible for the mechanical stability and to mount the modules to the cooling structure. The 16 readout chips are glued on top of the base strips with the sensor bump bonded on top. The top layer is the HDI which is glued as well to the sensor and connected via wire bonds to the readout chips. The detailed properties of the CMS barrel modules are given in Table 3.1 and a schematic drawing of the pixel module is given in Figure 3.9.

Size	66.6 mm $\times$ 26 mm
Weight	3.5 g
#ROCs	16
#Pixels per ROC	52$\times$80=4160
#Pixels	66560
Pixel size	100 μm $\times$ 150 μm
Sensor thickness	285 μm

Table 3.1: Properties of the CMS pixel barrel modules.

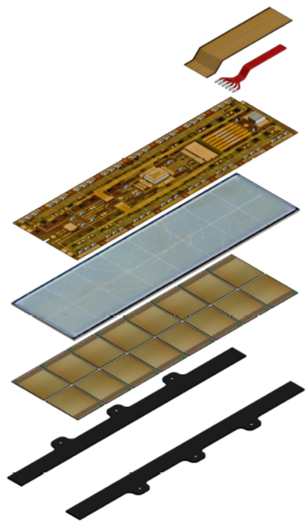


Figure 3.9: Schematic drawing of the CMS barrel pixel module, consisting of (from bottom to top) the base strips, the 16 readout chips, the silicon sensor, the HDI and the signal and power cables [94].

3.3.1 The sensor

Charged particles lose part of their energy due to elastic scattering when traversing the sensor. To detect them a depleted semiconductor is required. Charged particles crossing a sensor create electron-hole pairs. In silicon the energy needed to create an electron-hole pair is 3.6 eV. With a sensor thickness of 285 μm a minimum ionizing particle creates an average of 23000 electron-hole pairs inside the sensor. A reverse bias voltage is applied to increase the depletion zone and to collect these charge carriers. The bias voltage that is necessary to fully deplete an unirradiated sensor is around 150 V and can increase for irradiated sensors, although the maximum voltage is limited to 600 V due to the power cables. The charge carriers lead to a current pulse that is then detectable by the readout electronics.

The main challenge in the sensor design is the radiation hardness. The first pixel layer has to withstand a fluence of up to 1×10^{15} neq/cm^2. It was found that oxygen-enriched silicon substrate withstands higher fluences in comparison to low-oxygen silicon substrate [95]. In CMS a n-in-n sensor design is used, where the active volume is n-type, on the back-side of the sensor the sensor has a p-implant as a junction. In order to keep the pn-junction from the edge of the sensor, a guard ring structure is implemented on the back-side, keeping the edges of the sensor near ground potential and therefore preventing high voltage sparks between the sensor and the ROC. Consequently the sensor is double-sidedly processed.

On the top side of the sensor a p-stop technique was used for the forward pixel detector and a moderate p-spray technique in the barrel detector in order to isolate the n-implants from each other. More details on the sensor can be found in Ref [96].

3.3.2 The readout chips

The readout chips (ROCs) [94] handle the signal processing, buffering and the time-stamping of the detected hits. Each chip contains around 1.3 million transistors in an area of 7.9 mm × 9.8 mm. The chip itself can be divided into three areas: the array of pixel unit cells (PUC) organized in double columns, the double column periphery and a control and supply block.

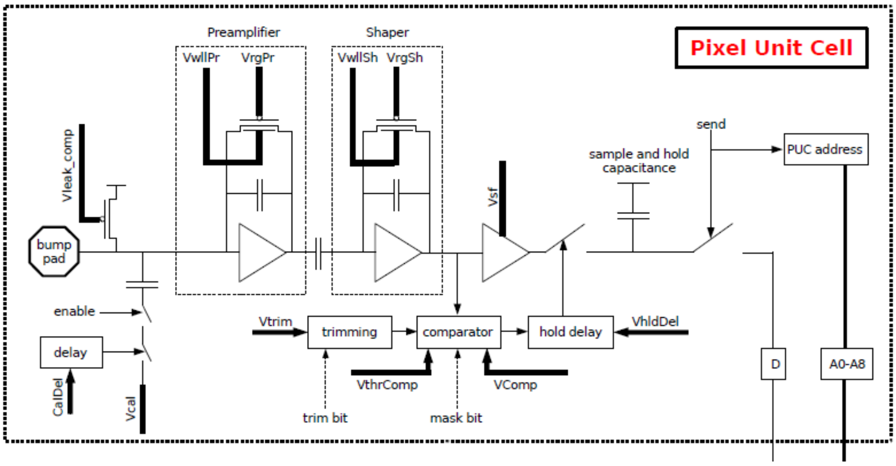


Figure 3.10: Schematic of the pixel unit cell [94].

The schematic of a PUC can be seen in Figure 3.10. The signal enters the pixel from the sensor through the bump bond. Afterwards the signal is amplified and shaped. It then enters the comparator, where the signal is compared to an adjustable threshold. The threshold is necessary since fake hits can be generated due to thermal effects. In order to minimize the number of fake hits one has to set a uniform threshold to all of the pixels of a module. If the signal passes the threshold it moves to the sample and hold capacitance and waits there for the readout.

Setting the threshold of each pixel is possible with the trimming procedure [97]. To set a global threshold to all pixels one has to set the VthrComp Digital Analog Converter (DAC) and the Vtrim DAC. Taking into account the variations of the pixels, it is possible to adjust the pixels with four trim bits which can be set individually per pixel. The trimming algorithm is able to unify the threshold with these three parameters. The threshold can be set to a value of Vcal DAC, which is an external calibration signal. The Vcal value can be related to the charge expressed in electrons with the x-ray calibration [see Section 4]. First the algorithm measures the threshold of each pixel. This is done by measuring the probability of a response of the pixel for different values of VthrComp. The value of VthrComp that has a 50% probability that the pixel responses is called the threshold of the pixel. The highest value of VthrComp, that is found, is set as a global value for the ROC. After the threshold is fixed in the first step the Vcal value is measured for each pixel. Vtrim is determined by the pixel with the highest Vcal response. In a last step the trim bits for each pixel are set, also determined by scanning different values of Vcal and measuring the response. In Figure 3.11 the threshold distribution of all pixels on one ROC before and after the trimming procedure are shown. One can see that the thresholds are much more uniform after the trimming procedure.

Furthermore, a mask bit for each pixel allows this pixel to be disabled in case of noise.
The double column periphery is responsible for the hit information from the pixels and stores the information in the double column buffers until the readout is requested by the trigger.

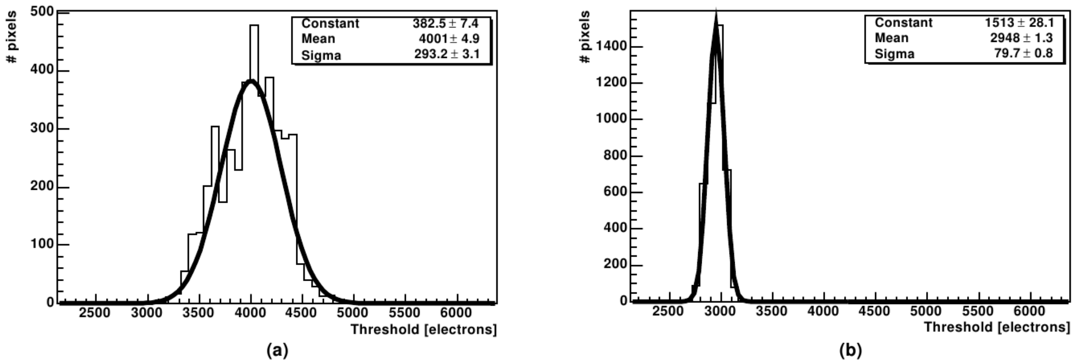


Figure 3.11: Threshold distribution for pixels on one ROC before (a) and after (b) the trimming procedure [97]. The threshold of the pixels get more uniform after the trimming procedure.

4 Phase 1 upgrade and x-ray calibration of CMS pixel detectors

In the first part of my Ph.D. time I was responsible for the x-ray calibration of the CMS pixel modules for the Phase 1 Upgrade in Hamburg. The CMS pixel detector will be exchanged during the so-called Phase 1 upgrade in 2016/2017. The new pixel detector will consist of four barrel layers and three endcap disks on each side. Half of the pixel modules for the fourth layer are produced by the University of Hamburg and DESY. This includes several production steps which will be explained in Section 4.1.1. One of the calibrations that is needed for the module production is the so-called x-ray calibration, which will be explained in Section 4.2. In the course of this work a setup for an automated calibration during the mass production of modules is developed. In Section 4.3, the setup and its construction will be explained. Furthermore, the dependence of the calibration procedure on temperature is investigated in Section 4.4. Finally, in Section 4.5 the stability of the procedure is tested and improved.

4.1 Phase 1 upgrade of the CMS pixel detector

After the long shutdown 1 (LS1) in 2013/2014 the design instantaneous luminosity of $1 \times 10^{34}\mathrm{cm}^{-2}\mathrm{s}^{-1}$ with 25 ns bunch spacing will be achieved. Consequently, the average event pile-up will increase up to more than 50. The increasing luminosity brings several challenges for the pixel detector [98]; the rate of the hits detected in the tracker will increase drastically.

The performance of the current CMS pixel detector in the period 2010-2011 is shown in Figure 4.1. With increasing instantaneous luminosity the track hit efficiency decreases, especially for the first layer. This effect is caused by the data buffers of the readout chips, which are filled and are not able to collect more data until the stored data is read out. Since the instantaneous luminosity will increase further, the tracker is expected to be even less efficient. One of the challenges for the Phase 1 upgrade is to improve the readout chips by increasing the buffer sizes and reduce their latencies to be capable to deal with the 25 ns bunch crossing.

During the Phase 1 upgrade in 2016/2017 the entire pixel detector will be replaced. The readout chip will be changed to a new digital readout chip with increased buffers to handle the higher hit rate due to higher pile-up. The digital readout is faster and therefore capable to deal with the increased instantaneous luminosity in Run II. Furthermore, a new cable allows a greater distance between the detector and the electronics, making it possible to reduce the material budget inside the detector. In Figure 4.2 the current module and the new module for the Phase 1 upgrade are shown. In comparison to the current pixel module the new pixel module has digital readout chips, a new HDI and just one cable for signal and power.

The overall of the design of the pixel detector will change, replacing the three existing layers as displayed in Figure 4.3. The first two new layers will be closer to the interaction

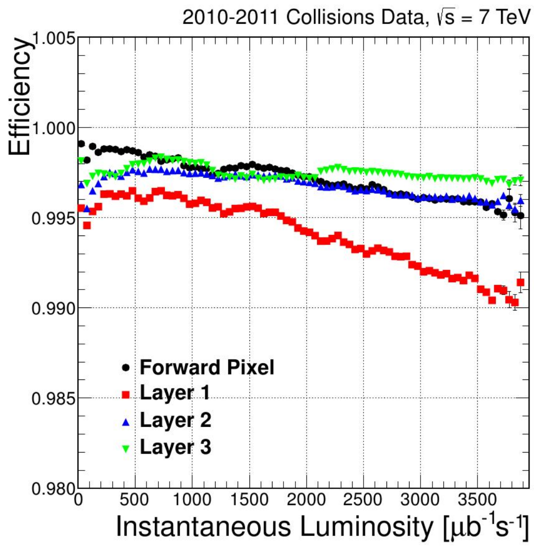


Figure 4.1: Tracking efficiency for the current CMS pixel detector as function of the instantaneous luminosity. All layers of the pixel detector, especially the first, show decreasing efficiency with increasing instantaneous luminosity [98].

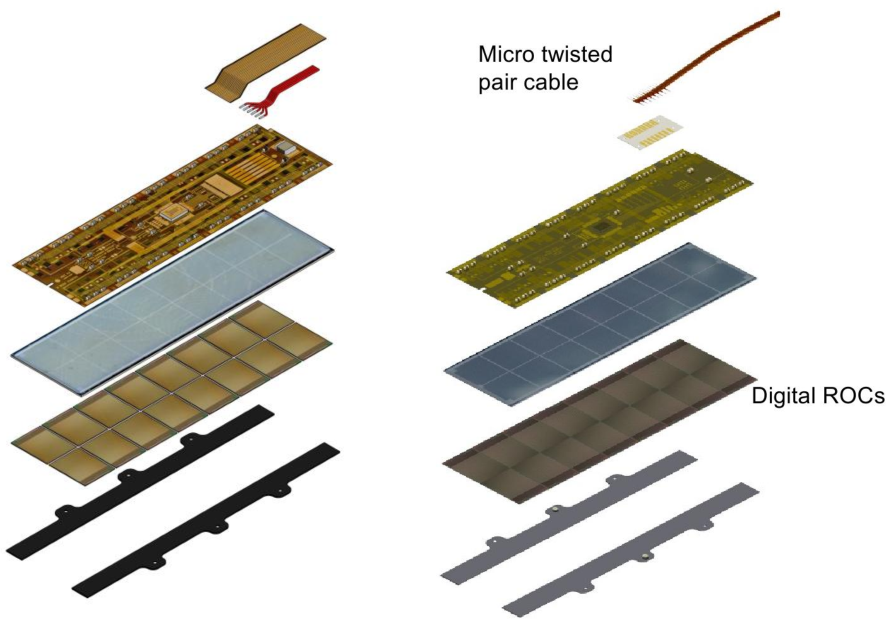


Figure 4.2: The current pixel module (left) [99] and the new pixel module (right) [100]: The analog readout chips are replaced by digital chips and the signal and power cable are replaced.

point at radii of 2.9 cm and 6.8 cm, whereas the third layer will be placed further away from the interaction point compared to the current third layer at a radius of 10.9 cm. The new fourth layer will be located at a distance of 16.0 cm. To bring the first layer closer to the interaction point the beam pipe is to be replaced. Moreover, the new pixel detector will consist of one more forward disc on each side at a distance of 51.6 cm.

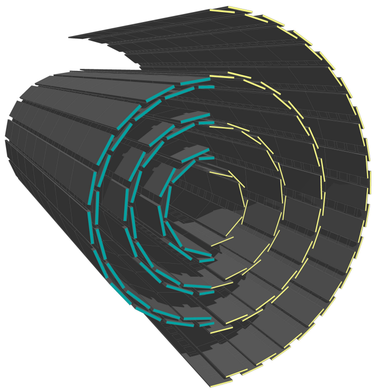


Figure 4.3: Schematic drawing of the current (left) and the new pixel detector (right), taken from [101]. The new pixel detector will contain an additional fourth layer and the position of other layers will change slightly.

The fourth additional layer will help to improve the b-quark identification efficiency shown in Figure 4.4. The graph shows the b-identification efficiency versus the misidentification rate (c and light jets) in simulation. An improvement of the performance due to improved secondary vertex resolution after the replacement of the current pixel detector is observed.

In addition the material inside the tracker will be reduced by moving the electronic boards outside the active tracking volume, replacing the support structure with lighter materials and changing to a new CO_2 cooling system. Figure 4.5 shows the amount of material inside the detector as a function of η for the current and the upgraded detector. The amount of material is reduced especially in the high η region.

4.1.1 Module production

For the Phase 1 upgrade of the CMS pixel detector the entire pixel detector will be replaced and therefore hundreds of modules have to be produced. The first two layers will be built in Switzerland, the third layer by CERN and Italy and the fourth layer in Germany. The University of Hamburg and DESY will build together one half of the fourth layer; thus, 384 modules have to be produced, including spare modules, by the end of 2016. In Table 4.1 an overview of the number of modules per institute is given.

In the following the steps for the module production are briefly described.

- In a first step the silicon sensors are produced and tested with current (IV) and capacitance (CV) measurements by CiS[1]. Afterwards the measurements are repeated by the University of Hamburg.

[1]Forschungsinstitut für Microsensorik GmbH, www.cismst.org/en/

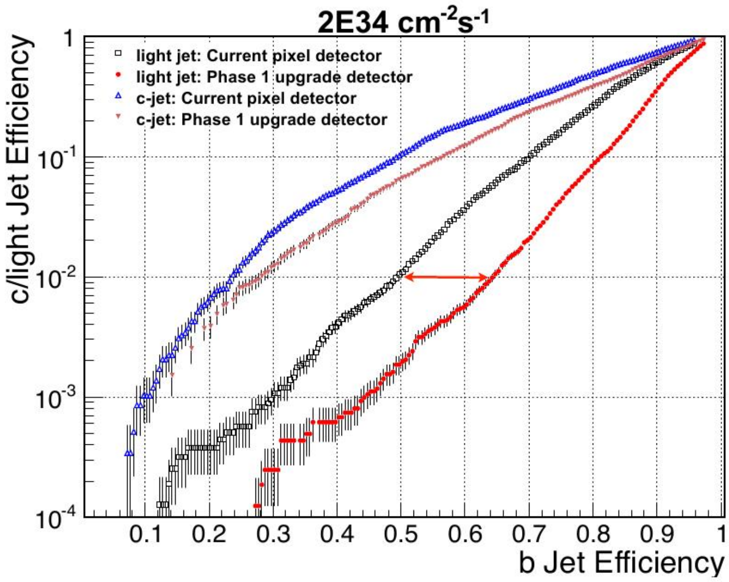


Figure 4.4: Misidentification of b quarks as c or light quarks as a function of the b-tagging efficiency for the current and upgraded detector in simulation [102].

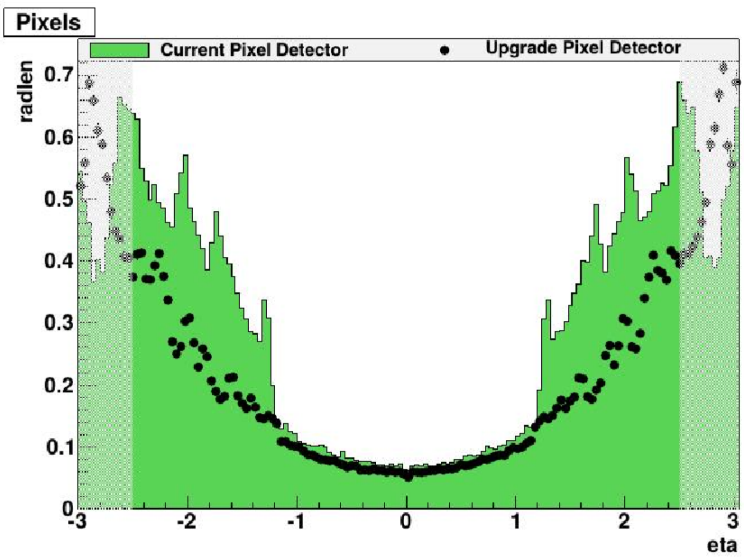


Figure 4.5: Amount of material inside the tracker in radiation lengths as a function of η for the current pixel detector (green) and the upgraded pixel detector (dots) [101].

Layer	# modules	Group
1	96	CH (PSI)
2	224	CH (PSI)
3	352	CERN, IT, TW, FI
4	512	DE

Table 4.1: Overview of the module production at the different institutes.

- In a next step the wafers are diced by PacTech[2] and the Under Bump Metallization (UBM) is done.
- After the dicing the sensor quality is tested again by IV and CV measurements.
- The readout chips are produced by IBM[3] and diced by PacTech, also these are tested before they undergo the next step.
- The sensor gets bump bonded to the 16 readout chips.
- In parallel an HDI functional test is performed and the Token Bit Manager (TBM), is glued to the HDI. The TBM is coordinating the readout of the 16 readout chips.
- In the next step the base strips are glued to the readout chips and the HDI is glued to the sensor.
- The readout chips and the TBM are wire bonded to the HDI.
- In a last step the power and signal cable is installed.

All production steps are done on the DESY campus except the dicing of the wafers and the UBM, these processes are done by external companies. By doing the other steps at one campus the shipping and movement of the modules is minimized to avoid possible damage.

After these steps are done, the module is complete and has to undergo several tests. The module has to withstand several thermal cycles with periods of warming and cooling, in order to simulate the situation in the detector. In a next step the readout chip is configured and a threshold is set to the module. Furthermore, a pulse height calibration is done, which will be explained in the following section. The last calibration is the x-ray calibration which will be explained as well in the following section.

4.2 X-ray calibration procedure

As a first step, before the Vcal injector is calibrated, the module has to be configured. Possible noisy pixels are masked and the threshold for the chip is set as described in Section 3.3.2.

In order to obtain the threshold or the energy deposition in physically meaningful units, the Vcal DAC is calibrated, since the threshold and the energy deposited are measured in units of Vcal.

[2] http://www.pactech.de/
[3] http://www.ibm.com

To calibrate this DAC the sensor can be irradiated with particles with known energy and that deposit their energy inside the sensor. The characteristic x-ray energies of given materials are well known, using the fluorescence x-rays of materials with different energies and the knowledge of how many electron-hole pairs are generated in the detector gives the possibility to express the Vcal DAC in units of electrons. The number of electron-hole pairs N (collected charge) in silicon is calculated with

$$N = \frac{E}{\epsilon} \tag{4.1}$$

$$\epsilon_{\text{silicon}} = 3.6\,\text{eV}, \tag{4.2}$$

where E is the energy of the particles and ϵ the specific energy that is needed to create one electron-hole pair in this material. The targets used for the calibration are Copper, Molybdenum, Silver, Tellurium and Barium (see Table 4.2) . The investigation of the temperature dependence in Section 4.4 is done with a radioactive source and the three targets Mo, Ag and Ba. Whereas in Section 4.5 the four targets Cu, Mo, Ag and Te are used with the fully working x-ray calibration setup.

Target	Energy [keV]	Collected charge in Si [e^-]
Cu	8.05	2236
Mo	17.48	4856
Ag	22.16	6156
Te	27.47	7631
Ba	32.19	8942

Table 4.2: Energy of the K_α line [103] and the corresponding collected charge in silicon.

In Figure 4.6 a hitmap for one readout chip is shown. A radioactive source is used as an x-ray source. The source is placed in the corner of the ROC. Some pixels are masked (white) since they delivered too much noise. The pixels on the edge obtain more hits since they have a larger size.

There is one intermediate step in the x-ray calibration procedure which is called the pulse-height calibration. Since the signal, in this case the characteristic x-rays, is measured in ADC units, there is a second calibration needed. Each pixel has its own ADC and its response differs slightly from pixel to pixel. Therefore, it is possible to convert the response of the ADC for each pixel into a response of an external capacitor Vcal. The external capacitor makes it possible to inject charge into each pixel of the ROC, simulating a real signal, with a selectable signal height.

During the pulse height calibration a signal is injected via the Vcal capacitor. This is done for all possible values of Vcal and for each value of Vcal, the pulse height measured, in ADC units is recorded. The calibration results in a conversion from ADC units to Vcal units.

In Figure 4.7 all steps of the x-ray calibration are shown. The characteristic fluorescence energy is recorded by the detector in ADC units (top right), in a separate measurement the relation between ADC and Vcal units for each pixel is determined (bottom left), in a last step the spectrum gets calculated in Vcal units (bottom right).

The Vcal distributions taken by the pixel detector are fitted with a Gaussian function

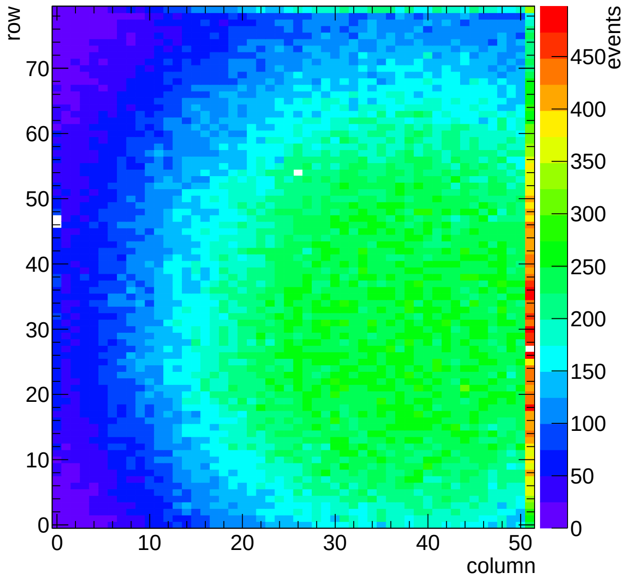


Figure 4.6: Hitmap for a ROC after illuminating with a radioactive source.

in two steps. The first step is to fit the data without any constraints, then the fit is redone using the width σ of the previous fit as an initial range for the new fit. It was found that in this way the fitting procedure is much more stable and not as sensitive as a single Gaussian fit to noise. In Figure 4.8 an example fit for one ROC is presented.

Repeating the fitting procedure with different fluorescence targets, the relation between the collected charge and Vcal units can be obtained as shown in Figure 4.9. For each target, the mean position of the Gaussian fit is taken as a point in the graph, and the width σ of the Gaussian fit is taken as its uncertainty.

The conversion that is found with the linear fit is

$$\text{Collected charge}(Vcal) = m_{\text{slope}} \cdot Vcal - b_{\text{offset}}. \tag{4.3}$$

For the calibration, shown in Figure 4.8, the following values are measured

$$m_{\text{slope}} = 51.07 \, \frac{\text{e}}{\text{Vcal}}, \tag{4.4}$$

$$b_{\text{offset}} = 815.3 \, \text{e}. \tag{4.5}$$

With Equation (4.3) it is possible to calculate the threshold that was set for this ROC in electrons. This is especially important for the track reconstruction in Monte Carlo simulation and therefore for b quark reconstruction. Furthermore, it is possible to observe the energy that is deposited inside the sensor, which can be used to estimate the lifetime of the module, since the radiation damage received by the sensor during the period the module is operated inside the detector decreases the amount of charge that can be collected.

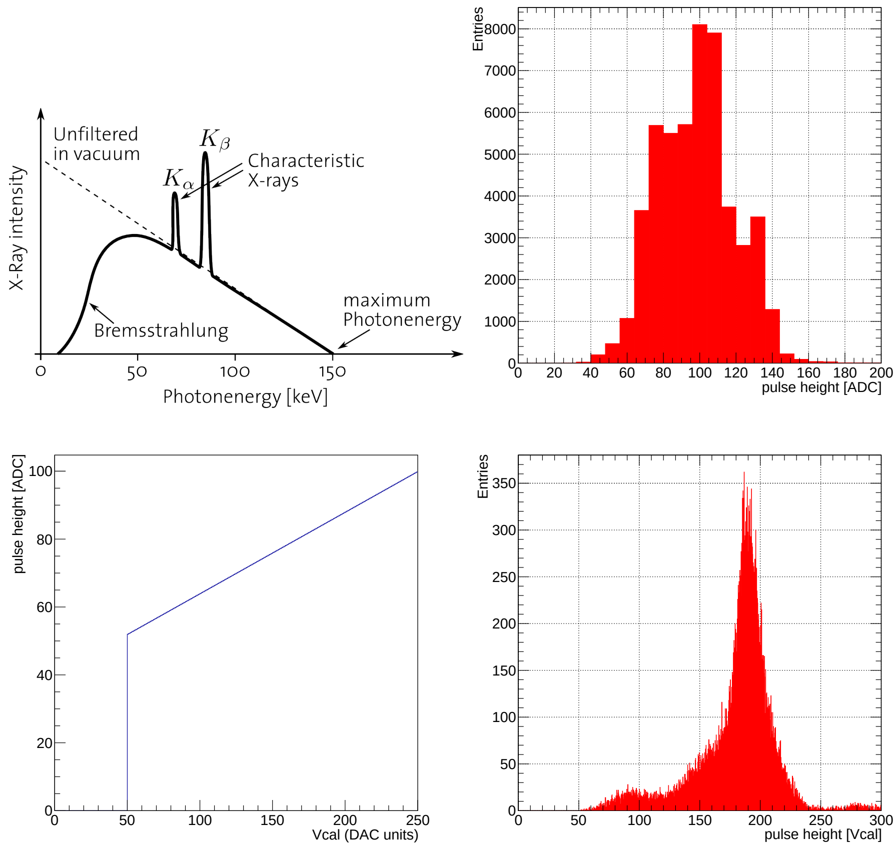


Figure 4.7: Top left: a schematic drawing of a fluorescence spectrum (taken from [104]), top right: the spectrum measured with the pixel detector in ADC units. Bottom left: the conversion between ADC units and Vcal units, which is almost linear. Bottom right: final spectrum in Vcal units using the conversion found in the bottom left graph.

4.3 Experimental setup

In order to obtain a stable and automated calibration for the mass production I developed the software and the corresponding setup. The requirements for the calibrations are:

- Mass production: Automated data taking for three to four different fluorescence materials. That means the targets have to change automatically without any human intervention with the setup.
- Safety: The setup has to be safe and be operable by technicians or students.
- High voltage: On the sensor of the module a bias voltage has to be applied.
- Read out: The test board needs to be connected to the module to read out the data.
- Stable temperature: The ROC has to be cooled and kept at a constant temperature during the calibration process.

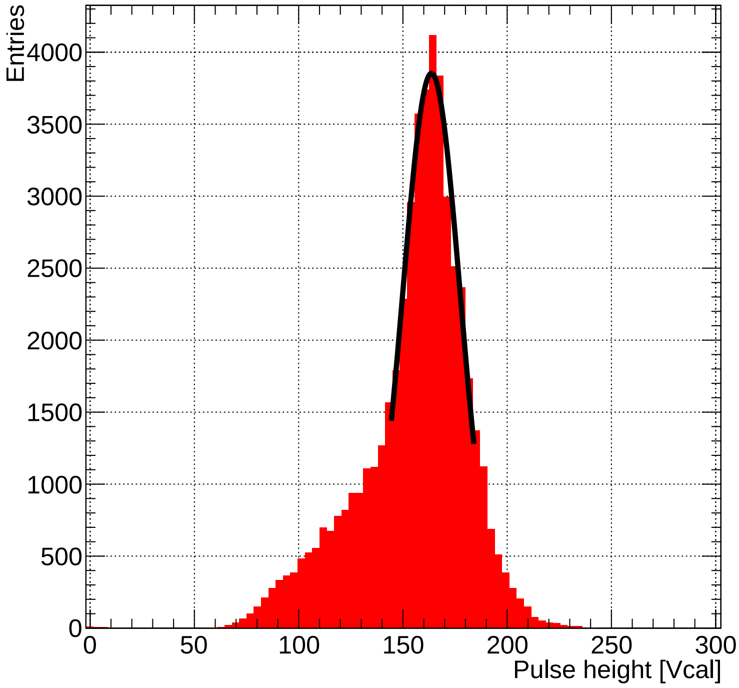


Figure 4.8: Pulse height spectrum in Vcal units fitted with two step Gaussian fit procedure, the χ^2/ndf is 18.7/8. The mean found by the fitting procedure is 163.7 ± 0.1 and the corresponding σ is 13.71 ± 0.15.

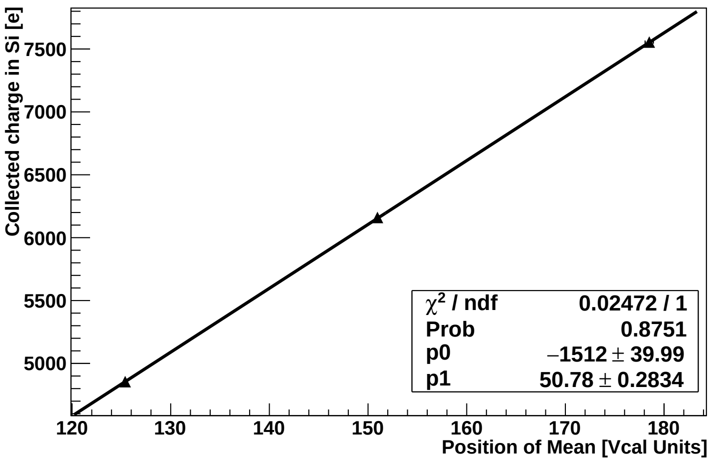


Figure 4.9: Linear fit through the points taken from the Gaussian fit shown in Figure 4.8. The uncertainties of the points are not visible since they are small.

- Dry environment: The environment of the setup should be as dry as possible to prevent condensation on the module.

- High rate test: For an additional test the module has to move into the direct beam.

As an x-ray source, the completely safe and via PC controllable PHYWE x-ray box 4.0 [105] is used, with a maximum voltage of 35 kV and a maximum tube current of 1.0 mA. The box has a target holder in the center for fluorescence targets that can turn by 360° (see Figure 4.10). Since the target has to be at an angle of 45° to the x-ray beam in order to reflect the beam at a 90° angle, one has the possibility to install a maximum of four targets and switch them by turning the target holder by 90°. Another advantage of this box is that it can also be used for student experiments easily when the box is not used for the calibration.

To supply the high voltage to the sensor of the module the voltage supply Keithley 248 is used, which is capable to provide a voltage up to 600 V.

To provide cooling to the module, a cold-block is designed, flushed by water mixed with glycol. A chiller cools down the liquid and pumps it through the block. To achieve fast and better cooling ($T < -20°$C), two peltier elements are installed between the cold-block and the module. Peltier elements are able to cool or to heat by inverting the current, making it possible to steer the heat transport accurately by setting the current. For the voltage support of the peltier elements a Hameg 2020 is used, connected via USB to the PC.

The temperature of the module is measured by placing a PT100 sensor next to the module on the cold-block. The readout of the sensor is done with an Arduino board[4] connected as well via USB to the PC.

Through the different steps of the calibration procedure, the readout chips of the modules are turned on and off and have different power consumptions. Consequently, their heat production is not constant and the change of the temperature has to be compensated quickly by the cooling system. Applying a constant voltage to the peltier elements will not lead to a constant temperature; therefore, the peltier elements have to be regulated according to the measured temperature. To solve this problem a PID (proportional-integral-derivative) controller is designed. In order to achieve stable temperature values the temperature sensor has to be read out fast enough and the voltage of the peltier elements has to be set at least every two seconds. Since the peltier elements transport the heat to their other side, an additional cooling by the chiller is needed. In this setup the temperature of the chiller is always set 5 K lower than the desired-temperature of the module.

During the cooling process condensation can appear on the module, which might cause an electrical short and therefore destroy the module. One has to ensure that the air is extremely dry when cooling down the module to temperatures below the freezing point. To avoid condensation a humidity sensor is installed inside the box. The humidity sensor is also read out by the Arduino board and in every read out loop the dew point T_{dp} is

[4] https://www.arduino.cc/

calculated

$$T_{\text{dp}} = \frac{c\gamma(T, RH)}{b - \gamma(T, RH)}, \tag{4.6}$$

$$\text{where } \gamma(T, RH) = \ln\left(\frac{RH}{100}\right) + \frac{bT}{c+T} \tag{4.7}$$

and T is the measured temperature, RH is the measured relative humidity and b (=17.67) and c (= 243.6°C) are constants.

Once the humidity is low enough (desired-temperature higher than the dew point), the cooling process can start.

In order to keep the module as dry as possible and to be able to use a low flush of dry air, a plastic cap is designed, covering only the module.

For another test (high-rate test), that will be performed after each x-ray calibration the module has to be placed in the direct beam. In order to execute this test, the arm that is holding the module can be rotated by 90°. To start the high-rate test the target holder has to be removed manually. In Figure 4.10 both configurations are shown in a schematic drawing; furthermore, a photo of the described setup is shown.

The readout of the module is done for this analysis with the PSI analog test board and the associated software PSI 46 [106].

In order to automate the calibration procedure, a software is developed to steer and control all the devices and readout the sensors. The measuring procedure is done in the following steps:

- dry the environment,
- cool the module to the desired temperature, therefore cool the chiller to 5 K lower and use the PID algorithm to stabilize the temperature on the module with the peltier elements,
- move the target holder to the desired fluorescence target,
- start the measurement,
- change target and start a new measurement.

4.4 Temperature dependence

As the planned operation temperature of the new pixel detector is -20°C, one of the challenges is to do the calibration at the planned operation temperature. However, this is only necessary if the x-ray calibration is dependent on the temperature, meaning that the linear behavior found in Equation (4.3) changes for different temperatures. Doing the calibration at -20°C is much more time consuming and complicated than doing it at room temperature. For the current analog chip the temperature dependence of the calibration was investigated by KIT[5] [107]. In Figure 4.11 the measurement of KIT is shown. On the x-axis the temperature is shown and on the y-axis the slope of the linear fit. The slope changes with the temperature, for this reason the temperature dependence of the x-ray calibration is tested with the new digital readout chip.

[5] Karlsruher Institut für Technologie

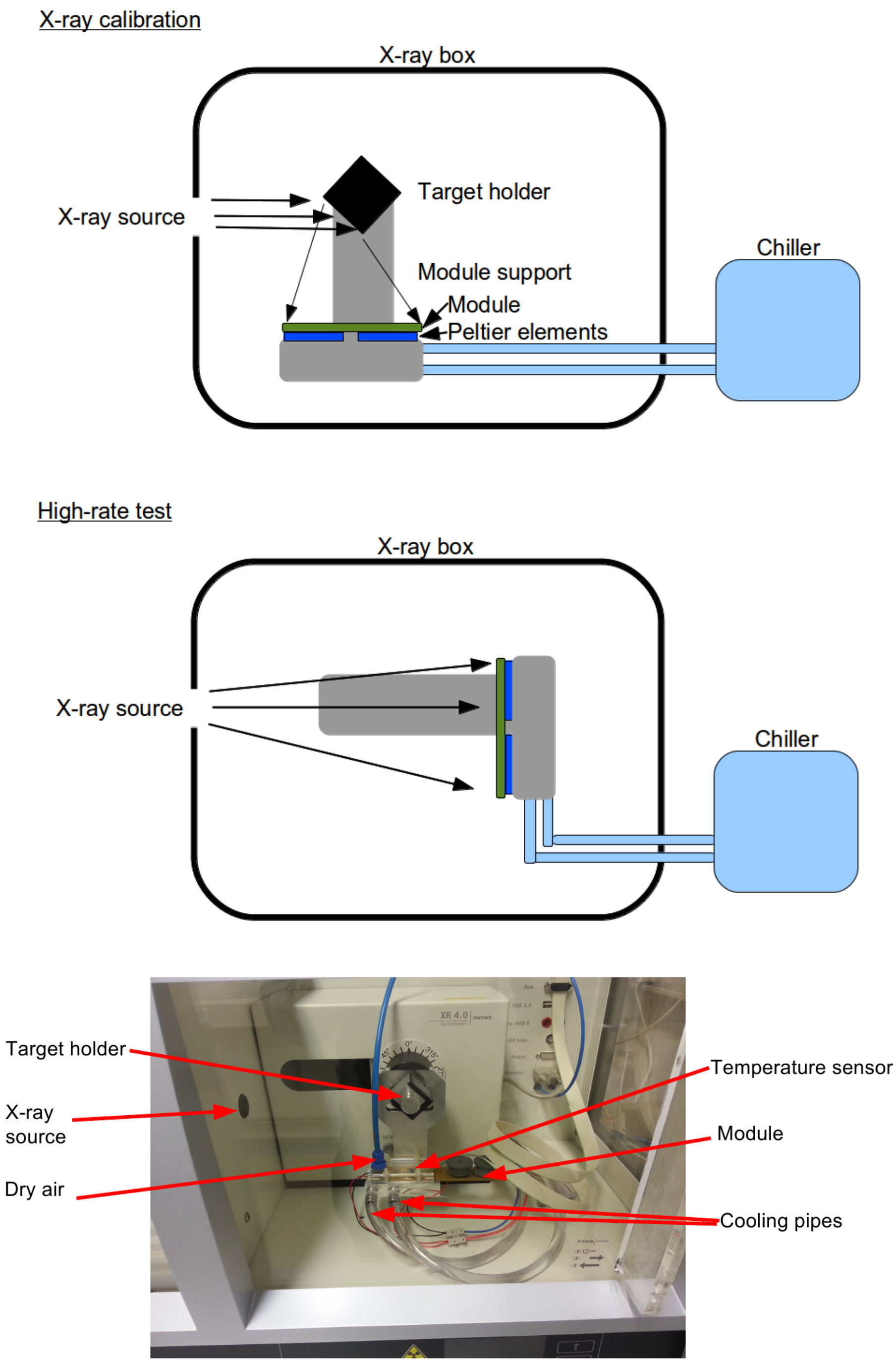


Figure 4.10: Schematic drawing of the setup configured to do the x-ray calibration (top), to do the high-rate test (middle) and photo of the setup (bottom). Here also the plastic cap of the module is shown.

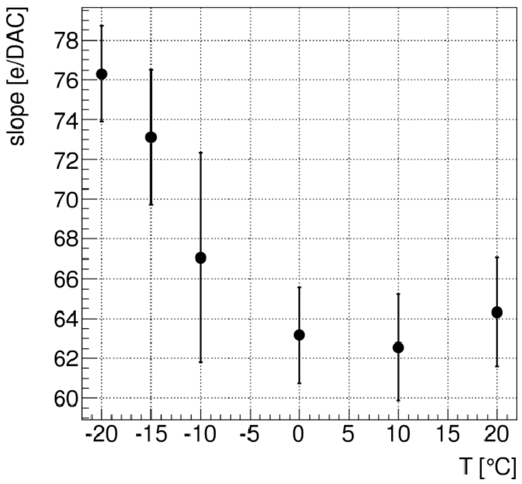


Figure 4.11: Temperature dependence observed for the analog readout chip by KIT [107].

To repeat the measurement with the digital readout chip, the setup is slightly modified. Since there were no full digital modules available, the measurements are done with a single chip module. A single chip module is a single ROC bump bonded to a small sensor.

The dependence of the temperature is investigated with the following steps:

1. wait for the temperature to stabilize,
2. do the pulse height calibration,
3. set the threshold of the readout chip to 30 Vcal units,
4. take data with three targets (Mo, Ag, Ba),
5. lower the temperature by 5 K and go back to step 1.

In Figure 4.12 the mean of the pulse height distribution in ADC units as a function of the temperature for the new digital readout chip is shown. The mean of the pulse height spectrum is shown before the pulse height calibration is applied. The mean of the pulse height spectrum is observed with the same fitting procedure, explained in the previous section. The uncertainty on the temperature is estimated to ± 1 K due to temperature fluctuations of the ROC that can not be regulated fast enough by the PID algorithm.

A clear dependence on the temperature is observed. With increasing temperature the pulse height distribution moves to higher values. Furthermore, the three targets show almost the same behavior with different offsets. This leads to the hypothesis that the linear calibration fit using the three targets might be independent of the temperature.

In Figure 4.13 the dependence of the mean of the Gaussian after the pulse height calibration at each temperature step is shown. The mean is given in Vcal units. The dependence observed in Figure 4.12 vanishes in Figure 4.13.

The linear fits for each temperature step after the pulse height calibration are shown in Figure 4.14. The slope differs only slightly whereas the offset changes more. This is caused by the linear fit through only three points. Little shifts of only one of these points can cause a significant change of the offset.

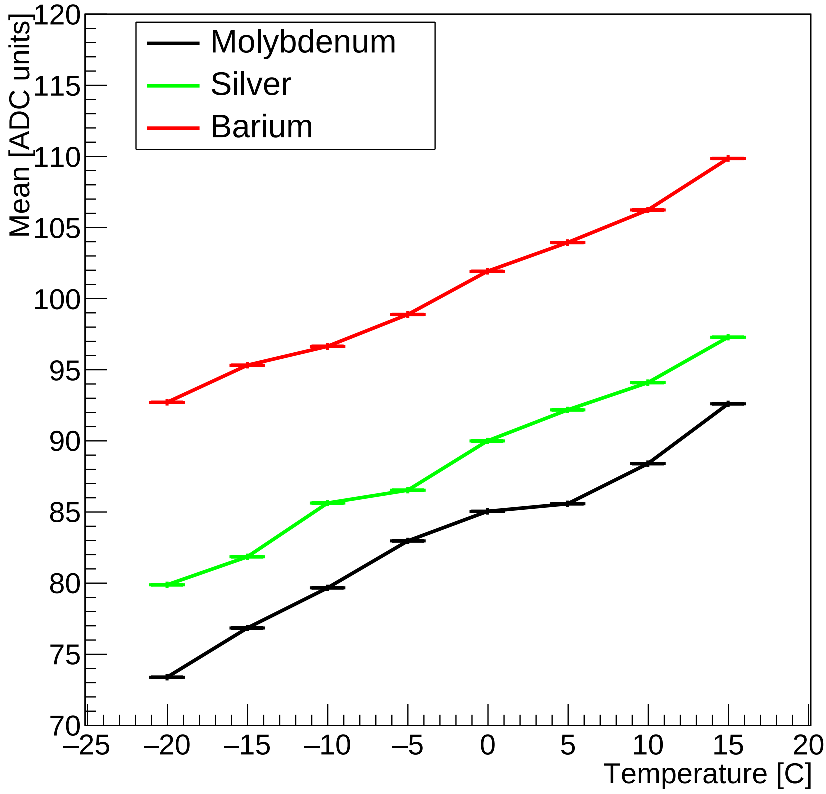


Figure 4.12: Gaussian mean in ADC units (before the pulse height calibration) as a function of the temperature for three different targets. The mean shifts with the change of temperature.

Figure 4.15 concludes the measurement on the temperature dependence of the calibration procedure. It is observed that the slope varies by ± 2e/Vcal but is not dependent on the temperature. The offset fluctuates stronger, which might be caused by statistical fluctuations or the usage of only three fitting points. It will be shown in Section 4.5 that this can be stabilized. Summarizing, it is possible to calibrate the modules at room temperature without the necessity to measure at the planned operating temperature of -20°C.

4.5 Stability and commissioning

Since the calibration will be done for 384 modules in a mass production, the goal is to achieve reliable and reproducible results, meaning that the calibration procedure itself should be very stable. For this reason the calibration of a single chip module is done several times at a constant temperature of 15°C on different days, as presented in Figure 4.16. For the average and the standard deviation the following values are measured:

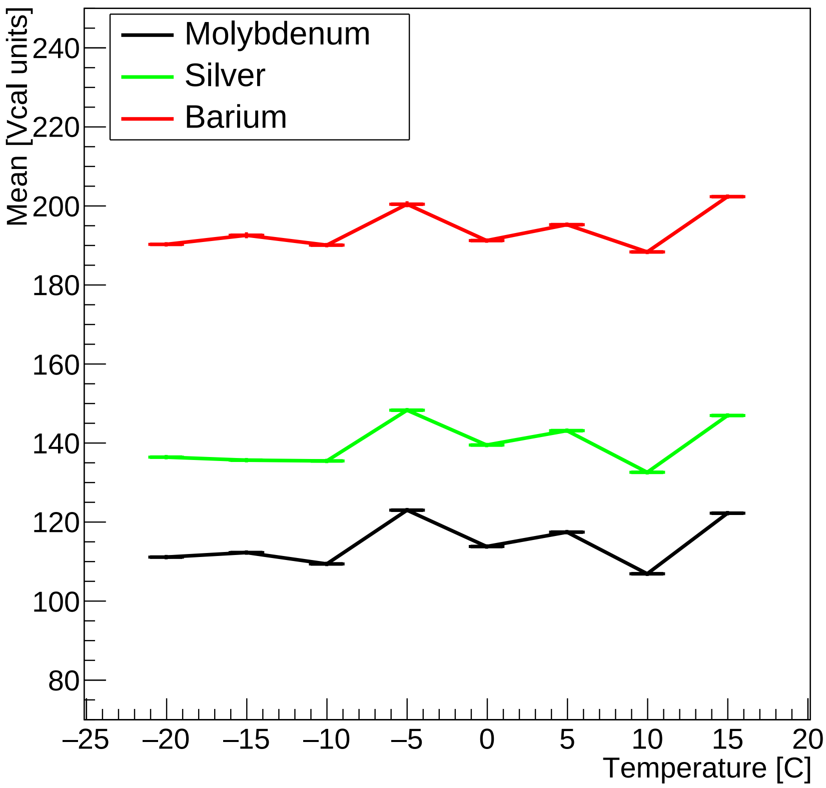


Figure 4.13: Gaussian mean in Vcal units as a function of the temperature after the pulse height calibration. No clear trend is observed.

$$m_{\text{slope}} = (53.9 \pm 1.1)\,\frac{\text{e}}{\text{Vcal}}. \tag{4.8}$$

The procedure is not stable with a large standard deviation.

In an earlier study it was found that the calibration is sensitive to the received rate by the ROC [107]. With increasing rate the position of the pulse height spectrum in Vcal units shifts to higher values. With this knowledge and the knowledge that different fluorescence targets lead to different rates measured at the module one can adjust the intensity of the x-rays by adjusting the tube current and therefore compensate the differences. In Ref. [108] it is shown that the most recent version of the readout chip is not sensitive anymore to the receiving rate.

In the automated measuring procedure another step is added, measuring the rate of the collected data for different x-ray tube currents and setting for each target the tube current, so that for all fluorescence targets the rate at the module is equal. Consequently, for each target a scan of the tube current is performed, also completely automatically. This scan has to be done only once after the threshold of the module is set and before the calibration procedure starts. The tube currents corresponding to their targets are shown in Table 4.3, the rates are adjusted to $20\,\text{kHz/cm}^2$. The systematic measurement is repeated

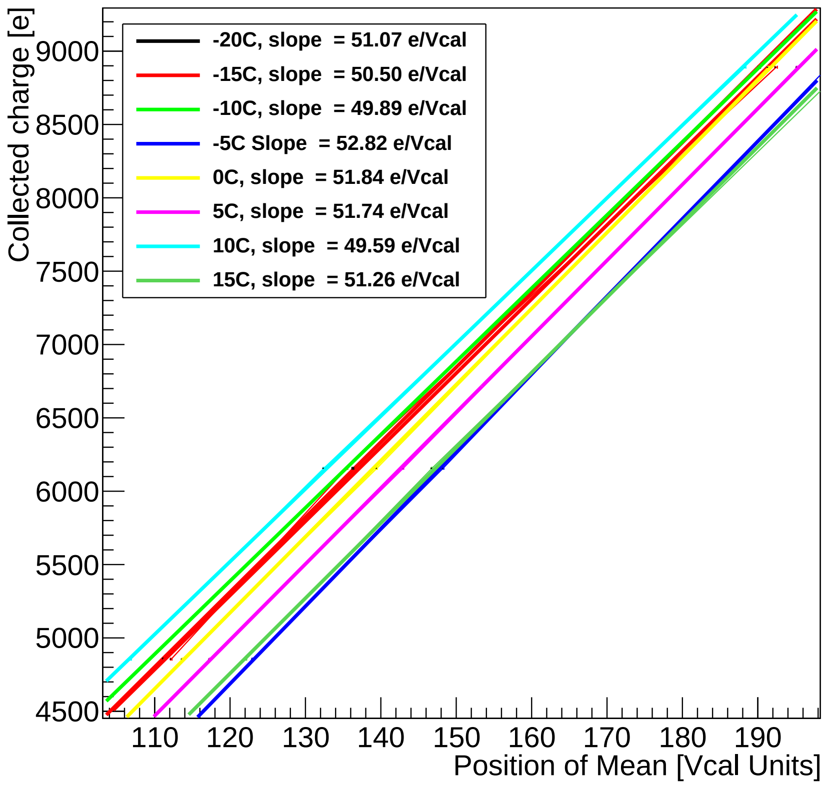


Figure 4.14: Linear fits of three different energies (Mo, Ag, Ba) for different temperatures.

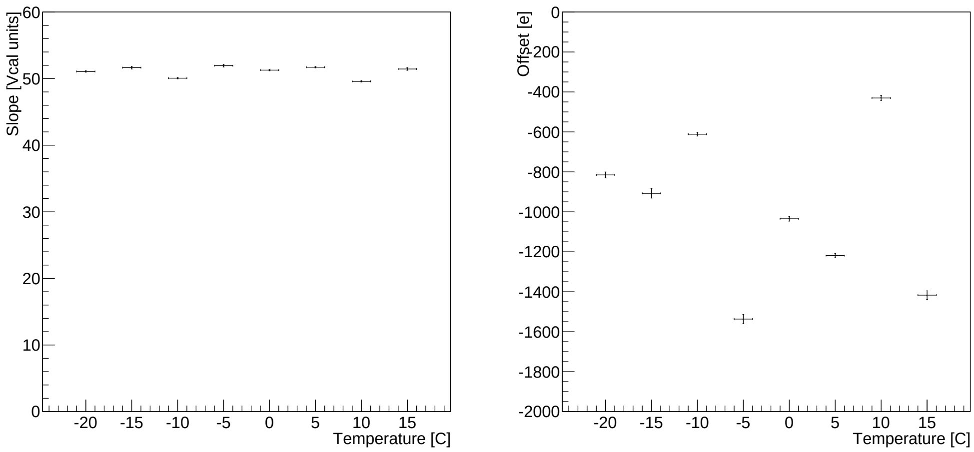


Figure 4.15: The slope (left) and the offset (right) as a function of the temperature.

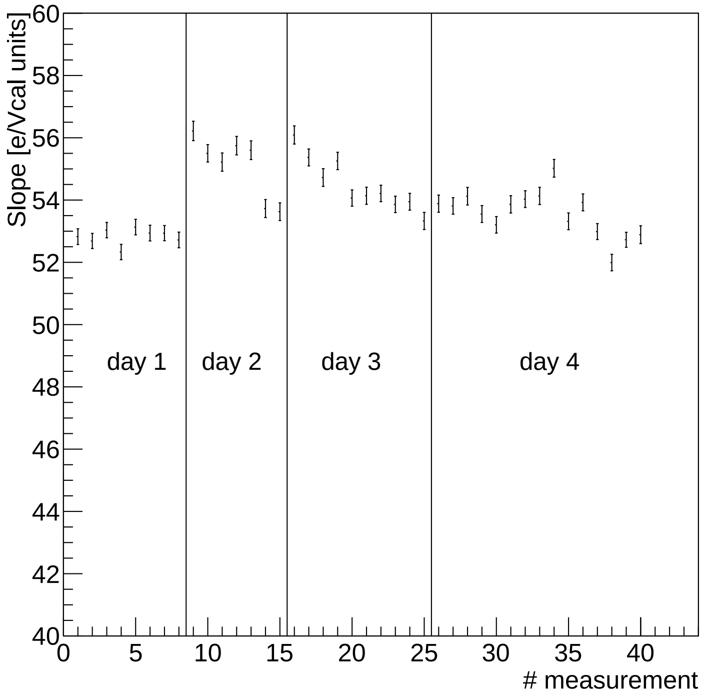


Figure 4.16: Stability test: The slope of the linear fit as a function of the number of measurements.

Target	Tube current [mA]
Cu	0.08
Mo	0.12
Ag	0.35
Te	1.0

Table 4.3: X-ray tube currents to adjust the rate for each target to $20\,\mathrm{kHz/cm^2}$.

after the tube current is adjusted (see Figure 4.17). The fluctuation is much smaller and the adjustment helps to get more reliable results. After the tube current adjustment the standard deviation is smaller and the measurement becomes more precise,

$$m_{\text{slope}} = (51.6 \pm 0.6)\,\frac{\mathrm{e}}{\mathrm{Vcal}}. \tag{4.9}$$

In order to achieve higher accuracy, the calibration is done with four targets instead of three. Copper is chosen as a fourth target, since it has a low energy K_α line and is significantly far away from the energies of the other three targets. Crosstalk reduction of the new digital readout chip allows to obtain lower energies. In Figure 4.18 the pulse height spectrum in Vcal units for all fluorescence targets and the calibration fit is shown. Furthermore, the pulse height distributions for all targets before and after the tube current adjustment is displayed. After the tube current adjustment, the integral of all distributions

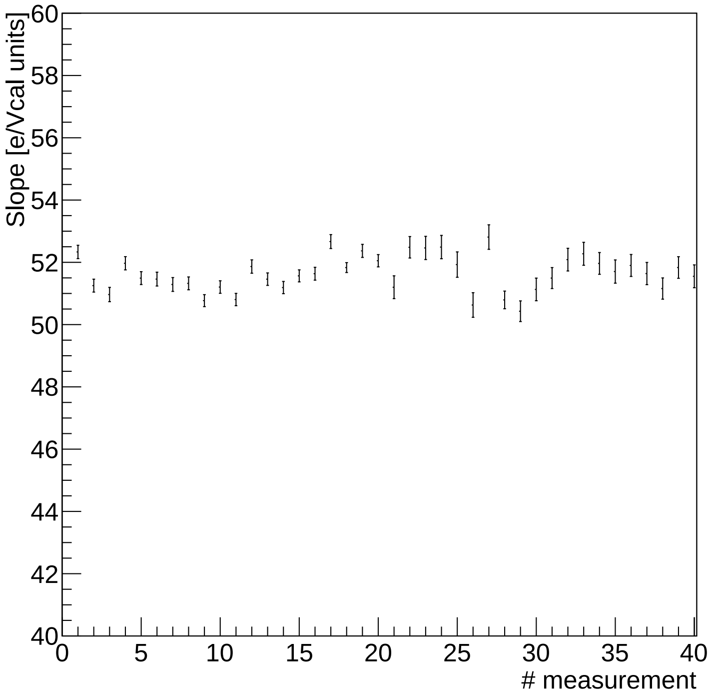


Figure 4.17: Stability of the automated measurements over three days after the adjustment of the rate.

is similar.

The four fit points make the linear fit more stable and little shifts of one of the points do not affect the calibration anymore. In Figure 4.19 it is shown that the procedure becomes stable after adding another target to the process. The standard deviation is reduced by another factor of two

$$m_{\text{slope}} = (51.7 \pm 0.3)\,\frac{\text{e}}{\text{Vcal}}. \tag{4.10}$$

The results have better reproducibility now and satisfy the requirements for the mass production of modules.

4.6 Conclusion

For the Phase 1 pixel upgrade and the replacement of the complete pixel detector a setup was developed to calibrate the internal Vcal DAC of the ROC with characteristic x-rays. Software and hardware were designed in a way that the calibration is done completely automated for the mass production. Moreover, it was shown that the x-ray calibration is not sensitive to the temperature. This makes it possible to process the calibration at room temperature, avoiding a cooling down process to the planned operating temperature of -20°C. In Section 4.5 the stability of the setup was tested, since the results for the

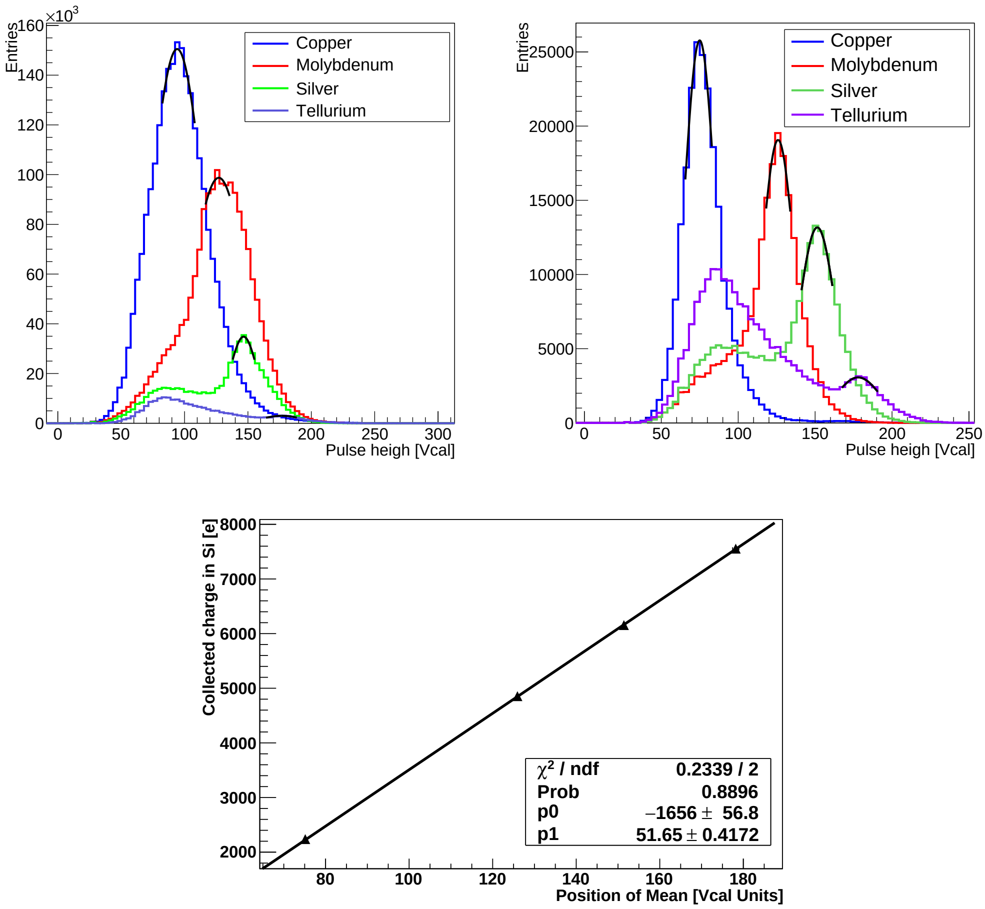


Figure 4.18: Pulse height spectrum for all fluorescence targets in Vcal units before the tube current adjustment (top left) and after (top right); calibration fit for all four fluorescence targets (bottom).

mass production of modules are required to be reliable. By adjusting the rate for each fluorescence target and performing the calibration with four fluorescence targets instead of three, the calibration process was significantly improved and stabilized. In this way, the calibration procedure is reasonably fast and the goal of two module calibrations per day is achieved.

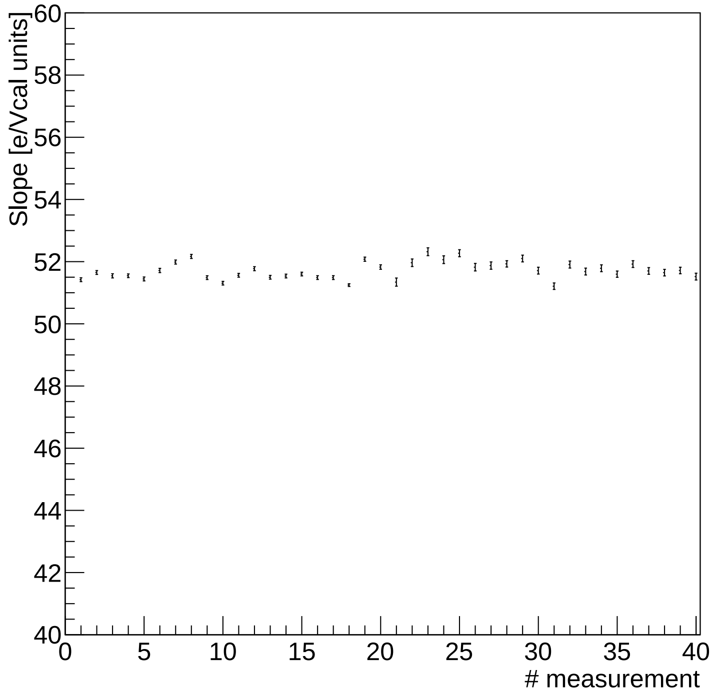


Figure 4.19: Stability of the automated measurements over three days after the adjustment of the rate and after adding a fourth target to the calibration.

5 Event simulation

In high energy physics, the simulation of colliding particles and their interactions and decays has a major importance since it provides the comparison between standard model predictions and measured data. Simulated events are used to interpret and validate actual collisions. Deviations from the standard model can indicate new physics beyond the standard model. Furthermore, one can also simulate the kinematics of possible new physics and can therefore exclude models or derive limits on models.

Simulated events are also used in order to improve measurements through new detector designs as it was also done for the Phase 1 upgrade described in Section 4.1. Event simulation can be used further as benchmarks for new detector concepts. Moreover, in simulated events the information about initial and final state particles is available (generator level) and can be used to explore the performance of top quark identification methods as can be seen in Section 9.

The simulation of high energy collisions is challenging since in each collision several hundreds of particles are created with a momentum distributed over a few orders of magnitude. Since the collisions of protons is a probabilistic process, numerical Monte Carlo (MC) techniques are used, creating pseudo-random numbers according to the probability functions from physics models.

There are different steps in the simulation of an event. In Figure 5.1 an event containing all steps from the hard scattering to the hadronization and the decay is shown. The event also contains an underlying event.

- Hard scattering: The hard scattering is the first step in the simulation chain and is responsible for the nominal parton-parton interaction. Two partons of the two protons interact with a momentum transfer Q. The probability for individual partons to take part in the hard interaction is parametrized with the parton distribution function (PDFs), which have been measured precisely at HERA [30]. The cross section for a given process can be calculated by the factorization theorem in Equation (2.29). Furthermore, since the energy transfer is high in comparison to the cutoff scale μ the calculation can be done perturbatively since the coupling constant of QCD α_S becomes significantly smaller than one.

- Parton shower and matching: In this step the final state particles, forming showers, are simulated. Furthermore, in this step initial and final state radiation are simulated, which correct for higher order contributions that were not addressed in the matrix element calculation. In initial state radiation (ISR) partons are radiated off before the hard scattering takes place, whereas the final state radiation (FSR) takes place afterwards.

  Since additional partons can be derived with matrix element calculations or arise from the parton showers, a double-counting is possible. In order to avoid double-counting these events get rejected by a matching algorithm (MLM algorithm [110]).

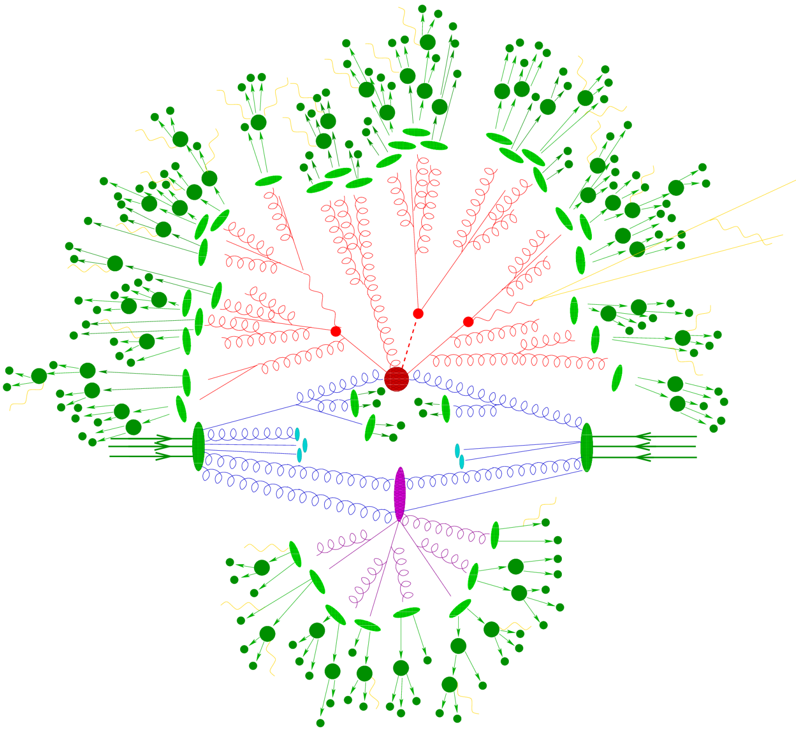


Figure 5.1: Illustration of the simulation steps of a proton-proton collision: In the center the hard scattering takes place, for the underlying event and for the main process. Several initial and final state radiations can be seen as well as hadronization and decay [109].

The algorithm clusters jets from all generated partons. Afterwards these jets are matched (above a certain p_{T}, which is the matching scale) to the partons from the matrix element calculations. If each matrix element parton matches exactly to one jet the event is kept, otherwise rejected. The rejection of an event is an approximate way to introduce a Sudakov form factor [111] to the hard process. In summary the parton shower should not generate emissions that would be double counted and were already generated by the matrix element calculations. Another method to avoid double counting is the CKKW algorithm explained in Ref. [112]. The DGLAP equations [113–115] are used in order to model the evolution of partons down to a scale of $\alpha_s \approx 1$. Non-perturbative hadronization starts at these scales.

- Hadronization: In the next step the transition from the final state partons to hadrons is simulated. In simulation this step can not be solved with perturbative calculations, since it takes place at a low energy scale. There are two main methods to model the hadronization: The cluster model [116, 117] and the string model [118]. Each of these models is used for different event generators. The cluster model splits all remaining gluons in quark and anti-quark pairs and afterwards they are combined to colorless clusters. These clusters can either decay to lighter clusters and form hadrons or directly form hadrons. The string model assumes that the potential between two partons increases linearly with their distance. If the distance between the two partons is high enough, a quark-anti quark pair is formed and can split further until the strings ends up as an on-shell hadron. In the last step the decay of unstable hadrons has to be modeled.

- Underlying event: In addition to the initial partons of the hard scattering other partons with smaller momentum transfer can interact. These additional interactions are called underlying events (UE). Underlying events can increase the number of particles and therefore enlarge the hadronic activity in the event.

- Pile-up: Due to high luminosities multiple proton-proton scatterings can take place in one bunch crossing. These interactions can introduce additional activity in the detector.

- Detector simulation: In a last step the detector and the interaction with the particles are simulated. This simulation is usually done with GEANT4 [119] with a detailed model of the CMS detector. Here the detector geometry, the magnetic field and the detector readout are simulated as well as the showers in the calorimeter. Afterwards the simulated events are saved in the same format as the data and provide therefore a good comparability. Due to computational power there are two types of detector simulation, the full simulation and the fast simulation. The full simulation includes all details of the detector and consumes a high amount of computing time, whereas the fast simulation consists of a more simple model of the detector with parametrized detector responses. Several jet related quantities can be calculated with almost the same accuracy in fast simulation as in full simulation.

5.1 Event generators

In the following the different event generators used in this thesis are introduced and explained:

- PYTHIA [120]: A commonly used and multi purpose event generator is PYTHIA. The generator can simulate all standard model processes and various models containing new physics. PYTHIA is capable of simulating up to $2 \to 3$ processes. Initial and final state radiation is modeled with the parton shower, which is p_{T} ordered. A string fragmentation model is used for the hadronization [118]. The initial hard scatter in PYTHIA is calculated at leading order.

- HERWIG++ [121]: Similar to PYTHIA, the event generator HERWIG is capable of simulating all standard model processes. It calculates the hard scattering at leading order. The main difference between the two generators is the hadronization, which is modeled in HERWIG with a cluster fragmentation model [116,117]. Furthermore, in HERWIG++ the parton shower is angular-ordered.

- POWHEG [122]: POWHEG is an event generator especially designed and optimized for proton-proton collisions. It calculates the matrix elements at next-to-leading order. Consequently, it includes the production of one more hard parton at matrix element level. In general POWHEG is not capable of simulating all SM processes, like MadGraph or PYTHIA.

- MadGraph [123]: Similar to POWHEG, MadGraph is specialized for proton-proton collisions. In principle it is able to calculate all processes $2 \to n$ at tree level or NLO with an unlimited number of final state particles. Since the complexity of the calculations increases strongly with the number of final state particles and therefore also the time consumption, the maximum number of final state particles is in general not higher than eight. Furthermore, the matrix element calculation includes additional hard partons.

- MC@NLO [124]: The MC@NLO formalism allows the incorporation of accurate NLO QCD calculation into the parton shower. It can be interfaced to both PYTHIA and HERWIG++. All standard model processes can be generated with MC@NLO, as well as some BSM processes.

5.2 Monte Carlo samples

In Table 5.2 all Monte Carlo samples used in this thesis are listed. For the top tagging studies in Monte Carlo simulation (Section 9) the $\mathrm{t\bar{t}}$ samples and the PYTHIA QCD multijet sample generated with a flat p_{T} spectrum are used. For the analysis in 8 TeV data (Section 10) all standard model predictions are used. In order to investigate tagging algorithms also in a high p_{T} range two $\mathrm{Z}' \to \mathrm{t\bar{t}}$ samples are used as a source for high energetic top quarks (Section 11). The processes are explained in the following:

- $\mathrm{t\bar{t}}$: This process describes top quark pair production as described in Section 2.3.1, which can be produced in association with additional hard jets. The top quark pair production simulation is done by interfacing POWHEG to PYTHIA 6.4.24. Another $\mathrm{t\bar{t}}$ sample is produced by interfacing MC@NLO to HERWIG++.

- Single top: As described in Section 2.3.1 top quarks can be produced also as single particles in association with an anti bottom quark (s-channel), a quark from the first or second generation (t-channel) as well as a W boson. To simulate single top production POWHEG is interfaced to PYTHIA 6.4.24.

- QCD: QCD multijet processes are initiated by the strong interaction. At the LHC QCD multijet events are the dominating background in many analyses. There are two types of samples used in the following analysis. One set of QCD multijet samples is muon enriched, meaning that an additional muon is produced in association with the QCD process, divided into different p_{T} ranges. The muon enriched samples are produced with PYTHIA 6.4.24. For comparison in this analysis QCD processes produced with HERWIG++ are used as well.

- Drell Yan (DY): The Drell Yan process describes the annihilation of two quarks from the scattering of two hadrons. Two leptons with opposite charge are produced through a photon or a Z boson.
  The Drell Yan samples are produced with MadGraph.

- W+jets: This process describes the leptonically decaying W boson in association with N hard jets. These samples contain only leptonically decaying W bosons, since the cross section of QCD multijet background processes is approximately three orders of magnitude higher. These samples are simulated with MadGraph as well.

- Vector boson pair production: The fusion of two gluons or quarks is responsible for vector boson pair production. These MC samples are produced with the event generator PYTHIA 6.4.24.
- Z': A possible new particle Z', predicted from several BSM models [51–59], decaying into a top quark pair, is simulated with different hypothetical masses and widths. This is done with MadGraph.

All samples have an additional tuning for the underlying event simulation, which is indicated in the sample names with "TuneZ2Star". These tunes use the PDFs provided by the CTEQ collaboration [125]. For the single top production the CT10 PDFs [126] are used, for the $t\bar{t}$ the CTEQ6M PDFs [127] are used and for the PYTHIA QCD multijet sample the CTEQ61L PDFs [125] are used, whereas the HERWIG++ QCD sample is generated using the MRST2001 PDFs [128]. In order to simulate the decay of τs PYTHIA and HERWIG++ are interfaced to TAULOA 27.121.5 [129].

Process	Monte Carlo samples (*/Summer12_DR53X-PU_S10_START53_V7A-v1/AODSIM)
DY + 1 jet	DY1JetsToLL_M-50_TuneZ2Star_8TeV-madgraph
DY + 2 jets	DY2JetsToLL_M-50_TuneZ2Star_8TeV-madgraph
DY + 3 jets	DY3JetsToLL_M-50_TuneZ2Star_8TeV-madgraph
DY + 4 jets	DY4JetsToLL_M-50_TuneZ2Star_8TeV-madgraph
W + 1 jet	W1JetsToLNu_TuneZ2Star_8TeV-madgraph
W + 2 jets	W2JetsToLNu_TuneZ2Star_8TeV-madgraph
W + 3 jets	W3JetsToLNu_TuneZ2Star_8TeV-madgraph
W + 4 jets	W4JetsToLNu_TuneZ2Star_8TeV-madgraph
WW	WW_TuneZ2star_8TeV_pythia6_tauola
WZ	WZ_TuneZ2star_8TeV_pythia6_tauola
ZZ	ZZ_TuneZ2star_8TeV_pythia6_tauola
Single top	T_t-channel_TuneZ2star_8TeV-powheg-tauola
Single top	T_tW-channel-DR_TuneZ2star_8TeV-powheg-tauola
Single top	T_s-channel_TuneZ2star_8TeV-powheg-tauola
Single top	Tbar_t-channel_TuneZ2star_8TeV-powheg-tauola
Single top	Tbar_tW-channel-DR_TuneZ2star_8TeV-powheg-tauola
Single top	Tbar_s-channel_TuneZ2star_8TeV-powheg-tauola
$t\bar{t}$	TT_CT10_TuneZ2star_8TeV-powheg-tauola
$t\bar{t}$	TT_Mtt-700to1000_CT10_TuneZ2star_8TeV-powheg-tauola
$t\bar{t}$	TT_Mtt-1000toInf_CT10_TuneZ2star_8TeV-powheg-tauola
$t\bar{t}$	TT_8TeV-mcatnlo
QCD	QCD_Pt-15to20_MuEnrichedPt5_TuneZ2star_8TeV_pythia6
QCD	QCD_Pt-20to30_MuEnrichedPt5_TuneZ2star_8TeV_pythia6
QCD	QCD_Pt-30to50_MuEnrichedPt5_TuneZ2star_8TeV_pythia6
QCD	QCD_Pt-50to80_MuEnrichedPt5_TuneZ2star_8TeV_pythia6
QCD	QCD_Pt-80to120_MuEnrichedPt5_TuneZ2star_8TeV_pythia6
QCD	QCD_Pt-120to170_MuEnrichedPt5_TuneZ2star_8TeV_pythia6
QCD	QCD_Pt-170to300_MuEnrichedPt5_TuneZ2star_8TeV_pythia6
QCD	QCD_Pt-300to470_MuEnrichedPt5_TuneZ2star_8TeV_pythia6
QCD	QCD_Pt-470to600_MuEnrichedPt5_TuneZ2star_8TeV_pythia6
QCD	QCD_Pt-600to800_MuEnrichedPt5_TuneZ2star_8TeV_pythia6
QCD	QCD_Pt-800to1000_MuEnrichedPt5_TuneZ2star_8TeV_pythia6
QCD	QCD_Pt-1000_MuEnrichedPt5_TuneZ2star_8TeV_pythia6
QCD	QCD_HT-100To250_TuneZ2star_8TeV-madgraph-pythia
QCD	QCD_HT-250To500_TuneZ2star_8TeV-madgraph-pythia6
QCD	QCD_HT-500To1000_TuneZ2star_8TeV-madgraph-pythia6
QCD	QCD_HT-1000ToInf_TuneZ2star_8TeV-madgraph-pythia6
QCD	QCD_Pt-15to3000_TuneEE3C_Flat_8TeV_herwigpp
QCD flat	QCD_Pt-15to3000_TuneZ2star_Flat_8TeV_pythia6
Z'	ZPrimeToTTJets_M4000GeV_W40GeV_TuneZ2star_8TeV-madgraph-tauola
Z'	ZPrimeToTTJets_M3000GeV_W30GeV_TuneZ2star_8TeV-madgraph-tauola
Z'	ZPrimeToTTJets_M3000GeV_W300GeV_TuneZ2star_8TeV-madgraph-tauola

Table 5.1: Monte Carlo samples used in this thesis, the CMS notation is used.

6 Object reconstruction and particle identification

The data that is collected with the CMS detector is stored in the same format as the output of the detector simulations, in this way they can be compared directly to each other. The first step of the analysis is to define and reconstruct the physics objects. In the following section the high-level object definition and reconstruction methods will be explained.

6.1 Primary vertex

The main primary vertex from the hard interaction is the vertex with the highest transverse momentum $\sum_{\text{tracks}} p_{\text{T}}^2$, which is calculated from the associated tracks. The tracks are clustered with a deterministic annealing algorithm [130] and each vertex has to satisfy the following criteria:

- To have the primary vertex in the center of the collision region the primary vertex has to fulfill $\sqrt{x^2 + y^2} < 2\,\text{cm}$ and $|z| < 24\,\text{cm}$,
- the weighted number of tracks has to be $N_{\text{DOF}} > 4$.

6.2 Particle flow algorithm

Since particles identification is challenging they need to be reconstructed by taking into account their characteristic behavior. For example a muon is bent by the strong magnetic field and is the only particle that reaches the outermost part of the detector. Electrons cause tracks in the silicon tracker and a shower in the electromagnetic calorimeter. In Figure 6.1 a cross-section of the CMS detector is shown and the interactions with particles are illustrated.

In CMS the Particle Flow algorithm [132] is used to reconstruct a full event taking into account all the detector components. This algorithm will be described in the following.

The Particle Flow algorithm can identify most of the possible particles like electrons, photons, charged and neutral hadrons and muons. For all particles the four-momentum is reconstructed by combining information from all detector components.

First fundamental objects in the subdetectors are analyzed. Charged particles in the silicon tracker are reconstructed with an iterative tracking algorithm [133]. Hits in the pixel detector, for example two or three hit clusters, are extrapolated to other layers of the tracker. The algorithm takes care of possible multiple scattering or the possible energy loss in material. The reconstructed track is removed by the algorithm and the seed criteria are loosened for further track reconstruction. In total the reconstruction efficiency for muons is about 99.5% and around 90% for charged hadrons with a misidentification rate of around 1%. However, the algorithm is not able to reconstruct charged particles with a transverse momenta below 150 MeV. In a second step the energy deposits in the calorimeters are

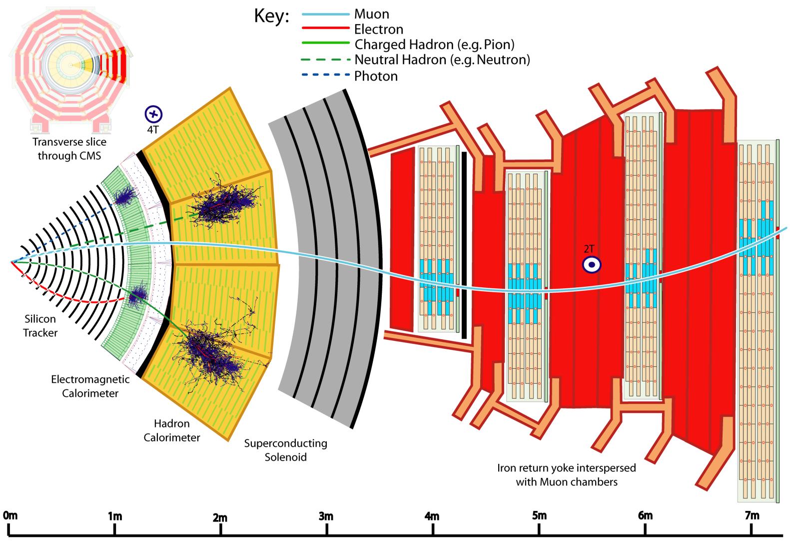


Figure 6.1: Cross-section of the CMS detector, particles with their characteristic behavior in the detector can be seen [131].

clustered as well, but in each component of the calorimeter separately. The seeds of the clustering have to be at least two standard deviations above the electronics noise.

In a third step the reconstruction made in the tracker system and the reconstruction of the energy deposits in the calorimeters have to be merged to avoid double counting of particles. As a first approach to link the tracks from the tracker to the clusters in the calorimeters the last hit in the tracker is extended one interaction length into the hadronic calorimeter. Afterwards all clusters in the calorimeters hit by the extended track are associated to the track. Since the track itself has an uncertainty due to gaps between cells, multiple scattering of low momentum particles and cracks between modules, the clusters in the calorimeters that are associated to the track can be extended by one cell width. In one last step a χ^2 fit is used to combine the tracks from the tracker with the ones from the muon system.

Since now all tracks in the different subdetectors are linked to clusters in the calorimeter the identification of the particles can start. Muons are the only particles that create tracks in the outermost part of the detector, the muon chambers. Consequently they are the most easily identified objects, and the algorithm starts with them and removes afterwards the energy deposit of these identified muons from all subdetector components. After the muon identification the electron identification follows by associating tracks in the electromagnetic calorimeter with the electrons. Similarly to the previous step, all energy deposits are removed from all subdetector components.

Remaining tracks with a good quality are considered to be charged hadrons. From the momenta derived from the tracks and from the energy deposit in the calorimeter cells further particles like neutral hadrons or photons can be reconstructed.

6.2.1 Charged hadron subtraction

The Particle Flow algorithm gives the possibility to reduce pile-up effects since tracks associated to a vertex with lower energy (pile-up vertices) can be easily removed, as well as corresponding secondary vertices. In this way it is possible to reduce the number of particles produced by other interactions in the analysis and therefore improve for example the reconstructed hadronic top quark mass in a jet.

6.2.2 Electron reconstruction

In order to reconstruct electrons an MVA (for details on MVAs see Section 8) is used, which was developed and tested by the Egamma POG [134].
The electrons have to satisfy the following criteria:

- The object has to pass the conversion veto.
- The number of tracker layers before the first hit belonging to the track (missing hits) has to be zero.
- The lower cut on the electron Triggering-MVA discriminator (mvaTrigV0) has to fulfill
  - mvaTrigV0 > 0.94 if $|\eta_{\mathrm{SC}}| < 0.8$,
  - mvaTrigV0 > 0.85 if $0.8 < |\eta_{\mathrm{SC}}| < 1.479$,
  - mvaTrigV0 > 0.92 if $1.479 < |\eta_{\mathrm{SC}}| < 2.5$,

  where η_{SC} is the pseudorapidity of the electron supercluster.

Furthermore, an electron candidate has to satisfy $p_{\mathrm{T}} > 35\,\mathrm{GeV}$ and $|\eta_{Sc}| < 2.5$ [135].

6.2.3 Muon reconstruction

The muon reconstruction [135] is done with a tight working point [136] and the following criteria:

- The candidate has to be reconstructed both as a Global Muon and as a Particle Flow Muon.
- The normalized χ^2 of the global-muon track fit has to fulfill $\chi^2/N_{\mathrm{DOF}} < 10$.
- At least one muon chamber hit has to be included in the global-muon track fit.
- In at least two muon stations segments have to match the global-muon track.
- The longitudinal distance of the muon inner track with respect to the primary vertex has to be $|\Delta z| < 0.5\,\mathrm{cm}$.
- The transverse impact parameter of the muon inner track w.r.t. the primary vertex has to be $|d_{xy}| < 0.2\,\mathrm{cm}$.
- At least one pixel hit has to be found for the inner track.
- Minimum five layers have to have hits.

In addition the muon has to have a $p_\mathrm{T} > 45\,\mathrm{GeV}$ and $|\eta| < 2.1$. For the muons no isolation criteria are applied since in a leptonically decaying top quark decay the muon might be reconstructed within the b-jet at high transverse momentum of the top quark, and the following analysis explicitly aims at investigating boosted top quarks.

6.3 Jet reconstruction algorithms

In CMS jets are reconstructed using the Particle Flow objects. In the following different jet reconstruction algorithms and their properties are introduced.

6.3.1 Cone algorithms

Cone algorithms cluster particles within a cone. The majority of cone algorithm need a seed particle, for instance the particle with the highest transverse momentum. One example for a cone algorithm is the iterative cone (IC) algorithm. The algorithm starts with a seed particle and clusters all particles i and j together with a distance ΔR_{ij} smaller than R_0

$$\Delta R_{ij}^2 < R_0. \tag{6.1}$$

Here R_0 is a parameter, that represents the conesize of the jet. After all particles within this cone are clustered the resulting direction of the jet is taken as a new seed for the next iteration step. The procedure is repeated until the seed direction and the jet direction are equal, meaning that the jet content does not change in an iteration step. In this case a stable cone is found and all particles within this cone are removed and the procedure starts again to reconstruct another jet. Consequently this methods avoids overlapping jets.

Since this algorithm uses a seed it is not IRC safe. In more detail it is not collinear safe since the splitting of a particle may change the seed and therefore the hard prongs of the jet. In Section 11.2.1 and 11.2.2 the infrared and collinear safety of this algorithm is tested.

A solution to this problem is the SISCone algorithm (more details in Ref. [137]), which is a cone algorithm without any seed. As well as the IRC safety of jets also the computation time is an important factor in high energy physics. The SISCone algorithm is IRC safe, but much slower than the sequential clustering, and hence it is not used in CMS any longer.

6.3.2 Sequential clustering

Sequential clustering algorithms [138] iteratively pair particles together according to their distance or their momenta and try to undo the evolution of the parton shower.

Sequential algorithms cluster objects to jets with the following procedure: For each pair of initial four-momenta, two variables, invariant under longitudinal boost, d_{ij} and $d_{i\mathrm{B}}$ are calculated,

$$d_{ij} = \min[p_{\mathrm{T},i}^{2n}, p_{\mathrm{T},j}^{2n}]\Delta R_{ij}^2, \tag{6.2}$$

$$d_{i\mathrm{B}} = p_{\mathrm{T},i}^{2n} R_0^2. \tag{6.3}$$

Here R_0 and n are free parameters, R_0 is called the distance parameter of the jet and is responsible for the size of the jet. If d_{ij} is the smallest variable, object i and j are clustered together and the algorithm calculates again for all pairs the distances d_{ij} and $d_{i\mathrm{B}}$. If the smallest variable is a value $d_{i\mathrm{B}}$, the object i is called a jet and is removed from the list of objects. Afterwards the algorithm continues searching for the smallest variable in the remaining objects. The cluster algorithm is finished when no objects are left in the list. In this way one can obtain jets with a distance parameter of R_0.

In Equation (6.2) and (6.3) n represents an integer number which corresponds to the algorithms shown in Table 6.1.

Algorithm	n
$\mathrm{k_T}$	1
Cambridge/Aachen	0
Anti-$\mathrm{k_T}$	-1

Table 6.1: Different types of sequential clustering algorithms

The algorithms differ since they cluster the objects according to their p_T in different ways, resulting in different jets. In CMS the most widely used algorithm is the anti-$\mathrm{k_T}$ algorithm [139], which clusters first the hard objects. The $\mathrm{k_T}$ algorithm [138,140] clusters first the soft particles and results in irregularly shaped jets (see Figure 6.2), and therefore makes the jet area calculation more difficult. The active area of a jet is determined by adding to the existing particles uniformly a large number of extremely soft particles (ghost particles). One can then calculate the active area of the jets $A(J)$ by determining the fraction of soft particles that were clustered to the jets [141]

$$A(J) = \frac{N_g(J)}{\nu_g} = A_g \cdot N_g(J). \tag{6.4}$$

Here N_g is the number of ghost particles that were clustered to the jet J and ν_g is the density of the ghost particles, A_g is the area of a single ghost particle. For a single hard particle the active areas for different jet clustering algorithms can be seen in Table 6.2. The anti-$\mathrm{k_T}$ algorithm has the active area of a circle.

Active area	anti-$\mathrm{k_T}$	$\mathrm{k_T}$	CA	SISCone
$< A > /\pi r^2$	1	0.81	0.81	0.25

Table 6.2: Active area for different jet clustering algorithms [141].

The Cambridge/Aachen (CA) algorithm [142,143] clusters the objects according to their spatial distance. In many top tagging applications the CA algorithm is used as will be seen in Section 7, since it is suited for analyses using jet substructure. The CA algorithm is preferred to the anti-$\mathrm{k_T}$ algorithm since it has a hierarchical substructure as the $\mathrm{k_T}$ algorithm. However, the subjets of the $\mathrm{k_T}$ algorithm tend to include more soft radiation and lead eventually to a poor resolution. Furthermore, the CA algorithm is preferred for substructure techniques since the angle-based clustering distance ensures that at the relevant splitting the radii of the jets of the two prongs are similar to the distance between the two prongs themselves. Due to the angular ordering these jets contain most of the

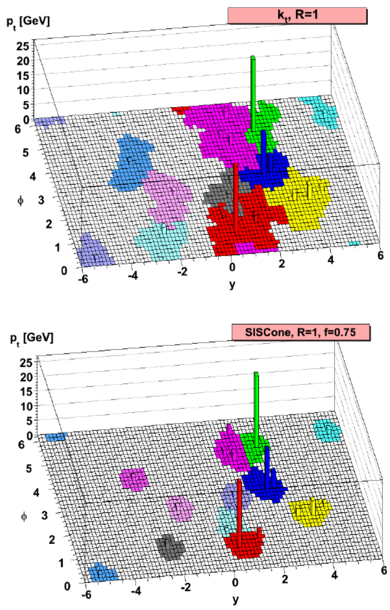


Figure 6.2: Areas and shapes of reconstructed jets using three different sequential clustering algorithms and the SiSCone algorithm for an example event [139].

radiation emitted by the decay products [45].

The sequential clustering is an IRC safe method as will be also shown in Section 11.2.1 and 11.2.2.

6.3.2.1 Variable R algorithm

The usual sequential cluster algorithms search for the smallest d_{ij} and $d_{i\mathrm{B}}$ as defined in Equation (6.2) and (6.3). In these equations R_0 was a fixed parameter which is called the distance parameter of a jet, and refers to the ΔR between the jet axis and the clustered object. However, a constant distance parameter as used in the standard sequential clustering algorithms is not the optimal choice since decay products will be collimated more strongly for high transverse momenta. In Figure 6.3 the maximum distance ΔR_{bjj} between the decay products of a top quark from a standard model $\mathrm{t\bar{t}}$ process as a function of the transverse momentum of the top quark is shown. One can observe that the maximum distance between the three decay products is decreasing with increasing transverse momentum. Using a constant conesize for the whole p_T range is not an optimal choice since the real jet distance parameter follows the following relation:

$$\Delta R \sim \frac{1}{p_\mathrm{T}}. \tag{6.5}$$

For a two prong decay with no spin effects [144] one can estimate

$$\Delta R \sim \frac{2m}{p_\mathrm{T}}. \tag{6.6}$$

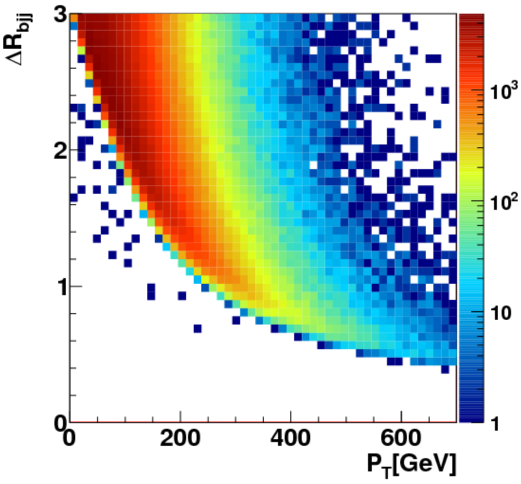


Figure 6.3: Maximum distance between the decay products of a top quark as a function of the transverse momentum of the top quark. One can see that the decay products are more collimated for higher values of p_{T} [145].

The Variable R algorithm [146] replaces the constant parameter in the sequential clustering by a p_{T} dependent R-parameter. A mass scale is introduced and the initial definition of $d_{i\mathrm{B}}$ is replaced by

$$d_{i\mathrm{B}} = p_{\mathrm{T}}^{2n} R_{\mathrm{eff}}^2(p_{\mathrm{T}i}), \quad (6.7)$$
$$\text{where } R_{\mathrm{eff}} = \frac{\rho}{p_{\mathrm{T}}}.$$

After modifying the clustering algorithms by introducing an effective distance parameter R_{eff}, which depends on the transverse momentum of the cluster we have a new mass scale ρ. In Equation (6.7) n is indicating the type of clustering algorithm as described in Section 6.3.

In each clustering step the distance d_{ij} is compared to $d_{i\mathrm{B}}$, which changes during the clustering process. For this reason the conesize of the jet changes in each clustering step and it is adjusted automatically with respect to the p_{T} of the jet.

A maximum R_{max} and a minimum radius R_{min} have to be set to avoid too large or small jets. In Figure 6.4 the jet clustering result for a single event is shown using the anti-k$_{\mathrm{T}}$ and the variable R anti-k$_{\mathrm{T}}$ algorithm. The distance parameter of the jet is constant for the usual clustering algorithm, whereas the distance parameter changes with respect to p_{T} for the variable R algorithm.

6.3.3 Jet energy corrections

Since the momentum of the reconstructed jet in the detector does not agree with the jets clustered from final state generator particles, the momentum of the jet has to be corrected. The reason for the differences between generator level jets and reconstructed

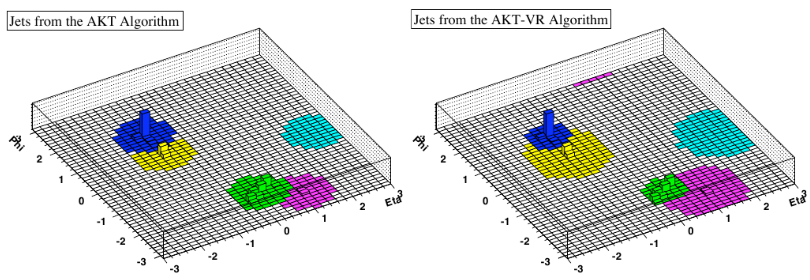


Figure 6.4: Clustering result for a single event with the anti-k_T algorithm (left) and the variable R anti-k_T algorithm (right) [146].

jets are mainly the non-uniform and non-linear response of the calorimeter, as well as from energy losses due to particles outside the jet area. Furthermore, additional particles from pile-up can be clustered to the reconstructed jet and undetected particles, like neutrinos, can cause a difference too. In order to correct for these effects in CMS a chain of factorized corrections is applied [147, 148]. This chain is visualized in Figure 6.5. In the following the steps applied are described briefly.

- The L1 correction corrects for pile-up and for instrumental noise separately for data and Monte Carlo simulation using the concept of jet areas [141, 149]. The transverse momentum is corrected depending of η, the active area A and the transverse momentum density ρ. Here ρ is an event-by-event median of the p_T^j/A_j, where the index j is the number of all jets in the event.

- The L2L3 corrections are calculated for the transverse momentum and the pseudorapidity from simulated QCD multijet events.

- In one last step the L2L3Res residual corrections on data are applied by comparing Z+jet, γ+jet and dijet events with simulation.

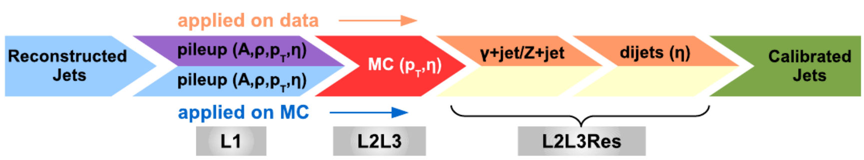


Figure 6.5: All steps of the jet energy corrections [150].

The remaining differences between data and simulation are corrected by applying scale factors (typically between 1 and 1.02 (η and p_T dependent)) to the measured four momenta.

Jet energy corrections are derived for anti-k_T jets. Furthermore, the corrections are calculated just for standard distance parameters of the jets. In order to correct also the jet energies of CA jets with a distance parameter of 0.8 (CA8) or CA jets with a distance parameter of 1.5 (CA15) the corrections of anti-k_T jets with a distance parameter of 0.7 (AK7) are applied to these jets. It was found that this is a good approximation and the description of jet energy distributions in data are reasonable [151]. However, clustering jets with non-standard distance parameters or other clustering algorithms can lead to a difference between data and simulation since the standard corrections are not accurate for these jets. In this analysis the START_V27 JEC are applied as described in more detail in Ref. [152].

6.3.4 Jet energy uncertainties

The main uncertainties originate from the differences in the flavor response simulated with PYTHIA6 Z2* and HERWIG++ EE-3C. The flavor uncertainties for gluon jets are around 2.5%, for light quarks in the order of 0.2% and for bottom jets around 0.5% [148]. The total JEC uncertainties in the barrel region are around 2% and are increasing with higher η up to 3%.

6.3.5 Jet resolutions

The jet resolution is measured in di-jet events where the asymmetry factor A is calculated

$$A = \frac{p_{\mathrm{T},1} - p_{\mathrm{T},2}}{p_{\mathrm{T},1} + p_{\mathrm{T},2}}. \tag{6.8}$$

Here $p_{\mathrm{T},1}$ and $p_{\mathrm{T},2}$ are the transverse momenta of the two jets in the event. Since the asymmetry is distributed around a mean with value zero one can write the standard deviation as

$$\sigma_A = \left|\frac{\partial A}{\partial} p_{\mathrm{T},1}\right| \cdot \sigma(p_{\mathrm{T},1}) + \left|\frac{\partial A}{\partial} p_{\mathrm{T},2}\right| \cdot \sigma(p_{\mathrm{T},2}). \tag{6.9}$$

If the two jets belong to the same η region one can write the jet p_{T} resolution $\sigma(p_{\mathrm{T}})$ as

$$\frac{\sigma(p_{\mathrm{T}})}{\langle p_{\mathrm{T}} \rangle} = \sqrt{2} \cdot \sigma_A. \tag{6.10}$$

This formula was already used at the Tevatron [153,154], in ATLAS [155] and in CMS [156]. The jet resolutions are calculated in data and in simulated events. The jet resolutions in data are generally worse than in simulation. In order to take this into account a matching with a standard ΔR criteria between the jet and the parton is applied ($\Delta R < 0.4$). If a jet is matched to a parton the following correction factor c can be calculated:

$$c = 1 + SF \cdot \frac{p_T^{\mathrm{jet}} - p_T^{\mathrm{parton}}}{p_T^{\mathrm{jet}}} \tag{6.11}$$

Here SF is an η-dependent scale factor. The correction factor c is then applied to simulated events in order to adjust the jet resolution to the data. In Figure 6.6 the jet resolution scale factors and their uncertainties are shown for the CMS detector. The scale factors are shown for the two data sets from 2010-2011 and a centre-of-mass energy of $\sqrt{s} = 7\,\mathrm{TeV}$ and the one from 2012 with a centre-of-mass energy of $\sqrt{s} = 8\,\mathrm{TeV}$. The scale factors and their uncertainties are increasing for both data sets with increasing η, due to the worse jet reconstruction in higher pseudorapidities regions.

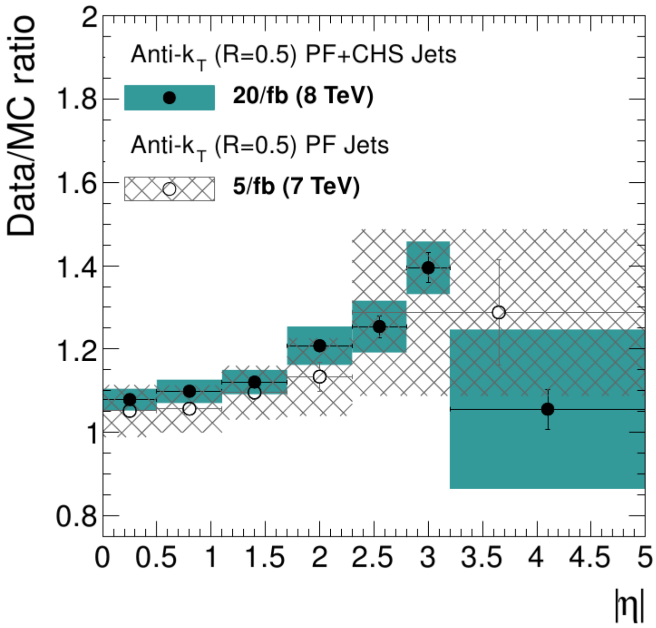


Figure 6.6: CMS jet resolution scale factors and their uncertainties, taken from [157].

6.4 Missing transverse energy

After all particles and jets are reconstructed and corrected, the missing transverse energy can be calculated from the negative sum of the four-vectors of all particles N, it is given by

$$\not{E}_{\mathrm{T}} = -\sum_{N} \vec{p}_i. \tag{6.12}$$

The missing energy is corrected in several steps as can be found in Ref. [158]. The typical resolution is around 20 GeV after the corrections were applied.

7 Top tagging algorithms

The identification of boosted hadronically decaying top quarks is challenging since for high transverse momenta the decay products of the top quarks are collimated within one large jet. Top tagging algorithms are using substructure techniques in order to distinguish these jets from background jets. In many cases the algorithms are trying to reconstruct the decay products of the jet by different subjet finding methods. The mass of the top quark is another important variable that can be accessed with different approaches.

Several algorithms with different approaches and different properties are already available. In the following the most widely used top tagging algorithms will be explained in detail and their advantages and disadvantages will be discussed.

7.1 CMS top tagger

The CMS top tagger [159,160] is a modified and optimized version of the algorithm developed by Kaplan et al. [161]. The algorithm uses jets clustered with the Cambridge/Aachen algorithm. In CMS the input of the tagger is a collection of jets with a distance parameter of 0.8 (CA8).

The algorithm finds subjets by a decomposition of the jet. It goes backwards in the cluster sequence and declusters in two separate steps, the primary and the secondary decomposition. The algorithm works in the following steps [160]:

1. The last clustering step to find the CA8 jet is undone and two subclusters are found.

2. If the two subclusters satisfy $\Delta R > 0.4 - A \times p_{\mathrm{T}}^{C}$ (adjacency criterion), where here p_{T}^{C} is the transverse momentum of the original jet in the primary decomposition or another subcluster in the secondary decomposition, the algorithm continues with step 3. The parameter $A = 0.0004\,\mathrm{GeV}^{-1}$ is a slope parameter which is optimized by using simulated events. If the two subclusters are too close together the decomposition fails. In this step it can be seen that the CMS top tagger has a p_{T} dependent parameter which takes the shrinking of the large jet with increasing p_{T} into account.

3. The two subclusters that pass the criteria in step 2 have to pass in addition a p_{T} fraction criterion: $p_{\mathrm{T}}^{\mathrm{Cluster}} > \delta_p \times p_{\mathrm{T}}^{\mathrm{hardjet}}$, where $p_{\mathrm{T}}^{\mathrm{Cluster}}$ is the transverse momentum of the subcluster and $p_{\mathrm{T}}^{\mathrm{hardjet}}$ the transverse momentum of the parent jet. The default parameter is $\delta_p = 0.05$. This parameter has been optimized by using simulated events.

4. In the case that just one of the two subclusters satisfies the criteria in step 3, the procedure is repeated on the passed subcluster from step 2 on. The constituents of the failed subcluster are removed from the algorithm.

5. The decomposition fails if there is no subcluster left that satisfies the p_{T} fraction criterion of step 3 after the iteration, or if the subcluster is a single constituent.

In summary, the algorithm declusters the jet to find subjets which are well separated (step 2) and which carry a significant fraction of the original jet's momentum (step 3). If the decomposition fails for all subclusters, the jet has just one subjet, the jet itself. After the primary decomposition the jet can have two subclusters. If the two subclusters do not satisfy the criteria for a further decomposition, the final jet will have two subjets. If just one of the two subclusters can be further decomposed, the jet will have three subjets, and if both can be further decomposed the jet will have four subjets. An overview of the clustering procedure is shown in a schematic drawing in Figure 7.1. After the

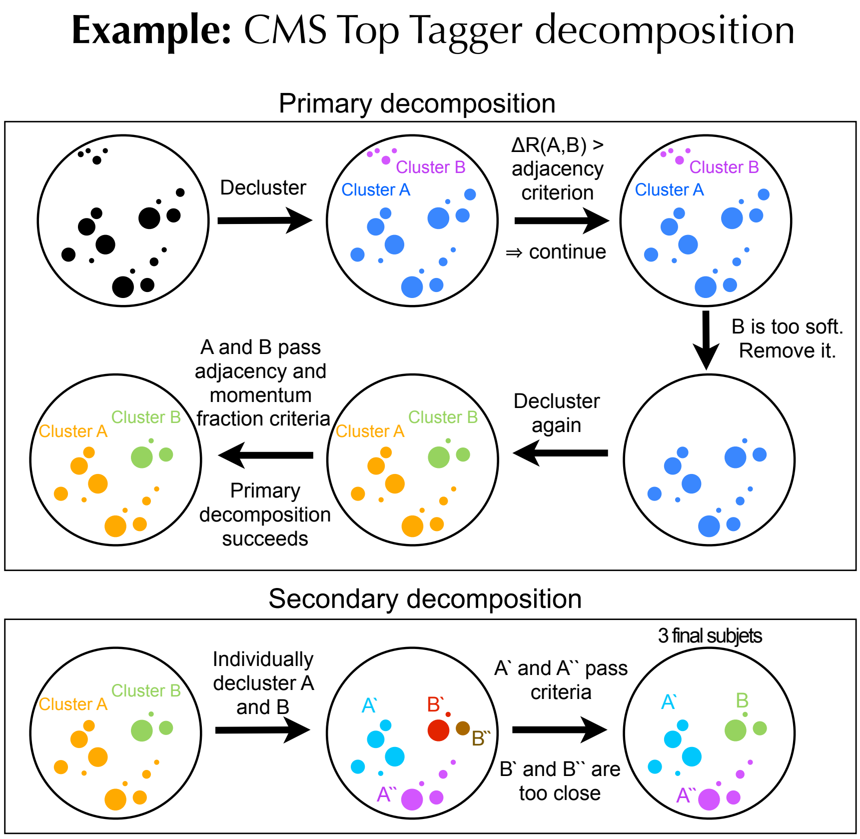


Figure 7.1: Primary and secondary decomposition phase of the CMS top tagger to identify subjets [162].

decomposition there are the following selection criteria applied:

1. Jet mass m_{jet}: A selection on the mass of the jet (the default selection requires the jet mass to fulfill $140 < m_{\mathrm{jet}} < 250$ GeV),

2. Number of subjets N_{subjets} (found by the decomposition): The standard selection is $N_{\mathrm{subjets}} \geq 3$, which aims at reconstructing the three decay products of the top quark;

3. Minimum pairwise mass m_{min}: The variable is calculated by building the pairwise invariant mass of the three highest-p_{T} subjets. The minimum pairwise mass is the minimum of the three mass combinations: $m_{\mathrm{min}} = \min(m_{12}, m_{13}, m_{23})$. Typically,

m_{min} is required to be larger than 50 GeV, which selects jets with the decay products of a hadronically decaying W boson inside.

The CMS top tagger is usually used in a high p_{T} region with $p_{\mathrm{T}} > 350$ GeV, since the decay products are not collimated enough at lower values of p_{T} to be fully contained in CA8 jets.

The CMS top tagger was used already in various analyses in Run I with a centre-of-mass energy of 7 and 8 TeV, for example in Ref. [64, 67, 163, 164].

7.2 HEP top tagger

The HEP top tagger [160, 165] is used in a lower p_{T} region and uses therefore larger jets than the CMS top tagger to ensure that all decay products are reconstructed inside the jet cone. The distance parameter of the Cambridge/Aachen jet is therefore 1.5 (CA15). The HEP top tagging algorithm finds subjets as described in the following [160]: Two subclusters are found by undoing the last clustering step that is used to find the CA15 jet. Two subclusters j_1 and j_2 are thus obtained with $m_{j1} > m_{j2}$.

1. The mass drop [166] criterion $m_{j1} < 0.8 \times m_j$ has to be satisfied. Here m_j is the mass of the original jet. If the criterion is not fulfilled, the subcluster j_2 is discarded and the mass drop criteria is proved for the subcluster j_1. If the subclusters have a mass that is less than 30 GeV, the subclusters are saved as subjets. In this step any number of subjets can be found. If there are less than three subjets found the declustered jet is not considered as a top quark candidate.

2. After the mass drop criterion has been checked, a filtering algorithm as described in Section 7.5.2 is used for each combination of three subjets. The algorithm is reclustering the constituents of the three subjets with the parameter $R_{\mathrm{filt}} = \min(0.3, \Delta R_{ij}/2)$. Here ΔR_{ij} is the smallest distance between the three subjets. By reclustering the three subjets any number of new subjets can be found.

3. Only the five subjets with the highest p_{T} are kept.

4. This is done for each combination of three subjets, consequently for each combination five filtered subjets are obtained.

5. For each set of the five filtered subjets the invariant mass is calculated. The set of subjets that has the mass closest to the top quark mass is kept. All other combinations are discarded.

6. Finally the constituents of the selected five filtered subjets are reclustered with the Cambridge/Aachen clustering algorithm. The final jet has then always exactly three subjets.

A schematic drawing for the clustering procedure for the HEP top tagger is given in Figure 7.2. Additional selection criteria are applied for the HEP top tagger:

- m_{123}: The invariant mass of the three final subjets. The cut aims on selecting jets which have a mass close to the top quark mass. The standard cut applied is $140\,\mathrm{GeV} < m_{123} < 250\,\mathrm{GeV}$.

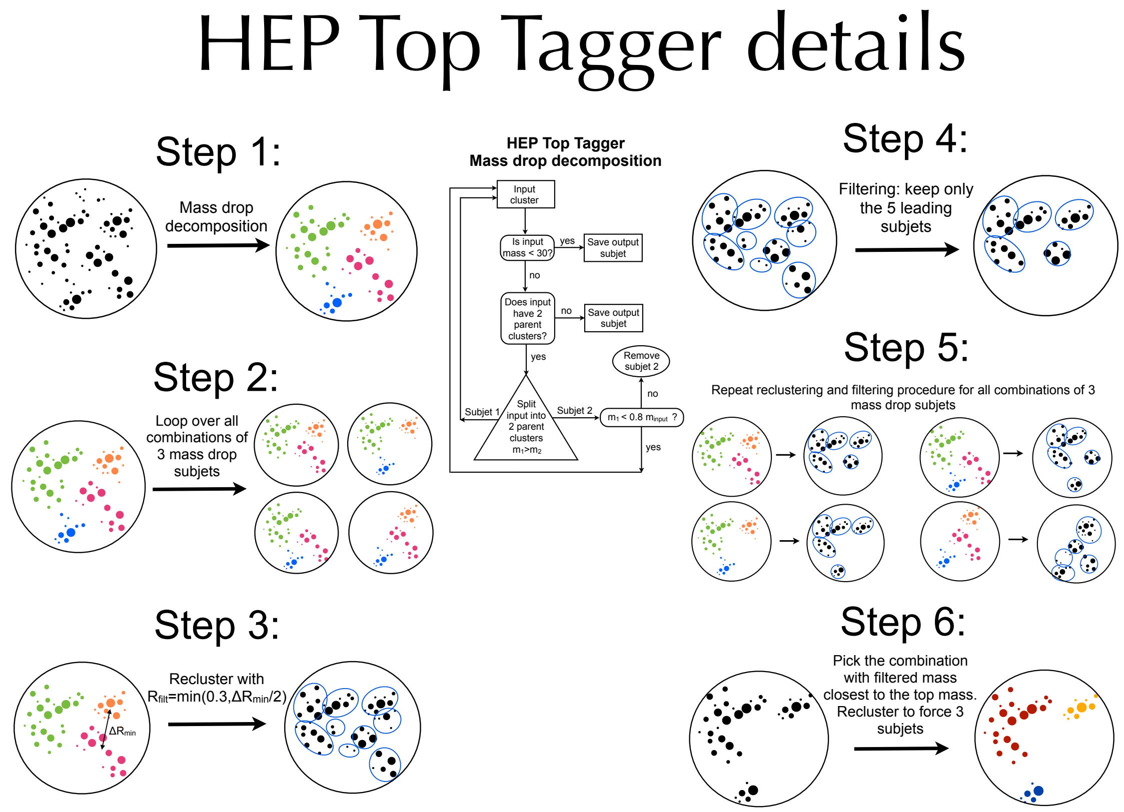


Figure 7.2: Schematic drawing of the clustering process and subjet finding of the HEP top tagger [162].

- m_{12}, m_{13}, m_{23}: Cuts on the invariant mass of combinations of two of the three subjets, aiming at reconstructing the W boson mass. Here the subjets are ordered in p_T, meaning that m_{12} is the invariant mass of the two leading subjets.

The cuts on the masses are more complex and are shown in Figure 7.3. An A-shaped region is selected, where the W boson mass selection has to satisfy one of the three criteria:

$$m_{13,\mathrm{min}} < \arctan(\frac{m_{13}}{m_{12}}) < m_{13,\mathrm{max}} \textbf{ and } R_\mathrm{min} < \frac{m_{23}}{m_{123}} < R_\mathrm{max}, \tag{7.1}$$

$$R^2_\mathrm{min}(1+(\frac{m_{13}}{m_{12}})^2) < 1-(\frac{m_{23}}{m_{123}})^2) < R^2_\mathrm{max}(1+(\frac{m_{13}}{m_{12}})^2) \textbf{ and } \frac{m_{23}}{m_{123}} > m_{23\mathrm{cut}}, \tag{7.2}$$

$$R^2_\mathrm{min}(1+(\frac{m_{12}}{m_{13}})^2) < 1-(\frac{m_{23}}{m_{123}})^2) < R^2_\mathrm{max}(1+(\frac{m_{12}}{m_{13}})^2) \textbf{ and } \frac{m_{23}}{m_{123}} > m_{23\mathrm{cut}}. \tag{7.3}$$

R_min and R_max are given by:

$$R_\mathrm{min} = (1-f_\mathrm{W}) \times \frac{m_\mathrm{W}}{m_\mathrm{t}}, \tag{7.4}$$

$$R_\mathrm{min} = (1+f_\mathrm{W}) \times \frac{m_\mathrm{W}}{m_\mathrm{t}}. \tag{7.5}$$

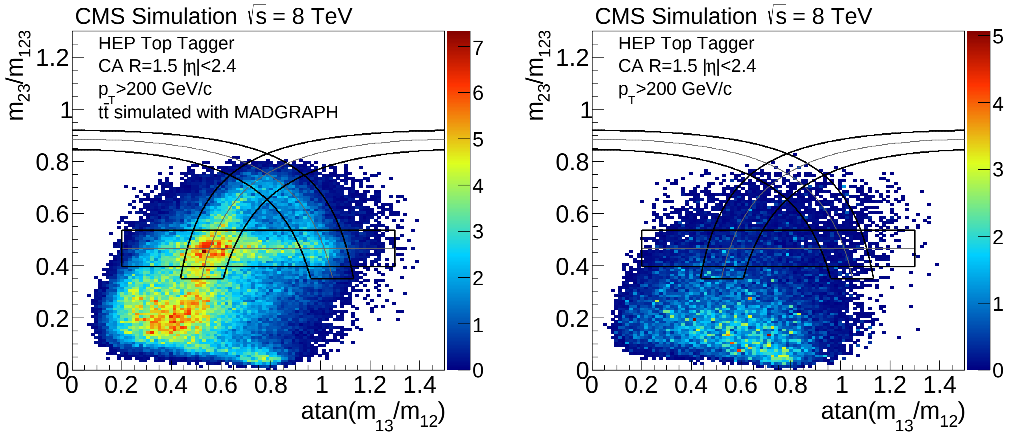


Figure 7.3: Mass ratio cuts for the HEP top tagger: The top quark mass and W boson mass are selected by cutting on m_{12}, m_{13}, m_{23} and m_{123} respectively for signal events (left) and background events (right) [162].

The parameter f_{W} is the width of the mass window, which is usually set to 0.15. The free parameters $m_{13,\mathrm{min}}$, $m_{13,\mathrm{max}}$ and $m_{23\mathrm{cut}}$ are usually set to:

$$m_{13,\mathrm{min}} = 0.2, \tag{7.6}$$

$$m_{13,\mathrm{max}} = 1.3, \tag{7.7}$$

$$m_{23\mathrm{cut}} = 0.35. \tag{7.8}$$

As will be shown in Sections 9.2 and 10.4.2 the HEP top tagger has a good performance in the low p_{T} range but its efficiency for higher transverse momenta is low in comparison to other top tagging algorithms. One of the reasons for this behavior is the distance parameter, that is used by the algorithm. Due to the large distance parameter the fraction of particles not belonging to the decay of the top quark increases, since the distance parameter stays constant, whereas the decay products will be stronger collimated with increasing p_{T}.

Together with the CMS top tagger the HEP top tagger is the most commonly used tagging algorithm in CMS and was used in some analyses that were kinematically close to the resolved scenario [64, 164].

7.3 OptimalR HEP top tagger

The OptimalR HEP top tagger [167] is a modified version of the HEP top tagger described in the previous section, using multiple conesizes. The OptimalR HEP top tagger starts with CA15 jets as the original HEP top tagger and reduces the conesize of the jet as long as the mass drop condition $m_{\mathrm{subjet}} > (1 - A) \cdot m_{\mathrm{jet}}$ is fulfilled, where A is the OptimalR threshold which is set here to 0.2. R_{opt} is the smallest radius where the mass drop criterion is still satisfied. To find the jets with smaller radii the last clustering step of the parent jet is reversed. Afterwards the HEP top tagger is run on each of the smaller jets. The mass of these new smaller jets is defined as the candidate mass after applying the HEP top tagger. To increase the number of valid jets the mass criteria of the HEP top tagger are relaxed (see Table 7.1).

In order to improve the background rejection, another variable is introduced: $R_{\rm opt}^{\rm calc}$. This is the optimal radius that is calculated from signal jets. This variable has a $p_{\rm T}$ dependence, since we expect the jets to become smaller with increasing $p_{\rm T}$. The fitted functions for $R_{\rm opt}^{\rm calc}$ are given by:
$p_{\rm T} < 500\,{\rm GeV}$:

$$R_{\rm opt}^{\rm calc} = 20.23 - 3.417 \cdot 10^{-2} \cdot p_{\rm T}[{\rm GeV}] + 2.269 \cdot 10^{-5} \cdot p_{\rm T}^2[{\rm GeV}^2] \quad (7.9)$$

$p_{\rm T} > 500\,{\rm GeV}$:

$$R_{\rm opt}^{\rm calc} = 15.95 - 1.870 \cdot 10^{-2} \cdot p_{\rm T}[{\rm GeV}] + 8.894 \cdot 10^{-6} \cdot p_{\rm T}^2[{\rm GeV}^2] \quad (7.10)$$

The OptimalR HEP top tagger in contrast to the HEP top tagger applies the mass ratio cuts before the set of subjets with the closest mass to the top quark mass is selected. In this way the artificially introduced mass resonance for background events that is caused by selecting the set of subjets with the closest mass to the top quark mass can be removed. The OptimalR HEP top tagger is based on three decision variables. Like the HEP top tagger, the OptimalR HEP top tagger selects events by applying cuts on $f_{\rm W}(R = R_{\rm opt})$ and $m(R = R_{\rm opt})$, but here the selection is performed only on the jet with the smallest radius. The algorithm also cuts then on $R_{\rm opt} - R_{\rm opt}^{\rm calc}$ for further background rejection.

In Table 7.1 the selection criteria and parameters that are used in this analysis are summarized.

Parameter	Value
min $R_{\rm opt}$	0.5
max $R_{\rm opt}$	1.5
OptimalR threshold A	0.2
top quark mass	172.3 GeV
W boson mass	80.4 GeV
top quark mass window m_{123}	[0, 2000] GeV
W boson mass ratio f_W	[0.396634, 0.536622] ([0.85%, 1.15%])
mass plane cuts	$(m_{23\rm cut}, m_{13,\rm min}, m_{13,\rm max}) = (0.35, 0.2, 1.3)$
mass drop threshold	0.8
max subjet mass	30 GeV
$R_{\rm filt}$	0.3
$n_{\rm filt}$	5
minimal subjet $p_{\rm T}$	20 GeV
minimal reconstructed $p_{\rm T}$	200 GeV

Table 7.1: Parameters used for the OptimalR HEP top tagger.

In Table 7.1 $R_{\rm filt}$ and $n_{\rm filt}$ are the filtering parameters which are explained in Section 7.5.2. In comparison to the HEP top tagger the OptimalR HEP top tagger also applies a selection on the transverse momentum of the subjets. By requiring the subjet $p_{\rm T}$ to be greater than 20 GeV the signal to background ratio is improved. As seen in Table 7.1 the OptimalR HEP top tagger considers only large jets with $p_{\rm T} > 200\,{\rm GeV}$.

7.4 N-subjettiness

N-subjettiness [168,169] is a jet-shape variable that estimates how compatible a jet is with the assumption of having N subjets. The N-subjettiness variable is defined as

$$\tau_N = \frac{1}{d_0} \sum_{i=1}^{N} p_{\mathrm{T},i} \min(\Delta R_{1,i}, \Delta R_{2,i}, ..., \Delta R_{N,i}). \tag{7.11}$$

The distances $\Delta R_{i,j}$ are the distances between all jet constituents j and a hypothetical subjet axis i. The subjet axis i is found by determining a one-pass optimization procedure which minimizes τ_N (for details see Ref. [168]). The $p_{\mathrm{T,i}}$ are the momenta of the jet constituents. The normalization d_0 is given by:

$$d_0 = \sum_i p_{\mathrm{T},i} R_0. \tag{7.12}$$

The parameter R_0 is the distance parameter of the large jet. Through the normalization, τ_N is bound to be in the interval $[0, 1]$. As can be seen from Equation (7.11) the smaller τ_N becomes the better the constituents are aligned with the subjet candidate axis i, which means that the probability that a jet consists of N subjets gets higher.

In top tagging a good decision variable is the ratio of τ_3/τ_2 since a signal like jet is expected to have three subjets where a background jet is expected to have less than three subjets. In other words, it is expected that for signal-like jets τ_3 is small and τ_2 is close to 1 so that the ratio τ_3/τ_2 will have smaller values than for background jets.

N-subjettiness is often combined with top tagging algorithms described in the previous section, where it usually gives additional separation power. Another advantage is that N-subjettiness is an infrared and collinear safe quantity.

In Run I N-subjettiness was used in many analyses together with the CMS top tagger or the HEP top tagger. Furthermore, it was also used to identify W bosons by using a selection on τ_2/τ_1 [170].

7.5 Grooming techniques

The general purpose of grooming techniques is to clean the large jets from pile-up contributions and additional soft radiation. Therefore, grooming methods are not used as actual top tagging algorithms, but are usually used in addition to reduce the pile-up dependence. Grooming algorithms can also improve the jet mass resolution and therefore lead to an improved top tagging performance.

In the following a few of these techniques are introduced.

7.5.1 Pruning

The pruning algorithm [171,172] reclusters the constituents of the large jet with the Cambridge/Aachen algorithm or with the $\mathrm{k_T}$-algorithm. During the cluster procedure before each combination step of the cluster j_1 and j_2 (with $p_{\mathrm{T}}^{j1} > p_{\mathrm{T}}^{j2}$), the following conditions are checked:

$$\frac{p_{\mathrm{T}}^{j2}}{p_{\mathrm{T}}^{j1+j2}} > z_{\mathrm{cut}}, \tag{7.13}$$

$$\Delta R_{12} < R_{\mathrm{cut}}(2m_{\mathrm{jet}}/p_{\mathrm{T,jet}}). \tag{7.14}$$

When at least one of the two conditions is fulfilled the algorithm combines j_1 and j_2 and checks the condition again in the next clustering step. If neither condition is satisfied the algorithm discards j_2 and continues with the next clustering step. Here z_{cut} and R_{cut} are free parameters of the algorithm.

The algorithm aims at removing soft particles and particles that are radiated in a wide angle.

7.5.2 Filtering

The filtering algorithm [173] reclusters all constituents of the large jet with the Cambridge/Aachen algorithm with a smaller radius R_{filt}. The new large jet is then the sum of the n_{filt} hardest jets. Typically R_{filt} is 0.3 and the three hardest jets are kept.

7.5.3 Trimming

The trimming procedure [174] clusters the constituents of the large jet again with the $\mathrm{k_T}$ algorithm to jets with smaller radii. Each of the new jets has to satisfy the following criterion:

$$\frac{p_{\mathrm{T,sub}}}{p_{\mathrm{T,jet}}} > f_{\mathrm{cut}}. \tag{7.15}$$

Here f_{cut} is a free parameter of the algorithm. If any of the new jets do not fulfill the criterion they are rejected. With this procedure the algorithm removes especially soft particles which come most likely from pile-up or final state radiation. QCD jets typically lose more of their mass than jets initiated by a top quark decay, which makes trimming also usable for top tagging.

7.5.4 Soft drop

Soft drop is a substructure technique that removes wide-angle soft radiation from jets, as described in [175]. It declusters a given CA jet by going backwards in the cluster sequence. After the declustering one can obtain two subclusters with p_{T1} and p_{T2} and a distance between these two subclusters of ΔR_{12}. If the two subclusters satisfy the following condition the final jet has been found, otherwise the softer subcluster will be removed and the declustering continues on the hard subcluster.

$$\frac{\min(\mathrm{p_{T1}}, \mathrm{p_{T2}})}{p_{\mathrm{T1}} + p_{\mathrm{T2}}} > z_{\mathrm{cut}} \left(\frac{\Delta R_{12}}{R_0}\right)^{\beta} \tag{7.16}$$

R_0 is the distance parameter of the original jet. It can be seen that for large angles between the two subclusters (large ΔR_{12}) the equation will not be satisfied.

The soft drop algorithm has two parameters: z_{cut} and β. For $\beta \to \infty$ the original jet will not be modified.

In contrast to other grooming techniques, soft drop is for $\beta > 0$ an infrared and collinear safe algorithm.

7.6 Other substructure variables

7.6.1 Jet shapes

The jet shape $\Psi(R)$ is a substructure variable that describes the fraction of the transverse momenta of all jet constituents with a distance smaller than R to the jet axis to the transverse momentum of the jet [176, 177]. In this analysis the jet shapes are calculated with subjets instead of jet constituents, here the subjets have to have a distance smaller than R

$$\Psi(R) = \frac{\sum p_{\mathrm{T,sub}<R}}{p_{\mathrm{T,jet}}}. \tag{7.17}$$

In this analysis the subjets, that are found with the CMS top tagger are used to calculate the jet shapes.

7.6.2 Helicity angle

The helicity angle θ_{ij} [178] is the angle between the primary quark i and the W boson before the decay in its rest frame, assuming the W boson decays into quark i and j. Since the quarks are reconstructed with subjets this variable is calculated on reconstruction level with the subjets.

7.6.3 Jet charge

The jet charge Q is defined by the sum of all jet constituents' charge,

$$Q = \frac{1}{N} \sum_{i}^{N} q_i. \tag{7.18}$$

7.6.4 Weighted jet charge

The weighted jet charge [179] is the sum of the charges of all particles multiplied with their respective energy E_i to the power of k:

$$Q_k = \frac{1}{E_{\mathrm{jet}}^k} \sum_{i}^{N} q_i E_i^k. \tag{7.19}$$

7.6.5 Jet moment

The first (n=1) and the second (n=2) jet moments γ_n are the sum of the distance between the jet axis and the constituents i,

$$\gamma_n = \frac{1}{N}\sum_i^N \Delta R^n_{\text{jet,i}}, \tag{7.20}$$

where N is the number of the jet constituents.

7.6.6 Q-jet volatility

Quantum-jet (Q-jet) volatility is a non-deterministic approach to cluster jets [180]. The constituents of the jet have to be reclustered for Q-jet volatility. As explained in Section 6.3.2, in the usual jet clustering the smallest distance among d_{ij} and $d_{i\text{B}}$ is calculated. For Q-jet volatility the variable ΔR_{ij} gets replaced by a weight ω_{ij}. With each weight a probability Ω_{ij} that the clusters i and j should be combined can be calculated

$$\Omega_{ij} = \frac{\omega_{ij}}{\sum_{ij}\omega_{ij}}, \tag{7.21}$$

with the weight

$$\omega^{\alpha}_{ij} = \exp\left(-\alpha\frac{d_{ij} - d_{ij,\text{min}}}{d_{ij,\text{min}}}\right). \tag{7.22}$$

Here α is a real parameter and $d_{ij,\text{min}}$ is the minimal d_{ij} in the clustering procedure.

Since now every combination step gets a probability by a randomized algorithm the jets differ with each clustering from each other. After clustering a jet several times with this clustering procedure all substructure variables have for a given jet a distribution and not a single value anymore.

One expects that the distribution of the substructure variables (for example the jet mass) is more sharply peaked in signal events than in background events, since for signal events one expects values around the top quark mass whereas for background events one does not expect any preferred value. Therefore, one can introduce the width of the mass distribution as

$$\Gamma_{\text{volatility}} = \frac{\sqrt{\langle m^2\rangle - \langle m\rangle^2}}{\langle m\rangle}. \tag{7.23}$$

In this analysis the width of the mass distribution is calculated, in general the Q-jet volatility can be calculated for any substructure variable.

7.7 Shower deconstruction

Shower deconstruction [181] is a new method to distinguish hadronically decaying top quarks from light quark jets, similar to the matrix element method [182]. To do this, the algorithm calculates a decision variable χ, which is a probability quotient. The nominator of this quotient is the probability that a given final state $\{p\}_N$ is created by a signal-like jet, and the denominator is the probability that the same final state $\{p\}_N$ is created by a background-like jet.

To understand how these probabilities can be calculated one can look into a specific example as shown in Figure 7.4. In the top drawing of Figure 7.4 a signal event is shown, in which some initial state radiation appears and a top quark (indicated with the star) decays into a W boson and a b quark. Before the decay takes place there is a gluon emitted. In the final state there are nine quarks and gluons, which in the experiment will be detected as microjets. If one looks at the bottom drawing of Figure 7.4 the same amount of microjets can be caused by a background event. In the QCD event also some initial state radiation takes place as well as some gluon splitting. In order to calculate the probability to find a certain configuration of microjets, the algorithm assigns a splitting probability and a decay probability to each vertex. These probabilities are the same as used in shower event generators. Furthermore, Sudakov factors [111] are used to calculate the probability that a splitting happened between one and another vertex.

For a given final state as shown in Figure 7.4 the probability density that this shower history is produced by a signal event is calculated. But since there are many possibilities to produce this final state, one has to sum over all possible shower histories. This is done for all showers that may be caused by a signal event $P(\{p\}_N|S)$ as well as for all possibilities that the final state $\{p\}_N$ is caused by background $P(\{p\}_N|B)$. The quotient of these two probabilities gives the decision variable χ of the shower deconstruction tagger,

$$\chi(\{p\}_N) = \frac{P(\{p\}_N|S)}{P(\{p\}_N|B)}. \tag{7.24}$$

Since in the detector one can not observe partons like shown in Figure 7.4, but jets, the final state $\{p\}_N$ is here the set of the momenta of the N microjets.

The microjets are found by reclustering the constituents of a large jet (typically CA8, CA15) with the k_T algorithm with a distance parameter around 0.2. Since the computational time of shower deconstruction increases almost exponentially with the number of microjets and microjets with low p_T carry little information, the maximum number of microjets is restricted to nine in this analysis. In addition, microjets with a p_T lower than 10 GeV are discarded.

In order to reduce the computational time, it is possible to introduce mass windows on the jet mass for the top quark and the W boson decay. In Table 7.2 all parameters are shown that are used for shower deconstruction in this analysis.

Shower deconstruction calculates the effective decay widths Γ_t and Γ_W with the parameters given in Table 7.2 as following [181]:

$$\Gamma_t = \frac{(m_t + |\sigma_t|)^2 - m_t^2}{N_{\mathrm{BW}} \cdot m_t}, \tag{7.25}$$

$$\Gamma_W = \frac{(m_W + |\sigma_w|)^2 - m_W^2}{N_{\mathrm{BW}} \cdot m_W}. \tag{7.26}$$

If the calculated top quark mass and W boson mass are not within the effective decay width, the shower histories are not calculated.

The shower deconstruction algorithm was already tested by the ATLAS experiment in Ref. [183] and in CMS within this analysis.

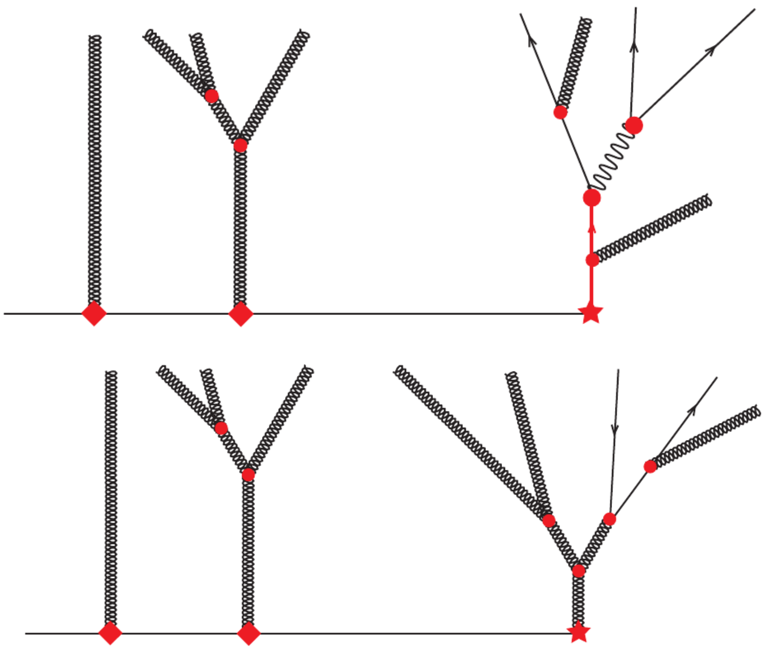


Figure 7.4: Shower histories for a signal (top) and a background (bottom) event with a similar final state. Stars are indicating the hard interaction, diamonds initial state radiation, large circles decays and small circles QCD splittings [181].

Parameter	Value
microjet R (k_T jets)	p_T dependent
max. number of leading microjets	9
microjet p_T	$> 10\,GeV$
Large jet R	0.8 / 1.5
Large jet p_T	$> 200\,GeV$
W boson mass m_W	80.4 GeV
W boson window σ_W	$\pm 12\,GeV$
Top quark mass m_t	160 GeV
Top quark mass window σ_t	$\pm 25\,GeV$
NBreitWigner N_{BW}	2

Table 7.2: Parameters used for the shower deconstruction tagger.

7.8 Mass jump algorithm

Subjet finding and the reconstruction of the decay products of the top quark are essential for a good top tagging algorithm as already seen in the former sections. The CMS top tagger, the HEP top tagger and the OptimalR HEP top tagger are declustering the jet in order to reconstruct the decay products. The shower deconstruction tagger clusters the particles of the jet again to smaller jets. The mass jump algorithm [184] is designed to find the subjets by reclustering the jet constituents of a large jet and is able to resolve the decay products also in a high p_T regime. The idea behind the algorithm is that mass jumps appear just between subjets. Like the VariableR jet clustering algorithm (Section 6.3.2.1) the mass jump algorithm is a modification of the jet reconstruction algorithms introduced in Section 6.3.

The algorithm distinguishes between two classes of jets: passive and active jets. In the beginning all particles are labeled as active jets. As in the usual clustering process the jet algorithm starts finding the smallest d_{ij} or d_{iB}. The algorithm has the following steps:

1. Find smallest distance among d_{ij} or d_{iB}. If it is given by d_{iB} mark jet i as passive and remove it from the list of active jets.
2. If d_{ij} is the smallest distance and the invariant mass of cluster i and j m_{ij} is smaller than a parameter μ cluster them together. As we will see later the parameter μ has to be chosen high enough to make the algorithm infrared safe.
3. If the invariant mass of the two clusters is already higher than μ check the mass jump criterion $\theta \cdot m_{ij} > \max[m_i, m_j]$:
   a) If the mass jump criterion is fulfilled label jet i and jet j as passive and go back to step 1.
   b) If the mass jump criterion is not fulfilled the criterion is checked also between active and passive jets:
      i. Find the closest passive jet k to jet i. Furthermore, this jet k should not be isolated, so it is not turned into a passive jet by step 1, but in step 3a.
      ii. Check if jet i and k would have been clustered if k would still be active $d_{ik} < d_{ij}$.
      iii. Check if there is a mass jump between the active and the passive jet $\theta \cdot m_{ik} > \max[m_i, m_k]$.
   c) If there appeared a mass jump between the active jet i and the passive jet k label i passive as well. Repeat the procedure for jet j.
4. If no mass jump is found cluster jet i and j.

The algorithm is inspired by the existing mass drop tagger and prevents the recombination of two hard prongs if their combined mass is significantly higher. With this method the algorithm resolves even highly boosted decays [184]. The mass jump algorithm has several similarities to the mass drop algorithm [166] that is used by the HEP top tagger to obtain subjets. The main differences between the two algorithms can be seen in Figure 7.5. The mass drop algorithm declusters the large jet and is grooming away soft radiation (1). If the mass drop criteria is fulfilled two subjets are found (2). This procedure continues (3+4) and removes more soft radiation (5). This iterative procedure continues until the subjet masses are below the threshold μ [184].

In contrast the mass jump algorithm clusters until the jet mass is higher than μ with the standard sequential clustering algorithms (a). The algorithm continues clustering (b) until a mass jump is reached (c). The two jets become passive and therefore subjets. The remaining active jets keep clustering (d) until another mass jump appears. This mass jump can appear also between a passive jet and an active jet (e). The algorithm is finished when all active clusters are turned passive (f) [184].

One of the main differences is that the mass jump algorithm collects more radiation below the mass jump threshold μ. In Figure 7.6 the background rejection for the HEP top tagger with the mass drop and the mass jump algorithm is shown as a function of the top tagging efficiency. Here the mass drop and the mass jump algorithms are used to obtain the subjets. The performance is shown on generator level and on reconstruction level. One can observe that for a given efficiency the background rejection of the mass jump algorithm is higher.

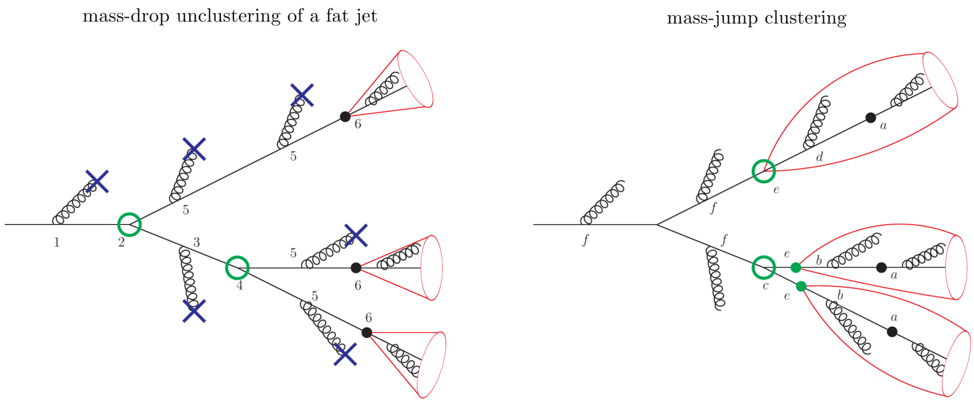


Figure 7.5: Subjet finding with the mass drop algorithm (left) and the mass jump algorithm (right) [184].

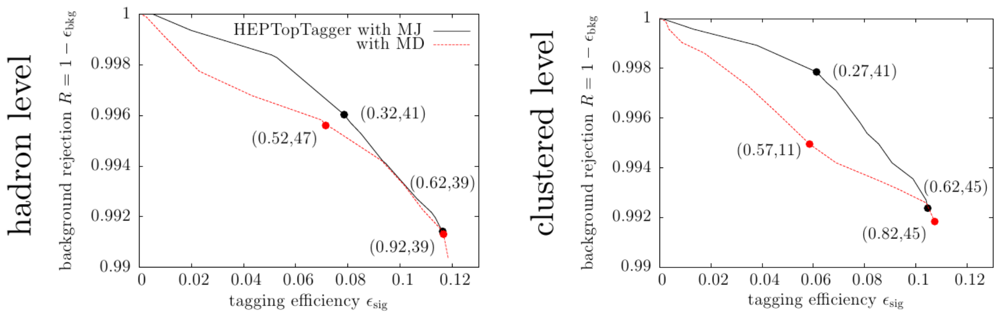


Figure 7.6: Top tagging performance for the HEP top tagger with mass drop subjet finding and mass jump subjet finding on generator level (left) and reconstruction level (right) [184]. Top quarks originate from a hypothetical Z' with a mass of 500 GeV decaying into top quark pairs. QCD multijet events are considered as background events.

7.9 b quark identification

The identification of b quarks [185, 186], also called b-tagging, can be used in addition to further improve the identification of a top quark, which decays into a W boson and a b quark. In QCD multijet production the fraction of b jets produced is small, such that the identification of b quarks can help to separate signal from background.

The B hadrons originating from the fragmentation of b quarks have a relatively long lifetime of about $\tau = \mathcal{O}(1\text{ps})$. During this time the B hadron can travel on average 1 mm from the primary vertex. The tracks caused by the decay of the B hadron are extrapolated back to reconstruct a secondary vertex. In Figure 7.7 a schematic drawing of the reconstruction is shown. The impact parameter indicated with d_0 in Figure 7.7 is one of the most discriminating variables for b tagging. Also the relatively high mass of the B hadrons can be used to improve the performance of b tagging algorithms.

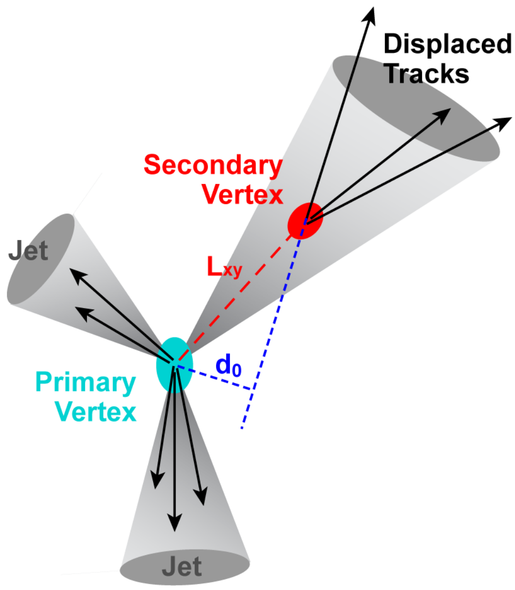


Figure 7.7: Illustration of the secondary vertex and the impact parameter d_0. Due to the long lifetime of the b quark it decays significantly further away from the primary vertex [187].

In the following analysis the Combined Secondary Vertex (CSV) [185] algorithm is used. The CSV algorithm combines the information of several discriminating variables like the impact parameter, the significance of tracks, the reconstructed secondary vertex and jet kinematics and projects them on just one discriminating variable with a multivariate analysis (see more details on multivariate analysis in Section 8). There are three working points available; the CSV loose working point corresponds to a b quark misidentification rate of 10%, the medium working point to a b quark misidentification rate of 1% and the tight working point to a mistag rate of 0.1%, respectively.

7.9.1 Subjet b-tagging

The application of b-tagging algorithms to highly collimated top quark jets has deficiencies because the jet axis has been calculated using the decay products from the fragmentation of three quarks, instead of the b quark alone. However, applying b tagging on the subjets found by the top tagging algorithms introduced above can help to re-gain a significant amount of efficiency [186]. This approach is known as subjet b-tagging and is used in the following to improve the performance of top tagging algorithms.

8 Top tagging with a multivariate approach

In the previous chapters all available top tagging algorithms in CMS are introduced and explained. All of the introduced algorithms use simple rectangular cuts. In this chapter it is analyzed how top tagging can be improved by using multivariate analysis methods and cuts on several substructure variables with the goal to use the performance as a benchmark for developing new top tagging algorithms. Multivariate methods were used already to determine benchmarks for tagging algorithms for several kind of particles, as for instance W and Z bosons in Ref. [188].

Rectangular cuts (also called 1-dimensional cuts) are acting on one variable at a time. That means the separation power of the second variable might be already reduced after the first variable is used. Multivariate analysis methods offer another approach to separate signal from background events. Furthermore, they are able to take into account the correlations between variables. Summarizing, the selection procedure with rectangular cuts provides good results if the decision variables have low correlations. If the correlations are large, 1-dimensional cuts are not effective and multivariate analysis methods perform better.

Based on earlier studies [189,190] the usage of boosted decision trees (BDTs) with substructure variables to identify hadronically decaying top quarks is further investigated. In the following section BDTs will be explained. The training of the BDT and the corresponding results will be explained in Section 8.2. In Section 9 the performance of the BDT in simulated events will be shown and compared with existing top tagging algorithms. The validation in 8 TeV data will be shown in Section 10.

8.1 Boosted decision trees

A BDT method consists of two main steps. First the BDT gets trained. This is usually done with simulated signal and background events. In this analysis $t\bar{t}$ events are used as signal and QCD multijet events as background. The simulated events get divided into two sets, one for the training and one for the testing. Furthermore, one has to provide decision variables as an input.

- Training: Each variable gets scanned and the so-called gini index with $p \cdot (1-p)$ is calculated, where p is the purity of the events. The minimal gini index give the optimal cut on this variable. The grid points for each variable are set in this analysis to 20, which means that each variable is separated in 20 bins. Afterwards the variable with the best separation power (symbolized with a node in the tree) is chosen. Now the tree gets split into two new branches, one with events that have a value higher than the optimized cut and others with lower values. The procedure is repeated for each of the branches and the branches can be split again, as seen in Figure 8.1. The iteration stops when a maximum number of branches is reached, or if there are too few events in one node left. Nodes that are not divided further are called leaves.

After the training each event is placed in a leaf and each leaf is either signal-like or background-like.

- Boosting: In order to reduce fluctuations for each training several trees are trained (typically 1000 decision trees). After the first tree is trained the events in the leaves are analyzed, and events that are misclassified (for example a signal event is classified as a background event) get high weights, events that are correctly assigned get low weights. The next tree is then trained with the new event weights and afterwards the weights are recalculated. In this analysis gradient boosting is used, for more details see Ref. [191].

- Testing: With the second half of the events the trained method is tested and evaluated.

Events that end up in signal-like leaves are assigned to a MVA value of 1, whereas events that end up in background-like leaves get a value of -1. In a boosted decision tree these values are summed up (including their weights) and are then normalized to the number of trees that are trained. With this method each event gets a MVA value between 1 (signal-like) and -1 (background-like). This value is the projection of the input decision variable and therefore the only decision variable left. In this analysis 1000 trees are trained with a maximum of 5 nodes. The number of nodes is restricted in order to avoid over-training. The analysis is done using the ROOT toolkit for multivariate data analysis TMVA [191].

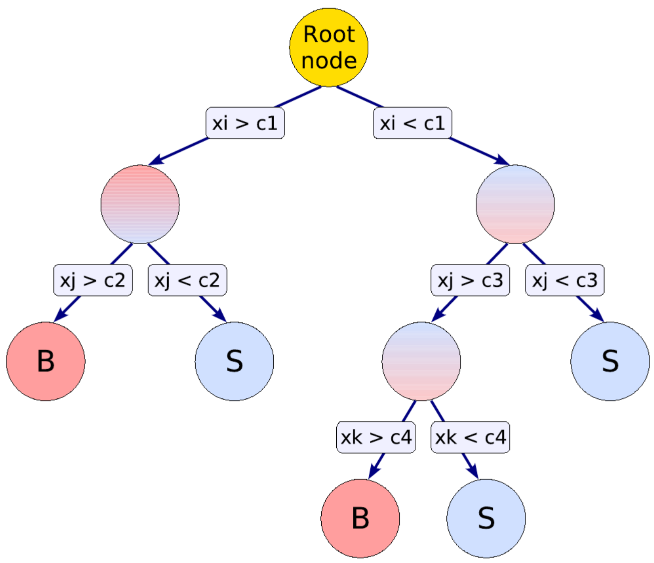


Figure 8.1: Schematic drawing of a decision tree, taken from [191].

8.2 Input decision variables

The BDT is trained in a first step with all calculated substructure variables shown in Table 8.1. The substructure variables are all calculated with CA8 jets and the subjets that are found by the CMS top tagger.

jet mass m_{jet}	pruned jet mass $m_{\text{jet pruned}}$
subjet masses $m_{\text{sub},i}$	pruned subjet masses $m_{\text{sub,pruned},i}$
subjet $p_{\text{T,i}}$	number of subjets N_{sub}
N-subjettiness τ_i	pruned N-subjettiness $\tau_{\text{p,i}}$
Q-jet volatility $\Gamma_{\text{volatility}}$	Helicity angle $\theta_{i,j}$
Jet charge Q	weighted jet charge Q_k
Number of jet constituents	Number of charged jet constituents
Jet shapes $\Psi(R)$	minimum pairwise mass m_{min}
1st jet moment	2nd jet moment

Table 8.1: All available substructure variables.

In a second training only the ten best variables (according to the TMVA ranking [191]) are chosen and used for further analysis in order to simplify the BDT. In TMVA the ranking of variables depends on how often a variable is used in order to split a node. Furthermore, each split gets weighted by the separation gain squared it has achieved and by the number of events in the node.

After removing the other variables from the training it is observed that these variables did not give any significant improvement anymore to the BDT top tagger. The variables that are used in this analysis are shown in Table 8.2. The TMVA ranking of the ten selected variables as a function of the jet p_{T} is shown in Figure 8.2. One can see that with increasing p_{T} the mass variables become more important. The reason for this behavior is that the decay products of the top quark are better reconstructed within a CA8 jet for higher transverse momenta. Furthermore, the W boson mass is better reconstructed by the invariant mass of the combined subjets. The other substructure variables change slightly their rank. Over the full p_{T} range the pruned N-subjettiness $\tau_{3,\text{p}}$ is ranked best by TMVA.

jet mass m_{jet}	number of subjets N_{sub}
invariant mass of subjet 1 & 2 mass $m_{1,2}$	invariant mass of subjet 2 & 3 mass $m_{2,3}$
p_{T} of the leading subjet $p_{\text{T},1}$	Q-jet volatility $\Gamma_{\text{volatility}}$
pruned N-subjettiness $\tau_{3,\text{p}}/\tau_{2,\text{p}}$	pruned N-subjettiness $\tau_{2,\text{p}}/\tau_{1,\text{p}}$
pruned N-subjettiness $\tau_{3,\text{p}}$	Helicity angle $\theta_{1,3}$

Table 8.2: Ten best substructure variables that are used for the final BDT training.

For each of the ten best substructure variables a Receiver Operating Characteristic (ROC) curve is determined by finding for each efficiency point the optimal selection on this variable that gives the minimum mistag rate. The ROC curves can be seen in Figure 8.3. As a single variable the transverse momentum of the leading subjet $p_{\text{T},1}$ and the Q-jet volatility $\Gamma_{\text{volatility}}$ are the best performing variables, whereas the pruned N-subjettiness ratio $\tau_{3,\text{p}}/\tau_{2,\text{p}}$ has the worst performance. One can observe that the best performing variable is not equal to the TMVA ranking. If variables have low correlations to other variables the TMVA ranking is lower since this variable will not be used to split many nodes.

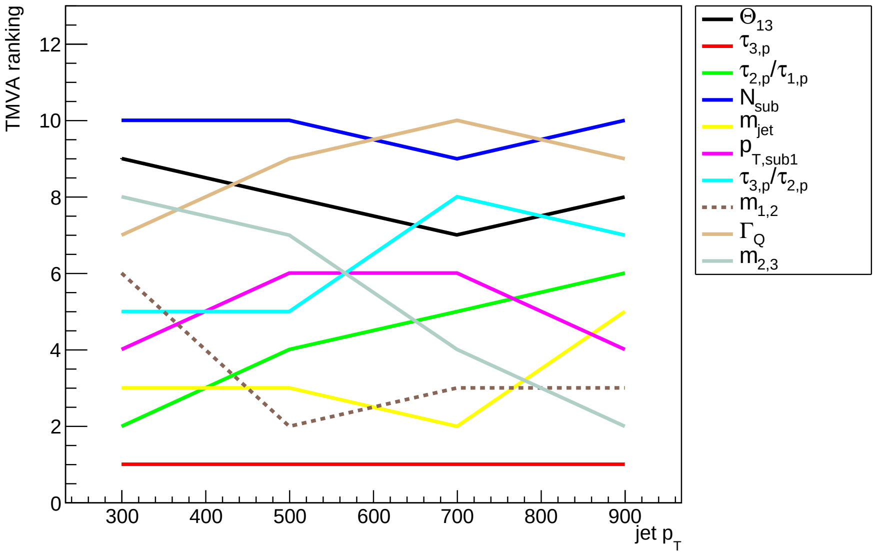


Figure 8.2: TMVA ranking of the ten substructure variables as a function of the jet p_T, where the best variable gets rank 1. The graph is divided into four bins of p_T (200-400 GeV; 400-600 GeV; 600-800 GeV and 800-1000 GeV).

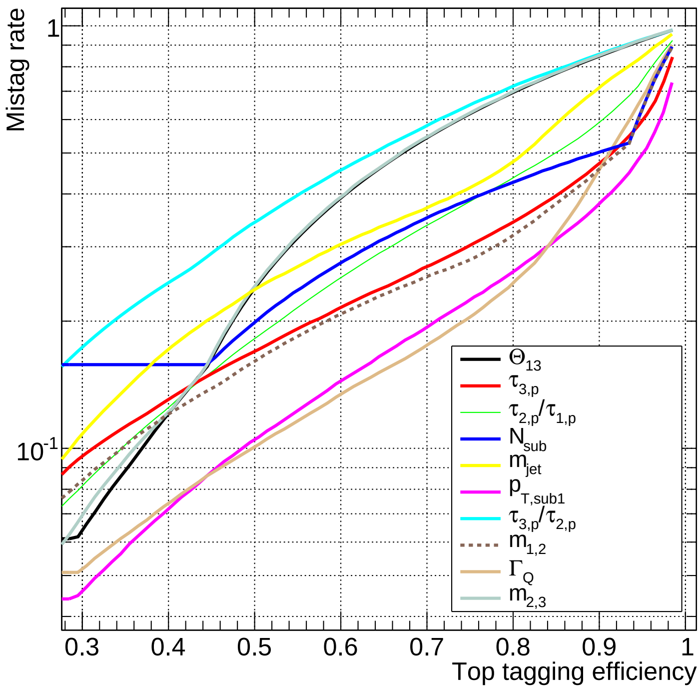


Figure 8.3: ROC curves for the ten best substructure variables.

The p_T distributions of the signal and background samples are reweighted to a flat p_T spectrum between 0 GeV and 2000 GeV in order to give all p_T bins the same importance and to prevent that the BDT is more sensitive to low transverse momenta.

For background the BDT is trained with two event generators. Similar results have been obtained for both event generators HERWIG++ and PYTHIA.

8.3 Linear correlations and training results

In Figures 8.4 and 8.5 the linear correlations between the ten substructure variables are shown. For signal events (Figure 8.4) the correlations between the variables are generally low. Highly correlated are the mass variables m_{jet}, $m_{1,2}$ and $m_{2,3}$ as well as the number of subjets N_{sub}. Since these variables are all created with the CMS top tagging algorithm by the declustering procedure a correlation is expected.

The N-subjettiness variables have a higher correlation to each other as well. Almost no correlations to any other variable can be observed for the Q-jet volatility and the leading subjet p_T.

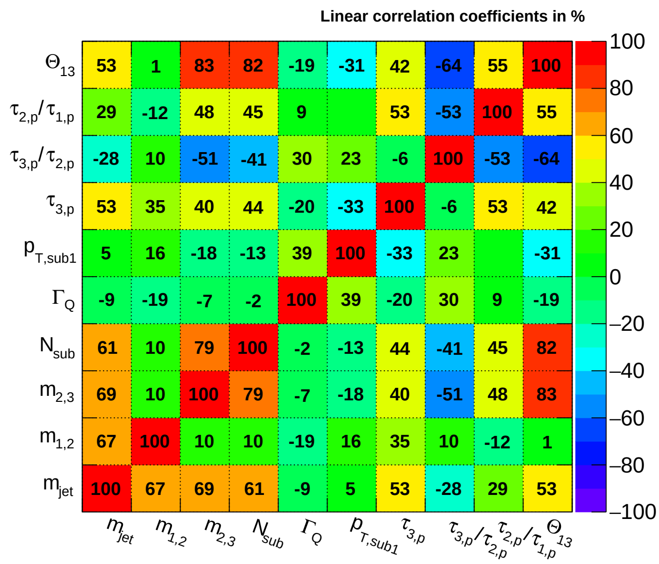


Figure 8.4: Linear correlations of the ten substructure variables, chosen for the BDT for signal events.

In Figure 8.6 the background rejection gain is shown for ten BDT trainings. In each training another variable is added (from left to right). The background rejection reaches a saturation after the usage of seven variables.

The result of the training with ten substructure variables can be seen in Figure 8.7 for low transverse momenta 400 GeV $< p_T <$ 600 GeV and high transverse momenta 800 GeV $< p_T <$ 1000 GeV. Signal and background events are clearly separated and signal events peak at 1, whereas background events peak at -1 as desired. The separation power of the BDT increases with increasing p_T since the shoulder of the signal events

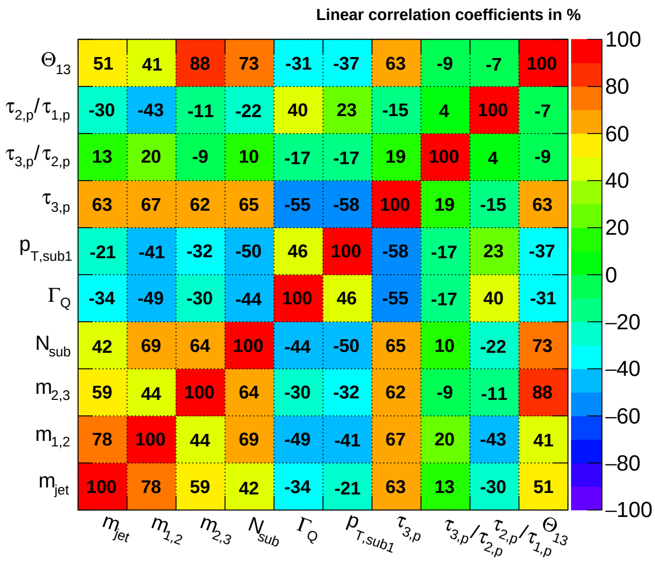


Figure 8.5: Linear correlations of the ten substructure variables, chosen for the BDT for background events.

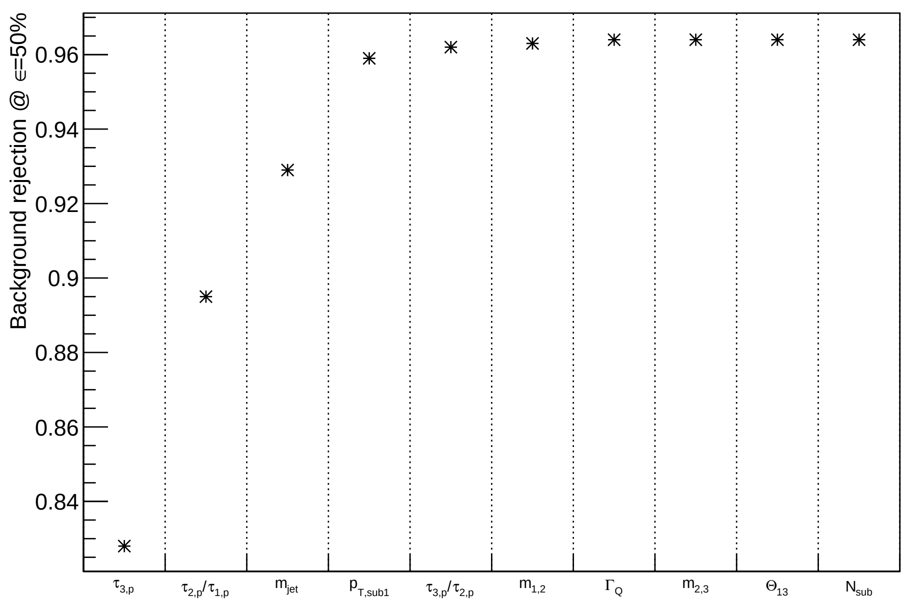


Figure 8.6: Background rejection for different BDTs. From left to right variables are added to the BDT and the background rejection is measured for a top tagging efficiency of 50%.

(Figure 8.7 top) vanishes with increasing p_{T} (Figure 8.7 bottom) and the signal events become more signal-like.

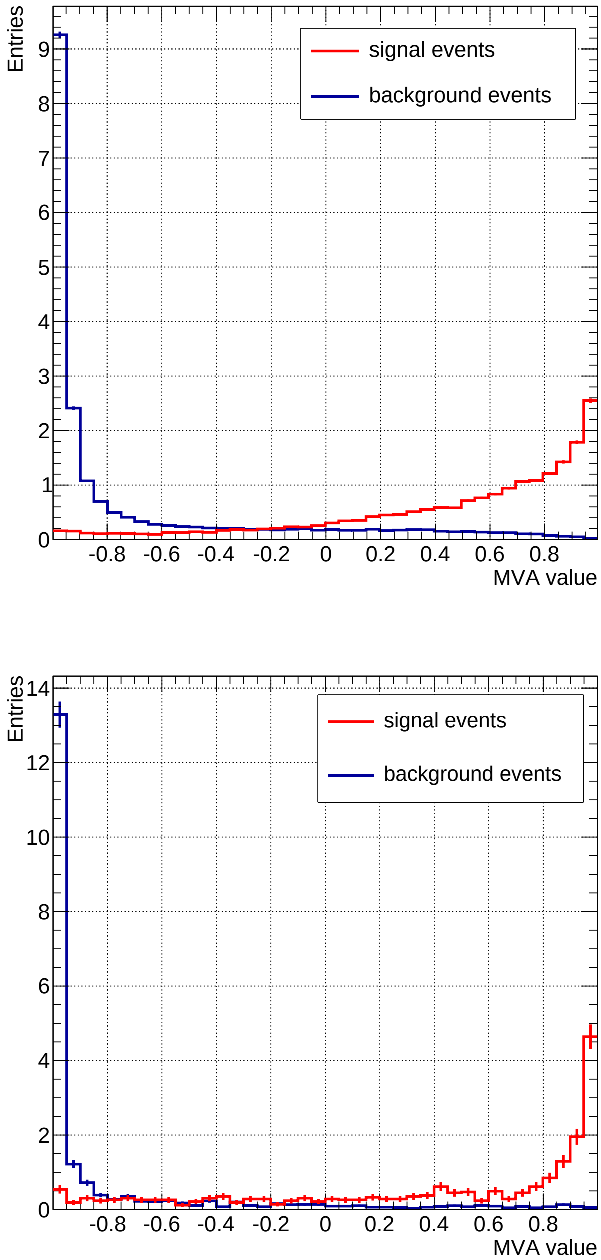


Figure 8.7: Decision variable of the BDT after the training with ten substructure variables for low transverse momenta (400 GeV $< p_{\mathrm{T}} <$ 600 GeV) (bottom) and high transverse momenta 800 GeV $< p_{\mathrm{T}} <$ 1000 GeV.

9 Comparison of top tagging algorithms in simulated events

In this chapter a comparison is made between the various top tagging algorithm introduced in Section 7. For these comparisons simulated events of the CMS detector with $\sqrt{s} = 8\,\text{TeV}$ are used. In the first part of this chapter the decision variables of the tagging algorithms are shown for signal and background events. These decision variables are used in Section 9.2 to determine the efficiency and mistag rate of the algorithms. In Section 9.2 the performance of the tagging algorithms are shown for different p_T ranges of the matched parton.

9.1 Decision variables

Decision variables are specific variables which provide a good separation power between signal events and background events. In order to operate top tagging algorithms one usually chooses rectangular cuts on these decision variables. In the following section the characteristic variables for each tagging algorithm are shown and explained. Each variable is shown for signal as well as for background events. For signal events a standard model $\mathrm{t\bar{t}}$ process generated with POWHEG interfaced to PYTHIA6 is used, with just hadronically decaying top quarks. The background events are simulated QCD multijet events generated with PYTHIA6.

9.1.1 CMS top tagger

As described in Section 7.1 the CMS top tagger performs a selection based on the decision variables N_subjets, m_min and m_jet. In Figure 9.1 the distributions of m_min and m_jet are shown for signal events and background events and a jet $p_\text{T} > 500\,\text{GeV}$. The minimum pairwise mass is shown here after a selection on the jet mass ($140 < m_\text{jet} < 250\,\text{GeV}$) and for more than two subjets. For both distributions only jets are considered which are matched to a top quark in case of signal events and to the highest p_T parton in case of the QCD multijet background. The peak of the jet mass distribution m_jet in $\mathrm{t\bar{t}}$ events is shifted to a higher value with respect to the actual value of the top quark mass, which is due to additional radiation collected by the jet algorithm. Also, the distribution exhibits a shoulder for low mass values. This is caused by unmerged jets, i.e. jets that contain just two of the three decay products. By requiring three or more subjets, the shoulder gets reduced. With increasing p_T this shoulder vanishes, as the jets are more collimated and all decay products are merged inside the CA8 jet. Comparing the distribution of the jet mass in $\mathrm{t\bar{t}}$ and QCD multijet production, good separation power is observed.

The minimum pairwise mass peaks for signal events at the position of the W boson mass, and has its peak at smaller values for QCD multijet production. Also for this variable good separation power is observed. The standard selection criteria (see Table 9.1) using these variables are close to optimal for high signal efficiency. In Table 9.1 the different

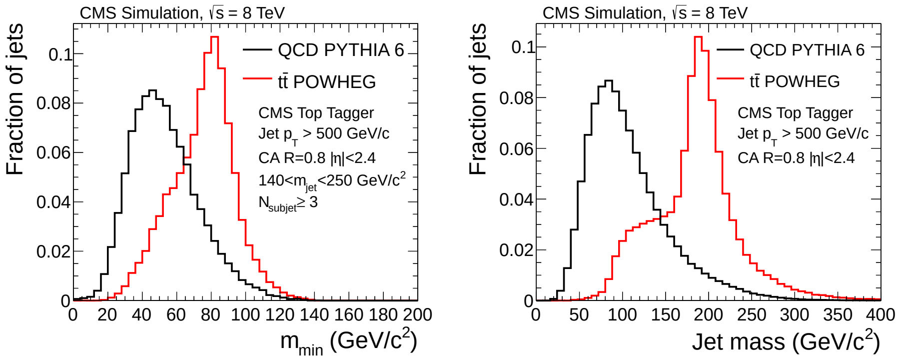


Figure 9.1: Minimum pairwise mass m_{min} for background and signal events (left) and jet mass m_{jet} (right) for the CMS top tagger, taken from [160].

working points for the CMS top tagger are listed, including working points where the CMS top tagger is combined with a subjet b-tag or N-subjettiness.

Working point	m_{jet} [GeV]	m_{min} [GeV]	subjet b-tag WP	τ_3/τ_2
CMS WP0	140-250	> 50	none	none
CMS combined WP1	140-250	> 50	CSV-loose	< 0.7
CMS combined WP2	140-250	> 50	CSV-loose	< 0.6
CMS combined WP3	140-250	> 50	CSV-medium	< 0.55
CMS combined WP4	140-250	> 65	CSV-medium	< 0.4

Table 9.1: Standard working points of the CMS top tagger [160].

9.1.2 HEP top tagger

In Figure 9.2 the invariant mass is shown of the three subjets of the HEP top tagger after the application of the filtering procedure. For signal events a clear top mass resonance is visible. A shoulder appears for low mass values due to unmerged jets. For background events, a resonance structure appears around the top quark mass. This is caused by combining the three subjets with the closest invariant mass to the top quark mass. This procedure leads to an artificial resonance structure for background events, and therefore decreases the discrimination power of this variable. The standard working points of the HEP top tagger and the HEP top tagger combined with a subjet b-tag and N-subjettiness are shown in Table 9.2.

Working point	m_{123} [GeV]	f_W	subjet b-tag WP	τ_3/τ_2
HTT WP0	140-250	> 50	0.495	none
HTT combined WP1	140-250	0.495	CSV-loose	none
HTT combined WP2	140-250	0.15	CSV-medium	none
HTT combined WP3	140-250	0.15	CSV-medium	< 0.63

Table 9.2: Standard working points of the HEP top tagger (HTT) [160].

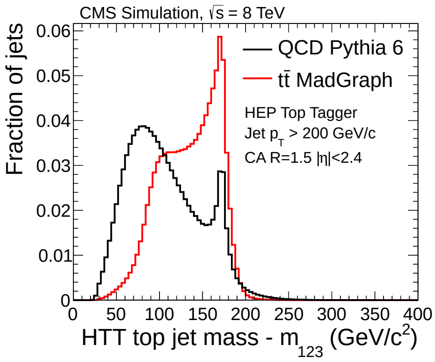


Figure 9.2: Invariant mass of the three subjets found by the HEP top tagger for signal and background events. The variable is shown for jets with $p_T > 200\,\mathrm{GeV}$, taken from [160].

9.1.3 OptimalR HEP top tagger

The OptimalR HEP top tagger is a modified version of the HEP top tagger using multiple conesizes (see Section 7.3). It uses the same decision variables as the HEP top tagger, but at the optimized radius. The mass at the optimal radius R_{opt} is shown in Figure 9.3 (right). Comparing the mass distributions with the one from the HEP top tagger (Figure 9.2) one can observe that the artificial resonance for background events disappeared. Another decision variable $R_{\mathrm{opt}} - R_{\mathrm{opt}}^{\mathrm{calc}}$ is used to provide additional separation power. In Figure 9.3 (left) this additional variable is shown. As expected the signal events are distributed around zero, since the difference between the observed R_{opt} and the expected radius $R_{\mathrm{opt}}^{\mathrm{calc}}$ is build and $R_{\mathrm{opt}}^{\mathrm{calc}}$ is calculated using signal events. Additional separation power is observed for this variable, as the distribution peaks at a value of about -0.4 for background events and at 0 for signal events.

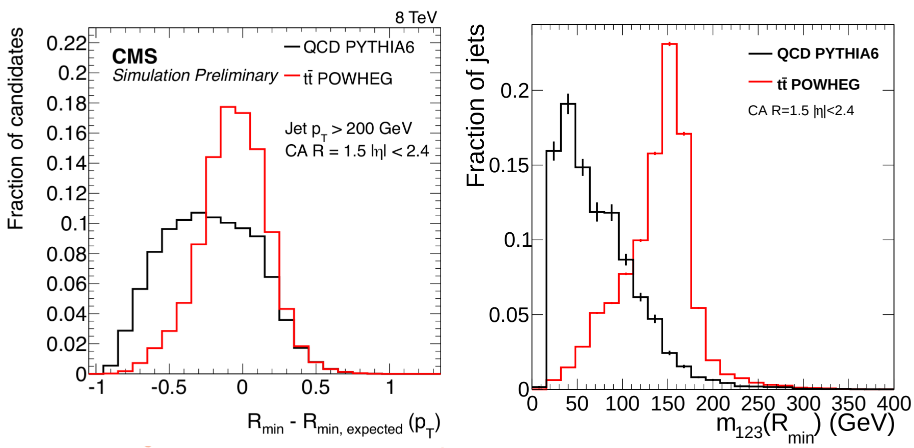


Figure 9.3: OptimalR HEP top tagger: $R_{\mathrm{op}} - R_{\mathrm{opt}}^{\mathrm{calc}}$ for signal and background events (left). The graph shows events for jets with $p_T > 200\,\mathrm{GeV}$, taken from [192]. Jet mass m_{123} at the optimal radius R_{opt} (right) for signal and background events.

9.1.4 Shower deconstruction

The shower deconstruction tagger selects jets based on a single decision variable χ. Preselection cuts are applied on the top quark mass, as well as on the W boson mass in order to reduce the computational time. For the same reason the number of microjets is restricted to a maximum of nine. The logarithm of the decision variable χ is shown in Figure 9.4. As expected, the distribution of signal events is distributed around higher values than for background events since the variable χ is a probability ratio between the probability that a configuration of microjets is caused by a signal event and the probability that a configuration of microjets is caused by a background event.

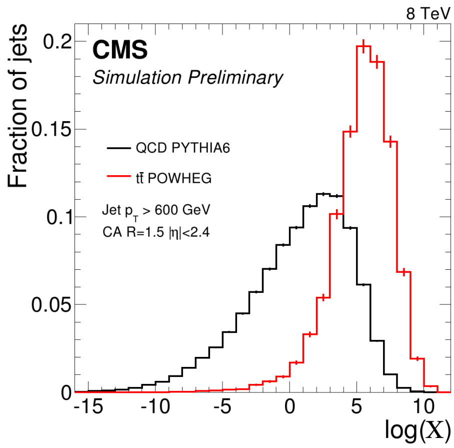


Figure 9.4: Decision variable χ for the shower deconstruction tagger for signal and background events and jets with $p_{\mathrm{T}} > 600\,\mathrm{GeV}$, published in Ref. [192].

9.1.5 N-subjettiness

For top tagging the most interesting decision variable based on N-subjettiness is the fraction τ_3/τ_2 (see Section 7.4). In Figure 9.5 this decision variable is shown for jets with $p_{\mathrm{T}} > 200\,\mathrm{GeV}$ and $p_{\mathrm{T}} > 500\,\mathrm{GeV}$. While for QCD multijet production there is no preferred subjet axis, the distribution of τ_3/τ_2 is accumulated close to one. Signal events are more likely to have three subjet axes than two, and therefore the fraction of τ_3/τ_2 is distributed at lower values.

N-subjettiness provides good separation power for high transverse momenta and can be used in addition to the usual top tagging algorithms to suppress background. This is facilitated by the fact that N-subjettiness has small correlations to most of the standard top tagging variables.

9.2 Comparison of top tagging algorithms

In order to characterize and compare different top tagging algorithms it is necessary to measure their efficiency and misidentification rate (mistag rate). Here the efficiency is a quantity that describes how many true top quarks are identified as top quarks by the respective algorithm. The mistag rate is a measure of how often a jet is wrongly identified

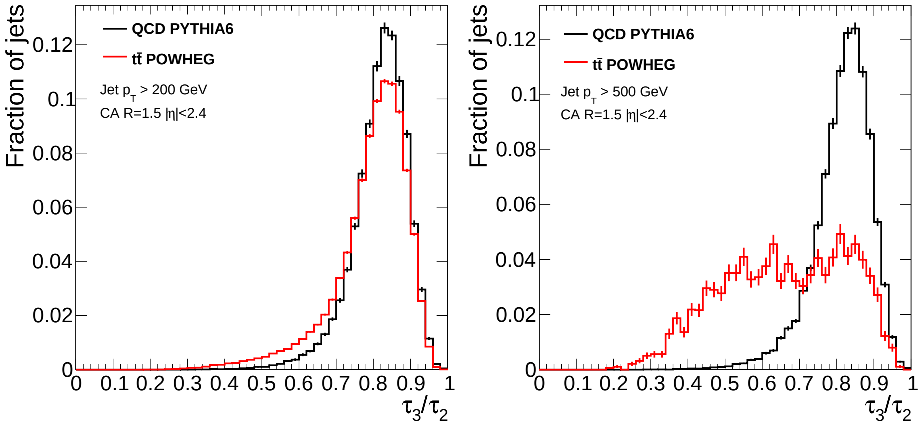


Figure 9.5: N-subjettiness ratio τ_3/τ_2 for signal and background events and with jets with $p_\mathrm{T} > 200\,\mathrm{GeV}$ (left) and $p_\mathrm{T} > 500\,\mathrm{GeV}$ (right).

to originate from a top quark decay, for example jets initiated by gluons or light quarks. In this section different top tagging algorithms that have been used in CMS physics analysis are compared, with new algorithms that will be available for Run II. These comparisons are done in simulation, where the generator information is available and the calculation of the efficiency and mistag rate is straightforward. The top tagging efficiencies and misidentification rates are calculated in accordance to Ref. [160], to ensure consistency. Here, the results from Ref. [160] were updated with additional ROC curves and published in Ref. [192].

The denominator for calculating the efficiency is the number of jets that pass the p_T selection and match to a generator top quark or anti-top quark. For the mistag rate the denominator is a matched gluon or light quark with the highest p_T from the hard scatter. The numerator is for both, efficiency and mistag rate, the number of jets that pass the criteria of the top tagging algorithm in addition to the matching. The efficiency is then calculated as

$$\epsilon_{\mathrm{eff,mis}} = \frac{N(\mathrm{Jets}_{\mathrm{tagged,matched}})}{N(\mathrm{Jets}_{\mathrm{matched}})}. \tag{9.1}$$

The matching criterion requires the distance between the jet axis and the top quark $\Delta R_{\mathrm{top,jet}}$ to be smaller than R_0. Since the conesize of CA15 jets is relatively large the matching criterion for these jets is $\Delta R_{\mathrm{top,jet}} < 1.0$. The matching efficiency depends on the conesize of the jet, large jets match already for low p_T quite often whereas smaller jets only match for higher transverse momenta. The matching efficiency in dependence of the matched parton p_T is shown in Figure 9.6. A jet is matched in Figure 9.6 if the corresponding angular distance between momentum axis of the top quark and the momentum axis of the jet is smaller than the distance parameter R_0 of the jet $\Delta R_{\mathrm{top,jet}} < R_0$. It can be seen that the highest matching efficiency can be achieved with the largest jet (here CA15). The matching efficiency for CA8 jets rises at slightly higher p_T. For reason of comparison also the matching efficiency for CA2 jets is shown, 100% efficiency is reached after $p_\mathrm{T} > 1000\,\mathrm{GeV}$.

To be able to compare the performance of different top taggers the working points have to be adjusted to the same efficiency values. For a given efficiency one has to compare

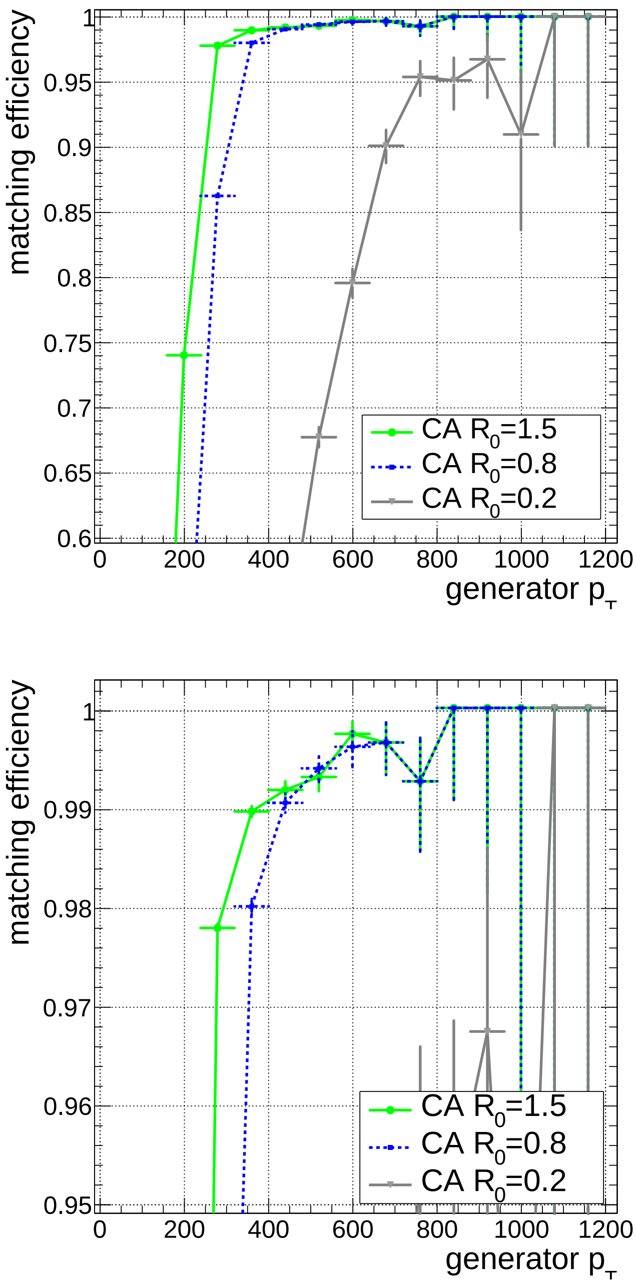


Figure 9.6: Matching efficiencies for Cambridge/Aachen jets with different conesizes, the matching criteria is always chosen such that the generator top quark axis has to be inside of the jet cone $\Delta R_{\text{jet,top}} < R_0$.

the mistag rates. The so-called Receiver Operating Characteristic curves (ROC curves) make it possible to compare different working points at the same time. In order to obtain the ROC curves for a given set of cuts on decision variables the efficiency is measured. For a given efficiency the minimum mistag rate is calculated. By varying the cuts on the decision variables the ROC curve of a top tagging algorithm can be derived.

The ROC curves for the CMS top tagger are found by scanning the minimum pairwise mass m_{min} and keeping the standard selections as described in Section 7.1 on the jet mass m_{jet} and the number of subjets N_{sub}. For the HEP top tagger a similar strategy is used, where the cut on the mass m_{123} is fixed and the W boson mass selection f_{W} is scanned. The ROC curve for the shower deconstruction tagger can be found by scanning the decision variable χ. Finally, the ROC curves for the OptimalR (MultiR) HEP top tagger are calculated by scanning three variables, each at the new working point of the optimal radius that is found $f_{\mathrm{W}}(R = R_{\mathrm{opt}})$, $m(R = R_{\mathrm{opt}})$ and $R_{\mathrm{opt}} - R_{\mathrm{opt}}^{\mathrm{calc}}$. Some top tagging algorithms are combined with N-subjettiness and a subjet b-tag. In this cases the scan is extended to one or two additional variables (τ_3/τ_2, CSV) [160, 192].

As introduced in Section 7, the different top tagging algorithms make use of jets with different conesizes, mostly CA8 and CA15 jets. As the matching efficiencies differ for the different jet conesizes, the ROC curves are split into categories of different jet conesizes. They are shown in Figure 9.7 - 9.9.

The ROC curves for different regions of p_{T} of the matched generator particle are produced and shown. This is important since the performance of tagging algorithms can differ from p_{T} region to p_{T} region and this is interesting for analyses that are sensitive to a specific p_{T} range. Some of the tagging algorithms do not perform well over the full p_{T} range.

In Figure 9.7 the ROC curves for the low p_{T} region are shown. The ROC curves of the HEP top tagger, the OptimalR HEP top tagger and a combination with N-subjettiness and b-tagging are displayed. In the low p_{T} region the use of multiple conesizes leads already to an improvement of the performance. The best performing combination in this p_{T} region is the HEP top tagger combined with N-subjettiness and a subjet b-tag. The standard working points are shown as markers. The combination of the OptimalR (MultiR) HEP top tagger with N-subjettiness and with an additional subjet b-tag are still under development and not shown here. The CMS top tagger as well as the shower deconstruction tagger do not work efficiently in this p_{T} region. Especially the CMS top tagger can not identify low p_{T} top quarks, since the decay products of the top quark will only rarely be collimated within a jet with a conesize of 0.8.

In Figure 9.8 three p_{T} regions are shown for tagging algorithms using a CA jet with a conesize of 1.5. In all of these p_{T} selections the shower deconstruction tagger combined with an additional subjet b-tag is the best performing algorithm. The improvement of the HEP top tagger due to the use of the OptimalR HEP top tagger is apparent. It can be seen that the OptimalR HEP top tagger reduces the mistag rate at a given efficiency by a factor of two in comparison to the original HEP top tagger. Furthermore, it can be observed that the plain shower deconstruction tagger and the plain OptimalR HEP top tagger show almost the same performance in these three p_{T} regions.

Algorithms using jets with a conesize of 0.8 are shown in Figure 9.9 for three different p_{T} regions. The best performing algorithm in all three p_{T} regions is the CMS combined tagger, which uses also a selection on N-subjettiness and a subjet b-tag. A similar performance is observed for the shower deconstruction tagger combined with a subjet b-tag. In this

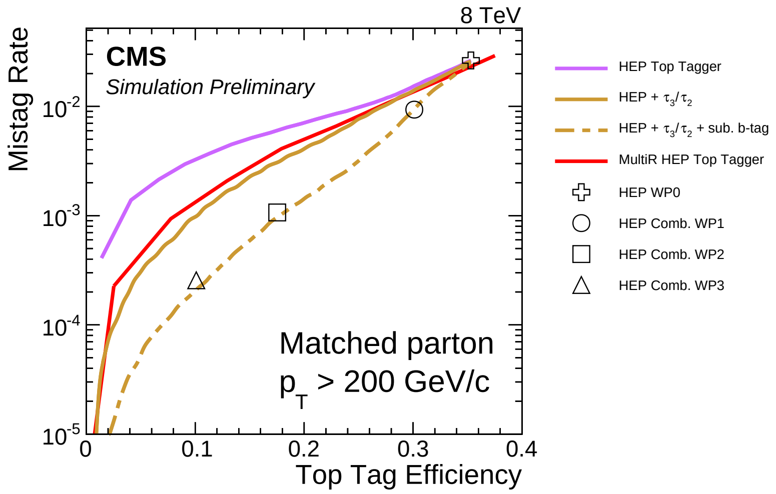


Figure 9.7: ROC curves for CA15 jets with a $p_{\mathrm{T}} > 200\,\mathrm{GeV}$, published in Ref. [192].

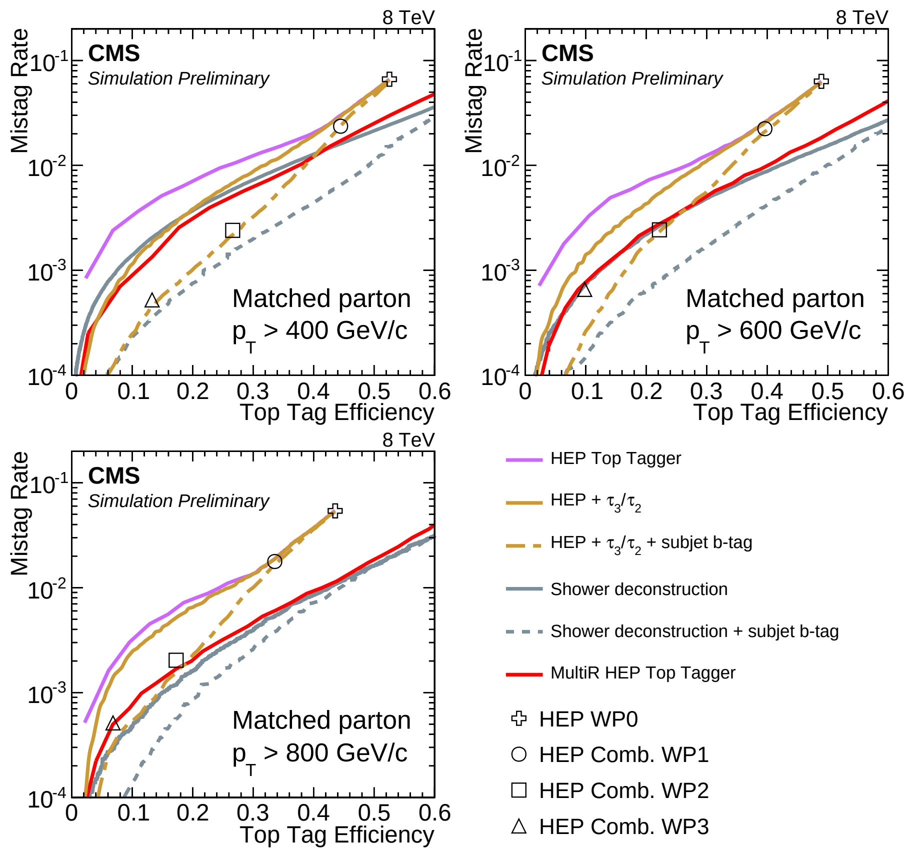


Figure 9.8: ROC curves for CA15 jets in different p_{T} slices, published in Ref. [192].

case the subjets reconstructed by the CMS top tagger are used as input for subjet b-tagging. The BDT, trained with ten substructure variables, shows a good performance in all p_T regions. It is the best tagging algorithm without using subjet b-tagging information. Furthermore, the BDT has the best performance in the high efficiency range, and can be especially interesting for analyses that need high signal efficiencies. Although the BDT performs in the lower and middle efficiency range just slightly better than the shower deconstruction tagger. Since these two top tagging algorithms use different information a combination of the BDT with shower deconstruction might be interesting.

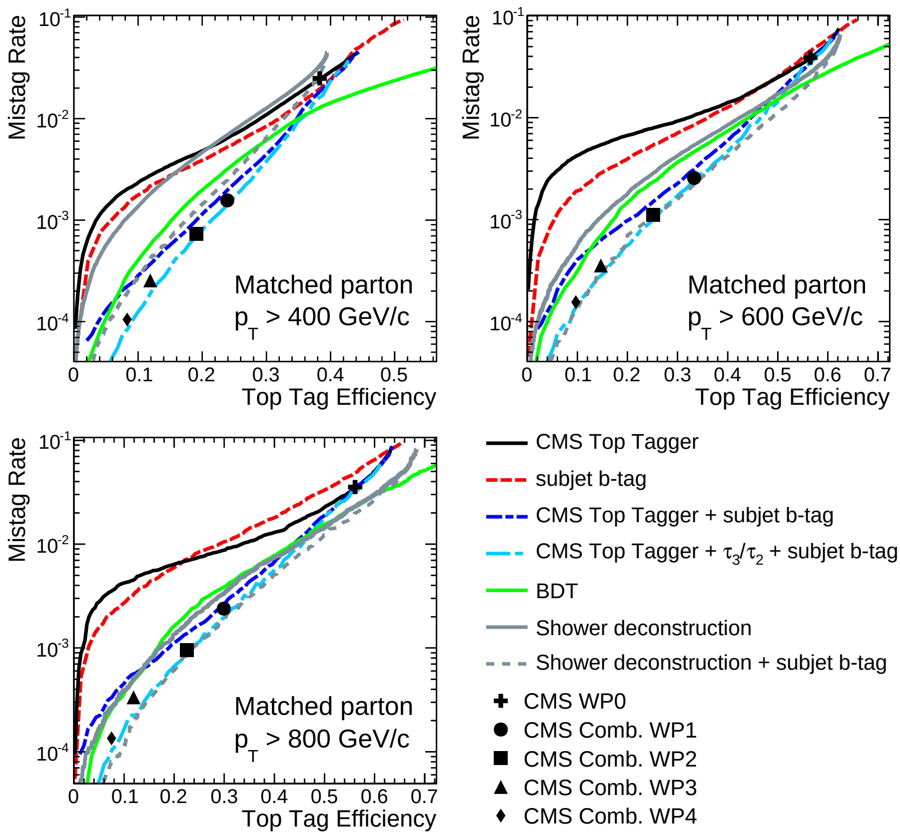


Figure 9.9: ROC curves for CA8 jets in different p_T regions, published in Ref. [192].

For very high $p_T > 800\,\text{GeV}$, the matching efficiency of the two jet conesizes (CA8, CA15) is almost 100% and therefore the categorization into CA8 and CA15 jets is not needed anymore. In this high p_T region the ROC curves are comparable for both jet conesizes. In Figure 9.10 the ROC curves from taggers applied to CA8 and CA15 jets are shown. It can be observed that the OptimalR HEP top tagger performs almost as well as the CMS top tagger combined with N-subjettiness. Especially in the high p_T regions the usage of multiple conesizes gives a significant improvement. As already seen in Figure 9.9, the best performing algorithms are the CMS combined top tagger and shower deconstruction combined with a subjet b-tag.

In summary, the ROC curves presented provide an easy way of comparing the per-

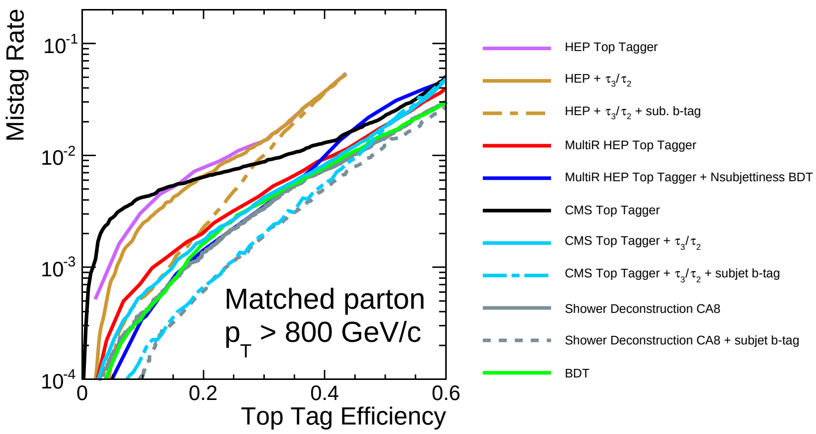


Figure 9.10: ROC curves for CA8 and CA15 jets with $p_{\mathrm{T}} > 800\,\mathrm{GeV}$, published in Ref. [192].

formance of different tagging algorithms. Over almost the whole p_{T} range the shower deconstruction tagger, the CMS top tagger combined with N-subjettiness, the OptimalR HEP top tagger and the BDT are the best performing algorithms. Focusing on Run II with high p_{T} top quarks, the ROC curves in Figure 9.10 are interesting and should be a guide to improved top tagging algorithms.

10 Measurement of top tagging efficiency and misidentification in 8 TeV data

In Section 9 different top tagging algorithms are compared in simulation, whereas in this chapter simulated events are compared to data with the aim of validating the top tagging algorithms and the simulation of substructure variables. In case some decision variables are not well described by the Monte Carlo simulations, it will cause differences in the performance of the tagging algorithms in data and simulated events. Distributions of substructure variables in a $t\bar{t}$ signal region and a QCD multijet background region are compared in order to assess possible differences that can cause deviations in the tagging performance. Furthermore, measurements of the efficiency and mistag rate of various taggers are presented, done with the data collected at a centre-of-mass energy of 8 TeV in 2012 and an integrated luminosity of 19.7 fb^{-1}. The data samples used for this analysis are shown in Table 10.1.

Data

Channel	Dataset	Int. Luminosity (pb^{-1})
muon	/SingleMu/Run2012A-22Jan2013-v1/AOD	876
	/SingleMu/Run2012B-22Jan2013-v1/AOD	4411
	/SingleMu/Run2012C-22Jan2013-v1/AOD	7055
	/SingleMu/Run2012D-22Jan2013-v1/AOD	7369
jet	/Jet/Run2012A-22Jan2013-v1/AOD	876
	/JetHT/Run2012B-22Jan2013-v1/AOD	4411
	/JetHT/Run2012C-22Jan2013-v1/AOD	7055
	/JetHT/Run2012D-22Jan2013-v2/AOD	7369

Table 10.1: CMS data sets used for the 8 TeV analysis, the first four data sets are collected with a single muon trigger and are used for the signal selection, the second four are used for the background selection.

The measurement of the efficiency and mistag rate in data is more challenging than the calculation in simulation since the generator information can not be used. It has to be ensured that only jets caused by a hadronically decaying top quarks are used to measure the efficiency or that only jets coming from QCD multijet events are used to measure the mistag rate. Therefore, signal and background regions are defined by selecting $t\bar{t}$ and QCD multijet events. In both samples the tag and probe method is used. In case of the background selection, events with a di-jet topology are selected, where one of the jets is chosen as tag and the other one as probe jet. In case of the signal selection, semileptonic $t\bar{t}$ decays are utilized, where the leptonic decay leg $t \to Wb \to l\nu$ is chosen as the tag and the hadronically decaying top quark as a probe jet. Since the top tagging algorithms are applied on jets with different distance parameters (see Section 7), each selection is done for two different CA jet collections (CA8, CA15).

10.1 Signal selection

The efficiency of top tagging algorithms can be measured in a sample enriched with $t\bar{t}$ events. The semileptonic channel is chosen to enrich the sample with $t\bar{t}$, meaning that one of the W bosons from the top quark decay is decaying hadronically and the other leptonically. This choice of the semileptonic decay channel provides an easy choice of the probe jet since it ensures in each event exactly one hadronically decaying top quark. The measurement is performed in the muon channel, since the best signal to background ratio is obtained there.

Events contained in the muon data stream are used, recorded by a single muon trigger without isolation requirement. The trigger name is `HTLMu40eta2p1v*`, and it selects events in which the reconstructed muon has a p_T greater than 40 GeV after the L1 reconstruction. The trigger efficiency is around 95% [193] with very good agreement between data and simulation. Residual differences are corrected for using muon measurements provided by the muon POG [193]. In order to reject events not originating from $t\bar{t}$ production, the following selection is applied:

- At least one good primary vertex with the criteria described in Section 6.1.
- Exactly one muon with $p_T > 45$ GeV and $|\eta| < 2.1$ and the criteria described in Section 6.2.3.
- Veto on electrons with $p_T > 35$ GeV and $|\eta| < 2.5$ (see additional criteria in Section 6.2.2).
- Lepton 2D-cut: the event is accepted if either $\Delta R(\mu, \mathrm{j}) > 0.5$ or $p_{T,\mathrm{rel}}(\mu, \mathrm{j}) > 25$ GeV, where j stands here for the nearest jet in ΔR to the muon candidate. The quantity $p_{T,\mathrm{rel}}(\mu, \mathrm{j})$ is the relative transverse momentum of the leading muon with respect to the closest jet in ΔR. Both quantities are calculated with respect to all anti-k_T jets with a distance parameter of 0.5 and with $p_T > 25$ GeV. This cut reduces the background from QCD multijet production, while retaining a sizable efficiency for high p_T top quarks. The efficiency of this cut has been studied in Ref. [64].
- Missing transverse energy: the missing transverse energy $\not{E}_T$ has to be at least 20 GeV.
- $H_T^{\mathrm{lep}} > 150$ GeV, where $H_T^{\mathrm{lep}} = \not{E}_T + p_T^{\mu}$.
- At least one b-tag at the CSV medium working point with a nominal background rejection of 99% [185] applied on an anti-k_T jets with a distance parameter of 0.5.
- At least one large CA jet with $p_T > 200$ GeV and $|\eta| < 2.4$.
  This jet is a candidate for the probe jet and usually contains the hadronically decaying top quark. The selection is performed once for CA8 jets, and once for CA15 jets. The differences in the two selections are mostly the kinematics, where for the CA15 selection the jet transverse momenta will be slightly higher on average.

Process	fraction of events [%]	
	CA8 selection	CA15 selection
$t\bar{t}$	72.8 ± 1.0	74.4 ± 1.0
Single Top	8.4 ± 0.4	7.6 ± 0.3
W+jets	16.0 ± 0.5	14.7 ± 0.4
Other	2.8 ± 0.2	2.8 ± 0.2

Table 10.2: Sample composition after the signal selection estimated from MC after scaling the $t\bar{t}$ POWHEG events by a factor of 0.9.

After this selection the sample is highly enriched with $t\bar{t}$ events. The event composition after the selection is summarized in Table 10.2. The signal selection is performed for two MC event generators, $t\bar{t}$ POWHEG interfaced to PYTHIA and $t\bar{t}$ MC@NLO interfaced to HERWIG++. The $t\bar{t}$ POWHEG sample is scaled by a factor of 0.9 to correct the normalization to the one observed in data, whereas no scaling is needed for the MC@NLO sample.

10.1.1 Reconstruction of the leptonically decaying top quark

The leptonically decaying top quark is reconstructed by building the four-momenta of the neutrino, the muon and one of the anti-k_T jets with a conesize of 0.5 with $p_T > 50\,\mathrm{GeV}$ [194]. The four-momentum of the neutrino is calculated by using the missing transverse energy $\not{E}_T$. It is mandatory that each event contains exactly one muon, to ensure the missing transverse energy comes from just one leptonically decaying W boson and therefore from the only neutrino in the event. The z-component of the momentum can be obtained by requiring momentum conservation,

$$M_W^2 = p_W^2 = (p_\mu + p_\nu)^2. \tag{10.1}$$

The parameters p_W, p_μ and p_ν are the four-momenta of the W boson, the muon and the neutrino. The x and y components of the neutrino are taken directly from the x and y components of the missing transverse energy. Solving Equation (10.1) leads to:

$$p_{z,\nu} = \alpha \frac{p_{z,\mu}}{p_{\mathrm{T},\mu}^2} \pm \sqrt{E_\mu^2 \frac{\alpha^2 - p_{\mathrm{T},\nu}^2 p_{\mathrm{T},\mu}^2}{p_{\mathrm{T},\mu}^4}}, \tag{10.2}$$

$$\alpha = \left(\frac{M_W^2 - M_\mu^2}{2} + p_{\mathrm{T},\mu} p_{\mathrm{T},\nu} \cos(\Delta\Phi) \right). \tag{10.3}$$

Equation (10.2) can have imaginary solutions, if the following condition is not fulfilled:

$$M_W^2 - M_\mu^2 = M_{W,\mathrm{eff}}^2 < 2 p_{\mathrm{T},\mu}^2 p_{\mathrm{T},\nu}^2 (1 - \cos(\Delta\Phi)). \tag{10.4}$$

If this happens, the term under the square root in Equation (10.2) is discarded and a single real solution is used. Considering all possible combinations of anti-k_T jets with

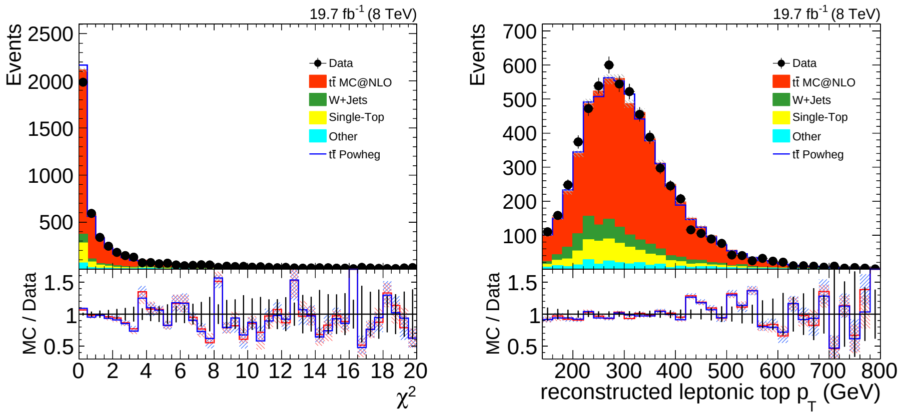


Figure 10.1: χ^2 distribution for the reconstruction of the leptonically decaying top quark (left) and p_T distribution of the reconstructed top quark (right) for the signal selection. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom panel, where the ratio to POWHEG (MC@NLO) is shown in blue (red). The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

$R = 0.5$, each event has $2N$ reconstruction hypotheses, where N is the number of jets and the factor of 2 comes from the ambiguity in the neutrino reconstruction. To find the best hypothesis, a discrimination variable χ^2 can be calculated as

$$\chi^2 = \frac{(m_{\text{top,lep}} - \overline{m}_{\text{top,lep}})^2}{\sigma^2_{\text{lep}}}. \tag{10.5}$$

The value $m_{\text{top,lep}}$ is the reconstructed mass of the top quark candidate, $\overline{m}_{\text{top,lep}}$ and σ_{lep} are the expected top quark mass and the standard deviation of this distribution, as obtained from simulation [64]. In each event only the hypothesis with the lowest value of χ^2 is kept, and used to calculate the four-momentum of the leptonically decaying top quark. In Figure 10.1 the χ^2 distribution and the p_T distribution of the reconstructed leptonic top quark are shown. It can be seen that the χ^2 distribution peaks at low values of χ^2, indicating that the leptonically decaying top quark is well reconstructed. Both distribution are well described by the simulation.

10.1.2 Tag and probe method

After reconstructing the leptonically decaying top quark, the CA jet on the other hemisphere is taken as a probe jet. The probe jet has to have a distance larger than 2.1 in the Φ-plane to the reconstructed leptonically decaying top quark. In Figure 10.2 a schematic drawing of a semi-leptonic event can be seen, indicating the tag and probe jets. In Figure 10.3 the ΔR distribution between the reconstructed leptonic top quark and the probe jet is shown. One can observe that the distribution peaks around π, showing that the reconstructed leptonically decaying top quark and the probe jet are on opposite hemispheres.

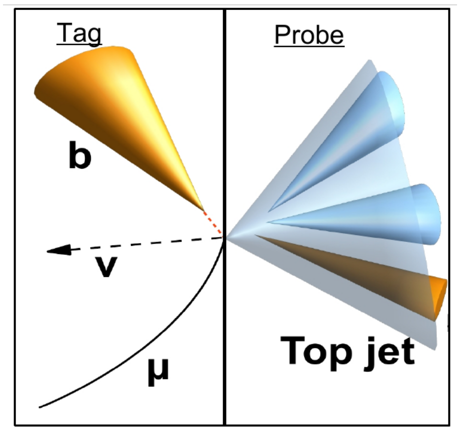


Figure 10.2: Semileptonic $t\bar{t}$ event, the leptonically decaying top quark is the tag, the jet on the other hemisphere the probe. Taken from [195] and modified.

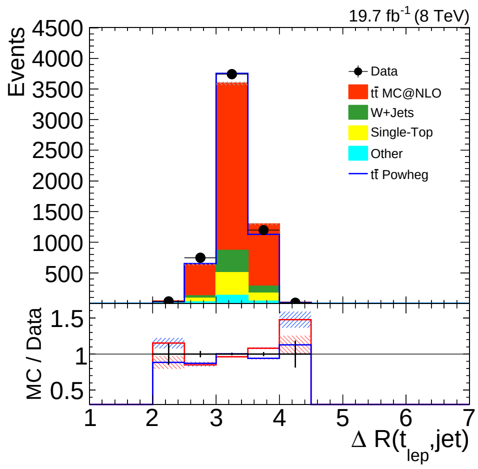


Figure 10.3: Distribution of ΔR between the reconstructed leptonic top quark and the probe jet in the signal selection. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom panel, where the ratio to POWHEG (MC@NLO) is shown in blue (red). The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

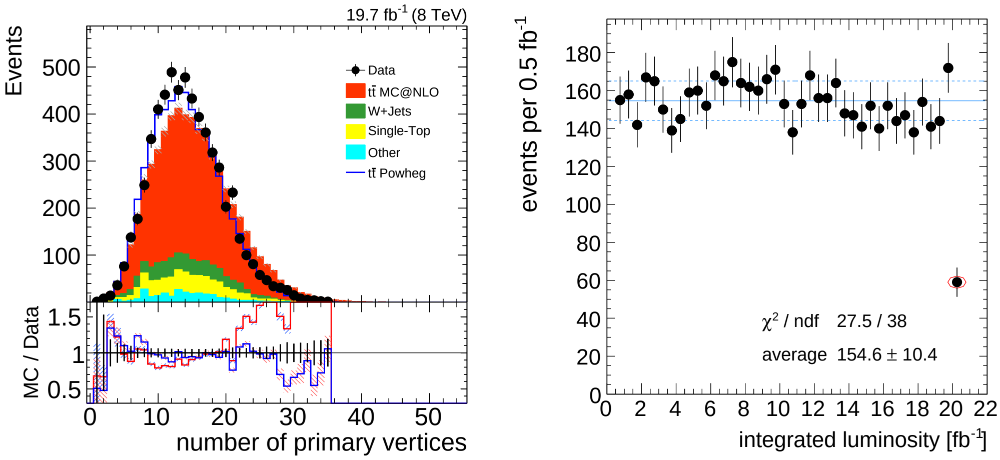


Figure 10.4: Number of primary vertices (left) and events per integrated luminosity (right) for the signal selection. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom panel, where the ratio to POWHEG (MC@NLO) is shown in blue (red). The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

10.1.3 Results of the $t\bar{t}$ signal selection

The description of the number of primary vertices N_{pv} is important since the efficiency as well as the mistag rate will be measured as function of N_{pv} in order to investigate their pile-up dependence. Ideally, the efficiency and the mistag rate are independent of the number of primary vertices. In Figure 10.4 (left) the distribution of the number of reconstructed primary vertices is shown. The number of primary vertices shows good agreement between data and simulation for the POWHEG sample, whereas the MC@NLO sample has difficulties to describe the data for a high number of primary vertices. Figure 10.4 (right) shows that the number of selected events is independent of the time the events are recorded, showing the stability of the CMS data taking over time.

Control distributions of the probe jet using CA8 jets are shown in Figure 10.5. Similar distributions for the probe CA15 jet using the CA15 selection are shown in Figure 10.6. The p_{T} and η spectra are well modeled by the simulation, after scaling the normalization of the $t\bar{t}$ POWHEG simulation by a factor of 0.9. This effect has already been observed in Ref. [160].

After the selection the signal sample consists of about 15-16% of events from W+jets production and about 8% of single top events. The top quarks from the single top production decay in 66% leptonically after the selection and are therefore considered as background events. These background events are subtracted for the efficiency measurement,

$$\epsilon = \frac{N(\text{tagged probe jets,}t\bar{t}\ ) - N(\text{tagged jets (background)})}{N(\text{all probe jets, }t\bar{t}) - N(\text{all jets (background)})}. \tag{10.6}$$

The denominator counts all probe jets, whereas the numerator includes the number of jets that also pass the top tagging requirements for a given tagging algorithm. In the denom-

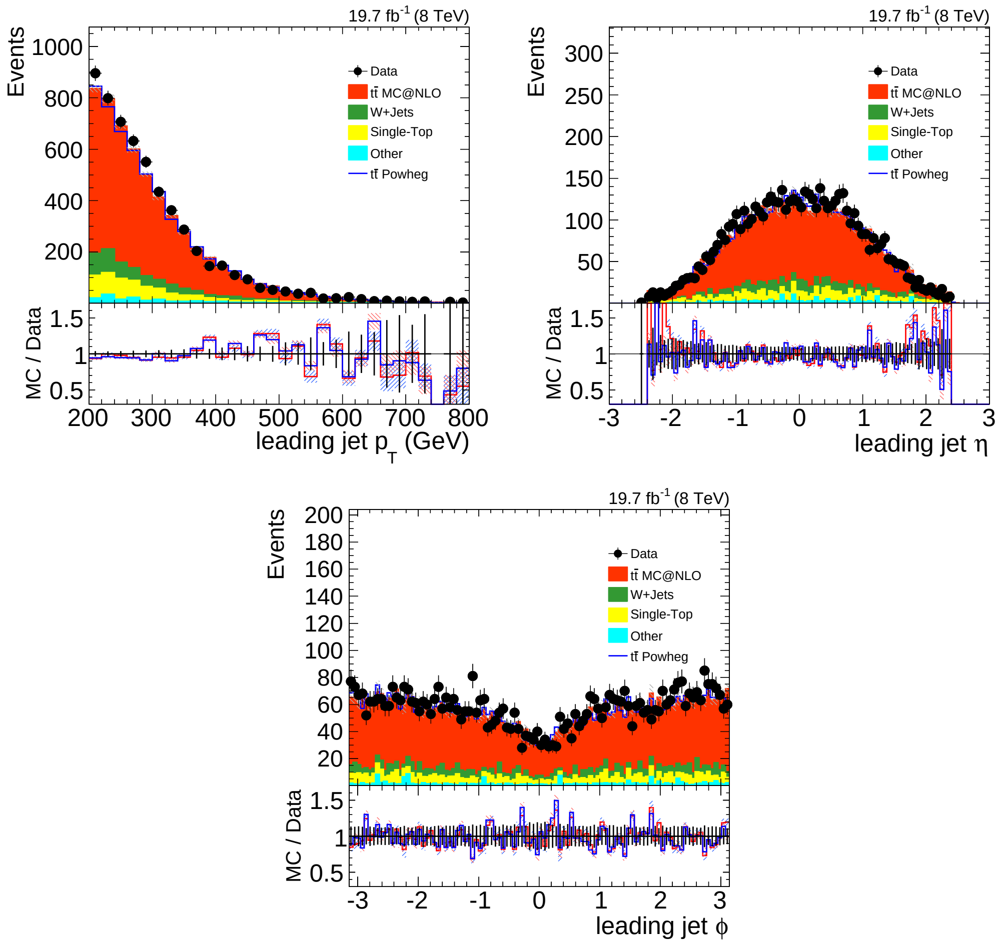


Figure 10.5: Transverse momentum p_T for the leading CA8 jet after the signal selection (top left), the η distribution of the leading CA8 jet (top right) and the Φ distribution (bottom middle) for the signal selection. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom panel, where the ratio to POWHEG (MC@NLO) is shown blue (red). The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

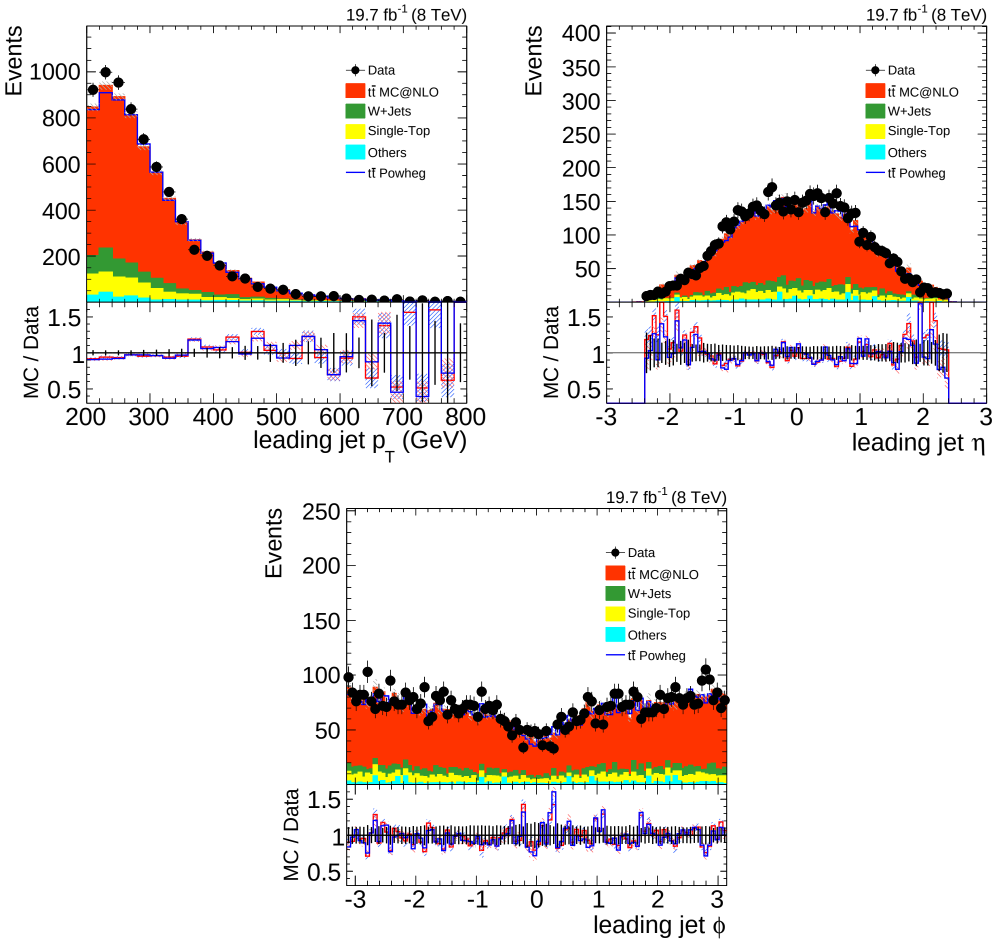


Figure 10.6: Transverse momentum p_T for the leading CA15 jet after the signal selection (top left), the η distribution of the leading CA15 jet (top right) and the Φ distribution (bottom middle) for the signal selection. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom panel, where the ratio to POWHEG (MC@NLO) is shown in blue (red). The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

inator as well as in the numerator, the W+jets and single top background is subtracted. Technically, for calculating the efficiency and mistag rate two histograms are filled (one for all probe jets and one for all tagged probe jets) with the p_{T} (N_{pv}, η) of the corresponding jet. Afterwards the quotient is built bin-wise in p_{T} (N_{pv}, η).

10.2 Background selection

The main background for top tagging originates from jets produced by gluons or light quarks. To measure the misidentification rate of the top tagging algorithms, a sample is enriched with mostly gluon or light quark jets by selecting QCD multijet events with di-jet-like topologies. Events from $\mathrm{t\bar{t}}$ production in the all-hadronic channel are suppressed by designated selection criteria. A trigger based on the scalar sum of the transverse momenta of all jets in the event is used, defined by

$$H_{\mathrm{T}} = \sum_{\mathrm{jets}} p_{\mathrm{T,jets}}. \tag{10.7}$$

The HLT trigger name is `HLT_HT750_v*` and it selects events with $H_{\mathrm{T}} > 750\,\mathrm{GeV}$.
The following cuts are applied to select di-jet events:

- At least one good primary vertex with the criteria described in Section 6.1.
- Veto on electrons and muons (see criteria in Sections 6.2.2 and 6.2.3).
- $H_{\mathrm{T}} > 1000\,\mathrm{GeV}$ to ensure a fully efficient trigger.
- At least two large CA jets with $p_{\mathrm{T}} > 400\,\mathrm{GeV}$ and $|\eta| < 2.4$ and a $\Delta\Phi$ between them of 2.1. Similar to the signal selection, two background selections are distinguished, one obtained by using CA8 jets and the other one using CA15 jets.

To measure the misidentification rate, one of the two jets in the event is selected randomly as a tag jet. The selected tag jet has to require the following criteria:

- The mass of the jet has to be in a window $140 < m_{\mathrm{jet}} < 250\,\mathrm{GeV}$ around the top quark mass, in order to achieve comparable kinematics as in $\mathrm{t\bar{t}}$ production.
- The minimum pairwise mass calculated with the subjets found by the CMS top tagger described in Section 7 has to be smaller than 50 GeV, to further reduce the contamination by hadronically decaying top quarks.

The second selection criterion is applied in order to remove any residual contribution from $\mathrm{t\bar{t}}$ production after the top tagging algorithms have been applied. If the criteria are fulfilled, the jet on the opposite hemisphere to the tag jet is taken as the probe jet and is used to measure the misidentification rate. Otherwise, the event is discarded.

The misidentification rate is calculated similar to the efficiency given in Equation (10.6). As the background selection is composed of only QCD multijet events (see Table 10.3), no background subtraction is needed.

In the following, the MC event generators PYTHIA [120] with the UE tune "TuneZ2star" and HERWIG++ [121] with the UE tune "TuneEE3C" are used in order to simulate the QCD multijet processes. The study is repeated in Ref. [196] with the PYTHIA 8 event

Process	fraction of events [%]	
	CA8 selection	CA15 selection
QCD	99.1 ± 0.2	99.2 ± 0.1
Others	0.9 ± 0.0	0.8 ± 0.0

Table 10.3: Sample composition after the di-jet background selection.

generator, the distributions can be found in Appendix B. Good agreement between data and simulated events is achieved by reweighting the p_T distribution of the leading jet, as described below in Section 10.2.2.

After the selection, the sample consists dominantly of events from QCD multijet production. In Figure 10.7 the number of primary vertices after the reweighting is shown, the agreement between data and simulated events is good for both event generators. The jet kinematics of the probe jet are well described after the reweighting, as shown in Figure 10.8 and 10.9.

10.2.1 W+jets background

To ensure that the sample is not contaminated by events from W+jets production a test is performed, where the W boson decays hadronically. Jets from these processes would have different substructure distributions and consequently different mistag rates. Since the available W+jets Monte Carlo samples are only available for the leptonically decay channel, a sample with $W \to qq'$ and a statistics corresponding to an integrated luminosity of $27.16\,\mathrm{pb}^{-1}$ is produced. Applying the background selection on this sample, no events

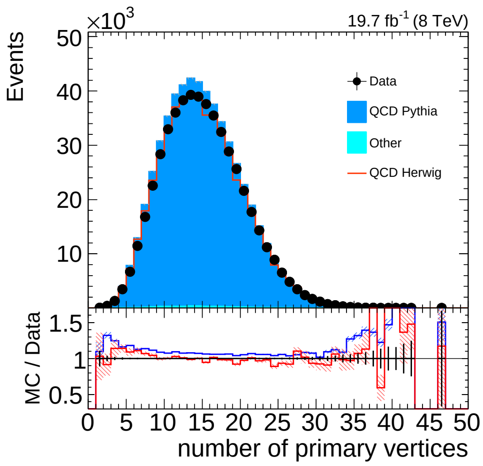


Figure 10.7: Number of primary vertices for the background selection. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom panel, where the ratio to PYTHIA is reported in blue while the one to HERWIG ++ is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

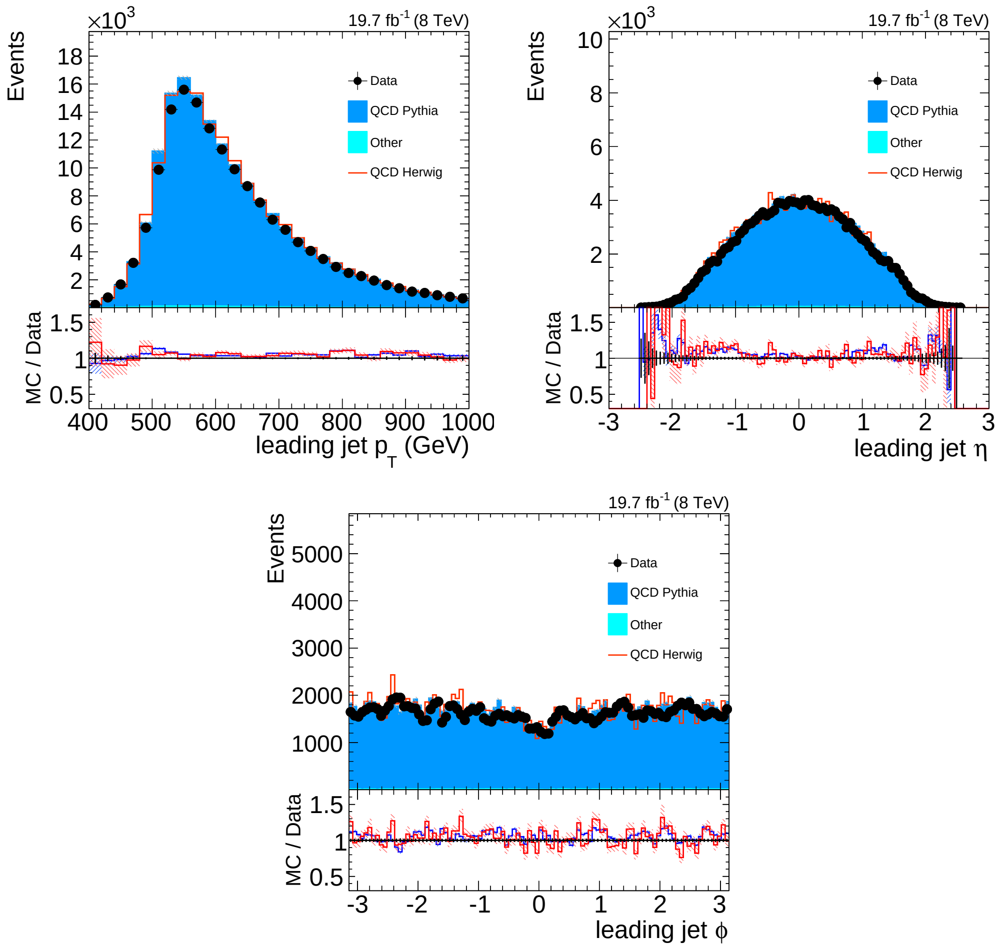


Figure 10.8: Distributions of p_{T} (left), η (right) and ϕ (bottom) for the CA8 probe jet after the background selection. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom panel, where the ratio to PYTHIA (HERWIG++) is shown in blue (red). The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

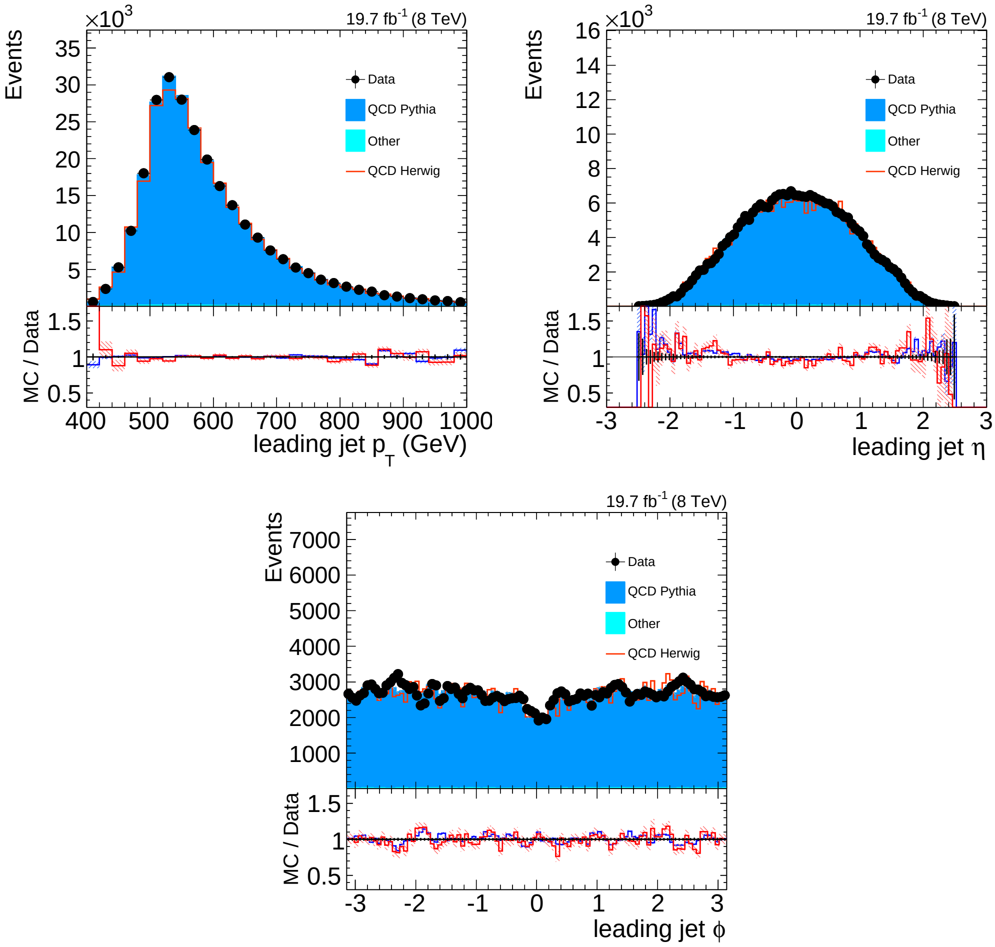


Figure 10.9: Distributions of p_{T} (left), η (right) and ϕ (bottom) for the CA15 probe jet after the background selection. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom panel, where the ratio to PYTHIA (HERWIG++) is shown in blue (red). The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

passing the selection are observed. This corresponds to a null measurement, with an uncertainty of about 2500 events, when scaled to the luminosity of the data. Since the selected di-jet sample consists of 163000 events, a maximum contamination of 1.5% from W+jets production is possible, which can be neglected.

10.2.2 Reweighting of the leading jet p_{T} for QCD multijet events

A mismodeling of the p_{T} distribution of the leading jet is observed for both MC simulations, as shown in Figures 10.10 - 10.13. Since the shapes of most substructure variables have a p_{T} dependence, this could cause a difference between data and simulation when comparing jet substructure distributions. As one is not interested in the description of the kinematics of the hard $2 \rightarrow 2$ process, the leading jet p_{T} distributions are reweighted. For each jet collection (CA8 and CA15) and each MC simulation, a weight function is derived from the ratio between data and simulated events. The resulting weight functions for CA8 jets are:

$$w_{\mathrm{Pythia}}(p_{\mathrm{T},1}) = (-0.0004 \cdot p_{\mathrm{T},1} + 1.59), \tag{10.8}$$

$$w_{\mathrm{Herwig}}(p_{\mathrm{T},1}) = (0.0013 \cdot p_{\mathrm{T},1} + 1.49), \tag{10.9}$$

and for the selection using CA15 jets:

$$w_{\mathrm{Pythia}}(p_{\mathrm{T},1}) = (1.1 \cdot e^{-0.5 \cdot ((p_{\mathrm{T},1}-451)/47.4)^2} + 1.37 - 3.04 \cdot 10^{-4} \cdot p_{\mathrm{T},1}), \tag{10.10}$$

$$w_{\mathrm{Herwig}}(p_{\mathrm{T},1}) = (0.0009 \cdot p_{\mathrm{T},1} + 1.48), \tag{10.11}$$

where $p_{\mathrm{T},1}$ is the transverse momentum of the leading jet. The distributions before and after the reweighting are shown in Figures 10.10 - 10.13. After the reweighting good agreement between data and simulated events is achieved. This is also observed for other kinematic quantities which are not shown here.

10.3 Systematic uncertainties

In the efficiency and mistag rate measurements statistical and systematic uncertainties are shown, where the systematic uncertainties come from jet energy correction, jet energy resolution and top p_{T} reweighting. Only uncertainties that change the shape of variables are considered since uncertainties that change the number of events have no impact on the efficiency and mistag rate measurement.

The uncertainties of the jet energies are provided by the JetMET group [197]. Technically, the uncertainties shift the reconstructed jet energy as a function of its transverse momentum and its pseudorapidity, where the full four-vector is changed. The shift is also propagated to the missing transverse energy.

The uncertainties of the jet energy resolution (described in Section 6.3.5) are shown in Table 10.4.

The distributions of the transverse momentum of the top quarks show discrepancies between data and simulated $\mathrm{t\bar{t}}$ events in the lepton+jets [76,199] and in the dilepton [199] channel. Based on this measurements the top quark p_{T} distribution can be reweighted by applying a weight ω to each $\mathrm{t\bar{t}}$ event [200],

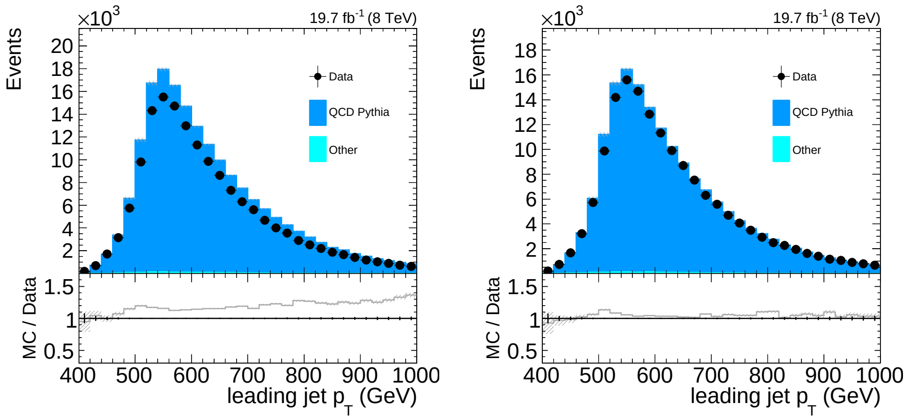


Figure 10.10: Distributions of the leading jet p_{T} before (left) and after (right) the reweighting of the PYTHIA sample for CA8 jets. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom panel, where the ratio to PYTHIA is reported in grey. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

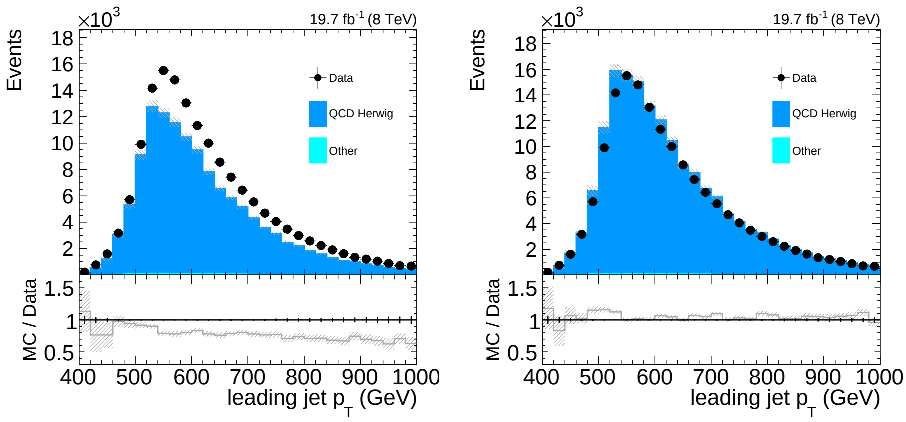


Figure 10.11: Distributions of the leading jet p_{T} before (left) and after (right) the reweighting of the HERWIG++ sample for CA8 jets. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom panel, where the ratio to HERWIG++ is reported in grey. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

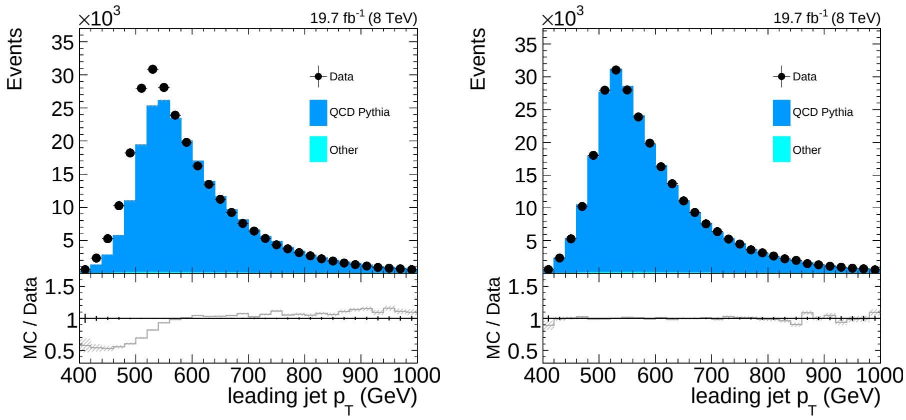


Figure 10.12: Distributions of the leading jet p_{T} before (left) and after (right) the reweighting of the PYTHIA sample for CA15 jets. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom panel, where the ratio to PYTHIA is reported in grey. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

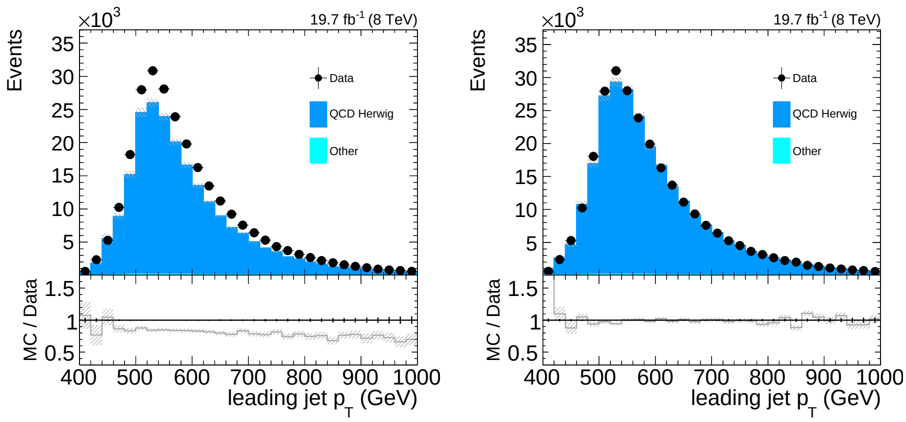


Figure 10.13: Distributions of the leading jet p_{T} before (left) and after (right) the reweighting of the HERWIG++ sample for CA15 jets. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom panel, where the ratio to HERWIG++ is reported in grey. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

η_{jet}	JER scale factor
$0 < \eta_{\text{jet}} < 0.5$	$1.052^{+0.063}_{-0.062}$
$0.5 < \eta_{\text{jet}} < 1.1$	$1.057^{+0.057}_{-0.056}$
$1.1 < \eta_{\text{jet}} < 1.7$	$1.096^{+0.065}_{-0.064}$
$1.7 < \eta_{\text{jet}} < 2.3$	$1.134^{+0.094}_{-0.092}$
$\eta_{\text{jet}} > 2.3$	$1.288^{+0.200}_{-0.199}$

Table 10.4: Jet energy resolution scale factors and uncertainties [198].

$$\omega = \sqrt{\exp(0.156 - 0.00137 \cdot p_{\text{T}}^{\text{t}}) \exp(0.156 - 0.00137 \cdot p_{\text{T}}^{\bar{\text{t}}})}, \tag{10.12}$$

where p_{T}^{t} and $p_{\text{T}}^{\bar{\text{t}}}$ are the transverse momenta of the generator top quark and anti-top quark. The uncertainty of the p_{T} top quark scale factor is 10% in a p_{T} range of $300 < p_{\text{T}} < 400\,\text{GeV}$, up to 20% in $400 < p_{\text{T}} < 600\,\text{GeV}$ and up to 50% and more for $p_{\text{T}} > 600\,\text{GeV}$.

10.4 Results of the data/MC comparison

10.4.1 CMS top tagger

In Figure 10.14 the jet mass distributions for the signal and background selection are shown. Since the CMS top tagger (CMSTT) applies a jet mass selection, a good description of this variable is necessary for a good simulation of the efficiency and mistag rate. For the signal selection the jet mass distribution has two peaks, one in the region of the W boson mass, the other one at the top quark mass. The agreement between data and MC is sufficient over the whole distribution. In case of the background selection, the event generator HERWIG++ seems to describe the data better than PYTHIA, although below 50 GeV the agreement is insufficient for both generators.

The minimum pairwise mass distribution is shown in Figure 10.15. It is well described for the signal selection for values above 60 GeV. A discrepancy is visible for low values of m_{min}, which has already been observed in earlier studies [160]. At around 80 GeV, a peak from the decay of the W boson can be observed. While the peak region in the background sample is not well described by PYTHIA, the HERWIG++ generator shows a better description. The opposite behavior is observed in the tail towards high values of m_{min}.

The number of subjets found by the CMS top tagger is shown in Figure 10.16. There is fair agreement between data and MC in the signal selection, albeit a slight slope in the ratio of data to simulation is observed. The description in the background region by HERWIG++ is good, while PYTHIA has problems to simulate the subjet multiplicity correctly. Comparing the signal with the background regions, it is confirmed that the number of subjets is a powerful variable to discriminate between jets originating from collimated top quark decays and light flavor jets. The three substructure variables are described similarly for the signal selection by the event generators POWHEG and MC@NLO.

The efficiencies and the mistag rates in dependence of the jet p_{T} are shown in Figure 10.17. For the measurement of the efficiency and the mistag rate the selection criteria

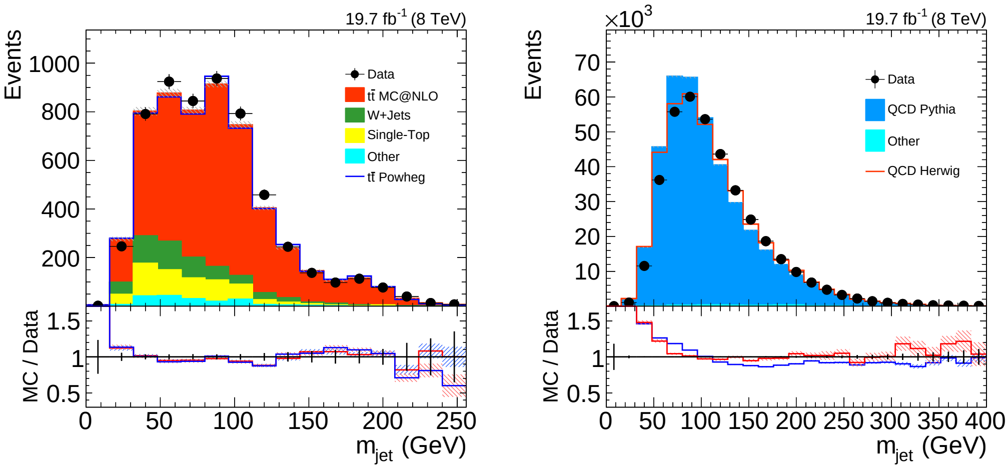


Figure 10.14: Jet mass for the signal selection (left) and the background selection (right). Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

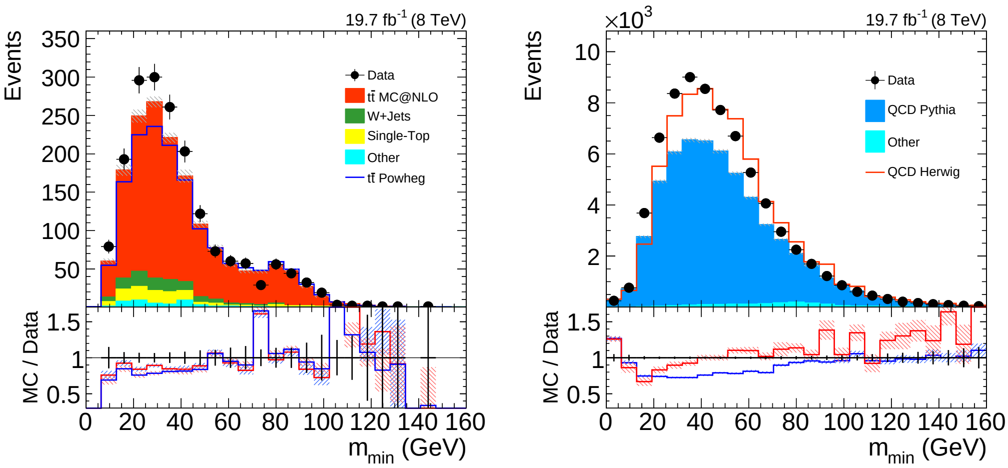


Figure 10.15: Minimum pairwise mass for the signal selection (left) and for the background selection (right). Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

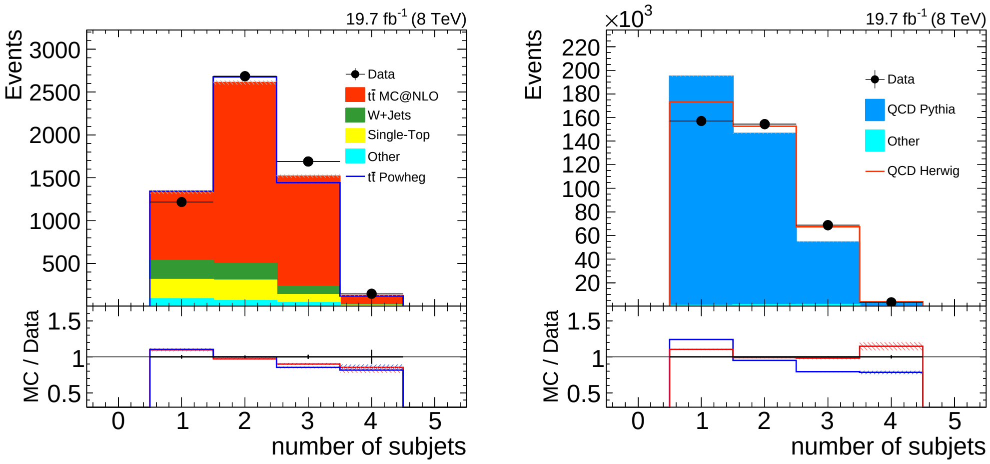


Figure 10.16: Number of subjets found by the CMS top tagger for the signal selection (left) and for the background selection (right). Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

Decision variable	selection
m_{jet}	$140\,\text{GeV} < m_{\text{jet}} < 250\,\text{GeV}$
m_{min}	$> 50\,\text{GeV}$
N_{sub}	≥ 3

Table 10.5: Selection criteria for the CMS top tagger.

listed in Table 10.5 are applied.

The efficiency of the CMS top tagger in data agrees well with the simulation within the statistical and systematic uncertainties. The efficiency increases with increasing p_{T} up to 50%. After $p_{\text{T}} > 700\,\text{GeV}$ the efficiency drops. The integrated efficiency for $p_{\text{T}} > 350$ GeV is 28%. The efficiency for jets with a p_{T} smaller than 400 GeV is very small. This is mostly due to the used distance parameter of the jets. The decay products of the top quark decay might not be collimated enough to be reconstructed within a jet with a distance parameter of 0.8 and therefore the mass will be reconstructed at lower values than the mass selection applied on m_{jet}. As expected, the mistag rate in data and MC agrees better with HERWIG++ than with PYTHIA simulation, since the substructure decision variables are better described by HERWIG++. The mistag rate for the CMS top tagger increases up to 6% at values of p_{T} around 800 GeV, and drops with increasing p_{T} similarly to the efficiency. The integrated mistag rate in HERWIG++ for $p_{\text{T}} > 350\,\text{GeV}$ is 5.4%, for PYTHIA 5.1%.

In Figure 10.18 the efficiency is shown as function of the reconstructed number of primary vertices. The efficiency, as well as the mistag rate, should be independent of the

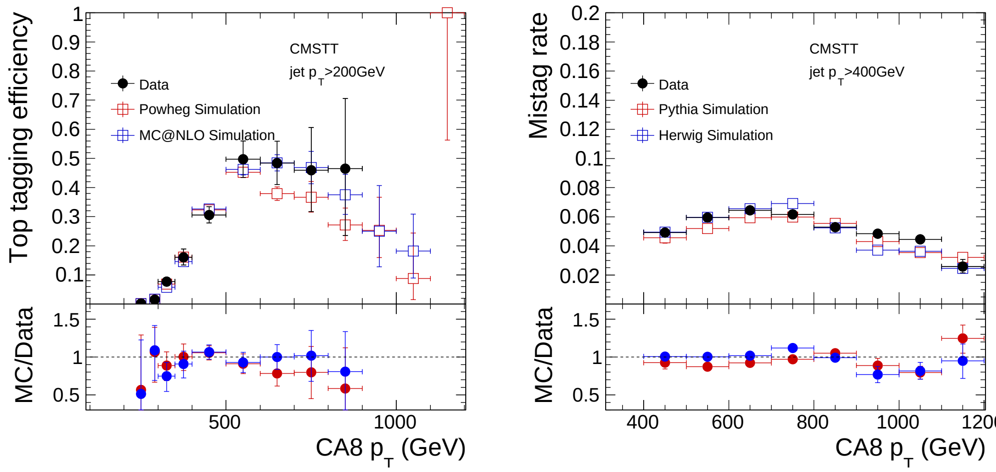


Figure 10.17: Efficiency (left) and mistag rate (right) of the CMS top tagger as function of p_{T}. The error bars show the combined statistical and systematic uncertainties. At the bottom of each panel the ratio of simulation to data is shown.

number of primary vertices, but since pile-up effects can introduce changes in the jet mass distribution, the efficiency is not flat. This is mostly due to the cut on the ungroomed jet mass used for the CMS top tagger. Since the upper cut on the jet mass window is rather high with a value of 250 GeV, signal jets are not as much affected as background jets. A slope towards higher values of the mistag rate with an increasing number of primary vertices is visible in data and simulation. The slope in data is well reproduced by the simulation.

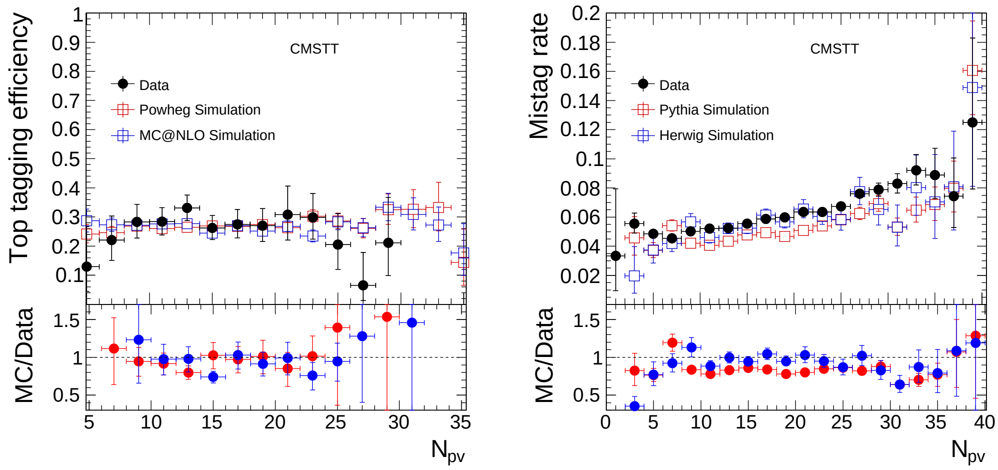


Figure 10.18: Efficiency (left) and mistag rate (right) of the CMS top tagger as function of the number of reconstructed primary vertices. The error bars show the combined statistical and systematic uncertainties. At the bottom of each panel the ratio of simulation to data is shown.

10.4.2 HEP top tagger

Most of the decision variables of the HEP top tagger are mass variables (see Section 7.2). In Figure 10.19 the invariant mass of the three filtered subjets is shown. The mass distribution is shown after applying the selection on the mass of 140 GeV $< m_{123} <$ 250 GeV. A clear top quark resonance in the signal selection can be observed. As discussed earlier, a peak around the top quark mass is introduced in the distribution of background jets, due to the ordering of subjets by the HEP top tagging algorithm. The agreement between data and MC is good for both selections and the different simulations.

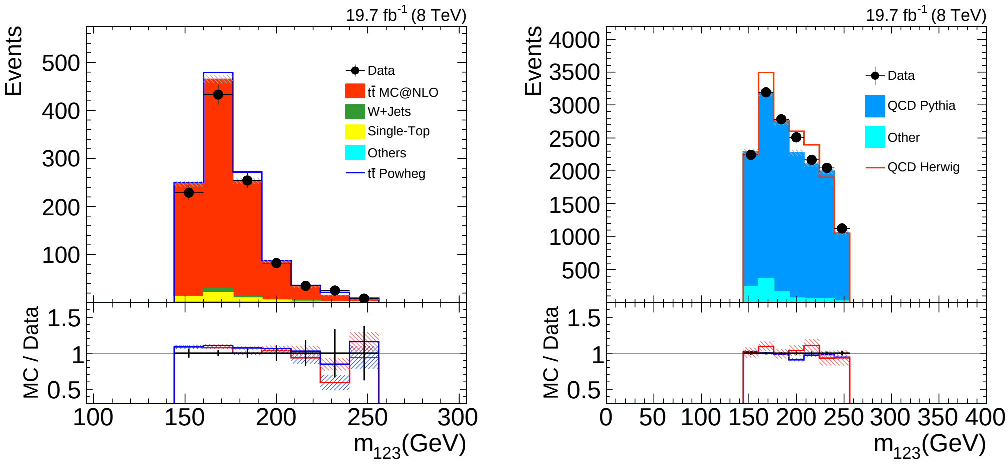


Figure 10.19: Filtered jet mass m_{123} for the signal selection (left), the background selection (right). Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

In Figures 10.20 - 10.22 the invariant mass of two combined subjets is shown for all possible subjet combinations. The simulation is able to describe the distributions observed in data. The W boson mass peak is visible for the signal selection in all three distributions.

The HEP top tagger aims at identifying top quarks with relatively low p_T (compare Sections 7.2 and 9). It has been designed for the use at the threshold to resolved analyses, where the decay products of the top quark are merged in a large jet with a distance parameter of 1.5. The efficiency of the HEP top tagger is shown in Figure 10.23. It is high in a p_T region below 600 GeV. The algorithm shows a non-zero efficiency already at 200 GeV and the plateau is reached around 300 GeV. The efficiency of the HEP top tagger decreases with increasing p_T after values of about 600 GeV, similarly as the efficiency of the CMS top tagger. The integrated efficiency for $p_T >$ 350 GeV is 28%. The mistag rate measured for the standard working point is below 5% and stays flat with increasing p_T. It is visible that the mistag rate in PYTHIA is smaller than the one in data by 0.2-0.5% (absolute), while HERWIG++ provides a good description. For HERWIG++ the integrated mistag rate for $p_T >$ 350 GeV is 2.9%, for PYTHIA it is 2.6%.

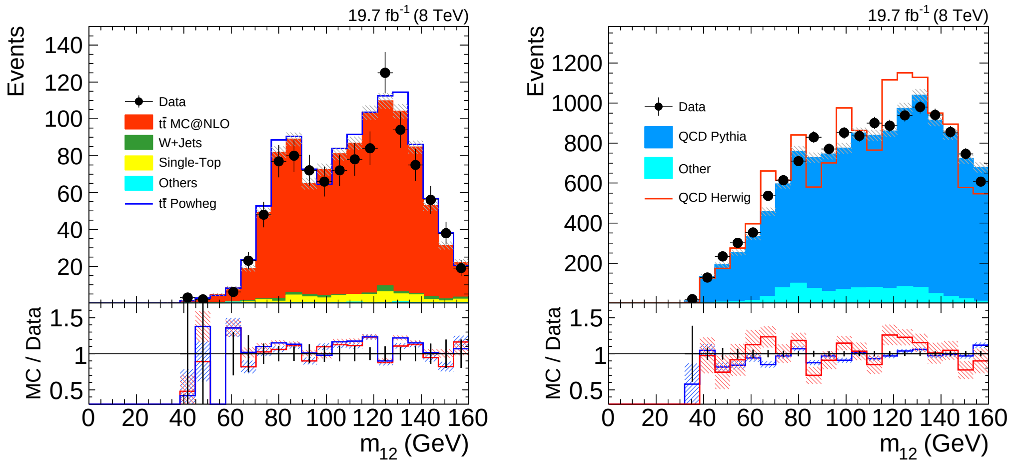


Figure 10.20: Invariant mass of subjet one and two m_{12} for the signal selection (left) and background selection (right). Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

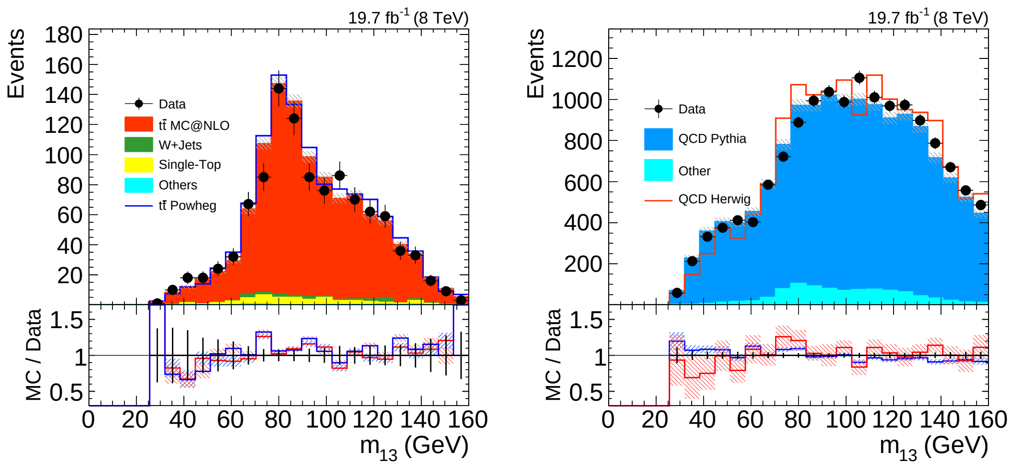


Figure 10.21: Invariant mass of subjet one and three m_{13} for the signal selection (left) and background selection (right). Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

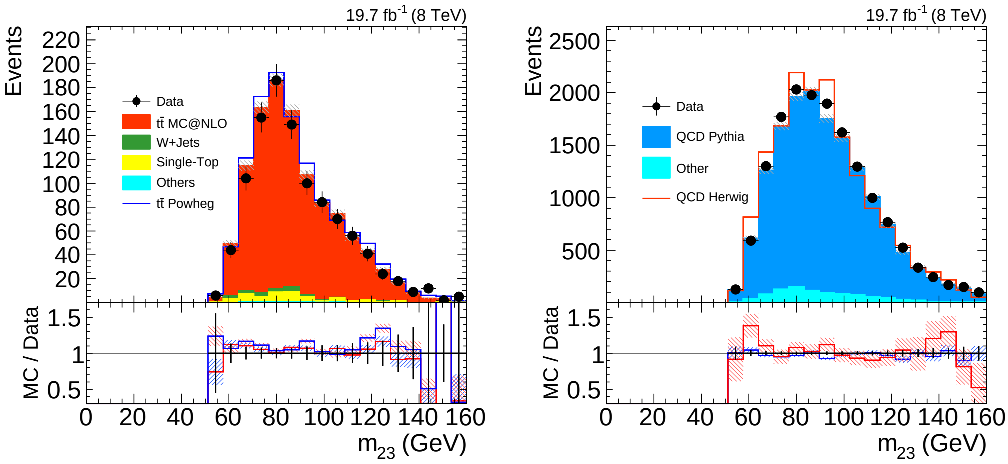


Figure 10.22: Invariant mass of subjet two and three m_{23} for the signal selection (left) and background selection (right). Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

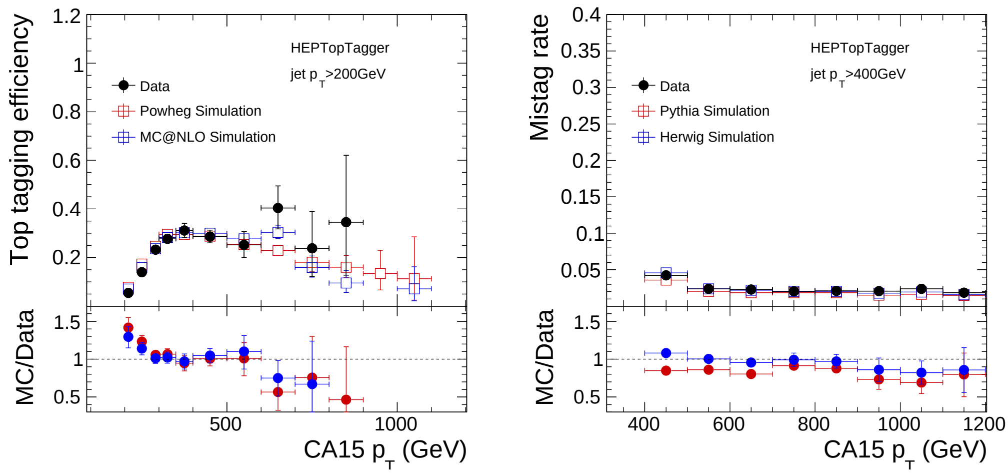


Figure 10.23: Top tagging efficiency (left) and the mistag rate (right) for the HEP top tagger as function of p_{T} . The error bars show the combined statistical and systematic uncertainties. At the bottom of each panel the ratio of simulation to data is shown.

The efficiency and mistag rate as function of the reconstructed number of primary vertices N_{pv} is shown in Figure 10.24. The efficiency is flat as function of N_{pv} for signal jets. In background events, the slope is less steep than the one observed for the CMS top tagger, which indicates that the algorithm is more stable with respect to pile-up. It is well described by the simulations.

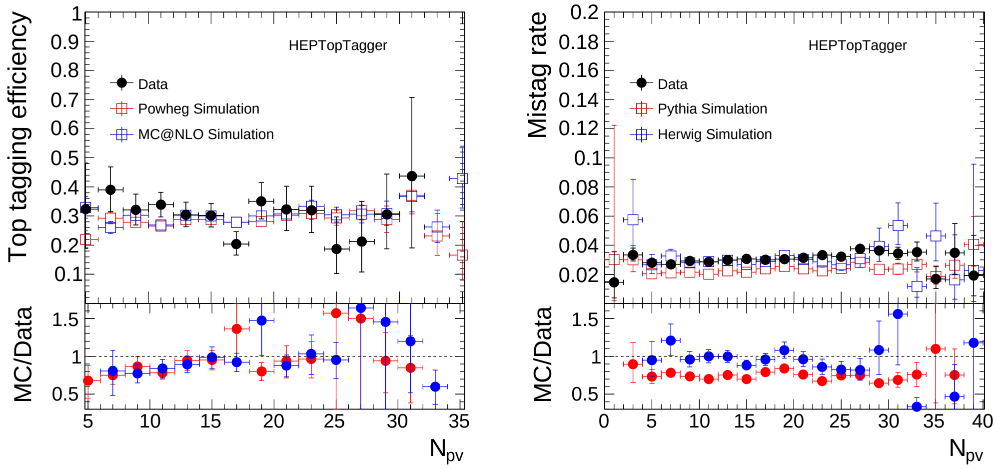


Figure 10.24: Top tagging efficiency (left) and mistag rate (right) for the HEP top tagger as function of the number of primary vertices N_{pv}. The error bars show the combined statistical and systematic uncertainties. At the bottom of each panel the ratio of simulation to data is shown.

10.4.3 OptimalR HEP top tagger

As the OptimalR HEP top tagger (HTT V2) is an improvement of the original HEP top tagger, this section focuses on the new variables introduced. The optimal radius R_{opt} as found by the algorithm is shown in Figure 10.25. In the signal selection the distribution exhibits a peak around 1.0. This indicates that without an additional cut on p_{T} of the jet, in most cases the jets have an optimal radius of 1.0. With increasing jet p_{T}, the peak shifts to lower values. This behavior is due to the decay products being more strongly collimated with increasing p_{T}. In the background selection no resonance peak is visible. Overall, the agreement between data and simulation is good for both selections.

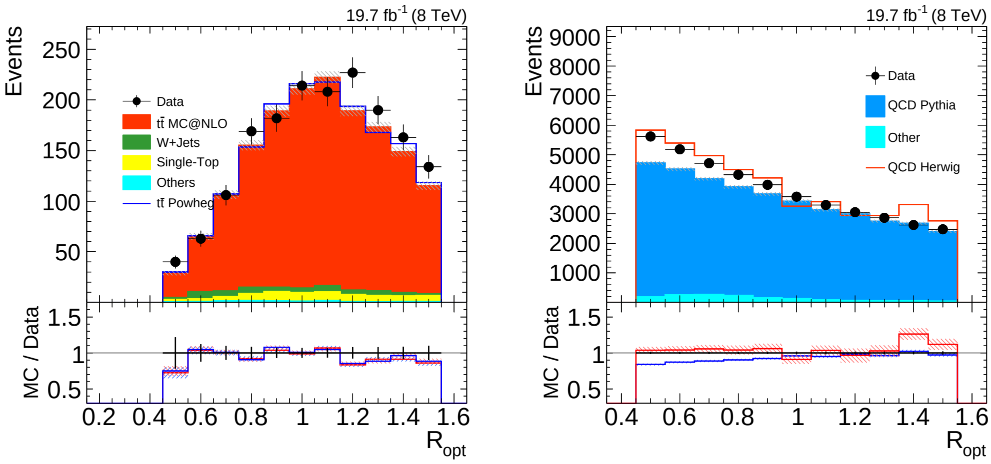


Figure 10.25: Minimum radius R_{min} of the OptimalR HEP top tagger for signal (left) and background (right) events. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

In Figure 10.26 the p_{T} distribution is shown for jets with the optimal radius. Some mismodeling is observed at small values of p_{T} in the signal and background regions. In the background selection, PYTHIA describes the region of high p_{T} well, while HERWIG++ does not provide a good description over the full range in p_{T}.

The mass m_{123} at the optimal radius R_{opt} is shown in Figure 10.27. For background events a smoothly falling distribution is observed. While the signal region is well described by the POWHEG $\mathrm{t\bar{t}}$ simulation and the distribution for background jets is well modeled by HERWIG++, PYTHIA does not provide a good description of the data for background jets.

The new decision variable $R_{\mathrm{opt}} - R_{\mathrm{opt}}^{\mathrm{calc}}$ is shown in Figure 10.28. This variable is well described by all Monte Carlo event generators and behaves as expected (see Section 9).

The efficiency and mistag rate as function of the jet p_{T} are shown in Figure 10.29. Here a tight working point is chosen, with the selection summarized in Table 10.6. The efficiency and mistag rate for other working points are shown in Appendix A.

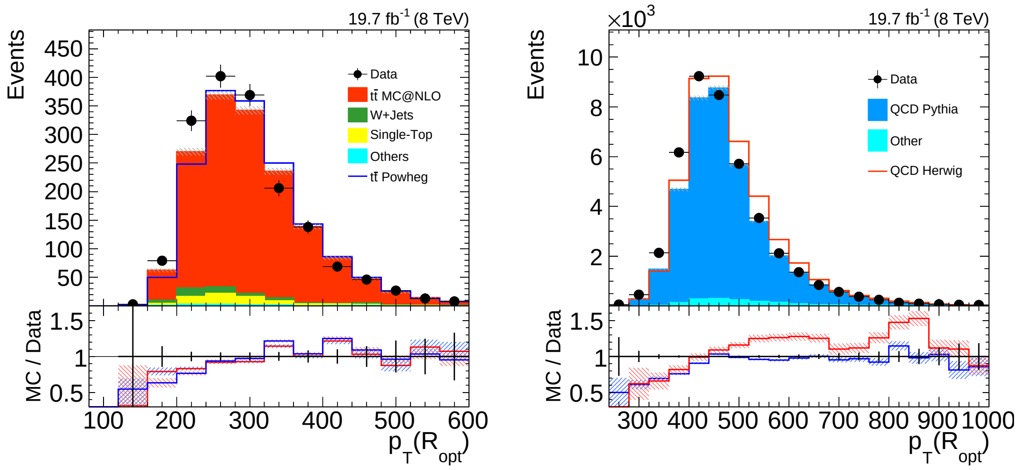


Figure 10.26: Distribution of p_{T} at minimal radius R_{min}. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

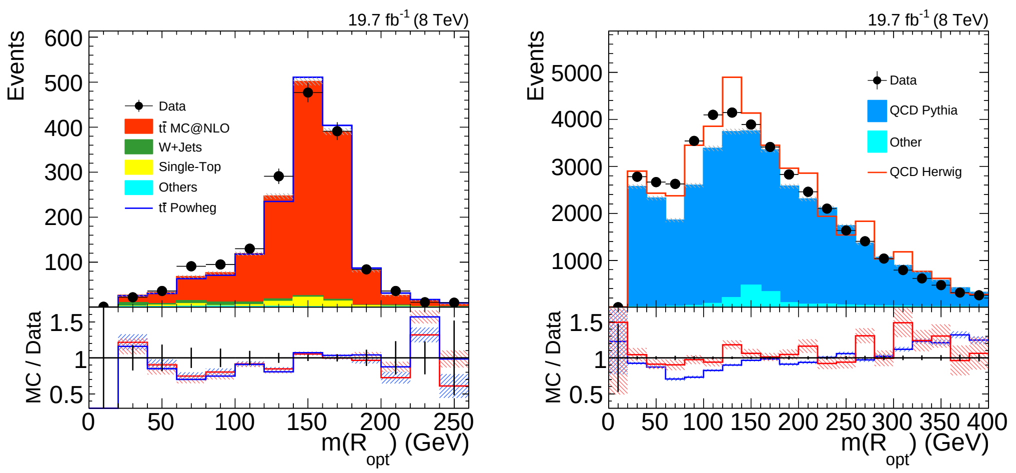


Figure 10.27: Mass at minimum radius R_{min}. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

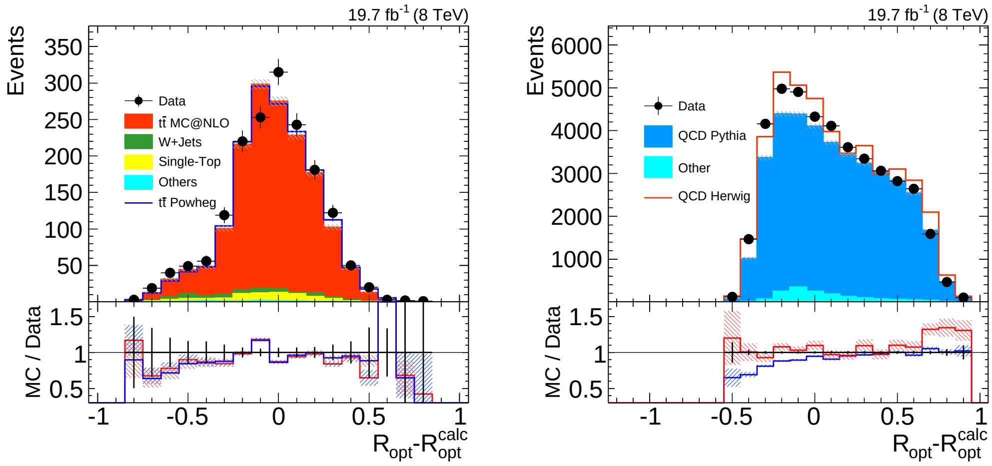


Figure 10.28: Decision variable $R_{\min} - R_{\min,\exp}$. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

Decision variable	selection
$m_{123}(R_{\min})$	$140\,\mathrm{GeV} < \mathrm{m}_{123}(\mathrm{R}_{\min}) < 250\,\mathrm{GeV}$
$f_W(R_{\min})$	0.15
$R_{\min} - R_{\min,\exp}$	> 0

Table 10.6: Selection criteria for the OptimalR HEP top tagger.

Similar to the taggers discussed above, the signal efficiency is well described by the simulation. The integrated efficiency for $p_{\mathrm{T}} > 350\,\mathrm{GeV}$ is 28%. The measured mistag rate is about 2% for $p_{\mathrm{T}} > 500\,\mathrm{GeV}$ and 4% for lower values of p_{T}. The average mistag rate for $p_{\mathrm{T}} > 350\,\mathrm{GeV}$ is 2.8% for HERWIG++ and 2.6% for PYTHIA. While HERWIG++ describes the mistag rate measurement well, the PYTHIA simulation underestimates the mistag rate by about 20%, which is similar to the results for the other top taggers.

The measurement of the efficiency and mistag rate as function of the number of primary vertices given in Figure 10.30, shows that the OptimalR HEP top tagger is fairly independent of pile-up. This behavior is well described by simulation.

The efficiency and mistag rate as a function of the pseudorapidity η are shown in Figure 10.31. The efficiency as well as the mistag rate are decreasing with increasing η. This effect is caused by the worse jet reconstruction to higher η, since the tracker covers a pseudorapidity range up to 2.5. The efficiency and mistag rate is approximately flat inside the barrel tracker region ($\eta < 1.3$), for higher η the Particle Flow algorithm does not reconstruct all jet constituents anymore. The agreement between data and MC is fair for the efficiency and mistag rate measurement.

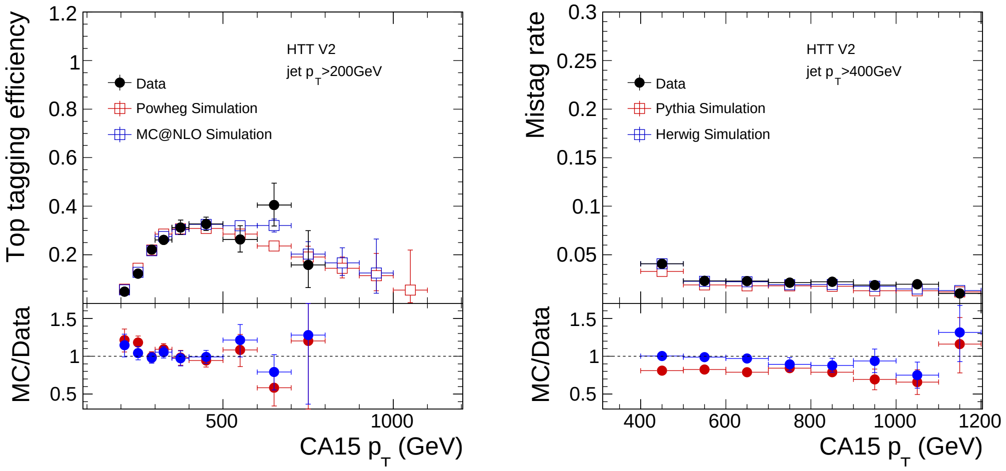


Figure 10.29: Measurement of the top tagging efficiency (left) and the mistag rate (right) for the OptimalR HEP top tagger as function of p_{T}. The error bars show the combined statistical and systematic uncertainties. At the bottom of each panel the ratio of simulation to data is shown.

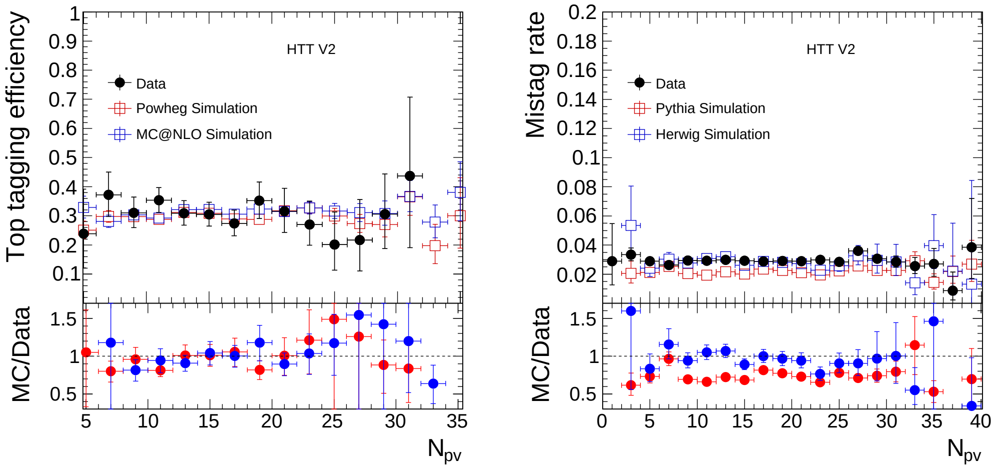


Figure 10.30: Measurement of the top tagging efficiency (left) and the mistag rate (right) for the OptimalR HEP top tagger as function of N_{pv}. The error bars show the combined statistical and systematic uncertainties. At the bottom of each panel the ratio of simulation to data is shown.

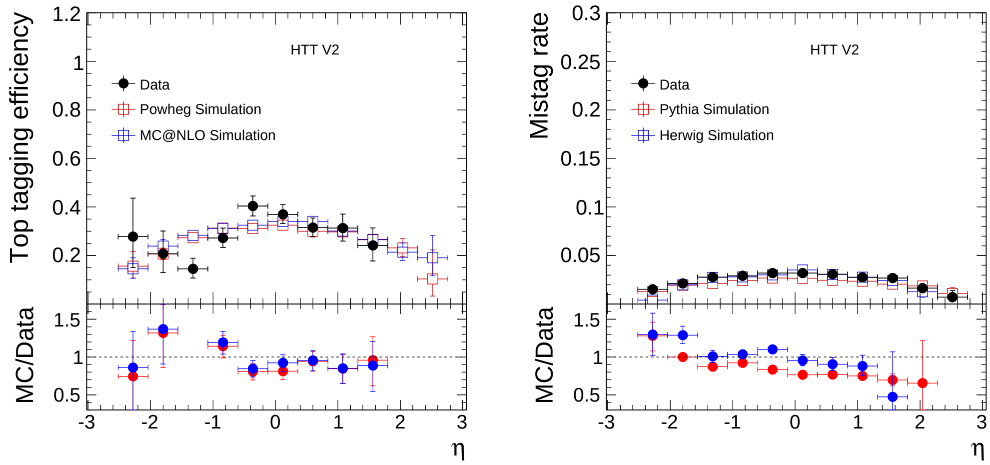


Figure 10.31: Measurement of the top tagging efficiency (left) and the mistag rate (right) for the OptimalR HEP top tagger as function of η. The error bars show the combined statistical and systematic uncertainties. At the bottom of each panel the ratio of simulation to data is shown.

10.4.4 Shower deconstruction algorithm

The shower deconstruction algorithm has just one decision variable χ. Since microjets are used as an input for the shower deconstruction tagger, the agreement between data and simulation for the kinematic properties of the microjets are investigated first. Here the focus is on the CA15 selection and consequently all following distributions are given for this selection.

Figure 10.32 shows the number of microjets as found by clustering the particle flow candidates again with the k_T algorithm. Good agreement is observed for the signal selection. It is observed that limiting the maximum number of microjets to nine, in order to reduce computational time, does not lead to a loss of signal since almost no large jets contain more than nine microjets. Neither PYTHIA nor HERWIG++ provide a good description of the microjet multiplicity for background jets. One has to note that there are no JECs applied to the microjets. The p_T distribution of the leading microjet is shown in Figure 10.33. Because of the selection criteria applied to obtain the signal and background regions (see Sections 10.1 and 10.2), the microjet p_T is larger on average in the background region than in the signal region. The simulations provide a good overall description of the distribution measured in data. A slight mismodeling is observed for microjets with low p_T. In Appendix A the p_T distributions of the other microjets are shown.

The invariant mass of all microjets is shown in Figure 10.34. The agreement between data and simulation is good for both selections. Only the very low mass region for background jets is not well modeled by either simulation program. If the maximum number of microjets is not set and no p_T cut (here $p_T > 10$ GeV) on the microjets is used, these distributions would be the same as the jet mass derived with all particle flow candidates of the entire CA15 jet.

The decision variable χ for jets with $p_T > 350$ GeV is shown in Figure 10.35. This variable is calculated using the microjets as input. The distribution is well modeled in the

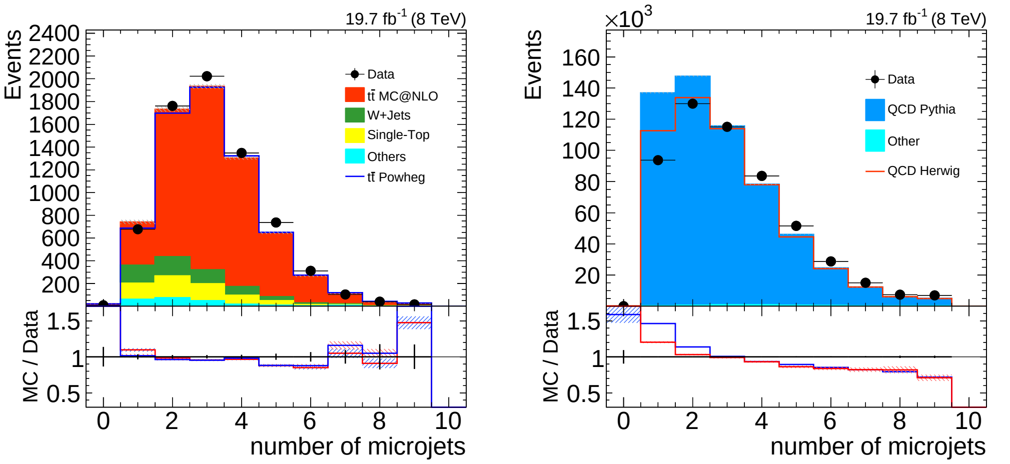


Figure 10.32: Number of microjets found by the k_T-algorithm for signal (left) and background jets (right). Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

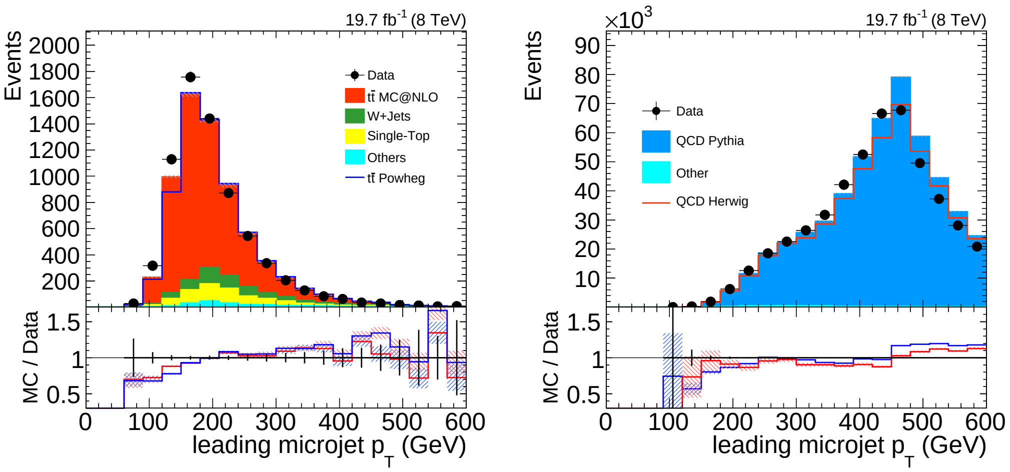


Figure 10.33: Distribution of p_T for the leading microjet for signal (left) and background events (right). Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

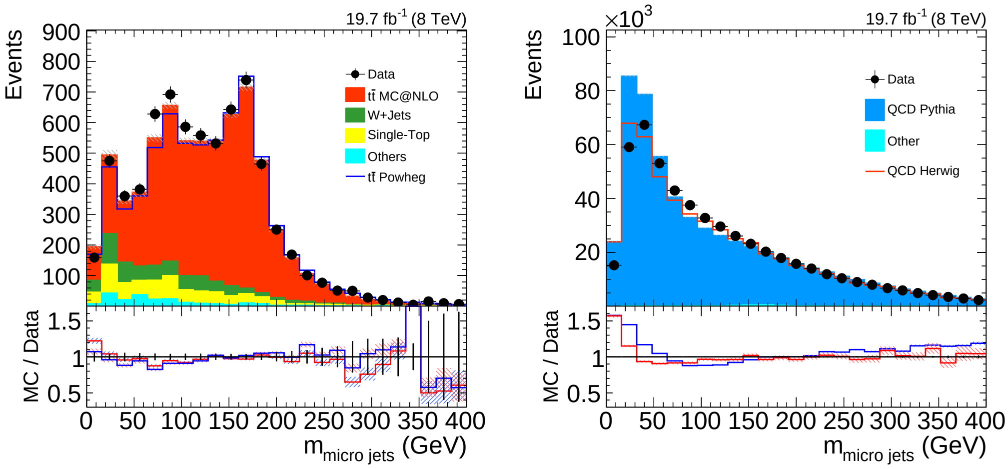


Figure 10.34: Invariant mass of all microjets for signal (left) and background events (right). Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

signal region, also confirming the good description of microjets in $t\bar{t}$ production. In the background region, some discrepancy between data and the simulation is observed. This can be traced back to the mismodeling of the microjet multiplicity and p_T shown above. However, the overall description of the χ distribution for background jets is fair.

The distribution of the decision variable for jets with a higher p_T requirement of $p_T >$ 600 GeV is shown in Figure 10.36. While the signal region is well described within the limited statistics available, a larger discrepancy between data and simulation than at low p_T is observed in the background region.

The measurement of the efficiency and the mistag rate for a specific working point of $\log(\chi) > 3.5$ is shown in Figure 10.37. Other working points can be found in Appendix A. The efficiency increases up to 45% at values of p_T of around 500 GeV. The shower deconstruction tagger is not very efficient in the low p_T region, and can therefore not be used in analysis kinematically close to a resolved decay of the top quark. Similar as for other tagging algorithms, the efficiency decreases with increasing transverse momenta. The integrated efficiency for $p_T > 350$ GeV is 23.1%. Data and simulation are in good agreement for the efficiency. The mistag rate is almost constant below 5%. It is better described by the HERWIG++ simulation.

Figure 10.38 shows the efficiency as function of the number of primary vertices. The agreement between data and simulation is good for the signal selection and no dependence on N_{pv} is observed. Also the mistag rate shows no significant dependence on the number of primary vertices. Again, the agreement between data and simulation is better for the HERWIG++ than for PYTHIA simulation. In Figure 10.39 is the efficiency and the mistag rate as function of the pseudorapidity η shown. Similar as for the OptimalR HEP top tagger the efficiency and mistag rate are decreasing with increasing η. The agreement

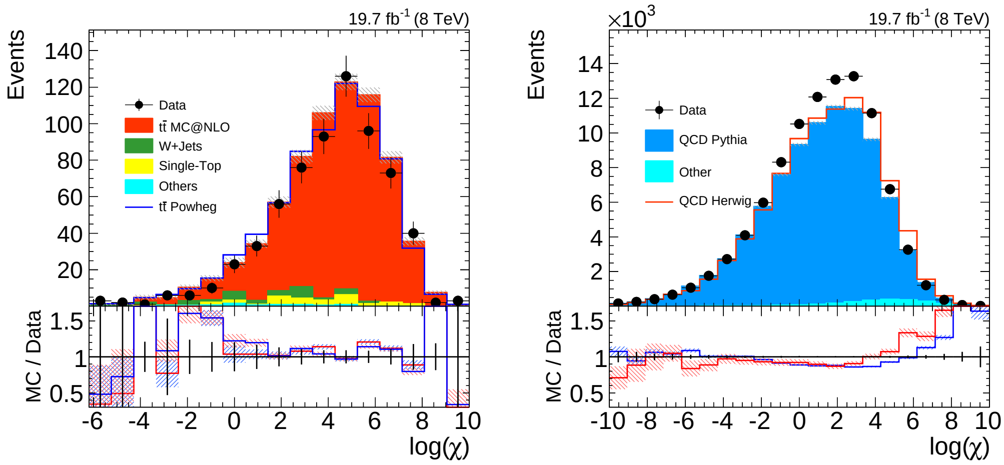


Figure 10.35: Decision variable χ for a jet $p_{\mathrm{T}} > 350\,\mathrm{GeV}$ for signal (left) and background events (right). Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

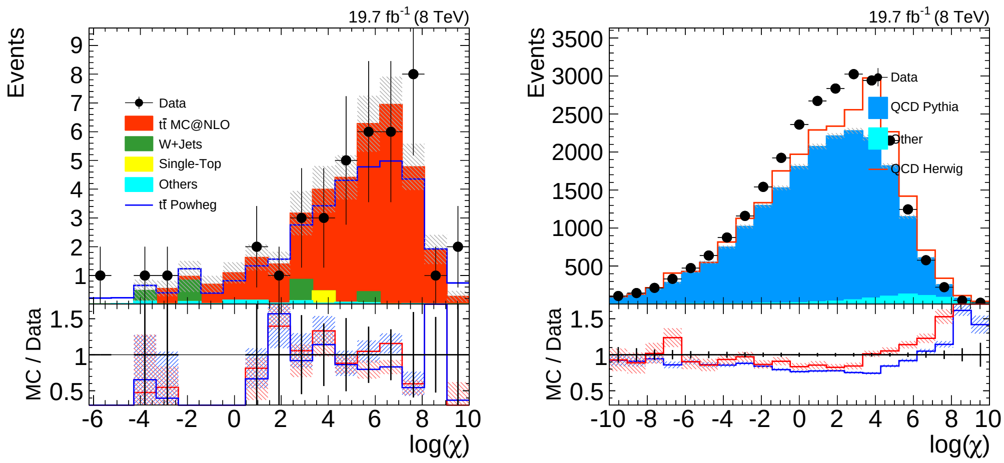


Figure 10.36: Decision variable χ for a jet $p_{\mathrm{T}} > 600\,\mathrm{GeV}$ for signal (left) and background events (right) . Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

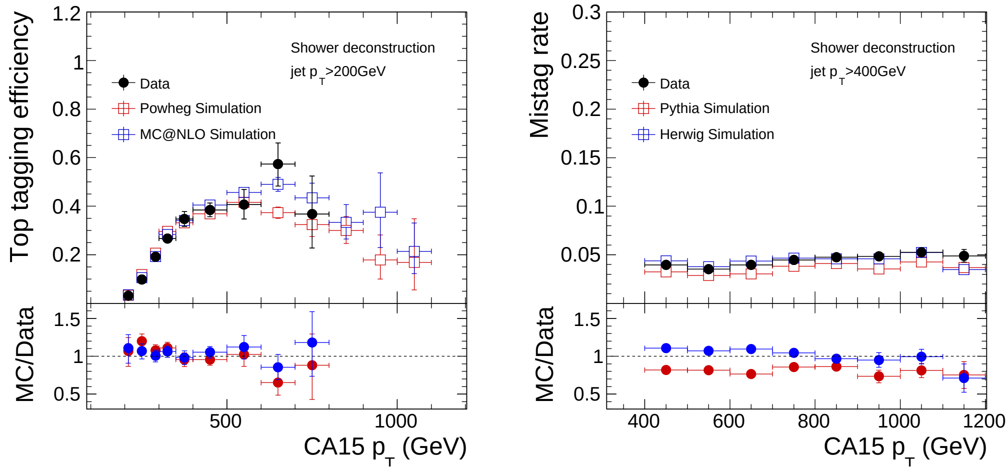


Figure 10.37: Top tagging efficiency (left) and mistag rate (right) for the shower deconstruction tagger at WP1 as function of p_{T}. The error bars show the combined statistical and systematic uncertainties. At the bottom of each panel the ratio of simulation to data is shown.

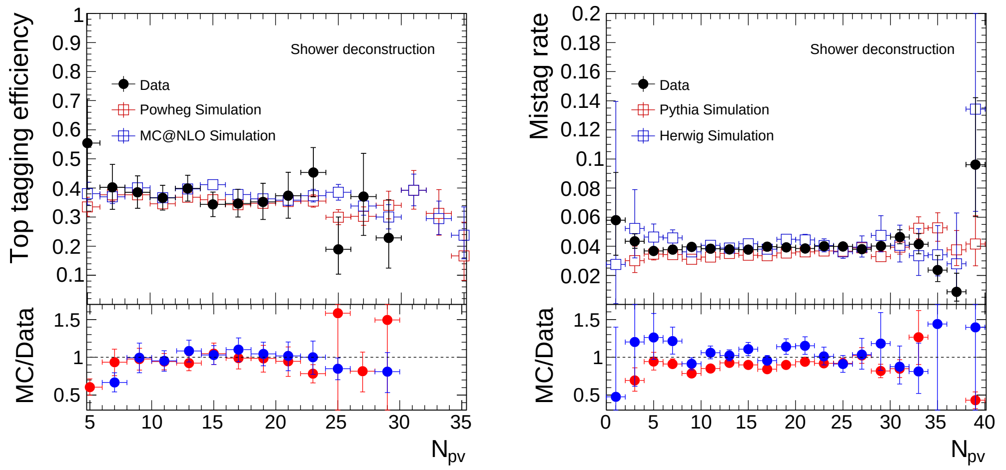


Figure 10.38: Top tagging efficiency (left) and mistag rate (right) for the shower deconstruction tagger at WP1 as function of N_{pv}. The error bars show the combined statistical and systematic uncertainties. At the bottom of each panel the ratio of simulation to data is shown.

between data and MC is reasonable for both selections.

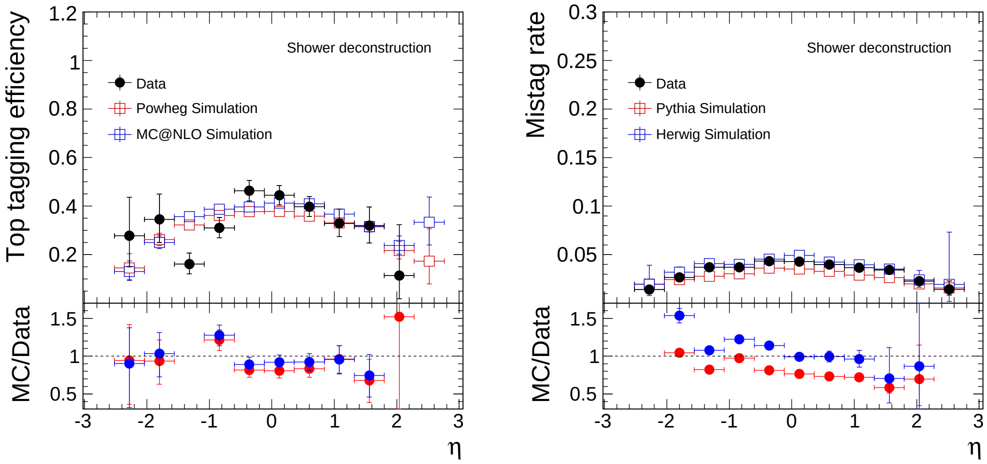


Figure 10.39: Top tagging efficiency (left) and mistag rate (right) for the shower deconstruction tagger at WP1 as function of η. The error bars show the combined statistical and systematic uncertainties. At the bottom of each panel the ratio of simulation to data is shown.

10.4.5 Soft drop

Recent studies [196] show that a very promising tool for improving the performance of top tagging algorithms is soft drop. In this section we explore how well the soft drop mass is modeled and thus if it is feasible to use it in future analysis.

The following parameters for soft drop are used (see Section 7.5.4):

$$z_{\text{cut}} = 0.2 \tag{10.13}$$

$$\beta = 1.0. \tag{10.14}$$

The selection is based on CA15 jets. The p_{T} distribution of the jets is shown in Figure 10.40. It changes after applying soft drop, shifting to lower values. In the signal and background regions, the agreement between data and simulation gets slightly worse (compare with Figures 10.6, 10.12 and 10.13), where the shift in data is larger than in simulation.

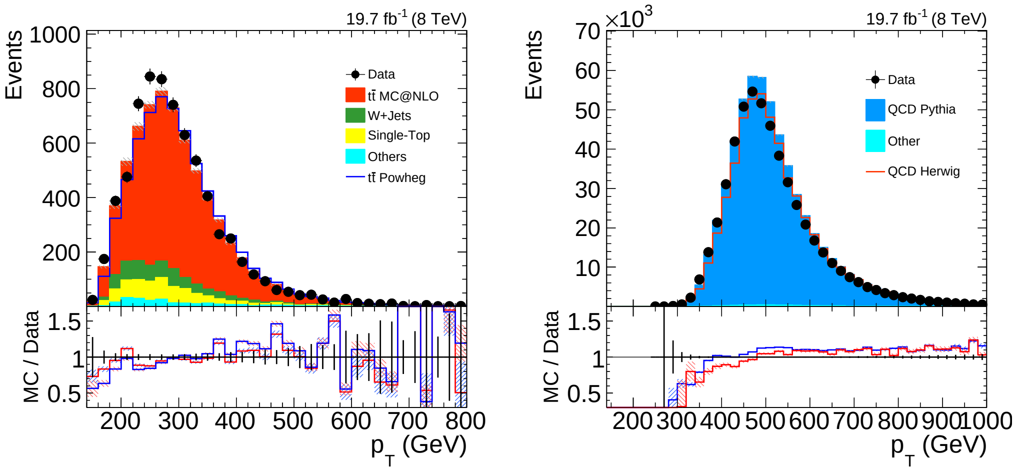


Figure 10.40: Distribution of the leading jet p_{T} after applying soft drop. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

In Figure 10.41 the mass of the jet after applying soft drop is shown. The jet mass is the decision variable of the soft drop algorithm. A clear resonance peak at the top quark mass is visible for the signal selection. In background events, the mass gets shifted to smaller values and good separation power between signal and background can be seen. The soft drop mass distribution is described well by HERWIG++, but a fair description with PYTHIA is obtained as well.

The efficiency and mistag rate using only the jet mass after applying soft drop as discriminating variable and once using the ungroomed jet mass, are shown in Figure 10.42. For both variables the mass selection criterion is $140\,\text{GeV} < m < 250\,\text{GeV}$, in order to

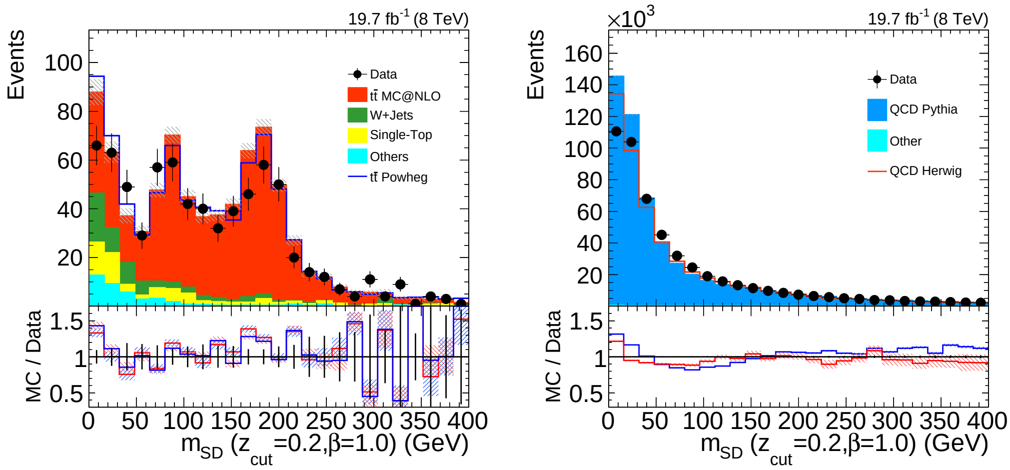


Figure 10.41: Jet mass after applying soft drop for a jet $p_{\mathrm{T}} > 400\,\mathrm{GeV}$. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

make the two variables comparable. For the ungroomed jet mass the efficiency increases up to 70% and the corresponding mistag rate increases up to 35%. The integrated efficiency for $p_{\mathrm{T}} > 350\,\mathrm{GeV}$ is 61% and the mistag rate is for PYTHIA 25% and for HERWIG++ 27%. In comparison, the efficiency using the soft drop mass is lower, similarly to the mistag rate. The integrated efficiency for soft drop and $p_{\mathrm{T}} > 350\,\mathrm{GeV}$ is 39% and the mistag rate is 9.7% for PYTHIA and 10% for HERWIG++. The efficiency decreased after applying soft drop by 40% and the mistag rate by 60% at the given mass selection. Consequently, the signal to background ratio improves after applying soft drop.
The efficiency and mistag rate of the ungroomed mass are well described by the simulation. The efficiency for soft drop is well described, whereas the mistag rate is under estimated by PYTHIA.

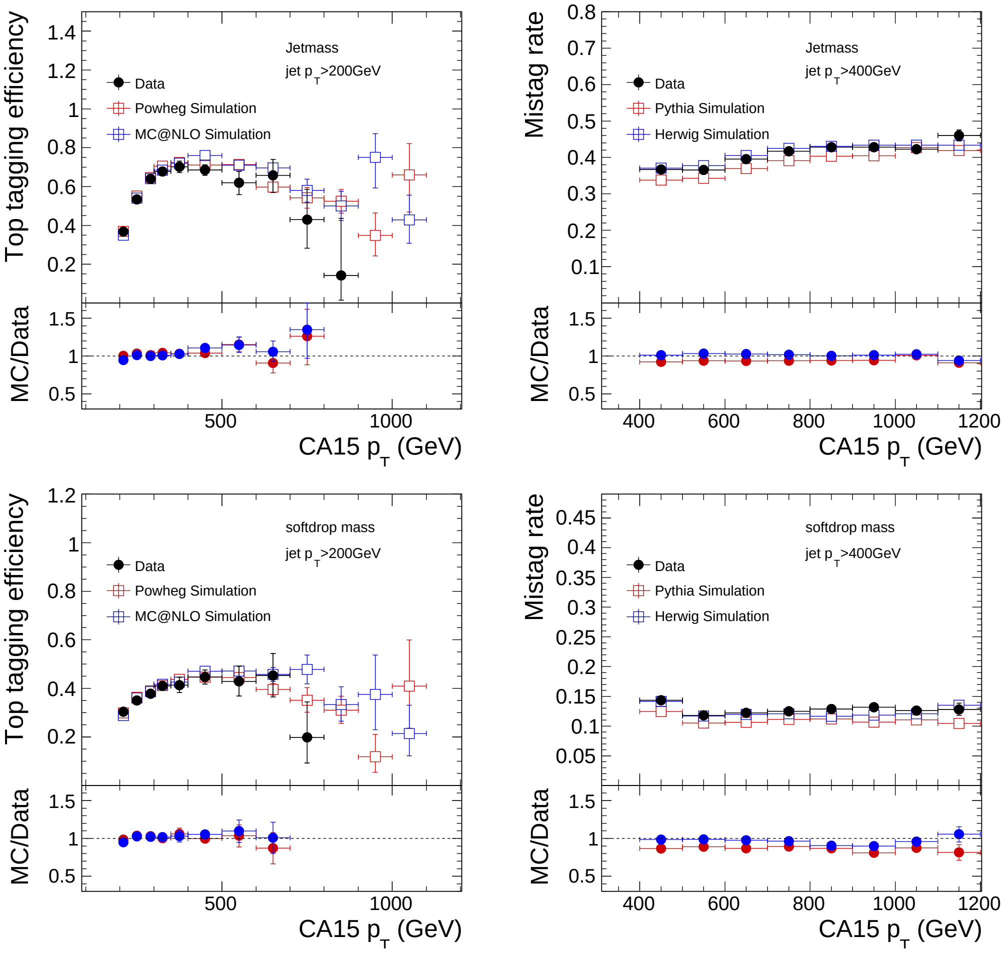


Figure 10.42: Top tagging efficiency (top left) and mistag rate (top right) for the ungroomed jet mass and top tagging efficiency (bottom left) and mistag rate (bottom right) for the soft drop mass as function of p_{T}. The error bars show the combined statistical and systematic uncertainties. At the bottom of each panel the ratio of simulation to data is shown.

10.4.6 Other substructure variables

In the following several distributions of further jet substructure variables are shown. The variables have been introduced in Section 7.6.

In Figure 10.43 the jet mass distributions after pruning are shown. The description of data is fair by all simulations, but discrepancies are seen. The background region is better described by HERWIG++, while PYTHIA provides a worse description of the data overall. The agreement between POWHEG and MC@NLO with the data is similar for all variables.

The invariant mass of two subjets m_{12} and m_{23} can be seen in Figure 10.44 (top and bottom row). The subjet mass m_{12} peaks for the signal selection at the W boson mass. For background events one can observe a smoothly falling distribution. The agreement between data and simulation for the signal selection is fair for mass values below 150 GeV.

PYTHIA and HERWIG++ are describing the data well in the background selection. The subjet mass m_{23} is well described by the simulation in the signal selection. Background-like events are for low mass values better described by HERWIG++ and for high mass values better by PYTHIA.

The transverse momentum of the leading subjet is shown in Figure 10.44 (bottom row). A fair description of the data is obtained in the signal selection, whereas the background selection is only well described in the low p_{T} range by both event generators.

In Figure 10.45 (top row) the Q-jet volatility is shown. The agreement between data and simulation is good for the signal selection, the background selection is well described by HERWIG++ only.

The jet shape variable Ψ is shown in Figure 10.45 (middle and bottom row). The distributions peak for both selections at a value of one. One can observe a larger tail towards lower values for the signal selection. The jet shape variables are described well for both selections by the simulation.

In Figure 10.46 the helicity angles between two subjets are shown. Background-like events are distributed at low values whereas signal-like events are distributed at slightly higher values. The simulation describes the data well.

The first and second jet moment can be seen in Figure 10.47 (top and middle row). The distributions look similar for the signal and the background selection. Good agreement is observed between data and simulation for signal events, but both event generators PYTHIA and HERWIG++ do not model jets in the background selection well.

In Figure 10.47 (bottom row) and 10.48 (top row) the number of constituents and the number of charged constituents, respectively, are shown. The variables have almost no separation power and have similar shapes for both selections. While signal-like events are described well, discrepancies are seen for background-like events for PYTHIA and HERWIG++. Especially for a large number of constituents the discrepancy becomes larger.

The jet charge and the weighted jet charge can be seen in Figure 10.48 (bottom row) and 10.49 (top row). The jet charge and the weighted jet charge do not provide strong separation power. The variables are well described in the signal selection. Deficiencies in the simulation of background events are observed for both, the PYTHIA and HERWIG++ generators. An improved simulation of the charged and neutral jet constituents may help to improve the simulation of other substructure variables.

One example of such a variable is N-subjettiness, where ratios are shown in Figure 10.49 (middle row) - 10.50 (bottom row). These distributions are shown before and after pruning. While the simulation describes the data for the signal selection well, discrepancies are observed for all distributions for background events. This has been reported in previous analyses [64,201] already and was one of the reasons that data driven background estimates were necessary.

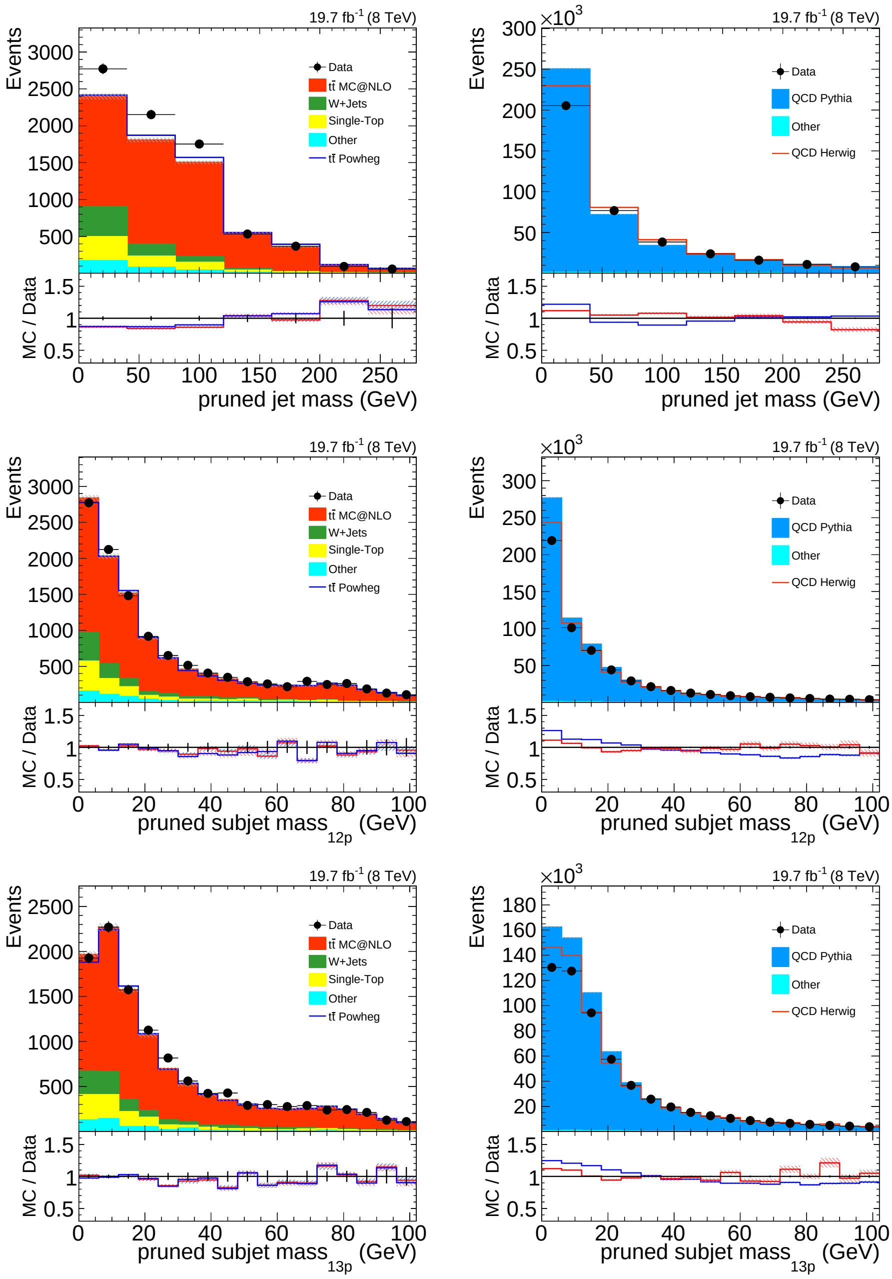


Figure 10.43: Jet mass after pruning (top row), invariant mass of subjet 1 and 2 (middle row) and of subjet 1 and 3 (bootom row) after pruning for signal (left) and background (right) events. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

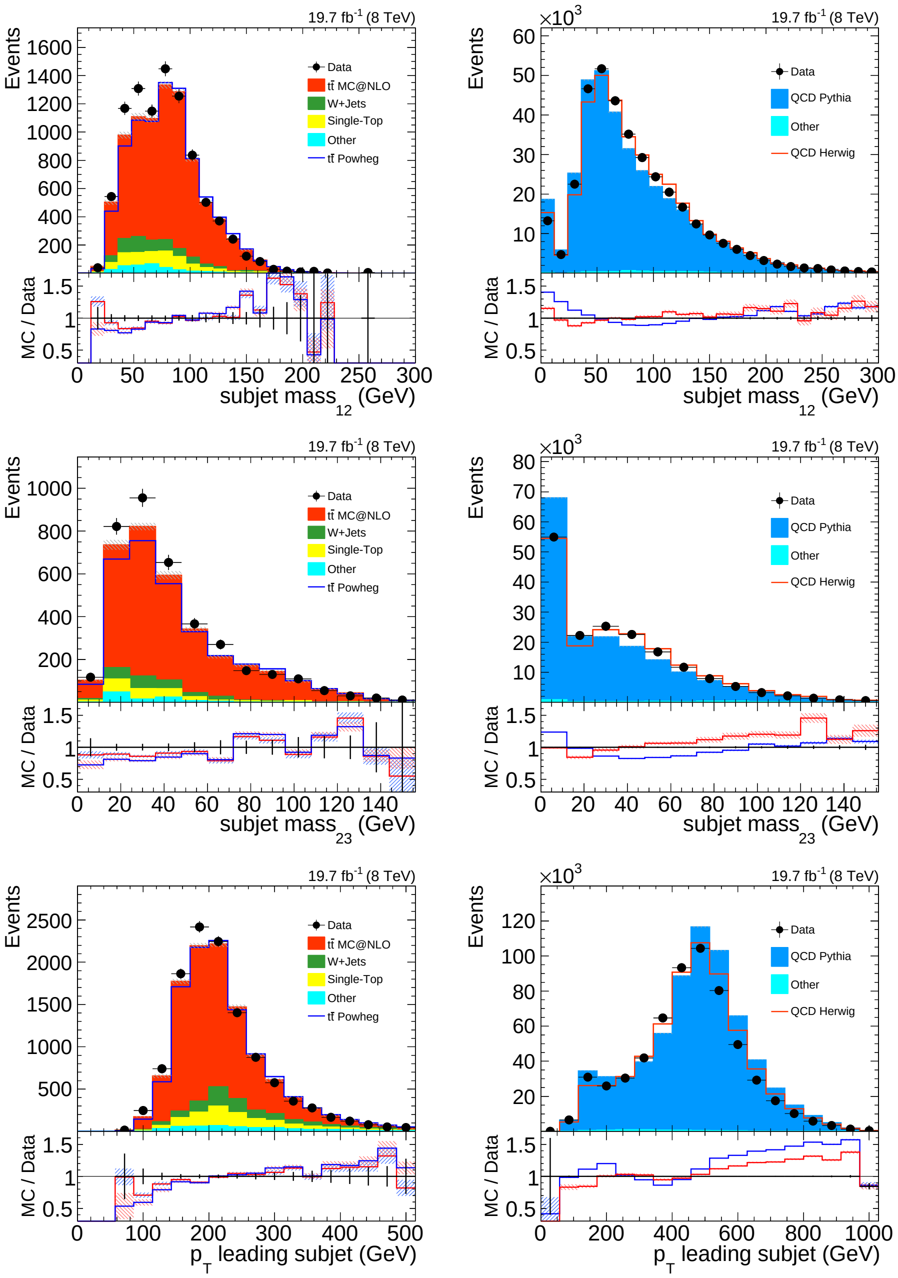


Figure 10.44: Invariant mass of subjet 1 and 2 (top row), invariant mass of subjet 2 and 3 (middle row) and of p_{T} distribution of the leading subjet (bootom row) for signal (left) and background (right) events. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

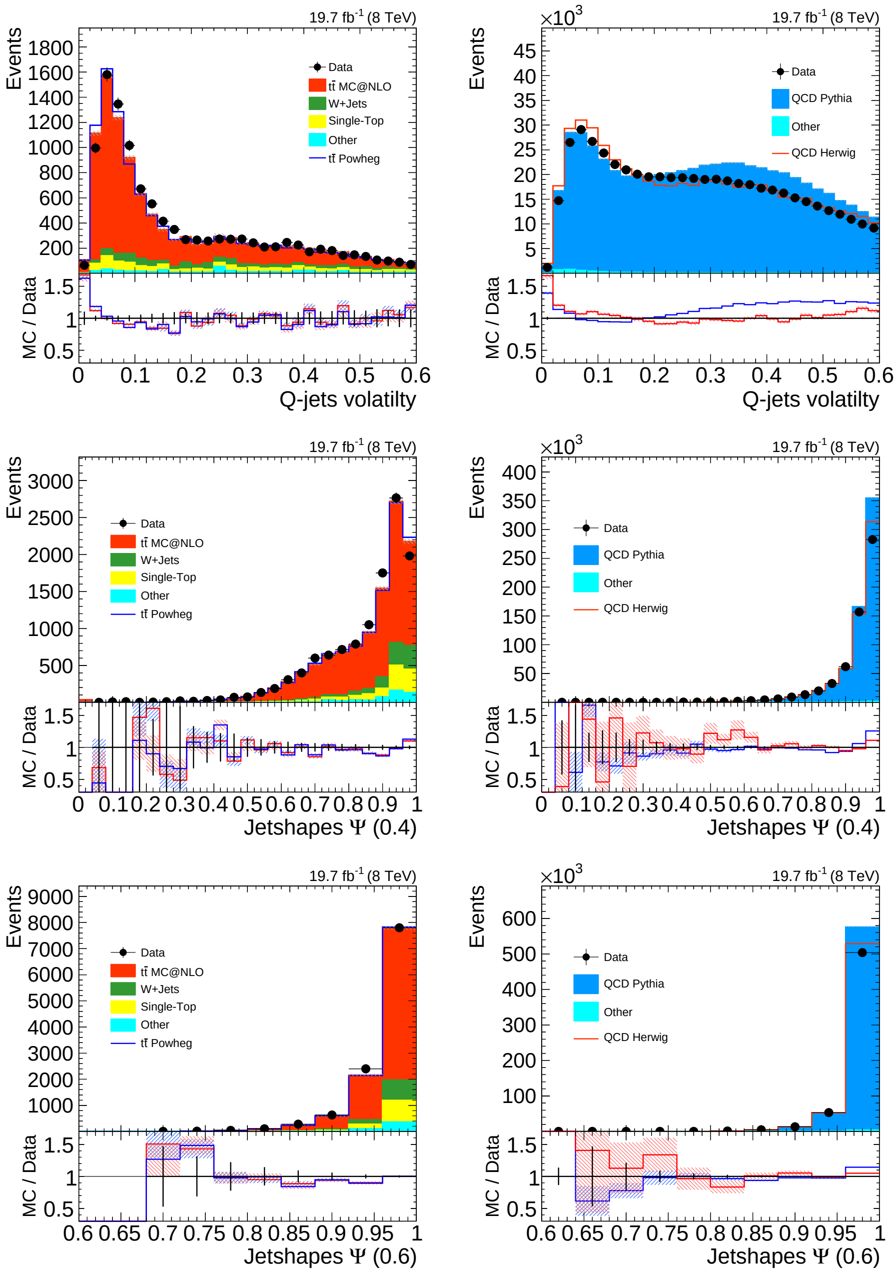


Figure 10.45: Q-jet volatility $\Gamma_{\text{volatility}}$ (top row), jet shape $\Psi(0.4)$ (miidle row) and jet shape $\Psi(0.6)$ (bootom row) for signal (left) and background (right) events. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

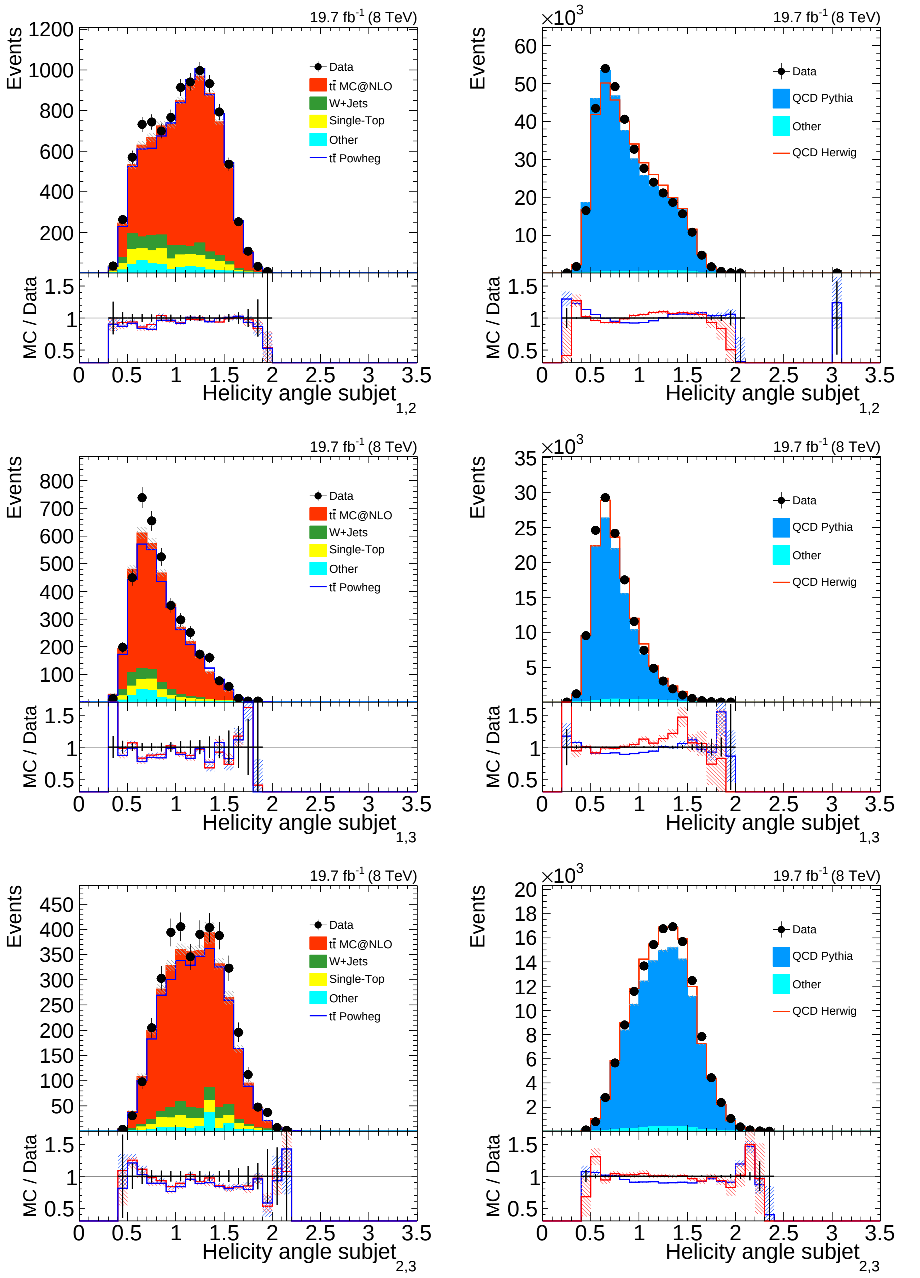


Figure 10.46: Helicity angle θ_{12} (top row), θ_{13} (middle row) and θ_{23} (bottom row) for signal (left) and background (right) events. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

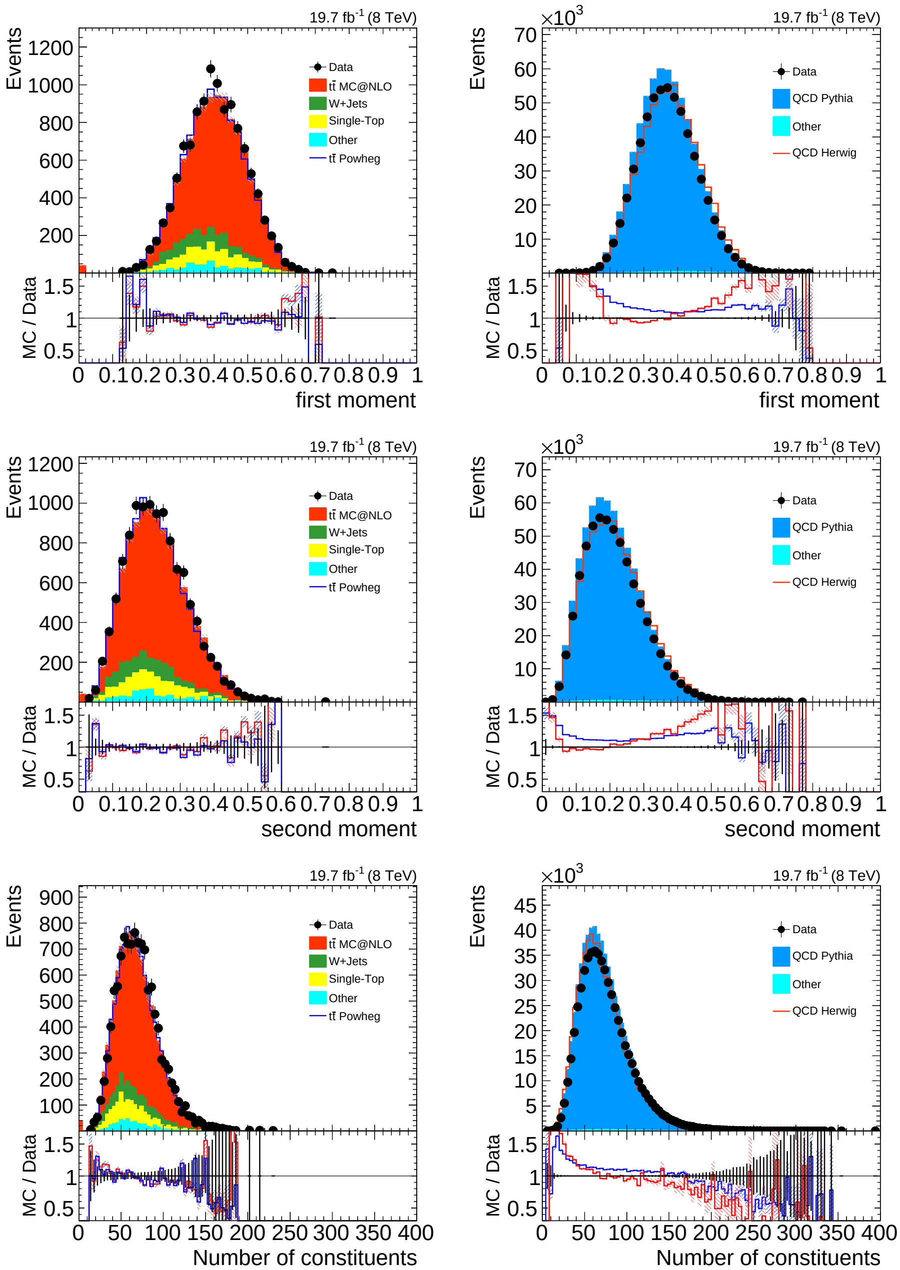


Figure 10.47: First (top row) and second jet moment (middle row) and number of jet constituents (bottom row) for signal (left) and background (right) events. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

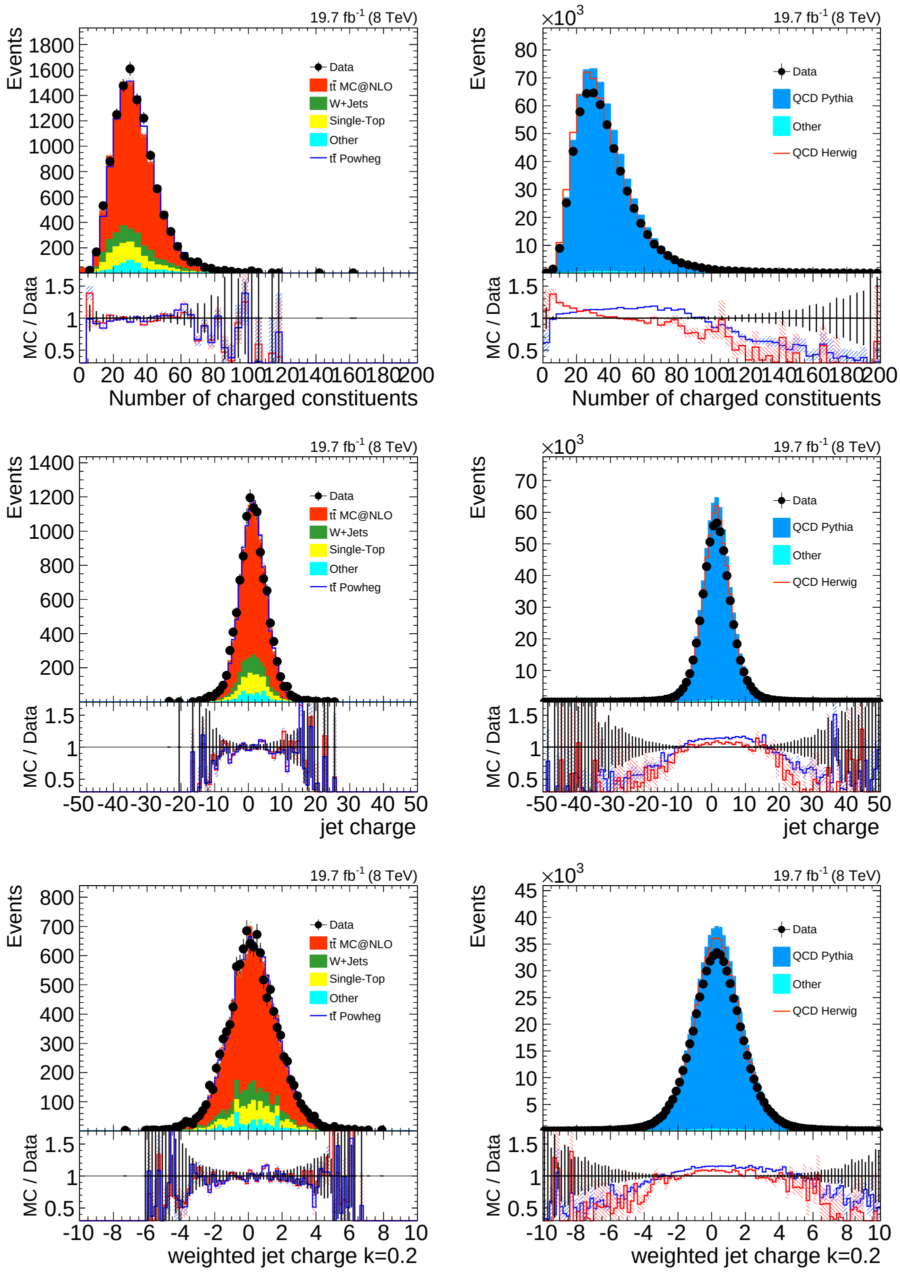


Figure 10.48: Number of charged constituents (top row), jet charge Q (middle row) and weighted jet charge Q_2 (bottom row) for signal (left) and background (right) events. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

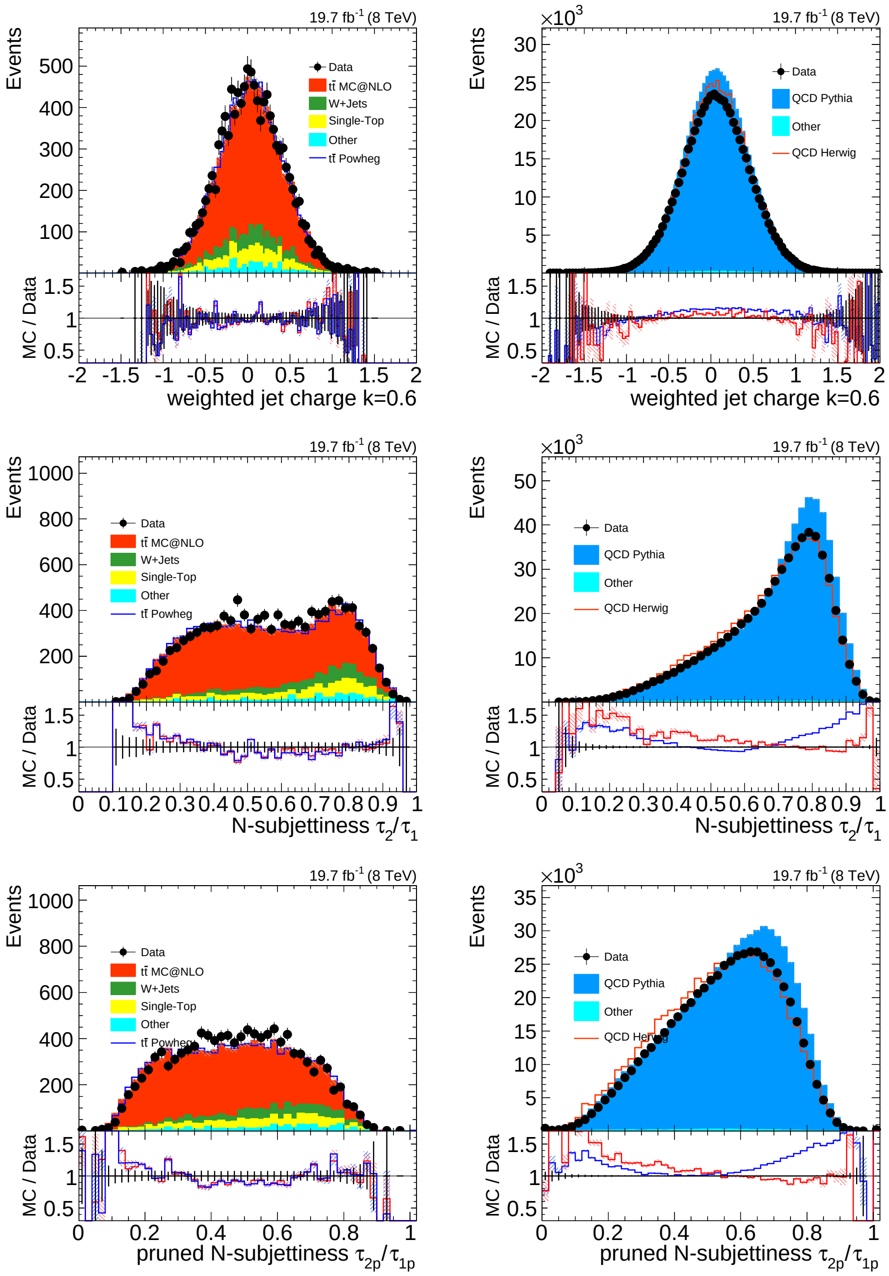


Figure 10.49: Weighted jet charge Q_6 (top row), N-subjettiness ratio τ_2/τ_1 (middle row) and pruned N-subjettiness ratio τ_2/τ_1 (bottom row) for signal (left) and background (right) events. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

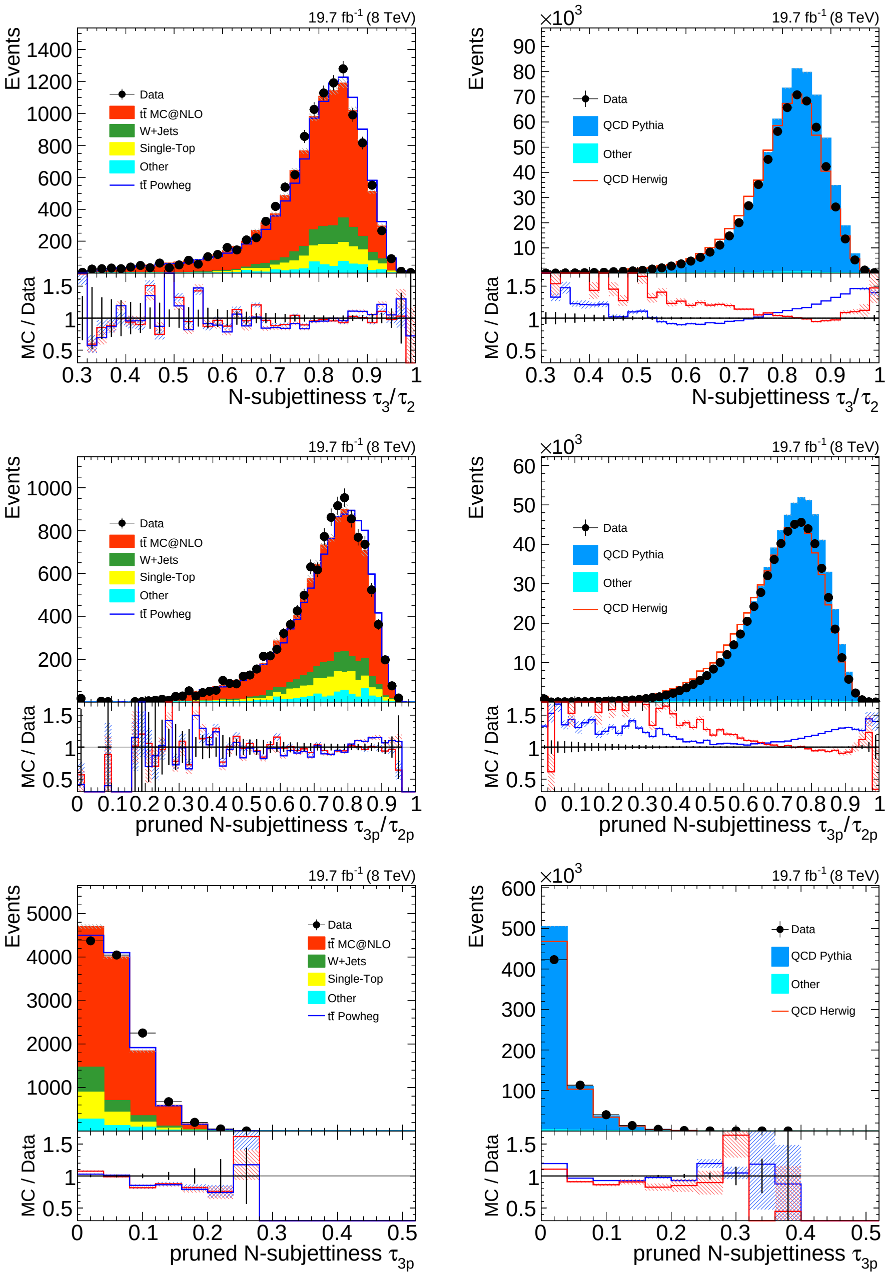


Figure 10.50: N-subjettiness ratio τ_3/τ_2 (top row), pruned N-subjettiness ratio τ_3/τ_2 (middle row) and pruned N-subjettiness τ_3 for signal (left) and background (right) events. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

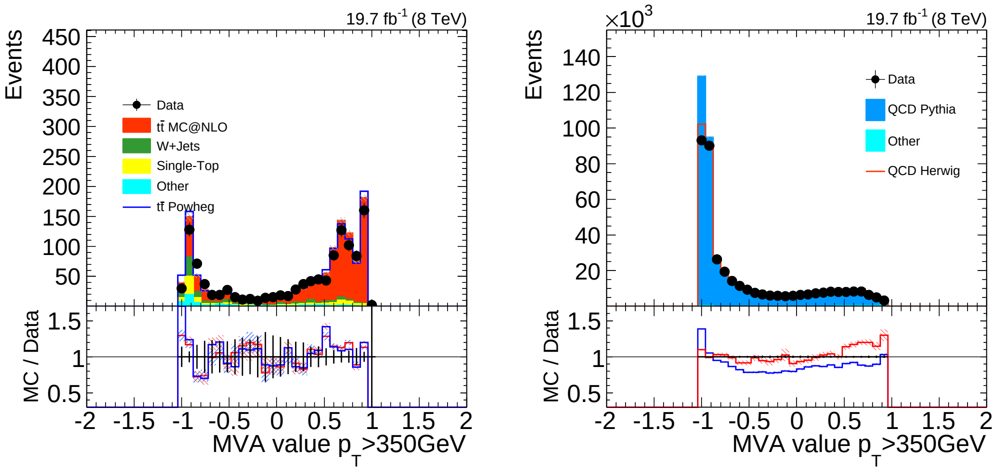


Figure 10.51: MVA values of the BDT after training with 10 substructure variables for a jet $p_T > 350$ GeV for signal (left) and background (right) events. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

10.4.7 Boosted decision tree

The results of the BDT, trained with 10 substructure variables (described in Section 8.2) can be seen in Figure 10.51. The variables used have been shown in the previous section (jet mass m_{jet} in Figure 10.14, number of subjets N_{sub} in Figure 10.16, invariant mass of two subjets $m_{1,2}$ and $m_{2,3}$ in Figure 10.44 (top row and middle row), transverse momentum of the leading subjet p_T in Figure 10.44 (bottom row), Q-jet volatility $\Gamma_{\text{volatility}}$ in Figure 10.45 (top row), pruned N-subjettiness $\tau_{2,\text{p}}/\tau_{1,\text{p}}$, $\tau_{3,\text{p}}/\tau_{2,\text{p}}$ and $\tau_{3,\text{p}}$ in Figure 10.49 (bottom row), 10.50 (middle and bottom row), helicity angle $\Theta_{1,3}$ in Figure 10.46). Here, the MVA value is the projection of the cuts made by the BDT. As mentioned already before, the MVA categorizes jets in signal-like (MVA value 1) and background-like (MVA value -1). In the signal selection values of 1 and -1 can be seen. The peak at -1 is caused by unmerged jets as is illustrated in Figure 10.52. Here the POWHEG $\text{t}\bar{\text{t}}$ is divided into two samples. Collimated means here that the angular distance of each of the generator decay products to the jet momentum axis is smaller than 0.8. In case one of the decay products has a larger distance than 0.8 to the jet momentum axis the jet is counted as not collimated. The decay products of the top quark for jets distributed at -1 are not collimated within the CA8 jet. The double peak structure at 1 is caused by jets containing not all decay products. Jets that contain all decay products are distributed at 1 and are categorized as signal-like, since the BDT is trained on jets with a distance parameter of 0.8. For low transverse momenta the decay products are not fully contained within the jet and are therefore categorized as background-like.

The measured efficiency and mistag rate are shown in Figure 10.53. The integrated

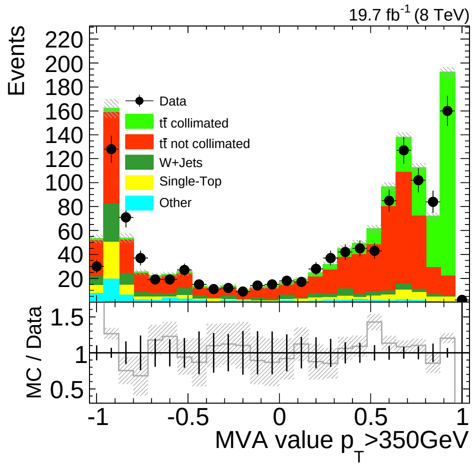


Figure 10.52: MVA values of the BDT after training with 10 substructure variables for a jet $p_T > 350\,\mathrm{GeV}$. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

efficiency for $p_T > 350\,\mathrm{GeV}$ is 28% and the mistag rate for PYTHIA is 2.2% and for HERWIG++ 2.8%, respectively. The working point of the BDT is adjusted in order to have the same efficiency as the CMS top tagger (compare Figure 10.17). This makes the mistag rates comparable and one can observe that the BDT misidentification rate is only half that of the CMS top tagger (mistag rate (CMS): 5.7% ; mistag rate (BDT): 2.8%).

The efficiency and mistag rate of the BDT is independent of the number of primary vertices, as seen in Figure 10.54. This is a remarkable result since the substructure variables used obtain a dependence on the number of primary vertices, e.g. the mass variables.

Summarizing, the BDT obtains a good performance with respect to other top tagging algorithms. The decision variable, the efficiency and the mistag rate are well described by simulation. The studies in this section and in Section 9 show that the BDT can be used as a tagging algorithm in physics analyses.

10.5 Conclusion

Performance studies of top tagging algorithms using the full CMS 8 TeV data set were presented.

Measurements of the efficiency and misidentification rate of all taggers with 8 TeV data have been used to validate the tagger performance. The efficiency is measured using a tag and probe method using CA8 and CA15 jets in a $t\bar{t}$ enriched sample. Misidentification rates have been obtained from light-flavour jets selected in a QCD multijet sample. Measurements of the decision variables of the tagging algorithms have been performed using the signal and background jets and were compared to simulation.

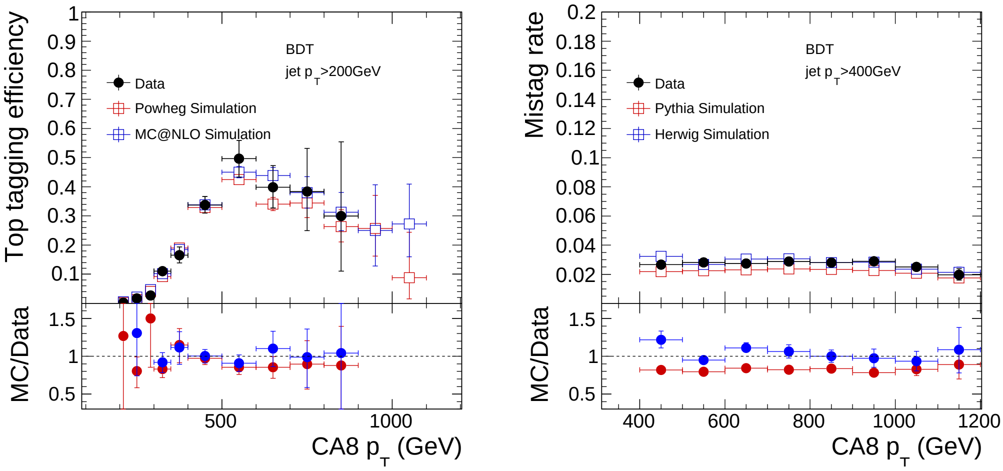


Figure 10.53: Top tagging efficiency (left) and mistag rate (right) as a function of p_T for a BDT trained with 10 substructure variables. The error bars show the combined statistical and systematic uncertainties. At the bottom of each panel the ratio of simulation to data is shown.

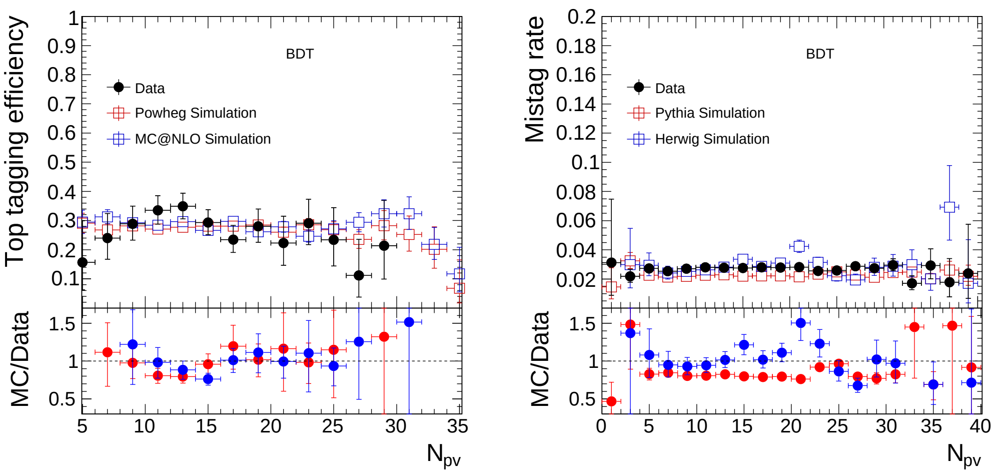


Figure 10.54: Top tagging efficiency (left) and mistag rate (right) for the BDT as function of N_{pv}. The error bars show the combined statistical and systematic uncertainties. At the bottom of each panel the ratio of simulation to data is shown.

In general, the agreement of simulation and data is sufficient. Especially in the signal region no significant discrepancies between data and simulated events by POHWEG and MC@NLO are observed. Two different QCD multijet simulations, PYTHIA 6 and HERWIG++, have been used for a comparison with the measured distributions in data in the background region. Overall, good agreement is found, while some deficiencies of the detailed modeling of substructure variables became apparent. While the misidentification rate is described well by HERWIG++ for all algorithms considered, the misidentification rate obtained by PYTHIA 6 tends to be about 0.2-0.5% too small.

The OptimalR HEP top tagger, the shower deconstruction tagger and the BDT show improved performance compared to top tagging algorithms used previously. Differences in the ratio of data to simulation for the efficiency and mistag rate measurements are small and never larger than for the standard taggers used in CMS. In summary, the performance measurement of the different tagging algorithms in data show the usability of these for future analyses.

11 Introduction to a new heavy object tagger

The top tagging algorithms studied in the previous chapters suffer from a decreasing efficiency with increasing transverse jet momentum. This behavior is of special disadvantage for LHC Run II, where the centre-of-mass energy increased to $\sqrt{s} = 13\,\text{TeV}$, leading to more top quarks with high transverse momentum produced and the stronger collimated decay products. Furthermore, at the high-luminosity run of the LHC, the in-time pile-up will rise, making the identification of top quarks even more difficult.

Most of the available top tagging algorithms use a fixed distance parameter and have been optimized in a specific p_T range (for example the HEP top tagger in the low p_T region and the CMS top tagger for $p_T > 400\,\text{GeV}$). The OptimalR HEP top tagger uses multiple conesizes by shrinking the jet cone and exceeds the performance of the original HEP top tagger. Furthermore, in Section 8 it is shown that substructure variables combined in a BDT lead to a top tagging algorithm with similar performance as the shower deconstruction tagger, the OptimalR HEP top tagger, or the CMS top tagger combined with N-subjettiness.

In this chapter, a novel top tagging algorithm is introduced using a p_T dependent distance parameter and a new subjet algorithm. In Section 11.1 the Heavy Object Tagger with a Variable R approach (HOTVR) is described which is based on the Variable R algorithm (Section 6.3.2.1), finding subjets inspired by the mass jump algorithm (Section 7.8). In Section 11.2 the decision variables, properties and a performance comparison of the HOTVR are presented on generator level. The properties, optimization, performance and validation of the HOTVR in 8 TeV data using the full CMS simulation are shown in Section 11.3.

11.1 Algorithm

In preparation for Run II with the increased centre-of-mass energy, the aim is to develop a tagging algorithm that is able to identify highly boosted top quarks without losing efficiency in the low and intermediate p_T regions. Since the decay products of the top quark are stronger collimated with increasing p_T, it is natural to use a jet algorithm with a variable distance parameter. The tagger introduced here is based on the Variable R algorithm [146].

In order to resolve the hard 3-prong decay, subjets are identified by an algorithm inspired by the mass jump algorithm [184] (compare Section 7.8). The effect of additional radiation and pile-up is reduced by discarding subjets with small fractional momenta. In the following, the HOTVR algorithm will be explained in more detail.

The HOTVR algorithm clusters particles with the sequential clustering method (see Section 6.3.2). Similar to the usual Cambridge/Aachen or anti-k_T algorithms, the HOTVR finds the smallest distance among d_{ij} and $d_{i\mathrm{B}}$, defined as:

$$d_{ij} = \min[p_{\mathrm{T,i}}^{2n}, p_{\mathrm{T,j}}^{2n}]\Delta R_{ij}^2, \tag{11.1}$$

$$d_{i\mathrm{B}} = p_{\mathrm{T}}^{2n} R_{\mathrm{eff}}^2(p_{\mathrm{T}i}), \tag{11.2}$$

$$R_{\mathrm{eff}} = \frac{\rho}{p_{\mathrm{T}}}.$$

Here $d_{i\mathrm{B}}$ includes the variable distance parameter R_{eff} of the Variable R algorithm [146] (Section 6.3.2.1). An algorithm similar to the mass jump algorithm [184] is used for the subjet finding. The difference to the original approach is explained in the following.

For each clustering step the following operations are performed:

- If the smallest variable is d_{ij} and the invariant mass of the clusters m_{ij} is smaller than μ, combine the cluster i and j.
- If the smallest variable is $d_{i\mathrm{B}}$, combine cluster i with the beam. In this case the jet is complete and will not be merged with other jets in subsequent steps. Consequently, as long as the invariant mass of clusters i and j is smaller than the mass jump threshold μ, the algorithm is equal to the Variable R algorithm.
- If the invariant mass of clusters i and j is larger than μ and the variable d_{ij} is smallest in the list check for the mass jump criterion $\theta \cdot m_{ij} > \max[m_i, m_j]$.
  - If a mass jump is found and the criterion is fulfilled, merge clusters i and j and save clusters i and j as subjets. In case clusters i or j have already subjets, merge cluster i and j and do not save i or j again, but keep the subjets previously found. Consequently, the large jet can contain an odd number of subjets.
  - If no mass jump is found, compare the mass of the two clusters and remove the lighter one. This step removes soft particles from the jet.

In summary, the algorithm has five parameters in total:

- R_{min}: The minimum distance parameter of the jet algorithm ($R_{\mathrm{min}} = 0.1$ in this analysis)
- R_{max}: The maximum distance parameter of the jet algorithm ($R_{\mathrm{max}} = 1.5$ in this analysis)
- ρ: Sets the slope of the increasing distance parameter with p_{T}^{-1}. The larger the value of ρ is, the faster the cone shrinks with p_{T}. For $\rho = 0$ the clustering becomes the usual sequential clustering procedure explained in Section 6.3.2 with $R_0 = R_{\mathrm{min}}$. For $\rho \to \infty$ the usual sequential clustering with $R_0 = R_{\mathrm{max}}$ is obtained. The parameter ρ is given in units of GeV.
- μ: Sets the mass jump threshold. Until this mass threshold is reached the algorithm behaves like the Variable R algorithm. The unit of μ is GeV.
- θ: Controls the strength of mass jump. The value of θ influences the number of subjets found. The parameter θ takes values in the interval $[0, 1]$. The larger the value of θ, the more subjets are found, but less soft clusters are rejected.

With this algorithm, any number of subjets can be found. Jets with only one subjet are removed from further consideration.

So far the algorithm is not specific for identifying top quarks. The following selection is applied at the standard working point:

- The subjet p_T of the first two leading jets has to be greater than 20 GeV.
- The number of subjets has to be greater or equal to three, $N_{sub} \geq 3$.
- The minimum pairwise mass has to be $m_{min} = \min(m_{12}, m_{13}, m_{23}) > 50$ GeV.
- The jet mass m_{jet}, which for the HOTVR is equal to the invariant mass of all subjets has to be $140\,\text{GeV} < m_{jet} < 220\,\text{GeV}$.
- The momentum fraction of the leading subjet to the large jet has to fulfill $p_{T,sub1}/p_{T,jet} < 0.8$, which ensures that the momentum is distributed among the subjets and not just carried by the leading subjet.

The HOTVR clustering can be performed with different values of n (see Equation (11.1) and (11.2)). In this analysis the Cambridge/Aachen algorithm is used with a value of $n = 0$. A schematic illustration of the algorithm is given in Figure 11.1.

11.2 Studies on generator level

For the following studies, no detector simulation is used. As an input for the tagging algorithms, the final state generator particles that do not decay or fragment further are used. This is done in order to have experiment-independent results since the HOTVR can be used in many high energy physics applications. To be independent of the experiment two Monte Carlo samples for $t\bar{t}$ and QCD multijet events are produced. For both samples the event generator PYTHIA 8 [202–204] is used with the multiple parton interaction tune Monash 2013 [205], the NNPDF2.3 QCD+QED [206] LO with $\alpha_S(M_Z) = 0.130$ PDFs and a centre-of-mass energy of $\sqrt{s} = 13$ TeV.

The decision variables of the HOTVR are shown in Figure 11.2 - 11.5. The distributions are shown for signal events ($t\bar{t}$) and background events (QCD multijet).

The jet mass is shown in Figure 11.2. The mass distribution is shown for different p_T selections and for signal and background events. Three peaks for each p_T selection are visible. The peak around 80 GeV is caused by partially merged jets that contain just the W boson decay products. In addition, the peak between 100 GeV and 150 GeV is caused by partially merged jets, when the b jet is merged with one of the light quarks from the W boson decay. The QCD multijet background shows a smoothly falling distribution. The jet mass variable provides good separation power between signal and background. This separation power increases with increasing p_T.

In Figure 11.3 the minimum pairwise mass can be seen. Signal events peak around the W boson mass and the distribution for background events is a smoothly falling distribution. The jump around 30 GeV is caused by the mass jump threshold μ.

The number of subjets found by the HOTVR is shown in Figure 11.4. Most of the signal events have three or more subjets whereas a large fraction of background events has just two subjets. The separation power of this variable increases with increasing transverse momentum.

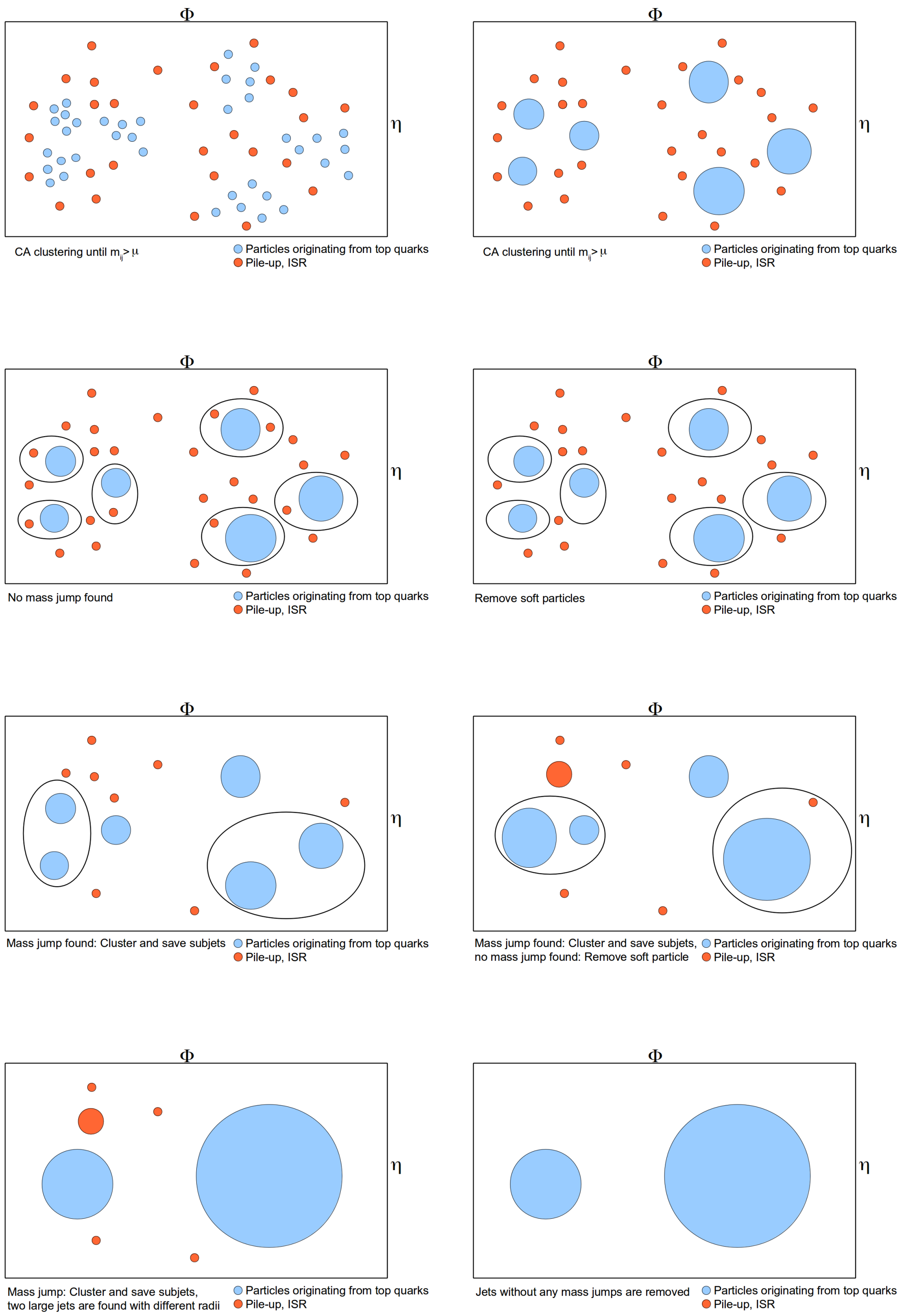


Figure 11.1: Schematic illustration of the HOTVR clustering procedure.

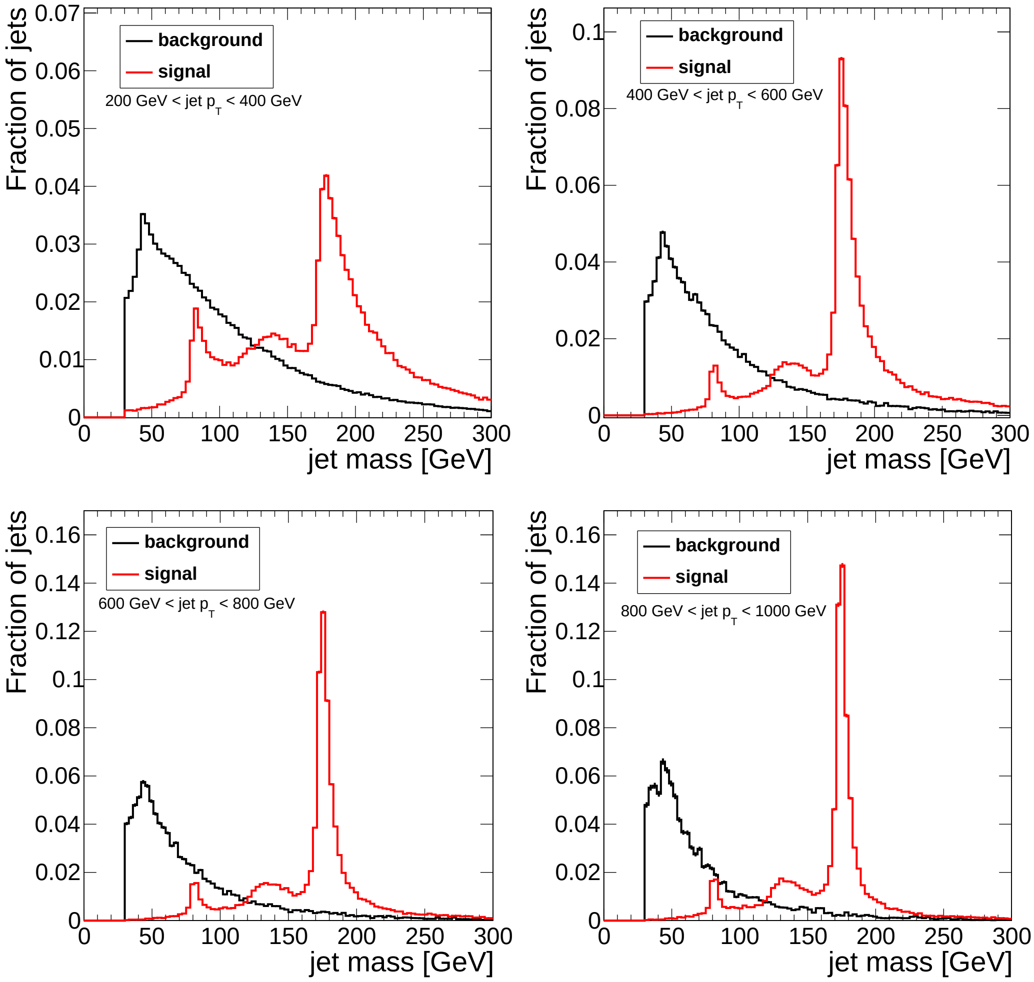


Figure 11.2: Jet mass for signal and background events for the HOTVR on generator level for different p_{T} selections.

Figure 11.5 shows the p_{T} fraction variable $p_{\mathrm{T,sub1}}/p_{\mathrm{T,jet}}$. For background events, the distribution shows that the leading subjet carries most of the jet p_{T} (distribution peaks around 1), whereas the transverse momentum of the jet is more distributed among all subjets for signal events.

In Figure 11.6 the jet mass and the minimum pairwise mass are shown after a selection on the other decision variables is made. The jet mass distribution is shown after selecting events with $N_{\mathrm{sub}} > 2$, $m_{\mathrm{min}} > 50\,\mathrm{GeV}$ and $p_{\mathrm{T,sub1}}/p_{\mathrm{T,jet}}$ <0.8. The jet mass of background events is shifted towards slightly higher values than before the selection, but good seration power is still achieveable. Selecting the top quark mass ($140 < m_{\mathrm{jet}} < 220\,\mathrm{GeV}$) rejects further background events. The minimum pairwise mass is shown after selecting events with $N_{\mathrm{sub}} > 2$, $140 < m_{\mathrm{jet}} < 220\,\mathrm{GeV}$ and $p_{\mathrm{T,sub1}}/p_{\mathrm{T,jet}}$ <0.8. The distribution is more signal-like than before the selection for background events. Nevertheless a good separation power is observed for this variable after applying a selection on the other decision variables.

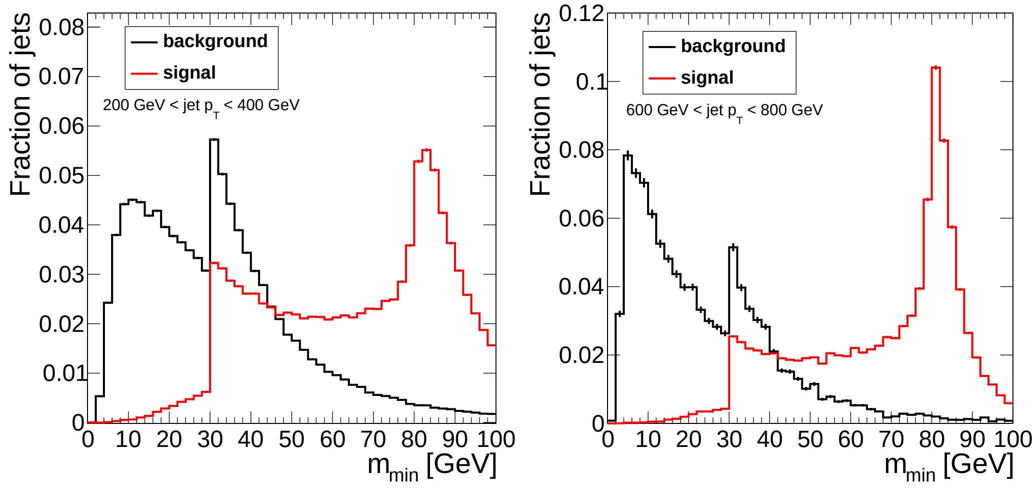


Figure 11.3: Minimum pairwise mass for the HOTVR on generator level for two different p_{T} selections.

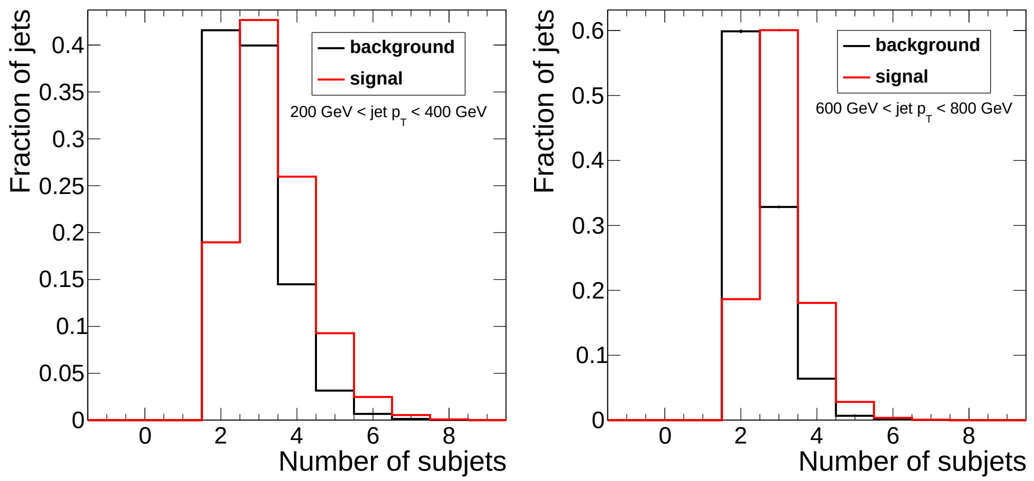


Figure 11.4: Number of subjets found by the HOTVR on generator level for two different p_{T} selections.

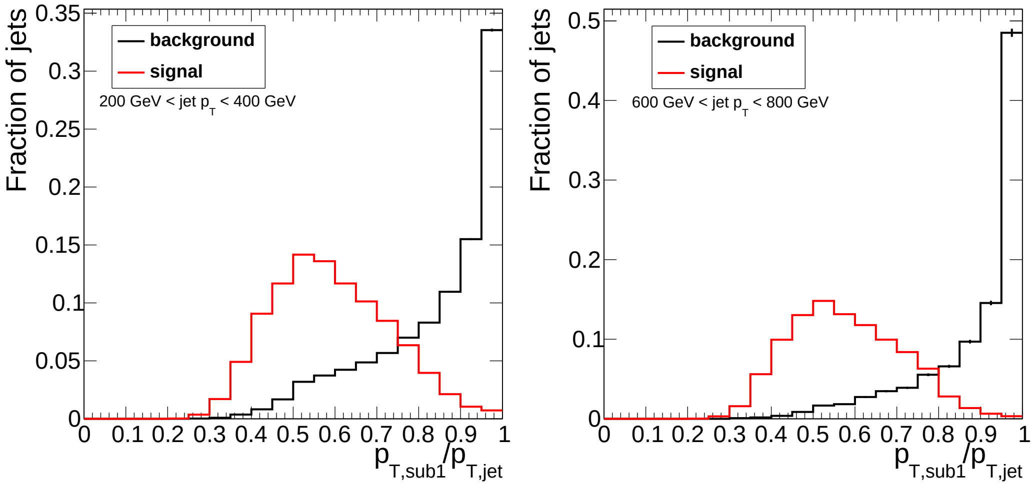


Figure 11.5: p_{T} fraction for the HOTVR on generator level for two different p_{T} selections.

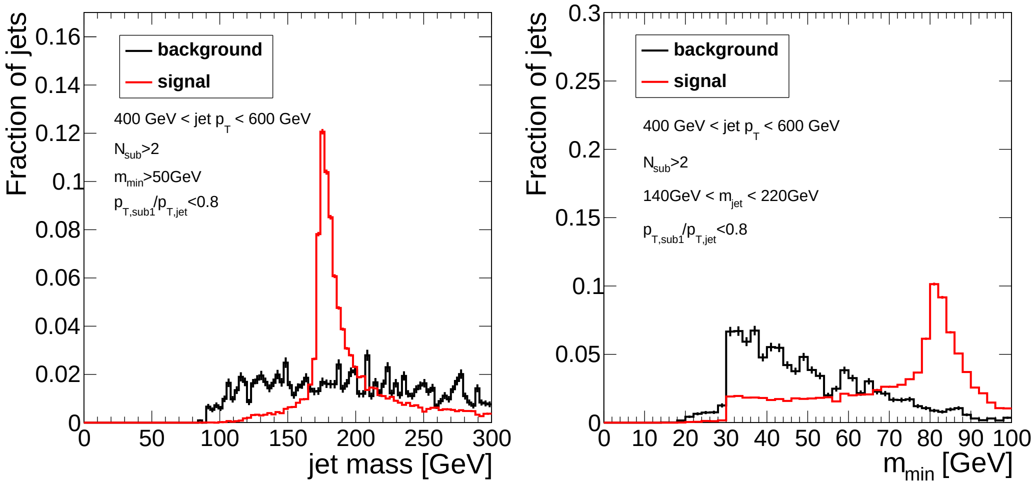


Figure 11.6: Jet mass (left) and minimum pairwise mass (right) for signal and background events after a selection on the other decision variables is made.

11.2.1 Infrared Safety

In the following the infrared (IR) safety, as described in Section 2.2, of the HOTVR is investigated. For this test, to the N final state generator particles that do not decay or fragment further a maximum of M soft particles are added, where M is a number between 0 and N. The properties of the soft particles are randomly distributed as follows:

$$p_{\mathrm{T}} \in [0, 10^{-10}]\,\mathrm{GeV}, \tag{11.3}$$

$$\phi \in [0, 2 \cdot \pi], \tag{11.4}$$

$$\eta \in [-5, 5], \tag{11.5}$$

$$E \in [0, 10^{-8}]\,\mathrm{GeV}. \tag{11.6}$$

In order to validate the IR safety several algorithms are tested and the results of Ref. [137] are reproduced (Midpoint & JetClu). In Figure 11.7 the fraction of events is shown that fail the infrared safety test for different algorithms.

In total each algorithm is tested on $3 \cdot 10^6$ different events. The percentage of the events that fails the test is shown on the x-axis. The test fails if just one of the jets in the events differs after adding the soft particles. As expected, the anti-k$_\mathrm{T}$ algorithm is infrared safe and none of the $3 \cdot 10^6$ events are failing the test.

To test the CMS top tagger, the infrared safety of the subjet algorithm is tested, meaning that the hard prongs of the subjets must stay the same after adding soft particles. The algorithm is tested for two choices of δ_p ($\delta_p = 0, \delta_p = 0.05$). One can see that the algorithm is not infrared safe for very small δ_p, as then one soft particle can be found as additional subjet in the declustering. The CMS top tagger is infrared safe for the standard parameter of $\delta_p = 0.05$.

The HOTVR algorithm is tested for several cases. As the HOTVR algorithm uses already a different jet finding algorithm (not the standard CA clustering) the infrared safety test is applied on the jet finder, testing the Variable R algorithm for infrared safety. It is found that the HOTVR jet finder is infrared safe, zero events failed the test.

In the next step the subjets found with the HOTVR are tested for several parameters. One can see that the algorithm is infrared safe for $\mu > 0$. However, it is not infrared safe for $\mu = 0$. In this case the algorithm finds mass jumps already between the soft particles, which changes the number of subjets found. For $\theta = 1.0$ a mass jump and consequently also subjets are found in each combination step, but since $\mu > 30\,\mathrm{GeV}$ the soft particles do not contribute.

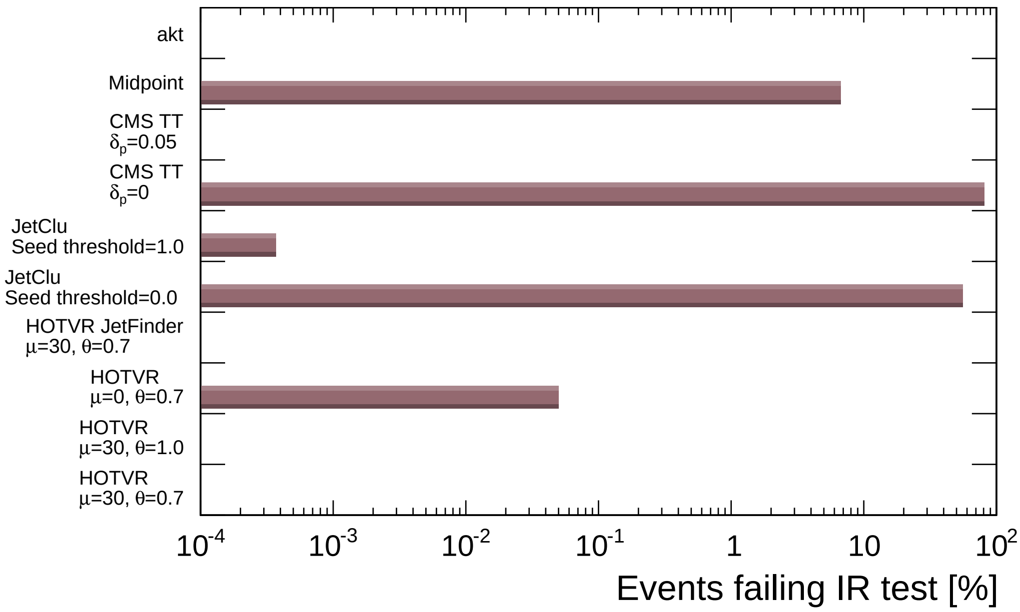


Figure 11.7: Infrared safety of several jet algorithms. On the x-axis the percentage of events is given that failed the test.

11.2.2 Collinear Safety

The collinear safety of the tagging algorithms is tested by deciding randomly with a 50% probability for each particle if a collinear splitting happens. In the case the particle gets split, the momenta of the outgoing particles are calculated as

$$p_{\text{original}} = (p_x, p_y, p_z, E), \tag{11.7}$$
$$p_1 = (L \cdot p_x, L \cdot p_y, L \cdot p_z, L \cdot E), \tag{11.8}$$
$$p_2 = ((1-L) \cdot p_x, (1-L) \cdot p_y, (1-L) \cdot p_z, (1-L) \cdot E). \tag{11.9}$$

Here the particle p_{original} is split into particles p_1 and p_2. The parameter L is here a uniformly distributed random number between 0 and 1.

For each algorithm $3 \cdot 10^6$ different events are used to test collinear safety. The test fails for an event if any of the jets or subjets changes after the splitting. In order to validate the test the CMS cone algorithm is tested as it is known to be collinear unsafe. In Figure 11.8 the percentage of events failing the collinear safety test can be seen for several algorithms. As expected the CMS cone algorithm does not pass the test.

The CMS top tagger passes the test even for $\delta_p = 0$. In case of the HOTVR, the jet finder is tested as well as the subjet algorithm. Similar to above $\mu = 0$ is the only case where the HOTVR does not pass the test. The HOTVR is consequently collinear safe for a sufficient high mass jump threshold μ. The reason is that in practice no single clusters (particle flow candidates) appear which are heavier than 30 GeV.

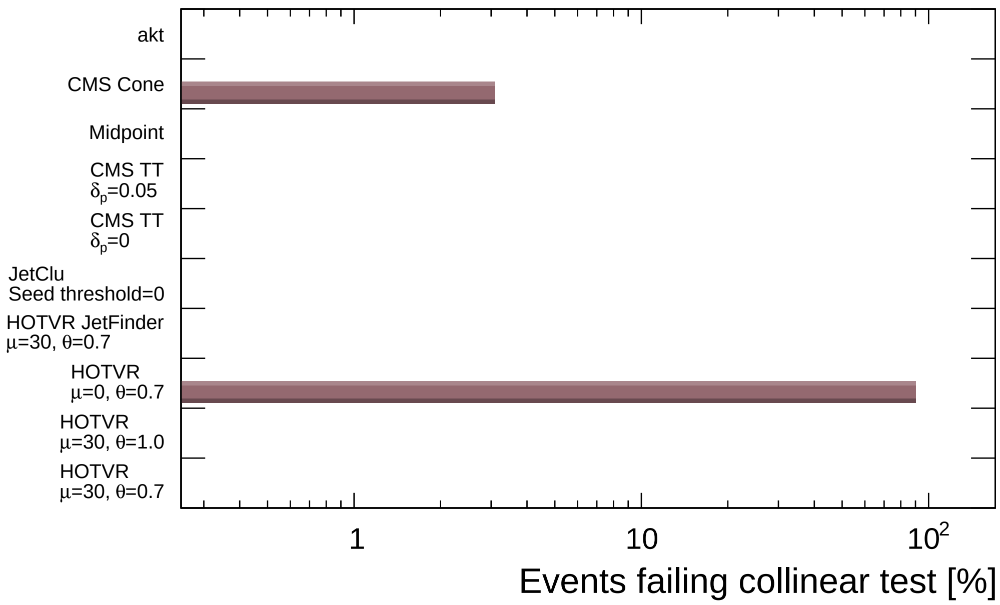


Figure 11.8: Collinear safety of several jet algorithms. On the x-axis the percentage of events is given that failed the test.

11.2.3 Computational time

Another important property of a jet algorithm is the computational time, as in high energy physics the amount of data analyzed can be very large.

In order to compare the existing top tagging algorithms with the HOTVR the computational time is measured for each tagging algorithm separately. Just the relevant part of the code is timed for 3×10^6 signal and background events. The measurement is repeated five times per algorithm and the average computational time is determined. Afterwards the CPU time is normalized to the time of the CMS top tagger.

The fastjet 3.0.5 framework [207] is used for the Cambridge/Aachen and anti-k_T sequential clustering algorithms. The CMS, HEP and OptimalR HEP top tagger are taken from the CMSSW 6.2.5 framework [208]. The Variable R algorithm and the mass jump algorithm are taken from fastjet contribs 1.017 [209].

The results can be seen in Table 11.2.3. The fastest top tagging algorithm is the CMS top tagger, using the Cambridge/Aachen jet clustering with a distance parameter of $R_0 = 0.8$. Clustering just the jets requires almost the same time as using in addition the CMS top tagger. The Cambridge/Aachen clustering with a distance parameter of $R_0 = 1.5$ is about 70% slower. The anti-k_T clustering performs similar to the CA clustering with a distance parameter of $R_0 = 0.8$, but about 20% faster for the large distance parameter. Top tagging algorithms using a large jet distance parameter are consequently slower than the ones using a small one. One of these algorithms is the HEP top tagger. With respect to the CMS top tagger it is 90% slower, comparing it with the CA clustering with a distance parameter of $R_0 = 1.5$ one obtains that the algorithm is 20% slower, caused by different filtering steps. The OptimalR HEP top tagger is in comparison to the HEP top tagger just 10% slower, but needs already twice the CPU time as the CMS top tagger. Nevertheless the usage of the updated version does makes sense taking into account the better performance and the just slightly increasing computing time.

The shower deconstruction tagger is the slowest algorithm. One observes a difference in the performance between the processing of $t\bar{t}$ events and QCD multijet events. In QCD multijet events, the number of microjets is higher and therefore the computing time increases significantly, due to the computing time increasing exponentially with the number of microjets. The shower deconstruction tagger takes more than 60 times more CPU time than the CMS top tagger. With an increasing centre-of-mass energy the number of microjets will increase and therefore potentially slow down shower deconstruction.

The Variable R clustering needs up to seven times more CPU time than the CA clustering with $R_0 = 0.8$. This is due the algorithms' implantation in the used fastjet contrib version.

Only about 30% slower than the CMS top tagger is the mass jump algorithm, but this is not an actual top tagging algorithm. The HOTVR performs about 50% slower than the CMS top tagger, but in comparison to the OptimalR HEP top tagger, which also operates with a variable distance parameter, it is about 50% faster.

11.2.4 Performance

In Section 9 it is shown that the comparison of top tagging algorithms, using different distance parameters, in simulated events is only possible if the matching efficiency of the different distance parameters is similar. The HOTVR algorithm uses a p_T dependent

Tagging algorithm	CPU time	
	$t\bar{t}$	QCD multijet
Cambridge/Aachen $R_0 = 0.8$	0.98 ± 0.01	0.99 ± 0.06
Cambridge/Aachen $R_0 = 1.5$	1.72 ± 0.04	1.76 ± 0.10
anti-k_T $R_0 = 0.8$	0.97 ± 0.07	0.95 ± 0.06
anti-k_T $R_0 = 1.5$	1.47 ± 0.02	1.57 ± 0.16
CMS top tagger	1.00 ± 0.06	1.00 ± 0.06
HEP top tagger	1.88 ± 0.07	1.90 ± 0.12
OptimalR HEP top tagger	1.99 ± 0.03	2.27 ± 0.23
Shower deconstruction $R_0 = 0.8$	2.28 ± 0.05	9.10 ± 0.59
Shower deconstruction $R_0 = 1.5$	59.81 ± 0.56	73.35 ± 4.13
Variable R clustering	7.11 ± 0.07	3.39 ± 0.19
Mass jump algorithm	1.22 ± 0.03	1.34 ± 0.09
HOTVR	1.51 ± 0.09	1.55 ± 0.10

Table 11.1: CPU time comparison, performed with 3×10^6 events, averaged over five runs and normalized to the CPU run time of the CMS top tagger.

distance parameter and a comparison over a large p_T range to other top tagging algorithms with a fixed distance parameter is not possible, since the matching efficiency of the HOTVR is p_T dependent. Therefore, now the efficiency and mistag rate will be measured as an effective efficiency, which is a convolution of the tagging efficiency ϵ_{tag} and the matching efficiency ϵ_{match},

$$\epsilon_{\text{eff}} = \epsilon_{\text{match}} \cdot \epsilon_{\text{tag}}. \tag{11.10}$$

The effective efficiency is a measure of how many top quarks are identified by the top tagging algorithms. In contrast the efficiency in Section 9 is a measure of how many matched top quarks are identified. The comparison with the definition of the effective efficiency is more reasonable, as it will be illustrated with the following example:

Imagining a top tagging algorithm that reconstructs top quarks within a small radius of 0.2 by applying only a selection on the jet mass. This algorithm identifies just extremely boosted top quarks, the tagging efficiency ϵ_{tag} will be around 100%, but since the radius of the jet is really small and most of the top quarks will be not collimated within a conesize of 0.2, the matching efficiency ϵ_{match} will be low. However, the effective efficiency from Equation (11.10) will give a moderate and more realistic efficiency of the tagger. One can calculate the effective efficiency and mistag rate as follows:

$$\epsilon_{\text{eff}} = \frac{N(\text{tops(partons)}_{\text{tag,match}})}{N(\text{all tops (partons)})}. \tag{11.11}$$

On generator level the denominator for the efficiency is now changed to be the number of all hadronically decaying top quarks (compare Equation (9.1)). For the mistag rate in Section 9 just the parton (light quark or gluon) with the highest p_T is matched. For the effective efficiency we are considering all partons (light quarks or gluons) with $p_T > 100\,\text{GeV}$ as denominator. Here one is not interested in matching extremely low p_T partons, since they may not have caused a signal-like jet.

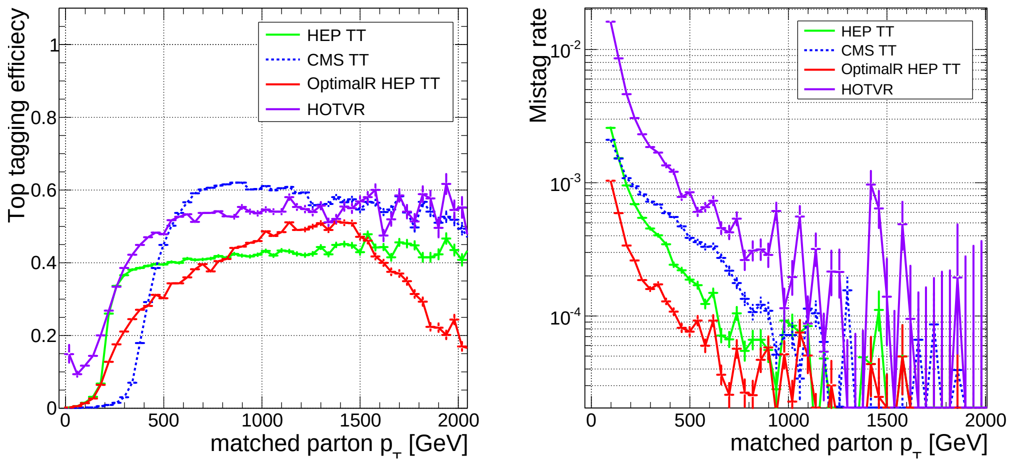


Figure 11.9: Efficiency (left) and mistag rate (right) as a function of the generator parton p_T for different top tagging algorithms.

In Table 11.2 the standard working points of the CMS, HEP and shower deconstruction top tagging algorithms are listed and will be used in the following sections. The standard working point of the HOTVR is defined above (Section 11.1) and is not listed again.

CMS top tagger	$140\,\text{GeV} < m_{\text{jet}} < 250\,\text{GeV}$ $m_{\text{min}} < 50\,\text{GeV}$ $N_{\text{sub}} > 2$
HEP top tagger	$140\,\text{GeV} < m_{123} < 250\,\text{GeV}$ $f_W = 0.15$
OptimalR HEP top tagger	$140\,\text{GeV} < m_{123} < 250\,\text{GeV}$ $f_W = 0.15$ $R_{\text{opt}} - R_{\text{opt}}^{\text{calc}} > 0$
Shower deconstruction	$\log(\chi) > 3.5$

Table 11.2: Standard working points of the top tagging algorithms.

In Figure 11.9 the efficiency and mistag rate as a function of the parton p_T is shown. The top tagging algorithms are used at their standard working points defined in Table 11.2. Below 500 GeV, the HOTVR obtains the highest efficiency in comparison to the other algorithms. The plateau of about 60% efficiency is reached for values of $p_T > 600\,\text{GeV}$. For the HOTVR the mistag rate in this p_T range is a factor two higher than for the CMS top tagger. The mistag rate for all top tagging algorithms decreases with increasing transverse momentum. The OptimalR HEP top tagger achieves the lowest mistag rate over the whole p_T range, although its efficiency is decreasing after 1400 GeV.

To compare top tagging algorithms at their optimized working points ROC curves are produced. To obtain the following ROC curves (Figure 11.10) the tagging algorithms are used at their standard working points and just the additional variable N-subjettiness τ_3/τ_2 is scanned.

This is done in order to get a more realistic comparison, since in practice the working points of the top tagging algorithms are technically not changed according to the jet p_T, but kept at their standard working points. Furthermore most of the analyses use a selection on the N-subjettiness ratio τ_3/τ_2 to improve the signal to background ratio.

In Figure 11.10 ROC curves are presented where N-subjettiness is scanned and the other variables are kept at the standard working points. One can observe that the mistag rate for all top tagging algorithms is on average one order of magnitude lower on generator level than on reconstruction level, due to the resolution of detector components. Figure 11.10 shows the ROC curves for different p_T selections. In the low p_T region ($200\,\text{GeV} < p_T < 400\,\text{GeV}$) the CMS top tagger is not working well due to unmerged jets. The OptimalR HEP top tagger and the HEP top tagger have a slightly better performance than the HOTVR. Already in the p_T region from 400 GeV to 600 GeV the HOTVR algorithm becomes the best performing algorithm. The intermediate p_T region from 600 GeV to 800 GeV shows that the HOTVR algorithm performs better with respect to the other algorithms. The CMS top tagger performs slightly worse than the HOTVR and it is better in the high efficiency region. In the highest p_T region ($800\,\text{GeV} < p_T < 1000\,\text{GeV}$) the HOTVR is the best algorithm in the lower efficiency range, whereas the CMS top tagger overtakes for high efficiencies. In this region the statistics are low and the statistical uncertainty on the ROC curves becomes large.

In summary, the HOTVR is the best overall algorithm when combined with N-subjettiness. It has just a slightly worse performance in the lowest p_T range, but performs better than other algorithms in all higher p_T selections. This makes the algorithm especially interesting for Run II as well as for analyses with top quarks with their p_T distributed over a large range.

11.3 Studies including the full CMS detector simulation

In the following section the HOTVR is studied on reconstruction level with the full CMS detector simulation. A $Z' \to t\bar{t}$ simulation is used in order to increase the statistics for signal events in the high p_T region. For background events, a QCD multijet sample is used. Both samples are produced with a centre-of-mass energy of $\sqrt{s} = 8\,\text{TeV}$. More details on the MC samples are given in Section 5.

11.3.1 Optimization of the algorithms parameters

The main parameters of the HOTVR clustering algorithm will be further optimized for top tagging in the following section. The linear correlations between the three parameters can be seen in Figure 11.11, the correlations between the parameters are low. In order to find the optimal values for ρ, θ and μ ROC curves based on a flat p_T spectrum of the parton are produced and the N-subjettiness ratio τ_3/τ_2 is scanned, while keeping the selections on the other decision variables constant as in Section 11.2.4.

Assuming the three parameters to be independent from each other, always one parameter is changed, whereas the others are kept fixed. Since the goal is a constant

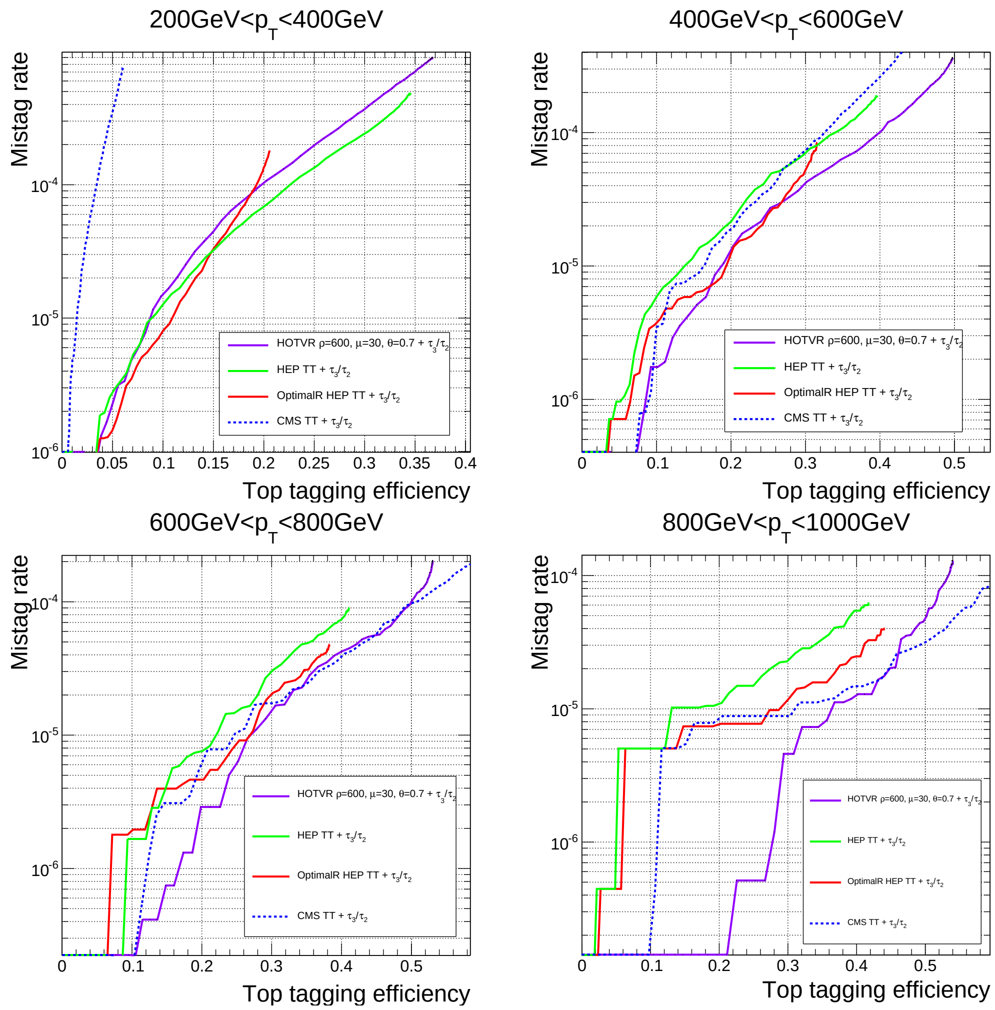


Figure 11.10: ROC curves on generator level for different top tagging algorithms at their standard working point and scanned N-subjettiness. The ROC curves are shown for different p_{T} selections.

and good performance over the whole p_{T} range, three p_{T} selections, one for low p_{T} ($200\,\mathrm{GeV} < p_{\mathrm{T}} < 400\,\mathrm{GeV}$), one for intermediate p_{T} ($600\,\mathrm{GeV} < p_{\mathrm{T}} < 800\,\mathrm{GeV}$), and one for high p_{T} ($1000\,\mathrm{GeV} < p_{\mathrm{T}} < 1200\,\mathrm{GeV}$) are chosen.

In Figure 11.12 the mass threshold μ is changed (20 GeV, 30 GeV, 40 GeV), whereas the other parameters are kept constant. The influence of the mass threshold μ on the performance is negligible. For all p_{T} selections the ROC curves show almost the same behavior. As we have seen in Section 11.2.1 the parameter should not be chosen to be too low in order to make the algorithm infrared safe. In this thesis the parameter is set to 30 GeV, unless it is stated differently.

The mass jump criterion parameter θ is changed (0.6, 0.7, 0.8), as shown in Figure 11.13. In the low p_{T} region the mass jump criteria of $\theta = 0.6$ and $\theta = 0.7$ give the best results. In the intermediate and high p_{T} region $\theta = 0.7$ and $\theta = 0.8$ are the best performing parameters. Therefore, the value of θ is chosen to be 0.7 since it has the best performance

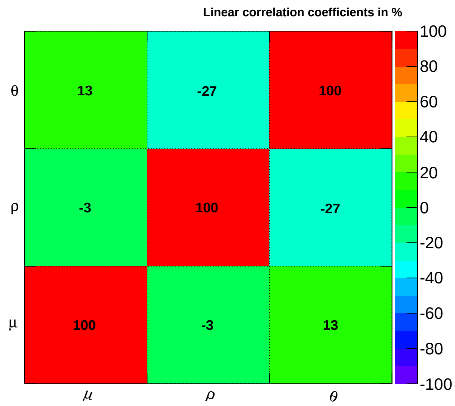


Figure 11.11: Linear correlation matrix for the three parameters of the HOTVR algorithm.

in all p_T regions.

Figure 11.14 shows the ROC curves for different parameters of ρ (400, 500, 600, 700). In the low p_T region the ROC curve with $\rho = 400$ is the best performing configuration, whereas for higher transverse momentum the performance with $\rho = 400$ is not the best anymore. In the intermediate and high p_T regions the best configuration is found with $\rho = 500$ or $\rho = 600$. The smaller the value of ρ is the smaller the distance parameter becomes with increasing p_T, leading to jets that do not contain all decay products of the top quark. Since $\rho = 500$ and $\rho = 600$ show almost the same performance in the higher p_T regions, in this analysis $\rho = 600$ is kept.

In summary the parameters of the algorithm are optimized, although further improvements can be achieved by scanning all parameters at the same time. Furthermore, the parameters are chosen to identify hadronically decaying top quarks. Other particles (like W, Z or H boson) can be identified with the algorithm by re-optimizing the parameters. In Table 11.3 the optimal parameters are listed.

μ_{opt}	30 GeV
θ_{opt}	0.7
ρ_{opt}	600 GeV

Table 11.3: Optimal parameters to identify hadronically decay top quarks with the HOTVR.

11.3.2 Characteristic properties

In the following the characteristic properties of the HOTVR using the optimized parameters determined in the previous section are shown.

In Figure 11.15 the influence of the parameters ρ, μ and θ on the jet mass for signal and background events is shown. The value of $\rho = \infty$ corresponds to a clustering algorithm without cone shrinking and it corresponds to a clustering with the Cambridge/Aachen

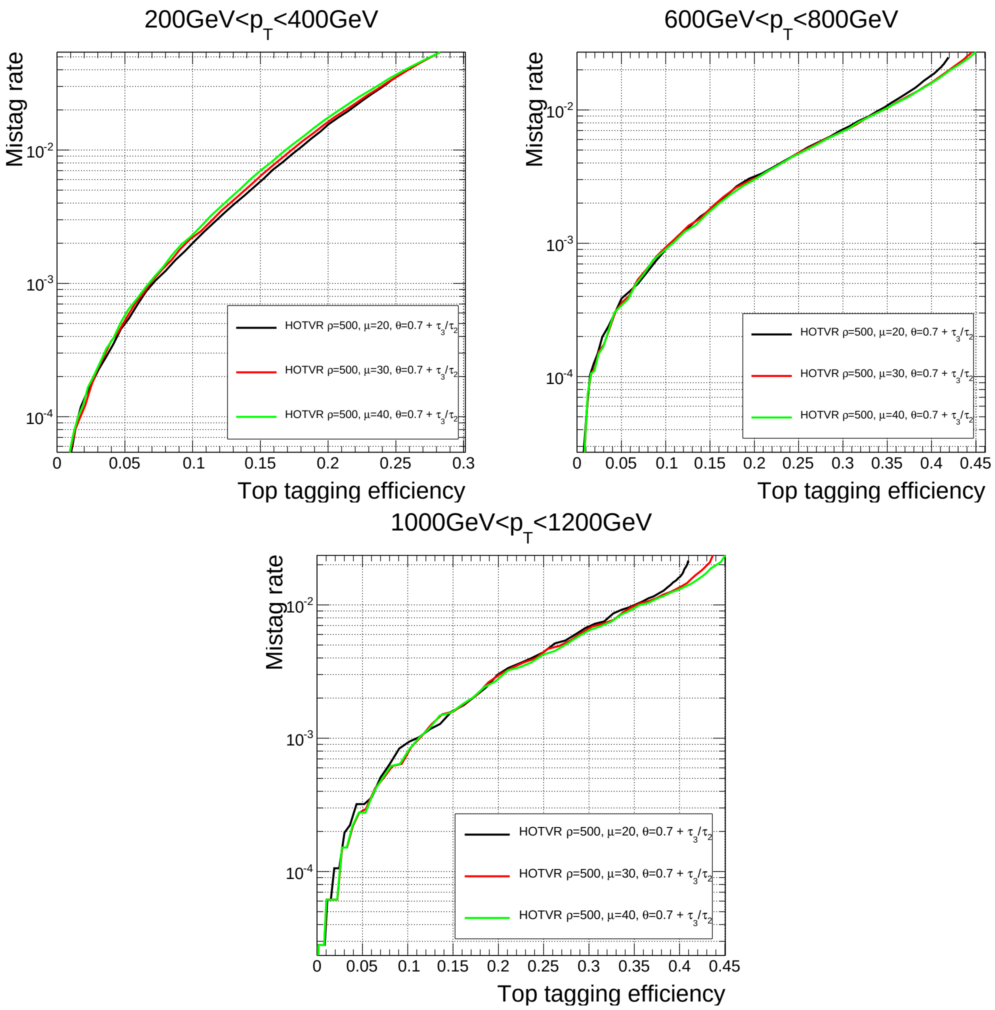


Figure 11.12: ROC curves for different values of μ and different $p_{\rm T}$ selections. The other parameters are kept constant.

algorithm with a distance parameter of 1.5. For this choice of ρ a resonance of the mass of the top quark is visible, but the distribution exhibits a large tail towards increasing values of the jet mass. This is caused by the large distance parameter, which results in particles from pile-up or initial state radiation clustered into the jet, making the jet heavier than the actual top quark. For background events a smoothly falling distribution is observed for masses $> 150\,\text{GeV}$. For $\theta = 1$ a mass jump and therefore a subjet is found in every clustering step. With this parameter setting, no soft cluster gets removed from the jet. In this case the jet has all its constituents as subjets. This means that the blue curve in Figure 11.15 shows how the cone shrinking affects the jet mass without using the subjet finding of the algorithm. In comparison to the black curve (without cone shrinking) one can see that the top quark mass resonance is much more pronounced and the tail to higher mass values is largely reduced. Background events get pushed towards lower values, out of the top quark mass window. The red curve shows the jet mass when the cone shrinking as well as the rejection of soft clusters and the subjet finding is activated by setting $\theta = 0.7$. The

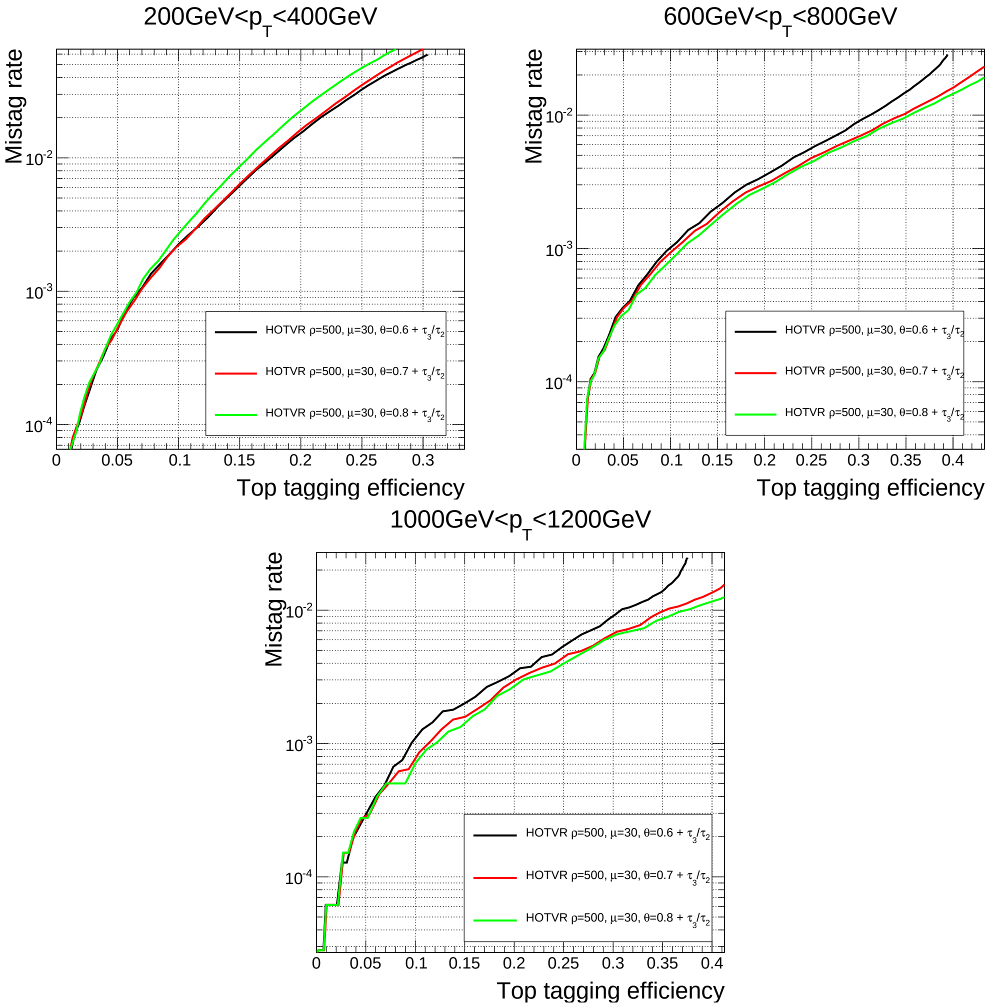


Figure 11.13: ROC curves for different values of θ and different p_{T} selections. The other parameters are kept constant.

peak value in signal events moves closer to the top quark mass, whereas for background events it moves further away from the top quark mass window.

In Figure 11.16 the effective conesizes of a jet clustered with the HOTVR algorithm as a function of the transverse momentum of the matched top quark is shown. The value of ρ is chosen to be 600 GeV. The jet conesize is determined by finding the clustered particle with the largest ΔR to the jet axis. As expected, the conesize shrinks with increasing p_{T}. The maximum distance at low p_{T} is larger than $R_{\mathrm{max}} = 1.5$, which is caused by the Cambridge/Aachen clustering, as seen in Section 6.3, Figure 6.2.

In the following two simulated $\mathrm{t\bar{t}}$ events are shown to compare the HOTVR with the anti-$\mathrm{k_T}$ clustering. Extremly soft ghost particles are used to make the jet area visible.

In Figure 11.17 hadronically decaying generator top quarks are shown in red and their decay products in yellow. Two top quarks with transverse momenta of 320 GeV and 520 GeV are contained in the event. In the event 16 primary vertices are reconstructed. The top graph shows the clustering with the anti-$\mathrm{k_T}$ algorithm with a constant distance

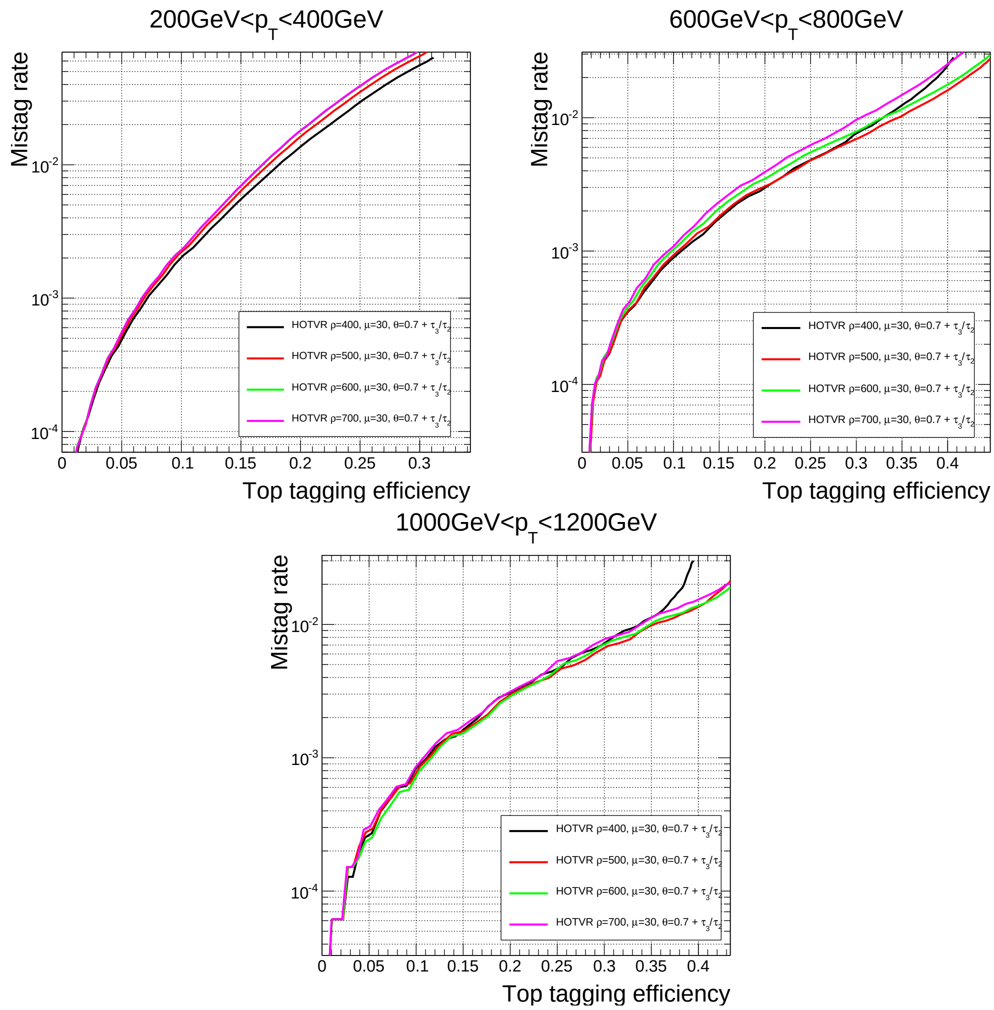


Figure 11.14: ROC curves for different values of ρ and different p_T selections. The other parameters are kept constant.

parameter of 0.8. Three jets are found with $p_T > 150$ GeV. All jets have almost the same size in the ϕ-η plane and have a circular shaped area. The leading jet (green) clusters all decay products of the top quark and reconstructs the top quark mass with 211 GeV. The second leading (grey) jet originates not from a top quark but from additional radiated gluons. This jet is reconstructed with a mass of 79 GeV. The third leading jet (blue) is reconstructed close to the other top quark in the event. The decay products of the top quarks are not clustered within the anti-k_T 0.8 jet. The reconstructed mass of this jet is only 40 GeV and is not reconstructed as a top quark.

The bottom graph in Figure 11.17 shows the same event clustered with the HOTVR. The leading jet (green) is obtained with a smaller area than the jet reconstructed with the anti-k_T algorithm. Consequently the reconstructed top quark mass is lower, around 179 GeV. The distance parameter is smaller than 0.8 and the jet clusters less pile-up and additional radiation, reconstructing the top quark therefore better. The area of the third leading jet is much larger in comparison to the anti-k_T clustering. Consequently,

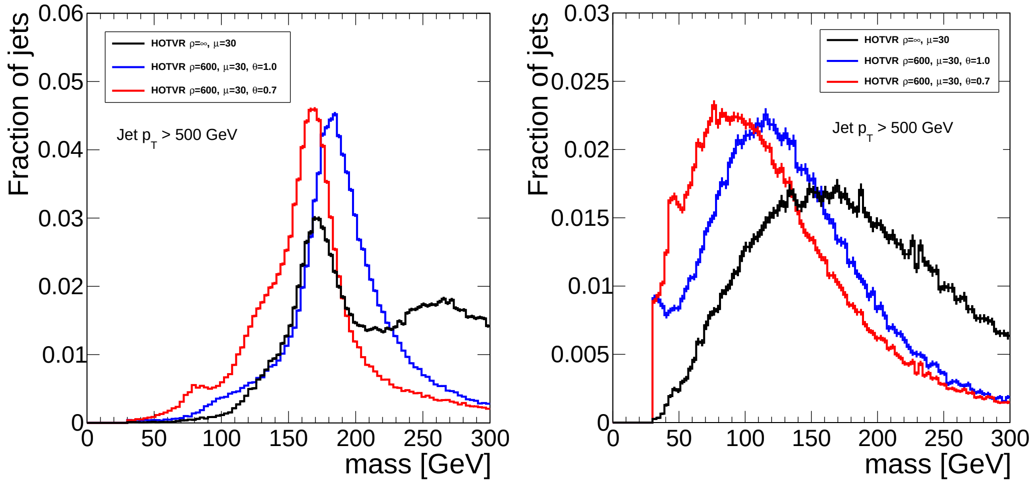


Figure 11.15: Jet mass for the HOTVR for different clustering parameters for signal events (left) and for background events (right).

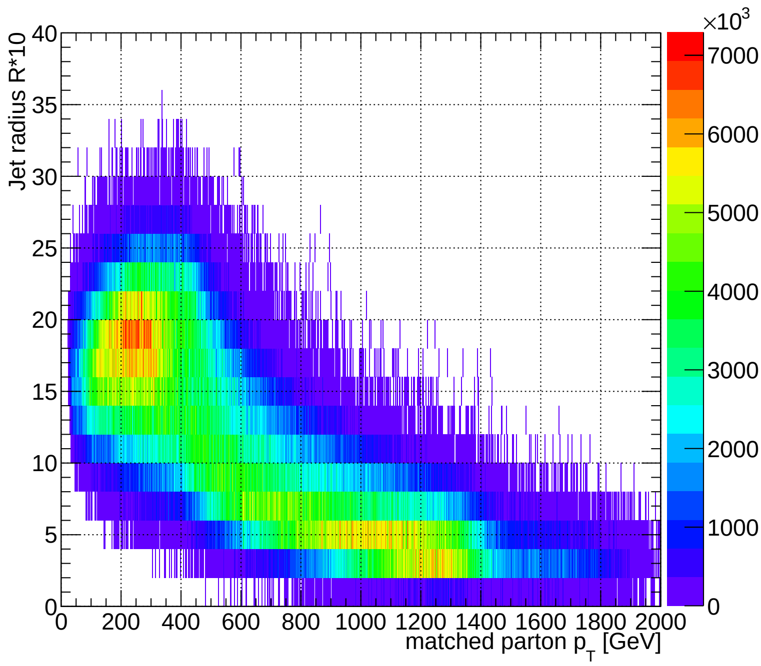


Figure 11.16: Effective conesize of the HOTVR jets as a function of the matched generator top quark p_{T}.

Tagger	$p_{\mathrm{T,top1}}$ [GeV]	$p_{\mathrm{T,top2}}$ [GeV]	m_{jet1} [GeV]	m_{jet2} [GeV]	m_{jet3} [GeV]
HOTVR	520	320	179	220	296
AKT	520	320	211	40	79
HOTVR	900	-	166	-	-
AKT	900	-	211	-	-

Table 11.4: Jet mass reconstruction by the anti-$\mathrm{k_T}$ and HOTVR algorithm for two events.

the clustering incorporate more particles and the reconstructed jet mass is 296 GeV. In contrast to the anti-$\mathrm{k_T}$ clustering the HOTVR clusters the decay products of the second top quark in the second leading jet. It reconstructs the mass with 220 GeV.

In Figure 11.18 the same event is displayed projected on the ϕ-η plane. Here only the jets found by the HOTVR are shown with the same color code as in Figure 11.17. In the top graph cluster are shown that are rejected by the algorithm due to the comparison with a harder jet (light blue). The clustering with the Variable R algorithm would have incorporated more particles. Especially for the leading jet just clusters are kept centered around the decay products (yellow). In pink clusters are shown that are clustered to jets that did not obtain a mass jump. In the white area another jet is reconstructed with $p_\mathrm{T} < 150$ GeV.

In the bottom graph of Figure 11.18 one can see the subjets found by the HOTVR algorithm. Here the subjets are indicated by different colors. The leading jet consists of exactly three subjets. The second leading jet that contains the second top quark in the event consists of five subjets, where two of the subjets have low transverse momenta and the three leading subjets are centered around the decay products.

A second event is displayed in Figure 11.19. This event is a semi leptonic event with a transverse momentum of the hadronically decaying top quark of about 900 GeV and 20 primary vertices. The decay products of the top quark, displayed in yellow, have a small angular distance to the top quark due to the high transverse momentum. The top graph shows the clustering with the anti-$\mathrm{k_T}$ algorithm with a distance parameter of 0.8. Only one jet is found with a transverse momentum $p_\mathrm{T} > 150$ GeV. The jet contains all decay products of the top quark and reconstructs the top quark mass with 211 GeV.

In the bottom graph the same event clustered with the HOTVR is displayed. As with the anti-$\mathrm{k_T}$ algorithm only one jet is found, but with a smaller area, leading to a smaller reconstructed top quark mass of 166 GeV.

In Figure 11.20 the event is displayed projected on the ϕ-η plane. The bottom graph shows in light blue the clusters, rejected by the algorithm due to the comparison to a harder cluster. Pink clusters belong to jets that did not obtain any mass jump and are therefore rejected by the algorithm. In the bottom graph the subjets of the jet are shown in different colors. The HOTVR algorithm identifies four subjets surrounding the decay products of the top quark.

In Table 11.4 the reconstruction of the two events by the anti-$\mathrm{k_T}$ algorithm and HOTVR is summarized.

In Figure 11.21 an example QCD multijet background event is shown containing one high p_T gluon (shown in red) and one high p_T light quark (shown in yellow). The event is clustered with the anti-$\mathrm{k_T}$ algorithm with a distance parameter of 0.8 (top) and the HOTVR algorithm (bottom). For both algorithms two jets are found with $p_\mathrm{T} > 150$ GeV.

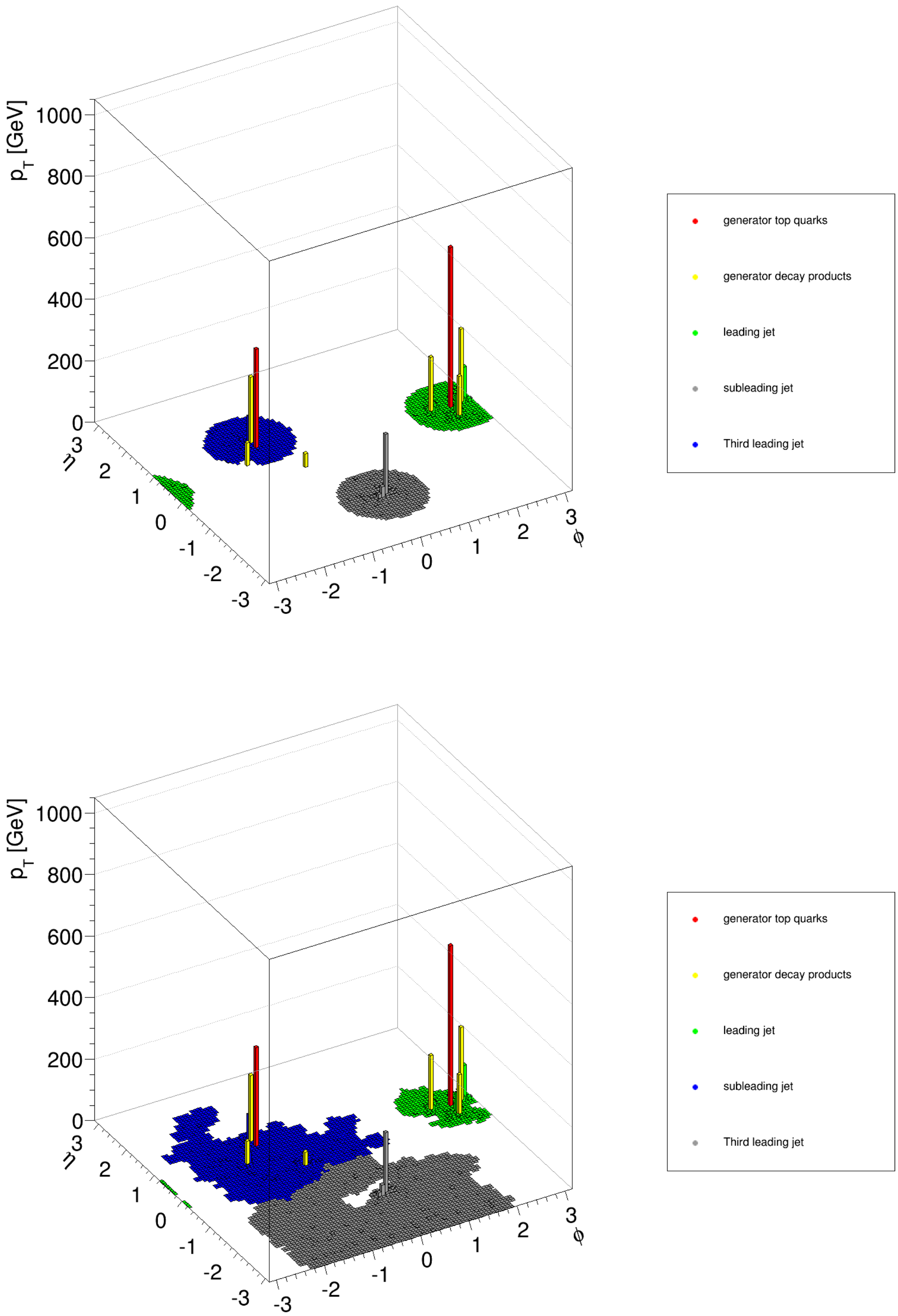


Figure 11.17: Reconstruction of a $t\bar{t}$ event with the anti-k_T algorithm (top) and the HOTVR (bottom). In red generator top quarks are shown and in yellow their decay products.

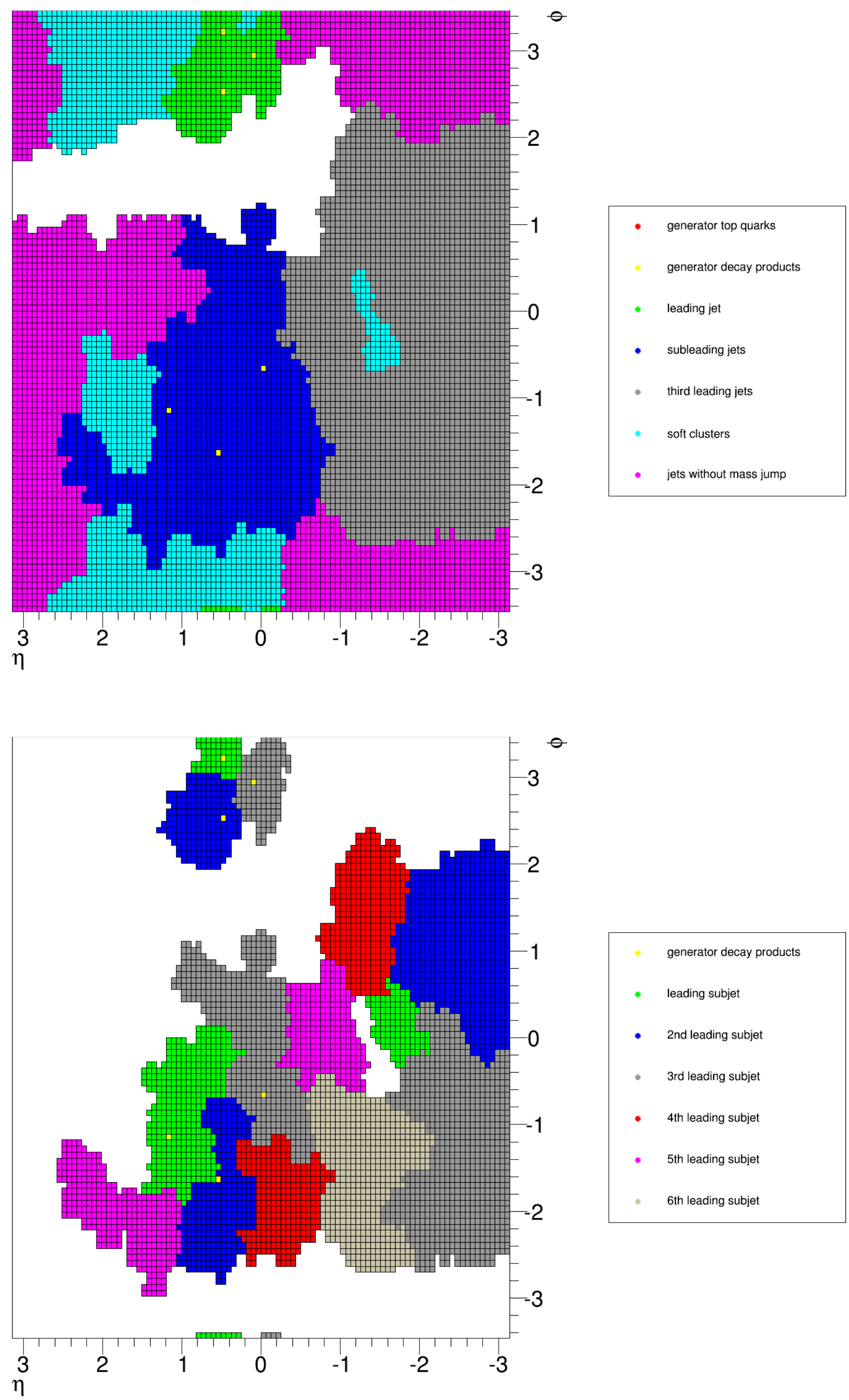
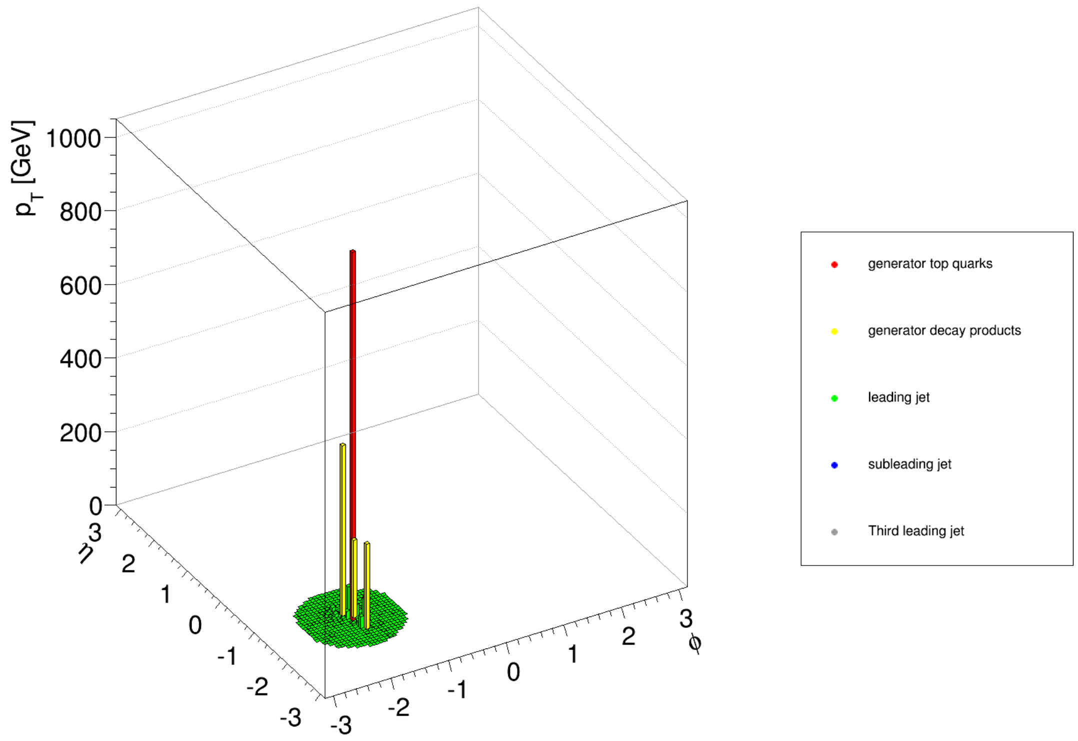


Figure 11.18: Clustered $t\bar{t}$ event with the HOTVR. Rejected clusters by the algorithm (top) and subjets found by the HOTVR algorithm (bottom).



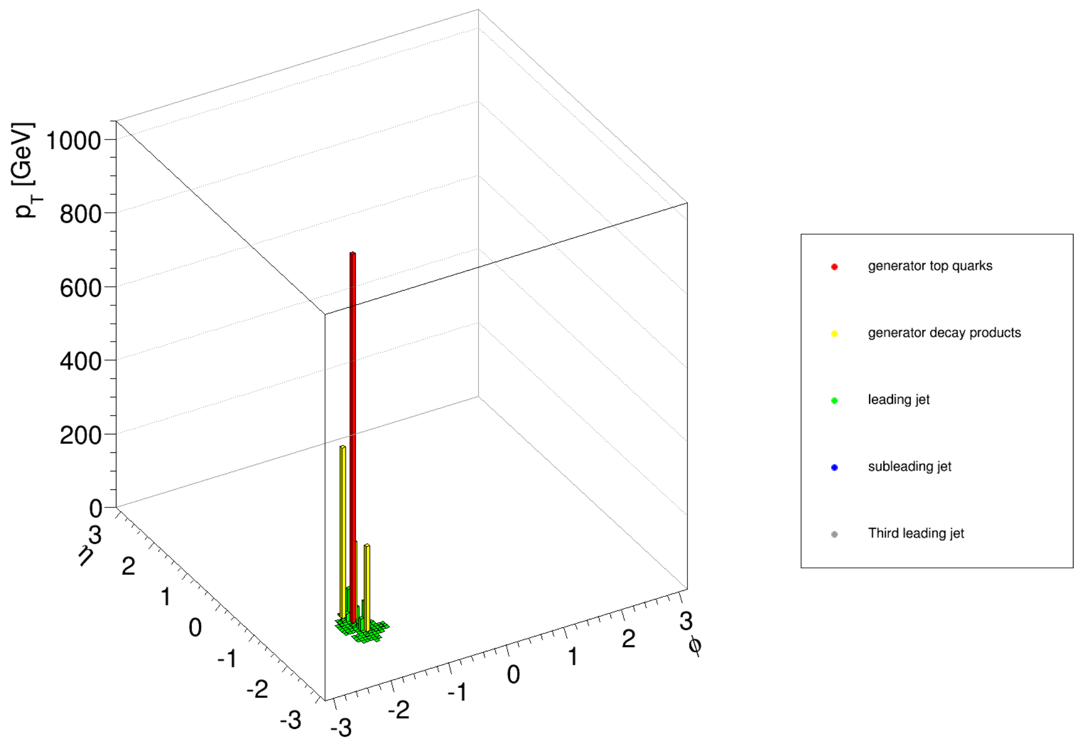
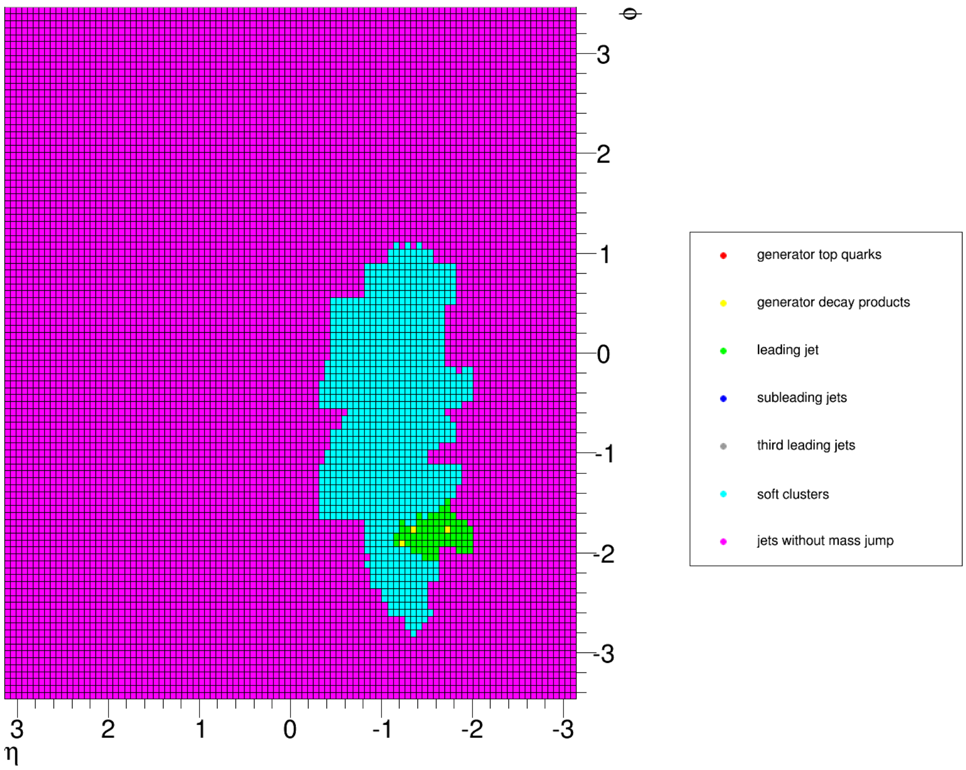


Figure 11.19: Reconstruction of a $t\bar{t}$ event with the anti-k_T algorithm (top) and the HOTVR (bottom). In red generator top quarks are shown and in yellow their decay products.



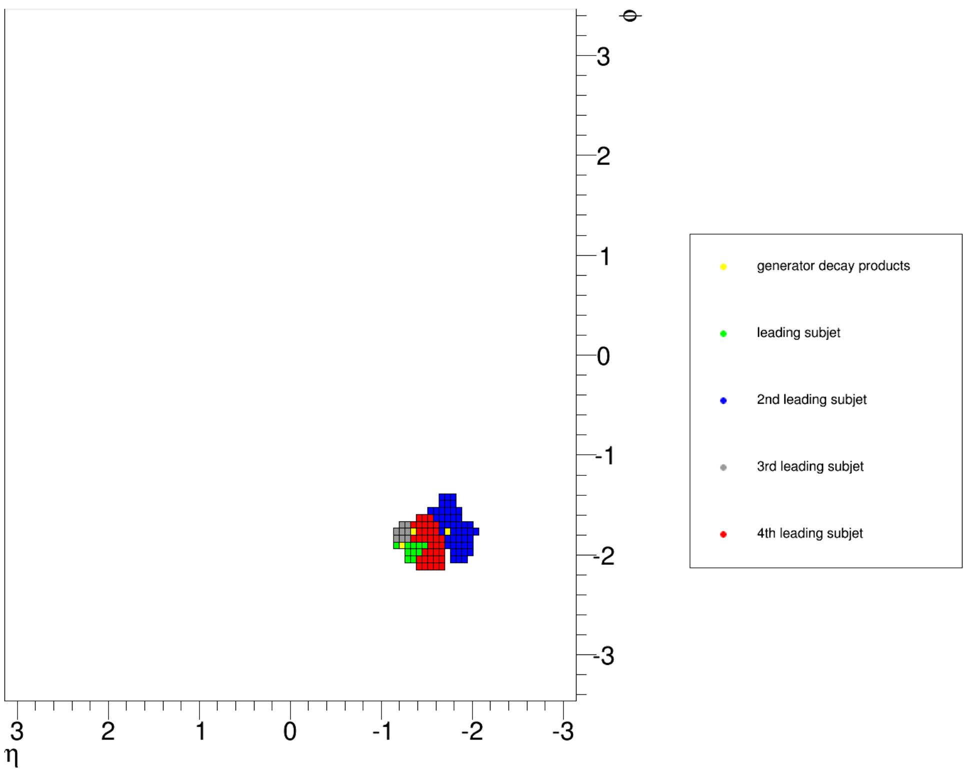


Figure 11.20: Clustered $t\bar{t}$ event with the HOTVR. Rejected clusters by the algorithm (top) and subjets found by the HOTVR (bottom).

The area of both jets is smaller with the HOTVR clustering than with the anti-$\mathrm{k_T}$ clustering. Two effects are causing a smaller jet size for HOTVR: The distance parameter is smaller with respect to the constant distance parameter of the anti-$\mathrm{k_T}$ jet, whereas the dominant effect making the jet area smaller is the rejection of light clusters in this event. The mass for the two reconstructed jets is therefore smaller with the HOTVR clustering than with the anti-$\mathrm{k_T}$ clustering.

Decision variables

In the following the decision variable of the HOTVR algorithm are shown and explained.

The number of subjets, found by the mass jump criteria, is shown in Figure 11.22. The two figures show distributions of the subjet multiplicities for two p_T selections ($200\,\mathrm{GeV} < p_\mathrm{T} < 400\,\mathrm{GeV}$ and $1000\,\mathrm{GeV} < p_\mathrm{T} < 1200\,\mathrm{GeV}$) for signal ($\mathrm{Z'} \to \mathrm{t\bar{t}}$) and background events (PYTHIA QCD multijet). The separation power of this variable increases with increasing transverse momentum, in the high p_T region the signal events peak at three subjets whereas the background events peak at two subjets. It should be noted that jets with just one subjet are discarded already in the clustering.

The jet mass for the HOTVR is shown in Figure 11.23. For reason of comparison also the jet mass of the CMS top tagger is shown, which is essentially the mass of a CA8 jet. The figure shows the jet mass for four different p_T selections. For the lowest p_T selection ($200\,\mathrm{GeV} < p_\mathrm{T} < 400\,\mathrm{GeV}$) the jet mass for the HOTVR shows a peak around 200 GeV, whereas the jet mass of the CA8 jet shows a peak at the W boson mass around 80 GeV. In this low p_T region the decay products are not collimated within a CA8 jet, so a top quark mass peak can not be expected. In this low p_T region the HOTVR clustering finds jets with rather large distance parameters, and can therefore cluster all decay products within one jet. However, there is still a shoulder of lower mass values, indicating that the decay products of some top quarks are not clustered into a single jet. Since the jet distance parameter is rather large, the top quark mass is reconstructed too high since pile-up and initial state radiation can contribute to the jet mass.

For the p_T selection of $400\,\mathrm{GeV} < p_\mathrm{T} < 600\,\mathrm{GeV}$ (upper right) most of the decay products of the top quark are already collimated within a CA8 jet and a peak at the top quark mass is visible also for the CMS top tagger. There is still a large shoulder visible for low values of the jet mass, which is due to unmerged jets. In this p_T region the HOTVR jet mass peaks at the top quark mass and the shoulder to low mass values almost vanishes. A tail towards high mass values is still visible.

With increasing p_T (bottom panels) the peak in the jet mass of the HOTVR algorithm is sharper and the mass resolution improves. The peak value is constant at the top quark mass, whereas the peak position of the CA8 jet moves to higher values with increasing p_T due to the constant distance parameter.

In Figure 11.24 the mass distribution for background events is shown, again in comparison with the jet mass of the CMS top tagger. In the lower p_T region of $400\,\mathrm{GeV} < p_\mathrm{T} < 600\,\mathrm{GeV}$ less events are in the top mass window for the CMS top tagger than for the HOTVR. In the very high p_T region this behavior is inverted.

In summary, the jet mass variable gains more separation power for higher transverse momentum compared to the mass of the CMS top tagger. In the low p_T region fully merged jets are reconstructed, which will be shown to result in a non-vanishing efficiency.

In Figure 11.25 the optimal mass selection as a function of the parton p_T is shown for

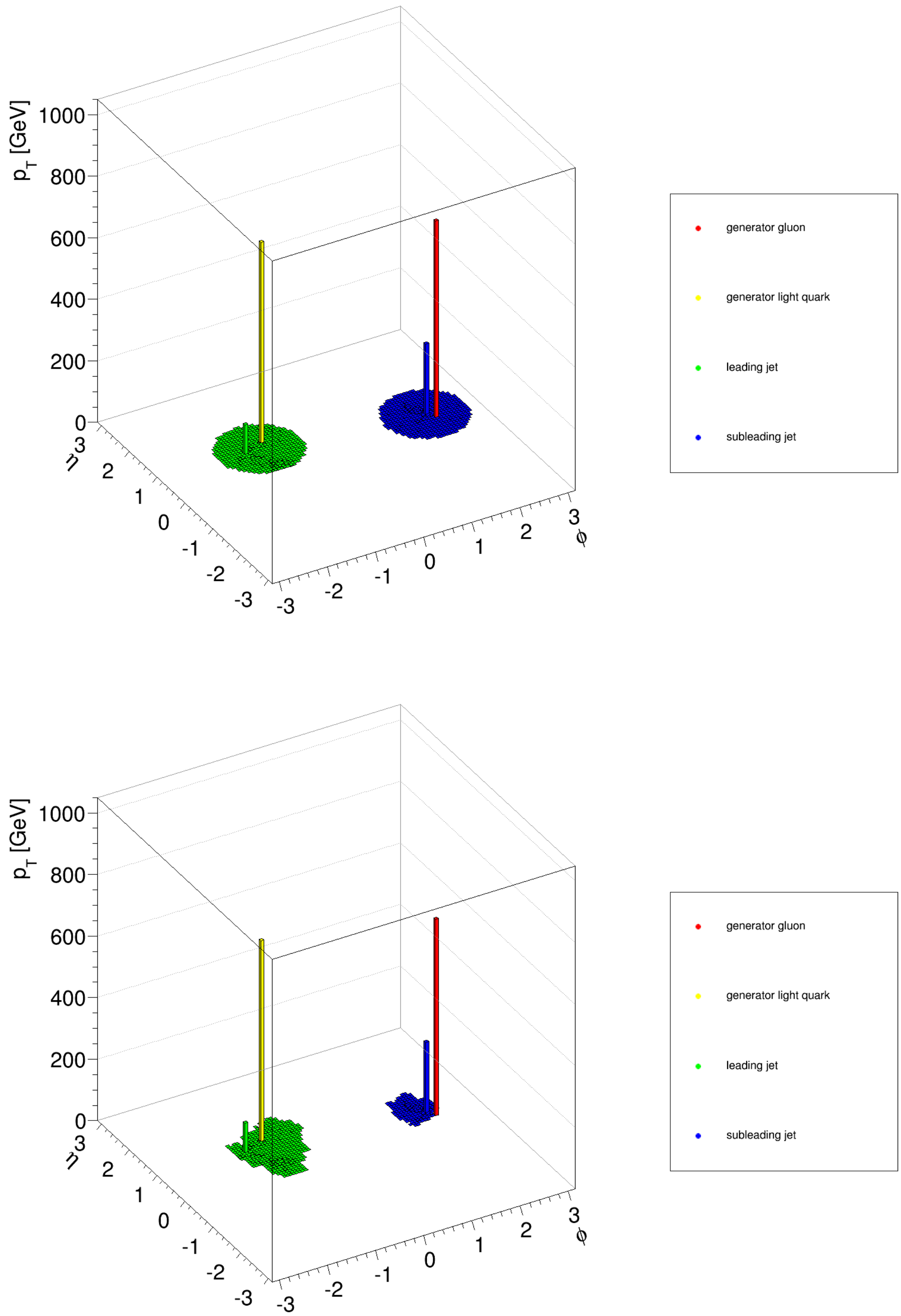


Figure 11.21: Reconstruction of a QCD multijet event with the anti-k_T algorithm (top) and the HOTVR (bottom). In red generator gluons are shown and in yellow light quarks.

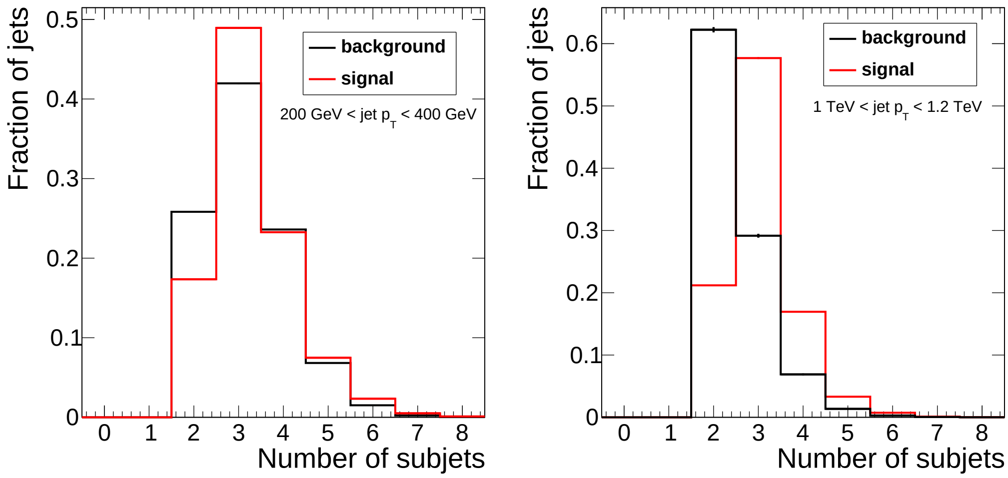


Figure 11.22: Number of subjets found with the HOTVR for two p_T selections.

an efficiency of 30%. Here the optimal mass window is found by maximizing the signal to background ratio for various mass selections. This is done for the HOTVR and the CMS top tagger. For low transverse momenta the mass window of the CMS top tagger is very large, caused by the small distance parameter of 0.8. Here a high efficiency can only be reached if the mass window is very broad. In this region the mass window of the HOTVR is much smaller but shifted to higher mass values. With increasing p_T the mass window of the HOTVR becomes smaller and is centered around the top quark mass. This behavior means that the mass window is very stable with transverse momentum and does not have to be changed for different p_T regions. The mass window for the CMS top tagger gets broader with increasing p_T and shifts to higher mass values.

In Figure 11.26 the minimum pairwise mass for background events and signal events is shown. The variable is calculated just for the leading three subjets. For signal events one can observe a peak around the W boson mass at 80 GeV, whereas the distribution is smoothly falling for background events. In both distributions one can observe a sharp cutoff at 30 GeV. This is caused by the mass jump threshold μ, which is set to 30 GeV. Subjets can only be identified if the invariant mass of clusters i and j is higher than μ. Since clusters i and j can be lighter than μ, one also obtains a shoulder below the mass jump threshold.

The p_T fraction of the leading subjet to the large jet for background and signal events can be seen in Figure 11.27. The background distribution peaks around 1, which shows that the leading subjet in background events carries a significant momentum fraction of the large jet, whereas the distribution for signal events peaks around 0.6. The transverse momentum is distributed more equally among the subjets for signal events than for background events.

The linear correlations between the four decision variables are shown in Figure 11.28. Correlations of about 50% and 70% between the mass variables m_{jet}, m_{min} and the number of subjets N_{sub} are observed. These correlations are due to the fact that the jet mass m_{jet} is calculated from all subjets and m_{min} from the pair of subjets with the lowest invariant mass. Therefore, the jet mass is correlated with the number of subjets and with the minimum pairwise mass. These variables show a smaller anti-correlation of about -30%

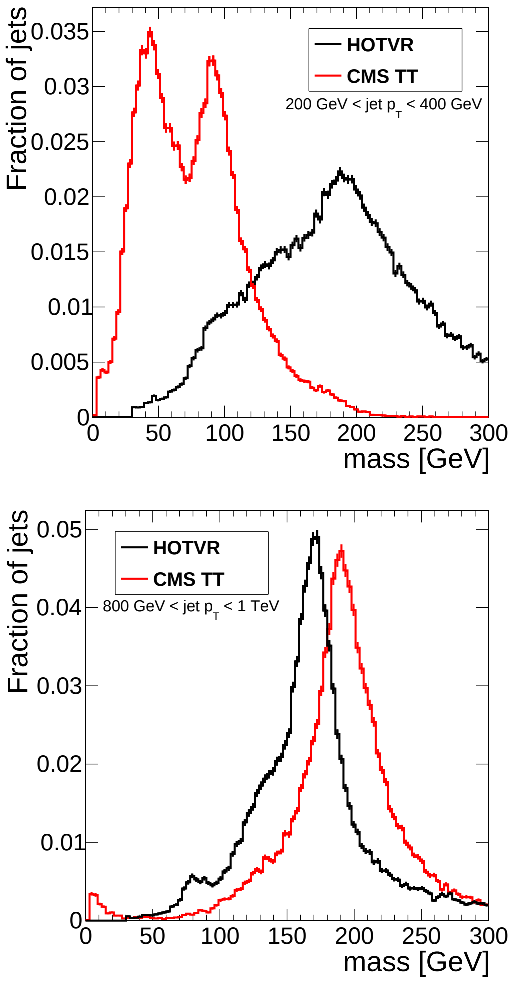


Figure 11.23: Jet mass of the HOTVR and the CMS top tagger for different p_T regions for signal events.

to the p_T fraction variable.

11.3.3 Performance

Figure 11.29 shows the efficiency and mistag rate as a function of the generator top quark p_T and the generator parton p_T, respectively. The efficiency and mistag rate are calculated using Equation (11.11). For low transverse momenta ($p_T < 400\,\text{GeV}$) an efficiency of about 20-40% is obtained with the HOTVR, the mistag rate in this p_T range is obtained between 6-8%. Using the CMS top tagger or the shower deconstruction tagger in this p_T range leads to an efficiency below 5%. For the CMS top tagger and the shower deconstruction tagger the efficiency and the mistag rate peak around 800 GeV and decrease with increasing p_T. The efficiency of the HOTVR decreases slower. The mistag rate stays higher than the mistag rates of the other algorithms.

Nevertheless one can see that the HOTVR has a good efficiency over the whole p_T range.

In Figure 11.30 the ROC curves for the CMS top tagger, the OptimalR HEP top tagger

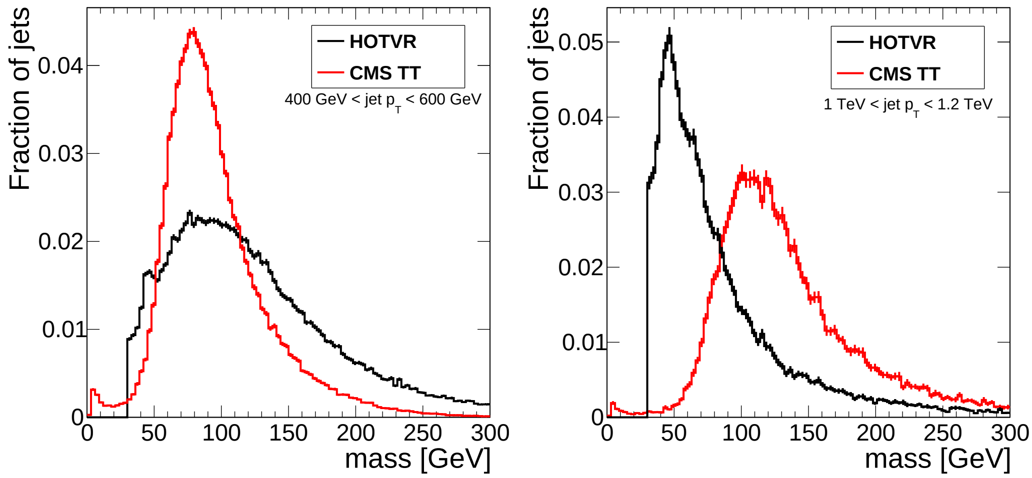


Figure 11.24: Jet mass of the HOTVR and the CMS top tagger for different p_{T} regions and background events.

and the HOTVR are shown with the modified efficiency and mistag rate definition. The ROC curves are obtained by scanning the parameters shown in Table 11.5. For all three tagging algorithms all decision variables are scanned. In addition the number of subjets is forced to be greater than two for all top tagging algorithms, in order to avoid contributions from W bosons from partially merged jets. For the HOTVR a selection on the transverse momentum of the first two leading jets is applied. The OptimalR and the CMS top

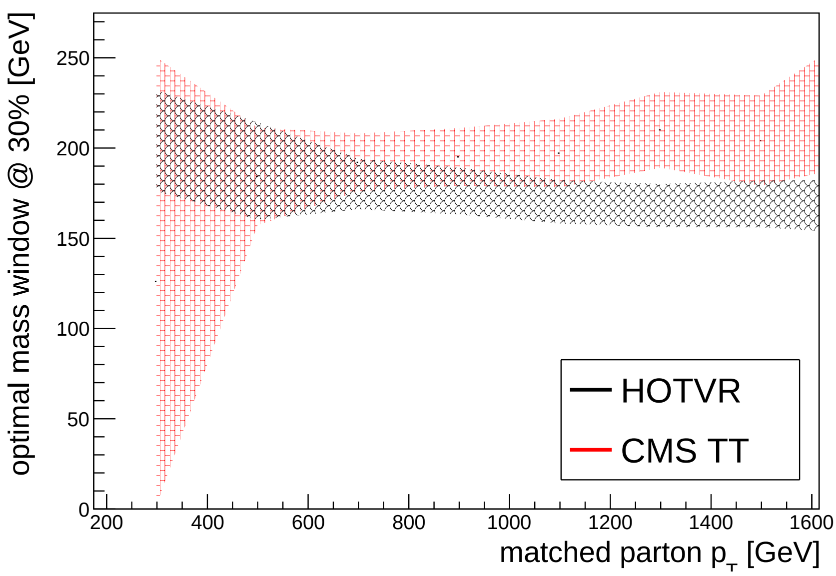


Figure 11.25: Optimal mass window for an efficiency of 30%, shown for the HOTVR and the CMS top tagger as a function of the parton p_{T}.

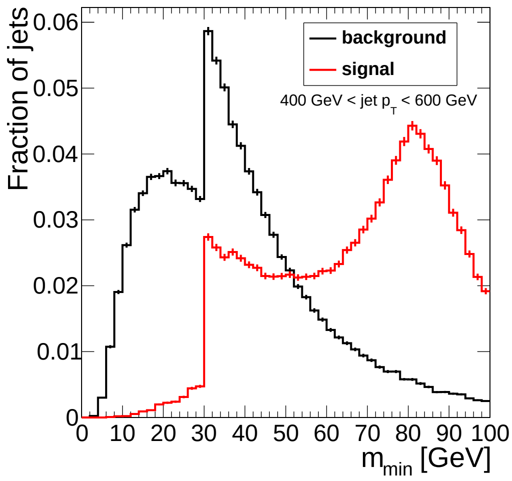


Figure 11.26: The minimum pairwise mass m_{min} for the HOTVR for signal and background events.

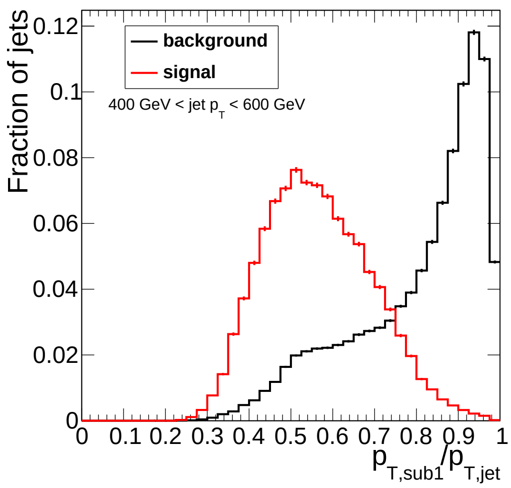


Figure 11.27: The p_{T} fraction of the leading subjet to the large jet for the HOTVR for signal and background events.

tagger contain similar selections in their subjet algorithms. The variable scan is done with TMVA [191] and the method "Cuts", which finds rectangular selections by a fitting method. This fit is done for each p_{T} selection, consequently the selections on the decision variables changes for each p_{T} selection.

Since for all tagging algorithms fixed selections are applied (see Table 11.5), the ROC curves after using TMVA are scaled by the signal efficiencies and background efficiencies of these selections, e.g. requiring the number of subjets $N_{\text{sub}} > 2$ results in different efficiencies and mistag rates for each algorithm. TMVA does not consider initial selections and the returned ROC curves have to be corrected. In Table 11.6 the scale factors for all p_{T} selections are shown.

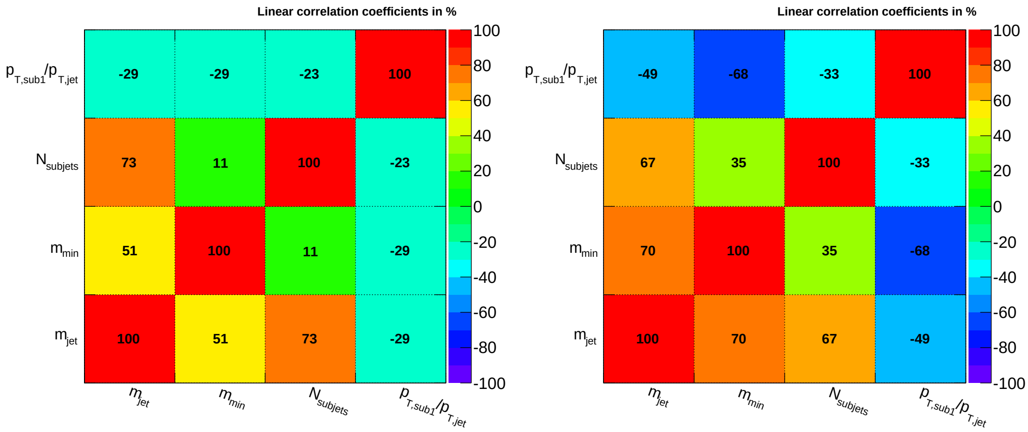


Figure 11.28: Linear correlation between the four decision variables for signal (left) and background (right) events.

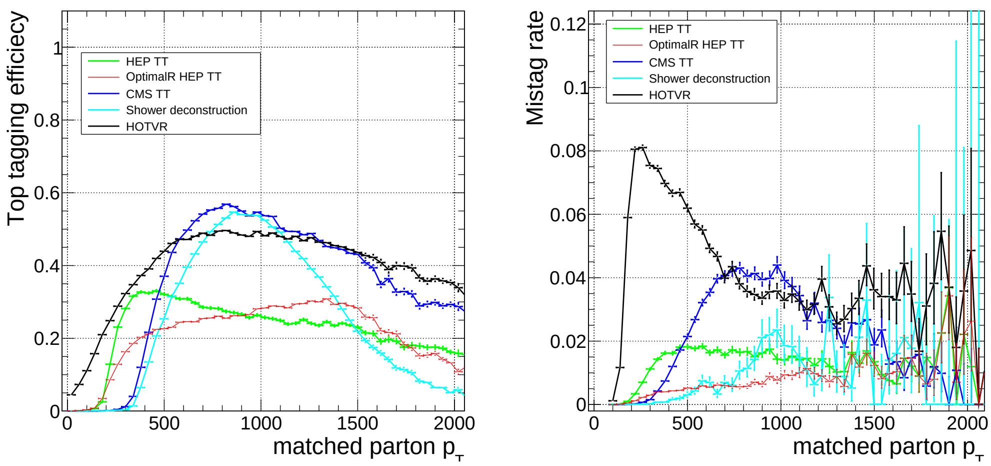


Figure 11.29: Efficiency (left) and mistag rate (right) as a function of the generator p_{T} for different top tagging algorithms.

The p_T distribution is reweighted to be approximately flat in order to give the same importance to all p_T regions.

The resulting ROC curves are shown for different p_T selections. In the low p_T region ($200\,\text{GeV} < p_T < 600\,\text{GeV}$) the OptimalR HEP top tagger is the best performing algorithm. The CMS top tagger is not working well in this region due to unmerged jets. The HOTVR performs comparable to the OptimalR HEP top tagger. In the intermediate p_T region ($600\,\text{GeV} < p_T < 1000\,\text{GeV}$) all tagging algorithms have approximately the same performance, which was already observed for the CMS top tagger and the OptimalR HEP top tagger in Ref. [210]. Even in the high p_T region ($1000\,\text{GeV} < p_T < 1200\,\text{GeV}$) all three algorithms have almost the same performance and no algorithm performs significantly better than the other ones.

In summary, the HOTVR has a reasonable good performance in the low p_T region and a similar performance in all other p_T regions. This is a remarkable result for an algorithm that does not include filtering procedures as e.g. the OptimalR HEP top tagger.

In Figure 11.31 the tagging algorithms are kept at their standard working points and the N-subjettiness ratio τ_3/τ_2 is scanned, as the taggers are often used in combination with N-subjettiness variables. The shower deconstruction tagger is used with a fixed microjet conesize of 0.2 and a fixed large jet distance parameter of 0.8. Only events with $\chi > 3.5$ are selected. The HOTVR is shown here in two versions ($\rho = 600\,\text{GeV}$; $\rho = \infty$). The value of $\rho = \infty$ corresponds to the HOTVR without cone shrinking and therefore to a fixed distance parameter of 1.5.

In the low p_T region ($200\,\text{GeV} < p_T < 400\,\text{GeV}$) the OptimalR HEP top tagger and the HEP top tagger are the best performing algorithms. The CMS top tagger and the shower deconstruction tagger are not capable to work in this region due to the small distance parameter of the large jet. The two versions of the HOTVR perform equally good in this region, which indicates that even with $\rho = 600\,\text{GeV}$ the distance parameter is kept at 1.5. The performance of the HOTVR is good and it can be used in this region as well.

In the p_T region from 400 GeV to 600 GeV the performance of the HOTVR combined with N-subjettiness increases significantly. Together with the CMS top tagger it has the best performance for high efficiencies. For low efficiencies ($\epsilon < 0.1$) shower deconstruction is the best performing algorithm. The performance of the OptimalR HEP top tagger and the HEP top tagger is worse in this region. The combination with N-subjettiness is improving the signal to background ratio of the other tagging algorithms but not the one of the HEP top tagger. Previous studies showed, that the linear correlations are higher

Tagger	Scanned variable	fixed selections
CMS top tagger	m_{jet}	$N_{\text{sub}} > 2$
	m_{min}	
OptimalR HEP top tagger	$m_{123}(R = R_{\text{min}})$	$N_{\text{sub}} > 2$
	$f_W(R = R_{\text{min}})$	
	$R_{\text{min}} - R_{\text{min,exp}}$	
HOTVR	m_{jet}	$N_{\text{sub}} > 2$
	m_{min}	$p_{\text{T,sub1}>20\,\text{GeV}}$
	$p_{\text{T,frac}}$	$p_{\text{T,sub2}>20\,\text{GeV}}$

Table 11.5: Scanned variables and fixed selections for the different top tagging algorithms.

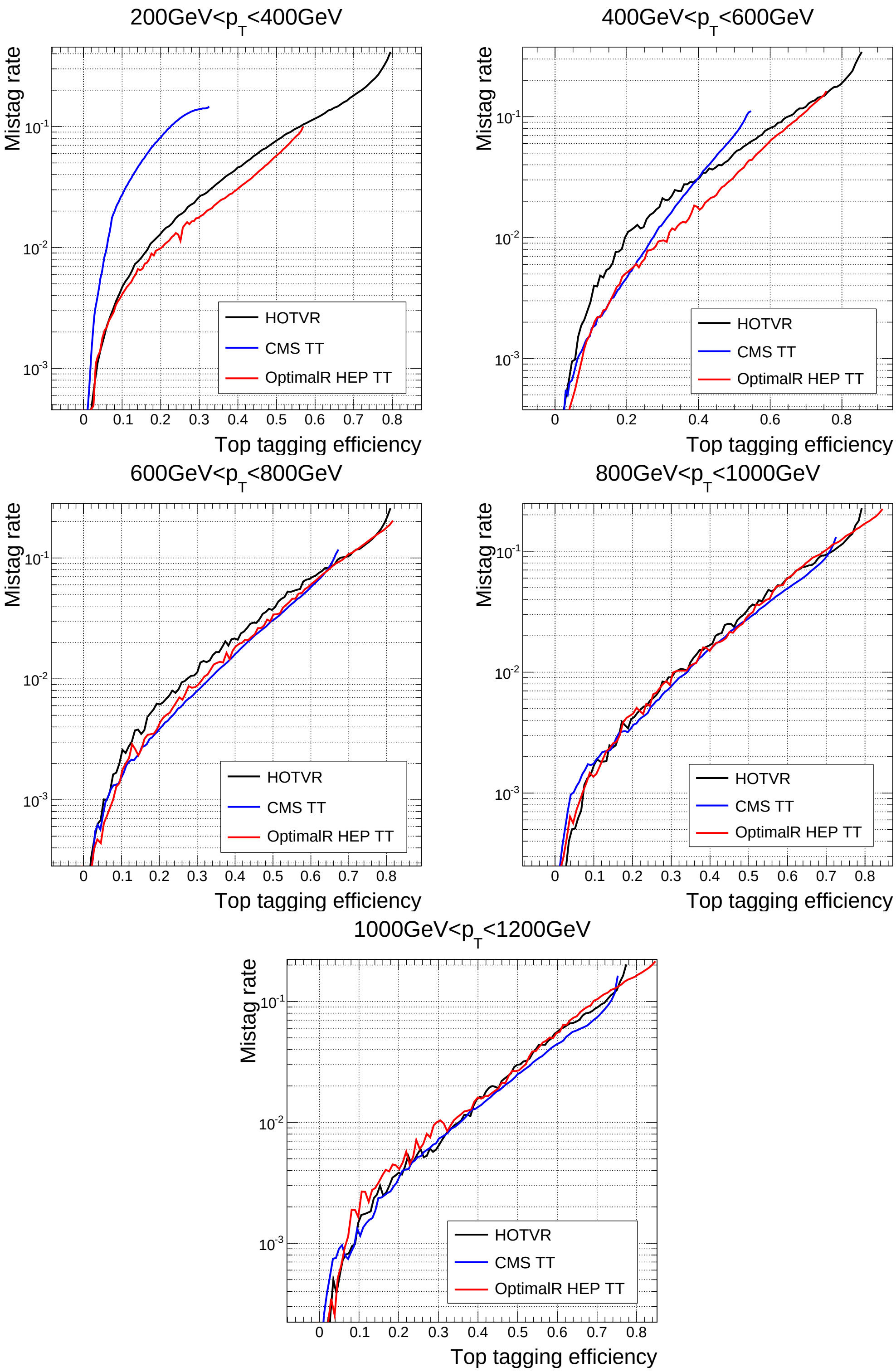


Figure 11.30: ROC curves after scanning all decision variables with TMVA for different p_{T} selections.

Algorithm	200-400 GeV	400-600 GeV	600-800 GeV	800-1000 GeV	1000-1200 GeV
CMS TT					
Signal	0.33	0.56	0.68	0.74	0.76
Background	0.15	0.11	0.12	0.14	0.23
OptimalR					
Signal	0.58	0.77	0.83	0.86	0.86
Background	0.15	0.25	0.30	0.32	0.31
HOTVR					
Signal	0.81	0.87	0.82	0.80	0.79
Background	0.59	0.58	0.45	0.38	0.34

Table 11.6: Scale factors to correct the ROC curves obtained with TMVA.

between N-subjettiness and OptimalR HEP top tagger or the HEP top tagger [167].

For a p_T selection between 600 GeV to 800 GeV the performance of the HOTVR is similar to the performance of the CMS top tagger, whereas shower deconstruction shows the best performance. In this p_T region a significant difference between the HOTVR with $\rho = 600\,\text{GeV}$ and $\rho = \infty$ is visible, which shows the importance of the cone shrinking.

For very high transverse momenta above 1 TeV the HOTVR combined with N-subjettiness obtains a better performance than all other algorithms in the whole efficiency range. One should note that the shower deconstruction tagger is performed in this analysis with a fixed microjet cone radius. One might improve the performance by using smaller conesizes of 0.1.

In summary, the HOTVR at the standard working point combined with N-subjettiness shows a great performance over the whole p_T range and outperforms in the highest p_T region other existing top tagging algorithms at their standard working points combined with N-subjettiness. This is a remarkable result for an algorithm that uses e.g. no filtering or other grooming techniques.

11.3.4 Validation in 8 TeV data

The validation is done with the data collected at a centre-of-mass energy of 8 TeV in 2012 with an integrated luminosity of $19.7\,\text{fb}^{-1}$ by the CMS detector. In order to achieve a pure $t\bar{t}$ sample the signal selection explained in Section 10.1 is used. The HOTVR jet on the other hemisphere ($\Delta\phi > 2.1$) of the leptonically decaying top quark is considered as a probe jet. In case there is more than one jet on the other hemisphere the jet that has the largest distance in the ϕ-η plane to the leptonically decaying top quark becomes the probe jet.

For the background selection, the HOTVR jet that is closest to the probe jet from the former selection (see Section 10.2) becomes the probe jet to achieve a comparable selection. The leading jet p_T distribution of the HOTVR is not reweighted again and the weights derived in Section 10.2.2 are used. In this analysis no jet energy corrections are used for the jets found by the HOTVR. Applying correct JECs is more challenging since jets with distance parameters between 0.1 and 1.5 can appear. As mentioned in Section 6.3.3 JECs are only available for AK5 and AK7 jets. It turns out that the appearance of different jet distance parameters is a disadvantage in the sense that for each distance parameter a

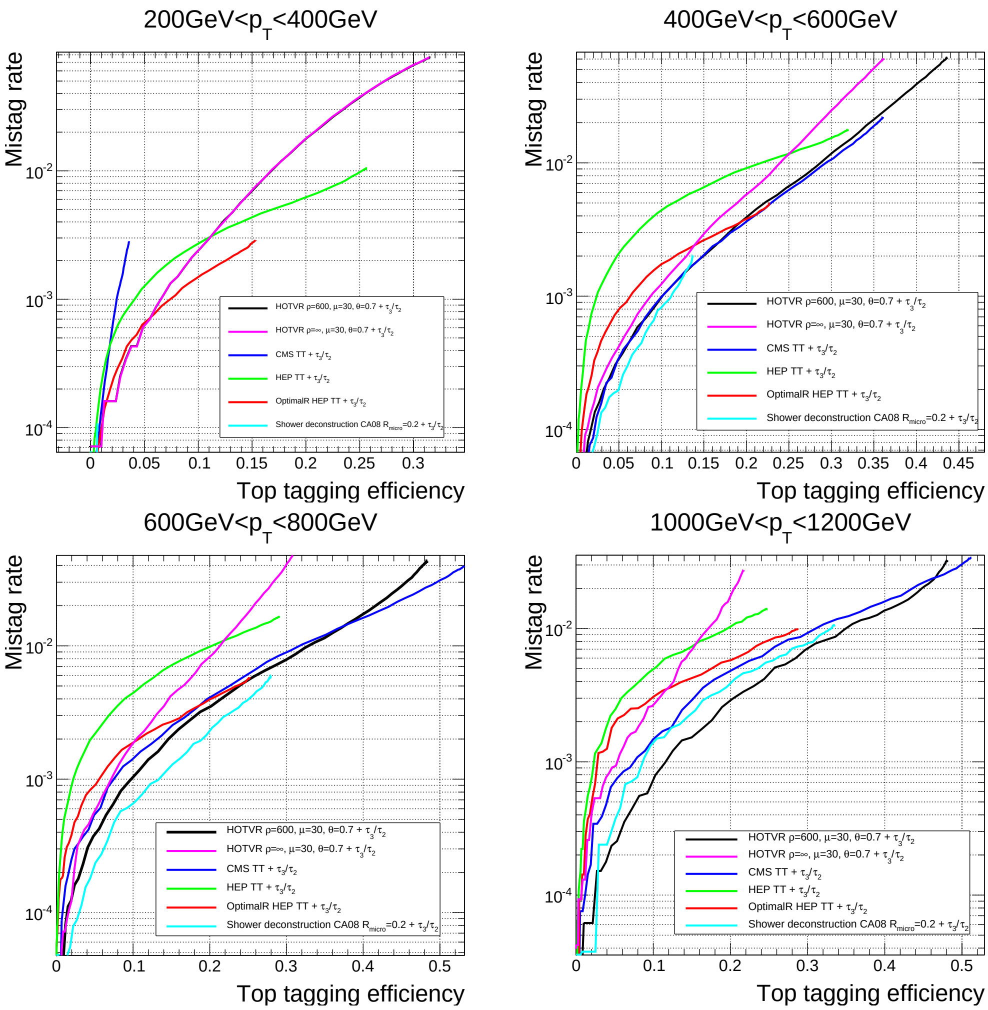


Figure 11.31: ROC curves for the different top tagging variables at their standard working point, the scanned variable is N-subjettiness for all algorithms.

JEC is needed or the JECs have to be interpolated.

In the following all substructure variables of the HOTVR are shown (Figure 11.32 - Figure 11.34) for p_T >400 GeV. Overall the agreement between data and MC is good. In the background selection just the PYTHIA event generator is used and is not compared to the HERWIG++ event generator here.

The mass distribution (Figure 11.32 top) shows for signal events a peak at the top quark mass and a smoothly falling distribution for $m_{\text{jet}} > 120$ GeV for the background selection. The separation power between signal and background increases with increasing transverse momentum. The distribution shows for signal and background events a good agreement between data and MC.

As expected the minimum pairwise mass distribution (Figure 11.32 middle) shows a resonance at the W boson mass for signal events and none for background events. The agreement of data and MC for this variable is better than the minimum pairwise mass calculated from the subjets found by the CMS top tagger (compare Figure 10.15). In the background selection the distribution is not well described for masses below 50 GeV and high masses above 150 GeV.

The number of subjets (Figure 11.32 bottom) is better described by PYTHIA for the HOTVR than for the CMS top tagger (compare Figure 10.16). All jets with only one subjet have been rejected earlier, therefore the separation power is even higher than the distribution shows.

The p_T fraction shows good agreement between data and MC in the signal selection (Figure 11.33 top) and a fair agreement for the background selection. For high values the simulation predicts less events than observed.

The leading subjet p_T is shown in Figure 11.33 (middle). The average p_T is higher for background events as it is expected due to the selections described in Section 10.1 and 10.2. For signal events the agreement between data and MC is good within the statistical uncertainties. For background events PYTHIA describes the data not well, especially for high transverse momenta.

The mass of the subjets are shown in the following since they are used to calculate the minimum pairwise mass. Discrepancies in the minimum pairwise mass can be traced back to discrepancies in the subjet mass variables. The mass of the leading subjet is shown in Figure 11.33 (bottom). The sharp cutoff at 30 GeV is caused by the mass jump threshold, since the clusters “grow” until the threshold is reached and afterwards the algorithm considers the mass jump criterion. In the case a mass jump appears the cluster is saved as a subjet with a mass below 30 GeV. The simulation for signal events predicts less jets than observed. The shape of the distribution is modeled well. The background simulation agrees well with the data.

Figure 11.34 (top) shows the transverse momentum of the second leading subjet. The signal selection describes the data well whereas the background selection has some discrepancies for high transverse momenta.

The mass of the second leading subjet is shown in Figure 11.34 (bottom). Here the agreement between data and MC is good for the background selection, for the signal selection the shape of the variable is not well described and in the ratio a slope is obtained.

The efficiency and mistag rate of the HOTVR are shown in Figure 11.35. The overall agreement with data is good. Among all top tagging algorithms that are tested, the HOTVR is the only tagging algorithm that reaches a plateau in the efficiency. All other tagging algorithms had a decreasing efficiency with increasing transverse momentum. At the standard working point the HOTVR reaches an efficiency up to 40%. It can be used in the whole p_T range since the efficiency increases already drastically around 200 GeV and reaches a plateau at 250 GeV. However, the mistag rate is about 8% for the low p_T range, but it decreases with increasing p_T. In order to improve the tagging algorithm in the low p_T range, one can use in addition an N-subjettiness selection, which is not shown here. The integrated efficiency for $p_T > 350$ GeV is 32% and the corresponding integrated mistag rate for $p_T > 350$ GeV is 5.7%.

The efficiency and mistag rate as a function of the number of primary vertices N_{pv} is shown in Figure 11.36. Here signal events are independent of N_{pv}, whereas background events show a slight dependence. The agreement of the simulation with data is fair for both selections.

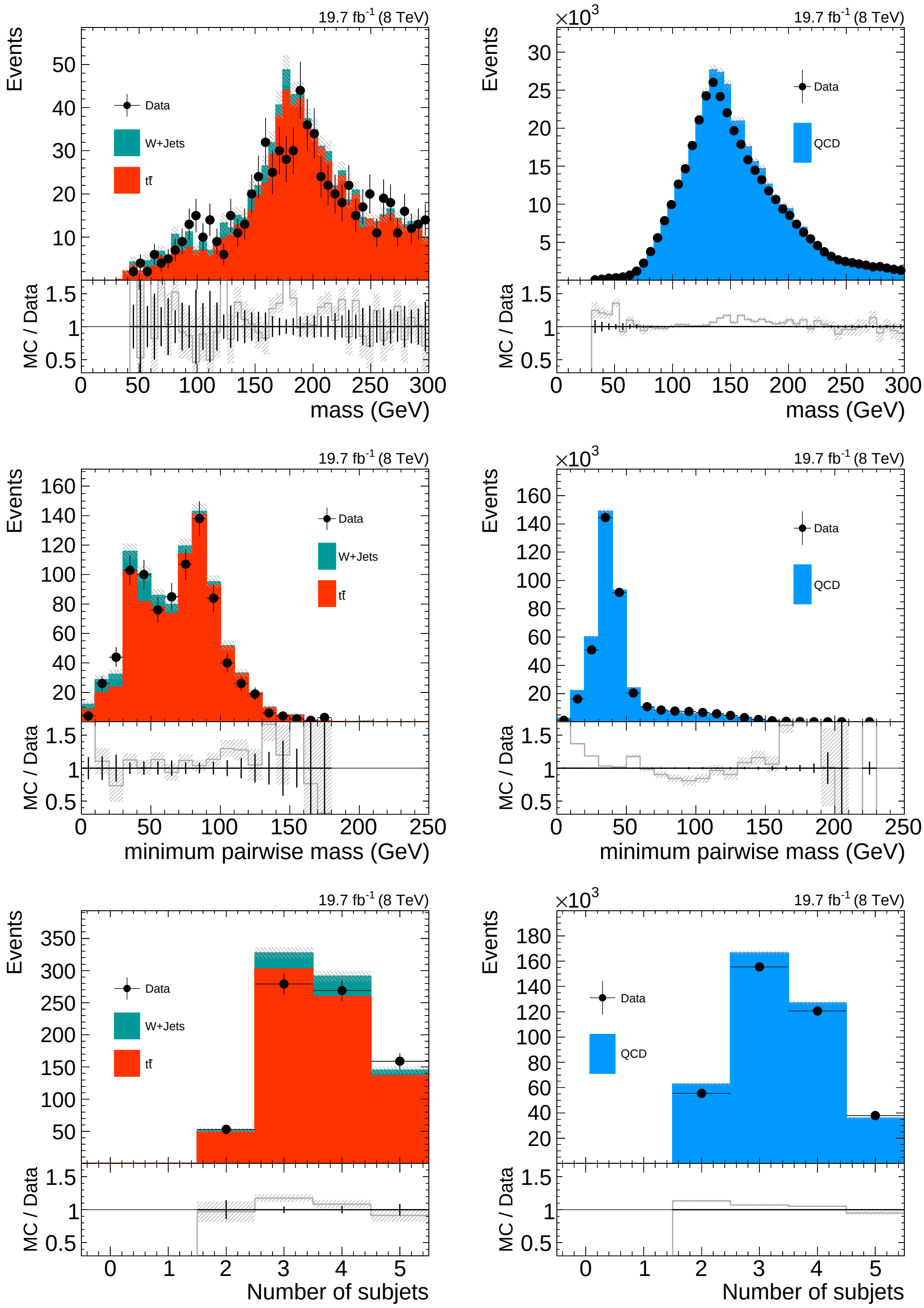


Figure 11.32: Jet mass (top), minimum pairwise mass (middle) and number of subjets (bottom) of the HOTVR for signal (left) and background events (right). Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom panel. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

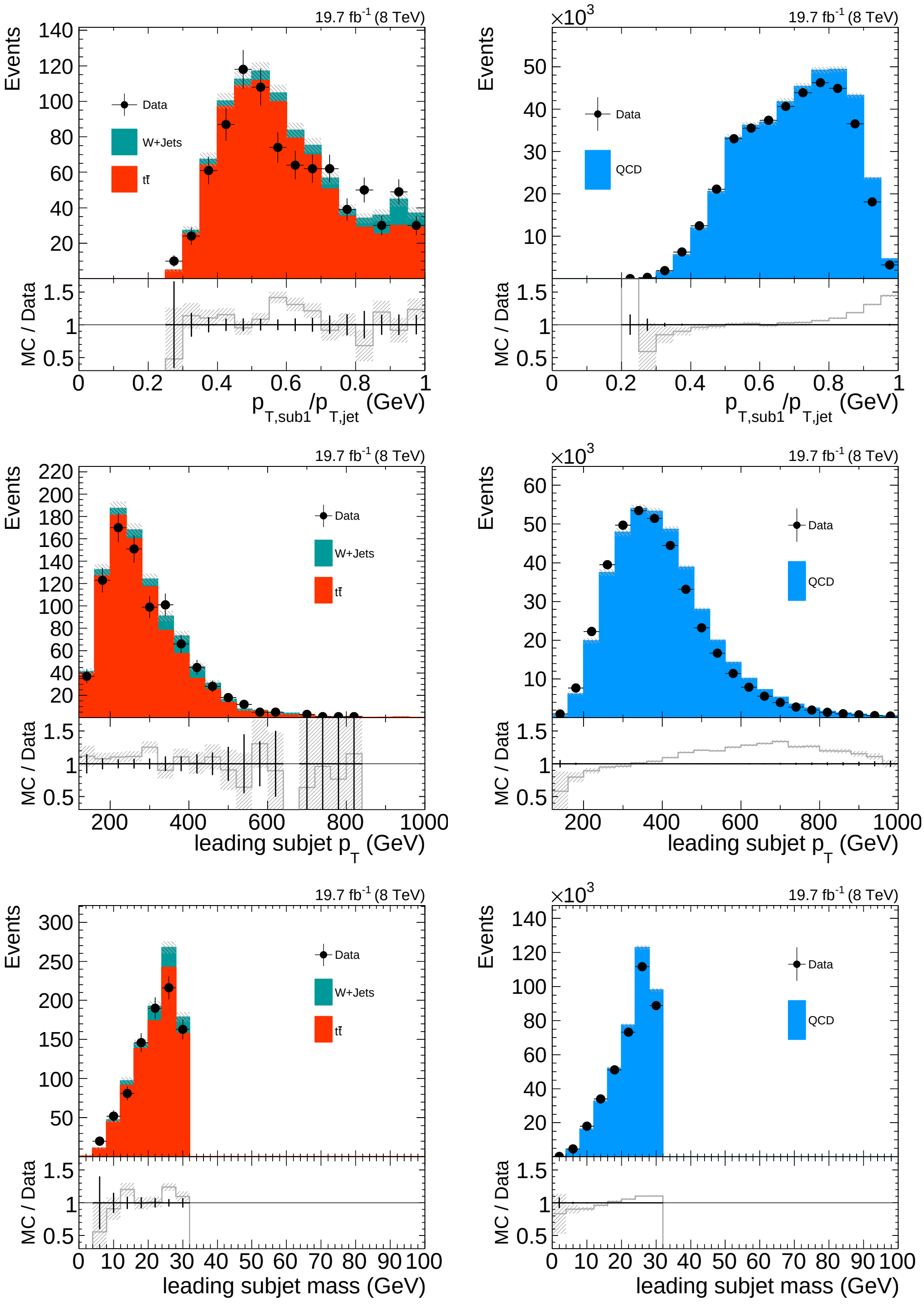


Figure 11.33: p_{T} fraction (top), leading subjet p_{T} (middle) and leading subjet mass (bottom) of the HOTVR for signal (left) and background events (right). Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom panel. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

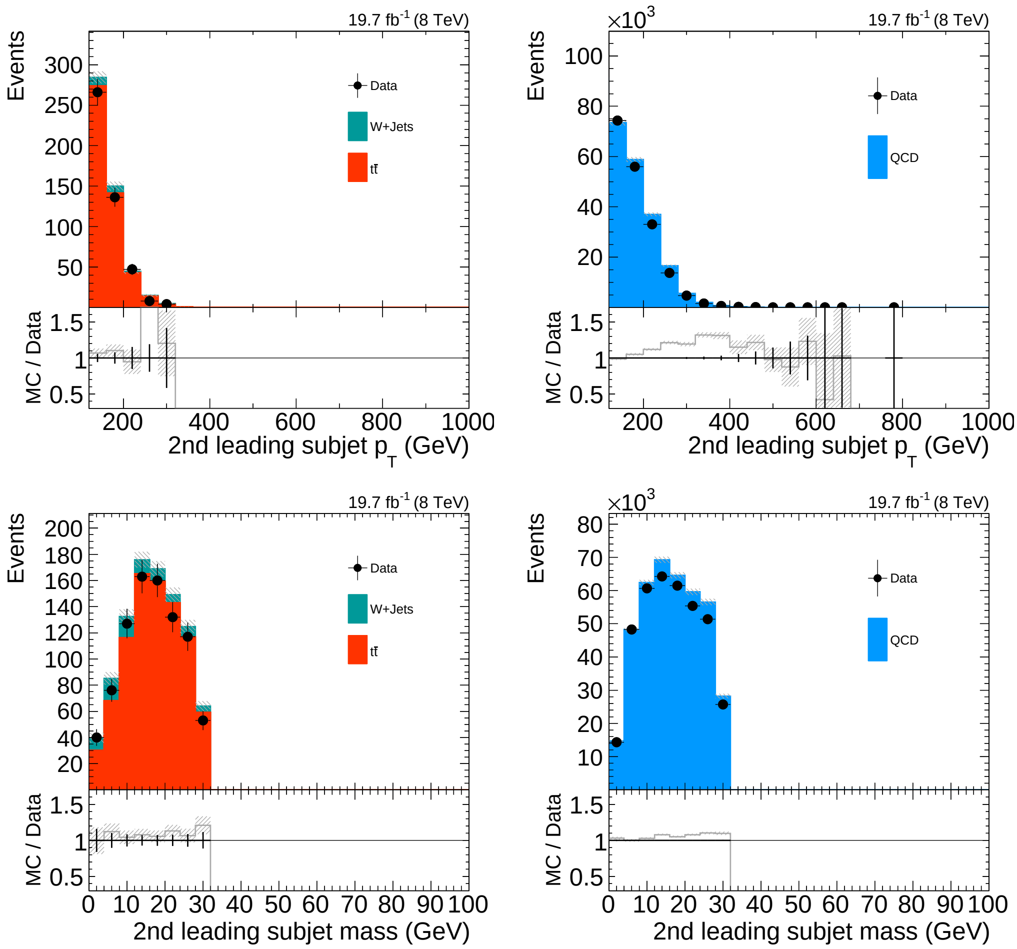


Figure 11.34: Second leading subjet p_{T} (top) and second leading mass (bottom) of the HOTVR for signal (left) and background events (right). Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom panel. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

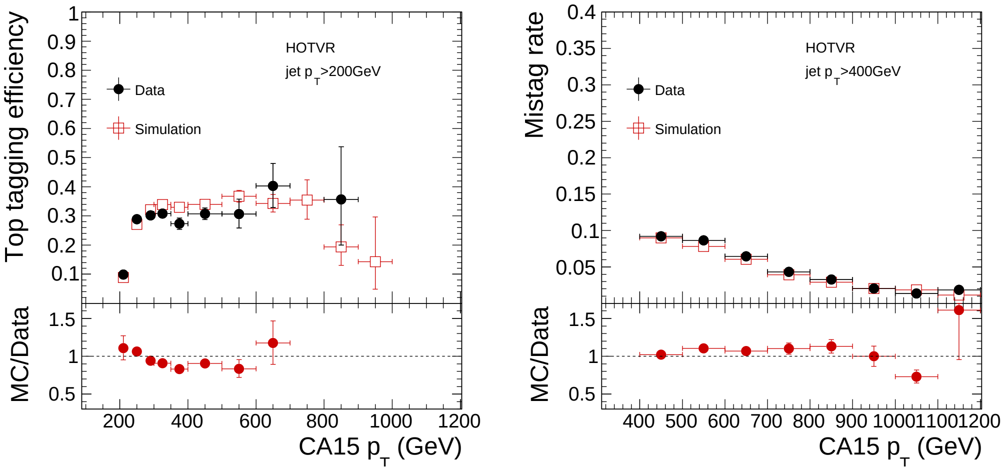


Figure 11.35: Efficiency and mistag rate as a function of p_{T} for the HOTVR in data and simulation. The efficiency reaches a plateau. The error bars show the combined statistical and systematic uncertainties. At the bottom of each panel the ratio of simulation to data is shown.

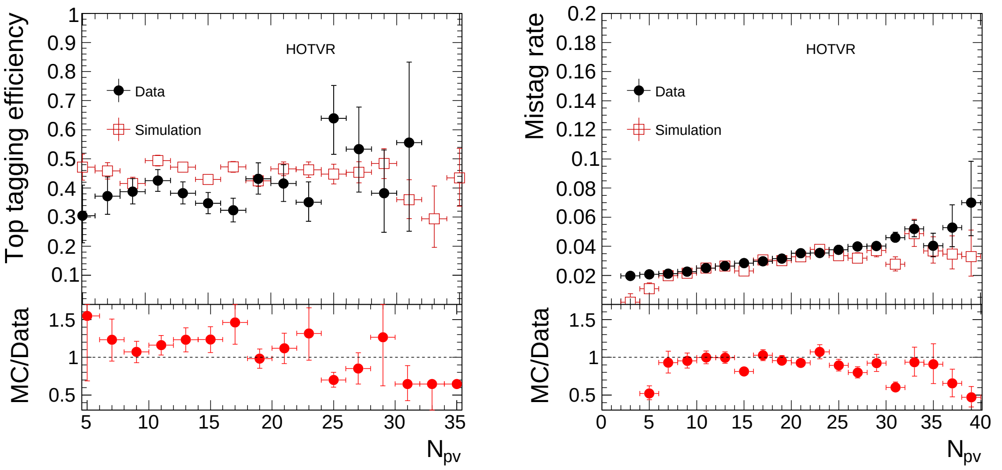


Figure 11.36: Efficiency and mistag rate as a funcion of N_{pv} for the HOTVR in data and simulation. The error bars show the combined statistical and systematic uncertainties. At the bottom of each panel the ratio of simulation to data is shown.

12 Conclusion and Outlook

In the search for new physics the identification of top quarks has major importance, since new heavy particles are predicted to couple strongly to top quarks. Many models beyond the standard model predict these heavy resonances. Furthermore, in high precision tests of the standard model the identification of boosted top quarks becomes increasingly important.

In the first part of this thesis, the Phase 1 upgrade of CMS and the x-ray calibration of pixel modules were presented. The increased luminosity makes the exchange of the pixel module readout chip necessary, with extended buffers in order to avoid too high occupancy. Furthermore, by adding a fourth pixel barrel layer to the existing three layers the b identification can be improved significantly. In this thesis the method to calibrate the Vcal circuit in the readout chip was presented. It was shown that the x-ray calibration of the new digital readout chips is independent of temperature and therefore the calibration for mass production can be done at room temperature. It was also shown that the stability of the calibration procedure can be significantly improved by adjusting the rate achieved by the readout chip for each fluorescence target. The highest stability was achieved by performing the calibration procedure with four instead of three fluorescence targets. By stabilizing the calibration method the standard deviation was reduced from 2% to 0.6%.

In the second part of this thesis current top tagging algorithms were compared with new approaches. The OptimalR HEP top tagger performs on average a factor two better than the original HEP top tagger, especially in the high p_{T} region. Furthermore, the performance of the shower deconstruction tagger was studied. Combined with an additional subjet b tag the algorithm shows the best performance in the high p_{T} region. Moreover, a BDT trained with ten substructure variables was introduced in this thesis. The BDT is the best performing algorithm for $p_{\mathrm{T}} > 400\,\mathrm{GeV}$ which does not use subjet b tagging information. The performance of all top tagging algorithms with new approaches is much better than the currently used ones in CMS. Especially with focus on Run II of the LHC and the increased centre-of-mass energy these are interesting and important results.

The new tagging algorithms were furthermore validated with the data set collected in 2012 with an integrated luminosity of $19.7\,\mathrm{fb}^{-1}$ and a centre-of-mass energy of $8\,\mathrm{TeV}$. A measurement of the top tagging efficiency and the mistag rate was performed with this data set. The overall agreement between data and simulated events was good for the signal selection. For background events the two event generators PYTHIA6 and HERWIG++ were compared with data. Some deficiencies of the detailed modeling of substructure variables became apparent for the background selection. While the misidentification rate is described well by HERWIG++ for all algorithms considered, the misidentification rate obtained by PYTHIA6 tends to be about 0.2-0.5% too small. Differences in the ratio of data to simulation for the efficiency and mistag rate measurements are small and never larger than for the standard taggers used in CMS. In summary, the performance measurement of the new algorithms in data show the usability of these in future analyses.

In the last part of this thesis, a novel top tagging approach was introduced. The HOTVR uses a variable R approach in order to minimize contributions from additional radiation to the jet and to take into account the collimation of the top quark decay products with increasing transverse momenta. Furthermore, it uses a mass jump criterion in order to identify subjets already during the clustering procedure. This makes the algorithm fast in comparison to other top tagging methods, like the shower deconstruction algorithm. It was shown that the HOTVR is infrared and collinear safe, making it useful for fixed-order pertubative QCD. The parameters of the algorithm were optimized to achieve an optimal efficiency while rejecting a maximum of background events. The performance of the algorithm was compared to other top tagging approaches and it was shown that the algorithm provides a good performance for low transverse momenta as well as for very high transverse momenta, it is therefore usable in a wide p_{T} range. Combined with a selection on N-subjettiness, the algorithm performs better than existing top tagging algorithms in the high p_{T} region, making it especially interesting for Run II with increased centre-of-mass energy. Furthermore, the algorithm was tested on generator level to present an experiment-independent comparison of the tagging algorithms. The performance on generator level is similar to the one on reconstruction level and can therefore be used also in other high energy experiments. The algorithm was validated in 8 TeV data of the CMS detector and the overall agreement between simulated events and data is good. By modifying the parameters and the selections on the decision variables the HOTVR algorithm could also be used to identify W, Z and Higgs bosons, as well as other heavy, hadronically decaying BSM particles.

A Additional material for the 8 TeV data/MC comparison

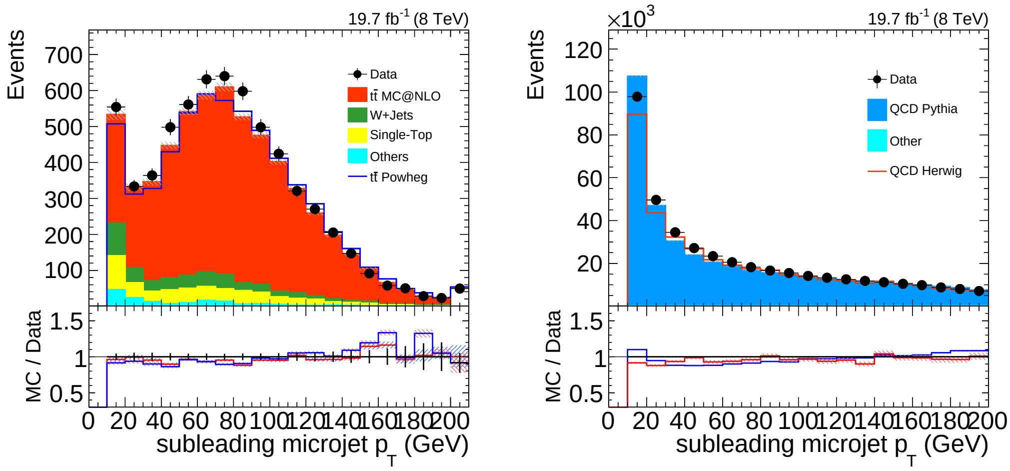


Figure A.1: p_{T} distribution of subleading microjet. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

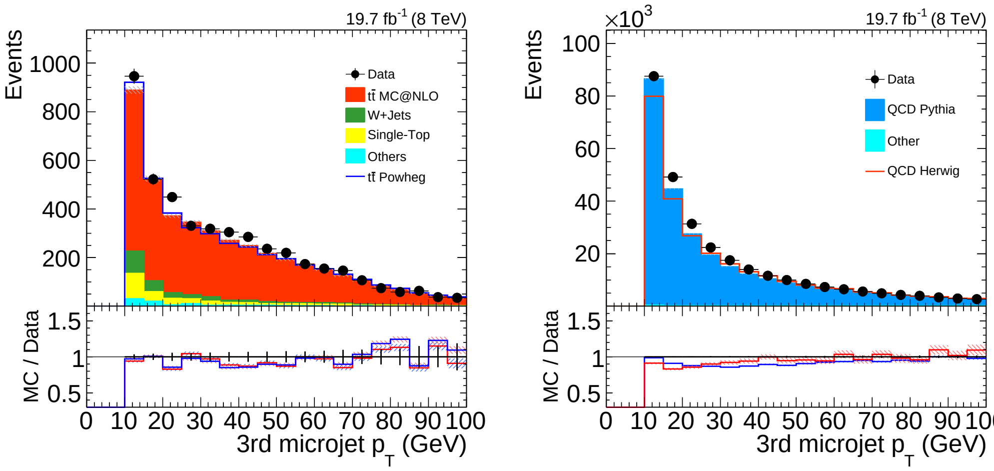


Figure A.2: p_T distribution of 3$^{\mathrm{rd}}$ leading microjet. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

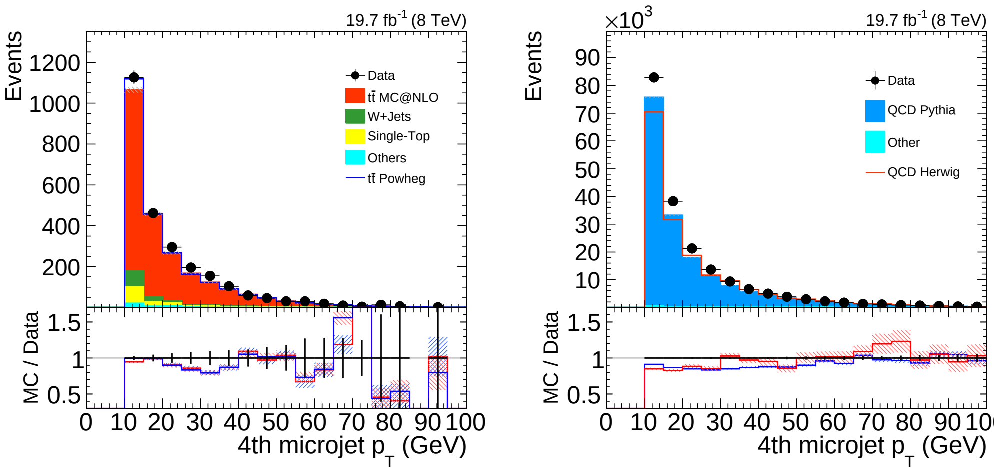


Figure A.3: p_T distribution of 4$^{\mathrm{th}}$ leading microjet. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

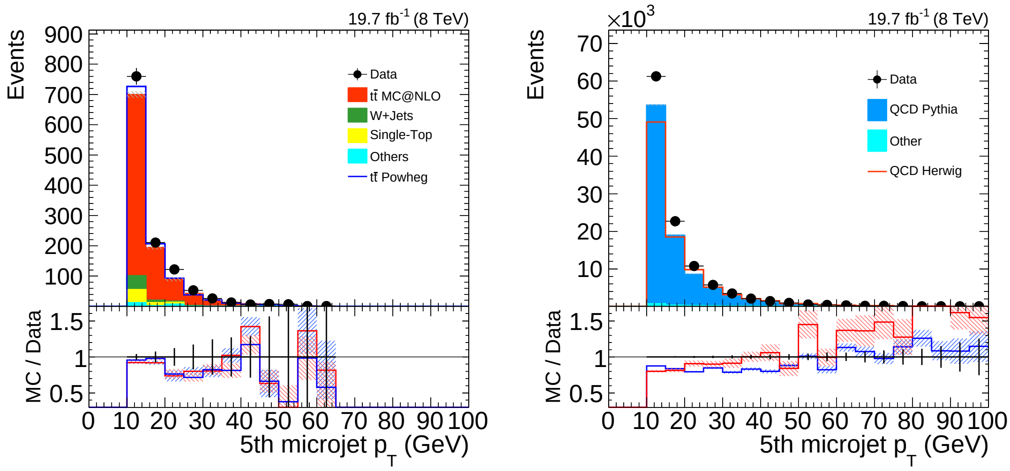


Figure A.4: p_T distribution of 5$^{\text{th}}$ leading microjet. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data.

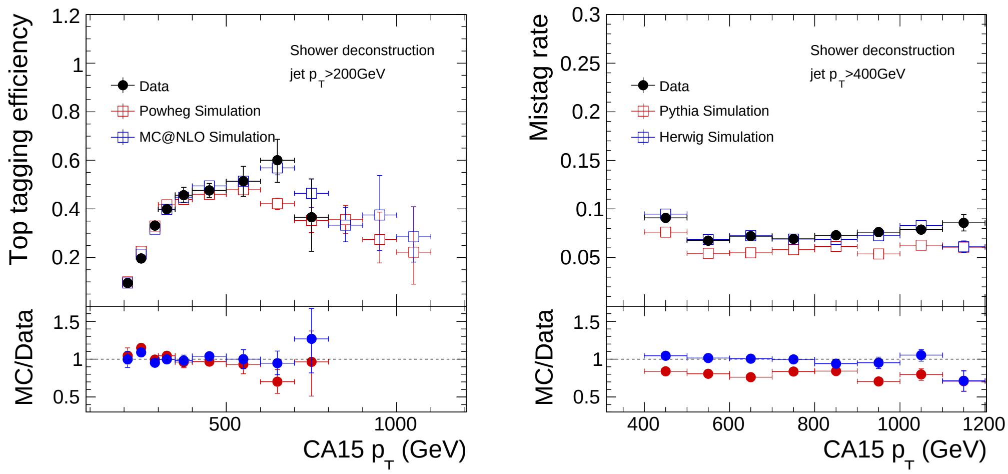


Figure A.5: Top tagging efficiency of the shower deconstruction tagger at WP0 ($\chi > 2$) (left) and mistag rate (right), as function of p_T. The error bars show the combined statistical and systematic uncertainties. At the bottom of each panel the ratio of simulation to data is shown.

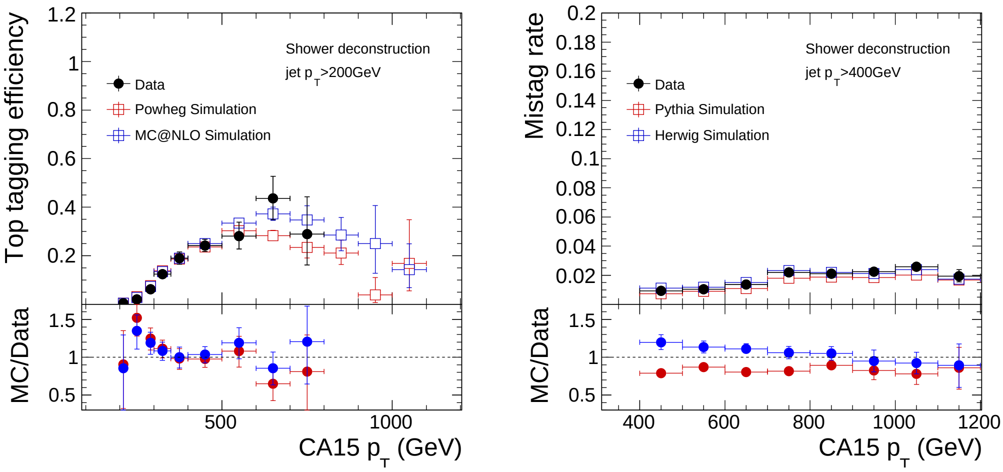


Figure A.6: Top tagging efficiency of the shower deconstruction tagger at WP2 ($\chi > 5$) (left) and mistag rate (right), as function of p_{T}. The error bars show the combined statistical and systematic uncertainties. At the bottom of each panel the ratio of simulation to data is shown.

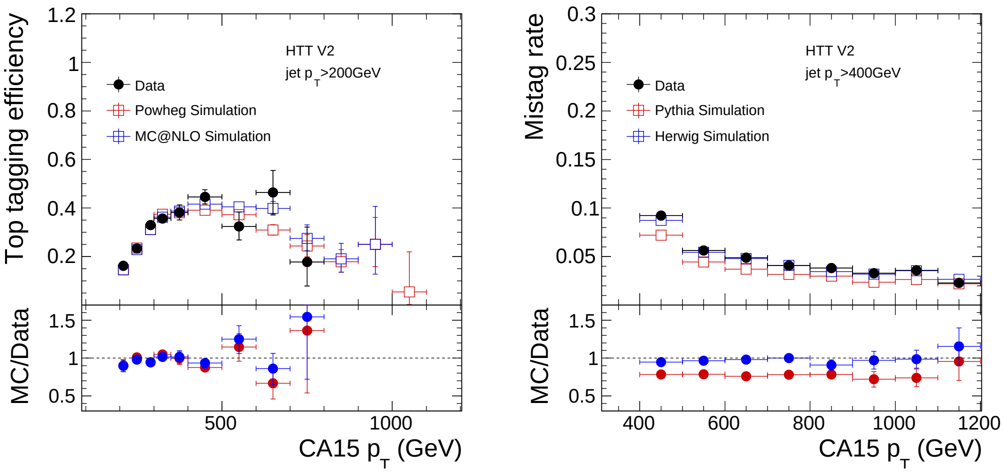


Figure A.7: Measurement of the top tagging efficiency of the OptimalR HEP top tagger WP0 ($R_{\min} - R_{\min,\exp} > 0.5$, no additonal cut on the candidate mass is apllied) (left) and the mistag rate (right) as function of p_{T}. The error bars show the combined statistical and systematic uncertainties. At the bottom of each panel the ratio of simulation to data is shown.

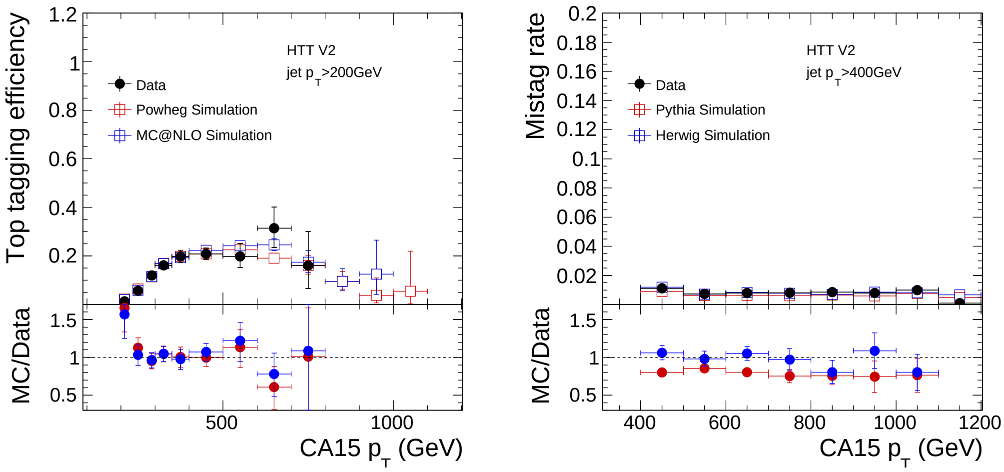


Figure A.8: Measurement of the top tagging efficiency of the OptimalR HEP top tagger WP2 ($R_{\text{min}} - R_{\text{min,exp}} > 1.0$ and the standard jet mass selection) (left) and the mistag rate (right) as function of p_{T}. The error bars show the combined statistical and systematic uncertainties. At the bottom of each panel the ratio of simulation to data is shown.

B Data/MC comparison with Pythia 8

In the following the results presented in Section 10 are repeated with the event generator PYTHIA 8.1. The p_{T} distribution of the leading jet was not reweighted, since good agreement with the data is observed. Pythia 8 describes the data similar to HERWIG++ and the agreement between data and MC is better with PYTHIA 8 than with PYTHIA 6. The studies using PYTHIA 8 were published in Ref. [196].

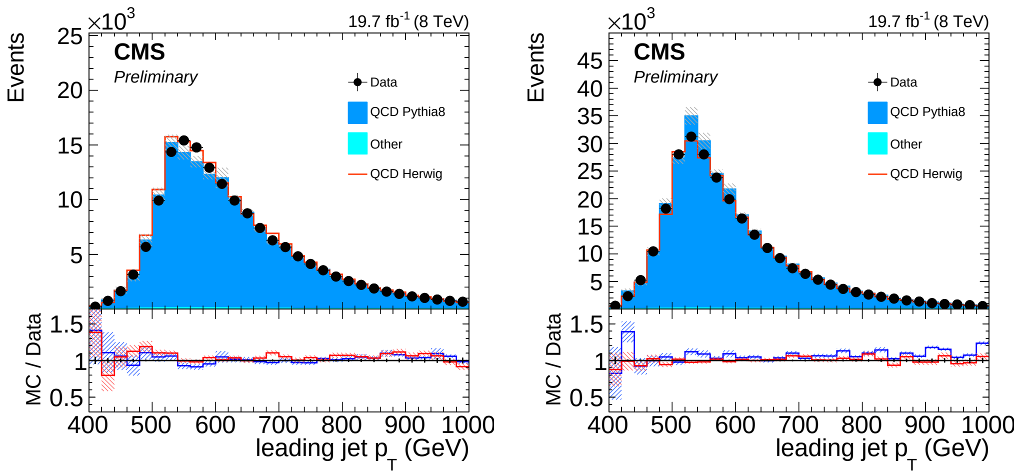


Figure B.1: p_{T} distribution for CA8 (left) and CA15 (right) jets after the background selection. Data is compared to simulated events by PYTHIA 8 and HERWIG++. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA 8) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data. Published in [196].

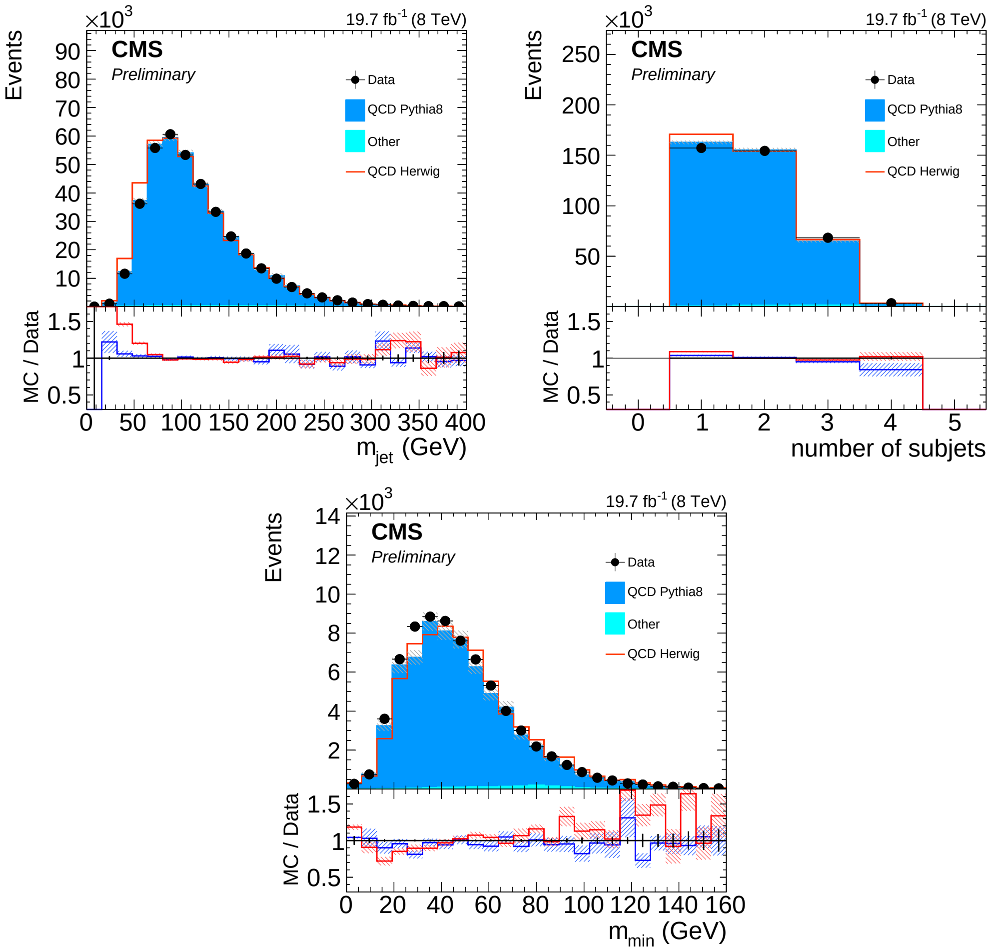


Figure B.2: Jet mass (top left), number of subjets (top right) and the minimum pairwise mass (bottom) for the background selection and reconstructed with the CMS top tagger. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA 8) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data. Published in [196].

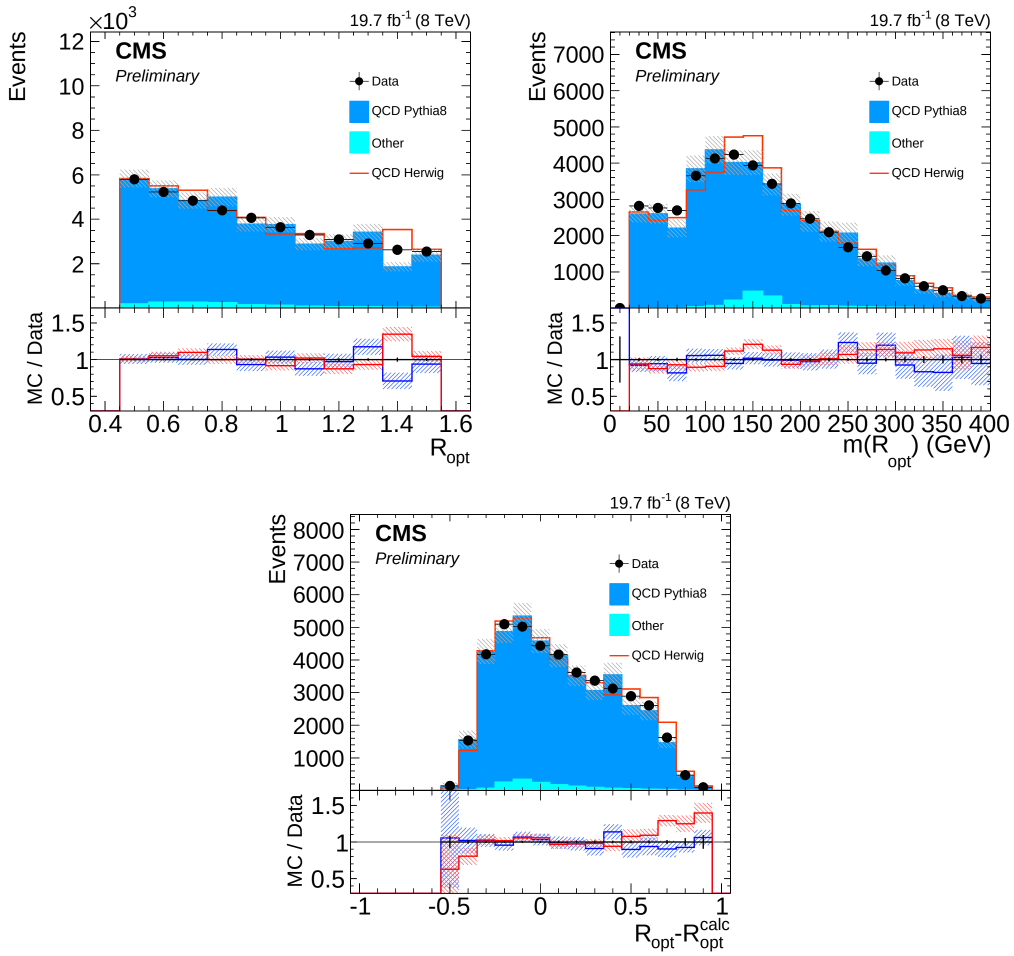


Figure B.3: Distribution of the optimal radius R_{opt} (top left), the mass m_{123} at the optimal radius R_{opt} (top right) and the decision variable for the Optimal R HEP top tagger $R_{\text{opt}} - R_{\text{opt}}^{\text{calc}}$ (bottom) for the background selection. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA 8) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data. Published in [196].

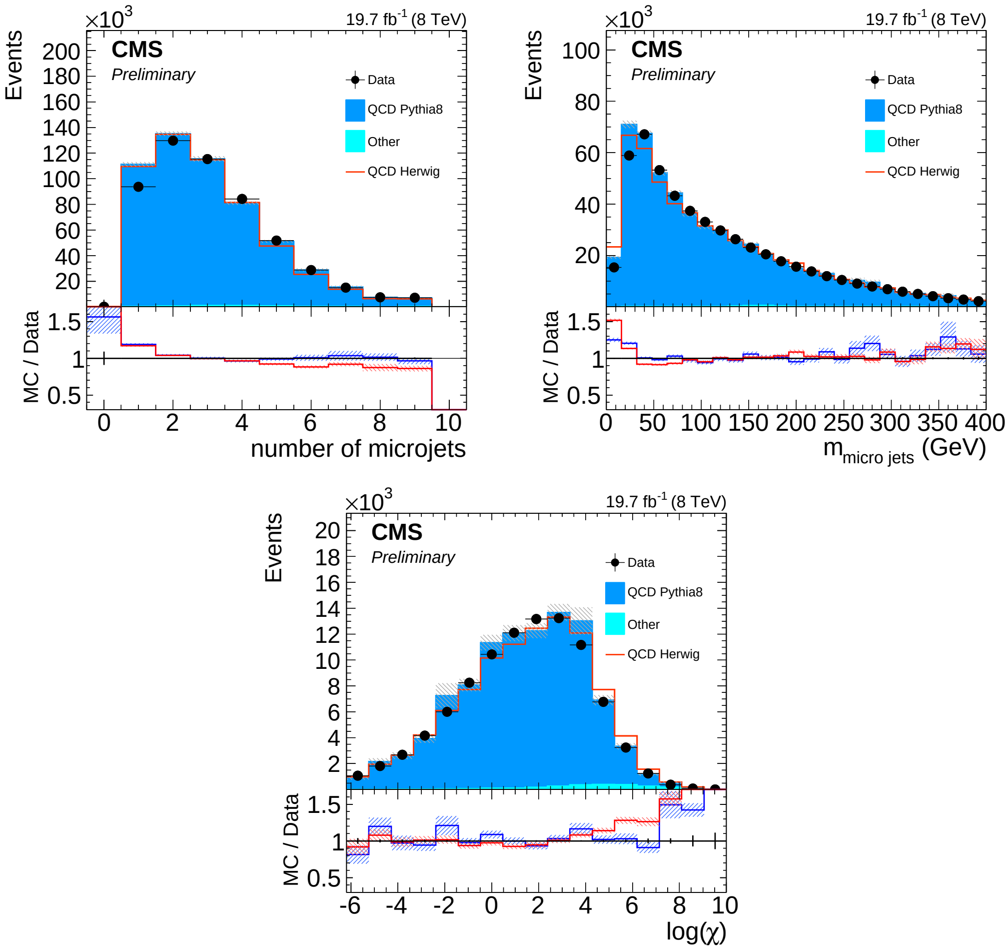


Figure B.4: Distribution of the microjet multiplicity (top left) , invariant mass of all microjets (top right) and the decison variable for shower deconstruction χ (bottom) for background events. Only statistical uncertainties are shown. The ratio of simulation to data is shown at the bottom of each panel. In the case of the signal (background) selection, the ratio to POWHEG (PYTHIA 8) is shown in blue while the one to MC@NLO (HERWIG++) is shown in red. The hashed bands depict the statistical uncertainty of the simulated samples, whereas the vertical bars show the statistical uncertainties of data. Published in [196].

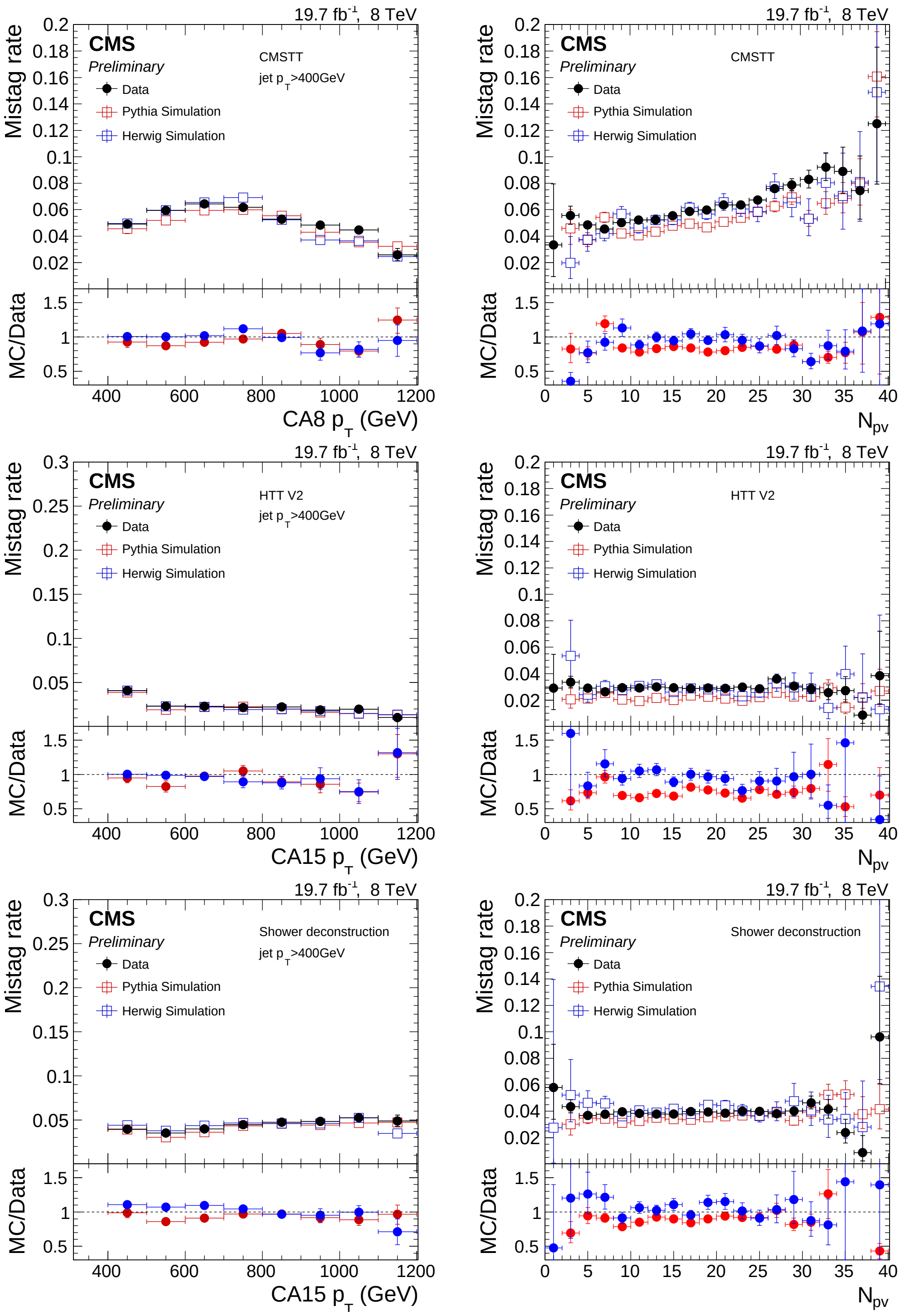


Figure B.5: Mistag rate as the function of p_{T} (right) and as a function of the number of primary vertices (left) for the CMS top tagger (top), the OptimalR HEP top tagger (WP1) (middle) and Shower deconstruction (WP1) (bottom). The error bars show the combined statistical and systematic uncertainties. At the bottom of each panel the ratio of simulation to data is shown. Published in [196].

Bibliography

[1] J. Dalton, “A New System of Chemical Philosophy”, 1808.

[2] J.J. Thomson, “Cathode rays”, *Philosophical Magazine 44 (269): 293* (1887).

[3] E. Rutherford, “The Scattering of and Particles by Matter and the Structure of the Atom”, *Philosophical Magazine 21 (125): 66988* (1911).

[4] N. Bohr, “Niels Bohr, The Nobel Prize in Physics 1922, Nobel Lecture”, 1922.

[5] J. Chadwick, M. Goldhaber, “A nuclear photo-effect: disintegration of the diplon by gamma rays”, *Nature 134: 237238* (1934).

[6] Svante Arrhenius, “Presentation speech for Nobel Prize in Physics”. Nobel speech, 1921.

[7] A. Einstein, “Concerning an heuristic point of view toward the emission and transformation of light”, *Annalen Phys.* **17** (1905) 132–148.

[8] E.D. Bloom et al., “High-Energy Inelastic ep Scattering at 6° and 10°”, *Physical Review Letters 23 (16): 930934* (1969).

[9] UA2 Collaboration Collaboration, “Evidence for $Z0 \to e + e-$ at the CERN $\overline{p}p$ collider”, *Phys. Lett. B* **129** (Aug, 1983) 130–140. 21 p.

[10] UA2 Collaboration, “Observation of Single Isolated Electrons of High Transverse Momentum in Events with Missing Transverse Energy at the CERN anti-p p Collider”, *Phys. Lett.* **B122** (1983) 476–485. doi:10.1016/0370-2693(83)91605-2.

[11] CDF Collaboration, “Observation of Top Quark Production in pp Collisions with the Collider Detector at Fermilab”, *Phys.Rev.Lett.74:2626-2631* (1995) arXiv:9503002. doi:10.1103/PhysRevLett.74.2626.

[12] G. ’t Hooft, “A confrontation with inifinity”, *Nobel lectures* (1999).

[13] F. Wilczek, “Asymptotic Freedom: From Paradox to Paradigm”, *Nobel lectures* (2004).

[14] M. Kobayashi, “CP Violation and Flavour Mixing”, *Nobel lectures* (2008).

[15] C. Biever, “It’s a boson! But we need to know if it’s the Higgs”. Nobel speech, 2012.

[16] M. Gell-Mann, “A Schematic Model of Baryons and Mesons”, *Phys. Lett.* **8** (1964) 214–215. doi:10.1016/S0031-9163(64)92001-3.

[17] G. Hiller, “Einführung in die Elementarteilchentheorie”. Vorlesung TU Dortmund, 2011.

[18] J.J. Sakuari, “Advanced Quantum Mechanics”. Addison Wesley, 1967.

[19] M.E. Peskin, D.V. Schroeder, “An Introduction to Quantum Field Theory”. Westview Press, 1995.

[20] “Standard Model of Elementary Paricles”. `https://cds.cern.ch/record/1473657/files/SMinfographic_image.png?subformat=`. Accessed: 2016-01-04.

[21] TOPAZ Collaboration, “Measurement of the Electromagnetic Coupling at Large Momentum Transfer”, *Phys. Rev. Lett.* **78** (Jan, 1997) 424–427. `doi:10.1103/PhysRevLett.78.424`.

[22] M. Dillig, “From Dimensional to Cut-Off Regularization”, `arXiv:nucl-th/0604062`.

[23] G. ’t Hooft and M. J. G. Veltman, “Regularization and Renormalization of Gauge Fields”, *Nucl. Phys.* **B44** (1972) 189–213. `doi:10.1016/0550-3213(72)90279-9`.

[24] J. Ward, “An identity in quantum electrodynamics”. Phys. Rev.78,182182, 1950.

[25] K. G. Wilson, “Confinement of Quarks”, *Phys. Rev.* **D10** (1974) 2445–2459. `doi:10.1103/PhysRevD.10.2445`.

[26] D. J. Gross and F. Wilczek, “Ultraviolet Behavior of Nonabelian Gauge Theories”, *Phys. Rev. Lett.* **30** (1973) 1343–1346. `doi:10.1103/PhysRevLett.30.1343`.

[27] H. D. Politzer, “Reliable Perturbative Results for Strong Interactions?”, *Phys. Rev. Lett.* **30** (1973) 1346–1349. `doi:10.1103/PhysRevLett.30.1346`.

[28] CMS Collaboration, “Measurement of the ratio of the inclusive 3-jet cross section to the inclusive 2-jet cross section in pp collisions at $\sqrt{s} = 7$ TeV and first determination of the strong coupling constant in the TeV range”, *Eur. Phys. J.* **C73** (2013), no. 10, 2604, `arXiv:1304.7498`. `doi:10.1140/epjc/s10052-013-2604-6`.

[29] D. E. Soper, “Basics of QCD perturbation theory”, in *The Strong interaction, from hadrons to partons: Proceedings, 24th SLAC Summer Institute on Particle Physics (SSI 96), Stanford, Calif., 19-30 Aug 1996.* 1996. `arXiv:hep-ph/9702203`.

[30] ZEUS, H1 Collaboration, “Combined Measurement and QCD Analysis of the Inclusive e+- p Scattering Cross Sections at HERA”, *JHEP* **01** (2010) 109, `arXiv:0911.0884`. `doi:10.1007/JHEP01(2010)109`.

[31] H1 and ZEUS Collaboration, “PDF fits including HERA-II high Q data (HERAPDF1.5)”, *H1 and ZEUS reports H1prelim-10-142, ZEUS-prel-10-018* (2013).

[32] C. S. Wu, E. Ambler, R. W. Hayward et al., “Experimental Test of Parity Conservation in Beta Decay”, *Phys. Rev.* **105** (1957) 1413–1414. `doi:10.1103/PhysRev.105.1413`.

[33] S. L. Glashow, "Partial Symmetries of Weak Interactions", *Nucl. Phys.* **22** (1961) 579–588. doi:10.1016/0029-5582(61)90469-2.

[34] A. Salam, "Weak and Electromagnetic Interactions", *Conf. Proc.* **C680519** (1968) 367–377.

[35] S. Weinberg, "A Model of Leptons", *Phys. Rev. Lett.* **19** (1967) 1264–1266. doi:10.1103/PhysRevLett.19.1264.

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

[37] P. W. Higgs, "Broken Symmetries and the Masses of Gauge Bosons", *Phys. Rev. Lett.* **13** (1964) 508–509. doi:10.1103/PhysRevLett.13.508.

[38] M. Kobayashi and T. Maskawa, "CP Violation in the Renormalizable Theory of Weak Interaction", *Prog. Theor. Phys.* **49** (1973) 652–657. doi:10.1143/PTP.49.652.

[39] "PDG CKM matrix elements". http://pdg.lbl.gov/2015/reviews/rpp2014-rev-ckm-matrix.pdf. Accessed: 2016-01-05.

[40] C. Buttar et al., "Standard Model Handles and Candles Working Group: Tools and Jets Summary Report", in *Physics at TeV colliders, La physique du TeV aux collisionneurs, Les Houches 2007 : 11-29 June 2007*, pp. 121–214. 2008. arXiv:0803.0678.

[41] G. P. Salam, "Towards Jetography", *Eur. Phys. J.* **C67** (2010) 637–686, arXiv:0906.1833. doi:10.1140/epjc/s10052-010-1314-6.

[42] A. Altheimer et al., "Jet Substructure at the Tevatron and LHC: New results, new tools, new benchmarks", *J. Phys.* **G39** (2012) 063001, arXiv:1201.0008. doi:10.1088/0954-3899/39/6/063001.

[43] CMS Collaboration, "Measurement of the Inclusive Jet Cross Section in pp Collisions at $\sqrt{s} = 7$ TeV", *Phys. Rev. Lett.* **107** (2011) 132001, arXiv:1106.0208. doi:10.1103/PhysRevLett.107.132001.

[44] ATLAS Collaboration, "Measurement of inclusive jet and dijet production in pp collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector", *Phys. Rev.* **D86** (2012) 014022, arXiv:1112.6297. doi:10.1103/PhysRevD.86.014022.

[45] M. Cacciari, "Understanding Hot & Dense QCD Matter, Jets, Lecture 1". 25th Indian-Summer School of Physics, 2013.

[46] M. Dasgupta, L. Magnea, and G. P. Salam, "Non-perturbative QCD effects in jets at hadron colliders", *JHEP* **02** (2008) 055, arXiv:0712.3014. doi:10.1088/1126-6708/2008/02/055.

[47] ATLAS, CDF, CMS, D0 Collaboration, "First combination of Tevatron and LHC measurements of the top-quark mass", arXiv:1403.4427.

[48] R. Barbieri and G. F. Giudice, "Upper Bounds on Supersymmetric Particle Masses", *Nucl. Phys.* **B306** (1988) 63. doi:10.1016/0550-3213(88)90171-X.

[49] Michal Czakon, Paul Fiedler, Alexander Mitov, "The total top quark pair production cross-section at hadron colliders through O(α_S^4)", *Phys. Rev. Lett. 110 252004* (2013) arXiv:1303.6254. doi:10.1103/PhysRevLett.110.252004.

[50] Particle Data Group Collaboration, "Review of Particle Physics (RPP)", *Phys. Rev.* **D86** (2012) 010001. doi:10.1103/PhysRevD.86.010001.

[51] C. T. Hill, "Topcolor: Top quark condensation in a gauge extension of the standard model", *Phys. Lett.* **B266** (1991) 419–424. doi:10.1016/0370-2693(91)91061-Y.

[52] C. T. Hill and S. J. Parke, "Top production: Sensitivity to new physics", *Phys. Rev.* **D49** (1994) 4454–4462, arXiv:hep-ph/9312324. doi:10.1103/PhysRevD.49.4454.

[53] C. T. Hill, "Topcolor assisted technicolor", *Phys. Lett.* **B345** (1995) 483–489, arXiv:hep-ph/9411426. doi:10.1016/0370-2693(94)01660-5.

[54] R. M. Harris and S. Jain, "Cross Sections for Leptophobic Topcolor Z' Decaying to Top-Antitop", *Eur. Phys. J.* **C72** (2012) 2072, arXiv:1112.4928. doi:10.1140/epjc/s10052-012-2072-4.

[55] J. L. Rosner, "Prominent decay modes of a leptophobic Z'", *Phys. Lett.* **B387** (1996) 113–117, arXiv:hep-ph/9607207. doi:10.1016/0370-2693(96)01022-2.

[56] K. R. Lynch, E. H. Simmons, M. Narain et al., "Finding Z' bosons coupled preferentially to the third family at LEP and the Tevatron", *Phys. Rev.* **D63** (2001) 035006, arXiv:hep-ph/0007286. doi:10.1103/PhysRevD.63.035006.

[57] M. Carena, A. Daleo, B. A. Dobrescu et al., "Z' gauge bosons at the Tevatron", *Phys. Rev.* **D70** (2004) 093009, arXiv:hep-ph/0408098. doi:10.1103/PhysRevD.70.093009.

[58] P. H. Frampton and S. L. Glashow, "Chiral Color: An Alternative to the Standard Model", *Phys. Lett.* **B190** (1987) 157. doi:10.1016/0370-2693(87)90859-8.

[59] D. Choudhury, R. M. Godbole, R. K. Singh et al., "Top production at the Tevatron/LHC and nonstandard, strongly interacting spin one particles", *Phys. Lett.* **B657** (2007) 69–76, arXiv:0705.1499. doi:10.1016/j.physletb.2007.09.057.

[60] D. Dicus, A. Stange, and S. Willenbrock, "Higgs decay to top quarks at hadron colliders", *Phys. Lett.* **B333** (1994) 126–131, arXiv:hep-ph/9404359. doi:10.1016/0370-2693(94)91017-0.

[61] L. Randall and R. Sundrum, "A Large mass hierarchy from a small extra dimension", *Phys. Rev. Lett.* **83** (1999) 3370–3373, arXiv:hep-ph/9905221. doi:10.1103/PhysRevLett.83.3370.

[62] K. Agashe, A. Belyaev, T. Krupovnickas et al., "LHC Signals from Warped Extra Dimensions", *Phys. Rev.* **D77** (2008) 015003, `arXiv:hep-ph/0612015`. `doi:10.1103/PhysRevD.77.015003`.

[63] H. Davoudiasl, J. L. Hewett, and T. G. Rizzo, "Phenomenology of the Randall-Sundrum Gauge Hierarchy Model", *Phys. Rev. Lett.* **84** (2000) 2080, `arXiv:hep-ph/9909255`. `doi:10.1103/PhysRevLett.84.2080`.

[64] CMS Collaboration, "Search for resonant $t\bar{t}$ production in proton-proton collisions at $\sqrt{s} = 8$TeV", *Phys. Rev.* **D93** (2016), no. 1, 012001, `arXiv:1506.03062`. `doi:10.1103/PhysRevD.93.012001`.

[65] M. Schmaltz and D. Tucker-Smith, "Little Higgs review", *Ann. Rev. Nucl. Part. Sci.* **55** (2005) 229–270, `arXiv:hep-ph/0502182`. `doi:10.1146/annurev.nucl.55.090704.151502`.

[66] T. Appelquist, H.-C. Cheng, and B. A. Dobrescu, "Bounds on universal extra dimensions", *Phys. Rev.* **D64** (2001) 035002, `arXiv:hep-ph/0012100`. `doi:10.1103/PhysRevD.64.035002`.

[67] CMS Collaboration, "Search for W' $\rightarrow$ tb in proton-proton collisions at $\sqrt{s} = 8$ TeV", `arXiv:1509.06051`.

[68] C. Cheung, A. L. Fitzpatrick, and L. Randall, "Sequestering CP Violation and GIM-Violation with Warped Extra Dimensions", *JHEP* **01** (2008) 069, `arXiv:0711.4421`. `doi:10.1088/1126-6708/2008/01/069`.

[69] A. L. Fitzpatrick, G. Perez, and L. Randall, "Flavor anarchy in a Randall-Sundrum model with 5D minimal flavor violation and a low Kaluza-Klein scale", *Phys. Rev. Lett.* **100** (2008) 171604, `arXiv:0710.1869`. `doi:10.1103/PhysRevLett.100.171604`.

[70] CMS Collaboration, "Search for the production of an excited bottom quark decaying to tW in proton-proton collisions at $\sqrt{s} = 8$ TeV", *JHEP* **01** (2016) 166, `arXiv:1509.08141`. `doi:10.1007/JHEP01(2016)166`.

[71] M. Beltran, D. Hooper, E. W. Kolb et al., "Maverick dark matter at colliders", *JHEP* **09** (2010) 037, `arXiv:1002.4137`. `doi:10.1007/JHEP09(2010)037`.

[72] K. Cheung, K. Mawatari, E. Senaha et al., "The Top Window for dark matter", *JHEP* **10** (2010) 081, `arXiv:1009.0618`. `doi:10.1007/JHEP10(2010)081`.

[73] J. Goodman, M. Ibe, A. Rajaraman et al., "Constraints on Dark Matter from Colliders", *Phys. Rev.* **D82** (2010) 116010, `arXiv:1008.1783`. `doi:10.1103/PhysRevD.82.116010`.

[74] CMS Collaboration, "Search for the production of dark matter in association with top-quark pairs in the single-lepton final state in proton-proton collisions at sqrt(s) = 8 TeV", *JHEP* **06** (2015) 121, `arXiv:1504.03198`. `doi:10.1007/JHEP06(2015)121`.

[75] ATLAS Collaboration, "Measurement of the differential cross-section of highly boosted top quarks as a function of their transverse momentum in $\sqrt{s} = 8$ TeV proton-proton collisions using the ATLAS detector", arXiv:1510.03818.

[76] CMS Collaboration, "Measurement of differential top-quark pair production cross sections in *pp* colisions at $\sqrt{s} = 7$ TeV", *Eur. Phys. J.* **C73** (2013), no. 3, 2339, arXiv:1211.2220. doi:10.1140/epjc/s10052-013-2339-4.

[77] CMS Collaboration Collaboration, "Measurement of the differential ttbar production cross section for high-pt top quarks in e/mu+jets final states at 8 TeV", *CMS Physics Analysis Summary* **CMS-PAS-TOP-14-012** (2015).

[78] LHC Study Group Collaboration, "The Large Hadron Collider: conceptual design", *Technical Report CERN-AC-95-05 LHC* (1995).

[79] LHCb Collaboration, "The LHCb Detector at the LHC", *JINST* **3** (2008) S08005. doi:10.1088/1748-0221/3/08/S08005.

[80] ALICE Collaboration, "The ALICE experiment at the CERN LHC", *JINST* **3** (2008) S08002. doi:10.1088/1748-0221/3/08/S08002.

[81] ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider", *JINST* **3** (2008) S08003. doi:10.1088/1748-0221/3/08/S08003.

[82] CMS Collaboration, "The CMS experiment at the CERN LHC", *JINST* **3** (2008) S08004. doi:10.1088/1748-0221/3/08/S08004.

[83] CMS Collaboration, "CMS Physics: Technical Design Report Volume I: Detector Performance and Software", *Technical Design Report Volume I CERN-LHCC-2006-001,CMS-TDR-008-1* (2006).

[84] CMS Collaboration, "CMS technical design report, volume II: Physics performance", *J.Phys. G34 995-1579* (2007). doi:10.1016/j.nima.2006.04.087.

[85] "CMS coordinate system". http://inspirehep.net/record/1294662/files/Figures_T_Coordinate.png. Accessed: 2015-12-16.

[86] CMS Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker", *JINST* **9** (2014), no. 10, P10009, arXiv:1405.6569. doi:10.1088/1748-0221/9/10/P10009.

[87] CMS Collaboration, "The CMS electromagnetic calorimeter project: Technical Design Report". Technical Design Report CMS. CERN, Geneva, 1997.

[88] P. Adzic et al., "Energy resolution of the barrel of the CMS electromagnetic calorimeter", *JINST* **2** (2007) P04004. doi:10.1088/1748-0221/2/04/P04004.

[89] CMS Collaboration Collaboration, "The CMS hadron calorimeter project: Technical Design Report". Technical Design Report CMS. CERN, Geneva, 1997.

[90] CMS Collaboration, "Performance of the CMS hadron calorimeter with cosmic ray muons and LHC beam data", *Journal of instrumentation* **5** (2010), no. 3, T03012–.

[91] CMS Collaboration Collaboration, "The CMS muon project: Technical Design Report". Technical Design Report CMS. CERN, Geneva, 1997.

[92] CMS Collaboration, "CMS TriDAS project: Technical Design Report, Volume 1: The Trigger Systems". Technical Design Report CMS.

[93] CMS Trigger and Data Acquisition Group Collaboration, "The CMS high level trigger", *Eur. Phys. J.* **C46** (2006) 605–667, `arXiv:hep-ex/0512077`. `doi:10.1140/epjc/s2006-02495-8`.

[94] A. Starodumov, W. Erdmann, R. Horisberger, H.Chr. Kaestli, D. Kotlinski, U. Langenegger, B. Meier, T. Rohe, P. Trueb, "Qualification of the CMS Barrel Pixel Detector Modules", *Nucl.Instrum.Meth. A565 67-72* (2006) `arXiv:0510165`. `doi:10.1016/j.nima.2006.04.087`.

[95] G. Lindström, "Radiation damage in silicon detectors", *Proceedings of the 9th European Symposium on Semiconductor Detectors: New Developments on Radiation Detectors* (2003).

[96] W. Erdmann, "The CMS pixel detector", *Int. J .Mod. Phys. A25 (2010) 1315* (2010).

[97] M. Waser, "Trimming of the CMS Pixel Detector Modules". Diploma thesis, ETH Zürich, 2007.

[98] CMS Collaboration, "Technical Proposal For The Upgrade Of The CMS Detector Through 2020", *Technical Report LHCC-P-004, CERN, (2011). CERN-LHCC-2011-006* (2011).

[99] CMS Collaboration, "The CMS Pixel Readout Chip for the Phase 1 Upgrade", *JINST* **10** (2015), no. 05, C05029. `doi:10.1088/1748-0221/10/05/C05029`.

[100] H. Kaestli, "Frontend electronics development for the CMS pixel detector upgrade",. `doi:10.1016/j.nima.2013.05.056`.

[101] CMS Collaboration, "CMS Technical Design Report for the Pixel Detector Upgrade", *CERN-LHCC-2012-016, CMS-TDR-011, FERMILAB-DESIGN-2012-02* (2012).

[102] CMS Collaboration, "CMS, tracker technical design report", *CERN-LHCC-98-06, CMS-TDR-5* (1998).

[103] "X-RAY DATA BOOKLET". `http://xdb.lbl.gov/`. Accessed: 2016-01-07.

[104] "Characteristic xray spectrum". `http://people.ee.ethz.ch/~cattin/MIA-ETH/01-Modalities-I-media/figs/Characteristic_Spectrum.png`. Accessed: 2015-08-09.

[105] Phywe, "Phywe x-ray box". `https://www.phywe.de/de/09057-99`. Accessed: 2015-11-13.

[106] H. C. Kastli, M. Barbero, W. Erdmann et al., "Design and performance of the CMS pixel detector readout chip", *Nucl. Instrum. Meth.* **A565** (2006) 188–194, arXiv:physics/0511166. doi:10.1016/j.nima.2006.05.038.

[107] B. Freund, "Calibration of the digital readout chip of the CMS Pixel Detector Phase I Upgrade". Diploma thesis, Karslruhe University, 2013.

[108] Matteo Centis Vignali, "Silicon sensors for the upgrades of the CMS pixel detector ". Ph.D. thesis, University of Hamburg, 2015.

[109] T. Gleisberg, S. Hoeche, F. Krauss et al, "Event generation with SHERPA 1.1", *JHEP 0902:007* (2009) arXiv:0811.4622. doi:10.1088/1126-6708/2009/02/007.

[110] M. L. Mangano, M. Moretti, F. Piccinini et al., "Matching matrix elements and shower evolution for top-quark production in hadronic collisions", *JHEP 0701:013* (2007) arXiv:0611129. doi:10.1088/1126-6708/2007/01/013.

[111] J. C. Collins, "Sudakov form-factors", *Adv. Ser. Direct. High Energy Phys.* **5** (1989) 573–614, arXiv:hep-ph/0312336. doi:10.1142/9789814503266_0006.

[112] S. Catani, F. Krauss, R. Kuhn et al., "QCD matrix elements + parton showers", *JHEP* **11** (2001) 063, arXiv:hep-ph/0109231. doi:10.1088/1126-6708/2001/11/063.

[113] V. N. Gribov and L. N. Lipatov, "Deep inelastic e p scattering in perturbation theory", *Sov. J. Nucl. Phys.* **15** (1972) 438–450. [Yad. Fiz.15,781(1972)].

[114] G. Altarelli and G. Parisi, "Asymptotic Freedom in Parton Language", *Nucl. Phys.* **B126** (1977) 298. doi:10.1016/0550-3213(77)90384-4.

[115] Y. L. Dokshitzer, "Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics.", *Sov. Phys. JETP* **46** (1977) 641–653. [Zh. Eksp. Teor. Fiz.73,1216(1977)].

[116] D. Amati and G. Veneziano, "Preconfinement as a Property of Perturbative QCD", *Phys. Lett.* **B83** (1979) 87. doi:10.1016/0370-2693(79)90896-7.

[117] A. Bassetto, M. Ciafaloni, and G. Marchesini, "Color Singlet Distributions and Mass Damping in Perturbative QCD", *Phys. Lett.* **B83** (1979) 207. doi:10.1016/0370-2693(79)90687-7.

[118] B. Andersson, G. Gustafson, G. Ingelman et al., "Parton Fragmentation and String Dynamics", *Phys. Rept.* **97** (1983) 31–145. doi:10.1016/0370-1573(83)90080-7.

[119] GEANT4 Collaboration, "GEANT4: A Simulation toolkit", *Nucl.Instrum.Meth. A506 (2003) 250-303* (2003).

[120] T. Sjöstrand, S. Mrenna, and P. Z. Skand, "PYTHIA 6.4 Physics and Manual", *JHEP 0605:026* (2006) arXiv:0603175. doi:10.1088/1126-6708/2006/05/026.

[121] M. Bähr, S. Gieseke, M. Gigg et al., “Herwig++ Physics and Manual”, *Eur.Phys.J.C58:639-707* (2008) arXiv:0803.0883. doi:10.1140/epjc/s10052-008-0798-9.

[122] C. Oleai, “The Powheg Box”, *Nucl.Phys.Proc.Suppl.205-206:36-41* (2010) arXiv:1007.3893. doi:10.1016/j.nuclphysbps.2010.08.016.

[123] J. Alwall, R. Frederix, S. Frixione et al, “The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations”, *JHEP07 079* (2014) arXiv:1405.0301. doi:10.1007/JHEP07(2014)079.

[124] S. Frixione, P. Nason, and C. Oleari, “Matching NLO QCD computations with Parton Shower simulations: the POWHEG method”, *JHEP* **11** (2007) 070, arXiv:0709.2092. doi:10.1088/1126-6708/2007/11/070.

[125] J. Pumplin, D. Stump, J. Huston et al., “New generation of parton distributions with uncertainties from global QCD analysis”, *JHEP 0207:012* (2002) arXiv:0201195. doi:10.1088/1126-6708/2002/07/012.

[126] H.-L. Lai, M. Guzzi, J. Huston et al., “New parton distributions for collider physics”, *Phys. Rev.* **D82** (2010) 074024, arXiv:1007.2241. doi:10.1103/PhysRevD.82.074024.

[127] P. M. Nadolsky, H.-L. Lai, Q.-H. Cao et al., “Implications of CTEQ global analysis for collider observables”, *Phys. Rev.* **D78** (2008) 013004, arXiv:0802.0007. doi:10.1103/PhysRevD.78.013004.

[128] A. D. Martin, R. G. Roberts, W. J. Stirling et al., “MRST2001: Partons and α_s from precise deep inelastic scattering and Tevatron jet data”, *Eur. Phys. J.* **C23** (2002) 73–87, arXiv:hep-ph/0110215. doi:10.1007/s100520100842.

[129] Z. Was, “TAUOLA the library for tau lepton decay, and KKMC KORALB KORALZ ... status report”, *Nucl.Phys.Proc.Suppl. 98 96-102* (2001) arXiv:0011305. doi:10.1016/S0920-5632(01)01200-2.

[130] W. Erdmann, “Offline Primary Vertex Reconstruction with Deterministic Annealing Clustering”, *CMS Internal Note* (2001).

[131] CMS Collaboration, “CMS slice”. https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/RetrieveFile?docid=4172&filename=CMS_Slice.gif&version=2. Accessed: 2015-11-13.

[132] CMS Collaboration, “Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET”, *CMS PAS PFT-09/001* (2009).

[133] CMS Collaboration, “Track Reconstruction in the CMS tracker”, *CMS Note* **CMS-NOTE-2006-041** (2006).

[134] CMS Collaboration, “Multivariate electron ID”. https://twiki.cern.ch/twiki/bin/viewauth/CMS/MultivariateElectronIdentification. Accessed: 2015-11-13.

[135] Johannes Haller, Roman Kogler, Tobias Lapsien, "Performance of the Shower Deconstruction and MultiR HEP TopTagger Algorithms in 8 TeV Data", *Analysis Note AN-15-001* (2015).

[136] CMS Collaboration, "MuonID". `https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideMuonId`. Accessed: 2015-11-13.

[137] Gavin P. Salam, Gregory Soyez, "A practical Seedless Infrared-Safe Cone jet algorithm", *JHEP 0705:086* (2007) `arXiv:0704.0292`. `doi:10.1088/1126-6708/2007/05/086`.

[138] Stephen D. Ellis, Davision E. Soper, "Successive Combination Jet Algorithm For Hadron Collisions", *Phys.Rev.D48:3160-3166* (1993) `arXiv:9305266`. `doi:10.1103/PhysRevD.48.3160`.

[139] Matteo Cacciari, Gavin P. Salam, Gregory Soyez, "The anti-k_t jet clustering algorithm", *JHEP 0804:063* (2008) `arXiv:0802.1189`. `doi:10.1088/1126-6708/2008/04/063`.

[140] S. Catani, Y. L. Dokshitzer, M. H. Seymour et al., "Longitudinally invariant K_t clustering algorithms for hadron hadron collisions", *Nucl. Phys.* **B406** (1993) 187–224. `doi:10.1016/0550-3213(93)90166-M`.

[141] M. Cacciari, G. P. Salam, and G. Soyez, "The Catchment Area of Jets", *JHEP* **04** (2008) 005, `arXiv:0802.1188`. `doi:10.1088/1126-6708/2008/04/005`.

[142] M. Wobisch and T. Wengler, "Hadronization corrections to jet cross-sections in deep inelastic scattering", in *Monte Carlo generators for HERA physics. Proceedings, Workshop, Hamburg, Germany, 1998-1999.* 1998. `arXiv:hep-ph/9907280`.

[143] Yu.L. Dokshitzer, G.D. Leder, S. Moretti, B.R. Webber, "Better Jet Clustering Algorithms", *JHEP 9708:001* (1997) `arXiv:9707323`. `doi:10.1088/1126-6708/1997/08/001`.

[144] M. Cacciari, "Understanding Hot & Dense QCD Matter, Jets, Lecture 3". 25th Indian-Summer School of Physics, 2013.

[145] T. Plehn and M. Spannowsky, "Top Tagging", *J. Phys.* **G39** (2012) 083001, `arXiv:1112.4441`. `doi:10.1088/0954-3899/39/8/083001`.

[146] David Krohn, Jesse Thaler, Lian-Tao Wang, "Jets with Variable R", *JHEP 0906:059* (2009) `arXiv:0903.0392`. `doi:10.1088/1126-6708/2009/06/059`.

[147] CMS Collaboration, "Plans for Jet Energy Corrections at CMS", *CMS Physics Analysis Summary* **CMS-PAS-JME-07-002** (2008).

[148] CMS Collaboration, "Jet Energy Calibration in the 8 TeV pp data ", *CMS Physics Analysis Summary* **CMS-PAS-JME-13-001** (2015).

[149] M. Cacciari and G. P. Salam, "Pileup subtraction using jet areas", *Phys. Lett.* **B659** (2008) 119–126, arXiv:0707.1378. doi:10.1016/j.physletb.2007.09.077.

[150] Henning Kirschenmann, "Jet Energy Scale Corrections and their Impact on Measurements of the Top-Quark Mass at CMS". Ph.D. thesis, University of Hamburg, 2014.

[151] J. Conway et al., "Search for BSM tt Production in the Boosted All-Hadronic Final State", *CMS Note 2011/194* (2011).

[152] CMS Collaboration, "Recommended Jet Energy Corrections and Uncertainties For Data and MC". https://twiki.cern.ch/twiki/bin/viewauth/CMS/JECDataMC. Accessed: 2016-01-08.

[153] D0 Collaboration, "High-p_T jets in $\bar{p}p$ collisions at $\sqrt{s} = 630$ GeV and 1800 GeV", *Phys. Rev.* **D64** (2001) 032003, arXiv:hep-ex/0012046. doi:10.1103/PhysRevD.64.032003.

[154] D0 Collaboration, "Measurement of the inclusive jet cross section in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV", *Phys. Rev.* **D85** (2012) 052006, arXiv:1110.3771. doi:10.1103/PhysRevD.85.052006.

[155] ATLAS Collaboration, "Jet energy resolution in proton-proton collisions at $\sqrt{s} = 7$ TeV recorded in 2010 with the ATLAS detector", *Eur. Phys. J.* **C73** (2013), no. 3, 2306, arXiv:1210.6210. doi:10.1140/epjc/s10052-013-2306-0.

[156] CMS Collaboration, "Determination of Jet Energy Calibration and Transverse Momentum Resolution in CMS", *JINST* **6** (2011) P11002, arXiv:1107.4277. doi:10.1088/1748-0221/6/11/P11002.

[157] Kristin Goebel, "Probing supersymmetry based on precise jet measurements at the CMS experiment". Ph.D. thesis, University of Hamburg, 2015.

[158] CMS Collaboration, "Missing transverse energy performance of the CMS detector", *JINST* **6** (2011) P09001, arXiv:1106.5048. doi:10.1088/1748-0221/6/09/P09001.

[159] CMS Collaboration, "A Cambridge-Aachen (C-A) based Jet Algorithm for boosted top-jet tagging", *CMS Physics Analysis Summary* **CMS-PAS-JME-09-001** (2009).

[160] CMS Collaboration, "Boosted Top Jet Tagging at CMS", *CMS Physics Analysis Summary* **CMS-PAS-JME-13-007** (2014).

[161] D. E. Kaplan, K. Rehermann, M. D. Schwartz and B. Tweed, "Top Tagging: A Method for Identifying Boosted Hadronically Decaying Top Quarks", *Phys.Rev.Lett.101:142001* (2008) arXiv:0806.0848. doi:10.1103/PhysRevLett.101.142001.

[162] CMS Collaboration, "CMS PAS JME-13-007 Public Twiki". https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsJME13007. Accessed: 2016-01-18.

[163] CMS Collaboration Collaboration, "Search for Anomalous Top Quark Pair Production in the Boosted All-Hadronic Final State using pp Collisions at sqrt(s) = 8 TeV", Technical Report CMS-PAS-B2G-12-005, CERN, Geneva, 2013.

[164] CMS Collaboration, "Search for Vector-Like Charge 2/3 T Quarks in Proton-Proton Collisions at $\sqrt{s} = 8$ TeV", arXiv:1509.04177.

[165] T. Plehn, M. Spannowsky, M. Takeuchi, and D. Zerwa, "Stop Reconstruction with Tagged Tops", *JHEP 1010:078* (2010) arXiv:1006.2833. doi:10.1007/JHEP10(2010)078.

[166] J. M. Butterworth, A. R. Davison, M. Rubin et al., "Jet substructure as a new Higgs search channel at the LHC", *Phys. Rev. Lett.* **100** (2008) 242001, arXiv:0802.2470. doi:10.1103/PhysRevLett.100.242001.

[167] G. Kasieczka, T. Plehn, T. Schell et al., "Resonance Searches with an Updated Top Tagger", *JHEP* **06** (2015) 203, arXiv:1503.05921. doi:10.1007/JHEP06(2015)203.

[168] J. Thaler and K. V. Tilbur, "Identifying Boosted Objects with N-subjettiness", *JHEP* **03** (2011) 015, arXiv:1011.2268. doi:10.1007/JHEP03(2011)015.

[169] J. Thaler and K. Van Tilbur, "Maximizing Boosted Top Identification by Minimizing N-subjettiness", arXiv:1108.2701. doi:10.1007/JHEP02(2012)093.

[170] CMS Collaboration, "Identification techniques for highly boosted W bosons that decay into hadrons", *JHEP* **12** (2014) 017, arXiv:1410.4227. doi:10.1007/JHEP12(2014)017.

[171] S. D. Ellis, C. K. Vermilion, and J. R. Wals, "Techniques for improved heavy particle searches with jet substructure", *Phys.Rev.D80:051501* (2009) arXiv:0903.5081. doi:10.1103/PhysRevD.80.051501.

[172] S. D. Ellis, C. K. Vermilion, and J. R. Wals, "Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches", *Phys.Rev.D81:094023* (2010) arXiv:0912.0033. doi:10.1103/PhysRevD.81.094023.

[173] M. Rubin, "Non-Global Logarithms in Filtered Jet Algorithms", arXiv:1002.4557. doi:10.1007/JHEP05(2010)005.

[174] D. Krohn, J. Thaler, and L.-T. Wang, "Jet Trimming", *JHEP* **02** (2010) 084, arXiv:0912.1342. doi:10.1007/JHEP02(2010)084.

[175] Andrew J. Larkoski, Simone Marzani, Gregory Soyez, Jesse Thaler, "Soft Drop", *JHEP 1405 146* (2014) arXiv:1402.2657. doi:10.1007/JHEP05(2014)146.

[176] F. Abe et al., “A Measurement of jet shapes in pp collisions at $\sqrt{s} = 1.8\,\text{TeV}$”, *Phys.Rev.Lett. 70 (1993) 713-717* (1993).

[177] J. Breitweg et al., “Measurement of jet shapes in photoproduction at HERA”, *Eur.Phys.J. C2 (1998) 61-75* (1998).

[178] S. Chatrchyan et al., “Measurement of the W-boson helicity in top-quark decays from $\mathrm{t\bar{t}}$ production in lepton+jets events in pp collisions at $\sqrt{s} = 7\,\text{TeV}$”, *JHEP 10 167* (2013) `arXiv:1308.3879`. `doi:10.1007/JHEP10(2013)167`.

[179] D. Krohn, M. D. Schwartz, T. Lin et al., “Jet Charge at the LHC”, *Phys. Rev. Lett.* **110** (2013), no. 21, 212001, `arXiv:1209.2421`. `doi:10.1103/PhysRevLett.110.212001`.

[180] S. D. Ellis, A. Hornig, D. Krohn, T. S. Roy, M. D. Schwartz, “Qjets: A Non-Deterministic Approach to Tree-Based Jet Substructure”, *PRL 108, 182003* (2012) `arXiv:1201.1914`. `doi:10.1103/PhysRevLett.108.182003`.

[181] Davison E. Soper, Michael Spannowsky, “Finding top quarks with shower deconstruction”, `arXiv:1211.3140`. `doi:10.1103/PhysRevD.87.054012`.

[182] Frank Fiedler, Alexander Grohsjean, Petra Haefner, Philipp Schieferdecker, “The Matrix Element Method and its Application to Measurements of the Top Quark Mass”, *Nucl.Instrum.Meth.A624:203-218* (2010) `arXiv:1003.1316`. `doi:10.1016/j.nima.2010.09.024`.

[183] ATLAS Collaboration, “Performance of shower deconstruction in ATLAS”, *ATLAS-CONF-2014-003* (2014).

[184] Martin Stoll, “Vetoed jet clustering: The mass-jump algorithm”, *JHEP04 111* (2015) `arXiv:1410.4637`. `doi:10.1007/JHEP04(2015)111`.

[185] CMS Collaboration, “Identification of b-quark jets with the CMS experiment”, *CMS Physics Analysis Summary* **CMS-PAS-BTV-11-001** (2011).

[186] CMS Collaboration, “Performance of b tagging at $\sqrt{s} = 8TeV$ in multijet, $t\bar{t}$ and boosted topology events”, *CMS Physics Analysis Summary* **CMS-PAS-BTV-13-001** (2013).

[187] A. B. Galtieri, F. Margaroli, and I. Volobouev, “Precision measurements of the top quark mass from the Tevatron in the pre-LHC era”, *Rept. Prog. Phys.* **75** (2012) 056201, `arXiv:1109.2163`. `doi:10.1088/0034-4885/75/5/056201`.

[188] CMS Collaboration Collaboration, “V Tagging Observables and Correlations”, *CMS Physics Analysis Summary* **CMS-PAS-JME-14-002** (2014).

[189] A. Reimers, “Optimization of top-tagging algorithms at the LHC”. Bachelor thesis, University of Hamburg, 2013.

[190] T. Dreyer, “Nutzung multivariater Analysemethoden zur Optimierung von Top-Tagging-Algorithmen”. Bachelor thesis, University of Hamburg, 2013.

[191] A. Hoecker et al., "TMVA - Toolkit for Multivariate Data Analysis", *PoS ACAT:040* (2007) arXiv:0703039.

[192] CMS Collaboration, "Boosted Top Jet Tagging at CMS", *CMS Note* **CMS-DP-2014-036** (2014).

[193] CMS Muon POG, "Reference muon id and isolation efficiencies". https://twiki.cern.ch/twiki/bin/viewauth/CMS/MuonReferenceEffs, 2012. Accessed: 2015-11-13.

[194] D. Gonzalez, "Suche nach $t\bar{t}$-Resonanzen mit dem CMS-Detektor". Master thesis, University of Hamburg, 2013.

[195] E. Usai, "Talk on Boost 2013".

[196] CMS Collaboration, "Top Tagging with New Approaches", *CMS Physics Analysis Summary* **CMS-PAS-JME-15-002** (2015).

[197] CMS Collaboration Collaboration, "8 TeV Jet Energy Corrections and Uncertainties based on 19.8 fb^{-1} of data in CMS", *CMS Note* (Oct, 2013).

[198] CMS Collaboration, "Jet Energy Resolution and Corrections (JERC)". https://twiki.cern.ch/twiki/bin/viewauth/CMS/JetEnergyScale. Accessed: 2016-02-02.

[199] CMS Collaboration, "Measurement of the differential cross section for top quark pair production in pp collisions at $\sqrt{s} = 8\,\text{TeV}$", *Eur. Phys. J.* **C75** (2015), no. 11, 542, arXiv:1505.04480. doi:10.1140/epjc/s10052-015-3709-x.

[200] CMS Collaboration, "pt(top-quark) based reweighting of ttbar MC". https://twiki.cern.ch/twiki/bin/viewauth/CMS/TopPtReweighting. Accessed: 2016-02-02.

[201] CMS Collaboration, "Search for massive resonances in dijet systems containing jets tagged as W or Z boson decays in pp collisions at $\sqrt{s} = 8$ TeV", *JHEP* **08** (2014) 173, arXiv:1405.1994. doi:10.1007/JHEP08(2014)173.

[202] T. Sjostrand, S. Mrenna, and P. Z. Skands, "PYTHIA 6.4 Physics and Manual", *JHEP* **05** (2006) 026, arXiv:hep-ph/0603175. doi:10.1088/1126-6708/2006/05/026.

[203] Sjöstrand, Torbjörn and Mrenna, Stephen and Skands, Peter Z., "A Brief Introduction to PYTHIA 8.1", *Comput.Phys.Commun.* **178** (2008) 852–867, arXiv:0710.3820. doi:10.1016/j.cpc.2008.01.036.

[204] T. Sjöstrand, S. Ask, J. R. Christiansen et al., "An Introduction to PYTHIA 8.2", *Comput. Phys. Commun.* **191** (2015) 159–177, arXiv:1410.3012. doi:10.1016/j.cpc.2015.01.024.

[205] P. Skands, S. Carrazza, and J. Rojo, "Tuning PYTHIA 8.1: the Monash 2013 Tune", *Eur. Phys. J.* **C74** (2014), no. 8, 3024, arXiv:1404.5630. doi:10.1140/epjc/s10052-014-3024-y.

[206] R. D. Ball et al., "Parton distributions with LHC data", *Nucl. Phys.* **B867** (2013) 244–289, arXiv:1207.1303. doi:10.1016/j.nuclphysb.2012.10.003.

[207] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet User Manual", *Eur. Phys. J.* **C72** (2012) 1896, arXiv:1111.6097. doi:10.1140/epjc/s10052-012-1896-2.

[208] "CMS Software". https://cmssdt.cern.ch/SDT/doxygen/index.php. Accessed: 2016-02-04.

[209] "Fastjet contribs". https://fastjet.hepforge.org/contrib/. Accessed: 2016-02-04.

[210] T. Strebler, "Substructure studies in preparation for a measurement of ttH". Master thesis, ETH Zürich, 2014.

Danksagung

Als erstes möchte ich mich bei Prof. Dr. Johannes Haller bedanken, der meine Doktorarbeit betreut hat, in wöchentlichen Meetings Resultate mit mir diskutiert hat und mir die Möglichkeit gab an Konferenzen und Schulen teilzunehmen.

Dr. Andreas Meyer möchte ich danken, der sich dazu bereit erklärt hat recht kurzfristig Zweitgutachter meiner Dissertation zu sein.

Prof. Dr. Erika Garutti möchte ich danken für die Betreuung während meiner Arbeit an dem Phase-1-Upgrade des Pixel Detektors, für viele Ideen und immer eine nette Arbeitsatmosphäre. Danke auch für die beruhigenden Worte, als das Licht während meiner Disputation ausfiel.

Der ganzen Prüfungskomission möchte ich für viele interessante Fragen während meiner Disputation danken.

Besonders möchte ich Roman Kogler danken, der mir stets mit guten Ratschlägen und großer Expertise im Bereich Top-Tagging und Substruktur geholfen hat. Weiterhin möchte ich ihm für das akribische Lesen meiner Arbeit danken.

Für das Lesen meiner Arbeit möchte ich ferner Thomas Peiffer, Jennifer Pöhlsen und Georg Steinbrück danken. Vielen Dank für die vielen nützlichen Korrekturvorschläge!

Meinen Büro-Kollegen Matteo, Jennifer, Marc und Arne möchte ich für die nette Arbeitsatmosphäre und den vielen kleinen Diskussionen danken. Weiterhin für die beliebten Runden in der Mittagspause.

Der gesamten Arbeitsgruppe am Institut möchte ich für die entspannte Arbeit und auch den ein oder anderen Freizeitaktivitäten, wie dem Fussball spielen danken.

Quisiera agradecer a Ligia, mi novia, por el maravilloso tiempo que he compartido con ella y por la motivación que me ha dado, en especial en estos últimos meses.

Danken möchte ich auch meiner Großmutter Annemarie, welche mich in der ersten Zeit meiner Promotion unterstützt hat und immer sehr Stolz war, jedoch den Tag meiner Dispuatation nicht mehr miterleben konnte.

Weiterhin möchte ich meiner Tante Angelika, meinem Onkel Walter und meiner Schwester Carina für die vielen motivierenden Worte während meiner Promotionszeit danken.

Ganz besonders möchte ich am Ende auch meinen Eltern Eva und Frank Lapsien danken, welche mich stets auf meinem Weg unterstützt und begleitet haben, oft aufmunternde und motivierende Worte für mich hatten und immer an mich gegeglaubt haben. Ohne eure Unterstützung hätte ich das nicht geschafft.